

# 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:

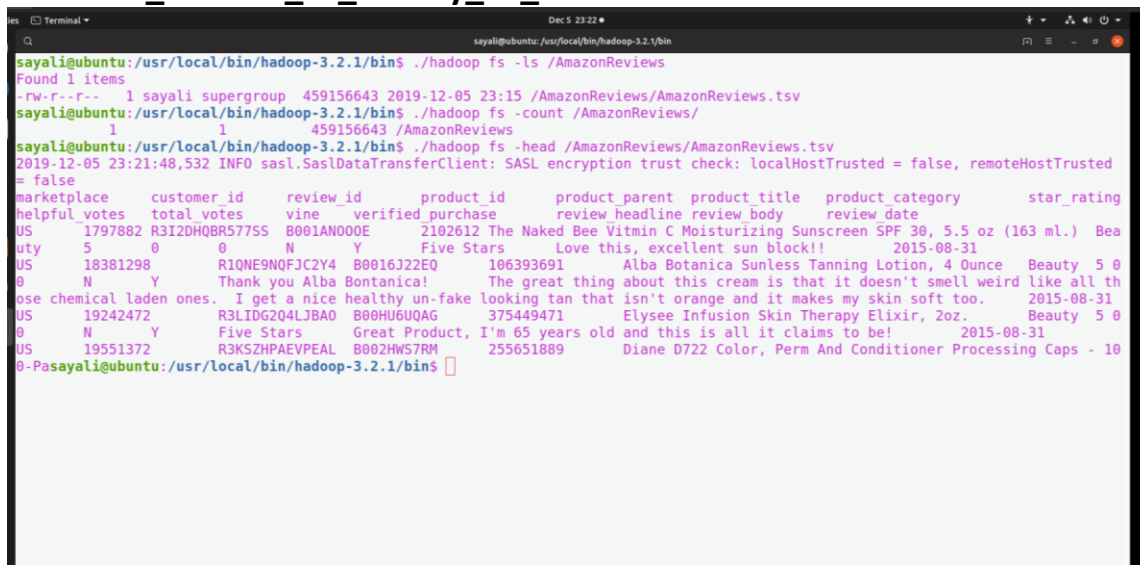
1. A collection of reviews written in the Amazon.com marketplace and associated metadata from 1995 until 2015. This is intended to facilitate study into the properties (and the evolution) of customer reviews potentially including how people evaluate and express their experiences with respect to products 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.



```
sayali@ubuntu:/usr/local/bin/hadoop-3.2.1/bin$ ./hadoop fs -ls /AmazonReviews
Found 1 items
-rw-r--r-- 1 sayali supergroup 459156643 2019-12-05 23:15 /AmazonReviews/AmazonReviews.tsv
sayali@ubuntu:/usr/local/bin/hadoop-3.2.1/bin$ ./hadoop fs -count /AmazonReviews/
1
459156643 /AmazonReviews
sayali@ubuntu:/usr/local/bin/hadoop-3.2.1/bin$ ./hadoop fs -head /AmazonReviews/AmazonReviews.tsv
2019-12-05 23:21:48,532 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted
= false
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 1797882 R3I2DHQBR5775S B001AN000E 2102612 The Naked Bee Vitmin C Moisturizing Sunscreen SPF 30, 5.5 oz (163 ml.) Bea
uty 5 0 0 N Y Five Stars Love this, excellent sun block!! 2015-08-31
US 18381298 R1QNE9NQFJC2Y4 B0016J22EQ 106393691 Alba Botanica Sunless Tanning Lotion, 4 Ounce Beauty 5 0
0 N Y Thank you Alba Botanica! The great thing about this cream is that it doesn't smell weird like all th
ose chemical laden ones. I get a nice healthy un-fake looking tan that isn't orange and it makes my skin soft too. 2015-08-31
US 19242472 R3LIDG2Q4LJBA0 B00HU6UQAG 375449471 Elysee Infusion Skin Therapy Elixir, 2oz. Beauty 5 0
0 N Y Five Stars Great Product, I'm 65 years old and this is all it claims to be! 2015-08-31
US 19551372 R3KSZHPAEVPEAL B002HWS7RM 255651889 Diane D722 Color, Perm And Conditioner Processing Caps - 10
0-Pasayali@ubuntu:/usr/local/bin/hadoop-3.2.1/bin$
```

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](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

amazon_reviews_us_Beauty_v1_00.csv - LibreOffice Calc					
Liberation Sans 10					
AT&TAMJ15					
	A	B	C	D	E
1	marketplace	customer_id	review_id	product_id	product_title
2	US	1707862	R3Q3DQKQ5W7755	B00LANC00E	210012 The Isolated Bee Vintm C Moisturizing Sunscreen SPF 30, 5.5 oz (163 ml)
3	US	10381298	R1Q2NEHQFQJ2CY4	B00LAJ2ZEQ	10643091 Alfa Botanica Solesens Tanning Lotion, 4 Ounce
4	US	15024737	R1G3Q2Q2L3J8A0	B00KALQQAQ	10644847 Elysse Infusion Skin Therapy Elux, 2oz
5	US	18951372	R3K5ZLPA4VPEAL	B002HW578M	25665389 Oars D712 Color, Perm And Conditioner Processing Caps - 100-Pack - Clear
6	US	14802487	R4Q2Q2Q2Q2Q2Q2	B005M99K9W	116256747 Bore UV Aqua Rich Watery Essence SPF30+PA+++ (back of 2)
7	US	2009389	R1J3QF4A88SP84	B000VYL125	186148615 Skatard Clarifying Cleanser
8	US	18397215	R3Q3KQJGQJ8QJPH	B0015YWTFG	111742328 CoverGel Queen Collection Perfect Point Plus Eyeliner Black Onyx 200
9	US	31952191	R1G3LUPVQ1QDMH	B009Z2V4M2	256620087 Alereba Mandan Eye Makeup Remover, 1.68oz
10	US	52216383	R8T7Y1JLX1J70	B00M5LWJ7K	248615458 Can You Handlebe All-Natural Secondary Moisturize Wax - Extra Strength, Firm Hold
11	US	10276218	R1G3Q3Q3Q3Q3Q3	B000YQ4115	8612905 Jemellee Great Lash Washable Mascara, Clear (1.05 fl. oz)
12	US	24605453	R2A3Q3Q3Q3Q3Q3	B005AQ3Q2V	282127037 12 New, High Quality, Amber 2 ml (0.68 Ounce) Glass Rods, with Office Reducer and Black Cap.
13	US	30788223	R13MGP1Y8Q23L	B00H7Q2QVJ	311308827 Welterpe Aquaviva Water Elixir
14	US	11297508	R1G3Q3Q3Q3Q3Q3	B00P7L8MA4	390303248 Pongee Agave Nectar Plus Silica Curling Cream Plus Debrazzant, 6 Ounce
15	US	29605778	R37CQJQ2Q3Q2QJH	B000R6V204	754959888 Shea Butter
16	US	41238422	R14VH8CKQ1Q728	B0079MA112	874223918 Exude House Fresh Cherry Tint # 02 RD3031
17	US	23620123	R34BQ654M2QNA	B00FVXBLHG	464001209 Vintage LF Sponge Holder
18	US	25964245	R3LUPQ3Q3Q3Q2H0	B000L8LCB0	208607870 Jenna Jenson Hair Breaker Perfume parent
19	US	861375	R3R3VH75R3Q2C8	B000GL4V05	620727388 Amaxon Cosmetics, Jia Kuo, X202 Original Hydra Serum, 30ml
20	US	947878	R5R4KLSRD48L78	B005MIR0W	351873146 Seick Hydro Silk TrimStyle Moisturizing Razor
21	US	18418407	R1Q2Q3Q3Q3Q3Q3	B003UK4HQ2	130347950 Cool-B Viality Easheer Rechargeable Electric Toothbrush 1 Count
22	US	42444737	R08F3K5Q3Q3Q3Q3	B00PATJURY	378370722 TBS-SEISMIC PERFECTLY (UN)DOONE Hair Spray 7.7 oz
23	US	4224848	R1Q2Q3Q3Q3Q3Q3	B009Z2V4M2	371877855 Joy & Karma Tropical Vitamin C Anti Aging Serum for Face with Hyaluronic Acid, 1.6 oz
24	US	49631440	R1J3Q3Q3Q3Q3Q3	B004MQC3Q0	296319484 TheGlossy Dry Mouth Lozenges
25	US	15548310	RCP2ZJ8J7WTF2	B000NLT01G	98576391 Mentur Edge MK 23C Long-Handled Traditional Double Edge Safety Razor - Excellent Comfort, Comfort and Design - 4.2 inches, Chrome Finish
26	US	1776078	R2R3W4VYQ2Q3Q3Q3	B000M8K1C6	227076018 Burt's Bees Eye Cream for Sensitive Skin, 0.5 Ounce
27	US	4029997	R38Q7Q3Q3Q3Q3Q3	B004HQH1Q1	790564533 Doting 5 X 2 Way Multifunction Doting Pen Set for Nail Art Manicure Pedicure, 4 Ounce
28	US	4113778	R1Q2Q3Q3Q3Q3Q3	B000M8K1C6	57378789 Goodies Burt's Bees Hair Wax Pump - Trio 2.4.6
29	US	1776078	R3Q3KALXQ3Q3Q3Q3	B000JUMEFY	343124386 Gpacks Eyelashes - 7475 by Christina
30	US	22944336	R13A4M8H9H9G4	B000VYVYH6	495838458 Dr Song Benzoyl Peroxide 10% Acne Cream Gel Treatment Lotion up to 8oz
31	US	1546105	R2R3W4VYQ2Q3Q3Q3	B00N4JAC9Y	218629672 Philips Sonicare Powerup Battery Toothbrush
32	US	106678	R1Q2Q3Q3Q3Q3Q3	B004ZP6D8	496393111 L.F. Powder Puffs
33	US	22944336	R2R3W4VYQ2Q3Q3Q3	B000JUMEFY	495729523 Body First Exfoliant Foot Peel, Lavender Scented, 2.4 Fl. Oz.
34	US	15322085	R1VHCW8Y8G5G8	B007Z2NATQ	240484503 Philips Sonicare Essence Rechargeable Electric Toothbrush
35	US	98605796	R137C4C7R3Q3Q3Q3	B0018A8D58	274266073 India Tempest - Song of India - 120 Stick Large Blue
36	US	11846474	R1E8L4VQ1YQ3Q3Q3	B0087TALVQ	60260712155 Professional Gilder Ceramic Tumblebar tone Flat Iron Hair Straightener ( Straightens & Curls with Adjustable Temp ) Incl Glove, Pouch, & Travel Size Arg
37	US	18591811	R3Q3KALXQ3Q3Q3Q3	B008SL15Y	64666057 Skulls and Roses Ed Hardy Eau De Parfum for Women

amazon_reviews_us_Beauty_v1_00.csv - LibreOffice Calc					
Liberation Sans 10					
AT&TAMJ15					
	A	B	C	D	E
1	star_rating	helpful_votes	total_votes	verified_purchase	review_headline
2	5	0	0	Y	Five Stars
3	5	0	0	Y	Thank you Alfa Botanica!
4	5	0	0	Y	Five Stars
5	5	0	0	Y	GOOD GOOD!
6	5	0	0	Y	This soaks in quick and provides a nice
7	5	0	0	Y	Five Stars
8	5	0	0	Y	Good buy
9	5	0	0	Y	Best makeup remover!
10	5	0	0	Y	Tame the wild mustache
11	5	0	0	Y	but it's like having nothing on them at all
12	5	0	0	Y	Good Product, I'm 65 years old and this is all it claims to be
13	5	0	0	Y	Optimum Oral Health
14	5	0	0	Y	Love this cream!
15	5	0	0	Y	It works so much better than store bought
16	5	0	0	Y	Five Stars
17	5	0	0	Y	Great product, fast delivery
18	5	0	0	Y	Does not smell cheap
19	5	0	0	Y	Five Stars
20	5	0	0	Y	Love this Razor/trimmer
21	5	0	0	Y	Five Stars
22	5	0	0	Y	Works Okay As A Styling Aid, But Does
23	5	0	0	Y	My hair feels so soft and smooth when using this
24	5	0	0	Y	Even my Dermat didn't know!
25	5	0	0	Y	Great shades
26	5	0	0	Y	Great sensitive eye cream
27	5	0	0	Y	Fun item
28	5	0	0	Y	Two Stars
29	5	0	0	Y	Great Lashes!
30	5	0	0	Y	Great product at a great value
31	5	0	0	Y	I use to get another one of the <b>Sonicare</b>
32	5	0	0	Y	They're okay
33	5	0	0	Y	Better Than Any <b>Brush</b>
34	5	0	0	Y	Great Product
35	5	0	0	Y	temple experienced
36	5	0	0	Y	I've had better and heads up surface it's just when u are in a hurry
37	5	0	0	Y	Five Stars

# 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.....)