Big Data Analytics in Cyber Security

Aarushi Arya, Harshit Malhotra
Student,
Dept. of Computer Science Engineering ,
HMR Institute of Technology and Management, Hamidpur, New Delhi, India

Dayanand
Research Scholar, Department of Computer Science and Information Technology, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, Uttar Pradesh, India

Wilson Jeberson
Professor, Department of Computer Science and Information Technology, Sam Higginbottom University of Agriculture, Technology and Sciences, Allahabad, Uttar Pradesh, India

***Abstract*-Big data analytics in security involves the ability to gather massive amounts of digital information to analyze, visualize and draw insights that can make it possible to predict and stop cyber attacks. Along with security technologies, it gives us stronger cyber defense posture. They allow organizations to recognize patterns of activity that represent network threats. In this paper, we focus on how Big Data can improve information security best practices.**

Keywords: Big Data, Cyber Security, Privacy, Database

I. INTRODUCTION

The term Big Data is defined for the data sets that are very large or complex that traditional data set processing application software is inadequate or are unable to deal with these complex or large data sets. The major difference between tradition and big data is in terms of volume, velocity and variation. Volume means amount of data that is been generated; velocity refers to the speed with which the data is been generated and variation means types of structured and non structured data.

Nowadays, big data is becoming an important topic for research in almost every field especially cyber security. The main sources of generation of this data are social media sites and smart devices. Generation of data at this speed leads to the various concern regarding the security of the data that is been created as it is very important to keep this data safe because this data also contain some important and sensitive data such as bank account number passwords credit card details etc so it is important to keep this data secure. Also, advances in Big Data analytics provide tools to extract and utilize this data, making violations of privacy easier. . As a result, along with developing Big Data tools, it is necessary to create safeguards to prevent abuse [2].

II. DEFINING AND ANALYTICS BIG DATA

The term big data is referred to massive amount information that is been stored and transmitted in a computer system.

Big Data is differentiated from traditional technology in 3 ways:

1. The amount of data (Volume) - Size: the volume of datasets is a critical factor, that is, how much amount of data that is been generated

2. The rate of data generation and transmission (Velocity) - Complexity: the structure, behaviour and permutations of datasets in critical factor.

3. The types of structured and unstructured data (Variety) - Technologies: tools and techniques that are been used to process a sizable or complex datasets is a crucial factor.

III. TECHNOLOGY MEGA TRENDS

Big data is generating an enormous amount of attention among business, media and even the consumers, along with the analytics, cloud based technologies. These all the part of the current eco-system created by technology megatrends.

Big data has become a major topic or the theme of the technology media, it has also made its way into many compliances and in internal audits. In EY's Global Forensic Data Analysis Survey 2014, 72% of respondents believe that emerging big data technologies can play a key role in fraud prevention and detection .yet only few about 7% of respondents were aware about any specific big data technologies, and only very few about 2%of them were actually using them. FDA (Forensic data analysis) technologies are available to help the companies to maintain the pace with increasing data at very high speed (volumes), as well as business complexities.

Big Data is broad and encompasses many trends and new technology developments, the top ten emerging technologies that are helping users cope with and handle Big Data in a cost-effective manner.

1. Column oriented database

Traditional, row oriented database are excellent for the online transaction processing with the high update speeds, but they fall short in the query performance as more data volume grows and as data becomes unstructured.

2. Schema less database or No Sql database

there are various database types that fit into this category, such as key value storage and document stores, which focus on storage and retrival of large volume of data which is either unstructured, semi-structured, or even structured data.

3. Map Reduce

This is a programming paradigm that allows for massive job execution scalability against thousands of servers or clusters of servers. Any Map Reduce implementation consists of two tasks:

The "Map" task, where an input dataset is converted into a different set of key/value pairs, or tuples. The "Reduce"

task, where several of the outputs of the "Map" task are combined to form a reduced set of tuples .

4. Hadoop

Hadoop is the best and the most popular implementation of map reduce, being an entirely an open source platform for handling of big data. It is flexible enough to be able to work with multiple data sources. It has several different applications, but one of the top use cases is for large volumes of constantly changing data, such as location-based data from weather or traffic sensors

5. Hive

It is a SQL-LIKE bridge that allows conventional BI application to run queries against a Hadoop cluster It was developed originally by Facebook, but has been made open source for some time now, and it's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store.

6. Pig

PIG was developed by Yahoo .PIG is bridge that tries to bring Hadoop closer to the realities of developers and business users, similar to Hive. Unlike Hive, however, PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, instead of a SQL-like language [9].

7. WibiData

Wibi data is a combination of web analytics with hadoop it is been built on the top of Hbase which itself a database layer on hadoop.

8. Sky Tree

It is a high performance machine learning and data analytics platform focussed specially on the handling of big data. machine learning is a very important part of big data, since the data volume make manual exploration.

IV. BIG DATA LIFE CYCLE

The big data life consist of three stages

1. Creation
2. Processing
3. Output

Creation

Certain type of data is not been able to be captured, but this type of data is rarely been used effectively 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).New and Advance technologies such as advanced sensor and specially customized software can now record this type of information for the purpose of analysis. Changes in the areas of communication in the way we communicate (e.g., social media vs. Telephone vs. text/SMS vs. email vs. letter) have also increased our ability to investigate areas such as consumer sentiment.

Processing

In present day scenario we have extremely large volume of data that have not been traditionally captured and processed for various reasons, mostly the reason is the cost to do the processing is far more greater than the value insights companies can drive from its analysis. That is why large amount of data is left unprocessed because cost involved in processing that data is very high.

However now some new technologies have lowered the cost and the technology barrier for effective data processing, allowing companies of all sizes, to be able to unlock the value contained in different data sources. For instance, it is difficult for conventional relational databases to handle the unstructured data.

Many organizations are looking for the cloud to provide the storage solution. Cloud computing enables companies to use prebuilt big data solutions, or quickly build and deploy a powerful array of servers, without the substantial costs involved in owning physical hardware.

Output

It is not easy and cheap to capture or gather data, store and process the data, it is not at all useful until the information is relevant; it must also be readily available when it is 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.

V. BIG DATA ANALYTICS FOR CYBER SECURITY

1. Big Data Analytics Used In Fraud Detection

Techniques used for fraud detection fall into two primary classes: 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 and associations among groups of data.

7. Matching algorithms to detect anomalies in the behaviour of transactions or users as compared to previously known models and profiles. Techniques are also needed to eliminate false alarms, estimate risks, and predict future of current transactions or users. Fraud management is a knowledge intensive activity.

The main AI techniques used for fraud management include [AI]:

1. Data mining to classify, cluster, and segment the data and automatically find associations and rules in the data that may signify interesting patterns, including those related to fraud.

2. Expert systems to encode expertise for detecting fraud in the form of rules.

3. Pattern recognition to detect approximate classes, clusters, or patterns of suspicious behaviour either automatically (unsupervised) or to match given inputs.

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

Anomaly detection algorithms are very simple to set and functions automatically. Some key performance indicators are for an event chosen and then thresholds are set. If a threshold is exceeded, then the event is signalled for further investigation. The effectiveness of this method is influenced by the choice of indicators to be monitored, of the analysis period, and of the threshold value settings.

Anomaly detection algorithms are very simple to set and functions without human intervention. The effectiveness of this method is influenced by the choice of parameters to be monitored, of the analysis period, and of the threshold value settings.

3. Provide Security Intelligence – They can reduce the time taken to correlate data for forensics purpose and generate actionable security response.

VI. 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

The goal of Big Data analytics for security is to obtain actionable intelligence in real time. Big Data can have a major impact on your current business in three ways. It can help you:

1. Discover hidden insights – For example, if you consider customer survey data when investigating a high service cancellation rate, you may detect a pattern or root cause that wasn't visible before and that you can eliminate to improve retention.

2. Improve decisions, by enriching information for decision makers – For example, if you consider a customer's social media profile, you can get a clearer picture of that customer and their place in the world and you can use that information to improve your response to service inquiries or to prioritize fraud alerts.

3. Automate business processes – For example, you can look at detailed stock trading information to identify patterns that lead to poorly executed trades and automate the process so that certain steps are taken when that pattern occurs again.

REFERENCES

[1] CLOUD SECURITY ALLIANCE Big Data Analytics for Security Intelligence
[2] Bryant, Katz, & Lazowska, 2008
[3] Big Data Analytics for Detection of Frauds in Matrimonial Websites Vemula Geeta et al | International Journal of Computer Science Engineering and Technology (IJCSET) | March 2015 | Vol 5, Issue 3, 57-61
[4] Big Data and Specific Analysis Methods for Insurance Fraud Detection Ana-Ramona BOLOGA, Razvan BOLOGA, Alexandra FLOREA University of Economic Studies, Bucharest, Romania
[5] Big Data Cyber security Analytics Research Report - Ponemon Institute© Research Report Date: August 2016
[6] Richard A.Derrig,"Insurance Fraud", The Journal of Risk and Insurance",2002,Vol.69,No.3,271-287
[7] Bresfelean, Vasile Paul, Mihaela Bresfelean, Nicolae Ghisoiu, and Calin-Adrian Comes. 2007. "Data Mining Clustering Techniques in Academia." In ICEIS (2), pp. 407-410.
[8] Bresfelean, V. P., Bresfelean, M., Ghisoiu, N., & Comes, C. A. 2008. Determining students' academic failure profile founded on data mining methods. In Information Technology Interfaces, IEEE, pp. 317-322.
[9] Data electronically available at http://www.ey.com/Publication/vwLUAssets/EY_Big_data:_changing_the_way_businesses_operate/%24FILE/EY-Insights-on-GRC-Big-data.pdf