Business Intelligence meets Big Data: An Overview on Security and Privacy

Claudio A. Ardagna and Ernesto Damiani

Università degli Studi di Milano, 26013 Crema, Italy
{firstname.lastname}@unimi.it

Abstract. Today big data are the target of many research activities focusing on big data management and analysis, definition of zero latency approaches to data analytics, and protection of big data security and privacy. In particular, security and privacy are two important, while contrasting, requirements. Big data security usually refers to the use of big data to implement solutions increasing security, reliability, and safety of a distributed system. Big data privacy, instead, focuses on the protection of big data from unauthorized use and unwanted inference. In this paper, we start from the manifesto on *Business Intelligence Meets Big Data* [8] and the notions of full data and zero-latency analysis to discuss new challenges in the context of big data security and privacy.

1 Introduction

Every minute, the world generates 1.7 million billion bytes of data, that is, over 6 megabytes for each human being on the planet Earth [10]. The business potential of these data is huge; as data technology advances, entire business domains are likely to be reshaped by analytics. The European Commission's recent *Communication on the data-driven economy* [9] describes its new strategy to promote the data-driven economy in the European Union. The *Communication* is a response to the European Council's call, in October 2013, for action to provide the right conditions for a single European market for big data and cloud computing. However, European stakeholders that rely on traditional business intelligence techniques, including many SMEs, may wonder how the *Communication*'s data strategy applies to their everyday operation.

Currently, many organizations worldwide collect and analyze their own business process data in order to improve their internal decision making. Traditionally, this type of business intelligence has involved data warehousing and sampling (i.e., selecting a - hopefully representative - subset of the entire data space), and a certain latency (i.e., some delay between the collection of process data and decision making based on them). Today, the twin notions of *full data* (a.k.a. *no-sampling*) and *zero latency* are being put forward as value propositions underlying a new notion of business intelligence. In general, zero-latency analysis of full data means better decisions and more accurate predictions; however, when this type of real time business intelligence is performed on Big Data, that is, large amounts of different types of data from heterogeneous data sources, a competitive gap exists between the “haves”, who have fast access to huge amounts of data, and the “have-nots”, who have to live with high latency and small data samples.

There is therefore an increasing demand both by research and industry communities of clear statements on *i)* some practical key scientific and technical challenges that need to be addressed to enable business intelligence on big data (as envisioned in the *Communication*) and *ii)* some solutions to tackle them.

This paper starts from the challeges identified in the manifesto on *Business Intelligence Meets Big Data* [8] and aims to provide a first answer to the above problems focusing on the security and privacy challenge in big data. More in detail, we start by summarizing the big data challenges and their impact on cost-effective, no-latency big data analytics (Section 2). Then, we focus our discussion on the analysis of the security and privacy challenge (Section 3).

2 Challenges

The application of business intelligence and analytics solutions to big data is introducing a set of new challenges that need to be carefully considered for the development of accurate and sound decision making and prediction techniques. According to the manifesto in [8], whose writing is involving big data experts and researchers with different affiliations (both academic and industrial), skill, and expertise, there are 9 main challenges as follows.

Challenge 1: Data preparation, quality, and trustworthiness. An important basis for the definition of strong and accurate techniques for big data analysis and management lies in the availability of high-quality, precise, and trustworthy data. Solutions for collection and preparation of these data are paramount for increasing the added value of information extracted by big data analytics techniques.

Challenge 2: Efficient distributed storage and search. Timeliness of data collection is fundamental to provide prompt analysis of big data especially in low-latency systems. In this context, there is an increasing need of reducing all potential sources of latency, providing efficient distributed storage with faster memories, fiber-optic channels with higher bit rate, and enhanced search algorithms.

Challenge 3: Effective online data analysis – BIG-OLAP. Big data analytics requires adapting existing solutions for traditional relational databases to the big data scenario. In particular, online analysis of multidimensional data becomes a must and a potential source of information for decision making. This could require to adapt existing OLAP approaches to big data (i.e., BIG-OLAP).

Challenge 4: Effective machine learning techniques for big data mining. Machine learning techniques have been used to build systems learning from data. Often these systems have been developed to predict some known properties/patterns/behaviors learned from training data. These techniques can also be used to increase the quality of data mining analysis, which is aimed at discovering new properties/patterns/behaviors from data. Machine learning and data mining should be adapted to big data to unleash the full potential of collected information.

Challenge 5: Efficient handling of big data streams. Similarly to Challenge 4, specific scenarios (e.g., stock exchange) would require analysis of data in the form of streams. Given the huge amount of sensors deployed in the cloud and in the Internet of Things, and in turn the huge dimension of data streams, fast and optimized solutions should be developed to make inference on big data streams.

Challenge 6: Semantic lifting techniques. Semantics of collected big data is less considered, though it represents an important aspect for future development of big data applications. Future approaches to big data analysis should be able to cope with their semantics.

Challenge 7: Programming models. Different programming models supporting management, reliability, scalability of big data infrastructures are available. Some examples include MapReduce [6] for processing and generating large data sets, Apache Hadoop [3] for developing reliable, scalable, and distributed computing. Additionally, this challenge might consider different approaches for storing and managing data.

Challenge 8: Social analytics. This challenge considers analytics solutions for modeling and filtering social interaction data and information. In this challenge, the ability to distinguish those data that can be trusted and comply with users' needs and preferences is important as well as difficult to achieve. Social analytics should then address this problem providing accurate and sound approaches to social data analysis.

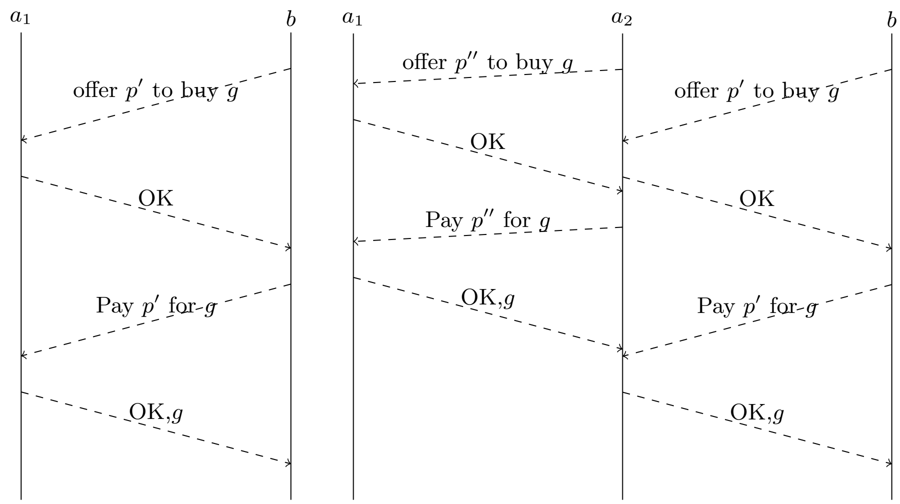
Challenge 9: Security and privacy. It considers all aspects related to two contrasting needs. On one side, big data are a priceless source of information at the basis of robust and accurate security solutions; on the other side, big data often contain sensitive information that needs to be protected from unauthorized access and release. Proper solutions should find a balance between the needs of security and privacy in a big data scenario.

3 Big Data Security and Privacy

According to the manifesto on *Business Intelligence Meets Big Data* [8], availability of trustworthy full data and zero latency are value propositions underlying a new notion of business intelligence for low-latency systems, where better decisions and more accurate predictions can be achieved. These two concepts are strictly interlaced and potentially conflicting with the concepts of security and privacy of data. This connection is clear when Big Data techniques are used for security applications like attack detection and classification. Intuitively, the more the data available for inference, the more the quality and precision of security techniques and, in turn, the security of the system implementing them.

In the extreme case of full data availability, though not yet proven in a big data environment, system security should achieve the optimum level. Full information on what is going on in the cyberspace can in fact support optimal countermeasures. However, the overhead given by data communication prior to analysis could invalidate the advantages given by the availability of full data. Also, privacy can be compromised by the availability of full data. Different solutions can be adopted to protect data privacy ranging from anonymity [4, 7, 12] to fragmentation techniques [1, 5], but their summarizing nature seems to conflict with full data assumption.

As an example, let us focus on vulnerability assessment and cyber security. The public awareness of cyberspace vulnerabilities has dramatically increased, pushing the issue of cyber security on the agenda of most European political and private stakeholders. At strategic level, cross-organizational collaboration and sharing of cyber security-relevant information are seen as imperatives for timely taking effective cyber security measures. While non real-time sharing of cyber security information is more or less state-of-the-art, an active defense against the increasing level of cyber threats requires new approaches for real time-sharing of cyber security data amongst a multitude of more or less trusted collaborating organizations. In this context, collaborative detection of attacks and anomalies aims at improving the effectiveness and efficiency of detection measures by sharing monitoring and other sensor data between the partners in a sharing network. As many events go by unnoticed, nobody acts upon them. It is usually the case that events being irrelevant for one entity might be of vital interest to others. Small events may contribute to a larger incident (or case) like pieces of a puzzle. In order to manage and respond in the best possible way, it is of critical importance to collect as many pieces as possible. Shared information can then enable the detection of events in a collaborative effort: each of the participating entities not only makes better use of existing resources, but also improves the chance to get critical and more complete information earlier on and therefore multiplies the extent of its detection ability. A solution addressing the above requirements needs fast and accurate algorithms for big data analysis and is more effective when more data are available. This solution, being able to reason on big data, could provide governments, enterprises, SMEs and individuals the benefit of a significantly increased cyber protection level by enabling truly collaborative and consistent detection and handling of cyber security incidents and threats. The other side of the coin is the increasing of privacy concerns due to the fact that several sensitive information is shared between potentially untrusted parties. The necessary exchange of data on security events raises a number of privacy and data protection issues, which need to be closely taken into account. One of the biggest challenges for implementing a collaborative detection effort like the one in our example is then to protect privacy of data by applying proper filtering, anonymization, sanitization, or pseudonymization methods, as well as to provide analysis methods performing the detection of attacks and anomalies based on the shared data.



(a) Normal operation (b) Zero-latency attack

Fig. 1. Influence of zero-latency on online auctioning

Performing analytics on full data is also strictly related to the need of zero latency solutions, where security information must be ready for online and real-time decisions. Zero latency in fact is difficult to achieve when big data are involved, while it can be mandatory in some case. Considering the previous example on cyber security and vulnerability assessment, it is clear that the more the information is fresh and up-to-date, the more the implemented defenses are effective and zero-day exploits less probable. However, zero latency can only be achieved if all components of a big data infrastructure provide extremely low overhead, and new and optimized algorithms are available for data analysis. Zero latency, while useful for detection, can also be exploited by malicious attackers [11], as for instance in the following auctioning scenario [2]. Let us consider an auctioneer a_1 selling a good g at price p, and a buyer b looking for g and willing to pay $p'>p$ for it. Now, let us consider auctioneer a_2 working at zero latency that *i)* offers p'', such that $p'>p''>p$, to a_1 for g, *ii)* buys g, and *iii)* sells g to b for p'. Thanks to zero latency, a_2 gains $p' - p''$. Figure 1 shows the normal operation of and a zero-latency attack on an online auctioning.

4 Conclusions

Security and privacy are among the most important requirements in a big data scenario. Unfortunately, these requirements are conflicting and need to be carefully balanced to find an optimum approach. In fact, while security can take advantage by an increasing amount of available data that can be analyzed in real time, privacy of users can be highly affected and suggests users to limit the free and unregulated sharing of their data. Next-generation business intelligence approaches for big data should revisit existing security and privacy approaches to find the best compromise between the need of security, the need of privacy protection, and the overhead posed by techniques for the analysis of big data.

Acknowledgments

This work was partly supported by the EU-funded project CUMULUS (contract n. FP7-318580).

References

1. Aggarwal, G., Bawa, M., Ganesan, P., Garcia-Molina, H., Kenthapadi, K., Motwani, R., Srivastava, U., Thomas, D., Xu, Y.: Two can keep a secret: A distributed architecture for secure database services. In: Proc. of the 2nd Biennial Conference on Innovative Data Systems Research (CIDR 2005). Asilomar, CA, USA (January 2005)
2. Anisetti, M., Ardagna, C., Bonatti, P., Damiani, E., Faella, M., Galdi, C., Sauro, L.: e-Auctions for multi-cloud service provisioning. In: Proc. of the 11th IEEE International Conference on Services Computing (SCC 2014). Anchorage, AL, USA (June–July 2014)
3. Apache Hadoop, `http://hadoop.apache.org/`
4. Ardagna, C., Conti, M., Leone, M., STEFA, J.: An anonymous end-to-end communication protocol for mobile cloud environments. IEEE Transactions on Services Computing (2014), (to appear)
5. Ciriani, V., De Capitani di Vimercati, S., Foresti, S., Jajodia, S., Paraboschi, S., Samarati, P.: Combining fragmentation and encryption to protect privacy in data storage. ACM Transactions on Information and System Security 13(3), 22:1–22:33 (July 2010)
6. Dean, J., Ghemawat, S.: Mapreduce: Simplified data processing on large clusters. Communications of the ACM 51(1) (January 2008)
7. Dingledine, R., Mathewson, N., Syverson, P.: Tor: The second-generation onion router. In: Proc. of the 13th USENIX Security Symposium. San Diego, CA, USA (August 2004)
8. E. Damiani *et al.*: Business intelligence meets big data: a manifesto. In: Proc. of the 3rd International Symposium on Data-Driven Process Discovery and Analysis (post proceedings). Riva del Garda, Italy (August 2013)
9. European Commission: Communication on Data-Driven Economy (July 2014), `https://ec.europa.eu/digital-agenda/en/news/communication-data-driven-economy`
10. European Commission: Helping SMEs Fish the Big Data Ocean (July 2014), `http://ec.europa.eu/digital-agenda/en/news/helping-smes-fish-big-data-ocean`
11. MacKenzie, D.: How to make money in microseconds. London Review of Books 33(10), 16–18 (2011)
12. Reiter, M., Rubin, A.: Crowds: Anonymity for web transactions. ACM Transactions on Information and System Security 1(1), 66–92 (1998)