21st International Research/Expert Conference
"Trends in the Development of Machinery and Associated Technology"
TMT 2018, Karlovy Vary, Czech Republic, 18th – 22nd September, 2018

LIGHT UP THE VALUE OF DARK DATA

Dražena Gašpar
Faculty of Economics, University of Mostar
Matice Hrvatske bb, Mostar
Bosnia and Herzegovina

Mirela Mabić
Faculty of Economics, University of Mostar
Matice Hrvatske bb, Mostar
Bosnia and Herzegovina

ABSTRACT
The paper analyses the potential value of dark data in organizations. Dark data refers to unused or unidentified data acquired during different business activities, but not used for any other purposes like business analytics or decision-making. Information technology enables today's organizations to collect and store enormous amounts of data on a daily basis. However, a gap exists between the ability to acquire and the ability to analyze stored data, at the expense of data analyzing. Dark data is widespread in most organizations, and it keeps busy valuable storage capacities producing more expenses. At first sight, dark data may seem irrelevant, but it represents huge opportunities that organization should not ignore. If the organization does not address dark data, it can lead to huge intelligence risk because of possibility to lose data on business plans, products, customers, financial status, market share, etc. Analyzing dark data is one of the ways to gain deep insight and uncover new opportunities. Different big data technologies like advanced content management, big data analytics, search systems, etc. are capable of the dark data analysis. The main objective of the paper is to identify how different big data technologies can be used to light up the value of dark data.
Keywords: dark data, Big Data, Big Data technologies, business analysis

1. INTRODUCTION
The modern world is based on technology-driven prosperity, both in business and private human life. The quantity of data stored in a digital format is rapidly expanding making data extremely important competitive currency. Main drivers of this data growth are the transition from analog to digital technologies in almost every aspect of human life; the increase of computer power and data storage capabilities; and the rapid growth of social media users, as well as its generated data. According to SINTEF [1], 90% of the world's data has been generated in the past two years. An enormous quantity of data is available on the Internet. The growing number of Internet users produce their data (Figure 1) transforming today's world into the digital universe. Estimates [2] show that the digital universe will grow 40% per year throughout the next decade. That includes an increase in the number of online users (both people and organizations), as well as things, i.e., smart devices connected to the Internet called the Internet of things (IoT). That will unleash a new wave of individual and organizational opportunities around the world. The expectation is that the quantity of data created and copied annually will reach 44 zettabytes or 44 trillion gigabytes by 2020 [2]. The term Big Data is coined to explain extensive requests for the storage and management of complex, dynamic, evolving, distributed, and heterogeneous data from different sources and platforms. Unstructured data makes Big Data issue more complicated. Today, 80% of data generated and/or gathered by organizations is unstructured. That means that data is highly dynamic and without a particular format. Data may exist in the form of e-mail attachments, images, pdf documents, medical records, x-rays, voicemail, graphics, video, audio, etc. It cannot be easily stored in row/column format as structured data.

Transforming this data to a format suitable for later analysis is a challenge in Big Data analysis. It is a driving force for adopting new technologies to deal with such data [3].

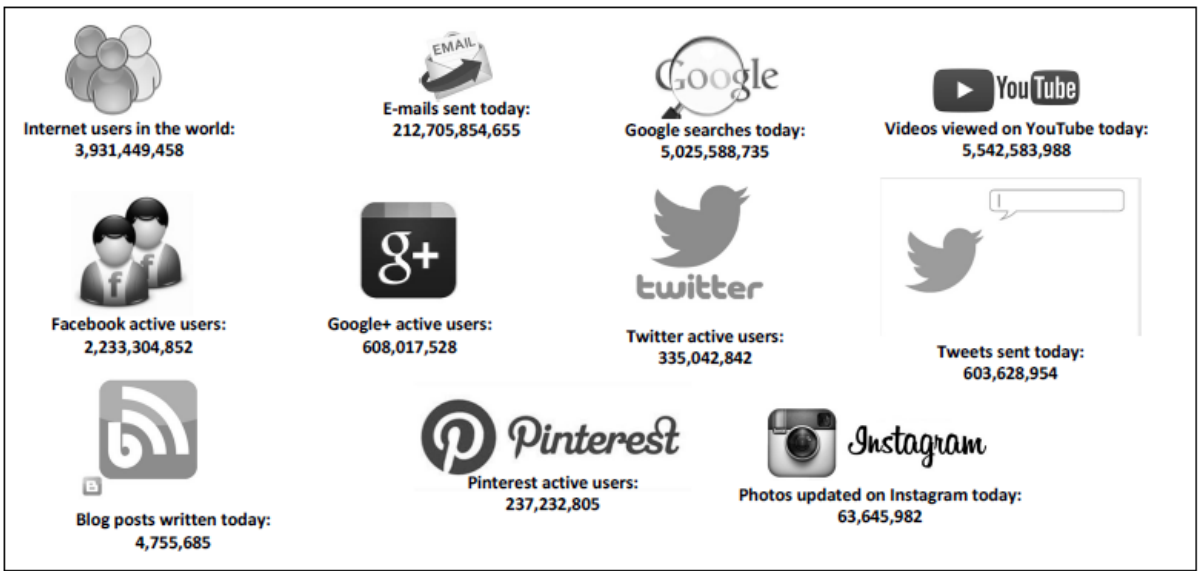

Figure 1 Internet data generated on a daily basis [4]

Therefore, the job of Chief Information Officers (CIOs) is today more complex than ever before. CIOs are faced with the necessity to integrate distributed data architecture, in-memory processing, machine learning, visualization, natural language processing, and cognitive analytics to answer questions and identify valuable patterns and insights that would have value for business users. However, dark data and "dark analytics" are becoming a huge challenge, both for technology and business development. Dark analytics is focused to the exploration of the huge universe of unstructured and "dark" data with the goal of unearthing the highly nuanced business, customer, and operational insights that structured data assets currently in their possession may not reveal [5]. It is evident that today's companies analyses only a small fraction of the overall digital universe. IDC estimates that by 2020, as much as 37 percent of the digital universe will contain information that might be valuable if analyzed [6]. That has been an open new research area for the dark data and dark data analytics, which is in focus of this paper. The paper explains dark data and their characteristics with the aim to identify how different big data technologies can be used to light up the value of dark data.

2. DARK DATA CHARACTERISTICS

One of the most cited definitions of dark data is Gartner's definition according to which "dark data is the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes" [7]. Dark data can be described as "unused or unidentified content that sits outside of the retention schedules, classification schemes and retrieval systems that organizations rely upon to meet compliance obligations, ease the burden of electronic discovery and ensure decisions are made with accurate and relevant information" [8].

Concerning business data, term dark data describes something that is hidden or undigested, mostly unstructured data, which may include things such as text messages, documents, call notes and meeting minutes, presentations, email, video and audio files, and still images [5].

The main characteristics resulting from the definition of dark data are:

- Unmanaged
- Uncategorized
- Unstructured
- Unknown/Hidden.

Unmanaged means that dark data is out of the control of Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), Business Intelligence (BI) tools or other data governance programs. It lay unattended and unprotected on old storage disks, keys or tapes [8]. The

organizations should be aware of the existence of dark data and put additional efforts to include it in IT infrastructure, enabling its searching and analyzing.
Uncategorized refers to the fact that dark data is not included in any classification schemes that exist in the organization. Data categorization is one of the first steps towards efficient management and use of data.
Dark data is typically unstructured meaning that it does not have pre-defined, i.e., recognizable structure. Unstructured data is textual data, video data, email data, audio data, social media data and so on.
Unknown/hidden means that data are stored on some old archive devices usually do not check or used for decades. However, in case that organization needs that data the costs of finding them could be enormous. Also, unknown/hidden can be any data buried in the deep web. Most companies do not browse or even can find deep web data. The reason for that lies in the fact that deep web data is not indexed by typical search engines (Google, Yahoo). The deep web can include data from academia, government agencies, user communities, as well as data on the dark web which consists of sites only accessible through special tools and that are largely untraceable [9]. It is impossible to calculate the deep web's size accurately, but by some estimates, it is 500 times larger than the surface web that most people search daily [10].
Since today data is valuable currency, these vast unexplored data resources could prove to be something like pure gold for organizations that succeed in its exploration. Today's organizations are forced to turn the huge volumes of dark data into manageable, categorized, usable and valuable data by using different data analytics tools.

3. LIGHTING UP DARK DATA

Even today, when analytic data potential is in question, most companies are still focused or limited to the structured data existing within different systems in the organization. Term dark analytics is used to stress that dark data also should be in focus of data analytics. Dark data analytics is focused on three dimensions [5]:
- Untapped data is already existing in the organization, especially data outside organizational IT systems (for example emails, notes, messages, logs, notifications from the Internet of Things (IoT) devices, etc.). This data mostly can be mined through traditional analytics tools.
- Nontraditional unstructured data (for example audio and video files, images, etc.). In order to make this data valuable organizations need to use advanced technology like computer vision, advanced pattern recognition, video and sound analytics. The ability to apply analytics to audio and video feeds in real time opens up profound new opportunities for signal detection and response [5].
- Data in the deep web is a huge challenge for analytics tools, both traditional and dark. For now, the solution is to bound and define the target in order to mine data in deep web. It is expecting that very soon business will be able to curate competitive intelligence using a variety of emerging search tools in order to target scientific research, activists data, as well as other interesting data from the deep web [5].

In order to light up the value of dark data, organizations should develop and implement the more formalized approach. Ryan [11] proposed four stages of lighting up data in an organization:

1. Identification – means that organization has to find out what data has and where it is stored.
2. Classification of data – stress the necessity of data classification and organization in order to reflect the fundamental organizational needs or processes within the organization.
3. Controls – ensures data managing and organizing to ensure its security and integrity. In that stage, data analysis is introduced to enable the organization to gain insights into the data for business intelligence.
4. Continuous monitoring has to ensure that proper procedures are put in place to provide that the data is continuously maintained to serve the needs of the organization. That stage should ensure return on investments (ROI) and identify areas of improvements and open issues.

Following this four stages organization could provide a formalized approach to illuminating dark data and making that data visible for searching, mining, and reporting, with final aim to ensure better decision-making process in the organization.

4. CONCLUSION

The driving force behind lighting up dark data is not just to help organizations to deal with dark data issues, but to empower them to efficiently and effectively respond to today's and future business challenges. Emerging technologies related to dark data analysis can bring the changes and lead to the new business solutions only if organizations become fully aware of the value of data they collect and find the ways to use that data to enhanced everyday business. It means that organizations must empower their employees by right data because in the today's business environment the employees are becoming the front line decision makers. To be successful in their regular activities employees need full insights about customers and they need to make right decisions and offer to customers what they want. Traditional and dark data analysis should ensure appropriate internal and external, structured and unstructured, both data and dark data in gaining valuable insights. That is still the main driving force in data analysis development because the main task of data analysis tools should be to ensure analysis of all data that business users need to be successful in performing their tasks.

5. REFERENCES

[1] SINTEF Big Data, for better or worse: 90% of world's data generated over last two years. ScienceDaily. 2013, May 22. Retrieved November 1, 2016. from www.sciencedaily.com/releases/2013/05/ 130522085217.htm

[2] IDC The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, 2014. Retrieved November 6, 2016. from http://www.emc.com/leadership/digital-universe/2014iview / executive-summary.htm

[3] Gašpar D, Ćorić I.: Bridging Relational and NoSQL Databases, IGI Global, USA, 2018.

[4] Internetlivestats: Internet live stats. Retrieved May 29, 2018, from http://www.internetlivestats.com/

[5] Kambies T, Mittal N., Roma P, Sharma S.K.: Dark Analytics – Illuminating opportunities hidden within unstructured data, Deloitte University Press, 2017. Retrieved April 2, 2018, from https://www2.deloitte.com/insights/us/en/focus/tech-trends/2017/dark-data-analyzing-unstructured-data.html

[6] Vesset D., Schubmehl D.: IDC FutureScape: Worldwide big data, business analytics, and cognitive software Predictions, International Data Corporation, December 2016. Retrieved May 11, 2018, from https://www.idc.com/getdoc.jsp?containerId=US41995816

[7] Gartner: Dark data, IT Glossary, 2018. Retrieved April 05, 2018 from https://www.gartner.com/it-glossary/ dark-data/

[8] Viewpointe: Dark Data, Big Data, Your Data: Creating an Action Plan for Information Governance, Viewpointe Archive Service, 2013. Retrieved April 05, 2018 from https://www.viewpointe.com /uploadedFiles/Viewpointe/PDFs/viewpointe-dark-data-white-paper.pdf

[9] McNulty K.: Expanding the Definition of Dark Data and Mining It with Dark Analytics, Prowess, June 30, 2017. Retrieved April 09, 2018 from http://www.prowesscorp.com/expanding-the-definition-of-dark-data-and-mining-it-with-dark-analytics/

[10] Goodman M.: Most of the web is invisible to Google. Here's what it contains, *Popular Science*, April 1, 2015, Retrieved May 10, 2018, from www.popsci.com/dark-web-revealed.

[11] Ryan S.: Illuminating Dark Dana, 2014. Retrieved May 10, 2018, from https://conferences.heanet.ie /2013/files/65/Lightning%20Talk%20-%20Shane%20Ryan%20-%20Illuminating%20Dark%20Data.pdf