Scheme of Evaluation
Internal Assessment Test 1 – October 2022

Sub:	Big Data Analytics							**Code:**	18CS72
Date**:**	21/10/2022	Duration:	90mins	Max Marks:	50	**Sem:**	VII	**Branch:**	ISE

Note: Answer Any five full questions.

Question #		**Description**	**Marks Distribution**		**Max Marks**
1	a)	Discuss the evolution of Big Data. Diagram Explanation	 3M 3M	6M	10M
1	b)	Explain the characteristics of Big Data Any 4 characteristics	1M*4	4M	
2	a)	With neat block diagram, explain Data Architecture Design Diagram Explanation of all layers	4M 6M	10M	
3	a)	Compare and contrast between grid computing and cluster computing Any 5 differences	1M*5	5M	10M
3	b)	Compare and contrast SQL and NoSQL databases. Any 5 differences	1M*5	5M	
4	a)	What are core components of Hadoop? Explain in brief its core components. Component names Diagram Explanation	1M 3M 6M	10M	10M

5	a)	Discuss the Hadoop system and ecosystem components in four layers Diagram All four Layer description	4M 6M	10	10M
6	a)	Write short note on Data Preprocessing and apply the following preprocessing steps in the dataset. Fill missing values using mean, median, and mode. After fill the missing values using any method apply Transformation techniques using standardization and normalization. Data Preprocessing Finding missing values Transformation	2M 4M 4M	10M	10M

Scheme Of Evaluation
Internal Assessment Test 1 – Oct 2022

Sub:	Big Data Analytics							Code:	18CS72
Date:	21/10/2022	Duration:	90mins	Max Marks:	50	Sem:	VII	Branch:	ISE

Note: Answer Any full five questions

Q1 a) Discuss the evolution of Big Data.

1.1.1 Need of Big Data

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 increase.

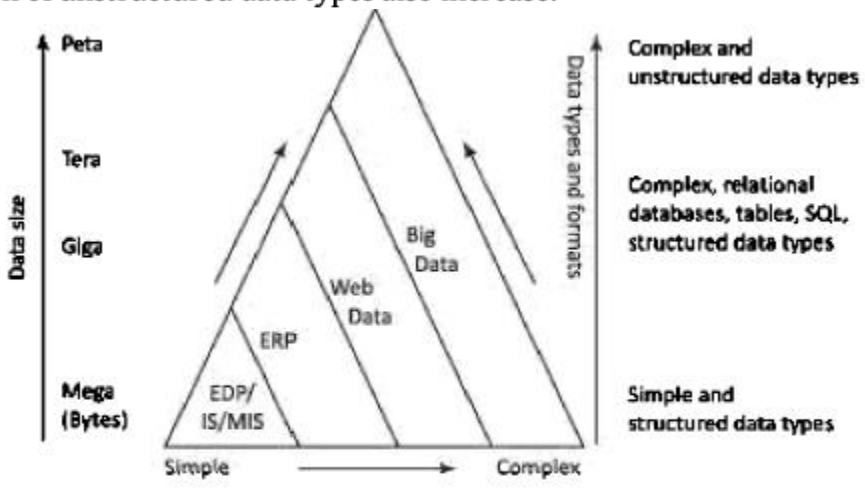


Figure 1.1 Evolution of Big Data and their characteristics

An example of a traditional tool for structured data storage and querying is RDBMS. Volume, velocity and variety (3Vs) of data need the usage of number of programs and tools for analyzing and processing at a very high speed. When integrated with the Internet of Things, sensors and machines data, the veracity of data is an additional V. (Section 1.2.3)

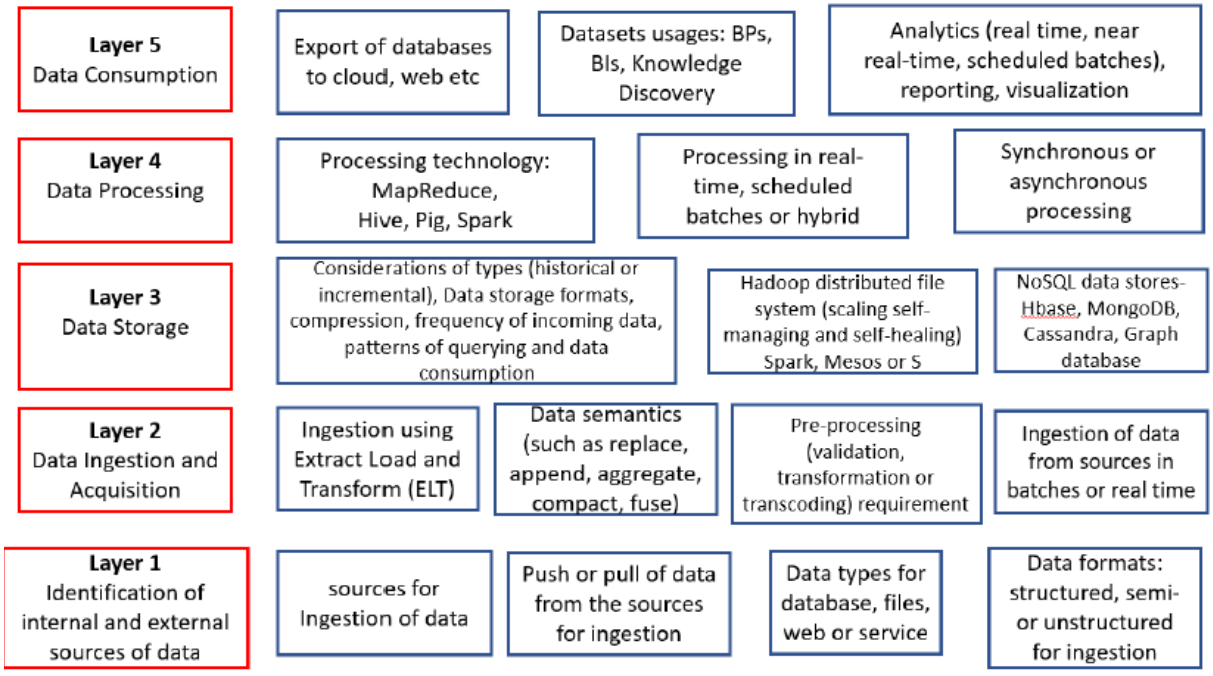
Big Data requires new tools for processing and analysis of a large volume of data. For

example, unstructured, NoSQL (not only SQL) data or Hadoop compatible system data.

Q.1 b) Explain the characteristics of Big Data.

Characteristics of Big Data, called 3Vs (and 4Vs also used) are:
Volume - The phrase Big Data' contains the term big, which is related to size of the data and hence the characteristic. Size defines the amount or quantity of data, which is generated from an application(s).
Velocity - The term velocity refers to the speed of generation of data. Velocity is a measure of how fast the data generates and processes.
Variety - Big Data comprises of a variety of data. Data is generated from multiple sources in a system. This introduces variety in data and therefore introduces 'complexity'. Data consists of various forms and formats.
Veracity – It is also considered an important characteristic to take into account the quality of data captured, which can vary greatly, affecting its accurate analysis.

Q. 2 a) With neat block diagram, explain Data Architecture Design



Data processing architecture consists of five layers:

(i) identification of data sources,

(ii) acquisition, ingestion, extraction, pre-processing, transformation of data,

(iii)data storage at files, servers, cluster or cloud,

(iv)data-processing, and

(v) data consumption

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.

L2 considers the following aspects:

- 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.

L3 considers the followings aspects:

- 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.

L4 considers the followings aspects:

- 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.

L5 considers the consumption of data for the following:

- Data integration

Datasets usages for reporting and visualization Analytics (real time, near real time, scheduled batches), BPs, Bls, knowledge discovery

- Export of datasets to cloud, web or other systems.

Q. 3 a) Compare and contrast between grid computing and cluster computing

Key	Cluster Computing	Grid Computing
Computer Type	Nodes or computers have to be of the same type, like same CPU, same OS. Cluster computing needs a homogeneous network.	Nodes or computers can be of same or different types. Grid computers can have homogeneous or heterogeneous network.
Task	Computers of Cluster Computing are dedicated to a single task and they cannot be used to perform any other task.	Computers of Grid Computing can leverage the unused computing resources to do other tasks.
Location	Computers of Cluster computing are co-located and are connected by high speed network bus cables.	Computers of Grid Computing can be present at different locations and are usually connected by the Internet or a low speed network bus.
Topology	Cluster computing network is prepared using a centralized network topology.	Grid computing network is distributed and have a decentralized network topology.
Task Scheduling	A centralized server controls the scheduling of tasks in cluster computing.	In Grid Computing, multiple servers can exist. Each node behaves independently without the need of any centralized scheduling server.
Resource Manager	Cluster Computing network has a dedicated centralized resource manager, managing the resources of all the nodes connected.	In Grid Computing, each node independently manages its own resources.

Q. 3b) Compare and contrast SQL and NoSQL databases.

Index	SQL	NoSQL
1)	Databases are categorized as Relational Database Management System (RDBMS).	NoSQL databases are categorized as Non-relational or distributed database system.
2)	SQL databases have fixed or static or predefined schema.	NoSQL databases have dynamic schema.
3)	SQL databases display data in form of tables so it is known as table-based database.	NoSQL databases display data as collection of key-value pair, documents, graph databases or wide-column stores.
4)	SQL databases are vertically scalable.	NoSQL databases are horizontally scalable.
5)	SQL databases use a powerful language "Structured Query	In NoSQL databases, collection of documents are used to query the data. It is also called unstructured

	Language" to define and manipulate the data.	query language. It varies from database to database.
6)	SQL databases are best suited for complex queries.	NoSQL databases are not so good for complex queries because these are not as powerful as SQL queries.
7)	SQL databases are not best suited for hierarchical data storage.	NoSQL databases are best suited for hierarchical data storage.
8)	MySQL, Oracle, Sqlite, PostgreSQL and MS-SQL etc. are the example of SQL database.	MongoDB, BigTable, Redis, RavenDB, Cassandra, Hbase, Neo4j, CouchDB etc. are the example of nosql database

Q.4 a) What are core components of Hadoop? Explain in brief its core components.

Hadoop Components:

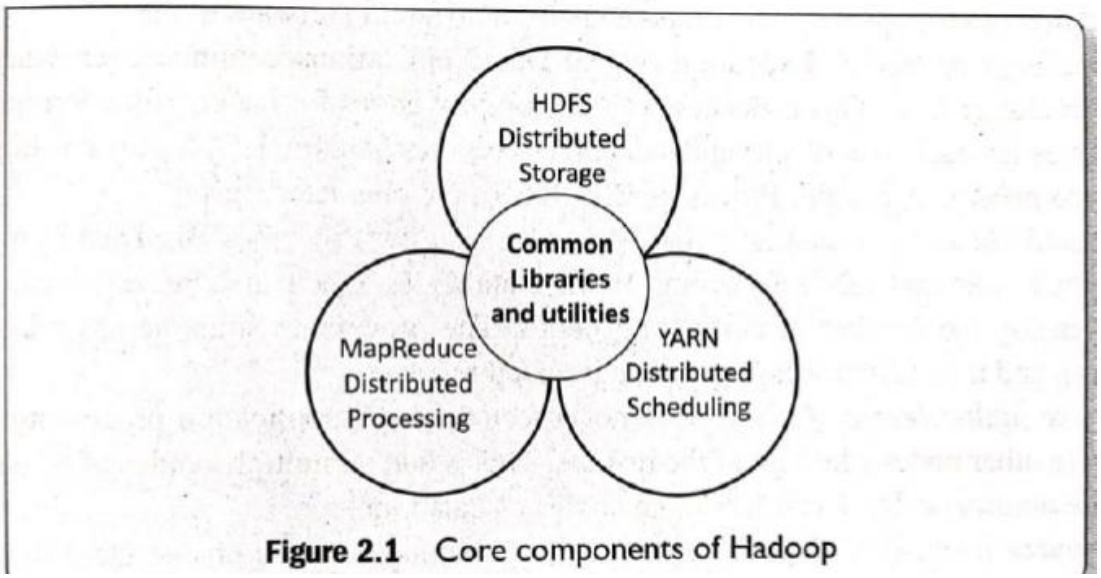


Figure 2.1 Core components of Hadoop

The Hadoop core components of the framework are:

1. 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.
2. Hadoop Distributed File System (HDFS) - A Java-based distributed file system which can store all kinds of data on the disks at the clusters.
3. MapReduce v1 - Software programming model in Hadoop 1 using Mapper and Reducer. The v1 processes large sets of data in parallel and in batches.
4. 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.
5. MapReduce v2 - Hadoop 2 YARN-based system for parallel processing of large datasets distributed processing of the application tasks.

Q. 5 a) Discuss the Hadoop system and ecosystem components in four layers.

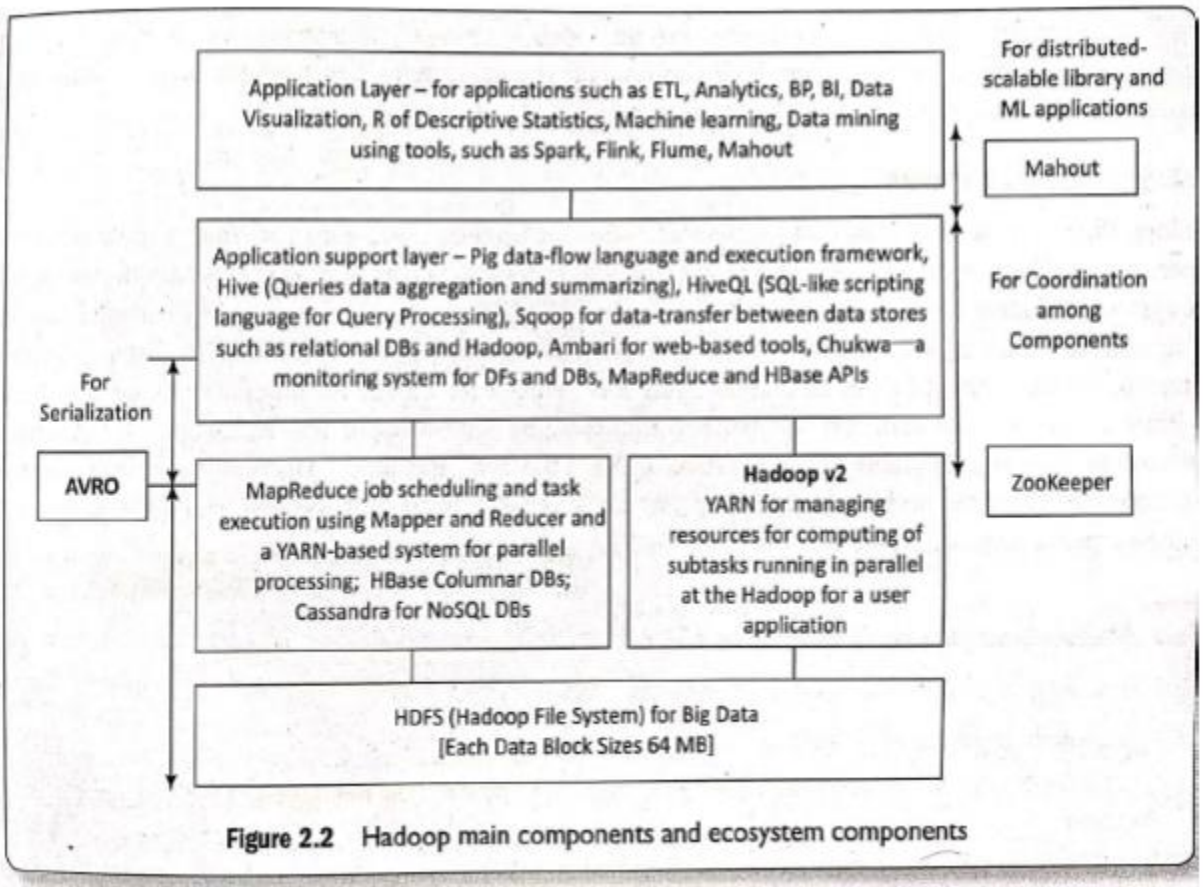


Figure 2.2 Hadoop main components and ecosystem components

Hadoop ecosystem refers to a combination of technologies. Hadoop ecosystem consists of own family of applications which tie up together with the Hadoop. The system components support the storage, processing, access, analysis, governance, security and operations for Big Data.

The system enables the applications which run Big Data and deploy HDFS The data store system consists of clusters, racks, DataNodes and blocks. Hadoop deploys application programming model, such as MapReduce and HBase, YARN manages resources and schedules sub-tasks of the application.

HBase uses columnar databases and does OLAP. Figure 2.2 shows Hadoop core components HDFS, MapReduce and YARN along with the ecosystem. Figure 2.2 also shows Hadoop ecosystem. The system includes the application support layer and application layer components-AVRO, Zookeeper, Pig. Hive, Sqoop, Ambari, Chukwa, Mahout, Spark, Flink and Flume. The figure also shows the components and their usages.

Q. 6 Write short note on Data Preprocessing and apply the following preprocessing steps in the dataset. Fill missing values using mean, median, and mode. After fill the missing values using any method apply Transformation techniques using standardization and normalization.

X1	X2	X3	X4	Y
78.5	67	1	0.2	73.2
78.5	67	0	0.2	69.2
78.5	67	0	0.2	69
78.5		0	0.2	69
75.5	66.5	1	0.2	73.5
75.5	66.5	1	0.4	
75.5	66.5	0	0.3	65.5
75.5	66.5	0	0.2	65.5
75		1		71
75	64	0	0.1	68
75	64	1	0.2	70.5

Solution:

Pre-processing needs are:

(1) Dropping out of range, inconsistent and outlier values

(ii) Filtering unreliable, irrelevant and redundant information

(ii)Data cleaning, editing, reduction and/or wrangling

(iv)Data validation, transformation or transcoding

(v) ELT processing.

Missing Value:

Using Mean: X2= 66.1, 66.1, X4=0.2, Y=69.4

Using Median: X2= 66.5, 66.5, X4=0.2, Y=69.1

Using Mode: X2=66.5, 66.5, X4=0.2, Y=69

Transformation: Input Table

X1	X2	X3	X4	Y
78.5	67	1	0.2	73.2
78.5	67	0	0.2	69.2
78.5	67	0	0.2	69
78.5	**66.1**	0	0.2	69
75.5	66.5	1	0.2	73.5
75.5	66.5	1	0.4	**69.4**
75.5	66.5	0	0.3	65.5
75.5	66.5	0	0.2	65.5
75	**66.1**	1	**0.2**	71
75	64	0	0.1	68
75	64	1	0.2	70.5

Standardization:

$$X' = \frac{X - \mu}{\sigma}$$

X = actual value

μ = mean

σ = standard deviation.

Example:

X=78.5

Mean for column =76.5

Standard Deviation of column =1.6

Standardization=(78.5-76.5)/1.6=**1.3**

Standardized Table:

X1	X2	X3	X4	Y
1.3	0.9	1.1	-0.3	1.5
1.3	0.9	-0.9	-0.3	-0.1
1.3	0.9	-0.9	-0.3	-0.2
1.3	0.0	-0.9	-0.3	-0.2
-0.6	0.4	1.1	-0.3	1.6
-0.6	0.4	1.1	2.5	0.0
-0.6	0.4	-0.9	1.1	-1.6
-0.6	0.4	-0.9	-0.3	-1.6
-0.9	0.0	1.1	-0.3	0.6
-0.9	-2.0	-0.9	-1.7	-0.6
-0.9	-2.0	1.1	-0.3	0.4

Normalization:

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Example:

X = 78.5

Min = 75

Max = 78.5

Normalization = (78.5-75)/(78.5-75) = **1.0**

Normalized Table:

X1	X2	X3	X4	Y
1.0	1.0	1.0	0.3	1.0
1.0	1.0	0.0	0.3	0.5
1.0	1.0	0.0	0.3	0.4
1.0	0.7	0.0	0.3	0.4
0.1	0.8	1.0	0.3	1.0
0.1	0.8	1.0	1.0	0.5
0.1	0.8	0.0	0.7	0.0
0.1	0.8	0.0	0.3	0.0
0.0	0.7	1.0	0.3	0.7
0.0	0.0	0.0	0.0	0.3
0.0	0.0	1.0	0.3	0.6