Jan 17th, 12:00 AM

Security issues in data analytical environments

Mehmet Emin Yazici
University Göttingen, Germany, mehmetemin.yazici@stud.uni-goettingen.de

Ilja Nastjuk
University Göttingen, Germany, ilja.nastjuk@wiwi.uni-goettingen.de

Security issues in data analytical environments

Mehmet Emin Yazici[1], Ilja Nastjuk[1]

[1] Chair of Information Security and Compliance, Göttingen, Germany
mehmetemin.yazici@stud.uni-goettingen.de
ilja.nastjuk@wiwi.uni-goettingen.de

Abstract. Nowadays, data is ubiquitous and gives businesses capabilities they did not have access to before. Data analytics helps organizations transform raw data into valuable insights and is, therefore, a critical asset to any organization as a baseline for any important tactical, operational, and strategic decisions. However, although data analytics provides many benefits, new security challenges have emerged that hamper the effectiveness of organizational analytics efforts. New approaches to security are required to address these challenges.

This research in progress paper provides an overview of security-related challenges surrounding data analytical solutions. In addition, the paper discusses shortcomings of current governance and security frameworks in addressing data analytics-specific security challenges and presents avenues for future research.

Keywords: data and analytics security issues, data warehouse, business intelligence, information security, data security

1 Introduction

Data-driven companies can differentiate themselves from competitors in their industries by their ability to analyze business processes using analytical systems [1]. According to Dedić and Stainer [2], data analytics is an umbrella term covering two technological solutions for getting new business insights. First, retrospective (descriptive) analytics with the primary goal that highlights current and past business processes, e.g., reporting of company revenue. In contrast, the primary goal of prospective (prescriptive) analytics is to predict *what could* and *will happen* through, for example, artificial intelligence (AI). Both systems are designed in special database ecosystems for processing data and applying algorithms to provide new business insights [3]. Subsequently, a data analytics solution is the collection of tools for data-driven evaluation [2]. There is no clear technological borderline between these two perspectives of data analytical solutions, but they are used commonly in heterogenic data landscapes driven by different system architecture [2]; for example, if a company wants to forecast customer behavior. Therefore, necessary information is distributed through several systems, which must be harmonized through a data analytics solution to provide it in a prepared business report. This approach of information collection is standardized and has been discussed in the literature as the Data Science Process Model

17th International Conference on Wirtschaftsinformatik,
February 2022, Nürnberg, Germany

(DASC-PM) [11] and Knowledge Discovery in Databases (KDD) [12]. However, one of the most popular and widely accepted models is the CRoss-Industry Standard Process of Data Mining framework (CRISP-DM) (see Figure 1) [13].

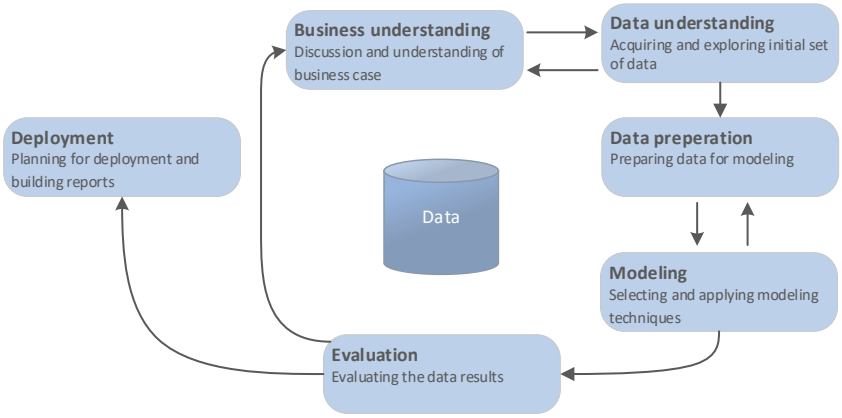

Figure 1: CRoss-Industry Standard Process of Data Mining (adapted from [13])

Although analytical systems have many benefits, there are also many risks involved in different layers of the aforementioned CRISP-DM framework (see Figure 1). For example, there are still unpatched operating server systems with vulnerabilities such as unencrypted interfaces or a lack of a well-designed authentication concept. The latter example can lead users to unintended data, e.g., aggregation mechanisms can be misused to compile company revenue, typically accessible only by management. Consequently, data distribution, which can involve sensitive data, is not sufficiently well-regulated by existing governance models [10]. In a worst-case scenario, sensitive data can be leaked internally as well as externally. Therefore, data infringement, such as unauthorized usage, can seriously impact [14], such as in a case reported in *The New York Times* [15]. Andrew Pole, a marketing analyst, concluded that pregnant customers are more open to buying new products. Therefore, he used a data analytics solution for a pregnancy prediction model based on customers' shopping carts. Based on this prediction model, promotions were linked with the probability of pregnancy. If the probability was very high, pregnant customers received promotional material for newborns. Consequently, teenage pregnancy was discovered before the young women's closest family members were informed. Thus, awareness of sensitive data is integral to managing data security issues, but existing governance and security frameworks do not consider such issues regarding the data analytics process sufficiently [16]. Furthermore, existing research papers investigate analytical solutions as a feature for making cyber security itself more intelligent [5, 6, 7, 8] or just consider it partially [9, 18, 19, 20, 21]. Thus, new security frameworks are required to address this issue holistically to data analytic solutions.

This paper aims to present an overview of important concepts that exist for holistically managing data analytical security issues and discuss avenues for future research. For our research purpose, we define security concepts as a set of practices and methodologies to keep data secure from unintended access [25, 26].

2 Related governance models

In general, governance models provide companies with a set of rules and business processes that enable strategic alignment and human capital performance. Besides governance models, we also consider cybersecurity frameworks as a common language of security postures in analytical solutions. The following sections describe related models and standardizations with references to Avery and Cheek [29] and briefly challenge their security gaps.

Data governance takes into account some of the challenges in data analytics solutions. Although there are several definitions of data governance, Wende's research [30] is the most commonly used. It is similar to several others: data governance is a "framework for decision rights and accountabilities to encourage desirable behavior in the use of data," e.g., [31] and [32]. Based on this definition, companies' data governance models are often driven by government regulations and compliance requirements such as the General Data Protection Regulation (GDPR) [17]. The GDPR was introduced by the European Union (EU) in 2018. This regulation applies to companies that process the personal data of any person in the EU. One of the important requirements of the GDPR is protecting data from unauthorized access, which transforms a data analytical solution from an asset into a liability. In contrast, information governance has limitations regarding data analytics because of the inability of the proprietary techniques to be scaled in terms of big data needs, as well as the uncontrolled usage of information in analytical solutions driven by business [33].

Another important model is information technology (IT) governance. The focus of this model is IT performance and managing risks to reach strategic objectives [39]. Compared with data and information governance, IT governance has been well researched, e.g., [34, 35] and standardized by the International Organization for Standardization (ISO) and the International Electronic Commission (IEC) in ISO 38504 [27]. In addition, control objectives for information and related technology provide a best practice framework. This framework, founded by the Information System Audit and Control Association, aims to connect a company's goals to IT-specific models, e.g., ITIL [28]. In essence, IT governance focuses on the technology itself and has limitations for data analytics solutions. For example, it does not consider employee behavior in relation to regulating sensitive data distribution.

3 Challenges and findings of data analytics security

Organizations waste capacities by implementing the full spectrum of governance models with related security frameworks to address comprehensive phases of data analytics solutions [24]. This proposal focuses on the security aspects of existing governance and cybersecurity framework issues based on a literature review on scientific search engines with findings of relevant publications.

Information security aims at protecting the confidentiality, integrity, and availability of data to minimize the business damage triggered by information security incidents [25]. As mentioned in the section before, data security is about securing the data itself.

According to R. von Solms [37], “because the security of underlying data is, to a large extent, reliant on the overall security of the information system on which the data resides, it can be argued that the term data security is in fact used in Dhillon [38] to refer to the same concept as that which ISO/IEC 13335-1 [26] calls ICT (information and communication technology) security.” Based on the aforementioned issues, we studied information and ICT security, which are scaled and illustrated in the data analytics perspectives described in Figure 2.

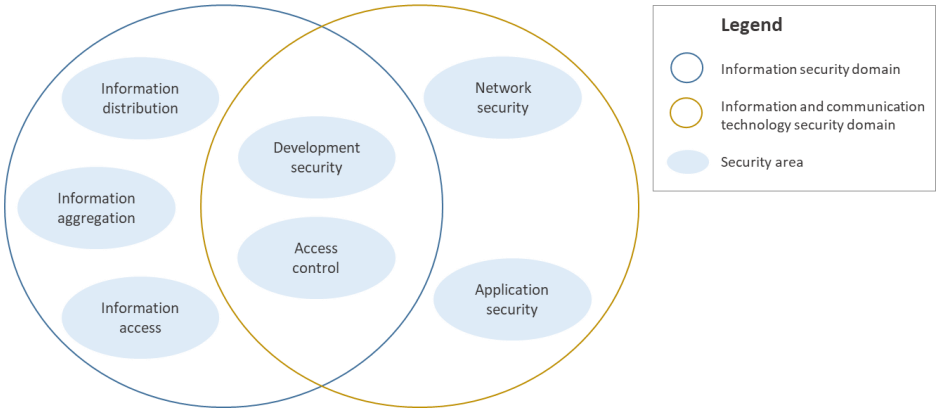

Figure 2: Data analytics security domains according to R. von Solms (adapted from [37])

Based on the foregoing illustration and related security dimension definitions [26, 37, 38], it is clear that data analytics solutions are at the center of the conflict. However, we have adapted the holistic security domain classification from von Solms [37]. On the one hand, information security is driven by the behavior of human capital, while on the other hand, ICT security impacts the information directly. Two subdomains, in particular, have a direct impact on information visibility: development security and access control. Information can be made visible to unauthorized users through incorrect development or authentication concept. The self-service approach, which increases vulnerabilities and risks, means business users can analyze and distribute data themselves. This impacts especially the last phase of CRISP-DM (Figure 1), where report distribution is not traceable anymore. In addition to that, unpatched operation systems impact almost all phases. However, data can quickly turn into sensitive information or critical data elements [18]. Therefore, information can be compiled in an uncontrolled manner, and consequently, either data infringements will occur (e.g., unauthorized distribution of sensitive information), or certain aggregation mechanisms that should only be accessible to high management level can also be compiled by people who are not allowed to do so. However, according to the ISO/IEC [26], multiple data analytics applications must be managed and secured because analytical capabilities are generally not built on a single application.

In summary, existing security concepts do not address all the comprehensive challenges of data analytics security. The increase in data diversity led by big data

requires a more comprehensive model to address and manage issues during the data lifecycle [36]. This short analysis reveals that governance models and best-practice frameworks are often driven by regulatory bodies and addressed from the perspective of liability rather than assets. Therefore, these capabilities are challenged by regulatory authorities. As a result of such directives and the increasing usage of analytical solutions, several security issues for protecting and governing significant amounts of data must be addressed. In addition, according to Dedić and Stainer [2], current research only considers one perspective of data analytics instead of taking a holistic investigation, e.g., [18, 19, 20, 22, 23].

4 Conclusion and Outlook

This research in progress provided a summary overview of existing and addressable security concepts to data analytics solutions and adjacent areas. The results contribute an overview of the state of the art in data analytics security issues in a holistic way that covers both the retro- and prospective spectrum of analytics. Specifically, technical and non-technical challenges are briefly investigated. The results were also obtained from several security domains, e.g., access control. Existing research and governance models have failed to address the factors of data analytical solutions in a holistic way for both perspectives [9, 19, 20, 26, 27]. In general, the approach like the CRoss-Industry Process of Data Mining framework [13] shows how to add value to data in a standardized way, but as mentioned by Consentino [10], there are additional capabilities required to reduce the risk of security issues, e.g., data infringement, unauthorized usage and access to sensitive data [14], that can happen through the aggregation of data which leads, for example, to companies revenue. Besides the human capital as a risk factor, also technical challenges must be managed according to ISO/IEC [26]. Moreover, regulatory authorities challenge these capabilities also through requirements for more transparency in data handling within companies, e.g., GDPR [17]. Data analytics solutions are at the center of the debate about information being an asset and a liability. Governance and compliance must move away from silo solutions and instead be addressed through a holistic analytics approach and related comprehensive phases to succeed in the digital age. In addition, business and IT functions must cooperate toward a common goal.

As an outlook and future work, we will perform a structured literature review to classify existing security concepts in data analytical solutions to provide a detailed overview in this field. In addition to that, we will further our research with expert interviews for contrasting existing theoretical frameworks with their usage in practice. We will incorporate this gained knowledge into a theoretical framework for managing data analytics security issues.

References

1. IBM Institute for Business Valuse and MIT Sloan Management Review: Analytics: The New Path to Value. Online (2010)
2. Dedić, N., Stainer, C.: Towards Differentiation Business Intelligence, Big Data, Data Analytics and Knowledge Discovery. ERP Future, P. 6 (2017)
3. Fang, H.: Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. IEEE International Conference on Cyber Technology in Automation (CYBER), pp. 820-824 (2015)
4. Dumslaff, U., Heimann, T.: Studie IT-Trends 2018. Capgemini, pp. 22 (2018)
5. Cárdenas, A. A., Manadhata, P. K., Rajan, S. P.: Big Data Analytics for Security. In IEEE Security & Privacy, vol. 11, no. 6, pp. 74-76 (2013)
6. Sarker, I. H., Kayes, A.S.M., Badsha, S.: Cybersecurity data science: an overview from machine learning perspective. J Big Data 7, 41 (2020)
7. Dev Mishra, A., Beer Singh, Y.: Big data analytics for security and privacy challenges. International Conference on Computing, Communication and Automation (ICCCA), pp. 50-53 (2016)
8. Alguliyev, R., Imamverdiyev, Y.: Big Data: Big Promises for Information Security. 2014. IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), pp. 1-4 (2014)
9. Gahi, Y., Guennoun, M., Mouftah, H. T.: Big Data Analytics: Security and privacy challenges. IEEE Symposium on Computers and Communication (ISCC), pp. 952-957 (2016)
10. Consentino, T.: Big Data Analytics Require Best Practices in Using Technology. Ventana Research, Online (2014)
11. Schulz, M., Neuhaus, U., Kaufmann, J., Badura, B., Kuehnel, S.: Introduction DASC-PM: A Data Science Process Model. ACIS (2020)
12. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P: From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37 (1996)
13. Wirth, R., Hipp, J.: CRISP-DM Towards a Standard Process Model for Data Mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining (2016)
14. Verhoef, P. C., Kooge, E., Walk, N.: Creating Value with Big Data Analytics – Making smarter marketing decisions. Routledge (1), pp. 105-116 (2016)
15. Duhigg, C.: How Companies Learns Your Secrets. The New York Times, Online (2012)
16. Kimball, R., Ross, M.: The Data Warehouse Toolkit. Wiley (3), pp. 490-496 (2013)
17. European Commission: General data protection regulation (GDPR). Online (2018)
18. Oracle: Enterprise Information Management: Best Practices in Data Governance. Whitepaper, Online (2011)
19. Nelson, B., Olovsson, T.: Security and privacy for big data: A systematic literature review. 2016 IEEE International Conference on Big Data, pp. 3693-3702 (2016)
20. Zhang, D.: Big Data Security and Privacy Protection. International Conference on Management and Computer Science (8), pp. 275-278 (2018)
21. Lin, L.: Security and Privacy Requirements Engineering Revisited in the Big Data Era. International Requirements Engineering Conference Workshops (24), pp. 55-55 (2016)
22. Yin, J., Zhao, D.: Data confidentiality challenges in big data applications. IEEE International Conference on Big Data, pp. 2886-2888 (2015)

23. Revathy, P., Mukesh, R.: Analysis of big data security practices. International Conference on Applied and Theoretical Computing and Communication Technology (3), pp. 264-267 (2017)
24. Petzold, B., Roggendorf, M., Rowshankish K., Sporleder, C: Designing data governance that delivers value. McKinsey Digital (2020)
25. ISO and IEC: ISO/IEC 27002:2013 Information technology Code of practice for information security control. International Organization for Standardization, Online (2013)
26. ISO and IEC: ISO/IEC 27005:2018 Information technology Security techniques - Information security risk management. International Organization for Standardization, Online (2018)
27. ISO and IEC: ISO/IEC 38504:2016 Governance of information technology – Guidance for principles-based standards in the governance of information technology. International Organization for Standardization, Online (2016)
28. Hochstein, A., Zarnekow, Brenner, R.: W. ITIL als Common-Practice-Referenzmodell für das IT-Service-Management — Formale Beurteilung und Implikationen für die Praxis. Wirtschaftsinformatik 46, 382–389 (2004)
29. Avery, A. A., Cheek, K.: Analytics Governance: Towards a Definition and Framework. Americas conference on Information Systems, pp. 1-8 (2015)
30. Wende, K.: A Model for Data Governance–Organizing Accountabilities for Data. Conference of Information Systems in Australia, pp. 416-425 (2007)
31. Cohen, R.: BI Strategy: What's in a Name? Data Governance Roles, Responsibilities, and Results Factors. Online (2006)
32. Newman, D., Logan, D.: Governance is an Essential Building Block for Enterprise Information System. Gartner Research, Online (2006)
33. Nasser, T., Tariq, R. S.: Big Data Challenges. Journal of Computer Engineering & Information Technology, pp. 6-7 (2015)
34. Xue, Y., Liang, H., Boulton, W.: Information Technoglogy Governance in Information Technology Investment Decision Processes: The Impact of Investment Characteristics, External Environment, and Internal Context. MIS Querterly, pp. 67-96 (2008)
35. Watson, H. J., Fuller, C., Ariyachandra, T.: Data Warehouse Governance: Best Practices at Blue Cross and Blue Shield of North Carolina. Decision Support System, pp. 433-450 (2003)
36. Kshetri, N.: Big data's impact on privacy, security and consumer welfare. Telecommunication Policy, pp. 1134-1145 (2014)
37. Solms, R., Niekerk, J.: From information security to cyber security. Computer & Security (38), pp. 97-102 (2013)
38. Dhillon, G.: Principles of information systems security. John Wiley & Sons (1) (2007)
39. Haes, S. De., Grembergen, W. V.: IT Governance and Its Mechanism. ISACA, Online (2004)