Akash Tanwani A20448831

CSP554—Big Data Technologies

Assignment #5 (Modules 05)

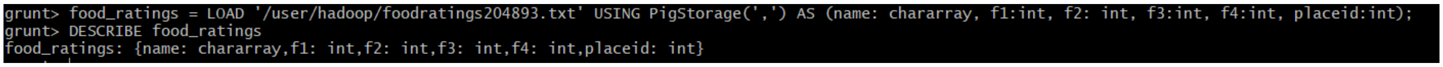
Worth: 15 points

Exercise 1) 2 points

Magic Number = 204893

Command:

food_ratings = LOAD '/home/hadoop/foodratings204893.txt' USING PigStorage(',') AS (name: chararray, f1:int, f2: int, f3:int, f4:int, placeid:int);
DESCRIBE food_ratings;

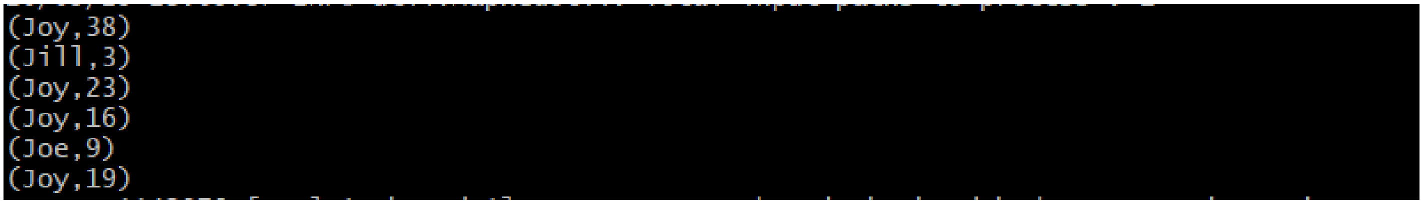


```
grunt> food_ratings = LOAD '/user/hadoop/foodratings204893.txt' USING PigStorage(',') AS (name: chararray, f1:int, f2: int, f3:int, f4:int, placeid:int);
grunt> DESCRIBE food_ratings
food_ratings: {name: chararray,f1: int,f2: int,f3: int,f4: int,placeid: int}
```

Exercise 2) 2 points

Command:

```
food_ratings_subset = FOREACH food_ratings GENERATE name,f4;
STORE food_ratings_subset INTO '/user/hadoop/fr_subset' USING PigStorage(',');
Result = LIMIT food_ratings_subset 6;
DUMP Result;
```

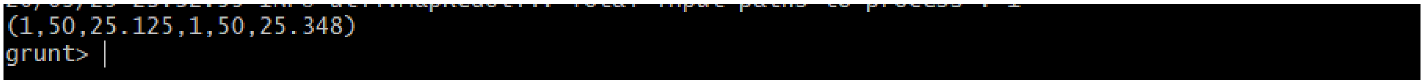


```
(Joy,38)
(Jill,3)
(Joy,23)
(Joy,16)
(Joe,9)
(Joy,19)
```

Exercise 3) 2 points

Command:

```
fr_relation = GROUP food_ratings ALL;
food_ratings_profile = FOREACH fr_relation GENERATE MIN(food_ratings.f2), MAX(food_ratings.f2),
AVG(food_ratings.f2), MIN(food_ratings.f3),MAX(food_ratings.f3), AVG(food_ratings.f3);
DUMP food_ratings_profile;
```

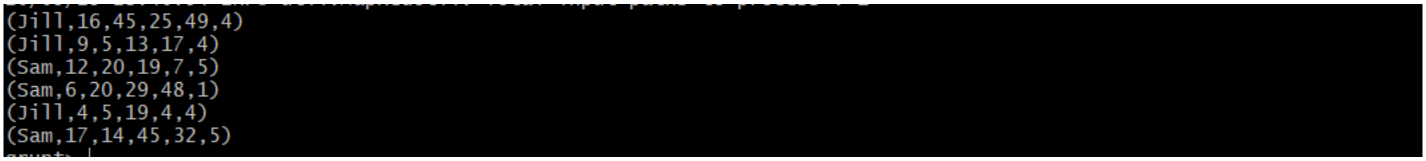


```
(1,50,25.125,1,50,25.348)
grunt> |
```

Exercise 4) 2 points

Command:

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



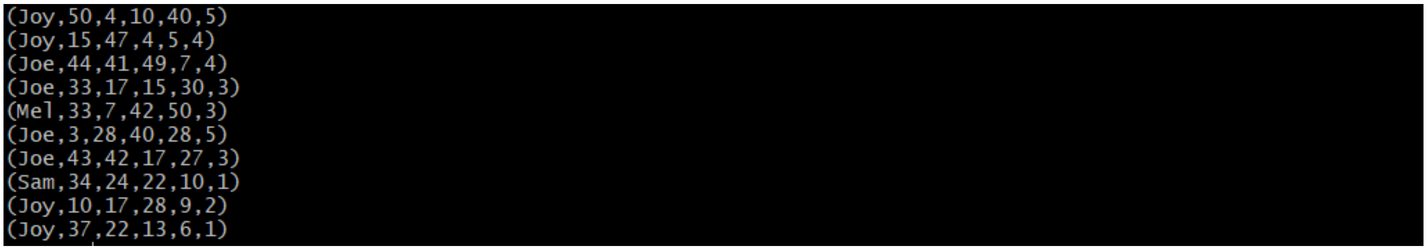
```
(Jill,16,45,25,49,4)
(Jill,9,5,13,17,4)
(Sam,12,20,19,7,5)
(Sam,6,20,29,48,1)
(Jill,4,5,19,4,4)
(Sam,17,14,45,32,5)
```

Exercise 5) 2 points

Command:

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

```
Result= LIMIT food_ratings_2percent 10;
DUMP Result;
```

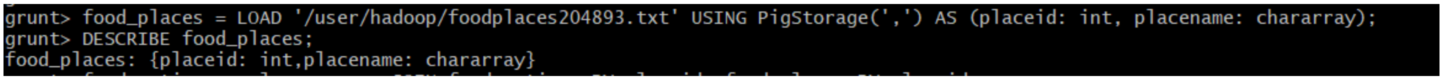


```
(Joy,50,4,10,40,5)
(Joy,15,47,4,5,4)
(Joe,44,41,49,7,4)
(Joe,33,17,15,30,3)
(Mel,33,7,42,50,3)
(Joe,3,28,40,28,5)
(Joe,43,42,17,27,3)
(Sam,34,24,22,10,1)
(Joy,10,17,28,9,2)
(Joy,37,22,13,6,1)
```

Exercise 6) 2 points

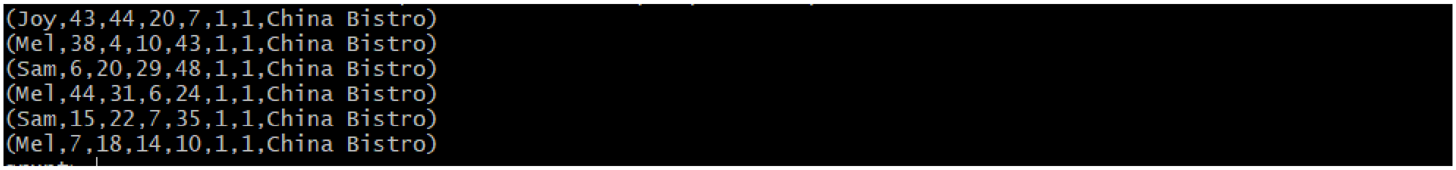
Command:

```
food_places = LOAD '/user/hadoop/foodplaces204893.txt' USING PigStorage(',') AS (placeid: int,
placename: chararray);
DESCRIBE food_places;
```



```
grunt> food_places = LOAD '/user/hadoop/foodplaces204893.txt' USING PigStorage(',') AS (placeid: int, placename: chararray);
grunt> DESCRIBE food_places;
food_places: {placeid: int,placename: chararray}
```

```
food_ratings_w_place_names= JOIN food_ratings BY placeid, food_places BY placeid;
Result= LIMIT food_ratings_w_place_names 6;
DUMP Result;
```



```
(Joy,43,44,20,7,1,1,China Bistro)
(Mel,38,4,10,43,1,1,China Bistro)
(Sam,6,20,29,48,1,1,China Bistro)
(Mel,44,31,6,24,1,1,China Bistro)
(Sam,15,22,7,35,1,1,China Bistro)
(Mel,7,18,14,10,1,1,China Bistro)
```

Exercise 7) (3 points) Identify the one correct answer for each the following questions. These questions are similar to the ones you might find on the mid-term covering Pig. Each is worth ½ point.

I. Which keyword is used to select a certain number of rows from a relation when forming a new relation?
Answer: **LIMIT**

Choices:

A. LIMIT
B. DISTINCT
C. UNIQUE
D. SAMPLE

II. Which keyword returns only unique rows for a relation when forming a new relation?
Answer: **DISTINCT**

Choices:

A. SAMPLE
B. FILTER
C. DISTINCT
D. SPLIT

III. Assume you have an HDFS file with a large number of records similar to the examples below

- Mel, 1, 2, 3

- Jill, 3, 4, 5

Which of the following would NOT be a correct pig schema for such a file?

Answer: **(f1: STRING, f2: INT, f3: INT, f4: INT)**

Choices:

A. (f1: CHARARRY, f2: INT, f3: INT, f4: INT)
B. (f1: STRING, f2: INT, f3: INT, f4: INT)
C. (f1, f2, f3, f4)
D. (f1: BYTEARRAY, f2: INT, f3: BYTEARRAY, f4: INT)

IV. Which one of the following statements would create a relation (relB) with two columns from a relation (relA) with 4 columns? Assume the pig schema for relA is as follows:
(f1: INT, f2, f3, f4: FLOAT)

Answer: **relB = FOREACH relA GENERATE $0, f3;**

Choices:

A. relB = GROUP relA GENERATE f1, f3;
B. relB = FOREACH relA GENERATE $0, f3;
C. relB = FOREACH relA GENERATE f1, f5;
D. relB = FOREACH relA SELECT f1, f3;

V. Pig Latin is a _______ language. Select the best choice to fill in the blank.
Answer: **data flow**

Choices:

A. functional
B. data flow
C. procedural
D. declarative

VI. Given a relation (relA) with 4 columns and pig schema as follows: (f1: INT, f2, f3, f4: FLOAT) which one statement will create a relation (relB) having records all of whose first field is less than 20

Answer: **relB = FILTER relA by $0 < 20**

Choices:

A. relB = FILTER relA by $0 < 20
B. relB = GROUP relA by f1 < 20
C. relB = FILTER relA by $1 < 20
D. relB = FOREACH relA GENERATE f1 < 20