CMP 7203 BIG DATA MANAGEMENT

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

BY

Joshua Fernandes

STUDENT NO**: 22169738**

MSc BIG DATA ANALYTICS

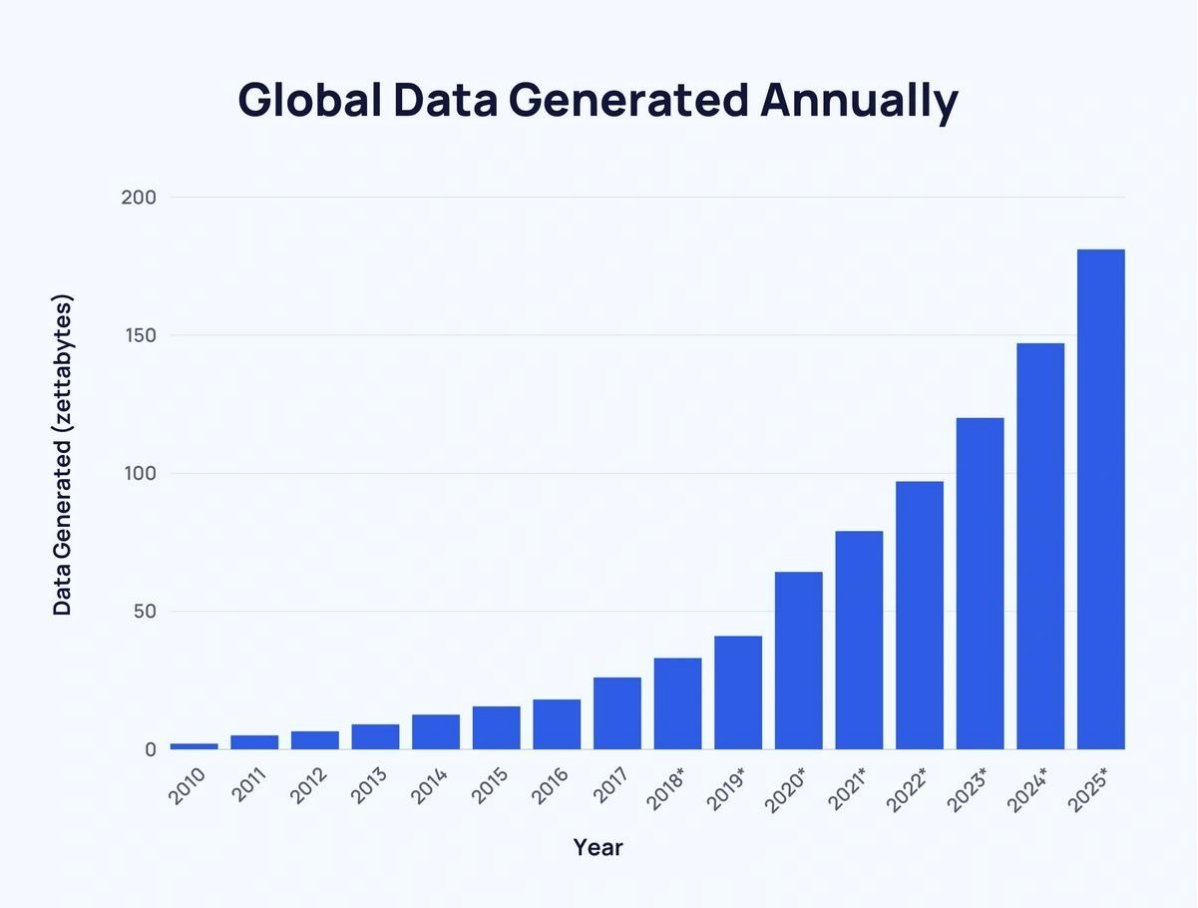
SUBMITTED MAY 19th 2023

Table of Contents

1. Introduction

Data is a very important topic in the world and almost every company uses data in one way or another.Data allows companies to determine what is causing problems in the company.It can be used as a tool for problem solving as well as streamlining processes and as a result help organisations make money. It also helps companies make money as it can help them identify where most of the expenses are coming from.

According to estimates 328.77 million terabytes of data are created every day which is approximately 120 Zettabytes of data per year, 10 zettabytes per month, 2.31 zettabytes per week, or 0.33 zettabytes every day.With this much data being created big data is important as we need to find optimal solutions to deal with all this data coming in and this report will show ways in which this can be done.



2. Big Data Processing Paradigms

2.1 Big Data Background

The definition of big data is data that large and complex sets of data that cannot be effectively processed or analysed using traditional data processing applications. Big data encompasses large volumes of complex structured, unstructured, and semi structured data which is beyond the processing of traditional databases. Big data possess key traits such as

volume, velocity, and variety of information. This refers to the 3Vs of big data. With big data the volume of data is very large. Most of this data is highly unstructured. A few examples of this data include things like Twitter data feeds, Sensor data, ticker data and surveillance. Velocity refers to the rate at which data is being received and acted on. Real time data streams directly from memory onto disk. Variety refers to the type of data. Traditional data types refer to structured data that fit neatly in a relational database. With big data emerging most data that is now being received is unstructured. Unstructured data is information that is not arranged according to a pre-set data model or schema, and therefore cannot be stored in a traditional relational database.

Big data paradigms refer to the fundamental models that guide the processing and analysis of complex unstructured data. These paradigms provide the framework which is used to address the problems caused by this data. The 3 main paradigms are Batch processing, Real Time processing and Hybrid processing. Batch processing refers to data being grouped together rather than being used individually or in real time. This collection of data is called a batch and is then processed together usually done by running a program or script on it. Real Time processing refers to the processing of data as it happens. Although batch processing works for some applications, others need a quicker response time. When certain aspects of batch processing and real time processing are used, it is referred to as hybrid processing.

2.2 Batch Processing Paradigm

The term batch processing dates back to the 1890s where an electronic tabulator was used to record information for the united states census Bureau. Census workers marked data cards now referred to as punch cards which were then put through an electromechanical device. From the 1960s developers began scheduling batch programs on magnetic tape to run sequentially throughout the day(AWS,2023).
Batch Processing is the method used by computers to complete high volume repetitive data jobs.

The main advantages of Batch Processing are efficiency, automation and most importantly scalability(Indeed,2022). Batch Processing has a greater efficiency as it only completes the task when it has the computational resources to do so. Automation is another big advantage of batch processing as by definition large parts of it don't require any human element to it. The biggest advantage of batch processing is scalability. One of the reasons batch processing is used for big data is because of its scalability.
This section will feature how scalability is achieved in batch processing using the MapReduce framework and it's three distributed data processing engines, Apache Hadoop, Apache Spark, Apache Flink. MapReduce is a framework introduced by Google in 2003 used to process and manage large datasets in a parallel and distributed cluster. This is done by splitting the data and distributing it among the cluster. The same operation is then done on each split of data simultaneously. The results are then aggregated and returned to the master node. If the task fails the framework manages all the task scheduling, monitoring and re-execution(Garcia-Gil et al., 2017).

There are many tools used to help with batch processing. The first one being Apache Hadoop. Hadoop is an open-source framework used to efficiently store and process large datasets. It implements the MapReduce algorithm (Lam, 2010) .A main module of Apache Hadoop is the Hadoop Distributed File System also known as HDFS.HDFS is a large distributed file system that uses commodity hardware and provides high throughput as well as fault tolerance to store data. HDFS stores the files as blocks and replicates them for fault tolerance. In addition to fault tolerance other advantages of HDFS include the ability to store large amounts of data and cost effectiveness as the data nodes that store the data rely on inexpensive hardware. Hadoop is also projected to store half the world's data. The limitations of Hadoop are the intensive disk usage, the inadequate in memory computation and the poor performance for online computing(Garcia-Gil et al., 2017).

Apache Spark can be used to improve upon the poor online and iterative computing. Apache spark is an open source cluster computing framework for large datasets(Shanahan & Dai, 2015).Spark maintains MapReduce's scalability and fault tolerance but improves it in more than a few ways. The first way is it is much faster as well as being much easier to program due to its APIs with Python, Java and Scala. It not only works with batch processing but with real time processing which will be looked at in the following sections. Spark is based on distributed data structures called Resilient Distributed Datasets which are both immutable and versatile meaning they can be used for both batch processing and real time processing(Garcia-Gil et al., 2017).

Another tool that can be used for batch processing is Apache Flink which similarly to Apache Spark can be used for bath processing and real time processing. The main library used for batch processing is Dataset and Data frame API.

Real world examples of batch processing include a bank that processes transaction using international money after hours. Another example would be a manufacturer producing a daily operational report for a production line that is run in a batch window

2.3 Real-time Processing Paradigm

Currently many applications require real time processing. It is a fast and prompt data processing technology that combines data capturing, data processing and data exportation all at once. Real time processing and stream processing are quite similar as they both process data with time constraints however they do have small differences. The difference between the two is that real time processing process data within a given space of time whereas stream processing must be immediate.

Real time processing deals with Velocity in a way that batch processing does using low latency processing and event driven architecture. Low latency processing can be defined as processing with minimal delay (Informatica,2023).This is achieved by using small amounts of

data. Event driven architecture can be defined as a computer program written to respond to actions generated by a user or system. The other features of real time and stream processing are scalability as well as parallel and distributed computing. Real time processing can scale vertically and horizontally however there are some limitations to scaling vertically as there are practical limitations to upgrading hardware.

Figure shows the lifecycle of big data processing. Data ingestion refers to the process in which data is imported from various sources into one database.In real time processing some of the tools used in data ingestion are Flume,Kafka and Nifi.Apache flume is an open source distributed data collection framework part of the Apache Hadoop ecosystem.Apache Kafka is an event streaming platform used to collect, process, store, and integrate data.Apache Nifi is a real time open source data ingestion framework made to manage data transfer between different sources and destinations.

The second area of the lifecycle is data processing also known as stream processing. The main tools used for stream processing are Storm,Spark streaming,S4,Flink and Samza. Apache Storm is a distributed real time framework for processing large volumes of high velocity data(Yang et al., 2013). Storm processes data extremely fast with a speed of a million records per second per node on a cluster dependent on it's size. Spark Streaming is an extension of the Spark API that enables scalable, fault-tolerant stream processing of live data streams. Apache S4 is a framework that analyzes streaming data. It is built in Java.Apache Flink as mentioned earlier can be used for both batch processing and real time processing.

The last area of the real time processing life cycle is Analytical Data storage and the tools it use are Hbase,Hive and Cassandra. HBase is a column-oriented non-relational database management system built on HDFS built by Apache Hadoop. Hive allows users to read, write, and manage data using standard query language. Cassandra is an open-source NoSQL distributed database that manages large amounts of data like Mongodb.

Real world examples of real time processing include near real time data and real time data. Real time data would be things like stats in online gaming where you get the data instantly as well as traffic monitoring where you would get the busiest areas with the most traffic instantly. Near real time data would be things like the weather as it would need to be collated from various weather stations and then using that it can be processed to forecast the weather. Another example of this would be things like live updates from news outlets and sports score as they would be processed quickly but not immediately

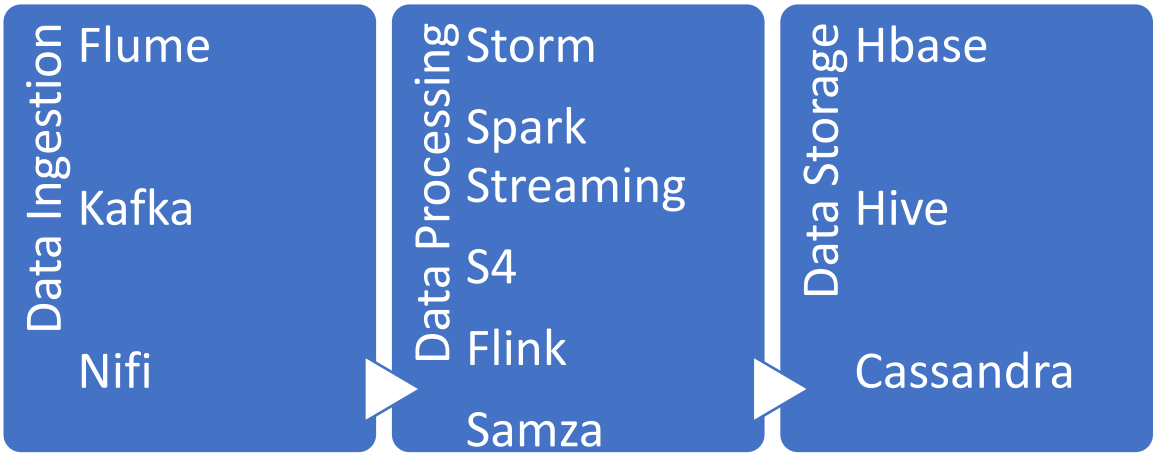


Figure 1Real Time Processing Lifecycle and frameworks

2.4 Hybrid Processing Paradigm

Hybrid processing helps solves the volume and velocity areas of big data. Hybrid processing leverages the use of batch processing and real time processing to fulfil certain goals. Hybrid processing deals with the limitations of batch processing and real time processing creating a unified platform to deal with both. The advantage of hybrid processing is flexibility as it can leverage the strength of both batch and real time processing.

An important framework for hybrid processing is lambda architecture. Lambda architecture is a data deployment model used for processing that is made up of a batch data pipeline and coupled with a fast-streaming data pipeline so that it can handle real time data all while using batch processing. The limitations of such a model is that it can be rather expensive as the architecture has many layers thus making it computationally expensive.

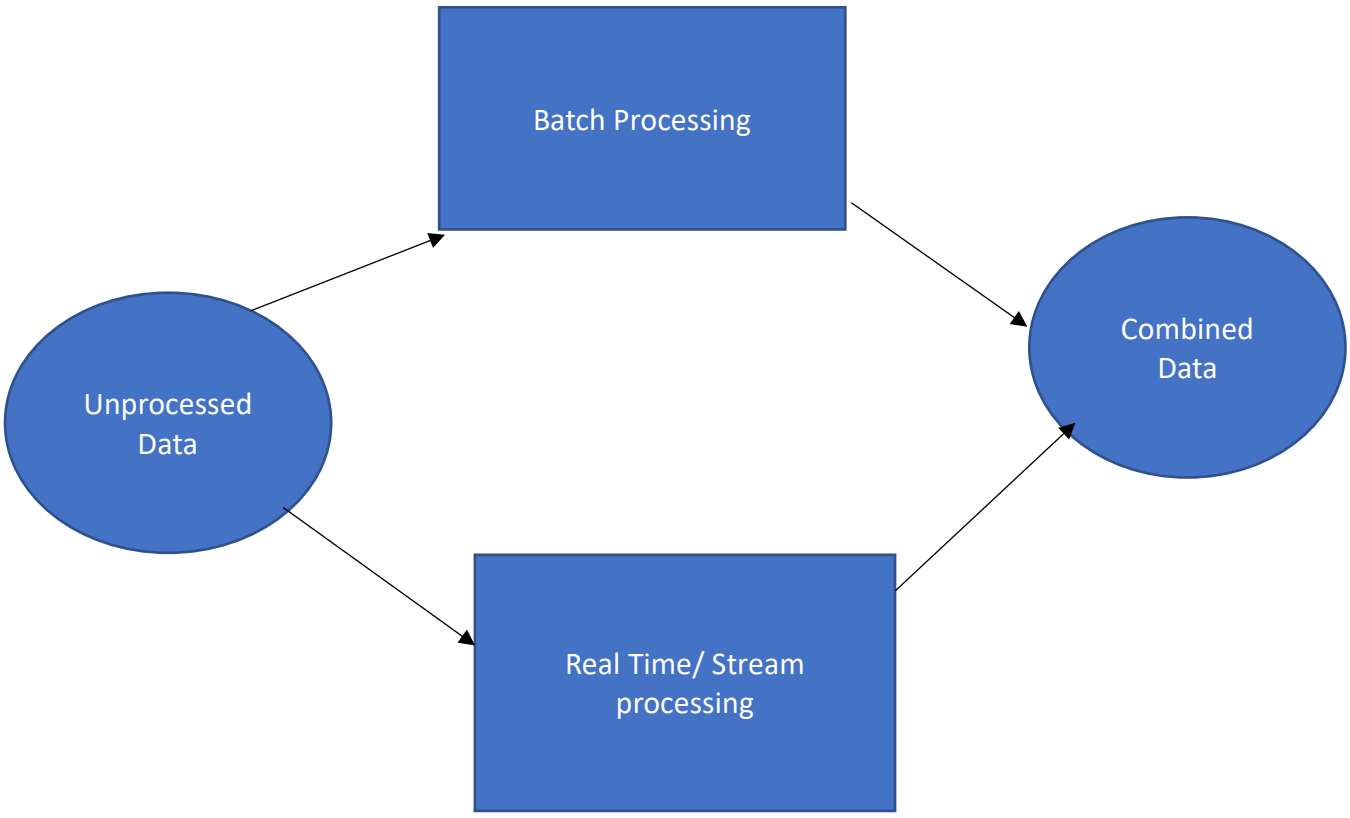


Figure 2 Hybrid Processing Model

2.5 Comparison of the three

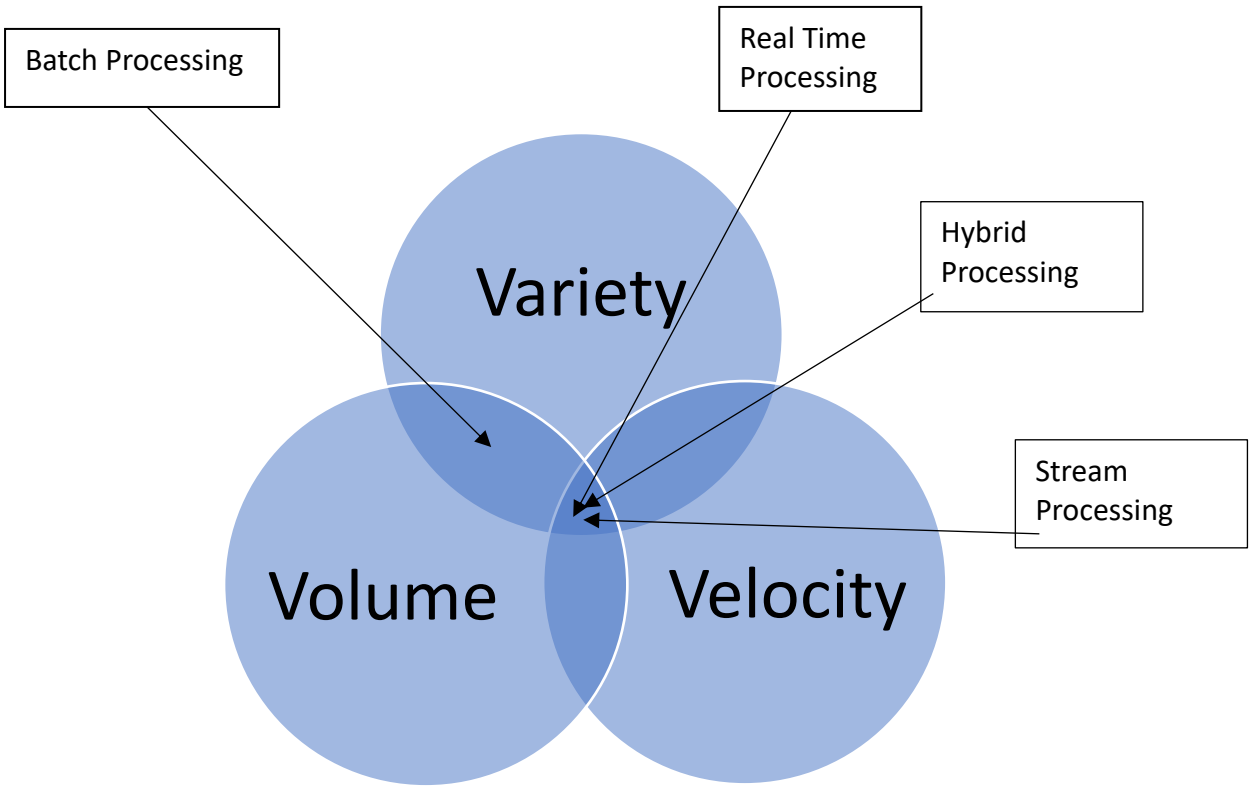


Figure 3 Problems with Big data and their solutions

To summarise the 3 processing paradigm all have its own strengths and limitations as well as dealing with the 3Vs of big data. Batch processing deals with Variety and Volume as during batch processing a lot of data comes in at once and in many different formats however it does not do well in dealing with fast. Real Time processing deals with data quickly as well as dealing with a lot of different data and in large amounts thus proving it can deal with all 3Vs.Similarly to real time processing hybrid processing deals with the 3Vs of big data however it can be more computationally expensive than the other 2.The one thing batch processing has over the other 2 is that it is more efficient as it completes multiple operations at once for large amounts of data. However the other 2 as mentioned earlier deal with data much faster and work with just as much data. This shows that when dealing with time sensitive data it is always better to use real time processing as it is immediate or near immediate depending on the situation. Hybrid processing is a way to leverage the 2 to be more flexible with data.

3. Exploratory Data Analysis

3.1 Flamingo Data Overview

Eglence Inc is an imaginary company known for making a product known as the mobile game called 'Catch the Pink Flamingo'.The objective of the game is to catch as many pink flamingos as they can by following the missions. As you catch more flamingos you level up and the complexity of the game increases.The game is a multiplayer game and you must collect at least one pink flamingo per player or per team in order to progress to the next level. The game gets more and more complex as it goes on.Users can also make purchases and communicate with their team.

The flamingo data set is made up of 3 areas. These sections are chat data,flamingo data and combined data.The flamingo data includes 8 datasets of which 5 were used.The chat data includes the 4 datasets, one for each action. The 4 actions are joining,leaving,mentioning and responding.This data will be analysed in the graph analysis section.

3.2 Data Pre-Processing

Data Acquisition

5 CSVs from the dataset were pre-processed in order to be visualised and to run machine learning .The first CSV was the combined-data csv which is data that contains certain columns from 3 the user-session,buy-clicks and game-clicks CSVs. The second CSV was the ad-clicks CSV which is the database of ads. The 3rd CSV was the buy-clicks CSV which was the database of purchases. The 4th CSV was the users CSV which was the database of the users. The last CSV was the team-assignments which was a record of each time a player joins a team.

Data Cleaning

To clean the data each dataset had the missing values removed. Using the schema the data type was checked and if it wasn't it was changed to the correct data type.The columns that weren't being used were dropped.

3.3 EDA Visualizations

Missing Values for combined data dataset

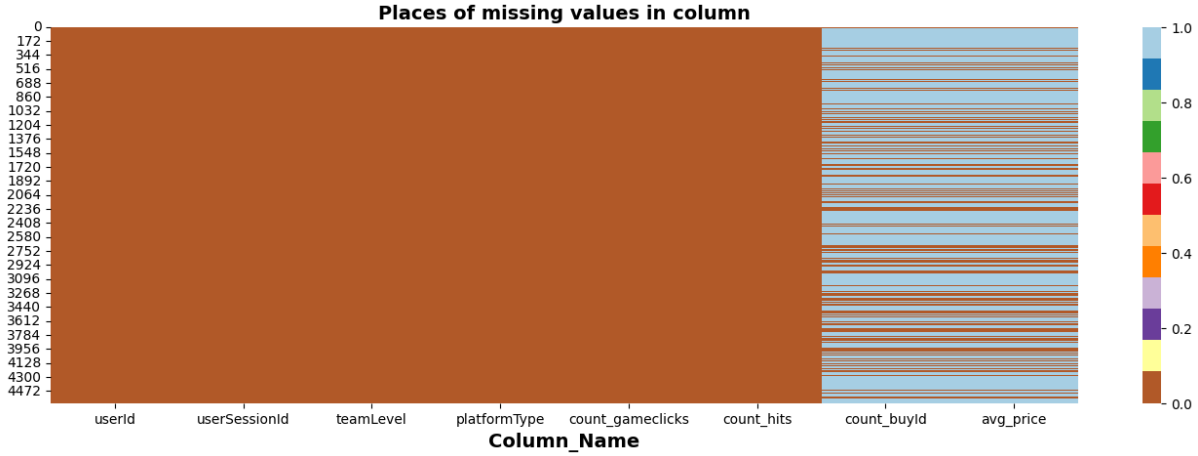


Figure 4Missing Values in combined data dataset

Figure 4 shows the missing data in the combined data dataset. As shown by the figure the only fields with missing data were count_buyId and avg_price. This shows us that the missing values are the fields in which no purchases were made. The EDA for this data was done before cleaning.

Adverts Insights for data

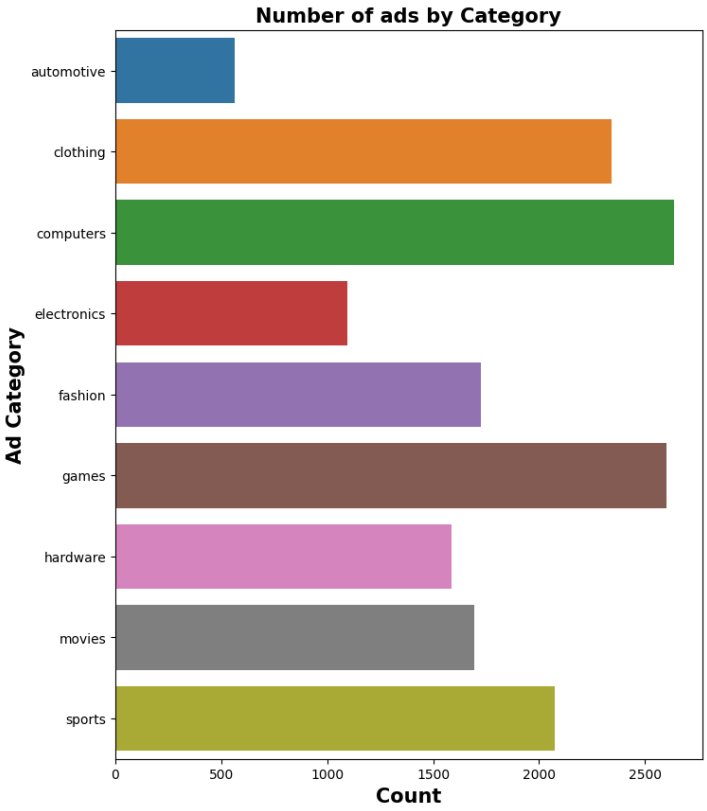


Figure 5Count Plot of Adverts by category

Figure 5 shows what category of adverts are most popular on the game. Data was taken from the database of adverts and cleaned leaving only the category column which was then counted to see how many each category was purchased. As is visible by figure 5 computers and games were very popular which makes sense as it is on a computer game.

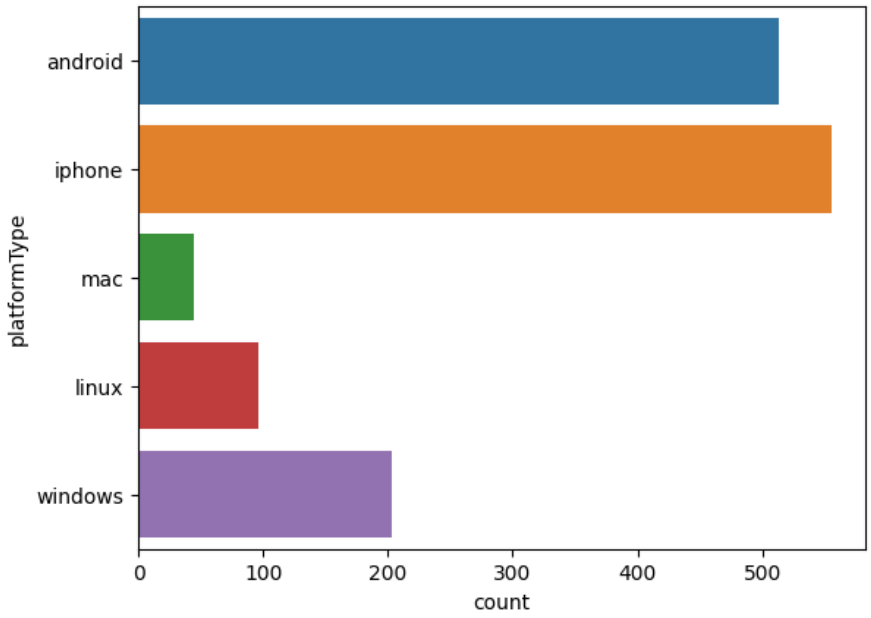


Figure 6Number of Users by platform

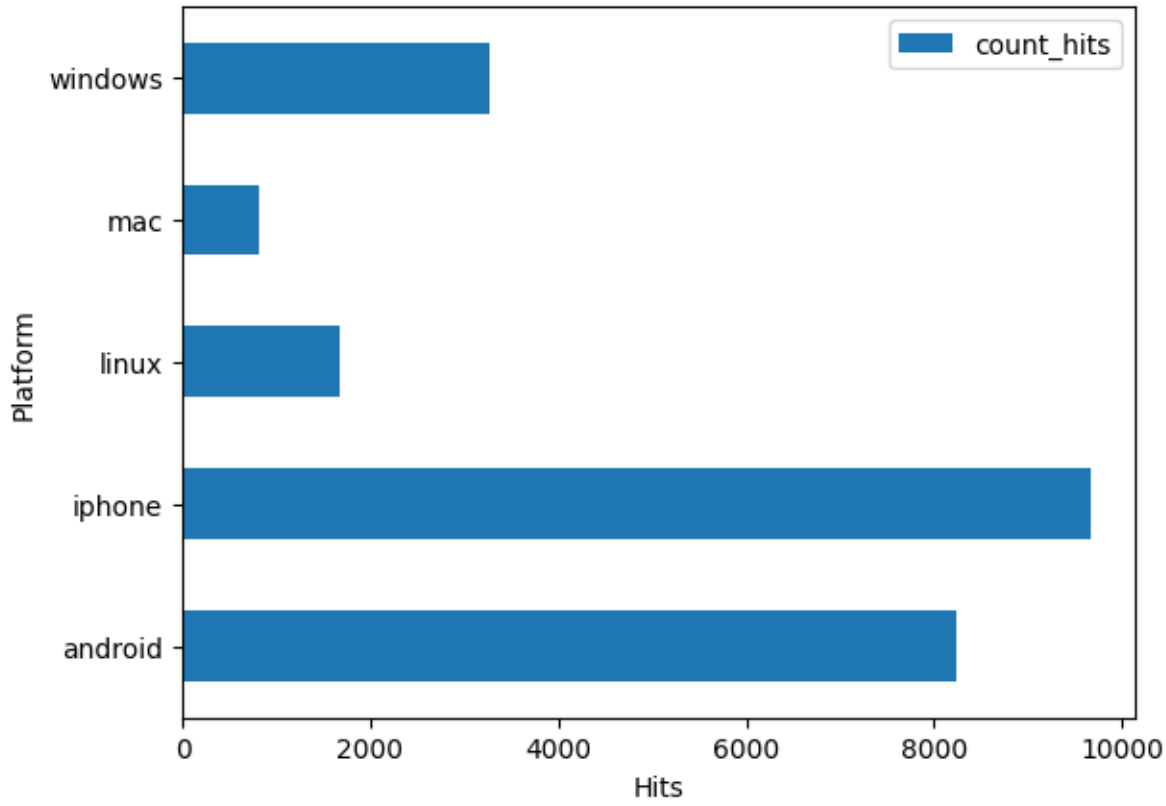


Figure 7Number of hits by platform

Figure 6 and Figure 7 show the number of users by platform and the number of hits by platform respectively.To graph figure 6 the platform type was grouped and the number of rows it has was summed before being plotted.To graph figure 7 the platform type was grouped again and for each platform the number of hits was summed.The number of users is an important metric as it tells us what platforms people are mainly playing the game on. Using this information the organisation can look at maintaining regular patches for iphone

for example as it has the most users along with android.The insights we can make from figure 7 is skill level by platform. However a limitation of this is that since one platform has more users they will have more hits. However when comparing both figures it is evident that the jump between bars is larger than in figure 6 revealing that it may not just be the number of users increasing the value.

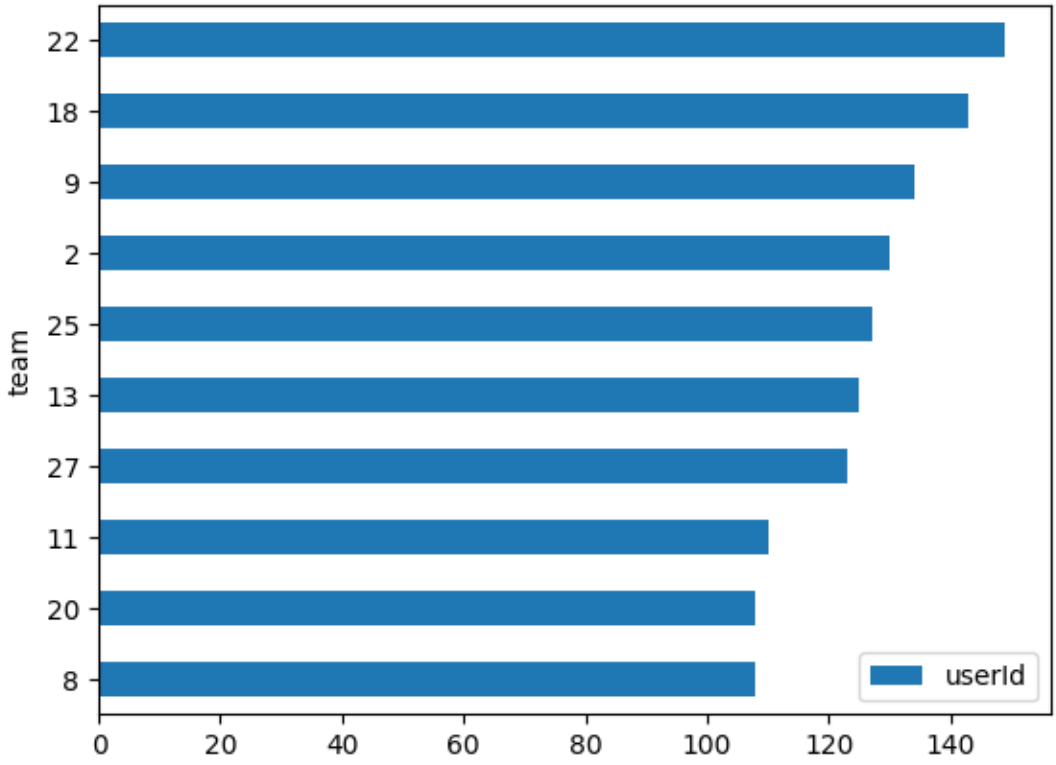


Figure 8Graph to show the teams with the most users

Figure 8 shows the 10 top teams with most users. To graph this each team was grouped and the number of rows counted for each group.Figure 8 helps us prepare the data for comparison to see if the team with most users equates to the best performance or not.

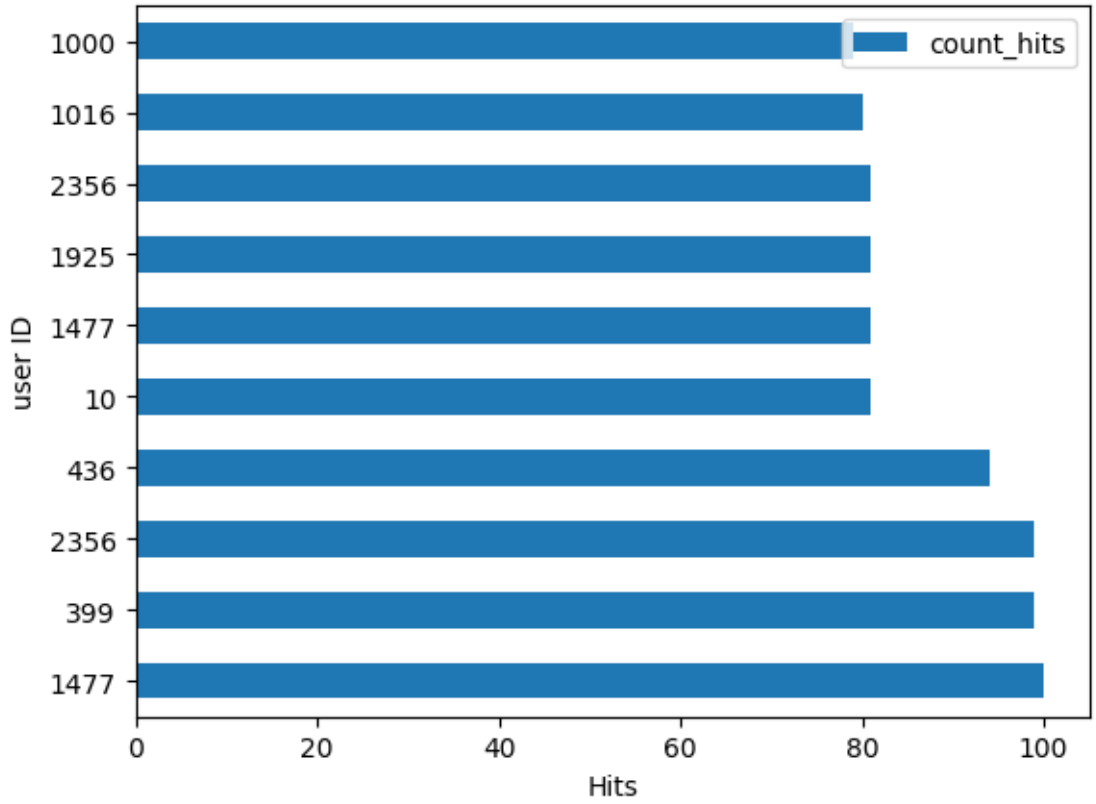


Figure 9Top Users by Hits

Figure 9 shows the 10 users with the most hits.To compute this graph the userID was grouped and the count hits was summed. This data prvides insight into the most skillful users in which going forward could set up some sort of ranked system or to create a leaderboard of some sort.

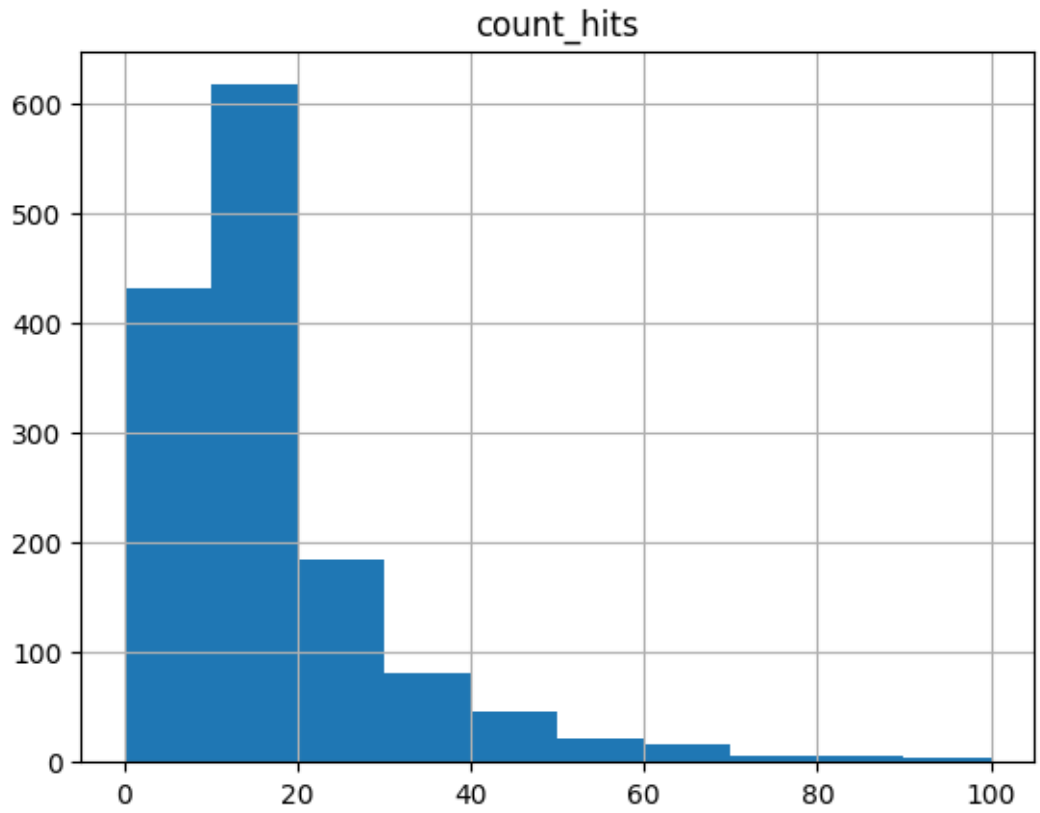


Figure 10Histogram for hit frequency

Figure 10 shows a histogram of the number of hits using the combined data dataset. The data has been split into 10 bins each of which contain 10 hits.This graph help us identify the distribution to see the average skill level.This graph shows normal distribution but with a right skewed distribution as is shown by the tail being on the right.This means that most of the values end up left of the mean which we know is around 16 according to the summary statistics show in figure 11.

	0	1	2	3	4
summary	count	mean	stddev	min	max
userId	4619	1189.9647109764019	691.0986309664246	0	2389
userSessionId	4619	17963.06798008227	7947.681126803095	5648	38722
teamLevel	4619	4.355704697986577	1.9246625516949587	1	8
platformType	4619	None	None	android	windows
count_gameclicks	4619	143.06300064949124	126.88339694316925	1	1207
count_hits	4619	15.705780471963628	13.986901900918744	0	121
count_buyId	1411	1.6832034018426647	0.9005088871085436	1	6
avg_price	1411	7.214323175053155	6.536501375588665	1.0	20.0

Figure 11Summary Statistics for combined data dataset

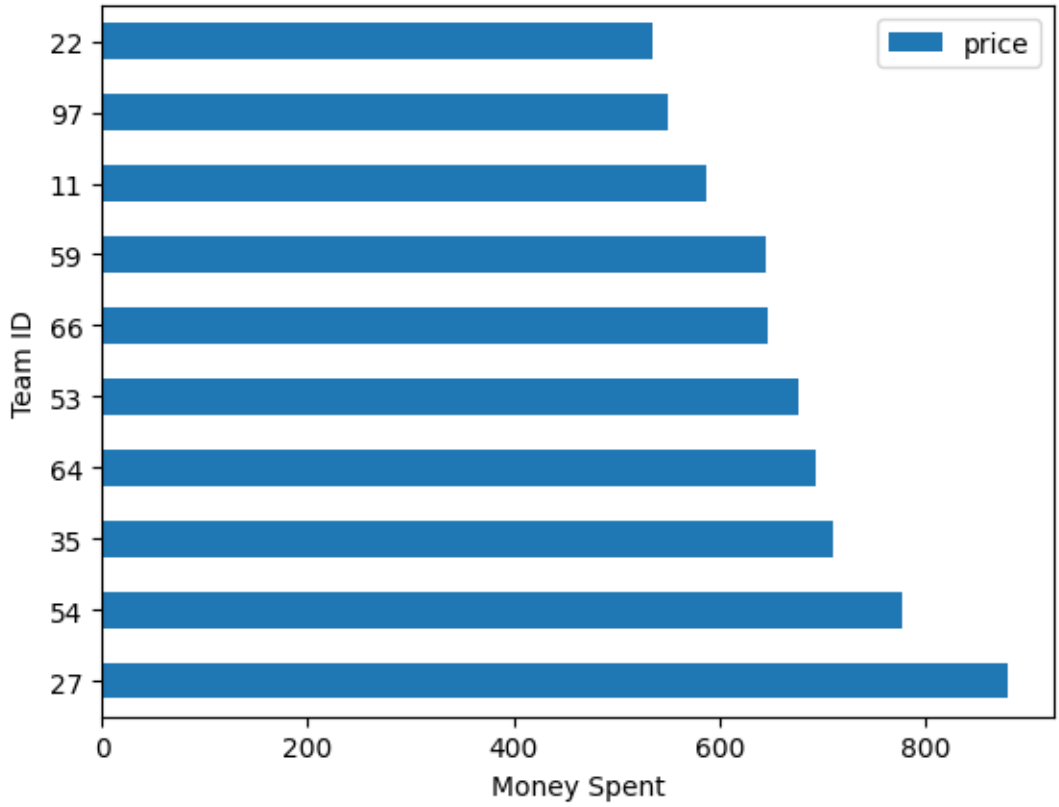


Figure 12Money Spent By team

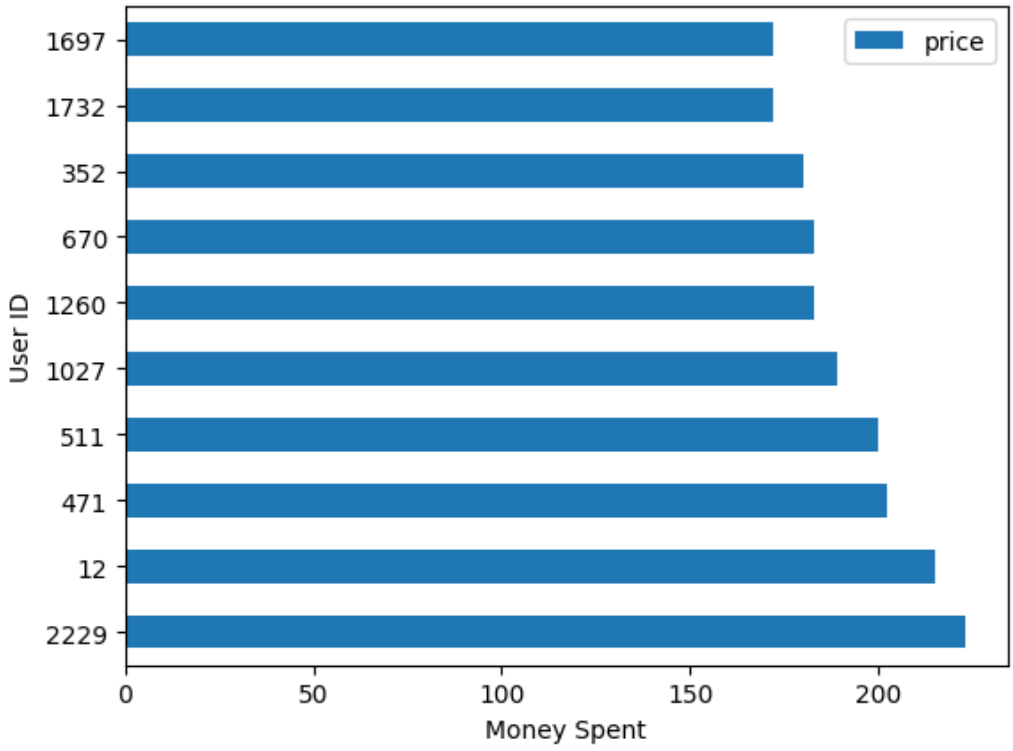


Figure 13Money Spent by User

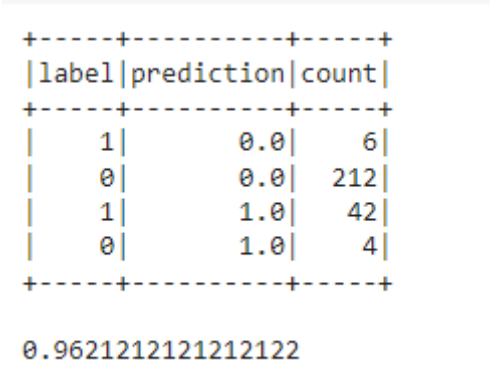
Figure 12 and 13 shows the money spent by users and by team. Figure 12 was processed for the buy-clicks data set and grouping the team ID then summing up the prices. Figure 13 was processed by grouping the user ID then summing up the prices.This data helps create insights for revenue. This helps identfify the teams and user that bring in the most money.

4. Machine Learning Models

4.1 Classification

4.1.1 Decision Tree

The goal for this machine learning model was to predict whether the player was a novice or an intermediate.To prepare the data for the decision tree the null values were first removed. The userId, userSessionId and the team level columns were dropped as they were not relevant to the machine learning model. A new column was added named label so that if the number of hits were greater than 25 then the column would return a 1 meaning they were an expert. If it was lower than a 25 the label column would show a 0 meaning that they were a novice.The platformType was then indexed as it was a string variable so it needed to be changed to a numeric variable. The columns being used for this model were the indexed value of the platform type,the number of gameclicksand the average price.The data was split into training and testing with an 80/20 split.The decision tree is then computed for the training dataset and then applied to the testing dataset.Figure 14 shows the confusion matrix and the accuracy which turned out to be 96% meaning it worked pretty well.Being able to predict the skill level of a user can help in multiplayer matchmaking and can even be used later to pull real time data from the game. A ranked system can also be implemented to make the game more competitive.



```
+-----+----------+-----+
|label|prediction|count|
+-----+----------+-----+
|    1|       0.0|    6|
|    0|       0.0|  212|
|    1|       1.0|   42|
|    0|       1.0|    4|
+-----+----------+-----+

0.9621212121212122
```

Figure 14Confusion Matrix for Decision Tree

4.1.2 Naïve Bayes

The goal for this machine was to predict whether the player was a big spender or not.The data had already been prepared from the decision tree but this time the main column being focused on was average price. If the average price was higher than 5 then the player would be considered a big spender but if the average price was lower than 5 then the player would be considered a low spender. The columns being used for this model were the indexed value of the platform type,the number of gameclicks and the number of hits.The data was split into training and testing with an 80/20 split.The naïve bayes algorithm is then computed for the training dataset and then applied to the testing dataset.Figure shows the confusion matrix and the accuracy which turned out to be 82.58% meaning it worked well but not as well as the decision tree and this could be for many reasons but a big one being that the data used was different.This can help us gather insights as to whether or not the users spend a lot of money on the game and if they do how to push more targeted ads to them in order to increase revenue

```
+-----+----------+-----+
|label|prediction|count|
+-----+----------+-----+
|    1|       0.0|   28|
|    0|       0.0|  121|
|    1|       1.0|   97|
|    0|       1.0|   18|
+-----+----------+-----+

0.8257575757575758
```

Figure 15Confusion Matrix for Naive Bayes Algorithm

4.2 Clustering

4.2.1 K means

For this section the Kmeans clustering algorithm was used in tandem with the silhouette method to get the ideal number of clusters.The data was first preprocessed and the columns that were to be used were game clicks,average price and hit count.Using a function in spark the ideal number of clusters were identified as shown in figure .As is visible 2 clusters had the highest value using the silhouette method.As we now have the ideal number of clusters we can run the k means clustering algorithm. This can then be placed in a dataframe with standardized values which can then be graphed in 3d as shown in figure.This shows that there are 2 main clusters with one being high in money spent but low game clicks and hits. This tells us that maybe they don't have time to play the game as much as they might work but have access to money so they can purchase as they see fit. The second cluster tells us that game clicks and hits are high but money spent is low. I think this may tell us that they have more free time on their hands and as a result have more time to focus on the game itself but less access to spendable money on the game.

```
Tested: 2 clusters: 0.8791124447845252
Tested: 3 clusters: 0.11592377959248912
Tested: 4 clusters: 0.16425527315275684
Tested: 5 clusters: 0.10958097753796746
Tested: 6 clusters: 0.24634170737002978
Tested: 7 clusters: 0.09846761768964858
Tested: 8 clusters: -0.03141173789374087
Tested: 9 clusters: 0.017989744261837862
```

Figure 16Silhouette score for each number of clusters

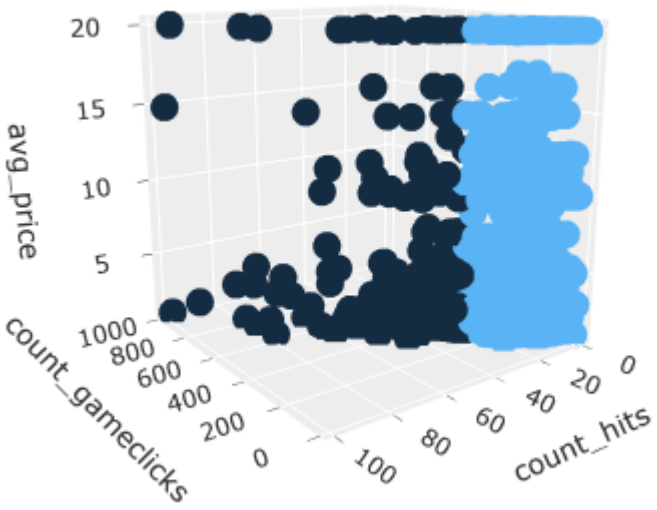


Figure 17 3d viz plot for K means clustering

4.2.2 Bisecting K means

Similarly to K means, Bisecting K means uses a similar method but has been shown a different way for the purpose of this assignment. Bisecting K means is a combination of K means and hierarchical clustering. The same columns were used for this algorithm. To begin the silhouette method was once again used and is shown in the form a graph in figure.As is visible by the graph 2 clusters is again shown as having the maximum value so that will be what is implemented in the algorithm.A pca scatter plot has been used to show the clustering for this one as shown in figure. The insights again are almost identical to the K means clustering but this is a good way to double check what insights are made from clustering.

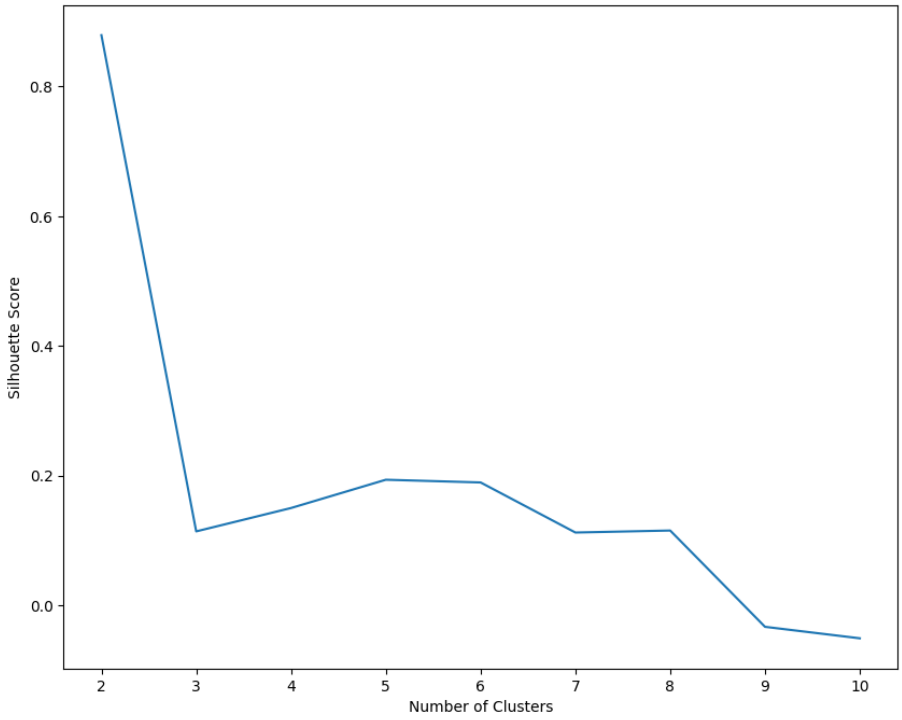


Figure 18Graph to show silhouette score

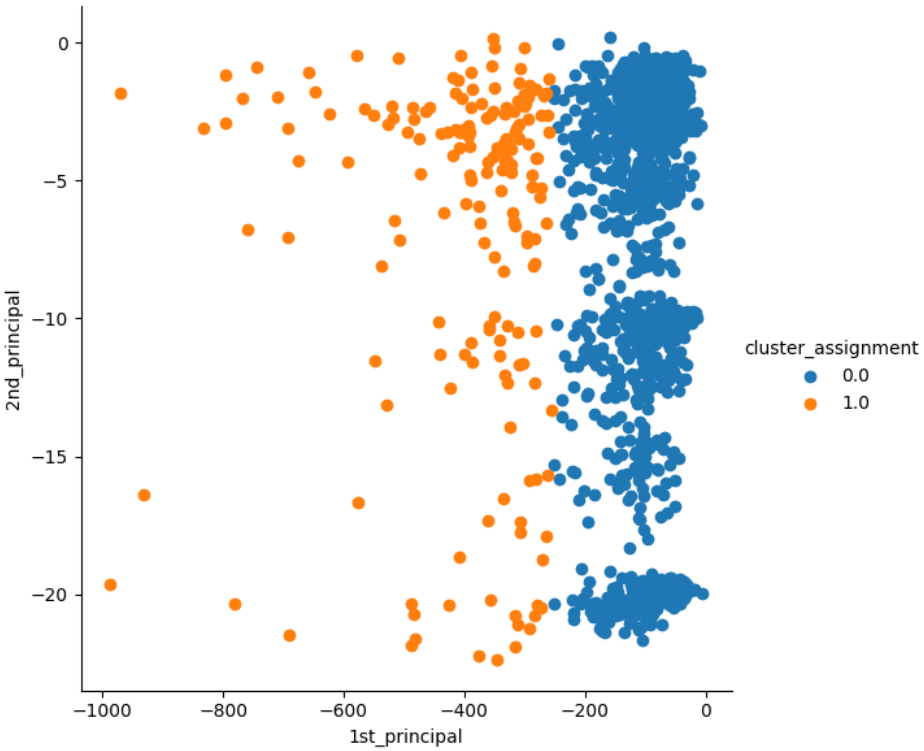


Figure 19Scatter plot showing Bisecting K means clustering

5. Graph Analysis

Graphs were created from 4 datasets coming from the chat data section. The 4 data sets were the join dataset,the leave data set,the mention dataset and the respond dataset. Each data set corresponds to a relationship. As is seen figure 20 Users can join and leave the chat session. Chat Item uses a user ID but is not treated as one in the graph. The Chat Item can mention a user. Chat Items are users that can respond to each other

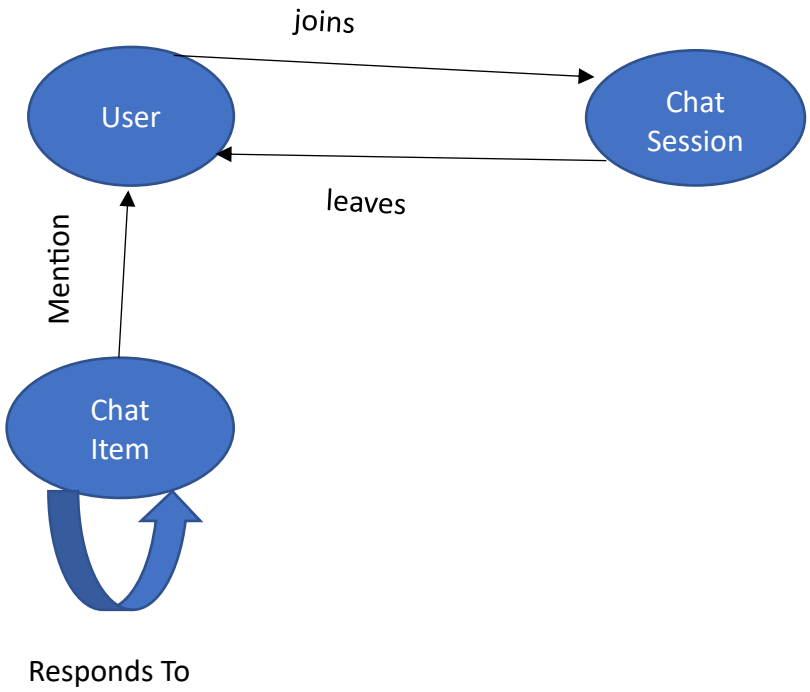


Figure 20Chat Data Schema

Figure 21 shows the 5 chat sessions with the most joins and leaves connected.Figure 22 shows the chat session node with the most relationships.Figure shows the relationship between one chat session node and many user nodes with the edges being the join relationship and leave relationship. This insight is important because it can be useful to adjust bandwith and resource delivered to the chat sessions with more activity.Figure 20 has 5 chat session nodes and 71 user nodes with 821 relationships.Figure 21 has 1 chat session node and 17 user nodes with 185 relationships.

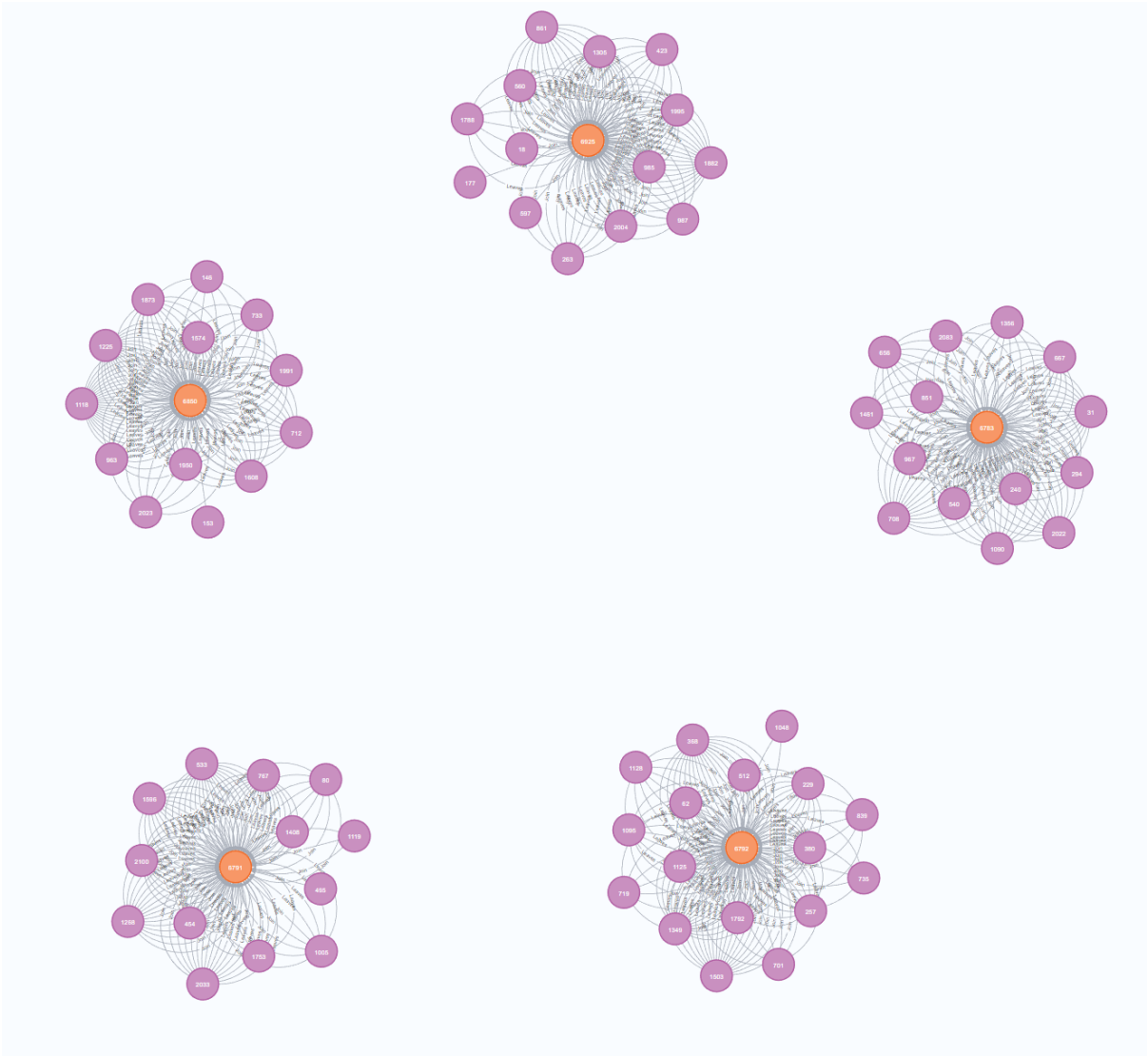


Figure 21Graphic visualisation of the 5 chat session most joins and leaves

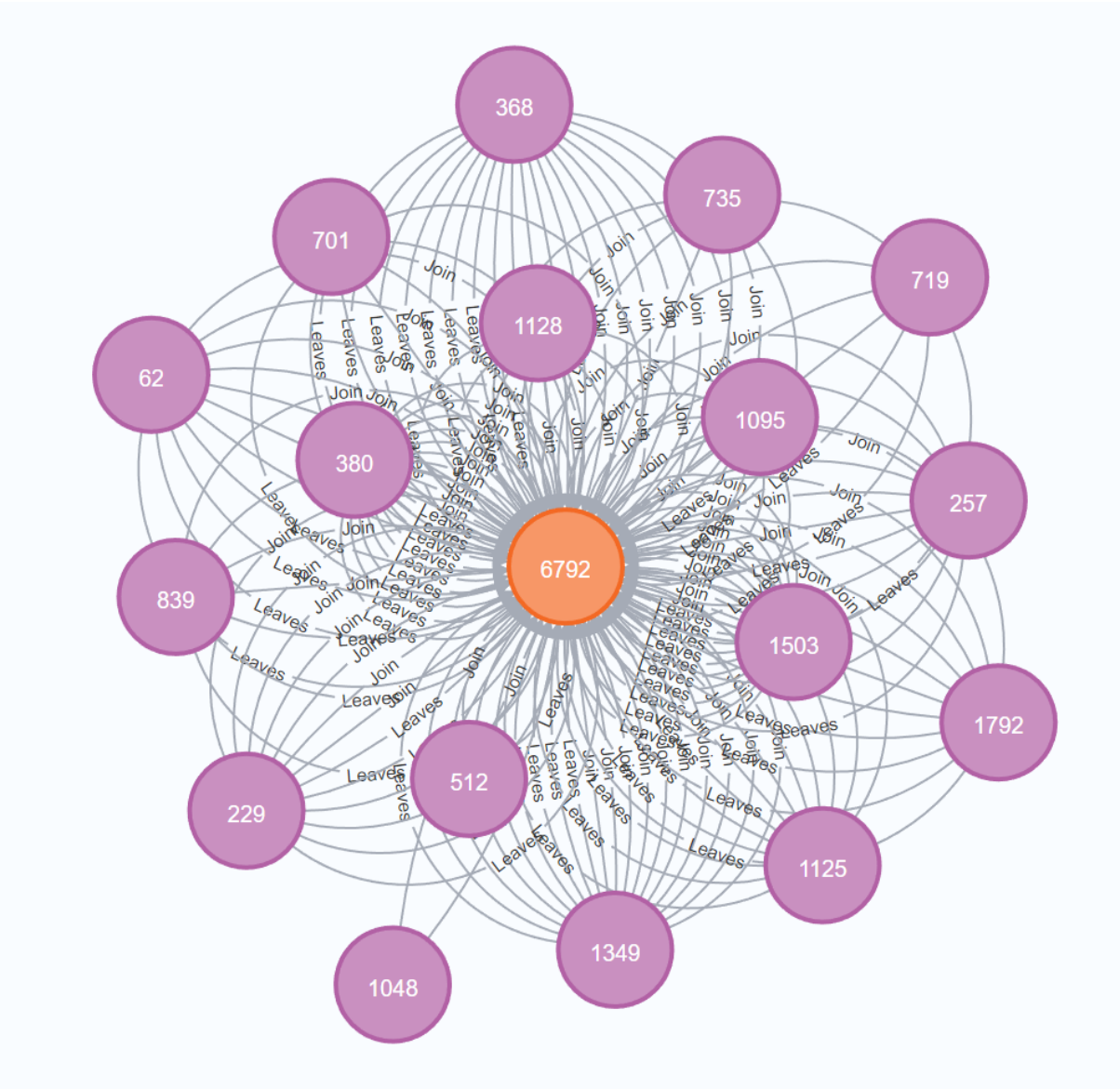


Figure 22Graphic Visualisation for the chat session with most joins and leaves

Table 1 shows the number of times a user is mentioned. This shows the most popular or active users and this can be used for ad insights as you can retrieve the OS information and started pushing through more ads in order to increase revenue. This could also mean that they spend more time on the game and therefore more money so pushing ways for him to spend money can be done.

	Users	chatitems
1	"131"	53
2	"1204"	47
3	"621"	47
4	"1428"	46
5	"1506"	46
6	"283"	42
7	"674"	42
8	"1482"	42
9	"1450"	42
10	"1989"	41

Table 1Chat Items By user

Figure 23 shows the mentioned edges together with the responding edges. This is because users can respond to other users but they can also mention these users therefore can be shown on the graph simultaneously. This gives us further insight on how active a user is as you can see if they are only responding or only mentioning or both.There are 10 user nodes denoted in pink. The other nodes are the users mentioning the user or responding to the user. Figure 24 shows this in greater detail whereas figure 23 shows the broad scope of it.

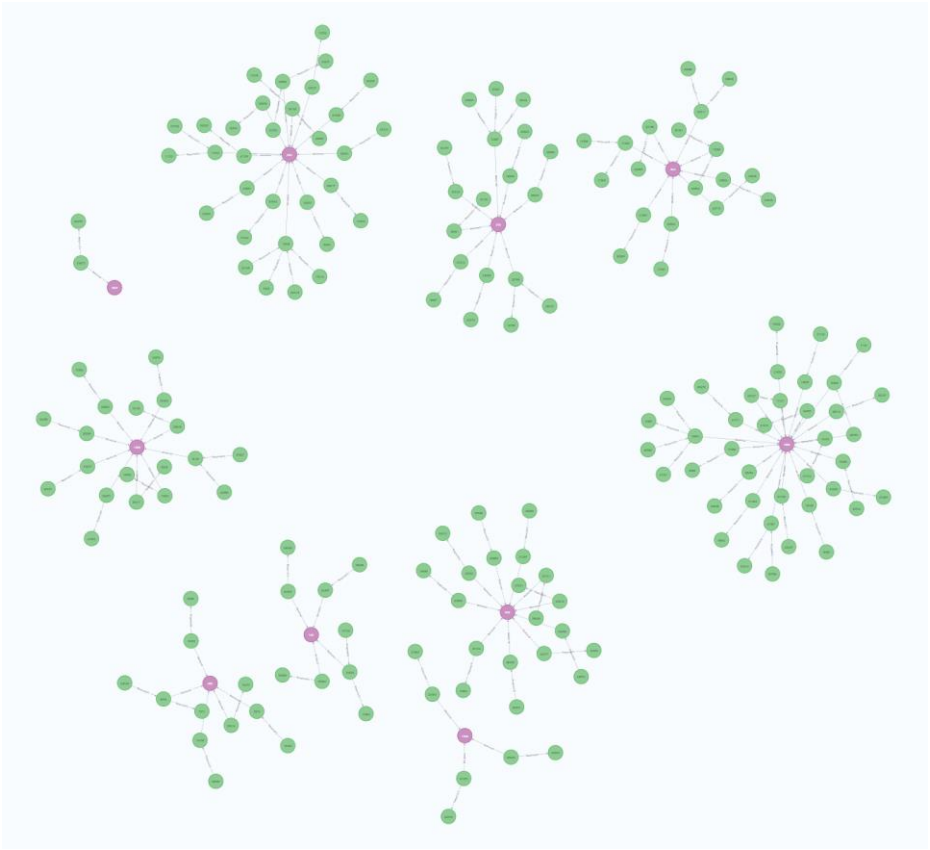


Figure 23Visualisation of mentioned and responded to edges

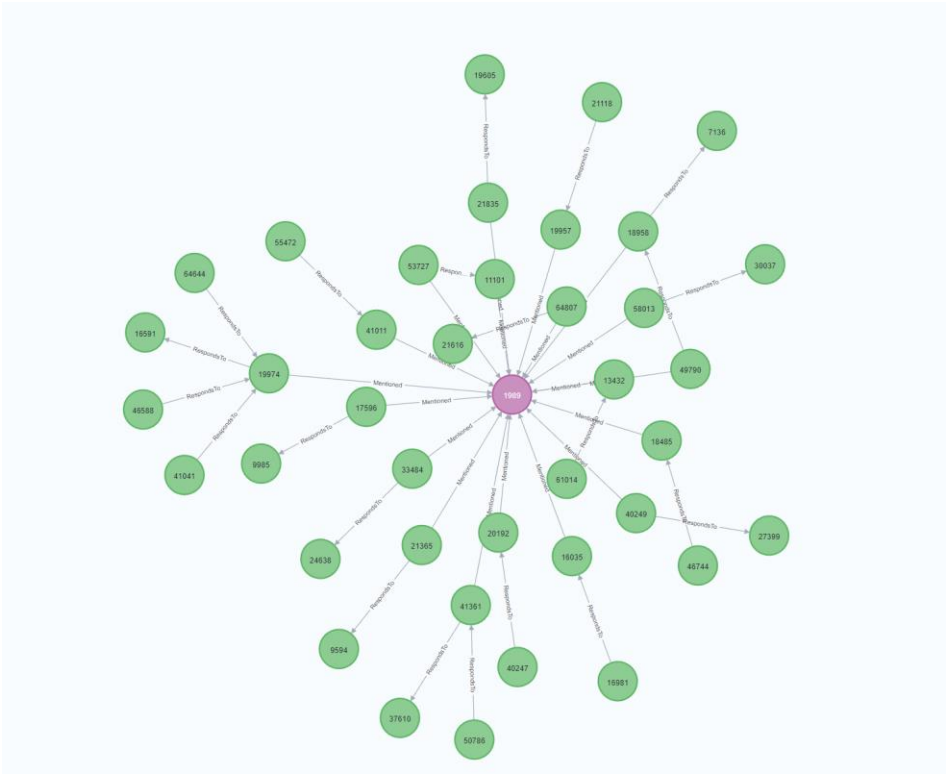


Figure 24Graphical Visualisation of user with most mentioned and responded to edges

Figure 25 shows the longest conversation chain which was 9 and as you can see it starts from node 52694 and ends at node 7803. It includes 10 nodes with 9 relationships.

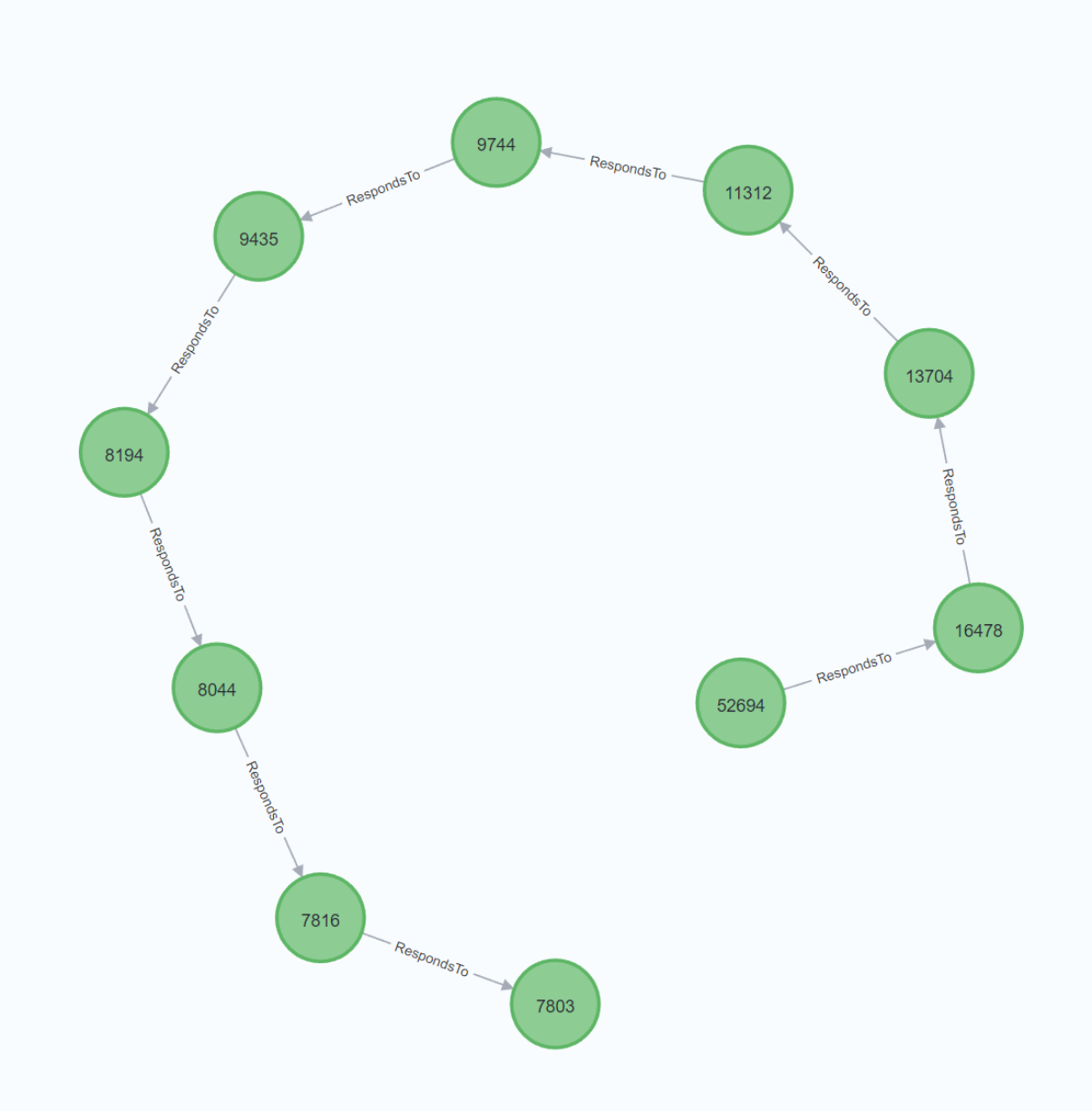


Figure 25Longest conversation thread

Table 2 shows the top 10 degree centralities for each of the relationship types.Degree Centrality can be defined as the number of edges a node has. The higher the degree the more central it is.

Join	Leave	Mention	Responds to
16	15	53	7
13	13	47	6
13	12	47	6
13	12	46	6
13	12	42	5
13	12	42	5
13	12	42	5
13	12	41	5
13	12	41	5
12	11	41	5

Table 2Degree Centrality by Relationship type

6. Big Data Ethics

Big data reveals quite a few ethical issues as there are various problems with having access to so much access to personal as well as data being so easily accessible to companies as well as using it for monetary value. In addition to that dealing with such a large volume of data can have many ethical implications.When talking about the ethics of big data it is important

The main ethical implications surround privacy and security and is an age old topic for debate.Privacy denotes and individual's willingness to offer up their personal information. However nowadays your privacy can be invaded via your digital footprint which is basically all your online activity. The use of digital footprints in data analytics raise ethical considerations related to not only privacy but consent and data protection.Privacy vs Security is one of the greatest debates as privacy focuses more on consent whereas security focuses more on protecting the data from malicious activities and data breaches such as hacks.on the opposite end a digital footprint can have advantages such as growing your brand or informing people on certain issues(Jain et al., 2016).

Another implication of big data relates to the bias and fairness of the data. The example chosen for this is the medical field. An example of this bias is when a Convolutional Neural Network was run for skin lesion classification in which black patients made up of 5-10% of patients.As a result of this when the network was run on the dataset the accuracy by about half when classifying black patients.This is especially a problem as black patients have the highest mortality rate for melanoma with an estimated survival rate of 70% vs 94% for white patients(Norori et al., 2021).

The last ethical implication of big data looked at is the economical and social implications of big data. On one hand there are benefits where big data can save companies lots of money as well as making them lots of money. For example a report from the Mckinsey global institute projected that big data could generate an additional $3 Trillion in value per year (Kennedy, 2022). On the other hand, it can have risks. The 2 big ones are the use of big data by bad people and the other is the unintentional misuse of data. Common examples of how

big data can be used to do bad things are activities such as phishing, bank fraud and insurance all of which stem from big data and is used frequently by organised crime groups(Hillier,2022). Unintentional misuse of big data refers to the systematic problems with big data. For example in 2018 a self driving uber killed a jaywalker as it could only predict objects near crosswalks.

7. Conclusion: Finding and Recommendation

In conclusion this report evaluated the batch processing,real time processing,stream processing and hybrid processing as well as it's frameworks and showed how it dealt with the 3Vs of big data.It showed the pros and cons of each paradigm. Following that EDA was done on the data for the imaginary game 'Catch the Flamingo' to prepare it to draw further insights as well as preparing it for the 4 machine learning models. The EDA was then visualised and analysed for insights. 4 machine learning techniques were then applied to the datasets of which 2 were classification techniques and 2 were clustering techniques. Following this graph analysis was done for chat data and finished off with an analysis of the ethical implications of Big data.

https://colab.research.google.com/drive/1rEYqWIk_H5dTP8iSUVROvwma78DNkgjf#scrollTo=RWhQmwYfWEC6 - Link to code

https://github.com/josh236916/Catch-the-pink-Flamingo - Github Link

References

Batch processing vs. real-time processing (2022) *IT Procedure Template*. Available at: https://www.it-procedure-template.com/batch-processing-vs-real-time-processing/ (Accessed: 19 May 2023).

Contributor, T. (2022) *What is event-driven application?: Definition from TechTarget, IT Operations*. Available at: https://www.techtarget.com/searchitoperations/definition/event-driven-application#:~:text=An%20event%2Ddriven%20application%20is,for%20system%20hardware%20or%20software. (Accessed: 19 May 2023).

Costa, C. and Santos, M.Y. (2017) *Página principal, Reposit*  *riUM*. Available at: https://repositorium.sdum.uminho.pt/handle/1822/46855 (Accessed: 19 May 2023).

García-Gil, D. *et al.* (2017) *A comparison on scalability for batch big data processing on Apache Spark and Apache Flink - Big Data Analytics, BioMed Central*. Available at: https://bdataanalytics.biomedcentral.com/articles/10.1186/s41044-016-0020-2#Sec12 (Accessed: 19 May 2023).

James G. Shanahan NativeX and UC Berkeley *et al.* (2015) *Large scale distributed data science using Apache Spark: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and data mining, ACM Conferences*. Available at: https://dl.acm.org/doi/abs/10.1145/2783258.2789993 (Accessed: 19 May 2023).

Lam, C. (2010) *Hadoop in action, Google Books*. Available at: https://books.google.co.uk/books?hl=en&lr=&id=8DozEAAAQBAJ&oi=fnd&pg=PT15&dq=hadoop&ots=ryHo430eHv&sig=FKTyzft8RvP0F_Qe1Y8Z8OC7IHE&redir_esc=y#v=onepage&q=hadoop&f=false (Accessed: 19 May 2023).

Low latency: What it is, meaning & definition: Informatica UK (2023) *Low Latency: What It Is, Meaning & Definition | Informatica UK*. Available at: https://www.informatica.com/gb/services-and-training/glossary-of-terms/low-latency-definition.html#:~:text=Low%20latency%20describes%20a%20computer,access%20to%20rapidly%20changing%20data. (Accessed: 19 May 2023).

Paulin, K. (2021) *What is batch processing and how has it evolved?, Stonebranch*. Available at: https://www.stonebranch.com/blog/what-is-batch-processing (Accessed: 19 May 2023).

Pfandzelter, T. and Bermbach, D. (2019) 'IOT data processing in the fog: Functions, streams, or batch processing?', *2019 IEEE International Conference on Fog Computing (ICFC)* [Preprint]. doi:10.1109/icfc.2019.00033.

Segal, T. (2022) *What is Big Data? definition, how it works, and uses, Investopedia*. Available at: https://www.investopedia.com/terms/b/big-data.asp (Accessed: 19 May 2023).

Tozzi, C. (2022) *Process streams vs. batch processing: When and why to use each, Precisely*. Available at: https://www.precisely.com/blog/big-data/big-data-101-batch-process-streams (Accessed: 19 May 2023).

What is ... (2022) *Amazon*. Available at: https://aws.amazon.com/what-is/batch-processing/ (Accessed: 19 May 2023).

What is batch processing? (advantages and disadvantages) (2022) *What is batch processing? (Advantages and disadvantages)*. Available at: https://uk.indeed.com/career-advice/career-development/what-is-batch-processing (Accessed: 19 May 2023).

What is Kafka, and how does it work? A tutorial for Beginners (no date) *Confluent*. Available at: https://developer.confluent.io/what-is-apache-kafka/ (Accessed: 19 May 2023).

Yang, W. *et al.* (2013) 'Big data real-time processing based on Storm', *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications* [Preprint]. doi:10.1109/trustcom.2013.247.

(No date) *Apache storm | cloudera*. Available at: https://www.cloudera.com/products/open-source/apache-hadoop/apache-storm.html (Accessed: 19 May 2023).

Norori, N. *et al.* (2021) 'Addressing bias in big data and AI for health care: A call for open science', *Patterns*, 2(10), p. 100347. doi:10.1016/j.patter.2021.100347.

Low latency: What it is, meaning & definition: Informatica UK (no date) *Low Latency: What It Is, Meaning & Definition | Informatica UK*. Available at: https://www.informatica.com/gb/services-and-training/glossary-of-terms/low-latency-definition.html#:~:text=Low%20latency%20describes%20a%20computer,access%20to%20rapidly%20changing%20data. (Accessed: 19 May 2023).

Jain, P., Gyanchandani, M. and Khare, N. (2016) 'Big Data Privacy: A Technological Perspective and Review', *Journal of Big Data*, 3(1). doi:10.1186/s40537-016-0059-y.

Kennedy, J. (2022) *Big Data's economic impact*, *Committee for Economic Development of The Conference Board*. Available at: https://www.ced.org/blog/entry/big-datas-economic-impact#:~:text=A%20report%20from%20McKinsey%20Global,would%20benefit%20the%20United%20States. (Accessed: 19 May 2023).

Kala Karun and K. Chitharanjan, "A review on hadoop — HDFS infrastructure extensions," 2013 IEEE Conference on Information & Communication Technologies, Thuckalay, India, 2013, pp. 132-137, doi: 10.1109/CICT.2013.6558077.

Hillier, W. *et al.* (2022) *Is big data dangerous? the risks uncovered [with examples]*, *CareerFoundry*. Available at: https://careerfoundry.com/en/blog/data-analytics/is-big-data-dangerous/ (Accessed: 19 May 2023).