VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of the UGC Act, 1956)

Reg. No. : 23MCA1030

Final Assessment Test (FAT) - May 2024

Programme	M.C.A.	Semester	WINTER SEMESTER 2023 - 24
Course Title	BIG DATA ANALYTICS	Course Code	PMCA607L
Faculty Name	Prof. MADHESWARI	Slot	E2+TE2
		Class Nbr	CH2023240501412
Time	3 Hours	Max. Marks	100

General Instructions:

- Write only Register Number in the Question Paper where space is provided (right-side at the top) & do not write any other details.

Answer all questions (10 X 10 Marks = 100 Marks)

01. Consider the scenario where individuals partake in various online activities, such as posting photos on Instagram, sending direct messages on Twitter, streaming music on Spotify, making online purchases on Amazon, and reading articles on a news website. [10]

a) Examine the types of data generated and received during these activities.(5 marks)

b) Provide a comprehensive analysis of their structures, relevance, and potential use cases. (5 marks)

02. a) Design a schema for user management with a table named users. Each user should have a **user_id** generated as a **UUID**, a username, and a set of **contact_ids** representing the user's contacts. Implement the schema along with appropriate data types and constraints. (5 marks) [10]

Sample Queries:

i. Write a query to insert a new user into the users table.

ii. Create a query to retrieve the username of a user given their user_id.

iii. Develop a query to update the contact_ids of a specific user.

iv. Design a query to delete a user record based on their user_id

b) Develop a schema for messaging functionality with a table named messages. Each message should be uniquely identified by a **message_id** generated as a **UUID**. The table should include columns for **sender_id**, **receiver_id**, **timestamp**, and a map of message metadata containing key-value pairs for message attributes. Construct the schema ensuring efficient storage and retrieval of messages.(5 marks)

Sample Queries:

i. Write a query to insert a new message into the messages table.

ii. Create a query to retrieve all messages sent by a particular user.

iii. Develop a query to retrieve messages between two users within a specified timeframe.

iv. Design a query to update the metadata of a specific message.

03. Outline the process of splitting a 500MB input file into blocks within HDFS, elucidating the splitting strategy and file storage with clear diagrams. Subsequently, discuss the ramifications of node failure in HDFS and the mechanisms implemented to address such failures, incorporating explanations of data replication, Namenode, and Datanode functionalities in ensuring fault tolerance. Provide examples illustrating the significance of replication and redundancy in maintaining data reliability and availability within HDFS. [10]

04. Consider a dataset containing transactions from an online retail platform. Each record comprises a customer ID and transaction details, including purchase, return, or browsing activity. For instance: [10]

Record 1: (Customer1, Purchase)
Record 2: (Customer2, Return)
Record 3: (Customer1, Purchase)
Record 4: (Customer3, Purchase)
Record 5: (Customer2, Purchase)
Record 6: (Customer1, Return)
Record 7: (Customer2, Purchase)
Record 8: (Customer1, Purchase)
Record 9: (Customer3, Return)
Record 10: (Customer2, Purchase)

a) Apply the Map function to transform records into key-value pairs, where the key represents the customer ID and the value denotes the transaction type.(5 marks)

b) Explain the shuffling and sorting of these pairs based on keys for the Reduce phase. Perform the Reduce operation to derive insights, computing relevant metrics or statistics from the aggregated results.(5 marks)

05. Delve into the comprehensive phases of the data analytics life cycle, employing a case study focused on predicting housing prices in a bustling metropolitan area. In your response, provide a detailed exploration of each phase, including data acquisition, data preprocessing, exploratory data analysis (EDA), model selection, model training, model evaluation, and deployment. [10]

06. Design a Hadoop MapReduce program to process the provided sample data for the Employee and Salary tables and accomplish the following tasks: [10]

a) Perform an inner join operation between the Employee and Salary tables.(6 marks)

b) Filter records where the employee's department is "Engineering". (2 marks)

c) Filter records where the years of experience are greater than 5 and the salary falls within the range of $50,000 to 100,000$. (2 marks)

Utilize the following sample data:

Employee Table:

Employee ID	Name	Department	Years of Experience
1001	Bob	Engineering	20
1002	Nisha	Medicine	10

Salary Table:

Employee ID	Salary
1001	85000
1002	70000

07. a) Describe the foundational components comprising the architecture of Hadoop 1.0, outlining their roles and interactions within the framework. Highlight essential elements like the Hadoop Distributed File System (HDFS), MapReduce processing model, and Hadoop Common libraries. (5 marks) [10]

b) Subsequently, delineate the step-by-step procedure for submitting and executing jobs in a Hadoop cluster, using a practical scenario of executing a Java program (analyzeData.java) aimed at analyzing 50,000 records within a 30-second timeframe. Detail the actions involved in preparing, packaging, and submitting the job to the cluster, along with Hadoop's mechanisms for distributing and executing the job across cluster nodes. (5 marks)

08. Consider a scenario where you're tasked with designing a sample table named "products" for a Cassandra database. The table schema is defined as follows: [10]

```
CREATE TABLE products (
  product_id UUID PRIMARY KEY,
  name TEXT,
  category TEXT,
  price DECIMAL,
  quantity INT
);
```

a) Design a keyspace and table in Cassandra to facilitate CRUD operations on product data. Define the schema for the table, specifying the primary key and any secondary indexes required for efficient querying. (5 marks)

b) Develop functions within your Cassandra program to implement the CRUD operations. Test your Cassandra program with diverse datasets, encompassing various scenarios and edge cases. Ensure to verify the results of each CRUD operation, validating that data is inserted, retrieved, updated, and deleted accurately. (5 marks)

09. a) Explain the fundamental components comprising big data in depth. Provide detailed insights into each element, elucidating their significance and interplay within the realm of big data analytics.(5marks) [10]

b) Explore the challenges and uses of big data in various industries. Discuss hurdles in managing and analyzing large data volumes, including privacy, security, and scalability concerns. Provide examples from healthcare, finance, retail, and manufacturing to illustrate how organizations overcome these obstacles to extract valuable insights from big data. (5 marks)

10. a) Perform agglomerative hierarchical clustering on the given dataset using the single-linkage method. (5 marks) [10]

b) Display the clustering steps and dendrogram.(5 marks)

Points	X coordinate	Y coordinate
P1	2	3
P2	7	2
P3	3	1
P4	4	5
P5	3	5
P6	9	2
P7	2	7