Data Analysis on Amazon Customer Reviews Dataset

Technologies:

Hadoop Map reduce

Apache Hive

Apache Mahout

Sayali Walke

NUID: 001417763

Summary

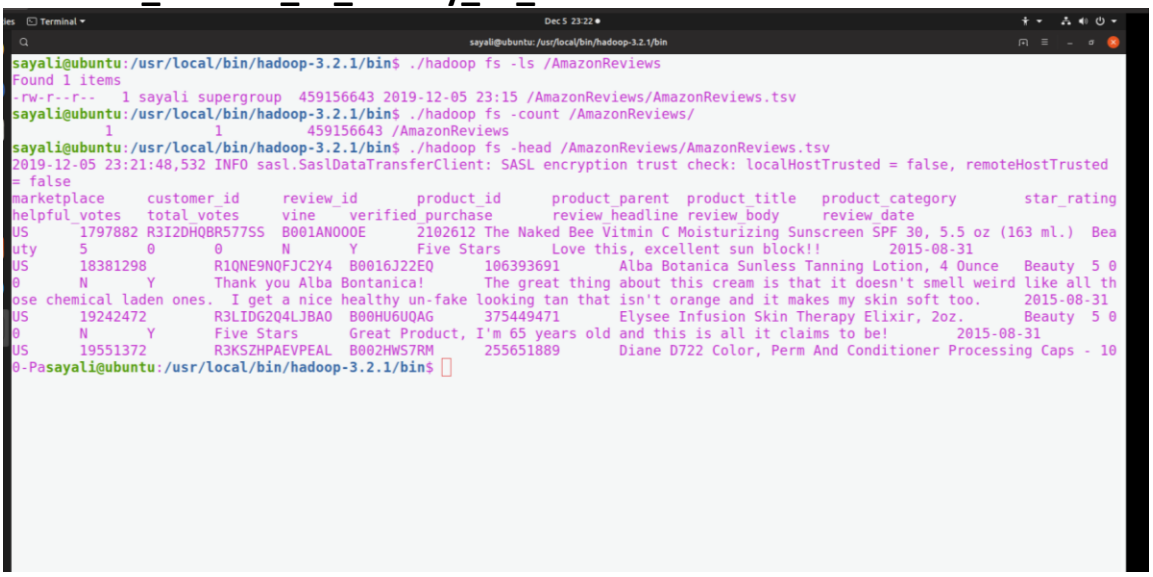
About the data:

The dataset contains the customer review text with accompanying metadata, consisting of three major components:

1. A collection of reviews written in the Amazon.com marketplace and associated metadata from 1995 until 2015. This is intended to facilitate study into the properties (and the evolution) of customer reviews potentially including how people evaluate and express their experiences with respect to products at scale. (130M+ customer reviews)
2. A collection of reviews about products in multiple languages from different Amazon marketplaces, intended to facilitate analysis of customers' perception of the same products and wider consumer preferences across languages and countries. (200K+ customer reviews in 5 countries)

Dataset used for analysis:

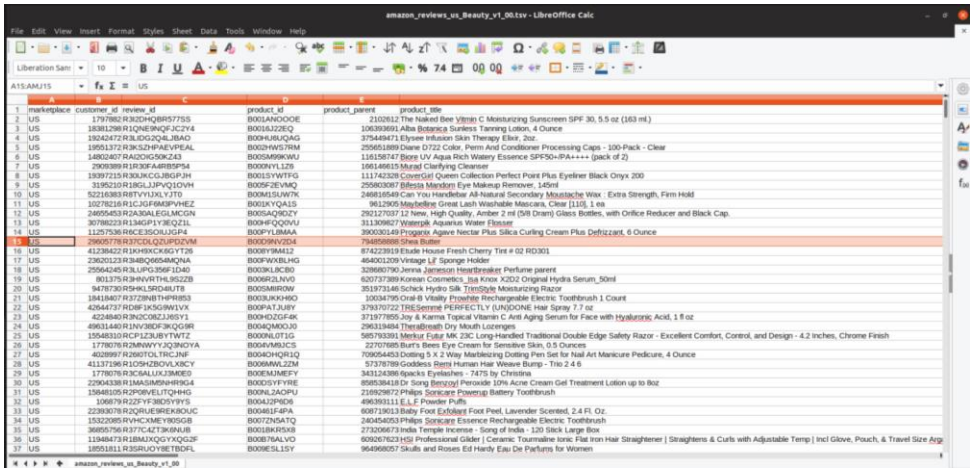
The Amazon Customer Reviews Dataset is a large dataset with size > 20GB.
However, for this analysis, we've used a subset of this dataset named
"amazon_reviews_us_Beauty_v1_00.tsv" Size of this dataset is around 500MB.

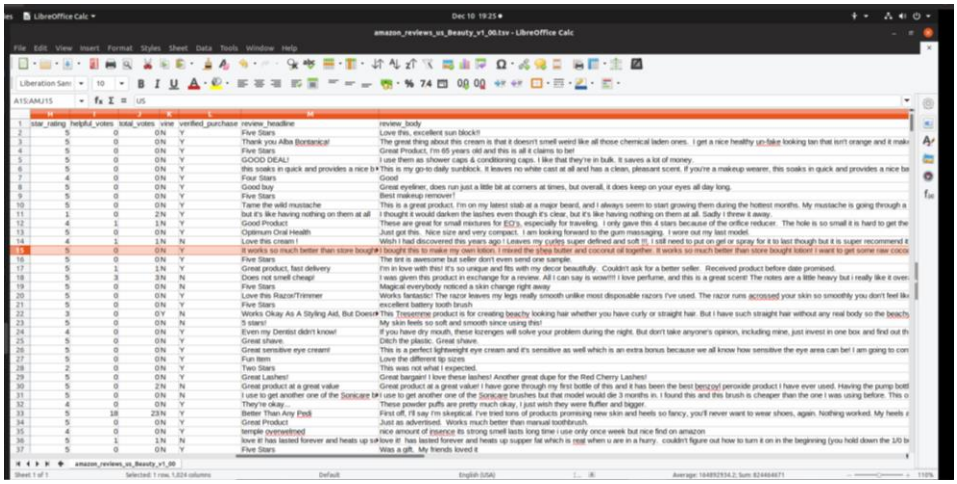


The link for the dataset can be found here
https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Beauty_v1_00.tsv.gz

Here’s the detailed description of dataset and its contents.
marketplace: 2 letter country code of the marketplace where the review was written.
customer_id: Random identifier that can be used to aggregate reviews written by a single author.
review_id: The unique ID of the review.
product_id: The unique Product ID the review pertains to. In the multilingual dataset the reviews for the same product in different countries can be grouped by the same product_id.
product_parent: Random identifier that can be used to aggregate reviews for the same product.
product_title: Title of the product.
product_category: Broad product category that can be used to group reviews
(also used to group the dataset into coherent parts).
star_rating: The 1-5 star rating of the review.
helpful_votes: Number of helpful votes.
total_votes: Number of total votes the review received.
Vine: Review was written as part of the Vine program.
verified_purchase: The review is on a verified purchase.
review_headline: The title of the review.
review_body: The review text.
review_date: The date the review was written.

Preview of Dataset





Hadoop

I have created a single node Hadoop cluster on virtual machine and carried out following data analysis using various Map-reduce Algorithms.

1. Top 100 Products based on average of reviews: (Filtering Pattern)
2. Average chaining and Sorting Of reviews: (Chaining and Sorting)
3. No of reviews per product: (Numerical Summarization)
4. Inner join on Average of reviews and no of reviews for each product: (Inner Join)
5. Customer list for each product: (Inverted Index)
6. Created 5 bins for 1,2,3,4,5 ratings: Binning (Organization Pattern)
7. Partitioned the data into different files for each day in 2015-08: Partitioning (Organization Pattern)
8. Distinct Reviews Counter: (Numerical Summarization Pattern)
9. Percentage of Helpful votes: (Numerical Summarization Pattern)

Apache HIVE

1. Top 10 Products based on Average ratings
2. Most Valuable Customer based on number of products bought
3. Most popular product based on number times product bought
4. Number of products bought per day
5. Number of products per ratings

Apache Pig

1. Number of reviews given per day
2. Number reviews given per product

MAHOUT

Created a recommender system using Mahout. It serves the functionality of recommending similar products based on the similar items bought by other customers.(People who bought this….also bought this…..)