Nirma University

Institute of Technology

Semester End Examination (IR), December 2025

M.Tech. CSE (Data Science) Semester-I

6CS271ME25 – BIG DATA SYSTEMS

Roll / Exam No. 25MCD005

Supervisor's initial with date

Time: 3 Hours

Max. Marks: 100

Instructions:
1. Attempt all questions.
2. Figures to the right indicate full marks.
3. Use a section-wise separate answer book.
4. Make suitable assumptions wherever necessary, and specify it correctly.

SECTION-I

Q-1. Answer the following. **[16]**

(a) **CO1, BL2** A healthcare analytics company collects patient data from multiple sources, including wearable devices, hospital databases, and electronic health records. The data volume is rapidly increasing, and the company aims to use this data to predict disease outbreaks and personalise treatment plans. Based on the above scenario, outline the significance of Big Data in enabling effective healthcare analytics and discuss the key challenges of Big Data in modern computing. [08]

(b) **CO2, BL3** Consider the following dataset stored as sales_data.csv file. Write answers in BOTH Pig Latin and Hive HQL. [08]

emp_id	name	department	sales	month	city
201	Amit	Electronics	120000	January	Ahmedabad
202	Neha	Clothing	85000	January	Surat
203	Rahul	Electronics	145000	February	Vadodara
204	Kriti	Grocery	65000	February	Ahmedabad
205	Mehul	Clothing	92000	March	Surat
206	Riya	Electronics	110000	March	Ahmedabad

a) List all employees who belong to the 'Electronics' department.
b) Find the total sales done in each department.
c) Display the employee details whose sales amount is more than 1,00,000.
d) Find the average sales for each city.

Q-2. Answer the following. **[16]**

(a) **CO1, BL1** Explain the architecture of the Hadoop Distributed File System (HDFS). Describe the roles of the NameNode and DataNode with a diagrammatic representation. [08]

(b) **CO4, BL5** Explain how the Shuffle and Sort phase can become a bottleneck in a MapReduce job. What factors influence its performance, and how can it be optimised? MapReduce provides fault tolerance by re-executing failed tasks. Critically analyse how frequent task failures affect job run time and cluster utilisation. [08]

Q.3 Answer the following. **[18]**

(a) **CO3, BL3** A distributed Hadoop cluster is processing a dataset of 2,500 GB using MapReduce. The HDFS block size is 256 MB. The replication factor is 3. Each map task processes exactly one block. On average, each map task takes 60 seconds to execute. The cluster can run 80 map tasks in parallel. Interpret the following: [08]

(i) How many map tasks will be created to process the dataset?
(ii) What is the total storage requirement in the cluster due to replication?

(b) **CO4, BL6** Design a high-level MapReduce algorithm to calculate the frequency of each unique word in a large corpus of text files. Clearly specify the Input, Mapper, Intermediate, and Reducer key-value pairs. Provide a suitable example with a diagram. [10]

SECTION -II

Q.4 Answer the following. **[16]**

(a) **CO2, BL3** Differentiate between SQL and NoSQL databases based on consistency model (ACID vs. BASE), scalability, and data structure. [08]

(b) **CO2, BL2** Explain the CRUD operations and data models used in Document Databases (like MongoDB). Give an example of a relevant use case where a Document Database is preferred over an RDBMS. [08]

Q.5 Answer the following. **[16]**

(a) **CO2, BL4** Discuss and differentiate the four major types of NoSQL databases (Document, Key-Value, Column-Family, Graph). Analyze the analytical factors that should guide the choice of one type over another. [08]

(b) **CO3, BL4** Describe the architecture and function of Apache Hive. Examine how Hive facilitates data querying in Hadoop by using the concept of a Metastore. [08]

Q.6 Answer the following. **[18]**

(a) **CO4, BL5** Evaluate the performance of Hadoop applications. Explain two key metrics for analyzing job execution efficiency and discuss the factors influencing job scheduling and data processing efficiency in YARN. [8]

(b) **CO4, BL6** Explain how CRUD operations in graph databases differ from document databases in terms of data relationships, indexing, and query expressiveness. Investigate whether graph queries are often more efficient for relationship-heavy datasets. [10]