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

Scheme of Evaluation with solutions
Internal Assessment Test 1 – MAY 2021

Sub:	Big data and Analytics					Sub Code:	18CS72	Branch:	ISE	
Date:	15/11/2021	Duration:	90 min's	Max Marks:	50	Sem / Sec:	VII A,B & C		OBE	

Answer any FIVE FULL Questions		MARKS	CO	RBT
1 (a)	**Write short note on Big Data and Analytics and its applications.** **Scheme:** Big Data Definition and application 3+2 = 5M **Solution:** Big Data is high-volume, high-velocity and/or high-variety information asset that requires new forms of processing for Data enhanced decision making, insight discovery and process optimization. Meanings and various Industry analyst Doug Laney described the "3Vs', i.e. volume, variety and/or velocity. 4Vs', i.e. volume, velocity, variety and veracity". A number of other definitions are available for Big Data, some of which are given below. "A collection of data sets so large or complex that traditional data processing applications are inadequate." "Data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges." "Big Data refers to data sets whose size is beyond the ability of typical database software tool to capture, store, manage and analyze." Applications: • Big Data in Marketing and Sales • Big Data Analytics in Detection of Marketing Frauds • Big Data Risks • Big Data Credit Risk Management • Big Data and Algorithmic Trading • Big Data and Healthcare • Big Data in Medicine • Big Data in Advertising	[05]	CO1	L1
(b)	**Explain the Characteristics and Classification of Big Data with example.** **Scheme:** Characteristics and Classification 3+2 = 5M **Solution:** 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.	[05]	CO1	L2

	Data can be classified as structured, semi-structured, multi-structured and unstructured. • Structured data It conforms and associate with data schemas and data models. Structured data are found in tables (rows and columns). Nearly 15-20% data are in structured or semi-structured form. Unstructured data do not conform and associate with any data models. • Semi-Structured Data Examples of semi-structured data are XML and JSON documents. Semi-structured data contain tags or other markers, which separate semantic elements and enforce hierarchies of records and fields within the data. • Multi-Structured Data Multi-structured data refers to data consisting of multiple formats of data, viz. structured, semi-structured and/or unstructured data. Multi-structured data sets can have many formats. • Unstructured data It does not possess data features such as a table or a database. Unstructured data are found in file types such as .TXT, .CSv. Data may be as key-value pairs, such as hash key-value pairs. Data may have internal structures, such as in e-mails.			
2	**Explain Scalability and Parallel Processing with example.** **Scheme**: Scalability =3M Parallel Processing and example= 7M **Solution:** Scalability enables increase or decrease in the capacity of data storage, processing and analytics. Scalability is the capability of a system to handle the workload as per the magnitude of the work. System capability needs increment with the increased workloads. When the workload and complexity exceed the system capacity, scale it up and scale it out. • Analytics Scalability to Big Data *Vertical scalability* means scaling up the given system's resources and increasing the system's analytics, reporting and visualization capabilities. This is an additional way to solve problems of greater complexities. *Scaling up* means designing the algorithm according to the architecture that uses resources efficiently. *Horizontal scalability* means increasing the number of systems working in coherence and scaling out the workload. Processing different datasets of a large dataset deploys horizontal scalability. *Scaling out* means using more resources and distributing the processing and storage tasks in parallel. Alternative ways for scaling up and out processing of analytics software and Big Data analytics deploy the Massively Parallel Processing Platforms (MPPS), cloud, grid, clusters, and distributed computing software. • Massively Parallel Processing Platforms Scaling uses parallel processing systems. Many programs are so large and/or complex that it is impractical or impossible to parallel and distributed execute them on a single computer system, especially in limited computer memory. Here, it is required to enhance (scale) up the computer system or use massive parallel processing (MPPS) platforms. Parallelization of tasks can be done at several levels: (i) distributing separate tasks onto separate threads on the same CPU, (ii) distributing separate tasks onto separate CPUS on the same computer	[10]	CO1	L2

(iii)distributing separate tasks onto separate computers.

The computational problem is broken into discrete pieces of sub-tasks that can be processed simultaneously. The system executes multiple program instructions or sub-tasks at any moment in time. Total time taken will be much less than with a single compute resource.

- Distributed Computing Model

A distributed computing model uses cloud, grid or clusters, which process and analyze big and large datasets on distributed computing nodes connected by high-speed networks.

- Cloud Computing

Cloud computing is a type of Internet-based computing that provides shared processing resources and data to the computers and other devices on demand. One of the best approach for data processing is to perform parallel and distributed computing in a cloud-computing environment. Cloud usages circumvent the single point failure due to failing of one node. Cloud design performs as a whole. Its multiple nodes perform automatically and interchangeably. It offers high data security compared to other distributed technologies.

Cloud resources can be Amazon Web Service (AWS) Elastic Compute Cloud (EC2), Microsoft Azure or Apache CloudStack and Amazon Simple Storage Service (S3).

Cloud computing features are:

(i) on-demand service

(ii) resource pooling,

(iii)scalability,

(iv)accountability,

(v) broad network access.

Cloud services can be accessed from anywhere and at any time through the Internet. A local private cloud can also be set up on a local cluster of computers. Cloud Services are: Infrastructure as a Service (laas), Platform as a Service (Paas) and Software as a Service (Saas).

- Grid and Cluster Computing

Grid Computing refers to distributed computing, in which a group of computers from several locations are connected with each other to achieve a common task. The computer resources are heterogeneously and geographically dispersed. A group of computers that might spread over remotely comprise a grid. A grid is used for a variety of purposes. A single grid of course, dedicates at an instance to a particular application only.

Cloud computing depends on sharing of resources (for example, networks, servers, storage, applications and services) to attain coordination and coherence among resources similar to grid computing. Similarly, grid also forms a distributed network for resource integration.

Drawbacks of Grid Computing Grid computing is the single point, which leads to failure in case of underperformance or failure of any of the participating nodes. A system's storage capacity varies with the number of users, instances and the amount of data transferred at a given time. Sharing resources among a large number of users helps in reducing infrastructure costs and raising load capacities.

- Cluster Computing

A cluster is a group of computers connected by a network. The group works together to accomplish the same task. Clusters are used mainly for load balancing. They shift processes between nodes to keep an even load on the group of connected computers.

- Volunteer Computing

Volunteers provide computing resources to projects of importance that use resources to do distributed computing and/or storage. Volunteer computing is a distributed computing paradigm which uses computing resources of the volunteers. Volunteers are organizations or members who own personal computers. Projects examples are science-related projects executed by universities or academia in general.

3 **Write short note on Data Preprocessing and apply the following preprocessi missing values using mean, median, and mode. After fill the missing value Transformation techniques using standardization and normalization.** [10] CO1 L3 **se d**

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

Scheme:
Data Preprocessing =2M Missing Value = 3M Transformation = 5M

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:

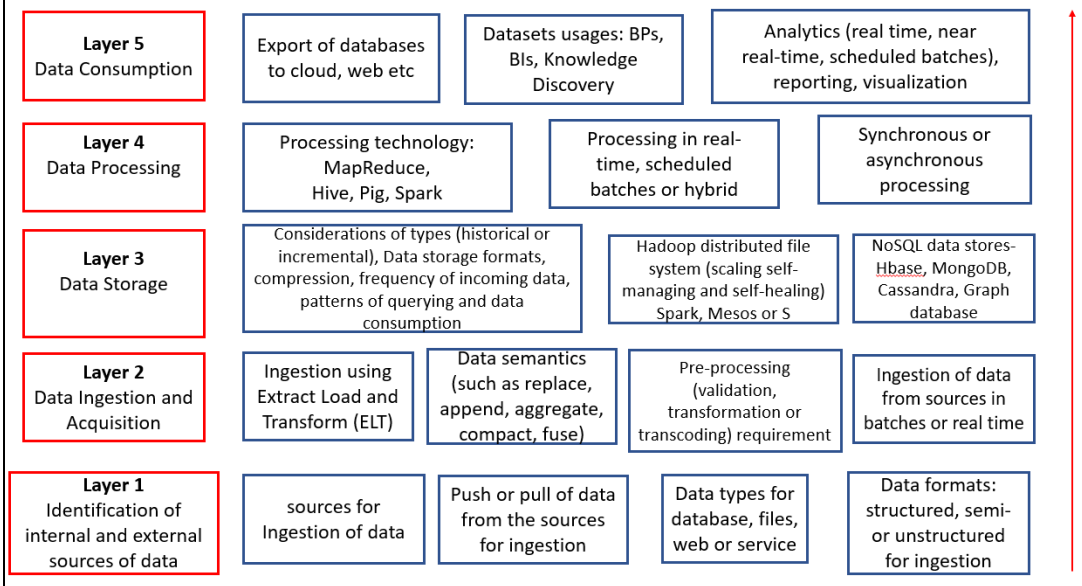
$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

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

4	**What is designing Data Architecture? Explain with neat diagram.** **Scheme:** Diagram and Explanation 2*5 =10M **Solution:**	[10]	CO1	L2



Data processing architecture consists of five layers:

(i) identification of data sources,

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

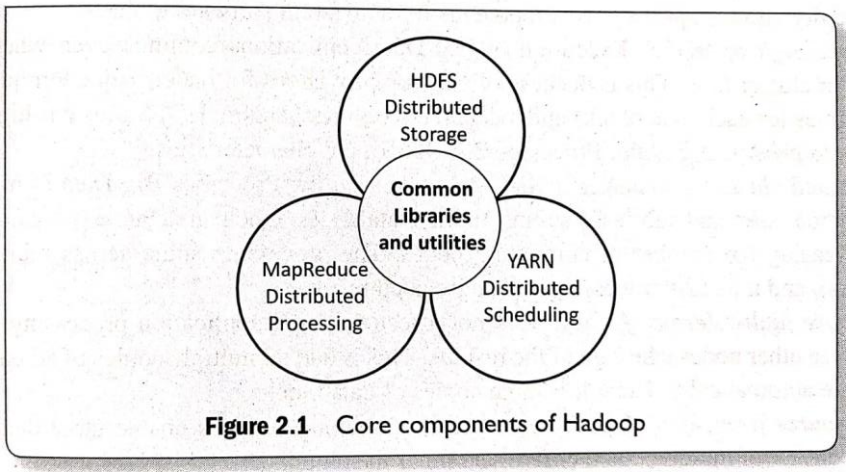
(iii)data storage at files, servers, cluster or cloud,

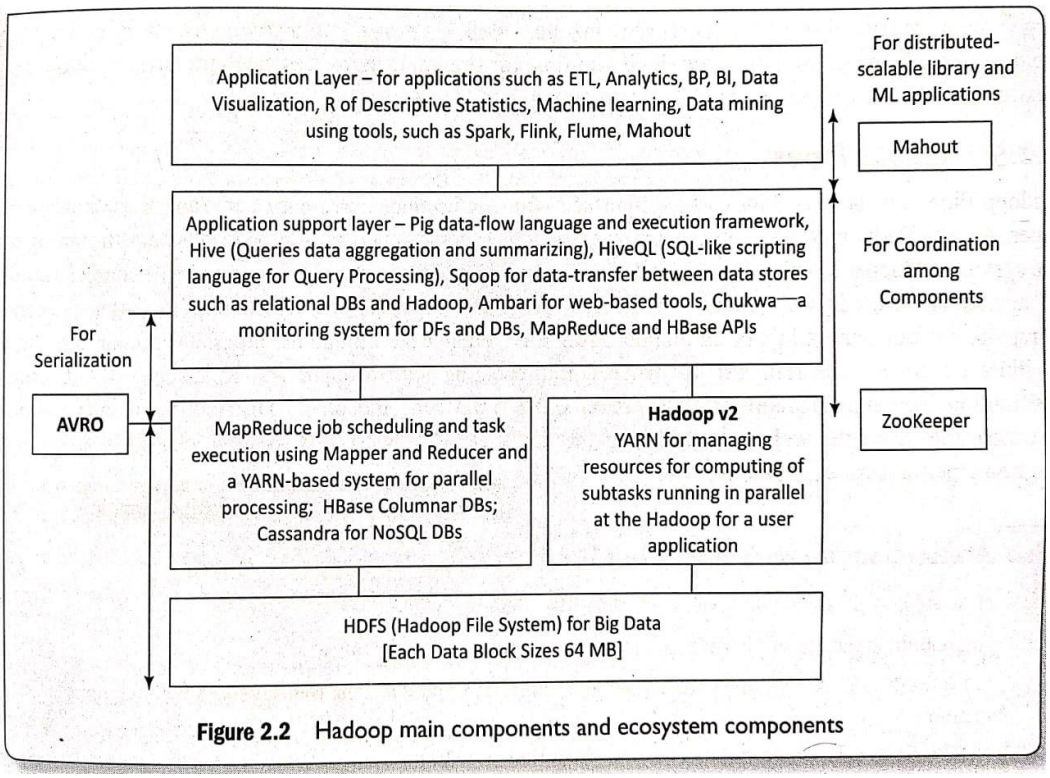
(iv)data-processing, and

(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.			
5 (a)	**Write short note on Hadoop.** **Scheme:** Hadoop Explanation 5M **Solution:** Apache initiated the project for developing storage and processing framework for Big Data storage and processing. Doug Cutting and Machael J. Cafarelle the creators named that framework as Hadoop. Cutting's son was fascinated by a stuffed toy elephant, named Hadoop, and this is how the name Hadoop was derived. The project consisted of two components, one of them is for data store in blocks in the clusters and the other is computations at each individual cluster in parallel with another. Hadoop components are written in Java with part of native code in C. The command line utilities are written in shell scripts. Hadoop is a computing environment in which input data stores, processes and stores the results. The environment consists of clusters which distribute at the cloud or set of servers. Each cluster consists of a string of data files constituting data blocks. The toy named Hadoop consisted of a stuffed elephant. The Hadoop system cluster stuffs files in data blocks. The complete system consists of a scalable distributed set of clusters. Infrastructure consists of cloud for clusters. A cluster consists of sets of computers or PCs. The Hadoop platform provides a low cost Big Data platform, which is open	[05]	CO2	L2

	source and uses cloud services. Tera Bytes of data processing takes just few minutes. Hadoop enables distributed processing of large datasets (above 10 million bytes) across clusters computers using a programming model called MapReduce. The system characteristics are scalable, self-manageable, self-healing and distributed file system. Scalable means can be scaled up (enhanced) by adding storage and processing units as per the requirements Self-manageable means creation of storage and processing resources which are used, scheduled and reduced or increased with the help of the system itself. Self-healing means that in case of faults, they are taken care of by the system itself. Self-healing enables functioning and resources availability, Software detect and handle failures at the task level. Software enable the service or task execution even in case of communication of node failure.			
(b)	**Explain the Features and Components of Hadoop with neat diagram.** **Scheme:** Features = 2M Components = 3M **Solution:** **Hadoop features** are as follows: • Fault-efficient scalable, flexible and modular design • Robust design of HDFS • Store and process Big Data • Distributed clusters computing model with data locality • Hardware fault-tolerant • Open-source framework • Java Linux based **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.	[05]	CO2	L2

	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.			
5	**Explain the Hadoop Ecosystem with neat diagram.** **Scheme:** Diagram =5M Explanation = 5M **Solution:**  **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.	[10]	CO2	L2