Assignment-1: Porting University RDBMS to MongoDB and Running Workloads using Apache Spark

Course: Big Data Analytics

Objective:

This assignment aims to give students hands-on experience in migrating a traditional relational database management system (RDBMS) to a NoSQL database (MongoDB) and evaluating queries on the migrated data using Apache Spark. This assignment will cover the following key areas:

1. Data modelling and transformation from a relational model to a document-based model.
2. Data migration and loading into MongoDB.
3. Querying the data using Apache Spark with MongoDB as the data source.
4. Performance analysis and optimization techniques.

Scenario:

We have a university's existing student information system in PostgreSQL. The system includes tables such as `Students`, `Courses`, `Enrollments`, `Instructors`, and `Departments`. The university wishes to migrate this data to a MongoDB database to leverage its flexibility and scalability.

Requirements:

1. Data Modeling and Schema Design (5 points)

- Analyze the existing RDBMS schema.
- Design a document-based schema in MongoDB that supports the same functionalities as the relational model.
- Ensure the schema design supports efficient querying and data retrieval per the following query workload.
    - Fetching all students enrolled in a specific course.
    - Calculating the average number of students enrolled in courses offered by a particular instructor at the university.
    - Listing all courses offered by a specific department.
    - Finding the total number of students per department.
    - Finding instructors who have taught all the BTech CSE core courses sometime during their tenure at the university.
    - Finding top-10 courses with the highest enrollments.

Deliverables:

- A report describing the mapping from the relational schema to the MongoDB schema. Include justifications for design choices and any denormalization performed.

2. Data Migration (5 points)

- Implement a data migration pipeline to transfer data from the RDBMS to MongoDB. This may include ETL (Extract, Transform, Load) processes.
- Ensure data consistency and integrity during migration.

Deliverables:

- A script or set of scripts used for the data migration process.
- A report on the data migration steps, including any data cleaning or transformation processes applied.

3. Query Implementation using Apache Spark (10 points)

- Set up an environment where Apache Spark can interact with the MongoDB database.
- Implement all the queries given above on the MongoDB data using Apache Spark.

Deliverables:

- The Spark scripts or notebooks containing the implemented queries.
- A report detailing the query results and any observations on performance.

4. Performance Analysis and Optimization (5 points)

- Analyze the performance of the queries executed in Apache Spark.
- Suggest and implement at least two optimization strategies (e.g., indexing in MongoDB, query optimization in Spark, data partitioning).
- Compare the performance before and after optimization.

Deliverables:

- A report detailing the performance analysis, optimization strategies implemented, and their impact on performance.

5. Documentation and Presentation (Mandatory to get marks)

- Compile all deliverables into a comprehensive report.
- You may be asked to present your findings in a 10-15 minute presentation to the instructor, highlighting key aspects of the migration, querying, and optimization processes.

Submission Guidelines:

- Submit all scripts, reports, and presentation materials to the course's submission portal in a single zip file.
- Ensure that all files are properly labelled and organized.