U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

June 2025 Semester End Main Examinations

Programme: B.E. **Semester: VI**

Branch: Artificial Intelligence and Machine Learning **Duration: 3 hrs.**

Course Code: 24AM6PEBDA **Max Marks: 100**

Course: Big Data Analytics

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

				CO	*PO*	Marks
			UNIT - I			
1	a)		Describe the key characteristics of big data with suitable examples for each.	*CO1*	*PO1*	**10**
	b)		As part of the Netflix analytics team, analyze data such as viewing history, user preferences, and watch times to propose specific Big Data-driven strategies for improving content recommendations and increasing user retention.	*CO1*	*PO2*	**10**
			OR			
2	a)		Differentiate between Structured, Semi-structured and Unstructured data with examples.	*CO1*	*PO1*	**8**
	b)		Compare ORC, Parquet, and Avro file formats based on their structure, performance, and suitability for various Big Data processing use cases.	*CO1*	*PO2*	**12**
			UNIT - II			
3	a)		Examine the HDFS file system and illustrate the interaction between clients, the NameNode, and DataNode during file operations.	*CO1*	*PO1*	**12**
	b)		Given a big data application that transfers large volumes of data between distributed systems, explain how you would implement data serialization and deserialization to optimize data storage and network communication? Illustrate the approach using examples of common serialization formats employed in Big Data environments.	*CO1*	*PO2*	**8**
			OR			
4	a)		Outline the architecture of YARN and examine how it facilitates resource management and task scheduling in Big Data processing?	*CO1*	*PO1*	**10**
	b)		Given a large collection of text files stored in HDFS, you need to count the frequency of each word across all files. Outline the implementation of the map and reduce functions in a Word Count application, highlighting their roles and how they process data to achieve the desired result.	*CO1*	*PO2*	**10**
			UNIT - III			
5	a)		Given an inventory database in Hive with the following Products and Sales tables:	*CO2*	*PO5*	**10**

			Product table: product_id INT, product_name STRING, category STRING, stock_quantity INT, price DECIMAL(10, 2) Sales table: sales_id INT, product_id INT, sales_date DATE, quantity_sold INT, sales_amount DECIMAL(10, 2) Write a Hive query to: i. Find the total stock value for each category (i.e., stock_quantity price). Group the results by category and display the category and the total stock_value in descending order. ii. Calculate the total quantity sold and total sales amount for each product. Group the results by product_name and display the product_name, total_quantity_sold, and total_sales_amount, sorted by total_sales_amount in descending order.			
		b)	In a system using HBase to store large-scale time-series sensor data, a user application needs to efficiently retrieve the latest data point for a given sensor ID. Based on the architecture and data flow of HBase, explain how the data retrieval request flows through HBase's components (from client to storage), and identify two architectural-level optimizations that can be applied to ensure low-latency reads?	*CO3*	*PO3*	**10**
			OR			
	6	a)	Peter is a data engineer at a company who handles large volumes of web server logs. The team has decided to use Apache Hive to manage and analyze these logs. The logs are stored in HDFS (Hadoop Distributed File System), and the task is to set up a Hive-based solution to query this data efficiently. Provide a diagram illustrating the Hive architecture and its integration with Hadoop components for managing and analyzing the logs. Write the query for the following: i. Create a Hive table to store the web server logs, assuming the logs are in a structured format with fields like 'timestamp', 'user_id', 'page_url', and 'response_time'. ii. Compute the average response time for each 'page_url' and list the top 5 pages with the highest average response time.	*CO2*	*PO3*	**10**
		b)	As a data engineer managing a real-time analytics system using Apache HBase, the task is to set up a new table called user_activity to store website clickstream data with the following structure: Row key: User ID + timestamp Column families: meta (to store browser, location) event (to store clicked URL, event type)	*CO3*	*PO5*	**10**

			Using appropriate HBase shell commands, describe how you would: i. Create the user_activity table with the specified column families. ii. Insert a record with user ID u123, timestamp 20250527T103000, browser Chrome, location NYC, clicked URL https://example.com/page1, and event type click. iii. Retrieve the full row for that user and timestamp. iv. List all tables in the system. v. Check the number of rows in the user_activity table. Provide relevant HBase shell commands for each step. Briefly explain what each command does.			
			UNIT - IV			
	7	a)	Ram has been assigned to migrate selected data from a MySQL database to Hadoop using Apache Sqoop. The MySQL database sales_db contains several tables, including customers, orders, and products. write the appropriate Sqoop commands for the following and briefly explain what it does.: i. Import the entire customers table into HDFS. Import only the orders from the last 7 days. ii. Perform an incremental import of the orders table based on the order_date column. iii. Import all tables from the sales_db database into a specified HDFS directory. iv. List all databases and all tables in sales_db using Sqoop.	*CO3*	*PO5*	**10**
		b)	The task is to set up a log collection system using Apache Flume to stream application logs from multiple servers to HDFS in near real-time. i. Describe the Flume architecture, clearly explaining the role of each of the following components: Agent, Source, Channel, and Sink. Use a diagram to illustrate the interaction between these components. ii. Outline the dataflow process in Flume from log generation to storage in HDFS. Include how events are handled internally by Flume and how Flume ensures delivery guarantees.	*CO2*	*PO2*	**10**
			OR			
	8	a)	The task involves designing a data ingestion pipeline for a large-scale log processing system using Apache Flume. The system must collect logs from multiple web servers and deliver them reliably to HDFS for long-term storage and analysis. i. Explain the key components of Flume architecture (Agent, Source, Channel, Sink) and their roles in the data flow. ii. Design a Flume flow that reads log data from web server log files, buffers the events in memory temporarily, and writes them to HDFS. Specify what types of Source, Channel, and Sink you would use. Justify your choices. Elucidate to ensure reliability and fault tolerance in this Flume configuration.	*CO3*	*PO5*	**10**

		b)	In a Big Data project where data from a traditional RDBMS needs to be integrated into Hadoop for analytics, the team suggests using Apache Sqoop for this task. i. What is Apache Sqoop, and why is it important in Big Data ecosystems? ii. List and briefly explain four key features of Apache Sqoop that make it suitable for large-scale data transfer. iii. Explain the working mechanism of Sqoop import and export operations, including how data moves between relational databases and the Hadoop ecosystem.	*CO3*	*PO3*	**10**
			UNIT - V			
	9	a)	Alex is a business analyst at a large e-commerce company. The company wants to leverage Apache Spark to analyze customer behavior and improve marketing strategies. Currently, the company collects vast amounts of data from multiple sources, including website clicks, purchase history, and customer reviews, stored in a distributed file system. i. Elucidate how Apache Spark's architecture can handle and process large-scale data from multiple sources in this scenario. ii. With a neat diagram, illustrate the working of Apache Spark.	*CO3*	*PO3*	**10**
		b)	Given the following data: val numberPairs: Array[(Int, Int)] = Array((1, 2), (3, 4), (5, 6), (7, 8)) i. Use a higher-order function and an anonymous function to compute the sum of each pair in the numberPairs array. Store the result in a new array. ii. Convert the result into a Map where the sum is the key and the original tuple is the value. iii. Filter the map to keep only those entries where the sum is greater than 10, using an anonymous function. iv. Explain the role of anonymous functions and higher-order functions in functional programming. Give one benefit of using them in data processing.	*CO4*	*PO3*	**10**
			OR			
	10	a)	Apache Spark is a widely-used Big Data processing framework known for its speed and ease of use. Describe the role of key components of Spark's architecture. Use a simple diagram to support your explanation. Also list and explain five major features of Apache Spark that make it suitable for large-scale data processing compared to traditional systems like MapReduce.	*CO2*	*PO5*	**10**
		b)	Write a Scala program that accepts a list of integers and performs the following operations: i. Declare a mutable variable to store the sum of even numbers. Use a for loop to iterate over the list and add only the even numbers to the sum variable. After the loop, use an if-else expression to print whether the sum is greater than 100 or not. Print "Large Sum" if it is greater than 100, otherwise "Small Sum". ii. Explain the difference between mutable (var) and immutable (val) variables in Scala, with examples.	*CO4*	*PO3*	**10**

****