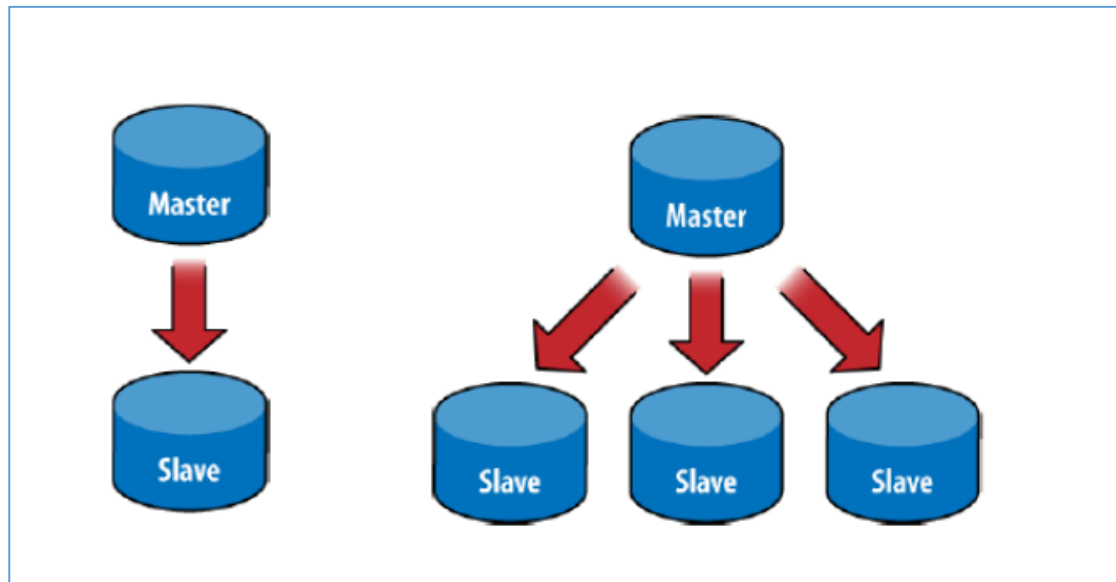


# 第一章 master-slave 集群

## 1.1 拓扑图



主从架构一般用于备份或者做读写分离。一般有一主一从或一主多从设计。

优点：

- 1) 主节点，可读可写，从节点可读不可写，实现了读写分离
- 2) 支持一主多从，大大缓解读的压力，适用于读操作比较多的场景

缺点：

- 1) 不能实现自动主从切换，即主挂后，从不能自动成为主，集群只可以读，不可写
- 2) 不支持链式结构，即 **Slave** 只能直接连 **Master**。而不像 **redis** 或者 **mysql** 的，**slave** 可从另一个 **slave** 同步。
- 3) 数据量大的情况下，主节点可能成为性能瓶颈

## 1.2 测试规划

### 1.2.1 规划清单

**OS:** CentOS 6.6 x64

**MongoDB:** 3.6.7 (4.0 之后不再支持主从模式，后面有介绍)

**Master:** 10.1.5.201

**Slave1:** 10.1.5.202

Slave2:10.1.5.203

数据目录: /mongodb-data/ (自定义)

日志文件: /var/log/mongodb/mongod.log (默认)

PID 文件: /var/run/mongodb/mongod.pid (默认)

PORT: 27017 (默认)

监听地址: 0.0.0.0 (自定义)

### 1.2.2 版本兼容

注意: mongodb4.0.1 不再支持主从模式的集群, 可见发行说明文档:

<https://docs.mongodb.com/manual/release-notes/4.0-compatibility/>

#### Remove Master-Slave Replication

MongoDB 4.0 removes support for the deprecated master-slave replication. Before you can upgrade to MongoDB 4.0, if your deployment uses master-slave replication, you must upgrade to a replica set.

安装了 4 之后, 通过查询帮助, 也得到类似的结论:

```
[root@hadoop01 ~]#  
[root@hadoop01 ~]# mongod --version  
db version v4.0.1  
git version: 54f1582fc6eb01de4d4c42f26fc133e623f065fb  
OpenSSL version: OpenSSL 1.0.1e-fips 11 Feb 2013  
allocator: tcmalloc  
modules: none  
build environment:  
  distmod: rhel62  
  distarch: x86_64  
  target_arch: x86_64  
[root@hadoop01 ~]#  
[root@hadoop01 ~]#  
[root@hadoop01 ~]# mongod -h | grep -A3 master  
--master                Master/slave replication no longer  
                        supported  
--slave                  Master/slave replication no longer  
                        supported  
[root@hadoop01 ~]#
```

## 1.3 安装和部署

master、slave1、slave2 上需要执行如下步骤。

### 1.3.1 添加 yum 源

```
[root@hadoop01 ~]# cat /etc/yum.repos.d/Mongodb-3.6.repo
```

```
[mongodb-org-3.6]
```

```
name = MongoDB Repository
```

```
baseurl = http://repo.mongodb.org/yum/redhat/$releasever/mongodb-org/3.6/x86_64/
```

```
gpgcheck = 0
```

```
enabled = 1
```

1.3.2 安装依赖包

```
yum install openssl libcurl
```

1.3.3 安装软件包

```
yum install mongodb-org
```

将自动安装以下 4 个依赖包：  
mongodb-org-server 包含 [mongodb](#) 守护程序以及关联的配置和 init 脚本。  
mongodb-org-mongos 包含 [mongos](#) 守护进程。  
mongodb-org-shell 包含 [mongo](#) shell。  
mongodb-org-tools 包含以下的 MongoDB 工具：  
[mongoimport](#), [bsondump](#), [mongodump](#), [mongoexport](#), [mongofiles](#),  
[mongoperf](#), [mongorestore](#), [mongostat](#) and [mongotop](#).

```

依赖关系解决
=====
软件包                                架构                                版本
-----
正在安装:
mongodb-org                          x86_64                              3.6.7-1.el6
为依赖而安装:
mongodb-org-mongos                   x86_64                              3.6.7-1.el6
mongodb-org-server                   x86_64                              3.6.7-1.el6
mongodb-org-shell                    x86_64                              3.6.7-1.el6
mongodb-org-tools                    x86_64                              3.6.7-1.el6

事务概要
=====
Install                               5 Package(s)

总下载量: 91 M
Installed size: 265 M
确定吗? [y/N]: y

```

1.3.4 数据目录准备

创建数据目，并修改权限

```
mkdir /mongodb-data  
chown mongod.mongod /mongodb-data/
```

1.4 修改配置文件

1.4.1 /etc/mongod.conf

master、slave1、slave2 上需要执行这一步。

默认配置文件内容如下：

```
# where to write logging data.
systemLog:
  destination: file
  logAppend: true
  path: /var/log/mongodb/mongod.log

# where and how to store data.
storage:
  dbPath: /var/lib/mongo
  journal:
    enabled: true
# engine:
# mmapv1:
# wiredTiger:

# how the process runs
processManagement:
  fork: true # fork and run in background
  pidFilePath: /var/run/mongodb/mongod.pid # location of
  timeZoneInfo: /usr/share/zoneinfo

# network interfaces
net:
  port: 27017
  bindIp: 127.0.0.1 # Enter 0.0.0.0,:: to bind to all IPv
```

配置文件使用分段形式配置，我们修改 dbpath 和 bindIP 选项，可使用如下语句快速修改：

```
sed -i 's#dbPath: /var/lib/mongo#dbPath: /mongodb-data#g' /etc/mongod.conf
sed -i 's#bindIp: 127.0.0.1#bindIp: 0.0.0.0#g' /etc/mongod.conf
```

#### 1.4.2 /etc/init.d/mongod

master 节点执行：

```
vim /etc/init.d/mongod
```

```
# NOTE: if you change any OPTIONS here, you get what you pay for
# this script assumes all options are in the config file.
CONFIGFILE="/etc/mongod.conf"
OPTIONS="-f $CONFIGFILE"
mongod=${MONGOD-/usr/bin/mongod}
```

修改为：

```
# NOTE: if you change any OPTIONS here, you get what you pay for
# this script assumes all options are in the config file.
CONFIGFILE="/etc/mongod.conf"
OPTIONS="-f $CONFIGFILE --slave --source 10.1.5.201:27017"
mongod=${MONGOD-/usr/bin/mongod}
```

可使用如下语句，快速修改：

```
sed -i 's/OPTIONS="-f $CONFIGFILE"/OPTIONS="-f $CONFIGFILE --slave --source 10.1.5.201:27017"/g' /etc/init.d/mongod
```

slave1、slave2 节点执行：

```
# NOTE: if you change any OPTIONS here, you get what you pay for
# this script assumes all options are in the config file.
CONFIGFILE="/etc/mongod.conf"
OPTIONS="-f $CONFIGFILE"
mongod=${MONGOD:-/usr/bin/mongod}
```

修改为:

```
# NOTE: if you change any OPTIONS here, you get what you pay for
# this script assumes all options are in the config file.
CONFIGFILE="/etc/mongod.conf"
OPTIONS="-f $CONFIGFILE --master"
mongod=${MONGOD:-/usr/bin/mongod}
```

可使用如下语句, 快速修改:

```
sed 's/OPTIONS="-f $CONFIGFILE"/OPTIONS="-f $CONFIGFILE --master"/g'
/etc/init.d/mongod
```

## 1.5 启动服务

master、slave1、slave2 上需要执行这一步, 其中 master 节点需要先启动。

```
service mongod start
chkconfig mongod on #添加开机启动, 默认已开启
```

## 1.6 测试

### 1.6.1 观察日志

less /var/log/mongodb/mongod.log

```
2018-08-22T11:51:32.874+0800 I CONTROL [initandlisten] target_arch: x86_64
2018-08-22T11:51:32.874+0800 I CONTROL [initandlisten] options: { config: "/etc/mongod.conf", net: { bindIp: "0.0.0.0", port: 27017 }, processManagement: { fork: true,
pidFilePath: "/var/run/mongodb/mongod.pid", timeZoneInfo: "/usr/share/zoneinfo" }, slave: true, source: "10.1.5.201:27017", storage: { dbPath: "/mongodb-data", jour
nal: { enabled: true } }, systemLog: { destination: "file", logAppend: true, path: "/var/log/mongodb/mongod.log" } }
2018-08-22T11:51:32.885+0800 I - [initandlisten] detected data files in /mongodb-data created by the 'wiredtiger' storage engine, so setting the active storage
engine to 'wiredtiger'.
2018-08-22T11:51:32.885+0800 I STORAGE [initandlisten] ** WARNING: Using the XFS filesystem is strongly recommended with the wiredTiger storage engine
2018-08-22T11:51:32.885+0800 I STORAGE [initandlisten] ** See http://dochub.mongodb.org/core/prodnotes-filesystem
2018-08-22T11:51:32.886+0800 I STORAGE [initandlisten] wiredtiger_open config: create,cache_size=140M,session_max=20000,eviction=(threads_min=4,threads_max=4),confi
base=false,statistics=(fast),cache_cursors=false,compatibility=(release="3.0",require_max="3.0"),log=(enabled=true,archive=true,path=journal,compressor=snappy),file
analyzer=(close_idle_time=10000),statistics_log=(wait=0),verbose=(recovery_progress)
2018-08-22T11:52:53.109+0800 I STORAGE [initandlisten] wiredtiger message [1534909973:109020][2417:0x7f67456bea80], txn-recover: Main recovery loop: starting at 1/26
68
2018-08-22T11:52:53.373+0800 I STORAGE [initandlisten] wiredtiger message [1534909973:373653][2417:0x7f67456bea80], txn-recover: Recovering log 1 through 2
2018-08-22T11:52:53.811+0800 I STORAGE [initandlisten] wiredtiger message [1534909973:811781][2417:0x7f67456bea80], txn-recover: Recovering log 2 through 2
2018-08-22T11:52:53.878+0800 I STORAGE [initandlisten] wiredtiger message [1534909973:878783][2417:0x7f67456bea80], txn-recover: Set global recovery timestamp: 0
2018-08-22T11:52:54.186+0800 I CONTROL [initandlisten] ** WARNING: Access control is not enabled for the database.
2018-08-22T11:52:54.186+0800 I CONTROL [initandlisten] ** Read and write access to data and configuration is unrestricted.
2018-08-22T11:52:54.187+0800 I CONTROL [initandlisten]
2018-08-22T11:52:54.197+0800 I CONTROL [initandlisten] ** WARNING: /sys/kernel/mm/transparent_hugepage/enabled is 'always'.
2018-08-22T11:52:54.197+0800 I CONTROL [initandlisten] ** We suggest setting it to 'never'.
2018-08-22T11:52:54.197+0800 I CONTROL [initandlisten]
2018-08-22T11:52:54.197+0800 I CONTROL [initandlisten] ** WARNING: /sys/kernel/mm/transparent_hugepage/defrag is 'always'.
2018-08-22T11:52:54.197+0800 I CONTROL [initandlisten] ** We suggest setting it to 'never'.
2018-08-22T11:52:54.198+0800 I CONTROL [initandlisten]
2018-08-22T11:52:54.198+0800 I CONTROL [initandlisten] ** WARNING: soft rlimits too low. rlimits set to 1024 processes, 64000 files. Number of processes should be at
least 32000 : 0.5 times number of files.
2018-08-22T11:52:54.198+0800 I CONTROL [initandlisten]
2018-08-22T11:52:54.198+0800 I CONTROL [initandlisten] ** WARNING: This mode was started in master-slave replication mode.
2018-08-22T11:52:54.198+0800 I CONTROL [initandlisten] ** Master-slave replication is deprecated and subject to be removed
2018-08-22T11:52:54.198+0800 I CONTROL [initandlisten] ** in a future version.
```

从启动日志中我们可以得到如下信息:

1) 启动参数, 比如 slave 的参数

```
2018-08-22T11:51:32.874+0800 I CONTROL [initandlisten] options: { config: "/etc/mongod.conf",
net: { bindIp: "0.0.0.0", port: 27017 }, processManagement: { fork: true, pidFilePath: "/var/
run/mongodb/mongod.pid", timeZoneInfo: "/usr/share/zoneinfo" }, slave: true, source: "10.1.5.2
01:27017", storage: { dbPath: "/mongodb-data", journal: { enabled: true } }, systemLog: { dest
ination: "file", logAppend: true, path: "/var/log/mongodb/mongod.log" } }
```

```
options:
{
```

```

config: "/etc/mongod.conf",
net:
  {
    bindIp: "0.0.0.0",
    port: 27017
  },
processManagement:
  {
    fork: true,
    pidFilePath: "/var/run/mongodb/mongod.pid",
    timeZoneInfo: "/usr/share/zoneinfo"
  },
slave: true,
source: "10.1.5.201:27017",
storage:
  {
    dbPath: "/mongodb-data",
    journal: { enabled: true }
  },
systemLog:
  {
    destination: "file",
    logAppend: true,
    path: "/var/log/mongodb/mongod.log"
  }
}

```

- 2) 警告 1, 告诉我们最好使用 XFS 的文件系统, 这个我们是测试环境, 暂时忽略, 但生产环境中, 建议格式化成 XFS, 可以提高性能

```

** WARNING: Using the XFS filesystem is strongly recommended with the WiredTiger storage engine
** See http://dochub.mongodb.org/core/prodnotes-filesystem

```

- 3) 警告 2, 告诉我们没有启用访问控制, 这个自己评估是否需要, 此处测试, 不再启用

```

WARNING: Access control is not enabled for the database.
Read and write access to data and configuration is unrestricted.

```

- 4) 警告 3, 告诉我们启用了大内存页功能, 而 mongodb 的运行环境建议应该取消, 这可以提高性能

```

WARNING: /sys/kernel/mm/transparent_hugepage/enabled is 'always'.
we suggest setting it to 'never'

```

如果需要修复, 可以执行如下语句:

```

echo never > /sys/kernel/mm/transparent_hugepage/enabled
echo never > /sys/kernel/mm/transparent_hugepage/defrag

```

并将其加入/etc/rc.local

- 5) 警告 5 rlimit 被设置为了 1024, 太小了对于 mongodb 来说

```

2018-08-22T11:52:54.198+0800 I CONTROL [initandlisten]
2018-08-22T11:52:54.198+0800 I CONTROL [initandlisten] ** WARNING: soft rlimits too low. rlimits set to 1024
processes, 64000 files. Number of processes should be at least 32000 : 0.5 times number of files.

```

如果要修复，可以使用如下语句：

```
echo "mongod soft nproc 65535" >> /etc/security/limits.conf
echo "mongod hard nproc 65535" >> /etc/security/limits.conf
```

并重新 ssh 服务器后并重启 mongod 服务。

6) 警告 6，告诉我们 master-slave 模式将来会取消支持，所以知道在 4.0 及其之后已经被取消了

```
WARNING: This node was started in master-slave replication mode.
Master-slave replication is deprecated and subject to be removed
in a future version.
```

### 1.6.2 观察主从复制

master 节点创建库，并插入一条数据：

```
[root@hadoop01 ~]# mongo 10.1.5.201:27017
MongoDB shell version v3.6.7
connecting to: mongodb://10.1.5.201:27017/test
MongoDB server version: 3.6.7
```

使用 mongos 连接 master

```
> db.printReplicationInfo()
configured oplog size: 2014.310791015625MB
log length start to end: 82072secs (22.8hrs)
oplog first event time: Tue Aug 21 2018 16:10:30 GMT+0800 (CST)
oplog last event time: Wed Aug 22 2018 14:58:22 GMT+0800 (CST)
now: Wed Aug 22 2018 14:58:28 GMT+0800 (CST)
>
> show dbs
admin 0.000GB
config 0.000GB
local 0.000GB
>
> use testdb
switched to db testdb
>
> db.testdb.insert({"name":"quzl"})
writeResult({ "nInserted" : 1 })
>
> show dbs
admin 0.000GB
config 0.000GB
local 0.000GB
testdb 0.000GB
```

命令解释：

db.printReplicationInfo()，查看主节点信息

oplog first event time：首次同步时间

oplog last event time：最后一次同步时间

show dbs，显示当前所有库

admin、config、local 是系统自带数据库

use testdb，如果不存在就创建该库，如果存在就进入该库

db.testdb.insert({"name":"quzl"})，在 testdb 库中插入一条文档数据

61617 成为 master 后，使用 8162 端口访问查看，发现队列仍然存在。

进入 slave 库查看数据是否被同步：

```
[root@hadoop01 ~]# mongo 10.1.5.202:27017
MongoDB shell version v3.6.7
connecting to: mongod://10.1.5.202:27017/test
MongoDB server version: 3.6.7
```

进入 slave1 节点

```
> show dbs
2018-08-22T15:09:11.686+0800 E QUERY [thread1] Error: listDatabases failed:{
  "ok" : 0,
  "errmsg" : "not master and slaveok=false",
  "code" : 13435,
  "codeName" : "NotMasterNoSlaveOk"
} :
_getErrorWithCode@src/mongo/shell/utils.js:25:13
Mongo.prototype.getDBs@src/mongo/shell/mongo.js:65:1
shellHelper.show@src/mongo/shell/utils.js:849:19
shellHelper@src/mongo/shell/utils.js:739:15
@(shellhelp2):1:1
>
> rs.slaveOk()
>
> show dbs
admin 0.000GB
config 0.000GB
local 0.000GB
testdb 0.000GB
```

第一次使用 show dbs，报 12435 的错误，那是因为 slave 库，默认是不可读的，使用 rs.slaveOk() 设置后就可以读取了。


```
> rs.printSlaveReplicationInfo()
source: 10.1.5.201:27017
  syncedTo: wed Aug 22 2018 15:24:52 GMT+0800 (CST)
  5 secs (0 hrs) behind the freshest member (no primary available at the moment)
>
```

使用 rs.printSlaveReplicationInfo() 查看从库，同步状态，即显示出上一次同步的时间。

**结论：主从复制正常**

### 1.6.3 从库不可写测试

```
> use testdb2
switched to db testdb2
>
>
> show dbs
admin 0.000GB
config 0.000GB
local 0.000GB
testdb 0.000GB
>
>
> db.testdb2.insert({"name":"guzl"})
writeResult({ "writeError" : { "code" : 10107, "errmsg" : "not master" } })
>
>
> show dbs
admin 0.000GB
config 0.000GB
local 0.000GB
testdb 0.000GB
>
```



如上图，从库不可写



#### 1.6.4 读写分离配置

**情况 1:** 使用 mongo shell 进入从库

使用如下命令中的任意一条, 即可实现从库可读, 但只对当前 shell 有效

- 1) `db.getMongo().setSlaveOk()`
- 2) `rs.slaveOk()`

**情况 2:** java 连接 mongodb:

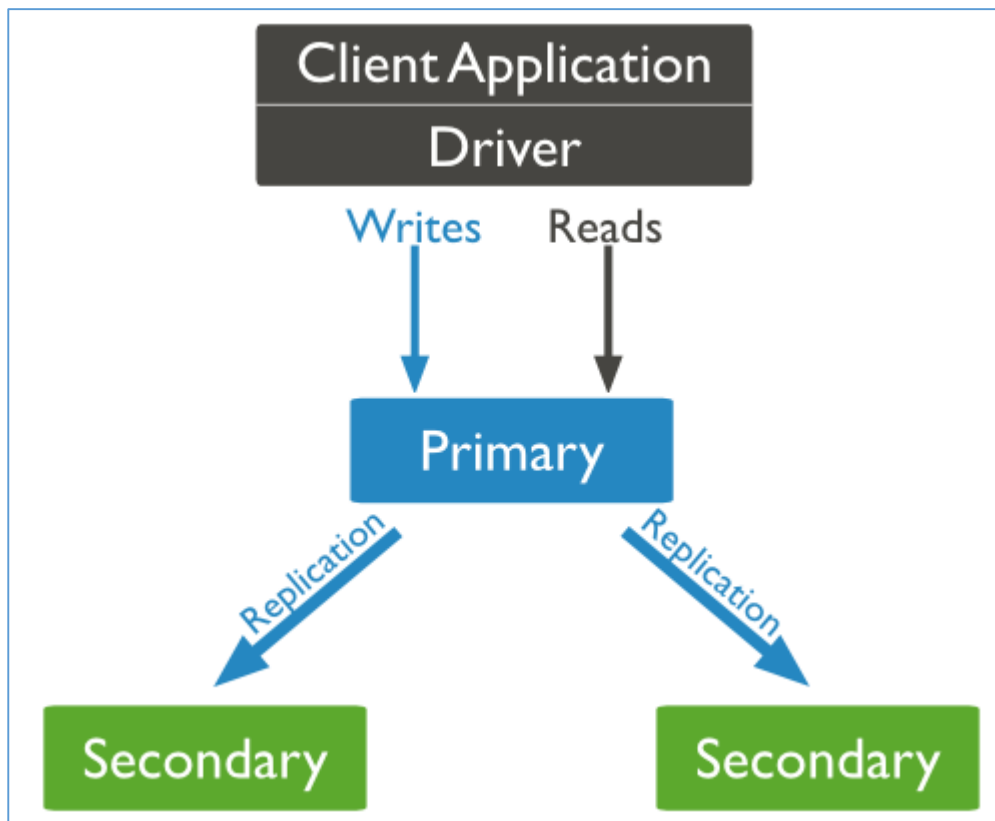
- 1) 在 java 代码中调用 `dbFactory.getDb().slaveOk()`;
- 2) 在 java 代码中调用  
`dbFactory.getDb().setReadPreference(ReadPreference.secondaryPreferred());`  
//在复制集中优先读 secondary, 如果 secondary 访问不了的时候就从 master 中读  
或  
`dbFactory.getDb().setReadPreference(ReadPreference.secondary());`  
//只从 secondary 中读, 如果 secondary 访问不了的时候就不能进行查询
- 3) 在配置 mongo 的时候增加 `slave-ok="true"` 也支持直接从 secondary 中读  

```
<mongo:mongo id="mongo" host="{mongodb.host}" port="{mongodb.port}">  
    <mongo:options slave-ok="true"/>  
</mongo:mongo>
```

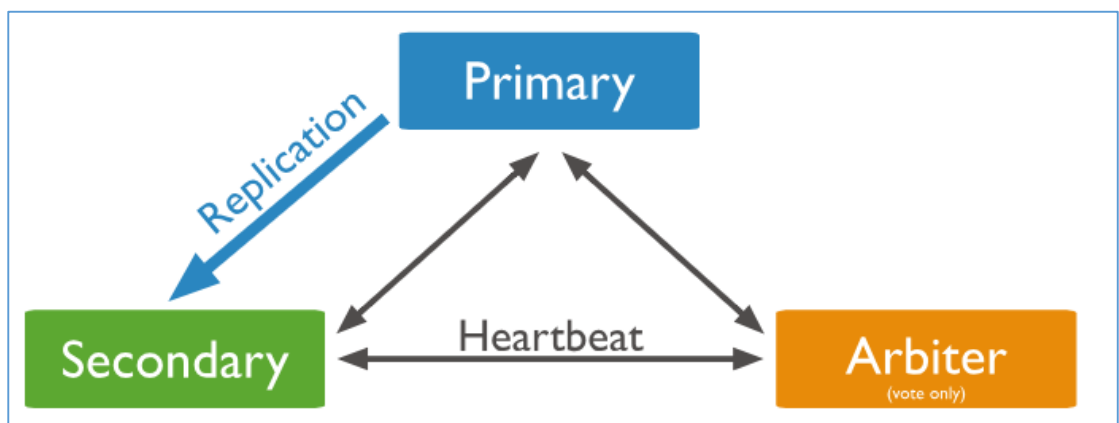
## 第二章 副本集集群

副本集集群，即 Replica Set 集群，可以是一主一从，也可以是一主多从，但为了保证有效性，整个集群节点个数最好是奇数个。

### 2.1 拓扑图



或者



(1) 主节点 (Primary)

接收所有的读、写请求，然后把修改同步到所有 **Secondary**。一个 **Replica Set** 只能有一个 **Primary** 节点，当 **Primary** 挂掉后，其他 **Secondary** 节点会重新选举出来一个成为主节点。

如果要将写请求发送到从节点，客户端可以指定读取首选项将读操作发送到辅助节点。对辅助节点的异步复制意味着从辅助节点读取可能会返回不反映主节点上数据状态的数据，即可能造成新增或修改的数据，无法获取到最新的状态。

## （2）副本节点（**Secondary**）

与主节点保持同样的数据集。当主节点挂掉的时候，参与选主。

## （3）仲裁者（**Arbiter**）

不保有数据，不参与选主，只进行选主投票。使用 **Arbiter** 可以减轻数据存储的硬件需求，**Arbiter** 跑起来几乎没什么大的硬件资源需求，如果您的副本集具有偶数个成员，请添加仲裁者以获得主要选举中的大多数投票，但重要的一点是，在生产环境下它和其他数据节点不要部署在同一台机器上。

优点：

- 1) 自动主从切换
- 2) 可配置读写分离

缺点：

- 1) 数据量大的情况下，主节点可能成为性能瓶颈

## 2.2 测试规划

这里采用 **Primary+ Secondary+ Arbiter** 的架构形式。

**OS:** CentOS 6.6 x64

**MongoDB:** 4.0.1

**Primary:** 10.1.5.201

**Secondary:** 10.1.5.202

**Arbiter:** 10.1.5.203

**数据目录:** /mongodb-data/ （自定义）

**日志文件:** /var/log/mongodb/mongod.log （默认）

**PID 文件:** /var/run/mongodb/mongod.pid （默认）

**PORT:** 27017 （默认）

**监听地址:** 0.0.0.0 （自定义）

## 2.3 安装和部署

本节 2.3 的所有操作，在 Primary、Secondary、Arbiter 三个节点都要执行。

### 2.3.1 添加 yum 源

```
[root@hadoop01 ~]# cat /etc/yum.repos.d/Mongodb-4.repo
[mongodb-org-4.0]
name = MongoDB Repository
baseurl = http://repo.mongodb.org/yum/redhat/$releasever/mongodb-org/4.0/x86_64/
gpgcheck = 0
enabled = 1
```

### 2.3.2 安装依赖包

```
yum install openssl libcurl
```

### 2.3.3 安装软件包

```
yum install mongodb-org
```

将自动安装以下 4 个依赖包：

mongodb-org-server 包含 [mongod](#) 守护程序以及关联的配置和 init 脚本。

mongodb-org-mongos 包含 [mongos](#) 守护进程。

mongodb-org-shell 包含 [mongo](#) shell。

mongodb-org-tools 包含以下的 MongoDB 工具：

[mongoimport](#), [bsondump](#), [mongodump](#), [mongoexport](#), [mongofiles](#),  
[mongorestore](#), [mongostat](#) and [mongotop](#)。

软件包	架构	版本
正在安装：		
mongodb-org	x86_64	4.0.1-1.el6
为依赖而安装：		
mongodb-org-mongos	x86_64	4.0.1-1.el6
mongodb-org-server	x86_64	4.0.1-1.el6
mongodb-org-shell	x86_64	4.0.1-1.el6
mongodb-org-tools	x86_64	4.0.1-1.el6
事务概要		
Install 5 Package(s)		
总下载量: 76 M		
Installed size: 232 M		
确定吗? [y/N]:		

### 2.3.4 数据目录准备

创建数据目，并修改权限

```
mkdir /mongodb-data
```

```
chown mongod.mongod /mongodb-data/
```

## 2.4 修改配置文件

本节 2.4 的所有操作，在 Primary、Secondary、Arbiter 三个节点都要执行。

### 2.4.1 /etc/mongod.conf

默认配置文件内容如下：

```
# where to write logging data.
systemLog:
  destination: file
  logAppend: true
  path: /var/log/mongodb/mongod.log

# where and how to store data.
storage:
  dbPath: /var/lib/mongo
  journal:
    enabled: true
# engine:
# mmapv1:
# wiredTiger:

# how the process runs
processManagement:
  fork: true # fork and run in background
  pidFilePath: /var/run/mongodb/mongod.pid # location of
  timeZoneInfo: /usr/share/zoneinfo

# network interfaces
net:
  port: 27017
  bindIp: 127.0.0.1 # Enter 0.0.0.0,:: to bind to all IPv4 and IPv6
```

配置文件使用分段形式配置，我们修改 dbpath 和 bindIP 选项，可使用如下语句快速修改：

```
sed -i 's#dbPath: /var/lib/mongo#dbPath: /mongodb-data#g' /etc/mongod.conf
sed -i 's#bindIp: 127.0.0.1#bindIp: 0.0.0.0#g' /etc/mongod.conf
```

### 2.4.2 /etc/mongod.conf

vim /etc/mongod.conf

```
# network interfaces
net:
  port: 27017
  bindIp: 0.0.0.0 # Enter 0.0.0.0,:: to bind to all IPv

#security:

#operationProfiling:

replication:
  replSetName: "my-rs1"

#sharding:
```

添加如下两行，其中 my-rs1 是副本集的名字，可以修改成其他的

```
replication:
  replSetName: "my-rs1"
```

## 2.5 启动集群

### 2.5.1 启动服务

在前面步骤都完成后，使用

```
service mongod start
```

启动三节点的 mongod 服务

### 2.5.2 初始化副本集-secondary 节点

使用 mongs 进入 201 节点，如果是本机，也可以直接输入 mongo

```
[root@hadoop01 ~]# mongo 10.1.5.201:27017
MongoDB shell version v3.6.7
connecting to: mongodb://10.1.5.201:27017/test
MongoDB server version: 3.6.7
```

使用如下命令初始化，其中注意 my-rs1 于 2.4.2 中配置的要一致，201、202、203 随意设置

```
rs.initiate( {
  _id : "my-rs1",
  members: [
    { _id: 201, host: "10.1.5.201:27017" },
    { _id: 202, host: "10.1.5.202:27017" }
  ]
})
```

```

> rs.initiate( {
...   _id : "my-rs1",
...   members: [
...     { _id: 201, host: "10.1.5.201:27017" },
...     { _id: 202, host: "10.1.5.202:27017" }
...   ]
... })
{
  "ok" : 1,
  "operationTime" : Timestamp(1534931616, 1),
  "$clusterTime" : {
    "clusterTime" : Timestamp(1534931616, 1),
    "signature" : {
      "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAA="),
      "keyId" : NumberLong(0)
    }
  }
}
my-rs1:SECONDARY>

```

输入

rs.conf()

可以查看副本集配置信息，目前只有两个节点：

```

my-rs1:PRIMARY> rs.conf()
{
  "_id" : "my-rs1",
  "version" : 1,
  "protocolVersion" : NumberLong(1),
  "writeConcernMajorityJournalDefault" : true,
  "members" : [
    {
      "_id" : 201,
      "host" : "10.1.5.201:27017",
      "arbiterOnly" : false,
      "buildIndexes" : true,
      "hidden" : false,
      "priority" : 1,
      "tags" : {

      },
      "slaveDelay" : NumberLong(0),
      "votes" : 1
    },
    {
      "_id" : 202,
      "host" : "10.1.5.202:27017",
      "arbiterOnly" : false,
      "buildIndexes" : true,
      "hidden" : false,
      "priority" : 1,
      "tags" : {

      },
    }
  ]
}

```

```

        "slaveDelay" : NumberLong(0),
        "votes" : 1
    }
],
"settings" : {
    "chainingAllowed" : true,
    "heartbeatIntervalMillis" : 2000,
    "heartbeatTimeoutSecs" : 10,
    "electionTimeoutMillis" : 10000,
    "catchUpTimeoutMillis" : -1,
    "catchUpTakeoverDelayMillis" : 30000,
    "getLastErrorModes" : {

    },
    "getLastErrorDefaults" : {
        "w" : 1,
        "wtimeout" : 0
    },
    "replicaSetId" : ObjectId("5b7d32a0339a21d84f409ab4")
}
}

```

输入

```
rs.status()
```

可以查看副本集状态信息，其中包含了主从信息，我们可以看出 201 节点处于 **PRIMARY** 状态，202 节点处于 **SECONDARY** 状态

```
my-rs1:PRIMARY> rs.status()
```

```

{
  "set" : "my-rs1",
  "date" : ISODate("2018-08-22T09:48:38.542Z"),
  "myState" : 1,
  "term" : NumberLong(1),
  "syncingTo" : "",
  "syncSourceHost" : "",
  "syncSourceId" : -1,
  "heartbeatIntervalMillis" : NumberLong(2000),
  "optimes" : {
    "lastCommittedOpTime" : {
      "ts" : Timestamp(1534931310, 1),
      "t" : NumberLong(1)
    },
    "readConcernMajorityOpTime" : {
      "ts" : Timestamp(1534931310, 1),
      "t" : NumberLong(1)
    }
  },

```



```

    "appliedOpTime" : {
      "ts" : Timestamp(1534931310, 1),
      "t" : NumberLong(1)
    },
    "durableOpTime" : {
      "ts" : Timestamp(1534931310, 1),
      "t" : NumberLong(1)
    }
  },
  "lastStableCheckpointTimestamp" : Timestamp(1534931294, 1),
  "members" : [
    {
      "_id" : 201,
      "name" : "10.1.5.201:27017",
      "health" : 1,
      "state" : 1,
      "stateStr" : "PRIMARY",
      "uptime" : 2615,
      "optime" : {
        "ts" : Timestamp(1534931310, 1),
        "t" : NumberLong(1)
      },
      "optimeDate" : ISODate("2018-08-22T09:48:30Z"),
      "syncingTo" : "",
      "syncSourceHost" : "",
      "syncSourceId" : -1,
      "infoMessage" : "",
      "electionTime" : Timestamp(1534928893, 1),
      "electionDate" : ISODate("2018-08-22T09:08:13Z"),
      "configVersion" : 2,
      "self" : true,
      "lastHeartbeatMessage" : ""
    },
    {
      "_id" : 202,
      "name" : "10.1.5.202:27017",
      "health" : 1,
      "state" : 2,
      "stateStr" : "SECONDARY",
      "uptime" : 2435,
      "optime" : {
        "ts" : Timestamp(1534931310, 1),
        "t" : NumberLong(1)
      },
    },
  ]
}

```

```

        "optimeDurable" : {
            "ts" : Timestamp(1534931310, 1),
            "t" : NumberLong(1)
        },
        "optimeDate" : ISODate("2018-08-22T09:48:30Z"),
        "optimeDurableDate" : ISODate("2018-08-22T09:48:30Z"),
        "lastHeartbeat" : ISODate("2018-08-22T09:48:36.715Z"),
        "lastHeartbeatRecv" : ISODate("2018-08-22T09:48:38.237Z"),
        "pingMs" : NumberLong(0),
        "lastHeartbeatMessage" : "",
        "syncingTo" : "",
        "syncSourceHost" : "",
        "syncSourceId" : -1,
        "infoMessage" : "",
        "configVersion" : 2
    }
],
"ok" : 1,
"operationTime" : Timestamp(1534931310, 1),
"$clusterTime" : {
    "clusterTime" : Timestamp(1534931310, 1),
    "signature" : {
        "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA="),
        "keyId" : NumberLong(0)
    }
}
}
}

```

### 2.5.3 初始化副本集-Arbiter 节点

接着上一步在 mongo shell 中，使用

```
rs.addArb("10.1.5.203:27017")
```

添加 arbiter 节点

```

my-rs1:PRIMARY> rs.addArb("10.1.5.203:27017")
{
    "ok" : 1,
    "operationTime" : Timestamp(1534935152, 1),
    "$clusterTime" : {
        "clusterTime" : Timestamp(1534935152, 1),
        "signature" : {
            "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA="),
            "keyId" : NumberLong(0)
        }
    }
}

```

```
}
```

添加完成之后，使用 `rs.status()` 查看集群状态，可以看出 201 节点处于 **PRIMARY** 状态，202 节点处于 **SECONDARY** 状态，203 节点处于 **ARBITER** 状态，符合规划，至此，副本集集群搭建完成。

```
my-rs1:PRIMARY> rs.status()
{
  "set" : "my-rs1",
  "date" : ISODate("2018-08-22T10:52:42.776Z"),
  "myState" : 1,
  "term" : NumberLong(1),
  "syncingTo" : "",
  "syncSourceHost" : "",
  "syncSourceId" : -1,
  "heartbeatIntervalMillis" : NumberLong(2000),
  "optimes" : {
    "lastCommittedOpTime" : {
      "ts" : Timestamp(1534935152, 1),
      "t" : NumberLong(1)
    },
    "readConcernMajorityOpTime" : {
      "ts" : Timestamp(1534935152, 1),
      "t" : NumberLong(1)
    },
    "appliedOpTime" : {
      "ts" : Timestamp(1534935152, 1),
      "t" : NumberLong(1)
    },
    "durableOpTime" : {
      "ts" : Timestamp(1534935152, 1),
      "t" : NumberLong(1)
    }
  },
  "lastStableCheckpointTimestamp" : Timestamp(1534935109, 1),
  "members" : [
    {
      "_id" : 201,
      "name" : "10.1.5.201:27017",
      "health" : 1,
      "state" : 1,
      "stateStr" : "PRIMARY",
      "uptime" : 3596,
      "optime" : {
```

```

        "ts" : Timestamp(1534935152, 1),
        "t" : NumberLong(1)
    },
    "optimeDate" : ISODate("2018-08-22T10:52:32Z"),
    "syncingTo" : "",
    "syncSourceHost" : "",
    "syncSourceId" : -1,
    "infoMessage" : "",
    "electionTime" : Timestamp(1534931627, 1),
    "electionDate" : ISODate("2018-08-22T09:53:47Z"),
    "configVersion" : 2,
    "self" : true,
    "lastHeartbeatMessage" : ""
},
{
    "_id" : 202,
    "name" : "10.1.5.202:27017",
    "health" : 1,
    "state" : 2,
    "stateStr" : "SECONDARY",
    "uptime" : 3545,
    "optime" : {
        "ts" : Timestamp(1534935152, 1),
        "t" : NumberLong(1)
    },
    "optimeDurable" : {
        "ts" : Timestamp(1534935152, 1),
        "t" : NumberLong(1)
    },
    "optimeDate" : ISODate("2018-08-22T10:52:32Z"),
    "optimeDurableDate" : ISODate("2018-08-22T10:52:32Z"),
    "lastHeartbeat" : ISODate("2018-08-22T10:52:40.785Z"),
    "lastHeartbeatRecv" : ISODate("2018-08-22T10:52:42.311Z"),
    "pingMs" : NumberLong(0),
    "lastHeartbeatMessage" : "",
    "syncingTo" : "",
    "syncSourceHost" : "",
    "syncSourceId" : -1,
    "infoMessage" : "",
    "configVersion" : 2
},
{
    "_id" : 203,
    "name" : "10.1.5.203:27017",

```

```

        "health" : 1,
        "state" : 7,
        "stateStr" : "ARBITER",
        "uptime" : 9,
        "lastHeartbeat" : ISODate("2018-08-22T10:52:40.786Z"),
        "lastHeartbeatRecv" : ISODate("2018-08-22T10:52:40.833Z"),
        "pingMs" : NumberLong(1),
        "lastHeartbeatMessage" : "",
        "syncingTo" : "",
        "syncSourceHost" : "",
        "syncSourceId" : -1,
        "infoMessage" : "",
        "configVersion" : 2
    },
    ],
    "ok" : 1,
    "operationTime" : Timestamp(1534935152, 1),
    "$clusterTime" : {
        "clusterTime" : Timestamp(1534935152, 1),
        "signature" : {
            "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA="),
            "keyId" : NumberLong(0)
        }
    }
}

```

#### 2.5.4 初始化的另一种方法

其实 2.5.2 和 2.5.3 可以合并起来，初始化时直接指定 arbiter 节点，如下：

```

rs.initiate( {
  _id : "my-rs1",
  members: [
    { _id: 201, host: "10.1.5.201:27017" },
    { _id: 202, host: "10.1.5.202:27017" },
    { _id: 203, host: "10.1.5.203:27017", arbiterOnly: true}
  ]
})

```

这种方法在 3.4.2 中也使用了。

## 2.6 其他集群操作

### 2.6.1 添加 secondary 节点

操作步骤:

- 1) 配置副本集名字于现有副本集名字一致, 参考 2.4.2
- 2) 启动服务
- 3) 使用 mongo 连接到 primary 节点, 执行如下命令

```
rs.add( { host: "IP:PORT", priority: 0, votes: 0 } )
```

命令解析:

priority: 优先级, 指是否有权限成为主, 比如 Arbiter 角色此值为 0

votes: 投票权权重, 一般应设置各 secondary 权限相同

**注意:** 初始化添加时, 不能设置 priority、votes 大于 0, 因为这可能引发问题, 详情见官方文档, <https://docs.mongodb.com/manual/tutorial/expand-replica-set/>

所以必须使用如下步骤更新这两个值

- 4) 其中 4 表示 rs.conf() 返回的 members 数组中, 新加入的成员的位置, 注意, 数组从 0 开始编号, 所以对于一个有 2 节点的副本集集群, 新加入的时第三位成员, 但在数组中是第 2 个成员

```
var cfg = rs.conf();
cfg.members[2].priority = 1
cfg.members[2].votes = 1
rs.reconfig(cfg)
```

另外, 注意这个操作会引发重新选举, 可能会有 10-20s 服务中断, 所以应在业务空闲时操作

- 5) 添加完成后, 使用 rs.conf() 查看是否符合需求

举例:

```
my-rs1:PRIMARY> rs.add( { host: "10.1.5.202:27017", priority: 0, votes: 0 } )
{
  "ok" : 1,
  "operationTime" : Timestamp(1534995218, 1),
  "$clusterTime" : {
    "clusterTime" : Timestamp(1534995218, 1),
    "signature" : {
      "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA"),
      "keyId" : NumberLong(0)
    }
  }
}
```

添加

```

"members" : [
  {
    "_id" : 201,
    "host" : "10.1.5.201:27017",
    "arbiterOnly" : false,
    "buildIndexes" : true,
    "hidden" : false,
    "priority" : 1,
    "tags" : {
    },
    "slaveDelay" : NumberLong(0),
    "votes" : 1
  },
  {
    "_id" : 203,
    "host" : "10.1.5.203:27017",
    "arbiterOnly" : true,
    "buildIndexes" : true,
    "hidden" : false,
    "priority" : 0,
    "tags" : {
    },
    "slaveDelay" : NumberLong(0),
    "votes" : 1
  },
  {
    "_id" : 204,
    "host" : "10.1.5.202:27017",
    "arbiterOnly" : false,
    "buildIndexes" : true,
    "hidden" : false,
    "priority" : 0,
    "tags" : {
    },
    "slaveDelay" : NumberLong(0),
    "votes" : 0
  }
],

```

查看发现，新加的成员位于数据的第 3 个，组引用位置是 2

```

my-rs1:PRIMARY> var cfg = rs.conf();
my-rs1:PRIMARY>
my-rs1:PRIMARY> cfg.members[2].votes = 1
1
my-rs1:PRIMARY> cfg.members[2].priority = 1
1
my-rs1:PRIMARY>
my-rs1:PRIMARY> rs.reconfig(cfg)
{
  "ok" : 1,
  "operationTime" : Timestamp(1534995133, 1),
  "$clusterTime" : {
    "clusterTime" : Timestamp(1534995133, 1),
    "signature" : {
      "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA"),
      "keyId" : NumberLong(0)
    }
  }
}
my-rs1:PRIMARY>

```

更新新成员权限

## 2. 6. 2 删除 secondary 节点

操作步骤：

1) 使用 mongo 连接到 primary 节点, 执行如下命令

```
rs.remove("IP:PORT")
```

根据官方文档, 还有其它方法, 但这种是最简单的。

2) 添加完成后, 使用 rs.conf() 查看是否添加成功

举例:

```
my-rs1:PRIMARY> rs.remove("10.1.5.202:27017")
{
  "ok" : 1,
  "operationTime" : Timestamp(1534994095, 1),
  "$clusterTime" : {
    "clusterTime" : Timestamp(1534994095, 1),
    "signature" : {
      "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAA"),
      "keyId" : NumberLong(0)
    }
  }
}
```

## 2.6.3 添加 Arbiter 节点

参考 2.5.3

## 2.6.4 删除 Arbiter 节点

参考 2.6.2

## 2.6.5 查看集群信息

```
rs.conf()
rs.status()
rs.isMaster()
```

## 2.7 测试

### 2.7.1 观察日志

1) 集群初始化的日志

```
2018-08-23T13:44:37.554+0800 I NETWORK [conn1] Skipping connection for connection # 2
2018-08-23T13:44:37.554+0800 I REPL [conn1] New replica set config in use: { _id: "my-rs1", version: 1, protocolVersion: 1, writeConcernMajorityJournalDefault: true, members: [ { _id: 201, host: "10.1.5.201:27017", arbiterOnly: false, buildIndexes: true, hidden: false, priority: 1.0, tags: {}, slaveDelay: 0, votes: 1 }, { _id: 202, host: "10.1.5.202:27017", arbiterOnly: false, buildIndexes: true, hidden: false, priority: 1.0, tags: {}, slaveDelay: 0, votes: 1 } ], settings: { chainingAllowed: true, heartbeatIntervalMillis: 2000, heartbeatTimeoutSecs: 10, electionTimeoutMillis: 10000, catchUpTimeoutMillis: -1, catchUpTakeoverDelayMillis: 30000, getLastErrorModes: {}, getLastErrorDefaults: { w: 1, wtimeout: 0 }, replicasetid: ObjectId("5b7e49c5569c9d2aeb577481") } }
2018-08-23T13:44:37.554+0800 I REPL [conn1] This node is 10.1.5.201:27017 in the config
2018-08-23T13:44:37.554+0800 I REPL [conn1] Starting replication storage threads
2018-08-23T13:44:37.555+0800 I REPL [replset-0] Member 10.1.5.202:27017 is now in state STARTUP
2018-08-23T13:44:37.556+0800 I REPL [conn1] transition to RECOVERING from STARTUP2
2018-08-23T13:44:37.556+0800 I REPL [conn1] Starting replication fetcher thread
2018-08-23T13:44:37.556+0800 I REPL [conn1] Starting replication applier thread
2018-08-23T13:44:37.556+0800 I REPL [conn1] Starting replication reporter thread
2018-08-23T13:44:37.556+0800 I REPL [rsSync-0] Starting oplog application
2018-08-23T13:44:37.556+0800 I COMMAND [conn1] command local.system.replset appName: "MongoDB Shell" command: replSetInitiate { replSetInitiate: { _id: "my-rs1", members: [ { _id: 201, host: "10.1.5.201:27017", arbiterOnly: false, buildIndexes: true, hidden: false, priority: 1.0, tags: {}, slaveDelay: 0, votes: 1 }, { _id: 202, host: "10.1.5.202:27017", arbiterOnly: false, buildIndexes: true, hidden: false, priority: 1.0, tags: {}, slaveDelay: 0, votes: 1 } ], settings: { chainingAllowed: true, heartbeatIntervalMillis: 2000, heartbeatTimeoutSecs: 10, electionTimeoutMillis: 10000, catchUpTimeoutMillis: -1, catchUpTakeoverDelayMillis: 30000, getLastErrorModes: {}, getLastErrorDefaults: { w: 1, wtimeout: 0 }, replicasetid: ObjectId("5b7e49c5569c9d2aeb577481") } } }
2018-08-23T13:44:37.556+0800 I NETWORK [conn6] skip closing connection for connection # 6
2018-08-23T13:44:37.556+0800 I NETWORK [conn6] received client metadata from 10.1.5.202:54033 conn6: { driver: { name: "NetworkInterfaceL", version: "4.0.1" }, os: { type: "Linux", name: "CentOS release 6.6 (final)", architecture: "x86_64", version: "Kernel 2.6.32-504.el6.x86_64" } }
2018-08-23T13:44:37.559+0800 I NETWORK [listener] connection accepted from 10.1.5.202:54032 #5 (3 connections now open)
2018-08-23T13:44:37.559+0800 I NETWORK [conn5] end connection 10.1.5.202:54032 (2 connections now open)
2018-08-23T13:44:37.559+0800 I REPL [replset-0] Member 10.1.5.202:27017 is now in state STARTUP2
2018-08-23T13:44:37.590+0800 I NETWORK [listener] connection accepted from 10.1.5.202:54033 #6 (3 connections now open)
2018-08-23T13:44:37.591+0800 I NETWORK [conn6] received client metadata from 10.1.5.202:54033 conn6: { driver: { name: "NetworkInterfaceL", version: "4.0.1" }, os: { type: "Linux", name: "CentOS release 6.6 (final)", architecture: "x86_64", version: "Kernel 2.6.32-504.el6.x86_64" } }
2018-08-23T13:44:37.602+0800 I NETWORK [listener] connection accepted from 10.1.5.202:54034 #7 (4 connections now open)
2018-08-23T13:44:37.603+0800 I NETWORK [conn7] received client metadata from 10.1.5.202:54034 conn7: { driver: { name: "NetworkInterfaceL", version: "4.0.1" }, os: { type: "Linux", name: "CentOS release 6.6 (final)", architecture: "x86_64", version: "Kernel 2.6.32-504.el6.x86_64" } }
```



## 2) 添加 Arbiter 节点的日志

```
2018-08-23T13:45:08.289+0800 I REPL [conn1] replSetReconfig admin command received from client: new config: { _id: "my-rs1", version: 2, protocolVersion: 1, writeConcernMajorityJournalDefault: true, members: [ { _id: 201, host: "10.1.5.201:27017", arbiterOnly: false, buildIndexes: true, hidden: false, priority: 1.0, tags: {}, slaveDelay: 0, votes: 1 }, { _id: 202, host: "10.1.5.202:27017", arbiterOnly: true }, { _id: 203, host: "10.1.5.203:27017", arbiterOnly: true } ], settings: { chainingAllowed: true, heartbeatIntervalMillis: 2000, heartbeatTimeoutSecs: 10, electionTimeoutMillis: 10000, catchUpTimeoutMillis: -1, catchUpTakeoverDelayMillis: 30000, getLastErrorModes: {}, getLastErrorDefaults: { w: 1, wtimeout: 0 }, replicasetId: ObjectId("5b7e49c5569cd2aeb577481") } }
2018-08-23T13:45:08.294+0800 I REPL [replset] New replica set config in use: { _id: "my-rs1", version: 2, protocolVersion: 1, writeConcernMajorityJournalDefault: true, members: [ { _id: 201, host: "10.1.5.201:27017", arbiterOnly: false, buildIndexes: true, hidden: false, priority: 1.0, tags: {}, slaveDelay: 0, votes: 1 }, { _id: 202, host: "10.1.5.202:27017", arbiterOnly: true, buildIndexes: true, hidden: false, priority: 0.0, tags: {}, slaveDelay: 0, votes: 1 }, { _id: 203, host: "10.1.5.203:27017", arbiterOnly: true, buildIndexes: true, hidden: false, priority: 0.0, tags: {}, slaveDelay: 0, votes: 1 } ], settings: { chainingAllowed: true, heartbeatIntervalMillis: 2000, heartbeatTimeoutSecs: 10, electionTimeoutMillis: 10000, catchUpTimeoutMillis: -1, catchUpTakeoverDelayMillis: 30000, getLastErrorModes: {}, getLastErrorDefaults: { w: 1, wtimeout: 0 }, replicasetId: ObjectId("5b7e49c5569cd2aeb577481") } }
2018-08-23T13:45:08.296+0800 I REPL [replset] This node is 10.1.5.201:27017 in the config
2018-08-23T13:45:08.298+0800 I REPL [replset] Member 10.1.5.203:27017 is now in state STARTUP
```

3) 启动日志，参考 1.6.1，及其重要，请务必仔细阅读，有些需要调整系统环境等，可提高性能

### 2.7.2 观察主从切换

连接 primary 节点

```
[root@hadoop01 ~]# mongo 10.1.5.201:27017
MongoDB shell version v4.0.1
connecting to: mongodb://10.1.5.201:27017/test
MongoDB server version: 4.0.1
```

插入一条数据

```
show dbs
use quzl
db.testdb.insert({"name":"quzl"})
```

```
my-rs1:PRIMARY> show dbs
admin 0.000GB
config 0.000GB
local 0.000GB
my-rs1:PRIMARY>
my-rs1:PRIMARY> use quzl
switched to db quzl
my-rs1:PRIMARY>
my-rs1:PRIMARY> db.quzl.insert({"name":"quzl"})
WriteResult({"nInserted" : 1 })
my-rs1:PRIMARY>
my-rs1:PRIMARY> show dbs
admin 0.000GB
config 0.000GB
local 0.000GB
quzl 0.000GB
my-rs1:PRIMARY>
```

停止 201 节点，并连接 202 节点查看

```
[root@hadoop01 ~]# service mongod stop
Stopping mongod:
[root@hadoop01 ~]#
[root@hadoop01 ~]#
[root@hadoop01 ~]#
[root@hadoop01 ~]# mongo 10.1.5.202:27017
MongoDB shell version v4.0.1
connecting to: mongodb://10.1.5.202:27017/test
MongoDB server version: 4.0.1
```

使用

```
rs.status()
```

查看，发现 202 成为了 primary 节点，而 201 处于不可用状态

```

{
  "_id" : 201,
  "name" : "10.1.5.201:27017",
  "health" : 0,
  "state" : 8,
  "stateStr" : "(not reachable/healthy)",
  "uptime" : 0,
  "optime" : {
    "ts" : Timestamp(0, 0),
    "t" : NumberLong(-1)
  },
  "optimeDurable" : {
    "ts" : Timestamp(0, 0),
    "t" : NumberLong(-1)
  },
  "optimeDate" : ISODate("1970-01-01T00:00:00Z"),
  "optimeDurableDate" : ISODate("1970-01-01T00:00:00Z"),
  "lastHeartbeat" : ISODate("2018-08-23T06:11:48.669Z"),
  "lastHeartbeatRecv" : ISODate("2018-08-23T06:10:49.038Z"),
  "pingMs" : NumberLong(0),
  "lastHeartbeatMessage" : "Error connecting to 10.1.5.201:27017 :: caused b",
  "syncingTo" : "",
  "syncSourceHost" : "",
  "syncSourceId" : -1,
  "infoMessage" : "",
  "configVersion" : -1
},
{
  "_id" : 202,
  "name" : "10.1.5.202:27017",
  "health" : 1,
  "state" : 1,
  "stateStr" : "PRIMARY",
  "uptime" : 1662
}

```

进入 quz1 库中查看，发现前面插入的数据仍然存在

```

my-rs1:PRIMARY>
my-rs1:PRIMARY> use quz1
switched to db quz1
my-rs1:PRIMARY>
my-rs1:PRIMARY> db.quz1.find()
{ "_id" : ObjectId("5b7e4ed87d106091d9d1b91c"), "name" : "quz1" }
my-rs1:PRIMARY>
my-rs1:PRIMARY>

```

如果我们查看 202 日志，会发现状态转换的相关日志记录

```

[replexec-11] election succeeded, assuming primary role in term 2
[replexec-11] transition to PRIMARY from SECONDARY
[replexec-11] Entering primary catch-up mode.

```

### 2.7.3 主从切换后操作

如果我们再次启动 201 节点，发现 201 自动成为了 secondary 节点，202 仍然是 primary 节点

```

members" : [
  {
    "_id" : 201,
    "name" : "10.1.5.201:27017",
    "health" : 1,
    "state" : 2,
    "stateStr" : "SECONDARY",
    "uptime" : 12,
    "optime" : {
      "ts" : Timestamp(1535006071, 1),
      "t" : NumberLong(2)
    },
    "optimeDate" : ISODate("2018-08-23T06:34:31Z"),
    "syncingTo" : "10.1.5.202:27017",
    "syncSourceHost" : "10.1.5.202:27017",
    "syncSourceId" : 202,
    "infoMessage" : "",
    "configversion" : 2,
    "self" : true,
    "lastHeartbeatMessage" : ""
  },
  {
    "_id" : 202,
    "name" : "10.1.5.202:27017",
    "health" : 1,
    "state" : 1,
    "stateStr" : "PRIMARY",
    "uptime" : 10,
    "optime" : {
      "ts" : Timestamp(1535006071, 1),
      "t" : NumberLong(2)
    },
    "optimeDurable" : {

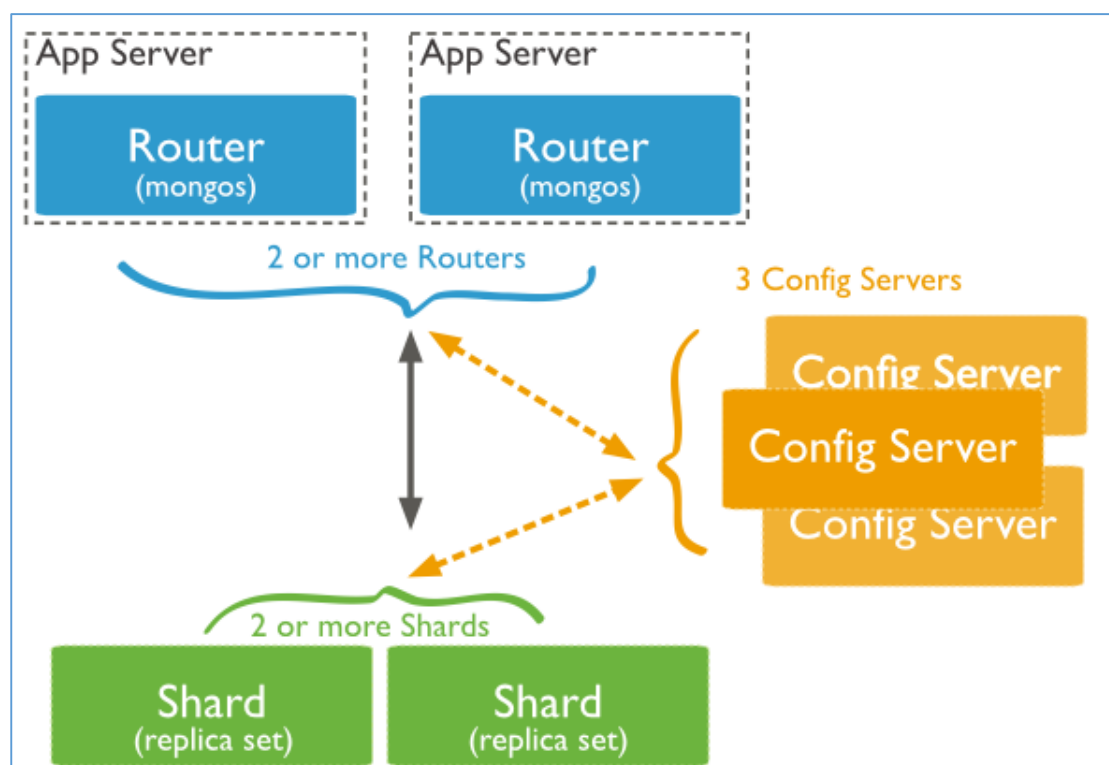
```

#### 特别注意:

如果是生产环境中，发生了主从切换，而且主机器永久性无法起来了，可以通过 2.6.2、2.6.1 的操作步骤，删除故障节点，添加新节点处理。

## 第三章 分片集群

### 3.1 拓扑图



- **Shard:**  
数据分片服务器。用于存储实际的数据块，可以是一个单独的 mongo 实例，也可以是一个副本集，在生产环境下 Shard 一般是一个 Replica Set，以防止该数据片的单点故障。
- **Config Server:**  
配置服务器集群，保存集群的元数据（metadata），包含各个 Shard 的路由规则（trunk 信息）。
- **Query Routers:**  
前端路由，客户端由此接入，且让整个集群看上去像单一数据库，前端应用可以透明使用，由 mongos 把读写请求路由到指定的 Shard 上去。一个 Sharding 集群，可以有一个 mongos，也可以有多 mongos 以减轻客户端请求的压力。

mongos 从配置服务器读取配置，这些信息只保留在内存中，不会有本地存储，当配置服务器的配置有变化时，会通知 mongos 更新。

## 3.2 测试规划

OS: CentOS 6.6 x64

MongoDB: 4.0.1

序号	类目	服务器1 10.1.5.201	服务器2 10.1.5.202	服务器3 10.1.5.203	集群名	端口	日志路径	数据目录	配置文件
1	mongos server	node 1	node 2	node 3	×	28100	/mongodb/mongos/mongs.log	×	×
2	config server	node 1	node 2	node 3	mycfg	28200	/mongodb/config/config.log	/mongodb/config/data	×
3	shard1	node 1	node 2	node 3 仲裁	rs1	27100	/mongodb/shard1/shard1.log	/mongodb/shard1/data	×
4	shard2	node 1 仲裁	node 2	node 3	rs2	27200	/mongodb/shard2/shard2.log	/mongodb/shard2/data	×
5	shard3	node 1	node 2 仲裁	node 3	rs3	27300	/mongodb/shard3/shard3.log	/mongodb/shard3/data	×

这里只是为了测试，使用了 3 台机器，部署了 15 个节点，实际生产环境中：

- 1) 将 mongos server、config server、shard 分开部署到不通机器
- 2) shard，部署成副本集形式，组的多少可横向扩展
- 3) mongos，应该是多节点的，不构成集群，各节点均可独立运行
- 4) config server，应该是多节点的，其实也是一个特殊的副本集集群
- 5) 我这里只是测试，没有单独的配置文件，使用参数启动，没指定的参数使用默认配置，简化了部署过程。生产中，应尽量使用配置文件启动，且定义为系统服务，方便维护，可参考《第二章 副本集集群》部分。

## 3.3 下载和部署

本节 3.3 的所有操作，在三台机器中都要执行。

### 3.3.2 添加 yum 源

```
[root@hadoop01 ~]# cat /etc/yum.repos.d/Mongodb-4.repo
[mongodb-org-4.0]
name = MongoDB Repository
baseurl = http://repo.mongodb.org/yum/redhat/$releasever/mongodb-org/4.0/x86_64/
gpgcheck = 0
enabled = 1
```

### 3.3.2 安装依赖包

```
yum install openssl libcurl
```

### 3.3.3 安装软件包

```
yum install mongodb-org
```

将自动安装以下 4 个依赖包：

mongodb-org-server 包含 `mongod` 守护程序以及关联的配置和 `init` 脚本。

mongodb-org-mongos 包含 `mongos` 守护进程。

mongodb-org-shell 包含 `mongo` shell。

mongodb-org-tools 包含以下的 MongoDB 工具：

`mongoimport`, `bsondump`, `mongodump`, `mongoexport`, `mongofiles`,  
`mongorestore`, `mongostat` and `mongotop`。

软件包	架构	版本
正在安装： mongodb-org	x86_64	4.0.1-1.el6
为依赖而安装： mongodb-org-mongos	x86_64	4.0.1-1.el6
mongodb-org-server	x86_64	4.0.1-1.el6
mongodb-org-shell	x86_64	4.0.1-1.el6
mongodb-org-tools	x86_64	4.0.1-1.el6

事务概要

Install 5 Package(s)

总下载量: 76 M  
Installed size: 232 M  
确定吗? [y/N]:

### 3.3.4 相关目录准备

可使用如下命令，一次性创建完成。

```
mkdir -pv /mongodb/{mongos/data,config/data,shard1/data,shard2/data,shard3/data}
```

如下图

```
[root@hadoop01 ~]# mkdir -pv /mongodb/{mongos/data,config/data,shard1/data,shard2/data,shard3/data}
mkdir: 已创建目录 "/mongodb/mongos"
mkdir: 已创建目录 "/mongodb/mongos/data"
mkdir: 已创建目录 "/mongodb/config"
mkdir: 已创建目录 "/mongodb/config/data"
mkdir: 已创建目录 "/mongodb/shard1"
mkdir: 已创建目录 "/mongodb/shard1/data"
mkdir: 已创建目录 "/mongodb/shard2"
mkdir: 已创建目录 "/mongodb/shard2/data"
mkdir: 已创建目录 "/mongodb/shard3"
mkdir: 已创建目录 "/mongodb/shard3/data"
[root@hadoop01 ~]#
[root@hadoop01 ~]#
[root@hadoop01 ~]# tree /mongodb
/mongodb
├── config
│   └── data
├── mongos
│   └── data
├── shard1
│   └── data
├── shard2
│   └── data
└── shard3
    └── data

10 directories, 0 files
```

这里是测试，直接使用 root 运行，如果不是 root，要修改目录权限，比如使用 mongo 运行

```
chmod -R mongo.mongd /mongodb
```

## 3.4 启动 Shard Server

该小节其实就是《第二章 副本集集群》的搭建过程，这里只写关键步骤，详细情况可参考前第二章。

### 3.4.1 启动所有 mongod 服务

三台机器都执行以下语句，初始化三个副本集，每个副本集 3 个节点

```
mongod --bind_ip 0.0.0.0 --port 27100 --replSet rs1 --shardsvr \  
--dbpath=/mongodb/shard1/data \  
--logpath=/mongodb/shard1/shard1.log \  
--logappend --fork
```

```
mongod --bind_ip 0.0.0.0 --port 27200 --replSet rs2 --shardsvr \  
--dbpath=/mongodb/shard2/data \  
--logpath=/mongodb/shard2/shard2.log \  
--logappend --fork
```

```
mongod --bind_ip 0.0.0.0 --port 27300 --replSet rs3 --shardsvr \  
--dbpath=/mongodb/shard3/data \  
--logpath=/mongodb/shard3/shard3.log \  
--logappend --fork
```

如下图，表示启动成功

```
mkdir: 已创建目录 '/mongodb/shard3/data'  
[root@hadoop03 ~]# mongod --bind_ip 0.0.0.0 --port 27100 --replSet rs1 --shardsvr \  
> --dbpath=/mongodb/shard1/data \  
> --logpath=/mongodb/shard1/shard1.log \  
> --logappend --fork  
2018-08-27T11:55:15.109+0800 I CONTROL [main] Automatically disabling TLS 1.0, to force-enable TLS 1.0  
about to fork child process, waiting until server is ready for connections.  
forked process: 1797  
child process started successfully, parent exiting  
[root@hadoop03 ~]#  
[root@hadoop03 ~]#  
[root@hadoop03 ~]#  
[root@hadoop03 ~]# mongod --bind_ip 0.0.0.0 --port 27200 --replSet rs2 --shardsvr \  
> --dbpath=/mongodb/shard2/data \  
> --logpath=/mongodb/shard2/shard2.log \  
> --logappend --fork  
2018-08-27T11:55:36.906+0800 I CONTROL [main] Automatically disabling TLS 1.0, to force-enable TLS 1.0  
about to fork child process, waiting until server is ready for connections.  
forked process: 1828  
child process started successfully, parent exiting  
[root@hadoop03 ~]#  
[root@hadoop03 ~]#  
[root@hadoop03 ~]#  
[root@hadoop03 ~]# mongod --bind_ip 0.0.0.0 --port 27300 --replSet rs3 --shardsvr \  
> --dbpath=/mongodb/shard3/data \  
> --logpath=/mongodb/shard3/shard3.log \  
> --logappend --fork  
2018-08-27T11:55:55.516+0800 I CONTROL [main] Automatically disabling TLS 1.0, to force-enable TLS 1.0  
about to fork child process, waiting until server is ready for connections.  
forked process: 1859  
child process started successfully, parent exiting  
[root@hadoop03 ~]#  
[root@hadoop03 ~]#
```

更多启动参数，使用

```
mongod --help
```

查看

**注意：**必须使用--shardsvr 参数，否则使用 mongos 连接后插入数，会提示错误如下

```
mongos>  
mongos> db.table1.save({id:1,"test1":"testval1"})  
writeResult({  
  "nInserted" : 0,  
  "writeError" : {  
    "code" : 193,  
    "errmsg" : "Cannot accept sharding commands if not started with --shardsvr"  
  }  
})  
mongos>  
mongos>
```

网上有些资料说，不需要添加这个参数，是错误的说法，也可能旧版本是可以的。经验证，4.0 版本，必须要添加这个参数。

### 3.4.2 创建三个分片集群

**副本集 1**，即 shard1,使用 mongo 连到 201 的 27100 端口。

按照规划，203:27100 是仲裁节点，集群 id 是 rs1，成员 id 随意写。

执行如下语句：

```
mongo 10.1.5.201:27100
```

```
rs.initiate( {
  _id : "rs1",
  members: [
    { _id: 0, host: "10.1.5.201:27100" },
    { _id: 1, host: "10.1.5.202:27100" },
    { _id: 2, host: "10.1.5.203:27100", arbiterOnly: true}
  ]
})
```

```
[root@hadoop01 ~]# mongo 10.1.5.201:27100
MongoDB shell version v4.0.1
connecting to: mongodb://10.1.5.201:27100/test
MongoDB server version: 4.0.1
Server has startup warnings:
```

```
> rs.initiate( {
...   _id : "rs1",
...   members: [
...     { _id: 0, host: "10.1.5.201:27100" },
...     { _id: 1, host: "10.1.5.202:27100" },
...     { _id: 2, host: "10.1.5.203:27100", arbiteronly: true}
...   ]
... })
{
  "ok" : 1,
  "operationTime" : Timestamp(1535096043, 1),
  "$clusterTime" : {
    "clusterTime" : Timestamp(1535096043, 1),
    "signature" : {
      "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA"),
      "keyId" : NumberLong(0)
    }
  }
}
```

**副本集 2**，即 shard2,使用 mongo 连到 202 的 27200 端口。

按照规划，201:27200 是仲裁节点，集群 id 是 rs2，成员 id 随意写。

执行如下语句：

```
mongo 10.1.5.202:27200
```

```
rs.initiate( {
  _id : "rs2",
  members: [
    { _id: 0, host: "10.1.5.201:27200", arbiterOnly: true},
    { _id: 1, host: "10.1.5.202:27200" },
    { _id: 2, host: "10.1.5.203:27200" }
  ]
})
```



```
[root@hadoop01 ~]# mongo 10.1.5.202:27200
MongoDB shell version v4.0.1
connecting to: mongodb://10.1.5.202:27200/test
MongoDB server version: 4.0.1
server has startup warnings:
```

```
> rs.initiate( {
...   _id : "rs2",
...   members: [
...     { _id: 0, host: "10.1.5.201:27200", arbiterOnly: true},
...     { _id: 1, host: "10.1.5.202:27200"},
...     { _id: 2, host: "10.1.5.203:27200"}
...   ]
... })
{
  "ok" : 1,
  "operationTime" : Timestamp(1535096881, 1),
  "$clusterTime" : {
    "clusterTime" : Timestamp(1535096881, 1),
    "signature" : {
      "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAA="),
      "keyId" : NumberLong(0)
    }
  }
}
```

副本集 3，即 shard3，使用 mongo 连到 203 的 27300 端口。  
按照规划，202:27300 是仲裁节点，集群 id 是 rs3，成员 id 随意写。  
执行如下语句：

```
mongo 10.1.5.203:27300
```

```
rs.initiate( {
  _id : "rs3",
  members: [
    { _id: 0, host: "10.1.5.201:27300"},
    { _id: 1, host: "10.1.5.202:27300", arbiterOnly: true},
    { _id: 2, host: "10.1.5.203:27300"}
  ]
})
```

```
[root@hadoop01 ~]# mongo 10.1.5.203:27300
MongoDB shell version v4.0.1
connecting to: mongodb://10.1.5.203:27300/test
MongoDB server version: 4.0.1
Server has startup warnings:
```

```
> rs.initiate( {
...   _id : "rs3",
...   members: [
...     { _id: 0, host: "10.1.5.201:27300"},
...     { _id: 1, host: "10.1.5.202:27300", arbiterOnly: true},
...     { _id: 2, host: "10.1.5.203:27300"}
...   ]
... })
{
  "ok" : 1,
  "operationTime" : Timestamp(1535097032, 1),
  "$clusterTime" : {
    "clusterTime" : Timestamp(1535097032, 1),
    "signature" : {
      "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAA="),
      "keyId" : NumberLong(0)
    }
  }
}
```

### 3.5 启动 config server

#### 3.5.1 启动 mongod 服务

三台机器都要执行，其实这一步和 3.4.1 执行的语句区别就是加上了

--configsvr

该参数表示，创建的是一个配置服务。网上有些教程说，不用添加这个参数，如果不加，创建 config 集群也可以成功，但 mongos 服务无法起来。

```
mongod --bind_ip 0.0.0.0 --port 28200 \  
--configsvr \  
--replSet mycfg \  
--dbpath=/mongodb/config/data \  
--logpath=/mongodb/config/config.log \  
--logappend --fork
```

如下图，表示启动成功

```
[root@hadoop01 ~]#  
[root@hadoop01 ~]# mongod --bind_ip 0.0.0.0 --port 28200 --configsvr --replSet mycfg --dbpath=/mongodb/config/  
data --logpath=/mongodb/config/config.log --logappend --fork  
2018-08-24T16:31:17.896+0800 I CONTROL [main] Automatically disabling TLS 1.0, to force-enable TLS 1.0 specif  
y --sslDisabledProtocols 'none'  
about to fork child process, waiting until server is ready for connections.  
forked process: 17082  
child process started successfully, parent exiting  
[root@hadoop01 ~]#
```

#### 3.5.2 创建 config 集群

使用 mongo 连到 201 的 28200 端口

按照规划，集群 id 是 mycfg，成员 id 随意写。

执行如下语句：

```
mongo 10.1.5.201:28200  
  
rs.initiate( {  
  _id : "mycfg",  
  members: [  
    { _id: 0, host: "10.1.5.201:28200"},  
    { _id: 1, host: "10.1.5.202:28200"},  
    { _id: 2, host: "10.1.5.203:28200"}  
  ]  
})
```

注意，因为 config 集群，存储的数据量很少，所以创建的是没有 arbiter 节点的副本集集群，至于区别可参考 2.1 节

```
[root@hadoop01 ~]#  
[root@hadoop01 ~]# mongo 10.1.5.201:28200  
MongoDB shell version v4.0.1  
connecting to: mongodb://10.1.5.201:28200/test  
MongoDB server version: 4.0.1
```

```
> rs.initiate( {
...   _id : "mycfg",
...   members: [
...     { _id: 0, host: "10.1.5.201:28200"},
...     { _id: 1, host: "10.1.5.202:28200"},
...     { _id: 2, host: "10.1.5.203:28200"}
...   ]
... })
{
  "ok" : 1,
  "operationTime" : Timestamp(1535098279, 1),
  "$clusterTime" : {
    "clusterTime" : Timestamp(1535098279, 1),
    "signature" : {
      "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAA="),
      "keyId" : NumberLong(0)
    }
  }
}
```

### 3.6 启动 mongos server

mongos server 也应该多台的，使用下面命令，在三台机器均执行

```
mongos --bind_ip 0.0.0.0 --port 28100 \
--configdb mycfg/10.1.5.201:28200,10.1.5.202:28200,10.1.5.203:28200 \
--logpath=/mongodb/mongos/mongos.log \
--logappend --fork
```

如下图所示，表示启动成功

```
[root@hadoop03 ~]#
[root@hadoop03 ~]# mongos --bind_ip 0.0.0.0 --port 28100 \
> --configdb mycfg/10.1.5.201:28200,10.1.5.202:28200,10.1.5.203:28200 \
> --logpath=/mongodb/mongos/mongos.log \
> --logappend --fork
2018-08-24T16:37:15.865+0800 I_CONTROL [main] Automatically disabling TLS 1.0, to force-enable TLS 1.0 specif
y --sslDisabledProtocols 'none'
about to fork child process, waiting until server is ready for connections.
forked process: 4265
child process started successfully, parent exiting
[root@hadoop03 ~]#
```

### 3.7 添加分片信息

#### 3.7.1 添加 rs1

连接 mongos，输入以下命令

```
mongo 10.1.5.201:28100
use admin
db.runCommand( { addshard : "rs1/10.1.5.201:27100,10.1.5.203:27100,10.1.5.202:27100"});
```

如下图：

```
[root@hadoop01 ~]#
[root@hadoop01 ~]# mongo 127.0.0.1:28100
MongoDB shell version v4.0.1
connecting to: mongodb://127.0.0.1:28100/test
MongoDB server version: 4.0.1
```

```

mongos> use admin
switched to db admin
mongos>
mongos> db.runCommand( { addshard : "rs1/10.1.5.201:27100,10.1.5.203:27100,10.1.5.202:27100"});
{
  "shardAdded" : "rs1",
  "ok" : 1,
  "operationTime" : Timestamp(1535103097, 3),
  "$clusterTime" : {
    "clusterTime" : Timestamp(1535103097, 3),
    "signature" : {
      "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAA="),
      "keyId" : NumberLong(0)
    }
  }
}

```

### 3.7.2 添加 rs2

接着上一步，输入以下命令

```
db.runCommand( { addshard : "rs2/10.1.5.201:27200,10.1.5.202:27200,10.1.5.203:27200"});
```

如下图：

```

mongos>
mongos> db.runCommand( { addshard : "rs2/10.1.5.201:27200,10.1.5.202:27200,10.1.5.203:27200"});
{
  "shardAdded" : "rs2",
  "ok" : 1,
  "operationTime" : Timestamp(1535103487, 2),
  "$clusterTime" : {
    "clusterTime" : Timestamp(1535103487, 2),
    "signature" : {
      "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAA="),
      "keyId" : NumberLong(0)
    }
  }
}

```

### 3.7.3 添加 rs3

接着上一步，输入以下命令

```
db.runCommand( { addshard : "rs3/10.1.5.201:27300,10.1.5.202:27300,10.1.5.203:27300"});
```

如下图：

```

mongos> db.runCommand( { addshard : "rs3/10.1.5.201:27300,10.1.5.202:27300,10.1.5.203:27300"});
{
  "shardAdded" : "rs3",
  "ok" : 1,
  "operationTime" : Timestamp(1535103535, 4),
  "$clusterTime" : {
    "clusterTime" : Timestamp(1535103535, 4),
    "signature" : {
      "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAA="),
      "keyId" : NumberLong(0)
    }
  }
}

```

### 3.7.4 相关说明

- 1) 只需要在一台 mongos 中添加配置信息即可
- 2) 配置信息会通过 mongos 同步到 config 服务器，其他 mongos 再读取 config 服务器的信息，从而获得同步
- 3) 各 mongos 可以独立工作

### 3.8 启用分片

目前搭建了 mongodb 配置服务器、路由服务器，各个分片服务器，不过应用程序连接到 mongos 路由服务器并不能使用分片机制，还需要在程序里设置分片配置，让分片生效。

#### 3.8.1 指定分片的库

使用 mongo shell 连接其中一台 mongos，注意是 mongos，端口是 28100

```
mongo 10.1.5.203:28100
```

```
[root@hadoop03 ~]#  
[root@hadoop03 ~]# mongo 10.1.5.203:28100  
MongoDB shell version v4.0.1  
connecting to: mongodb://10.1.5.203:28100/test  
MongoDB server version: 4.0.1  
Server has startup warnings:
```

使用如下语句

```
sh.enableSharding("com")
```

指定要使用分片功能的数据库

```
mongos>  
mongos> sh.enableSharding("com")  
{  
  "ok" : 1,  
  "operationTime" : Timestamp(1535350115, 2),  
  "$clusterTime" : {  
    "clusterTime" : Timestamp(1535350115, 2),  
    "signature" : {  
      "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA"),  
      "keyId" : NumberLong(0)  
    }  
  }  
}
```

使用如下语句查看是否配置成功。其中"partitioned": true 表示启用分片

```
use config
```

```
db.databases.find()
```

```
mongos> use config  
switched to db config  
mongos> db.databases.find()  
{ "_id" : "com", "primary" : "rs2", "partitioned" : true, "version" : { "uuid" : UUID("f08d2e37-463f-463f-bdbf-77d3b3f46161"), "lastMod" : 1 } }  
mongos>
```

#### 3.8.2 指定分片的集合

接着上一步，运行如下语句

```
sh.shardCollection('com.users',{'_id':'hashed'})
```

表示对 com 库的 users 集合启用分片，分片的 key 是 id，分片的方法是 hashed

另外，如果集合中已经存在数据，在选定作为 shard key 的键列必须提前创建索引；如果集合为空，mongodb 将在激活集合分片（sh.shardCollection）时创建索引。

至此，整个分片集群搭建完成。下面进行测试部分。

## 3.9 测试

### 3.9.1 观察日志

参考 1.6.1 和 2.7.1，这一步一定要看下，因为有些系统环境变量需要调整。

### 3.9.2 查看各副本集群配置

shard1:

```
[root@hadoop01 ~]# mongo 10.1.5.201:27100
MongoDB shell version v4.0.1
connecting to: mongodb://10.1.5.201:27100/test
MongoDB server version: 4.0.1
Server has startup warnings:

rs1:PRIMARY> rs.isMaster()
{
  "hosts" : [
    "10.1.5.201:27100",
    "10.1.5.202:27100"
  ],
  "arbiters" : [
    "10.1.5.203:27100"
  ],
  "primary" : "10.1.5.201:27100"
}
```

shard2:

```
[root@hadoop01 ~]# mongo 10.1.5.201:27200
MongoDB shell version v4.0.1
connecting to: mongodb://10.1.5.201:27200/test
MongoDB server version: 4.0.1
Server has startup warnings:

rs2:ARBITER> rs.isMaster()
{
  "hosts" : [
    "10.1.5.202:27200",
    "10.1.5.203:27200"
  ],
  "arbiters" : [
    "10.1.5.201:27200"
  ],
  "primary" : "10.1.5.202:27200"
}
```

shard3:

```
[root@hadoop01 ~]# mongo 10.1.5.201:27300
MongoDB shell version v4.0.1
connecting to: mongodb://10.1.5.201:27300/test
MongoDB server version: 4.0.1
Server has startup warnings:

rs3:SECONDARY> rs.isMaster()
{
  "hosts" : [
    "10.1.5.203:27300",
    "10.1.5.201:27300"
  ],
  "arbiters" : [
    "10.1.5.202:27300"
  ],
  "primary" : "10.1.5.203:27300"
}
```

经过查看，符合规划。

### 3.9.3 查看分片配置

使用

```
mongo 10.1.5.201:28100
```

登陆 mongos, 使用如下命令查看分片配置

```
use admin
```

```
db.runCommand({ listshards : 1 });
```

如下图所示, 注意每个 shard 集群中, 仲裁节点并未显示, 这是正常现象

```
[root@hadoop01 ~]#  
[root@hadoop01 ~]# mongo 127.0.0.1:28100  
MongoDB shell version v4.0.1  
connecting to: mongod://127.0.0.1:28100/test  
MongoDB server version: 4.0.1  
  
mongos>  
mongos> use admin  
switched to db admin  
mongos>  
mongos> db.runCommand({ listshards : 1 })  
{  
  "shards" : [  
    {  
      "_id" : "rs1",  
      "host" : "rs1/10.1.5.201:27100,10.1.5.202:27100",  
      "state" : 1  
    },  
    {  
      "_id" : "rs2",  
      "host" : "rs2/10.1.5.202:27200,10.1.5.203:27200",  
      "state" : 1  
    },  
    {  
      "_id" : "rs3",  
      "host" : "rs3/10.1.5.201:27300,10.1.5.203:27300",  
      "state" : 1  
    }  
  ],  
  "ok" : 1,  
  "operationTime" : Timestamp(1535103955, 1),  
  "$clusterTime" : {  
    "clusterTime" : Timestamp(1535103955, 1),  
    "signature" : {  
      "hash" : BinData(0,"AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA="),  
      "keyId" : NumberLong(0)  
    }  
  }  
}
```

同时也可以使用

```
sh.status()
```

查看当前的分片状态

```

mongos> sh.status()
mongos> --- Sharding Status ---
  sharding version: {
    "_id" : 1,
    "minCompatibleVersion" : 5,
    "currentVersion" : 6,
    "clusterId" : ObjectId("5b7fc2a3f4bef03b44f93f53")
  }
  shards:
    { "_id" : "rs1", "host" : "rs1/10.1.5.201:27100,10.1.5.202:27100", "state" : 1 }
    { "_id" : "rs2", "host" : "rs2/10.1.5.202:27200,10.1.5.203:27200", "state" : 1 }
    { "_id" : "rs3", "host" : "rs3/10.1.5.201:27300,10.1.5.203:27300", "state" : 1 }
  active mongoses:
    "4.0.1" : 3
  autosplit:
    Currently enabled: yes
  balancer:
    Currently enabled: yes
    Currently running: no
    Failed balancer rounds in last 5 attempts: 0
    Migration Results for the last 24 hours:
      No recent migrations
  databases:
    { "_id" : "config", "primary" : "config", "partitioned" : true }
      config.system.sessions
        shard key: { "_id" : 1 }
        unique: false
        balancing: true
        chunks:
          rs1 1
          { "_id" : { "$minkey" : 1 } } --> { "_id" : { "$maxkey" : 1 } } on : rs1 Timestamp(1, 0)
mongos>

```

经查，符合规划。

### 3.9.4 验证是否会分片

使用

mongo 10.1.5.201:28100

登陆 mongos

```

[root@hadoop01 ~]#
[root@hadoop01 ~]# mongo 127.0.0.1:28100
MongoDB shell version v4.0.1
connecting to: mongodb://127.0.0.1:28100/test
MongoDB server version: 4.0.1

```

切换到 com 库，创建 20W 条数据，其中 id 字段是变化的

use com

for(i=1;i<=200000;i++) db.users.insert({id:i,name:'hukey',age:23})

```

mongos>
mongos> use com
switched to db com
mongos>
mongos> for(i=1;i<=200000;i++) db.users.insert({id:i,name:'hukey',age:23})
writeResult({ "nInserted" : 1 })
mongos>
mongos>

```

使用如下命令，查看 users 表的信息

use admin

db.users.stats()



```

[root@hadoop01 ~]#
[root@hadoop01 ~]# mongo 10.1.5.201:28100
MongoDB shell version v4.0.1
connecting to: mongodb://10.1.5.201:28100/test
MongoDB server version: 4.0.1
Server has startup warnings:
2018-08-27T12:22:59.384+0800 I CONTROL [main]
2018-08-27T12:22:59.384+0800 I CONTROL [main] ** WARNING: Acces
2018-08-27T12:22:59.384+0800 I CONTROL [main] ** WARNING: Read
2018-08-27T12:22:59.384+0800 I CONTROL [main] ** WARNING: You
2018-08-27T12:22:59.384+0800 I CONTROL [main]
mongos>
mongos> use com
switched to db com
mongos>
mongos> db.users.stats()
{
  "sharded" : true,
  "capped" : false,
  "wiredTiger" : {
    "metadata" : {
      "formatVersion" : 1
    }
  }
}

```

如下面三图所示，20W 个 Key，被分配到了 3 个分片集群上，当然不是绝对 100%评价，这牵涉到分片的方法和技术，有兴趣的同学可以深入了解下。

```

"rs1" : {
  "ns" : "com.users",
  "size" : 4228245,
  "count" : 67115,
  "avgobjsize" : 63,
  "storageSize" : 1343488,
  "capped" : false,
  "wiredTiger" : {
    "metadata" : {
      "formatVersion" : 1
    }
  }
}

```

```

"rs2" : {
  "ns" : "com.users",
  "size" : 4184271,
  "count" : 66417,
  "avgobjsize" : 63,
  "storageSize" : 1327104,
  "capped" : false,
  "wiredTiger" : {
    "metadata" : {
      "formatVersion" : 1
    }
  }
}

```

```

"rs3" : {
  "ns" : "com.users",
  "size" : 4187484,
  "count" : 66468,
  "avgobjsize" : 63,
  "storageSize" : 1323008,
  "capped" : false,
  "wiredTiger" : {
    "metadata" : {
      "formatVersion" : 1
    }
  }
}

```

### 3.9.5 shared server 主从切换

参考 2.7.3

### 3.9.6 conf server 主从切换

其实配置服务器集群也是一个特殊的副本集集群，前面已多次说到。相关操作命令通用，可参考第二章相关内容。

主从切换的相关内容，参考 2.7.3