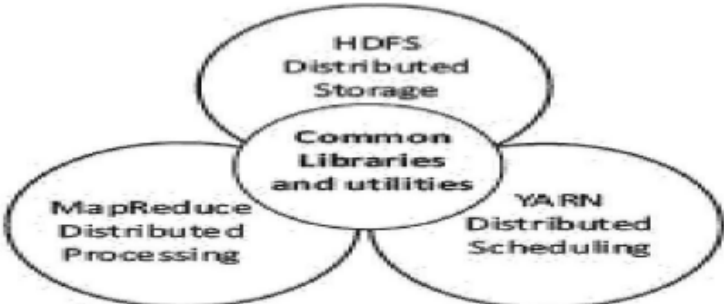


	<p>Fusing of existing data at an enterprise data warehouse with data from sources such as social media, websites, blogs, e-mails, and thus enriching existing data</p> <p>Using multiple sources of data and connecting with many applications</p> <p>Providing greater insights using querying of the multiple source data</p> <p>Analyzing data which enable structured reports and visualization</p> <p>Providing high volume data mining, new innovative applications and thus leading to new business intelligence and knowledge discovery</p>			
<p>2 (a)</p>	<p>List and explain with examples different Phases of Big-Data Analytics.</p> <p>List - 2Marks Explanation & Example -3 Marks Phases in analytics Analytics has the following phases before deriving the new facts, providing business intelligence and generating new knowledge.</p> <ol style="list-style-type: none"> 1. Descriptive analytics enables deriving the additional value from visualizations <p>SUNIL G L, A.P, DEPT. OF CSE, SVIT , BENGALURU 31 31 Big Data Analytics (18CS72)</p> <p>and reports</p> <ol style="list-style-type: none"> 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 undertake better decision. 	<p>[05]</p>	<p>CO 1</p>	<p>L3</p>
<p>(b)</p>	<p>Explain with a diagram the core components of the Hadoop ecosystem.</p> <p>Diagram - 2 Marks Explanation - 3 Marks</p>  <p>Figure 2.1 Core components of Hadoop</p> <p>Hadoop Common: contains the libraries and utilities that are required For example,</p>	<p>[05]</p>	<p>CO 1</p>	<p>L1</p>

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.

3 (a) Discuss the functions of each of the five layers in Big Data architecture design.

Diagram - 2 Marks

Explanation - 3 Marks

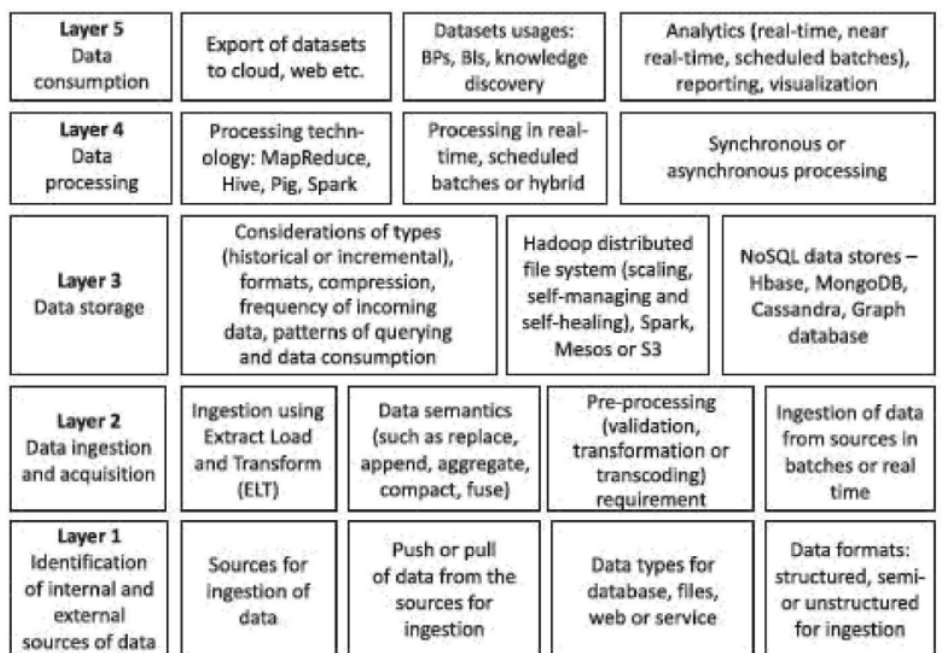


Figure 1.2 Design of logical layers in a data processing architecture, and functions in the layers

L1: Amount of data needed at ingestion layer L2.

Push from L1 or Pull by L2

Source data types : database , files , web or service

Source Formats : semi structured , unstructured or structured.

L2:

Ingestion processes either in real time

Store and use of data as generated or in batches

L3:

Data storage type

Historical or incremental

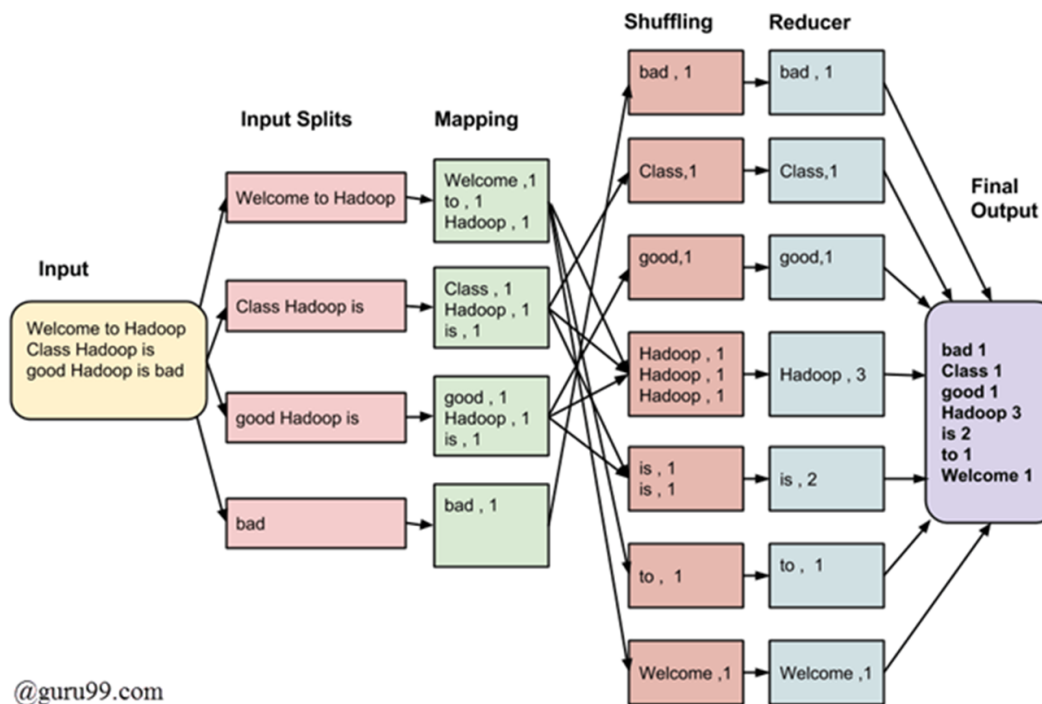
[06]

CO
1

L1

	Format Compression Incoming data frequency Using Hbase L4: Data Processing software such as MapReduce etc. Batch processing Real time etc			
(b)	List and explain the benefits of ZooKeeper Explanation -4 Marks Apache ZooKeeper is a service used by a cluster (group of nodes) to coordinate between themselves and maintain shared data with robust synchronization techniques. ZooKeeper is itself a distributed application and provides services for writing a distributed applications. Distributed applications offer a lot of benefits, but they throw a few complex and hard-to-crack challenges as well. ZooKeeper framework provides a complete mechanism to overcome all the challenges. - Race condition and deadlock are handled using a fail-safe synchronization approach. Another main drawback is the inconsistency of data- which ZooKeeper resolves with atomicity.	[04]	CO 2	L1

4 (a)	With diagram explains <i>the Map Reduce programming model</i> Diagram - 3 Marks Explanation - 3 Marks	[06]	CO 2	L3
-------	---	------	---------	----



Mapper:

Hadoop sends one record at a time to the mapper. After that, Mappers work on one record at a time and write the intermediate data to disk.

Reducer:

The function of a reducer is to aggregate the results, using the intermediate keys produced by the Mapper using aggregation, query or user-specified function. Reducers write the final concise output to the HDFS file system.

- Aggregation function means the function that groups the values of multiple rows together to result a single value of more significant meaning or measurement.
- For example, function such as count, sum, maximum, minimum, deviation and standard deviation.
- Querying function means a function that finds the desired values.
- For example, function for finding a best student of a class who has shown the best performance in examination.

MapReduce allows writing applications to process reliably the huge amounts of data, in parallel, on large clusters of servers.

(b)	<p>List and explain the Characteristics of Hadoop</p> <p>List - 4 Characteristics - 4 Marks</p> <ol style="list-style-type: none"> 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 continue 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 sub-tasks submit to the DataNodes. One can achieve more 	[04]	CO 1	L1
-----	--	------	---------	----

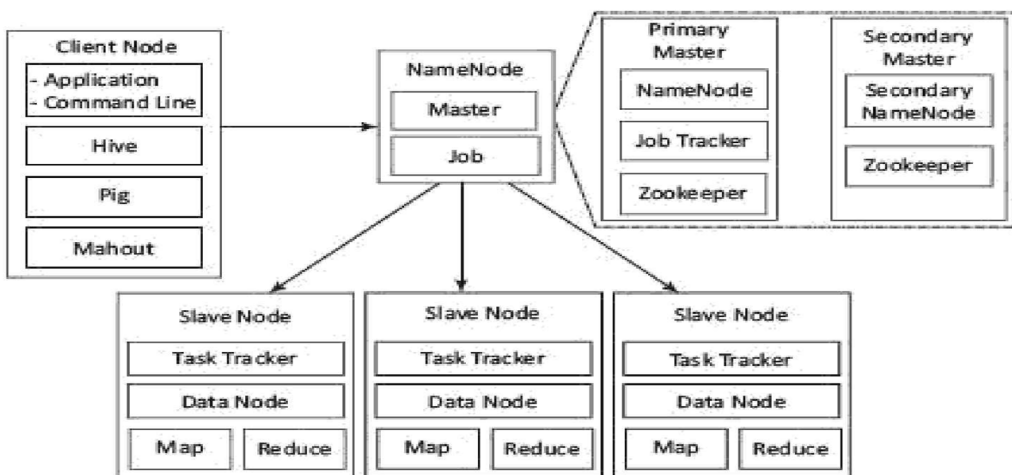
computing power by increasing the number of computing nodes. The processing splits across multiple DataNodes (servers), and thus fast processing and aggregated results.

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 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.

5 (a) Explain with diagram HDFS Name node and Data node.
 Diagram - 3 Marks
 Explanation - 3 Marks



NameNodes:

Very Few nodes in a Hadoop cluster act as NameNodes. These nodes are termed as MasterNodes (MN) or simply masters. The masters have high DRAM and processing power. The masters have much less local storage.

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.

DataNodes:

- Majority of the nodes in Hadoop cluster act as DataNodes and TaskTrackers.
- 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.

[6] CO 1 L2

(b) List the difference between Horizontal Scaling and Vertical Scaling
ist - 4 Differences - 4 Marks

Horizontal scalability

Means increasing the number of systems working in coherence. For example, using MPPs or number of servers as per the size of the dataset.

Vertical scalability

Means scaling up using the giving system resources and increasing the number of tasks in the system.

[2+2] CO 1 L2

	For example, extending analytics processing by including the reporting, business processing (BP), business intelligence (BI), data visualization, knowledge discovery and machine learning (ML) capabilities require the use of high capability hardware resources.			
6(a)	<p>Explain in detail 1)Mahout 2)Hbase</p> <p>Mahout(2 Marks) Hbase(3 Marks)</p> <p>Mahout is a project of Apache with library of scalable machine learning algorithms. Apache implemented Mahout on top of Hadoop. Apache used the MapReduce paradigm. Machine learning is mostly required to enhance the future performance of a system based on the previous outcomes. Mahout provides the learning tools to automate the finding of meaningful patterns in the Big Data sets stored in the HDFS. Similar to database, HBase is an Hadoop system database. HBase was created for large tables. HBase is an open-source, distributed, versioned and non-relational (NoSQL) database. Features of HBase features are:</p> <ol style="list-style-type: none"> 1. Uses a partial columnar data schema on top of Hadoop and HDFS. 2. Supports a large table of billions of rows and millions of columns. 3. Supports data compression algorithms. 4. Provisions in-memory column-based data transactions. 	[5]	CO 2	L1
(b)	<p>Explain with a diagram Berkeley Data Analytics Stack.</p> <p>Diagram - 2Marks Explanation - 2 Marks</p> <p>BDAS is an open-source data analytics stack for complex computations on Big Data. •It supports three fundamental processing requirements; accuracy, time and cost. Berkeley Data Analytics Stack (BDAS) consists of data processing, data management and resource management layers.</p> <ul style="list-style-type: none"> • Data processing software component provides in-memory processing which processes the data efficiently across the frameworks. • Data management software components does batching, streaming and interactive computations, backup, recovery • Resource management software component provides for sharing the infrastructure across various frameworks. 	[2+2]	CO 1	L2

CO PO Mapping

Course Outcomes		Modul es cover ed	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
			O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
CO1	Investigate Hadoop framework and Hadoop Distributed File system.	1	2	0	2	2	3	0	0	0	0	0	0	0	0	0	0	3
CO2	Illustrate the concepts of NoSQL using MongoDB and Cassandra for Big Data.	1,2	2	3	2	3	3	0	0	0	0	0	0	0	0	2	0	3
CO3	Demonstrate the MapReduce programming model to process the big data along with Hadoop tools.	3	2	2	3	3	3	0	0	0	0	0	0	0	0	2	0	3
CO4	Use Machine Learning algorithms for real world big data.	2,3,4	2	3	3	2	3	0	0	0	0	0	0	0	0	2	0	3
CO5	Analyze web contents and Social Networks to provide analytics with relevant visualization tools.	5	2	3	3	3	3	0	0	0	0	0	0	0	0	2	0	3
CO6	Investigate Hadoop framework and Hadoop Distributed File system.	5	2	3	2	2	3	0	0	0	0	0	0	0	0	2	0	3

COGNITIVE LEVEL	REVISED BLOOMS TAXONOMY KEYWORDS
L1	List, define, tell, describe, identify, show, label, collect, examine, tabulate, quote, name, who, when, where, etc.
L2	summarize, describe, interpret, contrast, predict, associate, distinguish, estimate, differentiate, discuss, extend
L3	Apply, demonstrate, calculate, complete, illustrate, show, solve, examine, modify, relate, change, classify, experiment, discover.
L4	Analyze, separate, order, explain, connect, classify, arrange, divide, compare, select, explain, infer.
L5	Assess, decide, rank, grade, test, measure, recommend, convince, select, judge, explain, discriminate, support, conclude, compare, summarize.

PROGRAM OUTCOMES (PO), PROGRAM SPECIFIC OUTCOMES (PSO)				CORRELATION LEVELS	
PO1	Engineering knowledge	PO7	Environment and sustainability	0	No Correlation
PO2	Problem analysis	PO8	Ethics	1	Slight/Low
PO3	Design/development of solutions	PO9	Individual and team work	2	Moderate/ Medium
PO4	Conduct investigations of complex problems	PO10	Communication	3	Substantial/ High
PO5	Modern tool usage	PO11	Project management and finance		
PO6	The Engineer and society	PO12	Life-long learning		
PSO1	Develop applications using different stacks of web and programming technologies				
PSO2	Design and develop secure, parallel, distributed, networked, and digital systems				
PSO3	Apply software engineering methods to design, develop, test and manage software systems.				
PSO4	Develop intelligent applications for business and industry				