**Big-Data-Analytics-on-Amazon-Customer-Reviews**

Performed analysis using Pyspark, chaining, secondary sorting, partitioning, summarization, and filtration patterns. Also made use of Collaborative Filtering to build the recommendation system.

**Abstract:**

Analysis of amazon is very crucial part when it comes to find an efficient way of getting insights on customer reviews about different products. Hence, this project is mainly aimed to analyse big data and produce an informative result about the customer reviews for the product Camera present on Amazon using Pyspark architecture, MLlib Collaborative Filtering and Delta Storage.

**Dataset:**

**Dataset Link:**

<https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Camera_v1_00.tsv.gz>

**Accessing the data:**

Data is present in the Amazon S3 bucket. It can be accessed as mentioned in below link, download the link and extract the Zip file – <https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Camera_v1_00.tsv.gz>

**About the data:**

DATA FORMAT - Tab ('\t') separated text file, without quote or escape characters. This dataset is 1GB in size. First line in the file is header; 1 line corresponds to 1 record.

**DATA COLUMNS:**

1. marketplace - 2 letter country code of the marketplace where the review was written.

2. customer\_id - Random identifier that can be used to aggregate reviews written by a single author.

3. review\_id - The unique ID of the review.

4. 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.

5. product\_parent - Random identifier that can be used to aggregate reviews for the same product.

6. product\_title - Title of the product.

7. product\_category - Broad product category that can be used to group reviews (also used to group the dataset into coherent parts).

8. star\_rating - The 1 to 5 star rating of the review.

9. helpful\_votes - Number of helpful votes.

10. total\_votes - Number of total votes the review received. vine - Review was written as part of the Vine program.

11. verified\_purchase - The review is on a verified purchase.

12. review\_headline - The title of the review.

13. review\_body - The review text.

14. review\_date - The date the review was written.

**Data Analysis/Insights:**

1. Find the total number of products present in the dataset.
2. Find the average product rating reviews for each product.
3. Find the topN reviewed products sorted by count.
4. Find total product rating for all those products which are present in the topN reviewed products.
5. Find all the users who has reviewed each product.
6. Find all the records partitioned by the date in which the product was reviewed.
7. Find all the product information for each unique star rating of the product by dividing it into different categories/bins.
8. Recommend the products to the user based on the star rating.
9. Find the count of the reviews grouped by date for each product.
10. Find the count of products for each product star rating.

**Techniques/Technologies Used:**

1. Pyspark and Databricks
2. Summarization Pattern – Numerical Summarization using Native Functions
3. Partitioning
4. Sorting
5. MLlib Recommendation System using Collaborative Filtering