CSCI 620/Section 03/Mior, Introduction to Big Data, Spring 2215

Assignment 4 – Document databases

Description

Create the Movies and the Members collection in MongoDB with appropriate types and populate it with IMDB data using the schema and data from Assignment 2.

```
Movies:
{       _id : _,
        type : _,
        title : _,
        originalTitle : _,
        startYear : _,
        endYear : _,
        runtime : _,
        avgRating : _,
        numVotes : _,
        genres : [Short, Documentary, …],
        actors : [ {    actor : _,
                        roles: [X, Y, Z,…] }, … ],
        directors : [ 1, 2, 3, … ],
        writers : [ 4, 5, 6, … ],
        producers : [ 7, 8, 9, … ]}
}
```

```
Members:
{       _id : _,
        name : _,
        birthYear : _,
        deathYear : _
}
```

Note that actors.actor in the Movies collection refers to _id in the Members collection. Note also that each value contained in directors, writers and producers refers to _id in the Members collection. When you have null attributes, the documents must not contain those fields, i.e., it is not valid to insert null or dummy values in MongoDB.

Your tasks

1. Provide a program to load the IMDB data from the relational database into MongoDB. Your program needs to load the whole database in approximately one hour using commodity hardware. **(25 points)**

2. Provide a program issuing the following queries over the MongoDB database. Each should consist of a single query (find and/or aggregation pipeline) and you need to report evidence that you retrieved what was expected. Report the time your queries took to run. **(10 points per query)**

    2.1. Alive actors whose name starts with “Phi” and did not participate in any movie in 2014.
    2.2. Producers who have produced more than 50 talk shows in 2017 and whose name contain “Gill”.
    2.3. Average runtime for movies that were written by members whose names contain “Bhardwaj” and are still alive.
    2.4. Alive producers with the greatest number of long-run movies produced (runtime greater than 120 minutes).
    2.5. Sci-Fi movies directed by James Cameron and acted in by Sigourney Weaver.

3. Provide a brief explanation of the execution plan for each of the previous queries.

    (Hint: https://docs.mongodb.com/manual/reference/explain-results)
    (10 points)

4. Taking the previous queries into account, create appropriate indexes where they are required. Document your decisions, provide the code to generate them, and analyze the performance differences using a program. That is, for each query, show your times with and without indexes.

    (Hint: See https://docs.mongodb.com/manual/indexes)
    (15 points)