Sardar Patel Institute of Technology

Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India
(Autonomous College Affiliated to University of Mumbai)

SET - I

April-May 2018

Max. Marks : 100
Class : BE Computer
Course Code : CPE 8035
Name of the Course: Big Data Analytics

Duration: 180 Minutes
Semester: VIII
Branch : Computer Engineering

Instructions:

(1) All Questions are Compulsory

(2) Draw neat diagrams

Question No.		Max. Marks	CO
Q1 A)	Explain 4 ways by which big data problems are handled by NoSQL A client needs a database design for his blog. Website has following requirements. a)Every post can have one or more tag. b)Every post has unique title, description and url c)Every post has name of its publisher and total no. of likes. d)On each post, there can be 0 or more comments. Design the normalized schema(tables and columns) for RDBMS and the MongoDB schema.	10	CO3
Q1 B)	How does clickstream analytics using Hadoop help sales and marketing team to prepare their campaigning strategies?	5	CO1
Q2)	How does Pagerank differ from HITS algorithm? Give 2 differences. Calculate the authority and hub scores for the given matrix that represent a graph of four vertices (n1 to n4) , using HITS algorithm with k = 2. Identify the best authority and hub nodes. $\begin{vmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{vmatrix}$ **OR**	10	CO1

	What are the categories the spammer divides the web pages into. Explain? Mathematically prove the improvement in pagerank of a target page using spam farm.		
Q3 A)	Write a Map Reduce program to find out Designation Wise Customer count like How many Doctors are there in the file or How many Lawyers are in the file. Consider following scenario for taking input data : Let there be a file : custs where col1 - cust_id col2 - first_name col3 - last_name col4 - age col5 - designation **OR** Write a Map Reduce program to find the total number of words from the file as given below : Let the input data is: **Apache Spark is a fast, in-memory data processing engine with elegant and expressive development APIs to allow.** final output will be = total words, 17	10	CO2
Q3 B)	How to compute RDBMS Selection Operation by Map Reduce. Clearly explain Map and Reduce functions and Give its pseudo code.	10	CO2
Q4)	Consider the following two texts, Find the similarity between 1 and 2. 1. Julie loves me more than Linda loves me 2. Jane likes me more than Julie loves me	5	CO2
Q5 A)	What is Data Stream Management System? Explain with block diagram.	10	CO4
Q5 B)	Explain with diagram how the PCY algorithm helps to perform frequent itemset mining for large datasets **OR** Explain the concept of CURE Clustering Algorithm with example.	10	CO4
Q6A)	Define Count Distinct Problem with constraints. Suppose data stream consists of the integers 3, 1, 4, 1, 5, 9, 2, 6, 5. Let the hash function being used is h(x) = 3x +1 mod 5;	10	CO4

	Show how Flajolet-Martin algorithm will estimate the number of distinct elements in this stream. **OR** Define following terms with respect to Association Rule Mining - A) Support of Item B) Confidence of Rule Consider a small database (table 1 below) with four items I = {Bread, Butter, Eggs, Milk} and four transactions as shown in following table. Suppose, that the minimum support and minimum confidence of an association rule are 40% and 60% respectively. Find out several potential association rules.		
Q6B)	What is a community in "Social Network Graph"? What are the standard Clustering Techniques for Graph Clustering? Explain one algorithm for finding communities in a Social Graph.	10	CO5
Q7)	By computing Pearson Coefficient , determine the rating of User-ID-3 for Item-ID-1 and Item-ID-6 using table 2 below.	10	CO5

Transaction ID	Items
T1	{Bread, Butter, Eggs}
T2	{Butter, Eggs, Milk}
T3	{Butter}
T4	{Bread, Butter}

Table 1

Item ID →	1	2	3	4	5	6
User ID↓						
1	7	6	7	4	5	6
2	6	7	?	4	3	4
3	?	3	3	1	1	?
4	1	2	2	3	3	4
5	1	?	1	2	3	3

Table 2