What is Apache Pig

Apache Pig is a high-level data flow platform for executing MapReduce programs of Hadoop. The language used for Pig is Pig Latin.

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.

Pig can handle any type of data, i.e., structured, semi-structured or unstructured and stores the corresponding results into Hadoop Data File System. Every task which can be achieved using PIG can also be achieved using java used in MapReduce.

Features of Apache Pig

Let's see the various uses of Pig technology.

1) Ease of programming

Writing complex java programs for map reduce is quite tough for non-programmers. Pig makes this process easy. In the Pig, the queries are converted to MapReduce internally.

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) Extensibility

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.

5) In-built operators

It contains various type of operators such as sort, filter and joins.

Differences between Apache MapReduce and PIG

| **Apache MapReduce** | **Apache PIG** |
| --- | --- |
| It is a low-level data processing tool. | It is a high-level data flow tool. |
| Here, it is required to develop complex programs using Java or Python. | It is not required to develop complex programs. |
| It is difficult to perform data operations in MapReduce. | It provides built-in operators to perform data operations like union, sorting and ordering. |
| It doesn't allow nested data types. | It provides nested data types like tuple, bag, and map. |

Advantages of Apache Pig

* Less code - The Pig consumes less line of code to perform any operation.
* Reusability - The Pig code is flexible enough to reuse again.
* Nested data types - The Pig provides a useful concept of nested data types like tuple, bag, and map.

# 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).

## Features of Hive

These are the following features of Hive:

* 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.
* It supports user-defined functions (UDFs) where user can provide its functionality.

## Limitations of Hive

* Hive is not capable of handling real-time data.
* It is not designed for online transaction processing.
* Hive queries contain high latency.

## Differences between Hive and Pig

| **Hive** | **Pig** |
| --- | --- |
| Hive is commonly used by Data Analysts. | Pig is commonly used by programmers. |
| It follows SQL-like queries. | It follows the data-flow language. |
| It can handle structured data. | It can handle semi-structured data. |
| It works on server-side of HDFS cluster. | It works on client-side of HDFS cluster. |
| Hive is slower than Pig. | Pig is comparatively faster than Hive |