

MIS 6346

Big Data Class Project – Part 1

Shwitaan Sreenivas Ravikumar

SXR190013

Introduction and Problem Description

In this project the amazon review dataset will be analysed and a business case will be developed. The analytics is only based on the reviews given by the users. This is not completely true because not all customers give their reviews. The most important reviews considered are given by the vine members because these reviewers are chosen by Amazon based on the reviews given by them and the accuracy of the ratings.

In this analysis the products are categorized based on different parameters and analysed with respect to the parameters. Initially the dataset is loaded into Hive as a table and by using Hive Query Language the required records are retrieved for analysis. At first, multiple customer reviews for the same product are deleted from the dataset because these reviews are not reliable and it becomes irrelevant.

In the first part of the analysis the work is done generally. The product categories considered are wireless, automotive, music, digital music purchase, sports, toys, digital video games, video games and only the products after 2005 are considered. In the second part, head to head analysis is done between 'Music' and 'Digital music' and between 'Video games' and 'Digital video games'.

The most important parameters considered are the star ratings and the product id. Because these parameters give information about the quality of the product and the most used products. The majority of the analysis is done based on these two parameters. The business models can be built based on these two parameters.

Vine membership also plays a significant part in the analysis. Because this benefit can change the behaviour of the customer and also the type of the products purchased. Products with good ratings given by vine members are brought at a larger scale as analysed in the dataset. Thus it would be more profitable if these products are stocked at warehouse.

The reviews are also being analysed based on the marketplace because the behavior of the customers changes with respect to their culture and the type of place they reside. Thus some products would be very popular in US and it may not be the favorite choice for Europeans. And some products are exclusively available to particular regions. The ratings with respect to marketplace are considered as a very important parameter especially if the particular product is available globally. Thus certain products must be customized based on the regions it is being sold.

Trend analysis is also done in this project which gives insight into the change in the customer behaviours over time. This shows the type of products which gained popularity and also we can predict the type of products which will be mostly sought after in the future. Trend analysis must be referred to produce the type of products which will cut the production cost and it will also reduce the wastage of products.

Step 1:

View of a table is created to consider the records from the year 2005:

```
CREATE VIEW amazon_review.amazon_reviews AS SELECT * FROM  
amazon_review.amazon_reviews_parquet WHERE year >= 2005;
```

Step 2:

View is created with all the duplicate records removed:

```
create view amazon_review.amazon_reviews_include as  
select s.marketplace,t.customer_id,s.review_id,t.product_id,s.product_parent,s.product_title,  
s.star_rating,s.helpful_votes,s.total_votes,s.vine,s.verified_purchase,s.review_headline,  
s.review_body,s.review_date,s.year,t.product_category  
from amazon_review.amazon_reviews s  
join (  
    SELECT  
        customer_id,product_id,product_category,  
        COUNT(*)  
FROM  
    amazon_review.amazon_reviews  
GROUP BY  
    customer_id,product_id,product_category  
HAVING  
    COUNT(*) == 1  
) t on s.customer_id = t.customer_id and  
s.product_id = t.product_id and  
s.product_category = t.product_category
```

Step 3:

Final table is created:

```
CREATE EXTERNAL TABLE amazon_review.amazon_reviews_v2(  
    `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');
```

Step 4:

Records are inserted into the table. Example is given below

```
insert overwrite table amazon_review.amazon_reviews_v2 partition(product_category='Wireless')
select marketplace,customer_id,review_id,product_id,product_parent,product_title,star_rating,
helpful_votes,total_votes,vine,verified_purchase,review_headline,review_body,review_date,year
from amazon_review.amazon_reviews_include where product_category='Wireless';
```

Basic exploratory analysis:

No. of reviews by category

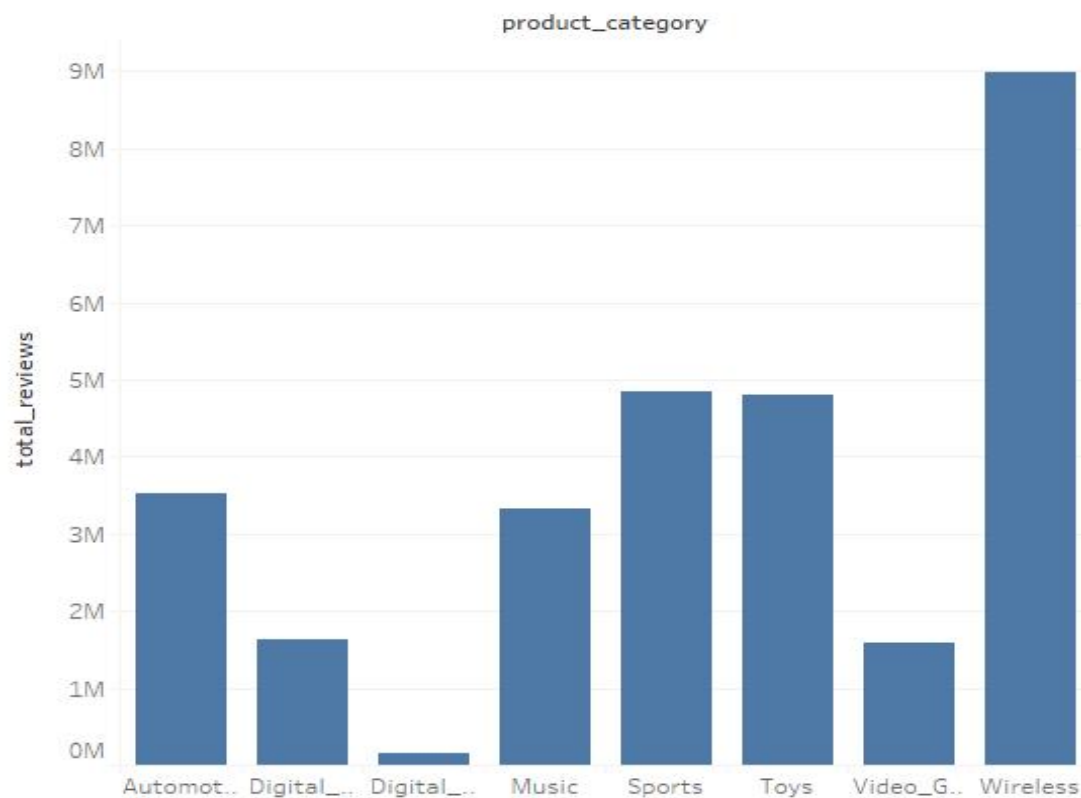
select product_category,count(*) as total_reviews from amazon_review.amazon_reviews_v2 group by product_category;

The screenshot shows a terminal window with the following content:

```
2006 287215
2007 403379
2008 459588
2009 587267
2010 761837
2011 1226751
2012 2244686
2013 5480137
2014 8515834
2015 8544084
Time taken: 30.698 seconds, Fetched: 11 row(s)
hive> set hive.cli.print.header=true;
hive> select product_category,count(*) as total_reviews from amazon_review.amazon_reviews_v2 group by product_category;
Query ID = hadoop_20200412194828_49675a7a-4bf8-4bad-95b8-d928c583c6f8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586702550022_0025)

-----
VERTICES    MODE             STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container    SUCCEEDED   30      30          0         0         0         0
Reducer 2 ... container    SUCCEEDED    2         2          0         0         0         0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 69.83 s
-----
OK
product_category    total_reviews
Automotive          3514995
Digital_Video_Games 145413
Music               3331361
Toys                4795936
Video_Games         1596278
Wireless            8978303
Digital_Music_Purchase 1636189
Sports              4847051
Time taken: 70.248 seconds, Fetched: 8 row(s)
hive>
```

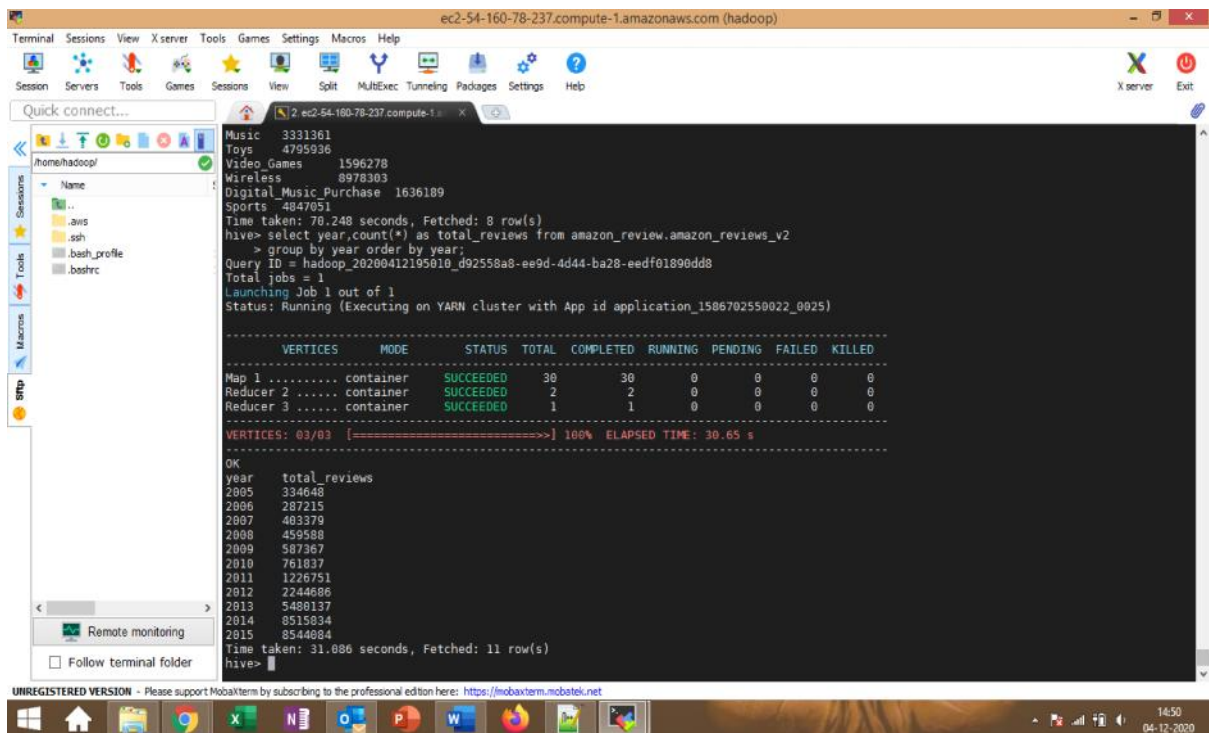
Total Reviews by product category



Sum of total_reviews for each product_category.

Trend analysis of no.of reviews

select year,count(*) as total_reviews from amazon_review.amazon_reviews_v2
group by year order by year;

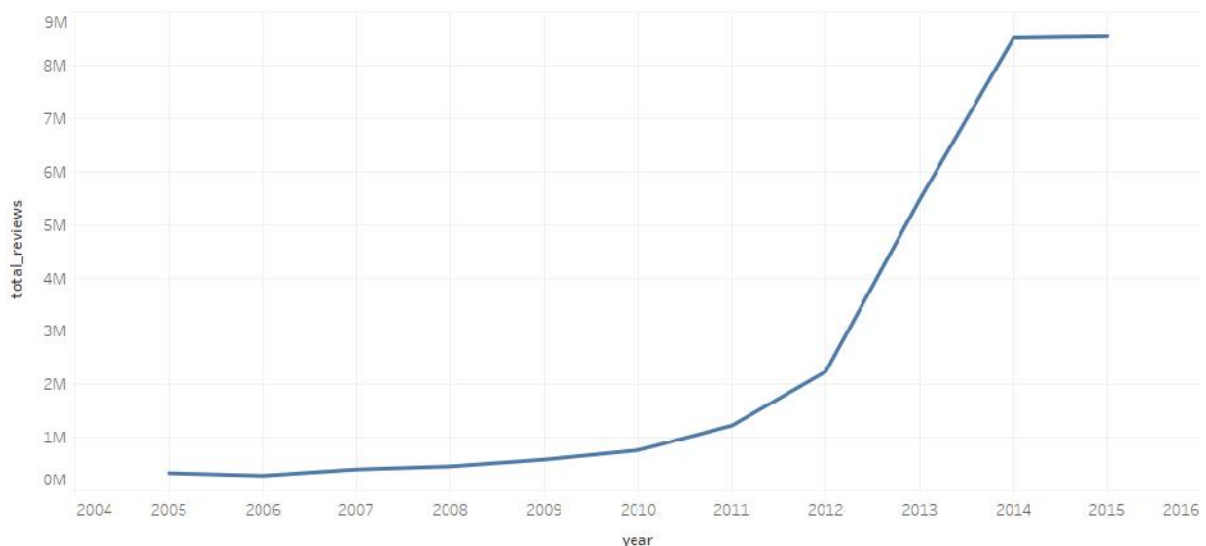


The screenshot shows a terminal window with the following content:

```
Music 3331361
Toys 4795936
Video_Games 1596278
Wireless 8978303
Digital_Music_Purchase 1636189
Sports 4847051
Time taken: 78.248 seconds, Fetched: 8 row(s)
hive> select year,count(*) as total_reviews from amazon_review.amazon_reviews_v2
> group by year order by year;
Query ID = hadoop_20200412195010_d92558a8-ee9d-4d44-ba28-ee9d01890dd8
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586702559022_0025)

-----
VERTICES    MODE    STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED   30         30           0         0         0
Reducer 2 ..... container  SUCCEEDED    2          2           0         0         0
Reducer 3 ..... container  SUCCEEDED    1          1           0         0         0
-----
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 30.65 s
OK
year    total_reviews
2005    334646
2006    287215
2007    483379
2008    459588
2009    587367
2010    761837
2011    1226751
2012    2244686
2013    5488137
2014    8515834
2015    8544884
Time taken: 31.086 seconds, Fetched: 11 row(s)
hive>
```

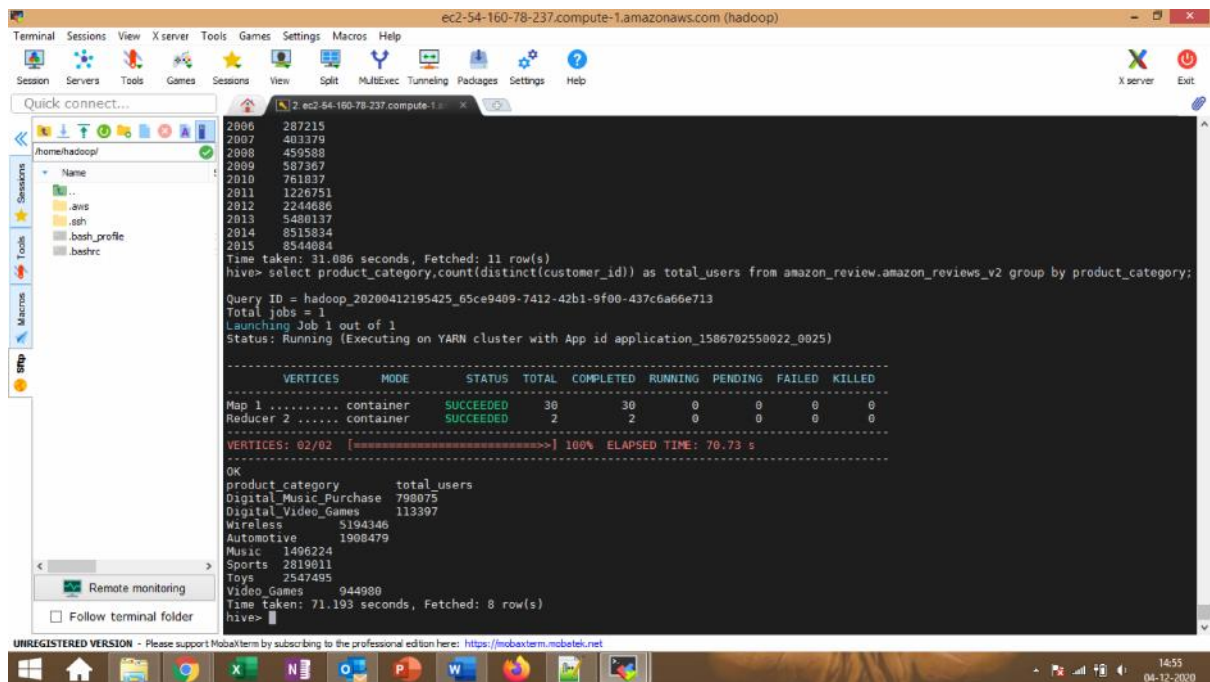
Trend analysis of number of reviews



The trend of sum of total_reviews for year..

No. of users by product_category

select product_category,count(distinct(customer_id)) as total_users from
amazon_review.amazon_reviews_v2 group by product_category;

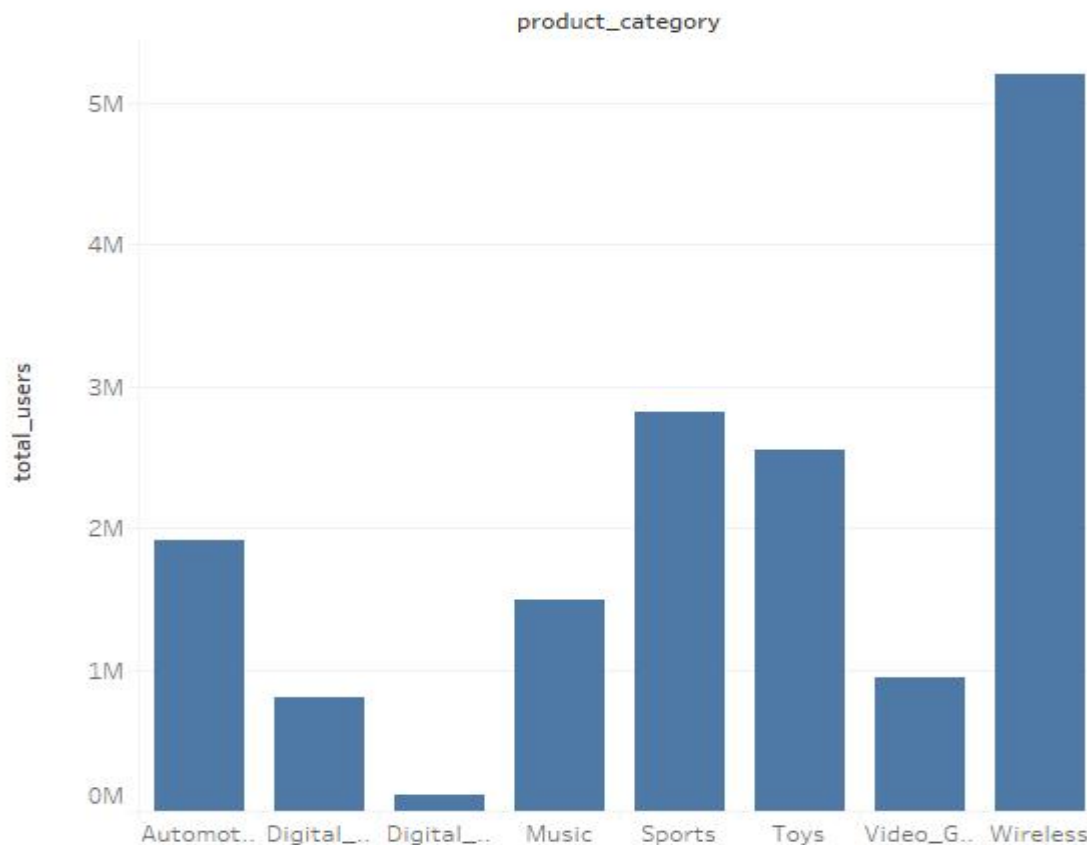


```
2006 287215
2007 481379
2008 459588
2009 587367
2010 761837
2011 1226751
2012 2244686
2013 5486137
2014 8915834
2015 8544084
Time taken: 31.086 seconds, Fetched: 11 row(s)
hive> select product_category,count(distinct(customer_id)) as total_users from amazon_review.amazon_reviews_v2 group by product_category;

Query ID = hadoop_20200412195425_65ce9409-7412-42b1-9f00-437c6a66e713
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586702550022_0025)

-----
VERTICES   MODE                STATUS      TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container    SUCCEEDED    30      30           0         0         0         0
Reducer 2 ... container    SUCCEEDED     2         2           0         0         0         0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 70.73 s
-----
OK
product_category      total_users
Digital_Music_Purchase 798075
Digital_Video_Games    113397
Wireless               5194346
Automotive             1908479
Music                 1496224
Sports                 2819011
Toys                   2547495
Video_Games           944980
Time taken: 71.193 seconds, Fetched: 8 row(s)
hive>
```

Number of users by product category



Sum of total_users for each product_category.

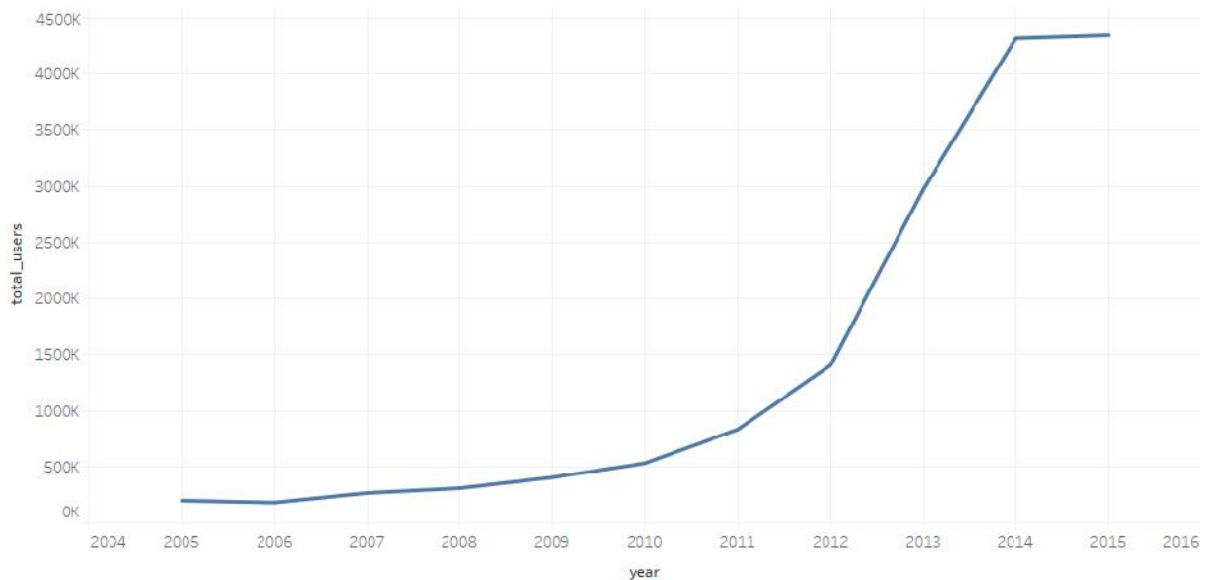
Trend analysis of no.of users

select year,count(distinct(customer_id)) as total_users from amazon_review.amazon_reviews
group by year order by year;

```
ec2-54-160-78-237.compute-1.amazonaws.com (hadoop)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MJExec Tunneling Packages Settings Help
Quick connect...
/home/hadoop/
Name
aws
ssh
.bash_profile
.bashrc
Remote monitoring
Follow terminal folder
Wireless 5194346
Automotive 1908479
Music 1496224
Sports 2019011
Toys 2547495
Video_Games 944980
Time taken: 71.193 seconds, Fetched: 8 row(s)
hive> select year,count(distinct(customer_id)) as total_users from amazon_review.amazon_reviews
> group by year order by year;
Query ID = hadoop_20200412195641_43c6e466-4ecb-46fa-9101-72379847f30f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586702550022_0025)

-----
VERTICES   MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----
Map 1 ..... container SUCCEEDED 30      30      0      0      0      0
Reducer 2 ..... container SUCCEEDED 13      13      0      0      0      0
Reducer 3 ..... container SUCCEEDED 1        1        0      0      0      0
-----
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 71.40 s
OK
year      total_users
2005      183787
2006      170839
2007      253330
2008      207008
2009      301263
2010      516934
2011      814049
2012      1372850
2013      2922165
2014      4258850
2015      4302430
Time taken: 71.965 seconds, Fetched: 11 row(s)
hive>
```

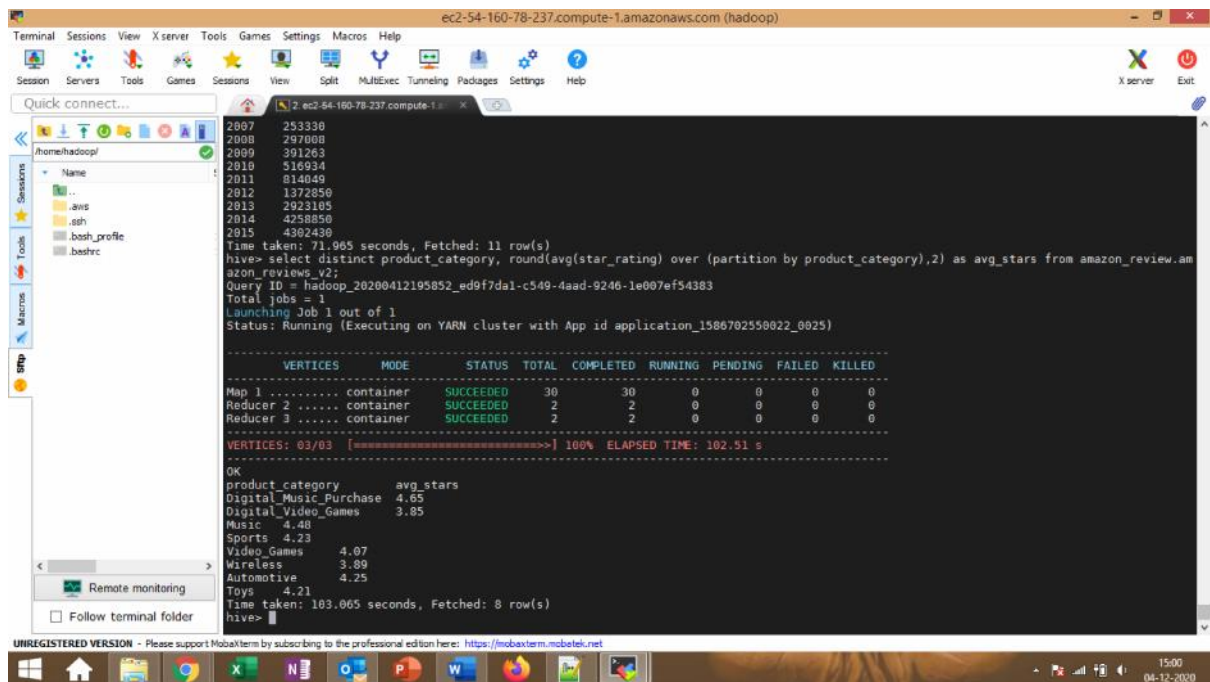
Trend analysis of number of users



The trend of sum of total_users for year.

Average stars by product_category

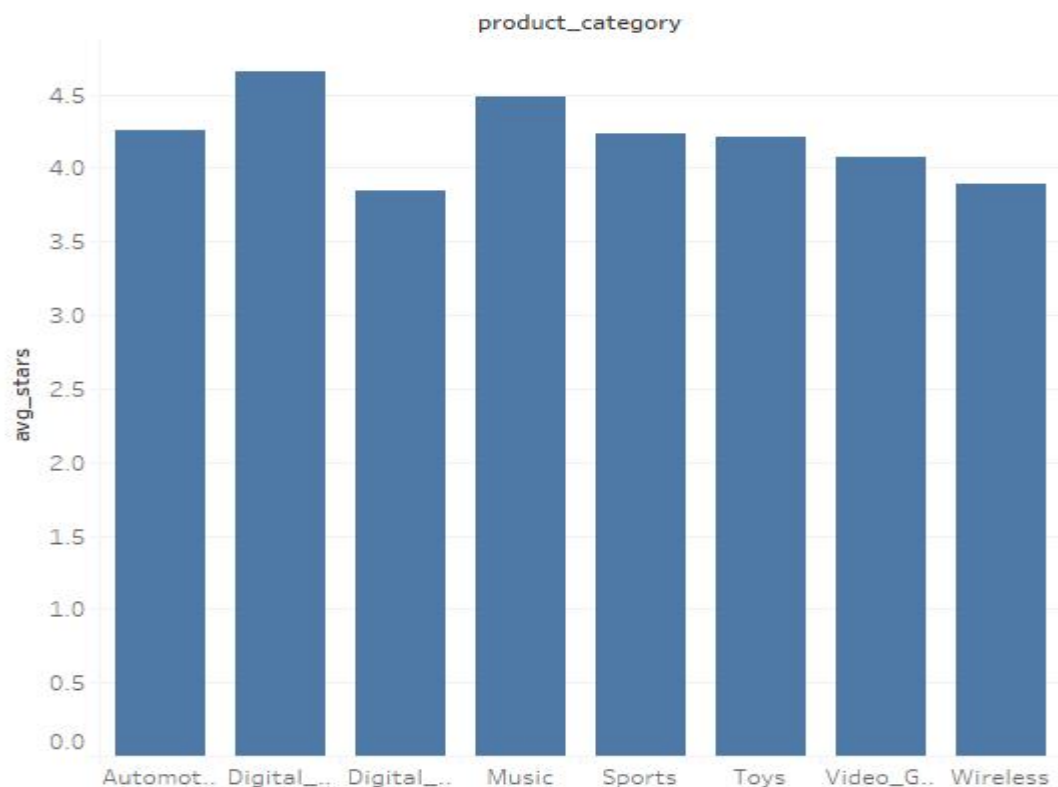
select distinct product_category, round(avg(star_rating) over (partition by product_category),2) as avg_stars from amazon_review.amazon_reviews_v2;



The screenshot shows a terminal window with a Hadoop job execution. The job is named 'hadoop_20200412195852_ed9f7dal-c549-4aad-9246-1e007ef54383'. The output shows the job is running on a YARN cluster. The job is completed with a status of 'SUCCEEDED'. The output shows the average stars for each product category.

product_category	avg_stars
Digital_Music_Purchase	4.65
Digital_Video_Games	3.85
Music	4.48
Sports	4.23
Video_Games	4.07
Wireless	3.89
Automotive	4.25
Toys	4.21

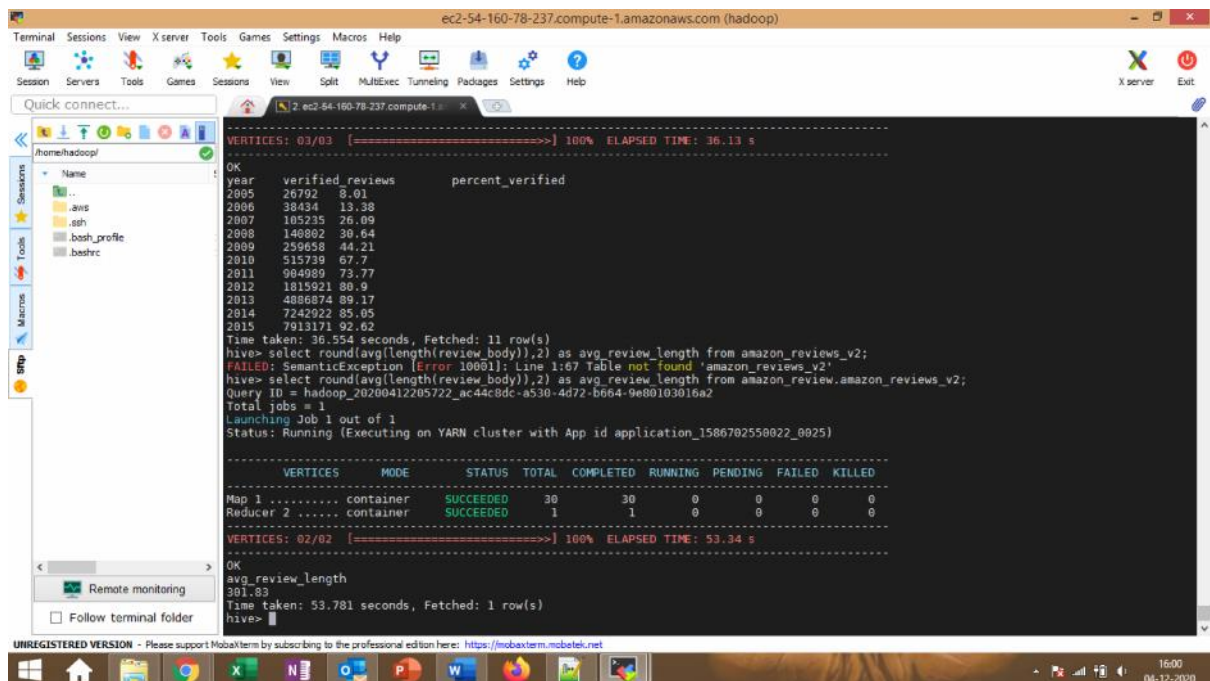
Average stars by product category



Sum of avg_stars for each product_category.

Average length of review

select round(avg(length(review_body)),2) as avg_review_length from
amazon_review.amazon_reviews_v2;

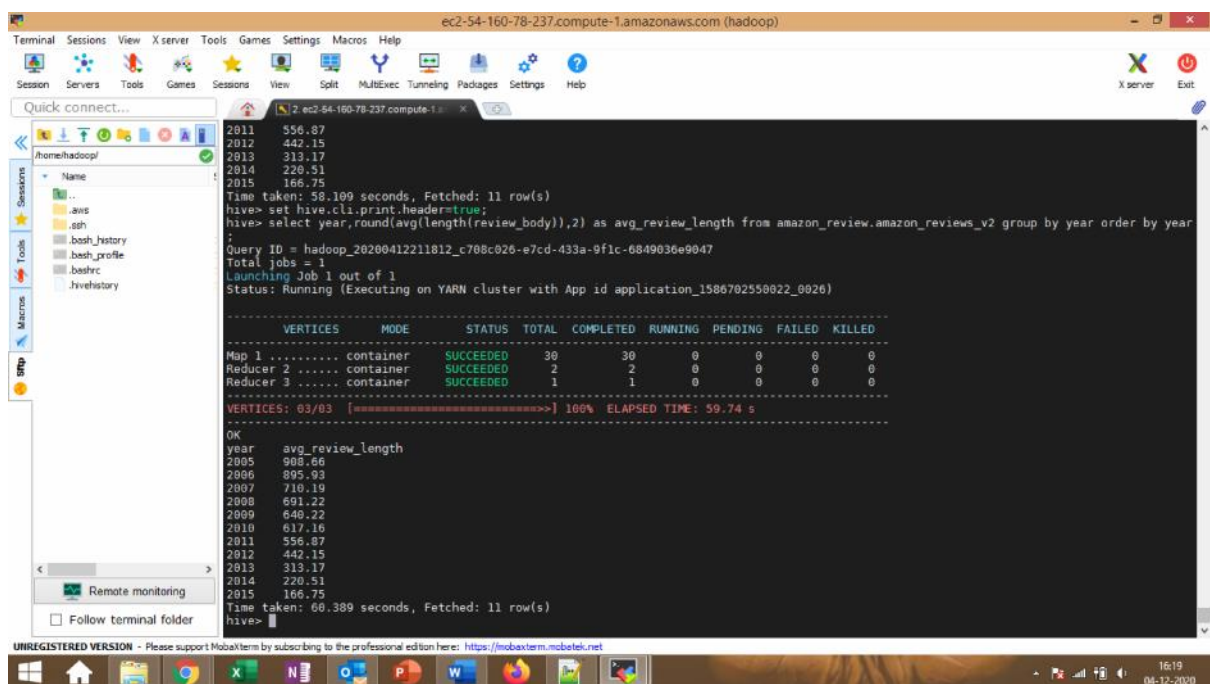


```
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 36.13 s
OK
year    verified_reviews    percent_verified
2005    26792    8.01
2006    38434    13.38
2007    105235    26.09
2008    148802    38.64
2009    259658    44.21
2010    515739    67.7
2011    904989    73.77
2012    1815921    80.9
2013    4886874    89.17
2014    7242922    85.05
2015    7913171    92.62
Time taken: 36.554 seconds, Fetched: 11 row(s)
hive> select round(avg(length(review_body)),2) as avg_review_length from amazon_reviews_v2;
FAILED: SemanticException [Error 10001]: Line 1:67 Table not found 'amazon_reviews_v2'
hive> select round(avg(length(review_body)),2) as avg_review_length from amazon_review.amazon_reviews_v2;
Query ID = hadoop_20200412205722_ac44c8dc-a530-4d72-b664-9e80103016a2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586702550022_0025)

-----
VERTICES    MODE    STATUS    TOTAL    COMPLETED    RUNNING    PENDING    FAILED    KILLED
-----
Map 1 ..... container    SUCCEEDED    30        30        0        0        0        0
Reducer 2 .... container    SUCCEEDED    1         1        0        0        0        0
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 53.34 s
OK
avg_review_length
301.83
Time taken: 53.781 seconds, Fetched: 1 row(s)
hive>
```

Trend analysis of average length of review

select year,round(avg(length(review_body)),2) as avg_review_length from
amazon_review.amazon_reviews_v2 group by year order by year;

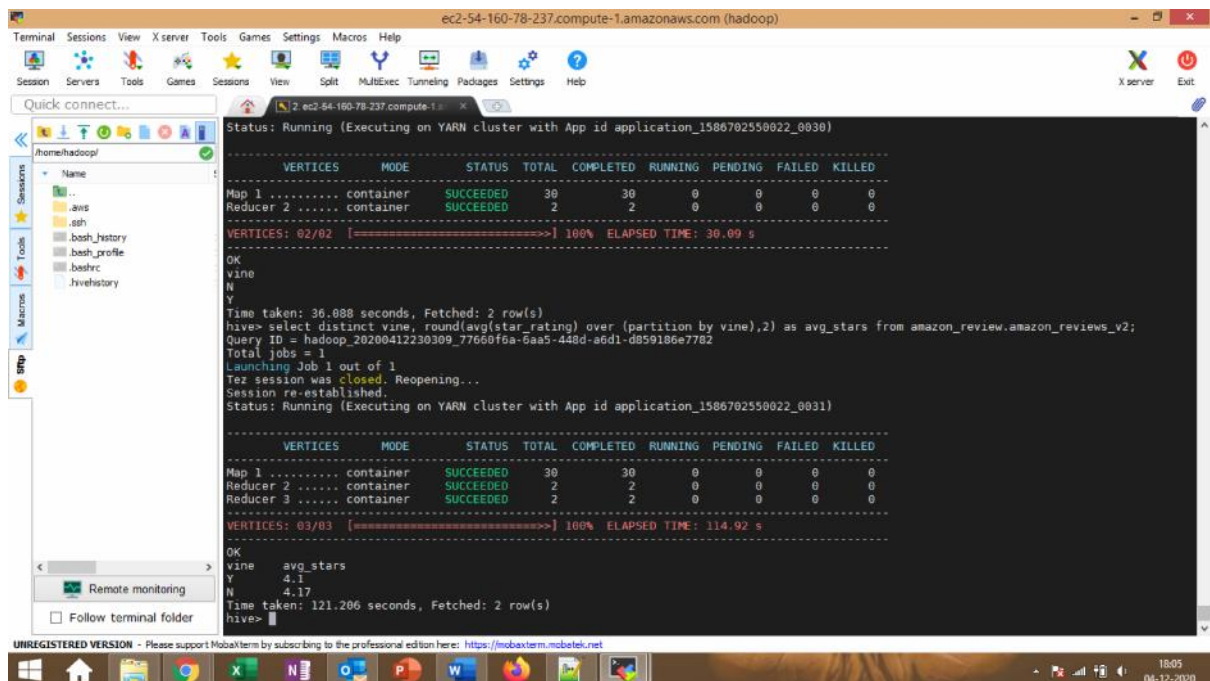


```
2011    556.87
2012    442.15
2013    313.17
2014    220.51
2015    166.75
Time taken: 58.109 seconds, Fetched: 11 row(s)
hive> set hive.cli.print.header=true;
hive> select year,round(avg(length(review_body)),2) as avg_review_length from amazon_review.amazon_reviews_v2 group by year order by year
;
Query ID = hadoop_20200412211812_c708c026-e7cd-433a-9f1c-6849036e9047
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586702550022_0026)

-----
VERTICES    MODE    STATUS    TOTAL    COMPLETED    RUNNING    PENDING    FAILED    KILLED
-----
Map 1 ..... container    SUCCEEDED    30        30        0        0        0        0
Reducer 2 .... container    SUCCEEDED    2         2        0        0        0        0
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 59.74 s
OK
year    avg_review_length
2005    908.66
2006    895.93
2007    710.19
2008    691.22
2009    640.22
2010    617.16
2011    556.87
2012    442.15
2013    313.17
2014    220.51
2015    166.75
Time taken: 68.389 seconds, Fetched: 11 row(s)
hive>
```

Average stars by vine membership

select distinct vine, round(avg(star_rating) over (partition by vine),2) as avg_stars from amazon_review.amazon_reviews_v2;

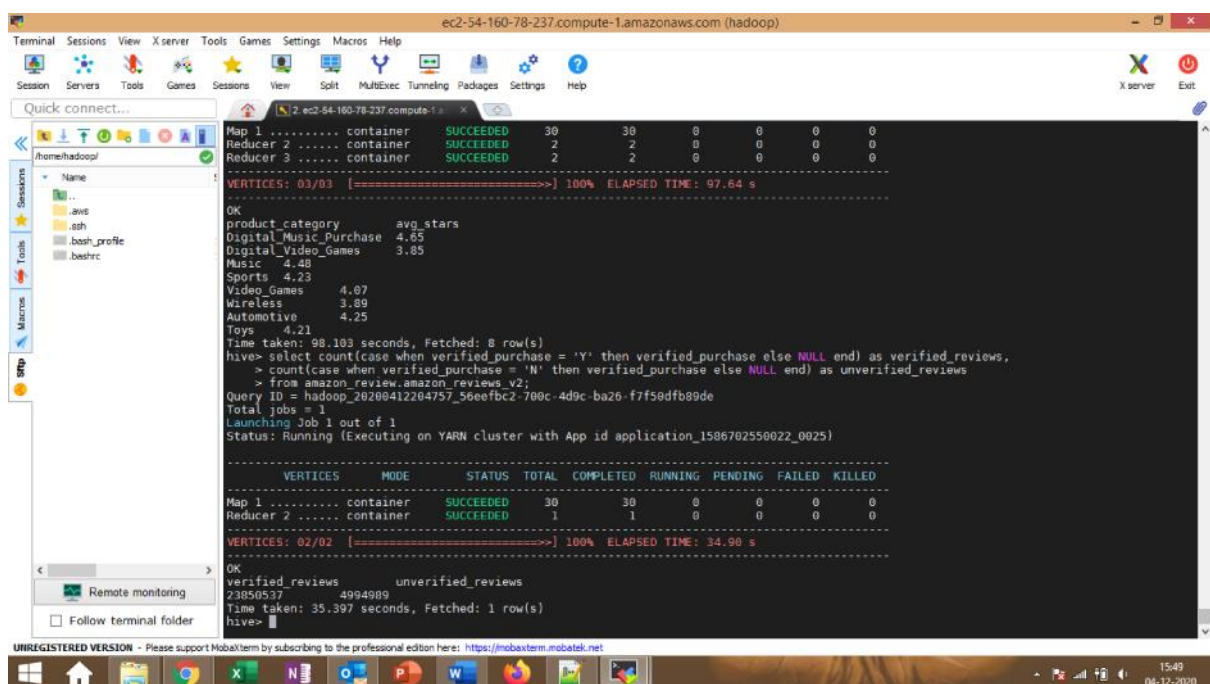


The screenshot shows a terminal window with a Hadoop job running. The job is titled "Status: Running (Executing on YARN cluster with App id application_1586702550022_0030)". The output shows the job progress, including the number of vertices (02/02), the status (SUCCEEDED), and the elapsed time (30.09 s). The job is executed on a YARN cluster with App id application_1586702550022_0031. The output shows the job progress, including the number of vertices (03/03), the status (SUCCEEDED), and the elapsed time (114.92 s). The job is executed on a YARN cluster with App id application_1586702550022_0031. The output shows the job progress, including the number of vertices (03/03), the status (SUCCEEDED), and the elapsed time (114.92 s).

```
ec2-54-160-78-237.compute-1.amazonaws.com (hadoop)
Terminal Sessions View Xserver Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
Quick connect...
/home/hadoop/
Name
...
.ssh
.bash_history
.bash_profile
.bashrc
.hivehistory
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net
18:05 04-12-2020
```

Number of verified versus unverified reviews

select count(case when verified_purchase = 'Y' then verified_purchase else NULL end) as verified_reviews, count(case when verified_purchase = 'N' then verified_purchase else NULL end) as unverified_reviews from amazon_review.amazon_reviews_v2;

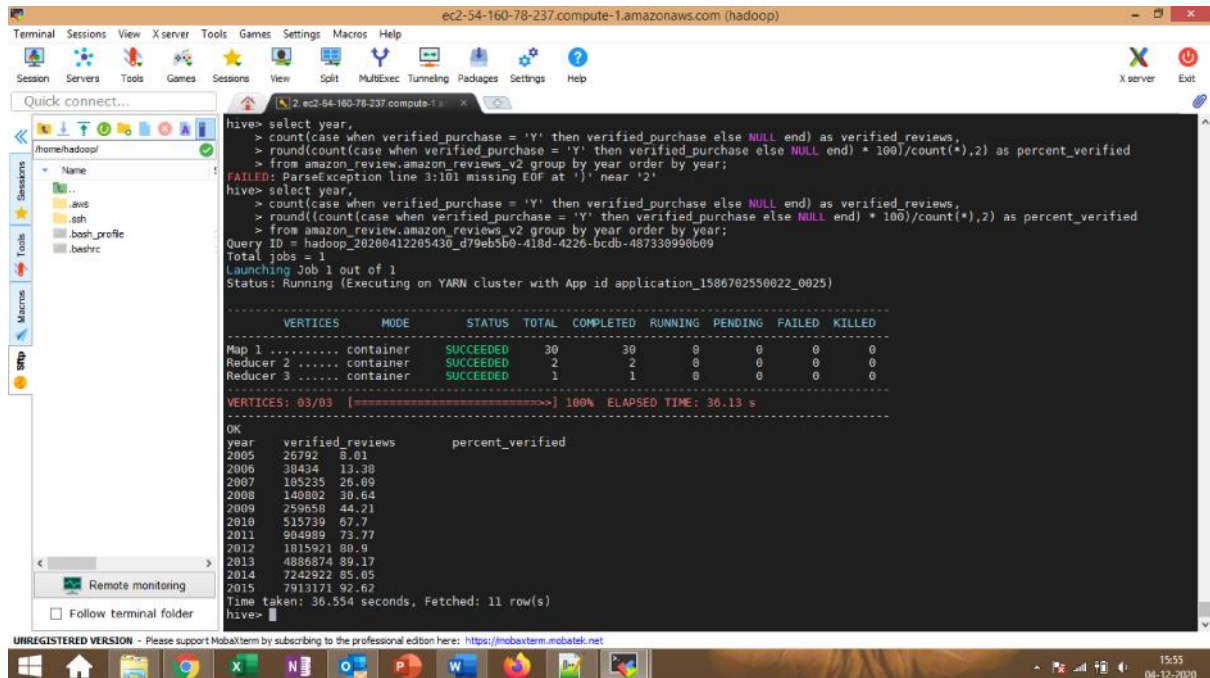


The screenshot shows a terminal window with a Hadoop job running. The job is titled "Status: Running (Executing on YARN cluster with App id application_1586702550022_0025)". The output shows the job progress, including the number of vertices (03/03), the status (SUCCEEDED), and the elapsed time (97.64 s). The job is executed on a YARN cluster with App id application_1586702550022_0025. The output shows the job progress, including the number of vertices (02/02), the status (SUCCEEDED), and the elapsed time (34.90 s). The job is executed on a YARN cluster with App id application_1586702550022_0025. The output shows the job progress, including the number of vertices (02/02), the status (SUCCEEDED), and the elapsed time (34.90 s).

```
ec2-54-160-78-237.compute-1.amazonaws.com (hadoop)
Terminal Sessions View Xserver Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
Quick connect...
/home/hadoop/
Name
...
.ssh
.bash_history
.bash_profile
.bashrc
.hivehistory
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net
15:49 04-12-2020
```

Trend analysis of no. of verified reviews

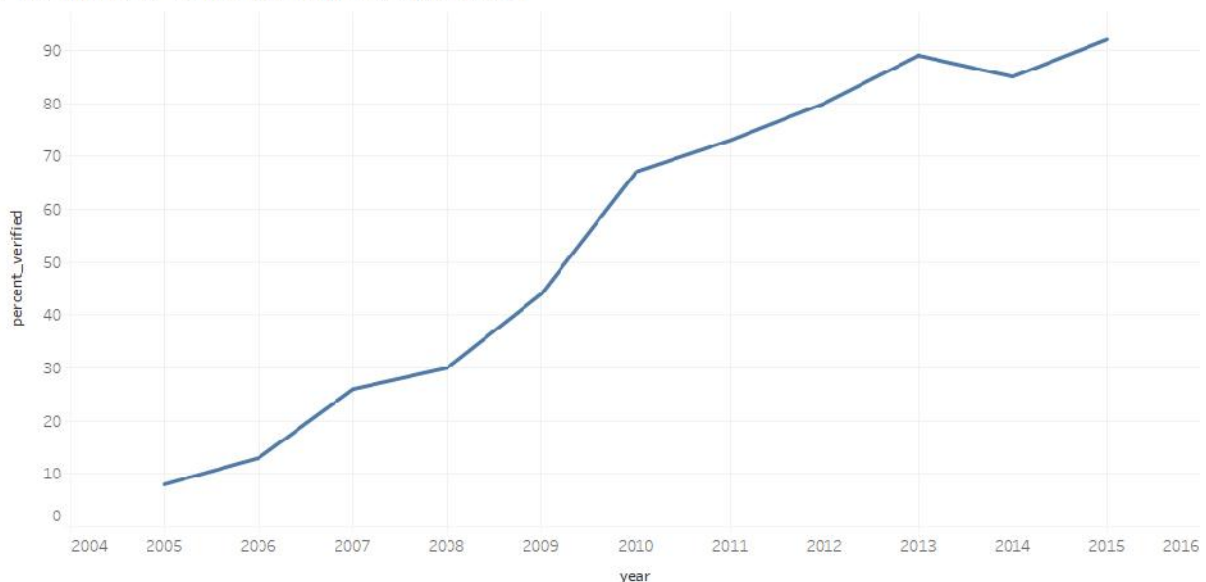
select year, count(case when verified_purchase = 'Y' then verified_purchase else NULL end) as verified_reviews, round((count(case when verified_purchase = 'Y' then verified_purchase else NULL end) * 100)/count(*),2) as percent_verified from amazon_review.amazon_reviews_v2 group by year order by year;



```
hive> select year,
> count(case when verified_purchase = 'Y' then verified_purchase else NULL end) as verified_reviews,
> round(count(case when verified_purchase = 'Y' then verified_purchase else NULL end) * 100)/count(*),2) as percent_verified
> from amazon_review.amazon_reviews_v2 group by year order by year;
FAILED: ParseException line 3:101 missing EOF at ') ' near '2'
hive> select year,
> count(case when verified_purchase = 'Y' then verified_purchase else NULL end) as verified_reviews,
> round(count(case when verified_purchase = 'Y' then verified_purchase else NULL end) * 100)/count(*),2) as percent_verified
> from amazon_review.amazon_reviews_v2 group by year order by year;
Query ID = hadoop_20200412205430_d79eb5b0-418d-4226-bcdb-487330990b09
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586702550022_0025)

-----
VERTICES    MODE        STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----
Map 1 ..... container SUCCEEDED 30      30      0      0      0      0
Reducer 2 .... container SUCCEEDED 2        2      0      0      0      0
Reducer 3 .... container SUCCEEDED 1         1      0      0      0      0
-----
VERTICES: 03/03 [=====] 100% ELAPSED TIME: 36.13 s
-----
OK
year    verified_reviews    percent_verified
2005    26792                8.01
2006    38434               13.38
2007    105235               26.09
2008    140802               30.64
2009    259650               44.21
2010    515739               67.7
2011    904989               73.77
2012    1015921              80.9
2013    4806874              89.17
2014    7242922              85.05
2015    7913171              92.67
Time taken: 36.554 seconds, Fetched: 11 row(s)
hive>
```

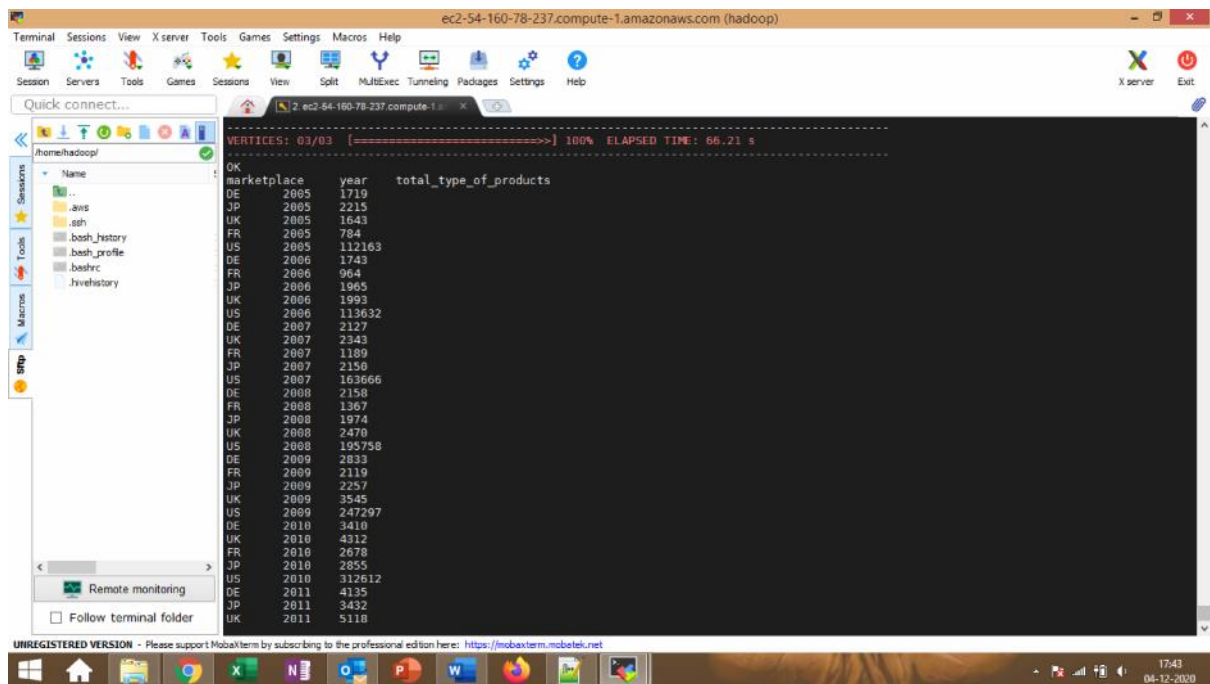
Percentage of verified reviews over the years



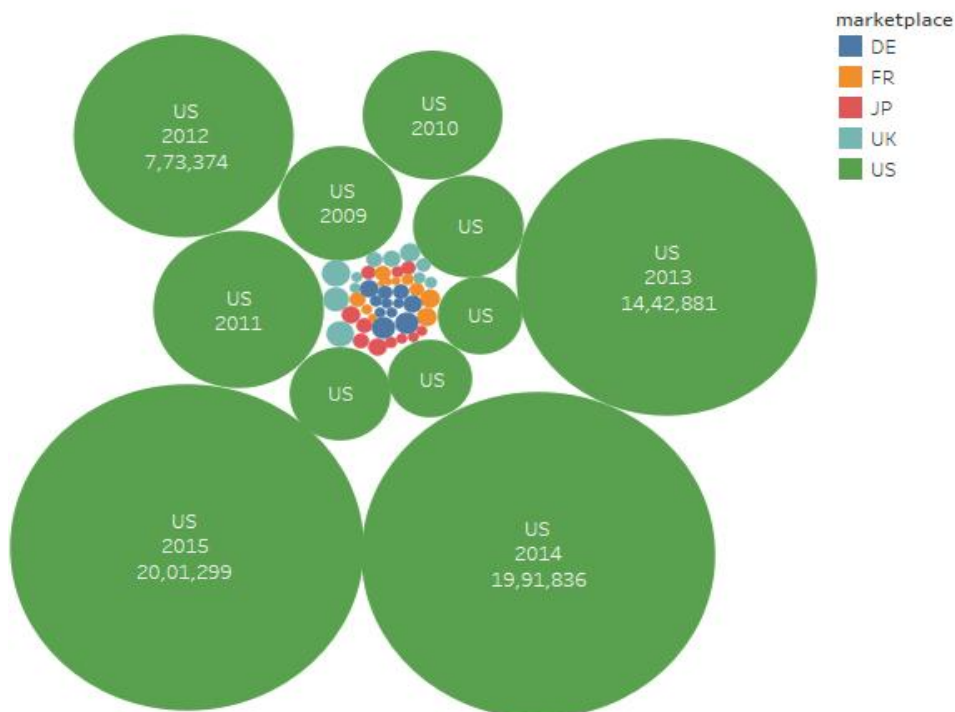
The trend of sum of percent_ver fied for year.

Trend analysis of no. of different types of products by marketplace

select marketplace,year,count(distinct(product_id)) as total_type_of_products from amazon_review.amazon_reviews_v2 group by marketplace,year order by year;



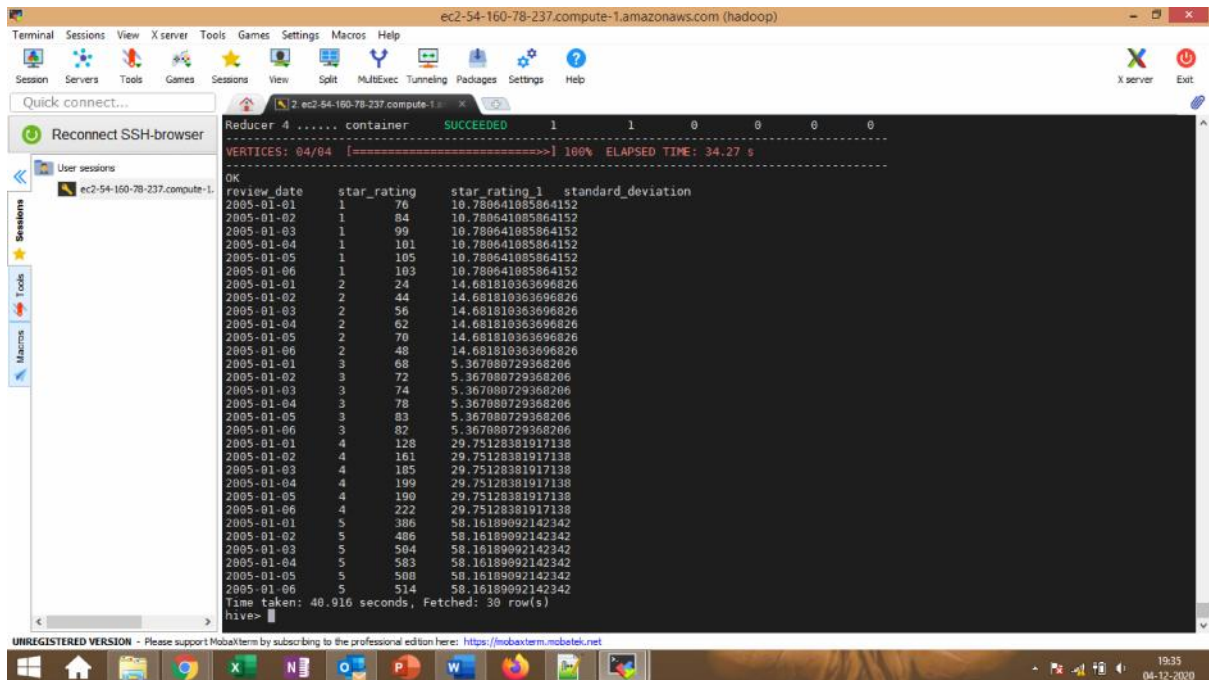
Sheet 1



Marketplace, year and sum of total_type_of_products. Colour shows details about marketplace. Size shows sum of total_type_of_products. The marks are labelled by marketplace, year and sum of total_type_of_products.

Standard deviation in star ratings

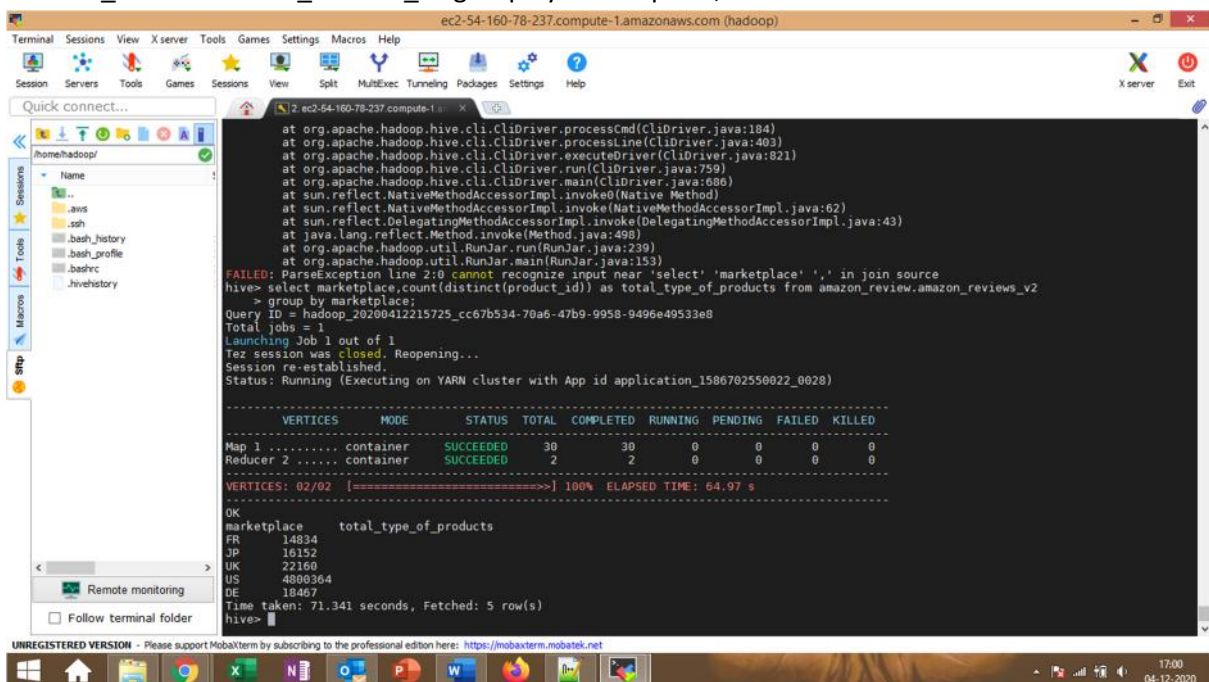
select review_date,star_rating,star_rating_1,stddev(star_rating_1) over (partition by star_rating order by review_date,star_rating asc rows between unbounded preceding and unbounded following) as standard_deviation from (select review_date,star_rating,count(star_rating) as star_rating_1 from amazon_review.amazon_reviews_v2 group by review_date,star_rating order by review_date,star_rating asc limit 30)s;



```
Reducer 4 ..... container SUCCEEDED 1 1 0 0 0 0
VERTICES: 04/04 [=====] 100% ELAPSED TIME: 34.27 s
OK
review_date  star_rating  star_rating_1  standard_deviation
2005-01-01    1          76    10.780641085864152
2005-01-02    1          84    10.780641085864152
2005-01-03    1          99    10.780641085864152
2005-01-04    1         101    10.780641085864152
2005-01-05    1         105    10.780641085864152
2005-01-06    1         103    10.780641085864152
2005-01-01    2          24    14.681810363696826
2005-01-02    2          44    14.681810363696826
2005-01-03    2          56    14.681810363696826
2005-01-04    2          62    14.681810363696826
2005-01-05    2          70    14.681810363696826
2005-01-06    2          48    14.681810363696826
2005-01-01    3          68    5.367080729368206
2005-01-02    3          72    5.367080729368206
2005-01-03    3          74    5.367080729368206
2005-01-04    3          78    5.367080729368206
2005-01-05    3          83    5.367080729368206
2005-01-06    3          82    5.367080729368206
2005-01-01    4         128    29.75128381917138
2005-01-02    4         161    29.75128381917138
2005-01-03    4         185    29.75128381917138
2005-01-04    4         199    29.75128381917138
2005-01-05    4         190    29.75128381917138
2005-01-06    4         222    29.75128381917138
2005-01-01    5         306    58.16189992142342
2005-01-02    5         486    58.16189992142342
2005-01-03    5         504    58.16189992142342
2005-01-04    5         583    58.16189992142342
2005-01-05    5         588    58.16189992142342
2005-01-06    5         514    58.16189992142342
Time taken: 40.916 seconds, Fetched: 39 row(s)
hive>
```

No. of different types of products by marketplace

select marketplace,count(distinct(product_id)) as total_type_of_products from amazon_review.amazon_reviews_v2 group by marketplace;



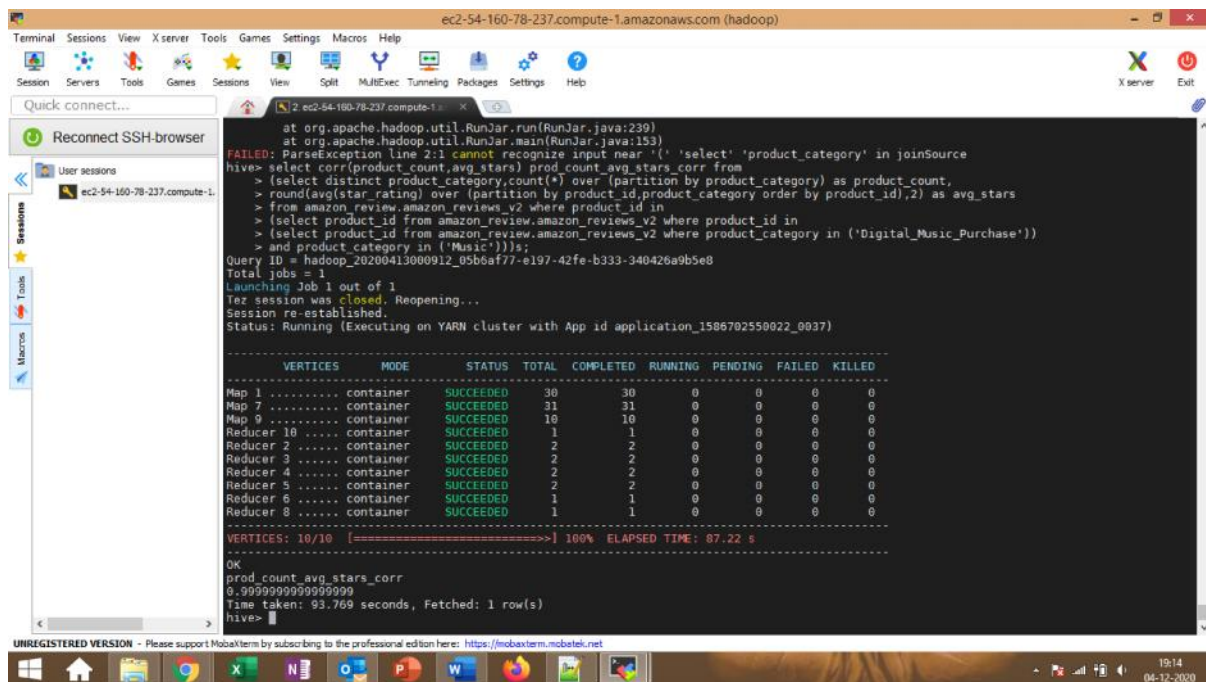
```
at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:184)
at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:403)
at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:686)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
at org.apache.hadoop.util.RunJar.main(RunJar.java:153)
FAILED: ParseException line 2:0 cannot recognize input near 'select' 'marketplace' ',' in join source
hive> select marketplace,count(distinct(product_id)) as total_type_of_products from amazon_review.amazon_reviews_v2
> group by marketplace;
Query ID = hadoop_20200412215725_cc67b534-70a6-47b9-9958-9496e49533e8
Total jobs = 1
Launching Job 1 out of 1
tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1586702550022_0028)

VERTICES    MODE    STATUS    TOTAL    COMPLETED    RUNNING    PENDING    FAILED    KILLED
Map 1 ..... container SUCCEEDED 30      30      0      0      0      0
Reducer 2 ..... container SUCCEEDED 2       2       0      0      0      0
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 64.97 s
OK
marketplace  total_type_of_products
FR           14834
JP           16152
UK           22160
US           4809364
DE           18467
Time taken: 71.341 seconds, Fetched: 5 row(s)
hive>
```

Detailed analysis of Music/Digital_Music_Purchase and Digital_Video_Games/Video_Games over time.

Correlation between Music and Digital_Music_Purchase

select corr(product_count,avg_stars) prod_count_avg_stars_corr from (select distinct product_category,count(*) over (partition by product_category) as product_count, round(avg(star_rating) over (partition by product_id,product_category order by product_id),2) as avg_stars from amazon_review.amazon_reviews_v2 where product_id in (select product_id from amazon_review.amazon_reviews_v2 where product_id in (select product_id from amazon_review.amazon_reviews_v2 where product_category in ('Digital_Music_Purchase')) and product_category in ('Music')));

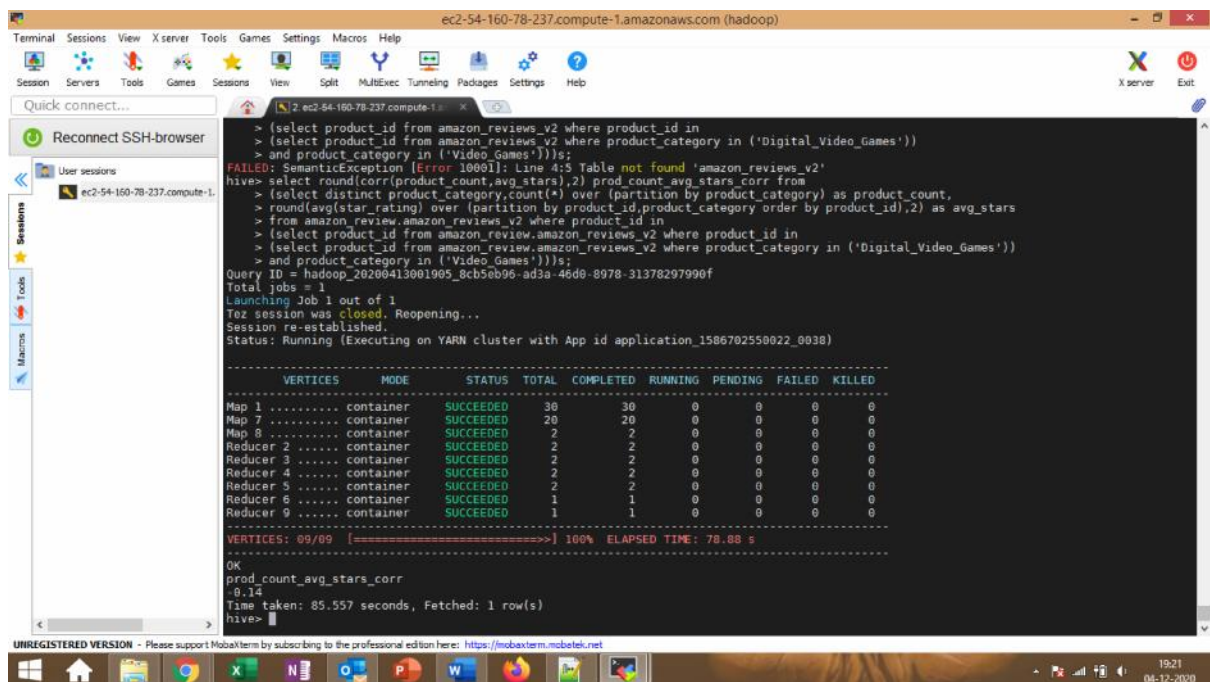


```
at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
at org.apache.hadoop.util.RunJar.main(RunJar.java:153)
FAILED: ParseException line 2:1 cannot recognize input near '(' 'select' 'product_category' in joinSource
hive> select corr(product_count,avg_stars) prod_count_avg_stars_corr from
> (select distinct product_category,count(*) over (partition by product_category) as product_count,
> round(avg(star_rating) over (partition by product_id,product_category order by product_id),2) as avg_stars
> from amazon_review.amazon_reviews_v2 where product_id in
> (select product_id from amazon_review.amazon_reviews_v2 where product_id in
> (select product_id from amazon_review.amazon_reviews_v2 where product_category in ('Digital_Music_Purchase'))
> and product_category in ('Music')));
Query ID = hadoop_20200413000912_05b6af77-e197-42fe-b333-340426a9b5e8
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1586702550022_0037)

-----
VERTICES    MODE             STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  30      30           0         0         0         0
Map 7 ..... container  SUCCEEDED  31      31           0         0         0         0
Map 9 ..... container  SUCCEEDED  10      10           0         0         0         0
Reducer 10 ... container  SUCCEEDED   1         1           0         0         0         0
Reducer 2 ..... container  SUCCEEDED   2         2           0         0         0         0
Reducer 3 ..... container  SUCCEEDED   2         2           0         0         0         0
Reducer 4 ..... container  SUCCEEDED   2         2           0         0         0         0
Reducer 5 ..... container  SUCCEEDED   2         2           0         0         0         0
Reducer 6 ..... container  SUCCEEDED   1         1           0         0         0         0
Reducer 8 ..... container  SUCCEEDED   1         1           0         0         0         0
-----
VERTICES: 10/10 [=====] 100% ELAPSED TIME: 07.22 s
-----
OK
prod_count_avg_stars_corr
0.9999999999999999
Time taken: 93.769 seconds, Fetched: 1 row(s)
hive>
```

Correlation between Digital_Video_Games and Video_Games

```
select round(corr(product_count,avg_stars),2) prod_count_avg_stars_corr from
(select distinct product_category,count(*) over (partition by product_category) as product_count,
round(avg(star_rating) over (partition by product_id,product_category order by product_id),2) as
avg_stars from amazon_reviews_v2 where product_id in (select product_id from
amazon_reviews_v2 where product_id in (select product_id from amazon_reviews_v2 where
product_category in ('Digital_Video_Games')) and product_category in ('Video_Games')));
```



```
ec2-54-160-78-237.compute-1.amazonaws.com (hadoop)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultExec Tunneling Packages Settings Help
Quick connect...
Reconnect SSH-browser
User sessions
ec2-54-160-78-237.compute-1
Sessions
Tools
Macros
ec2-54-160-78-237.compute-1
> (select product_id from amazon_reviews_v2 where product_id in
> (select product_id from amazon_reviews_v2 where product_category in ('Digital_Video_Games'))
> and product_category in ('Video_Games')));
FAILED: SemanticException [Error 10001]: Line 4:5 Table not found 'amazon_reviews_v2'
hive> select round(corr(product_count,avg_stars),2) prod_count_avg_stars_corr from
> (select distinct product_category,count(*) over (partition by product_category) as product_count,
> round(avg(star_rating) over (partition by product_id,product_category order by product_id),2) as avg_stars
> from amazon_review.amazon_reviews_v2 where product_id in
> (select product_id from amazon_review.amazon_reviews_v2 where product_id in
> (select product_id from amazon_review.amazon_reviews_v2 where product_category in ('Digital_Video_Games'))
> and product_category in ('Video_Games')));
Query ID = hadoop_20200413001905_0cb50b96-ad3a-46d0-8978-31378297990f
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1506702550022_0030)

-----
VERTICES      MODE        STATUS      TOTAL   COMPLETED   RUNNING   PENDING   FAILED   KILLED
-----
Map 1 ..... container   SUCCEEDED    30         30           0         0         0         0
Map 7 ..... container   SUCCEEDED    20         20           0         0         0         0
Map 8 ..... container   SUCCEEDED    2          2           0         0         0         0
Reducer 2 .... container   SUCCEEDED    2          2           0         0         0         0
Reducer 3 .... container   SUCCEEDED    2          2           0         0         0         0
Reducer 4 .... container   SUCCEEDED    2          2           0         0         0         0
Reducer 5 .... container   SUCCEEDED    2          2           0         0         0         0
Reducer 6 .... container   SUCCEEDED    1          1           0         0         0         0
Reducer 9 .... container   SUCCEEDED    1          1           0         0         0         0
-----
VERTICES: 09/09 [=====] 100% ELAPSED TIME: 78.88 s
-----
OK
prod_count_avg_stars_corr
0.14
Time taken: 85.557 seconds, Fetched: 1 row(s)
hive>
```


Total number of users reviewing in both Music and Digital_Music_Purchase

select count(distinct(customer_id)) from amazon_review.amazon_reviews_v2 where customer_id in (select customer_id from amazon_review.amazon_reviews_v2 where product_category in ('Digital_Music_Purchase')) and product_category in ('Music');

```
at sun.reflect.NativeMethodAccessorImpl.invoke(Native Method)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:62)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:239)
at org.apache.hadoop.util.RunJar.main(RunJar.java:153)
FAILED: ParseException line 1:102 cannot recognize input near '(' 'select' 'marketplace' in joinSource
hive> select count(distinct(customer_id)) from amazon_review.amazon_reviews_v2 where customer_id in
> (select customer_id from amazon_review.amazon_reviews_v2 where product_category in ('Digital_Music_Purchase'))
> and product_category in ('Music');
FAILED: SemanticException [Error 10001]: Line 1:41 Table not found 'amazon_reviews_v2'
hive> select count(distinct(customer_id)) from amazon_review.amazon_reviews_v2 where customer_id in
> (select customer_id from amazon_review.amazon_reviews_v2 where product_category in ('Digital_Music_Purchase'))
> and product_category in ('Music');
Query ID = hadoop_20200412231815_d7b1dc2-3194-4f49-a3bd-3bc99e50f281
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running [Executing on YARN cluster with App id application_1586702550022_0032]

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  31      31          0         0         0         0
Map 5 ..... container  SUCCEEDED  10      10          0         0         0         0
Reducer 2 ... container  SUCCEEDED   1         1          0         0         0         0
Reducer 3 ... container  SUCCEEDED   1         1          0         0         0         0
Reducer 4 ... container  SUCCEEDED   1         1          0         0         0         0
Reducer 6 ... container  SUCCEEDED   1         1          0         0         0         0
-----
VERTICES: 06/06 [=====] 100% ELAPSED TIME: 44.98 s
-----
OK
c0
140797
Time taken: 52.02 seconds, Fetched: 1 row(s)
hive>
```

Total number of users reviewing in both Digital_Video_Games and Video_Games

select count(distinct(customer_id)) from amazon_review.amazon_reviews_v2 where customer_id in (select customer_id from amazon_review.amazon_reviews_v2 where product_category in ('Digital_Video_Games')) and product_category in ('Video_Games');

```
Map 1 ..... container  SUCCEEDED  28      28          0         0         0         0
Map 5 ..... container  SUCCEEDED  10      10          0         0         0         0
Reducer 2 ... container  SUCCEEDED   1         1          0         0         0         0
Reducer 3 ... container  SUCCEEDED   1         1          0         0         0         0
Reducer 4 ... container  SUCCEEDED   1         1          0         0         0         0
Reducer 6 ... container  SUCCEEDED   1         1          0         0         0         0
-----
VERTICES: 06/06 [=====] 100% ELAPSED TIME: 42.99 s
-----
OK
c0
140797
Time taken: 46.246 seconds, Fetched: 1 row(s)
hive> select count(distinct(customer_id)) from amazon_review.amazon_reviews_v2 where customer_id in
> (select customer_id from amazon_review.amazon_reviews_v2 where product_category in ('Digital_Video_Games'))
> and product_category in ('Video_Games');
Query ID = hadoop_20200412233214_a22bbf17-b3f8-4be1-8dea-d7de61a557ad
Total jobs = 1
Launching Job 1 out of 1
Status: Running [Executing on YARN cluster with App id application_1586702550022_0034]

-----
VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container  SUCCEEDED  20      20          0         0         0         0
Map 4 ..... container  SUCCEEDED   2         2          0         0         0         0
Reducer 2 ... container  SUCCEEDED   1         1          0         0         0         0
Reducer 3 ... container  SUCCEEDED   1         1          0         0         0         0
Reducer 5 ... container  SUCCEEDED   1         1          0         0         0         0
-----
VERTICES: 05/05 [=====] 100% ELAPSED TIME: 35.94 s
-----
OK
c0
29762
Time taken: 36.604 seconds, Fetched: 1 row(s)
hive>
```

Average rating of similar products in Music and Digital_Music_Purchase

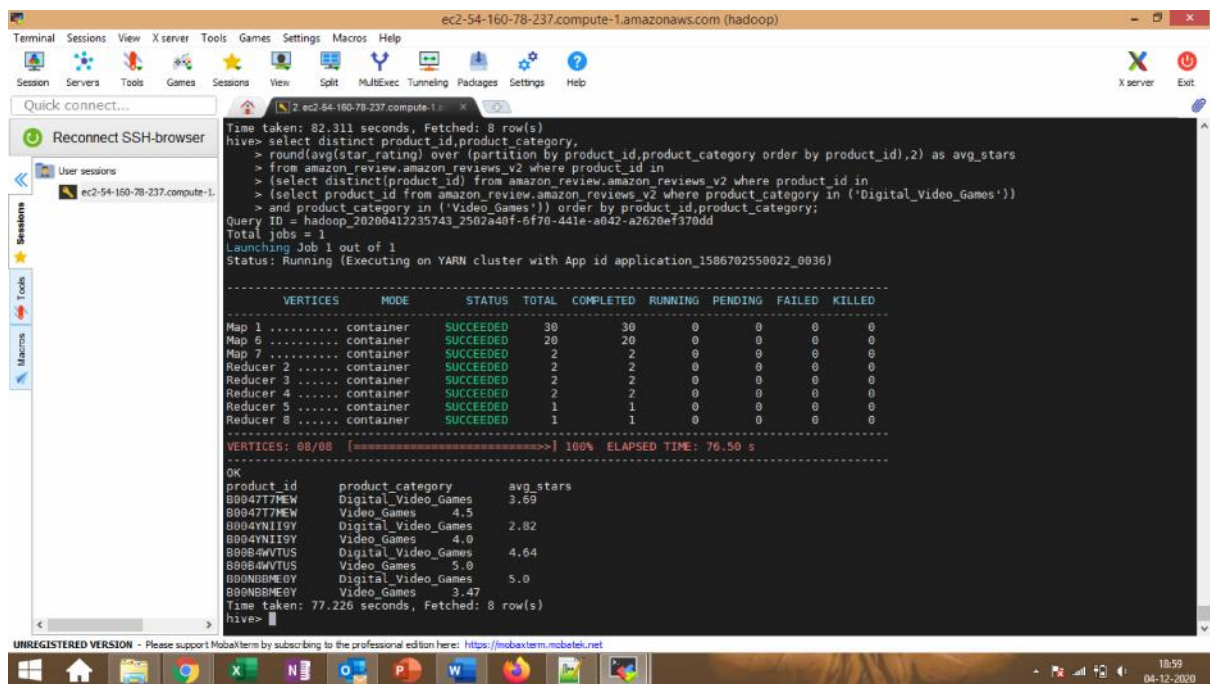
```
select distinct product_id,product_category, round(avg(star_rating) over (partition by
product_id,product_category order by product_id),2) as avg_stars from
amazon_review.amazon_reviews_v2 where product_id in (select distinct(product_id) from
amazon_review.amazon_reviews_v2 where product_id in (select product_id from
amazon_review.amazon_reviews_v2 where product_category in ('Music')) and product_category in
('Digital_Music_Purchase'));
```

```
OK
product_id
8004777MEW
8004YN1I9Y
800BdWVTU5
800NB8ME9Y
Time taken: 33.833 seconds, Fetched: 4 row(s)
hive> select distinct product_id,product_category,
> round(avg(star_rating) over (partition by product_id,product_category order by product_id),2) as avg_stars
> from amazon_review.amazon_reviews_v2 where product_id in
> (select distinct(product_id) from amazon_review.amazon_reviews_v2 where product_id in
> (select product_id from amazon_review.amazon_reviews_v2 where product_category in ('Music')))
> and product_category in ('Digital_Music_Purchase'));
Query ID = hadoop_20200412234510_9bdc4aeb-304b-48d2-a752-7ccc2f043f54
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1506702550022_0035)

-----
VERTICES      MODE        STATUS      TOTAL   COMPLETED   RUNNING   PENDING   FAILED   KILLED
-----
Map 1 ..... container  SUCCEEDED    30         30           0         0         0         0
Map 5 ..... container  SUCCEEDED    10         10           0         0         0         0
Map 7 ..... container  SUCCEEDED    31         31           0         0         0         0
Reducer 2 .... container  SUCCEEDED     2          2           0         0         0         0
Reducer 3 .... container  SUCCEEDED     2          2           0         0         0         0
Reducer 4 .... container  SUCCEEDED     2          2           0         0         0         0
Reducer 6 .... container  SUCCEEDED     1          1           0         0         0         0
Reducer 8 .... container  SUCCEEDED     1          1           0         0         0         0
-----
VERTICES: 08/08 [=====] 100% ELAPSED TIME: 07.58 s
-----
OK
product_id      product_category  avg_stars
80019M1ZJ5      Digital_Music_Purchase  5.0
80019M1ZJ5      Music              3.0
Time taken: 88.419 seconds, Fetched: 2 row(s)
hive>
```

Average rating of similar products in Digital_Video_Games and Video_Games

```
select distinct product_id,product_category, round(avg(star_rating) over (partition by
product_id,product_category order by product_id),2) as avg_stars from
amazon_review.amazon_reviews_v2 where product_id in (select distinct(product_id) from
amazon_review.amazon_reviews_v2 where product_id in (select product_id from
amazon_review.amazon_reviews_v2 where product_category in ('Digital_Video_Games'))
and product_category in ('Video_Games')) order by product_id,product_category;
```



The screenshot shows a terminal window with the following content:

```
Time taken: 82.311 seconds, Fetched: 8 row(s)
hive> select distinct product_id,product_category,
> round(avg(star_rating) over (partition by product_id,product_category order by product_id),2) as avg_stars
> from amazon_review.amazon_reviews_v2 where product_id in
> (select distinct(product_id) from amazon_review.amazon_reviews_v2 where product_id in
> (select product_id from amazon_review.amazon_reviews_v2 where product_category in ('Digital_Video_Games'))
> and product_category in ('Video_Games')) order by product_id,product_category;
Query ID = hadoop_20200412235743_2502a40f-6f70-441e-a042-a2620ef370dd
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1506702550022_0036)
```

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	30	30	0	0	0	0
Map 6	container	SUCCEEDED	20	20	0	0	0	0
Map 7	container	SUCCEEDED	2	2	0	0	0	0
Reducer 2	container	SUCCEEDED	2	2	0	0	0	0
Reducer 3	container	SUCCEEDED	2	2	0	0	0	0
Reducer 4	container	SUCCEEDED	2	2	0	0	0	0
Reducer 5	container	SUCCEEDED	1	1	0	0	0	0
Reducer 8	container	SUCCEEDED	1	1	0	0	0	0

```
VERTICES: 08/08 [=====] 100% ELAPSED TIME: 76.50 s
OK
product_id      product_category  avg_stars
0004777MEW      Digital_Video_Games  3.89
0004777MEW      Video_Games         4.5
0004YNII9Y      Digital_Video_Games  2.82
0004YNII9Y      Video_Games         4.0
00084MVTUS      Digital_Video_Games  4.64
00084MVTUS      Video_Games         5.0
000NBME0Y       Digital_Video_Games
000NBME0Y       Video_Games         3.47
Time taken: 77.226 seconds, Fetched: 8 row(s)
hive>
```


Number of vine users in Music and Digital_Music_Purchase

select product_category,count(distinct(customer_id)) as total_users from
amazon_review.amazon_reviews_v2 where product_category in ('Music','Digital_Music_Purchase')
and vine='Y' group by product_category;

The screenshot shows a MobaXterm terminal window with a Hive query execution. The query is: `select product_category,count(distinct(customer_id)) as total_users from amazon_review.amazon_reviews_v2 where product_category in ('Music','Digital_Music_Purchase') and vine='Y' group by product_category;` The output shows two rows: Music with 1109 total users and Digital_Music_Purchase with 109 total users. The execution time is 28.559 seconds.

```
Query ID = hadoop_20200413013519_4a6fb9dc-88d7-4a16-aa5d-af37f2ddf409
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1586702550022_0043)

-----
VERTICES    MODE        STATUS      TOTAL   COMPLETED   RUNNING   PENDING   FAILED   KILLED
Map 1 ..... container    SUCCEEDED    30       30           0         0         0         0
Reducer 2 ... container    SUCCEEDED     1         1           0         0         0         0
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 28.85 s
-----
OK
product_category    total_users
Music                1109
Time taken: 28.172 seconds, Fetched: 1 row(s)
hive> select product_category,count(distinct(customer_id)) as total_users from amazon_review.amazon_reviews_v2 where product_category in
('Digital_Music_Purchase') and vine='Y' group by product_category;
Query ID = hadoop_20200413013652_04af79c4-9326-4641-894b-3d15d1fbfdfe
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586702550022_0043)

-----
VERTICES    MODE        STATUS      TOTAL   COMPLETED   RUNNING   PENDING   FAILED   KILLED
Map 1 ..... container    SUCCEEDED    10        10           0         0         0         0
Reducer 2 ... container    SUCCEEDED     1         1           0         0         0         0
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 20.02 s
-----
OK
product_category    total_users
Digital_Music_Purchase 109
Time taken: 20.559 seconds
hive>
```

Number of vine users in Digital_Video_Games and Video_Games

select product_category,count(distinct(customer_id)) as total_users from
amazon_review.amazon_reviews_v2 where product_category in
('Video_Games','Digital_Video_Games') and vine='Y' group by product_category;

The screenshot shows a MobaXterm terminal window with a Hive query execution. The query is: `select product_category,count(distinct(customer_id)) as total_users from amazon_review.amazon_reviews_v2 where product_category in ('Video_Games','Digital_Video_Games') and vine='Y' group by product_category;` The output shows two rows: Video_Games with 1760 total users and Digital_Video_Games with 1760 total users. The execution time is 30.817 seconds.

```
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586702550022_0044)

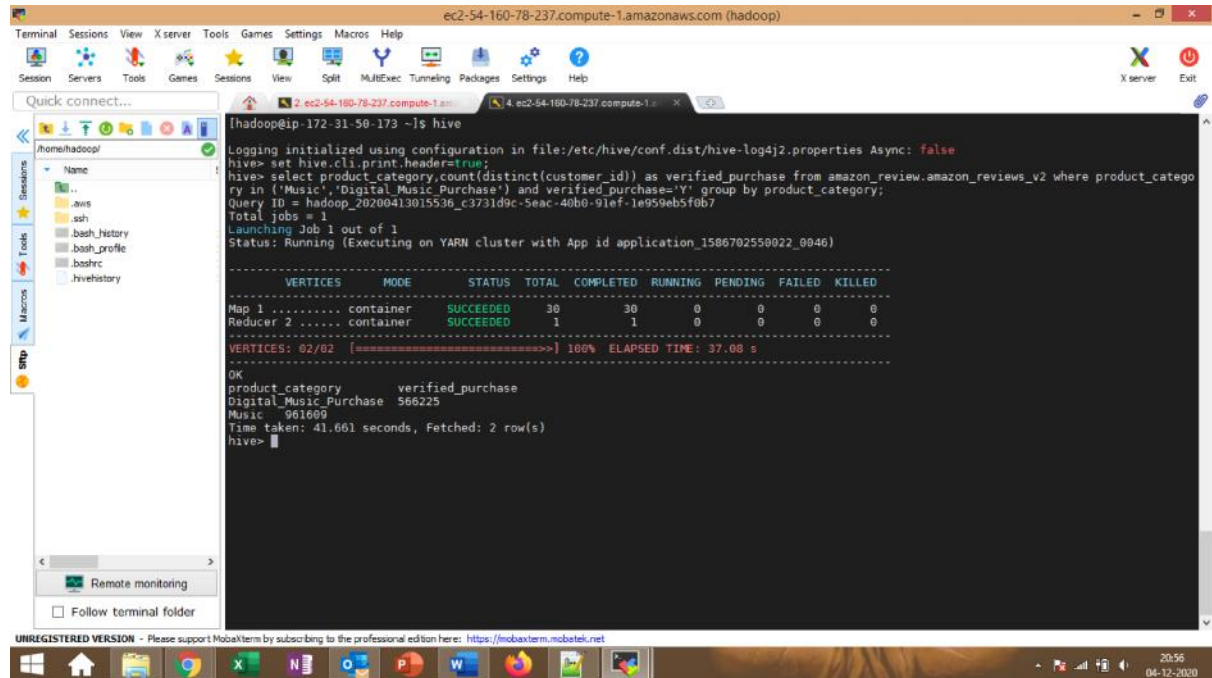
-----
VERTICES    MODE        STATUS      TOTAL   COMPLETED   RUNNING   PENDING   FAILED   KILLED
Map 1 ..... container    SUCCEEDED    22        22           0         0         0         0
Reducer 2 ... container    SUCCEEDED     1         1           0         0         0         0
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 26.44 s
-----
OK
Video_Games         1760
Time taken: 30.347 seconds, Fetched: 1 row(s)
hive> [hadoop@ip-172-31-50-173 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> set hive.cli.print.header=true;
hive> select product_category,count(distinct(customer_id)) as total_users from amazon_review.amazon_reviews_v2 where product_category in
('Video_Games','Digital_Video_Games') and vine='Y' group by product_category;
Query ID = hadoop_20200413015037_50233f11-0e8f-47e3-b7aa-57b724d6e2e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586702550022_0045)

-----
VERTICES    MODE        STATUS      TOTAL   COMPLETED   RUNNING   PENDING   FAILED   KILLED
Map 1 ..... container    SUCCEEDED    22        22           0         0         0         0
Reducer 2 ... container    SUCCEEDED     1         1           0         0         0         0
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 26.50 s
-----
OK
product_category    total_users
Video_Games         1760
Digital_Video_Games 1760
Time taken: 30.817 seconds, Fetched: 1 row(s)
hive>
```

Number of verified_reviews in Music and Digital_Music_Purchase

select product_category,count(distinct(customer_id)) as verified_purchase from
amazon_review.amazon_reviews_v2 where product_category in ('Music','Digital_Music_Purchase')
and verified_purchase='Y' group by product_category;



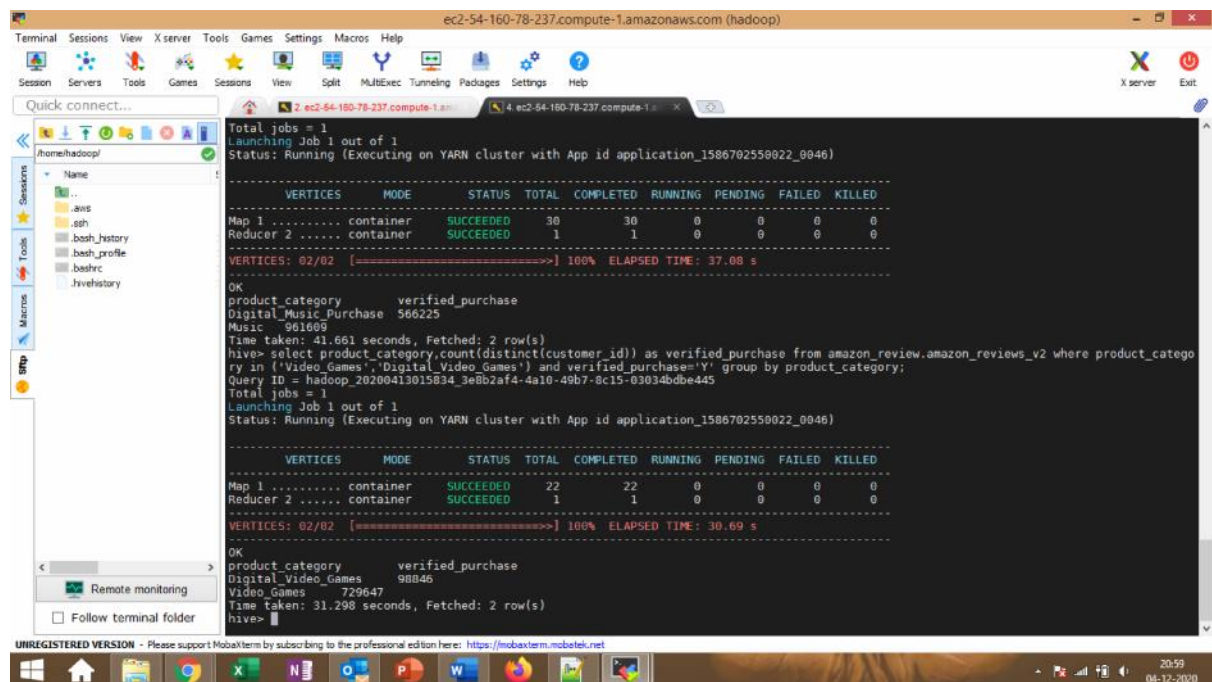
The screenshot shows a MobaXterm terminal window with a Hive query being executed. The query filters for 'Music' and 'Digital_Music_Purchase' categories where 'verified_purchase' is 'Y'. The execution status is 'SUCCEEDED' for both Map and Reduce tasks. The output shows 566225 verified purchases for Digital_Music_Purchase and 961609 for Music.

```
[hadoop@ip-172-31-50-173 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> set hive.cli.print.header=true;
hive> select product_category,count(distinct(customer_id)) as verified_purchase from amazon_review.amazon_reviews_v2 where product_category in ('Music','Digital_Music_Purchase') and verified_purchase='Y' group by product_category;
Query ID = hadoop_20200413015536_c3731d9c-Seac-40b0-91ef-1e959eb5f0b7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586702550022_0046)

-----
VERTICES    MODE    STATUS    TOTAL    COMPLETED    RUNNING    PENDING    FAILED    KILLED
-----
Map 1 ..... container    SUCCEEDED    30        30            0            0            0            0
Reducer 2 ..... container    SUCCEEDED    1          1            0            0            0            0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 37.08 s
-----
OK
product_category    verified_purchase
Digital_Music_Purchase    566225
Music    961609
Time taken: 41.661 seconds, Fetched: 2 row(s)
hive>
```

Number of verified_reviews in Digital_Video_Games and Video_Games

select product_category,count(distinct(customer_id)) as verified_purchase from
amazon_review.amazon_reviews_v2 where product_category in
('Video_Games','Digital_Video_Games') and verified_purchase='Y' group by product_category;



The screenshot shows a MobaXterm terminal window with a Hive query being executed. The query filters for 'Video_Games' and 'Digital_Video_Games' categories where 'verified_purchase' is 'Y'. The execution status is 'SUCCEEDED' for both Map and Reduce tasks. The output shows 98846 verified purchases for Digital_Video_Games and 729647 for Video_Games.

```
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586702550022_0046)

-----
VERTICES    MODE    STATUS    TOTAL    COMPLETED    RUNNING    PENDING    FAILED    KILLED
-----
Map 1 ..... container    SUCCEEDED    30        30            0            0            0            0
Reducer 2 ..... container    SUCCEEDED    1          1            0            0            0            0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 37.08 s
-----
OK
product_category    verified_purchase
Digital_Music_Purchase    566225
Music    961609
Time taken: 41.661 seconds, Fetched: 2 row(s)
hive> select product_category,count(distinct(customer_id)) as verified_purchase from amazon_review.amazon_reviews_v2 where product_category in ('Video_Games','Digital_Video_Games') and verified_purchase='Y' group by product_category;
Query ID = hadoop_20200413015834_3e8b2af4-4a10-49b7-8c15-03034bde445
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586702550022_0046)

-----
VERTICES    MODE    STATUS    TOTAL    COMPLETED    RUNNING    PENDING    FAILED    KILLED
-----
Map 1 ..... container    SUCCEEDED    22        22            0            0            0            0
Reducer 2 ..... container    SUCCEEDED    1          1            0            0            0            0
-----
VERTICES: 02/02 [=====] 100% ELAPSED TIME: 30.69 s
-----
OK
product_category    verified_purchase
Digital_Video_Games    98846
Video_Games    729647
Time taken: 31.298 seconds, Fetched: 2 row(s)
hive>
```

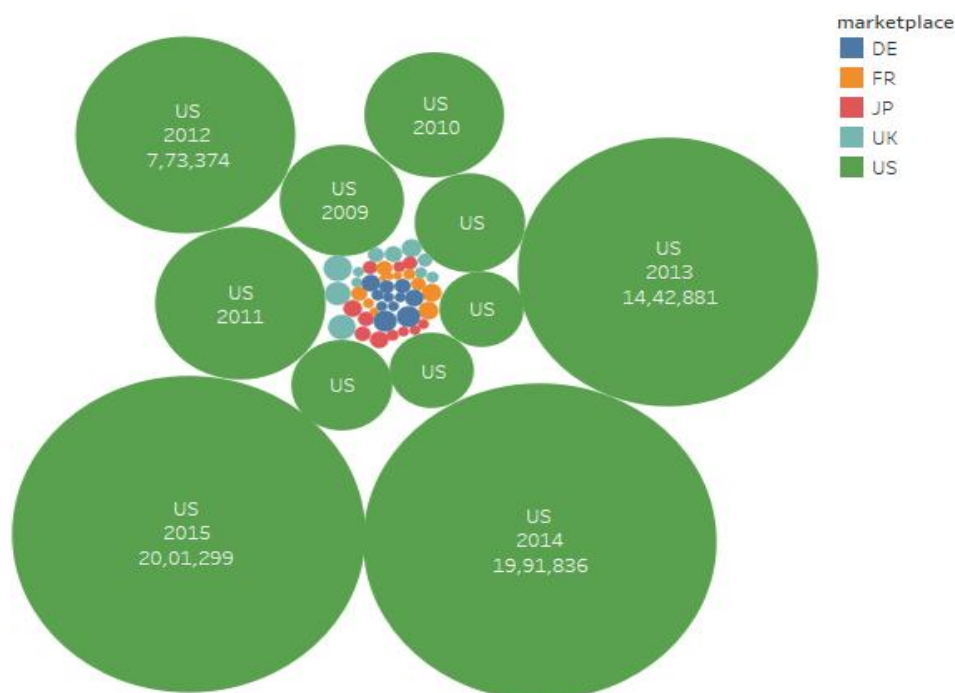
Conclusion

The finding provided about the dataset is with respect to the marketplace parameter. The marketplace gives detail about the customer behaviour based on their region. The following visualization shows the number of product types based on the marketplace. If you see the result, US has the greatest number of product types and the number of products sold in US is higher than any other region. It can be due to two factors which are the number of customers who shop online and the frequency of purchasing a product.

US is the largest consumer of consumer products in the world and it also the lowest to use recycled products. Consider European market for example, the Europeans recycle and reuse many products and thus the frequency of purchasing a product is less when compared to the US. And the number of people who shop online is larger in US when compared to other parts of the world.

Because the convenience in shopping online in US is significant when compared to Europe and Asia. This might be because of the distance of shops from the place of residence in Europe and Asia is less when compared to the US and this might be an important factor on the customer's choice of shopping. This can be significantly seen when comparing Japan and the US. As the proximity of shops in Japan is nearer when compared to the US, more people purchase in physical stores rather than online stores.

Sheet 1



Marketplace, year and sum of total_type_of_products. Colour shows details about marketplace. Size shows sum of total_type_of_products. The marks are labelled by marketplace, year and sum of total_type_of_products.

References

<https://minimaxir.com/2014/06/reviewing-reviews/>
<https://dzone.com/articles/100-shades-of-grey>