DATA8005 – Assignment 2 Part 1

Due: 9th December @ 23:59

Worth: 20% (subject to change)

Overview

We went through a demonstration of creating a cluster of 4 shards with 3 replicas each for a restaurant dataset. The shard key was based on cuisine and then borough. Your task is to create a similar cluster for the Tate data by updating the scripts given to you in cluster.zip.

This time there are to be 3 datacentres (London, Amsterdam and New York). London and Amsterdam have 3 replicas, while New York has 4. You are to choose a shard key for each collection containing one or more fields and explain why you chose the shard keys.

You should update the import script to import both artists.json and artworks.json. Note: artworks.json has been updated and re-uploaded to Blackboard and is larger now.

When imported, the tate database should be about 57 MB in size. Choose an appropriate chunk size.

Aggregation Framework Queries

You have freedom to devise your own queries. Explore the data and come up with some interesting results using the following commands:

- unwind
- project
- group (group by at least 2 fields and add a calculation)
- sort
- limit

Aim for diversity in your queries. The more interesting different sets of results you can produce, the better.

Finally, devise an aggregate pipeline query that uses each of the commands above, but does not replicate anything you did with the commands previously.

For each query you must describe in 1 or 2 sentences what problem you were trying to solve.

Marking Scheme

- 7 marks - Create a set of scripts based on the supplied example to simulate a sharded cluster for the Tate dataset.
- 0.5 mark - Unwind aggregate query
- 0.5 mark - Project aggregate query
- 3 marks - Group aggregate query
- 0.5 mark - Sort aggregate query
- 0.5 mark - Limit aggregate query
- 8 marks - Aggregate pipeline query

The level of complexity will have a bearing on the marks for the group and the pipeline query.