UNSTRUCTURED DATA: WHAT IS IT, HOW IS IT STORED, AND HOW IS IT BEING USED WITH EMERGING TECHNOLOGIES

Linh Dan Nguyen – 103488557

Faculty of Science, Engineering and Technology

Swinburne University of Technology

ABSTRACT

With the invention of several modern technologies, and online communication channels, such as social media, emails, and voice chat, structured data is no longer the dominant of the generated data. Instead, unstructured data has invaded the data society as all data generated from online channels does not have a specific format. Their formats are various, from text, image, and voice data to sensor data. Unstructured data is a newly popular type of data, which is considered challenging for many data scientists. This research paper will discuss different aspects of unstructured data, such as the definition, types, storage methods, and how it is applied combined with emerging technologies.

Keywords: Big Data, Unstructured Data, Structured Data, Data Analysis, Data storage, Technology

1. INTRODUCTION

Unstructured data is simply defined as data with images, text, random emails, and all unformatted data. (Feldman and Sanger, 2007). [1] The massive changes in the social media, e-commerce, and communication channels development in the current century have led to the increasing need for unstructured data. Research written by C. White proved that 80% of the data is unstructured [2]. Although unstructured data is hard to analyze, it is a valuable asset to any business. An article from IDC predicts that there will be 175 zettabytes (ZB) of data, while the amount of unstructured data will reach 90% in 2025 [3]. Dealing with a large amount of unstructured data is not an easy task, and the same problem occurs with data extraction. Unstructured data is important, but not many professional data scientists are willing to analyze this kind of data. This paper will introduce several types of unstructured data and give a brief discussion on why data scientists feel vague when trying to extract useful information from raw unstructured data. This paper also describes some cases the unstructured data accompany new technologies.

2. UNSTRUCTURED DATA & APPLICATIONS

In the fourth industrial revolution, data changed its storing method from storing in a block to flowing through different channels. There is no general format applied for all data to manage them. Each piece of data will have its structure based on the features of its original channels or

tools. The unstructured data includes semi-unstructured and unstructured data. However, semi-structured data is not powerful enough to replace unstructured data as a new dominant in the data field. It is estimated that nearly 2.5 quintillion bytes of data are generated from Internet content every day (Burghard & Schubmehl, 2012) [4] ,300 million photos are shared on social platforms per day, 231.4 million emails are sent each minute [5] The sources of unstructured data come from all content on the Internet, from posts, photos on the social media, study materials, tutorial videos, reference links on educational websites, to graphs and customers comments on the business websites. Unstructured data is always vague and chaotic. It is not commonplace data you can absorb immediately. It must go through an extracting process to be recognized and accepted.

2.1. Textual Data

Textual data is taking over the most percentage in the unstructured data total. It often appears in the world's daily life in the forms of posts on social media, articles on online news websites, products description, customer feedback on e-commerce websites, and learning documents. About 70-75% of unstructured data is text (He, Patel, Zhang, Chang) [6] Textual data generated on the Internet is the best resource that reflects users' behaviors. It acts as a spy observing users when users interact with different services on the Internet. Many giants on social media, such as Twitter, Facebook, Instagram, and e-commerce like Amazon and eBay, rely on textual data to identify their potential customers and upcoming social trends. In recent years, Facebook has been closer to users when they can quickly filter the search content based on customers' favorites and run more advertising about the products that users want to purchase. These giants analyze the data extracted from several zettabytes of unstructured data to collect information about customers' interests, and which content is their favorite. Social media applies the extracted customer data to display more content related to their interests and run more ads about the products that users are paying attention to. Not only social media platforms but the textual unstructured data is also used for Google search engine optimization. Every day, the search engine receives 8.5 billion searches with the desire to find different resources (Festic, Buchi, Latzer). [7] The search engine, as a result, gets massive amount of customers' desired data to determine suitable ads to show. Due to the complex unstructured data format, the data extracted to find the trend is not always correct. Data mining can pass a piece of data if it is too obscure. Human languages have many versions in dialects and expressions ways, and computers are still immature to master human minds. Language is a big challenge in unstructured data analysis. According to data research from Northeastern University, nearly 1.7 megabytes (MB) of data is created by a single user each second, so data mining will not have any second break in the data classifying process. [8] Some tools can assist the textual data classification, such as R,

Orange, and Python libraries. Overall, dealing with textual unstructured data is a tough task, which is a part of the Natural Language Processing field, it is widely applied in business analysis and is still in the development process to extract the data faster and more effectively.

2.2. Audio data

Besides textual content, audio is also a popular type of unstructured data. Audio data is growing and taking over a more important position in the data field. Businesses can extract the information from customer calls to identify the most popular issues that the customers mention in the phone calls. As a result, businesses can have more data about customers' attitudes, which can be positive or negative. The data extracted from phone calls can also help managers reflect the employees' attitudes toward their customers. The information extracted from unstructured audio data can support the marking process in education. Voice analytical tool is applied in online language testing exams, such as Duolingo, to automatically mark the test without human intervention. The audio analysis still has mistakes related to the meaning of the extracted data. It is better at evaluating the expressions in the audio files based on the volume level, pronunciation, and intonation than evaluating the content. Analyzing unstructured audio data is a challenging research field that belongs to the Artificial Intelligent field. The manual transcript of an audio file is obsolete and replaced by an automated transcript generated from the audio files. However, the development of audio data analysis still has not been able to extract all raw unstructured data to correct information. The audio analysis tools can extract most of the valuable information. However, it still needs more advanced improvements in deep learning to predict the trends from the extracted data with a higher accuracy level. Some Python libraries support audio data analysis, such as PyAudio, Torchaudio, SoundFile, and PyDub. Since 2007, Essentia - an open-source C++ library for audio analysis, was introduced by Affero with a collection of standard algorithms used for audio analysis. [9]

2.3. Image data

The third well-known unstructured data type is Image. This data type often accompanies textual data. They are especially widely used on social media platforms in form of the posts or statuses. Facebook has nearly 2.97 million active users at the beginning of 2023 (37% of the world's population). This number is still going up. Therefore, it is not surprising to know that each day Facebook receives 4 petabytes (PB) - a massive amount of data, and most are unstructured. By the end of 2020, 6.1 billion smartphones are sold all over the world, and every day 350 million photos are updated on Facebook, which is equal to 4051 photos per second. [10] Due to the storm of image data on the Internet, we need a breakthrough in image analytics to use the information extracted to serve business development. Based on the content of images that users viewed when they use the Internet, the data mining tool

will record that content and produce reports of users' trends in viewing as well as uploading photos: which image content can attract more views and why they can do that. Which photo styles or color palettes can be more successful compared to others?

Image data plays a vital role in assisting medical development. Medical images include medical documents, health check reports, and CT scans that need to be analyzed to find patients' health features. The patterns recognized from unstructured image data are beneficial in finding appropriate treatments for the patients. Python is the best programming language used for data analysis. Besides the libraries used for audio analytics, Python also provides libraries that concentrate on image analytics, such as OpenCV, Scikit-Image, SimpleITK, and Mahotas... Other tools like R and MATLAB are also developing to enter the image data analytics field.

2.4. Sensor data

"Sensor data quality plays a vital role in Internet of Things (IoT) applications as they are rendered useless if the data quality is bad" [11]. In research by Cisco, they estimated the amount of sensor data to be approximately 850 zettabytes. [12] Sensor data is unstructured data like other data types mentioned in the previous section. However, the popularity of sensor data in practical life is less than other data types due to its specialty in the data generating method. Textual, image and audio data are mainly produced thanks to human interactions. New textual, image, and audio data is generated when Internet users perform any actions on the user interface, such as uploading a photo, sending voice messages, writing a description for a product on the e-commerce website, or searching for content on the search engine. However, most sensor data follow the real-time update principles and is generated by different machines like weather data, location-based data on the map, natural resources data, radar waves, and gravitational waves. Other unstructured data types are static data. The data is recorded and stored in the unstructured database for future analysis. In contrast, sensor data is dynamic, and changes continuously based on the real-time process. Sensor data analysis is facing a challenging issue as the data analytics tools must perform the update action continuously without any interrupts. All extracted information must be clean and not have any data-losing errors.

Every second new sensor data is generated an updated, so the amount of data is massive and requires a high-powered database management system to archive. The data will be transferred from the sensors to the database for storage. These sensors can be a chip attached to smartphones (to generate the GPS signal), a satellite, or smart devices in your house. For example, the sensor in an air-conditioner is installed to record data about the frequency at that family members use the air-conditioner to recognize the trends of electricity usage in the family. In another example, the data about locations and traffic is collected to analyze the traffic

status of a country. The traffic data recorded using piezoelectric sensors can be the traffic congestion frequency or the average number of transportations in a specific period. Sensor data is unstructured data. It has the same general features as other unstructured types like lawless and turbulent. The sensor raw data must go through a transformation process to generate helpful information for trend prediction. Some tools to analyze data retrieved from the sensors include SensiML, PySensors, and R.

3. MANAGING UNSTRUCTURED DATA

Unstructured data is a data type without any specific format. Therefore, we cannot use relational databases to store and retrieve unstructured data. The failed attempt to store unstructured data using the relational database makes data analysis harder than ever. However, it is also the main reason for the appearance of new data storage. Creating a new database system to store unstructured data is a big challenge to all software developers, data analysts, and system managers in the earliest days of data analysis. After more than a decade with data, the developers brought new data analytics tools to the world to support the difficulties in unstructured data extraction. These tools include NoSQL, Hive Apache, DynamoDB, Cassandra, MongoDB, and Neo4J. The databases used for unstructured data storage are developing rapidly, and there are other databases besides some databases listed above. Hive Apache is the data storage of all unstructured data generated from Facebook content. Netflix uses the Cassandra database to store and analyze their customer's behaviors. Each unstructured database will have unique features, so businesses can consider and choose the most suitable database for their business demand. The advantage of Hive Apache is that it is suitable for complex data sets analysis, while Cassandra is an open-source database that follows the peer-to-peer architecture.

Due to the significant increase in the data, most businesses and organizations must spend more funds investing in the database to handle the speedy-growing unstructured data they receive day by day. Giants on the Internet like Facebook or Netflix never store all data in one repository due to the security and the massive data load. Instead, they divide the data into several repositories in different locations, which is often called a Distributed system. Storing data in a distributed system or the cloud platform is more beneficial than in only one local database because it can provide a higher data security level and allow the disaster recovery process to be faster.

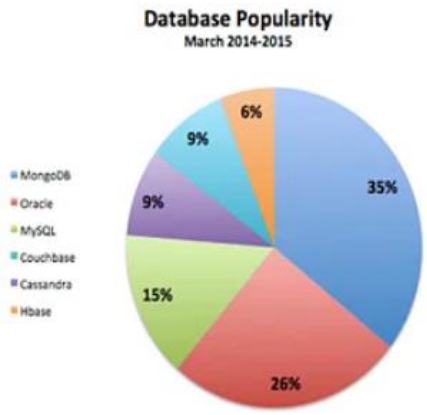

Figure 1: NoSQL database popularity (2014-2015)

Source: https://www.simplilearn.com

4. DISCUSSION ON THE GROWTH OF UNSTRUCTURED DATA

Andrew Ng. – the chief data scientist at Baidu stated: "Data is the new fuel that will power the digital generators of the AI revolution". As I discussed throughout the paper, unstructured data accounts for most data space in the world, with more than four-fifths of the total data, while structured data only takes over one-fifth of the total. The sources of unstructured data are various, from social media platforms, all websites on the Internet, and IoT devices to mobile applications. Each data source will produce unstructured data with some unique traits of the source. Inside a source, the generated data can be structured as follows a source general format. For example, all posts on social media platforms always have some textual content and images. When observing all generated data from all sources with a general view, the data is no longer structured because the data does not have any unified data format. Although the number of available tools for unstructured data analysis is less than for structured data, we cannot deny that unstructured data is growing constantly. It is on the way to achieving a higher position in the data society. Most human decisions nowadays are based on the trends or patterns recognized from the extracted unstructured data. If the unstructured is not extracted correctly, it is still a block of meaningless raw data. Otherwise, the information extracted from the raw data can lead the businesses to brilliant success if the business managers know how to convert the business-related information into new business strategies to boost business performance. Overall, the more organizations, businesses, and Internet users rely on the innovation of new technologies like the Internet, smart devices, and convenient apps, the more unstructured data generated, and the deeper unstructured data invades modern human society.

Table 1: Unstructured data growth rate

Company name	The amount of generated data
Facebook	4 petabytes of data per day (2020) 3.97 billion active users around the world (2022) Like button is clicked 5.2 billion times per day (2020) 500,000 photos uploaded per day (2020)
Instagram	1 billion active users around the world (2023) 50 billion photos have been shared since 2010 Like button is clicked 4.2 billion times per day (2023) 500 million stories are uploaded per day (2023)

Twitter	368 million active users around the world (2023) 6,000 tweets per day (2023) 200 billion tweets per year (2022)
Netflix	2.3 billion online videos (took over 17% of the world's total in 2021) 3,781 movies (2020)
Amazon	300 million active users around the world (2023) 350 million products (with product images and description)
Youtube	2.6 billion active users around the world (2023) 737.51 million comments (2022) 800 million videos (2023) 50 billion Shorts viewed per day (2023)
Google	40,000 search requests per second 3.5 billion search requests sent per day

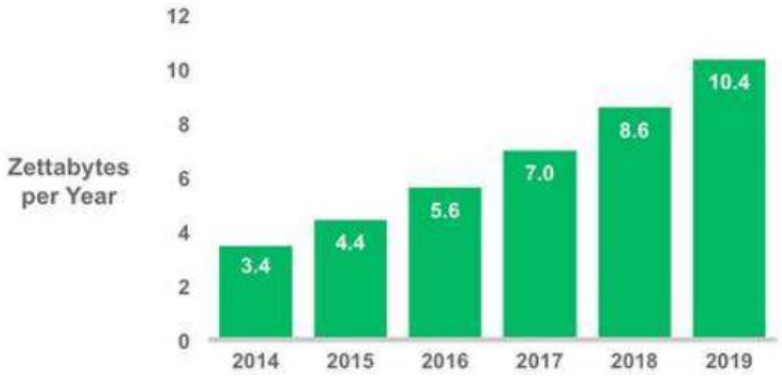

Figure 2: Annual data growth rate *Source: Cisco Global Cloud*

5. CONCLUSION

The twenty-first century is the era of new technologies. This era will continue many years later as it is making human life easier. Humans can make decisions more easily by looking at the necessary information extracted from unstructured data. Both static and dynamic unstructured data are indispensable parts of modern life as each data type will serve a specific human demand. Useful information extracted from the raw data is supportive for businesses, organizations, and also individuals. It will help the businesses create new policies that match the customer's desire. On the aspect of the impact on the individual users, unstructured data retrieved from their performed actions of them on the Internet help users quickly find the content they want magically by clicking on the recommendations suggested based on the extracted data from the raw unstructured

data. The sensor data retrieved from smart devices largely contributed to the weather forecast as it successfully saved many lives by identifying the upcoming natural disaster based on all weather-related data generated by the satellites and weather-predicting sensors. Unstructured data analysis is a potential emerging research field that focuses on all aspects and the applications of unstructured data in modern life. This paper successfully outlines the main types of unstructured data, describes some typical usage of the data in real life, suggests different databases used for unstructured data, and discusses the rocket growth of this data type.

REFERENCE

[1] Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press. (Accessed: April 7, 2023).

[2] C. White. Consolidating, Accessing, and AnalysingUnstructured Data. 2005 Dec. Business Intelligence Network article. Powell Media. LLC. (Accessed: April 7, 2023).

[3] Subramanian, K. (2022) Unstructured data management predictions for 2023, Dataversity. Available at: https://www.dataversity.net/unstructured-data-management-predictions-for-2023 (Accessed: April 7, 2023).

[4] Feldman, S., Hanover, J., Burghard, C., & Schubmehl, D. (2012). Unlocking the Power of Unstructured Data IDC Health Insigh. (Accessed: April 7, 2023).

[5] Thomas, O. (2012) Facebook: Users upload 300 million images a day (so please let us buy Instagram), Business Insider. Business Insider. Available at: https://www.businessinsider.com/facebook-images-a-day-instagram-acquisition-2012-7 (Accessed: April 7, 2023).

[6] He, B. et al. (2007) "Accessing the deep web," Communications of the ACM, 50(5), pp. 94–101. Available at: https://doi.org/10.1145/1230819.1241670.

[7] Festic, N., Büchi, M. and Latzer, M. (2021) "How long and what for? tracking a nationally representative sample to quantify internet use," Journal of Quantitative Description: Digital Media. Available at: https://doi.org/10.51685/jqd.2021.018.

[8] Bernard Marr, 20 Mind-Boggling Big Data Facts Everyone Must Read. Available at: https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/?sh=6a8014d817b1. (Accessed: April 7, 2023).

[9] Bogdanov, D. et al. (2013) "Essentia," Proceedings of the 21st ACM international conference on Multimedia [Preprint]. Available at: https://doi.org/10.1145/2502081.2502229.

[10] Andre, L. (2023) 53 important statistics about how much data is created every day, Financesonline.com. FinancesOnline.com. Available at: https://financesonline.com/how-much-data-is-created-every-day/ (Accessed: April 7, 2023).

[11] Teh, H.Y., Kempa-Liehr, A.W. and Wang, K.I.-K. (2020) "Sensor Data Quality: A Systematic Review," Journal of Big Data, 7(1). Available at: https://doi.org/10.1186/s40537-020-0285-1.

[12] Barnett, T. et al. Cisco Global Cloud index 2015–2020. (Accessed: April 6, 2023)