DEMYSTIFYING BIG DATA AND RECOGNISING ITS POTENTIAL IN VARIOUS INDUSTRIES

Prateek Chitpur[1]

[1]*UG student, Department of Computer Science and Engineering, JSSATE, Bengaluru*

Abstract

Twenty-first century has seen many advancements in the field of Technology and Engineering. One of the main driving forces of that has been the invention of Internet and its massive growth in the couple of decades. Now with this advancement and growth there also has been massive increase in the amount of data being produced. This data is being produced every minute and its amount has increased at such a large scale that it requires to be managed for its efficient utilization. According to IBM, approximately 2.5 exabytes of data was generated per day in 2012. Now this asks for some efficient method to deal with this mammoth amount of data. Big Data Analytics provides methods and tools to deal with that. Content of this paper illustrates about Big Data Analytics and how it can be applied to various fields[1].

Keywords: - *—Big Data, Big Data Analytics.*

----------***----------

1. INTRODUCTION

The term Big Data denotes the massive collection of data, both structured and unstructured. It is termed as "Big" Data because the velocity at which this data is being produced is tremendous and thus it increases its capacity manifold. The applications of this data, if used efficiently are boundless and that's where Big Data Analytics plays a pivotal role. The analysis of BigData involves applications of various algorithms such as MapReduce, Hashing etc.

Hadoop, a Java framework,is being used for storing and managing the huge amount of data in a very cheap and efficient manner. Hadoop mainly comprises of two parts: a data processing framework which is java-enabled and a distributed filesystem for data storage. The distributed filesystem consists of array of storage clusters which are also named Hadoop Distributed File System(HDFS) [2].

2. HISTORY OF BIG DATA

History says, the term Big Data has been coined around 2005, giving name and fame toJohnR. Mashey for its popularization. The current report claims, approximately 90% of existing data has been created in the past 2 years. However, acquisition and usability of data has beenaround much longer [3].

3. FORMAL DEFINITION

Basically, Big Data is vast cluster of complex data sets. These data sets don't show any specified structure or pattern, which is characterized by 6Vs of Big Data: Variety, Volume, Velocity, Validity, Veracity and Volatility.

3.1 Volume

Volume is concerned with the actual capacity of the data that is being stored for analyzing and processing. Since a vast amount of data is being produced everyday, it is of utmost importance to make sure the irrelevant data doesn't find entry to our storage systems. Otherwise this will lead to increased storage space and processing required to find the required structure and patterns to solve problems [4][5].

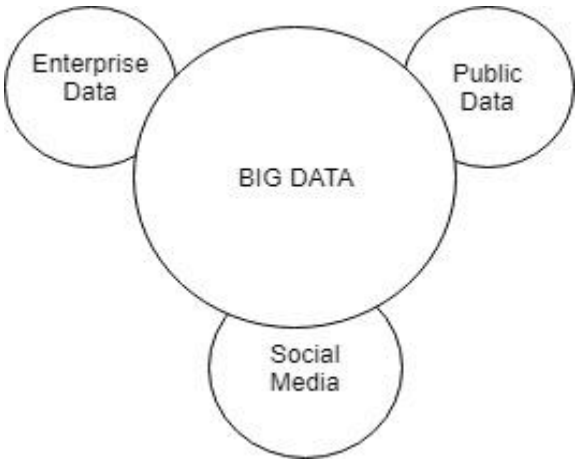

Fig 1: Representing Volume

3.2 Variety

Variety refers to data generated from varied set of sources that can range from social networking sites to online business transactions done on e-commerce websites. It is also concerned with the type of data whether its textual as in spreadsheets and databases, multimedia data such as videos and audios, photos etc. [4][5].

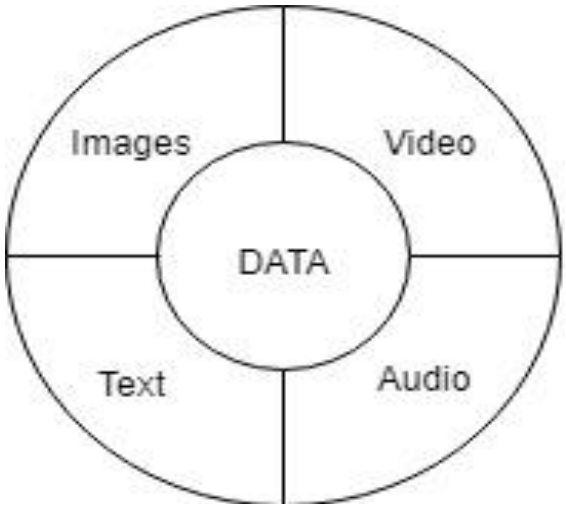

Fig 2: Showing Variety of Data Types

3.3 Velocity

Variety refers to the speed at which data emerges from various sources such as social networking websites, online business process, healthcare industry, Fintech industries etc. Since the advent of Internet, the velocity of data production has increased manifold and thus this can help in characterizing the datasets [4][5].

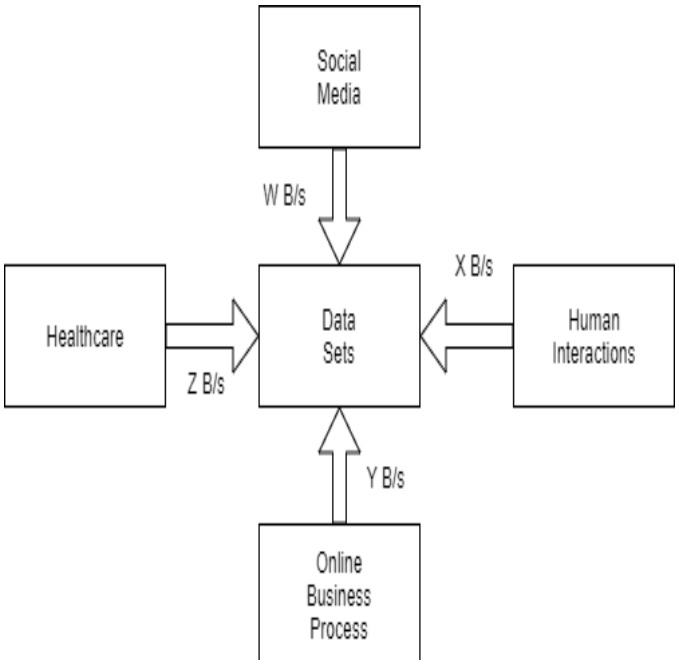

Fig 3: Illustrating Velocity of Data at Each Sector

3.4 Veracity

Veracity defines the abnormalities and disturbances such as noise, deviation. It is mainly concerned with whether the data being analyzed and processed and is concerned with the required information to solve some specific data problem. In order to apply Analytics to data efficiently it is of great importance to make sure that the data being stored for processing is clean and relevant [4][5].

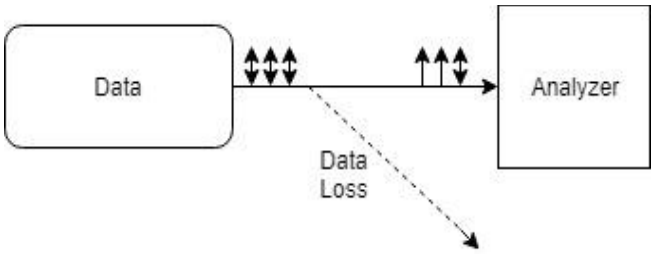

Fig 4: Depicting Veracity of Data

3.5 Validity

Validity is concerned with whether the data being used is correct and accurate for relevant problem. The correct and accurate data is very important to make relevant decisions. Ensuring the validity of data helps in making the best use of available data [4][5].

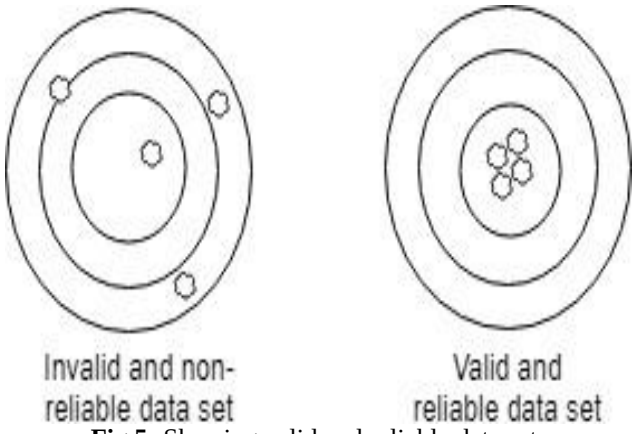

Fig 5: Showing valid and reliable data set

3.6 Volatility

Volatility is concerned with the time period for which the data being processed is valid and the time period for which that data should be stored. It helps in getting rid of data which is no longer valid or needed. That in turn increase the processing speed while reducing the storage space required. While dealing with huge volumes of data it is of utmost importance to manage the storage space and volatility helps in that [4][5].

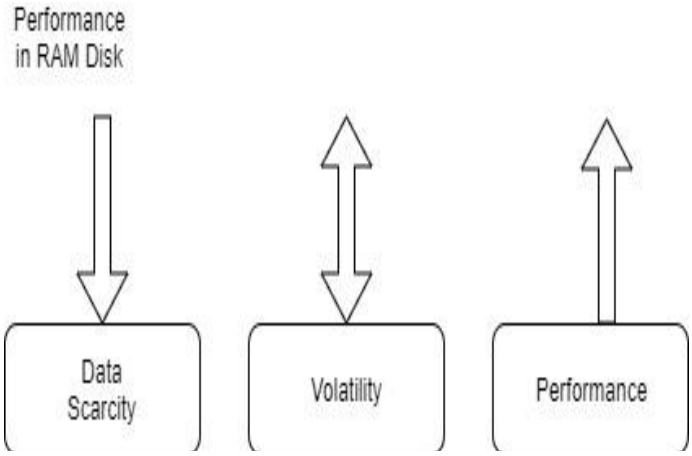

Fig 6: Representing Volatility and Performance in RAM Disk

4. TOOLS OF BIG DATA ANALYTICS

Big Data Analytics mainly comprises of two basic tools: Security Information Management(SIM) and Security Event Management(SEM), whose features has been combined together to form Security Information and Event Management(SIEM) capabilities. These capabilities basically comprise of:

a) Data Aggregation: Data is accumulated from various sources such as spreadsheets, databases, networks, servers, and applications to make sure that no crucial and relevant event is neglected.

b) Correlation: Data-sets from accumulated data is grouped and linked together to convert data into useful and relevant information.

c) Alerting: Automated alert systems make sure that any abnormality or security issues get reported immediately.

d) Retention: It is very uncommon that the cause of the data breach or any security issue will be known beforehand. Thus, it is imperative to store data and retain the useful information related to previous breaches.

e) Forensic Analysis: It involves the ability to search across very time periods for relevant data based on specified criteria [6].

5. APPLICATIONS OF BIG DATA IN VARIOUS INDUSTRIES

5.1 Healthcare Industry

Big Data is a buzzword in this modern world and its pace at healthcare industry is very useful. It is used in analyzing Electronic Health Records(EHRs), where doctors share EHRs which aggregates and analyzes data to reduce healthcare costs. As patients are examined and data shared among physicians and healthcare providers, duplicity of tests can be reduced and patient care is improved. Mainly EHR data isfor regulatory and security compliances, but finding this secure way to extract patient records can improve cost reduction and quality of care [7].

Nowadays, Big data is used in healthcare Internet of things(IoT), smart devices and sensors are interconnected, thedata they generate is moved between devices and finally to people. Many devices monitor each and every sort of patient behavior without physician intervention, for instance smart dispensers detect whether medicines are being given regularly at home. If not, they initiate call from providers to medicate patient properly [8].

For example, In Cancer treatment personalized medical care and facility can be provided to patient according to the type and stage of the disease the person is suffering through. It can also help in finding the right type of medicines andchemotherapy processes that suits the patients.

5.2 E-Commerce Industry

In this industry, the data is categorized into two types: structured and unstructured. Structured data is the regular data which includes address, name, age and sex that is been easily captured whereas unstructured data refers to pictures, videos, likes, tweets, customer preferences and many more collected from various organizations like Facebook, Twitter, Amazon, Flipkart, etc. [10].

Big data is widely utilized in the following major areas of unstructured data of e-commerce industry. They are:

Betterment of Customer Experience-improving customer experience is a major concern for e-commerce companies.Since customers have got various choices for same item hence it poses a tough competition for e-commerce companies. They analyze data from customers feedback and their buying paths, which is further processed for customer's superior treatment.

Pricing- based on demand for the products and competition among e-commerce companies, pricing is constantly monitored.

Personalization- this will help e-commerce companies to increase conversion and personalize emails very effectively.

Predictive Analytics- Amazon is the best example which uses predictive analytics to predict what their buyer or consumer will buy. This is not about the marketing ability of the Amazon, but it is about its ability of prediction [9][10].

5.3 Online Dating Services

Usage of big data in these days became a major contributor in finding one's spouse. There are number of online dating sites namely eHarmony, OkCupid.com, Match.com use big data.Citizens in countries like US, China use dating sites to find their perfect match. According to Match.com, it claimed that it has collected 70 terabytes of data from their clients, which created 5,00,000 companionships resulting 92,000 marriages, which, in turn, resulted 1million babies [11]. These dating sitesgather information by various methods from clients seeking their love. For instance, some online dating sites use match percentage, predictive compatibility model to decide on how similar to individuals are. Typically, posing questionnaires to clients turned up to be an efficient way in collecting data. The questionnaires will be like interests, education qualification, likes, dislikes and so on. After gathering data, it is analyzed and compiled using data base systems. And then it is fed to various algorithms, performing predictions to find perfect match [11].

5.4 Security and Fraudulent Detection

According to Verizon Annual Data breach 2017, 40,000 incidents of data breach has been analyzed of which 1,935 has been confirmed lately [12]. One of the rising causes of these breaches on careful analysis can be attributed a couple of factors which include malware-injection and hacking practices. Data Analytics along-with Predictive Analytics can be used to mitigate this problem.

Data Analytics can be used to accumulate data related to breach and then segregate it according to the reason of the breach. Predictive Analytics can be then used to generate

patterns, based on which reliable prediction can be made about what kind of attacks can be made in future on the security systems of various organizations. This will also help in designing and developing security systems which will be shelled by unauthorized access and attacks.

Till now this aspect of Data analytics hasn't been realized completely and thus largely unutilized. Data analytics offers a very robust and reliable security maintenance feature which needs to be utilized efficiently.

5.5 Travel and Tourism Industry

Travel and Tourism Industry is an area where the importance of data is being realized now. The end-user's history of services used in this department forms the most useful data-set. These data-sets can be used by airlines, travel agencies, hotels and restaurants various other allied sectors to their advantage. It can also help in improving the customer experience.

Let's take some examples: Storing data and processing it can tell us about customer choices such as which hotel customer chose for his previous stay, the rating of hotel, the duration of stay of the customer and other relevant parameters. This can help in predicting hotels next time for longer stay and others for shorter stay.

The security aspect surfaces here again. Data generated from online transaction while buying products can be used for wrongful means. Many instances have occurred when the credit card of the customer abroad gets blocked when the customer is trying to use it there. These conditions mainly occur due to some unauthorized interferences in the customer's credit card transactions data.Thus, this can be used to make the existing algorithms used for analytics better to deal with this kind of situation [13].

6. CONCLUSION

Data is of utmost importance and its value has increased manifold. Almost every other source now produces data at a large scale. Preserving the data for processing has become an important event. Efficient use of Hadoop framework and SIEM capabilities such as correlation and data retention can be used to one's advantage to find hidden patterns and structures for the profit of all the stakeholders. Predictive Analytics along with Data Analytics offers another major resource to enforce and bolster the security mechanism of organizations and secure the organization's important and private data ranging from customer account information to online business transactional processes. It's the most opportune time to make the best use of this large pool of available and continuously increasing data for the development of all the stakeholders.

REFERENCES

[1]. https://www-01.ibm.com/software/in/data/bigdata/
[2]. http://hadoop.apache.org/
[3]. Mukesh S B, Prasanna Kumar, Dr. D.V Ashoka "Big Data Applications in Personalised Marketing, Personalised Healthcare and Other Services" Available from https://ijircce.com/upload/2017/iccstar/6_D004.pdf
[4]. https://www.oracle.com/big-data/index.html
[5]. https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/
[6]. PrashasthavardhanaPrahlad, Sachin K. Rai, Dr.D.V.Ashoka "Use of Data Analytics to Recognize Threats in the Security Systems of Banks" Available from https://ijircce.com/upload/2017/iccstar/5_D003.pdf
[7]. http://attunelive.com/big-data-applications-healthcare/
[8]. https://mapr.com/blog/5-big-data-trends-healthcare-2017/
[9]. http://www.apparatus.io/The-application-of-Big-Data-for-ecommerce
[10]. http://blog.venturesity.com/what-is-big-data-analytics-and-its-application-in-e-commerce
[11]. http://sites.psu.edu/bigdataebook/chapter5/05-03/
[12]. http://www.verizonenterprise.com/verizon-insights-lab/dbir/2017/
[13]. https://www.futuremarketinsights.com/reports/big-data-analytics-in-tourism-overview-and-trends-analysis