Ex/CSE/T/424C/2018

B. E. COMPUTER SCIENCE & ENGINEERING EXAMINATION, 2018

(Fourth Year, Second Semester)

BIG DATA ANALYTICS

Time : Three Hours | Full Marks : 100

Answer *question no. 1* and any *four* from the rest
Special credit will be given to brief and to-the-point answers

1. (i) What are the characteristics of Big Data? Explain.

4

(ii) Explain the terms Confidence and Interest in the context of an Association Rule.

3

(iii) Explain how Mahalanobis Distance is defined. What are its assumptions?

4

(iv) Show with a diagram, how files are stored in the Hadoop Distributed File System.

4

(v) What do you mean by Analytics? Explain how Analytics can be useful in Advertising Industry.

2+3

2. What is a Data Stream? What are the challenges of Big Data Stream Processing?

What is a Bloom Filter? Explain its mechanism. Give some example applications of Bloom Filter.

Explain the mechanisms of Apache STORM for Data Stream Processing.

Show how the following problem can be solved by properly configuring the different components of STORM:

Three data streams are arriving from three temperature sensors from a room every minute. The system will output the average of the data from three sensors every minute, as well as the maximum and minimum reading of any sensor so far obtained, also every minute.

2 + 3 + 5 + 2 + 4 + 4

3. Explain in detail how the Map-Reduce Programming paradigm works.

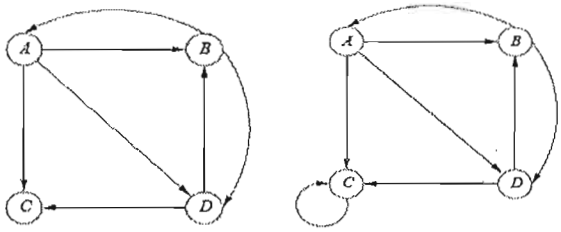
You are given a very large and sparse integer matrix and a large integer vector. Explain how you can multiply the matrix with the vector in the Map-Reduce paradigm.

What changes will you make in your scheme if the vector is so large that it does not fit into the main memory of the nodes in the Hadoop Cluster?

Describe some applications of this solution.

8 + 6 + 3 + 3

4. Find the transition matrices of the following two web graphs.



What are the problems in these two web graphs? Explain with the help of Markov process.

What is Link Spamming? How can you avoid it?

What do you mean by Hubs and Authorities? What are their uses?

Explain the recursive nature of the properties of Hubbiness and Authority of a web page. Explain how they are calculated.

8 + 2 + 2+ 2 + 2 + 4

5. How do you define a Frequent Item Set? Why is it required to find the Frequent Item Sets? Explain with the help of some application.

State and explain the Monotonicity property of Frequent Item Sets.

Explain in detail, how this property is used in the A-priori algorithm for finding out the frequent item sets. What are the challenges of its implementation in the Map-Reduce paradigm?

2 + 3 + 4 + 6 + 5

6. What do you mean by Euclidean Space and Non-Euclidean Space?

Differentiate between Hierarchical and Point Assignment methods of clustering.

Give the rationale of K-means Clustering.

Show how Map-Reduce paradigm can be used to implement K-means Clustering for a massive set of data points.

3 + 4 + 5 + 8

7. One million text files are to be analyzed to come up with a manageable set of candidate files for detailed matching. Explain the broad steps to get an acceptable solution.

Show how you can generate signatures of files such that for the similar files, equal signatures will be generated with high probability. Prove that "the probability that two signatures of two files are equal" is equal to "the similarity of the two files".

Is there any possibility of using the similarity detection mechanism for finding out similar images from a huge collection of image files? Explain.

$8 + 6 + 4 + 2$

8. Answer any four from the following: $4 \times 5 = 20$

(i) Briefly explain how Citation Analysis can be modeled as a Graph Mining Problem. What are the major issues in Large Graph Analysis?

(ii) Explain what you mean by Content Based Recommendation and Collaborative Filtering with some examples.

(iii) Briefly explain the Attribute-Value-Frequency (AVF) Algorithm for finding out the outliers and its Map-Reduce implementation.

(iv) Define the edit distance $d(x, y)$ between two strings x and y; show that the function d defined by you conforms to properties of distance functions.

(v) Considering the present technology trends and data generation trends, in your opinion, which are the directions of growth of Big Data Analytics?

----------X----------