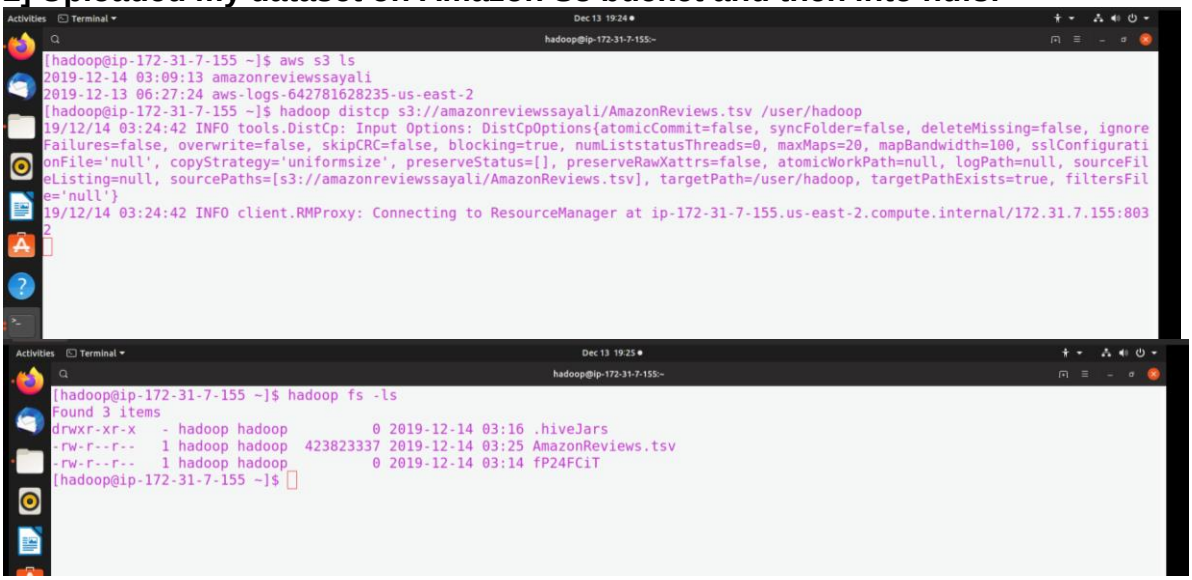
Apache Hive

Apache Hive provides a SQL-like interface to data stored in HDP(Hortonworks Data Platform) whose component is Hive.

Steps:

1] I have also created a Amazon EMR Hadoop cluster and I am going to use HIVE on it

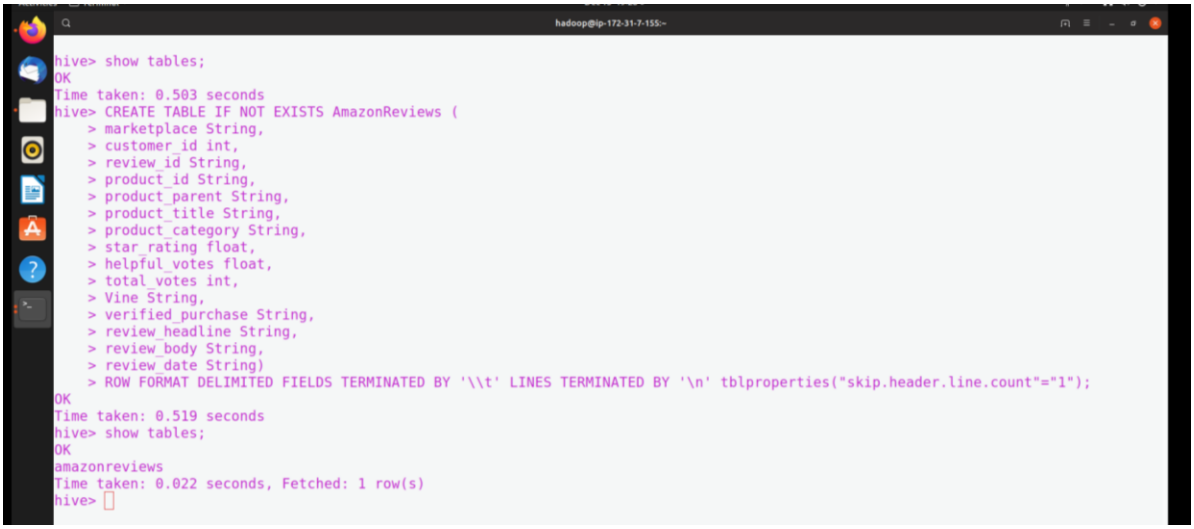
2] Uploaded my dataset on Amazon S3 bucket and then into hdfs:



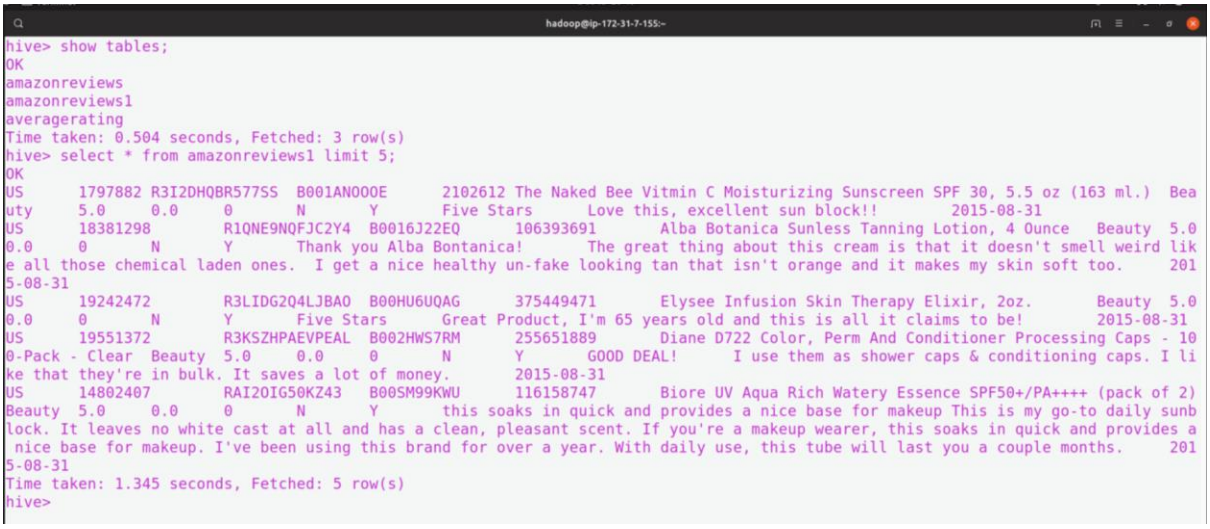
Amazon Review Data Analysis HIVE:

Create a table into Hive:

```
CREATE TABLE IF NOT EXISTS AmazonReviews (
marketplace String,
customer_id int,
review_id String,
product_id String,
product_parent String,
product_title String,
product_category String,
star_rating float,
helpful_votes float,
total_votes int,
Vine String,
verified_purchase String,
review_headline String,
review_body String,
review_date String)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LINES TERMINATED BY
'\n' tblproperties("skip.header.line.count"="1");
```



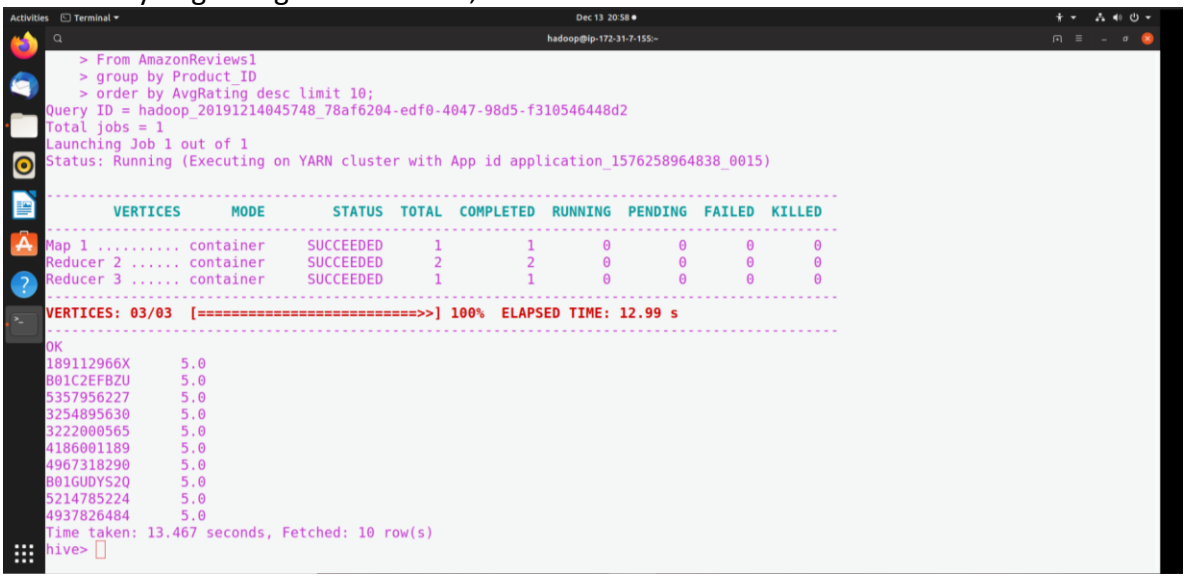
Load the dataset into hive table:



Top 10 Products based on average rating

I have calculated average rating for each product. I have demonstrated similar script on map reduce however hive gave the fastest performance.

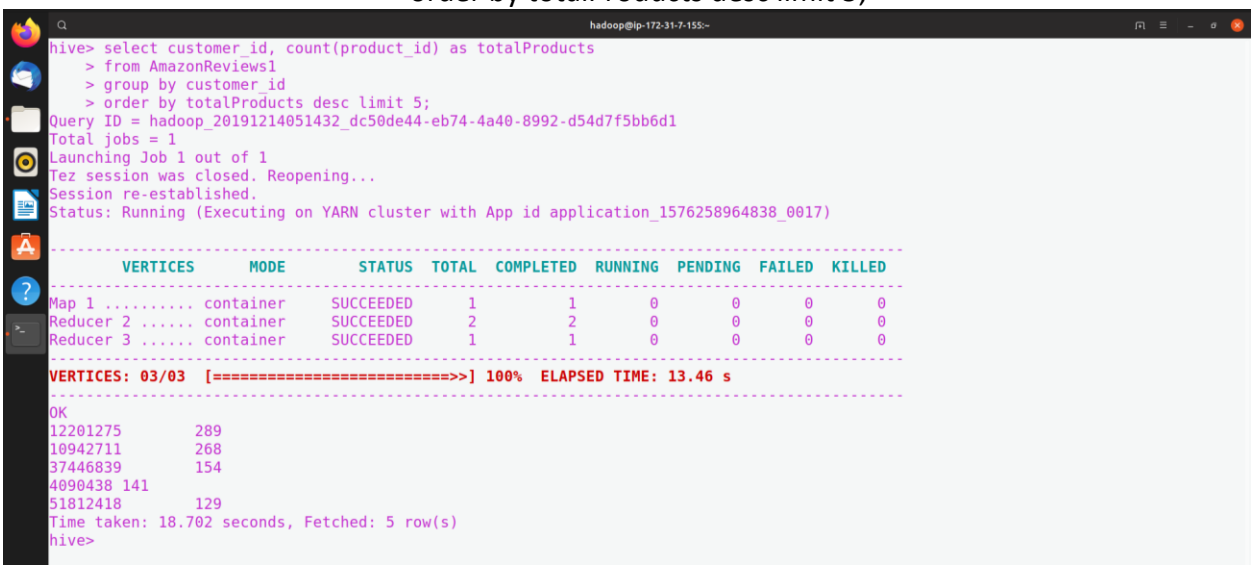
>Select product_id, avg(star_rating) as AvgRating
>From AmazonReviews1
>group by Product_ID
>order by AvgRating DESC limit 10;



Most valuable Customer

I have counted no of products ordered by each customer and Customer who ordered highest number of products is the most valuable customer.

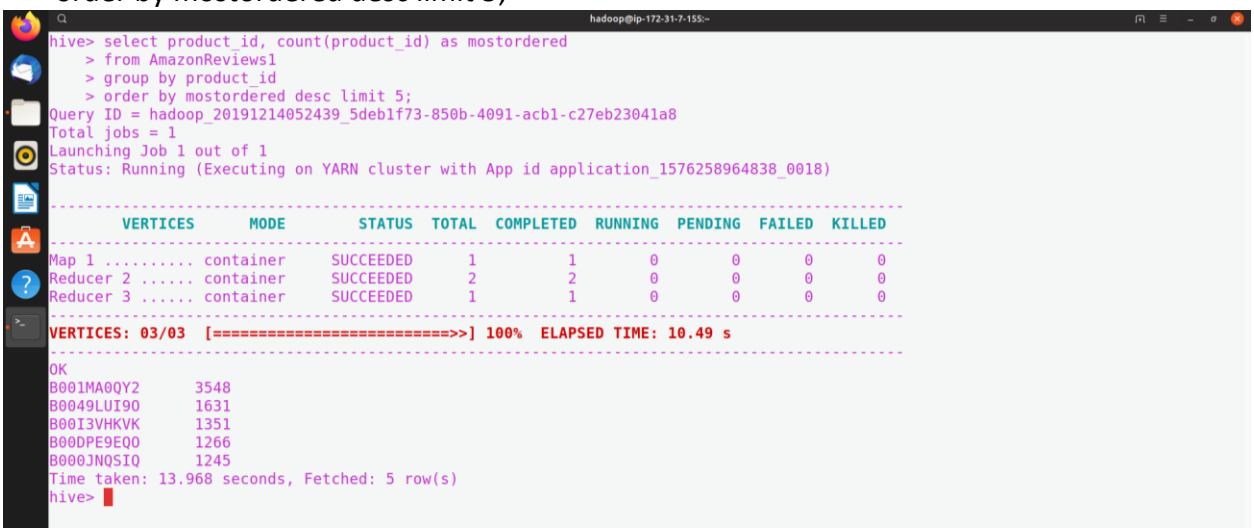
```
>select customer_id, count(product_id) as totalProducts
> from AmazonReviews1
> group by customer_id
                                        > order by totalProducts desc limit 5;
```



Most Popular Product

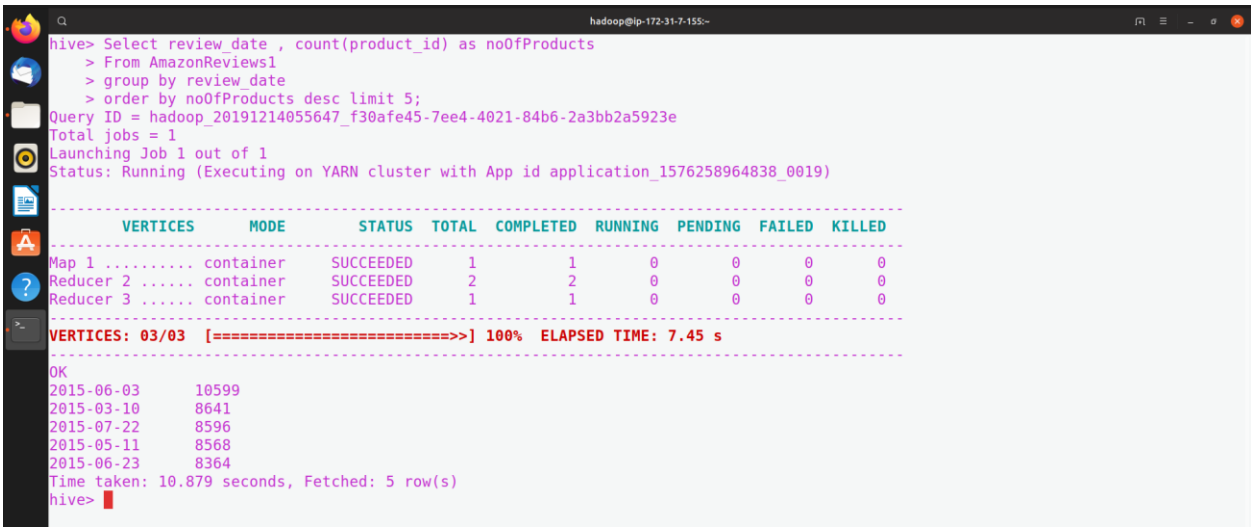
I have counted how many times each product was ordered.

```
>select product_id, count(product_id) as mostordered
> from AmazonReviews1
> group by product_id
> order by mostordered desc limit 5;
```



Number of Products Per day

> Select review_date , count(product_id) as noOfProducts
> From AmazonReviews1
> group by review_date
> order by noOfProducts desc limit 5;



Number of Products Per Rating

>select star_rating, count(product_id) as noOfProducts
> From AmazonReviews1
> group by star_rating
> order by noOfProducts desc limit 5;