CMP 7203 BIG DATA MANAGEMENT

EVALUATION OF BIG DATA PROCESSING PARADIGMS AND ANALYSIS OF "CATCH THE PINK FLAMINGO" GAME

BY

Faisal Ali

STUDENT NO**: 22168118**

MSc BIG DATA ANALYTICS

SUBMITTED MAY 19, 2023

Table of Contents

List of Figures

1. Introduction

Big data refers to a vast and diverse collection of information that organizations can extract and utilize for business purposes through advanced data analytics applications, machine learning, and predictive modeling. This concept has brought about a revolutionary transformation in the modern business landscape, despite being initially regarded as a mere buzzword. The impact of big data on the world has been truly tremendous. The historical roots of data analysis, which have paved the way for today's advanced big data analytics, can be traced back to 17th-century London. It was during this time that John Graunt introduced statistical data analysis while studying the bubonic plague. Fast forward to 1943, and the United Kingdom developed a theoretical computer and one of the earliest data-processing machines, marking significant milestones in data processing. In 2001, Doug Laney of Gartner coined the term " 3Vs" (volume, variety, and velocity) to define the fundamental characteristics of big data. (Phillips, 2021)

Figure 1: How batch processing works (Adejuwon Kehinde David,2018)

Big data is a term used to describe a large amount of data that exhibits the three characteristics known as the three Vs: volume, velocity, and variety. Volume refers to the amount of data, velocity is the speed at which data is generated, and variety refers to the different types of data available. Recently, two more Vs have been added to the definition, which is value and veracity. (OCI, 2023)

It is predicted that the overall quantity of data produced, stored, replicated, and utilized across the world will rise at a swift pace, reaching 64.2 zettabytes in 2020. From 2020 to 2025, the global data generation is anticipated to increase to over 180 zettabytes. In 2020, there was a record-breaking amount of data created and duplicated, with the expansion surpassing prior estimates, primarily due to the COVID-19 pandemic, which resulted in an increase in demand as more individuals worked, learned, and entertained themselves from home. (Statista, 2021)

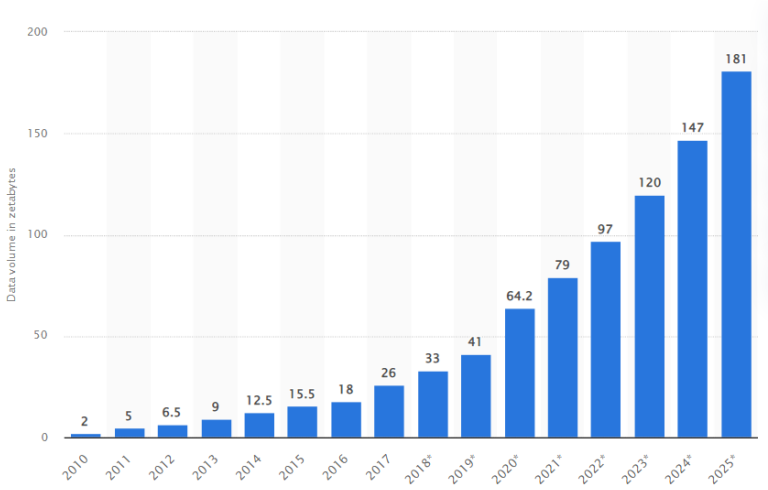

Figure 2: Data growth prediction between 2010 to 2025 (Statista,2021)

2. Big Data Processing Paradigms

The increase in data has resulted in a rising need for methods and tools that can aid in the collection, retention, and handling of these immense data sets. The big data paradigm has arisen as a solution to this problem, incorporating various technologies, methods, and strategies intended to tackle the obstacles associated with managing and analysing extensive and intricate data sets. The typical approaches to processing Big Data that are distinguished by the 3Vs (volume, velocity, and variety) consist of batch processing, real-time processing, and hybrid processing models.

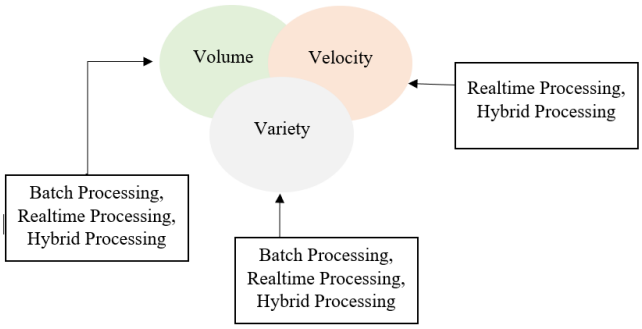

Figure 3: Challenges of big data and solutions (Simplilearn, 2021)

Batch processing is suitable for managing vast amounts of data, whereas real-time processing is better suited for handling rapidly changing data. A combination of batch and real-time processing is useful for managing both large static datasets and dynamic data. All three processing approaches, batch, real-time, and hybrid, are thought to be able to handle diverse types of data.

2.1 Batch Processing Paradigm

Batch processing is an important component of modern data processing systems and is used in a wide range of applications, including data warehousing, data integration, and data analysis. It has a long history that dates back to the early days of computing when computers were mainly used for scientific and engineering purposes. Due to scarce computing resources, batch processing was developed as a way of grouping jobs together to be processed efficiently. The first batch processing system was developed by IBM in the late 1950s, called the IBM 7090 Data Processing System, which processed large amounts of data quickly and efficiently.

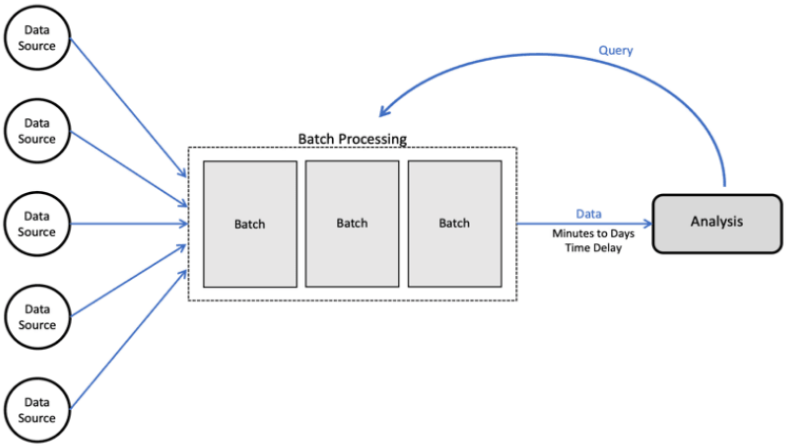

Figure 4: How batch processing works. (M, 2022)

Batch processing is a cost-effective and efficient method used for analyzing large data sets, improving the accuracy of data analysis. It involves collecting, cleaning, transforming, and analyzing structured and unstructured data, such as weblogs, social media data, and streaming data, later in one batch. This approach allows organizations to schedule their computing jobs during off-peak hours and reduce costs, relying on a parallel distributed processing framework, like MapReduce and Apache Spark for scalability. (Casado & Younas, 2014)

It deals with enormous amounts of data for implementing high-volume and repeating data jobs, each of which performs a specific operation without the need for user intervention.

2.2 Real-time/ Streaming Big Data Processing Paradigm:

Real-time processing is a method of computing that enables data to be analyzed and processed immediately as it is produced or received, without any delay. This means that the data is processed as it is being generated, rather than waiting for all of it to be collected before analyzing it. This paradigm is typically employed in applications that require prompt responses to events or actions, like financial transactions, industrial control systems, and online gaming. Real-time processing requires the use of highly optimized algorithms and hardware to ensure that data can be processed with speed and precision.

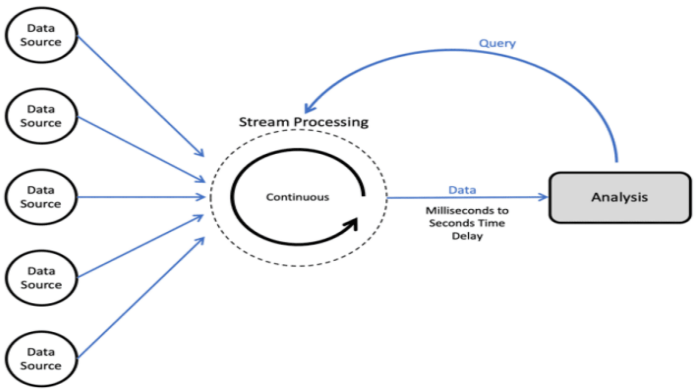

Figure 5: How stream processing works. (Roddewig, 2022)

There are several technologies (TIMOTH´EE DUBUC & ROESCH., 2021) that can be used for real-time stream processing that provide the infrastructure and tools needed to handle large volumes of streaming data and process it in real time with low latency.

2.3 Hybrid Processing Paradigm

Often, there is a requirement for both accuracy and speedy data processing. However, batch processing and stream processing can't meet these two needs simultaneously. Therefore, a hybrid approach that blends both approaches is employed.

In the context of Big Data, a Hybrid Processing Paradigm involves utilizing diverse processing techniques and technologies to efficiently manage large data sets. This method merges the benefits of both batch processing, which processes big data in bulk, and real-time processing, which processes data as it occurs.

Employing a hybrid processing paradigm enables companies to process significant amounts of data quickly and precisely while also instantly responding to any data changes. This can result in better insights and decision-making, a superior customer experience, and more efficient operations.
(Purohit, 2016)

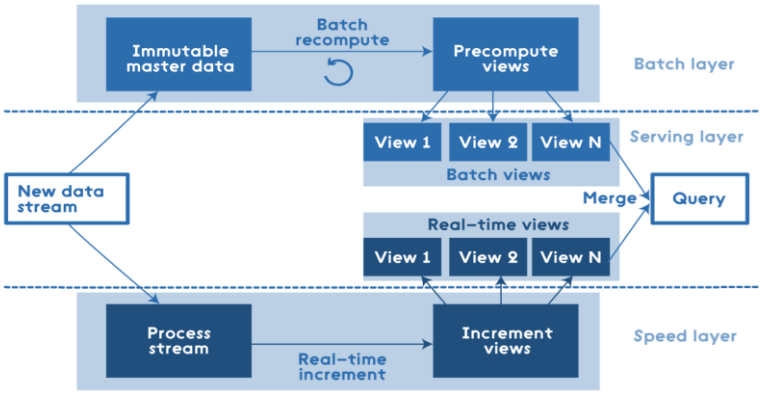

Figure 6: Lambda architecture. (Boulineau, 2018)

Lambda architecture is one of the architecture for hybrid data processing paradigm that integrates both batch and real-time processing methods to create a cohesive perspective of vast data sets. This combination approach empowers companies to handle large volumes of data rapidly and precisely, while promptly responding to data fluctuations in real-time. This enhances data-driven insights and decision-making processes.

3. Exploratory Data Analysis (EDA)

3.1 Flamingo Data Overview

For this analysis, seven different sets of data were collected and examined. These include a database of all the users who played the game (users), every time a user joined a team (team), every session a user played (user-session), information on how many times ads were clicked

(ad clicks), records of purchases made (buy clicks), every click a user made during the game (game clicks), every event that occurred in each level for a team (level events) and every team that exists within the game (team-assignments).

3.2 Data Cleaning:

There are multiple tools available for cleaning data in the field of big data. Some commonly used tools include Apache Spark, Hadoop MapReduce, Apache Pig, Apache Hive, Talend, and IBM InfoSphere DataStage.
In the current project, Spark was chosen as the tool to perform the essential data cleaning operations on the datasets. The initial step was to remove all incomplete rows, then to correctly format the date of birth (DOB) field into age. Furthermore, any duplicate columns across different csv files or datasets were removed. Finally, the various datasets were merged into one comprehensive dataframe, and any duplicate records were removed from it.

3.3 EDA Visualizations

3.3.1 Age Distribution

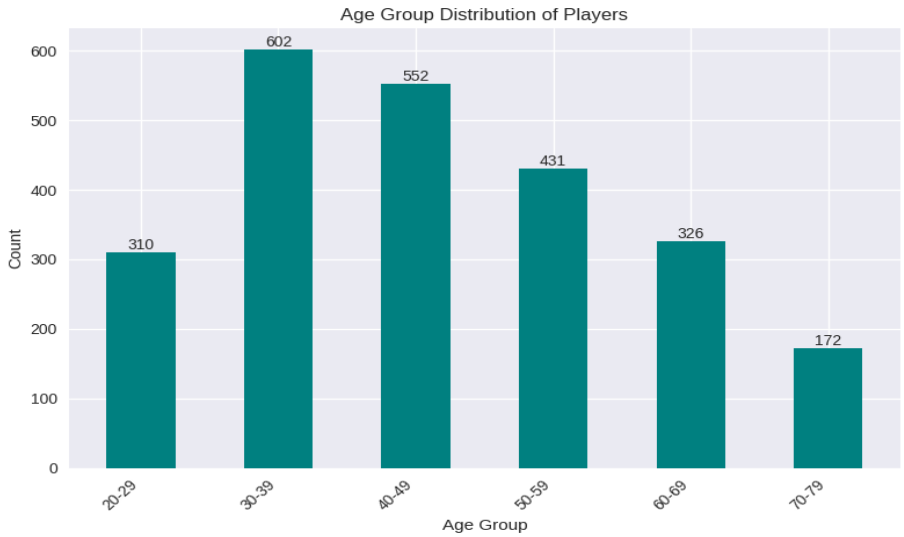

Figure 7: Players belongs to a different age.

The information presented in Figure above reveals that most players fell within the age range of 30 to 39, while there were significantly less player of young age like below 30 and it also seems that it attract less retired people of age above 60. To obtain these results, a new column called “age” was created by calculating the age based on the “DOB” (Date of Birth) column. Subsequently, the users were categorized into different age groups based on this calculation.

3.3.2 Purchased Items by User

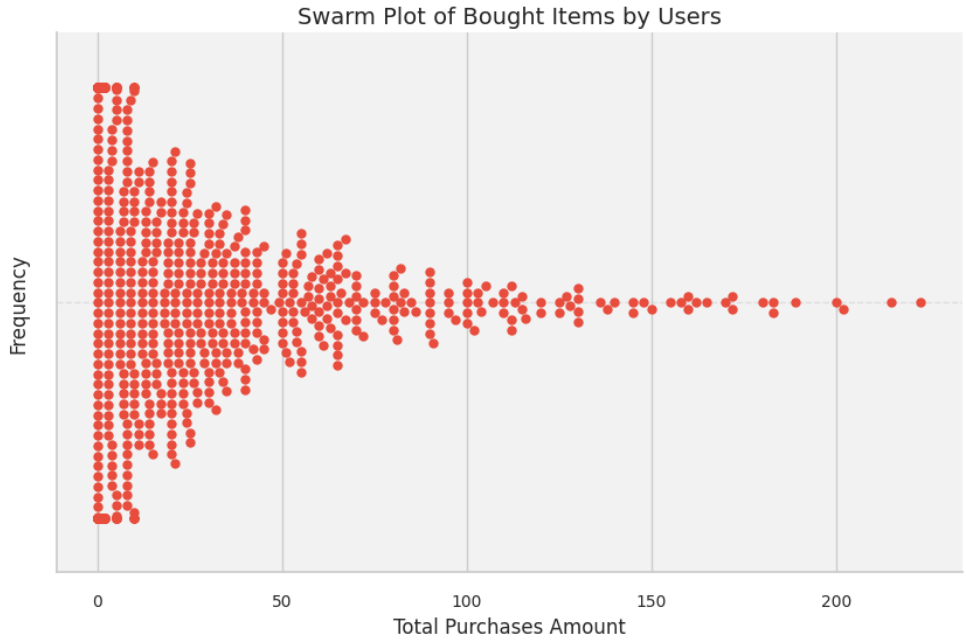

Figure 8: Distribution of purchased items among users

The swarm plot shown above visualizes the distribution of purchased items among users. Each point on the plot represents a user, and its position along the x-axis indicates the total amount they have spent on purchases. This visualization provides a clear representation of the distribution, highlighting that the range between 0 and 30 captures the highest cumulative purchase amounts observed among the users.

3.3.3 Game's Top 3 High-Performing Users

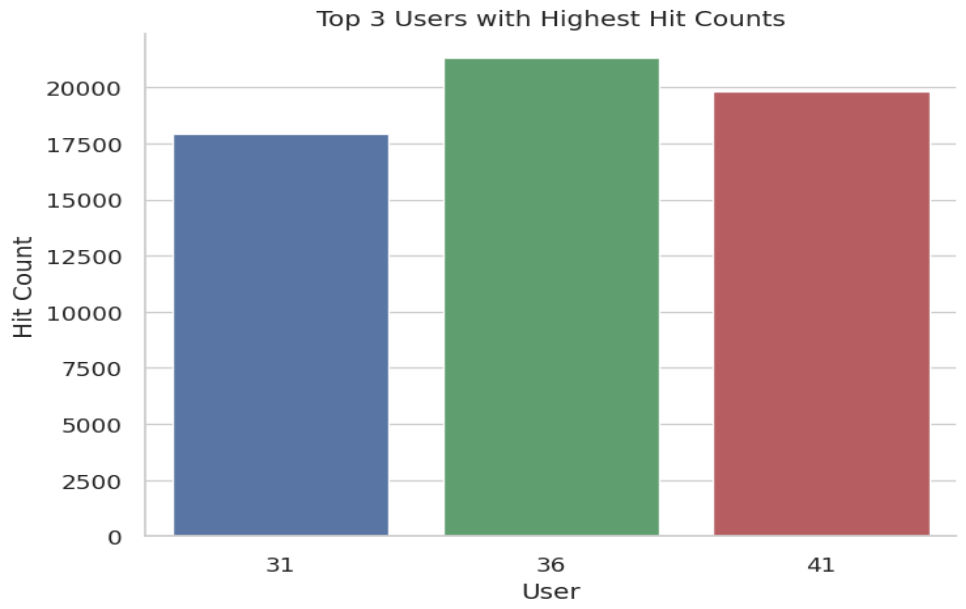

Figure 9: High-performance users with respect to number of Hits.

The graph presented above exhibits the users who performed exceptionally well in the Pink Flamingo game, as measured by their hit counts. It is evident that users with the IDs 36, 41, and 31 achieved the highest number of hits.

3.3.4 Top 3 Teams with Most User Count

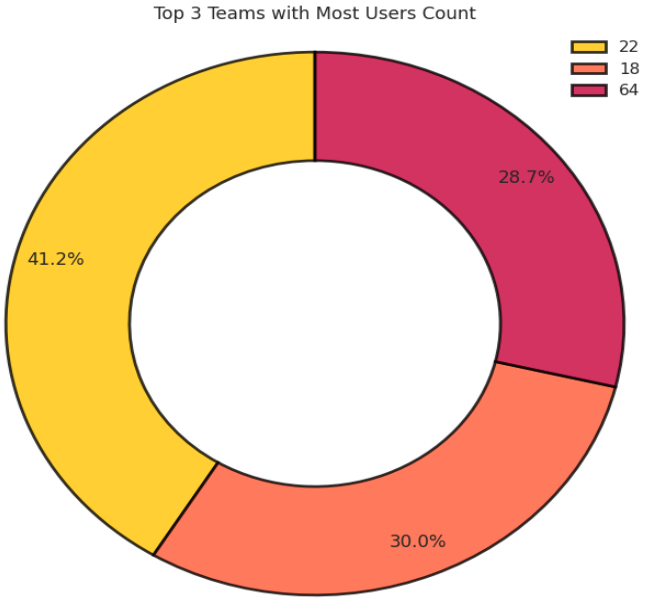

Figure 10: Top teams with most user counts

The pie chart depicted above represents the top three teams with the highest number of users and the maximum number of hits. It is apparent that Team 22 is at 40.8 percent, followed by Team 53 and Team 64.

3.3.5 Different Platforms used by Users

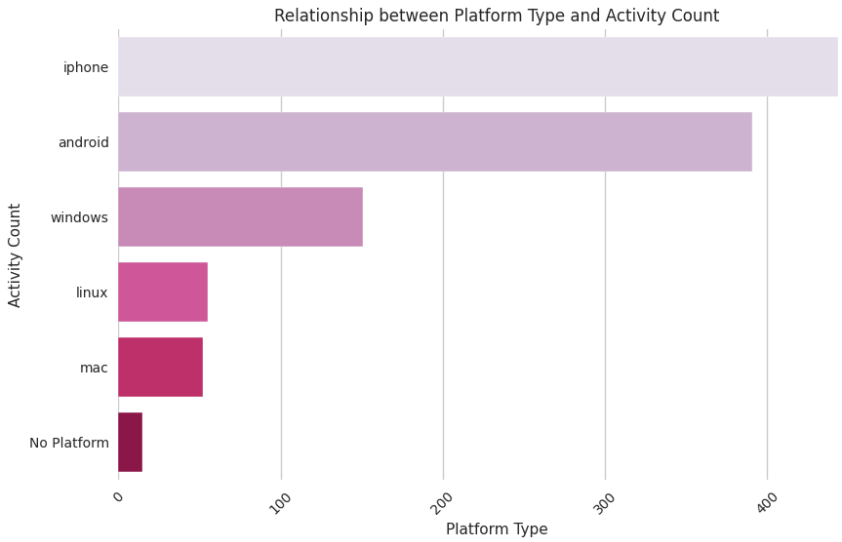

Figure 11: Platform used by the user to play the game.

Based on the figure, it is evident that players utilized various platforms to engage in the game. After conducting Exploratory Data Analysis (EDA), it was revealed that a significant number of individuals who played the Pink Flamingo game preferred using platforms such as iPhone, Android, and Windows.

3.3.6 Total purchase count from different platforms

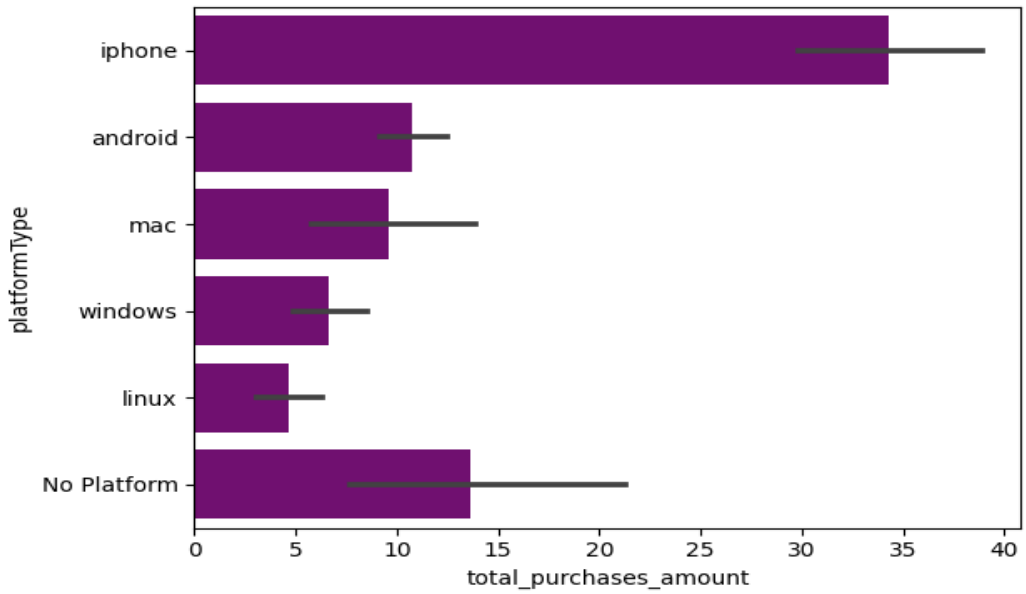

Figure 12: Platform used by the user to perform purchasing.

Based on the figure, it is evident that the most frequently used platform was iPhone. However, when considering spending patterns, Mac users spent more despite being less in number. In terms of duration of use, Mac users exhibited higher spending.

4. Machine Learning Models

4.1 Classification

4.1.1 Naive Bayes on Proposed Dataset

Naive Bayes (Vadapalli, 2022) is selected because it is considered to be the simplest yet highly effective algorithm. It demonstrates remarkable suitability for addressing multiclass prediction problems, which involve categorizing instances into multiple classes. Despite its straightforward nature, Naive Bayes consistently delivers accurate outcomes and performs well on intricate classification tasks. Its ease of implementation and versatility in handling diverse problem domains have contributed to its widespread adoption across various applications.

Initiated the training process for the Naive Bayes classification model by utilizing the merged dataset. To optimize the data representation, employed an assembler object to consolidate all predictor columns into a single composite column. Consequently, the data was converted into a numerical matrix format. The subsequent step involved dividing the dataset into two distinct subsets: the training dataset, which constituted 80 percent of the data, and the testing dataset, which made up the remaining 20 percent. With a predetermined seed value of 3000, the Naive Bayes algorithm was applied to the divided data for prediction purposes.

Finally, an evaluation of the classifier's performance was conducted using a MulticlassClassificationEvaluator, yielding an impressive accuracy score of **0.95**.

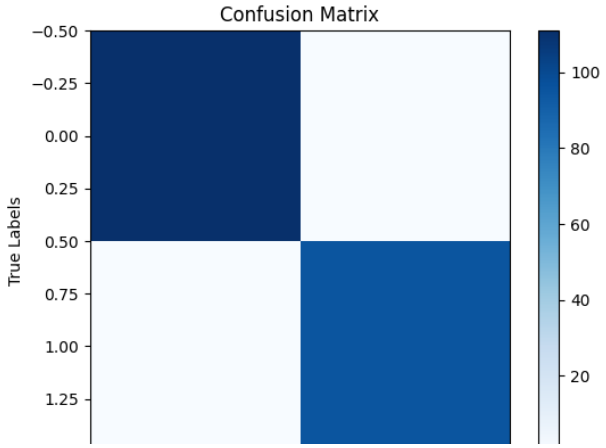

Figure 13: Confusion matrix gives results of naive Bayes.

The presented confusion matrix graph offers a visual depiction of the classification model's performance, providing an overview of its predictions on a test dataset and their comparison with the true labels of the data.

4.2 Clustering

4.2.1 K-Means Clustering on Proposed Dataset

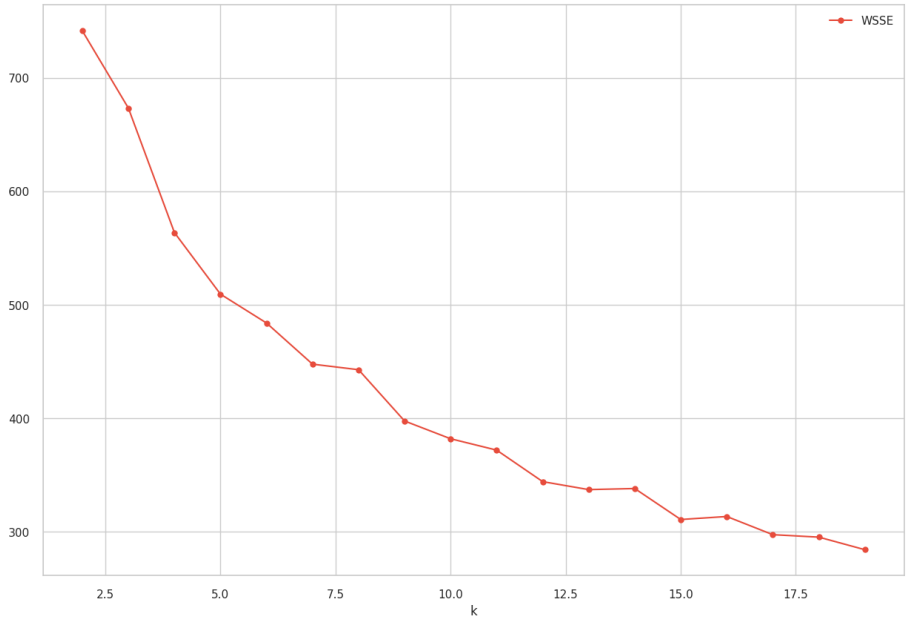

Figure 14: K-Means Algorithm.

To commence the algorithm, two cluster centers were chosen randomly. Each data point was subsequently assigned to its nearest cluster center. The primary objective was to minimize the within-cluster sum of squared distances, also known as WCSS (Within-Cluster Sum of Squares) or WSSE (Within-Cluster Sum of Squared Error). The aim was to identify the most suitable cluster centers that would effectively minimize the overall distance between data points and their respective cluster centers. Initially, 20 clusters were generated, yielding a WSSE value above 700. Notably, as the cluster size increased, the within-cluster sum of squared error exhibited a noticeable decrease.

5. Graph Analysis

Graph analytics, also known as network analysis, involves investigating the connections and associations among entities or nodes within a graph structure. This approach is applicable to various elements such as products, customers, operations, or devices. The growing adoption of graph analytics by businesses worldwide allows them to uncover valuable insights across diverse domains like marketing, fraud detection, supply chain management, and search engine optimization. (Kaley, 2021)

In the context of the "Catch the Pink Flamingo" chat dataset, graph analysis focuses on examining the relationships and patterns within the chat data, represented as a graph or network. The dataset likely contains user and team information,
as well as interactions, organized as nodes and edges.

Various popular tools are available for graph analysis, including NetworkX, Gephi, Neo4j, Cytoscape, Graph-tool, and Giraph. For this report, Neo4j desktop was utilized to conduct the analysis.

Table 1: Pink Flamingo chat dataset for graph analysis

Dataset	Description
chat join team chat.csv	Contains user ID, team, and date.
chat leave team chat.csv	Contains user ID, team, and leave date.
chat mention team chat.csv	Contains chat item, user id, and date.
chat respond team chat.csv	Contains chatid1, chatid2 and date

5.1 Users Join Activities

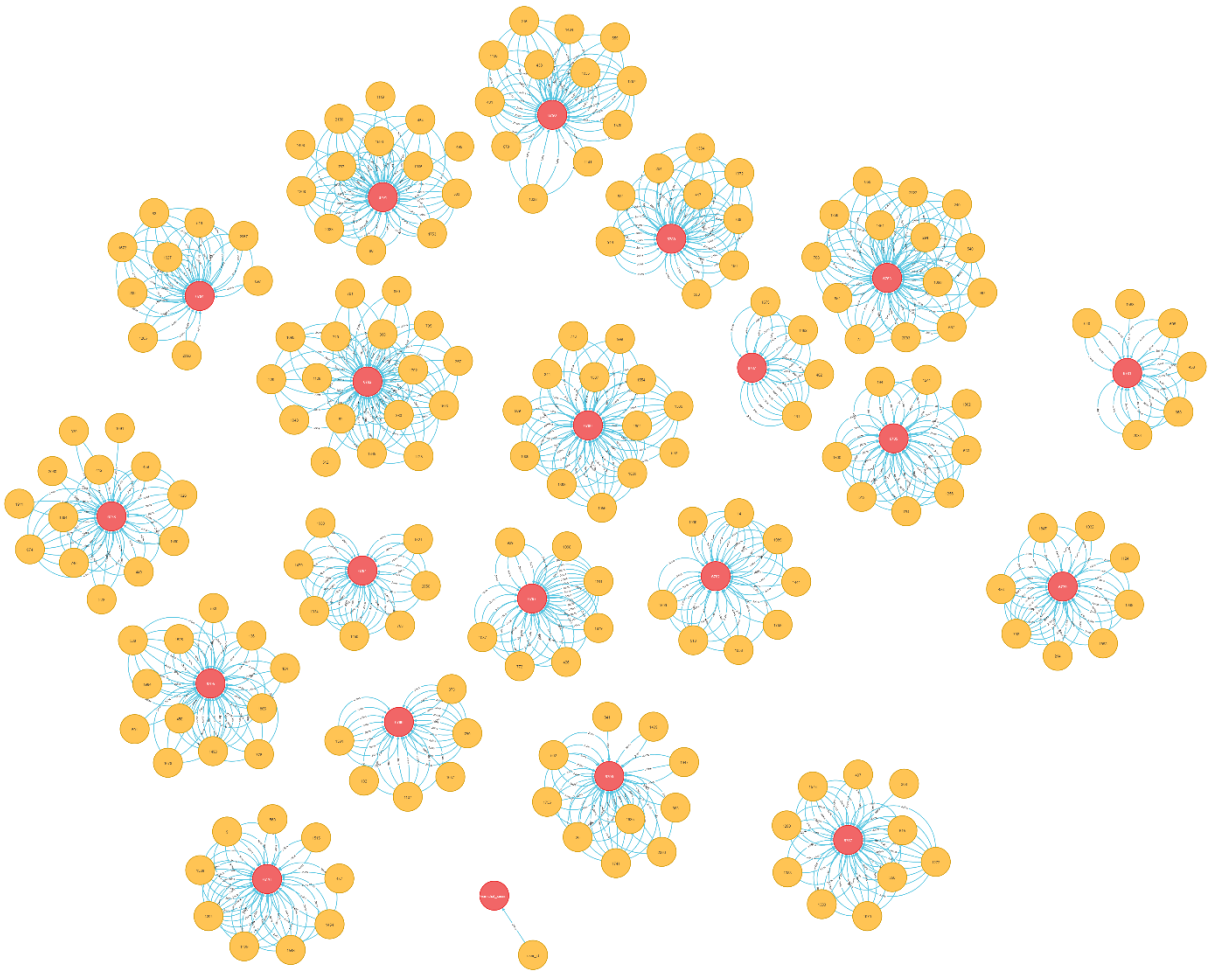

Figure 15: User joins different chats

It seems that team 143 has maximum number of users who join the chat.

5.2 Users Leave

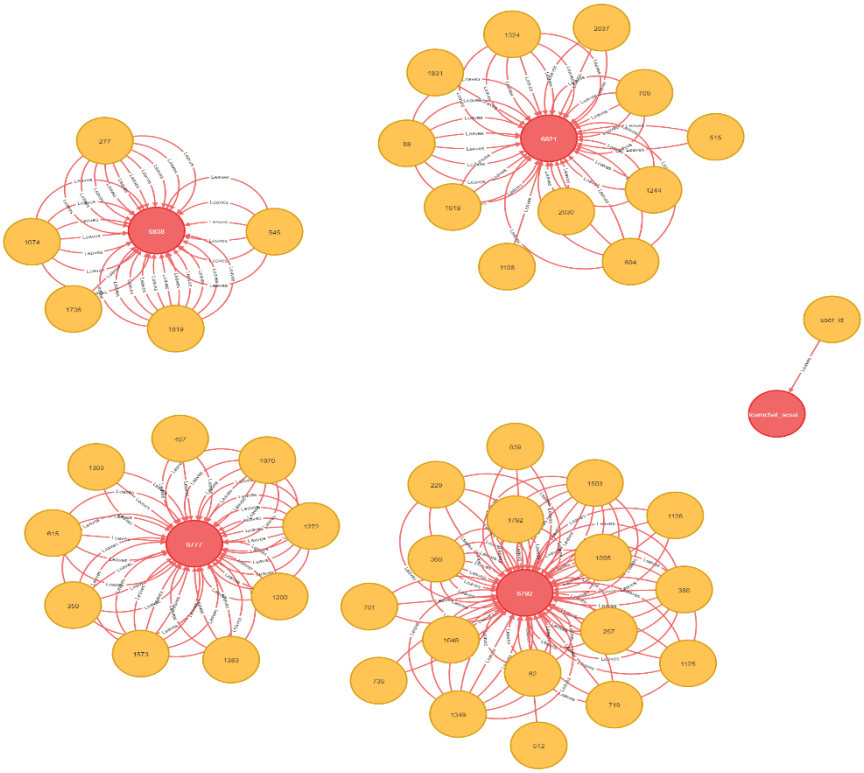

Figure 16: User leaves different chats

5.3 User Mentioned Datset

5.3.1 Number of Chat Mentioned per User

The table shows the top users with the highest mention counts. Such as user ” 131” has the highest mention count of 587,505, followed by users ” 621” and ” 1204” with mention counts of 520,995. This information highlights the active participation and involvement of these users in the chat conversations.

Table 2: Chat Mention Team Chat Analysis

n.user id	mention count
131	587,505
621	520,995
1204	520,995
1506	509,910
1428	509,910
1482	465,570
1450	465,570
283	465,570
674	465,570
88	454,485

Following results also give valuable insights for increasing revenue. Users with high chat mention count, such as those with user IDs ” 131” and ” 621,” can serve as influential players who have the potential to impact others. Targeting these influential users with special

promotions or rewards can drive engagement and boost revenue. Additionally, identifying highly engaged users, like those with user IDs ” 1204” and ” 1506,” allows game developers to understand their preferences and design personalized experiences to enhance user retention and encourage increased spending. Moreover, by examining the conversations around specific features mentioned by users with IDs ” 1428” and ” 1482,” developers can optimize in-game elements to satisfy players and create monetization opportunities.

5.3.2 Visualization of Mentioned Dataset

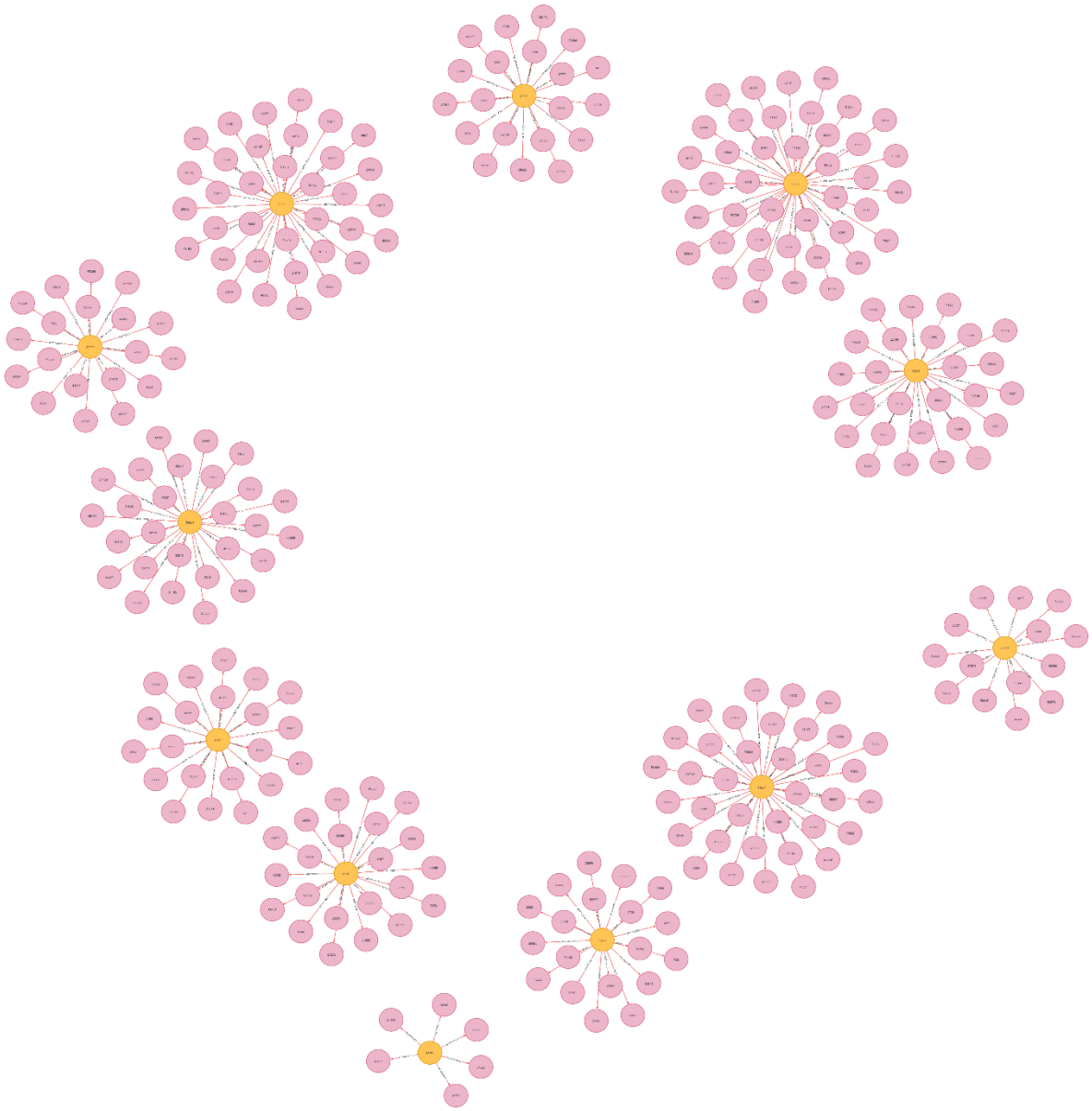

Figure 17: The graph shows the connections between users and

5.4 Influential Chats

Table 3: Top Chats with Response Counts

Chat ID	Response Count
6950	2
6955	2
6967	2
6980	2
6934	2

Following table identifies the top 5 influential chats based on the number of responses they received. Through these results, it is possible to gain insights into popular topics or discussions within the game. Targeting these topics in promotions or special events can help drive user engagement and potentially increase revenue.

5.5 Response Based Chat Connection

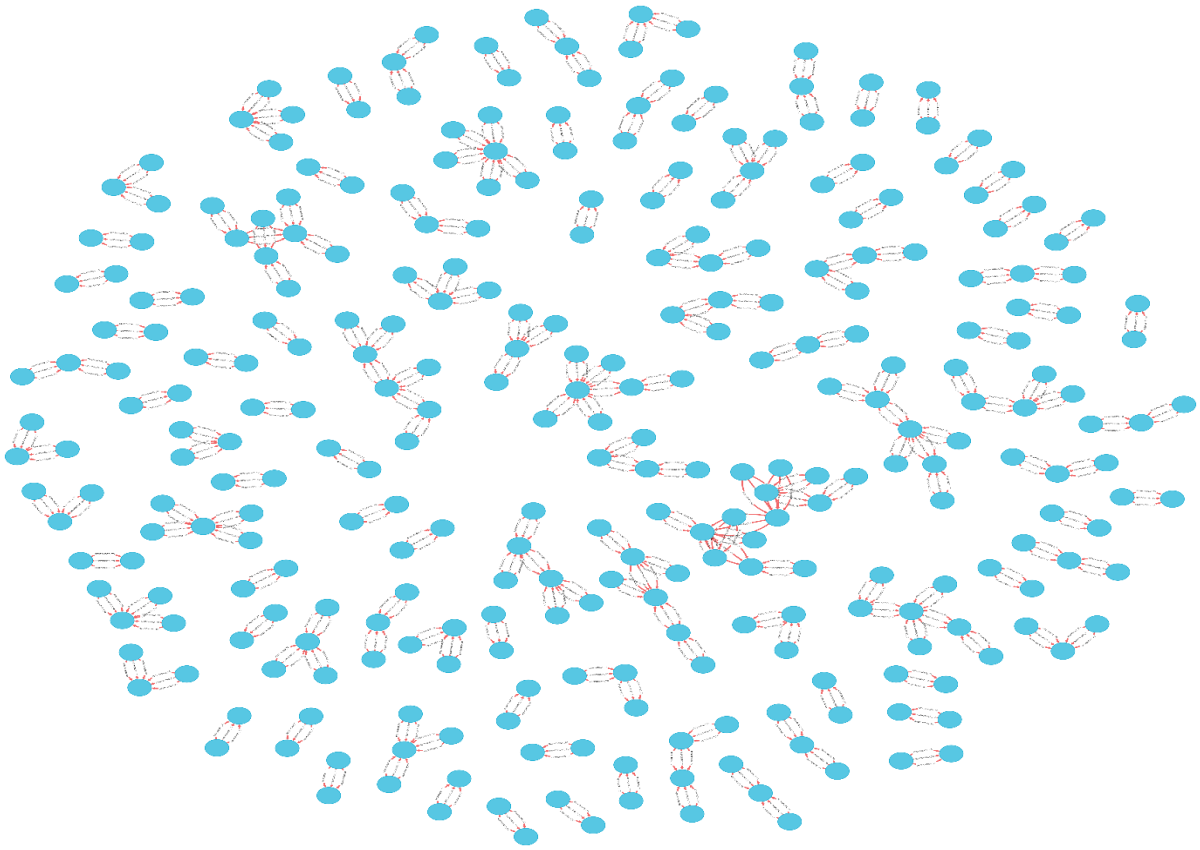

Figure 18: The graph visualization displays the connections between

The graph analysis uncovers valuable insights from chat interactions represented as nodes and their relationships that can enable the improvement of user engagement, identification.of popular discussions, and revenue generation in the “Catch the Pink Flamingo” game.

6. Big Data Ethics

Data ethics pertains to the ethical concerns linked to practices involving data that have the potential to harm individuals. It encompasses all stages of data, ranging from its creation, collection, analysis, to its distribution. Additionally, it ensures that individuals who use the internet give their consent for their data to be shared, and that organizations comply with relevant regulations and privacy laws, such as the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), the California Consumer Privacy Act (CCPA), and the Payment Card Industry Data Security Standard (PCI DSS).(PUR, 2023)

Given the increasing occurrence of cyberattacks, online users are becoming more cognizant of the issues surrounding the privacy of their data. Data ethics aids organizations in effectively managing the risks associated with data privacy, enabling them to enhance user experiences while upholding the confidentiality of personal data.

6.1 Data Storage and Security Ethics

When gathering information from individuals, it is crucial to guarantee the secure handling and storage of the collected data in alignment with legal frameworks governing data protection, organizational information governance obligations, and ethical and governance standards for research. The types of data that should be considered for storage encompass clinical recordings and research data. (PARKER,2019)

6.2 Data Sharing and Transfer Ethics

In certain instances, research data involving individuals cannot be publicly disclosed due to the potential risk of violating privacy. However, there exist ethical and legal avenues for sharing even highly sensitive information. Maintaining transparent communication with research participants is vital, ensuring they are fully informed about how their data will be utilized and shared both in the short and long term. It is important to emphasize that non-response from participants should not be interpreted as implied consent. Data sharing must always adhere to the terms outlined in the participant’s consent agreement.

If consent has not been obtained or if participants have explicitly withheld consent, data should only be shared if appropriate anonymization methods are employed. Stringent data protection laws, such as the European Union’ s General Data Protection Regulation (GDPR), establish guidelines for collecting, storing, and sharing personal information, clearly defining

acceptable practices and prohibited actions. Strict adherence to all relevant data legislation is essential.

In cases where someone does not own the research data, obtaining written permission from the data owner is imperative before publishing or sharing the data. (Francis, 2023)

6.3 Data Processing Ethics

Big data processing ethics involves the ethical factors and principles linked to the management, examination, and application of substantial amounts of data. With the ongoing advancements in big data technologies, it is imperative to acknowledge the ethical consequences tied to the gathering, processing, and utilization of extensive data sets.

Essential elements of big data processing ethics encompass guaranteeing the precision, dependability, and soundness of the data employed in big data processing. Additionally, it involves establishing robust security measures to preserve the confidentiality, integrity, and accessibility of the data, guarding against unauthorized access, breaches, and data leaks. (McCabe, 2023)

7. Conclusion

This report explores the historical evolution of Big Data and its exponential growth over time. It investigates the significance of various paradigms in handling big data and addresses the challenges associated with storing, processing, and retrieving vast amounts of information. The report also delves into the ethical considerations that arise from the storage and processing of big data, as well as the potential impact of new regulations that may profoundly impact the management of big data. Moreover, the report introduces a Big Data solution that utilizes Spark for Eglence Inc.'s fictional game, "Catch the Pink Flamingo". This solution encompasses conducting exploratory data analysis, employing machine learning techniques, and applying graph analysis.

8. Finding and Recommendation

- Following the exploratory data analysis (EDA), it was observed that a majority of the players fall within the age range of 30 to 39, while there is a notable scarcity of teenagers and individuals above the age of 59. It became evident that catching the pink flamingo is not excessively challenging; however, in order to engage a wider audience, particularly young individuals and retirees seeking leisure activities, it is essential to introduce more appealing features to the game.

- Based on the analysis of the swarm plot, it is evident that most users tend to spend less than 100 on in-game purchases. To cater to user preferences and enhance the overall experience, it is recommended to consider reducing the prices of various in-game items. items such as binoculars for spotting mission-specific flamingos, or special flamingos that yield higher grid points. By adjusting the pricing strategy for these in-

game purchases, the game can attract more users and encourage increased engagement.

- It has been discovered that there is currently a limited number of users who spend a significant amount of time in the game and achieve maximum hits. Since this game has no endpoint and continually presents more complex levels, Eglence Inc. faces the challenge of maintaining the game's appeal and keeping long time players engaged. To tackle this situation, Eglence Inc. must employ targeted strategies and incorporate features that specifically cater to the unique needs and interests of the player.

- The game was found to be accessed through various platforms, with Mac being less frequently used, while iPhone emerged as the most popular platform. To broaden the game's accessibility and cater to a wider user base, it is essential to incorporate the features available on iPhone into the Mac platform. By implementing this enhancement, players will have the option to enjoy the game on their Mac devices, thereby expanding their gaming experience beyond just the iPhone.

- Valuable insights from graph analysis indicate revenue growth potential by leveraging users with high chat mention counts, like "131" and "621," who hold influence. Targeting these influential users with promotions boosts engagement and revenue. Identifying highly engaged users, such as "1204" and "1506," allows for personalized experiences that enhance retention and spending. Analyzing conversations of users "1428" and "1482" optimizes in-game elements, satisfying players and creating monetization opportunities. This analysis maximizes engagement, revenue, and overall success in the gaming community.

- By conducting response-based chat connection analysis, significant insights can be uncovered from the chat interactions depicted as nodes and their relationships. These insights have the potential to enhance user engagement, identify popular discussions, and generate revenue within the "Catch the Pink Flamingo" game.

References

What is big data? Oracle, 2023. URL https://www.oracle.com/uk/big-data/what-is-big-data/#defined.

What is data ethics and how can storage improve ethics best practices? PURESTORAGE, 2023. URL https://www.purestorage.com/knowledge/what-is-data-ethics.html.

Adejuwon Kehinde David, J. O. F. Governance in the digitalera: An assessment of the effectiveness of big data onemergency management inlagos state, nigeria. Research-Gate, 2018. URL https://www.researchgate.net/publication/329277423_Governance_in_the_Digital_Era_An_Assessment_of_the_Effectiveness_of_Big_Data_on_Emergency_Management_in_Lagos_State_Nigeria.

Boulineau, J. Lambda architecture. jonboulineau,2018. URL https://jonboulineau.me/blog/architecture/lambda-architecture

Casado, R. and Younas, M. Emerging trends and technologies in big data processing. https://www.researchgate.net/publication/266373455_Emerging_trends_and_technologies_in_big_data_processing
,2014. [Online; accessed 4 April 2023].

Francis, T. . Data storage, information governance, and ethics. AUTHOR SERVICES, 2023. URL https://authorservices.taylorandfrancis.com/data-sharing/data-sharing-ethics/.

Kaley, A. What is graph analytics? Medium,2021. URL https://medium.com/swlh/what-is-graph-analytics-9223d71c26d8.

M, I. Apache spark batch processing: 5 easy steps. HEVO,2022. URL https://hevodata.com/learn/spark-batch-processing/.

McCabe, A. The ethics of big data. HURREE,2023. URL https://blog.hurree.co/the-ethics-of-big-data.

PARKER, R. Data storage, information governance,and ethics. WP, 2019. URL https://wp.lancs.ac.uk/dclinpsy/data-storage-information-governance-and-ethics/.

Phillips, A. A history and timeline of big data.TechTarget WhatIs.com, 2021. URL https://www.techtarget.com/whatis/feature/A-history-and-timeline-of-big-data.

Purohit, J. Hybrid data processing model for big data ‐a review. https://www.ijltet.org/journal/149656699651.1777.pdf
, 2016. [Online; accessed 4-April-2023].

Roddewig, S. What is a data pipeline? everything you need to know. HubSpot, 2022. URL https://blog.hubspot.com/website/data-pipeline.

Simplilearn. Challenges of big data. Simplilearn, 2021. URL https://www.simplilearn.com/challenges-of-big-data-article

Statista. Worldwide data created from 2010 to 2020, 2021. URL https://www.statista.com/statistics/871513/worldwide-data-created/ . [Accessed on April 4, 2023].

TIMOTH´EE DUBUC, F. S. and ROESCH., E. B. Mapping the big data landscape: Technologies, platforms and paradigms for real-time analytics of data streams. https://centaur.reading.ac.uk/95419/9/09300187.pdf , 2021. [Online; accessed 4-April-2023].

Vadapalli, P. Naive bayes explained: Function, advantages disadvantages, applications in 2023. upGrad,2022. URL https://www.upgrad.com/blog/naive-bayes-explained/.

A. Appendix:

A.1. Code Available at Github

https://github.com/faisal123-ali/Big-Data-Management-Assessment.git