# Tutorial 8
# Apache Pig

The commands and screenshots are provided for the deployment of Apache Pig tutorial. You can get the help of Linux commands (Tutorial 1) using the following commands

```
$help cd
$man mkdir
```

and similarly for other Linux commands.
The details of Apache Pig commands can be obtained from

- https://pig.apache.org/docs/latest/cmds.html

and further exploration can be found in the following book as mentioned below

- Programming Pig, 2nd Edition, Alan Gates, Daniel Dai, O'Reilly Media, Inc., November 2016, 368 pages.

1) Apache Pig is an open-source platform for creating programs that run on Apache Hadoop. Download the latest stable release of Pig from the Apache Pig release page http://ftp.heanet.ie/mirrors/www.apache.org/dist/pig/latest/pig-0.17.0.tar.gz


**Note:** The code and data files for this tutorial are available on Moodle.

```
$cd Downloads
$tar -xvf pig-0.17.0.tar.gz
```

2) Install the following commands as shown in the screenshot


```
$sudo mv ./pig-0.17.0  /usr/local
```

3) Create a symbolic link called **pig** to the pig-0.17.0 directory in the **/usr/local** directory:

```
$cd /usr/local/
$sudo ln -sf pig-0.17.0 pig
```

4) Change the ownership of the files in the **pig** directory so that the group is assigned to **Hadoop** and the owner is **hduser**:

```
$sudo chown -R hduser:hadoopgroup pig*
```

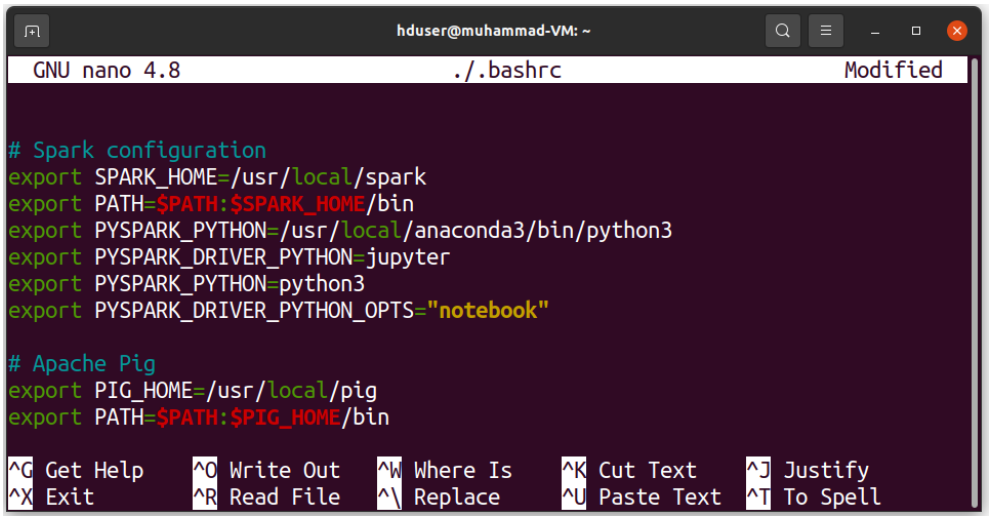
5) Add Pig environment variables to the **.bashrc** file from **/home/hduser**

```
hduser@muhammad-VM:/usr/local$ cd
hduser@muhammad-VM:~$ nano ./.bashrc
```

**$cd Hit Enter Key**
**$nano ./.bashrc**

Add the following two lines at the end of the file '**./.bashrc**' and the screen shot is also provided.

**export PIG_HOME=/usr/local/pig**
**export PATH=$PATH:$PIG_HOME/bin**




save the above lines in the bashrc file and use the command to load

**$source ./.bashrc**

```
hduser@muhammad-VM:~$ source ./.bashrc
hduser@muhammad-VM:~$
```

6) Create a file called **pig_tutorial_sample.txt** with the following content

**$cd Hit the Enter key**
**$cd Desktop**
**$nano pig_tutorial_sample.txt**

```
1,John,Montgomery,Alabama,US
2,David,Phoenix,Arizona,US
3,Sarah,Sacramento,California,US
4,Anoop,Montgomery,Alabama,US
5,Iqbal,Lahore,Punjab,Pakistan
```

Save the above data contents in the **pig_tutorial_sample.txt** file.

```
hduser@muhammad-VM:~$ cd Desktop
hduser@muhammad-VM:~/Desktop$ nano pig_tutorial_sample.txt
hduser@muhammad-VM:~/Desktop$
```

7) Create a file called **pig_tutorial_commands.pig** with the following Pig Latin command on your Ubuntu Desktop. Further create a Pig script file on the Desktop folder

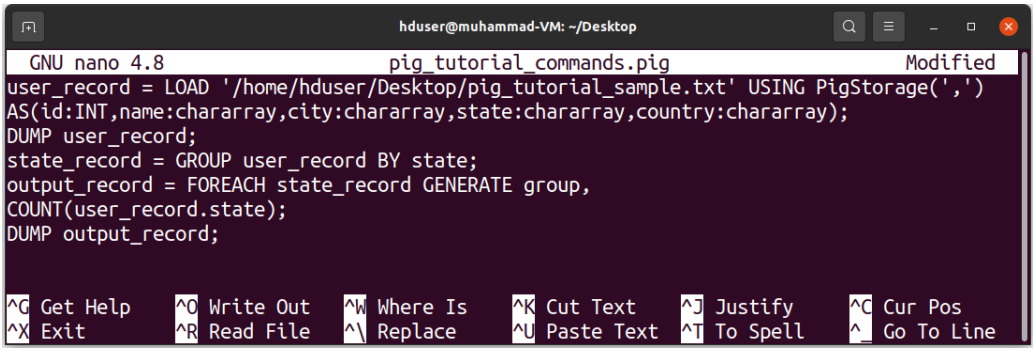
**$nano pig_tutorial_commands.pig**

and store the following lines in the file.

```
user_record = LOAD '/home/hduser/Desktop/pig_tutorial_sample.txt' USING
PigStorage(',')
AS(id:INT,name:chararray,city:chararray,state:chararray,country:chararray);
DUMP user_record;
state_record = GROUP user_record BY state;
output_record = FOREACH state_record GENERATE group,
COUNT(user_record.state);
DUMP output_record;
```

**Note:** Check the lecture notes for the understanding of these commands.

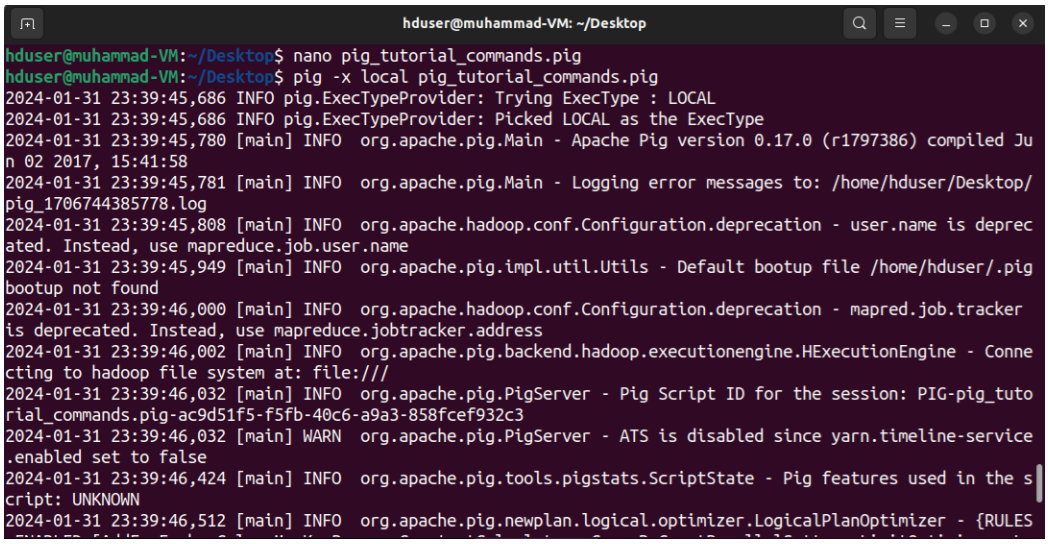
- Note: Use the path of your **local file system (Ubuntu OS)** in pig_tutorial_sample.txt, for example '**/home/hduser/Desktop/pig_tutorial_sample.txt**'



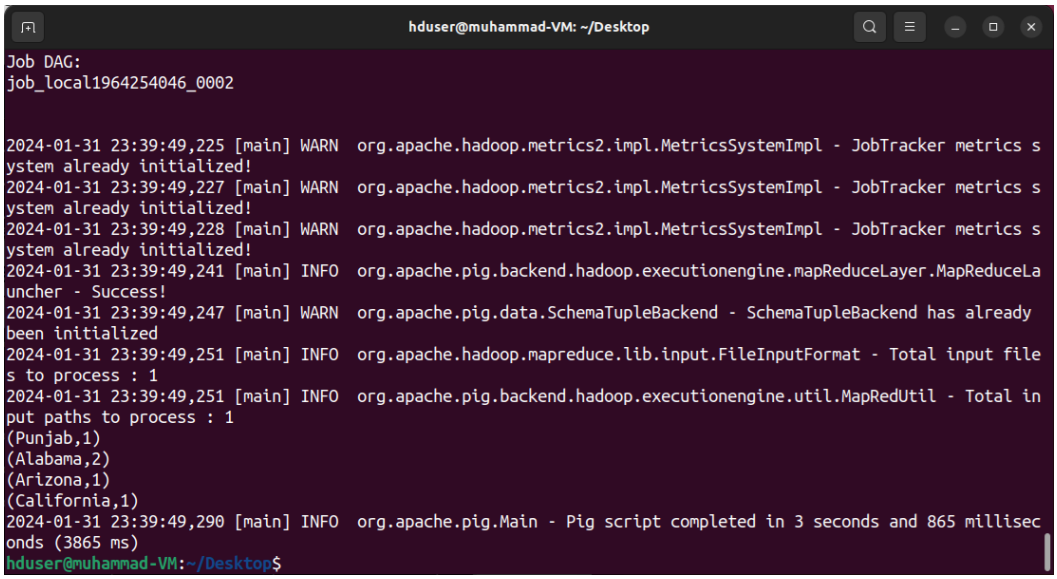

8) Start the Pig grunt using the **pig -x local** command, and then run the script with the command:

**$cd Desktop**

**$pig -x local pig_tutorial_commands.pig**

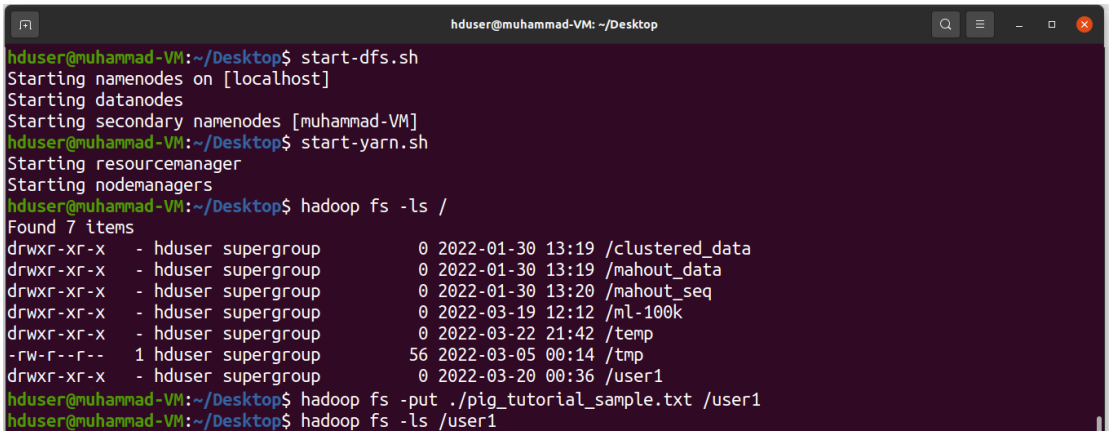



After execution of all steps, the following screen will show the output based on the aggregate function 'count' and It clear from the output as mentioned below




9) If the user would like to read and write the data from hadoop distributed file system (hdfs), we follow the procedure in the following steps.

10) Start hadoop (**start-dfs.sh and start-yarn.sh**) and move the already created text file '**pig_tutorial_sample.txt**' file from local system to hadoop distributed file system. The following screen shot provides the details for this file transfer.




The above screenshot showed that the file '**pig_tutorial_sample.txt**' is transferred to hdfs.

```
$start-dfs.sh
$start-yarn.sh
$hadoop fs -ls /
$hadoop fs -put ./pig_tutorial_sample.txt  /user1
$hadoop fs -ls /user1
```

**Note:** user1 folder has been created on hdfs in the previous tutorials. If you did not have a user1 folder on hdfs, you might see an error. You can use mkdir command to create the folder on hdfs.

11) Now write a script file on your Desktop folder on Ubuntu and copy to the hdfs file system as mentioned below


The code inside the '**pig_script.pig**' file is mentioned below

```
student_record = LOAD 'hdfs://localhost:9000/user1/pig_tutorial_sample.txt'
USING PigStorage(',') as
(id:int,name:chararray,city:chararray,state:chararray,country:chararray);
Dump student_record;
state_record = GROUP student_record BY state;
output_record = FOREACH state_record GENERATE group,
COUNT(student_record.state);
STORE output_record INTO 'hdfs://localhost:9000/user1/student_output/ ' USING
PigStorage(',');
```

12) Now start the **grunt** shell is started by starting pig locally as mentioned below

```
$pig -x local
```

```
hduser@muhammad-VM: ~/Desktop
hduser@muhammad-VM:~/Desktop$ pig -x local
2022-03-26 23:25:04,940 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2022-03-26 23:25:04,940 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2022-03-26 23:25:05,040 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02
 2017, 15:41:58
2022-03-26 23:25:05,042 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/hduser/Desktop/pig_
1648337105036.log
2022-03-26 23:25:05,074 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/hduser/.pigboot
up not found
2022-03-26 23:25:05,251 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is d
eprecated. Instead, use mapreduce.jobtracker.address
2022-03-26 23:25:05,253 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connectin
g to hadoop file system at: file:///
2022-03-26 23:25:05,419 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum i
s deprecated. Instead, use dfs.bytes-per-checksum
2022-03-26 23:25:05,458 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-a926
e753-7370-4e2b-8b33-886a42f5bb37
2022-03-26 23:25:05,461 [main] WARN  org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.ena
bled set to false
grunt>
```

```
grunt>fs -ls
grunt>clear
grunt>exec hdfs://localhost:9000/user1/pig_script.pig
```

The screenshot for the execution of command to read the data from hadoop and write the output on hadoop is mentioned below

```
grunt> exec hdfs://localhost:9000/user1/pig_script.pig
2022-03-27 00:00:18,110 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.check
sum is deprecated. Instead, use dfs.bytes-per-checksum
2022-03-27 00:00:19,187 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.check
sum is deprecated. Instead, use dfs.bytes-per-checksum
2022-03-27 00:00:19,371 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the
script: UNKNOWN
2022-03-27 00:00:19,417 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.check
sum is deprecated. Instead, use dfs.bytes-per-checksum
2022-03-27 00:00:19,457 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULE
S_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer,
LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, Predicate
PushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2022-03-27 00:00:19,546 [main] INFO  org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (Tenu
red Gen) of size 699072512 to monitor. collectionUsageThreshold = 489350752, usageThreshold = 489350752
```

```
hduser@muhammad-VM: ~/Desktop
grunt> exec hdfs://localhost:9000/user1/pig_script.pig
2022-03-27 00:00:18,110 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.ch
ecksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-03-27 00:00:19,187 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.ch
ecksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-03-27 00:00:19,371 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in t
he script: UNKNOWN
2022-03-27 00:00:19,417 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.ch
```

```
hduser@muhammad-VM: ~/Desktop
Output(s):
Successfully stored 4 records (42 bytes) in: "hdfs://localhost:9000/user1/student_output"

Counters:
Total records written : 4
Total bytes written : 42
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local11512335_0005
```

```
hduser@muhammad-VM: ~/Desktop
hduser@muhammad-VM:~/Desktop$ hadoop fs -ls /user1/
Found 3 items
-rw-r--r--   1 hduser supergroup        419 2022-03-27 00:20 /user1/pig_script.pig
-rw-r--r--   1 hduser supergroup        150 2022-03-26 23:50 /user1/pig_tutorial_sample.txt
drwxr-xr-x   - hduser supergroup          0 2022-03-27 00:20 /user1/student_output
hduser@muhammad-VM:~/Desktop$ hadoop fs -cat /user1/student_output
cat: `/user1/student_output': Is a directory
hduser@muhammad-VM:~/Desktop$ hadoop fs -cat /user1/student_output/*
Punjab,1
Alabama,2
Arizona,1
California,1
hduser@muhammad-VM:~/Desktop$
```

The above last screenshot (4 different terminals opened) showed the output read from the hadoop distributed file system (hdfs) as you did for Hadoop, HBase and Spark tutorials. Further exploration of Apache Pig can be obtained from the following references.

**References:**


- https://pig.apache.org/docs/latest/basic.html
- https://www.tutorialspoint.com/apache_pig/index.htm
- https://www.cloudduggu.com/pig/grunt-shell/
- https://www.youtube.com/watch?v=qr_awo5vz0g