Assigment -4

① what is Apache pig? Explain this features of pig.

Apache pig

Apache pig is a high-level platform for processing and analyzing large datasets on Apache Hadoop Developed by yahoo, it simplifies complex data tranformations using a high-level scripting language known as pig latin. pig provides an abstraction over the complexity of writing map Reduce programs, allowing data analyts and developers to perform data procesing it's commonly used in big data scenario for data extraction tranformation and loading (ETL) procesing.

Features

pig latin: A simple scripting language that allows users to to write data transformation without needing to be an expert in Java. pig latin is based on two componets: The language itelf, which provides operators for processing data and the runtime environment.

parallel execution: pig programs are structured to allow for substantial parallelization, which enables them to handle large data sets.

optimization: The system automatically optimizes the execution of tasks, allowing users to focus on semantics rather than efficiency

extensibility: users can create their own functions to do special-purpose procesing. These function can be written in Java, python, Javascript, Ruby or Groovy

Handles various data types: pig can use both structured and unstructured data as input.

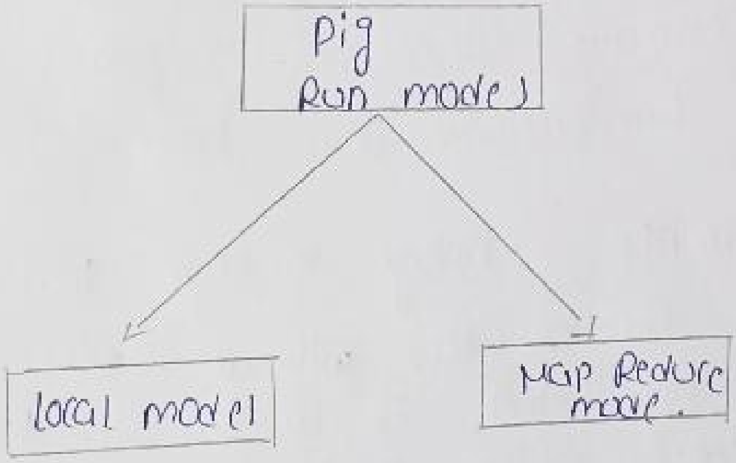
HDFS storage: pig uses HDFS to store the results of its analysis.

2) Explain the Apache pig Run modes with a neat diagram.

Apache pig executes in two models: They are

(1) Local mode

(2) Map Reduce mode.



Local Mode

* It executes in a single JVM and is used for development experimenting and prototyping
* Here files are installed and run using localhost.
* The local mode works on a local file system. The input and output data stored in the local file system.

The command for local mode grunt shell:

$pig-x local.

map Reduce mode:

* The map Reduce mode is also known as Hadoop mode

* it is the default mode.

* Here, the input and output data are present on HDFS

The command for map reduce mode

```
$ pig
 or
$pig -x mapreduce
```

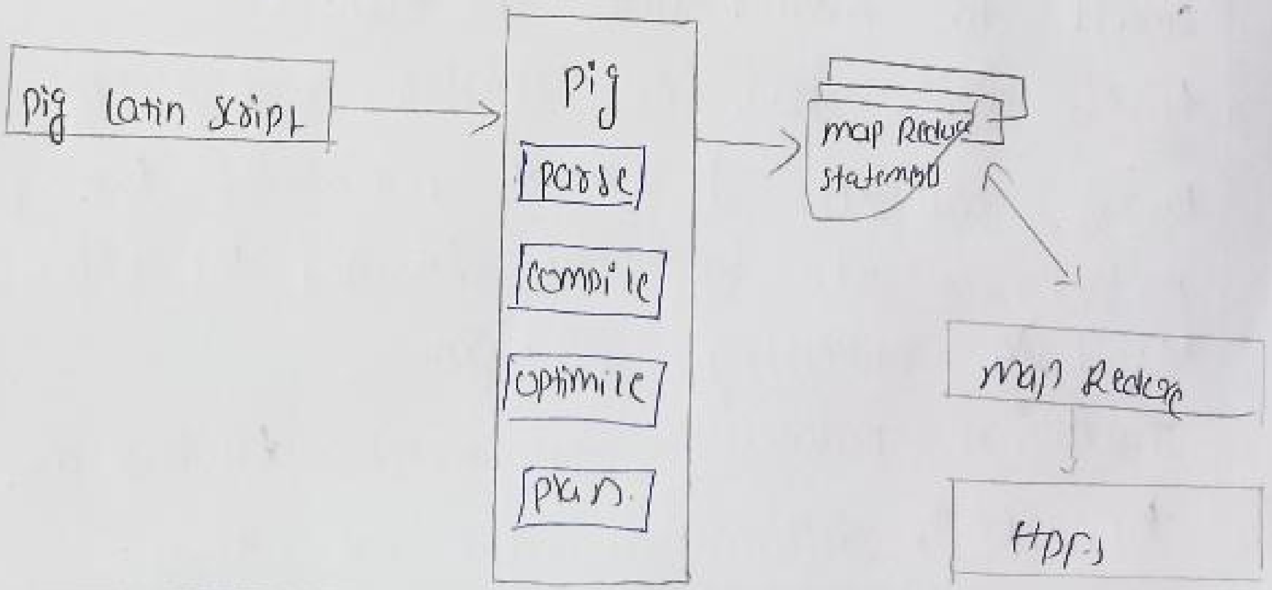
way to execute pig program

These are following ways of executing a pig program on local and mapReduce mode.

intractive mode: In this mode, the pig is executed in the Grunt shell. To invoke Grunt shell, run the pig command.

Batch mode: In this mode, we can run a script file having a .pig extension. These files contain pig latin commands.

embedded mode: In this mode, we can define our own functions. These functions can be called UDF, Here we use programing language like Java and Python.



3) compare Hive with Traditional database?

Comparing Hive with Traditional Databases

1. schema flexibility:-

Traditional Database: Traditional databases, often relational, require a well-defined schema upfront. Schema changes can be complex and may disrupt ongoing operations.

Hive: Hive provides schema-on-read, allowing users to define the structure of data during execution. This flexibility is advantageous when dealing with unstructured or semi-structured data.

2. Qery language:

Traditional Database: Relational databases use SQL as the standard query language.

Hive: Hive uses HQL, which closely resembles SQL

3. performance:

Traditional databases: Traditional Databases are optimized for transactional operations and perform well for small to medium-sized datasets.

Hive: while Hive is scalable and suitable for large datasets, it may not match the real time performance of traditional databases for ad-hoc queries.

4. Data processing paradigm:

Traditional Databases: Traditional database are optimized for OLAP, depending on the use case.

Hive: Hive is well-suited for batch processing and data warehousing scenario.

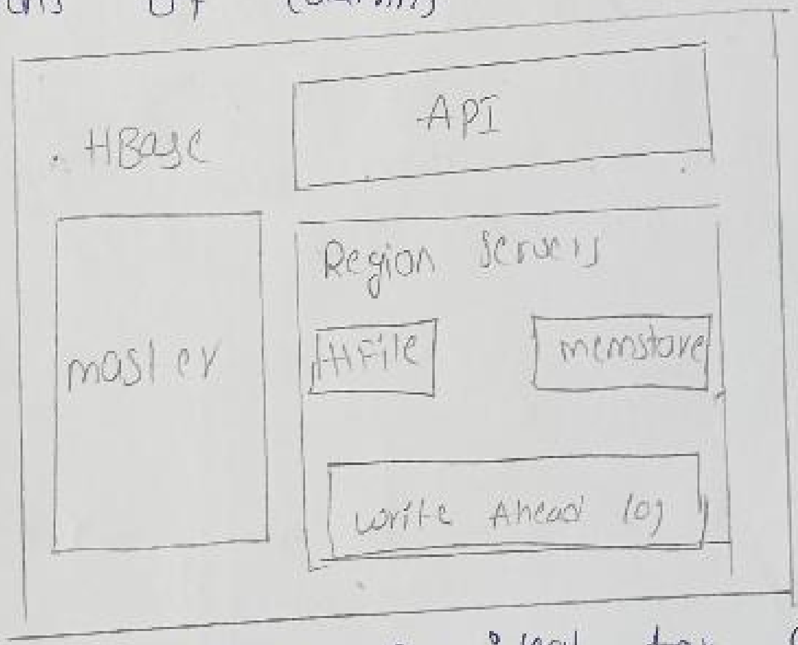
Use cases

Traditional Databases. Traditional databases excel in scenario where data is well-structured and real-time processing is critical

Hive: Hive is ideal for scenarios involving large-scale data processing, log analysis, social media analytics, and other Bigdata use cases.

4) explain HBase, clients, praxis, Zookeeper?

1. HBase: HBase is a distributed, scalable, NOSQL database built on top of Hadoop's HDFS. it's designed to handle large tables with billions of rows and millions of columns

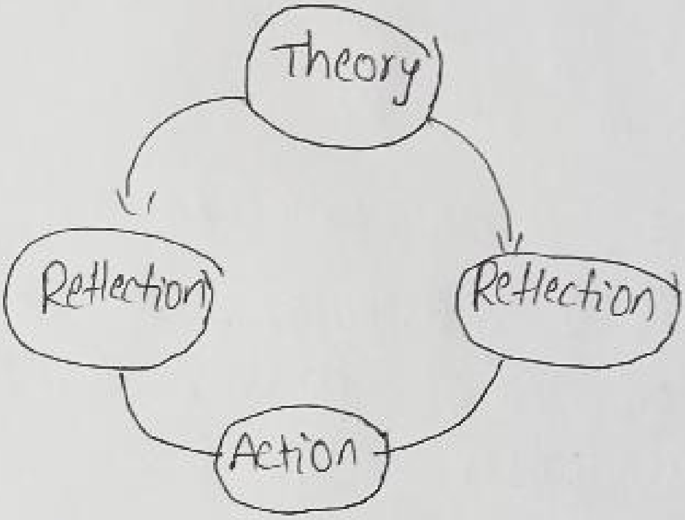


Applications: HBase is ideal for applications that requi real-time analytics.

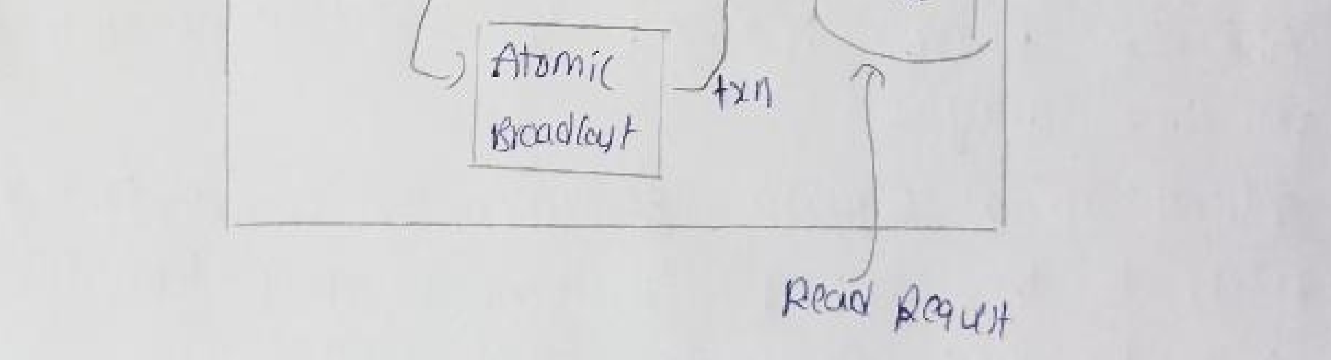
clients in a bigdata ecosystem, a "client" usually refers to the various applications or tools that interact with the data store or other components in the cluster

Types: HBase client can be used to perform CRUD (Create, Read, update, Delete) operations on data stored in HBase tables.

3. praxis: praxis is a broader term in big data environments, generally refering to best practices or practical application techniques in big data.



4. zookeeper: Apache zookeeper is a centralized service used to manage service used to manage and maintain configuration information, naming and provide distributed synchronication and group services.


zookeper service
write request
Request processor
txn
Atomic Broadcast
txn
Replicated database
response
Read Request