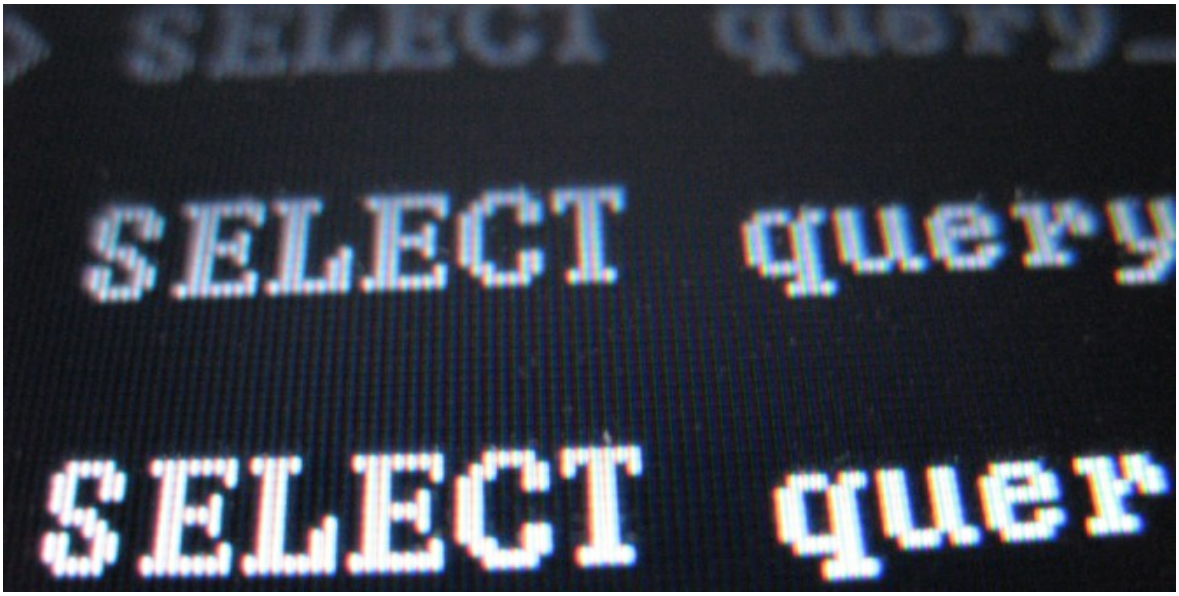
New big data platforms are more efficient, but pose a serious threat to privacy



The past few years have seen changes in the organisation of business models and work styles, caused by the rapid evolution of data analysis and data management systems. Business strategies are more and more driven by the integrated analysis of huge volumes of heterogeneous data, either directly or indirectly related to workflows. Data have become a key factor for making business decisions: we have entered the big data era.

The term big data refers to a phenomenon characterised by five “v’s”: datasets collect huge **volumes** of data with a high **variety** of formats; big data analytics platforms allow one to make predictions with high **velocity**, in a timely manner, and with high **veracity**, or low uncertainties; and with a high **value**, namely, with an expected significant gain.

The phenomenon has been pushed by numerous technological advancements. The most significant include the birth of NoSQL datastores, modern data management systems which, by means of innovative data models, provide highly efficient storage and analysis services for structured, unstructured, and semi structured data, such as transactions, electronic documents and emails. Also significant are distributed computational paradigms, like map-reduce, which have opened the way for the systematic analysis of semi-structured and unstructured data.

Overall, the support provided by big data platforms for the storage and analysis of huge and heterogeneous datasets cannot find a counterpart within traditional data management systems. In addition, the advantages of these new systems are not only related to the outstanding flexibility and efficacy of the analysis services. Big data platforms outperform traditional systems even with respect to performance and scalability.

However, the optimisation of these aspects goes to the detriment of data protection. As a matter of fact, for what data privacy and security are concerned, the majority of big data platforms integrate only quite basic enforcement mechanisms — in contrast with traditional systems, for which a variety of data protection frameworks exist.

Date originally posted: 2017-11-02

Permalink: http://blogs.lse.ac.uk/businessreview/2017/11/02/new-big-data-platforms-are-more-efficient-but-pose-a-serious-threat-to-privacy/

Blog homepage: http://blogs.lse.ac.uk/businessreview/

The unconstrained access to a high volume of data from multiple data sources, the sensitive contents of some data resources, and the advanced analysis and prediction capabilities of big data analytics platforms represent a serious threat to data privacy. For instance, the analysis capabilities can be exploited to derive correlations between sensitive and personal data. As an example, think about the retail sector, where the analysis of the purchases associated with customers' fidelity cards done for marketing purposes may allow the identification of individuals who suffer from food intolerances. As a consequence, although the potential benefits of big data analytics are indisputable, the lack of standard data protection tools open these services to potential privacy attackers.

The definition of proper data protection tools tailored for big data platforms appears as a very ambitious research challenge. State of the art enforcement techniques proposed for traditional systems cannot be used or adapted to the big data context for manifold reasons, such as the required support for semi structured and unstructured data (variety), the quantity of data to be protected (volume), and the very strict performance requirements (velocity). Therefore, the challenge is protecting privacy and confidentiality while preserving the user experience. Additional aspects contribute to raise the complexity of this goal, such as the variety of data models and data analysis and manipulation languages which are used by big-data platforms.

With the aim to do a first step to fill this void, we investigate the definition of a general framework for privacy-preserving data management within big data platforms. Several aspects are involved in the framework definition, such as the definition of a reference model for the specification and binding of privacy preferences, the design policy enforcement mechanisms, and the definition of proper tools for policy management and monitoring.

Fine grained access-control granularity has been recognised as a must for effective privacy protection. However, the related enforcement mechanisms need to be invented from scratch, as those proposed for traditional systems rely on data referring to known schema, while in this context data are heterogeneous and schema-less.

An aspect that should be considered is the support for context-based constraints, as these allow highly customised access-control forms. Access authorisations are granted when conditions are satisfied, which refer to properties of the environment within which an access request has been issued, the content of the protected resources, or the subject who has requested the access. For instance, one may authorise the access to all profile data which have not been classified as identifiable data.

For what enforcement mechanisms are concerned, we have proposed different solutions tailored for a wide class of modern data management systems. We believe that, with the aim to maximise the generality and applicability of the proposed enforcement mechanisms, these should be defined on top of analysis features shared by multiple datastores, such as a standard data manipulation language. A recent proposal of a unifying query language for traditional and modern data management systems, called SQL++, appears as a promising candidate for this purpose. Therefore we have started to investigate enforcement mechanisms for systems supporting SQL++. Finally, we are planning to define mechanisms which, on the basis of a model of the analysis features characterising a system, automatically generates customised enforcement mechanisms tailored for the considered platform.

Developing privacy-preserving big data platforms is a very challenging and fascinating task which requires addressing different possibly conflicting goals (e.g., user privacy vs efficiency and data availability), but it is the only way to make an ethical use of the enormous opportunities big data platforms offer.

Notes:

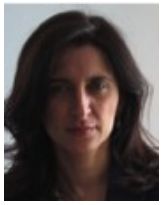
- *This blog post is based on the authors' paper Privacy Aware Access Control for Big Data: A Research Roadmap, in Big Data Research, December 2015.*
- *The post gives the views of the authors, not the position of LSE Business Review or the London School of Economics.*
- *Featured image credit: Learning SQL, by John Carter, under a CC-BY-NC-SA-2.0 licence*
- *When you leave a comment, you're agreeing to our Comment Policy.*

Date originally posted: 2017-11-02

Permalink: http://blogs.lse.ac.uk/businessreview/2017/11/02/new-big-data-platforms-are-more-efficient-but-pose-a-serious-threat-to-privacy/

Blog homepage: http://blogs.lse.ac.uk/businessreview/

Pietro Colombo is an assistant professor of Computer Science at the University of Insubria, Italy, where he works within the STRICT SociaLab of the Department of Theoretical and Applied Sciences. His most recent research activities are in the field of access control within NoSQL datastores, privacy aware data management, and data privacy within internet of things ecosystems, however he has also worked in the field of service availability and model driven engineering.



Elena Ferrari is a full professor of Computer Science at the University of Insubria, Italy and scientific director of the K&SM Research Center. Her research activities are related to access control, privacy and trust. In 2009, she received the IEEE Computer Society's Technical Achievement Award for "outstanding and innovative contributions to secure data management". She received a Google Award in 2010, and an IBM Faculty Award in 2014. She is an IEEE fellow and an ACM Distinguished Scientist.

Date originally posted: 2017-11-02

Permalink: http://blogs.lse.ac.uk/businessreview/2017/11/02/new-big-data-platforms-are-more-efficient-but-pose-a-serious-threat-to-privacy/

Blog homepage: http://blogs.lse.ac.uk/businessreview/