Big Data Security Challenges: An Overview and Application of User Behavior Analytics

Tanya Akutota[1] , Swarnava Choudhury[2]

*[1] UG Student, Computer Science Department
National Institute of Technology- Silchar
[2] UG Student, Electronics and Communication Department
National Institute of Technology-Silchar*

---***---

Abstract- *Automation and digitization of activities have resulted in a huge volume of data generated, called Big Data. Big Data helps many organizations gain useful insights, but at the same time, there are two types of risk involved: Security Risk to Big Data itself, and Privacy Risks of users and Individuals. In this paper, the characteristics of big data, its applications, and the security and privacy challenges that come with it are discussed. This paper also explores a novel Big Data Security Analytics method, called User Behavior Analytics, its functioning, use cases and advantages.*

Keywords: **Analytics, Big Data, Challenges, Security, SIEM, UBA**

1. INTRODUCTION

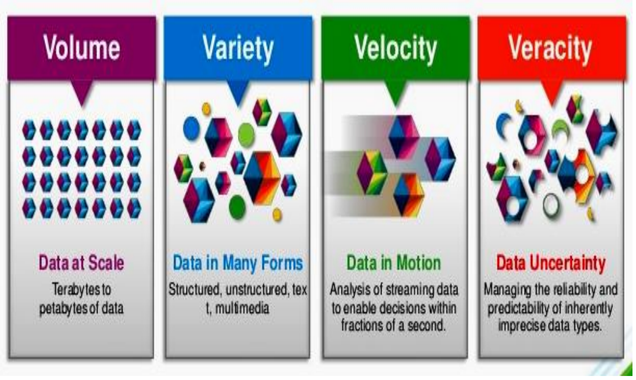
21st century has seen the human lives shifting towards digitalization; automated machines in industries, cellular phones, social networks, etc., all have led us to this. Such huge digitization means generation of huge, perhaps complex sets of data every day. These large and complex data maybe the data from sensors, browsing reports, users' statistics or anything which are increasing exponentially with each passing day. As the inventor of World-Wide-Web, Tim Berners-Lee said, 'Data is a precious thing as they last longer than systems', Big Data Analytics (or BDA) is the tool which actually helps us in realizing the power of such large and complex datasets. The conventional database tools are not able to process such amount of heterogeneous data. Whereas Big Data Analytics uses the power of parallel processing to extract an enormous amount of valuable information, like future trends of market, developments in life science, etc., from the data gathered from all possible and available sources.

A Big Data has many unique characteristics which set it apart from a conventional database system. The types of data they work upon varies. There are basically 3 major classes of data, namely:

1. Structured data- These data are present in the form of rigid relational models, with specific data types and sizes. Conventional database techniques are efficient at this level.
2. Semi-structured data- A type of structured data, but it is hierarchical in nature with the use of tags and markers. XML data is a perfect example of such data.
3. Unstructured data- It doesn't follow a predefined model. The data vary widely; this is where Big Data Analytics comes in.

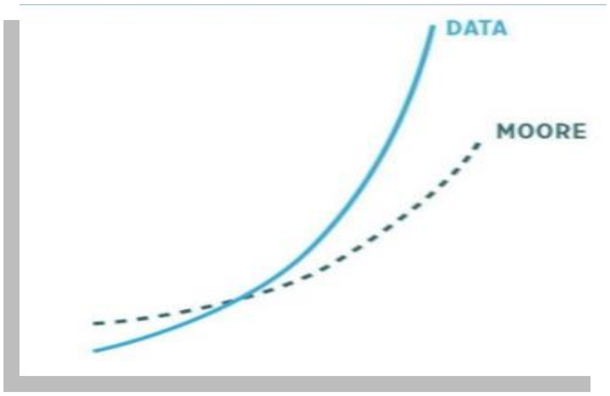
A Big Data can be best described using 5 characteristics, more popularly known as "the 5 V's":

- *Volume*- the scale of data; from Exabytes to Zettabytes!
- *Velocity*- rate at which streaming data is generated and analysed.
- *Variety*- different forms of data- from various external or internal sources.
- *Veracity*- the uncertainty of data, i.e., the different probabilities a value can take.
- *Value*- analysis and visualization of all the above components gives out the final data, the precious information referred to as the Value.


Volume
Variety
Velocity
Veracity
Data at Scale
Terabytes to petabytes of data
Data in Many Forms
Structured, unstructured, text, multimedia
Data in Motion
Analysis of streaming data to enable decisions within fractions of a second.
Data Uncertainty
Managing the reliability and predictability of inherently imprecise data types.

With all the big companies and industries adopting BDA more and more with time, data is being generated at a rapid rate. Let's look at this adjacent comparison. The rate at which data is being generated is much steeper than the famous curve of Moore's law, which says that the computer's capacity will double-fold every 2 years, at half the cost. Moore's law is partially responsible for this rise in data generated, along with other factors.

This wild growth in data overlooks the security threats that such huge well of information may attract to itself. Also, is our personal data getting compromised? These are the privacy issues that we need to worry about. We discuss the limitations BDA faces in the following section.



2. BIG DATA AND SECURITY

The digital data generated is so huge and random, that sometimes it may contain the personal information of the users, thus compromising their privacy. Also, the data generated needs to be kept safe and far from the reach of hackers who have the ability to use such vital information for their own benefits.

2.1 Challenges in BDA

The most important challenges faced by the big data technology are:

- Random Distribution- The distribution of data storage and processing is vital in parallel processing, failing which results in security problems.
- Privacy- Currently, Big Data Analytics treats all the data with same priority. Encoding the more valuable data may prevent any risk of a sniffing attack.
- Computations- The computations performed on big data determines crucial results. Any tampering with the computation may lead to deceiving results.
- Integrity- The raw data fed to the Big Data Analytics must be checked for genuineness of the data before relying on it.
- Communication- The nodes and clusters in Big Data Analytics communicate over ordinary networks, making the data vulnerable to being seized.
- Access Control- The addition or removal of nodes, and privileges among various nodes must be controlled and supervised.

2.2 Techniques to ensure security:

The above challenges can be tackled by taking up precautionary measures. There has been extensive research going on, to make the big data systems more secure:

- R. Toshniwal et al have presented in [8] an improved way of encrypting the big data. It emphasises on encrypting data selectively, instead of encrypting entire database.

- The use of Virtual Private Networks (VPNs) will prevent the chance of sniffing out data from the communication cables.

- A few of the nodes should be used as 'trap-nodes' or Honey-pots. The hacker is deceived and his behaviour can be analysed for improving security measures.

- Segregating the huge amount of data before it's analysed, so as to reject any sort of personal data collected from random sources.

- Nodes in a cluster should have proper authorisation. Authentication software like Kerberos distinguish a malicious node from an authorised one.

3. USER BEHAVIOR ANALYTICS

Behavior analysis systems first appeared in in the early 2000's as a tool to help marketing teams analyze and predict customer buying patterns- they used the data and information based on users behavior to customize and tailor their marketing strategies. User Behavior Analysis or UBA is a cyber-security tool that helps the detection of insider threats, targeted attacks, and financial fraud. UBA solutions look at patterns of human behavior, and then apply ML based algorithms and statistical analysis to detect meaningful anomalies that indicate potential threats Besides UBA, some of the other security terms are:

SIEM: Security Information and Event Management
DLP: Data Loss Prevention
NBA: Next Best Action
EDR: Endpoint Detection and Response
CASB: Cloud Access Security Brokers.

3.1 UBA for Security*:*

According to the research firm Gartner, "*User Behavior Analytics (UBA) [is] where the sources are variable (often logs feature prominently, of course), but the analysis is focused on users, user accounts, user identities — and not on, say, IP addresses or hosts. Some form of SIEM and DLP post-processing where the primary source data is SIEM and/or DLP outputs and enhanced user identity data as well as algorithms characterize these tools. So, these tools may collect logs and context data themselves or from a SIEM and utilize various analytic algorithms to create new insight from that data."* Through learning behavior and tracking anomalies, UBA makes it possible to detect and identify security risk or threats such as:

- Credential compromise
- Rogue / insecure Insiders
- Privileged user abuse
- Malicious hackers
- Breaches
- Password brute force attacks

Some of the popular UBA vendors in the present day market include: Caspida (Splunk), Exabeam, Fortscale, Gurucul, Rapid7, Securonix, ObserveIT, Microsoft ATA, namely.

3.2 Functioning of UBA:

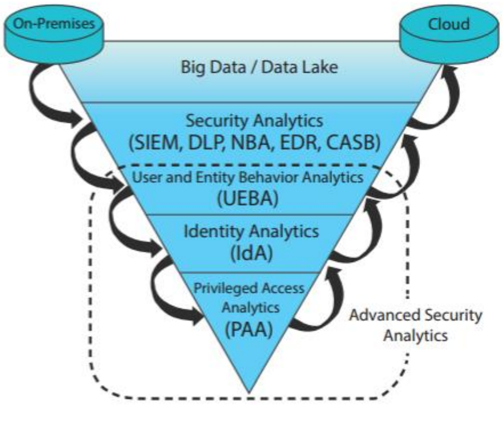
- First, UBA tools determine a baseline of normal activities specific to the organization and its individual users.
- Second, they identify deviations from normal. UBA uses big data and machine learning algorithms to assess these deviations in near-real time. They shed light on cases in which abnormal behavior is underway.

In most standalone UBA vendors these days, there's a core engine, running specialized analytics algorithms, that is fed data from existing sources and, and it analyzes the data. The Analytics Algorithms are the distinguishing factors of the tools. The findings are displayed on a dashboard, and the target is to provide actionable information. These tools don't take any defensive or corrective action themselves, they rather provide security operators with the insight to decide if an action is required. However, it is plausible for integrated tools, such as UBA + Firewall + Defensive systems to be available in the near future.

UBA collects various types of data, such as user roles and titles -- including access, accounts and permissions -- user activity and geographical location, and security alerts Machine learning algorithms allow UBA systems to eliminate false positives and provide clearer and more accurate actionable risk intelligence.

Here's a list of crucial features for UBA software that can function efficiently:

- Process vast amounts of user file and email activity: The UBA system should be able to search through and analyze the activity of many users across huge volumes of data.
- Determine a baseline of "normal" file and email access activities, based on historical data about the employees' activities. The UBA engine therefore should have intimate knowledge of file metadata with access times, users, permissions, etc. to produce accurate profiles of average user behavior.



Big Data Advanced Security Analytics Funnel

Source: "*User and Entity behavior Analytics Use cases*", Gurucul Predictive Analytics, 2017

- Real Time Alerts. The UBA software must be able to track file activities across a large user population in real-time. The decision algorithms must be fast enough to be on par with real-time activity.

3.2 How is UBA different from SIEM?

UBA is a close variant of SIEM. SIEM mainly relies on analyzing events captured in firewalls, OS, and other system logs in order to spot interesting correlations, usually through pre-defined rules. By prioritizing and focusing on user behavior instead of system events, UBA builds a profile of an employee based on their usage patterns, and sends out an alert if it sees abnormal user behavior

UBA tools utilize both basic and advance analytics approach ranging from rules-based models to Deep machine learning. A SIEM tool may or may not utilize these advances methods.
UBA engines work with narrow and highly relevant data for analysis. This results in higher quality of alerts with less false-negatives and false-positives. Whereas, SIEM tools take in an overwhelming amount of data only to generate more noise in their alerts

UBA tools build profiles for Users' behavior over a period of time and uses that as a baseline to detect any malicious actions by recording any abrupt change in their behavioral patterns. This functionality is not available in all SIEM tools.

3.3 Use Cases:

Situations and possible scenarios where UBA can play a key role are discussed hereunder. They are (but not limited to):

1. *Account hijacking and Credential Compromise:* Here, attackers exploit vulnerabilities through attacks such as Pass-the-Hash (PtH), Pass-the-Token, golden ticket, Brute Force and Remote Execution to gain access to user credentials. the underlying machine learning algorithms help detect these by inspecting various parameters like timestamp, location, IP, device, transaction patterns etc. to identify any deviation from the normal behavior of a particular account its activities.

2. *Privileged access compromise:* Privileged users account can be at risk particularly because they might have access to highly sensitive or classified information. UBA should be able to detect these scenarios, such as using HPA to assign special or elevated privileges to the user's own account or, transactions outside the window of checkout and check-in time window

3. *Insider Threat:* A rogue insider continues to be a source of data loss. Using ML behavior models Along with data risk monitoring and identifying high-risk profiles, UBA can reveal anomalies in data that humans could not otherwise recognize or detect.

4. *Data Exfiltration Alerts:* UBA solutions can identify known patterns such as: sensitive content downloaded and copied to external storage devices, large amounts of source code checked out from source code repositories and file uploads to cloud storage, emails to personal accounts, access to competitor websites, etc.

5. *Account Lockouts*: UBA should also help identify if an account lockout is an honest mistake or an attempt by hackers to compromise the details.

6. *Continuous Session Tracking*: More visibility is provided by tracking the user session state, even when a user navigates across different sources or applications, using different accounts at the same time. This helps reduce false positives.

7. *Anomalous behavior and watch lists*: UBA addresses anomalous behavior with watch lists to quickly profile and keep track of unknowns and apply escalating predictive risk scores. Machine learning behavior models are designed to deliver feedback on false positives

8. *Aggregating Risk Scores*: Unlike SIEM, UBA doesn't generate a huge number of alerts. Rather, it aggregates the user risk scores at the user level.

3.4 Advantages of UBA:

- More efficient than SIEM in terms of detecting malicious user behavior
- It doesn't collect data. Rather, it uses data collected by other security tools.
- As opposed to CASB gateways, UEBA actually tracks every online and offline transaction, activity, and logs
- UBA is designed to reduce false positives with new types of algorithms. These algorithms concentrate on aggregate of anomalies instead of each and every anomaly
- UBA is more efficient in pointing out and alerting about insider threats (E.g.: Such as Ed Snowden's theft of critical information)
- Allows more comprehensive management and risk handling of privileged accounts.

REFERENCES

[1] Jon Olstick, “Time to Consider User Behavior Analytics [UBA],” https://csooonline.com , January 25th , 2016

[2] Andy Green, “What is User Behavior Analytics?” https://blog.varonis.com/what-is-user-behavior-analytics ,July 21st , 2015

[3] Heather Howland,” What is UEBA and Why Does it Matter In Threat Detection? [Part 1 - Blog Series]” https://blog.preempt.com/part-1-what-is-ueba-and-why-does-it-matter-in-threat-detection-blog-series, September 22nd, 2016

[4] Amit Singh “Top 5 User Behaviour Analytics (UBA) Vendors at RSAC 2017” Fire Compass (https://www.firecompass.com/blog/top-5-user-behaviour-analytics-uba-vendors-at-rsa-conference-2017/), January 26th, 2017

[5] Johna Till Johnson “User behavioral analytics tools can thwart security attacks” TechTarget (http://searchsecurity.techtarget.com/feature/User-behavioral-analytics-tools-can-thwart-security-attacks)

[6] Margaret Rouse” User Behavior Analytics (UBA)” http://searchsecurity.techtarget.com/definition/user-behavior-analytics-UBA

[7] Gurucul – “User and Entity Behavior Analytics Use Cases” e Paper White Paper 2017

[8] Raghav Toshniwal “Big Data Security Issues and Challenges” IJIRAE Issue 2 Vol. 2, February 2015

[9] William El Kaim “Introduction to Big Data” , October 2016

(https://www.slideshare.net/welkaim/introduction-to-big-data-65870623)

[10] Youssef Gahi “Big Data Analytics: Security and Privacy Challenges” IEEE 2016

[11] K. Shanmugapriya “Security Issues Associated with Big Data in Cloud Computing” IJCSIT Vol. 6(6), 2015