

Go to the following link and download pig

<http://mirrors.estointernet.in/apache/pig/pig-0.16.0/>

To untar pig-0.16.0.tar.gz file run the following command:

```
$ tar xvfz pig-0.16.0.tar.gz
```

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 /usr/local/hadoop/pig
```

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 settings
export PIG_HOME=/usr/local/hadoop/pig-0.16.0
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop
export PIG_CONF_DIR=$PIG_HOME/conf
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH
#PIG setting ends
(to exit ctrl +x )
```

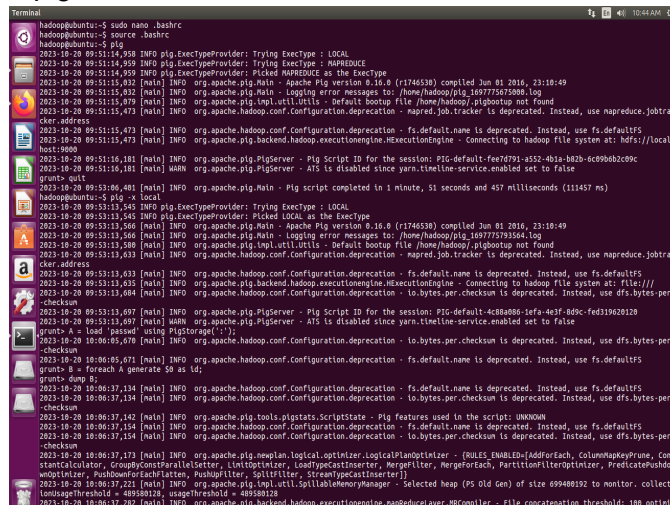
Run the following command to make the changes effective in the `.bashrc` file:

```
$ source .bashrc
```

```
start all Hadoop daemons
cd /usr/local/hadoop/bin
start-all.sh
jps
```

Now you can launch pig by executing the following command:

```
$ pig
```

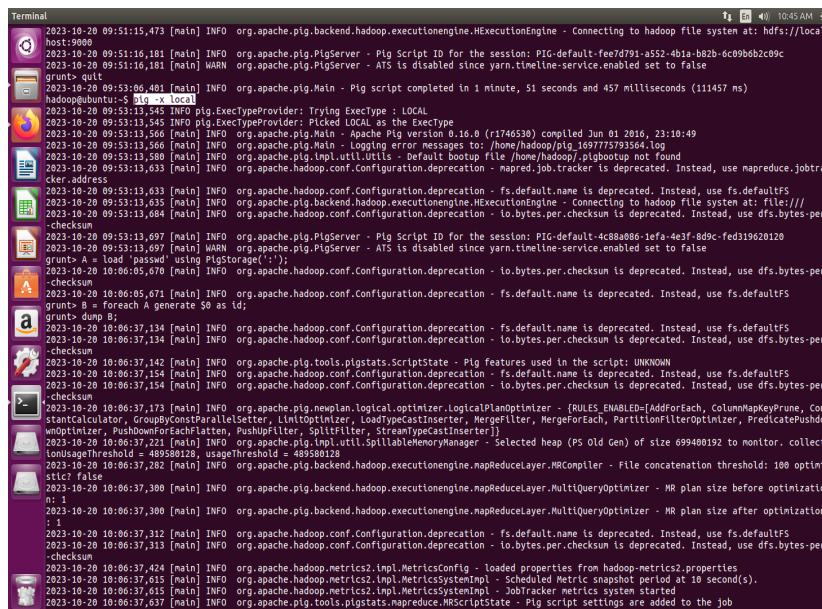


Now you are in pig and can perform your desired tasks on pig. You can come out of the pig by the quit command:

```
> quit;
```

Quit and run pig in local mode

```
pig -x local
```



```
2023-10-20 09:51:15,473 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://local
host:9000
2023-10-20 09:51:16,181 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-fee7d791-a552-4b1a-b82b-6c09b6b2c09c
2023-10-20 09:51:16,181 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> quit
2023-10-20 09:53:06,401 [main] INFO org.apache.pig.Main - Pig script completed in 1 minute, 51 seconds and 457 milliseconds (111457 ms)
hadoop@ubuntu:~$ pig -x local
2023-10-20 09:53:13,545 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2023-10-20 09:53:13,545 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2023-10-20 09:53:13,566 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746538) compiled Jun 01 2016, 23:10:49
2023-10-20 09:53:13,566 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoop/pig_1697775793564.log
2023-10-20 09:53:13,580 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/hadoop/.pigbootstrap not found
2023-10-20 09:53:13,633 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtra
cker.address
2023-10-20 09:53:13,633 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-10-20 09:53:13,635 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2023-10-20 09:53:13,684 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per
-checksum
2023-10-20 09:53:13,697 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-4c8a806-1afa-de3f-8d9c-fed319620120
2023-10-20 09:53:13,697 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> A = load 'passwd' using PigStorage('');
2023-10-20 10:06:05,670 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per
-checksum
2023-10-20 10:06:05,671 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = foreach A generate $0 as id;
2023-10-20 10:06:37,134 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-10-20 10:06:37,134 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per
-checksum
2023-10-20 10:06:37,142 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2023-10-20 10:06:37,154 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-10-20 10:06:37,154 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per
-checksum
2023-10-20 10:06:37,173 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - [RULES ENABLED=[AddForEach, ColumnMapKeyPrune, Con
stantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdo
wnOptimizer, PushdownForEachLatten, PushupFilter, SplitFilter, StreamTypeCastInserter]]
2023-10-20 10:06:37,221 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699408192 to monitor. collect
ionUsageThreshold = 489580128, usageThreshold = 489580128
2023-10-20 10:06:37,282 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimi
stic false
2023-10-20 10:06:37,300 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimizatio
n: 1
2023-10-20 10:06:37,300 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization
: 1
2023-10-20 10:06:37,312 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-10-20 10:06:37,313 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per
-checksum
2023-10-20 10:06:37,424 [main] INFO org.apache.hadoop.metrics2.impl.MetricsConfig - loaded properties from hadoop-metrics2.properties
2023-10-20 10:06:37,615 [main] INFO org.apache.hadoop.metrics2.impl.MetricsSystemImpl - Scheduled Metric snapshot period at 10 second(s).
2023-10-20 10:06:37,615 [main] INFO org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system started
2023-10-20 10:06:37,637 [main] INFO org.apache.pig.tools.pigstats.MapReduceScriptState - Pig script settings are added to the job
```

You can run Pig in interactive mode using the Grunt shell. Invoke the Grunt shell using the "pig" command and then enter your Pig Latin statements and Pig commands interactively at the command line.

Example

These Pig Latin statements extract all user IDs from the /etc/passwd file. First, copy the /etc/passwd file to your local working directory. Then, enter the Pig Latin statements interactively at the grunt prompt. The DUMP operator will display the results to your terminal screen.

```
grunt> A = load 'passwd' using PigStorage('');
```

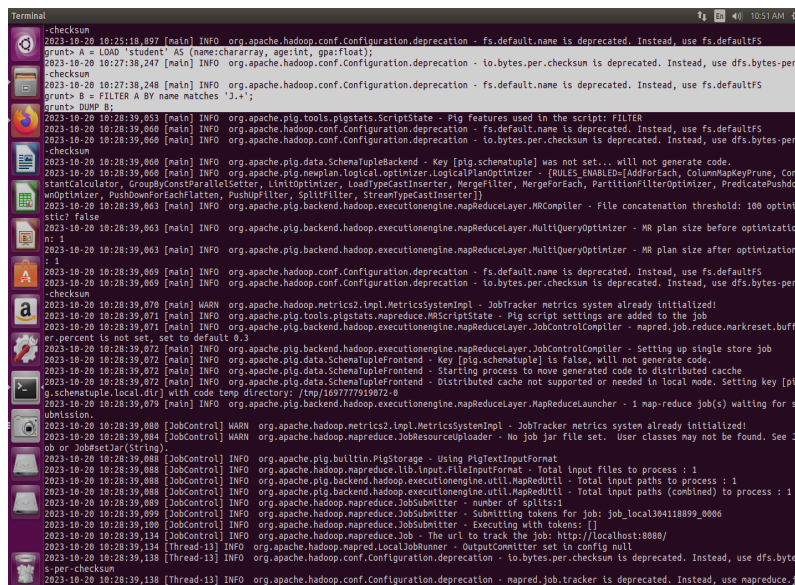

A = load 'passwd' using PigStorage(':'); -- load the passwd file
B = foreach A generate \$0 as id; -- extract the user IDs
store B into '/home/hadoop/id.out'; -- write the results to a file name id.out

A = LOAD 'student' AS (name:chararray, age:int, gpa:float);

DUMP A;
(John,18,4.0F)
(Mary,19,3.7F)
(Bill,20,3.9F)
(Joe,22,3.8F)
(Jill,20,4.0F)

B = FILTER A BY name matches 'J.+';

DUMP B;
(John,18,4.0F)
(Joe,22,3.8F)
(Jill,20,4.0F)

A terminal window showing the execution of a Pig script. The output includes various log messages from the Hadoop ecosystem, such as configuration deprecation warnings, Pig script state updates, and job execution progress. The script being executed is: A = LOAD 'student' AS (name:chararray, age:int, gpa:float); B = foreach A generate \$0 as id; store B into '/home/hadoop/id.out';. The output shows the script being loaded, the data being processed, and the results being stored in the specified location. The terminal also shows the execution of the 'checksum' command on the input and output files.

```
Terminal
- checksum
2023-10-20 10:25:18,897 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> A = LOAD 'student' AS (name:chararray, age:int, gpa:float);
2023-10-20 10:27:38,247 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per
- checksum
2023-10-20 10:27:38,248 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = FILTER A BY name matches 'J.+';
grunt> DUMP B;
2023-10-20 10:28:39,053 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2023-10-20 10:28:39,060 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-10-20 10:28:39,060 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per
- checksum
2023-10-20 10:28:39,060 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schemaTuple] was not set... will not generate code.
2023-10-20 10:28:39,060 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - [RULES_ENABLED:AddForEach, ColumnMapKeyPrune, Con
stantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdo
wnOptimizer, PushDownForEachFilter, PushUpFilter, SplitFilter, StreamTypeCastInserter]]
2023-10-20 10:28:39,063 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MRCompiler - File concatenation threshold: 100 optimi
stic? false
2023-10-20 10:28:39,063 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MultiQueryOptimizer - MR plan size before optimizatio
n: 1
2023-10-20 10:28:39,063 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MultiQueryOptimizer - MR plan size after optimizatio
n: 1
2023-10-20 10:28:39,069 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-10-20 10:28:39,069 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per
- checksum
2023-10-20 10:28:39,070 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2023-10-20 10:28:39,071 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2023-10-20 10:28:39,071 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buff
er.percent is not set, set to default 0.5
2023-10-20 10:28:39,072 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.JobControlCompiler - Setting up single store job
2023-10-20 10:28:39,072 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schemaTuple] is false, will not generate code.
2023-10-20 10:28:39,072 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2023-10-20 10:28:39,072 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pl
g.schemaTuple.local.dir] with code temp directory: /tmp/1697777919072-0
2023-10-20 10:28:39,079 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapreduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for s
ubmission.
2023-10-20 10:28:39,080 [JobControl] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2023-10-20 10:28:39,084 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User classes may not be found. See J
ob or JobSetup(String)
2023-10-20 10:28:39,088 [JobControl] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2023-10-20 10:28:39,088 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-10-20 10:28:39,088 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2023-10-20 10:28:39,088 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2023-10-20 10:28:39,089 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2023-10-20 10:28:39,089 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: Job local304118899_0006
2023-10-20 10:28:39,100 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Executing with tokens: []
2023-10-20 10:28:39,134 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost:8080/
2023-10-20 10:28:39,134 [Thread-13] INFO org.apache.hadoop.mapreduce.local.JobRunner - OutputCommitter set in config null
2023-10-20 10:28:39,138 [Thread-13] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.byte
s-per-checksum
2023-10-20 10:28:39,138 [Thread-13] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.j
```

A = LOAD 'student' AS (name:chararray, age:int, gpa:float);

B = GROUP A BY name;

C = FOREACH B GENERATE COUNT(A.age);

EXPLAIN C;

FILTER