BE/5/IT/Rev/VIII/CBGS/Bigdata analytics - / 09-12-2016

QP Code :725300



(3 Hours) [Total Marks 80 ]

i. Q. 1. is Compulsory.
ii. Attempt any three from the remaining.
iii. Assume suitable data.

Q. 1 (a) What are the three Vs of Big Data? Give two examples of big data case studies. Indicate which Vs are satisfied by these case studies. (5)

(b) What is the role of a "combiner " in the Map reduce framework? Explain with the help of one example. (5)

(c) Through an example illustrate how the triangular array can be used to optimally store and count pairs in a frequent itemset mining algorithm. (5)

(d) List the different issues and challenges in data stream query processing. (5)

Q. 2 (a) What are the different data architecture patterns in NOSQL? Explain "key value" store and "Document" store patterns with relevant examples. (10)

(b) Show Map Reduce implementation for the following two tasks using pseudocode. (10)

i. Multiplication of two matrices
ii. Computing Group-by and aggregation of a relational table.

Q. 3 (a) Give a formal definition of the Nearest Neighbor problem. Show how finding plagiarism in documents is Nearest Neighbor problem. What similarity measures can be used. (10)

(b) Clearly explain the concept of a Bloom Filter with the help of an example. (10)

Q. 4 (a) Suppose a data stream consists of the integers 3, 1, 4, 1, 5, 9, 2, 6, 5. Let the hash function being used is $h(x) = 3x + 1 \bmod 5$; Show how the Flajolet-Martin Algorithm will estimate the number of distinct element in this stream. (10)

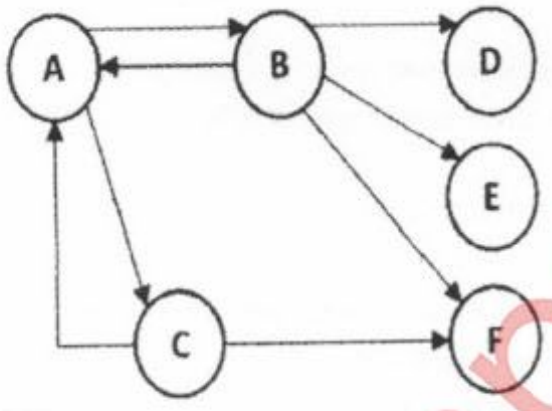
(b) Clearly explain how the CURE algorithm can be used to cluster big data sets. (10)



[Turn Over

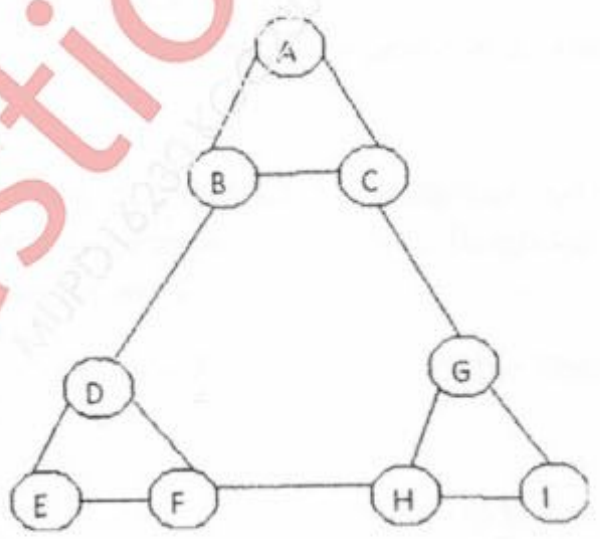
**Q. 5** **(a)** Define Collaborative filtering. Using an example of an e-commerce site like Flipkart or Amazon describe how it can be used to provide recommendations to users. **(10)**

**(b)** Define PageRank. Using the web graph shown below compute the PageRank at every node at the end of the second iteration. Use teleport factor = 0.8. **(10)**




**Q. 6** **(a)** Explain clearly with diagrams how the PCY algorithm helps to perform frequent itemset mining for large datasets. **(10)**

**(b)** For the graph given below use betweenness factor and find all communities **(10)**




---