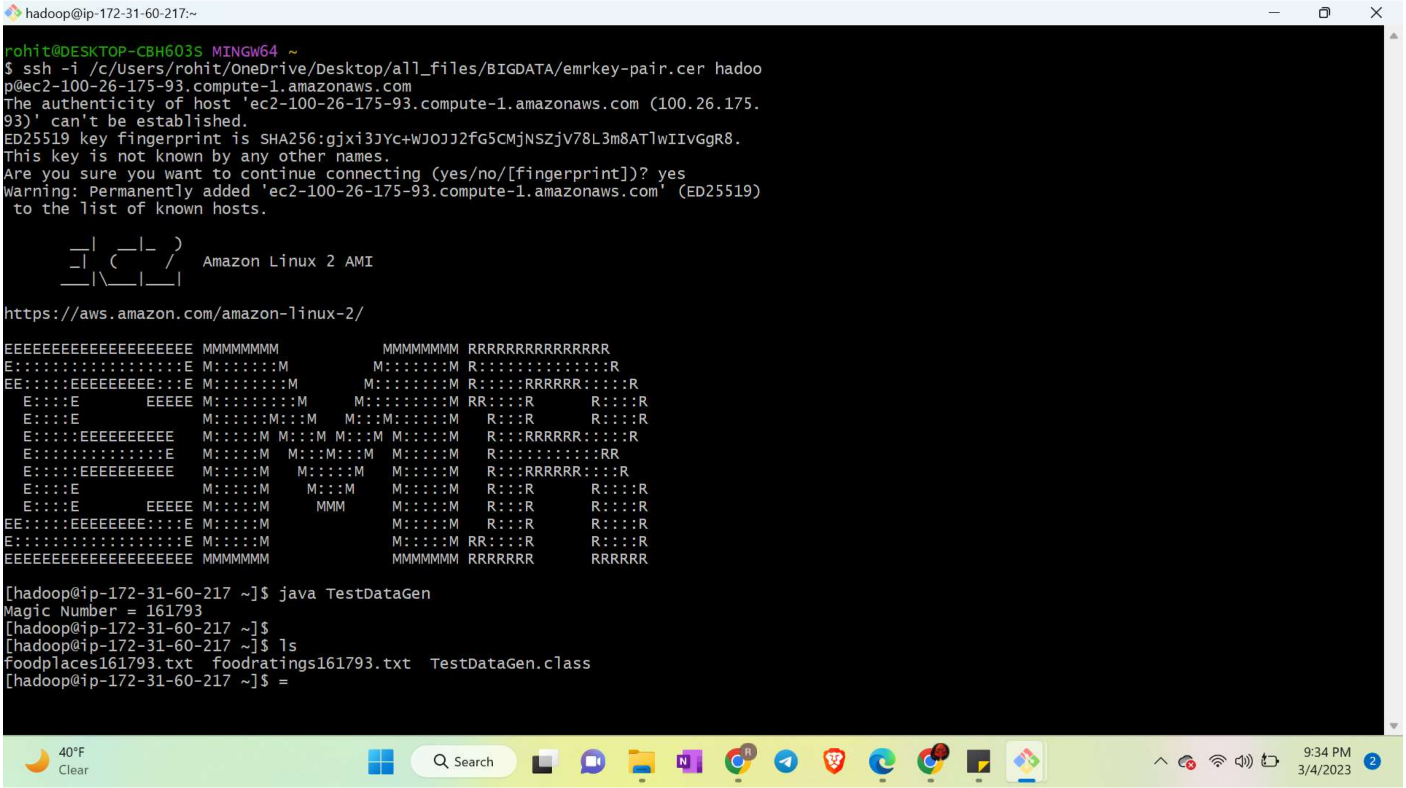
CWID: A20526972

CSP-554 Big Data Technologies

Bairi Rohith Reddy – Assignment 5

Magic Number :



```
rohit@DESKTOP-CBH603S MINGW64 ~
$ ssh -i /c/Users/rohit/OneDrive/Desktop/all_files/BIGDATA/emrkey-pair.cer hadoo
p@ec2-100-26-175-93.compute-1.amazonaws.com
The authenticity of host 'ec2-100-26-175-93.compute-1.amazonaws.com (100.26.175.
93)' can't be established.
ED25519 key fingerprint is SHA256:gjxi3JYc+WJOJJ2fG5CMjNSZjV78L3m8ATlwIIvGgR8.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'ec2-100-26-175-93.compute-1.amazonaws.com' (ED25519)
 to the list of known hosts.

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-60-217 ~]$ java TestDataGen
Magic Number = 161793
[hadoop@ip-172-31-60-217 ~]$
[hadoop@ip-172-31-60-217 ~]$ ls
foodplaces161793.txt  foodratings161793.txt  TestDataGen.class
[hadoop@ip-172-31-60-217 ~]$ =
```

Magic Number = 161793

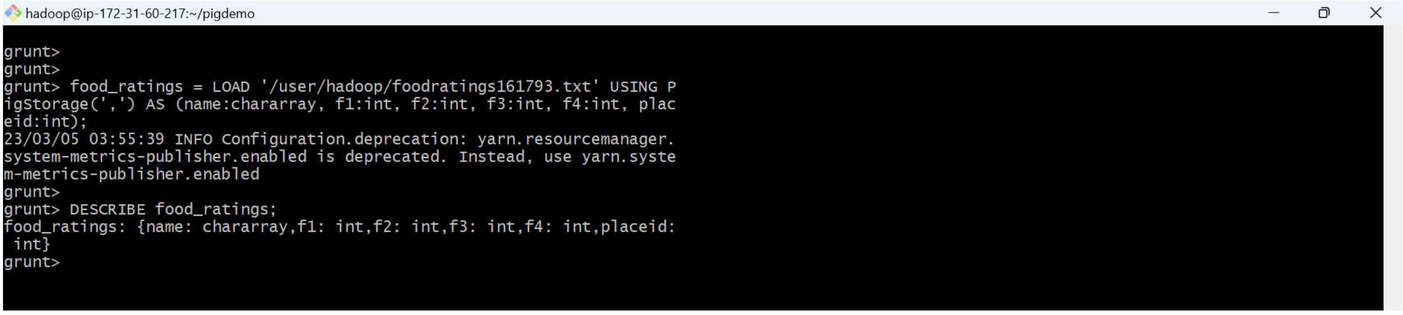
foodplaces161793.txt

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;



```
grunt>
grunt>
grunt> food_ratings = LOAD '/user/hadoop/foodratings161793.txt' USING P
igStorage(',') AS (name:chararray, f1:int, f2:int, f3:int, f4:int, plac
eid:int);
23/03/05 03:55:39 INFO Configuration.deprecation: yarn.resourcemanager.
system-metrics-publisher.enabled is deprecated. Instead, use yarn.syste
m-metrics-publisher.enabled
grunt>
grunt> DESCRIBE food_ratings;
food_ratings: {name: chararray,f1: int,f2: int,f3: int,f4: int,placeid:
 int}
grunt>
```

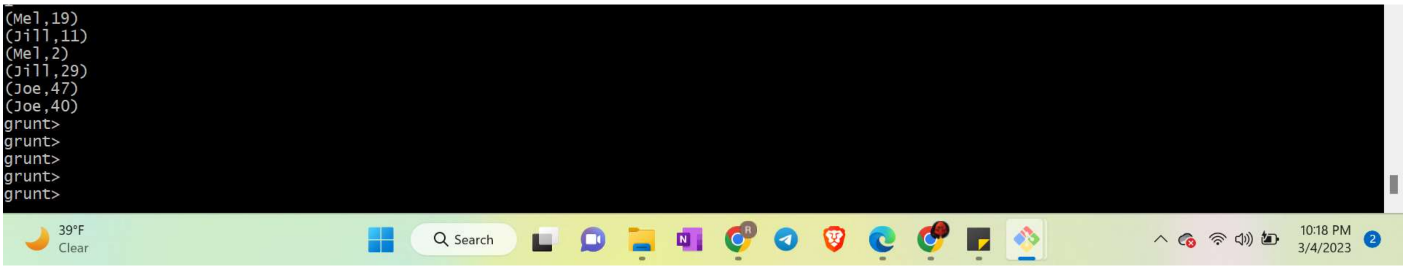
Exercise 2)

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;



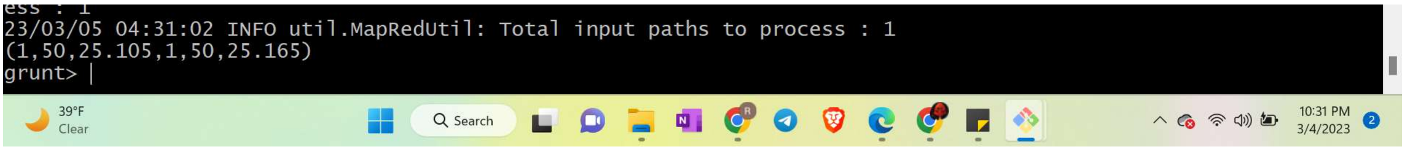
```
(Mel,19)
(Jill,11)
(Mel,2)
(Jill,29)
(Joe,47)
(Joe,40)
grunt>
grunt>
grunt>
grunt>
grunt>
```

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;



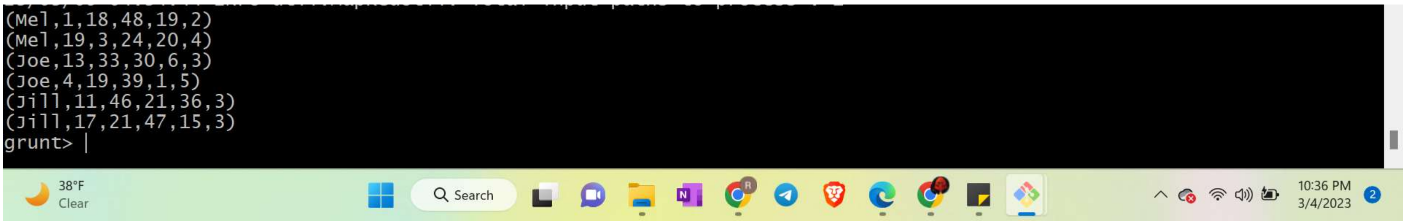
```
ess : 1
23/03/05 04:31:02 INFO util.MapRedUtil: Total input paths to process : 1
(1,50,25.105,1,50,25.165)
grunt> |
```

Exercise 4)

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

fr_filtered = LIMIT food_ratings_filtered 6;

DUMP fr_filtered;



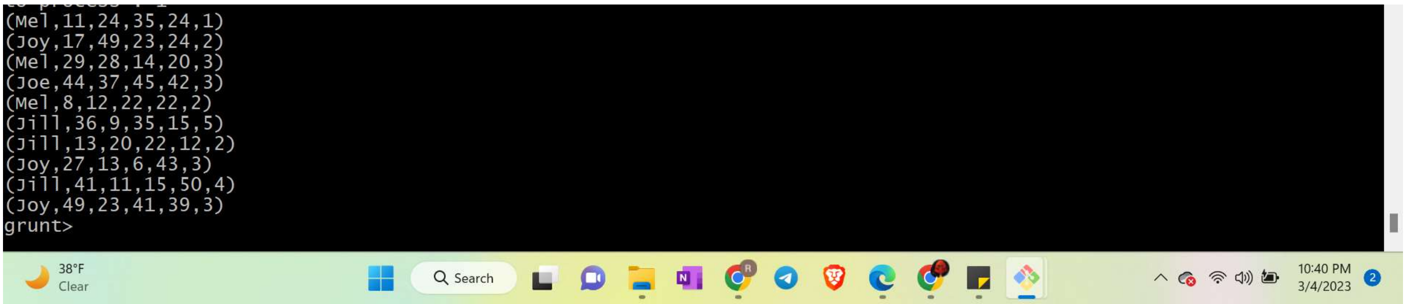
```
(Mel,1,18,48,19,2)
(Mel,19,3,24,20,4)
(Joe,13,33,30,6,3)
(Joe,4,19,39,1,5)
(Jill,11,46,21,36,3)
(Jill,17,21,47,15,3)
grunt> |
```

Exercise 5)

food_ratings_2percent = SAMPLE food_ratings 0.02;

filtered = LIMIT food_ratings_2percent 10;

DUMP filtered;

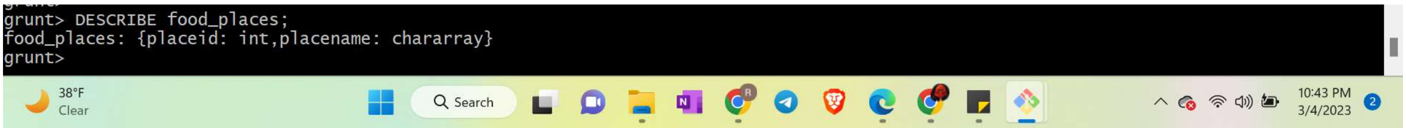


```
(Mel,11,24,35,24,1)
(Joy,17,49,23,24,2)
(Mel,29,28,14,20,3)
(Joe,44,37,45,42,3)
(Mel,8,12,22,22,2)
(Jill,36,9,35,15,5)
(Jill,13,20,22,12,2)
(Joy,27,13,6,43,3)
(Jill,41,11,15,50,4)
(Joy,49,23,41,39,3)
grunt>
```

Exercise 6)

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

DESCRIBE food_places;

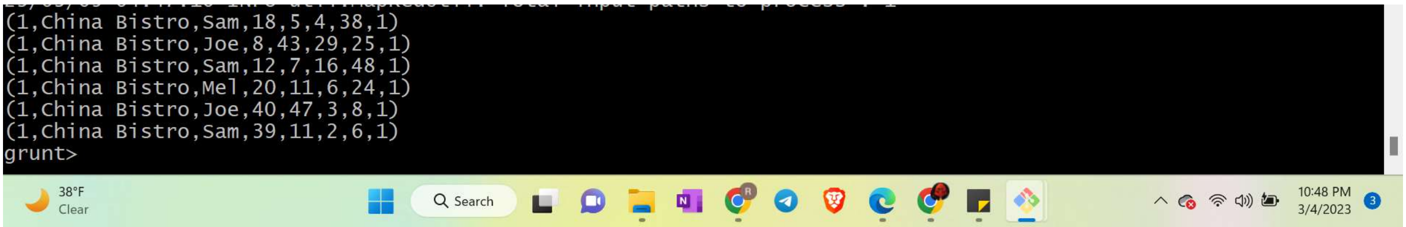


```
grunt> DESCRIBE food_places;
food_places: {placeid: int,placename: chararray}
grunt>
```

food_ratings_w_place_names= JOIN food_ratings BY placeid, food_places BY placeid;

fr_result= LIMIT food_ratings_w_place_names 6;

DUMP fr_result;



```
(1,China Bistro,Sam,18,5,4,38,1)
(1,China Bistro,Joe,8,43,29,25,1)
(1,China Bistro,Sam,12,7,16,48,1)
(1,China Bistro,Mel,20,11,6,24,1)
(1,China Bistro,Joe,40,47,3,8,1)
(1,China Bistro,Sam,39,11,2,6,1)
grunt>
```

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?**

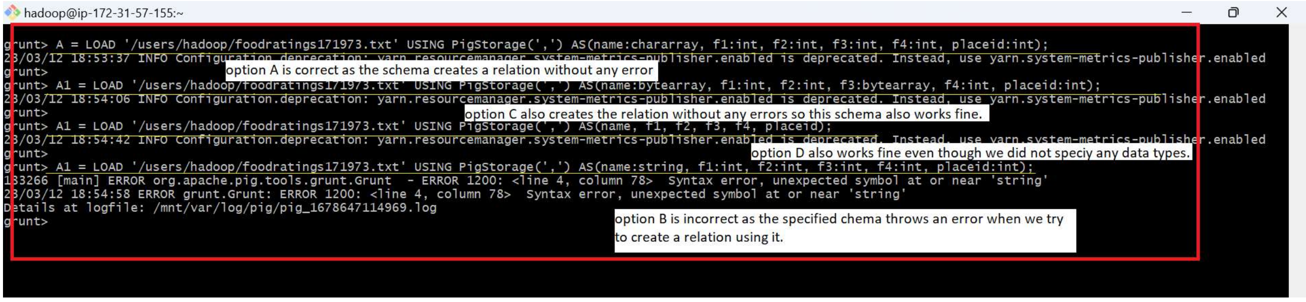
   Answer: **LIMIT**

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?**



hadoop@ip-172-31-57-155:~

```
grunt> A = LOAD '/users/hadoop/foodratings171973.txt' USING PigStorage(',') AS(name:chararray, f1:int, f2:int, f3:int, f4:int, placeid:int);
23/03/12 18:53:37 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt>
grunt> A1 = LOAD '/users/hadoop/foodratings171973.txt' USING PigStorage(',') AS(name:bytearray, f1:int, f2:int, f3:bytearray, f4:int, placeid:int);
23/03/12 18:54:06 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt>
grunt> A1 = LOAD '/users/hadoop/foodratings171973.txt' USING PigStorage(',') AS(name, f1, f2, f3, f4, placeid);
23/03/12 18:54:42 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt>
grunt> A1 = LOAD '/users/hadoop/foodratings171973.txt' USING PigStorage(',') AS(name:string, f1:int, f2:int, f3:int, f4:int, placeid:int);
133266 [main] ERROR org.apache.pig.tools.grunt.Grunt  - ERROR 1200: <line 4, column 78>  Syntax error, unexpected symbol at or near 'string'
23/03/12 18:54:58 ERROR grunt.Grunt: ERROR 1200: <line 4, column 78>  Syntax error, unexpected symbol at or near 'string'
Details at logfile: /mnt/var/log/pig/pig_1678647114969.log
grunt>
```

option A is correct as the schema creates a relation without any error

option C also creates the relation without any errors so this schema also works fine.

option D also works fine even though we did not speciy any data types.

option B is incorrect as the specified chema throws an error when we try to create a relation using it.

Answer: option B is incorrect as this throws an error when we try to create a relation with the given schema.

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;
```