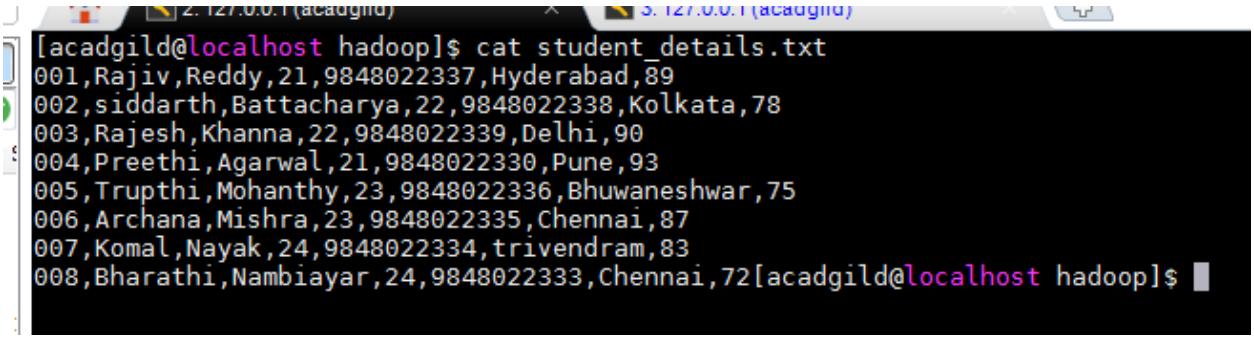
Assignment 4.2

Objective:-

Create a sample dataset and implement the below Pig Commands on the same dataset.

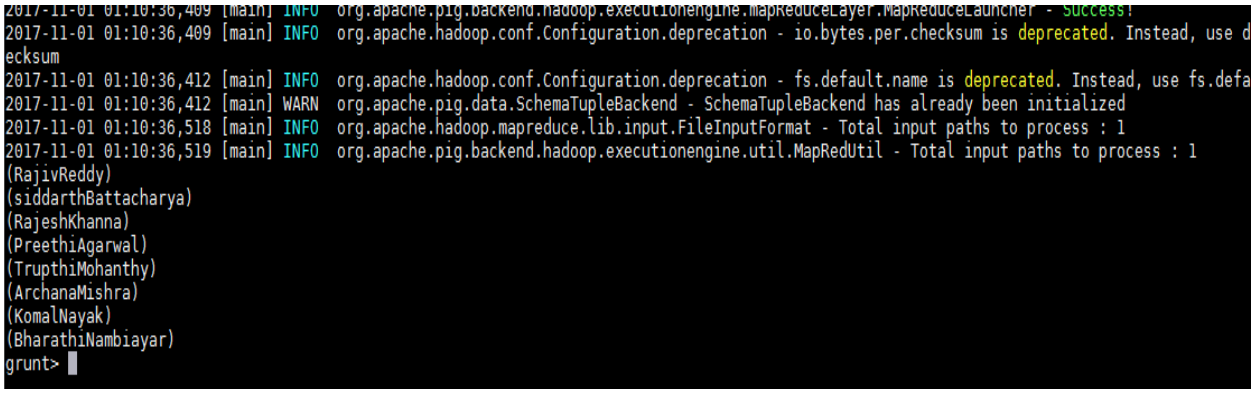
Sample Data Set:-



1. CONCAT:-

The CONCAT() function is used to concatenate two or more expressions of the same type.

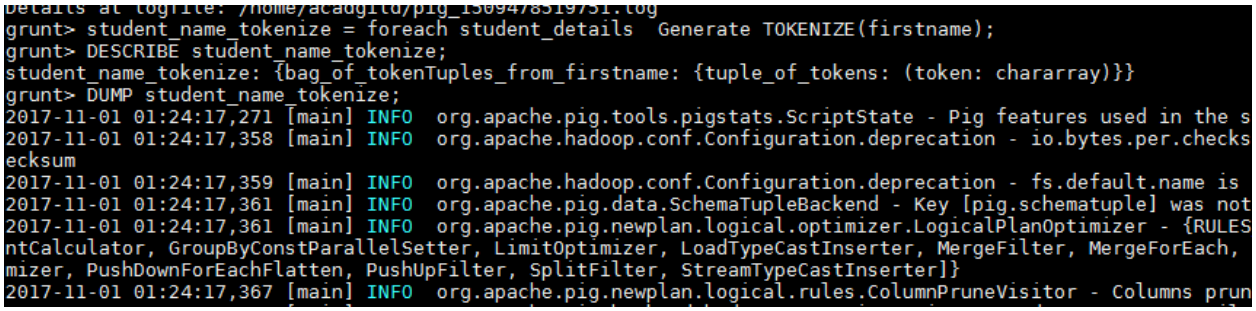
```
2017-11-01 01:06:34,350 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with pr
ready initialized
2017-11-01 01:06:34,358 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with pr
ready initialized
2017-11-01 01:06:34,366 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with pr
ready initialized
2017-11-01 01:06:34,389 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
2017-11-01 01:06:34,398 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is depr
ecksum
2017-11-01 01:06:34,399 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated
2017-11-01 01:06:34,401 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been ini
2017-11-01 01:06:34,566 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to pro
2017-11-01 01:06:34,566 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input path
(1,Rajiv,Reddy,21,9848022337,Hyderabad,89)
(2,siddarth,Battacharya,22,9848022338,Kolkata,78)
(3,Rajesh,Khanna,22,9848022339,Delhi,90)
(4,Preethi,Agarwal,21,9848022330,Pune,93)
(5,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar,75)
(6,Archana,Mishra,23,9848022335,Chennai,87)
(7,Komal,Nayak,24,9848022334,trivendram,83)
(8,Bharathi,Nambiayar,24,9848022333,Chennai,72)
grunt>
grunt> student_name_concat = foreach student_details Generate CONCAT (firstname, lastname);
grunt> DESCRIBE student_name_concat
student_name_concat: {chararray}
grunt> DUMP student_name_concat;
2017-11-01 01:10:34,851 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNK
2017-11-01 01:10:34,935 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is depr
```

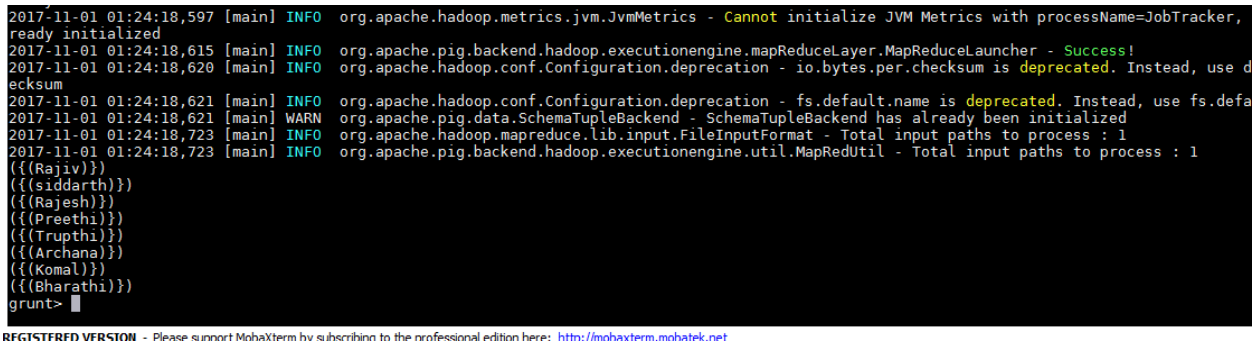


The above screen both first-name and last-name are concatenated together.

2. TOKENIZE:-

The TOKENIZE() function is used to split a string (which contains a group of words) in a single tuple and returns a bag which contains the output of the split operation.





3. SUM:-

SUM to get the total of the numeric values of a column in a single-column bag.

For calculating SUM of GPA,we first need to group the data using GROUP ALL function.

```
grunt> student_group = Group student_details all;
grunt> DESCRIBE student_group;
student_group: {group: chararray,student_details: {(id: int,firstname: chararray,lastname: chararray,age: int,phone: chararray,city: chararray,gpa: int)}}
grunt> DUMP student_group;
2017-11-01 01:35:59,134 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2017-11-01 01:35:59,292 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-ch
ecksum
```

```
2017-11-01 01:36:04,117 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(all,{(8,Bharathi,Nambiayar,24,9848022333,Chennai,72),(7,Komal,Nayak,24,9848022334,trivendram,83),(6,Archana,Mishra,23,9848022335,Chennai,87),(5,Trupthi,M
ohanthy,23,9848022336,Bhuwaneshwar,75),(4,Preethi,Agarwal,21,9848022330,Pune,93),(3,Rajesh,Khanna,22,9848022339,Delhi,90),(2,siddarth,Battacharya,22,98480
22338,Kolkata,78),(1,Rajiv,Reddy,21,9848022337,Hyderabad,89)})
grunt>
```

Now we can calculate the SUM of GPA as

```
grunt>
grunt>
grunt>
grunt>
grunt>
grunt>
grunt> student_gpa_sum = foreach student_group Generate (student_details.firstname,student_details.gpa),SUM(student_details.gpa);
grunt> DESCRIBE student_gpa_sum
student_gpa_sum: {org.apache.pig.builtin.totuple_21: ({(firstname: chararray)},{(gpa: int)}),long}
grunt> DUMP student_gpa_sum
```

```
2017-11-01 01:47:09,971 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-01 01:47:10,100 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-01 01:47:10,100 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(({(Bharathi),(Komal),(Archana),(Trupthi),(Preethi),(Rajesh),(siddarth),(Rajiv)},{(72),(83),(87),(75),(93),(90),(78),(89)}),667)
grunt>
grunt>
grunt>
grunt>
grunt>
```

Output is in the form as List of firstname,List of GPA,Sum of GPA.

4. MIN:-

The MIN is used to get the minimum (lowest) value (numeric or chararray) for a certain column in a single-column bag.

```
grunt>
grunt>
grunt>
grunt>
grunt>
grunt> student_gpa_min = foreach student_group  Generate (student_details.firstname, student_details.gpa), MIN(student_details.gpa);
grunt> DESCRIBE student_gpa_min
student_gpa_min: {org.apache.pig.builtin.totuple_49: ({(firstname: chararray)},{(gpa: int)}),int}
grunt> DUMP student_gpa_min
```

```
2017-11-01 01:53:43,566 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaul
2017-11-01 01:53:43,567 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-01 01:53:43,671 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-01 01:53:43,671 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(({(Bharathi),(Komal),(Archana),(Trupthi),(Preethi),(Rajesh),(siddarth),(Rajiv)},{(72),(83),(87),(75),(93),(90),(78),(89)}),72)
grunt>
```

5. MAX:-

MAX is used to calculate the highest value for a column (numeric values or chararrays) in a single-column bag.

```
grunt>
grunt> student_gpa_max = foreach student_group  Generate (student_details.firstname, student_details.gpa), MIN(student_details.gpa);
grunt> DESCRIBE student_gpa_max
student_gpa_max: {org.apache.pig.builtin.totuple_88: ({(firstname: chararray)},{(gpa: int)}),int}
grunt> DUMP student_gpa_max
```

```
2017-11-01 01:58:45,541 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-01 01:58:45,542 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2017-11-01 01:58:45,648 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-01 01:58:45,649 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(({(Bharathi),(Komal),(Archana),(Trupthi),(Preethi),(Rajesh),(siddarth),(Rajiv)},{(72),(83),(87),(75),(93),(90),(78),(89)}),93)
grunt>
```

6. LIMIT:

   The LIMIT operator is used to get a limited number of tuples from a relation.

```
grunt>
grunt>
grunt>
grunt>
grunt>
grunt> limit_data = LIMIT student_details 4;
grunt> DESCRIBE limit_data
limit_data: {id: int,firstname: chararray,lastname: chararray,age: int,phone: chararray,city: chararray,gpa: int}
grunt>
```

```
2017-11-01 02:04:14,788 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,Rajiv,Reddy,21,9848022337,Hyderabad,89)
(2,siddarth,Battacharya,22,9848022338,Kolkata,78)
(3,Rajesh,Khanna,22,9848022339,Delhi,90)
(4,Preethi,Agarwal,21,9848022330,Pune,93)
grunt>
```

7. STORE:-

We can store the loaded data in the file system using the store operator.

```
grunt>
grunt> STORE student_gpa_sum INTO '/home/acadgild/hadoop/student_gpa_sumout' USING PigStorage(',');
2017-11-01 02:09:20,442 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-ch
ecksum
2017-11-01 02:09:20,444 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-01 02:09:20,600 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapr
educe.output.textoutputformat.separator
2017-11-01 02:09:20,676 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2017-11-01 02:09:20,747 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-ch
ecksum
2017-11-01 02:09:20,748 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-11-01 02:09:20,751 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2017-11-01 02:09:20,751 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, Consta
ntCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOpti
mizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2017-11-01 02:09:20,760 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimisti
```

```
[acadgild@localhost hadoop]$ ls -lrt
total 3188
drwxr-xr-x. 3 acadgild acadgild    4096 Aug 12  2016 namenode
drwx------. 3 acadgild acadgild    4096 Aug 12  2016 datanode
-rw-rw-r--. 1 acadgild acadgild   21007 Oct  9 01:16 sample_temperature_dataset.csv
-rw-rw-r--. 1 acadgild acadgild    5445 Oct  9 23:45 mapreduce-0.0.1-SNAPSHOT.jar
-rw-rw-r--. 1 acadgild acadgild     210 Oct 11 22:24 max-temp.txt
-rw-rw-r--. 1 acadgild acadgild 3194099 Oct 14 13:21 NYSE_daily
drwxrwxr-x. 2 acadgild acadgild    4096 Oct 15 01:15 maxout
-rw-rw-r--. 1 acadgild acadgild    5338 Oct 15 01:22 pig_1508008101045.log
-rw-rw-r--. 1 acadgild acadgild      48 Oct 21 10:59 dept_data.csv
-rw-rw-r--. 1 acadgild acadgild     375 Nov  1 00:44 student_details.txt
drwxrwxr-x. 2 acadgild acadgild    4096 Nov  1 02:09 student_gpa_sumout
[acadgild@localhost hadoop]$ cd student_gpa_sumout/
[acadgild@localhost student_gpa_sumout]$ ls -lrt
total 4
-rw-r--r--. 1 acadgild acadgild 127 Nov  1 02:09 part-r-00000
-rw-r--r--. 1 acadgild acadgild   0 Nov  1 02:09 _SUCCESS
[acadgild@localhost student_gpa_sumout]$ cat part-r-00000
({(Bharathi),(Komal),(Archana),(Trupthi),(Preethi),(Rajesh),(siddarth),(Rajiv)},{(72),(83),(87),(75),(93),(90),(78),(89)}),667
[acadgild@localhost student_gpa_sumout]$
```

8. DISTINCT:-

The DISTINCT operator is used to remove redundant (duplicate) tuples from a relation.

As there is not duplicate data in our set, so Original set output and distinct set output will be same.

```
2017-11-01 02:15:42,285 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,Rajiv,Reddy,21,9848022337,Hyderabad,89)
(2,siddarth,Battacharya,22,9848022338,Kolkata,78)
(3,Rajesh,Khanna,22,9848022339,Delhi,90)
(4,Preethi,Agarwal,21,9848022330,Pune,93)
(5,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar,75)
(6,Archana,Mishra,23,9848022335,Chennai,87)
(7,Komal,Nayak,24,9848022334,trivendram,83)
(8,Bharathi,Nambiayar,24,9848022333,Chennai,72)
grunt> distinct_data = DISTINCT student_details;
grunt> DESCRIBE distinct_data
distinct_data: {id: int,firstname: chararray,lastname: chararray,age: int,phone: chararray,city: chararray,gpa: int}
grunt>
```

```
2017-11-01 02:17:20,250 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-11-01 02:17:20,250 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,Rajiv,Reddy,21,9848022337,Hyderabad,89)
(2,siddarth,Battacharya,22,9848022338,Kolkata,78)
(3,Rajesh,Khanna,22,9848022339,Delhi,90)
(4,Preethi,Agarwal,21,9848022330,Pune,93)
(5,Trupthi,Mohanthy,23,9848022336,Bhuwaneshwar,75)
(6,Archana,Mishra,23,9848022335,Chennai,87)
(7,Komal,Nayak,24,9848022334,trivendram,83)
(8,Bharathi,Nambiayar,24,9848022333,Chennai,72)
grunt>
```

Submitted by

Hardik Kaushik