BIG DATA SYSTEMS – ASSIGNMENT 1

Submission Date: 03 October 2020 11.50 PM

Weightage: 10%

1. Business Context

Bengaluru has more than 12,000 restaurants serving dishes from all over the world. With increasing demand, new restaurants are being opened every day resulting in intense competition to attract customers. The city being an IT capital of India with lot of younger population, most of them are dependent mainly on restaurant food. Sensing the overwhelming demand, "StreetFoodies" an upcoming online food delivery aggregator service planning to make a difference to the existing food delivery business by innovating with Data. The management team of the service is highly convinced about the data driven decision making.

2. Business Problem Understanding

In a stepwise manner, the team want to invest into the data platform that will help them to come up with a recommendation engine that will provide the food restaurant suggestions to the registered customers. For that purpose the first step identified is acquiring the data from external sources, preserving and using it in remaining stages of the data processing. Over the period of time, they are planning to acquire and preserve the data about their own customer base as well.

As any database architect knows, the first question they have to ask is whether to use a SQL or NoSQL database for application. SQL has had a large lead over the non-relational alternatives for decades, but NoSQL is quickly closing the gap with popular databases such as MongoDB, Redis, and Cassandra. SQL still holds 60% with rising demand for systems such as PostgreSQL. They are taking help of this report further to decide upon the database that they can harvest to fulfill their requirements, but has not reached to any conclusion yet.

3. Data Understanding

For this initial prototyping, the management is expecting data science team to explore the various SQL/ NoSQL database options available for the storage and querying the customer data. The customer data is available here. The data dictionary can also be read from the same source.

There are around 35000 customers data is available for the analysis equating to the size of 3 GB, consisting of interesting characteristics of the customers.

4. The Data Engineer Role

The data science team is looking for a savvy Data Engineer to join their growing team of analytics experts. The hire will be responsible for expanding and optimizing the data and data pipeline architecture, as well as optimizing data flow and collection for cross functional teams. The ideal candidate is an experienced data pipeline builder and data wrangler who enjoys optimizing data systems and building them from the ground up. The Data Engineer will support software developers, database architects, and data analysts and data scientists on data initiatives and will ensure optimal data delivery architecture is consistent throughout ongoing projects. The right candidate will be excited by the prospect of optimizing or even re-designing our company's data architecture to support their next generation of products and data initiatives.

5. Performance Considerations

Now after looking at your impressive profile and a rigorous interview process you have been selected as part of this team. The first task that has been given to you is to explore the various, commonly used SQL / NoSQL databases such as

- PostgreSQL
- MongoDB
- Cassandra

Your prototyping will involve loading the given data in these four databases and compare their performance against the below mentioned common database operations. Based upon the performance results that you have obtained, you need to recommend the database option the firm can deploy into their data pipeline.

The database operations that needs to be considered are:

- Write - If a given record / key-value pair is not found in the storage, then the pair is added to the storage. Otherwise, it updates the value for the given key in the storage. This operation therefore combines Create and Update operations of the CRUD model.
- Read - This reads the value corresponding to a given record / key from storage. This is the same as the Read operation of the CRUD (Create, Read, Update, and Delete) model commonly used to describe database operations.
- Delete - This deletes the record (i.e. key-value pair) corresponding to a given key from the key-value pair storage. This is the same as the Delete operation of the CRUD model.
- GroupBy and OrderBy – This operation involves grouping of the records and ordering by certain criteria.

6. Expected Outcomes

As a result of this prototyping exercise, you need to submit a report to the stakeholders interested for this database recommendation.

The report should consists of a document / presentation containing details about

a) The schema / structure / other representation used in each of the databases
b) The exact queries used for each of the operation mentioned above for each database
c) Various performance parameters considered (with reasons) while executing the queries and their analysis
d) Visualizations based on the performance parameters used for the queries and their interpretation
e) Tabular , quick summary of the prototyping exercise
f) Clearly spelt out recommendation
g) Any other observation / finding that you feel is important for this exercise

General Notes:

- Refer the document used while registering the groups. In case of discrepancies, write to me separately (copying all your group members) with subject line as "DSE BDS Group <your_group_number>". email - ppawar@wilp.bits-pilani.ac.in
- Using the Canvas, only the first member of group (as listed in the above mentioned doc) has to upload the file. No submission over email will be considered.
- Name the document file in format like "Grp_<your_group_number>.doc" only. Don't add anything into the file names.
- Make sure that you upload the file well ahead of deadline. At last moments, we have seen several groups have faced issues while doing the submissions.
- Note - As it's a group assignment, only one submission is expected from each group. Unnecessary don't upload the solution on individual basis. If it's observed, then the penalty (25% reduction) will be applicable on it.

7. References

- MongoDB documentation
- PostgreSQL docs
- Cassandra docs
- Redis docs
- A performance comparison of SQL and NoSQL databases
- Groups information