Makrufa S. Hajirahimova
Institute of Information Technology of ANAS, Baku, Azerbaijan
makrufa@science.az

DOI:10.25045/jpit.v07.i1.06

BIG DATA TECHNOLOGIES AND INFORMATION SECURITY CHALLENGES

The notion of Big Data, which is a new era in information processing, is widely discussed in the business environment, mass media and scientific literature. Big Data technology can make revolutionary changes in the field of governance, health, science, business, e-commerce, industry and others. On the one hand, it opens up new opportunities for the society; on the other hand, it causes new security problems. The article provides a brief description of the Big Data technology and investigates the use of Big Data analytics and some security issues. In addition, it considers the new ethical problems created by Big Data in terms of personal data protection and offers some recommendations.

Keywords: *big data, big data analytics, security, anonymisation, privacy, identification, de-identification, encryption*

Introduction

Big Data (BD) problem is not a new subject. With regard to the technical and technological development, the occurance rapid increase of data and its processing problems date back to the 40s of the previous century. Indeed, in comparison with that period, since the early twenty-first century, digital data has been increasing in geometric progression each year [1]. Analytical reports of companies such as IDC, Gartner and others also acknowledge this. [2]. Web, social networks, mobile devices, transactions made via credit cards and so on led to an expansion of the digital data flow, information abundance and the world was practically filled with information. Thus, in 2015, the amount of information around the world amounted 8.1 zettabyte (81x103 exabyte), while the figure for the data generation, since the existence of mankind in the world up to 2003. was about 5 exabyte (5x109Gbyte). This figure is predicted to increase by 40% each year, and reach 44 zettabyte by 2020 [2]. As a result, data processing, storage and use emerged the term *Big Data* reflecting a new era [3]. The term is associated with the sets of large-scale digital data created by the government agencies, large companies and etc., and subsequently are analyzed for different purposes, which can not be processed with the traditional databases and tools. Management of this data requires significant storage capacity and computing power. To solve this problem, one of the giants of the information technology industry, Google developed Google File System and MapReduce [4] software and hardware platform in 2004. Based on this platform, open-source Apache Hadoop and Hadoop File System [5] software platforms was developed for BD processing and analysis, which founded big data technologies.

Any technological achievement can be utilized for either prosperity or adverse purposes. As the other technologies, Big data is also the two folded – has its benefits and threats. On one hand, BD is a source of knowledge that is able to change radically all spheres of society, and it opens up new horizons for business. On the other hand, the more we digitize and accumulate additional information, the more information becomes available, and the more the number of users increases, hence we end up creating potential opportunities for attackers. In terms of information security, information theft and distortion, networks hacking, and personal data theft, and etc. by these attackers is eased [6]. Private data is analyzed without the consent of its owners. This, in turn, is ethically and legally unacceptable, and a very serious problem in terms of security and privacy. The problems can be reviewed from different perspectives: application of big data analytics for information security, or information security in big data analytics [7-11]. These are very topical issues for scientific researchers. The article aims at analyzing the key security problems of big data technologies and the problems of confidential data protection.

Big data analytics and information security issues

Today, the abundance of available information, called the Big Data, definitely exists. The "3Vs" is the first model which determines Big Data and distinguish it from others [12]. This model has great *Velocity*, and it reflects idea of acquiring more valuable information through the efficient usage, storage and analysis of a large *Volume* of data collected from *Variety* of sources. It should be noted that, IBM included the 4th "V" (veracity) by taking the authenticity of data into account, and Oracle involved the 5th "V" (value) by highlighting the value of BD [1,2,13]. Recently, experts has been increasing the number of "V"s.

The key idea of BD is the ability to process and analyze very large volumes of data in real-time. The increase in the volume of information should not disappoint people. On the contrary, it should be referred as the natural raw material and resource. As mentioned, BD is capable of changing the world with the revolutionary innovations in business and public administration. Because, this raw data includes in-depth knowledge that is conducive to scientific discoveries. However, maximum use of this resource requires the new generation of analytical technologies to add value for community and business. In this regard, decision makers and politicians of either business sector, media, or government agencies, focus on BD. Existance of BD is reviewed as a subject of scientific research in a new quality. The demand for real-time analysis of maximum data has led to the emergence of Big Data Analytics, which finds a correlation between different settings, features, events and so on, and classifies and develops analytical reports, and forecasts based on this analysis [1, 13-16]. It should be noted that BD analytics may also yield incorrect correlations.

BD analytics improves the efficiency of business processes in terms of corporate interests and expands marketing efforts. Ventures may control their incomes and expenses, and improve their financial performance and increase transparency with the help of BD collection and analysis. Joint analysis of (structured and unstructured) data, generated continuously from different sources as a result of the mutual interaction as Machine-to-machine (M2M), and obtaining new knowledge and useful information from them is extremely important for new scientific discoveries. This technology is very important for making the right decisions in public and private institutions, as well as for the protection of the laws, detection of national security and terrorism incidents, pre-determination of disease epidemics, revealing hidden behavior of the people, understanding their goals and intentions, and their interaction with the environment, as well as better understanding of economic risks of financial sector in national level, guiding politicians and regulatory bodies and better management of risk systems [1].

As research company Gartner notes, BD analysis will play a key role to identify crimes and safety violations, and 25% of large companies will use big data analytics in their cyber-security systems by 2016. Gartner experts state that Big Data analytics will ensure more reliable information security system [17].

However, the application of big data technologies revealed that existing security models are outdated. Security approach used 15 years ago is not adequate now. At the same time, volume, variety and velocity of BD intensify security and privacy problems [18,19]. The diversity of the sources and flow of data collection, a large-scale "cloud computing" infrastructure and migration of large-scale data to these "clouds" have revealed the vulnerabilities of security systems. Thus, traditional security mechanisms are not sufficient for BD's expansion. Because, the data flow requires a very flexible and fast security solutions.

Cloud Security Alliance (CSA) presents ten security problems of Big data systems and classifies them in four groups [20]:

- infrastructure safety (security measures of distributed application environment, the best security experience for the non-relational data warehouse);
- data confidentiality (scalable and composite data mining and analytics, access control carried out with cryptography, and secure communication, granular access control);

- data control (strengthening the supervision of the list of data storage and transfer, data origin, granular audits);
- integrity and responsive security (real-time security monitoring, access authorization/filter).

The classification shows that, to ensure the security infrastructure of big data systems, distributed computing and data warehouse must be protected. First and foremost, cryptography and granular access control tools should be used for the security of confidential data *per se*. Large-scale data management requires scalable and distributed solutions to control the data effectively and determine its origin. Finally, the integrity of data streaming out of various points should be checked, and security incidents should be analyzed in real-time.

In security and privacy problems, usually three concerns need to be addressed:

1. Modeling: formulating security model covering cyber-attacks or data leaks scenarios;
2. Analysis: finding possible solutions with the security model;
3. Realization: applying found solution in the existing infrastructure.

4-Layer Security model is applied in all Hadoop products to ensure safety [21] (figure 1).

Perimeter Security: responses to the user authentication and solves network security problems with the Kerberos network protocol created by the Massachusetts Institute of Technology. Kerberos is a tool designed to provide a reliable authentication in client-server applications with secret-key cryptography. In other words, it provides authentication and sustainable cryptography tools over the network to ensure the security of information systems. Kerberos is considered to be de-facto standard for the verification of authenticity, and available for everyone [10].

Data Access Security: designed to ensure users to access only the data, not services and resources.

Accountability: general purpose of this security layer is to promote the accountability. It allows the administrators to audit data access in Hadoop. In addition, it allows to identify the origin of the data, i.e., the source where data is derived from. Navigator was developed specifically to support the safety of this layer.

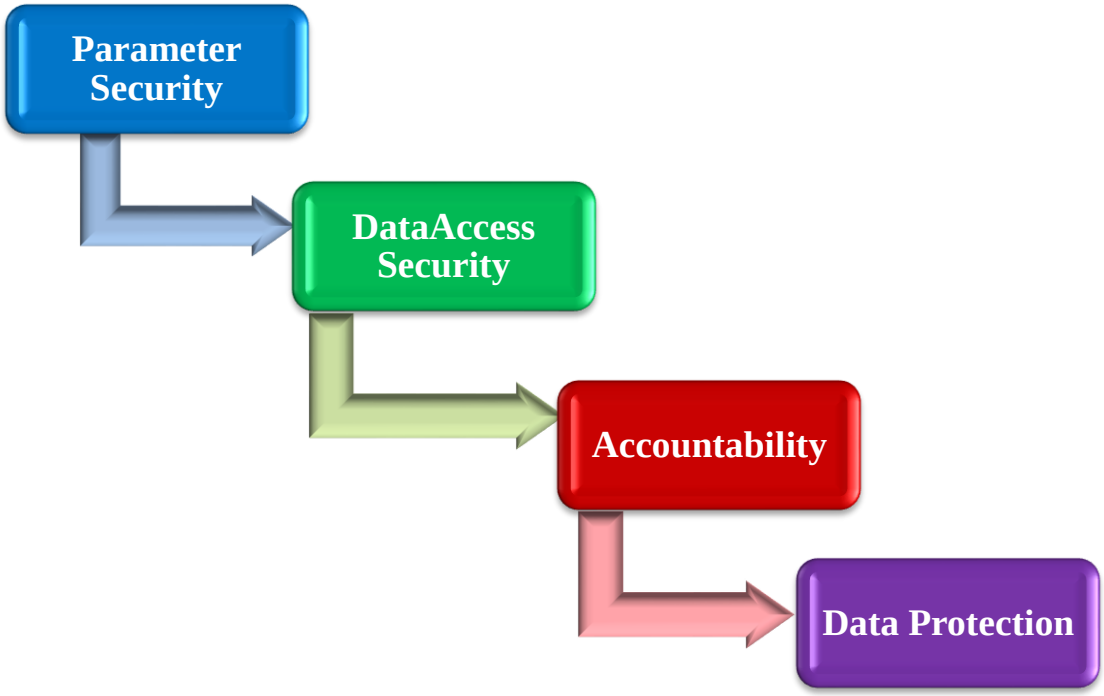

Figure 1. Data security model in Hadoop system

Data protection: the last security aspect that covers data encryption, masking and other fields [8].

Big data and data privacy

About 30 years ago, it was relatively easier to protect people's private lives and to ensure their anonymity. Because, there were fewer automated systems collecting and storing personal information, the Internet was very primitive, and too few people even knew about its existence. Everything has changed in recent years; digitization (GPS signals, mobile phones, e-mail, e-shopping, chats in social networks, e-medicine records, and etc.) has been widely spread. The confidentiality and anonymity of personal data has been left far behind by collecting information about the customers in marketing, banking, insurance, medical and other fields.

Although, Big data provides exceptionally useful information, it is culprit for new ethical challenges related to security and privacy. As Data Mining technology evolves and becomes widely used, acquisition of more in-depth knowledge from BD poses a threat to the protection of extremely sensitive personal data. The data collected about the people by both companies and public agencies, became uncontrolled, and the boundary between open and secret data is getting vanished. It is also a serious problem in terms of the data privacy [7].

Personal data has public and confidential categories. Public data includes the information unnamed in the specified manner, and announced as public data by a user or made available with his/her consent. The first name, last name and middle name of the individual are permanent public data. Unlike confidential information, public personal data do not require ensuring confidentiality. Confidential personal information, except in cases prescribed by law, may be transferred to the third parties only with the consent of user [22].

Data directly reflects human activity. People take medical treatments, use e-services, search on web sites and make phone calls. Their location, relations with family members, political activities, social coverage and interests are constantly monitored through smart phones and apps they use, and collected by companies. Today, the majority of this information is collected and used without the consent of the people. Insurance companies and banks refuse to give credits to their customers by knowing their debts; markets are aware of the pregnancy of daughter before the parents; doctor get the information about the patients based on e-health records, even if they try to hide it [23]. This also contradicts the principles and rules of data privacy.

One of the dangers that may arise is that people can be prosecuted as criminals by accident. Thus, the rights of any passers captured in the videos or audios unintentionally, which are shared in the social networks, might be violated. For example, while investigating the Boston explosions, several people, who were appeared in the photographs shared in social networks, were among the suspects in the terrorist attack.

Today, modern information technology factors as BD, analytics and cloud technologies are impossible to be imagined apart from each other. Cloud technologies are considered to be an extremely successful approach to computing and data storage [24]. However, at present, a safe and effective storage of data in the cloud is the biggest problem in this field. There is no reliable way to guarantee the protection of data stored in the cloud. [25] offered a method to create a safe computing environment to ensure the security of the cloud computing system. This method allows users to store data in the cloud safely and effectively. BD management and security problems are initially solved by encryption and compression data, when it is uploaded into the cloud. Investigation of encryption technologies protects the reliability of the system and improves security and the use of data storage systems.[25].

However, people still concern about the privacy of their personal information and storing in the clouds in general. The problem became more severe with the development of big data mining and analytics [16]. Because, as the other services, based on personalized and localized databases, it also requires personal information to achieve relevant results. Personal information is often subjected to the risks such as accurate inspection and profiling concerns, theft and loss [18].

The companies responsible for the security of personal data usually use de-identification methods, including, anonymization, pseudonymization, encryption, key-coding and so on.

Anonymization ensures privacy by deleting the name, address, and social insurance (protection) numbers, while pseudonymization replaces this information with pseudonym and artificial identification. Key-coding encodes personal information and creates the key for decoding. The inention to submit personal data to Big Data systems has its certain aspects. To this end, the following security solutions should be provided:

- collection and processing of Big Data ensuring privacy and safety;
- implementation of Big Data analysis in secure environment and ensuring privacy;
- implementation of data storage and storage policies of Big Data systems in secure (and privacy) mode.

Otherwise, the users hesitate to submit the data to Big Data systems. Confidential personal data should be protected by the owner, operator and users, who have access to this information, in compliance with the requirements specified by law. Personal data collection, processing and protection should be governed by regulations [22,26].

Big data analytics accurately monitors the behavior of individuals even without their consent. These opportunities contradict with two fundamental principles of data protection (data privacy and availability). Electronic medical data, new treatment techniques (with the help of sensors installed on the patient's body) is an important step forward in the field of medicine. However, most patients are extremely sensitive with this condition. Thus, even the analysis of large-scale registration data implemented anonymously via mobile devices, vanishes the individual nature of the data, and enables defining the user's (patient) personal parameters. The user's network parameters, spatial data and other indicators, which are accessible for analysis, are most likely to determine his/her identity.

Another issue is related to security in the field of computer technology (cyber security). Information attacks, threats and risks should be assessed from the perspectives of technical solutions that are adapted to big data phenomenon and to the results of possible crimes in the sphere of information. Information security policies and principles needs to be reconsidered. These policy and principles needs to be designed to follow data protection law strictly and respect private life.

Conclusion

At present, science, engineering and technology are producing exabytes and zettabytes of BD flows. The more the volume of digital information increases, the more its availability and the number of subjects who use it rises. Big data technology mainly promises to save lives, improve services and acknowledge the universe, by using the skills of researchers.

The use of Big data analytics technologies in the information security systems may reduce national security risks, terrorism, smuggling and so on in many areas. In this case, the used data may include the information transmitted through websites, medical devices, sensors and etc. Some of it may either be very simple (available for everyone), or reflecting people's private lives. This data can affect the person's insurance payments, online purchase payments, as well. Since the vast majority of personal information should be especially protected. Appropriate national and international legislation that guarantees the protection of personal data must be adopted. Analysis of private data without person's consent is unacceptable. Therefore, both researchers, and manufacturers should focus on security issues of processing and analysis of the information through Big Data technologies.

By summarizing the impact of Big Data technologies on the modern society, we conclude that as any other new technology, it also brings troubles. Although it is quite a powerful tool, it has certain drawbacks, limitations and hazards. Nevertheless, Big Data is a leap forward for humanity, and the right step towards development encompassing science, technology and business.

References

1. Aliguliyev R.M., Hajirahimova M.Sh. Big data phenomenon: Challenges and Opportunities // Information Technology, 2014, No 1, pp. 3-16.
2. The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study report, IDC, December 2012. www.emc.com/leadership/digital-universe.
3. Diebold F.X. Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting. Discussion Read to the Eighth World Congress of the Econometric Society, 2000.
4. Dean J., Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters / Proceedings of the Sixth Symposium on Operating System Design and Implementation, vol.6, 2004, pp.137–150.
5. Hadoop. http://hadoop.apache.org/.
6. Alguliyev R., Imamverdiyev Y., Yusifov F. Some conceptual views on information security of the society // Journal of Communication and Computer, 2012, vol.9, pp. 644-648.
7. Lei X.; Chunxiao J., Jian W., Jian Y., Yong R., Information Security in Big Data: Privacy and Data Mining // IEEE Access**,** 2014, vol.2, pp.1149–1176.
8. Big data and data protection.https://ico.org.uk/media/for-organisations/documents/1541/
9. Big data and privacy. A technological perspective. White House, May 2014.
10. Big data and privacy, MIT 2013. http://bigdata.csail.mit.edu
11. Alguliyev R., Imamverdiyev Y. Big Data: Big promises for information security / Proceedings of the IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), Astana, Kazakhstan 15-17 October, 2014, pp.216–219.
12. Laney D., 3D Data Management: Controlling Data Volume, Velocity and Variety. Technical report, META Group, Inc (now Gartner, Inc.), February 2001. http://blogs.gartner.com/
13. Gandomi A., Haider M. Beyond the hype: Big data concepts, methods, and analytics // Ted International Journal of Information Management, 2015, vol.35, pp.137–144.
14. Kambatla K., Kollias G., Kumar V., Grama A. Trends in Big Data Analytics // Parallel Distributed Computing, 2014, vol.74, no.7, pp.2561–2573.
15. Wu X., Zhu X., Wu G.Q., Ding W. Data mining with big data // IEEE Transactionson Knowledge and Data Engineering. 2014, vol.26, no.1, pp.97–107.
16. Che D., Safran M., Peng Z. From big data to big data mining: challenges, issues, and opportunities, in: B.Hong, X.Meng, L.Chen, W.Winiwarter, W.Song (Eds.), Database Systems for Advanced Applications, Springer, Berlin Heidelberg, 2013, pp.1–15.
17. http://cloudtimes.org/2014/02/12/gartner-report-big-data-will-revolutionize-the-cybersecurity-in-next-two-year/
18. Tene O., Polonetsky J. Big Data for All: Privacy and User Control in the age of analytics // Northwestern Journal of Technology and Intellectual Property, 2013, vol.11, no.5, pp.239–273.
19. Danger: 3 Reasons to Be Scared of Big Data. http://smartdatacollective.com/
20. Cloud Security Alliance, Expanded Top Ten Big Data Security and Privacy Challenges, 2013.
21. Malbrecht T., Prekopcsak Z. Big Data Security on Hadoop. www.rapidminer.com
22. Law of the Republic of Azerbaijan on Personal information, 2010, May 11. http://www.dmx.gov.az/userfiles/files/ferdimelumatlarqanun3.pdf
23. Duhigg C. How companies learn your secrets. New York Times, 2012, 16 February.
24. Agrawal D., Das S., Amr El Abbadi. Big Data and Cloud Computing: Current State and Future Opportunities / Proceedings of the 14th International Conference on Extending Database Technology, 2011, pp.530–533.
25. Yin C., Wang J., Xie C., et.al. Robot: An efficient model for big data storage systems based on erasure coding / IEEE International Conference on Big Data, 2013, pp.163–168.
26. Protection of personal data, 1995. http://eur-lex.europa.eu/legal-content/EN/