Steam Reviews Data Analysis with Tableau

Authors: Azim Pathan | Clever Lemus | Walter Giron | Yuner Paredes | Sipan Kouchian
apathan3@calstatela.edu, yparede2@calstatela.edu, skouchi@calstatela.edu, clemus28@calstatela.edu, wgiron2@calstatela.edu

Department of Information Systems, California State University Los Angeles
CIS 4560 - 02 Introduction to Big Data

Abstract: Our research explains the development as well as the usage of proficiency concepts in downloading data via Hadoop technologies and to obtain the necessary skill set of high-level data visualization analysis on our data. In order to achieve our research mission, we carried out three particular stages to meet our goals. The first stage was to identify a unique dataset, the second was to clean the data using Hadoop to download the data to the Oracle Linux server, however for us to download our data, we first had to upload it to Google Drive, which provided us access to the Hadoop wget code, which then allowed us to download data straight to the server. With this preparation being done, we went beyond and used Tableau and Excel to explore and visualize our data, resulting in the creation of maps, charts, and timelines of steam reviews on Steam's videogame library to provide us greater insight for the overall research. The data set used contains a storage size of **8.17Gb** and came from Kaggle. Following the completion of these three stages, we were able to obtain the results needed to explain and evaluate our findings on multiple portions of data, found using the full dataset and our own reliant flow of work.

1. Introduction

This project analyzes a data set covering Steam reviews which holds information about positive and negative game reviews and other helpful data from Steam, enabling us to review the data to depict changes based on users' review inputs.

Since Steam is a leading frontrunner in the gaming industry and is the main archive and place for many gamers to purchase and to play videogames, we chose a dataset based on game reviews from many countries to better provide organized critical feedback insight for gaming developers and companies, old and new, to figure out how they can improve their own products. In addition, they can benefit in creating a game and observe what consumers currently dislike and like which can also benefit the workflow of a game development process. Furthermore, long term competition from multiple competitors can give the upper hand regarding game development growth based on our data analysis.

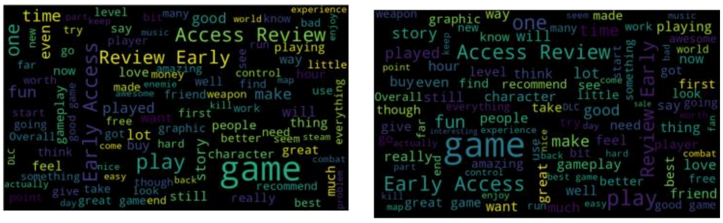
2. Related Work

Although Steam is a well-known videogame archive containing millions of historical games, there are quite a few sources which are publicly available based on our data other than from Steam. As our group reviewed other data sets and examples of the work, we came upon a user named Ali Marjani, who also examined a similar data set but took different approaches to review the data [1]. For example, Ali took the approach of clustering sentences and identifying purposes in a short text by analyzing these particular viewpoints of intent in reviews; this concept can lead to improved products or services, which is why our group decided to review the same data points but with different approaches.

In addition, we observed that Ali took the programming route to check his data, where he cleaned and formatted his data through the python language. Similarly, our team used the GitBash approach, a Unix-style command line environment that enabled us to access the Oracle Linux server.

We used similar tactics when creating our data; for example, our team along with Ali's method, grouped the data set so we could label and present our data and compare reviews of good and unfavorable remarks. The difference between how Ali and our groups showed our data is that David used Python, whereas our group used Tableau. This gave us a good understanding and visual representation of how to read our data, which gave us great insight into our findings.

The group also noted that a data Scientist at kredivo named Daniel Beltsazae Marpaung made a sentiment analysis using natural language. Daniel used sentiment-based terms with positive and negative comments to determine whether the outcomes of certain games he evaluated were excellent or poor. As stated in previous work, our group cleaned and analyzed the data with a Hadoop cluster and tableau. In contrast, Daniel installed NLP sentiment packages to conduct his Testing, and he led a WordCloud review to visualize the positive and negative words.
The following picture is Daniels's output.



We can observe that Daniel and our group had comparable discoveries but used distinct procedures, resulting in similar outputs but different data analysis methods.

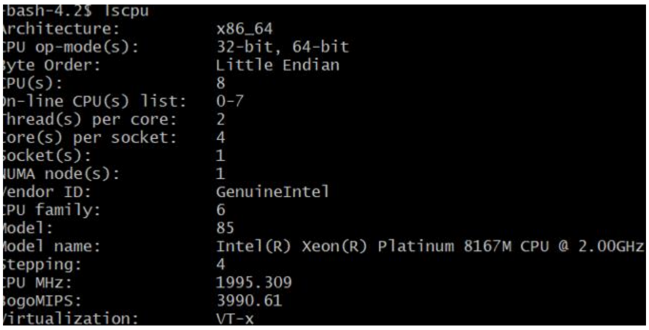
3. Specifications & Flow Chart

3.1 Specification

The data set is real publicly available reviews from steampowered.com, but the data can be kaggle.com. This data is from 2013-2021, with 23 columns, about 50 million rows, and over 300 games. The total size of the dataset is 8.17GB. Which was reduced and cleaned by us to 348MB.

3.2 Hardware

We used Hadoop 3.1.2 to combine computer resources which resulted in 2 master nodes and 3 worker nodes. With combined resources the processing power totaled up to (8) Intel Platinum 8167M CPUs and 58GB of RAM.

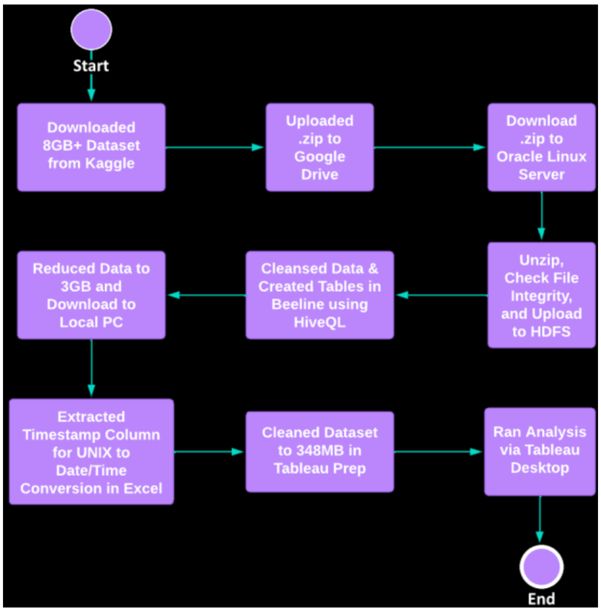


```
bash-4.2$ lscpu
rchitecture:           x86_64
PU op-mode(s):         32-bit, 64-bit
yte Order:             Little Endian
PU(s):                 8
n-line CPU(s) list:    0-7
hread(s) per core:     2
ore(s) per socket:     4
ocket(s):              1
UMA node(s):           1
endor ID:              GenuineIntel
PU family:             6
odel:                  85
odel name:             Intel(R) Xeon(R) Platinum 8167M CPU @ 2.00GHz
tepping:               4
PU MHz:                1995.309
ogoMIPS:               3990.61
irtualization:         VT-x
```

CPU	**Xeon Plantinum 8167M**
Cores (per CPU)	**4**
RAM	**58gb**
System Architecture	**64 bit**
Master Nodes	**2**
Worker Nodes	**3**
Operating System	**Linux**
Cloud Infrastructure	**Oracle**

3.3 Flow Chart

We created a flow chart of our journey during this project to simplify the understanding. We first moved the data to google drive to be able to download it to the server. Then once on the server we were able to unzip and extract the data and transfer it to hadoop. Using beeline we MapReduce to separate and reduce the data to a size of 3GB, this will allow us to analyze the data in our own personal computers.

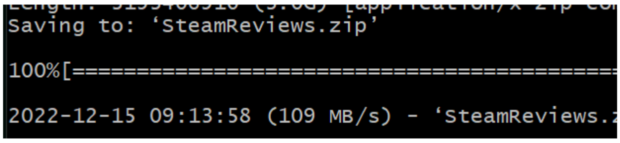

Start
Downloaded 8GB+ Dataset from Kaggle
Uploaded .zip to Google Drive
Download .zip to Oracle Linux Server
Unzip, Check File Integrity, and Upload to HDFS
Cleansed Data & Created Tables in Beeline using HiveQL
Reduced Data to 3GB and Download to Local PC
Extracted Timestamp Column for UNIX to Date/Time Conversion in Excel
Cleaned Dataset to 348MB in Tableau Prep
Ran Analysis via Tableau Desktop
End

After downloading the data, we cleaned and filtered several criteria in tableau, which decreased the size of our data even further. We were then able to perform more specific analysis on the cleansed data.

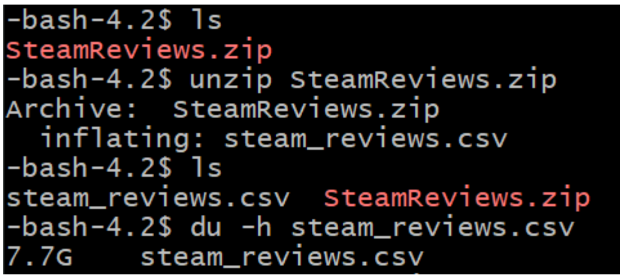
4. Cleaning & Processing

4.1 Hadoop

For our first step, we uploaded the dataset to Oracle Linux Server in order to download it using the Hadoop Distributed File System (HDFS). First thing we did was that we opened our command prompt on Windows or terminal on Apple Mac and used GitBash. Next, we accessed the Oracle Linux Server using the following command prompts, “ssh yourusername@144.24.14.145” and then another line will tell you to enter a password “password: yourpassword” (remember this should be private to the user only, this is only showing what you should type here). Step three is to copy and paste a wget command into the terminal, wait for the download to fully complete, and type the following in order into the Linux Terminal: 1. “ls” -to ensure that the file has been downloaded, 2. “unzip SteamReviews.zip" -to unzip the .zip folder and uncover the .csv within it, 3. “du -h steam_reviews.csv”- to check the size of the file and it should be 7.7GB; 5. “cat steam_reviews.csv | head -5” -(Optional) this can be used to further ensure the integrity of the .csv file. It helped us to check that the file was unzipped correctly and was not corrupted.



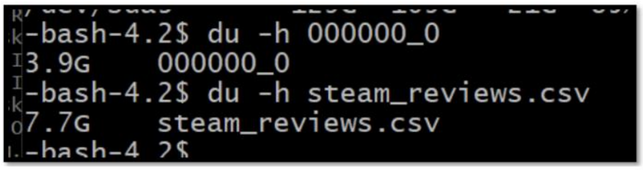
```
Saving to: ‘SteamReviews.zip’

100%[==========================================

2022-12-15 09:13:58 (109 MB/s) - ‘SteamReviews.z
```



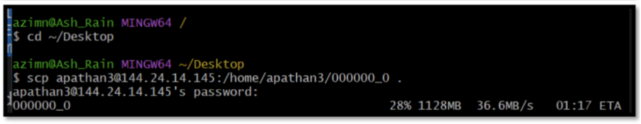
```
-bash-4.2$ ls
SteamReviews.zip
-bash-4.2$ unzip SteamReviews.zip
Archive:  SteamReviews.zip
  inflating: steam_reviews.csv
-bash-4.2$ ls
steam_reviews.csv  SteamReviews.zip
-bash-4.2$ du -h steam_reviews.csv
7.7G    steam_reviews.csv
```

Next step we accomplished was uploading the dataset to the Hadoop Distributed File System. Inorder for us to have done this, we made a directory in our HDFS and then uploaded the dataset file to that specific folder directory. Here is how we did it using the Linux terminal commands: 1. “hdfs dfs -mkdir /user/yourName/SteamReviews”- to make the SteamReviews directory folder in HDFS, 2. “hdfs dfs -mkdir /user/yourName/tmp”- to make the tmp directory folder in HDFS. If it already exists, skip this step, 3. “hdfs dfs -ls” - to ensure that the file has been downloaded, and 4. “hdfsdfs-putsteam_reviews.csv/user/yourName/SteamReviews”- to put the steam_reviews.csv dataset into the newly created folder. Next, we confirmed that the file was successfully uploaded to HDFS by using the command: ”hdfs dfs -ls SteamReviews/” .

Furthermore, we created multiple tables using beeline. The purpose of this was to create a table so that we can visualize the data we are working with. By doing so, it would be easier than using Excel, which actually crashed the application because it was not capable of viewing over 1 million rows normally. What we did next goes as follows: 1. “beeline” -this opens beeline prompt, 2. “create database SteamReviewer;”, 3. “use SteamReviewer;”, “SELECT * FROM SteamReviews LIMIT 3;”, and 4. “SELECT * COUNT (*) FROM SteamReviews”. Once this was done we were able to create and view the tables. Furthermore we cleaned the data using mapreduce and downloaded it into Linux and lastly we downloaded the clean dataset to our local computer. With all this done, we were ready to move onto Tableau to prepare the data visualizations.



```
-bash-4.2$ du -h 000000_0
3.9G    000000_0
-bash-4.2$ du -h steam_reviews.csv
7.7G    steam_reviews.csv
-bash-4.2$
```

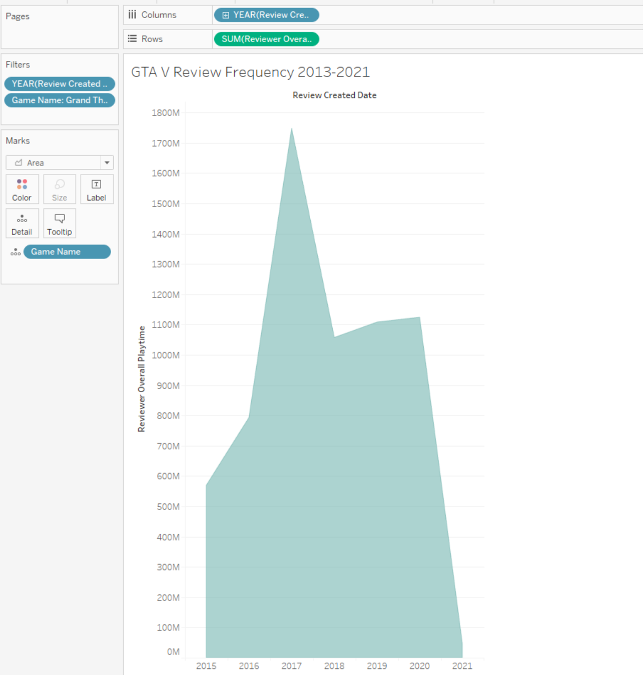


```
azimn@Ash_Rain MINGW64 /
$ cd ~/Desktop

azimn@Ash_Rain MINGW64 ~/Desktop
$ scp apathan3@144.24.14.145:/home/apathan3/000000_0 .
apathan3@144.24.14.145's password:
000000_0                                  28% 1128MB  36.6MB/s   01:17 ETA
```

4.2 Tableau

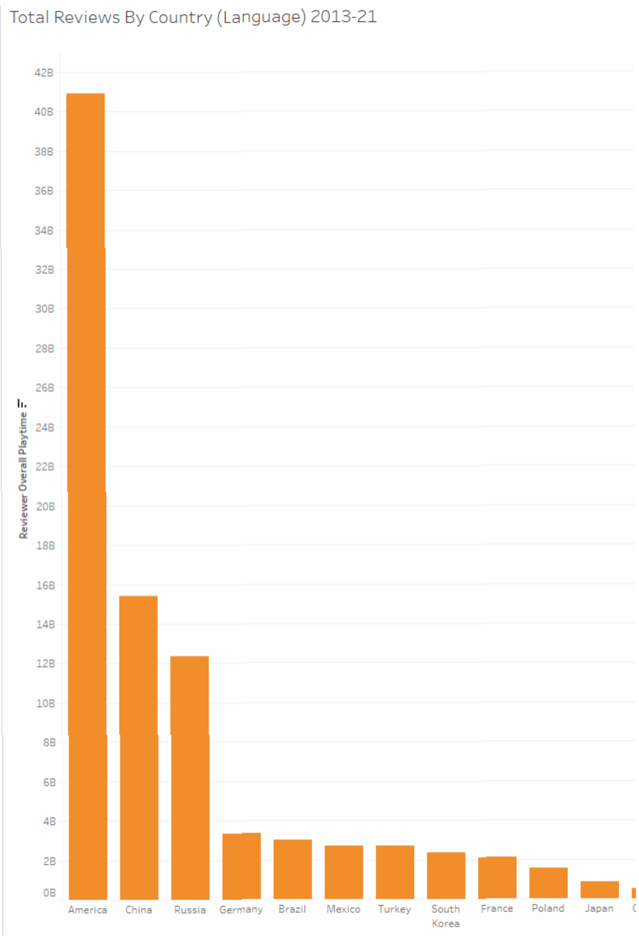
tableau
SOFTWARE

5. Data Analysis

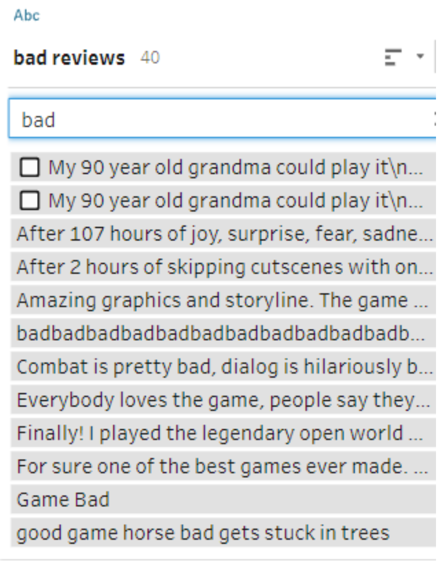


After analyzing the data, we were able to use Geospatial Visualization[1] of the entire dataset and understand that the total views by country from 2013-2021 was that the highest overall playtime was in America, with about 41 billion hours of playtime. Following America was China with about 16 billion, Russia with 12 billion, Germany with over 3 billion and Brazil with 3 billion.

Using temporal analysis from a business perspective on the game Grand Theft Auto V[2] we see that the reviewer's overall playtime was 945,560,343 hours. This playtime data however does not mean much since we are not sure if the feedback from the reviewers were positive or negative. So in order to find out if the game was actually doing well, we could not stop our analysis by just looking at the reviewer's overall playtime, instead we had to dive in deeper using natural language processing. Using natural language processing we can get a better understanding of our data.



Since we had so many reviews, that alone does not mean that the game was successful, we would need to understand if those reviews were positive or negative. For this reason using keywords like: “yes”, “good”, “great”, “amazing” and “bad”, “awful”, “no”. Using these keywords we can categorize each review as either being a positive review or a negative review. This would help us further understand the success of the game since it would show us if the vast amount of reviews were left for good or bad reasons. If there are thousands of reviews but most of them are negative, that does not mean that the game was successful so using temporal analysis was the best method of determining the actual success of the game given the reviews.

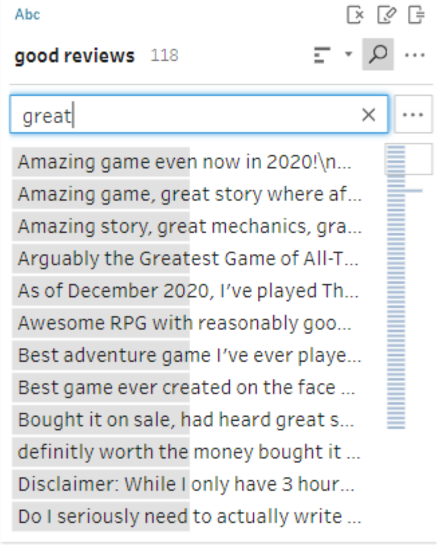


Abc

bad reviews 40

bad

☐ My 90 year old grandma could play it\n...
☐ My 90 year old grandma could play it\n...
After 107 hours of joy, surprise, fear, sadne...
After 2 hours of skipping cutscenes with on...
Amazing graphics and storyline. The game ...
badbadbadbadbadbadbadbadbadbadbadb...
Combat is pretty bad, dialog is hilariously b...
Everybody loves the game, people say they...
Finally! I played the legendary open world ...
For sure one of the best games ever made. ...
Game Bad
good game horse bad gets stuck in trees



Abc

good reviews 118

great

Amazing game even now in 2020!\n...
Amazing game, great story where af...
Amazing story, great mechanics, gra...
Arguably the Greatest Game of All-T...
As of December 2020, I've played Th...
Awesome RPG with reasonably goo...
Best adventure game I've ever playe...
Best game ever created on the face ...
Bought it on sale, had heard great s...
definitly worth the money bought it ...
Disclaimer: While I only have 3 hour...
Do I seriously need to actually write ...

6. Conclusion

To summarize, we took the steam reviews dataset and analyzed it in many ways. We experienced how to download the dataset, upload it to Oracle Linux Server and Hadoop Distributed File System, create tables using Beeline, clean data with MapReduce, and download it back to our local PC. From there we learned how to do incredible analysis on the data using Tableau Prep and Desktop. It was very interesting to gather so much data that could be useful to businesses. For example, we learned that reviewer's overall playtime was the highest in America, followed by China and then Russia. Using this kind of playtime data, developers can gear their development efforts with these larger markets in mind, being able to maximize player satisfaction, company profits, all while reducing costs and unnecessary efforts. Using the Natural Language Processing tool, we were able to further dive into the review text and learn more about consumer feedback.

References

Big Data Analytics. IBM. (n.d.). Retrieved December 18, 2022, from https://www.ibm.com/analytics/big-data-analytics

Big Data Analytics: What it is and why it matters. SAS. (n.d.). Retrieved December 18, 2022, from https://www.sas.com/en_us/insights/analytics/big-data-analytics.html

Team, D. F. (2022, January 27). *Top 15 hadoop analytics tools for 2022 - take a dive into analytics*. DataFlair. Retrieved December 18, 2022, from https://data-flair.training/blogs/hadoop-analytics-tools/

Babu, J. (2020, January 29). *Analyzing big data with Hadoop*. Open Source For You. Retrieved December 18, 2022, from https://www.opensourceforu.com/2018/02/analysing-big-data-hadoop/