Nirma University

Institute of Technology

Sessional Examination (Minor), Sep 2025

M Tech. (Data Science), Semester-I

6CS271ME25 - BIG DATA SYSTEMS

Roll / Exam No. 25MCD005

Supervisor's initial with date

Time: 2 Hours

Max. Marks:50

Instructions:
1. Attempt all questions.
2. Figures to right indicate full marks.
3. Draw neat sketches wherever necessary.
4. Assume dataset if needed.

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

A. CI01-BI02 How do you define "block" in HDFS? What is the default block size in Hadoop v1 and in Hadoop v2? Can it be changed? Describe in your words. (4)

B. CI02-BI03 You are at city shopping mall. You see few people are browsing the items. Some of them are looking for discounts. Some of them are filling feedback form. Few people are at billing counter. You may consider other things and events happening in this scenario. Think for while on the different types of data generated. Categorize each data source into appropriat category, by considering the Variety and velocity of each source. (6)

C. CI03-BI03 Draw HDFS Architecture. Explain following two commands of HDFS with syntax and at least one example of each. (6)
(i) copyFromLocal (ii) checksum

Q-2. Answers the following. **[18]**

A. What are the benefits of Big Data? Discuss challenges under Big Data. How Big Data Analytics can be useful in the development of smart cities? (6)

B. CI02-BI05 In a large-scale Hadoop cluster, how do replication strategies and NameNode high-availability mechanisms collectively ensure fault tolerance and continuous data availability, and what trade-offs arise between performance, storage overhead, and recovery time? (6)

C. CI02-BI03 In the YARN architecture, the ResourceManager allocates containers while the ApplicationMaster manages the lifecycle of an application and communicates with NodeManagers running on worker nodes. Analyze how the coordination among the Application Master, containers, and worker nodes ensures efficient resource utilization and fault tolerance in a Hadoop cluster. What challenges may arise in container allocation and worker node management under high workload conditions? (6)

Q-3. Answers the following. **[16]**

A. **C103-B104** Briefly explain the role of the Hadoop ecosystem in managing and analyzing big data. In a large-scale data analytics workflow, how do HDFS (storage), YARN (resource management), MapReduce (batch processing), and Hive (data warehousing) work together to ensure scalability and efficiency, and what performance trade-offs arise when integrating these four components? (8)

B. **C104-B105** Suppose you are analyzing a large web server log dataset (500 GB) to count the frequency of each unique IP address accessing the server. Using the MapReduce framework, explain step by step how the input is divided into splits, processed by Mapper, shuffled/sorted, and finally aggregated by Reducer. Illustrate your explanation with a block diagram of the MapReduce execution flow. (8)

****