**Big Data Class Project – Part 1**

**Amazon Customer Review Data Analysis**

**Introduction:**

Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency. Big data has one or more of the following characteristics: high volume, high velocity or high variety. Big data analytics is the often complex process of examining large and varied data sets, or big data, to uncover information -- such as hidden patterns, unknown correlations, market trends and customer preferences -- that can help organizations make informed business decisions.

**Problem description:**

In this project we are practicing and demonstrating out Big Data (BD) and Analytics skills. We are using Amazon customer review data which has more than 160 Million (160,796,570) observations. We will be using HDFS, Hive and AWS to handle and analyze the data and systematically extract information from.

**Data Set:**

Amazon reviews dataset: [*https://registry.opendata.aws/amazon-reviews/*](https://registry.opendata.aws/amazon-reviews/)

**Project Environment:**

AWS EMR - AWS Educate Class account.

Project tools:

1. AWS S3

2. AWS EMR - Hive and HDFS

3. AWS Athena - AWS alternative to Hive for the ﬁles stored in S3

**Dataset Requirements:**

A)Use the following product categories:

- Wireless

- Automotive

- Music

- Digital\_Music\_Purchase

- Sports

- Toys

- Digital\_Video\_Games

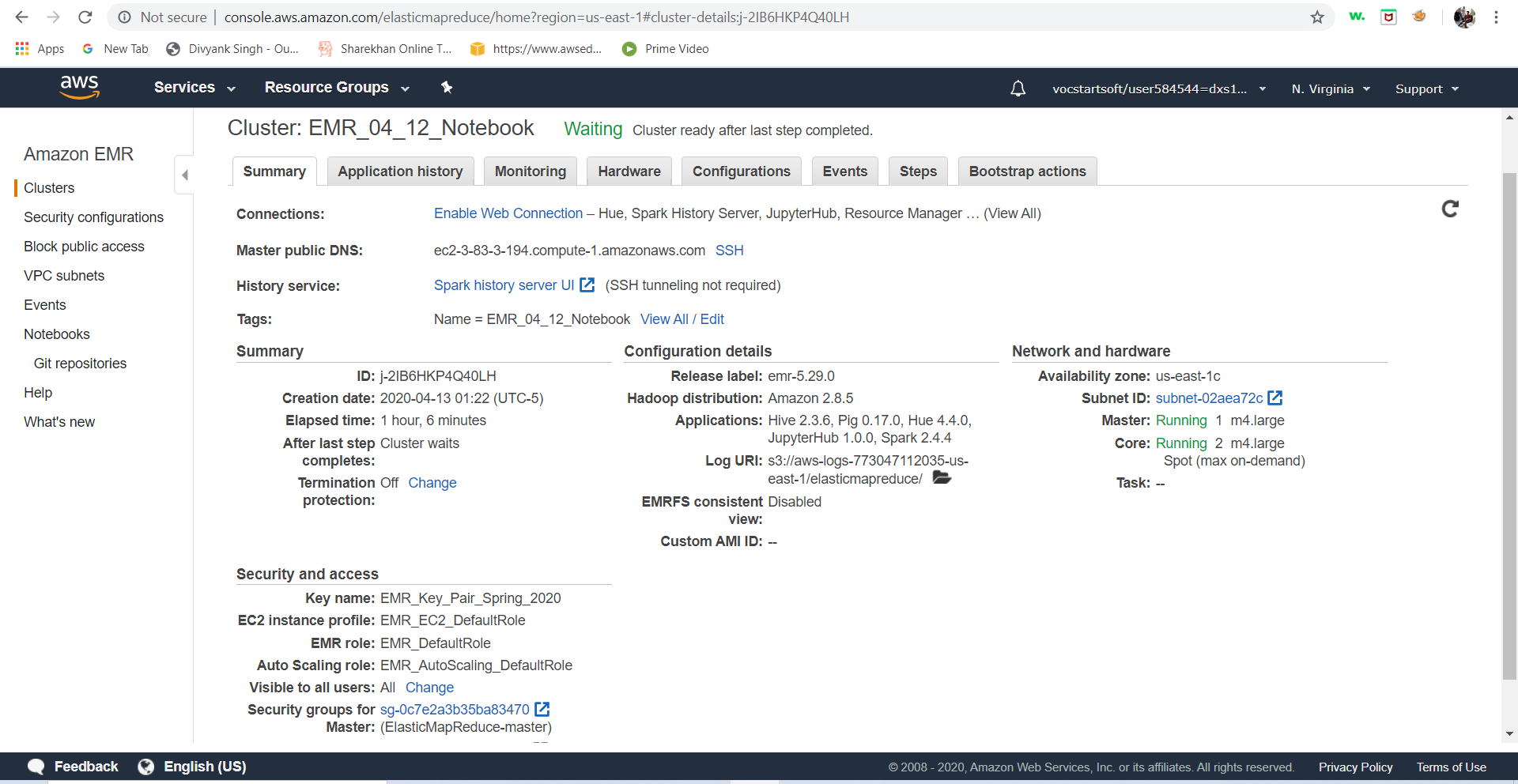
- Video\_Games

B)Start your analysis from year 2005.

C)Exclude multiple reviews by the same users for the same product. Each user should be allowed to review the product only once. To improve performance of your queries, create external table to point to HDFS/S3 ﬁle that will include all review-ids to be excluded.

**Step 1:- AWS EMR:**

Provisioned EMR with below details:-

****

**Step 2:-**

**A) Created directory for below each product category:-**

1. **Wireless**
2. **Automotive**
3. **Music**
4. **Digital\_Music\_Purchase**
5. **Sports**
6. **Toys**
7. **Digital\_Video\_Games**
8. **Video\_Games**

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product\_category=Wireless/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product\_category=Automotive/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product\_category=Music/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product\_category=Sports/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product\_category=Digital\_Music\_Purchase/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product\_category=Toys/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product\_category=Digital\_Video\_Games/

hdfs dfs -mkdir -p /hive/amazon-reviews-pds/parquet/product\_category=Video\_Games/

**B)Copying dataset from S3 in HDFS for each of the product category mentioned:-**

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product\_category=Wireless/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product\_category=Wireless/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product\_category=Automotive/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product\_category=Automotive/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product\_category=Music/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product\_category=Music/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product\_category=Digital\_Music\_Purchase/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product\_category=Digital\_Music\_Purchase/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product\_category=Sports/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product\_category=Sports/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product\_category=Toys/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product\_category=Toys/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product\_category=Digital\_Video\_Games/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product\_category=Digital\_Video\_Games/

s3-dist-cp --src=s3://amazon-reviews-pds/parquet/product\_category=Video\_Games/ --dest=hdfs:///hive/amazon-reviews-pds/parquet/product\_category=Video\_Games/

**Step3:- Creating database and tables:**

**A)Creating Database:-**

create database amazon\_review;

**B)Dropping table:-**

drop table amazon\_review.amazon\_reviews\_parquet;

**C)Creating external table:-**

CREATE EXTERNAL TABLE amazon\_review.amazon\_reviews\_parquet(

`marketplace` string,

`customer\_id` string,

`review\_id` string,

`product\_id` string,

`product\_parent` string,

`product\_title` string,

`star\_rating` int,

`helpful\_votes` int,

`total\_votes` int,

`vine` string,

`verified\_purchase` string,

`review\_headline` string,

`review\_body` string,

`review\_date` DATE,

`year` int)

PARTITIONED BY (

`product\_category` string)

--ROW FORMAT DELIMITED

--STORED AS PARQUET

ROW FORMAT SERDE

'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'

STORED AS INPUTFORMAT

'org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'

OUTPUTFORMAT

'org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat'

LOCATION

'hdfs:///hive/amazon-reviews-pds/parquet/'

TBLPROPERTIES (

'transient\_lastDdlTime'='1583454851');

Msck repair table amazon\_review.amazon\_reviews\_parquet;

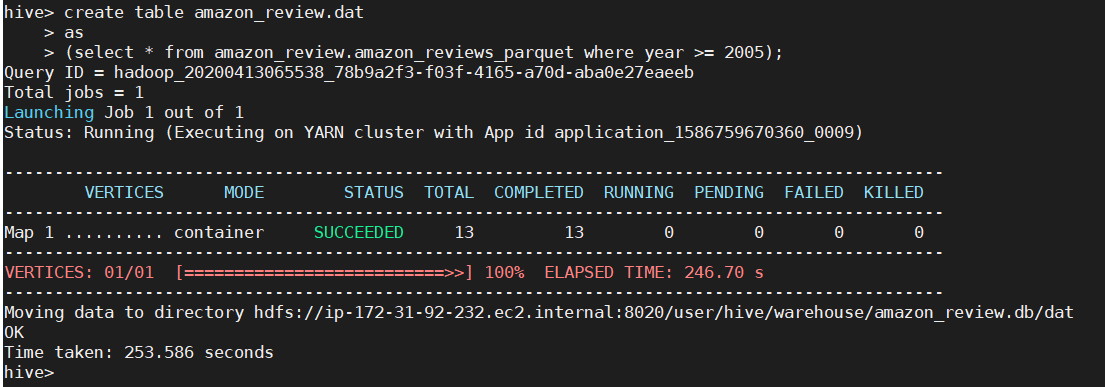
**D)Creating table which has year>=2005 and product category ('Wireless', 'Automotive', 'Music', 'Digital\_Music\_Purchase', 'Sports', 'Toys', 'Digital\_Video\_Games', 'Video\_Games')**

**Query:-**

create table amazon\_review.dat

as

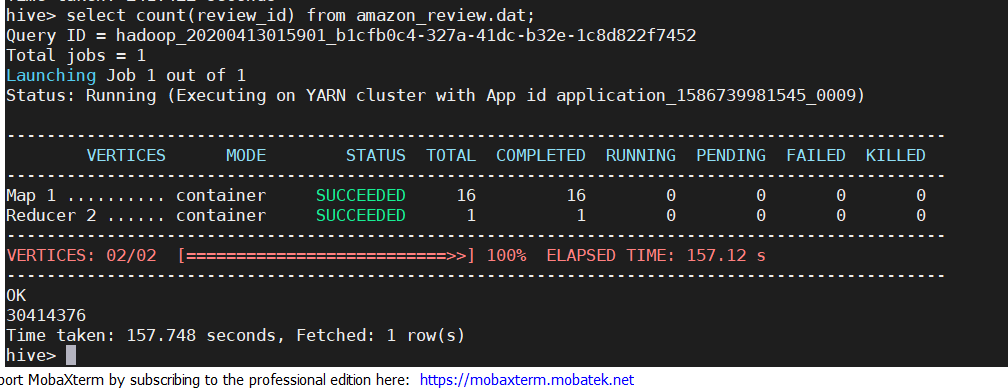
(select \* from amazon\_review.amazon\_reviews\_parquet where year >= 2005);



**E) Checking obs now:- 30414376**

**Query:-**

select count(review\_id) from amazon\_review.dat;



**F)Creating a table which has obs where a customer gave multiple reviews for a product:-**

**Query:-**

create table amazon\_review.excludedID

as

select distinct \* from amazon\_review.dat where

review\_id in (select review\_id from

(select \* from amazon\_review.dat

) tab1

inner join

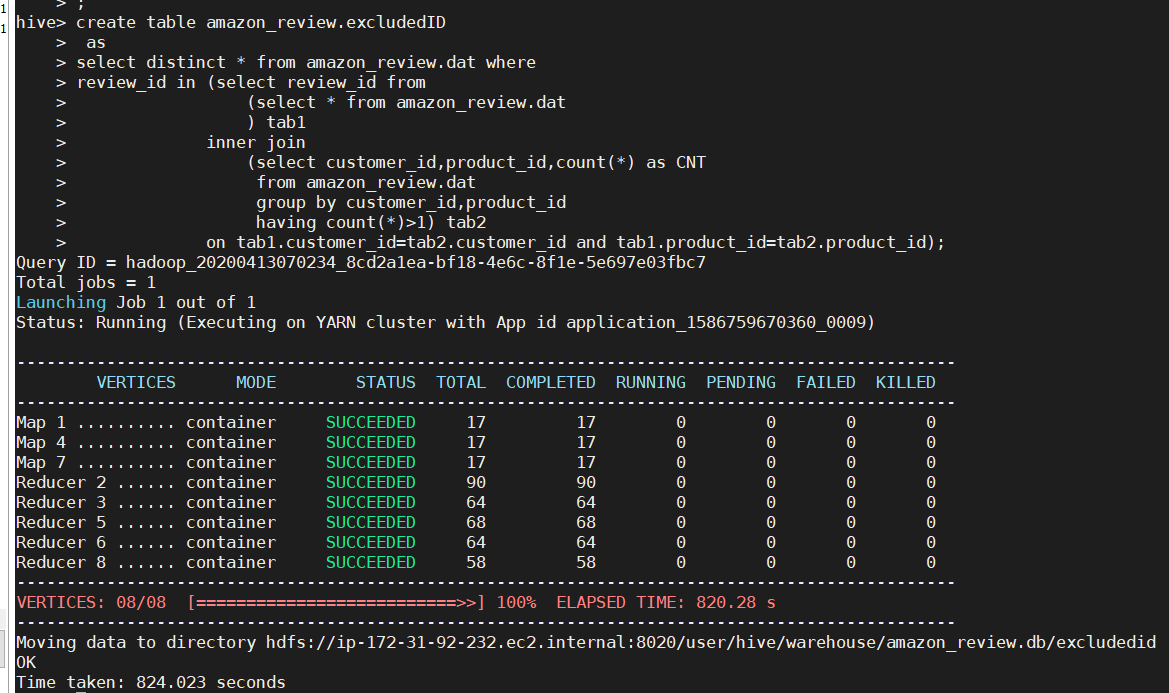
(select customer\_id,product\_id,count(\*) as CNT

from amazon\_review.dat

group by customer\_id,product\_id

having count(\*)>1) tab2

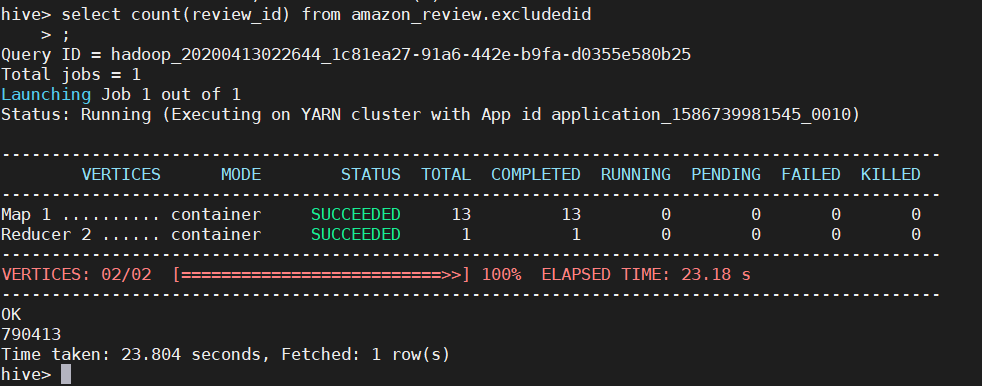
on tab1.customer\_id=tab2.customer\_id and tab1.product\_id=tab2.product\_id);



**G)Checking obs now:- 790413**

**Query:-**

select count(review\_id) from amazon\_review.excluded\_reviewid;



**H)Saving the obs where customer’s has only 1 review for a product:-**

**Query:-**

create table amazon\_review.final

as

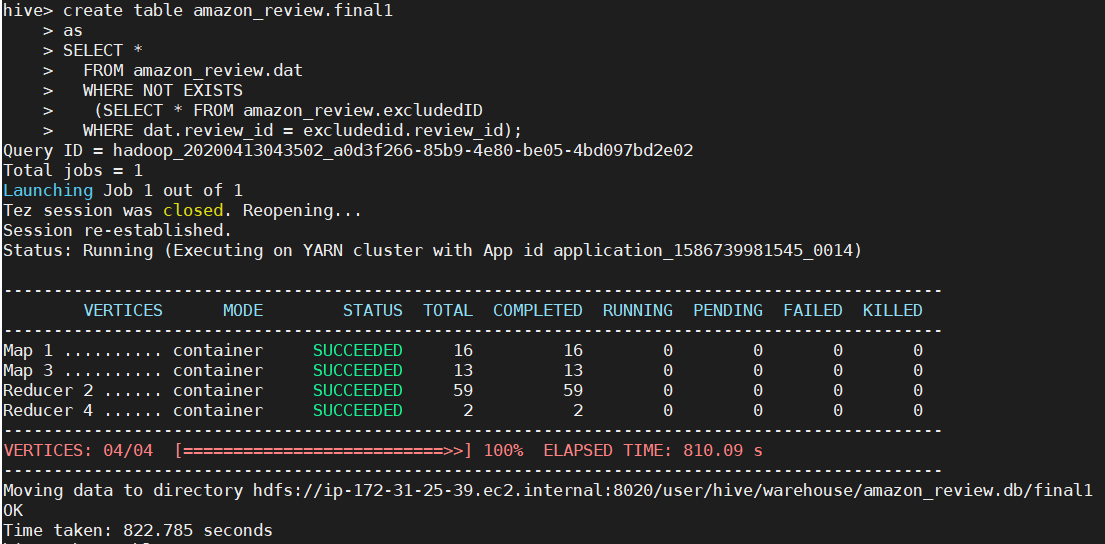
SELECT \*

  FROM amazon\_review.dat

  WHERE NOT EXISTS

   (SELECT \* FROM amazon\_review.excludedID

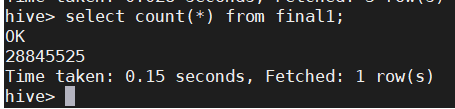
  WHERE dat.review\_id = excludedid.review\_id);



**I) Checking obs now:- 28845525**

**Query:-**

select count(review\_id) from amazon\_review.final;



**Step4:- Performing Exploratory Data Analysis**

**1. Explore the dataset and provide basic exploratory analysis:**

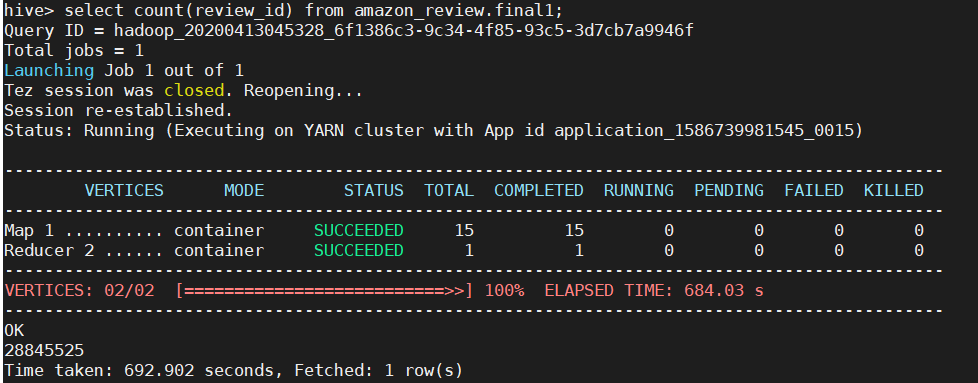
1. **Number of reviews**

**Query:-**

select count(review\_id) from amazon\_review.final;

**Output:-**

28845525



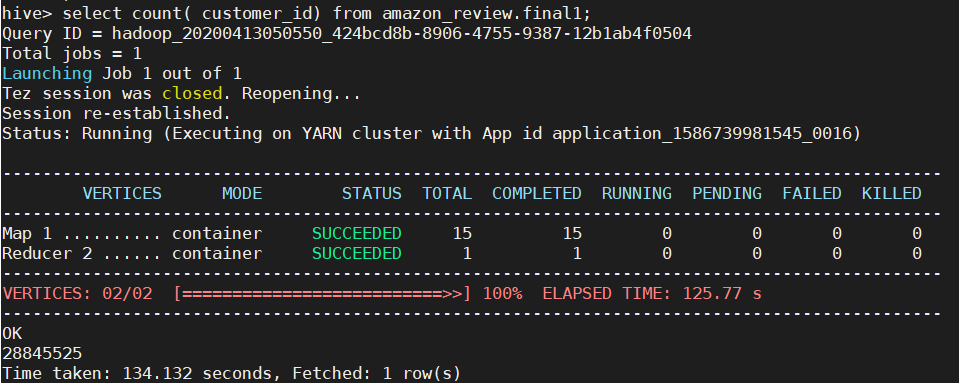
1. **Number of users**

**Query:-**

select count( customer\_id) from amazon\_review.final;

**Output:-**

28845525



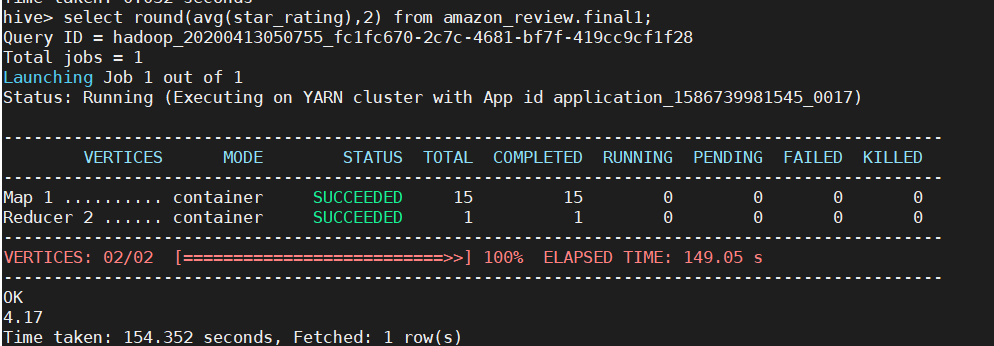
1. **Average review stars**

**Query:-**

select round(avg(star\_rating),2) from amazon\_review.final;

**Output:-**

4.17



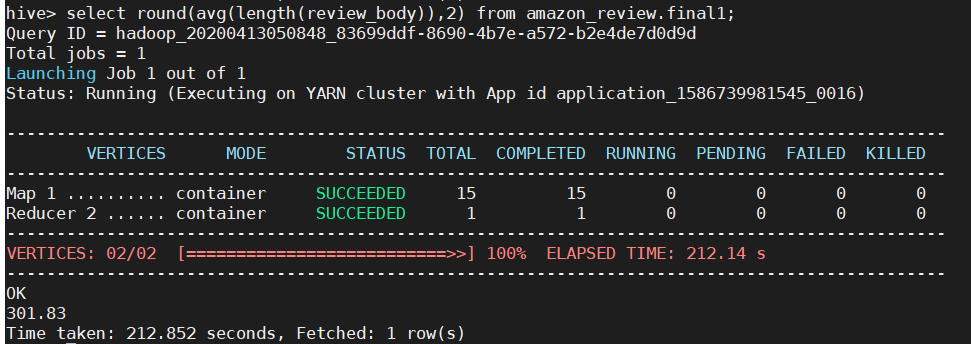
1. **Average length of the review**

**Query:-**

select round(avg(length(review\_body)),2) from amazon\_review.final;

**Output:-**

301.83



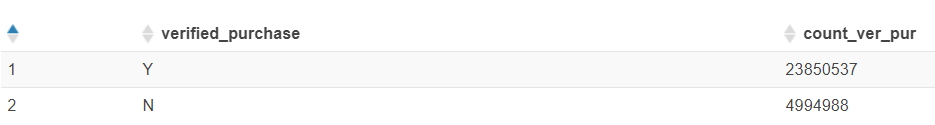
1. **Number of veriﬁed versus unveriﬁed reviews**

**Query:-**

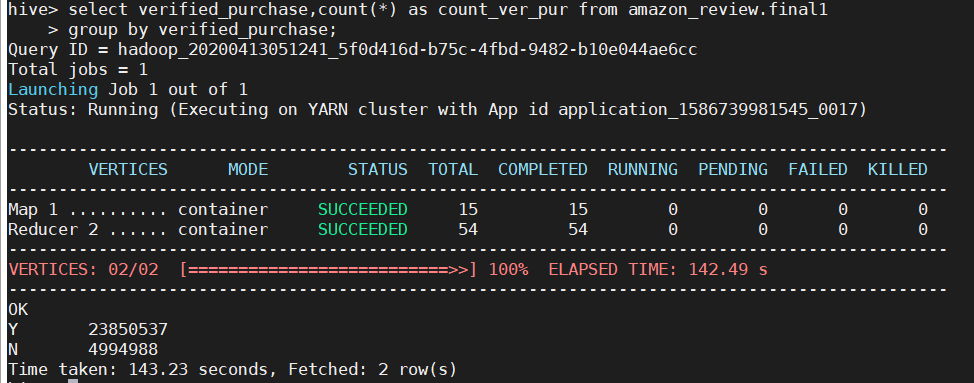
select verified\_purchase,count(\*) as count\_ver\_pur from amazon\_review.final

group by verified\_purchase;

**Output:-**



(Athena)



(Terminal)

1. **At least two more additional metrics:-**
2. **Count of each product\_category:-**

**Query:-**

Select product\_category, count\_prod\_cat, rank() over

(ORDER BY count\_prod\_cat DESC)

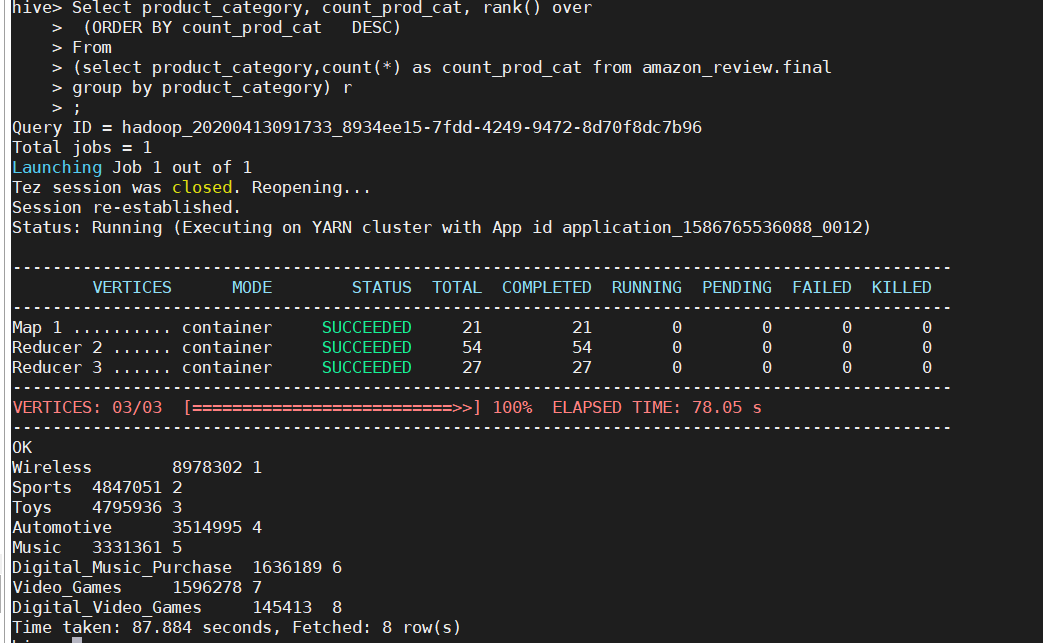
From

(select product\_category,count(\*) as count\_prod\_cat from amazon\_review.final

group by product\_category) r

;

**Output:-**



(Terminal)



|  |  |  |  |
| --- | --- | --- | --- |
| **num** | **product\_category** | **count\_prod\_cat** | **average** |
| 1 | Wireless | 8978302 | 31.13 |
| 2 | Sports | 4847051 | 16.8 |
| 3 | Toys | 4795936 | 16.63 |
| 4 | Automotive | 3514995 | 12.19 |
| 5 | Music | 3331361 | 11.55 |
| 6 | Digital\_Music\_Purchase | 1636189 | 5.67 |
| 7 | Video\_Games | 1596278 | 5.53 |
| 8 | Digital\_Video\_Games | 145413 | 0.5 |
|  | Sum | 28845525 |  |

**Interpretation:-**

We can see that more that 31% of the reviews are of wireless products, followed by Sports(16.8%) and Toys(16.63%), whereas Video games and digital video category constitutes only 6%.

1. **Count of each marketplace:-**

**Query:-**

Select marketplace, count\_mar\_plc, rank() over

(ORDER BY count\_mar\_plc DESC)

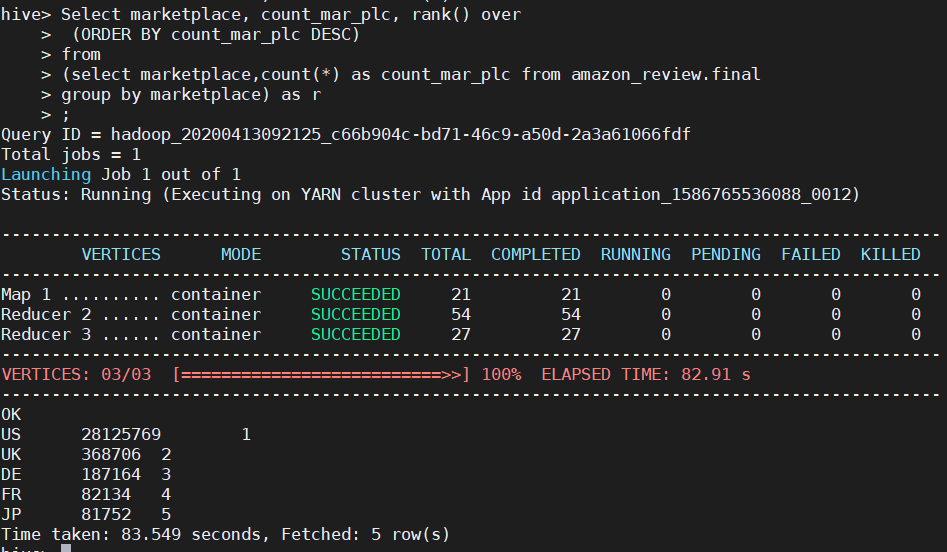
from

(select marketplace,count(\*) as count\_mar\_plc from amazon\_review.final

group by marketplace) as r

;

**Output:-**





|  |  |  |  |
| --- | --- | --- | --- |
|  | **marketplace** | **count\_mar\_plc** | per |
| 1 | JP | 81752 | 0.28341311 |
| 2 | FR | 82134 | 0.28473741 |
| 3 | DE | 187164 | 0.64884934 |
| 4 | UK | 368706 | 1.27820866 |
| 5 | US | 28125769 | 97.5047915 |
|  | sum | 28845525 |  |

**Interpretation:-**

I have grouped dataset based on marketplace and found that more than 97% of the customers are from just US.

1. **Finding mean,min,max,standard deviation per category:-**

**Query:-**

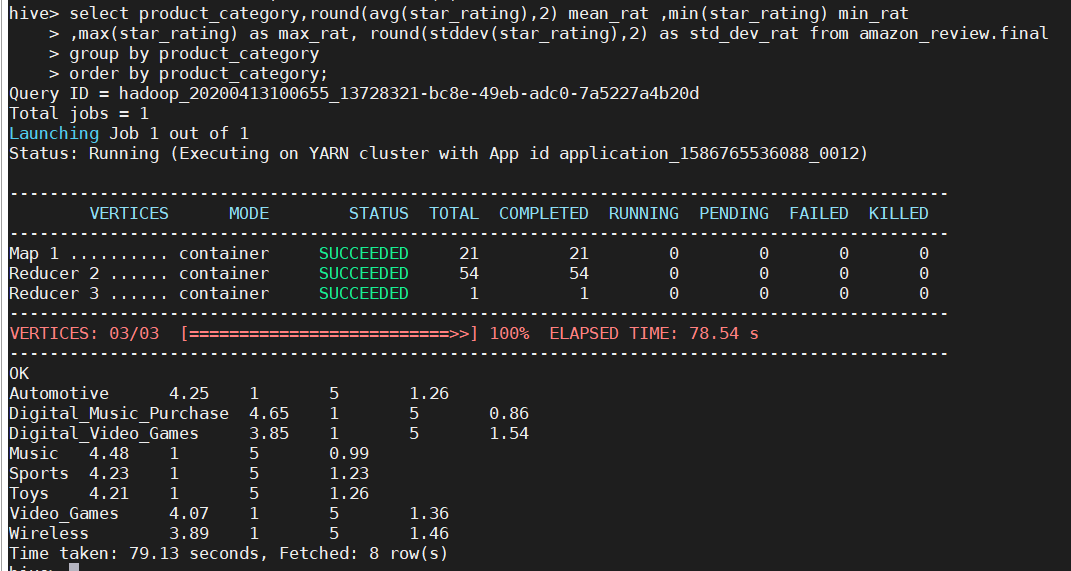
select product\_category,round(avg(star\_rating),2) mean\_rat ,min(star\_rating) min\_rat

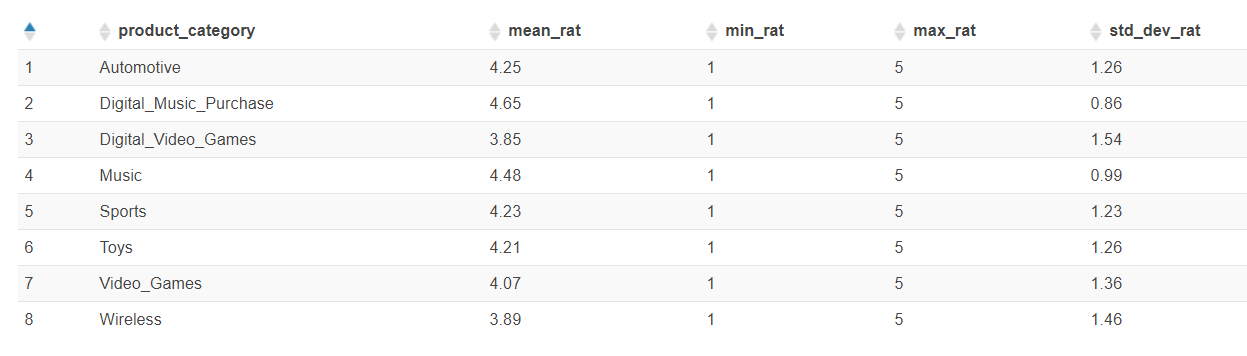
,max(star\_rating) as max\_rat, round(stddev(star\_rating),2) as std\_dev\_rat from amazon\_review.final

group by product\_category

order by product\_category;

**Output:-**





**Interpretation:-**

The output shows the minimum, maximum, mean and standard deviation for each of the product category.

1. **Provide trending (over time) analysis of each of the metrics above:-**

**Query:-**

select product\_category,year,count(\*) as count\_prod from amazon\_review.final

group by product\_category,year

order by product\_category,year

;

**Interpretation:-**

I have grouped dataset based on product category and year and imported the output in excel to get the above trendline, which shows that purchase count has definitely increased for all the product categories, though wireless category has noticed increase in purchase count rate highest.

**Query:-**

select marketplace,year,count(\*) as count\_marketplace from amazon\_review.final

group by marketplace,year

order by marketplace,year

;

**Interpretation:-**

I have grouped dataset based on marketplace and year and imported the output in excel to get the above trendline, which shows that purchase count rate has significantly increased in US only.

**2. Provide detailed analysis of Music/Digital\_Music\_Purchase and Digital\_Video\_Games/Video\_Games over time.**

**1. Do you see correlation (maybe negative) between the categories over time?**

**Part 1:-Checking between Music/Digital\_Music\_Purchase**

**Query:-**

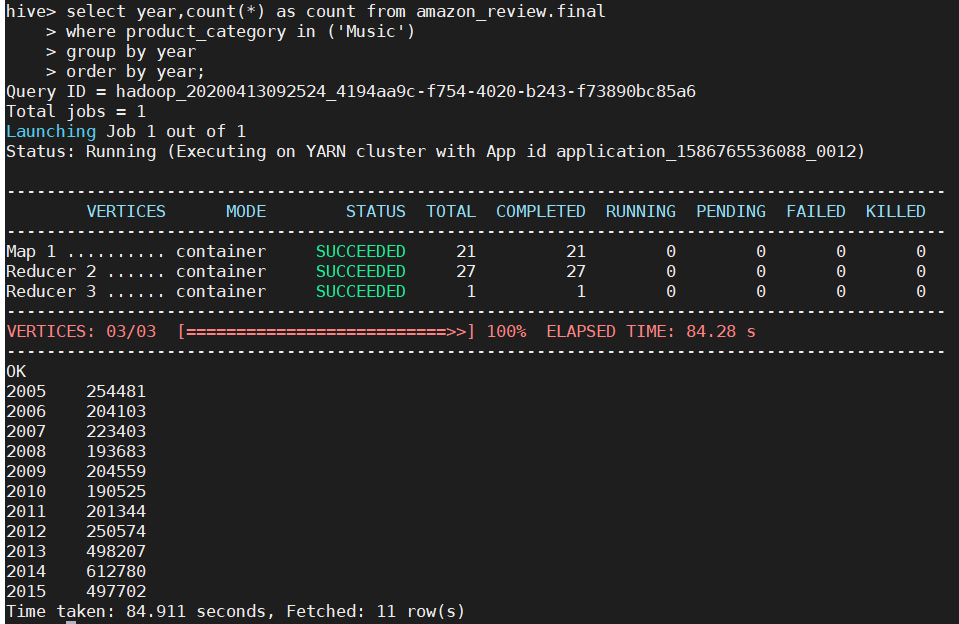
select year,count(\*) as count from amazon\_review.final

where product\_category in ('Music')

group by year

order by year;

**Output:-**





**Interpretation:-**

Here I have performed grouping dataset based on year for product category “Music”

**Query:-**

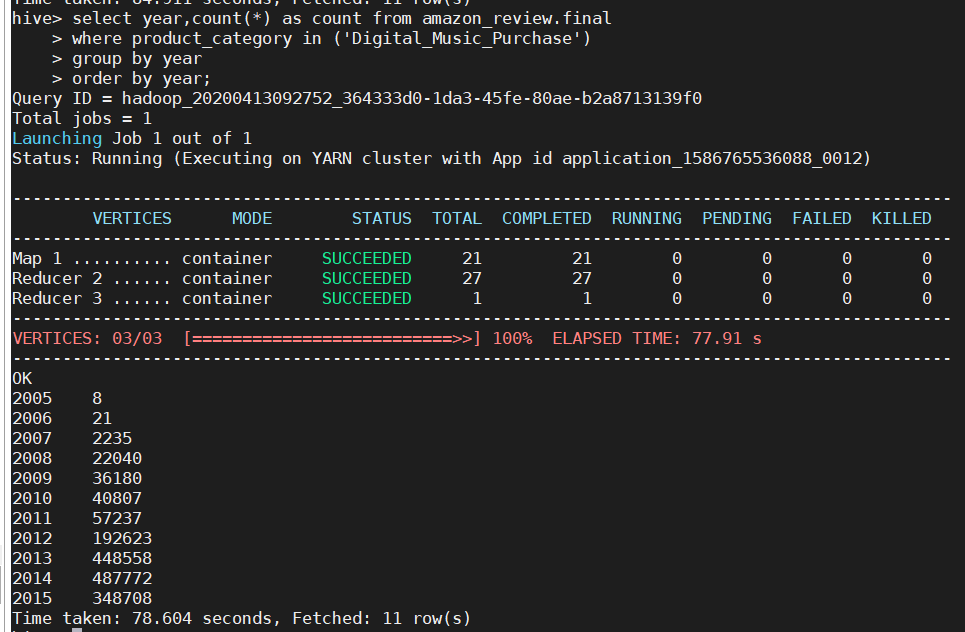
select year,count(\*) as count from amazon\_review.final

where product\_category in ('Digital\_Music\_Purchase')

group by year

order by year;

**Output:-**





Here I have performed grouping dataset based on year for product category “Music”

**Query:- (using corr function to find the correlation between music product category and year)**

select corr(count\_prod,year) as corr\_music from

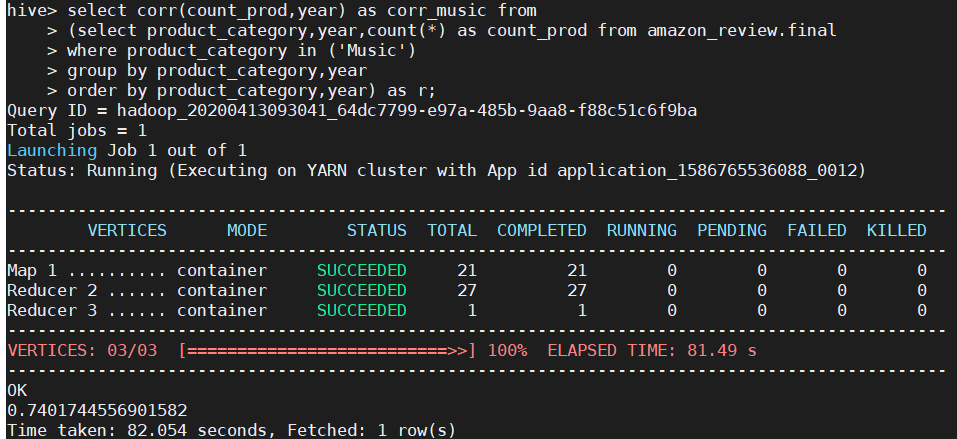
(select product\_category,year,count(\*) as count\_prod from amazon\_review.final

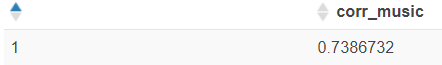
where product\_category in ('Music')

group by product\_category,year

order by product\_category,year) as r;

**Output:-**





I did a correlation test between music category purchase count and years and found that there is a high positive correlation between them, which states as year increased so does the purchase count.

**Query:- (using corr function to find the correlation between digital\_music\_category product category and year)**

select round(corr(count\_prod,year),2) as corr\_music from

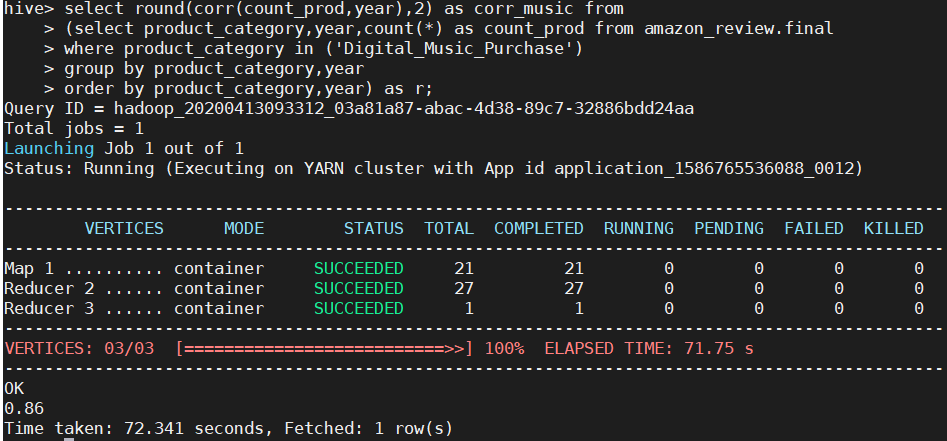
(select product\_category,year,count(\*) as count\_prod from amazon\_review.final

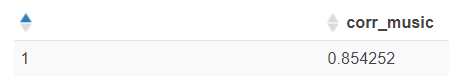
where product\_category in ('Digital\_Music\_Purchase')

group by product\_category,year

order by product\_category,year) as r;

**Output:-**





I did a correlation test between digital music category purchase count and years and found that there is a high positive correlation between them, which states as year increased so does the purchase count.

**Detailed analysis:-**

**Query:-**

select product\_category,year,count(\*) as count\_prod from amazon\_review.final

where product\_category in ('Music', 'Digital\_Music\_Purchase')

group by product\_category,year

order by product\_category,year

;

The above trendline between year and music/digital\_music\_category even shows positive correlation

**Part2:-Checking between Digital\_Video\_Games/Video\_Games**

**Query:-**

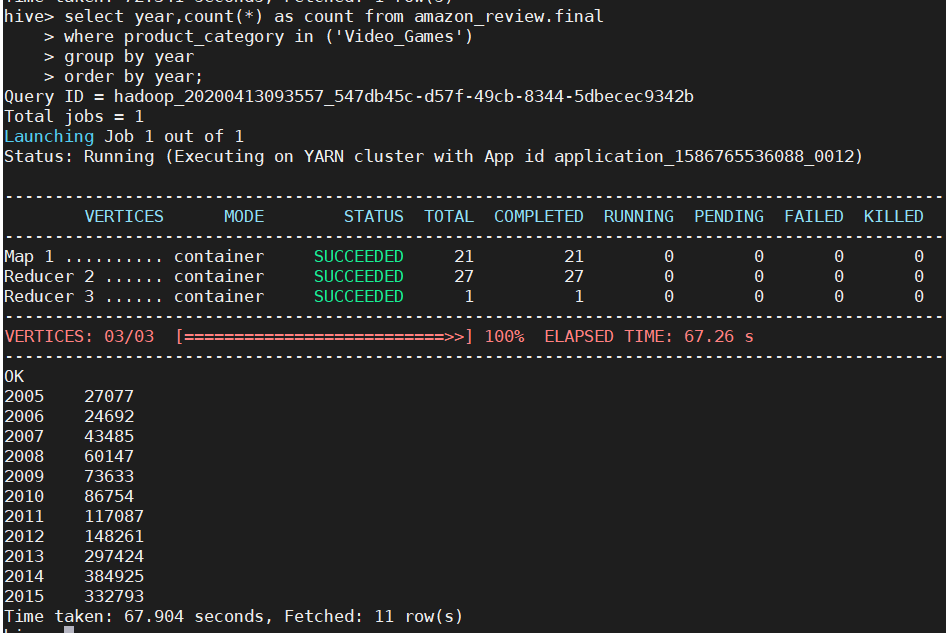
select year,count(\*) as count from amazon\_review.final

where product\_category in ('Video\_Games')

group by year

order by year;

**Output:-**

****

Here I have performed grouping dataset based on year for product category “Video\_Games”

**Query:-**

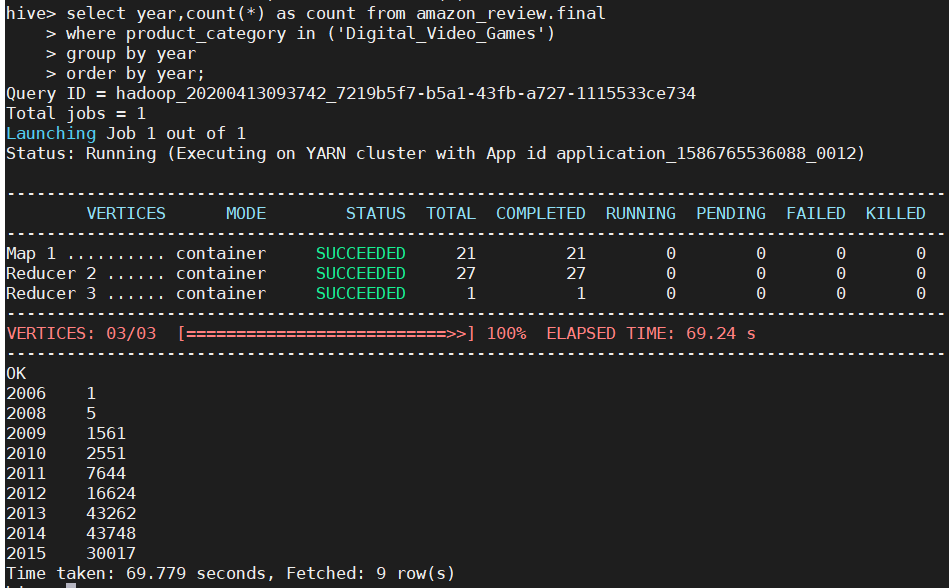
select year,count(\*) as count from amazon\_review.final

where product\_category in ('Digital\_Video\_Games')

group by year

order by year;

**Output:-**



Here I have performed grouping dataset based on year for product category “Digital\_Video\_Games”

**Query:- (using corr function to find the correlation between Video\_Games product category and year)**

select round(corr(count\_prod,year),2) as corr\_vid from

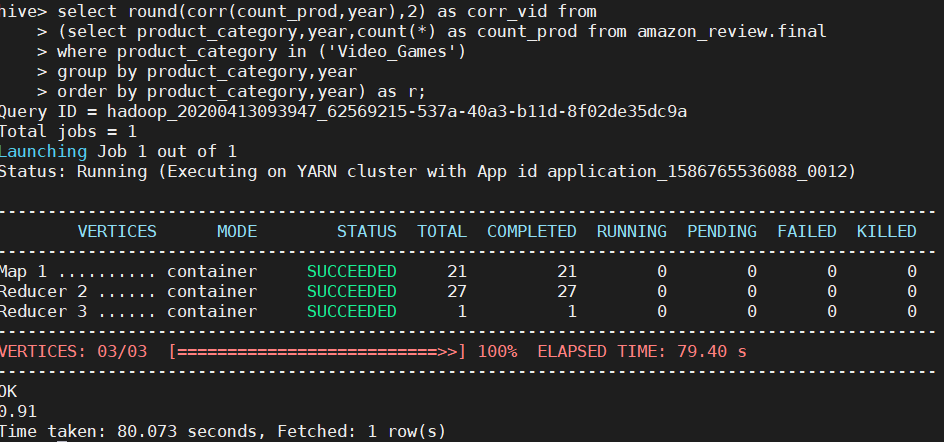
(select product\_category,year,count(\*) as count\_prod from amazon\_review.final

where product\_category in ('Video\_Games')

group by product\_category,year

order by product\_category,year) as r;

**Output:-**

****

I did a correlation test between video games category purchase count and years and found that there is a high positive correlation between them, which states as year increased so does the purchase count.

**Query:- (using corr function to find the correlation between Digital\_Video\_Games product category and year)**

select round(corr(count\_prod,year),2) as corr\_vid from

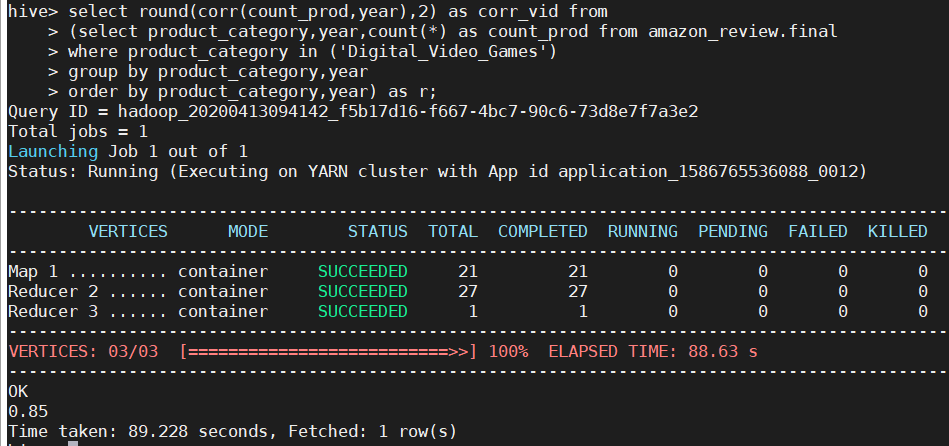
(select product\_category,year,count(\*) as count\_prod from amazon\_review.final

where product\_category in ('Digital\_Video\_Games')

group by product\_category,year

order by product\_category,year) as r;

**Output:-**

****

**Interpretation:-**

I did a correlation test between digital video games category purchase count and years and found that there is a high positive correlation between them, which states as year increased so does the purchase count.

**Detailed Analysis:-**

**Query:-**

select product\_category,year,count(\*) as count\_prod from amazon\_review.final

where product\_category in ('Video\_Games', 'Digital\_Video\_Games')

group by product\_category,year

order by product\_category,year

;

**Interpretation:-**

The above trendline between year and music/digital\_music\_category even shows positive correlation

**2. Are there same users reviewing in both categories?**

**Query:-**

select count(distinct dmp.customer\_id) as common\_cus from

(select customer\_id from amazon\_review.final

where product\_category = 'Digital\_Music\_Purchase') as dmp

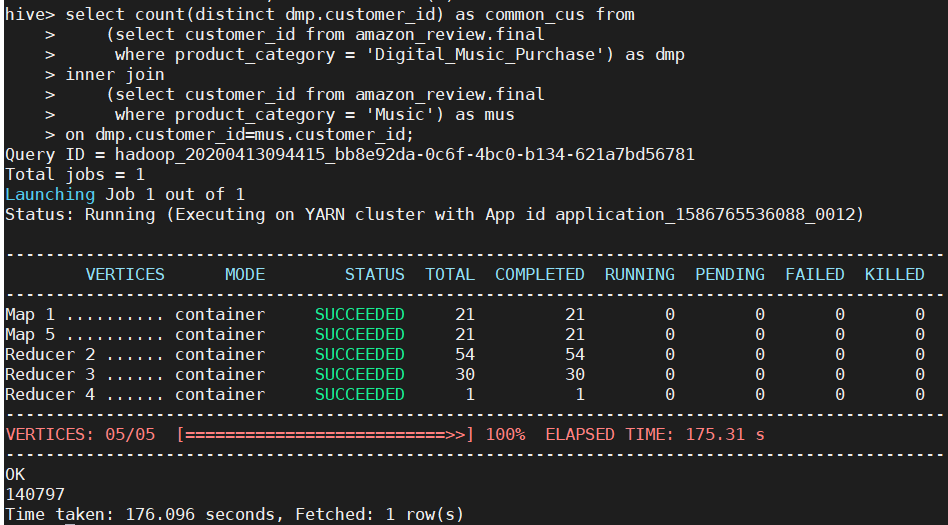
inner join

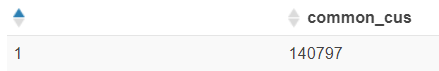
(select customer\_id from amazon\_review.final

where product\_category = 'Music') as mus

on dmp.customer\_id=mus.customer\_id;

**Output:-**





**Interpretation:-**

For Music and Digital\_Music\_Purchase there are 140797 common customers.

**Query:-**

select count(distinct dvs.customer\_id) as common\_cus from

(select customer\_id from amazon\_review.final

where product\_category = 'Digital\_Video\_Games') as dvs

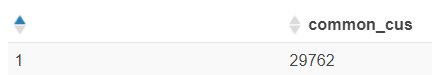
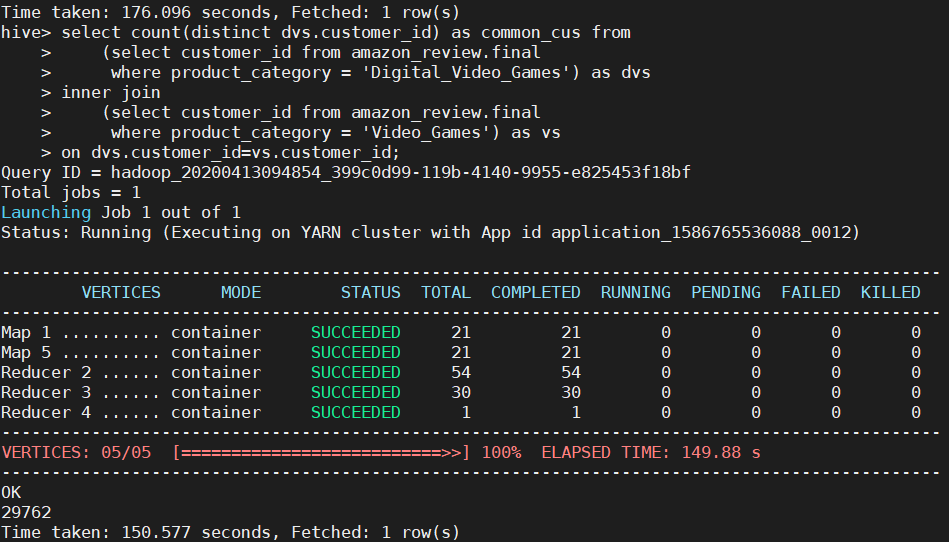
inner join

(select customer\_id from amazon\_review.final

where product\_category = 'Video\_Games') as vs

on dvs.customer\_id=vs.customer\_id;

**Output:-**



**Interpretation:-**

For Video\_Games and Digital\_Video\_Games there are 29,762 common customers.

**3. Can you identify similar items in both categories? Do they get same rating?**

**Part1:-Video\_Games and Digital\_Video\_Games**

1. **Finding similar items in both categories:-**

**Query:-**

select distinct vg.product\_id from

(select product\_id from final

where product\_category='Video\_Games') vg

inner join

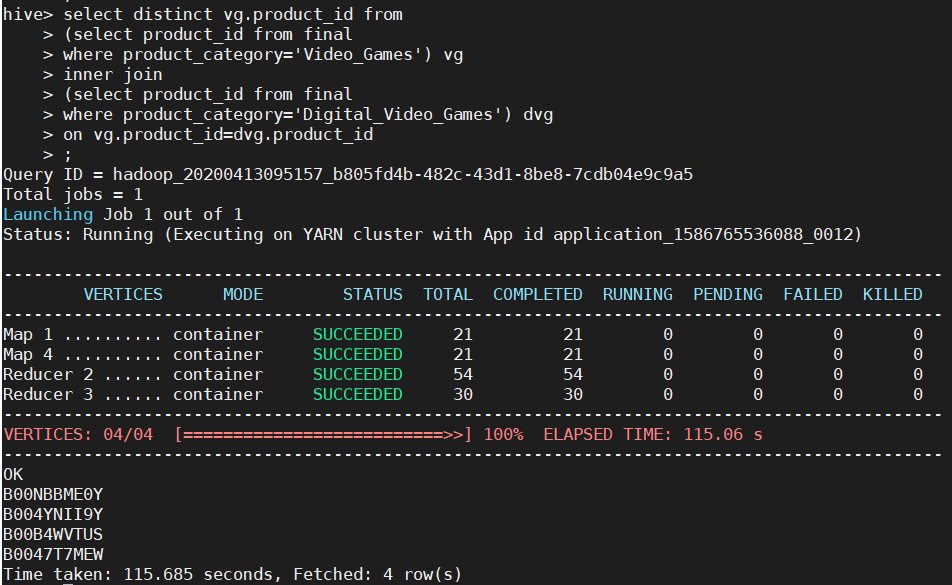
(select product\_id from final

where product\_category='Digital\_Video\_Games') dvg

on vg.product\_id=dvg.product\_id

;

**Output:-**



**Interpretation:-**

There is 4 similar product between Video\_Games and Digital\_Video\_Games

1. **Do they get similar ratings:-**

select vg.product\_id,vg\_avg\_rat,dvg\_avg\_rat,round((vg\_avg\_rat-dvg\_avg\_rat),2) as diff from

(select product\_id, round(avg(star\_rating),2) as vg\_avg\_rat from final

where product\_category='Video\_Games'

group by product\_id) vg

inner join

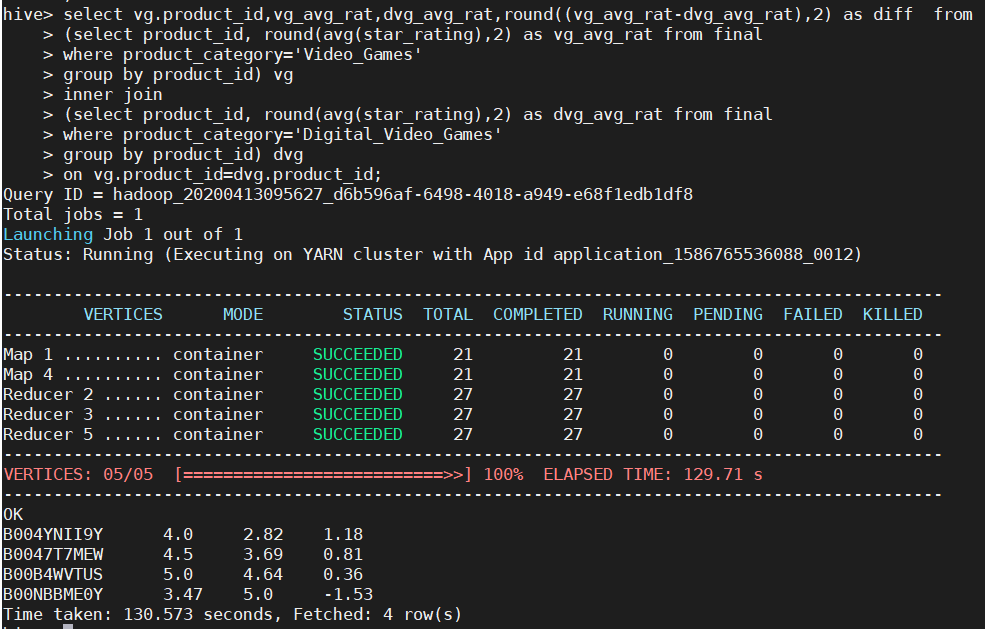
(select product\_id, round(avg(star\_rating),2) as dvg\_avg\_rat from final

where product\_category='Digital\_Video\_Games'

group by product\_id) dvg

on vg.product\_id=dvg.product\_id;

**Output:-**





**Interpretation:-**

The above table shows that they don’t get similar ratings.

**Part2:- Music v/s Digital\_Music\_Purchase**

**Query:-**

select distinct m.product\_id from

(select product\_id from final

where product\_category='Music') m

inner join

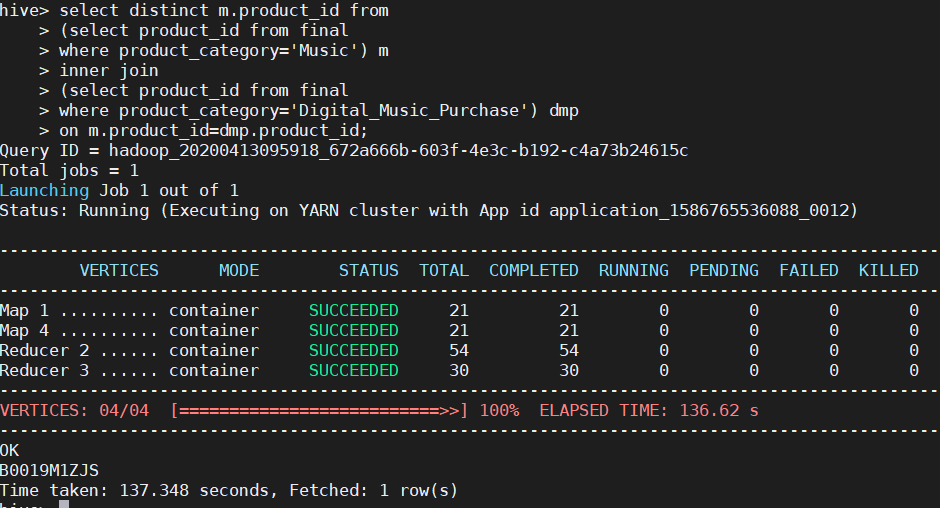
(select product\_id from final

where product\_category='Digital\_Music\_Purchase') dmp

on m.product\_id=dmp.product\_id;

;

**Output:-**





**Interpretation:-**

There is only 1 similar product between Video\_Games and Digital\_Video\_Games

**Query:-**

select m.product\_id,mus\_avg\_rat,dmp\_avg\_rat from

(select product\_id, avg(star\_rating) as mus\_avg\_rat from final

where product\_category='Music'

group by product\_id) m

inner join

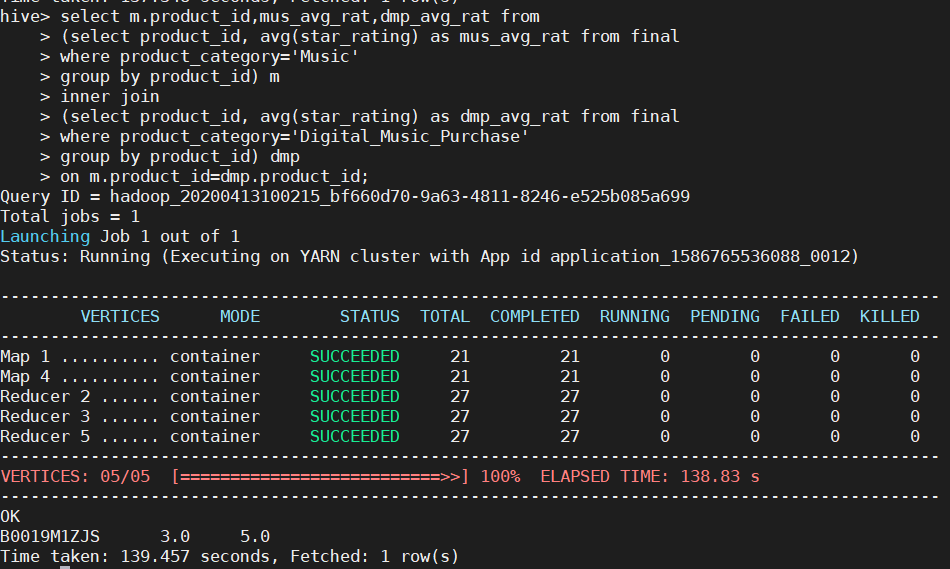
(select product\_id, avg(star\_rating) as dmp\_avg\_rat from final

where product\_category='Digital\_Music\_Purchase'

group by product\_id) dmp

on m.product\_id=dmp.product\_id;

**Output:-**





**Interpretation:-**

The above table shows that they don’t get similar ratings.

**4. You should cover additional questions and not limit yourself to the above questions**

**Query1:-**

select month(review\_date) as month,count(\*) from final

group by month(review\_date)

order by month(review\_date)

;

**Interpretation:-**

We can see that January month sees the highest purchase(3181577), then follows August July, December.

**Query2:-**

select product\_category,month(review\_date) as month,count(\*) from final

group by product\_category,month(review\_date)

order by product\_category,month(review\_date)

;

**Interpretation:-**

Here I have grouped the dataset based on Product category and month, to find the purchase count for each category per month. We can see that in August month wireless category saw maximum purchase whereas other categories saw in January and December.

**Query3:-**

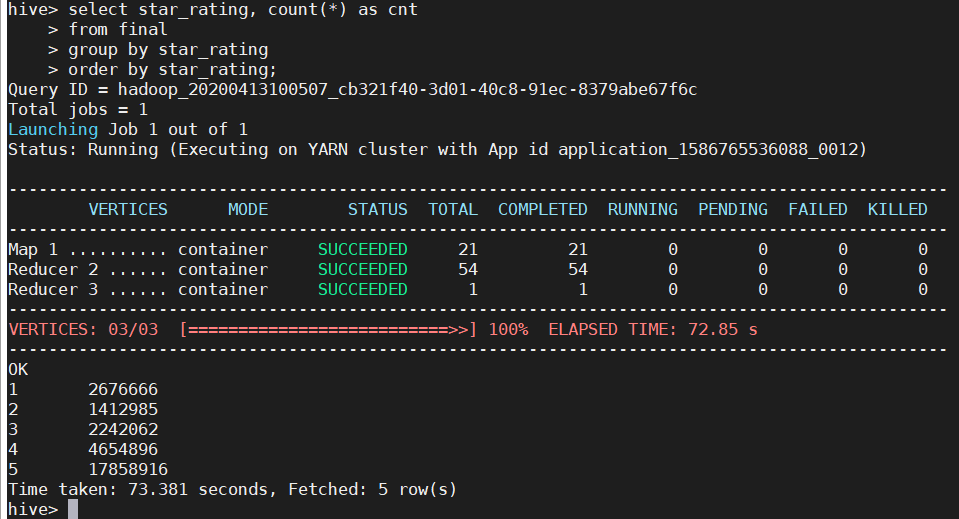
select star\_rating, count(\*) as cnt

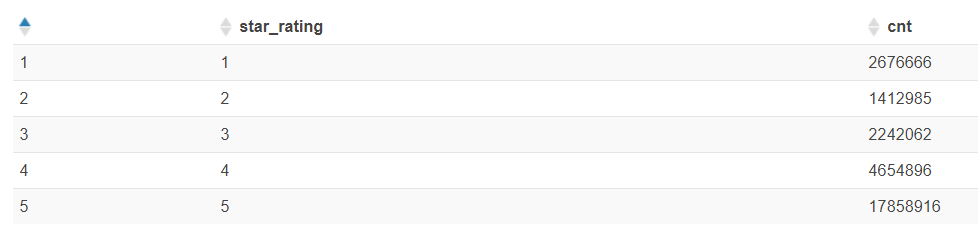
from final

group by star\_rating

order by star\_rating;

**Output:-**





**Interpretation:-**

In the above visualization we can say that we have most data we star rating 5. That shows most of the customers are highly satisfied with the products.

**Query4:-**

select product\_category,star\_rating as month,count(\*) from final

group by product\_category,star\_rating

order by product\_category,star\_rating

;

**Interpretation:-**

We can see that for wireless category, though purchase count is highest, customers are not highly satisfied with the products as their highest rating is 4, even after that rating 1 got maximum count.

Whereas for other product categories there is maximum purchase count for star rating 5 than other ratings, which shows that customers buying products other than wireless category are more satisfied.

**Conclusion:-**

1. We can see that more that 31% of the reviews are of wireless products, followed by Sports(16.8%) and Toys(16.63%), whereas Video games and digital video category constitutes only 6%.
2. 97% of the reviews are from just US.
3. There is a positive correlation between number of customers and years and this can be seen for all the product categories.
4. Wireless product category has noticed highest increase in reviews over the years compared to any other product category.
5. The number of customers have rapidly increased in the US than other countries.
6. There is high positive correlation between music category/ digital music category and years.
7. There is high positive correlation between video games/digital video games and years.
8. For all the product categories January and December are highest sale months whereas only wireless product sees August as the maximum sale month
9. Most of the customers are highly satisfied with the purchased products since most of the reviews are with 5 star rating.
10. We can see that for wireless category, though purchase count is highest, customers are not highly satisfied with the products.
11. Whereas for other product categories there is maximum purchase count for star rating 5 than other ratings, which shows that customers buying products other than wireless category are more satisfied.

**References:-**

<https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics>

<https://www.ibm.com/analytics/hadoop/big-data-analytics>

<https://aws.amazon.com/emr/faqs/>