HDP Security Overview

Date of Publish: 2018-07-15

http://docs.hortonworks.com

Contents

HDP Security Overview

Security is essential for organizations that store and process sensitive data in the Hadoop ecosystem. Many organizations must adhere to strict corporate security polices. Hadoop is a distributed framework used for data storage and large-scale processing on clusters using commodity servers. Adding security to Hadoop is challenging because not all of the interactions follow the classic client-server pattern.

- In Hadoop, the file system is partitioned and distributed, requiring authorization checks at multiple points.
- A submitted job is executed at a later time on nodes different than the node on which the client authenticated and submitted the job.
- Secondary services such as a workflow system access Hadoop on behalf of users.
- A Hadoop cluster scales to thousands of servers and tens of thousands of concurrent tasks.

A Hadoop-powered "Data Lake" can provide a robust foundation for a new generation of Big Data analytics and insight, but can also increase the number of access points to an organization's data. As diverse types of enterprise data are pulled together into a central repository, the inherent security risks can increase.

Hortonworks understands the importance of security and governance for every business. To ensure effective protection for its customers, Hortonworks uses a holistic approach based on five core security features:

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

This chapter provides an overview of the security features implemented in the Hortonworks Data Platform (HDP). Subsequent chapters in this guide provide more details on each of these security features.

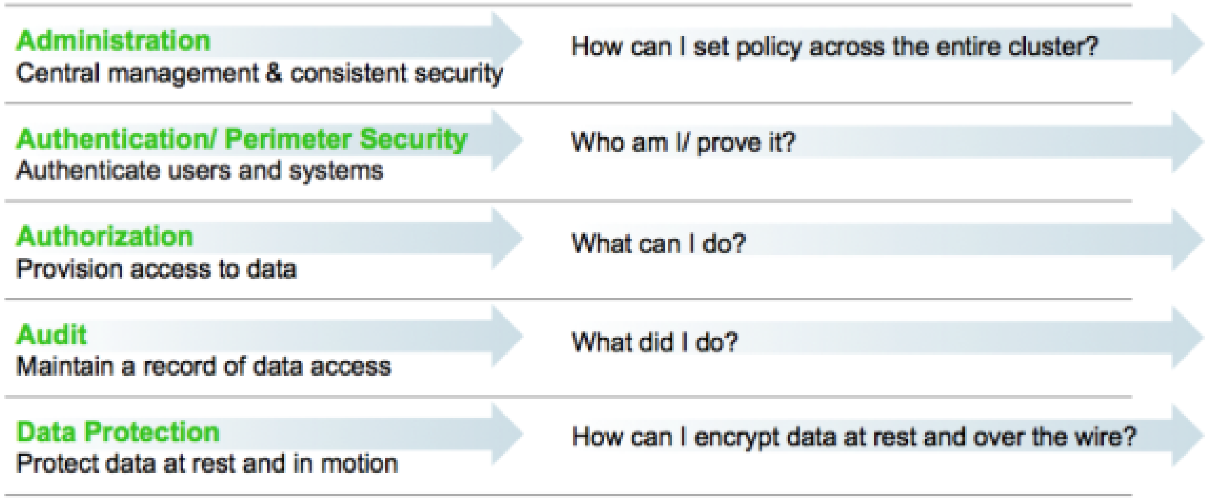
Understanding Data Lake Security

The successful Hadoop journey typically starts with data architecture optimization or new advanced analytic applications, which leads to the formation of what is known as a Data Lake. To prevent damage to the company’s business, customers, finances, and reputation, a Data Lake should meet the same high standards of security as any legacy data environment.

The general consensus in nearly every industry is that data is an essential new driver of competitive advantage. Hadoop plays a critical role in the modern data architecture by providing low-cost, large-scale data storage and processing. The successful Hadoop journey typically starts with data architecture optimization or new advanced analytic applications, which leads to the formation of what is known as a Data Lake. As new and existing types of data from machine sensors, server logs, clickstream data, 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 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 channeled into the Data Lake, it will store vital and often highly sensitive business data. 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, using methods of their own choosing, further increasing the risk of a breach. Any breach of this enterprise-wide data can result in catastrophic consequences: privacy violations, regulatory infractions, or the compromise of vital corporate intelligence. To prevent damage to the company’s business, customers, finances, and reputation, a Data Lake should meet the same high standards of security as any legacy data environment.

Piecemeal protections are no more effective for a Data Lake than they would be in a traditional repository. Effective Hadoop security depends on a holistic approach that revolves around five pillars of security: administration, authentication and perimeter security, authorization, auditing, and data protection.

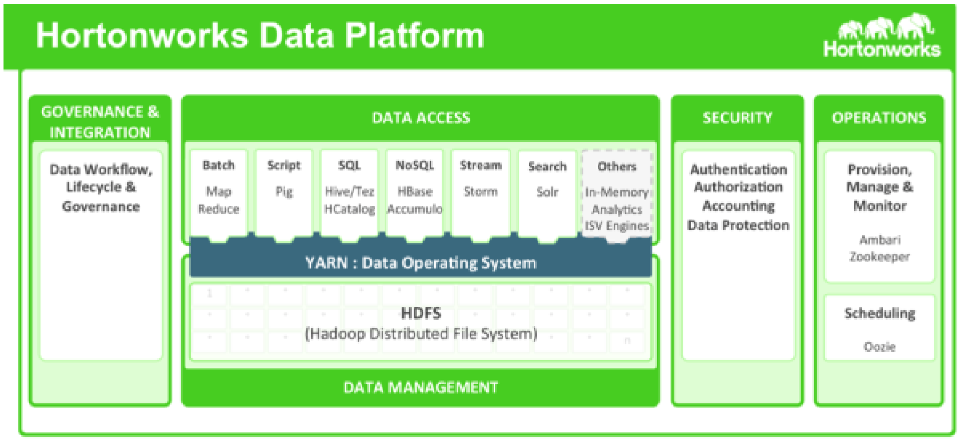


Administration Central management & consistent security	How can I set policy across the entire cluster?
Authentication/ Perimeter Security Authenticate users and systems	Who am I/ prove it?
Authorization Provision access to data	What can I do?
Audit Maintain a record of data access	What did I do?
Data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

Security administrators must address questions and provide enterprise-grade coverage across each of these areas as they design the infrastructure to secure data in Hadoop. If any of these pillars is vulnerable, it becomes a risk factor in the company's Big Data environment. A Hadoop security strategy must address all five pillars, with a consistent implementation approach to ensure effectiveness.

You cannot achieve comprehensive protection across the Hadoop stack by using an assortment of point solutions. Security must be an integral part of the platform on which your Data Lake is built. This bottom-up approach makes it possible to enforce and manage security across the stack through a central point of administration, thereby preventing gaps and inconsistencies. This approach is especially important for Hadoop implementations in which new applications or data engines are always emerging in the form of new Open Source projects — a dynamic scenario that can quickly exacerbate any vulnerability.

Hortonworks helps customers maintain high levels of protection for enterprise data by building centralized security administration and management into the infrastructure of the Hortonworks Data Platform. HDP provides an enterprise-ready data platform with rich capabilities spanning security, governance, and operations. HDP includes powerful data security functionality that works across component technologies and integrates with preexisting EDW, RDBMS, and MPP systems. By implementing security at the platform level, Hortonworks ensures that security is consistently administered to all of the applications across the stack, simplifying the process of adding or removing Hadoop applications.


Hortonworks Data Platform
Hortonworks
GOVERNANCE & INTEGRATION
Data Workflow, Lifecycle & Governance
DATA ACCESS
Batch
Map Reduce
Script
Pig
SQL
Hive/Tez HCatalog
NoSQL
HBase Accumulo
Stream
Storm
Search
Solr
Others
In-Memory Analytics ISV Engines
YARN : Data Operating System
HDFS
(Hadoop Distributed File System)
DATA MANAGEMENT
SECURITY
Authentication Authorization Accounting Data Protection
OPERATIONS
Provision, Manage & Monitor
Ambari Zookeeper
Scheduling
Oozie

What's New in This Release: Knox

New features and changes for Apache Knox have been introduced in Hortonworks Data Platform, along with documentation updates. New features are described in the following sections.

- Hortonworks Data Platform 3.0.0
  - Proxy: dynamic topology generation in the new Admin UI
  - New ambari-server setup-sso command for setting up Knox SSO
  - Support for multiple NameNodes in a federated cluster
  - Proxy for new component UIs: YARN, Oozie, Spark 2, HDFS, MapReduce2, Log Search, Livy (API only), and SmartSense
  - SSO for new components: Zeppelin, YARN, MapReduce2, HDFS, Oozie, and Log Search

What's New in This Release: Ranger

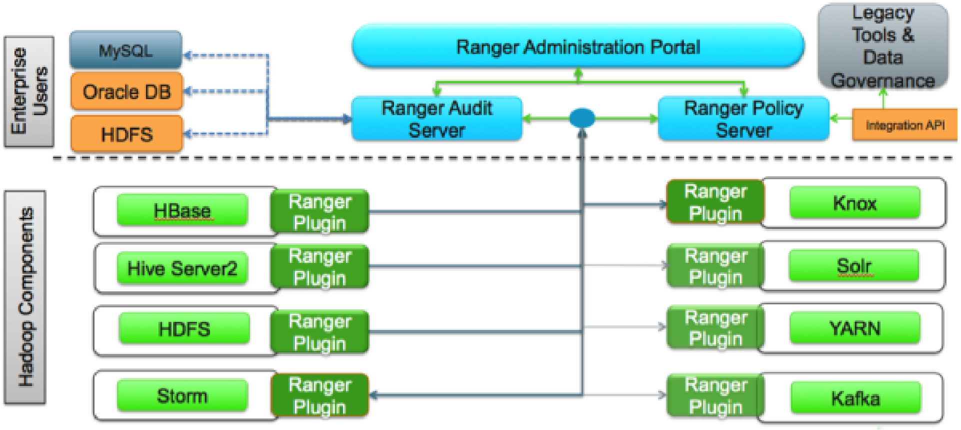
New features and changes for Apache Ranger have been introduced in Hortonworks Data Platform, along with documentation updates. New features are described in the following sections.

- Hortonworks Data Platform 3.0.0
  - Ranger included by default in Ambari installation
  - Support for multiple NameNodes in a federated cluster
  - Support for read-only Ranger admin users
  - Time-bound and temporary authorization policies and policy conditions
  - Auditing for Usersync operations
  - Ability to label policies, filter/search and show policies by labels

HDP Security Features

HDP uses Apache Ranger to provide centralized security administration and management. The Ranger Administration Portal is the central interface for security administration. You can use Ranger to create and update policies, which are then stored in a policy database.

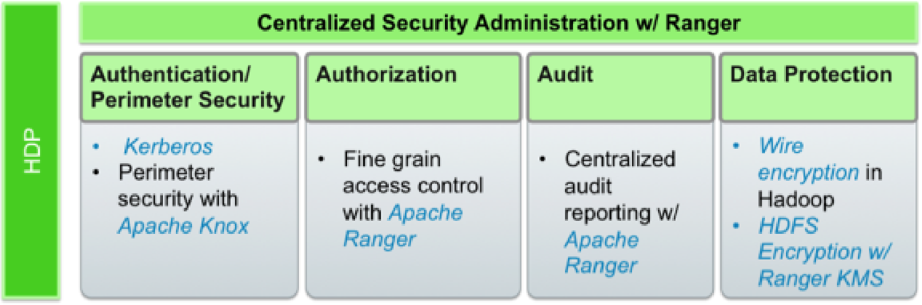
Ranger plug-ins (lightweight Java programs) are embedded within the processes of each cluster component. For example, the Ranger plug-in for Apache Hive is embedded within HiveServer2:



These plug-ins pull policies from a central server and store them locally in a file. When a user request comes through the component, these plug-ins intercept the request and evaluate it against the security policy. Plug-ins also collect data from the user request and follow a separate thread to send this data back to the audit server.

Administration

To deliver consistent security administration and management, Hadoop administrators require a centralized user interface they can use to define, administer and manage security policies consistently across all of the Hadoop stack components:

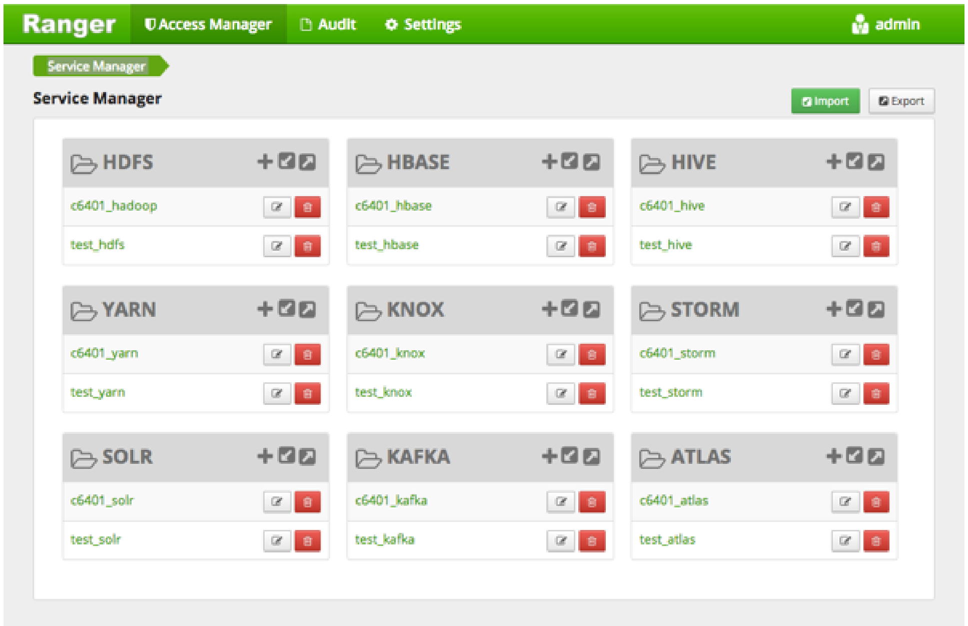


HDP

Centralized Security Administration w/ Ranger

Authentication/ Perimeter Security	Authorization	Audit	Data Protection
- Kerberos - Perimeter security with Apache Knox	- Fine grain access control with Apache Ranger	- Centralized audit reporting w/ Apache Ranger	- Wire encryption in Hadoop - HDFS Encryption w/ Ranger KMS

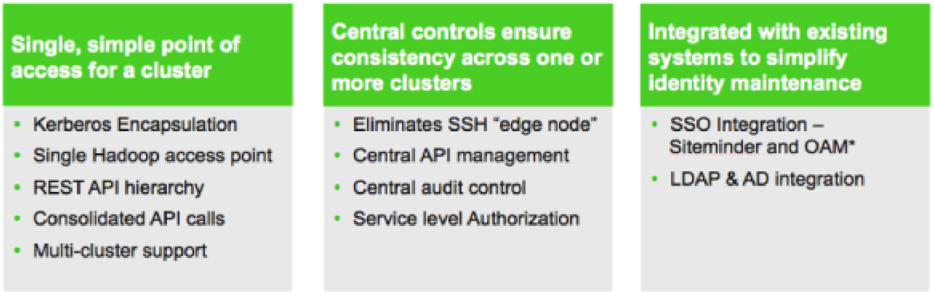
The Apache Ranger administration console provides a central point of administration for the other four pillars of Hadoop security.



Authentication and Secure Gateway

Establishing user identity with strong authentication is the basis for secure access in Hadoop. Users need to reliably identify themselves and then have that identity propagated throughout the Hadoop cluster to access cluster resources. Hortonworks uses Kerberos for authentication. Kerberos is an industry standard used to authenticate users and resources within a Hadoop cluster. HDP also includes Ambari, which simplifies Kerberos setup, configuration, and maintenance.

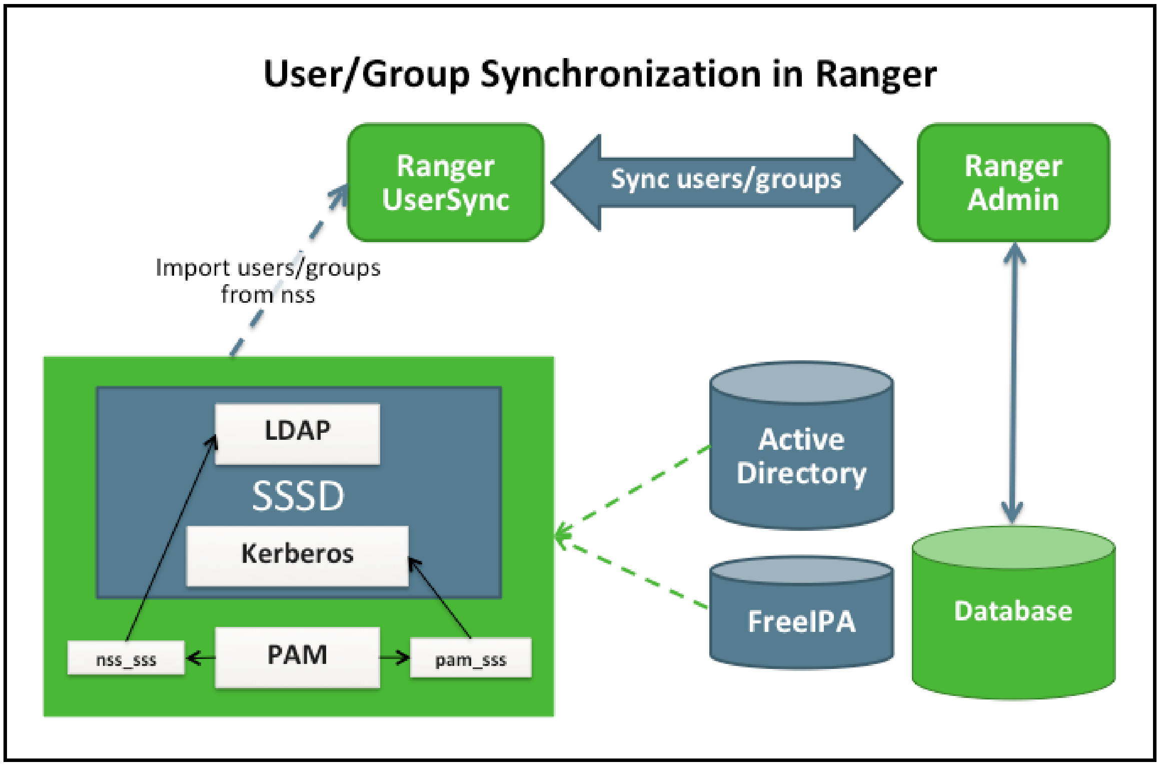
Apache Knox Gateway is used to help ensure perimeter security for Hortonworks customers. With Knox, enterprises can confidently extend the Hadoop REST API to new users without Kerberos complexities, while also maintaining compliance with enterprise security policies. Knox provides a central gateway for 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	Central controls ensure consistency across one or more clusters	Integrated with existing systems to simplify identity maintenance
• Kerberos Encapsulation • Single Hadoop access point • REST API hierarchy • Consolidated API calls • Multi-cluster support	• Eliminates SSH "edge node" • Central API management • Central audit control • Service level Authorization	• SSO Integration – Siteminder and OAM* • LDAP & AD integration

Authorization

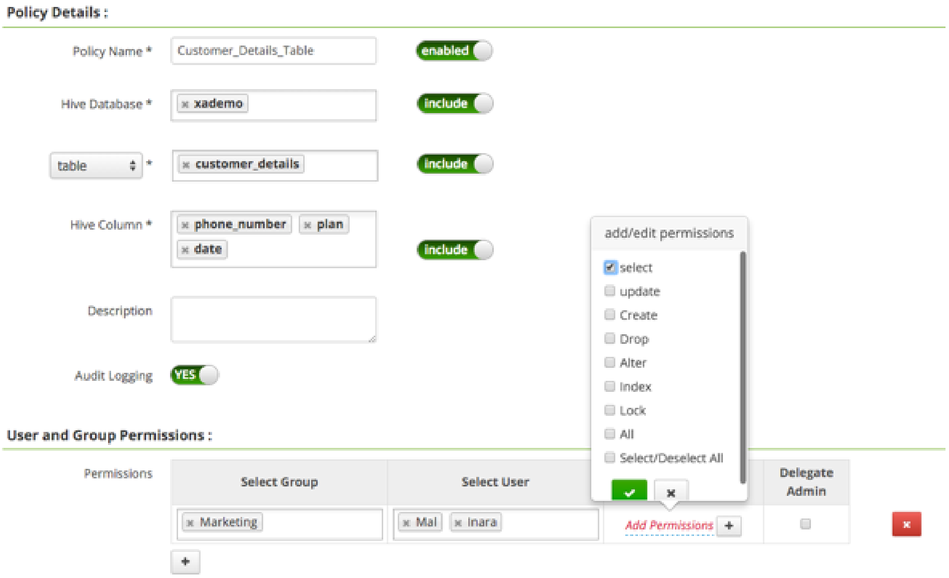
Ranger manages access control through a user interface that ensures consistent policy administration across Hadoop data access components. Security administrators can define security policies at the database, table, column, and file levels, and can administer permissions for specific LDAP-based groups or individual users. Rules based on dynamic conditions such as time or geolocation can also be added to an existing policy rule. The Ranger authorization model is pluggable and can be easily extended to any data source using a service-based definition.



Administrators can use Ranger to define a centralized security policy for the following Hadoop components:

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

Ranger works with standard authorization APIs in each Hadoop component and can enforce centrally administered policies for any method used to access the Data Lake.



Policy Details :

Policy Name * Customer_Details_Table enabled

Hive Database * xademo include

table * customer_details include

Hive Column * phone_number plan date include

Description

Audit Logging YES

add/edit permissions

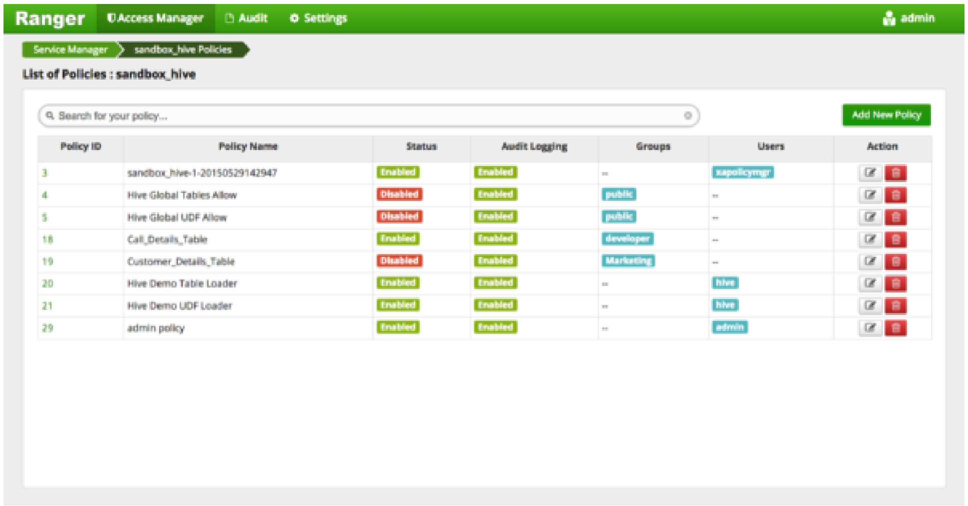
☑ select
☐ update
☐ Create
☐ Drop
☐ Alter
☐ Index
☐ Lock
☐ All
☐ Select/Deselect All

User and Group Permissions :

Permissions

Select Group	Select User		Delegate Admin
Marketing	Mal Inara	Add Permissions +	☐

Ranger provides administrators with the deep visibility into the security administration process that is required for auditing. The combination of a rich user interface and deep audit visibility makes Ranger highly intuitive to use, enhancing productivity for security administrators.



List of Policies : sandbox_hive

Search for your policy...

Add New Policy

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	

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 this reason, 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 named Apache Atlas. Apache Atlas is a set of core governance services that enables enterprises to meet their compliance requirements within Hadoop, while also enabling integration with the complete enterprise data ecosystem. These services include:

- Dataset search and lineage operations
- Metadata-driven data access control
- Indexed and searchable centralized auditing
- Data lifecycle management from ingestion to disposition
- Metadata interchange with other tools

Ranger also provides a centralized framework for collecting access audit history and reporting this data, including filtering on various parameters. HDP enhances audit information that is captured within different components within Hadoop and provides insights through this centralized reporting capability.

Data Protection

The data protection feature makes data unreadable both in transit over the network and at rest on a disk. HDP satisfies security and compliance requirements by using both transparent data encryption (TDE) to encrypt data for HDFS files, along with a Ranger-embedded open source Hadoop key management store (KMS). Ranger enables security administrators to manage keys and authorization policies for KMS. Hortonworks is also working extensively with its encryption partners to integrate HDFS encryption with enterprise-grade key management frameworks.

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.

Related Information

Hortonworks Establishes Data Governance Initiative

Dynamically Generating Knox Topology Files

Topology files can be dynamically generated from combinations of Provider Configurations and Descriptors, which can be defined using the Knox Admin UI.

Prior to HDP 3.0, you set up Knox proxy by editing topology files manually. Topology files consisted of 3 things:

- Provider configurations: e.g., authentication, federation, authentication, authorization, identity assertion, etc
- HA provider
- Services: component URLs you want to proxy

You configured each of these things in every topology file.

As of HDP 3.0, topology files are dynamically generated from combinations of Provider Configurations and Descriptors, defined using the Knox Admin UI. Additionally, these provider configurations and descriptors are now shared- you no longer have to specify configurations (e.g. authentication provider, identity assertion provider, or authorization provider) for each topology file- you define a Provider Configuration or Descriptor and they are shared across all topologies you choose. The Admin UI consists of 3 sections:

- Provider Configurations: A named set of providers, e.g., authentication, federation, authentication, authorization, identity assertion, etc. Provider configurations can be shared across descriptors/topologies.
- Descriptors: References the Provider Configurations to declare the policy (authentication, authorization, identity assertion, etc) that goes along with proxying that cluster. Descriptors cannot be shared across topologies; Descriptors and topologies are 1-to-1.
- Topologies: Dynamically generated based on the Provider Configurations and Descriptors you define.

However- the same topologies that were manageable in Ambari previously, still are. Within the Knox Admin UI, the topologies that are managed by Ambari should be read-only. Within an Ambari managed cluster, the Knox Admin UI is to be used for creating additional topologies. When a Knox instance is not managed by Ambari, all topology management will be done via the Knox Admin UI.

Securing Access to Hadoop Cluster: Apache Knox

The Apache Knox Gateway (“Knox”) is a system to extend the reach of Apache™ Hadoop® services to users outside of a Hadoop cluster without reducing Hadoop Security. Knox also simplifies Hadoop security for users who access the cluster data and execute jobs. The Knox Gateway is designed as a reverse proxy.

Establishing user identity with strong authentication is the basis for secure access in Hadoop. Users need to reliably identify themselves and then have that identity propagated throughout the Hadoop cluster to access cluster resources.

Layers of Defense for a Hadoop Cluster

- Authentication: Kerberos

  Hortonworks uses Kerberos for authentication. Kerberos is an industry standard used to authenticate users and resources within a Hadoop cluster. HDP also includes Ambari, which simplifies Kerberos setup, configuration, and maintenance.
- Perimeter Level Security: Apache Knox

  Apache Knox Gateway is used to help ensure 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.
- Authorization: Ranger

  OS Security: Data Encryption and HDFS

Apache Knox Gateway Overview

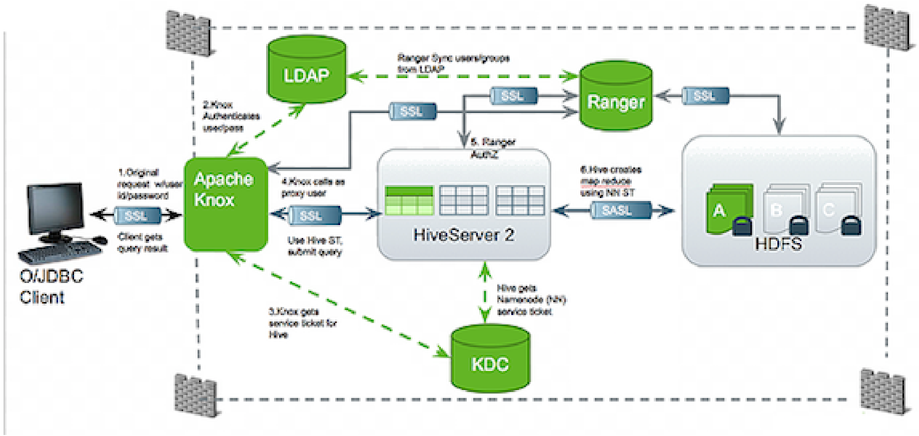
A conceptual overview of the Apache Knox Gateway, a reverse proxy.

Overview

Knox integrates with Identity Management and SSO systems used in enterprises and allows identity from these systems be used for access to Hadoop clusters.

Knox Gateways provides security for multiple Hadoop clusters, with these advantages:

- Simplifies access: Extends Hadoop’s REST/HTTP services by encapsulating Kerberos to within the Cluster.
- Enhances security: Exposes Hadoop’s REST/HTTP services without revealing network details, providing SSL out of the box.
- Centralized control: Enforces REST API security centrally, routing requests to multiple Hadoop clusters.
- Enterprise integration: Supports LDAP, Active Directory, SSO, SAML and other authentication systems.


LDAP
Ranger Sync users/groups from LDAP
Ranger
SSL
SSL
SSL
2.Knox Authenticates user/pass
5. Ranger AuthZ
1.Original request w/user id/password
Apache Knox
4.Knox calls as proxy user
SSL
SSL
Client gets query result
Use Hive ST, submit query
HiveServer 2
6.Hive creates map reduce using NN ST
SASL
A
B
C
HDFS
O/JDBC Client
3.Knox gets service ticket for Hive
Hive gets Namenode (NN) service ticket
KDC

Typical Security Flow: Firewall, Routed Through Knox Gateway

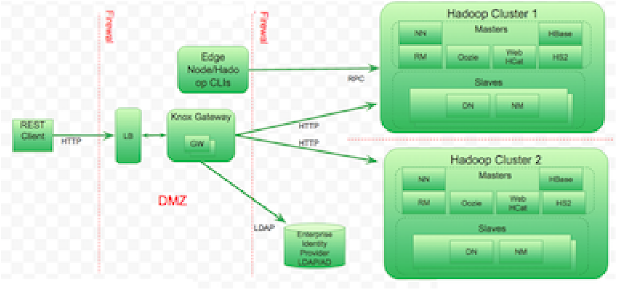
Knox can be used with both unsecured Hadoop clusters, and Kerberos secured clusters. In an enterprise solution that employs Kerberos secured clusters, the Apache Knox Gateway provides an enterprise security solution that:

- Integrates well with enterprise identity management solutions
- Protects the details of the Hadoop cluster deployment (hosts and ports are hidden from end users)
- Simplifies the number of services with which a client needs to interact

Knox Gateway Deployment Architecture

Users who access Hadoop externally do so either through Knox, via the Apache REST API, or through the Hadoop CLI tools.

The following diagram shows how Apache Knox fits into a Hadoop deployment.



NN=NameNode, RM=Resource Manager, DN=DataNote, NM=NodeManager

Knox Supported Services Matrix

A support matrix showing which services Apache Knox supports for Proxy and SSO, for both Kerberized and Non-Kerberized clusters.

Table 1: Knox Supported Components

Component	**SSO**	**Proxy (API)**	**Proxy (UI)**
Ambari	#	#	#
Ambari Metrics/Grafana			
Atlas	#	#	#
HBase		#[1]	
HDFS			#
Hive (via JDBC)		#	
Hive (via WebHCat)		#	
Livy		#	
Log Search	#		#
MapReduce2	#		#
Oozie	#	#	#
Ranger	#[2]	#	#[3]
SmartSense			
Spark 2/Spark History Server	#		#

[1] Stargate
[2] Admin Console
[3] Admin Console

Component	SSO	Proxy (API)	Proxy (UI)
WebHCat/Templeton		#	
WebHDFS		#	
YARN	#	#	#
Zeppelin	#	#	#

Note:

APIs, UIs, and SSO in the Apache Knox project that are not listed above are considered Community Features.

Community Features are developed and tested by the Apache Knox community but are not officially supported by Hortonworks. These features are excluded for a variety of reasons, including insufficient reliability or incomplete test case coverage, declaration of non-production readiness by the community at large, and feature deviation from Hortonworks best practices. Do not use these features in your production environments.