ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Security Aspects in Big Data

S. Hima Bindu, O. Gireesha, A. Naga Sahithi, A. Mounicama

M.Tech Scholar, Department of Information Technology, Sree Vidyanikethan Engineering College, Tirupati, India.

ABSTRACT: The measure of information in world is developing step by step. Information is developing due to utilization of web, smart phone and social network. Big data is a collection of data sets which is huge in size and also complex. Generally the data size is Petabyte and Exabyte. Traditional database systems were unable to analyze, capture and store the vast amount of this data. As the internet is growing, measure of enormous information keep on growing. Big data analytics afford new ways for organizations and government to dissect unstructured data. Now a days, big data is a standout amongst the most talked point in an Information Technology industry. It will play a critical part in future. Big data changes the way that information is overseen and utilized. The various applications like, banking, education, healthcare, retail, traffic management, etc., use a big data to store huge amount of data. In this paper, we discuss the characteristics of big data and security issues within it.

KEYWORDS: Big data, Privacy, security, Volume, Velocity.

I. INTRODUCTION

Now a days big data [1] is a buzz phrase in Information Technology. Joined with virtualization and cloud computing, big data is a technological capability that will drive data centres to essentially change and develop within the subsequent five years. Like virtualization, big data infrastructure is inimitable and can generate an architectural disruption in the manner systems, storage, and software infrastructure are joined and administered. Contrast to the preceding trade analytics solutions, the real-time competence of novel big data solutions provide mission critical business intelligence so as to alter the structure and speed of enterprise decision making without end. Therefore, the way in which IT infrastructure is associated and spread warrants a new and serious scrutiny.

Number of technological innovations drives the theatrical raise in data and data congregation [2], [3]. Because of this reason big data become a current area of tactical asset for IT organizations. For instance, the augment of mobile users increased enterprise conglomeration of user statistics—geographic, sensor, capability, data—that can, if properly synthesize and examine, give enormously dominant business intelligence. In accumulation, the expanded utilization of sensors for everything from activity designs, buying behaviours and real-time stock management is a primary example of the massive increase in data. This sort of information accumulates continuously and gives a one of a kind and intense open door in the event that it can be examined and followed up on rapidly. The unrecognized source of big data is Machine-to Machine interchange. The ascent of security information management (SIM) and the Security Information and Event Management (SIEM) industry is at the heart of congregation, analysing, and practically respond to event data commencing active machine log files. In real time it is able to respond, capture and analyse the data and its trends. It is clear that new advances and new types of individual correspondence are driving the big data trend, regard as the global Internet population increased by 6.5% commencing 2010 to 2011 and now represents over two billion people. This may be large, but suggests the enormous majority of the world's population has to connect. However, one cannot ever attain 100% of the world's populace online (because of asset imperatives, cost of goods, and limits to material flexibility), increasingly those that are online are further associated than ever. Before, it was reasonable to believe that numerous had a desktop (maybe at work) and perhaps a portable PC available to them. Conversely, now a days all are connected to Smartphone and even a tablet computing device. So, at present two billion connected people, among them many are attached to the enormous preponderance for their waking hours, each second generating data:

- In 2011 only, mankind frame d over 1.2 trillion Gigabytes about information.
- Data volume is expanded up to 50 times within the year 2020.
- Within a minute, Google acquire more than 2,000,000 search queries.
- In every minute, 72 hours of video are added to YouTube.
- At present 217 new mobile Internet users in every minute.
- Users of the twitter posts 100,000 of tweets within a minute (that's over 140 million every day).
- Companies, brands, and various organizations accept 34,000 "likes" on social networks in each and every minute.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

What Is Big Data

Big data refers to the collection and successive examination of any drastically great collection of data contains concealed insight or intelligence (e.g. user data, sensor data, and machine data) [8]. When analyzed appropriately, big data can delivers novel business insight, open latest markets, as well as generate competitive advantages.

1.1. What Does Big Data Mean to IT?

It is determined through a combination of technology innovations, growing open source software, commodity hardware, pervasive social networking, and persistent mobile devices; the increase of big data has formed a variation point building real-time data gathering and analysis mission critical for businesses today. However, the data and its structures are basically different; it is increasingly evident that the infrastructure, tools, and architectures to support real-time analysis and understanding from this information also should be distinctive. In an IT, big data mirror is the escalation in content and information source, and also the pervasiveness of innovation in our everyday lives. As more and more of what we do is both associated with and frequently engaged by a system and the devices that we connect to are themselves powered by a variety of sensors—we ought to expect that the progressing stream of data will grow. Within data centers, each hub (servers, storage, and applications) produces an enormous number of log files and isolated data streams that also can be gathered, grouped, and investigated. With storage costs dropping, the expense connected with sparing and utilizing even the most ordinary information turns into a nonissue.

a. Characteristics of Big Data:

Big Data is becoming a prominent term now a days. The term is utilized to portray the exponential development and accessibility of information. This incorporates organized, semi-organized and unstructured information. Huge information is imperative for the business since more information results into more exact examination. More precise examination results into better basic leadership. Better choice results into: better operational efficiencies, cost decreases and lessened danger. This implies the organization takes home more income.

In 2001 the Meta Group previously renowned Big Data by 3 V's: Volume, Velocity and Variety. At present those prolonged up to 7 V's, includes Validity, Veracity, Volatility and Value:

b. Volume:

Big data imply massive volume of data. It refers to the size of data that creates from various sources include text, social networking, medical data, space images, audio, video, weather forecasting, research studies, crime reports and natural disasters, etc. It is defined as the vast amount of data generates in every second. It encompasses Terabytes and Petabytes of the storage system meant for enterprises. As the database grows the application and architecture build to sustain the data requires re-evaluating.

c. Velocity:

It is defined as the speed in which new data is generated plus the information moves around. In each and every minute one can transfer hundred hours of video on you tube, send two hundred million emails and three hundred and thousand tweets. This type of data can be generated with the usage of World Wide Web and Smartphone's. Big data streams at very speed. Therefore, it must be doubtless in a timely manner. The speed of the data could be highly inconsistent. This is specially chilled in social media when something trends.

e. Variety:

It refers to the data that comes in different formats. For example, structured data resides in relational database. Unstructured data includes text documents, emails, video, audio, log files, etc. Semi-structured data is a combination of both structured and unstructured data. The big data has to be connected and correlated during the analysis phase in order to extract the useful information. The Variety of data straightforwardly affects the integrity of data. Whenever new applications are introduced then the new data formats exists.

f. Validity:

The input data is valid when the correct processing of data gives precise results. Validity of data is close to veracity of data. Through big data, one should be spare attentive regarding validity. For instance, in healthcare, data collected from clinical trial might relate to a patient's syndrome symptoms. However, a physician treats the person considering the clinical trial results as valid.

g. Veracity:

It deals with the quality and source of the data to ascertain whether it is conflicting or improve and trustworthy. When compared to volume and speed, veracity is the major challenge in data analysis. Veracity is of utmost concern to the big

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

data processing. The trusted data does not contain duplicates. Hence, it is crucial to consider cleansing of big data with tools and algorithms.

h. Volatility:

Within Big Data environment, volatility is familiar for data to modify constantly, and if it is not accounts for analytical results perhaps invalid the instant they are produced. This kind of state is true in industries for instance stock market or a telecom company (Call data records related to one day). Volatility is associated with the challenges in validity and veracity.

i. Value:

It refers that might be extract from definite data and how big data techniques increase the value. It is the result of big data processing. Value is a vital facet in the big data. It is costly to implement IT infrastructure systems to accumulate big data and businesses entail a Return on Investment (ROI).

II. BIG DATA EXAMPLES

The example of Big Data [10] includes E-Commerce and consumer marketing, Healthcare and engineering and Manufacturing.

a. E-Commerce and Consumer Marketing:

The problem of big data is storing, acting and managing on sentiment articulated by consumers consider a brand like Children's Tylenol or Toyota automobiles. Consumers when considers a brand can have millions of interactions with the electronic. This interaction may be directly with the company or through email to company or open source media, for instance Facebook or Twitter. In terms of volume, velocity and speed the electronic communication becomes a challenge. It requires faster and great decisive action

b. Healthcare:

The origin for both volume and value is human bodies and medical devices can sustain the mobility, health and safety. The intelligent electronic devices used by people at home and some travel with them capture and transmit data analysis to manage chronic diseases and conditions, deals with sleep disorders. The data about patient diseases increase the ability to have good clinical decisions.

c. Engineering and Manufacturing:

Electronic communication through consumers has many types of data present in larger volumes. For instance, the sensor devices in engineering and manufacturing processes collect enormous volumes of data. In a day, hundreds of sensors are used in manufacturing of car and for each second it generates data about the system of vehicle when it is in use. When the car is maintained and analysed to diagnose problems, data is stored in car.

III. BIG DATA SECURITY

Security and privacy concerns [9] are growing as big data becomes more and more accessible. The gathering and conglomeration of huge amounts of heterogeneous information are presently conceivable. Extensive scale information sharing is getting to be standard amongst researchers, clinicians, organizations, administrative offices, and citizens. However, the tools and technologies are developed to supervise these enormous data sets are regularly not designed to integrate sufficient security or privacy actions, in part as we lack adequate training and a elemental understanding of how to offer large-scale data security and privacy. We too lack satisfactory policies to guarantee compliance with present approach to security and privacy. Furthermore, presented innovative methodologies for security and privacy are progressively being breached, whether fortuitously or purposely, so necessitate the consistent reassessment plus upgrading of current methodologies to prevent data leakage.

The biggest challenge for big data from a security perspective is the protection of user's privacy. Enormous information as often as possible contains immense measures of individual identifiable information and therefore privacy of users is a huge concern.

3.1. Big Data Security and Privacy Challenges:

a. Secure Computations in Distributed Programming Framework (DPF):

To practice a vast amount of data, DPF (Distributed Programming Frameworks) exploit analogous computations. DPF (Distributed Programming Framework) make use of parallelism in computations with storage space to practice enormous amount of data. Let's consider an example of Map Reduce framework.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

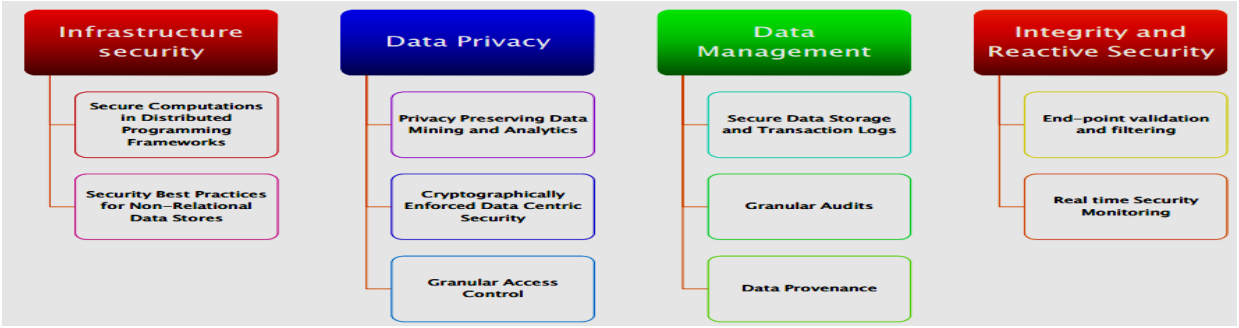

Figure 1: Top 10 Big Data Security and Privacy Challenges 4.

It consists of two phases. One is called the Mapper phase, which split an input file into various chunk. This phase stores key – value pairs, read the data and achieve some estimation. Another is Reducer phase, combine the values belong to every different key and output the end result. Attack prevention measures are majorly two types: securing the etiquette and securing the information in the occurrence of an untrusted way.

Fig. 2 Secure Computation in Distributed Programming Frameworks

b. Security Finest Practice for Non-Relational Data Stores:

Data stores in favour of non-relational data may acquaint security challenges due to their absence of capacity. Companies deal with huge unstructured data sets might assistance via migrates from a conventional relational database (RDB) to a NoSQL database. This includes the following scenarios:

1. Inefficient authorization mechanisms
2. Insider assaults
3. Lack of consistency
4. Lax authentication mechanisms
5. Susceptibility to injection assaults
6. Transactional Integrity
7. NoSQL DBMS are extremely scalable, but need identical interfaces.

Figure 3: Security Best Practices for Non – Relational Data Stores.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

c. Secure Data Storage and Transaction Logs:
While the data and transaction logs be capable of stored and managed across different storage media physically, the amount of such data means to facilitate computerized solutions are fetching more customary. Consequently such computerized solutions could not trail wherever the data is essentially stored introduce challenges within the application of security. For example, data is not often used to store on cheaper storage, on the other hand but this cheaper tier do not contain the equal security controls and the data are susceptible, subsequently a risk is introduced. Subsequently organizations ought to guarantee so as to their storage approach not merely consider the reclamation rate used for such data, excluding the warmth of data.

Figure 4: Secure Data Storage and Transaction Logs.

d. End Point Input Validation/Filtering:
A Big Data accomplishment is expected to assemble data from a large amount of sources but the confront will be attributing the level of trust associated with the data provide from such sources. Many big data use cases in enterprise setting necessitate information gathering from a few sources, like end-point devices. For instance, a SIEM accumulate event logs from millions of hardware devices and programming applications in an enterprise network. Data accumulate since a diversity of dissimilar sources. Thus it becomes fundamental to confirm the data itself along with the basis of data. What is the level of conviction of the data, what mechanism to utilize to confirm the foundation of the data is not spiteful, and how to eliminate spiteful data as of the accumulated data is a key dispute?

Figure 5: End Point Input Validation/Filtering.

e. Real –Time Security/Compliance Monitoring:
A key use case meant for Big Data is its capability on the way to aid within the security of further systems. This particular case includes mutually the monitoring of the Big Data infrastructure with by means of this equivalent infrastructure used for security monitoring. It is not ample to guard the infrastructure of the Big Data; nevertheless it is too vital to control Big Data analytics during civilizing the defense of other systems.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Figure 6: Real –Time Security Monitoring.

f. Scalable and Compassable Privacy-Preserving Data Mining And Analytics:

Boyd and Crawford [2] said, Big Data preserve potentially facilitate invasion of seclusion, enveloping marketing, decreased communal liberties, along with enlarged affirm and corporate control. There are chances in an attempt to that untrusted forecaster or business partner may extract private information of the consumers. So it becomes crucial to apply privacy preserving mining algorithms to evade confidentiality disclosure.

Figure 7: Scalable and Compostable Privacy-Preserving Data Mining and Analytics.

g. Cryptographically Enforced Access Control and Secure Communication:

There are two approaches to control the visibility of data to diverse entities, like systems, individuals and organizations. The first approach organizes the visibility of data through restrictive access to the principal system, for instance the OS (Operating System) or hypervisor. The second approach encapsulates the data itself in a defensive shell by means of cryptography. Both approaches enclose their remuneration and detriments. In the past, the first approach is simpler to execute and, as united among cryptographically-protected communication, be the standard on behalf of the preponderance of computing and communication infrastructure.

Figure 8: Cryptographically Enforced Access Control and Secure Communication.

h. Granular Access Control:

Enforcing the need-to-know standard is a significant foundation to achieve confidentiality in the data. The security assets that matter from the viewpoint of access control be secrecy-preventing access to information via people that must

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

not enclose access. The crisis with course-grained access mechanism is that information might otherwise be shared is regularly swept into a more preventive classification to ensure sound security granular access control gives data managers a scalpel rather than a sword to share data as much as possible without compromising secrecy.

Figure 9: Granular Access Control.

i. Granular Audits:

With real time security monitoring, we endeavor to notify the moment an assault takes place. In reality, this will not always be the case (e.g., latest attacks, missed true positives). So as to acquire the bottom of the missed attack, we require review data.

This is not only relevant because we need to comprehend what happened and what turned out badly, but also because compliance, regulation and forensics reasons in such manner, auditing is not something new, but the scope and granularity may be distinctive. For instance, we need to manage more data objects, which most likely are (however not so much) distributed.

Figure 10: Granular Audits.

j. Data Provenance:

Provenance is premeditated broadly in the earlier period in arts, literary works, scripts etc. Data provenance is defined as the extraction and descent of the data. It stores possession and process history with respect to data objects. The origin of data is vital for

- Auditing,
- Debugging,
- Validating,
- Evaluating the excellence of data and
- Determining reliability of data.

Provenance is naturally considering through the database, workflows and DS (Distributed System) communities. Users have to know about source of the data to guarantee its legitimacy meant for significant predictive activities

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Provenance metadata will mature within complexity attributable to great provenance graphs generate from provenance-enabled programming environments in big data application. Analysis of provenance graphs detects metadata dependency for security and/or confidentiality applications are computationally intensive.

Figure: 11 Data Provenance.

IV. CONCLUSION

Big Data change the way we perceive the world. The impact of big data has created and will continue to create can ripple through all facets of our life. Big data have various challenges related to security like-computation in distributed programming, security of data storage and transaction log, input filtering from client, scalable data mining and analytics, access control and secure communication.

REFERENCES

1. Kalyani Shirudkar, Dilip Motwani, March 2015, "Big-Data Security", International Journal of Advanced Research in Computer Science and Software Engineering, 2015; 5.
2. L Okman, N Gal-Oz, et al. Security Issues in NoSQL Databases" in Trust Com IEEE Conference on International Conference on Trust, Security and Privacy in Computing and Communications, 201; 541-547.
3. D Boyd, K Crawford, 10, 2012, Criticial Questions for Big Data, in Information, Communication & Society, 2012; 15: 662-675.
4. Cloud Security Alliance Top Ten Big Data Security and Privacy Challenges by CSA Big Data Working Group.
5. Y Yang, Z Xianghan, Type Based Keyword Search For Securing Big Data, International Conference On Cloud Computing And Big Data.2013.
6. H Xueli, D Xiaojiang, Achieving Big Data Privacy via Hybrid Cloud" in 2014 IEEE INFOCOM workshops: 2014 IEEE INFOCOM workshop on security and privacy in Big Data, 2014.
7. Min-Sheng Lin, Chien-Yi Chiu, et al. "Malicious URL Filtering-A Big Data Application" IEEE International Conference on Big Data.2013.
8. Roger Schell, Security –A Big Question for Big Data" in 2013 IEEE International Conference on Big Data, 2013.
9. M Katina, M Keith, 2013, Big Data: New Opportunities and New Challenges Published by the IEEE Computer Society 0018-9162/13/$31.00 © 2013 IEEE.
10. W Richard, Big Data: Business Opportunities, Requirements and Oracle's Approach, December 2011.