All about the Cluster Balancer in MongoDB

MongoDB Meetup Oct 2013

David Murphy
DBA, ObjectRocket
@dmurphy_data @objectrocket

Overview

1)What is sharding?

2)What does the balancer do?

3)Migration phases

4)What is an orphan, and why do I care?

5)Balanced chunks, but not disk space? What can you do?

What is sharding?

Definition A method to horizontally scale out software/database layers

Key terms

Shard	A specific replication set part of a larger cluster
Replication Set	Set of mongod servers setup as “Primary’s” and “Secondary's”
Chunk	Logical unit of documents used for moving data between shards
Balancer	Process on mongos that ensures even chunk counts over shards
Config Server	Stores metadata used by mongos
Mongos	Central router for queries to mongo shards

Why?

Needed to enable the linear scalability part to the mongo stack.

Servers have limited resources, enables using more servers to get more resources.

What does the Cluster Balancer do?

Does:

- ☑ Manage chunk count balance
- ☑ Clone chunks
- ☑ Apply oplog to clones
- ☑ Remove duplicate documents

Does not:

- ☐ Split chunks
- ☐ Merge chunk
- ☐ Balance disk space
- ☐ Remove orphans documents

Phase of a chunk migration

1. Parse arguments and prepare
2. Check config and lock
3. Find documents on donor and sort
4. Copy the chunk to destination shard
5. Global lock, log change, update config
6. Clean up and unlock

```
"step1 of 6" : 0,
"step2 of 6" : 202,
"step3 of 6" : 6,
"step4 of 6" : 3001,
"step5 of 6" : 128,
"step6 of 6" : 1652
```

Phase1 - Parse and prepare

1) Check if cluster is running healthy and majority

2) Ensure running in replset not master/slave

3) Ensure collection is replicated

4) Collect metadata for later phases

Phase2 - Check and lock

1)Check for active migration lock

2) Take distributed lock

3) Log event to “*config.changelog*”

4) Do final check of chunk boundaries and owning shard

5) Validate mongod versions match

Phase3 - Find and sort

1) Check locks and versions again

2) Gather all disk locations for documents in the chunk

3) Sort documents, to improve speed of cloning

4) Start copying documents from donor to receiver

Phase4 - Transfer chunk

1) Start a loop with 24 hour TTL
    a) Verify lock is still active
    b) sleeps between runs (time grow over iteration counts)
    c) Check status of chunk copy
    d) Check memory used for migrations, error >500MB

Phase 5 - Lock, log & update

1) Increase MetaVersions
2) Check config server status
    1) If good, update config map for chunks
    2) If not, rollback config and copied chunk
    3) Commit changes to config server, update mongos

Phase 6 - Clean up

1) Check if _waitfordelete usage
    A. added in 2.4 to not block on deletes
2) If false, Queue deletes and mark complete
3) Otherwise start *deleteNow* , regardless of successful delete return true
4) Mark Phase 6 complete in config.changelog

What is an orphan, and why do I care?

Orphan document - Any document that exists in a mongod, but the config map does not reference.

How the happen? - Anytime a balancer fails to fully clean up a chunk orphans can be left. The are very difficult to detect outside odd counts noticed by users/application code.

No Solutions for orphans, however -

10Gen Mongo 2.6+ will have new commands to help clean up orphans

ObjectRocket Utilities - https://github.com/objectrocket/Utils

Balanced chunks…non-balanced disk?

What does balanced mean?

- ObjectRocket - Even distribution of documents or disk space
- MongoDB Inc - Even distribution of chunk (ranges of documents) counts

Deal with balancing on spaces

1. Find large chunks using *"datasize"* command
2. Use "split" command to split the chunks

ObjectRocket Utilities - https://github.com/objectrocket/Utils

Contact

@dmurphy_data
@objectrocket
david@objectrocket.com
https://www.objectrocket.com

WE ARE HIRING! (DBA,DEVOPS, and more)
https://www.objectrocket.com/careers