Big Data Solution: Catch The Pink Flamingo By Eglence Inc

Hamza Waheed, ID: 21171402 [*1]

Word Count: 5781

Abstract

Growth of the digital universe has compelled companies to use Big Data technologies and change their way on how they store, operate and handle large amounts of data.

This report goes over the distinction between traditional data and big data. The need that gives rise to big data, and technologies that provide the big data solutions. The report then touches on the problems that arise when using these technologies and how empirical studies and solutions are tackling these problems.

There is also an emphasis on the ethical part and how future laws may impact the way companies deal with these big data solutions.

The report develops a big data solution using PySpark & Neo4j by doing Exploratory Data Analysis (EDA), Machine Learning & Graph analysis on the "Catch The Pink Flamingo" game of the imaginary company called "Eglence Inc". The recommendations from the analysis are given at the end of the report.

1. Introduction

"Big data" is a term often used in today's technological circles. It's considered to be the new buzzword. But it surrounds a lot of confusion as to what big data is. How it is different from the traditional data? Why separate tools are needed for handling big data?

Big data is classified as the type of data that's too big, too fast and too hard for the traditional tools like RDMS and queries to process (Madden, 2012).

By 2025, It is expected that the amount of data generated each day would reach 463 exabytes. Big companies like Google, Facebook, Microsoft, Amazon process petabytes of data on a daily basis. It is considered that these companies hold 1200 petabytes of data (Vuleta, 2022). Such humongous data gives rise to the need of new solutions that encompass the Big Data Paradigm.

Big data is characterized in the form of three "V"s. Volume, Velocity and Variety. These three components are shown in fig 1 (Sagiroglu & Sinanc, 2013).

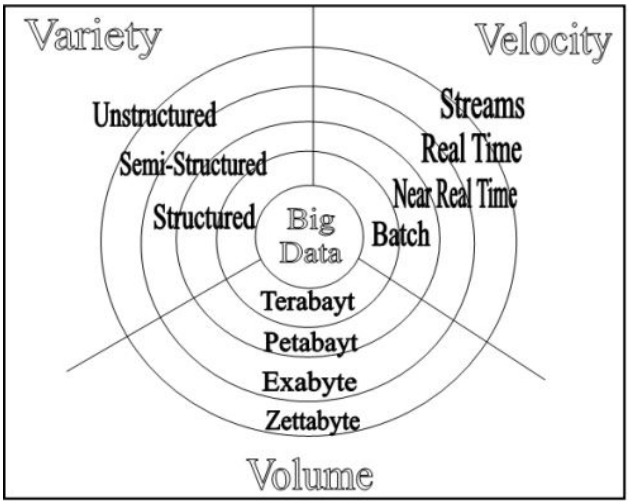

Figure 1. Three main components of Big Data

"Volume" suggests increasingly dealing with large amounts of data. The ever-increasing growth of data outrun traditional storage & analysis techniques.

"Velocity" refers to the rate at which data is processed. Big data systems requires real time data handling. Bank transactions, Fraud detection, Health are the examples of industries that require data to be processed immediately.

"Variety" highlights different forms & types of data i.e. Audio, Video, structured, unstructured, semi-structured. "Variety" shows that traditional storage methods that require strict schema regulations aren't suitable for handling & processing these types of data. Additionally, these traditional methods may not provide suitable insights out of the box

[1]M.Sc. Big Data Analytics, School of Computing and Digital Technology, Birmingham City University, UK. Correspondence to: Hamza Waheed <Hamza.Waheed2@mail.bcu.ac.uk>.

Big Data Management Report, CMP 7203 - Big Data Management.

for analysis of the data.

1.1. History

Performing operations on the dataset is a difficult task. In today's world, the amount of data expected to be generated is measured in exabytes. By 2025, the amount of data on the internet would exceed the brain capacity of everyone on the Earth (Davis, 2012).

As an example, in 2019, the number of google searches per minute were 3.8 million. Now take a look at a more industrial example. Imagine a query that performs running demographic analysis based on these searches. Running such a query would be extremely expensive and exhaustive. Especially, when the operations involve analysing the entire dataset. No single solution is available to tackle such problem without any pitfall. However, a series of different toolset can be employed to improve processing. One such framework is Map-Reduce.

With the realization of growing internet space, in 2004, Google introduced Map-Reduce framework for managing and handling big data. Map-Reduce is the most prominent framework when it comes to handling of big data. Map-Reduce allows many advantages over traditional methods. It's linearly scalable, fault-tolerant, and hides the complexity of managing many clusters (processing machines). These features have made Map-Reduce an industry standard, with many frameworks building on top of it (Shahrivari, 2014), (White, 2015).

Map-Reduce consists of three components that provide a distributed, scalable, fault-tolerant model. Google's Map-Reduce framework is composed of Map-Reduce engine, Google File System and distributed NoSQL database called BigTable (Shahrivari, 2014).

Apache also introduced it's distributed model called Hadoop, that builds on top of Map-Reduce framework. Apache Hadoop comprises a series of products that allow distributed processing across a cluster of servers. These technologies are, Hadoop Map-Reduce & Hadoop YARN as execution engine. HDFS and Hbase for database. Also, Hadoop provides many modules that provide storage, database, data warehouse and APIs to work with big data.

2. Literature Review

Big data processing paradigms categorize systems into batch, real-time (stream), graph and machine learning processing (Jain et al., 2016). This section gives the in-depth explanation of the these processing paradigms of big data.

2.1. Batch Processing

Batch processing is a Map-Reduce based model for parallel computing paradigm (Chandio et al., 2015).

Batch Processing solves the scalability and fault tolerance issue when dealing "Volume" part of big data. Batch processing systems like Hadoop comprise two main components, Hadoop Distributed File System (HDFS) and Map-Reduce. HDFS is a distributed, fault-tolerant system, while Map-Reduce is a computational framework that integrates with HDFS and scales horizontally. Horizontal scaling refers to more processing power that can be added to the system and Hadoop will take care of the complexity, hence the scalability aspect of Hadoop (Chandio et al., 2015).

Batch processing systems like Hadoop solve the processing problem in two steps. First, it creates an immutable copy of the original data. Second, it performs function (query) and store it's output on the disk (White, 2015).

Batch processing architectures are based on Master Slave model. It consists of two types of nodes, i.e. master, & slave. The master node divides the operation into smaller chunks and delegates it to worker nodes. After the worker nodes have performed their operations. The result is sent back to the master node. Master combines all the parts to provide the solution to the problem at hand (Yaqoob et al., 2016).

However, the problem with batch processing systems is, it deals with existing data and doesn't take into account any new data. Once the batch process starts, if the underlying data changes, it won't account for the new data. Thus, batch processing suffers with high latency issue. However, batch processing is suitable in scientific computation, graph and social network analysis where latency isn't required (Casado & Younas, 2014).

Table 1 provides a list of batch based processing tools.

Table 1. Batch Processing Technologies

Tools	Description
Hadoop	To process big data using Map-Reduce Framework.
Tableau	To process large amount of data for insights.
Karmasphere	For insights and customer analytics on Big Data.
Pentaho	For Data Mining, ETL and insights on Big Data.

2.2. Real Time Processing

Real time processing deals with the ”Velocity” aspect of big data and is based on the same principles as the batch processing. Real time processing deals with data streams that are captured in real time to generate near real-time response, It tackles the problems of high latency that arise with batch processing (EdPrice-MSFT).

The main pipelines for real-time processing involve ingestion, processing and storing the data chunks. The stream of data chunks is in the semi structured form, such as JSON, XML. Data arrives during the ingestion stage. After ingestion, transformation or processing on the data is applied. After necessary functions have been applied, the data is then stored for consumption.

Real Time Processing has low latency, high throughput streaming. It achieves so by processing in-memory data. Unlike batch processing, which was designed to store data on disk (Casado & Younas, 2014). Thus, fixing the issue of high latency with batch processing. Also, real-time processing has short processing time (Latency) measured in milliseconds.

In contest to batch processing, real-time processing often process data in small chunks. This is why it’s a real challenge in real-time processing, to perform intense computations on the incoming data stream without blocking the ingestion pipeline. Also, it’s important to provide the processed data in real-time for generating real-time insights (EdPrice-MSFT).

Many systems like Banking transaction, Fraud Detection, weather forecast, traffic routing systems, navigation and healthcare systems rely on real-time solutions. Table 2 gives a list of tools for real-time processing (Yaqoob et al., 2016).

Table 2. Real-time processing tools

Tools	Description
Storm	To process large amount of real-time data.
Splunk	Provides capturing real-time data for insights & report generation.
Apache Kafka	Provides event-store and stream processing platform for in-memory analytics for decision-making.
SQL Stream s-Server	Streaming based SQL platform for managing live data feeds on the cloud.

2.3. Hybrid Processing

Many applications require a combination of both batch and real-time processing. This model is called as hybrid processing and was first introduced as Lambda Architecture by Nathan Marz and James Warren in 2015. Lambda Architecture is based on building the Big Data System as layers, where each layer builds upon the layer beneath it (Marz & Warren, 2015).

It accomplishes so by pre-computing the arbitrary operations and saving it as views. Then, calculating the requested query on those views

In Lambda Architecture, instead of running ad-hoc query on the entire dataset. The queries are run on pre-computed results as a set of views and then query those views to get the result of the desired operation. The idea is, instead of running ad hoc queries on entire data. It is efficient to precompute the desired results as a set of views and then query those views (Liu et al., 2014).

Lambda Architecture comprises three layers, Batch Layer, Real-TIme Layer and Serving Layer. The batch layer computes the functions (Views) on Master dataset and this process is performed repeatedly. By the time the results are available by the batch layer, new data is received.

This new arrived data is handled by a parallel speed layer. Speed Layer main objective is to produce ”Views” on this new data that can be queried with low latency. Speed Layer achieves this by taking into account of the recent data and is updated shortly after the new data arrives. If computation is done on each update, for large applications this approach is still computationally expensive and has latency. To fix this various algorithms are employed, one technique is to reuse the previous real-time ”View” and update it as the new data comes. Building on top of the previous computation performed.

Any query against the data is answered by the serving layer, that is, by querying both the batch and speed layer views in the serving layer. The results of both the views are merged, and the result is returned to the user (Liu et al., 2014). A visual representation of Hybrid processing from the book (Marz & Warren, 2015) is shown in Fig 2.

Table 2.3 gives a list of tools based on hybrid-architecture.

Table 3. Tools based on Hybrid Processing

Tools	Description
Apache Flink	An open-source, unified stream and batch processing framework by Apache
Apache Spark	An open-source, unified analytics engine based on hybrid Architecture

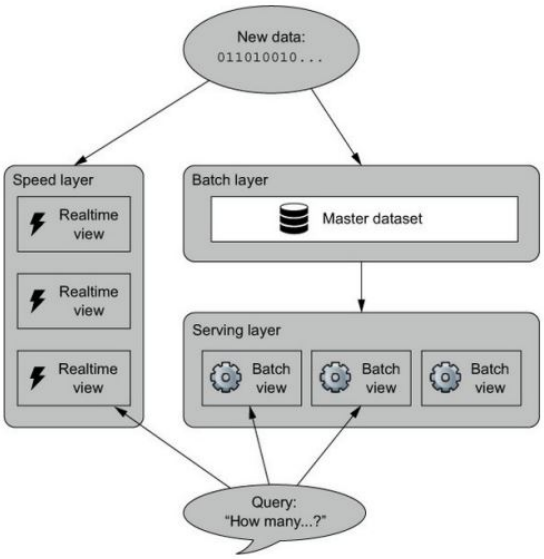

Figure 2. Workflow of Hybrid Processing

2.4. Machine Learning (ML)

Machine Learning is the discipline of knowledge discovery and making intelligent decisions by enabling machines to automatically improve themselves using sophisticated mathematical models (Oussous et al., 2018). The domain of Machine learning has led to huge social impact ranging from computer vision, natural language processing (NLP) and I.O.T.

Machine learning techniques have gained huge importance in data-intensive fields such as astronomy, music production and biology. Machine Learning has provided solutions for knowledge discovery and unravelling hidden patterns in these fields.

However, with the ever-growing data, the traditional ML methods of learning from databases and datasets have been pushed to their limits.

For ML algorithms to gain insights from large data system & storage, distributed & parallel learning systems are gaining attention, in which the learning is carried out in a distributed & parallel manner. Instead of gathering data into a single workstation, modelling is done on a multicore system and in parallel fashion, saving both cost and time. Decision rules, stacked generalization, meta-learning, and distributed boosting are the examples of distributed & parallel ML on Big Data (Qiu et al., 2016).

2.5. Graph Analysis

Graph analysis is one of the most profound techniques to model social networks, communities, interactions and complex patterns into graphical representation. With the rise in Big Data Technologies, graph technologies have been developed to leverage the power of parallel distributed computation for efficient processing and scaling.

Several models such as Message Passing Interface (MPI), Map-Reduce and Vertex-Centric Bulk Synchronous Parallel (BSP) has been developed to partition the graph into sub-graphs such that each piece is independently processed (Nisar et al., 2013). However, It's worth noting that graph partitioning and calculation is a NP-complete problem and has been an extensive area of research. Table 4 lists distributed Graph Analysis tools used;

Table 4. Distributed Graph Analysis Tools

Tools	Description
Apache Hama	A distributed computing framework based on bulk synchronous parallel processing for network & graph analysis.
Megalytic	Client reporting and dashboard tool for graph analysis.
GeoGebra	Academic tool for geometry, calculus and matrix learning.

3. Ethics

Like any new technology, that has its pros and cons. Big Data also doesn't come free from the ethical problems that come with the Volume, Variety and Velocity aspect of Big Data. With the ever-growing concerns about privacy. New laws like California Consumer Privacy Act (CCPA) and European Union's General Data Protection Regulation (GDPR) have regulated on the definition of "Personal Data" and how businesses collect and use these data (Matt, 2022). Upcoming future laws and pressure from regulating bodies have forced big data solution provider to enforce strict policies in handling of Big Data.

Big data like any technology is ethically neutral. This means that it doesn't come with its own sense of right and wrong. However, the organizations that use Big Data have their values. It's by asking questions that we can tackle the ethical issues that arise with Big Data (Davis, 2012).

Ethical practices, however, come from the ethical inquiry, that stems from the organization values. Specifically, when an organization asks, "What is the right thing to do?". However, ethical concerns that arise from Big Data Management can be addressed by organizations by ensuring standards in

three phases: Ethics in data generation, data storage & data processing.

3.1. Ethics in Data Generation

Data generation can be labelled as active & passive data generation. By active generation, the data owner explicitly give permission for the usage of data (Abadi et al., 2003). With the I.O.Ts, sensors and smart devices collect & stream data every single second. sensors like Proximity, Gyroscope and Magnetometer etc. stream data even in the idle state, without data ownerś knowledge.

Minimization of privacy violation is a serious challenge for the technology providers. The privacy violation is being tackled by major I.O.T providers like Google & Apple by introducing restriction options & abstraction of the data. "Sign in with Apple ID" is an example of such abstraction for the userś personal data.

3.2. Ethics in Data Storage

With the boom of cloud & big data storage systems, designing the storage systems that not only ensure data security but are also scalable to the distributed systems for processing is a serious computational & technological challenge.

Breach of the big data storage systems can expose the privacy of the user at risk. Therefore, ensuring privacy at the big data storage stage is of real importance. Data security of the individual on storage system has three dimension confidentiality, integrity and availability. The first two approaches deals with userś privacy right, while availability refers to consumption of the data for insights generation and processing by third parties (Jain et al., 2016).

Privacy of the individual on Big data storage systems, can involve conventional security measure at file, database and on application level that stores the data (CHENG Hongbing, 2015), (Jain et al., 2016). Additionally, existing mechanisms like public key encryption (PKE) where the senderś data is encrypted to only be decrypted by the valid recipients also strengthen the security of the userś data. (Jain et al., 2016) discuss approaches such as Attribute based encryption, Storage path encryption (Storage encryption of big data on cloud) and Hybrid cloud that utilizes a blend of on-premises, private cloud and third-party.

3.3. Ethics in Data Processing

Like the data generation & data storage phase, itś important to address the ethical concerns that may arise during Big Data processing phase, discussed in section 2.

The two main concerns in processing phase are to safeguard data from unsolicited disclosure and second concern is to ensure individual agnostic modelling of the data. For example, revealing personal information that can be harnessed to recreate the entire history of a person.

Any organization with the right access and tools can easily re-create the history of a user with the available data (big). Therefore, (Jain et al., 2016) suggests "De-Identification", that is sanitizing (stripping data of individual information) and transformation of data before machine learning modelling phase.

Techniques such as "k-anonymity", "l-diversity" and "t-closeness" replace quasi-identifiers in data with less particular, but semantically consistent values to reduce the impact of re-identification of the individual through modelling (Li et al., 2007), (Jain et al., 2016).

4. Methodology

"Catch The Pink Flamingo" is a multiplayer game developed by imaginary company "Eglence Inc". This report develops a Big Data solution using Spark & Neo4j to extract insights from the datasets that come from this game.

In this section, a brief introduction is given for the dataset used. The various preprocessing steps employed for doing Exploratory data analysis (EDA), applying Machine Learning techniques i.e. Clustering & Classification and graph analysis. The dataset is a combination of "csv" files. Table 5 gives a brief Introduction of "csv" files used for the exploratory data analysis. "combined-data.csv" file was used to perform Clustering & Classification tasks.

4.1. Pre-Processing

Different pre-processing & data transformation techniques were employed to transform the data for the analysis. A new column named "age" was added to "users.csv" file. "age" of the user was calculated using the "Date of birth" column. Furthermore, "age-group" column was added to segment the users into 8 subgroups. The "country" column of "user.csv" was converted into continent column. This step was used to check the demographics of the players. "team.csv" dataset and "buy-clicks.csv" files were joined together into one dataframe.

5. Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a statistical technique to analyse, summarize & extract key insights by using statistical graphs and different data visualization techniques. In this report, different statistical visualization techniques were employed to discover key insights that would reveal underlying trends and patterns to further improve the game.

Table 5. Dataset & summary of dataset

Dataset	Description
users.csv	This file holds the information about the users playing the game
ad-clicks.csv	This file holds information about the in-game ad clicks of the user.
buy-clicks.csv	When a user makes an in-app purchasea line is added in this file
team.csv	This file holds the information about all the teams in the game
team-assignments.csv	This file is updated each time a user starts or finishes a level
user-session.csv	This file holds the record every time a user starts or finishes a the game. Additionally, a new sessions starts each time a level is cleared
game-clicks.csv	A record is added into this file when user performs an in-game clic
combined-data.csv	This file holds the combined record of all user information like team, team level, game clicks buy id, average price

5.1. Age Group of Players

”users.csv” file was used to calculate the ”age” from ”date of birth” column. A new column named ”age-group” was generated by segmenting the users into 8 different subgroups. Table **??** shows the age group and number of players in each subgroup.

Fig 3 and table 6 highlight that the age of the large majority of players range between 30-39. Itś interesting to find out that the game is popular within the age range of 30 to 60 years. This implies that the game is beginner-friendly and has an easy learning curve for older audience to enjoy. The surprising insight from the analysis is, no player in the game has an age less than 20 years.

Table 6. Frequency of Players According to Age Group

Age Group	Age Count
20-29	378
30-39	592
40-49	536
50-59	436
60-69	304
70-79	147

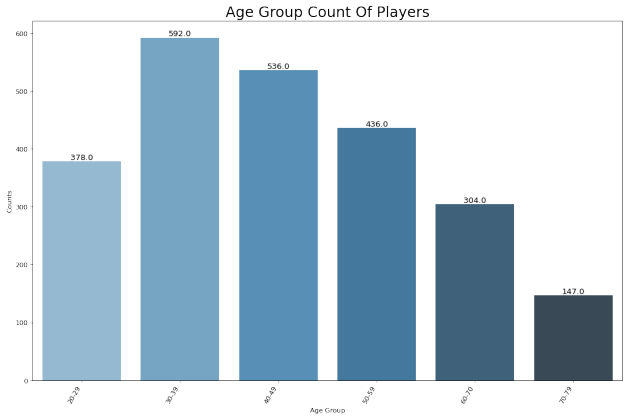

Figure 3. Age group frequency count of players

5.2. Top Spending Teams

To get the top spending teams, the ”teams.csv” was joined with ”buy-clicks.csv” to get the number of times a team purchased an item. Next, the team spending was aggregated using the ”teamId”. Fig 4 gives the top spending teams in the game.

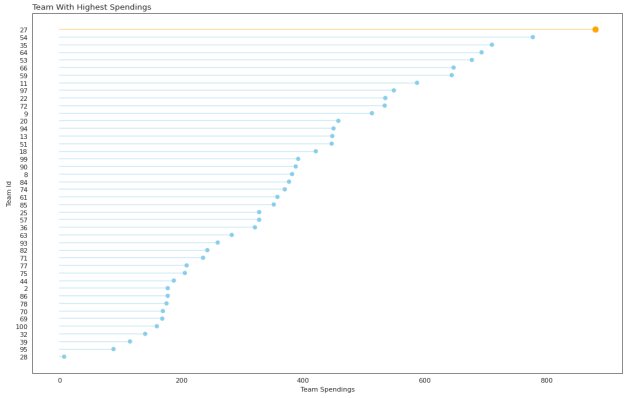

Figure 4. Top Spending Teams in Catch the Pink Flamingo

Fig 4 shows team no ”27” spent the highest on in-game purchases with little over ”800”, followed by team ”54” & ”35”.

5.3. Relation Between Team Spending & Team Strength

The relation between Team Spending & Team Strength was explored using scatter plot. ”teams.csv” & ”buy-click.csv” were joined and aggregated based on ”teamIds” & ”spending” to get ”total_spend”. The ”Strength” of teams and ”total_spend” were used to check the relation. A positive correlation value of 0.3967449 confirms the findings of Fig 5 suggesting somewhat weak correlation between spending and team strength.

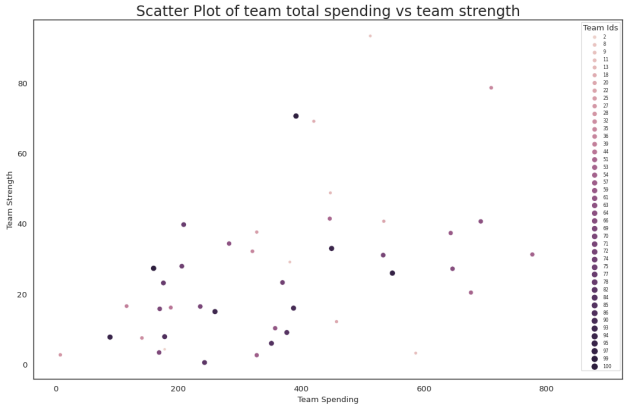

Figure 5. Relation b/w Team Strength & Team Spending

Fig 5 highlight a somewhat linear relation b/w team spending & team strength. However, not all teams who spend big are the top teams, like team ”2”.

5.4. Demographics With High Spending

The preprocessed ”user.csv” file with newly added ”continent” column, ”team.csv” & ”buy-clicks.csv”file were merged and aggregated to check which continent had the highest ratio of in-game spending. Fig 6, 7 show the continent with the highest spending percentage.

Fig 6, 7 illustrate ”Africa” & ”Asia” regions to be the top in-game spenders. It is worth noticing that ”Oceania” & ”South America” only account for 8.51% 8.54% of total spending respectively.

5.5. Highest Sold Product

”buy-click.csv” file holds the information for the in-game items that are sold and their selling price. ”price” column & ”buyId” that uniquely identifies a particular item being sold, was used to get the product that generated the most revenue. Fig 8 shows the top sold items in the game.

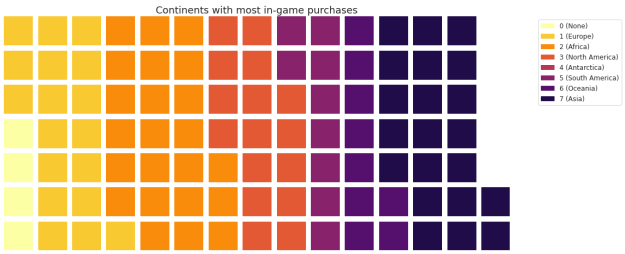

Figure 6. Continents with highest in-game spending

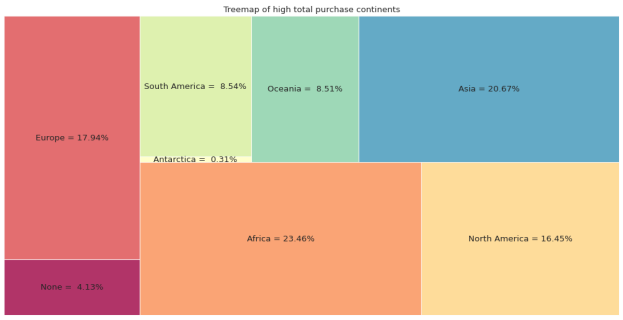

Figure 7. Percentage of Continents with highest in-game purchases

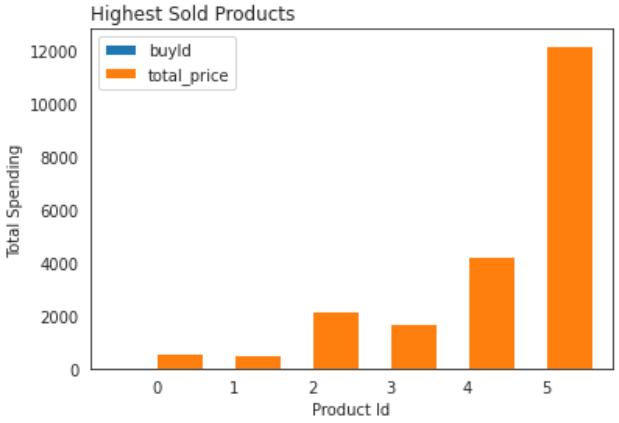

Figure 8. Items that generated the most revenue

According to the fig 8, Item with ID ”5” was the most popular among players with an accumulated revenue of, 12000 followed by item ”4”. Meanwhile, other items remain comparatively similar in popularity.

5.6. Popular Genre & Time Series of In-game Adclicks

”Ad-click.csv” file is updated each time a user clicks on an ad. This file was used to generate a time series of clicks over the entire history of the game since itś launch.

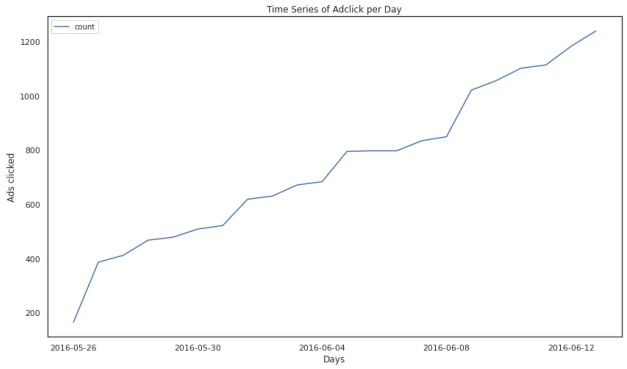

Figure 9. Time Series of Ad Clicks Since Game Laucnh

Fig 9 suggests an effective advertisement model with increase in daily ad clicks since launch date.

Similarly, the most popular ad genre among players was also investigated. Fig 10 suggests that ”Computers”, ”Gaming” & ”Clothing” were the genres that most players were interested in.

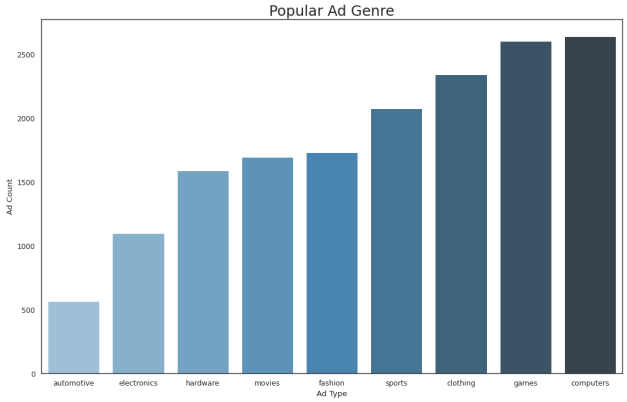

Figure 10. Popular Ad Genres

5.7. Popular Device Platform

file ”user-session.csv” holds the record of ”platformType” a player used for a particular in-game session. This file was used to aggregate the total number of devices in each category to check what are the most preferable devices people use for the game. According to the analysis shown in Fig 11, mobile devices are the most used platforms with ”iPhone” leading the race with ”41.88%” share followed by ”Android”. Meanwhile, PC platform ”Windows” is the most popular with ”13.41%” followed by ”Mac” & ”Linux”.

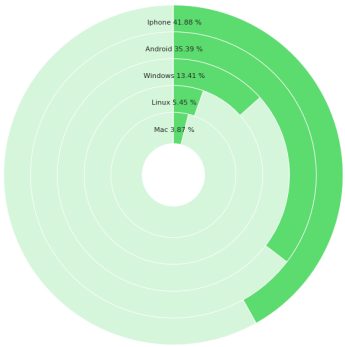

Figure 11. Most Popular Gaming Platform

5.8. Average Team Join

On average, how many times a player joins different teams, was calculated using the ”team-assigment.csv” that holds the record of each time a player joins a new team. The records in the file were aggregated using ”userId” and the mean value for each player was used to create a new column. This column was then used to create the box plot & violin plot. Fig 12 box plot shows that on average a player joins a team ”4” times. Violin plot gives an overall density distribution of 4 to 7 showing that majority of the players join 4 to 7 different teams on a whole.

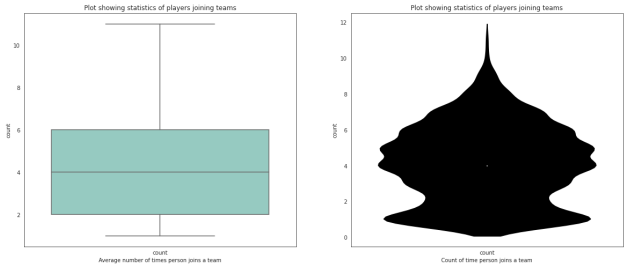

Figure 12. Number of times a player join different teams

5.9. Hit Ratio over different platforms

”user-session.csv” and ”game-clicks.csv” files were joined together to get the ”total clicks (Total Number of clicks)”, ”hit count (clicks that hit the target)” & ”Device Platform” columns. Fig 13 highlights the total number of hits vs hit counts and the overall percentage of hit counts of each platform.

According to the fig 13 all the platforms hit percentage range from 10% to 11%. This highlights that the game mechanics are well-designed for all platform types and encourages fair gameplay to all itś players.

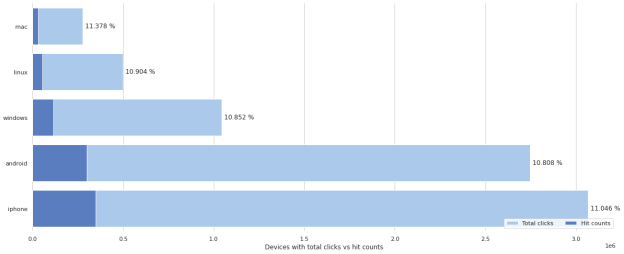

Figure 13. Device Hit Count Percentage

6. Machine Learning

Machine Learning algorithms were performed on ”combined-data.csv” file to get useful insights and segment the data into groups. The goal of applying Machine Learning techniques is to get insights that may help future business decisions in improving the in-game features. The goal is to target users based on their game playing abilities. For Machine Learning analysis, ”NULL” values of the ”average price” column that represent the userś average in-game spending were replaced with ”0”. Two new categorical columns were created to enable the analysis of players to be broken into ”Spender Non Spender” and ”Hitter & Non Hitter” categories.

For Spender & Non Spender category, the ”total average price” column was created by aggregating the sum of all the spending based on a player. The player is then placed into ”Spender”, ”Non Spender” class by checking if the player spending is more than 5.

Similarly, a player’s ”total game clicks” and ”total hit count” were aggregated. Then, the percentage of the of hit count is calculated by dividing the ”total game clicks” and ”total hit count”. The players are then placed into ”Hitter” class if hit count percentage is more than 10% otherwise ”Non Hitter” label is assigned.

6.1. Clustering Analysis

6.1.1. K-Means Clustering

K-Means clustering is an unsupervised machine learning algorithm. K-means algorithm clusters the data into similar groups based on Euclidean distance of feature-space from the centroid. The main goal of K-means is to maximize the inter-cluster similarity of the samples (Trevino, 2019).

K-means clustering scales well with large dataset & guarantees convergence of clusters to the feature space. The algorithm is versatile in a sense that it can be used in any type of grouping.

The crucial step in K-means clustering is to select the value of k. For the given dataset, Within Cluster Sum of square error (WSSE) was used to get the optimal value of k. Fig 14 shows the WSSE for different cluster sizes.

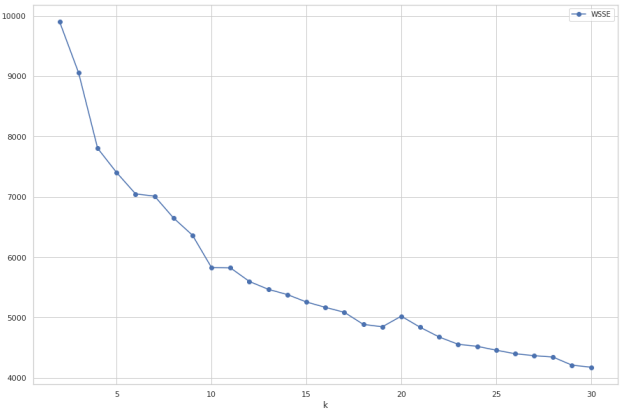

Figure 14. Elbow Method for WSSE

This method is also called as elbow method. The value of k is chosen based on the value of the x-axis, where the graph matches the shape of a human elbow. In this case, the elbow appears at cluster value of 5.

Groups of players that have different attributes like Spender, Non-Spender, hitter, Non-hiter & platformType may have different approaches to playing the game. For example, players using ”iPhone” devices may spend more on the in-game items.

For the given dataset, the following features were selected for cluster analysis ”platformType” ”teamLevel” ”hitter” ”spender” ”Average Price total” ”Average Hit Count” that segments the data into 5 distinct groups (k=5). Table 7 goes the features used in K-means clustering.

Table 6.1.1 shows the cluster centers K-means algorithm came up with. The cluster centre values are in the order ”team Level”, ”platform Type indexed”, ”avg price total”, ”count hits total”, ”spender” and ”hitter”. These clusters can be differentiated as follows:

Cluster 1 : includes the players that have the second last team level (Team level is -0.40354537 standard deviations below the mean). The cluster includes the highest spenders among all clusters (With average total spending of 0.56003647 & spender value of -0.09). The cluster includes low rank hitters (total hit count of -0.37215053 & hitter value of -0.18).

Cluster 2 : includes the players that have the lowest team level (Team level is -0.74546609 standard deviations below the mean). The cluster ranks fourth among spenders (With average total spending of 0.3891623 & spender value of 1.75 above the mean).The cluster includes the highest average hit

Table 7. Features Used For K-Means Clustering

Features	Rationale
platformType	PlatformType can be an important differentiator when it comes to game spending. For example, ”iPhone” & ”Mac” are expensive. So, the players using these platforms may not be averse to spending big on in-game items.
teamLevel	TeamLevel is also a discriminator. As, teams with high ranks may not require that much spending on items.
hitter	A player who is strong hitter might not be that much interested in item purchase.
spender	Spender property is used to investigate how spending fares with ”hitter” & ”Team Level”.
AveragetotalSpending	Average total spending per user is across all products is used.
Average hit Count	Average hit count per user is used to check individual hit count against the other properties.

Cluster	Properties Cluster Center
Cluster 1	[-0.40354537, -0.04157377, 0.56003647, -0.37215053, -0.09907657,-0.18002229]
Cluster 2	[-0.74546609, 0.29974088, 0.3891623, 2.60293323, 1.75300235, -0.02842629]
Cluster 3	[1.63562985, 0.03214219, 0.39206601, -0.32255074, -0.39555723,-0.15811977]
Cluster 4	[-0.03240632, 0.08333507, -1.78521144, -0.36291918, -0.45225236,-0.40737768]
Cluster 5	[0.10609881, -0.55201862, 0.44999551, -0.25554497, -0.06595115, 2.6295969]

count (total hit count of 2.60293 & hitter value of -0.028). This suggests their hit accuracy is related to in-game item purchase.

Cluster 3 : includes the players that have the highest team level (Team level is 1.63562985 standard deviations above the mean). The cluster ranks third among spenders (with average total spending of 0.392). Spender value of -0.39 suggest that not all players in this cluster are high spenders. Highlighting that some players are spending big in this cluster. The cluster includes the average hit count (total hit count of -0.3225). The cluster has a third-best hitter value of -0.15). All of this suggest that this cluster has mixture of players both high and low rank.

Cluster 4 : includes the players that have the third-best team level (Team level is -0.0324 standard deviations below the mean). The cluster ranks lowest among spenders (with mean total spending of -1.785 & spender value of -0.452 below the mean).The cluster includes the third-lowest average hit count (total hit count of -0.3629193 & hitter value of -0.407).

Cluster 5 : includes the players that have the second-best team level (Team level is 0.10609 standard deviations above the mean). The cluster ranks second highest among spenders with average total spending of 0.44 & spender value of -0.065 respectively. The cluster includes the best hitters among all clusters with hitter count of 2.629596 above the mean value.

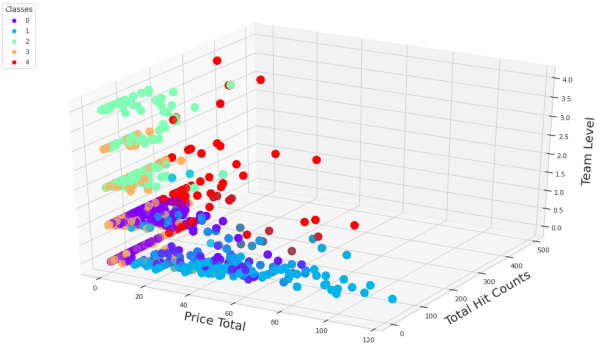

Figure 15. Cluster Analysis of Team Rank, Spending & Hits

Fig 15 visually draws the cluster comparison among these 5 clusters. Itś interesting to note that low rank teams represented by ”light blue & dark blue” cluster are high spenders. Whereas, teams that have high rank highlighted by ”green & red” have moderate hit count & spending.

This highlight that the game should advertise items more towards beginner level players.

6.2. Classification Analysis

Similar to clustering analysis, classification analysis was also performed on ”combined-data.csv” file. Binary classification models like Decision Tree & Logistic Regression

were used to predict the ”Spender & Non Spender” labels.

6.2.1. Decision Tree

In order to predict who is high spender and who is not based on the available attributes. A decision tree was used for the binary classification of the sample data into ”Spender” vs ”Non Spender”.

Decision tree is used with following features ”Total Hit Count”, ”Team Level”, ”PlatformType” to predict the aforementioned class labels of players. Itś important to predict which players are Spenders and which are not. This would help the game to design future game mechanics and in-game currencies based on that. Also, it would be helpful to design different marketing strategies based on that.

The Decision Tree model was fitted in PySpark with the features discussed before. The data was split into 80% training and 20% test data. The model obtained from the training set was used on test data, and the accuracy of the model was gauged using the confusion matrix given in table 6.2.1.

Table 8. Decision Tree Confusion Matrix of Spender

Spender	Prediction	Count
0 (True Negative)	0	810
1 (False Negative)	0	108
0 (False Positive)	1	11
1 (True Positive)	1	5

The accuracy of the model was calculated using $accuracy = \frac{(TN+TP)}{(TN+TP+FN+FP)}$. The accuracy of the decision tree classification was 87.25%. From the table 6.2.1, the model correctly predicted 810 Non Spenders and 5 Spenders. However, the model misclassified 108 Spenders as Non spenders. The model also mislabelled 11 Non Spenders as spenders.

A visual representation of the decision tree model is obtained by converting the PySpark model into JSON format. The model is shown in Fig 16.

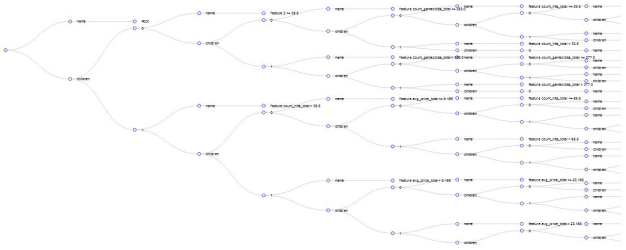

Figure 16. Decision Tree of Spender vs Non SPender

6.2.2. Logistic Regression

Logistic Regression is also a very popular binary classification technique. Unlike decision tree, that decides feature importance using GINI impurity for root node selection. Logistic regression probabilistically classify the samples based on the available feature space.

Logistic regression uses probability to classify the sample. If the probability of the sample under consideration falls above 0.5%. It classifies it as True, otherwise false.

Logistic regression model was employed for the classification of Players into ”Spender” vs ”Non Spender” classes. Same features as in decision tree model building ”Total Hit Count”, ”Team Level” & ”PlatformType” were used.

The model was fitted in PySpark with 80% training and 20% testing split. The model obtained from training data was used on the test set, and the accuracy of the model was gauged using the confusion matrix given below.

Table 9. Logistic Regression Confusion Matrix of Spender

Spender	Prediction	Count
0 (True Negative)	0	821
1 (False Negative)	0	112
1 (True Positive)	1	1

The accuracy of the model was calculated using $accuracy = \frac{(TN+TP)}{(TN+TP+FN+FP)}$. The accuracy of the Logistic Regression model was 88.01% that is slightly better than the decision tree model. However, the model incorrectly classified 112 Spenders into Non Spender label that is more than the decision tree misclassification in the same category. But the Logistic regression model showed improvement over decision tree when labelling Non Spenders with correct sample labelling of 821.

It turns out that Logistic regression is better at predicting Non Spenders, while decision tree is better at labelling Spenders. A suggested approach for ”Eglence Inc” would be to use an ensemble of both Decision Tree & Logistic Regression to improve the classification of players.

7. Graph Analysis

Graph Analysis is an important technique when identifying relations, communities, active users & interactions within communities. In order to increase the revenue for ”Catch The Pink Flamingo” game, Itś important to strategize resources and efforts on players and teams that are more active. The graph analysis was performed on the datasets described in table 10.The aim is to answer the questions about which teams and users are active. What are the longest conversations. This way, ”Eglence Inc” would be able to increase itś revenue by marketing itś products to the right audience

Table 10. Description of Data used in Neo4j

Dataset	Description
chat_join_ team_chat.csv	This file holds the userid, TeamChatSessionID, timestamp columns.A new line gets added when player "JOINS" a chat session.
chat_leave_ team$_chat.csv$	This file holds userid, TeamChatSessionID, timestamp columns. This file is updated when a player "LEAVES" a chat session.
_chat_mention_ team_chat.csv	This file holds ChatItem, userid, timestamp columns. A new line gets added when Player is "MENTIONED in the chat.
chat_respond_ team_chat.csv	This file holds the chatid1, chatid2, timestamp columns. When a player responds to itś "MENTIONED" chat. Both the "MENTIONED" chat & "Respondsto" chats get added into this file.

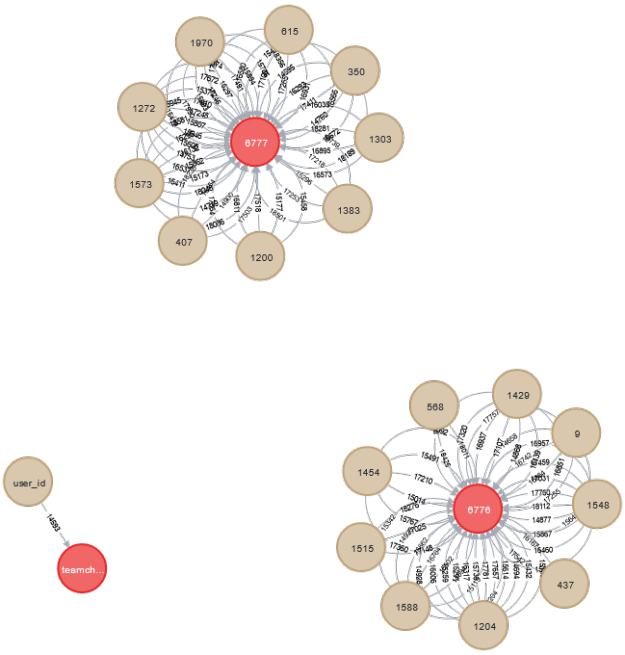

Figure 17. Join Relation between Player & ChatSession

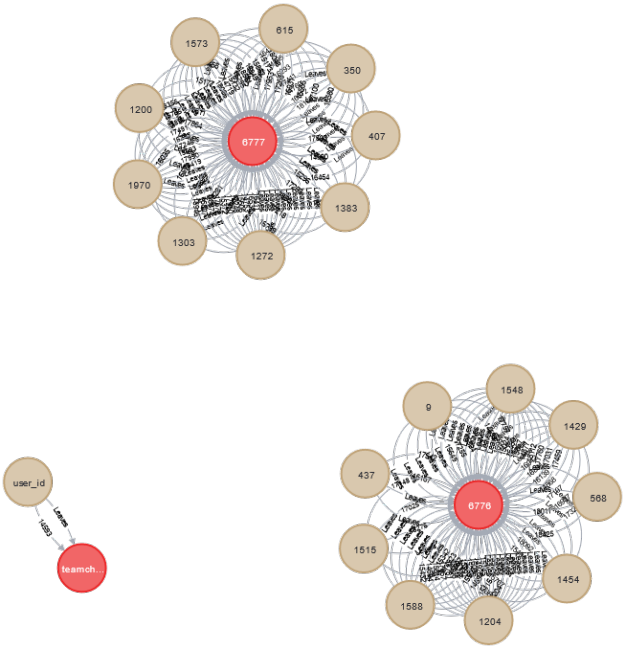

Figure 18. Graph of Players LEAVING ChatSession

who're participating actively.

7.1. Relations

First, following relationships were established to highlight the activity of the players.

7.1.1. JOIN RELATION

chat_join_team_chat.csv file was loaded into Neo4j. Two types of nodes were created using columns "userid" & "TeamChatSessionID" with an edge relation of "Joins" represented by timestamp column. Fig 17 shows the snapshot of the created graph.

7.1.2. LEAVES RELATION

"chat_leave_team_chat.csv" file is used to create "Leave" relationship. Two types of nodes were created using columns "userid" & "TeamChatSessionID" with an edge relation of "Leaves" represented by "timestamp" column. Fig 18 shows the snapshot of the created graph.

7.1.3. MENTIONED RELATION

"chat_mention_team_chat.csv" file is used to create "Mentioned" edge (relationship) between the player (UserID) & Chat (ChatItem) in which the user was mentioned. Fig 19

shows the snapshot of the created graph.

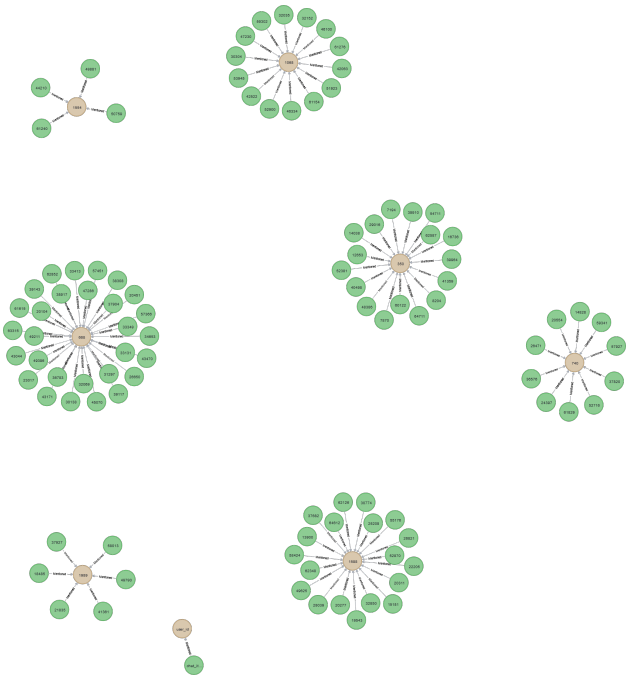

Figure 19. Graph of User Mention in Chat

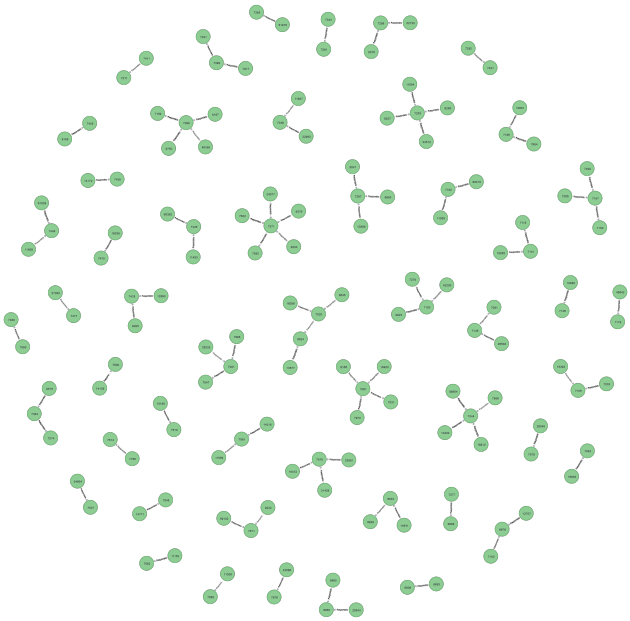

Figure 20. Graph of Chats that are replied to

7.1.4. Responded to relation

"chat_respond_team_chat.csv" file is used to create "Respondsto" edge (relationship) between the two chats (ChatItem). Fig 20 shows the snapshot of the created graph.

7.1.5. Active Teams

To better market the game, itś important to check who the active teams are. The teams that would have most "Join" activity are clearly going to be active. In order to check that, the teams that have the most "Joins" are considered. Fig 21 and table 11 represent the teams that are most active.

Table 11 & fig 21 represent the teams that have the most active chat session. Team "6792" is the most active with player joining frequency of 100. Similarly, Team "6783" is the second most active team with player joining frequency 91. Team "6925", "6850", "6791" are the next most active teams chat sessions joined by the players and so on.

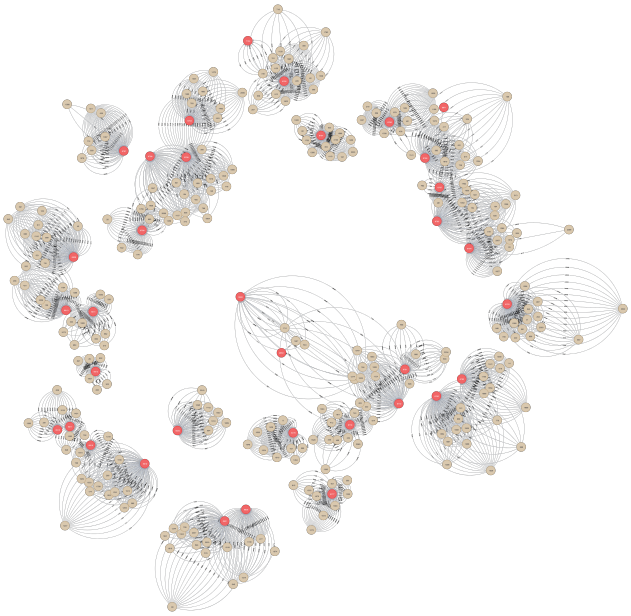

Figure 21. Most Active Team Chat Sessions

7.1.6. Active & Influential Users

Revenue of the game can be improved by marketing products directed towards users who are active and influential. In order to check the active & influential users. The mentioned relation discussed in section 7.1.3 is explored. The assumption is that users that are mentioned the most in the chat are likely to be more influential and active. Therefore, a list of most active users was obtained by running the query in Neo4j. The list of most mentioned players is given in table 12.

Table 12 highlights the players that are mentioned the most in the chat. Player "131" is mentioned the most, with "53"

Table 11. Player Join Count Of Team Chat Session

Team Id	Player Join Count
6792	100
6783	91
6925	87
6850	86
6791	81
6780	76
6809	72
6819	70
6790	67
6889	67

Table 12. List of most mentioned players

Player Id	Mention Count
131	53
1204	47
621	47
1428	46
1506	46
283	42
674	42
1482	42
1450	42
1127	41

times. Player ”1204” & ”621” are mentioned ”47” times. Followed by the players ”1428”, ”1506” and ”283” so on.

Itś a good idea for the game to market products directed at these players. As, these players are influential in the chat sessions. So itś likely that items marketed towards these players will likely bring huge interest from other players as well.

7.1.7. Longest Conversation

In order to better advertise items, itś important to know about how long a conversation lasts in the game. To check that, the length of the longest conversation chain is generated by exploring the relationship discussed in section 7.1.4. Fig 22 shows the list of chats that responded to each other.

Fig 22, shows the longest conversation chain of length 10.

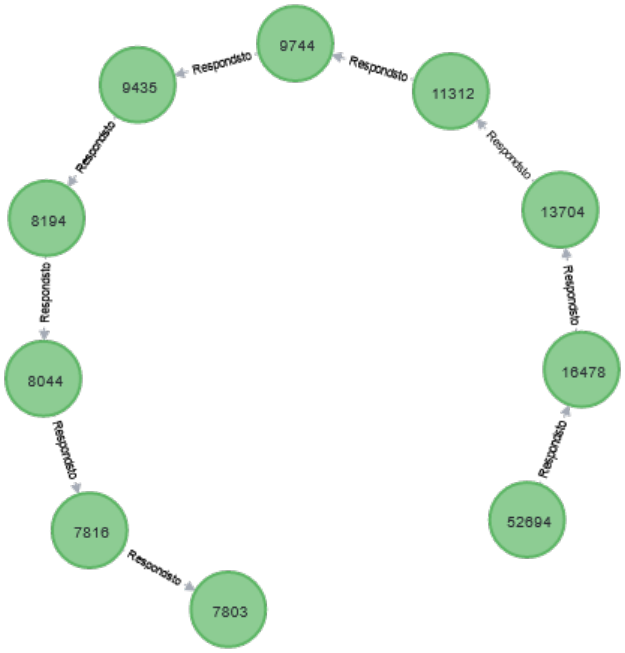

Figure 22. Longest Conversation Chain

This suggests that the longest linked conversation taken place between players had 10 messages.

This indicates that most players don’t spend that much time chatting with each other. So it would be good strategy to focus advertisement during the game sequences instead of the chat box.

8. Conclusion & Recommendations

This report analyse the emerging trends in Big data. It’s different processing paradigms, how they’re merging together to answer varied solutions to the difficult problems faced during storage, processing & fetching of data.

The report looks at the pressing ethical issues that arise from the storage & processing of big data. New laws that are likely to make a major shift in how Big data is handled.

The report then develops a Big data solution using Spark for the ”Catch the Pink Flamingo” game from the imaginary company named Eglence Inc. The solution performs EDA, Machine Learning & Graph analysis. The analysis has led to the following recommendations to improve the game.

8.1. Recommendations

This report concludes with the following list of recommendations to improve the game:

1. Majority of the users in the game have age 30 or more.

Also, there are no users age 19 or below.The game needs to add more gameplay mechanics and events that are tailored to younger audience.

2. Europe, North America and South America are the wealthiest region on the planet. However, most of the game revenue is coming from Asia & Africa. These regions need to be prioritized to further improve revenue.

3. Most of the item purchase revenue is coming from only two products. Eglence Inc needs to add more in-game items to introduce diversity. As currently, only 6 items are available to purchase.

4. An ensemble of both Logistic Regression & Decision is recommended to better classify users between Spenders & Non Spenders.

5. Low rank teams are likely to purchase more In-game items. Hence, more item advertisement should be directed towards low rank players.

6. Advertisements should be aimed at teams that are most active, like ”6792” & ”6783”.

7. Because the longest recorded chat has a length of 10. Instead of advertising inside the chat box, a better approach is to display ads before or after the gameplay ends.

8. The analysis suggests that lower rank teams and players are more prone to buying in-game items. So, the items should be developed and advertised by keeping this in mind.

References

Big data ethics. *Sage Journals*.

Abadi, D. J., Carney, D., Çetintemel, U., Cherniack, M., Convey, C., Lee, S., Stonebraker, M., Tatbul, N., and Zdonik, S. Aurora: A new model and architecture for data stream management. *The VLDB Journal*, 12(2): 120–139, aug 2003. ISSN 1066-8888. doi: 10.1007/s00778-003-0095-z. URL https://doi.org/10.1007/s00778-003-0095-z.

Casado, R. and Younas, M. Emerging trends and technologies in big data processing. *Concurrency and Computation: Practice and Experience*, 27:n/a–n/a, 09 2014. doi: 10.1002/cpe.3398.

Chandio, A., Tziritas, N., and Xu, C.-Z. Big-data processing techniques and their challenges in transport domain. *ZTE Communications*, 13:50–59, 02 2015. doi: 10.3969/j.issn.1673-5188.2015.01.007.

CHENG Hongbing, RONG Chunming, H. K. W. W. L. Y. Secure big data storage and sharing scheme for cloud tenants. *China Communications*, 12(6):106, 2015. URL http://www.cic-chinacommunications.cn/EN/abstract/article_127.shtml.

Davis, K. *Ethics of Big Data: Balancing Risk and Innovation*. O'Reilly Media, Inc., 2012. ISBN 1449311792.

EdPrice-MSFT. Real-time processing - azure architecture center. URL https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/real-time-processing.

Jain, P., Gyanchandani, M., and Khare, N. Big data privacy: a technological perspective and review. *Journal of Big Data*, 3, 11 2016. doi: 10.1186/s40537-016-0059-y.

Li, N., Li, T., and Venkatasubramanian, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pp. 106–115, 2007. doi: 10.1109/ICDE.2007.367856.

Liu, X., Iftikhar, N., and Xie, X. Survey of real-time processing systems for big data. In *Proceedings of the 18th International Database Engineering amp; Applications Symposium*, IDEAS '14, pp. 356–361, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450326278. doi: 10.1145/2628194.2628251. URL https://doi.org/10.1145/2628194.2628251.

Madden, S. From databases to big data. *IEEE Internet Computing*, 16(3):4–6, 2012. doi: 10.1109/MIC.2012.50.

Marz, N. and Warren, J. *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*. Manning Publications Co., USA, 1st edition, 2015. ISBN 1617290343.

Matt, K. Taking a regulation-agnostic approach to privacy, 2022. URL https://hyperproof.io/resource/regulation-agnostic-approach-privacy/.

Nisar, M. U., Fard, A., and Miller, J. A. Techniques for graph analytics on big data. In *2013 IEEE International Congress on Big Data*, pp. 255–262, 2013. doi: 10.1109/BigData.Congress.2013.78.

Oussous, A., Benjelloun, F.-Z., Ait Lahcen, A., and Belfkih, S. Big data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, 30(4):431–448, 2018. ISSN 1319-1578. doi: https://doi.org/10.1016/j.jksuci.2017.06.001. URL https://www.sciencedirect.com/science/article/pii/S1319157817300034.

Qiu, J., Wu, Q., Ding, G., Xu, Y., and Feng, S. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, 2016, 05 2016. doi: 10.1186/s13634-016-0355-x.

Sagiroglu, S. and Sinanc, D. Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pp. 42–47, 2013. doi: 10.1109/CTS.2013.6567202.

Shahrivari, S. Beyond batch processing: Towards real-time and streaming big data. *Computers*, 3(4):117–129, 2014. ISSN 2073-431X. URL https://www.mdpi.com/2073-431X/3/4/117.

Trevino, A. Introduction to k-means clustering, Dec 2019. URL https://blogs.oracle.com/ai-and-datascience/post/introduction-to-k-means-clustering.

Vuleta, B. How much data is created every day? [27 powerful stats], Feb 2022. URL https://seedscientific.com/how-much-data-is-created-every-day/.

White, T. *Hadoop: The Definitive Guide*. O'Reilly Media, Inc., 4th edition, 2015. ISBN 1491901632.

Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., and Vasilakos, A. V. Big data: From beginning to future. *International Journal of Information Management*, 36(6, Part B):1231–1247, 2016. ISSN 0268-4012. doi: https://doi.org/10.1016/j.ijinfomgt.2016.07.009. URL https://www.sciencedirect.com/science/article/pii/S0268401216304753.

A. Appendix

A.1. Code Available at Github

https://github.com/xahram/catch-the-pink-flamingo