



- **KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE**
- **DON'T WRITE ANYTHING ON THE QUESTION PAPER**

COs	CO Statements
CO1	Identify big data systems and design for analysis.
CO2	Store and analyse data in Hadoop.
CO3	Use graph models for solving real world problems.
CO4	Process Data using Spark and No SQL Databases.
CO5	Apply MapReduce based analysis.

BL – Blooms Taxonomy Level (1 – Remember, 2 – Understand, 3 – Apply, 4 – Analyse, 5 – Evaluate, 6 – Create)

Answer ALL Questions

(10 X 10 = 100 Marks)

1. a) Analyze the following scenarios and identify the most appropriate Data Analysis Method. [5] CO1 BL4

Scenario 1:

A financial analyst wants to determine how interest rates, GDP growth, and inflation impact stock prices to estimate future market performance.

Scenario 2:

A fitness app wants to evaluate how users who joined during a “New Year Resolution” campaign perform compared to users who joined during a mid-year discount offer, in terms of app activity and retention.

Scenario 3:

An education research group collects student feedback on various aspects of an online learning platform and wants to identify the hidden constructs (like usability, engagement, content clarity) that impact overall learning satisfaction.

Scenario 4:

A renewable energy startup wants to assess the financial risks of different investment strategies in solar farm deployment, considering uncertain weather conditions, energy demand variability, and price fluctuations.

- b) Analyze the following scenarios and identify the type of analytics [5]

Scenario 1:

A streaming platform wants to summarize the most-watched genres and peak viewing hours over the past six months.

Scenario 2:

A banking institution notices a sudden spike in loan application rejections and wants to analyze the contributing factors behind this trend.

Scenario 3:

A retail chain wants to forecast inventory requirements for the upcoming festive season based on historical sales patterns.

Scenario 4:

A flight scheduling system wants to recommend the best flight routes and departure timings to reduce fuel consumption and delays.

2. A fraud system reads the last 40 bits of a transaction stream (1 = suspicious, 0 = normal):
1,0,0,1,1,1,0,1,0,1,0,0,1,1,0,0,1,1,1,0,1,0,1,1,0,0,1,1,0,1,0,1,1,0,1,1,0,1
• Build the DGIM buckets for this 40-bit sliding window (assume the most recent bit is the rightmost).
• Use DGIM to estimate the number of 1s in the last 18 bits. Show the calculation and final estimate.
• Compute the exact count of 1s in that 18-bit suffix and report the relative error.
3. A stream of elements arrives:
p, q, r, p, s, t, r, u, v, w, q, x
A hash function H is defined as:
 $H(\text{element}) = (\text{Sum of ASCII values of characters in element} \times 7 + 13) \bmod 64$
Convert the result into a 6-bit binary number. Apply Flajolet–Martin algorithm. Estimate distinct elements. Compute the actual number of distinct elements in the stream and compare with the estimated count. Comment briefly on the accuracy of the estimation.
4. A Bloom filter is intended to store $n = 6000$ elements with a target false positive probability $P = 0.03$.
• Calculate the required size of the bit array (m).
• Calculate the optimal number of hash functions (k).
A Bloom filter of $m = 24$ bits and $k = 3$ hash functions is configured with the following hash formulas:
• Hash 1: $h_1(x) = (x \times 2) \% 24$
• Hash 2: $h_2(x) = (x \times 3 + 4) \% 24$
• Hash 3: $h_3(x) = (x \times 5 + 1) \% 24$
Perform:
Insert Operation: Insert elements {6, 11, 14} sequentially. Show the state of the bit array after each insertion.
Search Operation: Query elements {6, 8, 11, 20}. Explain whether each element is "definitely not present" or "possibly present" and why?
5. You are given a dataset containing short text snippets extracted from IoT security incident logs. Each log message contains keywords related to cyber-attacks detected in smart devices. Your task is to apply the MapReduce paradigm to compute the frequency of each attack-related keyword and explain the steps involved in solving this problem.
"IoT malware causes denial of service"
"Botnet malware spreads across IoT devices"
"DDoS attack is a form of denial of service"
"Securing IoT devices prevents malware attack"

CO2 BL4

CO2 BL3

CO2 BL3

CO5 BL3

6.

You are a Hadoop administrator managing a cluster. The `/user/hadoop/data/` directory contains various datasets, and you need to perform operations related to uploading, replication, block info, and cluster health.

- a) Task 1: Upload a local file `sales.csv` to HDFS directory `/input/sales/` with a replication factor of 4 during upload. [2]
 - Write the HDFS command to achieve this.
 - Explain briefly what happens in the cluster during this upload.
- b) Task 2: After uploading, you realize that the replication factor of `sales.csv` should be reduced to 2. [2]
 - Write the HDFS command to change the replication factor.
 - Explain how HDFS handles existing replicas during this change.
- c) Task 3: List all files in `/user/hadoop/data` along with block locations and replication details. [2]
 - Write the HDFS command for this operation.
 - Explain what information is displayed in the output.
- d) Task 4: You want to check the status of DataNodes and see if there are any missing or under-replicated blocks. [2]
 - Write the HDFS command(s) to perform this check.
 - Describe what the output indicates about cluster health.
- e) Task 5: You need to put the NameNode into safe mode for maintenance. [2]
 - Write the HDFS command to enter safe mode.
 - Explain why safe mode is important before performing operations like block verification or file deletion.

7.

A retail chain stores transactions across 3 regional servers (Node 1, Node 2, Node 3). Each transaction lists items purchased in a single order.

CO3 BL3

Node 1:

{TV, Remote, HDMI Cable}

{TV, Remote}

{TV, HDMI Cable}

{Remote, HDMI Cable}

{TV, Remote, HDMI Cable}

Node 2:

{TV, HDMI Cable}

{Remote, HDMI Cable}

{TV, Remote}

{TV, HDMI Cable}

{Remote, HDMI Cable}

Node 3:

{TV, Remote, HDMI Cable}

{Remote, HDMI Cable}

{TV, HDMI Cable}

{TV, Remote}

{HDMI Cable}

Global Minimum Support: 6

Compute 1-itemset and 2-itemset supports for each node. Identify locally frequent itemsets for each node. Merge local frequent itemsets to generate candidate global itemsets. Count global support for each candidate itemset. Identify globally frequent itemsets (support ≥ 6). Which items are frequently

bought together? If the retailer wants to create product bundles, which 2-item combinations are the best candidates?

CO4 BL3

8.

Write HBase commands to create a table Employee with column families Personal and Job. Insert the following data using HBase DML commands.

RowKey	Personal:Name	Personal:Age	Job:Title	Job:Dept
E101	Alice	30	Manager	HR
E102	Bob	25	Analyst	IT

Write commands to:

- Retrieve all columns for row E101
- Retrieve only Job:Title for all employees
- Delete Personal:Age column for E102

CO3 BL3

9.a)

Given a small directed graph with 4 pages (A, B, C, D) and the following links:

A → B, C
B → C
C → A
D → C

Compute the PageRank values after one iteration assuming equal initial rank (1/4) and damping factor 0.85. Show all intermediate calculations.

OR

9.b)

Consider a small directed web of 5 pages: P, Q, R, S, T with the following hyperlinks:

P → Q, R
Q → R
R → P, Q
S → R, P
T → (no outlinks — dangling page)

Compute the PageRank values after one iteration assuming equal initial rank (1/5) and damping factor 0.85. Show all intermediate calculations.

CO3 BL3

10.a)

A retail company receives 10 million sales transactions per day. Analyze whether batch processing or real-time processing is more suitable for.

- o End-of-day sales reports
- o Fraud detection during transactions

Justify your answers.

CO4 BL4

OR

10.b)

A smart energy management company collects sensor readings from 50,000 smart meters installed in homes, generating data every 5 seconds. For each of the following tasks, decide whether batch processing or real-time (stream) processing is more suitable and justify your answer with proper reasoning:

- Monthly energy consumption billing for customers.
- Detecting abnormal power usage spikes that may indicate equipment failure or electricity theft.

Explain by considering data volume, latency requirements, processing cost and business impact.

CO4 BL4

⇔⇔⇔ D/L/TY ⇔⇔⇔