BIG DATA ANALYTICS IN SUPPORT OF INFORMATION SECURITY

Popov Veselin
PhD, Associate Professor
"D.A.Tsenov" Academy of Economics, Svishtov
e-mail: v.popov@uni-svishtov.bg

Emilova Petya
PhD, Associate Professor
"D.A.Tsenov" Academy of Economics, Svishtov
e-mail: p.emilova@uni-svishtov.bg

Abstract
Along with the many opportunities that Big Data provides to users (business users and individuals), it brings many threats to society, businesses and individuals. At the same time, Big Data Anapytics has established itself as an effective tool in many areas of modern information processing, including the area of information security. Big Data technologies used in security systems are able to detect threats in advance.
By providing tools for collecting and analysing large amounts of digital information generated from different sources or recorded by different devices, BDA helps to detect and define patterns and trends of malicious behaviour, search for methods to track cybercriminals, predict and stop potential cyber-attacks.
This report focuses on the problems and challenges of protecting Big Data in these two aspects.

Keywords: *Big Data, Big Data Analytics, Information Security*

JEL Classification: *D80*

INTRODUCTION

The globalization of business and the increasing use of cloud and mobile services have brought with them significant challenges in information security and the need for new tools (other than traditional ones) to protect and detect malicious activities in corporate networks.

Big Data Analytics (BDA) has established itself as an effective tool in many areas of modern information processing, including information security. By providing more data and applying various analytical techniques, with BDA all real and potential risks can be analyzed very quickly, more alternatives for protection and counteraction can be assessed, more accurate forecasts for the future development can be made, more expert opinions can be collected. With this tool, IS security professionals can quickly and cost-effectively perform complex simulations and test multiple possible security scenarios.

This report explores the opportunities and challenges of Big Data and Big Data Analytics in terms of information security.

BIG DATA – SOURCES, FEATURES, PROTECTION, PLATFORMS

In the literature, the term "Big Data" refers to huge aggregates of information that are stored, transmitted and processed in computer systems. Today, the main sources for generating Big Data are social media, a huge variety of smart devices, video, digital images, sensors and records of commercial transactions (Techopedia). Apart from being huge, these sets of information are also complex, which is why traditional processing applications cannot handle them.

According to the National Institute of Standards and Technology (NIST, 2018), Big data are those data in which the volume, required speed of data processing or presentation

limits the possibilities for effective analysis with traditional relational approaches or requires significant horizontal scaling to ensure efficient processing. This created the need for a new generation of software applications and tools.

Big Data differs from traditional databases in four main characteristics related to: the volume of the data set; the speed of data generation and transmission; the diversity of data in the form of different types of structured and unstructured data; the complexity of the structure, behaviour and permutations of datasets when different critical factors are at work.

One of the most serious challenges for Big Data platforms is the protection that must be implemented at every stage of the platform's life cycle and uses a combination of *traditional* security tools, *new* tools and technologies, as well as *intelligent security monitoring processes*.

The Big Data lifecycle includes four stages: data organization, processing, data warehouse maintenance, processing, and production.

Data organization includes the creation and/or access to data. Big Data is a collection of various types of data that come from different sources. According to sources, Big Data is of three types: data generated by humans – text, photos, video; data generated by machines - computers operating documents, databases, multimedia, GPS, RFID, the so-called smart "homes"; and data generated by various digital devices (objects).

For example, user-generated data includes CRM data, data such as emails, telephones, SMS or social media posts, and more. There is a lot of transaction data and data stored in different databases. There is also a huge amount of data generated by custom software and sensors. All data transited from sources to the platform must be protected.

Storage in a data warehouse. Security tools such as encryption, user authentication, intrusion protection are used for protection. These security tools must operate in a distributed platform with many servers and branch nodes. Security tools must protect log files and analytics tools when they operate on the platform.

Many organizations use cloud services to secure a data storage solution. Cloud technologies also allow companies to use pre-built Big Data solutions, or quickly create or deploy powerful server complexes without significant hardware acquisition costs. In addition to advantages, new technologies pose challenges to protection.

Analytical processing (analysis and results) of different types of data is the essence of Big Data. The results obtained from this processing are directed to applications, reports, and dashboards are the target of attacks. Therefore, encryption of results, access control and traffic are extremely important.

The classic technologies used for protection in Big Data are: encryption, user access control, intrusion detection and counteraction, physical protection. Encryption protects data both when it is stored in repositories and when it is transited. It should be noted that encryption must work with different types of data and with different analytical tools, relational and non-relational databases, special file systems, etc. User access control is a key tool for protection at the network level. It is important for the Big Data platform; it must be very precise and based on well-defined policies. Intrusion detection and response systems are also important for the Big Data platform as they contribute to the timely detection of intrusion attempts. Physical protection is related to the protection of the building and rooms of the data centre of the organization or the cloud provider.

Hundreds of software tools and complex platforms are available on the Big Data processing *software market*. Some of them have a long history, while others have appeared recently. Large enterprise software solution manufacturers, such as Oracle, IBM, SAS,

SAP, Teradata, Microsoft and others, dominate this market. Their Big data products are characterized by being integrated with complex business management solutions.

Successful products are also offered by smaller software companies specializing in Big data such as Tableaau, RapidMiner, Pentaho, Alteryx, Alpine and others.

Open source products such as Apache Hadoop, MapReduce, GridGain, Storm and others are also available on the market. Some open source software solutions are often included in commercial vendor projects.

The best software tools for processing Big data, according to their purpose, functions performed and implemented level of protection are (import.io, 2018):

- for data storage and management – Haddop, Cloudera, MongoDB, Talend;
- for data cleaning – OpenRefine, DataCleaner;
- to extract knowledge from data – RapidMiner, IBM SPSS Modeler, Oracle data mining, Teradata, FramedData, Kaggle;
- for data analysis – Qubole, BigML, Statwing;
- for data visualization – Tableau, Silk, CartoDB, Plot.ly, Datawrapper;
- for data integration – Blockspring, Pentaho;
- data languages – R, Python, RegEx, XPath;
- for data collection – Import.io.

BIG DATA ANALYTICS FOR INFORMATION SECURITY

From a structural point of view, the information protection process includes three phases: *prevention – detection – response*. The BDA has a huge potential for the strategies and activities of the second phase – the phase of detecting crimes and potential threats.

The analysis of Big data is significantly more complex than that used in traditional databases. Big data have large volumes, non-aggregated, in different formats and their processing is difficult to do in the memory of only one computer. Big data processing includes mechanical processes and algorithms. The methods used for Big data analysis are of two main types - *responsive and predictive analysis* (Huang & W. Chaovalitwongse, 2015).

The response analysis aims to produce statistics on current and historical data and to provide information on what happened and why it happened. It includes methods such as statistical modelling, trend reporting, visualization, association and correlation analysis.

Predictive analysis focuses on the use of known data (training data), which include input data properties (attributes) and response values (target models) to build a predictable model (solution) to make predictions for invisible data (test data). It uses methods such as vector machines, linear regression / classification, nonlinear regression (generalized linear model), decision tree, Bayes theory, neural networks and others.

Big Data technologies used in security systems are able to detect threats in advance. For example, they can detect atypical behaviour on the network, predict an attack, and analyse the sources of an attack.

Big Data creates conditions for efficient and effective application of some fraud detection techniques. In the specialized literature, they are divided into two main groups: statistical techniques and techniques using *artificial intelligence*. Table 1 presents examples of these techniques.

Table 1. Examples of Big Data Analytics techniques

Statistical techniques
1. Techniques used in data pre-processing to detect, validate, correct errors, fill in missing and inaccurate data.
2. Calculation of various statistical parameters such as averages, valves, efficiency indicators, probability distribution, etc.
3. Models and probability distribution of different business activities.
4. Processing of user profiles.
5. Analysis of time-dependent data series.
6. Clustering and classification to detect possible models / schemes and dependencies / associations in data groups.
7. Combining algorithms to detect anomalies in the behavior of transactions or users
Techniques using artificial intelligence
1. Data mining
2. Expert systems.
3. Automatic pattern recognition
4. Machine learning techniques
5. Neural networks

Source: (BajpaiI & Arushi, 2018)

CHALLENGES

Big Data Working Group, defines four aspects of Big data protection, which are: infrastructure security; data confidentiality; data management, integrity and reactive security (Cloud Security Alliance, 2013). Each of the areas is associated with many problems. For example, infrastructure security has the following issues:

- Availability of single-layer protection - companies must include multi-layered defence within the company's defence strategy, which addresses both internal and external security threats..
- Data transfer across multiple devices, which requires additional levels of security and monitoring to ensure that data is not captured along the route from one device to another.
- Rapid development of big data technology and its supporting infrastructure (such as cloud services), which must be able to process data from an infinite number of points with speed, security and reliability.

The infrastructure must include security measures to keep information at every stage of the process.

In fact, in addition to enhancing business intelligence, Big Data provides an opportunity to enhance information security. This increases the existing problems and challenges that require attention and await solutions.

The main problems with Big Data security are related to: threats to data security; privacy risks; the need to confirm the authenticity; there is no technology to protect the confidentiality of big data.

These Big Data security issues require addressing many challenges, the most important of which are:

a) Acceptance of protection as a top priority for Big Data platforms, which will focus the attention of managers and developers in this direction.

b) Introduction of reactive and proactive protection.

c) The physical protection of devices and servers that contain sensitive information must be managed with special care and isolated from other devices.

d) Application protection is as important as device protection. In this regard, the protection of Data mining solutions is of particular importance. These solutions are the basis of Big Data platforms, contribute to the discovery of patterns of behaviour and development trends and on this basis offer business strategies. This importance of Data mining solutions requires that they be protected not only from external threats, but also from internal individuals who abuse their privilege to access sensitive information

e) High level of access control to be provided by encrypted authentication and to determine which individual can see what data.

f) Use of real-time protection tools. These tools generate a huge amount of information. The problem here is to ignore unimportant signals so that employees can focus on the real violations.

g) Data warehousing management. For the Big Data architecture, it is typical for data to be stored at multiple levels, depending on its importance to the business and the cost of storing it.

h) Carrying out a detailed audit, which can help determine when missed attacks can occur, what were the consequences, what to change in the current system.

i) Use of distributed systems. For faster analysis, most Big Data platforms actually distribute the huge amount of data processing work across many systems.

CONCLUSIONS

Along with the many opportunities that Big Data provides to consumers (business users and individuals), it brings with it many threats to society, business and individuals. At the same time, the BDA has established itself as an effective tool in many areas of modern information processing, including information security. This report addresses the challenges and challenges of Big Data protection in these two aspects.

BIBLIOGRAPHY

1. Bajpai, A., & Arushi, A. (2018). Big Data Analytics in Cyber Security. *International Journal of Computer Sciences and Engineering, 6*(7).
2. Cloud Security Alliance. (2013). *Expanded Top Ten Big Data Security and Privacy Challenges.* Retrieved from Expanded top Ten Big Date Security and Privacy Challenges: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf
3. Huang, S., & W. Chaovalitwongse, W. (2015). *Computational Optimization and Statistical Methods for Big Data Analytics: Applications in Neuroimaging.* INFORMS TutORials. Retrieved from Computational Optimization and Statistical Methods for Big Data Analytics: Applications in Neuroimaging: http://faculty.washington.edu/artchao/INFORMS_Tutorials_2015-Web.pdf
4. import.io. (2018). *All the best big data tools and how to use them.* Retrieved from import.io: https://www.import.io/post/all-the-best-big-data-tools-and-how-to-use-them/
5. NIST. (2018). *Big Data Information.* Retrieved from NIST: https://www.nist.gov/el/cyber-physical-systems/big-data-pwg
6. Techopedia. (n.d.). *Big Data Analytics.* Retrieved from Techopedia: https://www.techopedia.com/definition/28659/big-data-analytics