BIG DATA ANALYTICS

By

Uttara Naidu

Submitted to

The University of Liverpool

MASTER-OF-DATA-SCIENCE-AND-ARTIFICIAL-INTELLIGENCE

CSCK501 Global Trends in Computer Science October 2024

Word Count: 1993

16/12/2024

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Chapter 1. INTRODUCTION

With the advent of an era of smartphones 18 years ago, we have slid headlong into an exponential growth of technology. Our digital journey towards automation, connectivity and security is tailored to our journey towards an increasing demand of convenience and ease. As a consequence, the quantum of data that is spawned worldwide on a daily basis has exploded. As of 2024, the magnitude of data generated is:

2.5 quintillion bytes per day (Šprem, 2024)

This has led to coin the term 'Big Data'.

The term has obscure origins and is not restricted to a single organization or person. In 1980s, subtle hints were surfacing, but the concept started taking shape in 1996 when Silicon Graphics (SGI) advertised Big Data in their magazines in 1996, John Mashey presenting the concept of Big Data in 1998, and finally it reached META Group via Laney in 2001, who defined the three Vs of Big Data (volume, variety, and velocity). Although META entered the race at a much later stage, they took the concept leaps and bounds ahead.
(Diebold, 2021)

This essay looks deeper into the concept and explains how it works.

Chapter 2. BIG DATA

2.1 Definition

Big Data definition is supported by the 5Vs, as shown in figure 1 below:

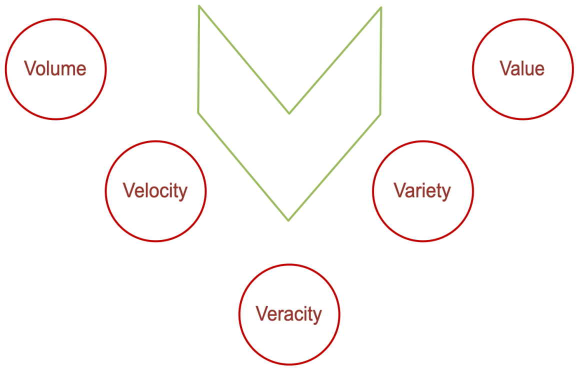

Figure 1. 5Vs of Big Data

Traditional approaches focused on the ability of storage devices to scale-up as per the quantity of incoming data. This approach was not suitable for large volumes of Big Data. Therefore, instead of restricting to a single storage device, the organizations came up with a method called *scale-out,* where the data is stored amongst a cluster of storage devices. This was the dawn of Big Data. (University of Liverpool, 2024)

2.1.1 Volume:

It refers to the arrival of high volumes of uninterrupted data from multiple sources that can be handled by a Big Data infrastructure. E.g. Walmart produces approximately 2.5 petabytes of data hourly. (Marr, 2021)

2.1.2 Velocity

This refers to the speed with which data is generated. E.g. In 2011, New York inhabitants were notified about an earthquake via Twitter, 30s before they felt the tremors themselves (Mohamed, 2020).

2.1.3 Veracity

It refers to the accuracy of Big Data. The authenticity of data is the foundation on which data analysis rests. Gaps or inaccuracies while transferring, or structuring data can lead to issues like data partisan, irregularities, and discrepancy. For example, Google's COMPAS miscategorized black people as more likely to reoffend as compared to white people. (Denison, 2024)

2.1.4 Variety

It refers to the types of data that are produced by different sources like sensors, transactions, streaming platforms etc. The data can be structured (schemas and meta data), unstructured (audio, video, images) or semi-structured (JSON, XML).

2.1.5 Value

It refers to the value that Big Data brings to organizations. A major part of Business decisions today is data driven. Consequently, organizations must critically brainstorm the requirements and expectations of the analysis outcome prior to beginning with the process of setting up the data model in big data infrastructure.

2.2 Big data Framework

The Big Data infrastructure is incorporated in Big Data Framework. The framework consists of three layers that are illustrated in *Figure 2.* To manage, process and analyze data, there are multiple tools and techniques tailored to the layers of Big Data. E.g. a single platform like Apache Kafka offers Apache Kafka Connect for ingestion and Apache Kafka Stream for processing (Šprem, 2024)

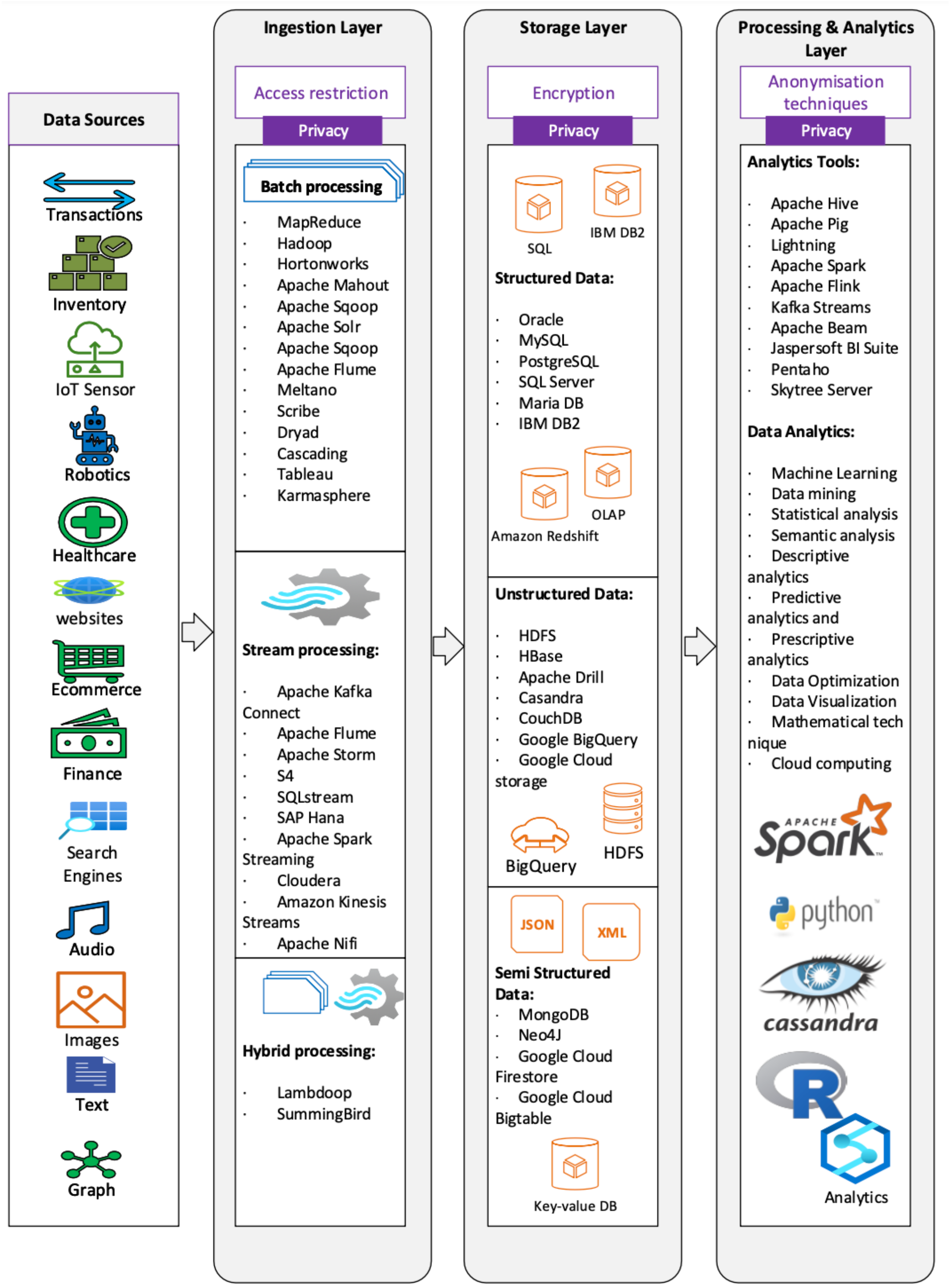

Figure 2. Big Data Framework (Mohamed, 2020)

2.2.1 Ingestion Layer

Data generated by multiple sources like sensors, social media, search engines and so on is consumed by the framework by a process called ingestion. Depending on the velocity of data, an appropriate ingestion technique - Batch processing, Stream processing or Hybrid processing is employed. (Sukumaran,2023)

In 2003, Google proposed a paradigm of MapReduce system i.e. storage followed by assessment called batch processing, whereas in 2010, Yahoo proposed a paradigm of S4 i.e. real-time streaming assessment (Mohamed, 2020). Additionally, this layer handles data transformation actions on raw data (Shiyal, 2021).

2.2.2 Storage Layer

Based on the type of data, an appropriate database i.e. relational database like SQL or non-relational database like NoSQL is used. The selection of database is crucial here since it also defines the accessibility, portability and compliance or manageability of data. In addition, the ease of interfacing with AI models and data extraction is considered while outlining the data model. (Sukumaran,2023)

2.2.3 Processing & Analytics layer

Once the data is stored in relational or non-relational database, it is retrieved to be processed and analyzed by an array of tools like Hike, Storm etc. The processing layer tools can provide an insight, tractability, and interpretability of data. Therefore, selection of a processing tool is directly proportional to the performance of overall framework. Scalability, choice of in-memory or external processing engine, efficiency, speed, adaptability are certain selection criteria. (Šprem, 2024)

2.2.4 Data Transformation

There are two methods of transforming the data (Shiyal, 2021):

1. ETL i.e. **e**xtract-**t**ransform-**l**oad is a traditional process where data is transformed before its storage in data warehouse.
2. ELT i.e. **e**xtract-**l**oad-**t**ransform is a newer process where data is first stored in a database and then transformed in an appropriate format.

It is a vital stage to make the data compatible with analytics tools.

Chapter 3. BIG DATA FOR DEPRESSION RECOGNITION

In current digital epoch, the usage and impact of social media on people's psychological health has received a growing attention. One of the flourishing areas of research is to gauge the mental health of people by monitoring their social media activities.

This chapter explains how Big Data can be leveraged to tackle the growing concern among people regarding mental health ailments like depression due to excessive time spent on online platforms.

3.1 Impact of social media

As the popularity of smartphones have grown, so has the access to and usage of social media. These platforms indulge 59.4% of the masses worldwide i.e. 5 billion inhabitants. (Kemp, 2023).

Social Media Platform	Monthly active users (in Millions)
Facebook	3065
YouTube	2504
Instagram	2000
TikTok	1582
WeChat	1343

Table.1. Monthly active users as of April 2024 (Statista, 2024)

According to Pew Research Center survey in 2022, the social media brought people closer in terms of connectivity (80%), emotional bonding (67%), encouragement to exhibit their creativeness (71%) and acceptance by diverse populace (58%). (Katella, 2024)

Despite the positives that it offers, use of social media is one of the leading causes of mental health decline in teenagers.

A teenager, on an average, spends 4.8 hours daily on multiple social media platforms. Amongst these, 41% claim to suffer from a depression and 10% express a tendency towards suicidal thoughts. (DeAngelis, 2024)

The worldwide public concern over the use of social media has propelled the governments to undertake strict measures and seek methods to curb mental health decline.

3.2 Big data analytics in social media:

The data generated by social media platforms is Big Data, predominantly representing the properties of velocity and volume i.e. velocity with which the data is generated, and large volumes of data generated. (Angskun, 2022)

The Big Data analytics offers the ability to determine the unseen associations, and concealed or veiled patterns in data, by intricate examination and processing of data. (Chaudhary, 2021)

The online platforms' data undergoes a pre-processing stage - as in the noisy, superfluous, and abnormal data is eliminated, and data is transformed via Normalization techniques and data reduction techniques. (Chaudhary, 2021)

3.3 Depression Recognition Model using Big Data framework:

Figure 3 shows the proposed depression recognition model.

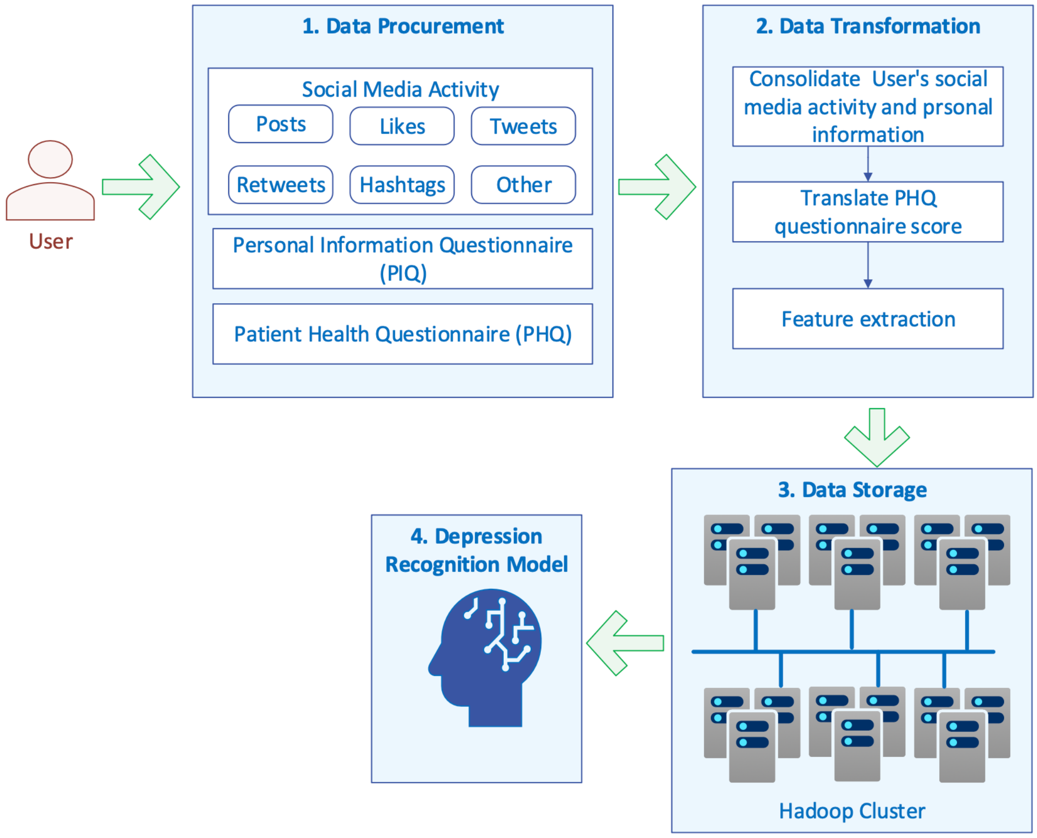
User
1. Data Procurement
Social Media Activity
Posts
Likes
Tweets
Retweets
Hashtags
Other
Personal Information Questionnaire (PIQ)
Patient Health Questionnaire (PHQ)
2. Data Transformation
Consolidate User's social media activity and prsonal information
Translate PHQ questionnaire score
Feature extraction
3. Data Storage
Hadoop Cluster
4. Depression Recognition Model

Figure 3. Depression Recognition Model (Angskun, 2022)

Data Sources

The model offers an option to choose the most frequently used online platforms. For example, the user is more active on Instagram, Twitter and YouTube as compared to other platforms like Facebook, TikTok, Snapchat etc.

As a second source of data, personal information of the user is fed to the model like age, gender, weight, income, number of family members, number of friends, employment status and so on.

A Patient Heath questionnaire (PHQ) is a set of questions aligned to determine the existing mental health level of the user. Certain points are allotted for each answer to determine depression signs.

Data Transformation

In the second stage, data transformation techniques are executed over the relevant data extracted from the first stage.

This data is fed to a language processor for sentiment analysis of user, based on the details of their online activity. It includes the reels that the user has shared or reacted to, updates about their life, retweets, hashtags used, videos streamed, frequency of tweets and so on. All this activity is split into a positive sentiment bucket and a negative sentiment bucket.

Additionally, the user's online account information like the number of followers, number of friends, the accounts/people they follow, time interval when the user is most active on online platforms i.e. mostly during the night or day are also taken into consideration.

Data cleaning is performed to delete irrelevant information like user's name etc.

Storage of Data

Hadoop is a prominent tool in the market for batch processing of data. The transformed data from the previous step is stored into a HDFS i.e. Hadoop Distributed File System where data is distributed amongst a nodes cluster. The nodes cluster comprises of worker nodes and master nodes for ease of access and maintaining anonymity. It offers the ability to process large magnitude of data using either distributed or parallel algorithms. (Mohamed, 2020)

AI Model

A python-based machine learning model is created to determine the level of depression. The feature parameters like user's consolidated online platform information, PIQ and PHQ are extracted from the database to perform analysis.

A classification technique is employed to determine the target variable i.e. the depression level of the user (Angskun, 2022):

Level 0	No depression
Level 2	Early signs of depression
Level 3	Depression

Table.2. Depression Level Description

Chapter 4. CONCLUSION

This essay explains the fundamentals of Big Data starting with the origins of term, definition, and the layers in its framework. With the help of real-life examples, the power of Big Data in terms of its properties, predominantly, volume and velocity are demonstrated. The essay further goes on to explain the end-to-end framework of Big Data incorporating the three layers and how each of this layer handles information. The essay then picks up a current topic, and discussed how the power of Big Data can be leveraged to mitigate the depression and self-harm risks associated with it. A Big Data based machine learning model is proposed to recognize early signs of depression. Finally, the essay ends with improvements that can be incorporated in the model in order to track the doom scrolling and its negative impact. The model can be improved by extracting further information of the user's online activities. Also, this model can be utilized by a medical professional to detect the patient's depression level and tailor the further treatment accordingly.

REFERENCES

Angskun, J., Tipprasert, S. and Angskun, T. (2022) ‘Big data analytics on social networks for real-time depression detection’, Journal of big data, 9(1), pp. 69–69. Available at: https://doi.org/10.1186/s40537-022-00622-2.

Buckley, G.J., Galea, S. and Wojtowicz, A. (eds) (2024) *Social media and adolescent health / Sandro Galea, Gillian J. Buckley, and Alexis Wojtowicz, editors ; Committee on the Impact of Social Media on Adolescent Health, Board on Population Health and Public Health Practice, Health and Medicine Division*. 1st ed. Washington, D.C: National Academies Press.

Chaudhary, K. et al. (2021) ‘Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics’, Journal of big data, 8(1), pp. 1–20. Available at: https://doi.org/10.1186/s40537-021-00466-2.

DeAngelis, T. (2024) *Teens are spending nearly 5 hours daily on social media. Here are the mental health outcomes.* Available at: https://www.apa.org/monitor/2024/04/teen-social-use-mental-health. (Accessed: 14 December 2024)

Diebold, F.X. (2021) ‘What’s the big idea? “Big Data” and its origins’, Significance (Oxford, England), 18(1), pp. 36–37. Available at: https://doi.org/10.1111/1740-9713.01490.

Denison, G. (2024) *8 shocking AI bias examples*. Available at: https://www.prolific.com/resources/shocking-ai-bias. (Accessed: 14 December 2024)

Katella, K. (2024). *How Social Media Affects Your Teen’s Mental Health: A Parent’s Guide*. Available at: https://www.yalemedicine.org/news/social-media-teen-mental-health-a-parents-guide. (Accessed:14 December 2024)

Martínez-Castaño, R., Pichel, J.C. and Losada, D.E. (2020) ‘A big data platform for real time analysis of signs of depression in social media’, International journal of environmental research and public health, 17(13), pp. 1–23. Available at: https://doi.org/10.3390/ijerph17134752.

Marr, B. (2021). *Walmart: Big Data analytics at the world’s biggest retailer*. Available at :https://bernardmarr.com/walmart-big-data-analytics-at-the-worlds-biggest-retailer/#:~:text=Walmart%20uses%20Big%20Data%20in,rapidly%20modelled%2C%20manipulated%20and%20visualised. (Accessed: 15 December 2024)

Mohamed, A., Najafabadi, M. K., Wah, Y. B., Zaman, E. A. K. and Maskat, R. (2020) 'The state of the art and taxonomy of big data analytics: view from new big data framework', Artificial Intelligence Review, 53(2), pp. 989-1037.

Rajeshwari, S. and Meenakshi, S. (2023) ‘The age of doom scrolling - Social media’s attractive addiction’, *Journal of Education and Health Promotion*, 12(1), pp. 21–21. Available at: https://doi.org/10.4103/jehp.jehp_838_22.

Shiyal, B. (2021) Beginning Azure synapse analytics : transition from data warehouse to data lakehouse / Bhadresh Shiyal. Place of publication not identified: Apress.

Šprem, Š. et al. (2024) ‘Building Advanced Web Applications Using Data Ingestion and Data Processing Tools’, Electronics (Basel), 13(4), pp. 709-. Available at: https://doi.org/10.3390/electronics13040709.

Statista, (2024). *Most popular social networks worldwide as of April 2024, by number of monthly active users (in millions).* Available at: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. (Accessed: 16 December 2024)

Sukumaran, A., Vergadia, P. and Narayanan, B. (2023) *Database Design and Modeling with Google Cloud: Learn Database Design and Development to Take Your Data to Applications, Analytics, and AI*. 1st edn. Birmingham: Packt Publishing, Limited.

University of Liverpool (2024) 'Lecturecast: Big Data Analytics', *Week 7: Big Data Analytics*. Available at: https://liverpool-online-study.com/course/view.php?id=2555&section=11. Accessed on: 6 December 2024

BIBLIOGRAPHY

Ahmed, O. et al. (2024) ‘Social media use, mental health and sleep: A systematic review with meta-analyses’, Journal of affective disorders, 367, pp. 701–712. Available at: https://doi.org/10.1016/j.jad.2024.08.193.

Dagan, D.T. and Wilkins, E.J. (2023) ‘What is “big data” and how should we use it? The role of large datasets, secondary data, and associated analysis techniques in outdoor recreation research’, Journal of outdoor recreation and tourism, 44, pp. 100668-. Available at: https://doi.org/10.1016/j.jort.2023.100668.

Lipschultz, J.H. (2024) Social media communication : concepts, practices, data, law and ethics / Jeremy Harris Lipschultz. Fourth edition. New York: Routledge.

Montag, C. et al. (2024) ‘Problematic social media use in childhood and adolescence’, Addictive behaviors, 153, pp. 107980-. Available at: https://doi.org/10.1016/j.addbeh.2024.107980.

National Academies of Sciences, Engineering, and Medicine. 2024. Social Media and Adolescent Health. Washington, DC: The National Academies Press. https://doi.org/10.17226/27396.

10-year BYU study shows elevated suicide risk from excess social media time for young teen girls. (2021, February 3). News. Available at: https://news.byu.edu/intellect/10-year-byu-study-shows-elevated-suicide-risk-from-excess-social-media-time-for-young-teen-girls. (Accessed: 14 December)