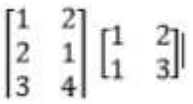
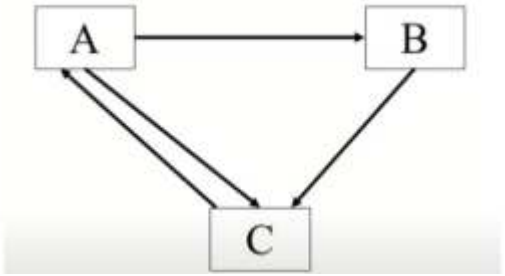
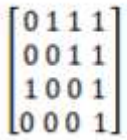
Sr No.	Questions	CO	Marks
1	How does Big Data create value for businesses?	**CO1**	2
2	What is the significance of velocity in Big Data?	**CO1**	2
3	List the tools used in big data analytics.	**CO1**	2
4	What are the challenges of storing Big Data?	**CO1**	2
5	Explain the concept of Big Data and identify various sources from which Big Data is generated.	**CO1**	2
6	Explain Unstructured Data with example.	**CO1**	2
7	Explain the "Volume" characteristics in Big Data	**CO1**	2
8	Describe the different types of Big Data.	**CO1**	5
9	Distinguish between Traditional Data Processing and Big Data Processing with example.	**CO1**	5
10	Examine the influence of the 'Volume' characteristic of Big Data on the effectiveness and scalability of Big Data analytics systems.	**CO1**	5
11	Discuss the new trends in Big Data analytics and their impact on various industries.	**CO1**	5
12	Outline the life cycle of big data analytics, detailing each phase and its significance	**CO1**	5
13	Describe how social media collects and uses big data to make decisions	**CO1**	5
14	How does the diversity of data types and formats within a dataset show the 'Variety' characteristic of big data?	**CO1**	5
15	Outline the five major challenges in managing Big Data and explore their effects on decision-making processes in organizations.	**CO1**	10
16	Discuss the distinguishing characteristics of Big Data compared to traditional data, using examples to illustrate these differences.	**CO1**	10
17	define big data analytics and explain the types of big data analytics with suitable example	**CO1**	10
18	a. List and explain examples of big data b. Discuss the challenges related to data privacy and security in Big Data, proposing viable solutions.	**CO1**	10
19	Construct a basic architecture for a Big Data solution focused on real-time analytics, using a Walmart case study as the foundation for your design.	**CO1**	10
20	Examine how Big Data is applied in healthcare, detailing its impact on patient outcomes with relevant examples or case studies.	**CO1**	10
21	Develop a Netflix case study in Big Data analytics, explaining the role of machine learning with specific examples.	**CO1**	10
22	What is Hadoop?	**CO2**	2
23	List the core components of Hadoop.	**CO2**	2
24	Explain the concept of fault tolerance in Hadoop.	**CO2**	2
25	Describe the role and significance of blocks in HDFS for managing large datasets	**CO2**	2
26	Illustrate what is a Secondary Name Node? Does it substitute the Name Node?	**CO2**	2
27	Classify why HDFS is more suited for applications having large datasets and not when there are small files?	**CO2**	2

28	List Advantages and Disadvantages of NoSQL	**CO2**	2
29	Explain how big data problems are handled by Hadoop system.	**CO2**	5
30	Compare and contrast RDBMS with NoSQL databases, emphasizing key differences and scenarios where each is most effective.	**CO2**	5
31	Examine the architecture of Pig, explaining how it processes data and integrates with the Hadoop ecosystem.	**CO2**	5
32	Analyze the architecture of HBase, focusing on its design principles and how it supports large-scale data storage and retrieval.	**CO2**	5
33	Examine the role of Apache Spark in the Hadoop ecosystem, discussing its architecture and analyzing different use cases."	**CO2**	5
34	Analyze basic Hadoop commands, explaining their functions and how they are used to manage Hadoop operations.	**CO2**	5
35	Evaluate the role of "FsImg" and "Edit logs" in HDFS, discussing their significance in ensuring data consistency and integrity.	**CO2**	5
36	Examine the architecture of YARN, describing its components and how it manages resources in a Hadoop cluster.	**CO2**	10
37	Design a data pipeline using Hadoop, where data is ingested using Sqoop, processed using Apache Spark, and stored in HBase. Include a diagram to illustrate the architecture	**CO2**	10
38	Analyze the components of the Hadoop Ecosystem, summarizing their roles and how they work together to handle Big Data.	**CO2**	10
39	Explore the process of reading and writing data in HDFS, detailing the steps involved and their significance in Hadoop's operation.	**CO2**	10
40	Examine RDDs in Spark, explaining their types and providing detailed examples of their usage in real-world applications.	**CO2**	10
41	Analyze the role of Apache Spark in the Hadoop ecosystem, discussing its architecture and analyzing different use cases.	**CO2**	10
42	Analyze the key features of different types of NoSQL databases and differentiate their capabilities and use cases. b. Analyze the benefits and limitations of using NoSQL databases like Cassandra in comparison to traditional RDBMS.	CO2	10
43	Describe the function of the Map task within the MapReduce framework	**CO3**	2
44	Define the term "Reducer" in the context of MapReduce.	**CO3**	2
45	Explain the concept of "Grouping by Key" in MapReduce.	**CO3**	2
46	Explain the concepts of union and intersection in the context of relational algebra operations.	**CO3**	2
47	What is Grouping & Aggregation in Relational Algebra Operations	**CO3**	2
48	Identify and describe a real-world example where MapReduce is effectively applied	**CO3**	2
49	Analyze and list the benefits associated with using the MapReduce process.	**CO3**	2
50	Describe the role of combiners in optimizing MapReduce jobs and provide an example.	**CO3**	5
51	Develop a pseudo code for implementing the Matrix-Vector Multiplication algorithm using MapReduce.	**CO3**	5
52	Compare and analyze traditional approaches to parallel and distributed processing, highlighting their methods and limitations.	**CO3**	5
53	Describe the detailed execution process of a MapReduce job, from the submission to the output.	**CO3**	5
54	Analyze how MapReduce can be used to perform matrix-vector multiplication.	**CO3**	5

<table>
<tr><td>55</td><td>Illustrate with an example how MapReduce can be used to compute the projection operation in relational algebra.</td><td>CO3</td><td>5</td></tr>
<tr><td>56</td><td>Write a map reduce pseudo code for word count problem. Illustrate with an example showing all the steps.</td><td>CO3</td><td>5</td></tr>
<tr><td>57</td><td>Analyze how relational algebra operations are computed and executed using MapReduce, including the role of Map and Reduce functions</td><td>CO3</td><td>10</td></tr>
<tr><td>58</td><td>i. Describe Map Reduce algorithm for Natural Join of two relations with example.
ii. Describe Map Reduce algorithm to perform Intersection of two sets with example.</td><td>CO3</td><td>10</td></tr>
<tr><td>59</td><td>Evaluate matrix vector multiplication by map reduce. Illustrate with an example ,Consider the following matrix.
</td><td>CO3</td><td>10</td></tr>
<tr><td>60</td><td>Explain the operation of grouping and aggregation using Map Reduce with Given example?
Map Worker 1
File 1
<table><tr><th>A</th><th>B</th><th>C</th><th>D</th></tr><tr><td>1</td><td>2</td><td>3</td><td>1</td></tr><tr><td>2</td><td>2</td><td>3</td><td>2</td></tr><tr><td>1</td><td>2</td><td>1</td><td>3</td></tr></table>File 2
<table><tr><th>A</th><th>B</th><th>C</th><th>D</th></tr><tr><td>4</td><td>2</td><td>1</td><td>3</td></tr><tr><td>6</td><td>8</td><td>4</td><td>4</td></tr><tr><td>3</td><td>2</td><td>2</td><td>4</td></tr></table>Map Worker 2
File 1
<table><tr><th>A</th><th>B</th><th>C</th><th>D</th></tr><tr><td>1</td><td>2</td><td>5</td><td>2</td></tr><tr><td>2</td><td>3</td><td>2</td><td>4</td></tr><tr><td>1</td><td>3</td><td>1</td><td>3</td></tr></table>File 2
<table><tr><th>A</th><th>B</th><th>C</th><th>D</th></tr><tr><td>3</td><td>2</td><td>5</td><td>2</td></tr><tr><td>2</td><td>3</td><td>9</td><td>2</td></tr><tr><td>3</td><td>4</td><td>2</td><td>1</td></tr></table></td><td>CO3</td><td>10</td></tr>
<tr><td>61</td><td>Demonstrate the Map Reduce program without combiner and with Combiner using suitable example</td><td>CO3</td><td>10</td></tr>
<tr><td>62</td><td>Create a detailed diagram that summarizes the MapReduce process in nutshell</td><td>CO3</td><td>10</td></tr>
<tr><td>63</td><td>Explain the operation of union using Map Reduce with Given example
Map Worker 1
Table 1
<table><tr><th>A</th><th>B</th></tr><tr><td>1</td><td>2</td></tr><tr><td>2</td><td>3</td></tr><tr><td>5</td><td>6</td></tr></table>Table 2
<table><tr><th>A</th><th>B</th></tr><tr><td>2</td><td>3</td></tr><tr><td>4</td><td>4</td></tr><tr><td>6</td><td>1</td></tr></table>Map Worker 2
Table 1
<table><tr><th>A</th><th>B</th></tr><tr><td>6</td><td>1</td></tr><tr><td>6</td><td>3</td></tr><tr><td>7</td><td>6</td></tr></table>Table 2
<table><tr><th>A</th><th>B</th></tr><tr><td>9</td><td>8</td></tr><tr><td>3</td><td>3</td></tr><tr><td>0</td><td>1</td></tr></table></td><td>CO3</td><td>10</td></tr>
<tr><td>64</td><td>What is Decaying Windows ? Explain in detail?</td><td>CO4</td><td>2</td></tr>
<tr><td>65</td><td>Discuss different issues in data stream processing</td><td>CO4</td><td>2</td></tr>
<tr><td>66</td><td>List and explain various data stream sources</td><td>CO4</td><td>2</td></tr>
<tr><td>67</td><td>How do sensor networks and social media platforms contribute as sources of data streams?</td><td>CO4</td><td>2</td></tr>
<tr><td>68</td><td>How does a stream query differ from a traditional SQL query?</td><td>CO4</td><td>2</td></tr>
</table>

69	What is data stream? Explain data stream opeartions in the context of big data?	**CO4**	2
70	Discuss the main components of the DGIM algorithm?	**CO4**	2
71	Explain count-distinct problem in stream processing with suitable example?	**CO4**	5
72	Explain the difference between batch processing and stream processing, highlighting the advantages of stream processing.	**CO4**	5
73	Investigate and discuss scenarios where a Bloom filter might produce false positives and false negatives. How can these issues be mitigated?	**CO4**	5
74	Explain what a Data Stream Management System (DSMS) is, and illustrate its architecture with a block diagram.	**CO4**	5
75	Explain what stream queries are and describe the different categories of stream queries with examples.	**CO4**	5
76	Imagine you're processing a binary stream to estimate the count of ones in a sliding window. Apply the concept of the DGIM algorithm to achieve this.	**CO4**	5
77	Compare the efficiency of reservoir sampling with other sampling techniques in stream processing.	**CO4**	5
78	Given a scenario where you need to check the existence of an item in a large dataset, explain how you would use a Bloom filter to perform this check?	**CO4**	10
79	Determine the distinct element in the stream using FM algorithm. Input stream of Integers= 1,3,2,1,2,3,4,3,1,2,3,1 h(x)=(6x+1)mod 5	**CO4**	10
80	Determine the distinct element in the stream using FM algorithm. Input stream of Integers= 4, 2, 5, 9, 1, 6, 3, 7 a) h(x) = 3x + 1 mod 32 b) h(x) = x + 6 mod 32	**CO4**	10
81	Analyze the operation of a Bloom filter in detail. Provide a thorough explanation of how it works and illustrate its functionality with a practical example.	**CO4**	10
82	Explain how the Flajolet-Martin algorithm can be used to estimate the number of distinct elements in a stream.Analyse the trade-offs between the Flajolet-Martin algorithm and the Datar-Gionis-Indyk-Motwani (DGIM) algorithm in terms of accuracy and memory space requirements.	**CO4**	10
83	Compare and contrast the challenges of estimating the count of distinct elements in a stream with the challenges of counting ones in a sliding window. How do the algorithms address these challenges?	**CO4**	10
84	A)Apply the concept of Bloom filters to a situation where you're processing incoming emails. Explain how a Bloom filter could be used to identify potential spam emails. B) A Bloom filter with m=1000 cells is used to store information about n=100 items, using k=4 hash functions. calculate the false positive probability of this instance	**CO4**	10
85	What is the main focus of frequent pattern mining, and why is it important in data analysis?	**CO5**	2
86	Describe the fundamental principle behind clustering algorithms and their purpose in data analysis.	**CO5**	2
87	Discuss the applications of Canopy Algorithm	**CO5**	2
88	Describe the main goals of visual data analysis techniques and how they facilitate better understanding of complex datasets.	**CO5**	2
89	Discuss the advantages of PCY algorithm?	**CO5**	2
90	How large data handled in main memory? Enlist the algorithm used to handle large datasets in main memory?	**CO5**	2

91	Explain all three pass of multistage algorithm	**CO5**	2
92	Given a large dataset and memory constraints, outline the steps of the basic Park, Chen, and Yu algorithm for frequent pattern mining.	**CO5**	5
93	List down two algorithms used for frequent pattern mining? Describe their key characteristics in detail.	**CO5**	5
94	Analyse the strengths and limitations of the CURE algorithm for clustering in comparison to traditional centroid-based clustering algorithms.	**CO5**	5
95	Compare the Canopy clustering approach to hierarchical agglomerative clustering. What are the computational and memory trade-offs between these methods?	**CO5**	5
96	Illustrate how the SON partition-based algorithm helps to perform frequent item set mining for large datasets. How does this algorithm avoid False negatives?	**CO5**	5
97	What is the primary function of visualization techniques in data analysis, and how do they aid in extracting insights from data?	**CO5**	5
98	Analyse the impact of varying the number of principal components retained after PCA on the quality of data representation and the explanatory power of the reduced dimensions.	**CO5**	5
99	Analyse clustering with MapReduce.	**CO5**	10
100	Illustrate how the PCY algorithm is used in big data analytics.	**CO5**	10
101	Apply PCY algorithm on the following transactions to find candidate set. use buckets and the concept of mapreduce to solve the problem. Threshhold:2, Hash Function: (i*j) mod 10 T1:{1,2,3} T2:{2,3,4} T3:{3,4,5} T4:{4,5,6} T5:{1,3,5} T6:{2,4,6} T7:{1,3,4} T8:{2,4,5} T9:{3,5,6} T10:{1,2,4} T11:{2,3,5} T12:{3,4,6}	**CO5**	10
102	Apply different classification algorithms to a sample dataset. Describe how you would implement algorithms such as Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), and Logistic Regression. Explain the steps involved in each algorithm and how you would evaluate their performance using metrics like accuracy, precision, and recall.	**CO5**	10
103	Compare and contrast the strengths and limitations of simple linear regression and multiple linear regression models in terms of their applicability and complexity.	**CO5**	10
104	Given a dataset with a high number of features, outline the steps you would take to perform dimension reduction using PCA. What considerations would guide your choices?	**CO5**	10
105	How does the SON (Savasere, Omiecinski, and Navathe) algorithm leverage MapReduce to mine frequent patterns? What advantage does MapReduce bring to this process?	**CO5**	10
106	What are dead ends in the context of link analysis on the web?	**CO6**	2
107	Explain the difference between content-based recommendations and collaborative filtering in recommendation engines?	**CO6**	2
108	What is clustering in the context of social network graphs? Name a common algorithm used for clustering social network graphs.	**CO6**	2
109	Define a dangling node in a web graph	**CO6**	2
110	List and briefly describe the common types of social network graphs?	**CO6**	2
111	Explain utility matrix with example	**CO6**	2
112	Explain how PageRank is used in search engines to rank web pages. What does a higher PageRank indicate?	**CO6**	2

113	Analyze the fundamental concept of the PageRank algorithm and evaluate its significance in web search and ranking of web pages.	**CO6**	5
114	Explain structure of web? Explain spider trap in detail?	**CO6**	5
115	Explain the fundamental characteristics that define social networks as graphs. What are nodes and edges in this context?	**CO6**	5
116	How does the process of direct discovery of communities work in the analysis of social network graphs?	**CO6**	5
117	What is the primary objective of recommendation systems, and why are they valuable for businesses and users?	**CO6**	5
118	Compare and contrast the direct discovery of communities in social network graphs with other community detection algorithms. What are the strengths and weaknesses of each approach?	**CO6**	5
119	Describe the concept of topic-sensitive PageRank and how it differs from the standard PageRank algorithm.	**CO6**	5
120	Compute the page rank of each page in the following graphs after 3 iterations 	**CO6**	10
121	Apply the principles of link spam detection. What are common strategies used by link spammers, and how can PageRank be used to identify them?	**CO6**	10
122	Apply different types of recommendation systems (e.g., content-based, collaborative filtering, hybrid) to recommend products in an online bookstore. Provide specific examples and discuss the expected benefits and limitations of each type in this context.	**CO6**	10
123	Analyse the importance of hubs and authorities in link analysis. How do these concepts relate to the HITS (Hyperlink-Induced Topic Search) algorithm?	**CO6**	10
124	Compute hub and authority score for the matrix using the HITS algorithm with k=3 and identify the best authority and hub node 	**CO6**	10
125	Consider the Web graph given below with six pages (A, B, C, D, E, F) with directed links as follows: • A → B, C • B → A, D, E, F • C → A, F Assume that the PageRank values for any page m at iteration 0 is PR(m) = 1 and teleportation factor for iterations is β = 0.85. Perform the PageRank algorithm and determine the rank for every page at iteration 2.	**CO6**	10
126	Apply the principles of collaborative filtering to recommend movies to a user based on their previous movie ratings and the ratings of other similar users.	**CO6**	10