

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 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
```

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.sh
```

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

pig

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

Hello

World vidhiya

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**

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
```

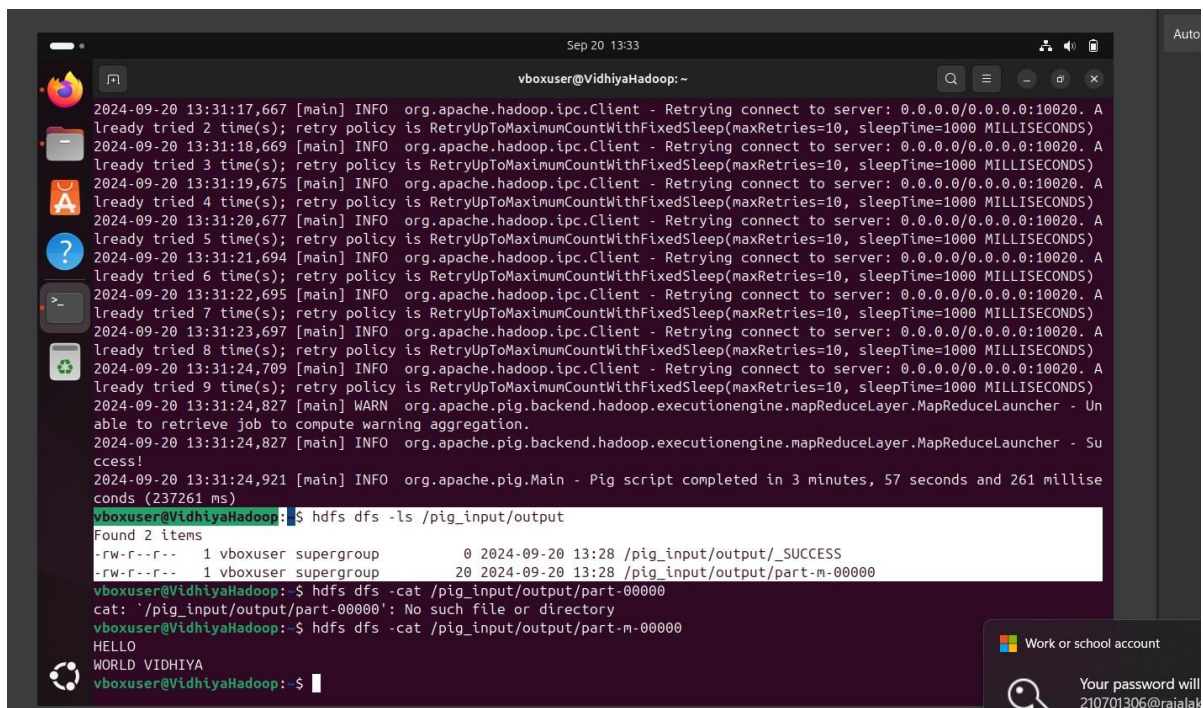
```
in sys.stdin:
```

```
    line = line.strip() result
```

```
    = uppercase(line)
```

```
    print(result)
```

OUTPUT



The screenshot shows a terminal window titled "vboxuser@VidhiyaHadoop: ~" with a dark background. The terminal displays a series of log messages from the Hadoop ecosystem, including warnings about retries and a successful completion of a Pig script. The user then runs the command `hdfs dfs -ls /pig_input/output`, which shows two files: `_SUCCESS` and `part-m-00000`. Subsequent commands `hdfs dfs -cat /pig_input/output/part-00000` and `hdfs dfs -cat /pig_input/output/part-m-00000` are executed, resulting in the output "HELLO WORLD VIDHIYA". A Windows login notification is visible in the bottom right corner of the terminal window.

```
Sep 20 13:33
vboxuser@VidhiyaHadoop: ~
2024-09-20 13:31:17,667 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. A
lready tried 2 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 13:31:18,669 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. A
lready tried 3 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 13:31:19,675 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. A
lready tried 4 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 13:31:20,677 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. A
lready tried 5 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 13:31:21,694 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. A
lready tried 6 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 13:31:22,695 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. A
lready tried 7 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 13:31:23,697 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. A
lready tried 8 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 13:31:24,709 [main] INFO org.apache.hadoop.ipc.Client - Retrying connect to server: 0.0.0.0/0.0.0.0:10020. A
lready tried 9 time(s); retry policy is RetryUpToMaximumCountWithFixedSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-09-20 13:31:24,827 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Un
able to retrieve job to compute warning aggregation.
2024-09-20 13:31:24,827 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Su
ccess!
2024-09-20 13:31:24,921 [main] INFO org.apache.pig.Main - Pig script completed in 3 minutes, 57 seconds and 261 millise
conds (237261 ms)
vboxuser@VidhiyaHadoop: ~$ hdfs dfs -ls /pig_input/output
Found 2 items
-rw-r--r-- 1 vboxuser supergroup 0 2024-09-20 13:28 /pig_input/output/_SUCCESS
-rw-r--r-- 1 vboxuser supergroup 20 2024-09-20 13:28 /pig_input/output/part-m-00000
vboxuser@VidhiyaHadoop: ~$ hdfs dfs -cat /pig_input/output/part-00000
cat: '/pig_input/output/part-00000': No such file or directory
vboxuser@VidhiyaHadoop: ~$ hdfs dfs -cat /pig_input/output/part-m-00000
HELLO
WORLD VIDHIYA
vboxuser@VidhiyaHadoop: ~$
```