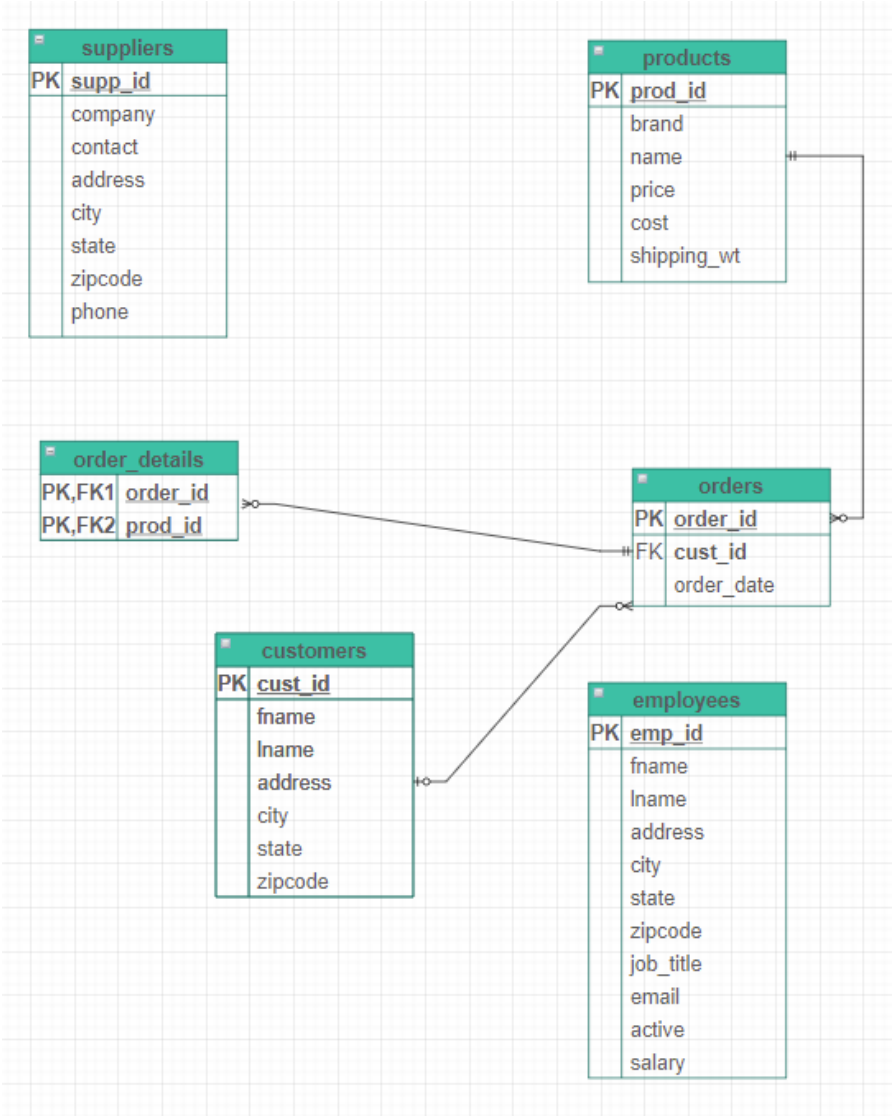
**Assignment 2: MapReduce Program Application**

COS20028 – Big data Architecture and Applications

Semester 2, 2021

Name: SYED OMAIR MAQDOOM MOHIUDDIN
Student ID: 102863768
Date: 30/10/2021

# Qestion 1 – ERD



# Question 2

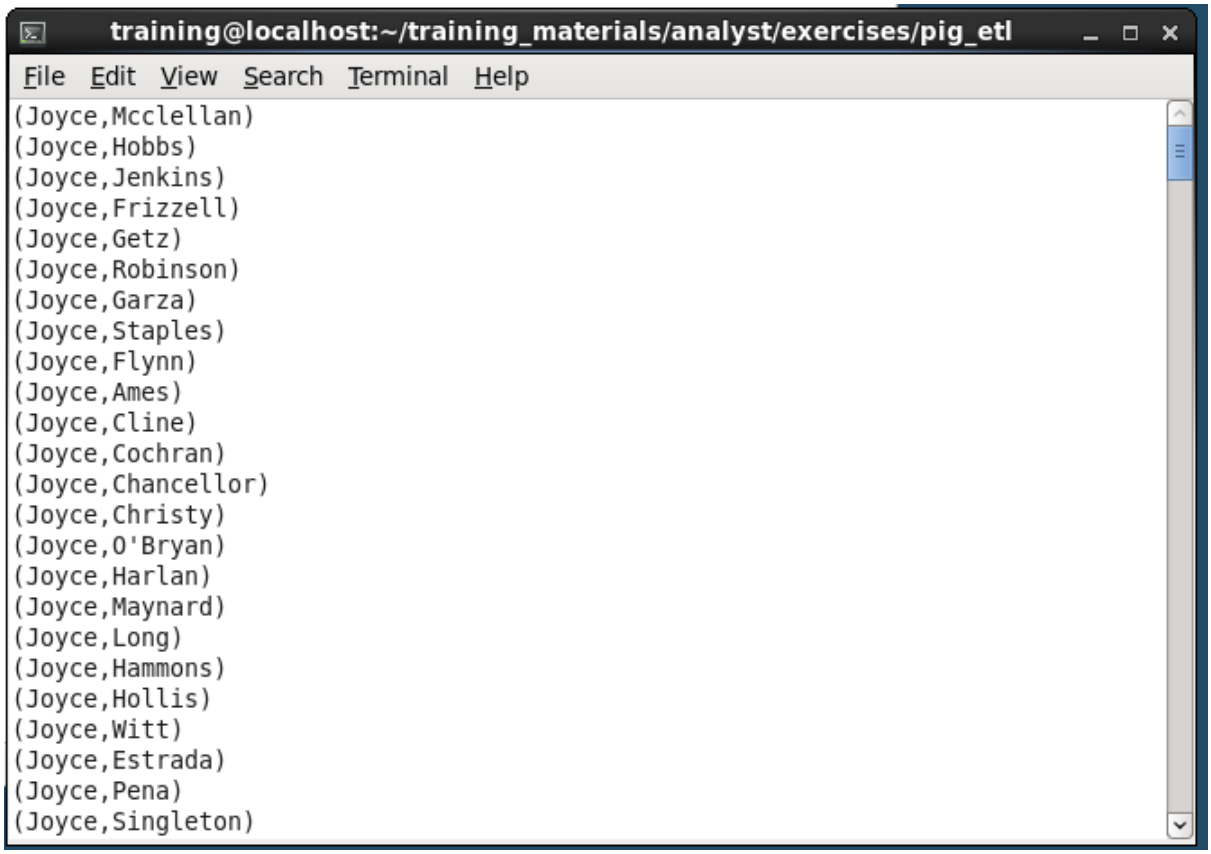
## a) Code:

2a.pig

```
data = LOAD '/dualcore/customers' AS (cust_id:int, fname:chararray, lname:chararray, address:chararray, city:chararray, state:chararray, zipcode:int);

data2 = FOREACH data GENERATE fname AS fname:chararray, lname AS lname:chararray;

data3 = ORDER data2 BY fname ASC;
DUMP data3;
```

## Result:

```
[training@localhost Assignment2]$ pig 2a.pig
2021-09-05 15:01:25,518 INFO org.apache.pig.Main: Apache Pig version 0.10.0-cdh4
.2.1 (rexported) compiled Apr 22 2013, 12:04:54
2021-09-05 15:01:25,519 INFO org.apache.pig.Main: Logging error messages to: /ho
me/training/training_materials/analyst/exercises/pig_etl/Assignment2/pig_1630868
485512.log
(Aaron,Ortega)
(Aaron,Parr)
(Aaron,Daniel)
(Aaron,Crawford)
(Aaron,Rios)
(Aaron,Bellows)
(Aaron,Darnell)
(Aaron,Hall)
(Aaron,Fairley)
(Aaron,Breckenridge)
(Aaron,Norman)
(Aaron,Matlock)
(Aaron,Lott)
(Aaron,Leyva)
(Aaron,Mercurio)
(Aaron,Gibbs)
```


training@localhost:~/training_materials/analyst/exercises/pig_etl
File Edit View Search Terminal Help
(Joyce,Mcclellan)
(Joyce,Hobbs)
(Joyce,Jenkins)
(Joyce,Frizzell)
(Joyce,Getz)
(Joyce,Robinson)
(Joyce,Garza)
(Joyce,Staples)
(Joyce,Flynn)
(Joyce,Ames)
(Joyce,Cline)
(Joyce,Cochran)
(Joyce,Chancellor)
(Joyce,Christy)
(Joyce,O'Bryan)
(Joyce,Harlan)
(Joyce,Maynard)
(Joyce,Long)
(Joyce,Hammons)
(Joyce,Hollis)
(Joyce,Witt)
(Joyce,Estrada)
(Joyce,Pena)
(Joyce,Singleton)

# b) Code:

Question2b.pig

```
data = LOAD '/dualcore/customers' AS (cust_id:int, fname:chararray, lname:chararray, address:chararray, city:chararray, state:chararray, zipcode:int);

data1 = FILTER data BY city == 'Louisville';


data2 = FOREACH data1 GENERATE fname AS fname:chararray,lname AS lname:chararray;
STORE data2 INTO 'dualcore/Customers_Louisville';
```

# Result:

```
[training@localhost Assignment2]$ pig Question2b.pig
2021-09-05 15:10:11,053 INFO org.apache.pig.Main: Apache Pig version 0.10.0-cdh4
.2.1 (rexported) compiled Apr 22 2013, 12:04:54
2021-09-05 15:10:11,054 INFO org.apache.pig.Main: Logging error messages to: /ho
me/training/training_materials/analyst/exercises/pig_etl/Assignment2/pig_1630869
011048.log
[training@localhost Assignment2]$
```

part-m-00000 (~/training_materials/analyst/exercises/pig_etl/Assignment2/Customers_Louisville) - gedit

File Edit View Search Tools Documents Help

Open Save Undo

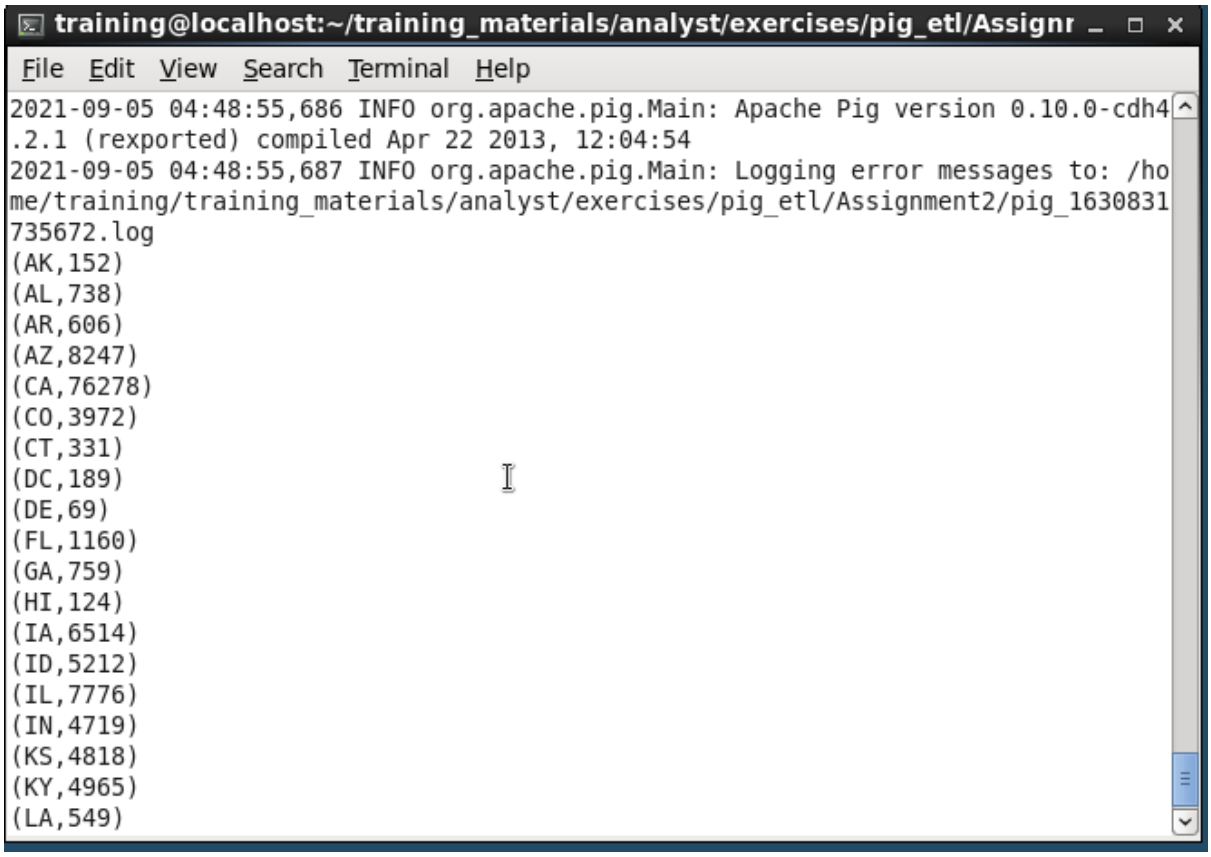
part-m-00000

```
Christie        Boudreaux
Federico        Wells
Jennifer        Cantu
Elizabeth       Roby
Darin   Dewitt
Mary    Brockway
Blaine  Javier
Shirley Dagostino
Traci   Hurd
Suzanne Simon
Leah    Younger
Joel    Hunt
Williams        Metz
Raymond Ingram
Nancy   Pena
Debra   Miller
Elizabeth       O'Neal
Marsha  Bouldin
Angelo  Gilliland
Mark    Nolan
```

# c) Code:

```
data = LOAD '/dualcore/customers' AS (cust_id:int, fname:chararray, lname:chararray, address:chararray, city:chararray, state:chararray, zipcode:int);

data1 = GROUP data BY state;

data2 = FOREACH data1 GENERATE group AS state, COUNT(data) as CustomerCount;

DUMP data2;
```

# Result:



# d) Code:

2d.pig

```
data = LOAD '/dualcore/products' AS (prod_id:int, brand:chararray, name:chararray, price:int, cost:int, shipping_wt:chararray);

data1 = DISTINCT data;

data2 = GROUP data1 BY brand;

data3 = FOREACH data2 GENERATE group AS brand;

data4 = ORDER data3 BY brand DESC;

STORE data4 INTO '/dualcore/products_brand';
```

# Result:

```
[training@localhost Assignment2]$ pig 2d.pig
2021-09-05 15:23:21,479 INFO org.apache.pig.Main: Apache Pig version 0.10.0-cdh4
.2.1 (rexported) compiled Apr 22 2013, 12:04:54
2021-09-05 15:23:21,480 INFO org.apache.pig.Main: Logging error messages to: /ho
me/training/training_materials/analyst/exercises/pig_etl/Assignment2/pig_1630869
801474.log
[training@localhost Assignment2]$
```

part-r-00000 (~/training_materials/analyst/exercises/pig_etl/Assignment2/products_brand) - gedit

File Edit View Search Tools Documents Help

Open Save Undo

part-r-00000

```
Yoyodyne
XYZ
Wolfpack
Whiteacre
Wernham
Weisenheimer
Weebits
United Digistuff
Ultramegaco
Tyrell
Tortoise
Texi
Terrapin Sands
TPS
SuperGamer
Sprite
Spindown
Sparky
Overtop
Orion
Olde-Gray
Megachango
McDowell
Lemmon
Krustybitz
Homertech
Gigabux
Foocorp
Electrosaurus
ElCheapo
Duff
Dualcore
```

# Question 3

## a) Code:

3a.pig

```
data = LOAD '/dualcore/orders' AS (order_id:int, cust_id:int, order_date:chararray);

data1 = LOAD '/dualcore/order_details' AS (order_id:int, prod_id:int);

data3 = FILTER data BY order_date MATCHES '^2013-05-.*$';

data4 = FILTER data1 BY prod_id == 1274348;

data5 = JOIN data1 BY order_id LEFT OUTER, data4 BY order_id;

data6 = FOREACH data5 GENERATE data1::order_id AS order_id:int, data1::prod_id AS  prod_id;

data7 = JOIN data3 BY order_id, data6 BY order_id;

data8 = FOREACH data7 GENERATE data3::order_id AS order_id:int, data6::prod_id AS prod_id:int;

data9 = DISTINCT data8;

data10 = GROUP data9 BY order_id;

data11 = FOREACH data10 GENERATE group, COUNT(data9.prod_id) AS Total:long;

data12 = GROUP data11 ALL;

data13 = FOREACH data12 GENERATE MAX(data11.Total);

DUMP data13;
```

## Result:

```
[training@localhost bonus_02]$ cd /home/training/training_materials/analyst/exer
cises/pig_etl/Assignment2/
[training@localhost Assignment2]$ pig 3a.pig
2021-09-05 09:46:11,555 INFO org.apache.pig.Main: Apache Pig version 0.10.0-cdh4
.2.1 (rexported) compiled Apr 22 2013, 12:04:54
2021-09-05 09:46:11,556 INFO org.apache.pig.Main: Logging error messages to: /ho
me/training/training_materials/analyst/exercises/pig_etl/Assignment2/pig_1630849
571549.log
(8)
[training@localhost Assignment2]$
```

# b) Code:

loyalty_program.pig

```
-- load the data sets
orders = LOAD '/dualcore/orders' AS (order_id:int,
             cust_id:int,
             order_dtm:chararray);

details = LOAD '/dualcore/order_details' AS (order_id:int,
             prod_id:int);

products = LOAD '/dualcore/products' AS (prod_id:int,
             brand:chararray,
             name:chararray,
             price:int,
             cost:int,
             shipping_wt:int);

orders_2012 = FILTER orders BY order_dtm MATCHES '^2012.*$';

grouping = GROUP orders BY cust_id;
orders_count = FOREACH grouping GENERATE group, COUNT(orders) AS num_orders;
orders_greater_6 = FILTER orders_count BY num_orders > 6;

orders_customer = JOIN orders_greater_6 BY group, orders_2012 BY cust_id;

order_details_customer = JOIN orders_customer BY orders_2012::order_id, details BY order_id;
order_product_customer = JOIN order_details_customer BY details::prod_id, products BY prod_id;

price_customer = FOREACH order_product_customer GENERATE orders_greater_6::group AS cust_id, products::price AS price;

customer_total_group = GROUP price_customer BY cust_id;

customer_total = FOREACH customer_total_group GENERATE group, SUM(price_customer.price)/100 AS Total;

SPLIT customer_total INTO
Platinum IF Total > 10000,
Gold IF Total >= 7000 AND Total < 10000,
Silver IF Total >= 5000 AND Total < 7000,
UnKnow IF Total <5000;

STORE Platinum INTO 'dualcore/loyalty/Platinum_Records';
STORE Gold INTO 'dualcore/loyalty/Gold_Records';
STORE Silver INTO 'dualcore/loyalty/Silver_Records';
STORE UnKnow INTO 'dualcore/loyalty/Unknow_Records';
```

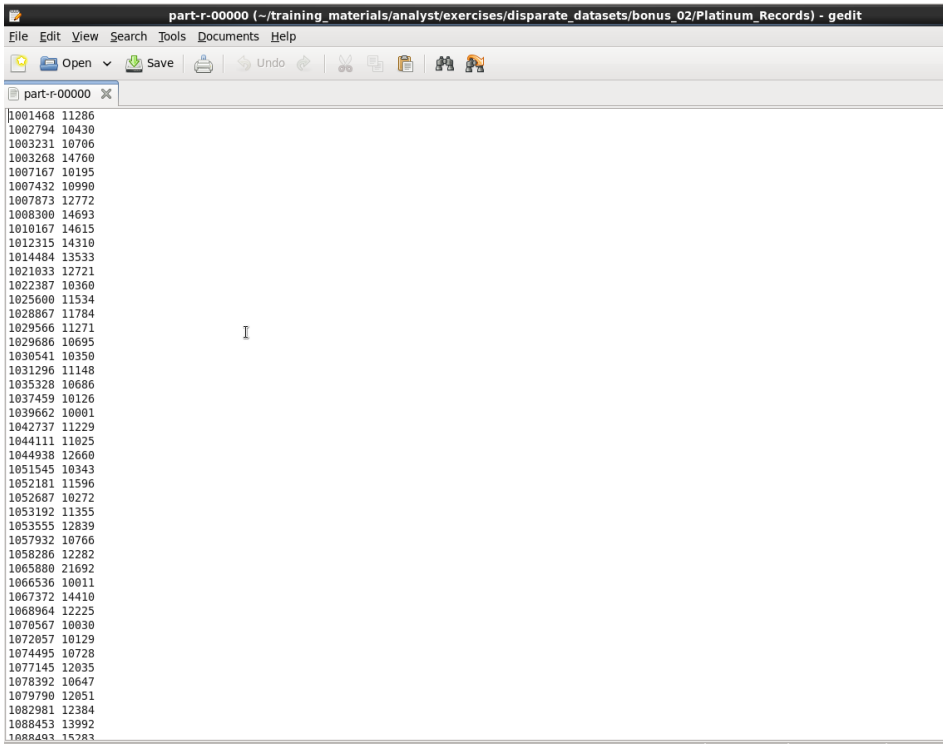
3b.pig

```
platinum = LOAD 'dualcore/loyalty/Platinum_Records/part-r-00000' AS (cust_id:int, price:long);
gold = LOAD 'dualcore/loyalty/Gold_Records/part-r-00000' AS (cust_id:int, price:long);
silver = LOAD 'dualcore/loyalty/Silver_Records/part-r-00000' AS (cust_id:int, price:long);
unknow = LOAD 'dualcore/loyalty/Unknow_Records/part-r-00000' AS (cust_id:int, price:long);

data = GROUP platinum ALL;
data1 = FOREACH data GENERATE 'Platinum_Count' AS groups:chararray, COUNT(platinum.cust_id) AS Total_Platinum:long;
DUMP data1;

data2 = GROUP gold ALL;
data3 = FOREACH data2 GENERATE 'Gold_Count' AS groups:chararray, COUNT(gold.cust_id) AS Total_Gold:long;
DUMP data3;

data4 = GROUP silver ALL;
data5 = FOREACH data4 GENERATE 'Siver_Count' As groups:chararray, COUNT(silver.cust_id) AS Total_Silver:long;
DUMP data5;

data6 = GROUP unknow ALL;
data7 = FOREACH data6 GENERATE 'Unknown' As groups:chararray, COUNT(unknow.cust_id) AS Total_Unknow:long;
DUMP data7;
```

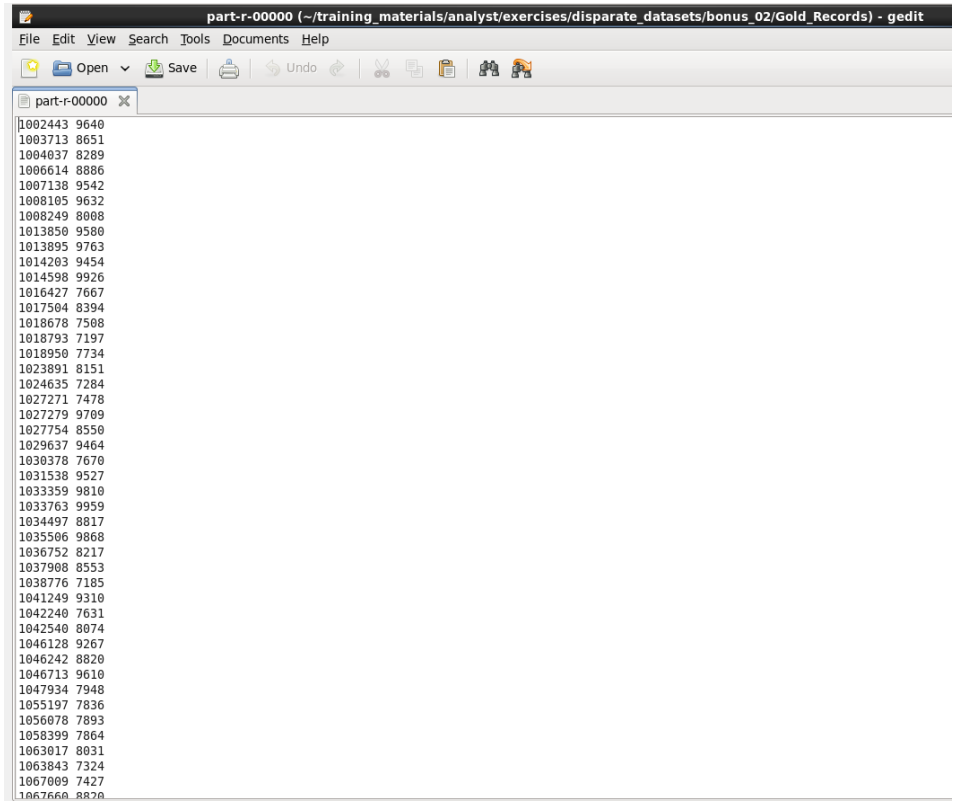
# Result:

```
[training@localhost bonus_02]$ pig 3b.pig
2021-09-06 16:44:03,474 INFO org.apache.pig.Main: Apache Pig version 0.10.0-cdh4
.2.1 (rexported) compiled Apr 22 2013, 12:04:54
2021-09-06 16:44:03,475 INFO org.apache.pig.Main: Logging error messages to: /ho
me/training/training_materials/analyst/exercises/disparate_datasets/bonus_02/pig
_1630961043469.log
(Platinum_Count,109)
(Gold_Count,156)
(Siver_Count,281)
(Unknown,110497)
[training@localhost bonus_02]$
```

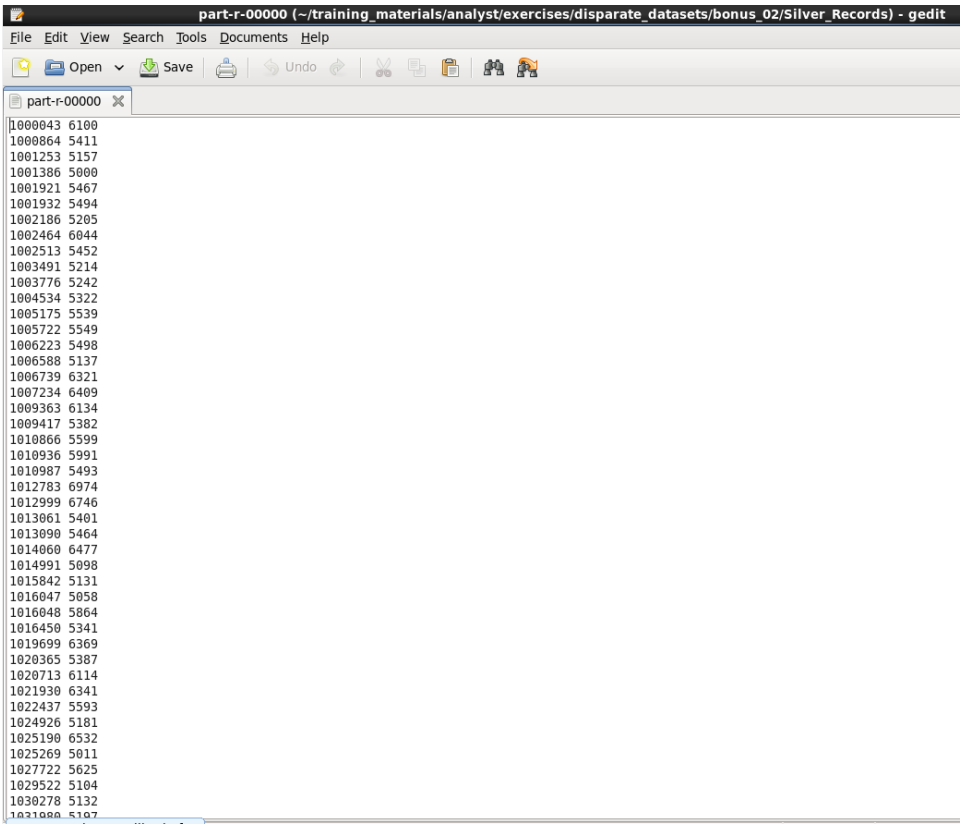
# Platinum:



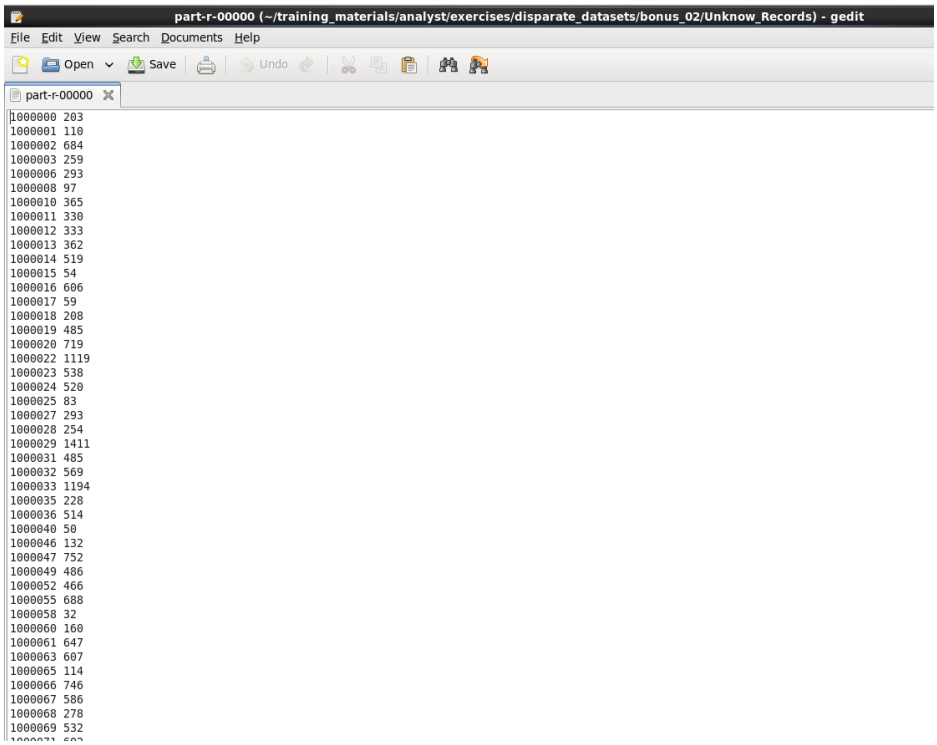

# Gold:



# Silver:




# Unknown:

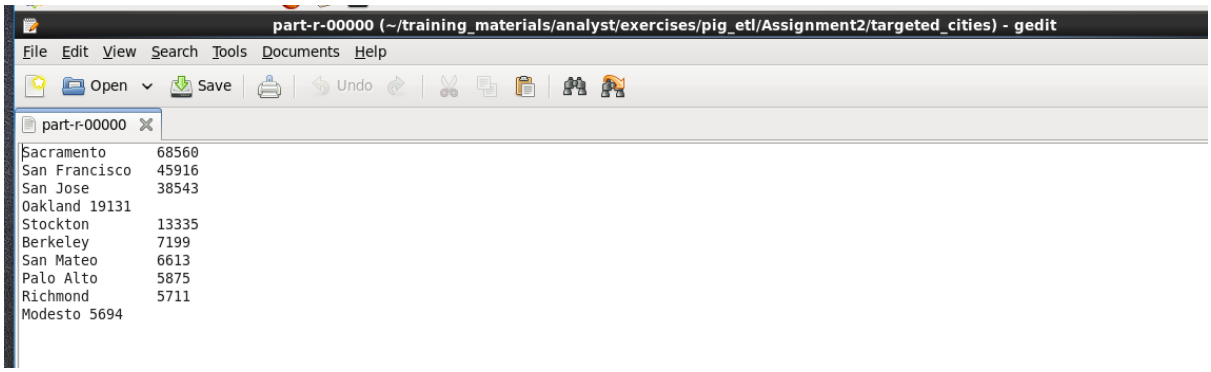


# c) Code:

top_ten_cities_customers.pig

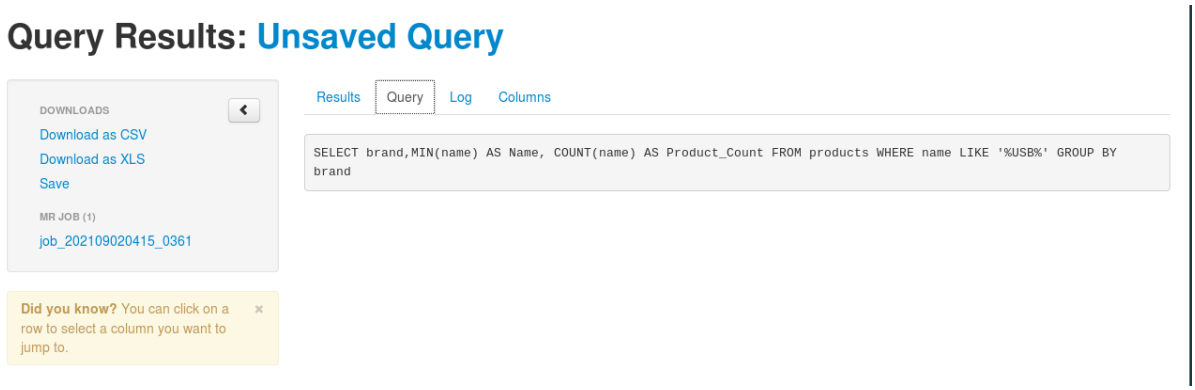
```
data = LOAD '/dualcore/customers' AS (cust_id:int, fname:chararray, lname:chararray, address:chararray, city:chararray, state:chararray, zipcode:int);

data1 = LOAD '/dualcore/orders' AS (order_id:int, cust_id:int, order_date:chararray);

data2 = DISTINCT data1;

data3 = JOIN data BY cust_id, data2 BY cust_id;

data4 = GROUP data3 BY data::city;

data5 = FOREACH data4 GENERATE group, COUNT(data3) AS Total:long;

data6 = ORDER data5 BY Total DESC;

data7 = LIMIT data6 10;

STORE data7 INTO 'dualcore/targeted_cities';
```

# Result:



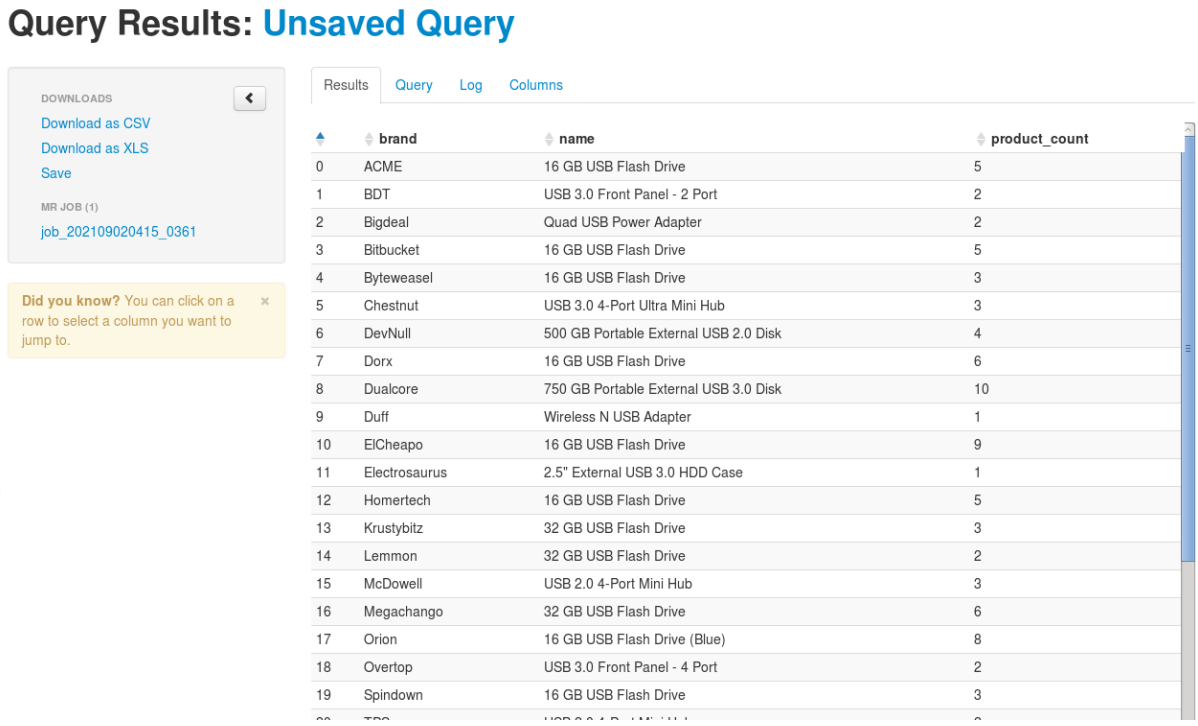
# Question 4

## a) Code:





## Result :




| | brand | name | product_count |
|---|---|---|---|
| 0 | ACME | 16 GB USB Flash Drive | 5 |
| 1 | BDT | USB 3.0 Front Panel - 2 Port | 2 |
| 2 | Bigdeal | Quad USB Power Adapter | 2 |
| 3 | Bitbucket | 16 GB USB Flash Drive | 5 |
| 4 | Byteweasel | 16 GB USB Flash Drive | 3 |
| 5 | Chestnut | USB 3.0 4-Port Ultra Mini Hub | 3 |
| 6 | DevNull | 500 GB Portable External USB 2.0 Disk | 4 |
| 7 | Dorx | 16 GB USB Flash Drive | 6 |
| 8 | Dualcore | 750 GB Portable External USB 3.0 Disk | 10 |
| 9 | Duff | Wireless N USB Adapter | 1 |
| 10 | ElCheapo | 16 GB USB Flash Drive | 9 |
| 11 | Electrosaurus | 2.5" External USB 3.0 HDD Case | 1 |
| 12 | Homertech | 16 GB USB Flash Drive | 5 |
| 13 | Krustybitz | 32 GB USB Flash Drive | 3 |
| 14 | Lemmon | 32 GB USB Flash Drive | 2 |
| 15 | McDowell | USB 2.0 4-Port Mini Hub | 3 |
| 16 | Megachango | 32 GB USB Flash Drive | 6 |
| 17 | Orion | 16 GB USB Flash Drive (Blue) | 8 |
| 18 | Overtop | USB 3.0 Front Panel - 4 Port | 2 |
| 19 | Spindown | 16 GB USB Flash Drive | 3 |

## Query Results: Unsaved Query

DOWNLOADS

Download as CSV

Download as XLS

Save

MR JOB (1)

job_202109020415_0495

Results | Query | Log | Columns

| | brand_count |
|---|---|
| 0 | 29 |

# b) Code:

## Query Results: Unsaved Query

DOWNLOADS

Download as CSV

Download as XLS

Save

MR JOBS (2)

job_202109020415_0366

job_202109020415_0367

Did you know? You can click on a row to select a column you want to jump to.

Results | Query | Log | Columns

```
SELECT COUNT(c.cust_id) AS Count_of_Orders,MIN(fname) AS Fname,MIN(lname)
AS Lname FROM customers c JOIN orders o ON (c.cust_id = o.cust_id)
WHERE fname LIKE 'Diana' AND lname LIKE 'Lamb'
```

# Result:

## Query Results: Unsaved Query

DOWNLOADS

Download as CSV

Download as XLS

Save

MR JOBS (2)

job_202109020415_0366

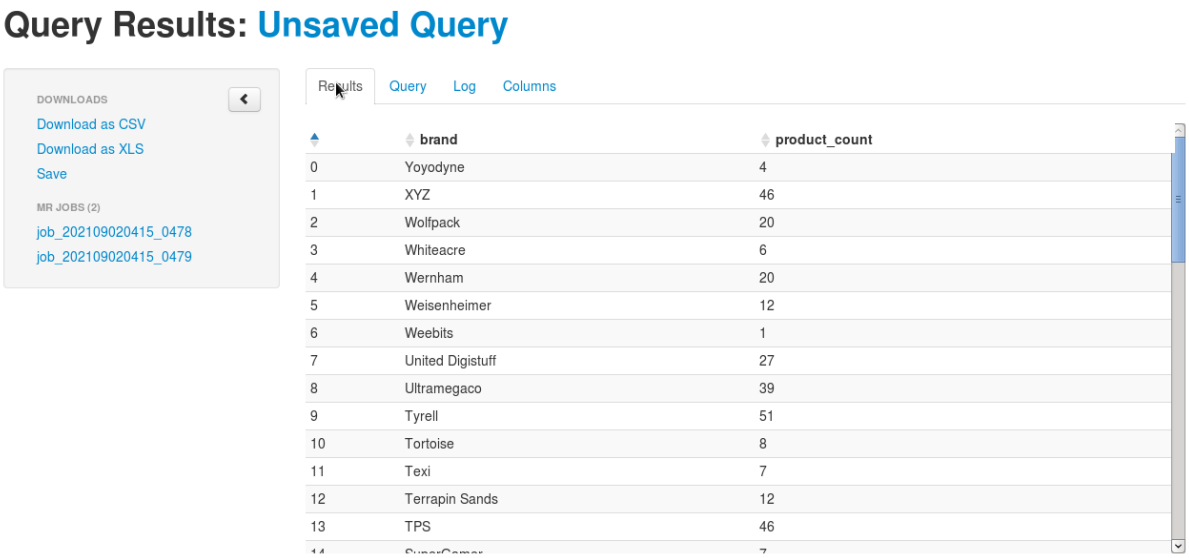
job_202109020415_0367

Results | Query | Log | Columns

| | count_of_orders | fname | lname |
|---|---|---|---|
| 0 | 10 | Diana | Lamb |

# c) Code:

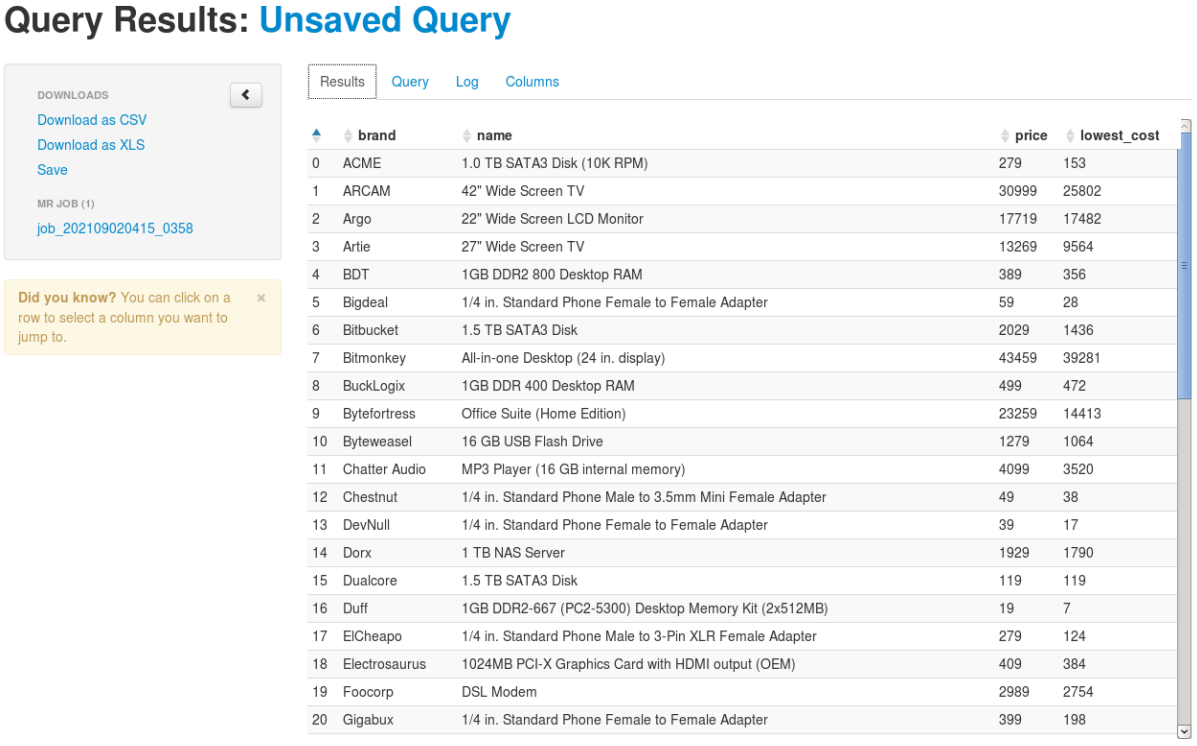

# Result:




| | brand | product_count |
|---|---|---|
| 0 | Yoyodyne | 4 |
| 1 | XYZ | 46 |
| 2 | Wolfpack | 20 |
| 3 | Whiteacre | 6 |
| 4 | Wernham | 20 |
| 5 | Weisenheimer | 12 |
| 6 | Weebits | 1 |
| 7 | United Digistuff | 27 |
| 8 | Ultramegaco | 39 |
| 9 | Tyrell | 51 |
| 10 | Tortoise | 8 |
| 11 | Texi | 7 |
| 12 | Terrapin Sands | 12 |
| 13 | TPS | 46 |

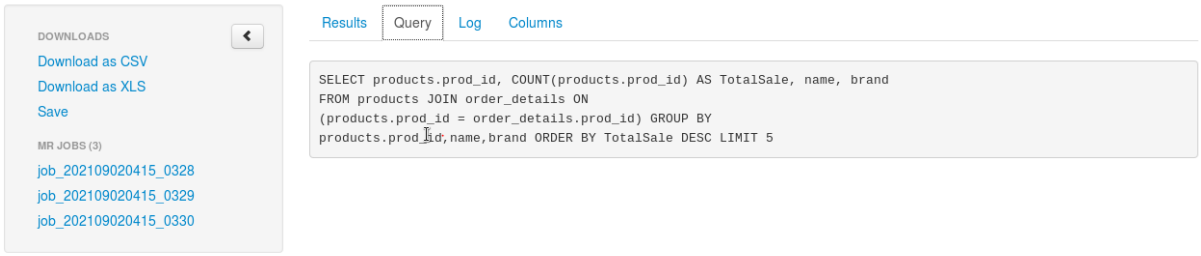
# d) Code:


# Result:

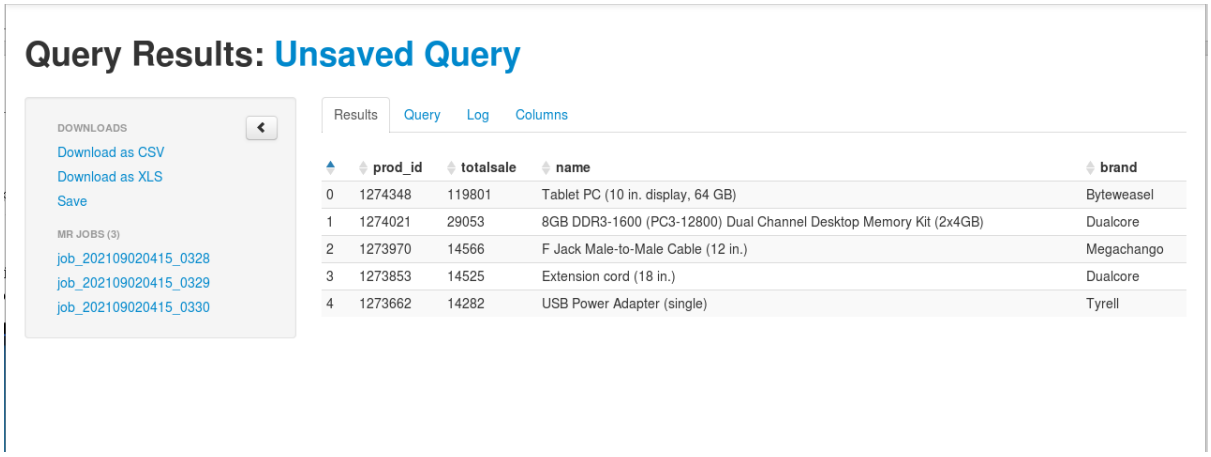



| | brand | name | price | lowest_cost |
|---|---|---|---|---|
| 0 | ACME | 1.0 TB SATA3 Disk (10K RPM) | 279 | 153 |
| 1 | ARCAM | 42" Wide Screen TV | 30999 | 25802 |
| 2 | Argo | 22" Wide Screen LCD Monitor | 17719 | 17482 |
| 3 | Artie | 27" Wide Screen TV | 13269 | 9564 |
| 4 | BDT | 1GB DDR2 800 Desktop RAM | 389 | 356 |
| 5 | Bigdeal | 1/4 in. Standard Phone Female to Female Adapter | 59 | 28 |
| 6 | Bitbucket | 1.5 TB SATA3 Disk | 2029 | 1436 |
| 7 | Bitmonkey | All-in-one Desktop (24 in. display) | 43459 | 39281 |
| 8 | BuckLogix | 1GB DDR 400 Desktop RAM | 499 | 472 |
| 9 | Bytefortress | Office Suite (Home Edition) | 23259 | 14413 |
| 10 | Byteweasel | 16 GB USB Flash Drive | 1279 | 1064 |
| 11 | Chatter Audio | MP3 Player (16 GB internal memory) | 4099 | 3520 |
| 12 | Chestnut | 1/4 in. Standard Phone Male to 3.5mm Mini Female Adapter | 49 | 38 |
| 13 | DevNull | 1/4 in. Standard Phone Female to Female Adapter | 39 | 17 |
| 14 | Dorx | 1 TB NAS Server | 1929 | 1790 |
| 15 | Dualcore | 1.5 TB SATA3 Disk | 119 | 119 |
| 16 | Duff | 1GB DDR2-667 (PC2-5300) Desktop Memory Kit (2x512MB) | 19 | 7 |
| 17 | ElCheapo | 1/4 in. Standard Phone Male to 3-Pin XLR Female Adapter | 279 | 124 |
| 18 | Electrosaurus | 1024MB PCI-X Graphics Card with HDMI output (OEM) | 409 | 384 |
| 19 | Foocorp | DSL Modem | 2989 | 2754 |
| 20 | Gigabux | 1/4 in. Standard Phone Female to Female Adapter | 399 | 198 |

# e) Code:



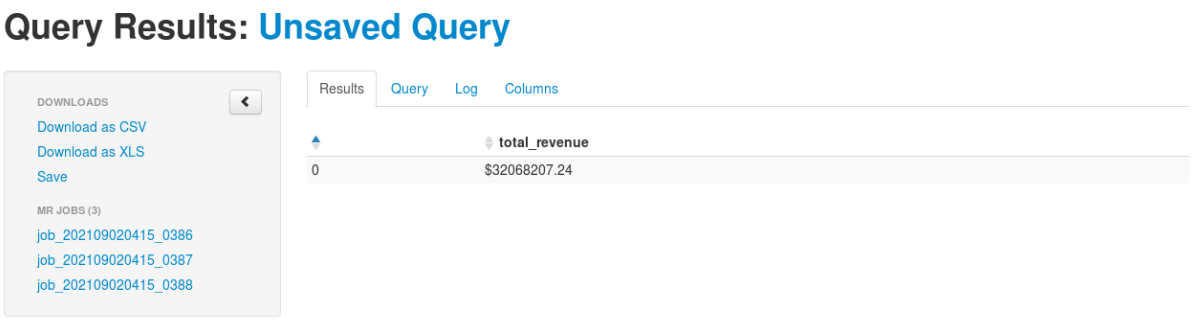

# Result:



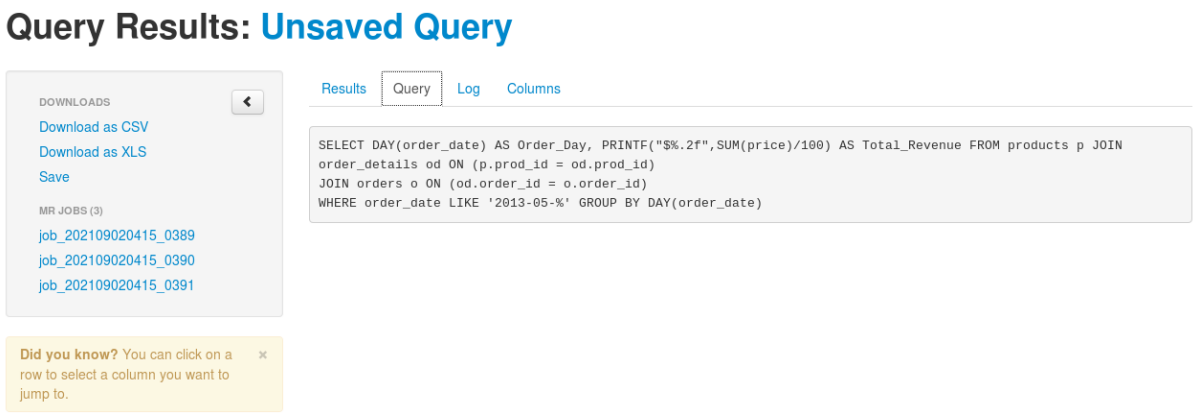

| | prod_id | totalsale | name | brand |
|---|---|---|---|---|
| 0 | 1274348 | 119801 | Tablet PC (10 in. display, 64 GB) | Byteweasel |
| 1 | 1274021 | 29053 | 8GB DDR3-1600 (PC3-12800) Dual Channel Desktop Memory Kit (2x4GB) | Dualcore |
| 2 | 1273970 | 14566 | F Jack Male-to-Male Cable (12 in.) | Megachango |
| 3 | 1273853 | 14525 | Extension cord (18 in.) | Dualcore |
| 4 | 1273662 | 14282 | USB Power Adapter (single) | Tyrell |

# f) Code:

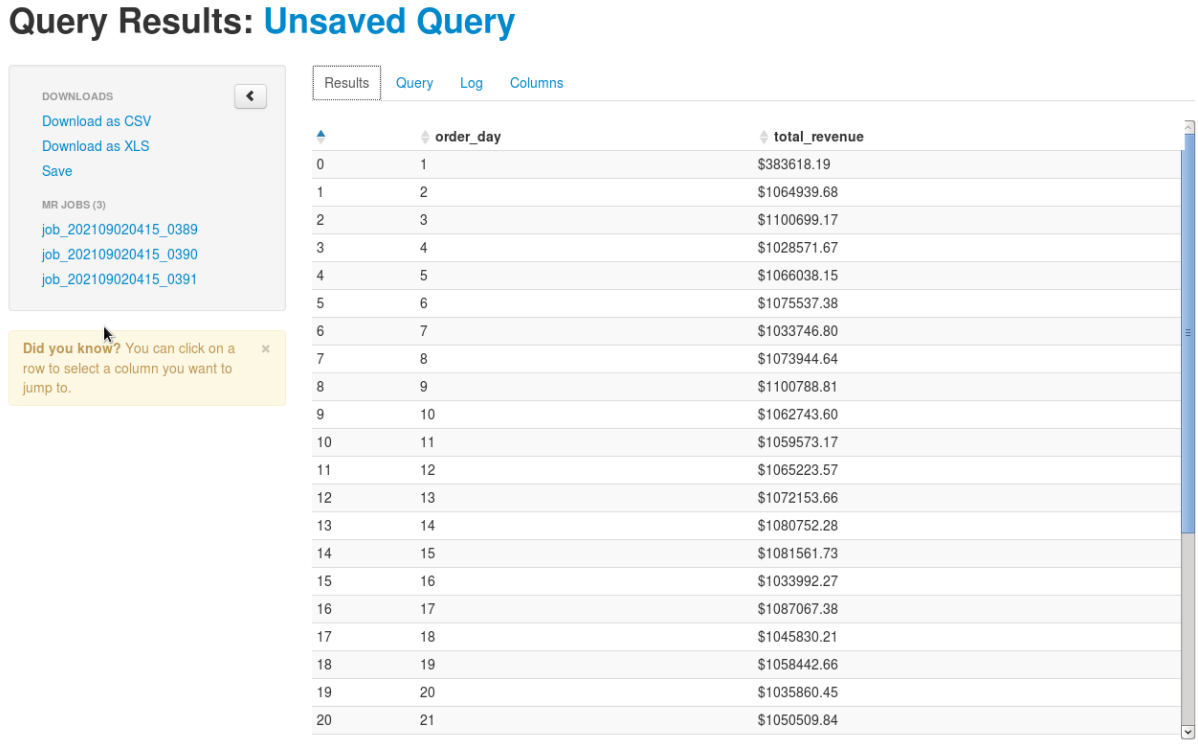

# Result:



# g) Code:

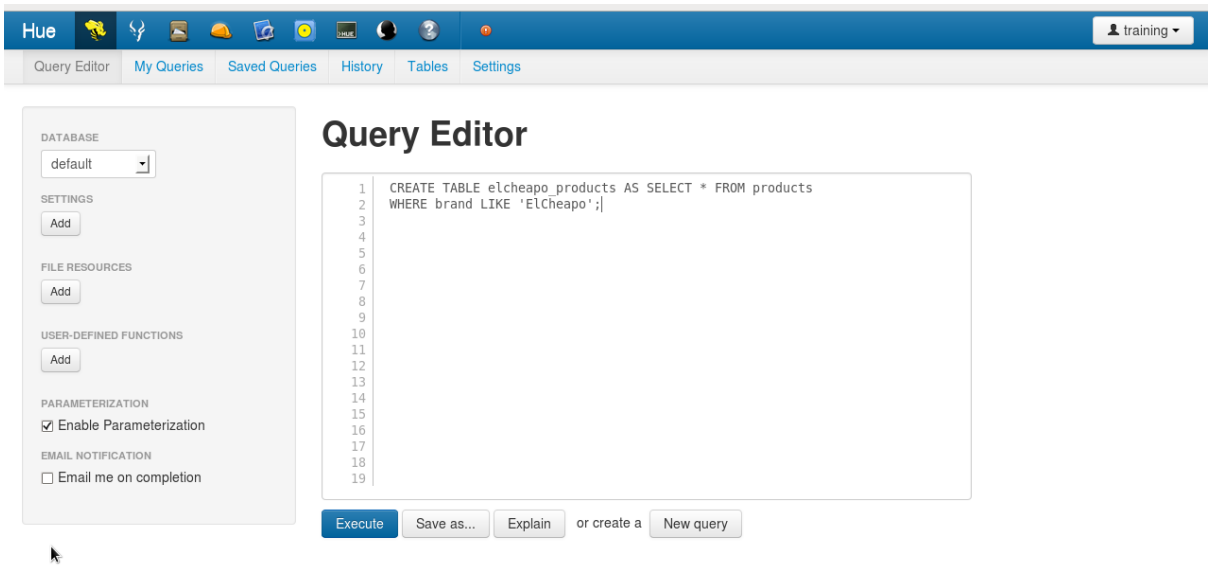



# Result:

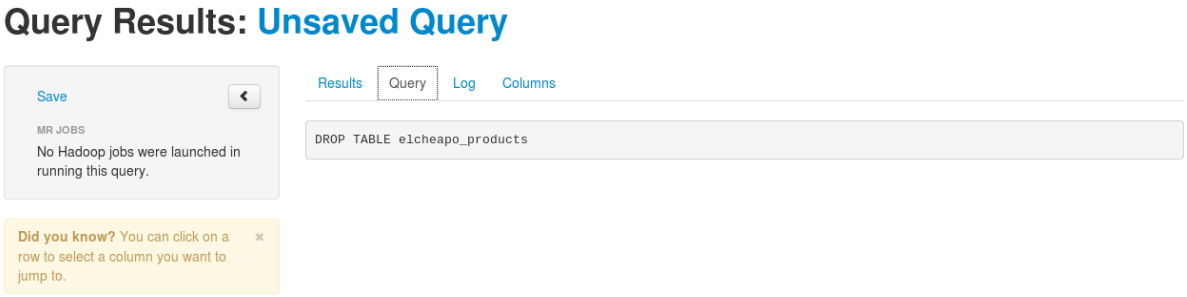



| | order_day | total_revenue |
|---|---|---|
| 0 | 1 | $383618.19 |
| 1 | 2 | $1064939.68 |
| 2 | 3 | $1100699.17 |
| 3 | 4 | $1028571.67 |
| 4 | 5 | $1066038.15 |
| 5 | 6 | $1075537.38 |
| 6 | 7 | $1033746.80 |
| 7 | 8 | $1073944.64 |
| 8 | 9 | $1100788.81 |
| 9 | 10 | $1062743.60 |
| 10 | 11 | $1059573.17 |
| 11 | 12 | $1065223.57 |
| 12 | 13 | $1072153.66 |
| 13 | 14 | $1080752.28 |
| 14 | 15 | $1081561.73 |
| 15 | 16 | $1033992.27 |
| 16 | 17 | $1087067.38 |
| 17 | 18 | $1045830.21 |
| 18 | 19 | $1058442.66 |
| 19 | 20 | $1035860.45 |
| 20 | 21 | $1050509.84 |

# Question 5
# (codes and table contents)







## Query Results: Unsaved Query

DOWNLOADS

Download as CSV

Download as XLS

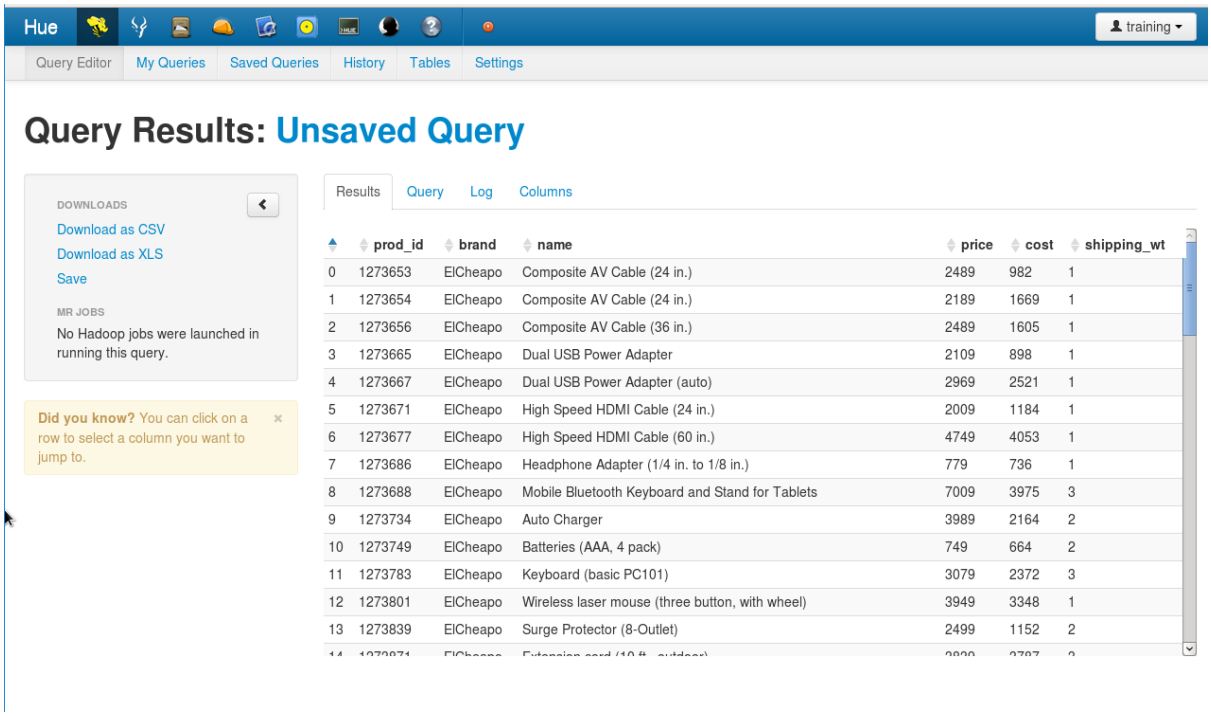
Save

MR JOBS

No Hadoop jobs were launched in running this query.

Results | Query | Log | Columns

| | tab_name |
|---|---|
| 0 | customers |
| 1 | elcheapo_products |
| 2 | order_details |
| 3 | orders |
| 4 | products |




## Query Results: Unsaved Query

DOWNLOADS

Download as CSV

Download as XLS

Save

MR JOBS

No Hadoop jobs were launched in running this query.

Results | Query | Log | Columns

| | tab_name |
|---|---|
| 0 | customers |
| 1 | order_details |
| 2 | orders |
| 3 | products |

Did you know? You can click on a row to select a column you want to jump to.