Unstructured Data: what is it, how is it stored, and how is it being used with emerging technologies.

Vi Luan Dang - 103802759

Faculty of Science, Engineering and Technology

Swinburne University of Technology

Ho Chi Minh, A35 Bach Dang, Tan Binh District

Abstract

Recent years have observed the ever-increasing momentum of digital data both structured and unstructured. This proved to be the challenge as well as a possibility for any organization since Business Intelligence plays an important role in the organization for gathering, integrating, and analyzing these data to be useful for informed decision making (Abdullah and Ahmad, 2015). The data collected by the organization can be categorized into three types: structured data, semi-structured data, and unstructured data. Unstructured data refers to data that does not have a predefined structure or data model (Kanimozhi and Venkatesan, 2015). Despite that, unstructured data is becoming increasingly important as it accounts for 80% of all the generated data (Taleb, Serhani and Dssouli, 2018) . A thorough understanding of unstructured data and its storage method will be crucial for harnessing the potential information locked in big data (Tanwar, Duggal and Khatri, 2015) . This literature review explores the relationship between big data and unstructured data, focusing on giving an accurate definition and how unstructured data is transformed, stored, and used. It will examine the challenges associated with transforming unstructured data into structured data and various approach of how emerging technologies such as artificial intelligence and machine learning are being used to harness the full potential of unstructured data.

Keywords: Unstructured data, big data, structured data, storage method, data analysis, business intelligence, data extraction.

1. Introduction

Due to the growth of the internet, there is an exponential increased volume of information being produced by every Internet user every minute. Due to the proliferation of data, big data analytics has been introduced, it refers to the process of acquiring, storing, and analyzing enormous datasets so as to uncover hidden patterns, insights, and relationships that can promote informed business decision (Gupta and Gosain, 2013) .

As organization amass petabytes of data from various touchpoints and systems, it is crucial to promote the importance of unstructured data. Research has shown that over 95% of the digital universe is unstructured data and 80% of all stored organizational data is unstructured (Tanwar, Duggal and Khatri, 2015). Unstructured data can be categorized into several types , including:

(i) Text Data: This includes data in the form of text, such as emails, social media posts, and customer reviews.

(ii) Image Data: This includes data in the form of images or photographs.

(iii) Audio Data: This includes data in the form of audio recordings, such as voices memos, call recording, and podcasts.

(iv) Video Data: This includes data in the form of videos or films.

(v) Social media Data: This includes data collected from social media platforms such as Facebook, Twitter, and Instagram.

Heterogeneity poses a signification challenge for big data analytics, and unstructured data is a primary contributing factor (Tanwar, Duggal and Khatri, 2015).. Unstructured information must be parsed and structured before analytical insights can be derived.

2. Big Data and Unstructured Data

Structural heterogeneity of Big Data can be segmented into structured, semi-structured and unstructured data as shown in Fig. 1(Chasupa and Paireekreng, 2021).

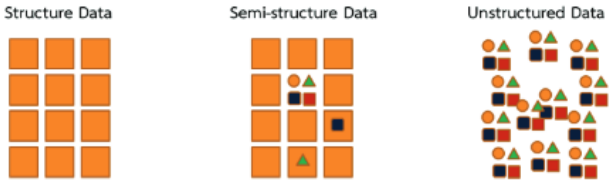

Figure 1: Structural heterogeneity of Big Data.

A. Structured Data:

Structured Data refers to data that is organized in relational database system. Instances include database tables, objects, tags, indexes and so on. It can be managed using SQL along with many Relational Database Management Systems (RDBMS). Some main characteristics of structured data include:

- Predefined schema: the data model and format are predefined so as to ensure data consistency, integrity and availability.
- Organized and categorized: Data is organized into rows, columns, tables, and relationship for ease of access, analysis, and aggregation.
- Integrated: Data from multiple sources can be integrated into a common schema to provide a unified and contiguous view of information.

B. Semi-structured data:

Semi-structured data refers to a structured information that lacks rigid data model or fails to conform to formal structure. It contains tags, markers or other identifiers to segregate semantic components and establish hierarchies of records within information (Kanimozhi and Venkatesan, 2015). Language such as XML, JSON or other markup languages are used to manage semi-structured data. Some main characteristics of semi-structured data include:

- Loose schema: The data model is loosely defined, allowing for flexibility as new data can be incorporated without rigidly enforcing a schema.
- Self-describing: Tags, markers and other identifiers are embedded in the data, therefore, the structure and meaning of data can be intuitively understood.

C. Unstructured data:

Unstructured data refers to information that lacks fixed schema or data model. It can come in the form of unformatted text files, email, videos, audio or sensors.

Unstructured data captures the nuance and depth of business operation, customer experiences, and other facets of an enterprise in a manner that structured information cannot emulate. It has also been stated that for decision making, the decision markers cannot rely solely on structured data as they would miss 95% of existing data (Gantz et al., 2007). Therefore, to gain a competitive advantage over their counterpart, organizations need to acknowledge the potential information locked in unstructured data (Wu and Lin, 2018). Some key characteristics of unstructured data include:

- Lack of schema: Unstructured data does not have a predefined data model or schema.
- Massive volume: Unstructured data is growing exponentially in size and prevalence. It makes up the majority of information in the digital world.
- Prone to redundancy and ambiguity: Unstructured data may contain repetitions or vagueness due to unstructured nature of information. This will further complicate the analyzing process and specialized approach are required to harvest its full potential.

In order to keep pace with the ever-evolving pace of information age, organizations will need to embrace the importance of unstructured data to extract patterns across complex and vague datasets. However, the most challenging in unstructured data analytic is that it tends to be noisy (Andriole, 2015) . Only after cleansing and preparation can unstructured data be effectively analyzed. Therefore, there is an imperative need for transforming unstructured data to structured data.

3. Transformation of Unstructured Data to Structured Data.

Unstructured data needs to be converted into more concrete and actionable structured format. This paper will propose two primary processes of identifying and labeling unstructured data into structured data.

3.1 Data Extraction

Analyzing and extracting unstructured data refers to identifying information from multiple sources and formats (Tanwar, Duggal and Khatri, 2015). As unstructured data comes in various representations, it is crucial to analyze formats and semantics to effectively extract meaningful information (Rao, 2003). There are two primary processes involved in extracting value from unstructured data:

- Entity extraction: Identifying entities such as names, places, products, organizations mentioned in unstructured data. This will helps obtain key details and concepts that can be used to classify, integrate and enrich unstructured information (Abdullah and Ahmad, 2015).
- Fact extraction: Understanding facts, key phrases, issues, content, relationships and other insightful information in unstructured data. This process aims to comprehend the significance, implication and role of information mentioned in unstructured data (Abdullah and Ahmad, 2015).

In order to perform the above processes, some analytical approaches and techniques can be used.

A. Text Analytics

This refers to the techniques used for extracting or retrieving information or insights from text-based data such as emails, documents, advertisement, forums, blogs, news content, and social network content (Tanwar, Duggal and Khatri, 2015). Text analytics aims to extract concepts, themes, topics, keywords, and other semantic concepts discussed in unstructured text data. Term extraction is also promoted in text analytic, important terms, technical terms, subject specific terms are extracted to develop taxonomies and metadata schemas for organizing unstructured text data. By extracting the key concepts and context of an unstructured text data, text analytics can produce targeted summaries that will reduce redundancy, highlight key point

and make large datasets more comprehensible (Hahn and Mani, 2000) .

B. Audio Analytics

This refers to the process of analyzing and extracting information from unstructured audio content or data. It is also known as Speech Analytics when applied to human conversation (Liu, 2012). It is usually implemented in areas such as customer call centers and video auto subtitles mechanism. Audio Analytics usually employs Transcript-based approach which match sounds with words from a predefined dictionary. If the system fails to identify the word, it will return the most similar word. Audio Analytic can enhance extraction of data from unstructured audio through various techniques and hence, promotes the importance of audio data to organizations.

C. Video Analytics

This refers to the process of monitoring, analyzing and gaining meaningful insights from videos. Video Analytics tracks the movement, location, trajectory, interactions, and behavioral patterns of object to determine events, identify key actions. This information can be used to create a summarization to reduces redundancy, enhance comprehension and facilitates extraction of video datasets (Tanwar, Duggal and Khatri, 2015).

Using these techniques will unlock deeper insights, reduce redundancy, and provide valuable context for Data Extraction. However, extracted unstructured data need to be classified to provide more meaningful context.

3.2 Data Classification

Data classification refers to the process of organizing information into categories and metadata labels for easier search, retrieval and management. This can be done by determining classification objectives related to the business scope and integrate metadata to the datasets (Abdullah and Ahmad, 2015). It is also important to choose a scheme structure that will promote searchability, integrability and flexibility for future process on the datasets. These enormous datasets will pose a tremendous storage challenge for any organization.

4. Storage solution for Unstructured data

The sheer volume and complex nature of unstructured datasets will be a major financial burden for any organization if left unmanaged (Samundiswary and Dongre, 2017).

Considering the advantages and disadvantages of unstructured datasets, it is considered more cost-efficient and with greater scalability and accessibility.

4.1 Cloud Computing

Cloud computing has become a significant development in the information and communication era (Tarigo Hashem, 2014). It is a scalable and cost-effective technology that offer Infrastructure-as-a-Service (IaaS), Platform-as-a-Service(PaaS), and Software-as-a-Service (SaaS). This will allow any organization to quickly and economically build system that can auto scaling their volume based on real-time changes.

4.2 Cloud Storage

Big data and cloud computing are constantly progressing (Tarigo Hashem, 2014), with unstructured data using large cluster servers and resources to manage its data. The cloud environment is a solution for processing and storing of large volumes and diversity of data. Cloud computing provides a vast pool of resources, storage, and networking, which can effectively cater to the needs of big data.

Cloud-based storage offers block, file storage, and object-based storage as well as traditional relational database for the storage of unstructured database.

Functionality	Google	Microsoft	Amazon
Large Data Storage	Google Cloud Storage	Azure Blob Storage	S3
Relational database	Cloud SQL	Azure SQL database	Relational Database Service
NoSQL	Google Cloud Bigtable	Azure Cosmos DB	DynamoDB

Table 1: Big Data Cloud Platform

Storing unstructured data requires a flexible and scalable approach that can handle large volume of data. Block storage, object storage, and file storage are all commonly used for storing unstructured data.

A. Block Storage

Block storage is a type of data storage that stores data in fixed-sized blocks, which are organized into volumes. Each block is identified as an individual hard drive with unique address, allowing the storage system to access and retrieve data quickly and efficiently (Samundiswary and Dongre, 2017). It is a flexible and scalable solution for storing structured and unstructured data.

B. File Storage

File Storage is a type of data storage that stores data as files organized into hierarchical directories and subdirectories (Samundiswary and Dongre, 2017). It is commonly used for storing unstructured data, such as documents, images and video. File storage can be implemented using various storage media that can facilitate data storage both on-premise and in cloud environment.

C. Object Storage

Object storage is a type of data storage that stores data as objects, which consist of both the data itself and metadata about the data. Each object is identified by a unique identifier (Samundiswary and Dongre, 2017). Object

storage is commonly used for storing unstructured data, such as multimedia files, archives and is ideal for distributing content or creating data lakes. Some of the well-known Object Storage including Amazon S3, Azure Blob Storage, or Google Cloud Storage.

Feature	File Storage	Object Storage	Block Storage
Data Format	Files organized in hierarchical directories and subdirectories	Objects consisting of data and metadata	Fixed-size blocks
Access Method	File-level protocols	Object-level protocols	Block-level protocols
Use Cases	File sharing, media and entertainment.	Unstructured data, static file data.	Transactional data
Scalability	Limited scalability for large data volumes	Highly scalable for large data volumes	Highly scalable for large data volumes
Performance	Good for sequential I/O and large files	Good for random I/O and small files	Good for random I/O and large files
Cost-effectiveness	Cost-effective for small to medium-sized data volumes	Cost-effective for large data volumes	Cost-effective for large data volumes
Data Retrieval	Fast for small files, slower for large files	Fast for all sizes of files	Fast for all sizes of files

Table 2: Storage Comparison

There is a wide range of storage option, however, the choice of which should be used will depend largely on the specific needs, requirements of the workload, and a careful evaluation of the cost for each service.

In essence, preparing and maintaining unstructured data can be difficult and costly, which raises doubts about whether it is worthwhile.

5. Leveraging Unstructured Data for Strategic Decision Making

The potential benefits of unstructured information, like data in text form, in audio or in video, are enormous. Organization can make significantly better decision, that are faster, more precise and grounded in more knowledge, by incorporating unstructured information in their decision making.

Emerging technologies can help organizations effectively incorporate unstructured information into their decision making in several ways:

- **Artificial Intelligence (AI) and Machine Learning (ML):** AI techniques like machine learning, natural language processing, text classification and sentiment analysis can analyze huge volumes of unstructured data to uncover insights to inform decisions (Gautam and Yadav, 2014). For example, ML models can detect trends in customer comments, identify trends mentioned on social media, or predict risks based on news report.
- **Big Data Analytics**: Integrating unstructured data sources into big data analytics platforms allows for scaled analysis and pattern detection. Tools can combine structured data, social media posts, news articles, emails, documents and more. Sophisticated analytics can then generate recommendations and highlight important information for decision makers.
- **Social Media Monitoring:** Dedicated tools and platforms for social media listening and analysis provide constant access to customer opinions, comments, feedback, experiences (positive or negative), questions, suggestions, and more—all which shapes realities and influences decisions (Aggarwal, 2011). Social data can drive important product changes, marketing adjustments, operational improvements, and strategic redirections.
- **Data analysis and mining:** Using techniques like text classification, semantic analysis, sentiment analysis, etc. to analyze large volumes of text data and extract meaningful insights and patterns. This can uncover new knowledge and perspectives to inform decisions.

6. Conclusion

Unstructured data continues to grow in volume and importance, propelling organizations to develop innovative method of managing and leveraging this valuable information asset. By embracing the importance of unstructured data, how it can be stored, and how it can be used, businesses can revolutionize how they make decisions and achieve edge advantage against their competitors.

The opportunities ahead for unstructured data are vast and largely untapped. But with continued progress in AI, ML, cloud computing and more, the potential of unstructured information will soon be fully realized, and business decision will stand to gain significantly.

7. Reference

Kanimozhi, K.V. and Venkatesan, M. (2015) *'Unstructured Data Analysis-A Survey'*, International Journal of Advanced Research in Computer and Communication Engineering, 4(3), pp. 223-227.

Tanwar, M., Duggal, R. and Khatri, S.K. (2015) *'Unravelling unstructured data: A wealth of information in big data'*, in: 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, India, pp. 1-6.

Taleb, I., Serhani, M.A. and Dssouli, R. (2018) *'Big Data Quality Assessment Model for Unstructured Data'*, in: 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, pp. 69-74.

Chasupa, T.-l. and Paireekreng, W. (2021) *'The Framework of Extracting Unstructured Usage for Big Data Platform'*, in: 2021 2nd International Conference on Big Data Analytics and Practices (IBDAP), Bangkok, Thailand, pp. 90-94.

Samundiswary, S. and Dongre, N.M. (2017) *'Object storage architecture in cloud for unstructured data'*, in: 2017 International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, pp. 1-6.

Abdullah, M.F. and Ahmad, K. (2015) *'Business intelligence model for unstructured data management'*, in: 2015 International Conference on Electrical Engineering and Informatics (ICEEI), Denpasar, Indonesia, pp. 473-477.

Gupta, V. and Gosain, A. (2013) *'A Comprehensive Review of Unstructured Data Management Approaches in Data Warehouse'*, in: 2013 International Symposium on Computational and Business Intelligence, New Delhi, India, pp. 64-67.

Kiefer, C. (2016) *'Assessing the Quality of Unstructured Data: An Initial Overview'*, Lernen, Wissen, Daten, Analysen.

Chen, M., Mao, S. and Liu, Y. (2014) *'Big Data: A Survey'*, Mobile Networks and Applications, 19(2), pp. 171-209.

Gantz, J. et al. (2007) *'The Expanding Digital Universe: A Forecast of World Wide Information Growth through 2010'*, IDC Whitepaper.

Wu, P.-J. and Lin, K.-C. (2018) *'Unstructured big data analytics for retrieving e-commerce logistics knowledge'*, Telematics and Informatics, 35(1), pp. 237-244. ISSN 0736-5853

Andriole, S. (2015) *'Unstructured Data: The Other Side of Analytics'*, Forbes. [Online]. Available at: http://www.forbes.com/sites/steveandriole/2015/03/05/the-other-side-of-analytics/ (Accessed: 6 April 2023).

Rao, R. (2003) 'From unstructured data to actionable intelligence', IT Professional, 5, pp. 29-35.

Hahn, U. and Mani, I. (2000) 'The challenges of automatic summarization', Computer, 33(11), pp. 29-36.

Liu, B. (2012) 'Sentiment analysis and opinion mining', Synthesis Lectures on Human Language Technologies, 5(1), pp. 1-167.

Tarigo Hashem, I.A. (2014) 'The Rise of Big data on cloud Computing: Review and open research issues', Elsevier.

Gautam, G. and Yadav, D. (2014) 'Sentiment analysis of twitter data using machine learning approaches and semantic analysis', in: Contemporary Computing (IC3), 2014 Seventh International Conference on, pp. 437-442.

Aggarwal, C.C. (2011) 'An introduction to social network data analytics', in: Aggarwal, C.C. (ed.) Social Network Data Analytics, Springer, pp. 1-15.