Beasca, Enriquez, Estrella, Pualengco
CS129.1 ReadMe File

1. **How to load the dataset**
    a. Download the file from https://www.kaggle.com/drgilermo/nba-players-stats
    b. The group figured out that fields there are numerous fields that are blank, making our data inaccurate
    c. The group also found out that only years 1980 to 2014 will be appropriate because it can be clustered into exactly seven parts
    d. Drop years that are blank in reference to the column 3P%, and the years exceeding 2014 and before 1980
    e. Drop players who did not take the 3P shot (0% is different from no percentage at all)

2. **How to setup the Replicate sets**
    a. Start two mongo servers on port 27017 and 27018 by running:

    ```
    mongod --replSet contempo --port 27017 --dbpath C:\data\db_proj1
    mongod --replSet contempo --port 27018 --dbpath C:\data\db_proj2
    ```

    b. Log-in to the mongo server on port 27017 and run rs.initiate:

    ```
    rs.initiate({
       "_id": "contempo",
       "version": 1,
       "members": [
          {
             "_id": 0,
             "host": "localhost:27017",
             "priority": 1
          },
          {
             "_id": 1,
             "host": "localhost:27018",
             "priority": 0
          }
       ]
    });
    ```

    c. On the other server, and the rest of the replicates, run rs.slaveOk()

3. **How to execute the MapReduce functions**

   The mapReduce functions are:

   ```
   db.nbastats.mapReduce(
      function () {
         yrStart = Math.trunc(this.Year / 5) * 5;
         yrEnd = yrStart + 4;
         pos = this.Pos.split('-')[0];

         emit ( {yrStart, yrEnd, pos}, this['3P%'] );
      },
      function (key, values) {
         return Array.sum(values) / values.length;
      },
      { out: "nbastats.avg3PtPctPerYrAndPos" }
   );

   db.nbastats.avg3PtPctPerYrAndPos.mapReduce(
      function() {
         emit (
            { yrStart: this._id.yrStart, yrEnd: this._id.yrEnd },
            this.value
         );
      },
      function (key, values) {
         return Array.sum(values) / values.length;
      },
      { out: "nbastats.avg3PtPctPerYr" }
   );
   ```

   These are two mapReduce functions. The first one outputs the Avg. 3 Pt. % Per Position to nbastats.avg3PtPctPerYrAndPos. Afterwards, this is mapReduced further to get the cumulative, which is then outputted to nbastats.avg3PtPctPerYr

4. **How to shard the MapReduce collection**

   a. To shard, kill start all servers with the --shardsvr parameter.

      ```
      mongod --shardsvr --port 27017 --dbpath C:\data\db_proj1
      mongod --shardsvr --port 27018 --dbpath C:\data\db_proj2
      ```

   b. Then, start a config server. The config server must be a replicate set. Thus, it will be set up as such.

      ```
      mongod --configsvr --replSet contempo_config --port 27019 --dbpath
      C:\data\db_proj_config
      ```

```
mongo --host localhost:27019

rs.initiate({
   "_id": "contempo_config",
   "version": 1,
   "configsvr": true,
   "members": [
      {
         "_id": 0,
         "host": "localhost:27019",
         "priority": 1
      }
   ]
});
```

c. Then, start a router server, and point it to the config server.

```
mongos --configdb contempo_config/localhost:27019 --port 27020
```

d. Log-in to the router server.

```
mongo --host localhost:2720
```

e. Add the two shard servers as shards of the collection.

```
sh.addShard("localhost:27017");
sh.addShard("localhost:27018");
```