Reg. No.																

Question Paper Code	13244

B.E. / B.Tech. - DEGREE EXAMINATIONS, NOV / DEC 2024

Fifth Semester

Information Technology

20ITPC502 - BIG DATA ESSENTIALS

Regulations - 2020

Duration: 3 Hours | Max. Marks: 100

PART - A (MCQ) (20 × 1 = 20 Marks)

Answer ALL Questions

		Marks	K – Level	CO
1.	Which Big Data technology is designed to handle large volumes of unstructured data? (a) SQL Databases (b) NoSQL Databases (c) File systems (d) Spreadsheets	1	K1	CO1
2.	Which of the following is NOT typically considered a Big Data characteristic? (a) Volume (b) Variety (c) Velocity (d) Viscosity	1	K1	CO1
3.	List a key characteristic of Big Data? (a) Small data sets (b) Structured data only (c) High volume, velocity, and variety (d) Low value	1	K1	CO1
4.	Which component of Hadoop is responsible for resource management and job scheduling? (a) HDFS (b) MapReduce (c)YARN (d)Hive	1	K1	CO2
5.	Which of the following is NOT a key characteristic of Hadoop? (a) Fault-tolerance (b) Real-time data processing (c) Scalability (d) Distributed computing	1	K1	CO2
6.	What is the main purpose of the NameNode in HDFS? (a) Store data in blocks (b) Manage the metadata and track the location of data blocks (c) Execute MapReduce jobs (d) Compress data	1	K1	CO2
7.	What are the two main components that must be implemented in a MapReduce application? (a) Mapper and Combiner (b) Mapper and Reducer (c) Reducer and Input Format (d) Combiner and Output Format	1	K1	CO3
8.	Which component in YARN is responsible for managing the execution of individual applications? (a) Node Manager (b) Job Tracker (c) Resource Manager (d) Application Master	1	K1	CO3
9.	Which of the following is a common input format used in MapReduce? (a) Sequence File Input Format (b) Avro Input Format (c) Text Output Format (d) All the above	1	K1	CO3
10.	What information does the Hive Metastore store? (a) Data processing results (b) Metadata about Hive tables (c) User credentials (d) Job execution logs	1	K1	CO4
11.	What type of system is Apache Hive primarily used for? (a) Real-time data analytics (b) Batch data processing and querying (c) Data streaming (d) Transaction processing	1	K1	CO4
12.	What is the primary purpose of Apache Pig? (a) Real-time data processing (b) Batch data processing (c) Data visualization (d) Data storage	1	K1	CO4
13.	Which Spark component is responsible for handling the execution of tasks on the cluster? (a) Spark Context (b) Driver Program (c)Executor (d) Cluster Manager	1	K1	CO5

14.	In Spark, which operation is used to transform a DataFrame by applying a function to each row? (a) map() (b) reduce() (c) filter() (d) groupBy()	1	K1	CO5
15.	Which of the following programming languages does Spark support natively? (a) Java (b) Python (c) R (d) All of the above	1	K1	CO5
16.	Which method is used to read data from a CSV file into a DataFrame in Spark? (a) read.csv() (b) spark.read.csv() (c) DataFrame.readCSV() (d) csv.read()	1	K1	CO5
17.	Which CUDA API function is used to allocate memory on the GPU? (a) cudaFree() (b) cudaMalloc() (c) cudaMemcpy() (d) cudaHostAlloc()	1	K1	CO6
18.	Which of the following is a key feature of the CUDA API? (a) It supports only single-thread execution. (b) It allows for dynamic parallelism. (c) It is limited to specific types of data. (d) It requires complex programming languages.	1	K1	CO6
19.	Which of the following describes a kernel in CUDA? (a) A small data structure used for storing results. (b) A function that runs on the host. (c) A function that executes on the GPU and is called from the host. (d) A process that manages memory allocation.	1	K1	CO6
20.	Which type of memory is fastest in CUDA? (a) Global memory (b) Shared memory (c) Constant memory (d) Local memory	1	K1	CO6

PART - B (10 × 2 = 20 Marks)

Answer ALL Questions

21.	What is the importance of data storage in Big Data analytics?	2	K2	CO1
22.	How does the velocity in big data affect decision-making?	2	K2	CO1
23.	Mention two key components of Hadoop and their functions.	2	K1	CO2
24.	Outline the purpose of serialization in Hadoop.	2	K1	CO2
25.	How does Hadoop handle failures in a MapReduce job?	2	K2	CO3
26.	What is the purpose of the 'split' operation in a MapReduce job?	2	K2	CO3
27.	How does Apache Pig differ from traditional relational databases?	2	K2	CO4
28.	What is the function of the Hive Metastore?	2	K2	CO4
29.	Define Resilient Distributed Dataset in Spark.	2	K2	CO5
30.	Why is it important to manage memory efficiently in CUDA programming?	2	K2	CO6

PART - C (6 × 10 = 60 Marks)

Answer ALL Questions

31.	a)	Describe the use cases of big data in different industries, such as healthcare, finance, and marketing. How do these industries leverage Big Data for decision-making and gaining competitive advantage?	10	K2	CO1
		OR			
	b)	Discuss the importance of data privacy and security in Big Data Analytics. How can organizations mitigate risks? Justify your answer with appropriate examples.	10	K2	CO1
32.	a)	Given a situation where your HDFS contains thousands of small files that are slowing down the performance of data processing jobs, explain how you would utilize Hadoop Archives (HAR) to manage these files effectively?	10	K3	CO2
		OR			
	b)	Discuss the benefits of using HDFS for this purpose, focusing on its architecture, data replication, and fault tolerance features. Provide examples of how these features would be advantageous in a scientific research environment.	10	K3	CO2

33. a) How the job scheduler in MapReduce framework make an impact on the performance and resource utilization of a Hadoop cluster. Provide examples of scenarios where different schedulers might be preferable. *10 K2 CO3*

OR

b) Explain the role of YARN in the MapReduce framework. Discuss its architecture and how it enhances resource management and job scheduling compared to the traditional MapReduce framework. *10 K2 CO3*

34. a) Consider a scenario where the team needs to create a table for storing user activity logs. Write a HiveQL statement to create a table named user_activity with the following columns: user_id (STRING), activity (STRING), timestamp (TIMESTAMP). Outline the steps to load data into this table from a CSV file located in HDFS. *10 K3 CO4*

OR

b) Analyze the customer feedback data stored in a Hive table. The table contains the following columns: *10 K3 CO4*

i. feedback_id (INT)
ii. customer_id (STRING)
iii. feedback_text (STRING)
iv. rating (INT)

The ratings range from 1 to 5, where 1 is the lowest and 5 is the highest. Write a HiveQL query to retrieve the average rating from the customer feedback table. And write a query to find the count of feedbacks for each rating value.

35. a) As a data engineer, analyze a customer transactions dataset (customer_transactions.csv). The dataset contains the following columns: *10 K3 CO5*

- transaction_id (STRING)
- customer_id (STRING)
- transaction_date (STRING in 'YYYY-MM-DD' format)
- amount (FLOAT)

Perform the following tasks using Apache Spark code,

1. Load the CSV file into a Spark DataFrame and convert the transaction_date column to a DateType.
2. Calculate the total transaction amount for each customer.
3. Find the customer with the highest total transaction amount and display their details.
4. Write the result to a new CSV file named top_customer.csv.

OR

b) As a large dataset of movie ratings stored in a CSV file named movie_ratings.csv. The dataset contains the following columns: *10 K3 CO5*

- user_id (INTEGER)
- movie_id (INTEGER)
- rating (FLOAT)
- timestamp (STRING in 'YYYY-MM-DD HH:MM' format)

Write a R code and perform the following,

1. Load the CSV file into a Spark DataFrame using the sparklyr package.
2. Calculate the average rating for each movie.
3. Identify the top five movies with the highest average ratings and display their movie_id and average rating.
4. Write the result to a new CSV file named top_movies.csv.

36. a) Describe various steps involved in the process of profiling and debugging a CUDA application from start to finish, including performance optimization techniques. *10 K2 CO6*

OR

b) List the advantages and challenges of using CUDA for high-performance computing applications. Explore the implications of CUDA's dependency on NVIDIA hardware. *10 K2 CO6*