Immediate Immediate Immediate		Q. Code	e: 83	1919	
Second Semester CP22204 - BIG DATA ANALYTICS (Computer Science and Engineering) (Regulation 2023) TIME: 3 HOURS MAX.MARKS: 100 COURSE MAX.MARKS: 100 O 1 Design algorithms for Big Data by deciding on the apt Features set. 3 COURSE Course consumption. 3 COURSE PART- A (20 x 2 = 40 Marks) COURSE PART- A (20 x 2 = 40 Marks) COURSE PART- A (20 x 2 = 40 Marks) COURSE PART- A (20 x 2 = 40 Marks) COURSE <th colspan<="" th=""><th></th><th>Reg. No.</th><th></th><th></th></th>	<th></th> <th>Reg. No.</th> <th></th> <th></th>		Reg. No.		
Second Semester CP22204 - BIG DATA ANALYTICS (Computer Science and Engineering) (Regulation 2023) TIME: 3 HOURS MAX.MARKS: 100 COURSE MAX.MARKS: 100 O 1 Design algorithms for Big Data by deciding on the apt Features set. 3 COURSE Course consumption. 3 COURSE PART- A (20 x 2 = 40 Marks) COURSE PART- A (20 x 2 = 40 Marks) COURSE PART- A (20 x 2 = 40 Marks) COURSE PART- A (20 x 2 = 40 Marks) COURSE <th colspan<="" th=""><th></th><th></th><th></th><th></th></th>	<th></th> <th></th> <th></th> <th></th>				
CP22204 - BIG DATA ANALYTICS (Computer Science and Engineering) (Regulation 2022) MAX. MARKS: 100 COURSE STATEMENT Image: Computer Science and Engineering) (Regulation 2022) MAX. MARKS: 100 COURSE STATEMENT Image: Computer Science and Engineering) (Regulation 2022) OLIVES MAX. MARKS: 100 COURSE STATEMENT Image: Computer Science and Engineering) (Regulation 2022) Advantage: Computer Science and Engineering) (Regulation 2022) COURSE COURSE SCIENCE and propose solutions for Big Data by optimizing main memory consumption. 3 CO S Design algorithms and propose solutions for Big Data by suggesting appropriate clustering techniques. CO PART- A (20 x 2 = 40 Marks) (Answer all Questions) CO RATE- A (20 x 2 = 40 Marks) (Answer all Questions) CO CO Pefine Bonferroni's principle. 1 1 . PART- A (20 x 2 = 40 Marks) (Answer all Questions) 1 . PART- A (20 x 2 = 40 Marks) (Answer all Questions) 1 . PART- A (20 x 2 = 40 Marks) (Answer all Questions) 1 . PART- A (20 x 2 = 40 Marks) (Answer					
(Computer Science and Engineering) (Regulation 2022) MAX.MARKS: 100 COURSE NAX.MARKS: 100 OPESIGN algorithms for Big Data by deciding on the apt Features set. 3 COURSE on Segunal algorithms for handling petabytes of datasets. 3 COURSE on Segunal algorithms and propose solutions for Big Data by optimizing main memory consumption. 3 COURSE on Segunate Device Segunations for Big Data by suggesting appropriate clustering techniques. COURSE CONSUMPTION COURSE ON COURSE ON COURSE OF COURSE ON COURSE ON COURSE OF COURSE ON					
(Regulation 2022) MAX. MARKS: 100 COURSE: STATEMENT Regulation 2022) COURSE: STATEMENT Regulation 2022) O 1 Design algorithms by employing Map Reduce technique for solving Big Data problems. 3 O 1 Design algorithms for Big Data by deciding on the apt Features set. 3 O 1 Design algorithms for handling petabytes of datasets. 3 O 2 Design algorithms and propose solutions for Big Data by optimizing main memory consumption. 3 COURSE: PART- A (20 x 2 = 40 Marks) (Answer all Questions) CO RE CO Efine Bonferroni's principle. 1 1 . PART- A (20 x 2 = 40 Marks) (Answer all Questions) CO RE . PART- A (20 x 2 = 40 Marks) . PART- A (20 x 2 = 40 Marks) . Define statistical model. <th col<="" td=""><td></td><td></td><td></td><td></td></th>	<td></td> <td></td> <td></td> <td></td>				
COURSE DECOMESSTATEMENTREF001Design algorithms by employing Map Reduce technique for solving Big Data problems.3022Design algorithms for Big Data by deciding on the apt Features set.303Design algorithms for handling petabytes of datasets.304Design algorithms and propose solutions for Big Data by optimizing main memory consumption.305Design algorithms and propose solutions for Big Data by suggesting appropriate clustering techniques.3PART- A (20 x 2 = 40 Marks) (Answer all Questions)CORE COPART- A (20 x 2 = 40 Marks) (Answer all Questions)CORE CORE LEVOptimizing main memory consumption.O besign solutions for problems in Big Data by suggesting appropriate clustering techniques.PART- A (20 x 2 = 40 Marks) (Answer all Questions)CORE RE LEVO befine Bonferroni's principle.11Define statistical model.1Describe about GFS and HDFS.1Define Hamming Distance.22How is an Efficient Minhashing achieved?22What is Bloom Filters?31					
Immonsion Imponsion Imponsion Imponsion Imponsion Imponsion Imponsion Imponsion Imponsion Imponsion Imponsion Imponsion Imponsion Imponsion Imponsion Imponsion Imponsion Imponsion Imponsion			₹KS:		
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 RB LEV CO RB Design solutions principle. 1 Define Bonferroni's principle. 1 1 Describe about GFS and HDFS. 1 1 Output the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, 2 3 Question Filters? 2 1 Output the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, 2 3 Question Filters? 2 1 Output the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, 2 3 Question Filters? 3 1 Question Filters? 3 1 Output the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, 2 2	OUTCO	MES		RBT LEVEL	
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 7 CO 4 (Answer all Questions) CO 8 PART- A (20 x 2 = 40 Marks) (Answer all Questions) CO 8 LEV CO 4 (Answer all Questions) CO 8 LEV CO 8 BERGIN A (20 x 2 = 40 Marks) (Answer all Questions) CO 8 LEV CO 8 LEV Define Bonferroni's principle. 1 Define statistical model. 1 Describe about GFS and HDFS. 1 Define Hamming Distance. 2 Compute the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, 2 A How is an Efficient Minhashing achieved? 2 <td>CO 1 CO 2</td> <td></td> <td>ems.</td> <td>3 3</td>	CO 1 CO 2		ems.	3 3	
consumption. 1 1 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 RB LEV (Answer all Questions) CO . Define Bonferroni's principle. 1 1 . Define statistical model. 1 1 . Define statistical model. 1 1 . Describe about GFS and HDFS. 1 2 . What is k-Shingles? 2 1 . Compute the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, 2 3 . Define Hamming Distance. 2 1 . How is an Efficient Minhashing achieved? 2 2 . What is Bloom Filters? 3 1 0. What are the different queries used in data-stream management system? 3 1 1. Define Decaying Windows. 3 1 1 3. What is meant by Page Rank? 4 1	CO 3	Design algorithms for handling petabytes of datasets.		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 RB LEV . Define Bonferroni's principle. 1 1 . Define statistical model. 1 1 . Define statistical model. 1 1 . What are the two approaches to modeling? 1 1 . Describe about GFS and HDFS. 1 2 . What is k-Shingles? 2 1 . Compute the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, 2 3 . Define Hamming Distance. 2 1 . How is an Efficient Minhashing achieved? 2 2 . What is Bloom Filters? 3 1 0. What are the different queries used in data-stream management system? 3 1 1. Define Decaying Windows. 3 1 2. What is a sampling data in a data stream? 3 1 3. What is meant by Page Rank? 4 1 <td>CO 4</td> <td colspan="2"></td> <td>3</td>	CO 4			3	
(Answer all Questions)CORBLEV.Define Bonferroni's principleDefine statistical modelWhat are the two approaches to modeling?.What are the two approaches to modeling?.Describe about GFS and HDFSWhat is k-Shingles? <td>CO 5</td> <td colspan="4">CO 5 Design solutions for problems in Big Data by suggesting appropriate clustering</td>	CO 5	CO 5 Design solutions for problems in Big Data by suggesting appropriate clustering			
CORBLEV11Define Bonferroni's principle.11Define statistical model.11What are the two approaches to modeling?11Describe about GFS and HDFS.12What is k-Shingles?21Compute the Jaccard Similarity of each pair of the following three sets: {1,2,3,4},2Compute the Jaccard Similarity of each pair of the following three sets: {1,2,3,4},2Define Hamming Distance.21How is an Efficient Minhashing achieved?22What is Bloom Filters?31Uwhat are the different queries used in data-stream management system?31What is a sampling data in a data stream?31What is meant by Page Rank?41					
.Define Bonferroni's principle.11.Define statistical model.11.What are the two approaches to modeling?11.Describe about GFS and HDFS.12.Describe about GFS and HDFS.12.What is k-Shingles?21.Compute the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, {2,3,5,7}, {2,4,6}.21.Define Hamming Distance.21.How is an Efficient Minhashing achieved?22.What is Bloom Filters?310.What are the different queries used in data-stream management system?311.Define Decaying Windows.312.What is a sampling data in a data stream?313.What is meant by Page Rank?41		(Answer all Questions)	CO	RBT	
 What are the two approaches to modeling? Describe about GFS and HDFS. Describe about GFS and HDFS. What is k-Shingles? Compute the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, {2,3,5,7}, {2,4,6}. Define Hamming Distance. How is an Efficient Minhashing achieved? What is Bloom Filters? What is Bloom Filters? What are the different queries used in data-stream management system? Define Decaying Windows. What is a sampling data in a data stream? What is meant by Page Rank? 	1.	Define Bonferroni's principle.		level 1	
 Describe about GFS and HDFS. What is k-Shingles? Compute the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, 2 1 Compute the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, 2 1 Compute the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, 2 1 Define Hamming Distance. How is an Efficient Minhashing achieved? What is Bloom Filters? What is Bloom Filters? What are the different queries used in data-stream management system? Define Decaying Windows. What is a sampling data in a data stream? What is meant by Page Rank? 	2.	Define statistical model.	1	1	
 What is k-Shingles? Compute the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, {2,3,5,7}, {2,4,6}. Define Hamming Distance. How is an Efficient Minhashing achieved? What is Bloom Filters? What is Bloom Filters? What are the different queries used in data-stream management system? Define Decaying Windows. What is a sampling data in a data stream? What is meant by Page Rank? Mat is meant by Page Rank? 	3.	What are the two approaches to modeling?	1	1	
 Compute the Jaccard Similarity of each pair of the following three sets: {1,2,3,4}, 2 {2,3,5,7}, {2,4,6}. Define Hamming Distance. How is an Efficient Minhashing achieved? What is Bloom Filters? What are the different queries used in data-stream management system? Define Decaying Windows. What is a sampling data in a data stream? What is meant by Page Rank? 4 	4.	Describe about GFS and HDFS.	1	2	
 {2,3,5,7}, {2,4,6}. Define Hamming Distance. How is an Efficient Minhashing achieved? What is Bloom Filters? What are the different queries used in data-stream management system? Define Decaying Windows. What is a sampling data in a data stream? What is meant by Page Rank? 	5.	What is k-Shingles?	2	1	
 How is an Efficient Minhashing achieved? What is Bloom Filters? What are the different queries used in data-stream management system? Define Decaying Windows. What is a sampling data in a data stream? What is meant by Page Rank? 		$\{2,3,5,7\}, \{2,4,6\}.$		3	
 What is Bloom Filters? What are the different queries used in data-stream management system? Define Decaying Windows. What is a sampling data in a data stream? What is meant by Page Rank? 	7.	Define Hamming Distance.	2	1	
0. What are the different queries used in data-stream management system?311. Define Decaying Windows.312. What is a sampling data in a data stream?313. What is meant by Page Rank?41	8.	How is an Efficient Minhashing achieved?	2	2	
1. Define Decaying Windows.312. What is a sampling data in a data stream?313. What is meant by Page Rank?41	9.	What is Bloom Filters?	3	1	
2. What is a sampling data in a data stream?313. What is meant by Page Rank?41	10.	What are the different queries used in data-stream management system?	3	1	
3. What is meant by Page Rank? 4 1					
4. What is meant by spider traps?41					
	14.	What is meant by spider traps?	4	1	

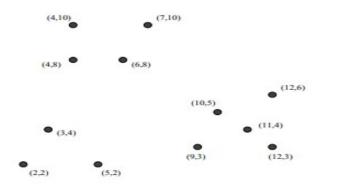
- Explain Association rule in mathematical notation 16.
- What is meant by curse of dimensionality? 17.
- Differentiate centroids and clusteroids. 18.
- Discuss about Collaborative and Content based Re 19.
- Define the CURE algorithm. 20.

PART- B (5 x 10 = 50 Marks)

		Marks	CO	RBT LEVEL
21. (a)	Demonstrate any two computational approaches to modeling in detail.	(10)	1	3
	(OR)			
(b)	(i) Illustrate in detail about any two algorithms using MapReduce.	(7)	1	3
	(ii) Apply Reduce Function for those two algorithms in MapReduce?	(3)	1	3
22. (a)	Illustrate about Shingling of Documents with a suitable example.	(10)	2	3
	(OR)			
(b)	(i) Demonstrate Matrix representation of sets with an example.	(4)	2	3
	(ii) Explain Minhashing in detail and how Minhashing is used along with	(6)	2	3
	Jaccard Similarity.			
23. (a)	Illustrate the architecture of data stream management system in detail.	(10)	3	3
	(OR)			
(b)	Explain briefly the methods applied for counting distinct elements.	(10)	3	3
24. (a)	(i) Demonstrate Early Search Engines and Term Spam in detail.	(5)	4	3
	(ii) Illustrate structure of the web with diagram.	(5)	4	3
	(OR)			
(b)	Explain algorithms used for handling larger datasets in main memory.	(10)	4	3
25. (a)	Perform a hierarchial clustering of the one-dimensional set of points 1, 4, 9,	(10)	5	4
	16, 25, 36, 49, 64, 81 assuming the clusters are represented by their centroid			
	(average), and at each step the cluster with the closest centroids are merged.			
	(OR)			
(b)	Compute the radius, in the sense used by the GRGPF Algorithm (square root	(10)	5	4
	Page 2 of 4			

	Q. Code: 831919		
	4 1		
ns.	4 2		
	5 1		
	5 2		
ecommendation.	5 2		
	5 1		

of the average square of the distance from the clustroid) for the cluster that is the five points in the lower right of given figure. Note that (11,4) is the clustroid.



PART- C (1 x 10 = 10 Marks)

(Q.No.26 is compulsory)

Marks	CO	RBT
		LEVEL
(10)	5	5

Cluster the following eight points (with (x, y) 26. three clusters:

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2). The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as- P(a, b) = |x2 - x1| + |y2 - y1|

Use K-Means Algorithm to find the three cluster centers after the second iteration.

Q. Code: 831919