Exp. No.: 4 **Create UDF in PIG**

Step-by-step installation of Apache Pig on Hadoop cluster on Ubuntu Pre-requisite:

· Ubuntu 16.04 or higher version running (I have installed Ubuntu on Oracle VM (Virtual Machine) VirtualBox),

· Run Hadoop on ubuntu (I have installed Hadoop 3.2.1 on Ubuntu 16.04). You may refer to my blog "How to install Hadoop installation" click here for Hadoop installation).

Pig installation steps

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/hadoop/pig-0.16.0 /home/hadoop/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

```
 GNU nano 7.2                              .bashrc

export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HADOOP_HOME=/home/hadoop/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"


# PIG settings
export PIG_HOME=/home/hadoop/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop
export PIG_CONF_DIR=$PIG_HOME/conf
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export PIG_CLASSPATH=$PIG_CONF_DIR:$PIG_CLASSPATH
# PIG settings end
```

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@lab-H410M-H:~/hadoop$ jps
14480 NodeManager
13857 SecondaryNameNode
14722 Jps
14122 ResourceManager
13418 NameNode
13615 DataNode
hadoop@lab-H410M-H:~/hadoop$
```

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

```
vamsee@ubuntu: $ pig
2024-09-13 08:58:02,246 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-13 08:58:02,248 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-13 08:58:02,248 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-13 08:58:02,295 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-13 08:58:02,296 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/vamsee/pig_1726198082289.log
2024-09-13 08:58:02,330 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/vamsee/.pigbootup not found
2024-09-13 08:58:02,625 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead,
se mapreduce.jobtracker.address
2024-09-13 08:58:02,625 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, us
fs.defaultFS
2024-09-13 08:58:02,625 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file sy
em at: hdfs://localhost:9000
2024-09-13 08:58:03,261 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, us
fs.defaultFS
2024-09-13 08:58:03,293 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-0eee5d5d-6cd3-407a-aba2
ea7b7f262fe
2024-09-13 08:58:03,293 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt>
```

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;

```
grunt> quit
2024-09-13 08:59:22,810 [main] INFO  org.apache.pig.Main - Pig script completed in 1 minute, 20 seconds and 588 milliseconds (80588
ms)
vamsee@ubuntu:~$
```

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:~/Documents$ nano sample.txt

Paste the below content to sample.txt

1,John
2,Jane
3,Joe
4,Emma

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

Create PIG File

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

paste the below the content to 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:~/Documents$ pig demo_pig.pig

```
2024-09-19 22:33:24,709 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2024-09-19 22:33:24,849 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.4.0   0.16.0  thrisha 2024-09-19 22:30:32     2024-09-19 22:33:24     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTim
e       MedianReducetime        Alias   Feature Outputs
job_1726761169955_0003  1       0       n/a     n/a     n/a     n/a     0       0       0       0       data    MAP_ONLY        hdfs
://localhost:9000/tmp/temp-231182825/tmp-950685089,

Input(s):
Successfully read 0 records from: "hdfs://localhost:9000/piginput/sample.txt"

Output(s):
Successfully stored 0 records in: "hdfs://localhost:9000/tmp/temp-231182825/tmp-950685089"

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

Job DAG:
job_1726761169955_0003
```

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)

Create the udfs folder on hadoop

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

put the upppercase_udf.py in to the abv folder

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

hadoop@Ubuntu:~/Documents$ nano udf_example.pig copy and paste the below content on udf_example.pig

-- Register the Python UDF script

REGISTER 'hdfs:///home/hadoop/udfs/uppercase_udf.py' USING jython AS udf;

-- Load some data

data = LOAD 'hdfs:///home/hadoop/sample.txt' AS (text:chararray);

-- Use the Python UDF

uppercased_data = FOREACH data GENERATE udf.uppercase(text) AS uppercase_text;

-- Store the result

STORE uppercased_data INTO 'hdfs:///home/hadoop/pig_output_data';

---

place sample.txt file on hadoop

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

To Run the pig file

hadoop@Ubuntu:~/Documents$ pig -f udf_example.pig

```
2024-09-19 22:34:27,383 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-09-19 22:34:27,390 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process :
1
(1,John)
(2,Jane)
(3,Joe)
(4,Emma)
2024-09-19 22:34:29,146 [main] INFO  org.apache.pig.Main - Pig script completed in 4 minutes, 1 second and 741 milliseconds (241741
ms)
```

To check the output file is created

hadoop@Ubuntu:~/Documents$ hdfs dfs -ls /home/hadoop/pig_output_data

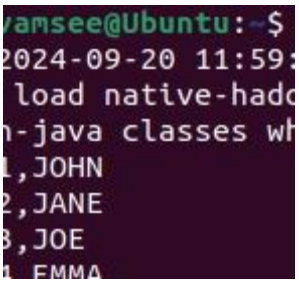
Found 2 items

If you need to examine the files in the output folder, use:

To view the output

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

```
1,JOHN
2,JANE
3,JOE
4,EMMA
```



Result:

Thus the program to create User Define Function in Apache Pig and execute it on map reduce has been done successfully.