Question Paper

Exam Date & Time: 22-Jun-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

6TH SEMESTER B.TECH MAKEUP EXAMINATIONS, JUNE 2024

BIG DATA ANALYTICS [ICT 4034]

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

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

1) A) Analyze the role and significance of the EditLog and FsImage components in the architecture of the Hadoop Distributed File System (HDFS). Evaluate how these components contribute to the reliability, fault tolerance, and consistency of file system metadata in HDFS (5)

B) Hospital ABC aims to enhance patient care and operational efficiency through big data analytics. Discuss how Hospital ABC can analyze electronic health records (EHR), medical imaging data, and patient demographics to identify trends, improve diagnosis accuracy, and optimize resource allocation. Illustrate the potential benefits of applying predictive analytics in healthcare settings. (3)

C) Given a scenario where a client needs to write a large dataset to HDFS, describe how pipelined writing can be utilized to optimize data transfer efficiency. (2)

2) A) Discuss the trade-offs between computational efficiency and resource utilization when processing prime numbers using MapReduce on a distributed system. Write a MapReduce program to count the number of prime numbers in a given range. (5)

B) How can organizations integrate multiple types of analytics such as descriptive, predictive and prescriptive to derive comprehensive insights? Provide an example of how integrating different types of analytics can lead to informed decision-making. (3)

C) Illustrate the underlying principles and mechanisms behind the representation of data in Apache Spark. (2)

3) A) Describe the characteristics of RDDs and their significance in Spark programming. How do RDDs enable fault-tolerant and distributed processing? Design a code snippet in Apache Spark to create an RDD from an existing collection in the driver program. Explain the steps involved in the process. (5)

B) An online retail store maintains a MongoDB database named "products" to store information about its products. The "products" collection has documents representing each product, following this schema: (3)

{ "id": ObjectId, "name": String, "category": String, "price": Number, "stock_quantity": Number }

Write a MongoDB query to retrieve all products in the "Electronics" category with a price greater than $500, while displaying their name, category, and price.

C) Analyze the impact of schema enforcement and type safety in DataFrames and Datasets on data quality and application robustness in Apache Spark. (2)

4) Consider a telecommunications company that wants to analyze its customer call data stored in (5)

A) Hadoop using Apache Pig. Design a Pig Latin script to accomplish these tasks efficiently. The call data is stored in a semi-structured format with the following fields: call_id, caller_number, callee_number, call_duration, and call_type. i. Calculate the total number of calls made by each caller. ii. Determine the average call duration for each call type (e.g., local, international). iii. Identify the top 10 most frequently dialled callee numbers. Explain the steps involved in the script using suitable comments.

B) Explore the relationship between transformers and estimators within the context of pipelines, using logistic regression as a case study. (3)

C) Discuss how Spark's resilient distributed datasets (RDDs) and directed acyclic graph (DAG) execution engine enable fault tolerance, data parallelism, and lazy evaluation in Spark applications. (2)

5)

A) Demonstrate how Transformers manipulate dataframes by adding, deleting, or updating existing features. Provide an example of a common Transformer, such as VectorAssembler, and describe its functionality in transforming dataframes. (5)

B) In a scenario where a company seeks to manage semi-structured data with adaptable schemas, propose an appropriate NoSQL metastore solution and justify your selection. (3)

C) How do DStreams facilitate real-time data processing? Describe a scenario where windowed operations are used with DStreams. (2)

-----End-----