Nirma University

Institute of Technology

Semester End Examination (IR), December - 2019

B. Tech. in Computer Engineering / Information Technology, Semester-VII

IT7C4 Big Data Analytics

Roll / Exam No.		Supervisor's initial with date	

Time: 3 Hours | Max. Marks : 100

Instructions:
1. Attempt all questions.
2. Figures to right indicate full marks.
3. Draw neat sketches wherever necessary.

SECTION – I

Q-1. Answers the following. **[18]**

A. CO2BL2 How do you differentiate Data science and Big Data analytics terms and its usage in the Industry? (3)

B. CO2BL2 What size of the data can be considered as Big Data? Is size of the data the only attribute of the data that makes it a Big Data? (3)

C. CO1BL2 Why do we use HDFS for applications having large data sets and not when there are a lot of small files? (6)

D. CO4BL3 What can be the real life applications of clustering in the Big data like Banking Data, Mega Store transaction data? What inferences can be made using clustering algorithms? (6)

Q-2. Answers the following. **[18]**

A. CO3BL5 We have data in Mongo DB which is the database of the Electricity Consumption. The sample document is as below (6)

Name: ABC, Cust-id: 1234, Area: Satellite, City: Ahmedabad, Month: January, Year: 2018, Units: 326, Amount: 5325.

Here we may have multiple rows with different months and year for the same customer. Write NoSQL Query in Mongo DB to find the following

1. Average consumption of the given customer year wise.
2. Highest amount of the bill year wise.

B. CO5BL4 State the reason why we can't perform "aggregation" (addition) in mapper? Why do we need the "reducer" for this? (6)

C. CO3BL2 When executing Hive queries in different directories, why is metastore_db created in all places from where Hive is launched? (6)

IT7C4

Q-3. Answers the following. **[14]**

A. CO3BL3 Consider the student data file (st.txt) Data in the following format Name, District, Age, gender: (6)

(i) Write a PIG Script to display Names of all Male Students.

(ii) Write a PIG Script to find the number of students from Vizianagaram district.

(iii) Write a PIG Script to display district wise count of all female students.

OR

A. CO3BL3 N dimensional numerical values are written in a row in the Text file. (6)

a) Write Map- Reduce pseudo code for implementing K Means clustering, clearly specify the key-Value pair.

b) What are the challenges in doing K Means clustering using Map Reduce?

B. CO3BL2 Explain the key features of Apache Spark. What are benefits of Spark over MapReduce? (8)

OR

B. CO3BL2 The NoSQL database cannot be used for the financial applications. Cite the reasons for the same. What could be implications of using No-SQL for financial transactions? (8)

SECTION – II

Q-4. Answers the following. **[16]**

A. CO1BL2 What are the various data sources available in Spark SQL? Why is there a need for broadcast variables when working with Apache Spark? (4)

B. CO1BL2 In HDFS file system in Hadoop frame work distributes the data over different nodes. What are the criteria for the block size of the data? What is the effect of having very small or very large block size? (6)

C. CO5BL3 Explain concept of Map Reduce using an example. Write Map Reduce pseudocode for "Group By" "aggregation" in a database. (6)

Q-5. Answers the following. **[18]**

A. CO1BL3 What is data leakage in the context of data analysis? What problems may arise from it? Which strategies can be applied to avoid it? (6)

B. CO4BL4 You are given a task of predicting whether it will rain tomorrow (Yes) or not (No) based on a sample of 1000 consecutive days up to today. For the given results of a classification algorithm in the confusion matrix below: (6)

		Predicted	
		Yes	No
Actual	Yes	40	200
	No	60	700

1. Compute accuracy, precision and recall with respect to "No" class.
2. Which of these metrics is a poor indicator of the overall performance of your algorithm? Which of these metrics is a good indicator of the overall performance? Give a brief reason for the same.

IT7C4

C. **CO4BL3**	Regression is typically referred to as least-square approach. Justify with example.	**(6)**

Q-6. Answers the following. **[16]**

A. **CO5BL2**	Differentiate between supervised and unsupervised learning approaches with examples of each.	(4)
B. **CO3BL4**	What are different transformations in RDD? How is Narrow transformation different than Wide Transformations?	(6)

OR

B. **CO3BL4**	What do you understand by SchemaRDD in Apache Spark RDD? How is Spark SQL different from HQL and SQL?	(6)
C. **CO5BL2**	The way we have spark shell and PySpark shell is it possible to have such shell in Map Reduce? If not specify the reason. How is MLib library and ML library different in Spark?	(6)

OR

C. **CO5BL2**	You are at city shopping mall. You see few people are browsing the items. Some of them are looking for discounts. Some of them are filling feedback form. Few people are at billing counter. You may consider other things and events happening in this scenario. Think for while on the different types of data generated. Categorize each data source into appropriat category, by considering the Variety and velocity of each source.	(6)