DR. BABASAHEB AMBEDKAR TECHNOLOGICAL UNIVERSITY, LONERE – RAIGAD -402 103
Mid Semester Examination – October - 2017

Branch: M.Tech (Automation) **Sem.:- I**

Subject with Subject Code: - Data Analytics for Automation, **ME-XX103**

Marks: 20

Date: - **Time: - 1 Hr.**

==========

Model Answer Scheme

(Marks)

Q.No.1 Attempt any one of the following

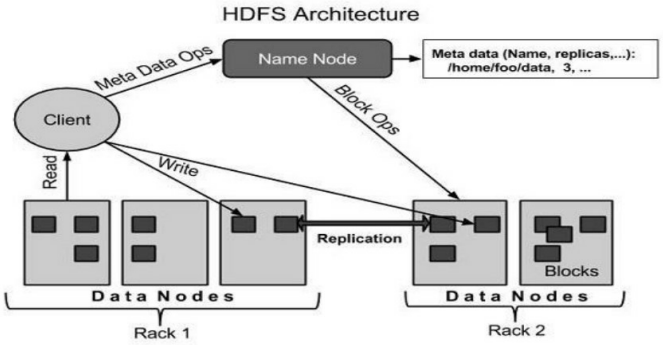
Marking Scheme: Distribution of 8 Marks

a) Definition and explanation (2M)
b) HDFS Architecture (namenode, Datanode & Blocks) (4M)
c) Goals (2M)

a.) Define HDFS. Explain HDFS in detail.

Answer: Hadoop File System was developed using distributed file system design. It is run on commodity hardware. Unlike other distributed systems, HDFS is highly fault-tolerant and designed using low-cost hardware. HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel Processing.

HDFS Architecture



- **Namenode:**

The namenode is the commodity hardware that contains the GNU/Linux operating system and the namenode software. It is software that can be run on commodity hardware. The system having the namenode acts as the master server and it does the following tasks: Manages the file system namespace. Regulates client's access to files. It also executes file system operations such as renaming, closing, and opening files and directories.

- **Datanode:**

The datanode is a commodity hardware having the GNU/Linux operating system and datanode software. For every node (Commodity hardware/System) in a cluster, there will be a datanode. These nodes manage the data storage of their system. Datanodes perform read-write operations on the file systems, as per client request. They also perform operations such as block creation, deletion, and replication according to the instructions of the namenode.

- **Blocks:**

Generally the user data is stored in the files of HDFS. The file in a file system will be divided into one or more segments and/or stored in individual data nodes. These file segments are called as blocks. In other words, the minimum amount of data that HDFS can read or write is called a Block. The default block size is 128MB, but it can be increased as per the need to change in HDFS configuration.

- **Goals:**

Fault detection and recovery: Since HDFS includes a large number of commodity hardware, failure of components is frequent. Therefore HDFS should have mechanisms for quick and automatic fault detection and recovery. Huge datasets: HDFS should have hundreds of nodes per cluster to manage the applications having huge datasets. Hardware at data: A requested task can be done efficiently, when the computation takes place near the data. Especially where huge datasets are involved, it reduces the network traffic and increases the throughput.

b.) Write any four Big Data Analytics Applications Area with example.

Marking Scheme: Distribution of 8 Marks

a) Social media (2M)
b) Retailer (2M)
c) Healthcare (2M)
d) Telecom (2M)

Answer:

- **Social Media**

Social Media Command Center combines automated search and display of consumer feedback expressed publicly on the social media. Often, the feedback is summarized in the form of “positive” or “negative” sentiment. Once the feedback is obtained, the marketer can respond to specific comments by entering into a conversation with the affected consumers, whether to respond to questions about an outage or obtain feedback about a new product offering.

- **Retailer**

A common example is that of a retailer being able to sift through tones of consumer data to derive insights on shopping preferences and direct targeted campaigns. This can even be extended to capture the personal preferences and likes of the shopper and provide customized offers, leading to increased hit rates and revenues. This is a win-win situation for both parties as the consumer gets information and offers that he is interested in and the retailer enjoys revenue growth and potential customer loyalty as well. Big Data analytics does not have to adopt a big-bang approach all the time and is equally useful and effective in behind-the-scene scenarios for retailers. It can be used for a dramatic reduction in processing time when comparing product information, which resides across multiple data sources. Analyzing data allows a retailer to make intelligent decisions and helps gain a competitive edge.

- **Healthcare**

Big Data analytics has immense potential in the field of healthcare too. Imagine if a hospital is able to go through its patient records and identify patterns in diseases. This can enable doctors to detect the onset of a disease much early on and the benefits of such an approach cannot be overstated. Throw in lifestyle data to gather additional insights and the possibilities are simply mind-boggling. Obvious gains are decreased mortality rates, better quality of life due to accurate prognosis, diagnosis and treatment, and lowered insurance costs. However, the challenge will be in overcoming regulatory and patient confidentiality issues.

- **Telecom**

Another example is from the telecom industry. Mobile connections are expected to exceed 6 billion globally and in India there are close to 750 million subscribers. In this highly connected world, the amount of data available is colossal and Telco's can cleverly mine this data to their advantage. The biggest impact can be experienced by studying subscriber persona and usage patterns and using that intelligence to devise targeted marketing campaigns. The analytics can also help Telco's determine what additional services are likely to find favor with subscribers and offer them appropriately. This also provides opportunities to offer value-added services such as location-based services leading to better customer service.

Q.No. 2 Attempt any three of the following: (12)

a.) Define the drivers for Big Data-Velocity, Variety, Volume, and Veracity.

Marking Scheme: Distribution of 4 Marks

a) Volume (1M)
b) Variety (1M)
c) Velocity (1M)
d) Veracity(1M)

Answer:

- **Volume**
- The quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.
- **Variety**

  The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.
- **Velocity**

  In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.
- **Veracity**

  The quality of captured data can vary greatly, affecting accurate analysis.

b.) Write any six Open Source Technologies for Big Data analytics in detail?

Marking Scheme: Distribution of 4 Marks

a) Apache Hadoop (2M)
b) Apache Storm (2M)
c) MOngoDB (2M)
d) R-Programming(2M)

Answer:

- **Apache Hadoop**

Hadoop has become synonymous with big data and is currently the most popular distributed data processing software. This powerful system is known for its ease of use and its ability to process extremely large data in both, structured and unstructured formats, as well as replicating chunks of data to nodes and making it available on the local processing machine. Apache has also introduced other technologies that accentuate Hadoop's capabilities such as Apache Cassandra, Apache Pig, Apache Spark and even ZooKeeper. You can learn this amazing technology using real world examples here.

- **Apache Storm**

Apache Storm can be used with or without Hadoop, and is an open source distributed realtime computation system. It makes it easier to process unbounded streams of data, especially for real-time processing. It is extremely simple and easy to use and can be configured with any programming language that the user is comfortable with. Storm is great for using in cases such as realtime analytics, continuous computation, online machine learning, etc. Storm is scalable and fast, making it perfect for companies that want fast and efficient results.

- **MOngoDB**

MongoDB is also a great tool to help store and analyze big data, as well as help make applications. It was originally designed to support humongous databases, with its name MongoDB, actually derived from the word humongous. MongoDB is a no SQL database that is written in C++ with document-oriented storage, full index support, replication and high availability, etc.

- **R- Programming**

R isn't just software, but also a programming language. Project R is the software that has been designed as a data mining tool, while R programming language is a high-level statistical language that is used for analysis. An open source language and tool, Project R is written is R language and is widely used among data miners for developing statistical software and data analysis. In addition to data mining it provides statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.

c.) Explain Map-reduce framework in detail?

Marking Scheme: Distribution of 4 Marks

a) Map and Reduce Introduction (1M)

b) Algorithm (2M)

c) Inserting Data (1M)

MapReduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner.MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce.

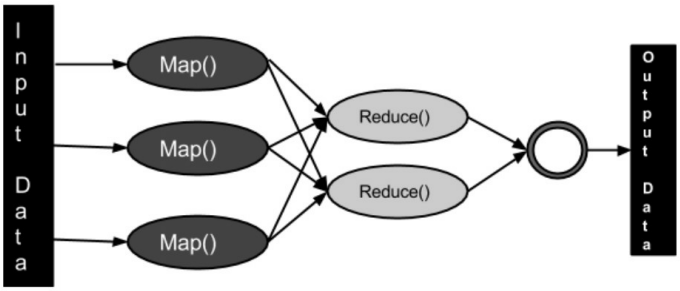
- Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).
- Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

- **The Algorithm**

MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

Map stage: The map or mapper's job is to process the inputdata. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.



- **Inserting Data into HDFS**

The MapReduce framework operates on <key, value> pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output of the job, conceivably of different types.

• The key and the value classes should be in serialized manner by the framework and hence, need to implement the Writable interface. Additionally, the key classes have to implement the Writable-Comparable interface to facilitate sorting by the framework.

• Input and Output types of a MapReduce job: (Input) <k1,v1> -> map -> <k2, v2>-> reduce -> <k3, v3> (Output).

	Input	Output
Map	<k1, v1>	list (<k2, v2>)
Reduce	<k2, list(v2)>	list (<k3, v3>)

d.) What is disparate data? How Hadoop Access disparate data?

Marking Scheme: Distribution of 4 Marks
a) Definition (1M)
b) Access method- any 6 (3M)

Answer:

Definition: Disparate data are any data that are essentially not alike, or are distinctly different in kind, quality, or character. They are unequal and cannot be readily integrated to meet the business information demand. They are low quality, defective, discordant, ambiguous, heterogeneous data.
Hadoop Access Data with the help of following Tools:

- The programming language **Pig** (which includes Pig Latin) is specifically intended to analyze any type of data without requiring the user to spend a lot of time setting up mapping and reduction programs;
- **HIVE**, a SQL-like query language that breaks a SQL statement down to a MapReduce job and distributes it across the cluster;
- **Flume**, which collects and aggregates large amounts of data from applications and moves them into the HDFS file system;
- **Spark** is an open-source cluster computing data analytics program that, in certain circumstances, can be 100 times faster than MapReduce;
- A data transfer program called **Sqoop** that can extract, load, and transform structured data;
- **Hbase**, a non-relational, non-SQL database that runs on top of HDFS and can support very large tables;
- **Avro**, which is a system that serializes data;
- **Tez**, which is a generalized data flow framework that uses the Hadoop module YARN to execute tasks in order to process data in both batch and interactive use modes.