**CMP 7203 BIG DATA MANAGEMENT**

**EVALUATION OF BIG DATA PROCESSING PARADIGMS AND ANALYSIS OF “CATCH THE PINK FLAMINGO” GAME**

**BY**

**Meera Mohan Mullamkuzhy**

STUDENT NO**: 22185279**

MSc BIG DATA ANALYTICS

**A picture containing graphical user interface

Description automatically generated**

**SUBMITTED ON MAY 19, 2022**

Table of Contents

[1. Introduction 4](#_Toc135395565)

[2. Big Data Processing Paradigms 5](#_Toc135395566)

[2.1 Big Data Background 5](#_Toc135395567)

[2.2 Batch Processing Paradigm 6](#_Toc135395568)

[2.3 Real-time Processing Paradigm 7](#_Toc135395569)

[2.4 Hybrid Processing Paradigm 8](#_Toc135395570)

[2.5 Comparison of the three paradigms 8](#_Toc135395571)

[3. Exploratory Data Analysis 9](#_Toc135395572)

[3.1 Flamingo Data Overview 9](#_Toc135395573)

[3.2 Data Pre-Processing 10](#_Toc135395574)

[3.3 EDA Visualizations 10](#_Toc135395575)

[3.3.1. Popular User Platform 10](#_Toc135395576)

[3.3.2 Age Group of the Players 11](#_Toc135395577)

[3.3.3 Top Spending Team 11](#_Toc135395578)

[3.3.4 Comparison between Team Total Spends and Team Strength 12](#_Toc135395579)

[3.3.5 Countries with Highest Spending Players 12](#_Toc135395580)

[3.3.6 Highest Revenue Generating Item 13](#_Toc135395581)

[3.3.7 Number of times an item is purchased 13](#_Toc135395582)

[3.3.8 Amount spend by top 10 users 14](#_Toc135395583)

[3.3.9 Time series of in-game ad clicks 14](#_Toc135395584)

[3.3.10 Average Team Join 15](#_Toc135395585)

[3.3.11 Users who never bought any item 16](#_Toc135395586)

[4. Machine Learning Models 16](#_Toc135395587)

[4.1 Classification or Supervised Learning 16](#_Toc135395588)

[4.1.1 Logistic Regression 16](#_Toc135395589)

[4.1.2 Decision Tree 17](#_Toc135395590)

[4.1.3 Comparison of Logistic Regression and Decision Tree 18](#_Toc135395591)

[4.2 Clustering or Unsupervised Learning 19](#_Toc135395592)

[4.2.1 K Means Clustering 19](#_Toc135395593)

[4.2.2 Hierarchical Clustering 20](#_Toc135395594)

[5. Graph Analysis 21](#_Toc135395595)

[5.1. Players JOIN the Chat 22](#_Toc135395596)

[5.2. Players LEAVE the Chat 23](#_Toc135395597)

[5.3. Players MENTIONED in the Chat 24](#_Toc135395598)

[5.4. Players RESPONDS to the Chat 25](#_Toc135395599)

[5.5. Identifying Long Conversations in the Chat 26](#_Toc135395600)

[5.6. Identifying Active Teams 27](#_Toc135395601)

[5.7. Identifying Active Users 28](#_Toc135395602)

[6. Big Data Ethics 28](#_Toc135395603)

[6.1 Ethics in Big Data Collection 29](#_Toc135395604)

[6.2 Ethics in Big Data Storage 29](#_Toc135395605)

[6.3 Ethics in Big Data Processing 30](#_Toc135395606)

[7. Conclusion: Finding and Recommendation 30](#_Toc135395607)

[Reference 31](#_Toc135395608)

[Appendix 34](#_Toc135395609)

# Introduction

Big data is the core of contemporary science and businesses. By 2003, there were 5 exabytes of data produced, and by 2015, there were 8 zettabytes (Sagiroglu and Sinanc, 2013). The data generated from social media, sensors, mobiles, click streams, science data, and other applications grew massively and became difficult to store, manage, process, and visualise. This problem can be handled with the capabilities that big data management offers.

Figure 1 shows the Vs of Big Data. Big data is characterised by three important components: volume, velocity, and variety. “Volume” indicates the dealing with large amounts of data. “Velocity” refers to the come fast the data is coming in.

A picture containing text, screenshot, circle, diagram

Description automatically generated

Figure : Important Vs of Big Data (Niculescu, 2020).

The other characteristics of Big Data are “Variability”, “Veracity”, “Visualisation”, and “Value”. “Variability” refers to the continuous change in data, “veracity” ensures the accuracy of the data, “visualisation” refers to how the data is presented for decision-making purposes, and “value” ensures that every user understand that there needs some value for efforts and resources spent on the aforementioned Vs.

This study includes learning the Big Data paradigms and the ethics in Big Data. Exploratory Data Analysis, machine learning, and graph analysis is carried out on the dataset of an imaginary game “Catch the Pink Flamingo”.

# Big Data Processing Paradigms

Big Data is the result of several technological advancements such as mobile, communication, and cloud services. Nowadays, Big Data analytics is gaining popularity among organisations to achieve insightful information to improve the growth of their business. Big data is not a new concept. It has existed for a long time. Big Data deals with data in different volumes and velocities. This section discusses about the history and processing paradigms of Big Data.

## Big Data Background

In recent years, big data has radically changed the nature of modern business. Machine learning, predictive modelling, and other cutting-edge data analytics tools can be used by organisations to mine big data for commercial goals.

The history of big data is ‘big’. John, who introduced statistical analysis for the plague in 1663, laid the foundation for big data. The beginning of big data can be dated to the 1990s, the internet era. In the era of the internet, Morris and Truskowski discussed digital data storage in 1996, the Google domain was registered in 1997, and Carlo developed NoSQL in 1998. In 2001, Doug Laney introduced the characteristics of big data as 3 Vs – Volume, Velocity, and Variety. Since then, more Vs, such as Veracity, Value, and Variability, have been added to the list. Figure 2 shows the possible solutions for big data issues. In 2005, Mike Cafarella and Doug Cutting developed Apache Hadoop, an open-source framework for storing and handling massive amounts of data. The core components of Hadoop are Hadoop distributed File system, YARN – Hadoop’s cluster resource manager, MapReduce, and Hadoop Common. In 2006, Amazon began to provide cloud computing services. Squeezing value out of data becomes essential as businesses deal with a period of economic volatility and uncertainty because of the Great Recession. Business intelligence, according to Gartner, was CIOs' top concern in 2009 (Phillips, 2021).

*A picture containing text, screenshot, font, diagram

Description automatically generated*

Figure : Possible solutions for big data issues (Casado and Younas, 2015).

In the 1970s, the use of relational databases and structured query language to extract data improved. Social media analytics and the management of unstructured data have become increasingly popular since the early 2000s because of the widespread use of social media. Data from sensors and mobile devices is the focus of the current big data phase (Big Data Framework, 2023).

Big data processing is essential for a firm to expand in the modern world. Big data processing paradigms can be categorised into three types, comprising, batch, real-time, and hybrid processing.

## Batch Processing Paradigm

Big data processing is an organised approach to handle large amounts of data. In this model, transactions happen over a certain period. The most popular framework for batch processing was Hadoop. Separate transactions are used for ingestion, processing, analytics, and reporting in batch processing. This paradigm involves ingesting, storing, and processing data. Analysis is done before producing the batch results(Gurcan and Berigel, 2018) .

The MapReduce framework, first developed by Google, is the most renowned solution for big data processing. Later, the open-source Apache MapReduce framework gained popularity. Hadoop MapReduce and Hadoop YARN as execution engines, the Hadoop Distributed File System, and HBase an alternative for BigTable (Shahrivari, 2014). Figure 3 shows the execution of a MapReduce program.

A diagram of a work flow

Description automatically generated with medium confidence

Figure : Execution of a MapReduce Program (Shahrivari, 2004)

Iterative runs are necessary for many analysis jobs in real-time systems, which MapReduce finds challenging to manage. Although batches can take longer to finish, the batch processing paradigm is more dependable. They are therefore unsuitable for low latency applications. Furthermore, if fresh data come in, batches cannot be stopped or immediately altered (Casado and Younas, 2015). Xhafa et al. (2015) also mention that although the MapReduce framework has been successful in processing Big Data, its batch mode processing demonstrates limits in handling Big Data Streams.

## Real-time Processing Paradigm

It is almost impossible to regulate the frequency of infinite amount of data produced from hardware sensors, servers, mobile devices, and other applications. Instantaneous handling of data streams is required for big data applications. Continuous data ingestion, processing, analysis, and reporting are the components of real-time processing (Gurcan and Berigel, 2018).

Gurcan and Berigel (2018) mention that low latency is supported throughout the process in real-time processing paradigm. Casado and Younas (2015) point out that paradigm analyses small data sets in memory rather than on secondary storages to achieve minimal latency.

Several frameworks like Storm, Spark, S4, Flink, and Samza, are available to support real-time data processing (Gurcan and Berigel, 2018). *‘Storm’* is a distributed real-time system for handling massive amounts of high velocity data. Apache's open-source project ‘*Kafka’* offers high availability, high scalability, and low latency. *‘Flume’* can serve as the real-time processing system's event backbone. A cloud-based solution for processing data in real-time is ‘*Amazon Kinesis’* (Buyya et al., 2016).

Figure 4 shows the life cycle of a real-time processing paradigm.

A picture containing text, screenshot, font, rectangle

Description automatically generated

Figure : Life cycle of real-time processing big data paradigm (Gurcan and Berigel, 2018).

Real-time processing of big data is necessary for numerous applications. Examples of real-time processing applications include banking systems, social networks, internet of things, radar systems, and smart cities.

## Hybrid Processing Paradigm

Hybrid processing paradigm is required big data applications that require both batch and real-time processing. In some scenarios, real-time applications restrict processing and analysis of data for a low response time. Hybrid paradigm was suggested to resolve this issue. The results obtained from batch and real time processing are analysed and combined to achieve the expected output. Apache Flink and Apache Spark are the common frameworks for hybrid processing (Gurcan and Berigel, 2018). The realisation that no single analytic platform is always best for all the requirements made hybrid model more dominant (Purohit, n.d.).

Hybrid processing was first introduced by Lambda architecture by Nathan Marz and James Warren in 2015 (Warren and Marz, 2015). Pandya et al. (n.d.) also mention about the Lambda architecture for hybrid paradigm. Lambda architecture as shown in Figure 5 has three major components. Batch layer manages the master dataset, and pre-compute query functions, called batch views. By indexing the batch views, serving layer provides low latency. Speed layer deals with recent data using fast algorithms (Purohit, n.d.).

A picture containing text, screenshot, font, line

Description automatically generated

Figure : Lambda Architecture (Purohit, n. Lambda Architecture d.)

Business agility, high availability, flexible scaling, and absence of server management, though maintenance of software is required, are the advantages of hybrid processing paradigm. High complexity of managing separate code bases for batch and streaming layers makes Lambda architecture challenging (databricks, n.d.).

## Comparison of the three paradigms

Batch processing deals with ‘Volume’ issue of big data, whereas real-time processing deals with the ‘Velocity’ issue. Hybrid processing paradigm resolves both the ‘Volume’ and ‘Velocity’ issues of big data. The characteristic ‘Variety’ is common in three paradigms (Casado and Younas, 2015).

Table 1 compares the three big data processing paradigms.

|  |  |  |  |
| --- | --- | --- | --- |
| **Features** | **Batch Processing Paradigm** | **Real-time processing paradigm** | **Hybrid processing paradigm** |
| Deals with | Volume | Velocity | Volume and Velocity |
| Data type | Static data | Data streams | Both historical and real-time data |
| Processing pattern | Data processed in a specific sequence. | Data processed immediately as it is received. | Combination of both. |
| Latency | High | Low | Low |
| Data storage | Requires large storage | Minimal storage requirements | Minimal storage requirements |
| Data Volume | Supports large volume of data | Ideal for small volume of data | Handles large volume of data |
| Cost | Affordable for small setups and expensive at scale. | Affordable when well optimised. Cost of large queries are avoided. | Affordable. |

Table : Comparison of three big data processing paradigm

# Exploratory Data Analysis

## Flamingo Data Overview

The goal of this study is to generate revenue for the fictitious game "Catch the Pink Flamingo." The dataset comprises of 4 files corresponding to game conversation data and 8 files concerning game-specific information, in-app purchases, and ad clicks. Table 2 provides information on game-specific files. In section 5, the specifics of the datasets connected to chat are covered in detail.

|  |  |
| --- | --- |
| **Dataset** | **Description** |
| ad-clicks.csv | Database of clicks on ads |
| buy-clicks.csv | Database of purchases. |
| game-clicks.csv | A record of each clicks a user performed during the game. |
| level-events.csv | A record of each level event for a team. Level events are recorded when a team ends or begins a new level |
| team-assignments.csv | A record of each time a user joins a team. |
| team.csv | A record of each team in the game. |
| user-session.csv | A record of each session a user plays.  When a team levels up, each current user session ends, and a new session begins with the new level. |
| users.csv | Database of the game users |
| combined-data.csv | Combined record of all user information. |

Table : Description of the game specific datasets in Flamingo dataset.

## Data Pre-Processing

On the datasets, various pre-processing approaches were used. The datasets were examined for NULL or NaN values. The existence of the duplication in the data was verified.

A new column names “age group” was generated using the “date of birth column” from the “users.csv” file. The datasets “teams.csv” and “buy-clicks.csv” were joined to find the total spend.

## EDA Visualizations

This section includes the results of different statistical visualisation techniques employed using PySpark to gain underlying insights about the data.

### 3.3.1. Popular User Platform

The *'platformType'* used by each user for the game is stored in the file *'user-session.csv'*. This data is used to determine the most popular gaming platform. Figure 6 makes it evident that most users chose *‘iPhone’* for the game. Users of *‘iPhone’* make up 41.9% of all users, while *‘android’* with 35.4% of users, is the second most popular platform. *'Mac'* is the least popular platform, with only 3.9% of users.

A pie chart with text

Description automatically generated with low confidence

Figure : Most used device platform by the users

### 3.3.2 Age Group of the Players

The data pre-processing mentioned in section 3.2 was carried out to determine the age group of the players.

A picture containing text, screenshot, diagram, number

Description automatically generated

Figure : Age group of the players

The number of players in each age group is depicted in Figure 7. As can be seen, most of the players are between the ages of 30-39, followed by those between the ages of 40 and 49. Players between the ages of 20 and 29 are less numerous than those between the ages of 60 and 70.

### 3.3.3 Top Spending Team

The teams with high spends are identified using the "buy-clicks.csv" dataset. The top 15 teams with high spends are displayed in Figure 8. Team ’27’ spends the most, followed by team ‘54’ and ‘35’.

A picture containing text, screenshot, number, font

Description automatically generated

Figure : Top 15 teams with high spends.

### 3.3.4 Comparison between Team Total Spends and Team Strength

The datasets “buy-clicks.csv” and “team.csv” are used to compare the total spend and strength of the teams. Figure 9 shows a scatterplot between teams’ total spend and strength. A p-value of 0.309 for Spearman's rho indicates that there is no strong evidence of a monotonic relationship between the total spend and strength of a team.

A picture containing screenshot, text, colorfulness, circle

Description automatically generated

Figure : Scatterplot to compare team total spend and strength.

### 3.3.5 Countries with Highest Spending Players

The datasets “buy-clicks.csv” and “users.csv” are used to identify the top 10 countries that provide high income to the company. From Figure 10, the country “Grenada (GD)” has the highest spending players with a value of “342”. The countries “Iraq (IQ)” and “Palau (PW)” holds the second and third positions with a spend of “328” and “324”, respectively.

A picture containing text, screenshot, colorfulness, plot

Description automatically generated

Figure : Top 10 countries with highest spend.

### 3.3.6 Highest Revenue Generating Item

The dataset “buy-clicks.csv” is used to find out the product that generates the most income to the company. Treemap in the Figure 11 shows that the item ‘5’ generates highest revenue of ‘12200’, followed by item ‘4’. As the least revenue-generating item, item "1" generates “538” in revenue.

A picture containing screenshot, rectangle, line, diagram

Description automatically generated

Figure : Item that generates highest revenue.

### 3.3.7 Number of times an item is purchased

The dataset “buy-clicks.csv” is used to find out the highly sold product so that the company can focus more on marketing the same. Figure 12 shows that item “2” is the highly sold product, followed by item “5” and item “0”. The least sold item is “1”.

A picture containing text, screenshot, line, number

Description automatically generated

Figure : Highly sold product.

### 3.3.8 Amount spend by top 10 users

The dataset “users.csv” is used to find out the highest spending users. The company can target these players to increase the revenue. From Figure 13, the user ‘YHxb6JJ’ holds the top position with a total spend of “223”, followed by the user “QmA8GoqXpmN” with a spend of 215.

A picture containing text, screenshot, line, font

Description automatically generated

Figure : Users with highest spend.

### 3.3.9 Time series of in-game ad clicks

A picture containing text, screenshot, line, diagram

Description automatically generated

Figure : In game ad clicks per day.

The dataset “ad-clicks.csv” is used to visualise the in-game ad clicks per day in the game. Figure 14 indicates the current advertisement model is effective with an increase in daily ad clicks.

### 3.3.10 Average Team Join

The dataset “team-assignments.csv” is used to find the number of times a player joins a team on average. The boxplot in the Figure 15 indicates that, on average a player joins a team 4 times.

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Figure : Boxplot for statistics of player joining teams.

The violin plot in the Figure 16 indicates that majority of the players join 4 to 7 different teams, as the density of the plot is distributed around 4 and 7.

A diagram of a violin plot

Description automatically generated with low confidence

Figure : Violin plot for statistics of players joining teams.

### 3.3.11 Users who never bought any item

The dataset “users.csv” is used to find the ratio of users who have never bought any item. From Figure 17, it is clear that “42.3%” of users have never bought any product, whereas 57.7% of the players bought at least an item. The organisation can try to attract these users with varieties of products and thus improve the revenue.

A blue circle with white text

Description automatically generated with low confidence

Figure : Users who never bought an item.

# Machine Learning Models

Learning from the data is the goal of machine learning. Based on sample input-output pairs, supervised learning involves learning an algorithm that maps an input to an output. In unsupervised learning, the algorithms are left to find and display the intriguing structure in the data on their own (Mahesh, 2018).

Machine learning algorithms were performed on “combined\_data.csv” file. Two new categorical columns “Spender” or “Non Spender”, and “Hitter” or “Non Hitter” are created. If the average price is higher than 5, the player is classified as a "Spender," otherwise, they are a "Non Spender." The players with average price as “NULL” are not considered for the analysis. The total number of game clicks and the total number of hits are computed, and the percentage of hits is obtained by dividing the total number of game clicks by the total number of hits. The player is labelled as a “Hitter” if the hit percentage is greater than 10%, otherwise, they are a “Non hitter”. The classification and clustering models are built using PySpark.

## 4.1 Classification or Supervised Learning

### 4.1.1 Logistic Regression

Logistic Regression is a classification model that works on an unusual dependent variable: a probability. The model measures how certain that an event of interest happens (Buis, 2017).

Logistic Regression was employed to classify the players as “Spender” or “Non Spender”. The features “team level”, “total hits”, and “Platform type” are used to build the model. The data was split into training and testing set in 70:30 ratio. The model was built using the training set data and employed on test set. The confusion matrix that presents the outcomes of predictions and results of the classifier is given Table 3.

|  |  |  |
| --- | --- | --- |
| **Spender** | **Prediction** | **Count** |
| 0 (True Negative) | 0 | 233 |
| 1 (False Negative) | 0 | 34 |
| 0 (False Positive) | 1 | 25 |
| 1 (True Positive) | 1 | 141 |

Table : Result of Logistic Regression model.

Equation : Equation to calculate Accuracy of a machine learning model.

The accuracy of the model is calculated using the above given formula and the Logistic Regression model obtained an accuracy of 86.37%. From Table 3, the model correctly classified 141 players as “Spenders” and 233 players as “Non spenders”. The model misclassified 34 “Spenders” as “Non spenders” and 25 “Non spenders” as “Spenders”.

Figure 18 is the visual representation of the confusion matrix of the built Logistic Regression model.

A picture containing text, screenshot, diagram, font

Description automatically generated

Figure : Confusion matrix of Logistic Regression model.

### 4.1.2 Decision Tree

Decision tree is a classification model with a tree structure that contains root nodes, branches, and leaf nodes. Every internal node in the model contains an attribute test, which has a branch as its outcome and a leaf node as its class label (Patel et al., 2018).

A decision tree model was built using the features “team level”, “total hits”, and “Platform type”, as same as the ones selected for building Logistic Regression model. The data was split into training and testing set in 70:30 ratio. The model to classify “Spenders” and “Non spenders” was built using the training set data and employed on test set. The confusion matrix that presents the outcomes of predictions and results of the classifier is given Table 4.

|  |  |  |
| --- | --- | --- |
| **Spender** | **Prediction** | **Count** |
| 0 (True Negative) | 0 | 234 |
| 1 (False Negative) | 0 | 36 |
| 0 (False Positive) | 1 | 24 |
| 1 (True Positive) | 1 | 139 |

Table : Result of Decision Tree model.

The model obtained an accuracy of 86.14% using the Equation 1. From Table 4, the model correctly classified 139 players as “Spenders” and 234 players as “Non spenders”. The model misclassified 36 “Spenders” as “Non spenders” and 24 “Non spenders” as “Spenders”. Figure 19 is the visualisation of the confusion matrix of the Decision Tree model.

A picture containing text, screenshot, diagram

Description automatically generated

Figure : Confusion matrix of Decision Tree model.

### 4.1.3 Comparison of Logistic Regression and Decision Tree

As mentioned in section 4.1.1 and 4.1.2, both Logistic Regression and Decision tree models gave almost similar accuracy of 86.37% and 86.14%, respectively. In this study, Logistic model is better at predicting “Spenders”, whereas Decision Tree model is better at predicting “Non spenders”. The organisation can use both the models to improve the classification of players and plan the marketing strategies accordingly.

## 4.2 Clustering or Unsupervised Learning

### 4.2.1 K Means Clustering

K-means algorithm is one of the most popular clustering techniques as it is relatively fast. Even though, there are a lot of other clustering algorithms, K-means remain the preferred tool in many real-world applications. K-means clusters similar group data of data based on Euclidean distance from the centroids, thus increasing the similarity between the samples in a cluster (Raykov et al., 2016).

The features such as “team level”, “average price total”, “total hits”, “platform type”, “hitter” or “Non hitter”, and “Spender” or “Non spender” are used to cluster the data.

Finding the right number of k is a crucial step in K-means algorithm. The k value is calculated using Silhouette score method. Silhouette score for each k is calculated. The highest scoring value of K was 2. As 2 clusters would not tell us much about the underlying structure of the data, the next value of k=4 with highest Silhouette score is considered.

The cluster centres generated by the model are displayed in Table 5.

|  |  |
| --- | --- |
| **Cluster** | **Cluster Centres** |
| Cluster 1 | [2.69221066, 0.52738379, 1.07493231, 1.51191111, 2.66953475, 0.05059769] |
| Cluster 2 | [2.82694309, 0.4789373, 0.82165603, 1.23303438, 0, 0.40313729] |
| Cluster 3 | [2.20342688, 0.98701034, 3.92231241, 1.12566665, 2.55244989, 0.5910989 ] |
| Cluster 4 | [2.75856681, 2.03427329, 1.24957437, 0.10939165, 2.66953475, 1.9192138 ] |

Table :Cluster centres of the clusters generated by K-means.

The clusters can be differentiated as follows.

1. **Cluster 1**:

The cluster includes the highest hitter. The second last average total price and total hits are included in this cluster. The highest hitter and the least spender is included in this cluster.

1. **Cluster 2**:

The cluster includes the highest team level. The least average price total and total hits are also present in this cluster sample. The cluster includes lowest rank hitters and spenders.

1. **Cluster 3**:

This cluster includes the lowest team level and the second highest average total price. The highest total hit is included in this cluster. This cluster also includes second highest hitter and spender.

1. **Cluster 4**:

The second highest team level is included in this cluster. The cluster contains the sample with highest average total price and second highest total hits. The highest spender and hitter is included in this cluster.

A picture containing colorfulness, child art, screenshot, pencil

Description automatically generated

Figure : Scatter plot for the clusters formed.

Figure 20 is the scatterplot for the generated clusters. The clusters with high total hits and average price total are highlighted in “yellow” and “orange” colours, respectively. The clusters with lower total hits and average total price are highlighted with “pink” and “blue” colours.

### 4.2.2 Hierarchical Clustering

In Spark, both K means and Hierarchical clustering are combined to be called as Bisecting K-means. It is a divisive hierarchical clustering algorithm (Abirami and Mayilvahanan, 2016).

In Bisecting K-means, the number of clusters is a hyperparameter to be tuned. The optimal number of clusters is found out using Silhouette method. Figure 21 shows Silhouette score mapped for each number of clusters. Like K-means, the local maximum is at k = 4. Four clusters will yield best results for the model.

A picture containing diagram, plot, line, text

Description automatically generated

Figure : Silhouette score plot for Hierarchical clustering.

A hierarchical clustering model is built and visualised using a dimensionality reduction technique called PCA. PCA is performed and extracted the principal components. Then the cluster assignments are retrieved from the bisecting k-means assignments.

A picture containing screenshot, text, diagram, map

Description automatically generated

Figure : Principal components plot for Hierarchical clustering.

Figure 22 shows the plot for principal components. There is a subtle difference between the clusters formed by the algorithm. The cluster 4 and 3 seems to be different from clusters 1 and 2.

# Graph Analysis

Chattier users, initiators of chats, users who belong to active teams are beneficial due to their ability to reach a bigger audience. As a result, the company can raise its revenue by selecting the proper marketing strategy to target such players by displaying pricier items to such individuals. Even if these users are not going to buy these items, they might persuade someone in their online circle to.

The Flamingo dataset includes 4 files that correlate to game conversation, as mentioned in section 3.1. Table 6 includes information on these files.

|  |  |
| --- | --- |
| **Dataset** | **Description** |
| chat\_join\_team\_chat.csv | This file contains the user ID, team chat session ID, and the timestamp. A new entry is added to the file when a user *‘JOINS’* the chat. |
| chat\_leave\_team\_chat.csv | This file contains the user ID, team chat session ID, and the timestamp. A new entry is added to the file when a user *‘LEAVES’* the chat. |
| chat\_mention\_team\_chat.csv | This file contains the user ID, chat item, and the timestamp. A new entry is added to the file when a user *‘MENTIONED’* in the chat. |
| chat\_respond\_team\_chat.csv | This file contains the chat ID 1, chat ID 2, and the timestamp. A new entry is added to the file when a user *‘RESPONDS’* to their mentioned chat.  The file will be updated with both the 'MENTIONED' and 'RESPONDS' entries. |

Table : Datasets used for graph analysis.

## 5.1. Players JOIN the Chat

‘*chat\_join\_team\_chat.csv’* file is used to find out the users who joins the team chats. The dataset is loaded into Neo4j. The nodes *'user\_id'* and *'teamchat\_session\_id'* were created with an edge *'Joins'* denoting the timestamp. Figure 23 and Figure 24 shows the graphs generated in Neo4j.

A picture containing fireworks, darkness, night, light

Description automatically generated

Figure : Graph of JOINS relation between players and chat sessions.

A picture containing drawing, flower, pattern, embroidery

Description automatically generated

Figure : Snapshot of the JOINS relation between Players and Chat sessions.

## 5.2. Players LEAVE the Chat

‘*chat\_leave\_team\_chat.csv’* file is used to find out the users who leaves team chats. The dataset is loaded into Neo4j. The nodes *'user\_id'* and *'teamchat\_session\_id'* were created with an edge *'Leaves'* denoting the timestamp. Figure 25 and Figure 26 shows the graphs for “LEAVES” relationship between user and team chat session.

A picture containing art, fractal art, circle, light

Description automatically generated

Figure : Graph for "LEAVES" relationship

A picture containing circle, design

Description automatically generated

Figure : Snapshot of the "LEAVES" relationship.

## 5.3. Players MENTIONED in the Chat

‘*chat\_mention\_team\_chat.csv’* file is used to find out the users who are mentioned in the team chats. The nodes *'user\_id'* and *'chat\_item'* were created with an edge *'Mentions'* denoting the timestamp. Figure 27 and Figure 28 shows the graphs for the “Mentions” relationship between user and chat item.

A picture containing fireworks, darkness, light

Description automatically generated

Figure : Graph for "MENTIONS" relationship.

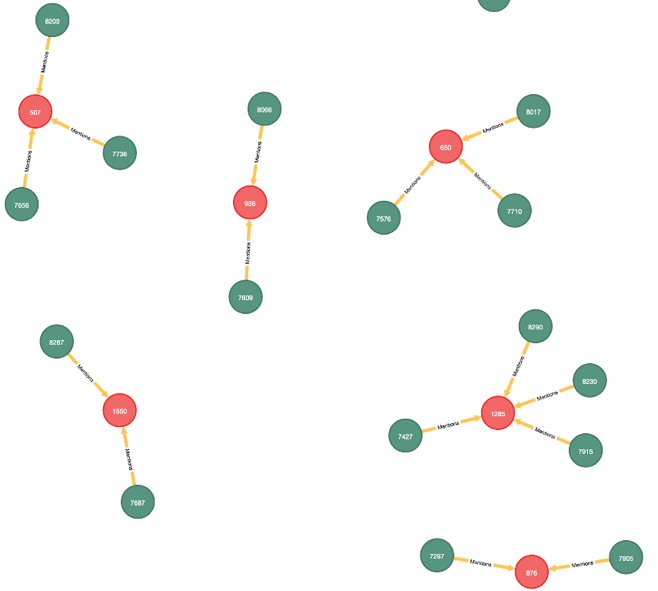


Figure : Snapshot of the "Mentions" relationship.

## 5.4. Players RESPONDS to the Chat

‘*chat\_respond\_team\_chat.csv’* file is used to find out the users who responds to the mentions in the team chats. The nodes *'chatid1'* and *'chatid2'* were created with an edge *'RespondsTo'* denoting the timestamp. Figure 29 and Figure 30 are the graphs generated in Neo4j for ‘RepsondsTo’ relationship.

A picture containing darkness, light, night

Description automatically generated

Figure : Graph for the "RespondsTo" relationship.

A picture containing circle

Description automatically generated

Figure : Snapshot of the "RespondsTo" relationship.

## 5.5. Identifying Long Conversations in the Chat

To increase the revenue of the organisation, it is crucial to determine when players are most active and to display advertisements during that moment. The 'RespondsTo' relationship indicated in section 5.4 is used to determine the length of the longest conversation.

The longest conversation between the players contained only 10 messages, as seen in Figure 31. This indicates that during the game, the players do not have further in-depth chats. Therefore, it will be preferable to show the advertisements during game rather than in chat.

A picture containing colorfulness

Description automatically generated

Figure : Longest conversation.

## 5.6. Identifying Active Teams

Planning a stronger marketing strategy necessitates identifying the teams who remain the most active. The team with the most 'Join' will be the most active team. The 'Join' relationship specified in section 5.1 is used to identify the active team.

Figure 32 and Table 7 show the most active teams. Figure 32 makes it evident that team ‘6792’ has the most participants joined, with a total of 100. Team ‘6783’ and ‘6925’, who had 91 and 87 players, respectively, are in second and third place.

|  |  |  |
| --- | --- | --- |
| **Rank** | **Team ID** | **Joined Player Count** |
| 1 | 6792 | 100 |
| 2 | 6783 | 91 |
| 3 | 6925 | 87 |
| 4 | 6850 | 86 |
| 5 | 6791 | 81 |
| 6 | 6780 | 76 |
| 7 | 6809 | 72 |
| 8 | 6819 | 70 |
| 9 | 6795 | 67 |
| 10 | 6889 | 67 |

Table : Most active teams.

A picture containing light

Description automatically generated

Figure : Graphical representation of most active teams.

## 5.7. Identifying Active Users

As previously mentioned, users who are more chatty or active are advantageous to the business because they greater sway on their peers. Advertising to these gamers would spark a lot of interest among other players as well.

The ‘Mentioned’ relationship given in section 5.3 was used to identify the active players. Table 8 lists the top 10 active users. Player ‘131’ is mentioned 53 times, followed by players ‘1204’ and ‘621’ with 47 mentions each. Based on these players' ranking and the fact that they have more influence over other players, it will be beneficial to promote more products to them.

|  |  |  |
| --- | --- | --- |
| **Rank** | **Player ID** | **Mention Count** |
| 1 | 131 | 53 |
| 2 | 1204 | 47 |
| 3 | 621 | 47 |
| 4 | 1428 | 46 |
| 5 | 1506 | 46 |
| 6 | 1482 | 42 |
| 7 | 1450 | 42 |
| 8 | 283 | 42 |
| 9 | 674 | 42 |
| 10 | 1035 | 41 |

Table : Active users.

# Big Data Ethics

Big data ethics is the study of conceptions of good and wrong behaviour when it comes to the use of data, with a focus on personal data. Big data ethics seeks to establish a moral and ethical standard for data use. Bigger does not necessarily equate to better, approachable does not equate to moral, and convenience does not equate to effectiveness (Chen et al., n.d.).

Due to the indiscriminate nature of big data, companies may have access to data that they never planned to gather. This might violate peoples' privacy. It might not be possible to eliminate the capacity to identify the individuals with such extensive data. Big Data analysis could quickly identify the real people whose data has been masked if data masking is not done properly (Herschel and Miori, 2017a).

European Union General Data Protection Regulation, which came into force in May 2018, regulates the collection, storage, and processing of personal data (Gruschka et al., 2018). California legislature enacted its own data protection law, the California Consumer Privacy Act in 2018 (Palmieri, 2020).

Big data ethics in the data collection, storage, and processing are discussed in the following sections.

## 6.1 Ethics in Big Data Collection

To access digital media, users must consent to the collection of the data they create while on a website. The amount of data gathered has increased dramatically, yet privacy awareness has not kept up. Access requires giving up privacy, which raises ethical concerns (Jurkiewicz, 2018). Big Data is the effect of individual actions, sensory data, and other applications. Ethical disadvantage arises from ignorance of the data obtained from online shoppers or mobile owners. Raytheon's Rapid Information Overlay Technology may track one person and make their daily activities transparent (Zwitter, 2014). Chessell (2014) mentions that the advancements in analytics and big data technology have widened the gap between what is possible and what is legally allowed, causing ethical issues in data collection. Big Data has been criticized as a breach of privacy and discriminatory in generating large and complex datasets (Martin, 2015).

Big data systems' challenges with data generation are frequently discussed in the media. An illustration of how linking datasets is unethically problematic is provided by the Beacon software from Facebook and the iBeacon software from Apple (Mittelstadt and Floridi, 2016). Big Data originates from people using Facebook, Google or Twitter, as well as from digital devices such as location sensors in phones and cars. These data are sold repeatedly until there is nothing left to be used (Someh et al., 2016).

Ethical awareness framework developed by the UK and Ireland Technical Consultancy Group helps to develop ethical policies on big data (Chessell, 2014).

## 6.2 Ethics in Big Data Storage

To handle the enormous volume of data, big data requires distributed systems with great processing and storage capacity. Due of the number of parties involved, there is a significant danger of privacy breach (MEHMOOD et al., 2016). To manage unstructured data, NoSQL data models were developed. It's debatable if modern databases can uphold integrity and confidentiality. The danger of security and privacy breaches has increased as database systems have transitioned from on-premises to distributed cloud-based systems (Samaraweera and Chang, 2021). Cloud storage can lower the cost of maintaining data and ease data administration. However, due to security and privacy concerns about independent cloud service providers, businesses are hesitant to shift their data to cloud storage (Li et al., 2016).

The difficulties of storing enormous amounts of data can be addressed through cloud computing (Liu, 2013). Xiao et al. (2013) identified confidentiality, integrity, availability, accountability, and privacy-preservability as the five most representative security and privacy attributes. Protecting a person's privacy is a fundamental necessity for big data storage systems. ‘Attribute based encryption’, ‘Homomorphic encryption’, ‘Storage path encryption’, and ‘Usage of Hybrid Clouds’ can be considered to safeguard the privacy of the data stored on the cloud systems (Jain et al., 2016).

Chen et al. (2021) proposed Holistic Big Data Integrated Artificial Intelligent Modelling to improve the privacy of data stored in cloud. Li et al. (2016) discussed about Oblivious RAM to preserve privacy for cloud stored data.

It is crucial to enable public auditability for cloud storage so that customers can use a third-party auditor to verify the integrity of data that has been outsourced. To support privacy-preserving public auditing, a secure cloud storage design was suggested (Wang et al., 2013).

## 6.3 Ethics in Big Data Processing

The processing of Big Data is fraught with ethical issues (Herschel and Miori, 2017). Results from big data are debatable in terms of ethics. The integrity of the information flows can be preserved, which will help with this problem (Johnson, 2014).

As mentioned in section 2, Big Data processing paradigms categorises the systems as batch, real-time, and hybrid. Ensuring the ethicality of these systems is crucial in Big Data. The first phase of privacy protection in data processing is to safeguard information from unsolicited disclosure, as it may contain sensitive information. The second phase is the extraction of meaningful information without violating the privacy (Jain et al., 2016).

Given how challenging it is to anticipate or even identify the location of big data processing, it is essential to implement some control measures, either to prevent the node from obtaining the processing results or to verify the node's legitimacy. Blind processing techniques based on homomorphic encryption can be considered as a solution for this (Gahi et al., 2016).

Lin et al. (2007) discussed the concepts of k-anonymity, l-diversity, and t-closeness for privacy protection, and HybrEx, a hybrid execution model, to ensure the confidentiality and privacy in cloud computing. The model uses organisation’s private cloud for processing the sensitive data. Individual privacy in Big Data can be ensured by running operations on encrypted data (Dwork, 2006).

# Conclusion: Finding and Recommendation

The study discussed about the Big Data processing paradigms such as batch, real time, and hybrid processing. The report also discusses about ethical issues in Big Data in data collection, storage, and processing, and the solutions.

EDA, machine learning, and graph analytics were performed on the dataset for the game “Catch the Pink Flamingo” and the results were recorded. The recommendations to improve the revenue of the organisation based on the analysis done are given below.

The data exploration showed that the top users and purchasers of most expensive items generate a large revenue for the company. Therefore, identifying and targeting such users will be beneficial for the company. It was also seen that the most used device platform is “iPhone”. Advertising the game to iPhone users will be more valuable than advertising to users in other platforms. Improving the user experience in iPhone and Android, the predominantly used platform, can attract more users, and thus improve the revenue. Targeting the players between the age 30 to 39, the highest spending teams such as “27”, “54”, and the players in the high spending countries like Grenada, Iraq, etc can improve the revenue. Steps can be taken to promote the highest revenue generating item “5” and the highly sold item “2”. The major portion of the users have never bought any item. The interests of these users can be identified, and more products can be introduced accordingly.

As these people frequently have greater influence among their peers, focusing on the chattier users identified by graph analytics may increase the company's value. It is clear from cluster analysis that different user groups have varied interests. The organisation needs to come up with various targeting techniques for various groups. Due to their tendency to make more in-game purchases, the users classified as "Spenders" can be identified and targeted to increase revenue.

# Reference

Abirami, K., Mayilvahanan, P., 2016. Performance Analysis of K-Means and Bisecting K-Means Algorithms in Weblog Data. International Journal of Emerging Technologies in Engineering Research (IJETER) 4.

Adaptive and Efficient Streaming Time Series Forecasting with Lambda Architecture and Spark [WWW Document], n.d. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9377947 (accessed 5.18.23).

Big Data Framework, 2023. A Short History of Big Data | Enterprise Big Data Framework [WWW Document]. URL https://www.bigdataframework.org/knowledge/a-short-history-of-big-data/ (accessed 5.15.23).

Buis, M.L., 2017. Logistic regression: When can we do what we think we can do? 1.

Buyya, R., Calheiros Rodrigo N., Dastjerdi Amir Vahid, 2016. Big Data: Principles and Paradigms - Google Books [WWW Document]. URL https://books.google.co.uk/books?hl=en&lr=&id=MfOeCwAAQBAJ&oi=fnd&pg=PP1&dq=real+time+big+data+processing+paradigm&ots=itT8\_dtDnL&sig=heysGQK\_aYrUjw8hJ3DHrpsaQvo&redir\_esc=y#v=onepage&q&f=false (accessed 5.18.23).

Casado, R., Younas, M., 2015. Emerging trends and technologies in big data processing. Concurr Comput 27, 2078–2091. https://doi.org/10.1002/CPE.3398

Chen, J., Ramanathan, L., Alazab, M., 2021. Holistic big data integrated artificial intelligent modeling to improve privacy and security in data management of smart cities. Microprocess Microsyst 81. https://doi.org/10.1016/J.MICPRO.2020.103722

Chen, W., Quan-Haase, A., Popham, J., Lavoie, J., Coomber, N., n.d. Big Data Ethics and Politics: Toward New Understandings. Soc Sci Comput Rev 2020, 3–9. https://doi.org/10.1177/0894439318810734

Chessell, M., 2014. Ethics for big data and analytics.

databricks, n.d. What Is Lambda Architecture? [WWW Document]. URL https://www.databricks.com/glossary/lambda-architecture (accessed 5.18.23).

Dwork, C., 2006. Differential privacy. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 4052 LNCS, 1–12. https://doi.org/10.1007/11787006\_1/COVER

Gahi, Y., Guennoun, M., Mouftah, H.T., 2016. Big Data Analytics: Security and Privacy Challenge [WWW Document]. IEEE Symposium on Computers and Communication (ISCC). URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7543859 (accessed 5.19.23).

Gruschka, N., Mavroeidis, V., Vishi, K., Jensen, M., 2018. Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR [WWW Document]. IEEE International Conference on Big Data (Big Data). URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8622621 (accessed 5.18.23).

Gurcan, F., Berigel, M., 2018. Real-Time Processing of Big Data Streams: Lifecycle, Tools, Tasks, and Challenges. ISMSIT 2018 - 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings. https://doi.org/10.1109/ISMSIT.2018.8567061

Herschel, R., Miori, V.M., 2017a. Ethics & Big Data. Technol Soc 49, 31–36. https://doi.org/10.1016/J.TECHSOC.2017.03.003

Herschel, R., Miori, V.M., 2017b. Ethics & Big Data. Technol Soc 49, 31–36. https://doi.org/10.1016/J.TECHSOC.2017.03.003

Jain, P., Gyanchandani, M., Khare, N., 2016. Big data privacy: a technological perspective and review. J Big Data 3, 1–25. https://doi.org/10.1186/S40537-016-0059-Y/TABLES/5

Johnson, J.A., 2014. The Ethics of Big Data in Higher Education. The International Review of Information Ethics 21, 3–10. https://doi.org/10.29173/IRIE365

Jurkiewicz, C.L., 2018. Big Data, Big Concerns: Ethics in the Digital Age. Public Integrity 20.

Li, N., Li, T., Venkatasubramanian, S., 2007. t-Closeness: Privacy Beyond k-Anonymity and -Diversity [WWW Document]. IEEE. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4221659 (accessed 5.19.23).

Li, P., Guo, S., Miyazaki, T., Xie, M., Hu, J., Zhuang, W., 2016. Privacy-Preserving Access to Big Data in the Cloud. IEEE Cloud Computing 3, 34–42. https://doi.org/10.1109/MCC.2016.107

Liu, S., 2013. Exploring the Future of Computing [WWW Document]. IEEE Computer Society. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6449386 (accessed 5.19.23).

Mahesh, B., 2018. Machine Learning Algorithms-A Review Machine Learning Algorithms-A Review View project Self Flowing Generator View project Batta Mahesh Independent Researcher Machine Learning Algorithms-A Review. International Journal of Science and Research. https://doi.org/10.21275/ART20203995

Martin, K.E., 2015. The Big Data Industry. MIS Quarterly Executive 67.

MEHMOOD, A., NATGUNANATHAN, I., XIANG, Y., HUA, G., GUO, S., 2016. Protection of Big Data Privacy [WWW Document]. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7460114 (accessed 5.18.23).

Mittelstadt, B.D., Floridi, L., 2016. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. Sci Eng Ethics 22, 303–341. https://doi.org/10.1007/S11948-015-9652-2/TABLES/3

Niculescu, V., 2020. On the Impact of High Performance Computing in Big Data Analytics for Medicine [WWW Document]. Appl Med Inform. URL https://www.researchgate.net/publication/341622174\_On\_the\_Impact\_of\_High\_Performance\_Computing\_in\_Big\_Data\_Analytics\_for\_Medicine (accessed 5.19.23).

Palmieri, N.F.I., 2020. Who Should Regulate Data: An Analysis of the California Consumer Privacy Act and Its Effects on Nationwide Data Protection Laws. Hastings Science and Technology Law Journal 11.

Patel, H., Prajapati, P., Patel, H.H., 2018. Study and Analysis of Decision Tree Based Classification Algorithms Extreme Multi-label Classification Problem View project Significance of the Transition to Outcome Based Education: Explore the Future View project Study and Analysis of Decision Tree Based Classification Algorithms. International Journal of Computer Sciences and Engineering Open Access Research Paper. https://doi.org/10.26438/ijcse/v6i10.7478

Phillips, A., 2021. A history and timeline of big data [WWW Document]. URL https://www.techtarget.com/whatis/feature/A-history-and-timeline-of-big-data (accessed 5.15.23).

Purohit, J., n.d. HYBRID DATA PROCESSING MODEL FOR BIG DATA – A REVIEW [WWW Document]. International Journal of Latest Trends in Engineering and Technology . URL https://www.ijltet.org/journal/149656699651.1777.pdf (accessed 5.18.23).

Raykov, Y.P., Boukouvalas, A., Baig, F., Little, M.A., 2016. What to Do When K-Means Clustering Fails: A Simple yet Principled Alternative Algorithm. PLoS One 11, e0162259. https://doi.org/10.1371/JOURNAL.PONE.0162259

Sagiroglu, S., Sinanc, D., 2013. Big data: A review. Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013 42–47. https://doi.org/10.1109/CTS.2013.6567202

Samaraweera, G.D., Chang, J.M., 2021. Security and Privacy Implications on Database Systems in Big Data Era: A Survey [WWW Document]. IEEE Trans Knowl Data Eng. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8765784 (accessed 5.18.23).

Shahrivari, S., 2014. Beyond Batch Processing: Towards Real-Time and Streaming Big Data. Computers 3, 117–129. https://doi.org/10.3390/computers3040117

Someh, I.A., Breidbach, C., Davern, M., 2016. ETHICAL IMPLICATIONS OF BIG DATA ANALYTICS Service Innovation and Technology View project Fintech Services View project.

Wang, C., Chow, S.S.M., Wang, Q., Ren, K., Lou, W., 2013. Privacy-preserving public auditing for secure cloud storage. IEEE Transactions on Computers 62, 362–375. https://doi.org/10.1109/TC.2011.245

Warren, J., Marz, N., 2015. Big Data: Principles and best practices of scalable realtime data systems [WWW Document]. URL https://books.google.co.uk/books?hl=en&lr=&id=XjszEAAAQBAJ&oi=fnd&pg=PT17&dq=big+data+principles+and+best+practices+marz+and+warren&ots=HWjsu-BsU8&sig=iVITmQURPVuyjr\_312m3K2DTS9Y&redir\_esc=y#v=onepage&q=big%20data%20principles%20and%20best%20practices%20marz%20and%20warren&f=false (accessed 5.19.23).

Xhafa, F., Naranjo, V., Caballé, S., 2015. Processing and analytics of big data streams with Yahoo!S4. Proceedings - International Conference on Advanced Information Networking and Applications, AINA 2015-April, 263–270. https://doi.org/10.1109/AINA.2015.194

Xiao, Z., Xiao, Y., 2013. Security and Privacy in Cloud Computing [WWW Document]. IEEE COMMUNICATIONS SURVEYS & TUTORIALS. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6238281 (accessed 5.19.23).

Zwitter, A., 2014. Big Data ethics. https://doi.org/10.1177/2053951714559253

# Appendix

The datasets and source code files are uploaded in the path <https://github.com/meeramullamkuzhy/Catch-the-Pink_flamingo>