Internal Assessment Test 1 – Nov 2021 **Scheme and Solution**

Sub:	Big Data Analytics					Sub Code:	18CS72	Branch:	CSE
Date:	15/11/21	Duration:	90 mins	Max Marks:	50	Sem / Sec:	VII/A,B,C	OBE	

		MARKS	CO	RBT
1	Compare distributed computing, cluster computing and grid computing?	[10]	CO1	L2
2	Briefly explain the functions of five layers in Big data architecture design. Architecture diagram-2 marks Layer functions-8 marks	[10]	CO1	L2
3	Explain the functions of data preprocessing. Explain different data enrichment methods in Big Data Analytics(BDA). Data preprocessing functions- 2 marks Data enrichment method-2 marks each	[10]	CO1	L2
4	Explain Hadoop based Big data environment with suitable diagram. Mention the Big data stack. Big data environement-2 marks Hadoop-2 marks Diagram and explanation – 6 marks	[10]	CO1	L2
5	Define data analytics. Explain phases of analytics in BDA. Data analytics definition- 2 marks phases of analytics in BDA- 8 marks	[10]	CO1	L3
6	Briefly explain any four applications of BDA. Each application – 2.5 marks	[10]	CO1	L2
7	Explain each core component in Hadoop with suitable diagram. Explain with diagram the Hadoop ecosystem components layers? Explain by an example. Hadoop core components and diagram- 4+1 marks Diagram and the Hadoop ecosystem components layers- 1+5 marks	[10]	CO2	L2
8	What are the functions of Name node, Data node, Slave node, and Master node? Explain with suitable diagram. Functions of Name node, Data node, Slave node, and Master node- 8 marks diagram- 2 marks	[10]	CO2	L2

1. Distributed computing

A distributed Computing model useS cloud, gria or clusters, which process and analyze big and large datasets on distributed computing nodes connected by high speed networks. Table 1.2 gives the requirements of processing and analyzing big, large and small to medium datasets on distributed computing nodes. Big Data processing uses a parallel, scalable and no-sharing program model, such as MapReduce, for computations on it.

Grid 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 comP.uter resources are heterogeneously and geographically dispersed. A group of computers that might spread over: remotely comprise a gri . A grid is used for a variety of purposes. A single grid of course, dedicated at an instance to a particular application onl . Grid computing provides large-scale resource free sharing which is flexible, coordinated and secure among its users irhe users consist of individuals, organizations and resources.

Grid computing suits data-intensive storage better than storage of small objects of a few millions of bytes. To achieve the maximum benefit from data grids, they should be used for a large amount of data which can be distributed over grid nodes. Besides data grid, the other variation of id, i.e., computational grid focuses on computationally intensive operations.

Features of Grid Computing Grid computing, similar to cloud computing, is scalable.

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 tlie single point, which leads to failure in case 0£ 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 gi:ven 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 node.s to keep an even load on the group of connected computers. Hadoop architecture uses the similar methods

Distributed computing on multiple processing nodes/clusters	Big Data> lOM	Large datasets below 10 M	Small to medium datasets up to 1 M
Distributed computing	Yes	Yes	No
Parallel computing	Yes	Yes	No

Scalable computing	Yes	Yes	No
Shared nothing (No in-between data sharing and inter-processor communication)	Yes	Limited sharing	No
Shared in-between between the distributed nodes/clusters	No	Limited sharing	Yes

Distributed computing	Cluster computing	Grid computing
•Loosely coupled •Heterogeneous •Single administration	•Tightly coupled •Homogeneous •Cooperative working	•Large scale •Cross organizational •Geographical distribution •Distributed management

---

2. Briefly explain the functions of five layers in Big 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 and

(iv) Data processing

(v) data consumption in number of programs and tools

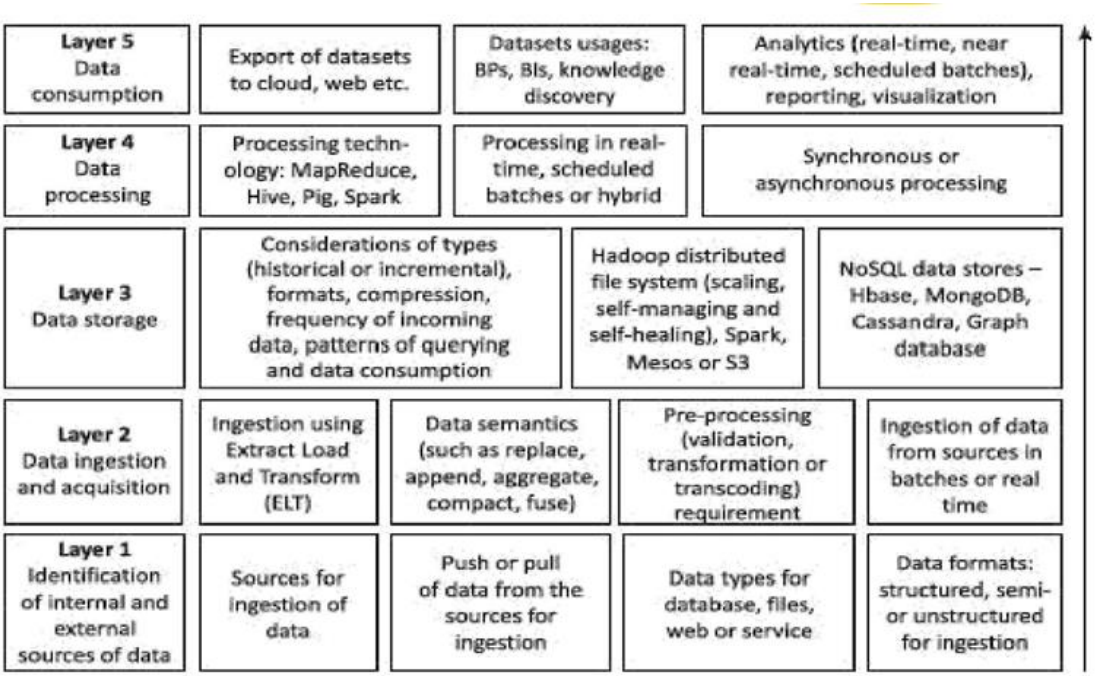

Figure 1.2 Design of logical layers in a data processing architecture, and functions in the layers

Logical layer 1 (Ll) is for identifying data sources, which are external, internal or both. The layer 2 (L2) is for data-ingestion.

Data ingestion means a process of absorbing information, just like the process of absorbing nutrients and medications into the body by eating or drinking them (Cambridge English Dictionary). Ingestion is the process of obtaining and importing data for immediate use or transfer. Ingestion may be in batches or in real time using pre-processing or semantics.

The L3 layer is for storage of data from the L2 layer. The L4 is for data processing using software, such as MapReduce, Hive, Pig or Spark. The top layer LS is for data consumption. Data is used in analytics, visualizations, reporting, export to cloud or web servers.

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 LS

- 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 LS. LS 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, Bis, knowledge discovery
- Export of datasets to cloud, web or other systems

---

3. Explain the functions of data preprocessing. Explain different data enrichment methods in Big Data Analytics(BDA).

Pre-processing needs are:

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

(ii) Filtering unreliable, irrelevant and redundant information (iii)Data cleaning, editing, reduction and/or wrangling (iv)Data validation, transformation or transcoding

(v) ELT processing.

Data Cleaning

Data cleaning refers to the process of removing or correcting incomplete, incorrect, inaccurate or irrelevant parts of the data after detecting them. For example, in Example correcting the grade outliers or mistakenly entered values means cleaning and correcting the data.

"Data enrichment refers to operations or processes which refine, enhance or improve the raw data."

Data Editing

Data editing refers to the process of reviewing and adjusting the acquired datasets. The editing controls the data quality. Editing methods are (i) interactive, (ii) selective, (iii) automatic, (iv) aggregating and (v) distribution.

Data Reduction

Data reduction enables the transformation of acquired information into an ordered, correct and simplified form. The reductions enable ingestion of meaningful data in the datasets. The basic concept is the reduction of multitudinous amount of data, and use of the meaningful parts. The reduction uses editing, scaling, coding, sorting, collating, smoothening, interpolating and preparing tabular summaries.

Data Wrangling

Data wrangling refers to the process of transforming and mapping the data. Results from analytics are then appropriate and valuable. For example, mapping enables data into another format, which makes it valuable for analytics and data visualizations.

Data Format used during Pre-Processing

Examples of formats for data transfer from (a) data storage, (b) analytics application, (b) service or (d) cloud can be:

(i) Comma-separated values CSV

(ii) Java Script Object Notation USON) as batches of object arrays or resource arrays

(iii) Tag Length Value (TLV) (iv)Key-value pairs

(v) Hash-key-value pairs

Data Format Conversions

Transferring the data may need pre-processing for data-format conversions. Data sources store need portability and usability. A number of different applications, services and tools need a specific format of data only. Pre-processing before their usages or storage on cloud services is a must.

---

4. Explain Hadoop based Big data environment with suitable diagram. Mention the Big data stack.

Bigdata platform should provision tools and services for:

1. storage, processing and analytics,
2. developing, deploying, operating and managing Big Data environment,
3. reducing the complexity of multiple data sources and integration of applications into one cohesive solution,
4. custom development, querying and integration with other systems, and
5. the traditional as well as Big Data techniques.

Hadoop

Storage can deploy Hadoop Distributed File System (HDFS), NoSQL data stores, such as HBase, MongoDB, Cassandra.

HDFS system is an open source storage system.

HDFS is a scaling, self-managing and self-healing file system.

The Hadoop system packages application-programming model.

Hadoop is a scalable and reliable parallel computing platform.

Hadoop manages Big Data distributed databases.

Small height cylinders represent MapReduce and big ones represent the Hadoop.

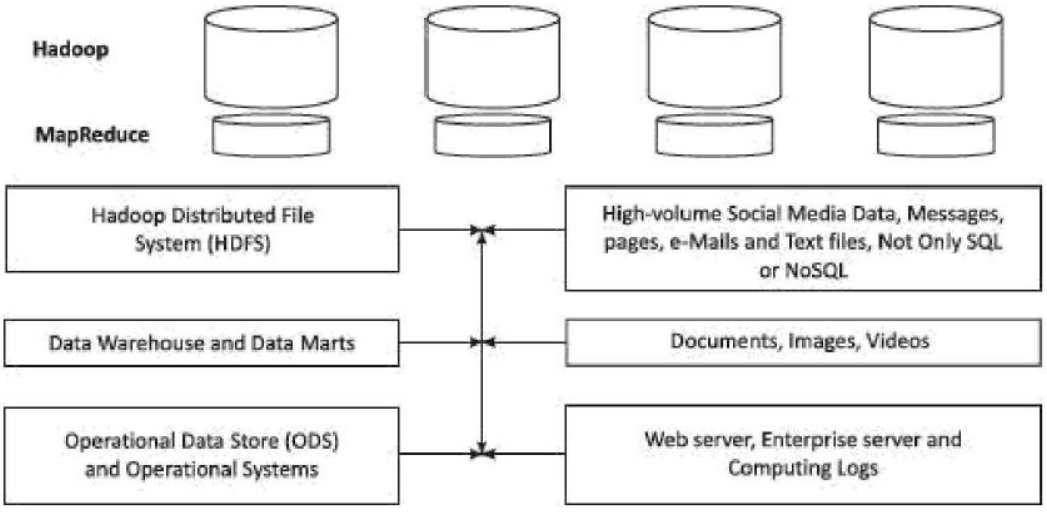

Figure 1.8 Hadoop based Big Data environment

A stack consists of a set of software components and data store units. Applications, machine-learning algorithms, analytics and visualization tools use Big Data Stack (BDS) at a cloud service, such as Amazon ECZ, Azure or private cloud. The stack uses cluster of high performance machines.

Types	Examples
MapReduce	Hadoop, Apache Hive, Apache Pig, Cascading, Cascalog, mrjob (Python MapReduce library), Apache S4, MapR, Apple Acunu, Apache Flume, Apache Kafka
NoSQL Databases	MongoDB, Apache CouchDB, Apache Cassandra, Aerospike, Apache HBase, Hypertable
Processing	Spark, IBM BigSheets, PySpark, R, Yahoo! Pipes, Amazon Mechanical Turk, Datameer, Apache Solr/Lucene, ElasticSearch
Servers	Amazon ECZ, S3, GoogleQuery, Google App Engine, AWS Elastic Beanstalk, Salesforce Heroku
Storage	Hadoop Distributed File System, Amazon S3, Mesos

---

5. Define data analytics. Explain phases of analytics in BDA.

Data Analytics can be formally defined as the statistical and mathematical data analysis that clusters, segments, ranks and predicts future possibilities. An important feature of data analytics is its predictive, forecasting and prescriptive capability.

Analytics has the following phases before deriving the new facts, providing business intelligence and generating new knowledge.

Descriptive analytics enables deriving the additional value from visualizations and reports

Predictive analytics is advanced analytics which enables extraction of new facts and knowledge, and then predicts/forecasts

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 decisions.

Analytics integrates with the enterprise server or data warehouse

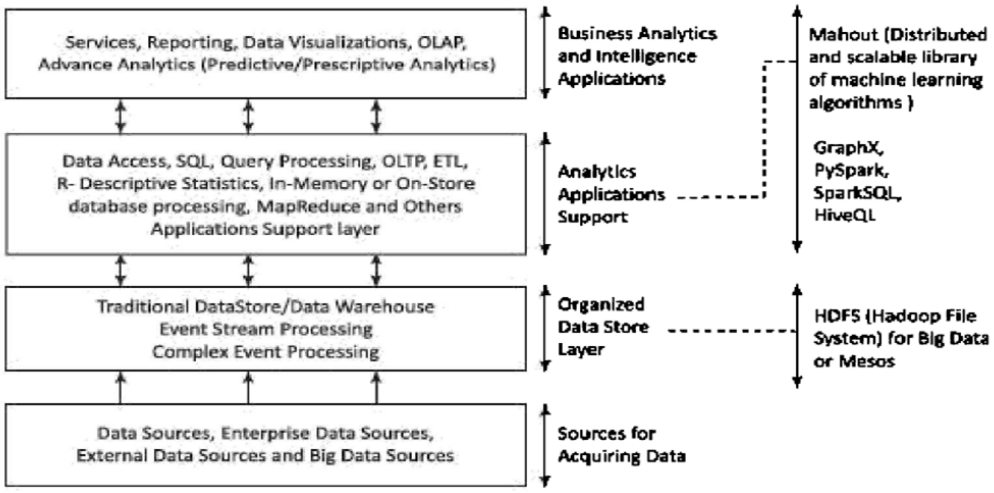

Figure 1.9 Traditional and Big Data analytics architecture reference model

The captured or stored data require a well-proven strategy to calculate, plan or analyze.

When Big Data combine with high-powered data analysis, enterprise achieve valued business-related tasks.

Examples are:

- Determine root causes of defects, faults and failures in minimum time.
- Deliver advertisements on mobiles or web, based on customer's location and buying habits.
- Detect offender before that affects the organization or society.

6. Briefly explain any four applications of BDA.

Big Data in Marketing and Sales

Data are important for most aspect of marketing, sales and advertising. Customer Value (CV) depends on three factors - quality, service and price. Big data analytics deploy large volume of data to identify and derive intelligence using predictive models about the individuals. The facts enable marketing companies to decide what products to sell.

A definition of marketing is the creation, communication and delivery of value to customers. Customer (desired) value means what a customer desires from a product. Customer (perceived) value means what the customer believes to have received from a product after purchase of the product. Customer value analytics (CVA) means analyzing what a customer really needs. CVA makes it possible for leading marketers, such as Amazon to deliver the consistent customer experiences. Following are the five application areas in order of the popularity of Big Data use cases:

1. CVA using the inputs of evaluated purchase patterns, preferences, quality, price and

post sales servicing requirements

2. Operational analytics for optimizing company operations
3. Detection of frauds and compliances
4. New products and innovations in service

5. Enterprise data warehouse optimization.

1.7.1.1 Big Data Analytics in Detection of Marketing Frauds

Fraud detection is vital to prevent financial loses to users. Fraud means someone deceiving deliberately. For example, mortgaging the same assets to multiple financial institutions, compromising

customer data and transferring customer information to third party, falsifying company information to financial institutions, marketing product with compromising quality, marketing product with service level different from the promised, stealing intellectual property, and much more.

Big Data analytics enable fraud detection. Big Data usages has the following features-for enabling detection and prevention of frauds:

1. 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

2. Using multiple sources of data and connecting with many applications

3. Providing greater insights using querying of the multiple source data

4. Analyzing data which enable structured reports and visualization

5. Providing high volume data mining, new innovative applications and thus leading to new business intelligence and knowledge discovery

6. Making it less difficult and faster detection of threats, and predict likely frauds by using various data and information publicly available.

7. Explain each core component in Hadoop with suitable diagram. Explain with diagram the Hadoop ecosystem components layers? Explain by an example.

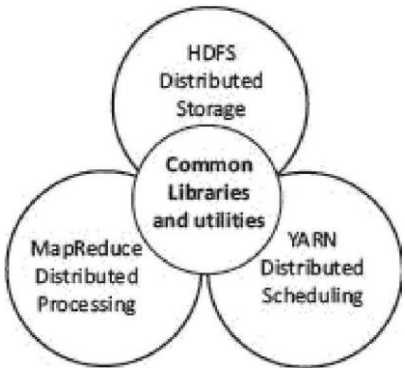

Figure 2.1 Core components of Hadoop

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 vl - Software programming model in Hadoop 1 using Mapper and Reducer. The vl 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 and distributed processing of the application tasks.

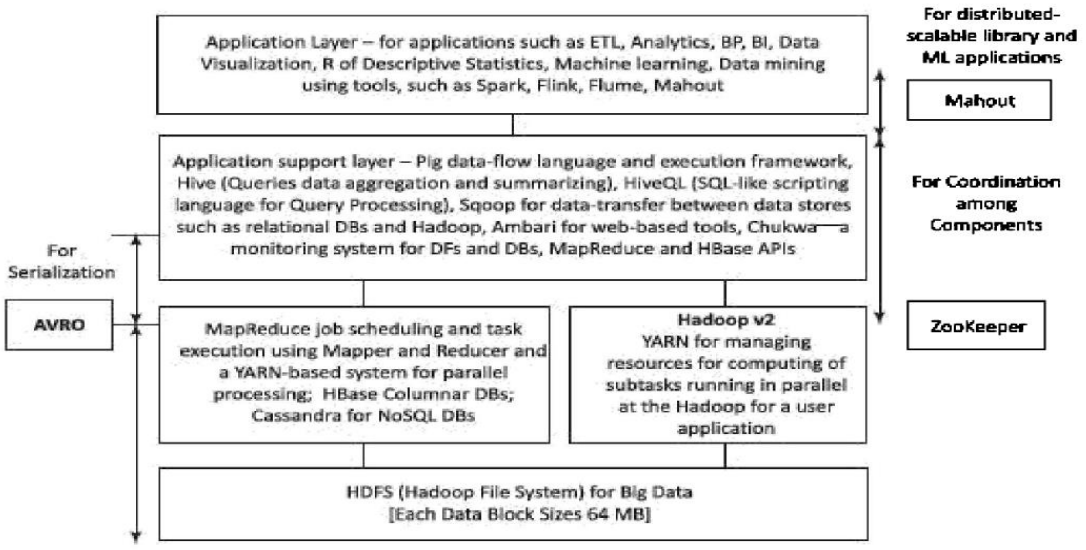

Figure 2.2 Hadoop main components and ecosystem components

The four layers in Figure 2.2 are as follows:

(i) Distributed storage layer

(ii) Resource-manager layer for job or application sub-tasks scheduling and execution

(iii) Processing-framework layer, consisting of Mapper and Reducer for the MapReduce process-flow

(iv) APis at application support layer (applications such as Hive and Pig). The codes communicate and run using MapReduce or YARN at processing framework layer. Reducer output communicate to APis (Figure 2.2).

Q8. What are the functions of Name node, Data node, Slave node, and Master node? Explain with suitable diagram.

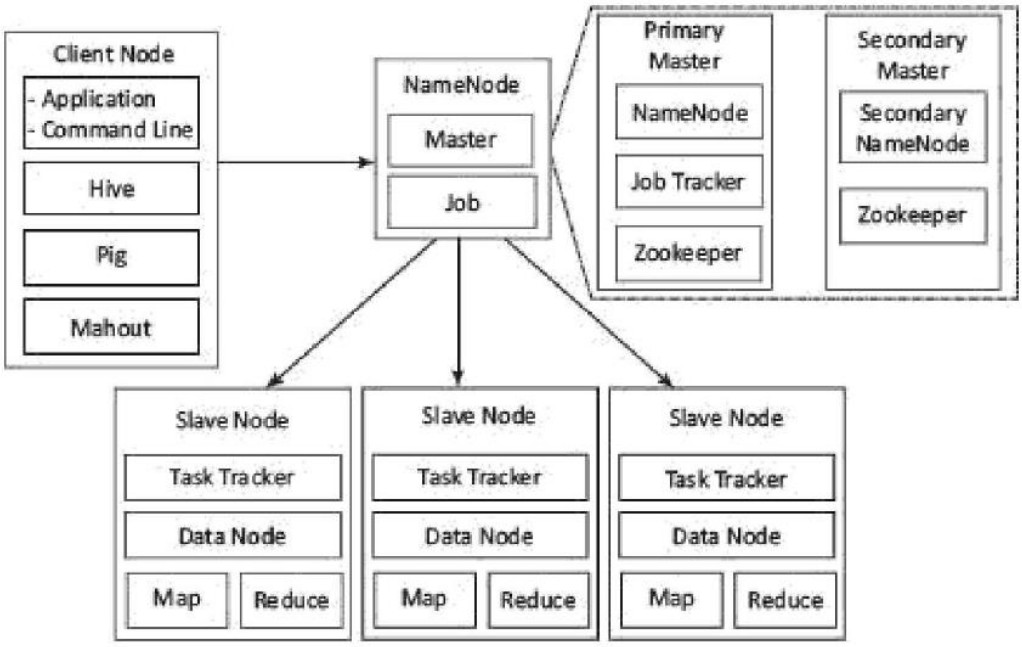

Figure 2.4 The client, master NameNode, MasterNodes and slave nodes

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 TaskTrackers. These nodes are referred to as slave nodes or slaves. The slaveshave 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 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 NameNode 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