cloudera®

WHITE PAPER

Data Security For Government Agencies

Version: Q115-101

Table of Contents

Abstract

Agencies are transforming data management with unified systems that combine distributed storage and computation at limitless scale, for any amount of data and any type, and with powerful, flexible analytics, from batch processing to interactive SQL to full-text search. Yet to realize their full potential, these enterprise data hub architectures require authentication, authorization, audit, and data protection controls.

Cloudera makes the enterprise data hub a reality for data-driven government by delivering applications and frameworks for deploying, managing, and integrating the necessary data security controls demanded by today's regulatory environments and compliance policies. With Cloudera, an enterprise data hub powered by Apache Hadoop includes Apache Sentry for role-based authorization, Cloudera Manager for centralized authentication management, and Cloudera Navigator for audit, lineage, and encryption capabilities.

Early Hadoop developers did not prioritize data security, as Hadoop's initial focus was on workloads and data sets not considered sensitive and subject to regulation. Security controls addressed user error, e.g. accidental deletion, and not malicious use.

Introduction

Government agencies at the forefront of information-driven decision-making recognize that the next generation of data management is a single, central system that stores all data in any form or amount and provides agency users with a wide array of tools and applications for working with this data. This evolution in data management is the enterprise data hub, and Apache Hadoop is at its core. Yet as agencies increasingly store data in this system, a growing portion is highly sensitive and subject to heavy regulations and governance controls designed to ensure data security. This information requires that Hadoop provide strong capabilities and measures to ensure security and compliance.

This paper aims to explore and summarize the common security approaches, tools, and practices found within Hadoop at large, and how Cloudera is uniquely positioned to provide, strengthen, and manage the data security required for an enterprise data hub.

Importance of Security

With national security on the line, data security is a top priority for government. Federal agencies operate within an environment awash in regulations that dictate controls and protection of data. For example, U.S. federal organizations and agencies must comply with the various information security mandates of the Federal Information Security Management Act of 2002 (FISMA), and laws like the Health Insurance Portability and Accountability Act of 1996 (HIPAA) protect and secure the public's health data.

Many agencies have mandates and objectives that establish internal information security standards to guard their data. For instance, agencies such as the Social Security Administration (SSA) focus on maintaining strict privacy of identifying citizen information, such as social security numbers, in order to maintain confidentiality and privacy for citizens. Agencies dealing with financial information may endeavor to maintain strict privacy in order to protect information vital to the health of our economy, or the personally identifying documents of individuals. Others may view data security as a wall protecting documents pertaining to national security or politically sensitive information. These internal policies can also be seen as insurance against the negative repercussions that may come as a result of a breach or accidental exposure.

Growing Pains

Historically, early Hadoop developers did not prioritize data security, as the use of Hadoop during this time was limited to small, internal audiences and designed for workloads and data sets that were not deemed overly sensitive and subject to regulation. The resulting early controls were designed to address user error - to protect against accidental deletion, 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. 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.

Employing Hadoop as the core of an enterprise data hub requires that it meets the security requirements of the government. Without such assurances, agencies cannot realize the full potential of an enterprise data hub for data management.

The data management environment and needs are vastly different today. The flexible, scalable, and cost-efficient storage capabilities of Hadoop now offer agencies the opportunity to acquire a greater range of data for analysis and the ability to retain this data for longer periods of time. This concentration of data assets makes Hadoop a natural target for those with malicious intent, but with the great advances in data security and security management in Hadoop today, agencies have robust and comprehensive capabilities to prevent intrusions and abuse.

The maturity of Hadoop today also means many data paths into and out of this information source. The breadth of ways in which an agency can work with data in Hadoop, from interactive SQL and full-text search to advanced machine-learning analysis, does offer agencies unprecedented technical capabilities and insights for their multiple business audiences and constituencies. However, the projects and products of this ecosystem are loosely coupled, and while they do often share the same data, metadata, and resources, data can and does move between them, and often each project will have separate channels for getting data into and out of the platform. Government security standards demand examination of each pathway. This poses challenges for IT teams, as some channels and methods may have the concept of fields, tables, and views, for example, while other channels might only address the same data as more course-grained files and directories.

Moreover, while all of these capabilities allow agencies the many means to explore data more fully and realize the potential value of their information, these acts of discovery might also uncover previously unknown sensitivity in the data. Agencies can no longer assume that critical and sensitive data has been identified prior to storage within Hadoop.

The powerful and far-reaching capabilities inherent in Hadoop also pose substantial challenges to federal agencies. While Hadoop stands as the core of the platform for big data, employing Hadoop as the foundation of an enterprise data hub requires that the data and processes of this platform meet government-mandated security requirements. Without such assurances, agencies simply cannot realize the full potential of and capitalize on an enterprise data hub for data management.

Security Requirements of the Government

How then can the government provide these assurances for data security? A common approach in understanding how agencies meet these challenges is to view data security as a series of mission-critical and operational capabilities, including:

- **Perimeter Security and Authentication**, which focuses on guarding access to the system, its data, and its various services;
- **Governance and Transparency**, which consists of the reporting and monitoring on the where, when, and how of data usage;
- **Entitlement and Access**, which includes the definition and enforcement of what users and applications can do with data; and
- **Data Protection**, which comprises the protection of data from unauthorized access either while at rest or in transit.

Government security requirements also demand that these controls are managed centrally and are automated across multiple systems and data sets in order to provide consistency of configuration and application according to policies, and to ease and streamline operations. This requirement is particularly critical for distributed systems where deployment and management events are common throughout the lifetime of the services, the mission, and underlying infrastructure.

Kerberos is a centralized network authentication system providing secure, single sign-on and data privacy for users and applications. Developed as a research project at the Massachusetts Institute of Technology (MIT) in the 1980s to tackle mutual authentication and encrypted communication across unsecured networks, Kerberos has become a critical part of any agency's security plan. The Kerberos protocol offers multiple encryption options that satisfy virtually all current security standards, ranging from single DES and DIGEST-MD5 to RC4-HMAC and the latest NIST encryption standard, the Advanced Encryption Standard (AES).

Lastly, government requires that systems and data integrate with existing IT tools such as directories, like Microsoft's Active Directory or the Lightweight Directory Access Protocol (LDAP), and logging and governance infrastructures, like security information and event management (SIEM) tools. Without strong integration, data security becomes fragmented. This fragmentation can lead to gaps in coverage and weak spots in the overall security umbrella, and thus creates entry points for breaches and exploitation.

Security Capabilities and Hadoop

Security in Hadoop today is very different from what was provided even two years ago. While Hadoop is still faced with the challenges outlined above, Cloudera has responded by providing security controls and management that meet the requirements of today's government. As a result of this effort, the functions of each of the separate 20+ common Hadoop projects are beginning to consolidate. For example, with Cloudera Manager, security for data-in-motion can be configured across multiple projects. With Cloudera Navigator, IT teams can now automatically gather events from the major Hadoop projects and send them to a single logging tool of their choice. And the Apache Sentry project enables administrators to manage centrally the permissions for data that can be viewed in multiple tools and computing engines.

The remainder of this paper focuses on how Hadoop now provides federal-grade security, and how IT teams can support deployments in environments that face frequent regulatory compliance audits and governance mandates. To do so, it is helpful to explore in greater detail the technical underpinnings, design patterns, and practices manifested in Hadoop for the four common factors of security: **Authentication, Authorization, Data Protection,** and **Auditing**.

Authentication

Authentication mitigates the risk of unauthorized usage of services, and the purpose of authentication in Hadoop, as in other systems, is simply to prove that a user or service is who he claims to be.

Typically, agencies 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 authentication system used by the government 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 over the network.

Both authentication schemes have been widely used for more than 20 years, and many enterprise applications and frameworks leverage them at their foundation. For example, Microsoft's Active Directory (AD) is an LDAP directory that also provides Kerberos for additional authentication security. Virtually all of the components of the Hadoop ecosystem are converging to use Kerberos authentication with the option to manage and store credentials in LDAP or AD. These additional capabilities further improve the government's security posture.

"The Security Assertion Markup Language (SAML) is an XML-based framework for communicating user authentication, entitlement, and attribute information... By defining standardized mechanisms for the communication of security and identity information between business partners, SAML makes federated identity, and the crossdomain transactions that it enables, a reality."

Organization for the Advancement of Structured Information Standards (OASIS)

Patterns and Practices

Hadoop, given its distributed nature, has many touchpoints, services, and operational processes that require robust authentication in order for a cluster to operate securely. Hadoop provides critical facilities for accomplishing this goal.

Perimeter Authentication

The first common approach concerns secured access to the Hadoop cluster itself; that is, access that is external to or on the perimeter of Hadoop. As mentioned above, there are two principle forms of external authentication commonly available within Hadoop: LDAP and Kerberos.

Best practices suggest that client-to-Hadoop services use Kerberos for authentication, and, as mentioned earlier, virtually all ecosystem clients now support Kerberos. When Hadoop security is configured to utilize Kerberos, client applications authenticate to the edge of the cluster via Kerberos RPC, and web consoles utilize HTTP SPNEGO Kerberos authentication. Despite the different methods used by clients to handle Kerberos tickets, all leverage a common core Kerberos and associated central directory of users and groups, thus maintaining continuity of identification across all services of the cluster and beyond.

To make it easier for agencies and enterprises to take advantage of Kerberos, Cloudera Manager integrates directly with Microsoft Active Directory servers. Additionally, Cloudera Manager also has a wizard that allows authorized users to configure Kerberos for individual servers (with recommended defaults) and trigger an automated workflow to secure their cluster. Finally, Cloudera Manager manages and deploys Kerberos client configurations and manages Hadoop Secured Socket Layer (SSL)-related configurations.

To help further aid agency IT teams with access integration, Cloudera has recently updated Hue and Cloudera Manager to offer additional Single Sign-On (SSO) options by employing Security Assertion Markup Language (SAML) for authentication.

Cluster and RPC Authentication

Virtually all intra-Hadoop services mutually authenticate each other using Kerberos RPC. These internal checks prevent rogue services from introducing themselves into cluster activities via impersonation and subsequently capturing potentially critical data as a member of a distributed job or service.

Distributed User Accounts

Many Hadoop projects operate at file-level granularity on data sets within HDFS, and this activity requires that a given user's account exist on all processing and storage nodes (NameNode, DataNode, JobTracker/Application Master, and TaskTracker/NodeManager) within a cluster in order to validate access rights and provide resource segmentation. The host operating system provides the mechanics to ensure user account propagation - for example, Pluggable Authentication Manager (PAM) or System Security Services Daemon (SSSD) on Linux hosts - allowing user accounts to be centrally managed via LDAP or AD, which facilitates the required process isolation. (See "Authorization: Process Isolation")

Hadoop has historically lacked centralized cross-component audit capabilities. Recent advances such as Cloudera Navigator add secured, real-time components to key data and access frameworks in Hadoop. For example, Navigator captures all activities within HDFS, Hive, HBase, Impala, and Sentry.

Authorization

Authorization is concerned with who or what has access or control over a given resource or service. Since Hadoop merges together the capabilities of multiple, varied, and previously separate IT systems as an enterprise data hub that stores and works on all data within an agency, it requires multiple authentication controls with varying granularities. Each control is modeled after controls already familiar to IT teams, allowing the easy selection of the right tool for the right job.

For example, common data flows in both Hadoop and traditional data management environments require different authorization controls at each stage of processing. First, a commercial ETL tool may gather data from a variety of sources and formats, ingesting that data into a common repository. During this stage, an ETL administrator has full access to view all elements of input data in order to verify formats and configure data parsing. Then the processed data is exposed to business intelligence (BI) tools through a SQL interface to the common repository. End users of the BI tools have varying levels of access to the data according to their assigned roles. In the first stage (data ingestion), file-level access control is in place, and in the second stage (data access through BI tools), field and row-level access control applies.

In cases where multiple authorization controls are needed, Hadoop management tools can increasingly simplify setup and maintenance by:

- Tying all users to groups, which can be specified in existing LDAP or AD directories;
- Providing a single policy control for similar interaction methods, like batch and interactive SQL and search queries. Apache Sentry permissions apply to Hive (HiveServer2), Impala, and Search, for example.

This consolidation continues to accelerate with anticipated features in Cloudera Manager that will expose permission configurations through its centralized management interface.

Patterns and Practices

Authorization for data access in Hadoop typically manifests in three forms: POSIX-style permissions on files and directories, Access Control Lists (ACL) for management of services and resources, and Role-Based Access Control (RBAC) for certain services with advanced access controls to data. These forms do share some similarities.

Process Isolation

Hadoop, as a shared services environment, relies on the isolation of computing processes from one another within the cluster, thus providing strict assurance that each process and user has explicit authorization to a given set of resources. This is particularly important within MapReduce, as the Tasks of a given Job can execute Unix processes (i.e. MR Streaming), individual Java VMs, and even arbitrary code on a host server.

In frameworks like MapReduce, the Task code is executed on the host server using the Job owner's UID, thus providing reliable process isolation and resource segmentation at the OS level of the Hadoop cluster.

POSIX Permissions

The majority of services within the Hadoop ecosystem, from client applications like the CLI shell to tools written to use the Hadoop API, directly access data stored within HDFS. HDFS uses POSIX-style permissions for directories and files; each directory and file is assigned a single owner and group. Each assignment has a basic set of permissions available; file permissions are simply read, write, and execute, and directories have an additional permission to determine access to child directories.

Role-Based Access Control (RBAC)

For finer-grained access to data accessible via schema - that is, data structures described by the Apache Hive Metastore and utilized by computing engines like Hive and Impala, as well as collections and indices within Cloudera Search - Cloudera developed Apache Sentry, which offers a highly modular, role-based privilege model for this data and its given schema. (Cloudera donated Apache Sentry to the Apache Foundation in 2013.)

Sentry governs access to each schema object in the Metastore via a set of privileges like SELECT and INSERT. The schema objects are common entities in data management, such as SERVER, DATABASE, TABLE, COLUMN, and URI, i.e. file location within HDFS. Cloudera Search has its own set of privileges, e.g. QUERY, and objects (down to the document-level), e.g. COLLECTION.

As with other RBAC systems that IT teams are already familiar with, Sentry provides for:

- Hierarchies of objects, with permissions automatically inherited by objects that exist within a larger umbrella object;
- Rules containing a set of multiple object/permission pairs;
- GRANT/REVOKE permissions via SQL commands;
- Groups that can be granted one or more roles;
- Users can be assigned to one or more groups.

Sentry is normally configured to deny access to services and data by default so that users have limited rights until they are assigned to a group that has explicit access roles.

Column-level Security, Row-level Security and Masked Access

Using the combination of Sentry-based permissions, SQL views, and User Defined Functions (UDFs), developers can gain a high degree of access control granularity for SQL computing engines through HiveServer2 and Impala, including:

- **Column-level security** - To limit access to only particular columns of entire tables, uses can access the data through a view, which contains either a subset of columns in the table, or have certain columns masked. For example, a view can filter a column to only the last four digits of a US Social Security number.
- **Row-level security** - To limit access by particular values, views can employ CASE statements to control rows to which a group of users has access. For example, a broker at a financial services firm may only be able to see data within her managed accounts.

Access Control Lists

Hadoop also maintains general access controls for the services themselves in addition to the data within each service and in HDFS. Service access control lists (ACL) range from NameNode access to client-to-DataNode communication. In the context of MapReduce and YARN, user and group identifiers form the basis for determining permission for job submission or modification. Cloudera has also added extended ACLs in HDFS to allow administrators to define multiple user groups to files (or sets of files) and to limit access on a per-group basis

Apache HBase also uses ACLs for data-level authorization. HBase ACLs authorize various operations (READ, WRITE, CREATE, ADMIN) by column, column family, column family qualifier, and down to cell-level. HBase ACLs are granted and revoked to both users and groups.

Data Protection

The goal of data protection is to ensure that only authorized users can view, use, and contribute to a data set. These security controls not only add another layer of protection against potential threats by end-users, but also thwart attacks from administrators and malicious actors on the network or in the data center. This is especially critical for the government, where insider threats pose arguably the greatest risk to data security. The means to achieving this goal consists of two elements: protecting data when it is persisted to disk or other storage mediums, commonly called "data-at-rest," and protecting data while it moves from one process or system to another, or "data in transit."

Several common compliance regulations call for data protection in addition to the other security controls discussed in this paper. Cloudera provides compliance-ready encryption and key management solutions for out-of-the-box security for both data-in-transit and data-at-rest through Cloudera Navigator Encrypt and Key Trustee.

Patterns and Practices

Hadoop, as a central data hub for many kinds of data within an organization, naturally has different needs for data protection depending on the audience and processes while using a given data set. Hadoop provides a number of facilities for accommodating these objectives.

Data Protection Technologies

There are several core data protection methods offering various degrees of coverage and security that can be applied within an agency's governance and compliance policies.

- **Encryption** - A mathematical transformation that can only be reversed through the use of a key. Strong encryption leverages NIST recognized methods such as AES 256-bit. Encryption may either use format-preserving (e.g. a 9-digit social security number encrypts to another 9-digit number) modes or other modes that expand the length of the value.
- **Tokenization** - A reversible transformation usually based on a randomly generated value that substitutes for the original value. Through the use of a lookup table, a given value, such as a credit card number, will always tokenize to the same random value. Tokens are usually format-preserving, so a 16-digit credit card number remains 16 digits. Tokenization has specific advantages for credit card data, for example, which is subject to PCI DSS compliance: systems that contain only tokenized credit card numbers may be removed from an audit scope, thereby simplifying annual audits.
- **Data Masking** - An irreversible transformation of data using a variety of methods, such as replacing a value with another value from a list of "realistic" values for a given data type (e.g. replacing the name "Steve" with "Mike"), or simply replacing certain digits with zeros or the letter "x" (e.g. replacing the first five digits of a US social security number with "x", leaving the last four digits unchanged).

Key Management

Encryption requires encryption keys. All of the above data protection techniques involve encryption – even tokenization and masking use internal tables that are protected using encryption. Key generation, storage, and access all need to be carefully managed to avoid a security breach or data loss. For example, keys must never be stored with the data, so that encrypted data remains useless if stolen. Cloudera Navigator Key Trustee provides enterprise-grade key management for securing encryption keys and other Hadoop security artifacts.

Data Granularity

Data protection can be applied at a number of levels within the Hadoop data infrastructure.

- **Field-level** - Encryption, tokenization, and data masking can all be applied to specific sensitive data fields - such as credit card numbers, social security numbers or names. When protection is applied at this level, it is generally applied only to specific sensitive fields, not to all data.
- **Filesystem-level** - Encryption can be applied at the operating system file system level to cover some or all files in a volume.
- **Network-level** - Encryption can be applied at this level to encrypt data just before it gets sent across a network and to decrypt it as soon as it is received. In Hadoop this means coverage for data sent from client user interfaces as well as service-to-service communication like remote procedure calls (RPCs).

Granularity	Advantages	Disadvantages
Field	• Protection travels with the data; it can protect data in motion and at rest. • Minimizes CPU overhead since protection is applied only to a fraction of data. • Protection can serve some of the same purposes as data masking– revealing certain digits (e.g. last 4 digits of SSN)– while concealing others when used as a reversible format-preserving method.	• Must be able to identify where the sensitive data exists in order to protect it. • Must find integration points in the data flow to intercept the sensitive data fields to protect and expose them. Due to the large number of data flow options inside or outside Hadoop, this can be a complex operation.
Filesystem	• All (or large chunks) of a data set are encrypted, so there are no need to identify specific fields that need protection. • Encryption is applied only at the "center" of a Hadoop ecosystem–HDFS – so there is only one logical place where data needs to be intercepted to encrypt and decrypt.	• Necessitates CPU overhead on all nodes, since all data is being encrypted and decrypted constantly. This may or may not be an issue, depending on workloads on the cluster. • Strictly protection for data-at-rest; guards against the recovery of data if a disk or host is removed from the data center for service. Does not protect data if copied from HDFS to a local machine or if exposed when used in report generation.
Network	• All data across the network is protected. No need to identify particular fields for protection. • Available now on virtually all transmissions within the Hadoop ecosystem. • Uses industry-standard protocols like SSL/TLS and SASL.	• Coverage limited to the network; must rely on other controls to protect data before and after transmission. • CPU overhead required to handle encryption and decrypting data transmissions.

Choosing Among the Protection Options

In designing a data protection strategy for Hadoop, most agencies strive to cover both data-at-rest and data in motion to address a large number of threats and regulatory compliance requirements. As the options above illustrate, there are multiple ways to address this challenge.

Below are examples of approaches that agencies may choose to secure their data, at rest and in transit. Note that while these approaches are common, these are not the only ways to reach a given objective, such as regulatory compliance.

Healthcare data governed by HIPAA - Often data in motion protection at the network level is employed for all Hadoop ecosystem tools used, and tools like Cloudera Manager largely automate the application of these protections. For data-at-rest protection, Cloudera is the only Hadoop vendor to offer built-in data encryption and key management. Navigator encrypt provides a transparent layer between the application and the file system, which dramatically reduces the performance impact of encryption

Insurance data used for modeling - Sensitive fields are irreversibly masked upon acquisition into HDFS and typically remain masked. Data is processed to be as “realistic” as possible in order to still be used to build models with no application changes. However, with this approach, use cases requiring original data are not an option.

Credit card data governed by PCI DSS - Credit card numbers are tokenized upstream before they are stored in HDFS. Most analysis can be done with tokenized data, but occasionally an application needs original credit card numbers and can call an API to de-tokenize. This approach simplifies PCI DSS audits by keeping Hadoop “out of scope”, but application changes are required to tokenize and de-tokenize data.

An alternative approach for credit card data involves encryption for data-at-rest (as with the healthcare example above), coupled with masking of credit card data so that the majority of users can only see the first 6 and last 4 digits, while the middle digits are replaced with X or 0. This approach enables PCI DSS compliance but without the audit scope reduction offered by tokenization.

Audit

The goal of auditing is to capture a complete and immutable record of all activity within a system. Auditing plays a central role in three key activities within the government.

First, auditing is part of a system’s security regime and can explain what happened, when, and to whom or what in case of a breach or other malicious intent. For example, if a rogue administrator deletes a user’s data set, auditing provides the details of this action, and the correct data may be retrieved from backup.

The second activity is compliance, and auditing participates in satisfying the core requirements of regulations associated with sensitive or personally identifiable data (PII), such as the Health Insurance Portability and Accountability Act (HIPAA) or the Federal Information Security Management Act of 2002 (FISMA). Auditing provides the touchpoints necessary to construct the trail of who, how, when, and how often data is produced, viewed, and manipulated.

Lastly, auditing provides the historical data and context for data forensics. Audit information leads to the understanding of how various populations use different data sets and can help establish the access patterns of these data sets. This examination, such as trend analysis, is broader in scope than compliance and can assist content and system owners in their data optimization and ROI efforts. This auditing context also provides chain-of-custody for investigative operations.

The risks facing auditing are the reliable, timely, and tamper-proof capture of all activity, including administrative actions. Until recently, the native Hadoop ecosystem has relied primarily on using log files. Log files are unacceptable for most audit use cases in the government as real-time monitoring is impossible, and log mechanics can be unreliable – a system crash before or during a write commit can compromise integrity and lose data. Other alternatives have included application-specific databases and other single-purpose data stores, and yet these approaches fail to capture all activity across the entire cluster.

In addition, any government audit solution must allow for simple central enforcement of logging policies, such as retention periods, high availability, tamper resistance, and common logging formats.

Patterns and Practices

The Hadoop ecosystem has multiple applications, services, and processes that constitute the corpus of activity, yet the following standard government auditing approaches and architectures must be applied across this ecosystem.

Event Capture

The lynchpin to these approaches is the efficient, reliable, and secure capture of event activity as it occurs. Government auditing should range from the creation, view, modification, and deletion of data to permission changes and executed queries from a number of systems and frameworks. As mentioned, Hadoop has historically lacked centralized cross-component audit capabilities, which greatly curtailed its use within a broader enterprise or agency context, especially with regulated and sensitive data and workloads.

However, advances such as Cloudera Navigator add secured, real-time components to key data and access frameworks in Hadoop. For example, Navigator captures all activity within HDFS, Hive, HBase, and now Impala and Sentry to enable an event stream that the Navigator Server captures, commits, and serves to both internal and external reporting systems. External systems could include SIEM products like RSA EnVision.

Prior to Cloudera Navigator, to gather a complete set of events, administrators needed to write scripts to gather logs from individual Hadoop project components, and then find ways to search across them and correlate events – a cumbersome and error-prone proposition given the variety of formats of the different log files and the distributed architecture of the platform.

Reporting

Activity reviews by auditors and systems occur with various scopes, filters, and timeframes, and arguably, the most ubiquitous is regular and ad-hoc report generation for formal audit procedures. Report generation comes in two forms. The first is characterized as an “assisted” report, where the auditor works with the system administrator to capture the data, screenshots, and other details required for quarterly reports, for example. The other approach provides separate access and user interfaces to the audit data. The auditor builds their own reports with no or limited assistance from the system administrator. Cloudera Navigator is the first native Hadoop application for audit reporting and provides a real-time dashboard and query interface for report generation and data gathering as well as export capabilities to external systems.

Monitoring and Alerting

Auditing also includes real-time monitoring and alerting of events as they occur. Typical SIEM tools have rule engines to alert appropriate parties to specific events.

Integration

The monitoring and reporting of Hadoop systems, while critical elements to its governmental usage, are only a part of a government's total audit infrastructure and data policy. Often these tools and policies require that all audit information route through a central interface to aid comprehensive reporting, and Hadoop-specific audit data can be integrated with these existing enterprise SIEM applications and other tools.

For example, Cloudera Navigator exposes Hadoop audit data through several delivery methods: via syslog, thus acting as a mediator between the raw event streams in Hadoop and the SIEM tools, via a REST API for custom tools, or simply exported to a file, such as a comma-delimited text file.

In addition, certain Cloudera partner vendors are actively working to add their event data into the Cloudera Navigator event stream so that applications and services closely tied to Hadoop can share events in a common framework.

Conclusion

Next-generation data management - the enterprise data hub - requires authentication, authorization, audit, and data protection controls in order to establish a place to store and operate on all data within the government, from batch processing to interactive SQL to advanced analytics. Hadoop is at its foundation, yet the core elements of Hadoop are only part of a complete solution for the government – especially given the rigorous security requirements necessary to keep sensitive governmental information safe. As more data and more variety of data is moved to the EDH, it is increasingly likely that this data will be subject to compliance and security mandates. Comprehensive and integrated security within Hadoop, then, is a keystone to establishing the new center of data management.

Cloudera is making it easier for agencies to deploy, manage, and integrate the necessary data security controls demanded by today's regulatory environments and necessary for realizing the enterprise data hub. With Cloudera, security in Hadoop has matured rapidly, with the development of services such as Apache Sentry for role-based authorization, Cloudera Manager for centralized authentication management, and Cloudera Navigator for audit and data encryption capabilities. It is clear that with Cloudera, agencies can consolidate the security functions of the Hadoop ecosystem, from authentication and authorization to data protection, encryption, key management, and auditing.

About Cloudera

Cloudera is revolutionizing data management with the first unified platform for big data, an enterprise data hub built on Apache Hadoop. Agencies today must be information-driven while managing risk and costs. Cloudera offers agencies a secure and cost-efficient place to store and analyze all their data, empowering them to derive new insights and correlation while extending the value of existing investments.

With Cloudera at the center of an agency's enterprise data hub (EDH), analysis and business users gain unprecedented visibility to contemporary and archival data and releases data sequestered in stand alone applications. An EDH offers a wide range of computing capabilities, including interactive SQL, search, and machine learning, and offers mission operators the right mix of flexible analysis and data depth and breadth to find the answers without sacrificing oversight and security to protect and govern the data and its use.

Cloudera helps agencies make the most of their data, their infrastructure, and their most valuable resource -- their people. Cloudera was the first and still is the leading provider and supporter of Hadoop for the public sector and offers software for mission critical data challenges including storage, cloud, security, management, and analysis. Cloudera works with over 1,450 hardware, software, and services partners to meet agencies' data goals.

cloudera.com

1-888-789-1488 or 1-650-362-0488
Cloudera, Inc. 1001 Page Mill Road, Palo Alto, CA 94304, USA