# Big Data Predictive Analysis of Amazon Product Review

**Monika Mishra, Jaydeep Chopde, Maitri Shah,**
**Pankti Parikh, Rakshith Chandan Babu, Jongwook Woo**
Computer Information Systems Department
California State University Los Angeles, California, United States
[e-mail: mmishra2@calstatela.edu, jchopde@calstatela.edu, mshah3@calstatela.edu,
pparikh6@calstatela.edu, rchanda@calstatela.edu, jwoo5@calstatela.edu,]
**Corresponding author: Monika Mishra*

## Abstract

The paper presents big data predictive analysis of the Amazon Products reviewed between the year 2005 and 2015. The data set is over 5GB, which is not easy to store and analyze in the traditional systems. Thus, Hadoop is adopted that can be even linearly scalable as the data size grows. The descriptive part of the analysis shows the various shopping patterns and customers' sentiments based on reviews and ratings provided by the customers on the online Amazon shopping site. It also tells about the shopping patterns by factors such as year, month and product category. In addition to the various valuable insights from the dataset, the predictive analysis is done to provide various recommendation prediction such as item recommendation and rating prediction of the Amazon products reviewed.

**Keywords**: Big Data, Hadoop, Microsoft Azure, machine learning, recommendation, sentiment analysis, HDFS

## 1. Introduction

With the quick growth of internet and it's increasing accessibility, e-commerce has developed rapidly in the past years. Because of the numerous advantages and benefits, more and more people these days prefer buying things online over the conventional method of going into stores [1]. Both businesses and customers have embraced online sales as a cheaper and more convenient way to shop, but just like anything associated with the Internet, there are benefits and dangers associated with shopping online [2]. The biggest disadvantage of e-commerce is that online shoppers do not have the ability to physically inspect or try on the items being considered for purchase [3]. Because of this constraint, most of the customers, on any online shopping sites, make purchasing decisions based on reviews and ratings. Therefore, it's very important for the business to know the various shopping patterns and customer's sentiments based upon the reviews and ratings.

Furthermore, the data set becoms too large to store and processed, which initiates Big Data. Big Data is defined as non-expensive frameworks, mostly on distributed parallel computing systems, which can store a large-scale data and process it in parallel. A large-scale data means a data of giga-bytes or more, which cannot be processed or expensive using traditional computing systems [4]. Hadoop is one of the popular Big Data platform and Hive is one of ecosystems for Big Data analysis.

Besides, the goal of creating a recommendation system is to recommend one or more "items" to "users" of the system.This model can learn about a user's preferences through observations made on how they rate products. Based on those observations, it recommends new items to the users when requested. It also predicts ratings for different products. These predictions are of great significance for the business.

This study was supported by Isaac Engineering and Oracle Cloud Innovation Accelerator. Oracle Big Data Cloud Service was used in the data analysis.

## 2. Related Work

Bhavesh [5] and Max [6] performed analysis on Amazon product review dataset but their goals and techniques were quite different from ours.

Bhavesh's work was to classify Amazon product review to positive and negative. He concentrated on just one product category – *baby*. He also performed sentimental analysis for one of the baby products. The tools used in his approach were Python, GraphLab and S Frame. The difference in our approach is that we adopted big data techniques, Hadoop and Hive, for data analysis. It is not only to store and process massive data set but also faster to analyze such a massive data set using these techniques. We did analysis based on ratings, reviews , timeperiod etc. Also we considered all of the Amazon products and not just one. Even with the machine learning, we implement recommendation model for predictions which was missing in Bhavesh's research.

Another similar research study was done by Max [6] who, unlike Bhavesh, didn't concentrate on just one product category. Max performed descriptive analysis using *Sparklyr* platform. In contrast, our research is about both descriptive and predictive analysis. We did a variety of descriptive analysis using Big Data on Cloud Computing. The Hive Query Language (HiveQL) was used to analyze structured data. We did the predictive analysis by building recommendation model while Max's work was restricted just to descriptive analysis.

## 3. Background/Existing Works

### 3.1 Big Data Analysis

The big data analysis is based on Hadoop. The data set is uploaded to Hadoop big data systems and transformed to be analyzed using Hive ecosystems. Big Data is defined as non-expensive frameworks, mostly on distributed parallel computing systems, which can store a large-scale data and process it in parallel. A large-scale data means a data of giga-bytes or more, which cannot be processed or expensive using traditional computing systems [6]. Hadoop is one of the popular Big Data platform and Hive is one of ecosystems for Big Data analysis. We imported the results of Hive queries into Microsoft Excel and created 3D Map chart to represent Spatial-Temporal chart even in the 3D world map.

### 3.2 Predictive Analysis

Our Matchbox Recommender module for predictive analysis is to implement the four score recommenders recommend different metrics: a) item recommendation, b) rating prediction, c) similar items and d) similar users. After this step, each metric is evaluated for accuracy.

## 4. Dataset Description

The dataset used for this paper is from an open data source of the Amazonaws site. The product reviewed is from Amazon online shopping site from four countries i.e. US (United States), UK (United Kingdom), FR (France), DE (Germany) between the year 2005 and 2015. The total file size is **5.26 GB**. The file format is a mix of Tab Separated Values (TSV) and Comma Separated Values (CSV). The total number of product reviews is 9.57 million. In addition to this, we have used dictionary files as well for different languages to compute customers' sentiments.

For predictive analysis we have used only the dataset from 'US' country which is **3.63 GB**.

## 5. System Specification

### 5.1 Big Data Analysis

We used services offered by Oracle cloud. We extensively worked on Hadoop system components. The Hadoop cluster details are given below:

Cluster Version: Oracle Big Data Compute Edition
Number of Nodes: 5
Memory size: 150 GB
CPU speed: 2.20 GHz
HDFS capacity: 147 GB
Storage: 678 GB

### 5.2 Predictive Analysis

For prediction, we have used Microsoft Azure Machine Learning Studio. The specification is provided below:

Platform used: Microsoft Azure Machine Learning Studio
Execution: Single Node
Storage Space Capacity: 10 GB

## 6. Architecture Workflow

### 6.1 Big Data Analysis

For the big data analysis we extracted the product review data for the country US, UK, DE, FR in TSV format from amazonaws site. As the reviews were in various languages, we also downloaded the English (TSV),German (TSV) and French (CSV) dictionaries. Then, we uploaded our data to the HDFS and used Hive queries to create external tables based on the data that was uploaded in HDFS. We then triggered Hive queries to analyze the data. For the visualization purpose, we used Tableau, Power BI and Excel 3D maps. At the end we found useful insights.

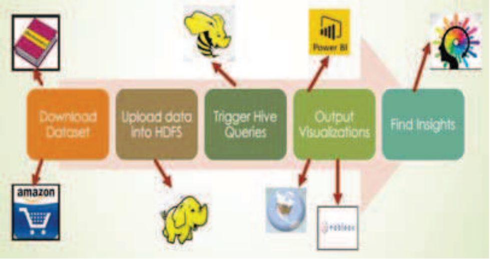


**Fig. 1.** Big Data Analysis - Flow Diagram

### 6.2 Predictive Analysis

For the predictive analysis we used only the dataset from the US country. As the US data was very large, we sampled the dataset and reduced its size to 111 MB. The reduced dataset was further worked upon for data cleaning, removing duplicates etc. The dataset was then split into two sets; one for training the algorithm, and another for evaluation purposes. At the end, the predictions were made, and the models were evaluated.

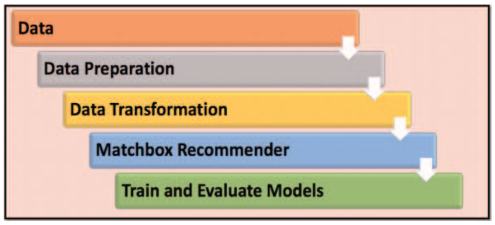


**Fig. 2.** Predictive Analysis - Flow Diagram

## 7. Data Analysis and Visualizations

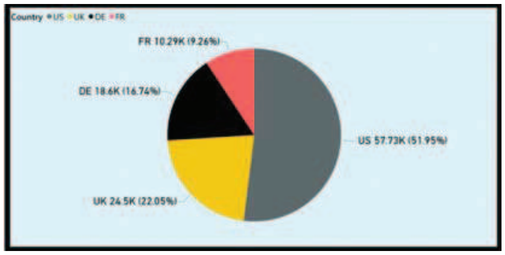


**Fig. 3.** Total review sentiment count by country

From the Figure 3, it shows that the overall sentiment count is maximum for USA which is 57.73K (51.95%) followed by UK (24.5K - 22.05%), Germany (18.6K - 16.74%), France (10.29K - 9.26%) between the year 2005 and 2015.

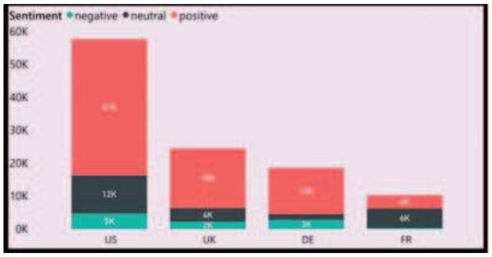


**Fig. 4.** Sentiment analysis by country

The Figure 4 shows the distribution of review sentiment - negative, neutral and positive among US, UK, DE and FR. It can be seen that all the countries except France have comparatively a positive view about the amazon products. Also, US has the maximum sentiment count followed by UK, Germany and France.

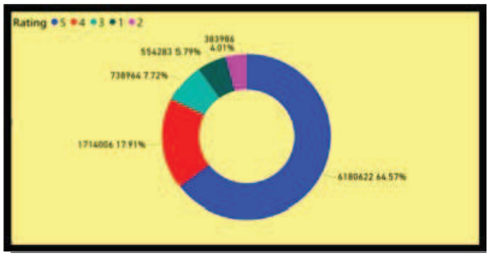


**Fig. 5.** Distribution of overall ratings for Amazon products

The **Fig. 5** shows the distribution of rating from 1 to 5 (5 being the maximum) on the amazon site of amazon products. It can be seen that the rating 5 has the maximum count (64.57%) followed by 4 (17.91%), 3 (7.72%), 1 (5.79%) and 2 (4.01%). So, it can be said that the customers are quite happy with the products that they buy from amazon.

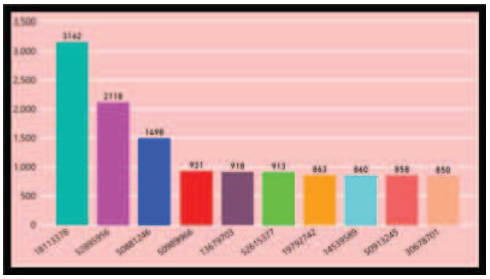


**Fig. 6.** Number of reviews given by unique users over ten years

**Fig. 6** shows the number of reviews given by unique users over ten years which indicates that amazon has customers who are not just using the products but also prefer to give reviews regularly as one of the users has given the review more than 3k over a span of ten years.

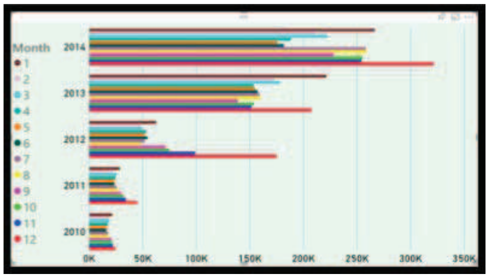


**Fig. 7.** Review count by year and month

The **Fig. 7** shows the reviews count by year (2010 to 2014) and month (January to December). From the above chart, it is clearly visible that the review count has significantly increased from 2010 to 2014. Also, the review count is more in the holiday months (November, December, January) as compared to other months.

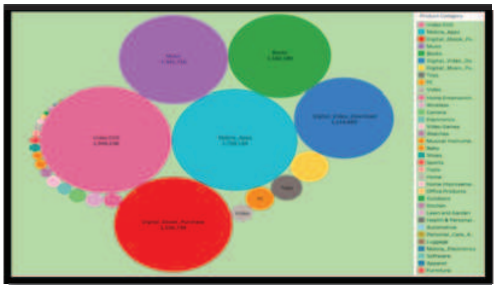


**Fig. 8.** Review count by product category

The **Fig. 8** shows the review count as per the product category which indicates that Video DVD was the maximum reviewed product which is 1,948,238. Then comes mobile apps with 1,759,144 and then the Digital_ebook_Purchase with 1,556,739 and so on. It can be seen that overall digital products have received more reviews than the non-digital products.

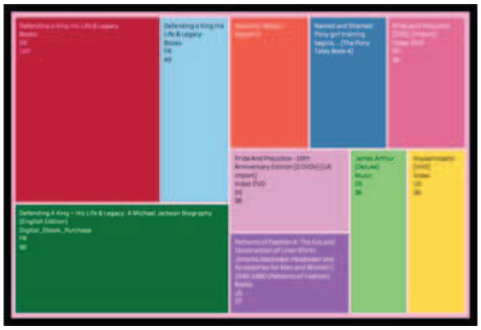


**Fig. 9.** Popular product based on average rating and review count

The tree map of **Fig. 9** reveals top 10 most popular product among the users based on user's average ratings and reviews. It can be concluded that Books is the most popular category. Among them Defending a king his life and legacy is top most trending book, which has the highest number of ratings and reviews. The same book is popular in the Digital e-book category as well.

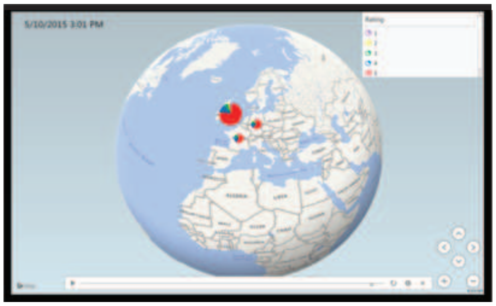
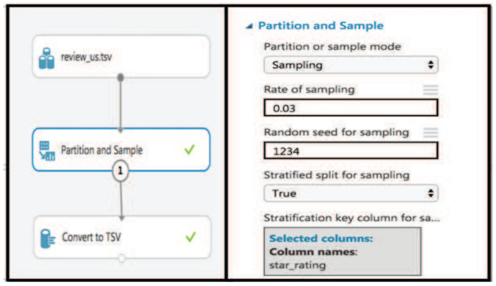


**Fig. 10.** Rating based geospatial representation for product category-baby

The 3D excel power map in Figure 10 shows that the product category "Baby" has received maximum rating (5) followed by 4, 3, 2, 1 in all the countries. This shows that users are overall happy with the baby products. Another thing which can be noticed is that the rating count has gradually increased over the years.

## 8. Predictive Analysis

### 8.1 Sampling



Partition and Sample

Partition or sample mode: Sampling

Rate of sampling: 0.03

Random seed for sampling: 1234

Stratified split for sampling: True

Stratification key column for sa...

Selected columns:
Column names:
star_rating

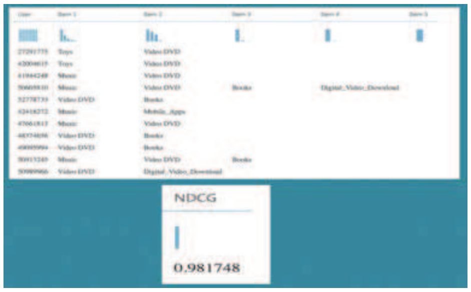
**Fig. 11.** Sampling original dataset

Since the original US dataset was 3.63 GB out of 5.26 GB, we sampled the dataset using "Partition and Sample" module in Microsoft Azure to reduce the size to 111.3 MB. The sampled dataset was further used for prediction. Stratified Split ensures that the output dataset contains a representative sample of the values in the selected column.

### 8.2 Matchbox Recommender

The goal of the recommender is to provide Amazon customers with recommendations for product categories based on their previous ratings, as well as the ratings of other users. Moreover, the model has a feature to predict the future ratings by user for a category.

After column were selected, the dataset was split into training and testing fractions by .75 to .25 ratio. We adopt Matchbox Recommender algorithm and generate two Score Recommender modules. Each of the two score recommenders represent different metrics: a) item recommendation and b) rating prediction.



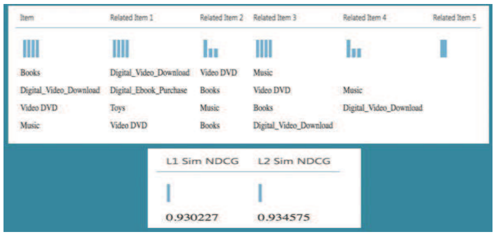
| User | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 |
|---|---|---|---|---|---|
| 27291775 | Toys | Video DVD | | | |
| 43004615 | Toys | Video DVD | | | |
| 41944248 | Music | Video DVD | | | |
| 50605810 | Music | Video DVD | Books | Digital_Video_Download | |
| 52778739 | Video DVD | Books | | | |
| 43418272 | Music | Mobile_Apps | | | |
| 47661813 | Music | Video DVD | | | |
| 48074696 | Video DVD | Books | | | |
| 49095994 | Video DVD | Books | | | |
| 50913249 | Music | Video DVD | Books | | |
| 50989966 | Video DVD | Digital_Video_Download | | | |

NDCG

0.981748

**Fig. 12.** Item Recommendation - From Rated Items (for model evaluation) and Evaluation

This option enables evaluation mode, and the module makes recommendations only from those items in the input dataset that have been rated. This model is evaluated by Normalized Discounted Cumulative Gain (NDCG) which in this case is 0.98. That is a very encouraging result.

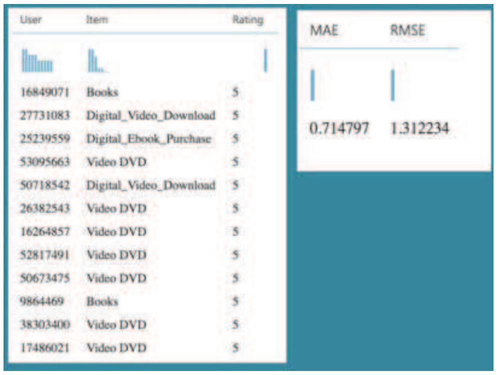


| Item | Related Item 1 | Related Item 2 | Related Item 3 | Related Item 4 | Related Item 5 |
|---|---|---|---|---|---|
| Books | Digital_Video_Download | Video DVD | Music | | |
| Digital_Video_Download | Digital_Ebook_Purchase | Books | Video DVD | Music | |
| Video DVD | Toys | Music | Books | Digital_Video_Download | |
| Music | Video DVD | Books | Digital_Video_Download | | |

| L1 Sim NDCG | L2 Sim NDCG |
|---|---|
| 0.930227 | 0.934575 |

**Fig. 13.** Related Items and Evaluation

This feature finds related items. By predicting related items, one can generate recommendations for users based on items that have already been rated. For pairs of related items that a group of users has rated, one can predict user rating for one item based on the rating of the other items. Results are evaluated by the similarity of the ratings using both L1(Manhattan) and L2 (Euclidian) average normalized discounted cumulative gain (NDCG) averaged over all the pairs selected. In this case it’s 0.93 for both which is pretty good.



| User | Item | Rating |
|---|---|---|
| 16849071 | Books | 5 |
| 27731083 | Digital_Video_Download | 5 |
| 25239559 | Digital_Ebook_Purchase | 5 |
| 53095663 | Video DVD | 5 |
| 50718542 | Digital_Video_Download | 5 |
| 26382543 | Video DVD | 5 |
| 16264857 | Video DVD | 5 |
| 52817491 | Video DVD | 5 |
| 50673475 | Video DVD | 5 |
| 9864469 | Books | 5 |
| 38303400 | Video DVD | 5 |
| 17486021 | Video DVD | 5 |

| MAE | RMSE |
|---|---|
| 0.714797 | 1.312234 |

**Fig. 14.** Rating prediction and Evaluation

This feature predicts ratings. When one predicts ratings, the model calculates how a given user will react to a particular item, given the training data. This model is evaluated by the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) which in this case is 0.714797 and 1.312234 respectively.

## 9. Conclusion

In the paper, Hadoop Big Data systems is adopted to store and analyze for over 5GB amazon data set, which is linearly scalable when the data set grow further. The insights of the analysis is as follows:

- Books is the most popular category based on ratings and reviews.
- Around 3,162 reviews(maximum) were written by one user in the span of 10 years.
- The review count has gradually increased over the years.
- The review count is maximum in the holiday months – November, December, January.
- 64.57% people gave the maximum rating 5 to the products.
- Video DVD received the maximum number of reviews.
- The consumer sentiments are mostly positive in US, UK and Germany country.
- Digital products have received more reviews than the non-digital products.

Besides, recommendation model is implemented to predict the item recommendation and rating prediction using Azure ML Studio as predictive analysis. It can help in finding customers with the preferred items.

## References

[1] A. S. Jain, “Top 10 Benefits of Online Shopping (and 10 Disadvantages),” *ToughNickel*, 2018. [Online]. Available: https://toughnickel.com/frugal-living/Online-shopping-sites-benefits.

[2] C. Morah, “Shopping Online: Convenience, Bargains And A Few Scams,” *Investopedia*, 2018. [Online]. Available: https://www.investopedia.com/articles/pf/08/buy-sell-online.asp.

[3] D. L. Montaldo, “When It Is Best to Shop Online and When It Is Not?,” *The Balance*, 2019. [Online]. Available: https://www.thebalance.com/the-pros-and-cons-of-online-shopping-939775.

[4] Woo, Jongwook, & Xu, Yuhang (July 18-21, 2011), Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing, The 2011 international Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2011), Las Vegas.

[5] B. Patel, "Predicting Amazon product reviews' ratings," *Towards Data Science*, 26-Apr-2017. [Online]. Available: https://towardsdatascience.com/predicting-sentiment-of-amazon-product-reviews-6370f466fa73.

[6] M. Woolf, "Playing with 80 Million Amazon Product Review Ratings Using Apache Spark," *minimaxir*, 02-Jan-2017. [Online]. Available: https://minimaxir.com/2017/01/amazon-spark/.