Reg. No.																

Question Paper Code	14207

M.E. / M.Tech. - DEGREE EXAMINATIONS, NOV / DEC 2025

First Semester

M.E. - Big Data Analytics

24PBDPC102 - BIG DATA MINING AND ANALYTICS

Regulations - 2024

Duration: 3 Hours Max. Marks: 100

PART - A (MCQ) (10 × 1 = 10 Marks)

Answer ALL Questions

Marks K – Level CO

1. A Distributed File System (DFS) is mainly designed to___________ *1 K1 CO1*
(a) Store data on a single local machine
(b) Manage large data across multiple connected nodes
(c) Compress and encrypt local files
(d) Remove duplicate datasets

2. What is the main purpose of the MapReduce framework? *1 K1 CO1*
(a) To store web documents (b) To process data in parallel over distributed systems
(c) To reduce redundancy in files (d) To analyse small data only

3. When using k-shingles on documents, increasing the value of k will generally: *1 K1 CO2*
(a) Increase sensitivity to small changes (b) Decrease sensitivity to small changes
(c) Produce random similarity (d) Have no effect

4. If cosine similarity between two vectors is 0.95, this means: *1 K2 CO2*
(a) They are highly similar (b) They are completely dissimilar
(c) They have orthogonal directions (d) They are identical copies

5. Why can't we store all data from a data stream? *1 K1 CO3*
(a) Because data streams are too small
(b) Because data streams are unbounded and arrive continuously
(c) Because storage devices are outdated
(d) Because data streams contain no useful information

6. Which approach helps maintain accuracy in fast data streams? *1 K1 CO3*
(a) Incremental model updates using recent data (b) Full database retraining every hour
(c) Ignoring new incoming data (d) Storing all samples offline

7. If a page receives more links from high-ranked pages, its PageRank will: *1 K2 CO4*
(a) Decrease (b) Increase (c) Remain constant (d) Become zero

8. A "support count" in association rule mining indicates: *1 K1 CO4*
(a) How often an itemset appears in transactions
(b) The total number of items in a database
(c) The accuracy of predictions
(d) The cost of transactions

9. Why does K-Means algorithm use centroids to represent clusters? *1 K2 CO5*
(a) Because centroids minimize the average distance within a cluster
(b) Because centroids represent outliers
(c) Because centroids increase randomness
(d) Because centroids store all raw data

10. Hierarchical clustering can be of two types: *1 K1 CO5*
(a) Agglomerative and Divisive (b) Linear and Non-linear
(c) Numeric and Categorical (d) Static and Dynamic

PART - B (12 × 2 = 24 Marks)

Answer ALL Questions

11. Outline the practical uses of feature extraction. *2 K2 CO1*
12. State the benefits of a combiner in a MapReduce job. *2 K2 CO1*
13. Identify the properties of Hamming distance. *2 K1 CO2*

K1 – Remember; K2 – Understand; K3 – Apply; K4 – Analyze; K5 – Evaluate; K6 – Create

14. Compare cosine and Euclidean distances. *2 K2 CO2*

15. Define a bucket which is used in DGIM algorithm. *2 K1 CO3*

16. Illustrate the process used to reduce errors while counting 1's in a window. *2 K2 CO3*

17. State the two possible outcomes generated by Toivonen's Algorithm. *2 K1 CO4*

18. Define PageRank and its basic concept in web structure analysis. *2 K2 CO4*

19. Mention the alternative rules that can be used to control hierarchical clustering. *2 K2 CO5*

20. Describe the competitive ratio used in online algorithm for advertising. *2 K2 CO5*

21. Interpret the role of hashing shingles in dimensionality reduction. *2 K2 CO2*

22. State the significance of multistage algorithm. *2 K1 CO4*

PART - C (6 × 11 = 66 Marks)

Answer ALL Questions

23. a) Describe the MapReduce framework in detail. Draw the architectural diagram for physical organization of computer nodes. *11 K2 CO1*

OR

b) Elaborate the use of mapReduce computing in grouping and aggregation operations. *11 K2 CO1*

11

24. a) Implement Locality-Sensitive Hashing (LSH) to handle different applications of similarity search. *11 K3 CO2*

OR

b) Use Jaccard similarity to identify textually similar documents on the web and demonstrate its necessity. *11 K3 CO2*

25. a) Illustrate the application of the moment estimation technique by computing different moments (e.g., mean, variance) using practical examples. *11 K3 CO3*

OR

b) Demonstrate how the Bloom filtering technique can be applied to efficiently select relevant tuples from a data stream. *11 K3 CO3*

26. a) Implement the PageRank calculation by outlining and applying the steps of the PageRank algorithm with an example. *11 K3 CO4*

OR

b) Use the Market Basket model to analyze customer purchase patterns with a practical example. *11 K3 CO4*

27. a) Demonstrate the two phases of the CURE algorithm by applying them to a given dataset. *11 K3 CO5*

OR

b) Apply the GRGPF algorithm to perform clustering on data in non-Euclidean spaces. *11 K3 CO5*

28. a) Apply the concept of a Data Stream Management System (DSMS) to demonstrate how continuous data streams are processed, and illustrate the constraints involved in handling such streams with suitable examples. *11 K3 CO3*

OR

b) Demonstrate the working of the Flajolet–Martin algorithm by counting distinct elements in a given stream. *11 K3 CO3*