Big Data Management
Final Report

Muhammad Ammar Shahid, ID: 23103156 [*1]
Word Count: 2968

1. Introduction

Today in the age of big data, where data is increasing exponentially regardless of field. Most of the data is generated by social media sites (Sharma, 2015). Organizations already use this data to get insights and make data-informed decisions to increase their profitability and innovation (Acciarini et al., 2023). According to the (Loi & Dehaye, 2017) world's most valuable resource, oil is replaced by data. The authors further stated that in 2017 the revenue generated by data is more than 55 billion dollars and is expected to increase by 103 billion dollars by 2027.

As big data is highly complex due to its structure (unstructured/semi-structured), organizations face many major challenges with analyzing, visualizing, and indexing the data (Naeem et al., 2022). However, (Munawar et al., 2022) discussed in their study, the development of various tools and techniques in the last two decades and suggest that machine learning (ML) techniques when coupled with big data can be very useful for the clustering and classification of the data.

This report will discuss the properties of big data and three processing paradigms in the section. Furthermore, data-set from the game (Catching The Flamingos) will be used for analysis and visualization purposes, clustering, and classification techniques will also be implemented on the same data for predictive analysis. In the section connections and communities within the user will be analyzed using graphs. The last section will elaborate on the ethics of big data.

2. Introduction to Big Data

Big data refers to massive and complex datasets that are produced by different organizations, individuals, and machines (Dumbill, 2012; Manyika et al., 2011; McAfee et al., 2012). These datasets are often too large and unstructured to be processed by traditional data processing applications. Therefore, new methods and technologies have emerged for managing and analyzing these data sets (De Mauro et al., 2016; Gantz & Reinsel, 2012). The potential to extract valuable insights from big data has become crucial for businesses and governments to make informed decisions, improve efficiency, and promote innovation (Manyika et al., 2011). However, on the other side privacy, security, and ethical concerns have emerged from the use of big data (Kshetri, 2014; Mayer-Schönberger & Cukier, 2013).

2.1. Big Data Processing paradigms

Big data is characterized by four Vs: volume, velocity, variety, and veracity. The volume of data refers to the large dataset size, which can range from terabytes to petabytes (Fan & Bifet, 2013). Velocity refers to the speed at which data is generated and needs to be processed, whether in real-time or through batch processing (Davenport & Dyché, 2013). Variety refers to the different types and formats of data, which can include structured, semi-structured, and unstructured data (Laney et al., 2001). Veracity refers to the accuracy and reliability of the data (Zikopoulos & Eaton, 2011). Processing paradigms, such as Hadoop and MapReduce, have been developed to manage and analyze big data, allowing organizations to gain valuable insights from these massive datasets (Kaisler et al., 2013).

2.1.1. Batch processing

Batch processing involves processing large volumes of data at regular intervals. This approach is suitable for processing large data sets that do not require immediate processing. Hadoop is a popular open-source framework used for batch processing of large data sets. According to (White, 2015), Hadoop provides a distributed file system (HDFS) that can store and manage large amounts of data. Hadoop uses a programming model called MapReduce to process data in parallel across a cluster of machines. (Lin & Dyer, 2010) provides a comprehensive guide to batch processing of text data using MapReduce, working is explained in Fig 1. Additionally, (Chambers & Zaharia, 2018) discuss Spark, a popular batch-processing engine that can work with Hadoop. Spark provides a more flexible and efficient alternative to

[1]M.Sc. Big Data Analytics, School of Computing and Digital Technology, Birmingham City University, UK. Correspondence to: <Muhammad.shahid3@mail.bcu.ac.uk>.

Modern Optimisation Report, CMP 7213 - Modern Optimisation.

MapReduce by performing in-memory operations, and it has gained popularity in recent years.

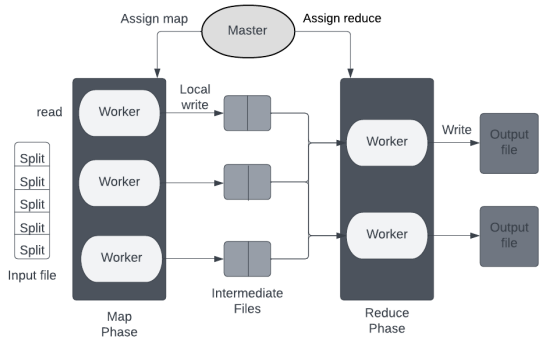

Figure 1. MapReduce working explained (modified from (Shahrivari, 2014))

Batch processing is commonly used in data warehousing and ETL (Extract, Transform, Load) processes. For example, a retail company may use batch processing to analyze customer data collected over a month to identify customer behavior patterns, popular products, and areas for improvement. Another example is a healthcare organization that may use batch processing to process large volumes of patient data to identify trends and make informed decisions about patient care.

2.1.2. Real Time Stream Processing

This approach is suitable for processing data that requires immediate processing, such as sensor data or financial transactions. Real-time processing engines can process data in real-time or near-real-time (Stream processing) and can provide immediate feedback or alerts. Apache Flink is a popular real-time processing engine that provides high-throughput and low-latency processing. (Vitorino et al., 2023) provides an introduction to Apache Flink, which is an open-source, distributed stream processing framework designed for high-performance, fault-tolerant, and real-time stream processing see Fig 2. It supports various stream processing use cases such as event-driven applications, real-time analytics, and machine learning. Kafka is another popular real-time processing platform used for building distributed streaming applications. (Narkhede et al., 2017) discuss Kafka, which is a distributed messaging system designed to handle high-throughput data streams. It is used for real-time streaming data processing, data integration, and messaging applications.

Real-time processing is commonly used in industries that require immediate action based on data, such as finance and transportation. For example, a stock trading company

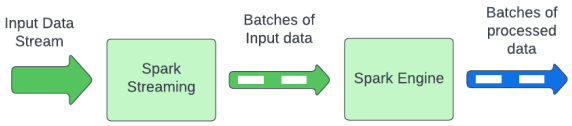

Figure 2. Spark working explained (modified from (Wingerath et al., 2016))

may use real-time processing to analyze stock prices and execute trades in real time. Similarly, an airline company may use real-time processing to analyze flight data and make decisions about flight delays or cancellations.

2.1.3. Hybrid Processing

Hybrid processing involves using a combination of batch and real-time processing to handle data. This approach is suitable for processing data that requires both immediate processing and historical analysis. Hybrid processing systems can provide real-time analytics and also store data for future analysis. The Lambda Architecture is a popular approach for building hybrid processing systems. According to (Warren & Marz, 2015), the Lambda Architecture consists of three layers: batch layer, speed layer, and serving layer. The batch layer stores and processes historical data, the speed layer processes real-time data, and the serving layer provides a unified view of the data see Fig 3. (Lee et al., 2022) discusses building hybrid cloud data pipelines and also provides guidance on building hybrid applications in the cloud.

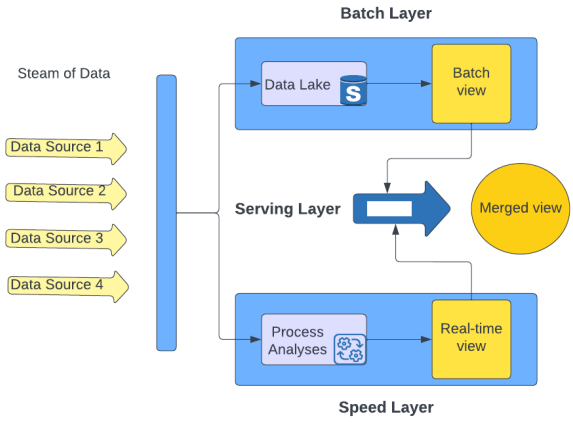

Figure 3. Lambda Architecture explained (modified from (Wingerath et al., 2016))

Table 1. comparison between processing paradigms

Paradigm	Batch Processing	Real-Time Processing	Hybrid Processing
Data processing	Process large volumes of data	Process data as it arrives	Process real-time and historical data
Latency	High latency	Low latency	Low latency
Use cases	Data warehousing, ETL, analytics	Finance, transportation	Social media, e-commerce, gaming
Tools and platforms	Hadoop, Spark with hadoop, MapReduce	Apache Flink, Kafka	Lambda Architecture, Kubernetes

Hybrid processing is commonly used in social media, e-commerce, and gaming industries, where there is a need for both real-time and batch processing. For example, a social media platform may use hybrid processing to process user data in real-time for immediate feedback and also store data for batch processing for insights into user behavior. Another example is an e-commerce company that may use hybrid processing to process real-time transactions and also analyze historical data to identify trends and provide personalized recommendations to customers.

2.1.4. Comparison

This report has provided an overview of three big data processing paradigms: batch processing, real-time processing, and hybrid processing. Each of these paradigms has its advantages and disadvantages and can be used to handle different types of data processing tasks. Batch processing is suitable for processing large data sets that do not require immediate processing, while real-time processing is suitable for processing data that require immediate processing. Hybrid processing provides a combination of batch and real-time processing and is suitable for processing data that requires both immediate processing and historical analysis. Choosing the right processing paradigm depends on the specific requirements of the data processing task. A detailed comparison is done in table 1. However, this report will use Apache Spark with Hadoop because flamingo data is not real-time.

3. Exploratory Data Analysis

3.1. Flamingo Data Overview

Catching the pink flamingo is a game, which is about catching as many flamingos as you can and gradually levels get tougher. Players can interact with each other, can make purchases, etc. The dataset includes information about players, items they purchased, teams, chats, and ad clicks. However, this report used subsets of the original dataset to perform different analyses.

3.2. Data Preprocessing and transformation

Mainly this report used combined data which includes, userid, sessionid, teamLevel, platformType, count gameclicks, count hits, count buyId, avg price. There were a total of 4620 records with a lot of null values. As shown in Fig 5, avg price and count buyId has a very large number of null values. To overcome this 2 other data files were merged including user data and user buy data, this made data more consistent users with no purchase history were dropped off as shown in Fig 6. Furthermore, several unnecessary columns were dropped, time and date format was cast correctly, the date of birth column was converted into age, and the difference between the user's first play date and purchase date was calculated to get useful insights. Full workflow can be seen in Fig 4.

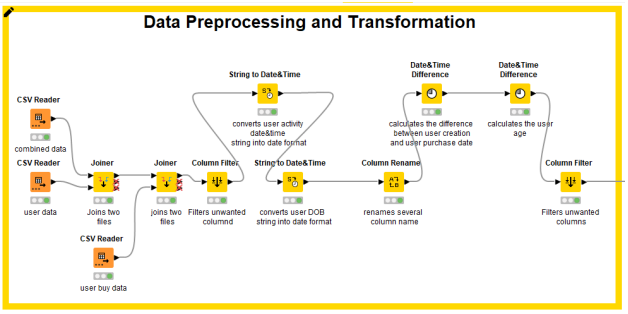

Figure 4. Workflow in Knime

Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
platformType	☐	0	5	iphone, android, windows, linux, mac	
count_buyId	☐	0	7	NULL, 1, 2, 3, 4, 5, 6	
avg_price	☐	0	65	NULL, 3.00, 1.00, 20.00, 10.00, [...], 2.60, 14.00, 5.17, 8.00, 14.40	

Figure 5. Data includes null values before preprocessing

Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
platformType	☐	0	5	iphone, android, windows, linux, mac	
Number_of_purchase(by a user)	☐	0	6	2, 1, 3, 4, 5, 6	
avg_Purchase_price	☐	0	64	3.00, 1.00, 20.00, 2.00, 10.00, [...], 4.33, 8.33, 6.67, 14.00, 8.00	

Figure 6. clean data

3.3. EDA Visualization

This section will discuss the different types of visualizations that elaborate interesting facts about the data.

3.3.1. Donut chart

Fig 7 elaborates that most of the players use iPhone devices. Out of the 2375 players, 935 uses iPhone and 849 uses Android phones, which clearly shows that the majority of players prefer to play game on their mobile phone.

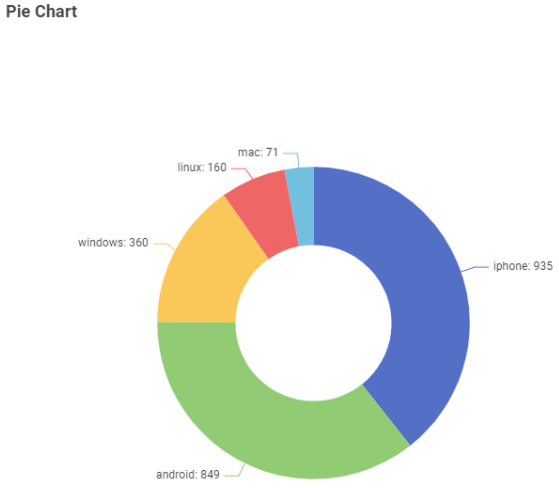

Figure 7. Platform comparison

3.3.2. 3D Scatter plot

Players up to the age of 45-46 made the most purchases, with average spending of 2, and most of these players purchased within the first 10 months fig 8 illustrates that.

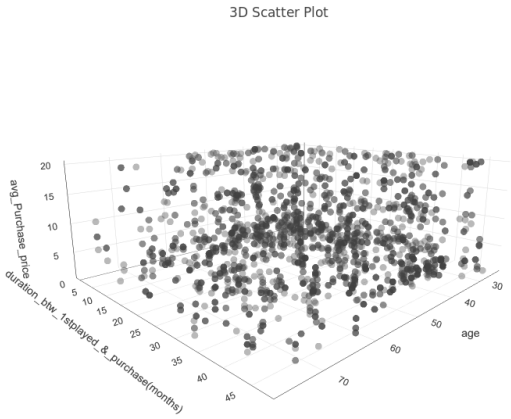

Figure 8. Comparison between age, first purchase, and average purchase price

3.3.3. Heat-map

As discussed in 3.3.1 most player uses iPhone, fig 9 shows that most players that use iPhone are on level 6 team and 231 players are on level 7 team, similarly most of the Android users are in level 7 team.

3.3.4. Box plot

Fig 10 presents detail about the hit counts of the players, age, and period of first purchase. The average (mean) age of the players is 45. Similarly, the average player purchases within the first 26 months and catches 13 flamingos on average. There are some outliers in the hits that show some players went above the majority.

3.3.5. Stacked area chart

Fig 11 shows that players at team level 2 have a higher number of hits and clicks, which means the level 2 team catches the higher number of flamingos.

3.3.6. 2D density plot

Players of age 63 (1853 time) and 43 (2024 time) clicked more ads as compared to others, in addition to that ads about computers and games were more clicked see fig 12

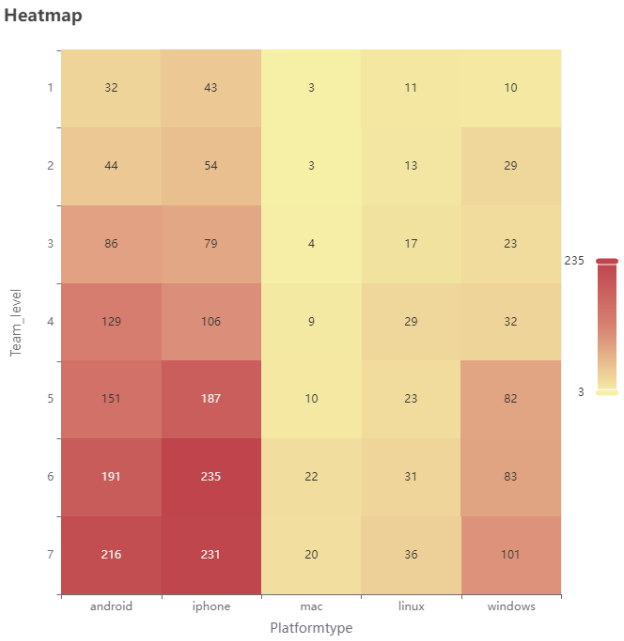

Figure 9. Platform vs team level

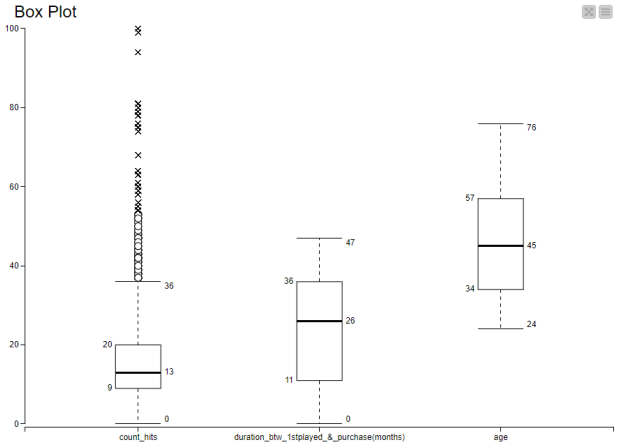

Figure 10. box plot

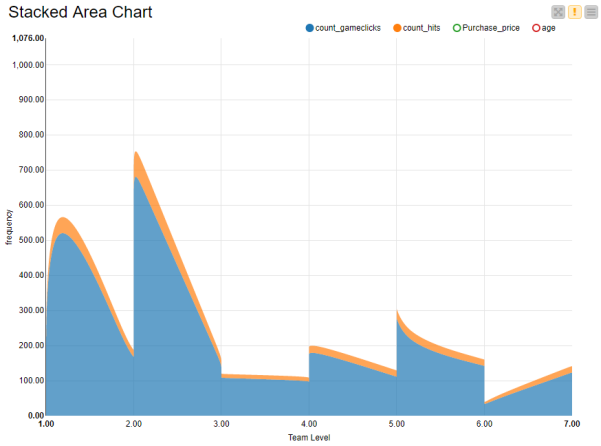

Figure 11. stacked area chart for game hits and game clicks with respect to team level

4. Machine Learning Models

After analyzing the data, the next step was to get the most out of data to predict the future. So for that purpose classification and clustering algorithms were applied to the data. Fig 13 shows the data columns that were used in the classification. Platform type was the target variable.

4.1. Classification

Two classification models were trained using Apache Spark with Hadoop. Both of them are discussed below.

4.1.1. DECISION TREE

Decision tree (DT) is capable of classifying multiple classes with high accuracy, so a DT model was trained to get the predictions about what type of platform a player might be using based on the independent variables (shown in fig 13). To train the model first categorical columns including platform type and country were indexed. After that, all the features were assembled into one column and label (target variable) in the 2nd column. The data set split ratio for training and testing was 80 and 20 respectively. The model achieved an accuracy of 92.09, and precision and recall were 88 and 96 percent respectively.

4.1.2. MULTI LAYER PERCEPTRON

It is a deep neural model. To train this model different combinations (hyperparameters) were tried; for instance, the number of layers, number of perceptrons, optimizer, number of epochs, block size, and learning rate. The best model achieved an accuracy of 78.88 percent. Four layers were used i-e 1 input, 1 output, and 2 hidden layers with 10,10,15, and 5 perceptrons respectively. Gradient descent was used as an optimizer with a learning rate of 0.1. The model was trained for 500 epochs. Code is provided in the 7

4.2. Clustering

The same data that was used for the classification was used for clustering as well but for clustering, the age column was

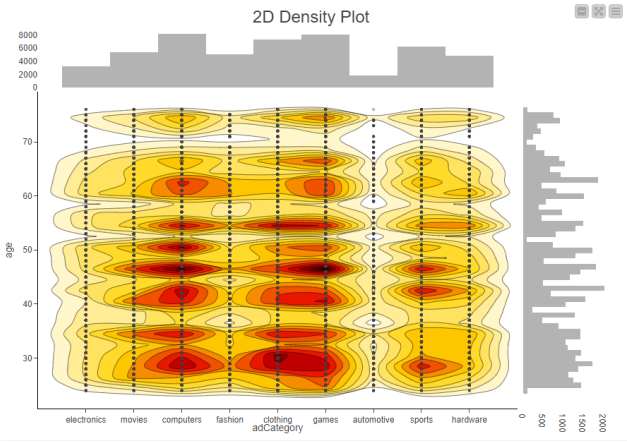

Figure 12. comparison of ad category with age

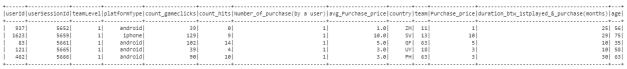

Figure 13. data head for classification and clustering

chosen. Clustering was performed in Knime, the number of clusters was 4. In addition to that hit counts were analyzed with these 4 clusters as shown in fig 14 on the x-axis age is mentioned. It is clear from the figure that in the first 2 clusters i-e up to age 50 performance (hits) is good. In comparison within the first cluster, the first cluster has slightly more hits.

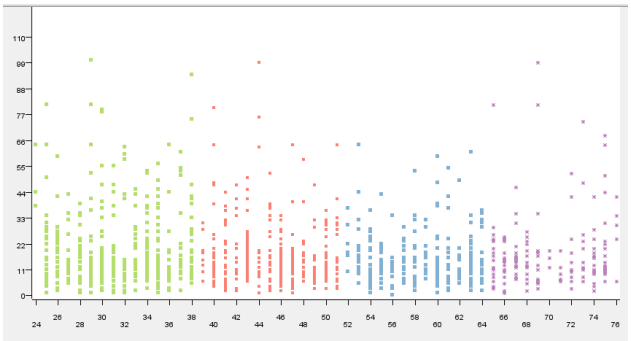

Figure 14. Clusters

5. Graph Analysis

For graph analysis, different data were used, and 8 different CSV files were merged including all the chat data, user data, and team data. Initially, there were more than 20,000 records in chat data but all data joined together data was reduced to 1900 as all the meaning less data was dropped during the joins. Finally, data about teams and their intercommunication was taken for the graph analysis. Graph for the teams and their communications can be seen in fig 15.

Data file was loaded into neo4j for analysis. Neo4j is a very powerful tool for graph analysis and it uses cypher query language (CQL). After loading the CSV data three types of nodes were created (Player, Teams, Chat) and their relationship were also defined using CQL.

Nodes and their relationship can be seen in fig 17. All different nodes have different colors. Furthermore, important nodes (teams) were analyzed based on the degree centrality and betweenness centrality. Most stronger teams (in terms of the number of players) are listed in descending order in accordance with degree centrality and betweenness centrality in Fig 16 and fig 18. Moreover, the average clustering coefficient (CC) was also calculated to check the strengths of the overall community or graph and CC was 0.9098 which means the community is strong.

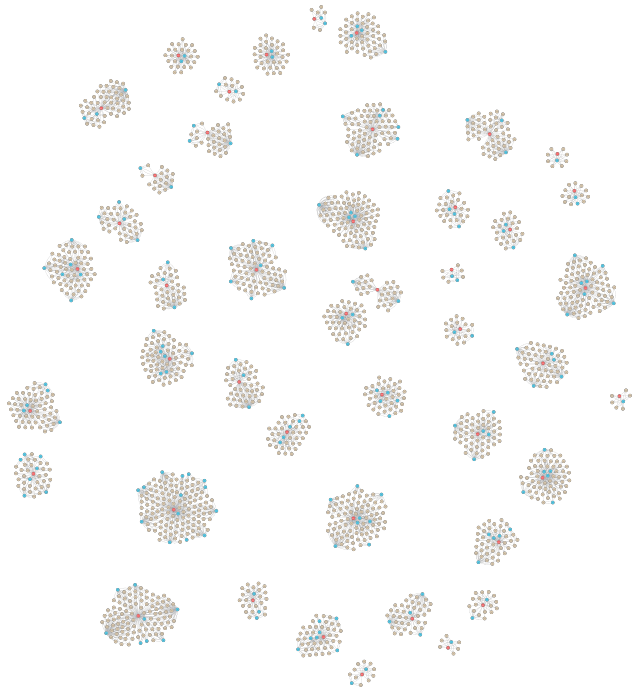

Figure 15. Graph for teams

6. Ethics of Big Data

The increasing use of big data in various fields has raised concerns about ethical issues associated with its collection, storage, analysis, and use. This report will discuss the current and foreseeable ethical issues related to big data, particularly in biomedical contexts, criminal justice, and

t.Name	degree_centrality
"Qj41nn"	120
"VpF5Lrtz7"	96
"EkwgEvX3h"	82
"8q0Umalp"	78
"LjcMqQ"	74
"b3PqpjmZ6TT"	66
"0G8HvkLE"	60
"mvigjg"	58
"Ev6gWyG3"	58

Figure 16. Important teams (degree centrality)

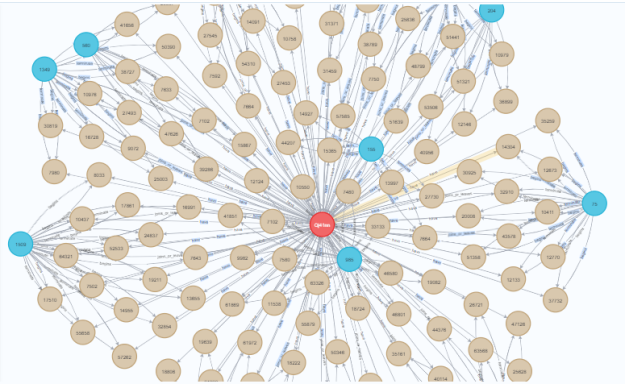

Figure 17. Closer look of team network

socioeconomic inequalities. The essay will also examine what data ethics is and the need for a framework to redress predictive privacy harms.

The rapid development of big data analytics has enabled biomedical researchers to extract valuable insights from large datasets. However, this has also raised ethical issues such as data privacy, consent, and ownership. (Mittelstadt & Floridi, 2016) argue that ethical issues in biomedical contexts are worsened by the increasing use of social media and other online platforms to collect data. This has resulted in data being collected and used without individuals' knowledge or consent, raising concerns about the right to privacy.

Data ethics, according to (Floridi, 2016), is concerned with the ethical and moral implications of the collection, analysis, distribution, and use of data. They suggest that data ethics should involve the protection of privacy, informed consent, and data ownership. This means that individuals have the right to control their data and how it is used, and organizations must obtain consent before collecting and using individuals' data.

In the criminal justice system, algorithms are increasingly used to make decisions about bail, sentencing, and parole. However, (Veale et al., 2018) argue that the use of these algorithms raises ethical issues related to transparency and accountability. They suggest that individuals have the right to an explanation of how algorithms are used to make decisions about them, particularly when those decisions have significant consequences.

The use of big data in socioeconomic contexts has also raised ethical concerns. (Eubanks, 2018) elaborates that the use of automated decision-making systems to determine eligibility for social services has led to the profiling and punishment of the poor. This has resulted in the negative impact on existing social inequalities. The Author suggests that data ethics should include the protection of vulnerable populations, such as low-income individuals.

Big data analytics has also raised concerns about due process and predictive privacy harms. (Crawford & Schultz, 2014) discuss that big data analytics may result in discriminatory outcomes and privacy violations. They suggest that a framework is needed to address these harms, which includes the right to notice, the right to contest, and the right to redress.

In summary, the use of big data has created ethical concerns regarding privacy, consent, transparency, accountability, and social fairness. Data ethics aims to protect people's rights to control their information and ensure openness and accountability in decision-making. To resolve privacy issues and prevent unfair outcomes, a framework that includes the right to be informed, to contest, and to get help is necessary. It is essential for organizations and policymakers to consider these ethical concerns when creating and implementing big data systems

node_name	score
"Qj41nn"	8487.0
"VpF5Lrtz7"	4708.0
"EkwgEvX3h"	3668.5
"8q0Umalp"	2958.2000000000007
"LjcMqQ"	2931.5
"b3PqpjmZ6TT"	2393.5
"0G8HvkLE"	1882.0
"mvigjg"	1865.5
"5aw5S13OD"	1703.0
"Ev6gWyG3"	1677.5

Figure 18. Important teams (betweennes centrality)

7. Link to the Colab

https://colab.research.google.com/drive/1nY6EG7z_4ZQRHLnDdaw80ghN8vIEOsSF#scrollTo=arH_LBjQ59TQ

References

Acciarini, C., Cappa, F., Boccardelli, P., and Oriani, R. How can organizations leverage big data to innovate their business models? a systematic literature review. *Technovation*, 123:102713, 2023. ISSN 0166-4972. doi: https://doi.org/10.1016/j.technovation.2023.102713. URL https://www.sciencedirect.com/science/article/pii/S016649722300024X.

Chambers, B. and Zaharia, M. *Spark: The definitive guide: Big data processing made simple*. " O'Reilly Media, Inc.", 2018.

Crawford, K. and Schultz, J. Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.*, 55:93, 2014.

Davenport, T. H. and Dyché, J. Big data in big companies. *International Institute for Analytics*, 3(1-31), 2013.

De Mauro, A., Greco, M., and Grimaldi, M. A formal definition of big data based on its essential features. *Library review*, 65(3):122–135, 2016.

Dumbill, E. What is big data? an introduction to the big data landscape. *Strata 2012: Making Data Work*, 2012.

Eubanks, V. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

Fan, W. and Bifet, A. Mining big data: current status, and forecast to the future. *ACM SIGKDD explorations newsletter*, 14(2):1–5, 2013.

Floridi, L. Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083): 20160112, 2016.

Gantz, J. and Reinsel, D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007(2012): 1–16, 2012.

Kaisler, S., Armour, F., Espinosa, J. A., and Money, W. Big data: Issues and challenges moving forward. In *2013 46th Hawaii international conference on system sciences*, pp. 995–1004. IEEE, 2013.

Kshetri, N. Big data s impact on privacy, security and consumer welfare. *Telecommunications Policy*, 38(11): 1134–1145, 2014.

Laney, D. et al. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001.

Lee, H., Noghabi, S., Noble, B., Furlong, M., and Cox, L. P. Bumblebee: Application-aware adaptation for edge-cloud orchestration. In *2022 IEEE/ACM 7th Symposium on Edge Computing (SEC)*, pp. 122–135. IEEE, 2022.

Lin, J. and Dyer, C. Data-intensive text processing with mapreduce. *Synthesis lectures on human language technologies*, 3(1):1–177, 2010.

Loi, M. and Dehaye, P. O. If data is the new oil, when is the extraction of value from data unjust? *Filosofia e Questioni Pubbliche*, 7(2):137–178, 2017.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A., et al. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, 2011. URL https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation.

Mayer-Schönberger, V. and Cukier, K. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., and Barton, D. Big data: the management revolution. *Harvard business review*, 90(10):60–68, 2012.

Mittelstadt, B. D. and Floridi, L. The ethics of big data: current and foreseeable issues in biomedical contexts. *The ethics of biomedical big data*, pp. 445–480, 2016.

Munawar, H. S., Ullah, F., Qayyum, S., and Shahzad, D. Big data in construction: Current applications and future opportunities. *Big Data and Cognitive Computing*, 6(1), 2022. ISSN 2504-2289. doi: 10.3390/bdcc6010018. URL https://www.mdpi.com/2504-2289/6/1/18.

Naeem, M., Jamal, T., Diaz-Martinez, J., Butt, S. A., Montesano, N., Tariq, M. I., De-la Hoz-Franco, E., and De-La-Hoz-Valdiris, E. Trends and future perspective challenges in big data. In Pan, J.-S., Balas, V. E., and Chen, C.-M. (eds.), *Advances in Intelligent Data Analysis and Applications*, pp. 309–325, Singapore, 2022. Springer Singapore. ISBN 978-981-16-5036-9.

Narkhede, N., Shapira, G., and Palino, T. *Kafka: the definitive guide: real-time data and stream processing at scale*. " O'Reilly Media, Inc.", 2017.

Shahrivari, S. Beyond batch processing: towards real-time and streaming big data. *Computers*, 3(4):117–129, 2014.

Sharma, S. Rise of big data and related issues. In *2015 Annual IEEE India Conference (INDICON)*, pp. 1–6, 2015. doi: 10.1109/INDICON.2015.7443346.

Veale, M., Binns, R., and Edwards, L. Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376 (2133):20180083, 2018.

Vitorino, J. P., Simão, J., Datia, N., and Pato, M. Ironedge: Stream processing architecture for edge applications. *Algorithms*, 16(2):123, 2023.

Warren, J. and Marz, N. *Big Data: Principles and best practices of scalable realtime data systems*. Simon and Schuster, 2015.

White, T. Hadoop: The definitive guide, storage and analysis at internet scale. *White–London: O'Reilly Media*, 2015.

Wingerath, W., Gessert, F., Friedrich, S., and Ritter, N. Real-time stream processing for big data. *it-Information Technology*, 58(4):186–194, 2016.

Zikopoulos, P. and Eaton, C. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.