BIG DATA SECURITY

Ms. Suman Madan
Assistant Professor (IT),
Jagan Institute of Management Studies,
New Delhi- 110085, INDIA.

Abstract : **Information today has gone from scarce to superfluous which brings gigantic new benefits but paired with big headache. We are all surrounded by just data and only data and we hardly have time to track that from where this data has flooded. In real world, to get the right answer, we must first have the right question to be asked and similarly to get the correct analytics from the data we must first have the real data rather than uncertain data. Earlier the amount of important data was less and hence easily maintainable, but today every bit of data is important and hence stored. Big data is a collection of data sets which is very large in size as well as complex, generally in Petabyte and Exabyte. Now-a-days the actual problem is that more the data more accurate analysis and forecast is possible but the dark truth behind it is that the actual data is surrounded by enormous amount of uncertain data. This uncertain data is replicating at a speed which is almost ten times more as compared to the real data because of use of internet, smart phone and social network. This paper throws light on important concepts of Big Data and discusses various aspects of big data including V's of big data. The introspection is done at the processes involved in data processing and assesses the security aspects of Big Data and proposing a encryption system for data.**

Keywords— Big data, Big data V's, security

I. INTRODUCTION

William Gibson once said that- "The future is here, but it's just not evenly distributed yet." According to The Economist [1]- In 2010 the digital universe was 1.2 zettabytes of data, in a decade the Digital Universe grew to 35 zettabytes, and in 2011 it shoot to 300 quadrillion files in which 90% of digital universe data is Unstructured- which is an alarming issue.

Alex Szalay, an astrophysicist at Johns Hopkins University, states that "The procreation of data is making them progressively unapproachable." He says that people must be trained not only the scientist or the computer professionals or the industry or government but including all of them who are contributing towards this never ending data creation must be trained in -How to make sense of all these data?

James Cortada of IBM said - "We are at a different period because of so much information." The information available is next to infinite which is getting difficult to manage and tackle. According to John Easton, IBM Distinguished Engineer-Advanced Analytics Infrastructures 80% of the data available by 2015 will be uncertain [2]-hence creating problem to handle (See figure 1). Joe Heller stein, a computer scientist at the University of California in Berkeley, calls it "the industrial revolution of data." The effect is being felt everywhere, from business to science, from government to the arts. Scientists and computer engineers have brainstormed a new term for the phenomenon: "BIG DATA".

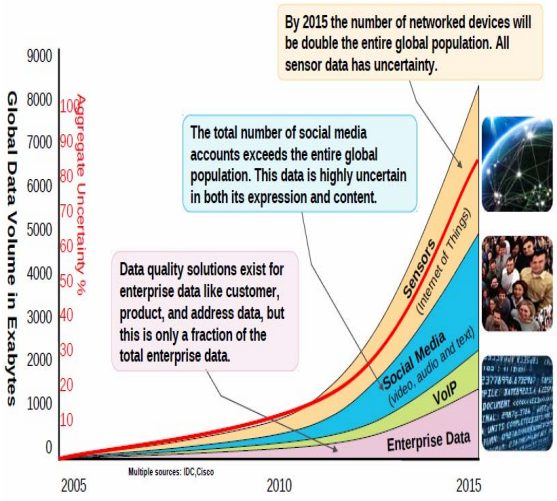

Figure 1: How Data Uncertainty is increasing [2]

The term "BIG DATA" was given by Roger Magoulas from O'Reilly media in 2005[3]. He explained that due to its whopping size and complexness, wide range of data sets is almost becoming insoluble to handle and manage through traditional data management tools. Big data can be seen in Banking and business for Inventory Management, Customer Behavior, Market Behavior. Big data is also seen in life sciences for analyzing and advance research in Genome sequencing, clinical data and patient data areas. Big data could also be seen in other areas like Astronomy [1,3] and Oceanography[3]. Retail-

giant Wal-Mart handles more than 1 million customers transaction every hour, fattening databases approximately to more than 2.5 Petabyte which is equivalent to 167 times the America's Library of Congress [1]. Social networking website Facebook is a home to about 40 billion photos. These examples conclude the same phenomenon: that the universe embraces an indescribable mammoth extent of digital information which is clutching gigantic more speedily. The amount of digital information accelerates tenfold every five years [1]. According to Moore's law, which IT sectors takes for granted, says that 'the processing power and storage capacity of computer chips double or their prices halves roughly every 18 months.'

Due to hasty outburst and combustion of data & the demand for digital collaboration everywhere, IT people know that the conventional data management techniques are no longer helpful in managing and utilizing their data, so they are moving towards finding advance solutions to secure their data. The surge in computational & storage power enables the gathering, storing and analyzing these Big Data sets and introduction of innovative technological solutions to Big Data analytics are blooming.

In order to create value from the big data, the data should be appropriately exploited by exuviating greater transparency in many fields. Three factors should be considered for this-

a. Users Control- it's time to give the users an upper hand on the control and access over the information held about them, also including with whom it is shared with, with various customizations allowed.

b. Taking Security Issues seriously- Organizations should start discussing and disclosing their hidden security policies now, in order to control security breaches within the organization.

c. Yearly examining Security points- making the security policies to be audited on yearly basis will help organizations to find out the loop holes and rectify them with more security measures for future furthermore it will help the organizations to keep their security measures up to date and help in taking a proactive step for any illegitimate entry.

Consecutively following above three factors the organizations can gain higher market initiative as compared to those not following them. Users will get more control over their data, will be secured and have freedom from complicated regulations that could starve them from innovations if, high level of transparency is provided to them.

The actual problem is not acquisition of hefty data but the undeniable need of this data. In simple words we can say large data is directly proportional to higher accurate analyses [4]. Higher accurate analyses may result in more dauntless decision making. And of course a smarter decision making can lead to an improved operational productivity, reduced costs, reduced time [5] and reduced risk. By entangling big data and high-powered analytics it is possible to-

a. Extracting the actual reason behind defects, flaws, failures in the real- time to save losses in future.

b. Analyze millions of Stock keeping units to determine prices that boost profit and clear all the stock.

c. Fully Optimize routes followed by package delivery vehicles while they are on the road.

d. Customers who have higher priority (more important) should be swiftly recognized.

e. Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.

f. Have a system for those analysis current and past purchases of various customers to generate attractive vouchers for them.

g. Reanalyze all the risk factors or points within minutes.

h. Use Click stream analysis and data mining to unmask fraud nature & behavior.

II. UNDERSTANDING BIG DATA & SECURITY

A. Components of Big Data

Big data analytics has the capacity to process any variety, volume and velocity of information and to derive an insight into data [6]. Big data analytics refers to the process of collecting, organizing and analyzing large sets of data "BIG DATA" to discover patterns and important information which in turn helps in discovering and extracting data useful in making future business decisions in addition to helping understanding the information within data. Big data analytics is about joining trusted, internal information with new data types to create value bringing new source of unstructured info to existing core data to create insight. What is this New data we are talking about -It's the information that is already there but we never used

it. Like Email, Blog, Stock Market, Sensors, Mobile Phone GPS etc.

The four V's of Big Data are-

i. ***Volume***- Increase in data volume is caused by many factors. Transaction-based data stored through the years. Unstructured data streaming in from social media. Increasing amounts of sensor and machine-to-machine data being collected. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data. Enterprises are flooded with ever-growing data of all types, easily accumulated to terabytes -even Petabyte -of information. The handle volume Big Data turns 12 terabytes of data (tweets) created each day into improved product sentiment analysis. Secondly, Converts 350 billion annual meter reading to better predict power consumptions etc.

ii. ***Velocity***- Data is streaming in at unprecedented speed and must be dealt with in a timely manner. Processing should be fast and quick for time- sensitive processes such as catching frauds. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations. Big data must be used to firstly, Scrutinize 5 million trade events created each day to identify potential frauds. Secondly, analyze 500 million daily call detail records in real-time to predict customer churn faster.

iii. ***Variety***- Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. Managing, merging and governing different varieties of data is something many organizations still grapple with. Big data must Firstly, monitor 100's of live feeds from surveillance cameras to target points of interest. Secondly, exploit the 80% data growth in images, videos and documents to improve customer satisfaction.

iv. ***Veracity***- In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved.

SAS introduced additional dimensions ***Complexity***. Complexity- Today's data comes from multiple sources. And it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

B. Misconception and Truth

A common misconception is when customers confuse the term Big Data with having to deal with lot of data. But the Truth is that volume is clearly a part of big data solution but Big Data is more about unlocking the potential of Structured & Unconstructed information, inside & potentially outside of our firewall & doing it in right time.

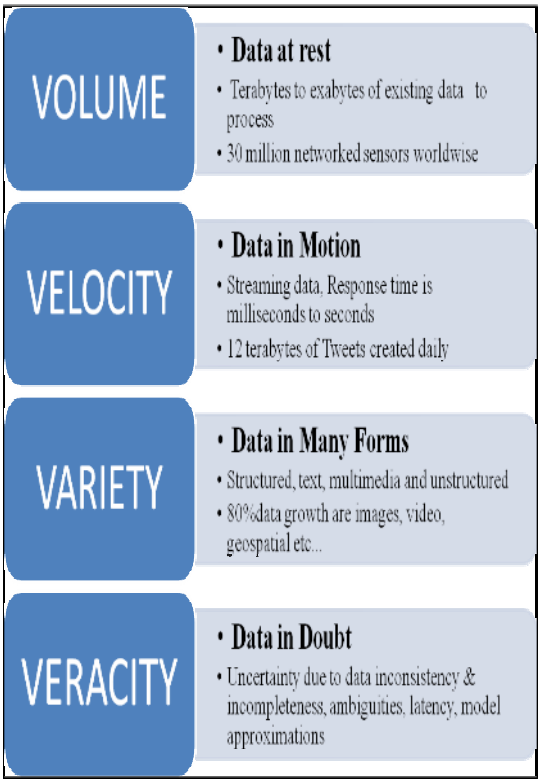

Figure 2: Four Dimensions of Big Data

C. Need of security in big data

Many of the businesses use big data for the marketing and research, but may lack the basic assets particularly from security side. Any security breach that occurs in big data would result in even more serious legal impact and reputational damage than at present. Many companies are using the technology to store and analyze Petabyte of data

about their company, business and their customers. Thus information classification becomes even more vital. For making big data secure, techniques such as encryption, logging, and honeypot detection must be necessary. In many organizations, the deployment of big data for fraud detection is very attractive and useful.

The challenge of detecting and preventing advanced threats and malicious intruders must be solved using big data style analysis. These techniques help in detecting the threats in the early stages using more sophisticated pattern analysis and analyzing multiple data sources. Like security, data privacy also challenges existing industries and federal organizations. With the increase in the use of big data in business, many companies are battling with privacy issues. Data privacy is a liability, thus companies must be on privacy defensive. But unlike security, privacy should be considered as an asset; therefore it becomes a selling point for both customers and other stakeholders. There should be a balance between data privacy and security.

III. OPERATIONAL VS ANALYTICAL BIG DATA TECHNOLOGY

In this era of Big Data, data is getting too massive for older generations of technology to handle, thus a new class of technologies sprouting up to meet the need i.e. Operational and analytical big data solutions. To succeed and pull away from the competition, a strong data management strategy is needed that involves the right mix of technologies that meet the requirements. Operational Big Data systems provide operational features to run real-time, interactive workloads that ingest and store data. MongoDB is a top technology for operational Big Data applications with over 10 million downloads of its open source software. Analytical Big Data technologies are useful for retrospective, sophisticated analytics of your data. Hadoop is the most popular example of an Analytical Big Data technology. The following table is a comparison between Operational and Analytical Systems in the field of Big Data.

Table 1: Operational Vs Analytical Technology

Parameters	**OPERATIONAL**	**ANALYTICAL**
Data scope	Operational	Retrospective
End user	Customer	Data Analysts/Expert
Latency	1ms – 100ms	1 min – 100 min
Concurrency	1,000 – 100,000	1 - 10
Access Pattern	Reads & Writes	Reads
Queries	Selective	Unselective
Technology	NoSQL	MapReduce, MPP Database

IV. SECURITY AND PRIVACY IN BIG DATA

Security and privacy concerns are on the increase as big data becomes more and more accessible. The collection and aggregation of gigantic quantities of heterogeneous data are now possible. However, the tools and technologies that are being developed to manage these gigantic data sets are often not designed to incorporate adequate security or privacy measures, in part because we lack sufficient training and a fundamental understanding of how to provide large-scale data security and privacy. We also lack adequate policies to ensure compliance with current approaches to security and privacy. Furthermore, existing technological approaches to security and privacy are increasingly being breached, whether accidentally or intentionally, thus necessitating the repetitive reassessment and updating of current approaches to prevent data leakage.

Data hackers have become more damaging in the era of big data due to the availability of large volumes of publically available data, the ability to store massive amounts of data on portable devices such as USB drives and laptops, and the accessibility of simple tools to acquire and integrate disparate data sources. According to the Open Security Foundation's DataLoss DB project [7], hacking accounts for 28% of all data breach incidents, with theft accounting for an additional 24%, fraud accounting for 12%, and web-related loss accounting for 9% of all data loss incidents. Greater than half (57%) of all data loss incidents involve external parties, but 10% involve malicious actions on the part of internal parties, and an additional 20% involve accidental actions by internal parties. Thus, Data leakage represents a major problem in today's era of big data. While stopping hackers from getting to data must be a goal, stopping hackers and authorized users from removing data from its authorized location is also a critical step to stop data leakage.

A. Existing solution for Protection of Big data

Data encryption technology is used for boost information privacy protection. However, traditional encryption primitives (such as symmetric key encryption and public key encryption) are not capable to ensure the usability and hinder even authorized users from searching

several keywords of encrypted files, it is difficult for the users to retrieve desired information from encrypted big data.

Figure 4: Query processing in Traditional Encryption System

Figure 4 shows how query is processed in traditional Encryption system. The complete database needs to be encrypted to process any query which takes significant amounts of time due to the slower rate of secure cryptographic techniques.

D. Proposed solution for Protection of Big data

It is essential to explore new cryptographic primitives to provide data encryption and ability to search for big data era. Figure 5 shows the proposed system of data encryption where the complete database is not encrypted and instead systematically encrypted to maintain little flexibility for query processing. This can be attained by encrypting individual parts of data rather than the complete database.

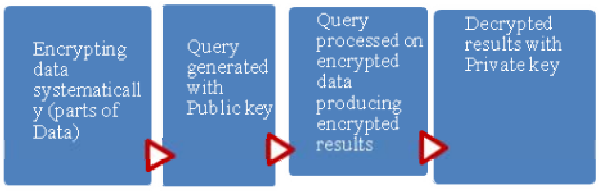

Figure 5: Proposed system of Data encryption

V. CONCLUSION

Big Data is changing the way we observe our world. The impact of big data will continue to create move through all facets of our life. Global Data is on the rise and by 2020 it would have quadrupled the data we generate every day. Existing technologies will continue to evolve as needs in data security and privacy are recognized and additional vulnerabilities are realized. Thus a new system of data encryption is suggested in this paper.

REFERENCES

[1] https://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf

[2] http://www.thebigdatainsightgroup.com/site/ system/ files/private_1

[3] http://www.researchtrends.com/wp-content/uploads/ 2012/09/Research_Trends_Issue30.pdf

[4] http://www.sas.com/en_in/insights/big-data/what-is-big-data.html

[5] http://www.sas.com/content/dam/SAS/en_us/ doc/ whitepaper2/bigdata-bigcompanies-106461.pdf

[6] http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf

[7] https://www-950.ibm.com/events/wwe/grp/grp037.nsf/ vLookupPDFs/Calgary_Keynote_%20David_%20Corrigan%20-%20v1/$file/Calgary_Keynote_%20David_%20 Corrigan%20-%20v1.pdf

[8] http://datalossdb.org

[9] http://www.computerweekly.com/feature/How-to-tackle-big-data-from-a-security-point-of-view#

[10] http://www.sas.com/en_in/insights/big-data/what-is-big-data.html

[11] Hindawi Publishing Corporation, The Scientific World Journal, Volume 2014,Article ID 712826, 18 pages. http://dx.doi.org/10.1155/2014/712826. http://www.researchgate.net/publication/256082290_Addressing_Big_Data_Issuesin_Scientific_Data_Infrastructure

[12] Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, Volume 275, Issue null, Pages 314-347 C.1L. Philip Chen, Chun-Yang Zhang, http://www.sciencedirect.com/science/article/pii/S0020025514000346