**How to Download and Install Pig**

Now in this Apache Pig tutorial, we will learn how to download and install Pig:

Before we start with the actual process, ensure you have Hadoop installed. Change user to ‘hduser’ (id used while Hadoop configuration, you can switch to the userid used during your Hadoop config)

Hadoop PIG Tutorial: Introduction, Installation & Example

**Step 1)** Download the stable latest release of Pig Hadoop from any one of the mirrors sites available at

<http://pig.apache.org/releases.html>



Select **tar.gz** (and not **src.tar.gz)** file to download.

**Step 2)**Once a download is complete, navigate to the directory containing the downloaded tar file and move the tar to the location where you want to setup Pig Hadoop. In this case, we will move to /usr/local

Hadoop PIG Tutorial: Introduction, Installation & Example

Move to a directory containing Pig Hadoop Files

cd /usr/local

Extract contents of tar file as below

sudo tar -xvf pig-0.12.1.tar.gz

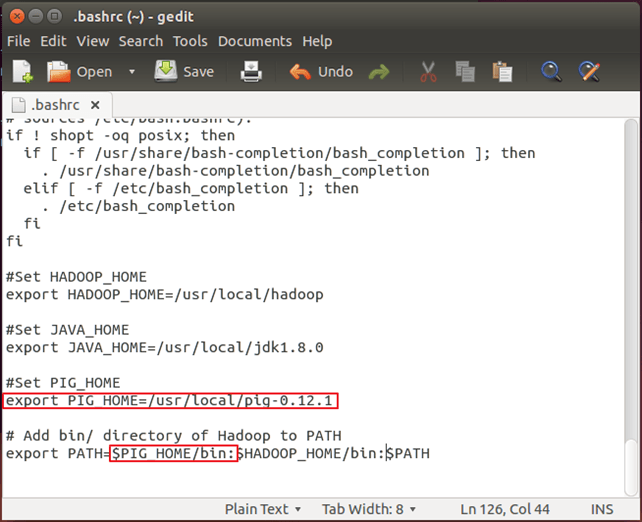
Hadoop PIG Tutorial: Introduction, Installation & Example

**Step 3).** Modify **~/.bashrc** to add Pig related environment variables

Open **~/.bashrc** file in any text editor of your choice and do below modifications-

export PIG\_HOME=<Installation directory of Pig>

export PATH=$PIG\_HOME/bin:$HADOOP\_HOME/bin:$PATH



**Step 4)** Now, source this environment configuration using below command

. ~/.bashrc

Hadoop PIG Tutorial: Introduction, Installation & Example

**Step 5)**We need to recompile **PIG** to support **Hadoop 2.2.0**

Here are the steps to do this-

Go to PIG home directory

cd $PIG\_HOME

Install Ant

sudo apt-get install ant

Hadoop PIG Tutorial: Introduction, Installation & Example

Note: Download will start and will consume time as per your internet speed.

Recompile PIG

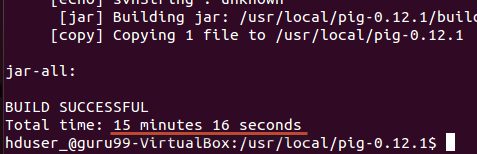
sudo ant clean jar-all -Dhadoopversion=23

Hadoop PIG Tutorial: Introduction, Installation & Example

Please note that in this recompilation process multiple components are downloaded. So, a system should be connected to the internet.

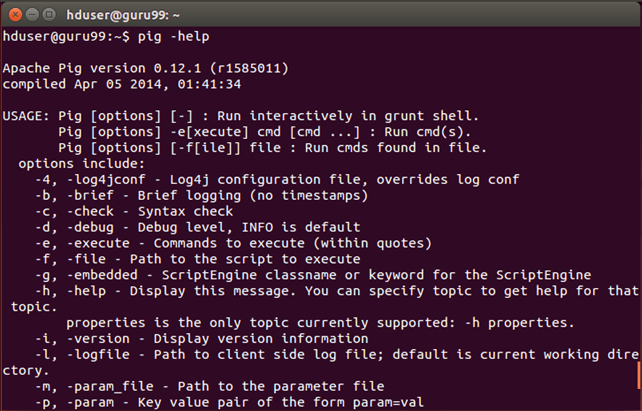
Also, in case this process stuck somewhere and you don’t see any movement on command prompt for more than 20 minutes then press **Ctrl + c** and rerun the same command.

In our case, it takes 20 minutes



**Step 6)** Test the **Pig** installation using the command

pig -help



**Example Pig Script**

We will use Pig Scripts to find the Number of Products Sold in Each Country.

**Input:** Our input data set is a CSV file, [SalesJan2009.csv](https://drive.google.com/uc?export=download&id=1tP8AJGSgDXwI12r2Ap07GyamMj1o0iDD)

**Step 1)**Start Hadoop

$HADOOP\_HOME/sbin/start-dfs.sh

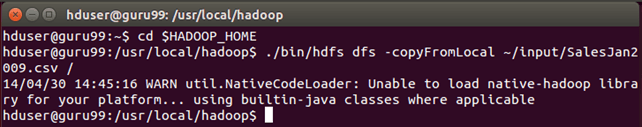
$HADOOP\_HOME/sbin/start-yarn.sh

**Step 2)**Pig in Big Data takes a file from HDFS in MapReduce mode and stores the results back to HDFS.

Copy file **SalesJan2009.csv** (stored on local file system, **~/input/SalesJan2009.csv**) to HDFS (Hadoop Distributed File System) Home Directory

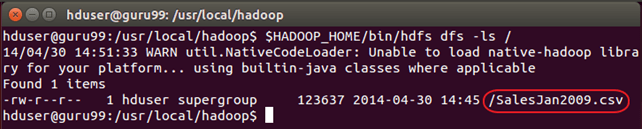
Here in this Apache Pig example, the file is in Folder input. If the file is stored in some other location give that name

$HADOOP\_HOME/bin/hdfs dfs -copyFromLocal ~/input/SalesJan2009.csv /



Verify whether a file is actually copied or not.

$HADOOP\_HOME/bin/hdfs dfs -ls /

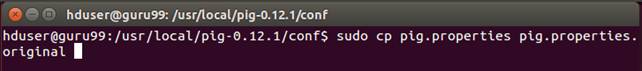


**Step 3)**Pig Configuration

First, navigate to $PIG\_HOME/conf

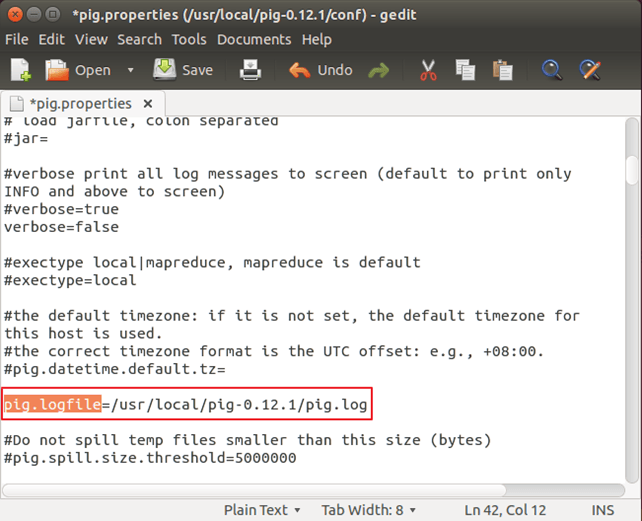
cd $PIG\_HOME/conf

sudo cp pig.properties pig.properties.original



Open **pig.properties**using a text editor of your choice, and specify log file path using **pig.logfile**

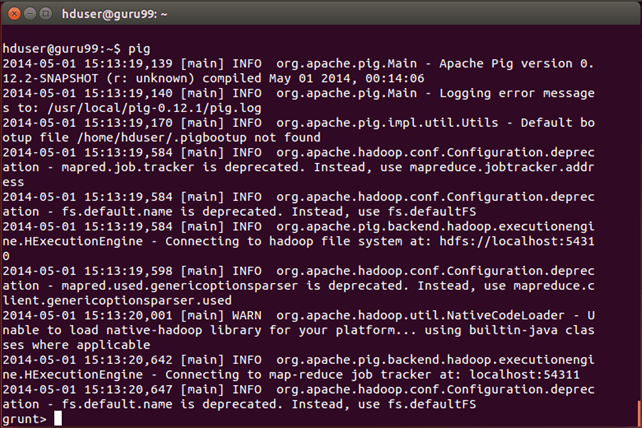
sudo gedit pig.properties



Loger will make use of this file to log errors.

**Step 4)** Run command ‘pig’ which will start Pig command prompt which is an interactive shell Pig queries.

pig

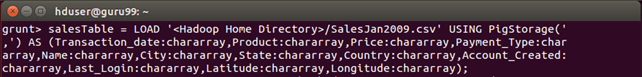


**Step 5)**In Grunt command prompt for Pig, execute below Pig commands in order.

— A. Load the file containing data.

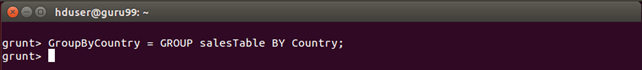
salesTable = LOAD '/SalesJan2009.csv' USING PigStorage(',') AS (Transaction\_date:chararray,Product:chararray,Price:chararray,Payment\_Type:chararray,Name:chararray,City:chararray,State:chararray,Country:chararray,Account\_Created:chararray,Last\_Login:chararray,Latitude:chararray,Longitude:chararray);

Press Enter after this command.



— B. Group data by field Country

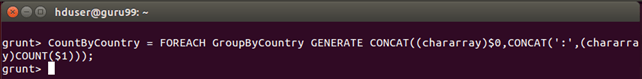
GroupByCountry = GROUP salesTable BY Country;



— C. For each tuple in **‘GroupByCountry’**, generate the resulting string of the form-> Name of Country: No. of products sold

CountByCountry = FOREACH GroupByCountry GENERATE CONCAT((chararray)$0,CONCAT(':',(chararray)COUNT($1)));

Press Enter after this command.

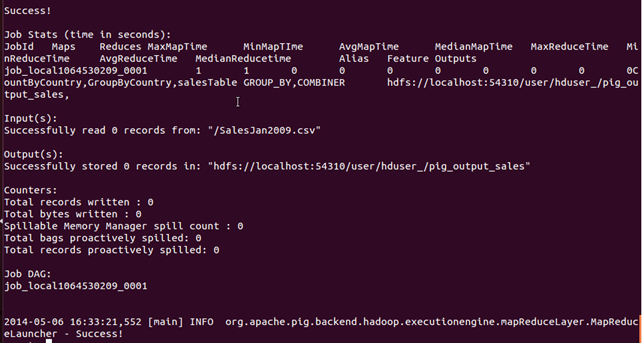


— D. Store the results of Data Flow in the directory **‘pig\_output\_sales’** on HDFS

STORE CountByCountry INTO 'pig\_output\_sales' USING PigStorage('\t');

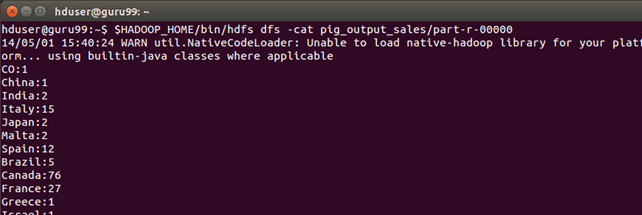
Hadoop PIG Tutorial: Introduction, Installation & Example

This command will take some time to execute. Once done, you should see the following screen



**Step 6)**Result can be seen through command interface as,

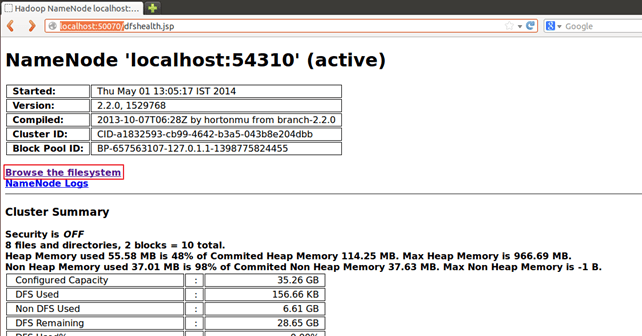
$HADOOP\_HOME/bin/hdfs dfs -cat pig\_output\_sales/part-r-00000



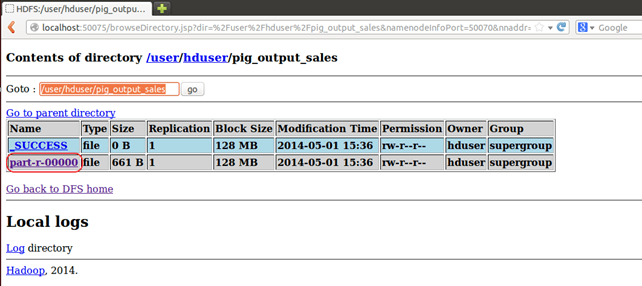
Results can also be seen via a web interface as-

**Results through a web interface-**

Open <http://localhost:50070/> in a web browser.



Now select **‘Browse the filesystem’**and navigate upto **/user/hduser/pig\_output\_sales**



Open **part-r-00000**

