Big Data: Big Promises for Information Security

Rasim Alguliyev
Institute of Information Technology
Azerbaijan National Academy of Sciences
Baku, Azerbaijan
director@iit.ab.az

Yadigar Imamverdiyev
Institute of Information Technology
Azerbaijan National Academy of Sciences
Baku, Azerbaijan
yadigar@lan.ab.az

***Abstract*—Big Data is related to technologies for collecting, processing, analyzing and extracting useful knowledge from very large volumes of structured and unstructured data generated by different sources at high speed. Big Data creates critical information security and privacy problems, at the same time Big Data analytics promises significant opportunities for prevention and detection of advanced cyber-attacks using correlated internal and external security data. We must address several challenges to realize true potential of Big Data for information security. The paper analyzes Big Data applications for information security problems, and defines research directions on Big Data analytics for security intelligence.**

***Index Terms*—Big Data, information security, analytics, Hadoop, data visualization.**

I. Introduction

There are different definitions of the "Big Data" term. The most popular definition is given by describing their three characteristics called "3V": Volume (the data volumes are very large which cannot be processed by traditional methods), Velocity (the data is produced with great velocity and must be captured and processed rapidly) and Variety (variety of data types: structured, semi-structured, and unstructured). Based on data quality, IBM has added a fourth V called: Veracity. However, Oracle has added a fifth V called: Value, highlighting the added value of Big Data [1].

Big Data is a relatively new term (it was only coined in 2008 [2]), but it became a very popular buzz word after publication of the report prepared by McKinsey Global Institute [3]. Now popular media is replete with publications on Big Data opportunities for government, business, healthcare, law enforcement, cyber security, research and development, etc. Industry is abuzz with the promise of Big Data [3]. National governments have recently announced significant programs on Big Data applications [4, 5]. Readers may have a misconception that big data can be used only by large companies. It should be noted that the convergence of Big Data and cloud computing technologies [6] allows small and medium enterprises using Big Data opportunities too.

While the promises of Big Data are real – they are proven by success primers of big companies like Google, Yahoo, Facebook – leading data science researchers are warning that there are many challenges at each step of Big Data analysis pipeline [7, 8].

The relation of information security and Big Data is twofold. Information security and privacy are among the most challenging issues of Big Data. At the same time, Big Data analytics promises significant opportunities for solving different information security problems. There are many reports, especially by big companies about Big Data opportunities for information security, but there are a few publications on challenges of Big Data for information security [9, 10].

The purpose of this paper is to analyze the-state-of-the-art Big Data research for information security, and to determine the most relevant research directions.

II. The Hadoop Ecosystem

Big Data framework for processing and analysis consists of a number of software tools. Currently the Hadoop software ecosystem (Fig. 1) is considered as a synonym for Big Data. Hadoop implements MapReduce technology of Google [11], which provides automatic data paralleling and processing on computer clusters. Many of the Hadoop components are open source software developed in various Apache projects [12].

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

HDFS (Hadoop Distributed File System) – a distributed file system for storage and management of date warehouses from a few terabytes to petabytes; it is the core of the Hadoop. HDFS splits the input data into blocks and allocates these blocks on servers in different places allocated to them. The TCP/IP level is used for communication. HDFS is fault tolerant, and failure of any component does not affect the overall system performance.

MapRedice – 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 [11]: 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, including key algorithms, such as classification, clustering, and recommendation and collaborative filtering. Basic algorithms are implemented with Map/Reduce paradigm on the Hadoop upper level.

Components like **Sqoop** and **Flume**, included in the ecosystem are used to transmit data to the Hadoop-clusters and vice versa.

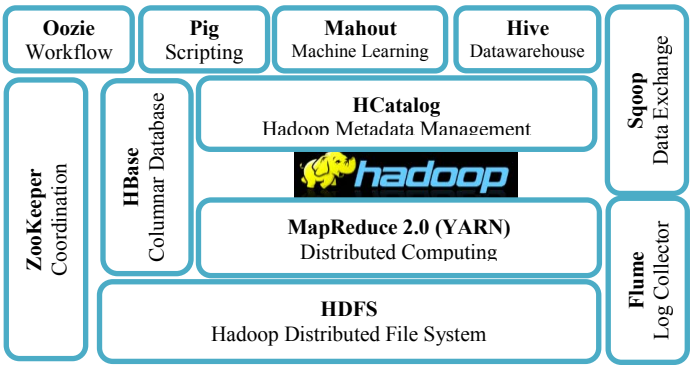

Fig. 3. The Hadoop ecosystem

Hadoop is often used in conjunction with standard data storage and processing technologies, it is sometimes added innovative solutions such as Storm, Dremel, Drill, etc. Moreover, almost all major producers of business intelligence tools add functionality to their products to analyze data permanently stored in Hadoop-clusters. We could extend the list of the Hadoop ecosystem components, because more and more companies are entering the market with products that have a connection with Hadoop.

III. 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 [13].

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 [13]:

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.

CSA distributed these tools into four groups: infrastructure security; data protection; data management; reactive security.

IV. Bird's Eye View of Cyber Threat Landscape and Security Tools

Advanced threats. Businesses and governments face an evolving threat landscape. One of the greatest challenges is presented by advanced persistent threats (APTs), which are sophisticated, long-term, multi-phase, multi-faceted attacks targeting a particular organization [14]. RSA, Google, NASA and some nation states have experienced large security breaches due to APTs. Mitigating the risk of APTs requires advances beyond traditional security defenses to include real-time threat management.

Data sources. Organizations collect a wide variety of data for security analysis and investigations: traffic data, log files (operation system, application, firewall, web access, etc.), log/event data from networking devices, DNS-specific logs/events, user activity, physical security activity data, firewall rule sets, asset data, and etc. In spite of all this, internal data collection and analysis is no longer enough. Strong risk management and incident detection/response practices are also being supplemented with growing volumes of external security data.

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 to correlate higher-level events (hacker intrusions, insider

actions, Trojan attacks) based on rules for simple events (triggering of IDS and antivirus signatures, firewall errors, incorrect passwords).

It should be noted that problems of testing and evaluation of IDSs found their reflection in many research papers [15]. Unfortunately, it is not case for SIEM systems. There are a few papers on SIEM systems testing (SC Magazine evaluations) and benchmarking [16].

V. Big Promises for Cyber Security

New Big Data technologies – such as the Hadoop ecosystem, stream mining, complex-event processing, and NoSQL databases – are enabling the analysis of large-scale, heterogeneous datasets at unprecedented scales and speeds. These technologies are allows extending traditional information security systems by facilitating the storage, maintenance, and analysis of security information. Analysis of data from different sources in different formats, the ability to compare these data, anomaly detection, and combating cyber threats in real time – all this has been made possible through the use of technologies for processing and analyzing Big Data.

Many companies offering security solutions published white papers, emphasizing the advantages and opportunities of Big Data for security [17, 18, 19]. 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 [20].

RSA recommends gradually move to the Intelligence-Driven Security model [19]. 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.

In general, highly scalable systems based on the principles of Intelligence-Driven Security, should have the following properties:

- use advanced monitoring subsystem to monitor a diverse array of sources and create synergies by combining information from different sources;
- include automated tools for collecting and processing Big Data and preparing results in a standardized format, accessible to other subsystems;
- include a central repository, powerful analytical tools and efficient visualization tools enabling to get useful knowledge from raw data.

VI. 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.

1) **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.

2) **APT detection by Big Data analytics.** There is need for new detection algorithms, capable of processing significant amounts of data from diverse data sources. Currently, a small number of proof of concept deployments that utilize Big Data analytics for security event detection exist, and show promising results [21, 22, 23].

3) **High performance cryptography –** encryption and decryption algorithms; encrypted data search, attribute-based encryption, attacks on the availability, reliability, and integrity of Big Data [24].

4) **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 [21].

5) **Data provenance problem.** Because Big Data lets us expand the data sources we use for processing, it's hard to be certain that each data source meets the trustworthiness that our analysis algorithms require to produce accurate results. Therefore, we need to reconsider the authenticity and integrity of data used in our tools. We can explore ideas from adversarial machine learning and robust statistics to identify and mitigate the effects of maliciously inserted data [25, 26].

6) **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 [27, 28]. There are open-source and commercial data visualization tools for security [29], but data visualization for security remains extremely elementary, dominated by pie charts, graphs, and Excel spreadsheet pivot tables.

7) **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.

VII. Conclusion

Big Data has recently emerged as a highly promising paradigm for analysis of the large volumes of heterogeneous data. Big Data technology is changing information security threat landscape and as well as security solutions.

However, despite the significant opportunities offered by Big Data for information security, many challenges described in this paper must be addressed before this potential can be realized fully. Many key challenges in this domain, including detection of advanced persistent attacks, detection of data leakage, incorporation of forensic, fraud and criminal intelligence, and security visualization are only starting to receive attention from the research community. Therefore, we believe there is still tremendous opportunity for researchers to make groundbreaking contributions in this field, and bring significant impact to their development in the industry.

In this paper, we survey the state-of-the-art of Big Data research for information security, covering its essential concepts, prominent characteristics, key technologies as well as

research directions. As the development of Big Data technology is still at an early stage, we hope our work will provide a better understanding of the research challenges of Big Data, and pave the way for further research in this area.

REFERENCES

[1] A. Baaziz, and L. Quoniam, "How to use Big Data technologies to optimize operations in Upstream Petroleum Industry," International Journal of Innovation, vol. 1, no. 1, pp. 19-29, 2013.

[2] Editorial: "Community cleverness required," Nature, Vol. 455, No. 7209, pp. 1-1, 4 September 2008. http://www.nature.com/news/specials/bigdata/index.html

[3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, May 2011.

[4] Big Data Research and Development Initiative. March 2012. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

[5] The Australian Public Service Big Data Strategy. August 2013 http://agict.gov.au/sites/default/files/Big%20Data%20Strategy.pdf.

[6] D. Agrawal, S. Das, and A. El Abbadi, "Big data and cloud computing: current state and future opportunities," Proc. of the 14th International Conference on Extending Database Technology, pp. 530-533, 2011.

[7] A. Labrinidis, and H. V. Jagadish, "Challenges and Opportunities with Big Data," Journal Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 2032-2033, August 2012.

[8] K. Michael, and K. Miller, "Big Data: New Opportunities and New Challenges," IEEE Security & Privacy, vol. 46, no. 6, pp. 22-24, June 2013.

[9] A. A. Cardenas, Manadhata P. K., and Rajan S. P., "Big Data Analytics for Security," IEEE Security & Privacy, vol. 11, no. 6, pp.74-76, June 2013.

[10] T. Mahmood, and U. Afzal, "Security Analytics: Big Data Analytics for cyber security: A review of trends, techniques and tools," Proc. of the 2nd National Conference on Information Assurance (NCIA), pp. 129-134, 2013.

[11] J. Dean, and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Proc. of the 6th Conference on Symposium on Opearting Systems Design & Implementation (OSDI'04), Vol. 6, pp. 137-150, 2004.

[12] T. White Hadoop: The definitive guide. O'Reilly Media, Inc., 2012.

[13] Cloud Security Alliance (CSA): Expanded Top Ten Big Data Security and Privacy Challenges, April 2013.

[14] A.K.Sood, R.J.Enbody, "Targeted Cyberattacks: A Superset of Advanced Persistent Threats," IEEE Security & Privacy, vol. 11, no. 1, pp. 54-61, 2013.

[15] S. Zanero, "Flaws and frauds in the evaluation of IDS/IPS technologies," 19th Annual Conference of the Forum for Incident Response and Security Teams (FIRST), 2007.

[16] J. M. Butler, Benchmarking Security Information Event Management (SIEM)," A SANS Whitepaper, February 2009.

[17] J. Oltsik, "IBM: An Early Leader across the Big Data Security Analytics Continuum". White paper, June 2013

[18] M. Bouchard, "Big Data for Advanced Threat Protection." White paper, 2012.

[19] S. Curry, E. Kirda, E. Schwartz, W. H. Stewart, and A. Yoran, "Big Data Fuels Intelligence-Driven Security". White paper, January 2013.

[20] Cloud Security Alliance (CSA): Big Data Analytics for Security Intelligence. September 2013. https://cloudsecurityalliance.org/download/big-data-analytics-for-security-intelligence

[21] Dumitras T., Shou D., Toward a Standard Benchmark for Computer Security Research: The Worldwide Intelligence Network Environment (WINE) / Proc. EuroSys BADGERS Workshop, ACM, 2011, pp. 89–96.

[22] J. François et al., BotCloud: Detecting Botnets Using MapReduce / Proc. Workshop Information Forensicsand Security, 2011, pp. 1–6.

[23] T.-F. Yen et al., "Beehive: Large-Scale Log Analysis for Detecting Suspicious Activity in Enterprise Networks," Proc. Ann. Computer Security ApplicationsConference (ACSAC 13), pp. 199-208, Dec. 2013.

[24] Ganugula U., Saxena A., "High Performance Cryptography: Need of the Hour," CSI Communications, pp. 16-17, September 2013.

[25] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, J. D. Tygar, "Adversarial machine learning," Proc. of the 4th ACM workshop on Security and artificial intelligence (AISec'11), pp. 43-58, 2011.

[26] P. J. Huber, E. M. Ronchetti, Robust Statistics. Wiley, 2009.

[27] A. Shiravi, H. Shiravi, and A.A. Ghorbani, "A Survey of Visualization Systems for Network Security," IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 8, pp. 1313-1329, Aug. 2012.

[28] Marty R., Applied Security Visualization, Addison-Wesley Professional; 1st edition, 2008. 552 p.

[29] Mittelstadt S., Behrisch M., Weber S., Schreck T. et al, "Visual analytics for the big data era – A comparative review of state-of-the-art commercial systems," Proc. of the IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 173-182, 2012.