APACHE PIG

Amazon Daily Review Count

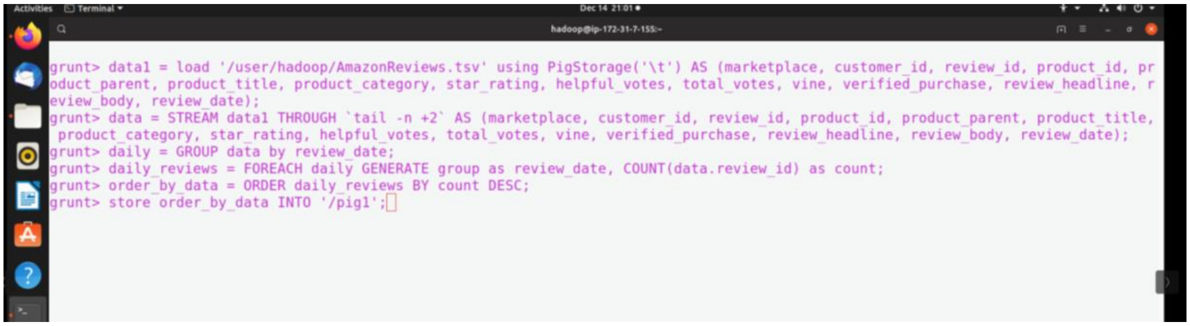
data1 = load '/user/hadoop/AmazonReviews.tsv' using PigStorage('\t') AS (marketplace, customer_id, review_id, product_id, product_parent, product_title, product_category, star_rating, helpful_votes, total_votes, vine, verified_purchase, review_headline, review_body, review_date);

data = STREAM data1 THROUGH `tail -n +2` AS (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);

daily = GROUP data by review_date;

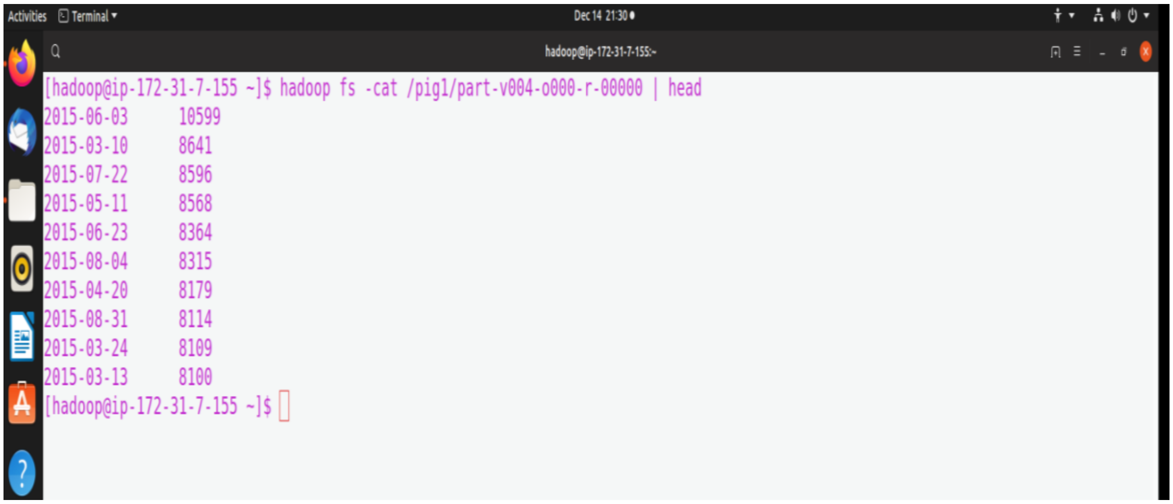
daily_reviews = FOREACH daily GENERATE group as review_date, COUNT(data.review_id) as count;
order_by_data = ORDER daily_reviews BY count DESC;

store order_by_data INTO '/pig1';

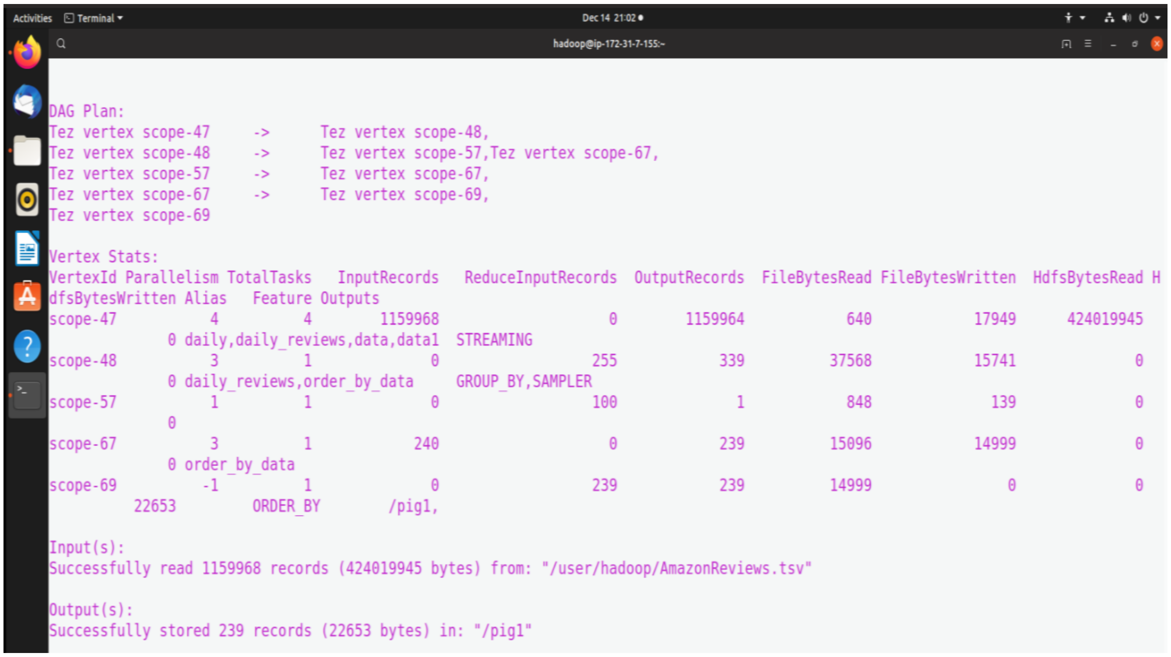


```
grunt> data1 = load '/user/hadoop/AmazonReviews.tsv' using PigStorage('\t') AS (marketplace, customer_id, review_id, product_id, pr
oduct_parent, product_title, product_category, star_rating, helpful_votes, total_votes, vine, verified_purchase, review_headline, r
eview_body, review_date);
grunt> data = STREAM data1 THROUGH `tail -n +2` AS (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);
grunt> daily = GROUP data by review_date;
grunt> daily_reviews = FOREACH daily GENERATE group as review_date, COUNT(data.review_id) as count;
grunt> order_by_data = ORDER daily_reviews BY count DESC;
grunt> store order_by_data INTO '/pig1';
```

OUTPUT:



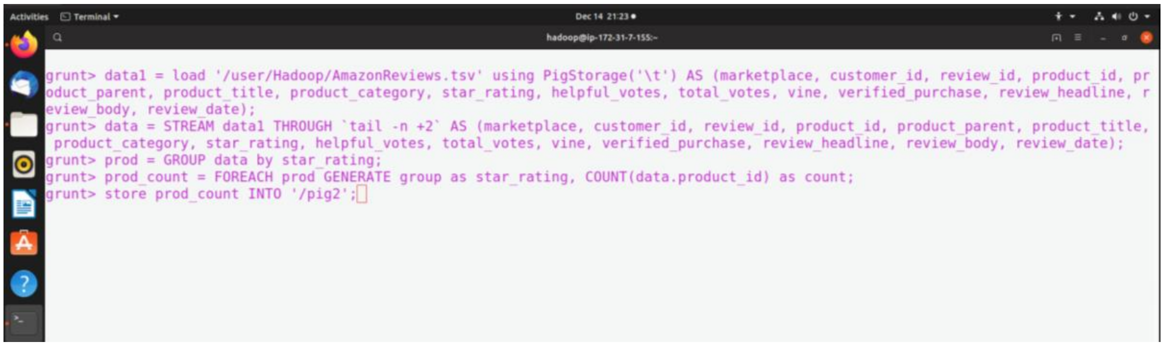
```
[hadoop@ip-172-31-7-155 ~]$ hadoop fs -cat /pig1/part-v004-o000-r-00000 | head
2015-06-03      10599
2015-03-10      8641
2015-07-22      8596
2015-05-11      8568
2015-06-23      8364
2015-08-04      8315
2015-04-20      8179
2015-08-31      8114
2015-03-24      8109
2015-03-13      8100
[hadoop@ip-172-31-7-155 ~]$ 
```



```
DAG Plan:
Tez vertex scope-47     ->      Tez vertex scope-48,
Tez vertex scope-48     ->      Tez vertex scope-57,Tez vertex scope-67,
Tez vertex scope-57     ->      Tez vertex scope-67,
Tez vertex scope-67     ->      Tez vertex scope-69,
Tez vertex scope-69

Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  HdfsBytesRead H
dfsBytesWritten Alias   Feature Outputs
scope-47           4          4        1159968                    0        1159964            640            17949      424019945
              0 daily,daily_reviews,data,data1  STREAMING
scope-48           3          1              0                  255            339          37568            15741              0
              0 daily_reviews,order_by_data     GROUP_BY,SAMPLER
scope-57           1          1              0                  100              1            848              139              0
              0
scope-67           3          1            240                    0            239          15096            14999              0
              0 order_by_data
scope-69          -1          1              0                  239            239          14999                0              0
          22653         ORDER_BY        /pig1,

Input(s):
Successfully read 1159968 records (424019945 bytes) from: "/user/hadoop/AmazonReviews.tsv"

Output(s):
Successfully stored 239 records (22653 bytes) in: "/pig1"
```

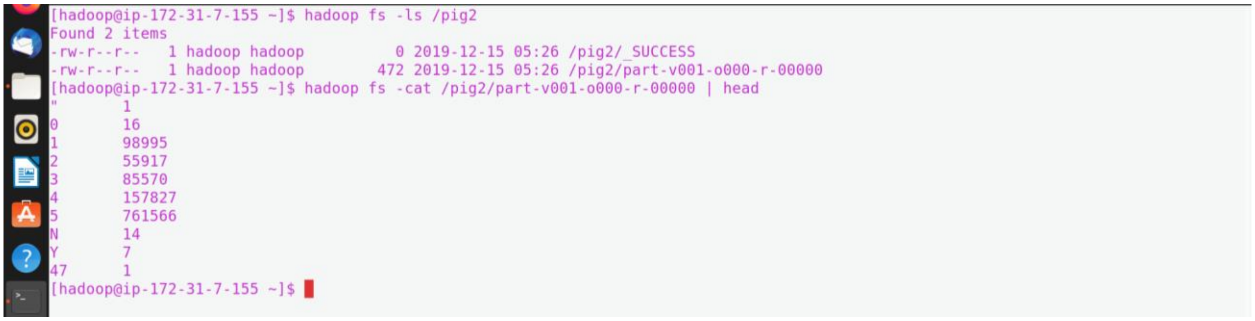
Amazon Total review count per product

```
data1 = load '/user/hadoop/AmazonReviews.tsv' using PigStorage('\t') AS (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);

data = STREAM data1 THROUGH `tail -n +2` AS (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);

prod = GROUP data by star_rating;

prod_count = FOREACH prod GENERATE group as star_rating, COUNT(data.product_id) as
count;

store prod_count INTO '/pig2';
```

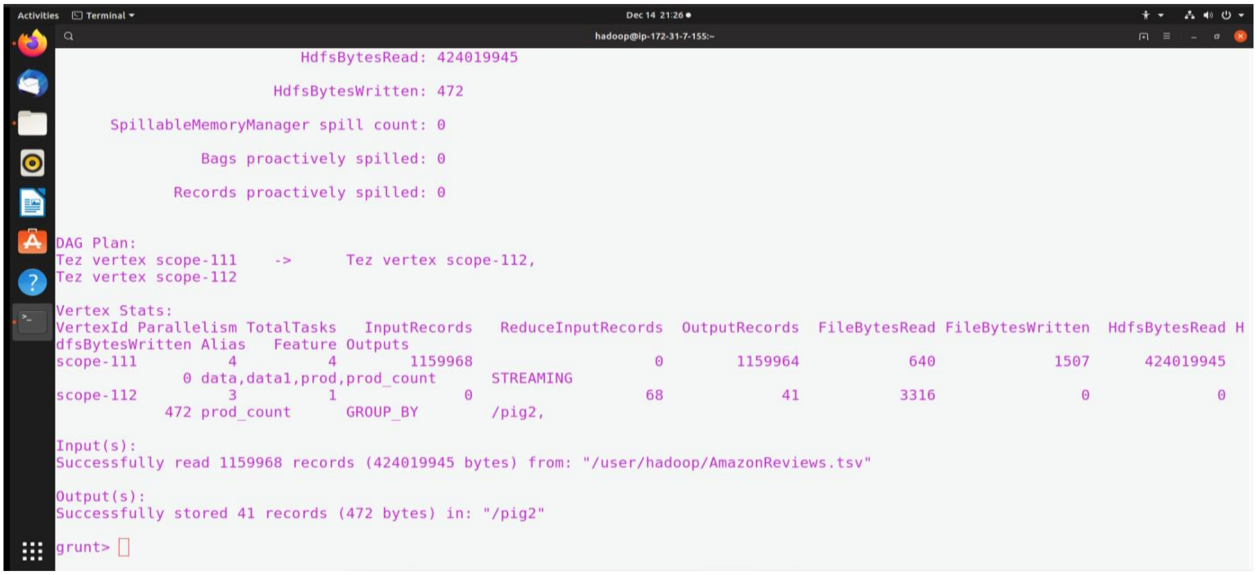


```
grunt> data1 = load '/user/Hadoop/AmazonReviews.tsv' using PigStorage('\t') AS (marketplace, customer_id, review_id, product_id, pr
oduct_parent, product_title, product_category, star_rating, helpful_votes, total_votes, vine, verified_purchase, review_headline, r
eview_body, review_date);
grunt> data = STREAM data1 THROUGH `tail -n +2` AS (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);
grunt> prod = GROUP data by star_rating;
grunt> prod_count = FOREACH prod GENERATE group as star_rating, COUNT(data.product_id) as count;
grunt> store prod_count INTO '/pig2';
```

OUTPUT :



```
[hadoop@ip-172-31-7-155 ~]$ hadoop fs -ls /pig2
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2019-12-15 05:26 /pig2/_SUCCESS
-rw-r--r--   1 hadoop hadoop        472 2019-12-15 05:26 /pig2/part-v001-o000-r-00000
[hadoop@ip-172-31-7-155 ~]$ hadoop fs -cat /pig2/part-v001-o000-r-00000 | head
"       1
0       16
1       98995
2       55917
3       85570
4       157827
5       761566
N       14
Y       7
47      1
[hadoop@ip-172-31-7-155 ~]$
```



```
                           HdfsBytesRead: 424019945

                        HdfsBytesWritten: 472

      SpillableMemoryManager spill count: 0

                Bags proactively spilled: 0

             Records proactively spilled: 0


DAG Plan:
Tez vertex scope-111    ->      Tez vertex scope-112,
Tez vertex scope-112

Vertex Stats:
VertexId Parallelism TotalTasks   InputRecords   ReduceInputRecords  OutputRecords  FileBytesRead FileBytesWritten  HdfsBytesRead H
dfsBytesWritten Alias   Feature Outputs
scope-111          4          4        1159968                    0        1159964            640             1507      424019945
              0 data,data1,prod,prod_count      STREAMING
scope-112          3          1              0                   68             41           3316                0              0
            472 prod_count      GROUP_BY        /pig2,

Input(s):
Successfully read 1159968 records (424019945 bytes) from: "/user/hadoop/AmazonReviews.tsv"

Output(s):
Successfully stored 41 records (472 bytes) in: "/pig2"

grunt> 
```