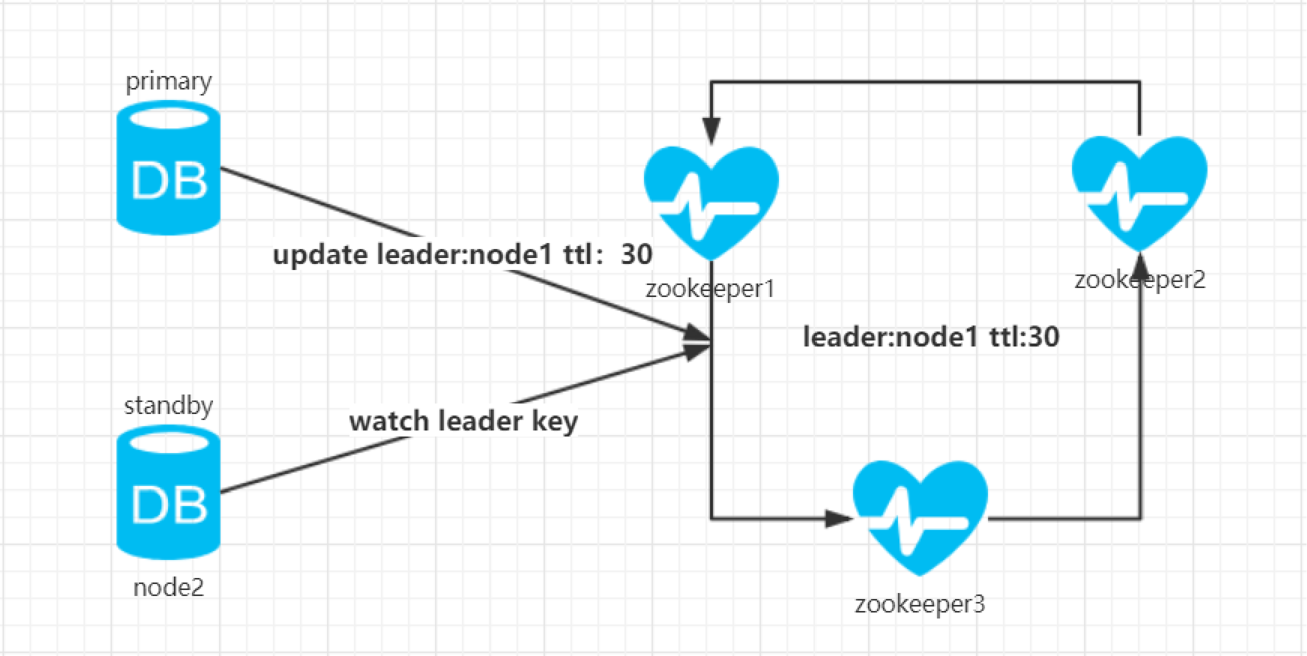
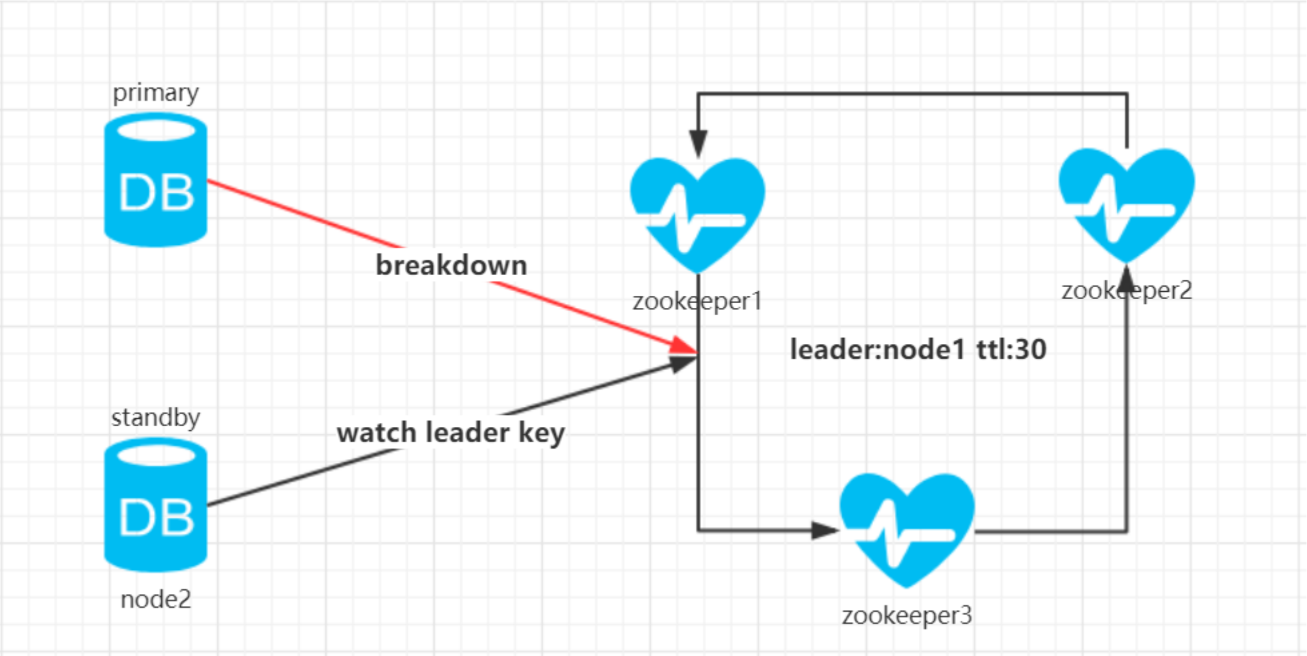
patroni原理分析

一、主库故障切换流程

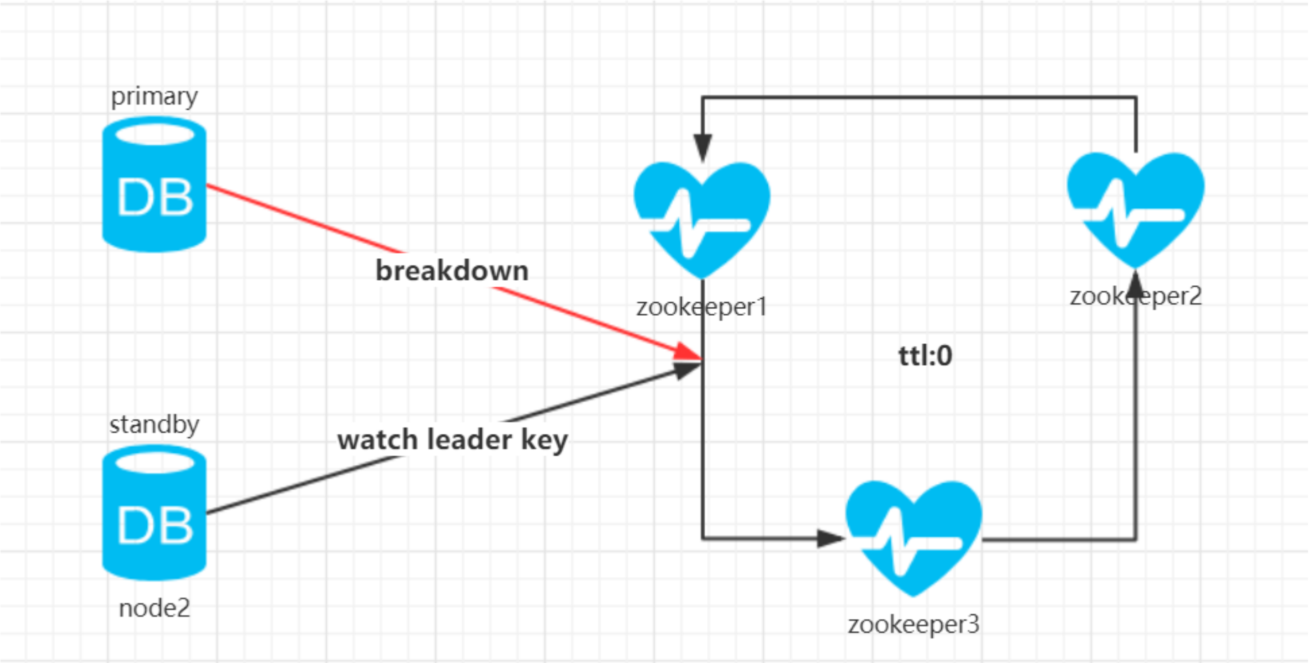
1、初始状态



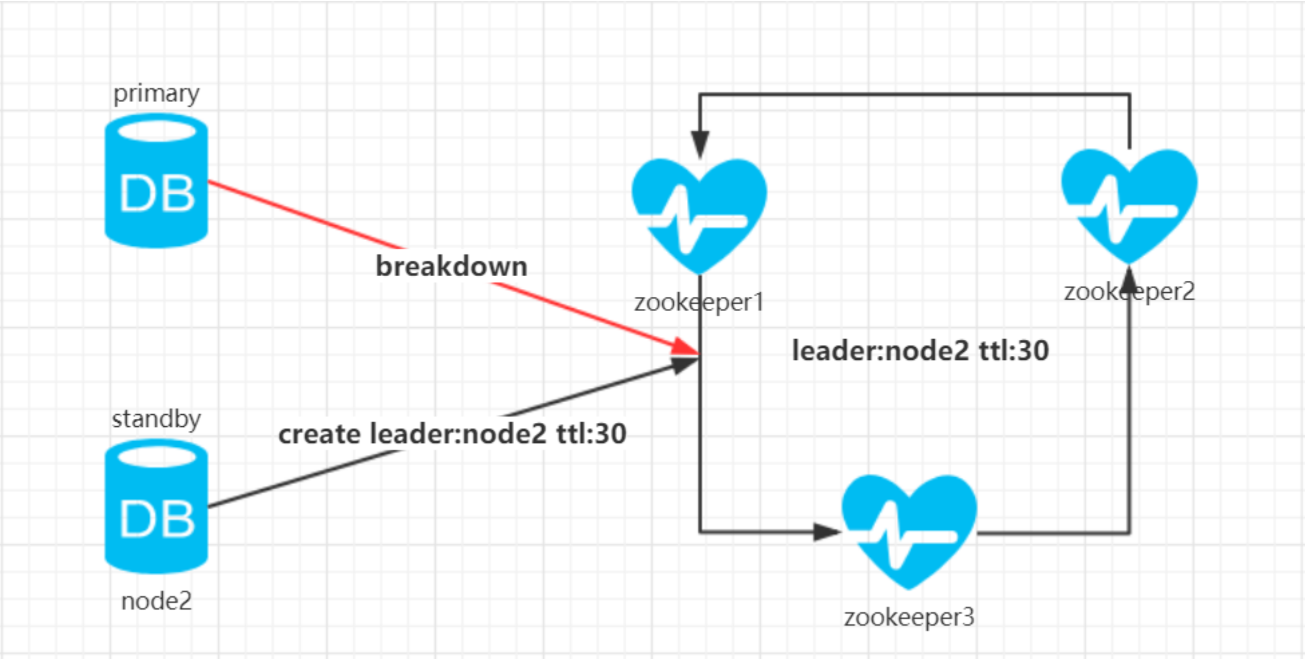
2、主库故障



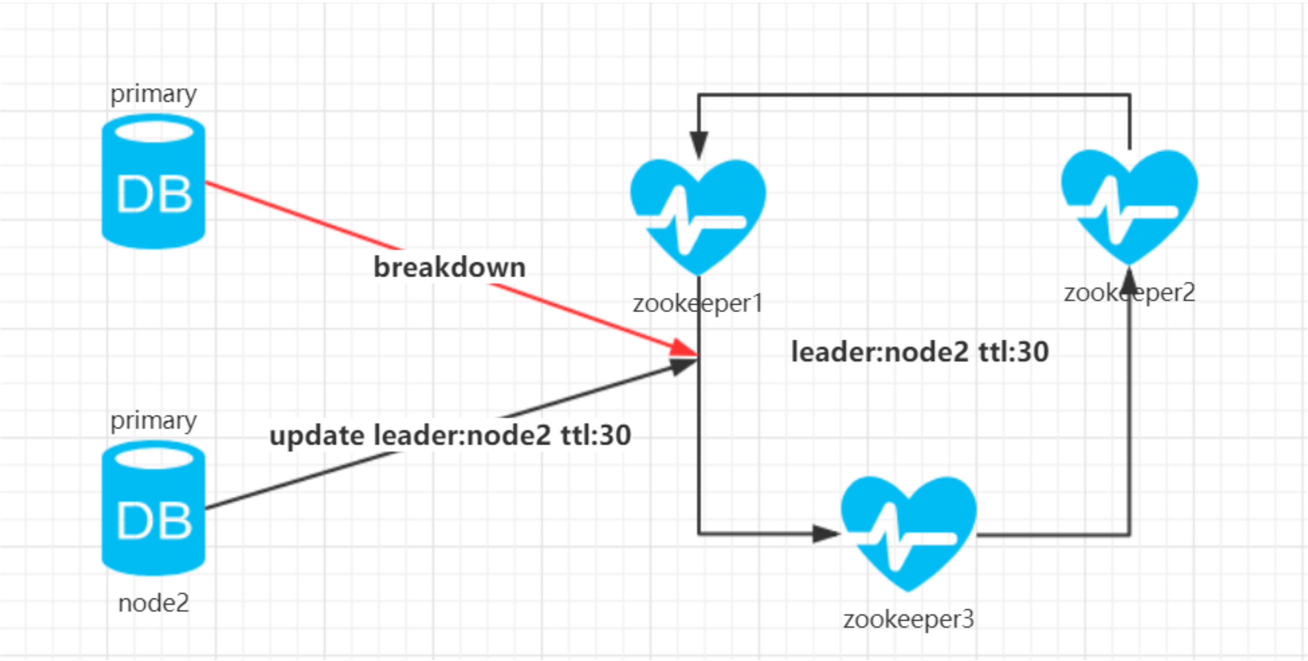
3、释放leader key



4、standby获得leader key



5、standby提升为primary



二、zookeeper中key值详解

相关key值如下：

```
[zk: localhost:2181(CONNECTED) 32] ls /service/batman
[leader, optime, failover, members, initialize, history, config, sync]
```

batman是patroni配置文件中scope的名字。

leader记录主节点的名字，是临时节点，当session时间超过ttl后未响应，zookeeper就会删除该节点。

```
[zk: localhost:2181(CONNECTED) 33] get /service/batman/leader
postgresql1
cZxid = 0x1000002ecc
ctime = Wed Nov 07 01:59:26 CST 2018
mZxid = 0x1000002ecc
mtime = Wed Nov 07 01:59:26 CST 2018
pZxid = 0x1000002ecc
cversion = 0
dataVersion = 0
aclVersion = 0
ephemeralOwner = 0x266e30f8ec70020
dataLength = 11
numChildren = 0
```

optime/leader是主库最后一次操作后的lsn位置，是持久节点不会因为session到期，删除该key值

```
[zk: localhost:2181(CONNECTED) 71] get /service/batman/optime/leader
1342177280
cZxid = 0x20000000e
ctime = Wed Oct 31 08:05:45 CST 2018
mZxid = 0x1000003320
mtime = Wed Nov 07 06:59:33 CST 2018
pZxid = 0x20000000e
cversion = 0
dataVersion = 269
aclVersion = 0
ephemeralOwner = 0x0
dataLength = 10
numChildren = 0
```

failover是记录计划的切换任务，是持久节点不会因为session到期，删除该key值

```
[postgres@node2 patroni-1.5.0]$ patronictl -c postgres0.yml switchover

Master [postgresql1]:
Candidate ['postgresql0'] []:
When should the switchover take place (e.g. 2015-10-01T14:30)  [now]: 2018-11-09 14:30
Current cluster topology
+---------+-------------+--------------+--------------+---------+-----------+
| Cluster |    Member   |     Host     |     Role     |  State  | Lag in MB |
+---------+-------------+--------------+--------------+---------+-----------+
|  batman | postgresql0 | 192.168.56.5 | Sync standby | running |           |
|  batman | postgresql1 | 192.168.56.4 |    Leader    | running |           |
+---------+-------------+--------------+--------------+---------+-----------+
Are you sure you want to switchover cluster batman, demoting current master
postgresql1? [y/N]: y
2018-11-07 07:15:54.59480 Switchover scheduled
+---------+-------------+--------------+--------------+---------+-----------+
| Cluster |    Member   |     Host     |     Role     |  State  | Lag in MB |
+---------+-------------+--------------+--------------+---------+-----------+
|  batman | postgresql0 | 192.168.56.5 | Sync standby | running |           |
|  batman | postgresql1 | 192.168.56.4 |    Leader    | running |           |
+---------+-------------+--------------+--------------+---------+-----------+
 Switchover scheduled at: 2018-11-09T14:30:00+08:00
                    from: postgresql1

[zk: localhost:2181(CONNECTED) 117] get /service/batman/failover
{"leader":"postgresql1","scheduled_at":"2018-11-09T14:30:00+08:00"}
cZxid = 0x20000014a
ctime = Wed Oct 31 08:34:15 CST 2018
mZxid = 0x10000033db
mtime = Wed Nov 07 07:15:54 CST 2018
pZxid = 0x20000014a
cversion = 0
dataVersion = 62
aclVersion = 0
ephemeralOwner = 0x0
dataLength = 67
```

```
numChildren = 0
```

members是分别记录了所有数据节点的连接信息和重要的状态信息，是临时节点，当session时间超过ttl后未响应，zookeeper就会删除该节点。

```
[zk: localhost:2181(CONNECTED) 120] get /service/batman/members/postgresql0
{"conn_url":"postgres://192.168.56.5:5432/postgres","api_url":"http://192.168.56.5:8008
/patroni","timeline":58,"state":"running","role":"replica","xlog_location":1426063952}
cZxid = 0x10000033c0
ctime = Wed Nov 07 07:13:56 CST 2018
mZxid = 0x10000033d5
mtime = Wed Nov 07 07:14:22 CST 2018
pZxid = 0x10000033c0
cversion = 0
dataVersion = 8
aclVersion = 0
ephemeralOwner = 0x366e30f8f100029
dataLength = 173
numChildren = 0

[zk: localhost:2181(CONNECTED) 121] get /service/batman/members/postgresql1
{"conn_url":"postgres://192.168.56.4:5432/postgres","api_url":"http://192.168.56.4:8009
/patroni","timeline":58,"state":"running","role":"master","xlog_location":1426063952}
cZxid = 0x10000033c4
ctime = Wed Nov 07 07:13:58 CST 2018
mZxid = 0x10000033d7
mtime = Wed Nov 07 07:14:23 CST 2018
pZxid = 0x10000033c4
cversion = 0
dataVersion = 4
aclVersion = 0
ephemeralOwner = 0x166e30f8ebb002a
dataLength = 172
numChildren = 0
```

initialize记录了数据库集群初始化的信息，是持久节点不会因为session到期，删除该key值

```
[zk: localhost:2181(CONNECTED) 123] get /service/batman/initialize
6618183861621602635
cZxid = 0x200000007
ctime = Wed Oct 31 08:05:35 CST 2018
mZxid = 0x200000007
mtime = Wed Oct 31 08:05:35 CST 2018
pZxid = 0x200000007
cversion = 0
dataVersion = 0
aclVersion = 0
ephemeralOwner = 0x0
dataLength = 19
numChildren = 0

[postgres@node2 ~]$ pg_controldata
```

```
pg_control version number:            1100
Catalog version number:               201809051
Database system identifier:           6618183861621602635
Database cluster state:               in archive recovery
pg_control last modified:             2018年11月07日 星期三 07时19分11秒
Latest checkpoint location:           0/550001A8
Latest checkpoint's REDO location:    0/55000170
Latest checkpoint's REDO WAL file:    0000003A0000000000000055
Latest checkpoint's TimeLineID:       58
Latest checkpoint's PrevTimeLineID:   58
Latest checkpoint's full_page_writes: on
Latest checkpoint's NextXID:          0:593
Latest checkpoint's NextOID:          16406
Latest checkpoint's NextMultiXactId:  1
Latest checkpoint's NextMultiOffset:  0
Latest checkpoint's oldestXID:        561
Latest checkpoint's oldestXID's DB:   1
Latest checkpoint's oldestActiveXID:  593
Latest checkpoint's oldestMultiXid:   1
Latest checkpoint's oldestMulti's DB: 1
Latest checkpoint's oldestCommitTsXid:0
Latest checkpoint's newestCommitTsXid:0
Time of latest checkpoint:            2018年11月07日 星期三 07时13分52秒
Fake LSN counter for unlogged rels:   0/1
Minimum recovery ending location:     0/55000218
Min recovery ending loc's timeline:   58
Backup start location:                0/0
Backup end location:                  0/0
End-of-backup record required:        no
```

history记录的是集群中时间线变化的过程，是持久节点不会因为session到期，删除该key值

```
[zk: localhost:2181(CONNECTED) 26] get /service/batman/history
[[1,67109464,"no recovery target specified"],[2,83886232,"no recovery target
specified"],[3,100663448,"no recovery target specified"],[4,218103960,"no recovery
target specified"],[5,251658392,"no recovery target specified"],[6,268435608,"no
recovery target specified"],[7,318767256,"no recovery target specified"],
[8,335544472,"no recovery target specified"],[9,352321688,"no recovery target
specified"],[10,369098904,"no recovery target specified"],[11,402653336,"no recovery
target specified"],[12,419430552,"no recovery target specified"],[13,436207768,"no
recovery target specified"],[14,452984984,"no recovery target specified"],
[15,469762200,"no recovery target specified"],[16,486539416,"no recovery target
specified"],[17,503316632,"no recovery target specified"],[18,536871064,"no recovery
target specified"],[19,553648280,"no recovery target specified"],[20,570425496,"no
recovery target specified"],[21,603979928,"no recovery target specified"],
[22,620757144,"no recovery target specified"],[23,637534360,"no recovery target
specified"],[24,654311576,"no recovery target specified"],[25,671088792,"no recovery
target specified","2018-11-05T16:44:52+08:00"],[26,704643224,"no recovery target
specified","2018-11-05T16:49:12+08:00"],[27,721420440,"no recovery target
specified","2018-11-05T23:21:12+08:00"],[28,771752088,"no recovery target
specified","2018-11-06T04:32:40+08:00"],[29,805306520,"no recovery target
specified","2018-11-06T05:22:45+08:00"],[30,822083736,"no recovery target
specified","2018-11-06T05:23:26+08:00"],[31,838860952,"no recovery target
specified","2018-11-06T05:44:02+08:00"],[32,855638168,"no recovery target
specified","2018-11-06T05:44:53+08:00"],[33,872415384,"no recovery target
specified","2018-11-06T05:46:17+08:00"],[34,889192600,"no recovery target
specified","2018-11-06T05:46:53+08:00"],[35,905969816,"no recovery target
specified","2018-11-06T05:49:39+08:00"],[36,922747032,"no recovery target
specified","2018-11-06T05:50:09+08:00"],[37,939524248,"no recovery target
specified","2018-11-06T05:51:36+08:00"],[38,956301464,"no recovery target
specified","2018-11-06T05:52:02+08:00"],[39,973078680,"no recovery target
specified","2018-11-06T06:00:37+08:00"],[40,989855896,"no recovery target
specified","2018-11-06T06:02:05+08:00"],[41,1006633112,"no recovery target
specified","2018-11-06T06:02:41+08:00"],[42,1023410328,"no recovery target
specified","2018-11-06T06:04:40+08:00"],[43,1040187544,"no recovery target
specified","2018-11-06T06:06:14+08:00"],[44,1090519192,"no recovery target
specified","2018-11-06T07:39:11+08:00"],[45,1107296408,"no recovery target
specified","2018-11-06T07:40:08+08:00"],[46,1140850840,"no recovery target
specified","2018-11-07T00:48:01+08:00"],[47,1157628056,"no recovery target
specified","2018-11-07T00:49:55+08:00"],[48,1224736920,"no recovery target
specified","2018-11-07T01:10:02+08:00"],[49,1241514136,"no recovery target
specified","2018-11-07T01:11:15+08:00"],[50,1241514632,"no recovery target
specified","2018-11-07T01:13:02+08:00"],[51,1275068568,"no recovery target
specified","2018-11-07T01:58:22+08:00"],[52,1275068896,"no recovery target
specified","2018-11-07T01:59:17+08:00"],[53,1358954648,"no recovery target
specified","2018-11-07T07:08:15+08:00"],[54,1375731864,"no recovery target
specified","2018-11-07T07:08:51+08:00"],[55,1392509080,"no recovery target
specified","2018-11-07T07:12:20+08:00"],[56,1409286296,"no recovery target
specified","2018-11-07T07:12:39+08:00"],[57,1426063512,"no recovery target
specified","2018-11-07T07:13:51+08:00"],[58,1442840728,"no recovery target
specified","2018-11-07T07:26:34+08:00"]]
cZxid = 0x200000159
ctime = Wed Oct 31 08:34:17 CST 2018
mZxid = 0x10000033eb
mtime = Wed Nov 07 07:26:34 CST 2018
```

```
pZxid = 0x200000159
cversion = 0
dataVersion = 60
aclVersion = 0
ephemeralOwner = 0x0
dataLength = 3628
numChildren = 0
```

config记录的是patroni配置文件的配置信息，是持久节点不会因为session到期，删除该key值

```
[zk: localhost:2181(CONNECTED) 28] get /service/batman/config
{"retry_timeout": 10, "postgresql": {"use_slots": true, "use_pg_rewind": true,
"parameters": {"hot_standby": "on", "wal_keep_segments": 8, "wal_level": "hot_standby",
"archive_command": "mkdir -p ../wal_archive && test ! -f ../wal_archive/%f && cp %p
../wal_archive/%f", "wal_log_hints": "on", "max_wal_senders": 10, "archive_timeout":
2000, "archive_mode": "on", "max_replication_slots": 10, "max_connections": 300},
"recovery_conf": {"restore_command": "cp ../wal_archive/%f %p"}}, "synchronous_mode":
true, "maximum_lag_on_failover": 1048576, "loop_wait": 10, "max_connection": 300,
"archive_timeout": "2000s", "ttl": 30, "max_connections": 300}
cZxid = 0x20000000a
ctime = Wed Oct 31 08:05:45 CST 2018
mZxid = 0x300000fec
mtime = Thu Nov 01 23:10:36 CST 2018
pZxid = 0x20000000a
cversion = 0
dataVersion = 26
aclVersion = 0
ephemeralOwner = 0x0
dataLength = 648
numChildren = 0
```

sync记录同步复制的状态，是持久节点不会因为session到期，删除该key值

```
[zk: localhost:2181(CONNECTED) 30] get /service/batman/sync
{"leader":"postgresql0","sync_standby":"postgresql1"}
cZxid = 0x3000002d5
ctime = Thu Nov 01 14:29:04 CST 2018
mZxid = 0x10000033ef
mtime = Wed Nov 07 07:26:47 CST 2018
pZxid = 0x3000002d5
cversion = 0
dataVersion = 145
aclVersion = 0
ephemeralOwner = 0x0
dataLength = 53
numChildren = 0
```

三、ha循环流程图

https://raw.githubusercontent.com/zalando/patroni/master/docs/ha_loop_diagram.png