配置mongodb分片群集(sharding cluster)

**Sharding cluster介绍**

这是一种可以水平扩展的模式，在数据量很大时特给力，实际大规模应用一般会采用这种架构去构建monodb系统。

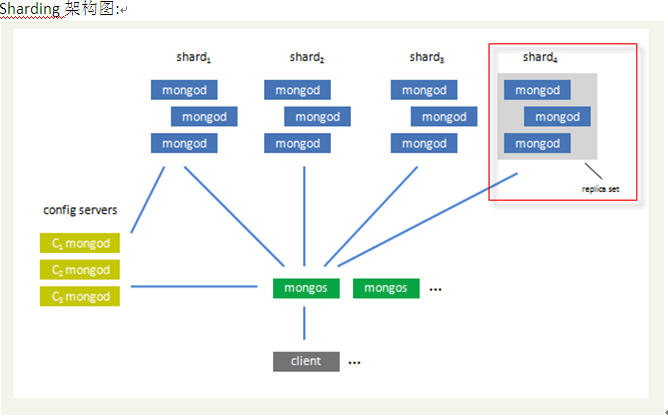
要构建一个 MongoDB Sharding Cluster，需要三种角色：

Shard Server: mongod 实例，用于存储实际的数据块，实际生产环境中一个shard server角色可由几台机器组个一个relica set承担，防止主机单点故障

Config Server: mongod 实例，存储了整个 Cluster Metadata，其中包括 chunk 信息。

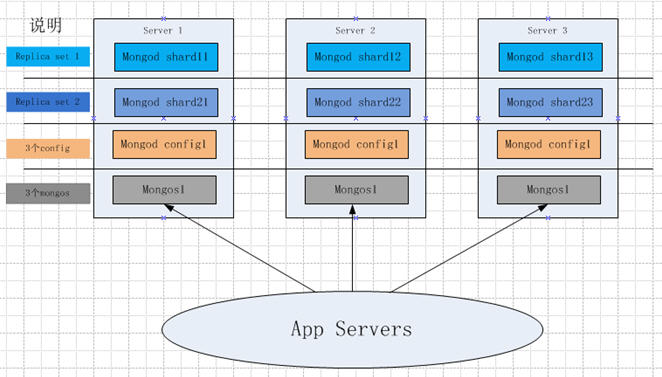
Route Server: mongos 实例，前端路由，客户端由此接入，且让整个集群看上去像单一数据库，前端应用可以透明使用。

Sharding架构图:

[](http://www.taobaodba.com/wp-content/dbauploads/2010/09/1.jpg)

**本例实际环境架构**

本例架构示例图：

[](http://www.taobaodba.com/wp-content/dbauploads/2010/09/mongodb-2.jpg)

1. 分别在3台机器运行一个mongod实例（称为mongod shard11，mongod shard12，mongod shard13）组织replica set1，作为cluster的shard1
2. 分别在3台机器运行一个mongod实例（称为mongod shard21，mongod shard22，mongod shard23）组织replica set2，作为cluster的shard2
3. 每台机器运行一个mongod实例，作为3个config server
4. 每台机器运行一个mongs进程，用于客户端连接

|  |  |  |
| --- | --- | --- |
| **主机** | **IP** | **端口信息** |
| Server1 | 10.1.1.1 | mongod shard11:27017 mongod shard12:27018 mongod config1:20000 mongs1:30000 |
| Server2 | 10.1.1.2 | mongod shard12:27017 mongod shard22:27018 mongod config2:20000 mongs2:30000 |
| Server3 | 10.1.1.3 | mongod shard13:27017 mongod shard23:27018 mongod config3:20000 mongs3:30000 |

**软件准备**

软件准备  
1. 创建用户  
groupadd -g 20001 mongodb  
useradd -u 20001 -g mongodb mongodb  
passwd mongodb

2. 安装monodb软件  
su – mongodb  
tar zxvf mongodb-linux-x86\_64-1.6.2.tar  
安装好后，目录结构如下：  
$ tree mongodb-linux-x86\_64-1.6.2  
mongodb-linux-x86\_64-1.6.2  
|– GNU-AGPL-3.0  
|– README  
|– THIRD-PARTY-NOTICES  
`– bin  
|– bsondump  
|– mongo  
|– mongod  
|– mongodump  
|– mongoexport  
|– mongofiles  
|– mongoimport  
|– mongorestore  
|– mongos  
|– mongosniff  
`– mongostat  
1 directory, 14 files

3. 创建数据目录  
根据本例sharding架构图所示，在各台sever上创建shard数据文件目录  
Server1:  
su – monodb  
cd /home/monodb  
mkdir -p data/shard11  
mkdir -p data/shard21  
Server2:  
su – monodb  
cd /home/monodb  
mkdir -p data/shard11  
mkdir -p data/shard22  
Server3:  
su – monodb  
cd /home/monodb  
mkdir -p data/shard13  
mkdir -p data/shard23

**配置relica sets**

1. 配置shard1所用到的replica sets:  
Server1:  
cd /home/mongodb/mongodb-linux-x86\_64-1.6.2/bin  
./mongod –shardsvr –replSet shard1 –port 27017 –dbpath /home/mongodb/data/shard11 –oplogSize 100 –logpath /home/mongodb/data/shard11.log –logappend –fork

Server2:  
cd /home/mongodb/mongodb-linux-x86\_64-1.6.2/bin  
./mongod –shardsvr –replSet shard1 –port 27017 –dbpath /home/mongodb/data/shard12 –oplogSize 100 –logpath /home/mongodb/data/shard12.log –logappend –fork

Server3:  
cd /home/mongodb/mongodb-linux-x86\_64-1.6.2/bin  
./mongod –shardsvr –replSet shard1 –port 27017 –dbpath /home/mongodb/data/shard13 –oplogSize 100 –logpath /home/mongodb/data/shard13.log –logappend –fork

初始化replica set  
用mongo连接其中一个mongod，执行:  
> config = {\_id: ‘shard1′, members: [  
{\_id: 0, host: '10.1.1.1:27017'},  
{\_id: 1, host: '10.1.1.2:27017'},  
{\_id: 2, host: '10.1.1.3:27017'}]  
}

> rs.initiate(config);

同样方法，配置shard2用到的replica sets:  
server1:  
cd /home/mongodb/mongodb-linux-x86\_64-1.6.2/bin  
./mongod –shardsvr –replSet shard2 –port 27018 –dbpath /home/mongodb/data/shard21 –oplogSize 100 –logpath /home/mongodb/data/shard21.log –logappend –fork

server2:  
cd /home/mongodb/mongodb-linux-x86\_64-1.6.2/bin  
./mongod –shardsvr –replSet shard2 –port 27018 –dbpath /home/mongodb/data/shard22 –oplogSize 100 –logpath /home/mongodb/data/shard22.log –logappend –fork

server3:  
cd /home/mongodb/mongodb-linux-x86\_64-1.6.2/bin  
./mongod –shardsvr –replSet shard2 –port 27018 –dbpath /home/mongodb/data/shard23 –oplogSize 100 –logpath /home/mongodb/data/shard23.log –logappend –fork

初始化replica set  
用mongo连接其中一个mongod，执行:  
> config = {\_id: ‘shard2′, members: [  
{\_id: 0, host: '10.1.1.1:27018'},  
{\_id: 1, host: '10.1.1.2:27018'},  
{\_id: 2, host: '10.1.1.3:27018'}]  
}

> rs.initiate(config);

到此就配置好了二个replica sets，也就是准备好了二个shards

**配置三台config server**

Server1:  
mkdir -p /home/mongodb/data/config  
./mongod –configsvr –dbpath /home/mongodb/data/config –port 20000 –logpath /home/mongodb/data/config.log –logappend –fork #config server也需要dbpath  
  
Server2:  
mkdir -p /home/mongodb/data/config  
./mongod –configsvr –dbpath /home/mongodb/data/config –port 20000 –logpath /home/mongodb/data/config.log –logappend –fork

Server3:  
mkdir -p /home/mongodb/data/config  
./mongod –configsvr –dbpath /home/mongodb/data/config –port 20000 –logpath /home/mongodb/data/config.log –logappend –fork

**配置mongs**

在server1,server2,server3上分别执行：  
./mongos –configdb 10.1.1.1:20000,10.1.1.2:20000,10.1.1.3:20000 –port 30000 –chunkSize 5 –logpath /home/mongodb/data/mongos.log –logappend –fork   
#mongs不需要dbpath

**Configuring the Shard Cluster**

连接到其中一个mongos进程，并切换到admin数据库做以下配置  
1. 连接到mongs，并切换到admin  
./mongo 10.1.1.1:30000/admin  
>db  
Admin  
2. 加入shards  
如里shard是单台服务器，用>db.runCommand( { addshard : “<serverhostname>[:<port>]” } )这样的命令加入，如果shard是replica sets，用replicaSetName/<serverhostname>[:port][,serverhostname2[:port],…]这样的格式表示，例如本例执行：  
>db.runCommand( { addshard : “shard1/10.1.1.1:27017,10.1.1.2:27017,10.1.1.3:27017″,name:”s1″,maxsize:20480} );  
>db.runCommand( { addshard : “shard2/10.1.1.1:27018,10.1.1.2:27018,10.1.1.3:27018″,name:”s2″,maxsize:20480} );  
注意：在添加第二个shard时，出现error:test database 已经存在的错误，这里用mongo命令连接到第二个replica set，用db.dropDatabase()命令把test数据库给删除然后就可加入

3. 可选参数  
Name:用于指定每个shard的名字，不指定的话系统将自动分配  
maxSize:指定各个shard可使用的最大磁盘空间，单位megabytes

4. Listing shards  
>db.runCommand( { listshards : 1 } )  
如果列出了以上二个你加的shards，表示shards已经配置成功

5. 激活数据库分片  
命令：  
> db.runCommand( { enablesharding : “<dbname>” } );  
通过执行以上命令，可以让数据库跨shard，如果不执行这步，数据库只会存放在一个shard，一旦激活数据库分片，数据库中不同的collection将被存放在不同的shard上，但一个collection仍旧存放在同一个shard上，要使单个collection也分片，还需单独对collection作些操作

**Collecton分片**

要使单个collection也分片存储，需要给collection指定一个分片key，通过以下命令操作：  
> db.runCommand( { shardcollection : “<namespace>”,key : <shardkeypatternobject> });  
注：  
a. 分片的collection系统会自动创建一个索引（也可用户提前创建好）  
b. 分片的collection只能有一个在分片key上的唯一索引，其它唯一索引不被允许  
One note: a sharded collection can have only one unique index, which must exist on the shard key. No other unique indexes can exist on the collection.

**分片collection例子**

>db.runCommand( { shardcollection : “test.c1″,key : {id: 1} } )  
>for (var i = 1; i <= 200003; i++) db.c1.save({id:i,value1:”1234567890″,value2:”1234567890″,value3:”1234567890″,value4:”1234567890″});  
> db.c1.stats()  
{  
“sharded” : true,  
“ns” : “test.c1″,  
“count” : 200003,  
“size” : 25600384,  
“avgObjSize” : 128,  
“storageSize” : 44509696,  
“nindexes” : 2,  
“nchunks” : 15,  
“shards” : {  
“s1″ : {  
“ns” : “test.c1″,  
“count” : 141941,  
“size” : 18168448,  
“avgObjSize” : 128,  
“storageSize” : 33327616,  
“numExtents” : 8,  
“nindexes” : 2,  
“lastExtentSize” : 12079360,  
“paddingFactor” : 1,  
“flags” : 1,  
“totalIndexSize” : 11157504,  
“indexSizes” : {  
“\_id\_” : 5898240,  
“id\_1″ : 5259264  
},  
“ok” : 1  
},  
“s2″ : {  
“ns” : “test.c1″,  
“count” : 58062,  
“size” : 7431936,  
“avgObjSize” : 128,  
“storageSize” : 11182080,  
“numExtents” : 6,  
“nindexes” : 2,  
“lastExtentSize” : 8388608,  
“paddingFactor” : 1,  
“flags” : 1,  
“totalIndexSize” : 4579328,  
“indexSizes” : {  
“\_id\_” : 2416640,  
“id\_1″ : 2162688  
},  
“ok” : 1  
}  
},  
“ok” : 1  
}

Categories: [数据库](http://www.taobaodba.com/html/category/database) Tags:

[Comments (11)](javascript:void(0);) [Trackbacks (1)](javascript:void(0);) [Leave a comment](http://www.taobaodba.com/html/525_525.html#respond) [Trackback](http://www.taobaodba.com/html/525_525.html/trackback)

1. http://1.gravatar.com/avatar/b520f39fcf51d3ed3ad579ad7067d77a?s=32&d=http%3A%2F%2F1.gravatar.com%2Favatar%2Fad516503a11cd5ca435acc9bb6523536%3Fs%3D32&r=G

frank

October 19th, 2010 at 17:34 | [#1](http://www.taobaodba.com/html/525_525.html#comment-15927)

[Reply](javascript:void(0);) | [Quote](javascript:void(0);)

各位DBA大侠们好！我按你们的介绍配置了下MongoDB分片集群，完成之后，我掐掉了primary master，之后会通过投票的方式产生一个新的primary master，但是，为什么当我在写入新的数据的时候，会报“specified is different configdb ”呢？是否需要在哪了配置下？望能指点下！！

1. http://1.gravatar.com/avatar/72fa1f6883edddd0c4a1377b866a5bbf?s=32&d=http%3A%2F%2F1.gravatar.com%2Favatar%2Fad516503a11cd5ca435acc9bb6523536%3Fs%3D32&r=G

[陶方](http://www.taobaodba.com)

November 1st, 2010 at 14:27 | [#2](http://www.taobaodba.com/html/525_525.html#comment-15960)

[Reply](javascript:void(0);) | [Quote](javascript:void(0);)

你的集群是sharding、ReplicaSet，还是混用？  
把集群的结构先贴一下吧。

1. http://1.gravatar.com/avatar/119500f49cbcc2d037d82a7eaccacf09?s=32&d=http%3A%2F%2F1.gravatar.com%2Favatar%2Fad516503a11cd5ca435acc9bb6523536%3Fs%3D32&r=G

soul

December 27th, 2010 at 13:47 | [#3](http://www.taobaodba.com/html/525_525.html#comment-16575)

[Reply](javascript:void(0);) | [Quote](javascript:void(0);)

请问，这三个mongos前端还能放一个lvs做他们的vip吗？  
如果web的数据库链接其中一个mongos地址，当这个mongos挂掉，不是会影响服务吗？

1. http://1.gravatar.com/avatar/72fa1f6883edddd0c4a1377b866a5bbf?s=32&d=http%3A%2F%2F1.gravatar.com%2Favatar%2Fad516503a11cd5ca435acc9bb6523536%3Fs%3D32&r=G

[陶方](http://www.taobaodba.com)

December 27th, 2010 at 16:24 | [#4](http://www.taobaodba.com/html/525_525.html#comment-16577)

[Reply](javascript:void(0);) | [Quote](javascript:void(0);)

在java版本的client api里面，是可以配置多个mongodb的ip的。  
只要配置多个mongos的ip，基本上不会出现影响服务的情况。

其他版本的client api是否支持，需要根据自己的应用情况查看一下文档。

1. http://1.gravatar.com/avatar/119500f49cbcc2d037d82a7eaccacf09?s=32&d=http%3A%2F%2F1.gravatar.com%2Favatar%2Fad516503a11cd5ca435acc9bb6523536%3Fs%3D32&r=G

soul

December 28th, 2010 at 13:03 | [#5](http://www.taobaodba.com/html/525_525.html#comment-16579)

[Reply](javascript:void(0);) | [Quote](javascript:void(0);)

您好，也就是说mongos前面确实是不可以放lvs的吧？

1. http://1.gravatar.com/avatar/72fa1f6883edddd0c4a1377b866a5bbf?s=32&d=http%3A%2F%2F1.gravatar.com%2Favatar%2Fad516503a11cd5ca435acc9bb6523536%3Fs%3D32&r=G

[陶方](http://www.taobaodba.com)

January 18th, 2011 at 23:03 | [#6](http://www.taobaodba.com/html/525_525.html#comment-16605)

[Reply](javascript:void(0);) | [Quote](javascript:void(0);)

mogos对于应用来说是无差别的，前置LVS应该没有问题

1. http://0.gravatar.com/avatar/aecb4e181600ff2ee457e004a5e1de4c?s=32&d=http%3A%2F%2F0.gravatar.com%2Favatar%2Fad516503a11cd5ca435acc9bb6523536%3Fs%3D32&r=G

soul

February 28th, 2011 at 18:11 | [#7](http://www.taobaodba.com/html/525_525.html#comment-16660)

[Reply](javascript:void(0);) | [Quote](javascript:void(0);)

您好，我想请教一下，如果做了表分片的话，那么对这个表请求的时候，mongos是随机将请求分到一台机器呢？还是就固定到一台机器，然后其他的表在固定分给其他的机器？（在不down机的情况下）  
因为现在我发现，我做了集群但是没有分片的情况下，虽然有3台服务器，但是实际上只是一台mongos的那台mongodb在工作，其余两台无非就是一个灾备的作用了。从而导致我这台机器的负载很大。。  
不知道能够给一个会解决的思路呢？

1. http://0.gravatar.com/avatar/2112f8ca637415bc6e6fa54252181beb?s=32&d=http%3A%2F%2F0.gravatar.com%2Favatar%2Fad516503a11cd5ca435acc9bb6523536%3Fs%3D32&r=G

青白眼

March 22nd, 2011 at 14:27 | [#8](http://www.taobaodba.com/html/525_525.html#comment-16684)

[Reply](javascript:void(0);) | [Quote](javascript:void(0);)

请问，在启动mongos的时候，为什么两个config server不能正常启动，必须是1个或者3个config server；我在sharding的文档中并没有找到这个描述。  
“A production shard cluster will have three config server processes, each existing on a separate machine.”也没有说到这个限制。