MongoDB. Home Task 2

Import Airline Collection

Follow these steps to import airlines collection to your local data base:

1. Save `airlines.csv` on your PC
2. Run local instance of MongoDB
3. Use `mongoimport` to import the collection to the database

```
mongoimport.exe -d frontcamp -c airlines --type csv --headerline --file
<path to airlines.csv>
```

4. Verify that the number of the documents in the `airlines` collection is 186648

Aggregating Airlines Collection

Look at a document from the collection to get familiar with the schema.

Answer the following questions and include both query and the result into your report:

1. How many records does each airline class have? Use `$project` to show result as `{ class: "Z", total: 999 }`
2. What are the top 3 **destination cities outside** of the United States (`destCountry` field, not included) with the **highest average** passengers count? Show result as `{ "avgPassengers" : 2312.380, "city" : "Minsk, Belarus" }`
3. Which carriers provide flights to Latvia (`destCountry`)? Show result as one document `{ "_id" : "Latvia", "carriers" : [ "carrier1", " carrier2", …] }`
4. What are the carriers which flue the most number of passengers from the United State to either Greece, Italy or Spain? Find top 10 carriers, but provide the last 7 carriers (do not include the first 3). Show result as `{ "_id" : "<carrier>", "total" : 999}`
5. Find the city (`originCity`) with the highest sum of passengers for each state (`originState`) of the United States (`originCountry`). Provide the city for the first 5 states ordered by state alphabetically (you should see the city for Alaska, Arizona and etc). Show result as `{ "totalPassengers" : 999, "location" : { "state" : "abc", "city" : "xyz" } }`

Restore Enron Collection

Follow these steps to restore enron collection:

1. Save enron.zip to your PC
2. Unzip the contents to a folder
3. Run local instance of MongoDB
4. Use `mongorestore` to import the collection:

```
mongorestore -d frontcamp -c enron <path to messages.bson>
```

Aggregate Enron Collection

Inspect a few of the documents to get a basic understanding of the structure. Enron was an American corporation that engaged in a widespread accounting fraud and subsequently failed.

In this dataset, each document is an email message. Like all Email messages, there is one sender but there can be multiple recipients.

For this task you will use the aggregation framework to figure out pairs of people that tend to communicate a lot. To do this, you will need to unwind the To list for each message.

This problem is a little tricky because a recipient may appear more than once in the To list for a message. You will need to fix that in a stage of the aggregation before doing your grouping and counting of (sender, recipient) pairs.

Which pair of people have the greatest number of messages in the dataset?

For you reference the number of messages from phillip.love@enron.co to sladana-anna.kulic@enron.com is 144.

Evaluation Criteria

1. Both collections were imported to the DB
2. 40% of tasks for airlines collection have been completed
3. 80% of tasks for airlines collection have been completed
4. All tasks for airlines collection have been completed
5. Same as 4 + task for Enron collection has been completed