1)Running TestDataGen

Magic Number: 142881

Files Generated: foodplaces142881.txt and foodratings142881.txt

Command Used: java TestDataGen

```
[hadoop@ip-172-31-44-188 ~]$ java TestDataGen
Magic Number = 142881
[hadoop@ip-172-31-44-188 ~]$ ls
foodplaces142881.txt  foodratings142881.txt  TestDataGen.class
[hadoop@ip-172-31-44-188 ~]$
```

```
[hadoop@ip-172-31-44-188 ~]$ hadoop fs -ls /user/csp554
Found 2 items
-rw-r--r--   1 hadoop hadoop         59 2020-10-01 19:28 /user/csp554/foodplaces
142881.txt
-rw-r--r--   1 hadoop hadoop      17518 2020-10-01 19:27 /user/csp554/foodrating
s142881.txt
[hadoop@ip-172-31-44-188 ~]$
```

Creating food_ratings Relation:

Command Used to Create Relation: food_ratings= LOAD 'user/hadoop/foodratings124881.txt' USING PigStorage(',') AS (name:CharArray,f1:int,f2:int,f3:int,f4:int,placeid:int);

Command Used to display Schema Relation: Describe food_ratings;

```
grunt> foodratingsv2= LOAD '/user/csp554/foodratings142881.txt' USING PigStorage
(',') AS (name:chararray,f1:int,f2:int,f3:int,f4:int,placeid:int);
20/10/01 19:41:22 INFO Configuration.deprecation: yarn.resourcemanager.system-me
trics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publishe
r.enabled
```

```
grunt> describe foodratingsv2;
foodratingsv2: {name: chararray,f1: int,f2: int,f3: int,f4: int,placeid: int}
grunt>
```

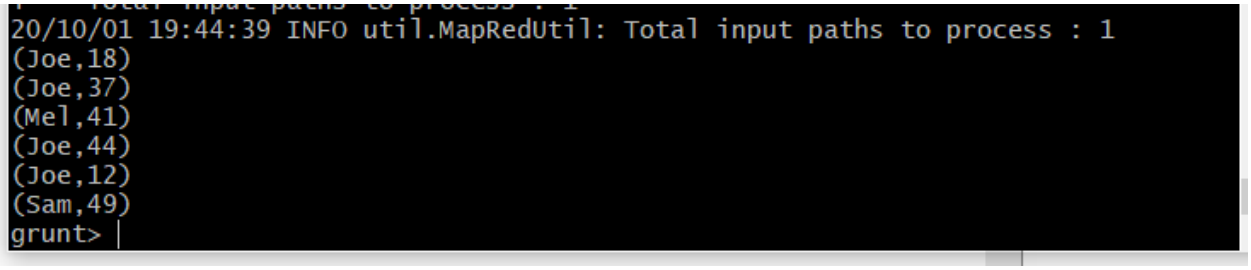
2) Creating a new relation from existing

Command Used: food_ratings_subset= FOREACH foodratings GENERATE name,f4;

```
grunt> food_ratings_subset= FOREACH foodratingsv2 GENERATE name,f4;
grunt> topC= limit food_ratings_subset 6;
grunt> dump topC;
941695 [main] INFO  org.apache.pig.tools.pigstats.ScriptState  - Pig features us
ed in the script: LIMIT
20/10/01 19:44:39 INFO pigstats.ScriptState: Pig features used in the script: LI
MIT
```

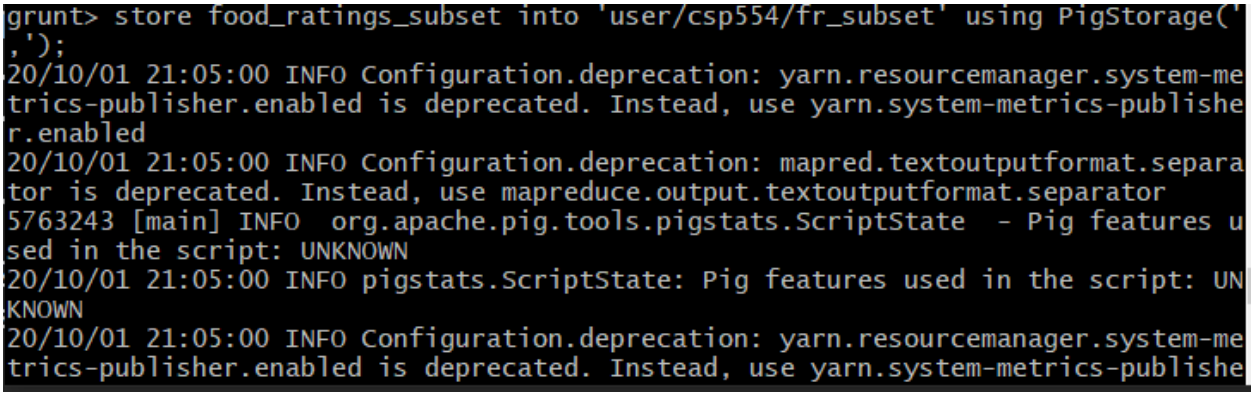
Printing top 6 rows

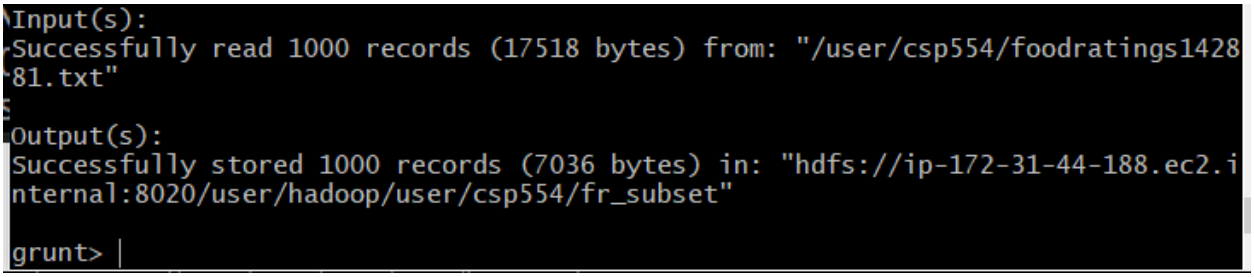
Command Used: `top6=limit food_ratings_subset 6;`



Storing into food_ratings_subset into HDFS

Command Used: `store food_ratings_subset into 'user/csp554/fr_subset' using PigStorage(',');`





3) ***Calculating MIN,MAX,AVG of the f2,f3 in foodratings***

COMMAND USED: `food_ratings_profile= FOREACH foodratingsAll generate MIN(foodratingsv2.f2) as f2min,MAX(food_ratingsv2.f2) as f2max,AVG(foodratingsv2.f2) as avgf2,MIN(foodratingsv2.f3) as f3min,MAX(foodratingsv2.f3) as f3max,AVG(foodratings.f3) as f3avg;`

```
Output(s):
Successfully stored 1 records (28 bytes) in: "hdfs://ip-172-31-44-188.ec2.internal:8020/tmp/temp-1421641141/tmp-201335866"

20/10/01 20:09:36 INFO input.FileInputFormat: Total input files to process : 1
2439148 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil  - Total input paths to process : 1
20/10/01 20:09:36 INFO util.MapRedUtil: Total input paths to process : 1
(1,50,26.073,1,50,24.917)
grunt>
```

4)*Filtering the records based on f1<20 and f3>5*

Command Used: `food_ratings_filter= FILTER foodratingsv2 by f1<20 and f3>5;`

```
grunt> food_ratings_filter= FILTER foodratingsv2 by f1<20 and f3>5;
grunt> foodRatFilterLimit= limit food_ratings_filter 6;
grunt> dump foodRatFilterLimit
```

Printing Top 6 rows

Command Used: `foodRatFilterLimit= limit food_ratings_filter 6; dump foodRatFilterLimit`

```
l  - Total input paths to process : 1
20/10/01 20:14:28 INFO util.MapRedUtil: Total input paths to process : 1
(Joe,1,48,21,12,2)
(Joe,4,49,44,50,4)
(Mel,16,30,30,3,4)
(Sam,17,48,33,2,2)
(Joy,9,37,32,8,4)
(Sam,7,36,34,17,2)
grunt>
```

5) *Sampling the 2 percent data randomly*

Command Used : `food_rating2percent= sample foodratingsv2 0.02;`

```
grunt> food_rating2percent= sample foodratingsv2 0.02;
grunt> foodtop10= limit food_rating2percent 10;
grunt> dump foodtop10;
```

```
l  - Total input paths to process : 1
0/10/01 20:17:30 INFO util.MapRedUtil: Total input paths to process : 1
Joe,6,29,38,27,2)
Sam,8,43,23,49,3)
Sam,20,13,3,36,3)
Mel,38,13,49,44,1)
Mel,3,16,49,33,5)
Jill,5,15,8,38,5)
Joy,42,19,10,34,3)
Jill,13,17,43,2,4)
Sam,42,3,29,23,1)
Sam,40,46,2,18,3)
runt>
```

6) ***Loading foodplaces Data***

Command Used: `foodplaces= load '/user/csp554/foodplaces142881.txt' using PigStorage(',') as (placeid:int,placeName:chararray);`

```
grunt> foodplaces= load '/user/csp554/foodplaces142881.txt' using PigStorage(','
) as (placeid:int,placeName:chararray);
20/10/01 20:21:32 INFO Configuration.deprecation: yarn.resourcemanager.system-me
trics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publishe
r.enabled
grunt> describe foodplaces;
foodplaces: {placeid: int,placeName: chararray}
grunt>
```

Performing join between foodplaces and foodratings

Command Used: `foodrating_foodplaces_join= join foodratingsv2 by placeid,foodplaces by placeid;`

```
grunt> foodrating_foodplaces_join= join foodratingsv2 by placeid,foodplaces by p
laceid;
grunt> describe foodrating_foodplaces_join;
foodrating_foodplaces_join: {foodratingsv2::name: chararray,foodratingsv2::f1: i
nt,foodratingsv2::f2: int,foodratingsv2::f3: int,foodratingsv2::f4: int,foodrati
ngsv2::placeid: int,foodplaces::placeid: int,foodplaces::placeName: chararray}
grunt>
```

```
grunt> topJoin= limit foodrating_foodplaces_join 6;
grunt> dump topJoin
```

```
20/10/01 20:27:07 INFO input.FileInputFormat: Total input files to process : 1
3490224 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUt
il  - Total input paths to process : 1
20/10/01 20:27:07 INFO util.MapRedUtil: Total input paths to process : 1
(Sam,15,33,6,22,1,1,China Bistro)
(Jill,46,1,49,4,1,1,China Bistro)
(Jill,40,35,26,3,1,1,China Bistro)
(Joy,11,42,28,13,1,1,China Bistro)
(Joy,26,29,29,47,1,1,China Bistro)
(Joe,42,15,39,24,1,1,China Bistro)
grunt>
```

7)

I. D
II. C
III. B
IV. B
V. B
VI. A