International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 1
January 2018

Dark Data and Its Future Prospects

Sarang Saxena
St. Joseph's Degree & Pg College

ABSTRACT

Dark data is data which is acquired through various computer network operations but not used in any manner to derive insights or for decision making. The ability of an organisation to collect data can exceed the throughput at which it can analyse the data. In some cases the organisation may not even be aware that the data is being collected. It's data that is ever present, unknown and unmanaged. For example, a company may collect data on how users use its products, internal statistics about software development processes, and website visits. However, a large portion of the collected data is never even analysed. In recent times, there are certain businesses which have realised the importance of such data and are working towards finding techniques, methods, processes and developing software so that such data is properly utilized

Through this paper I'm trying to find the scope and importance of dark data, its importance in the future, its implications on the smaller firms and organisations in need of relevant and accurate data which is difficult to find but is hoarded by giant firms who do not intend to reveal such information as a part of their practice or have no idea that the data even exists, the amount of data generatedby organisations and the percentage of which is actually utilised, the businesses formed around mining dark data and analysing it, the effects of dark data, the dangers and limitations of dark data.

INTRODUCTION

The term Dark Data also known as Big Data was first coined by Gartner Research Inc. who defined it as "Big data" is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

Gartner analyst Doug Laney came up with famous three Vs back in 2001.

The most interesting of 3Vs is variety: companies are digging out amazing insights from text, locations or log files. Elevator logs help to predict vacated real estate, shoplifters tweet about stolen goods right next to the store, emails contain communication patterns of successful projects. Most of this data already belongs to organizations, but it is sitting there unused — that's why Gartner calls it dark data. Similar to dark matter in physics, dark data cannot be seen directly, yet it is the bulk of the organizational universe.

Velocity is the most misunderstood data characteristic: it is frequently equated to real-time analytics. Yet, velocity is also about the rate of changes, about linking data sets that are coming with different speeds and about bursts of activities, rather than habitual steady tempo. It is important to realize that events in data arise out of the available data and that available data forms its own "social network". This means that some data serves as a "canary", other data influences and yet more data results in decisions. When the temporal relationship between two or more data sets changes (more data suddenly becomes less data), then everything else changes, even the definition of a "data event".

Volume is about the number of big data mentions in the press and social media.

Types of Dark Data

In simple terms structured data is neat and tidy, tabular in form,namely comprising of rows of columns in a database defined,and accessible. E.g. SQL-based relational databases have notonly become ubiquitous but also often regarded the 'linguafranca' of data storage, transaction processing, reporting andanalytics. Alternatively, data that doesn't conform easily to SQLbased structures represents the rebel counterpart, namely DarkData.The sources of dark data vary from company to company,organisation to organisation,

enterprise to enterprise; any ofthe following could fall into the category of dark data if viewedas redundant, and unstructured.Customer Information

Notes or Presentations

Legal Contracts

Raw Survey Data

Log Files

Financial Statements

Email Correspondences

Previous Employee Data

OBJECTIVE

1) To study the procedure that smaller firms follow to get access to data from bigger firms
2) To study if there is any regulatory body to help make the unrecognised data easily available to the smaller firms
3) To focus on the tactics adopted by businesses to harness dark data
4) To study the legal laws available in the constitution to curb the growing problems of secrecy and cyber hacking

Need Of The Study:

This paper aims at creating awareness among the common man about the concept of dark data. It aims to understand the various prospects that will arise through the utilization of dark data by various firms, the organizations being formed for the mining of dark data, the advent of AI for integrating and classifying unrecognised data by big firms like IBM, Google.

The challenges, security concerns thatmay arise with data logs, personal info, databases which are unmonitored and not even known to the firm sometimes. The need for regulation bodies exercising control on what can or may be accessed under various markets to ensure that there is not violation of any guidelines or business practices or misuse of sensitive information. The purpose for the data collection, the motto, the values and goals of such research based organisations venturing into tracking dark data through various methodologies.

Review Of Literature

In a paper published by Stanford university professors " Extracting Databases from Dark Data with DeepDive" they talk about the uses of the system DeepDive such as it can extract structured databases from dark data accurately and vastly more quickly than human beings. DeepDiveis distinctive because of its ability to produce databases of extremely high accuracy with reasonable amounts of human engineering effort. In a remarkable range of applications,DeepDive has been able to obtain data with precision thatmeets or beats that of human annotators. DeepDive's highquality is possible because of a unique design based around probabilistic inference and a specialized engineering development cycle; they are underpinned by several technical innovations for efficient statistical training and sampling.used it to extract high quality information in a number of disparate areas, including genomics, insurance, Web classified ads, materials science, palaeontology and others.

In a paper published by Harvard Business Review titled "Small Businesses Need Dark Data, Too" by Christiana Donnelly and Geoff Simmons conducted a research with Gillian Armstrong of the University of Ulster and Andrew Fearne of the University of Kent to build awareness among small firms about the value that data could have for their businesses. With funding from a regional UK government agency they were able to get over the cost barrier. It allowed them to get loyalty-card information from supermarket giant Tesco, free of charge, to seven firms in the northern Ireland region of the UK. The formalized structure of loyalty-card

data within a statistical format requires firms to take a more formalized and structured approach to marketing planning that's a challenge for small companies. One of the researchers helped owners-managers of small businesses to retrieve the most relevant data from the loyalty-card database and analyse the information. They found that prior to being exposed to the loyalty-card data, the small businesses tended to be dominated by their owner-managers, who made decisions on the basis of their past experiences and any consumer info they could get their hands on. Once they were given access to loyalty card, they were quick to adopt to a more formalized approach to marketing, long-range innovations, rather than reacting to competitors or the retailer's actions.

In a paper published by CommVault titled " Turning Dark Data into Smart Data" featuring research from Gartner Research inc It talks about the CommVaultSimpana Email and File Analytics which helps toControl the growth of data: By helping enterprises to clean up legacy and current data by moving it to lower cost storage and that which can be deleted.Keep only what has business, compliance or evidentiary value, cutting storage costs by 70%Support Compliance & eDiscovery: Automate and enforce retention and defensible deletion to reduce risk. Provide control over access rights and security processes.Extract Value:Simpana can support a well defined data strategy and be used to enforce information governance policies.Apply, audit and leverage data classifications for better insight, control and security of unstructured data.

Research Methodology

The Research Methodology used is secondary data which is collected from web sources, published paper, Journals, Magazines, Books etc.

Smaller firms

In a small or medium size business, chances are they feel that they do not need to invest in extensive customer data, relying instead on well-honed intuition to hold their own against data-rich, bigger competitors. Of course, for the smallest businesses, access to extensive consumer data can be prohibitively expensive. Some firms tend to find the whole concept daunting - they know they lack the expertise and the time resources to make good use of the information. Dark Data threatens to create a deep divide between the have - datas and the have no–datas, with big corporations gaining advantage by crunching the numbers and having a large source of dark data tap into it besides data which the organisation deems essential and small firms are left to stumble in the dark. For small and medium size businesses that do not manage to acquire consumer data, there's much work to be done. They need to be sure to encourage employees to participate in thinking about how to use data available competitively.

Big Firms

Big businesses exercise their power and dominance in the market to conduct wide spread research across nations. Although not most of it is utilised, aside from the objective for which the data was collected and the unrequired data goes into the stockpile of unfinished, unregulated data which the business expects to be of use in the future. Big businesses also enjoy great reputation and traction in the market as a whole, as aresult individuals, governments of other countries, prospective customers and existing customers are less circumspect about providing information to such firms whereas smaller firms or medium scale enterprises find it hard to get funding for their research and appropriate data.

Provision For Access to Dark Data

Businesses which cannot afford to engage in research activities at a larger scale or require information access which is beyond their reach and do not have the infrastructure or proper credentials to conduct such research should get help from big business houses to carry on their research on their behalf provided they get monetary benefit which is reasonable and in based on the width, subject matter, availability and difficulty. Also these business houses have a mass of information unrecognised i.e dark data which the company has amassed over a number of years, of which some might not be of use to them and they can grant access to such information through a deal where the company providing the information gets a royalty or a percentage of what the company employing the data or the information makes out of it, this will help further business growth and help in harnessing the otherwise scattered, unclassified and unused data. Governments and Universities can

play an important role in bridging the data divide, by providing funds and expertise so that small firms can get access to, and learn to interpret data.

Regulation of dark data

In most areas of the world, there are legislative rules and regulations for how companies must protect and manage personally identifiable information (PII), such as passport information, credit card and banking information and healthcare information. These laws have been inconsistent and have varied widely in the level of data protection they mandate, but it had not been a major concern as breaches of PII were relatively rare.

But in recent years, the amount of personal data stored by companies and governments has grown dramatically, prompting regulators to re-think their requirements for organizations. This explosion of digital PII data is the driver for GDPR, or new General Data Protection Regulation being enacted in Europe. It is a comprehensive set of new rules that mandate how PII data must be managed, not just for European companies, but for any company doing business in Europe or with European customers. GDPR was adopted April 27, 2016, and becomes enforceable on May 25, 2018, and penalties for non-compliance are harsh. They include fines upto 20,000,000 Euros or 4% of annual revenues for certain offenses, and companies are scrambling to set up and implement GDPR compliance initiatives. In May 2018, the new GDPR legislation will become effective, with new requirements for processing and protecting personal data. The main purpose of the GDPR assessment is a roadmap that prepares an organization for this GDPR legislation and to test risk factors in the organization of the client.

GDPR also allows for individuals to ask if their personal data is being captured and processed, and if it is, the organization must be able to produce copies of their personal data in electronic format. Organizations are also tasked with ensuring contracts contain provisions regarding the tasks and responsibilities of the data processor, including how and when data will be returned or deleted after processing, and the details of the processing, such as subject-matter, duration, nature, purpose, type of data and categories of data s

While GRPR includes a variety of components addressing the processing and management of PII, the extraction and analysis of data with contract documents plays a very important role in compliance. There are 3 specific areas where contract data plays a role, including:

Understanding where PII might be hidden, in particular in the “dark data” found in contract documents.

Ensuring data breach obligations, as indicated in contract documents, are understood and comply with GDPR requirements.

Confirming contractual agreements with data processors or other vendors that may come into contact with PII have the appropriate clauses and a defined scope.With GDPR on the horizon, this issue should be addressed. A great place to start may be to assess what personal information you have, where you keep it and what you are using it for. In this respect, GDPR can be an enabler to help transform your company into a truly data-driven organization. There are tools available to help support this process that facilitate you analyzing your structured and unstructured data by looking for patterns to identify and locate personal information and other kinds of data – no matter where it is stored.

Stanford University Professors developed DeepDive a system for extracting relational databases from dark data: the mass of text, tables, and images that are widely collected and stored but which cannot be exploited by standard relational tools. If the information in dark data — scientific papers, Web classified ads, customer service notes, and so on — were instead in a relational database, it would give analysts a massive and valuable new set of “big data.” DeepDive is distinctive when compared to previous information extraction systems in its ability to obtain very high precision and recall at reasonable engineering cost; in a number of applications, we have used DeepDive to create databases with accuracy that meets that of human annotators. To date we have successfully deployed DeepDive to create data-centric applications for insurance, materials science, genomics, paleontologists, law enforcement, and others. The data unlocked by DeepDive represents a massive opportunity for industry, government, and scientific researchers. DeepDive is enabled by an unusual design that combines large-scale probabilistic inference with a novel developer interaction cycle. This design is enabled by several core innovations around probabilistic training and inference.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 1
January 2018

SAP CEO on Dark Data

From a product perspective, users want their technology to be fast, secure, and smart, said Bill McDermott co-CEO of SAP, which leads to the push behind the HANA platform. “Smart in the way that it’s no longer about analysing data from the past – it’s the era of real-time, and predicting the future," explained McDermott. "Ninety-eight percent of your data is locked up somewhere you can’t access it – we call this dark data."HANA can unlock that and combine it with structured data to determine not only the sentiment of that data, but also predict the intent of that data.

“Secure in the way we need to ensure that merchants and banks are safe, and that consumer’s transactions are fully analysed. HANA tracks each card swipe and picks up on inaccuracies and uses innovation to bring peace of mind to the consumers – something we all need. “And finally, speed. We can all agree on one thing; slow kills companies fast.”

The company plays to move all of its line-of-business Cloud applications to the HANA Cloud Platform. Supporting this, it's done some re-branding so that SAP Financials OnDemand becomes SAP Cloud for Financials and SAP Travel OnDemand becomes SAP Cloud for Travel and so on.HANA not only represents the intellectual renewal of SAP, it is now the platform for every single thing the SAP company will do going forward."We knew big data would be big business for you," "And that's why we invented HANA, to give you intelligent data at the speed of thought. The SAP Business Suite on HANA is at least 2,000 times faster than any other product on the market today. And, SAP HANA has become the fastest growing software product in the world."

IBM’s Watson to help in Dark Data Analysis

According to IBM, around 90% of data gathered from sensors and analog-to-digital conversions never gets used. It means the majority of such data simply becomes dark data.

For instance, In a stock market, a trader will typically analyse an asset by creating a time chart that compares the asset’s value over a period of time. To get more useful information, that asset’s data is compared to other assets to figure out its relative value. One of the ways to simplify this process is using tool like IBM Watson to create multi-dimensional charts that can compare relative assets.

In June of 2017, Watson entered the exchange trade fund(ETF) business. The Equobot with Watson AI Total US ETF was filed to the US Securities and Exchange Commission. According to the file, Equobot will use Watson to “conduct an objective, fundamental analysis of US-listed common stocks and real estate investments trusts based on up to ten years of historical data and apply that analysis to recent economic and news data”

Spencer Fontein and Rob Williams of All Blue Solutions explain how dark data can be analysed for stock market trading using IBM Watson and Bluemix. Based on the analysis of messages that appear on TV and social media, they believe that AI can be utilized to get a better understanding of trade position fluctuations as events transpire in real time.

Risk Of Security Breach and Cyber Hacking

There is Legal and Regulatory Risk If data covered by mandate or regulation e.g. patient records,appear anywhere in dark data collections, its exposure couldinvolve legal and financial liabilities If dark data encompasses proprietary or sensitive informationreflective of business operations, practices, competitiveadvantages, important partnerships, joint ventures, etc.inadvertent disclosure could affect the bottom line or compromise important business activities and relationships. Any kind of data breach reflects badly on the organisationsaffected. This applies as much to dark data as to other kinds ofbreaches of particular concern should be where an organisationhas customer and/or operations stored in the cloud outsidetheir immediate control and maintenance.Another key challenge presented by dark data is determining itsreal value, if any at all.Much of dark data remains ‘unilluminated’ becauseorganisations simply do not know what it

contains. To destroy it may prove too risky (because of compliance issues), butanalysing it can be costly making it hard to justify the expenseif the potential value of the dark data is unknown.
One growing problem all businesses face in the present is constant threat to the secrecy and confidentiality of business operations, databases , files, documents , customer information and logistics. The danger is the fact that businesses are not aware that a large portion of data is generated and sometimes businesses do not know that it actually exists. As businesses are not aware of the existence of such data, it may well cause information to be used for malpractices, Medium to large organisations are generally able toprovide terabytes of file share storage space for theiremployees and departments to utilise. Employees drag anddrop all kinds of work related files, as well as personal files suchas personal photos, MP3 music files, personal communications,etc. In addition, PSTs and work station backup files. The clearmajority of these files are unmanaged and therefore neverlooked at again by the employee or anyone else. Hence, a security breach is a huge problem in case of unstructured dark data which the businesses have to keep track in order to avoid loss of reputation, intellectual property, copyrights which might be copied or replicated and release of personal customer and employee information which may cause outbreak.

Conclusions

In today's evolving information driven society one of the keyattributes of leading organisations is how deeply theyunderstand their market, customers, and competitors. As partof this 'information age' revolution organisations are gatheringexponentially vast amounts of data – 'big data'. Thus, Dark Data is going to have a huge impact in the future when holding data and utilising to its fullest potential will gain importance among the firms as there are already tell-tell signs of businesses adopting to the change and regulation and guidelines of maintaining data also being implemented, Proper assessment of Dark data will become a business imperative all around the world.

References

- //hbr.org/2013/12/small-businesses-need-big-data-too
- //www.altoros.com/blog/analyzing-dark-data-for-traders-using-ibm-watson-bluemix/
- //www.federalnewsradio.com/wp-content/uploads/pdfs/031115_gartner_co_branded_newsletter_turning_dark_data_into_smart_data.pdf
- //www.mycustomer.com/marketing/data/bigger-than-big-data-sap-ceo-declares-war-on-dark-data
- //www.highquestsolutions.com/darkdatadoc.pdf
- //The upside of GDPR: a potential remedy for your "dark" data - The Netherlands www.ibm.com
- Dark data in contracts poses hidden risk to GDPR compliance www.information-age.com
- Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s www.forbes.com
- //www-cs.stanford.edu/~chrismre/papers/modiv923-zhangA.pdf www-cs.stanford.edu