Unstructured Data: an overview of the data of Big Data

Adanma Cecilia Eberendu
Department of Computer Science, Madonna University, Nigeria

Abstract
With the emergence of new channels and technologies such as social networking, mobile computing, and online advertising, the data generated no longer have a standard format or structure like the conventional ones and cannot be processed using relational models. They come in the form of text, XML, emails, images, weblogs, videos, and so on resulting in a surge of new data types. This formless data is either semi-structured or unstructured data and makes searching and analysis complex. This paper gave an overview of this unstructured data that makes the backbone of predictive analysis. It outlined the sources and element of unstructured data and how organization benefits from gathering, analyzing and using unstructured data. The result concluded that organizations no longer neglect unstructured data nowadays; rather they are devising means of analyzing it to extract information.

Keywords: *Data, Unstructured Data, Big Data, Format, Structured Data*

Introduction

Data has been originally generated by organizational employees but recently it scaled-up to user generated and lately to machine generated given colossal amount of data that needs large storage and processing capability on a daily basis. The size of unstructured data generated by companies like those oil drilling, airlines, social networks, marketing, and others is tending to thousands of terabytes and the unstructured data is valuable that the companies are now devising method of extracting meaning from them. Due to this massive size of data and the variety of data types traditional data processing tools can no longer handle them. Data volume has grown exponentially because of the explosion of machine-generated data and from growing human engagement within the social networks. Cisco (2015) predicted that annual global IP traffic will pass the Zettabyte (1000 exabytes) threshold by the end of 2016, and will reach 2.3 Zettabytes per year by 2020. It went further to say that 2016 global IP traffic will reach 1.1 Zettabytes annually or 88.4 exabytes monthly, and 2019 global IP traffic will reach 2.0 Zettabytes annually or 168 exabytes monthly. Customer hourly transaction exceeds 1 million and data production will be 44 times greater than it was in 2009. The number of people calling, texting, tweeting, and browsing on mobile phones worldwide has already exceed 5 billion.

Big data is a term coined to address this massive volume of generated data, storage, and processing that is in the tone of gigabytes, terabytes, petabytes, exabytes, to zetabytes (Villars, Olofson & Eastwood, 2011). According to Dijcks (2013), big data typically includes traditional enterprise data, machine-generated /sensor data, digital streams, and social media. Small amount of data can turn to big data, for example trying to transfer 80MB of data through email attachment may result to big data because email attachment cannot accommodate such amount of data. Suppose there are digital images and video files amounting to 20TB to be processed in a conventional processing system within a given time frame. This is a big data because processing is difficult. Social network sites like Facebook, Google+, and LinkedIn emit volumes of data on a daily basis which cannot be processed with conventional system because the data is big data. Storing and processing becomes problematic as the number of users increases. Das and Kumar (2003) saw the Vs as characteristics of what the big data is all about. Although, they mention three (Volumes Variety, Velocity), as at today, there are eight Vs which can be used to characterized the big data. Analyzing big data with traditional data processing techniques pose performance problems due to the volume, veracity, velocity, variety, volatility, visualization and value of the data. The volume of electronic data available combined with multi-channel processes and transactions have increased dramatically in recent years.

Data has moved from stack to flow, static to dynamic. The amount of structured and unstructured data being generated and stored has exploded recently into exponential progression due to digitalization of data. The sources of both structured

and unstructured data include daily transactions, social media, sensor generated, digital images, videos, audios, and clickstreams which encompass contributions from organizations and individuals (Chen, Wang, Liu & Lin, 2009). Thus there is a need to analyze both structured and unstructured data in the day-to-day running of organization to determine customer reactions, product preference of consumers, product personalization, and other organizational requirements.

According to De Boe (2014) experts' estimation has shown that 85% of all existing data is in unstructured formats which are held in form of e-mails, contract documents, memos, clinical notes, legal briefs, social media feeds, etc. Structured data normally comprises quantitative data while important expert views and decisions are often concealed in these unstructured formats. Since these volumes of text are generated at unprecedented speed, this information cannot be made useful unless there is some process of synthesis or automation. Maluf and Tran (2008) opined that unstructured data usage is increasing enormously in real-world applications because organizations have realized the available potential that will be gained if unstructured data are analyzed and incorporated in decision making.

Growth Rate of Unstructured Data

Villars et al (2011) classified structured data as block and unstructured data as file. They found out that organizational structured data accounted for only 23.7% while unstructured was 61.8% showing a difference of nearly 70EB (exabyte or a billion billion). Therefore more data is stored in unstructured format than those stored in structured format. Sint, Schaffert, Stroka and Ferstl (2009) believed that nearly 80% of organizational data is unstructured, consisting of information from social networks, emails, customer calls and online comments, as well as diagnostic information logged by embedded and user devices, for example log files from devices. Unstructured data is gradually taken over the data format of organizations and Cisco (2015) found out that data is growing at unprecedented rate as depicted in figure 1.

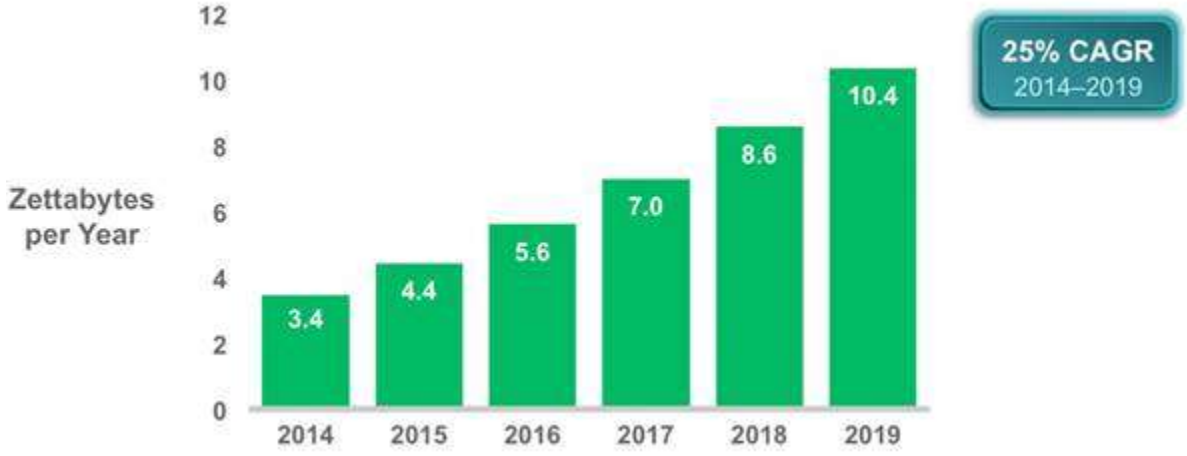

Figure 1: Annual data growth rate Source: Cisco Global Cloud Index, 2014–2019

Unstructured data cannot be sorted, searched, visualized, or analyzed in the same way as structured data; therefore new tools and processes are required to extort intelligence, share information, and deliver value (Hanig, Schierle & Trabold, 2010). Organizations spend much man-hour creating unstructured data, loosing time and money and the data does not reach the right people at the right time. Table 1 shows organizations with the rate at which they generate.

Table 1: Unstructured Data Growth Rate

Company	Generated Data
Digital Universe Study	1,227 Exabyte in 2010
	Predicted 1.8 Zettabyte data creation annually in 2011
	2.7 Zetabytes in 2013
YouTube	48 hours of new video uploaded every minute
Facebook	34,722 Likes received every minute
	100 terabytes uploaded daily
	30+ Petabytes (stores, accesses, and analyzes)
	30 Billion Pieces of content shared monthly
Domain Name	571 new websites are created every minute
Web store	More than 2.5 petabytes hourly
Twitter	Roughly 175 million tweets every day
	More than 465 million accounts
Boeing 787	40 terabytes (TB) per hour of flight
Oil drilling	Up to 2.4 TB per minute
Automated manufacturing facility	Approximately 1 TB per hour
Large retail store	Approximately 10 gigabytes (GB) per hour
Global data center IP traffic	8.6 Zettabytes annually
Data generated by IoE	400 ZB
The world	creates 2.5 quintillion bytes of data per day

Types of data

Structured data refers to data that has definite format and length, easy to store and analyze with high degree of organization. This means that the data is organized in identifiable structure to allow it response to queries to retrieve information for organizational use (Doan, Naughton, Baid, Chai, Chen, Chen & Huang, 2009). A typical example of structured data is relational database like structured query language (SQL) or Access, which contained organized numbers, dates, group of words and numbers called strings/text. Due to the database seamless structure, it is searchable with simple, straightforward search algorithms which might be by data type within the actual content. Traditional analytics focus had been on structured data in while neglecting larger amount of other types.

Semi-structured data is irregular data that may be incomplete and have a structure that changes rapidly or unpredictably but does not conform to a fixed or explicit schema. This means that it is not table-oriented as in a relational database model or sorted-graph as in object databases. According to Hanig, Schierle and Trabold (2010) semi-structured data model allows information from several sources, with related but different properties, to be fit together in one whole, for example, email, XML, Doc files.

On the contrary, unstructured data has no particular structure. Unstructured data typically includes bitmap images/objects, text, email and other data types that are not part of a database (Feldman & Sanger, 2007). Although emails are organized in a database format like in Lotus Notes and Microsoft Exchange, the body of the message is in text format without structure in any way. In other words, unstructured data comprises documents like PowerPoint used to describe company strategy, spreadsheets of lead list, emails between coworkers, and interactions of customers on social networks (Maluf & Tran, 2008). Word processing documents are another form of unstructured data, though with some formatting, the content is freeform text without any structure. Unstructured data dominates the modern business data and there is a clearer way of exploiting it. Dijcks (2013) discovered that unstructured data gained popularity from Big Data technologies exposing about 70 to 80 percent of unused data in organizations. According to Feldman, Hanover, Burghard & Schubmehl, (2012) unstructured data accounted for nearly 2.5 quintillion bytes of data per day from different sources like sensors, social media posts and digital photos showing that unstructured data is growing exponentially. Conventional data scientists will have to acquire new skills to analyze

unstructured data. Hänig et al (2010) categorized unstructured data as:

- Radar: oceanographic seismic, meteorological, and vehicular
- Static: e.g. printable files, PDF files, faxes, scanned documents
- Dynamic: this type is derived from documents that may be created, edited, reviewed, and approved by many people or groups such as white papers, procedures, policies, business documents and other office documents.
- Digital media: e.g. audio, video, animation, and images
- Communication documents: e-mail, social media contents, web document and instant messaging logs

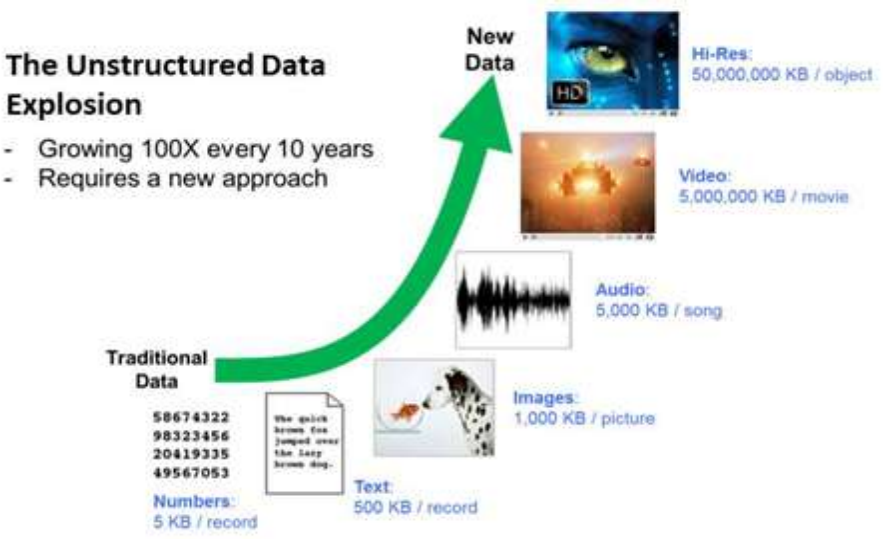

Figure 2: evolution of unstructured data
Source: www.imexresearch.com/newsletters/obs.html

Sources of Unstructured Data

Table 2: Sources of unstructured data

Source	Example
Social Media	Facebook, LinkedIn, Google+, Instagram, YouTube
Location/Geo Data	GPS, Weather, traffic
Machine-generated/Sensor	Call Detail Records, weblogs, smart meters, manufacturing sensors, equipment logs or digital exhaust, trading systems, data records
Digital Streams	Video, audio, and images
Text Documents	Email, PowerPoint, Spreadsheets, Word-processing
Logs	File Log, Clickstream
Transactions	customer information from CRM systems, web store, general ledger, transactional ERP
Micro-Blogging	Twitter, Customer feedback streams

Table 3: Elements within unstructured data

Types	Elements
Blog post	date and time it was posted, content, embedded links, author,comments
Social Media	opinions, preferences, comments, types of post, interests, needs and desires, date and time of posts, number of posts, mentions, fans, followers, views, likes, +1s, check-ins, pins
Geo Data	Date and time, initial data, real-time recording, location, number of flow,

Conclusion

Gathering and analyzing unstructured data gives organizations insight into their businesses and help them to increase competitive edge, enhance productivity, and create innovations. For instance, in policing, Nigerian Police can use information from social media like Facebook, twitter, text messages from mobile phones, calls from citizens to combat crimes. This will reduce the influx of people at the police stations and give quick and on the spot information. Oil companies install sensors in their rigs to return a stream of telemetry which reveals usage patterns, failure rates and other opportunities to reduce development costs. Supermarkets or online marketers can keep logs of who patronizes them by using social media and web log files from their ecommerce sites and even understand the reason for those who refused to patronize them. Information derived from this analysis will help organizations to re-strategize in order to increase their market share. Data from social network site like Facebook, Instagram, and LinkedIn would have been meaningless without unstructured data. Executives can get relevant information for decision making in less time with unstructured data analysis. Unstructured data has created room for fraudulent analysis, loyalty programmes that identifies and targets the consumers and customer segmentation based on stored behavior analysis.

References

1. Chen, Y., Wang, W., Liu, Z., & Lin, X. (2009, June). Keyword search on structured and semi-structured data. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data* (pp. 1005-1010). ACM.
2. Cisco (May, 2014). The Zettabyte Era—Trends and Analysis Cisco White Paper
3. Dalvi, N., Machanavajjhala, A., & Pang, B. (2012). An analysis of structured data on the web. *Proceedings of the VLDB Endowment*, *5*(7), 680-691.
4. Das, T., & Kumar, P. (2013). BIG Data Analytics: A Framework for Unstructured Data Analysis. *International Journal of Engineering and Technology (IJET) , 5* (1).
5. De Boe, B. (2014). Use Cases for Unstructured Data - Intersystems White Paper, InterSystems Corporation.
6. Dijcks, J.-P. (2013). *Oracle: Big Data for the Enterprise.* An Oracle White Paper Oracle Corporation
7. Doan, A., Naughton, J., Baid, A., Chai, X., Chen, F., Chen, T., ... & Huang, J. (2009). The case for a structured approach to managing unstructured data.*arXiv preprint arXiv:0909.1783*.
8. Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press
9. Feldman, S., Hanover, J., Burghard, C., & Schubmehl, D. (2012). Unlocking the Power of Unstructured Data IDC HEALTH INSIGH
10. Hänig, C., Schierle, M., & Trabold, D. (2010). Comparison of structured vs. unstructured data for industrial quality analysis. In *Proceedings of The World Congress on Engineering and Computer Science*.
11. Maluf, D. A., & Tran, P. B. (2008, March). Managing Unstructured Data with Structured Legacy Systems. In *Aerospace Conference, 2008 IEEE* (pp. 1-5). IEEE.
12. Rao, R. (2003). From unstructured data to actionable intelligence. *IT professional*, *5*(6), 29-35.
13. Sint, R., Schaffert, S., Stroka, S., & Ferstl, R. (2009, June). Combining unstructured, fully structured and semi-structured information in semantic wikis. In *Fourth Workshop on Semantic Wikis–The Semantic Wiki Web 6 th European Semantic Web Conference Hersonissos, Crete, Greece, June 2009* (p. 73).
14. Villars, R. L., Olofson, C. W., & Eastwood, M. (2011). Big data: What it is and why you should care. *White Paper, IDC*.