Assessment 1: Big Data Management: Application of Data Analysis, Graph Analysis and Machine Learning Techniques to Big Data

Love Agbanimu, ID: 21120404 [*1]
Word Count: 4673

Abstract

This report give a detailed information on what big data is, down to its charateristics, uses and classification. We also explained three types of big data processing paradigms. We performed Exploratory data analysis and visualisation on some big datasets, graph analysis was also performed along with Machine learning techniques was applied on some dataset. We used Decision tree algorithm in classifying a dataset and KMeans algorithm in identifying clusters. Spark (using google colaboratory notebook) was used for our exploratory data analysis, classifcation and clustering while Neo4j was used for the graph analysis. Hadoop was used to store the data while GitHub was used to store notebook used. Finally, we concluded our report with research ethics and data protection.

1. Introduction

1.1. What is Big Data?

The beginnings of enormous data sets may be traced back to the 1960s and 1970snwith the invention of the relational database and the first data centres. Open-source frameworks like Hadoop (and, more recently, Spark) have been critical to the proliferation of big data because they make it easier to deal with and store. Since then, the number of big data has exploded, while machine learning has produced even more data, and cloud computing has further broadened big data possibilities(Sun et al., 2018a). Big data is larger, more complex datasets, especially from new data sources. These datasets are so voluminous that traditional data processing software just cannot manage them. But these massive volumes of data can be used to address business problems that would be difficult to tackle. Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

Big Data is a collection of extremely massive and complicated data sets that traditional database systems are unable to process in the timeframe required. For illustration, at Twitter, collecting and processing millions of daily tweets necessitates extensive data storage, processing, and data analytics capabilities, such as the ability to detect correlations between millions of tweets or analyse user demographics. Even though traditional SQL-based databases have proven to be highly efficient, reliable, and consistent in terms of storing and processing structured (or relational) data, they fall short of processing Big Data, which is characterised by large volumes, variety, velocity, openness, inappropriate structure, and visualisation, among other characteristics [1].

1.2. Characteristics of Big Data

This section presents ten characteristics of big data, which is broken down into three levels: fundamental level, technological level, and socio-economic level. The fundamental level has four fundamental characteristics of big data, the technological level consists of three technological characteristics and the socioeconomic level has three socioeconomic characteristics(Sun et al., 2018b). These characteristics are:

1. **Volume**: This refers to the size of the dataset. Hadoop are storing, processing, and analysing massive amounts of data in the terabytes and even millions of gigabytes range.

2. **Velocity**: This refers to the throughput (i.e the high rate of data flowing in and out of interconnected systems in real time) and latency (requirement of business) of data.

3. **Variety**: This refers to the diversity of data sources, as well as the types of data that are available to everyone. They are further divided into three categories: structured (data stored in relational database systems like Oracle), semi-structured unstructured (data on the Web) data types. Web, blogs and social media data

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

Modern Optimisation Report, CMP 7213 - Modern Optimisation.

are unstructured because they contain a big number of slang phrases and a mix of languages in a multi ethnic, multi-language context.

4. **Veracity**: Data is characterized based on its accuracy and truthfulness.
5. **Intelligence**: This is a collection of ideas, technologies, systems, and tools that can mimic and augment human intelligence in the context of big data management and processing.
6. **Analytics**: Data is characterized based on the type of analytics performed on it such as descriptive, predictive and prescriptive analytics.
7. **Infrastructure**: They are characterized based on structures, systems and facilities serving big data processing in a country, city or area.
8. **Service**: This refers to big data service to hundreds of millions of people such as infrastructure services, cloud services, mobile services, big analytics services and social network services.
9. **Value**: This defines the significant economic value and potential of big data to transform a company into a more competitive global platform player.
10. **Market**: This is characteristics based on big data technologies, systems, platforms, tools and services.

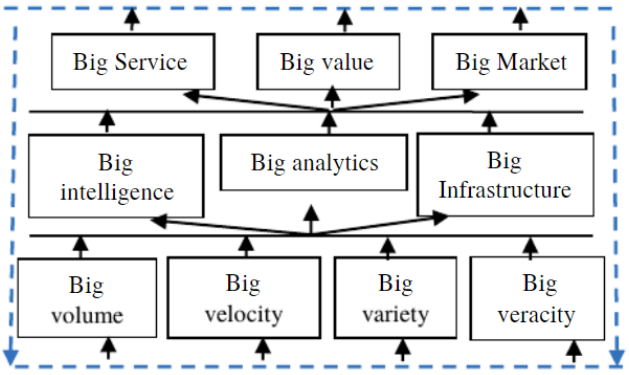

Figure 1. A unified framework of the 10 big data characteristics.

Although, all these are good characteristics, volume, velocity and variety are the most popular ones used in the big data world.

1.3. Uses of Big Data

Science, research, engineering, medicine, healthcare, finance, business, and, eventually, society itself are all benefits of Big Data(Barlow, 2013). It can be used to analyse and forecast business patterns, profit and loss, and identify real-time road traffic situations, healthcare, and weather data, among other things. Due to access to more information, big data allows permits one to get more full answers, and more complete answers mean more confidence in the data, which means an entirely different approach to problem-solving.

1.4. Classification of Big Data

In big data, it is vital to understand where the raw data comes from and how it needs to be processed before it can be analysed. Because there is so much of it, data extraction must be effective for the project to be worth working on. Structured data, unstructured data, and semi-structured data are the three classifications for big data(Casado & Younas, 2014).

2. Evaluation of Big Data Processing Paradigms

In this section, we will talk about the three processing paradigms which are real-time processing paradigm, batch processing paradigm and hybrid processing paradigm.

2.1. Real-Time Processing Paradigms

Real-time applications, typically at the second or millisecond level, use the streaming processing paradigm. Stream processing delivery patterns can be divided into three groups which are at-most-once, at-least-once and exactly-once. State management activities are another critical aspect of stream processing frameworks. In the construction of distributed and stream processing applications, scalable Message Oriented Middleware (MOM) is particularly significant. Although data must be streamed from Kafka to Hadoop, there are certain advantages to employing a Flume agent with Kafka producers to read data.

Real-time operating systems typically refer to the reactions to data. A system can be categorized as real-time if it can guarantee that the reaction will be within a tight real-world deadline, usually in a matter of seconds or milliseconds. An example of a real-time system are those used in the stock market. If a stock quote should come from the network within 10 milliseconds of being placed, this would be considered a real-time process. Whether this was achieved by using a software architecture that utilized stream processing or just processing in hardware is irrelevant; the guarantee of the tight deadline is what makes it real-time. Other situations where using real-time systems would be beneficial are:

- ATMs
- Air traffic control

- Anti-lock braking systems in your car

The real-time systems are extremely hard to implement using common software systems as these systems take control over the program execution, it brings an entirely new level of abstraction (i.e the distinction between the control-flow of your program and the source code is no longer apparent because the real-time system chooses which task to execute at that moment). This is beneficial, as it allows for higher productivity using higher abstraction and can make it easier to design complex systems, but it means less control overall, which can be difficult to debug and validate. Another common challenge with real-time operating systems is that the tasks are not isolated entities. The system decides which to schedule and sends out higher priority tasks before lower priority ones, thereby delaying their execution until all the higher priority tasks are completed.

More and more, some software systems are starting to go for a flavor of real-time processing (also known as soft real-time systems), where the deadline is not such an absolute as it is a probability. Generally, they are able to meet their deadline, although performance will begin to degrade if too many deadlines are missed.

2.2. Hybrid Processing Paradigm

Batch and real-time procedures are common in large data apps. This problem can be achieved with hybrid solutions. Hybrid computation in big data started with the introduction of Lambda Architecture (LA) which optimize costs by understanding parts of data having batch or real-time processing. Besides, the architecture allows to execute various calculation scripts on partitioned datasets.

Apache Storm is frequently used for the LA speed layer, and MongoDB is frequently used for the batch and serving layers. The development of two separate software and processing applications for the speed and batch layers resulted from the use of two different technologies. Both the speed and batch layers must use the same data processing technique however, the serving layer must be connected with both layers for data processing and data ingestion.

Stream processing is the process of being able to almost instantaneously analyze data that is streaming from one device to another. This method of continuous computation happens as data flows through the system with no compulsory time limitations on the output. With the almost instant flow, systems do not require large amounts of data to be stored. Stream processing is highly beneficial if the events you wish to track are happening frequently and close together in time. It is also best to utilize if the event needs to be detected right away and responded to quickly. Stream processing, then, is useful for tasks like fraud detection and cybersecurity. If transaction data is stream-processed, fraudulent transactions can be identified and stopped before they are even complete.

One of the biggest challenges that organizations face with stream processing is that the system's long-term data output rate must be just as fast, or faster, than the long-term data input rate otherwise the system will begin to have issues with storage and memory. Another challenge is trying to figure out the best way to cope with the huge amount of data that is being generated and moved. In order to keep the flow of data through the system operating at the highest optimal level, it is necessary for organizations to create a plan for how to reduce the number of copies, how to target compute kernels, and how to utilize the cache hierarchy in the best way possible.

2.3. Batch Processing Paradigm

Big data batch processing was started with Google File System which is a distributed file system and MapReduce programming framework for distributed computing. This is the initial paradigm Microsoft Dryad is another programming model for implementing parallel and distributed programs that can scale up capability. HPCC (High Performance Computing Cluster) Systems provide big data workflow management services.

In some systems, the changes are not made immediately but stored up and all performed in one go when the database is not in general use. This is called batch processing. This type of processing is used when it is not practical break the job into individual parts. Batch processing is ideal for: Recording attendance records in schools from OMR sheets Producing bills for electricity, gas and telephone companies Producing monthly bank and credit card statements In batch processing all the changes (Insertions, deletions, and amendments) are stored up in a transaction file. At a certain point the transaction file will be closed. To update the master file, the transaction and master files go through a merge process to create an updated master file.

Batch processing is the processing of a large volume of data all at once. The data easily consists of millions of records for a day and can be stored in a variety of ways (file, record, etc). The jobs are typically completed simultaneously in non-stop, sequential order. An example of a batch processing job is all of the transactions a financial firm might submit over the course of a week. Batching can also be used in:

- Payroll processes
- Line item invoices
- Supply chain and fulfillment

Batch data processing is an extremely efficient way to process large amounts of data that is collected over a period

of time. It also helps to reduce the operational costs that businesses might spend on labour as it doesn't require specialized data entry clerks to support its functioning. It can be used offline and gives managers complete control as to when to start the processing, whether it be overnight or at the end of a week or pay period.

As with anything, there are a few disadvantages to utilizing batch processing software. One of the biggest issues that businesses see is that debugging these systems can be tricky. If you don't have a dedicated IT team or professional, trying to fix the system when an error occurs could be detrimental, causing the need for an outside consultant to assist.

Another problem with batch processing is that companies usually implement it to save money, but the software and training requires a decent amount of expenses in the beginning. Managers will need to be trained to understand:

- How to schedule a batch
- What triggers them
- What certain notifications mean

These processes paradigms can be further explained in the figures below

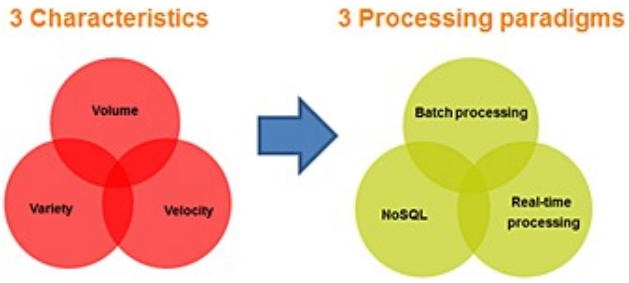

Figure 2. An inter-relationship between the characteristics of big data and processing paradigms.

3. Data Description

Data used is gotten from BCU Big Data Management CMP7203 Moodle page. Dataset comprises of other data. They are:

- chat-data
- combined-data; and
- flamingo-data

Dataset is described as thus:

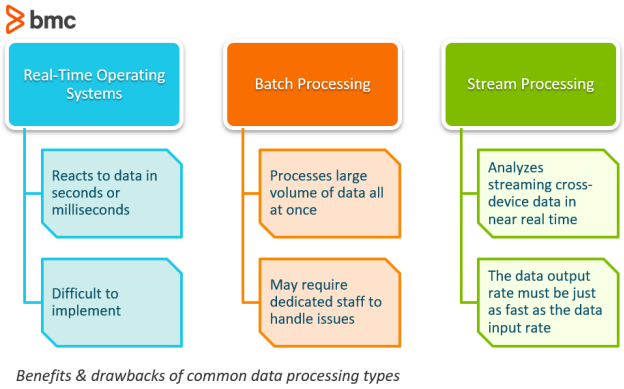

Figure 3. Benefits and drawbacks of common data processing types.

1. **chat-data:** This further comprises of four chat datasets. They are:
   - **chat join team chat:** It is a database of the join team chats. When a user joins a team, a new record is going to be added to this file which creates an edge labeled **Joins** from User to TeamChatSession. This has 4001 rows and 3 columns which comprises of the user id (displays the current user's ID), the teamchat session id (displays a uniqueID for the chat session) and the date (displays the time of the operation).
   - **chat leave team chat:** This is a database of the leave team chat. When a user leaves a team, a new record is going to be added to this file which creates an edge labeled **Leaves** from User to TeamChatSession. This has 3264 rows and 3 columns which comprises of the user id (displays the current user's ID), the teamchat session id (displays a uniqueID for the chat session) and the date (displays the time of the operation).
   - **chat mention team chat:** This is a database of the mention team chat. When a user gets a mention, a new record is going to be added to this file which creates an edge labeled **mentions** between the user node and chattext node. This has 11084 rows and 3 columns which comprises of chat item, userid (displays the current user's ID) and date (displays time of the operation).
   - **chat respond team chat:** This is a database of the respond team chat. When a player with chatid1 responds to a post by another player with chatid2, a new line is added in this file. This has 11074 rows and 3 columns which comprises of chat id1 (displays unique id for a chat), chat id2 (displays unique id for a chat) and date (displays time of

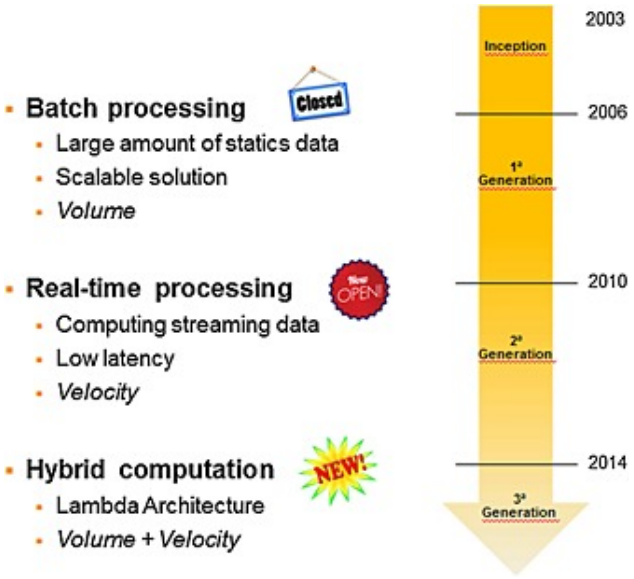

Figure 4. Trend description of the processing paradigms.

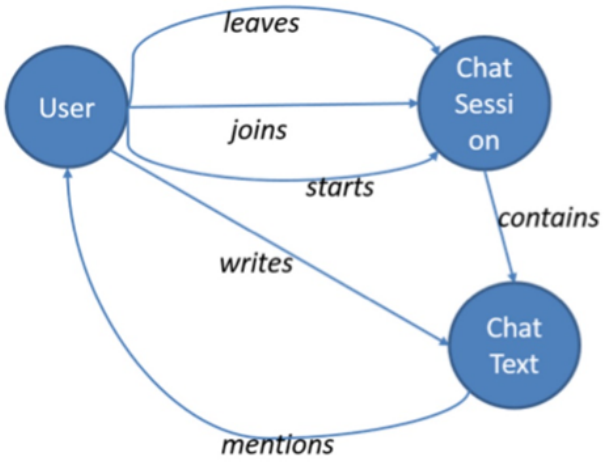

Figure 5. Illustration of and relationship between the chat datasets.

the chat).

2. **combined-data:** This is a game database which comprises of 4619 rows and 8 columns which includes userId, userSessionId, teamLevel, platformType, count gameclicks, count hits, count buyId and avg price.

3. **flamingo-data:** This is a game database which further comprises of eight flamingo dataset. They are:

   - **ad-clicks:** This is a database of clicks on ads. It has 16323 rows and 7 attributes. They are timestamp (displays when the click occurred), txid (displays a unique id within ad- clicks.log for the click), userSessionid (displays the id of the user session for the user who made the click), teamid (displays the current team id of the user who made the click), adId (displays the id of the ad clicked on) and adCategory (displays the category or type of ad clicked on).
   - **buy-clicks:** This is a database of purchases. It has 2947 rows and 7 attributes. They are timestamp (displays when the purchase was made), txId (displays a unique id within buy- clicks.log for the purchase), userSessionId (displays the id of the user session for the user who made the purchase), team (displays the current team id of the user who made the purchase), userId (displays the user id of the user who made the purchase), buyId (displays the id of the item purchased) and price (displays the price of the item purchased).
   - **game-clicks:** This is a record of each click a user performed during the game. It has 755806 rows and 7 attributes also. They are timestamp (displays when the click occurred), clickId (displays a unique id for the click), userId (displays the id of the user performing the click), userSessionId (displays the id of the session of the user when the click is performed), isHit (displays if the click was on a flamingo denoted by value 1 or missed the flamingo denoted by value 0), teamId (displays the id of the team of the user) and teamLevel (displays the current level of the team of the user).
   - **level-events:** This is a record of each level event for a team which are recorded when a team ends or begins a new level. it has 1254 rows and 5 attributes. They are timestamp (displays when the event occurred), eventide (displays a unique id for the event), teamId (displays the id of the team), teamLevel (displays the level started or completed) and eventType (displays the type of event, either start or end).
   - **team-assignments:** This is a record of each time a user joins a team. It has 9826 rows and 4 attributes. They are timestamp (displays when the user joined the team), team (displays the id of the team), userId (displays the id of the user) and assignmentId (displays a unique id for this assignment).
   - **Team:** This is a record of each team in the game. It has 109 rows and 6 attributes. They are teamId (displays the id of the team), name (displays the name of the team), teamCreationTime (displays

the timestamp when the team was created), teamEndTime (displays the timestamp when the last member left the team), strength (displays a measure of team strength, roughly corresponding to the success of a team) and currentLevel (displays the current level of the team).

- **user-session:** This is a record of each session a user plays. When a team levels up, each current user session ends and a new session begins with the new level. It has 9250 rows and 8 attributes. They are timestamp (displays a timestamp denoting when the event occurred), userSessionId (displays a unique id for the session), userId (displays the current user's ID), teamId (displays the current user's team), assignmentId (displays the team assignment id for the user to the team), sessionType (displays whether the event is the start or end of a session), teamLevel (displays the level of the team during this session) and platformType (displays the type of platform of the user during this session).
- **Users:** This is a database of the game users. It has 2393 rows and 6 attributes. They are timestamp (displays when user first played the game), userId (displays the user id assigned to the user), nick (displays the nickname chosen by the user), twitter (displays the twitter handle of the user), dob (displays the date of birth of the user) and country (displays the two-letter country code where the user lives).

3.1. Data Preprocessing

Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning. Data preprocessed is combined-data(Kotsiantis et al., 2006).

1. **Data quality assessment**
    - **Mismatched data types:** There is no mismatched data in all three data and sub data.
    - **Mixed data values:** There is no mixed data values in all three data and sub data.
    - **Data outliers:** There is no outliers in all three data and sub data.
    - **Missing data:** There is is no missing data for chat-data and flamingo-data but we did find missing data values in combined-data for attributes count buy and average price.
2. **Data cleaning:** We corrected and repaired missing data for combined-data where null values were dropped and replaced with the mean value of corresponding column (attribute).

userId	userSessionId	teamLevel	platformType	count_gameclicks	count_hits	count_buyId	avg_price	count_buyId_imputed	avg_price_imputed
812	5648	1	android	69	8	null	null	1	7.214323175053155
1658	5649	1	iphone	31	5	null	null	1	7.214323175053155
1589	5650	1	iphone	26	2	null	null	1	7.214323175053155
1863	5651	1	android	35	4	null	null	1	7.214323175053155
937	5652	1	android	39	0	1	1.0	1	1.0
342	5653	1	android	36	5	null	null	1	7.214323175053155
849	5654	1	iphone	40	5	null	null	1	7.214323175053155
1277	5655	1	windows	46	8	null	null	1	7.214323175053155
2203	5656	1	android	68	6	null	null	1	7.214323175053155
585	5657	1	iphone	76	9	null	null	1	7.214323175053155
1370	5658	1	iphone	69	6	null	null	1	7.214323175053155
1623	5659	1	iphone	129	9	1	10.0	1	10.0
881	5660	1	iphone	36	6	null	null	1	7.214323175053155
83	5661	1	android	102	14	1	5.0	1	5.0
453	5662	1	android	102	7	null	null	1	7.214323175053155
1966	5663	1	iphone	63	8	null	null	1	7.214323175053155
1073	5664	1	android	141	21	null	null	1	7.214323175053155
121	5665	1	android	39	4	1	3.0	1	3.0
462	5666	1	android	90	10	1	3.0	1	3.0
708	5667	1	iphone	32	2	null	null	1	7.214323175053155

Figure 6. Diagrammatic representation of top 20 clean dataframe.

The figure above shows the top 20 rows of our dataframe including the newly imputed cleaned values.

4. Exploratory Data Analysis

1. **ad-clicks data in flamingo-data** There is a strong positive correlation coefficient between txId and userSessionId of the ad clicks data in flamingo-data as correlation coefficient is 0.9848332193804368. Figure 7 shows the relationship between the two attributes and the data is groupedby adCategory which is visualised in figure 9 and 10.

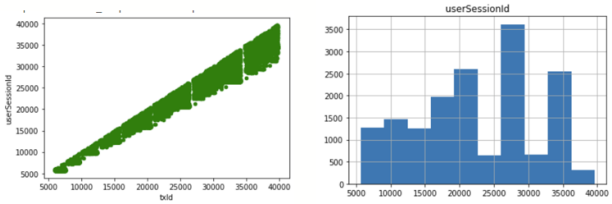

Figure 7. Relationship between txId and userSessionId in ad clicks data.

2. **buy-clicks data in flamingo-data** There is also a strong positive correlation between buyId and price (with correlation coefficient 0.9082396214188412) of the buy clicks data in flamingo-data. The scatter plot below shows their relationship. Figure 11 shows the relationship between the two attributes, figure 12 shows the distribution of price among users and the data is groupedby adCategory which is visualised in figure 14.

   buy-clicks data was also grouped using team. The figures below shows the top 10 team with the highest purchasing power and the per spending rate of the users.

adCategory	adId
automotive	16414
clothing	46397
computers	40653
electronics	2759
fashion	27403
games	40630
hardware	22549
movies	29487
sports	12906

adCategory	timestamp	txId	userSessionId	teamId	userId	adId
automotive	566	566	566	566	566	566
clothing	2340	2340	2340	2340	2340	2340
computers	2638	2638	2638	2638	2638	2638
electronics	1097	1097	1097	1097	1097	1097
fashion	1727	1727	1727	1727	1727	1727
games	2601	2601	2601	2601	2601	2601
hardware	1588	1588	1588	1588	1588	1588
movies	1692	1692	1692	1692	1692	1692
sports	2074	2074	2074	2074	2074	2074

Figure 8. ad-click data groupedby using adCategory.

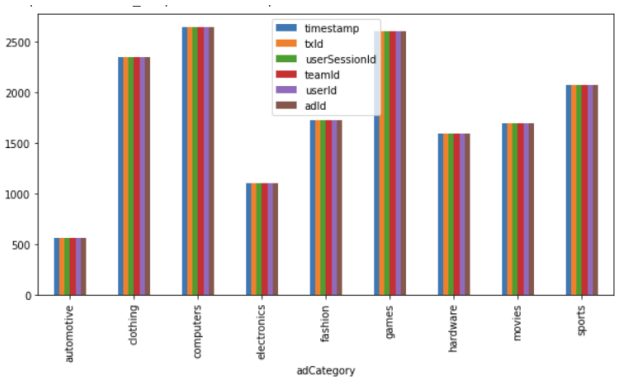

Figure 9. frequency distribution of attributes of ad-clicks data using adcategory.

The figure above shows the purchasing power of the top 10 team. Team 27 has the highest purchasing power of 880 with 103 more than the second highest purchasing team 54.

3. **game-clicks data in flamingo data** This data was groupedBy using teamLevel and we were able to visualise the top 10 team using our grouped data.

   This shows the top 10 most populated team level with teamlevel 6 as the hishest with a population of 122757 which is almost twice as populated as that of teamlevel 0.

4. **user-session data in flamingo data** This data was groupedBy using platformType which is also known as the devices used. Figure 15 shows the frequency of the 5 platformTypes used.

   This figure shows the distribution of device used in playing the flamingo game. Iphone is the most popular device used followed by android device with a difference of 600. Mac is not popular in this game since most game players prefer to play games on their phone since it is easy to carry around rather than the computer.

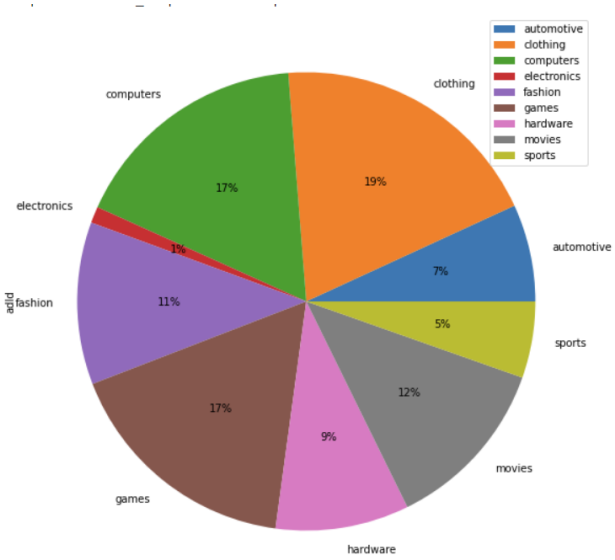

Figure 10. percentage of each adCategory.

5. Classification Results

Classification is defined as the process of recognition, understanding, and grouping of objects and ideas into preset categories. Classification algorithms used in machine learning utilize input training data for the purpose of predicting the likelihood or probability that the data that follows will fall into one of the predetermined categories.

Classification technique used was Decision Tree Classifier and dataset used is combined-data. A predictor column label (it predicts whether a user is a big or small player) was created using attributes teamLevel, count hits, count buyId imputed and avg price imputed to predict label.Data was split into train and test sets of 70:30 ratio, training ratio was confirmed to be 0.6958216064083135 (which is an approximation of 70 percent) was used on the test dataset Decisiontree classifier model is derived. The prediction model shows that there is an excellent accuracy score of 0.9074733096085409(which is also written as 90.74 percent).

DecisionTreeClassificationModel uid=DecisionTreeClassifier 21cf0670d8a9, depth=5, numNodes=39, numClasses=2, numFeatures=4

The Decision tree model feature importance is SparseVector(4, 0: 0.0037, 1: 0.9792, 2: 0.0071, 3: 0.0099)

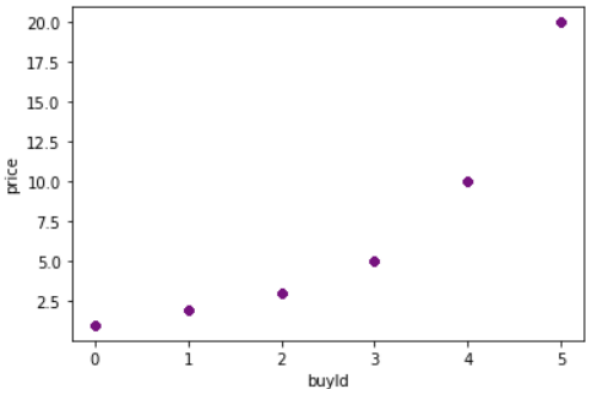

Figure 11. Relationship between buyId and price in buy-clicks data.

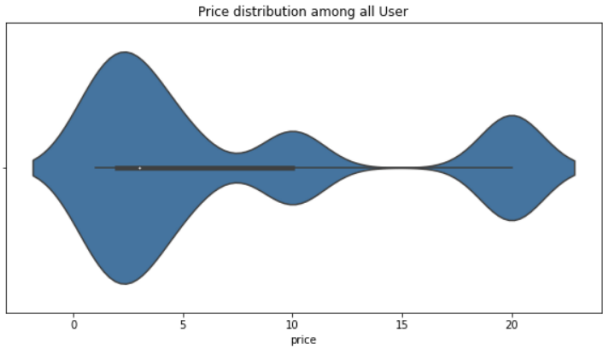

Figure 12. distribution of price among the users.

buyId	price
0	592.0
1	538.0
2	2142.0
3	1685.0
4	4250.0
5	12200.0

buyId	timestamp	txId	userSessionId	team	userId	price
0	592	592	592	592	592	592
1	269	269	269	269	269	269
2	714	714	714	714	714	714
3	337	337	337	337	337	337
4	425	425	425	425	425	425
5	610	610	610	610	610	610

Figure 13. buy-clicks data groupedby using buyId.

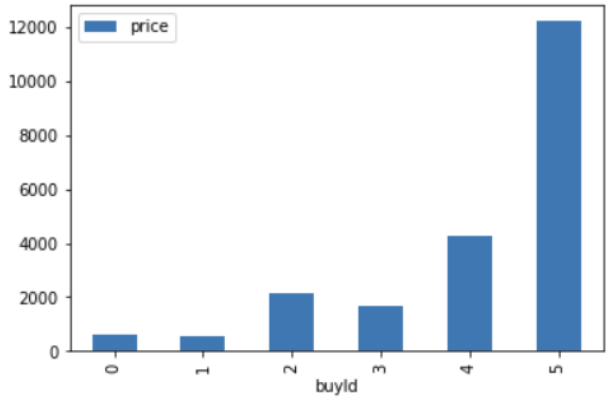

Figure 14. frequency distribution of price of buy-clicks data using buyId.

6. Clustering Results

Clustering algorithms are used to properly group data, so that it can be analyzed with like data. They're generally used in unsupervised learning, when not a lot is known about the relationships within your data.

Clustering method used is KMeans algorithm. Dataset used was combined-data and the clustering algorithm used was KMeans algorithm. Attributes selected for the clustering algorithm are count hits, count gameclicks and avg price imputed where we were able to assemble using features and derive the standardized values of the new dataframe. Two columns were created from the clean data to contain attributes needed for our clustering.

The figure above shows the dataframe to be used for clustering algorithm.

The figure above shows the values of k (used to determine the number of clusters) in descending order. Here we can see that the top features are three hence, the value for k with the with the largest score is 3.

Cluster Plot of Combined Dataset using KMeans Algorithm

7. —Graph Analysis

Data used was the chat-data and Neo4j was used to analyse the dataset.

1. **chat-respond-team chat data** It has 22146 nodes and 11073 relationship types. Property keys are join time, name, born, name, rating, released, roles, summary, tagline, timestamps and title. Nodes labels are ChatId1 and ChatId2, relationship type is responds.
2. **chat-mention-team chat data** It has 11801 nodes and 11084 relationship types. Property keys are join time, name, born, name, rating, released, roles, summary, tagline, timestamps and title. Nodes labels are Chats and Players, relationship type is MENTION.
3. **chat-leave-team chat data** It has 818 nodes and 3264 relationship types. Property keys are join time, name,

	team	sum_price
0	27	880.0
1	54	777.0
2	35	710.0
3	64	693.0
4	53	677.0
5	66	647.0
6	59	644.0
7	11	587.0
8	97	549.0
9	22	535.0

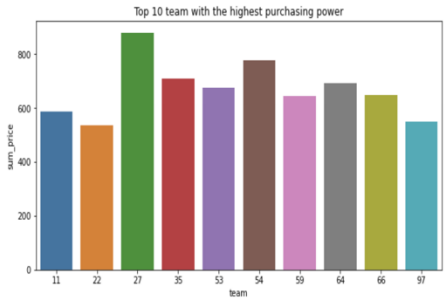

Figure 15. Top 10 team with the highest purchasing power for buy-clicks data.

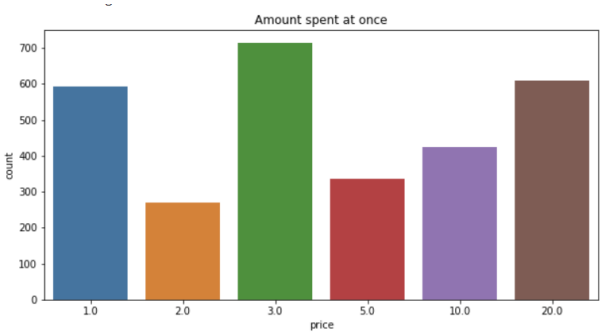

Figure 16. per spending rate of the users.

born, name, rating, released, roles, summary, tagline, timestamps and title. Nodes labels are User and teamChat, relationship type is LEAVE.

4. **chat-join-team data** It has 923 nodes and 4001 relationship types. Property keys are join time, name, born, name, rating, released, roles, summary, tagline, timestamps and title. Nodes labels are User and teamChat, relationship type is JOIN.

Node-relationship graph for all four data can be found on Github.

8. Ethics, Findings and Recommendations

8.1. Ethics

It is critical to think about research ethics and data protection from the start when planning for the management of data gathered from study participants, because how you handle information and consent processes can affect your capacity to share data afterwards. You also have an ethical and legal obligation to guarantee that confidential and personal data is stored and shared securely, and that it is not

	teamLevel	count
0	1	67271
1	6	122757
2	3	98823
3	5	117099
4	4	111176
5	8	51294
6	7	106436
7	2	80950

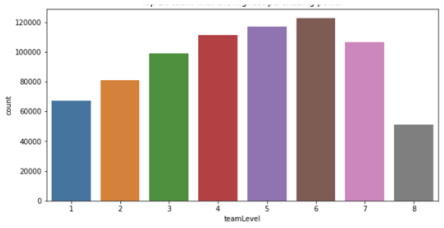

Figure 17. Top 10 team using teamLevel grouped by teamLevel.

	platformType	count
0	iphone	3874
1	android	3274
2	linux	504
3	mac	358
4	windows	1240

Figure 18. frequency of platformType used.

disclosed to unauthorized parties. Prior to sharing, proper consent procedures must be followed and identifying data is anonymized or minimized. In most circumstances, data collected from human participants can be made public or restricted (as needed). This ethics includes:

Personal data is any information relating to an identified or identifiable natural person. These data enjoy statutory protection under the General Data Protection Regulation 2016 and the Data Protection Act 2018. Under this legislation any personal data collected by you must be processed fairly and lawfully. Among other things you will be required to issue a privacy notice to your research participants, which explains the purpose(s) for which the data are being collected, your lawful basis for processing the data, who the data will be disclosed to, and the rights of the individuals in respect of their

```
+-----+----------+----------------------------------------+
|label|prediction|probability                             |
+-----+----------+----------------------------------------+
|1    |1.0       |[0.005454545454545455,0.9945454545454545]|
|0    |0.0       |[0.9925,0.0075]                         |
|0    |0.0       |[0.9137055837563451,0.08629441624365482]|
|0    |0.0       |[0.5764705882352941,0.4235294117647059] |
|0    |0.0       |[0.9925,0.0075]                         |
+-----+----------+----------------------------------------+
only showing top 5 rows
```

Figure 19. Top 5 predictions for test data usind DecisiontreeClassifier.

label	prediction	count
1	0.0	76
0	0.0	934
1	1.0	341
0	1.0	54

Figure 20. Confusion matrix for test data.

```
+---------------+---------+-----------------+------------------+------------------+
|count_gameclicks|count_hits|avg_price_imputed|          features|       standardized|
+---------------+---------+-----------------+------------------+------------------+
|             69|        8|7.214323175053155|[69.0,8.0,7.21432...|[0.54380637390174...|
|             31|        5|7.214323175053155|[31.0,5.0,7.21432...|[0.24431880566599...|
|             26|        2|7.214323175053155|[26.0,2.0,7.21432...|[0.20491254668761...|
|             35|        4|7.214323175053155|[35.0,4.0,7.21432...|[0.27584381284870...|
|             39|        0|              1.0|     [39.0,0.0,1.0]|[0.30736882003141...|
|             36|        5|7.214323175053155|[36.0,5.0,7.21432...|[0.28372506464438...|
|             40|        5|7.214323175053155|[40.0,5.0,7.21432...|[0.31525007182709...|
|             46|        8|7.214323175053155|[46.0,8.0,7.21432...|[0.36253758260116...|
|             68|        6|7.214323175053155|[68.0,6.0,7.21432...|[0.53592512210606...|
|             76|        9|7.214323175053155|[76.0,9.0,7.21432...|[0.59897513647148...|
|             69|        6|7.214323175053155|[69.0,6.0,7.21432...|[0.54380637390174...|
|            129|        9|             10.0|    [129.0,9.0,10.0]|[1.01668148164238...|
|             36|        6|7.214323175053155|[36.0,6.0,7.21432...|[0.28372506464438...|
|            102|       14|              5.0|    [102.0,14.0,5.0]|[0.80388768315909...|
|            102|        7|7.214323175053155|[102.0,7.0,7.2143...|[0.80388768315909...|
|             63|        8|7.214323175053155|[63.0,8.0,7.21432...|[0.49651886312767...|
|            141|       21|7.214323175053155|[141.0,21.0,7.214...|[1.11125650319051...|
|             39|        4|              3.0|     [39.0,4.0,3.0]|[0.30736882003141...|
|             90|       10|              3.0|     [90.0,10.0,3.0]|[0.70931266161096...|
|             32|        2|7.214323175053155|[32.0,2.0,7.21432...|[0.25220005746167...|
+---------------+---------+-----------------+------------------+------------------+
```

Figure 21. dataframe for clustering.

```
Silhouette Score: 0.7710754500882127
Silhouette Score: 0.7933619089102762
Silhouette Score: 0.7189021451892604
Silhouette Score: 0.5033523545832158
Silhouette Score: 0.4142152094525282
Silhouette Score: 0.6024525265056678
Silhouette Score: 0.6101847566755542
Silhouette Score: 0.6140983149901859
```

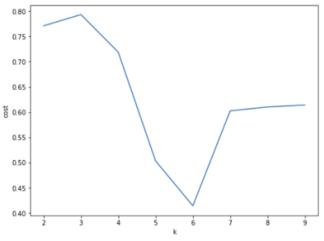

Figure 22. Silhouette values and visualisation of silhouette values to predict k using KMeans algorithm.

personal data. For certain kinds of research, for example involving the processing of sensitive data or human genetic data, you will need to complete a Data Protection Impact Assessment under the advice of the University Information Management and Policy Services Officer.

Data repositories such as the UK Data Service ReShare repository and the European Genome-phenome Archive, can manage controlled access to sensitive or confidential data. The University's Research Data Archive can also offer a restricted access option.

There is an obligation to protect the confidentiality of personal information provided by the research participants. Personal data should be destroyed when no longer required while clearly distinguishing between the personal data that will be held in confidence and ultimately destroyed, and the anonymised research data that will be retained indefinitely and made available to others.

Avoid making a specific commitment to destroy personal data by a set time as it is against the data protection law. Ensure that personal data are kept secure and are not disclosed to unauthorised persons. You should never promise to destroy research data gathered from participants in your application for ethical approval or in the material you provide to participants. You should also avoid sharing such data outside of the project because it may restrict you from sharing data in the future.

Always consider issues related to data protection and secure processing of information when you use instruments to collect data from research participants, including any online software services such as survey tools.

Working procedures should be designed to minimise the risk of inappropriate disclosure. When the study is complete and if there is no further need to link individuals to data, the linking key can be destroyed, so that the data become fully anonymised.

If data collected from human subjects have been fully anonymised, you do not need consent to share them, but it is good practice to inform your research participants how the data you collect from them will be used. Your information sheet should address this and your consent form should specifically allow the participant to indicate they have understood your intentions and agree to data sharing, by checking a statement such as

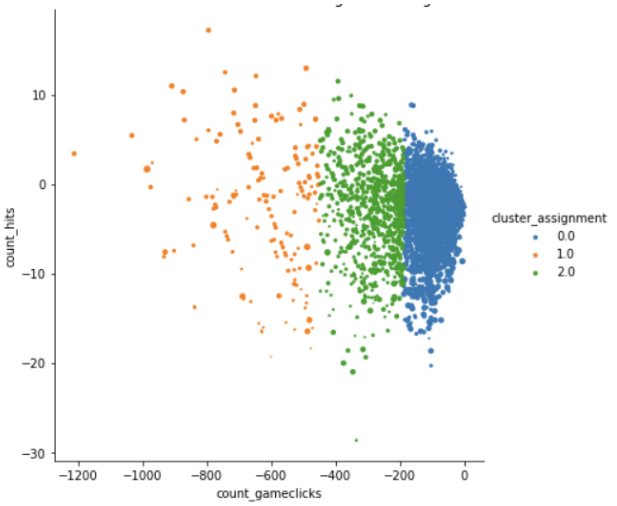

Figure 23. Cluster Plot of Combined Dataset using KMeans Algorithm.

> I understand that the data collected from me in this study will be preserved and made available in anonymised form, so that they can be consulted and re-used by others

.

8.2. Findings and Recommendations

From the report, we were able to understand what big data and various analysis that can be performed on it, visualise analysis and machine learning techniques made on our data and understand the ethics of data collection and usage.Spark due to its scalability and speed made it easy to use. Neo4j on the other hand, was used for graph analysis because the result was quicker to compute and more detailed but not all return nodes were being displayed due to Initial Node Display setting and previous query had to be deleted to give an appropriate result.

References

Alasadi, S. A. and Bhaya, W. S. Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, 12(16):4102–4107, 2017.

Alexandropoulos, S.-A. N., Kotsiantis, S. B., and Vrahatis, M. N. Data preprocessing in predictive data mining. *The Knowledge Engineering Review*, 34, 2019.

Barlow, M. *The culture of big data*. " O'Reilly Media, Inc.", 2013.

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.

Chaudhuri, S. and Dayal, U. An overview of data warehousing and olap technology. *SIGMOD Rec.*, 26:65–74, 1997.

Gray, J., Chambers, L., and Bounegru, L. *The data journalism handbook: How journalists can use data to improve the news*. " O'Reilly Media, Inc.", 2012.

Kotsiantis, S. B., Kanellopoulos, D., and Pintelas, P. E. Data preprocessing for supervised leaning. *International journal of computer science*, 1(2):111–117, 2006.

Meng, X., Bradley, J. K., Yavuz, B., Sparks, E. R., Venkataraman, S., Liu, D., Freeman, J., Tsai, D. B., Amde, M., Owen, S., Xin, D., Xin, R., Franklin, M. J., Zadeh, R. B., Zaharia, M. A., and Talwalkar, A. S. Mllib: Machine learning in apache spark. *J. Mach. Learn. Res.*, 17:34:1–34:7, 2016.

Quasim, M. T., Johri, P., Meraj, M., and Haider, S. 5v's of big data via cloud computing: uses and importance. *Sci. int (Lahore)*, 31(3):367–371, 2019.

Sedkaoui, S. *Data analytics and big data*. John Wiley & Sons, 2018.

Sun, Z., Strang, K., and Li, R. Big data with ten big characteristics. In *Proceedings of the 2nd International Conference on Big Data Research*, ICBDR 2018, pp. 56–61, New York, NY, USA, 2018a. Association for Computing Machinery. ISBN 9781450364768. doi: 10.1145/3291801.3291822. URL https://doi.org/10.1145/3291801.3291822.

Sun, Z., Strang, K., and Li, R. Big data with ten big characteristics. In *Proceedings of the 2nd International Conference on Big Data Research*, pp. 56–61, 2018b.

Tanasa, D. and Trousse, B. Advanced data preprocessing for intersites web usage mining. *IEEE Intelligent Systems*, 19(2):59–65, 2004.