

## Apache Pig Installation



The objective of this tutorial is to describe step by step process to install Pig (Version pig-0.17.0.tar.gz ) on Hadoop 3.1.2 version and the OS which we are using is Ubuntu 18.04.4 LTS (Bionic Beaver), once the installation is completed you can play with Pig.

### Platform

- **Operating System (OS).** You can use Ubuntu 18.04.4 LTS version or later version, also you can use other flavors of Linux systems like Redhat, CentOS, etc.
- **Hadoop.** We have already installed Hadoop 3.1.2 version on which we will run Pig (Please refer to the "[Hadoop Installation on Single Node](#)" tutorial and install Hadoop first before proceeding for Pig installation.)
- **Pig.** We have used the Apache Pig-0.17.0 version for installation.

### Download Software

- **Pig**  
<https://downloads.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz>


### Steps to Install Apache Pig version(0.17.0) on Ubuntu 18.04.4 LTS

Please follow the below steps to install Apache Pig.

**Step 1.** Please verify if Hadoop is installed.

**On Linux:** `$wget https://downloads.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz`

**On Windows:** `https://downloads.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz`

 cloudduggu@ubuntu: ~

```
cloudduggu@ubuntu:~$ wget https://downloads.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz
```

 cloudduggu@ubuntu: ~

```
cloudduggu@ubuntu:~$ wget https://downloads.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz
--2020-07-07 04:40:48-- https://downloads.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219, 2a01:4f8:10a:201a::2
Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|: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
2020-07-07 04:47:31 (563 KB/s) - 'pig-0.17.0.tar.gz' saved [230606579/230606579]

cloudduggu@ubuntu:~$
```

```
$mv pig-0.17.0 pig
```

```
cloudduggu@ubuntu: ~
```

```
cloudduggu@ubuntu:~$ tar -xzf pig-0.17.0.tar.gz  
cloudduggu@ubuntu:~$
```

```
cloudduggu@ubuntu: ~
```

```
cloudduggu@ubuntu:~$ mv pig-0.17.0 pig  
cloudduggu@ubuntu:~$
```

**Step 5.** Now edit the `.bashrc` file to update the environment variable of Apache Pig so that it can be accessed from any directory.

```
$nano .bashrc
```

```
export PATH=$PATH:/home/cloudduggu/pig/bin
```

```
export PIG_CLASSPATH=$HADOOP_HOME/etc/Hadoop
```

cloudduggu@ubuntu: ~

GNU nano 2.9.3

.bashrc

```
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

export PDSH_RCMD_TYPE=ssh
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/
export HADOOP_HOME=/home/cloudduggu/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export MAPRED_HOME=${HADOOP_HOME}
export HDFS_HOME=${HADOOP_HOME}
export YARN_HOME=${HADOOP_HOME}
export SPARK_HOME=/home/cloudduggu/spark
export HIVE_HOME=/home/cloudduggu/hive
export PATH=$PATH:$HIVE_HOME/bin

export PIG_HOME=/home/cloudduggu/pig
export PATH=$PATH:/home/cloudduggu/pig/bin
export PIG_CLASSPATH=$HADOOP_HOME/etc/Hadoop
```

**Step 6.** Run source command to update changes in the same terminal.

```
$source .bashrc
```

**Step 7.** Now run the pig version command to make sure that Pig is installed properly.

```
cloudduggu@ubuntu: ~
```

```
cloudduggu@ubuntu:~$ pig -version  
Apache Pig version 0.17.0 (r1797386)  
compiled Jun 02 2017, 15:41:58  
cloudduggu@ubuntu:~$
```

**Step 8.** Run pig help command to see all pig command options.



cloudduggu@ubuntu: ~

```
cloudduggu@ubuntu:~$ pig -help
```

```
Apache Pig version 0.17.0 (r1797386)
compiled Jun 02 2017, 15:41:58
```

```
USAGE: Pig [options] [-] : Run interactively in grunt shell.
      Pig [options] -e[execute] cmd [cmd ...] : Run cmd(s).
      Pig [options] [-f[file]] 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 ScriptEngine
- h, -help - Display this message. You can specify topic to get help for that topic. properties is the only topic currently supported: -h properties.
- i, -version - Display version information
- l, -logfile - Path to client side log file; default is current working directory.
- m, -param\_file - Path to the parameter file
- p, -param - Key value pair of the form param=val
- r, -dryrun - Produces script with substituted parameters. Script is not executed.
- t, -optimizer\_off - Turn optimizations off. The following values are supported:
  - ConstantCalculator - Calculate constants at compile time
  - SplitFilter - Split filter conditions
  - PushUpFilter - Filter as early as possible
  - MergeFilter - Merge filter conditions
  - PushDownForeachFlatten - Join or explode as late as possible
  - LimitOptimizer - Limit as early as possible
  - ColumnMapKeyPrune - Remove unused data
  - AddForEach - Add ForEach to remove unneeded columns
  - MergeForEach - Merge adjacent ForEach
  - GroupByConstParallelSetter - Force parallel 1 for "group all" statement
  - PartitionFilterOptimizer - Pushdown partition filter conditions to loader implementing LoadMetaData
  - PredicatePushdownOptimizer - Pushdown filter predicates to loader implementing LoadPredicatePushDown
  - All - Disable all optimizations

**Step 9.** Now start Pig grunt shell (grunt shell is used to execute Pig Latin scripts).

```
cloudduggu@ubuntu: ~  
cloudduggu@ubuntu:~$ pig  
2020-07-07 06:18:25,300 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL  
2020-07-07 06:18:25,310 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE  
2020-07-07 06:18:25,310 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType  
2020-07-07 06:18:25,672 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58  
2020-07-07 06:18:25,680 [main] INFO org.apache.pig.Main - Logging error messages to: /home/cloudduggu/pig_1594127905668.log  
2020-07-07 06:18:25,860 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/cloudduggu/.pigbootup not found  
2020-07-07 06:18:28,536 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2020-07-07 06:18:28,537 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000  
2020-07-07 06:18:33,415 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-791af914-342b-4aa5-a032-90f420796282  
2020-07-07 06:18:33,416 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false  
grunt> █
```