Data-Centric Security Key to Cloud and Digital Business

Ulf Mattsson

CTO

Compliance Engineering

INTRODUCTION

Recent breaches demonstrate the urgent need to secure enterprise identities against cyberthreats that target today's hybrid IT environment of cloud, mobile and on-premises. The rapid rise of cloud databases, storage and applications has led to unease among adopters over the security of their data. Whether it is data stored in a public, private or hybrid cloud, or used in third party SaaS applications, companies have good reason to be concerned. The biggest challenge in this interconnected world is merging data security with data value and productivity. If we are to realize the benefits promised by these new ways of doing business, we urgently need a data-centric strategy to protect the sensitive data flowing through these digital business systems.

DATA CENTRIC SECURITY

IT security has historically concentrated on protecting systems rather than the actual data, but since this has proven inadequate, organizations should consider a data-centric security approach that protects from a variety of attack vectors with different levels of granularity to control and monitor access to data as necessary without impacting business systems.

A good first step is comprehensive data discovery, determining what types of sensitive data exist in an enterprise and where they are. It is then possible to identify how the data is used, when it needs to be protected, and which method will be best to do that. Businesses often collect everything they can as they are starting out with their big data projects, anxious not to miss any precious detail. However, they quickly learn to their detriment that this approach is difficult to manage and that it is a much more complicated process to secure big data retrospectively. Its good practice to assess data for sensitivity as it enters the enterprise and treat it accordingly; there's no need to protect public information, but private information should remain so from the moment of capture. This is particularly important with big data because the faucet can be set all the way on, or filtered at various degrees in between, with an unknown conglomeration of data from inside and outside the enterprise.

Different data protection scenarios sometimes require different forms of protection.

Monitoring offers minimal security for lower risk data, but security teams might employ masking or strong encryption technology to protect more sensitive data at rest, in use, or in transit to mitigate reputational, legal, and financial damage should a breach occur.

More security-conscious and forward-thinking companies are turning to deidentification technologies such as tokenization to secure the data itself and protect privacy by replacing it with random fake data or tokens.

Once the data itself is secure, the next step in big data security evolution is actually one of the first mentioned earlier: data monitoring, auditing, and reporting of access and process attempts by whom, to what data, where, and when—using the big data systems themselves for security analysis to know if abuses of sensitive data are occurring.

DATA SECURITY TOOLS

Access Control and Authentication

The most common implementation of authentication in Hadoop is Kerberos. In access control and authentication, sensitive data are displayed in the clear during job functions— in transit and at rest. In addition, neither access control nor authentication provides much protection from privileged users, such as developers or system administrators, who can easily bypass them to abuse the data. For these reasons, many regulations, such as the Payment Card Industry Data Security Standard (PCI DSS) and the US Health Insurance Portability and Accountability Act (HIPAA) require security beyond them to be compliant.

Coarse-grained Encryption

Starting from a base of access controls and/or authentication, adding coarse-grained volume or disk encryption is the first choice typically for actual data security in Hadoop. This method requires the least amount of difficulty in implementation while still offering regulatory compliance.

Data are secure at rest (for archive or disposal), and encryption is typically transparent to authorized users and processes. The result is still relatively high levels of access, but data in transit, in use or in analysis are always in the clear and privileged users can still access sensitive data. This method protects only from physical theft.

Fine-grained Encryption

Adding strong encryption for columns or fields provides further security protecting data at rest, in transit and from privileged users, but it requires data to be revealed in the clear (decrypted) to perform job functions, including analysis, as encrypted data are unreadable to users and processes.

Format-preserving encryption preserves the ability of users and applications to read the protected data, but is one of the slowest performing encryption processes.

Implementing either of these methods can significantly impact performance, even with the fastest encryption/ decryption processes available, such that it negates many of the advantages of the Hadoop platform. As access is paramount, these methods tip the balance too far in the direction of security to be viable.

Some vendors offer a virtual file system above the Hadoop Distributed File System (HDFS), with role-based dynamic data encryption. While this provides some data security in use, it does nothing to protect data in analysis or from privileged users, who can access the operating system (OS) and layers under the virtual layer and get at the data in the clear.

Data Masking

Masking preserves the type and length of structured data, replacing it with an inert, worthless value. Because the masked data look and act like the original, they can be read by users and processes.

Static data masking (SDM) permanently replaces sensitive values with inert data. SDM is often used to perform job functions by preserving enough of the original data or de identifying the data. It protects data at rest, in use, in transit, in analysis and from privileged users. However, should the cleartext data ever be needed again (i.e., to carry out marketing operations or in health care scenarios), they are irretrievable. Therefore, SDM is utilized in test/development environments in which data that look and act like real data are needed for testing, but sensitive data are not exposed to developers or systems administrators. It is not typically used for data access in a production Hadoop environment.

Depending on the masking algorithms used and what data are replaced, SDM data may be subject to data inference and be de-identified when combined with other data sources.

Dynamic data masking (DDM) performs masking “on the fly.” As sensitive data are requested, policy is referenced and masked data are retrieved for the data the user or process is unauthorized to see in the clear, based on the user’s/process’s role. Much like dynamic data encryption and access control, DDM provides no security to data at rest or in transit and little from privileged users. Dynamically masked values can also be problematic to work with in production analytic scenarios, depending on the algorithm/method used.

Tokenization

Tokenization also replaces cleartext with a random, inert value of the same data type and length, but the process can be reversible. This is accomplished through the use of token tables, rather than a cryptographic algorithm. In vaultless tokenization, small blocks of the original data are replaced with paired random values from the token tables overlapping between blocks. Once the entire value has been tokenized, the process is run through again to remove any pattern in the transformation.

However, because the exit value is still dependent upon the entering value, a one-to-one relationship with the original data can still be maintained and, therefore, the tokenized data can be used in analytics as a replacement for the cleartext. Additionally, parts of the cleartext data can be preserved or “bled through” to the token, which is especially useful in cases where only part of the original data is required to perform a job.

Tokenization also allows for flexibility in the levels of data security privileges, as authority can be granted on a field-by field or partial field basis. Data are secured in all states: at rest, in use, in transit and in analytics.

Modern vaultless tokenization processes virtually eliminate scalability and performance issues, reducing bottlenecks caused by latency and fear of collisions that negatively impact business processes. Tokens can be embedded with business intelligence to preserve value, enable seamless secure analytics, and ensure business processes without the need to compromise security by detokenizing the data.

BRIDGING THE GAP

In comparing the methods of fine-grained data security, it becomes apparent that tokenization offers the greatest levels of accessibility and security. The randomized token values are worthless to a potential thief, as only those with authorization to access the token table and process can ever expect to return the data to their original value. The ability to use tokenized values in analysis presents added security and efficiency, as the data remain secure and do not require additional processing to unprotect or detokenize them.

This ability to securely extract value from de-identified sensitive data is the key to bridging the gap between privacy and access. Protected data remain useable to most users and processes, and only those with privileges granted through the data security policy can access the sensitive data in the clear.

CLOUD AND CORPORATE RISK

Many businesses—especially in financial services—adhere to internal security best practices that govern if, when, and how data can leave the organization or be viewed “in the clear,” with no exceptions. When sensitive information is sent into the cloud or consumed by application users, the company must keep control of the data at all times. A Cloud Gateway protects your data before it gets to the cloud, giving you the freedom to use cloud services without the risk of exposure. You're in control of your data, no matter what happens in the cloud—without impacting business processes or sacrificing SaaS functionality. A Cloud Gateway sits between cloud applications and users, replacing sensitive data with tokens or encrypted values before it is sent to the cloud. A gateway server cluster handles the traffic to and from the cloud, while a management server allows your security team to configure policies and protection methods.

WHAT'S AHEAD?

Historically, organizations have taken a reactive approach to data security in response to government regulations and industry standards. Big data requires organizations to be more proactive and flexible to the ever-changing nature of big data technology and threat landscape.

As Hadoop transitions to a more mission-critical role within the data center, vendors, tools, and regulations will continue to evolve along with it. To cope with the uncertainties of these developments and stay secure, organizations should remain open-minded; apply patches; run updates as available; and seek flexible, scalable enterprise security solutions that encourage interoperability and avoid lock in to solve the problems they are facing today and be forward-compatible.

The fast-paced evolution of big data technology has forced organizations to adapt, be agile, and seek security solutions with the same attributes. Data-centric security is not all that common but is slowly becoming more so as big data becomes more mature. And, generally, the more mature a company is, the more mature the security.

CONCLUSION

Following these best practices would enable organizations to securely extract sensitive data value and confidently adopt big data platforms with much lower risk of data breach. In addition, protecting and respecting the privacy of customers and individuals helps to protect the organization's brand and reputation.

The increasingly complex industry and federal regulatory compliance requirements are making it necessary for organizations to understand, measure, and validate the wide range of compliance initiatives. To do so, it is essential that they develop roadmaps and strategies that aim to build a reliable security program.

It is critical to connect and have a dialog with business executives about security metrics, costs, and compliance posture. Only through mutual understanding can goals be met, budgets be determined, and important initiatives be put on the executive's agenda.

The first step is to locate sensitive data in databases, file systems, and application environments and then identify the data's specific retention requirements and apply automated processes for secure deletion of data when it's no longer needed. With cost-effective approaches possibly based on agentless technologies and cloud based solutions, these goals are attainable.