Alexander Ashmore - Assignment 4

Labs: Gaining Insight with Sentiment Analysis, Data Transformation with Hive

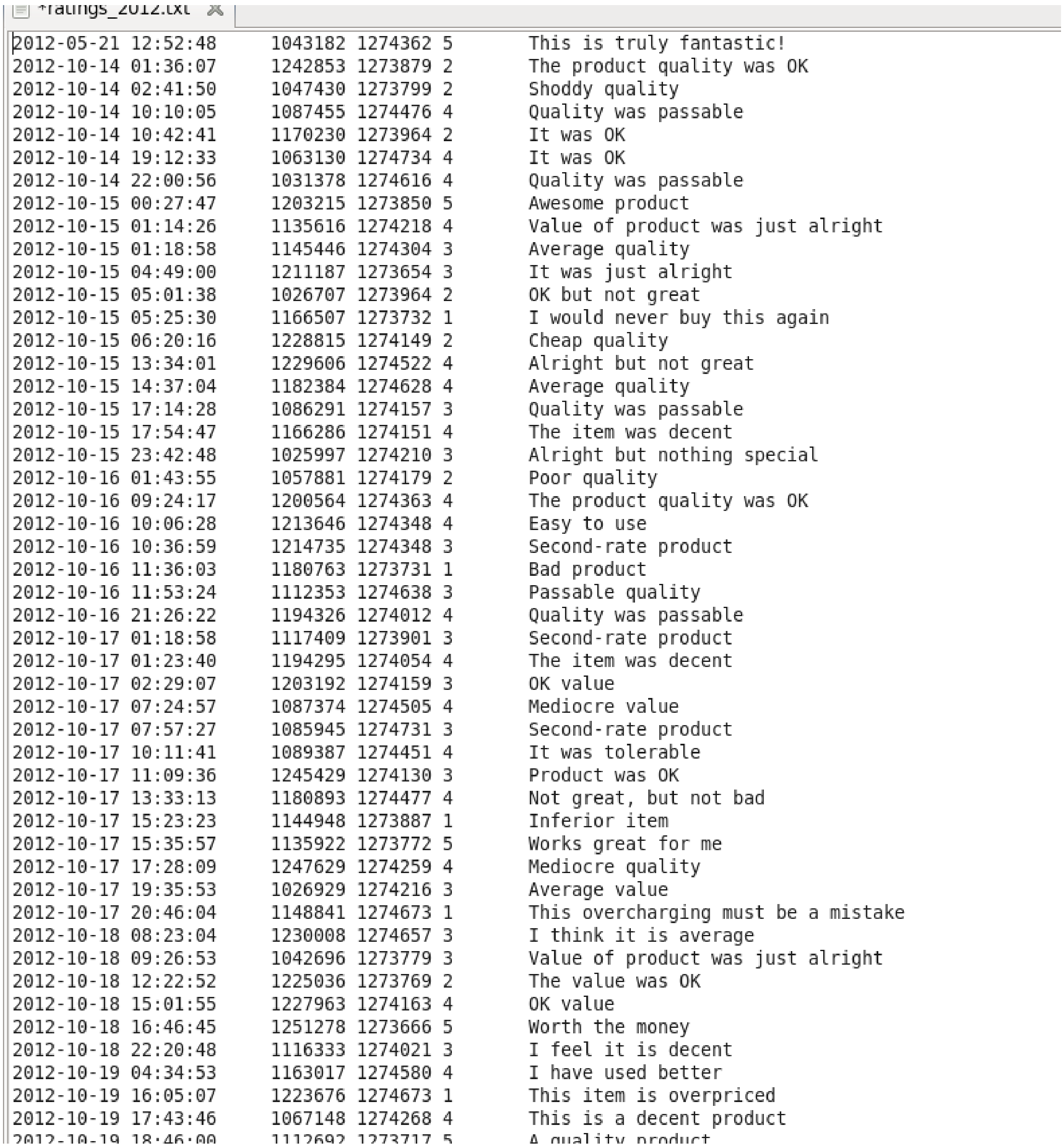
Lab: Gaining Insight with Sentiment Analysis

Project introduction, explain what is required in the project.

We are to use Hive’s text processing features to analyze the customer’s comments and product ratings. Doing so will uncover potential issues like overcharged items and then we can potentially provide solutions. This is based off of customer ratings and feedbacks being in free form text so we must sort through these by text processing.

The format of the data in the original data input file.

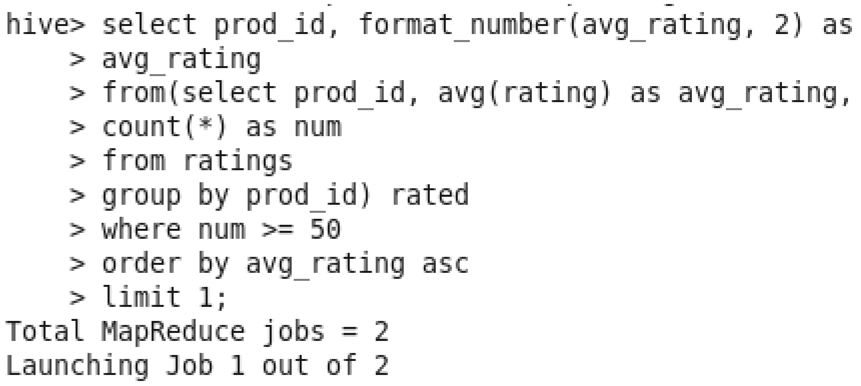
The format of the original data input file is free form text, along with other product identification information.



```
2012-05-21 12:52:48     1043182 1274362 5       This is truly fantastic!
2012-10-14 01:36:07     1242853 1273879 2       The product quality was OK
2012-10-14 02:41:50     1047430 1273799 2       Shoddy quality
2012-10-14 10:10:05     1087455 1274476 4       Quality was passable
2012-10-14 10:42:41     1170230 1273964 2       It was OK
2012-10-14 19:12:33     1063130 1274734 4       It was OK
2012-10-14 22:00:56     1031378 1274616 4       Quality was passable
2012-10-15 00:27:47     1203215 1273850 5       Awesome product
2012-10-15 01:14:26     1135616 1274218 4       Value of product was just alright
2012-10-15 01:18:58     1145446 1274304 3       Average quality
2012-10-15 04:49:00     1211187 1273654 3       It was just alright
2012-10-15 05:01:38     1026707 1273964 2       OK but not great
2012-10-15 05:25:30     1166507 1273732 1       I would never buy this again
2012-10-15 06:20:16     1228815 1274149 2       Cheap quality
2012-10-15 13:34:01     1229606 1274522 4       Alright but not great
2012-10-15 14:37:04     1182384 1274628 4       Average quality
2012-10-15 17:14:28     1086291 1274157 3       Quality was passable
2012-10-15 17:54:47     1166286 1274151 4       The item was decent
2012-10-15 23:42:48     1025997 1274210 3       Alright but nothing special
2012-10-16 01:43:55     1057881 1274179 2       Poor quality
2012-10-16 09:24:17     1200564 1274363 4       The product quality was OK
2012-10-16 10:06:28     1213646 1274348 4       Easy to use
2012-10-16 10:36:59     1214735 1274348 3       Second-rate product
2012-10-16 11:36:03     1180763 1273731 1       Bad product
2012-10-16 11:53:24     1112353 1274638 3       Passable quality
2012-10-16 21:26:22     1194326 1274012 4       Quality was passable
2012-10-17 01:18:58     1117409 1273901 3       Second-rate product
2012-10-17 01:23:40     1194295 1274054 4       The item was decent
2012-10-17 02:29:07     1203192 1274159 3       OK value
2012-10-17 07:24:57     1087374 1274505 4       Mediocre value
2012-10-17 07:57:27     1085945 1274731 3       Second-rate product
2012-10-17 10:11:41     1089387 1274451 4       It was tolerable
2012-10-17 11:09:36     1245429 1274130 3       Product was OK
2012-10-17 13:33:13     1180893 1274477 4       Not great, but not bad
2012-10-17 15:23:23     1144948 1273887 1       Inferior item
2012-10-17 15:35:57     1135922 1273772 5       Works great for me
2012-10-17 17:28:09     1247629 1274259 4       Mediocre quality
2012-10-17 19:35:53     1026929 1274216 3       Average value
2012-10-17 20:46:04     1148841 1274673 1       This overcharging must be a mistake
2012-10-18 08:23:04     1230008 1274657 3       I think it is average
2012-10-18 09:26:53     1042696 1273779 3       Value of product was just alright
2012-10-18 12:22:52     1225036 1273769 2       The value was OK
2012-10-18 15:01:55     1227963 1274163 4       OK value
2012-10-18 16:46:45     1251278 1273666 5       Worth the money
2012-10-18 22:20:48     1116333 1274021 3       I feel it is decent
2012-10-19 04:34:53     1163017 1274580 4       I have used better
2012-10-19 16:05:07     1223676 1274673 1       This item is overpriced
2012-10-19 17:43:46     1067148 1274268 4       This is a decent product
2012-10-19 18:46:00     1112692 1273717 5       A quality product
```

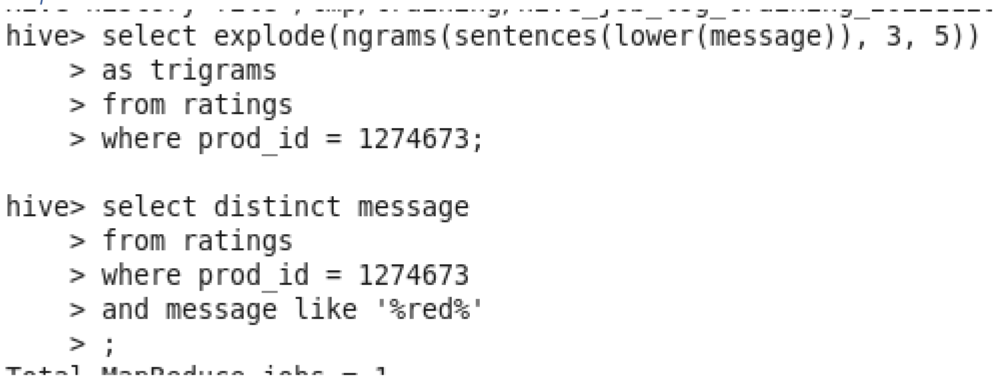
Data processing procedure.

Step 1



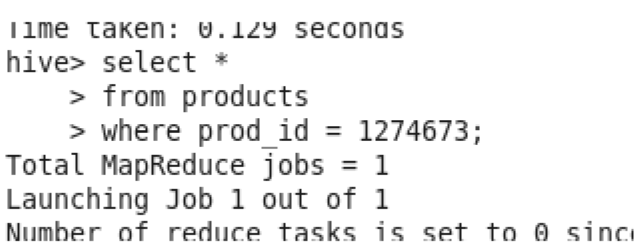
```
hive> select prod_id, format_number(avg_rating, 2) as
    > avg_rating
    > from(select prod_id, avg(rating) as avg_rating,
    > count(*) as num
    > from ratings
    > group by prod_id) rated
    > where num >= 50
    > order by avg_rating asc
    > limit 1;
Total MapReduce jobs = 2
Launching Job 1 out of 2
```

Step 2

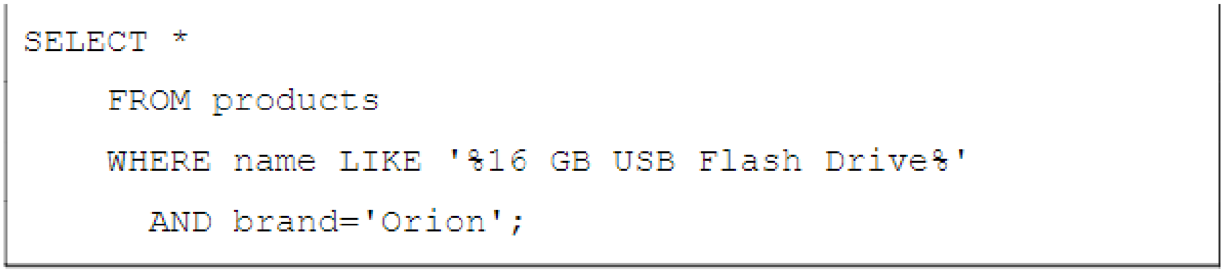


```
hive> select explode(ngrams(sentences(lower(message)), 3, 5))
    > as trigrams
    > from ratings
    > where prod_id = 1274673;

hive> select distinct message
    > from ratings
    > where prod_id = 1274673
    > and message like '%red%'
    > ;
Total MapReduce jobs = 1
```



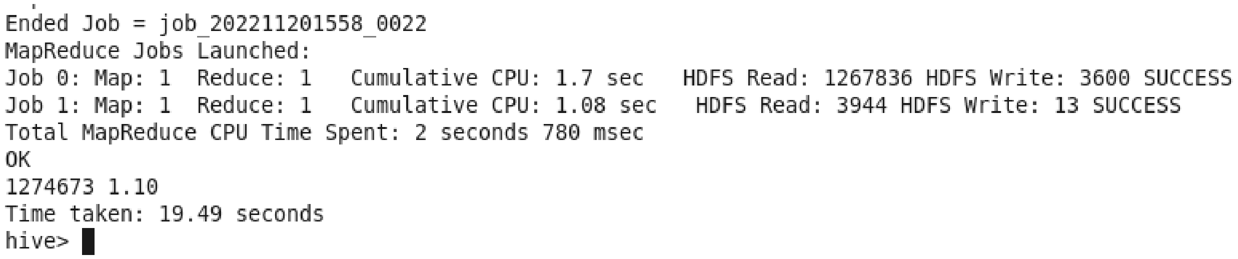
```
Time taken: 0.129 seconds
hive> select *
    > from products
    > where prod_id = 1274673;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 sinc
```



```
SELECT *
    FROM products
    WHERE name LIKE '%16 GB USB Flash Drive%'
      AND brand='Orion';
```

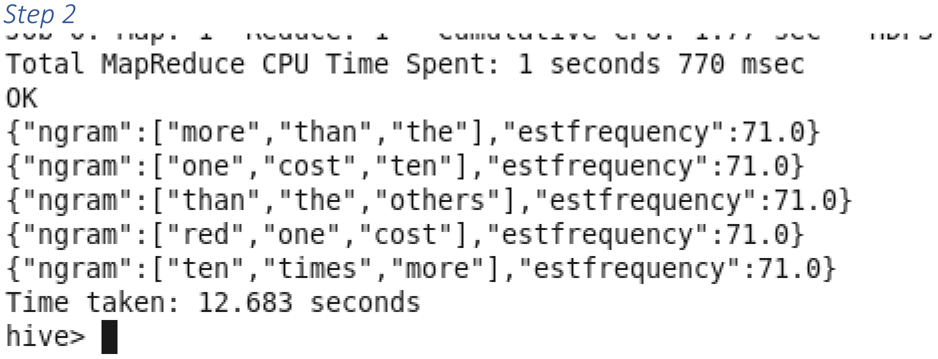
Data output and results.

Step 1

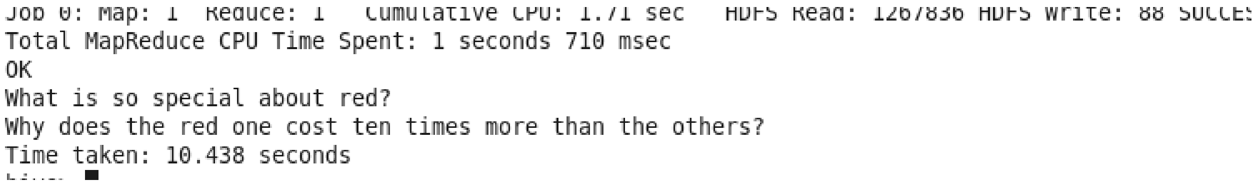


```
Ended Job = job_202211201558_0022
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 1.7 sec   HDFS Read: 1267836 HDFS Write: 3600 SUCCESS
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 1.08 sec   HDFS Read: 3944 HDFS Write: 13 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 780 msec
OK
1274673 1.10
Time taken: 19.49 seconds
hive> 
```

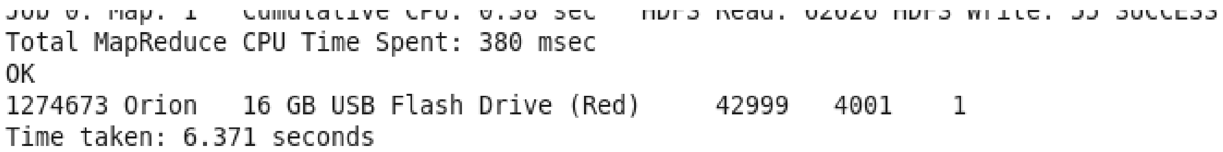
Step 2



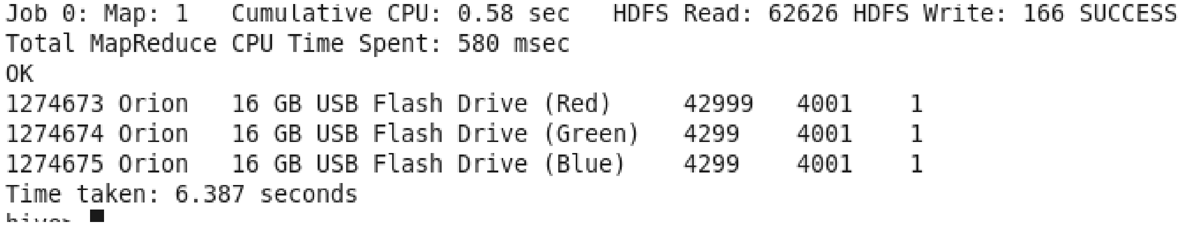
```
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 1.77 sec   HDFS
Total MapReduce CPU Time Spent: 1 seconds 770 msec
OK
{"ngram":["more","than","the"],"estfrequency":71.0}
{"ngram":["one","cost","ten"],"estfrequency":71.0}
{"ngram":["than","the","others"],"estfrequency":71.0}
{"ngram":["red","one","cost"],"estfrequency":71.0}
{"ngram":["ten","times","more"],"estfrequency":71.0}
Time taken: 12.683 seconds
hive> 
```



```
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 1.71 sec   HDFS Read: 1267836 HDFS Write: 88 SUCCES
Total MapReduce CPU Time Spent: 1 seconds 710 msec
OK
What is so special about red?
Why does the red one cost ten times more than the others?
Time taken: 10.438 seconds
hive> 
```



```
Job 0: Map: 1   Cumulative CPU: 0.38 sec   HDFS Read: 62626 HDFS Write: 55 SUCCESS
Total MapReduce CPU Time Spent: 380 msec
OK
1274673 Orion   16 GB USB Flash Drive (Red)     42999   4001    1
Time taken: 6.371 seconds
```



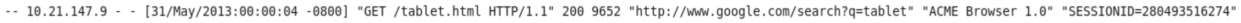
```
Job 0: Map: 1   Cumulative CPU: 0.58 sec   HDFS Read: 62626 HDFS Write: 166 SUCCESS
Total MapReduce CPU Time Spent: 580 msec
OK
1274673 Orion   16 GB USB Flash Drive (Red)     42999   4001    1
1274674 Orion   16 GB USB Flash Drive (Green)   4299    4001    1
1274675 Orion   16 GB USB Flash Drive (Blue)    4299    4001    1
Time taken: 6.387 seconds
hive> 
```

Lab: Data Transformation with Hive

Project introduction, explain what is required in the project.

We are to create several tables to analyze a problem of customers abandoning their shopping carts before completing the checkout process. This is done by creating and populating tables with log data from Dualcore's web server. Tables are created based on criteria that uses regex to sort data.

The format of the data in the original data input file.

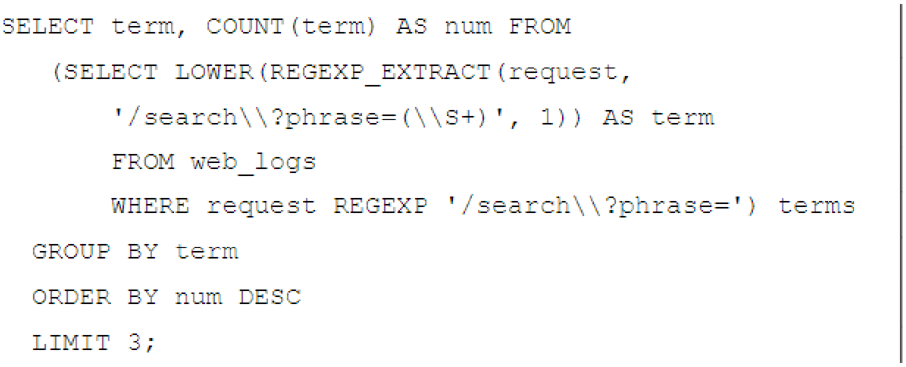


```
-- 10.21.147.9 - - [31/May/2013:00:00:04 -0800] "GET /tablet.html HTTP/1.1" 200 9652 "http://www.google.com/search?q=tablet" "ACME Browser 1.0" "SESSIONID=280493516274"
```

Above is an example of the original data input file formatting.

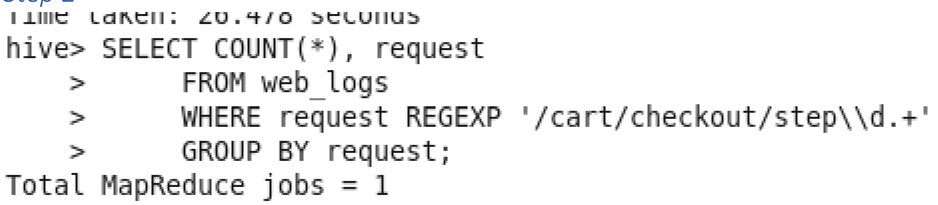
Data processing procedure.

Step 1

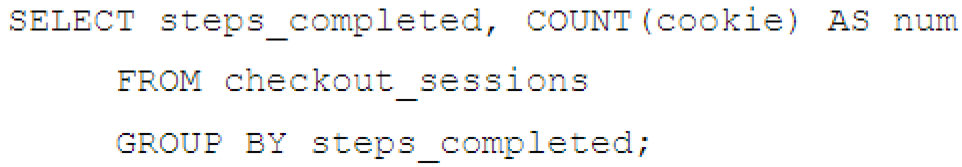


```
SELECT term, COUNT(term) AS num FROM
   (SELECT LOWER(REGEXP_EXTRACT(request,
       '/search\\?phrase=(\\S+)', 1)) AS term
       FROM web_logs
       WHERE request REGEXP '/search\\?phrase=') terms
  GROUP BY term
  ORDER BY num DESC
  LIMIT 3;
```

Step 2

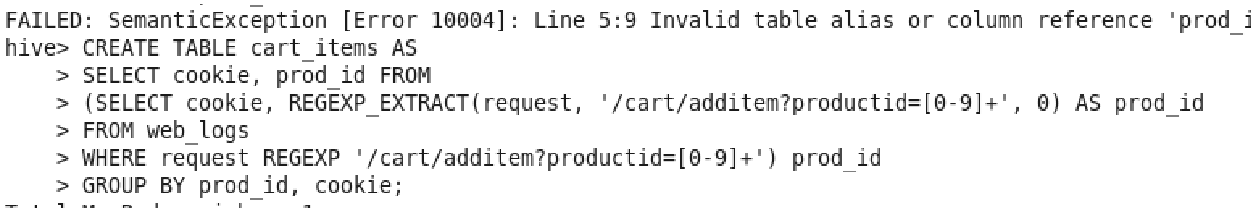


```
Time taken: 26.478 seconds
hive> SELECT COUNT(*), request
    >      FROM web_logs
    >      WHERE request REGEXP '/cart/checkout/step\\d.+'
    >      GROUP BY request;
Total MapReduce jobs = 1
```



```
SELECT steps_completed, COUNT(cookie) AS num
      FROM checkout_sessions
      GROUP BY steps_completed;
```

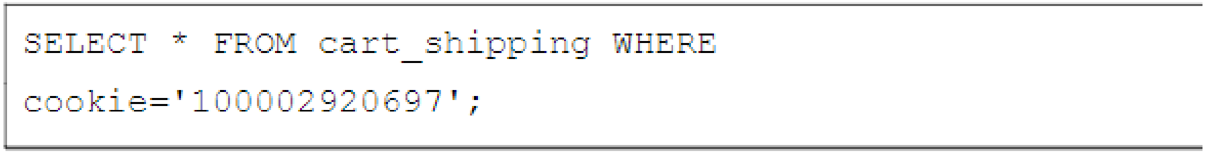
Step 4



```
FAILED: SemanticException [Error 10004]: Line 5:9 Invalid table alias or column reference 'prod_i
hive> CREATE TABLE cart_items AS
    > SELECT cookie, prod_id FROM
    > (SELECT cookie, REGEXP_EXTRACT(request, '/cart/additem?productid=[0-9]+', 0) AS prod_id
    > FROM web_logs
    > WHERE request REGEXP '/cart/additem?productid=[0-9]+') prod_id
    > GROUP BY prod_id, cookie;
Total MapReduce jobs = 1
```

This is incorrect. Was not able to figure out implementation.

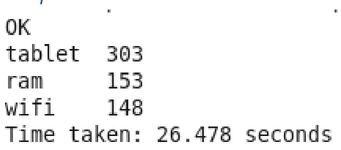
Step 6



```
SELECT * FROM cart_shipping WHERE
cookie='100002920697';
```

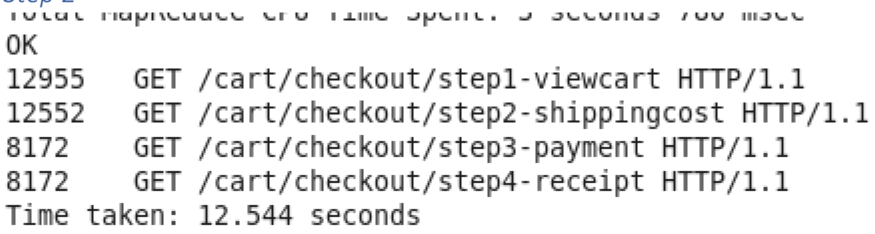
Data output and results.

Step 1

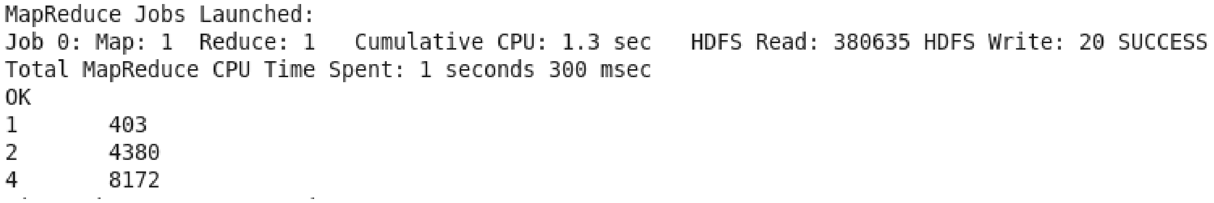


```
OK
tablet  303
ram     153
wifi    148
Time taken: 26.478 seconds
```

Step 2

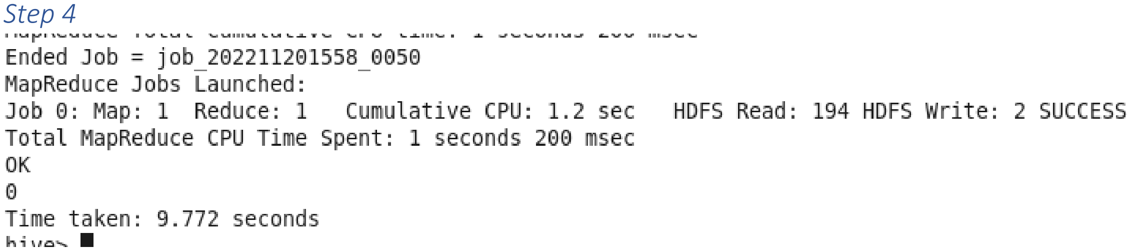


```
Total MapReduce CPU Time Spent: 5 seconds 780 msec
OK
12955   GET /cart/checkout/step1-viewcart HTTP/1.1
12552   GET /cart/checkout/step2-shippingcost HTTP/1.1
8172    GET /cart/checkout/step3-payment HTTP/1.1
8172    GET /cart/checkout/step4-receipt HTTP/1.1
Time taken: 12.544 seconds
```



```
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 1.3 sec   HDFS Read: 380635 HDFS Write: 20 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 300 msec
OK
1       403
2       4380
4       8172
```

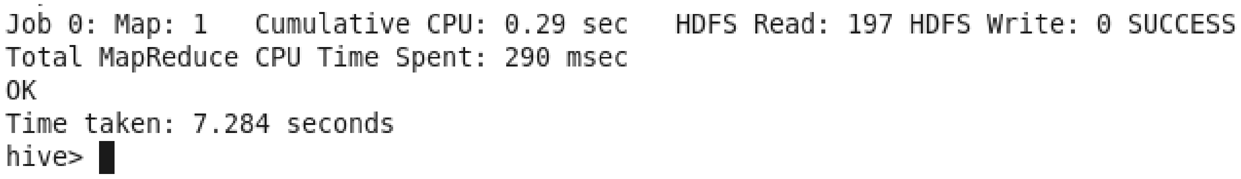
Step 4



```
MapReduce Total cumulative CPU time: 1 seconds 200 msec
Ended Job = job_202211201558_0050
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 1.2 sec   HDFS Read: 194 HDFS Write: 2 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 200 msec
OK
0
Time taken: 9.772 seconds
hive>
```

Output here is incorrect.

Step 6



```
Job 0: Map: 1   Cumulative CPU: 0.29 sec   HDFS Read: 197 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 290 msec
OK
Time taken: 7.284 seconds
hive>
```

Output here is incorrect because it is base cart_orders which is based on implementation of cart_items. I was not able to correctly implement cart_items so as it goes down the list it will continue to be wrong.