Experiment No:5

Title : PIG

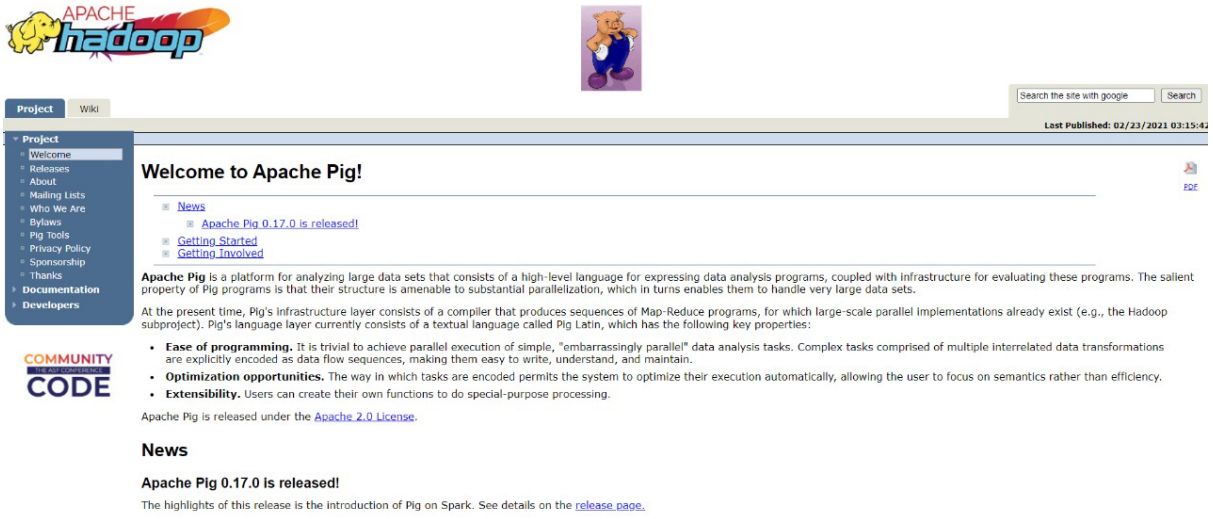
Aim: Working with operators in Pig - FOREACH, ASSERT, FILTER, GROUP, ORDERBY, DISTINCT, JOIN, LIMIT, SAMPLE, SPLIT, FLATTEN.

Download Apache Pig

First of all, download the latest version of Apache Pig from the following website – https://pig.apache.org/

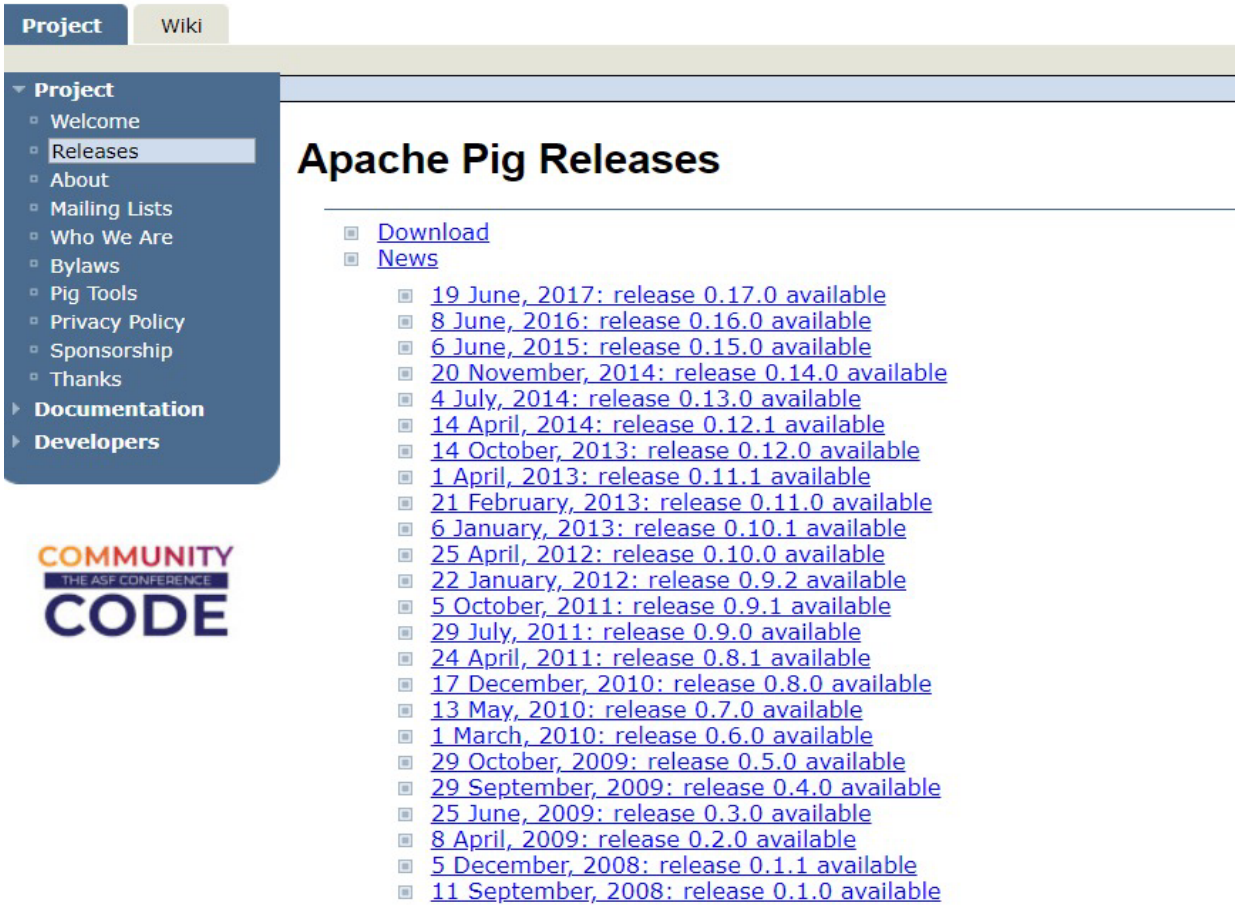
Step 1

Open the homepage of Apache Pig website. Under the section News, click on the link release page as shown in the following snapshot.



Step 2

On clicking the specified link, you will be redirected to the Apache Pig Releases page. On this page, under the Download section, you will have two links, namely, Pig 0.8 and later and Pig 0.7 and before. Click on the link Pig 0.8 and later, then you will be redirected to the page having a set of mirrors.



Step 3

These mirrors will take you to the Pig Releases page. This page contains various versions of Apache Pig. Click the latest version among them.

Pig Releases

Please make sure you're downloading from a nearby mirror site, not from www.apache.org.

Older releases are available from the archives.

Name	Last modified	Size	Description
Parent Directory		-	
latest/	2022-06-17 12:56	-	
pig-0.16.0/	2022-06-17 12:56	-	
pig-0.17.0/	2022-06-17 12:56	-	
KEYS	2017-06-19 08:12	11K	

Step 4

Within these folders, you will have the source and binary files of Apache Pig in various distributions. Download the tar files of the source and binary files of Apache Pig 0.15, pig0.15.0-src.tar.gz and pig-0.15.0.tar.gz.

Within these folders, you will have the source and binary files of Apache Pig in various distributions. Download the tar files of the source and binary files of Apache Pig 0.15, pig0.15.0-src.tar.gz and pig-0.15.0.tar.gz.

Install Apache Pig

After downloading the Apache Pig software, install it in your Linux environment by following the steps given below.

Step 1

Create a directory with the name Pig in the same directory where the installation directories of Hadoop, Java, and other software were installed. (In our tutorial, we have created the Pig directory in the user named Hadoop).

$ mkdir Pig

Step 2

Extract the downloaded tar files as shown below.

$ cd Downloads/

$ tar zxvf pig-0.15.0-src.tar.gz

$ tar zxvf pig-0.15.0.tar.gz

Step 3

Move the content of pig-0.15.0-src.tar.gz file to the Pig directory created earlier as shown below.

$ mv pig-0.15.0-src.tar.gz/* /home/Hadoop/Pig/

Ezoic

Configure Apache Pig

After installing Apache Pig, we have to configure it. To configure, we need to edit two files – bashrc and pig.properties.

.bashrc file

In the. bashrc file, set the following variables –

PIG_HOME folder to the Apache Pig's installation folder,

PATH environment variable to the bin folder, and

PIG_CLASSPATH environment variable to the etc (configuration) folder of your Hadoop installations (the directory that contains the core-site.xml, hdfs-site.xml and mapred-site.xml files).

export PIG_HOME = /home/Hadoop/Pig

export PATH = $PATH:/home/Hadoop/pig/bin

export PIG_CLASSPATH = $HADOOP_HOME/conf

```
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 HADOOP_PREFIX="/home/codegyani/hadoop-2.7.1/"
export PATH=$PATH:$HADOOP_PREFIX/bin
export PATH=$PATH:$HADOOP_PREFIX/sbin
export HADOOP_COMMON_HOME=${HADOOP_PREFIX}
export HADOOP_MAPRED_HOME=${HADOOP_PREFIX}
export HADOOP_HDFS_HOME=${HADOOP_PREFIX}
export YARN_HOME=${HADOOP_PREFIX}
export PIG_HOME=/home/codegyani/pig-0.16.0
export PATH=$PATH:$PIG_HOME/bin

^G Get Help  ^O WriteOut  ^R Read File ^Y Prev Page ^K Cut Text  ^C Cur Pos
^X Exit      ^J Justify   ^W Where Is  ^V Next Page ^U UnCut Text^T To Spell
```

Operations of PIG:

1) LOAD

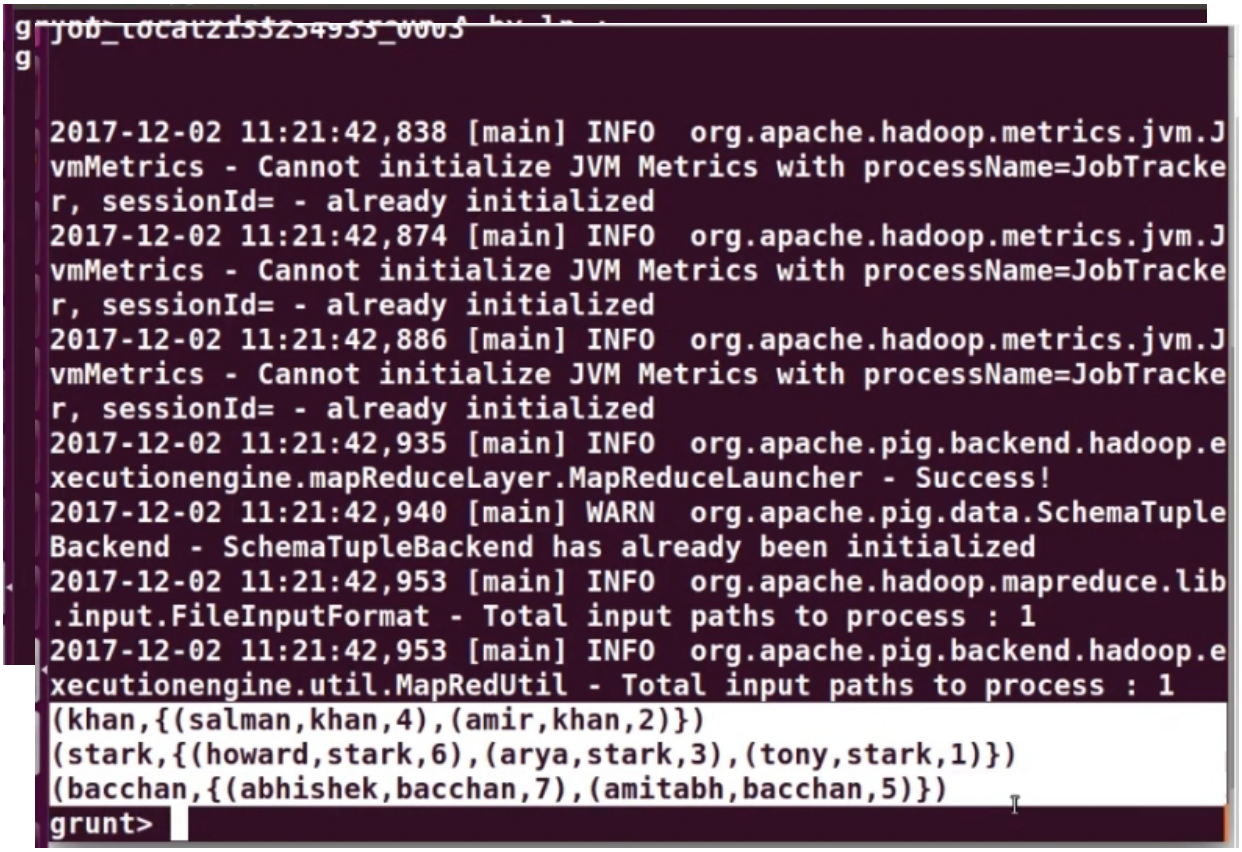
In Apache Pig, the LOAD operation is used to load data from a file or other storage into a relation for processing. This is typically the first step in a Pig script to read data into Pig's

```
grunt> fs -cat /pigdata/pig_dstab
tony    stark   1
amir    khan    2
arya    stark   3
salman  khan    4
amitabh bacchan 5
howard  stark   6
abhishek        bacchan 7
grunt> fs -cat /pigdata/pig_dscsv
tony,stark,1
amir,khan,2
arya,stark,3
salman,khan,4
amitabh,bacchan,5
howard,stark,6
abhishek,bacchan,7
grunt> B = load '/pigdata/pig_dscsv' using PigStorage(',') as (fn:ch
ararray,ln:chararray,idfld:int);
grunt> dump B
```

processing environment

2)GROUP

GROUP is used to group data in a relation by one or more fields.



3)FILTER

FILTER is used to filter out tuples based on a condition.

```
grunt> filterrowdata = filter A by ln != 'khan';
grunt> dump filterrowdata ;
```

```
2017-12-02 11:27:05,362 [main] INFO  org.apache.hadoop.metrics.jvm.J
vmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-12-02 11:27:05,375 [main] INFO  org.apache.hadoop.metrics.jvm.J
vmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-12-02 11:27:05,397 [main] INFO  org.apache.hadoop.metrics.jvm.J
vmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-12-02 11:27:05,404 [main] INFO  org.apache.pig.backend.hadoop.e
xecutionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-12-02 11:27:05,410 [main] WARN  org.apache.pig.data.SchemaTuple
Backend - SchemaTupleBackend has already been initialized
2017-12-02 11:27:05,468 [main] INFO  org.apache.hadoop.mapreduce.lib
.input.FileInputFormat - Total input paths to process : 1
2017-12-02 11:27:05,468 [main] INFO  org.apache.pig.backend.hadoop.e
xecutionengine.util.MapRedUtil - Total input paths to process : 1
(tony,stark,1)
(arya,stark,3)
(amitabh,bacchan,5)
(howard,stark,6)
(abhishek,bacchan,7)
grunt>
```

4)FOREACH

FOREACH is used to iterate over each element in a bag and generate transformations on columns.

```
grunt> filtercoldata = foreach A generate fn,ln ;
grunt> dump filtercoldata ;
```

```
r, sessionId= - already initialized
2017-12-02 11:29:12,165 [main] INFO  org.apache.hadoop.metrics.jvm.J
vmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-12-02 11:29:12,176 [main] INFO  org.apache.hadoop.metrics.jvm.J
vmMetrics - Cannot initialize JVM Metrics with processName=JobTracke
r, sessionId= - already initialized
2017-12-02 11:29:12,197 [main] INFO  org.apache.pig.backend.hadoop.e
xecutionengine.mapReduceLayer.MapReduceLauncher - Success!
2017-12-02 11:29:12,283 [main] WARN  org.apache.pig.data.SchemaTuple
Backend - SchemaTupleBackend has already been initialized
2017-12-02 11:29:12,329 [main] INFO  org.apache.hadoop.mapreduce.lib
.input.FileInputFormat - Total input paths to process : 1
2017-12-02 11:29:12,333 [main] INFO  org.apache.pig.backend.hadoop.e
xecutionengine.util.MapRedUtil - Total input paths to process : 1
(tony,stark)
(amir,khan)
(arya,stark)
(salman,khan)
(amitabh,bacchan)
(howard,stark)
(abhishek,bacchan)
grunt>
```

5)ORDER BY

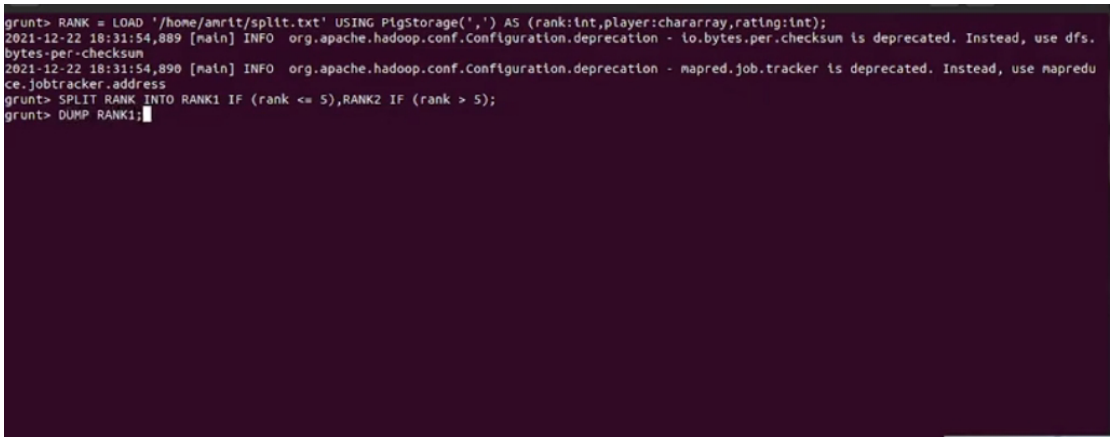
ORDER BY is used to sort the relation based on one or more fields.

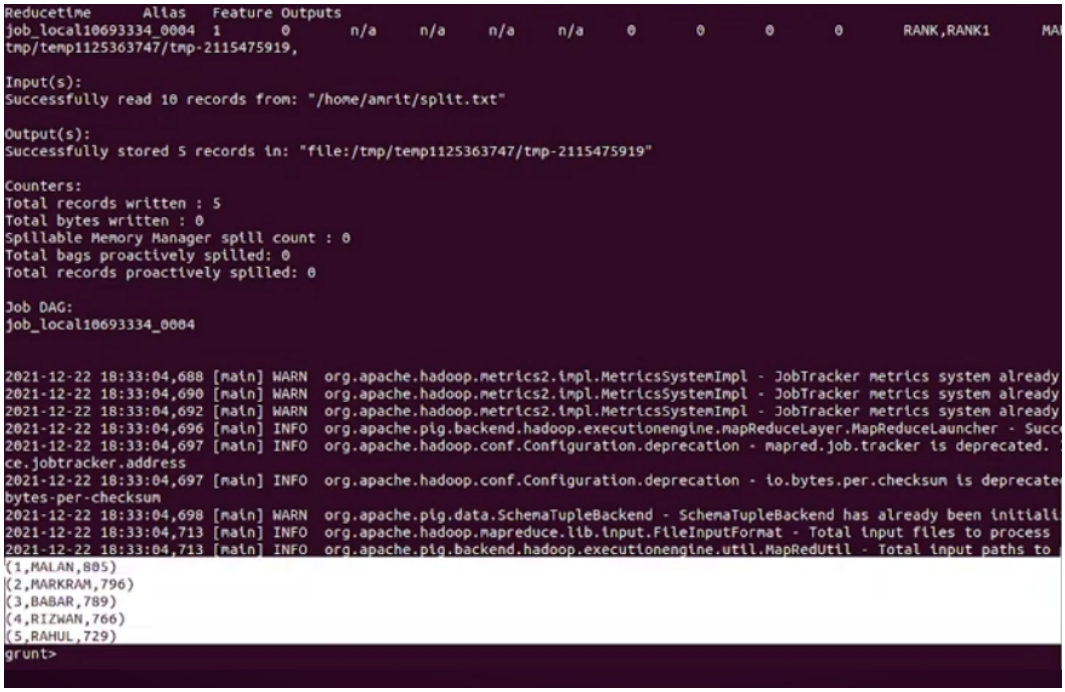
```
grunt> orderby_opr = ORDER gprec_data BY strength DESC;
grunt> DUMP orderby_opr;
```

```
        at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
2020-12-01 23:51:46,738 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate
- Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to j
ob history server
2020-12-01 23:51:47,432 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate
- Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to j
ob history server
2020-12-01 23:51:48,233 [main] INFO  org.apache.pig.backend.hadoop.executionengine.m
apReduceLayer.MapReduceLauncher - Success!
2020-12-01 23:51:48,234 [main] INFO  org.apache.hadoop.conf.Configuration.deprecatio
n - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-12-01 23:51:48,234 [main] INFO  org.apache.hadoop.conf.Configuration.deprecatio
n - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2020-12-01 23:51:48,236 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [p
ig.schematuple] was not set... will not generate code.
2020-12-01 23:51:48,252 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInput
Format - Total input paths to process : 1
2020-12-01 23:51:48,252 [main] INFO  org.apache.pig.backend.hadoop.executionengine.u
til.MapRedUtil - Total input paths to process : 1
(10, CSE,180)
(20, ECE,170)
(30, EEE,120)
grunt>
```

6)SPLIT

SPLIT is used to divide a relation into two or more relations based on a condition.





```
grunt> DUMP RANK2;
```

```
Reducetime      Alias   Feature Outputs
job_local673874123_0005 1       0       n/a     n/a     n/a     n/a     0       0       0       0       RANK,RANK2      MA
tmp/temp1125363747/tmp1088746893,

Input(s):
Successfully read 10 records from: "/home/amrit/split.txt"

Output(s):
Successfully stored 5 records in: "file:/tmp/temp1125363747/tmp1088746893"

Counters:
Total records written : 5
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local673874123_0005


2021-12-22 18:33:14,664 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already
2021-12-22 18:33:14,666 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already
2021-12-22 18:33:14,668 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already
2021-12-22 18:33:14,676 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succ
2021-12-22 18:33:14,677 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated.
ce.jobtracker.address
2021-12-22 18:33:14,679 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecate
bytes-per-checksum
2021-12-22 18:33:14,680 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initiali
2021-12-22 18:33:14,695 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process
2021-12-22 18:33:14,695 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to
(6,FINCH,709)
(7,CONWAY,703)
(8,BUTTLER,674)
(9,RVD,669)
(10,GUPTILL,658)
grunt>
```

7)LIMIT

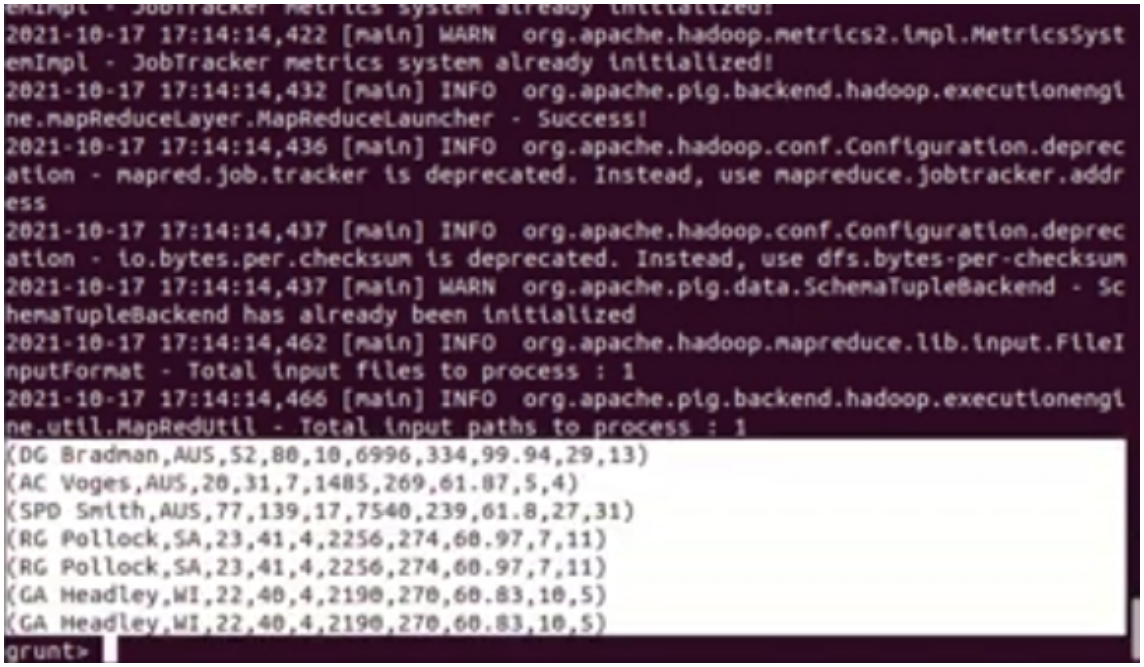
SPLIT is used to divide a relation into two or more relations based on a condition.

```
grunt> E = LIMIT RANK 4;
grunt> DUMP E;
```

```
grunt> E = LIMIT RANK 4;
grunt> DUMP E;
2021-12-22 18:34:33,668 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2021-12-22 18:34:33,692 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapredu
ce.jobtracker.address
2021-12-22 18:34:33,692 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.
bytes-per-checksum
2021-12-22 18:34:33,692 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-12-22 18:34:33,692 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyP
rune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, P
artitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2021-12-22 18:34:33,723 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 2
2021-12-22 18:34:33,723 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup _temporary
folders under output directory:false, ignore cleanup failures: false
2021-12-22 18:34:33,765 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-12-22 18:34:33,766 [main] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2021-12-22 18:34:33,766 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2021-12-22 18:34:33,766 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2021-12-22 18:34:33,772 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt__0001_m_000001
_1' to file:/tmp/temp1125363747/tmp1270611647
2021-12-22 18:34:33,782 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-12-22 18:34:33,795 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2021-12-22 18:34:33,795 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,MALAN,805)
(2,MARKRAM,796)
(3,BABAR,789)
(4,RIZWAN,766)
```

8)DISTINCT

Description: DISTINCT is used to remove duplicate records from a relation.


2021-10-17 17:14:14,422 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-10-17 17:14:14,432 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-10-17 17:14:14,436 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-10-17 17:14:14,437 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2021-10-17 17:14:14,437 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-10-17 17:14:14,462 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2021-10-17 17:14:14,466 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(DG Bradman,AUS,52,80,10,6996,334,99.94,29,13)
(AC Voges,AUS,20,31,7,1485,269,61.87,5,4)
(SPD Smith,AUS,77,139,17,7540,239,61.8,27,31)
(RG Pollock,SA,23,41,4,2256,274,60.97,7,11)
(RG Pollock,SA,23,41,4,2256,274,60.97,7,11)
(GA Headley,WI,22,40,4,2190,270,60.83,10,5)
(GA Headley,WI,22,40,4,2190,270,60.83,10,5)
grunt>


grunt> E = DISTINCT Q;
grunt> DUMP E;

```
Reducetime      Alias   Feature Outputs
job_local130806917_0003 1       1       n/a     n/a     n/a     n/a     n/a     n/a     n/a     n/a     Q       DISTINCT        file:/tmp/temp
-764148283/tmp-1553069399,

Input(s):
Successfully read 7 records from: "/home/amrit/test.txt"

Output(s):
Successfully stored 5 records in: "file:/tmp/temp-764148283/tmp-1553069399"

Counters:
Total records written : 5
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local130806917_0003


2021-10-17 17:16:12,598 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-10-17 17:16:12,605 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-10-17 17:16:12,607 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2021-10-17 17:16:12,619 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-10-17 17:16:12,620 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapredu
ce.jobtracker.address
2021-10-17 17:16:12,621 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.
bytes-per-checksum
2021-10-17 17:16:12,621 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-10-17 17:16:12,638 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2021-10-17 17:16:12,638 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(AC Voges,AUS,20,31,7,1485,269,61.87,5,4)
(SPD Smith,AUS,77,139,17,7540,239,61.8,27,31)
(DG Bradman,AUS,52,80,10,6996,334,99.94,29,13)
(GA Headley,WI,22,40,4,2190,270,60.83,10,5)
(RG Pollock,SA,23,41,4,2256,274,60.97,7,11)
```