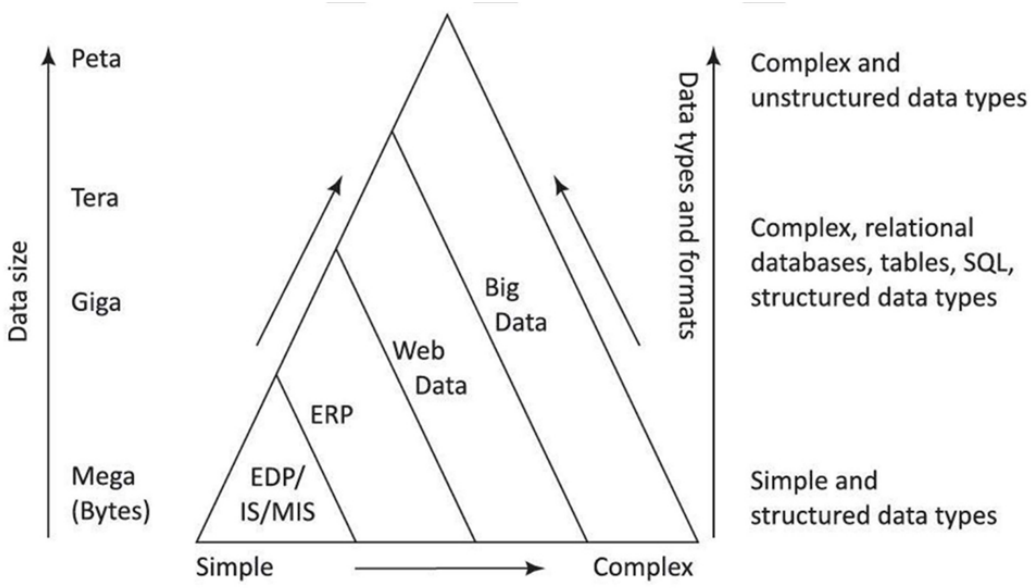
USN | | | | | | | | | | |

Internal Assessment Test 1 – Oct 2023

Iat 1 -BDA -Solution

Sub:	Big Data Analytics					Sub Code:	18CS72	Branch:	CSE
Date:	20/10/2022	Duration:	90 mins	Max Marks:	50	Sem / Sec:	7 –A/B/C		OBE

	Answer any FIVE FULL Questions	MARKS	CO	RBT
1 (a)	Consider satellite images of the Earth's atmosphere and its regions. The *Volume* of data from the satellites is large. A number of Indian satellites, such as KALPANA, INSAT-1A and INSAT-3D generate this data. Foreign satellites also generate voluminous data continuously. Satellites record the images of full disk and sectors, such as east and west Asia sectors and regions. In the above Example List all the 5 V's of Big data and Explain. • Volume: is related to size of the data The Volume of data from the satellites is large. • Velocity: refers to the speed of generation of data. • Variety: comprises of a variety of data Structured and unstructured data • **Veracity:** quality of data captured, which can vary greatly, affecting its accurate analysis , according to example - Image quality , data which we could able to capture due to hazards etc **Value:** Satellites data could be used to draw insights about weather , natural disaster , etc	[05]	CO 1	L2
(b)	Explain Evolution of Big Data and their characteristics	[05]	CO 1	L1



The rise in technology has led to the production and storage of voluminous amounts of data. Earlier megabytes (10^6 B) were used but nowadays petabytes (10^{15} B) are used for processing, analysis, discovering new facts and generating new knowledge. Conventional systems for storage, processing and analysis pose challenges in large growth in volume of data, variety of data, various forms and formats, increasing complexity, faster generation of data and need of quickly processing, analyzing and usage.

Figure 1.1 shows data usage and growth. As size and complexity increase, the proportion of unstructured data types also increases.

- *Volume: is related to size of the data*
- *Veloctiy: refers to the speed of generation of data.*
- *Variety: comprises of a variety of data*
- *Varacity: quality of data captured, which can vary greatly, affecting its accurate analysis*

2 (a)	List and explain ***usage of Big Data Analytics*** in a ***Company for car manufacturing, Marketing sales and maintenance of car service centers.*** ***Explanation for manufacturing -1 Mark*** ***Marketing Sales - 2 Marks*** ***Maintenance - 2 Marks***	[05]	CO 1	L3

27
a) CAR SERVICE COMPANY

→ To handle a scenario of a car service company, where the cars are manufacturing, sales are counted and also car care is provided, we use the Big data features (or) characteristics.

→ Variety: The types of cars, car models, the colors, service charges, manufacturing costs and labour charges are different formats of data. These types must be seggregated and put into a structured arrangement.

∴ This explains the variety attribute

CMR

→ Velocity: A car company must maintain a distributed storage unit to continuously monitor the sales of cars in the given time-frame.

-Also, velocity of data matters when the cars that are being serviced must be charged according to the corresponding services taken, so that the customers are not allowed to face service delays & billing delays.

→ Value: The amount of data that is being processed per day (or) per hour (or) any time-frame is recognized and constantly monitored.

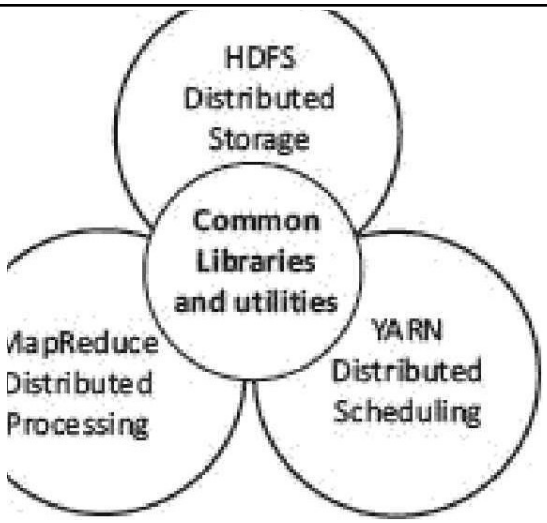
This monitoring helps in scaling the management system of car company.

→ Volume: Each sector of servicing; manufacturing, deploys materials and goods, sales sector produce huge amount of data where the volume attribute is considered. The data is processed / stored / analyzed and must be preserved also.

→ Veracity: The financial amount (or) the service charges schema must be accurate to specify the car company genuinity. The quality of products and cars is observed and must be taken into consideration.

Data loss is not entertained, so the information is automatically replicated into other components.

(b)	Explain the Hadoop ecosystem with diagrams.	[05]	CO 1	L1



The Hadoop core components of the framework are:

Hadoop Common - The common module contains the libraries and utilities that are required by the other modules of Hadoop. For example, Hadoop common provides various components and interfaces for distributed file system and general input/output. This includes serialization, Java RPC (Remote Procedure Call) and file-based data structures.

Hadoop Distributed File System (HDFS) - A Java-based distributed file system which can store all kinds of data on the disks at the clusters.

MapReduce vl - Software programming model in Hadoop 1 using Mapper and Reducer. The vl processes large sets of data in parallel and in batches.

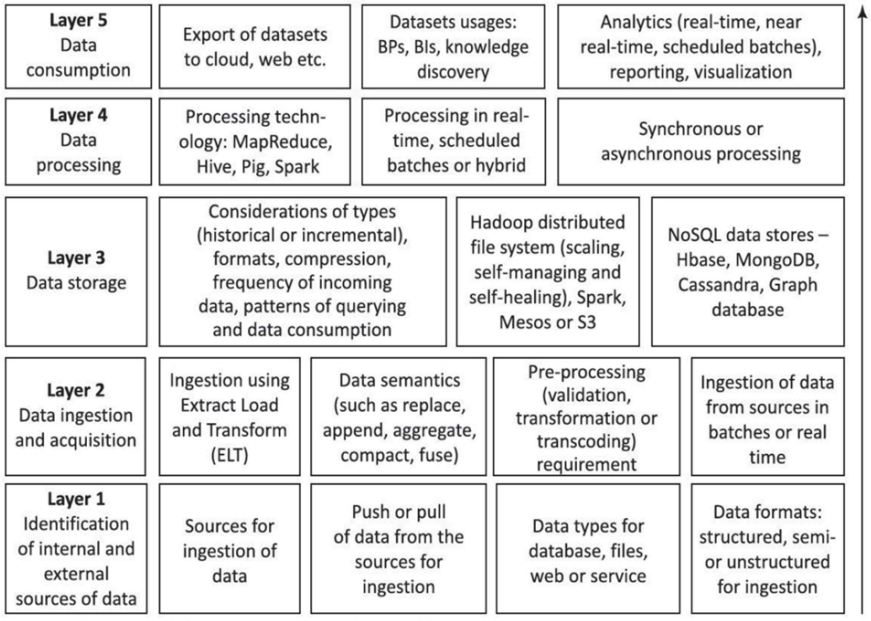
YARN - Software for managing resources for computing. The user application tasks or sub- tasks run in parallel at the Hadoop, uses scheduling and handles the requests for the resources in distributed running of the tasks.

MapReduce v2 - Hadoop 2 YARN-based system for parallel processing of large datasets and distributed processing of the application tasks.

Diagram - 2 Marks

Each component explanation -3 Marks

3 (a)	Discuss the functions of each of the five layers in Big Data architecture design.	[06]	CO 1	L1



- L1 considers the following aspects in a design
    - Amount of data needed at ingestion layer 2 (L2)
    - Push from L1 or pull by L2 as per the mechanism for the usages
    - Source data-types: Database, files, web or service
- Source formats, i.e., semi-structured, unstructured or structured.

Layer 2

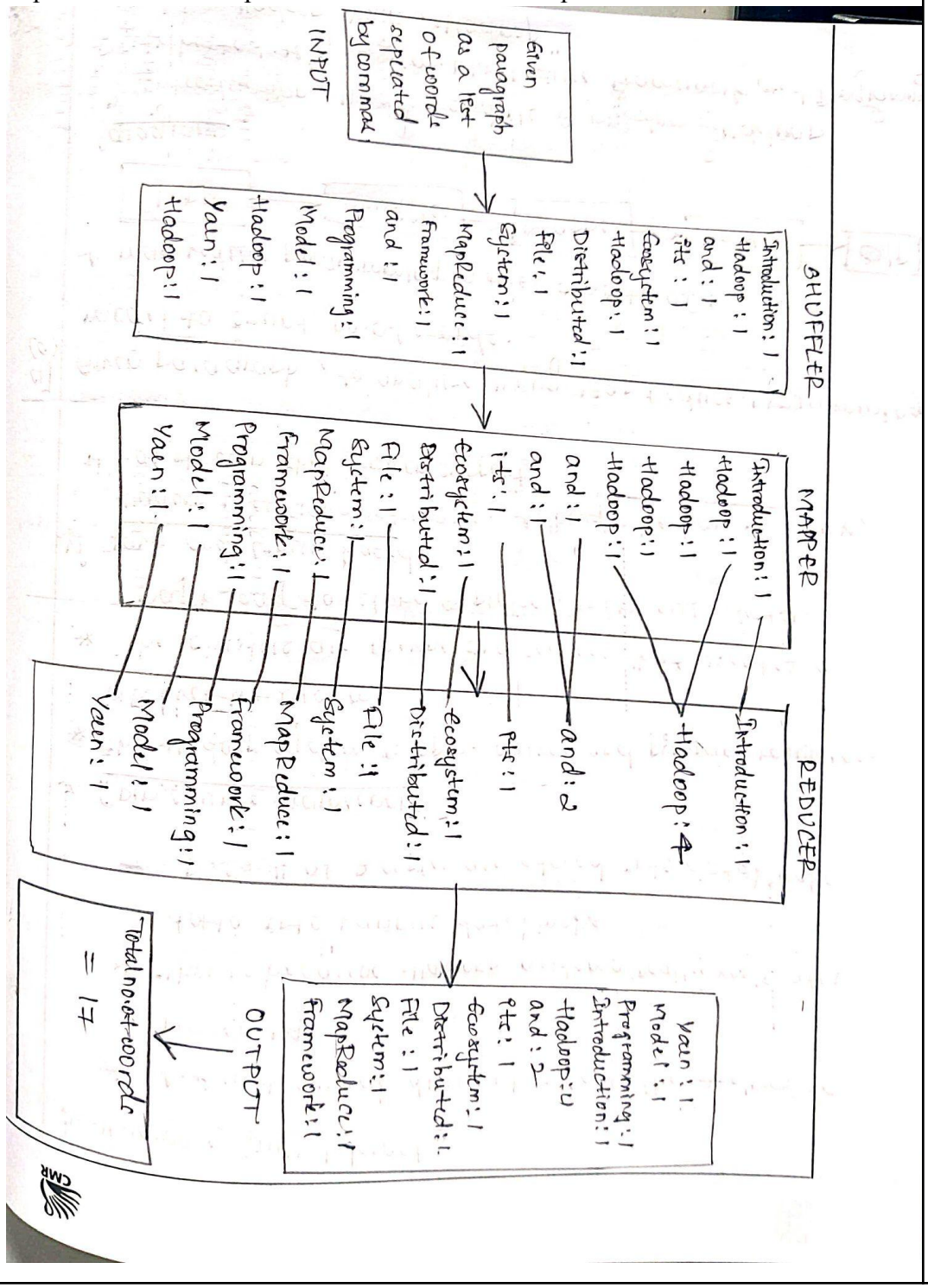
- Ingestion and ETL processes either in real time, which means store and use the data as generated, or in batches.
- Batch processing is using discrete datasets at scheduled or periodic intervals of time.

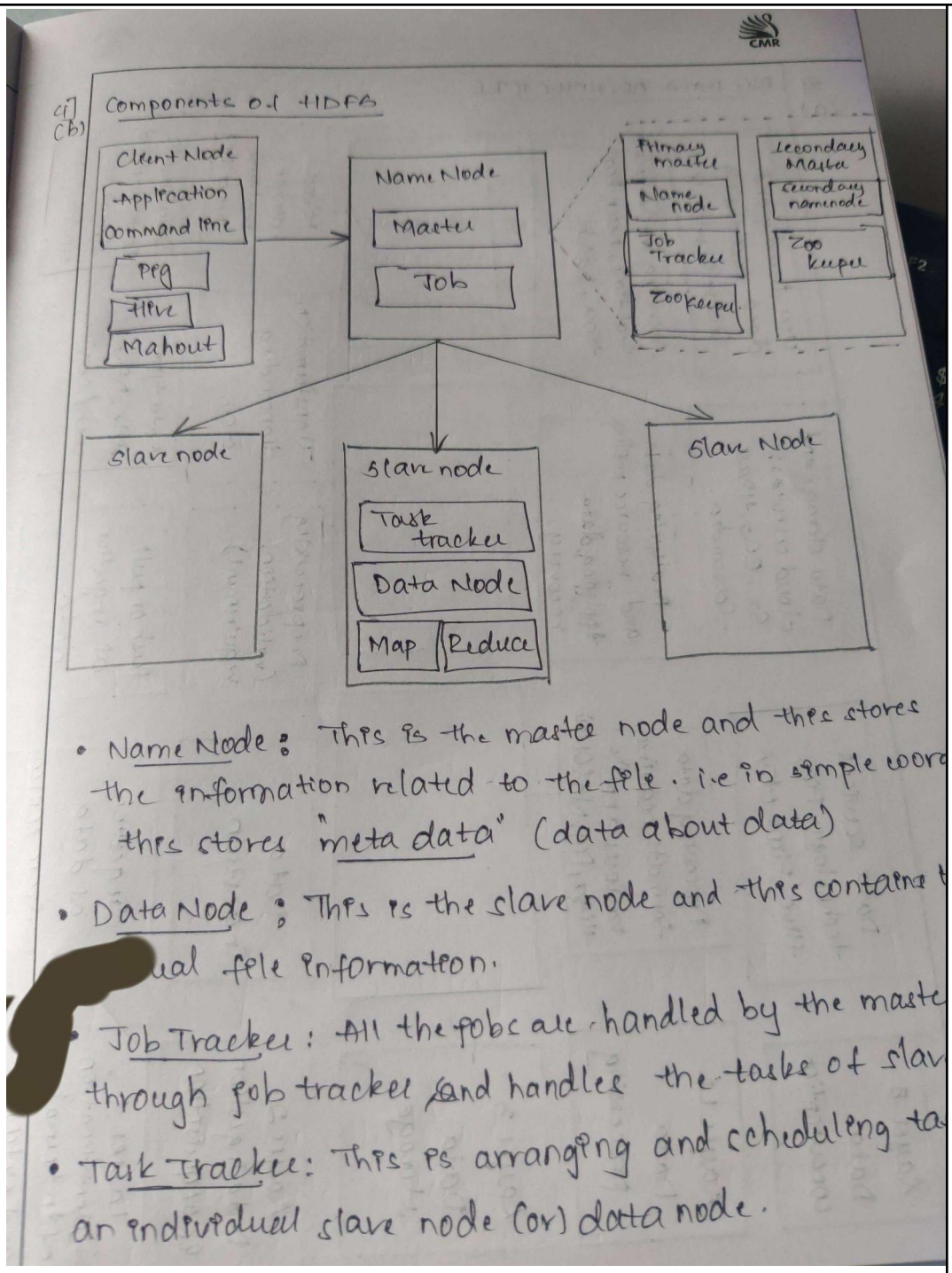
Layer 3

- Data storage type (historical or incremental), format, compression, incoming data
- frequency, querying patterns and consumption requirements for L4 or L5

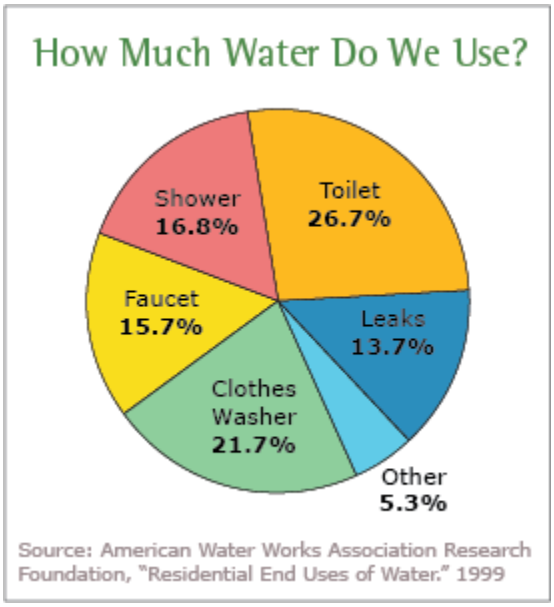
	• Data storage using Hadoop distributed file system or NoSQL data stores—HBase, Cassandra, MongoDB. **Layer 4** • Data processing software such as MapReduce, Hive, Pig, Spark, Spark Mahout, Spark Streaming • Processing in scheduled batches or real time or hybrid • Processing as per synchronous or asynchronous processing requirements at L5. **Layer 5** • Data integration • Datasets usages for reporting and visualization • Analytics (real time, near real time, scheduled batches), BPs, BIs, knowledge discovery • Export of datasets to cloud, web or other systems			
(b)	List and explain benefits of **YARN**. **Scalability:** The scheduler in Resource manager of YARN architecture allows Hadoop to extend and manage thousands of nodes and clusters. **Compatibility:** YARN supports the existing map-reduce applications without disruptions thus making it compatible with Hadoop 1.0 as well. **Cluster Utilization**:Since YARN supports Dynamic utilization of clusters in Hadoop, which enables optimized Cluster Utilization. **Multi-tenancy:** It allows multiple engine access thus giving organizations a benefit of multi-tenancy.	[04]	CO 2	L1
4 (a)	Taking example Paragraph ***" Introduction, Hadoop and its Ecosystem, Hadoop Distributed File System, MapReduce Framework and Programming Model, Hadoop Yarn, Hadoop "*** With diagrams explain ***Map reduce programming model To count no of words. Mapper and reducer diagrams should be drawn .***	[06]	CO 2	L3

(b)	Explain different components of HDFS with examples.	[04]	CO 1	L1



				
5 (a)	List and explain the different Phases in Analytics with examples.	[6]	CO 1	L2

	Analytics has the following phases before deriving the new facts, providing business intelligence and generating new knowledge. 1. *Descriptive analytics* enables deriving the additional value from visualizations andreports 2. *Predictive analytics* is advanced analytics which enables extraction of new facts and knowledge, and then predicts/forecasts 3. *Prescriptive analytics* enable derivation of the additional value and undertake better decisions for new option(s) to maximize the profits 4. *Cognitive analytics* enables derivation of the additional value and undertakes better decision.			
(b)	Explain how ***Visualization and Report*** helps in ***Descriptive analytics*** with examples. Descriptive analytics is the process of using current and historical data to identify trends and relationships. It's sometimes called the simplest form of data analysis because it describes trends and relationships but doesn't dig deeper. Descriptive analytics is relatively accessible and likely something your organization uses daily. Basic statistical software, such as Microsoft Excel or data visualization tools, such as Google Charts and Tableau, can help parse data, identify trends and relationships between variables, and visually display information.	[2+2]	CO 1	L2

	Descriptive analytics is especially useful for communicating change over time and uses trends as a springboard for further analysis to drive decision-making.  ***Pie chart shows 21.7% used water for clothes wash , 13.7% water wasted in leaks , 26.7% used in Toilet , 16.8% used in Shower , 15.7% used in Faucet and the remaining others.***			
6 (a)	List and Explain five applications of NoSQL Database.	[05]	CO2	L1
	1. Data Mining 2. Social Media Networking Sites 3. Software Development 4. Graph Database 5. Column-oriented Database A non-relational database that stores data in non-tabular relations, a NoSQL database management system is a thing of the 21st century. Referring to non-SQL or non-relational databases, a NoSQL Database can			

	store data in both traditional and non-traditional structural languages. Perhaps this is why it is also referred to as 'not only SQL'. Before relational databases, large data used to be stored in database management systems (DBMS) that had a few drawbacks such as functional complications and slower recovery of data.			
(b)	List and explain Characteristics of Hadoop. **1. Fault-efficient scalable, flexible and modular design** which uses simple and modular programming model. The system provides servers at high scalability. The system is scalable by adding new nodes to handle larger data. Hadoop proves very helpful in storing, managing, processing and analyzing Big Data. **2. Robust design of HDFS:** Execution of Big Data applications continues even when an individual server or cluster fails. This is because of Hadoop provisions for backup (due to replications at least three times for each data block) and a data recovery mechanism. HDFS thus has high reliability. **3. Store and process Big Data:** Processes Big Data of 3V characteristics. **4. Distributed clusters computing model with data locality:** Processes Big Data at high speed as the application tasks and subtasks submit to the DataNodes. One can achieve more computing power by increasing the number of computing nodes. The processing splits across multiple DataNodes (servers), and thus fast processing and aggregated results. **5. Hardware fault-tolerant:** A fault does not affect data and application processing. If a node goes down, the other nodes take care of the residue. This is	[05]	CO 1	L1

	due to multiple copies of all data blocks which replicate automatically. Default is three copies of data blocks. **6. Open-source framework:** Open source access and cloud services enable large data store. Hadoop uses a cluster of multiple inexpensive servers or the cloud. **7. Java and Linux based:** Hadoop uses Java interfaces. Hadoop base is Linux but has its own set of shell commands support.			