Biyani Institute of Science and Management
I Internal Examination(Solution) Sept. 2019
Class: MCA (V Semester)
Subject-Big Data Technologies & Analytics(MCA-504A)

MM: 20 **Set:A** **Time: 1 ½Hrs**

[I] Answer the following questions in one line only **(2*1=02)**

Q.1 What is Big Data?

Ans. Big Data is also data but with a huge size. Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time. In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently. Example: The statistic shows that 500+terabytes of new data get ingested into the databases of social media site Facebook, every day.

Q.2 What is Data Node and Name Node in HDFS?

Ans. NameNode: NameNode is the master node in the Apache Hadoop HDFS Architecture that maintains and manages the blocks present on the DataNodes (slave nodes). NameNode is a very highly available server that manages the File System Namespace and controls access to files by clients.

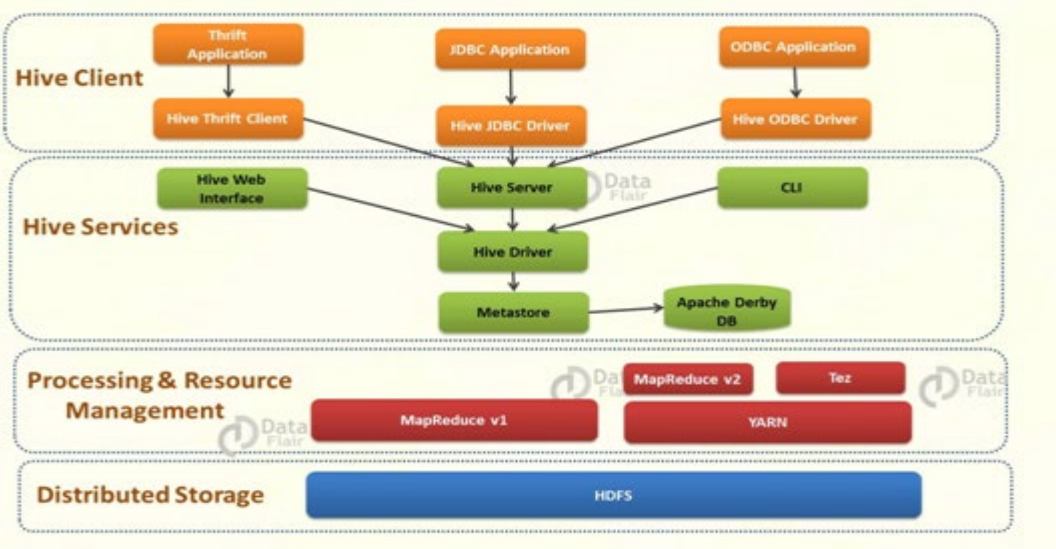
DataNode: DataNodes are the slave nodes in HDFS. Unlike NameNode, DataNode is a commodity hardware, that is, a non-expensive system which is not of high quality or high-availability.

[II]Answer the following questions in 50 words **(2*3=06)**

Q.1 Describe Architecture of Apache HIVE.

Ans. Hive is a Data warehousing tool developed on top of Hadoop Distributed File System (HDFS).

Hive Architecture:-



Hive Clients – Apache Hive supports all application written in languages like C++, Java, Python etc. using JDBC, Thrift and ODBC drivers. Thus, one can easily write Hive client application written in a language of their choice.

- **Hive Services** – Hive provides various services like web Interface, CLI etc. to perform queries.
- **Processing framework and Resource Management** – Hive internally uses Hadoop MapReduce framework to execute the queries.
- **Distributed Storage** – As seen above that Hive is built on the top of Hadoop, so it uses the underlying HDFS for the distributed storage.

Q.2 Define 4 V's of Big Data?

Ans. Volume

Big data always has a large volume of data. This is due to the building up of a volume of data from unstructured sources like social media interaction, posting or sharing reviews on the web page, mobile phones, and many more. Whenever a user visits the website using desktop, laptop, smartphones, PDAs, etc. generates the traffic.

Velocity

Velocity refers to the speed at which the data is coming in and how quickly the organizations are analysing and utilizing it. Processing the data using analytics tool can produce the answers to the queries through reports, dashboards, etc. With these results, a company can make suitable decisions that increase the efficiency and achieve customer-relation objectives such as developing applications that cater to the needs.

Variety

Different sources like social media, CRM systems, call center logins, emails, audio and video forms produce varied data. Managing such complex data is a big challenge for companies. However, to manage this big data, analytics tools are used to segregate groups based on sources and data generated. This would avoid mixing of data in the database. It is important to segregate new and old data coming from varied sources and must be able to make the changes according to customer behaviour.

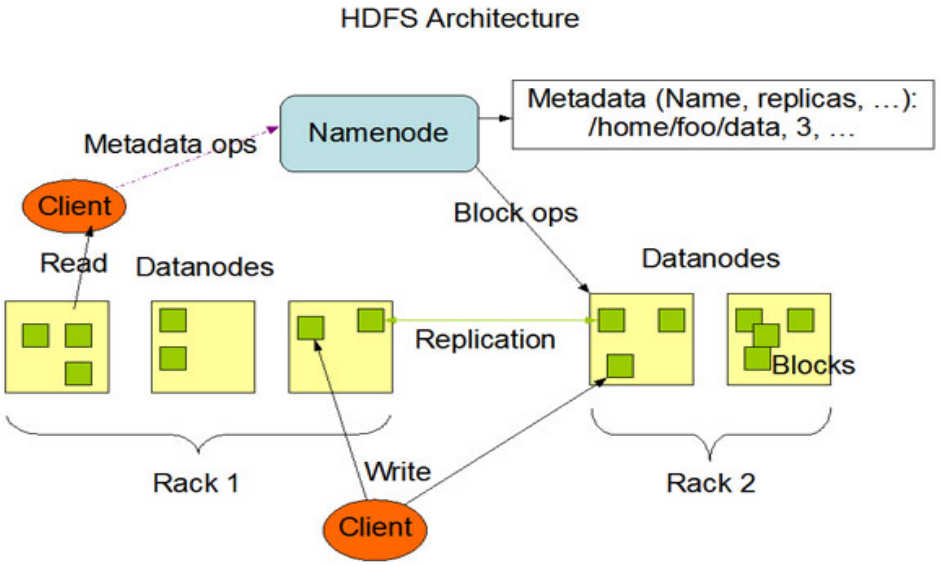
Veracity

Veracity refers to an uncertainty of data available, which makes it harder for the companies to react quickly and make appropriate solutions. Accuracy is the major issue in such a big data environment. Organizing the data according to groups, value and significance will enable you to have a better strategy to use the data. Avoid mixing to related and unrelated data as this reduce mixed interpretation. To make right decisions, the data must be clean, consistent and consolidated.

[III]Answer the following questions in 150 words. **(2*6=12)**

Q.1 Describe HDFS Architecture in detail?

Ans. Hadoop Distributed File System follows the master–slave data architecture. Each cluster comprises a single Namenode that acts as the master server in order to manage the file system namespace and provide the right access to clients. The next terminology in the HDFS cluster is the Datanode that is usually one per node in the HDFS cluster. The Datanode is assigned with the task of managing the storage attached to the node that it runs on. HDFS also includes a file system namespace that is being executed by the Namenode for general operations like file opening, closing, and renaming, and even for directories. The Namenode also maps the blocks to Datanodes.



Namenode and Datanode are actually Java programming codes that can be extensively run on commodity hardware machines. These machines could most probably be running on Linux OS or GNU. The entire HDFS is based on the Java programming language. The single Namenode on a Hadoop cluster centralizes all the arbitration and repository-related issues without creating any ambiguity.

Q.2 Difference between data analyst and data scientist

Ans.

- The job role of a data scientist strong business acumen and data visualization skills to converts the insight into a business story whereas a data analyst is not expected to possess business acumen and advanced data visualization skills.
- Data scientist explores and examines data from multiple disconnected sources whereas a data analyst usually looks at data from a single source like the CRM system.
- A data analyst will solve the questions given by the business while a data scientist will formulate questions whose solutions are likely to benefit the business.
- In many scenarios, data analysts are not expected have hands-on machine learning experience or build statistical models but the core responsibility of a data scientist is to build statistical models and be well-versed with machine learning.
- Most Data Scientists / Analysts get productive on their projects by having access to a ready-to-use library of sample solved code snippets.

Biyani Institute of Science and Management
I Internal Examination(Solution) Sept. 2019
MCA (V Semester)
Subject- Big Data Technologies & Analytics(MCA-504A)

MM: 20 **Set: B** **Time: 1 ½Hrs**

[I] Answer the following questions in one line only **(2*1=02)**

Q.1 What is BigData Analytics?

Ans. Big data analytics is the often complex process of examining large and varied data sets, or big data, to uncover information -- such as hidden patterns, unknown correlations, market trends and customer preferences -- that can help organizations make informed business decisions.

Q.2 What is the difference between Name Node and Secondary Name Node?

Ans. NameNode: NameNode is the master node in the Apache Hadoop HDFS Architecture that maintains and manages the blocks present on the DataNodes (slave nodes). NameNode is a very highly available server that manages the File System Namespace and controls access to files by clients.

DataNode: DataNodes are the slave nodes in HDFS. Unlike NameNode, DataNode is a commodity hardware, that is, a non-expensive system which is not of high quality or high-availability.

[II]Answer the following questions in 50 words **(2*3=06)**

Q.1 Describe the difference between PIG and Hive?

Ans. Difference Between Hive and Pig as follows:-

Pig	Hive
Procedural Data Flow Language	Declarative SQLish Language
For Programming	For creating reports
Mainly used by Researchers and Programmers	Mainly used by Data Analysts
Operates on the client side of a cluster.	Operates on the server side of a cluster.
Does not have a dedicated metadata database.	Makes use of exact variation of dedicated SQL DDL language by defining tables beforehand.
Pig is SQL like but	Directly leverages SQL and is easy to

varies to a great extent. learn for database experts.

Q.2 Describe different types of BigData Analytics?
Ans.

1. Predictive Data Analytics

Predictive analytics may be the most commonly used category of data analytics as it is used to identify trends, correlations, and causation. The category can be further broken down into predictive modeling and statistical modeling. But, it's important to know that these two really go hand in hand.

2. Prescriptive Data Analytics

Prescriptive analytics is where AI and big data meet to help predict outcomes and what actions to take. This category of analytics can be further broken down into optimization and random testing. Using advancements in machine learning, prescriptive analytics can help answer questions like "What if we try this?" and "What is the best action" without spending the time actually trying out each variable.

3. Diagnostic Data Analytics

Diagnostic data analytics is the process of examining data to understand cause and event, or why something happened. Techniques like drill-down, data discovery, data mining, and correlations are often employed.

4. Descriptive Data Analytics

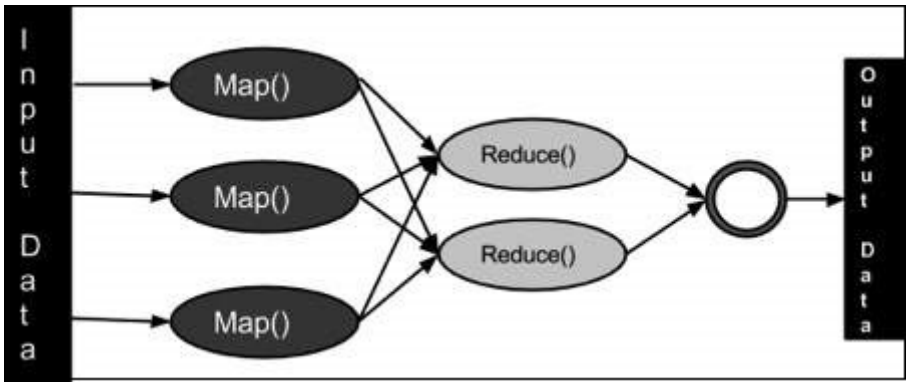
Descriptive analytics are the backbone of reporting—it's impossible to have BI tools and dashboards without it. It addresses your basic how many, when, where, and what questions. Once again, this can be further separated into two categories: ad hoc reporting and canned reports. A canned report is one that has been designed previously and contains information around a given subject. An example of this a monthly report sent by your ad agency or ad team that details performance metrics on your latest ad efforts.

[III]Answer the following questions in 150 words. **(2*6=12)**

Q.1 Describe Map Reduce Paradigm in detail?

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 major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into *mappers* and *reducers* is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.



The Algorithm

Generally MapReduce paradigm is based on sending the computer to where the data resides!

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 input data. 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.

During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.

The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.

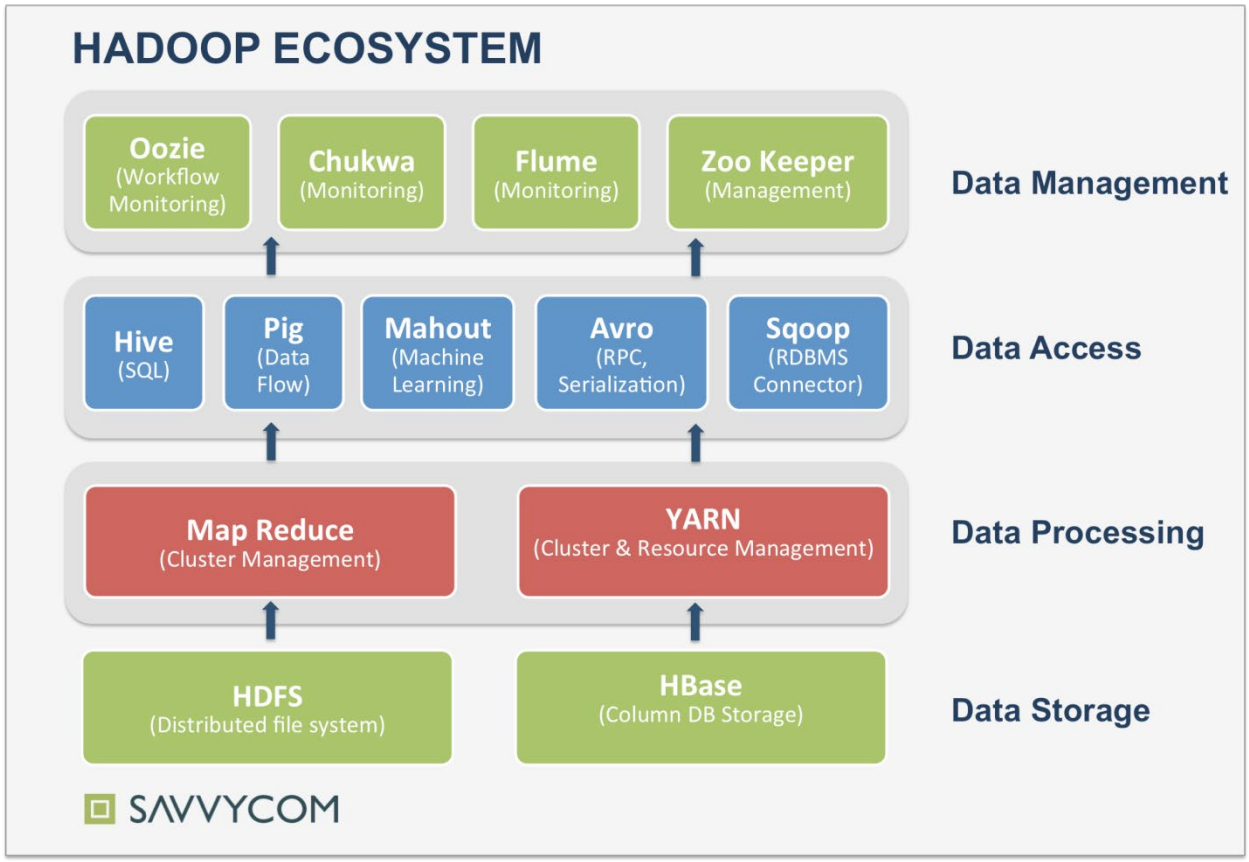
Most of the computing takes place on nodes with data on local disks that reduces the network traffic.

After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.

Q.2 Define Hadoop Ecosystem?

Ans. *Hadoop Ecosystem* is a platform or a suite which provides various services to solve the big data problems. It includes Apache projects and various commercial tools and solutions. There are *four major elements of Hadoop* i.e. **HDFS, MapReduce, YARN, and Hadoop Common**. Most of the tools or solutions are used to supplement or support these major elements. All these tools work collectively to provide services such as

absorption, analysis, storage and maintenance of data etc.



Following are the components that collectively form a Hadoop ecosystem:

- **HDFS:** Hadoop Distributed File System
- **YARN:** Yet Another Resource Negotiator
- **MapReduce:** Programming based Data Processing
- **Spark:** In-Memory data processing
- **PIG, HIVE:** Query based processing of data services
- **HBase:** NoSQL Database
- **Mahout, Spark MLLib:** Machine Learning algorithm libraries
- **Solar, Lucene:** Searching and Indexing
- **Zookeeper:** Managing cluster
- **Oozie:** Job Scheduling