Pig

MODULE - VI

SOUVIK

Introduction to Pig

- Pig is a high level Platform for processing and analyzing large datasets in Apache Hadoop.
- It Provides a Simple Scripting language called Pig Latin for expressing data transformations.
- Pig enables developers to write complete data Processing workflows without writing Java code.

Pig → Uses SQL like queries → Analyze Data.

→ Need for Pig

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

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

○ Grunt Shell

- Grunt is the interactive Shell Provided by Pig for executing Pig Latin Scripts and commands interactively.
- Developers can use Grunt to Prototype Pig Scripts, explore data, and test Queries before running them into Production.

○ Pig Data Model

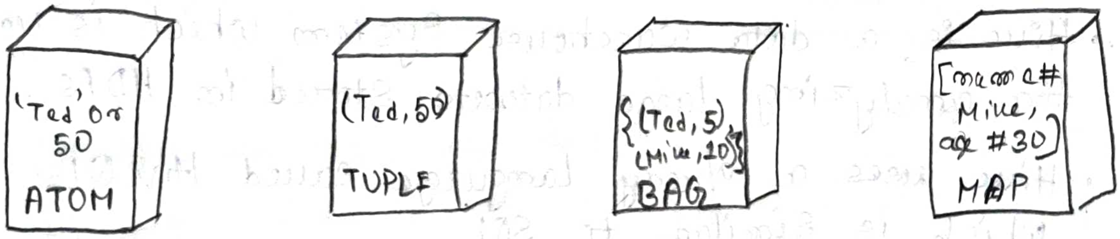
- Pig represents data as a collection of tuples (rows) and bags (Collection of tuples), allowing for Structured and Semi-Structured Data Processing.
- Example
  - '(1, 'John', 25)' represents a tuple with three fields: ID, name and Age.
  - '{(1, 'John', 25), (2, 'Alice', 30)}' represents a bag containing two tuples.

○ Pig Latin

- Pig Latin is a 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.
- It Provides operators for loading, filtering, grouping, joing and aggregating data.
- Example
  - 'Student_data = LOAD 'Students.csv' USING PigStorage(',') AS (id : int, name : chararray, age : int);
  - 'filtered_data = FILTER Student_data BY age > 18;'

Pig Latin Data Model



- Atom is a single Value of Primitive data type like int, float, String. It is always stored as string

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

- Bag is a collection of tuples. It is the same as a table in RDBMS, it is represented by '{}'

- Map is a set of key value pairs. Key is of char array type and value can be of any type. It is represented by '[]'.

○ Developing and Testing Pig Scripts

- Developers can develop and test Pig Scripts using IDEs like Apache Zeppelin or integrated development environments like Eclipse with Pig Plugins

- Unit testing frameworks like PigUnit enable developers to test Pig Scripts localy before deploying them to a Production environment.

HIVE

Introduction to 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

Hive → HiveQL → MapReduce Tasks

→ Why need Hive

Facebook found it hard to Process and analyze big data as not all the employees were well versed with High Level coding Languages.

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

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.

Hive Data Model

Tables   Partitions   Buckets

- Tables in Hive are similar to those in RDBMS
- Tables are grouped into Partitions to group the same kind of data based on the Partition key.
- Partitions are further divided into Buckets for better querying.

○ Data Types and File Formats

- Hive supports various data types including Primitive types (int, String, boolean), ~~Complex~~ complex types (array, map, struct) and custom types.
- It supports file formats such as Text file, Sequence file, ORC and Parquet for storing and Processing data efficiently.

○ HiveQL Data Definition

- HiveQL allows users to define and manipulate data structures such as tables, Partitions and buckets.
- Example

'CREATE TABLE employee (id INT, name STRING, age INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';'

○ HiveQL Data Manipulation

- HiveQL Provides commands for inserting, updating and deleting data in tables. as well as for loading data from external Sources.
- Example

'INSERT INTO employee VALUES (1, 'John', 25);'

○ HiveQL Queries

- Users can write SQL-like queries in HiveQL to perform data analysis, aggregation, filtering and join operations on Hive tables.
- Example

'SELECT name COUNT (*) FROM employee GROUP BY name;'

Difference b/w Pig and Hive

Feature	PIG	HIVE
• Language	Pig Latin, a Scripting language	Hive QL (Hive Query Language), SQL-like.
• Use Case	ETL, data processing	Data Warehousing, SQL based analytics
• Data Model	Procedural, tuples and bags	Declarative, tables and columns
• Scripting Style	Imperative	Declarative.
• Interactivity	Interactive Shell (Grunt)	Interactive Hive Shell.
• Schema	Schema-on-Read	Schema-on-Write.
• Performance	Generally faster for iterative and ad-hoc processing	Generally optimized for SQL-based queries and batch processing.