

Create UDF in PIG

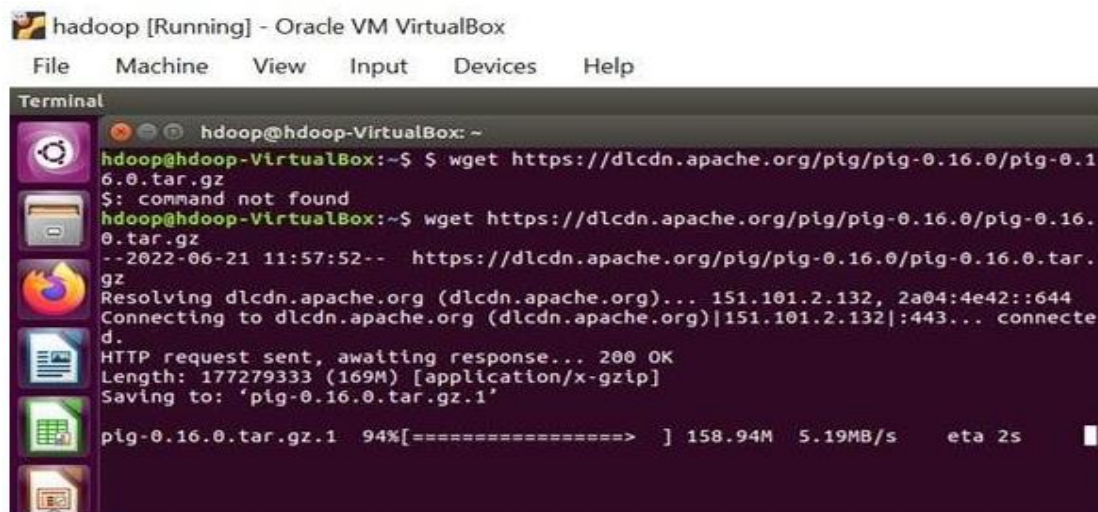
Step-by-step installation of Apache Pig on Hadoop cluster on Ubuntu

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 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Terminal
hadoop@hadoop-VirtualBox: ~
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>



```
hadoop [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Terminal
hadoop@hadoop-VirtualBox: ~
hadoop@hadoop-VirtualBox:~$
```

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.

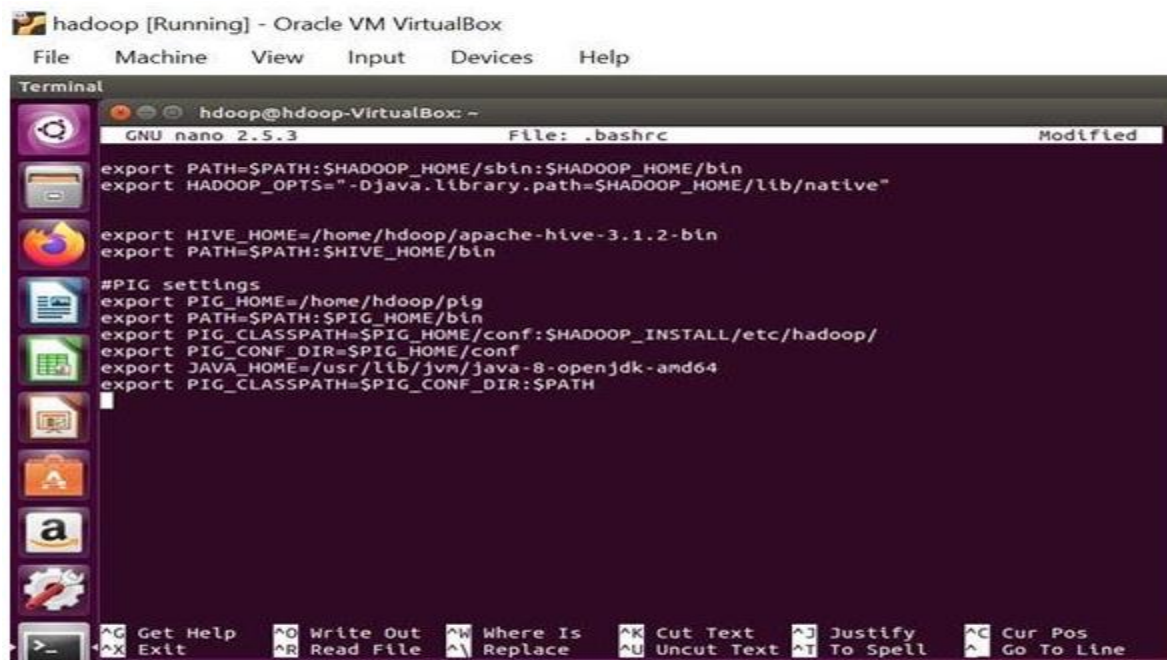
```
#PIG settingsexport PIG_HOME=/home/hadoop/pigexport
```

```
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-openjdk
```

```
amd64export 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
```

```
hadoop@hadoop-VirtualBox:~$ cd hadoop-3.2.1/sbin
hadoop@hadoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [hadoop-VirtualBox]
hadoop@hadoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@hadoop-VirtualBox:~/hadoop-3.2.1/sbin$ jps
4817 DataNode
5298 ResourceManager
5000 SecondaryNameNode
5450 NodeManager
4683 NameNode
5982 Jps
hadoop@hadoop-VirtualBox:~/hadoop-3.2.1/sbin$
```

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

```
$ pig
```

```
hadoop@dell-Inspiron-3443:~$ jps
18768 Jps
17776 DataNode
17636 NameNode
18022 SecondaryNameNode
18269 ResourceManager
18398 NodeManager
hadoop@dell-Inspiron-3443:~$ pig
2024-09-24 10:59:48,815 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-09-24 10:59:48,818 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2024-09-24 10:59:48,818 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2024-09-24 10:59:48,885 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2024-09-24 10:59:48,885 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoop/pig_1727155788875.log
2024-09-24 10:59:48,919 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/hadoop/.pigbootup not found
2024-09-24 10:59:49,345 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use
mapreduce.jobtracker.address
2024-09-24 10:59:49,345 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.
defaultFS
2024-09-24 10:59:49,345 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system
at: hdfs://localhost:9000
2024-09-24 10:59:50,160 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.
defaultFS
2024-09-24 10:59:50,208 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-3b9eaae4-0ca1-4883-93b3-c190
05d80173
2024-09-24 10:59:50,208 [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;
```