**Module 1**

Big Data

Characteristics of Big Data

Applications of Big data.

Difference between Traditional Database and Big Data

Differentiate structured, Semi structured and unstructured Database

Different V’s Supported by Big Data

give 2 examples of big data case studies indicate which V’s are satisfied by this case study

What are the problems/challenges with handling big data?

**Module 2**

Explain how Hadoop Framework handles BigData? - ecosystem wala ans

Hadoop and explain architecture of Hadoop

Explain core components of Hadoop.

Explain Hadoop Ecosystem with neat diagram

What are the advantages and limitations of Hadoop

explain the distributed storage system of hadoop with neat diagram

Explain architecture of HDFS with neat diagram.

Heartbeat Message in HDFS Architecture

Explain architecture and functions of HDFS ,YARN, and MAP Reduce

Explain functions of NameNode and Secondary NameNode in HDFS

or namenode and datanode in hdfs

Explain the concept of Blocks,default Block Size and replication factor and

Architecture of Spark

Explain MongoDB with its features.

What is NOSQL? Differentiate SQL and NoSQL

Explain different types of architecture pattern of NoSQL database.

Explain CAP theorem.

Explain the types of NoSQL data stores and their typical usage.

Explain CAP theorem and explain how NoSQL systems guarantees BASE property

**Module III: MapReduceParadigm**

What is MapReduce ? Explain How Map and Reduce Work with example..

Explain role of Combiner in MApReduce.

Write MAP Reduce algorithms for following operations:

Map Reduce algorithm for Word Count Frequency

Relational-Algebra Operations, any 5 with examples

Computing Selections by MapReduce,

Union, Intersection

Computing Natural Join by MapReduce,

Grouping and Aggregation by MapReduce,with example

Matrix Multiplication, Matrix Multiplication with One MapReduce Step.

Matrix Vector Multiplication

Illustrating use of MapReduce with use of real life databases and applications.

Explain Grouping and Aggregation algorithm using MapReduce. Support your answer with a suitable example

Write Map-Reduce algorithm to count occurrences of distinct words from a given text and apply same on following text: Hadoop is a big data framework. Hadoop manages distributed data using Hadoop Distributed File System.

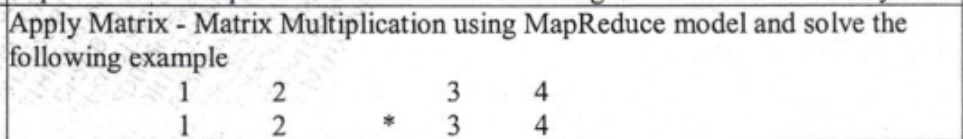
**sum :-**

Comment on {key, value) pair produced by Map Function while evaluating a) following query using Map Reduce.Query: Select avg(salary) from Employee group by dno

List Map-Reduce steps involved in Matrix multiplication and Find Matrix-Matrix Multiplication using 1-step MapReduce

A= [2 3] b= [ 2 4 3]

[ 4 5 ] [ 3 6 2]





**Module IV Mining Big Data Streams**

Explain Data stream and its characteristics

Explain Issues and challenges in Data Stream

Data stream mining sources and applications

Different operations on Data Stream

Sampling Techniques:  
 Filtering Steams:

Explain Bloom filter with example.

explain bloom filter for stream data mining Summarize Bloom’s filter with example and its applications. FM algorithm DGIM algorithm Explain the DGIM algorithm. State the rules used in DGIM that must be followed Sliding Window Give two applications for counting the number of 1's in a long stream of binary values. Using a stream of binary digits, illustrate how DGIM will find the number of l's.

**Numeric /Problems:**

Bloom Filter

FM algorithm

DGIM algorithm

Counting Distinct Elements in a Stream using FM algorithm.

Suppose a data stream consists of integers 1,3,5,4,6,1,5,9,3,2. Let the hash function used be:

i) h(x)= x+1 mod 16

ii) h(x)= 2x+3 mod 16

iii) h(x) = 3x + 1 mod 16

Show how the Flajolet-Martin algorithm will estimate the number of distinct elements in the stream

**Module V Big Data Mining Algorithms**

explain knn with an example

Need of finding frequent patterns in Big Data.

Applications of finding frequent patterns

Explain PCY algorithm.and its 2 types with neat labeled diagram.

Sons Algorithm

CURE Algorithm and advantages over traditional clustering

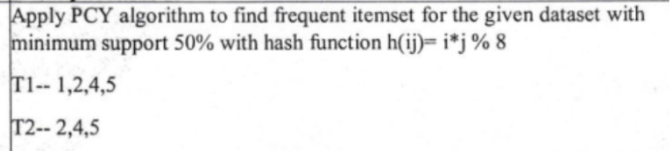
discuss all phases of cure algorithm for clustering with example

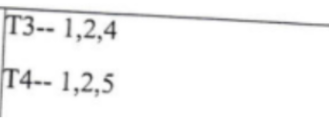
Cure algorithm, Clustering and classification algorithm,SVM,KNN etc

Numeric/Problems:

PCY algorithm

Explain clearly with diagrams the PCY method of finding frequent itemsets (pairs) in a large data set.





To apply the PCY (Park-Chen-Yu) algorithm to find frequent itemsets in the given dataset with a minimum support of 50% and using the hash function

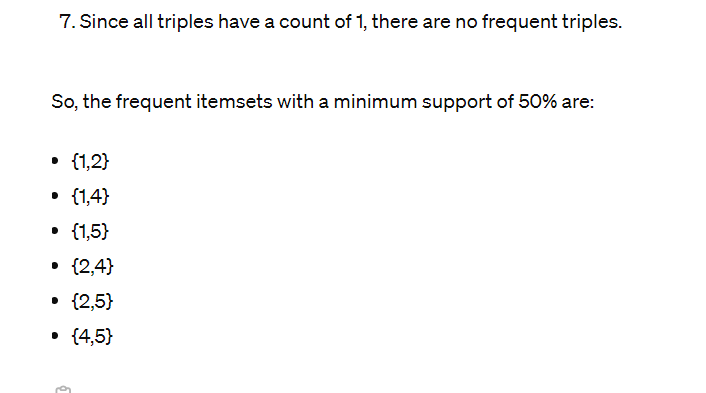
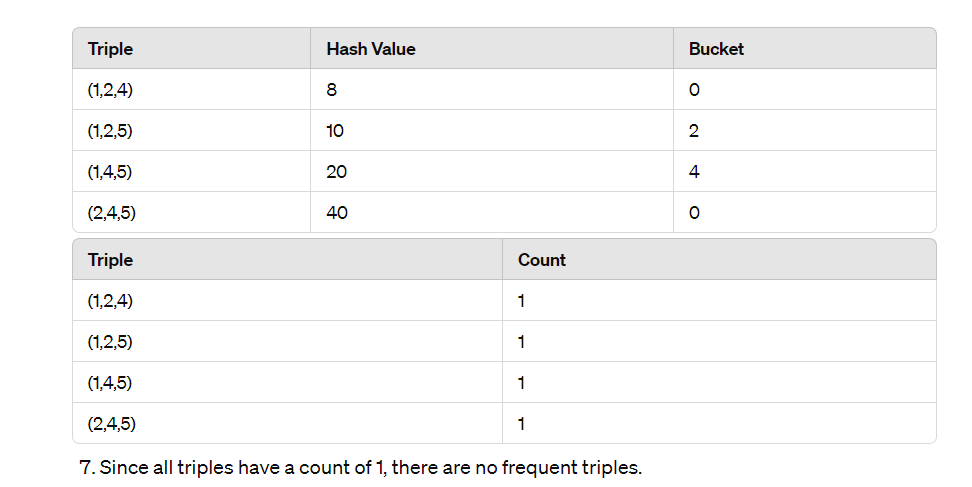
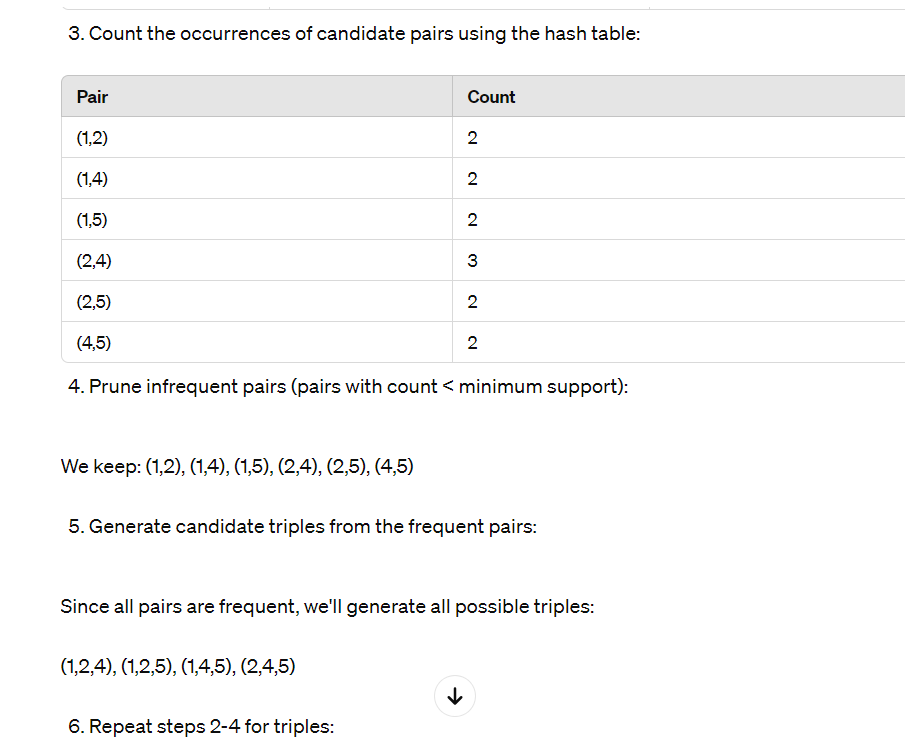
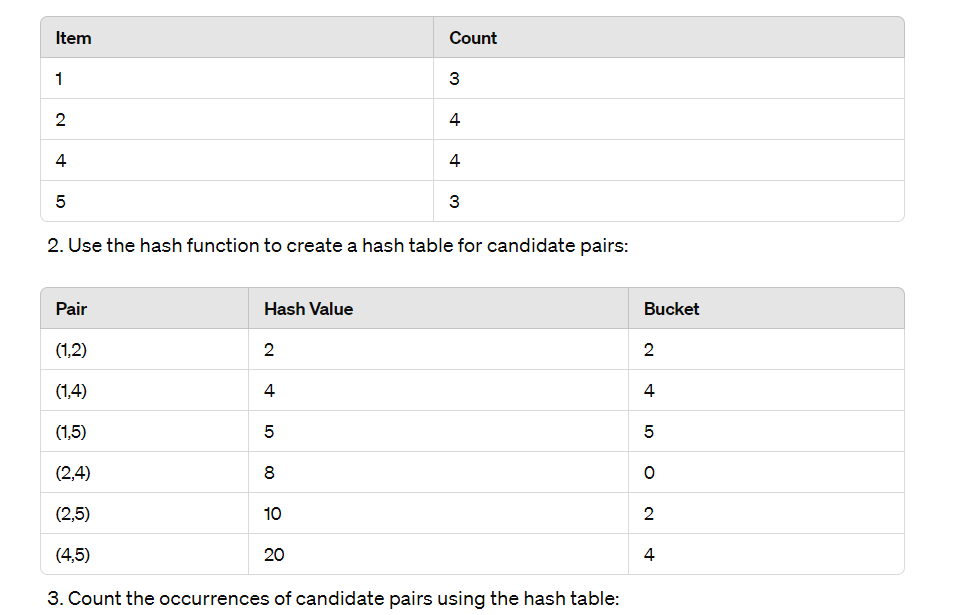
ℎ(𝑖𝑗)=(𝑖×𝑗)%8

*h*(*ij*)=(*i*×*j*)%8, we'll follow these steps:

1. Count the occurrences of each item.
2. Use the hash function to create a hash table for candidate pairs.
3. Count the occurrences of candidate pairs using the hash table.
4. Prune infrequent pairs.
5. Generate candidate triples from the frequent pairs.
6. Repeat steps 2-4 for triples.
7. Continue until no more frequent itemsets can be found.

Let's walk through these steps:

1. Count the occurrences of each item:



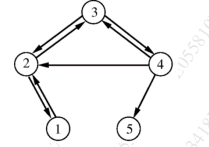
Module 6:

Structure of web

explain the steps of hits algorithm

PageRank Definition, Structure of the web,dead ends, Hubs and Authorities, HITS Algorithm etc.

Recall HITS algorithm. Generate Hub and Authority score after 2 iterations for the graph given here



For the graph given below use Clique percolation and find all communities.

Give two applications for counting the number of 1’s in a long stream of binary values. Using a stream of binary digits, illustrate how DGIM will find the number of 1’s

Explain what characteristics of social media makes it suitable for Big Data

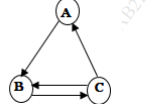
Explain the collaborative filtering based recommendation system

Explain collaborative filtering. How it is different from content based filtering?

Page Rank ,modified page rank,Spider Trap

Collaborative Filtering with applications

Define Hub and Authority. Compute the hub and Authority scores for the web:



option m

