Below is the Pig Latin Script that has been used to generate word count for the input file sample_temperature_dataset.csv which contains temperature for various years-

LoadFile = LOAD '/home/acadgild/hadoop/sample_temperature_dataset.csv' USING PigStorage(',') AS (full_date:chararray, zip:int, temp:int);

selRel = FOREACH LoadFile GENERATE SUBSTRING(full_date,0,4) AS year, temp;

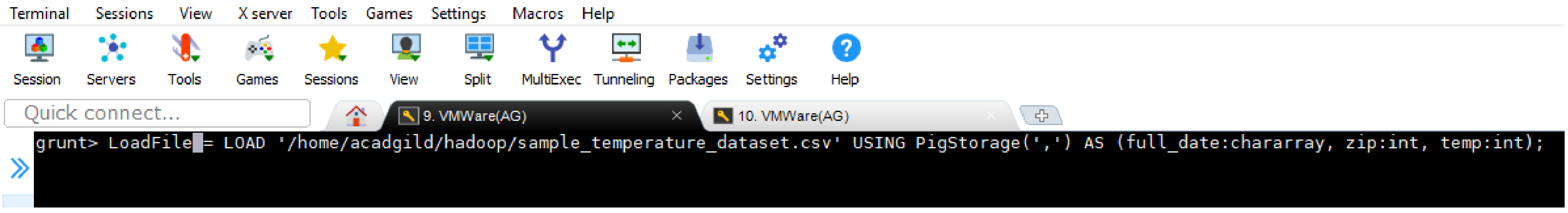
flattokenRel = FOREACH selRel GENERATE FLATTEN(TOKENIZE(year)) AS year;

GroupYear = GROUP flattokenRel BY year;

YearCount = FOREACH GroupYear GENERATE group, COUNT(flattokenRel);

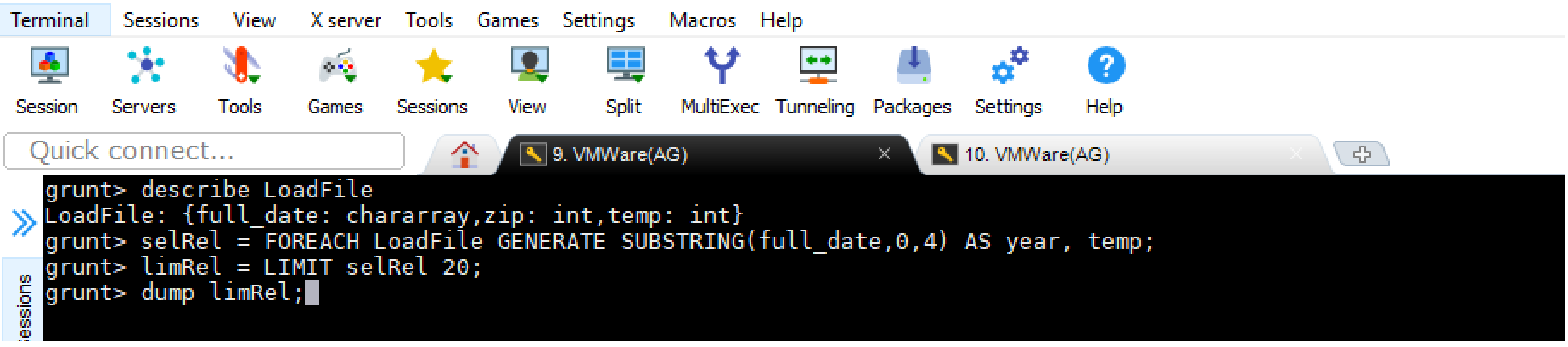
The above script is described below part by part using "dump" also in between to show the intermediate results-

1. Load the file-



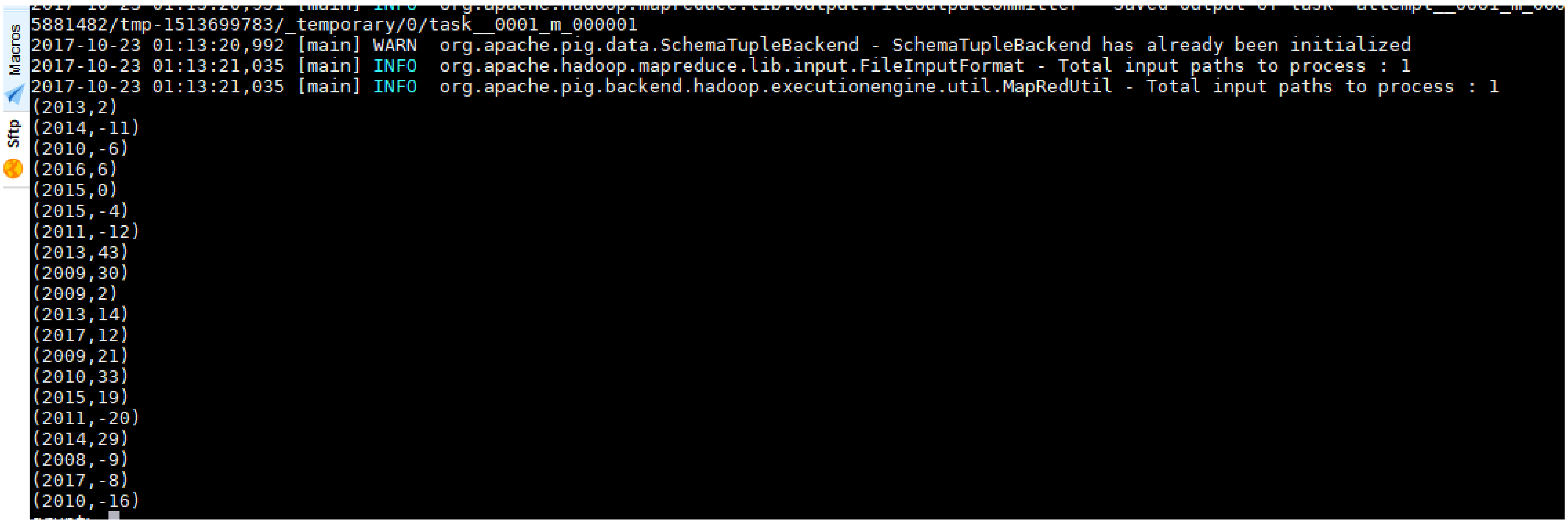
```
grunt> LoadFile = LOAD '/home/acadgild/hadoop/sample_temperature_dataset.csv' USING PigStorage(',') AS (full_date:chararray, zip:int, temp:int);
```

2. Generate the substring to extract only year in YYMM format-



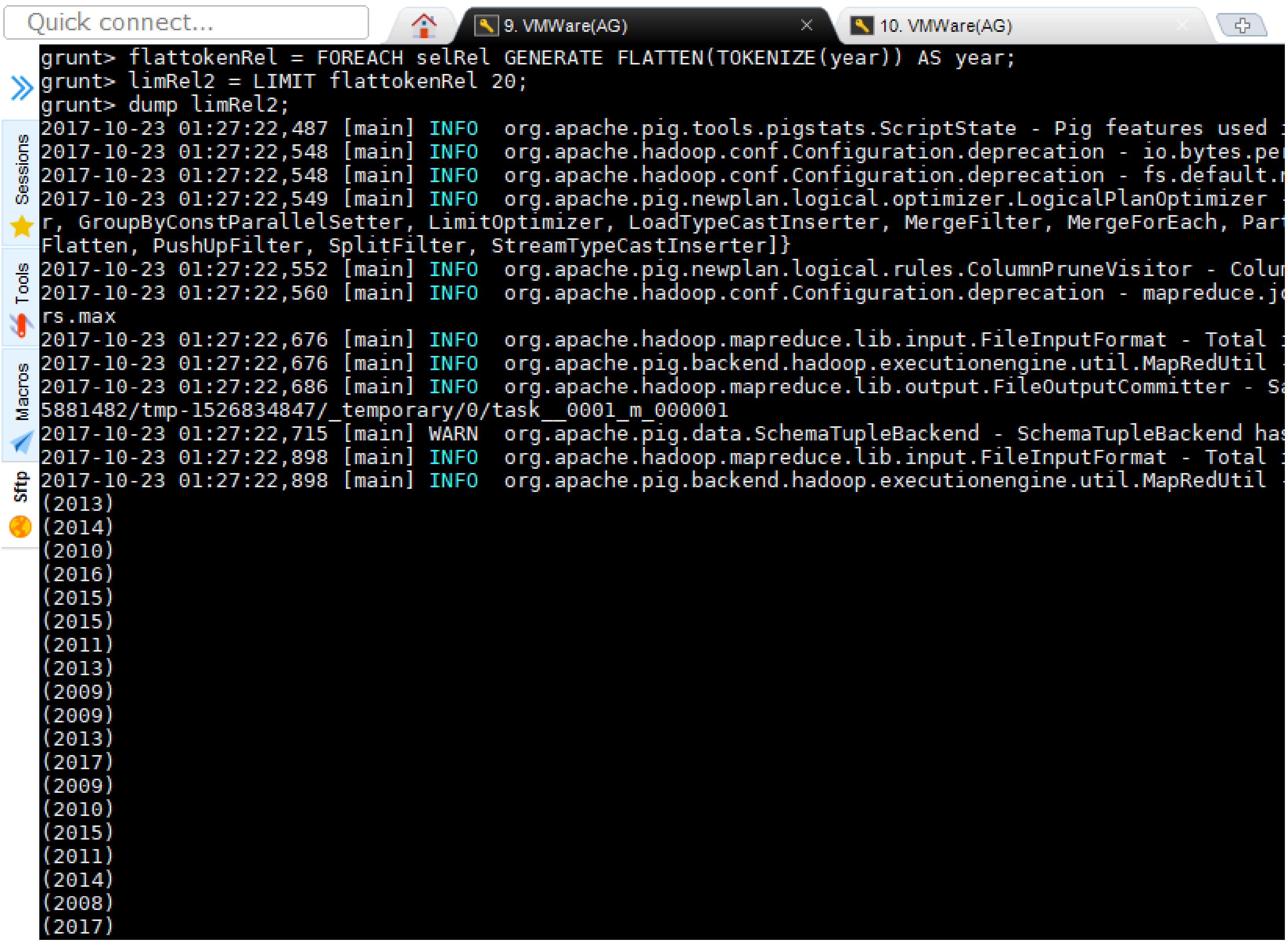
```
grunt> describe LoadFile
LoadFile: {full_date: chararray,zip: int,temp: int}
grunt> selRel = FOREACH LoadFile GENERATE SUBSTRING(full_date,0,4) AS year, temp;
grunt> limRel = LIMIT selRel 20;
grunt> dump limRel;
```

Results of above relation-



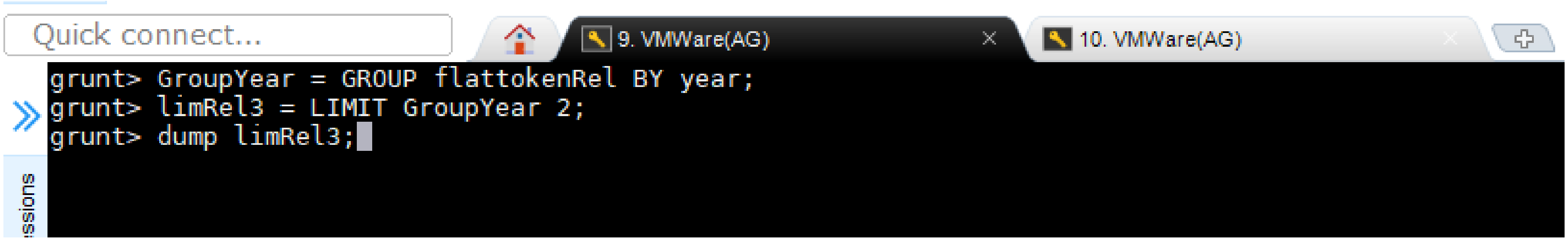
```
5881482/tmp-1513699783/_temporary/0/task__0001_m_000001
2017-10-23 01:13:20,992 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-10-23 01:13:21,035 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-10-23 01:13:21,035 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(2013,2)
(2014,-11)
(2010,-6)
(2016,6)
(2015,0)
(2015,-4)
(2011,-12)
(2013,43)
(2009,30)
(2009,2)
(2013,14)
(2017,12)
(2009,21)
(2010,33)
(2015,19)
(2011,-20)
(2014,29)
(2008,-9)
(2017,-8)
(2010,-16)
```

3. Tokenize the year field and flatten it to generate list of year appearing.

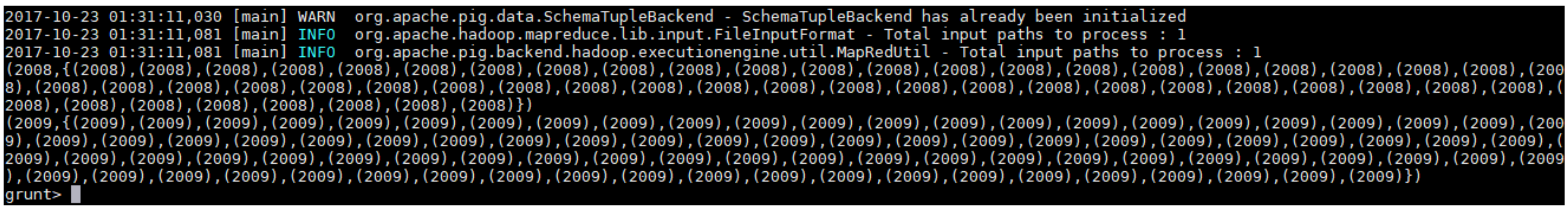


```
grunt> flattokenRel = FOREACH selRel GENERATE FLATTEN(TOKENIZE(year)) AS year;
grunt> limRel2 = LIMIT flattokenRel 20;
grunt> dump limRel2;
2017-10-23 01:27:22,487 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used i
2017-10-23 01:27:22,548 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.pe
2017-10-23 01:27:22,548 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.n
2017-10-23 01:27:22,549 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer -
r, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, Part
Flatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-10-23 01:27:22,552 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Colu
2017-10-23 01:27:22,560 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapreduce.jo
rs.max
2017-10-23 01:27:22,676 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total i
2017-10-23 01:27:22,676 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
2017-10-23 01:27:22,686 [main] INFO  org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Sa
5881482/tmp-1526834847/_temporary/0/task__0001_m_000001
2017-10-23 01:27:22,715 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has
2017-10-23 01:27:22,898 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total i
2017-10-23 01:27:22,898 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil -
(2013)
(2014)
(2010)
(2016)
(2015)
(2015)
(2011)
(2013)
(2009)
(2009)
(2013)
(2017)
(2009)
(2010)
(2015)
(2011)
(2014)
(2008)
(2017)
```

4. GROUP above relation by year to generate grouped data for years-

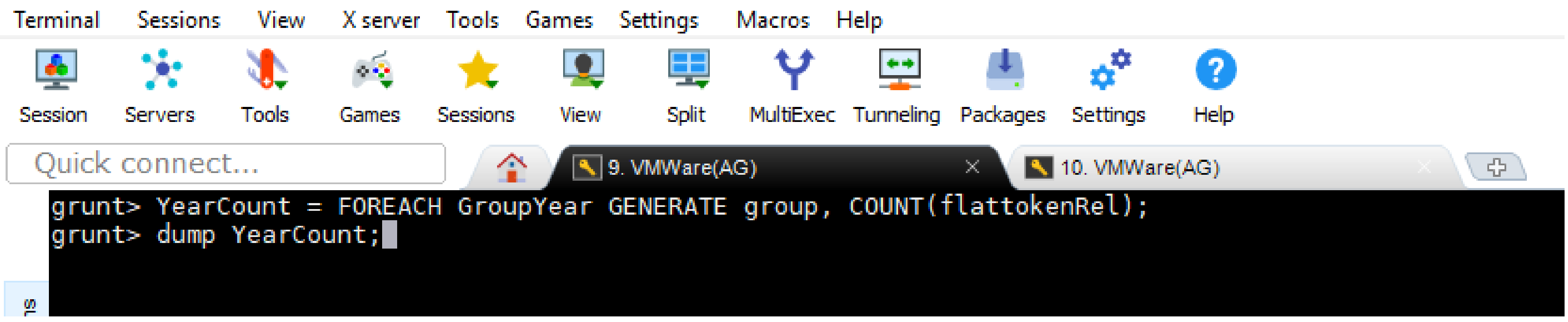


```
grunt> GroupYear = GROUP flattokenRel BY year;
grunt> limRel3 = LIMIT GroupYear 2;
grunt> dump limRel3;
```



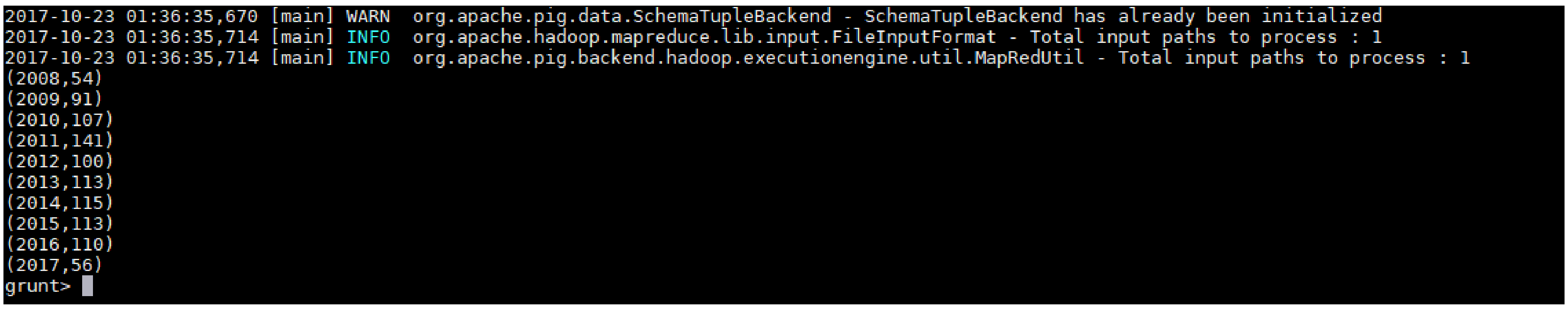
```
2017-10-23 01:31:11,030 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-10-23 01:31:11,081 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-10-23 01:31:11,081 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(2008,{(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(200
8),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008),(
2008),(2008),(2008),(2008),(2008),(2008),(2008),(2008)})
(2009,{(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(200
9),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(
2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009
),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009),(2009)})
grunt>
```

5. Calculate year count on the previous relation and include corresponding year by generate group-



```
grunt> YearCount = FOREACH GroupYear GENERATE group, COUNT(flattokenRel);
grunt> dump YearCount;
```

6. O/P-



```
2017-10-23 01:36:35,670 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-10-23 01:36:35,714 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-10-23 01:36:35,714 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(2008,54)
(2009,91)
(2010,107)
(2011,141)
(2012,100)
(2013,113)
(2014,115)
(2015,113)
(2016,110)
(2017,56)
grunt>
```