EXP 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/hdoop/pig-0.16.0 /home/hdoop/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:

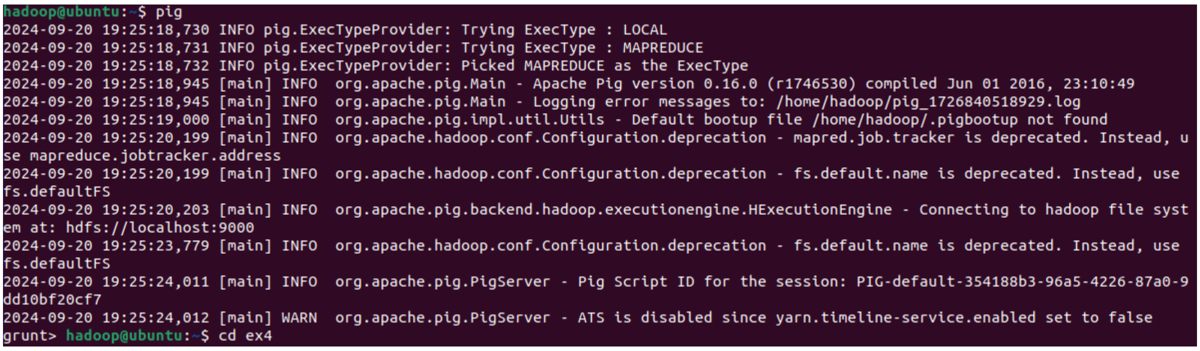
```
$ source .bashrc
```

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



```
hadoop@ubuntu:~$ pig
2024-09-20 19:25:18,730 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-20 19:25:18,731 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-20 19:25:18,732 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-20 19:25:18,945 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-20 19:25:18,945 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/hadoop/pig_1726840518929.log
2024-09-20 19:25:19,000 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup not found
2024-09-20 19:25:20,199 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, u
se mapreduce.jobtracker.address
2024-09-20 19:25:20,199 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-09-20 19:25:20,203 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file syst
em at: hdfs://localhost:9000
2024-09-20 19:25:23,779 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-09-20 19:25:24,011 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-354188b3-96a5-4226-87a0-9
dd10bf20cf7
2024-09-20 19:25:24,012 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> hadoop@ubuntu:~$ cd ex4
```

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,John

2,Jane

3,Joe

4,Emma

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

Create PIG File

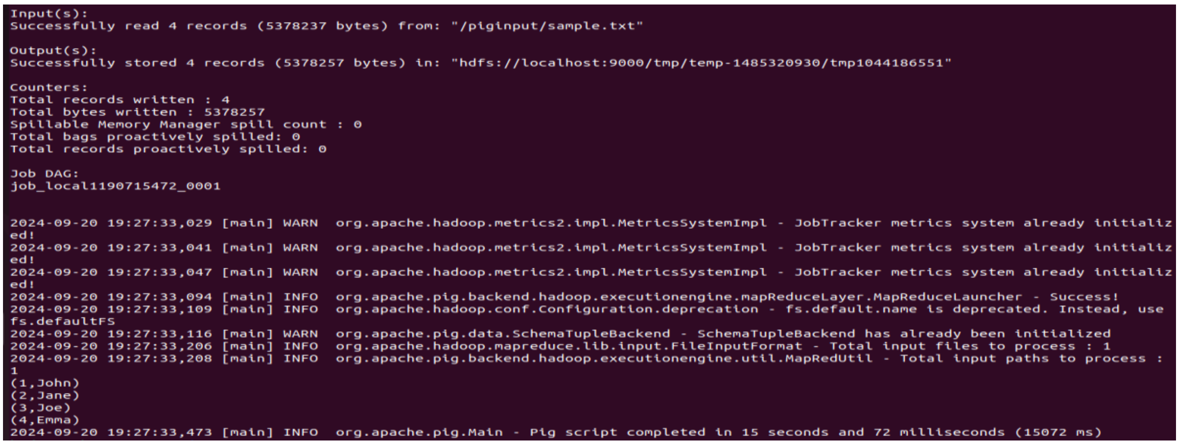
hadoop@Ubuntu:~/Documents$ nano demo_pig.pig

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
```



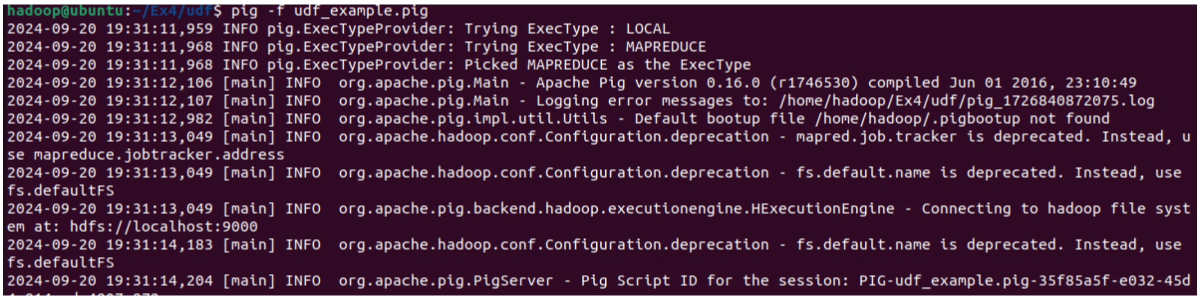
```
Input(s):
Successfully read 4 records (5378237 bytes) from: "/piginput/sample.txt"

Output(s):
Successfully stored 4 records (5378257 bytes) in: "hdfs://localhost:9000/tmp/temp-1485320930/tmp1044186551"

Counters:
Total records written : 4
Total bytes written : 5378257
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1190715472_0001


2024-09-20 19:27:33,029 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
2024-09-20 19:27:33,041 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
2024-09-20 19:27:33,047 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initializ
ed!
2024-09-20 19:27:33,094 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2024-09-20 19:27:33,109 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-09-20 19:27:33,116 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2024-09-20 19:27:33,206 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2024-09-20 19:27:33,208 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process :
1
(1,John)
(2,Jane)
(3,Joe)
(4,Emma)
2024-09-20 19:27:33,473 [main] INFO  org.apache.pig.Main - Pig script completed in 15 seconds and 72 milliseconds (15072 ms)
```

Create udf file an save as uppercase_udf.py

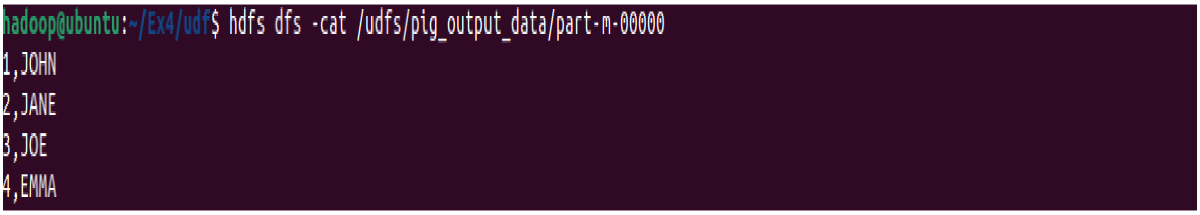
uppercase_udf.py

```
def uppercase(text): return text.upper()

if __name__ == "__main__":

import sys for line

in sys.stdin:

        line = line.strip() result

        = uppercase(line)

        print(result)
```

Output:



```
hadoop@ubuntu:~/Ex4/udf$ pig -f udf_example.pig
2024-09-20 19:31:11,959 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-20 19:31:11,968 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-20 19:31:11,968 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-20 19:31:12,106 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-20 19:31:12,107 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/hadoop/Ex4/udf/pig_1726840872075.log
2024-09-20 19:31:12,982 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup not found
2024-09-20 19:31:13,049 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, u
se mapreduce.jobtracker.address
2024-09-20 19:31:13,049 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-09-20 19:31:13,049 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file syst
em at: hdfs://localhost:9000
2024-09-20 19:31:14,183 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use
fs.defaultFS
2024-09-20 19:31:14,204 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-udf_example.pig-35f85a5f-e032-45d
```



```
hadoop@ubuntu:~/Ex4/udf$ hdfs dfs -cat /udfs/pig_output_data/part-m-00000
1,JOHN
2,JANE
3,JOE
4,EMMA
```