Reg.No.: 210701263

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

Reg.No.: 210701263

```
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

```
subbu@subbu:~$ jps
5970 RunJar
3700 NodeManager
6185 RunJar
11066 Jps
3146 DataNode
3578 ResourceManager
3019 NameNode
3374 SecondaryNameNode
subbu@subbu:~$
```

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

```
subbu@subbu: ~/exp4
subbu@subbu:~/exp3$ cd ..
subbu@subbu:~$ cd exp4
subbu@subbu:~/exp4$ pig
2024-09-21 15:38:24,170 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-21 15:38:24,238 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-21 15:38:24,241 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecT
ype
2024-09-21 15:38:24,788 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-21 15:38:24,788 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/subbu/exp4/pig_1726913304740.log
2024-09-21 15:38:25,039 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/subbu/.pigbootup not found
2024-09-21 15:38:27,133 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-09-21 15:38:27,133 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-21 15:38:27,133 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-21 15:38:32,492 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-21 15:38:32,716 [main] INFO  org.apache.pig.PigServer - Pig Script ID fo
r the session: PIG-default-a3280407-b0f8-464d-b0a2-fece712eaf8a
```

Reg.No.: 210701263

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;

```
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-21 15:38:32,716 [main] INFO  org.apache.pig.PigServer - Pig Script ID fo
r the session: PIG-default-a3280407-b0f8-464d-b0a2-fece712eaf8a
2024-09-21 15:38:32,717 [main] WARN  org.apache.pig.PigServer - ATS is disabled
since yarn.timeline-service.enabled set to false
grunt> quit
2024-09-21 15:38:53,443 [main] INFO  org.apache.pig.Main - Pig script completed
in 30 seconds and 327 milliseconds (30327 ms)
subbu@subbu:~/exp4$
```

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

Reg.No.: 210701263

```
subbu@subbu: ~/exp4

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMa
pTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        A
lias    Feature Outputs
job_1726912313635_0001  1       0       n/a     n/a     n/a     n/a     0       0
0       0       data    MAP_ONLY        hdfs://localhost:9000/tmp/temp-165198539
9/tmp-504640378,

Input(s):
Successfully read 0 records from: "/exp4/sample.txt"

Output(s):
Successfully stored 0 records in: "hdfs://localhost:9000/tmp/temp-1651985399/tmp
-504640378"

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:
```

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);

Reg.No.: 210701263

-- 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

```
ne.util.MapRedUtil - Total input paths to process : 1
(1,John)
(2,Jane)
(3,Joe)
(4,Emma)
2024-09-21 15:47:08,287 [main] INFO  org.apache.pig.Main - Pig script completed
in 4 minutes, 3 seconds and 41 milliseconds (243041 ms)
subbu@subbu:~/exp4$
```

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

```
subbu@subbu:~/exp4$ hdfs dfs -cat /exp4/output/part-m-00000
1,JOHN
2,JANE
3,JOE
4,EMMA
subbu@subbu:~/exp4$
```

Result:

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