IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 4.819

Volume 5, Issue 2, May 2021

Security of Data Science and Data Science for Security

Manish Kumar Kumawat
M.Sc. IT (Information Technology)
Sir Sitaram and Lady Shantabai Patkar College of Arts and Science, Mumbai, India

Abstract: *Security and data science are two of the fastest developing fields of IT and are being implemented more lately in different applications. In this paper, IN the first section we are going to discuss how security is important for our data, and in the second section, we are going to discuss how data science in respect to machine help in security and some of the applications of machine learning in security are listed below.*

Keywords: Security of Data Science, Data Science for Security, Information Security, machine learning.

I. Introduction

Security and computer science are two of the fastest developing fields of IT and areas related to each other. Data science combines multiple areas such as machine learning and high-performance computing and data management. The goal of data science is to analyze large-size heterogeneous data and find patterns and make predictions. Security is all about ensuring that data is not modified by an unauthorized person and controlling access to the data.

Methods and strategies in data science help solve some complicated challenges in the security area, such as the control of vast volumes of log data and the detection of irregularities or other measures which may identify behaviors that pose to an organization a danger. Consequently, development in the area of data analysis is not unexpected, leading to advancements in current security devices. For example, identifying irregularities in payment card transactions, network traffic, user activity, and other data forms leads to better products directly protecting companies today.

II. Security

Over the past half century, the information and communication technology (ICT) industry, omnibus and closely integrated into our modern society, has grown significantly. Security is known as the act of protecting ICT systems against different cyber threats or attacks. Several factors contributed to security: safety of information and communication technologies; the raw data and information it comprises and its processing and trans mission; the relevant virtual and physical elements of the systems; the standard of safety arising from the implementation of these measures. Security, according to After good et al., is a set of technologies and methods that secure computers, networks, programmes, and data from assaults and illegal access, modification, or deletion. [2].

2.1 Key Concepts of Information Security

A. CIA (Confidentiality, Integrity, Availability) of Information Security

1. Confidentiality: This is meant to provide access only to those people who are authorized to access it. It means only data is accessible by people who have permission to access the data. User ID, password, cryptography OTP, etc are examples of an attempt to ensure confidentiality of data.
2. Integrity: Integrity is meant to ensure that data is not altered or modified by an unauthorized person. One control of integrity is to block the authorized person to modify our data and restore the data to its original state. Hashing algorithms are a key process in providing integrity.
3. Availability: for any information to serve its purpose, information must available all the time when it is needed. Availability refers to the uptime of services, the assurance that services must be available all the time whenever needed.

DOI: 10.48175/568

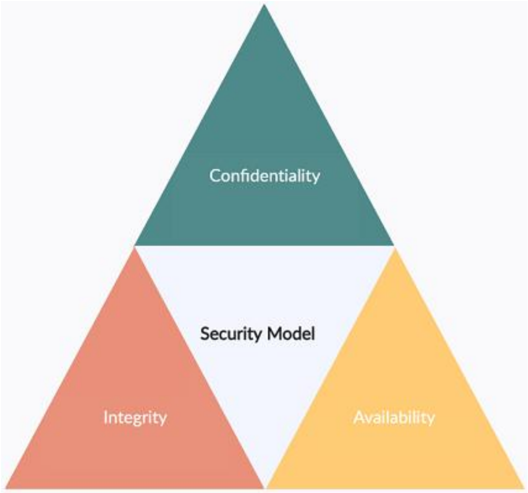

Figure 1

III. Security of Data Science

In this section, we will discuss some of the conditions where we need to secure our data and how to prevent them because in data science we mostly deal with data if our data is corrupted then the result might be different than expected and we address some of the issues and solution approaches.

Threats, i.e., who is attacking, vulnerabilities, i.e., the holes they are attacking, and impacts, i.e., what the assault does, are the three security factors that are commonly connected with any attack. [2]. A breach of the integrity, confidentiality, or availability of information and systems represents a security threat. Several incidents of the event have been turned into a security assault that is extremely destructive to the sources.

SQL injection attack: SQL injection is a weakness in web protection that enables an attacker to interact with queries made in the database of an application. It usually helps an intruder to access details they usually can't access this could involve details from other users or any other information accessible by the program itself. An attacker can in certain cases alter or erase this information, which causes the content to change in the application or the web.

If the SQL injection attacks are successful, unauthorized access to sensitive information, such as passwords, credit card information, or personal user information can occur. In recent years, many high-profile data breaches were caused by SQL injections which resulted in reputable damage and regulatory fines. In certain cases, an attacker can discover a continuous loophole in the processes of an organization, leading to a long-term compromise, which can be missed for a long period of time.

Take the following code lines, for example, where a SELECT query is being created and the user enters the input data

```
String query = "SELECT name, description from Product WHERE name LIKE '%" + userinput + "%'";
```

If the attacker built this query

```
UNION SELECT username, password FROM User—
```

Then this query is built

```
SELECT name, description from Product WHERE name LIKE '%' UNION SELECT username, password FROM User —‘%'
```

This query is syntactically correct and will return all products, as well as all usernames and passwords that are stored in the database Zhongding Dong et.al had defined a new role called “smart-driver” which is located between the database and the web portal. All types of information are sent to the user along with a random number for authorized user identity purposes and to protect their data from the attacker. The entire processing is performed by smart -driver. Whenever users access data from the database, a unique identifier is given to the user. Even though the attacker has cracked the identity number initially, the next step is protected with the newly generated random number which is issued to the user alone. Prevention is done with the help of this smart -driver[1].

3.1 Denial-of-Service (DoS)

Denial-of-Service is an attack meant to shut down a machine or network, making it inaccessible to its intended users by flooding the target with traffic that triggers a crash. The Denial -of-Service (DoS) attack typically uses one computer with an Internet connection, while distributed denial-of-service (DDoS) attack uses multiple computers and Internet connections to flood the targeted resource Usually, DoS attacks falls into multiple categories:

- Flood attacks: In this hacker send the multiple amount packets to the server unit, the server's storage capacity is full resulting in denial of servers. Developed by Cisco monitoring traffic patterns and DoS attacks is a very popular tool used by ISPs. The flow is defined as having unique attributes like source IP, Destination IP, Source port, Destination port, etc. Monitoring traffic in both directions all router interfaces must be monitored, including uplinks to the core router [8].
- Phishing: In phishing assaults, an attacker uses a psychological trick to persuade users to reveal personal information such as usernames, passwords, credit card numbers, bank account numbers, and so on. There are some common phishing attacks such as email, SMS, etc.
- Insider threat: The vulnerability danger that originates from inside the targeted company is an insider threat. It usually requires a current or former employee or business partner who is inside an organization's net work to have access to classified or privileged accounts, and who misuses this access. If an employee leaves the company, his account in Active Directory is automatically disabled since it is no longer needed. However, data accessed and stored on your computer (whether provided by the business or personal) must also be wiped clean. For e.g., you do not want a worker with disgust to also have access to corporate knowledge that is vital, sensitive, or priceless. Unauthorized access: In this unauthorized person can access sensitive or confidential information that cannot access by that person. A firewall is a general strategy for approved access to the network. Many firewall strategies are used to protect against unauthorized access. The network should also be configured to prevent unauthorized users from entering, viewing, or altering data.
- Data Protection: The collection of (large volumes of) data is a central task of data science. The data must be available in unencrypted form for most computing tasks. There are two big drawbacks to this. The first one is that criminals can simply steal the data and make use of any information it provides when authentication mechanisms such as access management fail. Data can always be kept in encrypted form to make this more difficult. This way, the attacker must steal the data when it is being processed or manage to steal the keys used to encrypt it.
- Encryption: Encryption is one of the most effective approaches to achieving data security. Encryption techniques hide the original content of the data so that the original information is only recovered using a key known as the decryption process. The goal of encryption is to protect or secure data from unauthorized access in terms of viewing or modifying the data. Encryption can be implemented using some substitute techniques, move techniques, or mathematical operations. Several symmetric key-based algorithms have been developed in the last year [3].
- Privacy Preservation: In certain scenarios, data science analyses individual data, for example, clinical data. Health data. In order to ensure that the privacy of people is secured, this data should be anonymized to answer legal and ethical obligations. Basically, data anonymization ensures that any data recorded in the collection should not be easy to link to a single user. Deletion is a fundamental way of anonymity by removing or swapping attributes with other values. The approach to blurred data is defined by replacing individual values with categories or ranges of values.

DOI: 10.48175/568

IV. Data Science for Security

In this section, we explain how data science with respect to machine help in security and some of the applications of data science in security are listed below.

Although in many instances Traditional security methods have their own merits, so much manual labor is required to keep up with the evolving security threat environment. On the contrary, data science will make a massive difference in technology and its practices, where machine learning algorithms can be used to understand or extract information from the training data for their identification and avoidance of security incident trends. For example, it is possible to use these techniques to detect ransomware or unusual patterns or to extract policy rules.

4.1 Machine Learning

Machine learning is a sub-field of artificial intelligence that aims to empower systems with the ability to use data to learn and improve without being explicitly programmed [7]. It uses mathematical models derived from data sets analysis, which are then used to make predictions on new input data. Machine training technologies include a wide variety of fields from e-commerce, in which software for machine learning is used to offer guidance on consumer behaviour, choice, and health care. Machine learning is used in forecasting epidemics or the possibility of people suffering from such illnesses, such as cancer, on the basis of their medical history.

4.2 Types of Machine Learning Algorithms

- Supervised learning
- Unsupervised Learning

A. Supervised Learning

In supervised learning, there is always a target variable, the value of which the machine learning model learns to predict using different learning algorithms e.g., based on an IP address location, frequencies of Web requests, and times of request, a machine learning model can predict if a given IP address was part of a Distributed Denial of Service (DDOS) attack. A variety of Machine learning algorithms fall under the umbrella of supervised learning, including Linear and Logistic Regression, Decision Tree, and Support Vector Machine (SVM) [7].

B. Unsupervised Learning

Unsupervised learning is used to find patterns of behavior that are very similar in the datasets. Computer systems detect such malware using clustering and correlation algorithms with similar working/behavior patterns.

4.3 Applications of Machine Learning In Security

- Threat detection and classification: Machine learning algorithms can be used in apps to detect and respond to cyber-attacks before they have a chance to take effect. [4]. This is typically done by examining large data sets of security incidents and detecting the patterns of malicious activity using a model developed. As a consequence, they are immediately dealt with when related events are observed. Usually, the training dataset of the models consists of previously defined and documented Compromise Indicators (IOC), which are then used to construct models and applications that can track, recognize and respond in real-time to threats. With the availability of IOC databases, machine learning classification algorithms can now be used to recognize the different malware activities in datasets and classify them accordingly.
- Network risk scoring: This applies to the use of predictive metrics that help organizations allocate cybersecurity services appropriately in terms of varying risk ratings for different parts of the network. Machine education can be used to simplify this method by reviewing the past data sets of cyber-attacks and assessing which network areas were mainly involved in those kinds of attacks. Using machine learning is advantageous in the sense that the resulting scores will not only be based on domain knowledge of the networks but most importantly, the scores will be data driven [4].

DOI: 10.48175/568

- Phishing Detection: In phishing attacks, an attacker used the psychological trick to manipulate users' to give their sensitive information such as usernames, password credit card details, bank account details, etc. There are some common phishing attacks such as email, SMS, etc. The use of cybersecurity tools with machine learning will prevent these phishing appeals. Emails can also be scanned by natural language processing to search whether there is any unusual information such as patterns or phrases that suggest that the email is a phishing attempt. Tessian is a trustworthy tech provider that offers email verification software to verify if an email is alleged phishing or abuse of records. The encoding of natural languages and the recognition of threats using suspicious detection technology.
- Automate routine security tasks and optimize human analysis: Repetitive tasks performed by security analysts during risk assessments can be automated by machine learning. This can be achieved by reviewing records/reports of previous steps taken by security experts to recognize and react effectively to such attacks and using that information to create a model that can identify similar attacks and respond appropriately without the presence of human beings. Machine learning can automate some aspects of analysis such as network log analysis, malware detection, network r isk analysis and automate the full process if a very difficult task. By integrating machine learning into the protection process, humans and machines will work together to do tasks that would otherwise be difficult at that same speed and efficiency.

V. Conclusion

In this paper, we discuss the relationship between data science and security and data science security Recently, data science has helped the security field to prevent multiple attacks and has the potential to grow more in the future through the use of data science in the security field. In this paper, on the other hand, there are multiple challenges in data science mostly concurs about t he security of the data, if our data is not secure and modified by the authorized person if we apply the data science method on this data then the result of data might the so we conclude that data science and security are very important to each other and with the help each other both field data science and security can grow.

References

[1] Rubidha Devi.D*, 2R.Venkatesan, 3Raghuraman.K, A Study On SQL Injection Techniques, International Journal of Pharmacy & Technology IJPT| Dec-2016 | Vol. 8 | Issue No.4 | 22405-22415 Page 22405

[2] Iqbal H. Sarker1,2*y, A. S. M. Kayes3, Shahriar Badsha4, Hamed Alqahtani5, Paul Watters3 and Alex Ng, Cybersecurity Data Science

[3] Ekta Agrawal1 , Dr. Parashu Ram Pal2 Research Scholar1 , Professor2, A Secure and Fast Approach for Encryption and Decryption of Message Communication, IJESC, Volume 7 Issue No.5

[4] Manjeet Rege, Raymond Blanch K. Mbah, Machine Learning for Cyber Defense and Attack, Data Analytics 2018 : The Seventh International Conference on Data Analytics

[5] Rubidha Devi.D*, 2R.Venkatesan, A Study On SQL Injection Techniques , International Journal of Pharmacy & Technology IJPT| Dec-2016 | Vol. 8 | Issue No.4 | 22405-22415 Page 22405

[6] S. Dolev and S. Lodha, “Cyber Security Cryptography and Machine Learning“, In Proceedings of the First International Conference, CSCML 2017, Beer-Sheva, Israel, June 29-30, 2017.

[7] Manjeet Rege, Raymond Blanch K. Mbah, Data Analytics 2018 : The Seventh International Conference on Data Analytics.

[8] Sushmita Chakraborty, 2Praveen Kumar, 3Dr. Bhawna Sinha “A Study On Ddos Attacks”.

DOI: 10.48175/568