

BIG DATA ANALYTICS

Project Part 1 -Analysis of Amazon Review Dataset**

Report by Vallabhan Kudlu Ramakrishnan

The University of Texas at Dallas**

April 2020

When purchasing a product only variable we look into are rating and reviews. Improving it is important to maintain current customers and will draw new customers. This project aims to identify relationships between different aspects of a review including review length, overall review star rating, review helpfulness, reviewer frequency, product price, etc.

This project uses advanced tools to research the different statistical patterns related to these aspects of the analysis. The Relation were analyzed using Hive,AWS Athena and Power BI

The workflow of this project involves four stages:

- Data collection and preparation
- Analysis using Hive
- Conclusions derived from analysis
- Visualizations

The Amazon Test Dataset is available online free of charge on the SNAP datasets website. The list contains customer reviews of items that are offered on the Amazon website. There are two pieces of this dataset. The first component includes about 35 million reviews spanning 13 years of data from product feedback. It includes user-related information that offers analysis, time-related information, and analysis features such as review length, description, etc.

In order to understand the structure and schema of the dataset, let us look at a sample Amazon Review.



- Summary : The title of the review
- Review text : The actual content of the review.
- Rating : User rating of the product on a scale of 1 to 5.
- Helpfulness : The number of people who found the review useful.

Analysis

```
create database amazon_review;
drop table amazon_review.amazon_reviews_parquet;

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;

set hive.cli.print.header=true;

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

**Final view**
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(*) AS 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

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');

Msck repair table amazon_review.amazon_reviews_v2;

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';

insert overwrite table amazon_review.amazon_reviews_v2
partition(product_category='Automotive')
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='Automotive';

insert overwrite table amazon_review.amazon_reviews_v2
partition(product_category='Music')
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='Music';

insert overwrite table amazon_review.amazon_reviews_v2
partition(product_category='Digital_Music_Purchase')
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='Digital_Music_Purchase';

insert overwrite table amazon_review.amazon_reviews_v2
partition(product_category='Sports')
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='Sports';

insert overwrite table amazon_review.amazon_reviews_v2
partition(product_category='Toys')
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='Toys';

insert overwrite table amazon_review.amazon_reviews_v2
partition(product_category='Digital_Video_Games')
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='Digital_Video_Games';

insert overwrite table amazon_review.amazon_reviews_v2
partition(product_category='Video_Games')
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='Video_Games';

```

1)How many unique customer from different marketplace?

US seem to have majority of business with new customers

```

select marketplace,count(distinct(customer_id))as users from amazon_reviews_v2
group by marketplace limit 10;

```

```

hive> select marketplace,count(distinct(customer_id)) as users from amazon_reviews_v2 group by marketplace limit 10;
Query ID = hadoop_20200412210656_ae05be0c-078a-4e5c-a589-1e92f6865c29
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0010)

-----  

 VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED    26     26      0       0       0       0       0  

Reducer 2 ..... container SUCCEEDED     2      2      0       0       0       0       0  

-----  

VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 77.92 s  

-----  

OK  

FR      45790  

JP      41114  

UK      227365  

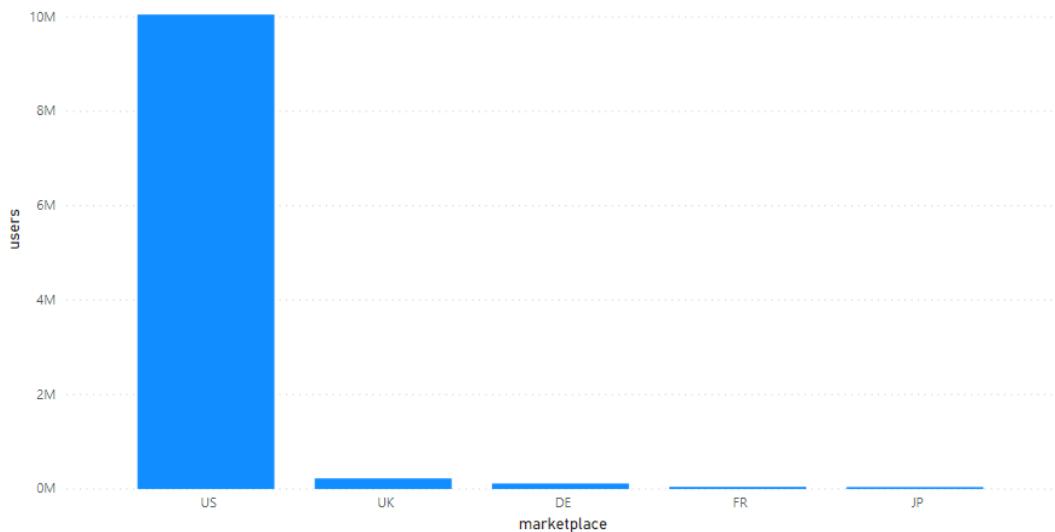
US      11407866  

DE      117430  

Time taken: 79.978 seconds, Fetched: 5 row(s)

```

users by marketplace



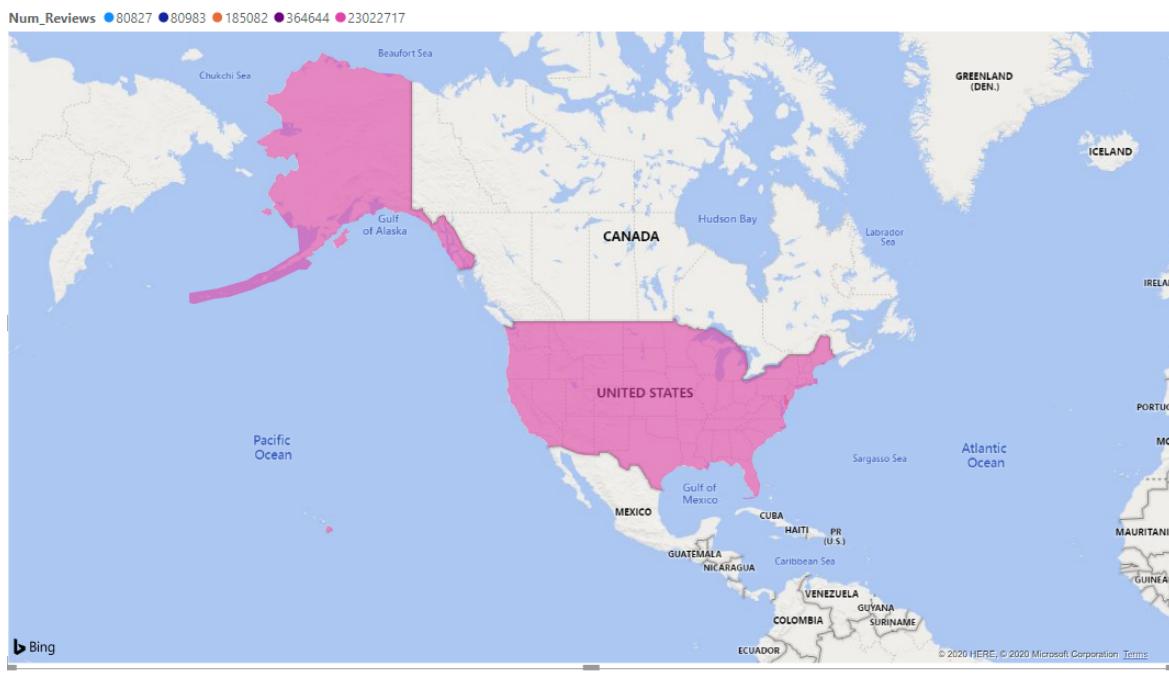
2) Which marketplace is reviewing the product most?

US seem to review lot of products

```
select marketplace, count(review_id) Num_Reviews from amazon_reviews_v2 group by marketplace order by Num_Reviews desc limit 10;
```

Marketplace	Num_Reviews
US	28125769
UK	368707
DE	187164
FR	82134
JP	81752
Others	80827, 80983, 185082, 364644, 23022717

Marketplace and Num_Reviews



3) Trend of star rating

UK is appears to give 4.5 average star rating

```
select marketplace, avg(star_rating) as avg_stars from amazon_reviews_v2 group by marketplace limit 10;
```

```

hive> select marketplace,avg(star_rating) as review_stars from amazon_reviews_v2 group by marketplace limit 10;
Query ID = hadoop_20200412212615_21ecf006-1a56-476f-ba54-3c7b5e0bfa72
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0012)

-----  

 VERTICES    MODE     STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED 24      24      0       0       0       0  

Reducer 2 .... container SUCCEEDED 2       2       0       0       0       0  

-----  

VERTICES: 02/02 [=====>>>] 100% ELAPSED TIME: 32.06 s  

-----  

OK  

DE      4.4272883674210854  

UK      4.535368192087484  

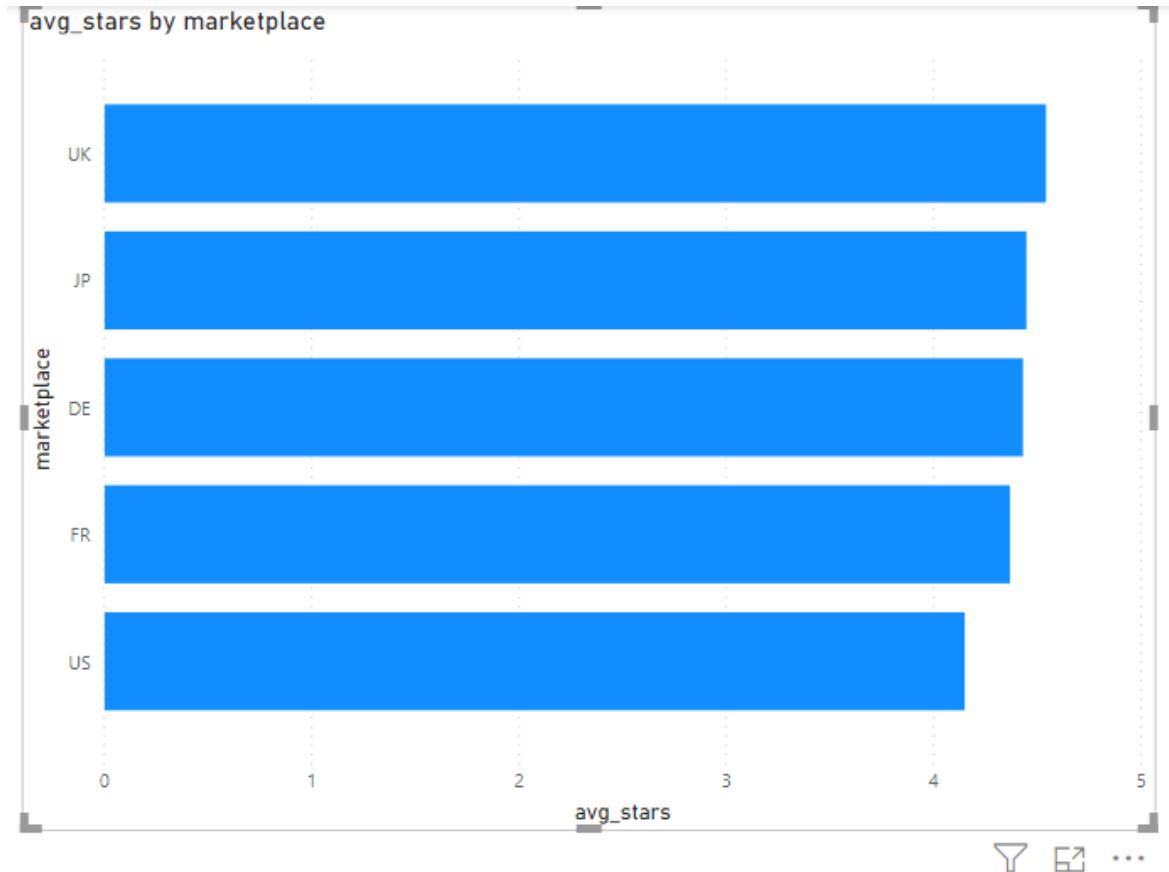
FR      4.363211337570312  

JP      4.4378608474410415  

US      4.157076309629981  

Time taken: 36.072 seconds, Fetched: 5 row(s)

```



4) what is the average length of reviews from different marketplace?

US reviewers likely to write a long review body when compared to other marketplace

```

select marketplace,avg(length(review_body)) as avg_review_length from
amazon_reviews_v2 group by marketplace order by avg_review_length desc limit 10;

```

```

hive> select marketplace,avg(length(review_body))avg_review_length from amazon_reviews_v2 group by marketplace order by avg_review_length desc limit 10;
Query ID = hadoop_20200412213525_049794fc-db09-496e-abaa-0e8ff5812f422
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_1586723064686_0013)

-----  

 VERTICES    MODE     STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED 24      24      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: 59.95 s  

-----  

OK  

DE      768.0271312859311  

FR      625.5894148586457  

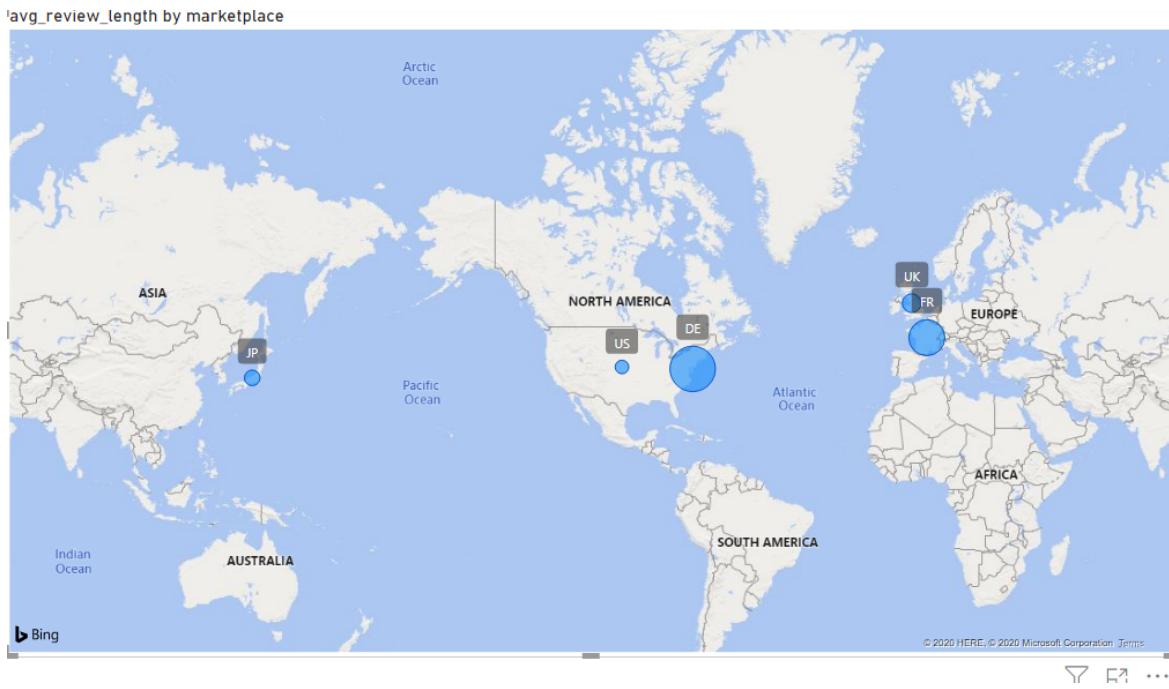
UK      374.37413108049306  

JP      324.7476625051375  

US      296.7596532062387  

Time taken: 67.025 seconds, Fetched: 5 row(s)

```



5) Has verification purchases changed over time for reviews?

Verification purchases seems to be increased from time to time

```
select year,sum(case when verified_purchase = 'Y' then 1 else 0 end)
verified_reviews,
      (sum(case when verified_purchase = 'Y' then 1 else 0
end)*100.00)/count(*) percent_verified
from amazon_reviews_v2 group by year order by year;
```

```
hive> select year,sum(case when verified_purchase = 'Y' then 1 else 0 end) verified_reviews,
    >           (sum(case when verified_purchase = 'Y' then 1 else 0 end)*100.00)/count(*) percent_verified
    >       from amazon_reviews_v2 group by year order by year;
Query ID = hadoop_20200412213826_6e147585-53f5-4bde-84c3-a81cdff22d02
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0013)

-----
 VERTICES     MODE     STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
-----  

Map 1 ..... container SUCCEEDED 24      24      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: 33.44 s
-----
OK
2005    26792   8.006024240395879
2006    38434   13.381613077311422
2007    105235  26.088368507036807
2008    140802   30.636570145434607
2009    259658   44.207114121154236
2010    515739   67.696764530995475
2011    904989   73.771205403541550
2012    1815921  80.898664668465879
2013    4886874  89.174303489128100
2014    7242922  85.052409429305456
2015    7913171  92.615791230516928
Time taken: 34.028 seconds, Fetched: 11 row(s)
```



6) How do Ratings Vary with verified Purchase?

As seen from the graph verified_purchase are on average rated 0.05 higher

```
select verified_purchase,
       round(avg(star_rating),2) avg_rating,
       count(*) count
  from amazon_reviews_v2
 group by verified_purchase;
```

```
hive> select verified_purchase,
       round(avg(star_rating),2) avg_rating,
       count(*) count
  from amazon_reviews_v2
 group by verified_purchase;
Query ID: hadoop_20290413023936_4bcd5ab0-8291-469f-b7b3-02e8e4edd0d4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586742509915_0009)

-----  

      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: 32.64 s  

-----  

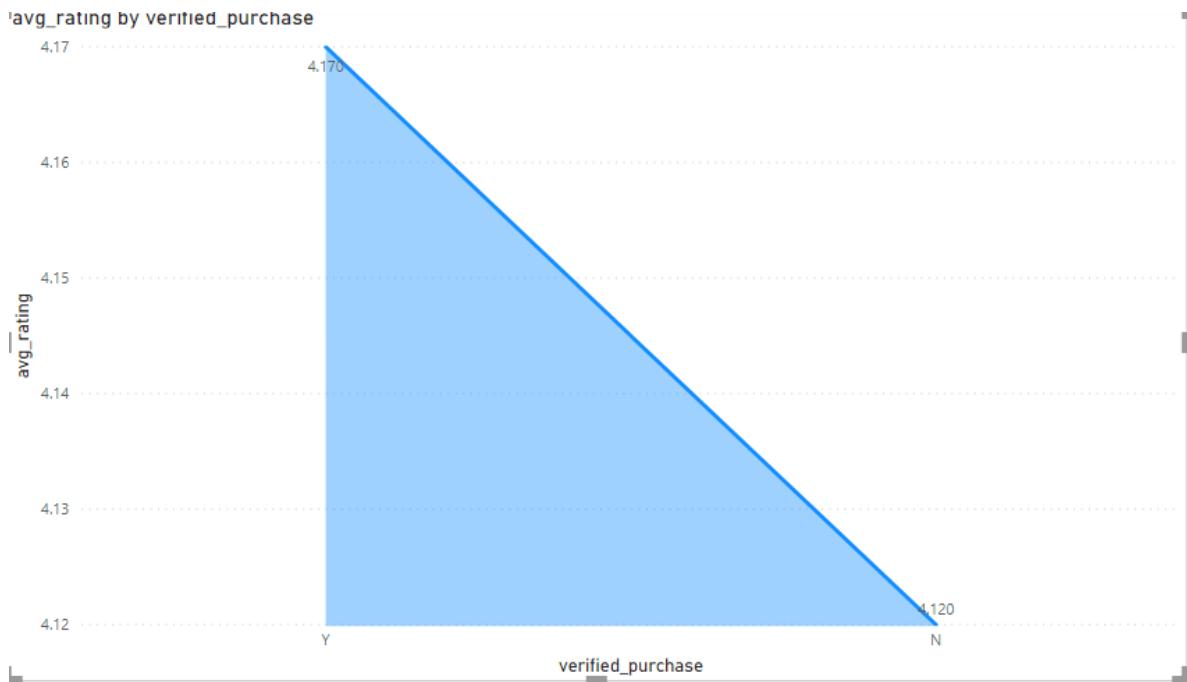
OK  

verified_purchase      avg_rating      count  

N          4.08        4994989  

Y          4.18        23850537  

Time taken: 33.095 seconds, Fetched: 2 row(s)
hive>
```



7) How do Ratings vary with vine membership?

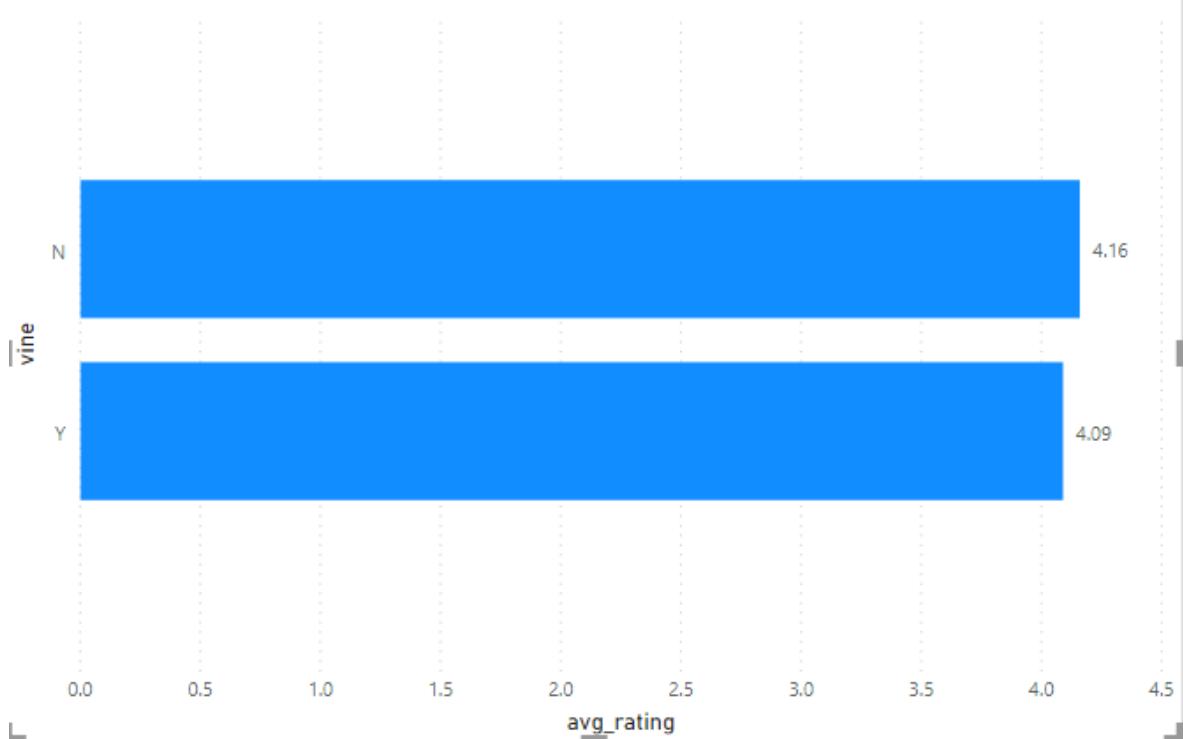
Vine Reviews are on average 0.16 lower than non-Vine

```
select vine,
       round(avg(star_rating),2) avg_rating,
       count(*) count
  from amazon_reviews_v2
 group by vine;
```

```
ec2-35-175-209-35.compute-1.amazonaws.com (hadoop) (1)
Terminal Sessions View Xserver Tools Games Sessions Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
Quick connect...
<< /home/hadoop/ >> X server Exit
Sessions
  Name
  aws
  ssh
  bash_profile
  bashrc
Tools Macros
Skip
  Sessions
  Tools
  Games
  Sessions
  View
  Split
  MultiExec
  Tunneling
  Packages
  Settings
  Help
  Quick connect...
<< /home/hadoop/ >>
> group by verified_purchase;
Query ID = hadoop_20200413023936_4bcd5ab0-8291-469f-b7b3-02e8e4edd6d4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586742509915_0009)
-----
  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: 32.64 s
-----
OK
verified_purchase    avg_rating    count
N        4.08    4994998
Y        4.18    22850537
Time taken: 33.095 seconds, Fetched: 2 row(s)
hive> select vine,
       >       round(avg(star_rating),2) avg_rating,
       >       count(*) count
       >   from amazon_reviews_v2
       >   group by vine;
Query ID = hadoop_20200413024055_47612059-eb5e-4137-b90b-3bbb117dd1eb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586742509915_0009)
-----
  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: 32.73 s
-----
OK
vine    avg_rating    count
N        4.17    28764123
Y        4.1     81483
Time taken: 33.196 seconds, Fetched: 2 row(s)
hive> 
```

UNREGISTERED VERSION - Please support Mobaxterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

avg_rating by vine



8) How do Ratings vary with Marketplace?

UK reviewers are the most positive ,with average rating of 0.3 higher than the US marketplace

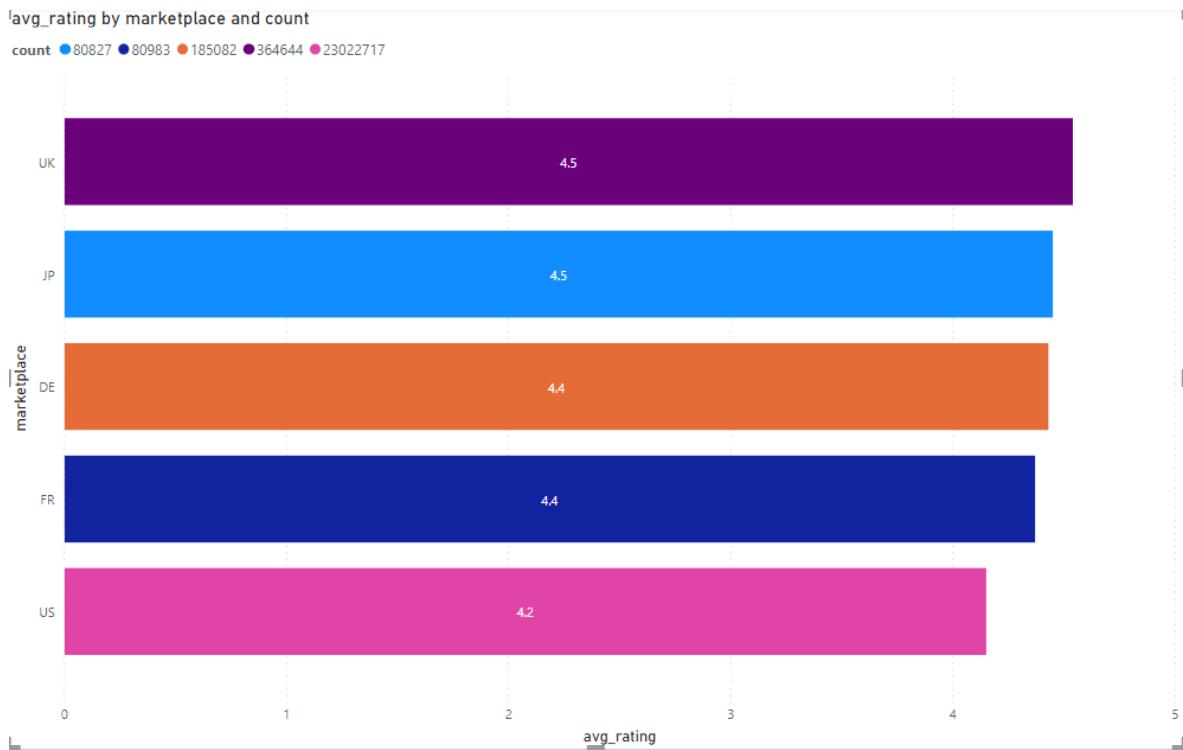
```
select marketplace,
       round(avg(star_rating),2) avg_rating,
       count(*) count
  from amazon_reviews_v2
 group by marketplace;
```

A screenshot of a terminal window titled "ec2-35-175-209-35.compute-1.amazonaws.com (hadoop) (1)". The window shows the execution of a Hive query to calculate average ratings by marketplace. The output includes statistics for vertices, mode, and status, followed by the results of the SELECT statement.

```

Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586742509915_0009)
-----
      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: 32.73 s
-----
OK
vine  avg_rating  count
N    4.17        28764123
Y    4.1          8149
Time taken: 33.538 seconds, Fetched: 2 row(s)
hive> select marketplace,
       >       round(avg(star_rating),2) avg_rating,
       >       count(*) count
       >   from amazon_reviews_v2
       >   group by marketplace;
Query ID = hadoop_20280413024244_d7f398a0-28a3-4fd7-ae07-721b6805e58c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586742509915_0009)
-----
      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: 33.05 s
-----
OK
marketplace  avg_rating  count
DE           4.43        187164
UK           4.54        368797
FR           4.36        82134
JP           4.44        81752
US           4.16        28125769
Time taken: 33.538 seconds, Fetched: 5 row(s)
hive> 
```

UNREGISTERED VERSION - Please support Mobaxterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

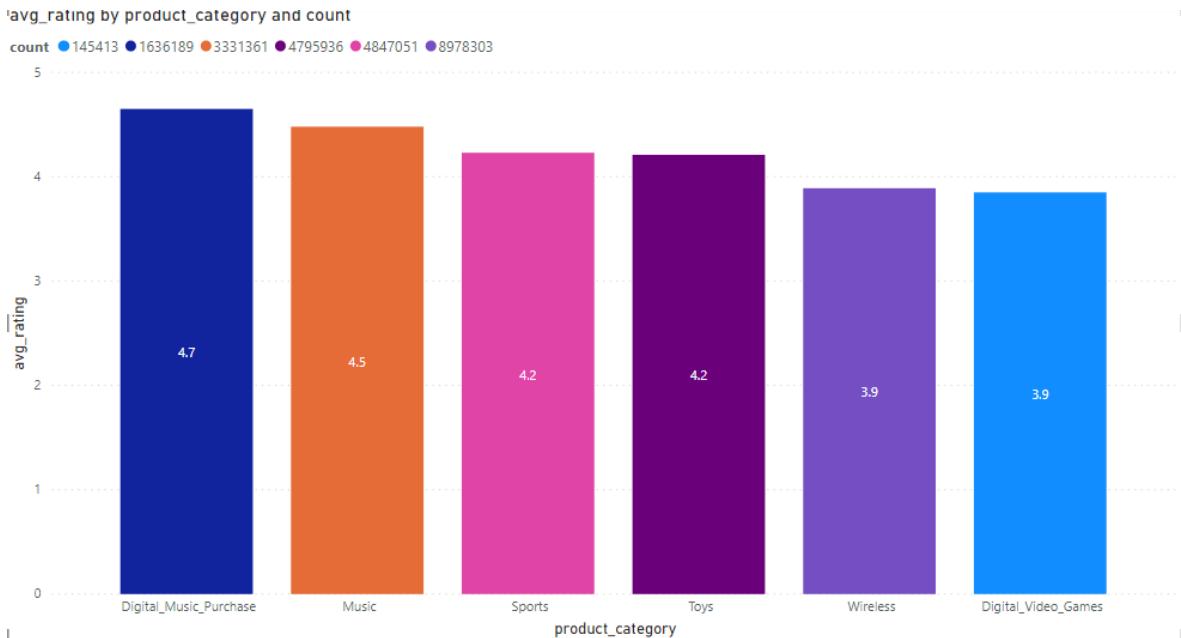


9) How do Ratings vary with product_category?

Digit Music purchase has highest average rating when compared to other product category

```
select product_category,
       round(avg(star_rating),2) avg_rating,
       count(*) count
  from amazon_reviews_v2
 group by product_category;
```

```
hive> select product_category,
       round(avg(star_rating),2) avg_rating,
       count(*) count
  from amazon_reviews_v2
 group by product_category;
OK
product_category      avg_rating      count
Automotive           4.25      3514995
Digital_Video_Games  3.85      145413
Music                4.48      3331361
Toys                 4.21      4795936
Video_Games          4.07      1596278
Wireless              3.89      8978303
Digital_Music_Purchase 4.65      1030189
Sports                4.23      4847051
Time taken: 32.651 seconds, Fetched: 8 row(s)
hive> 
```



10) Does length of review influence star rating?

As analyzed length of review has little influence on star rating

```
select length(review_body) length_of_review,
       round(avg(star_rating),2) avg_rating, count(*) count
  from amazon_reviews_v2
 group by length(review_body) order by length_of_review desc limit 10;
```

ec2-35-175-209-35.compute-1.amazonaws.com (hadoop) 1

Terminal Sessions View X server Tools Games Settings Macros Help

Session Servers Tools Games Sessions View Split MultiExec Tunneling Padangoes Settings Help

Quick connect...

File Edit View Insert X server Exit

Session Servers Tools Games Sessions View Split MultiExec Tunneling Padangoes Settings Help

2. ec2-35-175-209-35.compute-1.amazonaws.com (hadoop)

product_category avg_rating count

product_category	avg_rating	count
Automotive	4.25	3514995
Digital_Video_Games	3.85	145413
Music	4.48	3331361
Toys	4.21	479936
Videogames	4.87	1596278
Wireless	3.89	8978303
Digital_Music_Purchase	4.65	1636189
Sports	4.23	4847051

Time taken: 32.651 seconds, Fetched: 8 row(s)

hive: select length(review_body) length_of_review,
> round(avg(star_rating),2) avg_rating,count(*) count
> from amazon_reviews_v2
> group by length(review_body) order by length_of_review desc limit 10;

Query ID = mapred_202006130824710_c07fee8e-c099-4bfff-a1fd-b7bb4c74e9a9

Total jobs = 1

Launching Job 1 out of 1

Status: Running (Executing on YARN cluster with App id application_1586742509915_0000)

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	39	39	0	0	0	0
Reducer 2	container	SUCCEEDED	2	2	0	0	0	0
Reducer 3	container	SMPLETED	1	1	0	0	0	0

VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 61.78 s

OK

length_of_review	avg_rating	count
57815	5.0	1
51379	5.0	1
50789	5.0	1
50401	1.0	1
49849	5.0	1
49835	5.0	1
49792	4.0	1
49721	2.0	1
49454	5.0	1
49371	1.0	1

Time taken: 62.253 seconds, Fetched: 10 rows(s)

hive: |

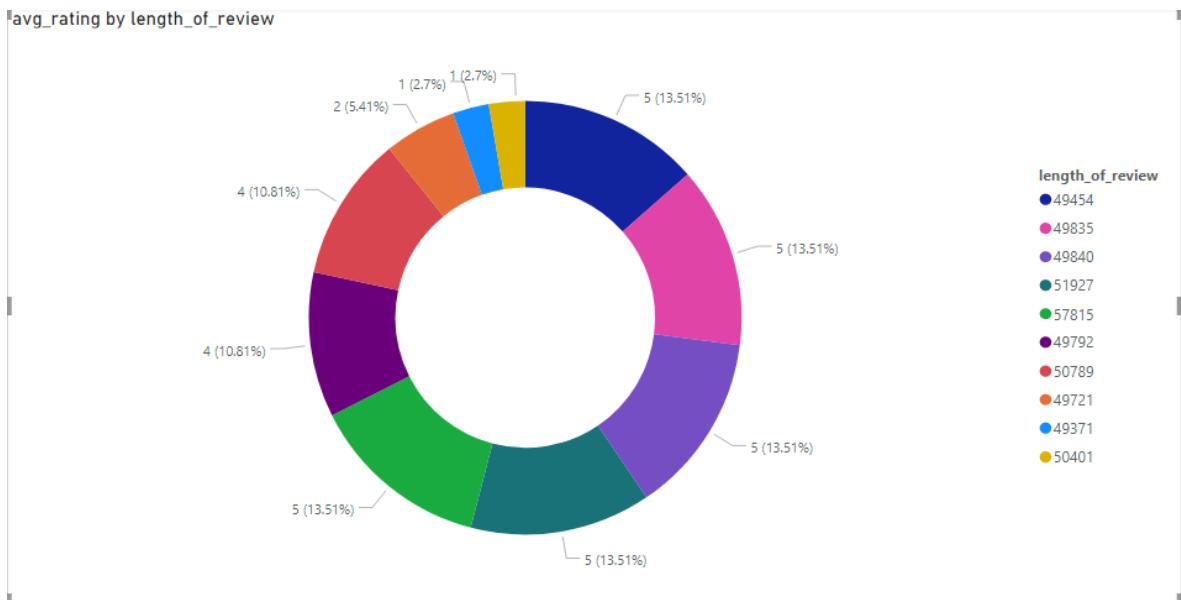
Remote monitoring

Follow terminal folder

UNREGISTERED VERSION - Please support MobateX by subscribing to the professional edition here: <https://mobatext.mobatek.net>

Type here to search

949 PM 4/12/2020



11) Looking at the Table we can easily check how helpful the average review is ,how many Vine Reviews they have and how many of their purchases are verified

```
select customer_id,
count(*) count,
count(distinct product_category) category,
round(avg(star_rating),2) avg_rating,
round(avg(helpful_votes),2) avg_help,
sum(case when star_rating = 1 then 1 else 0 end) one,
sum(case when star_rating = 2 then 1 else 0 end) two,
sum(case when star_rating = 3 then 1 else 0 end) three,
sum(case when star_rating = 4 then 1 else 0 end) four,
sum(case when star_rating = 5 then 1 else 0 end) five,
sum(case when vine = 'Y' then 1 else 0 end) vine_reviews,
sum(case when verified_purchase = 'Y' then 1 else 0 end) verified_purchases
from amazon_reviews_v2
group by customer_id
order by count desc;
```

```
ec2-35-175-209-35.compute-1.amazonaws.com (hadoop) (1)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunnelling Packages Settings Help
Quick connect...
File Edit View Insert Cell Session Help
hive> select customer_id,
> count(*) count,
> count(distinct product_category) category,
> round(avg(star_rating)/2) avg_rating,
> round(avg(helpful_votes),2) avg_helpful_votes,
> sum(case when star_rating = 1 then 1 else 0 end) one,
> sum(case when star_rating = 2 then 1 else 0 end) two,
> sum(case when star_rating = 3 then 1 else 0 end) three,
> sum(case when star_rating = 4 then 1 else 0 end) four,
> sum(case when star_rating = 5 then 1 else 0 end) five,
> sum(case when vine = 'Y' then 1 else 0 end) vine_reviews,
> sum(case when verified_purchase = 'Y' then 1 else 0 end) verified_purchases
from amazon_reviews
group by customer_id
order by count desc limit 10;
Query ID: hadoop_20280412215928_804a991e-0467-4c86-95a9-1ef3b9631826
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0014)

-----  

VERTICES      NODE      STATUS    TOTAL    COMPLETED    RUNNING    PENDING    FAILED    KILLED  

-----  

Map 1 ... container SUCCEEDED   24       24       0         0         0         0         0  

Reducer 2 ... container SUCCEEDED   2        2       0         0         0         0         0  

Reducer 3 ... container SUCCEEDED   1        1       0         0         0         0         0  

-----  

VERTICES: 03/03 [=====>>>>>] 100% ELAPSED TIME: 172.32 s  

-----  

OK  

38210533 5425 4 4.13 2.54 81 75 863 2422 1964 0 139  

31181997 5310 2 4.44 4.79 3 160 1295 1480 2372 0 958  

18116317 4137 5 4.1 1.83 52 69 560 2112 1315 4 528  

23267387 3616 7 3.9 2.38 7 96 876 1926 711 14 133  

7089939 3538 7 5.0 0.02 0 0 0 3538 0 14  

50736958 3218 5 4.28 1.16 124 198 260 830 1858 0 1  

14539589 2995 7 4.92 0.26 40 6 18 20 2911 0 131  

42418272 2981 8 3.37 4.18 84 465 901 1328 203 5 200  

52496677 2541 7 4.88 5.49 0 0 19 262 2260 2 61  

51381678 2494 2 3.88 0.26 0 49 759 1126 560 0 1  

Time taken: 176.793 seconds, Fetched: 10 row(s)  

hive>
```

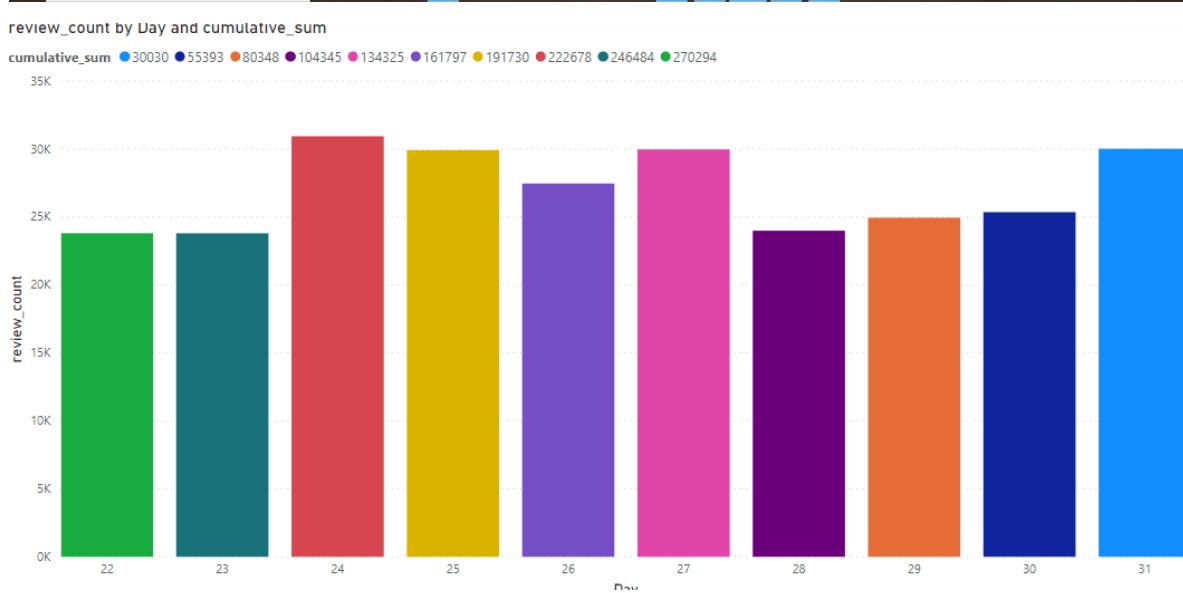
12)How many reviews per day?

We can see that there is variation in number of reviews that are given within 30 day period as seen from the visualization the cumulative sum hovers each day

```
select date(review_date)day ,count(review_id) review_count,
sum(count(review_id)) over (order by date(review_date) desc) as cumulative_sum
from amazon_reviews_v2 group by date(review_date) limit 10;
```

The terminal window shows the execution of a Hive query to calculate daily review counts and their cumulative sum for the last 10 days. The results are displayed in a table format.

Day	Review Count	Cumulative Sum
22	24,648	24,648
23	27,029	51,677
24	30,303	81,980
25	29,842	111,822
26	24,088	135,910
27	29,768	165,678
28	24,648	190,326
29	25,393	215,719
30	27,029	242,748
31	24,648	267,396



13)How many unique customer per day?

There seems to be increase in the new customers per day

```
select review_date,count(distinct(customer_id)) customer_count,
sum(count(distinct(customer_id))) over (order by review_date desc) as
cumulative_sum_customer
from amazon_reviews_v2 group by review_date order by review_date limit 10;
```

Sessions View Xserver Tools Games Settings Macros Help

Session Servers Tools Sessions View Split MultiExec Tunneling Packages Settings Help

Quick connect...

/home/hadoop/

```

at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1207)
at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:1222)
at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:233)
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.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:62)
at sun.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 1:24 cannot recognize input near 'date' ',' 'count' in function specification
hive> select date(review_date),
   > count(distinct(customer_id))customers
   > from amazon_reviews_v2 group by date(review_date) order by customers desc limit 5;
Query ID: hivesp_20200412221749_52b09331-cb4d-4781-95d0-66c7d0e0dc59
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_1586723064686_0016)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED 24 24 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: 69.16 s  

-----  

OK  

2015-01-03 39851  

2015-01-05 37019  

2015-01-07 35881  

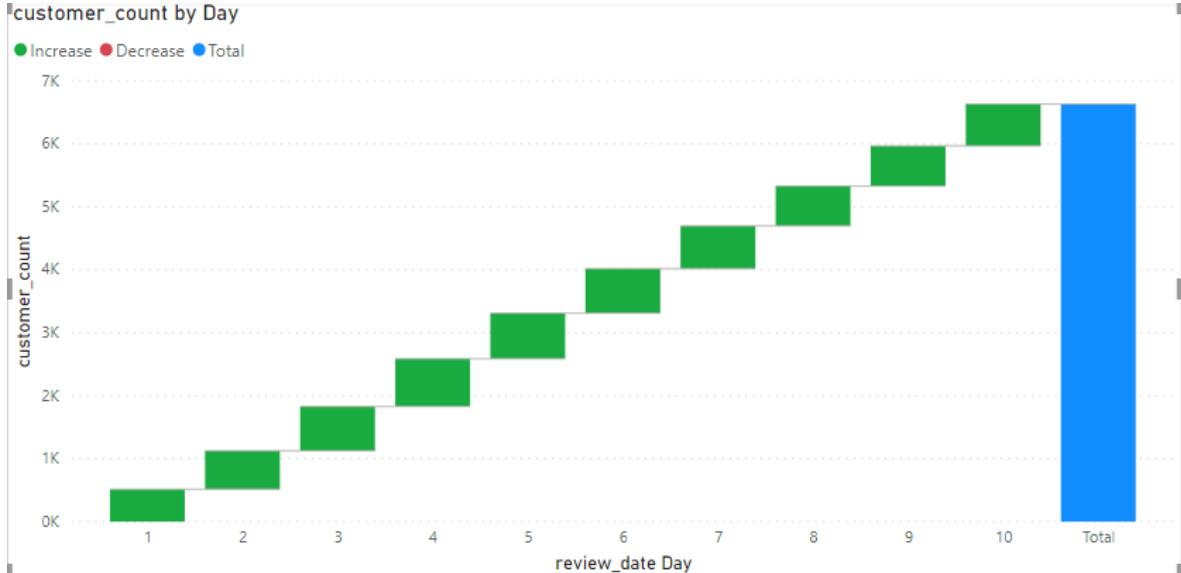
2015-06-03 35366  

2014-12-29 35294  

Time taken: 81.398 seconds, Fetched: 5 row(s)
hive>
```

UNREGISTERED VERSION - Please support MobaTerm by subscribing to the professional edition here: <https://mobaterm.mobatek.net>

Type here to search 5:19 PM 4/12/2020



14)What is the moving average star rating for 2 days?

```

select review_date,star_rating,star_rating_1,Avg(star_rating_1) over
(partition by star_rating order by review_date,star_rating asc rows between 2
preceding and current row) as avg_rating from (
  select review_date,star_rating,count(star_rating) as star_rating_1
  from amazon_reviews_v2 group by review_date,star_rating order by
  review_date,star_rating asc limit 30)s;
```

```

Time taken: 0.454 seconds
hive> select review_date,star_rating,sum(star_rating_1) over (order by review_date) as sum_rating_1 from (
    select review_date,star_rating,count(star_rating)
    as star_rating_1 from amazon_reviews_v2 where star_rating=5 group by
review_date,star_rating order by review_date asc limit 10)s
Query ID = hadoop_20260412223850_fe7414a0-a2df-43b3-ab70-ffe4151c1233
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0017)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED 24 24 0 0 0 0  

Reducer 2 ..... container SUCCEEDED 2 2 0 0 0 0  

Reducer 3 ..... container SUCCEEDED 1 1 0 0 0 0  

Reducer 4 ..... container SUCCEEDED 1 1 0 0 0 0  

-----  

VERTICES: 04/04 [=====>] 100% ELAPSED TIME: 32.81 s  

-----  

OK  

2005-01-01 1 76 76.0  

2005-01-02 1 84 80.0  

2005-01-03 1 99 86.33333333333333  

2005-01-04 1 101 94.66666666666667  

2005-01-05 1 105 101.66666666666667  

2005-01-06 1 103 100.0  

2005-01-01 2 24 24.0  

2005-01-02 2 44 34.0  

2005-01-03 2 56 41.33333333333336  

2005-01-04 2 62 54.0  

2005-01-05 2 70 62.666666666666664  

2005-01-06 2 48 60.0  

2005-01-01 3 68 68.0  

2005-01-02 3 72 70.0  

2005-01-03 3 74 71.33333333333333  

2005-01-04 3 78 74.66666666666667  

2005-01-05 3 83 78.33333333333333  

2005-01-06 3 82 81.0  

2005-01-01 4 128 128.0  

2005-01-02 4 161 144.5

```

UNREGISTERED VERSION - Please support MobaTerm by subscribing to the professional edition here: <https://mobaterm.mobatek.net>

15) Number of five star rating per day for time period of 10 days?

As analyzed the trend of five star rating seems to be increasing for the given time interval

```

select review_date,star_rating,sum(star_rating_1) over (order by review_date) as
sum_rating_1 from (
    select review_date,star_rating,count(star_rating)
    as star_rating_1 from amazon_reviews_v2 where star_rating=5 group by
review_date,star_rating order by review_date asc limit 10)s

```

```

Time taken: 0.454 seconds
hive> select review_date,star_rating,sum(star_rating_1) over (order by review_date) as sum_rating_1 from (
    select review_date,star_rating,count(star_rating)
    as star_rating_1 from amazon_reviews_v2 where star_rating=5 group by
review_date,star_rating order by review_date asc limit 10)s
Query ID = hadoop_20260412224230_68d5f85b-669c-45ab-b925-b83bd0e0c8f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0017)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED 24 24 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  

-----  

VERTICES: 04/04 [=====>] 100% ELAPSED TIME: 32.81 s  

-----  

OK  

2005-01-01 5 386  

2005-01-02 5 872  

2005-01-03 5 1376  

2005-01-04 5 1959  

2005-01-05 5 2467  

2005-01-06 5 2981  

2005-01-07 5 3531  

2005-01-08 5 4002  

2005-01-09 5 4464  

2005-01-10 5 4950  

Time taken: 32.768 seconds, Fetched: 10 row(s)
hive>

```

UNREGISTERED VERSION - Please support MobaTerm by subscribing to the professional edition here: <https://mobaterm.mobatek.net>

16) which product id has max reviews?

As we have top 10 product id with maximum reviews we can say that product id B00458F70M has the highest reviews

```

select product_id,max(sum_reviews)as max_reviews from (
  select product_id,count(review_id) as sum_reviews from amazon_reviews_v2
  group by product_id order by sum_reviews desc)
group by product_id order by max_reviews desc limit 10;

```

```

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 2:2 cannot recognize input near '(' 'select' 'distinct' in joinSource
hive> select product_id,max(sum_reviews)as max_reviews from amazon_reviews_v2
>   select distinct(product_id) as product_id,count(review_id) as sum_reviews from amazon_reviews_v2
>   group by product_id order by max_reviews desc limit 10;
>   select product_id,max(sum_reviews)as max_reviews from amazon_reviews_v2
>   group by product_id order by max_reviews desc limit 10;
>   group by product_id order by max_reviews desc limit 10;
Query ID = hadoop_20280412224631_acc2e62a-21e6-429c-951d-e7cfad4517f9
Total jobs: 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0017)

-----  

VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED 24    24    0     0     0     0     0  

Reducer 2 ..... container SUCCEEDED 2    2     0     0     0     0     0  

Reducer 3 ..... container SUCCEEDED 1    1     0     0     0     0     0  

Reducer 4 ..... container SUCCEEDED 1    1     0     0     0     0     0  

Reducer 5 ..... container SUCCEEDED 1    1     0     0     0     0     0  

-----  

VERTICES: 05/05 [=====>>>] 100% ELAPSED TIME: 79.44 s  

-----  

OK  

B00458F7QM 24273  

B009A50W2K 18345  

B009A53H4K 10235  

B0073FE1FO 9944  

B007FHX9OK 9470  

B0042FV2SI 8814  

B009SYZ8OC 8759  

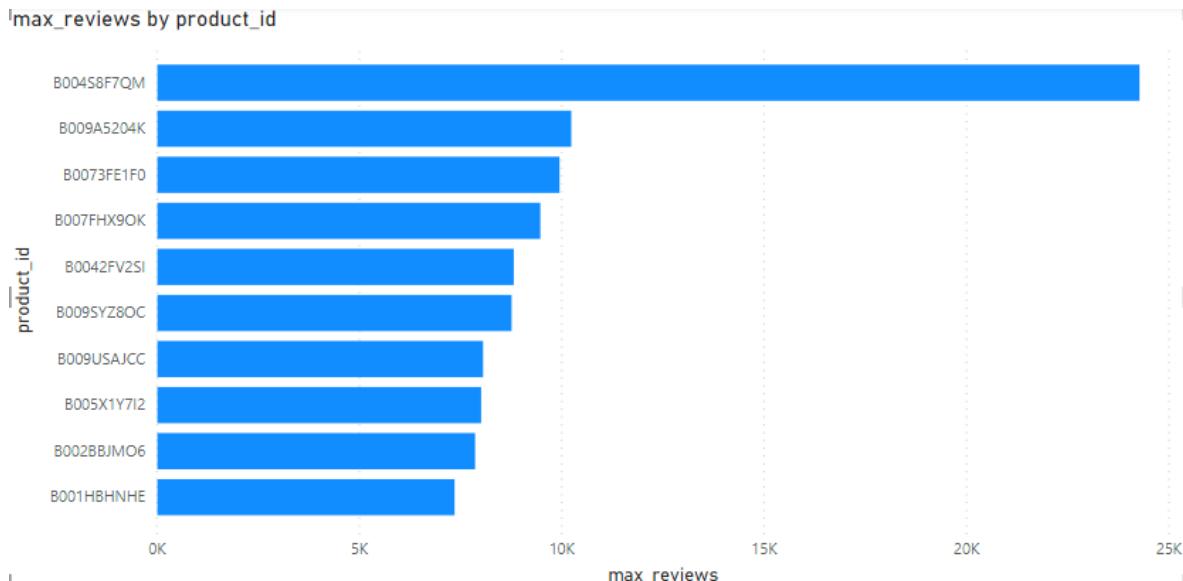
B009USAJCC 8055  

B005X1Y7I2 8007  

B002BBJMO6 7858  

Time taken: 80.017 seconds, Fetched: 10 row(s)
hive>

```



17)what is standard deviation for the star rating?

The standard deviation for one star rating - 10.70

The standard deviation for two star rating - 14.60

The standard deviation for three star rating - 5.36

The standard deviation for four star rating -29.75

```

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_reviews_v2 group by review_date,star_rating order by
review_date,star_rating asc limit 30)s

```

```

FAILED: ParseException line 2:41 cannot recognize input near 'date' ',' 'star_rating' in expression specification
hive> 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_reviews_v2 group by review_date,star_rating order by review_date,star_rating asc limit 30)s
Query ID = hadoop_20200412225031_907d1218-a9dd-41ba-b84e-6fa6db4b626d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_9017)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED 24 24 0 0 0 0  

Reducer 2 ..... container SUCCEEDED 2 2 0 0 0 0  

Reducer 3 ..... container SUCCEEDED 1 1 0 0 0 0  

Reducer 4 ..... container SUCCEEDED 1 1 0 0 0 0  

-----  

VERTICES: 04/04 [=====>>>] 100% ELAPSED TIME: 34.12 s  

OK  

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 5.367088729368206  

2005-01-02 3 72 5.367088729368206  

2005-01-03 3 74 5.367088729368206  

2005-01-04 3 78 5.367088729368206  

2005-01-05 3 83 5.367088729368206  

2005-01-06 3 62 5.367088729368206  

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 4 222 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 5 508 58.16189992142342  

2005-01-06 5 514 58.16189992142342
Time taken: 34.66 seconds, Fetched: 39 row(s)
hive>

```

18) Are there any common users between music and digital music purchase?

From the analysis there are 140797 users are common between two categories

```

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

```

```

hive> select count(distinct(customer_id))common_customers from amazon_reviews_v2 where customer_id in
> (select customer_id from amazon_reviews_v2 where product_category in ('Digital_Music_Purchase'))
> and product_category in ('Music');
Query ID = hadoop_2020041222525_714e438f-5999-40e2-b/bf-cd75dad65926
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_9017)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED 25 25 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: 44.14 s  

OK
140797
Time taken: 44.942 seconds, Fetched: 1 row(s)
hive>

```

19) Are there any common users between Video Games and digital video Games?

From the analysis there 29762 common users

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

The screenshot shows the MobaXterm interface on a Windows desktop. The terminal window displays the execution of a Hive query. The output shows two stages of the job, each with 66 vertices. The first stage has 29 Map tasks and 6 Reducer tasks. The second stage has 20 Map tasks and 5 Reducer tasks. Both stages completed successfully with an elapsed time of approximately 44 seconds.

```
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0017)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED 29 25 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: 66/66 [=====>>>] 100% ELAPSED TIME: 44.14 s  

-----  

OK
140797
Time taken: 44.942 seconds, Fetched: 1 rows(s)
hive> select count(distinct(customer_id)) from amazon_reviews_v2 where customer_id in
> (select customer_id from amazon_reviews_v2 where product_category in ('Digital_Video_Games'))
> and product category in ('Video_Games');
Query ID = hadoop_20280412225425_b4135529-1ccf-433a-b7d2-7f0ddaa1e83
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0017)

-----  

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 1 ..... 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: 35.13 s  

-----  

OK
29762
Time taken: 35.72 seconds, Fetched: 1 row(s)
hive>
```

20)Any common products between Digital Music Purchase and Music?

yes ,there seem to one common product between the above the category

```
select count(distinct(product_id)) from amazon_reviews_v2 where product_id in
(select product_id from amazon_reviews_v2 where product_category in
('Digital_Music_Purchase'))
and product_category in ('Music');
```

The screenshot shows the MobaXterm interface on a Windows desktop. The terminal window displays the execution of a Hive query. The output shows two stages of the job, each with 66 vertices. The first stage has 20 Map tasks and 5 Reducer tasks. The second stage has 25 Map tasks and 6 Reducer tasks. Both stages completed successfully with an elapsed time of approximately 45 seconds.

```
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0017)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 4 ..... container SUCCEEDED 20 20 0 0 0 0  

Map 5 ..... 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: 35.13 s  

-----  

OK
29762
Time taken: 35.72 seconds, Fetched: 1 row(s)
hive> select count(distinct(product_id)) from amazon_reviews_v2 where product_id in
> (select product_id from amazon_reviews_v2 where product_category in ('Digital_Music_Purchase'))
> and product category in ('Music');
Query ID = hadoop_20280412225656_0b3d79fe-0892-402b-bf07-25fe097d37c7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0017)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED 25 25 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 5 ..... container SUCCEEDED 1 1 0 0 0 0  

Reducer 6 ..... container SUCCEEDED 1 1 0 0 0 0  

VERTICES: 66/66 [=====>>>] 100% ELAPSED TIME: 45.19 s  

-----  

OK
1
Time taken: 45.776 seconds, Fetched: 1 row(s)
hive>
```

21)Any common product between Digital Video Games and Video Games?

Yes there are 4 common product between video games and Digital Video Games

```
select count(distinct(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');
```

The screenshot shows the MobaXterm interface with a terminal window open. The terminal displays the execution of a Hive query to find common products between 'Video Games' and 'Digital_Video_Games'. The output shows two sets of job statistics and their completion status.

```
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0017)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED 29 25 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: 66/66 [=====>>>] 100% ELAPSED TIME: 45.19 s  

-----  

OK  

1  

Time taken: 45.776 seconds, Fetched: 1 row(s)  

hive> select count(distinct(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');  

Query ID = hadoop_20200412225850_6285016d-9c11-4ad1-9089-0e1b66330fe  

Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0017)

-----  

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: 29.31 s  

-----  

OK  

4  

Time taken: 29.887 seconds, Fetched: 1 row(s)  

hive>
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

22)Is there correlation between average rating and length of review?

It seems to negatively correlated

```
select corr(review_length,avg_rating) as coeff from
(select length(review_headline) as review_length,avg(star_rating)as avg_rating
from amazon_reviews_v2 group by length(review_headline))s;
```

The screenshot shows the MobaXterm interface with a terminal window open. The terminal displays the execution of a Hive query to calculate the correlation coefficient between 'review_length' and 'avg_rating'. The output shows the query execution, job statistics, and the resulting correlation coefficient.

```
FAILED: ParseException line 2:1 cannot recognize input near '(' 'select' 'length' in joinSource
hive> select corr(review_length,avg_rating) as coeff from
> (select length(review_headline) as review_length,avg(star_rating)as avg_rating
from amazon_reviews_v2 group by length(review_headline))s;
Query ID = hadoop_20200412230103_2cb68446-f262-46a3-aebc-77d4823386bb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0017)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container SUCCEEDED 24 24 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: 37.99 s  

-----  

OK  

-.3877144515152911  

Time taken: 38.362 seconds, Fetched: 1 row(s)  

hive>
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

23)Correlation Between products and star rating of Music and Digital Music Purchase ?

From the analysis there is strong correlation

```
select corr(product_count,avg_stars) correlation 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_Music_Purchase'))
and product_category in ('Music'))))s;
```

The screenshot shows a terminal window titled 'ec2-35-175-209-35.compute-1.amazonaws.com (hadoop)' (Session 1). The terminal content includes:

- A file browser sidebar showing a directory structure under '/home/hadoop/'.
- The command: `hive> select corr(product_count,avg_stars) correlation from`
- An error message: `FAILED: ParseException line 2:1 cannot recognize input near '(' 'select' 'distinct' in joinSource`
- The full query: `hive> select corr(product_count,avg_stars) correlation 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_Music_Purchase'))
> and product_category in ('Music'))))s;`
- Query ID: `Query ID = hadoop_20200412230330_515b7698-afef-4ea9-8e27-64dc4b4cadae`
- Total jobs: `1`
- Launching job 1 out of 1
- Status: `Running (Executing on YARN cluster with App id application_1586723064686_0017)`
- Execution statistics table:

VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	24	24	0	0	0	0
Map 7	container	SUCCEEDED	25	25	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

- Completion message: `VERTICES: 10/10 [=====>] 100% ELAPSED TIME: 97.91 s`
- Success message: `OK`
- Time taken: `98.65 seconds, Fetched: 1 row(s)`
- Hive prompt: `hive>`

24)Correlation Between products and star rating of Digital Video Games and Video Games ?

It is weakly correlated

```
select corr(product_count,avg_stars) correlation 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'))))s;
```

```

ec2-35-175-209-35.compute-1.amazonaws.com (hadoop) (1)
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/ [3] ec2-35-175-209-35.compute-1.amazonaws.com (hadoop) (1)
1
Reducers: 3 ..... container SUCCEEDED 2 2 0 0 0 0 0
Reducers: 4 ..... container SUCCEEDED 2 2 0 0 0 0 0
Reducers: 5 ..... container SUCCEEDED 2 2 0 0 0 0 0
Reducers: 6 ..... container SUCCEEDED 1 1 0 0 0 0 0
Reducers: 8 ..... container SUCCEEDED 1 1 0 0 0 0 0
VERTICES: 10/10 [=====>] 100% ELAPSED TIME: 97.91 s
1
1 OK
1 0.9099999999999999
Time taken: 98.65 seconds, Fetched: 1 row(s)
hive> select corr(product_count,avg_stars) correlation 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')))
Query ID: hadoop_20260412230803_9dctf5f49-730c-42d9-f0e08862681b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0017)

----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 ..... container SUCCEEDED 24 24 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: 69/69 [=====>] 100% ELAPSED TIME: 99.35 s
----- OK
-0.14926298305179208
Time taken: 100.018 seconds, Fetched: 1 row(s)
hive>

```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

25)What is the prior day rating?

We can analyze the prior day rating with the help of following query which is one of the KPI's for the business purpose

```

select review_date,star_rating,count_star_rating_1,
lag(count_star_rating_1,1) over (partition by star_rating order by
review_date,star_rating asc) as prior_day_star_rating,
count_star_rating_1-coalesce(lag(count_star_rating_1,1) over (partition by
star_rating order by review_date,star_rating asc),0) as daily_squeeze
from (
  select review_date,star_rating,count(star_rating) as count_star_rating_1
  from amazon_reviews_v2 group by review_date,star_rating order by
  review_date,star_rating asc limit 30)s;

```

```

ec2-35-175-209-35.compute-1.amazonaws.com (hadoop) (1)
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/ [3] ec2-35-175-209-35.compute-1.amazonaws.com (hadoop) (1)
1
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.main(RunJar.java:239)
FAILED: ParseException line 1:13 cannot recognize input near 'date' ',' 'star_rating' in expression specification
hive> select review_date,star_rating,count_star_rating_1
> lag(count_star_rating_1,1) over (partition by star_rating order by review_date,star_rating asc) as prior_day_star_rating,
> count_star_rating_1-coalesce(lag(count_star_rating_1,1) over (partition by star_rating order by review_date,star_rating asc),0) as daily_squeeze
> from (
> select review_date,star_rating,count(star_rating) as count_star_rating_1
> from amazon_reviews_v2 group by review_date,star_rating order by review_date,star_rating asc limit 30)s;
Query ID: hadoop_20260412231203_d0940dfb-0c52-45b4-bccf-865aa6f4efb4
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1586723064686_0017)

----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 ..... container SUCCEEDED 24 24 0 0 0 0
Reducer 2 ..... container SUCCEEDED 2 2 0 0 0 0
Reducer 3 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 4 ..... container SUCCEEDED 1 1 0 0 0 0
VERTICES: 0/44 [=---->] 100% ELAPSED TIME: 32.38 s
----- OK
2005-01-01 1 76 NULL 76
2005-01-02 1 84 76 8
2005-01-03 1 90 84 15
2005-01-04 1 101 99 2
2005-01-05 1 105 101 4
2005-01-06 1 103 105 -2
2005-01-01 2 24 NULL 24
2005-01-02 2 44 24 20
2005-01-03 2 56 44 12
2005-01-04 2 62 56 6
2005-01-05 2 70 62 8
2005-01-06 2 48 70 -22
2005-01-01 3 68 NULL 68
2005-01-02 3 72 68 4
2005-01-03 3 74 72 2

```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: <https://mobaxterm.mobatek.net>

References:

Article 'Amazon Review Data: Spotting Trends and Fake Reviews' by Mark Chopping

'A Study of Amazon User Review Data using Visualization' by Preeti Bamane