****

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING IN**

**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

**Academic Year 2023-24 (ODD SEM)**

**QUESTION BANK**

**Class: BE/SEM VII Subject: BDA(CSC702)**

**MODULE 1 Introduction to Big Data and hadoop**

1. What are the three Vs of Big Data? Give two examples of big data case studies. Indicate which Vs are satisfied by these case studies. 5
2. What are the characteristics of big Data? 5
3. Explain Hadoop Architectural Model 5
4. write note on hadoop ecosystem 5
5. Explain the Hadoop Architecture with its features. 5
6. List & explain Big data :- 1) Characteristics 2) Types 3) Challenges
7. Explain Hadoop Ecosystem with core components. Explain its physical

architecture. State the limitations of Hadoop. 10

**MODULE 2 Hadoop HDFS and Map Reduce**

1. What is the role of a "combiner " in the Map reduce framework? Explain with the help of one example. (5)
2. Distinguish between Name node and Data node. 5

1. Explain selection and projection relational algebraic operation using Map reduce(10)
2. Explain matrix multiplication using 2 step map reduce model 10
3. Explain Map reduce programming model 5
4. Write a Map-Reduce Algorithm for Binary search tree. Explain the flow of execution. 10
5. List Relational-Algebra Operations. Explain any two using MapReduce.10
6. Explain "Shuffle & Sort" phase and "Reducer Phase" in Mapreduce **with** examples. 5
7. Write a Map reduce pseudo code to multiply two matrices. Illustrate with (10)

an example showing all the steps.

1. Show Map Reduce implementation for the following two tasks using

pseudocode.

i. joins two relations with an example.

ii. Multiplication of two matrices with one Map Reduce step 10

1. Show Map Reduce implementation for the following two tasks using

pseudocode.

i. Multiplication of two matrices

ii. Computing Group-by and aggregation of a relational table. 10

**MODULE 3 NoSQL**

1. What are the different data architecture patterns in NOSQL? Explain "key value" store and "Document" store patterns with relevant examples. 10
2. What are the different data architecture patterns in NOSQL? Explain "Graph Store" and "Column Family " store patterns with relevant examples. 10
3. Explain CAP Theorem? how it is different from ACID Properties.
4. Agility is a NoSQL business driver. Justify.5
5. List the different NoSQL data stores. Explain any two with diagram 10
6. Compare and contrast SQL and NoSQL. 5
7. Distinguish the following:document data store and column family data store. 5

**MODULE 4 Mining Data Streams**

1. List the different issues and challenges in data stream query processing. **(**5)
2. List the different issues in stream processing. **(05)**
3. Clearly explain the concept of a Bloom Filter for stream data mining with the help of an example. 10
4. Explain FM algorithm with an example? **(10)**
5. Describe any two sampling techniques for big data with the help of examples. 10
6. Suppose a data stream consists of the integers 3**,** 1, 4, 1, 5, 9, 2, 6, 5. Let the hash function being used is h(x) = 3x + 1 mod 5; Show how the Flajolet- Martin Algorithm will estimate the number of distinct element in this stream.10
7. Suppose a data stream consists of the integers 1,3,2,1,2,3,4,3,1,2,3,1. Let the Hash function being used is h(x) = (6x+1) mod 5; estimate the number of distinct in this stream using the Flajolet - Martin algorithm.
8. Using an example bit stream explains the working of the DGIM algorithms to count the number of 1’s(ones) in a data stream.
9. clearly explain how the CURE algorithms can be used to cluster big data sets. 10
10. explain DGIM algorithm for counting once in a stream with an example. 10
11. 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. (10)

1. Explain PCY algorithm with suitable examples. 10
2. Give the updating buckets approach of DGIM algorithm. 5
3. Explain Blooms filter for stream data mining. 5
4. Explain Flajolet Martin Algorithm with an example. 10
5. Through an example illustrate how triples can be used to optimally store and count pairs in a frequent itemset mining algorithm. 5
6. Distinguish the following:

i) DBMS and DSMS

ii) PCY, Multistage and Multi Hashing

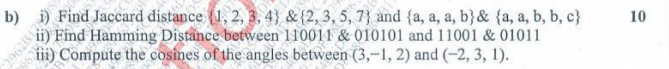
1. Suppose a stream consists of the integers 2,1,6,1,5,9,2,3,5. Let the hash functions all be of the form h(x)=ax+b mod 16 for some a & b. You should treat the result as a 4 bit binary integer. Determine the tail length for each stream element and the resulting estimate of the number of distinct elements if the hash function is :

a) h(x) = 2x +3 mod 16 b) h(x)=4x+1 mod 16 c)5x mod 16 [10]

**MODULE 5 Finding Similar Items and Clustering**

1. Clearly explain how the CURE algorithm can be used to cluster big data sets. 10
2. Define Edit distance. Explain with examples. **(05)**
3. Find the jaccard distance and cosine distance between the following pair of set

X=(0,1,2,4,5,3) and Y=(5,6,7,9,10,8).

1. explain social network graph clustering algorithm with example 10
2. 
3. Explain Different Distance Measures for Big Data. 5
4. write a note on different distance measures that can be used to find similarity / dis similarity between data points in a big data set. 10
5. Write steps of Girvan-Newman Algorithm. Explain clustering of Social-Network

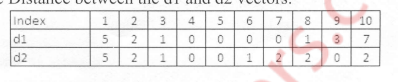
Graphs using GN algorithm with example?

1. Write a short note on

i) PCY Algorithm

ii) CURE algorithm [10]

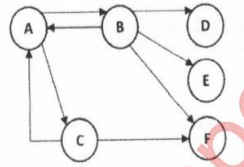
1. Find Cosine Distance between the d1 and d2 vectors:



11 Compute simplified page rank using damping factor d = 0.9 for web. **(10)**



**MODULE 6 Real Time Big Data Models**

1. Define Collaborative filtering. Using an example of an e-commerce site like Flipkart or Amazon describe how it can be used to provide recommendations to users. 10
2. What is the motivation to count triangles in a social network graph?outline one algorithm for counting triangles briefly. 5
3. Explain collaborative filtering systems. How is it different from a content based system? 10
4. What is the role and effect of page rank? **(05)**
5. Define PageRank. Using the web graph shown below compute the PageRank at every node at the end of the second iteration. Use teleport factor = 0.8.10
6. 
7. **(a)** Explain clearly with diagrams how the **PCY** algorithm helps to perform frequent itemset mining for large datasets. 10
8. explain model for recommendation system in detail.10
9. explain page ranking algorithm with suitable example.10
10. What are the shortcomings of the nearest neighbor technique in collaborative filtering method? Suggest some improvements. [10]
11. explain different types of recommendation systems with real time example.10
12. Explain the role and effect of the damping factor (teleportation) in PageRank computation. 5
13. 
14. What are different recommender systems? Explain any one with an example. [10]
15. Consider the web graph given below with six pages (A, B, C, D, E, F)

with directed links as follows.



Assume that the PageRank values for any page m at iteration 0 is PR(m)=1 and teleportation factor for iterations is ẞ=0.85. Perform the page rank algorithm and determine the rank for every page at iteration 2.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Course Incharge H.O.D

Prof. Snehal Gavale