30/12/23
Afn.

Roll No. | | | | | | | | | | |

ANNA UNIVERSITY (UNIVERSITY DEPARTMENTS)

B.E. / B. Tech / B. Arch (Full Time) - END SEMESTER EXAMINATIONS, NOV / DEC 2023

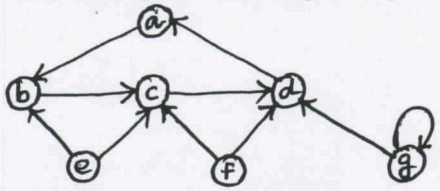
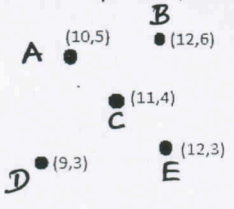
INFORMATION TECHNOLOGY
Fifth Semester
ITM502 & BIG DATA ANALYTICS
(Regulation 2019)

Time: 3hrs Max.Marks: 100

CO 1	Understand the basics of Big Data.
CO 2	Know about Hadoop and MapReduce.
CO 3	Know about Big Data technology, Tools and Algorithms.
CO 4	To analyze the Stream data and Link analysis.
CO 5	Know about the role of big data in Recommender systems and social network analysis.
CO 6	Design and Implementation of basic data intensive applications.

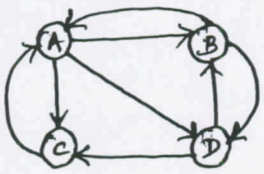
BL – Bloom's Taxonomy Levels
(L1 - Remembering, L2 - Understanding, L3 - Applying, L4 - Analyzing, L5 - Evaluating, L6 - Creating)

PART- A (10 x 2 = 20 Marks)
(Answer all Questions)

Q. No	Questions	Marks	CO	BL
1	Differentiate business intelligence and big data.	2	1	L2
2	What are the challenges of big data analytics?	2	1	L2
3	Mention some of the features of MongoDB?	2	2	L1
4	List the mapper and reduce formulas for matrix multiplication.	2	2	L3
5	Where the Apache Flume is used?	2	3	L4
6	What is Apache Oozie? Mention its purpose.	2	3	L1
7	Check whether the following graph having dead end, spider trap? And also give the corresponding matrix using binary values. 	2	4	L5
8	Apply the Datar-Gionis-Indyk-Motwani algorithm on the default window and estimate the number of 1's in the last k position, where k= 7 and 12.	2	4	L6
9	Consider the following random points, find out the closest points.  A = (10,5) B = (12,6) C = (11,4) D = (9,3) E = (12,3).	2	5	L5
10	What is recommendation system? Mention its different types.	2	5	L1

PART- B (5 x 13 = 65 Marks)

(Restrict to a maximum of 2 subdivisions)

<table>
<tr><th>Q. No</th><th>Questions</th><th>Marks</th><th>CO</th><th>BL</th></tr>
<tr><td>11 (a) (i)</td><td>Explain in detail about various classification of big data analytics.</td><td>6</td><td>1</td><td>L2</td></tr>
<tr><td>(ii)</td><td>A manufacturing company is looking to optimize its production processes using big data analytics. What data sources and analytics techniques would you use to identify areas for improvement and increase operational efficiency?</td><td>7</td><td>1</td><td>L4</td></tr>
<tr><td colspan="5">OR</td></tr>
<tr><td>11 (b) (i)</td><td>Briefly explain about some of the terminologies used in the big data environments.</td><td>6</td><td>1</td><td>L2</td></tr>
<tr><td>(ii)</td><td>A retail chain is experiencing a sudden drop in sales for a specific product category. How would you approach analyzing this situation using big data analytics? What steps would you take to identify the root causes?</td><td>7</td><td>1</td><td>L4</td></tr>
<tr><td>12 (a)</td><td>Explain the semantic representation of MapReduce algorithm and its workflow with suitable example.</td><td>13</td><td>2</td><td>L3</td></tr>
<tr><td colspan="5">OR</td></tr>
<tr><td>12 (b)</td><td>For the following matrix
<table>
<tr><th>Element</th><th>S1</th><th>S2</th><th>S3</th><th>S4</th></tr>
<tr><td>a</td><td>1</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>b</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>c</td><td>0</td><td>1</td><td>0</td><td>1</td></tr>
<tr><td>d</td><td>1</td><td>0</td><td>1</td><td>1</td></tr>
<tr><td>e</td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>f</td><td>0</td><td>1</td><td>1</td><td>1</td></tr>
</table>
i) Compute the min hash signature for the each column if we use the following hash functions
$h_1(x) = 2x + 4 \bmod 5$
$h_2(x) = 3x + 1 \bmod 5$
$h_3(x) = 7x - 1 \bmod 5$
ii) How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?</td><td>13</td><td>2</td><td>L3</td></tr>
<tr><td>13 (a) (i)</td><td>Briefly explain about Hadoop distributed file system with neat diagram.</td><td>8</td><td>3</td><td>L2</td></tr>
<tr><td>(ii)</td><td>Explain about Apache Pig architecture.</td><td>5</td><td>3</td><td>L2</td></tr>
<tr><td colspan="5">OR</td></tr>
<tr><td>13 (b) (i)</td><td>Shots notes on: HIVE and SPARK.</td><td>8</td><td>3</td><td>L2</td></tr>
<tr><td>(ii)</td><td>List out the various advantages of HDFS.</td><td>5</td><td>3</td><td>L2</td></tr>
<tr><td>14 (a)</td><td>Compute Trust Rank and Spam mass of each page for the following graph, by assuming A, B as the trusted pages.


i) Calculate the PageRank.
ii) Compute the Trust Rank.
iii) Compute the Spam mass.</td><td>13</td><td>4</td><td>L3</td></tr>
<tr><td colspan="5">OR</td></tr>
</table>

14 (b)	The stream consists of the integers 3, 1, 4, 1, 5, 9, 2, 6, 5. Determine the tail length of each stream element and the resulting estimate of the number of distinct elements if the hash function is i) h(x) = 2x+1 mod 32 ii) h(x) = 3x+7 mod 32 iii) h(x) = 4x mod32	13	4	L3
15 (a)	Explain in detail about Collaborating Filtering with suitable examples.	13	5	L5
	OR			
15 (b)	Define hierarchical clustering. Perform a hierarchical clustering of one dimensional set of points 1, 4, 9, 16, 25, 36, 49, 64, 81. Assuming clusters are represented by their centroid and at each step the clusters with the closest centroid are merged.	13	5	L5

PART- C (1 x 15 = 15 Marks)

(Q.No.16 is compulsory)

Q. No	Questions	Marks	CO	BL
16. (i)	Explain the Park-Chen and Yu (PCY)algorithm	5	4	L2
(ii)	For the following collection of 12 baskets. Each contains three of 6 items 1 through 6. {1,2,3} {2,3,4} {3,4,5} {4,5,6} {1,3,5} {2,4,6} {1,3,4} {2,4,5} {3,5,6} {1,2,4} {2,3,5} {3,4,6} The support threshold is 3. On the first pass of the PCY algorithm use a hash table with 11 buckets, and set {i, j} is hashed to buckets i*j mod 10. i) Compute the support for each item and each pair of items. ii) Which pairs hash to which buckets? iii) Which buckets are frequent? iv) Which pairs are counted on the second pass of the PCY algorithm?	10	4	L4