Towards Privacy aware Big Data analytics

Pietro Colombo, Barbara Carminati, and Elena Ferrari

Department of Theoretical and Applied Sciences,
University of Insubria,
Via Mazzini 5, 21100 - Varese, Italy
{pietro.colombo, barbara.carminati, elena.ferrari}@uninsubria.it

Abstract. Big Data platforms allow the integration and analysis of high volumes of data with heterogeneous format from different sources. Big Data analytics support the derivation of properties and correlations among data and are considered by companies a key asset to make business decisions. The analyzed data often include personal and sensitive information, thus the analysis implies threats to privacy, however, to the best of our knowledge, so far no Big Data analytics platform supports the specification and enforcement of privacy policies as a native service. Although the potential benefits of data analysis are manifold, the nonexistence of proper security and privacy protection mechanisms prevents the adoption of Big Data analytics by numerous companies. The inability of using analytics represents a potential economical loss since companies cannot derive information to enhance their business and management processes. Privacy-aware Big Data analytics are therefore required to address this issue. In this position paper, we discuss high level requirements and a roadmap to the development of a framework that integrates privacy policies management into Big Data analytics platforms.

Keywords: Big Data analytics, Privacy Policy, Enforcement

1 Introduction

Big Data analytics allow the joint analysis of large volumes of a variety of structured, semi structured, and unstructured data from different sources (e.g., electronic documents, emails and digital images), and the derivation of data properties and existing correlations which are considered a key asset to make business decisions. Traditional Data Warehouses (DWs) and Database Management Systems (DBMSs) do not support such advanced analytics. Big Data platforms also outdo traditional DWs and DBMSs wrt scalability, performance, and high availability of storage and analysis services. This is carried out by replicating and distributing data and computation over clusters of nodes, using computational paradigms such as MapReduce, and simple but effective data models (e.g., key-value and document based). Companies can either handle their own private Big Data clusters within local server farms or use cloud-based services.

Recent surveys show that, although the potential benefits of Big Data analytics attract numerous companies, many of them decide not to use these services

due to the lack of standard security and privacy protection tools [5]. Therefore, we believe that in order to ensure a wider diffusion of Big Data analytics, it is first required to fill this void. To this aim, in this paper we analyze possible strategies to enhance Big Data platforms with privacy protection capabilities. More precisely, this paper aims at discussing the foundations and development strategies of a framework that supports: 1) the specification of privacy policies regulating the access to data stored into target Big Data platforms, 2) the generation of efficient enforcement monitors for these policies, and 3) the integration of the generated monitors into the target analytics platforms.

Privacy issues are far more difficult to be addressed within Big Data management systems than in the context of traditional DBMSs. Enforcement techniques proposed for traditional DBMSs appear inadequate for the Big Data context due to the strict performance requirements needed to handle large data volumes, the heterogeneity of the data, the speed at which data must be analyzed, and the distributed nature of these systems [1, 6]. Recent surveys show that analytics are moving from batch to real-time [5], imposing even stricter performance requirements. In addition, no standard language and data model have emerged for Big Data platforms. The variety of query languages and data models proposed for different platforms and data stores make the development of a general enforcement solution even more ambitious.

The rest of the paper is organized as follows. Section 2 describes the framework definition goals and development roadmap; Section 3 shortly presents related work; finally, Section 4 concludes the paper.

2 The roadmap

In this section, we discuss the proposed strategies to the development of a framework supporting the integration of privacy policy management into Big Data platforms.

Privacy policies. In order to support the specification and enforcement of privacy policies, it is first required to introduce a domain model for privacy policies. This model aims at constraining the privacy policies for big data analytics one can actually express. For instance, a privacy policy might include authorizations based on purposes, or constrain the access to data based on the satisfaction of conditions or obligations. The development of the model requires the identification of the conceptual elements that concur to the definition of policies for the target domain (e.g., the type of analysis functions, purposes and authorizations), as well as the rules that constrain the specification. The identified elements can then be composed to form the domain model of privacy policies, using standard modeling language like UML or EMF, jointly used with OCL for the specification of constraints within the domain model.

Big Data platforms. The framework shall target the main Big Data analytics platforms currently existing on the market, that is, those beloging to the family

of MapReduce systems (e.g., Hadoop), as well as NoSQL data stores (e.g., MongoDB). These platforms should be investigated, analyzing the respective data models (e.g., key valued) and query languages. The analyzed platforms should be categorized wrt the type of data they are capable of processing (structured, semi-structured, and unstructured), the used data model (key-value, document oriented, column based), the query languages used for the analysis (SQL dialects, proprietary languages, Java), the current/potential diffusion, and the typical usage scenarios. A unified query domain model specifying the types of actions performed by queries on data shall also be defined to reduce as much as possible the effort required for the deployment of the privacy-preserving framework. The model shall be capable of representing the type of operations (e.g., type of data aggregations) and access that can be performed by the queries of each platform. The query domain model can be specified with the same approach and technologies as the ones used for the domain model of privacy policies.

Specification and encoding. A relevant issue related to policy management is identifying the granularity level to be used for policy specification and enforcement. Fine-grained policies have the potentiality to define personalized protection levels. However, they complicate the specification and enforcement mechanisms requiring more computational effort and memory consumption. We believe that the identification of the granularity level to be used for policy specifications depends on the data model of the target platforms and on the action types specified within the unified query domain model. The data model of MapReduce systems and NoSQL data stores could be brought back to data records, thus reaching the level of record fields. However, the granularity could also go beyond the limits imposed by the data structures. Indeed, an unstructured data field can be structured based on its actual content according to given patterns. For instance, a raw textual field could be seen as the serialized content of a data record. Each field of such a record could be derived with regular expressions and such fields could represent the finest granularity level for policy specification.

Proper encoding strategies that minimize memory consumption shall be considered for the specification of privacy policies. The memory must be proportional to the one used for storing the data for which the policies have been specified. This memory minimization requirement shall be handled in such a way that policies expressiveness is not compromised.

Enforcement. Policy enforcement mechanisms should then be implemented by enforcement monitors regulating the execution of query and analysis functions based on their compliance with privacy policies. In order to define such monitors, it is first required to define a function that checks the compliance of the actions executed on data with the privacy policies specified for the accessed data. This function shall be specified wrt the components of privacy policies (e.g. access control rules, purposes) and queries (e.g., action types, access purposes) specified within the respective domain models. OCL can be used as specification language for the compliance function. The number of policy compliance checks performed during the query execution can be even greater than the number of the accessed

data records (the policy granularity level can go beyond the level of field), and in the Big Data scenario, data sets can include up to hundreds of millions of data records. The execution time overhead required for the evaluation of policy compliance shall be minimized by defining compliance functions optimized for performance (e.g., implementing checks with bitwise operations).

The literature on access control policies for traditional data management systems presents approaches to rewrite queries in such a way that they can only access authorized data (e.g., [2]). A similar technique shall be considered for Big Data platforms. However, intrinsic properties of Big Data clusters complicate its definition. Indeed, the rewriting techniques depend on the query languages, and each platform adopts its own proprietary language. The enforcement technique also depends on the data model, and multiple model families have been proposed so far for Big Data platforms.We believe that the same policy should be enforced with different techniques on different platforms, although some common requirements can be identified for all these techniques. First of all, they shall be defined in such a way that the complexity of the enforcement does not compromise the usability of the hosting analytics platform. In addition, the proposed techniques must be preventive, blocking queries execution in case of insufficient permissions, and filtering the accessible data of the considered data sources. However, in some cases, the rewriting may not be applicable. For instance, Hadoop Jobs conceptually represent queries but they are provided as input to Hadoop as binary JAR packages. The source code of the jobs is not available, therefore in this case it is difficult to derive the instructions to rewrite them, and other techniques should be used. For instance, the enforcement could be achieved by pre-filtering the key-value pairs to be analyzed by the jobs.

Dashboard. Usability of the proposed framework can be achieved through the design of an administration dashboard embedding functionalities supporting the generation, integration, update, and removal of enforcement monitors into the target Big Data analytics platforms. A monitor is a software module appointed to enforce the privacy policies according to a selected enforcement technique. The monitor can act as a watch dog blocking the execution of queries in case of insufficient authorizations, or rewrite the queries based on the data model and languages of the considered platform. In addition to monitor management, the dashboard shall include functionalities to specify, remove or update privacy policies. Indeed, policy administration is one of the most expensive, error prone and time consuming task even in traditional settings. Filtering criteria shall allow the selection and the concurrent editing of policies specified for multiple data records. The dashboard will include tools to evaluate the time overhead due to the enforcement, and to measure the quantity of data that are filtered or modified.

3 Related work

Privacy protection approaches can be classified into anonymization techniques, and access control techniques based on privacy policies. Most of the work tar-

get data publishing [4] and belong to the first category (see for instance [10] and [7]). The framework briefly discussed in this position paper belongs to the second category, which so far only groups few recent proposals (e.g. [3]), most of which concern formal methods and are not tailored to data management systems. Indeed, privacy-based access control is still under investigation within relational DBMSs. At present, no DBMS natively supports privacy policy enforcement. Current DBMSs base access control on the discretionary model, the mandatory model or the role based model. Only a few recent work, such as [2], propose the integration of a purpose-based access control model into relational DBMSs. Big data clusters share most of the same security vulnerabilities as web applications and traditional data warehouses [8]. Most of Big Data platforms trade security and consistency for performance, scalability and flexibility [8]. They only introduce basic support for authentication and very simple authorization control without fine-grained control [8]. Only a few proposals, such as [9], provide recommendations to improve the security features of Big Data clusters (e.g, using Kerberos, file/OS layer encryption and key/certificate management). To the best of our knowledge, so far no framework has been proposed to enhance Big Data platforms with privacy protection capabilities.

4 Conclusions

Big Data are currently considered one of the great new frontiers of IT [5]. Big Data analytics platforms allow accessing and classifying large volumes of data according to different criteria, and deriving properties and relationships among data belonging to different sources with heterogeneous nature. These analysis features have considerable effects on decision-making processes such as investments or production processes. IT managers recognize the value of analytics, and consider the use of Big Data clusters an asset for their organizations [5]. However, based on 2012 survey involving 200 IT managers of US companies [5], the 95% of the interviewed managers see the poor security and privacy enforcement practices as a top obstacle. They need data security and privacy standards which still represent an open issue for the research community [1][6], as traditional security and privacy mechanisms tailored to small scale data are inadequate for Big Data [1]. The framework discussed in this position paper aims at filling this void, enhancing the limited privacy preserving capabilities of Big Data platforms.

References

1. Cloud Security Alliance. Top Ten Big Data Security and Privacy Challenges. https://cloudsecurityalliance.org/download/top-ten-big-data-security-and-privacy-challenges/
2. P. Colombo, E. Ferrari. Enforcement of Purpose Based Access Control within Relational Database Management Systems, IEEE TKDE, to appear.
3. Datta, J. Blocki, N. Christin, H. DeYoung, D. Garg, L. Jia, D. Kaynar, and A. Sinha. Understanding and Protecting Privacy: Formal Semantics and Principled Audit Mechanisms. In ICISS 2011.

4. B. Fung, K. Wang, R. Chen, and P. Yu. Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys, 42(4), 2010.
5. Intel Co. Intel's IT manager survey on how organizations are using big data. http://www.intel.com/content/dam/www/public/us/en/documents/reports/data-insights-peer- research-report.pdf
6. C. Kuner , F. Cate , C. Millard , D. Svantesson. The challenge of big data for data protection. International Data Privacy Law. 2(2), 2012.
7. Machanavajjhala, J. Gehrke, D. Kifer, and M.Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In ICDE 2006.
8. L. Okmsn, N. Gal-Oz, Y. Gonen, E. Gudes, J. Abramov. Security issues in NoSQL Databases. In IEEE TrustCom 2011
9. Securosis, LLC. Securing Big Data: Security Recommendations for Hadoop and NoSQL environments. 2012 https://securosis.com/assets/library/reports/SecuringBigData_FINAL.pdf
10. L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.