Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent unit of MAHE, Manipal)

II SEMESTER M.TECH. DEGREE END SEMESTER EXAMINATIONS, APRIL/MAY 2019

SUBJECT: OPEN ELECTIVE-BIG DATA ANALYTICS AND TECHNOLOGIES [ICT 5281]

REVISED CREDIT SYSTEM

(07/05/2019)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. Assume that the collection "EMPL" contains records shown in Figure Q.1A. Write the Mongodb queries to perform the following

1. To find document from the EMPL collection where name begins with R
2. MapReduce Mongodb query for displaying the count of number of persons with salary greater than or equal to 500.
3. To display a list of how many employees are working under each role.

```
{no:1,name:"ST",salary:2000,role:"OB"},
{no:2,name:"MSD",salary:1500,role:"WK"},
{no:3,name:"YS",salary:1000,role:"ALR"},
{no:4,name:"RD",salary:1000,role:"MOB"},
{no:5,name:"RS",salary:500,role:"OB"},
{no:6,name:"BK",salary:500,role:"MOB"},
{no:7,name:"VK",salary:300,role:"BW"},
{no:8,name:"JB",salary:400,role:"BW"},
{no:9,name:"HP",salary:400,role:"ALR"},
{no:10,name:"VS",salary:300,role:"OB"}
```

Figure Q.1A. **5**

1B. What is logistic regression? Write the gradient descent function for Logistic regression. **3**

1C. Given the input file as product.csv containing (year, product, quantity), write a hive script to select products whose quantity is greater than 1000 and year is 2001. **2**

2A. Provide two comparisons between HDFS and NFS. With respect to HDFS, explain the concepts Data Integrity, Staging and Replica selection **5**

2B. Let x, y be the variable of the dataset, x = {2.5, 3, 4, 7, 5, 1.8, 2, 4) y = {1.5, 3, 4.2, 6, 2.3, 3, 5, 2.3}, consider second and sixth points as initial centroids. Apply k-means algorithm to compute centroids after two iterations. **3**

2C. What are the criteria to tell machine is learning? Explain types of machine learning algorithms. **2**

3A. Consider the load data having fields (id, borrowerName, purpose, loanAmount, installmentsPaid) stored in loan.csv file. Write Map Reduce functions to display average loan amount for the purpose where more than 10 installments are paid. **5**

3B. Compare Complex event processing and Stream Processing. With respect to SPL explain Functor, Sink and Punctor operators. **3**

3C. In the bank marketing example, the training set includes 2,000 instances. An additional 100 instances are included as the testing set. Classifier is modelled to predict whether they would subscribe to the term deposit given. Construct the confusion matrix if 11 clients who subscribed to the term deposit, the model predicted 3 subscribed and 8 not subscribed. Similarly, of the 89 clients who did not subscribe to the term, the model predicted 2 subscribed and 87 not subscribed. Compute Precision and Specificity. **2**

4A. Given the data for stream analytics as shown in Table Q.4A. Write SPL script to compute highest and lowest marks in Maths. Draw the graph representation with different operators used in the script.

Table Q.4A.

R.no.	Maths	Physics	Chemistry	English
1	55	45	56	87
2	75	55	46	64
3	25	54	89	76
3	78	55	86	63
5	58	96	78	46

5

4B. Explain the components of Map Reduce program along with execution steps on Hadoop cluster. **3**

4C. Explain components of Pig Architecture with neat diagram. **2**

5A. Briefly explain the steps used for performing text analytics using BigInsight. Write an AQL extractor for fetching all the amounts appearing in a document such as $160.99 (billion/million). **5**

5B. Illustrate important properties of Big data? Justify "Hadoop is suitable for Big data Analytics". **3**

5C. Construct the root node of the decision tree for the data given in Table Q.5C. **2**

Table Q.5C.

	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rainy	3/9	2/5