Using a Roadmap in the Back Alleys of Dark Data

Vicky Seehusen
Metropolitan State University of Denver

Edgar Maldonado
Metropolitan State University of Denver

ABSTRACT:

Businesses and individuals continue to create massive amounts of data, and this exploding data volume is often overlooked by most business leaders. Veritas, an enterprise data protection company, predicts that by 2020, organizations around the world are on course to waste more than 3.3 trillion dollars storing data that is considered redundant, obsolete and trivial (ROT data) (Veritas, 2016). Most of the storage data has not been carefully indexed, so its value has not yet been identified, the term for this kind of information is Dark Data. This paper classifies the concerns about maintaining Dark Data according to current literature. The paper proposes a framework to study the Dark Data issue; some Dark Data might be useful for business decision making and predictive analysis. Non-useful data should be destroyed, freeing up corporate computing and storage assets. Managing, indexing and deleting non-useful Dark Data is an understudied topic in the academic world and thus solutions to the problems of Dark Data would benefit from additional research.

Keywords: Dark data, ROT data, data management, securing data, deleting data

Copyright statement: Authors retain the copyright to the manuscripts published in AABRI journals. Please see the AABRI Copyright Policy at http://www.aabri.com/copyright.html

INTRODUCTION

Since the proliferation of computer use in business, industry and personal lives, the creation of data has continued to grow. Society generates more than 2.5 quintillion bytes of data every day. Even more, by 2020 each person will create 1.7 MB of data every second (Domo, 2018). Unfortunately, a large amount of this data is not used by organizations to make business decisions.

Dark data was listed as one of the five hottest topics for information management in 2018 (Ismail, 2016). Ismail states that there are many reasons why companies refuse to delete or destroy data that is not considered useful; companies fear to delete something that might be important or they fear they might need this data "some" day. Therefore, data remains stored on the organization's network(s) but is ignored and thus "dark".

Unused data falls into two categories. The first category is Dark Data. Gartner, a global research company, originally coined the term "Dark Data", and defines it as "the information assets organizations collect, process, and store during regular business activities, but generally fail to use for other purposes." (Gartner, 2019). Kambie et al. states "in the context of business data, "dark" describes something that is hidden or undigested" (Kambie, Mittal, Roma, & Shamar, 2017, p. 35). In addition to the information storage in private servers, a big portion of the web is considered to be dark. There is a set of sites that have not been indexed (they are partially inaccessible), known as the deep web, estimated to be 500 times larger than the portion of the web used by most search engines (Goodman, 2015).

The second type of unused data is data that is redundant, obsolete and trivial (ROT). ROT data is defined as "digital documentation that an organization continues to retain even though the information that is documented has no business or legal value" (Rouse, 2016). ROT data is created when employees save multiple copies of the same information, or they save outdated information or any information that does not help an organization meet its goals. Although the costs associated with keeping this data are hard to come by, a company handling 500 terabytes of data (typical midsize company) wastes close to $1.5 million maintaining trivial files (photos, personal ID documents, music, videos, etc.) (Veritas, 2016).

In fact, Dark Data and ROT data often exceeds business-critical data (Veritas, 2016). Veritas, a backup and recovery company, coined the term "databerg" to classify data that is not business critical. In March 2016, Veritas released the 'Global Databerg Report'. This report is a culmination of survey data obtained from 2500 IT professionals in 22 countries (Veritas, 2016). The results of this survey indicate that 52% of the information organizations around the world currently store and process is considered Dark Data and that another 33% is considered ROT data. This means that only an average of 15% of the data that organizations store is considered mission critical. And the average costs or individual organizations to store non-critical information is estimated to be $650,000 annually. By 2020, this world-wide cumulative cost is estimated to be $3.3 trillion (Veritas, 2016).

STATEMENT OF THE PROBLEM

The current level of knowledge and understanding of information that is unused in organizations is as obscure as the name given to this type of data. The authors sought to explore the following questions using the available literature.

1. Definition: What is Dark Data?
2. Issues: What kind of negative repercussions bring Dark Data mismanagement?
3. Solutions: What kind of solutions to avoid/tap Dark Data are being developed/offered?

LITERATURE SURVEY

The literature survey in this work was done using two main keywords:

- Dark Data
- ROT Data

The authors used the following Databases:

- ACM Digital Library
- Compendex
- Computer Database
- Library, Information Science & Technology Abstracts (LISTA)
- Google Scholar

The search was limited to those articles written in the last 10 years. A review of academic journals and literature yielded very little about the topic. Nevertheless, the topic is not being ignored by business and industry, and the authors found articles in online magazines and business journals that are concerned with the "explosion" of Dark Data.

The total of articles found for this study is 24. Ten of those articles were published in academic venues/procedures, and the rest (14) comes from companies' studies or/and consultants. Additional sources included opinion-based blogs, forums, etc.

DISCUSSION

Dark Data: Definition

There is no one true definition of Dark Data, however in reviewing the literature, the authors found some common themes. Gartner originally coined the term and it defined Dark Data as information assets collected, processed and stored during regular business activities but generally unused for other purposes. (Austin, 2014). It can further be described as unstructured content. In fact, unstructured content makes up 90% of all the information used every day. And it is growing at a rate 3 times faster than structured data (Datskovsky, 2013). In another, simpler definition, Dark Data is any data that is available and stored but not being used. (Trajanov, Zdraveski, Stojanov, & Kocarev, 2018)

Given the vast array of Dark Data, the authors sought to categorize it as a first step to gaining some level of control over it. The result was two categories: data organizations know exists but might be difficult to access, query or extract answers from, even when it knows what it wants to query; and data the organization doesn't know exists or has forgotten about, such as old employee files, redundant data, and deep web data. Sometimes the lines become blurred between these two categories as evidenced below. The categorization is meant to be a starting point for rooting out the Dark Data within the organization.

Data the organization knows exists but is difficult to retrieve or use

As previously mentioned, there may be data that the organization knows exists but is hard to access. This could include paper files that have never been translated into electronic format, or data that is stored electronically but is in a format that is not conducive to extracting useful information for decision making or planning.

Some data is not even human generated. For example, there might be self-monitoring data collected by manufacturing equipment, much of which might be largely ignored if it has no impact on the day-to-day operations. This is the kind of information that may not be recognized as having organizational value. On occasion, even when an organization does know about the data, it does not know what questions to ask. Sometimes, an organization knows what questions it wants to ask but cannot extract the answers.

Sometimes, however, organizations do not know what data has business value. Some articles recognized there are challenges to deriving value and processing data before it goes dark. This processed data could then be used to create more complete pictures of the customers, vendors and even employees which whom organizations interact. The problems of deriving value become evident when an organization attempts to maintain unstructured and semi-structured data that lacks data tags, metadata, or other schema (Gutierrez, 2015). Gutierrez states "once this data gets into Hadoop or another big data store, it is difficult and costly in terms of both time and IT resources to understand its value without extensive analysis".

Data the organization has forgotten about, never considered, or does not know exists

Organizations store a great deal of legacy data. Newer employees and information technology professionals may have no idea this data is still residing on its systems. Because much of this data is unstructured and not used in day to day transactions, it "hides" in the background. Austin provides some specific examples of unstructured data. They include: customer information, log files, account information, data on ex-employees, financial statements, unprocessed survey data, email, notes or presentations, and old versions of relevant documents (Austin, 2014).

Organizations also store redundant data. This occurs when employees back up data to personal hard drives and memory sticks or copy the same documents to multiple folders or send copies to others in the organization who also store the same materials. As far as the organization is concerned, this data is "hidden" and these employees have created Dark Data (Grimm, 2018).

Another type of hidden data occurs when individuals don't "share". In his research of the scientific academic community, Hiedorn found that when asked, almost all scientists admitted they were holding on to data that had never been published or otherwise made available to the scientific community. Much of this data could be found in desk drawer that was not available to others until the scientist retired. It might include slides, photos, specimens and electronic media files (Heidorn, 2008). Presumably, electronic media files and other kinds or data to support a scientist's research were stored on disks or in the scientist's private electronic account. The organization cannot assign value to or create a tool for analyzing this hidden data. A different method of "not-sharing" can also result in Dark Data; Austin states that organizational silos may be responsible for creating data that goes dark such as when one department does not share data relevant to another department (Austin, 2014).

While reviewing articles with an emphasis on ROT or Dark Data, the authors found indications that companies have largely ignored information that has no business or legal value but is being saved by end-users on organizational devices such as desktop PC's and laptops, mobile devices, network servers and corporate cloud resources (Rouse, 2016).

Lastly, sometimes, data goes intentionally dark because the time and costs constraints placed on the organization are such that high priority data is analyzed while other data is ignored. (Grimm, 2018)

DARK DATA: ISSUES

The concealed nature of Dark Data makes it difficult to provide specifics on the issues that may surface when data is mismanaged or ignored. Nevertheless, the authors argue there are two main categories of issues that are obvious: Economic and Legal.

Economic

The low cost of digital storing solutions seems to be used as an excuse to hoard data the organization does not need. This leads to a philosophy that can be summed up, as "everything must be saved, since it will be used eventually." However, believing that storage costs will continue to go down and thus not pose a huge financial impact on the organization does not take into account recurrent costs such as energy and maintenance (Tallon, 2013). By 2020, ROT data – a form of Dark Data - could cost corporations close to $900 billion (Veritas, 2016).

Veritas observed that companies often adopt cloud applications under the misleading premise of 'almost free' storage. That belief goes along with the idea of accumulating data without paying attention to its business value and using an organization's infrastructure for personal use (Veritas, 2016). Additionally, organizations continue to procure IT infrastructure and information solutions to manage their growing data stores. Therefore, as the size of data goes up, companies need more technology investments to manage it (Tallon, 2013).

The other economic aspect of Dark Data comes from potentially missing business opportunities. Hasan states how Dark Data Analytics could provide opportunities in the Health Sector and retail market (Hasan, 2018). For example, in the Health Sector, health organizations have data points from patients that could be used to identify high-risk patients. In the retail sector, companies could provide new and better services/products given data points that usually are currently left untouched.

Legal

There is also a fear that deleting data might have legal ramifications. Beyond the costs organizations incur to store and maintain this data, all this data can hinder an employee's ability to comply with regulatory guidelines. Additionally, time spent sifting through inconsequential data means employees cannot quickly respond to information requests or make quick, data-driven decisions. ROT data is also vulnerable to security breaches and poses a liability risk because it could be used against an organization in audits or legal actions. (Rouse, 2016)

Dark Data: Solutions

Upon review of the literature, the authors derived two approaches to tackle the Dark Data topic: a policy approach and a computational approach. The first one will help to prevent the growth of Dark Data, and the second approach will assist in gaining insight from currently stored Dark Data.

Policy approach

The authors suggest that Dark Data management requires organizations develop policies to clean and purge data and organizational databases. Austin, suggests organizations complete regular audits of organizations' database and get rid of old, unneeded data (Austin, 2014). In that same vein, Datskovsky advises organizations determine if any piece of data qualify as a "business record", before cleaning a database (Datskovsky, 2013). On the other hand, Veritas does not mention the idea of deleting data (Veritas, 2016). The company's report advises the use of advanced visualization tools, make knowledge discovery a business-driven effort, hiring data scientists and asking the right questions; all these approaches focusing on gaining value from Dark Data.

The authors found very little currently written about creating policies to delete or eliminate data that has little or no significance to organizational goals or profitability. This absence of literature could be due, in large part to the emphasis on "Big Data" today. Additionally, as was stated before, the costs to purchase storage have decreased significantly in the past two decades. These two items to create complacency about what data organizations keep.

The authors are proponents of disposing of data when possible, instead hoarding it without limit. A simple three step process to do so is presented in the conclusion section.

Computational approach

There have been some work on the computational area to automatic identified and classify Dark Data. Data classification software, when in place, "automatically identifies, classifies, and tracks sensitive data from the moment it is created, modified, or transmitted. " (Data Classification, 2019). Some authors have taken advantage of their expertise on specific domains and used machine learning techniques to explore Dark Data. Examples of this specific domain Dark Data analytics research includes seismology (Wang, et al., 2019), chemical industry (Ambrayan & Pemmaraju, 2015), internet of things (Trajanov, Zdraveski, Stojanov, & Kocarev, 2018), and environmental science (Haddaway, Collins, Coughlin, & Kohl, 2017). In the case of specific domain studies, researches use Dark Data they have pre-categorized and use their expertise and algorithms to answer pertinent questions.

On the other hand, some research seems to focus on automating the discovery of structure from unstructured data. A good example of this kind of research is DeepDive (Zhang, Shin, Cafarella, & Niu, 2016). Zhang et al. developed a system that crawls through unstructured data and using statistical inference and machine learning algorithms, can provide labels and structure to the given data. The authors have called this kind of solution a 'General Domain' solution. This

approach can be applied to any domain to gain insight from data that has not been categorized at all.

CONCLUSIONS AND FUTURE RESEARCH

The three aspects of Dark Data described in prior section have been incorporated in Figure 1 (Appendix 1). This roadmap can help researchers and practitioners in the development of strategies to deal with the ROT Dark Data challenges their organizations might be facing.

Organizations should be encouraged to develop policies for what data should be saved on corporate resources and for how long. These policies will vary by industry and organization type. For example, government entities may be required to save pictures of land and construction projects longer than a company who develops on the land.

This could be an arduous task even if the organization is a small one; different employees and/or departments are going to come to different conclusions about what is important data now and in the future. This could result in a type of stand-off that results in IT employees being unable to automatically delete anything beyond the emails of a recently separated employee. Separate servers for marketing and transactional data are a way to begin to consciously think about the data organizational employees are keeping and why they are keeping it.

The authors suggest institutions must include guidelines for data deletion in their information management policies. The creation and implementation of these policies need to be addressed at the departmental level, since each unit is presumed to have the expertise to decide on the value of its own data. Prior to deciding what data is not valuable, units must classify all data generated and stored.

Reducing ROT and other Dark Data may be an arduous process for most organizations, however, there are auto-classification technologies in some content management systems that use algorithms to auto scan a document's contents and assign a records management classification code to it. (Holak, 2014) There are also predictive coding techniques that use keyword searches, filtering and sampling to reduce the number of irrelevant and non-responsive documents that would require manual reviewing. (Cole, 2015)

Before any of these things happen however, Rouse suggests the creation of an information governance plan which would include provisions for how to deal with ROT and data in an ongoing, proactive manner, create a company culture that encourages active information management and discourages data hoarding. (Rouse, 2016)

After analyzing the literature referred to in this paper, the authors propose a three-step strategy for information governance to minimize dark data within the organization. The steps are listed below:

1. Begin at the Departmental Level to establish a policy to create a data labeling schematic that stamps every piece of data that is saved in organizational resources along with pre-defined meta data codes. This step will create a "Data Dictionary" at the Department Level.
2. Integrate Department Level Dictionaries to create an Organizational Level dictionary.
3. Establish committees to determine the life span of the stored data (at departmental and organizational levels).

Finally, the Dark Data problem is a topic that deserves additional study and research within the business and academic communities. The authors of this paper have provided a roadmap framework that can be used as a starting point for communities.

BIBLIOGRAPHY

Ambrayan, B., & Pemmaraju, G. (2015). Marketing and Product Development Strategies in the Chemical Industry Using Dark Data & Data Sciences. *SPE ANTEC*, (pp. 1930-1933). Orlando.

Austin, B. (2014, December 8). *Dark Data: What is it and Why Should I Care*. Retrieved from R1soft: https://www.r1soft.com/blog/dark-data-what-is-it-and-why-should-i-care

Cole, B. (2015, September). *Predictive Coding.* Retrieved from Tech Target: https://searchcompliance.techtarget.com/definition/predictive-coding

Data Classification. (2019). Retrieved from Digital Guardian: https://digitalguardian.com/resources/data-security-knowledge-base/data-classification

Datskovsky, G. (2013, March/April). Harnessing Big Data for Competitive Advantage. *Information Management*, p. S6.

Domo. (2018, June 5). *Data Never Sleeps 6.0.* Domo. Retrieved from Domo: https://web-assets.domo.com/blog/wp-content/uploads/2018/05/18_domo_data-never-sleeps-6verticals.pdf

Gartner. (2019, July 1). *IT Glossary.* Retrieved from Gartner: https://www.gartner.com/it-glossary/dark-data

Goodman, M. (2015, April). Most of the web is Invisible to Google. Here's what it contains. *Popular Science*. Retrieved from www.posci.com/dark-web-revealed

Grimm, D. J. (2018). The Dark Data Quandary. *Am. UL Rev., 68*, 761.

Gutierrez, D. (2015, October 12). *Deriving Value from Data Before It Goes Dark*. Retrieved from Inside Big Data: https://insidebigdata.com/2015/10/12/deriving-value-from-data-before-it-goes-dark/

Haddaway, N., Collins, A., Coughlin, D., & Kohl, C. (2017). Including Non-Public Data and Studies in Systematic Reviews and Systemic Maps. *Environmental INternational, 99*, 351-355.

Hasan, A. (2018). Dark Data for Analytics. In *Data Analytics* (pp. 299-318). CRC Press.

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the long Tail of Science. *Library Trends*, 280-299.

Holak, B. (2014, November). *Auto Classification.* Retrieved from TechTarget: https://searchcompliance.techtarget.com/definition/autoclassification

Ismail, N. (2016, December 19). *5 Hot Topics for Information Management in 2018*. Retrieved from Information Age: https://www.information-age.com/5-hot-topics-information-management-2018-123463700/

Kambie, T., Mittal, N., Roma, P., & Shamar, S. K. (2017). Dark analytics: Illuminating. In *Deloitte Insights Tech Trends 2017* (pp. 21-33). Deloitte.

Rouse, M. (2016, September). *ROT (redundant, outdated, trivial information).* Retrieved from TechTarget: https://whatis.techtarget.com/definition/ROT-redundant-outdated-trivial-information

Tallon, P. P. (2013). Corporate Governance of Big Data: Perspectives on Value, Risks and Cost. *Computer*, 32-38.

Trajanov, D., Zdraveski, V., Stojanov, R., & Kocarev, L. (2018). Dark Data in Internet of Things (IoT): Challenges and Opportunities. *Proceedings of the 7th Small Systems Simulation Symposium.* Serbia.

Veritas. (2016, March 15). *The Databerg Report: See What Others Don't; Identify the Value, Risk and Cost of Your Data.* Veritas. Retrieved from Storage Review: https://www.storagereview.com/veritas_releases_us_databerg_report

Wang, K., Ellsworth, W., Beroza, G., Williams, G., Zhang, M., Schroeder, D., & Rubinstein, J. (2019). Seismology with Dark Data: Image-Based Processing of Analog Records Using Machine Learning for the Rangely Earthquake Control Experiment. *Seismological Research Letters, 90*(2A), 553-562.

Zhang, C., Shin, J., Cafarella, M., & Niu, F. (2016). Extracting Databases from Dark Data with DeepDive. *SIGMOD*, (pp. 847-859). San Francisco.

APPENDIX

Figure 1: