Apache Pig Installation on Linux

Recommended by 91 users

Awanish (https://www.edureka.co/blog/author/awanish/) | Jul 16,2018

In this post, I will talk about *Apache Pig installation on Linux*. Let's start off with the basic definition of Apache Pig and Pig Latin.

Apache Pig is a tool/platform for creating and executing Map Reduce program used with Hadoop. It is a tool/platform for analyzing large sets of data. You can say, Apache Pig is an abstraction over MapReduce. Programmers who are not so good at Java used to struggle working on Hadoop, majorly while writing MapReduce jobs. Apache Pig has its own language *Pig Latin* which is boon for poor programmers.

A basic introduction to Pig Latin will help you understand better:

The high-level procedural language used in Apache Pig platform is called *Pig Latin*. Apache Pig features 'Pig Latin' which is a relatively simpler language which can run over distributed datasets on Hadoop File System (HDFS). In Apache Pig, you need to write Pig scripts using Pig Latin language, which gets converted to MapReduce job when your run you Pig script. Apache Pig has various operators which are used to perform the tasks like reading, writing , processing the data. To learn about Apache Pig operators, go to our blog "***Operators in Apache Pig: Part 1- Relational Operators* (https://www.edureka.co/blog/operators-in-apache-pig/)**".

Now that you have basic understanding of Apache Pig, let us start with Apache Pig Installation on Linux.

Apache Pig Installation on Linux:

Below are the steps for Apache Pig Installation on Linux (ubuntu/centos/windows using Linux VM). I am using Ubuntu 16.04 in below setup.

Step 1: Download **Pig tar** file.

Command: wget http://www-us.apache.org/dist/pig/pig-0.16.0/pig-0.16.0.tar.gz

```
edureka@localhost:~$ wget http://www-us.apache.org/dist/pig/latest/pig-0
.16.0.tar.gz
--2016-11-18 17:46:31--  http://www-us.apache.org/dist/pig/latest/pig-0.
16.0.tar.gz
Resolving www-us.apache.org (www-us.apache.org)... 140.211.11.105
Connecting to www-us.apache.org (www-us.apache.org)|140.211.11.105|:80..
. connected.
HTTP request sent, awaiting response... 200 OK
Length: 177279333 (169M) [application/x-gzip]
Saving to: 'pig-0.16.0.tar.gz'

pig-0.16.0.tar.gz   2%[              ]   4.80M   149KB/s    eta 9m 0s
```

Step 2: Extract the **tar** file using tar command. In below tar command, **x** means extract an archive file, **z** means filter an archive through gzip, **f** means filename of an archive file.

Command: tar -xzf pig-0.16.0.tar.gz

Command: ls

```
edureka@localhost:~$ tar -xzf pig-0.16.0.tar.gz
edureka@localhost:~$ ls
apache-hive-2.1.0-bin          jdk-8u101-linux-i586.tar.gz
apache-hive-2.1.0-bin.tar.gz   Music
derby.log                      Pictures
Desktop                        pig-0.16.0
Documents                      pig-0.16.0.tar.gz
Downloads                      Public
examples.desktop               Templates
hadoop-2.7.3                   Videos
hadoop-2.7.3.tar.gz
edureka@localhost:~$
```

Step 3: Edit the "**.bashrc**" file to update the environment variables of Apache Pig. We are setting it so that we can access pig from any directory, we need not go to pig directory to execute pig commands. Also, if any other application is looking for Pig, it will get to know the path of Apache Pig from this file.

Command: sudo gedit .bashrc

Add the following at the end of the file:

export PIG_HOME=/home/edureka/pig-0.16.0
export PATH=$PATH:/home/edureka/pig-0.16.0/bin
export PIG_CLASSPATH=$HADOOP_CONF_DIR

Also, make sure that hadoop path is also set.

Run below command to make the changes get updated in same terminal.

Command: source .bashrc

Step 4: Check pig version. This is to test that Apache Pig got installed correctly. In case, you don't get the Apache Pig version, you need to verify if you have followed the above steps correctly.

Command: pig -version

```
edureka@localhost:~$ source .bashrc
edureka@localhost:~$ pig -version
Apache Pig version 0.16.0 (r1746530)
compiled Jun 01 2016, 23:10:49
```

Step 5: Check pig help to see all the pig command options.

Command: pig -help

```
edureka@localhost:~$ pig -help

Apache Pig version 0.16.0 (r1746530)
compiled Jun 01 2016, 23:10:49

USAGE: Pig [options] [-] : Run interactively in grunt shell.
       Pig [options] -e[xecute] cmd [cmd ...] : Run cmd(s).
       Pig [options] [-f[ile]] file : Run cmds found in file.
  options include:
    -4, -log4jconf - Log4j configuration file, overrides log conf
    -b, -brief - Brief logging (no timestamps)
    -c, -check - Syntax check
    -d, -debug - Debug level, INFO is default
    -e, -execute - Commands to execute (within quotes)
    -f, -file - Path to the script to execute
    -g, -embedded - ScriptEngine classname or keyword for the ScriptEngi
ne
```

Step 6: Run Pig to start the grunt shell. Grunt shell is used to run Pig Latin scripts.

Command: pig

```
edureka@localhost:~$ pig
16/11/18 18:23:05 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/11/18 18:23:05 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
16/11/18 18:23:05 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2016-11-18 18:23:05,903 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (
r1746530) compiled Jun 01 2016, 23:10:49
2016-11-18 18:23:05,903 [main] INFO  org.apache.pig.Main - Logging error messages to:
/home/edureka/pig_1479473585894.log
2016-11-18 18:23:06,035 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup f
ile /home/edureka/.pigbootup not found
2016-11-18 18:23:07,666 [main] WARN  org.apache.hadoop.util.NativeCodeLoader - Unable
to load native-hadoop library for your platform... using builtin-java classes where ap
plicable
2016-11-18 18:23:07,748 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2016-11-18 18:23:07,748 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-18 18:23:07,749 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HEx
ecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2016-11-18 18:23:09,138 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the
session: PIG-default-09da455b-2390-4800-91f8-9e642ee4ebfe
2016-11-18 18:23:09,139 [main] WARN  org.apache.pig.PigServer - ATS is disabled since
yarn.timeline-service.enabled set to false
grunt>
```

If you look at the above image correctly, Apache Pig has two modes in which it can run, by default it chooses MapReduce mode. The other mode in which you can run Pig is Local mode. Let me tell you more about this.

Execution modes in Apache Pig:

- *MapReduce Mode* – This is the default mode, which requires access to a Hadoop cluster and HDFS installation. Since, this is a default mode, it is not necessary to specify -x flag (you can execute *pig* OR *pig -x mapreduce*). The input and output in this mode are present on HDFS.
- *Local Mode* – With access to a single machine, all files are installed and run using a local host and file system. Here the local mode is specified using '-x flag' (*pig -x local*). The input and output in this mode are present on local file system.

Command: pig -x local

```
edureka@localhost:~$ pig -x local
16/11/22 18:53:12 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
16/11/22 18:53:12 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2016-11-22 18:53:12,375 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (
r1746530) compiled Jun 01 2016, 23:10:49
2016-11-22 18:53:12,375 [main] INFO  org.apache.pig.Main - Logging error messages to:
/home/edureka/pig_1479820992372.log
2016-11-22 18:53:12,458 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup f
ile /home/edureka/.pigbootup not found
2016-11-22 18:53:12,836 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2016-11-22 18:53:12,842 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- fs.default.name is deprecated. Instead, use fs.defaultFS
2016-11-22 18:53:12,844 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HEx
ecutionEngine - Connecting to hadoop file system at: file:///
2016-11-22 18:53:13,272 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation
- io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2016-11-22 18:53:13,328 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the
session: PIG-default-e8d202a8-bf5a-4fb6-b765-777af4fe7b18
2016-11-22 18:53:13,328 [main] WARN  org.apache.pig.PigServer - ATS is disabled since
yarn.timeline-service.enabled set to false
grunt>
```

You can go through below video to watch Apache Pig Installation on Linux:

Now that you are done with Apache Pig Installation on Linux, the next step forward is to try out some relational Pig operators on Pig Grunt shell. Hence, the next blog "***Operators in Apache Pig: Part 1- Relational Operators* (https://www.edureka.co/blog/operators-in-apache-pig/)**" will help you to master Pig operators.

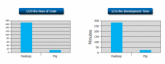
*Now that you have installed Apache Pig on Linux, check out the **Hadoop training (https://www.edureka.co/big-data-and-hadoop)** by Edureka, a trusted online learning company with a network of more than 250,000 satisfied learners spread across the globe. The Edureka Big Data Hadoop Certification Training course helps learners become expert in HDFS, Yarn, MapReduce, Pig, Hive, HBase, Oozie, Flume and Sqoop using real-time use cases on Retail, Social Media, Aviation, Tourism, Finance domain.*

Got a question for us? Please mention it in the comments section and we will get back to you.

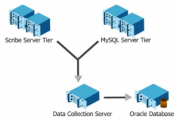
About Awanish (19 Posts (https://www.edureka.co/blog/author/awanish/))

Awanish is a Sr. Research Analyst at Edureka. He has rich expertise in Big Data technologies like Hadoop, Spark, Storm, Kafka, Flink. Awanish also holds good knowledge on Devops tools like Git, Jenkins, Docker, Puppet, Nagios. He is passionate about new technologies which provide realistic solutions.

Related Posts



Introduction to Pig
5.2K
(https://www.edureka.co/blog/introduction-to-pig/)



Pig Vs Hive
2.7K
(https://www.edureka.co/blog/pig-vs-hive/)

Operators in Apache Pig: Part 1- Relational Operators
15.5K
(https://www.edureka.co/blog/operators-in-apache-pig/)

Pig Programming: Create Your First Apache Pig Script
34K
(https://www.edureka.co/blog/pig-programming-create-your-first-apache-pig-script/)

Browse Categories

Big Data NoSQL (https://www.edureka.co/blog/category/big-data-nosql/) | Blockchain (https://www.edureka.co/blog/category/blockchain/) | Business Intelligence (https://www.edureka.co/blog/category/business-intelligence/) | Cloud Computing (https://www.edureka.co/blog/category/cloud-computing/) | Cyber Security (https://www.edureka.co/blog/category/cyber-security/) | Deep Learning (https://www.edureka.co/blog/category/deep-learning/) | Finance (https://www.edureka.co/blog/category/finance/) | Frameworks (https://www.edureka.co/blog/category/frameworks/) | Marketing (https://www.edureka.co/blog/category/marketing/) | Mobile Development (https://www.edureka.co/blog/category/mobile-development/) | Operations (https://www.edureka.co/blog/category/operations/) | Programming (https://www.edureka.co/blog/category/programming/) | Project Management (https://www.edureka.co/blog/category/project-management/) | Robotic Process Automation (https://www.edureka.co/blog/category/robotic-process-automation/) | Success Story (https://www.edureka.co/blog/category/success-story/) | Systems & Architecture (https://www.edureka.co/blog/category/systems-architecture/) | Systems Engineering (https://www.edureka.co/blog/category/systems-engineering/) | Testing (https://www.edureka.co/blog/category/testing/)

Related Blogs

Top Hadoop Interview Questions On Apache PIG For 2018 (https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-pig/)

(https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-pig/)

Introduction to Pig (https://www.edureka.co/blog/introduction-to-pig/)

(https://www.edureka.co/blog/introduction-to-pig/)

Pig Vs Hive (https://www.edureka.co/blog/pig-vs-hive/)

(https://www.edureka.co/blog/pig-vs-hive/)

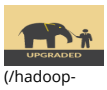
Operators in Apache Pig: Part 1- Relational Operators (https://www.edureka.co/blog/operators-in-apache-pig/)

(https://www.edureka.co/blog/operators-in-apache-pig/)

Big Data Analytics Courses


(/big-data-and-hadoop)

Apache Spark and Scala Certification Training (/apache-spark-scala-training)
(/apache-spark-scala-training)



Hadoop Administration Certification Training (/hadoop-admin)
(/hadoop-admin)


Splunk Power User & Admin Certification Training (/splunk)
(/splunk)