Q. Code: 131675

Reg. No.													

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

Second Semester

CP22204 – BIG DATA ANALYTICS

(Computer Science and Engineering)

(Regulation 2022)

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

COURSE OUTCOMES	STATEMENT	RBT LEVEL
CO 1	Design algorithms by employing Map Reduce technique for solving Big Data problems.	**3**
CO 2	Design algorithms for Big Data by deciding on the apt Features set.	**3**
CO 3	Design algorithms for handling petabytes of datasets.	**3**
CO 4	Design algorithms and propose solutions for Big Data by optimizing main memory consumption.	**3**
CO 5	Design solutions for problems in Big Data by suggesting appropriate clustering techniques.	**3**

PART- A (20 x 2= 40 Marks)

(Answer all Questions)

		CO	RBT LEVEL
1.	If you have an input file of 350 MB, how many input splits would HDFS create and what would be the size of each input split?	**1**	**3**
2.	List out the applications of Bonferroni's principle.	**1**	**2**
3.	Give some examples for distributed file systems.	**1**	**2**
4.	Suppose there is a repository of ten million documents. What is the IDF for a word that appears in 10,000 documents?	**1**	**3**
5.	Illustrate Minhashing.	**2**	**2**
6.	Compute the Jaccard similarity of each pair of the following three sets: {1,2,3,4}, {2,3,5,7} and {2,4,6}.	**2**	**3**
7.	What is the Hamming distance between the vectors 10101 and 11110?	**2**	**3**
8.	List out the applications of LSH.	**2**	**2**
9.	There are several ways that the bit-stream 1001011011101 could be partitioned into buckets. Find all of them.	**3**	**3**
10.	Analyse the issues in stream processing.	**3**	**3**
11.	What is the purpose for filtering streams? Mention the techniques used to filter streams.	**3**	**3**
12.	Give a note on the approach to find the most-common elements in the stream.	**3**	**3**

13.	How page rank helps in measuring of a web page within a set of similar entities?	**4**	**3**
14.	How do you identify spam mass in a page?	**4**	**2**
15.	How do you deal with dead ends of the graph?	**4**	**2**
16.	What are the limitations of apriori algorithm?	**4**	**3**
17.	Differentiate centroids and clusteroids.	**5**	**3**
18.	What are the benefits of using CURE clustering algorithms in data analytics?	**5**	**2**
19.	What is meant by adwords problem?	**5**	**2**
20.	How recommender systems use collaborative filtering?	**5**	**2**

PART- B (5x 10=50Marks)

		Marks	CO	RBT LEVEL
21(a)	Relate two variables in different ways by power laws that govern phenomena with examples.	**(10)**	**1**	**3**
	(OR)			
(b)	Illustrate the work flow of MapReduce. How node failures are handled in HDFS?	**(10)**	**1**	**3**
22(a)	Outline shingling of documents with a suitable example.	**(10)**	**2**	**3**
	(OR)			
(b)	Analyse the different ways to study the distance measures.	**(10)**	**2**	**3**
23. (a)	Compute the surprise number for the following stream: a, b, c, b, d, a, c, d, a, b, d, c, a, a, b. What is the third moment of this stream? For each possible value of i, if X_i is a variable starting position i, what is the value of X_i.value?	**(10)**	**3**	**3**
	(OR)			
(b)	Suppose the stream consists of the integers 3, 1, 4, 1, 5, 9, 2, 6, 5. Determine the number of distinct elements if the hash function is: h(x) = (3x + 7) mod 32. Assume the length of binary string as 5. Show all the steps of your solution using Flajolet-Martin algorithm.	**(10)**	**3**	**3**
24. (a)	Find the frequent itemsets and generate association rules on the following table. Assume that minimum support threshold (s= 33.33%) and minimum	**(10)**	**4**	**3**

confident threshold (c = 60%).

Transaction ID	Items
T1	Hot Dogs, Buns, Ketchup
T2	Hot Dogs, Buns
T3	Hot Dogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	Hot Dogs, Coke, Chips

(OR)

(b) Analyse the limitations of apriori algorithm. Also, outline any two algorithms to overcome it in limited passes to find the frequent itemsets. **(10)** **4** **3**

25. (a) Investigate how hierarchical clustering algorithm works in Euclidean space. **(10)** **5** **3**

(OR)

(b) Analyze the content-based architecture for a recommendation system. **(10)** **5** **3**

PART- C (1 x 10=10 Marks)

(Q.No.26 is compulsory)

Marks **CO** **RBT LEVEL**

26. Design MapReduce algorithms to take a very large file of integers and produce an output: **(10)** **1** **4**

i. The largest integer.

ii. The same set of integers, but with each integer appearing only once.

****