Dell Big Security for Big Data

With its maturing security features, Hadoop is now a platform that is, and continues to drive to be, in step with the data protection requirements of the enterprise.

A Dell Big Data White Paper
by Armando Acosta, SME, Product Manager, Dell Big Data Hadoop Solutions

The security challenge

Recent high-profile cyber-attacks underscore the need for more aggressive security in today's data centers. This need is particularly urgent with some Hadoop environments, which arose in a time when the Hadoop platform was young and had limited security features.

Hadoop was born as a data storage and processing system. For the Internet companies that adopted Hadoop early on, the focus was more on scalability than security. For example, Hadoop originally offered no capabilities for authentication of users or services. Anyone could submit code to be executed in Hadoop.

Over time, many enterprises adopted Hadoop to manage the onslaught of big data. In one notable trend, growing numbers of enterprises use Hadoop as a platform for consolidating diverse types of data and transforming it into a form that can be consumed by data analytics tools. Today, Hadoop is emerging as a de facto standard for capturing, storing, managing and manipulating big data in enterprise environments. In an *InformationWeek* survey, 42 percent of enterprises said they were either using Hadoop currently or had plans to adopt the platform.[1]

Hadoop's move into the enterprise has triggered a heightened focus on security. While the Internet providers who adopted Hadoop early on were more interested in securing software than in securing data, enterprises have an acute need to protect the data that is going into their Hadoop environments.

Today, there are two overarching reasons why security is essential in Hadoop:

- Hadoop contains sensitive data—anymore, all data is security-relevant. Improper usage or breaches of data can cause an enormous amount of damage to a business. Given this reality, Hadoop must be governed by the same security requirements as any other data center platform.

1 InformationWeek. "2015 Analytics, Business Intelligence, and Information Management Survey." October 2014.

"Many enterprises are embracing Hadoop because of the unique business benefits it provides, but, until now, this rapidly evolving big data technology hadn't always met enterprise security needs."

—UBM Tech[1]

1 UBM Tech. "Hadoop Security: Four Must-Haves." 2015.

- Data stored in Hadoop is subject to compliance adherence. Organizations must comply with various regulations that require protection of personal information, while simultaneously adhering to other corporate security policies.

These points are underscored in a UBM Tech paper on Hadoop security. "In order to mitigate the risk of data breaches, organizations must make certain they are securing their Hadoop clusters as well as protecting the rest of their sensitive information," the firm notes.[2]

Today, thanks to a sharp focus on security from a growing ecosystem, Hadoop now has a maturing security framework that is in step with the needs of today's enterprises.

The evolution of Hadoop security

A maturing security framework

Let's look at the evolution of security in the Hadoop platform, and how it has led to today's maturing security framework.

When Hadoop was a young platform, developers did not put a high priority on data security. That's because the use of Hadoop was largely limited to small, internal audiences with workloads and data sets that were not deemed highly sensitive and subject to regulation. Instead, the early controls were designed to address user error—to protect against accidental deletion of data, for example—and not to protect against misuse. For stronger protection, Hadoop relied on the surrounding security capabilities inherent to existing data management and networking infrastructure.

Authorization capabilities were added to Hadoop early on to help organizations avoid careless operation, but even then one user could easily impersonate other users. And as Hadoop matured, the various projects within the platform, such as HDFS, MapReduce, Oozie and Pig, addressed their own particular security needs. As is often the case with distributed, open source efforts, these projects established different methods for configuring the same types of security controls.

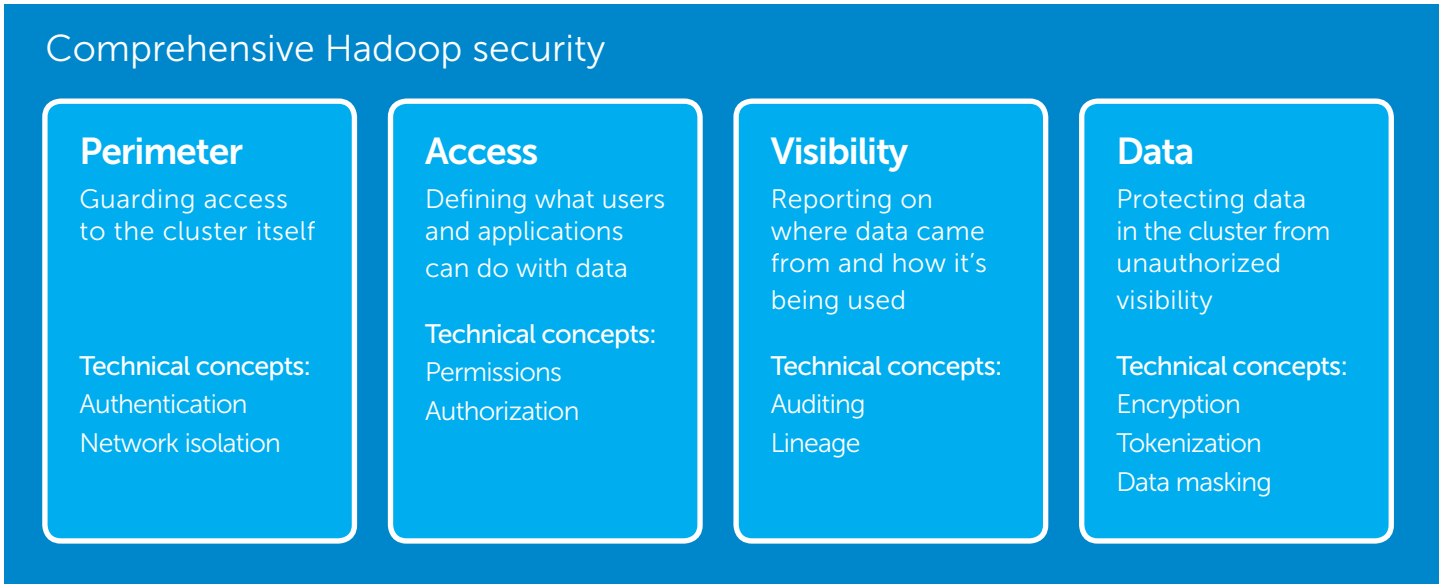
In 2009, Yahoo! added authentication to Hadoop, which was a step forward, although an imperfect one at that. The resulting security model was complex and easily misconfigured. In addition, there were no capabilities for encryption of data at rest, and only limited authorization capabilities.

In recent years, in a sign of the growing importance of Hadoop in the enterprise, technology providers and the security community have focused heavily on strengthening the platform's security features. These efforts are paying off today—and opening the door to broader usage of Hadoop.

In one notable initiative, Intel launched Project Rhino in early 2013. Project Rhino is an open source initiative dedicated to enhancing security in Hadoop. The project covers various security capabilities, including data protection and encryption, enterprise-grade authorization and single sign-on, authentication with role-based access controls, and enhanced auditing.

In 2014, Cloudera—the leading Hadoop provider—joined Project Rhino with its Apache Sentry project for unified authorization across all Hadoop frameworks. The Apache Sentry project enables administrators to manage centrally the permissions for data that can be viewed in multiple tools and computing engines. These role-based authorization capabilities make it even more viable to store sensitive data in Hadoop.

2 UBM Tech. "Hadoop Security: Four Must-Haves." 2015.



A security framework for Hadoop

Four keys to comprehensive security

Let's turn now to the requirements for a secure Hadoop environment. When it comes to data security, Hadoop is no different from other enterprise systems. We now recognize the importance of security being built into and around the core of the platform, and that it must encompass the four pillars of a comprehensive security framework: perimeter, access, visibility and data.

Each of these pillars comes with its own set of challenges and, without the proper system, can require excessive work from a security and operations point-of-view and involve manual, redundant tasks that can create error-prone security.

Let's walk through each of these security pillars, and what you should look for in a solution that bypasses the potential pitfalls on the path to a comprehensive Hadoop security framework.

Perimeter security

Perimeter security guards access to the cluster itself, drawing on technologies for authentication and network isolation. Authentication reduces the risk of unauthorized usage of services. With Hadoop, as with other enterprise systems, authentication is designed to prove that a user is who he or she claims to be.

Typically, enterprises manage identities, profiles and credentials through a single distributed system, such as a Lightweight Directory Access Protocol (LDAP) directory. LDAP authentication consists of straightforward username/password services backed by a variety of storage systems, ranging from file to database.

Another common enterprise-grade authentication system is Kerberos. Kerberos provides strong security benefits, including capabilities that render intercepted authentication packets unusable by an attacker.

Kerberos virtually eliminates the threat of impersonation present in earlier versions of Hadoop and never sends a user's credentials in the clear across the network.

Setting up and maintaining industry-standard security clusters that leverage Kerberos can be a challenging, time-consuming undertaking. To simplify the challenge, look for a solution that provides authentication for Hadoop as it is integrated with your existing systems, without limiting the flexibility of access.

Full PCI compliance

In a sign of the growing maturity of security features in the Hadoop platform, in 2014 Cloudera announced that its Cloudera Enterprise platform had become fully certified as compliant with Payment Card Industry Data Security Standards (PCI-DSS).[1] PCI-DSS is an industry-wide framework for protecting consumer credit card data. Any company that stores, processes or transmits credit card data must comply with PCI-DSS by properly securing and protecting the data.

1 "Cloudera Enterprise Data Management Platform Certified for Full PCI Compliance." Cloudera news release. October 22, 2014.

Perimeter security should:

- Preserve user choice
- Allow centrally managed authentication policies
- Enable implementation with existing standard systems

Access controls

Access controls define what users and applications can do with data—who or what has access or control over a given resource or service. These controls draw on tools that help you manage permissions and authorization.

Hadoop merges the capabilities of multiple, varied and previously separate IT systems as an enterprise data hub that stores and works on all data within an organization. Given this role, Hadoop requires multiple access controls with varying granularities. Each control is modeled after controls that should be already familiar to IT teams, controls that help administrators select the right tool for the right job.

Defining access controls for each process and path can be redundant, involve manual mirroring, and still leave the possibility for a rogue insider or a super-user breach. To address these concerns, look for a solution that opens up the unified data layer to users across the company, but with scalable access controls.

Access controls should:

- Provide users access to data needed to do their jobs
- Centrally manage access policies
- Leverage existing directory services

Visibility tools

Another building block of a Hadoop security framework revolves around visibility into the environment—specifically reporting on where data came from and how it is being used. Visibility is enabled in part by tools for auditing and tracking data lineage.

Not having a unified audit trail means that, in case of a breach, it can be extremely difficult to track point-in-time user access history. In addition, a system must have full data lineage to meet core data governance requirements. This means you need a solution that automatically stores audit logs and has robust lineage and discovery to meet your data governance requirements, without burdening your IT team.

Your visibility tools should:

- Include lineage and discovery capabilities
- Comply with policies for audit and classification
- Minimize the needs for new IT systems

Data security

This security pillar focuses on protecting the privacy of data in the Hadoop cluster, preventing any unauthorized visibility and warding off costly data breaches. Data security is enabled with tools for data encryption, tokenization and data masking.

The challenges associated with data protection include the costs of encrypting and securely storing keys. In addition, data protection tools can have limited integration with existing systems, which can put data-at-rest and data in-motion at risk. To overcome these challenges, your organization needs a solution that comprehensively protects data to meet compliance regulations without affecting analytic processes or adding latency.

Data protection tools should:

- Enable analytics on regulated data
- Encrypt data, conform to key management policies and protect from the root
- Integrate with your existing hardware security modules (HSMs)—the tools that protect and manage the digital keys used for authentication and crypto-processing.

Ready for a deeper dive?

Join subject matter experts from Dell, Cloudera and Intel for an interactive panel discussion on how to address your Hadoop security challenges. In this webinar, you will learn:

- How securing big data differs from traditional enterprise security
- Which tools and initiatives support Hadoop platform security
- Why a comprehensive solution must address the four pillars of security
- How real companies are building secure Hadoop solutions to manage and analyze big data

Webinar link: The Changing World of Big Data Security

About the author

Armando Acosta has been involved in the IT Industry over the last 15 years with experience in architecting IT solutions and product-marketing, management, planning and strategy. As a subject matter expert in big data and Hadoop, Armando's latest role is focused on addressing solutions that build new capabilities to meet emerging customer needs. He is responsible for defining the big data roadmap and working with customers to define their big data solutions. Armando graduated from the University of Texas at Austin and resides in Austin, Texas.

The Dell difference

Comprehensive security with Dell, Cloudera and Intel

Working closely with Cloudera and Intel, Dell can help your organization bring together the technologies and capacities necessary to create Hadoop environments with enterprise-class security that is integrated into your broader security environment. The Dell approach to protecting data in Hadoop covers all four of the pillars of data security.

In these efforts, Dell draws on Cloudera's comprehensive security framework for Hadoop. Built into the industry-leading core Hadoop platform, this framework provides transparent and compliance-ready security and governance across the four pillars of security.

- For perimeter security, Cloudera Manager automates Hadoop Kerberos configuration and Active Directory integration, as well as single sign-on. In addition, Dell offers an Active Directory plug-in that simplifies the process of integrating Kerberos and Active Directory environments.
- For access control, Apache Sentry provides unified authorization across all Hadoop frameworks. Through Project Rhino, Cloudera donated this software to the Hadoop community, while working closely with Intel to drive additional security enhancements in the Hadoop platform.
- For visibility, Cloudera Navigator provides an end-to-end governance solution for Hadoop for audit log aggregation, visual data lineage and data discovery.
- For data protection, Cloudera offers powerful encryption and enterprise-grade key management that is built into Hadoop through Navigator Encrypt and Navigator Key Trustee.

Drawing on technologies such as these, along with the expertise of the Dell professional services team, Dell can help you create a Hadoop environment with a level of security that is on par with the requirements of today's enterprise data centers.

To learn more, visit Dell.com/Hadoop | Dell.com/BigData

November 2015 | Version 1.0