Spring 2019 CS157C: NoSQL Database Systems

Take-Home Midterm Exam

Instructor: Dr. Kim

- Date of posting: Monday, March 18
- Due date: Friday, March 29, 4:00 pm in MH217 (No late mission will be accepted.)
- What to submit: a hard copy of your report containing descriptions about the process to accomplish the given tasks and corresponding screenshots.

A screenshot is expected to include the IP address of the virtual machine from which the screenshot is taken.

I. Problem Description

- Required: In AWS, setup a MongoDB cluster of 3 nodes, set up a sharding system consisting of three shards (one for each node), a replica set consisting of three config servers, and one mongos. (If the config servers are not in a replica set, a full credit may not be given for this task.) You may launch the replica set of config servers in nodes of your choice.
- Replicating shards is not required.

II. Tasks

For each step, describe steps/procedure and include screen-shot(s) to show the task is accomplished. It is your responsibility to select suitable screen-shots(s) to show your work.

1. (10 Points) Set up 3 nodes in AWS
2. (10 points) Access these instances (nodes) through SSH
3. (5 Points) Install MongoDB in each node (i.e. instance)
4. (3 points) Create a directory to store database in each node
5. (5 points) Specify Public and Private IP Addresses of three AWS instances used in your solution.
6. (6 points) Set up and launch three config servers in a replica set.
7. (5 points) Connect mongos to each config server.
8. (10 points) Set up and launch each of three shards: make sure to include the result of sh.status() before adding shards which will be done in the next task.
9. (5 points) Add shards: make sure to include the result of sh.status() after adding shards
10. (5points) Enable shards: explain the nature of the shard key (ascending, random, or location based) and the sharding strategy (range-based or hash based)

11. (10 points) Populate data in the cluster using a public data set : Explain your collection and include the code to populate data and sh.status() after populating data. Specify the URL to the dataset. (Consider the task 12 to choose an appropriate data set for the execution of the given queries. You are allowed to clean and reduce the public data set of your choice to populate a reasonable amount of data to be distributed across the shards. You may decide the reasonable amount.) You are NOT supposed to use zips.json given by the prior assignment.
12. (6 points) Generate the following queries for the populated data. For each query, show its execution time and also show which shard served the query.
    a. A range query to find documents in a given range.
    b. A query involving $elemMatch involving at least two conditions.
    c. A query involving $in, $nin, or $all
    d. A query involving aggregate()
    e. A query involving mapReduce()
    f. An insertion, delete OR update

III. Suggested Reference

The followings are provided for your reference. However, you are primarily responsible to find information to complete the given tasks throughout the midterm.

- MongoDB Sharding Tutorial by Eugine Chang from YouTube
- Launch a Linux Virtual Machine [in AWS] <https://aws.amazon.com/getting-started/tutorials/launch-a-virtual-machine/>
- Install MongoDB [Supported Platforms for MongoDB] <https://docs.mongodb.com/manual/installation/>
- Install MongoDB on Ubuntu <https://docs.mongodb.com/manual/tutorial/install-mongodb-on-ubuntu/>
- Deploy a Sharded Cluster [MongoDB] <https://docs.mongodb.com/manual/tutorial/deploy-shard-cluster/>
- Remotely connecting to MongoDB http interface on EC2 server <http://stackoverflow.com/questions/14653938/remotely-connecting-to-mongodb-http-interface-on-ec2-server>
- Modify Chunk Size in a Sharded Cluster <https://docs.mongodb.com/manual/tutorial/modify-chunk-size-in-sharded-cluster/>
- Import JSON file to mongos <http://stackoverflow.com/questions/38946129/import-json-file-to-mongos>

IV. Grading

Criteria	Maximum Obtainable Scores
12 Tasks	**80 points**
The use of public data set	**5 points**
On time submission	**5 points**
Clarity and Organization of report	**10 points**
Total	**100 points**