Hortonworks®

protegrity

Solving Hadoop® Security

Enhanced Security for Sensitive Data in Hadoop Data Lakes

A Hortonworks and
Protegrity White Paper

Contents

Overview

As companies rush to put Big Data to work for their business, new ways of operating can sometimes get ahead of IT's ability to digest their full implications. There's no question that the creation of a Hadoop-powered Data Lake can provide a robust foundation for a new generation of analytics and insight, but it's important to consider security before launching or expanding a Hadoop initiative. By making sure that data protection and governance are built into your Big Data environment, you can leverage the full value of advanced analytics without exposing your business to new risks.

Hortonworks and Protegrity understand the importance of security and governance for every business. To ensure effective protection for our customers, we use a holistic approach based on five pillars:

- Administration
- Authentication and perimeter security
- Authorization
- Audit
- Data protection

In each of these areas, Hortonworks together with Protegrity provide differentiated capabilities beyond those of other vendors to help customers achieve the highest possible level of protection. As a result, Big Data doesn't have to incur big risks—and companies can put it to work without sacrificing peace of mind.

Understanding the security implications of the Data Lake

The consensus is strong among leading companies in every industry: data is an essential new driver of competitive advantage. Hadoop plays a critical role in the modern data architecture by providing low-cost, scale-out data storage and value-add processing. The successful Hadoop journey typically starts with Data Architecture Optimization or new Advanced Analytic Applications, which leads to the formation of a Data Lake. As existing and new types of data from sensors and machines, server logs, clickstreams, and other sources flow into the Data Lake, it serves as a central repository based on shared Hadoop services that power deep organizational insights across a large, broad and diverse set of data.

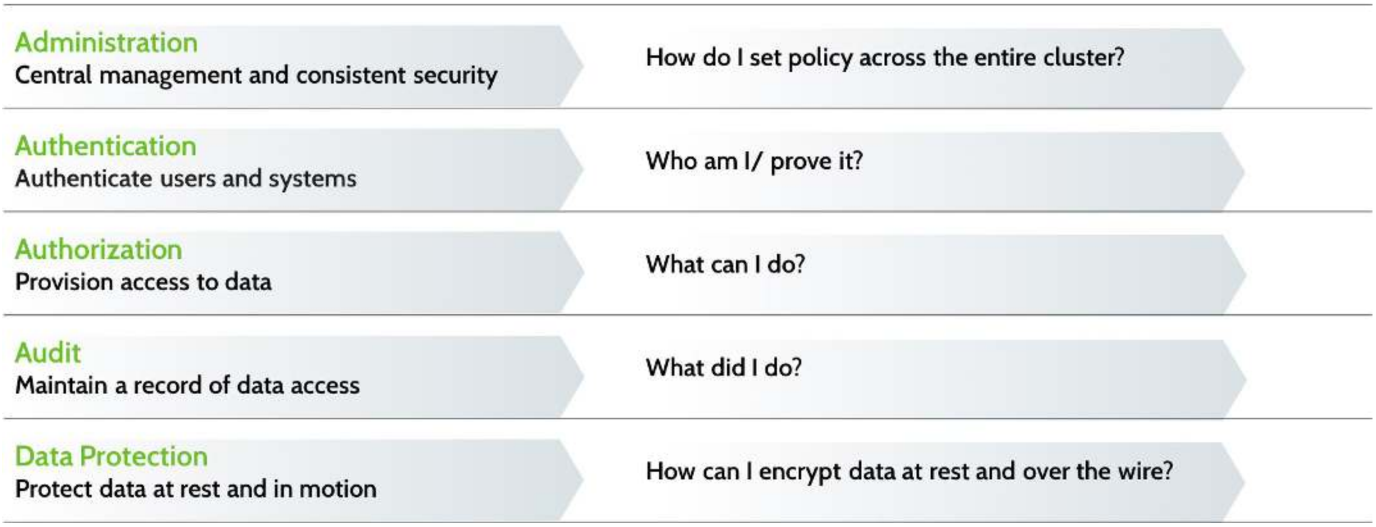
The need to protect the Data Lake with comprehensive security is clear. As large and growing volumes of diverse data are stored in the Data Lake, it comes to hold the crown jewels of your company—the vital and often highly sensitive data that has shaped and driven your business over a long history. However, the external ecosystem of data and operational systems feeding the Data Lake is highly dynamic and can introduce new security threats on a regular basis. Users across multiple business units can access the Data Lake freely and refine, explore and enrich its data at will, using methods of their own choosing, thereby increasing risks of exposure to unauthorized users. Any internal or external breach of this enterprise-wide data can be catastrophic, from privacy violations, to regulatory infractions, to damage to corporate image and long-term shareholder value. To prevent damage to the company's business, customers, finances and reputation, IT leaders must ensure that their Data Lake meets the same high standards of security as any legacy data environment.

Only as secure as the weakest link

Piecemeal protections are no more effective for a Data Lake than they would be in a traditional repository. There's no point in securing the primary access path to the data lake when a user can simply access the same data through a different path.

Hortonworks and Protegrity firmly believe that effective Hadoop security depends on a holistic approach. Our framework for comprehensive security revolves around five pillars: administration, authentication/ perimeter security, authorization, audit and data protection.

Five pillars of enterprise security



Administration Central management and consistent security	How do I set policy across the entire cluster?
Authentication Authenticate users and systems	Who am I/ prove it?
Authorization Provision access to data	What can I do?
Audit Maintain a record of data access	What did I do?
Data Protection Protect data at rest and in motion	How can I encrypt data at rest and over the wire?

Figure 1: Requirements for enterprise-grade security

Security administrators must address questions and provide enterprise-grade coverage across each of these pillars as they design the infrastructure to secure data in Hadoop. If any of these pillars remains weak, it introduces threat vectors to the entire data lake. In this light, your Hadoop security strategy must address all five pillars, with a consistent implementation approach to ensure their effectiveness.

Needless to say, you can't achieve comprehensive protection across the Hadoop stack through an ad-hoc approach. Security must be an integral part of the platform on which your Data Lake is built with a combination of bottom-up and top down approach. This makes it possible to enforce and manage security across the stack through a central point of administration and prevent gaps and inconsistencies. This approach is especially important for Hadoop implementations where new applications or data engines are always on the horizon in the form of new Open Source projects, a dynamic scenario that can quickly exacerbate any vulnerability.

Hortonworks and Protegrity help customers maintain the high levels of protection their enterprise data demands by building centralized security administration and management into the DNA of the Hortonworks Data Platform (HDP). HDP provides an enterprise-read data platform with rich capabilities spanning security, governance and operations. By implementing security at the platform level, Hortonworks ensures that security is consistently administered to any application built on top of the data platform, and makes it easier to build or retire data applications without impacting security. Protegrity enhances native Hortonworks security with additional data protection that provides advanced fine-grained tokenization and encryption capabilities to increase security while maintaining usability.

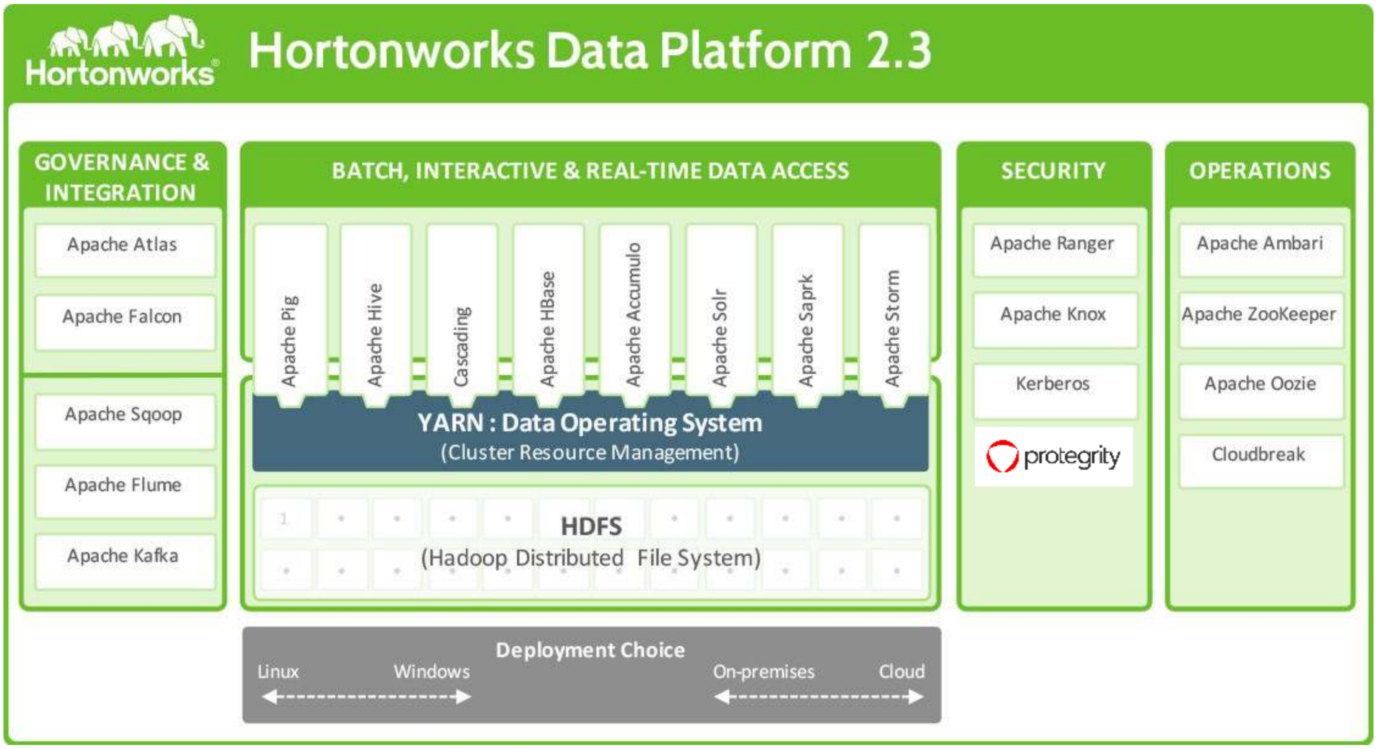


Figure 2: Hortonworks Data Platform and Protegrity

Commitment to enterprise-readiness

Hortonworks was founded with the objective to make Hadoop ready for the enterprise and has a strong legacy of significant contributions in this area. This goal of enterprise-readiness led the original Hadoop team at Yahoo! to develop Kerberos as the basis for strong authentication in Hadoop. Since that time, Hortonworks has continued to make significant investments in security.

In May 2014, Hortonworks acquired XA Secure, a leading data security company, to accelerate the delivery of a comprehensive approach to Hadoop security. To be consistent with its mission to develop, distribute and support 100% open source Apache Hadoop data platform, Hortonworks immediately incorporated the XA Secure technology into the Hortonworks Data Platform (HDP), while also converting the commercial solution into an open Apache community project called Apache Ranger.

Protegrity, a leading provider of data-centric enterprise data security solutions, partnered with Hortonworks to strengthen and expand the availability of data-centric protection and monitoring even further in the Hortonworks Data Platform (HDP). Protegrity Avatar™ for Hortonworks extends the capabilities of HDP native security with Protegrity Vaultless Tokenization (PVT) for Apache™ Hadoop®, Extended HDFS Encryption, and the Protegrity Enterprise Security Administrator (ESA), for advanced data protection policy, key management and auditing. Protegrity protects sensitive data in Hadoop from ingestion through consumption while also providing protection for other heterogeneous data sources under one single platform.

As part of HDP, Hortonworks features comprehensive security that spans across the five security pillars. Utilizing Protegrity components, HDP is enhanced further with advanced fine grained and coarse security capabilities to protect sensitive Hadoop data at use, in transit, or at rest. Together, Hortonworks and Protegrity enable IT to meet the requirements of Hadoop security better than any other solution available.

Administration

In order to deliver consistent security administration and management, Hadoop administrators require a centralized user interface—a single pane of glass that can be used to define, administer and manage security policies consistently across all the components of the Hadoop stack. Hortonworks addressed this requirement through Apache Ranger, an integral part of HDP, which provides a central point of administration for the other four functional pillars of Hadoop security. For central administration of Hadoop and other assets, Hortonworks utilizes Protegrity to provide heterogeneous capabilities centrally administer Hadoop together with other assets throughout the enterprise.

Ranger enhances the productivity of security administrators and reduces potential errors by empowering them to define security policy once and apply it to all the applicable components across the Hadoop stack from a central location.

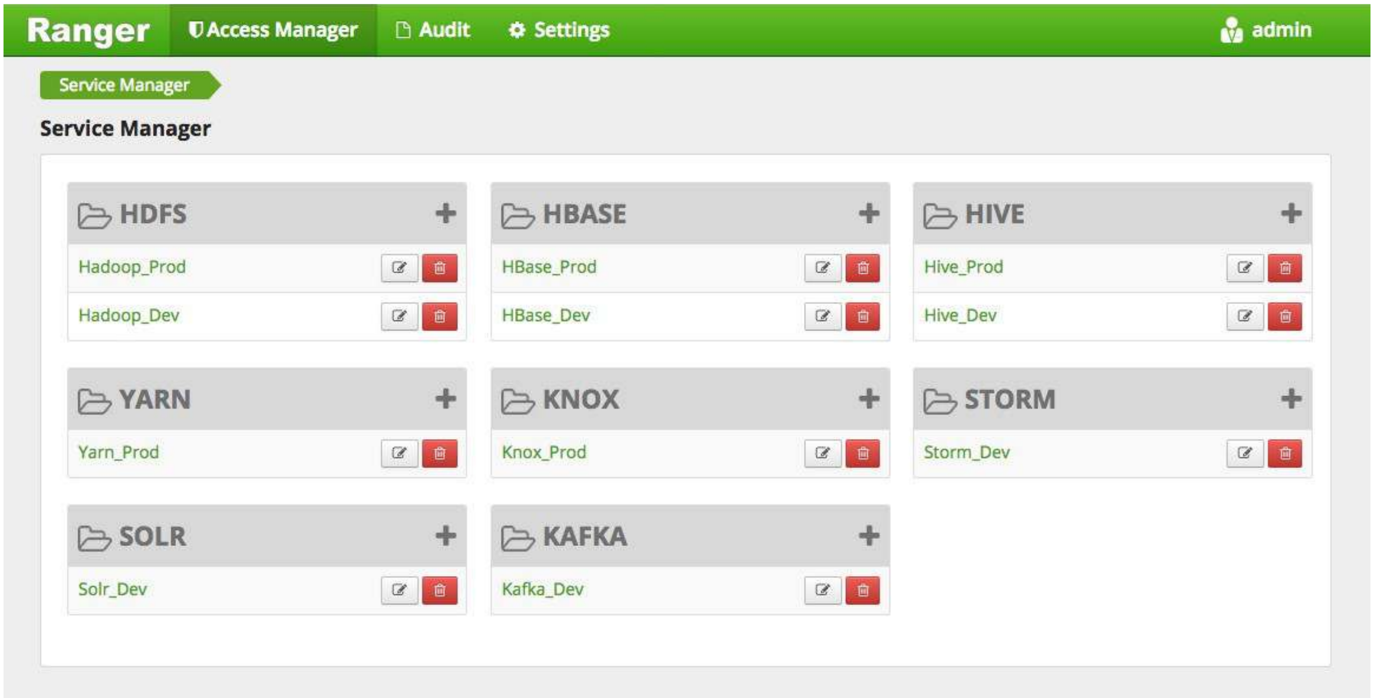


Figure 3: Apache Ranger provides a "single pane of glass" for the security administrator

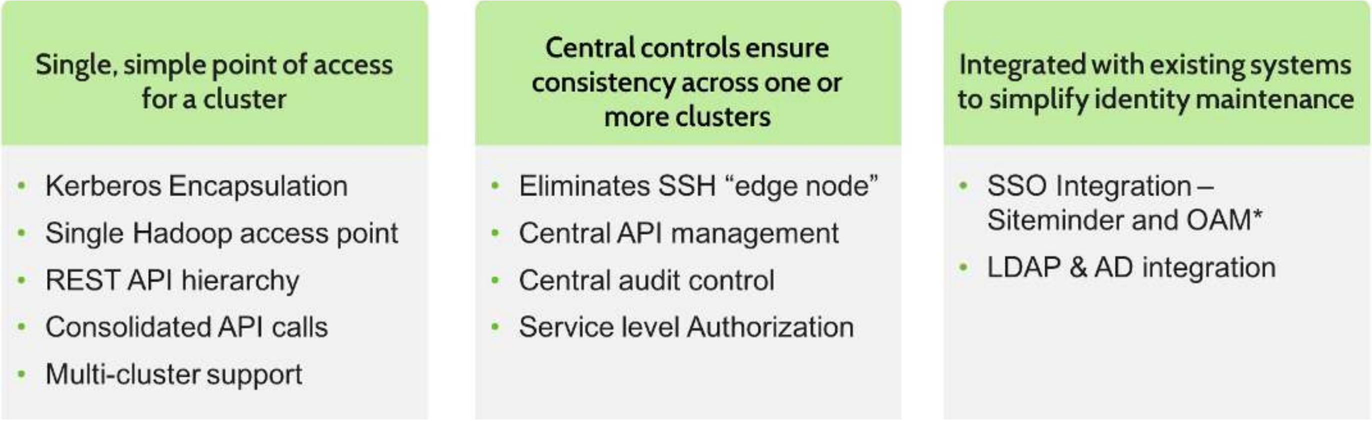
Other solutions for Hadoop enterprise security only offer partial administration across authentication, authorization, auditing and data protection/encryption—and lack the centralized administration and management needed for efficient and comprehensive security.

		APACHE RANGER
Centralized security administration	●	Apache Ranger provides a centralized platform for security policy administration

Authentication and perimeter security

Establishing user identity with strong authentication is the basis for secure access in Hadoop. Users need to reliably identify themselves and then have the identity propagated throughout the Hadoop cluster to access resources such as files and directories, and to perform tasks such as running MapReduce jobs. Hortonworks uses Kerberos, an industry standard, to authenticate users and resources within the Hadoop cluster. Hortonworks has also simplified Kerboeros setup, configuration and maintenance through Ambari 2.0.

Apache Knox Gateway ensures perimeter security for Hortonworks customers. With Knox, enterprises can confidently extend the Hadoop REST API to new users without Kerberos complexities, while also maintaining compliance with enterprise security policies. Knox provides a central gateway for Hadoop REST APIs that have varying degrees of authorization, authentication, SSL and SSO capabilities to enable a single access point for Hadoop.



Single, simple point of access for a cluster

- Kerberos Encapsulation
- Single Hadoop access point
- REST API hierarchy
- Consolidated API calls
- Multi-cluster support

Central controls ensure consistency across one or more clusters

- Eliminates SSH “edge node”
- Central API management
- Central audit control
- Service level Authorization

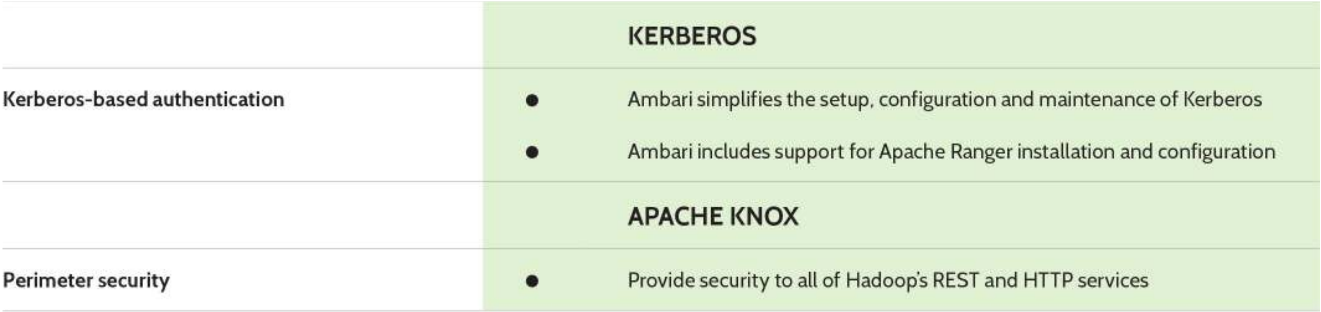
Integrated with existing systems to simplify identity maintenance

- SSO Integration – Siteminder and OAM*
- LDAP & AD integration

Figure 4: Perimeter security with Apache Knox

Other vendors fail to provide a comprehensive solution in this area, instead positioning Kerberos for perimeter security. Kerberos is an essential step for user authentication, but it is not sufficient in itself as it lacks the ability to hide cluster entry points and block access at the perimeter. By comparison, Apache Knox was built as a secure API gateway for Hadoop, with the ability to block services at the perimeter of the cluster. When using Apache Knox for REST APIs, cluster’s multiple access points are hidden from end users, adding another layer of protection for perimeter security.

Apache Knox is a pluggable framework, and a new REST API service can be added easily using a configurable services definition (Knox Stacks).



		KERBEROS
Kerberos-based authentication	●	Ambari simplifies the setup, configuration and maintenance of Kerberos
	●	Ambari includes support for Apache Ranger installation and configuration
		APACHE KNOX
Perimeter security	●	Provide security to all of Hadoop's REST and HTTP services

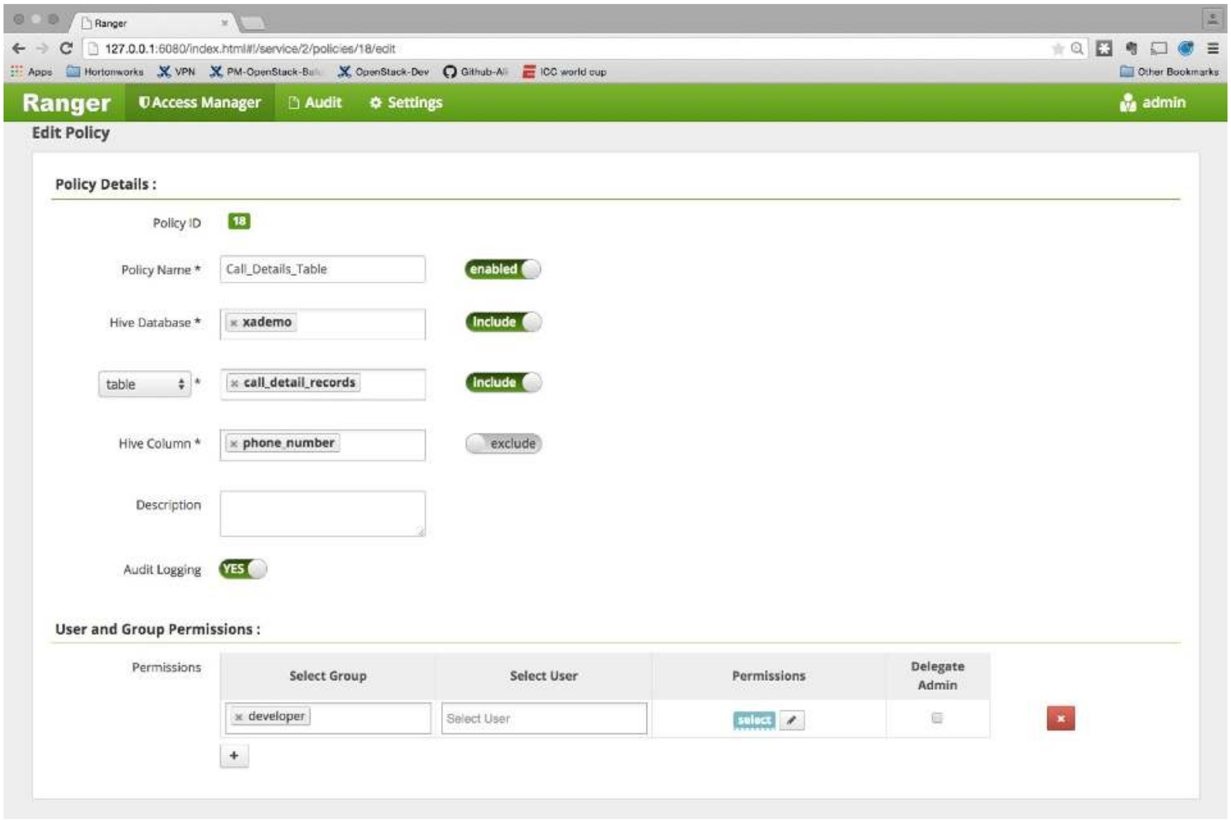
Authorization

Ranger manages fine-grained access control through a rich user interface that ensures consistent policy administration across Hadoop data access components. Security administrators have the flexibility to define security policies for a database, table and column or a file, and administer permissions for specific LDAP based groups or individual users. Rules based on dynamic conditions such as time or geography can also be added to an existing policy rule.

The **Ranger** authorization model is highly pluggable and can be easily extended to any data source using a service-based definition.Administrators can use Ranger to define centralized security policy for the following components:

- Apache Hadoop HDFS
- Apache Hadoop YARN
- Apache Hive
- Apache HBase
- Apache Storm
- Apache Knox
- Apache Solr
- Apache Kafka

Ranger works with standard authorization APIs in each Hadoop component and is able to enforce centrally administered policies for any method of accessing the data lake.



Edit Policy

Policy Details :

Policy ID 18

Policy Name * Call_Details_Table enabled

Hive Database * × xademo Include

table * × call_detail_records Include

Hive Column * × phone_number exclude

Description

Audit Logging YES

User and Group Permissions :

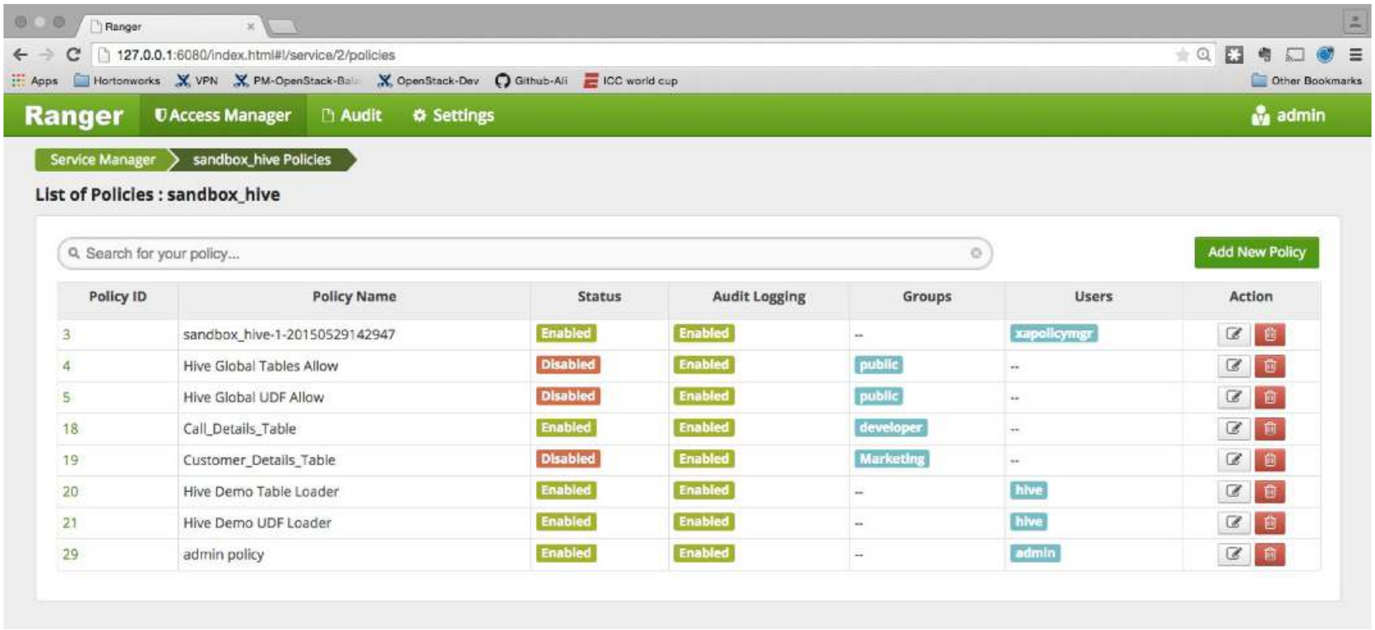
Permissions

Select Group	Select User	Permissions	Delegate Admin
× developer	Select User	select	□

Figure 5: Fine-grained security authorization policy definition with Apache Ranger

Solutions from other vendors lack the flexibility and rich user interface to enable administrators configure security policy for specific groups and individual users. In contrast, Ranger provides administrators with deep visibility into the security administration process that is required for auditing purposes. The combination of Ranger's rich user interface with deep audit visibility makes it highly intuitive to use, enhancing productivity for security administrators.

When extending policies beyond Hadoop, Hortonworks utilizes Protegrity to provide additional support for heterogeneous data sources such as files, enterprise applications, cloud applications, cloud storage and databases.



List of Policies : sandbox_hive

Policy ID	Policy Name	Status	Audit Logging	Groups	Users	Action
3	sandbox_hive-1-20150529142947	Enabled	Enabled	--	xapolicymgr	
4	Hive Global Tables Allow	Disabled	Enabled	public	--	
5	Hive Global UDF Allow	Disabled	Enabled	public	--	
18	Call_Details_Table	Enabled	Enabled	developer	--	
19	Customer_Details_Table	Disabled	Enabled	Marketing	--	
20	Hive Demo Table Loader	Enabled	Enabled	--	hive	
21	Hive Demo UDF Loader	Enabled	Enabled	--	hive	
29	admin policy	Enabled	Enabled	--	admin	

Figure 6: With Apache Ranger, administrators have complete visibility into the security administration process

		APACHE RANGER
Platform-wide coverage across Hadoop stack	●	Coverage across HDFS, YARN, Hive, HBase, Storm, Knox, Solr and Kafka
Fine grain authorization	●	Authorize security policies for a database, table and column or a file as well as LDAP based groups or individual user
Provide hooks for dynamic policy-based authorization	●	Specify dynamic conditions in service definitions Flexibility to define unique conditions by service (HDFS, Hive etc.)
Built on pluggable service-based model	●	Custom plugins can be created for any data store

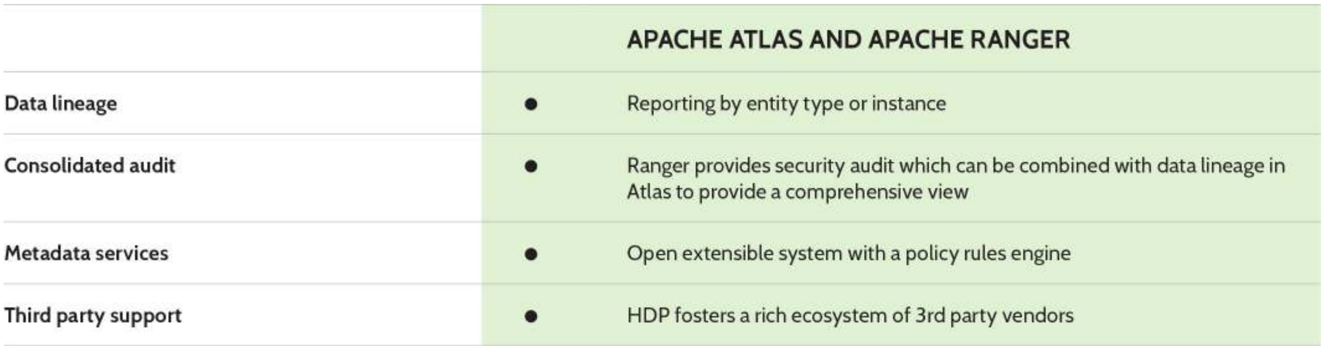
Audit

As customers deploy Hadoop into corporate data and processing environments, metadata and data governance must be vital parts of any enterprise-ready data lake. For these reasons, Hortonworks established the Data Governance Initiative (DGI) with Aetna, Merck, Target and SAS to introduce a common approach to Hadoop data governance into the open source community. This initiative has since evolved into a new open source project called Apache Atlas. Apache Atlas is a set of core foundational governance services that enables enterprises to effectively and efficiently meet their compliance requirements within Hadoop and allows integration with the complete enterprise data ecosystem. These services include:

- Search and lineage for datasets
- Metadata-driven data access control
- Indexed and searchable centralized auditing operational events
- Data lifecycle management from ingestion to disposition
- Metadata interchange with other tools

Ranger also provides a centralized framework for collecting access audit history and easily reporting on this data, including the ability to filter data based on various parameters. Together with Apache Atlas, this makes it possible for users to gain a comprehensive view of data lineage and access audit, with an ability to query and filter audit based on data classification, users or groups, and other filters.

For scenarios requiring centralized auditing of the enterprise, Hortonworks utilizes Protegrity enterprise security to achieve a complete view of enterprise security within Hadoop, a local enterprise assets, or resources in the cloud.



		APACHE ATLAS AND APACHE RANGER
Data lineage	●	Reporting by entity type or instance
Consolidated audit	●	Ranger provides security audit which can be combined with data lineage in Atlas to provide a comprehensive view
Metadata services	●	Open extensible system with a policy rules engine
Third party support	●	HDP fosters a rich ecosystem of 3rd party vendors

Data protection

Data protection adds a robust layer of security by making data unreadable to ensure that sensitive data is protected by internal and external threats. HDP supports encrypting network traffic as data into and through the Hadoop cluster over RPC, HTTP, Data Transfer Protocol (DTP) and JDBC. Network traffic over each of these protocols can be encrypted to provide privacy for data movement. The levels of protection, granularity of protection, and types of protection vary depending on the solution chosen. HDP with Ranger provides coarse-grained protection at the disk level to protect data at rest. This protects Hadoop data from being read in the clear outside of standard access controls. Protegrity provides coarse-grained protection and adds additional capabilities for different types of fine-grained protection to enable specific fields or data elements. The additional capabilities provide greater data protection while allowing for analysis of data without divulging sensitive data.

HDP with Ranger satisfies enterprise requirements for security and compliance through encryption for data at rest via coarse-grained protection. HDP supports the ability to encrypt files stored in Hadoop, including a Ranger-embedded open source Hadoop key management store (KMS). Ranger provides security administrators with the ability to manage keys and authorization policies for KMS. With HDP, our customers have the flexibility to leverage open source key management store (KMS) or use enterprise wide KMS solutions provided by others.

Encryption in HDFS, combined with KMS access policies maintained by Ranger, prevents rogue Linux or Hadoop administrators from accessing data and supports segregation of duties for both data access and encryption.

Protegrity adds additional protection to Hadoop with advanced coarse grained and fine-grained protection. All data is protected at rest, in use, and in transit to ensure data security throughout its lifecycle. Protegrity and HDP both utilize HDFS encryption for coarse-grained protection. However, Protegrity offers additional protection by encrypting or tokenizing the sensitive information in each record, assuring the security of the data from internal and external threats.

Protegrity's protection follows the data - guaranteeing its security regardless of where it is. The data is protected the moment it is created or imported, allowing the data to remain secure until it needs to be viewed in the clear by an authorized individual. When multiple systems feed a Hadoop data lake, Protegrity provides protection at the various points of entry as well as within the Hadoop cluster itself. Depending on the policies set, different users and systems may have different views of the data. All policies are enforced as the data moves between systems.

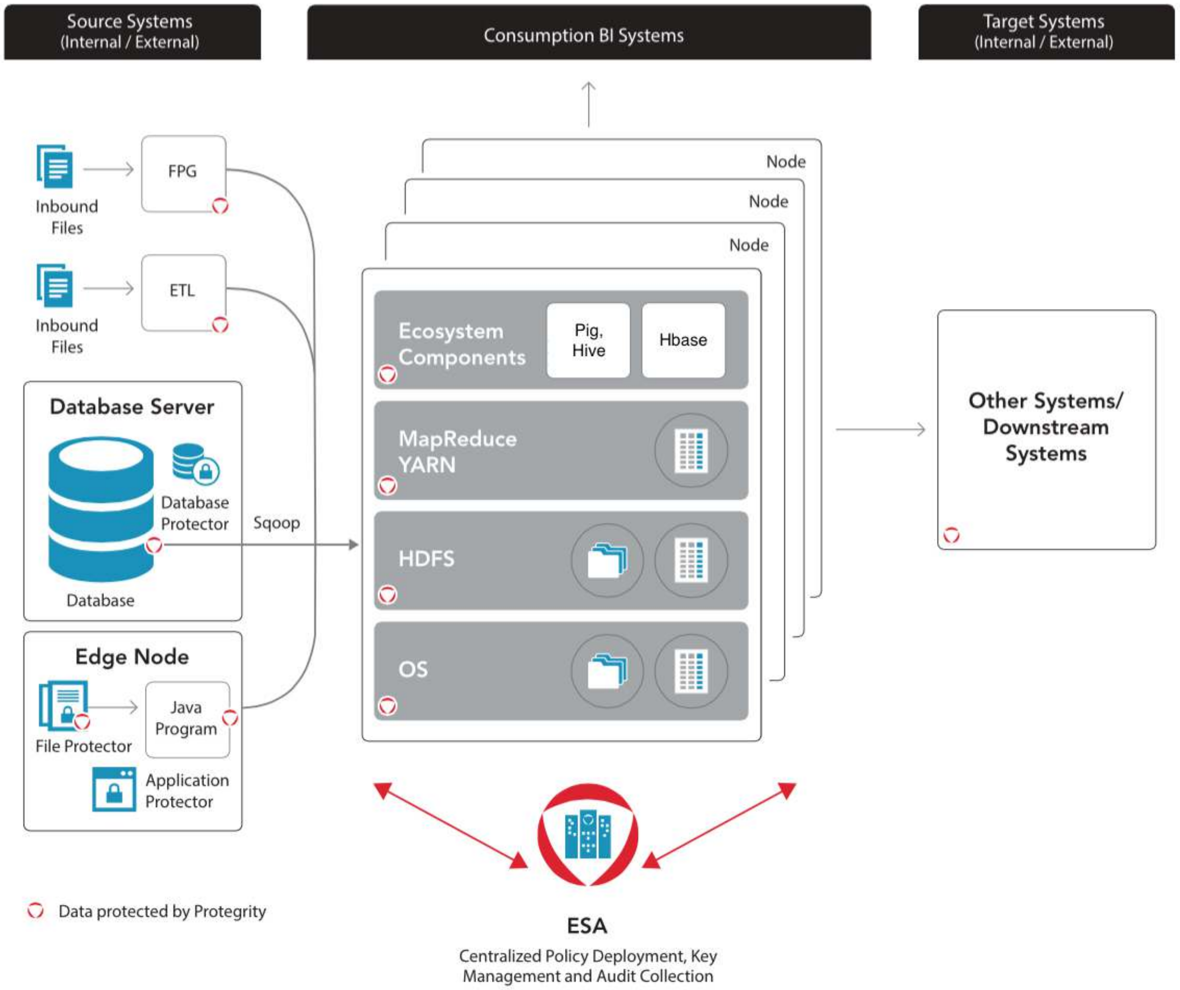


Figure 7: Protegrity secures data throughout its lifecycle regardless of source or destination in Hadoop.

Protegrity leverages Hadoop clusters to provide the highest performance encryption and tokenization available. By leveraging Hadoop, Protegrity scales with every node and distributes the workload across the cluster via its own cluster management capability.

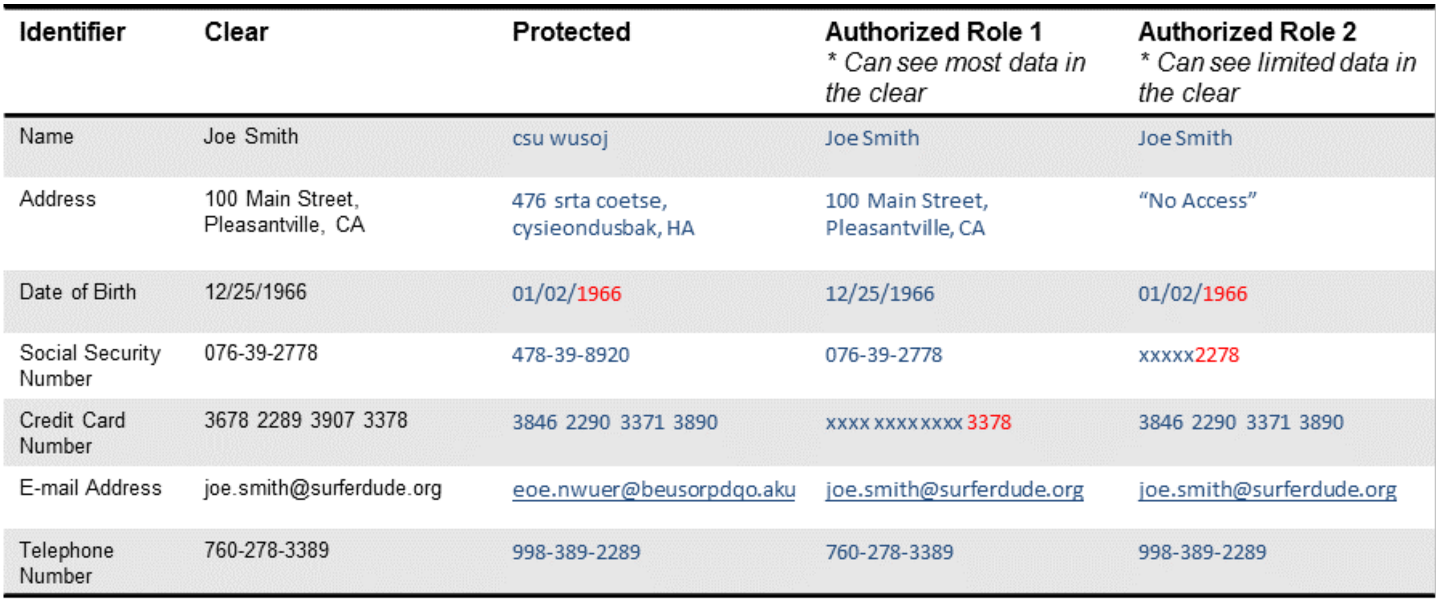
Protection Methods on Hadoop Nodes with Protegrity

- Fine-grained
    - Reversible
        - Vaultless Tokenization
        - Encryption
        - Format Preserving Encryption (FPE)
    - Irreversible
        - Masking
- Coarse Grained
    - HDFS Encryption

Different methods of fine-grained protection are available for sensitive fields and data. Tokenization, encryption, and masking can protect specific data within Hadoop. Tokenization capabilities provide the unique ability for existing consuming applications to utilize protected data without modification. Sensitive data is tokenized by Protegrity utilizing robust Vaultless Tokenization technology to provide significant performance gains over typical vault based tokenization, File Preserving Encryption, or AES encryption. This allows for de-identification of large amounts of sensitive data with minimal performance impact – a significant benefit for analysis within a large Hadoop Data Lake.

Protection of data within Hadoop can be specified at the row, column, label, cell, job, or file level. Protegrity's ability to protect Hadoop data at the row level is an industry first. Traditional column level protection allows for policies governing whether a field is encrypted or tokenized for a particular role or user. Row level protection adds additional capability to determine whether individual records may be viewed by a role or user, allowing for the protection of entire records in addition to specific fields of data. This type of functionality is ideal for multi-tenant situations in Hadoop to provide isolation of data by different tenants within the same Hadoop installation.

Protegrity expands security functionality with policy definitions that allow for specific criteria on how data is represented to a consuming application, user, or role. These policies can dictate what data can be seen, whether or not it is encrypted or tokenized, and how much of the data is shown to the user. This type of functionality allows for usability of data by different parties for analysis while preserving data security. Some roles may see all sensitive data while others may see tokenized, partially revealed, or masked data. In applications such as business intelligence where aggregate data is important, elements such as the year may be revealed while tokenizing the month and date for security.



Identifier	Clear	Protected	Authorized Role 1 * Can see most data in the clear	Authorized Role 2 * Can see limited data in the clear
Name	Joe Smith	csu wusoj	Joe Smith	Joe Smith
Address	100 Main Street, Pleasantville, CA	476 srta coetse, cysieondusbak, HA	100 Main Street, Pleasantville, CA	"No Access"
Date of Birth	12/25/1966	01/02/1966	12/25/1966	01/02/1966
Social Security Number	076-39-2778	478-39-8920	076-39-2778	xxxxx2278
Credit Card Number	3678 2289 3907 3378	3846 2290 3371 3890	xxxx xxxxxxxx 3378	3846 2290 3371 3890
E-mail Address	joe.smith@surferdude.org	eoe.nwuer@beusorpdqo.aku	joe.smith@surferdude.org	joe.smith@surferdude.org
Telephone Number	760-278-3389	998-389-2289	760-278-3389	998-389-2289

Figure 8: Data is protected and de-identified differently based on the role of the user

The Protegrity Enterprise Security Administrator centrally manages, audits, and reports on all the different forms of data protection inside and outside of Hadoop. As a heterogeneous solution, Protegrity allows the security of data to be consistently managed and administered across environments - minimizing impact while maximizing portability of the data regardless of where it is used.

	Hortonworks + Protegrity Protection
Robust Security with High Performance	Industry leading patented data protection utilizing Vaultless Tokenization and Encryption with high performance utilizing processing on Hadoop nodes with built in cluster management
Coarse Grained Data Protection	Full support for HDFS encryption with key management
Heterogeneous Enterprise Wide Protection	Protected data is supported across the data lifecycle across Hadoop, applications, databases, cloud, and file systems allowing for universal policies for data inside and outside of Hadoop
Flexible Security and Policies of Sensitive Data Elements for Multiple Use Cases	Advanced policies can be utilized to support complex use cases with advanced masking, encryption, and tokenization options to deal with different data types and rules governing what tokenized data should look like
Customizable Fine Grained Data Protection	• Column level protection • Row level protection • Cell level protection • File protection • Label level protection
Multiple Data Protection Methods	• Tokenization • AES Encryption • Format Preserving Encryption • Masking
Industry's First Row Level Protection for Hadoop	Row level protection of records for multi-tenant Hadoop installations allow for isolation and security of individual tenant data

Enabling extensibility through a pluggable framework

Apache Ranger and Knox provide a service-based architecture for policy administration and enforcement that enables customers to easily support new components or data engines. Applications that integrate with Ranger can simply use the pluggable architecture to leverage the existing security policies in Ranger without the need to redefine the security policy from scratch. Users can also create custom services as plug-ins to any data store and build and manage services centrally for their big data BI applications. Similarly, REST APIs can easily be integrated with Apache Knox using a service-based configuration. The configurable framework also enables partner solutions to easily work with Ranger and Knox to extend a new service to support a new component or data engine. Protegrity provides enhanced security for HDFS with full integration with Hortonworks and support for the Hadoop ecosystem. As new libraries and API's emerge for Hadoop, Protegrity is fully extensible to utilize these frameworks and services.

Comparison across security pillars

Point by point, pillar by pillar, HDP with Protegrity provides rich capabilities to protect sensitive data within your Hadoop Data Lake. In each aspect, Hortonworks with Protegrity exceeds what's possible with competing solutions. The ground-up platform design of the Hortonworks security architecture ensures that each pillar complements the others to ensure that no gaps remain. Protegrity's advanced tokenization and encryption technologies ensure that data is always protected throughout the lifecycle with Hadoop. The Hadoop Data Lake becomes fully usable while maintaining security. The centralization of administration and management enables IT to take a holistic approach to security, as well as increasing IT productivity and efficiency. By taking a approach based on a pluggable and completely 100% open source design, Hortonworks partnered with Protegrity to embed additional robust security capabilities to the HDP platform. In this way, companies can leverage comprehensive, easy to use security to their data and apply the same level of rigor to Hadoop as they would for traditional data environments—as it should be.

	HDP and PROTEGRITY	OTHER DISTRIBUTIONS
Administration	●	◑
Authentication	●	◑
Perimeter Security	●	○
Authorization	●	◑
Audit	●	◑
Data Protection	●	◑

Figure 9: Competitive analysis across security pillars

Summary

No business can afford to have Big Data insight come at the expense of enterprise security. As you plan your Hadoop strategy, make sure that the platform you choose provides a comprehensive and holistic approach to protect your Data Lake and the valuable information it contains. With Hortonworks and Protegrity, companies can implement a platform with all five pillars of Hadoop security woven into its architecture – administration, authentication/perimeter security, authorization, audit and data protection. Unlike other Hadoop platforms that offer only partial security functionality, HDP with Protegrity is all IT needs to support a complete strategy for secure Big Data – making its competitive benefits available to business users throughout the organization without exposing the company to risk.

About Hortonworks

Hortonworks develops, distributes and supports the only 100% open source Apache Hadoop data platform. Our team comprises the largest contingent of builders and architects within the Hadoop ecosystem who represent and lead the broader enterprise requirements within these communities. Hortonworks Data Platform deeply integrates with existing IT investments upon which enterprises can build and deploy Hadoop-based applications. Hortonworks has deep relationships with the key strategic data center partners that enable our customers to unlock the broadest opportunities from Hadoop. For more information, visit www.hortonworks.com.

About Protegrity

Protegrity is the only enterprise data security software platform that leverages scalable, data-centric encryption, tokenization and masking to help businesses secure sensitive information while maintaining data usability. Built for complex, heterogeneous business environments, the Protegrity Data Security Platform provides unprecedented levels of data security certified across applications, data warehouses, mainframes, big data, and cloud environments. Companies trust Protegrity to help them manage risk, achieve compliance, enable business analytics, and confidently adopt new platforms.

Protegrity is headquartered in Stamford, Connecticut, USA.

For additional information visit www.protegrity.com or call +1-203-326-7200.