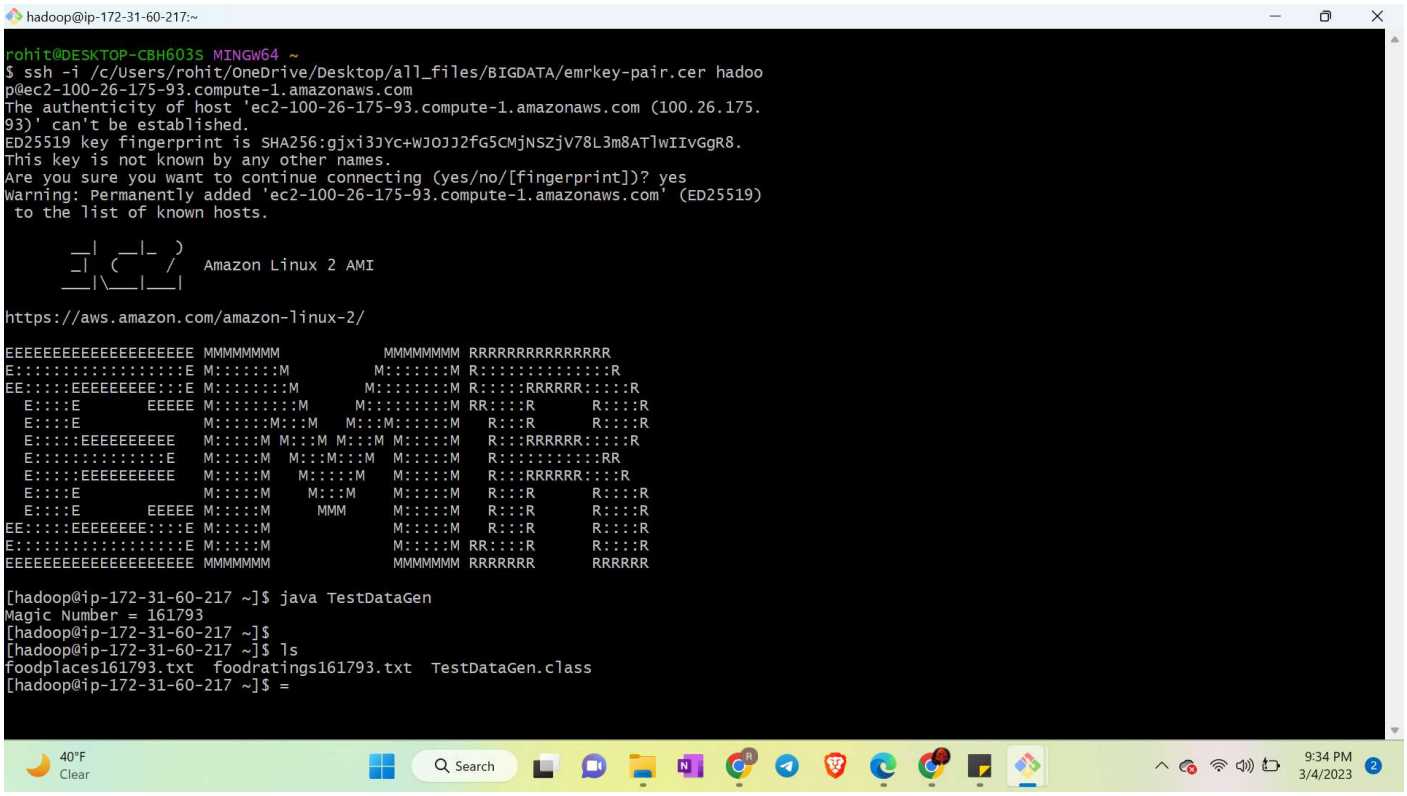
CWID: A20526972

CSP-554 Big Data Technologies

Bairi Rohith Reddy – Assignment 5

Magic Number :



Magic Number = 161793

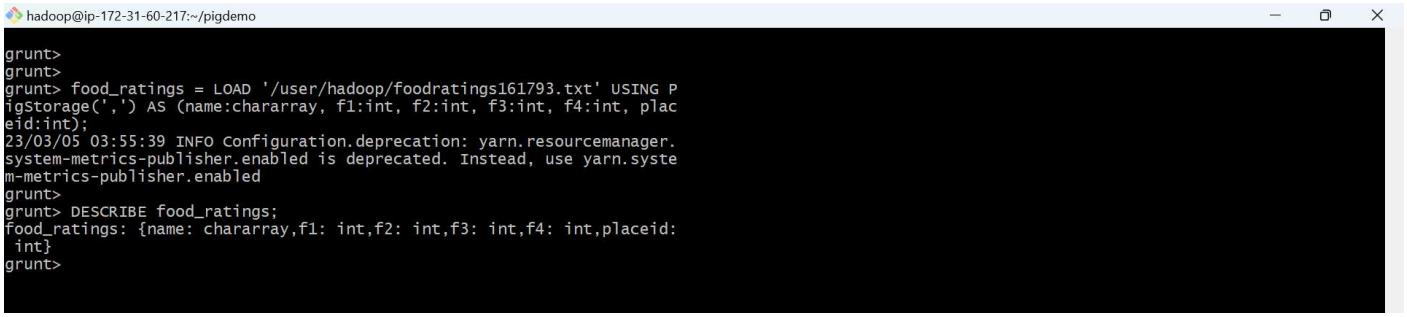
foodplaces161793.txt

foodratings161793.txt

Exercise 1)

food_ratings = LOAD '/user/hadoop/foodratings161793.txt' USING PigStorage(',') AS (name: chararray, f1:int, f2: int, f3:int, f4:int, placeid:int);

DESCRIBE food_ratings;



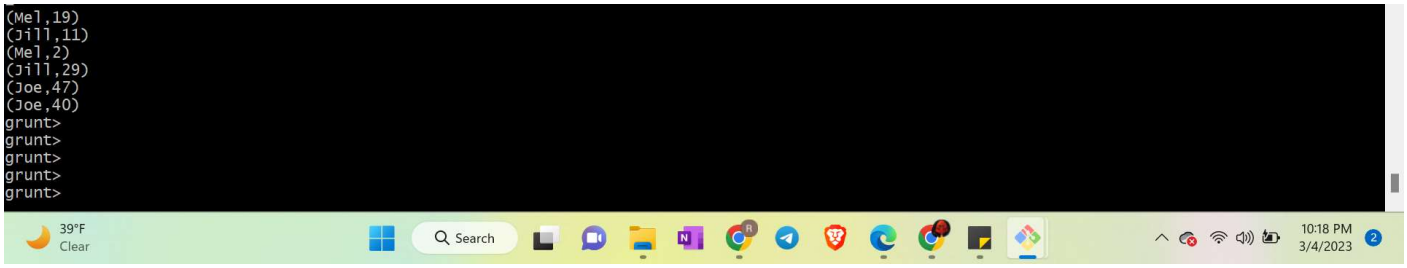
Exercise 2)

food_ratings_subset = FOREACH food_ratings GENERATE name,f4;

STORE food_ratings_subset INTO '/user/hadoop/fr_subset' USING PigStorage(',');

fr_output = LIMIT food_ratings_subset 6;

dump fr_output;



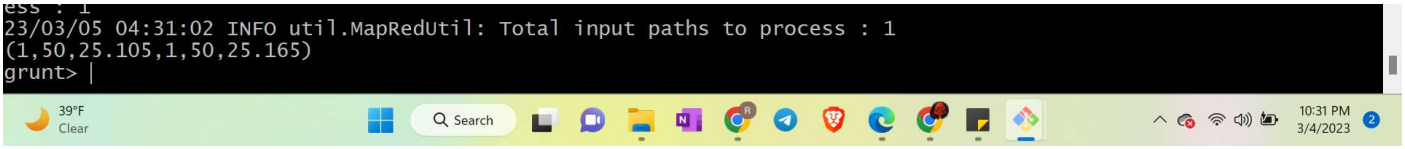
Exercise 3)

fr_profile = GROUP food_ratings ALL;

food_ratings_profile = FOREACH fr_profile 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;

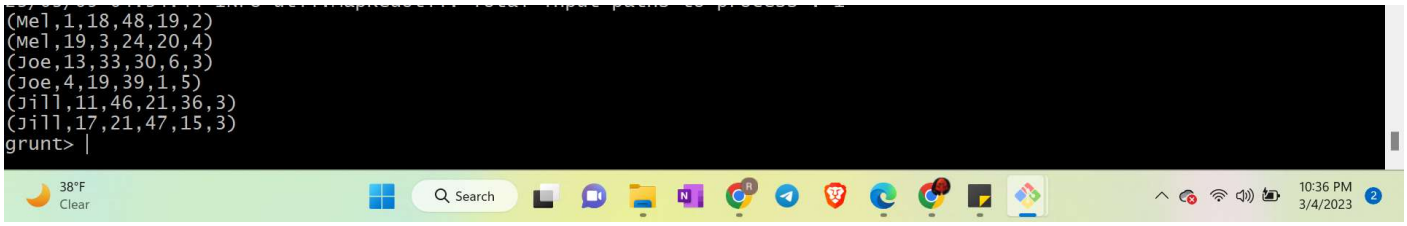


Exercise 4)

food_ratings_filtered = FILTER food_ratings BY (f1<20) AND (f3>5);

fr_filtered = LIMIT food_ratings_filtered 6;

DUMP fr_filtered;

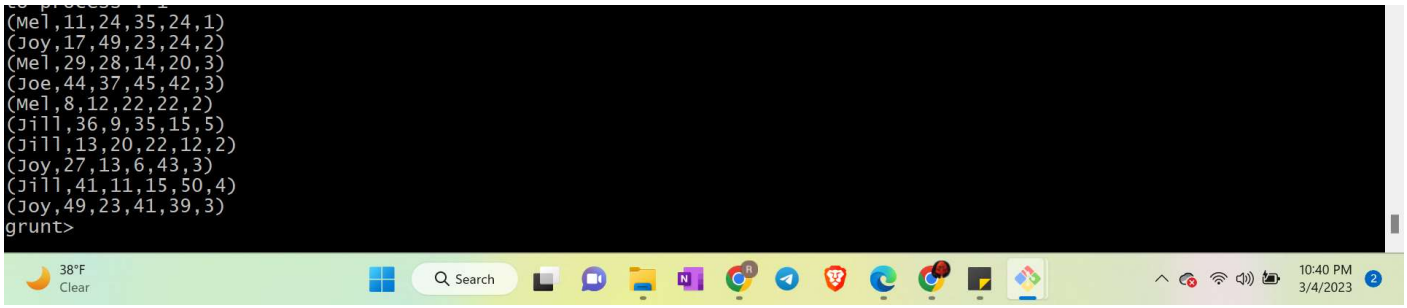


Exercise 5)

food_ratings_2percent = SAMPLE food_ratings 0.02;

filtered = LIMIT food_ratings_2percent 10;

DUMP filtered;



Exercise 6)

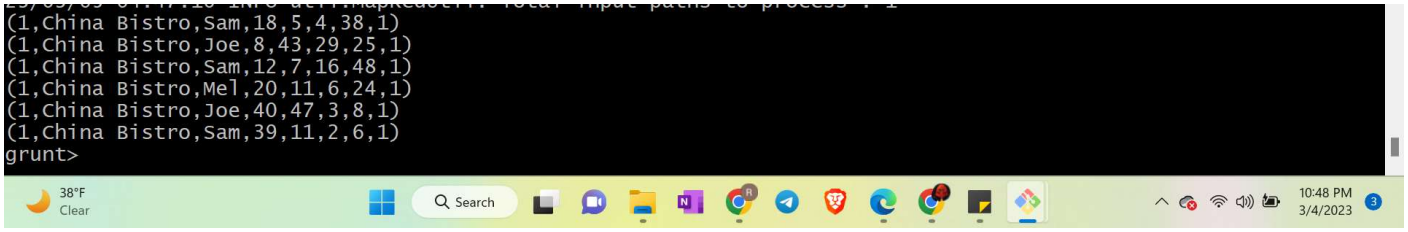
food_places = LOAD '/user/hadoop/foodplaces161793.txt' USING PigStorage(',') AS (placeid: int, placename: chararray);

DESCRIBE food_places;

food_ratings_w_place_names= JOIN food_ratings BY placeid, food_places BY placeid;

fr_result= LIMIT food_ratings_w_place_names 6;

DUMP fr_result;



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.

1. **Which keyword is used to select a certain number of rows from a relation when forming a new relation?**

   Answer: **LIMIT**

2. **Which keyword returns only unique rows for a relation when forming a new relation?**

   Answer: **DISTINCT**

3. **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, f2, f3, f4) is incorrect because a pig file cannot load data properly without explicitly mentioning of the data types for the columns.**

At the same time (f1:byte array, f2:integer, f3:byte array, f4:integer) also doesn't seem correct because the f3 column in the file contains an integer in decimal format but the data type used is byte array. This can cause errors while performing arithmetic operations as it pig has to convert this data using the ASCII values for storing and there is a possibility of errors while this is done.

But option A and B looks same and pig accepts Char array and String interchangeably.

4. **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;**

5. **Pig Latin is a _____ language. Select the best choice to fill in the blank.**

   Answer: **data flow**

6. **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**

Documentation of commands executed

Excercise 1:-

```
food_ratings = LOAD '/user/hadoop/foodratings161793.txt' USING PigStorage(',') AS (name:chararray, f1:int, f2:int, f3:int, f4:int, placeid:int);
DESCRIBE food_ratings;
```

Excercise 2:-

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

Excercise 3:-

```
fr_profile = GROUP food_ratings ALL;

food_ratings_profile = FOREACH fr_profile 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;
```

Excercise 4:-

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

fr_filtered = LIMIT food_ratings_filtered 6;

DUMP fr_filtered;
```

Excercise 5:-

```
food_ratings_2percent = SAMPLE food_ratings 0.02;

filtered = LIMIT food_ratings_2percent 10;

DUMP filtered;
```

Excercise 6:-

```
food_places = LOAD '/user/hadoop/foodplaces161793.txt' USING PigStorage(',') AS (placeid:int, placename:chararray);

DESCRIBE food_places;

food_ratings_w_place_names = JOIN food_places BY placeid, food_ratings BY placeid;

fr_result = LIMIT food_ratings_w_place_names 6;

DUMP fr_result;
```