Ex/CSE/PE/H/T/415B/2024

B. E. COMPUTER SCIENCE & ENGINEERING EXAMINATION, 2023

(Fourth Year, First Semester)

BIG DATA ANALYTICS (Hons.)

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. 3

(ii) Explain Bonferroni's Principle with an example. 5

(iii) Give three examples of applications of Outlier Detection. 3

(iv) Explain briefly how Fault Tolerance is achieved in Map-Reduce Programming Paradigm. 4

(v) What do you mean by Analytics? What are the types of Analytics? 2+2

2. What is meant by an Outlier? What are the challenges in the Outlier detection in Large Data Sets?

Explain the AVF algorithm for Outlier detection.

How can you implement the AVF algorithm in the Map-Reduce framework?

What are the sources of speedup in the M-R implementation of AVF algorithm? 2+2+4+9+3

3. Explain the mechanism of Map-Reduce Programming Framework.

Show in detail, how you will find the Natural Join of two relations R(A,B) and S(B,C) using M-R technique.

What are the factors affecting the efficiency of M-R algorithms?

What are the kinds of problems, that can be solved efficiently using the M-R paradigm? 6+8+3+3

[Turn over

4. What do you mean by Euclidean 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

5. Explain the architecture of Hadoop Distributed File System. What is the critical component in the HDFS for achieving speedup? Why?

Detail out how Replication Management is done in HDFS.

Show with figures, how the data reads and writes are executed in the File System.

What has been done to avoid single point failure of the File System?

4+3+3+6+4

6. An online store recommends movies when a user searches for books. What could be the flaw in the implementation of the recommendation system of the store?

Explain clearly the difference between Content-based recommendation system and Collaborative Filtering.

Explain in detail, how a system for recommending websites to web-surfers can be designed.

4 + 6 + 10

7. There are 200,000 webpages with an average size of 10 KB each. After 5-shingling, it was found that the total number unique shingles are 20,000. Explain the procedure of forming the Input Matrix using shingling. Calculate the percent reduction of the size of data to be handled by the above case of shingling for similarity detection.

Explain how the Input Matrix may be further compressed by using Minhashing without compromising the similarity properties of the files.

If there are 60 hash functions used in the Minhash algorithm, what is the percent reduction of the data to store the Sketches of the file?

6 + 4 + 7 + 3

8. Answer the following: 5 X 4 = 20

(i) How will you find Inverted Index of a large number of documents using Map-Reduce paradigm?

(ii) Explain what you mean by Hubs and Authorities in a collection of web pages.

(iii) Define Mahalanobis Distance. What are the assumptions for this distance function? Prove that this is indeed a distance function.

(iv) "Map-Reduce Programming Model and HDFS facilitate extremely high-throughput Batch processing of Big Data" – Explain in your own view

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