

Bharathwin MA

205229105

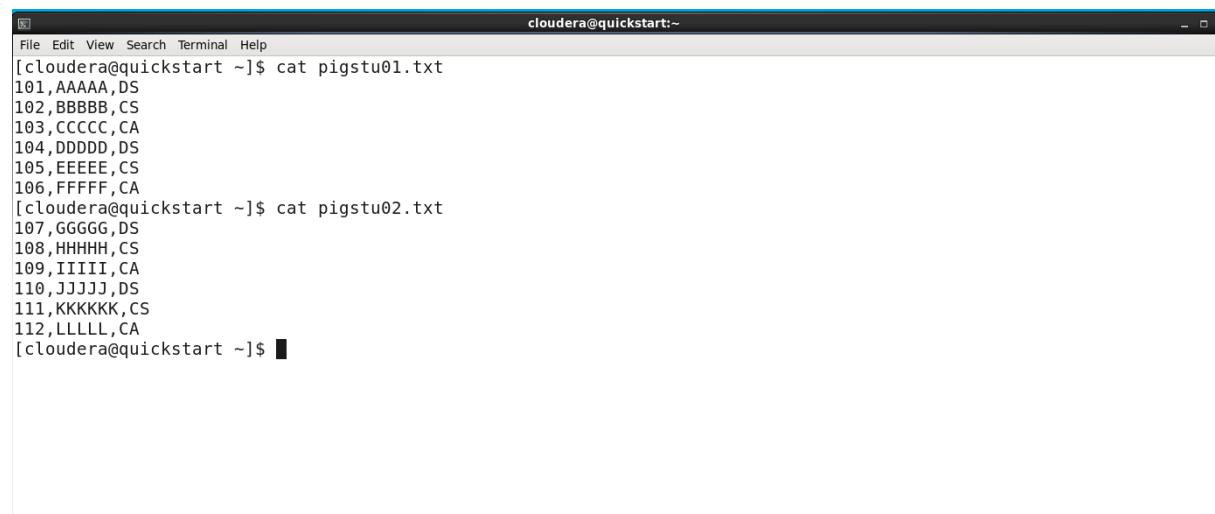
Exercise 08: Perform Sort, Group, Join, Split, and Filter Apache Pig Latin relational operations on a Student Data Set.

Here, we will be running Apache Pig Sample scripts using grunts. It is to just see the power of Apache Pig.

Step 1A: Start Grunt shell.

Open terminal and type pig

Step 1B,C: Create a file at /home/cloudera/pigstu01.txt with following content.



```
cloudera@quickstart:~$ cat pigstu01.txt
101,AAAAA,DS
102,BBBBB,CS
103,CCCCC,CA
104,DDDDD,DS
105,EEEE,CS
106,FFFF,CA
[cloudera@quickstart ~]$ cat pigstu02.txt
107,GGGG,DS
108,HHHH,CS
109,IIII,CA
110,JJJJ,DS
111,KKKKK,CS
112,LLLL,CA
[cloudera@quickstart ~]$
```

Step 2 a: Load the file stored in hadoop local with relation name 'stu1' and each line have to store in 'line' (comma separated file)

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
grunt> stu1 = LOAD '/home/cloudera/pigstu01.txt' USING PigStorage(',') as (stu_id:int,stu_name:chararray,stu_dep:chararray);  
grunt> dump stu1;  
2021-10-20 10:15:24,981 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN  
2021-10-20 10:15:24,983 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}  
2021-10-20 10:15:24,990 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false  
2021-10-20 10:15:24,992 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1  
2021-10-20 10:15:24,992 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1  
2021-10-20 10:15:24,993 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized  
2021-10-20 10:15:24,997 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job  
2021-10-20 10:15:25,011 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3  
2021-10-20 10:15:25,053 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
```

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
job_local158180278_0005  
  
2021-10-20 10:15:43,595 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!  
2021-10-20 10:15:43,596 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2021-10-20 10:15:43,596 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2021-10-20 10:15:43,596 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
2021-10-20 10:15:43,597 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized  
2021-10-20 10:15:43,634 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1  
2021-10-20 10:15:43,634 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1  
(101,AAAA,DS)  
(102,BBBBB,CS)  
(103,CCCC,CA)  
(104,DDDD,DS)  
(105,EEEE,CS)  
(106,FFFF,CA)  
grunt>
```

Step 2 b: Load the file stored in hadoop local with relation name 'stu2' and each line have to store in 'line' (comma separated file)

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
  
2021-10-20 09:29:08,253 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapRedu  
- Success!  
2021-10-20 09:29:08,296 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name i  
ed. Instead, use fs.defaultFS  
2021-10-20 09:29:08,297 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracke  
cated. Instead, use mapreduce.jobtracker.address  
2021-10-20 09:29:08,297 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.chec  
precated. Instead, use dfs.bytes-per-checksum  
2021-10-20 09:29:08,297 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has alre  
nitialized  
2021-10-20 09:29:08,478 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input  
rocess : 1  
2021-10-20 09:29:08,478 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Tot  
ths to process : 1  
(101,AAAAA,DS)  
(102,BBBBB,CS)  
(103,CCCC,CA)  
(104,DDDD,DS)  
(105,EEEE,CS)  
(106,FFFF,CA)  
grunt> stu2 = load '/home/cloudera/pigstu02.txt' as (line:chararray);  
grunt> dump stu2;
```

Step 3a: flatten the Stu_ID, Stu_Name,Stu_Department in each line from relation name 'stu1' and save separated words into relation name 'stu1foreach'

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
  
grunt> stu1foreach = foreach stu01 GENERATE FLATTEN(TOKENIZE(line',')) as word;  
2021-10-20 09:34:15,873 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: Pig script failed to  
<line 3, column 22> Undefined alias: stu01  
Details at logfile: /home/cloudera/pig 1634747218971.log  
grunt> stu1foreach = foreach stu1 GENERATE FLATTEN(TOKENIZE(line',')) as word;  
grunt> dump stu1foreach;  
2021-10-20 09:34:56,590 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the  
NKNOWN  
2021-10-20 09:34:56,599 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RUI  
=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSpli  
LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownFor  
n, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, F  
lterOptimizer]}  
2021-10-20 09:34:56,613 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRComp:  
concatenation threshold: 100 optimistic? false  
2021-10-20 09:34:56,615 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQ  
er - MR plan size before optimization: 1  
2021-10-20 09:34:56,615 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQ  
er - MR plan size after optimization: 1  
2021-10-20 09:34:56,617 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Me  
processName=JobTracker, sessionId= - already initialized  
2021-10-20 09:34:56,674 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are  
he job  
2021-10-20 09:34:56,700 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobConf
```

Step 3b: flatten the Stu_ID, Stu_Name,Stu_Department in each line from relation name 'stu2' and save separated words into relation name 'stu2foreach'

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
grunt> stu2foreach = foreach stu2 GENERATE FLATTEN(TOKENIZE(line, ',')) as word;  
grunt> dump stu2foreach;  
2021-10-20 09:43:44,404 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the  
UNKNOWN  
2021-10-20 09:43:44,408 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RUL  
=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplit  
LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownFor  
n, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, F  
lterOptimizer]}  
2021-10-20 09:43:44,416 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRComp  
concatenation threshold: 100 optimistic? false  
2021-10-20 09:43:44,419 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQu  
er - MR plan size before optimization: 1  
2021-10-20 09:43:44,419 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQu  
er - MR plan size after optimization: 1  
2021-10-20 09:43:44,421 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Met  
processName=JobTracker, sessionId= - already initialized  
2021-10-20 09:43:44,429 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are  
he job  
2021-10-20 09:43:44,444 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobCont  
r - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3  
2021-10-20 09:43:44,469 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobCont  
r - Setting up single store job  
2021-10-20 09:43:44,494 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is 1
```

Step 4a: Sort 'stu1foreach' data and save it into new relation name 'stu1sort'

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
  
2021-10-20 10:28:22,328 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher  
- Success!  
2021-10-20 10:28:22,329 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecate  
ed. Instead, use fs.defaultFS  
2021-10-20 10:28:22,329 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is depre  
cated. Instead, use mapreduce.jobtracker.address  
2021-10-20 10:28:22,329 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is de  
precated. Instead, use dfs.bytes-per-checksum  
2021-10-20 10:28:22,329 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been i  
nitialized  
2021-10-20 10:28:22,341 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to p  
rocess : 1  
2021-10-20 10:28:22,341 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input pa  
ths to process : 1  
(106,FFFFFF,CA)  
(105,EEEEEE,CS)  
(104,DDDDD,DS)  
(103,CCCC,CA)  
(102,BBBBB,CS)  
(101,AAAA,DS)  
grunt> stu2sort = ORDER stu2 BY stu_id DESC;  
grunt> dump stu2sort;
```

Step 4b: Sort 'stuforeach' data and save it into new relation name 'stu1sort'

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
job_local2063725508_0012  
  
2021-10-20 10:35:09,947 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher  
- Success!  
2021-10-20 10:35:09,961 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecate  
ed. Instead, use fs.defaultFS  
2021-10-20 10:35:09,961 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is depre  
cated. Instead, use mapreduce.jobtracker.address  
2021-10-20 10:35:09,962 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is de  
precated. Instead, use dfs.bytes-per-checksum  
2021-10-20 10:35:09,968 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been i  
nitialized  
2021-10-20 10:35:10,117 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to p  
rocess : 1  
2021-10-20 10:35:10,117 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input pa  
ths to process : 1  
(112,LLLLL,CA)  
(111,KKKKKK,CS)  
(110,JJJJ,DS)  
(109,IIIII,CA)  
(108,HHHH,CS)  
(107,GGGG,DS)  
grunt>
```

Step 5: Join relation 'stu1', relation 'stu2' and create relation name 'stujoin'

grunt> stujoin== UNION stu1 , stu2 ;

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
2021-10-20 10:35:09,947 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher  
- Success!  
2021-10-20 10:35:09,961 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecate  
ed. Instead, use fs.defaultFS  
2021-10-20 10:35:09,961 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is depre  
cated. Instead, use mapreduce.jobtracker.address  
2021-10-20 10:35:09,962 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is de  
precated. Instead, use dfs.bytes-per-checksum  
2021-10-20 10:35:09,968 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been i  
nitialized  
2021-10-20 10:35:10,117 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to p  
rocess : 1  
2021-10-20 10:35:10,117 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input pa  
ths to process : 1  
(112,LLLLL,CA)  
(111,KKKKKK,CS)  
(110,JJJJJ,DS)  
(109,IIIII,CA)  
(108,HHHHH,CS)  
(107,GGGGG,DS)  
grunt> stujoin = UNION stu1,stu2;  
grunt> dump stujoin;
```

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
ed. Instead, use fs.defaultFS  
2021-10-20 10:37:55,104 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is depre  
cated. Instead, use mapreduce.jobtracker.address  
2021-10-20 10:37:55,104 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is de  
precated. Instead, use dfs.bytes-per-checksum  
2021-10-20 10:37:55,104 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been i  
nitialized  
2021-10-20 10:37:55,179 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to p  
rocess : 2  
2021-10-20 10:37:55,179 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input pa  
ths to process : 2  
(101,AAAAA,DS)  
(102,BBBBBB,CS)  
(103,CCCCC,CA)  
(104,DDDDD,DS)  
(105,EEEEEE,CS)  
(106,FFFFFF,CA)  
(107,GGGGG,DS)  
(108,HHHHH,CS)  
(109,IIIII,CA)  
(110,JJJJJ,DS)  
(111,KKKKKK,CS)  
(112,LLLLL,CA)  
grunt>
```

Step 6: Split relation 'stujoin' as relation name 'studs' who all are belongs to 'DS'? And create relation name 'stucs' who all are belongs to 'CS'?

grunt> SPLIT student_details into studs if (Department=='DS'), stucs if (Department=='CS');

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
2021-10-20 10:37:55,104 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2021-10-20 10:37:55,104 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
2021-10-20 10:37:55,104 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized  
2021-10-20 10:37:55,179 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 2  
2021-10-20 10:37:55,179 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 2  
(101,AAAAA,DS)  
(102,BBBBB,CS)  
(103,CCCCC,CA)  
(104,DDDDD,DS)  
(105,EEEE,CS)  
(106,FFFFF,CA)  
(107,GGGGG,DS)  
(108,HHHHH,CS)  
(109,IIIII,CA)  
(110,JJJJJ,DS)  
(111,KKKKK,CS)  
(112,LLLLL,CA)  
grunt> SPLIT stujoin into studs if (stu_dep=='DS'), stucs if (stu_dep=='CS');  
grunt> dump studs;
```

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
  
Job DAG:  
job_local61563744_0014  
  
2021-10-20 10:40:39,328 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!  
2021-10-20 10:40:39,332 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2021-10-20 10:40:39,332 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2021-10-20 10:40:39,332 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum  
2021-10-20 10:40:39,333 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized  
2021-10-20 10:40:39,410 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 2  
2021-10-20 10:40:39,410 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 2  
(101,AAAAA,DS)  
(104,DDDDD,DS)  
(107,GGGGG,DS)  
(110,JJJJJ,DS)  
grunt>
```

Step 7: Filter relation 'stujoin' as relation name 'stufilter' who all are belongs to 'DS'?

grunt> stufilter = filter data by Department == 'DS';

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Job DAG:  
job_local61563744_0014  
  
2021-10-20 10:40:39,328 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher  
- Success!  
2021-10-20 10:40:39,332 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecate  
ed. Instead, use fs.defaultFS  
2021-10-20 10:40:39,332 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is depre  
cated. Instead, use mapreduce.jobtracker.address  
2021-10-20 10:40:39,332 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is de  
precated. Instead, use dfs.bytes-per-checksum  
2021-10-20 10:40:39,333 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been i  
nitialized  
2021-10-20 10:40:39,410 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to p  
rocess : 2  
2021-10-20 10:40:39,410 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input pa  
ths to process : 2  
(101,AAAAA,DS)  
(104,DDDDD,DS)  
(107,GGGGG,DS)  
(110,JJJJJ,DS)  
grunt> stufilter = FILTER stujoin by stu_dep=='DS';  
grunt> dump stufilter;
```

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
Job DAG:  
job_local1727104242_0015  
  
2021-10-20 10:43:02,195 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher  
- Success!  
2021-10-20 10:43:02,198 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecate  
ed. Instead, use fs.defaultFS  
2021-10-20 10:43:02,198 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is depre  
cated. Instead, use mapreduce.jobtracker.address  
2021-10-20 10:43:02,199 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is de  
precated. Instead, use dfs.bytes-per-checksum  
2021-10-20 10:43:02,199 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been i  
nitialized  
2021-10-20 10:43:02,355 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to p  
rocess : 2  
2021-10-20 10:43:02,355 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input pa  
ths to process : 2  
(101,AAAAA,DS)  
(104,DDDDD,DS)  
(107,GGGGG,DS)  
(110,JJJJJ,DS)  
grunt>
```