Devanshu Surana
1032210755
Roll no : 50 (BDT)
Batch - 2

BDT Lab Assignment 8

Problem statement:
Create pig database and perform data analytics on it.

Objectives:
1. To learn pig concept
2. To perform data analytics on it.

Theory:
Pig is a high-level platform for processing and analyzing large dataset in Apache hadoop. It provide an abstraction over Hadoop MapReduce, making it easier to work large-scale data. Pig latin is the language used for writing Pig Scripts.

Pig Architecture:
Pig Latin: The scripting language for defining data transformation and analysis.
Pig Execution Environment: Pig scripts are executed. It supports local mode and MapReduce mode.
UDFs: Custom functions to perform specific data processing tasks.

Pig Latin Scripts

↓

Grant Shell   Pig Server

Parser

Optimiser

Compiler

Execution Engine

↓

Map Reduce

↓

HDFS

Pig Functions :

Load: loads data

Filter: Filters records based on condition

Group: Groups the data

Foreach: Applies operations to each records

Join : Combines Data.

Store : Saves data

Platform: 64-bit Open Source Windows.

Conclusion: Hence, I learned to create Pig Latin Program to perform data analytics.

FAQ's

1. Write a Pig Scripts to perform JOIN operation.

→ - orders = LOAD 'orders_data' USING Pigstorage (',') As (order_id.int, order_data, chararray, customer_id: int)'

- Customers = LOAD 'customers_data' USING Pigstorage. (',')

As (customer_id: int, customer_name: chararray);
- joined_data = JOIN orders BY customer_id, customers BY customer_id;
- Store joined_data INTO 'output_data';

2. Explain complex data types in Pig
→ Pig Tuple: An ordered set of fields
Example: ' (1, 'Alice');

- Bag: An unordered collection of tuples
Ex: '{(1, 'Alice'}, (2, 'Bob')};

- Map: A key value pair collection
Ex: ['name # Alice, id # 1];

3. state examples of Pig technology which can be used with hadoop.

Ans. - Pig can be used with hadoop through various mechanisms, including Pig on Tez, integration with Hcatalog for metadata custom UDFs and Pig storage functions for different data formats.

24/11/23