KEY CONCEPTS AND APPROACHES OF BIG DATA ANALYTICS

Marius IORDAN

Group TI-221, Faculty of Computers, Informatics and Microelectronics,
Technical University of Moldova, Chișinău, Republic of Moldova

Corresponding author: Marius Iordan, marius.iordan@isa.utm.md

Coordinator: Corina TINTIUC, university assistant, Department of Foreign Languages, TUM

Abstract. *The growing volume and complexity of data generated by individuals and businesses have led to the emergence of big data analytics as a vital tool for gaining insights and making better decisions to keep today's data-ruled world intact. This paper aims to explore the key concepts and techniques of big data analytics in information technology and the challenges and opportunities that organizations face when working with large and complex datasets.*

Keywords: *data processing, insights, datasets, machine learning, algorithm*

Introduction

In recent years, big data has emerged as a keyword used to describe large datasets of high complexity that are difficult to process and analyze using traditional processing techniques. To the average population, the term "Big Data" may hint at regular types of data that corporations and businesses store, like the inflow of customers, profits and losses, and so on. Many people still need clarification on what big data is, how it works, and how it can be applied in different industries. Big data has the potential to provide organizations with valuable insights that can help them make intelligent decisions, improve efficiency, and innovate. However, to fully grasp the power of big data, it is essential to have a comprehensive understanding of its key concepts and applications.

Origins and evolution of Big data

Big data has been around briefly, and its origins are older than you may think. Moreover, its origins are not even related to the IT field; the first trace of big data was seen way back in 1663 when John Graunt dealt with overwhelming amounts of information while he studied the bubonic plague, Graunt being the first-ever person to use the statistical data analysis [1]. Around 2005, people began understanding how much data users generated through media platforms such as Facebook and YouTube. Hadoop (an open-source framework to store and analyze big datasets) finished development the same year. NoSQL also began to gain popularity during this time [2].

By now, you may already have a clue what big data is and what its gimmicks are, but to be more exact, big data is datasets that are so voluminous that traditional data processing software cannot manage them, being more precise, as estimated in 2021 people generate nearly 278108 petabytes of data monthly [3]. The predicted amount of data generated for the next year are at whopping 149 zettabytes (Fig 1). However, these massive volumes of data can be used to address business problems you would not have been able to tackle before.

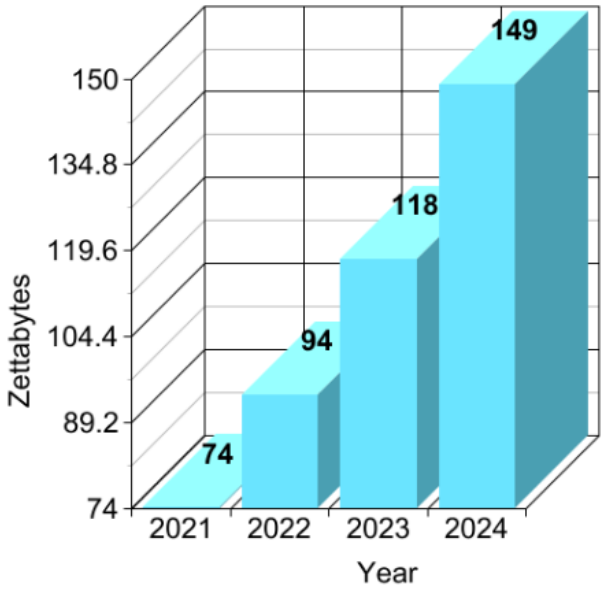

Figure 1. Zettabytes of data generated per year [5]

What are the characteristics of Big Data?

Three Vs characterize big data: **volume**, **velocity**, and **variety**.

- **Volume** is the amount of data that matters. With big data, you will have to process high volumes of low-density unstructured data (Fig. 2). The data always has an unknown value, such as Facebook posts, mobile app ratings, or sensor-enabled equipment. For some organizations, the sets of data come in different shapes and sizes, which might be terabytes of data. For others, it may be hundreds of petabytes [2].

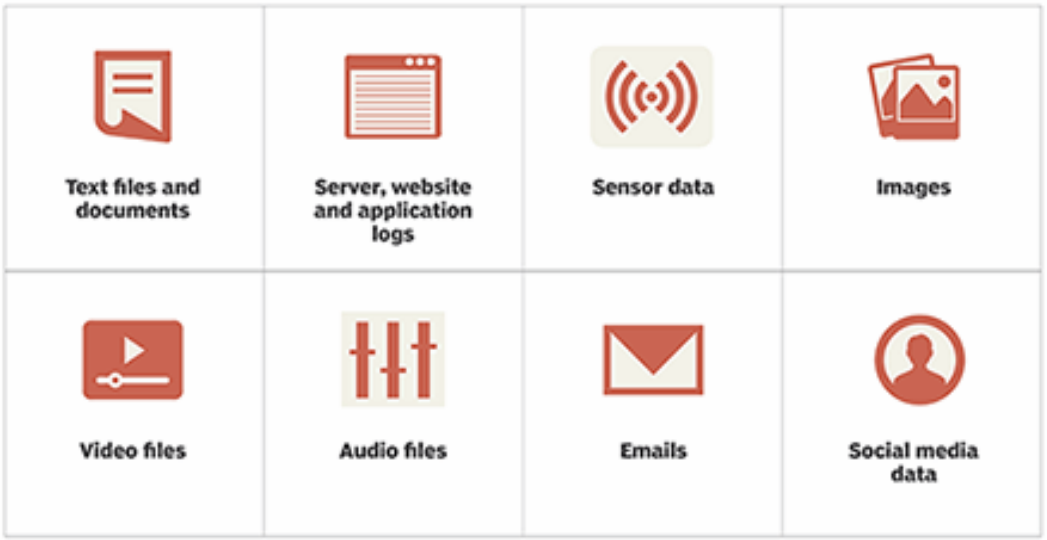

Figure 2. Unstructured data types

- **Velocity** is the fast rate at which the data is received or generated. Normally, the highest velocity of data is being streamed directly into memory instead of being written to disk. Some internet-enabled smart products or apps that operate in real-time will require real-time evaluation and action, some examples being apps like google maps, which gather information about traffic all over the world, updating the route status in real-time for 100 percent of the users.
- **Variety** refers to the types of data available that, in some way or another, affect the desired outcome or the offered service. Traditional data types were structured and fit neatly in a relational database (see Fig.3), which stores and provides access to data points that are related to one another [2].

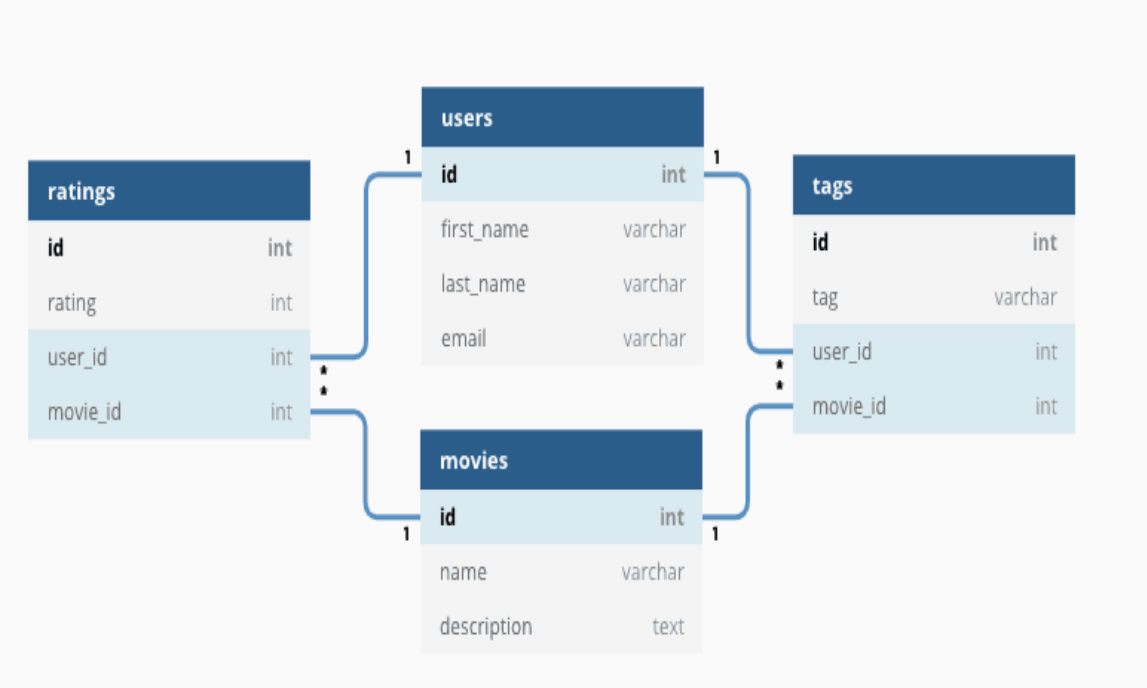

Figure 3. Relational Database [6]

With the ascending inflow of big data, data comes in new unstructured data types, which cannot be stored in a traditional database with defined types of data to be stored because the incoming data is somewhat random and has no defined type or category. Some examples are text, audio, and video, which all require additional preprocessing to derive meaning and support metadata.

Over the past few years, two more Vs have been implemented in the structure of big data. The new Vs are **value** and **veracity** [2]. Data has innate value. However, it has no use whatsoever until that value is discovered. Equally important is how much you can rely on your data and how trustworthy the dataset's origin is.

How does big data work?

Big data gives you new insights that open up new opportunities and business models. Getting started involves three key actions:

1. **Integrate.**
   Big data combines data from different sources. New strategies and technologies are needed to process and analyze datasets at a large scale.
2. **Manage.**
   Storage is crucial for big data. Data can be stored in any format, and processing can be applied on-demand in on-premises or cloud-based solutions.
3. **Analyze.**
   The primary goal is to gain insights and make informed decisions. Data analysis involves visual analysis, exploration, sharing, and building data models using machine learning and AI. The goal is to put data to work for business purposes.

How do we process and use big data in our favor?

When it comes to processing big data, there are multiple approaches: machine learning, association rule learning, social network analysis, etc.

Machine learning is a powerful tool for processing and analyzing big data because it enables computers to learn from large datasets and to make predictions or identify patterns being explicitly programmed to do so [3]. Some ways in which machine learning is useful in big data processing:

- Data processing: cleaning and preprocessing large datasets, including removing missing or inaccurate data, normalizing data, and transforming data into a format.
- Classification and clustering: automatically classifying or clustering data into groups based on similarities, which can help to identify patterns and relationships within large datasets.
- Predictive modeling: building predictive models that can be used to make predictions or identify trends in large datasets.
- Anomaly detection: identifying outliers or anomalies within large datasets, which can help to detect fraud or other unusual activity.
- Natural language processing: processing and analyzing large amounts of unstructured text data such as social media posts or customer reviews, which can help to identify trends or sentiment.

Real-life examples of what machine learning does are distinguishing spam and non-spam emails, learning user preferences and making recommendations based on this information, determining the probability of winning a case, and setting legal billing rates.

Association rule learning is a type of machine learning technique used for discovering interesting correlations between variables in large datasets. It was first used by major supermarket chains to discover interesting relations between products, using data from supermarket point-of-sale systems [3]. For instance, a grocery store might use association rule learning to identify that customers that buy bread are also likely to buy butter and use this information to make recommendations or adjust pricing strategies. The main algorithm behind association rule learning is *Apriori Algorithm* (Fig. 4) used in data mining. This algorithm identifies frequent itemsets, which are combinations of items that occur together in a specified percentage of transactions. Once frequent itemsets are identified, association rules are generated that describe the relationships between the items in the itemsets. These rules are typically expressed in the form of "if-then" statements. Even though this process may seem simple, it can be computationally expensive and require specialized hardware or software to process large datasets efficiently.

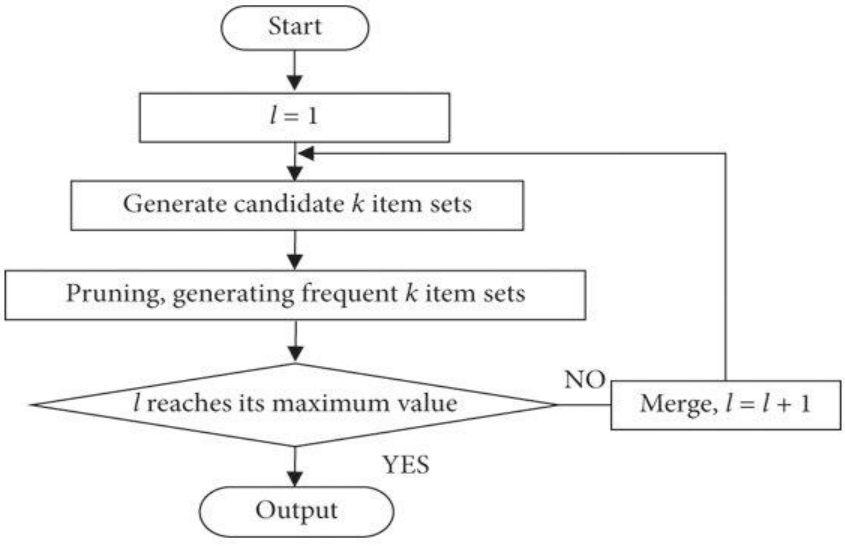

Figure 4. Apriori algorithm flowchart [7]

Social network analysis (SNA) is a technique that was first used in the telecommunications industry and quickly adopted by sociologists to study interpersonal relationships [3]. Social network analysis is the study of social networks and their underlying structure, including relationships between individuals, groups, organizations, and communities. It is a valuable tool for analyzing large datasets contained in big data. Some ways in which social network analysis is being used in big data processing:

- Building recommender systems that suggest relevant content or products to users based on their social connections.
- Detecting fraudulent activities, such as fake profiles, spam, or bot-driven activities.
- Identifying key influencers that could be targeted by businesses for marketing purposes.

Conclusions

In conclusion, big data analytics helps organizations make immediate, better-informed decisions and find more efficient ways of doing business. More companies have the opportunity to develop innovative new products to meet customers' changing needs.

As the field of big data is continuously growing, new challenges will arise, such as the ethical and legal implications of collecting, storing and analyzing large amounts of data. Therefore, it is essential that organizations adopt a responsible approach to big data and ensure that they comply with relevant regulations and standards.

Web references:

1. All About the Basics of Big Data: History, Types and Applications [online]. [accessed 03.04.2023] Disponibil: https://www.analyticsinsight.net/all-about-the-basics-ofbig-data-history-types-and-applications/#:~:text=History%20of%20big%20data,to%20use%20statistical%20data%20analysis
2. What Is Big Data? [online]. [accessed 03.04.2023] Disponibil: https://www.oracle.com/big-data/what-is-big-data/#
3. 7 Big Data Techniques That Create Bussines Value [online]. [accessed 03.04.2023] https://www.firmex.com/resources/blog/7-big-data-techniques-that-create-business-value/ [accessed 03.04.2023]
4. https://resources.useready.com/blog/why-analytics-is-the-future-of-fintech/ [accessed 03.04.2023]
5. https://www.techtarget.com/searchbusinessanalytics/definition/unstructured-data [accessed 03.04.2023]
6. https://www.heavy.ai/technical-glossary/relational-database [accessed 03.04.2023]
7. https://www.researchgate.net/figure/Flowchart-of-the-Apriori-algorithm_fig2_354189060 [accessed 03.04.2023]