B.E./B.TECH. Degree Examination, December 2020

Seventh Semester

IT16703 – Big Data Analytics

(Regulation 2016)

Time: Three hours Maximum : 80 Marks

Answer **ALL** questions

PART A - (8 X 2 = 16 marks)

1. According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies
   a) Big data management and data mining
   b) Data warehousing and business intelligence
   c) Management of Hadoop clusters
   d) Collecting and storing unstructured data
2. Hadoop is a framework that works with a variety of tools that includes
   a) Map-Reduce, Hive, HBase
   b) Map-Reduce, SQL, GoogleApps
   c) Map-Reduce, Heron, Trumpet
   d) Map-Reduce, Hummer, Iguana
3. Pig Latin statements are generally organized in one of the following ways :
   a) A LOAD statement to read data from the file system
   b) A series of "transformation" statements to process the data
   c) A DUMP statement to view results or a STORE statement to save the results
   d) All of the mentioned
4. We have a linear regression equation (Y = 5X +40) for the below table.

X	Y
5	45
6	76
7	78
8	87
9	79

Which of the following is a MAE (Mean Absolute Error) for this linear model?

   a) 8.4
   b) 10.29
   c) 42.5
   d) None of the above

5. List the Limitations of Pig?
6. On which conditions data is called by Big Data?
7. Why is it important for statistics to be one of the key disciplines for Big Data?
8. Differentiate between data streaming and traditional data processing

PART B - (4 X16 = 64 marks)

09. (a) (i) Why Mapreduce is called a Big Data technology and show how it supports Big Data? **(10)**

(ii) Illustrate on how cloud and Big Data related to each other. **(6)**

(OR)

(b) Identify the different statistics concepts required for Big Data and how would you compose the statistical concepts in inference? **(16)**

10. (a) Why RTAP is important in timely decision making? Explain it with a real time example. **(16)**

(OR)

(b) Suppose our stream consists of the integers 4,6,7,1,2,3,7,6,9. The hash function will all be of the form h(x)=ax+b mod 32 where a=2, b=3. Treat the result as a 5 bit integer. Determine the tail length for each stream element and the resulting, estimate the number of distinct elements. **(16)**

11. (a) Consider the dataset with the set of items as{Strawberry, Litchi, Apple, Oranges, Banana} assuming 30% as minimum support and 80% as minimum confidence. **(16)**

Trans ID	**Items Purchased**
101	Litchi, Apple, Strawberry, Banana
102	Strawberry, Apple, Oranges, Litchi, Banana
103	Oranges, Apple, Banana, Litchi
104	Banana, Apple, Strawberry
105	Strawberry
106	Strawberry ,Banana
107	Apple , Strawberry, Litchi
108	Banana , Oranges

Find all frequent item sets and association rule. Write a R code to perform association

(OR)

(b) Why Logistic regression is better than Linear Regression, differentiate Linear and Logistic Regression and also write R code for Linear and Logistic Regression. **(16)**

12. (a) With Neat sketch explain in detail Hadoop architecture and its components? Analyze how Hadoop is handling job execution and failures? **(16)**

(OR)

(b) With Neat sketch illustrate Hadoop Ecosystem. Differentiate Pig & Hive with their applications. **(16)**