

---

# Patroni Documentation

*Release 4.0.5*

**Patroni Contributors**

**Mar 12, 2025**



## CONTENTS:

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Development Status . . . . .	3
1.2	Technical Requirements/Installation . . . . .	3
1.3	Planning the Number of PostgreSQL Nodes . . . . .	3
1.4	Running and Configuring . . . . .	3
1.5	YAML Configuration . . . . .	4
1.6	Environment Configuration . . . . .	4
1.7	Replication Choices . . . . .	4
1.8	Applications Should Not Use Superusers . . . . .	4
1.9	Testing Your HA Solution . . . . .	4
<b>2</b>	<b>Installation</b>	<b>7</b>
2.1	Pre-requirements for Mac OS . . . . .	7
2.2	Psycpg . . . . .	7
2.3	General installation for pip . . . . .	7
2.4	Package installation on Linux . . . . .	8
2.5	Upgrading . . . . .	9
<b>3</b>	<b>Patroni configuration</b>	<b>11</b>
3.1	Dynamic Configuration Settings . . . . .	11
3.2	YAML Configuration Settings . . . . .	14
3.3	Environment Configuration Settings . . . . .	27
3.4	Important rules . . . . .	37
3.5	Configuration generation and validation . . . . .	40
<b>4</b>	<b>Patroni REST API</b>	<b>43</b>
4.1	Health check endpoints . . . . .	43
4.2	Monitoring endpoint . . . . .	45
4.3	Cluster status endpoints . . . . .	50
4.4	Config endpoint . . . . .	52
4.5	Switchover and failover endpoints . . . . .	54
4.6	Restart endpoint . . . . .	56
4.7	Reload endpoint . . . . .	56
4.8	Reinitialize endpoint . . . . .	57
<b>5</b>	<b>patronictl</b>	<b>59</b>
5.1	Configuration . . . . .	59
5.2	Usage . . . . .	60
<b>6</b>	<b>Replica imaging and bootstrap</b>	<b>93</b>
6.1	Bootstrap . . . . .	93

6.2	Building replicas . . . . .	94
<b>7</b>	<b>Replication modes</b>	<b>97</b>
7.1	Asynchronous mode durability . . . . .	97
7.2	PostgreSQL synchronous replication . . . . .	97
7.3	Synchronous mode . . . . .	98
7.4	Synchronous Replication Factor . . . . .	99
7.5	Maximum lag on synchronous node . . . . .	99
7.6	Synchronous mode implementation . . . . .	99
7.7	Quorum commit mode . . . . .	100
<b>8</b>	<b>Standby cluster</b>	<b>103</b>
<b>9</b>	<b>Watchdog support</b>	<b>105</b>
9.1	Setting up software watchdog on Linux . . . . .	106
<b>10</b>	<b>Pause/Resume mode for the cluster</b>	<b>107</b>
10.1	The goal . . . . .	107
10.2	The implementation . . . . .	107
10.3	User guide . . . . .	108
<b>11</b>	<b>DCS Failsafe Mode</b>	<b>109</b>
11.1	The problem . . . . .	109
11.2	Reasons for the current implementation . . . . .	109
11.3	DCS Failsafe Mode . . . . .	109
11.4	Low-level implementation details . . . . .	110
11.5	F.A.Q. . . . .	110
<b>12</b>	<b>Using Patroni with Kubernetes</b>	<b>111</b>
12.1	Use Endpoints . . . . .	111
12.2	Use ConfigMaps . . . . .	111
12.3	Configuration . . . . .	111
12.4	Examples . . . . .	112
<b>13</b>	<b>Citus support</b>	<b>113</b>
13.1	TL;DR . . . . .	113
13.2	patronictl . . . . .	114
13.3	Citus worker switchover . . . . .	115
13.4	Secondary nodes . . . . .	116
13.5	Peek into DCS . . . . .	116
13.6	Citus on Kubernetes . . . . .	117
13.7	Citus upgrades and PostgreSQL major upgrades . . . . .	119
<b>14</b>	<b>Convert a Standalone to a Patroni Cluster</b>	<b>121</b>
14.1	Procedure . . . . .	121
<b>15</b>	<b>Major Upgrade of PostgreSQL Version</b>	<b>123</b>
15.1	FAQ . . . . .	123
<b>16</b>	<b>Integration with other tools</b>	<b>125</b>
16.1	Barman . . . . .	125
<b>17</b>	<b>Security Considerations</b>	<b>127</b>
17.1	Protecting DCS . . . . .	127
17.2	Protecting the REST API . . . . .	127

<b>18 HA multi datacenter</b>	<b>129</b>
18.1 Synchronous Replication . . . . .	129
18.2 Asynchronous Replication . . . . .	130
<b>19 FAQ</b>	<b>131</b>
19.1 Comparison with other HA solutions . . . . .	131
19.2 DCS . . . . .	132
19.3 patronictl . . . . .	133
19.4 Configuration . . . . .	133
19.5 Postgres management . . . . .	136
19.6 Concepts and requirements . . . . .	136
19.7 Automatic failover . . . . .	137
19.8 Bootstrapping and standbys creation . . . . .	138
19.9 Monitoring . . . . .	138
<b>20 Release notes</b>	<b>139</b>
20.1 Version 4.0.5 . . . . .	139
20.2 Version 4.0.4 . . . . .	140
20.3 Version 4.0.3 . . . . .	141
20.4 Version 4.0.2 . . . . .	141
20.5 Version 4.0.1 . . . . .	142
20.6 Version 4.0.0 . . . . .	142
20.7 Version 3.3.2 . . . . .	144
20.8 Version 3.3.1 . . . . .	144
20.9 Version 3.3.0 . . . . .	145
20.10 Version 3.2.2 . . . . .	146
20.11 Version 3.2.1 . . . . .	147
20.12 Version 3.2.0 . . . . .	148
20.13 Version 3.1.2 . . . . .	149
20.14 Version 3.1.1 . . . . .	149
20.15 Version 3.1.0 . . . . .	150
20.16 Version 3.0.4 . . . . .	152
20.17 Version 3.0.3 . . . . .	153
20.18 Version 3.0.2 . . . . .	154
20.19 Version 3.0.1 . . . . .	155
20.20 Version 3.0.0 . . . . .	155
20.21 Version 2.1.7 . . . . .	156
20.22 Version 2.1.6 . . . . .	156
20.23 Version 2.1.5 . . . . .	157
20.24 Version 2.1.4 . . . . .	159
20.25 Version 2.1.3 . . . . .	160
20.26 Version 2.1.2 . . . . .	161
20.27 Version 2.1.1 . . . . .	163
20.28 Version 2.1.0 . . . . .	164
20.29 Version 2.0.2 . . . . .	166
20.30 Version 2.0.1 . . . . .	167
20.31 Version 2.0.0 . . . . .	168
20.32 Version 1.6.5 . . . . .	172
20.33 Version 1.6.4 . . . . .	175
20.34 Version 1.6.3 . . . . .	176
20.35 Version 1.6.2 . . . . .	176
20.36 Version 1.6.1 . . . . .	177
20.37 Version 1.6.0 . . . . .	179
20.38 Version 1.5.6 . . . . .	181

20.39 Version 1.5.5 . . . . .	182
20.40 Version 1.5.4 . . . . .	183
20.41 Version 1.5.3 . . . . .	184
20.42 Version 1.5.2 . . . . .	185
20.43 Version 1.5.1 . . . . .	185
20.44 Version 1.5.0 . . . . .	185
20.45 Version 1.4.6 . . . . .	187
20.46 Version 1.4.5 . . . . .	187
20.47 Version 1.4.4 . . . . .	188
20.48 Version 1.4.3 . . . . .	189
20.49 Version 1.4.2 . . . . .	190
20.50 Version 1.4.1 . . . . .	191
20.51 Version 1.4 . . . . .	191
20.52 Version 1.3.6 . . . . .	193
20.53 Version 1.3.5 . . . . .	193
20.54 Version 1.3.4 . . . . .	194
20.55 Version 1.3.3 . . . . .	195
20.56 Version 1.3.2 . . . . .	195
20.57 Version 1.3.1 . . . . .	195
20.58 Version 1.3 . . . . .	195
20.59 Version 1.2 . . . . .	197
20.60 Version 1.1 . . . . .	199
20.61 Version 1.0 . . . . .	201
20.62 Version 0.90 . . . . .	203
20.63 Version 0.80 . . . . .	204
<b>21 Contributing</b>	<b>207</b>
21.1 Contributing guidelines . . . . .	207
<b>22 Indices and tables</b>	<b>211</b>

Patroni is a template for high availability (HA) PostgreSQL solutions using Python. For maximum accessibility, Patroni supports a variety of distributed configuration stores like [ZooKeeper](#), [etcd](#), [Consul](#) or [Kubernetes](#). Database engineers, DBAs, DevOps engineers, and SREs who are looking to quickly deploy HA PostgreSQL in datacenters — or anywhere else — will hopefully find it useful.

We call Patroni a “template” because it is far from being a one-size-fits-all or plug-and-play replication system. It will have its own caveats. Use wisely. There are many ways to run high availability with PostgreSQL; for a list, see the [PostgreSQL Documentation](#).

Currently supported PostgreSQL versions: 9.3 to 17.

**Note to Citus users:** Starting from 3.0 Patroni nicely integrates with the [Citus](#) database extension to Postgres. Please check the [Citus support page](#) in the Patroni documentation for more info about how to use Patroni high availability together with a Citus distributed cluster.

**Note to Kubernetes users:** Patroni can run natively on top of Kubernetes. Take a look at the [Kubernetes](#) chapter of the Patroni documentation.





## INTRODUCTION

Patroni is a template for high availability (HA) PostgreSQL solutions using Python. Patroni originated as a fork of [Governor](#), the project from Compose. It includes plenty of new features.

For additional background info, see:

- [PostgreSQL HA with Kubernetes and Patroni](#), talk by Josh Berkus at KubeCon 2016 (video)
- [Feb. 2016 Zalando Tech blog post](#)

### 1.1 Development Status

Patroni is in active development and accepts contributions. See our [Contributing](#) section below for more details.

We report new releases information [here](#).

### 1.2 Technical Requirements/Installation

Go [here](#) for guidance on installing and upgrading Patroni on various platforms.

### 1.3 Planning the Number of PostgreSQL Nodes

Patroni/PostgreSQL nodes are decoupled from DCS nodes (except when Patroni implements RAFT on its own) and therefore there is no requirement on the minimal number of nodes. Running a cluster consisting of one primary and one standby is perfectly fine. You can add more standby nodes later.

### 1.4 Running and Configuring

The following section assumes Patroni repository as being cloned from <https://github.com/patroni/patroni>. Namely, you will need example configuration files `postgres0.yml` and `postgres1.yml`. If you installed Patroni with pip, you can obtain those files from the git repository and replace `./patroni.py` below with `patroni` command.

To get started, do the following from different terminals:

```
> etcd --data-dir=data/etcd --enable-v2=true
> ./patroni.py postgres0.yml
> ./patroni.py postgres1.yml
```

You will then see a high-availability cluster start up. Test different settings in the YAML files to see how the cluster's behavior changes. Kill some of the components to see how the system behaves.

Add more `postgres*.yaml` files to create an even larger cluster.

Patroni provides an [HAProxy](#) configuration, which will give your application a single endpoint for connecting to the cluster's leader. To configure, run:

```
> haproxy -f haproxy.cfg
```

```
> psql --host 127.0.0.1 --port 5000 postgres
```

## 1.5 YAML Configuration

Go [here](#) for comprehensive information about settings for etcd, consul, and ZooKeeper. And for an example, see `postgres0.yaml`.

## 1.6 Environment Configuration

Go [here](#) for comprehensive information about configuring(overriding) settings via environment variables.

## 1.7 Replication Choices

Patroni uses Postgres' streaming replication, which is asynchronous by default. Patroni's asynchronous replication configuration allows for `maximum_lag_on_failover` settings. This setting ensures failover will not occur if a follower is more than a certain number of bytes behind the leader. This setting should be increased or decreased based on business requirements. It's also possible to use synchronous replication for better durability guarantees. See [replication modes documentation](#) for details.

## 1.8 Applications Should Not Use Superusers

When connecting from an application, always use a non-superuser. Patroni requires access to the database to function properly. By using a superuser from an application, you can potentially use the entire connection pool, including the connections reserved for superusers, with the `superuser_reserved_connections` setting. If Patroni cannot access the Primary because the connection pool is full, behavior will be undesirable.

## 1.9 Testing Your HA Solution

Testing an HA solution is a time consuming process, with many variables. This is particularly true considering a cross-platform application. You need a trained system administrator or a consultant to do this work. It is not something we can cover in depth in the documentation.

That said, here are some pieces of your infrastructure you should be sure to test:

- Network (the network in front of your system as well as the NICs [physical or virtual] themselves)
- Disk IO

- file limits (nofile in Linux)
- RAM. Even if you have oomkiller turned off, the unavailability of RAM could cause issues.
- CPU
- Virtualization Contention (overcommitting the hypervisor)
- Any cgroup limitation (likely to be related to the above)
- `kill -9` of any postgres process (except postmaster!). This is a decent simulation of a segfault.

One thing that you should not do is run `kill -9` on a postmaster process. This is because doing so does not mimic any real life scenario. If you are concerned your infrastructure is insecure and an attacker could run `kill -9`, no amount of HA process is going to fix that. The attacker will simply kill the process again, or cause chaos in another way.



## INSTALLATION

### 2.1 Pre-requirements for Mac OS

To install requirements on a Mac, run the following:

```
brew install postgresql etcd haproxy libyaml python
```

### 2.2 Psycopg

Starting from [psycopg2-2.8](#) the binary version of `psycopg2` will no longer be installed by default. Installing it from the source code requires C compiler and postgres+python dev packages. Since in the python world it is not possible to specify dependency as `psycopg2` OR `psycopg2-binary` you will have to decide how to install it.

There are a few options available:

1. Use the package manager from your distro

```
sudo apt-get install python3-psycopg2 # install psycopg2 module on Debian/Ubuntu
sudo yum install python3-psycopg2    # install psycopg2 on RedHat/Fedora/CentOS
```

2. Specify one of `psycopg`, `psycopg2`, or `psycopg2-binary` in the *list of dependencies* when installing Patroni with pip.

### 2.3 General installation for pip

Patroni can be installed with pip:

```
pip install patroni[dependencies]
```

where *dependencies* can be either empty, or consist of one or more of the following:

**etcd or etcd3**

*python-etcd* module in order to use Etcd as Distributed Configuration Store (DCS)

**consul**

*py-consul* module in order to use Consul as DCS

**zookeeper**

*kazoo* module in order to use Zookeeper as DCS

**exhibitor**

*kazoo* module in order to use Exhibitor as DCS (same dependencies as for Zookeeper)

**kubernetes**

*kubernetes* module in order to use Kubernetes as DCS in Patroni

**raft**

*pysyncobj* module in order to use python Raft implementation as DCS

**aws**

*boto3* in order to use AWS callbacks

**jsonlogger**

*python-json-logger* module in order to enable *logging* in json format

**all**

all of the above (except psycopg family)

**psycopg**

*psycopg[binary]>=3.0.0* module

**psycopg2**

*psycopg2>=2.5.4* module

**psycopg2-binary**

*psycopg2-binary* module

For example, the command in order to install Patroni together with psycopg3, dependencies for Etcd as a DCS, and AWS callbacks is:

```
pip install patroni[psycopg3,etcd3,aws]
```

Note that external tools to call in the replica creation or custom bootstrap scripts (i.e. WAL-E) should be installed independently of Patroni.

## 2.4 Package installation on Linux

Patroni packages may be available for your operating system, produced by the Postgres community for:

- RHEL, RockyLinux, AlmaLinux;
- Debian and Ubuntu;
- SUSE Enterprise Linux.

You can also find packages for direct dependencies of Patroni, like python modules that might not be available in the official operating system repositories.

For more information see the [PGDG repository](#) documentation.

If you are on a RedHat Enterprise Linux derivative operating system you may also require packages from EPEL, see [EPEL repository](#) documentation.

Once you have installed the PGDG repository for your OS you can install patroni.

**Note**

Patroni packages are not maintained by the Patroni developers, but rather by the Postgres community. If you require support please first try connecting on [Postgres slack](#).

### 2.4.1 Installing on Debian derivatives

With PGDG repo installed, see [above](#), install Patroni via apt run:

```
apt-get install patroni
```

### 2.4.2 Installing on RedHat derivatives

With PGDG repo installed, see [above](#), install patroni with an etcd DCS via dnf on RHEL 9 (and derivatives) run:

```
dnf install patroni patroni-etcd
```

You can install etcd from PGDG if your RedHat derivative distribution does not provide packages. On the nodes that will host the DCS run:

```
dnf install 'dnf-command(config-manager)'
dnf config-manager --enable pgdg-rhel9-extras
dnf install etcd
```

You can replace the version of RHEL with 8 in the repo to make *pgdg-rhel8-extras* if needed. The repo name is still *pgdg-rhelN-extras* on RockyLinux, AlmaLinux, Oracle Linux, etc...

### 2.4.3 Installing on SUSE Enterprise Linux

You might need to enable the SUSE PackageHub repositories for some dependencies. see [SUSE PackageHub](#) documentation.

For SLES 15 with PGDG repo installed, see [above](#), you can install patroni using:

```
zypper install patroni patroni-etcd
```

With the SUSE PackageHub repo enabled you can also install etcd:

```
SUSEConnect -p PackageHub/15.5/x86_64
zypper install etcd
```

## 2.5 Upgrading

Upgrading patroni is a very simple process, just update the software installation and restart the Patroni daemon on each node in the cluster.

However, restarting the Patroni daemon will result in a Postgres database restart. In some situations this may cause a failover of the primary node in your cluster, therefore it is recommended to put the cluster into maintenance mode until the Patroni daemon restart has been completed.

To put the cluster in maintenance mode, run the following command on one of the patroni nodes:

```
patronictl pause --wait
```

Then on each node in the cluster, perform the package upgrade required for your OS:

```
apt-get update && apt-get install patroni patroni-etcd
```

Restart the patroni daemon process on each node:

```
systemctl restart patroni
```

Then finally resume monitoring of Postgres with patroni to take it out of maintenance mode:

```
patronictl resume --wait
```

The cluster will now be full operational with the new version of Patroni.



## PATRONI CONFIGURATION

### 3.1 Dynamic Configuration Settings

Dynamic configuration is stored in the DCS (Distributed Configuration Store) and applied on all cluster nodes.

In order to change the dynamic configuration you can use either *patronictl edit-config* tool or Patroni *REST API*.

- **loop\_wait**: the number of seconds the loop will sleep. Default value: 10, minimum possible value: 1
- **ttl**: the TTL to acquire the leader lock (in seconds). Think of it as the length of time before initiation of the automatic failover process. Default value: 30, minimum possible value: 20
- **retry\_timeout**: timeout for DCS and PostgreSQL operation retries (in seconds). DCS or network issues shorter than this will not cause Patroni to demote the leader. Default value: 10, minimum possible value: 3

#### Warning

when changing values of **loop\_wait**, **retry\_timeout**, or **ttl** you have to follow the rule:

```
loop_wait + 2 * retry_timeout <= ttl
```

- **maximum\_lag\_on\_failover**: the maximum bytes a follower may lag to be able to participate in leader election.
- **maximum\_lag\_on\_syncnode**: the maximum bytes a synchronous follower may lag before it is considered as an unhealthy candidate and swapped by healthy asynchronous follower. Patroni utilize the max replica lsn if there is more than one follower, otherwise it will use leader's current wal lsn. Default is -1, Patroni will not take action to swap synchronous unhealthy follower when the value is set to 0 or below. Please set the value high enough so Patroni won't swap synchronous follower frequently during high transaction volume.
- **max\_timelines\_history**: maximum number of timeline history items kept in DCS. Default value: 0. When set to 0, it keeps the full history in DCS.
- **primary\_start\_timeout**: the amount of time a primary is allowed to recover from failures before failover is triggered (in seconds). Default is 300 seconds. When set to 0 failover is done immediately after a crash is detected if possible. When using asynchronous replication a failover can cause lost transactions. Worst case failover time for primary failure is:  $\text{loop\_wait} + \text{primary\_start\_timeout} + \text{loop\_wait}$ , unless **primary\_start\_timeout** is zero, in which case it's just **loop\_wait**. Set the value according to your durability/availability tradeoff.
- **primary\_stop\_timeout**: The number of seconds Patroni is allowed to wait when stopping Postgres and effective only when **synchronous\_mode** is enabled. When set to  $> 0$  and the **synchronous\_mode** is enabled, Patroni sends SIGKILL to the postmaster if the stop operation is running for more than the value set by **primary\_stop\_timeout**. Set the value according to your durability/availability tradeoff. If the parameter is not set or set  $\leq 0$ , **primary\_stop\_timeout** does not apply.

- **synchronous\_mode**: turns on synchronous replication mode. Possible values: `off`, `on`, `quorum`. In this mode the leader takes care of management of `synchronous_standby_names`, and only the last known leader, or one of synchronous replicas, are allowed to participate in leader race. Synchronous mode makes sure that successfully committed transactions will not be lost at failover, at the cost of losing availability for writes when Patroni cannot ensure transaction durability. See [replication modes documentation](#) for details.
- **synchronous\_mode\_strict**: prevents disabling synchronous replication if no synchronous replicas are available, blocking all client writes to the primary. See [replication modes documentation](#) for details.
- **synchronous\_node\_count**: if `synchronous_mode` is enabled, this parameter is used by Patroni to manage the precise number of synchronous standby instances and adjusts the state in DCS and the `synchronous_standby_names` parameter in PostgreSQL as members join and leave. If the parameter is set to a value higher than the number of eligible nodes, it will be automatically adjusted. Defaults to 1.
- **failsafe\_mode**: Enables [DCS Failsafe Mode](#). Defaults to `false`.
- **postgresql**:
  - **use\_pg\_rewind**: whether or not to use `pg_rewind`. Defaults to `false`. Note that either the cluster must be initialized with data page checksums (`--data-checksums` option for `initdb`) and/or `wal_log_hints` must be set to `on`, or `pg_rewind` will not work.
  - **use\_slots**: whether or not to use replication slots. Defaults to `true` on PostgreSQL 9.4+.
  - **recovery\_conf**: additional configuration settings written to `recovery.conf` when configuring follower. There is no `recovery.conf` anymore in PostgreSQL 12, but you may continue using this section, because Patroni handles it transparently.
  - **parameters**: configuration parameters (GUCs) for Postgres in format `{max_connections: 100, wal_level: "replica", max_wal_senders: 10, wal_log_hints: "on"}`. Many of these are required for replication to work.
  - **pg\_hba**: list of lines that Patroni will use to generate `pg_hba.conf`. Patroni ignores this parameter if `hba_file` PostgreSQL parameter is set to a non-default value.
    - \* - **host all all 0.0.0.0/0 md5**
    - \* - **host replication replicator 127.0.0.1/32 md5**: A line like this is required for replication.
  - **pg\_ident**: list of lines that Patroni will use to generate `pg_ident.conf`. Patroni ignores this parameter if `ident_file` PostgreSQL parameter is set to a non-default value.
    - \* - **mapname1 systemname1 pguser1**
    - \* - **mapname1 systemname2 pguser2**
- **standby\_cluster**: if this section is defined, we want to bootstrap a standby cluster.
  - **host**: an address of remote node
  - **port**: a port of remote node
  - **primary\_slot\_name**: which slot on the remote node to use for replication. This parameter is optional, the default value is derived from the instance name (see function `slot_name_from_member_name`).
  - **create\_replica\_methods**: an ordered list of methods that can be used to bootstrap standby leader from the remote primary, can be different from the list defined in [PostgreSQL](#)
  - **restore\_command**: command to restore WAL records from the remote primary to nodes in a standby cluster, can be different from the list defined in [PostgreSQL](#)
  - **archive\_cleanup\_command**: cleanup command for standby leader
  - **recovery\_min\_apply\_delay**: how long to wait before actually apply WAL records on a standby leader

- **member\_slots\_ttl**: retention time of physical replication slots for replicas when they are shut down. Default value: *30min*. Set it to *0* if you want to keep the old behavior (when the member key expires from DCS, the slot is immediately removed). The feature works only starting from PostgreSQL 11.
- **slots**: define permanent replication slots. These slots will be preserved during switchover/failover. Permanent slots that don't exist will be created by Patroni. With PostgreSQL 11 onwards permanent physical slots are created on all nodes and their position is advanced every **loop\_wait** seconds. For PostgreSQL versions older than 11 permanent physical replication slots are maintained only on the current primary. The logical slots are copied from the primary to a standby with restart, and after that their position advanced every **loop\_wait** seconds (if necessary). Copying logical slot files performed via **libpq** connection and using either rewind or superuser credentials (see **postgresql.authentication** section). There is always a chance that the logical slot position on the replica is a bit behind the former primary, therefore application should be prepared that some messages could be received the second time after the failover. The easiest way of doing so - tracking **confirmed\_flush\_lsn**. Enabling permanent replication slots requires **postgresql.use\_slots** to be set to **true**. If there are permanent logical replication slots defined Patroni will automatically enable the **hot\_standby\_feedback**. Since the failover of logical replication slots is unsafe on PostgreSQL 9.6 and older and PostgreSQL version 10 is missing some important functions, the feature only works with PostgreSQL 11+.
  - **my\_slot\_name**: the name of the permanent replication slot. If the permanent slot name matches with the name of the current node it will not be created on this node. If you add a permanent physical replication slot which name matches the name of a Patroni member, Patroni will ensure that the slot that was created is not removed even if the corresponding member becomes unresponsive, situation which would normally result in the slot's removal by Patroni. Although this can be useful in some situations, such as when you want replication slots used by members to persist during temporary failures or when importing existing members to a new Patroni cluster (see *Convert a Standalone to a Patroni Cluster* for details), caution should be exercised by the operator that these clashes in names are not persisted in the DCS, when the slot is no longer required, due to its effect on normal functioning of Patroni.
    - \* **type**: slot type. Could be **physical** or **logical**. If the slot is logical, you have to additionally define database and plugin.
    - \* **database**: the database name where logical slots should be created.
    - \* **plugin**: the plugin name for the logical slot.
- **ignore\_slots**: list of sets of replication slot properties for which Patroni should ignore matching slots. This configuration/feature/etc. is useful when some replication slots are managed outside of Patroni. Any subset of matching properties will cause a slot to be ignored.
  - **name**: the name of the replication slot.
  - **type**: slot type. Can be **physical** or **logical**. If the slot is logical, you may additionally define database and/or plugin.
  - **database**: the database name (when matching a logical slot).
  - **plugin**: the logical decoding plugin (when matching a logical slot).

Note: **slots** is a hashmap while **ignore\_slots** is an array. For example:

```
slots:
  permanent_logical_slot_name:
    type: logical
    database: my_db
    plugin: test_decoding
  permanent_physical_slot_name:
    type: physical
  ...
ignore_slots:
```

(continues on next page)

(continued from previous page)

```
- name: ignored_logical_slot_name
  type: logical
  database: my_db
  plugin: test_decoding
- name: ignored_physical_slot_name
  type: physical
...
```

Note: When running PostgreSQL v11 or newer Patroni maintains physical replication slots on all nodes that could potentially become a leader, so that replica nodes keep WAL segments reserved if they are potentially required by other nodes. In case the node is absent and its member key in DCS gets expired, the corresponding replication slot is dropped after `member_slots_ttl` (default value is *30min*). You can increase or decrease retention based on your needs. Alternatively, if your cluster topology is static (fixed number of nodes that never change their names) you can configure permanent physical replication slots with names corresponding to the names of the nodes to avoid slots removal and recycling of WAL files while replica is temporarily down:

```
slots:
  node_name1:
    type: physical
  node_name2:
    type: physical
  node_name3:
    type: physical
...
```

#### Warning

Permanent replication slots are synchronized only from the `primary/standby_leader` to replica nodes. That means, applications are supposed to be using them only from the leader node. Using them on replica nodes will cause indefinite growth of `pg_wal` on all other nodes in the cluster. An exception to that rule are physical slots that match the Patroni member names (created and maintained by Patroni). Those will be synchronized among all nodes as they are used for replication among them.

#### Warning

Setting `nostream` tag on standby disables copying and synchronization of permanent logical replication slots on the node itself and all its cascading replicas if any.

## 3.2 YAML Configuration Settings

### 3.2.1 Global/Universal

- **name:** the name of the host. Must be unique for the cluster.
- **namespace:** path within the configuration store where Patroni will keep information about the cluster. Default value: `"/service"`
- **scope:** cluster name

### 3.2.2 Log

- **type**: sets the format of logs. Can be either **plain** or **json**. To use **json** format, you must have the *jsonlogger* installed. The default value is **plain**.
- **level**: sets the general logging level. Default value is **INFO** (see [the docs for Python logging](#))
- **traceback\_level**: sets the level where tracebacks will be visible. Default value is **ERROR**. Set it to **DEBUG** if you want to see tracebacks only if you enable **log.level=DEBUG**.
- **format**: sets the log formatting string. If the log type is **plain**, the log format should be a string. Refer to [the LogRecord attributes](#) for available attributes. If the log type is **json**, the log format can be a list in addition to a string. Each list item should correspond to LogRecord attributes. Be cautious that only the field name is required, and the `%()` should be omitted. If you wish to print a log field with a different key name, use a dictionary where the dictionary key is the log field, and the value is the name of the field you want to be printed in the log. Default value is `%(asctime)s %(levelname)s: %(message)s`
- **dateformat**: sets the datetime formatting string. (see the [formatTime\(\) documentation](#))
- **static\_fields**: add additional fields to the log. This option is only available when the log type is set to **json**.
- **max\_queue\_size**: Patroni is using two-step logging. Log records are written into the in-memory queue and there is a separate thread which pulls them from the queue and writes to stderr or file. The maximum size of the internal queue is limited by default by **1000** records, which is enough to keep logs for the past 1h20m.
- **dir**: Directory to write application logs to. The directory must exist and be writable by the user executing Patroni. If you set this value, the application will retain 4 25MB logs by default. You can tune those retention values with *file\_num* and *file\_size* (see below).
- **mode**: Permissions for log files (for example, `0644`). If not specified, permissions will be set based on the current umask value.
- **file\_num**: The number of application logs to retain.
- **file\_size**: Size of patroni.log file (in bytes) that triggers a log rolling.
- **loggers**: This section allows redefining logging level per python module
  - **patroni.postmaster**: **WARNING**
  - **urllib3**: **DEBUG**

Here is an example of how to config patroni to log in json format.

```
log:
  type: json
  format:
    - message
    - module
    - asctime: '@timestamp'
    - levelname: level
  static_fields:
    app: patroni
```

### 3.2.3 Bootstrap configuration

#### **i** Note

Once Patroni has initialized the cluster for the first time and settings have been stored in the DCS, all future changes to the `bootstrap.dcs` section of the YAML configuration will not take any effect! If you want to change them please use either *patronictl edit-config* or the Patroni *REST API*.

- **bootstrap:**

- **dcs:** This section will be written into `/<namespace>/<scope>/config` of the given configuration store after initializing the new cluster. The global dynamic configuration for the cluster. You can put any of the parameters described in the *Dynamic Configuration settings* under `bootstrap.dcs` and after Patroni has initialized (bootstrapped) the new cluster, it will write this section into `/<namespace>/<scope>/config` of the configuration store.
- **method:** custom script to use for bootstrapping this cluster.  
See *custom bootstrap methods documentation* for details. When `initdb` is specified revert to the default `initdb` command. `initdb` is also triggered when no `method` parameter is present in the configuration file.
- **initdb:** (optional) list options to be passed on to `initdb`.
  - \* - **data-checksums:** Must be enabled when `pg_rewind` is needed on 9.3.
  - \* - **encoding: UTF8:** default encoding for new databases.
  - \* - **locale: UTF8:** default locale for new databases.
- **post\_bootstrap** or **post\_init:** An additional script that will be executed after initializing the cluster. The script receives a connection string URL (with the cluster superuser as a user name). The `PGPASSFILE` variable is set to the location of `pgpass` file.

### 3.2.4 Citus

Enables integration Patroni with *Citus*. If configured, Patroni will take care of registering Citus worker nodes on the coordinator. You can find more information about Citus support *here*.

- **group:** the Citus group id, integer. Use `0` for coordinator and `1, 2, etc...` for workers
- **database:** the database where `citus` extension should be created. Must be the same on the coordinator and all workers. Currently only one database is supported.

### 3.2.5 Consul

Most of the parameters are optional, but you have to specify one of the **host** or **url**

- **host:** the host:port for the Consul local agent.
- **url:** url for the Consul local agent, in format: `http(s)://host:port`.
- **port:** (optional) Consul port.
- **scheme:** (optional) **http** or **https**, defaults to **http**.
- **token:** (optional) ACL token.
- **verify:** (optional) whether to verify the SSL certificate for HTTPS requests.

- **cacert**: (optional) The ca certificate. If present it will enable validation.
- **cert**: (optional) file with the client certificate.
- **key**: (optional) file with the client key. Can be empty if the key is part of **cert**.
- **dc**: (optional) Datacenter to communicate with. By default the datacenter of the host is used.
- **consistency**: (optional) Select consul consistency mode. Possible values are default, consistent, or stale (more details in [consul API reference](#))
- **checks**: (optional) list of Consul health checks used for the session. By default an empty list is used.
- **register\_service**: (optional) whether or not to register a service with the name defined by the scope parameter and the tag master, primary, replica, or standby-leader depending on the node's role. Defaults to **false**.
- **service\_tags**: (optional) additional static tags to add to the Consul service apart from the role (primary/replica/standby-leader). By default an empty list is used.
- **service\_check\_interval**: (optional) how often to perform health check against registered url. Defaults to '5s'.
- **service\_check\_tls\_server\_name**: (optional) override SNI host when connecting via TLS, see also [consul agent check API reference](#).

The token needs to have the following ACL permissions:

```
service_prefix "${scope}" {
    policy = "write"
}
key_prefix "${namespace}/${scope}" {
    policy = "write"
}
session_prefix "" {
    policy = "write"
}
```

### 3.2.6 Etcd

Most of the parameters are optional, but you have to specify one of the **host**, **hosts**, **url**, **proxy** or **srv**

- **host**: the host:port for the etcd endpoint.
- **hosts**: list of etcd endpoint in format host1:port1,host2:port2,etc... Could be a comma separated string or an actual yaml list.
- **use\_proxies**: If this parameter is set to true, Patroni will consider **hosts** as a list of proxies and will not perform a topology discovery of etcd cluster.
- **url**: url for the etcd.
- **proxy**: proxy url for the etcd. If you are connecting to the etcd using proxy, use this parameter instead of **url**.
- **srv**: Domain to search the SRV record(s) for cluster autodiscovery. Patroni will try to query these SRV service names for specified domain (in that order until first success): `_etcd-client-ssl`, `_etcd-client`, `_etcd-ssl`, `_etcd`, `_etcd-server-ssl`, `_etcd-server`. If SRV records for `_etcd-server-ssl` or `_etcd-server` are retrieved then ETCD peer protocol is used to query ETCD for available members. Otherwise hosts from SRV records will be used.
- **srv\_suffix**: Configures a suffix to the SRV name that is queried during discovery. Use this flag to differentiate between multiple etcd clusters under the same domain. Works only with conjunction with **srv**.

For example, if `srv_suffix: foo` and `srv: example.org` are set, the following DNS SRV query is made: `_etcd-client-ssl-foo._tcp.example.com` (and so on for every possible ETCD SRV service name).

- **protocol:** (optional) http or https, if not specified http is used. If the **url** or **proxy** is specified - will take protocol from them.
- **username:** (optional) username for etcd authentication.
- **password:** (optional) password for etcd authentication.
- **cacert:** (optional) The ca certificate. If present it will enable validation.
- **cert:** (optional) file with the client certificate.
- **key:** (optional) file with the client key. Can be empty if the key is part of **cert**.

### 3.2.7 Etcdrv3

If you want that Patroni works with Etcd cluster via protocol version 3, you need to use the `etcd3` section in the Patroni configuration file. All configuration parameters are the same as for `etcd`.

#### Warning

Keys created with protocol version 2 are not visible with protocol version 3 and the other way around, therefore it is not possible to switch from `etcd` to `etcd3` just by updating Patroni config file. In addition, Patroni uses Etcd's gRPC-gateway (proxy) to communicate with the V3 API, which means that TLS common name authentication is not possible.

### 3.2.8 ZooKeeper

- **hosts:** List of ZooKeeper cluster members in format: `['host1:port1', 'host2:port2', 'etc...']`.
- **use\_ssl:** (optional) Whether SSL is used or not. Defaults to `false`. If set to `false`, all SSL specific parameters are ignored.
- **cacert:** (optional) The CA certificate. If present it will enable validation.
- **cert:** (optional) File with the client certificate.
- **key:** (optional) File with the client key.
- **key\_password:** (optional) The client key password.
- **verify:** (optional) Whether to verify certificate or not. Defaults to `true`.
- **set\_acls:** (optional) If set, configure Kazoo to apply a default ACL to each ZNode that it creates. ACLs will assume 'x509' schema and should be specified as a dictionary with the principal as the key and one or more permissions as a list in the value. Permissions may be one of `CREATE`, `READ`, `WRITE`, `DELETE` or `ADMIN`. For example, `set_acls: {CN=principal1: [CREATE, READ], CN=principal2: [ALL]}`.
- **auth\_data:** (optional) Authentication credentials to use for the connection. Should be a dictionary in the form that *scheme* is the key and *credential* is the value. Defaults to empty dictionary.

#### Note

It is required to install `kazoo>=2.6.0` to support SSL.



### 3.2.9 Exhibitor

- **hosts:** initial list of Exhibitor (ZooKeeper) nodes in format: 'host1,host2,etc...'. This list updates automatically whenever the Exhibitor (ZooKeeper) cluster topology changes.
- **poll\_interval:** how often the list of ZooKeeper and Exhibitor nodes should be updated from Exhibitor.
- **port:** Exhibitor port.

### 3.2.10 Kubernetes

- **bypass\_api\_service:** (optional) When communicating with the Kubernetes API, Patroni is usually relying on the *kubernetes* service, the address of which is exposed in the pods via the *KUBERNETES\_SERVICE\_HOST* environment variable. If *bypass\_api\_service* is set to *true*, Patroni will resolve the list of API nodes behind the service and connect directly to them.
- **namespace:** (optional) Kubernetes namespace where Patroni pod is running. Default value is *default*.
- **labels:** Labels in format {label1: value1, label2: value2}. These labels will be used to find existing objects (Pods and either Endpoints or ConfigMaps) associated with the current cluster. Also Patroni will set them on every object (Endpoint or ConfigMap) it creates.
- **scope\_label:** (optional) name of the label containing cluster name. Default value is *cluster-name*.
- **role\_label:** (optional) name of the label containing role (*primary*, *replica*, or other custom value). Patroni will set this label on the pod it runs in. Default value is *role*.
- **leader\_label\_value:** (optional) value of the pod label when Postgres role is *primary*. Default value is *primary*.
- **follower\_label\_value:** (optional) value of the pod label when Postgres role is *replica*. Default value is *replica*.
- **standby\_leader\_label\_value:** (optional) value of the pod label when Postgres role is *standby\_leader*. Default value is *primary*.
- **tmp\_role\_label:** (optional) name of the temporary label containing role (*primary* or *replica*). Value of this label will always use the default of corresponding role. Set only when necessary.
- **use\_endpoints:** (optional) if set to *true*, Patroni will use Endpoints instead of ConfigMaps to run leader elections and keep cluster state.
- **pod\_ip:** (optional) IP address of the pod Patroni is running in. This value is required when *use\_endpoints* is enabled and is used to populate the leader endpoint subsets when the pod's PostgreSQL is promoted.
- **ports:** (optional) if the Service object has the name for the port, the same name must appear in the Endpoint object, otherwise service won't work. For example, if your service is defined as {Kind: Service, spec: {ports: [{name: postgresql, port: 5432, targetPort: 5432}]}}, then you have to set *kubernetes.ports: [{"name": "postgresql", "port": 5432}]* and Patroni will use it for updating subsets of the leader Endpoint. This parameter is used only if *kubernetes.use\_endpoints* is set.
- **cacert:** (optional) Specifies the file with the CA\_BUNDLE file with certificates of trusted CAs to use while verifying Kubernetes API SSL certs. If not provided, patroni will use the value provided by the ServiceAccount secret.
- **retriable\_http\_codes:** (optional) list of HTTP status codes from K8s API to retry on. By default Patroni is retrying on 500, 503, and 504, or if K8s API response has *retry-after* HTTP header.

### 3.2.11 Raft (deprecated)

- **self\_addr**: `ip:port` to listen on for Raft connections. The `self_addr` must be accessible from other nodes of the cluster. If not set, the node will not participate in consensus.
- **bind\_addr**: (optional) `ip:port` to listen on for Raft connections. If not specified the `self_addr` will be used.
- **partner\_addrs**: list of other Patroni nodes in the cluster in format: `['ip1:port', 'ip2:port', 'etc...']`
- **data\_dir**: directory where to store Raft log and snapshot. If not specified the current working directory is used.
- **password**: (optional) Encrypt Raft traffic with a specified password, requires `cryptography` python module.

Short FAQ about Raft implementation

- Q: How to list all the nodes providing consensus?  
A: `syncobj_admin -conn host:port -status` where the `host:port` is the address of one of the cluster nodes
- Q: Node that was a part of consensus and has gone and I can't reuse the same IP for other node. How to remove this node from the consensus?  
A: `syncobj_admin -conn host:port -remove host2:port2` where the `host2:port2` is the address of the node you want to remove from consensus.
- Q: Where to get the `syncobj_admin` utility?  
A: It is installed together with `pysyncobj` module (python RAFT implementation), which is Patroni dependency.
- Q: it is possible to run Patroni node without adding in to the consensus?  
A: Yes, just comment out or remove `raft.self_addr` from Patroni configuration.
- Q: It is possible to run Patroni and PostgreSQL only on two nodes?  
A: Yes, on the third node you can run `patroni_raft_controller` (without Patroni and PostgreSQL). In such a setup, one can temporarily lose one node without affecting the primary.

### 3.2.12 PostgreSQL

- **postgresql**:
  - **authentication**:
    - \* **superuser**:
      - **username**: name for the superuser, set during initialization (`initdb`) and later used by Patroni to connect to the postgres.
      - **password**: password for the superuser, set during initialization (`initdb`).
      - **sslmode**: (optional) maps to the `sslmode` connection parameter, which allows a client to specify the type of TLS negotiation mode with the server. For more information on how each mode works, please visit the [PostgreSQL documentation](#). The default mode is `prefer`.
      - **sslkey**: (optional) maps to the `sslkey` connection parameter, which specifies the location of the secret key used with the client's certificate.
      - **sslpassword**: (optional) maps to the `sslpassword` connection parameter, which specifies the password for the secret key specified in `sslkey`.

- **sslcert**: (optional) maps to the `sslcert` connection parameter, which specifies the location of the client certificate.
- **sslrootcert**: (optional) maps to the `sslrootcert` connection parameter, which specifies the location of a file containing one or more certificate authorities (CA) certificates that the client will use to verify a server's certificate.
- **sslcrll**: (optional) maps to the `sslcrll` connection parameter, which specifies the location of a file containing a certificate revocation list. A client will reject connecting to any server that has a certificate present in this list.
- **sslcrldir**: (optional) maps to the `sslcrldir` connection parameter, which specifies the location of a directory with files containing a certificate revocation list. A client will reject connecting to any server that has a certificate present in this list.
- **sslnegotiation**: (optional) maps to the `sslnegotiation` connection parameter, which controls how SSL encryption is negotiated with the server, if SSL is used.
- **gssencmode**: (optional) maps to the `gssencmode` connection parameter, which determines whether or with what priority a secure GSS TCP/IP connection will be negotiated with the server
- **channel\_binding**: (optional) maps to the `channel_binding` connection parameter, which controls the client's use of channel binding.

\* **replication**:

- **username**: replication username; the user will be created during initialization. Replicas will use this user to access the replication source via streaming replication
- **password**: replication password; the user will be created during initialization.
- **sslmode**: (optional) maps to the `sslmode` connection parameter, which allows a client to specify the type of TLS negotiation mode with the server. For more information on how each mode works, please visit the [PostgreSQL documentation](#). The default mode is `prefer`.
- **sslkey**: (optional) maps to the `sslkey` connection parameter, which specifies the location of the secret key used with the client's certificate.
- **sslpassword**: (optional) maps to the `sslpassword` connection parameter, which specifies the password for the secret key specified in `sslkey`.
- **sslcert**: (optional) maps to the `sslcert` connection parameter, which specifies the location of the client certificate.
- **sslrootcert**: (optional) maps to the `sslrootcert` connection parameter, which specifies the location of a file containing one or more certificate authorities (CA) certificates that the client will use to verify a server's certificate.
- **sslcrll**: (optional) maps to the `sslcrll` connection parameter, which specifies the location of a file containing a certificate revocation list. A client will reject connecting to any server that has a certificate present in this list.
- **sslcrldir**: (optional) maps to the `sslcrldir` connection parameter, which specifies the location of a directory with files containing a certificate revocation list. A client will reject connecting to any server that has a certificate present in this list.
- **sslnegotiation**: (optional) maps to the `sslnegotiation` connection parameter, which controls how SSL encryption is negotiated with the server, if SSL is used.
- **gssencmode**: (optional) maps to the `gssencmode` connection parameter, which determines whether or with what priority a secure GSS TCP/IP connection will be negotiated with the server

- **channel\_binding**: (optional) maps to the [channel\\_binding](#) connection parameter, which controls the client's use of channel binding.
- \* **rewind**:
  - **username**: (optional) name for the user for `pg_rewind`; the user will be created during initialization of postgres 11+ and all necessary [permissions](#) will be granted.
  - **password**: (optional) password for the user for `pg_rewind`; the user will be created during initialization.
  - **sslmode**: (optional) maps to the [sslmode](#) connection parameter, which allows a client to specify the type of TLS negotiation mode with the server. For more information on how each mode works, please visit the [PostgreSQL documentation](#). The default mode is `prefer`.
  - **sslkey**: (optional) maps to the [sslkey](#) connection parameter, which specifies the location of the secret key used with the client's certificate.
  - **sslpassword**: (optional) maps to the [sslpassword](#) connection parameter, which specifies the password for the secret key specified in `sslkey`.
  - **sslcert**: (optional) maps to the [sslcert](#) connection parameter, which specifies the location of the client certificate.
  - **sslrootcert**: (optional) maps to the [sslrootcert](#) connection parameter, which specifies the location of a file containing one or more certificate authorities (CA) certificates that the client will use to verify a server's certificate.
  - **sslcrl**: (optional) maps to the [sslcrl](#) connection parameter, which specifies the location of a file containing a certificate revocation list. A client will reject connecting to any server that has a certificate present in this list.
  - **sslcrlidir**: (optional) maps to the [sslcrlidir](#) connection parameter, which specifies the location of a directory with files containing a certificate revocation list. A client will reject connecting to any server that has a certificate present in this list.
  - **sslnegotiation**: (optional) maps to the [sslnegotiation](#) connection parameter, which controls how SSL encryption is negotiated with the server, if SSL is used.
  - **gssencmode**: (optional) maps to the [gssencmode](#) connection parameter, which determines whether or with what priority a secure GSS TCP/IP connection will be negotiated with the server
  - **channel\_binding**: (optional) maps to the [channel\\_binding](#) connection parameter, which controls the client's use of channel binding.
- **callbacks**: callback scripts to run on certain actions. Patroni will pass the action, role and cluster name. (See `scripts/aws.py` as an example of how to write them.)
  - \* **on\_reload**: run this script when configuration reload is triggered.
  - \* **on\_restart**: run this script when the postgres restarts (without changing role).
  - \* **on\_role\_change**: run this script when the postgres is being promoted or demoted.
  - \* **on\_start**: run this script when the postgres starts.
  - \* **on\_stop**: run this script when the postgres stops.
- **connect\_address**: IP address + port through which Postgres is accessible from other nodes and applications.
- **proxy\_address**: IP address + port through which a connection pool (e.g. `pgbouncer`) running next to Postgres is accessible. The value is written to the member key in DCS as `proxy_url` and could be used/useful for service discovery.

- **create\_replica\_methods**: an ordered list of the create methods for turning a Patroni node into a new replica. “basebackup” is the default method; other methods are assumed to refer to scripts, each of which is configured as its own config item. See [custom replica creation methods documentation](#) for further explanation.
- **data\_dir**: The location of the Postgres data directory, either *existing* or to be initialized by Patroni.
- **config\_dir**: The location of the Postgres configuration directory, defaults to the data directory. Must be writable by Patroni.
- **bin\_dir**: (optional) Path to PostgreSQL binaries (pg\_ctl, initdb, pg\_controldata, pg\_basebackup, postgres, pg\_isready, pg\_rewind). If not provided or is an empty string, PATH environment variable will be used to find the executables.
- **bin\_name**: (optional) Make it possible to override Postgres binary names, if you are using a custom Postgres distribution:
  - \* **pg\_ctl**: (optional) Custom name for pg\_ctl binary.
  - \* **initdb**: (optional) Custom name for initdb binary.
  - \* **pgcontroldata**: (optional) Custom name for pg\_controldata binary.
  - \* **pg\_basebackup**: (optional) Custom name for pg\_basebackup binary.
  - \* **postgres**: (optional) Custom name for postgres binary.
  - \* **pg\_isready**: (optional) Custom name for pg\_isready binary.
  - \* **pg\_rewind**: (optional) Custom name for pg\_rewind binary.
- **listen**: IP address + port that Postgres listens to; must be accessible from other nodes in the cluster, if you’re using streaming replication. Multiple comma-separated addresses are permitted, as long as the port component is appended after to the last one with a colon, i.e. `listen: 127.0.0.1,127.0.0.2:5432`. Patroni will use the first address from this list to establish local connections to the PostgreSQL node.
- **use\_unix\_socket**: specifies that Patroni should prefer to use unix sockets to connect to the cluster. Default value is `false`. If `unix_socket_directories` is defined, Patroni will use the first suitable value from it to connect to the cluster and fallback to tcp if nothing is suitable. If `unix_socket_directories` is not specified in `postgresql.parameters`, Patroni will assume that the default value should be used and omit host from the connection parameters.
- **use\_unix\_socket\_repl**: specifies that Patroni should prefer to use unix sockets for replication user cluster connection. Default value is `false`. If `unix_socket_directories` is defined, Patroni will use the first suitable value from it to connect to the cluster and fallback to tcp if nothing is suitable. If `unix_socket_directories` is not specified in `postgresql.parameters`, Patroni will assume that the default value should be used and omit host from the connection parameters.
- **pgpass**: path to the `.pgpass` password file. Patroni creates this file before executing pg\_basebackup, the post\_init script and under some other circumstances. The location must be writable by Patroni.
- **recovery\_conf**: additional configuration settings written to recovery.conf when configuring follower.
- **custom\_conf** : path to an optional custom `postgresql.conf` file, that will be used in place of `postgresql.base.conf`. The file must exist on all cluster nodes, be readable by PostgreSQL and will be included from its location on the real `postgresql.conf`. Note that Patroni will not monitor this file for changes, nor backup it. However, its settings can still be overridden by Patroni’s own configuration facilities - see [dynamic configuration](#) for details.
- **parameters**: configuration parameters (GUCs) for Postgres in format `{ssl: "on", ssl_cert_file: "cert_file"}`.

- **pg\_hba**: list of lines that Patroni will use to generate `pg_hba.conf`. Patroni ignores this parameter if `hba_file` PostgreSQL parameter is set to a non-default value. Together with *dynamic configuration* this parameter simplifies management of `pg_hba.conf`.
  - \* - **host all all 0.0.0.0/0 md5**
  - \* - **host replication replicator 127.0.0.1/32 md5**: A line like this is required for replication.
- **pg\_ident**: list of lines that Patroni will use to generate `pg_ident.conf`. Patroni ignores this parameter if `ident_file` PostgreSQL parameter is set to a non-default value. Together with *dynamic configuration* this parameter simplifies management of `pg_ident.conf`.
  - \* - **mapname1 systemname1 pguser1**
  - \* - **mapname1 systemname2 pguser2**
- **pg\_ctl\_timeout**: How long should `pg_ctl` wait when doing `start`, `stop` or `restart`. Default value is 60 seconds.
- **use\_pg\_rewind**: try to use `pg_rewind` on the former leader when it joins cluster as a replica. Either the cluster must be initialized with `data page checksums` (`--data-checksums` option for `initdb`) and/or `wal_log_hints` must be set to `on`, or `pg_rewind` will not work.
- **remove\_data\_directory\_on\_rewind\_failure**: If this option is enabled, Patroni will remove the PostgreSQL data directory and recreate the replica. Otherwise it will try to follow the new leader. Default value is **false**.
- **remove\_data\_directory\_on\_diverged\_timelines**: Patroni will remove the PostgreSQL data directory and recreate the replica if it notices that timelines are diverging and the former primary can not start streaming from the new primary. This option is useful when `pg_rewind` can not be used. While performing timelines divergence check on PostgreSQL v10 and older Patroni will try to connect with replication credential to the “postgres” database. Hence, such access should be allowed in the `pg_hba.conf`. Default value is **false**.
- **replica\_method**: for each `create_replica_methods` other than `basebackup`, you would add a configuration section of the same name. At a minimum, this should include “command” with a full path to the actual script to be executed. Other configuration parameters will be passed along to the script in the form “parameter=value”.
- **pre\_promote**: a fencing script that executes during a failover after acquiring the leader lock but before promoting the replica. If the script exits with a non-zero code, Patroni does not promote the replica and removes the leader key from DCS.
- **before\_stop**: a script that executes immediately prior to stopping postgres. As opposed to a callback, this script runs synchronously, blocking shutdown until it has completed. The return code of this script does not impact whether shutdown proceeds afterwards.

### 3.2.13 REST API

- **restapi**:
  - **connect\_address**: IP address (or hostname) and port, to access the Patroni’s *REST API*. All the members of the cluster must be able to connect to this address, so unless the Patroni setup is intended for a demo inside the localhost, this address must be a non “localhost” or loopback address (ie: “localhost” or “127.0.0.1”). It can serve as an endpoint for HTTP health checks (read below about the “listen” REST API parameter), and also for user queries (either directly or via the REST API), as well as for the health checks done by the cluster members during leader elections (for example, to determine whether the leader is still running, or if there is a node which has a WAL position that is ahead of the one doing the query; etc.) The `connect_address` is put in the member key in DCS, making it possible to translate the member name into the address to connect to its REST API.



- **listen**: IP address (or hostname) and port that Patroni will listen to for the REST API - to provide also the same health checks and cluster messaging between the participating nodes, as described above. to provide health-check information for HAProxy (or any other load balancer capable of doing a HTTP “OPTION” or “GET” checks).
- **authentication**: (optional)
  - \* **username**: Basic-auth username to protect unsafe REST API endpoints.
  - \* **password**: Basic-auth password to protect unsafe REST API endpoints.
- **certfile**: (optional): Specifies the file with the certificate in the PEM format. If the certfile is not specified or is left empty, the API server will work without SSL.
- **keyfile**: (optional): Specifies the file with the secret key in the PEM format.
- **keyfile\_password**: (optional): Specifies a password for decrypting the keyfile.
- **cafile**: (optional): Specifies the file with the CA\_BUNDLE with certificates of trusted CAs to use while verifying client certs.
- **ciphers**: (optional): Specifies the permitted cipher suites (e.g. “ECDHE-RSA-AES256-GCM-SHA384:DHE-RSA-AES256-GCM-SHA384:ECDHE-RSA-AES128-GCM-SHA256:DHE-RSA-AES128-GCM-SHA256:!SSLv1:!SSLv2:!SSLv3:!TLSv1:!TLSv1.1”)
- **verify\_client**: (optional): none (default), optional or required. When none REST API will not check client certificates. When required client certificates are required for all REST API calls. When optional client certificates are required for all unsafe REST API endpoints. When required is used, then client authentication succeeds, if the certificate signature verification succeeds. For optional the client cert will only be checked for PUT, POST, PATCH, and DELETE requests.
- **allowlist**: (optional): Specifies the set of hosts that are allowed to call unsafe REST API endpoints. The single element could be a host name, an IP address or a network address using CIDR notation. By default allow all is used. In case if allowlist or allowlist\_include\_members are set, anything that is not included is rejected.
- **allowlist\_include\_members**: (optional): If set to true it allows accessing unsafe REST API endpoints from other cluster members registered in DCS (IP address or hostname is taken from the members api\_url). Be careful, it might happen that OS will use a different IP for outgoing connections.
- **http\_extra\_headers**: (optional): HTTP headers let the REST API server pass additional information with an HTTP response.
- **https\_extra\_headers**: (optional): HTTPS headers let the REST API server pass additional information with an HTTP response when TLS is enabled. This will also pass additional information set in http\_extra\_headers.
- **request\_queue\_size**: (optional): Sets request queue size for TCP socket used by Patroni REST API. Once the queue is full, further requests get a “Connection denied” error. The default value is 5.

Here is an example of both **http\_extra\_headers** and **https\_extra\_headers**:

```
restapi:
  listen: <listen>
  connect_address: <connect_address>
  authentication:
    username: <username>
    password: <password>
  http_extra_headers:
    'X-Frame-Options': 'SAMEORIGIN'
    'X-XSS-Protection': '1; mode=block'
```

(continues on next page)

(continued from previous page)

```
'X-Content-Type-Options': 'nosniff'
cafile: <ca file>
certfile: <cert>
keyfile: <key>
https_extra_headers:
  'Strict-Transport-Security': 'max-age=31536000; includeSubDomains'
```

### Warning

- The `restapi.connect_address` must be accessible from all nodes of a given Patroni cluster. Internally Patroni is using it during the leader race to find nodes with minimal replication lag.
- If you enabled client certificates validation (`restapi.verify_client` is set to `required`), you also **must** provide **valid client certificates** in the `ctl.certfile`, `ctl.keyfile`, `ctl.keyfile_password`. If not provided, Patroni will not work correctly.

## 3.2.14 CTL

- **ctl:** (optional)
  - **authentication:**
    - \* **username:** Basic-auth username for accessing protected REST API endpoints. If not provided *patronictl* will use the value provided for REST API “username” parameter.
    - \* **password:** Basic-auth password for accessing protected REST API endpoints. If not provided *patronictl* will use the value provided for REST API “password” parameter.
  - **insecure:** Allow connections to REST API without verifying SSL certs.
  - **cacert:** Specifies the file with the CA\_BUNDLE file or directory with certificates of trusted CAs to use while verifying REST API SSL certs. If not provided *patronictl* will use the value provided for REST API “cafile” parameter.
  - **certfile:** Specifies the file with the client certificate in the PEM format.
  - **keyfile:** Specifies the file with the client secret key in the PEM format.
  - **keyfile\_password:** Specifies a password for decrypting the client keyfile.

## 3.2.15 Watchdog

- **mode:** `off`, `automatic` or `required`. When `off` watchdog is disabled. When `automatic` watchdog will be used if available, but ignored if it is not. When `required` the node will not become a leader unless watchdog can be successfully enabled.
- **device:** Path to watchdog device. Defaults to `/dev/watchdog`.
- **safety\_margin:** Number of seconds of safety margin between watchdog triggering and leader key expiration.



### 3.2.16 Tags

- **clonefrom:** true or false. If set to true other nodes might prefer to use this node for bootstrap (take `pg_basebackup` from). If there are several nodes with `clonefrom` tag set to true the node to bootstrap from will be chosen randomly. The default value is false.
- **noloadbalance:** true or false. If set to true the node will return HTTP Status Code 503 for the GET / replica REST API health-check and therefore will be excluded from the load-balancing. Defaults to false.
- **replicatefrom:** The name of another replica to replicate from. Used to support cascading replication.
- **nosync:** true or false. If set to true the node will never be selected as a synchronous replica.
- **nofailover:** true or false, controls whether this node is allowed to participate in the leader race and become a leader. Defaults to false, meaning this node `_can_` participate in leader races.
- **failover\_priority:** integer, controls the priority that this node should have during failover. Nodes with higher priority will be preferred over lower priority nodes if they received/replayed the same amount of WAL. However, nodes with higher values of receive/replay LSN are preferred regardless of their priority. If the `failover_priority` is 0 or negative - such node is not allowed to participate in the leader race and to become a leader (similar to `nofailover: true`).
- **nostream:** true or false. If set to true the node will not use replication protocol to stream WAL. It will rely instead on archive recovery (if `restore_command` is configured) and `pg_wal/pg_xlog` polling. It also disables copying and synchronization of permanent logical replication slots on the node itself and all its cascading replicas. Setting this tag on primary node has no effect.

#### Warning

Provide only one of `nofailover` or `failover_priority`. Providing `nofailover: true` is the same as `failover_priority: 0`, and providing `nofailover: false` will give the node priority 1.

In addition to these predefined tags, you can also add your own ones:

- **key1:** true
- **key2:** false
- **key3:** 1.4
- **key4:** "RandomString"

Tags are visible in the [REST API](#) and `patronictl list`. You can also check for an instance health using these tags. If the tag isn't defined for an instance, or if the respective value doesn't match the querying value, it will return HTTP Status Code 503.

## 3.3 Environment Configuration Settings

It is possible to override some of the configuration parameters defined in the Patroni configuration file using the system environment variables. This document lists all environment variables handled by Patroni. The values set via those variables always take precedence over the ones set in the Patroni configuration file.

### 3.3.1 Global/Universal

- **PATRONI\_CONFIGURATION**: it is possible to set the entire configuration for the Patroni via `PATRONI_CONFIGURATION` environment variable. In this case any other environment variables will not be considered!
- **PATRONI\_NAME**: name of the node where the current instance of Patroni is running. Must be unique for the cluster.
- **PATRONI\_NAMESPACE**: path within the configuration store where Patroni will keep information about the cluster. Default value: `"/service"`
- **PATRONI\_SCOPE**: cluster name

### 3.3.2 Log

- **PATRONI\_LOG\_TYPE**: sets the format of logs. Can be either **plain** or **json**. To use **json** format, you must have the *jsonlogger* installed. The default value is **plain**.
- **PATRONI\_LOG\_LEVEL**: sets the general logging level. Default value is **INFO** (see [the docs for Python logging](#))
- **PATRONI\_LOG\_TRACEBACK\_LEVEL**: sets the level where tracebacks will be visible. Default value is **ERROR**. Set it to **DEBUG** if you want to see tracebacks only if you enable **PATRONI\_LOG\_LEVEL=DEBUG**.
- **PATRONI\_LOG\_FORMAT**: sets the log formatting string. If the log type is **plain**, the log format should be a string. Refer to [the LogRecord attributes](#) for available attributes. If the log type is **json**, the log format can be a list in addition to a string. Each list item should correspond to LogRecord attributes. Be cautious that only the field name is required, and the `%`( and ) should be omitted. If you wish to print a log field with a different key name, use a dictionary where the dictionary key is the log field, and the value is the name of the field you want to be printed in the log. Default value is `%(asctime)s %(levelname)s: %(message)s`
- **PATRONI\_LOG\_DATEFORMAT**: sets the datetime formatting string. (see the [formatTime\(\) documentation](#))
- **PATRONI\_LOG\_STATIC\_FIELDS**: add additional fields to the log. This option is only available when the log type is set to **json**. Example `PATRONI_LOG_STATIC_FIELDS="{app: patroni}"`
- **PATRONI\_LOG\_MAX\_QUEUE\_SIZE**: Patroni is using two-step logging. Log records are written into the in-memory queue and there is a separate thread which pulls them from the queue and writes to stderr or file. The maximum size of the internal queue is limited by default by **1000** records, which is enough to keep logs for the past 1h20m.
- **PATRONI\_LOG\_DIR**: Directory to write application logs to. The directory must exist and be writable by the user executing Patroni. If you set this env variable, the application will retain 4 25MB logs by default. You can tune those retention values with `PATRONI_LOG_FILE_NUM` and `PATRONI_LOG_FILE_SIZE` (see below).
- **PATRONI\_LOG\_MODE**: Permissions for log files (for example, `0644`). If not specified, permissions will be set based on the current umask value.
- **PATRONI\_LOG\_FILE\_NUM**: The number of application logs to retain.
- **PATRONI\_LOG\_FILE\_SIZE**: Size of patroni.log file (in bytes) that triggers a log rolling.
- **PATRONI\_LOG\_LOGGERS**: Redefine logging level per python module. Example `PATRONI_LOG_LOGGERS="{patroni.postmaster: WARNING, urllib3: DEBUG}"`

### 3.3.3 Citus

Enables integration Patroni with [Citus](#). If configured, Patroni will take care of registering Citus worker nodes on the coordinator. You can find more information about Citus support [here](#).

- **PATRONI\_CITUS\_GROUP**: the Citus group id, integer. Use 0 for coordinator and 1, 2, etc... for workers
- **PATRONI\_CITUS\_DATABASE**: the database where `ci_tus` extension should be created. Must be the same on the coordinator and all workers. Currently only one database is supported.

### 3.3.4 Consul

- **PATRONI\_CONSUL\_HOST**: the host:port for the Consul local agent.
- **PATRONI\_CONSUL\_URL**: url for the Consul local agent, in format: `http(s)://host:port`
- **PATRONI\_CONSUL\_PORT**: (optional) Consul port
- **PATRONI\_CONSUL\_SCHEME**: (optional) **http** or **https**, defaults to **http**
- **PATRONI\_CONSUL\_TOKEN**: (optional) ACL token
- **PATRONI\_CONSUL\_VERIFY**: (optional) whether to verify the SSL certificate for HTTPS requests
- **PATRONI\_CONSUL\_CACERT**: (optional) The ca certificate. If present it will enable validation.
- **PATRONI\_CONSUL\_CERT**: (optional) File with the client certificate
- **PATRONI\_CONSUL\_KEY**: (optional) File with the client key. Can be empty if the key is part of certificate.
- **PATRONI\_CONSUL\_DC**: (optional) Datacenter to communicate with. By default the datacenter of the host is used.
- **PATRONI\_CONSUL\_CONSISTENCY**: (optional) Select consul consistency mode. Possible values are `default`, `consistent`, or `stale` (more details in [consul API reference](#))
- **PATRONI\_CONSUL\_CHECKS**: (optional) list of Consul health checks used for the session. By default an empty list is used.
- **PATRONI\_CONSUL\_REGISTER\_SERVICE**: (optional) whether or not to register a service with the name defined by the `scope` parameter and the tag `master`, `primary`, `replica`, or `standby-leader` depending on the node's role. Defaults to **false**
- **PATRONI\_CONSUL\_SERVICE\_TAGS**: (optional) additional static tags to add to the Consul service apart from the role (`primary/replica/standby-leader`). By default an empty list is used.
- **PATRONI\_CONSUL\_SERVICE\_CHECK\_INTERVAL**: (optional) how often to perform health check against registered url
- **PATRONI\_CONSUL\_SERVICE\_CHECK\_TLS\_SERVER\_NAME**: (optional) override SNI host when connecting via TLS, see also [consul agent check API reference](#).

### 3.3.5 Etcd

- **PATRONI\_ETCD\_PROXY**: proxy url for the etcd. If you are connecting to the etcd using proxy, use this parameter instead of **PATRONI\_ETCD\_URL**
- **PATRONI\_ETCD\_URL**: url for the etcd, in format: `http(s)://(username:password@)host:port`
- **PATRONI\_ETCD\_HOSTS**: list of etcd endpoints in format `'host1:port1','host2:port2',etc...`
- **PATRONI\_ETCD\_USE\_PROXIES**: If this parameter is set to true, Patroni will consider **hosts** as a list of proxies and will not perform a topology discovery of etcd cluster but stick to a fixed list of **hosts**.
- **PATRONI\_ETCD\_PROTOCOL**: http or https, if not specified http is used. If the **url** or **proxy** is specified - will take protocol from them.
- **PATRONI\_ETCD\_HOST**: the host:port for the etcd endpoint.
- **PATRONI\_ETCD\_SRV**: Domain to search the SRV record(s) for cluster autodiscovery. Patroni will try to query these SRV service names for specified domain (in that order until first success): `_etcd-client-ssl`, `_etcd-client`, `_etcd-ssl`, `_etcd`, `_etcd-server-ssl`, `_etcd-server`. If SRV records for `_etcd-server-ssl` or `_etcd-server` are retrieved then ETCD peer protocol is used do query ETCD for available members. Otherwise hosts from SRV records will be used.
- **PATRONI\_ETCD\_SRV\_SUFFIX**: Configures a suffix to the SRV name that is queried during discovery. Use this flag to differentiate between multiple etcd clusters under the same domain. Works only with conjunction with **PATRONI\_ETCD\_SRV**. For example, if `PATRONI_ETCD_SRV_SUFFIX=foo` and `PATRONI_ETCD_SRV=example.org` are set, the following DNS SRV query is made: `_etcd-client-ssl-foo._tcp.example.com` (and so on for every possible ETCD SRV service name).
- **PATRONI\_ETCD\_USERNAME**: username for etcd authentication.
- **PATRONI\_ETCD\_PASSWORD**: password for etcd authentication.
- **PATRONI\_ETCD\_CACERT**: The ca certificate. If present it will enable validation.
- **PATRONI\_ETCD\_CERT**: File with the client certificate.
- **PATRONI\_ETCD\_KEY**: File with the client key. Can be empty if the key is part of certificate.

### 3.3.6 Etcdv3

Environment names for Etcdv3 are similar as for Etcd, you just need to use `ETCD3` instead of `ETCD` in the variable name. Example: `PATRONI_ETCD3_HOST`, `PATRONI_ETCD3_CACERT`, and so on.

#### Warning

Keys created with protocol version 2 are not visible with protocol version 3 and the other way around, therefore it is not possible to switch from Etcd to Etcdv3 just by updating Patroni configuration. In addition, Patroni uses Etcd's gRPC-gateway (proxy) to communicate with the V3 API, which means that TLS common name authentication is not possible.

### 3.3.7 ZooKeeper

- **PATRONI\_ZOOKEEPER\_HOSTS:** Comma separated list of ZooKeeper cluster members: “‘host1:port1’, ‘host2:port2’, etc...’”. It is important to quote every single entity!
- **PATRONI\_ZOOKEEPER\_USE\_SSL:** (optional) Whether SSL is used or not. Defaults to `false`. If set to `false`, all SSL specific parameters are ignored.
- **PATRONI\_ZOOKEEPER\_CACERT:** (optional) The CA certificate. If present it will enable validation.
- **PATRONI\_ZOOKEEPER\_CERT:** (optional) File with the client certificate.
- **PATRONI\_ZOOKEEPER\_KEY:** (optional) File with the client key.
- **PATRONI\_ZOOKEEPER\_KEY\_PASSWORD:** (optional) The client key password.
- **PATRONI\_ZOOKEEPER\_VERIFY:** (optional) Whether to verify certificate or not. Defaults to `true`.
- **PATRONI\_ZOOKEEPER\_SET\_ACLS:** (optional) If set, configure Kazoo to apply a default ACL to each ZN-ode that it creates. ACLs will assume ‘x509’ schema and should be specified as a dictionary with the principal as the key and one or more permissions as a list in the value. Permissions may be one of `CREATE`, `READ`, `WRITE`, `DELETE` or `ADMIN`. For example, `set_acls: {CN=principal1: [CREATE, READ], CN=principal2: [ALL]}`.
- **PATRONI\_ZOOKEEPER\_AUTH\_DATA:** (optional) Authentication credentials to use for the connection. Should be a dictionary in the form that *scheme* is the key and *credential* is the value. Defaults to empty dictionary.

#### Note

It is required to install `kazoo>=2.6.0` to support SSL.

### 3.3.8 Exhibitor

- **PATRONI\_EXHIBITOR\_HOSTS:** initial list of Exhibitor (ZooKeeper) nodes in format: ‘host1,host2,etc...’. This list updates automatically whenever the Exhibitor (ZooKeeper) cluster topology changes.
- **PATRONI\_EXHIBITOR\_PORT:** Exhibitor port.

### 3.3.9 Kubernetes

- **PATRONI\_KUBERNETES\_BYPASS\_API\_SERVICE:** (optional) When communicating with the Kubernetes API, Patroni is usually relying on the *kubernetes* service, the address of which is exposed in the pods via the `KUBERNETES_SERVICE_HOST` environment variable. If `PATRONI_KUBERNETES_BYPASS_API_SERVICE` is set to `true`, Patroni will resolve the list of API nodes behind the service and connect directly to them.
- **PATRONI\_KUBERNETES\_NAMESPACE:** (optional) Kubernetes namespace where the Patroni pod is running. Default value is *default*.
- **PATRONI\_KUBERNETES\_LABELS:** Labels in format `{label1: value1, label2: value2}`. These labels will be used to find existing objects (Pods and either Endpoints or ConfigMaps) associated with the current cluster. Also Patroni will set them on every object (Endpoint or ConfigMap) it creates.
- **PATRONI\_KUBERNETES\_SCOPE\_LABEL:** (optional) name of the label containing cluster name. Default value is *cluster-name*.

- **PATRONI\_KUBERNETES\_ROLE\_LABEL:** (optional) name of the label containing role (*primary*, *replica* or other custom value). Patroni will set this label on the pod it runs in. Default value is *role*.
- **PATRONI\_KUBERNETES\_LEADER\_LABEL\_VALUE:** (optional) value of the pod label when Postgres role is *primary*. Default value is *primary*.
- **PATRONI\_KUBERNETES\_FOLLOWER\_LABEL\_VALUE:** (optional) value of the pod label when Postgres role is *replica*. Default value is *replica*.
- **PATRONI\_KUBERNETES\_STANDBY\_LEADER\_LABEL\_VALUE:** (optional) value of the pod label when Postgres role is *standby\_leader*. Default value is *primary*.
- **PATRONI\_KUBERNETES\_TMP\_ROLE\_LABEL:** (optional) name of the temporary label containing role (*primary* or *replica*). Value of this label will always use the default of corresponding role. Set only when necessary.
- **PATRONI\_KUBERNETES\_USE\_ENDPOINTS:** (optional) if set to true, Patroni will use Endpoints instead of ConfigMaps to run leader elections and keep cluster state.
- **PATRONI\_KUBERNETES\_POD\_IP:** (optional) IP address of the pod Patroni is running in. This value is required when *PATRONI\_KUBERNETES\_USE\_ENDPOINTS* is enabled and is used to populate the leader endpoint subsets when the pod's PostgreSQL is promoted.
- **PATRONI\_KUBERNETES\_PORTS:** (optional) if the Service object has the name for the port, the same name must appear in the Endpoint object, otherwise service won't work. For example, if your service is defined as `{Kind: Service, spec: {ports: [{name: postgresql, port: 5432, targetPort: 5432}]}}`, then you have to set *PATRONI\_KUBERNETES\_PORTS*='[{"name": "postgresql", "port": 5432}]' and Patroni will use it for updating subsets of the leader Endpoint. This parameter is used only if *PATRONI\_KUBERNETES\_USE\_ENDPOINTS* is set.
- **PATRONI\_KUBERNETES\_CACERT:** (optional) Specifies the file with the CA\_BUNDLE file with certificates of trusted CAs to use while verifying Kubernetes API SSL certs. If not provided, patroni will use the value provided by the ServiceAccount secret.
- **PATRONI\_RETRIABLE\_HTTP\_CODES:** (optional) list of HTTP status codes from K8s API to retry on. By default Patroni is retrying on 500, 503, and 504, or if K8s API response has *retry-after* HTTP header.

### 3.3.10 Raft (deprecated)

- **PATRONI\_RAFT\_SELF\_ADDR:** *ip:port* to listen on for Raft connections. The *self\_addr* must be accessible from other nodes of the cluster. If not set, the node will not participate in consensus.
- **PATRONI\_RAFT\_BIND\_ADDR:** (optional) *ip:port* to listen on for Raft connections. If not specified the *self\_addr* will be used.
- **PATRONI\_RAFT\_PARTNER\_ADDRS:** list of other Patroni nodes in the cluster in format "'ip1:port1', 'ip2:port2'". It is important to quote every single entity!
- **PATRONI\_RAFT\_DATA\_DIR:** directory where to store Raft log and snapshot. If not specified the current working directory is used.
- **PATRONI\_RAFT\_PASSWORD:** (optional) Encrypt Raft traffic with a specified password, requires cryptography python module.

### 3.3.11 PostgreSQL

- **PATRONI\_POSTGRES\_LISTEN**: IP address + port that Postgres listens to. Multiple comma-separated addresses are permitted, as long as the port component is appended after to the last one with a colon, i.e. `listen: 127.0.0.1, 127.0.0.2:5432`. Patroni will use the first address from this list to establish local connections to the PostgreSQL node.
- **PATRONI\_POSTGRES\_CONNECT\_ADDRESS**: IP address + port through which Postgres is accessible from other nodes and applications.
- **PATRONI\_POSTGRES\_PROXY\_ADDRESS**: IP address + port through which a connection pool (e.g. pgbouncer) running next to Postgres is accessible. The value is written to the member key in DCS as `proxy_url` and could be used/useful for service discovery.
- **PATRONI\_POSTGRES\_DATA\_DIR**: The location of the Postgres data directory, either existing or to be initialized by Patroni.
- **PATRONI\_POSTGRES\_CONFIG\_DIR**: The location of the Postgres configuration directory, defaults to the data directory. Must be writable by Patroni.
- **PATRONI\_POSTGRES\_BIN\_DIR**: Path to PostgreSQL binaries. (`pg_ctl`, `initdb`, `pg_controldata`, `pg_basebackup`, `postgres`, `pg_isready`, `pg_rewind`) The default value is an empty string meaning that `PATH` environment variable will be used to find the executables.
- **PATRONI\_POSTGRES\_BIN\_PG\_CTL**: (optional) Custom name for `pg_ctl` binary.
- **PATRONI\_POSTGRES\_BIN\_INITDB**: (optional) Custom name for `initdb` binary.
- **PATRONI\_POSTGRES\_BIN\_PG\_CONTROLDATA**: (optional) Custom name for `pg_controldata` binary.
- **PATRONI\_POSTGRES\_BIN\_PG\_BASEBACKUP**: (optional) Custom name for `pg_basebackup` binary.
- **PATRONI\_POSTGRES\_BIN\_POSTGRES**: (optional) Custom name for `postgres` binary.
- **PATRONI\_POSTGRES\_BIN\_IS\_READY**: (optional) Custom name for `pg_isready` binary.
- **PATRONI\_POSTGRES\_BIN\_PG\_REWIND**: (optional) Custom name for `pg_rewind` binary.
- **PATRONI\_POSTGRES\_PGPASS**: path to the `.pgpass` password file. Patroni creates this file before executing `pg_basebackup` and under some other circumstances. The location must be writable by Patroni.
- **PATRONI\_REPLICATION\_USERNAME**: replication username; the user will be created during initialization. Replicas will use this user to access the replication source via streaming replication
- **PATRONI\_REPLICATION\_PASSWORD**: replication password; the user will be created during initialization.
- **PATRONI\_REPLICATION\_SSLMODE**: (optional) maps to the `sslmode` connection parameter, which allows a client to specify the type of TLS negotiation mode with the server. For more information on how each mode works, please visit the [PostgreSQL documentation](#). The default mode is `prefer`.
- **PATRONI\_REPLICATION\_SSLKEY**: (optional) maps to the `sslkey` connection parameter, which specifies the location of the secret key used with the client's certificate.
- **PATRONI\_REPLICATION\_SSLPASSWORD**: (optional) maps to the `sslpassword` connection parameter, which specifies the password for the secret key specified in `PATRONI_REPLICATION_SSLKEY`.
- **PATRONI\_REPLICATION\_SSLCERT**: (optional) maps to the `sslcert` connection parameter, which specifies the location of the client certificate.
- **PATRONI\_REPLICATION\_SSLROOTCERT**: (optional) maps to the `sslrootcert` connection parameter, which specifies the location of a file containing one or more certificate authorities (CA) certificates that the client will use to verify a server's certificate.



- **PATRONI\_REPLICATION\_SSLCRL:** (optional) maps to the [sslcrl](#) connection parameter, which specifies the location of a file containing a certificate revocation list. A client will reject connecting to any server that has a certificate present in this list.
- **PATRONI\_REPLICATION\_SSLCRLDIR:** (optional) maps to the [sslcrlidir](#) connection parameter, which specifies the location of a directory with files containing a certificate revocation list. A client will reject connecting to any server that has a certificate present in this list.
- **PATRONI\_REPLICATION\_SSLNEGOTIATION:** (optional) maps to the [sslnegotiation](#) connection parameter, which controls how SSL encryption is negotiated with the server, if SSL is used.
- **PATRONI\_REPLICATION\_GSSENCMODE:** (optional) maps to the [gssencmode](#) connection parameter, which determines whether or with what priority a secure GSS TCP/IP connection will be negotiated with the server
- **PATRONI\_REPLICATION\_CHANNEL\_BINDING:** (optional) maps to the [channel\\_binding](#) connection parameter, which controls the client's use of channel binding.
- **PATRONI\_SUPERUSER\_USERNAME:** name for the superuser, set during initialization (initdb) and later used by Patroni to connect to the postgres. Also this user is used by `pg_rewind`.
- **PATRONI\_SUPERUSER\_PASSWORD:** password for the superuser, set during initialization (initdb).
- **PATRONI\_SUPERUSER\_SSLMODE:** (optional) maps to the [sslmode](#) connection parameter, which allows a client to specify the type of TLS negotiation mode with the server. For more information on how each mode works, please visit the [PostgreSQL documentation](#). The default mode is `prefer`.
- **PATRONI\_SUPERUSER\_SSLKEY:** (optional) maps to the [sslkey](#) connection parameter, which specifies the location of the secret key used with the client's certificate.
- **PATRONI\_SUPERUSER\_SSLPASSWORD:** (optional) maps to the [sslpassword](#) connection parameter, which specifies the password for the secret key specified in `PATRONI_SUPERUSER_SSLKEY`.
- **PATRONI\_SUPERUSER\_SSLCERT:** (optional) maps to the [sslcert](#) connection parameter, which specifies the location of the client certificate.
- **PATRONI\_SUPERUSER\_SSLROOTCERT:** (optional) maps to the [sslrootcert](#) connection parameter, which specifies the location of a file containing one or more certificate authorities (CA) certificates that the client will use to verify a server's certificate.
- **PATRONI\_SUPERUSER\_SSLCRL:** (optional) maps to the [sslcrl](#) connection parameter, which specifies the location of a file containing a certificate revocation list. A client will reject connecting to any server that has a certificate present in this list.
- **PATRONI\_SUPERUSER\_SSLCRLDIR:** (optional) maps to the [sslcrlidir](#) connection parameter, which specifies the location of a directory with files containing a certificate revocation list. A client will reject connecting to any server that has a certificate present in this list.
- **PATRONI\_SUPERUSER\_SSLNEGOTIATION:** (optional) maps to the [sslnegotiation](#) connection parameter, which controls how SSL encryption is negotiated with the server, if SSL is used.
- **PATRONI\_SUPERUSER\_GSSENCMODE:** (optional) maps to the [gssencmode](#) connection parameter, which determines whether or with what priority a secure GSS TCP/IP connection will be negotiated with the server
- **PATRONI\_SUPERUSER\_CHANNEL\_BINDING:** (optional) maps to the [channel\\_binding](#) connection parameter, which controls the client's use of channel binding.
- **PATRONI\_REWIND\_USERNAME:** (optional) name for the user for `pg_rewind`; the user will be created during initialization of postgres 11+ and all necessary [permissions](#) will be granted.
- **PATRONI\_REWIND\_PASSWORD:** (optional) password for the user for `pg_rewind`; the user will be created during initialization.



- **PATRONI\_REWIND\_SSLMODE:** (optional) maps to the `sslmode` connection parameter, which allows a client to specify the type of TLS negotiation mode with the server. For more information on how each mode works, please visit the [PostgreSQL documentation](#). The default mode is `prefer`.
- **PATRONI\_REWIND\_SSLKEY:** (optional) maps to the `sslkey` connection parameter, which specifies the location of the secret key used with the client's certificate.
- **PATRONI\_REWIND\_SSLPASSWORD:** (optional) maps to the `sslpassword` connection parameter, which specifies the password for the secret key specified in `PATRONI_REWIND_SSLKEY`.
- **PATRONI\_REWIND\_SSLCERT:** (optional) maps to the `sslcert` connection parameter, which specifies the location of the client certificate.
- **PATRONI\_REWIND\_SSLROOTCERT:** (optional) maps to the `sslrootcert` connection parameter, which specifies the location of a file containing one or more certificate authorities (CA) certificates that the client will use to verify a server's certificate.
- **PATRONI\_REWIND\_SSLCRL:** (optional) maps to the `sslcrl` connection parameter, which specifies the location of a file containing a certificate revocation list. A client will reject connecting to any server that has a certificate present in this list.
- **PATRONI\_REWIND\_SSLCRLDIR:** (optional) maps to the `sslcrlidir` connection parameter, which specifies the location of a directory with files containing a certificate revocation list. A client will reject connecting to any server that has a certificate present in this list.
- **PATRONI\_REWIND\_SSLNEGOTIATION:** (optional) maps to the `sslnegotiation` connection parameter, which controls how SSL encryption is negotiated with the server, if SSL is used.
- **PATRONI\_REWIND\_GSSENCMODE:** (optional) maps to the `gssencmode` connection parameter, which determines whether or with what priority a secure GSS TCP/IP connection will be negotiated with the server
- **PATRONI\_REWIND\_CHANNEL\_BINDING:** (optional) maps to the `channel_binding` connection parameter, which controls the client's use of channel binding.

### 3.3.12 REST API

- **PATRONI\_RESTAPI\_CONNECT\_ADDRESS:** IP address and port to access the REST API.
- **PATRONI\_RESTAPI\_LISTEN:** IP address and port that Patroni will listen to, to provide health-check information for HAProxy.
- **PATRONI\_RESTAPI\_USERNAME:** Basic-auth username to protect unsafe REST API endpoints.
- **PATRONI\_RESTAPI\_PASSWORD:** Basic-auth password to protect unsafe REST API endpoints.
- **PATRONI\_RESTAPI\_CERTFILE:** Specifies the file with the certificate in the PEM format. If the certfile is not specified or is left empty, the API server will work without SSL.
- **PATRONI\_RESTAPI\_KEYFILE:** Specifies the file with the secret key in the PEM format.
- **PATRONI\_RESTAPI\_KEYFILE\_PASSWORD:** Specifies a password for decrypting the keyfile.
- **PATRONI\_RESTAPI\_CAFILE:** Specifies the file with the CA\_BUNDLE with certificates of trusted CAs to use while verifying client certs.
- **PATRONI\_RESTAPI\_CIPHERS:** (optional) Specifies the permitted cipher suites (e.g. "ECDHE-RSA-AES256-GCM-SHA384:DHE-RSA-AES256-GCM-SHA384:ECDHE-RSA-AES128-GCM-SHA256:DHE-RSA-AES128-GCM-SHA256:!SSLv1:!SSLv2:!SSLv3:!TLSv1:!TLSv1.1")
- **PATRONI\_RESTAPI\_VERIFY\_CLIENT:** `none` (default), `optional` or `required`. When `none` REST API will not check client certificates. When `required` client certificates are required for all REST API calls. When `optional` client certificates are required for all unsafe REST API endpoints. When `required` is used, then

client authentication succeeds, if the certificate signature verification succeeds. For `optional` the client cert will only be checked for PUT, POST, PATCH, and DELETE requests.

- **PATRONI\_RESTAPI\_ALLOWLIST**: (optional): Specifies the set of hosts that are allowed to call unsafe REST API endpoints. The single element could be a host name, an IP address or a network address using CIDR notation. By default `allow_all` is used. In case if `allowlist` or `allowlist_include_members` are set, anything that is not included is rejected.
- **PATRONI\_RESTAPI\_ALLOWLIST\_INCLUDE\_MEMBERS**: (optional): If set to `true` it allows accessing unsafe REST API endpoints from other cluster members registered in DCS (IP address or hostname is taken from the members `api_url`). Be careful, it might happen that OS will use a different IP for outgoing connections.
- **PATRONI\_RESTAPI\_HTTP\_EXTRA\_HEADERS**: (optional) HTTP headers let the REST API server pass additional information with an HTTP response.
- **PATRONI\_RESTAPI\_HTTPS\_EXTRA\_HEADERS**: (optional) HTTPS headers let the REST API server pass additional information with an HTTP response when TLS is enabled. This will also pass additional information set in `http_extra_headers`.
- **PATRONI\_RESTAPI\_REQUEST\_QUEUE\_SIZE**: (optional): Sets request queue size for TCP socket used by Patroni REST API. Once the queue is full, further requests get a “Connection denied” error. The default value is 5.

#### Warning

- The `PATRONI_RESTAPI_CONNECT_ADDRESS` must be accessible from all nodes of a given Patroni cluster. Internally Patroni is using it during the leader race to find nodes with minimal replication lag.
- If you enabled client certificates validation (`PATRONI_RESTAPI_VERIFY_CLIENT` is set to `required`), you also **must** provide **valid client certificates** in the `PATRONI_CTL_CERTFILE`, `PATRONI_CTL_KEYFILE`, `PATRONI_CTL_KEYFILE_PASSWORD`. If not provided, Patroni will not work correctly.

### 3.3.13 CTL

- **PATRONICTL\_CONFIG\_FILE**: (optional) location of the configuration file.
- **PATRONI\_CTL\_USERNAME**: (optional) Basic-auth username for accessing protected REST API endpoints. If not provided *patronictl* will use the value provided for REST API “username” parameter.
- **PATRONI\_CTL\_PASSWORD**: (optional) Basic-auth password for accessing protected REST API endpoints. If not provided *patronictl* will use the value provided for REST API “password” parameter.
- **PATRONI\_CTL\_INSECURE**: (optional) Allow connections to REST API without verifying SSL certs.
- **PATRONI\_CTL\_CACERT**: (optional) Specifies the file with the CA\_BUNDLE file or directory with certificates of trusted CAs to use while verifying REST API SSL certs. If not provided *patronictl* will use the value provided for REST API “cafile” parameter.
- **PATRONI\_CTL\_CERTFILE**: (optional) Specifies the file with the client certificate in the PEM format.
- **PATRONI\_CTL\_KEYFILE**: (optional) Specifies the file with the client secret key in the PEM format.
- **PATRONI\_CTL\_KEYFILE\_PASSWORD**: (optional) Specifies a password for decrypting the client keyfile.

There are 3 types of Patroni configuration:

- **Global *dynamic configuration*.**  
These options are stored in the DCS (Distributed Configuration Store) and applied on all cluster nodes. Dynamic configuration can be set at any time using *patronictl edit-config* tool or Patroni *REST API*. If

the options changed are not part of the startup configuration, they are applied asynchronously (upon the next wake up cycle) to every node, which gets subsequently reloaded. If the node requires a restart to apply the configuration (for [PostgreSQL parameters](#) with context postmaster, if their values have changed), a special flag `pending_restart` indicating this is set in the `members.data` JSON. Additionally, the node status indicates this by showing `"restart_pending": true`.

- **Local *configuration file* (`patroni.yml`).**

These options are defined in the configuration file and take precedence over dynamic configuration. `patroni.yml` can be changed and reloaded at runtime (without restart of Patroni) by sending SIGHUP to the Patroni process, performing `POST /reload` REST-API request or executing `patronictl reload`. Local configuration can be either a single YAML file or a directory. When it is a directory, all YAML files in that directory are loaded one by one in sorted order. In case a key is defined in multiple files, the occurrence in the last file takes precedence.

- ***Environment configuration.***

It is possible to set/override some of the “Local” configuration parameters with environment variables. Environment configuration is very useful when you are running in a dynamic environment and you don’t know some of the parameters in advance (for example it’s not possible to know your external IP address when you are running inside docker).

## 3.4 Important rules

### 3.4.1 PostgreSQL parameters controlled by Patroni

Some of the PostgreSQL parameters **must hold the same values on the primary and the replicas**. For those, **values set either in the local patroni configuration files or via the environment variables take no effect**. To alter or set their values one must change the shared configuration in the DCS. Below is the actual list of such parameters together with the default values:

- `max_connections`: 100
- `max_locks_per_transaction`: 64
- `max_worker_processes`: 8
- `max_prepared_transactions`: 0
- `wal_level`: `hot_standby`
- `track_commit_timestamp`: `off`

For the parameters below, PostgreSQL does not require equal values among the primary and all the replicas. However, considering the possibility of a replica to become the primary at any time, it doesn’t really make sense to set them differently; therefore, **Patroni restricts setting their values to the *dynamic configuration***.

- `max_wal_senders`: 10
- `max_replication_slots`: 10
- `wal_keep_segments`: 8
- `wal_keep_size`: 128MB

These parameters are validated to ensure they are sane, or meet a minimum value.

There are some other Postgres parameters controlled by Patroni:

- `listen_addresses` - is set either from `postgresql.listen` or from `PATRONI_POSTGRESQL_LISTEN` environment variable

- **port** - is set either from `postgresql.listen` or from `PATRONI_POSTGRESQL_LISTEN` environment variable
- **cluster\_name** - is set either from `scope` or from `PATRONI_SCOPE` environment variable
- **hot\_standby: on**

To be on the safe side parameters from the above lists are not written into `postgresql.conf`, but passed as a list of arguments to the `pg_ctl start` which gives them the highest precedence, even above `ALTER SYSTEM`

There also are some parameters like **`postgresql.listen`**, **`postgresql.data_dir`** that **can be set only locally**, i.e. in the Patroni *config file* or via *configuration* variable. In most cases the local configuration will override the dynamic configuration.

When applying the local or dynamic configuration options, the following actions are taken:

- The node first checks if there is a `postgresql.base.conf` file or if the `custom_conf` parameter is set.
- If the `custom_conf` parameter is set, the file it specifies is used as the base configuration, ignoring `postgresql.base.conf` and `postgresql.conf`.
- If the `custom_conf` parameter is not set and `postgresql.base.conf` exists, it contains the renamed “original” configuration and is used as the base configuration.
- If there is no `custom_conf` nor `postgresql.base.conf`, the original `postgresql.conf` is renamed to `postgresql.base.conf` and used as the base configuration.
- The dynamic options (with the exceptions above) are dumped into the `postgresql.conf` and an include is set in `postgresql.conf` to the base configuration (either `postgresql.base.conf` or the file at `custom_conf`). Therefore, we would be able to apply new options without re-reading the configuration file to check if the include is present or not.
- Some parameters that are essential for Patroni to manage the cluster are overridden using the command line.
- If an option that requires restart is changed (we should look at the context in `pg_settings` and at the actual values of those options), a `pending_restart` flag is set on that node. This flag is reset on any restart.

The parameters would be applied in the following order (run-time are given the highest priority):

1. load parameters from file `postgresql.base.conf` (or from a `custom_conf` file, if set)
2. load parameters from file `postgresql.conf`
3. load parameters from file `postgresql.auto.conf`
4. run-time parameter using `-o -name=value`

This allows configuration for all the nodes (2), configuration for a specific node using `ALTER SYSTEM` (3) and ensures that parameters essential to the running of Patroni are enforced (4), as well as leaves room for configuration tools that manage `postgresql.conf` directly without involving Patroni (1).

### 3.4.2 PostgreSQL parameters that touch shared memory

PostgreSQL has some parameters that determine the size of the shared memory used by them:

- **`max_connections`**
- **`max_prepared_transactions`**
- **`max_locks_per_transaction`**
- **`max_wal_senders`**
- **`max_worker_processes`**

Changing these parameters require a PostgreSQL restart to take effect, and their shared memory structures cannot be smaller on the standby nodes than on the primary node.

As explained before, Patroni restrict changing their values through *dynamic configuration*, which usually consists of:

1. Applying changes through *patronictl edit-config* (or via REST API /config endpoint)
2. Restarting nodes through *patronictl restart* (or via REST API /restart endpoint)

**Note:** please keep in mind that you should perform a restart of the PostgreSQL nodes through *patronictl restart* command, or via REST API /restart endpoint. An attempt to restart PostgreSQL by restarting the Patroni daemon, e.g. by executing `systemctl restart patroni`, can cause a failover to occur in the cluster, if you are restarting the primary node.

However, as those settings manage shared memory, some extra care should be taken when restarting the nodes:

- If you want to **increase** the value of any of those settings:
  1. Restart all standbys first
  2. Restart the primary after that
- If you want to **decrease** the value of any of those settings:
  1. Restart the primary first
  2. Restart all standbys after that

**Note:** if you attempt to restart all nodes in one go after **decreasing** the value of any of those settings, Patroni will ignore the change and restart the standby with the original setting value, thus requiring that you restart the standbys again later. Patroni does that to prevent the standby to enter in an infinite crash loop, because PostgreSQL quits with a *FATAL* message if you attempt to set any of those parameters to a value lower than what is visible in `pg_controldata` on the Standby node. In other words, we can only decrease the setting on the standby once its `pg_controldata` is up-to-date with the primary in regards to these changes on the primary.

More information about that can be found at [PostgreSQL Administrator's Overview](#).

### 3.4.3 Patroni configuration parameters

Also the following Patroni configuration options **can be changed only dynamically**:

- `ttl`: 30
- `loop_wait`: 10
- `retry_timeouts`: 10
- `maximum_lag_on_failover`: 1048576
- `max_timelines_history`: 0
- `check_timeline`: false
- `postgresql.use_slots`: true

Upon changing these options, Patroni will read the relevant section of the configuration stored in DCS and change its run-time values.

Patroni nodes are dumping the state of the DCS options to disk upon for every change of the configuration into the file `patroni.dynamic.json` located in the Postgres data directory. Only the leader is allowed to restore these options from the on-disk dump if these are completely absent from the DCS or if they are invalid.

## 3.5 Configuration generation and validation

Patroni provides command-line interfaces for a Patroni *local configuration* generation and validation. Using the `patroni` executable you can:

- Create a sample local Patroni configuration;
- Create a Patroni configuration file for the locally running PostgreSQL instance (e.g. as a preparation step for the *Patroni integration*);
- Validate a given Patroni configuration file.

### 3.5.1 Sample Patroni configuration

```
patroni --generate-sample-config [configfile]
```

#### Description

Generate a sample Patroni configuration file in `yaml` format. Parameter values are defined using the *Environment configuration*, otherwise, if not set, the defaults used in Patroni or the `#FIXME` string for the values that should be later defined by the user.

Some default values are defined based on the local setup:

- **postgresql.listen**: the IP address returned by `gethostname` call for the current machine's hostname and the standard 5432 port.
- **postgresql.connect\_address**: the IP address returned by `gethostname` call for the current machine's hostname and the standard 5432 port.
- **postgresql.authentication.rewind**: is only defined if the PostgreSQL version can be defined from the binary and the version is 11 or later.
- **restapi.listen**: IP address returned by `gethostname` call for the current machine's hostname and the standard 8008 port.
- **restapi.connect\_address**: IP address returned by `gethostname` call for the current machine's hostname and the standard 8008 port.

#### Parameters

`configfile` - full path to the configuration file used to store the result. If not provided, the result is sent to `stdout`.

### 3.5.2 Patroni configuration for a running instance

```
patroni --generate-config [--dsn DSN] [configfile]
```

## Description

Generate a Patroni configuration in `yaml` format for the locally running PostgreSQL instance. Either the provided DSN (takes precedence) or PostgreSQL [environment variables](#) will be used for the PostgreSQL connection. If the password is not provided, it should be entered via prompt.

All the non-internal GUCs defined in the source Postgres instance, independently if they were set through a configuration file, through the postmaster command-line, or through environment variables, will be used as the source for the following Patroni configuration parameters:

- **scope:** `cluster_name` GUC value;
- **postgresql.listen:** `listen_addresses` and `port` GUC values;
- **postgresql.datadir:** `data_directory` GUC value;
- **postgresql.parameters:** `archive_command`, `restore_command`, `archive_cleanup_command`, `recovery_end_command`, `ssl_passphrase_command`, `hba_file`, `ident_file`, `config_file` GUC values;
- **bootstrap.dcs:** all other gathered PostgreSQL GUCs.

If `scope`, `postgresql.listen` or `postgresql.datadir` is not set from the Postgres GUCs, the respective [Environment configuration](#) value is used.

Other rules applied for the values definition:

- **name:** `PATRONI_NAME` environment variable value if set, otherwise the current machine's hostname.
- **postgresql.bin\_dir:** path to the Postgres binaries gathered from the running instance.
- **postgresql.connect\_address:** the IP address returned by `gethostname` call for the current machine's hostname and the port used for the instance connection or the `port` GUC value.
- **postgresql.authentication.superuser:** the configuration used for the instance connection;
- **postgresql.pg\_hba:** the lines gathered from the source instance's `hba_file`.
- **postgresql.pg\_ident:** the lines gathered from the source instance's `ident_file`.
- **restapi.listen:** IP address returned by `gethostname` call for the current machine's hostname and the standard 8008 port.
- **restapi.connect\_address:** IP address returned by `gethostname` call for the current machine's hostname and the standard 8008 port.

Other parameters defined using [Environment configuration](#) are also included into the configuration.

## Parameters

### configfile

Full path to the configuration file used to store the result. If not provided, result is sent to `stdout`.

### dsn

Optional DSN string for the local PostgreSQL instance to get GUC values from.

### 3.5.3 Validate Patroni configuration

```
patroni --validate-config [configfile] [--ignore-listen-port | -i]
```

#### Description

Validate the given Patroni configuration and print the information about the failed checks.

#### Parameters

##### **configfile**

Full path to the configuration file to check. If not given or file does not exist, will try to read from the PATRONI\_CONFIG\_VARIABLE environment variable or, if not set, from the *Patroni environment variables*.

##### **--ignore-listen-port | -i**

Optional flag to ignore bind failures for listen ports that are already in use when validating the configfile.



## PATRONI REST API

Patroni has a rich REST API, which is used by Patroni itself during the leader race, by the *patronictl* tool in order to perform failovers/switchovers/reinitialize/restarts/reloads, by HAProxy or any other kind of load balancer to perform HTTP health checks, and of course could also be used for monitoring. Below you will find the list of Patroni REST API endpoints.

### 4.1 Health check endpoints

For all health check GET requests Patroni returns a JSON document with the status of the node, along with the HTTP status code. If you don't want or don't need the JSON document, you might consider using the HEAD or OPTIONS method instead of GET.

- The following requests to Patroni REST API will return HTTP status code **200** only when the Patroni node is running as the primary with leader lock:
  - GET /
  - GET /primary
  - GET /read-write
- GET /standby-leader: returns HTTP status code **200** only when the Patroni node is running as the leader in a *standby cluster*.
- GET /leader: returns HTTP status code **200** when the Patroni node has the leader lock. The major difference from the two previous endpoints is that it doesn't take into account whether PostgreSQL is running as the **primary** or the **standby\_leader**.
- GET /replica: replica health check endpoint. It returns HTTP status code **200** only when the Patroni node is in the state **running**, the role is **replica** and **noheartbeat** tag is not set.
- GET /replica?lag=<max-lag>: replica check endpoint. In addition to checks from **replica**, it also checks replication latency and returns status code **200** only when it is below specified value. The key `cluster.last_leader_operation` from DCS is used for Leader wal position and compute latency on replica for performance reasons. `max-lag` can be specified in bytes (integer) or in human readable values, for e.g. 16kB, 64MB, 1GB.
  - GET /replica?lag=1048576
  - GET /replica?lag=1024kB
  - GET /replica?lag=10MB
  - GET /replica?lag=1GB

- GET /replica?tag\_key1=value1&tag\_key2=value2: replica check endpoint. In addition, It will also check for user defined tags key1 and key2 and their respective values in the **tags** section of the yaml configuration management. If the tag isn't defined for an instance, or if the value in the yaml configuration doesn't match the querying value, it will return HTTP Status Code 503.

In the following requests, since we are checking for the leader or standby-leader status, Patroni doesn't apply any of the user defined tags and they will be ignored.

- GET /?tag\_key1=value1&tag\_key2=value2
- GET /leader?tag\_key1=value1&tag\_key2=value2
- GET /primary?tag\_key1=value1&tag\_key2=value2
- GET /read-write?tag\_key1=value1&tag\_key2=value2
- GET /standby\_leader?tag\_key1=value1&tag\_key2=value2
- GET /standby-leader?tag\_key1=value1&tag\_key2=value2
- GET /read-only: like the above endpoint, but also includes the primary.
- GET /synchronous or GET /sync: returns HTTP status code **200** only when the Patroni node is running as a synchronous standby.
- GET /read-only-sync: like the above endpoint, but also includes the primary.
- GET /quorum: returns HTTP status code **200** only when this Patroni node is listed as a quorum node in synchronous\_standby\_names on the primary.
- GET /read-only-quorum: like the above endpoint, but also includes the primary.
- GET /asynchronous or GET /async: returns HTTP status code **200** only when the Patroni node is running as an asynchronous standby.
- GET /asynchronous?lag=<max-lag> or GET /async?lag=<max-lag>: asynchronous standby check endpoint. In addition to checks from asynchronous or async, it also checks replication latency and returns status code **200** only when it is below specified value. The key cluster.last\_leader\_operation from DCS is used for Leader wal position and compute latency on replica for performance reasons. max-lag can be specified in bytes (integer) or in human readable values, for e.g. 16kB, 64MB, 1GB.
  - GET /async?lag=1048576
  - GET /async?lag=1024kB
  - GET /async?lag=10MB
  - GET /async?lag=1GB
- GET /health: returns HTTP status code **200** only when PostgreSQL is up and running.
- GET /liveness: returns HTTP status code **200** if Patroni heartbeat loop is properly running and **503** if the last run was more than ttl seconds ago on the primary or 2\*ttl on the replica. Could be used for livenessProbe.
- GET /readiness: returns HTTP status code **200** when the Patroni node is running as the leader or when PostgreSQL is up and running. The endpoint could be used for readinessProbe when it is not possible to use Kubernetes endpoints for leader elections (OpenShift).

Both, readiness and liveness endpoints are very light-weight and not executing any SQL. Probes should be configured in such a way that they start failing about time when the leader key is expiring. With the default value of ttl, which is 30s example probes would look like:

```

readinessProbe:
  httpGet:
    scheme: HTTP
    path: /readiness
    port: 8008
  initialDelaySeconds: 3
  periodSeconds: 10
  timeoutSeconds: 5
  successThreshold: 1
  failureThreshold: 3
livenessProbe:
  httpGet:
    scheme: HTTP
    path: /liveness
    port: 8008
  initialDelaySeconds: 3
  periodSeconds: 10
  timeoutSeconds: 5
  successThreshold: 1
  failureThreshold: 3

```

## 4.2 Monitoring endpoint

The GET `/patroni` is used by Patroni during the leader race. It also could be used by your monitoring system. The JSON document produced by this endpoint has the same structure as the JSON produced by the health check endpoints.

**Example:** A healthy cluster

```

$ curl -s http://localhost:8008/patroni | jq .
{
  "state": "running",
  "postmaster_start_time": "2024-08-28 19:39:26.352526+00:00",
  "role": "primary",
  "server_version": 160004,
  "xlog": {
    "location": 67395656
  },
  "timeline": 1,
  "replication": [
    {
      "username": "replicator",
      "application_name": "patroni2",
      "client_addr": "10.89.0.6",
      "state": "streaming",
      "sync_state": "async",
      "sync_priority": 0
    },
    {
      "username": "replicator",
      "application_name": "patroni3",
      "client_addr": "10.89.0.2",

```

(continues on next page)

(continued from previous page)

```

    "state": "streaming",
    "sync_state": "async",
    "sync_priority": 0
  }
],
"dcs_last_seen": 1692356718,
"tags": {
  "clonefrom": true
},
"database_system_identifier": "7268616322854375442",
"patroni": {
  "version": "4.0.0",
  "scope": "demo",
  "name": "patroni1"
}
}

```

**Example:** An unlocked cluster

```

$ curl -s http://localhost:8008/patroni | jq .
{
  "state": "running",
  "postmaster_start_time": "2024-08-28 19:39:26.352526+00:00",
  "role": "replica",
  "server_version": 160004,
  "xlog": {
    "received_location": 67419744,
    "replayed_location": 67419744,
    "replayed_timestamp": null,
    "paused": false
  },
  "timeline": 1,
  "replication": [
    {
      "username": "replicator",
      "application_name": "patroni2",
      "client_addr": "10.89.0.6",
      "state": "streaming",
      "sync_state": "async",
      "sync_priority": 0
    },
    {
      "username": "replicator",
      "application_name": "patroni3",
      "client_addr": "10.89.0.2",
      "state": "streaming",
      "sync_state": "async",
      "sync_priority": 0
    }
  ],
  "cluster_unlocked": true,
  "dcs_last_seen": 1692356928,

```

(continues on next page)

(continued from previous page)

```

"tags": {
  "clonefrom": true
},
"database_system_identifier": "7268616322854375442",
"patroni": {
  "version": "4.0.0",
  "scope": "demo",
  "name": "patroni1"
}
}

```

**Example:** An unlocked cluster with *DCS failsafe mode* enabled

```

$ curl -s http://localhost:8008/patroni | jq .
{
  "state": "running",
  "postmaster_start_time": "2024-08-28 19:39:26.352526+00:00",
  "role": "replica",
  "server_version": 160004,
  "xlog": {
    "location": 67420024
  },
  "timeline": 1,
  "replication": [
    {
      "username": "replicator",
      "application_name": "patroni2",
      "client_addr": "10.89.0.6",
      "state": "streaming",
      "sync_state": "async",
      "sync_priority": 0
    },
    {
      "username": "replicator",
      "application_name": "patroni3",
      "client_addr": "10.89.0.2",
      "state": "streaming",
      "sync_state": "async",
      "sync_priority": 0
    }
  ],
  "cluster_unlocked": true,
  "failsafe_mode_is_active": true,
  "dcs_last_seen": 1692356928,
  "tags": {
    "clonefrom": true
  },
  "database_system_identifier": "7268616322854375442",
  "patroni": {
    "version": "4.0.0",
    "scope": "demo",
    "name": "patroni1"
  }
}

```

(continues on next page)

(continued from previous page)

```
}
}
```

**Example:** A cluster with the *pause mode* enabled

```
$ curl -s http://localhost:8008/patroni | jq .
{
  "state": "running",
  "postmaster_start_time": "2024-08-28 19:39:26.352526+00:00",
  "role": "replica",
  "server_version": 160004,
  "xlog": {
    "location": 67420024
  },
  "timeline": 1,
  "replication": [
    {
      "username": "replicator",
      "application_name": "patroni2",
      "client_addr": "10.89.0.6",
      "state": "streaming",
      "sync_state": "async",
      "sync_priority": 0
    },
    {
      "username": "replicator",
      "application_name": "patroni3",
      "client_addr": "10.89.0.2",
      "state": "streaming",
      "sync_state": "async",
      "sync_priority": 0
    }
  ],
  "pause": true,
  "dcs_last_seen": 1724874295,
  "tags": {
    "clonefrom": true
  },
  "database_system_identifier": "7268616322854375442",
  "patroni": {
    "version": "4.0.0",
    "scope": "demo",
    "name": "patroni1"
  }
}
```

Retrieve the Patroni metrics in Prometheus format through the GET `/metrics` endpoint.

```
$ curl http://localhost:8008/metrics

# HELP patroni_version Patroni semver without periods. \
# TYPE patroni_version gauge
```

(continues on next page)

(continued from previous page)

```

patroni_version{scope="batman",name="patroni1"} 040000
# HELP patroni_postgres_running Value is 1 if Postgres is running, 0 otherwise.
# TYPE patroni_postgres_running gauge
patroni_postgres_running{scope="batman",name="patroni1"} 1
# HELP patroni_postmaster_start_time Epoch seconds since Postgres started.
# TYPE patroni_postmaster_start_time gauge
patroni_postmaster_start_time{scope="batman",name="patroni1"} 1724873966.352526
# HELP patroni_primary Value is 1 if this node is the leader, 0 otherwise.
# TYPE patroni_primary gauge
patroni_primary{scope="batman",name="patroni1"} 1
# HELP patroni_xlog_location Current location of the Postgres transaction log, 0 if this_
↳node is not the leader.
# TYPE patroni_xlog_location counter
patroni_xlog_location{scope="batman",name="patroni1"} 22320573386952
# HELP patroni_standby_leader Value is 1 if this node is the standby_leader, 0 otherwise.
# TYPE patroni_standby_leader gauge
patroni_standby_leader{scope="batman",name="patroni1"} 0
# HELP patroni_replica Value is 1 if this node is a replica, 0 otherwise.
# TYPE patroni_replica gauge
patroni_replica{scope="batman",name="patroni1"} 0
# HELP patroni_sync_standby Value is 1 if this node is a sync standby replica, 0_
↳otherwise.
# TYPE patroni_sync_standby gauge
patroni_sync_standby{scope="batman",name="patroni1"} 0
# HELP patroni_quorum_standby Value is 1 if this node is a quorum standby replica, 0_
↳otherwise.
# TYPE patroni_quorum_standby gauge
patroni_quorum_standby{scope="batman",name="patroni1"} 0
# HELP patroni_xlog_received_location Current location of the received Postgres_
↳transaction log, 0 if this node is not a replica.
# TYPE patroni_xlog_received_location counter
patroni_xlog_received_location{scope="batman",name="patroni1"} 0
# HELP patroni_xlog_replayed_location Current location of the replayed Postgres_
↳transaction log, 0 if this node is not a replica.
# TYPE patroni_xlog_replayed_location counter
patroni_xlog_replayed_location{scope="batman",name="patroni1"} 0
# HELP patroni_xlog_replayed_timestamp Current timestamp of the replayed Postgres_
↳transaction log, 0 if null.
# TYPE patroni_xlog_replayed_timestamp gauge
patroni_xlog_replayed_timestamp{scope="batman",name="patroni1"} 0
# HELP patroni_xlog_paused Value is 1 if the Postgres xlog is paused, 0 otherwise.
# TYPE patroni_xlog_paused gauge
patroni_xlog_paused{scope="batman",name="patroni1"} 0
# HELP patroni_postgres_streaming Value is 1 if Postgres is streaming, 0 otherwise.
# TYPE patroni_postgres_streaming gauge
patroni_postgres_streaming{scope="batman",name="patroni1"} 1
# HELP patroni_postgres_in_archive_recovery Value is 1 if Postgres is replicating from_
↳archive, 0 otherwise.
# TYPE patroni_postgres_in_archive_recovery gauge
patroni_postgres_in_archive_recovery{scope="batman",name="patroni1"} 0
# HELP patroni_postgres_server_version Version of Postgres (if running), 0 otherwise.
# TYPE patroni_postgres_server_version gauge

```

(continues on next page)

(continued from previous page)

```

patroni_postgres_server_version{scope="batman",name="patroni1"} 160004
# HELP patroni_cluster_unlocked Value is 1 if the cluster is unlocked, 0 if locked.
# TYPE patroni_cluster_unlocked gauge
patroni_cluster_unlocked{scope="batman",name="patroni1"} 0
# HELP patroni_postgres_timeline Postgres timeline of this node (if running), 0
↳ otherwise.
# TYPE patroni_postgres_timeline counter
patroni_failsafe_mode_is_active{scope="batman",name="patroni1"} 0
# HELP patroni_postgres_timeline Postgres timeline of this node (if running), 0
↳ otherwise.
# TYPE patroni_postgres_timeline counter
patroni_postgres_timeline{scope="batman",name="patroni1"} 24
# HELP patroni_dcs_last_seen Epoch timestamp when DCS was last contacted successfully by
↳ Patroni.
# TYPE patroni_dcs_last_seen gauge
patroni_dcs_last_seen{scope="batman",name="patroni1"} 1724874235
# HELP patroni_pending_restart Value is 1 if the node needs a restart, 0 otherwise.
# TYPE patroni_pending_restart gauge
patroni_pending_restart{scope="batman",name="patroni1"} 1
# HELP patroni_is_paused Value is 1 if auto failover is disabled, 0 otherwise.
# TYPE patroni_is_paused gauge
patroni_is_paused{scope="batman",name="patroni1"} 1

```

### 4.3 Cluster status endpoints

- The GET /cluster endpoint generates a JSON document describing the current cluster topology and state:

```

$ curl -s http://localhost:8008/cluster | jq .
{
  "members": [
    {
      "name": "patroni1",
      "role": "leader",
      "state": "running",
      "api_url": "http://10.89.0.4:8008/patroni",
      "host": "10.89.0.4",
      "port": 5432,
      "timeline": 5,
      "tags": {
        "clonefrom": true
      }
    },
    {
      "name": "patroni2",
      "role": "replica",
      "state": "streaming",
      "api_url": "http://10.89.0.6:8008/patroni",
      "host": "10.89.0.6",
      "port": 5433,
      "timeline": 5,

```

(continues on next page)



(continued from previous page)

```

    "tags": {
      "clonefrom": true
    },
    "lag": 0
  }
],
"scope": "demo",
"scheduled_switchover": {
  "at": "2023-09-24T10:36:00+02:00",
  "from": "patroni1",
  "to": "patroni3"
}
}

```

- The GET /history endpoint provides a view on the history of cluster switchovers/failovers. The format is very similar to the content of history files in the pg\_wal directory. The only difference is the timestamp field showing when the new timeline was created.

```

$ curl -s http://localhost:8008/history | jq .
[
  [
    1,
    25623960,
    "no recovery target specified",
    "2019-09-23T16:57:57+02:00"
  ],
  [
    2,
    25624344,
    "no recovery target specified",
    "2019-09-24T09:22:33+02:00"
  ],
  [
    3,
    25624752,
    "no recovery target specified",
    "2019-09-24T09:26:15+02:00"
  ],
  [
    4,
    50331856,
    "no recovery target specified",
    "2019-09-24T09:35:52+02:00"
  ]
]

```

## 4.4 Config endpoint

GET /config: Get the current version of the dynamic configuration:

```
$ curl -s http://localhost:8008/config | jq .
{
  "ttl": 30,
  "loop_wait": 10,
  "retry_timeout": 10,
  "maximum_lag_on_failover": 1048576,
  "postgresql": {
    "use_slots": true,
    "use_pg_rewind": true,
    "parameters": {
      "hot_standby": "on",
      "wal_level": "hot_standby",
      "max_wal_senders": 5,
      "max_replication_slots": 5,
      "max_connections": "100"
    }
  }
}
```

PATCH /config: Change the existing configuration.

```
$ curl -s -XPATCH -d \
    '{"loop_wait":5,"ttl":20,"postgresql":{"parameters":{"max_connections":"101"}}}' \
    http://localhost:8008/config | jq .
{
  "ttl": 20,
  "loop_wait": 5,
  "maximum_lag_on_failover": 1048576,
  "retry_timeout": 10,
  "postgresql": {
    "use_slots": true,
    "use_pg_rewind": true,
    "parameters": {
      "hot_standby": "on",
      "wal_level": "hot_standby",
      "max_wal_senders": 5,
      "max_replication_slots": 5,
      "max_connections": "101"
    }
  }
}
```

The above REST API call patches the existing configuration and returns the new configuration.

Let's check that the node processed this configuration. First of all it should start printing log lines every 5 seconds (loop\_wait=5). The change of "max\_connections" requires a restart, so the "pending\_restart" flag should be exposed:

```
$ curl -s http://localhost:8008/patroni | jq .
{
```

(continues on next page)

(continued from previous page)

```

"database_system_identifier": "6287881213849985952",
"postmaster_start_time": "2024-08-28 19:39:26.352526+00:00",
"xlog": {
  "location": 2197818976
},
"timeline": 1,
"dcs_last_seen": 1724874545,
"database_system_identifier": "7408277255830290455",
"pending_restart": true,
"pending_restart_reason": {
  "max_connections": {
    "old_value": "100",
    "new_value": "101"
  }
},
"patroni": {
  "version": "4.0.0",
  "scope": "batman",
  "name": "patroni1"
},
"state": "running",
"role": "primary",
"server_version": 160004
}

```

Removing parameters:

If you want to remove (reset) some setting just patch it with null:

```

$ curl -s -XPATCH -d \
    '{"postgresql":{"parameters":{"max_connections":null}}}' \
    http://localhost:8008/config | jq .
{
  "ttl": 20,
  "loop_wait": 5,
  "retry_timeout": 10,
  "maximum_lag_on_failover": 1048576,
  "postgresql": {
    "use_slots": true,
    "use_pg_rewind": true,
    "parameters": {
      "hot_standby": "on",
      "unix_socket_directories": ".",
      "wal_level": "hot_standby",
      "max_wal_senders": 5,
      "max_replication_slots": 5
    }
  }
}

```

The above call removes `postgresql.parameters.max_connections` from the dynamic configuration.

PUT `/config`: It's also possible to perform the full rewrite of an existing dynamic configuration unconditionally:

```
$ curl -s -XPUT -d \
    '{"maximum_lag_on_failover":1048576,"retry_timeout":10,"postgresql":{"use_slots
→:true,"use_pg_rewind":true,"parameters":{"hot_standby":"on","wal_level":"hot_standby",
→"unix_socket_directories":".", "max_wal_senders":5}},"loop_wait":3,"ttl":20}' \
    http://localhost:8008/config | jq .
{
  "ttl": 20,
  "maximum_lag_on_failover": 1048576,
  "retry_timeout": 10,
  "postgresql": {
    "use_slots": true,
    "parameters": {
      "hot_standby": "on",
      "unix_socket_directories": ".",
      "wal_level": "hot_standby",
      "max_wal_senders": 5
    },
    "use_pg_rewind": true
  },
  "loop_wait": 3
}
```

## 4.5 Switchover and failover endpoints

### 4.5.1 Switchover

`/switchover` endpoint only works when the cluster is healthy (there is a leader). It also allows to schedule a switchover at a given time.

When calling `/switchover` endpoint a candidate can be specified but is not required, in contrast to `/failover` endpoint. If a candidate is not provided, all the eligible nodes of the cluster will participate in the leader race after the leader stepped down.

In the JSON body of the POST request you must specify the `leader` field. The `candidate` and the `scheduled_at` fields are optional and can be used to schedule a switchover at a specific time.

Depending on the situation, requests might return different HTTP status codes and bodies. Status code **200** is returned when the switchover or failover successfully completed. If the switchover was successfully scheduled, Patroni will return HTTP status code **202**. In case something went wrong, the error status code (one of **400**, **412**, or **503**) will be returned with some details in the response body.

DELETE `/switchover` can be used to delete the currently scheduled switchover.

**Example:** perform a switchover to any healthy standby

```
$ curl -s http://localhost:8008/switchover -XPOST -d '{"leader":"postgresql1"}'
Successfully switched over to "postgresql2"
```

**Example:** perform a switchover to a specific node

```
$ curl -s http://localhost:8008/switchover -XPOST -d \
    '{"leader":"postgresql1","candidate":"postgresql2"}'
Successfully switched over to "postgresql2"
```

**Example:** schedule a switchover from the leader to any other healthy standby in the cluster at a specific time.

```
$ curl -s http://localhost:8008/switchover -XPOST -d \
    '{"leader":"postgresql0","scheduled_at":"2019-09-24T12:00+00"}'
Switchover scheduled
```

## 4.5.2 Failover

`/failover` endpoint can be used to perform a manual failover when there are no healthy nodes (e.g. to an asynchronous standby if all synchronous standbys are not healthy enough to promote). However there is no requirement for a cluster not to have leader - failover can also be run on a healthy cluster.

In the JSON body of the POST request you must specify the `candidate` field. If the `leader` field is specified, a switchover is triggered instead.

**Example:**

```
$ curl -s http://localhost:8008/failover -XPOST -d '{"candidate":"postgresql1"}'
Successfully failed over to "postgresql1"
```

### Warning

*Be very careful* when using this endpoint, as this can cause data loss in certain situations. In most cases, *the switchover endpoint* satisfies the administrator's needs.

POST `/switchover` and POST `/failover` endpoints are used by *patronictl switchover* and *patronictl failover*, respectively.

DELETE `/switchover` is used by *patronictl flush cluster-name switchover*.

Table 1: Failover/Switchover comparison

	Failover	Switchover
Requires leader specified	no	yes
Requires candidate specified	yes	no
Can be run in pause	yes	yes (only to a specific candidate)
Can be scheduled	no	yes (if not in pause)

## 4.5.3 Healthy standby

There are a couple of checks that a member of a cluster should pass to be able to participate in the leader race during a switchover or to become a leader as a failover/switchover candidate:

- be reachable via Patroni API;
- not have `nofailover` tag set to `true`;
- have watchdog fully functional (if required by the configuration);
- in case of a switchover in a healthy cluster or an automatic failover, not exceed maximum replication lag (`maximum_lag_on_failover` *configuration parameter*);
- in case of a switchover in a healthy cluster or an automatic failover, not have a timeline number smaller than the cluster timeline if `check_timeline` *configuration parameter* is set to `true`;

- in *synchronous mode*:
  - In case of a switchover (both with and without a candidate): be listed in the `/sync` key members;
  - For a failover in both healthy and unhealthy clusters, this check is omitted.

#### Warning

**In case of a manual failover in a cluster without a leader, a candidate will be allowed to promote even if:**

- it is not in the `/sync` key members when synchronous mode is enabled;
- its lag exceeds the maximum replication lag allowed;
- it has the timeline number smaller than the last known cluster timeline.

## 4.6 Restart endpoint

- `POST /restart`: You can restart Postgres on the specific node by performing the `POST /restart` call. In the JSON body of POST request it is possible to optionally specify some restart conditions:
  - **restart\_pending**: boolean, if set to `true` Patroni will restart PostgreSQL only when restart is pending in order to apply some changes in the PostgreSQL config.
  - **role**: perform restart only if the current role of the node matches with the role from the POST request.
  - **postgres\_version**: perform restart only if the current version of postgres is smaller than specified in the POST request.
  - **timeout**: how long we should wait before PostgreSQL starts accepting connections. Overrides `primary_start_timeout`.
  - **schedule**: timestamp with time zone, schedule the restart somewhere in the future.
- `DELETE /restart`: delete the scheduled restart

`POST /restart` and `DELETE /restart` endpoints are used by *patronictl restart* and *patronictl flush cluster-name restart* respectively.

## 4.7 Reload endpoint

The `POST /reload` call will order Patroni to re-read and apply the configuration file. This is the equivalent of sending the `SIGHUP` signal to the Patroni process. In case you changed some of the Postgres parameters which require a restart (like `shared_buffers`), you still have to explicitly do the restart of Postgres by either calling the `POST /restart` endpoint or with the help of *patronictl restart*.

The reload endpoint is used by *patronictl reload*.

## 4.8 Reinitialize endpoint

POST /reinitialize: reinitialize the PostgreSQL data directory on the specified node. It is allowed to be executed only on replicas. Once called, it will remove the data directory and start `pg_basebackup` or some alternative *replica creation method*.

The call might fail if Patroni is in a loop trying to recover (restart) a failed Postgres. In order to overcome this problem one can specify `{"force": true}` in the request body.

The reinitialize endpoint is used by *patronictl reinit*.





## PATRONICTL

Patroni has a command-line interface named `patronictl`, which is used basically to interact with Patroni's REST API and with the DCS. It is intended to make it easier to perform operations in the cluster, and can easily be used by humans or scripts.

## 5.1 Configuration

`patronictl` uses 3 sections of the configuration:

- **ctl**: how to authenticate against the Patroni REST API, and how to validate the server identity. Refer to *ctl settings* for more details;
- **restapi**: how to authenticate against the Patroni REST API, and how to validate the server identity. Only used if `ctl` configuration is not enough. `patronictl` is mainly interested in `restapi.authentication` section (in case `ctl.authentication` is missing) and `restapi.cacfile` setting (in case `ctl.cacert` is missing). Refer to *REST API settings* for more details;
- DCS (e.g. **etcd**): how to contact and authenticate against the DCS used by Patroni.

Those configuration options can come either from environment variables or from a configuration file. Look for the above sections in *Environment Configuration Settings* or *YAML Configuration Settings* to understand how you can set the options for them through environment variables or through a configuration file.

If you opt for using environment variables, it's a straight forward approach. `Patronictl` will read the environment variables and use their values.

If you opt for using a configuration file, you have different ways to inform `patronictl` about the file to be used. By default `patronictl` will attempt to load a configuration file named `patronictl.yaml`, which is expected to be found under either of these paths, according to your system:

- Mac OS X: `~/Library/Application Support/patroni`
- Mac OS X (POSIX): `~/ .patroni`
- Unix: `~/ .config/patroni`
- Unix (POSIX): `~/ .patroni`
- Windows (roaming): `C:\Users\<user>\AppData\Roaming\patroni`
- Windows (not roaming): `C:\Users\<user>\AppData\Local\patroni`

You can override that behavior either by:

- Setting the environment variable `PATRONICTL_CONFIG_FILE` with the path to a custom configuration file;
- Using the `-c / --config-file` command-line argument of `patronictl` with the path to a custom configuration file.

**Note**

If you are running `patronictl` in the same host as `patroni` daemon is running, you may just use the same configuration file if it contains all the configuration sections required by `patronictl`.

## 5.2 Usage

`patronictl` exposes several handy operations. This section is intended to describe each of them.

Before jumping into each of the sub-commands of `patronictl`, be aware that `patronictl` itself has the following command-line arguments:

**-c / --config-file**

As explained before, used to provide a path to a configuration file for `patronictl`.

**-d / --dcs-url / --dcs**

Provide a connection string to the DCS used by Patroni.

This argument can be used either to override the DCS and namespace settings from the `patronictl` configuration, or to define it if it's missing in the configuration.

The value should be in the format `DCS://HOST:PORT/NAMESPACE`, e.g. `etcd3://localhost:2379/service` to connect to etcd v3 running on `localhost` with Patroni cluster stored under `service` namespace. Any part that is missing in the argument value will be replaced with the value present in the configuration or with its default.

**-k / --insecure**

Flag to bypass validation of REST API server SSL certificate.

This is the synopsis for running a command from the `patronictl`:

```
patronictl [ { -c | --config-file } CONFIG_FILE ]
           [ { -d | --dcs-url | --dcs } DCS_URL ]
           [ { -k | --insecure } ]
           SUBCOMMAND
```

**Note**

This is the syntax for the synopsis:

- Options between square brackets are optional;
- Options between curly brackets represent a “choose one of set” operation;
- Options with `[ , ... ]` can be specified multiple times;
- Things written in uppercase represent a literal that should be given a value to.

We will use this same syntax when describing `patronictl` sub-commands in the following sub-sections. Also, when describing sub-commands in the following sub-sections, the commands' synopsis should be seen as a replacement for the `SUBCOMMAND` in the above synopsis.

In the following sub-sections you can find a description of each command implemented by `patronictl`. For sake of example, we will use the configuration files present in the GitHub repository of Patroni (files `postgres0.yml`, `postgres1.yml` and `postgres2.yml`).

## 5.2.1 patronictl dsn

### Synopsis

```
dsn
[ CLUSTER_NAME ]
[ { { -r | --role } { leader | primary | standby-leader | replica | standby | any } |
→ { -m | --member } MEMBER_NAME } ]
[ --group CITUS_GROUP ]
```

### Description

`patronictl dsn` gets the connection string for one member of the Patroni cluster.

If multiple members match the parameters of this command, one of them will be chosen, prioritizing the primary node.

### Parameters

#### CLUSTER\_NAME

Name of the Patroni cluster.

If not given, `patronictl` will attempt to fetch that from the scope configuration, if it exists.

#### -r / --role

Choose a member that has the given role.

Role can be one of:

- **leader**: the leader of either a regular Patroni cluster or a standby Patroni cluster; or
- **primary**: the leader of a regular Patroni cluster; or
- **standby-leader**: the leader of a standby Patroni cluster; or
- **replica**: a replica of a Patroni cluster; or
- **standby**: same as **replica**; or
- **any**: any role. Same as omitting this parameter; or

#### -m / --member

Choose a member of the cluster with the given name.

**MEMBER\_NAME** is the name of the member.

#### --group

Choose a member that is part of the given Citus group.

**CITUS\_GROUP** is the ID of the Citus group.

## Examples

Get DSN of the primary node:

```
$ patronictl -c postgres0.yml dsn batman -r primary
host=127.0.0.1 port=5432
```

Get DSN of the node named postgresql1:

```
$ patronictl -c postgres0.yml dsn batman --member postgresql1
host=127.0.0.1 port=5433
```

## 5.2.2 patronictl edit-config

### Synopsis

```
edit-config
[ CLUSTER_NAME ]
[ --group CITUS_GROUP ]
[ { -q | --quiet } ]
[ { -s | --set } CONFIG="VALUE" [, ... ] ]
[ { -p | --pg } PG_CONFIG="PG_VALUE" [, ... ] ]
[ { --apply | --replace } CONFIG_FILE ]
[ --force ]
```

### Description

`patronictl edit-config` changes the dynamic configuration of the cluster and updates the DCS with that.

#### Note

When invoked through a TTY the command attempts to show a diff of the dynamic configuration through a pager. By default, it attempts to use either `less` or `more`. If you want a different pager, set the `PAGER` environment variable with the desired one.

### Parameters

#### **CLUSTER\_NAME**

Name of the Patroni cluster.

If not given, `patronictl` will attempt to fetch that from the `scope` configuration, if it exists.

#### **--group**

Change dynamic configuration of the given Citus group.

If not given, `patronictl` will attempt to fetch that from the `citus.group` configuration, if it exists.

`CITUS_GROUP` is the ID of the Citus group.

#### **-q / --quiet**

Flag to skip showing the configuration diff.

**-s / --set**

Set a given dynamic configuration option with a given value.

CONFIG is the name of the dynamic configuration path in the YAML tree, with levels joined by . .

VALUE is the value for CONFIG. If it is null, then CONFIG will be removed from the dynamic configuration.

**-p / --pg**

Set a given dynamic Postgres configuration option with the given value.

It is essentially a shorthand for --s / --set with CONFIG prepended with postgresql.parameters..

PG\_CONFIG is the name of the Postgres configuration to be set.

PG\_VALUE is the value for PG\_CONFIG. If it is null, then PG\_CONFIG will be removed from the dynamic configuration.

**--apply**

Apply dynamic configuration from the given file.

It is similar to specifying multiple -s / --set options, one for each configuration from CONFIG\_FILE.

CONFIG\_FILE is the path to a file containing the dynamic configuration to be applied, in YAML format. Use - if you want to read from stdin.

**--replace**

Replace the dynamic configuration in the DCS with the dynamic configuration specified in the given file.

CONFIG\_FILE is the path to a file containing the new dynamic configuration to take effect, in YAML format. Use - if you want to read from stdin.

**--force**

Flag to skip confirmation prompts when changing the dynamic configuration.

Useful for scripts.

**Examples**

Change max\_connections Postgres GUC:

```
patronictl -c postgres0.yml edit-config batman --pg max_connections="150" --force
---
+++
@@ -1,6 +1,8 @@
loop_wait: 10
maximum_lag_on_failover: 1048576
postgresql:
+ parameters:
+   max_connections: 150
pg_hba:
- host replication replicator 127.0.0.1/32 md5
- host all all 0.0.0.0/0 md5

Configuration changed
```

Change loop\_wait and ttl settings:

```
patronictl -c postgres0.yml edit-config batman --set loop_wait="15" --set ttl="45" --
↪ force
```

(continues on next page)

(continued from previous page)

```

---
+++
@@ -1,4 +1,4 @@
-loop_wait: 10
+loop_wait: 15
maximum_lag_on_failover: 1048576
postgresql:
  pg_hba:
@@ -6,4 +6,4 @@
- host all all 0.0.0.0/0 md5
  use_pg_rewind: true
retry_timeout: 10
-ttl: 30
+ttl: 45

Configuration changed

```

Remove `maximum_lag_on_failover` setting from dynamic configuration:

```

patronictl -c postgres0.yml edit-config batman --set maximum_lag_on_failover="null" --
↪ force
---
+++
@@ -1,5 +1,4 @@
loop_wait: 10
-maximum_lag_on_failover: 1048576
postgresql:
  pg_hba:
- host replication replicator 127.0.0.1/32 md5

Configuration changed

```

### 5.2.3 patronictl failover

#### Synopsis

```

failover
[ CLUSTER_NAME ]
[ --group CITUS_GROUP ]
--candidate CANDIDATE_NAME
[ --force ]

```

## Description

`patronictl failover` performs a manual failover in the cluster.

It is designed to be used when the cluster is not healthy, e.g.:

- There is no leader; or
- There is no synchronous standby available in a synchronous cluster.

It also allows to fail over to an asynchronous node if synchronous mode is enabled.

### Note

Nothing prevents you from running `patronictl failover` in a healthy cluster. However, we recommend using `patronictl switchover` in those cases.

### Warning

Triggering a failover can cause data loss depending on how up-to-date the promoted replica is in comparison to the primary.

## Parameters

### CLUSTER\_NAME

Name of the Patroni cluster.

If not given, `patronictl` will attempt to fetch that from the `scope` configuration, if it exists.

### --group

Perform a failover in the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

### --candidate

The node to be promoted on failover.

CANDIDATE\_NAME is the name of the node to be promoted.

### --force

Flag to skip confirmation prompts when performing the failover.

Useful for scripts.

## Examples

Fail over to node `postgresql2`:

```
$ patronictl -c postgres0.yml failover batman --candidate postgresql2 --force
Current cluster topology
+ Cluster: batman (7277694203142172922) +-+-----+-----+-----+
| Member      | Host          | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0 | 127.0.0.1:5432 | Leader | running | 3  |           |
| postgresql1 | 127.0.0.1:5433 | Replica | streaming | 3  | 0         |
```

(continues on next page)

(continued from previous page)

```

| postgresql2 | 127.0.0.1:5434 | Replica | streaming | 3 | 0 |
+-----+
2023-09-12 11:52:27.50978 Successfully failed over to "postgresql2"
+ Cluster: batman (7277694203142172922) -+-----+
| Member      | Host           | Role   | State   | TL | Lag in MB |
+-----+
| postgresql0 | 127.0.0.1:5432 | Replica | stopped |    | unknown    |
| postgresql1 | 127.0.0.1:5433 | Replica | running | 3  | 0          |
| postgresql2 | 127.0.0.1:5434 | Leader  | running | 3  |           |
+-----+

```

## 5.2.4 patronictl flush

### Synopsis

```

flush
  CLUSTER_NAME
  [ MEMBER_NAME [, ... ] ]
  { restart | switchover }
  [ --group CITUS_GROUP ]
  [ { -r | --role } { leader | primary | standby-leader | replica | standby | any } ]
  [ --force ]

```

### Description

`patronictl flush` discards scheduled events, if any.

### Parameters

#### CLUSTER\_NAME

Name of the Patroni cluster.

#### MEMBER\_NAME

Discard scheduled events for the given Patroni member(s).

Multiple members can be specified. If no members are specified, all of them are considered.

#### Note

Only used if discarding scheduled restart events.

#### restart

Discard scheduled restart events.

#### switchover

Discard scheduled switchover event.

#### --group

Discard scheduled events from the given Citus group.

CITUS\_GROUP is the ID of the Citus group.



**-r / --role**

Discard scheduled events for members that have the given role.

Role can be one of:

- **leader**: the leader of either a regular Patroni cluster or a standby Patroni cluster; or
- **primary**: the leader of a regular Patroni cluster; or
- **standby-leader**: the leader of a standby Patroni cluster; or
- **replica**: a replica of a Patroni cluster; or
- **standby**: same as **replica**; or
- **any**: any role. Same as omitting this parameter.

**Note**

Only used if discarding scheduled restart events.

**--force**

Flag to skip confirmation prompts when performing the flush.

Useful for scripts.

**Examples**

Discard a scheduled switchover event:

```
$ patronictl -c postgres0.yml flush batman switchover --force
Success: scheduled switchover deleted
```

Discard scheduled restart of all standby nodes:

```
$ patronictl -c postgres0.yml flush batman restart -r replica --force
+ Cluster: batman (7277694203142172922) +-----+-----+-----+-----+
↪-----+
| Member      | Host          | Role   | State   | TL | Lag in MB | Scheduled_
↪restart      |               |        |         |    |           |
+-----+-----+-----+-----+-----+-----+
↪-----+
| postgresql0 | 127.0.0.1:5432 | Leader | running | 5 |           | 2023-09-
↪12T17:17:00+00:00 |
| postgresql1 | 127.0.0.1:5433 | Replica | streaming | 5 | 0 | 2023-09-
↪12T17:17:00+00:00 |
| postgresql2 | 127.0.0.1:5434 | Replica | streaming | 5 | 0 | 2023-09-
↪12T17:17:00+00:00 |
+-----+-----+-----+-----+-----+-----+
↪-----+
Success: flush scheduled restart for member postgresql1
Success: flush scheduled restart for member postgresql2
```

Discard scheduled restart of nodes postgresql0 and postgresql1:

```
$ patronictl -c postgres0.yml flush batman postgresql0 postgresql1 restart --force
+ Cluster: batman (7277694203142172922) -+-----+-----+-----+-----+
↪-----+
| Member      | Host           | Role   | State   | TL | Lag in MB | Scheduled_ |
↪restart      |
+-----+-----+-----+-----+-----+-----+-----+
↪-----+
| postgresql0 | 127.0.0.1:5432 | Leader | running | 5 |          | 2023-09-
↪12T17:17:00+00:00 |
| postgresql1 | 127.0.0.1:5433 | Replica | streaming | 5 |          | 2023-09-
↪12T17:17:00+00:00 |
| postgresql2 | 127.0.0.1:5434 | Replica | streaming | 5 |          | 2023-09-
↪12T17:17:00+00:00 |
+-----+-----+-----+-----+-----+-----+-----+
↪-----+
Success: flush scheduled restart for member postgresql0
Success: flush scheduled restart for member postgresql1
```

## 5.2.5 patronictl history

### Synopsis

```
history
  [ CLUSTER_NAME ]
  [ --group CITUS_GROUP ]
  [ { -f | --format } { pretty | tsv | json | yaml } ]
```

### Description

`patronictl history` shows a history of failover and switchover events from the cluster, if any.

The following information is included in the output:

#### TL

Postgres timeline at which the event occurred.

#### LSN

Postgres LSN at which the event occurred.

#### Reason

Reason fetched from the Postgres `.history` file.

#### Timestamp

Time when the event occurred.

#### New Leader

Patroni member that has been promoted during the event.

## Parameters

### CLUSTER\_NAME

Name of the Patroni cluster.

If not given, `patronictl` will attempt to fetch that from the `scope` configuration, if it exists.

### --group

Show history of events from the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

If not given, `patronictl` will attempt to fetch that from the `citus.group` configuration, if it exists.

### -f / --format

How to format the list of events in the output.

Format can be one of:

- `pretty`: prints history as a pretty table; or
- `tsv`: prints history as tabular information, with columns delimited by `\t`; or
- `json`: prints history in JSON format; or
- `yaml`: prints history in YAML format.

The default is `pretty`.

### --force

Flag to skip confirmation prompts when performing the flush.

Useful for scripts.

## Examples

Show the history of events:

```
$ patronictl -c postgres0.yml history batman
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
↪-----+
| TL |      LSN | Reason                                     | Timestamp                                     | New_↪
↪Leader |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
↪-----+
|  1 | 24392648 | no recovery target specified | 2023-09-11T22:11:27.125527+00:00 | ↪
↪postgres10 |
|  2 | 50331864 | no recovery target specified | 2023-09-12T11:34:03.148097+00:00 | ↪
↪postgres10 |
|  3 | 83886704 | no recovery target specified | 2023-09-12T11:52:26.948134+00:00 | ↪
↪postgres12 |
|  4 | 83887280 | no recovery target specified | 2023-09-12T11:53:09.620136+00:00 | ↪
↪postgres10 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
↪-----+
```

Show the history of events in YAML format:

```
$ patronictl -c postgres0.yml history batman -f yaml
- LSN: 24392648
  New Leader: postgresql0
  Reason: no recovery target specified
  TL: 1
  Timestamp: '2023-09-11T22:11:27.125527+00:00'
- LSN: 50331864
  New Leader: postgresql0
  Reason: no recovery target specified
  TL: 2
  Timestamp: '2023-09-12T11:34:03.148097+00:00'
- LSN: 83886704
  New Leader: postgresql2
  Reason: no recovery target specified
  TL: 3
  Timestamp: '2023-09-12T11:52:26.948134+00:00'
- LSN: 83887280
  New Leader: postgresql0
  Reason: no recovery target specified
  TL: 4
  Timestamp: '2023-09-12T11:53:09.620136+00:00'
```

## 5.2.6 patronictl list

### Synopsis

```
list
[ CLUSTER_NAME [, ... ] ]
[ --group CITUS_GROUP ]
[ { -e | --extended } ]
[ { -t | --timestamp } ]
[ { -f | --format } { pretty | tsv | json | yaml } ]
[ { -W | { -w | --watch } TIME } ]
```

### Description

`patronictl list` shows information about Patroni cluster and its members.

The following information is included in the output:

#### Cluster

Name of the Patroni cluster.

#### Member

Name of the Patroni member.

#### Host

Host where the member is located.

#### Role

Current role of the member.

Can be one among:

- **Leader:** the current leader of a regular Patroni cluster; or
- **Standby Leader:** the current leader of a Patroni standby cluster; or
- **Sync Standby:** a synchronous standby of a Patroni cluster with synchronous mode enabled; or
- **Replica:** a regular standby of a Patroni cluster.

**State**

Current state of Postgres in the Patroni member.

Some examples among the possible states:

- **running:** if Postgres is currently up and running;
- **streaming:** if a replica and Postgres is currently streaming WALs from the primary node;
- **in archive recovery:** if a replica and Postgres is currently fetching WALs from the archive;
- **stopped:** if Postgres had been shut down;
- **crashed:** if Postgres has crashed.

**TL**

Current Postgres timeline in the Patroni member.

**Lag in MB**

Amount worth of replication lag in megabytes between the Patroni member and its upstream.

Besides that, the following information may be included in the output:

**System identifier**

Postgres system identifier.

**Note**

Shown in the table header.

Only shown if output format is `pretty`.

**Group**

Citus group ID.

**Note**

Shown in the table header.

Only shown if a Citus cluster.

**Pending restart**

\* indicates that the node needs a restart for some Postgres configuration to take effect. An empty value indicates the node does not require a restart.

**Note**

Shown as a member attribute.

Shown if:

- Printing in `pretty` or `tsv` format and with extended output enabled; or

- If node requires a restart.

### Scheduled restart

Timestamp at which a restart has been scheduled for the Postgres instance managed by the Patroni member. An empty value indicates there is no scheduled restart for the member.

#### Note

Shown as a member attribute.

Shown if:

- Printing in `pretty` or `tsv` format and with extended output enabled; or
- If node has a scheduled restart.

### Tags

Contains tags set for the Patroni member. An empty value indicates that either no tags have been configured, or that they have been configured with default values.

#### Note

Shown as a member attribute.

Shown if:

- Printing in `pretty` or `tsv` format and with extended output enabled; or
- If node has any custom tags, or any default tags with non-default values.

### Scheduled switchover

Timestamp at which a switchover has been scheduled for the Patroni cluster, if any.

#### Note

Shown in the table footer.

Only shown if there is a scheduled switchover, and output format is `pretty`.

### Maintenance mode

If the cluster monitoring is currently paused.

#### Note

Shown in the table footer.

Only shown if the cluster is paused, and output format is `pretty`.

## Parameters

### CLUSTER\_NAME

Name of the Patroni cluster.

If not given, `patronictl` will attempt to fetch that from the scope configuration, if it exists.

### --group

Show information about members from the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

### -e / --extended

Show extended information.

Force showing `Pending restart`, `Scheduled restart` and `Tags` attributes, even if their value is empty.

#### Note

Only applies to `pretty` and `tsv` output formats.

### -t / --timestamp

Print timestamp before printing information about the cluster and its members.

### -f / --format

How to format the list of events in the output.

Format can be one of:

- `pretty`: prints history as a pretty table; or
- `tsv`: prints history as tabular information, with columns delimited by `\t`; or
- `json`: prints history in JSON format; or
- `yaml`: prints history in YAML format.

The default is `pretty`.

### -W

Automatically refresh information every 2 seconds.

### -w / --watch

Automatically refresh information at the specified interval.

TIME is the interval between refreshes, in seconds.

## Examples

Show information about the cluster in pretty format:

```
$ patronictl -c postgres0.yml list batman
+ Cluster: batman (7277694203142172922) +-----+-----+-----+
| Member      | Host              | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0 | 127.0.0.1:5432    | Leader | running | 5 |           |
| postgresql1 | 127.0.0.1:5433    | Replica | streaming | 5 |           0 |
| postgresql2 | 127.0.0.1:5434    | Replica | streaming | 5 |           0 |
+-----+-----+-----+-----+-----+-----+
```

Show information about the cluster in pretty format with extended columns:

```
$ patronictl -c postgres0.yml list batman -e
+ Cluster: batman (7277694203142172922) -+-----+-----+-----+-----+
↪+-----+-----+
| Member      | Host          | Role   | State   | TL | Lag in MB | Pending restart |
↪| Scheduled restart | Tags |
+-----+-----+-----+-----+-----+-----+-----+
↪+-----+-----+
| postgresql0 | 127.0.0.1:5432 | Leader | running | 5 |          |                 |
↪|              |              |
| postgresql1 | 127.0.0.1:5433 | Replica | streaming | 5 |          0 |                 |
↪|              |              |
| postgresql2 | 127.0.0.1:5434 | Replica | streaming | 5 |          0 |                 |
↪|              |              |
+-----+-----+-----+-----+-----+-----+-----+
↪+-----+-----+
```

Show information about the cluster in YAML format, with timestamp of execution:

```
$ patronictl -c postgres0.yml list batman -f yaml -t
2023-09-12 13:30:48
- Cluster: batman
  Host: 127.0.0.1:5432
  Member: postgresql0
  Role: Leader
  State: running
  TL: 5
- Cluster: batman
  Host: 127.0.0.1:5433
  Lag in MB: 0
  Member: postgresql1
  Role: Replica
  State: streaming
  TL: 5
- Cluster: batman
  Host: 127.0.0.1:5434
  Lag in MB: 0
  Member: postgresql2
  Role: Replica
  State: streaming
  TL: 5
```

## 5.2.7 patronictl pause

### Synopsis

```
pause
[ CLUSTER_NAME ]
[ --group CITUS_GROUP ]
[ --wait ]
```



## Description

`patronictl pause` temporarily puts the Patroni cluster in maintenance mode and disables automatic failover.

## Parameters

### CLUSTER\_NAME

Name of the Patroni cluster.

If not given, `patronictl` will attempt to fetch that from the `scope` configuration, if it exists.

### --group

Pause the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

If not given, `patronictl` will attempt to fetch that from the `citus.group` configuration, if it exists.

### --wait

Wait until all Patroni members are paused before returning control to the caller.

## Examples

Put the cluster in maintenance mode, and wait until all nodes have been paused:

```
$ patronictl -c postgres0.yml pause batman --wait
'pause' request sent, waiting until it is recognized by all nodes
Success: cluster management is paused
```

## 5.2.8 patronictl query

### Synopsis

```
query
[ CLUSTER_NAME ]
[ --group CITUS_GROUP ]
[ { { -r | --role } { leader | primary | standby-leader | replica | standby | any } |
↪ { -m | --member } MEMBER_NAME } ]
[ { -d | --dbname } DBNAME ]
[ { -U | --username } USERNAME ]
[ --password ]
[ --format { pretty | tsv | json | yaml } ]
[ { { -f | --file } FILE_NAME | { -c | --command } SQL_COMMAND } ]
[ --delimiter ]
[ { -W | { -w | --watch } TIME } ]
```

## Description

`patronictl query` executes a SQL command or script against a member of the Patroni cluster.

## Parameters

### **CLUSTER\_NAME**

Name of the Patroni cluster.

If not given, `patronictl` will attempt to fetch that from the `scope` configuration, if it exists.

### **--group**

Query the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

### **-r / --role**

Choose a member that has the given role.

Role can be one of:

- **leader**: the leader of either a regular Patroni cluster or a standby Patroni cluster; or
- **primary**: the leader of a regular Patroni cluster; or
- **standby-leader**: the leader of a standby Patroni cluster; or
- **replica**: a replica of a Patroni cluster; or
- **standby**: same as **replica**; or
- **any**: any role. Same as omitting this parameter.

### **-m / --member**

Choose a member that has the given name.

MEMBER\_NAME is the name of the member to be picked.

### **-d / --dbname**

Database to connect and run the query.

DBNAME is the name of the database. If not given, defaults to `USERNAME`.

### **-U / --username**

User to connect to the database.

USERNAME name of the user. If not given, defaults to the operating system user running `patronictl query`.

### **--password**

Prompt for the password of the connecting user.

As Patroni uses `libpq`, alternatively you can create a `~/.pgpass` file or set the `PGPASSWORD` environment variable.

### **--format**

How to format the output of the query.

Format can be one of:

- **pretty**: prints query output as a pretty table; or
- **tsv**: prints query output as tabular information, with columns delimited by `\t`; or
- **json**: prints query output in JSON format; or
- **yaml**: prints query output in YAML format.

The default is `tsv`.

**-f / --file**

Use a file as source of commands to run queries.

`FILE_NAME` is the path to the source file.

**-c / --command**

Run the given SQL command in the query.

`SQL_COMMAND` is the SQL command to be executed.

**--delimiter**

The delimiter when printing information in `tsv` format, or `\t` if omitted.

**-W**

Automatically re-run the query every 2 seconds.

**-w / --watch**

Automatically re-run the query at the specified interval.

`TIME` is the interval between re-runs, in seconds.

## Examples

Run a SQL command as `postgres` user, and ask for its password:

```
$ patronictl -c postgres0.yml query batman -U postgres --password -c "SELECT now()"
Password:
now
2023-09-12 18:10:53.228084+00:00
```

Run a SQL command as `postgres` user, and take password from `libpq` environment variable:

```
$ PGPASSWORD=patroni patronictl -c postgres0.yml query batman -U postgres -c "SELECT
↪now()"
now
2023-09-12 18:11:37.639500+00:00
```

Run a SQL command and print in pretty format every 2 seconds:

```
$ patronictl -c postgres0.yml query batman -c "SELECT now()" --format pretty -W
+-----+
| now                |
+-----+
| 2023-09-12 18:12:16.716235+00:00 |
+-----+
| now                |
+-----+
| 2023-09-12 18:12:18.732645+00:00 |
+-----+
| now                |
+-----+
| 2023-09-12 18:12:20.750573+00:00 |
+-----+
```

Run a SQL command on database test and print the output in YAML format:

```
$ patronictl -c postgres0.yml query batman -d test -c "SELECT now() AS column_1, 'test' AS column_2" --format yaml
- column_1: 2023-09-12 18:14:22.052060+00:00
  column_2: test
```

Run a SQL command on member postgresql2:

```
$ patronictl -c postgres0.yml query batman -m postgresql2 -c "SHOW port"
port
5434
```

Run a SQL command on any of the standbys:

```
$ patronictl -c postgres0.yml query batman -r replica -c "SHOW port"
port
5433
```

## 5.2.9 patronictl reinit

### Synopsis

```
reinit
  CLUSTER_NAME
  [ MEMBER_NAME [, ... ] ]
  [ --group CITUS_GROUP ]
  [ --wait ]
  [ --force ]
```

### Description

`patronictl reinit` rebuilds a Postgres standby instance managed by a replica member of the Patroni cluster.

### Parameters

#### **CLUSTER\_NAME**

Name of the Patroni cluster.

#### **MEMBER\_NAME**

Name of the replica member for which the Postgres instance will be rebuilt.

Multiple replica members can be specified. If no members are specified, the command does nothing.

#### **--group**

Rebuild a replica member of the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

#### **--wait**

Wait until the reinitialization of the Postgres standby node(s) is finished.

**--force**

Flag to skip confirmation prompts when rebuilding Postgres standby instances.

Useful for scripts.

**Examples**

Request a rebuild of all replica members of the Patroni cluster and immediately return control to the caller:

```
$ patronictl -c postgres0.yml reinit batman postgresql1 postgresql2 --force
+ Cluster: batman (7277694203142172922) -+-----+-----+-----+
| Member      | Host              | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0 | 127.0.0.1:5432    | Leader | running | 5  |           |
| postgresql1 | 127.0.0.1:5433    | Replica | streaming | 5  | 0         |
| postgresql2 | 127.0.0.1:5434    | Replica | streaming | 5  | 0         |
+-----+-----+-----+-----+-----+-----+
Success: reinitialize for member postgresql1
Success: reinitialize for member postgresql2
```

Request a rebuild of postgresql2 and wait for it to complete:

```
$ patronictl -c postgres0.yml reinit batman postgresql2 --wait --force
+ Cluster: batman (7277694203142172922) -+-----+-----+-----+
| Member      | Host              | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0 | 127.0.0.1:5432    | Leader | running | 5  |           |
| postgresql1 | 127.0.0.1:5433    | Replica | streaming | 5  | 0         |
| postgresql2 | 127.0.0.1:5434    | Replica | streaming | 5  | 0         |
+-----+-----+-----+-----+-----+-----+
Success: reinitialize for member postgresql2
Waiting for reinitialize to complete on: postgresql2
Reinitialize is completed on: postgresql2
```

**5.2.10 patronictl reload****Synopsis**

```
reload
  CLUSTER_NAME
  [ MEMBER_NAME [, ... ] ]
  [ --group CITUS_GROUP ]
  [ { -r | --role } { leader | primary | standby-leader | replica | standby | any } ]
  [ --force ]
```

## Description

`patronictl reload` requests a reload of local configuration for one or more Patroni members. It also triggers `pg_ctl reload` on the managed Postgres instance, even if nothing has changed.

## Parameters

### CLUSTER\_NAME

Name of the Patroni cluster.

### MEMBER\_NAME

Request a reload of local configuration for the given Patroni member(s).

Multiple members can be specified. If no members are specified, all of them are considered.

### --group

Request a reload of members of the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

### -r / --role

Select members that have the given role.

Role can be one of:

- **leader**: the leader of either a regular Patroni cluster or a standby Patroni cluster; or
- **primary**: the leader of a regular Patroni cluster; or
- **standby-leader**: the leader of a standby Patroni cluster; or
- **replica**: a replica of a Patroni cluster; or
- **standby**: same as **replica**; or
- **any**: any role. Same as omitting this parameter.

### --force

Flag to skip confirmation prompts when requesting a reload of the local configuration.

Useful for scripts.

## Examples

Request a reload of the local configuration of all members of the Patroni cluster:

```
$ patronictl -c postgres0.yml reload batman --force
+ Cluster: batman (7277694203142172922) -+-----+-----+-----+
| Member      | Host           | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0 | 127.0.0.1:5432 | Leader | running | 5 |           |
| postgresql1 | 127.0.0.1:5433 | Replica | streaming | 5 |           0 |
| postgresql2 | 127.0.0.1:5434 | Replica | streaming | 5 |           0 |
+-----+-----+-----+-----+-----+-----+
Reload request received for member postgresql0 and will be processed within 10 seconds
Reload request received for member postgresql1 and will be processed within 10 seconds
Reload request received for member postgresql2 and will be processed within 10 seconds
```

### 5.2.11 patronictl remove

#### Synopsis

```
remove
  CLUSTER_NAME
  [ --group CITUS_GROUP ]
  [ { -f | --format } { pretty | tsv | json | yaml } ]
```

#### Description

`patronictl remove` removes information of the cluster from the DCS.

It is an interactive action.

#### Warning

This operation will destroy the information of the Patroni cluster from the DCS.

#### Parameters

##### CLUSTER\_NAME

Name of the Patroni cluster.

##### --group

Remove information about the Patroni cluster related with the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

##### -f / --format

How to format the list of members in the output when prompting for confirmation.

Format can be one of:

- `pretty`: prints members as a pretty table; or
- `tsv`: prints members as tabular information, with columns delimited by `\t`; or
- `json`: prints members in JSON format; or
- `yaml`: prints members in YAML format.

The default is `pretty`.

#### Examples

Remove information about Patroni cluster `batman` from the DCS:

```
$ patronictl -c postgres0.yml remove batman
+ Cluster: batman (7277694203142172922) -+-----+-----+-----+
| Member      | Host              | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0 | 127.0.0.1:5432    | Leader | running | 5  |           |
| postgresql1 | 127.0.0.1:5433    | Replica | streaming | 5  | 0         |
```

(continues on next page)

(continued from previous page)

```
| postgresql2 | 127.0.0.1:5434 | Replica | streaming | 5 | 0 |
+-----+-----+-----+-----+-----+
Please confirm the cluster name to remove: batman
You are about to remove all information in DCS for batman, please type: "Yes I am aware
↪": Yes I am aware
This cluster currently is healthy. Please specify the leader name to continue:
↪postgresql0
```

## 5.2.12 patronictl restart

### Synopsis

```
restart
  CLUSTER_NAME
  [ MEMBER_NAME [, ...] ]
  [ --group CITUS_GROUP ]
  [ { -r | --role } { leader | primary | standby-leader | replica | standby | any } ]
  [ --any ]
  [ --pg-version PG_VERSION ]
  [ --pending ]
  [ --timeout TIMEOUT ]
  [ --scheduled TIMESTAMP ]
  [ --force ]
```

### Description

`patronictl restart` requests a restart of the Postgres instance managed by a member of the Patroni cluster.

The restart can be performed immediately or scheduled for later.

### Parameters

#### CLUSTER\_NAME

Name of the Patroni cluster.

#### --group

Restart the Patroni cluster related with the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

#### -r / --role

Choose members that have the given role.

Role can be one of:

- **leader**: the leader of either a regular Patroni cluster or a standby Patroni cluster; or
- **primary**: the leader of a regular Patroni cluster; or
- **standby-leader**: the leader of a standby Patroni cluster; or
- **replica**: a replica of a Patroni cluster; or
- **standby**: same as **replica**; or



- any: any role. Same as omitting this parameter.

**--any**

Restart a single random node among the ones which match the given filters.

**--pg-version**

Select only members which version of the managed Postgres instance is older than the given version.

PG\_VERSION is the Postgres version to be compared.

**--pending**

Select only members which are flagged as Pending restart.

**timeout**

Abort the restart if it takes more than the specified timeout, and fail over to a replica if the issue is on the primary.

TIMEOUT is the amount of seconds to wait before aborting the restart.

**--scheduled**

Schedule a restart to occur at the given timestamp.

TIMESTAMP is the timestamp when the restart should occur. Specify it in unambiguous format, preferably with time zone. You can also use the literal `now` for the restart to be executed immediately.

**--force**

Flag to skip confirmation prompts when requesting the restart operations.

Useful for scripts.

**Examples**

Restart all members of the cluster immediately:

```
$ patronictl -c postgres0.yml restart batman --force
+ Cluster: batman (7277694203142172922) -+-----+-----+-----+
| Member      | Host          | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0 | 127.0.0.1:5432 | Leader | running | 6 |           |
| postgresql1 | 127.0.0.1:5433 | Replica | streaming | 6 |           0 |
| postgresql2 | 127.0.0.1:5434 | Replica | streaming | 6 |           0 |
+-----+-----+-----+-----+-----+-----+
Success: restart on member postgresql0
Success: restart on member postgresql1
Success: restart on member postgresql2
```

Restart a random member of the cluster immediately:

```
$ patronictl -c postgres0.yml restart batman --any --force
+ Cluster: batman (7277694203142172922) -+-----+-----+-----+
| Member      | Host          | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0 | 127.0.0.1:5432 | Leader | running | 6 |           |
| postgresql1 | 127.0.0.1:5433 | Replica | streaming | 6 |           0 |
| postgresql2 | 127.0.0.1:5434 | Replica | streaming | 6 |           0 |
+-----+-----+-----+-----+-----+-----+
Success: restart on member postgresql1
```

Schedule a restart to occur at 2023-09-13T18:00-03:00:

```
$ patronictl -c postgres0.yml restart batman --scheduled 2023-09-13T18:00-03:00 --force
+ Cluster: batman (7277694203142172922) -+-----+-----+-----+
| Member      | Host              | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0 | 127.0.0.1:5432    | Leader | running | 6  |           |
| postgresql1 | 127.0.0.1:5433    | Replica | streaming | 6  | 0         |
| postgresql2 | 127.0.0.1:5434    | Replica | streaming | 6  | 0         |
+-----+-----+-----+-----+-----+-----+
Success: restart scheduled on member postgresql0
Success: restart scheduled on member postgresql1
Success: restart scheduled on member postgresql2
```

### 5.2.13 patronictl resume

#### Synopsis

```
resume
[ CLUSTER_NAME ]
[ --group CITUS_GROUP ]
[ --wait ]
```

#### Description

`patronictl resume` takes the Patroni cluster out of maintenance mode and re-enables automatic failover.

#### Parameters

##### CLUSTER\_NAME

Name of the Patroni cluster.

If not given, `patronictl` will attempt to fetch that from the `scope` configuration, if it exists.

##### --group

Resume the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

If not given, `patronictl` will attempt to fetch that from the `citus.group` configuration, if it exists.

##### --wait

Wait until all Patroni members are unpaused before returning control to the caller.

## Examples

Put the cluster out of maintenance mode:

```
$ patronictl -c postgres0.yml resume batman --wait
'resume' request sent, waiting until it is recognized by all nodes
Success: cluster management is resumed
```

### 5.2.14 patronictl show-config

#### Synopsis

```
show-config
[ CLUSTER_NAME ]
[ --group CITUS_GROUP ]
```

#### Description

`patronictl show-config` shows the dynamic configuration of the cluster that is stored in the DCS.

#### Parameters

##### CLUSTER\_NAME

Name of the Patroni cluster.

If not given, `patronictl` will attempt to fetch that from the `scope` configuration, if it exists.

##### --group

Show dynamic configuration of the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

If not given, `patronictl` will attempt to fetch that from the `citus.group` configuration, if it exists.

## Examples

Show dynamic configuration of cluster batman:

```
$ patronictl -c postgres0.yml show-config batman
loop_wait: 10
postgresql:
  parameters:
    max_connections: 250
  pg_hba:
    - host replication replicator 127.0.0.1/32 md5
    - host all all 0.0.0.0/0 md5
  use_pg_rewind: true
retry_timeout: 10
ttl: 30
```

## 5.2.15 patronictl switchover

### Synopsis

```
switchover
[ CLUSTER_NAME ]
[ --group CITUS_GROUP ]
[ { --leader | --primary } LEADER_NAME ]
--candidate CANDIDATE_NAME
[ --force ]
```

### Description

`patronictl switchover` performs a switchover in the cluster.

It is designed to be used when the cluster is healthy, e.g.:

- There is a leader;
- There are synchronous standbys available in a synchronous cluster.

#### Note

If your cluster is unhealthy you might be interested in `patronictl failover` instead.

### Parameters

#### **CLUSTER\_NAME**

Name of the Patroni cluster.

If not given, `patronictl` will attempt to fetch that from the `scope` configuration, if it exists.

#### **--group**

Perform a switchover in the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

#### **--leader / --primary**

Indicate who is the leader to be demoted at switchover time.

LEADER\_NAME should match the name of the current leader in the cluster.

#### **--candidate**

The node to be promoted on switchover, and take the primary role.

CANDIDATE\_NAME is the name of the node to be promoted.

#### **--scheduled**

Schedule a switchover to occur at the given timestamp.

TIMESTAMP is the timestamp when the switchover should occur. Specify it in unambiguous format, preferably with time zone. You can also use the literal `now` for the switchover to be executed immediately.

#### **--force**

Flag to skip confirmation prompts when performing the switchover.

Useful for scripts.

## Examples

Switch over with node postgresql2:

```
$ patronictl -c postgres0.yml switchover batman --leader postgresql0 --candidate_
↪ postgresql2 --force
Current cluster topology
+ Cluster: batman (7277694203142172922) -+-----+-----+-----+
| Member      | Host              | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0 | 127.0.0.1:5432    | Leader | running | 6 |           |
| postgresql1 | 127.0.0.1:5433    | Replica | streaming | 6 |           0 |
| postgresql2 | 127.0.0.1:5434    | Replica | streaming | 6 |           0 |
+-----+-----+-----+-----+-----+-----+
2023-09-13 14:15:23.07497 Successfully switched over to "postgresql2"
+ Cluster: batman (7277694203142172922) -+-----+-----+-----+
| Member      | Host              | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0 | 127.0.0.1:5432    | Replica | stopped |   | unknown   |
| postgresql1 | 127.0.0.1:5433    | Replica | running | 6 |           0 |
| postgresql2 | 127.0.0.1:5434    | Leader | running | 6 |           |
+-----+-----+-----+-----+-----+-----+
```

Schedule a switchover between postgresql0 and postgresql2 to occur at 2023-09-13T18:00:00-03:00:

```
$ patronictl -c postgres0.yml switchover batman --leader postgresql0 --candidate_
↪ postgresql2 --scheduled 2023-09-13T18:00-03:00 --force
Current cluster topology
+ Cluster: batman (7277694203142172922) -+-----+-----+-----+
| Member      | Host              | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0 | 127.0.0.1:5432    | Leader | running | 8 |           |
| postgresql1 | 127.0.0.1:5433    | Replica | streaming | 8 |           0 |
| postgresql2 | 127.0.0.1:5434    | Replica | streaming | 8 |           0 |
+-----+-----+-----+-----+-----+-----+
2023-09-13 14:18:11.20661 Switchover scheduled
+ Cluster: batman (7277694203142172922) -+-----+-----+-----+
| Member      | Host              | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0 | 127.0.0.1:5432    | Leader | running | 8 |           |
| postgresql1 | 127.0.0.1:5433    | Replica | streaming | 8 |           0 |
| postgresql2 | 127.0.0.1:5434    | Replica | streaming | 8 |           0 |
+-----+-----+-----+-----+-----+-----+
Switchover scheduled at: 2023-09-13T18:00:00-03:00
                        from: postgresql0
                        to: postgresql2
```

## 5.2.16 patronictl topology

### Synopsis

```
topology
[ CLUSTER_NAME [, ... ] ]
[ --group CITUS_GROUP ]
[ { -W | { -w | --watch } TIME } ]
```

### Description

`patronictl topology` shows information about the Patroni cluster and its members with a tree view approach.

The following information is included in the output:

#### Cluster

Name of the Patroni cluster.

##### Note

Shown in the table header.

#### System identifier

Postgres system identifier.

##### Note

Shown in the table header.

#### Member

Name of the Patroni member.

##### Note

Information in this column is shown as a tree view of members in terms of replication connections.

#### Host

Host where the member is located.

#### Role

Current role of the member.

Can be one among:

- **Leader:** the current leader of a regular Patroni cluster; or
- **Standby Leader:** the current leader of a Patroni standby cluster; or
- **Sync Standby:** a synchronous standby of a Patroni cluster with synchronous mode enabled; or
- **Replica:** a regular standby of a Patroni cluster.

#### State

Current state of Postgres in the Patroni member.

Some examples among the possible states:

- **running**: if Postgres is currently up and running;
- **streaming**: if a replica and Postgres is currently streaming WALs from the primary node;
- **in archive recovery**: if a replica and Postgres is currently fetching WALs from the archive;
- **stopped**: if Postgres had been shut down;
- **crashed**: if Postgres has crashed.

#### TL

Current Postgres timeline in the Patroni member.

#### Lag in MB

Amount worth of replication lag in megabytes between the Patroni member and its upstream.

Besides that, the following information may be included in the output:

#### Group

Citus group ID.

#### Note

Shown in the table header.

Only shown if a Citus cluster.

#### Pending restart

\* indicates the node needs a restart for some Postgres configuration to take effect. An empty value indicates the node does not require a restart.

#### Note

Shown as a member attribute.

Shown if node requires a restart.

#### Scheduled restart

Timestamp at which a restart has been scheduled for the Postgres instance managed by the Patroni member. An empty value indicates there is no scheduled restart for the member.

#### Note

Shown as a member attribute.

Shown if node has a scheduled restart.

#### Tags

Contains tags set for the Patroni member. An empty value indicates that either no tags have been configured, or that they have been configured with default values.

#### Note

Shown as a member attribute.

Shown if node has any custom tags, or any default tags with non-default values.

### Scheduled switchover

Timestamp at which a switchover has been scheduled for the Patroni cluster, if any.

#### Note

Shown in the table footer.

Only shown if there is a scheduled switchover.

### Maintenance mode

If the cluster monitoring is currently paused.

#### Note

Shown in the table footer.

Only shown if the cluster is paused.

## Parameters

### CLUSTER\_NAME

Name of the Patroni cluster.

If not given, `patronictl` will attempt to fetch that from the `scope` configuration, if it exists.

### --group

Show information about members from the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

### -W

Automatically refresh information every 2 seconds.

### -w / --watch

Automatically refresh information at the specified interval.

TIME is the interval between refreshes, in seconds.

## Examples

Show topology of the cluster `batman` – `postgresql1` and `postgresql2` are replicating from `postgresql0`:

```
$ patronictl -c postgres0.yml topology batman
+ Cluster: batman (7277694203142172922) ---+-----+-----+-----+
| Member          | Host              | Role   | State   | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| postgresql0     | 127.0.0.1:5432    | Leader | running | 8  |           |
| + postgresql1   | 127.0.0.1:5433    | Replica | streaming | 8  | 0         |
| + postgresql2   | 127.0.0.1:5434    | Replica | streaming | 8  | 0         |
+-----+-----+-----+-----+-----+-----+
```



## 5.2.17 patronictl version

### Synopsis

```
version
[ CLUSTER_NAME [, ... ] ]
[ MEMBER_NAME [, ... ] ]
[ --group CITUS_GROUP ]
```

### Description

`patronictl version` gets the version of `patronictl` application. Besides that it may also include version information about Patroni clusters and their members.

### Parameters

#### CLUSTER\_NAME

Name of the Patroni cluster.

#### MEMBER\_NAME

Name of the member of the Patroni cluster.

#### --group

Consider a Patroni cluster with the given Citus group.

CITUS\_GROUP is the ID of the Citus group.

### Examples

Get version of `patronictl` only:

```
$ patronictl -c postgres0.yml version
patronictl version 4.0.0
```

Get version of `patronictl` and of all members of cluster `batman`:

```
$ patronictl -c postgres0.yml version batman
patronictl version 4.0.0

postgresql0: Patroni 4.0.0 PostgreSQL 16.4
postgresql1: Patroni 4.0.0 PostgreSQL 16.4
postgresql2: Patroni 4.0.0 PostgreSQL 16.4
```

Get version of `patronictl` and of members `postgresql1` and `postgresql2` of cluster `batman`:

```
$ patronictl -c postgres0.yml version batman postgresql1 postgresql2
patronictl version 4.0.0

postgresql1: Patroni 4.0.0 PostgreSQL 16.4
postgresql2: Patroni 4.0.0 PostgreSQL 16.4
```



## REPLICA IMAGING AND BOOTSTRAP

Patroni allows customizing creation of a new replica. It also supports defining what happens when the new empty cluster is being bootstrapped. The distinction between two is well defined: Patroni creates replicas only if the `initialize` key is present in DCS for the cluster. If there is no `initialize` key - Patroni calls bootstrap exclusively on the first node that takes the `initialize` key lock.

### 6.1 Bootstrap

PostgreSQL provides `initdb` command to initialize a new cluster and Patroni calls it by default. In certain cases, particularly when creating a new cluster as a copy of an existing one, it is necessary to replace a built-in method with custom actions. Patroni supports executing user-defined scripts to bootstrap new clusters, supplying some required arguments to them, i.e. the name of the cluster and the path to the data directory. This is configured in the `bootstrap` section of the Patroni configuration. For example:

```
bootstrap:
  method: <custom_bootstrap_method_name>
  <custom_bootstrap_method_name>:
    command: <path_to_custom_bootstrap_script> [param1 [, ...]]
    keep_existing_recovery_conf: False
    no_params: False
    recovery_conf:
      recovery_target_action: promote
      recovery_target_timeline: latest
      restore_command: <method_specific_restore_command>
```

Each bootstrap method must define at least a `name` and a `command`. A special `initdb` method is available to trigger the default behavior, in which case `method` parameter can be omitted altogether. The `command` can be specified using either an absolute path, or the one relative to the `patroni` command location. In addition to the fixed parameters defined in the configuration files, Patroni supplies two cluster-specific ones:

<b>--scope</b>	Name of the cluster to be bootstrapped
<b>--datadir</b>	Path to the data directory of the cluster instance to be bootstrapped

Passing these two additional flags can be disabled by setting a special `no_params` parameter to `True`.

If the bootstrap script returns `0`, Patroni tries to configure and start the PostgreSQL instance produced by it. If any of the intermediate steps fail, or the script returns a non-zero value, Patroni assumes that the bootstrap has failed, cleans up after itself and releases the `initialize` lock to give another node the opportunity to bootstrap.

If a `recovery_conf` block is defined in the same section as the custom bootstrap method, Patroni will generate a `recovery.conf` before starting the newly bootstrapped instance (or set the recovery settings on Postgres config-

uration if running PostgreSQL  $\geq 12$ ). Typically, such recovery configuration should contain at least one of the `recovery_target_*` parameters, together with the `recovery_target_timeline` set to `promote`.

If `keep_existing_recovery_conf` is defined and set to `True`, Patroni will not remove the existing `recovery.conf` file if it exists (PostgreSQL  $\leq 11$ ). Similarly, in that case Patroni will not remove the existing `recovery.signal` or `standby.signal` if either exists, nor will it override the configured recovery settings (PostgreSQL  $\geq 12$ ). This is useful when bootstrapping from a backup with tools like `pgBackRest` that generate the appropriate recovery configuration for you.

Besides that, any additional key/value pairs informed in the custom bootstrap method configuration will be passed as arguments to `command` in the format `--name=value`. For example:

```
bootstrap:
  method: <custom_bootstrap_method_name>
  <custom_bootstrap_method_name>:
    command: <path_to_custom_bootstrap_script>
    arg1: value1
    arg2: value2
```

Makes the configured `command` to be called additionally with `--arg1=value1 --arg2=value2` command-line arguments.

#### Note

Bootstrap methods are neither chained, nor fallen-back to the default one in case the primary one fails

As an example, you are able to bootstrap a fresh Patroni cluster from a Barman backup with a configuration like this:

```
bootstrap:
  method: barman
  barman:
    keep_existing_recovery_conf: true
    command: patroni_barman --api-url https://barman-host:7480 recover
    barman-server: my_server
    ssh-command: ssh postgres@patroni-host
```

#### Note

`patroni_barman recover` requires that you have both Barman and `pg-backup-api` configured in the Barman host, so it can execute a remote `barman recover` through the backup API. The above example uses a subset of the available parameters. You can get more information running `patroni_barman recover --help`.

## 6.2 Building replicas

Patroni uses tried and proven `pg_basebackup` in order to create new replicas. One downside of it is that it requires a running leader node. Another one is the lack of ‘on-the-fly’ compression for the backup data and no built-in cleanup for outdated backup files. Some people prefer other backup solutions, such as `WAL-E`, `pgBackRest`, `Barman` and others, or simply roll their own scripts. In order to accommodate all those use-cases Patroni supports running custom scripts to clone a new replica. Those are configured in the `postgresql` configuration block:

```

postgresql:
  create_replica_methods:
    - <method name>
  <method name>:
    command: <command name>
    keep_data: True
    no_params: True
    no_leader: 1

```

example: wal\_e

```

postgresql:
  create_replica_methods:
    - wal_e
    - basebackup
  wal_e:
    command: patroni_wale_restore
    no_leader: 1
    envdir: {{WALE_ENV_DIR}}
    use_iam: 1
  basebackup:
    max-rate: '100M'

```

example: pgbackrest

```

postgresql:
  create_replica_methods:
    - pgbackrest
    - basebackup
  pgbackrest:
    command: /usr/bin/pgbackrest --stanza=<scope> --delta restore
    keep_data: True
    no_params: True
  basebackup:
    max-rate: '100M'

```

example: Barman

```

postgresql:
  create_replica_methods:
    - barman
    - basebackup
  barman:
    command: patroni_barman --api-url https://barman-host:7480 recover
    barman-server: my_server
    ssh-command: ssh postgres@patroni-host
  basebackup:
    max-rate: '100M'

```

#### Note

`patroni_barman recover` requires that you have both Barman and `pg-backup-api` configured in the Barman host, so it can execute a remote `barman recover` through the backup API. The above example uses a subset of the

available parameters. You can get more information running `patroni_barman recover --help`.

The `create_replica_methods` defines available replica creation methods and the order of executing them. Patroni will stop on the first one that returns 0. Each method should define a separate section in the configuration file, listing the command to execute and any custom parameters that should be passed to that command. All parameters will be passed in a `--name=value` format. Besides user-defined parameters, Patroni supplies a couple of cluster-specific ones:

<b>--scope</b>	Which cluster this replica belongs to
<b>--datadir</b>	Path to the data directory of the replica
<b>--role</b>	Always 'replica'
<b>--connstring</b>	Connection string to connect to the cluster member to clone from (primary or other replica). The user in the connection string can execute SQL and replication protocol commands.

A special `no_leader` parameter, if defined, allows Patroni to call the replica creation method even if there is no running leader or replicas. In that case, an empty string will be passed in a connection string. This is useful for restoring the formerly running cluster from the binary backup.

A special `keep_data` parameter, if defined, will instruct Patroni to not clean PGDATA folder before calling restore.

A special `no_params` parameter, if defined, restricts passing parameters to custom command.

A `basebackup` method is a special case: it will be used if `create_replica_methods` is empty, although it is possible to list it explicitly among the `create_replica_methods` methods. This method initializes a new replica with the `pg_basebackup`, the base backup is taken from the leader unless there are replicas with `clonefrom` tag, in which case one of such replicas will be used as the origin for `pg_basebackup`. It works without any configuration; however, it is possible to specify a `basebackup` configuration section. Same rules as with the other method configuration apply, namely, only long (with `-`) options should be specified there. Not all parameters make sense, if you override a connection string or provide an option to create tar-ed or compressed base backups, patroni won't be able to make a replica out of it. There is no validation performed on the names or values of the parameters passed to the `basebackup` section. Also note that in case symlinks are used for the WAL folder it is up to the user to specify the correct `--waldir` path as an option, so that after replica buildup or re-initialization the symlink would persist. This option is supported only since v10 though.

You can specify `basebackup` parameters as either a map (key-value pairs) or a list of elements, where each element could be either a key-value pair or a single key (for options that does not receive any values, for instance, `--verbose`). Consider those 2 examples:

```
postgresql:
  basebackup:
    max-rate: '100M'
    checkpoint: 'fast'
```

and

```
postgresql:
  basebackup:
    - verbose
    - max-rate: '100M'
    - waldir: /pg-wal-mount/external-waldir
```

If all replica creation methods fail, Patroni will try again all methods in order during the next event loop cycle.

## REPLICATION MODES

Patroni uses PostgreSQL streaming replication. For more information about streaming replication, see the [Postgres documentation](#). By default Patroni configures PostgreSQL for asynchronous replication. Choosing your replication schema is dependent on your business considerations. Investigate both async and sync replication, as well as other HA solutions, to determine which solution is best for you.

### 7.1 Asynchronous mode durability

In asynchronous mode the cluster is allowed to lose some committed transactions to ensure availability. When the primary server fails or becomes unavailable for any other reason Patroni will automatically promote a sufficiently healthy standby to primary. Any transactions that have not been replicated to that standby remain in a “forked timeline” on the primary, and are effectively unrecoverable<sup>1</sup>.

The amount of transactions that can be lost is controlled via `maximum_lag_on_failover` parameter. Because the primary transaction log position is not sampled in real time, in reality the amount of lost data on failover is worst case bounded by `maximum_lag_on_failover` bytes of transaction log plus the amount that is written in the last `t1` seconds (`loop_wait/2` seconds in the average case). However typical steady state replication delay is well under a second.

By default, when running leader elections, Patroni does not take into account the current timeline of replicas, what in some cases could be undesirable behavior. You can prevent the node not having the same timeline as a former primary become the new leader by changing the value of `check_timeline` parameter to `true`.

### 7.2 PostgreSQL synchronous replication

You can use Postgres’s [synchronous replication](#) with Patroni. Synchronous replication ensures consistency across a cluster by confirming that writes are written to a secondary before returning to the connecting client with a success. The cost of synchronous replication: increased latency and reduced throughput on writes. This throughput will be entirely based on network performance.

In hosted datacenter environments (like AWS, Rackspace, or any network you do not control), synchronous replication significantly increases the variability of write performance. If followers become inaccessible from the leader, the leader effectively becomes read-only.

To enable a simple synchronous replication test, add the following lines to the `parameters` section of your YAML configuration files:

---

<sup>1</sup> The data is still there, but recovering it requires a manual recovery effort by data recovery specialists. When Patroni is allowed to rewind with `use_pg_rewind` the forked timeline will be automatically erased to rejoin the failed primary with the cluster. However, for `use_pg_rewind` to function properly, either the cluster must be initialized with `data page checksums` (`--data-checksums` option for `initdb`) and/or `wal_log_hints` must be set to on.

```
synchronous_commit: "on"  
synchronous_standby_names: "*/"
```

When using PostgreSQL synchronous replication, use at least three Postgres data nodes to ensure write availability if one host fails.

Using PostgreSQL synchronous replication does not guarantee zero lost transactions under all circumstances. When the primary and the secondary that is currently acting as a synchronous replica fail simultaneously a third node that might not contain all transactions will be promoted.

## 7.3 Synchronous mode

For use cases where losing committed transactions is not permissible you can turn on Patroni's `synchronous_mode`. When `synchronous_mode` is turned on Patroni will not promote a standby unless it is certain that the standby contains all transactions that may have returned a successful commit status to client<sup>2</sup>. This means that the system may be unavailable for writes even though some servers are available. System administrators can still use manual failover commands to promote a standby even if it results in transaction loss.

Turning on `synchronous_mode` does not guarantee multi node durability of commits under all circumstances. When no suitable standby is available, primary server will still accept writes, but does not guarantee their replication. When the primary fails in this mode no standby will be promoted. When the host that used to be the primary comes back it will get promoted automatically, unless system administrator performed a manual failover. This behavior makes synchronous mode usable with 2 node clusters.

When `synchronous_mode` is on and a standby crashes, commits will block until next iteration of Patroni runs and switches the primary to standalone mode (worst case delay for writes `t_t_l` seconds, average case `loop_wait/2` seconds). Manually shutting down or restarting a standby will not cause a commit service interruption. Standby will signal the primary to release itself from synchronous standby duties before PostgreSQL shutdown is initiated.

When it is absolutely necessary to guarantee that each write is stored durably on at least two nodes, enable `synchronous_mode_strict` in addition to the `synchronous_mode`. This parameter prevents Patroni from switching off the synchronous replication on the primary when no synchronous standby candidates are available. As a downside, the primary is not be available for writes (unless the Postgres transaction explicitly turns off `synchronous_mode`), blocking all client write requests until at least one synchronous replica comes up.

You can ensure that a standby never becomes the synchronous standby by setting `nosync` tag to true. This is recommended to set for standbys that are behind slow network connections and would cause performance degradation when becoming a synchronous standby. Setting tag `nostream` to true will also have the same effect.

Synchronous mode can be switched on and off using `patronictl edit-config` command or via Patroni REST interface. See [dynamic configuration](#) for instructions.

Note: Because of the way synchronous replication is implemented in PostgreSQL it is still possible to lose transactions even when using `synchronous_mode_strict`. If the PostgreSQL backend is cancelled while waiting to acknowledge replication (as a result of packet cancellation due to client timeout or backend failure) transaction changes become visible for other backends. Such changes are not yet replicated and may be lost in case of standby promotion.

---

<sup>2</sup> Clients can change the behavior per transaction using PostgreSQL's `synchronous_commit` setting. Transactions with `synchronous_commit` values of `off` and `local` may be lost on fail over, but will not be blocked by replication delays.



## 7.4 Synchronous Replication Factor

The parameter `synchronous_node_count` is used by Patroni to manage the number of synchronous standby databases. It is set to 1 by default. It has no effect when `synchronous_mode` is set to `off`. When enabled, Patroni manages the precise number of synchronous standby databases based on parameter `synchronous_node_count` and adjusts the state in DCS & `synchronous_standby_names` in PostgreSQL as members join and leave. If the parameter is set to a value higher than the number of eligible nodes it will be automatically reduced by Patroni.

## 7.5 Maximum lag on synchronous node

By default Patroni sticks to nodes that are declared as `synchronous`, according to the `pg_stat_replication` view, even when there are other nodes ahead of it. This is done to minimize the number of changes of `synchronous_standby_names`. To change this behavior one may use `maximum_lag_on_syncnode` parameter. It controls how much lag the replica can have to still be considered as “synchronous”.

Patroni utilizes the max replica LSN if there is more than one standby, otherwise it will use leader’s current wal LSN. The default is `-1`, and Patroni will not take action to swap a synchronous unhealthy standby when the value is set to `0` or less. Please set the value high enough so that Patroni won’t swap synchronous standbys frequently during high transaction volume.

## 7.6 Synchronous mode implementation

When in synchronous mode Patroni maintains synchronization state in the DCS (`/sync` key), containing the latest primary and current synchronous standby databases. This state is updated with strict ordering constraints to ensure the following invariants:

- A node must be marked as the latest leader whenever it can accept write transactions. Patroni crashing or PostgreSQL not shutting down can cause violations of this invariant.
- A node must be set as the synchronous standby in PostgreSQL as long as it is published as the synchronous standby in the `/sync` key in DCS..
- A node that is not the leader or current synchronous standby is not allowed to promote itself automatically.

Patroni will only assign one or more synchronous standby nodes based on `synchronous_node_count` parameter to `synchronous_standby_names`.

On each HA loop iteration Patroni re-evaluates synchronous standby nodes choice. If the current list of synchronous standby nodes are connected and has not requested its synchronous status to be removed it remains picked. Otherwise the cluster members available for sync that are furthest ahead in replication are picked.

### 7.6.1 Example:

`/config` key in DCS

```
synchronous_mode: on
synchronous_node_count: 2
...
```

#### /sync key in DCS

```
{
  "leader": "node0",
  "sync_standby": "node1,node2"
}
```

#### postgresql.conf

```
synchronous_standby_names = 'FIRST 2 (node1,node2)'
```

In the above examples only nodes `node1` and `node2` are known to be synchronous and allowed to be automatically promoted if the primary (`node0`) fails.

## 7.7 Quorum commit mode

Starting from PostgreSQL v10 Patroni supports quorum-based synchronous replication.

In this mode, Patroni maintains synchronization state in the DCS, containing the latest known primary, the number of nodes required for quorum, and the nodes currently eligible to vote on quorum. In steady state, the nodes voting on quorum are the leader and all synchronous standbys. This state is updated with strict ordering constraints, with regards to node promotion and `synchronous_standby_names`, to ensure that at all times any subset of voters that can achieve quorum includes at least one node with the latest successful commit.

On each iteration of HA loop, Patroni re-evaluates synchronous standby choices and quorum, based on node availability and requested cluster configuration. In PostgreSQL versions above 9.6 all eligible nodes are added as synchronous standbys as soon as their replication catches up to leader.

Quorum commit helps to reduce worst case latencies, even during normal operation, as a higher latency of replicating to one standby can be compensated by other standbys.

The quorum-based synchronous mode could be enabled by setting `synchronous_mode` to `quorum` using `patronictl edit-config` command or via Patroni REST interface. See [dynamic configuration](#) for instructions.

Other parameters, like `synchronous_node_count`, `maximum_lag_on_syncnode`, and `synchronous_mode_strict` continue to work the same way as with `synchronous_mode=on`.

### 7.7.1 Example:

#### /config key in DCS

```
synchronous_mode: quorum
synchronous_node_count: 2
...
```

### /sync key in DCS

```
{  
  "leader": "node0",  
  "sync_standby": "node1,node2,node3",  
  "quorum": 1  
}
```

### postgresql.conf

```
synchronous_standby_names = 'ANY 2 (node1,node2,node3)'
```

If the primary (**node0**) failed, in the above example two of the **node1**, **node2**, **node3** will have the latest transaction received, but we don't know which ones. To figure out whether the node **node1** has received the latest transaction, we need to compare its LSN with the LSN on **at least** one node (**quorum=1** in the **/sync** key) among **node2** and **node3**. If **node1** isn't behind of at least one of them, we can guarantee that there will be no user visible data loss if **node1** is promoted.



## STANDBY CLUSTER

Patroni also support running cascading replication to a remote datacenter (region) using a feature that is called “standby cluster”. This type of clusters has:

- “standby leader”, that behaves pretty much like a regular cluster leader, except it replicates from a remote node.
- cascade replicas, that are replicating from standby leader.

Standby leader holds and updates a leader lock in DCS. If the leader lock expires, cascade replicas will perform an election to choose another leader from the standbys.

There is no further relationship between the standby cluster and the primary cluster it replicates from, in particular, they must not share the same DCS scope if they use the same DCS. They do not know anything else from each other apart from replication information. Also, the standby cluster is not being displayed in *patronictl list* or *patronictl topology* output on the primary cluster.

For the sake of flexibility, you can specify methods of creating a replica and recovery WAL records when a cluster is in the “standby mode” by providing *create\_replica\_methods* key in *standby\_cluster* section. It is distinct from creating replicas, when cluster is detached and functions as a normal cluster, which is controlled by *create\_replica\_methods* in *postgresql* section. Both “standby” and “normal” *create\_replica\_methods* reference keys in *postgresql* section.

To configure such cluster you need to specify the section *standby\_cluster* in a patroni configuration:

```
bootstrap:
  dcs:
    standby_cluster:
      host: 1.2.3.4
      port: 5432
      primary_slot_name: patroni
      create_replica_methods:
        - basebackup
```

Note, that these options will be applied only once during cluster bootstrap, and the only way to change them afterwards is through DCS.

Patroni expects to find *postgresql.conf* or *postgresql.conf.backup* in PGDATA of the remote primary and will not start if it does not find it after a basebackup. If the remote primary keeps its *postgresql.conf* elsewhere, it is your responsibility to copy it to PGDATA.

If you use replication slots on the standby cluster, you must also create the corresponding replication slot on the primary cluster. It will not be done automatically by the standby cluster implementation. You can use Patroni’s permanent replication slots feature on the primary cluster to maintain a replication slot with the same name as *primary\_slot\_name*, or its default value if *primary\_slot\_name* is not provided.

In case the remote site doesn’t provide a single endpoint that connects to a primary, one could list all hosts of the source cluster in the *standby\_cluster.host* section. When *standby\_cluster.host* contains multiple hosts separated by commas, Patroni will:

- add `target_session_attrs=read-write` to the `primary_conninfo` on the standby leader node.
- use `target_session_attrs=read-write` when trying to determine whether we need to run `pg_rewind` or when executing `pg_rewind` on all nodes of the standby cluster.
- It is important to note that for `pg_rewind` to operate successfully, either the cluster must be initialized with `data page checksums` (`--data-checksums` option for `initdb`) and/or `wal_log_hints` must be set to on. Otherwise, `pg_rewind` will not function properly.

There is also a possibility to replicate the standby cluster from another standby cluster or from a standby member of the primary cluster: for that, you need to define a single host in the `standby_cluster.host` section. However, you need to beware that in this case `pg_rewind` will fail to execute on the standby cluster.

## WATCHDOG SUPPORT

Having multiple PostgreSQL servers running as primary can result in transactions lost due to diverging timelines. This situation is also called a split-brain problem. To avoid split-brain Patroni needs to ensure PostgreSQL will not accept any transaction commits after leader key expires in the DCS. Under normal circumstances Patroni will try to achieve this by stopping PostgreSQL when leader lock update fails for any reason. However, this may fail to happen due to various reasons:

- Patroni has crashed due to a bug, out-of-memory condition or by being accidentally killed by a system administrator.
- Shutting down PostgreSQL is too slow.
- Patroni does not get to run due to high load on the system, the VM being paused by the hypervisor, or other infrastructure issues.

To guarantee correct behavior under these conditions Patroni supports watchdog devices. Watchdog devices are software or hardware mechanisms that will reset the whole system when they do not get a keepalive heartbeat within a specified timeframe. This adds an additional layer of fail safe in case usual Patroni split-brain protection mechanisms fail.

Patroni will try to activate the watchdog before promoting PostgreSQL to primary. If watchdog activation fails and watchdog mode is required then the node will refuse to become leader. When deciding to participate in leader election Patroni will also check that watchdog configuration will allow it to become leader at all. After demoting PostgreSQL (for example due to a manual failover) Patroni will disable the watchdog again. Watchdog will also be disabled while Patroni is in paused state.

By default Patroni will set up the watchdog to expire 5 seconds before TTL expires. With the default setup of `loop_wait=10` and `ttl=30` this gives HA loop at least 15 seconds (`ttl - safety_margin - loop_wait`) to complete before the system gets forcefully reset. By default accessing DCS is configured to time out after 10 seconds. This means that when DCS is unavailable, for example due to network issues, Patroni and PostgreSQL will have at least 5 seconds (`ttl - safety_margin - loop_wait - retry_timeout`) to come to a state where all client connections are terminated.

Safety margin is the amount of time that Patroni reserves for time between leader key update and watchdog keepalive. Patroni will try to send a keepalive immediately after confirmation of leader key update. If Patroni process is suspended for extended amount of time at exactly the right moment the keepalive may be delayed for more than the safety margin without triggering the watchdog. This results in a window of time where watchdog will not trigger before leader key expiration, invalidating the guarantee. To be absolutely sure that watchdog will trigger under all circumstances set up the watchdog to expire after half of TTL by setting `safety_margin` to `-1` to set watchdog timeout to `t1 // 2`. If you need this guarantee you probably should increase `t1` and/or reduce `loop_wait` and `retry_timeout`.

Currently watchdogs are only supported using Linux watchdog device interface.

## 9.1 Setting up software watchdog on Linux

Default Patroni configuration will try to use `/dev/watchdog` on Linux if it is accessible to Patroni. For most use cases using software watchdog built into the Linux kernel is secure enough.

To enable software watchdog issue the following commands as root before starting Patroni:

```
modprobe softdog
# Replace postgres with the user you will be running patroni under
chown postgres /dev/watchdog
```

For testing it may be helpful to disable rebooting by adding `soft_noboot=1` to the `modprobe` command line. In this case the watchdog will just log a line in kernel ring buffer, visible via `dmesg`.

Patroni will log information about the watchdog when it is successfully enabled.



## **PAUSE/RESUME MODE FOR THE CLUSTER**

### **10.1 The goal**

Under certain circumstances Patroni needs to temporarily step down from managing the cluster, while still retaining the cluster state in DCS. Possible use cases are uncommon activities on the cluster, such as major version upgrades or corruption recovery. During those activities nodes are often started and stopped for reasons unknown to Patroni, some nodes can be even temporarily promoted, violating the assumption of running only one primary. Therefore, Patroni needs to be able to “detach” from the running cluster, implementing an equivalent of the maintenance mode in Pacemaker.

### **10.2 The implementation**

When Patroni runs in a paused mode, it does not change the state of PostgreSQL, except for the following cases:

- For each node, the member key in DCS is updated with the current information about the cluster. This causes Patroni to run read-only queries on a member node if the member is running.
- For the Postgres primary with the leader lock Patroni updates the lock. If the node with the leader lock stops being the primary (i.e. is demoted manually), Patroni will release the lock instead of promoting the node back.
- Manual unscheduled restart, manual unscheduled failover/switchover and reinitialize are allowed. No scheduled action is allowed. Manual switchover is only allowed if the node to switch over to is specified.
- If ‘parallel’ primaries are detected by Patroni, it emits a warning, but does not demote the primary without the leader lock.
- If there is no leader lock in the cluster, the running primary acquires the lock. If there is more than one primary node, then the first primary to acquire the lock wins. If there are no primary altogether, Patroni does not try to promote any replicas. There is an exception in this rule: if there is no leader lock because the old primary has demoted itself due to the manual promotion, then only the candidate node mentioned in the promotion request may take the leader lock. When the new leader lock is granted (i.e. after promoting a replica manually), Patroni makes sure the replicas that were streaming from the previous leader will switch to the new one.
- When Postgres is stopped, Patroni does not try to start it. When Patroni is stopped, it does not try to stop the Postgres instance it is managing.
- Patroni will not try to remove replication slots that don’t represent the other cluster member or are not listed in the configuration of the permanent slots.

## 10.3 User guide

patronictl supports *pause* and *resume* commands.

One can also issue a PATCH request to the {namespace}/{cluster}/config key with {"pause": true/false/null}

## DCS FAILSAFE MODE

### 11.1 The problem

Patroni is heavily relying on Distributed Configuration Store (DCS) to solve the task of leader elections and detect network partitioning. That is, the node is allowed to run Postgres as the primary only if it can update the leader lock in DCS. In case the update of the leader lock fails, Postgres is immediately demoted and started as read-only. Depending on which DCS is used, the chances of hitting the “problem” differ. For example, with Etcd which is only used for Patroni, chances are close to zero, while with K8s API (backed by Etcd) it could be observed more frequently.

### 11.2 Reasons for the current implementation

The leader lock update failure could be caused by two main reasons:

1. Network partitioning
2. DCS being down

In general, it is impossible to distinguish between these two from a single node, and therefore Patroni assumes the worst case - network partitioning. In the case of a partitioned network, other nodes of the Patroni cluster may successfully grab the leader lock and promote Postgres to primary. In order to avoid a split-brain, the old primary is demoted before the leader lock expires.

### 11.3 DCS Failsafe Mode

We introduce a new special option, the `failsafe_mode`. It could be enabled only via global *dynamic configuration* stored in the DCS `/config` key. If the failsafe mode is enabled and the leader lock update in DCS failed due to reasons different from the version/value/index mismatch, Postgres may continue to run as a primary if it can access all known members of the cluster via Patroni REST API.

## 11.4 Low-level implementation details

- We introduce a new, permanent key in DCS, named `/failsafe`.
- The `/failsafe` key contains all known members of the given Patroni cluster at a given time.
- The current leader maintains the `/failsafe` key.
- The member is allowed to participate in the leader race and become the new leader only if it is present in the `/failsafe` key.
- If the cluster consists of a single node the `/failsafe` key will contain a single member.
- In the case of DCS “outage” the existing primary connects to all members presented in the `/failsafe` key via the `POST /failsafe` REST API and may continue to run as the primary if all replicas acknowledge it.
- If one of the members doesn’t respond, the primary is demoted.
- Replicas are using incoming `POST /failsafe` REST API requests as an indicator that the primary is still alive. This information is cached for `t1` seconds.

## 11.5 F.A.Q.

- Why MUST the current primary see ALL other members? Can’t we rely on quorum here?

This is a great question! The problem is that the view on the quorum might be different from the perspective of DCS and Patroni. While DCS nodes must be evenly distributed across availability zones, there is no such rule for Patroni, and more importantly, there is no mechanism for introducing and enforcing such a rule. If the majority of Patroni nodes ends up in the losing part of the partitioned network (including primary) while minority nodes are in the winning part, the primary must be demoted. Only checking ALL other members allows detecting such a situation.

- What if node/pod gets terminated while DCS is down?

If DCS isn’t accessible, the check “are ALL other cluster members accessible?” is executed every cycle of the heartbeat loop (every `loop_wait` seconds). If pod/node is terminated, the check will fail and Postgres will be demoted to a read-only and will not recover until DCS is restored.

- What if all members of the Patroni cluster are lost while DCS is down?

Patroni could be configured to create the new replica from the backup even when the cluster doesn’t have a leader. But, if the new member isn’t present in the `/failsafe` key, it will not be able to grab the leader lock and promote.

- What will happen if the primary lost access to DCS while replicas didn’t?

The primary will execute the failsafe code and contact all known replicas. These replicas will use this information as an indicator that the primary is alive and will not start the leader race even if the leader lock in DCS has expired.

- How to enable the Failsafe Mode?

Before enabling the `failsafe_mode` please make sure that Patroni version on all members is up-to-date. After that, you can use either the `PATCH /config REST API` or `patronictl edit-config -s failsafe_mode=true`

## USING PATRONI WITH KUBERNETES

Patroni can use Kubernetes objects in order to store the state of the cluster and manage the leader key. That makes it capable of operating Postgres in Kubernetes environment without any consistency store, namely, one doesn't need to run an extra Etcd deployment. There are two different type of Kubernetes objects Patroni can use to store the leader and the configuration keys, they are configured with the *kubernetes.use\_endpoints* or *PATRONI\_KUBERNETES\_USE\_ENDPOINTS* environment variable.

### 12.1 Use Endpoints

Despite the fact that this is the recommended mode, it is turned off by default for compatibility reasons. When it is on, Patroni stores the cluster configuration and the leader key in the *metadata: annotations* fields of the respective *Endpoints* it creates. Changing the leader is safer than when using *ConfigMaps*, since both the annotations, containing the leader information, and the actual addresses pointing to the running leader pod are updated simultaneously in one go.

### 12.2 Use ConfigMaps

In this mode, Patroni will create *ConfigMaps* instead of *Endpoints* and store keys inside meta-data of those *ConfigMaps*. Changing the leader takes at least two updates, one to the leader *ConfigMap* and another to the respective *Endpoint*.

To direct the traffic to the Postgres leader you need to configure the Kubernetes Postgres service to use the label selector with the *role\_label* (configured in patroni configuration).

Note that in some cases, for instance, when running on OpenShift, there is no alternative to using *ConfigMaps*.

### 12.3 Configuration

Patroni Kubernetes *settings* and *environment variables* are described in the general chapters of the documentation.

### 12.3.1 Customize role label

By default, Patroni will set corresponding labels on the pod it runs in based on node's role, such as `role=primary`. The key and value of label can be customized by `kubernetes.role_label`, `kubernetes.leader_label_value`, `kubernetes.follower_label_value` and `kubernetes.standby_leader_label_value`.

Note that if you migrate from default role labels to custom ones, you can reduce downtime by following migration steps:

1. Add a temporary label using original role value for the pod with `kubernetes.tmp_role_label` (like `tmp_role`). Once pods are restarted they will get following labels set by Patroni:

```
labels:
  cluster-name: foo
  role: primary
  tmp_role: primary
```

2. After all pods have been updated, modify the service selector to select the temporary label.

```
selector:
  cluster-name: foo
  tmp_role: primary
```

3. Add your custom role label (e.g., set `kubernetes.leader_label_value=primary`). Once pods are restarted they will get following new labels set by Patroni:

```
labels:
  cluster-name: foo
  role: primary
  tmp_role: primary
```

4. After all pods have been updated again, modify the service selector to use new role value.

```
selector:
  cluster-name: foo
  role: primary
```

5. Finally, remove the temporary label from your configuration and update all pods.

```
labels:
  cluster-name: foo
  role: primary
```

## 12.4 Examples

- The `kubernetes` folder of the Patroni repository contains examples of the Docker image, and the Kubernetes manifest to test Patroni Kubernetes setup. Note that in the current state it will not be able to use PersistentVolumes because of permission issues.
- You can find the full-featured Docker image that can use Persistent Volumes in the [Spilo Project](#).
- There is also a [Helm chart](#) to deploy the Spilo image configured with Patroni running using Kubernetes.
- In order to run your database clusters at scale using Patroni and Spilo, take a look at the [postgres-operator](#) project. It implements the operator pattern to manage Spilo clusters.

## CITUS SUPPORT

Patroni makes it extremely simple to deploy [Multi-Node Citus](#) clusters.

### 13.1 TL;DR

There are only a few simple rules you need to follow:

1. [Citus](#) database extension to PostgreSQL must be available on all nodes. Absolute minimum supported Citus version is 10.0, but, to take all benefits from transparent switchovers and restarts of workers we recommend using at least Citus 11.2.
2. Cluster name (`scope`) must be the same for all Citus nodes!
3. Superuser credentials must be the same on coordinator and all worker nodes, and `pg_hba.conf` should allow superuser access between all nodes.
4. [REST API](#) access should be allowed from worker nodes to the coordinator. E.g., credentials should be the same and if configured, client certificates from worker nodes must be accepted by the coordinator.
5. Add the following section to the `patroni.yaml`:

```
citus:  
  group: X # 0 for coordinator and 1, 2, 3, etc for workers  
  database: citus # must be the same on all nodes
```

After that you just need to start Patroni and it will handle the rest:

0. Patroni will set `bootstrap.dcs.synchronous_mode` to [quorum](#) if it is not explicitly set to any other value.
1. `citus` extension will be automatically added to `shared_preload_libraries`.
2. If `max_prepared_transactions` isn't explicitly set in the global [dynamic configuration](#) Patroni will automatically set it to `2*max_connections`.
3. The `citus.local_hostname` GUC value will be adjusted from `localhost` to the value that Patroni is using in order to connect to the local PostgreSQL instance. The value sometimes should be different from the `localhost` because PostgreSQL might be not listening on it.
4. The `citus.database` will be automatically created followed by `CREATE EXTENSION citus`.
5. Current superuser [credentials](#) will be added to the `pg_dist_authinfo` table to allow cross-node communication. Don't forget to update them if later you decide to change superuser username/password/sslcert/sslkey!
6. The coordinator primary node will automatically discover worker primary nodes and add them to the `pg_dist_node` table using the `citus_add_node()` function.

- Patroni will also maintain `pg_dist_node` in case failover/switchover on the coordinator or worker clusters occurs.

## 13.2 patronictl

Coordinator and worker clusters are physically different PostgreSQL/Patroni clusters that are just logically grouped together using the `Citus` database extension to PostgreSQL. Therefore in most cases it is not possible to manage them as a single entity.

It results in two major differences in `patronictl` behaviour when `patroni.yaml` has the `citus` section comparing with the usual:

- The `list` and the `topology` by default output all members of the Citus formation (coordinators and workers). The new column `Group` indicates which Citus group they belong to.
- For all `patronictl` commands the new option is introduced, named `--group`. For some commands the default value for the group might be taken from the `patroni.yaml`. For example, `patronictl pause` will enable the maintenance mode by default for the group that is set in the `citus` section, but for example for `patronictl switchover` or `patronictl remove` the group must be explicitly specified.

An example of `patronictl list` output for the Citus cluster:

```
postgres@coord1:~$ patronictl list demo
+ Citus cluster: demo -----+-----+-----+-----+-----+
| Group | Member | Host       | Role           | State  | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| 0     | coord1 | 172.27.0.10 | Replica        | running | 1 | 0          |
| 0     | coord2 | 172.27.0.6  | Quorum Standby | running | 1 | 0          |
| 0     | coord3 | 172.27.0.4  | Leader         | running | 1 |            |
| 1     | work1-1 | 172.27.0.8  | Quorum Standby | running | 1 | 0          |
| 1     | work1-2 | 172.27.0.2  | Leader         | running | 1 |            |
| 2     | work2-1 | 172.27.0.5  | Quorum Standby | running | 1 | 0          |
| 2     | work2-2 | 172.27.0.7  | Leader         | running | 1 |            |
+-----+-----+-----+-----+-----+-----+
```

If we add the `--group` option, the output will change to:

```
postgres@coord1:~$ patronictl list demo --group 0
+ Citus cluster: demo (group: 0, 7179854923829112860) -----+
| Member | Host       | Role           | State  | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| coord1 | 172.27.0.10 | Replica        | running | 1 | 0          |
| coord2 | 172.27.0.6  | Quorum Standby | running | 1 | 0          |
| coord3 | 172.27.0.4  | Leader         | running | 1 |            |
+-----+-----+-----+-----+-----+-----+

postgres@coord1:~$ patronictl list demo --group 1
+ Citus cluster: demo (group: 1, 7179854923881963547) -----+
| Member | Host       | Role           | State  | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| work1-1 | 172.27.0.8  | Quorum Standby | running | 1 | 0          |
| work1-2 | 172.27.0.2  | Leader         | running | 1 |            |
+-----+-----+-----+-----+-----+-----+
```



## 13.3 Citus worker switchover

When a switchover is orchestrated for a Citus worker node, Citus offers the opportunity to make the switchover close to transparent for an application. Because the application connects to the coordinator, which in turn connects to the worker nodes, then it is possible with Citus to *pause* the SQL traffic on the coordinator for the shards hosted on a worker node. The switchover then happens while the traffic is kept on the coordinator, and resumes as soon as a new primary worker node is ready to accept read-write queries.

An example of *patronictl switchover* on the worker cluster:

```
postgres@coord1:~$ patronictl switchover demo
+ Citus cluster: demo -----+-----+-----+-----+
| Group | Member | Host          | Role          | State | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+-----+
| 0 | coord1 | 172.27.0.10 | Replica       | running | 1 | 0 |
| 0 | coord2 | 172.27.0.6  | Quorum Standby | running | 1 | 0 |
| 0 | coord3 | 172.27.0.4  | Leader        | running | 1 |  |
| 1 | work1-1 | 172.27.0.8  | Leader        | running | 1 |  |
| 1 | work1-2 | 172.27.0.2  | Quorum Standby | running | 1 | 0 |
| 2 | work2-1 | 172.27.0.5  | Quorum Standby | running | 1 | 0 |
| 2 | work2-2 | 172.27.0.7  | Leader        | running | 1 |  |
+-----+-----+-----+-----+-----+-----+-----+

Citus group: 2
Primary [work2-2]:
Candidate ['work2-1'] []:
When should the switchover take place (e.g. 2024-08-26T08:02 ) [now]:
Current cluster topology
+ Citus cluster: demo (group: 2, 7179854924063375386) --+-----+
| Member | Host          | Role          | State | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| work2-1 | 172.27.0.5 | Quorum Standby | running | 1 | 0 |
| work2-2 | 172.27.0.7 | Leader        | running | 1 |  |
+-----+-----+-----+-----+-----+-----+

Are you sure you want to switchover cluster demo, demoting current primary work2-2? [y/
↵N]: y
2024-08-26 07:02:40.33003 Successfully switched over to "work2-1"
+ Citus cluster: demo (group: 2, 7179854924063375386) -----+
| Member | Host          | Role          | State | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+
| work2-1 | 172.27.0.5 | Leader        | running | 1 |  |
| work2-2 | 172.27.0.7 | Replica       | stopped |  | unknown |
+-----+-----+-----+-----+-----+-----+

postgres@coord1:~$ patronictl list demo
+ Citus cluster: demo -----+-----+-----+-----+
| Group | Member | Host          | Role          | State | TL | Lag in MB |
+-----+-----+-----+-----+-----+-----+-----+
| 0 | coord1 | 172.27.0.10 | Replica       | running | 1 | 0 |
| 0 | coord2 | 172.27.0.6  | Quorum Standby | running | 1 | 0 |
| 0 | coord3 | 172.27.0.4  | Leader        | running | 1 |  |
| 1 | work1-1 | 172.27.0.8  | Leader        | running | 1 |  |
| 1 | work1-2 | 172.27.0.2  | Quorum Standby | running | 1 | 0 |
| 2 | work2-1 | 172.27.0.5  | Leader        | running | 2 |  |
```

(continues on next page)

(continued from previous page)

	2		work2-2		172.27.0.7		Quorum Standby		running		2		0	
+-----+-----+-----+-----+-----+-----+-----+-----+														

And this is how it looks on the coordinator side:

```
# The worker primary notifies the coordinator that it is going to execute "pg_ctl stop".
2024-08-26 07:02:38,636 DEBUG: query(BEGIN, ())
2024-08-26 07:02:38,636 DEBUG: query(SELECT pg_catalog.citus_update_node(%s, %s, %s,
↳ true, %s), (3, '172.19.0.7-demoted', 5432, 10000))
# From this moment all application traffic on the coordinator to the worker group 2 is
↳ paused.

# The old worker primary is assigned as a secondary.
2024-08-26 07:02:40,084 DEBUG: query(SELECT pg_catalog.citus_update_node(%s, %s, %s,
↳ true, %s), (7, '172.19.0.7', 5432, 10000))

# The future worker primary notifies the coordinator that it acquired the leader lock in
↳ DCS and about to run "pg_ctl promote".
2024-08-26 07:02:40,085 DEBUG: query(SELECT pg_catalog.citus_update_node(%s, %s, %s,
↳ true, %s), (3, '172.19.0.5', 5432, 10000))

# The new worker primary just finished promote and notifies coordinator that it is ready
↳ to accept read-write traffic.
2024-08-26 07:02:41,485 DEBUG: query(COMMIT, ())
# From this moment the application traffic on the coordinator to the worker group 2 is
↳ unblocked.
```

## 13.4 Secondary nodes

Starting from Patroni v4.0.0 Citus secondary nodes without `no_loadbalance_tag` are also registered in `pg_dist_node`. However, to use secondary nodes for read-only queries applications need to change `citus.use_secondary_nodes` GUC.

## 13.5 Peek into DCS

The Citus cluster (coordinator and workers) are stored in DCS as a fleet of Patroni clusters logically grouped together:

```
/service/batman/                # scope=batman
/service/batman/0/              # citus.group=0, coordinator
/service/batman/0/initialize
/service/batman/0/leader
/service/batman/0/members/
/service/batman/0/members/m1
/service/batman/0/members/m2
/service/batman/1/              # citus.group=1, worker
/service/batman/1/initialize
/service/batman/1/leader
/service/batman/1/members/
/service/batman/1/members/m3
```

(continues on next page)

(continued from previous page)

```
/service/batman/1/members/m4
...
```

Such an approach was chosen because for most DCS it becomes possible to fetch the entire Citus cluster with a single recursive read request. Only Citus coordinator nodes are reading the whole tree, because they have to discover worker nodes. Worker nodes are reading only the subtree for their own group and in some cases they could read the subtree of the coordinator group.

## 13.6 Citus on Kubernetes

Since Kubernetes doesn't support hierarchical structures we had to include the citus group to all K8s objects Patroni creates:

```
batman-0-leader # the leader config map for the coordinator
batman-0-config # the config map holding initialize, config, and history "keys"
...
batman-1-leader # the leader config map for worker group 1
batman-1-config
...
```

I.e., the naming pattern is: `${scope}-${citus.group}-${type}`.

All Kubernetes objects are discovered by Patroni using the `label selector`, therefore all Pods with Patroni&Citus and Endpoints/ConfigMaps must have similar labels, and Patroni must be configured to use them using Kubernetes `settings` or `environment variables`.

A couple of examples of Patroni configuration using Pods environment variables:

1. for the coordinator cluster

```
apiVersion: v1
kind: Pod
metadata:
  labels:
    application: patroni
    citus-group: "0"
    citus-type: coordinator
    cluster-name: citusdemo
  name: citusdemo-0-0
  namespace: default
spec:
  containers:
  - env:
    - name: PATRONI_SCOPE
      value: citusdemo
    - name: PATRONI_NAME
      valueFrom:
        fieldRef:
          apiVersion: v1
          fieldPath: metadata.name
    - name: PATRONI_KUBERNETES_POD_IP
      valueFrom:
```

(continues on next page)

(continued from previous page)

```

    fieldRef:
      apiVersion: v1
      fieldPath: status.podIP
  - name: PATRONI_KUBERNETES_NAMESPACE
    valueFrom:
      fieldRef:
        apiVersion: v1
        fieldPath: metadata.namespace
  - name: PATRONI_KUBERNETES_LABELS
    value: '{application: patroni}'
  - name: PATRONI_CITUS_DATABASE
    value: citus
  - name: PATRONI_CITUS_GROUP
    value: "0"

```

2. for the worker cluster from the group 2

```

apiVersion: v1
kind: Pod
metadata:
  labels:
    application: patroni
    citus-group: "2"
    citus-type: worker
    cluster-name: citusdemo
  name: citusdemo-2-0
  namespace: default
spec:
  containers:
  - env:
    - name: PATRONI_SCOPE
      value: citusdemo
    - name: PATRONI_NAME
      valueFrom:
        fieldRef:
          apiVersion: v1
          fieldPath: metadata.name
    - name: PATRONI_KUBERNETES_POD_IP
      valueFrom:
        fieldRef:
          apiVersion: v1
          fieldPath: status.podIP
    - name: PATRONI_KUBERNETES_NAMESPACE
      valueFrom:
        fieldRef:
          apiVersion: v1
          fieldPath: metadata.namespace
    - name: PATRONI_KUBERNETES_LABELS
      value: '{application: patroni}'
    - name: PATRONI_CITUS_DATABASE
      value: citus
    - name: PATRONI_CITUS_GROUP

```

(continues on next page)

(continued from previous page)

**value:** "2"

As you may noticed, both examples have `citus-group` label set. This label allows Patroni to identify object as belonging to a certain Citus group. In addition to that, there is also `PATRONI_CITUS_GROUP` environment variable, which has the same value as the `citus-group` label. When Patroni creates new Kubernetes objects ConfigMaps or Endpoints, it automatically puts the `citus-group: ${env.PATRONI_CITUS_GROUP}` label on them:

```
apiVersion: v1
kind: ConfigMap
metadata:
  name: citusdemo-0-leader # Is generated as ${env.PATRONI_SCOPE}-${env.PATRONI_CITUS_
↪GROUP}-leader
  labels:
    application: patroni # Is set from the ${env.PATRONI_KUBERNETES_LABELS}
    cluster-name: citusdemo # Is automatically set from the ${env.PATRONI_SCOPE}
    citus-group: '0' # Is automatically set from the ${env.PATRONI_CITUS_GROUP}
```

You can find a complete example of Patroni deployment on Kubernetes with Citus support in the `kubernetes` folder of the Patroni repository.

There are two important files for you:

1. `Dockerfile.citus`
2. `citus_k8s.yaml`

## 13.7 Citus upgrades and PostgreSQL major upgrades

First, please read about upgrading Citus version in the [documentation](#). There is one minor change in the process. When executing upgrade, you have to use `patronictl restart` instead of `systemctl restart` to restart PostgreSQL.

The PostgreSQL major upgrade with Citus is a bit more complex. You will have to combine techniques used in the Citus documentation about major upgrades and Patroni documentation about [PostgreSQL major upgrade](#). Please keep in mind that Citus cluster consists of many Patroni clusters (coordinator and workers) and they all have to be upgraded independently.



## CONVERT A STANDALONE TO A PATRONI CLUSTER

This section describes the process for converting a standalone PostgreSQL instance into a Patroni cluster.

To deploy a Patroni cluster without using a pre-existing PostgreSQL instance, see [Running and Configuring](#) instead.

### 14.1 Procedure

You can find below an overview of steps for converting an existing Postgres cluster to a Patroni managed cluster. In the steps we assume all nodes that are part of the existing cluster are currently up and running, and that you *do not* intend to change Postgres configuration while the migration is ongoing. The steps:

1. Create the Postgres users as explained for [authentication](#) section of the Patroni configuration. You can find sample SQL commands to create the users in the code block below, in which you need to replace the usernames and passwords as per your environment. If you already have the relevant users, then you can skip this step.

```
-- Patroni superuser
-- Replace PATRONI_SUPERUSER_USERNAME and PATRONI_SUPERUSER_PASSWORD accordingly
CREATE USER PATRONI_SUPERUSER_USERNAME WITH SUPERUSER ENCRYPTED PASSWORD 'PATRONI_
↳SUPERUSER_PASSWORD';

-- Patroni replication user
-- Replace PATRONI_REPLICATION_USERNAME and PATRONI_REPLICATION_PASSWORD accordingly
CREATE USER PATRONI_REPLICATION_USERNAME WITH REPLICATION ENCRYPTED PASSWORD
↳'PATRONI_REPLICATION_PASSWORD';

-- Patroni rewind user, if you intend to enable use_pg_rewind in your Patroni_
↳configuration
-- Replace PATRONI_REWIND_USERNAME and PATRONI_REWIND_PASSWORD accordingly
CREATE USER PATRONI_REWIND_USERNAME WITH ENCRYPTED PASSWORD 'PATRONI_REWIND_PASSWORD
↳';
GRANT EXECUTE ON function pg_catalog.pg_ls_dir(text, boolean, boolean) TO PATRONI_
↳REWIND_USERNAME;
GRANT EXECUTE ON function pg_catalog.pg_stat_file(text, boolean) TO PATRONI_REWIND_
↳USERNAME;
GRANT EXECUTE ON function pg_catalog.pg_read_binary_file(text) TO PATRONI_REWIND_
↳USERNAME;
GRANT EXECUTE ON function pg_catalog.pg_read_binary_file(text, bigint, bigint,
↳boolean) TO PATRONI_REWIND_USERNAME;
```

2. Perform the following steps on all Postgres nodes. Perform all steps on one node before proceeding with the next node. Start with the primary node, then proceed with each standby node:

1. If you are running Postgres through systemd, then disable the Postgres systemd unit. This is performed as Patroni manages starting and stopping the Postgres daemon.
2. Create a YAML configuration file for Patroni. You can use *Patroni configuration generation and validation tooling* for that.
  - **Note (specific for the primary node):** If you have replication slots being used for replication between cluster members, then it is recommended that you enable `use_slots` and configure the existing replication slots as permanent via the `slots` configuration item. Be aware that Patroni automatically creates replication slots for replication between members, and drops replication slots that it does not recognize, when `use_slots` is enabled. The idea of using permanent slots here is to allow your existing slots to persist while the migration to Patroni is in progress. See *YAML Configuration Settings* for details.
3. Start Patroni using the `patroni` systemd service unit. It automatically detects that Postgres is already running and starts monitoring the instance.
3. Hand over Postgres “start up procedure” to Patroni. In order to do that you need to restart the cluster members through `patronictl restart cluster-name member-name` command. For minimal downtime you might want to split this step into:
  1. Immediate restart of the standby nodes.
  2. Scheduled restart of the primary node within a maintenance window.
4. If you configured permanent slots in step 1.2., then you should remove them from `slots` configuration through `patronictl edit-config cluster-name member-name` command once the `restart_lsn` of the slots created by Patroni is able to catch up with the `restart_lsn` of the original slots for the corresponding members. By removing the slots from `slots` configuration you will allow Patroni to drop the original slots from your cluster once they are not needed anymore. You can find below an example query to check the `restart_lsn` of a couple slots, so you can compare them:

```
-- Assume original_slot_for_member_x is the name of the slot in your original
-- cluster for replicating changes to member X, and slot_for_member_x is the
-- slot created by Patroni for that purpose. You need restart_lsn of
-- slot_for_member_x to be >= restart_lsn of original_slot_for_member_x
SELECT slot_name,
       restart_lsn
FROM pg_replication_slots
WHERE slot_name IN (
    'original_slot_for_member_x',
    'slot_for_member_x'
)
```



## MAJOR UPGRADE OF POSTGRESQL VERSION

The only possible way to do a major upgrade currently is:

1. Stop Patroni
2. Upgrade PostgreSQL binaries and perform `pg_upgrade` on the primary node
3. Update `patroni.yml`
4. Remove the initialize key from DCS or wipe complete cluster state from DCS. The second one could be achieved by running `patronictl remove cluster-name`. It is necessary because `pg_upgrade` runs `initdb` which actually creates a new database with a new PostgreSQL system identifier.
5. If you wiped the cluster state in the previous step, you may wish to copy `patroni.dynamic.json` from old data dir to the new one. It will help you to retain some PostgreSQL parameters you had set before.
6. Start Patroni on the primary node.
7. Upgrade PostgreSQL binaries, update `patroni.yml` and wipe the `data_dir` on standby nodes.
8. Start Patroni on the standby nodes and wait for the replication to complete.

Running `pg_upgrade` on standby nodes is not supported by PostgreSQL. If you know what you are doing, you can try the `rsync` procedure described in <https://www.postgresql.org/docs/current/pgupgrade.html> instead of wiping `data_dir` on standby nodes. The safest way is however to let Patroni replicate the data for you.

### 15.1 FAQ

- During Patroni startup, Patroni complains that it cannot bind to the PostgreSQL port.

You need to verify `listen_addresses` and `port` in `postgresql.conf` and `postgresql.listen` in `patroni.yml`. Don't forget that `pg_hba.conf` should allow such access.

- After asking Patroni to restart the node, PostgreSQL displays the error message `could not open configuration file "/etc/postgresql/10/main/pg_hba.conf": No such file or directory`

It can mean various things depending on how you manage PostgreSQL configuration. If you specified `postgresql.config_dir`, Patroni generates the `pg_hba.conf` based on the settings in the `bootstrap` section only when it bootstraps a new cluster. In this scenario the PGDATA was not empty, therefore no bootstrap happened. This file must exist beforehand.



## INTEGRATION WITH OTHER TOOLS

Patroni is able to integrate with other tools in your stack. In this section you will find a list of examples, which although not an exhaustive list, might provide you with ideas on how Patroni can integrate with other tools.

### 16.1 Barman

Patroni delivers an application named `patroni_barman` which has logic to communicate with `pg-backup-api`, so you are able to perform Barman operations remotely.

This application currently has a couple of sub-commands: `recover` and `config-switch`.

#### 16.1.1 `patroni_barman recover`

The `recover` sub-command can be used as a custom bootstrap or custom replica creation method. You can find more information about that in *Replica imaging and bootstrap*.

#### 16.1.2 `patroni_barman config-switch`

The `config-switch` sub-command is designed to be used as an `on_role_change` callback in Patroni. As an example, assume you are streaming WALs from your current primary to your Barman host. In the event of a failover in the cluster you might want to start streaming WALs from the new primary. You can accomplish this by using `patroni_barman config-switch` as the `on_role_change` callback.

#### **Note**

That sub-command relies on the `barman config-switch` command, which is in charge of overriding the configuration of a Barman server by applying a pre-defined model on top of it. This command is available since Barman 3.10. Please consult the Barman documentation for more details.

This is an example of how you can configure Patroni to apply a configuration model in case this Patroni node is promoted to primary:

```
postgresql:
  callbacks:
    on_role_change: >
      patroni_barman
      --api-url YOUR_API_URL
      config-switch
```

(continues on next page)

(continued from previous page)

```
--barman-server YOUR_BARMAN_SERVER_NAME
--barman-model YOUR_BARMAN_MODEL_NAME
--switch-when promoted
```

#### Note

`patroni_barman config-switch` requires that you have both Barman and `pg-backup-api` configured in the Barman host, so it can execute a remote `barman config-switch` through the backup API. Also, it requires that you have pre-configured Barman models to be applied. The above example uses a subset of the available parameters. You can get more information running `patroni_barman config-switch --help`, and by consulting the Barman documentation.

## SECURITY CONSIDERATIONS

A Patroni cluster has two interfaces to be protected from unauthorized access: the distributed configuration storage (DCS) and the Patroni REST API.

### 17.1 Protecting DCS

Patroni and *patronictl* both store and retrieve data to/from the DCS.

Despite DCS doesn't contain any sensitive information, it allows changing some of Patroni/Postgres configuration. Therefore the very first thing that should be protected is DCS itself.

The details of protection depend on the type of DCS used. The authentication and encryption parameters (tokens/basic-auth/client certificates) for the supported types of DCS are covered in *settings*.

The general recommendation is to enable TLS for all DCS communication.

### 17.2 Protecting the REST API

Protecting the REST API is a more complicated task.

The Patroni REST API is used by Patroni itself during the leader race, by the *patronictl* tool in order to perform failovers/switchovers/reinitialize/restarts/reloads, by HAProxy or any other kind of load balancer to perform HTTP health checks, and of course could also be used for monitoring.

From the point of view of security, REST API contains safe (GET requests, only retrieve information) and unsafe (PUT, POST, PATCH and DELETE requests, change the state of nodes) endpoints.

The unsafe endpoints can be protected with HTTP basic-auth by setting the `restapi.authentication.username` and `restapi.authentication.password` parameters. There is no way to protect the safe endpoints without enabling TLS.

When TLS for the REST API is enabled and a PKI is established, mutual authentication of the API server and API client is possible for all endpoints.

The `restapi` section parameters enable TLS client authentication to the server. Depending on the value of the `verify_client` parameter, the API server requires a successful client certificate verification for both safe and unsafe API calls (`verify_client: required`), or only for unsafe API calls (`verify_client: optional`), or for no API calls (`verify_client: none`).

The `ctl` section parameters enable TLS server authentication to the client (the *patronictl* tool which uses the same config as patroni). Set `insecure: true` to disable the server certificate verification by the client. See *settings* for a detailed description of the TLS client parameters.

Protecting the PostgreSQL database proper from unauthorized access is beyond the scope of this document and is covered in <https://www.postgresql.org/docs/current/client-authentication.html>

## HA MULTI DATACENTER

The high availability of a PostgreSQL cluster deployed in multiple data centers is based on replication, which can be synchronous or asynchronous ([replication\\_modes](#)).

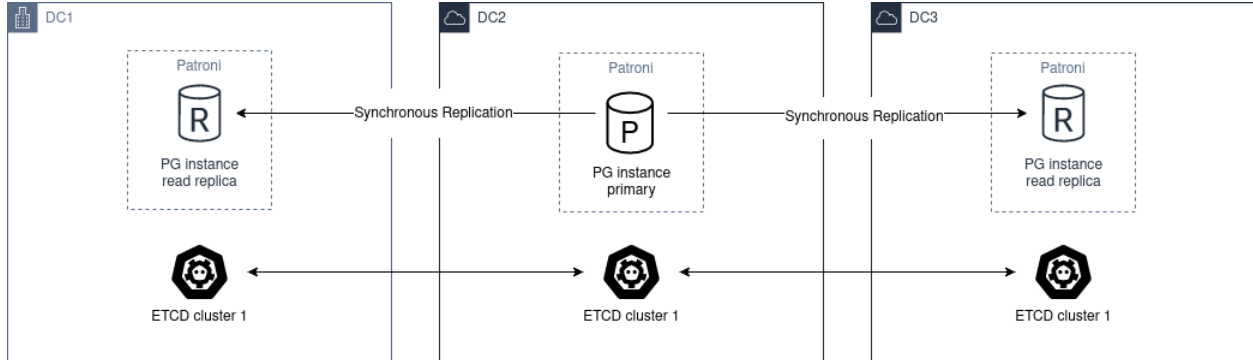
In both cases, it is important to be clear about the following concepts:

- Postgres can run as primary or standby leader only when it owns the leading key and can update the leading key.
- You should run the odd number of etcd, ZooKeeper or Consul nodes: 3 or 5!

### 18.1 Synchronous Replication

To have a multi DC cluster that can automatically tolerate a zone drop, a minimum of 3 is required.

The architecture diagram would be the following:



We must deploy a cluster of etcd, ZooKeeper or Consul through the different DC, with a minimum of 3 nodes, one in each zone.

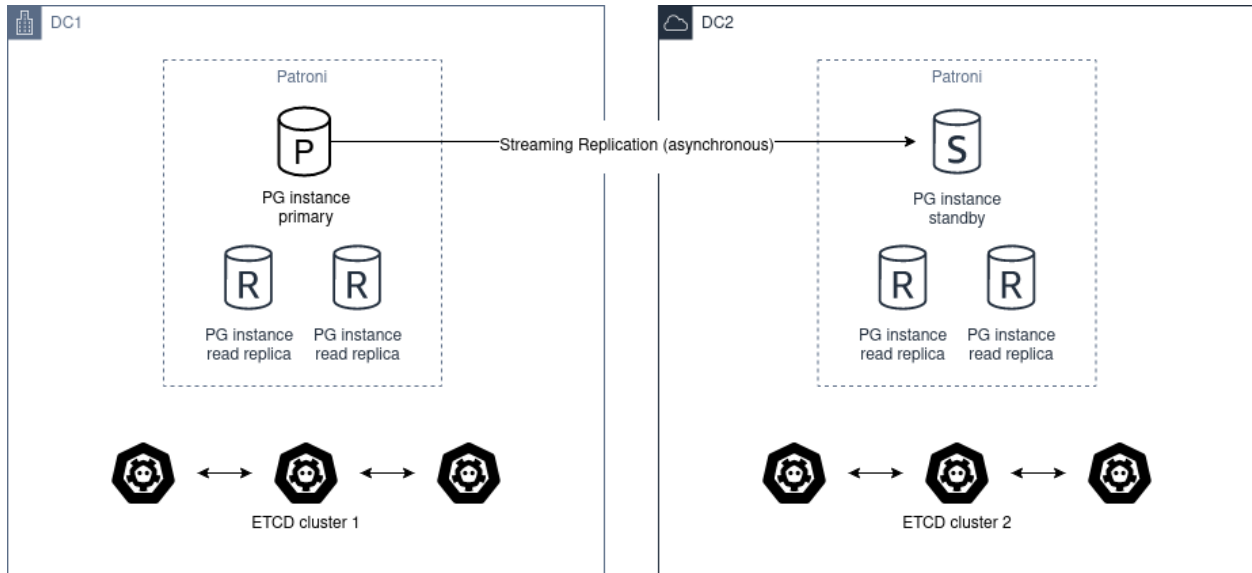
Regarding postgres, we must deploy at least 2 nodes, in different DC. Then you have to set `synchronous_mode: true` in the global [dynamic configuration](#).

This enables sync replication and the primary node will choose one of the nodes as synchronous.

## 18.2 Asynchronous Replication

With only two data centers it would be better to have two independent etcd clusters and run Patroni *standby cluster* in the second data center. If the first site is down, you can MANUALLY promote the *standby\_cluster*.

The architecture diagram would be the following:



Automatic promotion is not possible, because DC2 will never be able to figure out the state of DC1.

You should not use `pg_ctl promote` in this scenario, you need “manually promote” the healthy cluster by removing *standby\_cluster* section from the *dynamic configuration*.

### Warning

If the source cluster is still up and running and you promote the standby cluster you create a split-brain.

In case you want to return to the “initial” state, there are only two ways of resolving it:

- Add the *standby\_cluster* section back and it will trigger `pg_rewrite`; however, for `pg_rewrite` to function properly, either the cluster must be initialized with `data page checksums` (`--data-checksums` option for `initdb`) and/or `wal_log_hints` must be set to on, but there are still chances that `pg_rewrite` might fail due to other factors.
- Rebuild the standby cluster from scratch.

Before promoting standby cluster one have to manually ensure that the source cluster is down (STONITH). When DC1 recovers, the cluster has to be converted to a standby cluster.

Before doing that you may manually examine the database and extract all changes that happened between the time when network between DC1 and DC2 has stopped working and the time when you manually stopped the cluster in DC1.

Once extracted, you may also manually apply these changes to the cluster in DC2.



In this section you will find answers for the most frequently asked questions about Patroni. Each sub-section attempts to focus on different kinds of questions.

We hope that this helps you to clarify most of your questions. If you still have further concerns or find yourself facing an unexpected issue, please refer to [Chatting](#) and [Reporting bugs](#) for instructions on how to get help or report issues.

## 19.1 Comparison with other HA solutions

### **Why does Patroni require a separate cluster of DCS nodes while other solutions like repmgr do not?**

There are different ways of implementing HA solutions, each of them with their pros and cons.

Software like repmgr performs communication among the nodes to decide when actions should be taken.

Patroni on the other hand relies on the state stored in the DCS. The DCS acts as a source of truth for Patroni to decide what it should do.

While having a separate DCS cluster can make you bloat your architecture, this approach also makes it less likely for split-brain scenarios to happen in your Postgres cluster.

### **What is the difference between Patroni and other HA solutions in regards to Postgres management?**

Patroni does not just manage the high availability of the Postgres cluster but also manages Postgres itself.

If Postgres nodes do not exist yet, it takes care of bootstrapping the primary and the standby nodes, and also manages Postgres configuration of the nodes. If the Postgres nodes already exist, Patroni will take over management of the cluster.

Besides the above, Patroni also has self-healing capabilities. In other words, if a primary node fails, Patroni will not only fail over to a replica, but also attempt to rejoin the former primary as a replica of the new primary. Similarly, if a replica fails, Patroni will attempt to rejoin that replica.

That is way we call Patroni as a “template for HA solutions”. It goes further than just managing physical replication: it manages Postgres as a whole.

## 19.2 DCS

### Can I use the same etcd cluster to store data from two or more Patroni clusters?

Yes, you can!

Information about a Patroni cluster is stored in the DCS under a path prefixed with the namespace and scope Patroni settings.

As long as you do not have conflicting namespace and scope across different Patroni clusters, you should be able to use the same DCS cluster to store information from multiple Patroni clusters.

### What occurs if I attempt to use the same combination of namespace and scope for different Patroni clusters that point to the same DCS cluster?

The second Patroni cluster that attempts to use the same namespace and scope will not be able to manage Postgres because it will find information related with that same combination in the DCS, but with an incompatible Postgres system identifier. The mismatch on the system identifier causes Patroni to abort the management of the second cluster, as it assumes that refers to a different cluster and that the user has misconfigured Patroni.

Make sure to use different namespace / scope when dealing with different Patroni clusters that share the same DCS cluster.

### What occurs if I lose my DCS cluster?

The DCS is used to store basically status and the dynamic configuration of the Patroni cluster.

The very first consequence is that all the Patroni clusters that rely on that DCS will go to read-only mode – unless *DCS Failsafe Mode* is enabled.

### What should I do if I lose my DCS cluster?

There are three possible outcomes upon losing your DCS cluster:

1. The DCS cluster is fully recovered: this requires no action from the Patroni side. Once the DCS cluster is recovered, Patroni should be able to recover too;
2. The DCS cluster is re-created in place, and the endpoints remain the same. No changes are required on the Patroni side;
3. A new DCS cluster is created with different endpoints. You will need to update the DCS endpoints in the Patroni configuration of each Patroni node.

If you face scenario 2. or 3. Patroni will take care of creating the status information again based on the current status of the cluster, and recreate the dynamic configuration on the DCS based on a backup file named `patroni.dynamic.json` which is stored inside the Postgres data directory of each member of the Patroni cluster.

### What occurs if I lose majority in my DCS cluster?

The DCS will become unresponsive, which will cause Patroni to demote the current read/write Postgres node.

Remember: Patroni relies on the state of the DCS to take actions on the cluster.

You can use the *DCS Failsafe Mode* to alleviate that situation.

## 19.3 patronictl

### Do I need to run *patronictl* in the Patroni host?

No, you do not need to do that.

Running *patronictl* in the Patroni host is handy if you have access to the Patroni host because you can use the very same configuration file from the *patroni* agent for the *patronictl* application.

However, *patronictl* is basically a client and it can be executed from remote machines. You just need to provide it with enough configuration so it can reach the DCS and the REST API of the Patroni member(s).

### Why did the information from one of my Patroni members disappear from the output of *patronictl list* command?

Information shown by *patronictl list* is based on the contents of the DCS.

If information about a member disappeared from the DCS it is very likely that the Patroni agent on that node is not running anymore, or it is not able to communicate with the DCS.

As the member is not able to update the information, the information eventually expires from the DCS, and consequently the member is not shown anymore in the output of *patronictl list*.

### Why is the information about one of my Patroni members not up-to-date in the output of *patronictl list* command?

Information shown by *patronictl list* is based on the contents of the DCS.

By default, that information is updated by Patroni roughly every `loop_wait` seconds. In other words, even if everything is normally functional you may still see a “delay” of up to `loop_wait` seconds in the information stored in the DCS.

Be aware that that is not a rule, though. Some operations performed by Patroni cause it to immediately update the DCS information.

## 19.4 Configuration

### What is the difference between dynamic configuration and local configuration?

Dynamic configuration (or global configuration) is the configuration stored in the DCS, and which is applied to all members of the Patroni cluster. This is primarily where you should store your configuration.

Settings that are specific to a node, or settings that you would like to overwrite the global configuration with, you should set only on the desired Patroni member as a local configuration. That local configuration can be specified either through the configuration file or through environment variables.

See more in *Patroni configuration*.

### What are the types of configuration in Patroni, and what is the precedence?

The types are:

- Dynamic configuration: applied to all members;
- Local configuration: applied to the local member, overrides dynamic configuration;
- Environment configuration: applied to the local member, overrides both dynamic and local configuration.

**Note:** some Postgres GUCs can only be set globally, i.e., through dynamic configuration. Besides that, there are GUCs which Patroni enforces a hard-coded value.

See more in *Patroni configuration*.

### Is there any facility to help me create my Patroni configuration file?

Yes, there is.

You can use `patroni --generate-sample-config` or `patroni --generate-config` commands to generate a sample Patroni configuration or a Patroni configuration based on an existing Postgres instance, respectively.

Please refer to [Sample Patroni configuration](#) and [Patroni configuration for a running instance](#) for more details.

**I changed my parameters under `bootstrap.dcs` configuration but Patroni is not applying the changes to the cluster members. What is wrong?**

The values configured under `bootstrap.dcs` are only used when bootstrapping a fresh cluster. Those values will be written to the DCS during the bootstrap.

After the bootstrap phase finishes, you will only be able to change the dynamic configuration through the DCS.

Refer to the next question for more details.

**How can I change my dynamic configuration?**

You need to change the configuration in the DCS. That is accomplished either through:

- `patronictl edit-config`; or
- A PATCH request to [Config endpoint](#).

**How can I change my local configuration?**

You need to change the configuration file of the corresponding Patroni member and signal the Patroni agent with `SIGHUP`. You can do that using either of these approaches:

- Send a POST request to the REST API [Reload endpoint](#); or
- Run `patronictl reload`; or
- Locally signal the Patroni process with `SIGHUP`:
  - If you started Patroni through `systemd`, you can use the command `systemctl reload PATRONI_UNIT.service`, `PATRONI_UNIT` being the name of the Patroni service; or
  - If you started Patroni through other means, you will need to identify the `patroni` process and run `kill -s HUP PID`, `PID` being the process ID of the `patroni` process.

**Note:** there are cases where a reload through the `patronictl reload` may not work:

- Expired REST API certificates: you can mitigate that by using the `-k` option of the `patronictl`;
- Wrong credentials: for example when changing `restapi` or `ctl` credentials in the configuration file, and using that same configuration file for Patroni and `patronictl`.

**How can I change my environment configuration?**

The environment configuration is only read by Patroni during startup.

With that in mind, if you change the environment configuration you will need to restart the corresponding Patroni agent.

Take care to not cause a failover in the cluster! You might be interested in checking [patronictl pause](#).

**What occurs if I change a Postgres GUC that requires a reload?**

When you change the dynamic or the local configuration as explained in the previous questions, Patroni will take care of reloading the Postgres configuration for you.

**What occurs if I change a Postgres GUC that requires a restart?**

Patroni will mark the affected members with a flag of `pending restart`.

It is up to you to determine when and how to restart the members. That can be accomplished either through:

- `patronictl restart`; or
- A POST request to [Restart endpoint](#).

**Note:** some Postgres GUCs require a special management in terms of the order for restarting the Postgres nodes. Refer to [PostgreSQL parameters that touch shared memory](#) for more details.

**What is the difference between etcd and etcd3 in Patroni configuration?**

etcd uses the API version 2 of etcd, while etcd3 uses the API version 3 of etcd.

Be aware that information stored by the API version 2 is not manageable by API version 3 and vice-versa.

We recommend that you configure etcd3 instead of etcd because:

- API version 2 is disabled by default from Etcd v3.4 onward;
- API version 2 will be completely removed on Etcd v3.6.

**I have use\_slots enabled in my Patroni configuration, but when a cluster member goes offline for some time, the replication slot used by that member is dropped on the upstream node. What can I do to avoid that issue?**

There are two options:

1. You can tune `member_slots_ttl` (default value 30min, available since Patroni 4.0.0 and PostgreSQL 11 onwards) and replication slots for absent members will not be removed when the members downtime is shorter than the configured threshold.
2. You can configure permanent physical replication slots for the members.

Since Patroni 3.2.0 it is now possible to have member slots as permanent slots managed by Patroni.

Patroni will create the permanent physical slots on all nodes, and make sure to not remove the slots, as well as to advance the slots' LSN on all nodes according to the LSN that has been consumed by the member.

Later, if you decide to remove the corresponding member, it's **your responsibility** to adjust the permanent slots configuration, otherwise Patroni will keep the slots around forever.

**Note:** on Patroni older than 3.2.0 you could still have member slots configured as permanent physical slots, however they would be managed only on the current leader. That is, in case of failover/switchover these slots would be created on the new leader, but that wouldn't guarantee that it had all WAL segments for the absent node.

**Note:** even with Patroni 3.2.0 there might be a small race condition. In the very beginning, when the slot is created on the replica it could be ahead of the same slot on the leader and in case if nobody is consuming the slot there is still a chance that some files could be missing after failover. With that in mind, it is recommended that you configure continuous archiving, which makes it possible to restore required WALs or perform PITR.

**What is the difference between loop\_wait, retry\_timeout and ttl?**

Patroni performs what we call a HA cycle from time to time. On each HA cycle it takes care of performing a series of checks on the cluster to determine its healthiness, and depending on the status it may take actions, like failing over to a standby.

`loop_wait` determines for how long, in seconds, Patroni should sleep before performing a new cycle of HA checks.

`retry_timeout` sets the timeout for retry operations on the DCS and on Postgres. For example: if the DCS is unresponsive for more than `retry_timeout` seconds, Patroni might demote the primary node as a security action.

`ttl` sets the lease time on the leader lock in the DCS. If the current leader of the cluster is not able to renew the lease during its HA cycles for longer than `ttl`, then the lease will expire and that will trigger a leader race in the cluster.

**Note:** when modifying these settings, please keep in mind that Patroni enforces the rule and minimal values described in *Dynamic Configuration Settings* section of the docs.

## 19.5 Postgres management

### Can I change Postgres GUCs directly in Postgres configuration?

You can, but you should avoid that.

Postgres configuration is managed by Patroni, and attempts to edit the configuration files may end up being frustrated by Patroni as it may eventually overwrite them.

There are a few options available to overcome the management performed by Patroni:

- Change Postgres GUCs through `$PGDATA/postgresql.base.conf`; or
- Define a `postgresql.custom_conf` which will be used instead of `postgresql.base.conf` so you can manage that externally; or
- Change GUCs using `ALTER SYSTEM` / `ALTER DATABASE` / `ALTER USER`.

You can find more information about that in the section *Important rules*.

In any case we recommend that you manage all the Postgres configuration through Patroni. That will centralize the management and make it easier to debug Patroni when needed.

### Can I restart Postgres nodes directly?

No, you should **not** attempt to manage Postgres directly!

Any attempt of bouncing the Postgres server without Patroni can lead your cluster to face failovers.

If you need to manage the Postgres server, do that through the ways exposed by Patroni.

### Is Patroni able to take over management of an already existing Postgres cluster?

Yes, it can!

Please refer to *Convert a Standalone to a Patroni Cluster* for detailed instructions.

### How does Patroni manage Postgres?

Patroni takes care of bringing Postgres up and down by running the Postgres binaries, like `pg_ctl` and `postgres`.

With that in mind you **MUST** disable any other sources that could manage the Postgres clusters, like the systemd units, e.g. `postgresql.service`. Only Patroni should be able to start, stop and promote Postgres instances in the cluster. Not doing so may result in split-brain scenarios. For example: if the node running as a primary failed and the unit `postgresql.service` is enabled, it may bring Postgres back up and cause a split-brain.

## 19.6 Concepts and requirements

### Which are the applications that make part of Patroni?

Patroni basically ships a couple applications:

- `patroni`: This is the Patroni agent, which takes care of managing a Postgres node;
- `patronictl`: This is a command-line utility used to interact with a Patroni cluster (perform switchovers, restarts, changes in the configuration, etc.). Please find more information in *patronictl*.

### What is a standby cluster in Patroni?

It is a cluster that does not have any primary Postgres node running, i.e., there is no read/write member in the cluster.

These kinds of clusters exist to replicate data from another cluster and are usually useful when you want to replicate data across data centers.

There will be a leader in the cluster which will be a standby in charge of replicating changes from a remote Postgres node. Then, there will be a set of standbys configured with cascading replication from such leader member.

**Note:** the standby cluster doesn't know anything about the source cluster which it is replicating from – it can even use `restore_command` instead of WAL streaming, and may use an absolutely independent DCS cluster.

Refer to *Standby cluster* for more details.

#### What is a leader in Patroni?

A leader in Patroni is like a coordinator of the cluster.

In a regular Patroni cluster, the leader will be the read/write node.

In a standby Patroni cluster, the leader (AKA standby leader) will be in charge of replicating from a remote Postgres node, and cascading those changes to the other members of the standby cluster.

#### Does Patroni require a minimum number of Postgres nodes in the cluster?

No, you can run Patroni with any number of Postgres nodes.

Remember: Patroni is decoupled from the DCS.

#### What does pause mean in Patroni?

Pause is an operation exposed by Patroni so the user can ask Patroni to step back in regards to Postgres management.

That is mainly useful when you want to perform maintenance on the cluster, and would like to avoid that Patroni takes decisions related with HA, like failing over to a standby when you stop the primary.

You can find more information about that in *Pause/Resume mode for the cluster*.

## 19.7 Automatic failover

#### How does the automatic failover mechanism of Patroni work?

Patroni automatic failover is based on what we call leader race.

Patroni stores the cluster's status in the DCS, among them a leader lock which holds the name of the Patroni member which is the current leader of the cluster.

That leader lock has a time-to-live associated with it. If the leader node fails to update the lease of the leader lock in time, the key will eventually expire from the DCS.

When the leader lock expires, it triggers what Patroni calls a leader race: all nodes start performing checks to determine if they are the best candidates for taking over the leader role. Some of these checks include calls to the REST API of all other Patroni members.

All Patroni members that find themselves as the best candidate for taking over the leader lock will attempt to do so. The first Patroni member that is able to take the leader lock will promote itself to a read/write node (or standby leader), and the others will be configured to follow it.

#### Can I temporarily disable automatic failover in the Patroni cluster?

Yes, you can!

You can achieve that by temporarily pausing the cluster. This is typically useful for performing maintenance.

When you want to resume the automatic failover of the cluster, you just need to unpause it.

You can find more information about that in *Pause/Resume mode for the cluster*.

## 19.8 Bootstrapping and standbys creation

### How does Patroni create a primary Postgres node? What about a standby Postgres node?

By default Patroni will use `initdb` to bootstrap a fresh cluster, and `pg_basebackup` to create standby nodes from a copy of the `leader` member.

You can customize that behavior by writing your custom bootstrap methods, and your custom replica creation methods.

Custom methods are usually useful when you want to restore backups created by backup tools like `pgBackRest` or `Barman`, for example.

For detailed information please refer to *[Bootstrap](#)* and *[Building replicas](#)*.

## 19.9 Monitoring

### How can I monitor my Patroni cluster?

Patroni exposes a couple handy endpoints in its *[Patroni REST API](#)*:

- `/metrics`: exposes monitoring metrics in a format that can be consumed by Prometheus;
- `/patroni`: exposes the status of the cluster in a JSON format. The information shown here is very similar to what is shown by the `/metrics` endpoint.

You can use those endpoints to implement monitoring checks.



## RELEASE NOTES

### 20.1 Version 4.0.5

Released 2025-02-20

#### Stability improvements

- Compatibility with `python-json-logger`  $\geq 3.1$  (Alexander Kukushkin)  
Get rid of the warnings produced by the old API usage.
- Compatibility with Python 3.13 (Alexander Kukushkin)  
Run tests against Python 3.13.
- Compatibility with `pyinstaller`  $\geq 4.4$  (Joe Jensen)  
Fall back to the default `iter_modules` if `pyinstaller toc` attribute is not present.
- Fix issues with PostgreSQL 9.5 support (Alexander Kukushkin)
  - Properly handle `pg_rewind` output format.
  - Consider `synchronous_standby_names` format not supporting “num” specification.
- Compatibility with the latest changes in `urlparse` (Alexander Kukushkin)  
`urlparse` doesn't accept multiple hosts with `[]` character in URL anymore. To mitigate the problem, switch to the native wrappers of `PQconninfoParse()` from `libpq`, when it is possible, and use our implementation only for older `psycpg2` versions that are linked with an outdated version of `libpq`.

#### Bugfixes

- Show only the members to be restarted upon restart confirmation (András Váczí)  
Previously, when doing `patronictl restart <clustername> --pending`, the confirmation listed all members, regardless of whether their restart is pending.
- Cancel long-running jobs on Patroni stop and remove data directory on replica bootstrap failure (Alexander Kukushkin)  
Previously, Patroni could be doing replica bootstrap, while `pg_basebackup / wal-g / pgBackRest / barman` or similar keep running.
- Properly handle cluster names with a slash in `patronictl edit-config` (Antoni Mur)  
Replace a forward slash in `cluster_name` with an underscore.
- Avoid dropping physical slots too early (Alexander Kukushkin)  
Postpone removal of physical replication slots containing `xmin` after a failover: on the new primary – until this member is promoted, on replicas – until there is a leader in the cluster.

- Handle all exceptions raised by subprocess in `controldata()` (Alexander Kukushkin)  
Patroni was not properly handling all exceptions possibly raised when calling `pg_controldata` utility.
- Fix bug with a slot for a former leader not retained on failover (Alexander Kukushkin)  
Avoid falsely relying on members being present in DCS, while on failover `/member` key for the former leader is expiring exactly at the same time.
- Fix a couple of bugs in the quorum state machine (Alexander Kukushkin)
  - When evaluating whether there are healthy nodes for a leader race, before demoting we need to take into account quorum requirements. Without it, the former leader may end up in recovery surrounded by asynchronous nodes.
  - `QuorumStateResolver` wasn't correctly handling the case when a replica node quickly joined and disconnected.

### Improvements

- Improve error on an empty or non-dictionary configuration file (Julian)  
Throw a more explicit exception when validating if Patroni configuration file contains a valid `Mapping` object.

## 20.2 Version 4.0.4

Released 2024-11-22

### Stability improvements

- Add compatibility with the `py-consul` module (Alexander Kukushkin)  
`python-consul` module is unmaintained for a long time, while `py-consul` is the official replacement. Backward compatibility with `python-consul` is retained.
- Add compatibility with the `prettytable>=3.12.0` module (Alexander Kukushkin)  
Address deprecation warnings.
- Compatibility with the `ydiff==1.4.2` module (Alexander Kukushkin)  
Fix compatibility issues for the latest version, constrain version in `requirements.txt`, and introduce latest version compatibility test.

### Bugfixes

- Run `on_role_change` callback after a failed primary recovery (Polina Bungina, Alexander Kukushkin)  
Additionally run `on_role_change` callback for a primary that failed to start after a crash to increase chances the callback is executed, even if the further start as a replica fails.
- Fix a thread leak in `patronictl list -W` (Alexander Kukushkin)  
Cache DCS instance object to avoid thread leak.
- Ensure only supported parameters are written to the connection string (Alexander Kukushkin)  
Patroni used to pass parameters introduced in newer versions to the connection string, which had been leading to connection errors.

## 20.3 Version 4.0.3

Released 2024-10-18

### Bugfixes

- Disable `pgaudit` when creating users not to expose password (kviset)  
Patroni was logging `superuser`, `replication`, and `rewind` passwords on their creation when `pgaudit` extension was enabled.
- Fix issue with mixed setups: primary on pre-Patroni v4 and replicas on v4+ (Alexander Kukushkin)  
Use `xlog_location` extracted from `/members` key instead of trying to get a member's slot position from `/status` key if Patroni version running on the leader is pre-4.0.0. Not doing so has been causing WALs accumulation on replicas.
- Do not ignore valid PostgreSQL GUCs that don't have Patroni validator (Polina Bungina)  
Still check against `postgres --describe-config` if a GUC does not have a Patroni validator but is, in fact, a valid GUC.

### Improvements

- Recheck annotations on 409 status code when reading leader object in K8s (Alexander Kukushkin)  
Avoid an additional update if PATCH request was canceled by Patroni, while the request successfully updated the target.
- Add support of `sslnegotiation` client-side connection option (Alexander Kukushkin)  
`sslnegotiation` was added to the final PostgreSQL 17 release.

## 20.4 Version 4.0.2

Released 2024-09-17

### Bugfixes

- Handle exceptions while discovering configuration validation files (Alexander Kukushkin)  
Skip directories for which Patroni does not have sufficient permissions to perform list operations.
- Make sure inactive hot physical replication slots don't hold `xmin` (Alexander Kukushkin, Polina Bungina)  
Since version 3.2.0 Patroni creates physical replication slots for all members on replicas and periodically moves them forward using `pg_replication_slot_advance()` function. However if for any reason `hot_standby_feedback` is enabled and the primary is demoted to replica, the now inactive slots have NOT NULL `xmin` value propagated back to the new primary. This results in `xmin` horizon not being moved forward and vacuum not being able to clean up dead tuples. With this fix, Patroni recreates the physical replication slots that are supposed to be inactive but have NOT NULL `xmin` value.
- Fix unhandled `DCSError` during the startup phase (Waynerv)  
Ensure DCS connectivity before trying to check the uniqueness of the node name.
- Explicitly include `CMDLINE_OPTIONS` GUCs when querying `pg_settings` (Alexander Kukushkin)  
Make sure all GUCs that are passed to postmaster as command line parameters are restored when Patroni is joining a running standby. This is a follow-up for the bug fixed in Patroni 3.2.2.

- Fix bug in `synchronous_standby_names` quoting logic (Alexander Kukushkin)

According to PostgreSQL documentation, `ANY` and `FIRST` keywords are supposed to be double-quoted, which Patroni did not do before.

- Fix keepalive connection out-of-range issue (hadizamani021)

Ensure that `keepalive` option value calculated based on the `t1` set does not exceed the maximum allowed value for the current platform.

## 20.5 Version 4.0.1

Released 2024-08-30

### Bugfix

- Patroni was creating unnecessary replication slots for itself (Alexander Kukushkin)

It was happening if name contains upper-case or special characters.

## 20.6 Version 4.0.0

Released 2024-08-29

### Warning

- This version completes work on getting rid of the “master” term, in favor of “primary”. This means a couple of breaking changes, please read the release notes carefully. Upgrading to the Patroni 4+ will work reliably only if you run Patroni 3.1.0 or newer. Upgrading from an older version directly to 4+ is possible but may lead to unexpected behavior if the primary fails while the rest of the nodes are running on other Patroni versions.

### Breaking changes

- The following breaking changes were introduced when getting rid of the non-inclusive “master” term in the Patroni code:
  - On Kubernetes, Patroni by default will set role label to `primary`. In case if you want to keep the old behavior and avoid downtime or lengthy complex migrations, you can configure parameters `kubernetes.leader_label_value` and `kubernetes.standby_leader_label_value` to `master`. Read more [here](#).
  - Patroni role is written to DCS as `primary` instead of `master`.
  - Patroni role returned by Patroni REST API has been changed from `master` to `primary`.
  - Patroni REST API no longer accepts `role=master` in requests to `/switchover`, `/failover`, `/restart` endpoints.
  - `/metrics` REST API endpoint will no longer report `patroni_master` metric.
  - `patronictl` no longer accepts `--master` option for any command. `--leader` or `--primary` options should be used instead.
  - `no_master` option in the declarative configuration of custom replica creation methods is no longer treated as a special option, please use `no_leader` instead.
  - `patroni_wale_restore` script doesn't accept `--no_master` option anymore.

- `patroni_barman` script doesn't accept `--role=master` option anymore.
- All callback scripts are executed with `role=primary` option passed instead of `role=master`.
- `patronictl failover` does not accept `--leader` option that was deprecated since Patroni 3.2.0.
- User creation functionality (`bootstrap.users` configuration section) deprecated since Patroni 3.2.0 has been removed.

### New features

- Quorum-based failover (Ants Aasma, Alexander Kukushkin)

The feature implements quorum-based synchronous replication (available from PostgreSQL v10) which helps to reduce worst-case latencies, even during normal operation, as a higher latency of replicating to one standby can be compensated by other standbys. Patroni implements additional safeguards to prevent any user-visible data loss by choosing a failover candidate based on the latest transaction received.
- Register Citus secondaries in `pg_dist_node` (Alexander Kukushkin)

Patroni now maintains the list of nodes with `role==replica`, `state==running` and without `noloadbalance` *tag* in `pg_dist_node`.
- Configurable retention of members' replication slots (Alexander Kukushkin)

Implements support of `member_slots_ttl` global configuration parameter that controls for how long member replication slots should be kept around when the member key is absent.
- Make permissions of log files created by Patroni configurable (Alexander Kukushkin)

Allows to set specific permissions for log files created by Patroni. If not specified, permissions are set based on the current `umask` value.
- Compatibility with PostgreSQL 17 beta3 (Alexander Kukushkin)

GUC's validator rules were extended. Patroni handles all the new auxiliary backends during shutdown and sets `dbname` in `primary_conninfo`, as it is required for logical replication slots synchronization.
- Implement `--ignore-listen-port` option for Patroni config validation (Sahil Naphade)

Make it possible to ignore already bound ports when running `patroni --validate-config`.

### Improvements

- Make `wal_log_hints` configurable (Paul\_Kim)

Allows to avoid the overhead of `wal_log_hints` configuration being enabled in case `use_pg_rewind` is set to `off`.
- Log `pg_basebackup` command in `DEBUG` level (Waynerv)

Facilitates failed initialization debugging.

### Bugfixes

- Advance permanent slots for cascading nodes while in failsafe (Alexander Kukushkin)

Ensure that slots for cascading replicas are properly advanced on the primary when failsafe mode is activated. It is done by extending replicas response on `POST /failsafe` REST API request with their `xlog_location`.
- Don't let the current node be chosen as synchronous (Alexander Kukushkin)

There may be "something" streaming from the current primary node with `application_name` that matches the name of the current primary. Patroni was not properly handling this situation, which could end up in the primary being declared as a synchronous node and consequently was blocking switchovers.
- Ignore `restapi.allowlist_include_members` for `POST /failsafe` (Alexander Kukushkin)

- Improve GUCs validation (Polina Bungina)

Due to additional validation through running `postgres --describe-config` command, it was previously not possible to set GUCs not listed there through Patroni configuration. This limitation is now removed.

- Add line with `localhost` to `.pgpass` file when unix sockets are detected (Alexander Kukushkin)

Patroni will add an additional line to `.pgpass` file if `host` parameter specified starts with `/` character. This allows to cover a corner case when `host` matches the default socket directory path.

- Fix logging issues (Waynerv)

Defined proper request URL in failsafe handling logs and fixed the order of timestamps in postmaster check log.

## 20.7 Version 3.3.2

Released 2024-07-11

### Bugfixes

- Fix plain Postgres synchronous replication mode (Israel Barth Rubio)

Since `synchronous_mode` was introduced to Patroni, the plain Postgres synchronous replication was not working. With this bugfix, Patroni sets the value of `synchronous_standby_names` as configured by the user, if that is the case, when `synchronous_mode` is disabled.

- Handle logical slots invalidation on a standby (Polina Bungina)

Since PG16 logical replication slots on a standby can be invalidated due to horizon: from now on, Patroni forces copy (i.e., recreation) of invalidated slots.

- Fix race condition with logical slot advance and copy (Alexander Kukushkin)

Due to this bug, it was a possible situation when an invalidated logical replication slot was copied with PostgreSQL restart more than once.

## 20.8 Version 3.3.1

Released 2024-06-17

### Stability improvements

- Compatibility with Python 3.12 (Alexander Kukushkin)

Handle a new attribute added to `logging.LogRecord`.

### Bugfixes

- Fix infinite recursion in `replicatefrom` tags handling (Alexander Kukushkin)

As a part of this fix, also improve `is_physical_slot()` check and adjust documentation.

- Fix wrong role reporting in standby clusters (Alexander Kukushkin)

`synchronous_standby_names` and synchronous replication only work on a real primary node and in the case of cascading replication are simply ignored by Postgres. Before this fix, `patronictl list` and `GET /cluster` were falsely reporting some nodes as synchronous.

- Fix availability of the `allow_in_place_tablespaces` GUC (Polina Bungina)

`allow_in_place_tablespaces` was not only added to PostgreSQL 15 but also backpatched to PostgreSQL 10-14.

## 20.9 Version 3.3.0

Released 2024-04-04

### Warning

All older Patroni versions are not compatible with `ydiff>=1.3`.

There are the following options available to “fix” the problem:

1. upgrade Patroni to the latest version
2. install `ydiff<1.3` after installing Patroni
3. install `cdiff` module

### New features

- Add ability to pass `auth_data` to Zookeeper client (Aras Mumcuyan)  
It allows to specify the authentication credentials to use for the connection.
- Add a contrib script for Barman integration (Israel Barth Rubio)  
Provide an application `patroni_barman` that allows to perform Barman operations remotely and can be used as a custom bootstrap/custom replica method or as an `on_role_change` callback. Please check [here](#) for more information.
- Support JSON log format (alisalemmi)  
Apart from `plain` (default), Patroni now also supports `json` log format. Requires `python-json-logger>=2.0.2` library to be installed.
- Show `pending_restart_reason` information (Polina Bungina)  
Provide extended information about the PostgreSQL parameters that caused `pending_restart` flag to be set. Both `patronictl list` and `/patroni` REST API endpoint now show the parameters names and their “diff” as `pending_restart_reason`.
- Implement `nostream` tag (Grigory Smolkin)  
If `nostream` tag is set to `true`, the node will not use replication protocol to stream WAL but instead rely on archive recovery (if `restore_command` is configured). It also disables copying and synchronization of permanent logical replication slots on the node itself and all its cascading replicas.

### Improvements

- Implement validation of the `log` section (Alexander Kukushkin)  
Until now validator was not checking the correctness of the logging configuration provided.
- Improve logging for PostgreSQL parameters change (Polina Bungina)  
Convert old values to a human-readable format and log information about the `pg_controldata` vs Patroni global configuration mismatch.

### Bugfixes

- Properly filter out not allowed `pg_basebackup` options (Israel Barth Rubio)  
Due to a bug, Patroni was not properly filtering out the not allowed options configured for the basebackup replica bootstrap method, when provided in the `- setting: value` format.

- Fix etcd3 authentication error handling (Alexander Kukushkin)

Always retry one time on etcd3 authentication error if authentication was not done right before executing the request. Also, do not restart watchers on reauthentication.

- Improve logic of the validator files discovery (Waynerv)

Use `importlib` library to discover the files with available configuration parameters when possible (for Python 3.9+). This implementation is more stable and doesn't break the Patroni distributions based on zip archives.

- Use `target_session_attrs` only when multiple hosts are specified in the `standby_cluster` section (Alexander Kukushkin)

`target_session_attrs=read-write` is now added to the `primary_conninfo` on the standby leader node only when `standby_cluster.host` section contains multiple hosts separated by commas.

- Add compatibility code for `ydiff` library version 1.3+ (Alexander Kukushkin)

Patroni is relying on some API from `ydiff` that is not public because it is supposed to be just a terminal tool rather than a python module. Unfortunately, the API change in 1.3 broke old Patroni versions.

## 20.10 Version 3.2.2

Released 2024-01-17

### Bugfixes

- Don't let replica restore initialize key when DCS was wiped (Alexander Kukushkin)

It was happening in the method where Patroni was supposed to take over a standalone PG cluster.

- Use consistent read when fetching just updated sync key from Consul (Alexander Kukushkin)

Consul doesn't provide any interface to immediately get `ModifyIndex` for the key that we just updated, therefore we have to perform an explicit read operation. Since stale reads are allowed by default, we sometimes used to get an outdated version of the key.

- Reload Postgres config if a parameter that requires restart was reset to the original value (Polina Bungina)

Previously Patroni wasn't updating the config, but only resetting the `pending_restart`.

- Fix erroneous inverted logic of the confirmation prompt message when doing a failover to an async candidate in synchronous mode (Polina Bungina)

The problem existed only in `patronictl`.

- Exclude leader from failover candidates in `patronictl` (Polina Bungina)

If the cluster is healthy, failing over to an existing leader is no-op.

- Create Citus database and extension idempotently (Alexander Kukushkin, Zhao Junwang)

It will allow to create them in the `post_bootstrap` script in case if there is a need to add some more dependencies to the Citus database.

- Don't filter our contradictory `nofailover` tag (Polina Bungina)

The configuration `{nofailover: false, failover_priority: 0}` set on a node didn't allow it to participate in the race, while it should, because `nofailover` tag should take precedence.

- Fixed PyInstaller frozen issue (Sophia Ruan)

The `freeze_support()` was called after `argparse` and as a result, Patroni wasn't able to start Postgres.



- Fixed bug in the config generator for `patronictl` and Citus configuration (Israel Barth Rubio)  
It prevented `patronictl` and Citus configuration parameters set via environment variables from being written into the generated config.
- Restore recovery GUCs and some Patroni-managed parameters when joining a running standby (Alexander Kukushkin)  
Patroni was failing to restart Postgres v12 onwards with an error about missing `port` in one of the internal structures.
- Fixes around `pending_restart` flag (Polina Bungina)  
Don't expose `pending_restart` when in custom bootstrap with `recovery_target_action = promote` or when someone changed `hot_standby` or `wal_log_hints` using for example `ALTER SYSTEM`.

## 20.11 Version 3.2.1

Released 2023-11-30

### Bugfixes

- Limit accepted values for `--format` argument in `patronictl` (Alexander Kukushkin)  
It used to accept any arbitrary string and produce no output if the value wasn't recognized.
- Verify that replica nodes received checkpoint LSN on shutdown before releasing the leader key (Alexander Kukushkin)  
Previously in some cases, we were using LSN of the SWITCH record that is followed by CHECKPOINT (if archiving mode is enabled). As a result the former primary sometimes had to do `pg_rewind`, but there would be no data loss involved.
- Do a real HTTP request when performing node name uniqueness check (Alexander Kukushkin)  
When running Patroni in containers it is possible that the traffic is routed using `docker-proxy`, which listens on the port and accepts incoming connections. It was causing false positives.
- Fixed Citus support with Etcd v2 (Alexander Kukushkin)  
Patroni was failing to deploy a new Citus cluster with Etcd v2.
- Fixed `pg_rewind` behavior with Postgres v16+ (Alexander Kukushkin)  
The error message format of `pg_waldump` changed in v16 which caused `pg_rewind` to be called by Patroni even when it was not necessary.
- Fixed bug with custom bootstrap (Alexander Kukushkin)  
Patroni was falsely applying `--command` argument, which is a bootstrap command itself.
- Fixed the issue with REST API health check endpoints (Sophia Ruan)  
There were chances that after Postgres restart it could return `unknown` state for Postgres because connections were not properly closed.
- Cache `postgres --describe-config` output results (Waynerv)  
They are used to figure out which GUCs are available to validate PostgreSQL configuration and we don't expect this list to change while Patroni is running.

## 20.12 Version 3.2.0

Released 2023-10-25

### Deprecation notice

- The `bootstrap.users` support will be removed in version 4.0.0. If you need to create users after deploying a new cluster please use the `bootstrap.post_bootstrap` hook for that.

### Breaking changes

- Enforce `loop_wait + 2*retry_timeout <= ttl` rule and hard-code minimal possible values (Alexander Kukushkin)

Minimal values: `loop_wait=2`, `retry_timeout=3`, `ttl=20`. In case values are smaller or violate the rule they are adjusted and a warning is written to Patroni logs.

### New features

- Failover priority (Mark Pekala)

With the help of `tags.failover_priority` it's now possible to make a node more preferred during the leader race. More details in the documentation (ref tags).

- Implemented `patroni --generate-config [--dsn DSN]` and `patroni --generate-sample-config` (Polina Bungina)

It allows to generate a config file for the running PostgreSQL cluster or a sample config file for the new Patroni cluster.

- Use a dedicated connection to Postgres for Patroni REST API (Alexander Kukushkin)

It helps to avoid blocking the main heartbeat loop if the system is under stress.

- Enrich some endpoints with the name of the node (sskserk)

For the monitoring endpoint `name` is added next to the `scope` and for metrics endpoint the `name` is added to `tags`.

- Ensure strict failover/switchover difference (Polina Bungina)

Be more precise in log messages and allow failing over to an asynchronous node in a healthy synchronous cluster.

- Make permanent physical replication slots behave similarly to permanent logical slots (Alexander Kukushkin)

Create permanent physical replication slots on all nodes that are allowed to become the leader and use `pg_replication_slot_advance()` function to advance `restart_lsn` for slots on standby nodes.

- Add capability of specifying namespace through `--dcs` argument in `patronictl` (Israel Barth Rubio)

It could be handy if `patronictl` is used without a configuration file.

- Add support for additional parameters in custom bootstrap configuration (Israel Barth Rubio)

Previously it was only possible to add custom arguments to the `command` and now one could list them as a mapping.

### Improvements

- Set `citus.local_hostname` GUC to the same value which is used by Patroni to connect to the Postgres (Alexander Kukushkin)

There are cases when Citus wants to have a connection to the local Postgres. By default it uses `localhost`, which is not always available.

### Bugfixes

- Ignore `synchronous_mode` setting in a standby cluster (Polina Bungina)

Postgres doesn't support cascading synchronous replication and not ignoring `synchronous_mode` was breaking a switchover in a standby cluster.

- Handle `SIGCHLD` for `on_reload` callback (Alexander Kukushkin)

Not doing so results in a zombie process, which is reaped only when the next `on_reload` is executed.

- Handle `AuthOldRevision` error when working with Etcd v3 (Alexander Kukushkin, Kenny Do)

The error is raised if Etcd is configured to use JWT and when the user database in Etcd is updated.

## 20.13 Version 3.1.2

Released 2023-09-26

### Bugfixes

- Fixed bug with `wal_keep_size` checks (Alexander Kukushkin)

The `wal_keep_size` is a GUC that normally has a unit and Patroni was failing to cast its value to `int`. As a result the value of `bootstrap.dcs` was not written to the `/config` key afterwards.

- Detect and resolve inconsistencies between `/sync` key and `synchronous_standby_names` (Alexander Kukushkin)

Normally, Patroni updates `/sync` and `synchronous_standby_names` in a very specific order, but in case of a bug or when someone manually reset `synchronous_standby_names`, Patroni was getting into an inconsistent state. As a result it was possible that the failover happens to an asynchronous node.

- Read GUC's values when joining running Postgres (Alexander Kukushkin)

When restarted in `pause`, Patroni was discarding the `synchronous_standby_names` GUC from the `postgresql.conf`. To solve it and avoid similar issues, Patroni will read GUC's value if it is joining an already running Postgres.

- Silenced annoying warnings when checking for node uniqueness (Alexander Kukushkin)

`WARNING` messages are produced by `urllib3` if Patroni is quickly restarted.

## 20.14 Version 3.1.1

Released 2023-09-20

### Bugfixes

- Reset failsafe state on promote (ChenChangAo)

If switchover/failover happened shortly after failsafe mode had been activated, the newly promoted primary was demoting itself after failsafe becomes inactive.

- Silence useless warnings in `patronictl` (Alexander Kukushkin)

If `patronictl` uses the same `patroni.yaml` file as Patroni and can access `PGDATA` directory it might have been showing annoying warnings about incorrect values in the global configuration.

- Explicitly enable synchronous mode for a corner case (Alexander Kukushkin)

Synchronous mode effectively was never activated if there are no replicas streaming from the primary.

- Fixed bug with 0 integer values validation (Israel Barth Rubio)  
In most cases, it didn't cause any issues, just warnings.
- Don't return logical slots for standby cluster (Alexander Kukushkin)  
Patroni can't create logical replication slots in the standby cluster, thus they should be ignored if they are defined in the global configuration.
- Avoid showing docstring in `patronictl --help` output (Israel Barth Rubio)  
The `click` module needs to get a special hint for that.
- Fixed bug with `kubernetes.standby_leader_label_value` (Alexander Kukushkin)  
This feature effectively never worked.
- Returned cluster system identifier to the `patronictl list` output (Polina Bungina)  
The problem was introduced while implementing the support for Citus, where we need to hide the identifier because it is different for coordinator and all workers.
- Override `write_leader_optime` method in Kubernetes implementation (Alexander Kukushkin)  
The method is supposed to write shutdown LSN to the leader Endpoint/ConfigMap when there are no healthy replicas available to become the new primary.
- Don't start stopped postgres in pause (Alexander Kukushkin)  
Due to a race condition, Patroni was falsely assuming that the standby should be restarted because some recovery parameters (`primary_conninfo` or similar) were changed.
- Fixed bug in `patronictl query` command (Israel Barth Rubio)  
It didn't work when only `-m` argument was provided or when none of `-r` or `-m` were provided.
- Properly treat integer parameters that are used in the command line to start postgres (Polina Bungina)  
If values are supplied as strings and not casted to integer it was resulting in an incorrect calculation of `max_prepared_transactions` based on `max_connections` for Citus clusters.
- Don't rely on `pg_stat_wal_receiver` when deciding on `pg_rewind` (Alexander Kukushkin)  
It could happen that `received_tli` reported by `pg_stat_wal_receiver` is ahead of the actual replayed timeline, while the timeline reported by `DENIFY_SYSTEM` via replication connection is always correct.

## 20.15 Version 3.1.0

Released 2023-08-03

### Breaking changes

- Changed semantic of `restapi.keyfile` and `restapi.certfile` (Alexander Kukushkin)  
Previously Patroni was using `restapi.keyfile` and `restapi.certfile` as client certificates as a fallback if there were no respective configuration parameters in the `ctl` section.

#### Warning

If you enabled client certificates validation (`restapi.verify_client` is set to `required`), you also **must** provide **valid client certificates** in the `ctl.certfile`, `ctl.keyfile`, `ctl.keyfile_password`. If not provided, Patroni will not work correctly.

## New features

- Make Pod role label configurable (Waynerv)

Values could be customized using `kubernetes.leader_label_value`, `kubernetes.follower_label_value` and `kubernetes.standby_leader_label_value` parameters. This feature will be very useful when we change the master role to the primary. You can read more about the feature and migration steps [here](#).

## Improvements

- Various improvements of `patroni --validate-config` (Alexander Kukushkin)

Improved parameter validation for different DCS, `bootstrap.dcs`, `ctl`, `restapi`, and `watchdog` sections.

- Start Postgres not in recovery if it crashed during recovery while Patroni is running (Alexander Kukushkin)

It may reduce recovery time and will help to prevent unnecessary timeline increments.

- Avoid unnecessary updates of `/status` key (Alexander Kukushkin)

When there are no permanent logical slots Patroni was updating the `/status` on every heartbeat loop even when LSN on the primary didn't move forward.

- Don't allow stale primary to win the leader race (Alexander Kukushkin)

If Patroni was hanging during a significant time due to lack of resources it will additionally check that no other nodes promoted Postgres before acquiring the leader lock.

- Implemented visibility of certain PostgreSQL parameters validation (Alexander Kukushkin, Feike Steenberg)

If validation of `max_connections`, `max_wal_senders`, `max_prepared_transactions`, `max_locks_per_transaction`, `max_replication_slots`, or `max_worker_processes` failed Patroni was using some sane default value. Now in addition to that it will also show a warning.

- Set permissions for files and directories created in PGDATA (Alexander Kukushkin)

All files created by Patroni had only owner read/write permissions. This behaviour was breaking backup tools that run under a different user and relying on group read permissions. Now Patroni honors permissions on PGDATA and correctly sets permissions on all directories and files it creates inside PGDATA.

## Bugfixes

- Run `archive_command` through shell (Waynerv)

Patroni might archive some WAL segments before doing crash recovery in a single-user mode or before `pg_rewind`. If the `archive_command` contains some shell operators, like `&&` it didn't work with Patroni.

- Fixed “on switchover” shutdown checks (Polina Bungina)

It was possible that specified candidate is still streaming and didn't received shut down checking but the leader key was removed because some other nodes were healthy.

- Fixed “is primary” check (Alexander Kukushkin)

During the leader race replicas were not able to recognize that Postgres on the old leader is still running as a primary.

- Fixed `patronictl list` (Alexander Kukushkin)

The Cluster name field was missing in `tsv`, `json`, and `yaml` output formats.

- Fixed `pg_rewind` behaviour after pause (Alexander Kukushkin)

Under certain conditions, Patroni wasn't able to join the false primary back to the cluster with `pg_rewind` after coming out of maintenance mode.

- Fixed bug in Etcd v3 implementation (Alexander Kukushkin)  
Invalidate internal KV cache if key update performed using `create_revision/mod_revision` field due to revision mismatch.
- Fixed behaviour of replicas in standby cluster in pause (Alexander Kukushkin)  
When the leader key expires replicas in standby cluster will not follow the remote node but keep `primary_conninfo` as it is.

## 20.16 Version 3.0.4

Released 2023-07-13

### New features

- Make the replication status of standby nodes visible (Alexander Kukushkin)  
For PostgreSQL 9.6+ Patroni will report the replication state as `streaming` when the standby is streaming from the other node or in `archive recovery` when there is no replication connection and `restore_command` is set. The state is visible in member keys in DCS, in the REST API, and in `patronictl list` output.

### Improvements

- Improved error messages with Etcd v3 (Alexander Kukushkin)  
When Etcd v3 cluster isn't accessible Patroni was reporting that it can't access `/v2` endpoints.
- Use quorum read in `patronictl` if it is possible (Alexander Kukushkin)  
Etcd or Consul clusters could be degraded to read-only, but from the `patronictl` view everything was fine. Now it will fail with the error.
- Prevent splitbrain from duplicate names in configuration (Mark Pekala)  
When starting Patroni will check if node with the same name is registered in DCS, and try to query its REST API. If REST API is accessible Patroni exits with an error. It will help to protect from the human error.
- Start Postgres not in recovery if it crashed while Patroni is running (Alexander Kukushkin)  
It may reduce recovery time and will help from unnecessary timeline increments.

### Bugfixes

- REST API SSL certificate were not reloaded upon receiving a SIGHUP (Israel Barth Rubio)  
Regression was introduced in 3.0.3.
- Fixed integer GUCs validation for parameters like `max_connections` (Feike Steenberg)  
Patroni didn't like quoted numeric values. Regression was introduced in 3.0.3.
- Fix issue with `synchronous_mode` (Alexander Kukushkin)  
Execute `txid_current()` with `synchronous_commit=off` so it doesn't accidentally wait for absent synchronous standbys when `synchronous_mode_strict` is enabled.

## 20.17 Version 3.0.3

Released 2023-06-22

### New features

- Compatibility with PostgreSQL 16 beta1 (Alexander Kukushkin)  
Extended GUC's validator rules.
- Make PostgreSQL GUC's validator extensible (Israel Barth Rubio)  
Validator rules are loaded from YAML files located in `patroni/postgresql/available_parameters/` directory. Files are ordered in alphabetical order and applied one after another. It makes possible to have custom validators for non-standard Postgres distributions.
- Added `restapi.request_queue_size` option (Andrey Zhidenkov, Aleksei Sukhov)  
Sets request queue size for TCP socket used by Patroni REST API. Once the queue is full, further requests get a "Connection denied" error. The default value is 5.
- Call `initdb` directly when initializing a new cluster (Matt Baker)  
Previously it was called via `pg_ctl`, what required a special quoting of parameters passed to `initdb`.
- Added before stop hook (Le Duane)  
The hook could be configured via `postgresql.before_stop` and is executed right before `pg_ctl stop`. The exit code doesn't impact shutdown process.
- Added support for custom Postgres binary names (Israel Barth Rubio, Polina Bungina)  
When using a custom Postgres distribution it may be the case that the Postgres binaries are compiled with different names other than the ones used by the community Postgres distribution. Custom binary names could be configured using `postgresql.bin_name.*` and `PATRONI_POSTGRESQL_BIN_*` environment variables.

### Improvements

- Various improvements of `patroni --validate-config` (Polina Bungina)
  - Make `bootstrap.initdb` optional. It is only required for new clusters, but `patroni --validate-config` was complaining if it was missing in the config.
  - Don't error out when `postgresql.bin_dir` is empty or not set. Try to first find Postgres binaries in the default PATH instead.
  - Make `postgresql.authentication.rewind` section optional. If it is missing, Patroni is using the superuser.
- Improved error reporting in `patronictl` (Israel Barth Rubio)  
The `\n` symbol was rendered as it is, instead of the actual newline symbol.

### Bugfixes

- Fixed issue in Citus support (Alexander Kukushkin)  
If the REST API call from the promoted worker to the coordinator failed during switchover it was leaving the given Citus group blocked during indefinite time.
- Allow `etcd3` URL in `-dcs-url` option of `patronictl` (Israel Barth Rubio)  
If users attempted to pass a `etcd3` URL through `-dcs-url` option of `patronictl` they would face an exception.

## 20.18 Version 3.0.2

Released 2023-03-24

### Warning

Version 3.0.2 dropped support of Python older than 3.6.

### New features

- Added sync standby replica status to `/metrics` endpoint (Thomas von Dein, Alexander Kukushkin)  
Before were only reporting `primary/standby_leader/replica`.
- User-friendly handling of `PAGER` in `patronictl` (Israel Barth Rubio)  
It makes pager configurable via `PAGER` environment variable, which overrides default `less` and `more`.
- Make K8s retrieable HTTP status code configurable (Alexander Kukushkin)  
On some managed platforms it is possible to get status code `401 Unauthorized`, which sometimes gets resolved after a few retries.

### Improvements

- Set `hot_standby` to `off` during custom bootstrap only if `recovery_target_action` is set to `promote` (Alexander Kukushkin)  
It was necessary to make `recovery_target_action=pause` work correctly.
- Don't allow `on_reload` callback to kill other callbacks (Alexander Kukushkin)  
`on_start/on_stop/on_role_change` are usually used to add/remove Virtual IP and `on_reload` should not interfere with them.
- Switched to `IMDSFetcher` in `aws` callback example script (Polina Bungina)  
The `IMDSv2` requires a token to work with and the `IMDSFetcher` handles it transparently.

### Bugfixes

- Fixed `patronictl switchover` on Citus cluster running on Kubernetes (Lukáš Lalinský)  
It didn't work for namespaces different from `default`.
- Don't write to `PGDATA` if major version is not known (Alexander Kukushkin)  
If right after the start `PGDATA` was empty (maybe wasn't yet mounted), Patroni was making a false assumption about PostgreSQL version and falsely creating `recovery.conf` file even if the actual major version is `v10+`.
- Fixed bug with Citus metadata after coordinator failover (Alexander Kukushkin)  
The `citus_set_coordinator_host()` call doesn't cause metadata sync and the change was invisible on worker nodes. The issue is solved by switching to `citus_update_node()`.
- Use `etcd` hosts listed in the config file as a fallback when all `etcd` nodes "failed" (Alexander Kukushkin)  
The `etcd` cluster may change topology over time and Patroni tries to follow it. If at some point all nodes became unreachable Patroni will use a combination of nodes from the config plus the last known topology when trying to reconnect.



## 20.19 Version 3.0.1

Released 2023-02-16

### Bugfixes

- Pass proper role name to an `on_role_change` callback script'. (Alexander Kukushkin, Polina Bungina)  
Patroni used to erroneously pass `promoted` role to an `on_role_change` callback script on promotion. The passed role name changed back to `master`. This regression was introduced in 3.0.0.

## 20.20 Version 3.0.0

Released 2023-01-30

This version adds integration with [Citux](#) and makes it possible to survive temporary DCS outages without demoting primary.

### Warning

- Version 3.0.0 is the last release supporting Python 2.7. Upcoming release will drop support of Python versions older than 3.7.
- The RAFT support is deprecated. We will do our best to maintain it, but take neither guarantee nor responsibility for possible issues.
- This version is the first step in getting rid of the “master”, in favor of “primary”. Upgrading to the next major release will work reliably only if you run at least 3.0.0.

### New features

- DCS failsafe mode (Alexander Kukushkin, Polina Bungina)  
If the feature is enabled it will allow Patroni cluster to survive temporary DCS outages. You can find more details in the [documentation](#).
- Citux support (Alexander Kukushkin, Polina Bungina, Jelte Fennema)  
Patroni enables easy deployment and management of [Citux](#) clusters with HA. Please check [here](#) page for more information.

### Improvements

- Suppress recurring errors when dropping unknown but active replication slots (Michael Banck)  
Patroni will still write these logs, but only in `DEBUG`.
- Run only one monitoring query per HA loop (Alexander Kukushkin)  
It wasn't the case if synchronous replication is enabled.
- Keep only latest failed data directory (William Albertus Dembo)  
If bootstrap failed Patroni used to rename `$PGDATA` folder with timestamp suffix. From now on the suffix will be `.failed` and if such folder exists it is removed before renaming.
- Improved check of synchronous replication connections (Alexander Kukushkin)  
When the new host is added to the `synchronous_standby_names` it will be set as synchronous in DCS only when it managed to catch up with the primary in addition to `pg_stat_replication.sync_state = 'sync'`.

### Removed functionality

- Remove `patronictl scaffold` (Alexander Kukushkin)

The only reason for having it was a hacky way of running standby clusters.

## 20.21 Version 2.1.7

Released 2023-01-04

### Bugfixes

- Fixed little incompatibilities with legacy python modules (Alexander Kukushkin)

They prevented from building/running Patroni on Debian buster/Ubuntu bionic.

## 20.22 Version 2.1.6

Released 2022-12-30

### Improvements

- Fix annoying exceptions on ssl socket shutdown (Alexander Kukushkin)

The HAProxy is closing connections as soon as it got the HTTP Status code leaving no time for Patroni to properly shutdown SSL connection.

- Adjust example Dockerfile for arm64 (Polina Bungina)

Remove explicit `amd64` and `x86_64`, don't remove `libnss_files.so.*`.

### Security improvements

- Enforce `search_path=pg_catalog` for non-replication connections (Alexander Kukushkin)

Since Patroni is heavily relying on superuser connections, we want to protect it from the possible attacks carried out using user-defined functions and/or operators in `public` schema with the same name and signature as the corresponding objects in `pg_catalog`. For that, `search_path=pg_catalog` is enforced for all connections created by Patroni (except replication connections).

- Prevent passwords from being recorded in `pg_stat_statements` (Feike Steenberg)

It is achieved by setting `pg_stat_statements.track_utility=off` when creating users.

### Bugfixes

- Declare `proxy_address` as optional (Denis Laxalde)

As it is effectively a non-required option.

- Improve behaviour of the `insecure` option (Alexander Kukushkin)

Ctl's `insecure` option didn't work properly when client certificates were used for REST API requests.

- Take watchdog configuration from `bootstrap.dcs` when the new cluster is bootstrapped (Matt Baker)

Patroni used to initially configure watchdog with defaults when bootstrapping a new cluster rather than taking configuration used to bootstrap the DCS.

- Fix the way file extensions are treated while finding executables in WIN32 (Martín Marqués)

Only add `.exe` to a file name if it has no extension yet.

- Fix Consul TTL setup (Alexander Kukushkin)

We used `t+1/2.0` when setting the value on the `HTTPClient`, but forgot to multiply the current value by 2 in the `class` property. It was resulting in Consul TTL off by twice.

### Removed functionality

- Remove `patronictl configure` (Polina Bungina)

There is no more need for a separate `patronictl` config creation.

## 20.23 Version 2.1.5

Released 2022-11-28

This version enhances compatibility with PostgreSQL 15 and declares Etcd v3 support as production ready. The Patroni on Raft remains in Beta.

### New features

- Improve `patroni --validate-config` (Denis Laxalde)

Exit with code 1 if config is invalid and print errors to stderr.

- Don't drop replication slots in pause (Alexander Kukushkin)

Patroni is automatically creating/removing physical replication slots when members are joining/leaving the cluster. In pause slots will no longer be removed.

- Support the HEAD request method for monitoring endpoints (Robert Cutajar)

If used instead of GET Patroni will return only the HTTP Status Code.

- Support behave tests on Windows (Alexander Kukushkin)

Emulate graceful Patroni shutdown (SIGTERM) on Windows by introduce the new REST API endpoint `POST /sigterm`.

- Introduce `postgresql.proxy_address` (Alexander Kukushkin)

It will be written to the member key in DCS as the `proxy_url` and could be used/useful for service discovery.

### Stability improvements

- Call `pg_replication_slot_advance()` from a thread (Alexander Kukushkin)

On busy clusters with many logical replication slots the `pg_replication_slot_advance()` call was affecting the main HA loop and could result in the member key expiration.

- Archive possibly missing WALs before calling `pg_rewind` on the old primary (Polina Bungina)

If the primary crashed and was down during considerable time, some WAL files could be missing from archive and from the new primary. There is a chance that `pg_rewind` could remove these WAL files from the old primary making it impossible to start it as a standby. By archiving ready WAL files we not only mitigate this problem but in general improving continues archiving experience.

- Ignore 403 errors when trying to create Kubernetes Service (Nick Hudson, Polina Bungina)

Patroni was spamming logs by unsuccessful attempts to create the service, which in fact could already exist.

- Improve liveness probe (Alexander Kukushkin)

The liveness problem will start failing if the heartbeat loop is running longer than `t+1` on the primary or `2*t+1` on the replica. That will allow us to use it as an alternative for `watchdog` on Kubernetes.

- Make sure only sync node tries to grab the lock when switchover (Alexander Kukushkin, Polina Bungina)

Previously there was a slim chance that up-to-date async member could become the leader if the manual switchover was performed without specifying the target.

- Avoid cloning while bootstrap is running (Ants Aasma)

Do not allow a create replica method that does not require a leader to be triggered while the cluster bootstrap is running.

- Compatibility with kazoo-2.9.0 (Alexander Kukushkin)

Depending on python version the `SequentialThreadingHandler.select()` method may raise `TypeError` and `IOError` exceptions if `select()` is called on the closed socket.

- Explicitly shut down SSL connection before socket shutdown (Alexander Kukushkin)

Not doing it resulted in `unexpected eof while reading` errors with OpenSSL 3.0.

- Compatibility with *prettytable* >= 2.2.0 (Alexander Kukushkin)

Due to the internal API changes the cluster name header was shown on the incorrect line.

### Bugfixes

- Handle expired token for Etcd lease\_grant (monsterxx03)

In case of error get the new token and retry request.

- Fix bug in the GET /read-only-sync endpoint (Alexander Kukushkin)

It was introduced in previous release and effectively never worked.

- Handle the case when data dir storage disappeared (Alexander Kukushkin)

Patroni is periodically checking that the PGDATA is there and not empty, but in case of issues with storage the `os.listdir()` is raising the `OSError` exception, breaking the heart-beat loop.

- Apply `master_stop_timeout` when waiting for user backends to close (Alexander Kukushkin)

Something that looks like user backend could be in fact a background worker (e.g., Citus Maintenance Daemon) that is failing to stop.

- Accept `*:<port>` for `postgresql.listen` (Denis Laxalde)

The `patroni --validate-config` was complaining about it being invalid.

- Timeouts fixes in Raft (Alexander Kukushkin)

When Patroni or `patronictl` are starting they try to get Raft cluster topology from known members. These calls were made without proper timeouts.

- Forcefully update consul service if token was changed (John A. Lotoski)

Not doing so results in errors “rpc error making call: rpc error making call: ACL not found”.

## 20.24 Version 2.1.4

Released 2022-06-01

### New features

- Improve `pg_rewind` behavior on typical Debian/Ubuntu systems (Gunnar “Nick” Bluth)  
On Postgres setups that keep `postgresql.conf` outside of the data directory (e.g. Ubuntu/Debian packages), `pg_rewind --restore-target-wal` fails to figure out the value of the `restore_command`.
- Allow setting `TLSServerName` on Consul service checks (Michael Gmelin)  
Useful when checks are performed by IP and the Consul `node_name` is not a FQDN.
- Added `ppc64le` support in watchdog (Jean-Michel Scheiwiler)  
And fixed watchdog support on some non-x86 platforms.
- Switched `aws.py` callback from `boto` to `boto3` (Alexander Kukushkin)  
`boto 2.x` is abandoned since 2018 and fails with python 3.9.
- Periodically refresh service account token on K8s (Haitao Li)  
Since Kubernetes v1.21 service account tokens expire in 1 hour.
- Added `/read-only-sync` monitoring endpoint (Dennis4b)  
It is similar to the `/read-only` but includes only synchronous replicas.

### Stability improvements

- Don’t copy the logical replication slot to a replica if there is a configuration mismatch in the logical decoding setup with the primary (Alexander Kukushkin)  
A replica won’t copy a logical replication slot from the primary anymore if the slot doesn’t match the `plugin` or `database` configuration options. Previously, the check for whether the slot matches those configuration options was not performed until after the replica copied the slot and started with it, resulting in unnecessary and repeated restarts.
- Special handling of recovery configuration parameters for PostgreSQL v12+ (Alexander Kukushkin)  
While starting as replica Patroni should be able to update `postgresql.conf` and restart/reload if the leader address has changed by caching current parameters values instead of querying them from `pg_settings`.
- Better handling of IPv6 addresses in the `postgresql.listen` parameters (Alexander Kukushkin)  
Since the `listen` parameter has a port, people try to put IPv6 addresses into square brackets, which were not correctly stripped when there is more than one IP in the list.
- Use `replication` credentials when performing divergence check only on PostgreSQL v10 and older (Alexander Kukushkin)  
If `rewind` is enabled, Patroni will again use either `superuser` or `rewind` credentials on newer Postgres versions.

### Bugfixes

- Fixed missing import of `dateutil.parser` (Wesley Mendes)  
Tests weren’t failing only because it was also imported from other modules.
- Ensure that `optime` annotation is a string (Sebastian Hasler)  
In certain cases Patroni was trying to pass it as numeric.

- Better handling of failed `pg_rewind` attempt (Alexander Kukushkin)

If the primary becomes unavailable during `pg_rewind`, `$PGDATA` will be left in a broken state. Following that, Patroni will remove the data directory even if this is not allowed by the configuration.

- Don't remove slots annotations from the leader `ConfigMap/Endpoint` when PostgreSQL isn't ready (Alexander Kukushkin)

If slots value isn't passed the annotation will keep the current value.

- Handle concurrency problem with K8s API watchers (Alexander Kukushkin)

Under certain (unknown) conditions watchers might become stale; as a result, `attempt_to_acquire_leader()` method could fail due to the HTTP status code 409. In that case we reset watchers connections and restart from scratch.

## 20.25 Version 2.1.3

Released 2022-02-18

### New features

- Added support for encrypted TLS keys for `patronictl` (Alexander Kukushkin)

It could be configured via `ctl.keyfile_password` or the `PATRONI_CTL_KEYFILE_PASSWORD` environment variable.

- Added more metrics to the `/metrics` endpoint (Alexandre Pereira)

Specifically, `patroni_pending_restart` and `patroni_is_paused`.

- Make it possible to specify multiple hosts in the standby cluster configuration (Michael Banck)

If the standby cluster is replicating from the Patroni cluster it might be nice to rely on client-side failover which is available in `libpq` since PostgreSQL v10. That is, the `primary_conninfo` on the standby leader and `pg_rewind` setting `target_session_attrs=read-write` in the connection string. The `pgpass` file will be generated with multiple lines (one line per host), and instead of calling `CHECKPOINT` on the primary cluster nodes the standby cluster will wait for `pg_control` to be updated.

### Stability improvements

- Compatibility with legacy `psycpg2` (Alexander Kukushkin)

For example, the `psycpg2` installed from Ubuntu 18.04 packages doesn't have the `UndefinedFile` exception yet.

- Restart `etcd3` watcher if all `Etcd` nodes don't respond (Alexander Kukushkin)

If the watcher is alive the `get_cluster()` method continues returning stale information even if all `Etcd` nodes are failing.

- Don't remove the leader lock in the standby cluster while paused (Alexander Kukushkin)

Previously the lock was maintained only by the node that was running as a primary and not a standby leader.

### Bugfixes

- Fixed bug in the standby-leader bootstrap (Alexander Kukushkin)

Patroni was considering bootstrap as failed if Postgres didn't start accepting connections after 60 seconds. The bug was introduced in the 2.1.2 release.

- Fixed bug with failover to a cascading standby (Alexander Kukushkin)

When figuring out which slots should be created on cascading standby we forgot to take into account that the leader might be absent.

- Fixed small issues in Postgres config validator (Alexander Kukushkin)

Integer parameters introduced in PostgreSQL v14 were failing to validate because min and max values were quoted in the validator.py

- Use replication credentials when checking leader status (Alexander Kukushkin)

It could be that the `remove_data_directory_on_diverged_timelines` is set, but there is no `rewind_credentials` defined and superuser access between nodes is not allowed.

- Fixed “port in use” error on REST API certificate replacement (Ants Aasma)

When switching certificates there was a race condition with a concurrent API request. If there is one active during the replacement period then the replacement will error out with a port in use error and Patroni gets stuck in a state without an active API server.

- Fixed a bug in cluster bootstrap if passwords contain % characters (Bastien Wirtz)

The bootstrap method executes the DO block, with all parameters properly quoted, but the `cursor.execute()` method didn’t like an empty list with parameters passed.

- Fixed the “AttributeError: no attribute ‘leader’” exception (Hrvoje Milković)

It could happen if the synchronous mode is enabled and the DCS content was wiped out.

- Fix bug in divergence timeline check (Alexander Kukushkin)

Patroni was falsely assuming that timelines have diverged. For `pg_rewind` it didn’t create any problem, but if `pg_rewind` is not allowed and the `remove_data_directory_on_diverged_timelines` is set, it resulted in reinitializing the former leader.

## 20.26 Version 2.1.2

Released 2021-12-03

### New features

- Compatibility with `psycopg>=3.0` (Alexander Kukushkin)

By default `psycopg2` is preferred. `psycopg>=3.0` will be used only if `psycopg2` is not available or its version is too old.

- Add `dcslastseen` field to the REST API (Michael Banck)

This field notes the last time (as unix epoch) a cluster member has successfully communicated with the DCS. This is useful to identify and/or analyze network partitions.

- Release the leader lock when `pg_controldata` reports “shut down” (Alexander Kukushkin)

To solve the problem of slow switchover/shutdown in case `archive_command` is slow/failing, Patroni will remove the leader key immediately after `pg_controldata` started reporting PGDATA as `shut down` cleanly and it verified that there is at least one replica that received all changes. If there are no replicas that fulfill this condition the leader key is not removed and the old behavior is retained, i.e. Patroni will keep updating the lock.

- Add `sslcrlidir` connection parameter support (Kostiantyn Nemchenko)

The new connection parameter was introduced in the PostgreSQL v14.

- Allow setting ACLs for ZNodes in Zookeeper (Alwyn Davis)

Introduce a new configuration option `zookeeper.set_acls` so that Kazoo will apply a default ACL for each ZNode that it creates.

## Stability improvements

- Delay the next attempt of recovery till next HA loop (Alexander Kukushkin)

If Postgres crashed due to out of disk space (for example) and fails to start because of that Patroni is too eagerly trying to recover it flooding logs.

- Add log before demoting, which can take some time (Michael Banck)

It can take some time for the demote to finish and it might not be obvious from looking at the logs what exactly is going on.

- Improve “I am” status messages (Michael Banck)

no action. I am a secondary ( $\{0\}$ ) vs no action. I am ( $\{0\}$ ), a secondary

- Cast to `int wal_keep_segments` when converting to `wal_keep_size` (Jorge Solórzano)

It is possible to specify `wal_keep_segments` as a string in the global *[dynamic configuration](#)* and due to Python being a dynamically typed language the string was simply multiplied. Example: `wal_keep_segments: "100"` was converted to `100100100100100100100100100100100100100100100100100100100MB`.

- Allow switchover only to sync nodes when synchronous replication is enabled (Alexander Kukushkin)

In addition to that do the leader race only against known synchronous nodes.

- Use cached role as a fallback when Postgres is slow (Alexander Kukushkin)

In some extreme cases Postgres could be so slow that the normal monitoring query does not finish in a few seconds. The `statement_timeout` exception not being properly handled could lead to the situation where Postgres was not demoted on time when the leader key expired or the update failed. In case of such exception Patroni will use the cached `role` to determine whether Postgres is running as a primary.

- Avoid unnecessary updates of the member ZNode (Alexander Kukushkin)

If no values have changed in the members data, the update should not happen.

- Optimize checkpoint after promote (Alexander Kukushkin)

Avoid doing `CHECKPOINT` if the latest timeline is already stored in `pg_control`. It helps to avoid unnecessary `CHECKPOINT` right after initializing the new cluster with `initdb`.

- Prefer members without `nofailover` when picking sync nodes (Alexander Kukushkin)

Previously sync nodes were selected only based on the replication lag, hence the node with `nofailover` tag had the same chances to become synchronous as any other node. That behavior was confusing and dangerous at the same time because in case of a failed primary the failover could not happen automatically.

- Remove duplicate hosts from the etcd machine cache (Michael Banck)

Advertised client URLs in the etcd cluster could be misconfigured. Removing duplicates in Patroni in this case is a low-hanging fruit.

## Bugfixes

- Skip temporary replication slots while doing slot management (Alexander Kukushkin)

Starting from v10 `pg_basebackup` creates a temporary replication slot for WAL streaming and Patroni was trying to drop it because the slot name looks unknown. In order to fix it, we skip all temporary slots when querying `pg_stat_replication_slots` view.



- Ensure `pg_replication_slot_advance()` doesn't timeout (Alexander Kukushkin)

Patroni was using the default `statement_timeout` in this case and once the call failed there are very high chances that it will never recover, resulting in increased size of `pg_wal` and `pg_catalog` bloat.

- The `/status` wasn't updated on demote (Alexander Kukushkin)

After demoting PostgreSQL the old leader updates the last LSN in DCS. Starting from 2.1.0 the new `/status` key was introduced, but the `optime` was still written to the `/optime/leader`.

- Handle DCS exceptions when demoting (Alexander Kukushkin)

While demoting the master due to failure to update the leader lock it could happen that DCS goes completely down and the `get_cluster()` call raises an exception. Not being handled properly it results in Postgres remaining stopped until DCS recovers.

- The `use_unix_socket_repl` didn't work in some cases (Alexander Kukushkin)

Specifically, if `postgresql.unix_socket_directories` is not set. In this case Patroni is supposed to use the default value from `libpq`.

- Fix a few issues with Patroni REST API (Alexander Kukushkin)

The `clusters_unlocked` sometimes could be not defined, what resulted in exceptions in the `GET /metrics` endpoint. In addition to that the error handling method was assuming that the `connect_address` tuple always has two elements, while in fact there could be more in case of IPv6.

- Wait for newly promoted node to finish recovery before deciding to rewind (Alexander Kukushkin)

It could take some time before the actual promote happens and the new timeline is created. Without waiting replicas could come to the conclusion that rewind isn't required.

- Handle missing timelines in a history file when deciding to rewind (Alexander Kukushkin)

If the current replica timeline is missing in the history file on the primary the replica was falsely assuming that rewind isn't required.

## 20.27 Version 2.1.1

Released 2021-08-19

### New features

- Support for ETCD SRV name suffix (David Pavlicek)

Etd allows to differentiate between multiple Etd clusters under the same domain and from now on Patroni also supports it.

- Enrich history with the new leader (huiyalin525)

It adds the new column to the `patronictl history` output.

- Make the CA bundle configurable for in-cluster Kubernetes config (Aron Parsons)

By default Patroni is using `/var/run/secrets/kubernetes.io/serviceaccount/ca.crt` and this new feature allows specifying the custom `kubernetes.cacert`.

- Support dynamically registering/deregistering as a Consul service and changing tags (Tommy Li)

Previously it required Patroni restart.

### Bugfixes

- Avoid unnecessary reload of REST API (Alexander Kukushkin)

The previous release added a feature of reloading REST API certificates if changed on disk. Unfortunately, the reload was happening unconditionally right after the start.

- Don't resolve cluster members when `etcd.use_proxies` is set (Alexander Kukushkin)

When starting up Patroni checks the healthiness of Etcd cluster by querying the list of members. In addition to that, it also tried to resolve their hostnames, which is not necessary when working with Etcd via proxy and was causing unnecessary warnings.

- Skip rows with NULL values in the `pg_stat_replication` (Alexander Kukushkin)

It seems that the `pg_stat_replication` view could contain NULL values in the `replay_lsn`, `flush_lsn`, or `write_lsn` fields even when `state = 'streaming'`.

## 20.28 Version 2.1.0

Released 2021-07-06

This version adds compatibility with PostgreSQL v14, makes logical replication slots to survive failover/switchover, implements support of allowlist for REST API, and also reducing the number of logs to one line per heart-beat.

### New features

- Compatibility with PostgreSQL v14 (Alexander Kukushkin)

Unpause WAL replay if Patroni is not in a “pause” mode itself. It could be “paused” due to the change of certain parameters like for example `max_connections` on the primary.

- Failover logical slots (Alexander Kukushkin)

Make logical replication slots survive failover/switchover on PostgreSQL v11+. The replication slot if copied from the primary to the replica with restart and later the `pg_replication_slot_advance()` function is used to move it forward. As a result, the slot will already exist before the failover and no events should be lost, but, there is a chance that some events could be delivered more than once.

- Implemented allowlist for Patroni REST API (Alexander Kukushkin)

If configured, only IP's that matching rules would be allowed to call unsafe endpoints. In addition to that, it is possible to automatically include IP's of members of the cluster to the list.

- Added support of replication connections via unix socket (Mohamad El-Rifai)

Previously Patroni was always using TCP for replication connection what could cause some issues with SSL verification. Using unix sockets allows exempt replication user from SSL verification.

- Health check on user-defined tags (Arman Jafari Tehrani)

Along with *predefined tags*: it is possible to specify any number of custom tags that become visible in the `patronictl list` output and in the REST API. From now on it is possible to use custom tags in health checks.

- Added Prometheus `/metrics` endpoint (Mark Mercado, Michael Banck)

The endpoint exposing the same metrics as `/patroni`.

- Reduced chattiness of Patroni logs (Alexander Kukushkin)

When everything goes normal, only one line will be written for every run of HA loop.

### Breaking changes

- The old permanent logical replication slots feature will no longer work with PostgreSQL v10 and older (Alexander Kukushkin)

The strategy of creating the logical slots after performing a promotion can't guaranty that no logical events are lost and therefore disabled.

- The `/leader` endpoint always returns 200 if the node holds the lock (Alexander Kukushkin)

Promoting the standby cluster requires updating load-balancer health checks, which is not very convenient and easy to forget. To solve it, we change the behavior of the `/leader` health check endpoint. It will return 200 without taking into account whether the cluster is normal or the `standby_cluster`.

### Improvements in Raft support

- Reliable support of Raft traffic encryption (Alexander Kukushkin)

Due to the different issues in the `PySyncObj` the encryption support was very unstable

- Handle DNS issues in Raft implementation (Alexander Kukushkin)

If `self_addr` and/or `partner_addrs` are configured using the DNS name instead of IP's the `PySyncObj` was effectively doing resolve only once when the object is created. It was causing problems when the same node was coming back online with a different IP.

### Stability improvements

- Compatibility with `psycpg2-2.9+` (Alexander Kukushkin)

In `psycpg2` the `autocommit = True` is ignored in the `with connection` block, which breaks replication protocol connections.

- Fix excessive HA loop runs with Zookeeper (Alexander Kukushkin)

Update of member ZNodes was causing a chain reaction and resulted in running the HA loops multiple times in a row.

- Reload if REST API certificate is changed on disk (Michael Todorovic)

If the REST API certificate file was updated in place Patroni didn't perform a reload.

- Don't create `pgpass` dir if kerberos auth is used (Kostiantyn Nemchenko)

Kerberos and password authentication are mutually exclusive.

- Fixed little issues with custom bootstrap (Alexander Kukushkin)

Start Postgres with `hot_standby=off` only when we do a PITR and restart it after PITR is done.

### Bugfixes

- Compatibility with `kazoo-2.7+` (Alexander Kukushkin)

Since Patroni is handling retries on its own, it is relying on the old behavior of `kazoo` that requests to a Zookeeper cluster are immediately discarded when there are no connections available.

- Explicitly request the version of Etcd v3 cluster when it is known that we are connecting via proxy (Alexander Kukushkin)

Patroni is working with Etcd v3 cluster via gPRC-gateway and it depending on the cluster version different endpoints (`/v3`, `/v3beta`, or `/v3alpha`) must be used. The version was resolved only together with the cluster topology, but since the latter was never done when connecting via proxy.

## 20.29 Version 2.0.2

Released 2021-02-22

### New features

- Ability to ignore externally managed replication slots (James Coleman)  
Patroni is trying to remove any replication slot which is unknown to it, but there are certainly cases when replication slots should be managed externally. From now on it is possible to configure slots that should not be removed.
- Added support for cipher suite limitation for REST API (Gunnar “Nick” Bluth)  
It could be configured via `restapi.ciphers` or the `PATRONI_RESTAPI_CIPHERS` environment variable.
- Added support for encrypted TLS keys for REST API (Jonathan S. Katz)  
It could be configured via `restapi.keyfile_password` or the `PATRONI_RESTAPI_KEYFILE_PASSWORD` environment variable.
- Constant time comparison of REST API authentication credentials (Alex Brasetvik)  
Use `hmac.compare_digest()` instead of `==`, which is vulnerable to timing attack.
- Choose synchronous nodes based on replication lag (Krishna Sarabu)  
If the replication lag on the synchronous node starts exceeding the configured threshold it could be demoted to asynchronous and/or replaced by the other node. Behaviour is controlled with `maximum_lag_on_syncnode`.

### Stability improvements

- Start postgres with `hot_standby = off` when doing custom bootstrap (Igor Yanchenko)  
During custom bootstrap Patroni is restoring the basebackup, starting Postgres up, and waiting until recovery finishes. Some PostgreSQL parameters on the standby can't be smaller than on the primary and if the new value (restored from WAL) is higher than the configured one, Postgres panics and stops. In order to avoid such behavior we will do custom bootstrap without `hot_standby` mode.
- Warn the user if the required watchdog is not healthy (Nicolas Thauvin)  
When the watchdog device is not writable or missing in required mode, the member cannot be promoted. Added a warning to show the user where to search for this misconfiguration.
- Better verbosity for single-user mode recovery (Alexander Kukushkin)  
If Patroni notices that PostgreSQL wasn't shutdown clearly, in certain cases the crash-recovery is executed by starting Postgres in single-user mode. It could happen that the recovery failed (for example due to the lack of space on disk) but errors were swallowed.
- Added compatibility with `python-consul2` module (Alexander Kukushkin, Wilfried Roset)  
The good old `python-consul` is not maintained since a few years, therefore someone created a fork with new features and bug-fixes.
- Don't use `bypass_api_service` when running `patronictl` (Alexander Kukushkin)  
When a K8s pod is running in a non-default namespace it does not necessarily have enough permissions to query the kubernetes endpoint. In this case Patroni shows the warning and ignores the `bypass_api_service` setting. In case of `patronictl` the warning was a bit annoying.
- Create `raft.data_dir` if it doesn't exist or make sure that it is writable (Mark Mercado)  
Improves user-friendliness and usability.

### Bugfixes

- Don't interrupt restart or promote if lost leader lock in pause (Alexander Kukushkin)

In pause it is allowed to run postgres as primary without lock.

- Fixed issue with `shutdown_request()` in the REST API (Nicolas Limage)

In order to improve handling of SSL connections and delay the handshake until thread is started Patroni overrides a few methods in the `HTTPServer`. The `shutdown_request()` method was forgotten.

- Fixed issue with sleep time when using Zookeeper (Alexander Kukushkin)

There were chances that Patroni was sleeping up to twice longer between running HA code.

- Fixed invalid `os.symlink()` calls when moving data directory after failed bootstrap (Andrew L'Ecuyer)

If the bootstrap failed Patroni is renaming data directory, `pg_wal`, and all tablespaces. After that it updates symlinks so filesystem remains consistent. The symlink creation was failing due to the `src` and `dst` arguments being swapped.

- Fixed bug in the `post_bootstrap()` method (Alexander Kukushkin)

If the superuser password wasn't configured Patroni was failing to call the `post_init` script and therefore the whole bootstrap was failing.

- Fixed an issues with `pg_rewind` in the standby cluster (Alexander Kukushkin)

If the superuser name is different from Postgres, the `pg_rewind` in the standby cluster was failing because the connection string didn't contain the database name.

- Exit only if authentication with Etcd v3 explicitly failed (Alexander Kukushkin)

On start Patroni performs discovery of Etcd cluster topology and authenticates if it is necessarily. It could happen that one of etcd servers is not accessible, Patroni was trying to perform authentication on this server and failing instead of retrying with the next node.

- Handle case with `psutil.cmdline()` returning empty list (Alexander Kukushkin)

Zombie processes are still postmasters children, but they don't have `cmdline()`

- Treat `PATRONI_KUBERNETES_USE_ENDPOINTS` environment variable as boolean (Alexander Kukushkin)

Not doing so was making impossible disabling `kubernetes.use_endpoints` via environment.

- Improve handling of concurrent endpoint update errors (Alexander Kukushkin)

Patroni will explicitly query the current endpoint object, verify that the current pod still holds the leader lock and repeat the update.

## 20.30 Version 2.0.1

Released 2020-10-01

### New features

- Use `more` as pager in `patronictl edit-config` if `less` is not available (Pavel Golub)

On Windows it would be the `more.com`. In addition to that, `cdiff` was changed to `ydiff` in `requirements.txt`, but `patronictl` still supports both for compatibility.

- Added support of `raft.bind_addr` and `password` (Alexander Kukushkin)

`raft.bind_addr` might be useful when running behind NAT. `raft.password` enables traffic encryption (requires the `cryptography` module).

- Added `sslpassword` connection parameter support (Kostiantyn Nemchenko)

The connection parameter was introduced in PostgreSQL 13.

### Stability improvements

- Changed the behavior in pause (Alexander Kukushkin)
  1. Patroni will not call the `bootstrap` method if the `PGDATA` directory is missing/empty.
  2. Patroni will not exit on `sysid` mismatch in pause, only log a warning.
  3. The node will not try to grab the leader key in pause mode if Postgres is running not in recovery (accepting writes) but the `sysid` doesn't match with the initialize key.

- Apply `master_start_timeout` when executing crash recovery (Alexander Kukushkin)

If Postgres crashed on the leader node, Patroni does a crash-recovery by starting Postgres in single-user mode. During the crash-recovery the leader lock is being updated. If the crash-recovery didn't finish in `master_start_timeout` seconds, Patroni will stop it forcefully and release the leader lock.

- Removed the `secure` extra from the `urllib3` requirements (Alexander Kukushkin)

The only reason for adding it there was the `ipaddress` dependency for python 2.7.

### Bugfixes

- Fixed a bug in the `Kubernetes.update_leader()` (Alexander Kukushkin)

An unhandled exception was preventing demoting the primary when the update of the leader object failed.

- Fixed hanging `patronictl` when RAFT is being used (Alexander Kukushkin)

When using `patronictl` with Patroni config, `self_addr` should be added to the `partner_addrs`.

- Fixed bug in `get_guc_value()` (Alexander Kukushkin)

Patroni was failing to get the value of `restore_command` on PostgreSQL 12, therefore fetching missing WALs for `pg_rewind` didn't work.

## 20.31 Version 2.0.0

Released 2020-09-02

This version enhances compatibility with PostgreSQL 13, adds support of multiple synchronous standbys, has significant improvements in handling of `pg_rewind`, adds support of Etcd v3 and Patroni on pure RAFT (without Etcd, Consul, or Zookeeper), and makes it possible to optionally call the `pre_promote` (fencing) script.

### PostgreSQL 13 support

- Don't fire `on_reload` when promoting to `standby_leader` on PostgreSQL 13+ (Alexander Kukushkin)

When promoting to `standby_leader` we change `primary_conninfo`, update the role and reload Postgres. Since `on_role_change` and `on_reload` effectively duplicate each other, Patroni will call only `on_role_change`.

- Added support for `gssencmode` and `channel_binding` connection parameters (Alexander Kukushkin)

PostgreSQL 12 introduced `gssencmode` and 13 `channel_binding` connection parameters and now they can be used if defined in the `postgresql.authentication` section.

- Handle renaming of `wal_keep_segments` to `wal_keep_size` (Alexander Kukushkin)

In case of misconfiguration (`wal_keep_segments` on 13 and `wal_keep_size` on older versions) Patroni will automatically adjust the configuration.

- Use `pg_rewind` with `--restore-target-wal` on 13 if possible (Alexander Kukushkin)  
On PostgreSQL 13 Patroni checks if `restore_command` is configured and tells `pg_rewind` to use it.

### New features

- [BETA] Implemented support of Patroni on pure RAFT (Alexander Kukushkin)  
This makes it possible to run Patroni without 3rd party dependencies, like Etcd, Consul, or Zookeeper. For HA you will have to run either three Patroni nodes or two nodes with Patroni and one node with `patroni_raft_controller`. For more information please check the [documentation](#).
- [BETA] Implemented support for Etcd v3 protocol via gPRC-gateway (Alexander Kukushkin)  
Etcd 3.0 was released more than four years ago and Etcd 3.4 has v2 disabled by default. There are also chances that v2 will be completely removed from Etcd, therefore we implemented support of Etcd v3 in Patroni. In order to start using it you have to explicitly create the `etcd3` section in the Patroni configuration file.
- Supporting multiple synchronous standbys (Krishna Sarabu)  
It allows running a cluster with more than one synchronous replicas. The maximum number of synchronous replicas is controlled by the new parameter `synchronous_node_count`. It is set to 1 by default and has no effect when the `synchronous_mode` is set to `off`.
- Added possibility to call the `pre_promote` script (Sergey Dudoladov)  
Unlike callbacks, the `pre_promote` script is called synchronously after acquiring the leader lock, but before promoting Postgres. If the script fails or exits with a non-zero exitcode, the current node will release the leader lock.
- Added support for configuration directories (Floris van Nee)  
YAML files in the directory loaded and applied in alphabetical order.
- Advanced validation of PostgreSQL parameters (Alexander Kukushkin)  
In case the specific parameter is not supported by the current PostgreSQL version or when its value is incorrect, Patroni will remove the parameter completely or try to fix the value.
- Wake up the main thread when the forced checkpoint after promote completed (Alexander Kukushkin)  
Replicas are waiting for checkpoint indication via member key of the leader in DCS. The key is normally updated only once per HA loop. Without waking the main thread up, replicas will have to wait up to `loop_wait` seconds longer than necessary.
- Use of `pg_stat_wal_receiver` view on 9.6+ (Alexander Kukushkin)  
The view contains up-to-date values of `primary_conninfo` and `primary_slot_name`, while the contents of `recovery.conf` could be stale.
- Improved handling of IPv6 addresses in the Patroni config file (Mateusz Kowalski)  
The IPv6 address is supposed to be enclosed into square brackets, but Patroni was expecting to get it plain. Now both formats are supported.
- Added Consul `service_tags` configuration parameter (Robert Edström)  
They are useful for dynamic service discovery, for example by load balancers.
- Implemented SSL support for Zookeeper (Kostiantyn Nemchenko)  
It requires `kazoo>=2.6.0`.
- Implemented `no_params` option for custom bootstrap method (Kostiantyn Nemchenko)  
It allows calling `wal-g`, `pgBackRest` and other backup tools without wrapping them into shell scripts.

- Move WAL and tablespaces after a failed init (Feike Steenberg)

When doing `reinit`, Patroni was already removing not only PGDATA but also the symlinked WAL directory and tablespaces. Now the `move_data_directory()` method will do a similar job, i.e. rename WAL directory and tablespaces and update symlinks in PGDATA.

### Improved in `pg_rewind` support

- Improved timeline divergence check (Alexander Kukushkin)

We don't need to rewind when the replayed location on the replica is not ahead of the switchpoint or the end of the checkpoint record on the former primary is the same as the switchpoint. In order to get the end of the checkpoint record we use `pg_waldump` and parse its output.

- Try to fetch missing WAL if `pg_rewind` complains about it (Alexander Kukushkin)

It could happen that the WAL segment required for `pg_rewind` doesn't exist in the `pg_wal` directory anymore and therefore `pg_rewind` can't find the checkpoint location before the divergence point. Starting from PostgreSQL 13 `pg_rewind` could use `restore_command` for fetching missing WALs. For older PostgreSQL versions Patroni parses the errors of a failed rewind attempt and tries to fetch the missing WAL by calling the `restore_command` on its own.

- Detect a new timeline in the standby cluster and trigger rewind/reinitialize if necessary (Alexander Kukushkin)

The `standby_cluster` is decoupled from the primary cluster and therefore doesn't immediately know about leader elections and timeline switches. In order to detect the fact, the `standby_leader` periodically checks for new history files in `pg_wal`.

- Shorten and beautify history log output (Alexander Kukushkin)

When Patroni is trying to figure out the necessity of `pg_rewind`, it could write the content of the history file from the primary into the log. The history file is growing with every failover/switchover and eventually starts taking up too many lines, most of which are not so useful. Instead of showing the raw data, Patroni will show only 3 lines before the current replica timeline and 2 lines after.

### Improvements on K8s

- Get rid of `kubernetes` python module (Alexander Kukushkin)

The official python `kubernetes` client contains a lot of auto-generated code and therefore very heavy. Patroni uses only a small fraction of K8s API endpoints and implementing support for them wasn't hard.

- Make it possible to bypass the `kubernetes` service (Alexander Kukushkin)

When running on K8s, Patroni is usually communicating with the K8s API via the `kubernetes` service, the address of which is exposed in the `KUBERNETES_SERVICE_HOST` environment variable. Like any other service, the `kubernetes` service is handled by `kube-proxy`, which in turn, depending on the configuration, is either relying on a userspace program or `iptables` for traffic routing. Skipping the intermediate component and connecting directly to the K8s master nodes allows us to implement a better retry strategy and mitigate risks of demoting Postgres when K8s master nodes are upgraded.

- Sync HA loops of all pods of a Patroni cluster (Alexander Kukushkin)

Not doing so was increasing failure detection time from `t1` to `t1 + loop_wait`.

- Populate `references` and `nodename` in the subsets addresses on K8s (Alexander Kukushkin)

Some load-balancers are relying on this information.

- Fix possible race conditions in the `update_leader()` (Alexander Kukushkin)

The concurrent update of the leader configmap or endpoint happening outside of Patroni might cause the `update_leader()` call to fail. In this case Patroni rechecks that the current node is still owning the leader lock and repeats the update.



- Explicitly disallow patching non-existent config (Alexander Kukushkin)

For DCS other than `kubernetes` the `PATCH` call is failing with an exception due to `cluster.config` being `None`, but on Kubernetes it was happily creating the config annotation and preventing writing bootstrap configuration after the bootstrap finished.

- Fix bug in `pause` (Alexander Kukushkin)

Replicas were removing `primary_conninfo` and restarting Postgres when the leader key was absent, but they should do nothing.

### Improvements in REST API

- Defer TLS handshake until worker thread has started (Alexander Kukushkin, Ben Harris)

If the TLS handshake was done in the API thread and the client-side didn't send any data, the API thread was blocked (risking DoS).

- Check `basic-auth` independently from client certificate in REST API (Alexander Kukushkin)

Previously only the client certificate was validated. Doing two checks independently is an absolutely valid use-case.

- Write double CRLF after HTTP headers of the `OPTIONS` request (Sergey Burladyan)

HAProxy was happy with a single CRLF, while Consul health-check complained about broken connection and unexpected EOF.

- `GET /cluster` was showing stale members info for Zookeeper (Alexander Kukushkin)

The endpoint was using the Patroni internal cluster view. For Patroni itself it didn't cause any issues, but when exposed to the outside world we need to show up-to-date information, especially replication lag.

- Fixed health-checks for standby cluster (Alexander Kukushkin)

The `GET /standby-leader` for a master and `GET /master` for a `standby_leader` were incorrectly responding with 200.

- Implemented `DELETE /switchover` (Alexander Kukushkin)

The REST API call deletes the scheduled switchover.

- Created `/readiness` and `/liveness` endpoints (Alexander Kukushkin)

They could be useful to eliminate “unhealthy” pods from subsets addresses when the K8s service is used with label selectors.

- Enhanced `GET /replica` and `GET /async` REST API health-checks (Krishna Sarabu, Alexander Kukushkin)

Checks now support optional keyword `?lag=<max-lag>` and will respond with 200 only if the lag is smaller than the supplied value. If relying on this feature please keep in mind that information about WAL position on the leader is updated only every `loop_wait` seconds!

- Added support for user defined HTTP headers in the REST API response (Yogesh Sharma)

This feature might be useful if requests are made from a browser.

### Improvements in `patronictl`

- Don't try to call non-existing leader in `patronictl pause` (Alexander Kukushkin)

While pausing a cluster without a leader on K8s, `patronictl` was showing warnings that member “None” could not be accessed.

- Handle the case when member `conn_url` is missing (Alexander Kukushkin)

On K8s it is possible that the pod doesn't have the necessary annotations because Patroni is not yet running. It was making `patronictl` to fail.

- Added ability to print ASCII cluster topology (Maxim Fedotov, Alexander Kukushkin)

It is very useful to get overview of the cluster with cascading replication.

- Implement `patronictl flush switchover` (Alexander Kukushkin)

Before that `patronictl flush` only supported cancelling scheduled restarts.

### Bugfixes

- Attribute error during bootstrap of the cluster with existing PGDATA (Krishna Sarabu)

When trying to create/update the `/history` key, Patroni was accessing the `ClusterConfig` object which wasn't created in DCS yet.

- Improved exception handling in Consul (Alexander Kukushkin)

Unhandled exception in the `touch_member()` method caused the whole Patroni process to crash.

- Enforce `synchronous_commit=local` for the `post_init` script (Alexander Kukushkin)

Patroni was already doing that when creating users (`replication`, `rewind`), but missing it in the case of `post_init` was an oversight. As a result, if the script wasn't doing it internally on it's own the bootstrap in `synchronous_mode` wasn't able to finish.

- Increased `maxsize` in the Consul pool manager (ponvenkates)

With the default `size=1` some warnings were generated.

- Patroni was wrongly reporting Postgres as running (Alexander Kukushkin)

The state wasn't updated when for example Postgres crashed due to an out-of-disk error.

- Put `*` into `pgpass` instead of missing or empty values (Alexander Kukushkin)

If for example the `standby_cluster.port` is not specified, the `pgpass` file was incorrectly generated.

- Skip physical replication slot creation on the leader node with special characters (Krishna Sarabu)

Patroni appeared to be creating a dormant slot (when `slots` defined) for the leader node when the name contained special chars such as `'-'` (for e.g. `"abc-us-1"`).

- Avoid removing non-existent `pg_hba.conf` in the custom bootstrap (Krishna Sarabu)

Patroni was failing if `pg_hba.conf` happened to be located outside of the `pgdata` dir after custom bootstrap.

## 20.32 Version 1.6.5

Released 2020-08-23

### New features

- Master stop timeout (Krishna Sarabu)

The number of seconds Patroni is allowed to wait when stopping Postgres. Effective only when `synchronous_mode` is enabled. When set to value greater than 0 and the `synchronous_mode` is enabled, Patroni sends `SIGKILL` to the postmaster if the stop operation is running for more than the value set by `master_stop_timeout`. Set the value according to your durability/availability tradeoff. If the parameter is not set or set to non-positive value, `master_stop_timeout` does not have an effect.

- Don't create permanent physical slot with name of the primary (Alexander Kukushkin)

It is a common problem that the primary recycles WAL segments while the replica is down. Now we have a good solution for static clusters, with a fixed number of nodes and names that never change. You just need to list the names of all nodes in the slots so the primary will not remove the slot when the node is down (not registered in DCS).

- First draft of Config Validator (Igor Yanchenko)

Use `patroni --validate-config patroni.yaml` in order to validate Patroni configuration.

- Possibility to configure max length of timelines history (Krishna Sarabu)

Patroni writes the history of failovers/switchovers into the `/history` key in DCS. Over time the size of this key becomes big, but in most cases only the last few lines are interesting. The `max_timelines_history` parameter allows to specify the maximum number of timeline history items to be kept in DCS.

- Kazoo 2.7.0 compatibility (Danyal Prout)

Some non-public methods in Kazoo changed their signatures, but Patroni was relying on them.

### Improvements in patronictl

- Show member tags (Kostiantyn Nemchenko, Alexander Kukushkin)

Tags are configured individually for every node and there was no easy way to get an overview of them

- Improve members output (Alexander Kukushkin)

The redundant cluster name won't be shown anymore on every line, only in the table header.

```
$ patronictl list
+ Cluster: batman (6813309862653668387) +-----+-----+-----+-----+-----+-----+-----+
↪+
| Member | Host | Role | State | TL | Lag in MB | Tags |
↪|
+-----+-----+-----+-----+-----+-----+-----+
↪+
| postgresql0 | 127.0.0.1:5432 | Leader | running | 3 | | clonefrom: true |
↪|
| | | | | | | noloadbalance: true |
↪|
| | | | | | | nosync: true |
↪|
+-----+-----+-----+-----+-----+-----+-----+
↪+
| postgresql1 | 127.0.0.1:5433 | | running | 3 | 0.0 | |
↪|
+-----+-----+-----+-----+-----+-----+-----+
↪+
```

- Fail if a config file is specified explicitly but not found (Kaarel Moppel)

Previously `patronictl` was only reporting a `DEBUG` message.

- Solved the problem of not initialized K8s pod breaking `patronictl` (Alexander Kukushkin)

Patroni is relying on certain pod annotations on K8s. When one of the Patroni pods is stopping or starting there is no valid annotation yet and `patronictl` was failing with an exception.

### Stability improvements

- Apply 1 second backoff if LIST call to K8s API server failed (Alexander Kukushkin)  
It is mostly necessary to avoid flooding logs, but also helps to prevent starvation of the main thread.
- Retry if the `retry-after` HTTP header is returned by K8s API (Alexander Kukushkin)  
If the K8s API server is overwhelmed with requests it might ask to retry.
- Scrub KUBERNETES\_ environment from the postmaster (Feike Steenberg)  
The KUBERNETES\_ environment variables are not required for PostgreSQL, yet having them exposed to the postmaster will also expose them to backends and to regular database users (using pl/perl for example).
- Clean up tablespaces on reinitialize (Krishna Sarabu)  
During reinit, Patroni was removing only PGDATA and leaving user-defined tablespace directories. This is causing Patroni to loop in reinit. The previous workaround for the problem was implementing the *custom bootstrap* script.
- Explicitly execute CHECKPOINT after promote happened (Alexander Kukushkin)  
It helps to reduce the time before the new primary is usable for `pg_rewind`.
- Smart refresh of Etcd members (Alexander Kukushkin)  
In case Patroni failed to execute a request on all members of the Etcd cluster, Patroni will re-check A or SRV records for changes of IPs/hosts before retrying the next time.
- Skip missing values from `pg_controldata` (Feike Steenberg)  
Values are missing when trying to use binaries of a version that doesn't match PGDATA. Patroni will try to start Postgres anyway, and Postgres will complain that the major version doesn't match and abort with an error.

### Bugfixes

- Disable SSL verification for Consul when required (Julien Riou)  
Starting from a certain version of `urllib3`, the `cert_reqs` must be explicitly set to `ssl.CERT_NONE` in order to effectively disable SSL verification.
- Avoid opening replication connection on every cycle of HA loop (Alexander Kukushkin)  
Regression was introduced in 1.6.4.
- Call `on_role_change` callback on failed primary (Alexander Kukushkin)  
In certain cases it could lead to the virtual IP remaining attached to the old primary. Regression was introduced in 1.4.5.
- Reset rewind state if postgres started after successful `pg_rewind` (Alexander Kukushkin)  
As a result of this bug Patroni was starting up manually shut down postgres in the pause mode.
- Convert `recovery_min_apply_delay` to ms when checking `recovery.conf`  
Patroni was indefinitely restarting replica if `recovery_min_apply_delay` was configured on PostgreSQL older than 12.
- PyInstaller compatibility (Alexander Kukushkin)  
PyInstaller freezes (packages) Python applications into stand-alone executables. The compatibility was broken when we switched to the `spawn` method instead of `fork` for multiprocessing.

## 20.33 Version 1.6.4

Released 2020-01-27

### New features

- Implemented `--wait` option for `patronictl reinit` (Igor Yanchenko)  
Patronictl will wait for `reinit` to finish if the `--wait` option is used.
- Further improvements of Windows support (Igor Yanchenko, Alexander Kukushkin)
  1. All shell scripts which are used for integration testing are rewritten in python
  2. The `pg_ctl kill` will be used to stop postgres on non posix systems
  3. Don't try to use unix-domain sockets

### Stability improvements

- Make sure `unix_socket_directories` and `stats_temp_directory` exist (Igor Yanchenko)  
Upon the start of Patroni and Postgres make sure that `unix_socket_directories` and `stats_temp_directory` exist or try to create them. Patroni will exit if failed to create them.
- Make sure `postgresql.pgpass` is located in the place where Patroni has write access (Igor Yanchenko)  
In case if it doesn't have a write access Patroni will exit with exception.
- Disable Consul `serfHealth` check by default (Kostiantyn Nemchenko)  
Even in case of little network problems the failing `serfHealth` leads to invalidation of all sessions associated with the node. Therefore, the leader key is lost much earlier than `t1` which causes unwanted restarts of replicas and maybe demotion of the primary.
- Configure tcp keepalives for connections to K8s API (Alexander Kukushkin)  
In case if we get nothing from the socket after TTL seconds it can be considered dead.
- Avoid logging of passwords on user creation (Alexander Kukushkin)  
If the password is rejected or logging is configured to verbose or not configured at all it might happen that the password is written into postgres logs. In order to avoid it Patroni will change `log_statement`, `log_min_duration_statement`, and `log_min_error_statement` to some safe values before doing the attempt to create/update user.

### Bugfixes

- Use `restore_command` from the `standby_cluster` config on cascading replicas (Alexander Kukushkin)  
The `standby_leader` was already doing it from the beginning the feature existed. Not doing the same on replicas might prevent them from catching up with standby leader.
- Update timeline reported by the standby cluster (Alexander Kukushkin)  
In case of timeline switch the standby cluster was correctly replicating from the primary but `patronictl` was reporting the old timeline.
- Allow certain recovery parameters be defined in the `custom_conf` (Alexander Kukushkin)  
When doing validation of recovery parameters on replica Patroni will skip `archive_cleanup_command`, `promote_trigger_file`, `recovery_end_command`, `recovery_min_apply_delay`, and `restore_command` if they are not defined in the patroni config but in files other than `postgresql.auto.conf` or `postgresql.conf`.

- Improve handling of postgresql parameters with period in its name (Alexander Kukushkin)

Such parameters could be defined by extensions where the unit is not necessarily a string. Changing the value might require a restart (for example `pg_stat_statements.max`).

- Improve exception handling during shutdown (Alexander Kukushkin)

During shutdown Patroni is trying to update its status in the DCS. If the DCS is inaccessible an exception might be raised. Lack of exception handling was preventing logger thread from stopping.

## 20.34 Version 1.6.3

Released 2019-12-05

### Bugfixes

- Don't expose password when running `pg_rewind` (Alexander Kukushkin)

Bug was introduced in the [#1301](#)

- Apply connection parameters specified in the `postgresql.authentication` to `pg_basebackup` and custom replica creation methods (Alexander Kukushkin)

They were relying on url-like connection string and therefore parameters never applied.

## 20.35 Version 1.6.2

Released 2019-12-05

### New features

- Implemented `patroni --version` (Igor Yanchenko)

It prints the current version of Patroni and exits.

- Set the `user-agent` http header for all http requests (Alexander Kukushkin)

Patroni is communicating with Consul, Etcd, and Kubernetes API via the http protocol. Having a specifically crafted `user-agent` (example: `Patroni/1.6.2 Python/3.6.8 Linux`) might be useful for debugging and monitoring.

- Make it possible to configure log level for exception tracebacks (Igor Yanchenko)

If you set `log.traceback_level=DEBUG` the tracebacks will be visible only when `log.level=DEBUG`. The default behavior remains the same.

### Stability improvements

- Avoid importing all DCS modules when searching for the module required by the config file (Alexander Kukushkin)

There is no need to import modules for Etcd, Consul, and Kubernetes if we need only e.g. Zookeeper. It helps to reduce memory usage and solves the problem of having INFO messages `Failed to import smth`.

- Removed python `requests` module from explicit requirements (Alexander Kukushkin)

It wasn't used for anything critical, but causing a lot of problems when the new version of `urllib3` is released.

- Improve handling of `etcd.hosts` written as a comma-separated string instead of YAML array (Igor Yanchenko)

Previously it was failing when written in format `host1:port1, host2:port2` (the space character after the comma).

### Usability improvements

- Don't force users to choose members from an empty list in `patronictl` (Igor Yanchenko)  
If the user provides a wrong cluster name, we will raise an exception rather than ask to choose a member from an empty list.
- Make the error message more helpful if the REST API cannot bind (Igor Yanchenko)  
For an inexperienced user it might be hard to figure out what is wrong from the Python stacktrace.

### Bugfixes

- Fix calculation of `wal_buffers` (Alexander Kukushkin)  
The base unit has been changed from 8 kB blocks to bytes in PostgreSQL 11.
- Use `passfile` in `primary_conninfo` only on PostgreSQL 10+ (Alexander Kukushkin)  
On older versions there is no guarantee that `passfile` will work, unless the latest version of `libpq` is installed.

## 20.36 Version 1.6.1

Released 2019-11-15

### New features

- Added `PATRONICTL_CONFIG_FILE` environment variable (msvechla)  
It allows configuring the `--config-file` argument for `patronictl` from the environment.
- Implement `patronictl history` (Alexander Kukushkin)  
It shows the history of failovers/switchovers.
- Pass `-c statement_timeout=0` in `PGOPTIONS` when doing `pg_rewind` (Alexander Kukushkin)  
It protects from the case when `statement_timeout` on the server is set to some small value and one of the statements executed by `pg_rewind` is canceled.
- Allow lower values for PostgreSQL configuration (Soulou)  
Patroni didn't allow some of the PostgreSQL configuration parameters be set smaller than some hardcoded values. Now the minimal allowed values are smaller, default values have not been changed.
- Allow for certificate-based authentication (Jonathan S. Katz)  
This feature enables certificate-based authentication for superuser, replication, rewind accounts and allows the user to specify the `sslmode` they wish to connect with.
- Use the `passfile` in the `primary_conninfo` instead of password (Alexander Kukushkin)  
It allows to avoid setting `600` permissions on `postgresql.conf`
- Perform `pg_ctl reload` regardless of config changes (Alexander Kukushkin)  
It is possible that some config files are not controlled by Patroni. When somebody is doing a reload via the REST API or by sending `SIGHUP` to the Patroni process, the usual expectation is that Postgres will also be reloaded. Previously it didn't happen when there were no changes in the `postgresql` section of Patroni config.
- Compare all recovery parameters, not only `primary_conninfo` (Alexander Kukushkin)  
Previously the `check_recovery_conf()` method was only checking whether `primary_conninfo` has changed, never taking into account all other recovery parameters.

- Make it possible to apply some recovery parameters without restart (Alexander Kukushkin)

Starting from PostgreSQL 12 the following recovery parameters could be changed without restart: `archive_cleanup_command`, `promote_trigger_file`, `recovery_end_command`, and `recovery_min_apply_delay`. In future Postgres releases this list will be extended and Patroni will support it automatically.

- Make it possible to change `use_slots` online (Alexander Kukushkin)

Previously it required restarting Patroni and removing slots manually.

- Remove only `PATRONI_` prefixed environment variables when starting up Postgres (Cody Coons)

It will solve a lot of problems with running different Foreign Data Wrappers.

### Stability improvements

- Use LIST + WATCH when working with K8s API (Alexander Kukushkin)

It allows to efficiently receive object changes (pods, endpoints/configmaps) and makes less stress on K8s master nodes.

- Improve the workflow when PGDATA is not empty during bootstrap (Alexander Kukushkin)

According to the `initdb` source code it might consider a PGDATA empty when there are only `lost+found` and `.dotfiles` in it. Now Patroni does the same. If PGDATA happens to be non-empty, and at the same time not valid from the `pg_controldata` point of view, Patroni will complain and exit.

- Avoid calling expensive `os.listdir()` on every HA loop (Alexander Kukushkin)

When the system is under IO stress, `os.listdir()` could take a few seconds (or even minutes) to execute, badly affecting the HA loop of Patroni. This could even cause the leader key to disappear from DCS due to the lack of updates. There is a better and less expensive way to check that the PGDATA is not empty. Now we check the presence of the `global/pg_control` file in the PGDATA.

- Some improvements in logging infrastructure (Alexander Kukushkin)

Previously there was a possibility to loose the last few log lines on shutdown because the logging thread was a daemon thread.

- Use `spawn` multiprocessing start method on python 3.4+ (Maciej Kowalczyk)

It is a known [issue](#) in Python that threading and multiprocessing do not mix well. Switching from the default method `fork` to the `spawn` is a recommended workaround. Not doing so might result in the Postmaster starting process hanging and Patroni indefinitely reporting `INFO: restarting after failure in progress`, while Postgres is actually up and running.

### Improvements in REST API

- Make it possible to check client certificates in the REST API (Alexander Kukushkin)

If the `verify_client` is set to `required`, Patroni will check client certificates for all REST API calls. When it is set to `optional`, client certificates are checked for all unsafe REST API endpoints.

- Return the response code 503 for the GET `/replica` health check request if Postgres is not running (Alexander Anikin)

Postgres might spend significant time in recovery before it starts accepting client connections.

- Implement `/history` and `/cluster` endpoints (Alexander Kukushkin)

The `/history` endpoint shows the content of the `history` key in DCS. The `/cluster` endpoint shows all cluster members and some service info like pending and scheduled restarts or switchovers.

### Improvements in Etcd support



- Retry on Etcd RAFT internal error (Alexander Kukushkin)

When the Etcd node is being shut down, it sends `response code=300, data='etcdserver: server stopped'`, which was causing Patroni to demote the primary.

- Don't give up on Etcd request retry too early (Alexander Kukushkin)

When there were some network problems, Patroni was quickly exhausting the list of Etcd nodes and giving up without using the whole `retry_timeout`, potentially resulting in demoting the primary.

### Bugfixes

- Disable `synchronous_commit` when granting execute permissions to the `pg_rewind` user (kremius)

If the bootstrap is done with `synchronous_mode_strict: true` the `GRANT EXECUTE` statement was waiting indefinitely due to the non-synchronous nodes being available.

- Fix memory leak on python 3.7 (Alexander Kukushkin)

Patroni is using `ThreadingMixIn` to process REST API requests and python 3.7 made threads spawn for every request non-daemon by default.

- Fix race conditions in asynchronous actions (Alexander Kukushkin)

There was a chance that `patronictl reinit --force` could be overwritten by the attempt to recover stopped Postgres. This ended up in a situation when Patroni was trying to start Postgres while basebackup was running.

- Fix race condition in `postmaster_start_time()` method (Alexander Kukushkin)

If the method is executed from the REST API thread, it requires a separate cursor object to be created.

- Fix the problem of not promoting the sync standby that had a name containing upper case letters (Alexander Kukushkin)

We converted the name to the lower case because Postgres was doing the same while comparing the `application_name` with the value in `synchronous_standby_names`.

- Kill all children along with the callback process before starting the new one (Alexander Kukushkin)

Not doing so makes it hard to implement callbacks in bash and eventually can lead to the situation when two callbacks are running at the same time.

- Fix 'start failed' issue (Alexander Kukushkin)

Under certain conditions the Postgres state might be set to 'start failed' despite Postgres being up and running.

## 20.37 Version 1.6.0

Released 2019-08-05

This version adds compatibility with PostgreSQL 12, makes it possible to run `pg_rewind` without superuser on PostgreSQL 11 and newer, and enables IPv6 support.

### New features

- Psycopg2 was removed from requirements and must be installed independently (Alexander Kukushkin)

Starting from 2.8.0 `psycopg2` was split into two different packages, `psycopg2`, and `psycopg2-binary`, which could be installed at the same time into the same place on the filesystem. In order to decrease dependency hell problem, we let a user choose how to install it. There are a few options available, please consult the [documentation](#).

- Compatibility with PostgreSQL 12 (Alexander Kukushkin)

Starting from PostgreSQL 12 there is no `recovery.conf` anymore and all former recovery parameters are converted into GUC. In order to protect from `ALTER SYSTEM SET primary_conninfo` or similar, Patroni will parse `postgresql.auto.conf` and remove all standby and recovery parameters from there. Patroni config remains backward compatible. For example despite `restore_command` being a GUC, one can still specify it in the `postgresql.recovery_conf.restore_command` section and Patroni will write it into `postgresql.conf` for PostgreSQL 12.

- Make it possible to use `pg_rewind` without superuser on PostgreSQL 11 and newer (Alexander Kukushkin)

If you want to use this feature please define `username` and `password` in the `postgresql.authentication.rewind` section of Patroni configuration file. For an already existing cluster you will have to create the user manually and `GRANT EXECUTE` permission on a few functions. You can find more details in the [PostgreSQL documentation](#).

- Do a smart comparison of actual and desired `primary_conninfo` values on replicas (Alexander Kukushkin)

It might help to avoid replica restart when you are converting an already existing primary-standby cluster to one managed by Patroni

- IPv6 support (Alexander Kukushkin)

There were two major issues. Patroni REST API service was listening only on `0.0.0.0` and IPv6 IP addresses used in the `api_url` and `conn_url` were not properly quoted.

- Kerberos support (Ajith Vilas, Alexander Kukushkin)

It makes possible using Kerberos authentication between Postgres nodes instead of defining passwords in Patroni configuration file

- Manage `pg_ident.conf` (Alexander Kukushkin)

This functionality works similarly to `pg_hba.conf`: if the `postgresql.pg_ident` is defined in the config file or DCS, Patroni will write its value to `pg_ident.conf`, however, if `postgresql.parameters.ident_file` is defined, Patroni will assume that `pg_ident` is managed from outside and not update the file.

### Improvements in REST API

- Added `/health` endpoint (Wilfried Roset)

It will return an HTTP status code only if PostgreSQL is running

- Added `/read-only` and `/read-write` endpoints (Julien Riou)

The `/read-only` endpoint enables reads balanced across replicas and the primary. The `/read-write` endpoint is an alias for `/primary`, `/leader` and `/master`.

- Use `SSLContext` to wrap the REST API socket (Julien Riou)

Usage of `ssl.wrap_socket()` is deprecated and was still allowing soon-to-be-deprecated protocols like TLS 1.1.

### Logging improvements

- Two-step logging (Alexander Kukushkin)

All log messages are first written into the in-memory queue and later they are asynchronously flushed into the `stderr` or file from a separate thread. The maximum queue size is limited (configurable). If the limit is reached, Patroni will start losing logs, which is still better than blocking the HA loop.

- Enable debug logging for GET/OPTIONS API calls together with latency (Jan Tomsa)

It will help with debugging of health-checks performed by HAProxy, Consul or other tooling that decides which node is the primary/replica.

- Log exceptions caught in Retry (Daniel Kucera)

Log the final exception when either the number of attempts or the timeout were reached. It will hopefully help to debug some issues when communication to DCS fails.

### Improvements in patronictl

- Enhance dialogues for scheduled switchover and restart (Rafia Sabih)

Previously dialogues did not take into account scheduled actions and therefore were misleading.

- Check if config file exists (Wilfried Roset)

Be verbose about configuration file when the given filename does not exist, instead of ignoring silently (which can lead to misunderstanding).

- Add fallback value for EDITOR (Wilfried Roset)

When the EDITOR environment variable was not defined, `patronictl edit-config` was failing with *PatroniCtlException*. The new strategy is to try `editor` and then `vi`, which should be available on most systems.

### Improvements in Consul support

- Allow to specify Consul consistency mode (Jan Tomsa)

You can read more about consistency mode [here](#).

- Reload Consul config on SIGHUP (Cameron Daniel Kucera, Alexander Kukushkin)

It is especially useful when somebody is changing the value of `token`.

### Bugfixes

- Fix corner case in switchover/failover (Sharoon Thomas)

The variable `scheduled_at` may be undefined if REST API is not accessible and we are using DCS as a fallback.

- Open trust to localhost in `pg_hba.conf` during custom bootstrap (Alexander Kukushkin)

Previously it was open only to `unix_socket`, which was causing a lot of errors: `FATAL: no pg_hba.conf entry for replication connection from host "127.0.0.1", user "replicator"`

- Consider synchronous node as healthy even when the former leader is ahead (Alexander Kukushkin)

If the primary loses access to the DCS, it restarts Postgres in read-only, but it might happen that other nodes can still access the old primary via the REST API. Such a situation was causing the synchronous standby not to promote because the old primary was reporting WAL position ahead of the synchronous standby.

- Standby cluster bugfixes (Alexander Kukushkin)

Make it possible to bootstrap a replica in a standby cluster when the `standby_leader` is not accessible and a few other minor fixes.

## 20.38 Version 1.5.6

Released 2019-08-03

### New features

- Support work with etcd cluster via set of proxies (Alexander Kukushkin)

It might happen that etcd cluster is not accessible directly but via set of proxies. In this case Patroni will not perform etcd topology discovery but just round-robin via proxy hosts. Behavior is controlled by `etcd.use_proxies`.

- Changed callbacks behavior when role on the node is changed (Alexander Kukushkin)

If the role was changed from *master* or *standby\_leader* to *replica* or from *replica* to *standby\_leader*, *on\_restart* callback will not be called anymore in favor of *on\_role\_change* callback.

- Change the way how we start postgres (Alexander Kukushkin)

Use *multiprocessing.Process* instead of executing itself and *multiprocessing.Pipe* to transmit the postmaster pid to the Patroni process. Before that we were using pipes, what was leaving postmaster process with stdin closed.

#### Bug fixes

- Fix role returned by REST API for the standby leader (Alexander Kukushkin)

It was incorrectly returning *replica* instead of *standby\_leader*

- Wait for callback end if it could not be killed (Julien Tachaires)

Patroni doesn't have enough privileges to terminate the callback script running under *sudo* what was cancelling the new callback. If the running script could not be killed, Patroni will wait until it finishes and then run the next callback.

- Reduce lock time taken by *dcs.get\_cluster* method (Alexander Kukushkin)

Due to the lock being held DCS slowness was affecting the REST API health checks causing false positives.

- Improve cleaning of PGDATA when *pg\_wal/pg\_xlog* is a symlink (Julien Tachaires)

In this case Patroni will explicitly remove files from the target directory.

- Remove unnecessary usage of *os.path.relpath* (Ants Aasma)

It depends on being able to resolve the working directory, what will fail if Patroni is started in a directory that is later unlinked from the filesystem.

- Do not enforce ssl version when communicating with Etcd (Alexander Kukushkin)

For some unknown reason python3-etcd on debian and ubuntu are not based on the latest version of the package and therefore it enforces TLSv1 which is not supported by Etcd v3. We solved this problem on Patroni side.

## 20.39 Version 1.5.5

Released 2019-02-15

This version introduces the possibility of automatic reinit of the former master, improves *patronictl* list output and fixes a number of bugs.

#### New features

- Add support of *PATRONI\_ETCD\_PROTOCOL*, *PATRONI\_ETCD\_USERNAME* and *PATRONI\_ETCD\_PASSWORD* environment variables (Étienne M)

Before it was possible to configure them only in the config file or as a part of *PATRONI\_ETCD\_URL*, which is not always convenient.

- Make it possible to automatically reinit the former master (Alexander Kukushkin)

If the *pg\_rewind* is disabled or can't be used, the former master could fail to start as a new replica due to diverged timelines. In this case, the only way to fix it is wiping the data directory and reinitializing. This behavior could be changed by setting *postgresql.remove\_data\_directory\_on\_diverged\_timelines*. When it is set, Patroni will wipe the data directory and reinitialize the former master automatically.

- Show information about timelines in `patronictl list` (Alexander Kukushkin)

It helps to detect stale replicas. In addition to that, *Host* will include `:{port}` if the port value isn't default or there is more than one member running on the same host.

- Create a headless service associated with the `$SCOPE-config` endpoint (Alexander Kukushkin)

The “config” endpoint keeps information about the cluster-wide Patroni and Postgres configuration, history file, and last but not the most important, it holds the *initialize* key. When the Kubernetes master node is restarted or upgraded, it removes endpoints without services. The headless service will prevent it from being removed.

### Bug fixes

- Adjust the read timeout for the leader watch blocking query (Alexander Kukushkin)

According to the Consul documentation, the actual response timeout is increased by a small random amount of additional wait time added to the supplied maximum wait time to spread out the wake up time of any concurrent requests. It adds up to *wait / 16* additional time to the maximum duration. In our case we are adding *wait / 15* or 1 second depending on what is bigger.

- Always use `replication=1` when connecting via replication protocol to the postgres (Alexander Kukushkin)

Starting from Postgres 10 the line in the `pg_hba.conf` with `database=replication` doesn't accept connections with the parameter `replication=database`.

- Don't write `primary_conninfo` into `recovery.conf` for wal-only standby cluster (Alexander Kukushkin)

Despite not having neither *host* nor *port* defined in the *standby\_cluster* config, Patroni was putting the *primary\_conninfo* into the *recovery.conf*, which is useless and generating a lot of errors.

## 20.40 Version 1.5.4

Released 2019-01-15

This version implements flexible logging and fixes a number of bugs.

### New features

- Improvements in logging infrastructure (Alexander Kukushkin, Lucas Capistrant, Alexander Anikin)

Logging configuration could be configured not only from environment variables but also from Patroni config file. It makes it possible to change logging configuration in runtime by updating config and doing reload or sending SIGHUP to the Patroni process. By default Patroni writes logs to stderr, but now it becomes possible to write logs directly into the file and rotate when it reaches a certain size. In addition to that added support of custom dateformat and the possibility to fine-tune log level for each python module.

- Make it possible to take into account the current timeline during leader elections (Alexander Kukushkin)

It could happen that the node is considering itself as a healthiest one although it is currently not on the latest known timeline. In some cases we want to avoid promoting of such node, which could be achieved by setting *check\_timeline* parameter to *true* (default behavior remains unchanged).

- Relaxed requirements on superuser credentials

Libpq allows opening connections without explicitly specifying neither username nor password. Depending on situation it relies either on pgpass file or trust authentication method in `pg_hba.conf`. Since `pg_rewind` is also using libpq, it will work the same way.

- Implemented possibility to configure Consul Service registration and check interval via environment variables (Alexander Kukushkin)

Registration of service in Consul was added in the 1.5.0, but so far it was only possible to turn it on via `patroni.yaml`.

### Stability Improvements

- Set `archive_mode` to off during the custom bootstrap (Alexander Kukushkin)  
We want to avoid archiving wals and history files until the cluster is fully functional. It really helps if the custom bootstrap involves `pg_upgrade`.
- Apply five seconds backoff when loading global config on start (Alexander Kukushkin)  
It helps to avoid hammering DCS when Patroni just starting up.
- Reduce amount of error messages generated on shutdown (Alexander Kukushkin)  
They were harmless but rather annoying and sometimes scary.
- Explicitly secure rw perms for `recovery.conf` at creation time (Lucas Capistrant)  
We don't want anybody except `patroni/postgres` user reading this file, because it contains replication user and password.
- Redirect `HTTPServer` exceptions to logger (Julien Riou)  
By default, such exceptions were logged on standard output messing with regular logs.

### Bug fixes

- Removed `stderr` pipe to `stdout` on `pg_ctl` process (Cody Coons)  
Inheriting `stderr` from the main Patroni process allows all Postgres logs to be seen along with all patroni logs. This is very useful in a container environment as Patroni and Postgres logs may be consumed using standard tools (docker logs, kubectl, etc). In addition to that, this change fixes a bug with Patroni not being able to catch postmaster pid when postgres writing some warnings into `stderr`.
- Set Consul service check deregister timeout in Go time format (Pavel Kirillov)  
Without explicitly mentioned time unit registration was failing.
- Relax checks of `standby_cluster` cluster configuration (Dmitry Dolgov, Alexander Kukushkin)  
It was accepting only strings as valid values and therefore it was not possible to specify the port as integer and `create_replica_methods` as a list.

## 20.41 Version 1.5.3

Released 2018-12-03

Compatibility and bugfix release.

- Improve stability when running with python3 against zookeeper (Alexander Kukushkin)  
Change of `loop_wait` was causing Patroni to disconnect from zookeeper and never reconnect back.
- Fix broken compatibility with postgres 9.3 (Alexander Kukushkin)  
When opening a replication connection we should specify `replication=1`, because 9.3 does not understand `replication='database'`
- Make sure we refresh Consul session at least once per HA loop and improve handling of consul sessions exceptions (Alexander Kukushkin)  
Restart of local consul agent invalidates all sessions related to the node. Not calling session refresh on time and not doing proper handling of session errors was causing demote of the primary.

## 20.42 Version 1.5.2

Released 2018-11-26

Compatibility and bugfix release.

- Compatibility with kazoo-2.6.0 (Alexander Kukushkin)

In order to make sure that requests are performed with an appropriate timeout, Patroni redefines `create_connection` method from `python-kazoo` module. The last release of `kazoo` slightly changed the way how `create_connection` method is called.

- Fix Patroni crash when Consul cluster loses the leader (Alexander Kukushkin)

The crash was happening due to incorrect implementation of `touch_member` method, it should return boolean and not raise any exceptions.

## 20.43 Version 1.5.1

Released 2018-11-01

This version implements support of permanent replication slots, adds support of pgBackRest and fixes number of bugs.

### New features

- Permanent replication slots (Alexander Kukushkin)

Permanent replication slots are preserved on failover/switchover, that is, Patroni on the new primary will create configured replication slots right after doing promote. Slots could be configured with the help of *patronictl edit-config*. The initial configuration could be also done in the *bootstrap.dcs*.

- Add pgbackrest support (Yogesh Sharma)

pgBackrest can restore in existing `$PGDATA` folder, this allows speedy restore as files which have not changed since last backup are skipped, to support this feature new parameter `keep_data` has been introduced. See *replica creation method* section for additional examples.

### Bug fixes

- A few bugfixes in the “standby cluster” workflow (Alexander Kukushkin)

Please see <https://github.com/patroni/patroni/pull/823> for more details.

- Fix REST API health check when cluster management is paused and DCS is not accessible (Alexander Kukushkin)

Regression was introduced in <https://github.com/patroni/patroni/commit/90cf930036a9d5249265af15d2b787ec7517cf57>

## 20.44 Version 1.5.0

Released 2018-09-20

This version enables Patroni HA cluster to operate in a standby mode, introduces experimental support for running on Windows, and provides a new configuration parameter to register PostgreSQL service in Consul.

### New features



- Standby cluster (Dmitry Dolgov)

One or more Patroni nodes can form a standby cluster that runs alongside the primary one (i.e. in another datacenter) and consists of standby nodes that replicate from the master in the primary cluster. All PostgreSQL nodes in the standby cluster are replicas; one of those replicas elects itself to replicate directly from the remote master, while the others replicate from it in a cascading manner. More detailed description of this feature and some configuration examples can be found at [here](#).

- Register Services in Consul (Pavel Kirillov, Alexander Kukushkin)

If `register_service` parameter in the consul [configuration](#) is enabled, the node will register a service with the name `scope` and the tag `master`, `replica` or `standby-leader`.

- Experimental Windows support (Pavel Golub)

From now on it is possible to run Patroni on Windows, although Windows support is brand-new and hasn't received as much real-world testing as its Linux counterpart. We welcome your feedback!

### Improvements in patronictl

- Add `patronictl -k/--insecure` flag and support for restapi cert (Wilfried Roset)

In the past if the REST API was protected by the self-signed certificates `patronictl` would fail to verify them. There was no way to disable that verification. It is now possible to configure `patronictl` to skip the certificate verification altogether or provide CA and client certificates in the `ctl:` section of configuration.

- Exclude members with `nofailover` tag from `patronictl switchover/failover` output (Alexander Anikin)

Previously, those members were incorrectly proposed as candidates when performing interactive switchover or failover via `patronictl`.

### Stability improvements

- Avoid parsing non-key-value output lines of `pg_controldata` (Alexander Anikin)

Under certain circumstances `pg_controldata` outputs lines without a colon character. That would trigger an error in Patroni code that parsed `pg_controldata` output, hiding the actual problem; often such lines are emitted in a warning shown by `pg_controldata` before the regular output, i.e. when the binary major version does not match the one of the PostgreSQL data directory.

- Add member name to the error message during the leader election (Jan Mussler)

During the leader election, Patroni connects to all known members of the cluster and requests their status. Such status is written to the Patroni log and includes the name of the member. Previously, if the member was not accessible, the error message did not indicate its name, containing only the URL.

- Immediately reserve the WAL position upon creation of the replication slot (Alexander Kukushkin)

Starting from 9.6, `pg_create_physical_replication_slot` function provides an additional boolean parameter `immediately_reserve`. When it is set to `false`, which is also the default, the slot doesn't reserve the WAL position until it receives the first client connection, potentially losing some segments required by the client in a time window between the slot creation and the initial client connection.

- Fix bug in strict synchronous replication (Alexander Kukushkin)

When running with `synchronous_mode_strict: true`, in some cases Patroni puts `*` into the `synchronous_standby_names`, changing the sync state for most of the replication connections to `potential`. Previously, Patroni couldn't pick a synchronous candidate under such circumstances, as it only considered those with the state `async`.



## 20.45 Version 1.4.6

Released 2018-08-14

### Bug fixes and stability improvements

This release fixes a critical issue with Patroni API /master endpoint returning 200 for the non-master node. This is a reporting issue, no actual split-brain, but under certain circumstances clients might be directed to the read-only node.

- Reset is\_leader status on demote (Alexander Kukushkin, Oleksii Kliukin)  
Make sure demoted cluster member stops responding with code 200 on the /master API call.
- Add new “cluster\_unlocked” field to the API output (Dmitry Dolgov)  
This field indicates whether the cluster has the master running. It can be used when it is not possible to query any other node but one of the replicas.

## 20.46 Version 1.4.5

Released 2018-08-03

### New features

- Improve logging when applying new postgres configuration (Don Seiler)  
Patroni logs changed parameter names and values.
- Python 3.7 compatibility (Christoph Berg)  
async is a reserved keyword in python3.7
- Set state to “stopped” in the DCS when a member is shut down (Tony Sorrentino)  
This shows the member state as “stopped” in “patronictl list” command.
- Improve the message logged when stale postmaster.pid matches a running process (Ants Aasma)  
The previous one was beyond confusing.
- Implement patronictl reload functionality (Don Seiler)  
Before that it was only possible to reload configuration by either calling REST API or by sending SIGHUP signal to the Patroni process.
- Take and apply some parameters from controldata when starting as a replica (Alexander Kukushkin)  
The value of *max\_connections* and some other parameters set in the global configuration may be lower than the one actually used by the primary; when this happens, the replica cannot start and should be fixed manually. Patroni takes care of that now by reading and applying the value from *pg\_controldata*, starting postgres and setting *pending\_restart* flag.
- If set, use LD\_LIBRARY\_PATH when starting postgres (Chris Fraser)  
When starting up Postgres, Patroni was passing along PATH, LC\_ALL and LANG env vars if they are set. Now it is doing the same with LD\_LIBRARY\_PATH. It should help if somebody installed PostgreSQL to non-standard place.
- Rename create\_replica\_method to create\_replica\_methods (Dmitry Dolgov)  
To make it clear that it’s actually an array. The old name is still supported for backward compatibility.

### Bug fixes and stability improvements

- Fix condition for the replica start due to `pg_rewind` in paused state (Oleksii Kliukin)  
Avoid starting the replica that had already executed `pg_rewind` before.
- Respond 200 to the master health-check only if `update_lock` has been successful (Alexander Kukushkin)  
Prevent Patroni from reporting itself a master on the former (demoted) master if DCS is partitioned.
- Fix compatibility with the new consul module (Alexander Kukushkin)  
Starting from v1.1.0 python-consul changed internal API and started using *list* instead of *dict* to pass query parameters.
- Catch exceptions from Patroni REST API thread during shutdown (Alexander Kukushkin)  
Those uncaught exceptions kept PostgreSQL running at shutdown.
- Do crash recovery only when Postgres runs as the master (Alexander Kukushkin)  
Require *pg\_controldata* to report ‘in production’ or ‘shutting down’ or ‘in crash recovery’. In all other cases no crash recovery is necessary.
- Improve handling of configuration errors (Henning Jacobs, Alexander Kukushkin)  
It is possible to change a lot of parameters in runtime (including *restapi.listen*) by updating Patroni config file and sending SIGHUP to Patroni process. This fix eliminates obscure exceptions from the ‘restapi’ thread when some of the parameters receive invalid values.

## 20.47 Version 1.4.4

Released 2018-05-22

### Stability improvements

- Fix race condition in `poll_failover_result` (Alexander Kukushkin)  
It didn’t affect directly neither failover nor switchover, but in some rare cases it was reporting success too early, when the former leader released the lock, producing a ‘Failed over to “None”’ instead of ‘Failed over to “desired-node”’ message.
- Treat Postgres parameter names as case insensitive (Alexander Kukushkin)  
Most of the Postgres parameters have snake\_case names, but there are three exceptions from this rule: `DateStyle`, `IntervalStyle` and `TimeZone`. Postgres accepts those parameters when written in a different case (e.g. `timezone = ‘some/tzn’`); however, Patroni was unable to find case-insensitive matches of those parameter names in `pg_settings` and ignored such parameters as a result.
- Abort start if attaching to running postgres and cluster not initialized (Alexander Kukushkin)  
Patroni can attach itself to an already running Postgres instance. It is imperative to start running Patroni on the master node before getting to the replicas.
- Fix behavior of `patronictl scaffold` (Alexander Kukushkin)  
Pass dict object to `touch_member` instead of json encoded string, DCS implementation will take care of encoding it.
- Don’t demote master if failed to update leader key in pause (Alexander Kukushkin)  
During maintenance a DCS may start failing write requests while continuing to responds to read ones. In that case, Patroni used to put the Postgres master node to a read-only mode after failing to update the leader lock in DCS.

- Sync replication slots when Patroni notices a new postmaster process (Alexander Kukushkin)

If Postgres has been restarted, Patroni has to make sure that list of replication slots matches its expectations.

- Verify sysid and sync replication slots after coming out of pause (Alexander Kukushkin)

During the *maintenance* mode it may happen that data directory was completely rewritten and therefore we have to make sure that *Database system identifier* still belongs to our cluster and replication slots are in sync with Patroni expectations.

- Fix a possible failure to start not running Postgres on a data directory with postmaster lock file present (Alexander Kukushkin)

Detect reuse of PID from the postmaster lock file. More likely to hit such problem if you run Patroni and Postgres in the docker container.

- Improve protection of DCS being accidentally wiped (Alexander Kukushkin)

Patroni has a lot of logic in place to prevent failover in such case; it can also restore all keys back; however, until this change an accidental removal of `/config` key was switching off pause mode for 1 cycle of HA loop.

- Do not exit when encountering invalid system ID (Oleksii Kliukin)

Do not exit when the cluster system ID is empty or the one that doesn't pass the validation check. In that case, the cluster most likely needs a reinit; mention it in the result message. Avoid terminating Patroni, as otherwise reinit cannot happen.

### Compatibility with Kubernetes 1.10+

- Added check for empty subsets (Cody Coons)

Kubernetes 1.10.0+ started returning *Endpoints.subsets* set to *None* instead of *[]*.

### Bootstrap improvements

- Make deleting `recovery.conf` optional (Brad Nicholson)

If `bootstrap.<custom_bootstrap_method_name>.keep_existing_recovery_conf` is defined and set to `True`, Patroni will not remove the existing `recovery.conf` file. This is useful when bootstrapping from a backup with tools like pgBackRest that generate the appropriate `recovery.conf` for you.

- Allow options to the basebackup built-in method (Oleksii Kliukin)

It is now possible to supply options to the built-in basebackup method by defining the *basebackup* section in the configuration, similar to how those are defined for custom replica creation methods. The difference is in the format accepted by the *basebackup* section: since `pg_basebackup` accepts both `-key=value` and `-key` options, the contents of the section could be either a dictionary of key-value pairs, or a list of either one-element dictionaries or just keys (for the options that don't accept values). See [replica creation method](#) section for additional examples.

## 20.48 Version 1.4.3

Released 2018-03-05

### Improvements in logging

- Make log level configurable from environment variables (Andy Newton, Keyvan Hedayati)

`PATRONI_LOGLEVEL` - sets the general logging level `PATRONI_REQUESTS_LOGLEVEL` - sets the logging level for all HTTP requests e.g. Kubernetes API calls See *the docs for Python logging* <https://docs.python.org/3.6/library/logging.html#levels> to get the names of possible log levels

### Stability improvements and bug fixes

- Don't rediscover etcd cluster topology when watch timed out (Alexander Kukushkin)

If we have only one host in etcd configuration and exactly this host is not accessible, Patroni was starting discovery of cluster topology and never succeeding. Instead it should just switch to the next available node.

- Write content of bootstrap.pg\_hba into a pg\_hba.conf after custom bootstrap (Alexander Kukushkin)

Now it behaves similarly to the usual bootstrap with *initdb*

- Single user mode was waiting for user input and never finish (Alexander Kukushkin)

Regression was introduced in <https://github.com/patroni/patroni/pull/576>

## 20.49 Version 1.4.2

Released 2018-01-30

### Improvements in patronictl

- Rename scheduled failover to scheduled switchover (Alexander Kukushkin)

Failover and switchover functions were separated in version 1.4, but *patronictl list* was still reporting *Scheduled failover* instead of *Scheduled switchover*.

- Show information about pending restarts (Alexander Kukushkin)

In order to apply some configuration changes sometimes it is necessary to restart postgres. Patroni was already giving a hint about that in the REST API and when writing node status into DCS, but there were no easy way to display it.

- Make show-config to work with cluster\_name from config file (Alexander Kukushkin)

It works similar to the *patronictl edit-config*

### Stability improvements

- Avoid calling pg\_controldata during bootstrap (Alexander Kukushkin)

During initdb or custom bootstrap there is a time window when pgdata is not empty but pg\_controldata has not been written yet. In such case pg\_controldata call was failing with error messages.

- Handle exceptions raised from psutil (Alexander Kukushkin)

cmdline is read and parsed every time when *cmdline()* method is called. It could happen that the process being examined has already disappeared, in that case *NoSuchProcess* is raised.

### Kubernetes support improvements

- Don't swallow errors from k8s API (Alexander Kukushkin)

A call to Kubernetes API could fail for a different number of reasons. In some cases such call should be retried, in some other cases we should log the error message and the exception stack trace. The change here will help debug Kubernetes permission issues.

- Update Kubernetes example Dockerfile to install Patroni from the master branch (Maciej Szulik)

Before that it was using *feature/k8s*, which became outdated.

- Add proper RBAC to run patroni on k8s (Maciej Szulik)

Add the Service account that is assigned to the pods of the cluster, the role that holds only the necessary permissions, and the rolebinding that connects the Service account and the Role.

## 20.50 Version 1.4.1

Released 2018-01-17

### Fixes in patronictl

- Don't show current leader in suggested list of members to failover to. (Alexander Kukushkin)  
patronictl failover could still work when there is leader in the cluster and it should be excluded from the list of member where it is possible to failover to.
- Make patronictl switchover compatible with the old Patroni api (Alexander Kukushkin)  
In case if POST /switchover REST API call has failed with status code 501 it will do it once again, but to /failover endpoint.

## 20.51 Version 1.4

Released 2018-01-10

This version adds support for using Kubernetes as a DCS, allowing to run Patroni as a cloud-native agent in Kubernetes without any additional deployments of Etcd, Zookeeper or Consul.

### Upgrade notice

Installing Patroni via pip will no longer bring in dependencies for (such as libraries for Etcd, Zookeeper, Consul or Kubernetes, or support for AWS). In order to enable them one need to list them in pip install command explicitly, for instance `pip install patroni[etcd,kubernetes]`.

### Kubernetes support

Implement Kubernetes-based DCS. The endpoints meta-data is used in order to store the configuration and the leader key. The meta-data field inside the pods definition is used to store the member-related data. In addition to using Endpoints, Patroni supports ConfigMaps. You can find more information about this feature in the [Kubernetes chapter of the documentation](#)

### Stability improvements

- Factor out postmaster process into a separate object (Ants Aasma)  
This object identifies a running postmaster process via pid and start time and simplifies detection (and resolution) of situations when the postmaster was restarted behind our back or when postgres directory disappeared from the file system.
- Minimize the amount of SELECT's issued by Patroni on every loop of HA cycle (Alexander Kukushkin)  
On every iteration of HA loop Patroni needs to know recovery status and absolute wal position. From now on Patroni will run only single SELECT to get this information instead of two on the replica and three on the master.
- Remove leader key on shutdown only when we have the lock (Ants Aasma)  
Unconditional removal was generating unnecessary and misleading exceptions.

### Improvements in patronictl

- Add version command to patronictl (Ants Aasma)  
It will show the version of installed Patroni and versions of running Patroni instances (if the cluster name is specified).

- Make optional specifying `cluster_name` argument for some of `patronictl` commands (Alexander Kukushkin, Ants Aasma)

It will work if `patronictl` is using usual Patroni configuration file with the `scope` defined.

- Show information about scheduled switchover and maintenance mode (Alexander Kukushkin)

Before that it was possible to get this information only from Patroni logs or directly from DCS.

- Improve `patronictl reinit` (Alexander Kukushkin)

Sometimes `patronictl reinit` refused to proceed when Patroni was busy with other actions, namely trying to start postgres. `patronictl` didn't provide any commands to cancel such long running actions and the only (dangerous) workaround was removing a data directory manually. The new implementation of `reinit` forcefully cancels other long-running actions before proceeding with `reinit`.

- Implement `--wait` flag in `patronictl pause` and `patronictl resume` (Alexander Kukushkin)

It will make `patronictl` wait until the requested action is acknowledged by all nodes in the cluster. Such behaviour is achieved by exposing the `pause` flag for every node in DCS and via the REST API.

- Rename `patronictl failover` into `patronictl switchover` (Alexander Kukushkin)

The previous `failover` was actually only capable of doing a switchover; it refused to proceed in a cluster without the leader.

- Alter the behavior of `patronictl failover` (Alexander Kukushkin)

It will work even if there is no leader, but in that case you will have to explicitly specify a node which should become the new leader.

#### **Expose information about timeline and history**

- Expose current timeline in DCS and via API (Alexander Kukushkin)

Store information about the current timeline for each member of the cluster. This information is accessible via the API and is stored in the DCS

- Store promotion history in the `/history` key in DCS (Alexander Kukushkin)

In addition, store the timeline history enriched with the timestamp of the corresponding promotion in the `/history` key in DCS and update it with each promote.

#### **Add endpoints for getting synchronous and asynchronous replicas**

- Add new `/sync` and `/async` endpoints (Alexander Kukushkin, Oleksii Kliukin)

Those endpoints (also accessible as `/synchronous` and `/asynchronous`) return 200 only for synchronous and asynchronous replicas correspondingly (excluding those marked as *noloadbalance*).

#### **Allow multiple hosts for Etcd**

- Add a new `hosts` parameter to Etcd configuration (Alexander Kukushkin)

This parameter should contain the initial list of hosts that will be used to discover and populate the list of the running etcd cluster members. If for some reason during work this list of discovered hosts is exhausted (no available hosts from that list), Patroni will return to the initial list from the `hosts` parameter.

## 20.52 Version 1.3.6

Released 2017-11-10

### Stability improvements

- Verify process start time when checking if postgres is running. (Ants Aasma)

After a crash that doesn't clean up postmaster.pid there could be a new process with the same pid, resulting in a false positive for `is_running()`, which will lead to all kinds of bad behavior.

- Shutdown postgresql before bootstrap when we lost data directory (ainlolcat)

When data directory on the master is forcefully removed, postgres process can still stay alive for some time and prevent the replica created in place of that former master from starting or replicating. The fix makes Patroni cache the postmaster pid and its start time and let it terminate the old postmaster in case it is still running after the corresponding data directory has been removed.

- Perform crash recovery in a single user mode if postgres master dies (Alexander Kukushkin)

It is unsafe to start immediately as a standby and not possible to run `pg_rewind` if postgres hasn't been shut down cleanly. The single user crash recovery only kicks in if `pg_rewind` is enabled or there is no master at the moment.

### Consul improvements

- Make it possible to provide datacenter configuration for Consul (Vilius Okockis, Alexander Kukushkin)

Before that Patroni was always communicating with datacenter of the host it runs on.

- Always send a token in X-Consul-Token http header (Alexander Kukushkin)

If `consul.token` is defined in Patroni configuration, we will always send it in the 'X-Consul-Token' http header. python-consul module tries to be "consistent" with Consul REST API, which doesn't accept token as a query parameter for [session API](#), but it still works with 'X-Consul-Token' header.

- Adjust session TTL if supplied value is smaller than the minimum possible (Stas Fomin, Alexander Kukushkin)

It could happen that the TTL provided in the Patroni configuration is smaller than the minimum one supported by Consul. In that case, Consul agent fails to create a new session. Without a session Patroni cannot create member and leader keys in the Consul KV store, resulting in an unhealthy cluster.

### Other improvements

- Define custom log format via environment variable `PATRONI_LOGFORMAT` (Stas Fomin)

Allow disabling timestamps and other similar fields in Patroni logs if they are already added by the system logger (usually when Patroni runs as a service).

## 20.53 Version 1.3.5

Released 2017-10-12

### Bugfix

- Set role to 'uninitialized' if data directory was removed (Alexander Kukushkin)

If the node was running as a master it was preventing from failover.

### Stability improvement

- Try to run postmaster in a single-user mode if we tried and failed to start postgres (Alexander Kukushkin)

Usually such problem happens when node running as a master was terminated and timelines were diverged. If `recovery.conf` has `restore_command` defined, there are really high chances that postgres will abort startup and leave controldata unchanged. It makes impossible to use `pg_rewind`, which requires a clean shutdown.

### Consul improvements

- Make it possible to specify health checks when creating session (Alexander Kukushkin)

If not specified, Consul will use “serfHealth”. From one side it allows fast detection of isolated master, but from another side it makes it impossible for Patroni to tolerate short network lags.

### Bugfix

- Fix watchdog on Python 3 (Ants Aasma)

A misunderstanding of the `ioctl()` call interface. If `mutable=False` then `fcntl.ioctl()` actually returns the arg buffer back. This accidentally worked on Python2 because int and str comparison did not return an error. Error reporting is actually done by raising `IOError` on Python2 and `OSError` on Python3.

## 20.54 Version 1.3.4

Released 2017-09-08

### Different Consul improvements

- Pass the consul token as a header (Andrew Colin Kissa)

Headers are now the preferred way to pass the token to the consul [API](#).

- Advanced configuration for Consul (Alexander Kukushkin)

possibility to specify `scheme`, `token`, `client` and `ca` certificates [details](#).

- compatibility with python-consul-0.7.1 and above (Alexander Kukushkin)

new python-consul module has changed signature of some methods

- “Could not take out TTL lock” message was never logged (Alexander Kukushkin)

Not a critical bug, but lack of proper logging complicates investigation in case of problems.

### Quote synchronous\_standby\_names using quote\_ident

- When writing `synchronous_standby_names` into the `postgresql.conf` its value must be quoted (Alexander Kukushkin)

If it is not quoted properly, PostgreSQL will effectively disable synchronous replication and continue to work.

### Different bugfixes around pause state, mostly related to watchdog (Alexander Kukushkin)

- Do not send keepalives if watchdog is not active
- Avoid activating watchdog in a pause mode
- Set correct postgres state in pause mode
- Do not try to run queries from API if postgres is stopped



## 20.55 Version 1.3.3

Released 2017-08-04

### Bugfixes

- synchronous replication was disabled shortly after promotion even when `synchronous_mode_strict` was turned on (Alexander Kukushkin)
- create empty `pg_ident.conf` file if it is missing after restoring from the backup (Alexander Kukushkin)
- open access in `pg_hba.conf` to all databases, not only postgres (Franco Bellagamba)

## 20.56 Version 1.3.2

Released 2017-07-31

### Bugfix

- `patronictl edit-config` didn't work with ZooKeeper (Alexander Kukushkin)

## 20.57 Version 1.3.1

Released 2017-07-28

### Bugfix

- failover via API was broken due to change in `_MemberStatus` (Alexander Kukushkin)

## 20.58 Version 1.3

Released 2017-07-27

Version 1.3 adds custom bootstrap possibility, significantly improves support for `pg_rewind`, enhances the synchronous mode support, adds configuration editing to `patronictl` and implements watchdog support on Linux. In addition, this is the first version to work correctly with PostgreSQL 10.

### Upgrade notice

There are no known compatibility issues with the new version of Patroni. Configuration from version 1.2 should work without any changes. It is possible to upgrade by installing new packages and either restarting Patroni (will cause PostgreSQL restart), or by putting Patroni into a *pause mode* first and then restarting Patroni on all nodes in the cluster (Patroni in a pause mode will not attempt to stop/start PostgreSQL), resuming from the pause mode at the end.

### Custom bootstrap

- Make the process of bootstrapping the cluster configurable (Alexander Kukushkin)

Allow custom bootstrap scripts instead of `initdb` when initializing the very first node in the cluster. The bootstrap command receives the name of the cluster and the path to the data directory. The resulting cluster can be configured to perform recovery, making it possible to bootstrap from a backup and do point in time recovery. Refer to the [documentation page](#) for more detailed description of this feature.

### Smarter `pg_rewind` support

- Decide on whether to run `pg_rewind` by looking at the timeline differences from the current master (Alexander Kukushkin)

Previously, Patroni had a fixed set of conditions to trigger `pg_rewind`, namely when starting a former master, when doing a switchover to the designated node for every other node in the cluster or when there is a replica with the `nofailover` tag. All those cases have in common a chance that some replica may be ahead of the new master. In some cases, `pg_rewind` did nothing, in some other ones it was not running when necessary. Instead of relying on this limited list of rules make Patroni compare the master and the replica WAL positions (using the streaming replication protocol) in order to reliably decide if rewind is necessary for the replica.

### **Synchronous replication mode strict**

- Enhance synchronous replication support by adding the strict mode (James Sewell, Alexander Kukushkin)

Normally, when `synchronous_mode` is enabled and there are no replicas attached to the master, Patroni will disable synchronous replication in order to keep the master available for writes. The `synchronous_mode_strict` option changes that, when it is set Patroni will not disable the synchronous replication in a lack of replicas, effectively blocking all clients writing data to the master. In addition to the synchronous mode guarantee of preventing any data loss due to automatic failover, the strict mode ensures that each write is either durably stored on two nodes or not happening altogether if there is only one node in the cluster.

### **Configuration editing with `patronictl`**

- Add configuration editing to `patronictl` (Ants Aasma, Alexander Kukushkin)

Add the ability to `patronictl` of editing dynamic cluster configuration stored in DCS. Support either specifying the parameter/values from the command-line, invoking the `$EDITOR`, or applying configuration from the yaml file.

### **Linux watchdog support**

- Implement watchdog support for Linux (Ants Aasma)

Support Linux software watchdog in order to reboot the node where Patroni is not running or not responding (e.g because of the high load) The Linux software watchdog reboots the non-responsive node. It is possible to configure the watchdog device to use (`/dev/watchdog` by default) and the mode (on, automatic, off) from the watchdog section of the Patroni configuration. You can get more information from the [watchdog documentation](#).

### **Add support for PostgreSQL 10**

- Patroni is compatible with all beta versions of PostgreSQL 10 released so far and we expect it to be compatible with the PostgreSQL 10 when it will be released.

### **PostgreSQL-related minor improvements**

- Define `pg_hba.conf` via the Patroni configuration file or the dynamic configuration in DCS (Alexander Kukushkin)

Allow to define the contents of `pg_hba.conf` in the `pg_hba` sub-section of the `postgresql` section of the configuration. This simplifies managing `pg_hba.conf` on multiple nodes, as one needs to define it only ones in DCS instead of logging to every node, changing it manually and reload the configuration.

When defined, the contents of this section will replace the current `pg_hba.conf` completely. Patroni ignores it if `hba_file` PostgreSQL parameter is set.

- Support connecting via a UNIX socket to the local PostgreSQL cluster (Alexander Kukushkin)

Add the `use_unix_socket` option to the `postgresql` section of Patroni configuration. When set to true and the PostgreSQL `unix_socket_directories` option is not empty, enables Patroni to use the first value from it to connect to the local PostgreSQL cluster. If `unix_socket_directories` is not defined, Patroni will assume its default value and omit the `host` parameter in the PostgreSQL connection string altogether.

- Support change of superuser and replication credentials on reload (Alexander Kukushkin)

- Support storing of configuration files outside of PostgreSQL data directory (@jouis)

Add the new configuration `postgresql` configuration directive `config_dir`. It defaults to the data directory and must be writable by Patroni.

### Bug fixes and stability improvements

- Handle `EtcdEventIndexCleared` and `EtcdWatcherCleared` exceptions (Alexander Kukushkin)

Faster recovery when the watch operation is ended by Etcd by avoiding useless retries.

- Remove error spinning on Etcd failure and reduce log spam (Ants Aasma)

Avoid immediate retrying and emitting stack traces in the log on the second and subsequent Etcd connection failures.

- Export locale variables when forking PostgreSQL processes (Oleksii Kliukin)

Avoid the *postmaster became multithreaded during startup* fatal error on non-English locales for PostgreSQL built with NLS.

- Extra checks when dropping the replication slot (Alexander Kukushkin)

In some cases Patroni is prevented from dropping the replication slot by the WAL sender.

- Truncate the replication slot name to 63 (`NAMEDATALEN - 1`) characters to comply with PostgreSQL naming rules (Nick Scott)

- Fix a race condition resulting in extra connections being opened to the PostgreSQL cluster from Patroni (Alexander Kukushkin)

- Release the leader key when the node restarts with an empty data directory (Alex Kerney)

- Set asynchronous executor busy when running bootstrap without a leader (Alexander Kukushkin)

Failure to do so could have resulted in errors stating the node belonged to a different cluster, as Patroni proceeded with the normal business while being bootstrapped by a bootstrap method that doesn't require a leader to be present in the cluster.

- Improve WAL-E replica creation method (Joar Wandborg, Alexander Kukushkin).

- Use `csv.DictReader` when parsing WAL-E base backup, accepting ISO dates with space-delimited date and time.
- Support fetching current WAL position from the replica to estimate the amount of WAL to restore. Previously, the code used to call system information functions that were available only on the master node.

## 20.59 Version 1.2

Released 2016-12-13

This version introduces significant improvements over the handling of synchronous replication, makes the startup process and failover more reliable, adds PostgreSQL 9.6 support and fixes plenty of bugs. In addition, the documentation, including these release notes, has been moved to <https://patroni.readthedocs.io>.

### Synchronous replication

- Add synchronous replication support. (Ants Aasma)

Adds a new configuration variable `synchronous_mode`. When enabled, Patroni will manage `synchronous_standby_names` to enable synchronous replication whenever there are healthy standbys available. When synchronous mode is enabled, Patroni will automatically fail over only to a standby that was

synchronously replicating at the time of the master failure. This effectively means that no user visible transaction gets lost in such a case. See the [feature documentation](#) for the detailed description and implementation details.

### Reliability improvements

- Do not try to update the leader position stored in the `leader_optime` key when PostgreSQL is not 100% healthy. Demote immediately when the update of the leader key failed. (Alexander Kukushkin)
- Exclude unhealthy nodes from the list of targets to clone the new replica from. (Alexander Kukushkin)
- Implement retry and timeout strategy for Consul similar to how it is done for Etcd. (Alexander Kukushkin)
- Make `--dcs` and `--config-file` apply to all options in `patronictl`. (Alexander Kukushkin)
- Write all postgres parameters into `postgresql.conf`. (Alexander Kukushkin)

It allows starting PostgreSQL configured by Patroni with just `pg_ctl`.

- Avoid exceptions when there are no users in the config. (Kirill Pushkin)
- Allow pausing an unhealthy cluster. Before this fix, `patronictl` would bail out if the node it tries to execute pause on is unhealthy. (Alexander Kukushkin)
- Improve the leader watch functionality. (Alexander Kukushkin)

Previously the replicas were always watching the leader key (sleeping until the timeout or the leader key changes). With this change, they only watch when the replica's PostgreSQL is in the `running` state and not when it is stopped/starting or restarting PostgreSQL.

- Avoid running into race conditions when handling SIGCHLD as a PID 1. (Alexander Kukushkin)

Previously a race condition could occur when running inside the Docker containers, since the same process inside Patroni both spawned new processes and handled SIGCHLD from them. This change uses `fork/execs` for Patroni and leaves the original PID 1 process responsible for handling signals from children.

- Fix WAL-E restore. (Oleksii Kliukin)

Previously WAL-E restore used the `no_master` flag to avoid consulting with the master altogether, making Patroni always choose restoring from WAL over the `pg_basebackup`. This change reverts it to the original meaning of `no_master`, namely Patroni WAL-E restore may be selected as a replication method if the master is not running. The latter is checked by examining the connection string passed to the method. In addition, it makes the retry mechanism more robust and handles other minutia.

- Implement asynchronous DNS resolver cache. (Alexander Kukushkin)

Avoid failing when DNS is temporary unavailable (for instance, due to an excessive traffic received by the node).

- Implement starting state and master start timeout. (Ants Aasma, Alexander Kukushkin)

Previously `pg_ctl` waited for a timeout and then happily trodded on considering PostgreSQL to be running. This caused PostgreSQL to show up in listings as running when it was actually not and caused a race condition that resulted in either a failover, or a crash recovery, or a crash recovery interrupted by failover and a missed rewind. This change adds a `master_start_timeout` parameter and introduces a new state for the main HA loop: `starting`. When `master_start_timeout` is 0 we will failover immediately when the master crashes as soon as there is a failover candidate. Otherwise, Patroni will wait after attempting to start PostgreSQL on the master for the duration of the timeout; when it expires, it will failover if possible. Manual failover requests will be honored during the crash of the master even before the timeout expiration.

Introduce the `timeout` parameter to the `restart` API endpoint and `patronictl`. When it is set and restart takes longer than the timeout, PostgreSQL is considered unhealthy and the other nodes becomes eligible to take the leader lock.

- Fix `pg_rewind` behavior in a pause mode. (Ants Aasma)

Avoid unnecessary restart in a pause mode when Patroni thinks it needs to rewind but rewind is not possible (i.e. `pg_rewind` is not present). Fallback to default `libpq` values for the `superuser` (default OS user) if `superuser` authentication is missing from the `pg_rewind` related Patroni configuration section.

- Serialize callback execution. Kill the previous callback of the same type when the new one is about to run. Fix the issue of spawning zombie processes when running callbacks. (Alexander Kukushkin)
- Avoid promoting a former master when the leader key is set in DCS but update to this leader key fails. (Alexander Kukushkin)

This avoids the issue of a current master continuing to keep its role when it is partitioned together with the minority of nodes in Etcd and other DCSs that allow “inconsistent reads”.

### Miscellaneous

- Add `post_init` configuration option on bootstrap. (Alejandro Martínez)

Patroni will call the script argument of this option right after running `initdb` and starting up PostgreSQL for a new cluster. The script receives a connection URL with `superuser` and sets `PGPASSFILE` to point to the `.pgpass` file containing the password. If the script fails, Patroni initialization fails as well. It is useful for adding new users or creating extensions in the new cluster.

- Implement PostgreSQL 9.6 support. (Alexander Kukushkin)

Use `wal_level = replica` as a synonym for `hot_standby`, avoiding `pending_restart` flag when it changes from one to another. (Alexander Kukushkin)

### Documentation improvements

- Add a Patroni main [loop workflow diagram](#). (Alejandro Martínez, Alexander Kukushkin)
- Improve README, adding the Helm chart and links to release notes. (Lauri Apple)
- Move Patroni documentation to Read the Docs. The up-to-date documentation is available at <https://patroni.readthedocs.io>. (Oleksii Kliukin)

Makes the documentation easily viewable from different devices (including smartphones) and searchable.

- Move the package to the semantic versioning. (Oleksii Kliukin)

Patroni will follow the major.minor.patch version schema to avoid releasing the new minor version on small but critical bugfixes. We will only publish the release notes for the minor version, which will include all patches.

## 20.60 Version 1.1

Released 2016-09-07

This release improves management of Patroni cluster by bring in pause mode, improves maintenance with scheduled and conditional restarts, makes Patroni interaction with Etcd or Zookeeper more resilient and greatly enhances `patronictl`.

### Upgrade notice

When upgrading from releases below 1.0 read about changing of credentials and configuration format at 1.0 release notes.

### Pause mode

- Introduce pause mode to temporary detach Patroni from managing PostgreSQL instance (Murat Kabilov, Alexander Kukushkin, Oleksii Kliukin).

Previously, one had to send `SIGKILL` signal to Patroni to stop it without terminating PostgreSQL. The new pause mode detaches Patroni from PostgreSQL cluster-wide without terminating Patroni. It is similar to the maintenance mode in Pacemaker. Patroni is still responsible for updating member and leader keys in DCS, but it

will not start, stop or restart PostgreSQL server in the process. There are a few exceptions, for instance, manual failovers, reinitializes and restarts are still allowed. You can read [a detailed description of this feature](#).

In addition, patronictl supports new `pause` and `resume` commands to toggle the pause mode.

### **Scheduled and conditional restarts**

- Add conditions to the restart API command (Oleksii Kliukin)

This change enhances Patroni restarts by adding a couple of conditions that can be verified in order to do the restart. Among the conditions are restarting when PostgreSQL role is either a master or a replica, checking the PostgreSQL version number or restarting only when restart is necessary in order to apply configuration changes.

- Add scheduled restarts (Oleksii Kliukin)

It is now possible to schedule a restart in the future. Only one scheduled restart per node is supported. It is possible to clear the scheduled restart if it is not needed anymore. A combination of scheduled and conditional restarts is supported, making it possible, for instance, to schedule minor PostgreSQL upgrades in the night, restarting only the instances that are running the outdated minor version without adding postgres-specific logic to administration scripts.

- Add support for conditional and scheduled restarts to patronictl (Murat Kabilov).

patronictl restart supports several new options. There is also patronictl flush command to clean the scheduled actions.

### **Robust DCS interaction**

- Set Kazoo timeouts depending on the loop\_wait (Alexander Kukushkin)

Originally, ping\_timeout and connect\_timeout values were calculated from the negotiated session timeout. Patroni loop\_wait was not taken into account. As a result, a single retry could take more time than the session timeout, forcing Patroni to release the lock and demote.

This change set ping and connect timeout to half of the value of loop\_wait, speeding up detection of connection issues and leaving enough time to retry the connection attempt before losing the lock.

- Update Etcd topology only after original request succeed (Alexander Kukushkin)

Postpone updating the Etcd topology known to the client until after the original request. When retrieving the cluster topology, implement the retry timeouts depending on the known number of nodes in the Etcd cluster. This makes our client prefer to get the results of the request to having the up-to-date list of nodes.

Both changes make Patroni connections to DCS more robust in the face of network issues.

### **Patronictl, monitoring and configuration**

- Return information about streaming replicas via the API (Feike Steenbergen)

Previously, there was no reliable way to query Patroni about PostgreSQL instances that fail to stream changes (for instance, due to connection issues). This change exposes the contents of pg\_stat\_replication via the /patroni endpoint.

- Add patronictl scaffold command (Oleksii Kliukin)

Add a command to create cluster structure in Etcd. The cluster is created with user-specified sysid and leader, and both leader and member keys are made persistent. This command is useful to create so-called master-less configurations, where Patroni cluster consisting of only replicas replicate from the external master node that is unaware of Patroni. Subsequently, one may remove the leader key, promoting one of the Patroni nodes and replacing the original master with the Patroni-based HA cluster.

- Add configuration option bin\_dir to locate PostgreSQL binaries (Ants Aasma)

It is useful to be able to specify the location of PostgreSQL binaries explicitly when Linux distros that support installing multiple PostgreSQL versions at the same time.

- Allow configuration file path to be overridden using `custom_conf` of (Alejandro Martínez)

Allows for custom configuration file paths, which will be unmanaged by Patroni, [details](#).

### Bug fixes and code improvements

- Make Patroni compatible with new version schema in PostgreSQL 10 and above (Feike Steenberg)

Make sure that Patroni understand 2-digits version numbers when doing conditional restarts based on the PostgreSQL version.

- Use `pkgutil` to find DCS modules (Alexander Kukushkin)

Use the dedicated python module instead of traversing directories manually in order to find DCS modules.

- Always call `on_start` callback when starting Patroni (Alexander Kukushkin)

Previously, Patroni did not call any callbacks when attaching to the already running node with the correct role. Since callbacks are often used to route client connections that could result in the failure to register the running node in the connection routing scheme. With this fix, Patroni calls `on_start` callback even when attaching to the already running node.

- Do not drop active replication slots (Murat Kabilov, Oleksii Kliukin)

Avoid dropping active physical replication slots on master. PostgreSQL cannot drop such slots anyway. This change makes possible to run non-Patroni managed replicas/consumers on the master.

- Close Patroni connections during start of the PostgreSQL instance (Alexander Kukushkin)

Forces Patroni to close all former connections when PostgreSQL node is started. Avoids the trap of reusing former connections if postmaster was killed with `SIGKILL`.

- Replace invalid characters when constructing slot names from member names (Ants Aasma)

Make sure that standby names that do not comply with the slot naming rules don't cause the slot creation and standby startup to fail. Replace the dashes in the slot names with underscores and all other characters not allowed in slot names with their unicode codepoints.

## 20.61 Version 1.0

Released 2016-07-05

This release introduces the global dynamic configuration that allows dynamic changes of the PostgreSQL and Patroni configuration parameters for the entire HA cluster. It also delivers numerous bugfixes.

### Upgrade notice

When upgrading from v0.90 or below, always upgrade all replicas before the master. Since we don't store replication credentials in DCS anymore, an old replica won't be able to connect to the new master.

### Dynamic Configuration

- Implement the dynamic global configuration (Alexander Kukushkin)

Introduce new REST API endpoint `/config` to provide PostgreSQL and Patroni configuration parameters that should be set globally for the entire HA cluster (master and all the replicas). Those parameters are set in DCS and in many cases can be applied without disrupting PostgreSQL or Patroni. Patroni sets a special flag called "pending restart" visible via the API when some of the values require the PostgreSQL restart. In that case, restart should be issued manually via the API.

Patroni `SIGHUP` or `POST` to `/reload` will make it re-read the configuration file.



See the [Patroni configuration](#) for the details on which parameters can be changed and the order of processing difference configuration sources.

The configuration file format *has changed* since the v0.90. Patroni is still compatible with the old configuration files, but in order to take advantage of the bootstrap parameters one needs to change it. Users are encourage to update them by referring to the [dynamic configuration documentation page](#).

### **More flexible configuration\***

- Make postgresql configuration and database name Patroni connects to configurable (Misja Hoebe)  
Introduce *database* and *config\_base\_name* configuration parameters. Among others, it makes possible to run Patroni with PipelineDB and other PostgreSQL forks.
- Implement possibility to configure some Patroni configuration parameters via environment (Alexander Kukushkin)  
Those include the scope, the node name and the namespace, as well as the secrets and makes it easier to run Patroni in a dynamic environment, i.e. Kubernetes Please, refer to the [supported environment variables](#) for further details.
- Update the built-in Patroni docker container to take advantage of environment-based configuration (Feike Steenberg).
- Add Zookeeper support to Patroni docker image (Alexander Kukushkin)
- Split the Zookeeper and Exhibitor configuration options (Alexander Kukushkin)
- Make patronictl reuse the code from Patroni to read configuration (Alexander Kukushkin)  
This allows patronictl to take advantage of environment-based configuration.
- Set application name to node name in primary\_conninfo (Alexander Kukushkin)  
This simplifies identification and configuration of synchronous replication for a given node.

### **Stability, security and usability improvements**

- Reset sysid and do not call pg\_controldata when restore of backup in progress (Alexander Kukushkin)  
This change reduces the amount of noise generated by Patroni API health checks during the lengthy initialization of this node from the backup.
- Fix a bunch of pg\_rewind corner-cases (Alexander Kukushkin)  
Avoid running pg\_rewind if the source cluster is not the master.  
In addition, avoid removing the data directory on an unsuccessful rewind, unless the new parameter *remove\_data\_directory\_on\_rewind\_failure* is set to true. By default it is false.
- Remove passwords from the replication connection string in DCS (Alexander Kukushkin)  
Previously, Patroni always used the replication credentials from the Postgres URL in DCS. That is now changed to take the credentials from the patroni configuration. The secrets (replication username and password) and no longer exposed in DCS.
- Fix the asynchronous machinery around the demote call (Alexander Kukushkin)  
Demote now runs totally asynchronously without blocking the DCS interactions.
- Make patronictl always send the authorization header if it is configured (Alexander Kukushkin)  
This allows patronictl to issue “protected” requests, i.e. restart or reinitialize, when Patroni is configured to require authorization on those.



- Handle the SystemExit exception correctly (Alexander Kukushkin)  
Avoids the issues of Patroni not stopping properly when receiving the SIGTERM
- Sample haproxy templates for confd (Alexander Kukushkin)  
Generates and dynamically changes haproxy configuration from the patroni state in the DCS using confd
- Improve and restructure the documentation to make it more friendly to the new users (Lauri Apple)
- API must report role=master during pg\_ctl stop (Alexander Kukushkin)  
Makes the callback calls more reliable, particularly in the cluster stop case. In addition, introduce the *pg\_ctl\_timeout* option to set the timeout for the start, stop and restart calls via the *pg\_ctl*.
- Fix the retry logic in etcd (Alexander Kukushkin)  
Make retries more predictable and robust.
- Make Zookeeper code more resilient against short network hiccups (Alexander Kukushkin)  
Reduce the connection timeouts to make Zookeeper connection attempts more frequent.

## 20.62 Version 0.90

Released 2016-04-27

This releases adds support for Consul, includes a new *noloadbalance* tag, changes the behavior of the *clonefrom* tag, improves *pg\_rewind* handling and improves *patronictl* control program.

### Consul support

- Implement Consul support (Alexander Kukushkin)  
Patroni runs against Consul, in addition to Etcd and Zookeeper. the connection parameters can be configured in the YAML file.

### New and improved tags

- Implement *noloadbalance* tag (Alexander Kukushkin)  
This tag makes Patroni always return that the replica is not available to the load balancer.
- Change the implementation of the *clonefrom* tag (Alexander Kukushkin)  
Previously, a node name had to be supplied to the *clonefrom*, forcing a tagged replica to clone from the specific node. The new implementation makes *clonefrom* a boolean tag: if it is set to true, the replica becomes a candidate for other replicas to clone from it. When multiple candidates are present, the replicas picks one randomly.

### Stability and security improvements

- Numerous reliability improvements (Alexander Kukushkin)  
Removes some spurious error messages, improves the stability of the failover, addresses some corner cases with reading data from DCS, shutdown, demote and reattaching of the former leader.
- Improve systems script to avoid killing Patroni children on stop (Jan Keirse, Alexander Kukushkin)  
Previously, when stopping Patroni, *systemd* also sent a signal to PostgreSQL. Since Patroni also tried to stop PostgreSQL by itself, it resulted in sending to different shutdown requests (the smart shutdown, followed by the fast shutdown). That resulted in replicas disconnecting too early and a former master not being able to rejoin after demote. Fix by Jan with prior research by Alexander.

- Eliminate some cases where the former master was unable to call `pg_rewind` before rejoining as a replica (Oleksii Kliukin)

Previously, we only called `pg_rewind` if the former master had crashed. Change this to always run `pg_rewind` for the former master as long as `pg_rewind` is present in the system. This fixes the case when the master is shut down before the replicas managed to get the latest changes (i.e. during the “smart” shutdown).

- Numerous improvements to unit- and acceptance- tests, in particular, enable support for Zookeeper and Consul (Alexander Kukushkin).
- Make Travis CI faster and implement support for running tests against Zookeeper (Exhibitor) and Consul (Alexander Kukushkin)

Both unit and acceptance tests run automatically against Etcd, Zookeeper and Consul on each commit or pull-request.

- Clear environment variables before calling PostgreSQL commands from Patroni (Feike Steenberg)

This prevents a possibility of reading system environment variables by connecting to the PostgreSQL cluster managed by Patroni.

### Configuration and control changes

- Unify `patronictl` and Patroni configuration (Feike Steenberg)

`patronictl` can use the same configuration file as Patroni itself.

- Enable Patroni to read the configuration from the environment variables (Oleksii Kliukin)

This simplifies generating configuration for Patroni automatically, or merging a single configuration from different sources.

- Include database system identifier in the information returned by the API (Feike Steenberg)
- Implement `delete_cluster` for all available DCSs (Alexander Kukushkin)

Enables support for DCSs other than Etcd in `patronictl`.

## 20.63 Version 0.80

Released 2016-03-14

This release adds support for *cascading replication* and simplifies Patroni management by providing *scheduled failovers*. One may use older versions of Patroni (in particular, 0.78) combined with this one in order to migrate to the new release. Note that the scheduled failover and cascading replication related features will only work with Patroni 0.80 and above.

### Cascading replication

- Add support for the *replicatefrom* and *clonefrom* tags for the patroni node (Oleksii Kliukin).

The tag *replicatefrom* allows a replica to use an arbitrary node as a source, not necessarily the master. The *clonefrom* does the same for the initial backup. Together, they enable Patroni to fully support cascading replication.

- Add support for running replication methods to initialize the replica even without a running replication connection (Oleksii Kliukin).

This is useful in order to create replicas from the snapshots stored on S3 or FTP. A replication method that does not require a running replication connection should supply `no_master: true` in the yaml configuration. Those scripts will still be called in order if the replication connection is present.

### Patronictl, API and DCS improvements

- Implement scheduled failovers (Feike Steenberg).

Failovers can be scheduled to happen at a certain time in the future, using either `patronictl`, or API calls.

- Add support for `dbuser` and `password` parameters in `patronictl` (Feike Steenberg).
- Add PostgreSQL version to the health check output (Feike Steenberg).
- Improve Zookeeper support in `patronictl` (Oleksandr Shulgin)
- Migrate to `python-etcd` 0.43 (Alexander Kukushkin)

### Configuration

- Add a sample systems configuration script for Patroni (Jan Keirse).
- Fix the problem of Patroni ignoring the superuser name specified in the configuration file for DB connections (Alexander Kukushkin).
- Fix the handling of CTRL-C by creating a separate session ID and process group for the postmaster launched by Patroni (Alexander Kukushkin).

### Tests

- Add acceptance tests with *behave* in order to check real-world scenarios of running Patroni (Alexander Kukushkin, Oleksii Kliukin).

The tests can be launched manually using the *behave* command. They are also launched automatically for pull requests and after commits.

Release notes for some older versions can be found on [project's github page](#).



## CONTRIBUTING

Resources and information for developers can be found in the pages below.

### 21.1 Contributing guidelines

#### 21.1.1 Chatting

If you have a question, looking for an interactive troubleshooting help or want to chat with other Patroni users, join us on channel `#patroni` in the [PostgreSQL Slack](#).

#### 21.1.2 Reporting bugs

Before reporting a bug please make sure to **reproduce it with the latest Patroni version!** Also please double check if the issue already exists in our [Issues Tracker](#).

#### 21.1.3 Running tests

Requirements for running behave tests:

1. PostgreSQL packages including `contrib` modules need to be installed.
2. PostgreSQL binaries must be available in your *PATH*. You may need to add them to the path with something like `PATH=/usr/lib/postgresql/11/bin:$PATH python -m behave`.
3. If you'd like to test with external DCSs (e.g., Etcd, Consul, and Zookeeper) you'll need the packages installed and respective services running and accepting unencrypted/unprotected connections on localhost and default port. In the case of Etcd or Consul, the behave test suite could start them up if binaries are available in the *PATH*.

Install dependencies:

```
# You may want to use Virtualenv or specify pip3.
pip install -r requirements.txt
pip install -r requirements.dev.txt
```

After you have all dependencies installed, you can run the various test suites:

```
# You may want to use Virtualenv or specify python3.

# Run flake8 to check syntax and formatting:
python setup.py flake8
```

(continues on next page)

(continued from previous page)

```
# Run the pytest suite in tests/:
python setup.py test

# Moreover, you may want to run tests in different scopes for debugging purposes,
# the -s option include print output during test execution.
# Tests in pytest typically follow the pattern: FILEPATH::CLASSNAME::TESTNAME.
pytest -s tests/test_api.py
pytest -s tests/test_api.py::TestRestApiHandler
pytest -s tests/test_api.py::TestRestApiHandler::test_do_GET

# Run the behave (https://behave.readthedocs.io/en/latest/) test suite in features/;
# modify DCS as desired (raft has no dependencies so is the easiest to start with):
DCS=raft python -m behave
```

### 21.1.4 Testing with tox

To run tox tests you only need to install one dependency (other than Python)

```
pip install tox>=4
```

If you wish to run *behave* tests then you also need docker installed.

Tox configuration in *tox.ini* has “environments” to run the following tasks:

- lint: Python code lint with *flake8*
- test: unit tests for all available python interpreters with *pytest*, generates XML reports or HTML reports if a TTY is detected
- dep: detect package dependency conflicts using *pipdeptree*
- type: static type checking with *pyright*
- black: code formatting with *black*
- docker-build: build docker image used for the *behave* env
- docker-cmd: run arbitrary command with the above image
- docker-behave-etc: run tox for behave tests with above image
- py\*behave: run behave with available python interpreters (without docker, although this is what is called inside docker containers)
- docs: build docs with *sphinx*

## Running tox

To run the default env list; dep, lint, test, and docs, just run:

```
tox
```

The *test* envs can be run with the label *test*:

```
tox -m test
```

The *behave* docker tests can be run with the label *behave*:

```
tox -m behave
```

Similarly, docs has the label *docs*.

All other envs can be run with their respective env names:

```
tox -e lint
tox -e py39-test-lin
```

It is also possible to select partial env lists using *factors*. For example, if you want to run all envs for python 3.10:

```
tox -f py310
```

This is equivalent to running all the envs listed below:

```
$ tox -l -f py310
py310-test-lin
py310-test-mac
py310-test-win
py310-type-lin
py310-type-mac
py310-type-win
py310-behave-etcd-lin
py310-behave-etcd-win
py310-behave-etcd-mac
```

You can list all configured combinations of environments with tox (>=v4) like so

```
tox l
```

The envs *test* and *docs* will attempt to open the HTML output files when the job completes, if tox is run with an active terminal. This is intended to be for benefit of the developer running this env locally. It will attempt to run *open* on a mac and *xdg-open* on Linux. To use a different command set the env var *OPEN\_CMD* to the name or path of the command. If this step fails it will not fail the run overall. If you want to disable this facility set the env var *OPEN\_CMD* to the *: no-op* command.

```
OPEN_CMD=: tox -m docs
```

## Behave tests

Behave tests with *-m behave* will build docker images based on PG\_MAJOR version 11 through 16 and then run all behave tests. This can take quite a long time to run so you might want to limit the scope to a select version of Postgres or to a specific feature set or steps.

To specify the version of postgres include the full name of the dependent image build env that you want and then the behave env name. For instance if you want Postgres 14 use:

```
tox -e pg14-docker-build,pg14-docker-behave-etcd-lin
```

If on the other hand you want to test a specific feature you can pass positional arguments to behave. This will run the watchdog behave feature test scenario with all versions of Postgres.

```
tox -m behave -- features/watchdog.feature
```

Of course you can combine the two.

### 21.1.5 Contributing a pull request

1. Fork the repository, develop and test your code changes.
2. Reflect changes in the user documentation.
3. Submit a pull request with a clear description of the changes objective. Link an existing issue if necessary.

You'll get feedback about your pull request as soon as possible.

Happy Patroni hacking ;-)



## INDICES AND TABLES

- `genindex`
- `search`