Reg. No.															

Question Paper Code	13586

B.E. / B.Tech. - DEGREE EXAMINATIONS, APRIL / MAY 2025

Sixth Semester

Artificial Intelligence and Data Science

(Common to Computer Science and Engineering (AIML))

20AIPW602 – BIG DATA ANALYTICS WITH LABORATORY

Regulations - 2020

Duration: 3 Hours | Max. Marks: 100

PART - A (MCQ) (10 × 1 = 10 Marks)

Answer ALL Questions

		Marks	K – Level	CO
1.	On which of the following platforms does Hadoop run? (a) Debian (b) Cross Platform (c) Bare metal (d) Unix	1	K1	CO1
2.	Input to the _____ is the sorted output of the mappers. (a) Reducee (b) Mapper (c) Shuffle (d) All of the above	1	K1	CO1
3.	Which one is a usual way people use to work with big data? (a) Lambda (b) Monolithic (c) Centralized (d) Client-server	1	K2	CO2
4.	Which NoSQL database is most commonly used for handling document-based data? (a) Cassandra (b) MongoDB (c) Redis (d) Neo4j	1	K2	CO2
5.	_____ is the architectural center of Hadoop that allows multiple data processing engines. (a) YARN (b) Hive (c) Incubator (d) Chuckwa	1	K1	CO3
6.	Use the _____ command to run a Pig script that can interact with the Grunt shell. (a) Fetch (b) Declare (c) Run (d) Set	1	K2	CO3
7.	Avro schemas are defined with _____ (a) JSON (b) XML (c) JAVA (d) C	1	K1	CO4
8.	What component in the Hadoop ecosystem is used for real-time processing of streaming data? (a) MapReduce (b) Hive (c) Spark (d) Flume	1	K1	CO4
9.	Which tool is commonly used for creating interactive and dynamic data visualizations? (a) Tableau (b) Microsoft (c) power BI (d) D3.js	1	K1	CO5
10.	Although the Hadoop framework is implemented in Java, MapReduce applications need not be written in _____ (a) Java (b) C (c) C# (d) VB	1	K2	CO6

PART - B (12 × 2 = 24 Marks)

Answer ALL Questions

		Marks	K – Level	CO
11.	Compare the use of mean, median and mode in descriptive analytics.	2	K2	CO1
12.	Why is Big Data Analytics important for businesses?	2	K2	CO1
13.	How Does MongoDB Ensure High Availability and Scalability?	2	K2	CO2
14.	Differentiate between the deleteOne() and deleteMany() functions in MongoDB.	2	K2	CO2
15.	Name any two data processing operators used in Pig.	2	K1	CO3
16.	State the role of the Hive Metastore in Hive architecture and why it is important for managing metadata.	2	K2	CO3
17.	Give the command to copy a local file named data.txt to HDFS.	2	K1	CO4
18.	List down the three types of schedulers in Hadoop.	2	K1	CO4
19.	Define a scatter plot and mention one of its typical use cases.	2	K1	CO5
20.	Why a box plot might be a better choice than a bar chart for showing the spread of test scores?	2	K2	CO5

21.		After installing Hive, what command do you type to open the Hive shell?	*2*	*K2*	*CO6*
22.		Identify the prerequisites required for installing Hive on a Hadoop cluster.	*2*	*K2*	*CO6*

PART - C (6 × 11 = 66 Marks)

Answer ALL Questions

23.	a)	Explain the functionality of Hadoop Streaming and evaluate its significance in enabling non-Java developers to use Hadoop.	*11*	*K2*	*CO1*
		OR			
	b)	Relate how IBM's incorporates the Big Data Strategy with respect to its integration of AI. How does IBM differentiate itself from open-source alternatives?	*11*	*K2*	*CO1*
24.	a)	Infer a MongoDB CRUD application for a student grading system. Create the schema, demonstrate inserting multiple records, updating a grade, reading student details, and deleting a student entry.	*11*	*K2*	*CO2*
		OR			
	b)	Show how a MapReduce function is used to remove duplicate records from a large dataset. Explain the logic used in the Mapper and Reducer to identify and eliminate duplicates.	*11*	*K2*	*CO2*
25.	a)	Describe the HiveQL queries for the following, Create a table for employee records, Load data into the table, Retrieve all employees from a specific department and Find the average salary	*11*	*K2*	*CO3*
		OR			
	b)	Illustrate the Pig Latin script to do the following, Load student records from a file, Filter students who have marks greater than 50, Group the filtered students by their department and calculate the average marks for each department.	*11*	*K2*	*CO3*
26.	a)	Describe the architecture of the Hadoop Distributed File System (HDFS). Explain the functions of the NameNode and DataNodes, and summarize how data is stored, replicated and retrieved in HDFS.	*11*	*K2*	*CO4*
		OR			
	b)	What do you mean by AVRO Serialization in Hadoop? Explain it with its relevant advantages & Disadvantages.	*11*	*K2*	*CO4*
27.	a)	Compare and contrast the use of a distribution plot versus a histogram for displaying the distribution of a large dataset. In which cases would one be preferred over the other?	*11*	*K3*	*CO5*
		OR			
	b)	Relate the appropriateness of using a map chart for visualizing data related to population density. What are the limitations and possible improvements to this approach?	*11*	*K3*	*CO5*
28.	a)	Construct a complete solution for implementing matrix multiplication using Hadoop MapReduce. Create the necessary MapReduce job configuration, input and output formats, and sample datasets.	*11*	*K3*	*CO6*
		OR			
	b)	Prepare a scalable Hive architecture for a real-time data processing pipeline. Describe how you would handle big data workloads, partition data, optimize queries, and use the Hive metastore efficiently.	*11*	*K3*	*CO6*