

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**


```
hadoop@hadoop-VirtualBox:~$ $ wget https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.gz
$: command not found
hadoop@hadoop-VirtualBox:~$ wget https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.gz
--2022-06-21 11:57:52-- 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... connected.
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  94%[=====>] 158.94M  5.19MB/s   eta 2s
```

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 command 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 settings
export PIG_HOME=/home/hadoop/pig
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop/export
export PIG_CONF_DIR=$PIG_HOME/conf
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PIG_CLASSPATH=$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
```

```

srinathi@srinathi-VirtualBox:~$ jps
4992 NameNode
6225 Jps
5447 SecondaryNameNode
5803 NodeManager
5660 ResourceManager
5167 DataNode

```

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

```
srinathi@srinathi-VirtualBox:~$ pig
2024-09-19 18:22:12,652 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-19 18:22:12,653 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-19 18:22:12,654 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-19 18:22:12,708 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01
2016, 23:10:49
2024-09-19 18:22:12,709 [main] INFO org.apache.pig.Main - Logging error messages to: /home/srinathi/pig_1726750
332695.log
2024-09-19 18:22:12,750 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/srinathi/.pigboo
tup not found
2024-09-19 18:22:13,202 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is de
precated. Instead, use mapreduce.jobtracker.address
2024-09-19 18:22:13,202 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depre
cated. Instead, use fs.defaultFS
2024-09-19 18:22:13,202 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting
to hadoop file system at: hdfs://localhost:9000
2024-09-19 18:22:14,058 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depre
cated. Instead, use fs.defaultFS
2024-09-19 18:22:14,104 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-64004
d12-1aa6-4a02-9741-d037862a77ba
2024-09-19 18:22:14,104 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enab
led 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,SRI

2,VAISH

3,SUBHI

4,PRIYA

5,SWEATHA

```
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
```



```

srinathi@srinathi-VirtualBox:~$ nano sample.txt
srinathi@srinathi-VirtualBox:~$ hadoop fs -mkdir -p /home/srinathi/piginput
srinathi@srinathi-VirtualBox:~$ hadoop fs -put sample.txt /home/srinathi/piginput
srinathi@srinathi-VirtualBox:~$ ls /home/srinathi/piginput
Found 1 items
-rw-r--r-- 3 srinathi supergroup 40 2024-09-19 21:11 /home/srinathi/piginput/sample.txt
srinathi@srinathi-VirtualBox:~$ nano demo_pig.pig
srinathi@srinathi-VirtualBox:~$ pig demo_pig.pig
2024-09-19 21:13:52,971 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-19 21:13:52,972 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-19 21:13:52,972 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-19 21:13:53,028 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-19 21:13:53,028 [main] INFO org.apache.pig.Main - Logging error messages to: /home/srinathi/pig_1726760633017.log
2024-09-19 21:13:53,269 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/srinathi/.pigbootstrap not found
2024-09-19 21:13:53,309 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce
.jobtracker.address
2024-09-19 21:13:53,309 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-19 21:13:53,309 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs:
//localhost:9000
2024-09-19 21:13:53,599 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-19 21:13:53,629 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-demo_pig.pig-5044271b-d1f5-4e4e-a78e-e6b8e09f
ef03
2024-09-19 21:13:53,629 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-09-19 21:13:53,913 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-19 21:13:54,099 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-09-19 21:13:54,193 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-19 21:13:54,207 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-09-19 21:13:54,235 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}

```

Create udf file and 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 uppercase_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 uppercase_data = FOREACH data GENERATE
```

```
udf.uppercase(text) AS uppercase_text;
```

```
-- Store the result
```

```
STORE uppercase_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
```

```
srinathi@srinathi-VirtualBox:~$ nano uppercase_udf.py
srinathi@srinathi-VirtualBox:~$ hdfs dfs -mkdir /home/srinathi/udfs
srinathi@srinathi-VirtualBox:~$ hdfs dfs -put uppercase_udf.py /home/srinathi/udfs/
srinathi@srinathi-VirtualBox:~$ nano udf_example.pig
srinathi@srinathi-VirtualBox:~$ hadoop fs -put sample.txt /home/srinathi/
srinathi@srinathi-VirtualBox:~$ pig -f udf_example.pig
2024-09-19 21:21:40,131 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-19 21:21:40,131 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-19 21:21:40,132 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-19 21:21:40,186 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-19 21:21:40,187 [main] INFO org.apache.pig.Main - Logging error messages to: /home/srinathi/pig_1726761100177.log
2024-09-19 21:21:40,421 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/srinathi/.pigbootup not found
2024-09-19 21:21:40,475 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce
.jobtracker.address
2024-09-19 21:21:40,475 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-19 21:21:40,475 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs:
//localhost:9000
2024-09-19 21:21:40,746 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-19 21:21:40,766 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-udf_example.pig-926b4e31-e64c-48a7-9a55-ffc0e
2a9d6cc
2024-09-19 21:21:40,766 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-09-19 21:21:40,799 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-19 21:21:41,037 [main] INFO org.apache.pig.scripting.jython.JythonScriptEngine - created tmp python.cachedir=/tmp/pig_jython_1708844982
9640389629
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.python.core.PySystemState (file:/home/srinathi/pig/lib/jython-standalone-2.7.0.jar) to method java.io.
Console.encoding()
WARNING: Please consider reporting this to the maintainers of org.python.core.PySystemState
```

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

```
srinathi@srinathi-VirtualBox:~$ hdfs dfs -ls /home/srinathi/pig_output_data
Found 2 items
-rw-r--r--  3 srinathi supergroup    0 2024-09-19 21:21 /home/srinathi/pig_output_data/_SUCCESS
-rw-r--r--  3 srinathi supergroup   40 2024-09-19 21:21 /home/srinathi/pig_output_data/part-m-00000
srinathi@srinathi-VirtualBox:~$ hdfs dfs -cat /home/srinathi/pig_output_data/part-m-00000
1,SRI
2,VAISH
3,SUBHI
4,PRIYA
5,SWEATHA
srinathi@srinathi-VirtualBox:~$
```

Result:

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