AI3621A

USN 1 R V 2 2 A I 0 6 7

RV COLLEGE OF ENGINEERING®

(An Autonomous Institution Affiliated to VTU)

VI Semester B. E. Regular Examinations August-2025

Artificial Intelligence and Machine Learning

BIG DATA TECHNOLOGIES

Time: 03 Hours *Maximum Marks: 100*

Instructions to candidates:

1. Answer all questions from Part A. Part A questions should be answered in first three pages of the answer book only.
2. Answer FIVE full questions from Part B. In Part B question number 2 is compulsory. Answer any one full question from 3 and 4, 5 and 6, 7 and 8, 9 and 10.

PART-A

			M	BT	CO
1	1.1	Mention two key differences between HDFS and traditional RDBMS.	02	1	1
	1.2	What is the purpose of the NameNode in HDFS? How does it help in maintaining file metadata?	02	1	1
	1.3	Define HDFS Federation. How does it help in scaling a Hadoop cluster?	02	1	1
	1.4	What happens when a mapper fails during a MapReduce job?	02	2	1
	1.5	Differentiate between Schema-on-Read and Schema-on-Write with respect to Hive and traditional RDBMS.	02	2	1
	1.6	What is the role of partitions in Hive? How do they impact query performance?	02	2	2
	1.7	What are interceptors in Apache Flume? Mention one use case.	02	2	1
	1.8	List any two delivery guarantees provided by Flume.	02	2	1
	1.9	What are wide transformations in Spark? How do they influence job stages?	02	2	1
	1.10	State the significance of lineage in Spark RDDs.	02	2	1

PART-B

2	a	Explain the anatomy of a file read operation in HDFS. Include data locality and block access.	06	1	1
	b	Discuss the high availability feature in HDFS. Explain how automatic failover is achieved.	10	1	1
3	a	Write a Java MapReduce program to count the number of lines containing a specific keyword in a text file.	10	1	4
	b	Describe the stages of task execution in a MapReduce job with a labeled diagram.	06	2	1
		OR			
4	a	What is the purpose of the Combiner function in MapReduce? Explain with an example scenario.	06	2	4
	b	Write a Hadoop Streaming example using Python to compute word counts.	10	2	1
5	a	Draw and explain the architecture of Hive. Include the roles of Driver, Compiler, Execution Engine and Metastore.	06	2	4

	b	Create a Hive table for employee attendance data and write HiveQL for: Consider relevant attributes to the table i) Total attendance in a specific month. ii) List employees with perfect attendance. iii) Total late entries in the year 2023. iv) Number of distinct employees from Karnataka. v) Employees who were absent for more than 10 days. vi) Average attendance per department.	10	3	2
		OR			
6	a	How does bucketing work in Hive? Illustrate with a sample use case.	06	2	4
	b	For a dataset containing wildlife census (AnimalID, Species, Region, Count, Survey_Date) create a Hive table and write queries for : i) Total animal counter per region. ii) Most common species. iii) Regions with surveys in the last 3 years. iv) Species found only in one region. v) Average count per species vi) Species count in the "Western Ghats" region	10	3	2
7	a	Explain the configuration of a Spooling Directory Source and File Channel in Flume with an example.	10	2	3
	b	What are multiplexing selectors in Flume? How do they route events to different sinks?	06	2	3
		OR			
8	a	Explain the difference between Replicating and Multiplexing fan-out in Flume. Provide a configuration for each.	10	2	3
	b	Describe how event flows through multi-agent Flume tiers with an appropriate diagram.	06	2	3
9	a	Write a Spark program in Scala to compute the word count for a given file using RDD transformations.	10	1	3
	b	With a diagram explain the lifecycle of a Spark job from job submission to task execution.	06	1	5
		OR			
10	a	Explain how spark executes a job with a neat diagram.	10	2	3
	b	Differentiate between narrow and wide transformations in Spark with examples.	06	2	5