MUMBAI EDUCATIONAL TRUST

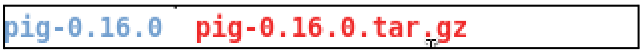
MET INSTITUTE OF COMPUTER SCIENCE

Program Number	PIG Part 1
Roll Number	1546
Title of program	Download and install pig and perform basic operations
Program	Download and install pig and perform basic operations

1. Extract Pig Tar File

- **command:** `tar -xvf pig-0.16.0.tar.gz`

- **purpose:** Extracts the Pig compressed archive into the current directory so files can be used.
- **output:** After successful extraction and running “ls”

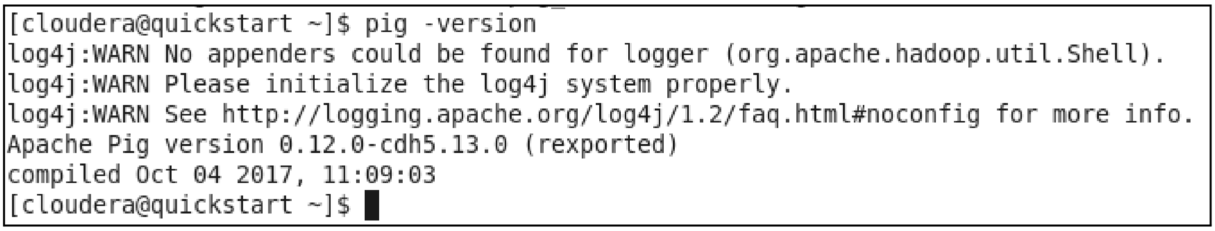


```
pig-0.16.0  pig-0.16.0.tar.gz
```

2. Check Pig Version

- **command:** `pig -version`

- **purpose:** This command verifies the installation by displaying the currently installed Pig version.
- **Output:**



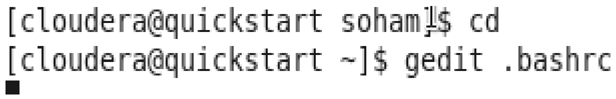
```
[cloudera@quickstart ~]$ pig -version
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Apache Pig version 0.12.0-cdh5.13.0 (rexported)
compiled Oct 04 2017, 11:09:03
[cloudera@quickstart ~]$ 
```

3. Open Bash Configuration File

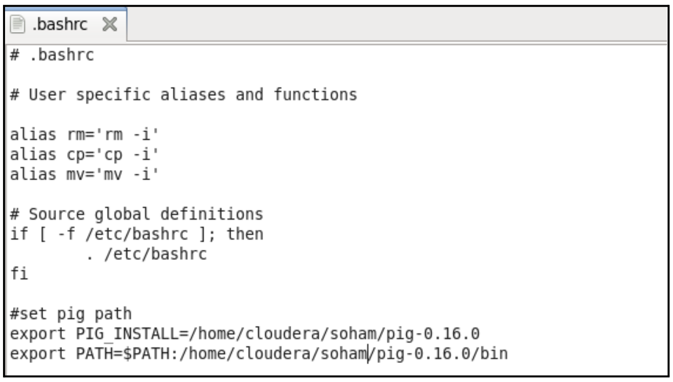
- **command:** `gedit .bashrc`

- **purpose:** This command opens the hidden `.bashrc` file in the text editor to configure environment variables. Set the pig path where the extracted file is located.

- **output:**



```
[cloudera@quickstart soham]$ cd
[cloudera@quickstart ~]$ gedit .bashrc
```



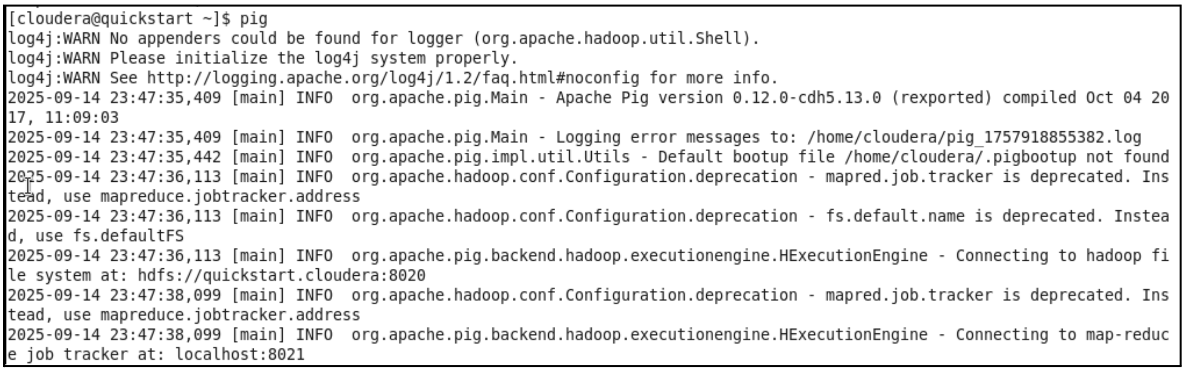
```
# .bashrc

# User specific aliases and functions

alias rm='rm -i'
alias cp='cp -i'
alias mv='mv -i'

# Source global definitions
if [ -f /etc/bashrc ]; then
        . /etc/bashrc
fi

#set pig path
export PIG_INSTALL=/home/cloudera/soham/pig-0.16.0
export PATH=$PATH:/home/cloudera/soham/pig-0.16.0/bin
```

4. Start Pig Grunt Shell

- **command:** `pig`

- **purpose:** This command launches the Pig interactive shell (Grunt) to execute Pig Latin commands. By default, it starts in local mode.

- **output:**



```
[cloudera@quickstart ~]$ pig
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2025-09-14 23:47:35,409 [main] INFO  org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.13.0 (rexported) compiled Oct 04 20
17, 11:09:03
2025-09-14 23:47:35,409 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/cloudera/pig_1757918855382.log
2025-09-14 23:47:35,442 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/cloudera/.pigbootup not found
2025-09-14 23:47:36,113 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
tead, use mapreduce.jobtracker.address
2025-09-14 23:47:36,113 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instea
d, use fs.defaultFS
2025-09-14 23:47:36,113 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop fi
le system at: hdfs://quickstart.cloudera:8020
2025-09-14 23:47:38,099 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Ins
tead, use mapreduce.jobtracker.address
2025-09-14 23:47:38,099 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduc
e job tracker at: localhost:8021
```

5. Upload emp.csv to HDFS

- **command:** `hdfs dfs -copyFromLocal emp.csv /user/cloudera/`

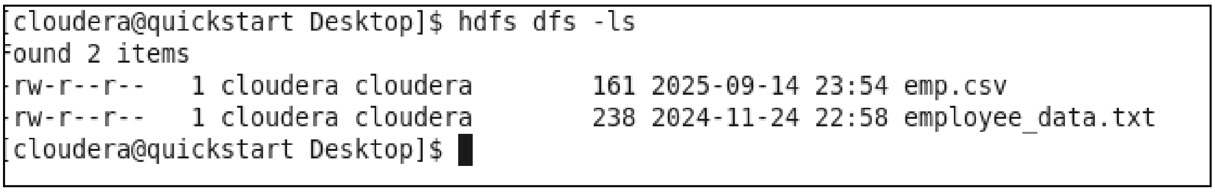
- **purpose:** This command copies the local file `emp.csv` from the desktop to the HDFS directory `/user/cloudera/` for further processing with Pig.
- **data in** `emp.csv`:
  101,ajay,2500,30
  102,vijay,3500,10
  103,sanjay,500,30
  104,Gavrav,1500,20
  105,Nita,300,20
  106,rita,3800,20
  107,reena,18000,10
  108,seeta,1800,20
  109,vijaya,3000,30

6. List Files in HDFS Directory

- **command:** `hdfs dfs -ls`

- **purpose:** This command lists all files and directories in the current HDFS path to verify that `emp.csv` has been uploaded successfully.

- **output:**



```
[cloudera@quickstart Desktop]$ hdfs dfs -ls
Found 2 items
-rw-r--r--   1 cloudera cloudera        161 2025-09-14 23:54 emp.csv
-rw-r--r--   1 cloudera cloudera        238 2024-11-24 22:58 employee_data.txt
[cloudera@quickstart Desktop]$ 
```

7. Load emp.csv into Pig Relation with Schema

- **command:** `empdata2 = load '/user/cloudera/emp.csv' using PigStorage(',') as (empid:int, ename:chararray, sal:int, dept:int);`

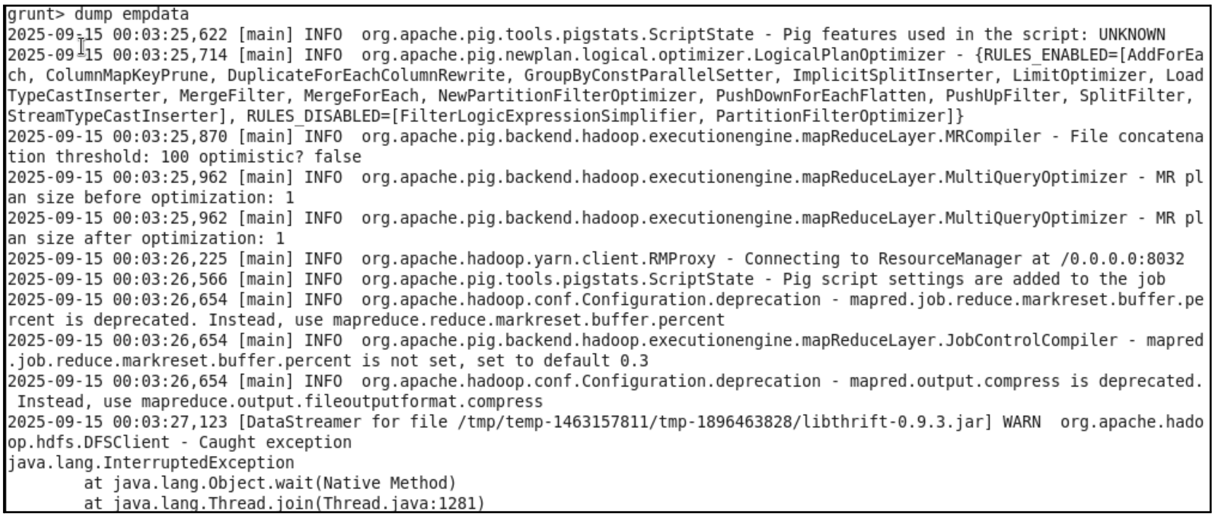
- **purpose:** This command loads the `emp.csv` file from HDFS into a Pig relation named `empdata2`, using comma as the delimiter and explicitly defining the schema (empid, ename, salary, dept).

8. Display empdata Relation

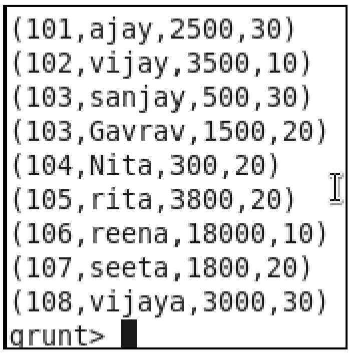
- **command:** `dump empdata2`

- **purpose:** This command displays all the tuples from the `empdata` relation to verify that the CSV data has been successfully loaded into Pig.

- **output:**



```
grunt> dump empdata
2025-09-15 00:03:25,622 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2025-09-15 00:03:25,714 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEa
ch, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, Load
TypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter,
StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2025-09-15 00:03:25,870 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatena
tion threshold: 100 optimistic? false
2025-09-15 00:03:25,962 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size before optimization: 1
2025-09-15 00:03:25,962 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR pl
an size after optimization: 1
2025-09-15 00:03:26,225 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2025-09-15 00:03:26,566 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2025-09-15 00:03:26,654 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.pe
rcent is deprecated. Instead, use mapreduce.reduce.markreset.buffer.percent
2025-09-15 00:03:26,654 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred
.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2025-09-15 00:03:26,654 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated.
 Instead, use mapreduce.output.fileoutputformat.compress
2025-09-15 00:03:27,123 [DataStreamer for file /tmp/temp-1463157811/tmp-1896463828/libthrift-0.9.3.jar] WARN  org.apache.hado
op.hdfs.DFSClient - Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
```



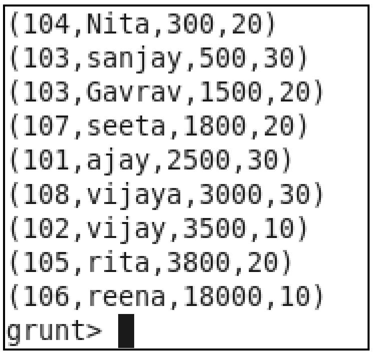
```
(101,ajay,2500,30)
(102,vijay,3500,10)
(103,sanjay,500,30)
(103,Gavrav,1500,20)
(104,Nita,300,20)
(105,rita,3800,20)
(106,reena,18000,10)
(107,seeta,1800,20)
(108,vijaya,3000,30)
grunt>
```

9. Order Employees by Salary

- **command:** orderbysal = ORDER empdata2 BY sal; dump orderbysal;

- **purpose:** These commands first create a new relation `orderbysal` by sorting the employee records in ascending order of salary, and then display the sorted results.

- **output:**



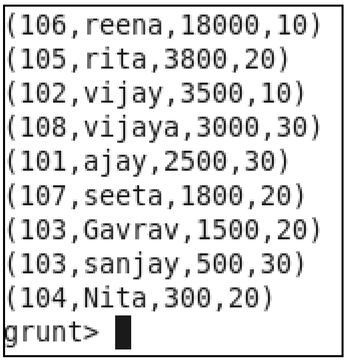
```
(104,Nita,300,20)
(103,sanjay,500,30)
(103,Gavrav,1500,20)
(107,seeta,1800,20)
(101,ajay,2500,30)
(108,vijaya,3000,30)
(102,vijay,3500,10)
(105,rita,3800,20)
(106,reena,18000,10)
grunt>
```

10. Order Employees by Salary (Descending)

- **command:** orderbysaldesc = ORDER empdata2 BY sal DESC; dump orderbysaldesc;

- **purpose:** These commands sort the employee records in descending order of salary and then display the highest-paid employees first.

- **output:**



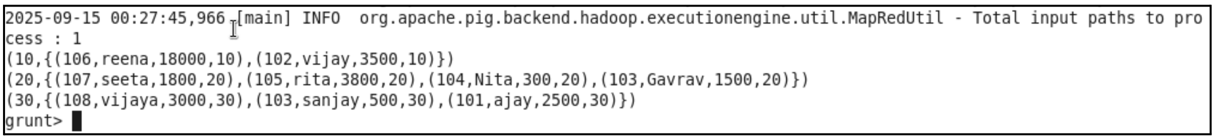
```
(106,reena,18000,10)
(105,rita,3800,20)
(102,vijay,3500,10)
(108,vijaya,3000,30)
(101,ajay,2500,30)
(107,seeta,1800,20)
(103,Gavrav,1500,20)
(103,sanjay,500,30)
(104,Nita,300,20)
grunt>
```

11. Group Employees by Department

- **command:** grpbydept = GROUP empdata2 BY dept; dump grpbydept;

- **purpose:** These commands group employee records by the department number and then display each group with its associated records.

- **output:**



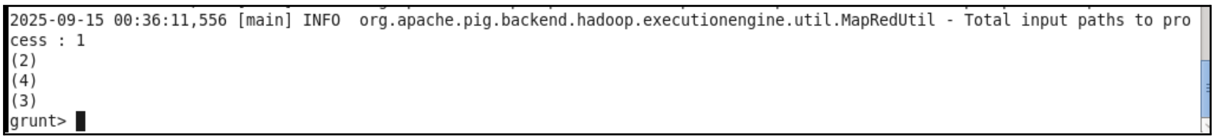
```
2025-09-15 00:27:45,966 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(10,{(106,reena,18000,10),(102,vijay,3500,10)})
(20,{(107,seeta,1800,20),(105,rita,3800,20),(104,Nita,300,20),(103,Gavrav,1500,20)})
(30,{(108,vijaya,3000,30),(103,sanjay,500,30),(101,ajay,2500,30)})
grunt>
```

12. Count Employees in Each Department

- **command:** cntEmp = FOREACH grpbydept GENERATE COUNT(empdata2); dump cntEmp;

- **purpose:** These commands calculate the total number of employees in each department group and display the counts.

- **output:**



```
2025-09-15 00:36:11,556 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pro
cess : 1
(2)
(4)
(3)
grunt> 
```