| Reg. No. | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|

# Manipal Institute of Technology, Manipal

(A Constituent Institute of Manipal University)

## VII SEMESTER B.TECH (COMPUTER SCIENCE AND ENGINEERING)

## END SEMESTER EXAMINATIONS, NOV/DEC 2015

SUBJECT: BIG DATA ANALYTCS [CSE 449]

**REVISED CREDIT SYSTEM**

Time: 3 Hours DATE: 08-12-2015 MAX. MARKS: 50

**Instructions to Candidates:**

- Answer **ANY FIVE FULL** questions.
- Missing data, if any, may be suitably assumed.

1A. What is Big Data Analytics? Discuss a case study on Big Data Analytics. 4M

1B. Give the importance of P(>|t|) **a**nd F-Statistic in the analysis of linear regression model. Implement RHadoop script for multilinear regression model using a 2000 X 10 matrix of random numbers with normal distribution. 4M

1C. What is recommendation algorithm? What are the different types of recommendations? List the basic steps to generate recommendation. 2M

2A. What is CAP theorem? Illustrate with an example. How different NOSQL Databases follow CAP theorem? 3M

2B. Describe the following in the context of Cassandra:

i. Writes ii. Hinted Handoffs 3M

2C. Create a column family to store basic information of Students such as Roll No, Student Name, Student Date of Birth and Student Address with two to three student records.

i. Alter the table to include the subject preferences and hobbies of each student. There should be a minimum of two subject preferences and a maximum of four. The order of preferences as given by the student should be preserved. The hobbies as given by the student should be arranged in alphabetical order.

ii. Give a query to replace the first subject preference by 'Big Data and Analytics' for Roll No=1. 4M

3A. Write an R script for the following:

i. Generate vector empid of 1 to 10.
ii. Generate vector empnamesof 10 names
iii. Generatae a vector salary of 10 numeric values
iv. Create a dataframe employee from the above vectors
v. Display total , max , min and average salary
vi. Display employee names who draws max. salary
vii. Select and list the employees with salary greater than average salary 3M

3B. Give a R script to visualize the Iris(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, Species) data set as follows in four parts of single graphics panel.

i. Histogram of Sepal.Length with appropriate labels for X and Y-axis
ii. Boxplot of Sepal.Length for all the three subsets of Iris data set partitioned upon class, Species, label(setosa, versicolor, virginicia).
iii. Boxplot of all the four attributes excluding class(Species) label
iv. Scatter plot of Sepal.Length vs Petal.Width 3M

3C. With necessary diagrams, explain the architecture of following R and Hadoop techniques:

i. Hadoop streaming in R and ii. RHIPE 4M

4A. Explain any four advantages of bucketing in Hive 2M

4B. Explain any three situations where PIG can be used. 3M

4C. With a neat diagram explain the anatomy of a MapReduce job run in YARN. 5M

5A. Consider two text files (first.txt and second.txt) with lots of words separated by spaces. Write a PIG script to display the words which have the same number of occurrence in both the files 5M

5B. With a neat diagram explain how client accesses data from HDFS. 3M

5C. With a neat diagram explain architecture of Hive. 2M

6A. Explain any six characteristics that are important when selecting a data serialization format 3M

6B. Explain with pseudo code how iterative message passing can be performed using MapReduce. 4M

6C. Consider a text file which contains one JSON object per line. Each JSON object contains student's name and his score in one subject.
Write the java code for Driver, Mapper and Reducer(Use only one mapper and reducer) to find the total marks of each student. 3M

*********