Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

B.Tech Degree S8 (R,S) Examinations April 2026 (2019 Scheme)

Course Code: AIT426
Course Name: MINING OF MASSIVE DATASETS

Max. Marks: 100 **Duration: 3 Hours**

PART A

Answer all questions, each carries 3 marks. Marks

1		What is the TF.IDF measure, and how is it useful in data mining?	(3)
2		Describe the importance of feature extraction in data mining and give an example.	(3)
3		What is a "Distributed File System" (DFS)? Name two examples of DFS implementations.	(3)
4		Explain the purpose of the "Reduce" task in the MapReduce programming model.	(3)
5		How do different techniques for sampling data in streams contribute to managing large datasets?	(3)
6		What are the fundamental components and functions that characterize a Data-Stream-Management System?	(3)
7		Explain the two fundamental strategies of the clustering algorithm.	(3)
8		How to merge two consecutive buckets? What things are done in clustering?	(3)
9		What are online algorithms and how are they relevant in various aspects of computing and advertising?	(3)
10		How is "mining social-network graphs" utilized in advertising strategies, and what significance does it hold in marketing campaigns?	(3)

PART B

Answer any one full question from each module, each carries 14 marks.

Module I

11	a)	What are hash functions, and why are they important in data mining? Provide an example of their use.	(7)
	b)	What is Bonferroni's Principle and how does it apply to data mining? Illustrate with an example.	(7)

OR

12	a)	Describe the computational approaches to modeling in data mining and how they differ from traditional statistical methods.	(7)
	b)	Explain power law and provide an example of how it limits data mining.	(7)

Module II

13 a) Describe how MapReduce processes data using the Map and Reduce functions, highlighting their respective roles in data transformation and aggregation. (7)

b) What are the relational algebra operations used in database queries of map reduce? (7)

OR

14 a) How is Matrix-Vector Multiplication implemented using the MapReduce framework? (7)

b) Explain the step-by-step sequence involved in the execution of MapReduce. (7)

Module III

15 a) Explain the Datar-Gionis-Indyk-Motwani Algorithm. (7)

b) What are the fundamental challenges associated with stream Processing? (7)

OR

16 a) Explain the basic principles behind Bloom's Filter and discuss its primary applications in data management systems. (7)

b) Why the Alon-Matias-Szegedy Algorithm Works. (7)

Module IV

17 a) Illustrate K-means Algorithm. (7)

b) Explain Clustering in non-Euclidean spaces. (7)

OR

18 a) Describe the rules for bucket sizes in the BDMO Algorithm. How do these rules help in organizing the data efficiently? (7)

b) What is the difference between hierarchical and point assignment algorithms in the case of clustering strategies? (7)

Module V

19 a) Explain social networks as graphs with examples. (7)

b) How does the Girvan-Newman algorithm iteratively identify and remove edges from a network to detect community structure? (7)

OR

20 a) Explain the greedy algorithm for maximal matching. (7)

b) Discuss the different types of social networks based on their primary functionalities. (7)

****