CDP Private Cloud Data Services 1.5.3

CDP Private Cloud Experiences Security Overview

Date published: 2023-12-16
Date modified: 2024-03-05

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 4

Handling of sensitive data in CDP 4

Secure in-bound communication 4

Data Lake Security 5

TLS Connections in CDP Private Cloud 5

- CDW 7
- CML 8
- CDE 9
- Compute Service 9

No wildcard DNS/TLS setup 9

CDP user management

The Cloudera Data Platform (CDP) Private Cloud Management Console includes a user management system that allows you to integrate your LDAP identity provider and manage user access to CDP resources.

When CDP Private Cloud is installed, a CDP account administrator user is created. A CDP account administrator has all privileges and can perform any task in CDP. Administrators can create other administrators by assigning the EnvironmentAdmin role to users. CDP users with the EnvironmentAdmin role can also register environments and create Data Lake clusters.

The CDP Private Cloud Management Console also enables account administrators to federate access to CDP by configuring an external LDAP identity provider. CDP users can include users synched with an external LDAP identity provider, or machine users. Machine users can be assigned roles and resource roles, but cannot log in to the web console.

Handling of sensitive data in CDP

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

CDP Private cloud uses the Embedded/External Vault to store user credentials and some critical system data pertaining to the Control Plane that includes the machine user password used to create the control plane, the contents of the kubeconfig file which provides admin access to the Kubernetes Control Plane, and the LDAP credentials configured for the ECS Cluster.

The CDP Private Cloud installer can install Vault, but for OpenShift environments, typically this is a pre-existing customer-managed external Vault deployment.

- For more information on how to install an external HashiCorp Vault, see Install Vault.

  Vault install notes:

  - Supported Vault version: 1.4.0
  - Secrets engine: kv-v2
  - Auth type: kubernetes
- For more information on how to configure an external HashiCorp Vault for CDP Private Cloud, see External Vault Requirements.

Secure in-bound communication

CDP uses Vault to encrypt sensitive data (such as tokens, passwords, certificates, and encryption keys). The CDP Private Cloud installer can install Vault, but typically this is a pre-existing customer-managed Vault deployment.

Data Warehouse communication endpoints

The Data Warehouse service 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. The Data Warehouse service does not 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. The Machine Learning service does not make connections to endpoints in the cluster.

Data Lake Security

CDP Private Cloud security and governance are managed by Apache Ranger and Apache Atlas.

A Data Lake refers to the shared security and governance services in a CDP Private Cloud Base cluster linked to a CDP Private Cloud environment, and managed by Cloudera Manager. This set of shared services is also referred to as SDX (Shared Data eXperience).

Data Lake services

Data Lake services are managed by Cloudera Manager, and can include the following services:

- Hive MetaStore (HMS) -- table metadata
- Apache Ranger -- fine-grained authorization policies, auditing
- Apache Atlas -- metadata management and governance: lineage, analytics, attributes

Security in all workload clusters created in an environment is managed by these shared security and governance services.

Links to the Atlas and Ranger web UIs are provided on each environment home page. A link to the Cloudera Manager instance provides access to configuration settings.

Apache Ranger

Apache Ranger manages access control through a user interface that ensures consistent policy administration in CDP 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.

Apache Atlas

Apache Atlas provides a set of metadata management and governance services that enable you to manage CDP 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.

TLS Connections in CDP Private Cloud

TLS is used to secure data in-flight in CDP Private Cloud.

Network connections made in CDP Private Cloud fall into the following categories:

- Traffic between the Control Plane or Data Services and end users: All endpoints exposed by CDP Private Cloud take the form of Kubernetes ingresses and must go through the Ingress Controller and Gateway. The Ingress Controller is configured with a TLS certificate and guarantees that the traffic must use HTTPS, while the Gateway guarantees that the traffic must be authenticated.
- Traffic between the Control Plane and Data Services: These two reside in separate Kubernetes namespaces and are treated as external to each other. As such, connections from Data Services to the Control Plane go through the Ingress Controller, guaranteeing the use of HTTPS.
- Traffic between the Control Plane or Data Services and the Kubernetes API server: API calls to the Kubernetes API server always use HTTPS and are authenticated using the pod's service account tokens, thus being subject to RBAC.
- Traffic between the Control Plane or Data Services and the Base Cluster: The Base Cluster must have TLS enabled, meaning that all connections will be encrypted.

The Control Plane has a network policy which prevents traffic from the outside from reaching inside it unless it goes through the Ingress Controller and Gateway. This effectively ensures that all traffic going into the Control Plane must be encrypted with TLS and authenticated.

There is no analogous network policy for egress traffic, but all connections from the Control Plane and Data Services to the outside (such as pulling images from an external Docker registry) use TLS. CDP Private Cloud supports running in an air gap environment where connections to external networks are forbidden.

Traffic within the same cluster and namespace is trusted, and is not currently encrypted.

Base Cluster TLS Requirements

The Base cluster must have TLS enabled, either using manual TLS configuration or Auto-TLS.

The truststore configured in Cloudera Manager will be automatically imported into CDP Private Cloud during installation of the Control Plane, but afterwards can be managed separately from the Base Cluster's truststore. This truststore is imported so that the Control Plane and workloads can trust TLS connections to the Base cluster.

For the most seamless experience, ensure that this truststore trusts:

- The Cloudera Manager server certificate
- The LDAP server certificate
- The Postgres database server certificate of all Hive Metastores that expect to be used with Private Cloud

Instead of specifying individual certificates, it is recommended to import a root CA certificate that signed all of the above. CA certificates can be updated or rotated after Control Plane installation.

Unified Truststore

After Control Plane installation, management of trusted CA certificates can be done from the Administration page of the Control Plane. The set of trusted CA certificates forms a "unified truststore" which is then propagated to workload clusters.

The unified truststore contains certificates for:

- Cloudera Manager and base cluster services
- Control Plane
- LDAP
- Control Plane Database
- Docker Registry (if external)
- Vault (if external)

Again, it is not likely that separate certificates are required for each entry above. In most cases, an organization will only have one CA certificate signing the certificates for the above entities.

External Endpoints

At the time of ECS installation, an ingress certificate and private key must be provided. OpenShift-based Private Cloud installations use the OpenShift ingress certificate instead. This ingress certificate is used as the TLS server certificate for the ingress controller that serves the Control Plane and Data Services workload endpoints. It is recommended that the ingress certificate be signed by the organization's trusted Certificate Authority. The ingress certificate must be a wildcard certificate of the format *.apps.<domain.com> where "domain.com" is the domain of the ECS cluster.

If an ingress certificate is not provided during ECS installation, a self-signed certificate will be generated. This is recommended only for Proof-of-Concept (PoC) installations.

CDW and CDE workload clusters will also use the ingress certificate for their exposed endpoints.

CML workload clusters do not use the ingress certificate for their exposed endpoints. This is because CML requires multiple levels of subdomains underneath *.apps.<domain.com>, while the TLS standard prescribes that only a single level of subdomain be represented by a wildcard certificate. CML uses multiple levels of subdomains for isolation and web security.

The following sections describe in detail the TLS endpoints of different Data Services.

CDW

Connections to Control Plane

CDW workload clusters run in the same Kubernetes/Openshift cluster as the Control Plane, but in different namespaces. CDW workload clusters connect to Control Plane services (thunderhead-environment, thunderhead-usermanagement, thunderhead-kerberosmgmt-api, thunderhead-servicediscoverysimple, cluster-proxy) through the ingress controller, and are encrypted and authenticated.

Ingress Connections

Openshift route / Kubernetes ingress names are given below. The example hostnames are given in parentheses as if the cluster hostname was cluster.example.com.

- Hive VW (example hostnames are given as if the Hive VW name was hive1)
  - hiveserver2-route / hiveserver2-ingress (hs2-hive1.apps.cluster.example.com)
  - hue-route / hue-ingress (hue-hive1.apps.cluster.example.com)
  - das-route / das-ingress (hive1.apps.cluster.example.com; deprecated – disabled by default in PvC 1.4.1)
- Impala VW (example hostnames are given as if the Impala VW name was impala1)
  - coordinator-debug-route / coordinator-debug-ingress (coordinator-web-impala1.apps.cluster.example.com)
  - hue-route / hue-ingress (hue-impala1.apps.cluster.example.com)
  - impala-catalogd-route / catalogd-ingress (catalogd-web-impala1.apps.cluster.example.com)
  - impala-coordinator-route / coordinator-ingress (coordinator-impala1.apps.cluster.example.com)
  - statestored-route / statestored-ingress (statestored-web-impala1.apps.cluster.example.com)
- DataViz (example hostname is given as if the DataViz instance name was dataviz1)
  - viz-webapp-route / viz-webapp-ingress (viz-dataviz1.apps.cluster.example.com)

All Openshift route definitions have the setting spec.tls.insecureEdgeTerminationPolicy: Redirect so that Openshift automatically redirects HTTP requests to HTTPS. ECS ingress definitions are set up to handle HTTPS traffic, but they handle HTTP traffic as well and there is no automatic redirection (unlike to Openshift).

Egress Connections

- Base cluster: Connection settings inherit the SSL configuration settings from the base cluster and will use CM's truststore. You can specify additional certificates in the installation wizard and on the Management Console.
  - HDFS
  - Zookeeper
  - Ranger
  - Atlas
  - Kafka(in the Atlas hooks)
- LDAP : When specifying the settings on the Management Console / Authorization page, you have to specify an ldaps URL and upload the CA certificate of the LDAP server to make the connection secure.
- KDC : Encrypted by default, can be customized in the config file krb5.conf as it is inherited from the base cluster.
- Relational Databases
  - For DWX, Hue and HMS in a custom DBC : This is either the embedded Postgres DB instance that runs in the Control Plane in the Kubernetes/Openshift cluster or the external DB provided by the customer during the PvC installation. In either case it must be Postgres , and SSL is mandatory.
  - For HMS in the default DBC : This is the same relational DB instance that is used by HMS on the base cluster. Postgres, Mysql, MariaDB and Oracle are supported, SSL is mandatory.

CML

Connection	TLS (Yes/ No/ Optional)	Certificate	Comments
End user to Control Plane	Yes	Ingress Controller Certificate	Customizable by user
Control Plane to Vault	Yes	Autogenerated Certificate	
End user to CML workspace	Yes (CML supports workspace creation without TLS. But for production, TLS is strongly recommended)	User-generated certificate	It is used to access the CML workspace. This is a user-generated certificate and can be loaded to CML via the procedure mentioned here: See 'Deploy an ML Workspace with Support for TLS '
CML workspace to git repository	Yes	User-generated certificate	User generated certificate added to the Control Plane in Administration CA certificates.
CML workspace to AMP repository / AMP catalog	Yes	User-generated certificate	User generated certificate added to the Control Plane in Administration CA certificates.
CML workspace to external docker registry	Yes	User-generated certificate	User generated certificate added to the Control Plane in Administration CA certificates.

Related Information

Deploy an ML Workspace with Support for TLS

CDE

Connection	TLS (Yes/ No/ Optional)	Certificate	Comments
End user to control plane	Yes	Ingress Controller Certificate	Customizable by user
Control plane to Vault	Yes	Autogenerated certificate	
End user to CDE Jobs UI	Yes	User generated certificate	Used to access CDE jobs UI. Customer uploads certificate using cde-utils.sh.
End user through CDE CLI to Jobs API endpoint	Yes	User generated certificate	Used to access CDE jobs UI. Customer uploads certificate using cde-utils.sh.
End user to CDE Service	Yes	User uploaded certificate	Used to access CDE jobs UI. Customer uploads certificate using cde-utils.sh.

Note: There are two levels of chaining of ingress objects in CDE. This means that when a user sends a request to the <jobs-url>/dex/*, it first is caught by the cluster ingress controller and redirected to the respective CDE nginx controller present in the CDE Service. The primary ingress here is TLS enabled whereas the secondary ingress (within the CDE namespace) is not TLS enabled. There is no way to directly access the secondary ingress controller as every request is first caught by the cluster ingress controller due to its ingress rule <jobs-urls>/*

Compute Service

Connection	TLS (Yes/ No/ Optional)	Certificate	Comments
End user to Liftie	Yes	Ingress controller Certificate	Customizable by user
Liftie to Vault	Yes	Auto generated certifcate	
Liftie to Postgres Database	Yes	Auto generated certifcate	
End user to resource-management End point	Yes	Ingress controller Certificate	Customizable by user

No wildcard DNS/TLS setup

This guide documents the required entries that must be present in DNS and TLS certificates when not using wildcards. This is meant to reflect customer setups where wildcard DNS and TLS are not allowed.

Only the Control Plane and Cloudera Data Warehouse (CDW) support this workflow currently. All entries specified in the Control Plane and CDW sections must be present in DNS and the Ingress controller TLS certificate.

Entries required by Control Plane

Let APPDOMAIN be the base app domain for the ECS cluster, not including the ".apps" subdomain.

```
For example, if your console URL is "console-cdp.apps.cloudera.com", then th
e APPDOMAIN is "cloudera.com".
```

Note: For consistency with Cloudera Manager, the ECS cluster base app domain should not include ".apps".

OpenShift Container Platform (OCP) :

- console-<namespace>.apps.APPDOMAIN
- validation-<namespace>.apps.APPDOMAIN

Embedded Container Service (ECS) :

- console-cdp.apps.APPDOMAIN
- prometheus-cp.apps.APPDOMAIN
- infra-prometheus.apps.APPDOMAIN
- validation-cdp.apps.APPDOMAIN
- kube-dashboard.apps.APPDOMAIN
- longhorn.apps.APPDOMAIN
- fluent-console-cdp.apps.APPDOMAIN

Entries required by CDW

Let APPDOMAIN be the base app domain for the ECS cluster.

```
For example, if your console URL is "console-cdp.apps.cloudera.com", then th
e APPDOMAIN is "cloudera.com".
```

Let VWHNAME be the name of the CDW Virtual Warehouse. This must match the name the user provides when creating a new Virtual Warehouse (VW).

Endpoints of Hive VW:

- hue-VWHNAME.apps.APPDOMAIN
- hs2-VWHNAME.apps.APPDOMAIN

Endpoints of Impala VW:

- hue-VWHNAME.apps.APPDOMAIN
- coordinator-VWHNAME.apps.APPDOMAIN
- admissiond-web-VWHNAME.apps.APPDOMAIN
- catalogd-web-VWHNAME.apps.APPDOMAIN
- coordinator-web-VWHNAME.apps.APPDOMAIN
- statestored-web-VWHNAME.apps.APPDOMAIN
- impala-proxy-VWHNAME.apps.APPDOMAIN
- impala-autoscaler-web-VWHNAME.apps.APPDOMAIN

Endpoints of Viz:

- viz-VWHNAME.apps.APPDOMAIN

Adding DNS entries

For each entry in the certificate, create an 'A' record pointing to the IP address of the host running the ECS Ingress Controller (should be the same host running the ECS server role). When creating additional virtual warehouses, create additional DNS entries.

Adding TLS certificate entries

You must construct a single TLS certificate with all of the entries as SubjectAltName (SAN) fields. This certificate and corresponding private key (in PEM format) must be placed on the Cloudera Manager server host, and the paths to those files must be specified in the Ingress Controller TLS certificate and private key configurations when creating the ECS cluster.

When creating additional virtual warehouses, you must sign a new certificate with all existing SANs plus the SANs for the new virtual warehouse. Place the new certificate on the Cloudera Manager server host (overwriting the old one if desired), and set the Ingress Controller TLS certificate and private key configurations in the ECS service to the new file paths (if required). Then run the Cloudera Manager command to rotate the Ingress Controller TLS certificate.