**PIG Commands - Lab Manual**

**Aim:**

To understand and implement data processing using Apache Pig, optimizing large-scale data handling in Hadoop.

**Understanding Apache Pig:**

Apache Pig is a high-level scripting language that translates scripts into MapReduce jobs, simplifying large-scale data analysis.

**1. Loading Data into Pig**

**Definition:** Loading data from HDFS into a Pig relation.

**Command:**

employees = LOAD '/hdfs\_path/employees.csv'

USING PigStorage(',')

AS (employee\_id:int, name:chararray, department:chararray, salary:float, year:int);

**2. Filtering Data**

**Definition:** Extracting specific rows based on conditions.

**Command:**

high\_salary = FILTER employees BY salary > 70000;

**3. Grouping Data**

**Definition:** Grouping records by a specific column.

**Command:**

grouped\_by\_dept = GROUP employees BY department;

**4. Aggregating Data**

**Definition:** Performing aggregations like sum, average, etc.

**Command:**

avg\_salary = FOREACH grouped\_by\_dept GENERATE group AS department, AVG(employees.salary) AS avg\_salary;

**Output:**

| **department** | **avg\_salary** |
| --- | --- |
| HR | 60000 |
| IT | 75000 |

**5. Storing Results**

**Definition:** Writing processed data back to HDFS.

**Command:**

STORE avg\_salary INTO '/hdfs\_path/output' USING PigStorage(',');

**result:**

Apache Pig simplifies large-scale data processing by converting scripts into MapReduce jobs. It provides an easy-to-use alternative to writing complex Java-based MapReduce programs, enhancing efficiency in big data analytics.