Submitted by: Kuhan
Roll No.: 2024MCS110028

Lab Assessment - PIG

Task 1: PIG Installation Compatible with Hadoop Version

Step 1: Install PIG

Download Apache Pig:

kuhan@kuhan-virtual-machine:~$ wget https://downloads.apache.org/pig/latest/pig-0.17.0.tar.gz

Extract the package:

kuhan@kuhan-virtual-machine:~$ tar -xvzf pig-0.17.0.tar.gz

Move it to /usr/local/:

kuhan@kuhan-virtual-machine:~$ sudo mv pig-0.17.0 /usr/local/pig

Set environment variables (add to ~/.bashrc):

kuhan@kuhan-virtual-machine:~$ nano ~/.bashrc

Add the following lines:

export PIG_HOME=/usr/local/pig

export PATH=$PIG_HOME/bin:$PATH

export PIG_CLASSPATH=$HADOOP_HOME/etc/hadoop

Apply changes:

kuhan@kuhan-virtual-machine:~$ source ~/.bashrc

Verify installation:

kuhan@kuhan-virtual-machine:~$ pig -version

Task 2: Basic Operation on PIG

Step 1: Create a CSV File

Create sample_data.csv with the following content:

kuhan@kuhan-virtual-machine:~$ nano sample_data.csv

Inside the file:

Name,Age,Salary,Address

John,25,50000,New York

Emma,30,60000,California

Liam,28,55000,Texas

Sophia,35,75000,Florida

Noah,40,90000,Washington

Step 2: Upload CSV File to HDFS

Create a directory in HDFS:

kuhan@kuhan-virtual-machine:~$ hdfs dfs -mkdir /pig_data

Upload the file to HDFS:

kuhan@kuhan-virtual-machine:~$ hdfs dfs -put sample_data.csv /pig_data/

Verify the file is uploaded:

kuhan@kuhan-virtual-machine:~$ hdfs dfs -ls /pig_data/

Step 3: Load Data in Pig

Start Pig in MapReduce mode:

kuhan@kuhan-virtual-machine:~$ pig -x mapreduce

Run the following Pig script:

grunt> data = LOAD '/pig_data/sample_data.csv' USING PigStorage(',') AS (Name:chararray, Age:int, Salary:int, Address:chararray);

grunt> DUMP data;

Step 4: Perform Operations

FILTER: Employees with Salary > 55000

```
grunt> high_salary = FILTER data BY Salary > 55000;
```

```
grunt> DUMP high_salary;
```

LIMIT: Display Only 3 Records

```
grunt> limited_data = LIMIT data 3;
```

```
grunt> DUMP limited_data;
```

ORDER BY: Sort Data by Salary in Descending Order

```
grunt> ordered_data = ORDER data BY Salary DESC;
```

```
grunt> DUMP ordered_data;
```

Summary

Installed Pig and configured it with Hadoop.

Created a sample CSV file and uploaded it to HDFS.

Loaded data in Pig and performed operations (LOAD, FILTER, LIMIT, ORDER BY).
