Question Paper

Exam Date & Time: 07-Jul-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

VI SEMESTER B.TECH MAKE UP EXAMINATIONS, JUNE/JULY 2023

BIG DATA ANALYTICS [ICT 4034]

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

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

1)		In what ways can big data help businesses in making better decisions? What strategies can be used based on characteristic of big data that impact the decision making?	(5)
	A)		
	B)	Describe the working of Map reduce framework and illustrate its working process with a neat diagram.	(3)
	C)	Highlight the difference between data analysis and data analytics with a suitable example?	(2)
2)		Consider a dataset of employees with following schema (emp_id:int, department: char array, name:chararray, designation:chararray, salary:int); Write a pig script to	(5)
	A)	i. Compute the total number of employees with salary > 50,000 ii. Display the employee ids whose designation = "Project Manager ".	
	B)	Write a hive script to create table and display all the products in another file. Consider the input file as below. Laptop, 45000, Computers Pencils, 2, Stationery Rice, 64.45, Grocery Furniture, 65000, Interiors	(3)
	C)	Identify and describe the three important components responsible for the query execution in Hive?	(2)
3)		What is the characteristic of Resilient Distributed Datasets(RDD) that enables it to be fault tolerant? Write a pySpark code to create a RDD and then convert the created RDD to dataframe.	(5)
	A)		
	B)	Discuss what happens when a DataNode fails during the write process in HDFS?	(3)
	C)	Illustrate the significance of SQL interpreter and optimizer in the spark SQL framework.	(2)
4)		Write a pyspark code to create a ML pipeline which consists of three stages namely tokeniser ,hashingTF and logistic regression. Consider the dataframe is a list of (id, text, label),maxIterations=10, regParam =0.001.	(5)
	A)		
	B)	Compare and contrast different output operations and their effectiveness when performed on	(3)

DStreams data?

C) How do you infer that spark is more beneficial than mapreduce. (2)

5) For the schema structure given in figure Q5A explain how a relational user model can be modelled. Write the mongodb code to perform the possible CRUD operations. (5)

A)

Users

ID	first_name	surname	cell	city	location_x	location_y
1	Paul	Miller	447557505611	London	45.123	47.232

Professions

ID	user_id	profession
10	1	banking
11	1	finance
12	1	trader

Cars

ID	user_id	model	year
20	1	Bentley	1973
21	1	Rolls Royce	1965

Figure Q5A

B) How does the process of data processing work in Spark Streaming? Analyze the importance of window operations in spark streaming. (3)

C) Discuss the NoSQL data stores and their characteristic features. (2)

-----End-----