Reg. No.									

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
A Constituent Institution of Manipal University

VII SEMESTER B. TECH. (COMPUTER SCIENCE & ENGINEERING)

END SEMESTER EXAMINATIONS, NOV/DEC 2016

SUBJECT: BIG DATA ANALYTICS [CSE 449]

REVISED CREDIT SYSTEM
(06/12/2016)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ANY FIVE FULL** questions.
- Missing data may be suitable assumed.

1A. What is Big Data? What are its characteristics? What is the sources of Big Data **3M**

1B. Why NOSQL? What are its advantages and disadvantages? What is the role of CAP theorem in NOSQL? **3M**

1C. Give CQL operations in Cassandra for the following:

i. Import data from 'D:/Cassandra/Students.csv' with the fields: RegNo, Branch,and CGPA to Student table.
ii. List first five student's information with each column value displayed in separate line
iii. Alter the Student table schema to add the following columns:
  a. Sname of type name (first_name text, last_name text);
  b. SAddress of type address(street text, city text, state text);
  c. Languages of type list
  d. Todo of type map(int, text)
iv. Add 'Hindi' to Languages of student with RegNo=1000 such that it becomes the first element in the list
v. Update the languages of student with RegNo=1005 to change the 3rd element in the languages list to 'Tamil'
vi. List the students who knows 'Bengali' language
vii. List the students who has 'skating' in value part of todo.
viii. List the students who has '5' in key part of todo. **4M**

2A. Illustrate the usage of mapreduce and cursor constructs in MongoDB with necessary examples **3M**

2B. Write an R script for the following:

i. Generate vectors, empid of 101 to 110, ename of 10 names, salary of 10 numeric values
ii. Create a dataframe employee from the above vectors
iii. Use sapply() to increase the employee salary: 10% for less than 5000: 15% for greater than 5000.
iv. Display a) the change in Average salary after the revision, b) the no. of employees whose salary is greater than new average salary and, c) the change in no. of employees, whose salary is greater than average salary, before and after the revision. **3M**

2C. Give an R script to develop a tree based model for predicting whether the customer will take pep(class label) using the customer profile data given in bank-data(age, sex, region, income, married, children , car, save_act, current_act, mortgage, pep). **4M**

Use 80% of data to develop the model and validate the model using the remaining 20% of data. Display the tree model and find accuracy.

3A. Consider iris(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, Species) dataset with class label, Species: Setosa, Versicolor or Virginica. Write R script for the following

Graphical analysis in multiple panes of a window:

i. Histogram of all the attributes excluding class label

ii. Partition the data set based upon class label and draw Boxplot of Petal. Length for each partition of the data set.

iii. Scatter plot for every pair of attributes **3M**

3B. Why to interface R and Hadoop at cluster level? What are the available options? Explain the architecture of each option in detail. **3M**

3C. Give RHadoop mapreduce script for the following:

i. Given m and n, generate k(=n-m+1) random number sets of size m to n and find maximum and minimum of each set.

ii. Compute the frequency of stock market change for the company ‘XYZ.BO’ using stock market data(date, open, high, low, close, Volume, Adj.Close) available in yahoo.finance.com **4M**

4A. With a neat diagram, explain the anatomy of a YARN application run. **4M**

4B. Write an application in MapReduce to process the input dataset to find the name of highest salaried employee in each gender in different departments. Assume that there are only four departments.

Id	Name	Dept	Gender	Salary
1201	gopal	CSE	Male	50,000
1202	manisha	CSE	Female	45,000
1203	khalil	IT	Male	40,000
1204	prasanth	CSE	Male	30,000

Write the java code for Driver, Mapper, Reducer and Partitioner (Use only one mapper and one reducer). **4M**

4C. Write an algorithm to calculate PageRank using MapReduce. **2M**

5A. Explain any six characteristics that are important when selecting a data serialization format. **3M**

5B. Explain six advantages of Hadoop. **3M**

5C. Consider a csv file **'student.csv'** located inside HDFS at **'/user/cloudera/pig/'.** The file contains name and city as shown in Fig.Q.5C. Write a pig script with UDF to display the length of city. (eg. If city is "MANIPAL" then output should be "7").

```
John,[city#BANGALORE]
Jack,[city#PUNE]
James,[city#MANIPAL]
```

Fig.Q.5C **4M**

6A. With a neat diagram explain architecture of Hive. Explain four Hive data units. **5M**

6B. With a neat diagram explain the anatomy of File read and File write in HDFS. **3M**

6C. What is the use of secondary name node in Hadoop. **2M**