Q. Code:297854

Reg. No. | | | | | | | | | | | | | |

B.E./ B. TECH. DEGREE EXAMINATIONS, MAY 2024

Seventh Semester

CS18003 – DATA ANALYTICS

(Computer Science and Engineering)

(Regulation 2018 / 2018A)

TIME:3 HOURS **MAX. MARKS: 100**

COURSE OUTCOMES	STATEMENT	RBT LEVEL
CO 1	Students will be able to illustrate the importance of data and data analysis.	**2**
CO 2	Students will be able to interpret the probabilistic models for data.	**3**
CO 3	Students will be able to apply the knowledge of hypothesis, uncertainty principle in data mining streams.	**3**
CO 4	Students will be able to interpret the evaluation of regression analysis and various clustering algorithms on item sets and frequency count datasets.	**4**
CO 5	Students will be able to investigate Hadoop framework and Hadoop Distributed File system and to illustrate the concepts of NoSQL using MongoDB and Cassandra for Big Data.	**4**

PART- A (10 x 2=20 Marks)

(Answer all Questions)

		CO	RBT LEVEL
1.	State the advantages of EADS over traditional ADS.	1	2
2.	Give examples for structured, semi structured and unstructured data.	1	2
3.	Mention the rule to represent fuzzy logic.	2	3
4.	What is multivariate analysis of variance?	2	1
5.	Compare data streams and traditional DBMS.	3	3
6.	What are DGIM’s maximum error boundaries?	3	1
7.	Differentiate multistage and PCY.	4	2
8.	Infer the different ways to choose clustroid.	4	4
9.	What will happen when one of the nodes on which a map task is running fails?	5	2
10.	Identify the type of S3 storage service. What is S3 storage used for?	5	2

PART- B (5 x 14=70Marks)

		Marks	CO	RBT LEVEL
11. (a)	Discuss the challenges of conventional systems in handling big data.	**(14)**	**1**	**2**
	(OR)			
(b)	Discuss the importance of sampling distributions in big data analytics.	**(14)**	**1**	**2**

12. (a)	Illustrate the different models for time series analysis.	**(14)**	**2**	**3**
	(OR)			
(b)	What is regression modeling? Explain how it is used to analyze data with appropriate illustration.	**(14)**	**2**	**3**
13. (a)	Suppose the stream consists of the integers 1, 2, 1, 3, 2, 5, 7, 6, 1, 4. Determine the number of distinct elements if the hash function is: h(x) = (2x + 2) mod 4. Assume the length of binary string as 3. Show all the steps of your solution using Flajolet-Martin algorithm. Also, list out the advantages and disadvantages of the algorithm.	**(14)**	**3**	**3**
	(OR)			
(b)	Compute the surprise number (second moment) for the stream 3, 1, 4, 1, 3, 4, 2, 1, 2. What is the third moment of this stream? For each possible value of i, if Xi is a variable starting position i, what is the value of Xi.value?	**(14)**	**3**	**3**
14. (a)	Analyze the efficiency of hierarchical clustering algorithms.	**(14)**	**4**	**4**
	(OR)			
(b)	Analyze the approach to form the representation of a cluster in non-Euclidean spaces.	**(14)**	**4**	**4**
15. (a)	Illustrate the different approaches to sharded architectures. Where do all non sharded collections get stored in a sharded cluster? Also, discuss the limitations of sharding.	**(14)**	**5**	**3**
	(OR)			
(b)	Demonstrate the visual data analysis techniques with its datatypes, visualization and interaction techniques.	**(14)**	**5**	**3**

<u>PART- C (1 x 10=10 Marks)</u>

(Q.No.16 is compulsory)

		Marks	CO	RBT LEVEL
16.	Illustrate the working of bloom filter with examples. How do you reduce the false positive rate ?	**(10)**	**3**	**3**

****