MongoDB Exercise. (20 study points)

Setup.

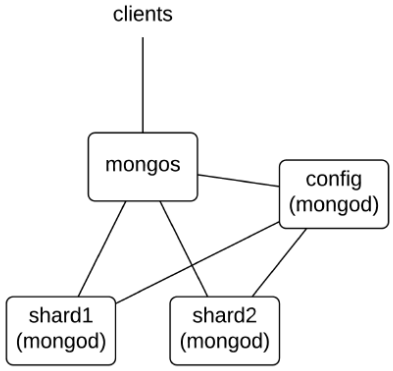
1. Download MongoDB from: https://www.mongodb.com/download-center/community
2. Run MongoDB deamon
3. Download required Dataset from: https://github.com/ozlerhakan/mongodb-json-files

This is a group assignment. Min. 3 preferably 4 members in a group.

Each member setup their computer with MongoDB.

Set each computer except 1 up to be a shard in MongoDB that is shared across each computer.
OPTIONAL: It is optional if the shards is setup on localhost only or as a network of computers described above.

The one exception is setup to be the mongos and mongo config machine this will be the one entry point for the entire setup.



1. Load the Dataset in to 1 of the MongoDB's
2. Setup the sharding environment mentioned in the description above.

Exercise Description.

Setup a small simple website that through the push of a button can display the first 10 document json data from MongoDB and can push some data to the MongoDB Database system created above.

1. Deliver the site code in Github.
2. Together with a document answering the below points.

For this task you need to download twitter dataset from the link mentioned in 3 of the setup section. This time you have to answer query "what are the top 10 hashtags used in the given tweets". To answer this you need to use MapReduce. You can look at the scheme of the collection using db.collection.findOne(). It will print one record with scheme information. Also you can use function like this.hasOwnProperty('field_name') to check if a field exist in the record. (if the field does not exist you will get error.

The following is to be answered in this assignment.

a) What is sharding in mongoDB?
b) What are the different components required to implement sharding?
c) Explain architecture of sharding in mongoDB?
d) Provide implementation of map and reduce function
e) Provide execution command for running MapReduce or the aggregate way of doing the same
f) Provide top 10 recorded out of the sorted result. (hint: use sort on the result returned by MapReduce or the aggregate way of doing the same)

Optional Questions:

g) Show what happens to the data when one shard is turned off.
h) Show what happens to the data when the shard rejoins.
i) Explain how you could introduce redundancy to the setup above.