1)Explain Hadoop architecture in detail and operators in?

Hadoop is an open source framework from Apache and is used to store process and analyze data which are very huge in volume. Hadoop is written in Java and is not OLAP (online analytical processing). It is used for batch/offline processing.It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more

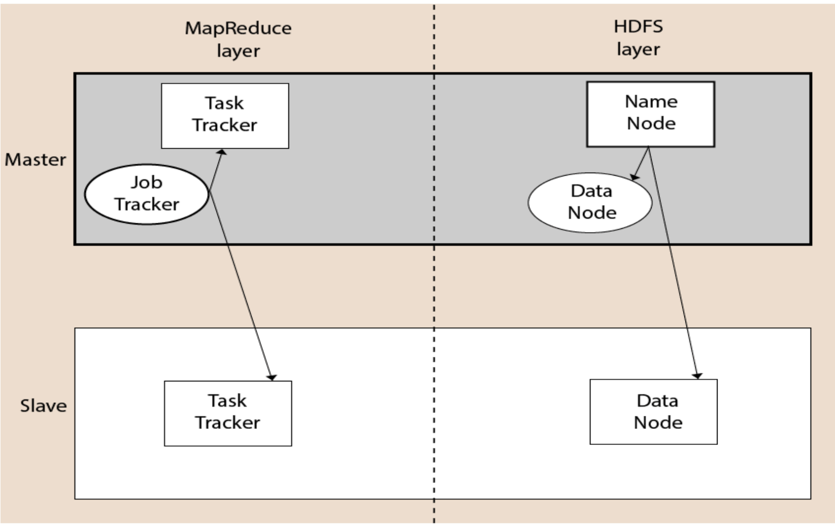
Modules of Hadoop -1)**HDFS:** Google published its paper GFS and on the basis of that HDFS was developed.

2)Yarn: Yet another Resource Negotiator is used for job scheduling and manage the cluster.

3)Map Reduce: This is a framework which helps Java programs to do the parallel computation on data using key value pair.

4)Hadoop Common: These Java libraries are used to start Hadoop and are used by other Hadoop modules.

Hadoop Architecture---The Hadoop architecture is a package of the file system, MapReduce engine and the HDFS (Hadoop Distributed File System). The MapReduce engine can be MapReduce/MR1 or YARN/MR2. A Hadoop cluster consists of a single master and multiple slave nodes. The master node includes Job Tracker, Task Tracker, NameNode, and DataNode whereas the slave node includes DataNode and TaskTracker.



2) Explain features of Hive and application of Hive?

Features of Apache Hive-

1. Open-source: Apache Hive is an open-source tool. We can use it free of cost.
2. Query large datasets: Hive can query and manage huge datasets stored in Hadoop Distributed File System.
3. Multiple-users: Multiple users can query the data using Hive Query Language simultaneously.
4. Backward compatible: Apache Hive perfectly fits the low level interface requirement of Apache Hadoop.
5. Partitioning and Bucketing: Apache Hive supports partitioning and bucketing of data at the table level to improve performance.
6. File-formats: Hive provides support for various file formats such as textFile, ORC, Avro Files, SequenceFile, Parquet, RCFile, LZO Compression etc.
7. Hive Query Language: Hive uses Hive Query Language which is similar to SQL. We do not require any knowledge of programming languages to work with Hive
8. Built-In function: Hive provides various Built-In functions.
9. User-Defined Functions: It also provides support for User-Defined Functions for the tasks like data cleansing and filtering. We can define UDFs according to our requirements

When to use Hive

- Most suitable for data warehouse applications where relatively static data is analyzed.
- Fast response time is not required.
- Data is not changing rapidly.
- An abstraction to underlying MR program.
- Hive of course is a good choice for queries that lend themselves to being

3) Expalin HBase storage mechanism and HBase architecture?

HBase architecture has 3 main components: HMaster, Region Server, Zookeeper.

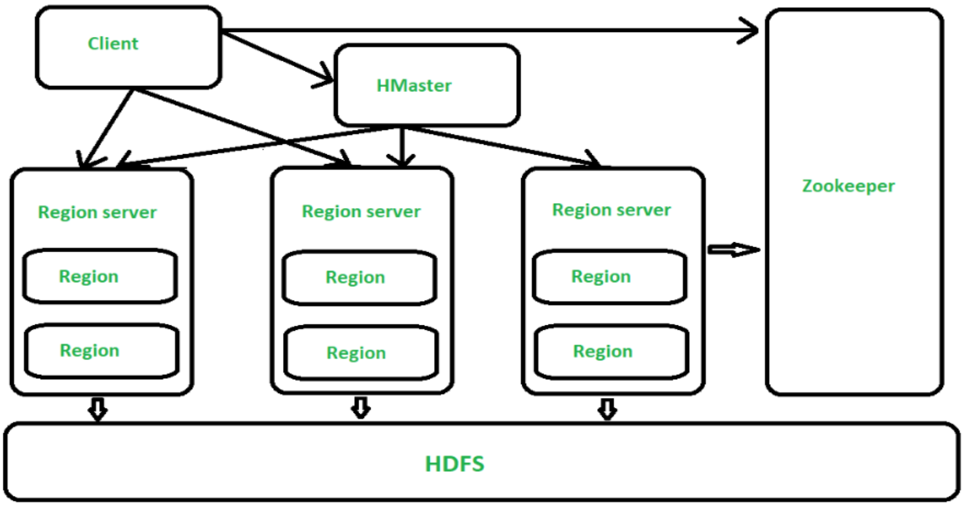


Figure – Architecture of HBase

1. **HMaster** – The implementation of Master Server in HBase is HMaster. It is a process in which regions are assigned to region server as well as DDL (create, delete table) operations. It monitor all Region Server instances present in the cluster HMaster has many features like controlling load balancing, failover etc.

2. **Region Server** – HBase Tables are divided horizontally by row key range into Regions. **Regions** are the basic building elements of HBase cluster that consists of the distribution of tables and are comprised of Column families. Region Server runs on HDFS DataNode which is present in Hadoop cluster. Regions of Region Server are responsible for several things, like handling, managing, The default size of a region is 256 MB.

3. **Zookeeper** – \It is like a coordinator in HBase. It provides services like maintaining configuration information, naming, providing distributed synchronization, server failure notification etc. Clients communicate with region servers via zookeeper.

4)Define Bigdata ? Explain characteristics and need of big data?,

•The definition of big data is data that contains greater variety, arriving in increasing volumes and with more velocity.

•Put simply, big data is larger, more complex data sets, especially from new data sources.

Characteristics Of Big Data

(i) Volume – The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.

(ii) Variety – The next aspect of Big Data is its **variety**. Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications.

(iii) Velocity – The term **'velocity'** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

(iv) Variability – This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

Need of Big Data

•To understand Where, When and Why their customers buy

•Protect the company's client base with improved loyalty programs

•Grabbing cross-selling and upselling opportunities

•Provide targeted promotional information

•Optimize Workforce planning and operations

5) What is NOSQL ? Explain NOSQL storage architecture in detail.

When people use the term “NoSQL database”, they typically use it to refer to any non-relational database. Some say the term “NoSQL” stands for “non SQL” while others say it stands for “not only SQL”. Either way, most agree that NoSQL databases are databases that store data in a format other than relational tables.

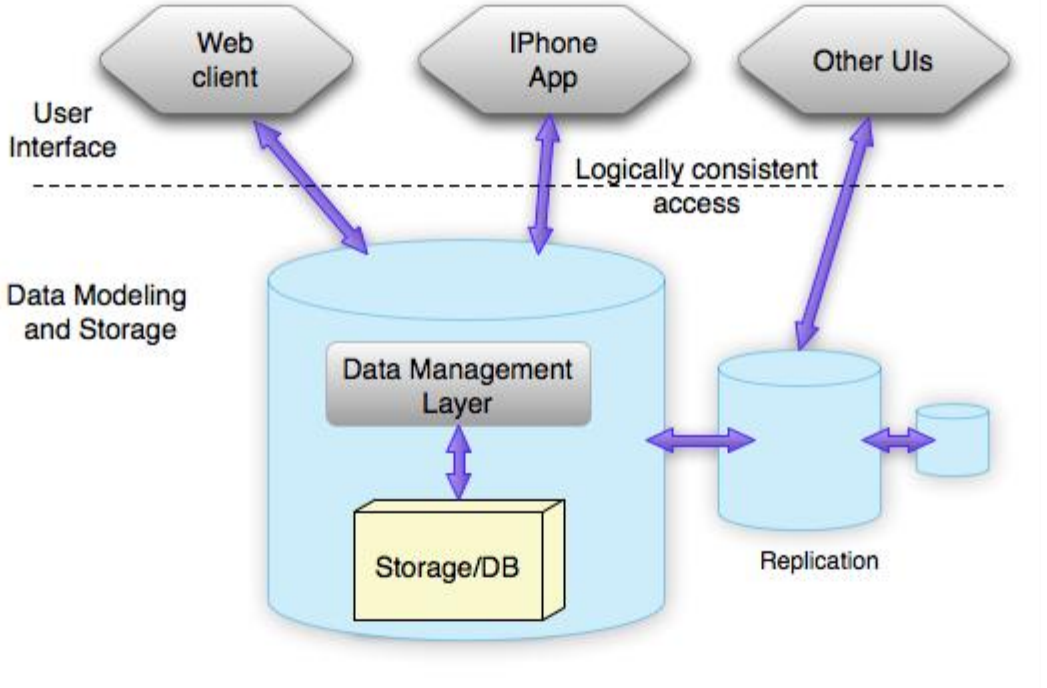
•NoSQLdatabases are generally classified into four main categories:

•Document databases: These databases store data as semi-structured documents, such as JSON or XML, and can be queried using document-oriented query languages.

•Key-value stores: These databases store data as key-value pairs, and are optimized for simple and fast read/write operations.

•Column-family stores: These databases store data as column families, which are sets of columns that are treated as a single entity. They are optimized for fast and efficient querying of large amounts of data.

•Graph databases: These databases store data as nodes and edges, and are designed to handle complex relationships between data.



6)Write difference between RDBMS and NOSQL.

1. Relational Database :
RDBMS stands for Relational Database Management Systems. It is most popular database. In it, data is store in the form of row that is in the form of tuple. It contain numbers of table and data can be easily accessed because data is store in the table. This Model was proposed by E.F. Codd.

2. NoSQL :
NoSQL Database stands for a non-SQL database. NoSQL database doesn’t use table to store the data like relational database. It is used for storing and fetching the data in database and generally used to store the large amount of data. It supports query language and provides better performance.

Relational Database	NoSQL
It is used to handle data coming in low velocity.	It is used to handle data coming in high velocity.
It gives only read scalability.	It gives both read and write scalability.
It manages structured data.	It manages all type of data.
Data arrives from one or few locations.	Data arrives from many locations.
It supports complex transactions.	It supports simple transactions.
It has single point of failure.	No single point of failure.
It handles data in less volume.	It handles data in high volume.
Transactions written in one location.	Transactions written in many locations.
support ACID properties compliance	doesn’t support ACID properties
Its difficult to make changes in database once it is defined	Enables easy and frequent changes to database

7)What is Hive? Explain features and applictions of Hive.

What is HIVE

Hive is a data warehouse system which is used to analyze structured data. It is built on the top of Hadoop. It was developed by Facebook.

Hive provides the functionality of reading, writing, and managing large datasets residing in distributed storage. It runs SQL like queries called HQL (Hive query language) which gets internally converted to MapReduce jobs.

Using Hive, we can skip the requirement of the traditional approach of writing complex MapReduce programs. Hive supports Data Definition Language (DDL), Data Manipulation Language (DML), and User Defined Functions (UDF).

Hive's Features

- Hive is fast and scalable.
- It provides SQL-like queries (i.e., HQL) that are implicitly transformed to MapReduce or Spark jobs.
- It is capable of analyzing large datasets stored in HDFS.
- It allows different storage types such as plain text, RCFile, and HBase.
- It uses indexing to accelerate queries.
- It can operate on compressed data stored in the Hadoop ecosystem.

Applicatons

- It supports user-defined functions (UDFs) where user can provide its functionality.
- Most suitable for data warehouse applications where relatively static data is analyzed.
- Fast response time is not required.
- Data is not changing rapidly.
- An abstraction to underlying MR program.

8)Application of big Data

Big Data Applications refer to the various ways in which vast volumes of data are collected, analysed, and utilised to extract valuable insights and drive decision-making across diverse industries. These applications encompass fields such as healthcare, finance, marketing, manufacturing, transportation, education, and more.

Big Data in e-commerce -The e-commerce industry has experienced a paradigm shift with the advent of Big Data Applications. E-commerce platforms now have access to vast amounts of customer data, including browsing history, purchase behaviour, and preferences.

Big Data in social media -Big Data Applications have become the lifeblood of social media platforms, shaping how users interact and engage with content. By analysing vast amounts of user data, including preferences, behaviour, and interactions, social media platforms can deliver personalised content and advertisements that resonate with users.

Big Data in education -In the education sector, Big Data has opened up new possibilities for personalised learning and educational improvements. By analysing data on student performance, attendance, and engagement, educators can gain insights into individual learning needs and adapt their teaching methods accordingly.

Big Data in healthcare - Big Data's transformative impact on the healthcare industry has been nothing short of revolutionary. With the digitisation of patient records and the proliferation of advanced medical devices

Big Data in finance **-** The finance industry has embraced Big Data Applications as a powerful tool for gaining valuable insights into customer behaviour and financial markets

Big Data in marketing - The marketing domain has undergone a seismic shift with the integration of Big Data Applications.

9) What is HDFS

Hadoop comes with a distributed file system called HDFS. In HDFS data is distributed over several machines and replicated to ensure their durability to failure and high availability to parallel application.

Apache Hadoop Distributed File System (HDFS) is a distributed file system that handles large data sets running on commodity hardware. It is used to scale a single Apache Hadoop cluster to hundreds (and even thousands) of nodes.

Where to use HDFS

Very Large Files: Files should be of hundreds of megabytes, gigabytes or more.

Streaming Data Access: The time to read whole data set is more important than latency in reading the first. HDFS is built on write-once and read-many-times pattern.

Commodity Hardware: It works on low cost hardware.

HDFS Concepts

Blocks: A Block is the minimum amount of data that it can read or write.HDFS blocks are 128 MB by default and this is configurable.Files n HDFS are broken into block-sized chunks,which are stored as independent units.Unlike a file system,

Name Node: HDFS works in master-worker pattern where the name node acts as master.Name Node is controller and manager of HDFS as it knows the status and the metadata of all the files in HDFS;

Data Node: They store and retrieve blocks when they are told to; by client or name node. They report back to name node periodically, with list of blocks that they are storing.

10)Explain Architecture of Pig and features of Pig ?

Hadoop stores raw data coming from various sources like IOT, websites, mobile phones, etc. and preprocessing is done in Map-reduce. Pig framework converts any pig job into Map-reduce hence we can use the pig to do the ETL (Extract Transform and Load) process on the raw data. Apache pig can handle large data stored in Hadoop to perform data analysis and its support file formats like text, CSV, Excel, RC, etc.

1. Parser: Any pig scripts or commands in the grunt shell are handled by the parser. Parse will perform checks on the scripts like the syntax of the scripts, do type checking and perform various other checks

2. Optimizer: As soon as parsing is completed and DAG is generated, It is then passed to the logical optimizer to perform logical optimization like projection and pushdown.

3. Compiler: The optimized logical plan generated above is compiled by the compiler and generates a series of Map-Reduce jobs.

4. Execution Engine: Finally, all the MapReduce jobs generated via compiler are submitted to Hadoop in sorted order

Features of Pig

1) **Ease of programming**-Writing complex java programs for map reduce is quite tough for non-programmers. Pig makes this process easy.

2) **Optimization opportunities-It** is how tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.

3) **Extensibilityn** -A user-defined function is written in which the user can write their logic to execute over the data set.

4) **Flexible** -It can easily handle structured as well as unstructured data.

11) MangoDB

MongoDB is an open source NoSQL database management program. NoSQL (Not only SQL) is used as an alternative to traditional relational databases. NoSQL databases are quite useful for working with large sets of distributed data. MongoDB is a tool that can manage document-oriented information, store or retrieve information.

MongoDB is used for high-volume data storage, helping organizations store large amounts of data while still performing rapidly. Organizations also use MongoDB for its ad-hoc queries, indexing, load balancing, aggregation, server-side JavaScript execution and other features.

Structured Query Language (SQL) is a standardized programming language that is used to manage relational databases. SQL normalizes data as schemas and tables, and every table has a fixed structure.

Features of MongoDB

Replication. A replica set is two or more MongoDB instances used to provide high availability. Replica sets are made of primary and secondary servers

Scalability. MongoDB supports vertical and horizontal scaling. Vertical scaling works by adding more power to an existing machine, while horizontal scaling works by adding more machines to a user's resources.

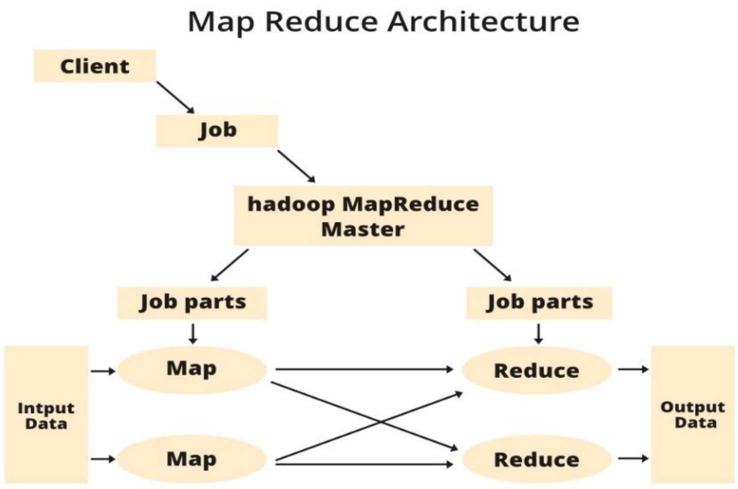
Load balancing. MongoDB handles load balancing without the need for a separate, dedicated load balancer, through either vertical or horizontal scaling.

Schema-less. MongoDB is a schema-less database, which means the database can manage data without the need for a blueprint.

12)Map Reduce

MapReduce is a Java-based, distributed execution framework within the Apache Hadoop Ecosystem. It takes away the complexity of distributed programming by exposing two processing steps that developers implement: 1) Map and 2) Reduce. In the Mapping step, data is split between parallel processing tasks. Transformation logic can be applied to each chunk of data. Once completed, the Reduce phase takes over to handle aggregating data from the Map set.. In general, MapReduce uses Hadoop Distributed File System (HDFS) for both input and output.

MapReduce Architecture:



Client: The MapReduce client is the one who brings the Job to the MapReduce for processing.
Job: The MapReduce Job is the actual work that the client wanted to do which is comprised of so many smaller tasks that the client wants to process or execute.
Hadoop MapReduce Master: It divides the particular job into subsequent job-parts.
Job-Parts: The task or sub-jobs that are obtained after dividing the main job.
Input Data: The data set that is fed to the MapReduce for processing.
Output Data: The final result is obtained after the processing.
\

13)Hadoop daemons

1. NameNode NameNode works on the Master System. The primary purpose of Namenode is to manage all the MetaData. Metadata is the list of files stored in HDFS(Hadoop Distributed File System). As we know the data is stored in the form of blocks in a Hadoop cluster. All information regarding the logs of the transactions happening in a Hadoop cluster (when or who read/wrote the data) will be stored in MetaData. MetaData is stored in the memory.
2. DataNode - DataNode works on the Slave system. The NameNode always instructs DataNode for storing the Data. DataNode is a program that runs on the slave system that serves the read/write request from the client. As the data is stored in this DataNode, they should possess high memory to store more Data.
3. Secondary NameNode - Secondary NameNode is used for taking the hourly backup of the data. In case the Hadoop cluster fails, or crashes, the secondary Namenode will take the hourly backup or checkpoints of that data and store this data into a file name *fsimage*. This file then gets transferred to a new system. A new MetaData is assigned to that new system and a new Master is created with this MetaData, and the cluster is made to run again correctly.
4. Resource Manager - Resource Manager is also known as the Global Master Daemon that works on the Master System. The Resource Manager Manages the resources for the applications that are running in a Hadoop Cluster. The Resource Manager Mainly consists of 2 things. **A. ApplicationsManager**
B. Scheduler
5. Node Manager
The Node Manager works on the Slaves System that manages the memory resource within the Node and Memory Disk. Each Slave Node in a Hadoop cluster has a single NodeManager Daemon running in it. It also sends this monitoring information to the Resource Manager.