

104-24



VIT

Vellore Institute of Technology

Slot : G2+TG2

# School of Computer Science Engineering and Information Systems

FALL Semester 2025-2026

Continuous Assessment Test – I

Programme Name &amp; Branch: B.Tech. IT

Course Name &amp; code: BITE411L BIG DATA ANALYTICS

Class Number (s): VL2025260107397

Faculty Name (s): Dr. Swarna Priya RM

Exam Duration: 90 Min.

Maximum Marks: 50

General instruction(s):

Answer ALL questions.

| Q.No. | Question                                                                                                                                                                                                                                                                                                                                                                           | Max Marks | CO  | BL  |
|-------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|-----|-----|
| 1.    | Analyze the following scenarios and identify the type of analytics (Descriptive, Diagnostic, Predictive or Prescriptive) that is best suited to address the problem. Justify your choice for each scenario.<br>Scenario 1: A grocery delivery company wants to determine which neighborhoods experienced the highest growth in delivery requests over the past year.               | 4         | CO1 | BL4 |
| (a)   | Scenario 2: A manufacturing firm notices an unexpected increase in defective products over the last month. They want to uncover the root cause of the defects.                                                                                                                                                                                                                     |           |     |     |
| (b)   | Analyze the following scenarios and identify the most appropriate data analysis method (Regression Analysis, Cohort Analysis, Factor Analysis, or Monte Carlo Simulation) for each. Justify your choice.<br>Scenario 1: A marketing team wants to understand the relationship between advertising spend, social media engagement, and total sales revenue to predict future sales. | 6         |     |     |
|       | Scenario 2: A subscription-based online service wants to compare the retention rates of users who joined during different promotional campaigns.                                                                                                                                                                                                                                   |           |     |     |
|       | Scenario 3: A research team is analyzing survey responses to uncover latent variables (such as customer satisfaction drivers) influencing overall customer ratings.                                                                                                                                                                                                                |           |     |     |
|       | Scenario 4: A logistics company is evaluating the risks of supply chain disruptions under various random events, such as delays, demand fluctuations, and transportation failures.                                                                                                                                                                                                 |           |     |     |

|           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |    |     |     |
|-----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|-----|-----|
| 2.        | <p>A bank's fraud detection system monitors transactions in a bitstream, where 1 indicates a potentially fraudulent transaction and 0 indicates a normal transaction. The system keeps a sliding window of the last 60 transactions:</p> <p>0,1,0,1,1,0,1,0,1,1,1,0,1,1,0,1,0,1,1,0,1,1,0,1,1,0,<br/>0,1,1,0,1,1,1,0,1,1,1,0,1,0,1,1,1,0,1,1,1,0,1,1,1,0,1,1,1,0</p> <p>(a) Store the bitstream using DGIM buckets.</p> <p>(b) Estimate the number of fraudulent transactions (1s) in the last 32 transactions using DGIM.</p> <p>(c) Discuss the accuracy and memory efficiency of DGIM compared to storing the entire stream.</p>                                                                                                                                                                                 | 10 | CO2 | BL4 |
| 3.<br>(a) | <p>A Bloom filter is used to store up to <math>n=5000</math> elements with a desired false positive probability of <math>P=0.05</math>. Calculate the required size of the bit array (<math>m</math>) and the optimal number of hash functions (<math>k</math>). (4 Marks)</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 10 | CO2 | BL3 |
| (b)       | <p>A Bloom filter is configured with the following parameters: (6 Marks) Size of the bit array (<math>m</math>): 20 bits. Number of hash functions (<math>k</math>): 3. The hash functions produce the following indices:</p> <p>Hash 1: <math>\text{index}=(\text{element}\%7)</math></p> <p>Hash 2: <math>\text{index}=(\text{element}\times3\%11)</math></p> <p>Hash 3: <math>\text{index}=(\text{element}\times5\%13)</math></p> <p>Perform the following operations:</p> <p>Insert Operation: Insert the elements {3, 7, 10} into the Bloom filter. Show the state of the bit array after each insertion.</p> <p>Search Operation: Check if the following elements are present in the Bloom filter: {3, 5, 7, 12}. Explain whether each element is "definitely not present" or "possibly present" and why?</p> | 10 | CO2 | BL3 |
| 4.        | <p>Consider an HDFS client writing a file named "MIS.csv" of 370 MB with a block size of 128 MB and a replication factor of 3. Explain how HDFS handles reading and writing across multiple nodes with different IP addresses, and illustrate the process with a neat sketch. Assume the NameNode assigns three different DataNodes to store each block. If the replication factor is set to 3, determine how many copies of each block will exist in HDFS. Additionally, analyze what happens if a DataNode storing Block 2 fails, and describe how HDFS handles such failures. Finally, discuss the impact of increasing the replication factor to 4 instead of 3, considering aspects</p>                                                                                                                        | 10 | CO2 | BL3 |

| Block Number | DataNode (Primary) | Replicas                   |
|--------------|--------------------|----------------------------|
| Block 1      | 192.168.1.10       | 192.168.1.20, 192.168.1.30 |
| Block 2      | 192.168.1.15       | 192.168.1.25, 192.168.1.35 |
| Block 3      | 192.168.1.40       | 192.168.1.50, 192.168.1.60 |

such as storage requirements, data availability, fault tolerance, network traffic, and performance.

|                                                                                                                                                                                                                                                                                                                                                                                                                            |    |     |     |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|-----|-----|
| <p>You are given a dataset containing text from AI research discussions. The dataset consists of short phrases related to machine learning. Your task is to apply MapReduce to analyze word frequency and explain the steps involved in solving this problem. Input Text: "machine learning is powerful" "deep learning is a subset of machine learning" "simulated neural networks, are a subset of machine learning"</p> | 10 | CO5 | BL3 |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|-----|-----|

NOTE\*: Please refer below to the BL – Bloom's Taxonomy Levels and mention the respective level in the questions.

| Bloom's Levels | Taxonomy | Category      |
|----------------|----------|---------------|
| BL1            |          | Remembering   |
| BL2            |          | Understanding |
| BL3            |          | Applying      |
| BL4            |          | Analyzing     |
| BL5            |          | Evaluating    |
| BL6            |          | Creating      |