

Big Data Analytics

Exercise Sheet MongoDB (5 points).

Setup and getting ready:

- 1) (Download mongoDB from <https://www.mongodb.com/download-center/community>)
 - a) Run mongoDB daemon
 - mongod
 - sudo service mongod start
 - b) Start mongoDB shell
 - mongo

Basic mongoDB tutorial <https://docs.mongodb.com/getting-started/shell/introduction/>
- 2) Download required dataset from <https://github.com/ozlerhakan/mongodb-json-files>
There will be two types of dataset available i.e. json and bson.
 - a) From terminal (cmd) To load json dataset you need

```
mongoimport --db earth --collection countries --drop --file countries.json
```
 - b) For bson dataset you need

```
mongorestore --db twitter --collection tweet twitter/tweets.bson
```
- 3) mongoDB and MapReduce <https://docs.mongodb.com/manual/core/map-reduce/>
<https://docs.mongodb.com/manual/sharding/>
- 4) For scripting you can use any of the listed editors at <https://docs.mongodb.com/manual/>

Exercise: Sharding in MongoDB

What is sharding in mongoDB?

What are the different components required to implement sharding?

Explain architecture of sharding in mongoDB?

Exercise: MapReduce with mongoDB (warmup)

As a first exercise you are required to load reddit data from the link mentioned in 2). With help of map and reduce you need to find top 10 "lang" (language) of the documents in reddit.

- a) Provide implementation of map and reduce function
- b) Provide execution command for running MapReduce
- c) Provide top 10 recorded out of the sorted result. (hint: use sort on the result returned by MapReduce)

Exercise: MapReduce with mongoDB (hashtag query) (5 points)

For this task you need to download twitter dataset from the link mentioned in 2). 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.

- a) Provide implementation of map and reduce function
- b) Provide execution command for running MapReduce
- c) Provide top 10 recorded out of the sorted result. (hint: use sort on the result returned by MapReduce)