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

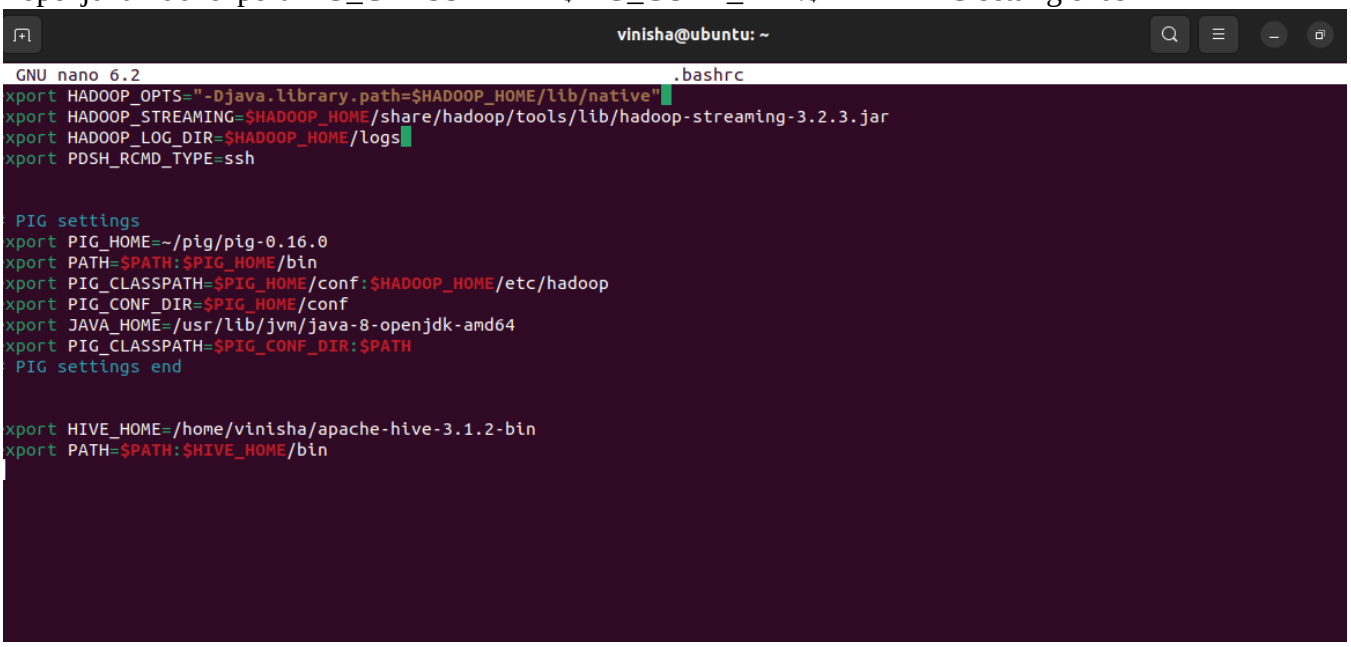
Step 5: Now open the .bashrc file to edit the path and variables/settings for pig. Run the following command:

$ 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

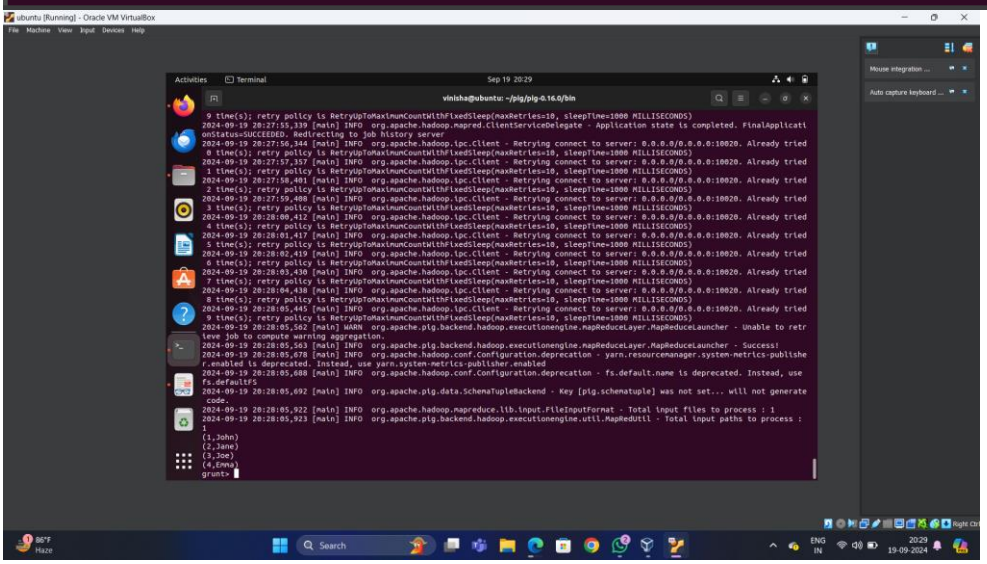


Step 6: Run the following command to make the changes effective in the .bashrc file:

$ 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: $ pi

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

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

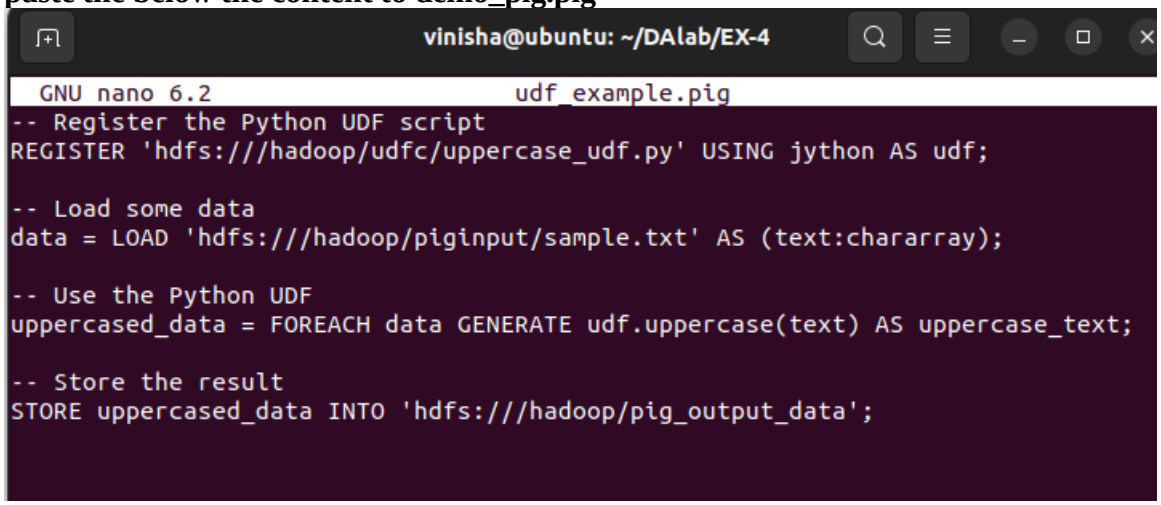
Paste the below content to sample.txt

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

Create PIG File

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

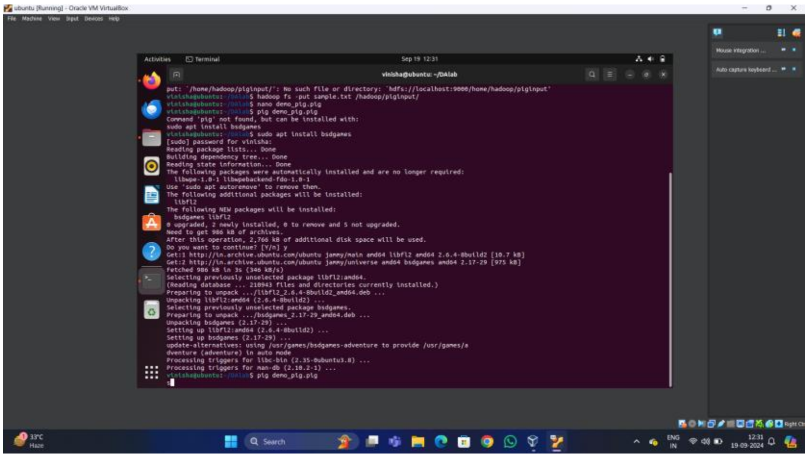
paste the below the content to demo_pig.pig



```
-- 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
```



```
put: `/home/hadoop/piginput/': No such file or directory: `hdfs://localhost:9000/home/hadoop/piginput'
vinisha@ubuntu:~/DAlab$ hadoop fs -put sample.txt /hadoop/piginput/
vinisha@ubuntu:~/DAlab$ nano demo_pig.pig
vinisha@ubuntu:~/DAlab$ pig demo_pig.pig
Command 'pig' not found, but can be installed with:
sudo apt install bsdgames
vinisha@ubuntu:~/DAlab$ sudo apt install bsdgames
[sudo] password for vinisha:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following packages were automatically installed and are no longer required:
  libwpe-1.0-1 libwpebackend-fdo-1.0-1
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
  libfl2
The following NEW packages will be installed:
  bsdgames libfl2
0 upgraded, 2 newly installed, 0 to remove and 5 not upgraded.
Need to get 986 kB of archives.
After this operation, 2,766 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://in.archive.ubuntu.com/ubuntu jammy/main amd64 libfl2 amd64 2.6.4-8build2 [10.7 kB]
Get:2 http://in.archive.ubuntu.com/ubuntu jammy/universe amd64 bsdgames amd64 2.17-29 [975 kB]
Fetched 986 kB in 3s (346 kB/s)
Selecting previously unselected package libfl2:amd64.
(Reading database ... 210943 files and directories currently installed.)
Preparing to unpack .../libfl2_2.6.4-8build2_amd64.deb ...
Unpacking libfl2:amd64 (2.6.4-8build2) ...
Selecting previously unselected package bsdgames.
Preparing to unpack .../bsdgames_2.17-29_amd64.deb ...
Unpacking bsdgames (2.17-29) ...
Setting up libfl2:amd64 (2.6.4-8build2) ...
Setting up bsdgames (2.17-29) ...
update-alternatives: using /usr/games/bsdgames-adventure to provide /usr/games/a
dventure (adventure) in auto mode
Processing triggers for libc-bin (2.35-0ubuntu3.8) ...
Processing triggers for man-db (2.10.2-1) ...
vinisha@ubuntu:~/DAlab$ pig demo_pig.pig
$
```

- -

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';
```

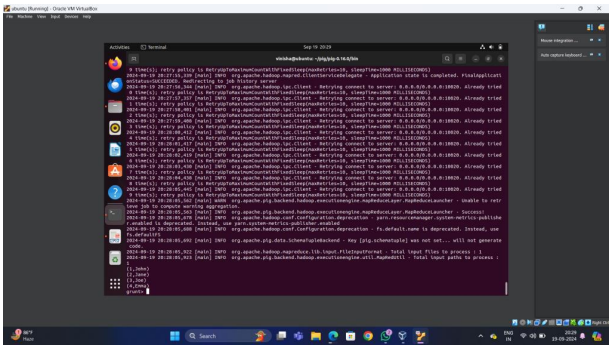
---

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



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

```
vinisha@ubuntu:~/DAlab$ nano sample.txt
vinisha@ubuntu:~/DAlab$ hdfs dfs -cat /hadoop/pig_output_data/part-m-00000
1.JOHN
2.JANE
3.JOE
4.EMMA
vinisha@ubuntu:~/DAlab$
```

Result:

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