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@sudharsan-sundar-VirtualBox:~/hadoop/sbin$ nano .bashrc
hadoop@sudharsan-sundar-VirtualBox:~/hadoop/sbin$ cd
hadoop@sudharsan-sundar-VirtualBox:~$ nano .bashrc
hadoop@sudharsan-sundar-VirtualBox:~$ source .bashrc
hadoop@sudharsan-sundar-VirtualBox:~$ jps
3203 SecondaryNameNode
2852 NameNode
2983 DataNode
6509 Jps
3437 ResourceManager
```

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

```
hadoop@sudharsan-sundar-VirtualBox:~$ pig
2024-09-19 19:40:19,916 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-19 19:40:19,930 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-19 19:40:19,930 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-19 19:40:20,006 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15
:41:58
2024-09-19 19:40:20,009 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/hadoop/pig_1726755020000.log
2024-09-19 19:40:20,058 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup not fo
und
2024-09-19 19:40:20,385 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated
. Instead, use mapreduce.jobtracker.address
2024-09-19 19:40:20,387 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hado
op file system at: hdfs://localhost:9000
2024-09-19 19:40:21,303 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-e4750850-acc8
-4f0f-99ea-284200aceb80
2024-09-19 19:40:21,303 [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;

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,Srimathy

2,Subhikshaa

3,Sudharsan

4,Vaisharly

5,Swetha

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

```
hadoop@sudharsan-sundar-VirtualBox:~$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2024-09-19 17:31 /weatherdata
drwxr-xr-x   - hadoop supergroup          0 2024-09-19 13:57 /word_count_in_python
hadoop@sudharsan-sundar-VirtualBox:~$ hadoop fs -mkdir -p /home/hadoop/piginput/
hadoop@sudharsan-sundar-VirtualBox:~$ hadoop fs -put sample.txt/home/hadoop/piginput/
put: `.': No such file or directory: `hdfs://localhost:9000/user/hadoop'
hadoop@sudharsan-sundar-VirtualBox:~$ hadoop fs -put sample.txt /home/hadoop/piginput/
hadoop@sudharsan-sundar-VirtualBox:~$ nano demo_pig.pig
hadoop@sudharsan-sundar-VirtualBox:~$ pig demo_pig.pig
2024-09-19 20:00:33,820 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-19 20:00:33,824 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-19 20:00:33,827 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-19 20:00:33,944 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15
:41:58
2024-09-19 20:00:33,947 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/hadoop/pig_1726756233937.log
2024-09-19 20:00:34,379 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup not fo
und
2024-09-19 20:00:34,458 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated
. Instead, use mapreduce.jobtracker.address
2024-09-19 20:00:34,459 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hado
op file system at: hdfs://localhost:9000
2024-09-19 20:00:35,308 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-demo_pig.pig-4776ba00
-872d-4455-9f07-d46ac0064d18
2024-09-19 20:00:35,309 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set
to false
2024-09-19 20:00:36,469 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOW
N
2024-09-19 20:00:36,537 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will
not generate code.
```

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

```
hadoop@sudharsan-sundar-VirtualBox:~$ nano uppercase_udf.py
hadoop@sudharsan-sundar-VirtualBox:~$ hadoop fs -mkdir /home/hadoop/udfs
hadoop@sudharsan-sundar-VirtualBox:~$ hadoop fs -put uppercase_udf.py /home/hadoop/udfs/
hadoop@sudharsan-sundar-VirtualBox:~$ nano udf_example.pig
hadoop@sudharsan-sundar-VirtualBox:~$ hadoop fs -put sample.txt /home/hadoop/
put: `/home/hadoop/sample.txt': File exists
hadoop@sudharsan-sundar-VirtualBox:~$ pig -f udf_example.pig
2024-09-19 20:43:31,377 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-19 20:43:31,378 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-19 20:43:31,378 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-19 20:43:31,460 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15
:41:58
2024-09-19 20:43:31,460 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/hadoop/pig_1726758811453.log
2024-09-19 20:43:31,689 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup not fo
und
2024-09-19 20:43:31,747 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated
. Instead, use mapreduce.jobtracker.address
2024-09-19 20:43:31,747 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hado
op file system at: hdfs://localhost:9000
2024-09-19 20:43:32,192 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-udf_example.pig-c898c
261-5349-4c41-9ccf-c51056100cfd
2024-09-19 20:43:32,193 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set
to false
2024-09-19 20:43:32,655 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine - created tmp python.cachedir=/t
mp/pig_jython_8840826539398102556
2024-09-19 20:43:36,746 [main] WARN  org.apache.pig.scripting.jython.JythonScriptEngine - pig.cmd.args.remainders is emp
```

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

```
hadoop@sudharsan-sundar-VirtualBox:~$ hadoop fs -ls /home/hadoop/pig_output_data
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2024-09-19 20:43 /home/hadoop/pig_output_data/_SUCCESS
-rw-r--r--   1 hadoop supergroup         66 2024-09-19 20:43 /home/hadoop/pig_output_data/part-m-00000
hadoop@sudharsan-sundar-VirtualBox:~$ hadoop fs -cat /home/hadoop/pig_output_data/part-m-00000
1.SRIMATHY
2.SUBHIKASHAA
3.SUDHARSAN
4.VAISHARLY
5.SWETHA
6.PRIYA
hadoop@sudharsan-sundar-VirtualBox:~$
```

Result:

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