6. Implement a pig script to analyze the data using UDFs.

To analyze data using Pig (a high-level platform for creating programs that run on Apache Hadoop), you'll need to create a Pig script and User Defined Functions (UDFs) to perform specific tasks. Below is an example of how to do this:

**Example Data**

Let's assume we have a dataset called data.txt with the following format:

1,John,2000

2,Jane,3000

3,Bob,1500

4,Alice,2500

5,Tom,1200

The dataset represents records with the following fields: id, name, and salary.

### Pig Script

1. **Load the data**
2. **Register the UDF**
3. **Use the UDF to analyze the data**

#### Step 1: Write the UDF

We'll write a simple UDF in Java to calculate a bonus based on salary. For example, let's give a 10% bonus to each salary.

Create a Java file BonusCalculator.java:

import java.io.IOException;

import org.apache.pig.EvalFunc;

import org.apache.pig.data.Tuple;

public class BonusCalculator extends EvalFunc<Double> {

@Override

public Double exec(Tuple input) throws IOException {

if (input == null || input.size() == 0) {

return null;

}

try {

Double salary = (Double) input.get(0);

return salary \* 0.10;

} catch (Exception e) {

throw new IOException("Caught exception processing input row ", e);

}

}

}

Compile the Java code and package it into a JAR file.

#### Step 2: Write the Pig Script

Create a Pig script analyze\_data.pig:

-- Load the data

data = LOAD 'data.txt' USING PigStorage(',') AS (id:int, name:chararray, salary:double);

-- Register the UDF

REGISTER 'path/to/your/udf.jar';

-- Define the UDF

DEFINE BonusCalculator BonusCalculator();

-- Calculate the bonus

data\_with\_bonus = FOREACH data GENERATE id, name, salary, BonusCalculator(salary) AS bonus;

-- Store the result

STORE data\_with\_bonus INTO 'output' USING PigStorage(',');

### Steps to Execute

1. **Compile and package the UDF**

javac -cp /path/to/pig.jar BonusCalculator.java

jar -cvf udf.jar BonusCalculator.class

Run the Pig script

pig analyze\_data.pig

**Explanation**

1. **Loading Data**: The LOAD function reads data from data.txt.
2. **Registering UDF**: The REGISTER command makes the UDF available to the Pig script.
3. **Defining UDF**: DEFINE assigns a name to the UDF.
4. **Using UDF**: The FOREACH statement applies the BonusCalculator UDF to each record's salary to calculate the bonus.
5. **Storing Results**: The STORE function writes the result to an output directory.