Exp.No.: 4 **Create UDF in PIG**

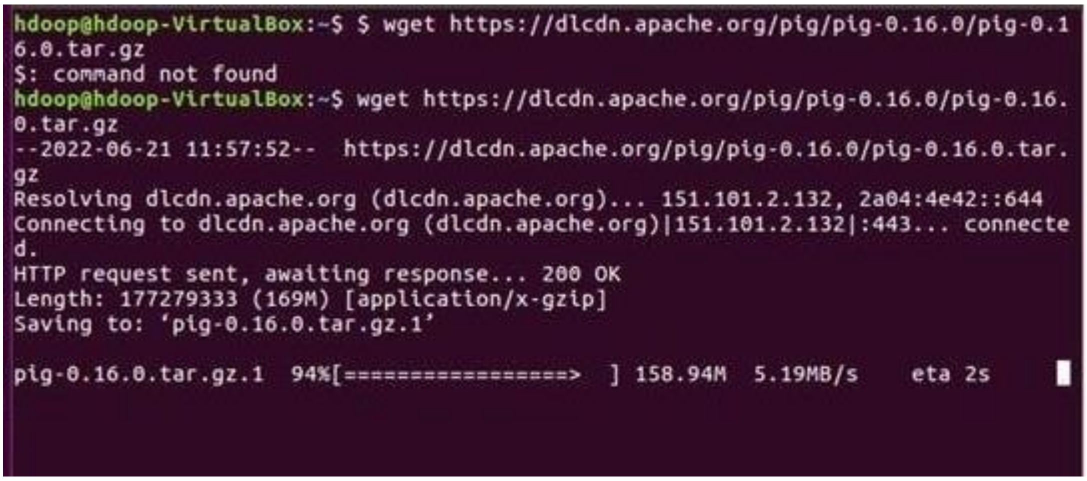
Step-by-step installation of Apache Pig on Hadoop cluster on Ubuntu Pre-requisite:

· 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



```
hdoop@hdoop-VirtualBox:~$ $ wget https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.1
6.0.tar.gz
$: command not found
hdoop@hdoop-VirtualBox:~$ 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... 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  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 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:

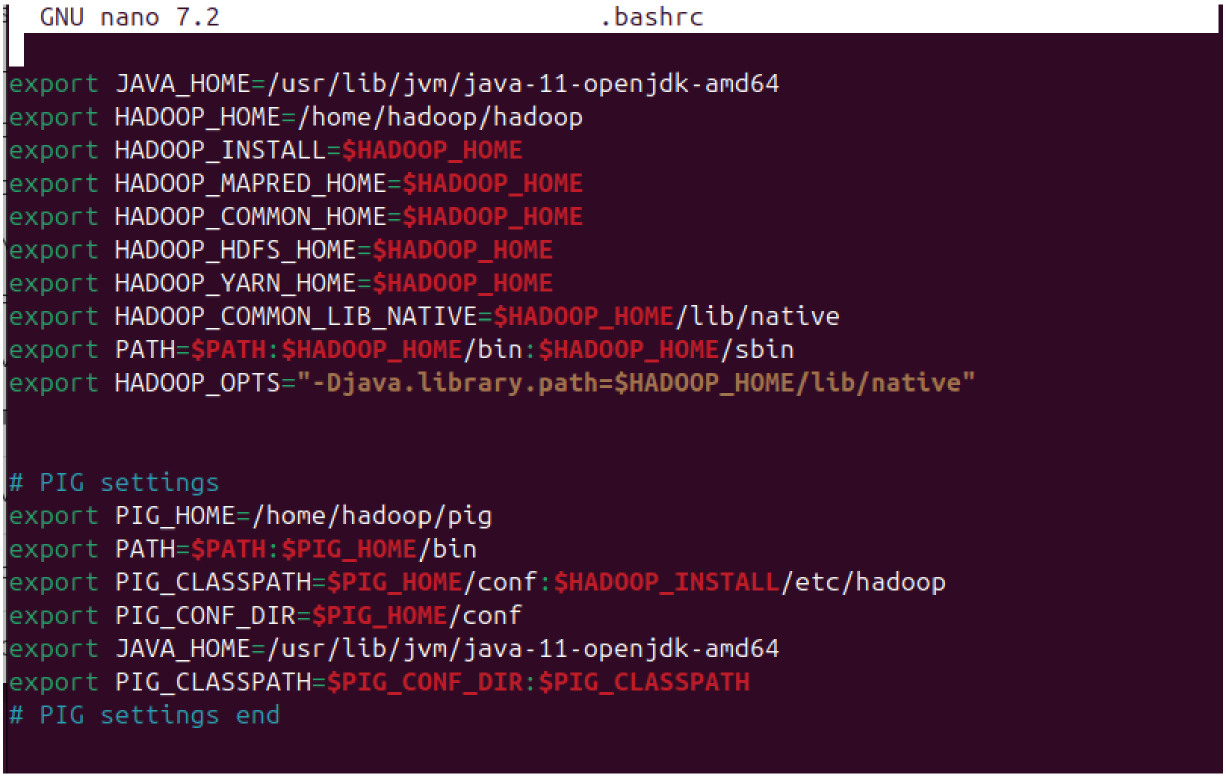
$ sudo mv /home/hadoop/pig-0.16.0 /home/hadoop/pig

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
```



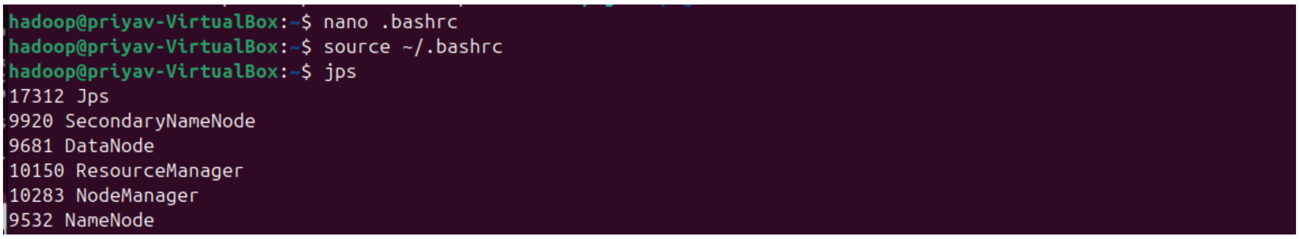
```
  GNU nano 7.2                          .bashrc

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

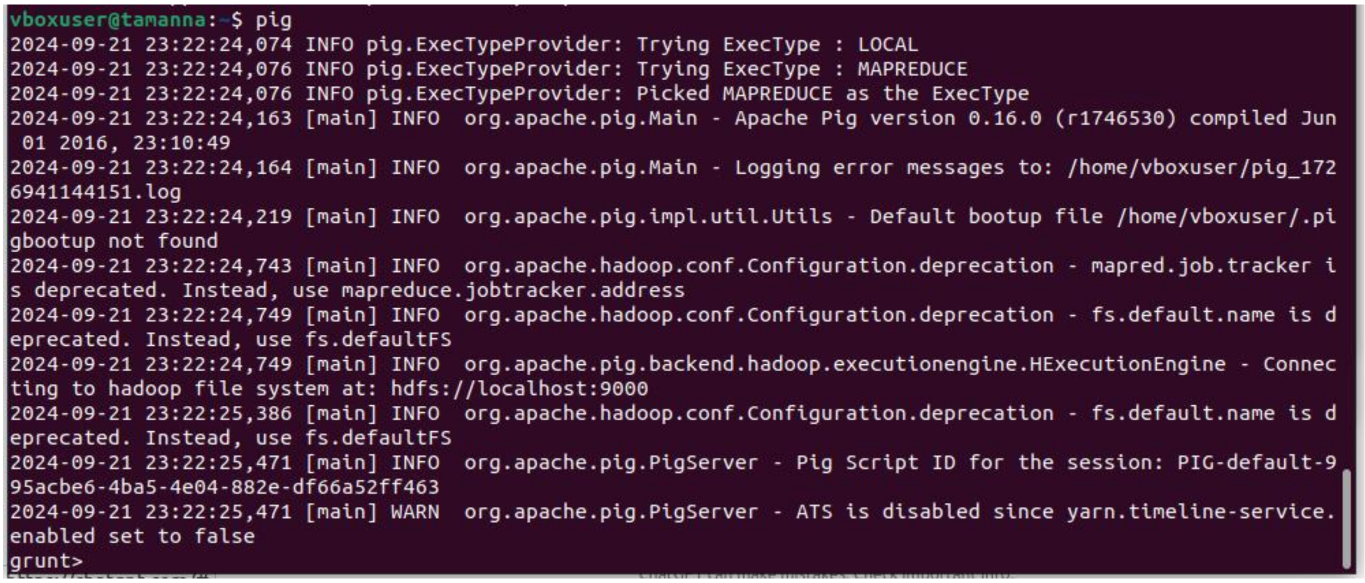
Step 7: To start all Hadoop daemons, navigate to the hadoop-3.2.1/sbin folder and run the following commands:

$ ./start-dfs.sh$ ./start-yarn$ jps



```
hadoop@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



```
vboxuser@tamanna:~$ pig
2024-09-21 23:22:24,074 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-21 23:22:24,076 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-21 23:22:24,076 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-21 23:22:24,163 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun
 01 2016, 23:10:49
2024-09-21 23:22:24,164 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/vboxuser/pig_172
6941144151.log
2024-09-21 23:22:24,219 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/vboxuser/.pi
gbootup not found
2024-09-21 23:22:24,743 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker i
s deprecated. Instead, use mapreduce.jobtracker.address
2024-09-21 23:22:24,749 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is d
eprecated. Instead, use fs.defaultFS
2024-09-21 23:22:24,749 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connec
ting to hadoop file system at: hdfs://localhost:9000
2024-09-21 23:22:25,386 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is d
eprecated. Instead, use fs.defaultFS
2024-09-21 23:22:25,471 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-9
95acbe6-4ba5-4e04-882e-df66a52ff463
2024-09-21 23:22:25,471 [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,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
```

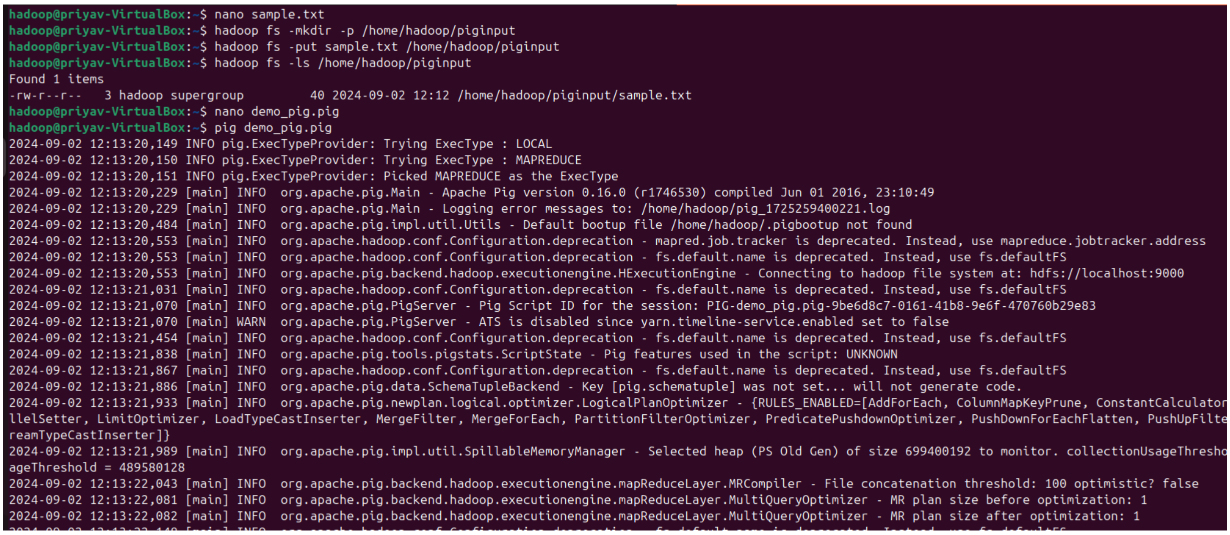
paste the below the content to demo_pig.pig

```
-- 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@priyav-VirtualBox:~$ nano sample.txt
hadoop@priyav-VirtualBox:~$ hadoop fs -mkdir -p /home/hadoop/piginput
hadoop@priyav-VirtualBox:~$ hadoop fs -put sample.txt /home/hadoop/piginput
hadoop@priyav-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@priyav-VirtualBox:~$ nano demo_pig.pig
hadoop@priyav-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 bootup file /home/hadoop/.pigbootup 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
llelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilte
reamTypeCastInserter]}
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';

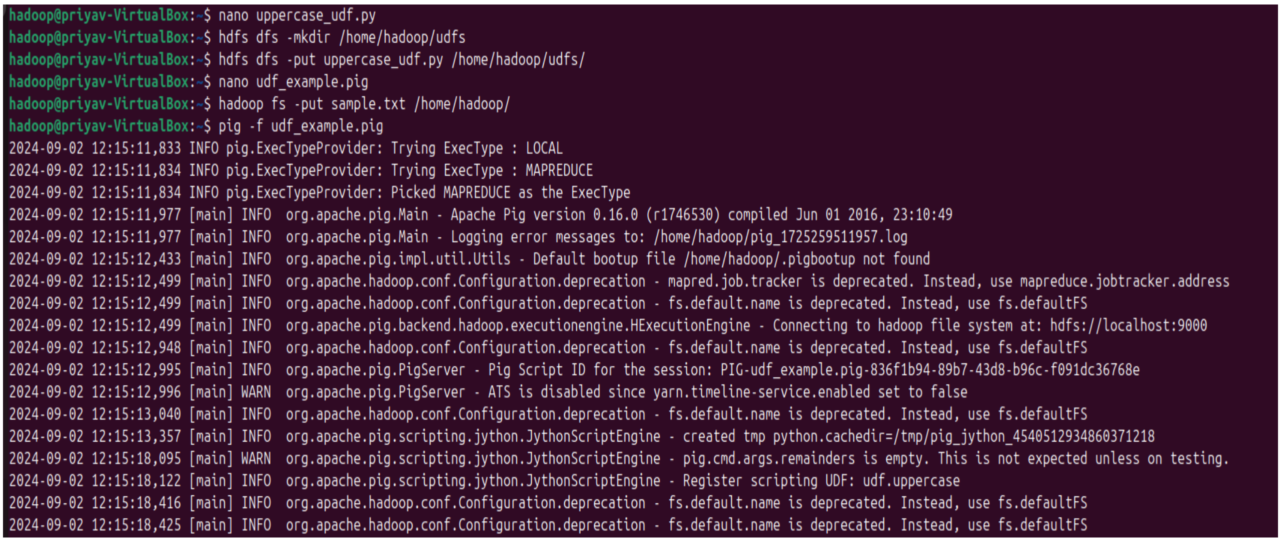
----

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-f091dc36768e
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.remainders 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

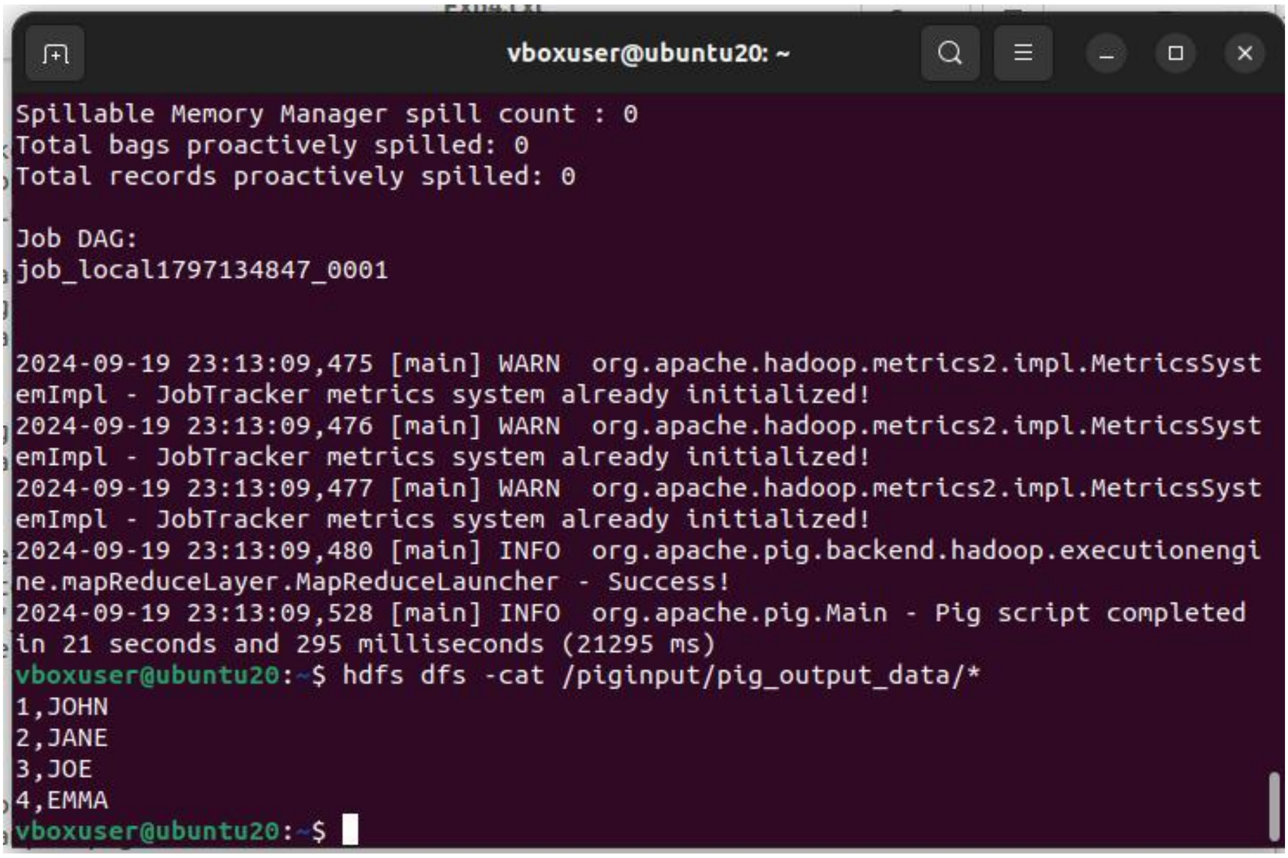
hadoop@Ubuntu:~/Documents$ hdfs dfs -ls /home/hadoop/pig_output_data

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



```
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1797134847_0001


2024-09-19 23:13:09,475 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSyst
emImpl - JobTracker metrics system already initialized!
2024-09-19 23:13:09,476 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSyst
emImpl - JobTracker metrics system already initialized!
2024-09-19 23:13:09,477 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSyst
emImpl - JobTracker metrics system already initialized!
2024-09-19 23:13:09,480 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2024-09-19 23:13:09,528 [main] INFO  org.apache.pig.Main - Pig script completed
in 21 seconds and 295 milliseconds (21295 ms)
vboxuser@ubuntu20:~$ hdfs dfs -cat /piginput/pig_output_data/*
1,JOHN
2,JANE
3,JOE
4,EMMA
vboxuser@ubuntu20:~$
```

Result:

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