

Apache PIG

Notes By Mannan UI Haq

Introduction

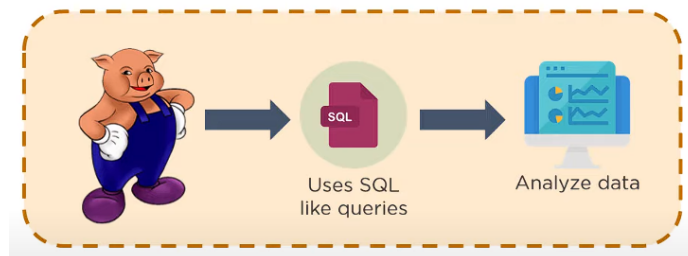
- Yahoo used Pig as a solution to handle the growing big data.
- As we know, Hadoop uses MapReduce for processing data. MapReduce required users to write long codes.
- Not all users were well versed with Java and other coding languages. This proved to be a disadvantage for them.
- They faced issues in incorporating map, sort, reduce fundamentals of MapReduce while creating a program.

What is Pig?

Pig is a scripting platform that runs on Hadoop clusters, designed to process and analyze large datasets using a high level language called Pig Latin.

Pig was designed for performing a long series of data operations, making it ideal for three categories of Big Data jobs:

1. Extract-transform-load (ETL) data pipelines,
2. Research on raw data, and
3. Iterative data processing



Benefits

1. **Ease of Use:** Pig's declarative language lets users focus on data processing logic rather than low-level implementation details.
2. **Extensibility:** Users can enhance Pig's functionality by writing custom functions in Java, Python, or other languages, enabling them to extend Pig's functionality.
3. **Optimization:** Pig automatically optimizes execution plans, and it offers tools for performance tuning, helping to improve the efficiency and speed of data processing tasks.

Integration With Hadoop

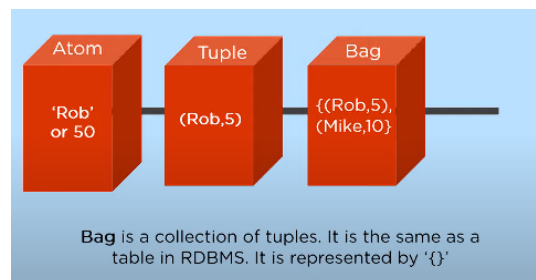
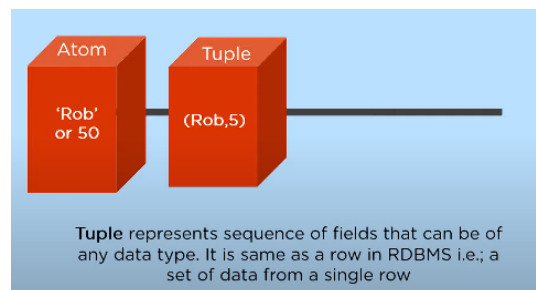
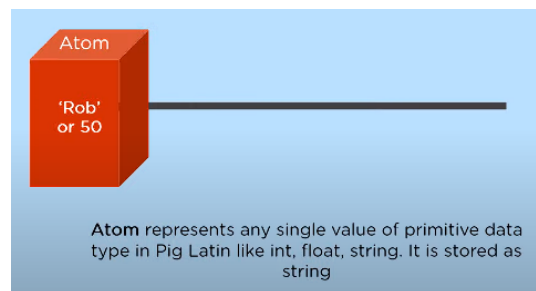
1. **Pig and MapReduce:** Pig translates Pig Latin scripts into MapReduce jobs, enabling them to execute on Hadoop clusters. This integration leverages Hadoop's distributed processing capabilities for scalable data processing tasks.
2. **Integration with Hive:** Pig seamlessly integrates with Hive, allowing users to perform SQL-like operations on structured data stored in HDFS.
3. **Pig and Spark:** Apache Pig can also run on Apache Spark, enabling in-memory data processing. This integration offers faster performance for certain workloads, especially those that benefit from Spark's distributed computing and caching capabilities.

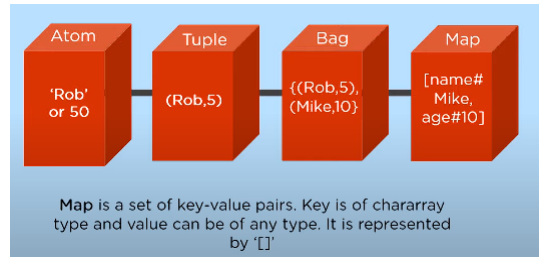
Pig Data Types

Primitive Types:

int, long, float, double, chararray, bytearray

Complex Types:








Basic Operations

1. LOAD
2. STORE
3. FILTER
4. GROUP
5. JOIN

MapReduce vs Hive vs Pig

 hadoop Map Reduce	VS	 HIVE	VS	
Compiled language		SQL like query		Scripting language
Need to write long complex codes		No need to write complex codes		No need to write complex codes
Can process structured, semi structured and unstructured data		Can process only structured data		Can process structured, semi structured and unstructured data
Lower level of abstraction		Higher level of abstraction		Higher level of abstraction



VS



VS



Supports partitioning feature	Supports partitioning feature	No concept of partitioning in Pig
MapReduce uses Java and Python	Hive uses a SQL like query language known as HiveQL	Pig Latin is used which is a procedural data flow language
MapReduce is used by programmers	Hive is used by data analysts	Pig is used by researchers and programmers
Code performance is good	Code performance is lesser than MapReduce and Pig	Code performance is lesser than MapReduce but better than Hive