BH/K/TX

Reg. No: 21BCE2629

VIT
Vellore Institute of Technology

Final Assessment Test – November 2024

Course: **BCSE402L - Big Data Analytics**

Class NBR(s): **1795/1803/1810**

Slot: **B2+TB2**

Time: **Three Hours**

Max. Marks: **100**

- **KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE**
- **DON'T WRITE ANYTHING ON THE QUESTION PAPER**

Answer ALL Questions

(10 X 10 = 100 Marks)

1. Give three examples of unstructured machine-generated data. Explain the unique characteristics of each type and evaluate how they relate hierarchically to relational data within the context of big data processing.

2. Illustrate the steps involved in basic file system operations in HDFS, such as reading, writing, and deleting files. How do these operations handle the challenges of distributed data storage, and what mechanisms are in place to ensure data reliability?

3. Describe how YARN improves resource utilization compared to the earlier versions of Hadoop (pre-YARN). What are the key features of YARN that allow it to handle multiple types of workloads, such as batch processing and interactive querying, simultaneously?

4. Consider an ice cream manufacturing firm selling six flavors (IC1 to IC6), where hourly sales data is logged across 3,000 machines, and 18 files are generated every 8 hours. Each file contains key-value pairs representing hourly sales for each flavour. Describe how you would design a MapReduce model to process and analyze this large-scale sales data efficiently. Also, provide the steps how Map phase would handle the distribution of sales data from the 3,000 machines and how the Reduce phase would aggregate the results.

5. Explain how the Partitioner in the MapReduce framework ensures even distribution of data across Reducers. Why is this distribution critical for maintaining the efficiency of a MapReduce job with addressing issues such as data skew and workload imbalance.

6. When can we use HBase instead of HDFS in a distributed systems that scale to hundreds or thousands of nodes? How is Persistence and data availability addressed in HBase?

7. Consider the data given below.

ProductCategory	ProductId	ProductName
Toy_Airplane	10725	Lost Temple
Toy_Airplane	31047	Propeller Plane
Toy_Airplane	31049	Twin Spin Helicopter
Toy_Train	31054	Blue Express
Toy_Train	10254	Winter Holiday Toy_Train

i. Create a Schema RDD for a table named product_details with three columns: Product Categories, ProductId, and Product Name.

ii. How does Spark RDD programming gather objects? Describe the methods for creating RDDs using Spark operations.

8. Consider a scenario where a transportation company stores data on vehicle trips, including vehicle ID, start time, end time, distance traveled, and fuel consumed. Using Spark's transformations and actions, explain how you would:

i. Apply transformations to filter trips that exceed a certain distance threshold (e.g., more than 100 km) and map the results to show only the vehicle ID and fuel consumed.

ii. Use actions to compute the total number of long-distance trips and the average fuel consumption for those trips.

Describe the specific transformation and action methods used and explain how they contribute to the efficient processing of the trip data.

9.a) A log management system collects IP addresses from different users accessing an online service. The system records the IP addresses at different time intervals to track usage patterns. The goal is to determine how many unique users (IP addresses) accessed the system within a specific time period.

Dataset-1 (IP addresses over a day):

(192.168.1.1, 192.168.1.2, 192.168.1.3, 192.168.1.1, 192.168.1.4, 192.168.1.2, 192.168.1.5)

Dataset-2 (IP addresses over a week):

(192.168.1.1, 192.168.1.2, 192.168.1.3, 192.168.1.5, 192.168.1.6, 192.168.1.7, 192.168.1.2, 192.168.1.8, 192.168.1.1, 192.168.1.9)

Using the Count Distinct Algorithm, calculate the number of distinct IP addresses for each dataset. Explain how the algorithm works and describe how it efficiently computes the distinct IP addresses in each dataset.

OR

9.b) In the context of Data Stream Management, explain how sampling and filtering techniques can be applied to an industrial monitoring system collecting high-frequency sensor data (e.g., temperature, pressure, vibration) from machines. How do these techniques help optimize real-time data processing and analysis?

10.a) Consider a scenario where a library system categorizes books by genre, author, and publisher. Each book has a unique ID, a title, and belongs to a specific genre. Using this scenario:

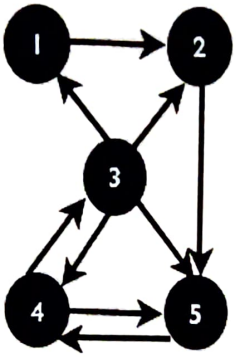
Category	ID	Name
Book	10101	HarryPotter
Book	10102	Lord of the Rings
Book	10103	The Hobbit
Magazine	10201	Nat Geo
Magazine	10202	TIME
Newspaper	10301	The Times
Newspaper	10302	The Guardian

i. Draw a property graph model representing the relationships between the different categories (Books, Magazines, Newspapers), their respective IDs, and names. Each node should represent either a Category (Book, Magazine, Newspaper) or an Item (specific books, magazines, newspapers), and the relationships should show how items belong to categories.

ii. Create triples in the subject-predicate-object format to represent the data in the table.

OR

10.b) A directed graph G has the set of nodes {1, 2, 3, 4, 5} with the edges arranged as follows. Let damping factor is 0.85(d). Compute the PageRank score at 2nd iteration.



⇔⇔⇔ **BH/K/TX** ⇔⇔⇔