Data Protection is a Big Deal

IT departments might not “own” data security anymore, but many are still tasked with implementing it. Organizations want to store and provide diverse data to more users, continually scaling upwards. Unfortunately, doing so has traditionally meant managing data sets scattered over disparate servers and different platforms, which can be a huge undertaking to secure and make accessible.

That’s where Apache Hadoop comes in. Hadoop enables unlimited storage, processing, and analytics of all types of data through a single platform. As enterprises continue to use Hadoop to handle more critical workloads, the need for more security and governance in the platform has also increased.

In traditional databases—with fully structured and limited data—it is comparatively easy to see what is and isn’t sensitive. These traditional systems have also had more security added to them over the years, as security became a larger concern. With unstructured data, however, sensitive information may arrive in the data hub unnoticed. Additionally, this unstructured data is arriving in larger volumes, more frequently, meaning that securing this data must happen at a much larger scale than traditional systems.

The Four Pillars of Security

Cloudera’s enterprise data hub (powered by Hadoop) has comprehensive security built into the core - enabling enterprises to store sensitive data in Hadoop while maintaining Hadoop’s flexibility for unified storage and broad user access. Cloudera’s solution addresses all the traditional security concerns around authentication, authorization, audit, and compliance— for a full compliance-ready stack across the four pillars of security (see Figure 1). Cloudera is also the first and only Hadoop distribution to achieve PCI compliance.
Four Pillars of Security: Perimeter

One benefit of Hadoop is that it provides both data flow and user access with multiple entry points. These multiple entry points mean that security must go beyond traditional firewall and single access gate though. Perimeter security for Hadoop, which addresses what services can have access to the cluster itself, must preserve the choice around having multiple services available to business users while still leveraging industry standard authentication systems. Active Directory (AD) and Kerberos are authentication staples within the enterprise and Cloudera’s platform allows enterprises to continue to leverage these existing standards, while making it easy to set up and manage at scale.

Cloudera Manager, the enterprise administration tool for Hadoop, automates much of this setup and management. With direct AD Kerberos integration, there’s no need to set up a separate Kerberos server. Additionally, users can continue to authenticate directly against AD, Hadoop Services are defined directly in AD as part of the overall service management layer, and user access to Hadoop services is controlled via AD groups. Cloudera Manager also automates the process of Kerberizing the cluster, eliminating a tedious, error-prone process and allowing for strong, scalable authentication that preserves the agility of Hadoop.

Four Pillars of Security: Access

The next step in comprehensive security is access control, which defines what users and applications can do with data. Especially with Hadoop opening up this data to more users, access policies must be scalable and centrally managed, without inhibiting the ability for users to do their job.

With many different frameworks and users accessing data, managing access controls can be quite an undertaking. That is why Cloudera created Apache Sentry, the open standard for unified authorization in Hadoop. Sentry provides fine-grained, role-based access controls (RBAC) for HDFS, Hive, Impala, and Search, with the goal of providing it for all Hadoop services and third-party applications. Sentry can be configured to use AD to determine a user’s group assignment so any changes to group assignment centrally managed, without inhibiting the ability for users to do their job.

First, for auditing and access management, Cloudera Navigator offers a full audit history for Hadoop to easily identify who has been accessing what data and review access control permissions in Sentry (to see what the permissions were at a point in time and how the permissions have been changing), all in a single place. For metadata discovery and exploration, Cloudera Navigator has complete metadata storage for technical and business metadata. Not only does it store all technical metadata associated with the data in the cluster, it also has the capability to automatically tag data based on the external sources entering the cluster (such as automatically tagging data from external ETL processes when it enters the cluster) and supports user-based tagging for business metadata. This allows for users to easily discover, classify, and locate data to not only support governance and compliance, but also user discovery within the cluster.

Lastly, Cloudera Navigator features lineage capabilities. This means, for every data source, there’s the ability to drill down into every column within that data source to see what the precise upstream data sources were and the transforms performed to produce it. Cloudera Navigator provides an automatic collection and easy visualization of upstream and downstream data lineage.

Four Pillars of Security: Data Protection

Lastly, for data protection, Cloudera protects all data inside the Hadoop cluster—at rest and in motion—with transparent encryption and enterprise-grade key management built into the platform with Cloudera Navigator Encrypt and Cloudera Navigator Key Trustee. Navigator Encrypt provides massively scalable, high-performance encryption that’s critical for sensitive Hadoop data and compliance regulations. Its high-performance encryption stems from its optimization with the Intel chipset, including the AES-NI co-processors. Navigator Key Trustee—Cloudera’s key management solution for managing encryption keys, certificates, and other sensitive Hadoop security assets—not only keeps keys separate from encrypted data but also integrates with Hardware Security Modules (HSMs) to conform with existing security policies.

Cloudera also established the Cloudera Center for Security Excellence to continually improve platform security with a security testing and certification lab, security partner enablement, and joint work with Intel for hardware and virtualization optimizations.
The Future of Data Security

Hadoop's popularity and usage throughout enterprises of all sizes and industries is not slowing down any time soon - especially with systems such as Cloudera's enterprise data hub that open up more data to more users, for faster time-to-value. As these systems power more critical workloads and store more sensitive data, security requirements will continue to increase in importance. That is why Cloudera, partnered with Intel, will continue to lead security innovation, for security that's comprehensive, transparent, and compliance-ready.

About Cloudera

Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data, an enterprise data hub built on Apache Hadoop. Cloudera offers enterprises one place to store, access, process, secure, and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. Cloudera's open source big data platform is the most widely adopted in the world, and Cloudera is the most prolific contributor to the open source Hadoop ecosystem. As the leading educator of Hadoop professionals, Cloudera has trained over 22,000 individuals worldwide. Over 1,400 partners and a seasoned professional services team help deliver greater time to value. Finally, only Cloudera provides proactive and predictive support to run an enterprise data hub with confidence. Leading organizations in every industry plus top public sector organizations globally run Cloudera in production.