Total Marks of Question no.		Examiner 1	A.Y. 2024-25

Space for Marks	Question No.	START WRITING HERE (odd)

Department : CMPN
Semester : 7
Subject : Big Data Analytics (BDA)
Solution of : MSE 1
Exam date : 12/8/24

M.S.E-1 Solution

	Data Mining	Big Data
1a) i)	Data source includes ERP data, CRM data, Web transactions, Financial data of the organization	Data source includes social media data, sensor data, log data, etc... from organization and other sources.
ii)	Volume of data is in Gigabytes to Terabytes	Volume of data is in hundreds of Terabytes to Petabytes Exabytes & even to Zettabytes.
iii)	Generally does not require immediate response & deals with batch or near real time data	often real time data with high velocity that requires immediate response.
iv)	Structured & Semi structured data	High variety of structured & Unstructured data.
v)	Useful for BI, analysis & reporting	Useful for complex BI & predictive analysis.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	1 B)	Volume is data at rest which is in hundreds of tera bytes or to exabytes. In Big Data, this data is obtained from organization as well as from other sources.
		Variety is data in many forms such as structured, unstructured, images, audio, etc.. In Big Data, data is complex due to its variety.
		Eventually, the volume & variety/complexity determines the ~~key~~ value that can be generated from it. ~~For~~ If data is small (organizational data) and is ~~data~~ structured (from RDBMS & organization), then value generated from an application like ERP can be simple pattern analysis ~~to~~ or report generation. It can atmost perform market segmentation based on CRM data. If data is high in volume & variety/complexity, the corresponding application can perform complex analysis like click stream or perform complex transformation like speech to text conversion & complex analysis like Sentiment Analysis. Hence the 2 are key characteristics for Big Data.

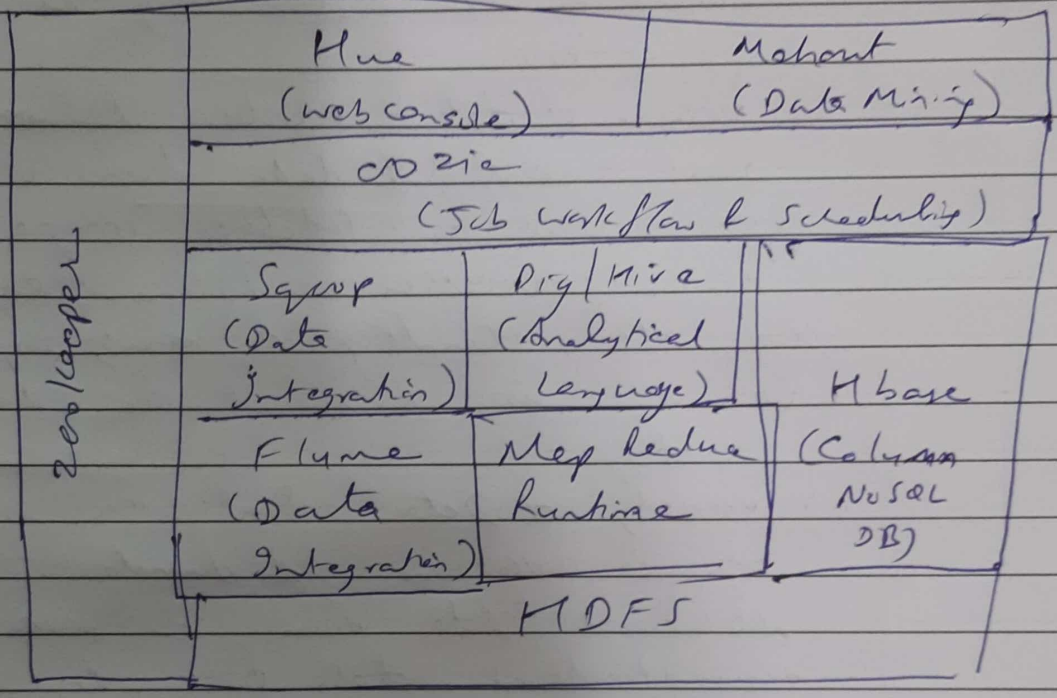
Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	1 c)	One real life application for Big Data is Smart farming.
		Justification of smart farming as Big Data application.
		1) Volume of data: Modern farms generate vast amount of data from various sources such as satellite image, soil sensors, weather conditions, equipment log, etc. Handling & analyzing such volume is crucial for decision making.
		2) Variety of Data: Data in smart farming are in various formats such as structured data from database, semi-structured data from logs, unstructured data from images & sensors. etc.. It is important to integrate this diverse data to provide comprehensive views.
		3) Velocity of Data: Data from sensors and machinery is generated continously and needs to processed in real-time.
		4) Value: Big Data helps in managing risk associated with farming, enable precision in agriculture, allows advanced analytics and optimizes resource utilization.
		Thus, smart farming is a big data application.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE

2A) Architecture of Hadoop Ecosystem



Zookeeper is a centralized service for maintaining configuration information. It provides distributed synchronization. It contains a set of tools to build distributed applications that can safely handle partial failures.

Oozie is a workflow management project. Oozie allows developers to create a workflow of Map Reduce jobs including dependencies between jobs.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	2 B)	Core components of Hadoop are:
		1) Hadoop Distributed File System (HDFS). It provides reliable & scalable storage. It consists of Name Node (master) that manages the metadata and file system name space and it consists of Data Node (slave) that stores the actual data and serves read/write operations.
		2) YARN (Yet Another Resource Negotiator) It manages & schedules resources in cluster. Its main components are Resource Manager that allocates resources among applications and a Node Manager that manages resources on individual nodes.
		3) MapReduce: It is a programming model for processing large data. It consist of Job Tracker (Master), Task Tracker (Slave), Map Task and Reduce Task.
		4) Hadoop Common: It provides common utilities and libraries that support other modules. Example: Libraries for file systems, RPC and serialization.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	2c)	CAP theorem states that only two of the following three can be guaranteed by any distributed database system:
		Consistency
		Availability &
		Partition Tolerance
		Consistency ensures that all nodes see the same data simultaneously.
		Availability guarantees that every request receives a response (success/fail).
		Partition Tolerance ensures that system continues to operate successfully despite network partitions.
		ACID properties (Atomicity, Consistency, Isolation & Durability) ensure reliable transactions in traditional SQL databases. They mainly favour Consistency & Partition tolerance over Availability.
		BASE properties (Basically Available, Soft state and Eventually consistent) are used in NoSQL database. They mainly favour Availability and Partition tolerance over Consistency.
		BASE allows more flexible & scalable systems in contrast to strict consistency as guarantees by ACID.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE
	3A	Map Reduce algorithm to compute total sales of each product category in last one year. Assume: I/p data is in form of csv file. Each record in the csv file has 4 fields date, product_id, category and amount. Example 2024-08-06, 001, Clothing, 300 2024-08-06, 002, Books, 200 2024-08-07, 003, Electronics, 600 2024-08-07, 004, Clothing, 200 2024-08-07, 005, Electronics, 500 2024-08-07, 006, Clothing, 150 and so on...
		Map Task: For each record generate <key, value> pair with key as category and value as amount. For above example: <Clothing, 300> <Books, 200> <Electronics, 600> <Clothing, 200> <Electronics, 500> <Clothing, 150>, etc -- and so on...
		This o/p of mapper goes to next stage.

Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE

which is Shuffle and Sort.
Shuffle & Sort stage groups all Key-value pairs based on Key which is category.
Example o/p of Shuffle & Sort stage

<Clothing, [300, 200, 150]>
<Books, [200]>
<Electronics, [600, 500]>
:

Last stage of Map-Reduce is the Reduce stage.
Reduce Task:
For each Key and list of values, reducer sums the values in the list and generates it with Key as o/p.

Example: o/p of for earlier example will be

<Clothing, 650>
<Books, 200>
<Electronics, 1100>
:
and so on

Thus the Map-Reduce effectively computes total sales of each product category in last one year.

Total Marks of Question no.		Examiner 1	

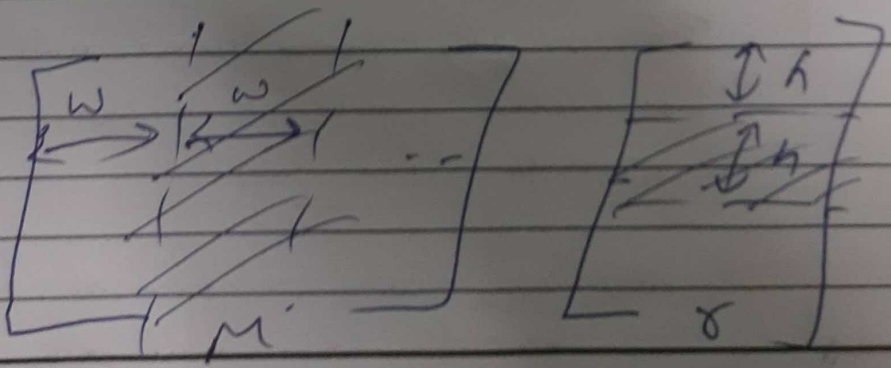
Space for Marks	Question No.	START WRITING HERE
	3 B)	Consider n web pages to be ranked as vector. These pages. The links are represented using a matrix M. Hence, each element m_{ij} of matrix M represents a link from page i to page j. This is to be multiplied with rank vector. V. The new rank will be computed as

$$r = \sum_{j=1}^{n} m_{ij} * v_j$$

Here, the size of matrix M $(n \times n)$ is square of number of web pages and size of rank vector V $(n \times 1)$ is of order number of web pages.

As number of web pages are large (in billions), we cannot fit rank of all web pages in memory of single system.

Hence we split the ranks of vector and with same width split the matrix



Total Marks of Question no.		Examiner 1	

Space for Marks	Question No.	START WRITING HERE

Each Map task gets a strip of matrix M and the corresponding ranks of from the rank vector (as shown shaded for one map task)

Map Task :

Each map task operates on a strip K and has corresponding values of vector v.

For each matrix element M_{ij}, Map task produces key-value pair as $(i, m_{ij} * v_j)$

Shuffle & Sort stage :

It groups all products $m_{ij} * v_j$ and produces a list of products for a specific i.

$(i, [m_{ij} * v_j, m_{i2} * v_2, m_{i3} * v_3, \ldots])$

Reduce task

It simply sums all the values associated with a given key i and generates (i, x_i).

In this way, using matrix-vector multiplication, we can compute new rank for billions of webpages