KISORE R G
210701124

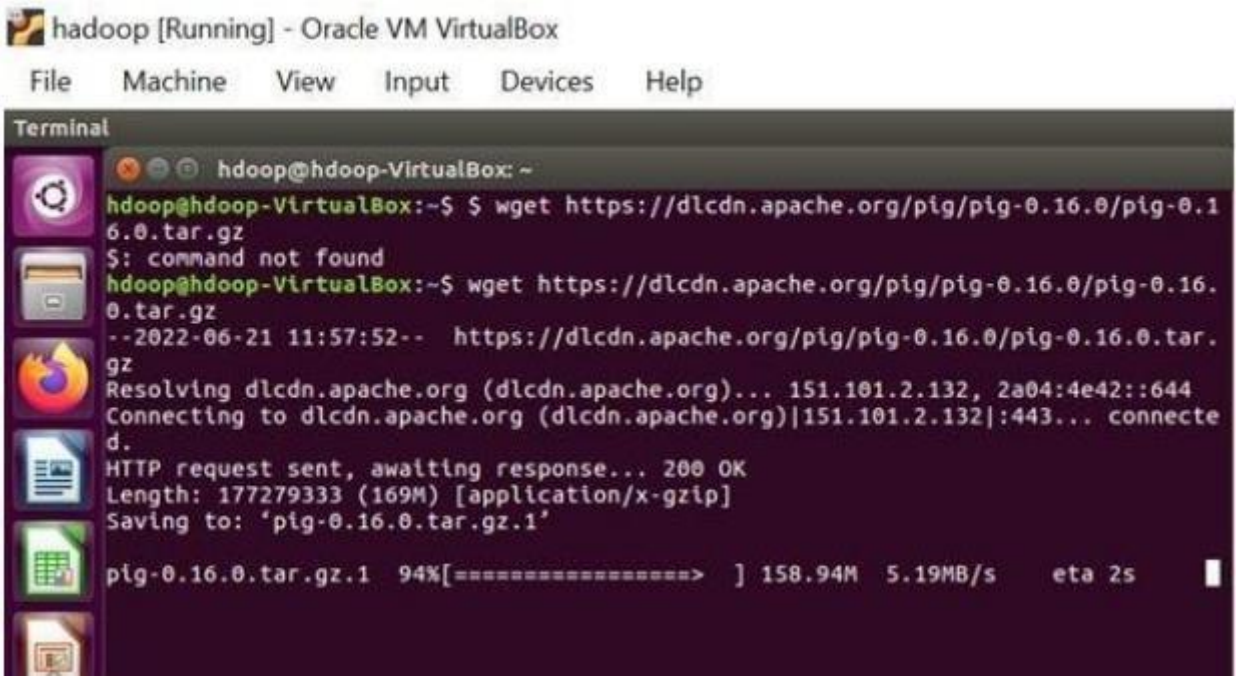
EXP NO:4 **Create UDF in PIG**

AIM:

Step-by-step installation of Apache Pig on Hadoop cluster on Ubuntu.

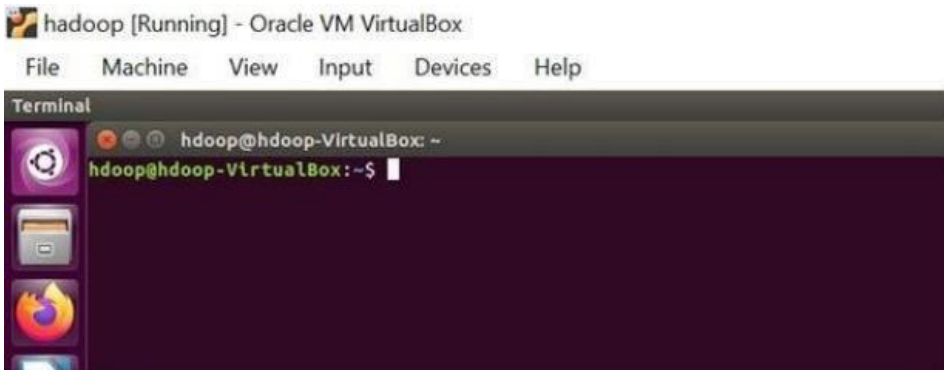
PROCEDURE:

Step 1 : **Login into Ubuntu**.



Step 2: Go to https://pig.apache.org/releases.html and copy the path of the latest version of pig that you want to install. Run the following comment to download Apache Pig in Ubuntu:

$ wget https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.gz



Step 3: To untar pig-0.16.0.tar.gz file run the following command:

$ tar xvzf pig-0.16.0.tar.gz

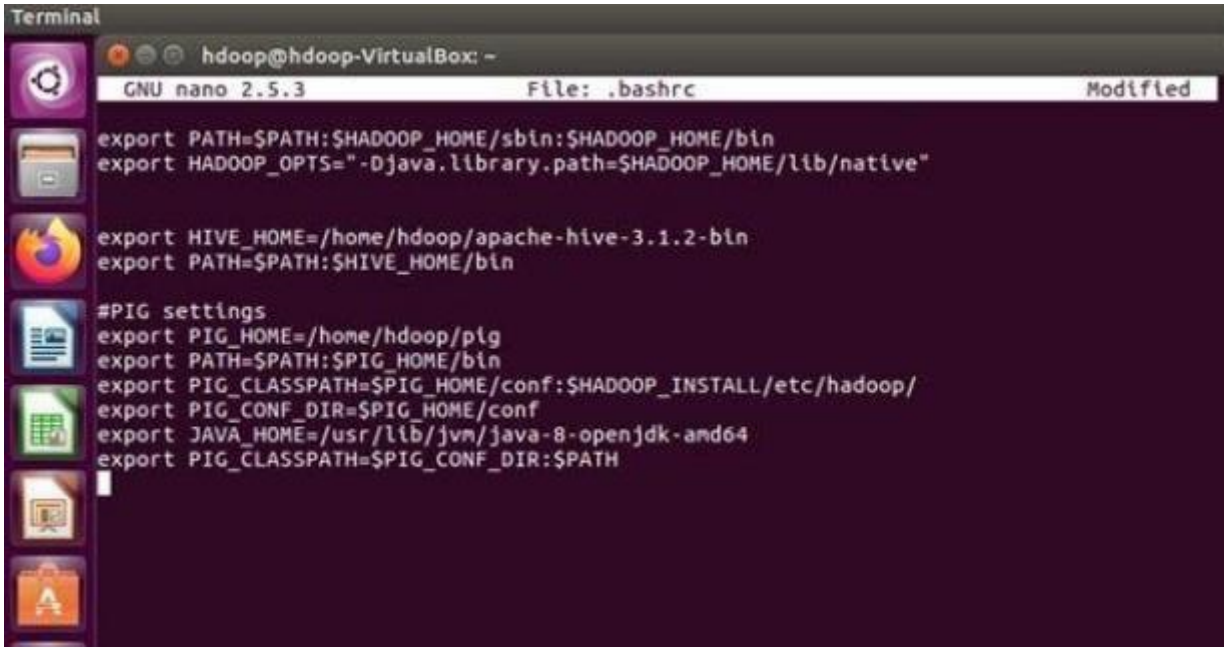
Step 4: To create a pig folder and move pig-0.16.0 to the pig folder, execute the following command:

$ sudo mv /home/hdoop/pig-0.16.0 /home/hdoop/pig

Step 5: Now open the .bashrc file to edit the path and variables/settings for pig. Run the following command:

$ sudo nano .bashrc

Add the below given to .bashrc file at the end and save the file.
#PIG settingsexport PIG_HOME=/home/hdoop/pigexport PATH=$PATH:$PIG_HOME/binexport PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop/export PIG_CONF_DIR=$PIG_HOME/confexport JAVA_HOME=/usr/lib/jvm/java-8-openjdkamd64export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH#PIG setting ends



Step 6: Run the following command to make the changes effective in the .bashrc file:

$ source .bashrc

Step 7: To start all Hadoop daemons, navigate to the hadoop-3.2.1/sbin folder and run the following commands:

$ ./start-dfs.sh$ ./start-yarn$ jps

```
hadoop@ubuntu:~/hadoop$ jps
51489 Jps
33059 SecondaryNameNode
32887 DataNode
32652 NameNode
33325 ResourceManager
33453 NodeManager
hadoop@ubuntu:~/hadoop$
```

Step 8: Now you can launch pig by executing the following command:

$ pig

```
hadoop@ubuntu:/$ pig
2024-09-17 19:24:37,302 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-17 19:24:37,317 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-17 19:24:37,317 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-17 19:24:37,422 WARN pig.Main: Cannot write to log file: //pig_1726581277422.lo
g
2024-09-17 19:24:37,441 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r
1746530) compiled Jun 01 2016, 23:10:49
2024-09-17 19:24:37,516 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup fi
le /home/hadoop/.pigbootup not found
2024-09-17 19:24:38,197 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
 mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-17 19:24:38,198 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
 fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-17 19:24:38,198 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExe
cutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-17 19:24:39,206 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
 fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-17 19:24:39,295 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the s
ession: PIG-default-c76725b6-2bd0-4223-8347-da37b42cbd12
2024-09-17 19:24:39,300 [main] WARN  org.apache.pig.PigServer - ATS is disabled since y
arn.timeline-service.enabled set to false
```

Step 9: Now you are in pig and can perform your desired tasks on pig. You can come out of the pig by the quit command: > quit;

KISORE R G
210701124

CREATE USER DEFINED FUNCTION(UDF)

AIM : To create User Define Function in Apache Pig and execute it on map reduce.

PROCEDURE:
Create a sample text file

hadoop@Ubuntu:$ nano sample.txt
sample.txt
1,John
2,Jane
3,Joe
4,Emma

Create PIG File

hadoop@Ubuntu:~/Documents$ nano demo_pig.pig

```
-- Load the data from HDFS
data = LOAD '/home/hadoop/piginput/sample.txt' USING PigStorage(',') AS (id:int>
-- Dump the data to check if it was loaded correctly
DUMP data;
```

Run the above file
hadoop@Ubuntu:$ pig demo_pig.pig

```
- Total input paths to process : 1
(1,John)
(2,Jane)
(3,Joe)
(4,Emma)
```

Create udf file an save as uppercase_udf.py

uppercase_udf.py:

```
def uppercase(text):
return text.upper()
if __name__ == "__main__":
import sys
for line in sys.stdin:
        line = line.strip()
        result = uppercase(line)
        print(result)
```

KISORE R G
210701124

Create the udfs folder on hadoop

hadoop@Ubuntu:$ hadoop fs -mkdir /home/hadoop/udfs

put the upppercase_udf.py in to the abv folder

hadoop@Ubuntu:$ hdfs dfs -put uppercase_udf.py /home/hadoop/udfs/

hadoop@Ubuntu:$ nano udf_example.pig

place sample.txt file on hadoop

hadoop@Ubuntu:$ hadoop fs -put sample.txt /home/hadoop/

To Run the pig file

hadoop@Ubuntu:$ pig -f udf.pig

INPUT:

```
Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime
   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feat
ure Outputs
job_local1563652978_0001        1       0       n/a     n/a     n/a     n/a     0
       0       0       0       data    MAP_ONLY        hdfs://localhost:9000/tmp/temp92
3286826/tmp982664124,

Input(s):
Successfully read 5 records (5378235 bytes) from: "/udf_pig/sample.txt"
```

OUTPUT:

KISORE R G
210701124

```
Output(s):
Successfully stored 5 records (5378263 bytes) in: "hdfs://localhost:9000/tmp/temp923286
826/tmp982664124"

Counters:
Total records written : 5
Total bytes written : 5378263
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1563652978_0001


2024-09-13 08:44:46,888 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl
- JobTracker metrics system already initialized!
2024-09-13 08:44:46,896 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl
- JobTracker metrics system already initialized!
2024-09-13 08:44:46,900 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl
- JobTracker metrics system already initialized!
2024-09-13 08:44:46,930 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapR
educeLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 1 time(
s).
2024-09-13 08:44:46,932 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapR
educeLayer.MapReduceLauncher - Success!
2024-09-13 08:44:46,948 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
 fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-13 08:44:46,948 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTup
leBackend has already been initialized
2024-09-13 08:44:47,017 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFor
mat - Total input files to process : 1
2024-09-13 08:44:47,018 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util
```

To view the output

hadoop@Ubuntu:$ hdfs dfs -cat /home/hadoop/pig_output_data/part-m00000

```
.MapRedUtil - Total input paths to process : 1
(1,John)
(2,Jane)
(3,Joe)
(4,Emma)
(,)
```

RESULT:

Thus, pig installation and a program has been executed successfully.