# Question Paper

Exam Date & Time: 06-May-2024 (02:30 PM - 05:30 PM)

## MANIPAL ACADEMY OF HIGHER EDUCATION

6th SEMESTER B.TECH END SEMESTER EXAMINATIONS, APRIL / MAY 2024

**BIG DATA ANALYTICS [ICT 4034]**

**Marks: 50** **Duration: 180 mins.**

**Answer all the questions.**

Instructions to Candidates: Answer ALL questions Missing data may be suitably assumed

1) A) Given a scenario where a large dataset needs to be stored in HDFS, design a replication strategy that balances fault tolerance with storage efficiency. Justify your choice of replication factor and explain how it aligns with the requirements of the scenario. (5)

B) Company JKL operates a global supply chain network and seeks to enhance efficiency and reduce costs. Explore how Company JKL can utilise various types of big data analytics to achieve these objectives. Discuss how predictive analytics forecast demand and optimise inventory levels, and prescriptive analytics to streamline supply chain processes and enhance supplier collaboration. (3)

C) Given a scenario where a client needs to read data from HDFS, outline the specific steps the client and the HDFS architecture undertake to complete the read operation. Apply your understanding of HDFS read operations to address the requirements of the scenario effectively. (2)

2) A) How can a MapReduce job be implemented to efficiently compute the average number of likes for each publisher using the dataset schema: (Publisher, Title, Author, Genre, Pages, Rating, Likes)? Write the MapReduce code for the same with appropriate comments. (5)

B) Consider a social media platform as an example of a Big Data generator. Justify the types of data they generate and the significance of this data for decision-making. Additionally, discuss the different types of Big Data and provide suitable examples to illustrate each type. (3)

C) Analyze how different data structures such as RDDs, DataFrames and Datasets facilitate distributed data processing in Spark applications. (2)

3) A) Analyze the role of RDD lineage and dependencies in achieving fault tolerance and data recovery in Apache Spark. How does Spark utilize lineage information to reconstruct lost RDDs in case of failures? (5)

B) Consider employee data in a MongoDB database for a company. The database contains a collection named "staff" where each document represents an employee and adheres to the following schema: {"id": Number, "name": String, age": Number, "department": String, "position": String, "salary": Number}
i. Write a MongoDB query to retrieve all employees working in the "Marketing" department who earn a salary greater than $50,000. Display their names, positions and salaries in the results. (3)

C) Discuss the interpretation of query plan generation and optimization stages performed by the Catalyst optimizer. (2)

4) A) What role does the Metastore play in Hive architecture and why is it important? Identify the key differences between Hive tables and traditional relational database tables, highlighting the unique features of Hive's data model. (5)

B) Apply the concept of labelled points in building a logistic regression model for binary classification using Apache Spark. (3)

C) Compare and contrast the communication models used in Apache Spark and traditional MapReduce frameworks such as Apache Hadoop. (2)

5) A) Compare and contrast the performance of a machine learning pipeline consisting of logistic regression with feature scaling using StandardScaler versus logistic regression without any preprocessing. Provide insights into how the choice of preprocessing techniques impacts model performance, interpretability, and computational efficiency. (5)

B) Consider a scenario where a company needs to manage tabular data with nested groups of interrelated information. Justify the selection of a document storage device, such as MongoDB, over a traditional relational database management system (RDBMS). Provide two advantages specific to the scenario. (3)

C) Describe the major components of Spark Streaming and their roles in the streaming data processing pipeline. Illustrate the flow of data through these components. (2)

-----End-----