
VS
Hive
Pig

As we all know, Hadoop uses MapReduce to process and analyze big data. Processing big data consumed more time using traditional methods; Hadoop MapReduce was used to process big data faster

Before

After

Processing Big Data consumed more time

Processing Big Data was faster using Mapreduce

As we all know, Hadoop uses MapReduce to process and analyze big data. Processing big data consumed more time using traditional methods; Hadoop MapReduce was used to process big data faster

Before

After

MapReduce is primarily implemented using Java codes. Lengthy complex codes were written by programmers to process data

Processing Big Data consumed more time

Processing Big Data was faster using Mapreduce

- Need for Hive & Pig
- What is Hive & Pig
- HiveQL & Pig Latin
- Data Models
- Execution Modes
- Features
- Commands

Need for Hive

Problem

Facebook found it hard to process and analyze big data as not all the employees were well versed with high-level coding languages

Solution

They required a language similar to SQL, which was easier to write. Hence, Hive was developed with a vision to include the concepts of tables, columns just like SQL

Need for Pig

Problem

Similarly, Yahoo also found it hard to process and analyze big data using MapReduce as not all the employees were well versed with complex Java codes

Solution

There was a necessity to process data using a language which was easier than Java. Yahoo researchers developed Pig, which was used to process data quickly and easily

- Need for Hive & Pig
- What is Hive & Pig
- HiveQL & Pig Latin
- Data Models
- Execution Modes
- Features
- Commands

What is Hive?

Hive is a data warehouse system which is used for analyzing large datasets stored in HDFS. Hive uses a query language called HiveQL which is similar to SQL

HiveQL

MapReduce tasks

What is Pig?

Hive is a data warehouse system which is used for analyzing large datasets stored in HDFS. Hive uses a query language called HiveQL which is similar to SQL

Uses SQL like queries

Analyze data

- Need for Hive & Pig
- What is Hive & Pig
- HiveQL & Pig Latin
- Data Models
- Execution Modes
- Features
- Commands

HiveQL

- Hive Query Language (HiveQL) is a query language used by Hive to process and analyze data
- Declarative language which is exactly similar to SQL
- HiveQL works on structured data

Pig Latin

- Pig Latin is the procedural data flow language used in Pig to analyze data
- Pig Latin is similar to SQL but varies greatly
- It is used for structured, semi-structured and unstructured data. 10 lines of Pig Latin code = 200 lines in Java

- Need for Hive & Pig
- What is Hive & Pig
- HiveQL & Pig Latin
- Data Models
- Execution Modes
- Features
- Commands

Hive Data Model

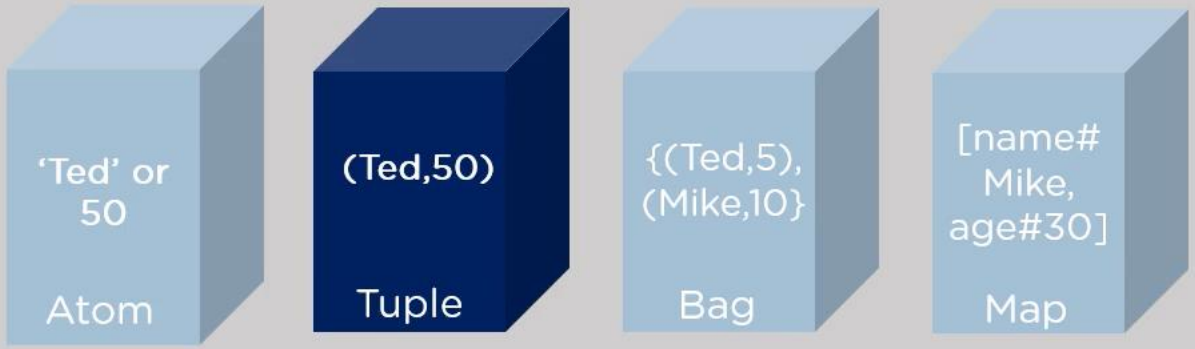
Tables

Partitions

Buckets

Partitions are further divided into buckets for better querying

Pig Latin Data Model



Tuple represents sequence of fields that can be of any data type. It is same as a row in RDBMS

- Need for Hive & Pig
- What is Hive & Pig
- HiveQL & Pig Latin
- Data Models
- Execution Modes
- Features
- Commands

Hive Execution modes

Hive operates in two modes depending on the number and size of data nodes

Local Mode

MapReduce Mode

It is used when there are multiple datanodes and the data is large

Pig Execution Modes

Depending on where the data is residing and where the Pig script is going to run, Pig works in two modes

Local Mode

MapReduce Mode

In this mode, queries written in Pig Latin are translated into MapReduce jobs and are run on a Hadoop cluster. Pig runs on this mode by default

- Need for Hive & Pig
- What is Hive & Pig
- HiveQL & Pig Latin
- Data Models
- Execution Modes
- Features
- Commands

- Used by analysts

- Used by programmers and researchers

- Need for Hive & Pig
- What is Hive & Pig
- HiveQL & Pig Latin
- Data Models
- Execution Modes
- Features
- Commands

- Used by analysts
- HiveQL is the language used

HiveQL

- Used by programmers and researchers
- Pig Latin is the language used

Pig Latin

- Need for Hive & Pig
- What is Hive & Pig
- HiveQL & Pig Latin
- Data Models
- Execution Modes
- Features
- Commands

- Used by analysts
- HiveQL is the language used
- Works on structured data. Does not work on other types of data

- Used by programmers and researchers
- Pig Latin is the language used
- Works on structured, semi-structured and unstructured data

- Need for Hive & Pig
- What is Hive & Pig
- HiveQL & Pig Latin
- Data Models
- Execution Modes
- Features
- Commands

- Used by analysts
- HiveQL is the language used
- Works on structured data. Does not work on other types of data
- Works on the server side of the cluster

- Used by programmers and researchers
- Pig Latin is the language used
- Works on structured, semi-structured and unstructured data
- Works on the client side of the cluster

- Need for Hive & Pig
- What is Hive & Pig
- HiveQL & Pig Latin
- Data Models
- Execution Modes
- Features
- Commands

- Used by analysts
- HiveQL is the language used
- Works on structured data. Does not work on other types of data
- Works on the server side of the cluster
- Hive does not support Avro

- Used by programmers and researchers
- Pig Latin is the language used
- Works on structured, semi-structured and unstructured data
- Works on the client side of the cluster
- Pig supports Avro

- Need for Hive & Pig
- What is Hive & Pig
- HiveQL & Pig Latin
- Data Models
- Execution Modes
- Features
- Commands

- Used by analysts
- HiveQL is the language used
- Works on structured data. Does not work on other types of data
- Works on the server side of the cluster
- Hive does not support Avro
- Hive supports partitions
- Hive has web interface

- Used by programmers and researchers
- Pig Latin is the language used
- Works on structured, semi-structured and unstructured data
- Works on the client side of the cluster
- Pig supports Avro
- Pig does not support partitions although there is an option for filtering
- Pig does not support web interface

Few Hive Commands

- create database database_name // used to create a new database
- show databases; //shows the list of existing databases
- Now, to create a table inside the database

  create table table_name(ID INT, Name STRING, DEPT STRING, YOJ INT) row format delimited fields terminated by ',';
- show tables; //Gives list of the created table
- hive> SELECT round(2.3) from temp; //Rounds off the value to the nearest highest integer -> 2.3 – 2
- hive> SELECT floor(2.3) from temp; //Rounds off any positive or negative decimal value down to the next least integer value -> 2.3 – 2
- hive> SELECT ceil(2.3) from temp; //This function is used to get the smallest integer which is greater than, or equal to, the specified numeric expression -> 2.3 - 3

- Need for Hive & Pig
- What is Hive & Pig
- HiveQL & Pig Latin
- Data Models
- Execution Modes
- Features
- Commands

Few Pig Commands

- hadoop dfs -put 'path_name' /pigInput //For file to be moved into HDFS
- pig // To start the grunt shell mode
- relation1 = LOAD '/pigInput' USING PigStorage(',') AS (Id:chararray,Name:chararray,Profession:chararray,Age:chararray); //Loads the file from HDFS into Pig
- dump relation1; //The results from the previous load command is displayed using dump
- relation1_filter = filter relation1 by column_name == 'attribute_name';
- dump relation1_filter; //Filter command shows the result for that particular filter that we give


THANK YOU