

Data Science and Big Data Analytics for Security and Privacy Challenges

Tushar Chauhan¹ Prof. Sejal Thakkar² Prof. Pruthvi Patel³

¹M.Tech Student ^{2,3}Assistant Professor

¹Department of (Data Science) Computer Engineering ^{2,3}Department of Computer Engineering

^{1,2,3}Indus Institute of Technology & Engineering, Gujarat, India

Abstract— Here in this paper, we present a brief overview of important topic regarding the connection of security and data science, big data. Here we discuss a selection of security aspects that data scientist and big data analyst should consider to make their service and product more secure. And also discuss about applications where security plays important role in real life application this includes detailed looks about security issues related to data science and big data analysis. The Term Big Data Analytics for Security intelligence refers to a process of analyzing and mining large amounts of data. The scope of big data security is limited not only to the current data set but also historical data to identified threats, anomalies, fraud so that network can be safe from targeted attacks. Many institutions are taking steps to focus the growing problems of advanced persistent threats, attacks and fraud.

Keywords: Data Science, Security and Privacy Challenges

I. INTRODUCTION

Concept of big data analytics refers to large scale information management and analysis technologies that beyond the scope of traditional data processing technologies. Big data can be differentiated from traditional technologies in four types.

- 1) The amount of data (volume).
- 2) The types of structured and unstructured data (variety).
- 3) The rate of data generation and transmission (velocity).
- 4) The economic value of different data varies significantly (Value).

Big data analytics are being used more widely every day for so many reasons. These new methods of applying analytics certainly can bring innovative Ease of Use improvements for business. For example: retailers shoppers are using big data analytics frequently to detect items whose sale is frequent in each season and to predict geographical areas where demand will be high. The benefits of big data analytics is so great that in addition to all the positive business possibilities, there are just as many new privacy concerns being created. Traditional security technologies were not too much capable to detect and protect against fraud, cyber-attacks and threat. Smart cyber criminals can easily penetrate organization operation to get sensitive information such as intellectual property, credit card numbers, and customer database to damage the enterprise.

Big data analytics may have provided a chance to anomaly and fraud detection based on security analytics. Since data is not always transferred from one organization to another organization, the analysis may be done in a distributed environment. A large variety of data types store in diverse system so the infrastructure required for analyzing big data must be able to support deeper analytics such as data mining and statistical analysis. Collecting and storing large amount of data is one side of coin but the other side of coin is to protect massive amount of data from unauthorized access

which is more challenging. Big data is applied heavily in improving security and enabling law enforcement

Giants like Sony, Yahoo and Anthem Inc., the second-largest US health insurance company, heavily rely on big data and machine learning systems to efficiently store and process huge amounts of data. But large enterprises are not the only ones; there are more and more startups and SMEs whose business model focuses on data-centric services and products. Unfortunately, where there is valuable data, there are also hackers that want to get it or manipulate it for fun and profit. It is therefore important that data scientists are aware of the fact that new services or data products should be designed with security in mind. Many of the popular technologies and algorithms used in their domain are not secure by default. They have to be used with care. For example, recent research showed that access to the public API of a classification service (e.g., face recognition) might be sufficient to steal or invert the underlying model. We refer to these aspects as security of data science, i.e., issues related to the security of data science methods and applications.

Data science methods and techniques help to address some of the most challenging problems in this field such as the management of huge amounts of log data and the identification of anomalies or other clues that might pinpoint activities posing a risk for an organization. It is therefore not surprising that advancements in the field of data science led to improvements of existing security products. For instance, becoming better at detecting anomalies in credit card transactions, network traffic, user behavior, and other types of data directly results in improved products to protect today's businesses. However, improvements to existing products is not the only outcome of the already fruitful relation of data science and security. It also led to the development of completely new solutions such as next-generation anti-virus products.

II. SECURITY CHALLENGES IN BIG DATA ANALYSIS AND DATA SCIENCE:

Various big data analytics for security issues and privacy challenges are discussed here. In recent research threat environment big data can be differentiated from traditional technologies in four types: volume, variety, velocity and value. Security issues challenges are Amplified by velocity, variety and volume of big data such as very large-scale cloud framework, distinction of data source and pattern, cascading nature of data acquisition. These are amplifying at a rapid rate and has required a shift in how protected vendors manage threat, attacks.

- 1) Protected database storage and transaction log file
- 2) Privacy issues for non-relational data stores
- 3) End-point input validation/filtering
- 4) Extensible and pasture able privacy preserving data mining and analysis
- 5) Real-time security and compliance monitoring

- 6) Metadata provenance
- 7) Information security
- 8) Cryptographically enforced access control and secure communication

Since any data product needs an infrastructure to run on, a piece of software that implements it, data that fuels it, and customers that feel comfortable using it, we provide a brief overview and references to more in-depth material on

- 1) Infrastructure security
- 2) Software security
- 3) Data protection
- 4) Data anonymization.

III. DATA SCIENCE AND BIG DATA ANALYTICS FOR SECURITY:

After having discussed some of the security challenges a data scientist might face when developing and using modern data science technologies, this section deals with the opportunities of using data science to help solve major challenges in information security. In this context, we are looking at three general application areas:

Anomaly detection

- Malware detection and classification
- Threat detection.
- Anomaly Detection

Most frequently use drivers of big data analysis tools is to analyze and hold trend dataset collected by business purpose. Big Data analysis tools and technology are used to analytics for anomalies or fraud identify in different domain such as insurance, healthcare, credit card, net banking and so on. Big Data analytics can be used to analyze network traffic, financial transaction and log files to finding fraud and suspicious activities, and to tally multiple sources of information into a rational view. However, analyzing log files, network route and the system events for forensics and encroachment detection have traditionally been a remarkable problem

- 1) Advanced Big Data technology such as database related to Hadoop system and stream processing are validating the storage and analysis of very large mixed data at an unprecedented scale and speed. These types of technologies will transform by the security analytics:
- 2) Gathering data at very large scale from many external and internal enterprise sources such as susceptibility databases.
- 3) Performing broad analytics on the data.
- 4) Giving a buildup view of security related information and accomplishing real time analysis of streaming data.

IV. DATA SCIENCE AND BIG DATA ANALYTICS FOR PRIVACY:

Big data analytics for privacy concern holds great promise for inspiring improving upon all fields of organization, significant innovations, and bringing benefit to individual in so many ways. However, organization that choose to use big data analytics must determine the associate privacy and information security impacts before they actually put analytics into use.

- 1) Anonymization could become impossible
- 2) Bid data analytics are not complete accurate
- 3) Legal protection exists for the involved individuals
- 4) Security intelligence and Compliance audit
- 5) Unethical actions based on interpretations
- 6) Privacy breaches and fraud incidents
- 7) Discrimination

Data science analyzes data of human individuals. This data should be anonymized to make sure the privacy of the individuals is protected. Data anonymization basically means that any data record in the data set should not be easily linkable to a particular individual. Obvious solutions include stripping the real name or the detailed address of individuals from the records, but experience teaches that this is usually not enough to truly anonymize the data.

V. CONCLUSION

In this paper, we try to summarize some security and privacy related issues that need to focus for constructing big data processing and computing infrastructure extra secure. Security is now a big data problem because the data that has security context is huge. If we are ignoring some of that data or can't analyze it, big data security analyses tools are not security properly. The objective of big data analytics for security is to obtain actionable intelligence in real time. Although big data analytics have significant promises, there are a number of challenges that must be overcome to realize its true potential.

Big data analytics for security and privacy issues focus on the research challenges, leading to greater security and privacy in big data platforms. Following are some of them that need to be addressed:

- 1) Data provenance
- 2) Securing big data stores
- 3) Human computer interaction
- 4) Privacy Information security as a big data issue.

With respect to security, data science is a double-edged sword. On the one side, it offers many new opportunities and a lot of potential to significantly improve traditional security algorithms and solutions. Recent advances in challenging domains such as anomaly detection, malware detection, and threat detection underline the tremendous potential of security data science