SMIT R PATEL

19162121031

SEM 5

BDA

PRACTICAL 7

In general, Apache Pig works on top of Hadoop. It is an analytical tool that analyzes large datasets that exist in the **H**adoop **F**ile **S**ystem. To analyze data using Apache Pig, we have to initially load the data into Apache Pig. This chapter explains how to load data to Apache Pig from HDFS.

Preparing HDFS

In MapReduce mode, Pig reads (loads) data from HDFS and stores the results back in HDFS. Therefore, let us start HDFS and create the following sample data in HDFS.

Student ID	**First Name**	**Last Name**	**Phone**	**City**
001	Rajiv	Reddy	9848022337	Hyderabad
002	siddarth	Battacharya	9848022338	Kolkata
003	Rajesh	Khanna	9848022339	Delhi
004	Preethi	Agarwal	9848022330	Pune

005	Trupthi	Mohanthy	9848022336	Bhuwaneshwar
006	Archana	Mishra	9848022335	Chennai

The above dataset contains personal details like id, first name, last name, phone number and city, of six students.

The input file of Pig contains each tuple/record in individual lines. And the entities of the record are separated by a delimiter (In our example we used **","**).

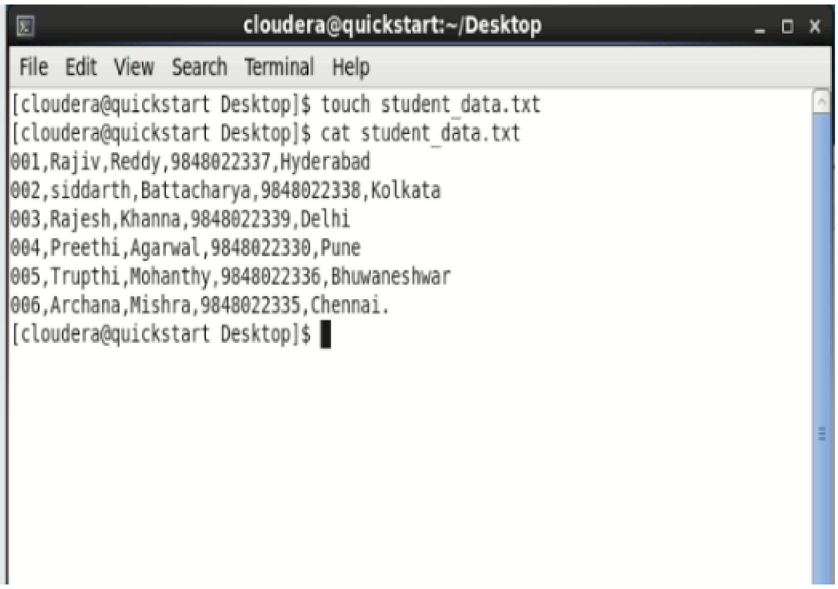
In the local file system, create an input file **student_data.txt** containing data as shown below.

```
001,Rajiv,Reddy,9848022337,Hyderabad
002,siddarth,Battacharya,9848022338,Kolkata
003,Rajesh,Khanna,9848022339,Delhi
004,Preethi,Agarwal,9848022330,Pune
005,Trupthi,Mohanthy,9848022336,Bhuwaneshwar
006,Archana,Mishra,9848022335,Chennai.
```

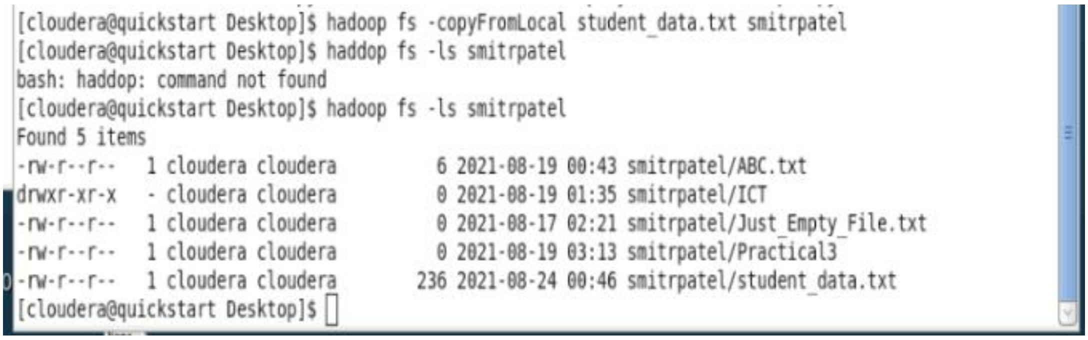
Now, move the file from the local file system to HDFS.

Verify whether the file has been moved into the HDFS.

You can load data into Apache Pig from the file system (HDFS/ Local) using **LOAD** operator of **Pig Latin**.


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
[cloudera@quickstart Desktop]$ touch student_data.txt
[cloudera@quickstart Desktop]$ cat student_data.txt
001,Rajiv,Reddy,9848022337,Hyderabad
002,siddarth,Battacharya,9848022338,Kolkata
003,Rajesh,Khanna,9848022339,Delhi
004,Preethi,Agarwal,9848022330,Pune
005,Trupthi,Mohanthy,9848022336,Bhuwaneshwar
006,Archana,Mishra,9848022335,Chennai.
[cloudera@quickstart Desktop]$
```



```
[cloudera@quickstart Desktop]$ hadoop fs -copyFromLocal student_data.txt smitrpatel
[cloudera@quickstart Desktop]$ haddop fs -ls smitrpatel
bash: haddop: command not found
[cloudera@quickstart Desktop]$ hadoop fs -ls smitrpatel
Found 5 items
-rw-r--r--   1 cloudera cloudera          6 2021-08-19 00:43 smitrpatel/ABC.txt
drwxr-xr-x   - cloudera cloudera          0 2021-08-19 01:35 smitrpatel/ICT
-rw-r--r--   1 cloudera cloudera          0 2021-08-17 02:21 smitrpatel/Just_Empty_File.txt
-rw-r--r--   1 cloudera cloudera          0 2021-08-19 03:13 smitrpatel/Practical3
-rw-r--r--   1 cloudera cloudera        236 2021-08-24 00:46 smitrpatel/student_data.txt
[cloudera@quickstart Desktop]$
```

Syntax

The load statement consists of two parts divided by the “=” operator. On the left-hand side, we need to mention the name of the relation **where** we want to store the data, and on the right-hand side, we have to define **how** we store the data. Given below is the syntax of the **Load** operator.

```
Relation_name = LOAD 'Input file path' USING
function as schema;
```

Where,

- **relation_name** − We have to mention the relation in which we want to store the data.
- **Input file path** − We have to mention the HDFS directory

where the file is stored. (In MapReduce mode)

- **function** – We have to choose a function from the set of load functions provided by Apache Pig (**BinStorage, JsonLoader, PigStorage, TextLoader**).
- **Schema** – We have to define the schema of the data. We can define the required schema as follows –

```
(column1 : data type, column2 : data type, column3 : data
type);
```

Note – We load the data without specifying the schema. In that case, the columns will be addressed as $01, $02, etc… (check).

Example

As an example, let us load the data in **student_data.txt** in Pig under the schema named **Student** using the **LOAD** command.

```
grunt> student = LOAD
'hdfs://localhost:9000/pig_data/student_data.txt'
 USING PigStorage(',')
 as ( id:int, firstname:chararray, lastname:chararray,
phone:chararray,
 city:chararray );
```

Relation name	We have stored the data in the schema **student**.
Input file path	We are reading data from the file **student_data.txt,** which is in the /pig_data/ directory of HDFS.
Storage function	We have used the **PigStorage()** function. It loads and stores data as structured text files. It takes a delimiter using which each entity of a tuple is separated, as a parameter. By default, it takes '\t' as a parameter.
schema	We have stored the data using the following schema. column id firstname lastname phone city datatype int char array char array char array char array

Following is the description of the above statement.

Note − The **load** statement will simply load the data into the specified relation in Pig.

Dump Operator

The **Dump** operator is used to run the Pig Latin statements and display the results on the screen. It is generally used for debugging Purpose.

Syntax

Given below is the syntax of the **Dump** operator.

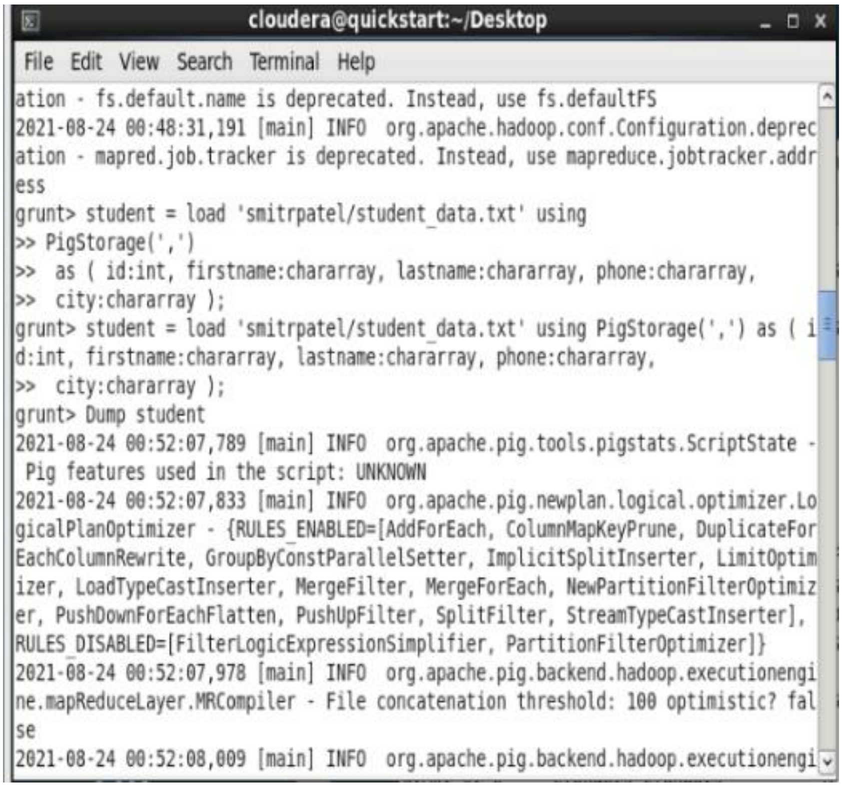
```
grunt> Dump Relation_Name
```

Now, let us print the contents of the relation using the **Dump operator** as shown below.

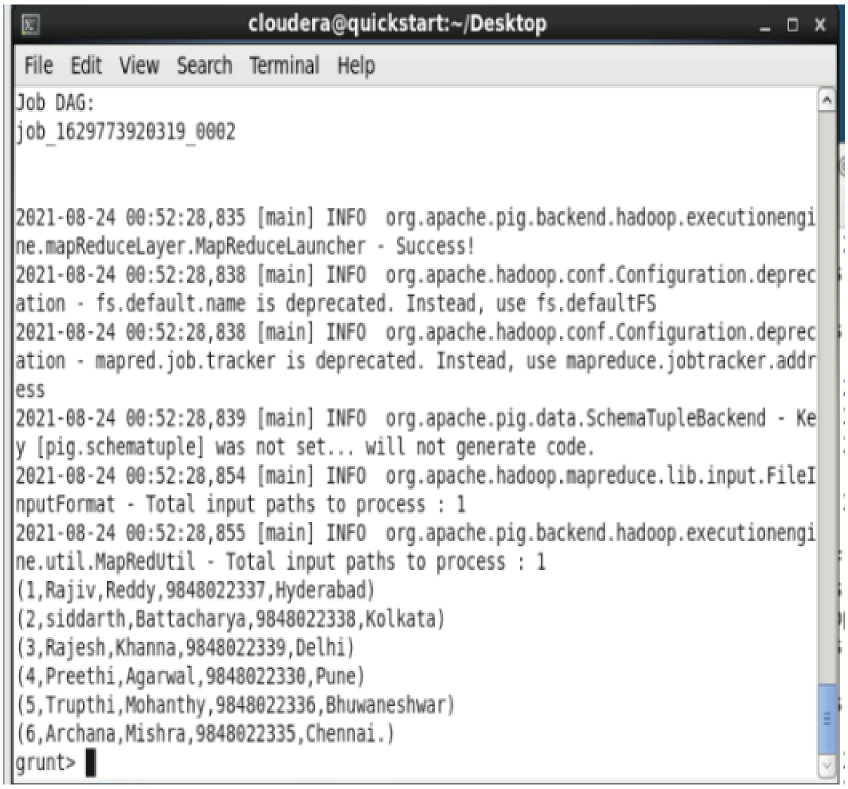
```
grunt> Dump student
```

Once you execute the above **Pig Latin** statement, it will start a MapReduce job to read data from HDFS.

```
grunt> student = load 'smitrpatel/student_data.txt' using
PigStorage(',') as ( id:int, firstname:chararray,
lastname:chararray, phone:chararray, city:chararray );
```


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-08-24 00:48:31,191 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
grunt> student = load 'smitrpatel/student_data.txt' using
>> PigStorage(',')
>>  as ( id:int, firstname:chararray, lastname:chararray, phone:chararray,
>>  city:chararray );
grunt> student = load 'smitrpatel/student_data.txt' using PigStorage(',') as ( i
d:int, firstname:chararray, lastname:chararray, phone:chararray,
>>  city:chararray );
grunt> Dump student
2021-08-24 00:52:07,789 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: UNKNOWN
2021-08-24 00:52:07,833 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-08-24 00:52:07,978 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2021-08-24 00:52:08,009 [main] INFO  org.apache.pig.backend.hadoop.executionengi
```



cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
Job DAG:
job_1629773920319_0002


2021-08-24 00:52:28,835 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2021-08-24 00:52:28,838 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-08-24 00:52:28,838 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-08-24 00:52:28,839 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2021-08-24 00:52:28,854 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-08-24 00:52:28,855 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(1,Rajiv,Reddy,9848022337,Hyderabad)
(2,siddarth,Battacharya,9848022338,Kolkata)
(3,Rajesh,Khanna,9848022339,Delhi)
(4,Preethi,Agarwal,9848022330,Pune)
(5,Trupthi,Mohanthy,9848022336,Bhuwaneshwar)
(6,Archana,Mishra,9848022335,Chennai.)
grunt> 
```

Describe Operator

The **describe** operator is used to view the schema of a relation.

Syntax

The syntax of the **describe** operator is as follows –

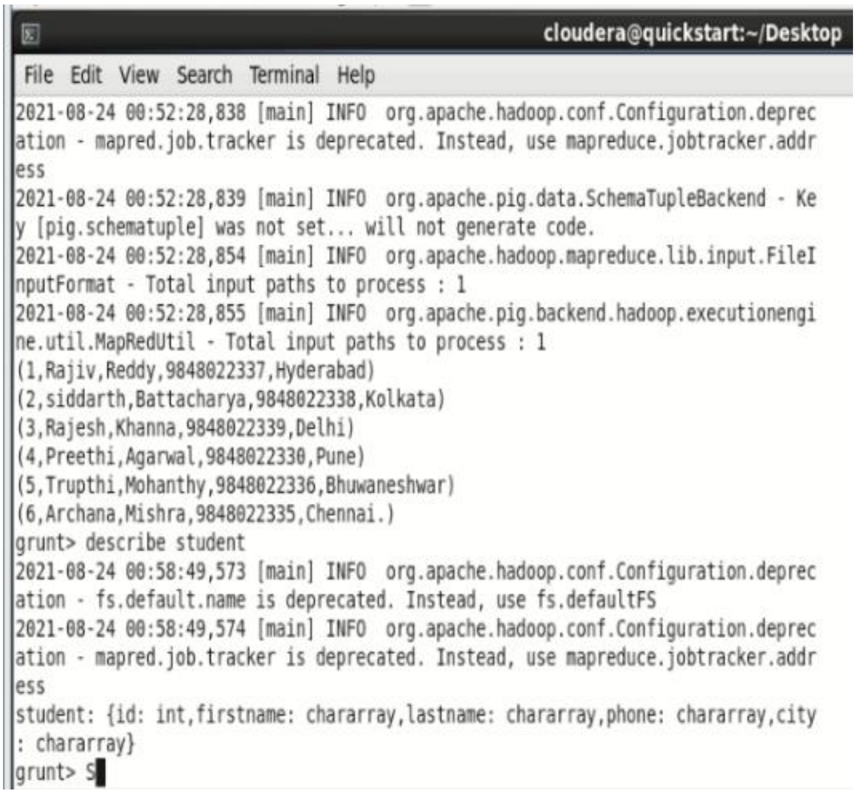
```
grunt> Describe Relation_name
```

let us describe the relation named **student** and verify the schema as shown below. `grunt> describe student;`

Output

Once you execute the above **Pig Latin** statement, it will produce the following output.

```
grunt> student: { id: int,firstname:
chararray,lastname: chararray,phone:
chararray,city: chararray }
```



cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
2021-08-24 00:52:28,838 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-08-24 00:52:28,839 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2021-08-24 00:52:28,854 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-08-24 00:52:28,855 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(1,Rajiv,Reddy,9848022337,Hyderabad)
(2,siddarth,Battacharya,9848022338,Kolkata)
(3,Rajesh,Khanna,9848022339,Delhi)
(4,Preethi,Agarwal,9848022330,Pune)
(5,Trupthi,Mohanthy,9848022336,Bhuwaneshwar)
(6,Archana,Mishra,9848022335,Chennai.)
grunt> describe student
2021-08-24 00:58:49,573 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-08-24 00:58:49,574 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
student: {id: int,firstname: chararray,lastname: chararray,phone: chararray,city
: chararray}
grunt> S
```

Explain Operator

The **explain** operator is used to display the logical, physical, and MapReduce execution plans of a relation.

Syntax

Given below is the syntax of the **explain** operator.

```
grunt> explain Relation_name;
```

let us explain the relation named student using the **explain** operator as shown below.

```
grunt> explain student;
```

Output

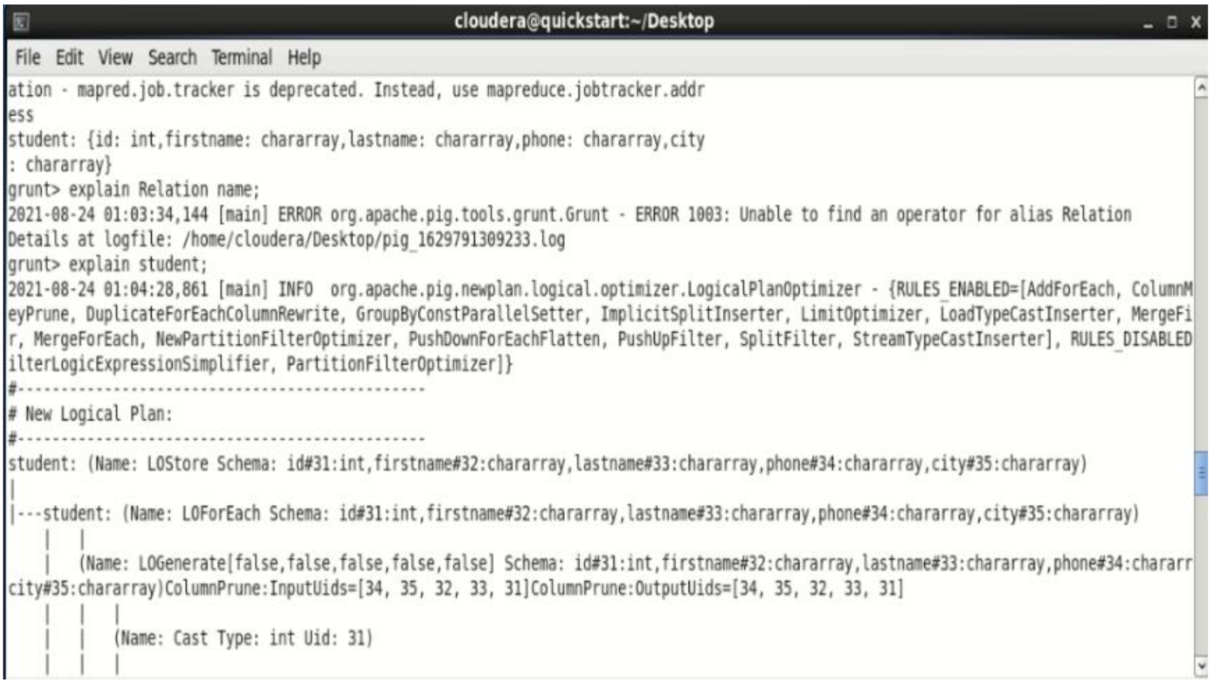
It will produce the following output.

```
grunt> explain student;

2021-08-24 01:04:28,861 [main] INFO
org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimize
r - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune,
DuplicateForEachColumnRewrite, GroupByConstParallelSetter,
ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter,
MergeFilter, MergeForEach, NewPartitionFilterOptimizer,
PushDownForEachFlatten, PushUpFilter, SplitFilter,
StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier,
PartitionFilterOptimizer]}

#-----------------------------------------------

# New Logical Plan:

#-----------------------------------------------

student: (Name: LOStore Schema:
id#31:int,firstname#32:chararray,lastname#33:chararray,phone
#34:chararray,city#35:chararray)

|

|---student: (Name: LOForEach Schema:
```

```
id#31:int,firstname#32:chararray,lastname#33:chararray,phone
#34:chararray,city#35:chararray)

    |   |

    |   (Name: LOGenerate[false,false,false,false,false]
Schema:
id#31:int,firstname#32:chararray,lastname#33:chararray,phone
#34:chararray,city#35:chararray)ColumnPrune:InputUids=[34,
35, 32, 33, 31]ColumnPrune:OutputUids=[34, 35, 32, 33, 31]

    |   |   |

    |   |   (Name: Cast Type: int Uid: 31)

    |   |   |

    |   |   |---id:(Name: Project Type: bytearray Uid: 31
Input: 0 Column: (*))

    |   |   |

    |   |   (Name: Cast Type: chararray Uid: 32)

    |   |   |

    |   |   |---firstname:(Name: Project Type: bytearray
Uid: 32 Input: 1 Column: (*))

    |   |   |

    |   |   (Name: Cast Type: chararray Uid: 33)

    |   |   |

    |   |   |---lastname:(Name: Project Type: bytearray Uid:
33 Input: 2 Column: (*))

    |   |   |

    |   |   (Name: Cast Type: chararray Uid: 34)

    |   |   |

    |   |   |---phone:(Name: Project Type: bytearray Uid: 34
Input: 3 Column: (*))

    |   |   |

    |   |   (Name: Cast Type: chararray Uid: 35)

    |   |   |

    |   |   |---city:(Name: Project Type: bytearray Uid: 35
Input: 4 Column: (*))

     |    |

     |    |---(Name: LOInnerLoad[0] Schema: id#31:bytearray)

     |    |

     |    |---(Name: LOInnerLoad[1] Schema:
firstname#32:bytearray)

     |    |

     |    |---(Name: LOInnerLoad[2] Schema:
lastname#33:bytearray)

     |    |

     |    |---(Name: LOInnerLoad[3] Schema:
phone#34:bytearray)

     |    |

     |    |---(Name: LOInnerLoad[4] Schema: city#35:bytearray)

     |

     |---student: (Name: LOLoad Schema:
id#31:bytearray,firstname#32:bytearray,lastname#33:bytearray
,phone#34:bytearray,city#35:bytearray)RequiredFields:null


#-----------------------------------------------

# Physical Plan:

#-----------------------------------------------

student: Store(fakefile:org.apache.pig.builtin.PigStorage) -
scope-36

|

|---student: New For
Each(false,false,false,false,false)[bag] - scope-35

     |    |

     |    Cast[int] - scope-21

     |    |

     |    |---Project[bytearray][0] - scope-20
```

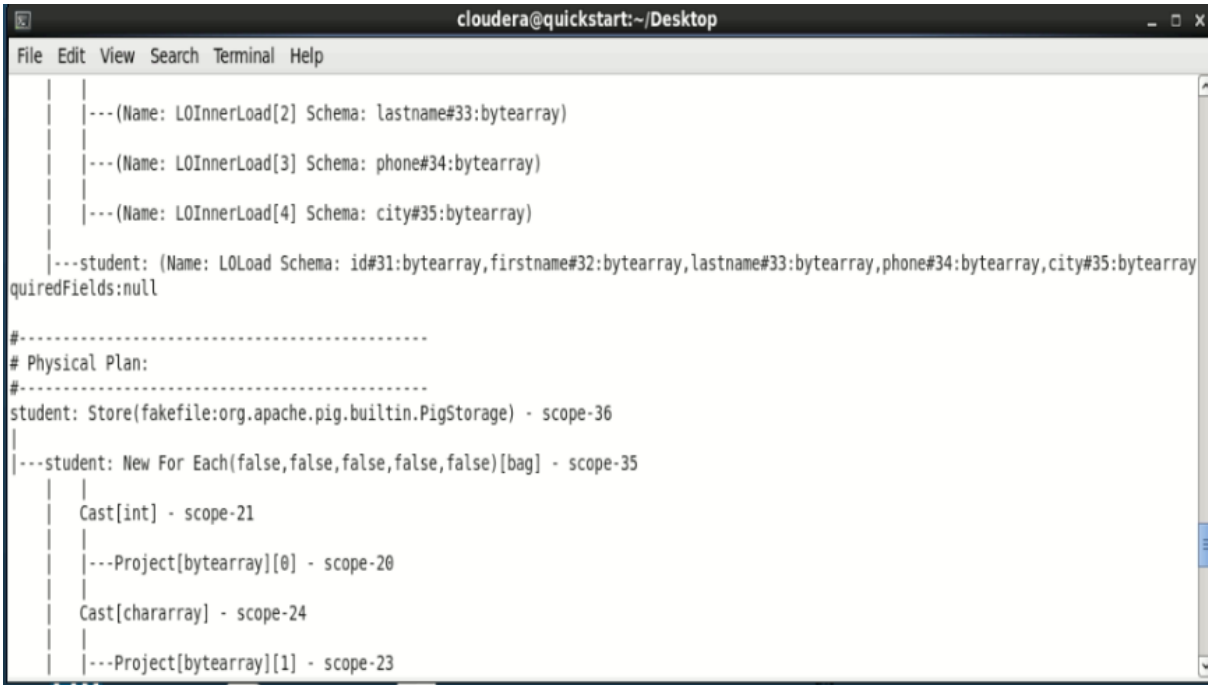
```
    |   |
    |   Cast[chararray] - scope-24
    |   |
    |   |---Project[bytearray][1] - scope-23
    |   |
    |   Cast[chararray] - scope-27
    |   |
    |   |---Project[bytearray][2] - scope-26
    |   |
    |   Cast[chararray] - scope-30
    |   |
    |   |---Project[bytearray][3] - scope-29
    |   |
    |   Cast[chararray] - scope-33
    |   |
    |   |---Project[bytearray][4] - scope-32
    |
    |---student:
Load(hdfs://quickstart.cloudera:8020/user/cloudera/smitrpate
l/student_data.txt:PigStorage(',')) - scope-19



2021-08-24 01:04:28,869 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.MRCompiler - File concatenation threshold: 100 optimistic?
false

2021-08-24 01:04:28,870 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.MultiQueryOptimizer - MR plan size before optimization: 1

2021-08-24 01:04:28,870 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.MultiQueryOptimizer - MR plan size after optimization: 1

#-----------------------------------------------------
```

```
# Map Reduce Plan
#--------------------------------------------------
MapReduce node scope-37
Map Plan
student: Store(fakefile:org.apache.pig.builtin.PigStorage) - scope-36
|
|---student: New For Each(false,false,false,false,false)[bag] - scope-35
    |   |
    |   Cast[int] - scope-21
    |   |
    |   |---Project[bytearray][0] - scope-20
    |   |
    |   Cast[chararray] - scope-24
    |   |
    |   |---Project[bytearray][1] - scope-23
    |   |
    |   Cast[chararray] - scope-27
    |   |
    |   |---Project[bytearray][2] - scope-26
    |   |
    |   Cast[chararray] - scope-30
    |   |
    |   |---Project[bytearray][3] - scope-29
    |   |
    |   Cast[chararray] - scope-33
    |   |
    |    |---Project[bytearray][4] - scope-32

    |

    |---student:
Load(hdfs://quickstart.cloudera:8020/user/cloudera/smitrpate
l/student_data.txt:PigStorage(',')) - scope-19--------

Global sort: false

----------------
```

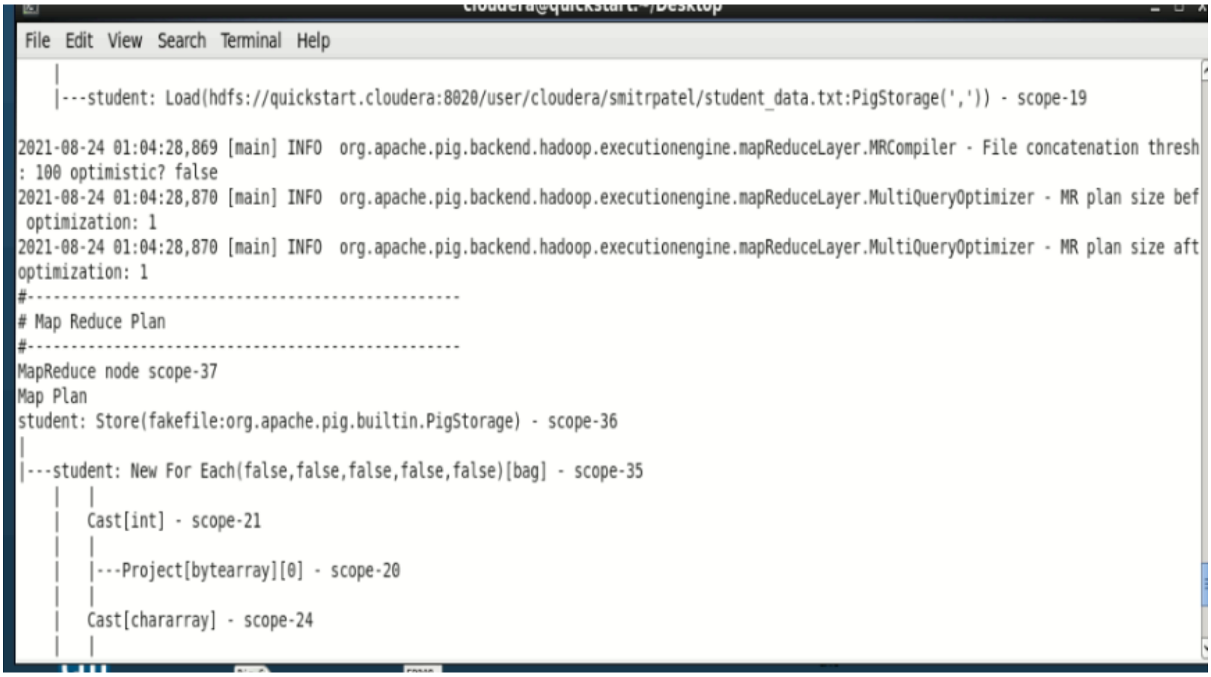


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

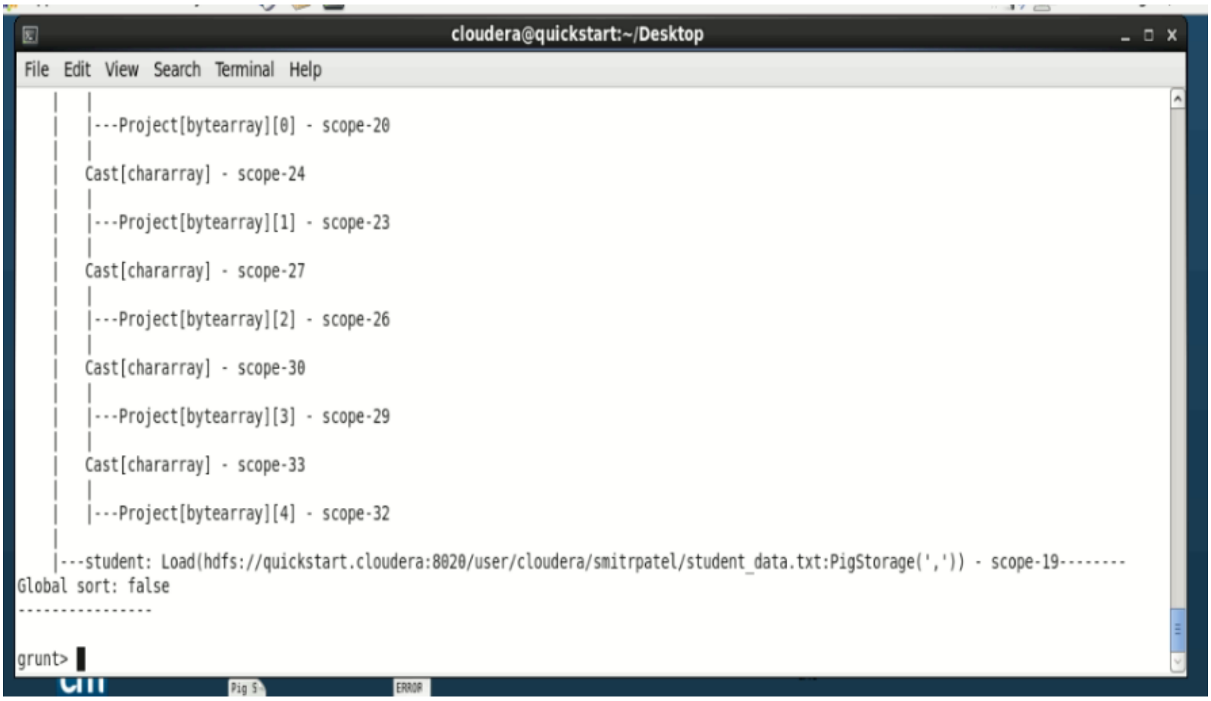
```
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
student: {id: int,firstname: chararray,lastname: chararray,phone: chararray,city
: chararray}
grunt> explain Relation name;
2021-08-24 01:03:34,144 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1003: Unable to find an operator for alias Relation
Details at logfile: /home/cloudera/Desktop/pig_1629791309233.log
grunt> explain student;
2021-08-24 01:04:28,861 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnM
eyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFi
r, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED
ilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
#-----------------------------------------------
# New Logical Plan:
#-----------------------------------------------
student: (Name: LOStore Schema: id#31:int,firstname#32:chararray,lastname#33:chararray,phone#34:chararray,city#35:chararray)
|
|---student: (Name: LOForEach Schema: id#31:int,firstname#32:chararray,lastname#33:chararray,phone#34:chararray,city#35:chararray)
    |   |
    |   (Name: LOGenerate[false,false,false,false,false] Schema: id#31:int,firstname#32:chararray,lastname#33:chararray,phone#34:chararr
city#35:chararray)ColumnPrune:InputUids=[34, 35, 32, 33, 31]ColumnPrune:OutputUids=[34, 35, 32, 33, 31]
    |   |   |
    |   |   (Name: Cast Type: int Uid: 31)
    |   |   |
```


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
    |   |
    |   |---(Name: LOInnerLoad[2] Schema: lastname#33:bytearray)
    |   |
    |   |---(Name: LOInnerLoad[3] Schema: phone#34:bytearray)
    |   |
    |   |---(Name: LOInnerLoad[4] Schema: city#35:bytearray)
    |
    |---student: (Name: LOLoad Schema: id#31:bytearray,firstname#32:bytearray,lastname#33:bytearray,phone#34:bytearray,city#35:bytearray
quiredFields:null

#-----------------------------------------------
# Physical Plan:
#-----------------------------------------------
student: Store(fakefile:org.apache.pig.builtin.PigStorage) - scope-36
|
|---student: New For Each(false,false,false,false,false)[bag] - scope-35
    |   |
    |   Cast[int] - scope-21
    |   |
    |   |---Project[bytearray][0] - scope-20
    |   |
    |   Cast[chararray] - scope-24
    |   |
    |   |---Project[bytearray][1] - scope-23
```


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
    |
    |---student: Load(hdfs://quickstart.cloudera:8020/user/cloudera/smitrpatel/student_data.txt:PigStorage(',')) - scope-19

2021-08-24 01:04:28,869 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thresh
: 100 optimistic? false
2021-08-24 01:04:28,870 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size bef
 optimization: 1
2021-08-24 01:04:28,870 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size aft
optimization: 1
#--------------------------------------------------
# Map Reduce Plan
#--------------------------------------------------
MapReduce node scope-37
Map Plan
student: Store(fakefile:org.apache.pig.builtin.PigStorage) - scope-36
|
|---student: New For Each(false,false,false,false,false)[bag] - scope-35
    |   |
    |   Cast[int] - scope-21
    |   |
    |   |---Project[bytearray][0] - scope-20
    |   |
    |   Cast[chararray] - scope-24
    |   |
```


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
    |    |
    |    |---Project[bytearray][0] - scope-20
    |    |
    |    Cast[chararray] - scope-24
    |    |
    |    |---Project[bytearray][1] - scope-23
    |    |
    |    Cast[chararray] - scope-27
    |    |
    |    |---Project[bytearray][2] - scope-26
    |    |
    |    Cast[chararray] - scope-30
    |    |
    |    |---Project[bytearray][3] - scope-29
    |    |
    |    Cast[chararray] - scope-33
    |    |
    |    |---Project[bytearray][4] - scope-32
    |
    |---student: Load(hdfs://quickstart.cloudera:8020/user/cloudera/smitrpatel/student_data.txt:PigStorage(',')) - scope-19--------
Global sort: false
----------------

grunt> 
```

Illustrate Operator

The **illustrate** operator gives you the step-by-step execution of a sequence of statements.

Syntax

Given below is the syntax of the **illustrate** operator.

```
grunt> illustrate Relation_name;
```

let us illustrate the relation named student as shown below.

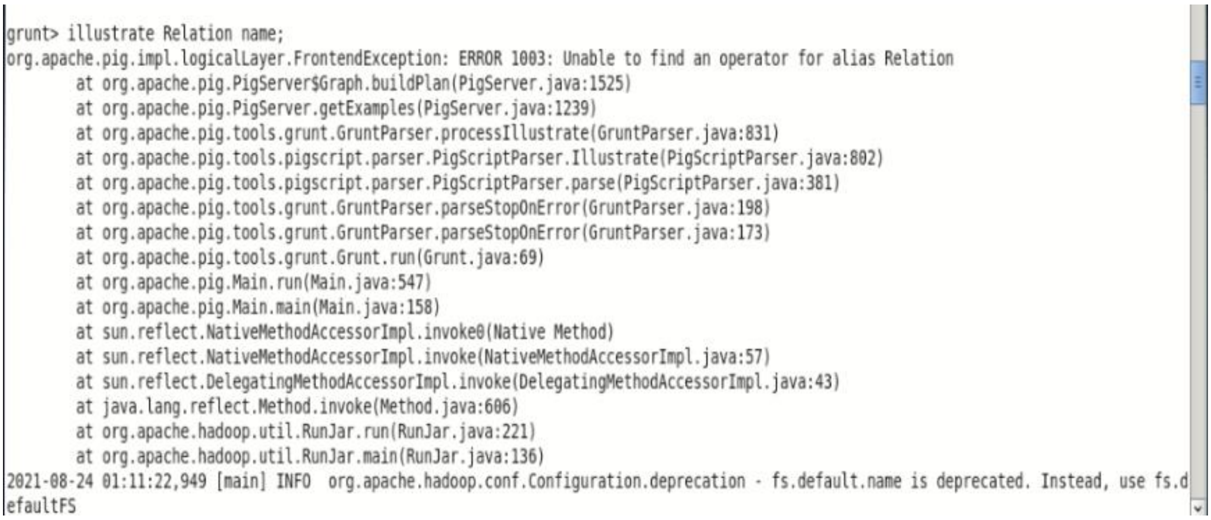
```
grunt> illustrate student;
```

Output

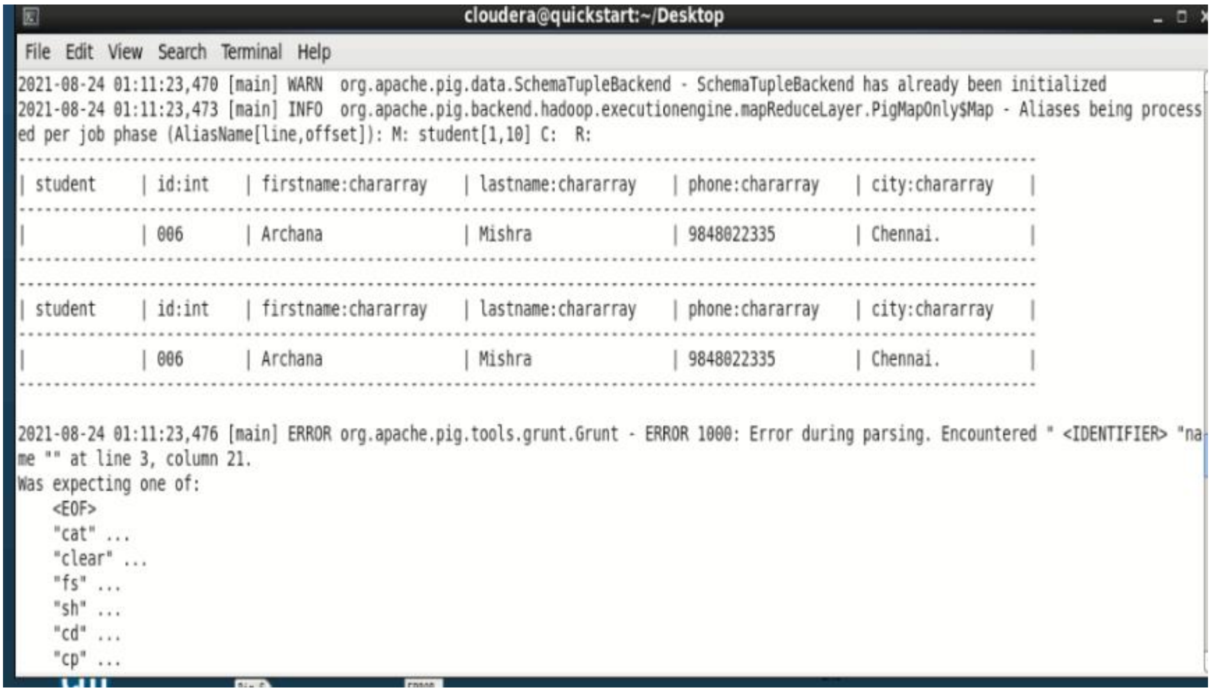
On executing the above statement, you will get the following output.

```
grunt> illustrate student;
```

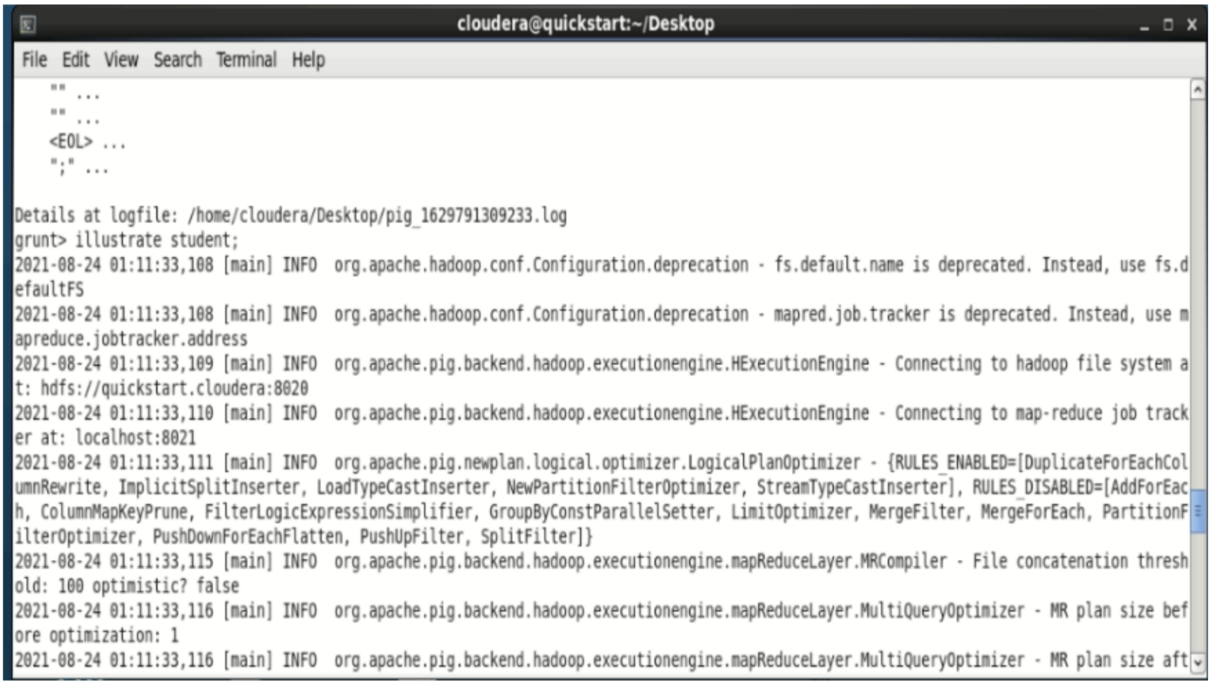
```
2021-08-24 01:11:33,330 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.PigMapOnly$Map - Aliases being processed per job phase
(AliasName[line,offset]): M: student[8,10] C:  R:

------------------------------------------------------------
---------------------------------------------------------

| student      | id:int     | firstname:chararray    |
lastname:chararray    | phone:chararray    | city:chararray
|

------------------------------------------------------------
---------------------------------------------------------

|              | 003        | Rajesh                 | Khanna
| 9848022339         | Delhi             |

------------------------------------------------------------
---------------------------------------------------------
```



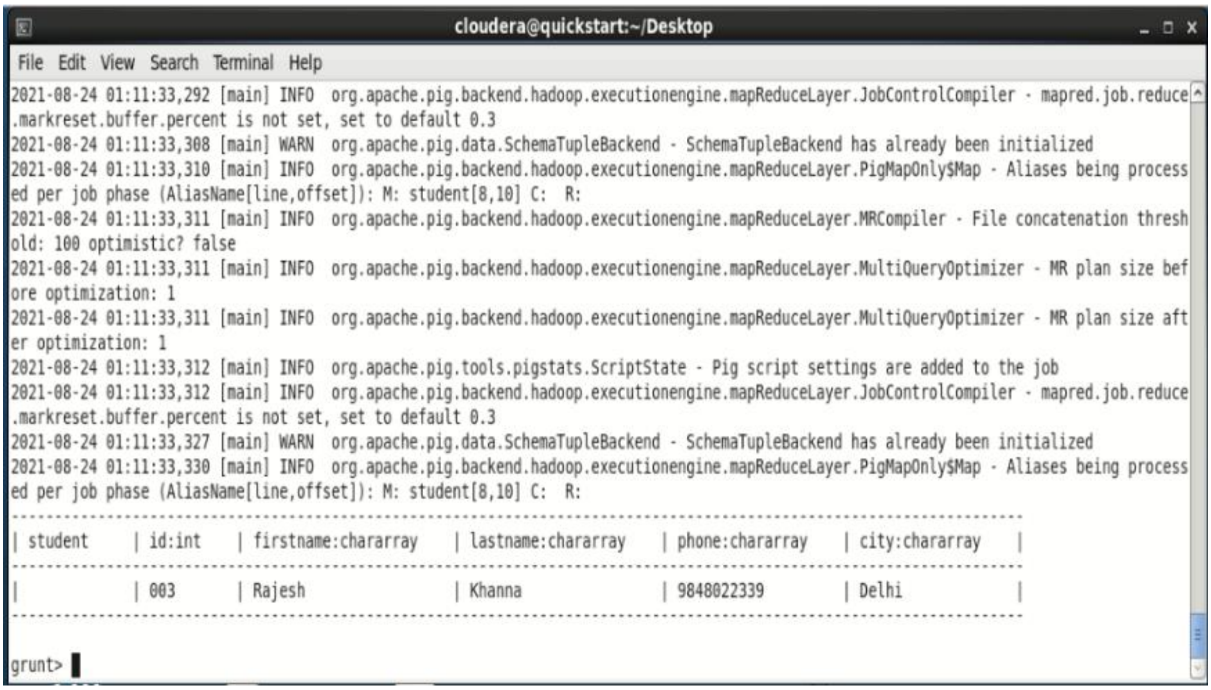
```
grunt> illustrate Relation name;
org.apache.pig.impl.logicalLayer.FrontendException: ERROR 1003: Unable to find an operator for alias Relation
        at org.apache.pig.PigServer$Graph.buildPlan(PigServer.java:1525)
        at org.apache.pig.PigServer.getExamples(PigServer.java:1239)
        at org.apache.pig.tools.grunt.GruntParser.processIllustrate(GruntParser.java:831)
        at org.apache.pig.tools.pigscript.parser.PigScriptParser.Illustrate(PigScriptParser.java:802)
        at org.apache.pig.tools.pigscript.parser.PigScriptParser.parse(PigScriptParser.java:381)
        at org.apache.pig.tools.grunt.GruntParser.parseStopOnError(GruntParser.java:198)
        at org.apache.pig.tools.grunt.GruntParser.parseStopOnError(GruntParser.java:173)
        at org.apache.pig.tools.grunt.Grunt.run(Grunt.java:69)
        at org.apache.pig.Main.run(Main.java:547)
        at org.apache.pig.Main.main(Main.java:158)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:606)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
2021-08-24 01:11:22,949 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.d
efaultFS
```


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
2021-08-24 01:11:23,470 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-08-24 01:11:23,473 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being process
ed per job phase (AliasName[line,offset]): M: student[1,10] C:  R:
--------------------------------------------------------------------------------------------------------------------
| student     | id:int     | firstname:chararray     | lastname:chararray     | phone:chararray     | city:chararray     |
--------------------------------------------------------------------------------------------------------------------
|             | 006        | Archana                 | Mishra                 | 9848022335          | Chennai.           |
--------------------------------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------------------------------
| student     | id:int     | firstname:chararray     | lastname:chararray     | phone:chararray     | city:chararray     |
--------------------------------------------------------------------------------------------------------------------
|             | 006        | Archana                 | Mishra                 | 9848022335          | Chennai.           |
--------------------------------------------------------------------------------------------------------------------

2021-08-24 01:11:23,476 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1000: Error during parsing. Encountered " <IDENTIFIER> "na
me "" at line 3, column 21.
Was expecting one of:
    <EOF>
    "cat" ...
    "clear" ...
    "fs" ...
    "sh" ...
    "cd" ...
    "cp" ...
```


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
    "" ...
    "" ...
    <EOL> ...
    ";" ...

Details at logfile: /home/cloudera/Desktop/pig_1629791309233.log
grunt> illustrate student;
2021-08-24 01:11:33,108 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.d
efaultFS
2021-08-24 01:11:33,108 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use m
apreduce.jobtracker.address
2021-08-24 01:11:33,109 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system a
t: hdfs://quickstart.cloudera:8020
2021-08-24 01:11:33,110 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job track
er at: localhost:8021
2021-08-24 01:11:33,111 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[DuplicateForEachCol
umnRewrite, ImplicitSplitInserter, LoadTypeCastInserter, NewPartitionFilterOptimizer, StreamTypeCastInserter], RULES_DISABLED=[AddForEac
h, ColumnMapKeyPrune, FilterLogicExpressionSimplifier, GroupByConstParallelSetter, LimitOptimizer, MergeFilter, MergeForEach, PartitionF
ilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter]}
2021-08-24 01:11:33,115 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thresh
old: 100 optimistic? false
2021-08-24 01:11:33,116 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size bef
ore optimization: 1
2021-08-24 01:11:33,116 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size aft
```


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
2021-08-24 01:11:33,292 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce
.markreset.buffer.percent is not set, set to default 0.3
2021-08-24 01:11:33,308 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-08-24 01:11:33,310 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being process
ed per job phase (AliasName[line,offset]): M: student[8,10] C:  R:
2021-08-24 01:11:33,311 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation thresh
old: 100 optimistic? false
2021-08-24 01:11:33,311 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size bef
ore optimization: 1
2021-08-24 01:11:33,311 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size aft
er optimization: 1
2021-08-24 01:11:33,312 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2021-08-24 01:11:33,312 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce
.markreset.buffer.percent is not set, set to default 0.3
2021-08-24 01:11:33,327 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-08-24 01:11:33,330 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapOnly$Map - Aliases being process
ed per job phase (AliasName[line,offset]): M: student[8,10] C:  R:
-------------------------------------------------------------------------------------------------------------------------------
| student     | id:int    | firstname:chararray    | lastname:chararray     | phone:chararray    | city:chararray      |
-------------------------------------------------------------------------------------------------------------------------------
|             | 003       | Rajesh                 | Khanna                 | 9848022339         | Delhi               |
-------------------------------------------------------------------------------------------------------------------------------

grunt> 
```

Group Operator

The GROUP operator is used to group the data in one or more relations. It collects the data having the same key.

Syntax

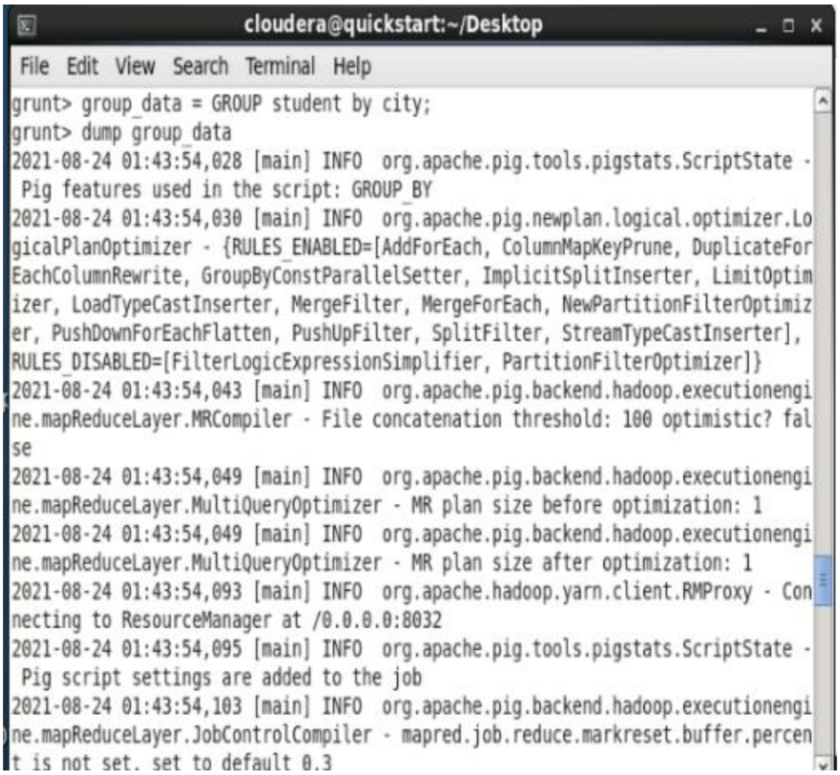
Given below is the syntax of the group operator.

grunt> Group_data = GROUP Relation_name BY age;

let us group the records/tuples in the relation by age as shown below. grunt> group_data = GROUP student_details by age;

grunt> group_data = GROUP student by city;

grunt> dump group_data


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
grunt> group_data = GROUP student by city;
grunt> dump group_data
2021-08-24 01:43:54,028 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: GROUP_BY
2021-08-24 01:43:54,030 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-08-24 01:43:54,043 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2021-08-24 01:43:54,049 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-08-24 01:43:54,049 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-08-24 01:43:54,093 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /0.0.0.0:8032
2021-08-24 01:43:54,095 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig script settings are added to the job
2021-08-24 01:43:54,103 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percen
t is not set. set to default 0.3
```

Verification

Verify the relation **group_data** using the **DUMP** operator as shown below.

```
grunt>  Dump  group_data;
```

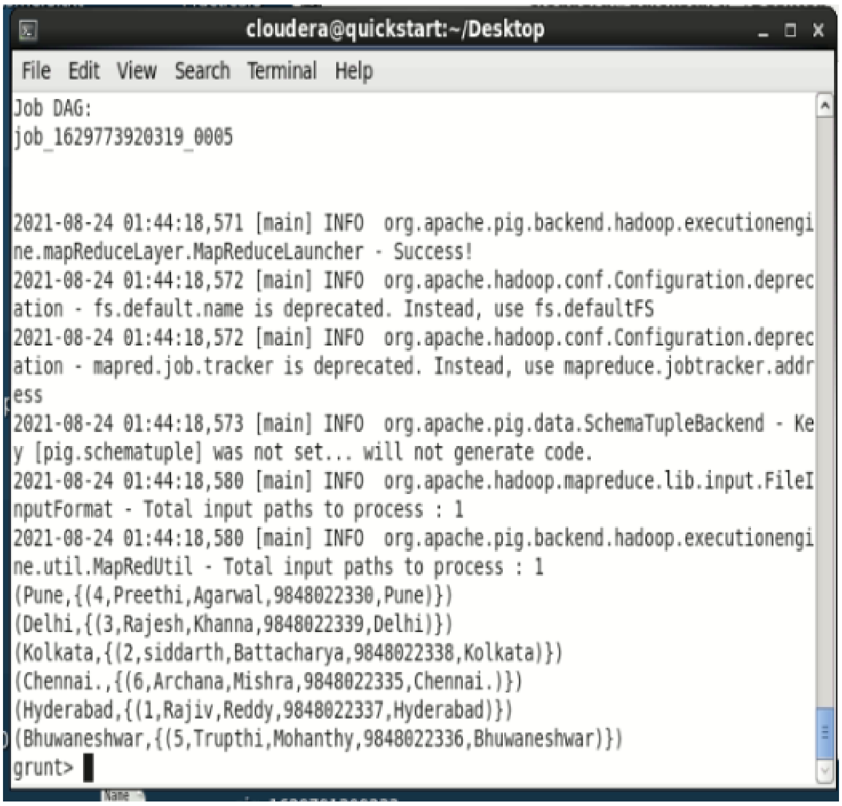
Output

Then you will get output displaying the contents of the relation named **group_data** as shown below. Here you can observe that the resulting schema has two columns −

- One is **age**, by which we have grouped the relation.
- The other is a **bag**, which contains the group of tuples, student records with the respective age.

```
(21,{(4,Preethi,Agarwal,21,9848022330,Pune),(1,Raji
v,Reddy,21,984 8022337,Hydera bad)})

(22,{(3,Rajesh,Khanna,22,9848022339,Delhi),(2,sidda
rth,Battachary a,22,984802233 8,Kolkata)})

(23,{(6,Archana,Mishra,23,9848022335,Chennai),(5,Tr
upthi,Mohanthy ,23,9848022336 ,Bhuwaneshwar)})

(24,{(8,Bharathi,Nambiayar,24,9848022333,Chennai),(
7,Komal,Nayak, 24,9848022334, trivendram)})
```

You can see the schema of the table after grouping the data using the **describe** command as shown below.

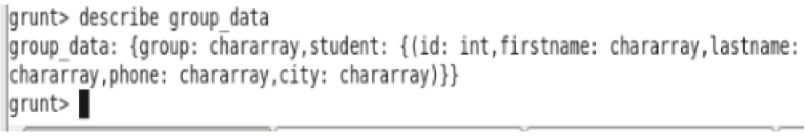


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
Job DAG:
job_1629773920319_0005


2021-08-24 01:44:18,571 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2021-08-24 01:44:18,572 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-08-24 01:44:18,572 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-08-24 01:44:18,573 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2021-08-24 01:44:18,580 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-08-24 01:44:18,580 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Pune,{(4,Preethi,Agarwal,9848022330,Pune)})
(Delhi,{(3,Rajesh,Khanna,9848022339,Delhi)})
(Kolkata,{(2,siddarth,Battacharya,9848022338,Kolkata)})
(Chennai.,{(6,Archana,Mishra,9848022335,Chennai.)})
(Hyderabad,{(1,Rajiv,Reddy,9848022337,Hyderabad)})
(Bhuwaneshwar,{(5,Trupthi,Mohanthy,9848022336,Bhuwaneshwar)})
grunt>
```

grunt> Describe group_data;

group_data: {group: int,student_details: {(id: int,firstname: chararray, lastname: chararray,age: int,phone: chararray,city: chararray)}}



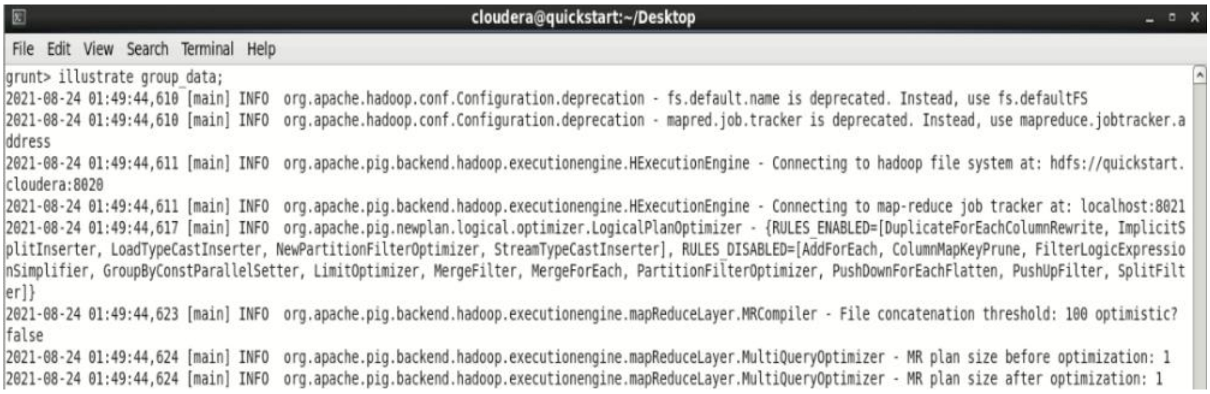
```
grunt> describe group_data
group_data: {group: chararray,student: {(id: int,firstname: chararray,lastname:
chararray,phone: chararray,city: chararray)}}
grunt>
```

In the same way, you can get the sample illustration of the schema using the illustrate command as shown below.

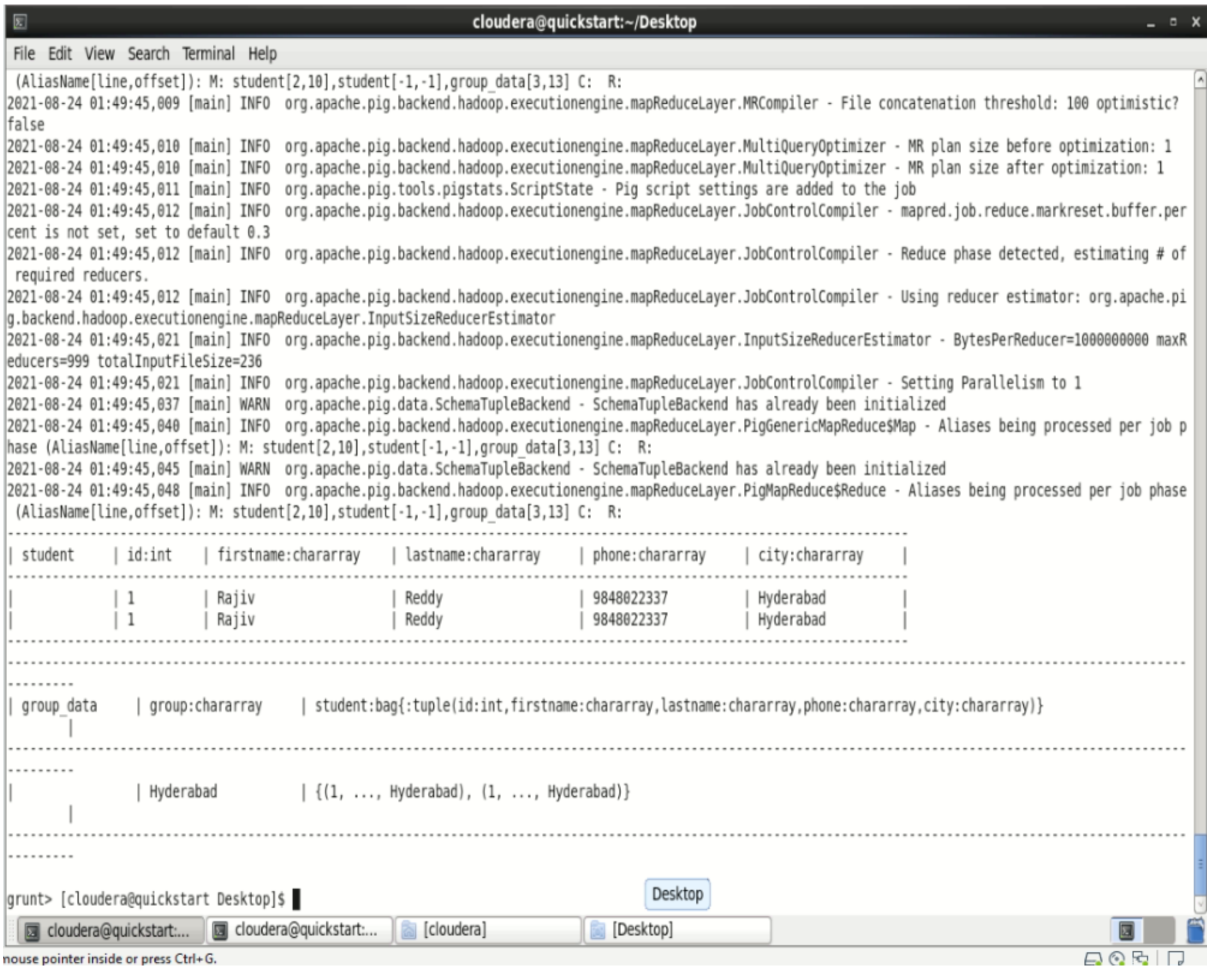
```
grunt>illustrate group_data;
```

It will produce the following output –

```
WARN  org.apache.pig.data.SchemaTupleBackend -
SchemaTupleBackend has already been initialized

2021-08-24 01:49:45,048 [main] INFO
org.apache.pig.backend.hadoop.executionengine.mapRedu
ceLayer.PigMapReduce$Reduce - Aliases being processed
per job phase (AliasName[line,offset]): M:
student[2,10],student[-1,-1],group_data[3,13] C:  R:

-----------------------------------------------------
-----------------------------------------------------
--------------

| student     | id:int    | firstname:chararray    |
lastname:chararray     | phone:chararray     |
city:chararray     |

-----------------------------------------------------
-----------------------------------------------------
--------------

|             | 1         | Rajiv                  |
Reddy                  | 9848022337          |
Hyderabad          |

|             | 1         | Rajiv                  |
Reddy                  | 9848022337          |
Hyderabad          |

-----------------------------------------------------
-----------------------------------------------------
--------------

-----------------------------------------------------
-----------------------------------------------------
-----------------------------------------------------
-------

| group_data     | group:chararray     |
student:bag{:tuple(id:int,firstname:chararray,lastnam
e:chararray,phone:chararray,city:chararray)}
|

-----------------------------------------------------
-----------------------------------------------------
-----------------------------------------------------
-------

|                | Hyderabad           | {(1, ...,
Hyderabad), (1, ..., Hyderabad)}
|
```


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
grunt> illustrate group_data;
2021-08-24 01:49:44,610 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-08-24 01:49:44,610 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.a
ddress
2021-08-24 01:49:44,611 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://quickstart.
cloudera:8020
2021-08-24 01:49:44,611 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: localhost:8021
2021-08-24 01:49:44,617 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[DuplicateForEachColumnRewrite, ImplicitS
plitInserter, LoadTypeCastInserter, NewPartitionFilterOptimizer, StreamTypeCastInserter], RULES_DISABLED=[AddForEach, ColumnMapKeyPrune, FilterLogicExpressio
nSimplifier, GroupByConstParallelSetter, LimitOptimizer, MergeFilter, MergeForEach, PartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilt
er]}
2021-08-24 01:49:44,623 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic?
false
2021-08-24 01:49:44,624 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-08-24 01:49:44,624 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
```

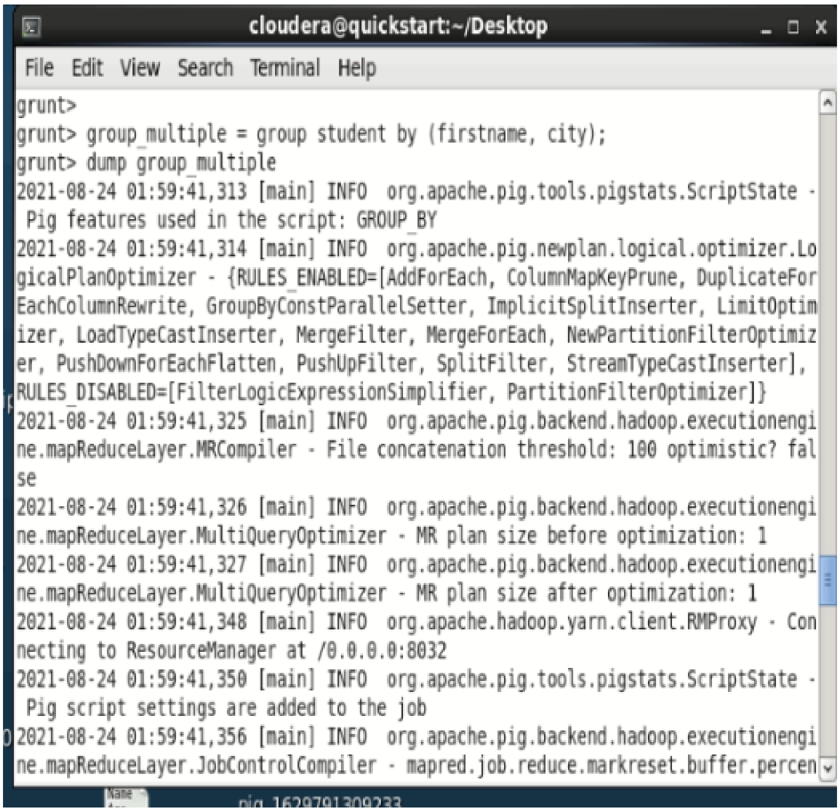

cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Desktop
cloudera@quickstart:...
cloudera@quickstart:...
[cloudera]
[Desktop]
nouse pointer inside or press Ctrl+G.

```
 (AliasName[line,offset]): M: student[2,10],student[-1,-1],group_data[3,13] C:  R:
2021-08-24 01:49:45,009 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic?
false
2021-08-24 01:49:45,010 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-08-24 01:49:45,010 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-08-24 01:49:45,011 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2021-08-24 01:49:45,012 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.per
cent is not set, set to default 0.3
2021-08-24 01:49:45,012 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Reduce phase detected, estimating # of
 required reducers.
2021-08-24 01:49:45,012 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Using reducer estimator: org.apache.pi
g.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator
2021-08-24 01:49:45,021 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.InputSizeReducerEstimator - BytesPerReducer=1000000000 maxR
educers=999 totalInputFileSize=236
2021-08-24 01:49:45,021 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2021-08-24 01:49:45,037 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-08-24 01:49:45,040 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigGenericMapReduce$Map - Aliases being processed per job p
hase (AliasName[line,offset]): M: student[2,10],student[-1,-1],group_data[3,13] C:  R:
2021-08-24 01:49:45,045 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2021-08-24 01:49:45,048 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.PigMapReduce$Reduce - Aliases being processed per job phase
 (AliasName[line,offset]): M: student[2,10],student[-1,-1],group_data[3,13] C:  R:
---------------------------------------------------------------------------------------------------------
| student     | id:int    | firstname:chararray    | lastname:chararray     | phone:chararray    | city:chararray     |
---------------------------------------------------------------------------------------------------------
|             | 1         | Rajiv                  | Reddy                  | 9848022337         | Hyderabad          |
|             | 1         | Rajiv                  | Reddy                  | 9848022337         | Hyderabad          |
---------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------------------
--------
| group_data     | group:chararray     | student:bag{:tuple(id:int,firstname:chararray,lastname:chararray,phone:chararray,city:chararray)}
       |
--------------------------------------------------------------------------------------------------------------------------------------------
--------
|                | Hyderabad           | {(1, ..., Hyderabad), (1, ..., Hyderabad)}
       |
--------------------------------------------------------------------------------------------------------------------------------------------
--------

grunt> [cloudera@quickstart Desktop]$ █
```

Grouping by Multiple Columns

Let us group the relation by age and city as shown below.

```
grunt> group_multiple = GROUP student_details by (age,
city);
```



cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
grunt>
grunt> group_multiple = group student by (firstname, city);
grunt> dump group_multiple
2021-08-24 01:59:41,313 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: GROUP_BY
2021-08-24 01:59:41,314 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-08-24 01:59:41,325 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2021-08-24 01:59:41,326 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-08-24 01:59:41,327 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-08-24 01:59:41,348 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /0.0.0.0:8032
2021-08-24 01:59:41,350 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig script settings are added to the job
2021-08-24 01:59:41,356 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percen
```

You can verify the content of the relation named **group_multiple** using the Dump operator as shown below.

```
grunt> dump group_multiple;
```

2021-08-24 02:00:05,184 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

((Rajiv,Hyderabad),{(1,Rajiv,Reddy,9848022337,Hyderabad)})

((Rajesh,Delhi),{(3,Rajesh,Khanna,9848022339,Delhi)})

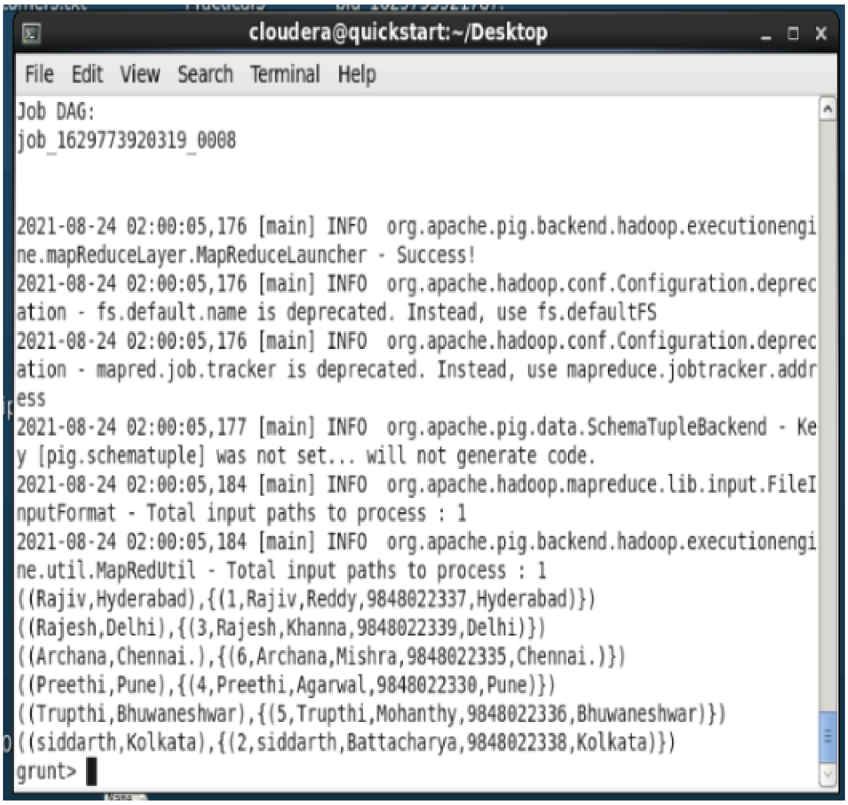
((Archana,Chennai.),{(6,Archana,Mishra,9848022335,Chennai.)})

((Preethi,Pune),{(4,Preethi,Agarwal,9848022330,Pune)})

((Trupthi,Bhuwaneshwar),{(5,Trupthi,Mohanthy,9848022336,Bhuwaneshwar)})

((siddarth,Kolkata),{(2,siddarth,Battacharya,9848022338,Kolkata)})

grunt>



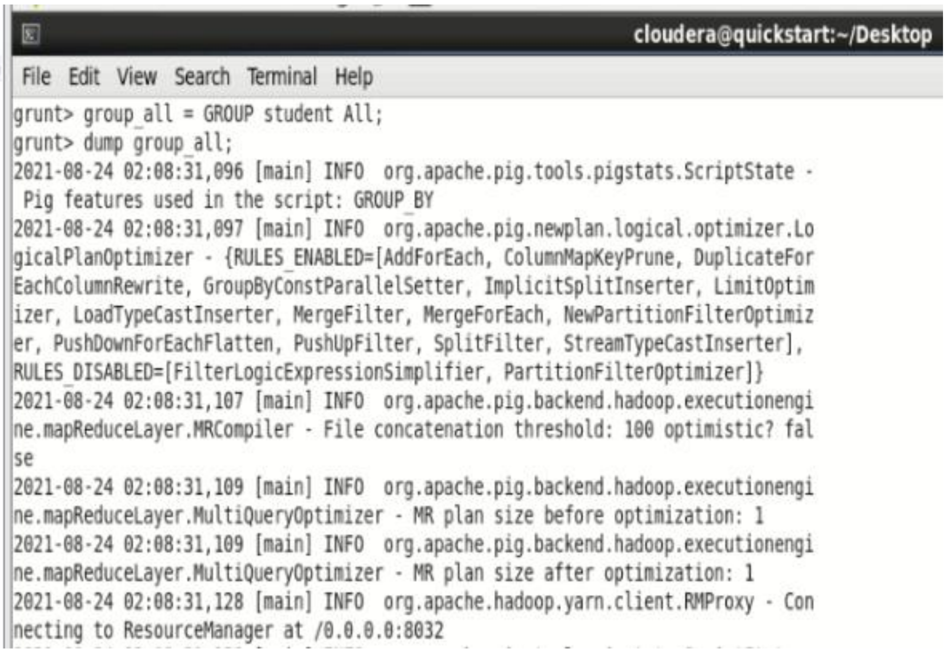
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
Job DAG:
job_1629773920319_0008


2021-08-24 02:00:05,176 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2021-08-24 02:00:05,176 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-08-24 02:00:05,176 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-08-24 02:00:05,177 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2021-08-24 02:00:05,184 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-08-24 02:00:05,184 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
((Rajiv,Hyderabad),{(1,Rajiv,Reddy,9848022337,Hyderabad)})
((Rajesh,Delhi),{(3,Rajesh,Khanna,9848022339,Delhi)})
((Archana,Chennai.),{(6,Archana,Mishra,9848022335,Chennai.)})
((Preethi,Pune),{(4,Preethi,Agarwal,9848022330,Pune)})
((Trupthi,Bhuwaneshwar),{(5,Trupthi,Mohanthy,9848022336,Bhuwaneshwar)})
((siddarth,Kolkata),{(2,siddarth,Battacharya,9848022338,Kolkata)})
grunt>
```

Group All

You can group a relation by all the columns as shown below.

```
grunt> group_all = GROUP student_details All;
```



cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
grunt> group_all = GROUP student All;
grunt> dump group_all;
2021-08-24 02:08:31,096 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: GROUP_BY
2021-08-24 02:08:31,097 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-08-24 02:08:31,107 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2021-08-24 02:08:31,109 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-08-24 02:08:31,109 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-08-24 02:08:31,128 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /0.0.0.0:8032
```

Now, verify the content of the relation **group_all** as shown below.

```
grunt> Dump group_all;
```

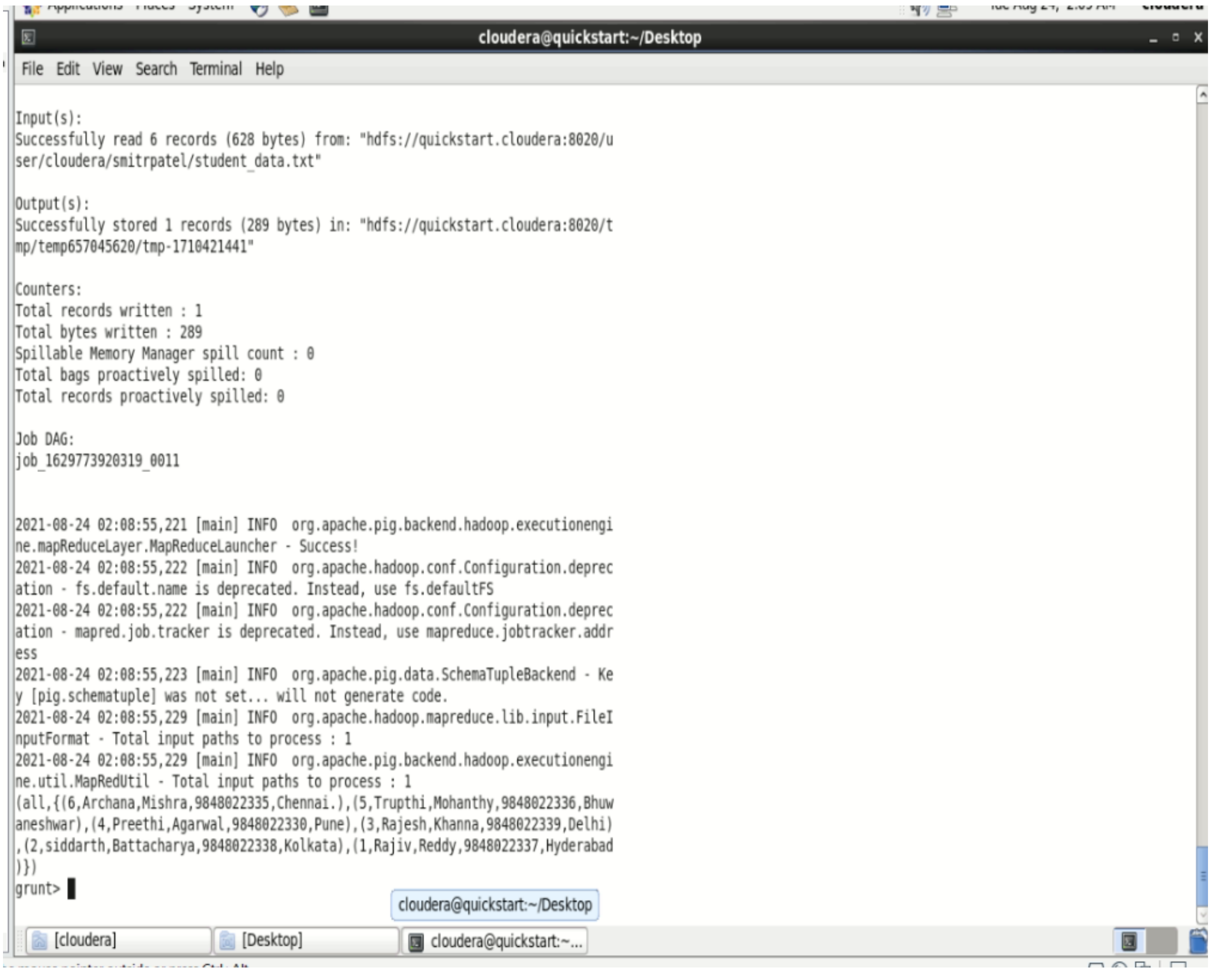
2021-08-24 02:08:55,223 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.

2021-08-24 02:08:55,229 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1

2021-08-24 02:08:55,229 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

(all,{(6,Archana,Mishra,9848022335,Chennai.),(5,Trupthi,Mohanthy,9848022336,Bhuwaneshwar),(4,Preethi,Agarwal,9848022330,Pune),(3,Rajesh,Khanna,9848022339,Delhi),(2,siddarth,Battacharya,9848022338,Kolkata),(1,Rajiv,Reddy,9848022337,Hyderabad)})

grunt>


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
cloudera@quickstart:~/Desktop
[cloudera]
[Desktop]
cloudera@quickstart:~...

```
Input(s):
Successfully read 6 records (628 bytes) from: "hdfs://quickstart.cloudera:8020/u
ser/cloudera/smitrpatel/student_data.txt"

Output(s):
Successfully stored 1 records (289 bytes) in: "hdfs://quickstart.cloudera:8020/t
mp/temp657045620/tmp-1710421441"

Counters:
Total records written : 1
Total bytes written : 289
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1629773920319_0011


2021-08-24 02:08:55,221 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2021-08-24 02:08:55,222 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-08-24 02:08:55,222 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-08-24 02:08:55,223 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2021-08-24 02:08:55,229 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-08-24 02:08:55,229 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(all,{(6,Archana,Mishra,9848022335,Chennai.),(5,Trupthi,Mohanthy,9848022336,Bhuw
aneshwar),(4,Preethi,Agarwal,9848022330,Pune),(3,Rajesh,Khanna,9848022339,Delhi)
,(2,siddarth,Battacharya,9848022338,Kolkata),(1,Rajiv,Reddy,9848022337,Hyderabad
)})
grunt> 
```

Cogroup Operator

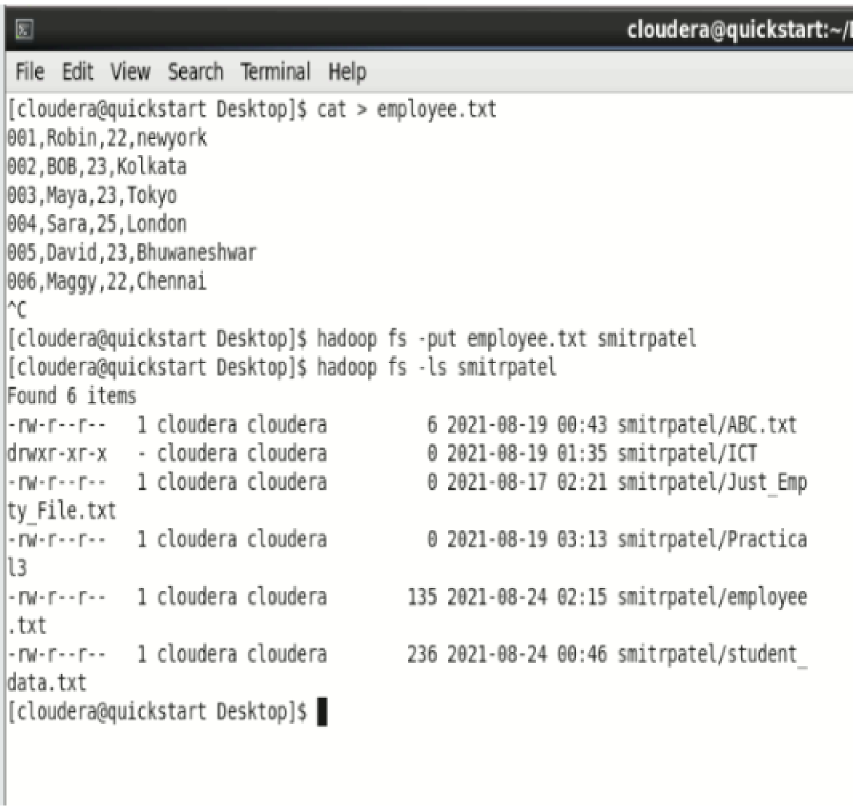
The **COGROUP** operator works more or less in the same way as the GROUP operator. The only difference between the two operators is that the **group** operator is normally used with one relation, while the **cogroup** operator is used in statements involving two or more relations.

Grouping Two Relations using Cogroup

Assume that we have two files namely **student_details.txt** and **employee_details.txt** in the HDFS directory **/pig_data/**

employee_details.txt

```
001,Robin,22,newyork

002,BOB,23,Kolkata

003,Maya,23,Tokyo

004,Sara,25,London

005,David,23,Bhuwaneshwar

006,Maggy,22,Chennai
```

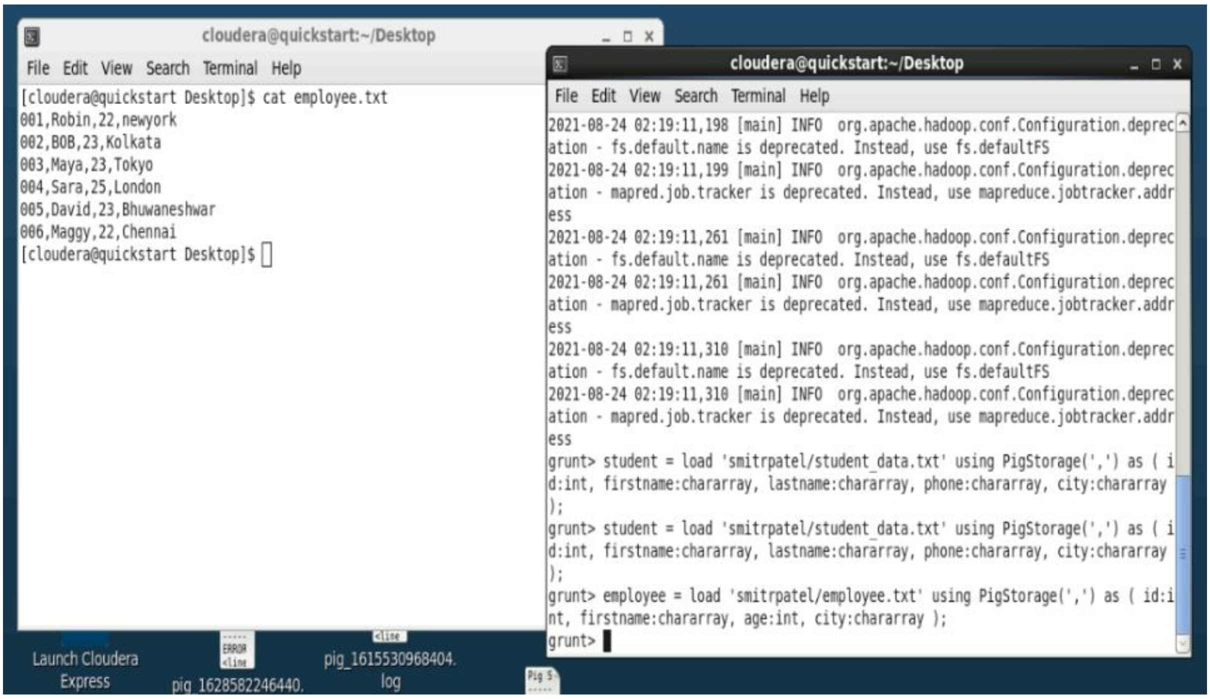


```
[cloudera@quickstart Desktop]$ cat > employee.txt
001,Robin,22,newyork
002,BOB,23,Kolkata
003,Maya,23,Tokyo
004,Sara,25,London
005,David,23,Bhuwaneshwar
006,Maggy,22,Chennai
^C
[cloudera@quickstart Desktop]$ hadoop fs -put employee.txt smitrpatel
[cloudera@quickstart Desktop]$ hadoop fs -ls smitrpatel
Found 6 items
-rw-r--r--   1 cloudera cloudera          6 2021-08-19 00:43 smitrpatel/ABC.txt
drwxr-xr-x   - cloudera cloudera          0 2021-08-19 01:35 smitrpatel/ICT
-rw-r--r--   1 cloudera cloudera          0 2021-08-17 02:21 smitrpatel/Just_Emp
ty_File.txt
-rw-r--r--   1 cloudera cloudera          0 2021-08-19 03:13 smitrpatel/Practica
l3
-rw-r--r--   1 cloudera cloudera        135 2021-08-24 02:15 smitrpatel/employee
.txt
-rw-r--r--   1 cloudera cloudera        236 2021-08-24 00:46 smitrpatel/student_
data.txt
[cloudera@quickstart Desktop]$ 
```

And we have loaded these files into Pig with the relation names **student_details** and **employee_details** respectively

Now, let us group the records/tuples of the relations **student_details** and **employee_details** with the key age, as shown below.

```
grunt> cogroup_data = COGROUP student_details by
age, employee_details by age;
```



```
[cloudera@quickstart Desktop]$ cat employee.txt
001,Robin,22,newyork
002,BOB,23,Kolkata
003,Maya,23,Tokyo
004,Sara,25,London
005,David,23,Bhuwaneshwar
006,Maggy,22,Chennai
[cloudera@quickstart Desktop]$ 
```

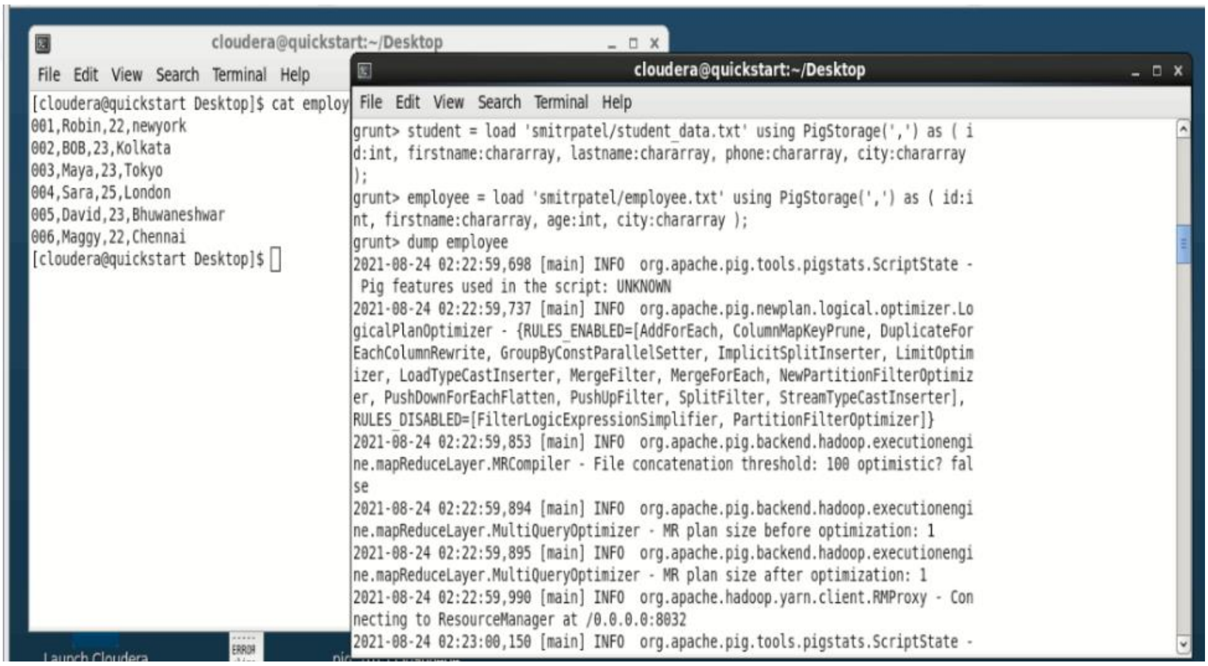
```
2021-08-24 02:19:11,198 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-08-24 02:19:11,199 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-08-24 02:19:11,261 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-08-24 02:19:11,261 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-08-24 02:19:11,310 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-08-24 02:19:11,310 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
grunt> student = load 'smitrpatel/student_data.txt' using PigStorage(',') as ( i
d:int, firstname:chararray, lastname:chararray, phone:chararray, city:chararray
);
grunt> student = load 'smitrpatel/student_data.txt' using PigStorage(',') as ( i
d:int, firstname:chararray, lastname:chararray, phone:chararray, city:chararray
);
grunt> employee = load 'smitrpatel/employee.txt' using PigStorage(',') as ( id:i
nt, firstname:chararray, age:int, city:chararray );
grunt> 
```

cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Launch Cloudera Express
pig_1628582246440.
pig_1615530968404.log
Pig 5

grunt>student = load 'smitrpatel/student_data.txt' using PigStorage(',') as (id:int, firstname:chararray, lastname:chararray, phone:chararray, city:chararray);

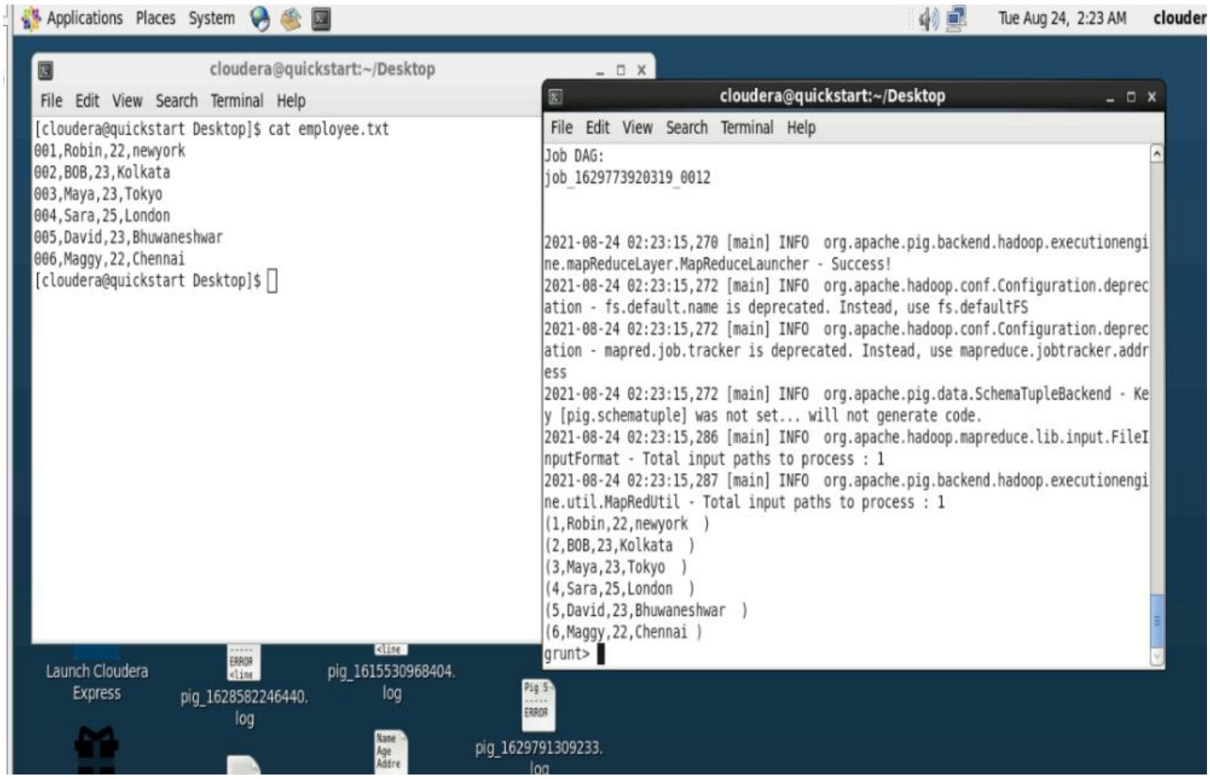
grunt>employee = load 'smitrpatel/employee.txt' using PigStorage(',') as (id:int, firstname:chararray, age:int, city:chararray);

grunt>dump employee


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Launch Cloudera

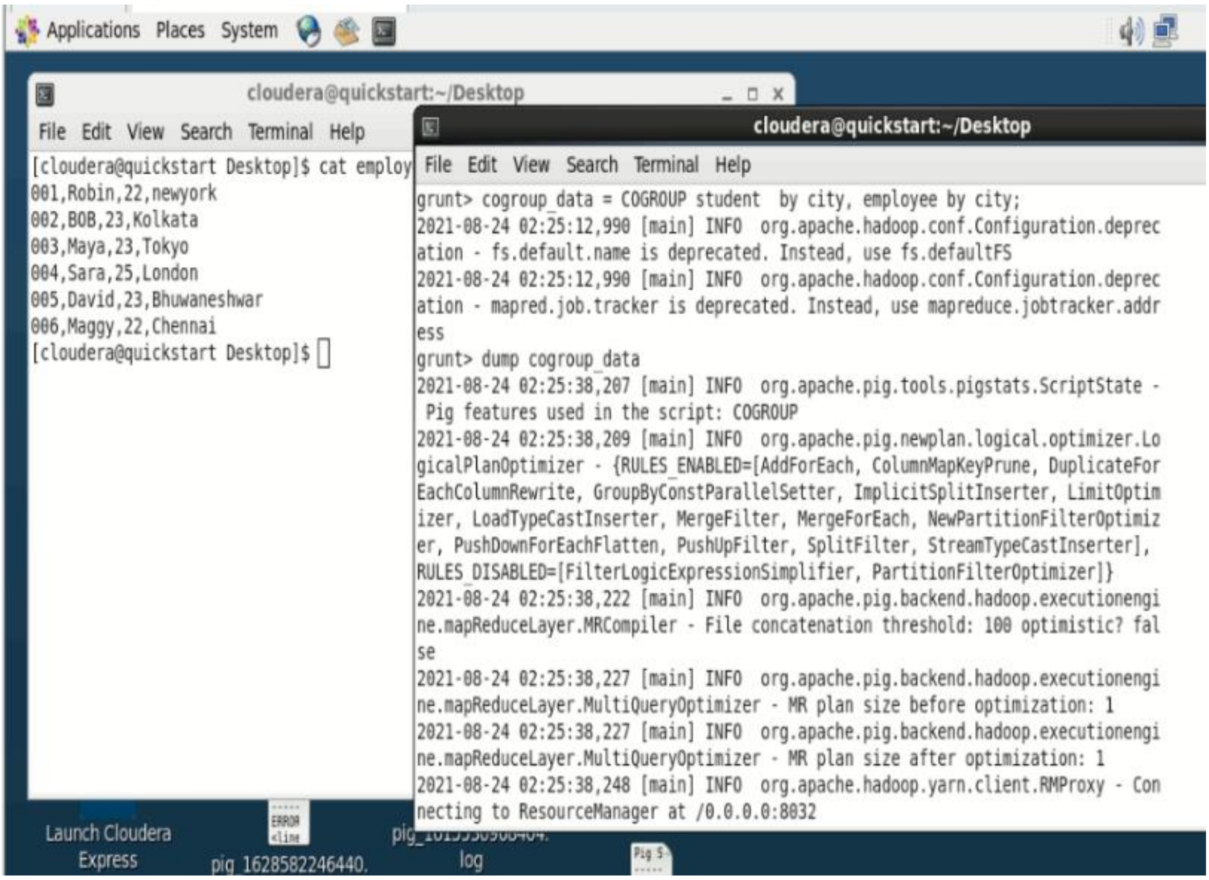
```
[cloudera@quickstart Desktop]$ cat employ
001,Robin,22,newyork
002,BOB,23,Kolkata
003,Maya,23,Tokyo
004,Sara,25,London
005,David,23,Bhuwaneshwar
006,Maggy,22,Chennai
[cloudera@quickstart Desktop]$ 
```

```
grunt> student = load 'smitrpatel/student_data.txt' using PigStorage(',') as ( i
d:int, firstname:chararray, lastname:chararray, phone:chararray, city:chararray
);
grunt> employee = load 'smitrpatel/employee.txt' using PigStorage(',') as ( id:i
nt, firstname:chararray, age:int, city:chararray );
grunt> dump employee
2021-08-24 02:22:59,698 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: UNKNOWN
2021-08-24 02:22:59,737 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-08-24 02:22:59,853 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2021-08-24 02:22:59,894 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-08-24 02:22:59,895 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-08-24 02:22:59,990 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /0.0.0.0:8032
2021-08-24 02:23:00,150 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
```


Applications Places System
Tue Aug 24, 2:23 AM
clouder
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Launch Cloudera Express
pig_1628582246440.log
pig_1615530968404.log
pig_1629791309233.log

```
[cloudera@quickstart Desktop]$ cat employee.txt
001,Robin,22,newyork
002,BOB,23,Kolkata
003,Maya,23,Tokyo
004,Sara,25,London
005,David,23,Bhuwaneshwar
006,Maggy,22,Chennai
[cloudera@quickstart Desktop]$ 
```

```
Job DAG:
job_1629773920319_0012


2021-08-24 02:23:15,270 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2021-08-24 02:23:15,272 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-08-24 02:23:15,272 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-08-24 02:23:15,272 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2021-08-24 02:23:15,286 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-08-24 02:23:15,287 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(1,Robin,22,newyork  )
(2,BOB,23,Kolkata  )
(3,Maya,23,Tokyo  )
(4,Sara,25,London  )
(5,David,23,Bhuwaneshwar  )
(6,Maggy,22,Chennai )
grunt> 
```


Applications Places System
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Launch Cloudera Express
pig_1628582246440.
log

```
[cloudera@quickstart Desktop]$ cat employ
001,Robin,22,newyork
002,BOB,23,Kolkata
003,Maya,23,Tokyo
004,Sara,25,London
005,David,23,Bhuwaneshwar
006,Maggy,22,Chennai
[cloudera@quickstart Desktop]$
```

```
grunt> cogroup_data = COGROUP student  by city, employee by city;
2021-08-24 02:25:12,990 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-08-24 02:25:12,990 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
grunt> dump cogroup_data
2021-08-24 02:25:38,207 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: COGROUP
2021-08-24 02:25:38,209 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateFor
EachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptim
izer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimiz
er, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter],
RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2021-08-24 02:25:38,222 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? fal
se
2021-08-24 02:25:38,227 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2021-08-24 02:25:38,227 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2021-08-24 02:25:38,248 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Con
necting to ResourceManager at /0.0.0.0:8032
```

Verification

Verify the relation **cogroup_data** using the **DUMP** operator as shown below.

```
grunt> Dump cogroup_data;
```

Output

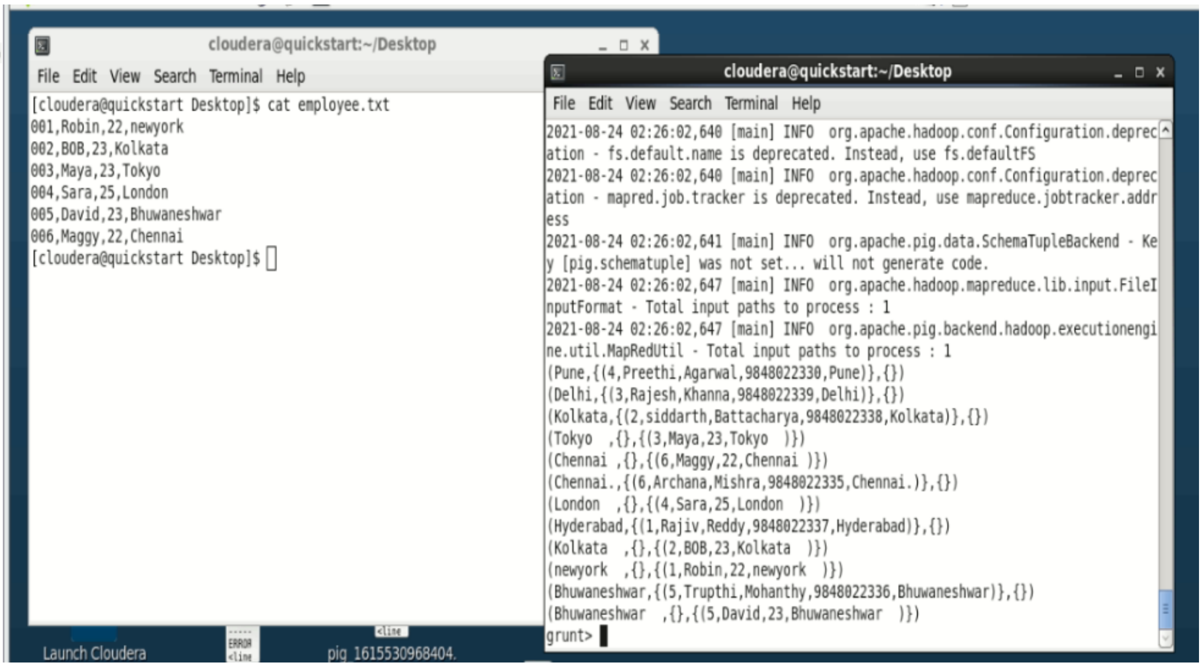
It will produce the following output, displaying the contents of the relation named **cogroup_data** as shown below.

2021-08-24 02:26:02,647 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1

(Pune,{(4,Preethi,Agarwal,9848022330,Pune)},{})

(Delhi,{(3,Rajesh,Khanna,9848022339,Delhi)},{})

```
(Kolkata,{(2,siddarth,Battacharya,9848022338,Kolkata)},{})

(Tokyo  ,{},{(3,Maya,23,Tokyo  )})

(Chennai ,{},{(6,Maggy,22,Chennai )})

(Chennai.,{(6,Archana,Mishra,9848022335,Chennai.)},{})

(London  ,{},{(4,Sara,25,London  )})

(Hyderabad,{(1,Rajiv,Reddy,9848022337,Hyderabad)},{})

(Kolkata  ,{},{(2,BOB,23,Kolkata  )})

(newyork  ,{},{(1,Robin,22,newyork  )})

(Bhuwaneshwar,{(5,Trupthi,Mohanthy,9848022336,Bhuwaneshwar)},{})

(Bhuwaneshwar  ,{},{(5,David,23,Bhuwaneshwar  )})

grunt>
```


cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Launch Cloudera
pig_1615530968404.

```
[cloudera@quickstart Desktop]$ cat employee.txt
001,Robin,22,newyork
002,BOB,23,Kolkata
003,Maya,23,Tokyo
004,Sara,25,London
005,David,23,Bhuwaneshwar
006,Maggy,22,Chennai
[cloudera@quickstart Desktop]$
```

cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help

```
2021-08-24 02:26:02,640 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-08-24 02:26:02,640 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-08-24 02:26:02,641 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2021-08-24 02:26:02,647 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-08-24 02:26:02,647 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Pune,{(4,Preethi,Agarwal,9848022330,Pune)},{})
(Delhi,{(3,Rajesh,Khanna,9848022339,Delhi)},{})
(Kolkata,{(2,siddarth,Battacharya,9848022338,Kolkata)},{})
(Tokyo  ,{},{(3,Maya,23,Tokyo  )})
(Chennai ,{},{(6,Maggy,22,Chennai )})
(Chennai.,{(6,Archana,Mishra,9848022335,Chennai.)},{})
(London  ,{},{(4,Sara,25,London  )})
(Hyderabad,{(1,Rajiv,Reddy,9848022337,Hyderabad)},{})
(Kolkata  ,{},{(2,BOB,23,Kolkata  )})
(newyork  ,{},{(1,Robin,22,newyork  )})
(Bhuwaneshwar,{(5,Trupthi,Mohanthy,9848022336,Bhuwaneshwar)},{})
(Bhuwaneshwar  ,{},{(5,David,23,Bhuwaneshwar  )})
grunt>
```

The **cogroup** operator groups the tuples from each relation according to age where each group depicts a particular age value.

For example, if we consider the 1st tuple of the result, it is grouped by age 21. And it contains two bags –

- the first bag holds all the tuples from the first relation (**student_details** in this case) having age 21, and
- the second bag contains all the tuples from the second relation

(**employee_details** in this case) having age 21.

In case a relation doesn't have tuples having the age value 21, it returns an empty bag.