Top Ten Big Data Security and Privacy Challenges

November 2012

Contents

Acknowledgments

CSA Big Data Working Group Co-Chairs

Lead: Sreeranga Rajan, Fujitsu
Co-Chair: Wilco van Ginkel, Verizon
Co-Chair: Neel Sundaresan, eBay

Contributors

Alvaro Cardenas Mora, Fujitsu
Yu Chen, SUNY Binghamton
Adam Fuchs, Sqrrl
Adrian Lane, Securosis
Rongxing Lu, University of Waterloo
Pratyusa Manadhata, HP Labs
Jesus Molina, Fujitsu
Praveen Murthy, Fujitsu
Arnab Roy, Fujitsu
Shiju Sathyadevan, Amrita University

CSA Global Staff

Aaron Alva, Graduate Research Intern
Luciano JR Santos, Research Director
Evan Scoboria, Webmaster
Kendall Scoboria, Graphic Designer
John Yeoh, Research Analyst

1.0 Abstract

Security and privacy issues are magnified by velocity, volume, and variety of big data, such as large-scale cloud infrastructures, diversity of data sources and formats, streaming nature of data acquisition, and high volume inter-cloud migration. Therefore, traditional security mechanisms, which are tailored to securing small-scale static (as opposed to streaming) data, are inadequate. In this paper, we highlight top ten big data-specific security and privacy challenges. Our expectation from highlighting the challenges is that it will bring renewed focus on fortifying big data infrastructures.

2.0 Introduction

The term big data refers to the massive amounts of digital information companies and governments collect about us and our surroundings. Every day, we create 2.5 quintillion bytes of data—so much that 90% of the data in the world today has been created in the last two years alone. Security and privacy issues are magnified by velocity, volume, and variety of big data, such as large-scale cloud infrastructures, diversity of data sources and formats, streaming nature of data acquisition and high volume inter-cloud migration. The use of large scale cloud infrastructures, with a diversity of software platforms, spread across large networks of computers, also increases the attack surface of the entire system

Traditional security mechanisms, which are tailored to securing small-scale static (as opposed to streaming) data, are inadequate. For example, analytics for anomaly detection would generate too many outliers. Similarly, it is not clear how to retrofit provenance in existing cloud infrastructures. Streaming data demands ultra-fast response times from security and privacy solutions.

In this paper, we highlight the top ten big data specific security and privacy challenges. We interviewed Cloud Security Alliance members and surveyed security practitioner-oriented trade journals to draft an initial list of high-priority security and privacy problems, studied published research, and arrived at the following top ten challenges:

1. Secure computations in distributed programming frameworks
2. Security best practices for non-relational data stores
3. Secure data storage and transactions logs
4. End-point input validation/filtering
5. Real-time security/compliance monitoring
6. Scalable and composable privacy-preserving data mining and analytics
7. Cryptographically enforced access control and secure communication
8. Granular access control
9. Granular audits
10. Data provenance

In the rest of the paper, we provide brief descriptions and narrate use cases.

3.0 Secure Computations in Distributed Programming Frameworks

Distributed programming frameworks utilize parallelism in computation and storage to process massive amounts of data. A popular example is the MapReduce framework, which splits an input file into multiple chunks. In the first phase of MapReduce, a Mapper for each chunk reads the data, performs some computation, and outputs a list of key/value pairs. In the next phase, a Reducer combines the values belonging to each distinct key and outputs the result. There are two major attack prevention measures: securing the mappers and securing the data in the presence of an untrusted mapper.

3.1 Use Cases

Untrusted mappers could return wrong results, which will in turn generate incorrect aggregate results. With large data sets, it is next to impossible to identify, resulting in significant damage, especially for scientific and financial computations.

Retailer consumer data is often analyzed by marketing agencies for targeted advertising or customer-segmenting. These tasks involve highly parallel computations over large data sets, and are particularly suited for MapReduce frameworks such as Hadoop. However, the data mappers may contain intentional or unintentional leakages. For example, a mapper may emit a very unique value by analyzing a private record, undermining users' privacy.

4.0 Security Best Practices for Non-Relational Data Stores

Non-relational data stores popularized by NoSQL databases are still evolving with respect to security infrastructure. For instance, robust solutions to NoSQL injection are still not mature. Each NoSQL DBs were built to tackle different challenges posed by the analytics world and hence security was never part of the model at any point of its design stage. Developers using NoSQL databases usually embed security in the middleware. NoSQL databases do not provide any support for enforcing it explicitly in the database. However, clustering aspect of NoSQL databases poses additional challenges to the robustness of such security practices.

4.1 Use Cases

Companies dealing with big unstructured data sets may benefit by migrating from a traditional relational database to a NoSQL database in terms of accommodating/processing huge volume of data. In general, the security philosophy of NoSQL databases relies in external enforcing mechanisms. To reduce security incidents, the company must review security policies for the middleware adding items to its engine and at the same time toughen NoSQL database itself to match its counterpart RDBs without compromising on its operational features.

5.0 Secure Data Storage and Transactions Logs

Data and transaction logs are stored in multi-tiered storage media. Manually moving data between tiers gives the IT manager direct control over exactly what data is moved and when. However, as the size of data set has been, and continues to be, growing exponentially, scalability and availability have necessitated auto-tiering for big data storage management. Auto-tiering solutions do not keep track of where the data is stored, which poses new challenges to secure data storage. New mechanisms are imperative to thwart unauthorized access and maintain the 24/7 availability.

5.1 Use Cases

A manufacturer wants to integrate data from different divisions. Some of this data is rarely retrieved, while some divisions constantly utilize the same data pools. An auto-tier storage system will save the manufacturer money by pulling the rarely utilized data to a lower (and cheaper) tier. However, this data may consist in R&D results, not popular but containing critical information. As lower-tier often provides decreased security, the company should study carefully tiering strategies.

6.0 End-Point Input Validation/Filtering

Many big data use cases in enterprise settings require data collection from many sources, such as end-point devices. For example, a security information and event management system (SIEM) may collect event logs from millions of hardware devices and software applications in an enterprise network. A key challenge in the data collection process is input validation: how can we trust the data? How can we validate that a source of input data is not malicious and how can we filter malicious input from our collection? Input validation and filtering is a daunting challenge posed by untrusted input sources, especially with the bring your own device (BYOD) model.

6.1 Use Cases

Both data retrieved from weather sensors and feedback votes sent by an iPhone application share a similar validation problem. A motivated adversary may be able to create “rogue” virtual sensors, or spoof iPhone IDs to rig the results. This is further complicated by the amount of data collected, which may exceed millions of readings/votes. To perform these tasks effectively, algorithms need to be created to validate the input for large data sets.

7.0 Real-time Security/Compliance Monitoring

Real-time security monitoring has always been a challenge, given the number of alerts generated by (security) devices. These alerts (correlated or not) lead to many false positives, which are mostly ignored or simply “clicked away,” as humans cannot cope with the shear amount. This problem might even increase with big data, given the volume and velocity of data streams. However, big data technologies might also provide an opportunity, in the sense that these technologies do allow for fast processing and analytics of different types of data. Which in its turn can be used to provide, for instance, real-time anomaly detection based on scalable security analytics.

7.1 Use Cases

Most industries and government (agencies) will benefit from real-time security analytics, although the use cases may differ. There are use cases which are common, like, "Who is accessing which data from which resource at what time"; "Are we under attack?" or "Do we have a breach of compliance standard C because of action A?" These are not really new, but the difference is that we have more data at our disposal to make faster and better decisions (e.g., less false positives) in that regard. However, new use cases can be defined or we can redefine existing use cases in lieu of big data. For example, the health industry largely benefits from big data technologies, potentially saving billions to the tax-payer, becoming more accurate with the payment of claims and reducing the fraud related to claims. However, at the same time, the records stored may be extremely sensitive and have to be compliant with HIPAA or regional/local regulations, which call for careful protection of that same data. Detecting in real-time the anomalous retrieval of personal information, intentional or unintentional, allows the health care provider to timely repair the damage created and to prevent further misuse.

8.0 Scalable and Composable Privacy-Preserving Data Mining and Analytics

Big data can be seen as a troubling manifestation of Big Brother by potentially enabling invasions of privacy, invasive marketing, decreased civil freedoms, and increase state and corporate control.

A recent analysis of how companies are leveraging data analytics for marketing purposes identified an example of how a retailer was able to identify that a teenager was pregnant before her father knew. Similarly, anonymizing data for analytics is not enough to maintain user privacy. For example, AOL released anonymized search logs for academic purposes, but users were easily identified by their searchers. Netflix faced a similar problem when users of their anonymized data set were identified by correlating their Netflix movie scores with IMDB scores.

Therefore, it is important to establish guidelines and recommendations for preventing inadvertent privacy disclosures.

8.1 Use Cases

User data collected by companies and government agencies are constantly mined and analyzed by inside analysts and also potentially outside contractors or business partners. A malicious insider or untrusted partner can abuse these datasets and extract private information from customers.

Similarly, intelligence agencies require the collection of vast amounts of data. The data sources are numerous and may include chat-rooms, personal blogs and network routers. Most collected data is, however, innocent in nature, need not be retained, and anonymity preserved.

Robust and scalable privacy preserving mining algorithms will increase the chances of collecting relevant information to increase user safety.

9.0 Cryptographically Enforced Access Control and Secure Communication

To ensure that the most sensitive private data is end-to-end secure and only accessible to the authorized entities, data has to be encrypted based on access control policies. Specific research in this area such as attribute-based encryption (ABE) has to be made richer, more efficient, and scalable. To ensure authentication, agreement and fairness among the distributed entities, a cryptographically secure communication framework has to be implemented.

9.1 Use Cases

Sensitive data is routinely stored unencrypted in the cloud. The main problem to encrypt data, especially large data sets, is the all-or-nothing retrieval policy of encrypted data, disallowing users to easily perform fine grained actions such as sharing records or searches. ABE alleviates this problem by utilizing a public key cryptosystem where attributes related to the data encrypted serve to unlock the keys. On the other hand, we have unencrypted less sensitive data as well, such as data useful for analytics. Such data has to be communicated in a secure and agreed-upon way using a cryptographically secure communication framework.

10.0 Granular Access Control

The security property that matters from the perspective of access control is secrecy—preventing access to data by people that should not have access. The problem with course-grained access mechanisms is that data that could otherwise be shared is often swept into a more restrictive category to guarantee sound security. Granular access control gives data managers a scalpel instead of a sword to share data as much as possible without compromising secrecy.

10.1 Use Cases

Big data analysis and cloud computing are increasingly focused on handling diverse data sets, both in terms of variety of schemas and variety of security requirements. Legal and policy restrictions on data come from numerous sources. The Sarbanes-Oxley Act levees requirements to protect corporate financial information, and the Health Insurance Portability and Accountability Act includes numerous restrictions on sharing personal health records. Executive Order 13526 outlines an elaborate system of protecting national security information.

Privacy policies, sharing agreements, and corporate policy also impose requirements on data handling. Managing this plethora of restrictions has so far resulted in increased costs for developing applications and a walled garden approach in which few people can participate in the analysis. Granular access control is necessary for analytical systems to adapt to this increasingly complex security environment.

11.0 Granular Audits

With real-time security monitoring (see section 12.0), we try to be notified at the moment an attack takes place. In reality, this will not always be the case (e.g., new attacks, missed true positives). In order to get to the bottom of a missed attack, we need audit information. This is not only relevant because we want to understand what happened and what went wrong, but also because compliance, regulation and forensics reasons. In that regard, auditing is not something new, but the scope and granularity might be different. For example, we have to deal with more data objects, which probably are (but not necessarily) distributed.

11.1 Use Cases

Compliance requirements (e.g., HIPAA, PCI, Sarbanes-Oxley) require financial firms to provide granular auditing records. Additionally, the loss of records containing private information is estimated at $200/record. Legal action – depending on the geographic region – might follow in case of a data breach. Key personnel at financial institutions require access to large data sets containing PI, such as SSN. Marketing firms want access, for instance, to personal social media information to optimize their customer-centric approach regarding online ads.

12.0 Data Provenance

Provenance metadata will grow in complexity due to large provenance graphs generated from provenance-enabled programming environments in big data applications. Analysis of such large provenance graphs to detect metadata dependencies for security/confidentiality applications is computationally intensive.

12.1 Use Cases

Several key security applications require the history of a digital record – such as details about its creation. Examples include detecting insider trading for financial companies or to determine the accuracy of the data source for research investigations. These security assessments are time sensitive in nature, and require fast algorithms to handle the provenance metadata containing this information. In addition, data provenance complements audit logs for compliance requirements, such as PCI or Sarbanes-Oxley.

13.0 Conclusion

Big data is here to stay. It is practically impossible to imagine the next application without it consuming data, producing new forms of data, and containing data-driven algorithms. As compute environments become cheaper, applications environments become networked, and system and analytics environments become shared over the cloud, security, access control, compression and encryption and compliance introduce challenges that have to be addressed in a systematic way. The Cloud Security Alliance (CSA) Big Data Working Group (BDWG) recognizes these challenges and has a mission for addressing these in a standardized and systematic way.

In this paper, we have highlighted the top ten security and privacy problems that need to be addressed for making big data processing and computing infrastructure more secure. Some common elements in this list of top ten issues that are specific to big data arise from the use of multiple infrastructure tiers (both storage and computing) for processing big data, the use of new compute infrastructures such as NoSQL databases (for fast throughput necessitated by big data volumes) that have not been thoroughly vetted for security issues, the non-scalability of encryption for large data sets, non-scalability of real-time monitoring techniques that might be practical for smaller volumes of data, the heterogeneity of devices that produce the data, and confusion with the plethora of diverse legal and policy restrictions that leads to ad hoc approaches for ensuring security and privacy. Many of the items in the list of top ten challenges also serve to clarify specific aspects of the attack surface of the entire big data processing infrastructure that should be analyzed for these types of threats. We plan to use OpenMobius, an open-source, large scale, distributed data processing, analytics, and tools platform from eBay Research Labs as an experimental test bed.

Our hope is that this paper will spur action in the research and development community to collaboratively increase focus on the top ten challenges, leading to greater security and privacy in big data platforms.