Formal Methods for Preserving Privacy for Big Data Extraction Software

M. Brian Blake and Iman Saleh

University of Miami, Coral Gables, FL

Abstract

Given the inexpensive nature and increasing availability of information storage media, businesses, government agencies, healthcare professionals, and individuals worldwide have exponentially increased their production and persistence of large amounts of data whether such data are captured as text, images, or sound. As such, it is no surprise that just coining the term “Big Data” has generated a viral level of energy in industry and the research community. In order to facilitate the benefits of Big Data, users must capture and store their information in systematic and predictable data stores using trusted software that extract, transform and load (ETL) the data. This position paper focuses on challenges associated with the protection of sensitive information that, when mishandled or correlated with other pieces of data, can jeopardize user’s privacy. In this paper, we motivate the need for enhanced formal approaches that will enable - (1) the creation of a customized, contemporary software lifecycle and framework for big data testing, (2) development of new approaches for automated test generation including approaches that leverage the efforts of crowds/social networks, and (3) development of algorithms for continuously ensuring privacy as ETL applications and data evolve.

Introduction

Businesses and government agencies are generating and continuously collecting huge amounts of data. The current increased focus on large amounts of data will undoubtedly create opportunities and avenues for understanding the processing of such data across many varying domains. In this position paper, we use the term, Big Data, to define many terabyte and larger repositories of interconnected (sometimes with unknown relationships) data that might be realized as logs, product information, transactions, medical or biological collections. Data can be structured and unstructured and can contain complex types such as text documents, graphs, images, audio and video files.

Several examples of Big Data may include repositories containing data such as information captured from social media on the Internet, large databases of genomic information (a domain that we have leveraged via our ongoing projects), or consumer’s buying history and logs. Analysis of these huge repositories of data introduces fascinating new opportunities for discovering new patterns, relationships and insights that contribute to different branches of science. Researchers, for example, have analyzed data from over 7 million electronic medical records and used them to illustrate in a powerful visualization the (sometimes surprising) relationships between medical conditions on the basis of the frequency of co-occurrences [1]. The visualized network is called the “Health InfoScape” (Figure 1). Such visualizations raise significant HIPAA concerns [2].

Figure 1. “Health InfoScape”: a visualization of the relationships between medical conditions on the basis of the frequency of co-occurrences.
The potential of Big Data comes however with a price; the users’ privacy is often at risk. Guarantees of conformance to privacy terms and regulations are limited in current Big Data analytics and mining practices. Furthermore, current practices do not ensure that privacy terms are enforced as the data and underlying software evolve. In recent events, the government healthcare website exemplified the challenges faced by Big Data applications. The website requests enrollees to provide personally identifiable information such as date of birth, Social Security number and income information. According to the website’s privacy policy, that information is shared among many government agencies, including the IRS, Social Security Administration and Department of Homeland Security. However, the website has been criticized that there have been no guarantees that this information is used solely for the purpose of verifying the enrollee’s income, immigration status and healthcare eligibility. Users expect that the information should not be used, shared or stored beyond that purpose. This example and many others [17][18] show that there exists a distinct need to build trust into applications that extract, translate, and load (ETL) this scale of data. Furthermore, these processes must ensure the protection of sensitive information. Developers should be able to verify that their applications comply with privacy agreements and that sensitive information is kept private regardless of changes in the applications and/or privacy regulations.

Overview of Big Data Loading and Testing

To address these challenges, we have identified a need for new contributions in the areas of formal methods and testing procedures. A 2013 article on Big Data Testing [15] identifies the main testing areas, depicted in Figure 2, on a Hadoop architecture; the de-facto standard for Big Data. In our work, we have discovered the need for new paradigms for privacy conformance testing to the four areas of the ETL process as follows:

1) **Pre-Hadoop Process Validation:** This step represents the data loading process. At this step, the privacy specifications define the sensitive pieces of data that can uniquely identify a user or an entity. Privacy terms can also specify which pieces of data can be stored and for how long. At this step, schema restrictions can take place as well.

2) **Map-Reduce Process Validation:** This process manipulates big data assets to efficiently respond to a query. Privacy terms can dictate the minimum number of returned records required to conceal individual values, in addition to constraints on data sharing between different processes.

3) **ETL Process Validation:** Similar to step (2), warehousing logic need to be verified at this step for compliance with privacy terms. Some data values may be aggregated anonymously or not included in the warehouse if that implies high probability of identifying individuals.

4) **Reports Testing:** reports are another form of queries, potentially with higher visibility and wider audience. Privacy terms that define the notion of ‘purpose’ are essential to verify that sensitive data is not reported except for specified uses.

![Figure 2. Big Data Architecture and Main Testing Areas [15].](image-url)
Challenges for Formal Validation in Loading Big Data

Big Data applications have evolved through commercial use but they are still lacking the theoretical rigor to provide confidence and repeatability. Formal approaches, when applied to Big Data and its applications, will ensure the development of trusted systems in terms of compliance with privacy regulations. Through our work, we introduce techniques that leverage the efforts of social networks. Crowdsourcing techniques can help with subjectivity of privacy rules while providing a solution to dealing with the size of data when verifying privacy rules [3][20]. There are several high-level research challenges that must be addressed in order to enable our investigations and the implementation of our formalization framework, namely:

- **Formal methods and specifications must be customized to capture the privacy requirements of Big Data.** Any formal methods [4][10][12] [13] to address these problems must provide the necessary rigor to specify Big Data privacy regulations, given its unstructured and highly variable nature. The specification needs to capture the sensitivity level of different pieces of data, both individually and combined. It also needs to incorporate the notion of differential privacy where the risk of exposing personal information is probabilistically minimized. At the same time, the formalization language requires reasonable ease-of-use for application developers. Techniques and tools must be developed that reason about the specification of rules in order to automatically generate compliance tests.

- **Human input (i.e. crowdsourcing techniques) should be leveraged to efficiently and effectively guide automatic code and rules repair.** With the huge amount of data and the risk involved in automatically correcting sensitive data, we believe that leveraging crowds (perhaps of socially-linked groups [3][20]) can increase the trust in the data-driven applications. We also believe that users' input is valuable in defining and refining privacy rules considering its subjective nature.

- **Current privacy-preserving mining algorithms for Big Data must facilitate verification of privacy compliance.** Given the high volume, and the mix of structured and unstructured data, we need a set of new algorithms for managing and analyzing Big Data privately. These algorithms may build on current privacy-preserving data techniques [5][6][11] [14][19]. Leveraging the appropriate formal specification, we propose a family of algorithms that can (1) comply with the advertised privacy regulations and (2) automatically adapt to changes in the application logic and the privacy rules.

- **New software lifecycle (or customizations to existing contemporary lifecycles) must be created to incorporate the proposed approaches for developing trusted software for Big Data.** Privacy considerations must be integrated into popular software development cycle [16] to ensure wide adoption of these approaches. Privacy must be built in as a property to the applications rather than an afterthought. Consequently, there must be a redefinition of the phases of existing software development processed to integrate formalization that define privacy. Subsequently, innovations are required that extend current testing practices and the underlying processes for handling changes in privacy rules and the application logic.

**Conclusion**

As Big Data analytics continue to become prevalent in the operations of industry and government organizations, principle methods must ensure privacy at present time and as the data evolve. Aforementioned challenges require research coordination in the general areas of software engineering, data management, and machine learning.
References


