

CDP Security Overview

Date published: 2019-08-22

Date modified:

CLOUDEXERA

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 security FAQs.....	4
CDP identity management.....	5
CDP user management system.....	5
FreeIPA identity management.....	6
Cloud identity federation.....	7
Authentication with Apache Knox.....	9
Access to customer resources.....	10
Handling of sensitive data in CDP.....	11
User access to clusters.....	11
Secure inbound communication.....	11
Cloudera Private Link Network Overview.....	15
Data Lake security.....	16

CDP security FAQs

This topic covers frequently asked questions related to the communication between the CDP Control Plane and workload subnets.

Data security within workload subnets and virtual networks

CDP provides comprehensive security features to ensure that minimal network configuration is needed on workload clusters and that only metrics, logs, and control signals go in and out of the customer's network. No data from the workload clusters is ever accessed by the CDP Control Plane.

To operate at the highest level of security, Cloudera recommends running CDP workload clusters in private subnets. This means that nodes in the workload clusters do not have public IP addresses and all outbound traffic to the internet goes through a gateway or a firewall. This allows your security operations team to ensure that hosts that are allowed to communicate with the clusters are legitimate and pose no security threat to the assets. A common best practice is to avoid inbound connections of any sort directly into the workload subnet.

Cloudera recommends that no inbound connections be allowed into the private network and that you use your sec-ops approved method(s) to provide outbound access to the list of hosts specified in the [AWS Azure GCP](#) outbound network access destinations.

Communication between CDP and workload subnets

Customers use CDP Management Console to operate and manage workload clusters running in their own VPCs. For every operation performed through the Management Console on the workload clusters, a control signal is sent to the hosts in the private network. This is achieved through a feature called Cluster Connectivity Manager (CCM).

CCM eliminates the need to configure inbound connections into your secure private network. CCM, which is set up during cluster creation, configures a reverse tunnel from the customer's private network to the CDP Control Plane. All control signals to create/delete clusters, stop/start environments and other management actions related to workload clusters go through the CCM tunnel. CCM allows customers to avoid configuring any inbound connections/routes to their private workload subnets, thus providing a better security posture for their public cloud assets.

More information on CCM is available in the [Cluster Connectivity Manager](#) documentation.

Cloudera Private Link Network for additional privacy and security

Users who are concerned with privacy can utilize the Cloudera Private Link Network to establish private and secure connections from their workloads to the Cloudera Control Plane without using the public internet.

Cloudera Private Link Network is designed to provide seamless, private connectivity between your cloud workloads and the Cloudera Control Plane. For further information, see [Cloudera Private Link Network Overview](#).

Egress network setup for better control and visibility on outbound traffic

Large security-conscious enterprises typically inspect all traffic going in and out of their private networks. This is generally achieved in the following way:

1. Identify a single egress virtual network that is used to provide internet access to all other subnets and virtual networks. Route outbound (Internet) traffic from all other subnets to this egress network.
2. Purpose-built technologies, such as web proxies, next-generation firewalls and cloud access security brokers are deployed in the egress VPC to monitor for anomalous outbound behavior. Network analyzers and forensic recorders can also be used.

Cloudera recommends a similar topology to configure outbound traffic rules that are needed for normal operation of the clusters in a private network. The alternative is to set up egress rules within the same VPC that hosts the private subnet.

Information exchanged between CDP and workload subnets

Irrespective of the security posture used by the customer to secure their assets in the workload subnet, a minimum set of communications is needed for normal operations of a CDP environment. The following set of points summarize the data exchanged between CDP Control Plane and workload subnets:

1. All user actions on the CDP Control Plane that are related to interacting with the workload clusters in the workload subnets happen via CCM.
2. Metering and billing information is sent at regular intervals from the customer's workload subnet to the Cloudera Databus as specified in the [AWS Azure GCP](#) outbound network access destinations.
3. If used, an optional component called Cloudera Observability also leverages the Cloudera Databus to communicate with the Control Plane.
4. Diagnostic bundles for troubleshooting are generally sent by customers to engage in support cases. The diagnostic bundles can either be sent on demand or on a scheduled basis to Cloudera Customer Support. This feature also uses the Cloudera Databus channel. Refer to [Sending Usage and Diagnostic Data to Cloudera](#) page to read more about diagnostic bundles.
5. Customers generally share logs with Customer Support in a self-service fashion through CDP Control Plane. No customer data is ever shared in the diagnostics bundle. To ensure sensitive data does not show up inadvertently in the logs, we recommend that customers follow the directions specified in [Redaction of Sensitive Information from Diagnostic Bundles](#).

CDP identity management

CDP Identity Management includes the CDP user management system and Knox authentication.

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.

For more information, refer to the following documentation:

Related Information

[User Management](#)

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.

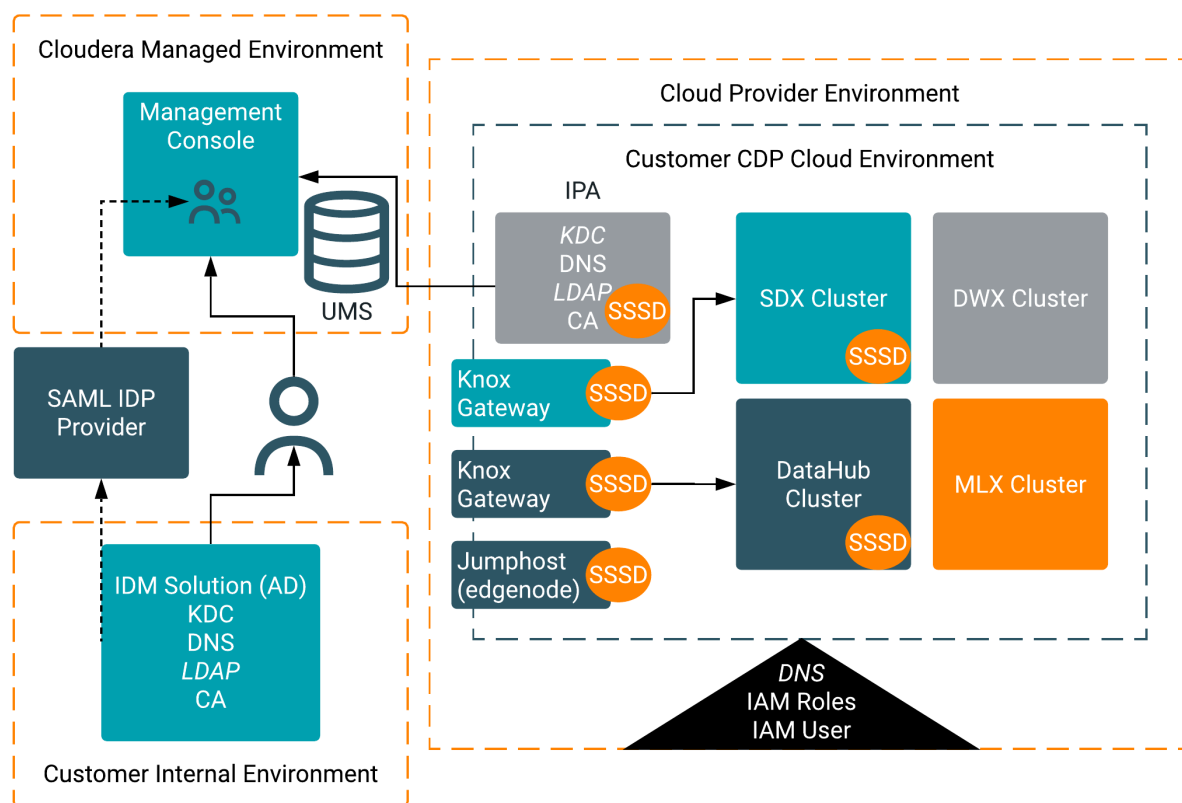
FreeIPA prerequisites

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?

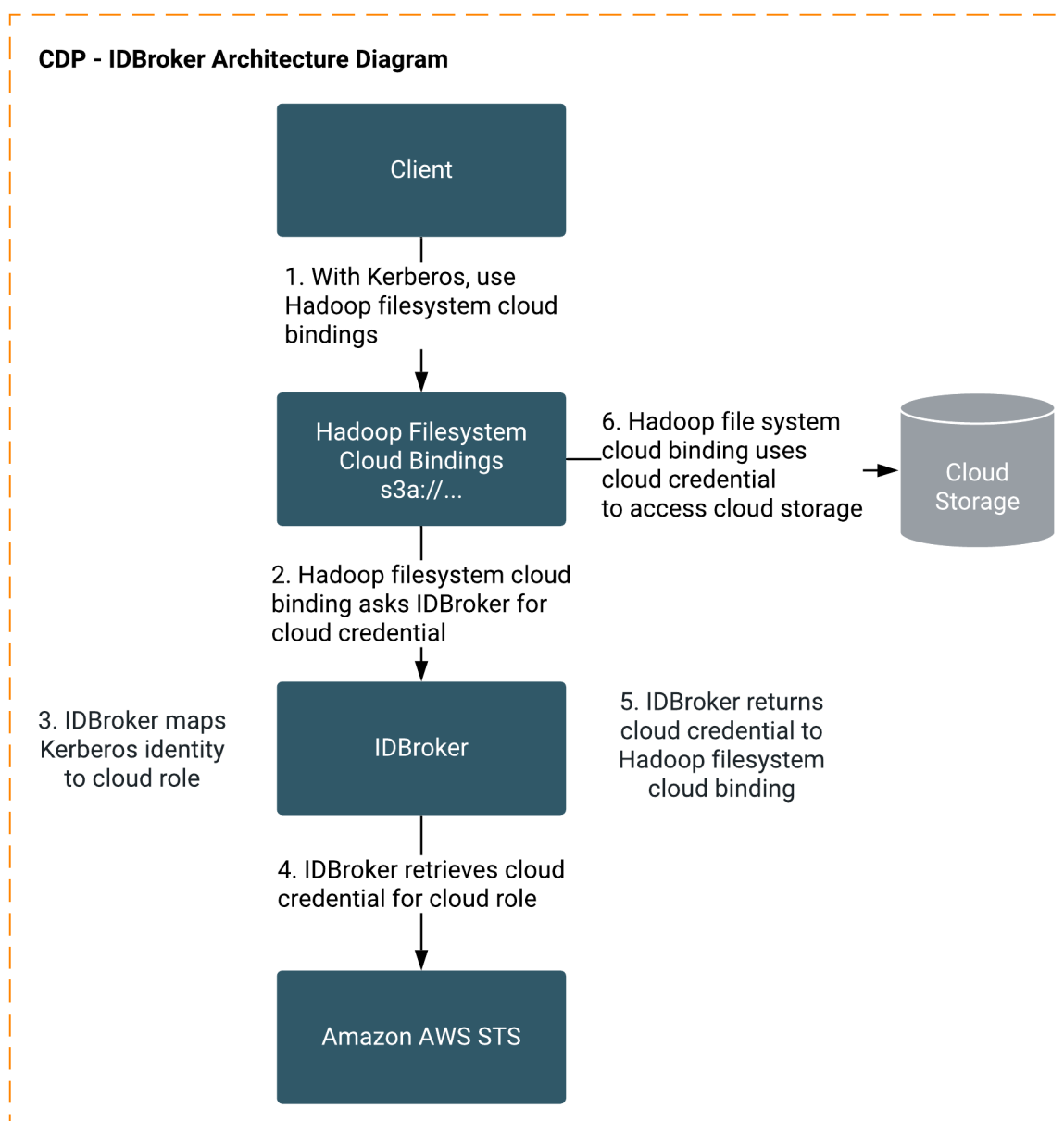
IDBroker is a REST API built as part of Apache Knox's authentication services. It allows an authenticated and authorized user to exchange a set of credentials or a token for cloud vendor access tokens.

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

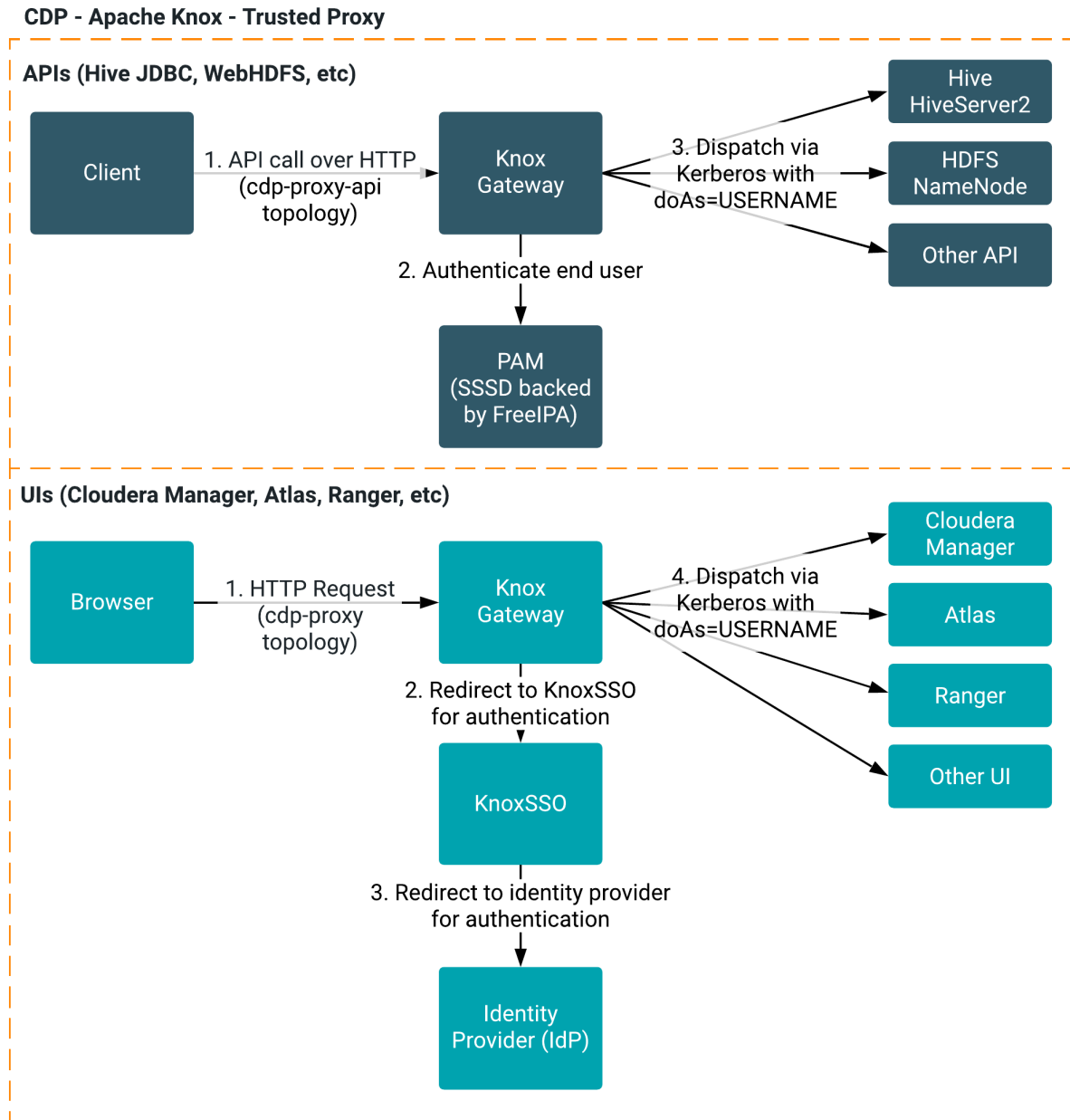
Apache Knox Gateway is a reverse proxy that authenticates and provides a single access point for REST and HTTP interactions with the CDP Data Hub clusters.

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.

Related Information

[Knox Supported Services Matrix](#)

Access from CDP to customer resources

CDP creates clusters and runs jobs in your cloud provider account on your behalf.

CDP requires your permission to be able to use the resources required by the clusters and jobs in your cloud provider account. To allow CDP to create clusters or run jobs in your cloud provider account:

- Your AWS administrator must create a cross-account access IAM role and grant CDP access to the role as a trusted principal. The policy defined for the cross-account access IAM role must include permissions to allow CDP to create and manage resources and to perform the tasks and access the resources required for the CDP clusters and jobs.
- Your Azure account administrator must create an app registration and assign a role to it with permissions allowing CDP to create and manage resources and to perform the tasks and access the resources required for the CDP clusters and jobs.
- Your GCP account administrator must create a service account and assign permissions allowing CDP to create and manage resources and to perform the tasks and access the resources required for the CDP clusters and jobs.

For more information about credentials and security groups, refer to the following documentation:

Related Information

[Role-based credential on AWS](#)

[App-based credential on Azure](#)

[Provisioning credential for Google Cloud](#)

[Default security group settings](#)

Handling of sensitive data in CDP

CDP uses [Vault](#) to encrypt sensitive data (such as tokens, passwords, certificates, and encryption keys) .

Classic cluster credentials

During HDP, CDH, or CDP PvC Base "classic cluster" registration in the Management Console, CDP asks you to enter cluster credentials. This is required for authenticating requests on the cluster side. CDP uses these credentials to access cluster APIs during and after cluster registration. During registration CDP stores these credentials securely in the vault and later whenever CDP makes an API call to the cluster, CDP reads these credentials from the vault and inject them inside the request.

Related Information

[Enabling CCM in the Management Console](#)

[Upgrading a classic cluster from CCMv1 to CCMv2](#)

User access to clusters

Access to Cloudera Manager and other cluster UIs and endpoints (such as JDBC) is always via the secure Knox gateway on port 443. Users are automatically logged in with their CDP as a Service credentials.

Secure inbound communication

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

To operate in public subnets, CDP secures inbound 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 ports:

- Identifying authentic connections: Use security group rules to ensure inbound connections originate from the set of stable IP addresses that belong to your organization.
- Identifying authentic ports: Inbound 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

Inbound communication for shared services support CDP management of FreeIPA and Data Lake services.

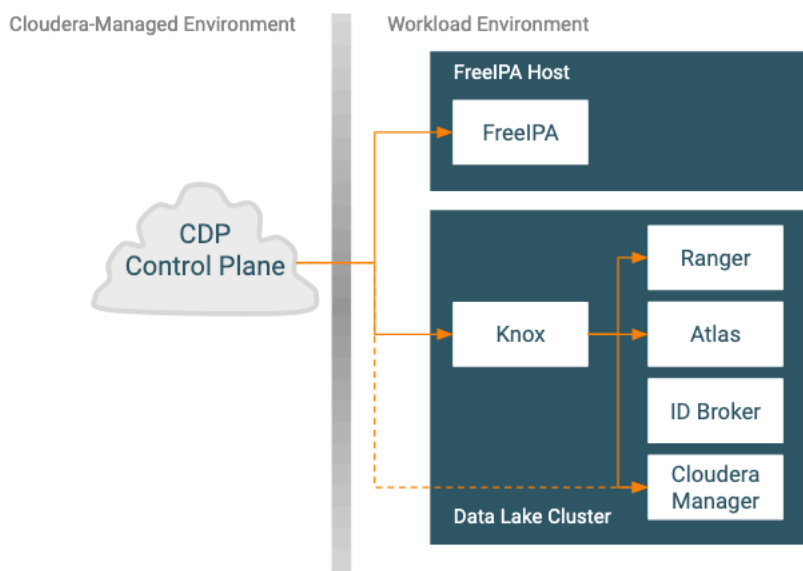
Each Environment contains a deployment of FreeIPA and a Data Lake cluster that support the following security and governance activities:

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 inbound 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

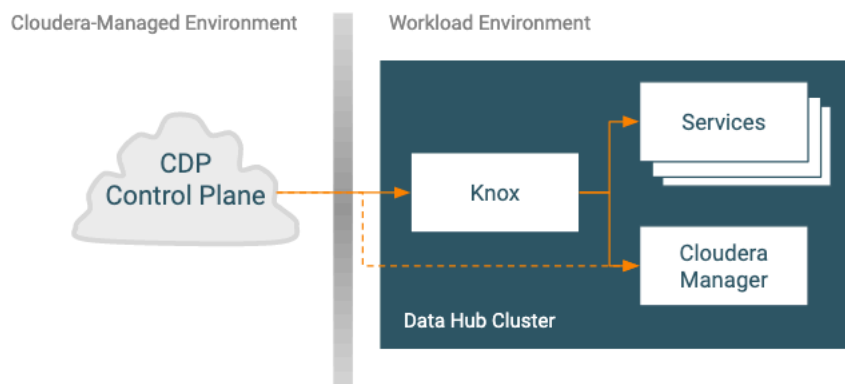
- User and Group synchronization
- Service Principal management
- Retrieving the CA root certificate

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

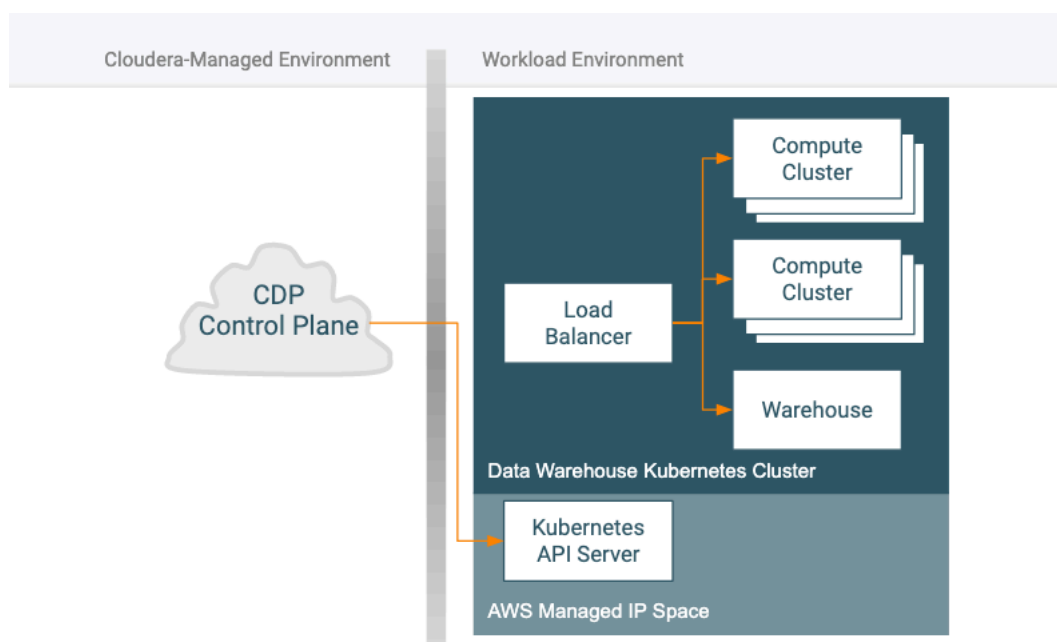
- User and Group synchronization
- Service Principal management
- Retrieving the CA root certificate

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

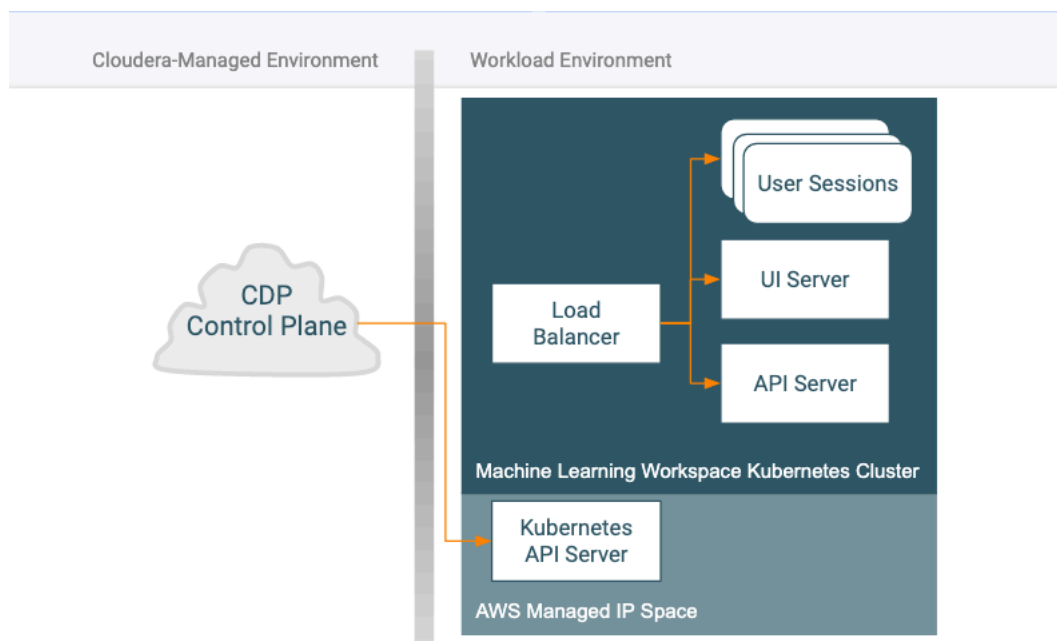
The Data Warehouse service operates significantly differently from Data Hub, as it runs on top of a Kubernetes cluster and does not include a Cloudera Manager instance.



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.



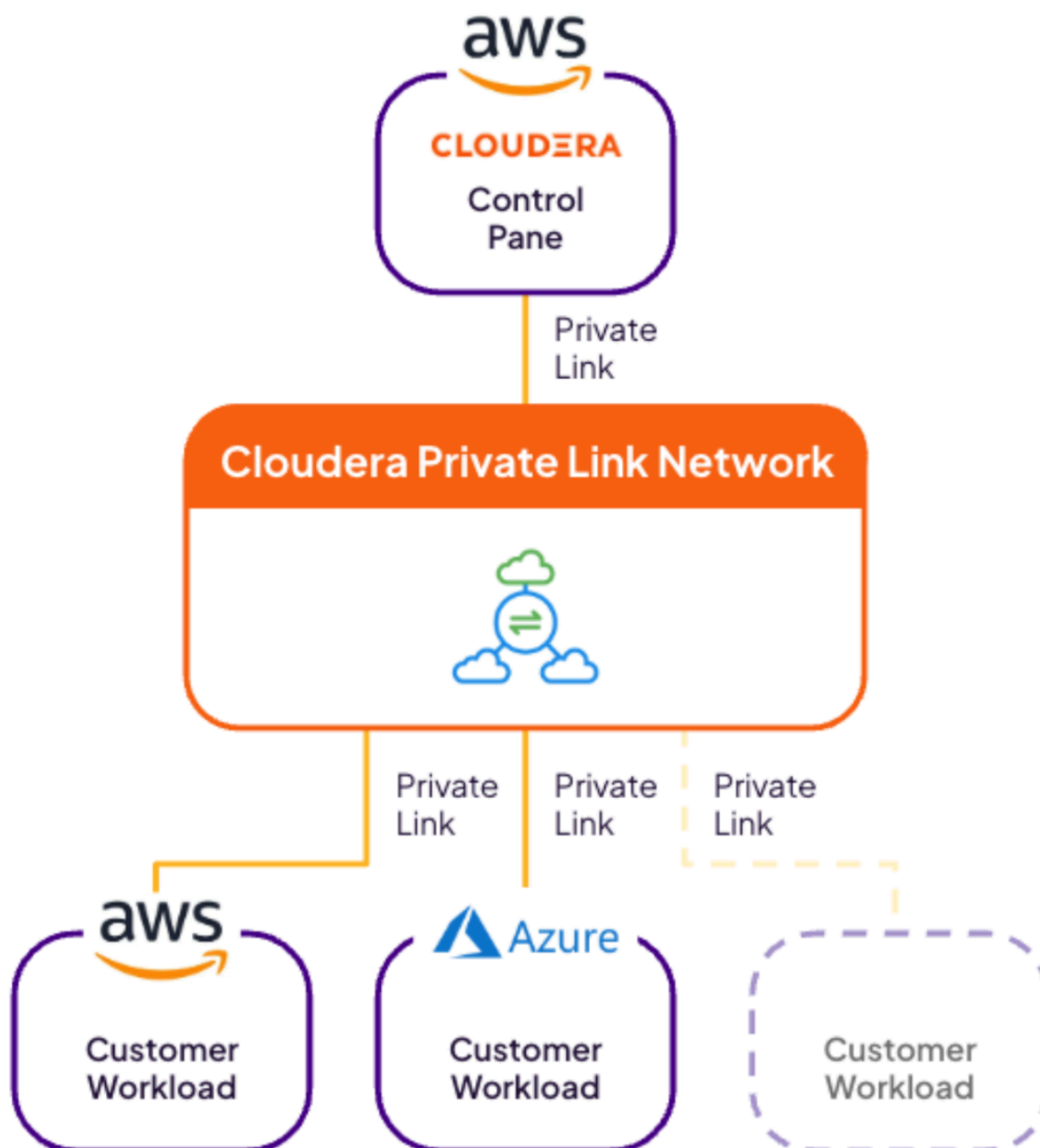
Primary command and control communication goes to the Kubernetes API server. This endpoint is specific to a particular Kubernetes cluster, but it is provisioned by the cloud provider outside of the customer VPC. Whether

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.

Cloudera Private Link Network Overview

Users who are concerned with privacy can utilize the Cloudera Private Link Network to establish private and secure connections from their workloads to the Cloudera Control Plane without using the public internet.

Cloudera Private Link Network is designed to provide seamless, private connectivity between your cloud workloads and the Cloudera Control Plane.



For more information, see [Cloudera Private Link Network for AWS](#).

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.

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.

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.

Related Information

[Data Lakes](#)