8157 – Project Milestone 2

**Ali Hasan, Md Tafsirul Islam, Muzakkir Qadri Mohammed, Abhilash Harsh Saksena**

**Problem statement:**

Trend analysis is a process of computationally identifying shopping trends with respect to brands to strategize business. In this project we are doing trend analysis on reviews data to analyze whether the shopping seasons are predictive and if larger product portfolios result in more data points for a manufacturer. Review frequency and positive reviews with recommendations can be utilized as a reference to improve service quality, future product development, reduce user churn and improve how the product is presented. This project can possibly be used to provide specific marketing strategies. We also plan to show gathered information in graphs.

**Design:**

**UML Diagram:**

In the UML diagram below we have two types of actors here- the users, and the administrators. Users will view products and possibly leave a review. Our task as administrators is to-   
1. Get the reviews by extracting them using scripts

2. Store data in HDFS

3. Execute several querries and saving those results

4. Show results to manager in a graphical format



**Implementation Details:**

**Pyspark:**

PySpark is an interface for Apache Spark in Python. It not only allows you to write Spark applications using Python APIs, but also provides the PySpark shell for interactively analyzing your data in a distributed environment. Spark is a unified analytics engine for large-scale data processing. It provides high-level APIs in Scala, Java, Python, and R, and an optimized engine that supports general computation graphs for data analysis. It also supports a rich set of higher-level tools including Spark SQL for SQL and Data Frames, MLlib for machine learning, GraphX for graph processing, and Structured Streaming for stream processing.

**Installation Steps:**

1. Download and install Gnu on windows (GOW). GOW allows you to use linux commands on windows. In this install, we will need curl, gzip, tar which GOW provides.
2. Download and install Anaconda.
3. Open a new command line (CMD).
4. Go to the Apache Spark website

a) Choose a Spark release

b) Choose a package type

c) Choose a download type: (Direct Download)

d) Download Spark. If you download a newer version, you will need to modify the remaining commands for the file you downloaded.

1. Move the file to where you want to unzip it.
2. Unzip the file. Use the bolded commands below

**gzip -d spark-2.1.0-bin-hadoop2.7.tgz**

**tar xvf spark-2.1.0-bin-hadoop2.7.tar**

1. Download winutils.exe into your **spark-2.1.0-bin-hadoop2.7\bin**

**curl -k -L –o winutils.exe**[**https://github.com/steveloughran/winutils/blob/master/hadoop-2.6.0/bin/winutils.exe?raw=true**](https://github.com/steveloughran/winutils/blob/master/hadoop-2.6.0/bin/winutils.exe?raw=true)

1. Make sure you have [Java 7+](https://www.java.com/en/) installed on your machine.
2. Next, we will edit our environmental variables so we can open a spark notebook in any directory.

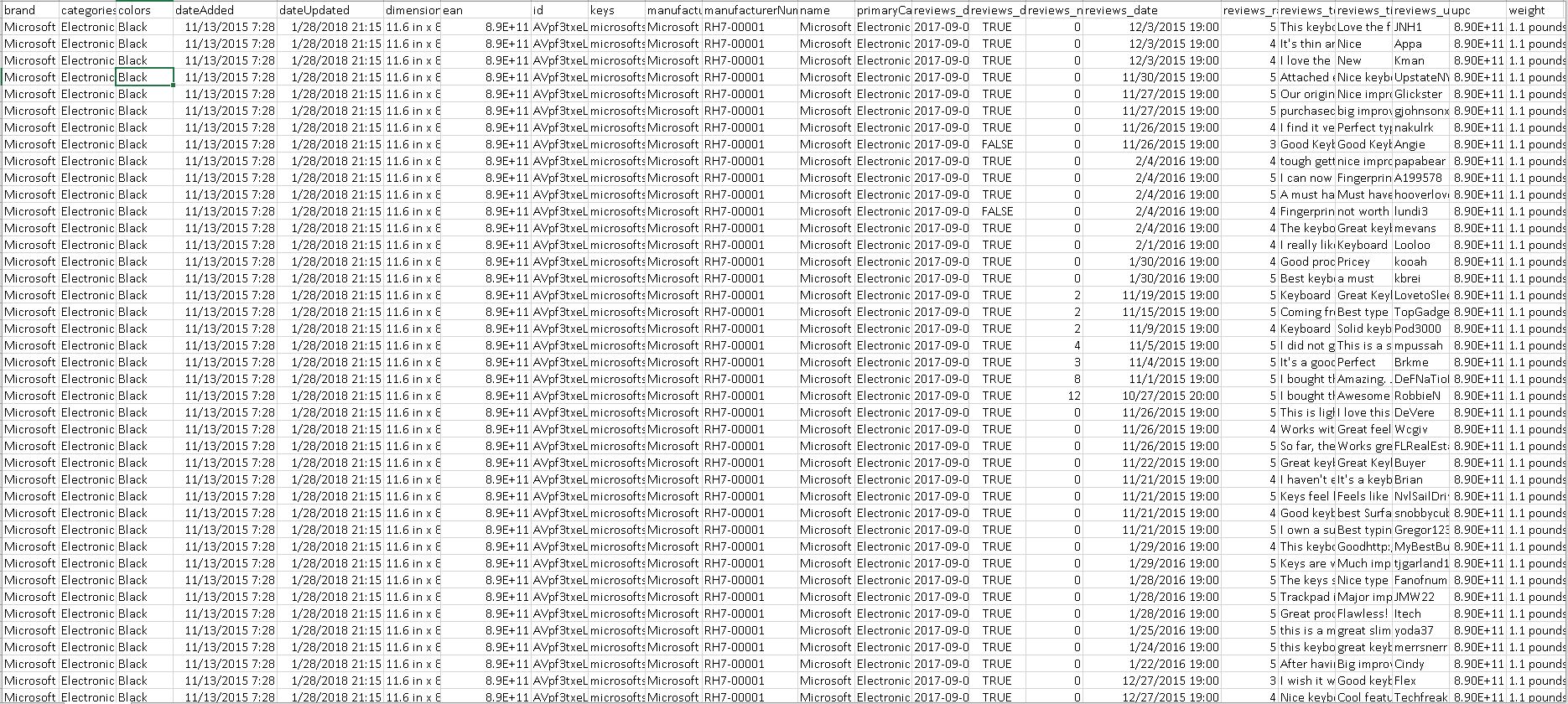
* setx SPARK\_HOME C:\opt\spark\spark-2.1.0-bin-hadoop2.7
* setx HADOOP\_HOME C:\opt\spark\spark-2.1.0-bin-hadoop2.7
* setx PYSPARK\_DRIVER\_PYTHON ipython(or Jupyter for newer anaconda versions
* setx PYSPARK\_DRIVER\_PYTHON\_OPTS notebook
* Add C:\opt\spark\spark-2.1.0-bin-hadoop2.7\bin to your path.

1. Close your terminal and open a new one. Type the command below.

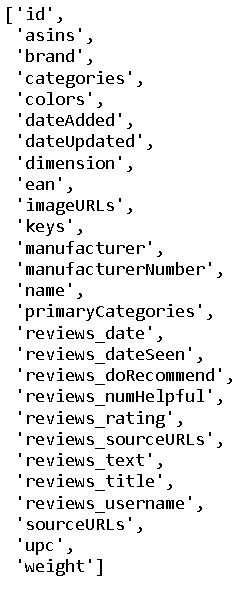
* Pyspark --master local[2]

**Challenges and implementation of modules:**

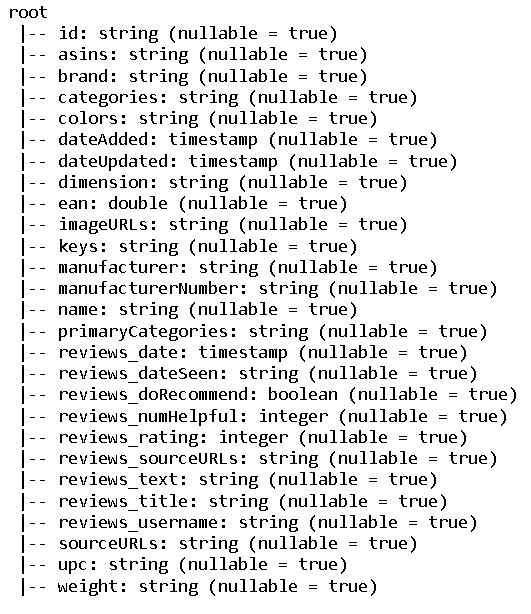
**Data Representation:**



This dataset contains the reviews of electronic products from BestBuy and has the following columns:



The following is the schema of the dataframe after importing the dataset into pyspark:



**Data-Preprocessing with Pyspark:**

Data preprocessing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance. Data-Preprocessing is necessary to has the following steps:

1. loading data (e.g., load csv file)
2. Exploring data (e.g., summary statistics, data visualization, etc.)
3. Cleaning data (e.g., handle missing data)
4. Transforming data (e.g., features engineering, scaling, reformatting as Numpy array or Spark RDD (Resilient Distributed Dataset))

As the data that is being collected is reviews which are by nature non numerical, no transformations are applied. All null values have been removed from the dataset to get clean data for further processing.

**Data visualization:**

Data visualization in this project is achieved using pre-processed data as mentioned in the previous steps. The data is visualized using Tableau, a product focused on business intelligence with various features built in to gain insights from data. It uses “measures” and “dimensions” for quantitative and qualitative data. It also gives us the ability to combine “measures” into “bins” which are then used as “dimensions” to gain better insights. Different color pallets can be used to gain insight into data, such as how many positive reviews happened withing the last year to understand the public sentiment about a brand and can also be used to figure out if there’s any trend in popularity of the brand and if the trend can be co-related to any specific product launches or controversies.

Installation steps for tableau:

For this project we were using the free public version of tableau. It was downloaded from <https://public.tableau.com/en-us/s/>

Graphical user interface

Description automatically generated

Then download the exe file, run as administrator, and install the app.

Graphical user interface, text, application

Description automatically generated

Findings from the data we analyzed, we found some trends.

1. Manufacturers that have a larger portfolio have more reviews. The data below shows that a larger portfolio of products results in a larger online presence on e-commerce platforms. If we look closer, it also shows that the number of reviews is in line with number of sales which directly corelates to the entry price point of a specific brand. For example, a company like Bose which has a higher entry price point and serves a niche market of audio products has considerably less reviews than say Logitech which has a much lower entry price point and a higher price range, starting with keyboards as low as $20 to high end sim racing products which can go upwards of $1000. A similar comparison can be made between Yamaha and Microsoft, one a very niche company vs one that has a large product portfolio serving many areas of tech.

Chart, bar chart

Description automatically generated

1. Reviews trend is in line with shopping seasons. The graph below shows as such. We see that purchases go up right around back to school and holiday seasons. Such data helps manufacturers forecast the demand, adjust their production levels, and arrange for logistics. This however is a very high-level overview, as a detailed analysis would involve forecasting on every product separately, categorized by the location of sales and the price points that will make the most sense for the bottom line.

Chart, line chart

Description automatically generated

1. Users who bought more, left more reviews. Turns out people like voicing their opinions about products bought out of convenience, while sitting on the living room couch in PJs. The trends show that users leave a review almost every time they buy something. Needless to say, we found some spikes for some users who seem to be shopaholics, no judgement.

Chart

Description automatically generated

1. Reviews rose steadily during early 2010s through the rise of e-commerce and based on the dataset. The sharp drop at the tail end of the data is because of the duration between which the data was collection. Alternate, but highly implausible theory is that they dropped sharply when the whole controversy about fake reviews was flagged on e-commerce websites.

Chart, line chart

Description automatically generated

1. Number of recommendations for a brand depend on how large their product portfolio is. Just like how many reviews a brand gets. Since we are dealing with name brands here, product quality is usually up to the mark, therefore the graphs for recommendations and number of reviews look strikingly similar. The same reasoning for more reviews applies here for number of recommendations, larger the product portfolio, lower the entry price point, the greater number of recommendations a brand gets.

Chart, bar chart

Description automatically generated