Special Issue - 2016
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
NSDMCC - 2015 Conference Proceedings

Big Data and Security

Loshima Lohi
Asst. Professor
Carmel College, Mala

Greeshma K V
Asst. Professor
Carmel College, Mala

I. INTRODUCTION

The term big data refers to the massive amount of digital information companies and governments collect about us and our surroundings. This data is not only generated by traditional information exchange and software use via desktop computers, mobile phones and so on, but also from the myriads of sensors of various types embedded in various environments, whether in city streets (cameras, microphones) or jet engines (temperature sensors), and the soon-toproliferate Internet of Things, where virtually every electrical device will connect to the Internet and produce data. 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 (as of 2011 [1]). The issues of storing, computing, security and privacy, and analytics are all magnified by the 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.

II. BIG DATA, BIG SECURITY CHALLENGES

Big data presents a tremendous opportunity for enterprises across industries. By tapping into new volumes and varieties of data, scientists, executives, product managers, marketers, and a range of others can start making more informed plans and decisions, discover new opportunities for optimization, and deliver breakthrough innovations.Without the right security and encryption solution in place, however, big data can mean big problems.
Securing big data comes with its own unique challenges beyond being a high-value target. It's not that big data security is fundamentally different from traditional data security. Big data security challenges arise because of incremental differences, not fundamental ones. The differences between big data environments and traditional data environments include:

- The data collected, aggregated, and analyzed for big data analysis
- The infrastructure used to store and house big data
- The technologies applied to analyze structured and unstructured big data

A. *The Data*

The variety, velocity and volume of big data amplifies security management challenges that are addressed in traditional security management. Big data repositories will likely include information deposited by various sources across the enterprise. This variety of data makes secure access management a challenge. Each data source will likely have its own access restrictions and security policies, making it difficult to balance appropriate security for all data sources with the need to aggregate and extract meaning from the data. For example, a big data environment may include a dataset with proprietary research information, a dataset requiring regulatory compliance, and a separate dataset with personally identifiable information (PII). A researcher might want to correlate their research with a dataset including PII data, but what restrictions should be in-place to ensure adequate security? Protecting big data requires balancing analysis like this with security requirements on a case-by-case basis.

In addition, many of the repositories collect data at high volumes and velocity from a number of different data sources, and they all might have their own data transfer workflows. These connections to multiple repositories can increase the attack surface for an adversary. A big data system receiving feeds from 20 different data sources may present an attacker with 20 viable vectors to attempt to gain access to a cluster.

B. *The Infrastructure*

Another big data challenge is the distributed nature of big data environments. Compared with a single high-end database server, distributed environments are more complicated and vulnerable to attack. When big data environments are distributed geographically, physical security controls need to be standardized across all accessible locations. When data scientists across the organization want access to information, perimeter protection becomes important and complicated to ensure access to users while protecting the system from a possible attack. With a large number of servers, there is an increased possibility that the configuration of servers may not be consistent – and that certain systems may remain vulnerable.

C. *The Technology*

An additional big data security challenge is that big data programming tools, including Hadoop and NoSQL databases, were not originally designed with security in mind. For example, Hadoop originally didn't authenticate services or users, and didn't encrypt data that's transmitted between nodes in the environment. This creates vulnerabilities for authentication and network security. NoSQL databases lack some of the security features provided by traditional databases, such as role-based access control. The advantage of NoSQL is that it allows for the flexibility to include new data types on the fly, but defining security policies for this new data is not straightforward with these technologies.

The biggest challenge for big data from a security point of view is the protection of user's privacy. Big data frequently contains huge amounts of personal identifiable information and therefore privacy of users is a huge concern.

Because of the big amount of data stored, breaches affecting big data can have more devastating consequences than the data breaches we normally see in the press. This is because a big data security breach will potentially affect a much larger number of people, with consequences not only from a reputational point of view, but with enormous legal repercussions.

III. THE MASSIVE SCOPE OF BIG DATA SECURITY

To establish comprehensive big data security, executives and administrators have to address the following areas:

A. Data sources:

To most fully exploit the advantages of big data, organizations leverage various forms of data, including both structured data in a range of heterogeneous applications and databases and unstructured data that comes in a number of file types. Organizations may leverage data from enterprise resource planning systems, customer relationship management platforms, video files, spreadsheets, social media feeds, and many other sources. Further, more data sources are added all the time. Today, you don't know where new data sources may come from tomorrow, but you can have some certainty that there will be more to contend with and more diversity to accommodate. These big data sources can include personally identifiable information, payment card data, intellectual property, health records, and much more. Consequently, the data sources being compiled need to be secured in order to address security policies and compliance mandates.

B. Big data frameworks:

Within the big data environment itself—whether it's powered by Hadoop, MongoDB, NoSQL, Teradata, or another system—massive amounts of sensitive data may be managed at any given time. Sensitive assets don't just reside on big data nodes, but they can come in the form of system logs, configuration files, error logs, and more.

C. Analytics:

The ultimate fruit of a big data initiative is the output, the analytics that help the business optimize and innovate. This information can be presented in dashboards and reports, and accessed via on-demand queries. In some businesses, big data analytics represent the most sensitive asset of all, intelligence that provides a critical competitive differentiator—and a huge competitive exposure if it falls into the wrong hands.

It is important to recognize that the attributes that make big data valuable to the business also make it valuable to others—whether they're hardened cyber criminals or a disgruntled system administrator looking to make a quick, illicit buck. Establishing effective security across the categories above—and the massive number of specific outputs, systems, and services that fall into each category—is both critical and challenging.

Further, given the massive, widely fluctuating processing demands associated with big data environments, many organizations are leveraging cloud-based services and platforms to support their big data initiatives. For those organizations running big data environments in the cloud, the task of managing security grows even more difficult. In the cloud, security teams have to contend with the threats of vendor's infrastructure administrators, potential exposure to other tenants, and a number of other additional risks.

IV. SECURING BIG DATA ENVIRONMENTS WITH VORMETRIC

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

Vormetric solutions for big data security enable organizations to maximize the benefits of big data analytics—while maximizing the security of their sensitive data and addressing the requirements of their compliance office. The Vormetric Data Security Platform offers the granular controls, robust encryption, and comprehensive coverage that organizations need to secure sensitive data across their big data environments—including big data sources, big data infrastructure, and big data analytic results. By delivering a single security solution that offers coverage of these areas, Vormetric enables security teams to leverage centralized controls that optimize efficiency and compliance adherence.

The Vormetric Data Security Platform offers capabilities for big data encryption, key management, and access control—featuring several product offerings that share a common, extensible infrastructure. Further, the solution generates security intelligence on data access by users, processes, and applications.

Protecting Big Data Sources

As outlined earlier, organizations can leverage data from a broad array of sources, both structured and unstructured, for their big data initiatives. Data from databases, data warehouses, system logs, spreadsheets, and many other diverse systems may be fed into a big data environment.

To establish data security for these diverse data sources, organizations can use the following Vormetric solutions:

A.Vormetric Transparent Encryption:

This product encrypts and controls access at the file-system level. This encryption solution is easy to deploy because it doesn’t require any changes to applications.

B. *Vormetric Application Encryption:*

With this encryption product, you can encrypt specific columns in an application before it writes the field to a database. By encrypting a specific column, you can ensure a specific sensitive field will remain unreadable, even after it is imported into, and processed within, the big data environment.

V. SECURING BIG DATA FRAMEWORKS

In big data environments, data is routinely replicated and migrated among a large number of nodes. In addition, sensitive information can be stored in system logs, configuration files, disk caches, error logs, and so on. Vormetric Transparent Encryption efficiently protects data across all these areas, delivering encryption, privileged user access control, and security intelligence. In addition, with Vormetric Protection for Teradata Database, your organization can gain the comprehensive, granular controls required to secure the most sensitive assets across your Teradata environments, while enabling you to maximize the business benefits of your big data investments.

VI. SAFEGUARDING BIG DATA ANALYTICS

Big data output comes in many forms, including on-demand dashboards, automated reports, and ad hoc queries. Very often, these outputs contain intellectual property that is very valuable to an organization—and a potential target of attack. To provide big data analytics security for these confidential assets, security teams can use the solutions Vormetric Transparent Encryption and Vormetric Application Encryption.

VII. CONCLUSION

This paper has highlighted the Big data science, infrastructure related issues that have not been thoroughly vetted for security, the non-scalability real time monitoring techniques, etc. This clarifies attack surface of Big data security. In the upcoming years, Big data will have high impact on entire security spectrum which includes, anti-malware, data loss, network monitoring, user authentication & authorization, identity management, fraud detection, governance, risk and compliance. I hope that this paper will spur interest in R & D community to collaboratively focus on

REFERENCES

[1] Big Data Security Implications – http://www.academia.edu/3331728/ Big_Data_-_Security_implications

[2] Top 10 Big Data Security and Privacy Challenge http://www.isaca.org/groups/professional-english/big-data/groupdocuments/big_data_top_ten_v1.pdf

[3] Big Data Security – Article in vormetric.com – http://www.vormetric.com/data-security-solutions/use-cases/big-data-security