Nirma University

Institute of Technology

Semester End Examination (IR), December - 2018

B. Tech. in Computer Engineering / Information Technology, Semester-VII

IT7C4 Big Data Analytics

Roll / Exam No.		Supervisor's initial with date	

Time: 3 Hours | Max. Marks : 100

Instructions:
1. Attempt all questions of Section I and II separately in same Answerbook.
2. Figures to right indicate full marks.
3. Draw neat sketches wherever necessary.

SECTION – I

Q-1. **[16]**

a. What is Big Data? What are the characteristics of Big Data? (4)

b. How do you differentiate Data science and Big Data analytics terms and its usage in the Industry? (4)

c. What are the major changes in the computing paradigm required to meet the challenges put in by Big Data? (4)

d. What are the greatest challenges that prevent business from capitalizing on big data? (4)

Q-2. **[18]**

a. What is Z score and importance of the score in statistical analysis. Let (3, 5, 7, 11) be a sample data, calculate the correct standardization using z-score for all the values. (6)

b. Zip fir Tire Company stocks three brands of tire: Brand A; Brand B and Brand C. 40% are Brand A, 35% are Brand B and 25% are Brand C. The percentage of defective tires are 2% of Brand A, 1% of Brand B and 3% of Brand C. If a tire is picked at random what is the probability that it is defective? (4)

OR

b. For the data given below draw the box plot. Also define how to identify the outliers in the data, data :{45,44,36,23,38,10,50,3,37,35,36,39}

c. For each part a and b, compare distributions (1) and (2) based on their means and standard deviations. You do not need to calculate these statistics; simply state how the means and the standard deviations compare. Make sure to explain your reasoning. For comparison sketch dot plots of the distributions. (4)

(a) (1) 3, 5, 5, 5, 8, 11, 11, 11, 13
(2) 3, 5, 5, 5, 8, 11, 11, 11, 20

b) (1) -20, 0, 0, 0, 15, 25, 30, 30
(2) -40, 0, 0, 0, 15, 25, 30, 30

IT 7C4

d. In an ANN a 4-input neuron has weights 1, 2, 3 and 4. The activation function f(x) = 2x. The inputs are 4, 10, 5 and 20 respectively. What will be the output of the neuron? (4)

Q-3. **[16]**

a. We have data in Mongo DB which is the database of the Electricity Consumption. The sample document is as below (4)

Name: ABC, Cust-id: 1234, Area: Satellite, City: Ahmedabad, Month: January, Year: 2018, Units: 326, Amount: 5325.

Here we may have multiple rows with different months and year for the same customer. Write NoSQL Query in Mongo DB to find the following

1. Average consumption of the given customer year wise.
2. Highest amount of the bill year wise.

b. Take an example to demonstrate the application of Spark SQL. Also compare the efforts in absence of Spark SQL. (4)

c. What is CAP theorem? Give examples from the available solutions and place them in the appropriate vertical . (4)

OR

c. What are graph databases and what types of applications need graph databases?

d. What are ACID and BASE property in terms of Data Bases? (4)

OR

d. What is specific need of Big Data that cannot be accommodated by RDBMS? What is the solution?

SECTION – II

Q-4. **[16]**

a. What are the components of Spark Core? What are the advantages of running an application on spark over map reduce? (4)

b. The way we have spark shell and PySpark shell is it possible to have such shell in Map Reduce? If not specify the reason. (4)

c. How is MLib library and ML library different in Spark? (4)

d. What are the limitation of RDD in spark? (4)

OR

d. In a twitter sentiment analysis application what all component of the spark frame work will be used? Make a block diagram to represent the same.

Q-5. **[18]**

a. In HDFS file system in Hadoop frame work distributes the data over different nodes. What are the criteria for the block size of the data? What is the effect of having very small or very large block size? (5)

b. In a distributed environment over commodity software reliability is a big question. How is reliability and availability ensured in Hadoop environment. (5)

c. The following description the matrix is defined as below:- (8)
M is a matrix with element m in row i and column j.

N is a matrix with element n in row j and column k.
P is a matrix = MN with element p in row i and column k, where P=M*N.
The Matrix being sparse in nature is saved as relational Representation as below.
M with tuples (i, j, m)
N with tuples (j, k, n)

(i) Write the Map-reduce psudo code for doing matrix multiplication. Clearly specify the key-Value pair.
(ii) Can this be done without Map Reduce? Explain your answer

OR

c. N dimensional numerical values are written in a row in the Text file.

(i) Write Map- Reduce pseudo code for implementing K Means clustering, clearly specify the key-Value pair.
(ii) What are the challenges in doing K Means clustering using Map Reduce?

Q-6. **[16]**

a. Considering an application of song repository where the users can create play list and play the song over a large data of songs. Identify the storage solution for this data and explain it pointwise. What can be done specifically for the data modelling to enhance the performance? (6)

b. Consider banking System which intend to use data from various sources like Historical data, Transaction data, customer feedback data, IOT measuring the footfall at various counters. Categorize each data source into appropriat category, by considering the Variety and velocity of each source. (6)

c. Match the following (4)

Column-1(NoSQL type)	Column – 2(Example)
Key-Value Store	Neo4J
Graph database	CouchDB
Document oriented database	Cassandra
Column oriented database	Dynamo