Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

Big Data Analytics for Security to Obtain Actionable Intelligence in Real Time

[1]K.Valli Madhavi, [2]Dr.Y.Venkateswarlu, [3]Varsha Sharma
[1]Research Scholar, [2]Professor, [3]Research Scholar
[1,3]Computer Science Engineering Department, Mewar University, Udaipur, India
[2]Computer Science Engineering Department, GIET Engineering College, Rajahmundry, India,

***Abstract*—**Advances in information technology and its widespread magnification in several areas of business, engineering, medical, and scientific studies are resulting in information/data explosion. Erudition revelation and decision-making from such rapidly growing voluminous data are a challenging task in terms of data organization and processing, which is an emerging trend kenned as astronomically immense data computing, an incipient paradigm that amalgamates sizable voluminous-scale compute, incipient data intensive techniques, and mathematical models to build data analytics. Sizable voluminous data computing demands an immensely colossal storage and computing for data curation and processing that could be distributed from on-premise or clouds infrastructures. With information solidly close by and with the capacity given by Big Data Technologies to prosperously store and break down this information, we can discover answers to these inquiries and work to streamline each part of our conduct. With the appearance of numerous computerized modalities this information has developed to astronomically immense information is still on the ascent. Eventually Immensely Colossal Data innovations can subsist to enhance rudimental leadership and to give more eminent insights...faster when required yet with the drawback of loss of information security.

Keywords: *Data Analytics, Big Data, Data Computing, Data Explosion.*

I. INTRODUCTION

Society is becoming increasingly more instrumented and as a result, organisations are producing and storing vast amounts of data. Managing and gaining insights from the produced data is a challenge and key to competitive advantage. Analytics solutions that mine structured and unstructured data are important as they can help organisations gain insights not only from their privately acquired data, but also from large amounts of data publicly avail-able on the Web [118]. The ability to cross-relate private information on consumer preferences and products with information from tweets, blogs, product evaluations, and data from social networks opens a wide range of possibilities for organisations to understand the needs of their customers, predict their wants and demands, and optimise the use of resources. This paradigm is being popularly termed as Big Data. Despite the popularity on analytics and Big Data, putting them into practice is still a complex and time consuming endeavour.

The term Big Data alludes to extensive scale data administration and examination innovations that surpass the capacity of conventional information preparing advancements. Big Data is separated from conventional innovations in three ways: the measure of information (volume), the rate of information era and transmission (speed), and the sorts of organized and unstructured information. Human creatures now make 2.5 quintillion bytes of information every day. The rate of information creation has expanded so much that 90% of the information on the planet today has been made in the most recent two years alone. This speeding up in the generation of data has made a requirement for new innovations to dissect monstrous information sets. The criticalness for community research on Big Data themes is underscored by the U.S. national government's late $200 million financing activity to bolster Big Data research. This record portrays how the consolidation of Big Data is changing security examination by giving new apparatuses and chances to utilizing expansive amounts of organized and unstructured information.

II. BIG DATA ANALYTICS

Big data refers to large-scale data architectures and facilitates tools addressing new requirements in handling data volume, velocity, and variability. Traditional databases (data warehousing) assume data are organized in rows and columns and employ data-cleansing methods on the data, while the data volumes grow over a time period and often lack on handling such large-scale data processing. Traditional data base/warehousing systems were designed to address smaller volumes of structure data, with the predictable updates and consistent data structure, which mostly operate on single server and lead to operational expenses with the increased data volume. However, big data comes in a variety of diverse formats with both batch and stream processing in several areas such as geospatial data, 3D data, audio and video, structured data, unstructured text including log files,

Big Data analytics – the process of analyzing and mining Big Data – can produce operational and business knowledge at an unprecedented scale and specificity. The need to analyze and leverage trend data collected by businesses is one of the main drivers for Big Data analysis tools. The technological advances in storage, processing, and analysis of Big Data include

- The rapidly decreasing cost of storage and CPU power in recent years;
- The flexibility and cost-effectiveness of datacenters and cloud computing for elastic computation and storage; and
- The development of new frameworks such as Hadoop, which allow users to take advantage of these distributed computing systems storing large quantities of data through flexible parallel processing.

These advances have created several differences between traditional analytics and Big Data analytics which is shown in figure1.

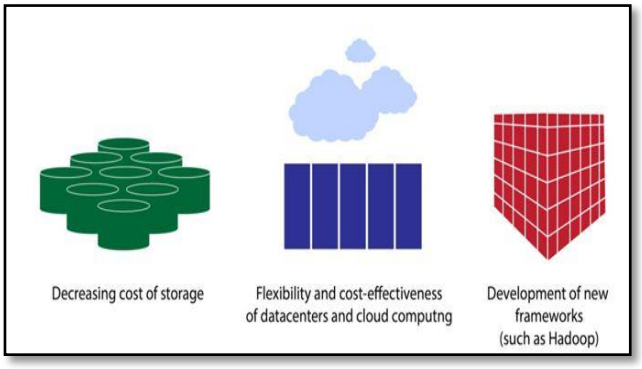

Figure 1: Drivers of Big Data

Storage cost has dramatically decreased in the last few years. Therefore, while traditional data warehouse operations retained data for a specific time interval, Big Data applications retain data indefinitely to understand long historical trends.

Big Data tools such as the Hadoop ecosystem and NoSQL databases provide the technology to increase the processing speed of complex queries and analytics.

Extract, Transform, and Load (ETL) in traditional data warehouses is rigid because users have to define schemas ahead of time. As a result, after a data warehouse has been deployed, incorporating a new schema might be difficult. With Big Data tools, users do not have to use predefined formats. They can load structured and unstructured data in a variety of formats and can choose how best to use the data.

Big Data technologies can be divided into two groups: batch processing, which are analytics on data at rest, and stream processing, which are analytics on data in motion. Real-time processing does not always need to reside in memory, and new interactive analyses of large-scale data sets through new technologies like Drill and Dremel provide new paradigms for data analysis; however, Figure 2 still represents the general trend of these technologies.

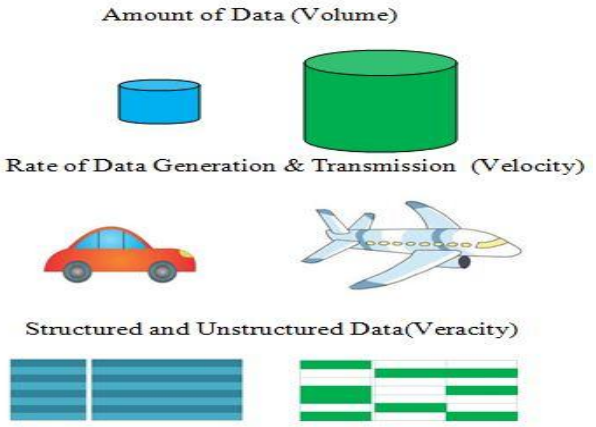

Figure 2: Traditional database Vs Big Data

A. *Hadoop*

The explosion of Big Data has forced the companies to use the technologies that could help them manage the complex and unstructured data in such a way that maximum information could be extracted and analyzed without any loss and delay. This necessity sprouted the development of big data technologies that are able to process multiple operations at once without a failure. However, what are those special features that made Hadoop widely accepted among other technologists? Read below:

a) *Capable of storing and processing complex datasets:* With increasing volumes of data, increases the possibility of data loss and failure. However Hadoop's ability to store and process large and complex unstructured datasets makes it somewhat special.

b) *Great computational ability :* Its distributed computational model enables fast processing of Big Data with multiple nodes running in parallel.

c) *Lesser faults:* Implementing it leads to lesser number of failures as the jobs are automatically redirected to other nodes as and when one node fails. This ultimately causes the system to respond in real-time without failure.

d) *No pre-processing required :* Enormous data can be stored and retrieved at once, including both structured and unstructured data without having to preprocess before storing into the database.

e) *Highly scalable :* It is a highly scalable big data tool as you can raise the size of cluster from single machine to thousands of servers without having to administer extensively.

f) *Cost-effective:* Open source technologies come free of cost and hence require lesser amount of money for implementing them.

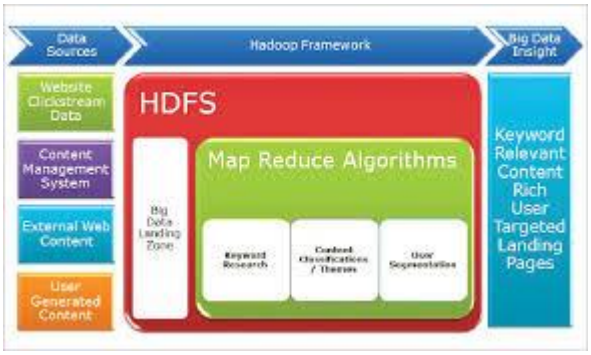

Figure 3: Hadoop Framework

B. *Map Reduce*

- Procedure interminable measures of data(multi-terabytedata-sets)in-parallel.
- Accomplishes unrivaled on broad gatherings (countless) of item hardware in a reliable, deficiency tolerant way.
- Parts the data set into free knots.
- Sorts the yields of the maps, which are then commitment to there duce assignments.
- Takes thought of booking assignments, watching the mandre executes the failed endeavours.

The MapReduce framework works exclusively on <key, value> sets, that is, the structure sees the commitment to the business as a plan of <key, value> consolidates and makes a course of action of <key, value> sets as the yield of the occupation, perhaps of different types. The key and regard classes must be serializable by the structure and thusly need to realize the Writableinterface. Additionally, the key classes need to complete the WritableComparableinterface to empower sorting by the framework. Data and Yield sorts of a MapReduce job:(input) <k1, v1> - > map-> <k2, v2> - > join > <k2, List(v2)> - > diminish > <k3, v3> (yield).

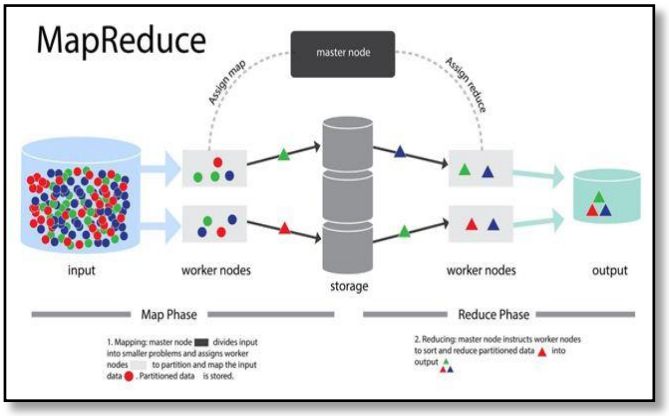

Figure 4: Map Reduce

III. BIG DATA ANALYTICS FOR SECURITY

This section expounds how Immensely colossal Data is transmuting the analytics landscape. In particular, Astronomically Immense Data analytics can be leveraged to ameliorate information security and circumstantial cognizance. For example, Sizably voluminous Data analytics can be employed to analyze financial transactions, log files, and network traffic to identify anomalies and suspicious activities, and to correlate multiple sources of information into a coherent view.

Data-driven information security dates back to bank fraud detection and anomaly-predicated intrusion detection systems. Fraud detection is one of the most visible uses for Sizably Voluminous Data analytics. Credit card companies have conducted fraud detection for decades. However, the custom-built infrastructure to mine immensely colossal Data for fraud detection was not economical to acclimate for other fraud detection uses. Off-the-shelf Immensely Colossal Data implements and techniques are now bringing attention to analytics for fraud detection in healthcare, indemnification, and other fields.

In the context of data analytics for intrusion detection, the following evolution is anticipated:

- 1st generation: Intrusion detection systems – Security architects realized the need for layered security (e.g., reactive security and breach response) because a system with 100% protective security is impossible.
- 2nd generation: Security information and event management (SIEM) – Managing alerts from different intrusion detection sensors and rules was a big challenge in enterprise settings. SIEM systems aggregate and filter alarms from many sources and present actionable information to security analysts.
- 3rd generation: Big Data analytics in security (2nd generation SIEM) – Big Data tools have the potential to provide a significant advance in actionable security intelligence by reducing the time for correlating, consolidating, and contextualizing diverse security event information, and also for correlating long-term historical data for forensic purposes.

Analyzing logs, network packets, and system events for forensics and intrusion detection has traditionally been a significant problem; however, traditional technologies fail to provide the tools to support long-term, large-scale analytics for several reasons:

- Storing and retaining an astronomically immense quantity of data was not economically feasible. As a result, most event logs and other recorded computer activity were effaced after a fine-tuned retention period (e.g., 60 days).
- Performing analytics and intricate queries on astronomically immense, structured data sets was inefficient because traditional implements did not leverage immensely colossal Data technologies.
- Traditional implements were not designed to analyze and manage unstructured data. As a result, traditional implements had rigid, defined schemas. Astronomically Immense Data implements (e.g., Pig Latin scripts and conventional expressions) can query data in flexible formats.
- Big Data systems use cluster computing infrastructures. As a result, the systems are more reliable and available, and provide guarantees that queries on the systems are processed to completion.

New Big Data technologies, such as databases related to the Hadoop ecosystem and stream processing, are enabling the storage and analysis of large heterogeneous data sets at an unprecedented scale and speed. These technologies will transform security analytics by: (a) collecting data at a massive scale from many internal enterprise sources and external sources such as vulnerability databases; (b) performing deeper analytics on the data; (c) providing a consolidated view of security-related information; and (d) achieving real-time analysis of streaming data. It is important to note that Big Data tools still require system architects and analysts to have a deep knowledge of their system in order to properly configure the Big Data analysis tools.

IV. THE WINE PLATFORM FOR EXPERIMENTING WITH BIG DATA ANALYTICS IN SECURITY

The Worldwide Intelligence Network Environment (WINE) provides a platform for conducting data analysis at scale, using field data collected at Symantec (e.g., anti-virus telemetry and file downloads), and promotes rigorous experimental methods (Dumitras & Shoue, 2011). WINE loads, samples, and aggregates data feeds originating from millions of hosts around the world and keeps them up-to-date. This allows researchers to conduct open-ended, reproducible experiments in order to, for example, validate new ideas on real-world data, conduct empirical studies, or compare the performance of different algorithms against reference data sets archived in WINE. WINE is currently used by Symantec's engineers and by academic researchers.

A. Data Sharing and Provenance

Experimental research in cyber security is rarely reproducible because today's data sets are not widely available to the research community and are often insufficient for answering many open questions. Due to scientific, ethical, and legal barriers to publicly disseminating security data, the data sets used for validating cyber security research are often mentioned in a single publication and then forgotten. The "data wishlist" (Camp, 2009) published by the security research community in 2009 emphasizes the need to obtain data for research purposes on an ongoing basis.

WINE provides one possible model for addressing these challenges. The WINE platform continuously samples and aggregates multiple petabyte-sized data sets, collected around the world by Symantec from customers who agree to share this data. Through the use of parallel processing techniques, the

platform also enables open-ended experiments at scale. In order to protect the sensitive information included in the data sets, WINE can only be accessed on-site at Symantec Research Labs. To conduct a WINE experiment, academic researchers are first required to submit a proposal describing the goals of the experiment and the data needed. When using the WINE platform, researchers have access to the raw data relevant to their experiment. All of the experiments carried out on WINE can be attributed to the researchers who conducted them and the raw data cannot be accessed anonymously or copied outside of Symantec's network.

WINE provides access to a large collection of malware samples and to the contextual information needed to understand how malware spreads and conceals its presence, how malware gains access to different systems, what actions malware performs once it is in control, and how malware is ultimately defeated. The malware samples are collected around the world and are used to update Symantec's anti-virus signatures. Researchers can analyze these samples in an isolated "red lab," which does not have inbound/outbound network connectivity in order to prevent viruses and worms from escaping this isolated environment. A number of additional telemetry data sets, received from hosts running Symantec's products, are stored in a separate parallel database. Researchers can analyze this data using SQL queries or by writing MapReduce tasks.

These data sets include anti-virus telemetry and intrusion-protection telemetry, which record occurrences of known host-based threats and network-based threats, respectively. The binary reputation data set provides information on unknown binaries that are downloaded by users who participate in Download Insight, Symantec's reputation-based security program. The history of binary reputation submissions can reveal when a particular threat has first appeared and how long it existed before it was detected. Similarly, the binary stability data set is collected from the users who participate in the Performance Insight program, which reports the health and stability of applications before users download them. This telemetry data set reports application and system crashes, as well as system lifecycle events (e.g., software installations and uninstallations). Telemetry submission is an optional feature of Symantec products and users can opt out at any time.

These data sets are collected at high rates and the combined data volume exceeds 1 petabyte. To keep the data sets up-to-date and to make them easier to analyze, WINE stores a representative sample from each telemetry source. The samples included in WINE contain either all of the events recorded on a host or no data from that host at all, allowing researchers to search for correlations among events from different data sets.

This operational model also allows Symantec to record metadata establishing the provenance of experimental results (Dumitras & Efstathopoulos, 2012), which ensures the reproducibility of past experiments conducted on WINE. The WINE data sets include provenance information, such as when an attack was first observed, where it has spread, and how it was detected. Moreover, experimentation is performed in a controlled environment at Symantec Research Labs and all intermediate and final results are kept within the administrative control of the system.

B. Big data security

Big data project can uncover tremendous value for an enterprise, by revealing customer buying habits, detecting or preventing fraud, or monitoring real-time events. However, a poorly run big data project can be a security and compliance nightmare, leading to data breaches. Big data must be protected, to ensure that only the right people have appropriate access to it. Big data security addresses several mechanisms for large-scale high-volume rapidly growing varied forms of data, analytics, and large-scale compute infrastructure. As the data volumes and compute infrastructures are very large, traditional methods of computing and data security mechanisms, which are tailored for securing small-scale data and infrastructure, are inadequate. Also, the use of large-scale cloud infrastructures, with a diversity of software platforms, spreads across large networks of computers and also increases the attacks. The onion model of defense for big data security is depicted in Figure 6, and the several elements are described in the succeeding texts.

- Distributed computing infrastructure: mechanisms for providing security while data are analyzed over multiple distributed systems. Big data setup would be either confined to an enter-prise or could be a large collection of several enterprises, social, and scientific collection of disparate sources distributed system. Privacy, security, and confidentiality – not revealing private and confidential information to unauthorized users. For example, in a mailing system, secrecy is concerned about preventing the users from finding out the passwords of other users.
- Large-scale distributed data: privacy-preserving mechanisms, encryption techniques for the data stored on large-scale distributed systems, role-based access and control mechanisms, and security of column, document, key-value, and graph data models to be evolved. In order to maintain fast access for the data, NoSQL databases come with little built-in security; due to their BASE properties, rather than requiring consistency after every transaction, the data base just needs to eventually reach a consistent state.
- Analytics security: developing frameworks that are secured that allow organizations for publish and use the analytics securely based on several authentication mechanisms such as one-time passwords, multi-level authentications, and role-based access mechanisms.
- Users' privacy and security: confidentiality, integrity, and authentication mechanisms to validate the users.

CONCLUSION

Big data computing is an emerging platform for data analytics to address large-scale multidimensional data for knowledge discovery and decision-making. In this paper, we have studied, characterized, and categorized several aspects of big data computing systems. Big data technology is evolving and changing the present traditional data bases with effective data organization, large computing, and data workloads processing with new innovative analytics tools bundled with statistical and machine-learning techniques.

The goal of Big Data analytics for security is to obtain actionable intelligence in real time. Although Big Data analytics have significant promise, there are a number of challenges that must be overcome to realize its true potential. The following are only some of the areas that need to be addressed:

- Data provenance: authenticity and integrity of data used for analytics. As Big Data expands the sources of data it can use, the trustworthiness of each data

source needs to be verified and the inclusion of ideas such as adversarial machine learning must be explored in order to identify maliciously inserted data.

- Privacy: we need regulatory incentives and technical mechanisms to minimize the amount of inferences that Big Data users can make. CSA has a group dedicated to privacy in Big Data and has liaisons with NIST'sBig Data working group on security and privacy. We plan to produce new guidelines and white papers exploring the technical means and the best principles for minimizing privacy invasions arising from Big Data analytics.
- Securing Big Data stores: this document focused on using Big Data for security, but the other side of the coin is the security of Big Data. CSA has produced documents on security in Cloud Computing and also has working groups focusing on identifying the best practices for securing Big Data.
- Human-computer interaction: Big Data might facilitate the analysis of diverse sources of data, but a human analyst still has to interpret any result. Compared to the technical mechanisms developed for efficient computation and storage, the human-computer interaction with Big Data has received less attention and this is an area that needs to grow. A good first step in this direction is the use of visualization tools to help analysts understand the data of their systems.

References

[1] Bilge, L. & T. Dumitras. (2012, October) Before We Knew It: An empirical study of zero-day attacks in the real world. Paper presented at the ACM Conference on Computer and Communications Security (CCS), Raleigh, NC.

[2] Bryant, R., R. Katz & E. Lazowska. (2008). Big-Data Computing: Creating revolutionary breakthroughs in commerce, science and society. Washington, DC: Computing Community Consortium.

[3] Camp, J. (2009). Data for Cybersecurity Research: Process and "whish list". Retrieved July 15, 2013, from http://www.gtisc.gatech.edu/files_nsf10/data-wishlist.pdf

[4] Cugoala, G. & Margara, A. (2012). Processing Flows of Information: From Data Stream to Complex Event Processing. ACM Computing Surveys 44, no. 3:15.

[5] Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. Stamford, CT: META Group.

[6] Apache Hadoop Project (http://hadoop.apache.org)

[7] http://www.esgglobal.com/blogs/strong-opportunities-and-some-challenges-for-bigdatasecurity-analytics-in-2014/

[8] www.computereducation.org

[9] Big Data by Viktor Mayer-Schonberger.