Big Data Storage – Final project

Max. Group members: 5

The percentage for the final score: 45%

Delivery date:

- June 3, 2024, until 23H59.

Project Presentation:

- June 4, 2024.

Final Deliverables:

- A report containing:
  1. A cover page with the **team members' names and student numbers**
  2. One page outlining the design decisions taken by the group and discussing the advantages and disadvantages of the options taken.
  3. Deliver a full backup of the database named **group_xxx.bson**
  4. Deliver a text file with all your queries and aggregations named **group_xxx.txt**
  5. PowerPoint presentation to be presented by the group.
  6. Upload a zip file in Moodle with all the above documents.

NOTES:

- Deliveries are via email. Only one group member should send and add in cc the rest of the members.
- For every day delayed in the delivery, you will be penalised 1 point (up to 5).
- A reference solution for this project will not be available.
- Presentation is mandatory.
- MongoDB Compass is the recommended tool for this assignment.

Description

A. Think about any commercial business process of a product or service that needs a MongoDB database. Describe it in 1 page. Explain why you chose the specific company and why this dataset fits in the MongoDB structure.

B. Select a dataset to be modelled in MongoDB. If your dataset is not in a JSON or CSV file format, it must be converted before being imported into MongoDB. Your database should have at most 10 different collections.

C. Validate that imported data is assigned the correct data type and identify any needed operations to improve the data quality.

D. Add at least 5 validation rules to ensure imported data is valid.

E. Define at least 10 different queries that output interesting business insights from the imported dataset.

F. Optimize your queries using indexes. Include in the report how the usage of indexes impacted query performance.

G. Use at least 5 different aggregations with one or more stages that improve the information of the dataset.

Databases examples:
https://www.kaggle.com/datasets
https://www.kaggle.com/datasets/ylchang/coffee-shop-sample-data-1113
https://www.kaggle.com/datasets/lokeshparab/amazon-products-dataset?select=Badminton.csv
https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset
https://www.kaggle.com/datasets/datasnaek/youtube-new