DATA ANALYTICS USED IN THE FIELD OF BIG DATA SECURITY INTELLIGENCE

M.GEETHANJALI[1]
Assistant.Professor,
Dept. Computer Science
St. Joseph's College of Arts & Science for Women, Hosur

STUDENTS:
M.RABUNI[2]
Dept. Computer Science
St. Joseph's College of Arts & Science for Women, Hosur

L.PRIYADHARSHINI[3]
Dept. Computer Science
St. Joseph's College of Arts & Science for Women, Hosur

ABSTRACT:

The term Big Data refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies. The acceleration in the production of information has created a need for new technologies to analyze massive data sets. Data is growing because of use of internet, smart phone and social network. Big data is a collection of data sets which is very large in size as well as complex. Generally size of the data is Petabyte and Exabyte. Big data analytics provide new ways for businesses and government to analyze unstructured data. Now a day, Big data is one of the most talked topic in IT industry. It is going to play important role in future. Big data changes the way that data is managed and used. Some of the applications are in areas such as healthcare, traffic management, banking, retail, education and so on. Organizations are becoming more flexible and more open. The present paper highlights important concepts of Big Data. In this write up we discuss various aspects of big data. We define Big Data and discuss the parameters along which Big Data is defined.

Keywords: Exabyte, Petabyte, Veracity, Database, Velocity, Data Analytics, Big data, Traditional approach, Collaborative approach, Structured and Unstructured data.

1. INTRODUCTION

Big data is a collective term referring to data that is so **large** and **complex** that it exceeds the processing capability of conventional data management systems and software techniques. However with big data come big values. Data becomes **big data** when individual data stops mattering and only a large collection of it or analyses derived from it are of value. The term Big Data appeared for the first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey having the title Big Data and the Next Wave of Infra Stress. The first book mentioning Big Data is a data mining book that came to fore in 1998 too by Weiss and Indrukya. The first academic paper having the word Big Data in the title appeared in the year 2000 in a paper by Diebold. The era of Big Data has bought with it a plethora of opportunities for the advancement of science, improvement of health care, promotion of economic growth, enhancement of education system and more ways of social interaction and entertainment. But as is said everything has its flip side as well big data too has its issues. Security and privacy

are great issues in big data due to its huge volume, high velocity, large variety like large scale cloud infrastructure, variety in data sources and formats, data acquisition of streaming data, inter cloud migration and others. The use of large scale cloud infrastructure having a varied number of software platforms across large networks of computers increases the region of attack to an all new level of the entire system. The various challenges related to big data and cloud computing and its security and privacy issues and the reasons why they crop up are explained later in details.

2. BIG DATA ANALYTICS

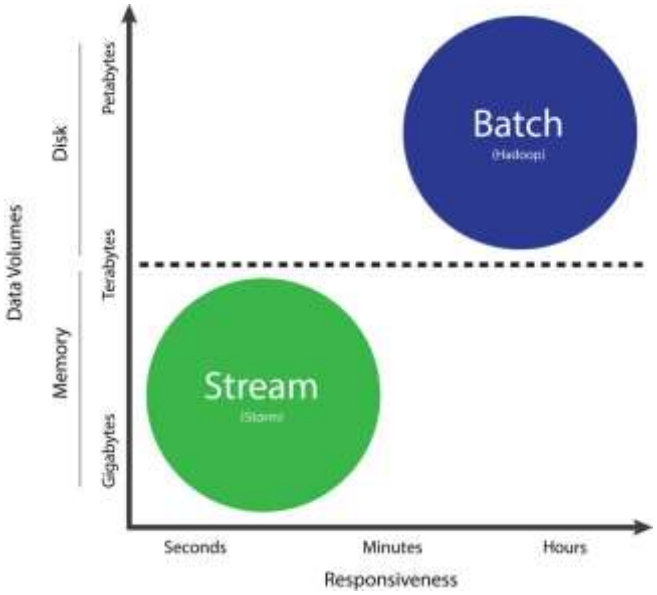
The analysis of big data is very essential aspect now a day. Big Data analytics can produce operational and business knowledge at an unprecedented scale and specificity .The need to analyze and leverage trend data collected by Businesses is one of the main drivers for Big Data analysis tools. The technological advances in storage, processing, and Analysis of Big Data include:

(a) The rapidly decreasing Cost of storage and CPU power in recent years;

(b) The flexibility and cost-effectiveness of datacenters and cloud computing for elastic computation and storage and

(c) The development of new frameworks such as Hadoop, which allow users to take advantage of these, distributed computing systems storing large quantities of data through flexible parallel processing.

Hence, by using this approach, the traditional approach is now a day's no longer used.



(a) Batch and stream processing

Hadoop is one of the most popular technologies for batch processing. The Hadoop framework provides developers with the Hadoop Distributed File System for storing large files Figure (a), which is tailored for frequently occurring large-scale data processing problems that can be distributed and parallelized.

3. ISSUES AND CHALLENGES IN BIG DATA

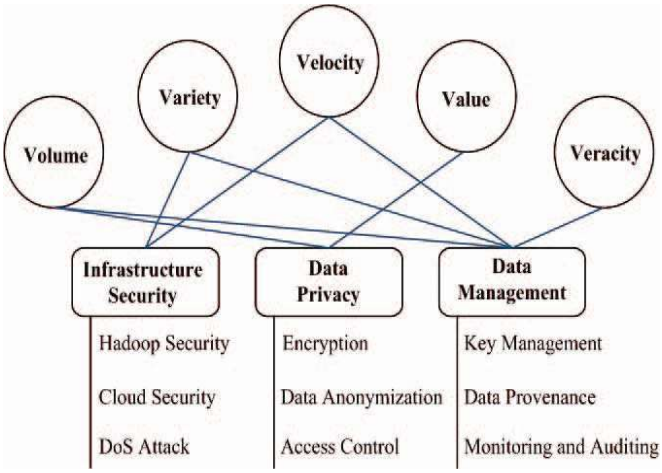
Big Data Issues and Challenges Related to Characteristics of Big Data

- **Data volume:** When data volume is thought of the very first issue that occurs is storage. As data volume increases so the amount of space required to store data efficiently also increases. Not only that the huge volumes of data needs to be retrieved at a fast speed to extract results from them. Networking, bandwidth, cost of storing like in-house versus cloud storing are other areas to be looked after. With the increase in volume of data the value of data records tend to decrease in proportion to age, type, richness and quality. Such volumes of data are difficult to be handled using existing traditional databases.

- **Data velocity:** Computer systems are creating more and more data, both operational and analytical at increasing speeds and the number of consumers of that data are growing. People want all of the data and they

want it as soon as possible leading to what is trending as high-velocity data. High velocity data can mean millions of rows of data per second. Data generated by both devices and actions of human beings like log files, website click stream data like in E-commerce, twitter feeds can't be collected because the state of the art technology can't handle that data.

- **Data variety:** Big data comes in many a form like messages, updates and images in social media sites, GPS signals from sensors and cell phones and a whole lot more. Many of these sources of big data are virtually new or rather as old as the networking sites themselves. Smart phones and other mobiles devices can be bracketed in the same category. As these devices are ubiquitous the traditional databases that store most corporate information until recently are found to be ill suited to these data. Much of these data are unstructured and unwieldy and noisy which requires rigorous technique for decision making based on the data.



(b) Security challenges in big data

- **Data value:** Data are stored by different organizations to gain insights from them and use them for analytics for business intelligence. This storing produces a gap between the business leaders and the IT professionals. The business leaders are concerned with adding value to their business and obtaining profits from it. More the data more are the insights. This however doesn't go well with the IT professionals as they have to deal with the technicalities related to storing and processing the huge amounts of data.

Big Data Technical Issues and Challenges

- **Fault Tolerance:** With the advent of technologies like cloud computing the aim must remain such that whenever failure occurs the damage done must occur within acceptable threshold rather than the entire work requiring to be redone. Fault-tolerant computing is tedious and requires extremely complex algorithms. A foolproof, cent percent reliable fault tolerant machine or software is simply a far-fetched idea. To reduce the probability of failure to an acceptable level we can do:

1) **Divide the entire computation to be done into tasks** and assign these tasks to different nodes for computation.

2) **Keep a node as a supervising node** and look over all the other assigned nodes as to whether they are working properly or not. If a glitch occurs the particular task is restarted. There are however certain scenario where the entire computation can't be divided into separate tasks as a task can be recursive in nature and requires the output of the previous computation to find the present result. These tasks can't be restated in case of an error. Here checkpoints are applied to keep the state of the system at certain intervals of time so that computation can restart from the last checkpoint so recorded.

3) **Data Heterogeneity:** 80% of data in today's world are unstructured data. It encompassed almost every kind of data we produce on a daily basis like social media interaction, document sharing, fax transfers, emails, messages and a lot more. Working with unstructured data is inconvenient and expensive too. Converting these to structured data is unfeasible as well.

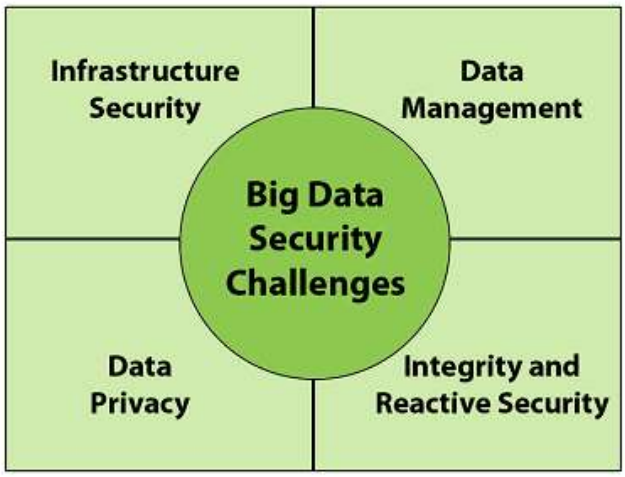
4) **Data Quality:** A storage of big data is very expensive and there is always a tiff between business leaders and IT professionals regarding the amount of data the company or the organization is storing. The quality of data is an important factor to be looked into here. There is no point in storing very large data sets that are irrelevant as better result and conclusions can't be drawn from them. Ensuring whether the amount of data is enough for a particular conclusion to be drawn or whether the data is relevant at all are further queries.

5) **Scalability:** The challenge in scalability of big data has led to cloud computing. It is capable of aggregating multiple different workloads with different performance goals into very large clusters. This needs high level of sharing of resources that is quite expensive and brings along with it various challenges like executing various jobs so that the goal of every workload is met successfully. It also has to deal with system failures in an efficient manner as it is quite common when working with large clusters. Hard disk drives being replaced by solid state drives and phase change technology do not have the same performance between sequential and random data transfer. The kind of storage device to be used is thus a large question looming around big data storage issue.

PRIVACY AND SECURITY ISSUES AND CHALLENGES WITH BIG DATA

• **Secure Computations in Distributed Programming Frameworks** Distributed programming frameworks use parallel computing and data storage for massive amounts of data. An example of this is MapReduce framework. As has been mentioned earlier MapReduce framework divides an input file into many chunks and then a mapper for each chunk reads the data, does computations and provides outputs in the form of key/value pairs. A reducer then combines the values belonging to each Unique key and outputs the results. The main concerns here are: securing the mappers and securing the data from a malicious mapper. Mappers returning incorrect results are difficult to detect and it eventually results in incorrect aggregate outputs. With very large data sets malicious mappers are too hard to be detected as well and they eventually damage essential data. Mappers leaking, intentionally or unintentionally, Private records are also an issue of concern. MapReduce computations are often subjected to replay attack, man-in-the-middle attack and denial-of-service attack Rogue data nodes can be added to a cluster, and in turn receive replicated data or deliver altered MapReduce code.

• **Security Best Practices for Non-Relational Data Stores** Non-relational databases used to store big data, mainly NoSQL databases, handle many challenges of big data analytics without concerning much over security issues. NoSQL databases consist of security embedded in the middleware and no explicit security enforcement is provided. Transactional integrity maintenance is very lax in NoSQL databases. Complex integrity constrains can't be inculcated in NoSQL databases as it hampers with its functioning of providing better performance and scalability. NoSQL databases have weak authentication techniques and weak password storage mechanisms. Authorization techniques in NoSQL provide authorization at higher layers only. It provides authorization on a per database level rather than at the level where the data are collected. NoSQL databases are subjected to inside attacks as well due to lenient security mechanisms. They may go unnoticed due to poor logging and log analysis methods along with other fundamental security mechanisms.



(c) Big data security

• **Secure Data Storage And Transaction Logs** Data and transactions logs used to be kept in multi-tiered storage media. As data size grew scalability and accessibility became an issue hence auto-tiering for big data storage came to the fore. It doesn't keep track of where the data are stored unlike in previous multi-tiered storage media where IT managers knew which data resided where and when. This gave rise to many new challenges for data security storage. Untrustworthy storage service providers often search for clues that help them correlate user activities and data sets and get to know certain properties, which can well prove vital to them. They however are not able to break into the data overcoming the encipherment. As the data owner stores the cipher text in an auto-storage system and distributes the private key to each user, he gives the right to access data of certain portions to certain users, he being unauthorized to access the data. The service provider can instigate roll back attack on users in case of a multi-user environment. He may serve outdated versions of data while the updated ones are already uploaded in the database. Data tampering and data loss resulted by malicious users often results in disputes between the data storage provider or amongst users.

• **End Point Input Validation/ Filtering**
Organizations collect data from a variety of sources including hardware devices, software applications and endpoint devices. As and when collecting these data, validation of the data as well as the source is a challenge. Often mischievous users tamper with the device from where the data are collected or tamper with the data collecting application installed in the device so that malicious data gets input into the central data collecting system. Fake IDs may be created by malicious users and provide malicious data as input into the central data collecting system. ID cloning attacks like Sybil attacks are predominant in a Bring Your Own Device (BYOD) scenario where a malicious user brings his own device, faked as a trusted device and provides malicious input from there into the central data collecting system. Input sources of sensory data can be manipulated as well like artificially changing the temperature from a temperature sensor and inputting malicious input into the temperature collection process. GPS signals can be manipulated much the same way. The malicious user may change data while it is in transmission from a generous source to the central data collection system. It's a man-in-the middle attack in a sense.

• **Real-Time Security Monitoring**
Real-time security monitoring has been an ongoing challenge in the big data analysis scenario mainly due to the number of alerts generated by security devices. Security monitoring requires that the Big Data infrastructure or platform be inherently secure. Threats to a Big Data infrastructure include rogue admin access to applications or nodes, (web) application threats, and eavesdropping on the line. Infrastructure which is mostly an ecosystem of different components, the security of each component and the security integration of the components must be considered. In case of a Hadoop cluster run in a public cloud the security of the public cloud, itself being an ecosystem of components consisting of computing, storage and network components, needs to be considered. The security of the Hadoop cluster, the security of the nodes, the interconnection among the nodes and the security of the data stored in a node needs to be considered. The security of the monitoring application including applicable correlation rules that should follow secure coding principles, must be considered as well. The security of the input source from where the data comes from too must be taken into account.

(d) Security and privacy issues

1.	Infrastructure security	• Secure computations in distributed programming • Security best practices for non-relational data stores
2.	Data Privacy	• Privacy-preserving data mining and analytics • Cryptographically enforced data centric security • Granular access control
3.	Data Management	• Secure data storage and transactions logs • Granular audits • Data provenance and verification
4.	Integrity/Reactive security	• End-point input validation and filtering • Real-time security monitoring

• Scalable and Composable Privacy-Preserving Data Mining and Analytics

Big data are subjected to appropriation of privacy, invasive marketing, reduction of civil liberty and increase in state and corporate control. An employee of a company in charge of the big data store can misuse his power and violate privacy policies. For example: He can stalk people by monitoring through chats, if the company is a social networking one that facilitates chatting. An untrustworthy business partner can infiltrate into private information and take it up into the cloud as cloud infrastructure is handled by the owners of data.

• Cryptographically Enforced Data-Centric Security

There exist two fundamental approaches of controlling visibility of data to individuals, organizations and systems. The first one being restricting access to underlying systems like operating systems or hypervisor. The second is encapsulating the data itself in a protective shell by virtue of cryptography. There are many attacks like buffer overflow and privilege escalation attack that bypass access control implementations and access the data. Protecting data end-to-end by encryption provides a much smaller well-defined attacking surface. Various threats associated with cryptographically enforced access control method using encryption are: It should not be identifiable by the adversary, the corresponding plaintext data looking at the cipher text even if he has to choose between a correct and an incorrect plain text. The cryptographic protocol must also ensure that adversary must not be able to forge data that came from the claimed source for this may well be false hence affecting integrity of data.

• Granular audits

Real-time security monitoring notification at the very moment an attack takes place is a real challenge. There may often be new attacks or missed true positives. In order to discover a missed attack audit information is required. Audit information from any device must be complete or rather it must give us details about what exactly happened and what went wrong. It must give timely access, so that it serves the purpose of Getaneh Berie Tarekegn and Yirga Yayeh Munaye.

FUTURE SCOPE AND DEVELOPMENT:

As far as the future of big data is concerned it is for certain that data volumes will continue to grow and the prime reason for that would be the drastic increment in the number of hand held devices and internet connected devices, which is expected to grow in an exponential order. SQL will remain as the standard for data analysis and Spark, which is emerging, will emerge as the complimentary tool for data analysis. Tools for analysis without the presence of an analyst are set to take over, with Microsoft and Sales force both recently announcing features letting non-coders to create apps for viewing business data. As per IDC half of all business analytics software will include intelligence where it is needed by 2020. In other words it can be said that prescriptive analytics will be built into business software. Programs like Kafka and Spark will enable users to make decisions in real time. Machine learning will have a far bigger role to play for data preparation and predictive analysis in businesses in the coming days. Privacy and security challenges related to big data will grow and by 2018, 50% of business

ethics violations will be related to data. Chief Data Officer will be a common sight in companies in the recent future though it is thought that it won't last long. Autonomous agents and things like robots, autonomous vehicles, virtual personal assistant and smart devices will be a huge trend in the future. Big data talent crunch as is seen these days will reduce in the coming days. The International Institute for Analytics predicts that companies will use recruiting and internal training to budding data scientists to get their own problems done. Businesses will soon be able to buy algorithms rather than program them by themselves and add their own data to it. Existing services like Algorithmic, DataXu, and Kaggle will grow in a large scale that is algorithm markets will emerge. More companies will try to derive their revenue from their data. The gap between insight and action in big data is going to reduce and more energy will be given to obtaining insights and execution rather than collecting big data. Fast and actionable data will replace big data.

CONCLUSION:

The goal of Big Data analytics for security is to obtain actionable intelligence in real time. Although Big Data analytics have significant promise, there are a number of challenges that must be overcome to realize its true potential. The following are only some of the questions that need to be addressed:

1. Data provenance: Authenticity and integrity of data used for analytics. As Big Data expands the sources of data it can use, the trustworthiness of each data source needs to be verified and the inclusion of ideas such as adversarial machine learning must be explored in order to identify maliciously inserted data.

2. Privacy: we need regulatory incentives and technical mechanisms to minimize the amount of inferences that Big Data users can make. CSA has a group dedicated to privacy in Big Data and has liaisons with NIST's Big Data working group on security and privacy. We plan to produce new guidelines and white papers exploring the technical means and the best principles for minimizing privacy invasions arising from Big Data analytics.

3. Securing Big Data stores: This document focused on using Big Data for security, but the other side of the coin is the security of Big Data. CSA has produced documents on security in Cloud Computing and also has working groups focusing on identifying the best practices for securing Big Data.

4. Human-computer interaction: Big Data might facilitate the analysis of diverse sources of data, but a human analyst still has to interpret any result. Compared to the technical mechanisms developed for efficient computation and storage, the human-computer interaction with Big Data has received less attention and this is an area that needs to grow. A good first step in this direction is the use of visualization tools to help analysts understand the data of their systems.

REFERENCES

1. *Introduction to Big Data*. (n.d.). Retrieved from www.coursera.org: https://www.coursera.org/learn/big-data-introduction
2. Guess, A. R. (2014, July 15). *The Most Common Big Data Management Issues (And Their Solutions)*. Retrieved from http://www.dataversity.net/common-big-data-management-issues-solutions /
3. Iswarya, K. (2014). "Security Issues Associated With Big Data in Cloud Computing". *1* (8).
4. McAfee, A., & Brynjolfsson, E. (2012). *"Big Data: The Management Revolution"* . Harvard Business Review.
5. Russom, P. (2013). "Managing Big Data".
6. Sing, R., & Ali, K. A. (2016). "Challenges and Security Issues in Big Data Analysis" . *5* (1).
7. Upadhye, G., & Dange, T. (2014). "Efficient Data Processing Using Hadoop". *International Journal of Computer Engineering and Technology* , 11-16.