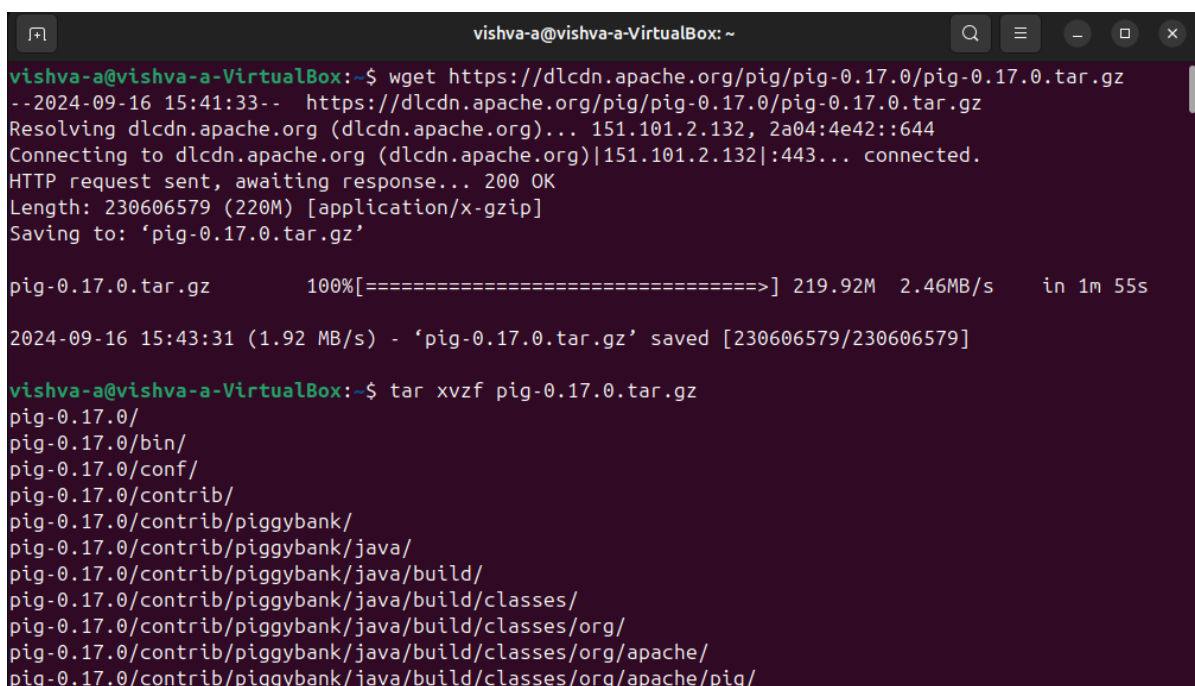


EXP 4: Create UDF in PIG**Step-by-step installation of Apache Pig on Hadoop cluster on Ubuntu Pre-requisite:**

- 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**A terminal window titled 'vishva-a@vishva-a-VirtualBox: ~' showing the process of downloading and extracting Apache Pig. The user runs 'wget https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz', which successfully downloads the file. Then, they run 'tar xvzf pig-0.17.0.tar.gz', which lists the contents of the archive, including directories like bin, conf, contrib, and various Java build directories.

```
vishva-a@vishva-a-VirtualBox: ~  
vishva-a@vishva-a-VirtualBox:~$ wget https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz  
--2024-09-16 15:41:33-- https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.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: 230606579 (220M) [application/x-gzip]  
Saving to: 'pig-0.17.0.tar.gz'  
  
pig-0.17.0.tar.gz      100%[=====] 219.92M  2.46MB/s   in 1m 55s  
  
2024-09-16 15:43:31 (1.92 MB/s) - 'pig-0.17.0.tar.gz' saved [230606579/230606579]  
  
vishva-a@vishva-a-VirtualBox:~$ tar xvzf pig-0.17.0.tar.gz  
pig-0.17.0/  
pig-0.17.0/bin/  
pig-0.17.0/conf/  
pig-0.17.0/contrib/  
pig-0.17.0/contrib/piggybank/  
pig-0.17.0/contrib/piggybank/java/  
pig-0.17.0/contrib/piggybank/java/build/  
pig-0.17.0/contrib/piggybank/java/build/classes/  
pig-0.17.0/contrib/piggybank/java/build/classes/org/  
pig-0.17.0/contrib/piggybank/java/build/classes/org/apache/  
pig-0.17.0/contrib/piggybank/java/build/classes/org/apache/pig/
```

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 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
$ jps
```

```
vishva-a@vishva-a-VirtualBox: ~  
WARNING: resourcemanager did not stop gracefully after 5 seconds: Trying to kill with kill -9  
vishva-a@vishva-a-VirtualBox:~$ cd hadoop-3.3.6/sbin  
vishva-a@vishva-a-VirtualBox:~/hadoop-3.3.6/sbin$ ./start-dfs.sh  
Starting namenodes on [localhost]  
Starting datanodes  
Starting secondary namenodes [vishva-a-VirtualBox]  
vishva-a@vishva-a-VirtualBox:~/hadoop-3.3.6/sbin$ ./start-yarn.sh  
Starting resourcemanager  
Starting nodemanagers  
vishva-a@vishva-a-VirtualBox:~/hadoop-3.3.6/sbin$ jps  
14884 NameNode  
15686 Jps  
15446 ResourceManager  
15575 NodeManager  
15180 SecondaryNameNode  
15005 DataNode  
vishva-a@vishva-a-VirtualBox:~/hadoop-3.3.6/sbin$ cd  
vishva-a@vishva-a-VirtualBox:~$ pig  
2024-09-16 16:00:59,660 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL  
2024-09-16 16:00:59,670 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE  
2024-09-16 16:00:59,670 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType  
2024-09-16 16:00:59,908 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) com  
piled Jun 02 2017, 15:41:58  
2024-09-16 16:00:59,908 [main] INFO org.apache.pig.Main - Logging error messages to: /home/vishva-
```

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

pig

```
vimal@vimal-VirtualBox:~/pig-0.16.0$ cd  
vimal@vimal-VirtualBox:~$ sudo mkdir -p /usr/local/pignew  
[sudo] password for vimal:  
vimal@vimal-VirtualBox:~$ cd pig-0.16.0/  
vimal@vimal-VirtualBox:~/pig-0.16.0$ sudo mv * /usr/local/pignew  
vimal@vimal-VirtualBox:~/pig-0.16.0$ ls  
vimal@vimal-VirtualBox:~/pig-0.16.0$ cd  
vimal@vimal-VirtualBox:~$ nano .bashrc  
vimal@vimal-VirtualBox:~$ source .bashrc  
vimal@vimal-VirtualBox:~$ pig  
2024-09-20 19:57:46,080 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL  
2024-09-20 19:57:46,088 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE  
2024-09-20 19:57:46,088 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType  
2024-09-20 19:57:46,445 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49  
2024-09-20 19:57:46,447 [main] INFO org.apache.pig.Main - Logging error messages to: /home/vimal/pig_1726842466427.log  
2024-09-20 19:57:46,614 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/vimal/.pigbootstrap not found  
2024-09-20 19:57:47,742 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2024-09-20 19:57:47,743 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2024-09-20 19:57:47,743 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000  
2024-09-20 19:57:50,145 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-7f72bae6-e5ce-42df-a096-7e2097b8b9ce  
2024-09-20 19:57:50,145 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false  
grunt> quit  
2024-09-20 19:58:04,190 [main] INFO org.apache.pig.Main - Pig script completed in 18 seconds and 221 milliseconds (18221 ms)  
vimal@vimal-VirtualBox:~$
```

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:~/ 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 script.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>
```

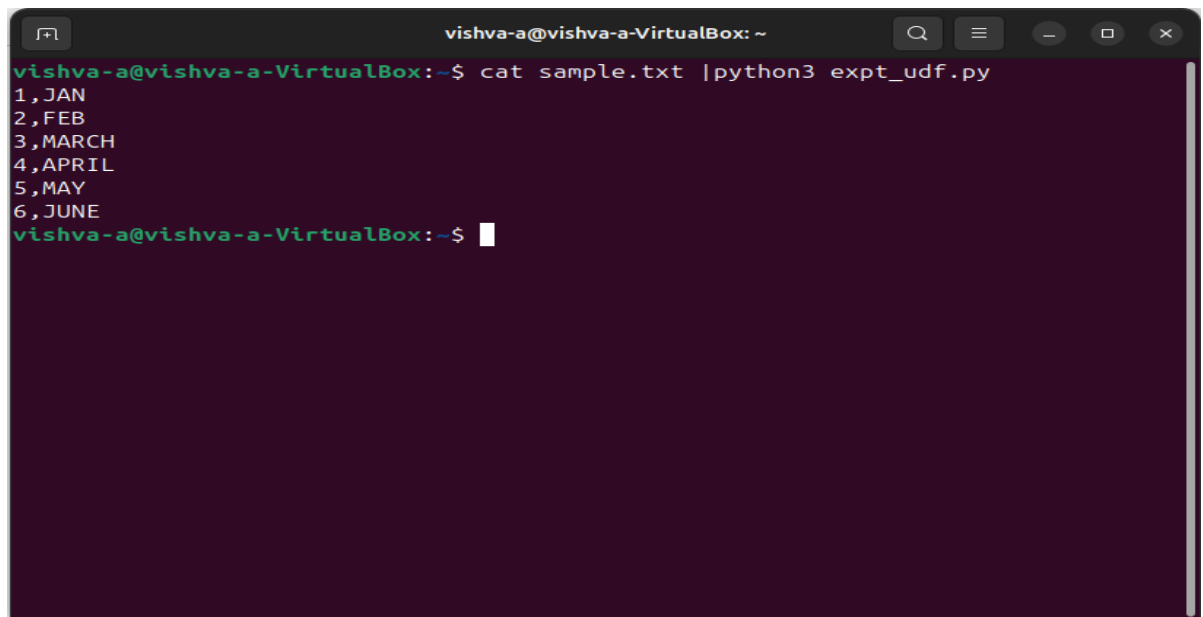
-- Dump the data to check if it was loaded correctly

```
DUMP data;
```

Create a user defined file named as expt_udf.py.

Run the following command,

```
cat sample.txt | python3 expt_udf.py
```

A terminal window titled 'vishva-a@vishva-a-VirtualBox: ~' with standard window controls. The prompt is 'vishva-a@vishva-a-VirtualBox:~\$'. The command 'cat sample.txt |python3 expt_udf.py' has been executed, resulting in the following output:

```
1,JAN
2,FEB
3,MARCH
4,APRIL
5,MAY
6,JUNE
vishva-a@vishva-a-VirtualBox:~$
```

```
vishva-a@vishva-a-VirtualBox:~$ cat sample.txt |python3 expt_udf.py
1,JAN
2,FEB
3,MARCH
4,APRIL
5,MAY
6,JUNE
vishva-a@vishva-a-VirtualBox:~$
```