CSP 554 Assignment 5

Atharva Tanaji Kadam (A20467229)

Magic Number = 11799

Create new versions of the foodratings and foodplaces files by using TestDataGen (as described in assignment #4) and copy them to HDFS.

```
[hadoop@ip-172-31-33-233 ~]$ java TestDataGen
Magic Number = 11799
[hadoop@ip-172-31-33-233 ~]$ hdfs fs -ls
Error: Could not find or load main class fs
[hadoop@ip-172-31-33-233 ~]$ hdfs fs -ls /user/hadoop
Error: Could not find or load main class fs
[hadoop@ip-172-31-33-233 ~]$ hdfs  -ls /user/hadoop
Unrecognized option: -ls
Error: Could not create the Java Virtual Machine.
Error: A fatal exception has occurred. Program will exit.
[hadoop@ip-172-31-33-233 ~]$ hdfs dfs -ls /user/hadoop
[hadoop@ip-172-31-33-233 ~]$ hdfs dfs -ls /user/hadoop/
[hadoop@ip-172-31-33-233 ~]$ hadoop fs -ls /user/hadoop
[hadoop@ip-172-31-33-233 ~]$ ls
foodplaces11799.txt  foodratings11799.txt  TestDataGen.class
[hadoop@ip-172-31-33-233 ~]$ hdfs dfs -copyFromLocal foodratings11799.txt /user/hadoop
[hadoop@ip-172-31-33-233 ~]$ hdfs dfs -copyFromLocal foodplaces11799.txt /user/hadoop
[hadoop@ip-172-31-33-233 ~]$ hdfs dfs -ls /user/hadoop
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup         59 2022-10-10 01:17 /user/hadoop/foodplaces11799.txt
-rw-r--r--   1 hadoop hdfsadmingroup      17461 2022-10-10 01:17 /user/hadoop/foodratings11799.txt
[hadoop@ip-172-31-33-233 ~]$
```

Question 1:

```
grunt> food_ratings = LOAD '/user/hadoop/foodratings11799.txt' USING PigStorage(',')
>> AS (name:chararray,
>> f1:int,
>> f2:int,
>> f3:int,
>> f4:int,
>> placeid:int);
22/10/10 01:28:26 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt> DESCRIBE food_ratings;
food_ratings: {name: chararray,f1: int,f2: int,f3: int,f4: int,placeid: int}
grunt>
```

food_ratings = LOAD '/user/hadoop/foodratings11799.txt' USING PigStorage(',')
AS (name:chararray,
f1:int,

f2:int,
f3:int,
f4:int,
placeid:int);

DESCRIBE food_ratings;

Question 2 -

```
grunt> food_ratings_subset = FOREACH food_ratings GENERATE name, f4;
grunt> STORE food_ratings_subset INTO 'food_ratings_subset' USING PigStorage ('|');
22/10/10 01:29:47 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is depre
568191 [main] INFO  org.apache.pig.tools.pigstats.ScriptState  - Pig features used in the script: UNKNOWN
22/10/10 01:29:47 INFO pigstats.ScriptState: Pig features used in the script: UNKNOWN
22/10/10 01:29:47 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is depre
568212 [main] INFO  org.apache.pig.data.SchemaTupleBackend  - Key [pig.schematuple] was not set... will not gene
22/10/10 01:29:47 INFO data.SchemaTupleBackend: Key [pig.schematuple] was not set... will not generate code.
568213 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer  - {RULES_ENABLED=[AddForEach,
imitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer,
itFilter, StreamTypeCastInserter]}
22/10/10 01:29:47 INFO optimizer.LogicalPlanOptimizer: {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCa
erter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, Pu
er]}
568223 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor  - Columns pruned for food_ratings:
22/10/10 01:29:47 INFO rules.ColumnPruneVisitor: Columns pruned for food_ratings: $1, $2, $3, $5
22/10/10 01:29:47 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.atte
22/10/10 01:29:47 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is depre
568247 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher  - Tez staging directory is /t
22/10/10 01:29:47 INFO tez.TezLauncher: Tez staging directory is /tmp/temp1844803879 and resources directory is
568248 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.plan.TezCompiler  - File concatenation thr
```

```
Success!


DAG 0:
                                Name: PigLatin:DefaultJobName-0_scope-1
                       ApplicationId: job_1665362784034_0001
                   TotalLaunchedTasks: 1
                        FileBytesRead: 0
                     FileBytesWritten: 0
                        HdfsBytesRead: 17461
                     HdfsBytesWritten: 7008
      SpillableMemoryManager spill count: 0
                Bags proactively spilled: 0
              Records proactively spilled: 0

DAG Plan:
Tez vertex scope-19

Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  HdfsBytesRead HdfsBytesWritten Alias        Feature Outputs
scope-19           1          1          1000                   0           1000             0               0          17461            7008 food_ratings,food_ratings_subset h
dfs://ip-172-31-33-233.ec2.internal:8020/user/hadoop/food_ratings_subset,

Input(s):
Successfully read 1000 records (17461 bytes) from: "/user/hadoop/foodratings11799.txt"

Output(s):
Successfully stored 1000 records (7008 bytes) in: "hdfs://ip-172-31-33-233.ec2.internal:8020/user/hadoop/food_ratings_subset"
```

```
[hadoop@ip-172-31-33-233 ~]$ hdfs dfs -ls /user/hadoop/*subset*
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2022-10-10 01:30 /user/hadoop/food_ratings_subset/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup       7008 2022-10-10 01:30 /user/hadoop/food_ratings_subset/part-v000-o000-r-00000
[hadoop@ip-172-31-33-233 ~]$
```

food_ratings = LOAD '/user/hadoop/foodratings11799.txt' USING PigStorage(',') AS (name:chararray, f1:int, f2:int, f3:int, f4:int, placeid:int);

food_ratings_subset = FOREACH food_ratings GENERATE name, f4;

STORE food_ratings_subset INTO 'food_ratings_subset' USING PigStorage ('|');

```
Details at logfile: /mnt/var/log/pig/pig_1665365881184.log
grunt> food_ratings_limit = LIMIT food_ratings_subset 6;
grunt> DUMP food_ratings_limit;
656711 [main] INFO  org.apache.pig.tools.pigstats.ScriptState  - Pi
22/10/10 01:48:57 INFO pigstats.ScriptState: Pig features used in t
22/10/10 01:48:57 INFO Configuration.deprecation: yarn.resourcemana
657785 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil  - Total input paths to
22/10/10 01:48:58 INFO util.MapRedUtil: Total input paths to process : 1
(Sam,46)
(Sam,6)
(Jill,26)
(Joy,28)
(Joy,14)
(Mel,18)
grunt>
```

Question 3 :

```
grunt> food_ratings_profile = FOREACH food_ratings_all GENERATE MIN(food_ratings.f2) as MIN_F2, MAX(food_ratings.f2) as MAX_F2, AVG(food_ratings.f2) as AVG_F2, MIN(food_ratings.f3)
as MIN_F3, MAX(food_ratings.f3) as MAX_F3, AVG(food_ratings.f3) as AVG_F3;
grunt> DESCRIBE food_ratings_profile;
food_ratings_profile: {MIN_F2: int,MAX_F2: int,AVG_F2: double,MIN_F3: int,MAX_F3: int,AVG_F3: double}
grunt> DUMP food_ratings_profile
1199128 [main] INFO  org.apache.pig.tools.pigstats.ScriptState  - Pig features used in the script: GROUP_BY
22/10/10 01:58:00 INFO pigstats.ScriptState: Pig features used in the script: GROUP_BY
22/10/10 01:58:00 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
1199146 [main] INFO  org.apache.pig.data.SchemaTupleBackend  - Key [pig.schematuple] was not set... will not generate code.
22/10/10 01:58:00 INFO data.SchemaTupleBackend: Key [pig.schematuple] was not set... will not generate code.


                    FileBytesWritten: 87
                       HdfsBytesRead: 17461
                    HdfsBytesWritten: 28
      SpillableMemoryManager spill count: 0
                Bags proactively spilled: 0
             Records proactively spilled: 0

DAG Plan:
Tez vertex scope-60     ->      Tez vertex scope-61,
Tez vertex scope-61

Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  HdfsBytesRead HdfsBytesWritten Alias        Feature Outputs
scope-60           1          1          1000                   0           1000            64              87          17461               0 food_ratings,food_ratings_all,food
_ratings_profile
scope-61           1          1             0                   1              1            87               0              0              28 food_ratings_profile GROUP_BY    h
dfs://ip-172-31-33-233.ec2.internal:8020/tmp/temp-1740288465/tmp2057289421,

Input(s):
Successfully read 1000 records (17461 bytes) from: "/user/hadoop/foodratings11799.txt"

Output(s):
Successfully stored 1 records (28 bytes) in: "hdfs://ip-172-31-33-233.ec2.internal:8020/tmp/temp-1740288465/tmp2057289421"

22/10/10 01:58:18 INFO input.FileInputFormat: Total input files to process : 1
1217185 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil  - Total input paths to process : 1
22/10/10 01:58:18 INFO util.MapRedUtil: Total input paths to process : 1
(1,50,25.285,1,50,25.351)
grunt>
```

food_ratings_all = GROUP food_ratings ALL;
food_ratings_profile = FOREACH food_ratings_all GENERATE MIN(food_ratings.f2) as MIN_F2, MAX(food_ratings.f2) as MAX_F2, AVG(food_ratings.f2) as AVG_F2, MIN(food_ratings.f3) as MIN_F3, MAX(food_ratings.f3) as MAX_F3, AVG(food_ratings.f3) as AVG_F3;

DESCRIBE food_ratings_profile;
DUMP food_ratings_profile;

‘

Question 4:

```
grunt> food_ratings_filtered = FILTER food_ratings BY (f1 < 20) AND (f3 < 5);
grunt> food_ratings_filtered_six = LIMIT food_ratings_filtered 6;
grunt> DESCRIBE food_ratings_filtered_six;
food_ratings_filtered_six: {name: chararray,f1: int,f2: int,f3: int,f4: int,placeid: int}
grunt> DUMP food_ratings_filtered_six;
1519950 [main] INFO  org.apache.pig.tools.pigstats.ScriptState  - Pig features used in the script: FILTER,LIMIT
22/10/10 02:03:21 INFO pigstats.ScriptState: Pig features used in the script: FILTER,LIMIT
22/10/10 02:03:21 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
1519997 [main] INFO  org.apache.pig.data.SchemaTupleBackend  - Key [pig.schematuple] was not set... will not generate
22/10/10 02:03:21 INFO data.SchemaTupleBackend: Key [pig.schematuple] was not set... will not generate code.
1520004 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer  - {RULES_ENABLED=[AddForEach, Colu
LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, Predi
litFilter, StreamTypeCastInserter]}
22/10/10 02:03:21 INFO optimizer.LogicalPlanOptimizer: {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculat
erter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDown
er]}
22/10/10 02:03:21 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
1520045 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.TezLauncher  - Tez staging directory is /tmp/te
22/10/10 02:03:21 INFO tez.TezLauncher: Tez staging directory is /tmp/temp-927182701 and resources directory is /tmp/t
1520046 [main] INFO  org.apache.pig.backend.hadoop.executionengine.tez.plan.TezCompiler  - File concatenation threshol
```

```
,food_ratings_filtered_six
scope-143          1          1              6                    0              6            151               0
dfs://ip-172-31-33-233.ec2.internal:8020/tmp/temp-1740288465/tmp-1377844105,

Input(s):
Successfully read 212 records (17461 bytes) from: "/user/hadoop/foodratings11799.txt"

Output(s):
Successfully stored 6 records (117 bytes) in: "hdfs://ip-172-31-33-233.ec2.internal:8020/tmp/temp-1740288465/tmp-13778

22/10/10 02:03:29 INFO input.FileInputFormat: Total input files to process : 1
1528436 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil  - Total input paths to process : 1
22/10/10 02:03:29 INFO util.MapRedUtil: Total input paths to process : 1
(Joe,13,27,3,15,1)
(Jill,11,19,2,7,5)
(Mel,5,25,1,23,1)
(Sam,8,8,2,38,1)
(Joy,16,36,3,20,3)
(Jill,19,20,1,26,5)
grunt>
```

food_ratings_filtered = FILTER food_ratings BY (f1 < 20) AND (f3 > 5);
food_ratings_filtered_six = LIMIT food_ratings_filtered 6;
DESCRIBE food_ratings_filtered_six;
DUMP food_ratings_filtered_six;

Question 5 :

```
grunt> food_ratings_2percent = SAMPLE food_ratings 0.02;
grunt> DESCRIBE food_ratings_2percent
food_ratings_2percent: {name: chararray,f1: int,f2: int,f3: int,f4: int,placeid: int}
grunt> food_ratings_2percent_ten = LIMIT food_ratings_2percent 10;
grunt> DESCRIBE food_ratings_2percent_ten;
food_ratings_2percent_ten: {name: chararray,f1: int,f2: int,f3: int,f4: int,placeid: int}
grunt> DUMP food_ratings_2percent_ten;
1748212 [main] INFO  org.apache.pig.tools.pigstats.ScriptState  - Pig features used in the script: FILTER,LIMIT
22/10/10 02:07:09 INFO pigstats.ScriptState: Pig features used in the script: FILTER,LIMIT
22/10/10 02:07:09 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
1748225 [main] INFO  org.apache.pig.data.SchemaTupleBackend  - Key [pig.schematuple] was not set... will not generate
22/10/10 02:07:09 INFO data.SchemaTupleBackend: Key [pig.schematuple] was not set... will not generate code.
1748226 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer  - {RULES_ENABLED=[AddForEach, Colu
LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, Predi
litFilter, StreamTypeCastInserter]}
22/10/10 02:07:09 INFO optimizer.LogicalPlanOptimizer: {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculat
erter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDowr
er]}
22/10/10 02:07:09 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
```

```
Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  HdfsBytesRead HdfsBytesWritten Alias        Feature Outputs
scope-172          1          1            904                   0             10              0             202          17461                0 food_ratings,food_ratings_2percent
,food_ratings_2percent_ten
scope-174          1          1             10                   0             10            202               0              0              203 food_ratings_2percent_ten    LIMIh
dfs://ip-172-31-33-233.ec2.internal:8020/tmp/temp-1740288465/tmp1655843618,

Input(s):
Successfully read 904 records (17461 bytes) from: "/user/hadoop/foodratings11799.txt"

Output(s):
Successfully stored 10 records (203 bytes) in: "hdfs://ip-172-31-33-233.ec2.internal:8020/tmp/temp-1740288465/tmp1655843618"

22/10/10 02:07:16 INFO input.FileInputFormat: Total input files to process : 1
1755461 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil  - Total input paths to process : 1
22/10/10 02:07:16 INFO util.MapRedUtil: Total input paths to process : 1
(Jill,48,48,41,6,2)
(Mel,26,17,45,33,2)
(Joy,39,34,48,20,3)
(Joe,5,4,4,17,5)
(Jill,10,44,22,6,1)
(Mel,33,47,32,9,3)
(Jill,15,15,43,11,3)
(Joe,9,20,12,9,4)
(Mel,24,46,44,42,5)
(Jill,34,6,5,12,2)
grunt>
```

Question 6:

```
(Jill,34,6,5,12,2)
grunt> food_places = LOAD '/user/hadoop/foodplaces11799.txt' USING PigStorage(',') as (placeid:int, placename:chararray);
22/10/10 02:09:46 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Inste
grunt> DESCRIBE food_places;
food_places: {placeid: int,placename: chararray}
grunt> food_ratings_w_place_names = JOIN food_ratings BY placeid, food_places by placeid;
grunt> DECRIBE food_ratings_w_place_names;
```

```
Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  HdfsBytesRead HdfsBytesWritten Alias        Feature Outputs
scope-218          1          1           1000                   0           1000            112             9784          17461                0 food_ratings,food_ratings_w_place_
names
scope-219          1          1              5                   0              5            112              200             59                0 food_places,food_ratings_w_place_n
ames
scope-220          2          1              0                 194              6          16036              168              0                0 food_ratings_w_place_names,food_ra
tings_w_place_names_six HASH_JOIN
scope-222          1          1              6                   0              6            168                0              0              211 food_ratings_w_place_names_six   L
IMIT    hdfs://ip-172-31-33-233.ec2.internal:8020/tmp/temp-1740288465/tmp-963346877,

Input(s):
Successfully read 5 records (59 bytes) from: "/user/hadoop/foodplaces11799.txt"
Successfully read 1000 records (17461 bytes) from: "/user/hadoop/foodratings11799.txt"

Output(s):
Successfully stored 6 records (211 bytes) in: "hdfs://ip-172-31-33-233.ec2.internal:8020/tmp/temp-1740288465/tmp-963346877"

22/10/10 02:15:58 INFO input.FileInputFormat: Total input files to process : 1
2277359 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil  - Total input paths to process : 1
22/10/10 02:15:58 INFO util.MapRedUtil: Total input paths to process : 1
(Jill,9,36,8,8,1,1,China Bistro)
(Sam,40,19,41,38,1,1,China Bistro)
(Joe,14,33,17,32,1,1,China Bistro)
(Jill,33,1,11,45,1,1,China Bistro)
(Mel,12,8,50,41,1,1,China Bistro)
(Joe,26,37,10,49,1,1,China Bistro)
grunt>
```

```
food_ratings_2percent = SAMPLE food_ratings 0.02;
DESCRIBE food_ratings_2percent;

food_ratings_2percent_ten = LIMIT food_ratings_2percent 10;
DESCRIBE food_ratings_2percent_ten;
DUMP food_ratings_2percent_ten;
```

Question 7:

1. A. LIMIT
2. C. DISTINCT
3. B. (f1: STRING, f2: INT, f3: INT, f4: INT)
4. B.relB = FOREACH relA GENERATE $0, f3;
5. B. Data Flow
6. A. relB= FILTER relA by $0<20