REG.NO.:

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE ENGINEERING AND INFORMATION SYSTEMS
CONTINUOUS ASSESSMENT TEST - II
WINTER SEMESTER 2025-2026

SLOT: B1 + TB1

Programme Name & Branch	**: B.Tech. Information Technology**
Course Code and Course Name	**: BITE411L – Big Data Analytics**
Faculty Name(s)	**: Dr. Balasubramani M, Dr. GN.Vivekananda, Dr. MohanaPriya M**
Class Number(s)	**: VL2025260502133, VL2025260502139, VL2025260502136**
Date of Examination	**: 16/03/2026 (AN)**
Exam Duration	**: 90 minutes**

Maximum Marks: 50

General instruction(s):

- Answer All Questions
- M - Max mark; CO – Course Outcome; BL – Blooms Taxonomy Level (1 – Remember, 2 – Understand, 3 – Apply, 4 – Analyse, 5 – Evaluate, 6 – Create)

Course Outcomes:

4. Process Data using Spark and No SQL Databases.
5. Apply MapReduce based analysis.

Q. No	Question	M	CO	BL
1.	A government organization maintains a large national census dataset stored in HDFS, containing millions of records with attributes such as region, occupation, and income level. To analyze employment distribution across regions, the dataset is processed using Hadoop MapReduce running on YARN within a distributed cluster environment. Explain how the YARN architecture components coordinate to execute this analytics task. In your response, discuss how components manage cluster resources and job execution, and demonstrate how a MapReduce program can be designed to process the census dataset to determine the total number of individuals belonging to each occupation category across different regions.	10	5	L3
2.	A financial services company processes large volumes of stock market transactions and real-time trading data every minute. To analyze patterns such as price volatility, trading trends, and market anomalies, the organization deploys Apache Spark on YARN within its Hadoop ecosystem so that Spark applications can run alongside other distributed workloads in the cluster. Analyze how Apache Spark applications operate when deployed on YARN to process this high-volume financial data. Illustrate the Spark architecture, and explain how RDD operations can be applied to analyze stock transaction data, and evaluate how Spark's in-memory distributed processing model improves performance and scalability compared to traditional disk-based processing frameworks such as Hadoop MapReduce.	10	5	L4
3.	A smart city infrastructure platform collects continuous data from traffic sensors installed at multiple intersections, including vehicle count, average speed, location ID, and timestamp. Because the system must handle massive volumes of streaming sensor data and support fast random-access queries, the data is stored using HBase on top of HDFS. Based on the above scenario, explain how HBase architecture supports the storage and processing of large-scale traffic sensor data.	10	4	L4

REG.NO.:

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE ENGINEERING AND INFORMATION SYSTEMS
CONTINUOUS ASSESSMENT TEST - II
WINTER SEMESTER 2025-2026

SLOT: B1 + TB1

	Write appropriate HBase commands for the following tasks: • Create a table named traffic_data with suitable column families for storing sensor information. • Insert a record containing vehicle count and average speed for a specific location and timestamp. • Retrieve all traffic records for a given location ID. • Display the structure of the created table. • Scan the table to view all stored traffic sensor data.			
4.	A large e-commerce company collects semi-structured customer activity data such as user ID, product ID, product views, purchases, and timestamps from multiple web and mobile applications. Since the data is generated from different services, it is stored in JSON format in HDFS. Based on the above scenario, analyze the architecture of Hive and explain how its components work together to process analytical queries on large datasets. Further, explain how JSON files can be handled and analyzed using Hive, including how semi-structured JSON and XML data can be integrated into Hive tables and processed for large-scale data analysis in this environment.	10	4	L4
5.	A digital news analytics company is developing an automated system to classify online articles into three categories: Technology, Business, and Health. The system uses a Complementary Naïve Bayes classifier to categorize new articles based on the occurrence of specific keywords in the training dataset. Consider the following word frequency counts obtained from the training dataset:	10	4	L5

Word	Technology	Business	Health
algorithm	20	3	2
investment	2	18	3
vaccine	1	2	22
data	15	4	3

A new article contains the following words: {algorithm, data, investment}.

Assume the vocabulary size |V| = 4 and apply Laplace smoothing (+1) while computing probabilities. For the Complementary Naïve Bayes classifier, compute word probabilities using the complement of each class. Using the CNB approach, compute the class scores for Technology, Business, and Health and determine the most appropriate category for the article.
Further, critically analyze the differences between the Naïve Bayes and CNB in terms of probability estimation, class representation, and handling of feature distributions, and evaluate how these differences influence classification performance when applied to large-scale text datasets.

****