Reg. No.

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL
(A constituent unit of MAHE, Manipal)

VI SEMESTER B.TECH. (INFORMATION TECHNOLOGY/COMPUTER AND COMMUNICATION ENGINEERING)

MAKEUP EXAMINATIONS, JUNE 2019

PROGRAM ELECTIVE II - BIG DATA ANALYTICS [ICT 4005]

REVISED CREDIT SYSTEM

(14/06/2019)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ALL** the questions.
- Missing data, if any, may be suitably assumed.

1A. Given a collection of objects each with n measurable attributes, write an algorithm to cluster the values using K-means algorithm for a chosen value of k. **5**

1B. What are the benefits of doing a pilot program before a full-scale rollout of a new analytical methodology? **3**

1C. With a neat diagram explain the file read operation in HDFS. **2**

2A. Describe the syntax of variants of apply function. Give one example for each. Write a R Script to count the average of even numbers in the vector which has consecutive number from 10 to 19. **5**

2B. Explain how moving average be implemented using window function in In-database analytics. Write the query for the same. **3**

2C. Mention the important terms used in Decision tree implementation. **2**

3A. What are the major criteria's to be considered by data analytics for big data projects. Explain what activities the data analytic team has to undergo to develop an idea of the scope of the data needed and validate that idea with the domain experts on the project. **5**

3B. Consider the confusion matrix given in Table Q.3B. Calculate accuracy, specificity and sensitivity. State the formula for each.

Table Q.3B

	Predicted Class			
Actual Class		Good	Bad	Total
	Good	671	29	300
	Bad	38	262	700
Total		709	291	1000

3

3C. When do you say the given time series is a stationary time series? **2**

4A. Consider the customer dataset with schema (cid, c_name, age, and city). Write hadoop map and reduce functions to compute the number of customers above age 40 and residing in city = Mumbai. **5**

4B. Consider a real world scenario and derive the hypothesis for the same. Differentiate between type 1 and type 2 error and justify your statements with respect to the scenario expressed. **3**

4C. Illustrate the importance of bag of words in text analytics with example. **2**

5A. i. Explain in brief the components in the pig architecture. Write the importance of Parser, optimizer and compiler of pig.

ii. Consider a dataset of employees with following schema (emp_id: int, department: char array, name: char array, designation: char array, salary: int); Write a pig script to

a) Compute the total number of employees with salary > 50,000

b) Calculate the total number of employees in each department. **5**

5B. What are the key outputs from the successful analytics project. Discuss how these outputs are managed by the corresponding key players. **3**

5C. Differentiate between ETL and ELT process. **2**