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. 3, Issue 12, December 2015

A Study on Security and Privacy in Big Data Processing

C.Yosepu¸ P Srinivasulu¸ Bathala Subbarayudu

Assistant Professor, Dept of CSE, St.Martin's Engineering College, Hyderabad, India

Assistant Professor, Dept of IT, St.Martin's Engineering College, Hyderabad, India

Assistant Professor, Dept of IT, St.Martin's Engineering College, Hyderabad, India

ABSTRACT : Big data refers to huge amount of digital information collected from multiple and different sources. Since a key point of big data is to access data from multiple and different domains security and privacy will play an important role in big data research and technology. Traditional security mechanisms, which are used to secure small-scale static data, are inadequate. So the question is which security and privacy technology is adequate for efficient access to big data. In this paper, we focused on big data specific security and privacy challenges. Main expectation from the focused challenges is that it will bring a novel focus on the big data infrastructure.

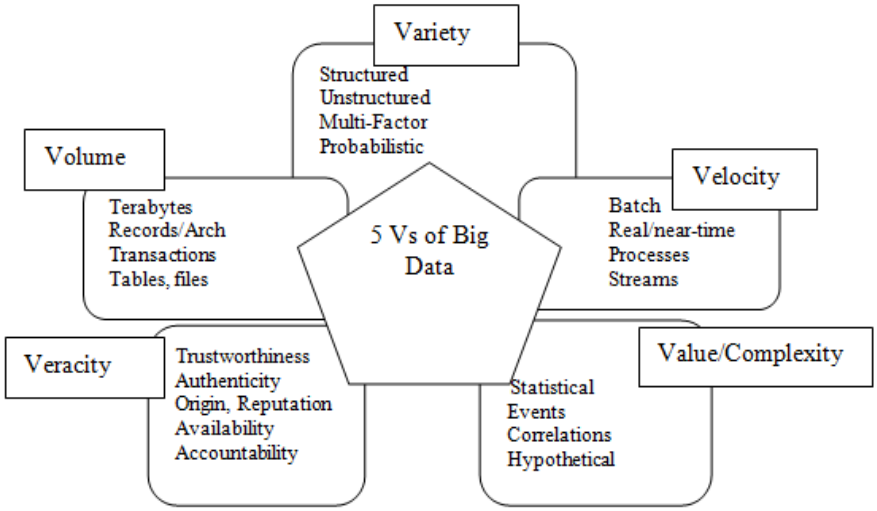
KEYWORDS: Big data, security, privacy

I. INTRODUCTION

The term big data refers to the huge amount of data organization and government collect about us and our surroundings. Big data "size" is constantly growing because every day we create quintillion bytes of data. At present 90% of the data in the world, has been created in the last two years only. It becomes complex to process big data using traditional data processing applications. With advanced big data analyzing technologies, we can make efficient decisions for critical development areas such as economic productivity, healthcare, energy and natural disaster production.

II. CHARACTERISTICS OF BIG DATA

The big data can be described by the following properties:



1. Volume: The volume of big data is very important in the context. Many factors contribute to e increase in data volume online transactions data, live streaming data from social media, customer feedback, data produced by

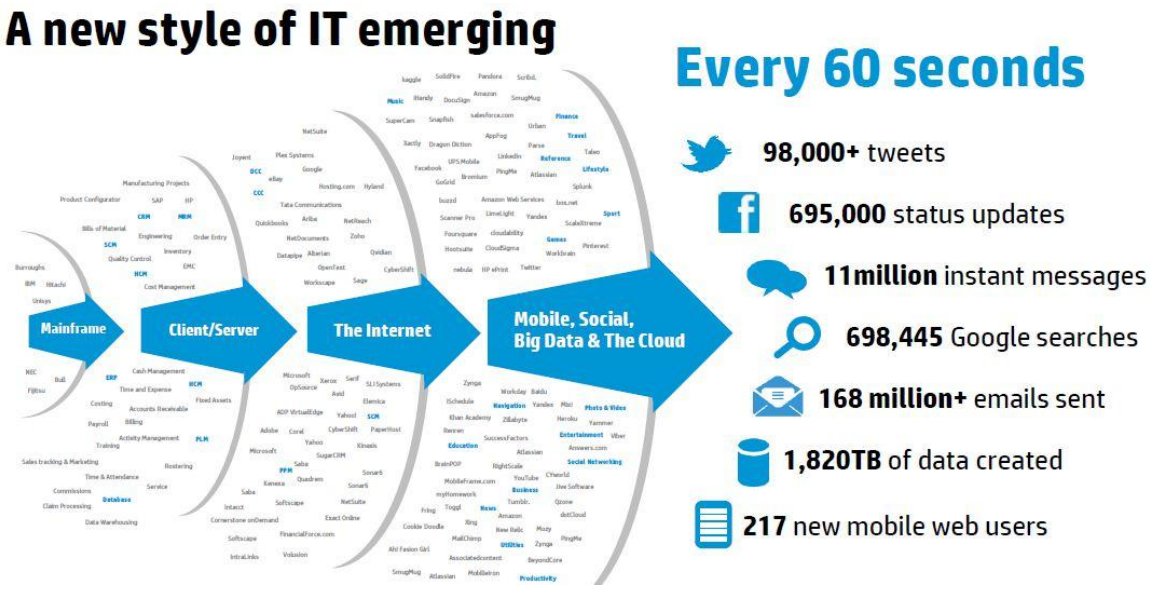
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

employee, contractors, partners, and suppliers using social networking sites and data collected from sensors etc.. In the past, excessive data volume was a storage issue. Other issues are how to determine relevant data within large data volumes and how to use big data analytics to generate value from relevant data.



2. Variety: The next characteristic of big data is variety. Today data available in different types of formats(ex. Structural Data and Unstructured data). Structural data such as numeric data in traditional data bases and information created from business applications unstructured data such as text documents, email, video, audio, online transaction. Merging managing and accessing different varieties of data is not easy task still.
3. Velocity: The term velocity means how fast the data is being produced and how fast the data needs to be processed to meet the demand and challenges. Data velocity is a challenge for most enterprises.
4. Variability: The term variability of big data refers to inconsistency of data. Along with the velocity and varieties of data, data flows can be highly inconsistent with periodic peaks.
5. Complexity: Complexity of data needs to be considered, especially when large amount of data come from multiple sources. The data must be cleaned ,merged, matched and transformed into required format before actual processing.

III. NEED OF SECURITY AND PRIVACY IN BIG DATA

Many of the organizations, uses big data, but may not have the efficient mechanism, particularly from a security perspective if a security problem occurs to big data, it causes more serious legal consequences and reputational damage than at present. In the present era, many organizations are using traditional security mechanisms which are used to secure small scale static data, are inadequate. To provide security for big data, techniques such as encryption, logging honey pot detection must be necessary. Security and privacy issues are magnified by volume; variety and velocity of big data, such as large scale, cloud infrastructure, multiple sources and different formats, streaming nature of data acquisition and high volume inter cloud migration.

IV. SECURITY AND PRIVACY CHALLENGES IN BIG DATA

In this paper we focused on the big data security and privacy challenges. We studied survival security practitioner oriental trade journals to focus an initial list of high- priority security and privacy problems and arrived at the following top ten challenges.

1. Secure computations in distributed programming frameworks
2. Security best practices for non-relational data stores
3. Secure data storage and transactions logs

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. 3, Issue 12, December 2015

4. End-point input validation/filtering
5. Real-time security monitoring
6. Scalable and composable privacy-preserving data mining and analytics
7. Cryptographically enforced data centric security
8. Granular access control
9. Granular audits
10. Data provenance

In the rest of the paper, we provide brief description of security and privacy challenges

1. Secure computations in distributed programming frameworks

Distributed programming frameworks uses parallelism in computation and storage to process huge amount of data. MapReduce is an example of distributed programming framework. It is used for data intensive computation in a large cluster environment. MapReduce has become popular for analyzing the large data sets. It provides a simple programming framework and is accountable for distributed execution of computation, fault tolerance, and load balancing. However, many relational data based applications need parsing the relational data iteratively and need to operate on these data through many iterations. MapReduce do not have built-in support for the iterative process. A new framework iMapReduce, which supports iterative processing. iMapReduce permit users to specify the iterative operations with map and reduce functions, and it support the iterative processing automatically without the need of users' involvement.

2. Security best practices for non-relational data storage

Non relational data stores are popularized by NOSQL databases are still developing with respect to security infrastructure. NoSQL is a database used to store huge amount of data. NoSQL databases are distributed, open source, non-relational and are horizontally scalable. NoSQL does not follow property of ACID as we follow in SQL. NoSQL is horizontally scalable which leads to high performance in a linear way. It is having more flexible structure. NoSQL databases are desirable and popular among Web--- based companies, due to their demonstrated advantages in data flexibility, scalability and performance. Security issues of NoSQL in general remain to be improved. There are only a few NoSQL (e.g., Cassandra) that currently meet the data security requirements of PCI---DSS, e.g., data--- at---rest and data---in---motion. However, enhanced security is expected to come at the expense of performance.

3. Secure Data Storage and Transactions Logs

Secure Data storage and transaction logs are stored in multi-tiered storage devices. Manually moving data between tiers helps the IT manager to control what data is moved and when. However, as the volume of data set continue to increase exponentially, availability and scalability have necessitated auto-tiering for Big Data storage management. Auto-tiering solutions do not maintain information about where the data is stored, which creates new challenges to secure data storage. New mechanisms are imperative to prevent unauthorized access and maintain constant availability.

4.End-Point Input Validation/Filtering

Many Big Data uses in organization settings require data collection from multiple sources, including end-point devices. For example, a security information system may gather event logs from millions of software applications and hardware devices in an enterprise network. A key challenge in the data collection process is input validation: how can we trust the data? How can we confirm that a source of input data is not harmful? And how can we filter harmful input from our collection? Validation and filtering of input is a daunting challenge posed by untrusted input sources, especially with the bring-your-own-device (BYOD) model.

5. Real-Time Security Monitoring

Big Data and security do not only meet the protection of Big Data infrastructures, but also at the leveraging of Big Data analytics to improve the security of other systems.

Real-time security monitoring is one of the most challenging Big Data analytics problems, which consists of two main angles: (a) monitoring the Big Data infrastructure itself (b) use the same infrastructure for data analytics. An example

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. 3, Issue 12, December 2015

of (a) is the monitoring of the performance and strength of all the nodes that construct the Big Data infrastructure. An example of (b) would be a health care provider using monitoring tools to look for false claims or a cloud provider using similar Big Data tools to get better real-time alert and compliance monitoring. This enhancement could provide a reduction in the number of false positives and/or an increase in the quality of the positives. In this paper, we focus on both angles.

Real-time security monitoring is a challenge due to the number of alerts produced by security devices. These alerts (correlated or not) lead to a huge number of false positives, which are often ignored because of limited human capacity for analysis. This problem may increase with Big Data, given the volume and velocity of data streams. However, Big Data technologies may provide an opportunity to fast process and analyze different types of data. These technologies can be used to provide, for example, real-time inconsistency detection based on scalable security analytics.

6.Scalable and Composable Privacy-Preserving Data Mining and Analytics

Big Data can potentially allow invasions of privacy, decreased civil liberties ,invasive marketing, and increased state and corporate control. A recent analysis of how organizations are leveraging data analytics for marketing purposes included an example of how a vendor was able to recognize a teen's pregnancy before her father learned of it. Similarly, anonymizing data for analytics is not adequate to maintain user privacy. For example, AOL released anonymized search logs for academic purposes, but users were easily recognized by their searches . Netflix faced a similar problem when anonymized users in their data set were recognized by correlating Netflix movie scores with IMDB scores.

7.Cryptographically Enforced Data-Centric Security

There are two approaches to control the visibility of data to different entities, such as individuals, systems and organizations.. The first approach manage the visibility of data by limiting access to the underlying system, such as the operating system. The second approach uses cryptography to encapsulates the data itself in a protective shell. Both approaches have their merits and detriments. Historically, the first approach has been simpler to implement and when combined with cryptographically-protected communication, is the standard for the majority of computing and communication infrastructure.

However, the system-based approach possibly exposes a much larger attack surface. The literature on system security is replete with attacks on the primary systems to avoid access control implementations and access the data directly.

8 Granular Access Control

The security property that matters from the view of access control is privacy – preventing access to data by unauthorized people. The problem of the course-grained access mechanisms is that data that could otherwise be shared is often swept into a more preventive category to guarantee sound security. Granular access control helps data managers more precision when sharing data, without compromising secrecy.

9.Granular Audits

With real-time security monitoring notification at the moment an attack takes place is the goal. Actually this will not be the case always (e.g., new attacks, missed true positives). In order to determine a missed attack, audit information is required. Audit information is crucial to understand what happened and what went wrong. It is also needed due to compliance, regulation and forensic investigation. Auditing is not something new, but the scope and granularity might be diverse in real-time security contexts. For example, in these perspective there are more data objects, which are probably (but not necessarily) distributed.

10 Data Provenance

Provenance metadata will increase in complexity because of large provenance graphs produced from provenance-enabled programming environments in Big Data applications. Analysis of such large provenance graphs to identify metadata dependencies for confidentiality or security applications is computationally intensive.

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. 3, Issue 12, December 2015

V CONCLUSION

In this paper, we have focused on the security and privacy problems that needs to provide more secure for Big Data processing and computing infrastructure. Common elements of Big Data arise from the use of multiple infrastructure tiers for processing Big Data. The use of new compute infrastructures such as NoSQL databases (for higher performance necessitated by Big Data volumes) that have not been thoroughly examine for security issues; the non-scalability of encryption for huge data sets; the non-scalability of real-time monitoring techniques that might be practical for small amount of data; the heterogeneity of devices that produce the data; and the confusion surrounding the different legal and policy restrictions that lead to ad hoc approaches for ensuring security and privacy. Many of the elements in this list serve to illuminate specific aspects of the attack surface of the entire Big Data processing infrastructure that should be analyzed for these threats.

REFERENCES

1. ”Cloud Security Alliance Top Ten Big Data Security And Privacy Challenges “by CSA Big Data Working Group.
2. Kalyani Shirudkar, Dilip Motwani "Big-Data Security" published in International Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 3, March 2015.
3. Venkata Narasimha Inukollu1 , Sailaja Arsi1 and Srinivasa Rao Ravuri "SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUD COMPUTING" International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014.
4. Vatika Sharma1, Meenu Dave2 " SQL and NoSQL Databases " published in International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 8, August 2012.
5. www.fidelissecurity.com/files/NDFInsightsWhitePaper.pdf.
6. www.isaca.org/Groups/Professional-English /bigdata /GroupDocuments /Big_Data_Top _ Ten_v1.pdf .

BIOGRAPHY

C.Yosepu is an Assistant Professor in the Department of Computer Science and Engineering, St.Martin's Engineering College,JNT University,Hyderabad. He received Master of Technology (Computer Science and Engineering) degree in 2013 from SSIT, JNT University,Hyderabad,Telanagana,India.His research interestes are Design and Analysis of Algorithms, Big Data,Dataware Housing and Data Mining,Computer Netwroks

P Srinivasulu is an Assistant Professor in the Department of Information Technology, St.Martin's Engineering College,JNT University,Hyderabad. He received Master of Technology (Computer Science and Engineering) degree in 2014 from AITS, JNT University,Anathapur,AP,India. His research interestes are,Computer Netwroks , Big Data and Mobile Computing.

Bathala Subbarayudu is an Assistant Professor in the Department of Information Technology,St.Martin's Engineering College,JNT University,Hyderabad. He received Master of Technology (Computer Science and Engineering) degree in 2014 from AITS, JNT University,Anathapur,AP,India. His research interestes are Network Security Computer Netwroks and Big Data.