Accepted Manuscript

Privacy aware access control for Big Data: A research roadmap

Pietro Colombo, Elena Ferrari

PII: S2214-5796(15)00048-9
DOI: http://dx.doi.org/10.1016/j.bdr.2015.08.001
Reference: BDR 28

To appear in: *Big Data Research*

Received date: 15 April 2015
Revised date: 12 June 2015
Accepted date: 3 August 2015

Please cite this article in press as: P. Colombo, E. Ferrari, Privacy aware access control for Big Data: A research roadmap, *Big Data Research* (2015), http://dx.doi.org/10.1016/j.bdr.2015.08.001

Privacy aware access control for Big Data: a research roadmap

Pietro Colombo[a,1,*], Elena Ferrari[a]

[a]*Department of Theoretical and Applied Sciences,*
University of Insubria,
Via Mazzini 5, 21100 - Varese, Italy

Abstract

Big Data is an emerging phenomenon that is rapidly changing business models and work styles [1]. Big Data platforms allow the storage and analysis of high volumes of data with heterogeneous format from different sources. This integrated analysis allows the derivation of properties and correlations among data that can then be used for a variety of purposes, such as making predictions that can profitably affect decision processes. As a matter of fact, nowadays Big Data analytics are generally considered an asset for making business decisions. Big Data platforms have been specifically designed to support advanced form of analytics satisfying strict performance and scalability requirements. However, no proper consideration has been devoted so far to data protection. Indeed, although the analyzed data often include personal and sensitive information, with relevant threats to privacy implied by the analysis, so far Big Data platforms integrate quite basic form of access control, and no support for privacy policies. Although the potential benefits of data analysis are manifold, the lack of proper data protection mechanisms may prevent the adoption of Big Data analytics by several companies. This motivates the fundamental need to integrate privacy and security awareness into Big Data platforms. In this paper, we do a first step to achieve this ambitious goal, discussing research issues related to the definition of a framework that supports the integration of privacy aware access control features into existing Big Data platforms.

Keywords: Big Data, NoSQL datastores, MapReduce systems, Data management, Data analytics, Privacy policies, Access control enforcement, MongoDB

1. Introduction

In the recent years we have entered the Big Data era, which is rapidly and radically changing the way we live, work and think [1]. The Big Data term denotes a data management and analytics paradigm featuring 5V: huge data Volume, high Velocity (i.e., timely response requirements), high Variety of data formats, low Veracity (i.e., uncertainties in the data), and high Value [2]. Jin et al. [2] believe that Big Data are helping people to better understand the present and that this enhanced perception allows one to better predict the future. Big data play a key role for industrial upgrades. They will not uniquely sustain the growth of information industry, rather they will become a mean for improving their competitiveness. Big Data are source of business for IT giants, which are taking huge profit from the development of services and technologies (e.g., cloud-based storage and analysis services) providing the infrastructure to Big Data analytics and management. The availability of these services is continuously growing and is evolving towards real-time and ready-to-use solutions that radically change user experience wrt analysis and prediction. For instance, IBM Watson analytics[2] has been specifically designed to provide advanced, but easy to use, analytics and prediction services to managers.

Several companies which have integrated the use of Big Data analytics services into their processes, and which represent the users of the above mentioned services, are also significantly improving their business. For instance, companies that have invested on Internet of Things [3] technologies are radically changing the management of logistics and production processes getting benefit from the analysis of high volumes of sensed data and from advanced prediction features. According to Jin et al. [2], Big Data will even play a key role for national development, and according to their vision, data sovereignty of a country will be in the great power-game space together with resources such as land, sea and air.

Data managed by Big Data analytics platforms are of heterogeneous formats [4] and they can come from any source which can be sensed features of the physical world as well as information referring to social networks, internet, business, finance, economics or others. They can be structured, unstructured and semi structured data, such as transactions, electronic documents and emails. The joint analysis of data from different sources allows deriving valuable information, such as data correlations and properties that can be profitably used for making business decisions. Due to innovative computational paradigms that distribute data and analysis tasks over clusters of nodes, and simple but effective data models, Big Data analytics platforms outdo traditional Data Warehouses (DWs) and Database Management Systems (DBMSs) wrt scalability, performance, and high availability. Moreover, due to the high availability of cloud-based

*Corresponding author
Email addresses: pietro.colombo@uninsubria.it (Pietro Colombo), elena.ferrari@uninsubria.it (Elena Ferrari)
[1]Tel. +39 0332218927 Fax. +39 0332218919
[2]http://www-03.ibm.com/software/products/it/watson-content-analytics