

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
```

```

hadoop@priyav-VirtualBox:~$ nano .bashrc
hadoop@priyav-VirtualBox:~$ source ~/.bashrc
hadoop@priyav-VirtualBox:~$ jps
17312 Jps
9920 SecondaryNameNode
9681 DataNode
10150 ResourceManager
10283 NodeManager
9532 NameNode

```

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

```
hadoop@priyav-VirtualBox:~$ pig
2024-09-02 11:55:06,758 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-02 11:55:06,762 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-02 11:55:06,762 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-02 11:55:06,851 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 20
16, 23:10:49
2024-09-02 11:55:06,852 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoop/pig_17252583068
34.log
2024-09-02 11:55:06,911 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup
not found
2024-09-02 11:55:07,459 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is depr
ecated. Instead, use mapreduce.jobtracker.address
2024-09-02 11:55:07,460 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depreca
ted. Instead, use fs.defaultFS
2024-09-02 11:55:07,460 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting t
o hadoop file system at: hdfs://localhost:9000
2024-09-02 11:55:08,852 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is depreca
ted. Instead, use fs.defaultFS
2024-09-02 11:55:08,920 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-cc78940
d-6226-4ed6-96e0-1e0f8f8b5502
2024-09-02 11:55:08,920 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enable
d 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

```
hadoop@prijay-VirtualBox:~$ nano sample.txt
hadoop@prijay-VirtualBox:~$ hadoop fs -mkdir -p /home/hadoop/piginput
hadoop@prijay-VirtualBox:~$ hadoop fs -put sample.txt /home/hadoop/piginput
hadoop@prijay-VirtualBox:~$ hadoop fs -ls /home/hadoop/piginput
Found 1 items
-rw-r--r-- 3 hadoop supergroup 40 2024-09-02 12:12 /home/hadoop/piginput/sample.txt
hadoop@prijay-VirtualBox:~$ nano demo_pig.pig
hadoop@prijay-VirtualBox:~$ pig demo_pig.pig
2024-09-02 12:13:20,149 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-02 12:13:20,150 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-02 12:13:20,151 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-02 12:13:20,229 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-02 12:13:20,229 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoop/pig_1725259400221.log
2024-09-02 12:13:20,484 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/hadoop/pigbootstrap not found
2024-09-02 12:13:20,553 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-02 12:13:20,553 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-02 12:13:20,553 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-02 12:13:21,031 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-02 12:13:21,070 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-demo_pig.pig-9be6d8c7-0161-41b8-9e6f-470760b29e83
2024-09-02 12:13:21,070 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-09-02 12:13:21,454 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-02 12:13:21,838 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2024-09-02 12:13:21,867 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-02 12:13:21,886 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2024-09-02 12:13:21,933 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator,
lletSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter,
reanTypeCastInserter]}
2024-09-02 12:13:21,989 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThresho
ageThreshold = 489580128
2024-09-02 12:13:22,043 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2024-09-02 12:13:22,081 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2024-09-02 12:13:22,082 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
```

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@priyav-VirtualBox:~$ nano uppercase_udf.py
hadoop@priyav-VirtualBox:~$ hdfs dfs -mkdir /home/hadoop/udfs
hadoop@priyav-VirtualBox:~$ hdfs dfs -put uppercase_udf.py /home/hadoop/udfs/
hadoop@priyav-VirtualBox:~$ nano udf_example.pig
hadoop@priyav-VirtualBox:~$ hadoop fs -put sample.txt /home/hadoop/
hadoop@priyav-VirtualBox:~$ pig -f udf_example.pig
2024-09-02 12:15:11,833 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-02 12:15:11,834 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-02 12:15:11,834 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-02 12:15:11,977 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-02 12:15:11,977 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoop/pig_1725259511957.log
2024-09-02 12:15:12,433 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup not found
2024-09-02 12:15:12,499 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2024-09-02 12:15:12,499 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-02 12:15:12,499 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2024-09-02 12:15:12,948 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-02 12:15:12,995 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-udf_example.pig-836f1b94-89b7-43d8-b96c-f091dc36760e
2024-09-02 12:15:12,996 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
2024-09-02 12:15:13,040 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-02 12:15:13,357 [main] INFO org.apache.pig.scripting.jython.JythonScriptEngine - created tmp python.cachedir=/tmp/pig_jython_4540512934860371218
2024-09-02 12:15:18,095 [main] WARN org.apache.pig.scripting.jython.JythonScriptEngine - pig.cmd.args.reminders is empty. This is not expected unless on testing.
2024-09-02 12:15:18,122 [main] INFO org.apache.pig.scripting.jython.JythonScriptEngine - Register scripting UDF: udf.uppercase
2024-09-02 12:15:18,416 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2024-09-02 12:15:18,425 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
```

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@priyav-VirtualBox:~$ hdfs dfs -ls /home/hadoop/pig_output_data
Found 2 items
-rw-r--r--   3 hadoop supergroup      0 2024-09-02 12:15 /home/hadoop/pig_output_data/_SUCCESS
-rw-r--r--   3 hadoop supergroup    40 2024-09-02 12:15 /home/hadoop/pig_output_data/part-m-00000
hadoop@priyav-VirtualBox:~$ hdfs dfs -cat /home/hadoop/pig_output_data/part-m-00000
1,SRI
2,VAISH
3,SUBHI
4,PRIYA
5,SWEATHA
hadoop@priyav-VirtualBox:~$
```

Result:

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