**  
Báo cáo Sharding trong MongoDB**

*~Phạm Danh Vũ~*



1. **Một số khái niệm cơ bản trong việc sharding Mongo**

Three different components are involved in sharding as follows:

**shard**

A shard is a container that holds a subset of a collection’s data.  A shard is either a single mongod server (for development/testing), or a replica set (for production).

**mongos**

This is the router process.  It routes requests and aggregates responses.  It doesn’t store any data or configuration information, although it does cache information from the config servers.

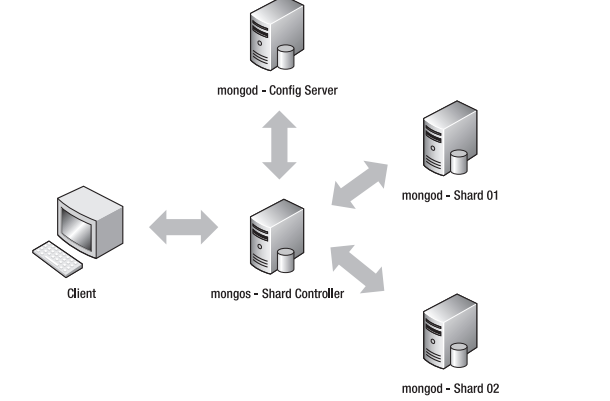
**config server**

Config servers store the configuration of the cluster.  For example, which data is located on which shard.  Used by mongos to determine request routing.

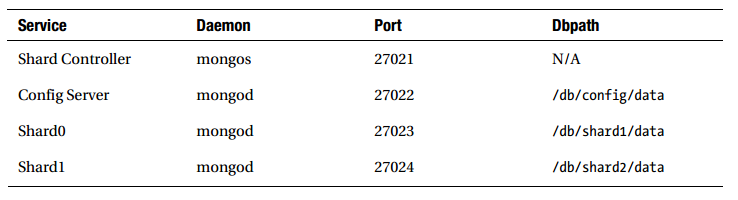
1. **Thực hành**

* Làm việc trên Local host trước khi test trên máy ảo.

Mô hình phân tán:



Chuẩn bị các thư mục và thông tin cấu hình như sau:



**Lưu ý:** Cứ mỗi bước sau phải mở mới 1 terminal chứ không được tắt services của nó.

**Bước 1: Config server**

Set biến môi trường PATH=C:\mongodb\bin

Trong teminal gõ:

C:\Users\PhamDanhVu> mongod.exe --port 27022 --dbpath /db/config/data --configsvr

**Bước 2: Setup the shard controller (Mongos)**

Trong teminal gõ:

C:\Users\PhamDanhVu>mongos.exe --configdb localhost:27022 --port 27021 --chunkSize 1

**Bước 3: Bring up the two servers**

--Tạo thêm thư mục db/shard0/data

Mở terminal thứ 1: gõ

C:\Users\PhamDanhVu>mongod.exe --port 27023 --dbpath /db/shard0/data --shardsvr

Mở termianl thứ 2 : gõ

C:\Users\PhamDanhVu>mongod.exe --port 27024 --dbpath /db/shard1/data –shardsvr

**Bước 4: Config location two shard with sharding system.**

C:\Users\PhamDanhVu> mongo.exe localhost:27021

> use admin

switched to db admin

> db.runCommand( { addshard : "localhost:27023", allowLocal : true } )

{ "added" : "localhost:27023", "ok" : 1 } // ket qua thong bao

> db.runCommand( { addshard : "localhost:27024", allowLocal : true } )

{ "added" : "localhost:27024", "ok" : 1 } //ket qua thong bao

**Bước 5: View infor config the shards**

> db.runCommand({listshards:1})

{

"shards" : [

{

"\_id" : "shard0",

"host" : "localhost:27023"

},

{

"\_id" : "shard1",

"host" : "localhost:27024"

}

],

"ok" : 1

}

**Bước 6: tạo mới 1 Database tên là testdb, trong testdb tạo 1 collection.**

Ở đây ta sử dụng PHP để random dữ liệu khoảng 200 000 document.

**Bước 6.1: Cấu hình PHP - Mongo**

* Phải add driver PHP và Mongo:
  + Lên trang Mongodb.org kiếm driver
    - 🡪 Down driver
    - 🡪 Giải nén 🡪
    - Add 2 cái file .dll đó vào thư viện của PHP ở đường dẫn C:\wamp\bin\php\php5.3.8\ext
    - 🡪 bật WARM và sửa trong file php.ini: thêm 2 dòng này vào

extension=php\_mongo-1.2.10-5.3-vc9-nts-x86\_64.dll

extension=php\_mongo-1.2.10-5.3-vc9-x86\_64.dll

* + Restart services

**Bước 6.2: Tạo file Php random dữ liệu**

<?php

// Open a database connection to the mongos daemon

$mongo = new Mongo("localhost:27021");

// Select the test database

$db = $mongo->selectDB('testdb');

// Select the TestIndex collection

$collection = $db->testcollection;

for($i=0; $i < 200000 ; $i++){

$data=array();

$data['testkey'] = rand(1,100000);

$data['testtext'] = "Because of the nature of MongoDB, many of the more "

. "traditional functions that a DB Administrator "

. "would perform are not required. Creating new databases, "

. "collections and new fields on the server are no longer

necessary, "

. "as MongoDB will create these elements on-the-fly as you access

them."

. "Therefore, for the vast majority of cases managing databases

and "

. "schemas is not required.";

$collection->insert($data);

}

?>

Lưu file Php và chạy file đó lên để insert dữ liệu vào testdb

**Bước 7: Kiểm tra tình trạng dữ liệu insert và dữ liệu phân tán.**

C:\Users\PhamDanhVu> mongo.exe localhost:27021

>use testdb

>db.testcollection.count()

200000 // kết quả

C:\Users\PhamDanhVu> mongo.exe localhost:27023

>use testdb

>db.testcollection.count()

149875

$mongo localhost:27024

>use testdb

>db.testcollection.count()

25125

**Một số vấn đề gặp phải:**

1. Dữ liệu mất mát: shard 1 + shard 2 = 149875 + 25125 = 175000 🡪 Mất đâu 25000
2. Việc phân tán dữ liệu dựa trên cơ chế balance giữa các shard và random. Chưa biết cách tùy biến cái shard key theo ý mình, tức là tôi muốn đặt ở shard 1 các document từ 1 -100 còn shard là các document từ 100 -> chẳng hạn. Bài toán này là phân tán dữ liệu trong nhà hàng. Khó có thể ép buộc shard theo cách đó. Hoặc chưa biết.
3. Triển khai trên là thực hiện trên local host với cách phân tán dữ liệu trên các phân vùng khác nhau. Nếu muốn triển khai sharding trên mạng máy tính thì thay cái vị từ locahost trong các bước cấu hình là địa chỉ IP của máy đó. (Chú ý test connect giữa 2 máy trước khi cấu hình bằng lệnh ping <IP> )