Work Integrated
Learning Programmes

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Computer Science & Information Systems

# Big Data Systems – Lab Sheet

# PIG-Latin

## Objectives

Students should be able to

A. Gain understanding about PIG
B. Process data using various PIG operators

## Introduction to PIG

Pig is an open-source high level data flow system. It provides a simple language called Pig Latin, for queries and data manipulation, which are then compiled in to MapReduce jobs that run on Hadoop.

Pig was originally developed by Yahoo in 2006, for researchers to have an ad-hoc way of creating and executing MapReduce jobs on very large data sets. It was created to reduce the development time through its multi-query approach. Pig makes the job easier for professionals with non-programming background. It is easy write pig scripts if you are familiar with SQL.

Pig provides data operations like filters, joins, ordering for analyzing large datasets.

Pig can be run in two modes:

- **MapReduce mode** - In this mode, Pig loads and processes the data stored on HDFS. Pig Latin statements invoke a MapReduce job to perform the processing. It is the recommended mode in a production environment.
- **Local mode -** In this mode, Pig accesses files stored on the local file system. Data processing happens on the local machine. This mode is generally used for testing locally and speeding up development.

Pig scripts are run on the pig shell known as grunt shell. In order to access the grunt shell you can use the following commands

In order to launch Pig in local mode you can type the following command.

```
>pig –x local
```

In order to launch Pig in mapreduce mode you can type the following command.

```
>pig -x mapreduce
```

## Operators in PIG

In Pig Latin various operators are used for analyzing the data. In this section we will see the usage of some of the operators.

In order to run any pig script, first step is to access the grunt shell either in local mode or in mapreduce mode.

## LOAD

The LOAD statement is used to read the data from the file system. The data that requires to be manipulated is first loaded. The data file may be stored in HDFS or in local file system.

Consider an input file student_data.txt as follows

```
001,Rajiv,Hyderabad,M,8.10
002,siddarth,Kolkata,M,7.25
003,Rajesh,Delhi,M,8.52
004,Seema,Pune,F,9.25
005,Pooja,Mumbai,F,8.43
```

In order to load the data from the input file ‘student_data.txt’, you can use the following command

```
grunt>students = LOAD 'student_data.txt' USING PigStorage(',')
as (Id: int, Name: chararray, City: chararray,
gender:chararray, cgpa: double);
```

## DESCRIBE

The describe command is used to view the schema of a relation. In order to view the schema of students you can use describe as follows

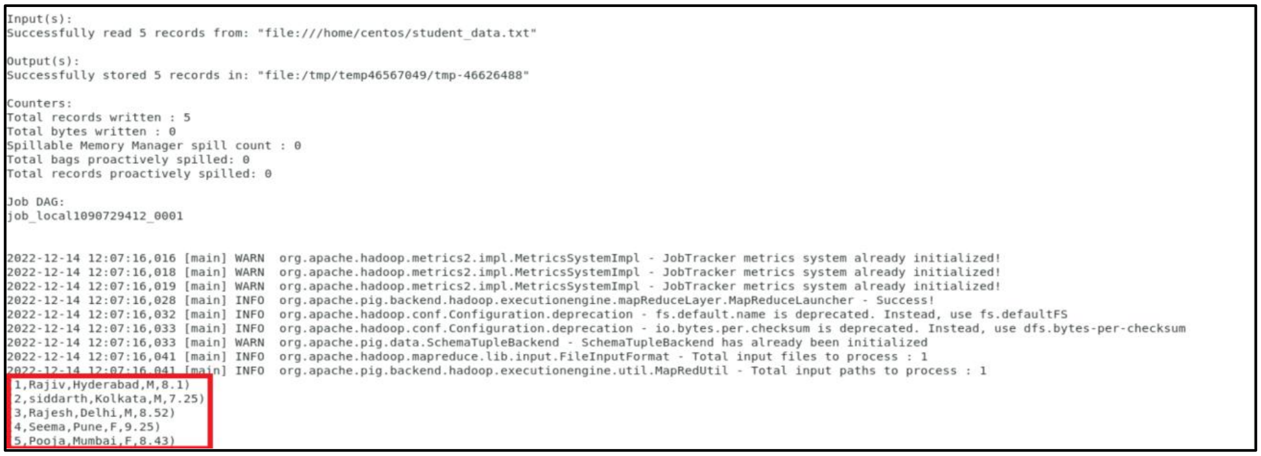
```
grunt>describe students;
output: students: {Id: int,Name: chararray,City:
chararray,gender: chararray,cgpa: double}
```

## DUMP

The dump command is used to display results. In order to display records of students schema, you can use dump as follows

```
grunt>dump students;
```

### output



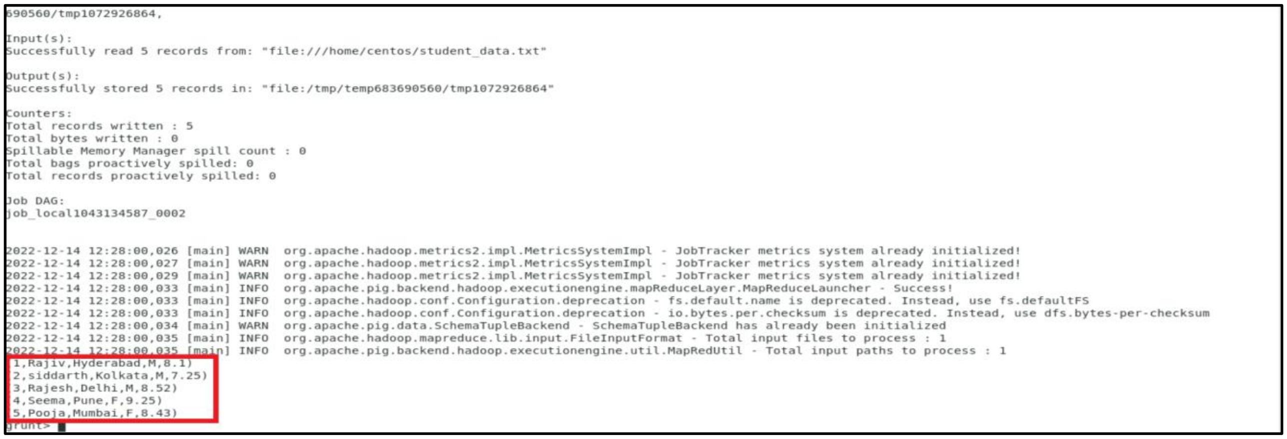
```
Input(s):
Successfully read 5 records from: "file:///home/centos/student_data.txt"

Output(s):
Successfully stored 5 records in: "file:/tmp/temp46567049/tmp-46626488"

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_local1090729412_0001


2022-12-14 12:07:16,016 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:07:16,018 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:07:16,019 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:07:16,028 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-14 12:07:16,032 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 12:07:16,033 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 12:07:16,033 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 12:07:16,041 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 12:07:16,041 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
1,Rajiv,Hyderabad,M,8.1)
2,siddarth,Kolkata,M,7.25)
3,Rajesh,Delhi,M,8.52)
4,Seema,Pune,F,9.25)
5,Pooja,Mumbai,F,8.43)
```

## FOREACH

The FOREACH operator performs iterations over every record and generate a new collection of records. You can use FOREACH on students as follows

```
grunt>studentDetails= FOREACH students generate *;
grunt> dump studentDetails;
```

Here * means to pick each attribute of students. The above statement will create studentDetails similar to students. Dump statement is used to display the results.

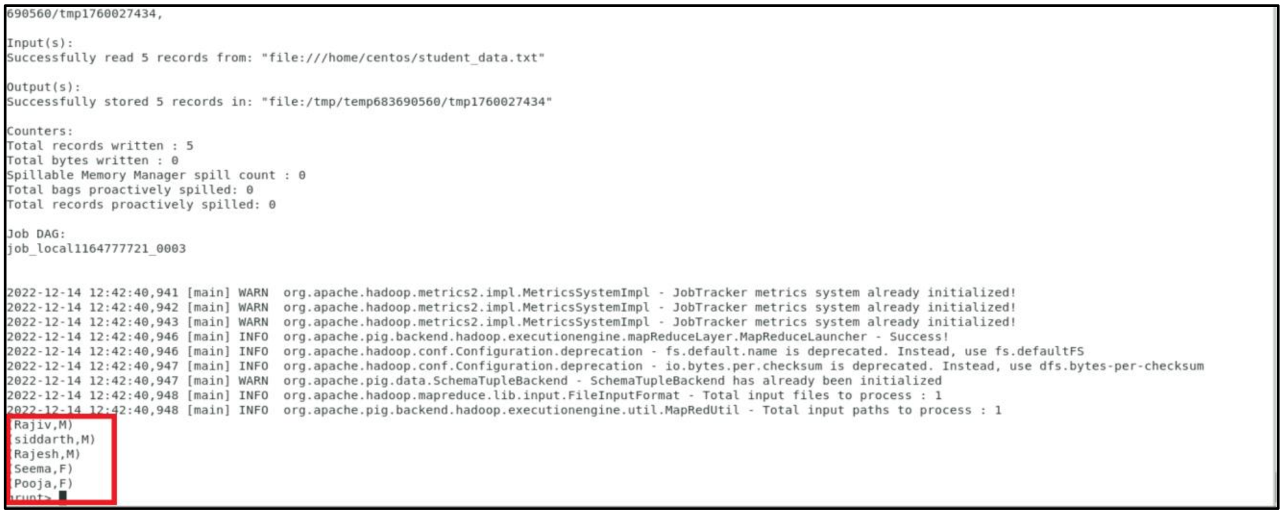


```
690560/tmp1072926864,

Input(s):
Successfully read 5 records from: "file:///home/centos/student_data.txt"

Output(s):
Successfully stored 5 records in: "file:/tmp/temp683690560/tmp1072926864"

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_local1043134587_0002


2022-12-14 12:28:00,026 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:28:00,027 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:28:00,029 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:28:00,033 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-14 12:28:00,033 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 12:28:00,033 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 12:28:00,034 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 12:28:00,035 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 12:28:00,035 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,Rajiv,Hyderabad,M,8.1)
(2,siddarth,Kolkata,M,7.25)
(3,Rajesh,Delhi,M,8.52)
(4,Seema,Pune,F,9.25)
(5,Pooja,Mumbai,F,8.43)
grunt> 
```

We can also select few attributes and generate another relation as follows

```
grunt>studentDetails= FOREACH students generate name, $3 *;
grunt> dump studentDetails;
```

The above statement will generate a studentDetails with 2 attributes name and gender. One way to specify the attributes is to specify the attribute name and another way to specify using $. The $3 in the above statement indicates 4th attribute which is gender. The dump can be used to display the results as follows.



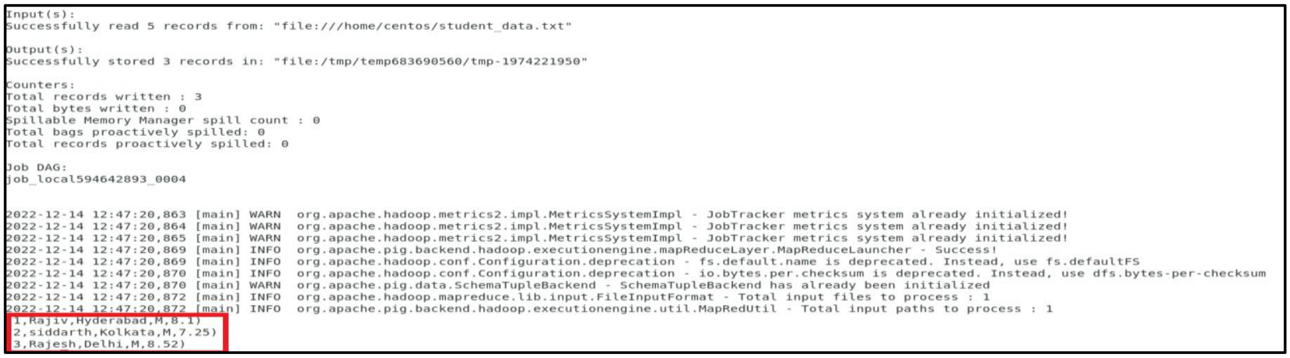
```
690560/tmp1760027434,

Input(s):
Successfully read 5 records from: "file:///home/centos/student_data.txt"

Output(s):
Successfully stored 5 records in: "file:/tmp/temp683690560/tmp1760027434"

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_local1164777721_0003


2022-12-14 12:42:40,941 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:42:40,942 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:42:40,943 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:42:40,946 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-14 12:42:40,946 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 12:42:40,947 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 12:42:40,947 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 12:42:40,948 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 12:42:40,948 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(Rajiv,M)
(siddarth,M)
(Rajesh,M)
(Seema,F)
(Pooja,F)
grunt> 
```

## FILTER

The FILTER operator enables you to filter the records based on a predicate. The records for which the predicate returns true are retained and others are discarded.

```
grunt>studentMales= FILTER students by gender == 'M';
grunt> dump studentMales;
```

The above statement will filter the male candidates from students. The dump is used to display the results.



```
Input(s):
Successfully read 5 records from: "file:///home/centos/student_data.txt"

Output(s):
Successfully stored 3 records in: "file:/tmp/temp683690560/tmp-1974221950"

Counters:
Total records written : 3
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local594642893_0004


2022-12-14 12:47:20,863 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:47:20,864 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:47:20,865 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:47:20,869 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-14 12:47:20,869 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 12:47:20,870 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 12:47:20,870 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 12:47:20,872 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 12:47:20,872 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
 1,Rajiv,Hyderabad,M,8.1)
 2,siddarth,Kolkata,M,7.25)
 3,Rajesh,Delhi,M,8.52)
```

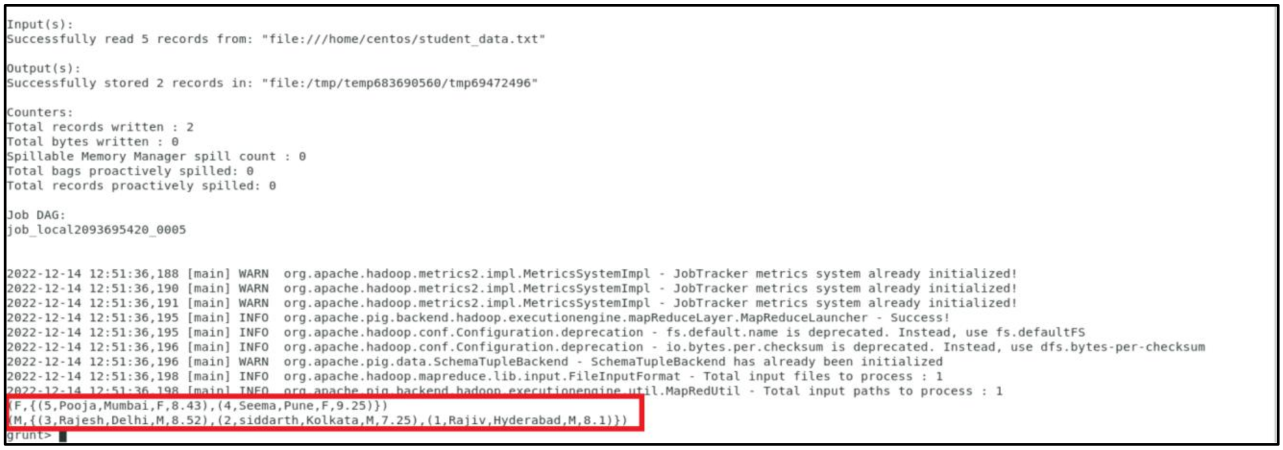
## GROUP

The GROUPoperator is used for grouping the data in relation.

In-order to group students by gender, you can use the following command

```
grunt>studentGender= Group students by gender;
grunt> dump studentGender;
```

The above statement will group the records based on gender. The dump is used to display the results



```
Input(s):
Successfully read 5 records from: "file:///home/centos/student_data.txt"

Output(s):
Successfully stored 2 records in: "file:/tmp/temp683690560/tmp69472496"

Counters:
Total records written : 2
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local2093695420_0005


2022-12-14 12:51:36,188 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:51:36,190 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:51:36,191 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 12:51:36,195 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-14 12:51:36,195 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 12:51:36,196 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 12:51:36,196 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 12:51:36,198 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 12:51:36,198 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(F,{(5,Pooja,Mumbai,F,8.43),(4,Seema,Pune,F,9.25)})
(M,{(3,Rajesh,Delhi,M,8.52),(2,siddarth,Kolkata,M,7.25),(1,Rajiv,Hyderabad,M,8.1)})
grunt> 
```

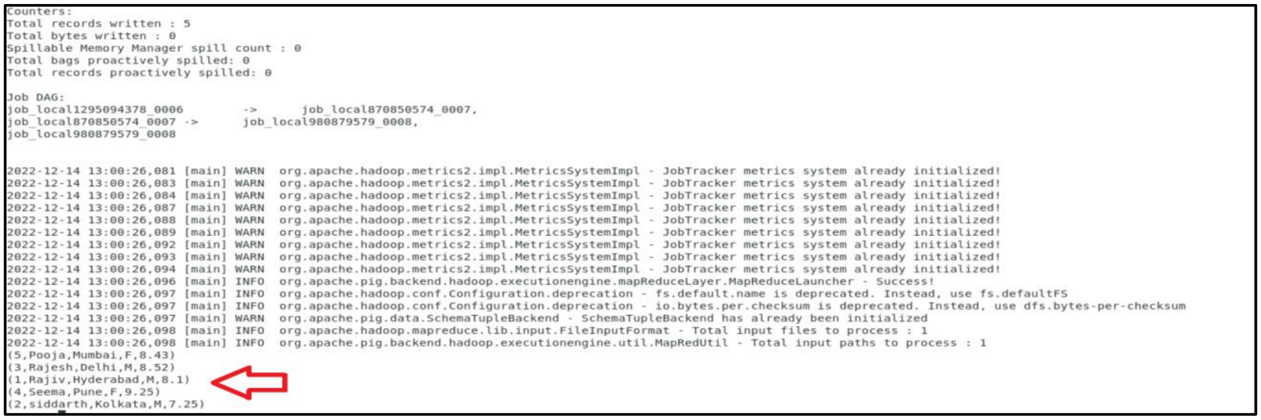
## ORDER BY

The ORDER BY operator is used for sorting depending on one or more fields.

In-order to sort the students by name in ascending order, you can use the following command

```
grunt>orderByName= ORDER students by name;
grunt>dump orderByName;
```

The above statement will sort the students records by name. the dump is used to display the results.

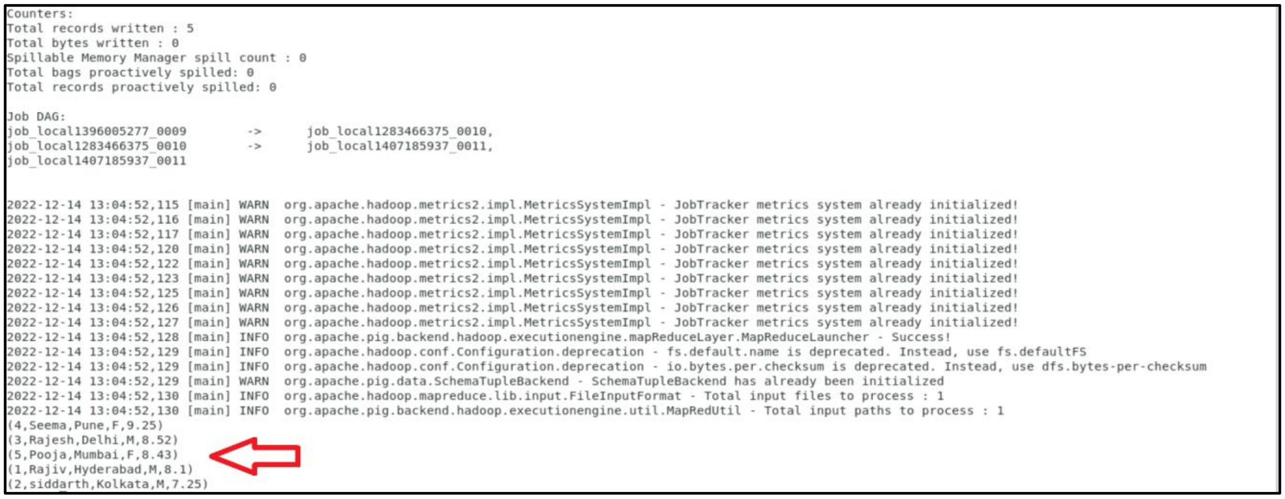


```
Counters:
Total records written : 5
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1295094378_0006        ->      job_local870850574_0007,
job_local870850574_0007 ->      job_local980879579_0008,
job_local980879579_0008


2022-12-14 13:00:26,081 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:00:26,083 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:00:26,084 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:00:26,087 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:00:26,088 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:00:26,089 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:00:26,092 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:00:26,093 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:00:26,094 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:00:26,096 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-14 13:00:26,097 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 13:00:26,097 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 13:00:26,097 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 13:00:26,098 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 13:00:26,098 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(5,Pooja,Mumbai,F,8.43)
(3,Rajesh,Delhi,M,8.52)
(1,Rajiv,Hyderabad,M,8.1)
(4,Seema,Pune,F,9.25)
(2,siddarth,Kolkata,M,7.25)
```

In-order to sort the in descending order, you can use the DESC.

```
grunt>orderByCgpa= ORDER students by cgpa DESC;
grunt> dump orderByCgpa;
```

The above statement will sort the students in descending order of CGPA. The dump is used display the results as follows.



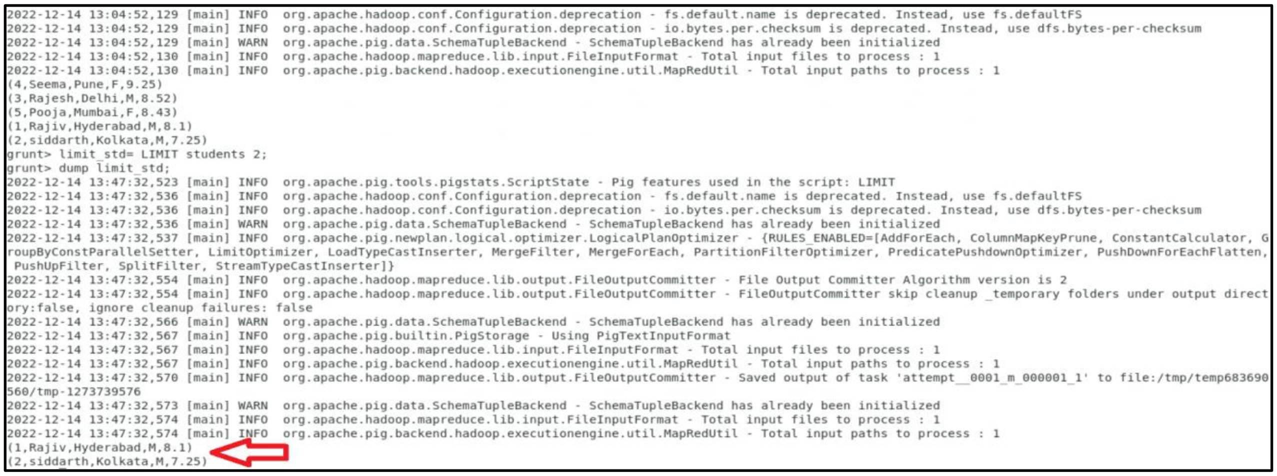
```
Counters:
Total records written : 5
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1396005277_0009        ->      job_local1283466375_0010,
job_local1283466375_0010        ->      job_local1407185937_0011,
job_local1407185937_0011


2022-12-14 13:04:52,115 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:04:52,116 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:04:52,117 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:04:52,120 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:04:52,122 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:04:52,123 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:04:52,125 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:04:52,126 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:04:52,127 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:04:52,128 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-14 13:04:52,129 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 13:04:52,129 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 13:04:52,129 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 13:04:52,130 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 13:04:52,130 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(4,Seema,Pune,F,9.25)
(3,Rajesh,Delhi,M,8.52)
(5,Pooja,Mumbai,F,8.43)
(1,Rajiv,Hyderabad,M,8.1)
(2,siddarth,Kolkata,M,7.25)
```

## LIMIT

The LIMIT operator allows a user to limit the number of records to be displayed as output

```
grunt>limit_std= LIMIT students 2;
grunt>dump limit_std;
```

The above statement will limit the number of records to 2 to be displayed. The dump is used to display the output as follows.



```
2022-12-14 13:04:52,129 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 13:04:52,129 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 13:04:52,129 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 13:04:52,130 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 13:04:52,130 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(4,Seema,Pune,F,9.25)
(3,Rajesh,Delhi,M,8.52)
(5,Pooja,Mumbai,F,8.43)
(1,Rajiv,Hyderabad,M,8.1)
(2,siddarth,Kolkata,M,7.25)
grunt> limit_std= LIMIT students 2;
grunt> dump limit_std;
2022-12-14 13:47:32,523 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2022-12-14 13:47:32,536 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 13:47:32,536 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 13:47:32,536 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 13:47:32,537 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, G
roupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten,
 PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2022-12-14 13:47:32,554 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 2
2022-12-14 13:47:32,554 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup _temporary folders under output direct
ory:false, ignore cleanup failures: false
2022-12-14 13:47:32,566 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 13:47:32,567 [main] INFO  org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2022-12-14 13:47:32,567 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 13:47:32,567 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2022-12-14 13:47:32,570 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt__0001_m_000001_1' to file:/tmp/temp683690
560/tmp-1273739576
2022-12-14 13:47:32,573 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 13:47:32,574 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 13:47:32,574 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,Rajiv,Hyderabad,M,8.1)
(2,siddarth,Kolkata,M,7.25)
```

## SPLIT

The SPLIT operator partitions a given relation into 2 or more relations

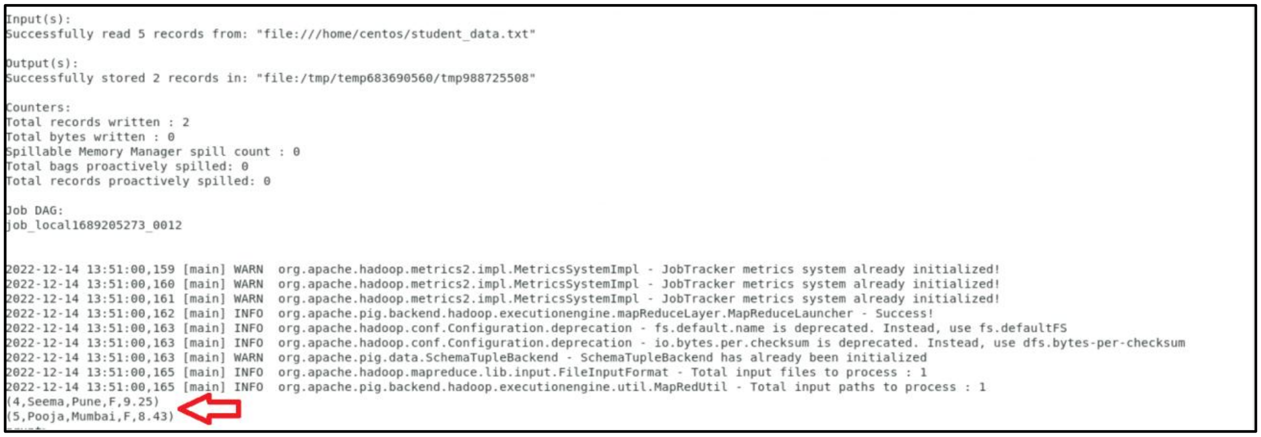
```
grunt>SPLIT students into femalestudents IF gender=='F',
malestudents IF gender=='M';
```

The above command will split the student relation into femalestudents and malestudents based on gender,

You can display new relations as follows.

```
grunt>dump femalestudents;
```

The output of female students is as follows

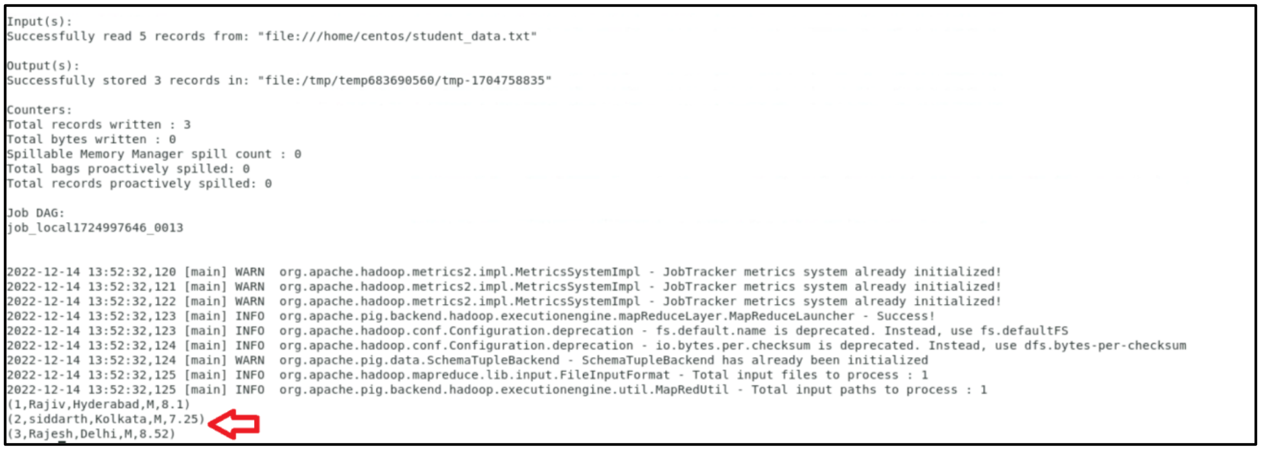


```
Input(s):
Successfully read 5 records from: "file:///home/centos/student_data.txt"

Output(s):
Successfully stored 2 records in: "file:/tmp/temp683690560/tmp988725508"

Counters:
Total records written : 2
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1689205273_0012


2022-12-14 13:51:00,159 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:51:00,160 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:51:00,161 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:51:00,162 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-14 13:51:00,163 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 13:51:00,163 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 13:51:00,163 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 13:51:00,165 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 13:51:00,165 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(4,Seema,Pune,F,9.25)
(5,Pooja,Mumbai,F,8.43)
```

You can display malestudents as follows.

```
grunt> dump malestudents;
```

The output of male students is as follows



```
Input(s):
Successfully read 5 records from: "file:///home/centos/student_data.txt"

Output(s):
Successfully stored 3 records in: "file:/tmp/temp683690560/tmp-1704758835"

Counters:
Total records written : 3
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1724997646_0013


2022-12-14 13:52:32,120 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:52:32,121 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:52:32,122 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 13:52:32,123 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-14 13:52:32,123 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 13:52:32,124 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 13:52:32,124 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 13:52:32,125 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 13:52:32,125 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,Rajiv,Hyderabad,M,8.1)
(2,siddarth,Kolkata,M,7.25)
(3,Rajesh,Delhi,M,8.52)
```

## COUNT

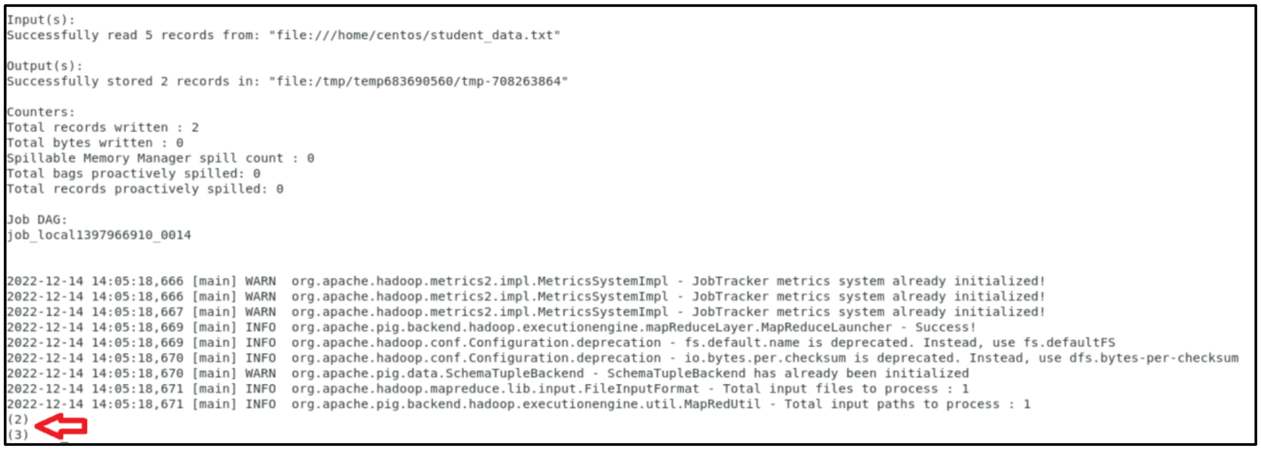
The COUNT operator counts the number of elements in a bag.

```
Grunt>studentGender= Group students by gender;
grunt>stdcount= FOREACH studentGender generate COUNT
(students.cgpa)
grunt>dump stdcount
```

In the first GROUP statement, we are creating a bags from students relation based on gender. It will create 2 bags one for males and another for females;

Further the count operation is applied to count the records in each bag. We have passed an attribute cgpa to count the records in each bag.

The dump can be used to display the results as follows.



```
Input(s):
Successfully read 5 records from: "file:///home/centos/student_data.txt"

Output(s):
Successfully stored 2 records in: "file:/tmp/temp683690560/tmp-708263864"

Counters:
Total records written : 2
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1397966910_0014


2022-12-14 14:05:18,666 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 14:05:18,666 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 14:05:18,667 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 14:05:18,669 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-14 14:05:18,669 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 14:05:18,670 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 14:05:18,670 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 14:05:18,671 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 14:05:18,671 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(2)
(3)
```

## AVERAGE (AVG)

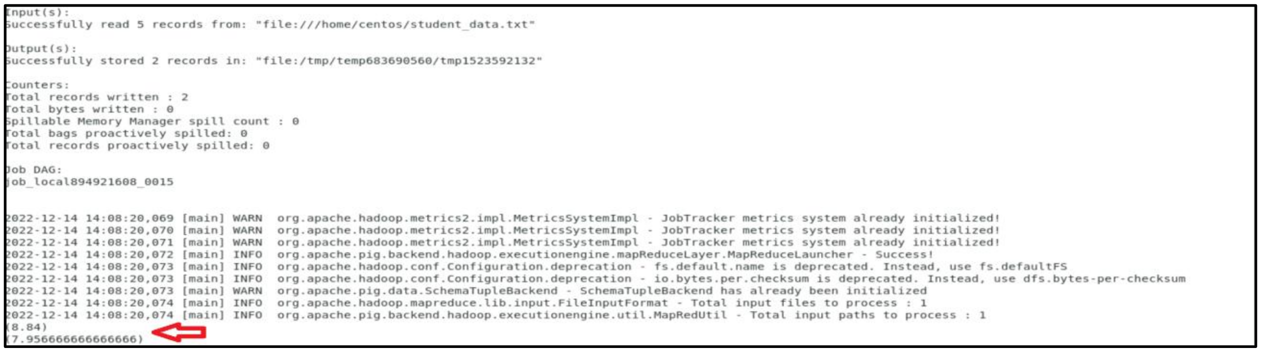
The  AVG operator calculates the averages of numeric values of in a bag.

```
Grunt>studentGender= Group students by gender;
grunt>avgcgpa= FOREACH studentGender generate AVG
(students.cgpa)
dumpavgcgpa;grunt> dump avgcgpa
```

In the first GROUP statement, we are creating a bags from students relation based on gender. It will create 2 bags one for males and another for females;

Further the AVG operation is specified on each bag which will calculate average cgpa of males and females.

The dump can be used to display the results as follows.



```
Input(s):
Successfully read 5 records from: "file:///home/centos/student_data.txt"

Output(s):
Successfully stored 2 records in: "file:/tmp/temp683690560/tmp1523592132"

Counters:
Total records written : 2
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local894921608_0015


2022-12-14 14:08:20,069 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 14:08:20,070 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 14:08:20,071 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 14:08:20,072 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-14 14:08:20,073 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 14:08:20,073 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 14:08:20,073 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 14:08:20,074 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 14:08:20,074 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(8.84)
(7.956666666666666)
```

## MAXIMUN (MAX)

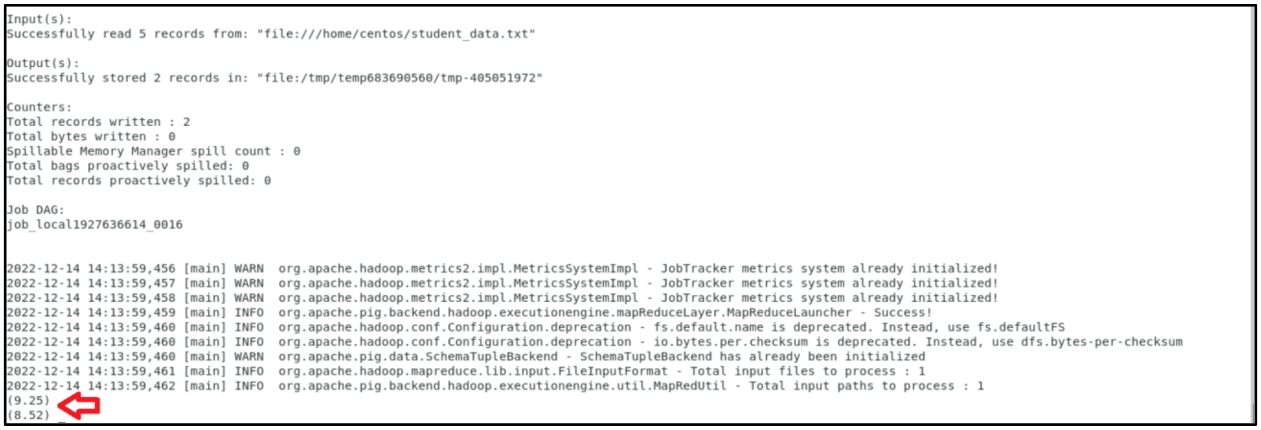
The MAX operator calculates maximum of numeric values in the column bags.

```
Grunt>studentGender= Group students by gender;
grunt>maxcgpa= FOREACH studentGender generate MAX
(students.cgpa)
dumpavgcgpa;grunt> dump maxcgpa
```

In the first GROUP statement, we are creating a bags from students relation based on gender. It will create 2 bags one for males and another for females;

Further the MAX operation is specified on each bag which will calculate maximum cgpa of males and females

The dump can be used to display the results as follows.



```
Input(s):
Successfully read 5 records from: "file:///home/centos/student_data.txt"

Output(s):
Successfully stored 2 records in: "file:/tmp/temp683690560/tmp-405051972"

Counters:
Total records written : 2
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1927636614_0016


2022-12-14 14:13:59,456 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 14:13:59,457 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 14:13:59,458 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 14:13:59,459 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-14 14:13:59,460 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 14:13:59,460 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 14:13:59,460 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 14:13:59,461 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 14:13:59,462 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(9.25)
(8.52) _
```

## MINIMUM (MIN)

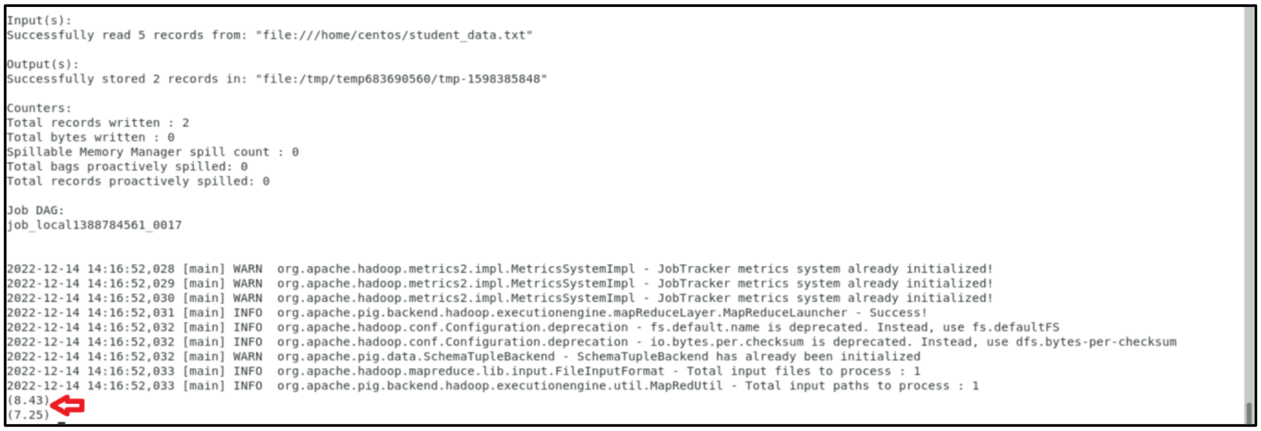
The  MIN operator calculates minimum of numeric values in the column bags.

```
Grunt>studentGender= Group students by gender;
grunt>maxcgpa= FOREACH studentGender generate MAX
(students.cgpa)
dumpavgcgpa;grunt> dump maxcgpa
```

In the first GROUP statement, we are creating a bags from students relation based on gender. It will create 2 bags one for males and another for females;

Further the MIN operation is specified on each bag which will calculate minimumcgpa of males and females

The dump can be used to display the results as follows.



```
Input(s):
Successfully read 5 records from: "file:///home/centos/student_data.txt"

Output(s):
Successfully stored 2 records in: "file:/tmp/temp683690560/tmp-1598385848"

Counters:
Total records written : 2
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1388784561_0017


2022-12-14 14:16:52,028 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 14:16:52,029 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 14:16:52,030 [main] WARN  org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialized!
2022-12-14 14:16:52,031 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-14 14:16:52,032 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-14 14:16:52,032 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-12-14 14:16:52,032 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-14 14:16:52,033 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-14 14:16:52,033 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(8.43)
(7.25)
```

## JOIN

The join operator in pig latin can be used to join two relations. The joining of two relations is possible in case they have common key. There are two types of join in pig latin.

- Inner join
- Outer join

## Inner join

This operator is used to return the matching rows from 2 relations on a common attribue

Consider the following input files customers.txt (id, name, age, address, salary) and orders.txt (oid, date, customer_id, amount)

Customers.txt

```
1,Ramesh,32,Ahmedabad,2000.00
2,Khilan,25,Delhi,1500.00
3,kaushik,23,Kota,2000.00
4,Chaitali,25,Mumbai,6500.00
5,Hardik,27,Bhopal,8500.00
6,Komal,22,MP,4500.00
7,Muffy,24,Indore,10000.00
```

Orders.txt

```
102,2009-10-08 00:00:00,3,3000
100,2009-10-08 00:00:00,3,1500
101,2009-11-20 00:00:00,2,1560
103,2008-05-20 00:00:00,4,2060
```

Start the history server using the following command on terminal (Required for pig mapreduce mode)

```
[centos@master~]$ cd $HADOOP_HOME/sbin
[centos@masterSbin]$./mr-jobhistory-daemon.sh start
```

Launch Pig in mapreduce mode.

```
>pig –x mapreduce
```

We can create the customers and orders relations using LOAD as follows

```
grunt> customers = LOAD '/user/pig/customers.txt' USING
PigStorage(',') as (id:int, name:chararray, age:int,
address:chararray, salary:int);
```

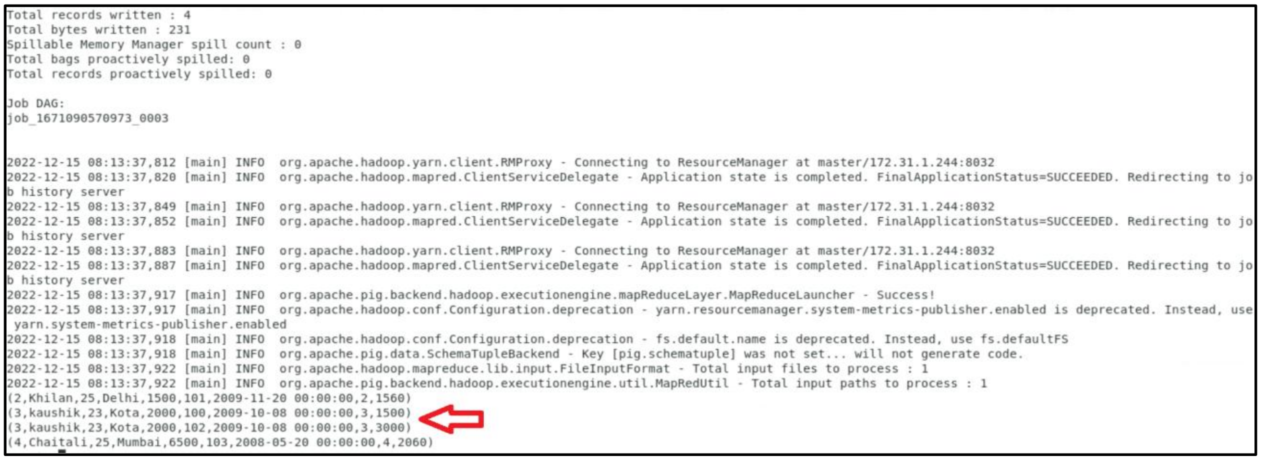
```
grunt> orders = LOAD '/user/pig/orders.txt' USING
PigStorage(',') as (oid:int, date:chararray, customer_id:int,
amount:int);
```

The above LOAD commands read the files stored in HDFS.

We can apply join on customers and orders as follows.

```
grunt>customer_orders = JOIN customers BY id, orders BY
customer_id;
grunt> dump customer_orders;
```

The dump can be used to display the results as follows.



```
Total records written : 4
Total bytes written : 231
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1671090570973_0003


2022-12-15 08:13:37,812 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:13:37,820 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:13:37,849 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:13:37,852 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:13:37,883 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:13:37,887 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:13:37,917 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-15 08:13:37,917 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use
 yarn.system-metrics-publisher.enabled
2022-12-15 08:13:37,918 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-15 08:13:37,918 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-12-15 08:13:37,922 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-15 08:13:37,922 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(2,Khilan,25,Delhi,1500,101,2009-11-20 00:00:00,2,1560)
(3,kaushik,23,Kota,2000,100,2009-10-08 00:00:00,3,1500)
(3,kaushik,23,Kota,2000,102,2009-10-08 00:00:00,3,3000)
(4,Chaitali,25,Mumbai,6500,103,2008-05-20 00:00:00,4,2060)
```

## Outer Join

**outer join** returns all the rows (even non matching) from at least one of the relations. An outer join operation is carried out in three ways −

- Left outer join
- Right outer join
- Full outer join

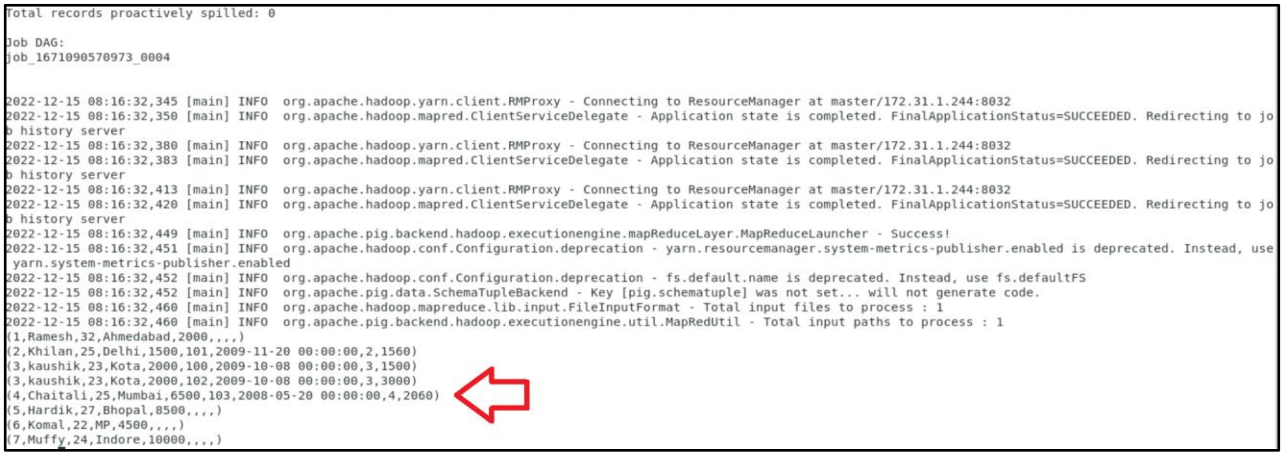
## Left outer join

The **left outer Join** operation returns all rows from the left table, even if there are no matches in the right relation. In non-matching rows the attribute values from other tables are filled in with null values.

We can apply join on customers and orders as follows.

```
grunt>outer_left = JOIN customers BY id LEFT OUTER, orders BY
customer_id;
grunt> dump outer_left;
```

The dump can be used to display the results as follows.



```
Total records proactively spilled: 0

Job DAG:
job_1671090570973_0004


2022-12-15 08:16:32,345 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:16:32,350 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:16:32,380 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:16:32,383 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:16:32,413 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:16:32,420 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:16:32,449 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-15 08:16:32,451 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use
 yarn.system-metrics-publisher.enabled
2022-12-15 08:16:32,452 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-15 08:16:32,452 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-12-15 08:16:32,460 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-15 08:16:32,460 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,Ramesh,32,Ahmedabad,2000,,,,)
(2,Khilan,25,Delhi,1500,101,2009-11-20 00:00:00,2,1560)
(3,kaushik,23,Kota,2000,100,2009-10-08 00:00:00,3,1500)
(3,kaushik,23,Kota,2000,102,2009-10-08 00:00:00,3,3000)
(4,Chaitali,25,Mumbai,6500,103,2008-05-20 00:00:00,4,2060)
(5,Hardik,27,Bhopal,8500,,,,)
(6,Komal,22,MP,4500,,,,)
(7,Muffy,24,Indore,10000,,,,)
```

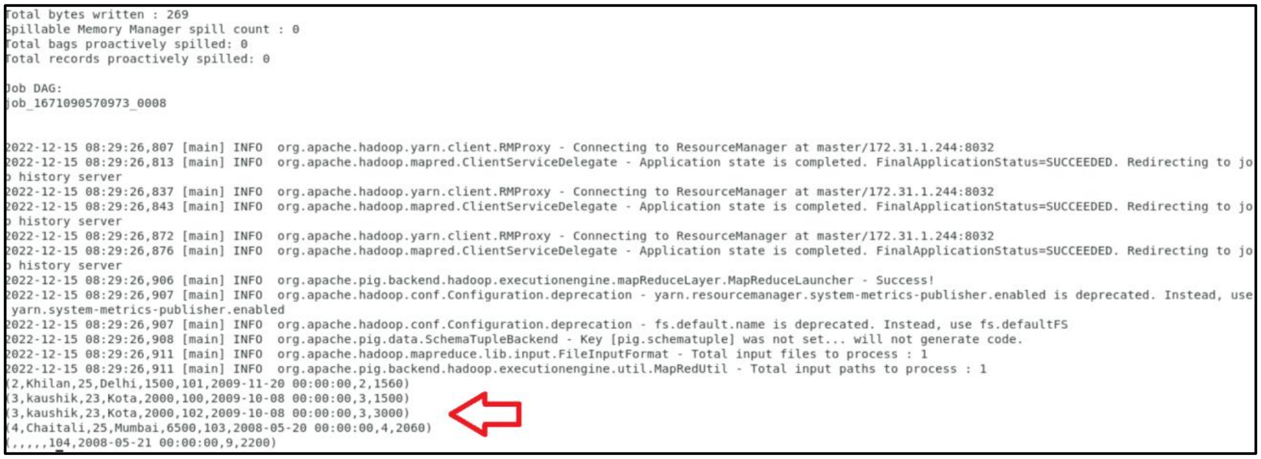
## Right outer join

The **right outer Join** operation returns all rows from the right table, even if there are no matches in the left relation. In non-matching rows the attribute values from other tables are filled in with null values.

We can apply join on customers and orders as follows.

```
grunt>outer_right = JOIN customers BY id RIGHT, orders BY
customer_id;
grunt> dump outer_right;
```

The dump can be used to display the results as follows.



```
Total bytes written : 269
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1671090570973_0008


2022-12-15 08:29:26,807 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:29:26,813 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:29:26,837 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:29:26,843 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:29:26,872 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:29:26,876 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:29:26,906 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-15 08:29:26,907 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use
 yarn.system-metrics-publisher.enabled
2022-12-15 08:29:26,907 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-15 08:29:26,908 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-12-15 08:29:26,911 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-15 08:29:26,911 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(2,Khilan,25,Delhi,1500,101,2009-11-20 00:00:00,2,1560)
(3,kaushik,23,Kota,2000,100,2009-10-08 00:00:00,3,1500)
(3,kaushik,23,Kota,2000,102,2009-10-08 00:00:00,3,3000)
(4,Chaitali,25,Mumbai,6500,103,2008-05-20 00:00:00,4,2060)
(,,,,,104,2008-05-21 00:00:00,9,2200)
```

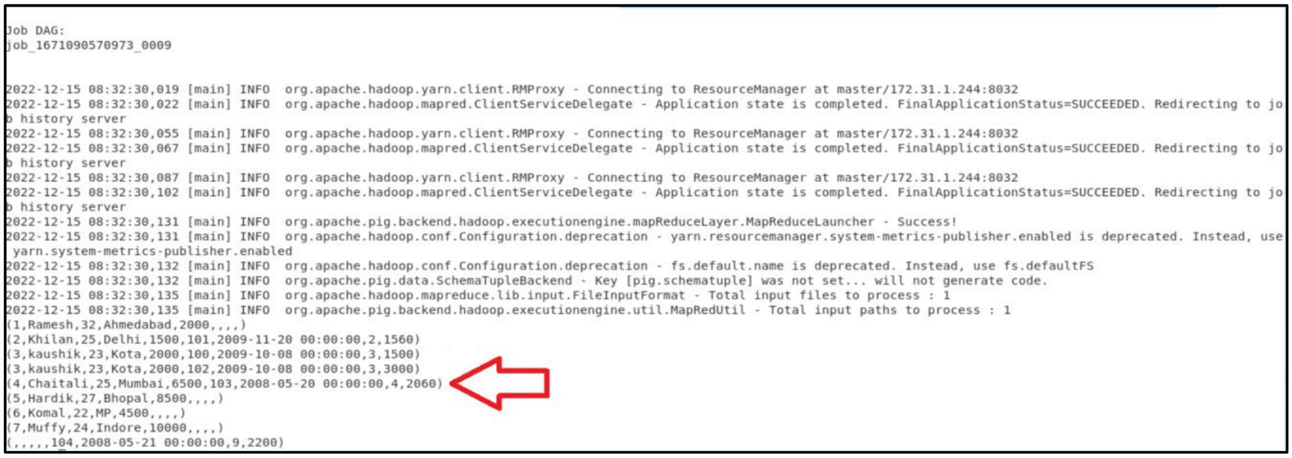
## Full outer join

The **full outer Join** operation returns matching and non-matching rows from both the tables. In case of non-matching rows the attributes from other table are filled in with null values.

We can apply join on customers and orders as follows.

```
grunt>outer_full = JOIN customers BY id FULL OUTER, orders BY
customer_id;
grunt> dump outer_full;
```

The dump can be used to display the results as follows.



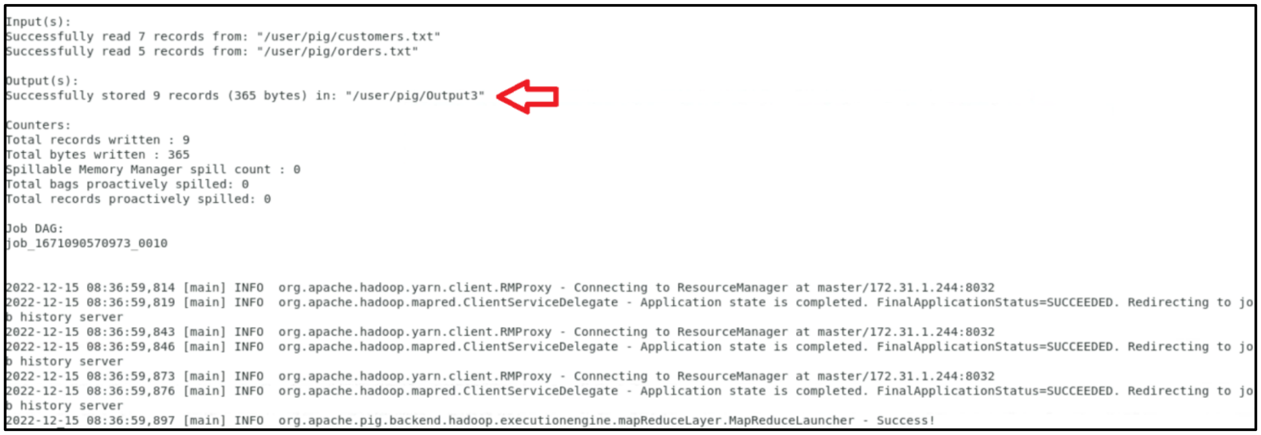
```
Job DAG:
job_1671090570973_0009


2022-12-15 08:32:30,019 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:32:30,022 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:32:30,055 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:32:30,067 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:32:30,087 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:32:30,102 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:32:30,131 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-15 08:32:30,131 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use
 yarn.system-metrics-publisher.enabled
2022-12-15 08:32:30,132 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-15 08:32:30,132 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-12-15 08:32:30,135 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-15 08:32:30,135 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,Ramesh,32,Ahmedabad,2000,,,,)
(2,Khilan,25,Delhi,1500,101,2009-11-20 00:00:00,2,1560)
(3,kaushik,23,Kota,2000,100,2009-10-08 00:00:00,3,1500)
(3,kaushik,23,Kota,2000,102,2009-10-08 00:00:00,3,3000)
(4,Chaitali,25,Mumbai,6500,103,2008-05-20 00:00:00,4,2060)
(5,Hardik,27,Bhopal,8500,,,,)
(6,Komal,22,MP,4500,,,,)
(7,Muffy,24,Indore,10000,,,,)
(,,,,,104,2008-05-21 00:00:00,9,2200)
```

## STORE

The STORE operator is used to store the output in a file. The following command will store the output in /user/pig/output directory in HDFS.

```
grunt>STORE outer_full INTO '/user/pig/Output/ ' USING
PigStorage (',');
```

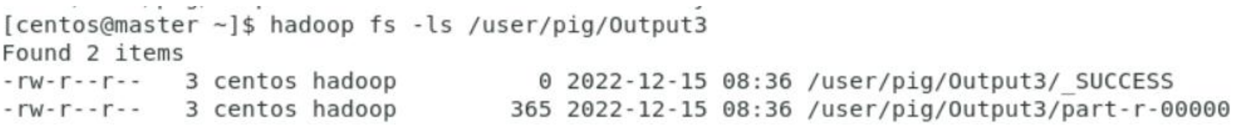
The output of STORE is as follows.



```
Input(s):
Successfully read 7 records from: "/user/pig/customers.txt"
Successfully read 5 records from: "/user/pig/orders.txt"

Output(s):
Successfully stored 9 records (365 bytes) in: "/user/pig/Output3"

Counters:
Total records written : 9
Total bytes written : 365
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1671090570973_0010


2022-12-15 08:36:59,814 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:36:59,819 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:36:59,843 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:36:59,846 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:36:59,873 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at master/172.31.1.244:8032
2022-12-15 08:36:59,876 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to jo
b history server
2022-12-15 08:36:59,897 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
```

You can list files in the output directory using following command

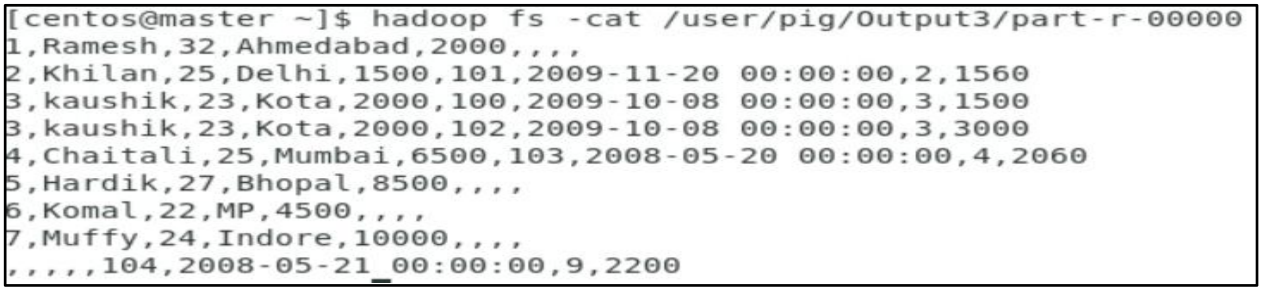
```
[centos@master~]$hadoop fs –ls /user/pig/output3
```



```
[centos@master ~]$ hadoop fs -ls /user/pig/Output3
Found 2 items
-rw-r--r--   3 centos hadoop          0 2022-12-15 08:36 /user/pig/Output3/_SUCCESS
-rw-r--r--   3 centos hadoop        365 2022-12-15 08:36 /user/pig/Output3/part-r-00000
```

You can display the contents of output as follows

```
[centos@master~]$hadoop fs –cat /user/pig/output3/part-r-00000
```



```
[centos@master ~]$ hadoop fs -cat /user/pig/Output3/part-r-00000
1,Ramesh,32,Ahmedabad,2000,,,,
2,Khilan,25,Delhi,1500,101,2009-11-20 00:00:00,2,1560
3,kaushik,23,Kota,2000,100,2009-10-08 00:00:00,3,1500
3,kaushik,23,Kota,2000,102,2009-10-08 00:00:00,3,3000
4,Chaitali,25,Mumbai,6500,103,2008-05-20 00:00:00,4,2060
5,Hardik,27,Bhopal,8500,,,,
6,Komal,22,MP,4500,,,,
7,Muffy,24,Indore,10000,,,,
,,,,,104,2008-05-21 00:00:00,9,2200
```

## Outputs/Results

- Students should be able to appreciate usage of pig Latin scripts
- Students should be able execute various PIG commands and verify the output

## Observations

Students should carefully observe the syntax of PIG commands and their usage

## REFERENCES

- Apache Pig Documentation
- Edureka
- Tutorials Point