Logo

Description automatically generated

**ALY 6110: Data Management and Big Data**

**Final Project Proposal**

**CRN: 70633**

**Week 3: Assignment**

**NUID: 002938209**

**Email: jain.abhin@northeastern.edu**

**Submitted by: Abhinav Jain**

**Date: 10/06/2022**

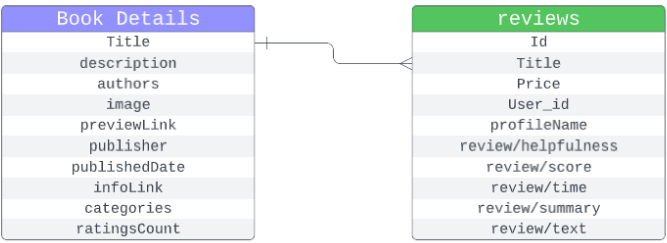
**Submitted to: Prof. Valeriy Shevchenko**

**Introduction:**

According to each student's particular guidelines for this project, a suggestion for the final dataset is included in this task. To manage the 3 million user dataset with around 200 000 unique books that have feedback or reviews on Amazon, I have picked the dataset from Amazon Books Reviews. The 142 million reviews in this dataset were written between May 1996 and July 2014. On the other hand, this information includes numerous descriptions and book IDs.

The ID of the book, the title of the book, the price of the book, the User Id of the person who rated the book, and the profile name of the book reviewed by the person are the attributes or features that are included in the dataset. Additionally, links are provided where users may get information about a book on Google Books to find out whether it is now available or to read reviews.A picture containing company name

Description automatically generated

Fig 1: Various Features of the Dataset (Amazon Book Reviews)

**Summary**

In this assignment, generating insightful insight from the data that benefit a business’s growth. Big data help in market analysis, research development, and various other organizational developments in the industry. The obtain exact results after using the big data, can be used to evaluate the performance of the data.

By examining the book reviews that the book society is interested in, this dataset will be used to assess user sentiment. The whole Amazon reviews will provide you with a comprehensive understanding of data processing. However, this project's primary goal is to gather data-related information based on user reviews.

We can think more about the various big data technologies, including. The data should be preprocessed using big information or insights. Apache Sparks may be used to collect information on the data, which can be saved to the HDFS (Hadoop Distribution File System). However, we can visualize the data using the big sheets by confirming the finding of the data. There may be some changes as per the requirements of the project.

**Research:** This dataset was chosen from Kaggle, which was simple to locate, and when I opened it in Excel, I saw that it had a lot of data. To manage that enormous data, I would want to employ community-based data bricks, where we can manage data brick services. SLAs, Amazon webservices, and Azure are a few alternatives. to manage the workload

We may utilize Apache Spark as a large data distributed processing system to get a clear picture of how huge data can be handled.

Graphical user interface, application

Description automatically generatedFig: Managing Bigdata with Databricks

**Analysis:** We can provide a better understanding after implementing the project start. Let us first start by discussing between the team.

**Conclusions summary**:

Between May 1996 and July 2014, 142 million reviews in this dataset were written. However, this data also contains multiple descriptions and book IDs. The elements or features that are included in the dataset are the ID of the book, the title of the book, the price of the book, the User Id of the person who rated the book, and the profile name of the book that person reviewed. Additionally, there are links that direct consumers to Google Books, where they may read reviews or find out if a book is currently available. If there will be any requirement of the different tools we can use while doing the project we can use other tools also.

**References:**

[1] **EDA for Amazon books reviews**

Mohamedbakhet

[**https://www.kaggle.com/code/mohamedbakhet/eda-for-amazon-books-reviews/data**](https://www.kaggle.com/code/mohamedbakhet/eda-for-amazon-books-reviews/data)

[2] **Comparing Databricks to Apache Spark**

[**https://www.databricks.com/spark/comparing-databricks-to-apache-spark**](https://www.databricks.com/spark/comparing-databricks-to-apache-spark)

[3] **Big Data**

Schmarzo

[**https://aws.amazon.com/big-data/what-is-spark/**](https://aws.amazon.com/big-data/what-is-spark/)