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

Internal Assessment Test 5 – February 2022

Sub:	BIG DATA AND ANALYTICS					Sub Code:	18CS72	Branch:	ISE
Date:	05/02/2022	Duration:	90 min's	Max Marks:	50	Sem / Sec:	VII / A, B & C		OBE

	Answer any FIVE FULL Questions	MARKS	CO	RBT
1	**Explain in detail about HDFS design features with suitable diagram.** **Scheme:** Design Features + Diagram 6+4 = 10M **Solution:** • Google File System (GFS) • Hadoop Distributed File System (HDFS) • HDFS block size is typically 64MB or 128MB Big Data analytics applications are software applications that leverage large scale data. The applications analyze Big Data using massive parallel processing frameworks HDFS is a core component of Hadoop. HDFS is designed to run on a cluster of computers and servers at cloud-based utility services. HDFS stores Big Data which may range from GBs to PBs. HDFS stores the data in a distributed manner in order to compute fast. The distributed data store in HDFS stores data in any format regardless of schema HDFS provides high throughput access to data-centric applications that require large-scale data processing workloads. HDFS Data Storage Hadoop data store concept implies storing the data at a number of chatters. Each cluster has a number of data stores, called racks. Each rack stores a number of DataNodes. Each DataNode has a large number of data blocks. The racks distribute across a cluster. The nodes have processing and storage capabilities. The nodes have the data in data blocks to run the application tasks. The data blocks replicate by default at least on three DataNodes in same or remote nodes. Data at the stores enable running the distributed applications including analytics, data mining, OLAP using the clusters. A file, containing the data divides into data blocks. A data block default size is 64 MBs (HDFS division of files concept is similar to Linux or virtual memory page Intel x86 and Pentium processors where the block size is fixed and is of 4 KB). Hadoop HDFS features are as follows: (i) Create, append, delete, rename and attribute modification functions (ii) Content of individual file cannot be modified or replaced but appended with new data at the end of the file. (iii) Write once but read many times during usages and processing (iv) Average file size can be more than 500 MB. The following is an example how the files store at a Hadoop cluster. Hadoop Physical Organization The conventional file system uses directories. A directory consists of folders. A folder consists of files. When data processes, the data sources identify by pointers for the	[10]	CO2	L2

resources. A data-dictionary stores the resource pointers. Master tables at the dictionary store at a central location. The centrally stored tables enable administration easier when the data sources change during processing. Similarly, the files, DataNodes and blocks need the identification during processing at HDFS. HDFS use the NameNodes and DataNodes. A NameNode stores the file's meta data. Meta data gives information about the file of user application, but does not participate in the computations. The DataNode stores the actual data files in the data blocks.

Few nodes in a Hadoop cluster act as NameNodes. These nodes are termed as MasterNodes or simply masters. The masters have a different configuration supporting high DRAM and processing power. The masters have much less local storage. Majority of the nodes in Hadoop cluster act as DataNodes and Task Trackers. These nodes are referred to as slave nodes or slaves. The slaves have lots of disk storage and moderate amounts of processing capabilities and DRAM. Slaves are responsible to store the data and process the computation tasks submitted by the clients.

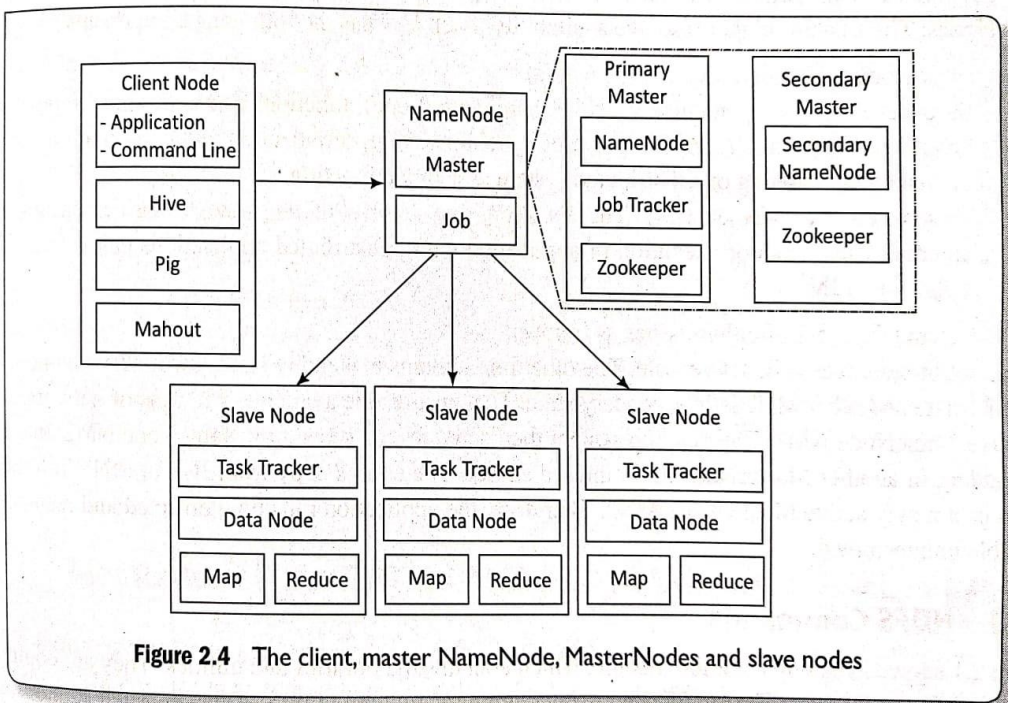


Figure 2.4 The client, master NameNode, MasterNodes and slave nodes

Figure 2.4 shows the client, master NameNode, primary and secondary MasterNodes and slave nodes in the Hadoop physical architecture. Clients as the users run the application with the help of Hadoop ecosystem projects. For example. Hive, Mahout and Pig are the ecosystem's projects. They are not required to be present at the Hadoop cluster. A single MasterNode provides HDFS, MapReduce and Hbase using threads in small to medium sized clusters. When the cluster size is large, multiple servers are used, such as to balance the load. The secondary NameNode provides NameNode management services and Zookeeper is used by HBase for metadata storage.

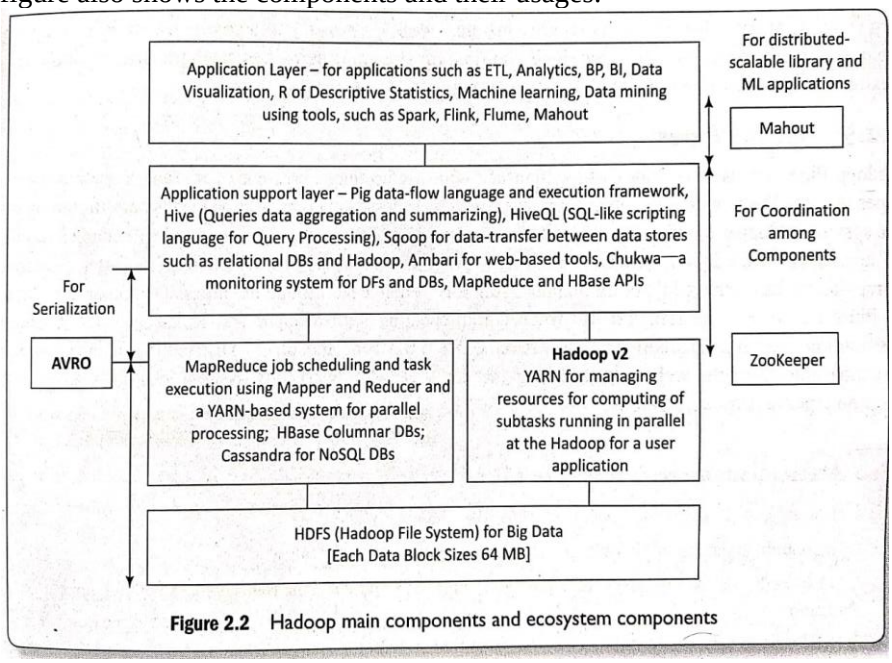
The MasterNode fundamentally plays the role of a coordinator. The MasterNode receives client connections, maintains the description of the global file system namespace, and the allocation of file blocks. It also monitors the state of the system in order to detect any failure. The Masters consists of three components NameNode, Secondary NameNode and JobTracker. The NameNode stores all the file system related information such as:

- The file section is stored in which part of the cluster
- Last access time for the files

	• User permissions like which user has access to the file. Secondary Name Node is an alternate for NameNode. Secondary node keeps a copy of NameNode meta data. Thus, stored meta data can be rebuilt easily, in case of NameNode failure. The JobTracker coordinates the parallel processing of data. Masters and slaves, and Hadoop client (node) load the data into cluster, submit the processing job and then retrieve the data to see the response after the job completion.			

2	**Write elaborately the Commands of HDFS with example.**
	Scheme:
	Commands of HDFS = 10M
	Solution:
	General HDFS Commands
	• hdfs version
	• hdfs dfs
	Generic options supported are
	• -conf <configuration file> specify an application configuration file
	• -D <property=value> use value for given property
	• -fs <locallnamenode:port> specify a namenode
	• -jt <localiresourcemanagenport> specify a ResourceManager
	• -files <comma separated list of files> specify comma separated
	• files to be copied to the map reduce cluster
	• -libjars <comma separated list of jars> specify comma separated jar files to include in the classpath.
	• -archives <comma separated list of archives> specify comma separated archives to be unarchived on the compute machines.
	List Files in HDFS
	To list the files in the root HDFS directory
	$ hdfs dfs -Is /
	To list files in your home directory
	$ hdfs dfs -Is
	The same result can be obtained by issuing the following command
	hdfs dfs -Is /user/hdfs
	Make a Directory in HDFS
	To make a directory in HDFS
	hdfs dfs -mkdir stuff
	Copy Files to HDFS
	To copy a file from your current local directory into HDFS
	hdfs dfs -put test stuff
	Copy Files from HDFS
	Files can be copied back to your local file system
	hdfs dfs -get stuff/test test-local
	Copy Files within HDFS
	The following command will copy a file in HDFS
	hdfs dfs -cp stuff/test test.hclis
	Delete a File within HDFS
	To delete the HDFS file test.dhfs that was created previously
	$ hdfs dfs -rm test.hdfs
	Moved:'hdfs://limulus:8020/user/hdfs/stuff/tesr
	totrashathdfs://limulus:8020/user/hdfs/.T
	rash/CurrentNote that when the fs.trash.interval option is set to a non-zero value in core-site.xml, all deleted files are moved to the user's .Trash directory. This can be avoided by including the - skipTrash option.
	$ hdfs dfs -rm -skipTrash stuff/test
		[10]	CO2	L2

	Deleted stuff/test
	Delete a Directory in HDFS
	To delete the HDFS directory stuff and all its contents
	$ hdfs dfs -rm -r -skipTrash stuff
	Deleted stuff
	Get an HDFS Status Report
	users can get an abbreviated HDFS status report using the following command
	$ hdfs dfsadmin -report			

3	**Write short notes on Hadoop Ecosystem Tools. Explain Sqoop and Oozie with neat diagram.** **Scheme:** Hadoop Ecosystem Tools Sqoop + Oozie 6+4 = 10M **Solution:** 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.  **Figure 2.2** Hadoop main components and ecosystem components **Sqoop** The loading of data into Hadoop clusters becomes an important task during data analytics. Apache a tool that is built for loading efficiently the voluminous amount of data between Hadoop and external data repositories that resides on enterprise application servers or relational databases Sqoop works with relational databases such as Oracle, MySQL, PostgreSQL and DB2.	[10]	CO2	L2

	Sqoop provides the mechanism to import data from external Data Stores into HDPS, Sqoop relates to Hadoop eco-system components, such as Hive and HBase Sqoop can extract data from Hadoop or other ecosystem components. Sqoop provides command line interface to its users. Sqoop can also be accessed using Java API The tool allows defining the schema of the data for import. Sqoop exploits MapReduce framework to import and export the data, and transfers for parallel processing of sub-tasks. Sqoop provisions for fault tolerance. Parallel transfer of data results in parallel results and fast data transfer. Sqoop initially parses the arguments passed in the command line and prepares the map task. The map task initializes multiple Mappers depending on the number supplied by the user in the command line. Each map task will be assigned with part of data to be imported based on key defined in the command line. Sqoop distributes the input data equally among the Mappers. Then each Mapper creates a connection with the database using JDBC and fetches the part of data assigned by Sqoop and writes it into HDFS/Hive/HBase as per the choice provided in the command line. **Oozie** Apache Oozie is an open-source project of Apache that schedules Hadoop jobs. An efficient process for job handling is required. Analysis of Big Data requires creation of multiple jobs and sub-tasks in a process. Oozie design provisions the scalable processing of multiple jobs. Thus, Oozie provides a way to package and bundle multiple coordinator and workflow jobs, and manage the lifecycle of those jobs. The two basic Oozie functions are: • Oozie workflow jobs are represented as Directed Acrylic Graphs (DAGs), specifying a sequence of actions to execute. • Oozie coordinator jobs are recurrent Oozie workflow jobs that are triggered by time and data of availability. Oozie provisions for the following: 1. Integrates multiple jobs in a sequential manner 2. Stores and supports Hadoop jobs for MapReduce, Hive, Pig, and Sqoop 3. Runs workflow jobs based on time and data triggers 4. Manages batch coordinator for the applications 5. Manages the timely execution of tens of elementary jobs lying in thousands of workflows in a Hadoop cluster			
4	**Explain about Outliers, Variances, Probability Distributions and Correlations in detail.** **Scheme:** Outlier + Variance + PD + Correlation =2+2+3+3 = 10M **Solution:** **Variance** • A Random Variable is a variable whose possible values are outcomes of unpredictable processes to numerical quantities. • Variance measures by the sum of squares of the difference in values of a variable with respect to the expected value. • It indicates how widely data pointes in a dataset vary. • If data points vary greatly from the mean value in a dataset, the variance is large. Otherwise, the variance is less. $$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N}$$ **Probabilistic Distribution of Variables:**	[10]	CO5	L2

- Probability is the chance of observing a dependent variable value with respect to some independent variable.
- Probability Distribution is the distribution of P values as a function of all possible independent values, variables, situations, distances.

P is given by a function P(x), then P varies as x changes. Variations in P(x) with x can be discrete or continuous. The values of P are normalized such that sum of all P values is 1. Assuming distribution is around the expected value $\bar{x}$, the standard normal distribution formula is:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\bar{x})^2}{2\sigma^2}} \quad (6.3)$$

Normal distribution relates to Gaussian function. Figure 6.3 shows a PDF with normal distribution around x = $\underline{x}$ standard deviation = s and variance = s^2.

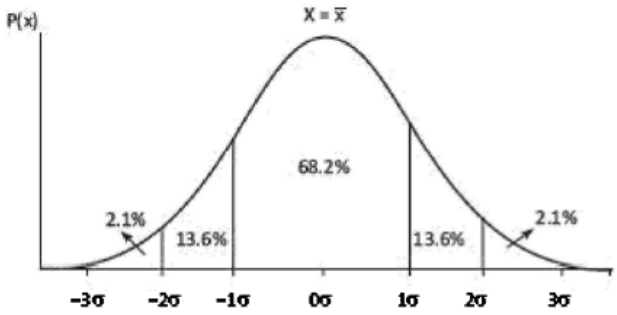
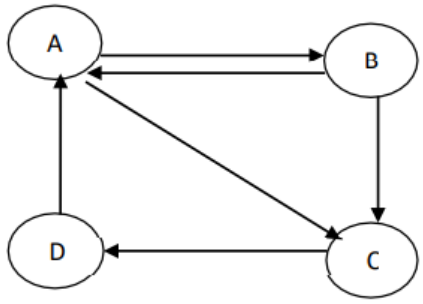


Figure 6.3 Probability distribution function as a function of x assuming normal distribution around $x = \bar{x}$, and standard deviation = s

Correlation is a statistical technique that measures and describes the 'strength' and 'direction' of the relationship between two variables.

Correlation means analysis which lets us find the association or the absence of the relationship between two variables, x and y. Correlation gives the strength of the relationship between the model and the dependent variable on a convenient 0-100% scale.

R-Square R is a measure of correlation between the predicted values y and the observed values of x. *R-squared* (R^2) is a goodness-of-fit measure in linear-regression model. It is also known as the coefficient of determination. R^2 is the square of R, the coefficient of multiple correlations, and includes additional independent (explanatory) variables in regression equation.

5	**Compute the Rank values for the nodes for the following network. Which the highest rank node after computation?**  **Scheme:** Solution for Problem = 10M	[10]	CO5	L3

Solution :

a) Compute the Influence matrix (rank matrix)

- Assign the variables for influence value for each node, as Ra, Rb, Rc, Rd.
- There are two bound links from node A to nodes B and C. Thus, both B and C receives half of node A's influence. Similarly, there are two outbound links from node B to nodes C and A, So both C and A received half of node B's influence.

$$Ra = 0.5*Rb + Rd$$
$$Rb = 0.5*Ra$$
$$Rc = 0.5*Ra + 0.5*Rb$$
$$Rd = Rc$$

b) Set the initial set of rank values such as 1/n (n is number of nodes). As 4 nodes are there, initial rank values for all nodes are ¼ i.e 0.25

Variables	Initial Values
Ra	0.25
Rb	0.25
Rc	0.25
Rd	0.25

c) Compute the rank values for 1st iteration and theniteratively compute new rank values till they stabilized.

Variables	Initial Values	Iteration 1
Ra	0.25	0.375
Rb	0.25	0.125
Rc	0.25	0.250
Rd	0.25	0.250

Variables	Initial Values	Iteration 1	Iteration 2
Ra	0.25	0.375	0.3125
Rb	0.25	0.125	0.1875
Rc	0.25	0.250	0.250
Rd	0.25	0.250	0.250

Variables	Initial Values	Iteration 1	Iteration 2	--------	Iteration 8
Ra	0.25	0.375	0.3125	…….	0.333
Rb	0.25	0.125	0.1875	……	0.167
Rc	0.25	0.250	0.250	……	0.250
Rd	0.25	0.250	0.250	…….	0.250

The Final rank shows of node A is highest at 0.333

6	**Explain in detail about Social Network Analytics.** **Scheme:** Explanation of SNA = 10M **Solution:** A social network is a social structure made of individuals (or organizations) called "nodes," which are tied (connected) by one or more specific types of inter-dependency, such as friendship, kinship, financial exchange, dislike or relationships of beliefs, knowledge or prestige. **Social Network as Graphs**	[10]	CO5	L2

	Social network as graphs provide a number of metrics for analysis. The metrics enable the application of the graphs in a number of fields. Network topological analysis tools compute the degree, closeness, betweenness, egonet, K-neighbourhood, top-K shortest paths, PageRank, clustering, SimRank **Centralities, Ranking and Anomaly Detection** Important metrics are degree (centrality), closeness (centrality), betweenness (centrality) and eigenvector (centrality). Eigenvector consists of elements such as status, rank and other properties. Social graph-network analytics discovers the degree of interactions, closeness, betweenness, ranks, probabilities, beliefs and potentials. Social network characteristics from observations in the organizations are as follows: 1. Three-step neighbourhoods show positive correlation between a person and high performance. Betweenness between vertices and bridges between numbers of structures are not helpful to the organization. Too many strong links of a person may have a negative correlation with the performance. 2. Social network of a person shows high performance outcome when the network exhibits structural diversity. Person with a social network with an abundant number of structural holes exhibits higher performance. This is because having diverse relations help an organization. Social network analysis enables detection of an anomaly. An example is detection of one dominant edge which other sub-graphs are follow (succeed). Ego network is another example. The network structure is such that a given vertex corresponds to a sub-graph where only its adjacent neighbours and their mutual links are included.			

Faculty Signature CCI Signature HOD Signature