MAT JOURNALS

Journal of Data Mining and Management
Volume 1 Issue 2

Review on Big Data Promises for Information Security

Anika Gupta

Department of Computer Science and Engineering, Punjab Technical University, Jalandhar, India

E-mail: anika.mit90@yahoo.com

Abstract

Big information is expounded to technologies for assembling, processing, analyzing and extracting helpful information from terribly giant volumes of structured and unstructured information generated by totally different sources at high speed. Huge information creates essential info security and privacy issues, at identical time huge information analytics guarantees important opportunities for hindrance and detection of advanced cyber-attacks victimisation correlate internal and external security information. We tend to address many challenges to appreciate true potential of massive information for info security. The paper analyzes huge information applications for info security issues, and defines analysis directions on huge information analytics for counterintelligence.

Keywords: *Big data, technologies, analyzing, extracting, information*

INTRODUCTION

There are completely different definitions of the "Big Data" term. The foremost fashionable definition is given by describing their three characteristics known as "3V": Volume (the information volumes are terribly large that cannot be processed by ancient methods), speed (the information is made with nice speed and should be captured and processed rapidly) and selection (variety of knowledge types: structured, semi-structured, and unstructured). Supported information quality, IBM has accessorial a fourth V called: truthfulness. However, Oracle has accessorial a fifth V called: worth, light the accessorial worth of huge information [1]. Huge information could be a comparatively new term (it was solely coined in 2008), however, it became a really fashionable buzz word once publication of the report ready by McKinsey international Institute [2, 3]. Currently, fashionable media is replete with publications on huge information opportunities for state, business,

healthcare, enforcement, cyber security, analysis and development, etc. trade is buzzing with the promise of huge information [3]. National governments have recently proclaimed vital programs on huge information applications [4, 5]. Readers could have a thought that huge information will be used solely by massive corporations. It ought to be noted that the convergence of huge information and cloud computing technologies permits little and medium enterprises exploitation huge information opportunities too [6]. Whereas, the guarantees of huge information are real–they are tested by success primers of huge corporations like Google, Yahoo, Facebook–leading information science researchers are warning that there are several challenges at every step of huge information analysis pipeline. The relation of knowledge security and massive information is twofold. Data security and privacy are among the foremost difficult problems with huge information. At an equivalent time, huge information analytics guarantees important opportunities for determination totally different data security issues. There are several reports, particularly by huge firms concerning huge information opportunities for data security; however, there are a couple of publications on challenges of huge information for data security. The aim of this paper is to investigate the-state-of-the-art huge information research for data security, and to see the foremost relevant analysis directions.

THE HADOOP ECOSYSTEM

Big information framework for process and analysis consists of variety of software package tools. Presently, the Hadoop software package system is taken into account as an equivalent word for giant information. Hadoop implements MapReduce technology of Google, that provides automatic information paralleling and process on laptop clusters. Several of the Hadoop elements area unit open source software package developed in varied Apache comes.

Below a brief description of some components of the Hadoop ecosystem is given:

HDFS (Hadoop Distributed File System)

A distributed classification system for storage and management of date warehouses from many terabytes to petabytes; it is the core of the Hadoop. HDFS splits the computer file into blocks and allocates these blocks on servers in numerous places allotted to them. The TCP/IP level is employed for communication. HDFS is fault tolerant,

and failure of any part does not have an effect on the system performance.

MapReduce

Implements (in Java) Google's distributed computing model for parallel computing with very large data, several petabytes, in computer clusters. A MapReduce job consists of two steps: Map and Reduce. On the Map-step the input data is pre-processed. To do this, one of the computers (known as the master node) receives input data of the problem, divides them into parts and transfers to other computers (worker nodes) for pre-processing. On the Reduce-step the pre-processed data is collected. The master node receives responses from the working nodes and forms the solution on the basis of their results.

Apache Pig component consists of a compiler that generates a sequence of MapReduce programs, and language 'Pig Latin'. Provides support for performing SQL-like queries
to distributed databases to Hadoop.

Hive

A data warehouse infrastructure, used to refer to large data placed in the Hadoop file system through SQL.

HCatalog

Provides storage management service and data tables created in Hadoop. It supports sustainable functioning of the Hadoop components, such as Pig, MapReduce, Streaming
and Hive.

HBase (Hadoop DataBase)

A distributed, columnar database (derived from Google's BigTable).

Zookeeper

Its main function is to store the coordination information, naming, providing distributed synchronization, and group services, which are very important for a variety of
distributed systems.

Mahout

Software for machine learning, as well as key algorithms, akin to classification, clustering, and recommendation and cooperative filtering. Basic algorithms square measure enforced with Map/Reduce paradigm on the Hadoop higher level. Elements like Sqoop and Flume, enclosed within the scheme square measure wont to transmit knowledge to the Hadoop-clusters and contrariwise. Hadoop is commonly employed in conjunction with common place knowledge storage

and process technologies, it is generally further innovative solutions akin to Storm, Dremel, Drill, etc. Moreover, the majority major producers of business intelligence tools add practicality to their merchandise to investigate knowledge for good keep in Hadoop-clusters. We tend to may extend the list of the Hadoop scheme elements, as a result of a lot of and a lot of firms square measure getting into the market with merchandise that have a reference to Hadoop.

BIG DATA SECURITY

Although, information security and is critical issues for Big Data, these issues have attracted little attention until now. Some researchers point out that due to big volumes Big Data is
unattractive for the attackers for now. But Big Data creates new threats to information security, and ideology of protection adopted for traditional security measures, is no longer adequate for Big Data. Cloud Security Alliance (CSA), a working group which studies Big Data security issues recently prepared a document that lists the tools to protect Big Data systems:

1. Secure computations in distributed programming frameworks.
2. Security best practices for non-relational data stores.
3. Secure data storage and transactions logs.
4. End-point input validation/filtering.
5. Real-time security monitoring.
6. Scalable and composable privacy-preserving data mining and analytics.
7. Cryptographically enforced data centric security.
8. Granular access control.
9. Granular audits.
10. Data provenance.

BIRD'S EYE VIEW OF CYBER THREAT LANDSCAPE AND SECURITY TOOLS

Advanced Threats

Businesses associated governments face an evolving threat landscape. One amongst the best challenges is given by advanced persistent threats (APTs), those square measure subtle, long-term, multi-phase, multi-faceted attacks targeting a selected organization. RSA, Google, National Aeronautics and Space Administration and a few nation states have toughened giant security breaches because of APTs. Mitigating the chance of APTs needs advances on the far side ancient security defences to incorporate time period threat management.

Data Sources

Organizations collect a large sort of knowledge for security analysis and investigations: traffic knowledge, log files (operation system, application, firewall, web access, etc.), log/event knowledge from networking devices, DNS-specific logs/events, user activity, physical security activity knowledge, firewall rule sets, asset data, and etc. In spite of all this, internal knowledge assortment and analysis is not any longer enough. Risk management and incident detection/response practices are being supplemented with growing volumes of external security knowledge.

Data-Driven Information Security Tools

Data-driven information security dates back to anomaly-based intrusion detection systems (IDSs). The next stage is development of Security Information and Event Management (SIEM) systems.

Intrusion Detection System (IDS)

IDSs collect and analyze network traffic and use predominantly signature approach to intrusion detection. Main limitations of IDSs are limited use of external events and "flat" event model. IDSs suffer from high values of false positives and false negatives.

Security Information and Event Management (SIEM)

SIEM systems collect, aggregate, and filter alarms from many intrusion detection sensors and other sources and present actionable information to security analysts. SIEM systems use
external data sources extensively. The tree event model allows correlating higher-level events.

BIG PROMISES FOR CYBER SECURITY

New huge knowledge technologies–reminiscent of the Hadoop scheme, stream mining, complex-event process, and NoSQL databases-square measure facultative the analysis of large-scale, heterogeneous datasets at new scales and speeds. These technologies square measure permits extending ancient info security systems by facilitating the storage, maintenance, and analysis of security info. Analysis of information from totally different sources in several formats, the flexibility to match these knowledge, anomaly detection, and combating cyber threats in real time-all this has been created doable through the employment of technologies for process and analyzing huge knowledge. Many companies offering security solutions published white papers, emphasizing the advantages and

opportunities of Big Data for security. The CSA working group's report, "Big Data Analytics for Security Intelligence" focuses on big data's role in security, and highlights possible research directions. RSA recommends gradually move to the Intelligence-Driven Security model. Compared with conventional SIEM systems the advantage of the Intelligence-Driven Security model is the ability to analyze a much larger extent than before, the most diverse, not used before the data.

THE BIG DATA CHALLENGES FOR CYBER SECURITY

Although the application of Big Data analytics to cyber security problems has significant promise, we must address several challenges to realize its true potential.

Privacy

The Big data analytics makes privacy violations easier. It is a fact that the implications of privacy exposure to end-users are not yet fully understood. We must develop privacy preserving Big Data applications.

APT Detection by Big Data Analytics

There is want for brand new detection algorithms, capable of process vital amounts knowledge from numerous data sources. Currently, tiny low varieties of proof of construct deployments that utilize massive knowledge analytics for security event detection exist, and show promising results.

High Performance Cryptography

Encryption and decryption algorithms; encrypted data search, attribute-based encryption, attacks on the availability, reliability, and integrity of Big Data.

Big Data Datasets for Security Research

There is a significant amount of cybersecurity data that exists, but understanding ground truth is nearly impossible from data that is organically gathered. These datasets can contain a tremendous amount of activity, but knowing what is benign and/or where attack data are to be found is very difficult.

Data Provenance Problem

Because huge knowledge lets USA expand the info sources we tend to use for process, it is laborious to make sure that every knowledge supply meets the trustiness that our analysis algorithms need to provide correct results. Therefore, we want to rethink the believability and integrity of information employed in our tools. We are able to explore concepts from adversarial machine learning and sturdy statistics to

spot and mitigate the consequences of maliciously inserted knowledge.

Security Visualization

Visualization leverages human's extraordinary ability to detect patterns in images. Visualization technology is an emerging area today but there is an increasing amount of research and development. There are open source and commercial data visualization tools for security, but data visualization for security remains extremely elementary, dominated by pie charts, graphs, and Excel spreadsheet pivot tables.

Skilled Personnel

Appropriately skilled personnel are a critical element for successful implementation of Big Data for information security. One of the challenges in this regard is the relative shortage of such staff. Specific skills include data management expertise, data analysis expertise, and threat analysis expertise. These skills are unlikely to be found in any one person, and this means that collaborative teams of specialists will need to be formed to allow organizations to

achieve optimal results from their Big Data efforts.

CONCLUSION

Big information has recently emerged as an extremely promising paradigm for analysis of the massive volumes of heterogeneous information. Huge information technology is ever-changing info security threat landscape and furthermore as security solutions. However, despite the many opportunities offered by huge information for info security, several challenges delineate during this paper should be addressed before this potential is realised absolutely. Several key challenges during this domain, together with detection of advanced persistent attacks, detection of information escape, incorporation of rhetorical, fraud and criminal intelligence, and security visual image area unit solely getting down to receive attention from the analysis community. Therefore, we tend to believe there is still tremendous chance for researchers to form groundbreaking contributions during this field, and convey vital impact to their development within the business. During this paper, we tend to survey the progressive of massive knowledge analysis for info security, covering its essential concepts, distinguished characteristics, key technologies additionally as analysis directions. Because the development of massive knowledge technology remains at

associate in nursing early stage, we tend to hope our work can give a more robust understanding of the analysis challenges of massive knowledge, and pave the manner for any analysis during this space.

REFERENCES

1. Available at: https://en.wikipedia.org/wiki/Big_data.
2. Available at: http://www.sas.com/en_us/insights/big-data/what-is-big-data.html.
3. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2703083.
4. Available at: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_Intelligence.pdf.
5. Available at: https://www.infoq.com/articles/bigdata-analytics-for-security.
6. Available at: https://www.infoq.com/articles/bigdata-analytics-for-security.