DEPARTMENT OF CSE

**SUBJECT CODE : U23CST61**

**SUBJECT NAME : BIG DATA ANALYTICS**

**YEAR/SEM : III/V**

**STAFF NAME : ARUN G & RAMYA K**

## UNIT-1

## UNDERSTANDING BIG DATA

### Part A

**1.What is big data analytics?**

Big Data analytics is a process used to extract meaningful insights, such as hidden patterns, unknown correlations, market trends, and customer preferences.

Big Data analytics provides various advantages—it can be used for better decision making, preventing fraudulent activities, among other things.

**2.What is big data?**

Big Data is a massive amount of data sets that cannot be stored, processed, or analyzed using traditional tools.

**For example**, in a regular Excel sheet, data is classified as structured data—with a definite format. In contrast, emails fall under semi-structured, and your pictures and videos fall under unstructured data. All this data combined makes up Big Data.

**3. Why is big data analytics important**?

Big data analytics assists organizations in harnessing their data and identifying new opportunities. As a result, smarter business decisions are made, operations are more efficient, profits are higher, and customers are happier.

**4. How does big data analytics work?**

Gather information. Once data has been collected and saved, it must be correctly organized in order to produce reliable answers to analytical queries, especially when the data is huge and unstructured. Then, clean and analyze the data.

**5. Who uses big data analytics?**

Industries that include big data analytics are Banking and Securities, Healthcare Providers, Communications, Media and Entertainment, Education, Government, Retail and Wholesale trade, Manufacturing Natural Resources, and Insurance.

**6. What are the five types of big data analytics?**

The five types of big data analytics are Prescriptive Analytics, Diagnostic Analytics, Cyber Analytics, Descriptive Analytics, and Predictive Analytics.

7**. What are the 3 types of big data?**

There are three types of big data: Data that is **structured**, Data that is **unstructured**, and Data that is **semi-structured.**

**8. What are the advantages of big data?**

Businesses can tailor products to customers based on big data instead of spending a fortune on ineffective advertising. Businesses may use big data to study consumer patterns by tracking POS transactions and internet purchases.

**9. Why do we need big data analytics?**

Organizations may harness their data and utilize big data analytics to findnew possibilities. This results in wiser company decisions, more effective operations, more profitability, and happier clients. Businesses that employ big data and advanced analytics benefit in a variety of ways, including cost reduction.

**10. What is big data in simple words?**

Big data is a collection of large, complex, and voluminous data that traditional data management tools cannot store or process.

**11. What is the meaning of big data analytics?**

Big data analytics refers to the complex process of analyzing big data to reveal information such as correlations, hidden patterns, market trends, and customer preferences.

.

**12.What are the applications of Big data**?

- Transportation
- Advertising and Marketing
- Banking and Financial Services
- Government
- Media and Entertainment
- Meteorology
- Healthcare
- Cyber security and Education

**13.What are the industry examples of big data?**

- Retail
- Banking
- Manufacturing
- Education
- Government
- Health Care

**14.What are 4 types of big data technologies?**

- Big data technologies can be categorized into four main types:
- data storage,
- data mining,
- data analytics, and
- data visualization

**15.What is unstructured data?**

Unstructured data is the data which does not conforms to a data model and has no easily identifiable structure such that it cannot be used by a computer program easily. Unstructured data is not organised in a pre-defined manner or does not have a pre-defined data model, thus it is not a good fit for a mainstream relational database

**16.What are the advantages and disadvantages of unstructured data?**

**Advantages of Unstructured Data:**

- Its supports the data which lacks a proper format or sequence
- The data is not constrained by a fixed schema
- Very Flexible due to absence of schema.
- Data is portable and very scalable
- It can deal easily with the heterogeneity of sources.

**Disadvantages of Unstructured data:**

- It is difficult to store and manage unstructured data due to lack of schema and structure and Ensuring security to data is difficult task.

**17.What is web analytics?**

Web analytics is the process of analyzing the behavior of visitors to a website. This involves tracking, reviewing and reporting data to measure web activity, including the use of a website and its components, such as webpages, images and videos.

**18.What is Hadoop?**

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed *storage* and *computation* across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

**19.What is mapreduce?**

MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The MapReduce program runs on Hadoop which is an Apache open-source framework.

**20. What are the advantages and disadvantages of Hadoop?**

**Advantages of Hadoop**

- Hadoop framework allows the user to quickly write and test distributed systems.
- It is efficient, and it automatic distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.

**Disadvantages of Hadoop:**

- Not very effective for small data.
- Hard cluster management.
- Has stability issues and Security concerns

## PART B

**1.Explain in detail about the architecture of Hadoop and HDFS.**

- Definition
- Hadoop Architecture
- Hadoop Distributed File System
- *Modules and Features of Hadoop*
- Hadoop Distributed File System
- Advantages of Hadoop
- Disadvantages:

**2.Explain in detail about industry examples of big data.**

- Definition
- Examples of big data.
    1.) Retail
    2.) Banking
    3.) Manufacturing
    4.) Education
    5.) Government andHealth Care

**3.Explain in detail about Unstructured data**

- Definition
- Characteristics of Unstructured Data:
- Sources of Unstructured Data:
- Problems faced in storing unstructured data:
- Possible solution for storing Unstructured data:
- Extracting information from unstructured Data:
- Advantages of Unstructured Data.
- Disadvantages of Unstructured data:

**4.Define web analytics and explain it.**

- Definition
- Steps of web analytics
- Main categories of web analytics
- Web analytics tools and examples

**5.Explain in detail about open source technologies.**

- Opensource technologies (Cassandra,Greenplum,MongoDB, CouchDB,MariaDB)
- Features
- Pros
- Cons

## UNIT-2
## NOSQL DATA MANAGEMENT

### PART A

**1.Define NoSQL database**.

NoSQL Database is used to refer a non-SQL or non relational database.It provides a mechanism for storage and retrieval of data other than tabular relations model used in relational databases.NoSQL database doesn't use tables for storing data.It is generally used to store big data and real-time web applications.

**2.Define aggregate data models**

The term aggregate means a collection of objects that we use to treat as a unit. An aggregate is a collection of data that we interact with as a unit. These units of data or aggregates form the boundaries for ACID operation.

**3.Define key-value data model**.

A key-value data model or database is also referred to as a key-value store. It is a non-relational type of database.

**4.Define graph databases**.

A graph database is a type of NoSQL database that is designed to handle data with complex relationships and interconnections.In a graph database, data is stored as nodes and edges, where nodes represent entities and edges represent the relationships between those entities.

**5.Define schemaless databases**

A schemaless database, like MongoDB, does not have these up-front constraints, mapping to a more 'natural' database. Even when sitting on top of a data lake, each document is created with a partial schema to aid retrieval.Any data, formatted or not, can be stored in a non-tabular NoSQL type of database. At the same time, using the right tools in the form of a schemaless database can unlock the value of all of your structured and unstructured data types.

**6.Define materialized views**

A materialized view takes the regular view described above and materializes it by proactively computing the results and storing them in a "virtual" table.A view can be "materialized" by storing the tuples of the view in the database. Index structures can be built on the materialized view.

**7.Define distribution models**

Aggregate oriented databases make distribution of data easier, since the distribution mechanism has to move the aggregate and not have to worry about related data, as all the related data is contained in the aggregate. There are two styles of distributing data:Sharding and Replication:

**8.Define Cassandra**

Cassandra deals with unstructured data. Cassandra has a flexible schema.Tables or column families are the entity of a keyspace. Row is a unit of replication in Cassandra. Column is a unit of storage in Cassandra.

**9.Define Cassandra clients.**

The Cassandra client is used to connect, manage and develop your Cassandra database. The database client is used to manage your Cassandra database with actions like insert, delete and update table

**10.What is Apache Cassandra?**

Apache Cassandra is a highly scalable, high-performance distributed database designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure. It is a type of NoSQL database

**11.List the types of NOSQL databases**.

Types of NoSQL database: Types of NoSQL databases and the name of the databases system that falls in that category are:

- **Graph Databases**: Examples – Amazon Neptune, Neo4j

**Key value store**: Examples – Memcached, Redis, Coherence

**Tabular**: Examples – Hbase, Big Table, Accumulo

**Document based**: Examples – MongoDB, CouchDB, Cloudant

**12.What are the components of Cassandra?**

Node ,Data center ,Cluster ,Commit log ,Mem-table SSTable ,Bloom filter

**13.Differentiate between relational database and NOSQL database.**

| Relational Database | NoSql Database |
|---|---|
| Supports powerful query language. | Supports very simple query language. |
| It has a fixed schema. | No fixed schema. |
| Follows ACID (Atomicity, Consistency, Isolation, and Durability). | It is only “eventually consistent”. |
| Supports transactions. | Does not support transactions. |

**14.What is sharding and replication**.

**Sharding**: Sharding distributes different data across multiple servers, so each server acts as the single source for a subset of data.

**Replication**: Replication copies data across multiple servers, so each bit of data can be found in multiple places.

**15.What are the types of replication?**

Replication comes in two forms,

**Master-slave replication** makes one node the authoritative copy that handles writes while slaves synchronize with the master and may handle reads.

**Peer-to-peer replication** allows writes to any node; the nodes coordinate to synchronize their copies of the data.

**16.Define consistency .**

Data Consistency in DBMS is the state of data that is consistent across all systems. Integrity ensures that the data is correct. Consistency ensures that the data format is correct, or that the data is correct with respect to other data.

**17.What is SPOF?**

A single point of failure (SPOF) is a part of a system that, if it fails, will stop the entire system from working. SPOFs are undesirable in any system with a goal of high availability or reliability, be it a business practice, software application, or other industrial system.

**18.Differentiate between RDBMS and Cassandra.**

| RDBMS | Cassandra |
|---|---|
| RDBMS deals with structured data and fixed schema | Cassandra deals with unstructured data and flexible schema |
| In RDBMS, a table is an array of arrays. (ROW x COLUMN) | In Cassandra, a table is a list of "nested key-value pairs". (ROW x COLUMN key x COLUMN value) |
| Database is the outermost container that contains data corresponding to an application. | Keyspace is the outermost container that contains data corresponding to an application. |
| Tables are the entities of a database. | Tables or column families are the entity of a keyspace. |

**19.What is keyspace in Cassandra?**

Keyspace is the outermost container for data in Cassandra. The basic attributes of a Keyspace in Cassandra are – Replication factor,placement strategy,column families.

**20.Is Cassandra a SQL database.**

Cassandra is designed to be scalable and can handle large amounts of data. Cassandra is a NoSQL database, which means it does not use the traditional table structure found in SQL databases. This can make Cassandra more flexible and easier to use for certain types of data

## PART B

**1.Explain in detail about NOSQL databases**.

- Definition
- Introduction and key features of NoSQL
- Features of NoSQL Databases
- Types of NoSQL database
- Advantages of NoSQL
- Disadvantages of NoSQL

**2.Explain in detail about Aggregate Data Model.**

- Definition
- Example of Aggregate Data Model:
- Consequences of Aggregate Orientation
- Advantage
- Disadvantage

**3.Write short notes on key-value and document data models**

- Definition
- Features:
- Advantages:
- Disadvantages:
- Some examples **of key-value databases:**
    1. Couchbase
    2. Amazon DynamoDB
    3. Riak**.**
    4. Aerospike
    5. Berkeley DB

**4.Explain detail about graph databases**

- Definition:
- Types of Graph Databases:
- Property Graphs:
- RDF Graphs:
- Example of Graph Database:
- Advantages of Graph Database:
- Disadvantages of Graph Database:
- Future of Graph Database**:**

.

**5.Explain in detail about Cassandra data model.**

- *Definition*
- *Components of Cassandra*
- Column Family,Super Column
- *Data Models of Cassandra and RDBMS*
- Cassandra clients
- NoSQL vs. Relational Database
- Apache Cassandra
- Features of Cassandra

## UNIT-III
## MAP REDUCE APPLICATIONS

### Part A

**1.What is map reduce?**

Map Reduce is a terminology that comes with **Map Phase** and **Reducer Phase**. The map is used for Transformation while the Reducer is used for aggregation kind of operation. The terminology for Map and Reduce is derived from some functional programming languages like Lisp, Scala, etc.

**2.What are the components of map reuce?**

1. our **Driver code**,
2. **Mapper**(For Transformation), and
3. **Reducer**(For Aggregation).

**3. What is shuffling and sorting**?

The data are **shuffled** between/within nodes so that it moves out from the map and get ready to process for reduce function. Sometimes, the shuffling of data can take much computation time.The **sorting** operation is performed on input data for Reduce function.

**4. What is reduce function?**

The Reduce function is assigned to each unique key. These keys are already arranged in sorted order. The values associated with the keys can iterate the Reduce and generates the corresponding output

**5. What is YARN?**

YARN stands for "***Yet Another Resource Negotiator***". It was introduced in Hadoop 2.0 to remove the bottleneck on Job Tracker which was present in Hadoop 1.0. YARN architecture basically separates resource management layer from the processing layer

**6. What is MRunit?**

MRUnit is a JUnit-based Java library that allows us to unit test Hadoop MapReduce programs. This makes it easy to develop as well as to maintain Hadoop MapReduce code bases. MRUnit supports testing Mappers and Reducers separately as well as testing MapReduce computations as a whole.

7**. What are the features of YARN?**

**YARN Features:** YARN gained popularity because of the following features

- Scalability
- Compatibility
- Cluster Utilization
- Multi-tenancy

**8. What are the three types of failures in mapreduce?**

- Task failure
- TaskTracker Failure
- JobTracker Failure

**9. What are the reasons for task failures?**

- **Limited memory:** A task can fail if it runs out of memory while processing data
- **Failures of disk:** If the disk that stores data or intermediate results fails, tasks that depend on that data may fail.

**10. What are the types of job scheduling in mapreduce?**

- FIFO
- Capacity scheduler
- Job scheduler

**11. What are the types of Mapreduce?**

- Input formats
- Output formats

**12.What is mapping?**

*Mapping* is the core technique of processing a list of data elements that come in pairs of keys and values. The map function applies to individual elements defined as key-value pairs of a list and produces a new list.

.

**13.Define data locality.**

Map-Reduce comes with a feature called **Data-Locality**. Data Locality is the potential to move the computations closer to the actual data location on the machines.

**14.What are the industry examples of big data?**

- Retail
- Banking
- Manufacturing
- Education
- Government
- Health Care

**15.What are the data flow in mapreduce?**

- Input reader
- Map function,
- Partition function
- Shuffling and sorting,
- Reduce function and
- Output writer

**16.How to call mapreduce job?**

To run this job ,with a single method call,**submit()** on a Job object (you can likewise call **waitForCompletion( )**, which presents the activity on the off chance that it hasn't been submitted effectively, at that point sits tight for it to finish).

**17.What is shuffling and sorting?**

**Shuffling** is the process by which it transfers **mappers** intermediate output to the **reducer.** Reducer gets 1 or more keys and associated values on the basis of

reducers. The intermediated key – value generated by mapper is sorted automatically by key. In Sort phase merging and sorting of map output takes place.

**18.What is heartbeat in bigdata?**

A 'heartbeat' is a signal sent between a Data Node and Name Node. This signal is taken as a sign of vitality. If there is no response to the signal, then it is understood that there are certain health issues/ technical problems with the Data Node or the Task Tracker.

**19.What is scaling out in big data?**

Horizontal scaling, or scaling out or in, where you add more databases or divide your large database into smaller nodes, using a data partitioning approach called sharding which can be managed faster and more easily across servers.

**20.What is Hadoop streaming?**

Hadoop streaming is a utility that comes with the Hadoop distribution. The utility allows you to create and run Map/Reduce jobs with any executable or script as the mapper and/or the reducer.

**PART B**

**1.Explain in detail about Hadoop MapReduce framework.**

- Definition
- Mapreduce
- Steps of Data flow diagram
- Brief Working of Mapper
- Brief Working Of Reducer

**2.Explain in detail about the architecture of YARN .**

- Definition
- YARN
  - Scalability: .
  - Compatibility:
  - Cluster Utilization:.
  - Multi-tenancy
- Hadoop YARN Architecture

The main components of YARN architecture include:

1.Client
2.Resource Manager
3.Node Manager
4.Application Master
5.Container

- Advantages
- Disadvantages

**3. Discuss the failures in classic Map-reduce**

- Definition
- Failures in map reduce

- Task failure
- Tasktracker failure
- Jobtracker failure
- Reasons for task failure
- How to overcome Task failure.

4. **Discuss about Job Scheduling in MapReduce**

- ***Definition***
- Map reduce algorithm
- Hadoop Schedulers
- Types of job scheduling in Mapreduce
  - FIFO
  - Capacity scheduler
  - Fair Scheduler
- Advantages
- Disadvantages

**5.Explain in details about MapReduce types.**

- Definition:
- MapReduce Types
- The Java API:
- Input Formats
- *Output Formats*

## *UNIT IV*

## *BASICS OF HADOOP*

### *PART A*

***1.Define Hadoop.***

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

**2.What are the three components of Hadoop?**

- Hadoop HDFS - Hadoop Distributed File System (HDFS) is the storage unit of Hadoop.
- Hadoop MapReduce - Hadoop MapReduce is the processing unit of Hadoop.
- Hadoop YARN - Hadoop YARN is a resource management unit of Hadoop.

***3.What are the MODULES OF HADOOP?***

Hadoop is made up of 4 core modules:

- the Hadoop Distributed File System (HDFS),
- Yet *Another* Resource Negotiator (YARN),
- Hadoop Common and
- MapReduce

**4.What is scaling out?**

It is also referred as "scale out" is basically the addition of more machines or setting up the cluster. In horizontal scaling instead of increasing hardware capacity of individual machine you add more nodes to existing cluster and most importantly, you can add more machines without stopping the system.

**5.What are the types of Scaling ?**

**Vertical scaling** : When we add more resources to a single machine when the load increases. For example you need 20gb of ram but currently your server has 10 GB of ram so you add extra ram to the same server to meet the needs.

**Horizontal scaling** of scaling out : when you add more machines to match the resources need it's called horizontal scaling. So if I have a machine of already 10 GB I'll add an extra machine with 10 GB ram.

**6.Define Hadoop streaming** –

Hadoop Streaming is a part of the Hadoop Distribution System. It facilitates ease of writing Map Reduce programs and codes. Hadoop Streaming supports almost all types of programming languages such as Python, C++, Ruby, Perl etc. The entire Hadoop Streaming framework runs on Java.

**7.Define Avro.**

Avro is a language-neutral data serialization system. It was developed by Doug Cutting, the father of Hadoop. Since Hadoop writable classes lack language portability, Avro becomes quite helpful, as it deals with data formats that can be processed by multiple languages. Avro is a preferred tool to serialize data in Hadoop.

***8. What is Hadoop Pipes?***

Hadoop Pipes is a C++ API that enables developers to write MapReduce programs for Hadoop. The API allows the use of existing C++ libraries with Hadoop and enables the development of high-performance MapReduce programs. Hadoop Pipes is a popular choice for developers who need to process large data sets efficiently.

**9.Define DFS**.

DFS stands for the distributed file system, it is a concept of storing the file in multiple nodes in a distributed manner. DFS actually provides the Abstraction for a single large system whose storage is equal to the sum of storage of other nodes in a cluster.

**10. What are the types of File-based data structures?**

Two file formats:

1, Sequencefile

2, MapFile

**11.What is java interface?**

In Java, an interface specifies the behavior of a class by providing an abstract type. As one of Java's core concepts, abstraction, polymorphism, and multiple inheritance are supported through this technology. Interfaces are used in Java to achieve abstraction

**12.What is compression?**

Java object compression is done using the GZIPOutputStream class (this class implements a stream filter for writing compressed data in the GZIP file format) and passes it to the ObjectOutputStream class (this class extends OutputStream and implements ObjectOutput, ObjectStreamConstants) .

**13.What is serialization?**

**Serialization in Java** is a mechanism of *writing the state of an object into a byte-stream*. It is mainly used in Hibernate, RMI, JPA, EJB and JMS technologies

**14.Define Hadoop integration**.

Hadoop architecture is designed to be easily integrated with other systems. Integration is very important because although we can process the data efficiently in Hadoop, but we should also be able to send that result to another system to move the data to another level.

**15.What are the advantages and disadvantages of Hadoop?**

**Advantages:**

- Cost
- Flexibility
- Speed
- scalability
- Fault tolerance

**Disadvantages:**

- Vulnerability
- Lack of security
- High up processing
- Supports only batch processing

**16.What are the modes of operations.**

- Standalone Mode
- Pseudo-distributed Mode
- Fully-Distributed Mode

**17.What are the types of schedulers in Hadoop?**

- FIFO (First In First Out) Scheduler.
- Capacity Scheduler.
- Fair Scheduler.

**18. Where is Hadoop used in real life?**

Hadoop can be used to analyze transaction data to detect fraudulent activities and improve risk management. It can also be used to analyze market data to identify investment opportunities and make informed trading decisions

**19.Which databse is used in hadoop?**

Hadoop is not a type of database, but rather a software ecosystem that allows for massively parallel computing. It is an enabler of certain types NoSQL distributed databases (such as HBase), which can allow for data to be spread across thousands of servers with little reduction in performance.

**20.What language that Hadoop does?**

Java programming languageThe Hadoop framework itself is mostly written in the Java programming language, with some native code in C and command line utilities written as shell scripts.

## PART B

**1.Explain in detail about file based data structure.**

- Definition
- File-based data structures
  Two file formats:
  1, Sequence file
  2, Map File
- Read/write Sequence file/Map file
- Benefits of the Sequence file/Map file file format:
- Disadvantages of the Sequence file /Map file file format:

**2.Explain in detail about integration of Hadoop**.

- Definition
- R integration with Hadoop and Diagram
- *R Hadoop Integration Method*
- Hadoop Streaming
- RHIPE, ORCH

**3.Explain in detail about HDFS architecture**.

- Definition
- Types
- Read and write operation in HDFS
- Hadoop Streaming and pipes
- Design of Hadoop distributed file system (HDFS)
- Distributed File System Processing:
- Overview – HDFS
- *Read / Write Operation In HDFS*

**4.Explain in detail about Hadoop I/O.**

- Definition
- Data integrity
- Hadoop local file system
- Compression
- Serialization
- Avro

**5.Explain in detail about Hadoop streaming and Pipes**

- Definition
- Features and Code execution process
- Hadoop pipes
- Execution of streaming and pipes

## UNIT-V

## HADOOP RELATED TOOLS

## PART A

**1.Define Hbase.**

Hbase is an open source and sorted map data built on Hadoop. It is column oriented and horizontally scalable.It is based on Google's Big Table

**2.Define hbase clients.**

It describes the java client API for HBase that is used to perform CRUD operations on HBase tables. HBase is written in Java and has a Java Native API. Therefore it provides programmatic access to Data Manipulation Language (DML).

**3.Define pig latin data model.**

Pig Latin data model allows Pig to handle any kind of data. Pig Latin data model is fully nested and can treat both atomic like integer, float, and non-atomic complex data types such as Map and tuple.

**4.What is pig latin?**

Pig is a high-level data flow platform for executing Map Reduce programs of Hadoop. It was developed by Yahoo. The language for Pig is pig Latin.

**5.Define pig latin Scripts**.

The Pig scripts get internally converted to Map Reduce jobs and get executed on data stored in HDFS. Apart from that, Pig can also execute its job in Apache Tez or Apache Spark.

**6.What is pig latin statements?**

The Pig Latin statements are used to process the data. It is an operator that accepts a relation as an input and generates another relation as an output.

**7.What are the types of modes in pig?**

Apache Pig executes in two modes: Local Mode and MapReduce Mode.

**8.Define HiveQL queries.**

HiveQL is a query language for Hive to analyze and process structured data in a Meta-store. It is a mixture of SQL-92, MySQL, and Oracle's SQL. It is very much similar to SQL and highly scalable.

**9.Define 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.

**10.What are steps involved in pig latin componets?**

- Parser
- Optimizer
- Compiler
- Execution engine

**11.Why Pig is used in big data?**

Pig Represents Big Data as data flows. Pig is a high-level platform or tool which is used to process the large datasets. It provides a high-level of abstraction for processing over the MapReduce. It provides a high-level scripting language, known as Pig Latin which is used to develop the data analysis codes

**12.What are the features of pig data model?**

Allows programmers to write fewer lines of codes.Apache Pig multi-query approach reduces the development time.Apache pig has a rich set of datasets for performing operations like join, filter, sort, load, group, etc.Pig Latin language is very similar to SQL.

**13.Define HiveQL data definition language**.

Hive Data Definition Language (DDL) is a subset of Hive SQL statements that describe the data structure in Hive by creating, deleting, or altering schema objects such as databases, tables, views, partitions, and buckets. Most Hive DDL statements start with the keywords CREATE , DROP , or ALTER .

**14.Define HiveQL data manipulation language**

Hive DML (Data Manipulation Language) commands are used to insert, update, retrieve, and delete data from the Hive table once the table and database schema has been defined using Hive DDL commands. The various Hive DML commands are: LOAD.

**15.What are the types of file formats in Hive?**

- TEXTFILE.
- SEQUENCEFILE.

- RCFILE.
- ORC.
- PARQUET.
- AVRO.

**16.What are the clients of HBase?**

Kundera – the object mapper. The REST client. The Thrift client. The Hadoop ecosystem client.

**17.What are HiveQueries?**

Hive enables data summarization, querying, and analysis of data. Hive queries are written in HiveQL, which is a query language similar to SQL. Hive allows you to project structure on largely unstructured data. After you define the structure, you can use HiveQL to query the data without knowledge of Java or MapReduce.

**18.What is Pig Grunt?**

Apache Pig Grunt is an interactive shell that enables users to enter Pig Latin interactively and provides a shell to interact with HDFS and local file system commands. You can enter Pig Latin commands directly into the Grunt shell for execution.

**19. What is Praxis in big data?**

Praxis is driven by the purpose of creating resources that will lead India's transformation into the tech and data-driven digital world.

**20.What are the data types is used in HIVE?**

- Column Types
- Literals
- Null Values
- Complex Types

**PART –B**

**1.Explain in detail about the architecture of HBASE with neat diagram.**

- Definition
- Architecture diagram
- Features
- Hbase Read and Write
- Advantages and Disadvantages of HBASE.

**2.Explain in detail about the architecture of PIG with neat diagram.**

- Definition
- Architecture diagram
- Features

- Pig latin ,Conventions, Statements
- Advantages and Disadvantages of PIG

.

**3.Explain in detail about the architecture of HIVE with neat diagram.**

- Definition
- Architecture diagram
- Features
- Limitations
- Advantages and Disadvantages of HIVE

**4.Explain in detail about the data types of HIVE with examples.**

- Definition
- Types
- Syntax and examples.

**5.Explain in detail about the Pig latin statements and conventions with examples.**

- Definition
- Types
- Tabular column
- Syntax and examples

**********

.