Cloudera Runtime 7.0.2

CDP Security Overview

Date published: 2019-11-01
Date modified:

https://docs.cloudera.com/

Legal Notice

© Cloudera Inc. 2024. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

CDP user management system

CDP Management Console includes a user management system that allows you to integrate your identity provider and manage user access to CDP resources.

During the initial setup of a Cloudera Data Platform (CDP) subscription, Cloudera designates a user account as a CDP account administrator. A CDP account administrator has all privileges and can perform any task in CDP. Administrators can create other CDP administrators by assigning the PowerUser role to users. CDP administrators can also register environments and create Data Lake clusters.

CDP administrators can create users and groups and then assign roles and resource roles to users or groups. The CDP Management Console also enables CDP administrators to federate access to CDP by configuring an external identity provider. CDP users can include users corresponding to an actual living person within the organization or machine users.

In addition to the SSO credentials mentioned above, CDP uses another set of credentials that must be used for accessing some CDP components (for example accessing Data Hub clusters via SSH).

To access to the CDP CLI or SDK, each user must have an API access key and private key. Each user must generate this key pair using the Management Console, and CDP creates a credentials file based on the API access key. When you use the CDP CLI or SDK, CDP uses the credentials file to get the cluster connection information and verify your authorization.

Data Lake security

Data Lake security and governance is managed by a shared set of services referred to as a Data Lake cluster.

Data Lake cluster services

A Data Lake cluster is managed by Cloudera Manager, and includes the following services:

- Hive MetaStore (HMS) -- table metadata
- Apache Ranger -- fine-grained authorization policies, auditing
- Apache Atlas -- metadata management and governance: lineage, analytics, attributes
- Apache Knox:
  - Authenticating Proxy for Web UIs and HTTP APIs -- SSO
  - IDBroker -- identity federation; cloud credentials

Currently there is one Data Lake cluster for each CDP environment. Security in all DataHub clusters created in a Data Lake is managed by these shared security and governance services.

Links to the Atlas and Ranger web UIs are provided on each DataLake home page. A link to the Data Lake cluster Cloudera Manager instance provides access to Data Lake cluster settings.

Apache Ranger

Apache Ranger manages access control through a user interface that ensures consistent policy administration across Data Lake components and DataHub clusters.

Security administrators can define security policies at the database, table, column, and file levels, and can administer permissions for groups or individual users. Rules based on dynamic conditions such as time or geolocation can also be added to an existing policy rule. Ranger security zones enable you to organize service resources into multiple security zones.

Ranger also provides a centralized framework for collecting access audit history and reporting data, including filtering on various parameters.

Note: Authorization through Apache Ranger is just one element of a secure production cluster: Cloudera supports Ranger only when it runs on a cluster where Kerberos is enabled to authenticate users.

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 run 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.

Apache Atlas

Apache Atlas provides a set of metadata management and governance services that enable you to manage data lake and DataHub cluster assets.

- Search and Proscriptive Lineage – facilitates pre-defined and ad hoc exploration of data and metadata, while maintaining a history of data sources and how specific data was generated.
- Ranger plugin for metadata-driven data access control.
- Flexible modeling of both business and operational data.
- Data Classification – helps you understand the nature of the data within Hadoop and classify it based on external and internal sources.

Apache Knox

Knox SSO provides web UI SSO (Single Sign-on) capabilities to Data Lakes and associated environments. Knox SSO enables users to log in once and gain access to Data Lake and DataHub cluster resources.

Knox IDBroker is an identity federation solution that provides temporary cloud credentials in exchange for various tokens or authentication.

CDP identity management

CDP Identity Management includes CDP user management system, Free IPA, identity federation, and Knox authentication.

FreeIPA identity management

Federating identity management with users/groups maintained in FreeIPA and passwords authenticated via SSO to an SAML-compliant identity provider (IDP) provides the necessary backbone infrastructure needed for CDP services, without requiring you to expose your on-prem identity management system over the network.

What is FreeIPA?

FreeIPA is an open-source product that combines four identity management capabilities:

- LDAP directory: a common user directory so that all services in both the SDX and workload clusters can consistently resolve users.
- Kerberos KDC: a single common Kerberos realm so that services can authenticate each other, within and between clusters. Kerberos is also used as a user authentication mechanism by some services.
- DNS server: a relatively simple way to discover and reach shared services in an SDX cluster from various workloads.
- Certificate Authority (CA): some services secure communication channels with TLS, which means they need certificates. A shared CA allows CDP to establish a common trusted root for all connected workloads.

Identity management with FreeIPA

IPA is an identity management framework used to assert who a user is. A subset of users and groups are replicated into IPA (and propagated to the nodes via SSSD). Making the users and groups available on the nodes with consistent user names enables security policies to be migrated from on-prem to the cloud. Users and groups are imported from on-prem and principally managed from the Control Plane UMS (User Management System,) with IPA providing the backend propagation.

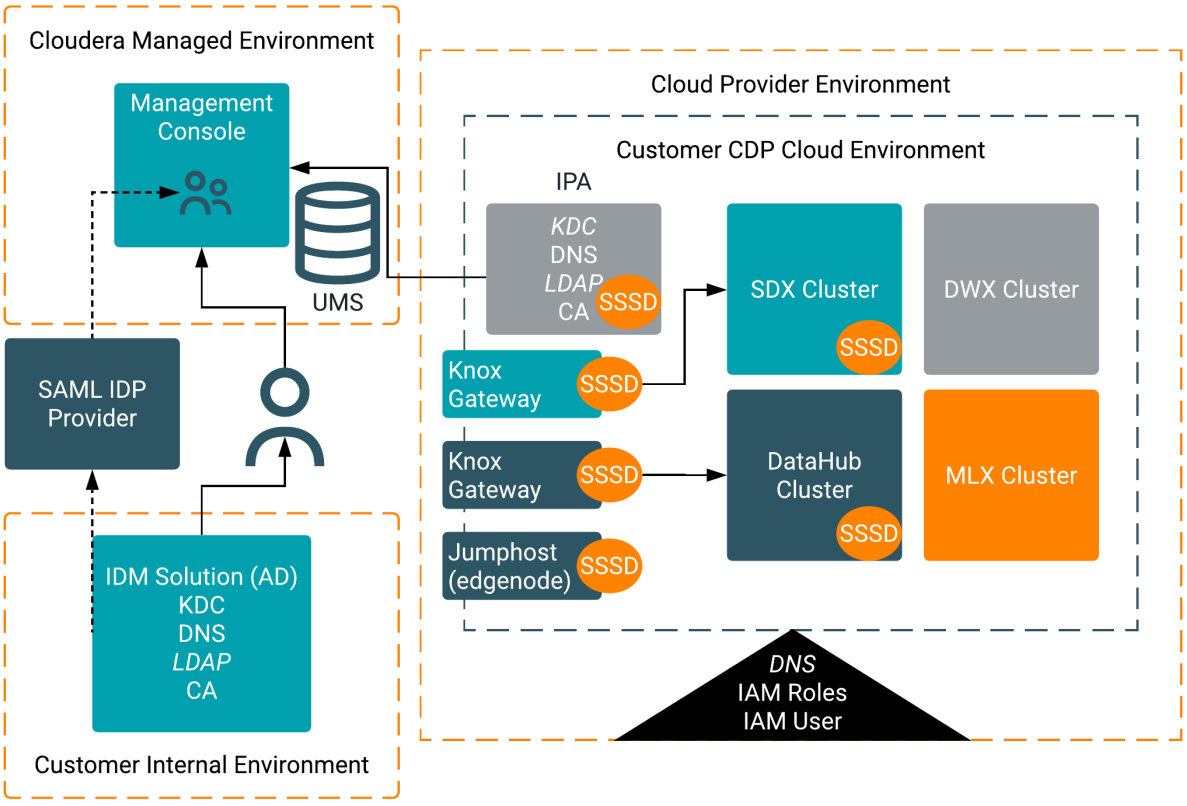
FreeIPA prerequisites

For IPA to work, you must have:

- An AD on-prem or a central LDAP where relationships between users and groups are maintained.
- A SAML identity provider (e.g., Okta or KeyCloak) that can be leveraged to authenticate users and import their groups.

How FreeIPA works

The following diagram illustrates how FreeIPA works:



Cloud identity federation

When accessing cloud storage in CDP, credentials are provided by Knox IDBroker, an identity federation solution that exchanges cluster authentication for temporary cloud credentials.

What is IDBroker?

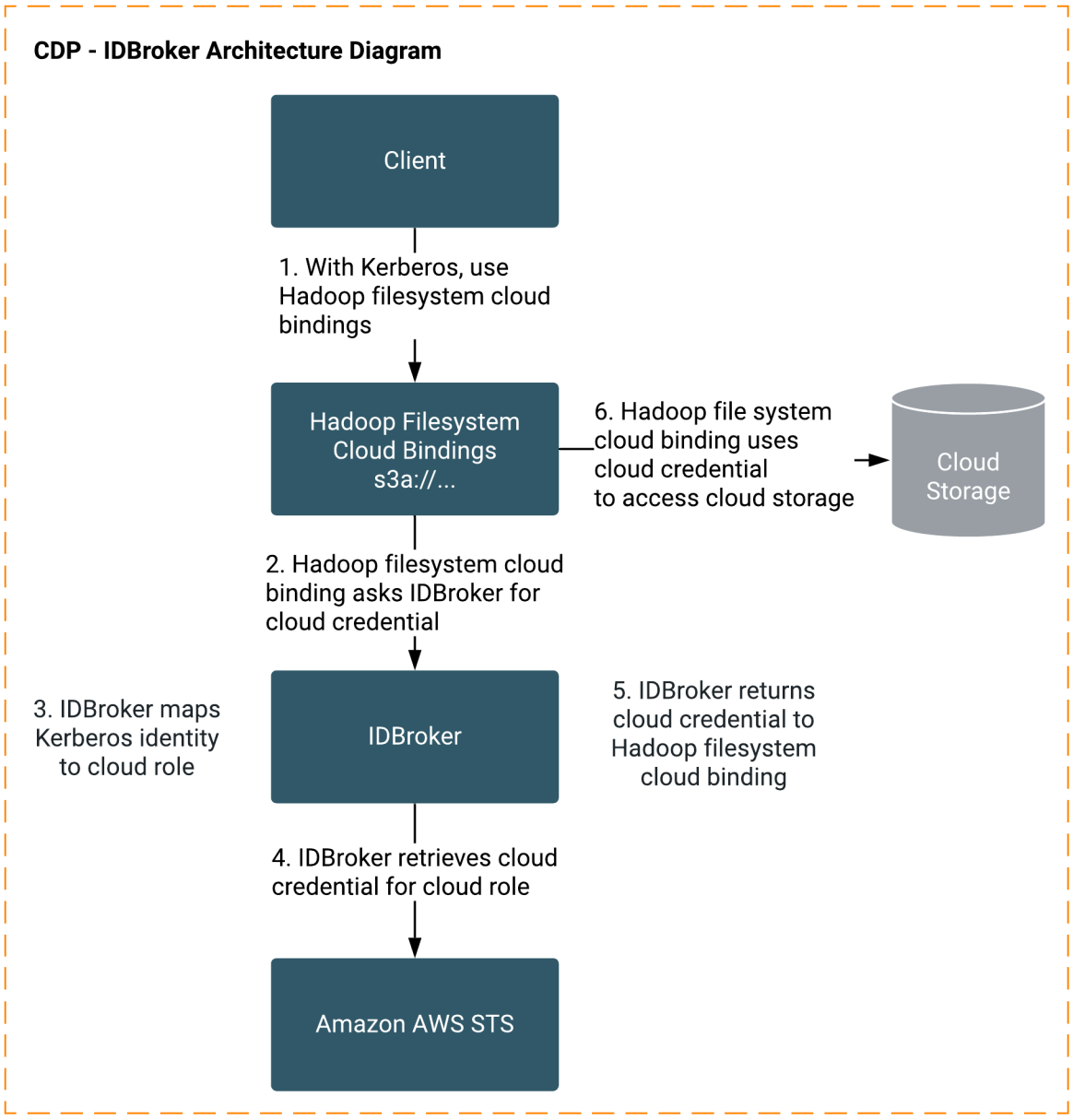
IDBroker is a REST API built as part of Apache Knox’s authentication services. It allows an authenticated and authorized user to exchange a set of credentials or a token for cloud vendor access tokens.

IDBroker is automatically configured by Cloudera Manager in CDP deployments, where Knox is installed. Cloud roles can be mapped to users or groups that exist in UMS to help control authorization.

How IDBroker works

Object store data access permissions are managed within the native cloud provider’s authorisation systems, not within CDP. So, when Knox IDBroker considers a particular user, it doesn't automatically have credentials for them to access the requested data. IDBroker addresses this problem by giving us a way to map UMS users and groups to cloud provider identities (roles) with associated permissions. Then, when data access operations happen via a Hadoop storage connector (eg: s3a), the connector will connect to IDBroker and ask for short lived access credentials which can be passed to the cloud provider when making data access calls. IDBroker will use the mappings to decide which credentials to obtain and return.

Knox IDBroker can help bridge the identity gap by requesting short term tokens from the cloud provider on behalf of a user that has successfully authenticated with Knox, and who is specified as being a part of a group mapped to a role that can access the bucket.



User and group mapping

IDBroker creates mappings between CDP users and groups (imported from corporate LDAP/AD, stored in UMS) and native cloud platform roles (e.g. in AWS: the IAM roles associated with policies). The mappings are specified in the Control Plane and synchronised to deployed clusters.

IDBroker authentication delegation tokens lifetime

By default, IDBroker authentication delegation tokens, used to request cloud credentials, have a lifetime of 7 days. This lifetime can be adjusted by modifying the idbroker_knox_token_ttl_ms configuration property.

Authentication with Apache Knox

Apache Knox handles proxy for web UIs and APIs, and Trusted Proxy propagates the authenticated end user to the backend service.

Knox Gateway

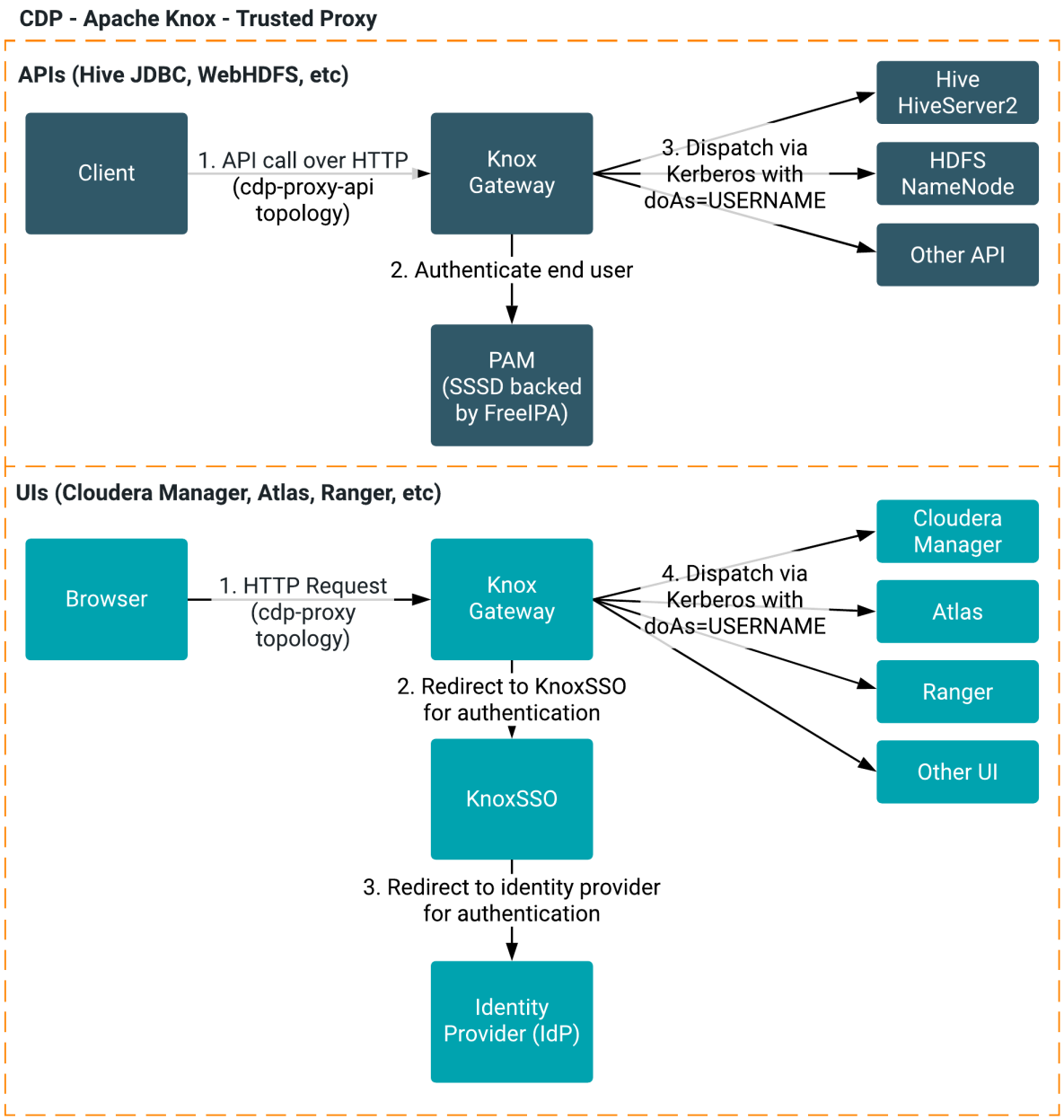
Apache Knox Gateway is a reverse proxy that authenticates and provides a single access point for REST and HTTP interactions with the CDP Data Hub clusters.

Trusted proxy

Knox Trusted Proxy is useful in cloud deployments when you need the seamless and uniform authentication benefits of both proxy and SSO. Trusted Proxy is automatically configured by Cloudera Manager in CDP deployments.

Knox Trusted Proxy propagates the authenticated end user to the backend service. The request is "trusted" in that the given backend/service is able to validate that the request came from a certain place and was allowed to make the request. A backend in this case is any service that Knox is acting as a proxy for (e.g., Cloudera Manager, Hive JDBC, Ranger UI, etc). Each of these services have a mechanism to ensure that the 1) request IP address and 2) request user matches what it expects. If the request matches those two things, then the service will not have to authenticate again and can trust that Knox sent the request.

When making requests to the cluster, Knox first authenticates the end user, and then adds that user as a query parameter to the request (?doAs=USERNAME) to the backend. The backend then checks that the request is trusted (request IP and request user) and extracts the end user (USERNAME) from the query parameter. The backend service then does whatever is necessary as that backend user. Knox and the proxied services authenticate to each other via Kerberos.



SSO via Knox

Knox SSO is configured automatically.

TLS encryption using auto-TLS

Auto-TLS greatly simplifies the process of enabling and managing TLS encryption on your cluster.

Auto-TLS automates the creation of an internal certificate authority (CA) and deployment of certificates across all cluster hosts. It can also automate the distribution of existing certificates, such as those signed by a public CA. Adding new cluster hosts or services to a cluster with auto-TLS enabled automatically creates and deploys the required certificates.

In CDP, auto-TLS is enabled by default.

When TLS is enabled for the Cloudera Manager Admin Console, web requests now include the Strict-Transport-Security header. For more details about this header, see "Strict-Transport-Security (Mozilla)".

Related Information

Strict-Transport-Security (Mozilla)

Secure in-bound communication

The CDP Control Plane reaches out to workload environments for various command and control purposes. These connections currently go over the Internet to the workload environment hosts. As a consequence, CDP currently deploys workloads into public (Internet routable) subnets.

To operate in public subnets, CDP secures in-bound communication to the listening ports using TLS encryption. Connections are authenticated using environment- or cluster-specific credentials that CDP manages internally. As appropriate, administrators can further secure listening ports by authenticating IP addresses and/or ports:

- Identifying authentic connections: Use security group rules to ensure in-bound connections originate from the set of stable IP addresses that belong to your organization.
- Identifying authentic ports: In-bound connections take place on an identifiable set of ports, which can be used to additionally narrow the allowlist.

In addition to the CDP management traffic, your organization's use cases may specify connecting to the workload hosts over the Internet. These communications may involve different endpoints, protocols, ports, and credentials than CDP management traffic.

Endpoints fall into four categories today, associated with our shared Environment services on the Data Lake and on Data Hub, Data Warehouse, and Machine Learning workloads. The following sections enumerate the communications for each of those categories.

Data Lake communication endpoints

In-bound communication for shared services support CDP management of FreeIPA and Data Lake services.

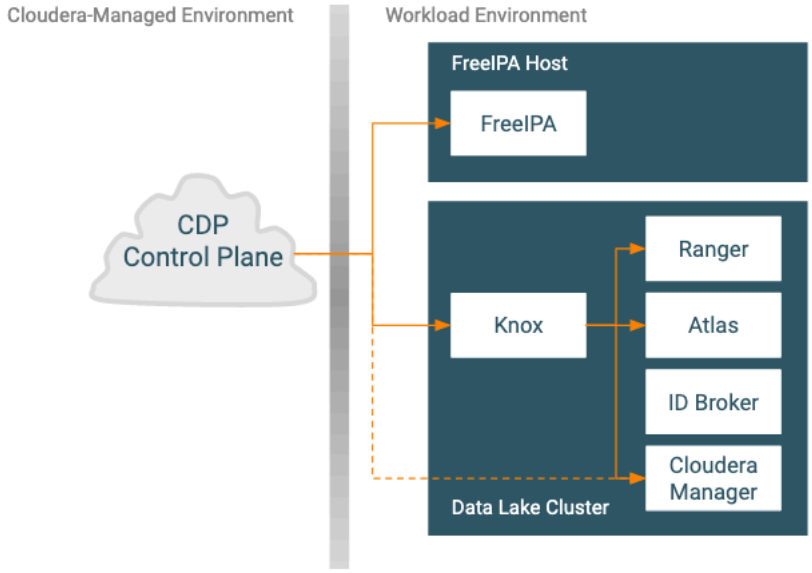
Each Environment contains a deployment of FreeIPA and a Data Lake cluster that support the following security and governance activities:

Free IPA identity and security services

- LDAP: User Directory
- Kerberos KDC: Manages a Kerberos Realm for the Environment
- DNS: Provides internal resolution of workload hostnames
- Certificate Authority: Issues TLS certificates for internal use, and for the in-bound CDP control connections

Data Lake cluster services

- Hive Metastore: Tabular metadata storage
- Ranger: Security policies and audit trail
- Atlas: Data lineage, tagging, analytics
- ID Broker: Mapping of CDP identities to cloud provider identities
- Knox: A proxy gateway for access to cluster services
- Cloudera Manager: Local management for the data lake services



Communication to the CDP Control Plane includes the following:

Free IPA identity and security services

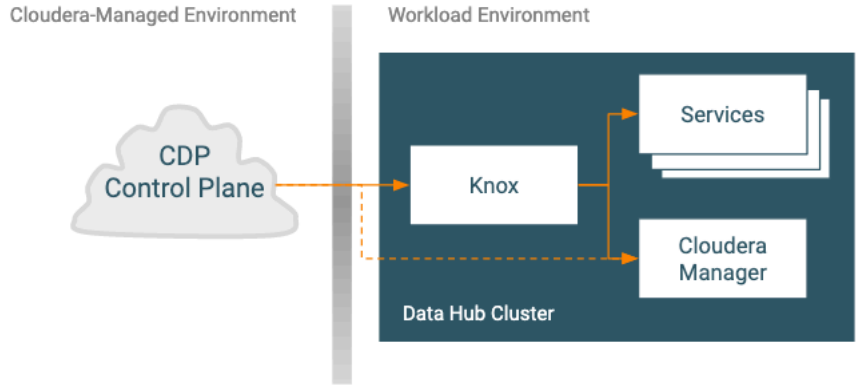
- User and Group synchronization
- Service Principal management
- Retrieving the CA root certificate

Data Lake cluster services

- Core lifecycle management operations (sent directly between CDP and Cloudera Manager to include the Knox proxy as one of the managed entities; shown as a dashed line in the picture)
- Communication (via Knox proxy)
  - General Cloudera Manager operations
  - Ranger operations to manage repositories for workloads
  - ID Broker mappings (updated via Cloudera Manager)
  - Data Catalog communicates with Atlas and Ranger to surface information

Data Hub communication endpoints

Data Hub clusters are built on the same underlying technology as the Data Lake cluster and so present a similar connectivity profile.



Communication to the CDP Control Plane includes the following:

Free IPA identity and security services

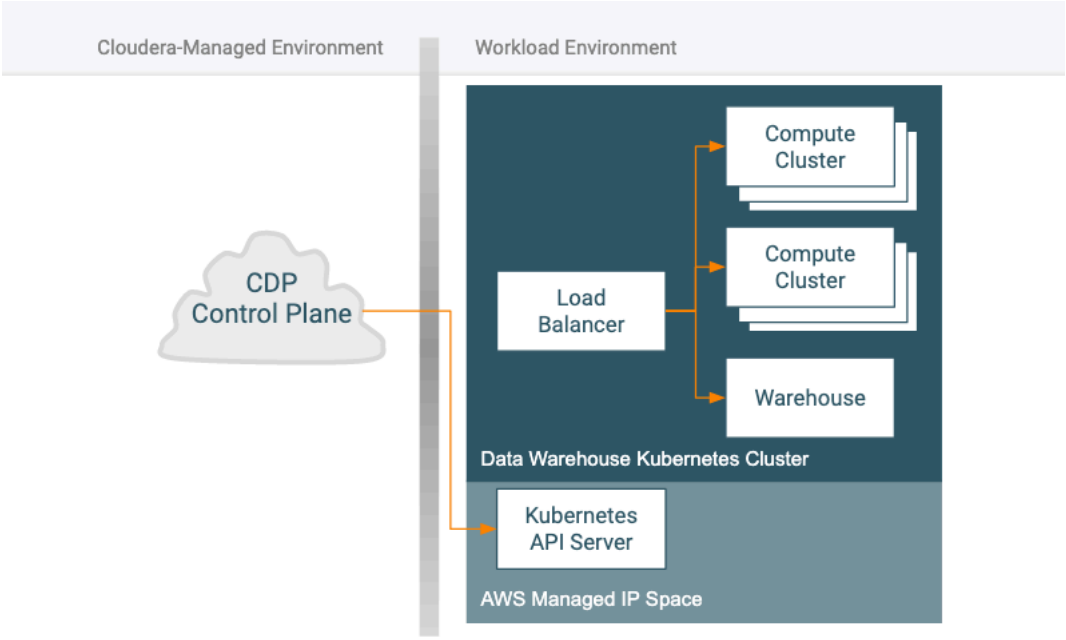
- User and Group synchronization
- Service Principal management
- Retrieving the CA root certificate

Data Lake cluster services

- Core lifecycle management operations (sent directly between CDP and Cloudera Manager because the Knox proxy is one of the managed entities; shown as a dashed line in the previous picture)
- Communication (via Knox proxy)
  - General Cloudera Manager operations
  - Service-specific communication depending on the specific DataHub cluster in question

Data Warehouse communication endpoints

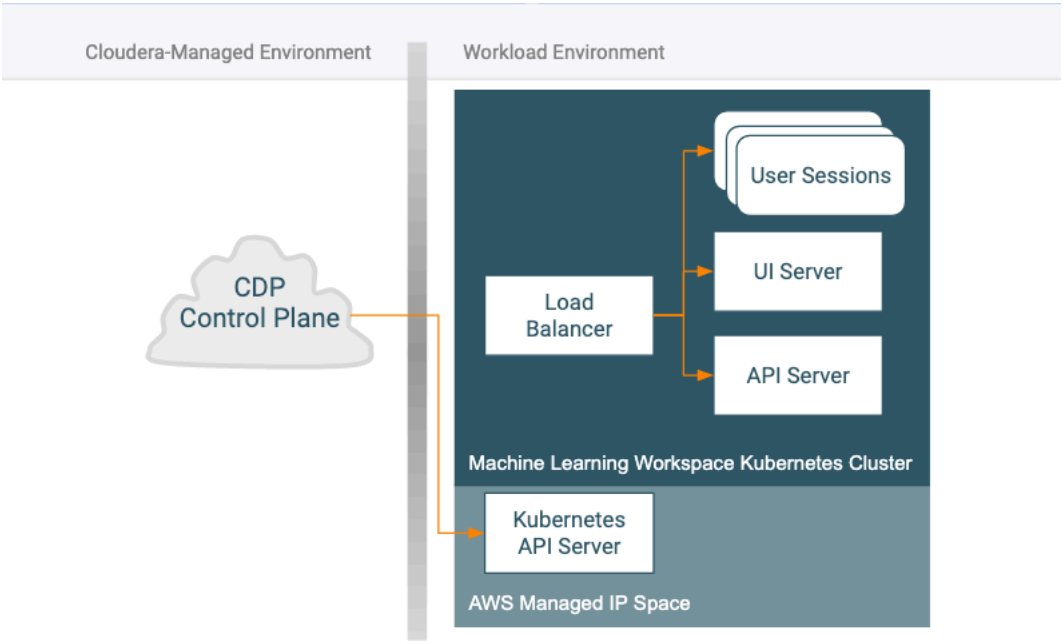
The Data Warehouse service operates significantly differently from Data Hub, as it runs on top of a Kubernetes cluster and does not include a Cloudera Manager instance.



Primary command and control communication goes to the Kubernetes API server. This endpoint is specific to a particular Kubernetes cluster, but it is provisioned by the cloud provider outside of the customer VPC. Whether or not it is Internet-facing is independent of the VPC configuration. The Data Warehouse service doesn't make connections to endpoints in the cluster.

Machine Learning communication endpoints

In terms of communication, a Machine Learning Workspace looks very similar to a Data Warehouse workspace in that it is also a Kubernetes cluster, although the contents differ.



Primary command and control communication goes to the Kubernetes API server. This endpoint is specific to a particular Kubernetes cluster, but it is provisioned by the cloud provider outside of the customer VPC. Whether or not it is Internet-facing is independent of the VPC configuration. The Machine Learning service doesn't make connections to endpoints in the cluster.

Security terminology

The following terminology is key to understanding CDP security:

Table 1: Management Console terminology

Term	Description
Credential	A credential allows CDP to authenticate with your cloud provider account and obtain authorization to provision cloud provider resources on your behalf.
Environment	In CDP, an environment is a logical subset of your cloud provider account including a specific virtual network. A credential provides CDP with access to an environment.
Data Lake	Data Lake is a service for creating safe, secure, and governed Data Lake which provides a protective ring around the data stored in a cloud object store.
Virtual network	An environment corresponds to a single private virtual network (for example VPC on AWS) into which all your CDP resources are deployed.
Security access settings	Security access settings refer to security groups that are created on your cloud provider account to allow communication via specific ports.

Table 2: Knox terminology

Term	Description
FreeIPA	FreeIPA is an open-source product that combines four identity management capabilities: LDAP directory, Kerberos KDC, DNS server, and Certificate Authority (CA).
IDBroker	An identity federation solution that exchanges cluster authentication for temporary cloud credentials.
Knox Gateway	A reverse proxy that authenticates proxy for web UIs and HTTP APIs.

Term	Description
Trusted Proxy	Propagates the authenticated end user to the backend service.