EXPERIMENT NO: 5

AIM: Experiment on PIG

Theory:

Pig is a high-level platform and scripting language built on top of the Hadoop ecosystem. It simplifies the process of data analysis and manipulation in Hadoop by providing a more abstract and user-friendly way to express data transformations. In this 400-word explanation, we'll delve into Pig and its significance in the world of big data analytics.

Key Features and Concepts:

1. **Ease of Use**: Pig is designed to make big data processing more accessible for those who may not be well-versed in complex Java or MapReduce programming. It uses a simple and English-like scripting language, known as Pig Latin, to express data transformations.
2. **Abstraction Layer**: Pig provides a higher-level abstraction layer that sits on top of Hadoop, making it easier for analysts and data scientists to work with large datasets without getting deep into the low-level details of MapReduce.
3. **Extensibility**: Pig is extensible, allowing users to write their own user-defined functions (UDFs) in Java and incorporate them into their Pig scripts for custom data processing.
4. **Optimization**: Pig includes an optimizer that aims to improve the performance of data processing tasks by automatically optimizing and reordering operations in the script.
5. **Ecosystem Integration**: Pig works seamlessly with other Hadoop ecosystem components like HDFS, HBase, Hive, and more, enabling it to handle a wide variety of data sources and types.

Pig is a versatile tool suitable for various data processing tasks, including:

1. **ETL (Extract, Transform, Load)**: Pig is commonly used for ETL tasks, where data is ingested, cleaned, and transformed before being loaded into a data warehouse or analytical system.
2. **Data Cleansing**: Pig can handle the cleaning and preprocessing of data, removing inconsistencies, missing values, or irrelevant information.

1. Start pig and fs .

```
[training@localhost ~]$ pig
2023-10-13 21:56:54,482 [main] INFO  org.apache.pig.Main - Lo
2023-10-13 21:56:54,607 [main] INFO  org.apache.pig.backend.h
8020
2023-10-13 21:56:54,758 [main] INFO  org.apache.pig.backend.h
1
grunt> fs -ls
Found 62 items
-rw-r--r--   1 training supergroup       1390 2015-10-02 20:2
```

2. Copy from local

```
grunt> copyFromLocal /home/training/file.txt /user/training/
```

```
grunt> cat file.txt
Hello!!
This is Big Data Analytics
grunt> copyFromLocal /home/training/student.txt /user/training/
grunt> cat student.txt
1, disha, 68.2, F,female
2, harsh, 70, F,male
3, sahil, 83.33, F,male
4, Ishwar, 62, F,male
5, vikas, 68, F,male
6, ankita, 91, F,female
```

3. Create database and dump

```
grunt> student = load '/user/training/student.txt' USING PigStorage(',') as (rollno:int,name:chararray,percentage:float,sex:chararray);
grunt> dump student;
2023-10-13 22:11:26,106 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
```

```
2023-10-13 22:11:33,234 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-10-13 22:11:33,240 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-10-13 22:11:33,240 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1, disha,68.2, F)
(2, harsh,70.0, F)
(3, sahil,83.33, F)
(4, Ishwar,62.0, F)
(5, vikas,68.0, F)
(6, ankita,91.0, F)
```

4. Projections

```
grunt> Studentname = foreach student generate name;
grunt> Dump Studentname;
2023-10-13 22:13:13,276 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
```

```
2023-10-13 22:13:20,146 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-10-13 22:13:20,160 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-10-13 22:13:20,160 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
( disha)
( harsh)
( sahil)
( Ishwar)
( vikas)
( ankita)
```

```
2023-10-13 22:22:57,294 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-10-13 22:22:57,302 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-10-13 22:22:57,302 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(10, disha, engineer,50000)
(20, harsh, clerk,45000)
(30, sahil, scientist,50000)
(40, Ishwar, chemist,50000)
(50, vikas, biochemist,50000)
(60, ankita, engineer,50000)
```

5. Cross

```
grunt> x = cross student,dept;
grunt> dump x;
2023-10-13 22:26:21,100 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - P
2023-10-13 22:26:21,100 [main] INFO  org.apache.pig.backend.hadoop.executionengine
```

```
2023-10-13 22:26:38,085 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - S
2023-10-13 22:26:38,090 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to proce
2023-10-13 22:26:38,090 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths
(1, disha,68.2, F,20, harsh, clerk,45000)
(1, disha,68.2, F,40, Ishwar, chemist,50000)
(1, disha,68.2, F,30, sahil, scientist,50000)
(1, disha,68.2, F,60, ankita, engineer,50000)
(1, disha,68.2, F,50, vikas, biochemist,50000)
(1, disha,68.2, F,10, disha, engineer,50000)
(4, Ishwar,62.0, F,40, Ishwar, chemist,50000)
(4, Ishwar,62.0, F,20, harsh, clerk,45000)
(4, Ishwar,62.0, F,30, sahil, scientist,50000)
(5, vikas,68.0, F,40, Ishwar, chemist,50000)
(5, vikas,68.0, F,20, harsh, clerk,45000)
(5, vikas,68.0, F,30, sahil, scientist,50000)
```

6. Create a dummy numeric file and find distinct values

```
grunt> A = load '/user/training/dumA.txt' USING PigStorage(',') as (a1:int,a2:int,a3:int,a4:int);
grunt> dump A;
2023-10-13 22:29:57,454 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNK
2023-10-13 22:29:57,454 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - pig.usenewlogic
```

```
grunt> z = distinct A;
grunt> dump z;
2023-10-13 22:33:34,438 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: DISTINCT
2023-10-13 22:33:34,438 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - pig.usenewlogicalpla
```

7. Filter command

```
grunt> y = FILTER A By a2 == 2;
grunt> dump y;
2023-10-13 22:41:37,136 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
```

```
2023-10-13 22:46:18,391 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-10-13 22:46:18,394 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-10-13 22:46:18,394 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(10, disha)
(20, harsh)
(30, sahil)
(40, Ishwar)
(50, vikas)
(60, ankita)
```