# Assignment 2.2

Input Data:

```
acadgild@localhost:~
[acadgild@localhost ~]$ cat input.csv
1,Cameron,Larkcum,clarkcum0@youku.com,Male,38.184.231.111
2,Carlye,Whitington,cwhitington1@artisteer.com,Female,253.240.112.101
3,Jasun,Anstee,janstee2@latimes.com,Male,126.161.252.141
4,Tome,Joselin,tjoselin3@mlb.com,Male,103.12.209.61
5,Terrence,Brolan,tbrolan4@bbb.org,Male,166.181.35.71
6,Dal,Porte,dporte5@patch.com,Male,22.28.81.135
7,Annmaria,Seeley,aseeley6@independent.co.uk,Female,34.9.187.58
8,Mallory,Moreinu,mmoreinu7@pcworld.com,Female,41.197.110.62
9,Bax,Liddel,bliddel8@over-blog.com,Male,157.87.97.11
10,Del,Quilter,dquilter9@noaa.gov,Male,218.250.204.131
11,Webb,Geroldini,wgeroldinia@cnet.com,Male,209.226.45.138
12,Ilise,Saur,isaurb@disqus.com,Female,180.125.107.20
13,Keefer,Cay,kcayc@mit.edu,Male,76.104.44.154
14,Trever,Whyard,twhyardd@nih.gov,Male,188.225.247.186
15,Burgess,Martine,bmartinee@nydailynews.com,Male,100.243.145.200
[acadgild@localhost ~]$
```

Entered grunt shell with **‘ pig -x LOCAL ’**

Loading data:

grunt> contact_info = LOAD '/home/acadgild/input.csv' USING PigStorage(',') as (id:int, firstname:chararray, lastname:chararray,email:chararray, gender:chararray,ip:chararray);

```
grunt> contact_info = LOAD '/home/acadgild/input.csv' USING PigStorage(',') as (id:int, firstname:chararray, lastname
:chararray,email:chararray, gender:chararray,ip:chararray);
```

Displaying data in contact_info:

dump contact_info

```
Input(s):
Successfully read 15 records from: "/home/acadgild/input.csv"

Output(s):
Successfully stored 15 records in: "file:/tmp/temp1532720645/tmp1654789236"

Counters:
Total records written : 15
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1326997257_0001


2018-05-19 15:35:22,336 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
me=JobTracker, sessionId= - already initialized
2018-05-19 15:35:22,341 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
me=JobTracker, sessionId= - already initialized
2018-05-19 15:35:22,343 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processNa
me=JobTracker, sessionId= - already initialized
2018-05-19 15:35:22,363 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Succes
s!
2018-05-19 15:35:22,368 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated.
 Instead, use dfs.bytes-per-checksum
2018-05-19 15:35:22,369 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Inste
ad, use fs.defaultFS
2018-05-19 15:35:22,369 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialize
d
2018-05-19 15:35:22,391 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process :
1
2018-05-19 15:35:22,391 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to pr
ocess : 1
(1,Cameron,Larkcum,clarkcum0@youku.com,Male,38.184.231.111)
(2,Carlye,Whitington,cwhitington1@artisteer.com,Female,253.240.112.101)
(3,Jasun,Anstee,janstee2@latimes.com,Male,126.161.252.141)
(4,Tome,Joselin,tjoselin3@mlb.com,Male,103.12.209.61)
(5,Terrence,Brolan,tbrolan4@bbb.org,Male,166.181.35.71)
(6,Dal,Porte,dporte5@patch.com,Male,22.28.81.135)
(7,Annmaria,Seeley,aseeley6@independent.co.uk,Female,34.9.187.58)
(8,Mallory,Moreinu,mmoreinu7@pcworld.com,Female,41.197.110.62)
(9,Bax,Liddel,bliddel8@over-blog.com,Male,157.87.97.11)
(10,Del,Quilter,dquilter9@noaa.gov,Male,218.250.204.131)
(11,Webb,Geroldini,wgeroldinia@cnet.com,Male,209.226.45.138)
(12,Ilise,Saur,isaurb@disqus.com,Female,180.125.107.20)
(13,Keefer,Cay,kcayc@mit.edu,Male,76.104.44.154)
(14,Trever,Whyard,twhyardd@nih.gov,Male,188.225.247.186)
(15,Burgess,Martine,bmartinee@nydailynews.com,Male,100.243.145.200)
grunt>
```

## 1. Concat

Concatenating firstname and lastname separated by a space.

Using foreach loop to do the same for all rows.

```
grunt> first_last_name = foreach contact_info Generate CONCAT(firstname,' ',lastname) as name;
grunt> dump first_last_name;
```

```
(Cameron Larkcum)
(Carlye Whitington)
(Jasun Anstee)
(Tome Joselin)
(Terrence Brolan)
(Dal Porte)
(Annmaria Seeley)
(Mallory Moreinu)
(Bax Liddel)
(Del Quilter)
(Webb Geroldini)
(Ilise Saur)
(Keefer Cay)
(Trever Whyard)
(Burgess Martine)
grunt>
```

## 2. Tokenize

Using tokenize on the first_last_name to split them.

```
grunt> tokenized_name = foreach first_last_name GENERATE TOKENIZE(name);
grunt> dump tokenized_name;
```

```
({(Cameron),(Larkcum)})
({(Carlye),(Whitington)})
({(Jasun),(Anstee)})
({(Tome),(Joselin)})
({(Terrence),(Brolan)})
({(Dal),(Porte)})
({(Annmaria),(Seeley)})
({(Mallory),(Moreinu)})
({(Bax),(Liddel)})
({(Del),(Quilter)})
({(Webb),(Geroldini)})
({(Ilise),(Saur)})
({(Keefer),(Cay)})
({(Trever),(Whyard)})
({(Burgess),(Martine)})
grunt>
```

## 3. Sum

In this we group the contact info based on gender.
Then we find the sum of id for this grouped data.

Output is the sum of id grouped by gender.

```
grunt> contact_id_grouped = GROUP contact_info BY gender;
grunt> id_sum = foreach contact_id_grouped GENERATE SUM(contact_info.id);
grunt> dump id_sum;
```

```
(91)
(29)
grunt>
```

4. **Min**
5. **Max**

Here we calculate minimum and maximum id based on gender.

We can see that male has min 1 and max 15 as id.
female has min 2 and max 12 id.

```
grunt> id_min_max = foreach contact_id_grouped GENERATE MAX(contact_info.id),MIN(contact_info.id),con
tact_info.gender;
grunt> dump id_min_max;
```

```
(15,1,{(Male),(Male),(Male),(Male),(Male),(Male),(Male),(Male),(Male),(Male),(Male)})
(12,2,{(Female),(Female),(Female),(Female)})
grunt>
```

6. **Limit**

This is similar to SQL limit operation where the output is limited to given number

```
grunt> first_5 = limit contact_info 5;
grunt> dump first_5;
```

```
(1,Cameron,Larkcum,clarkcum0@youku.com,Male,38.184.231.111)
(2,Carlye,Whitington,cwhitington1@artisteer.com,Female,253.240.112.101)
(3,Jasun,Anstee,janstee2@latimes.com,Male,126.161.252.141)
(4,Tome,Joselin,tjoselin3@ml.com,Male,103.12.209.61)
(5,Terrence,Brolan,tbrolan4@bbb.org,Male,166.181.35.71)
grunt>
```

7. **Store**

Used for storing the output into a file.

Here I have stored the data into local system.

```
grunt> STORE first_5 into '/home/acadgild/first_5' USING PigStorage(',');
```

```
[acadgild@localhost ~]$ cat first_5/part-r-00000
1,Cameron,Larkcum,clarkcum0@youku.com,Male,38.184.231.111
2,Carlye,Whitington,cwhitington1@artisteer.com,Female,253.240.112.101
3,Jasun,Anstee,janstee2@latimes.com,Male,126.161.252.141
4,Tome,Joselin,tjoselin3@mlb.com,Male,103.12.209.61
5,Terrence,Brolan,tbrolan4@bbb.org,Male,166.181.35.71
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

## 8. Distinct

Finds out the distinct/unique values.

```
(Male)
(Female)
(Male)
(Male)
(Male)
(Male)
(Female)
(Female)
(Male)
(Male)
(Male)
(Female)
(Male)
(Male)
(Male)
grunt> distinct_genders = DISTINCT genders;
grunt> dump distinct_genders;
```

```
(Male)
(Female)
```

## 9. Flatten

This flattens the nested schema into a unnested schema.

```
grunt> flattened_contact = foreach contact_id_grouped GENERATE FLATTEN(contact_info);
grunt> dump flattened_contact;
```

```
(1,Cameron,Larkcum,clarkcum0@youku.com,Male,38.184.231.111)
(14,Trever,Whyard,twhyardd@nih.gov,Male,188.225.247.186)
(13,Keefer,Cay,kcayc@mit.edu,Male,76.104.44.154)
(11,Webb,Geroldini,wgeroldinia@cnet.com,Male,209.226.45.138)
(10,Del,Quilter,dquilter9@noaa.gov,Male,218.250.204.131)
(9,Bax,Liddel,bliddel8@over-blog.com,Male,157.87.97.11)
(15,Burgess,Martine,bmartinee@nydailynews.com,Male,100.243.145.200)
(6,Dal,Porte,dporte5@patch.com,Male,22.28.81.135)
(5,Terrence,Brolan,tbrolan4@bbb.org,Male,166.181.35.71)
(4,Tome,Joselin,tjoselin3@mlb.com,Male,103.12.209.61)
(3,Jasun,Anstee,janstee2@latimes.com,Male,126.161.252.141)
(12,Ilise,Saur,isaurb@disqus.com,Female,180.125.107.20)
(8,Mallory,Moreinu,mmoreinu7@pcworld.com,Female,41.197.110.62)
(7,Annmaria,Seeley,aseeley6@independent.co.uk,Female,34.9.187.58)
(2,Carlye,Whitington,cwhitington1@artisteer.com,Female,253.240.112.101)
grunt>
```

## 10.IsEmpty

To find if empty maps/bags.

```
grunt> empty_entry = filter contact_id_grouped by IsEmpty(contact_info);
grunt> dump empty_entry;
```

The dump empty_entry has no value as there were no empty bags.

Now trying to find which are NOT empty.

```
grunt> non_empty_entry = filter contact_id_grouped by not IsEmpty(contact_info);
grunt> dump non_empty_entry;
```

```
Total input paths to process : 1
(Male,{(1,Cameron,Larkcum,clarkcum0@youku.com,Male,38.184.231.111),(14,Trever,Whyard,twhyardd@nih.gov
,Male,188.225.247.186),(13,Keefer,Cay,kcayc@mit.edu,Male,76.104.44.154),(11,Webb,Geroldini,wgeroldini
a@cnet.com,Male,209.226.45.138),(10,Del,Quilter,dquilter9@noaa.gov,Male,218.250.204.131),(9,Bax,Lidde
l,bliddel8@over-blog.com,Male,157.87.97.11),(15,Burgess,Martine,bmartinee@nydailynews.com,Male,100.24
3.145.200),(6,Dal,Porte,dporte5@patch.com,Male,22.28.81.135),(5,Terrence,Brolan,tbrolan4@bbb.org,Male
,166.181.35.71),(4,Tome,Joselin,tjoselin3@mlb.com,Male,103.12.209.61),(3,Jasun,Anstee,janstee2@latime
s.com,Male,126.161.252.141)})
(Female,{(12,Ilise,Saur,isaurb@disqus.com,Female,180.125.107.20),(8,Mallory,Moreinu,mmoreinu7@pcworld
.com,Female,41.197.110.62),(7,Annmaria,Seeley,aseeley6@independent.co.uk,Female,34.9.187.58),(2,Carly
e,Whitington,cwhitington1@artisteer.com,Female,253.240.112.101)})
grunt>
```