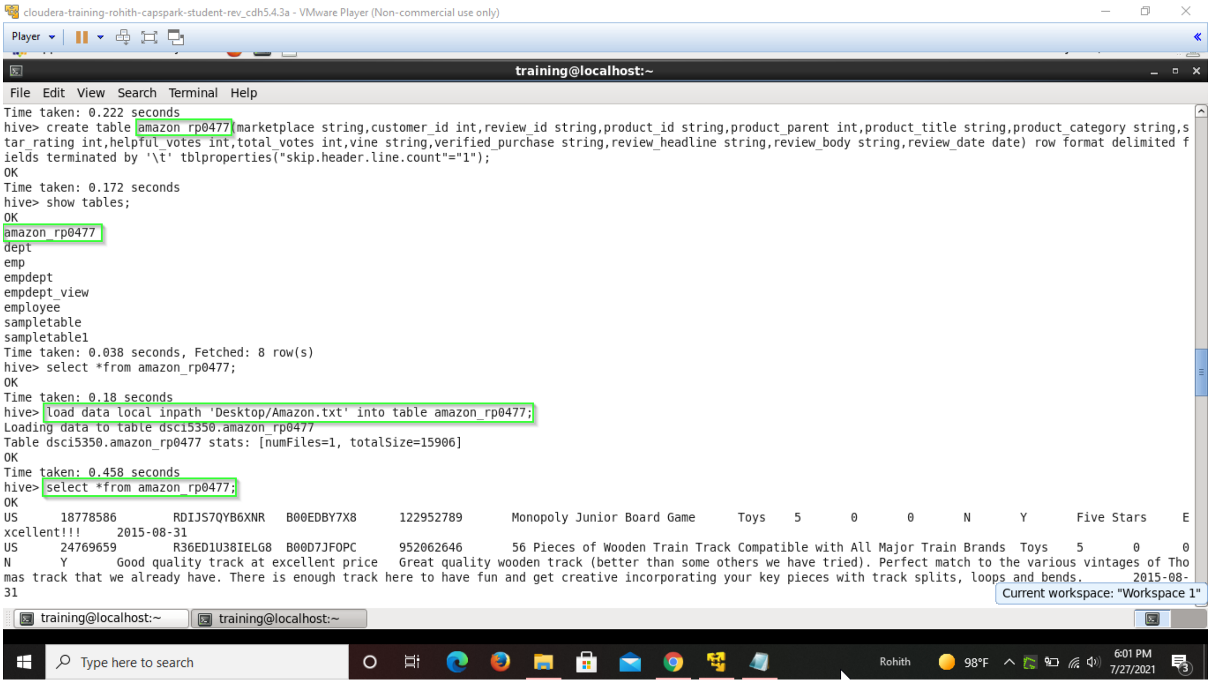
Detecting Sentiment Label for amazon customer reviews

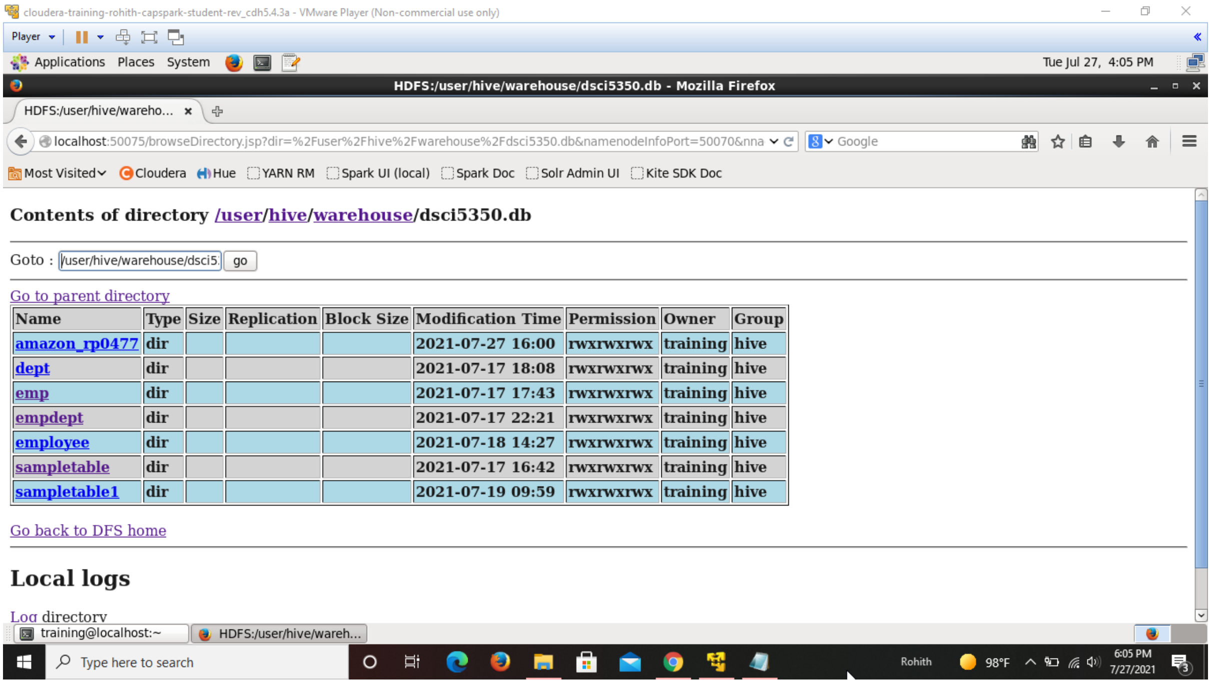
In the first step of my analysis, I have used an exising data base with course name dsci5350 and I have created a table with name as 'amazon_rp0477' with the column names and data types as shown below in the query. As there is a chance of displaying null values for the column headers, I have set the table properties to skip the header line . In the second query, I have ran show tables to see whether the table has been created or not.



cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
training@localhost:~
File Edit View Search Terminal Help

```
Time taken: 0.222 seconds
hive> create table amazon_rp0477(marketplace string,customer_id int,review_id string,product_id string,product_parent int,product_title string,product_category string,s
tar_rating int,helpful_votes int,total_votes int,vine string,verified_purchase string,review_headline string,review_body string,review_date date) row format delimited f
ields terminated by '\t' tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.172 seconds
hive> show tables;
OK
amazon_rp0477
dept
emp
empdept
empdept_view
employee
sampletable
sampletable1
Time taken: 0.038 seconds, Fetched: 8 row(s)
hive> select *from amazon_rp0477;
OK
Time taken: 0.18 seconds
hive> load data local inpath 'Desktop/Amazon.txt' into table amazon_rp0477;
Loading data to table dsci5350.amazon_rp0477
Table dsci5350.amazon_rp0477 stats: [numFiles=1, totalSize=15906]
OK
Time taken: 0.458 seconds
hive> select *from amazon_rp0477;
OK
US      18778586        RDIJS7QYB6XNR   B00EDBY7X8      122952789       Monopoly Junior Board Game      Toys    5       0       0       N       Y       Five Stars      E
xcellent!!!     2015-08-31
US      24769659        R36ED1U38IELG8  B00D7JFOPC      952062646       56 Pieces of Wooden Train Track Compatible with All Major Train Brands  Toys    5       0       0
N       Y       Good quality track at excellent price   Great quality wooden track (better than some others we have tried). Perfect match to the various vintages of Tho
mas track that we already have. There is enough track here to have fun and get creative incorporating your key pieces with track splits, loops and bends.       2015-08-
31
```

Then I have loaded the table with the file that I saved on my local desktop using 'load data local inpath 'Desktop/Amazon.txt' into table amazon_rp0477'. Using this command I have loaded the data into the table.



cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Tue Jul 27, 4:05 PM
HDFS:/user/hive/warehouse/dsci5350.db - Mozilla Firefox
HDFS:/user/hive/wareho...
localhost:50075/browseDirectory.jsp?dir=%2Fuser%2Fhive%2Fwarehouse%2Fdsci5350.db&namenodeInfoPort=50070&nna
Google
Most Visited
Cloudera
Hue
YARN RM
Spark UI (local)
Spark Doc
Solr Admin UI
Kite SDK Doc

Contents of directory /user/hive/warehouse/dsci5350.db

Goto : /user/hive/warehouse/dsci5 go

Go to parent directory

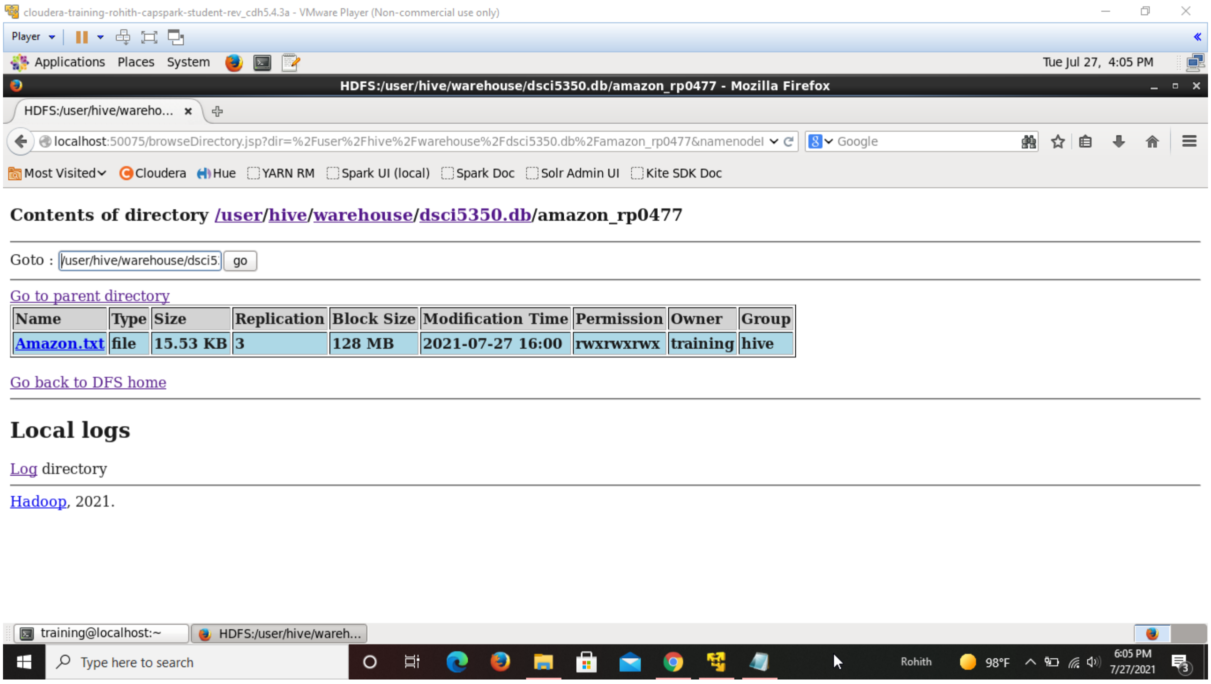
Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
amazon_rp0477	dir				2021-07-27 16:00	rwxrwxrwx	training	hive
dept	dir				2021-07-17 18:08	rwxrwxrwx	training	hive
emp	dir				2021-07-17 17:43	rwxrwxrwx	training	hive
empdept	dir				2021-07-17 22:21	rwxrwxrwx	training	hive
employee	dir				2021-07-18 14:27	rwxrwxrwx	training	hive
sampletable	dir				2021-07-17 16:42	rwxrwxrwx	training	hive
sampletable1	dir				2021-07-19 09:59	rwxrwxrwx	training	hive

Go back to DFS home

Local logs

Log directory

Now, I'm checking whether the file that is sitting on desktop has been loaded into amazon_rp0477 table or not. From the below screenshot, it is evident that amazon.txt file is loaded into amazon_rp0477.


cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Tue Jul 27, 4:05 PM
HDFS:/user/hive/warehouse/dsci5350.db/amazon_rp0477 - Mozilla Firefox
HDFS:/user/hive/wareho...
localhost:50075/browseDirectory.jsp?dir=%2Fuser%2Fhive%2Fwarehouse%2Fdsci5350.db%2Famazon_rp0477&namenodel
Google
Most Visited
Cloudera
Hue
YARN RM
Spark UI (local)
Spark Doc
Solr Admin UI
Kite SDK Doc
training@localhost:~
HDFS:/user/hive/wareh...
Type here to search
Rohith
98°F
6:05 PM
7/27/2021

Contents of directory /user/hive/warehouse/dsci5350.db/amazon_rp0477

Goto : /user/hive/warehouse/dsci5 go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
Amazon.txt	file	15.53 KB	3	128 MB	2021-07-27 16:00	rwxrwxrwx	training	hive

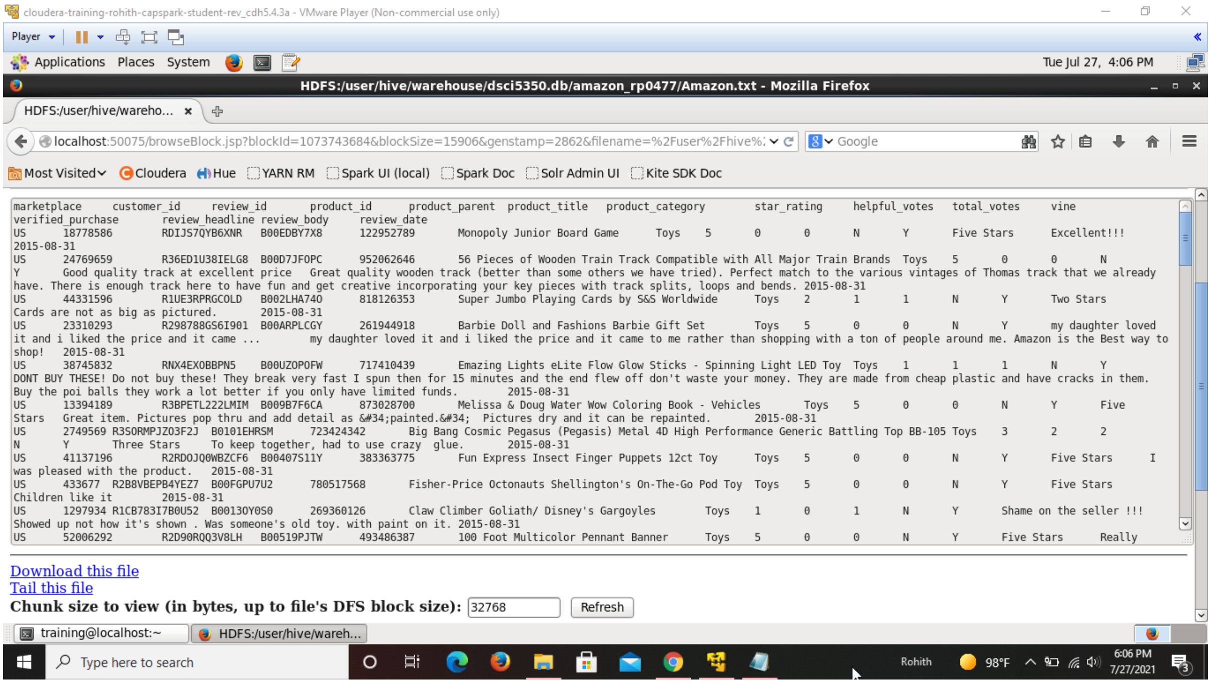
Go back to DFS home

Local logs

Log directory

Hadoop, 2021.

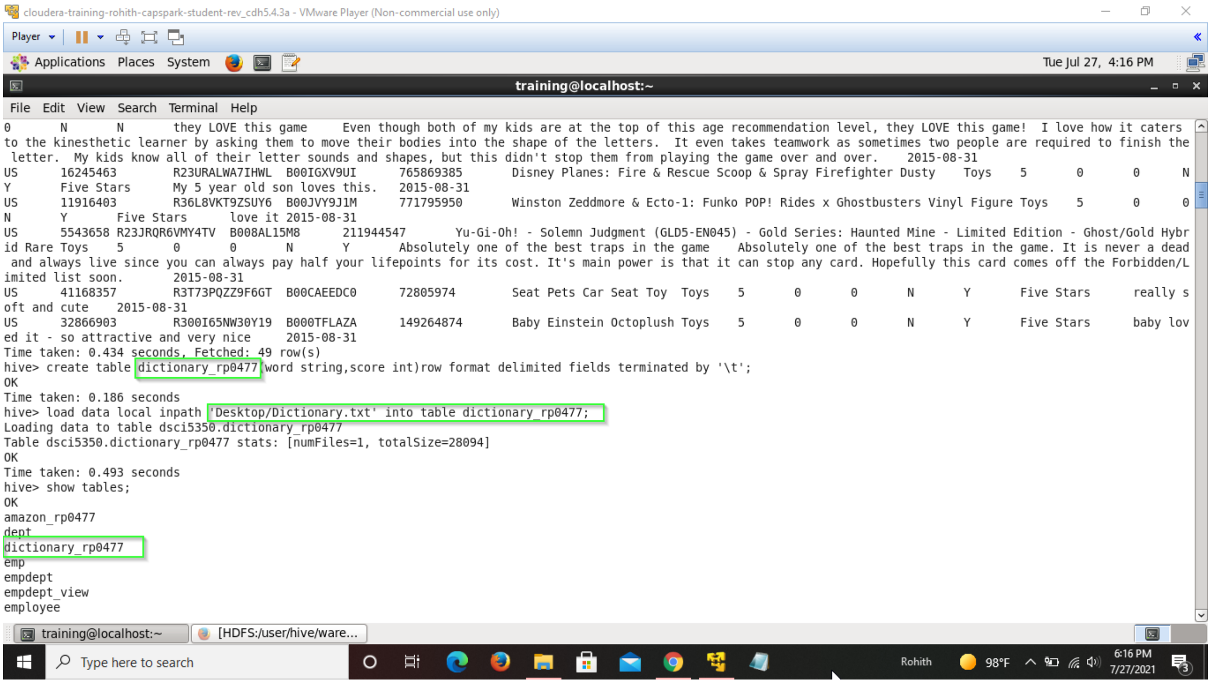
Below is the data that can be viewed when clicked on the amazon.txt file.


cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Tue Jul 27, 4:06 PM
HDFS:/user/hive/warehouse/dsci5350.db/amazon_rp0477/Amazon.txt - Mozilla Firefox
HDFS:/user/hive/wareho...
localhost:50075/browseBlock.jsp?blockId=1073743684&blockSize=15906&genstamp=2862&filename=%2Fuser%2Fhive%2
Google
Most Visited
Cloudera
Hue
YARN RM
Spark UI (local)
Spark Doc
Solr Admin UI
Kite SDK Doc
training@localhost:~
HDFS:/user/hive/wareh...
Type here to search
Rohith
98°F
6:06 PM
7/27/2021

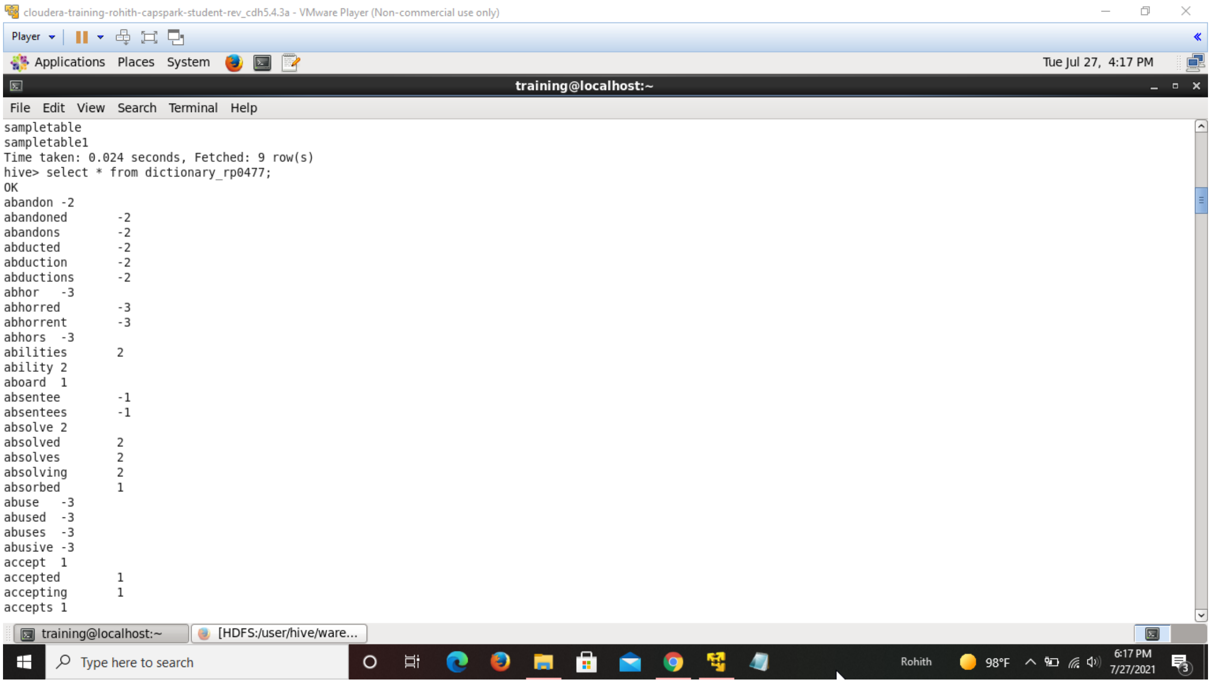
```
marketplace     customer_id     review_id       product_id      product_parent  product_title   product_category        star_rating     helpful_votes   total_votes     vine
verified_purchase       review_headline review_body     review_date
US      18778586        RDIJS7QYB6XNR   B00EDBY7X8      122952789       Monopoly Junior Board Game      Toys    5       0       0       N       Y       Five Stars      Excellent!!!
2015-08-31
US      24769659        R36ED1U38IELG8  B00D7JFOPC      952062646       56 Pieces of Wooden Train Track Compatible with All Major Train Brands  Toys    5       0       0       N
Y       Good quality track at excellent price   Great quality wooden track (better than some others we have tried). Perfect match to the various vintages of Thomas track that we already
have. There is enough track here to have fun and get creative incorporating your key pieces with track splits, loops and bends. 2015-08-31
US      44331596        R1UE3RPRGCOLD   B002LHA740      818126353       Super Jumbo Playing Cards by S&S Worldwide      Toys    2       1       1       N       Y       Two Stars
Cards are not as big as pictured.       2015-08-31
US      23310293        R298788GS6I901  B00ARPLCGY      261944918       Barbie Doll and Fashions Barbie Gift Set        Toys    5       0       0       N       Y       my daughter loved
it and i liked the price and it came ...        my daughter loved it and i liked the price and it came to me rather than shopping with a ton of people around me. Amazon is the Best way to
shop!   2015-08-31
US      38745832        RNX4EXOBBPN5    B00UZOPOFW      717410439       Emazing Lights eLite Flow Glow Sticks - Spinning Light LED Toy  Toys    1       1       1       N       Y
DONT BUY THESE! Do not buy these! They break very fast I spun then for 15 minutes and the end flew off don't waste your money. They are made from cheap plastic and have cracks in them.
Buy the poi balls they work a lot better if you only have limited funds.        2015-08-31
US      13394189        R3BPETL222LMIM  B009B7F6CA      873028700       Melissa & Doug Water Wow Coloring Book - Vehicles       Toys    5       0       0       N       Y       Five
Stars   Great item. Pictures pop thru and add detail as &#34;painted.&#34;  Pictures dry and it can be repainted.       2015-08-31
US      2749569 R3SORMPJZO3F2J  B0101EHRSM      723424342       Big Bang Cosmic Pegasus (Pegasis) Metal 4D High Performance Generic Battling Top BB-105 Toys    3       2       2
N       Y       Three Stars     To keep together, had to use crazy  glue.       2015-08-31
US      41137196        R2RDOJQ0WBZCF6  B00407S11Y      383363775       Fun Express Insect Finger Puppets 12ct Toy      Toys    5       0       0       N       Y       Five Stars      I
was pleased with the product.   2015-08-31
US      433677  R2B8VBEPB4YEZ7  B00FGPU7U2      780517568       Fisher-Price Octonauts Shellington's On-The-Go Pod Toy  Toys    5       0       0       N       Y       Five Stars
Children like it        2015-08-31
US      1297934 R1CB783I7B0U52  B0013OY0S0      269360126       Claw Climber Goliath/ Disney's Gargoyles        Toys    1       0       1       N       Y       Shame on the seller !!!
Showed up not how it's shown . Was someone's old toy. with paint on it. 2015-08-31
US      52006292        R2D90RQQ3V8LH   B00519PJTW      493486387       100 Foot Multicolor Pennant Banner      Toys    5       0       0       N       Y       Five Stars      Really
```

Download this file
Tail this file
Chunk size to view (in bytes, up to file's DFS block size): 32768 Refresh

I have created another table for dictionary.txt file as dictionary_rp0477


cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Tue Jul 27, 4:16 PM
training@localhost:~
File Edit View Search Terminal Help
training@localhost:~
[HDFS:/user/hive/ware...
Type here to search
Rohith
98°F
6:16 PM
7/27/2021

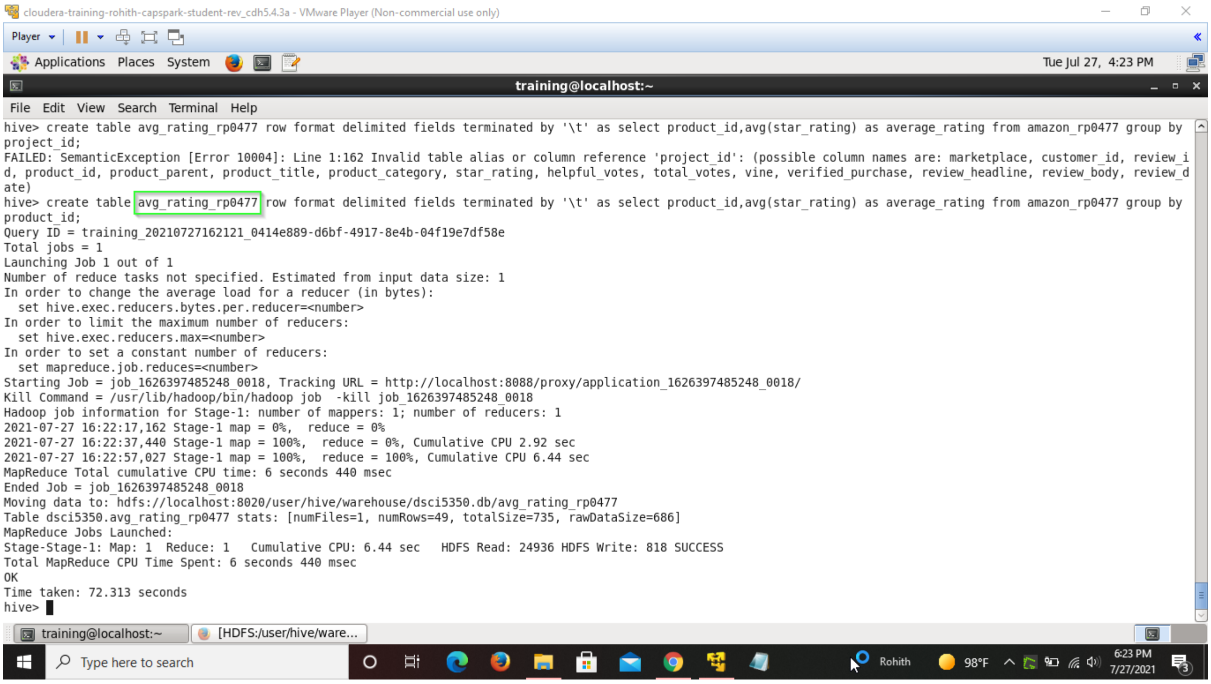
```
0       N       N       they LOVE this game     Even though both of my kids are at the top of this age recommendation level, they LOVE this game!  I love how it caters
to the kinesthetic learner by asking them to move their bodies into the shape of the letters.  It even takes teamwork as sometimes two people are required to finish the
 letter.  My kids know all of their letter sounds and shapes, but this didn't stop them from playing the game over and over.    2015-08-31
US      16245463        R23URALWA7IHWL  B00IGXV9UI      765869385       Disney Planes: Fire & Rescue Scoop & Spray Firefighter Dusty    Toys    5       0       0       N
Y       Five Stars      My 5 year old son loves this.   2015-08-31
US      11916403        R36L8VKT9ZSUY6  B00JVY9J1M      771795950       Winston Zeddmore & Ecto-1: Funko POP! Rides x Ghostbusters Vinyl Figure Toys    5       0       0
N       Y       Five Stars      love it 2015-08-31
US      5543658 R23JRQR6VMY4TV  B008AL15M8      211944547       Yu-Gi-Oh! - Solemn Judgment (GLD5-EN045) - Gold Series: Haunted Mine - Limited Edition - Ghost/Gold Hybr
id Rare Toys    5       0       0       N       Y       Absolutely one of the best traps in the game    Absolutely one of the best traps in the game. It is never a dead
 and always live since you can always pay half your lifepoints for its cost. It's main power is that it can stop any card. Hopefully this card comes off the Forbidden/L
imited list soon.       2015-08-31
US      41168357        R3T73PQZZ9F6GT  B00CAEEDC0      72805974        Seat Pets Car Seat Toy  Toys    5       0       0       N       Y       Five Stars      really s
oft and cute    2015-08-31
US      32866903        R300I65NW30Y19  B000TFLAZA      149264874       Baby Einstein Octoplush Toys    5       0       0       N       Y       Five Stars      baby lov
ed it - so attractive and very nice     2015-08-31
Time taken: 0.434 seconds, Fetched: 49 row(s)
hive> create table dictionary_rp0477(word string,score int)row format delimited fields terminated by '\t';
OK
Time taken: 0.186 seconds
hive> load data local inpath 'Desktop/Dictionary.txt' into table dictionary_rp0477;
Loading data to table dsci5350.dictionary_rp0477
Table dsci5350.dictionary_rp0477 stats: [numFiles=1, totalSize=28094]
OK
Time taken: 0.493 seconds
hive> show tables;
OK
amazon_rp0477
dept
dictionary_rp0477
emp
empdept
empdept_view
employee
```


cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Tue Jul 27, 4:17 PM
training@localhost:~
File Edit View Search Terminal Help
training@localhost:~
[HDFS:/user/hive/ware...
Type here to search
Rohith
98°F
6:17 PM
7/27/2021

```
sampletable
sampletable1
Time taken: 0.024 seconds, Fetched: 9 row(s)
hive> select * from dictionary_rp0477;
OK
abandon -2
abandoned       -2
abandons        -2
abducted        -2
abduction       -2
abductions      -2
abhor   -3
abhorred        -3
abhorrent       -3
abhors  -3
abilities       2
ability 2
aboard  1
absentee        -1
absentees       -1
absolve 2
absolved        2
absolves        2
absolving       2
absorbed        1
abuse   -3
abused  -3
abuses  -3
abusive -3
accept  1
accepted        1
accepting       1
accepts 1
```

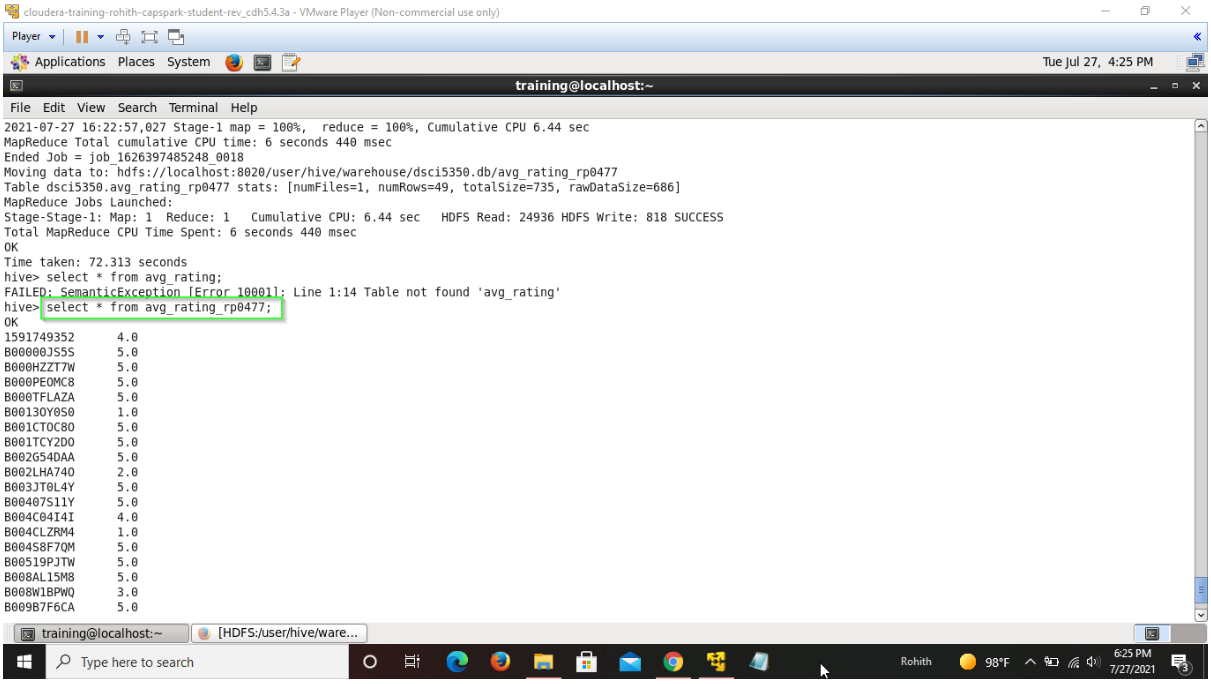
The above screenshot displays all the data that is contained in the file or loaded from the dictionary file into the table which we created.

2. A) Now, In order to find the products that has the average rating less than 2; we need to create a table that has average of the star rating column. So we created another table avg_rating_rp0477 with product id and their avg ratings.



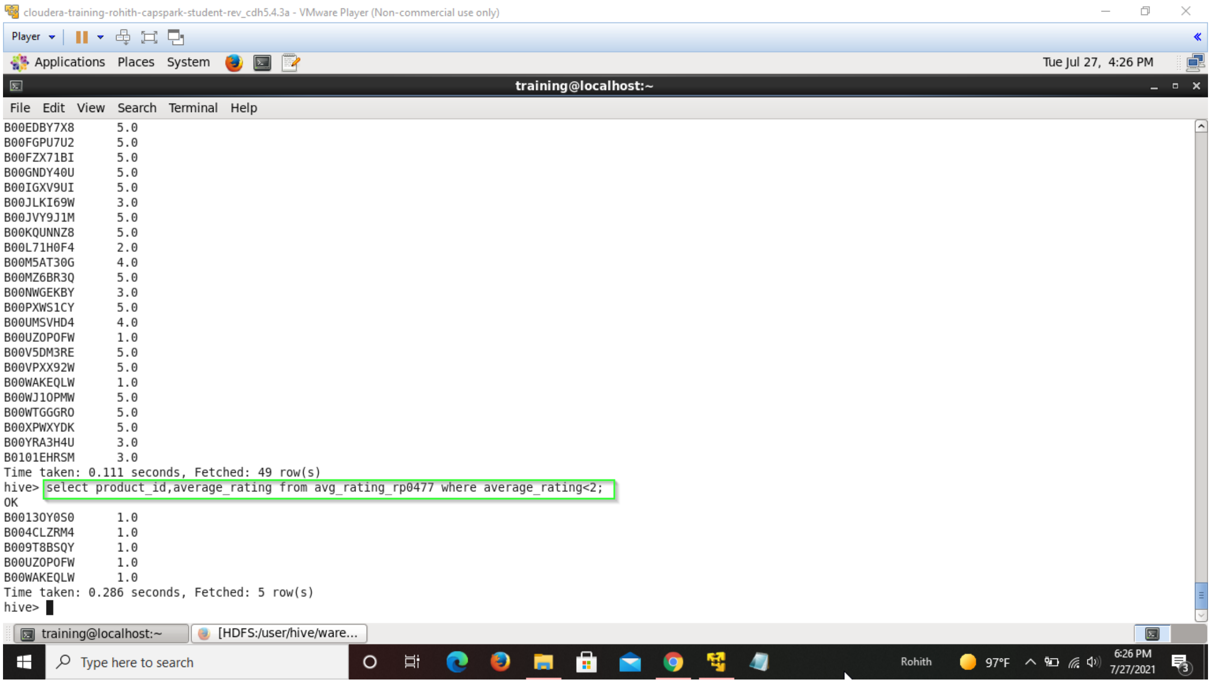
```
hive> create table avg_rating_rp0477 row format delimited fields terminated by '\t' as select product_id,avg(star_rating) as average_rating from amazon_rp0477 group by project_id;
FAILED: SemanticException [Error 10004]: Line 1:162 Invalid table alias or column reference 'project_id': (possible column names are: marketplace, customer_id, review_id, product_id, product_parent, product_title, product_category, star_rating, helpful_votes, total_votes, vine, verified_purchase, review_headline, review_body, review_date)
hive> create table avg_rating_rp0477 row format delimited fields terminated by '\t' as select product_id,avg(star_rating) as average_rating from amazon_rp0477 group by product_id;
Query ID = training_20210727162121_0414e889-d6bf-4917-8e4b-04f19e7df58e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1626397485248_0018, Tracking URL = http://localhost:8088/proxy/application_1626397485248_0018/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1626397485248_0018
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-27 16:22:17,162 Stage-1 map = 0%,  reduce = 0%
2021-07-27 16:22:37,440 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.92 sec
2021-07-27 16:22:57,027 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.44 sec
MapReduce Total cumulative CPU time: 6 seconds 440 msec
Ended Job = job_1626397485248_0018
Moving data to: hdfs://localhost:8020/user/hive/warehouse/dsci5350.db/avg_rating_rp0477
Table dsci5350.avg_rating_rp0477 stats: [numFiles=1, numRows=49, totalSize=735, rawDataSize=686]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.44 sec   HDFS Read: 24936 HDFS Write: 818 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 440 msec
OK
Time taken: 72.313 seconds
hive> 
```

The above query created a table for average reviews.



```
2021-07-27 16:22:57,027 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.44 sec
MapReduce Total cumulative CPU time: 6 seconds 440 msec
Ended Job = job_1626397485248_0018
Moving data to: hdfs://localhost:8020/user/hive/warehouse/dsci5350.db/avg_rating_rp0477
Table dsci5350.avg_rating_rp0477 stats: [numFiles=1, numRows=49, totalSize=735, rawDataSize=686]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.44 sec   HDFS Read: 24936 HDFS Write: 818 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 440 msec
OK
Time taken: 72.313 seconds
hive> select * from avg_rating;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'avg_rating'
hive> select * from avg_rating_rp0477;
OK
1591749352      4.0
B00000JS5S      5.0
B000HZZT7W      5.0
B000PEOMC8      5.0
B000TFLAZA      5.0
B0013OY0S0      1.0
B001CTOC8O      5.0
B001TCY2DO      5.0
B002G54DAA      5.0
B002LHA74O      2.0
B003JT0L4Y      5.0
B00407S11Y      5.0
B004C04I4I      4.0
B004CLZRM4      1.0
B004S8F7QM      5.0
B00519PJTW      5.0
B008AL15M8      5.0
B008W1BPWQ      3.0
B009B7F6CA      5.0
```

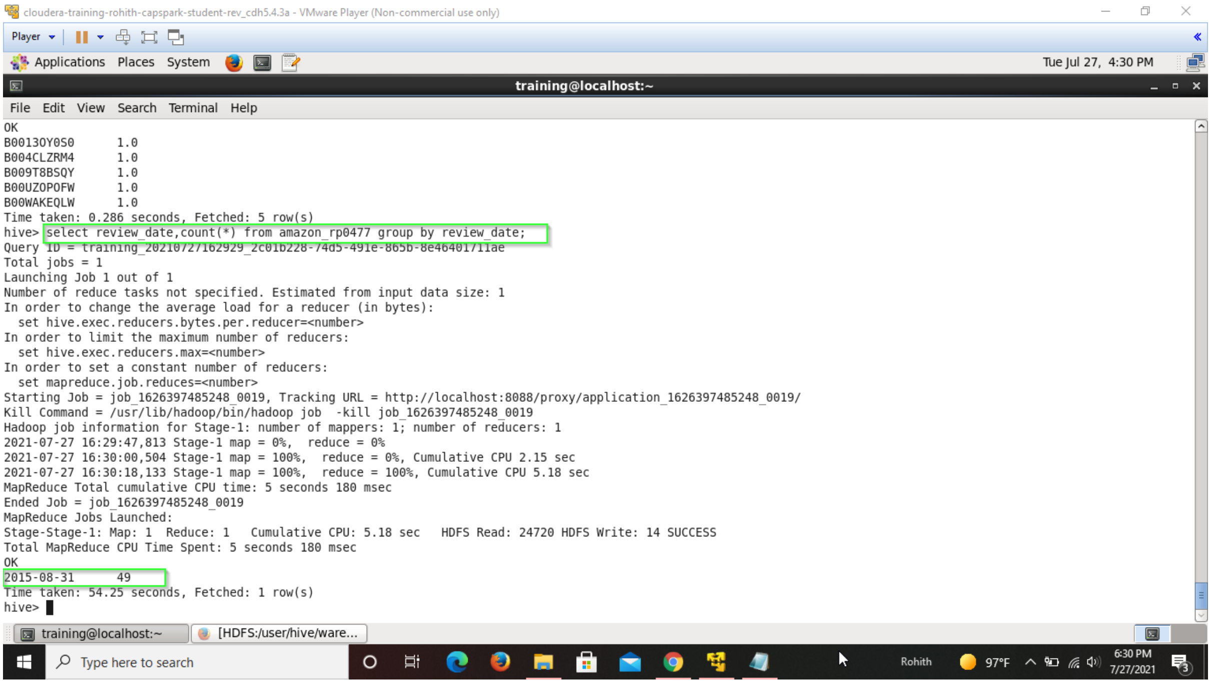
Here, we can see all the data i.e. with the product_id and avg_rating


cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Tue Jul 27, 4:26 PM
training@localhost:~
File Edit View Search Terminal Help
training@localhost:~
[HDFS:/user/hive/ware...
Type here to search
Rohith
97°F
6:26 PM
7/27/2021

```
B00EDBY7X8      5.0
B00FGPU7U2      5.0
B00FZX71BI      5.0
B00GNDY40U      5.0
B00IGXV9UI      5.0
B00JLKI69W      3.0
B00JVY9J1M      5.0
B00KQUNNZ8      5.0
B00L71H0F4      2.0
B00M5AT30G      4.0
B00MZ6BR3Q      5.0
B00NWGEKBY      3.0
B00PXWS1CY      5.0
B00UMSVHD4      4.0
B00UZOPOFW      1.0
B00V5DM3RE      5.0
B00VPXX92W      5.0
B00WAKEQLW      1.0
B00WJ1OPMW      5.0
B00WTGGGRO      5.0
B00XPWXYDK      5.0
B00YRA3H4U      3.0
B0101EHRSM      3.0
Time taken: 0.111 seconds, Fetched: 49 row(s)
hive> select product_id,average_rating from avg_rating_rp0477 where average_rating<2;
OK
B0013OY0S0      1.0
B004CLZRM4      1.0
B009T8BSQY      1.0
B00UZOPOFW      1.0
B00WAKEQLW      1.0
Time taken: 0.286 seconds, Fetched: 5 row(s)
hive> 
```

In the above screenshot, I have ran a query to find the average rating that is less than 2 and could see that there are 5 products that have rating less than 2.

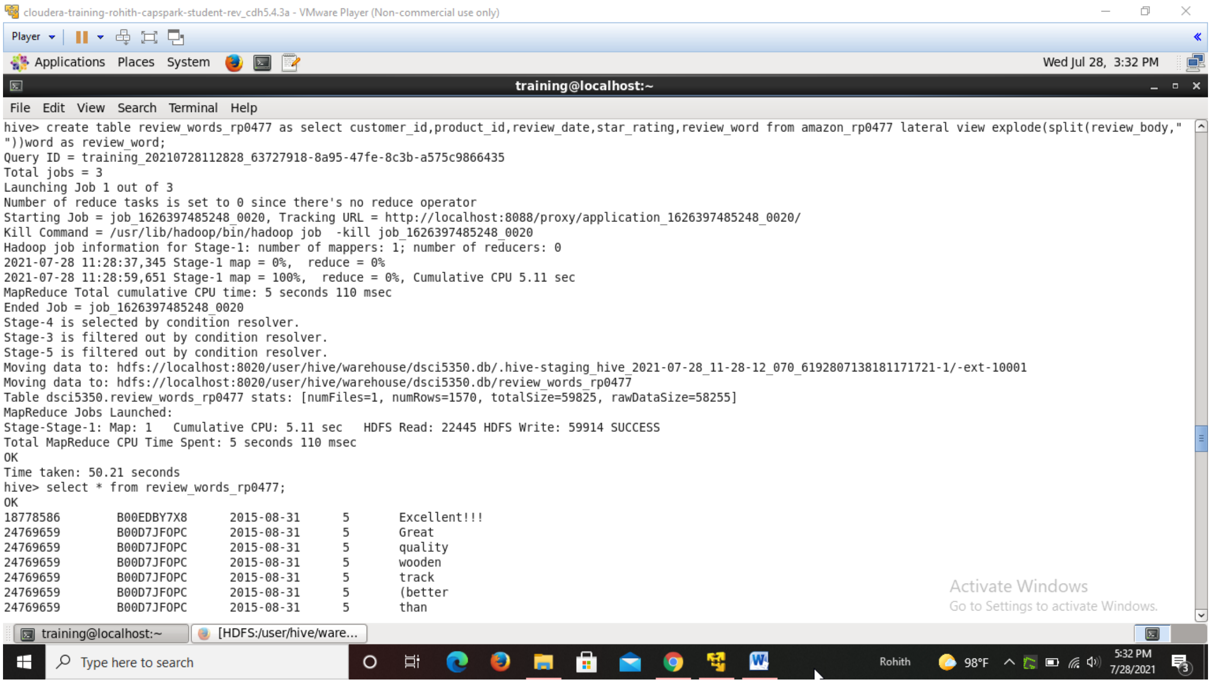
2. B) Here, we are asked to display the total number of products that are reviewed per day. I have ran a sql query to display and we have got 49 total number of products that are reviewed per day.


cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Tue Jul 27, 4:30 PM
training@localhost:~
File Edit View Search Terminal Help
training@localhost:~
[HDFS:/user/hive/ware...
Type here to search
Rohith
97°F
6:30 PM
7/27/2021

```
OK
B0013OY0S0      1.0
B004CLZRM4      1.0
B009T8BSQY      1.0
B00UZOPOFW      1.0
B00WAKEQLW      1.0
Time taken: 0.286 seconds, Fetched: 5 row(s)
hive> select review_date,count(*) from amazon_rp0477 group by review_date;
Query ID = training_20210727162929_2c01b228-74d5-491e-865b-8e46401711ae
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1626397485248_0019, Tracking URL = http://localhost:8088/proxy/application_1626397485248_0019/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1626397485248_0019
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-27 16:29:47,813 Stage-1 map = 0%,  reduce = 0%
2021-07-27 16:30:00,504 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.15 sec
2021-07-27 16:30:18,133 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.18 sec
MapReduce Total cumulative CPU time: 5 seconds 180 msec
Ended Job = job_1626397485248_0019
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.18 sec   HDFS Read: 24720 HDFS Write: 14 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 180 msec
OK
2015-08-31      49
Time taken: 54.25 seconds, Fetched: 1 row(s)
hive> 
```

2. C) Here we are asked to use sentiment analysis and display the top 5 good reviews amazon received for their products. In addition to this, we need to include customer_id, product_id, review_date, star_rating, total score and sentiment label in a new table.

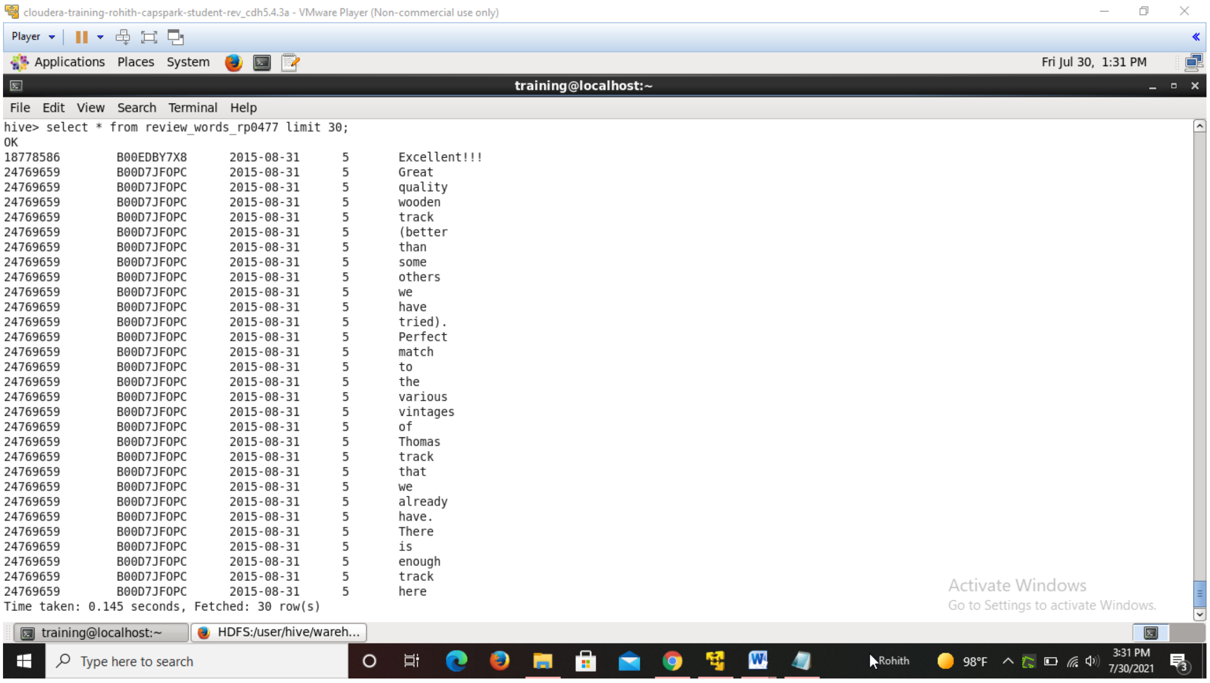
Initially we need to have two new columns that as per the question. We need to find the total score and sentiment label using the rest of the columns. In order to find the total score, first we need to find the individual score of each word and then add them to sum up for the total score. Hence we can run a split query that basically splits all the words from a review body or sentence. Below is the query that is needed to run the query.



```
hive> create table review_words_rp0477 as select customer_id,product_id,review_date,star_rating,review_word from amazon_rp0477 lateral view explode(split(review_body,"
"))word as review_word;
Query ID = training_20210728112828_63727918-8a95-47fe-8c3b-a575c9866435
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1626397485248_0020, Tracking URL = http://localhost:8088/proxy/application_1626397485248_0020/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1626397485248_0020
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2021-07-28 11:28:37,345 Stage-1 map = 0%,  reduce = 0%
2021-07-28 11:28:59,651 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.11 sec
MapReduce Total cumulative CPU time: 5 seconds 110 msec
Ended Job = job_1626397485248_0020
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://localhost:8020/user/hive/warehouse/dsci5350.db/.hive-staging_hive_2021-07-28_11-28-12_070_6192807138181171721-1/-ext-10001
Moving data to: hdfs://localhost:8020/user/hive/warehouse/dsci5350.db/review_words_rp0477
Table dsci5350.review_words_rp0477 stats: [numFiles=1, numRows=1570, totalSize=59825, rawDataSize=58255]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 5.11 sec   HDFS Read: 22445 HDFS Write: 59914 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 110 msec
OK
Time taken: 50.21 seconds
hive> select * from review_words_rp0477;
OK
18778586        B00EDBY7X8      2015-08-31      5       Excellent!!!
24769659        B00D7JFOPC      2015-08-31      5       Great
24769659        B00D7JFOPC      2015-08-31      5       quality
24769659        B00D7JFOPC      2015-08-31      5       wooden
24769659        B00D7JFOPC      2015-08-31      5       track
24769659        B00D7JFOPC      2015-08-31      5       (better
24769659        B00D7JFOPC      2015-08-31      5       than
```

cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 28, 3:32 PM
training@localhost:~
File Edit View Search Terminal Help
Activate Windows
Go to Settings to activate Windows.
training@localhost:~
[HDFS:/user/hive/ware...
Type here to search
Rohith
98°F
5:32 PM
7/28/2021

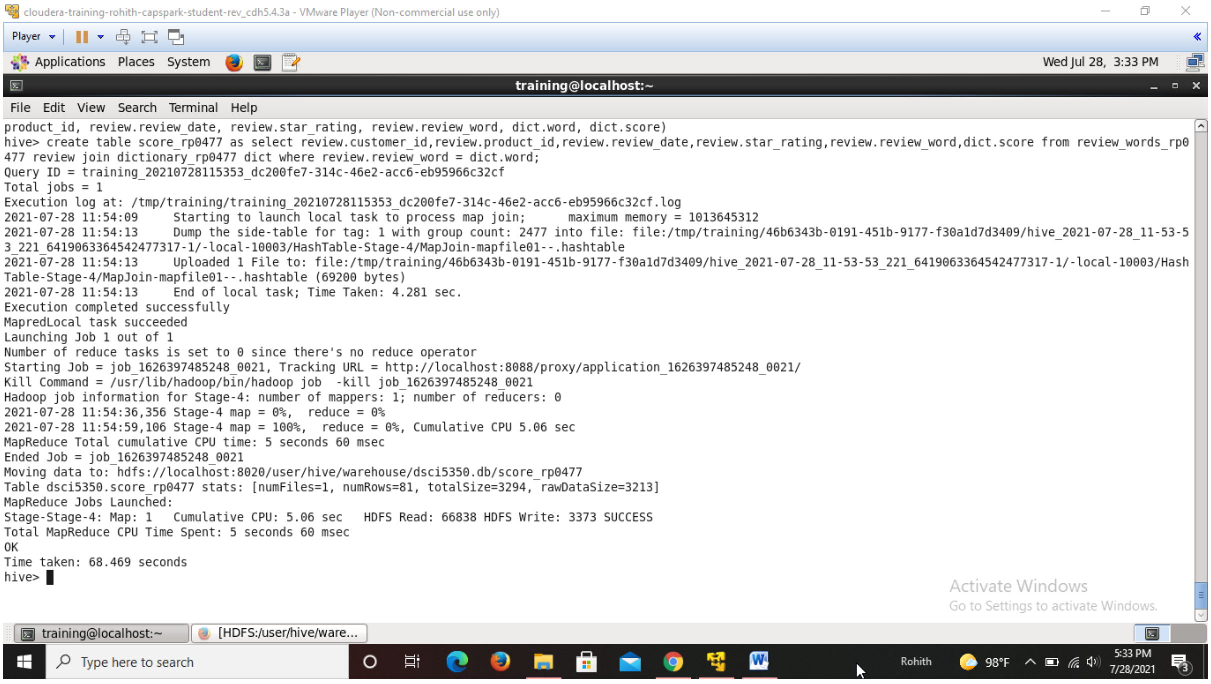
After creating a table using the query mentioned above, we display the data using select command and the above screenshot displays the data. In the below screenshot we can see the top 30 data as we have used limit 30



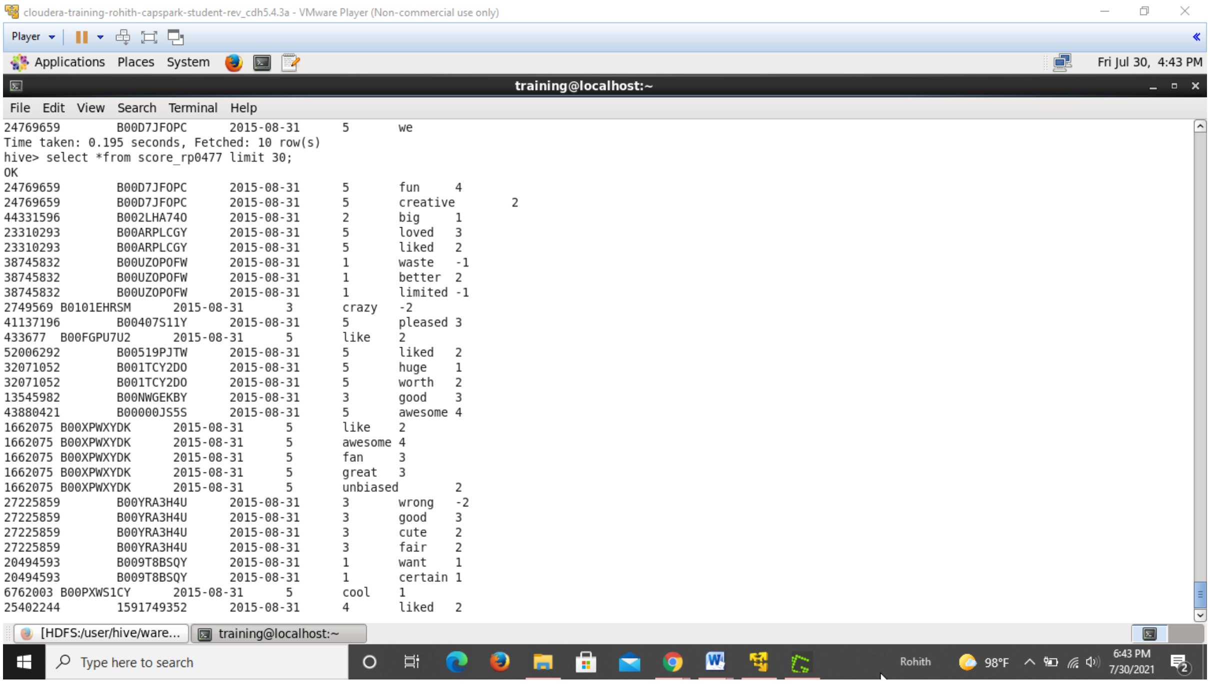
```
hive> select * from review_words_rp0477 limit 30;
OK
18778586        B00EDBY7X8      2015-08-31      5       Excellent!!!
24769659        B00D7JFOPC      2015-08-31      5       Great
24769659        B00D7JFOPC      2015-08-31      5       quality
24769659        B00D7JFOPC      2015-08-31      5       wooden
24769659        B00D7JFOPC      2015-08-31      5       track
24769659        B00D7JFOPC      2015-08-31      5       (better
24769659        B00D7JFOPC      2015-08-31      5       than
24769659        B00D7JFOPC      2015-08-31      5       some
24769659        B00D7JFOPC      2015-08-31      5       others
24769659        B00D7JFOPC      2015-08-31      5       we
24769659        B00D7JFOPC      2015-08-31      5       have
24769659        B00D7JFOPC      2015-08-31      5       tried).
24769659        B00D7JFOPC      2015-08-31      5       Perfect
24769659        B00D7JFOPC      2015-08-31      5       match
24769659        B00D7JFOPC      2015-08-31      5       to
24769659        B00D7JFOPC      2015-08-31      5       the
24769659        B00D7JFOPC      2015-08-31      5       various
24769659        B00D7JFOPC      2015-08-31      5       vintages
24769659        B00D7JFOPC      2015-08-31      5       of
24769659        B00D7JFOPC      2015-08-31      5       Thomas
24769659        B00D7JFOPC      2015-08-31      5       track
24769659        B00D7JFOPC      2015-08-31      5       that
24769659        B00D7JFOPC      2015-08-31      5       we
24769659        B00D7JFOPC      2015-08-31      5       already
24769659        B00D7JFOPC      2015-08-31      5       have.
24769659        B00D7JFOPC      2015-08-31      5       There
24769659        B00D7JFOPC      2015-08-31      5       is
24769659        B00D7JFOPC      2015-08-31      5       enough
24769659        B00D7JFOPC      2015-08-31      5       track
24769659        B00D7JFOPC      2015-08-31      5       here
Time taken: 0.145 seconds, Fetched: 30 row(s)
```

cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Fri Jul 30, 1:31 PM
training@localhost:~
File Edit View Search Terminal Help
Activate Windows
Go to Settings to activate Windows.
training@localhost:~
HDFS:/user/hive/wareh...
Type here to search
Rohith
98°F
3:31 PM
7/30/2021

Now, we need to match the score of the review with the dictionary file and find the total score, hence I have created a table score_rp0477. In the below screenshot we can see the query that finds the total score for the review_words.

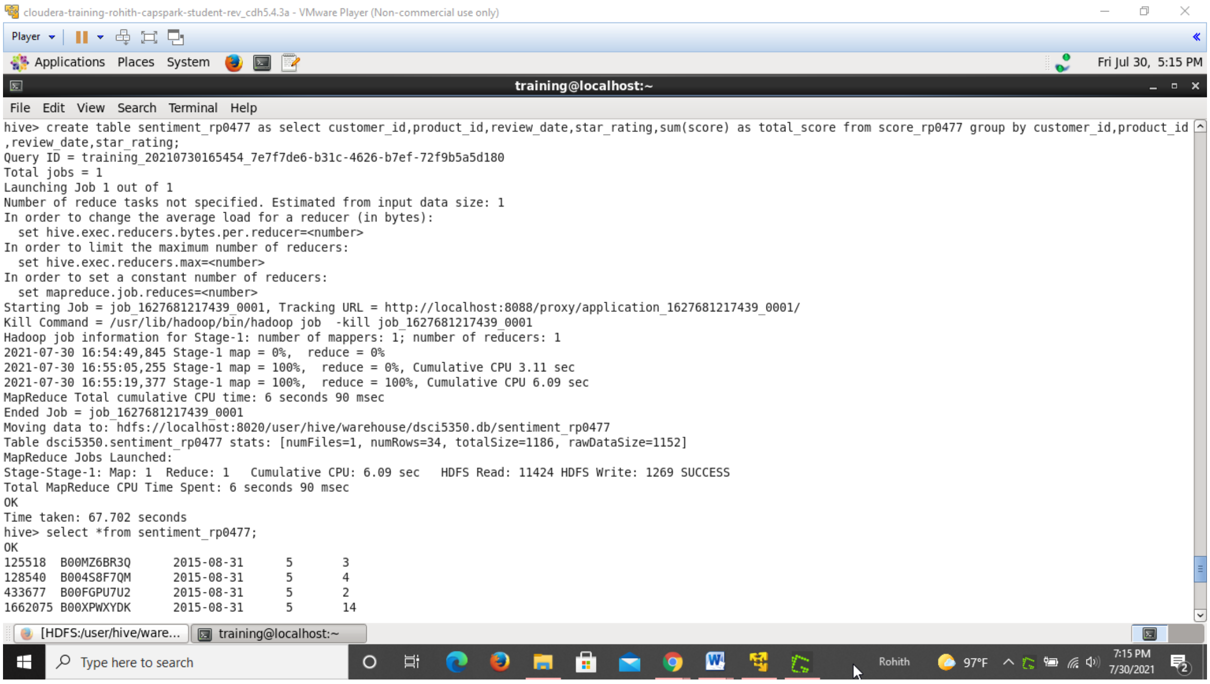

cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 28, 3:33 PM
training@localhost:~
File Edit View Search Terminal Help
Activate Windows
Go to Settings to activate Windows.
training@localhost:~
[HDFS:/user/hive/ware...
Type here to search
Rohith
98°F
5:33 PM
7/28/2021

```
product_id, review.review_date, review.star_rating, review.review_word, dict.word, dict.score)
hive> create table score_rp0477 as select review.customer_id,review.product_id,review.review_date,review.star_rating,review.review_word,dict.score from review_words_rp0
477 review join dictionary_rp0477 dict where review.review_word = dict.word;
Query ID = training_20210728115353_dc200fe7-314c-46e2-acc6-eb95966c32cf
Total jobs = 1
Execution log at: /tmp/training/training_20210728115353_dc200fe7-314c-46e2-acc6-eb95966c32cf.log
2021-07-28 11:54:09     Starting to launch local task to process map join;      maximum memory = 1013645312
2021-07-28 11:54:13     Dump the side-table for tag: 1 with group count: 2477 into file: file:/tmp/training/46b6343b-0191-451b-9177-f30a1d7d3409/hive_2021-07-28_11-53-5
3_221_6419063364542477317-1/-local-10003/HashTable-Stage-4/MapJoin-mapfile01--.hashtable
2021-07-28 11:54:13     Uploaded 1 File to: file:/tmp/training/46b6343b-0191-451b-9177-f30a1d7d3409/hive_2021-07-28_11-53-53_221_6419063364542477317-1/-local-10003/Hash
Table-Stage-4/MapJoin-mapfile01--.hashtable (69200 bytes)
2021-07-28 11:54:13     End of local task; Time Taken: 4.281 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1626397485248_0021, Tracking URL = http://localhost:8088/proxy/application_1626397485248_0021/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1626397485248_0021
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 0
2021-07-28 11:54:36,356 Stage-4 map = 0%,  reduce = 0%
2021-07-28 11:54:59,106 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 5.06 sec
MapReduce Total cumulative CPU time: 5 seconds 60 msec
Ended Job = job_1626397485248_0021
Moving data to: hdfs://localhost:8020/user/hive/warehouse/dsci5350.db/score_rp0477
Table dsci5350.score_rp0477 stats: [numFiles=1, numRows=81, totalSize=3294, rawDataSize=3213]
MapReduce Jobs Launched:
Stage-Stage-4: Map: 1   Cumulative CPU: 5.06 sec   HDFS Read: 66838 HDFS Write: 3373 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 60 msec
OK
Time taken: 68.469 seconds
hive> 
```


cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Fri Jul 30, 4:43 PM
training@localhost:~
File Edit View Search Terminal Help
[HDFS:/user/hive/ware...
training@localhost:~
Type here to search
Rohith
98°F
6:43 PM
7/30/2021

```
24769659        B00D7JFOPC      2015-08-31      5       we
Time taken: 0.195 seconds, Fetched: 10 row(s)
hive> select *from score_rp0477 limit 30;
OK
24769659        B00D7JFOPC      2015-08-31      5       fun     4
24769659        B00D7JFOPC      2015-08-31      5       creative        2
44331596        B002LHA740      2015-08-31      2       big     1
23310293        B00ARPLCGY      2015-08-31      5       loved   3
23310293        B00ARPLCGY      2015-08-31      5       liked   2
38745832        B00UZOPOFW      2015-08-31      1       waste   -1
38745832        B00UZOPOFW      2015-08-31      1       better  2
38745832        B00UZOPOFW      2015-08-31      1       limited -1
2749569 B0101EHRSM      2015-08-31      3       crazy   -2
41137196        B00407S11Y      2015-08-31      5       pleased 3
433677  B00FGPU7U2      2015-08-31      5       like    2
52006292        B00519PJTW      2015-08-31      5       liked   2
32071052        B001TCY2DO      2015-08-31      5       huge    1
32071052        B001TCY2DO      2015-08-31      5       worth   2
13545982        B00NWGEKBY      2015-08-31      3       good    3
43880421        B00000JS5S      2015-08-31      5       awesome 4
1662075 B00XPWXYDK      2015-08-31      5       like    2
1662075 B00XPWXYDK      2015-08-31      5       awesome 4
1662075 B00XPWXYDK      2015-08-31      5       fan     3
1662075 B00XPWXYDK      2015-08-31      5       great   3
1662075 B00XPWXYDK      2015-08-31      5       unbiased        2
27225859        B00YRA3H4U      2015-08-31      3       wrong   -2
27225859        B00YRA3H4U      2015-08-31      3       good    3
27225859        B00YRA3H4U      2015-08-31      3       cute    2
27225859        B00YRA3H4U      2015-08-31      3       fair    2
20494593        B009T8BSQY      2015-08-31      1       want    1
20494593        B009T8BSQY      2015-08-31      1       certain 1
6762003 B00PXWS1CY      2015-08-31      5       cool    1
25402244        1591749352      2015-08-31      4       liked   2
```

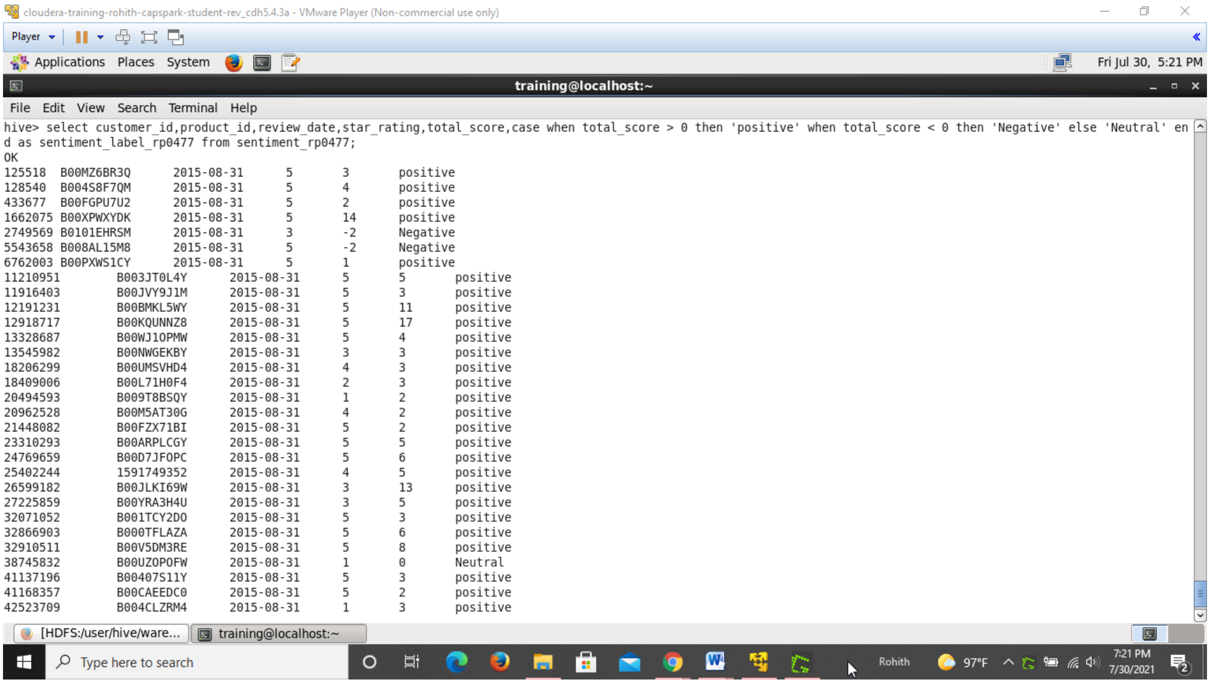
Now, we need to create a sentiment table that has a column total_score i.e. sum of all indiviual row score for the same product is calculated as total_score. Hence I have ran a query shown below and output can also be seen in the below screenshot.



```
hive> create table sentiment_rp0477 as select customer_id,product_id,review_date,star_rating,sum(score) as total_score from score_rp0477 group by customer_id,product_id
,review_date,star_rating;
Query ID = training_20210730165454_7e7f7de6-b31c-4626-b7ef-72f9b5a5d180
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1627681217439_0001, Tracking URL = http://localhost:8088/proxy/application_1627681217439_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1627681217439_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-30 16:54:49,845 Stage-1 map = 0%,  reduce = 0%
2021-07-30 16:55:05,255 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.11 sec
2021-07-30 16:55:19,377 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.09 sec
MapReduce Total cumulative CPU time: 6 seconds 90 msec
Ended Job = job_1627681217439_0001
Moving data to: hdfs://localhost:8020/user/hive/warehouse/dsci5350.db/sentiment_rp0477
Table dsci5350.sentiment_rp0477 stats: [numFiles=1, numRows=34, totalSize=1186, rawDataSize=1152]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.09 sec   HDFS Read: 11424 HDFS Write: 1269 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 90 msec
OK
Time taken: 67.702 seconds
hive> select *from sentiment_rp0477;
OK
125518  B00MZ6BR3Q      2015-08-31      5       3
128540  B004S8F7QM      2015-08-31      5       4
433677  B00FGPU7U2      2015-08-31      5       2
1662075 B00XPWXYDK      2015-08-31      5       14
```

cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Fri Jul 30, 5:15 PM
training@localhost:~
File Edit View Search Terminal Help
[HDFS:/user/hive/ware...
training@localhost:~
Type here to search
Rohith
97°F
7:15 PM
7/30/2021

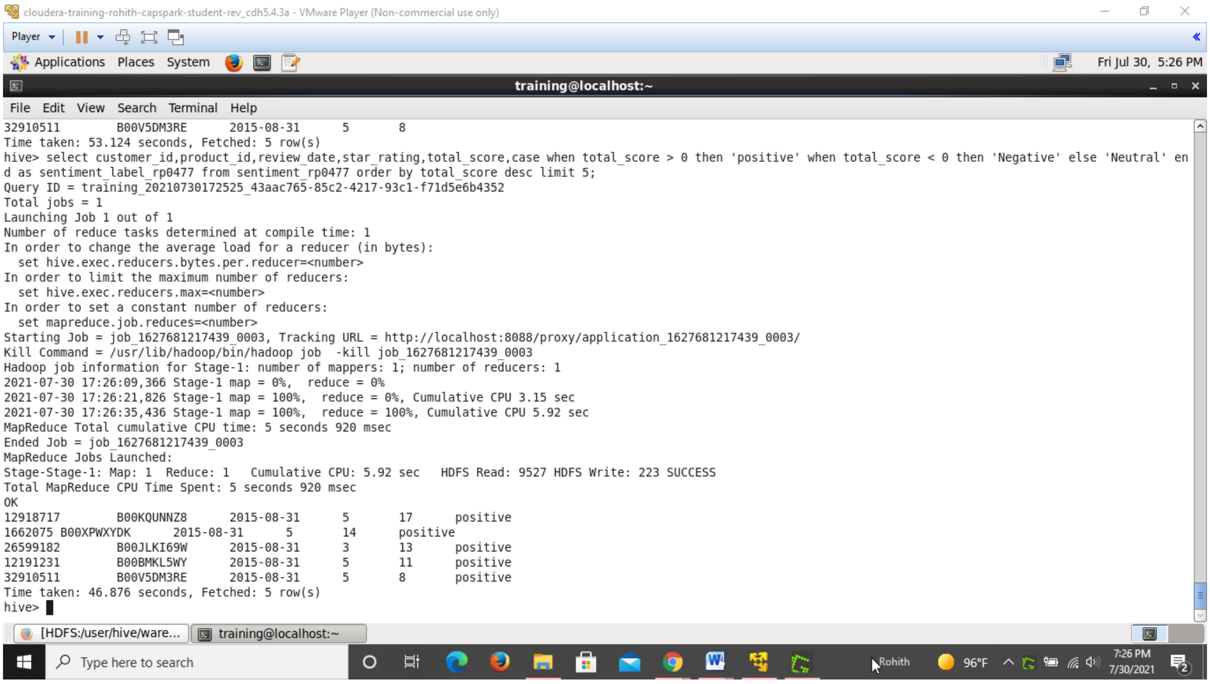
Now, we can see a column that is added beside the star rating i.e. total_score it received. Then in order to do sentiment analysis, I have ran a case statement using conditions if total_score > 0, display positive, total_score < 0 display negative else display neutral. In the below screenshot it is evident of running a query.



```
hive> select customer_id,product_id,review_date,star_rating,total_score,case when total_score > 0 then 'positive' when total_score < 0 then 'Negative' else 'Neutral' en
d as sentiment_label_rp0477 from sentiment_rp0477;
OK
125518  B00MZ6BR3Q      2015-08-31      5       3       positive
128540  B004S8F7QM      2015-08-31      5       4       positive
433677  B00FGPU7U2      2015-08-31      5       2       positive
1662075 B00XPWXYDK      2015-08-31      5       14      positive
2749569 B0101EHRSM      2015-08-31      3       -2      Negative
5543658 B008AL15M8      2015-08-31      5       -2      Negative
6762003 B00PXWS1CY      2015-08-31      5       1       positive
11210951        B003JT0L4Y      2015-08-31      5       5       positive
11916403        B00JVY9J1M      2015-08-31      5       3       positive
12191231        B00BMKL5WY      2015-08-31      5       11      positive
12918717        B00KQUNNZ8      2015-08-31      5       17      positive
13328687        B00WJ1OPMW      2015-08-31      5       4       positive
13545982        B00NWGEKBY      2015-08-31      3       3       positive
18206299        B00UMSVHD4      2015-08-31      4       3       positive
18409006        B00L71H0F4      2015-08-31      2       3       positive
20494593        B009T8BSQY      2015-08-31      1       2       positive
20962528        B00M5AT30G      2015-08-31      4       2       positive
21448082        B00FZX71BI      2015-08-31      5       2       positive
23310293        B00ARPLCGY      2015-08-31      5       5       positive
24769659        B00D7JFOPC      2015-08-31      5       6       positive
25402244        1591749352      2015-08-31      4       5       positive
26599182        B00JLKI69W      2015-08-31      3       13      positive
27225859        B00YRA3H4U      2015-08-31      3       5       positive
32071052        B001TCY2DO      2015-08-31      5       3       positive
32866903        B000TFLAZA      2015-08-31      5       6       positive
32910511        B00V5DM3RE      2015-08-31      5       8       positive
38745832        B00UZOPOFW      2015-08-31      1       0       Neutral
41137196        B00407S11Y      2015-08-31      5       3       positive
41168357        B00CAEEDC0      2015-08-31      5       2       positive
42523709        B004CLZRM4      2015-08-31      1       3       positive
```

cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Fri Jul 30, 5:21 PM
training@localhost:~
File Edit View Search Terminal Help
[HDFS:/user/hive/ware...
training@localhost:~
Type here to search
Rohith
97°F
7:21 PM
7/30/2021

In the question it is asked to display the top 5 results for the good reviewed products.



cloudera-training-rohith-capspark-student-rev_cdh5.4.3a - VMware Player (Non-commercial use only)
Player
Applications Places System
Fri Jul 30, 5:26 PM
training@localhost:~
File Edit View Search Terminal Help
[HDFS:/user/hive/ware...
training@localhost:~
Type here to search
Rohith
96°F
7:26 PM
7/30/2021

```
32910511        B00V5DM3RE      2015-08-31      5       8
Time taken: 53.124 seconds, Fetched: 5 row(s)
hive> select customer_id,product_id,review_date,star_rating,total_score,case when total_score > 0 then 'positive' when total_score < 0 then 'Negative' else 'Neutral' en
d as sentiment_label_rp0477 from sentiment_rp0477 order by total_score desc limit 5;
Query ID = training_20210730172525_43aac765-85c2-4217-93c1-f71d5e6b4352
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1627681217439_0003, Tracking URL = http://localhost:8088/proxy/application_1627681217439_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1627681217439_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-07-30 17:26:09,366 Stage-1 map = 0%,  reduce = 0%
2021-07-30 17:26:21,826 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.15 sec
2021-07-30 17:26:35,436 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.92 sec
MapReduce Total cumulative CPU time: 5 seconds 920 msec
Ended Job = job_1627681217439_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.92 sec   HDFS Read: 9527 HDFS Write: 223 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 920 msec
OK
12918717        B00KQUNNZ8      2015-08-31      5       17      positive
1662075 B00XPWXYDK      2015-08-31      5       14      positive
26599182        B00JLKI69W      2015-08-31      3       13      positive
12191231        B00BMKL5WY      2015-08-31      5       11      positive
32910511        B00V5DM3RE      2015-08-31      5       8       positive
Time taken: 46.876 seconds, Fetched: 5 row(s)
hive> 
```

In the above screenshot, we can see the results of top 5 good reviews that are received for the amazon products.

3. From the output of 2.c, we can infer that out of total reviews we have got only two negative reviews based on total score. Basically total score represents how negative or positive the total review is about. The drawback of this approach is whatever dictionary we tend to use has a chance of missing words that are present in the reviews. The words that are not present in the dictionary become useless and meaningless in the analysis. In the given dictionary file, we can see that same word is present with different verbs such as disregard, disregarded, disregarding, performance or accuracy might increase when for a single word for all variations instead of multiple words for a single variation. Though we add certain words based on our research and give a proper score to it, it works to certain extent but if the data is huge, we cannot read the data manually and assign a score to the words individually. So the best approach would be to have an individual dictionary type for different industries. For a retail industry like amazon, flipkart, alibaba, there should be certain different type of dictionary that helps to maximize the potential of the sentimental analysis. In addition to this we have got many speech processing techniques for analysis. The text can be converted to speech and processed for sentimental analysis to yield a better result.