Unit - 5

PART - A

1. What is NoSQL? List its types.
2. Compare RDBMS and Mongo DB.
3. Provide an example of a query written in Hive Query Language (HQL) to retrieve data from a Hive table.
4. Differentiate between a Document store and a Graph store.
5. Justify the need for sharding in database management system.
6. Differentiate between partitioning and buckets in Hive.
7. Provide an example of a query written in Hive Query Langauage (HQL) to retrieve data from a Hive table.
8. State the three components of the CAP theorem.
9. List the features of Hive.
10. What is CAP theorem?
11. Justify the need for sharding.
12. State the limitations of Hive.

PART – B

1. Discuss NoSQL databases, including their characteristics and various types. Name and describe current tools and technologies used in the NoSQL ecosystem.A data scientist is working on a project that involves merging data sets from different sources. Discuss the challenges they might encounter during the data munging process and propose strategies to address these challenges

2. Illustrate the architecture of Apache Hive in detail, explaining its components and how they interact. Discuss Hive Query Language (HQL) with examples to demonstrate its usage.

3. A content management system needs to store and retrieve a variety of content types including text ,images and videos. Explain how MongoDB's support for flexible schema and various data types makes it suitable choice for this scenario.

4. A rapidly growing online retail store is experiencing increased database loads. Propose a sharding strategy for MongoDB to handle the scalability requirements of the growing dataset. Consider factors such as data distribution and query patterns.

Part – C

1. Ecolab is a chemical company that wants to go the machine learning way of doing things to improve productivity. Since Ecolab not familiarize with ML they wish to automate the ML process. Suggest a solution architecture so that they can improve productivity and faster deployment of models.

2. A healthcare organization is developing a distributed system to manage patient records, appointment scheduling, and real-time monitoring of vital signs across multiple locations. The system must ensure that healthcare providers can access up-to-date patient information at all times, even during network issues. Evaluate the trade-offs between Consistency, Availability, and Partition Tolerance in the design of this distributed healthcare system.

3. A company is handling a large volume of user generated data, including user profiles, preferences and interaction history. Recommend a suitable type of NOSQL for efficiently managing this diverse and dynamic dataset. Justify your choice.