Big Data Analytics in Cyber Security

Mrs. Preeti Rani
Assistant Professor of Computer Science
Kanya Mahavidyalaya, Kharkhoda

ABSTRACT
The ability to compile enormous volumes of digital data, analyze it, visualize it, and derive insights that can help forecast and thwart cyber attacks is known as big data analytics in security. It improves our cyber defence posture together with security technology. They make it possible for businesses to identify patterns of behaviour that indicate network dangers. We concentrate on how Big Data can enhance information security best practises in this article.
Keywords: Big Data, Cyber Security, Privacy, Database

INTRODUCTION

Big Data is a term used to describe data sets that are so massive or complicated that typical data set processing application software is insufficient for or unable to handle them. Big data differs significantly from conventional data in terms of volume, velocity, and variation. Volume denotes the quantity of data generated, Velocity the rate at which the data is produced, and Variation the categories of organized and unstructured data.

Big data is now a hot topic for research across practically all disciplines, especially cyber security. Social media websites and mobile devices are the primary sources of this data creation. Since data is being generated at such a rapid rate, many people are concerned about the security of the newly created data. It is crucial to keep this data secure since it contains critical information like credit card numbers and bank account numbers. Additionally, improvements in big data analytics offer ways to collect and use this data, making privacy infractions simpler. As a result, in addition to creating Big Data technologies, it is essential to prevent abuse.

DEFINING AND ANALYTICS BIG DATA

Massive amounts of data that are exchanged and stored in computer systems are referred to as "big data."

Big Data is differentiated from traditional technology in 3 ways:

1. The amount of data (Volume) - Size: The volume of datasets, or how much data has been generated, is an important element.
2. The speed at which data is generated and transmitted (Velocity). The structure, behaviour, and permutations of datasets have a crucial role in complexity.
3. The different categories of organised and unstructured data (Variety). Technologies: the methods and instruments applied to handle large or complicated datasets are an important consideration.

TECHNOLOGY MEGA TRENDS

Along with analytics and cloud-based technology, big data is receiving a tonne of attention from industry, the media, and even consumers. All of them are a component of the contemporary eco-system that technology megatrends have produced.

Big data has emerged as a dominant topic or theme in the technology media. It has also been used into numerous compliances and internal audits. According to 72% of participants in EY's Global Forensic Data Analysis Survey 2014, developing big data technologies can be crucial in the prevention and detection of fraud. However, just a small percentage of respondents about 7% knew about any specific big data technologies, and only a very small percentage about 2% were really employing them. FDA (Forensic data analysis) solutions are available to help businesses keep up with the pace of rapidly growing data quantities and organizational complexity.

The top ten developing technologies are assisting users in coping with and handling Big Data in a cost-effective manner. Big Data is vast and incorporates numerous trends and new technology advances.

1. Column oriented database

Traditional, row-oriented databases perform well for online transaction processing with fast update rates, but as data volume increases and as data becomes more unstructured, they struggle with query performance.

2. Schema less database or No Sql database

Many other database formats fall under this heading, including key value storage and document stores, which concentrate on the storage and retrieval of substantial amounts of data that are either unstructured, semi-structured, or even structured.

3. Map Reduce

This programming paradigm enables extremely large job execution scalability across tens of thousands of computers or server clusters. Two activities must be completed for any Map Reduce implementation:

The "Map" task involves transforming an input dataset into a new set of key/value pairs, or "tuples," The "Reduce" task combines a number of the "Map" task's results into a smaller set of tuples.

4. Hadoop

Since Hadoop is a wholly open source platform for managing large amounts of data, it is the best and most widely used implementations of map reduce. It is adaptable enough to function with many data sources. Although it has several uses, one of the most prevalent ones is for massive amounts of dynamic data, such as location-based information from weather or traffic sensors.

5. Hive

It is a SQL-LIKE bridge that enables queries to be conducted against a Hadoop cluster from a traditional BI application. It was initially created by Facebook, but it has long been available as open source. It is a higher-level abstraction of the Hadoop framework that enables anyone to do queries on data stored in a Hadoop cluster exactly as if they were working with a traditional data store.

6. Pig

Yahoo was the creator of PIG. Similar to Hive, PIG is a bridge that aims to connect Hadoop to the realities of developers and business users. The "Perl-like" language used by PIG, as opposed to Hive's SQL-like language, enables query execution on data stored on a Hadoop cluster.

7. WibiData

Web analytics and Hadoop are combined in Wibi Data, which is developed on top of Hbase, a database layer for Hadoop.

8. Sky Tree

It is a powerful platform for machine learning and data analytics that has been designed with handling massive data in mind. Big data is a crucial component since the volume of data makes manual exploration difficult.

BIG DATA LIFE CYCLE

The big data life consist of three stages

1. Creation
2. Processing
3. OutputCreation

Some types of data are impossible to gather, yet they have rarely been used productively up until now (one of general example is, the location of the person at any particular movement of time, the number of steps a person takes every day). These types of data can now be recorded for examination using new and advanced technology like sophisticated sensors and specifically designed software. Changes in the ways we communicate (e.g., social media versus telephone versus text/SMS versus email versus letter) have also improved our capacity to research topics like customer opinion.

Processing

We currently have a very significant amount of data that hasn't been traditionally recorded and processed for a number of reasons, most notably because the cost of processing is far more than the value insights businesses can derive from it. Due to the high expense of processing such data, a significant amount of data is left unprocessed.

The financial and technological barrier for efficient data processing have recently been lowered by new technologies, enabling businesses of all sizes to realize the potential hidden in various data sources. For instance, handling unstructured data is challenging for traditional relational databases.

Many businesses are going to the cloud for their storage needs. Without the high expenses associated with purchasing physical gear, cloud computing enables businesses to employ prebuilt big data solutions or quickly construct and deploy a powerful array of servers.

Output

Data collection, storage, and processing are not simple or inexpensive tasks, and until the data is relevant, it is of no use. The data must also be easily accessible when needed.

There are three key enablers:

- Mobile — Established mobile networks have allowed for easier distribution of information in real-time.
- Visual/interactive — Technologies have brought the ability to review large and complex data sets into the realm of the average business user.
- Human resource — There is a new breed of employees with the knowledge to handle the complexities of big data and with the ability to simplify the output for daily use.

BIG DATA ANALYTICS FOR CYBER SECURITY

1. Big Data Analytics Used In Fraud Detection Techniques used for fraud detection fall into two primaryclasses: statistical techniques and artificial intelligence.

Examples of statistical data analysis techniques are:

1. Data pre-processing techniques for detection, validation, error correction, and filling up of missing or incorrect data.
2. Calculation of various statistical parameters such as averages, quintiles, performance metrics, probability distributions, and so on.
3. Models and probability distributions of various business activities either in terms of various parameters or probability distributions.
4. Computing user profiles.
5. Time-series analysis of time-dependent data.
6. Clustering and classification to find patterns andassociations among groups of data.
7. Algorithms that compare data to previously collected models and profiles to find abnormalities in user or transaction behaviour. Techniques are also required to detect false alarms, calculate risks, and forecast future behaviour of present users or transactions. Management of fraud requires a lot of knowledge.

The main AI techniques used for fraud management include:

1. Using data mining to categorize, cluster, and segment the data as well as automatically discover associations and rules in the data that may represent intriguing patterns, including fraud-related trends.
2. Expert systems to encode expertise for detecting fraud inthe form of rules.
3. Using pattern recognition to match inputs or automatically (unsupervised) identify approximative classes, clusters, or patterns of suspicious behaviour.
4. Machine learning techniques to automatically identify characteristics of fraud.
5. Neural networks that can learn suspicious patterns from samples and used later to detect them.

2. Big Data Analytics Used To Detect Anamoly- Based Intrusion

Algorithms for anomaly detection are fairly easy to set up and work automatically. Thresholds are set once some key performance indicators are selected for an event. An incident is flagged for further inquiry if a threshold is surpassed. The selection of the monitored indicators, the analysis time, and the threshold value settings all have an impact on the method's performance.

Algorithms for anomaly detection require no human intervention and are relatively easy to set up. The parameters used for monitoring, the analysis time, and the threshold value settings all have an impact on how effective this strategy is.

3. Offer security intelligence - They can speed up the process of forensic data correlation and produce security responses that can be put into practise.

CHALLENGES

1. Some organizations may not be data driven. They do not understand the benefits of analytics and hesitant regarding big data analytics.
2. Organizations may think of big data analytics as a way to create value from data. But it is more about finding the right use case related to intended business objective.
3. Analytics team and the users work together in the various phases of analytics process from scope definition to data extraction and delivery.

4. The management may not be able to trust the analytics outcome as it is difficult to understand how data can generate such outcomes.
5. Limited number of well trained and experienced data scientists.
6. Security issues of big data.

CONCLUSION

Real-time actionable intelligence acquisition is the main objective of Big Data analytics for security. In three different ways, big data might significantly affect your present business. It can assist you in:

1. Uncover hidden insights - For instance, when looking at a high service cancellation rate, customer survey data may reveal a pattern or fundamental cause that wasn't previously apparent and that you may eliminate to increase retention.
2. Make better decisions by providing decision makers with richer information. For instance, if you take into account a customer's social media profile, you can get a better understanding of that customer and their place in the world. You can then use this information to enhance your response to service requests or to rank fraud alerts.
3. Automate business operations - For instance, you can analyse comprehensive stock trading data to spot trends that result in shoddy transaction executions and automate the process so that certain actions are taken when that pattern reappears.

REFERENCES

1. Cloud Security Alliance Analytics of Big Data for Security Intelligence
2. Bryant, Katz, and Lazowska (2008)
3. Big Data Analytics for Matrimonial Website Fraud Detection Vemula Geeta et al., International Journal of Computer Science Engineering and Technology (IJCSET), March 2015, Vol. 5, No. 3, 57–61.
4. Using Big Data and Specific Analysis Techniques to Find Insurance Fraud University of Economic Studies, Bucharest, Romania, Ana-Ramona BOLOGA, Razvan BOLOGA, and Alexandra FLOREA
5. Ponemon Institute Research Report on Big Data Cybersecurity Analytics, August 2016.Richard A.Derrig,"Insurance Fraud", The Journal of Risk and Insurance",2002,Vol.69,No.3,271-287
6. Bresfelean, Vasile Paul, Calin-Adrian Comes, Mihaela Bresfelean, and Nicolae Ghisoiu. 2007. Data Mining Clustering Methods in Research. Pages 407–410 of ICEIS (2).
7. Ghisoiu, N., Bresfelean, V. P., and Comes, C. A. 2008. identifying the profile of academic failure among students using data mining techniques. IEEE, pp. 317–322 in Information Technology Interfaces.
8. Information that can be found electronically at http://www.ey.com/Publication/vwLUAssets/ EY Big data: changin g the way businesses operate/%24FILE/EY-Insights-on-GRC-Big-data.pdf
9. National Initiative of Cyber-security Careers and Studies (NICCS), USA, Cyber-Security Definitions. https://niccs.us-cert.gov/glossary; Access date: March 31, 2016.
10. Kaspersky Security Bulletin 2015: Kaspersky Corporation's 2015 global statistics.
11. Big Data and Predictive Analytics: On the Cyber-security Front Line: White Paper,