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

```
vaisharli@vaisharli:~$ wget https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.16.0.t
ar.gz
--2024-09-14 13:37:31--  https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.
gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connecte
d.
HTTP request sent, awaiting response... 200 OK
Length: 177279333 (169M) [application/x-gzip]
Saving to: 'pig-0.16.0.tar.gz.1'

pig-0.16.0.tar.gz.1 100%[===================>] 169.07M  6.29MB/s    in 28s

2024-09-14 13:38:14 (6.13 MB/s) - 'pig-0.16.0.tar.gz.1' saved [177279333/1772793
33]
```

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:

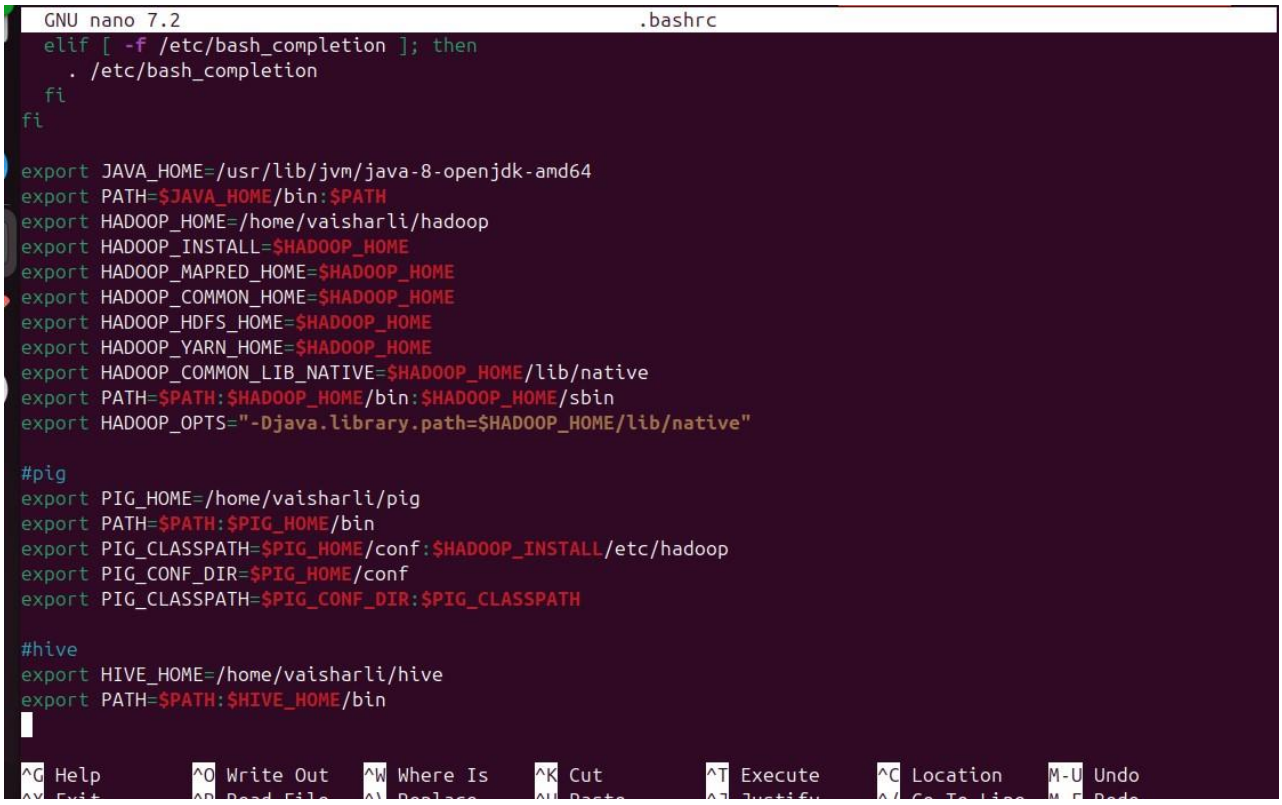
$ sudo mv /home/hadoop/pig-0.16.0 /home/hadoop/pig

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



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

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

```
2024-09-14 13:49:52,623 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-14 13:49:52,625 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-14 13:49:52,625 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecT
ype
2024-09-14 13:49:52,674 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-14 13:49:52,675 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/vaisharli/pig_1726301992661.log
2024-09-14 13:49:52,707 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/vaisharli/.pigbootup not found
2024-09-14 13:49:53,031 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2024-09-14 13:49:53,031 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 13:49:53,031 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-14 13:49:54,041 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-14 13:49:54,080 [main] INFO  org.apache.pig.PigServer - Pig Script ID fo
r the session: PIG-default-e35446d9-ef37-482c-93d6-db43f6cf1e0c
2024-09-14 13:49:54,088 [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,harry

2,ron

3,hermoine

4,cedric

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

```
vaisharli@vaisharli:~$ hdfs dfs -mkdir -p /home/hadoop/piginput
vaisharli@vaisharli:~$ hdfs dfs -ls /home/hadoop/piginput
vaisharli@vaisharli:~$ hadoop fs -put sample.txt /home/hadoop/piginput/
vaisharli@vaisharli:~$ hdfs dfs -ls /home/hadoop/piginput
Found 1 items
-rw-r--r--   1 vaisharli supergroup         34 2024-09-20 10:33 /home/hadoop/piginput/sample.txt
vaisharli@vaisharli:~$ nano demo_pig.pig
vaisharli@vaisharli:~$ pig demo_pig.pig
2024-09-20 10:37:01,071 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-20 10:37:01,074 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-20 10:37:01,075 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-20 10:37:01,269 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun
1 2016, 23:10:49
2024-09-20 10:37:01,275 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/vaisharli/pig_172
808821248.log
2024-09-20 10:37:01,897 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/vaisharli/.pi
bootup not found
2024-09-20 10:37:02,028 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is
deprecated. Instead, use mapreduce.jobtracker.address
2024-09-20 10:37:02,029 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is de
recated. Instead, use fs.defaultFS
2024-09-20 10:37:02,030 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connect
ng to hadoop file system at: hdfs://localhost:9000
2024-09-20 10:37:02,846 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is de
recated. Instead, use fs.defaultFS
2024-09-20 10:37:02,882 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-demo_pig.p
g-2b213e9a-4c58-4657-be57-b6a40d5a0016
2024-09-20 10:37:02,883 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.e
abled set to false
2024-09-20 10:37:03,497 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is de
```

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

```
vaisharli@vaisharli:~$ nano udf_example.pig
vaisharli@vaisharli:~$ pig -f udf_example.pig
2024-09-20 10:40:19,861 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-20 10:40:19,863 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-20 10:40:19,863 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-20 10:40:19,912 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun
1 2016, 23:10:49
2024-09-20 10:40:19,912 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/vaisharli/pig_172
809019903.log
2024-09-20 10:40:20,287 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/vaisharli/.pi
bootup not found
2024-09-20 10:40:20,404 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is
deprecated. Instead, use mapreduce.jobtracker.address
2024-09-20 10:40:20,405 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is de
recated. Instead, use fs.defaultFS
2024-09-20 10:40:20,405 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connect
ng to hadoop file system at: hdfs://localhost:9000
2024-09-20 10:40:20,975 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is de
recated. Instead, use fs.defaultFS
2024-09-20 10:40:21,003 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-udf_exampl
.pig-4a59e5c8-9e66-4f7c-9a4d-898fc94dcb5a
2024-09-20 10:40:21,003 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.e
abled set to false
2024-09-20 10:40:21,063 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is de
recated. Instead, use fs.defaultFS
2024-09-20 10:40:21,502 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine - created tmp python.
achedir=/tmp/pig_jython_4329229878308954211
2024-09-20 10:40:24,588 [main] WARN  org.apache.pig.scripting.jython.JythonScriptEngine - pig.cmd.args.remain
ers is empty. This is not expected unless on testing.
2024-09-20 10:40:24,595 [main] INFO  org.apache.pig.scripting.jython.JythonScriptEngine - Register scripting
DF: udf.uppercase
```

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

```
vaisharli@vaisharli:~$ hdfs dfs -cat /home/hadoop/pig_output_data/part-m-00000
HARRY
RON
HERMOINE
CEDRIC
vaisharli@vaisharli:~$
```

Result:

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