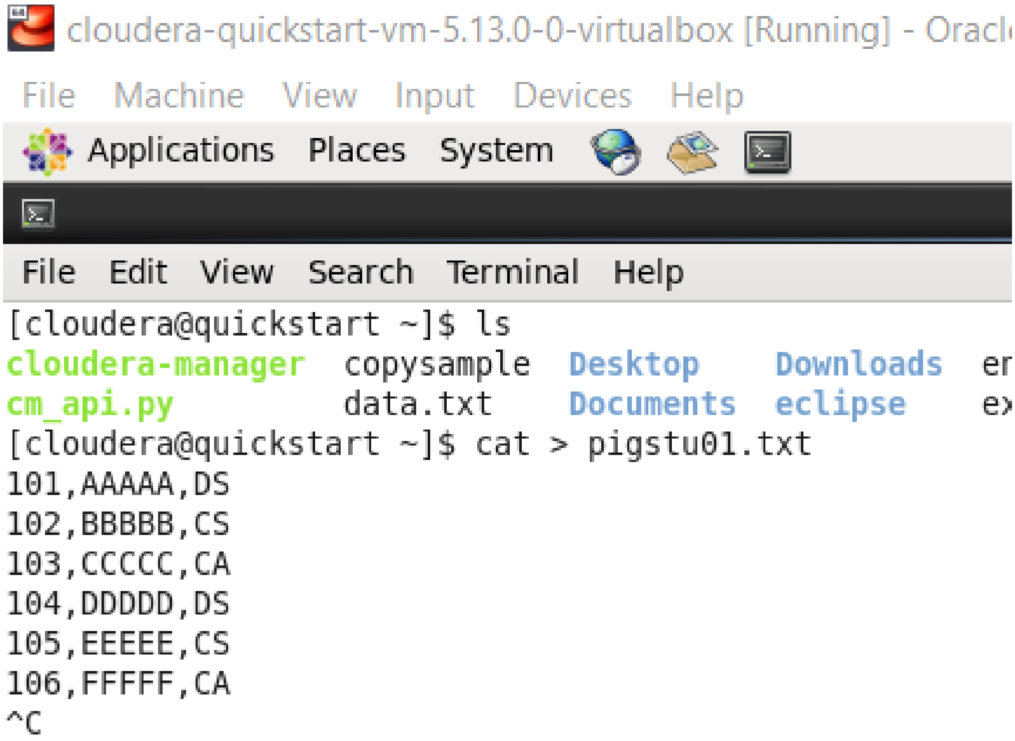
Exercise 05: Perform Sort, Group, Join, Split, and Filter Apache Pig Latin relational operations on a Student Data Set.

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*

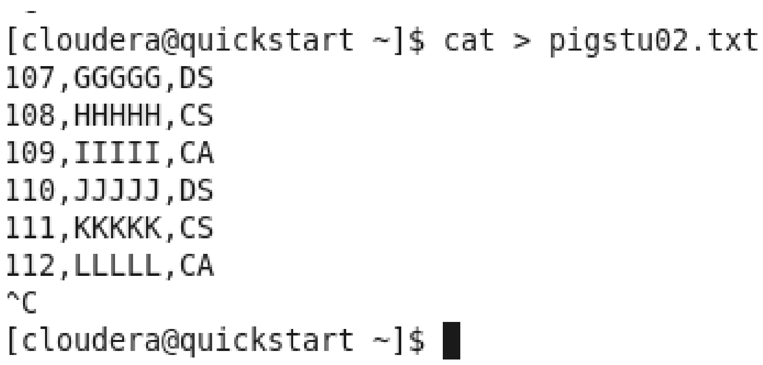
Step 1B: Create a file at /home/cloudera/pigstu01.txtwithfollowing content.


cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracl
File Machine View Input Devices Help
Applications Places System
File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ ls
cloudera-manager  copysample  Desktop    Downloads  er
cm_api.py         data.txt    Documents  eclipse    e>
[cloudera@quickstart ~]$ cat > pigstu01.txt
101,AAAAA,DS
102,BBBBB,CS
103,CCCCC,CA
104,DDDDD,DS
105,EEEEE,CS
106,FFFFF,CA
^C
```

101, AAAAA, DS
102,BBBBB,CS
103,CCCCC,CA
104,DDDDD,DS
105,EEEEE,CS
106,FFFFF,CA

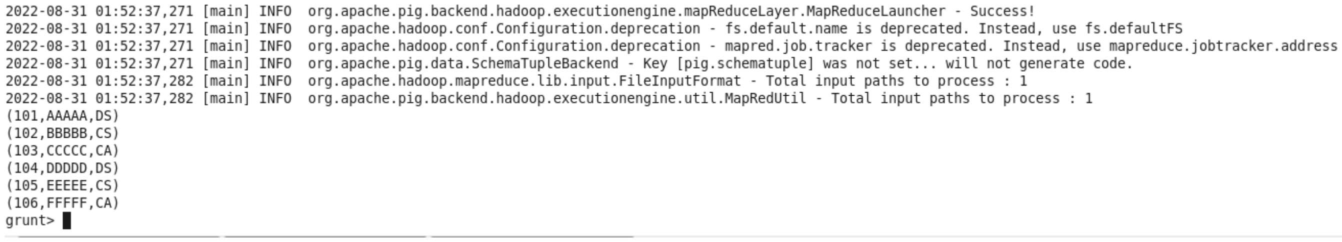
Step 1C: Create a file at /home/cloudera/pigstu02.txtwithfollowing content.



```
[cloudera@quickstart ~]$ cat > pigstu02.txt
107,GGGGG,DS
108,HHHHH,CS
109,IIIII,CA
110,JJJJJ,DS
111,KKKKK,CS
112,LLLLL,CA
^C
[cloudera@quickstart ~]$ 
```

```
[cloudera@quickstart ~]$cat > pigstu02.txt
107,GGGGG,DS
108,HHHHH,CS
109,IIIII,CA
110,JJJJJ,DS
111,KKKKKK,CS
112,LLLLL,CA
^ctrlC
[cloudera@quickstart ~]$cat pigstu02.txt
```

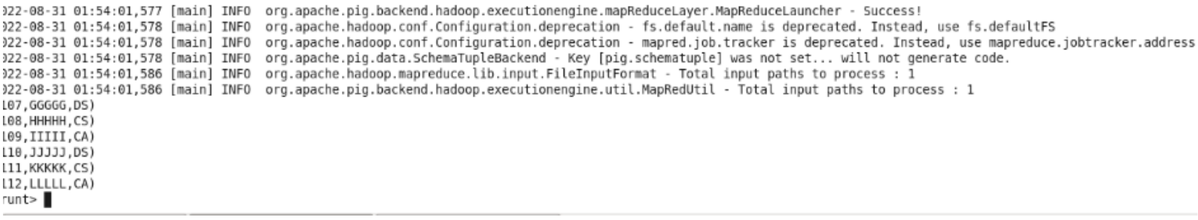
Step 2 a: Load the file stored in hadoop local with relation name ‘stu1’ and each line have to store in ‘line’ (comma separated file)



```
2022-08-31 01:52:37,271 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-31 01:52:37,271 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-31 01:52:37,271 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-31 01:52:37,271 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-31 01:52:37,282 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-08-31 01:52:37,282 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101,AAAAA,DS)
(102,BBBBB,CS)
(103,CCCCC,CA)
(104,DDDDD,DS)
(105,EEEEE,CS)
(106,FFFFF,CA)
grunt> 
```

(101)(AAAAA)(DS)
(102)(BBBBB)(CS)
(103)(CCCCC)(CA)
(104)(DDDDD)(DS)
(10)(,EEEEE)(CS)
(106)(FFFFF)(CA)

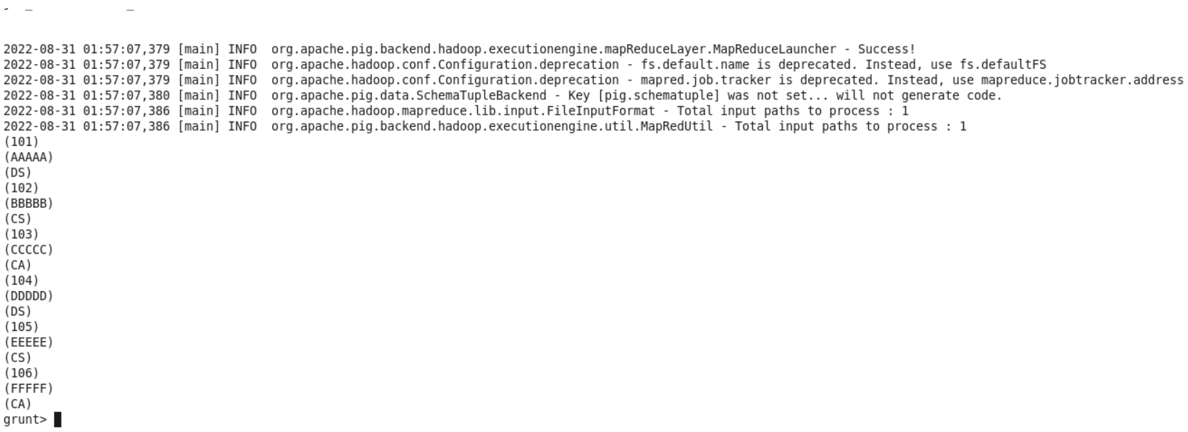
Step 2 a: Load the file stored in hadoop local with relation name ‘stu2’ and each line have to store in ‘line’ (commaseparated file)



```
)22-08-31 01:54:01,577 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
)22-08-31 01:54:01,578 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
)22-08-31 01:54:01,578 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
)22-08-31 01:54:01,578 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
)22-08-31 01:54:01,586 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
)22-08-31 01:54:01,586 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
l07,GGGGG,DS)
l08,HHHHH,CS)
l09,IIIII,CA)
l10,JJJJJ,DS)
l11,KKKKK,CS)
l12,LLLLL,CA)
runt> 
```

(107)(GGGGG)(DS)
(108)(HHHHH)(CS)
(109)(IIIII)(CA)
(110)(JJJJJ)(DS)
(111)(KKKKKK)(CS)
(112)(LLLLL)(CA)

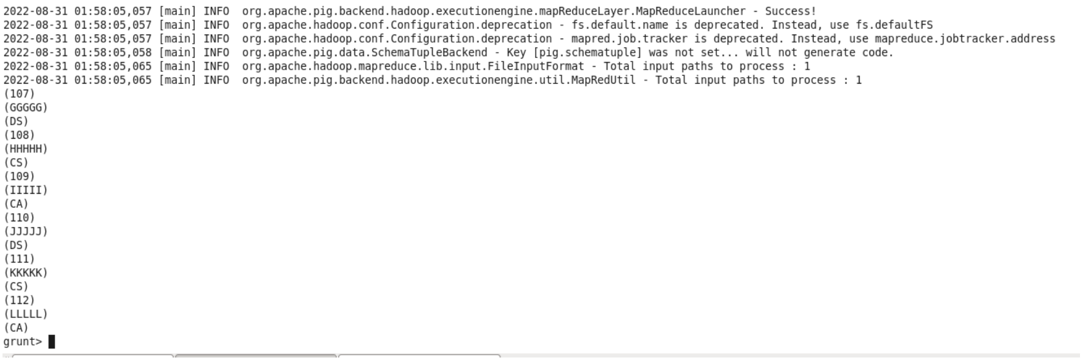
Step 3a: flatten the Stu_ID, Stu_Name,Stu_Department in each line from relation name 'stu1' and save separated words into relation name 'stu1foreach'



```
2022-08-31 01:57:07,379 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-31 01:57:07,379 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-31 01:57:07,379 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-31 01:57:07,380 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-31 01:57:07,386 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-08-31 01:57:07,386 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(101)
(AAAAA)
(DS)
(102)
(BBBBB)
(CS)
(103)
(CCCCC)
(CA)
(104)
(DDDDD)
(DS)
(105)
(EEEEE)
(CS)
(106)
(FFFFF)
(CA)
grunt>
```

(101, AAAAA, DS)
(102,BBBBB,CS)
(103,CCCCC,CA)
(104,DDDDD,DS)
(105,EEEEE,CS)
(106,FFFFF,CA)

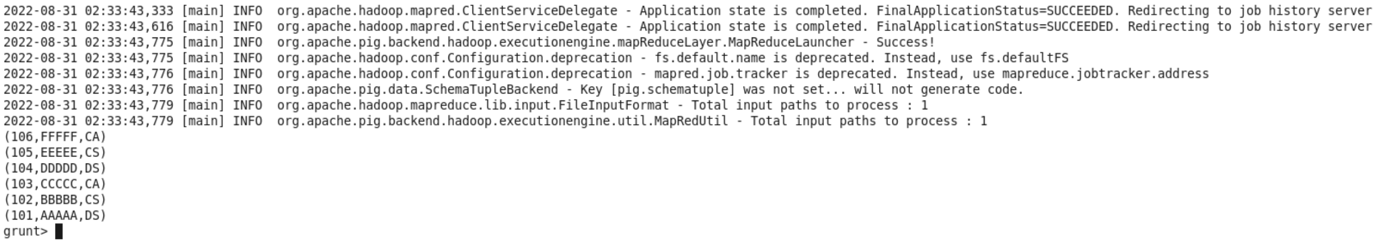
Step 3b: flatten the Stu_ID, Stu_Name,Stu_Department in each line from relation name 'stu1' and save separated words into relation name 'stu1foreach'



```
2022-08-31 01:58:05,057 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-31 01:58:05,057 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-31 01:58:05,057 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-31 01:58:05,058 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-31 01:58:05,065 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-08-31 01:58:05,065 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(107)
(GGGGG)
(DS)
(108)
(HHHHH)
(CS)
(109)
(IIIII)
(CA)
(110)
(JJJJJ)
(DS)
(111)
(KKKKK)
(CS)
(112)
(LLLLL)
(CA)
grunt>
```

(107,GGGGG,DS)
(108,HHHHH,CS)
(109,IIIII,CA)
(110,JJJJJ,DS)
(111,KKKKKK,CS)
(112,LLLLL,CA)

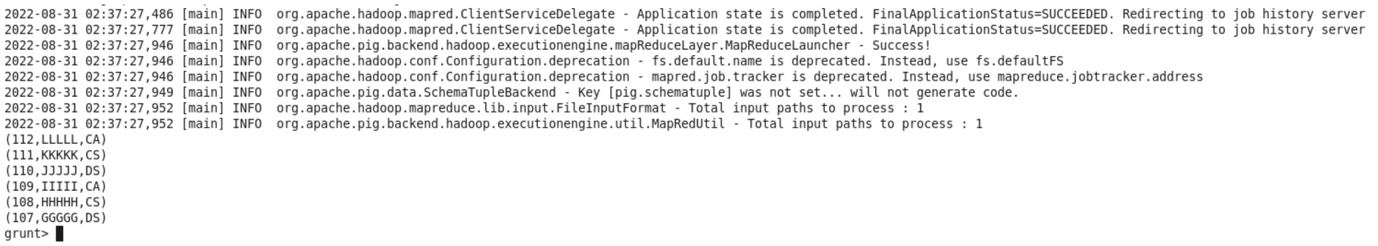
Step 4a: Sort ‘stu1foreach’ data and save it into new relation name ‘stu1sort’



```
2022-08-31 02:33:43,333 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-31 02:33:43,616 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-31 02:33:43,775 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-31 02:33:43,775 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-31 02:33:43,776 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-31 02:33:43,776 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-31 02:33:43,779 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-08-31 02:33:43,779 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(106,FFFFF,CA)
(105,EEEEE,CS)
(104,DDDDD,DS)
(103,CCCCC,CA)
(102,BBBBB,CS)
(101,AAAAA,DS)
grunt> 
```

(106,FFFFF,CA)
(105,EEEEE,CS)
(104,DDDDD,DS)
(103,CCCCC,CA)
(102,BBBBB,CS)
(101, AAAAA, DS)

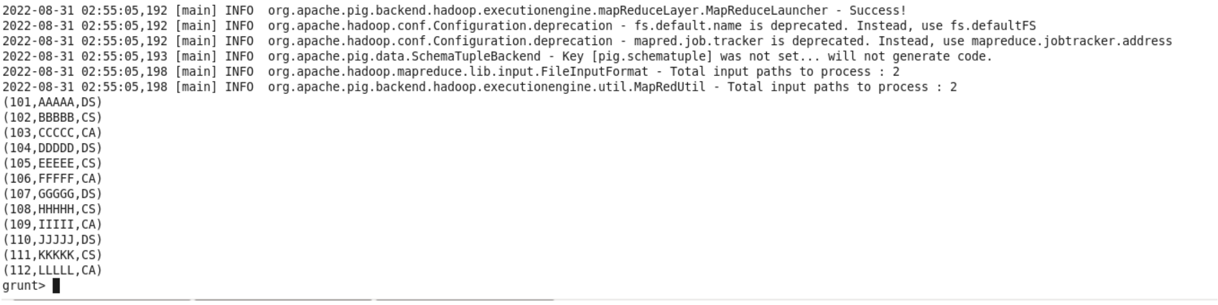
Step 4b: Sort ‘stuforeach’ data and save it into new relation name ‘stu1sort’



```
2022-08-31 02:37:27,486 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-31 02:37:27,777 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-31 02:37:27,946 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-31 02:37:27,946 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-31 02:37:27,946 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-31 02:37:27,949 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-31 02:37:27,952 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-08-31 02:37:27,952 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(112,LLLLL,CA)
(111,KKKKK,CS)
(110,JJJJJ,DS)
(109,IIIII,CA)
(108,HHHHH,CS)
(107,GGGGG,DS)
grunt> 
```

(112,LLLLL,CA)
(111,KKKKKK,CS)
(110,JJJJJ,DS)
(109,IIIII,CA)
(108,HHHHH,CS)
(107,GGGGG,DS)

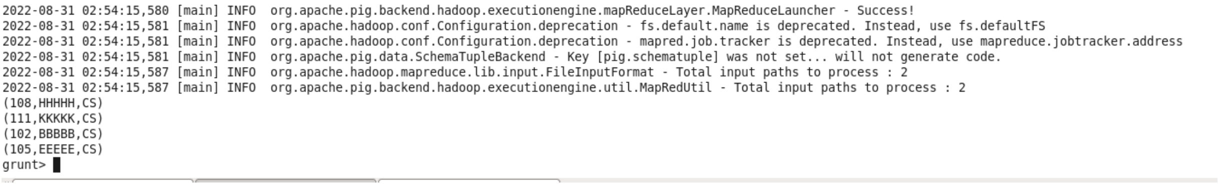
Step 5: Join relation ‘stu1’, relation ‘stu2’ and create relation name ‘stujoin’



```
2022-08-31 02:55:05,192 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-31 02:55:05,192 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-31 02:55:05,192 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-31 02:55:05,193 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-31 02:55:05,198 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 2
2022-08-31 02:55:05,198 [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,EEEEE,CS)
(106,FFFFF,CA)
(107,GGGGG,DS)
(108,HHHHH,CS)
(109,IIIII,CA)
(110,JJJJJ,DS)
(111,KKKKK,CS)
(112,LLLLL,CA)
grunt> █
```

```
grunt> stujoin== UNION stu1 , stu2 ;
```

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’?

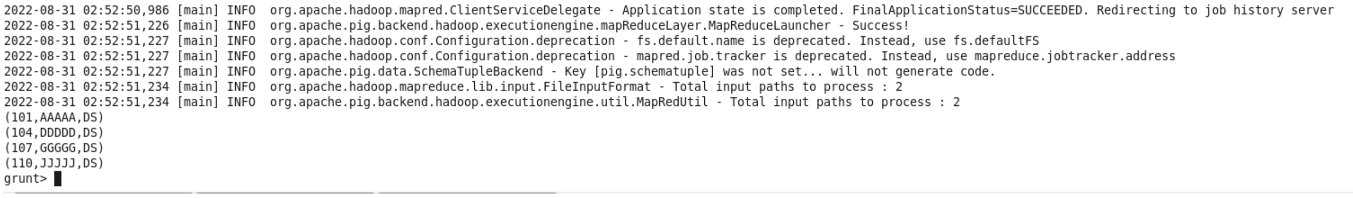


```
2022-08-31 02:54:15,580 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-31 02:54:15,581 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-31 02:54:15,581 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-31 02:54:15,581 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-31 02:54:15,587 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 2
2022-08-31 02:54:15,587 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 2
(108,HHHHH,CS)
(111,KKKKK,CS)
(102,BBBBB,CS)
(105,EEEEE,CS)
grunt> █
```

```
grunt> SPLIT student_details into studs if (Department==’DS’),
stucs  if (Department==’DS’);
```

Step 7:

Filter relation ‘stujoin’ as relation name ‘stufilter’ who all are belongs to ‘DS’?



```
2022-08-31 02:52:50,986 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2022-08-31 02:52:51,226 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-08-31 02:52:51,227 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-08-31 02:52:51,227 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-08-31 02:52:51,227 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-08-31 02:52:51,234 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 2
2022-08-31 02:52:51,234 [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> █
```

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