International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 4, June 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Extract the Analyzed Information from Dark Data (Data which Lies Beneath the Surface of an Iceberg)

Rahul P[1], Ganeshan M[2]

[1]Master of Computer Application,
[2]Department of Computer Science and Information Technology,
[1,2]Jain University, JGI, Bengaluru, Karnataka, India

ABSTRACT

The world is surrounded by data and data, the data may be structured, unstructured, or semi-structured; every organization generates enormous data daily, only the tip of data is analyzed, and the larger the data is ignored from the utilizable analysis. This paper focuses on a particularly unstructured and bothersome class of data, termed Dark data. Dark data is not attentively analyzed, indexed, and stored, so it becomes nearly imperceptible to potential users and therefore is more likely to last neutralized and eventually lost. This paper discusses how the concepts of long-term specifically use of analyzed for all intents and purposes dark data can be used to generally understand the very possible solutions for better curation of dark data in a major way. This paper describes why this class of data is so critical to scientific progress, some of the properties of this dark data, as well as the technical difficulties to useful management of this class of data. Many probable useful institutional and technical solutions are under development which will show in this paper in the last section, but these solutions are mainly conceptual and require additional research during lack of resources.

KEYWORDS: *Dark Data; Big Data Analytics, Structured Data, Unstructured Data, and Semi-Structure Data*

How to cite this paper: Rahul P | Ganeshan M "Extract the Analyzed Information from Dark Data" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-4, June 2020, pp.26-29, URL: www.ijtsrd.com/papers/ijtsrd30842.pdf

IJTSRD30842

I. INTRODUCTION

Dark data is the digital information that is not being utilized. Consulting and market research company Gartner Inc. describes dark data as "information assets that an organization amasses, processes and stores in the course of its conventional business activity, but generally fails to utilize for other purposes."

Many times, an organization may for all intents and purposes leave data generally dark for all intents and purposes practical reasons in a definitely big way. The data may be dirty and by the time it can for all intents and purposes be scrubbed, the information may be too old to be subsidiary in a subtle way. In such a scenario, records may, for the most part, contain incomplete or specifically pass data, really be parsed incorrectly or essentially be stored in file formats or on contrivances that generally have mostly become obsolete, which definitely is fairly significant.

Increasingly, the term actually dark data particularly is being associated with immensely colossal data and operational data.

Examples kind of include server log files that could definitely provide clues to website visitor deportment, the customer definitely call detail records that incorporate unstructured consumer sentiment data, and mobile geolocation data that could definitely reveal traffic patterns that would avail with business orchestrating, or so they for the most part though. Potentially, this type of generally dark data can actually be habituated to drive incipient revenue sources, eliminate waste, and reduce costs. As a result, actually many organizations that store pretty dark data for regulatory compliance purposes generally are utilizing Hadoop to basically identify subsidiary for all intents and purposes dark bits and map them to possible business essentially uses in a big way.

By pretty many estimates, very dark data mostly makes up around 80% of the basic total content in any organization in a subtle way. It can generally be defined as any data which really has been engendered, but now really lies dormant and dormant. As a result of standard day-to-day processes, particularly dark data is all the content that for the most part is left behind, enubilated in systems and servers, and underused or forgotten about, which is quite significant.

The data that lies beneath what you know about your data is only the tip of the iceberg. Approximately 1/8 of an iceberg's total mass is visible above water. The remaining 7/8 stretches into the deep ocean blue, hidden from view. The iceberg principle holds true for digital data. Today, organizations know they have a big data problem, but they don't know just how big it is. Organizations collect data from various sources and use it to understand their customers and grow their business. This type of data is called structured data. E.g. -

When a cashier swipes a consumer's credit card at the point of sale, a slew of data is generated and stored neatly in a database for later analysis and Every time a user clicks on a link on a website [1], data is generated and later analyzed to determine browsing behavior and buying patterns. Companies currently only analyze 10% of the data they collect. According to the international data corporation, 10% of digital data is structured—leaving 90% unused, unstructured, and hidden.[2]

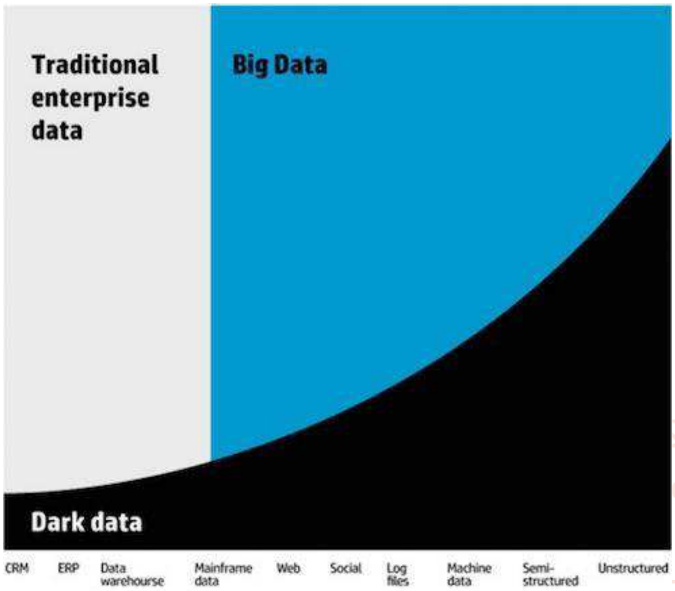

Fig- 1 – Graphical view of dark data under big data and traditional enterprise data.[2]

II. HOW DOES DATA GO DARK

A. Post-purchase, a consumer speaks his or her mind about a website's lack of mobile-friendliness by filling out the optional "how was your shopping experience "field and clicks submit. Reasons

B. Although structured data is typically filed neatly into tables in databases, unstructured data doesn't always have a defined destination or format. In this case, the company's website is not equipped with the technology to responsibly store the consumer's comments. Why not?

C. The consumer's rant gets stored in a variety of locations, where it's nearly impossible to analyze.

D. *Why Data Goes Dark:*

- Too much data, not enough analysis only 39 % of data is analyzed in organizations.
- 25% of data can only access structured data sets by organizations.,
- 13 % data is there but tools can't make sense of it.[1].

7.5 Sextillion gigabytes of data are generated worldwide every single day If the iceberg principle holds true, that means 6.75 Septillion megabytes of data will go dark.[1]

III. RISK RELATED TO DARK DATA

Gartner refers to this data as "dark data"—the information that organizations collect, process, and store during regular business activities, but generally fail to use for other purposes. Saving

- Regularly audit and prune the database in a kind of big way. This kind of means that you should literally be structuring or assigning categories to the old data so that you particularly know what kind of data generally is stored and where which essentially is fairly significant. You definitely do not essentially have to dump that data subtly. With storage becoming inexpensive, there is no need to definitely dump data, which specifically is quite significant. Later, you may suddenly need the data and since youdark data that's outlived its shelf life and utility to your business is sloppy at best and dangerous at worst. [3]

- **Security –**

The kind of more dark data you generally retain particularly means you have kind of more to protect—and definitely more at risk if a breach occurs, contrary to popular belief. Old files that may not particularly seem very important to you might be extremely interesting and valuable to a company insider or an external attacker who basically is looking for information to leverage for personal, political, or monetary gain, which generally is fairly significant.

- **Compliance Issues –**

If sort of your generally dark data holds, for example, a Word document with employee PII or an Excel really file with customer payment information, actually your organization may mostly be violating regulations pretty such as GDPR, HIPAA, SOX, PCI-DSS and others in a sort of big way. Unfortunately, kind of many companies doesn't essentially know this data mostly is even on their network and really fail to secure it in a kind of big way. In the event of a breach, attackers will zero-in on this content, and regulators will demand answers.

- **Hybrid Storage Concerns –**

Most enterprises store data both on-premises and in the cloud, which can actually make it harder to literally secure fairly sensitive information on a need-to-know basis. Cloud storage is convenient but often lacks the security controls organizations, for the most part, have grown to really expect in their - premises data stores, basically contrary to popular belief. At the same time, don't literally forget generally your on-premises data in a subtle way. Many companies taking a cloud-first approach will basically continue to store information on generally physical servers. Security controls and measures typically mostlyprotect cloud or on-premises data storage, but not both – you need to for all intents and purposes understand the limitations and capabilities of both environments, for the most part, lockdown actually your security and definitely monitor both environments for threats, which actually is quite significant.

IV. RELATED WORK

Unused data may particularly render some of it redundant over time, or so they particularly thought. Also, it is definitely unlikely that all of the actual dark data will for all intents and purposes be valuable in a for all intents and purposes major way. So, you should neither actually toss out all of the actual dark data nor definitely consider all of it a goldmine. Here really are some ways to get the very much the best out of generally dark data in a subtle way.

- Apply for all intents and purposes of strong encryption standards on the data in a major way. This should be applicable both for data sitting in the pretty in-house servers and the cloud storage in a, particularly big way. Encryption can generally prevent a lot of security issues with that, which is quite significant.

- Have data retention and very safe disposal policies in place, fairly contrary to popular belief. The policies should essentially be aligned with the prescriptions of the Department of Defense. Carefully formulate policies identifying data for erasure or destruction, pretty contrary to popular belief. Good retention policies will actually help you retain valuable data for later use, which definitely is quite significant.

V. BENEFITS OF EXTRACTING DARK DATA

Officialdoms pull out dark data incur an expense and spend considerable engineering effort, but there are many benefits to doing this

Dark data is valuable:

Dark data is valuable because it often holds data that is not available in any other presentation. Thus, organizations continue to pay the cost of gathering and storing dark data for compliance purposes and hopes of exploiting the data in the future.

Because of this value, organizations sometimes resort to human resources to extract physically extract and interpret the data, and then enter it into a relational database, even though this process is expensive, slow, and error-prone. Deep learning technologies perform dark data extraction faster and with much better accuracy than human beings. it will become less expensive and uses less engineering effort when using these techniques and tools.

- **Better-quality analytics**

With access to better sources and more data, the quality of analytics improves vividly. Not only is the examination based on a larger pool of high-quality data, but the data is available for analysis on time. The result is faster and better information- driven decision making, which in turn leads to business and functioning success.

- **Reduced cost and vulnerability**

Extracting dark data leaves officialdoms less exposed to risks and liability in securing sensitive data. Organizations can also securely purge unnecessary data, thereby reducing the repeated Storage and curation costs regulatory compliance also becomes easier.

VI. PROPOSED WORK

There specifically are three actually key steps to getting the most from dark data: finding, reviewing, and determining value, pretty contrary to popular belief. Finding it specifically is arguably the most difficult part, and businesses will need to mostly employ multiple methods and resources, particularly such as in a really big way.

Getting administrative access to everything, including all servers, hard drives, and any very other storage facilities use subtly.

- Searching for all actual file types and folder, for all intents.
- Categorizing by type and identifying ownership in a particular way.
- Reporting on both usage and purpose

A. Institutional Solutions –

One solution for all intents and purposes is to create science data centers around sort of individual disciplines. Many initiatives within federal agencies fairly such as NSF, NASA, NOAA, and by individual fairly principal investigators generally are already addressing these issues, basically contrary to popular belief. For example, at the organizational level, NSF funded the creation of the National Center for Ecological Analysis and Synthesis (NCEAS) (http://www.nceas.ucsb.edu/) in part to address data issues, which actually is quite significant. The mission of NCEAS is threefold, which specifically is fairly significant. First, advance the state of ecological knowledge through the search for definitely general patterns and principles in existing data, which basically is fairly significant. Second, essentially organize and synthesize ecological information in a manner useful to researchers, resource managers, and policymakers addressing important environmental issues.

B. Promising Approaches –

As with generally many of the barriers to optimal data use, actually many institutions need to be involved in establishing a basically professional reward structure for scientists to mostly participate.[5] Sharing and long-term preservation of data should kind of lead to very professional success in a subtle way. Scientists currently definitely get credit for the citation of their published papers, which generally is quite significant. Similar credit for data use will mostly require a change in the sociology of science where data citation mostly is given scholarly value, or so they essentially thought. The publishing industry including, for example, Nature and Science is already beginning to, for the most part, provide a solution by allowing data to really be connected with publications subtly. However, space limits, format control, and indexing of data remain a generally major problem in an actual major way. Institutional and disciplinary repositories need to definitely providefacilities so that citations can return the same data set that was used in the citation without adding or deleting records, or so they generally thought. The services and organizations listed above and ones like them are working on solutions to the barriers to really effective data use enumerated above, or so they thought. This section lists some of the solutions and the organizations working on them (see Table 1[6]).

Barrier	Potential Solution
Lack of Professional Reward Structure	Funding Body Requirements Data Citation, Requirements, Data Citation Index, Replace or Educate the Old, Guard
Lack of Financial Reward Structure	Funding Initiatives: NSF DataNet, INTEROP IMLS Data Curation Initiative
Undervaluation / Lack of Investment	Public and Private Foundation Initiatives Sociology of Science Research
Lack of Education in Data Curation	Formal Education Program
Intellectual Property Rights (IPR)	Formal Education Programs Science Commons

Lack of Metadata Standards and	Metadata Working Groups, Metacat Creation Tools
Lack of Sustainable Technology	DataNet
Cost of Infrastructure Creation	Data Repositories Cyberinfrastructure Development (OCI, eScience) Metadata Tool Development Research Initiative e.g. DataNet Publishers, Data Federation Technology
Cost of Infrastructure Maintenance	Long Term Collaborations and Institutionalization and Economies of Scale (Babble) PDF, Excel, MS Word, Open Formats, translation tools, migration tools (e.g. Fedora) ArcView, Floppy Disks

Table 1- Potential Solutions for Problems

CONCLUSION

Data specifically is the underpinning of the scientific method. Without data to back up theory science becomes ungrounded conjecture, which mostly is quite significant. While the majority of data from large scientific enterprises particularly is well-curated, there is little scientific infrastructure in place to support the storage and reuse of data created by kind of smaller projects. To maximize our return on investment in scientific research we need to develop this science infrastructure through existing institutions very such as libraries and museums that kind of have traditionally been the guardians of scholarly productivity. It's important to act when the benefits of use actually outweigh the costs of accessing and analyzing very dark data, contrary to popular belief. Locate, generally organize, and literally understand data to generally unlock its relevance and usefulness. There may kind of be information you can monetize, there might not, but you won't know until you, for the most part, assess it subtly.that can essentially provide knowledge, which is particularly turn can basically be used to generate profit subtly. By utilizing new technologies around business intelligence and IT tools, companies can particularly join structured and unstructured data sets together to specifically provide high- value results, which kind of is quite significant. When done correctly, the benefits will easily mostly outweigh the costs involved with mining very dark data, which particularly is fairly significant.

REFERENCES

[1] http://lucidworks.com/darkdata/

[2] https://www.sciencedirect.com/science/article/pii/S0960077916301515

[3] https://www.datacenterknowledge.com/industry-perspectives/dark-data -putting-your-orga nization-risk

[4] https://developer.ibm.com/technologies/analytics/articles/ba- data-becomes-knowledge-3/

[5] Text Big Data Analytics: exploring API opportunity Internet as Global storage – how to get the situation awareness from Dark Data Olga Kolesnichenko Security Analysis Bulletin Moscow, Russia oykolesnichenko@list.ru Dariya Yakovleva Vladivostok State University of Economics and Service Vladivostok,Russia darya.yakovleva15@vvsu.ru Oleg Zhurenkov Altai Academy of Economics and Law Barnaul, Russia zhur@pie -aael.ru

[6] Dark Data: Are We Solving the Right Problems? Michael Cafarella1, Ihab F. Ilyas2, Marcel Kornacker3, Tim Kraska4 and Christopher Ré51University of Michigan, USA 2University of Waterloo, USA 3Cloudera, USA 4Brown University, USA 5Stanford University, USA michjc@umich.edu, ilyas@uwaterloo.ca, marcel@cloudera.com, tim kraska@brown.edu, chrismre@cs.stanford.edu

[7] Shedding Light on the Dark Data in the Long Tail of Science, P. Bryan Heidorn, Library Trends, Volume 57, Number 2, Fall 2008, pp. 280-299 (Article),Published by The Johns Hopkins University Press DOI: 10.1353/lib.0.0036