

CDP Public Cloud

Azure Reference Network Architecture

Date published: 2019-08-22

Date modified:

CLOUDERA

<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 Public Cloud reference network architecture for Azure.....4**
- Taxonomy of network architectures.....5**
 - Management Console to customer cloud network.....6
 - Customer on-prem network to cloud network.....8
- Network architecture.....8**
 - Architecture diagrams.....9
 - Component description..... 11
 - VNet..... 11
 - Subnets..... 11
 - Gateways and route tables.....12
 - Private endpoints.....12
 - Security groups..... 13
 - Security groups for Data Lakes.....13
 - Additional rules for AKS-based workloads..... 13
 - Outbound connectivity requirements.....14
 - Domain names for the endpoints..... 14
 - DNS..... 16

CDP Public Cloud reference network architecture for Azure

This topic includes a conceptual overview of the CDP Public Cloud network architecture for Azure, its use cases, and personas who should be using it.

Overview

CDP Public Cloud allows customers to set up cloud Data Lakes and compute workloads in their cloud accounts on AWS, Azure, and Google Cloud. It maps a cloud account to a concept called the environment into which all CDP workloads, including Data Hubs (compute workload clusters) and data services (such as Cloudera Data Engineering (CDE), Cloudera Data Warehouse (CDW), Cloudera Machine Learning (CML), Cloudera Operational Database (COD), Cloudera DataFlow (CDF)) are launched. For these Data Lakes, Data Hubs, and data services to function correctly, several elements of the cloud architecture need to be configured appropriately: access permissions, networking setup, cloud storage and so on. Broadly, these elements can be configured in one of two ways:

- CDP can set up these elements for the customer

Usually, this model helps to set up a working environment quickly and try out CDP. However, many enterprise customers prefer or even mandate specific configurations of a cloud environment for Infosec or compliance reasons. Setting up elements such as networking and cloud storage requires prior approvals and they would generally not prefer, or even actively prevent, a third party vendor like Cloudera to set up these elements automatically.

- CDP can work with pre-created elements provided by the customer

In this model, the flow for creating the cloud Data Lakes accepts pre-created configurations of the cloud environment and launches workloads within those boundaries. This model is clearly more aligned with enterprise requirements. However, it brings with it the risk that the configuration might not necessarily play well with CDP requirements. As a result, customers might face issues launching CDP workloads and the turnaround time to get to a working environment might be much longer and involve many tedious interactions between Cloudera and the customer cloud teams.

From our experience in working with several enterprise customers, the most complicated element of the cloud environment setup is the cloud network configuration. The purpose of this document is to clearly articulate the networking requirements needed for setting up a functional CDP Public Cloud environment into which the Data Lakes and compute workloads of different types can be launched. It attempts to establish the different points of access to these workloads and establishes how the given architecture helps to accomplish this access.

Along with this document, you can use the “cloudera-deploy tool” to automatically set up a model of this reference architecture, which can then be reviewed for security and compliance purposes.



Note:

Currently this document only covers network architecture required for registering a CDP environment (with a Data Lake and FreeIPA) and deploying CDW, and CML in the environment. It does not currently cover Data Hubs and the remaining data services (CDE, CDF, and COD).

Use cases

CDP Public Cloud allows customers to process data in the cloud storage under a secure and governed Data Lake using different types of compute workloads that are provisioned via Data Hub or data services. Typically the lifecycle of these workloads is as follows:

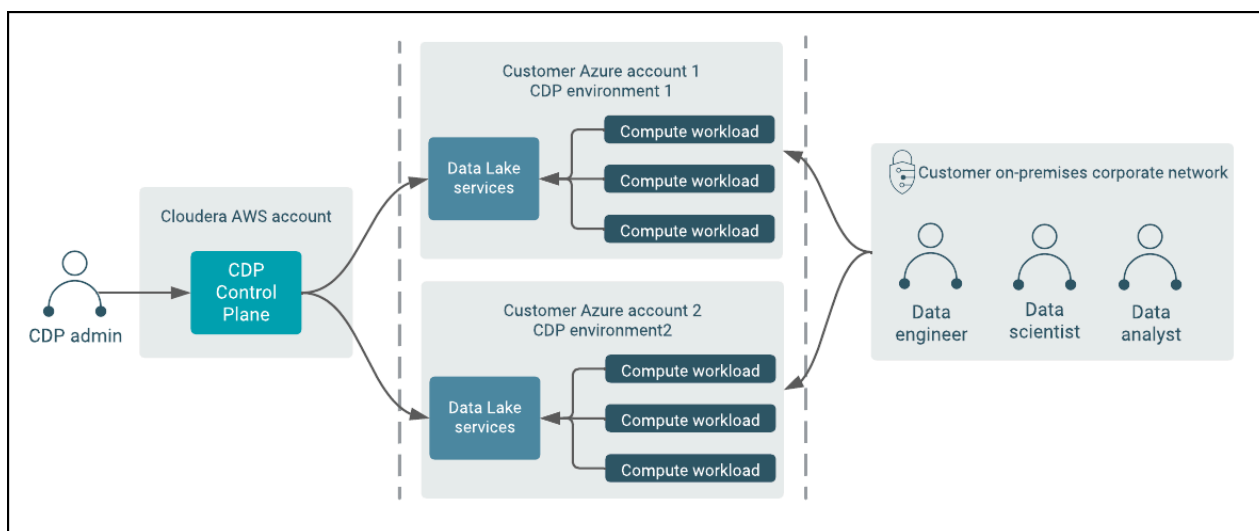
- A CDP environment is set up by a CDP admin using their cloud account. This sets up a cloud Data Lake cluster and FreeIPA cluster with security and governance services and an identity provider for this environment. The CDP admin may need to work with a cloud administrator to create all the cloud provider resources (including networking resources) that are required by CDP.

- Then one or more Data Hubs and data services can be launched, linked to the Data Lake. Each of these Data Hubs and data services typically serves a specific purpose such as data ingestion, analytics, machine learning and so on.
- These Data Hubs and data services are accessed by data consumers such as data engineers, analysts or scientists. This is the core purpose of using CDP on the public cloud.
- These compute workload clusters and data services can be long-running or ephemeral, depending on the customer needs.

There are two types of CDP users who interact with the product for different purposes:

- CDP admins - These persons are usually concerned with the launch and maintenance of the cloud environment, and the Data Lake, FreeIPA, Data Hubs, and data services running inside the environment. They use a Management Console running in the Cloudera AWS account to perform these operations of managing the environment.
- Data consumers - These are the data scientists, data analysts, and data engineers who use the Data Hubs and data services to process data. They mostly interact directly with the compute workloads (Data Hubs and data services) running in their cloud account. They could access these either from their corporate networks (typically through a VPN) or other cloud networks their corporate owns.

These two types of users and their interaction with CDP are represented in the following diagram:



Related Information

[cloudera-deploy tool](#)

Taxonomy of network architectures

This topic provides a high-level overview of each type of network architecture that CDP supports.

At a high level, there are several types of network architectures that CDP supports. As can be expected, each type brings a unique trade-off among various aspects, such as ease of setup, security provided, workloads supported, and so on. This section only provides a high level overview of each type. The characteristics of each type are explained under appropriate sections in the rest of the document. The users must review the advantages and disadvantages of each of these taxonomies in detail before making a choice suitable to their needs.

Name	Description	Trade-offs
Publicly accessible networks	Deploys customer workloads to hosts with public IP addresses. Security groups must be used to restrict access only to corporate networks as needed.	Easy to set up for POCs. Low security levels.

Name	Description	Trade-offs
Semi-private networks	Deploys customer workloads to private subnets, but exposes services to which data consumers need access over a load balancer with a public IP address. Security groups or allow-lists (of IP addresses or ranges) on load balancers must be used to restrict access to these public services only to corporate networks as needed.	This option is fairly easy to set up too, but it may not solve all the use cases of access (in semi private networks). The surface of exposure is reduced, and it is reasonably secure.
Fully private networks	Deploys customer workloads to private subnets, and the services to which data consumers need access are only on private IPs. Requires connectivity to corporate networks to be provided using solutions like VPN gateways, and so on.	Complex to set up depending on prior experience of establishing such connectivity, primarily due to the way the customer has to solve the corporate network peering problem. But it is very secure.
Fully private outbound restricted networks	This is the same as fully private networks; Except, in addition, Cloudera provides a mechanism for users to configure an outbound proxy or firewall to monitor or restrict the communication outside their networks.	Most complex to set up, mainly considering the varied needs that data consumers would have to connect outside the VNet on an evolving basis. It is also the most secure for an enterprise.

Management Console to customer cloud network

This topic explains the possible ways in which the CDP Control Plane can communicate with the compute infrastructure in the customer network, in the context of the Management Console.

As described previously, the CDP admin would typically use the CDP Management Console that runs in the CDP Control Plane to launch CDP environments with Data Lakes, FreeIPA, Data Hubs, and data services into their cloud accounts. In order to accomplish this, the CDP Control Plane and the compute infrastructure in the customer network (such as VMs, AKS clusters) should be able to communicate with each other. Depending on the chosen network architecture, this communication can occur in the ways described below.

Publicly accessible networks

In this model of publicly accessible networks, the compute infrastructure must be reachable over the public internet from the Management Console. While this is fairly easy to set up, it is usually not preferred by enterprise customers, as it implies that the VM nodes or AKS nodes are assigned public IP addresses. While the access control rules for these nodes can still be restricted to the IP addresses of the Management Console components, it is still considered insecure for each of the network architectures described earlier.

Semi-private networks

Publicly accessible networks are easy to set up for connectivity, both from the CDP Control Plane and the customer on-prem network, but have a large surface area of exposure as all compute infrastructure has public IP addresses. In contrast, fully private networks need special configuration to enable connectivity from the customer on-prem network, due to having no surface area of exposure to any of the compute infrastructure. While very secure, it is more complex to establish.

There is a third configuration supported by CDP, semi-private networks, that provides some trade-offs between these two options. In this configuration, the user deploys the worker nodes of the compute infrastructure on fully private networks as described above. However, the user chooses to expose UIs or APIs of the services fronting these worker nodes over a public network load balancer. By using this capability, the data consumers can access the UIs or APIs of the compute infrastructure through these load balancers. It is also possible to restrict the IP ranges from which such access is allowed using security groups.

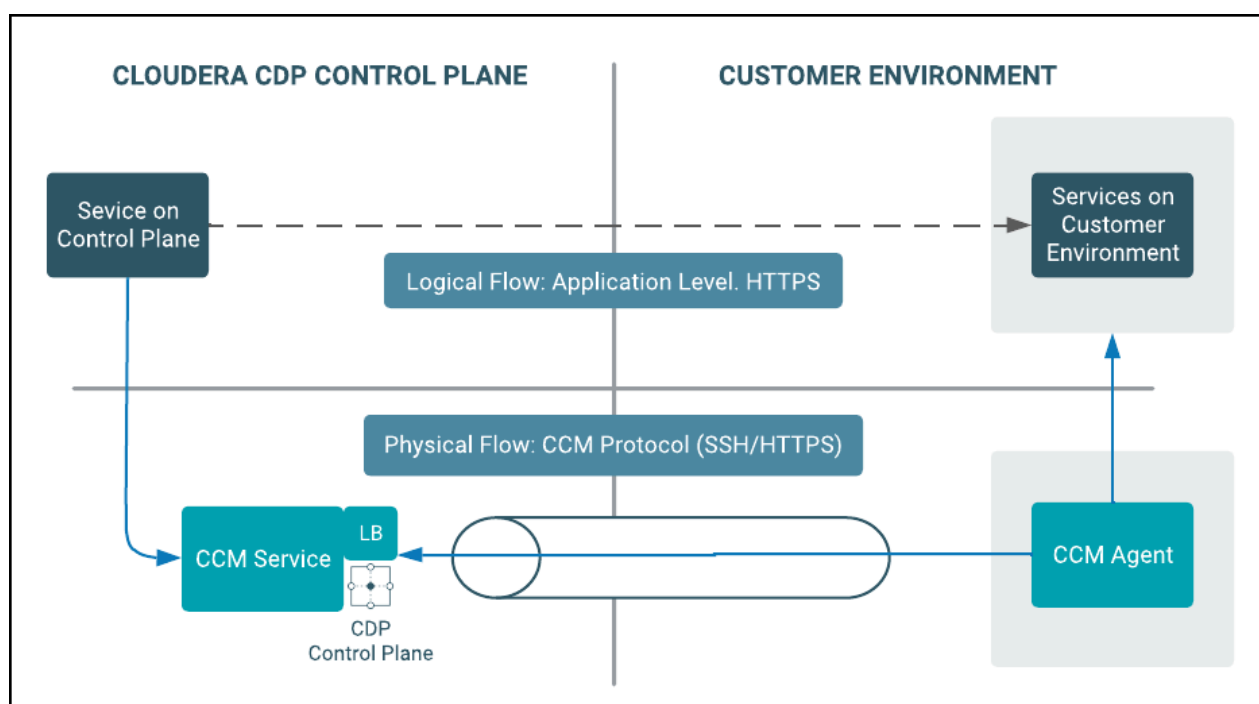
While this option provides a trade-off between ease of setup and exposure levels, it may not satisfy all use cases related to communication between various endpoints. For example, some compute workloads involving Kafka or NiFi would not benefit from having a simple publicly exposed load balancer. It is recommended that customers evaluate their use cases against the trade-off and choose an appropriately convenient and secure model of setup.

Fully private networks

In this model of fully private networks, the compute infrastructure is not assigned any public IP addresses. In this case, communication between the CDP Control Plane and compute infrastructure is established using a “tunnel” that originates from the customer network to the CDP Control Plane. All communication from the CDP Control Plane to the compute nodes is then passed through this tunnel. From experience, Cloudera has determined that this is the preferred model of communication for customers.

To elaborate on the tunneling approach, Cloudera uses a solution called “Cluster Connectivity Manager” (CCM). At a high level, the solution uses two components, an agent (CCM agent) that runs on a VM provisioned in the customer network and a service (CCM service) that runs on the CDP Control Plane. The CCM agent, at start-up time, establishes a connection with the CCM service. This connection forms the tunnel. This tunnel is secured by asymmetric encryption. The private key is shared with the agent over cloud specific initialization mechanisms, such as a user-data script in Azure.

When any service on the CDP Control Plane wants to send a request to a service deployed on the customer environment (depicted in the below diagram as the “logical flow”), it physically sends a request to the CCM service running in the CDP Control Plane. The CCM agent and CCM service collaborate over the established tunnel to accept the request, forward it to the appropriate service, and send a response over the tunnel to be handed over the calling service on the CDP Control Plane.



Currently, all AKS clusters provisioned by various CDP data services are enabled with public and private cluster endpoints. The AKS public endpoint is needed to facilitate the interactions between CDP Control Plane and the AKS cluster while worker nodes and Kubernetes control plane interact over private API endpoints. CDW supports private AKS endpoints today (see “Enabling a private CDW environment in Azure Kubernetes Service”). There are plans to support private AKS endpoints for other data services in the future. When this occurs, the documentation will be updated to reflect the same.

Fully private outbound restricted networks

Fully private outbound restricted networks is a variant of the fully private network where customers would like to pass outbound traffic originating from their cloud account through a proxy or firewall and explicitly allow-list URLs that are allowed to pass through. CDP Public Cloud supports such configuration. If such network architecture is chosen, the customer must ensure the following:

- Users configure a proxy for the environment via CDP, as documented in “Using a non-transparent proxy”.

- Compute resources (such as VMs used by Data Hubs and data services) can connect to the proxy or firewall via appropriate routing rules.
- The proxy or firewall is set up to allow connections to all hosts, IP ranges, ports, and protocol types that are documented in “Azure outbound network access destinations”.

**Note:**

Given that fully private networks is the recommended option of connectivity in most cases, this document describes the architecture assuming a fully private network setup.

Related Information

[Cluster Connectivity Manager](#)

[Enabling a private CDW environment in Azure Kubernetes Service](#)

[Using a non-transparent proxy](#)

[Azure outbound network access destinations](#)

Customer on-prem network to cloud network

After Data Hubs and data services are launched in the customer’s cloud network, data consumers such as data engineers, data scientists, and data analysts access services running in these CDP data services. Sometimes, CDP admins who set up and operate these clusters might need this access to diagnose any issues the clusters face.

Examples of these include:

- Web UIs such as:
 - Hue: For running SQL queries in Hive tables
 - CML Workspaces: For accessing Cloudera Machine Learning projects, models, notebooks, and so on
 - Cloudera Manager: For Data Hubs and Data Lakes
 - Atlas and Ranger: For metadata, governance, and security in the Data Lake
- JDBC endpoints: Customers can connect tools such as Tableau using a JDBC URL pointing to the Hive server.
- SSH access: Data engineers might log in to nodes on the compute workload clusters and data services to run data processing jobs using YARN, Spark, or other data pipeline tools.
- Kube API access: CDP data services that run on AKS (such as Cloudera Data Warehouse and Cloudera Machine Learning) also provide admin access to Kubernetes for purposes of diagnosing issues.
- API access: Customers can use APIs for accessing many of the services exposed via the web UIs for purposes of automation and integration with other tools, applications, or other workloads they have. For example, CML exposes the CML API v2 to work with Cloudera Machine Learning projects and other entities. See [CML API v2](#).

These services are accessed by these consumers from within a corporate network inside a VPN. These services typically have endpoints that have a DNS name, the format of which is described more completely in the DNS section of this reference architecture documentation. These DNS names resolve to IP addresses assigned to the nodes, or load balancers fronting the ingest controllers of Kubernetes clusters. Note that these IP addresses are usually private IPs; Therefore, in order to be able to connect to these IPs from the on-premise network within a VPN, some special connectivity setup would be needed, typically accomplished using technologies like VPN peering, DirectConnect, transit gateways, and so on. While there are many options possible here, this document describes one concrete option of achieving this connectivity.

Related Information

[CML API v2](#)

Network architecture

Cloudera recommends that customers configure their cloud networks as fully private networks, as described in this chapter. This will help on-boarding CDP Data Lakes, Data Hubs, and data services smoothly.

**Note:**

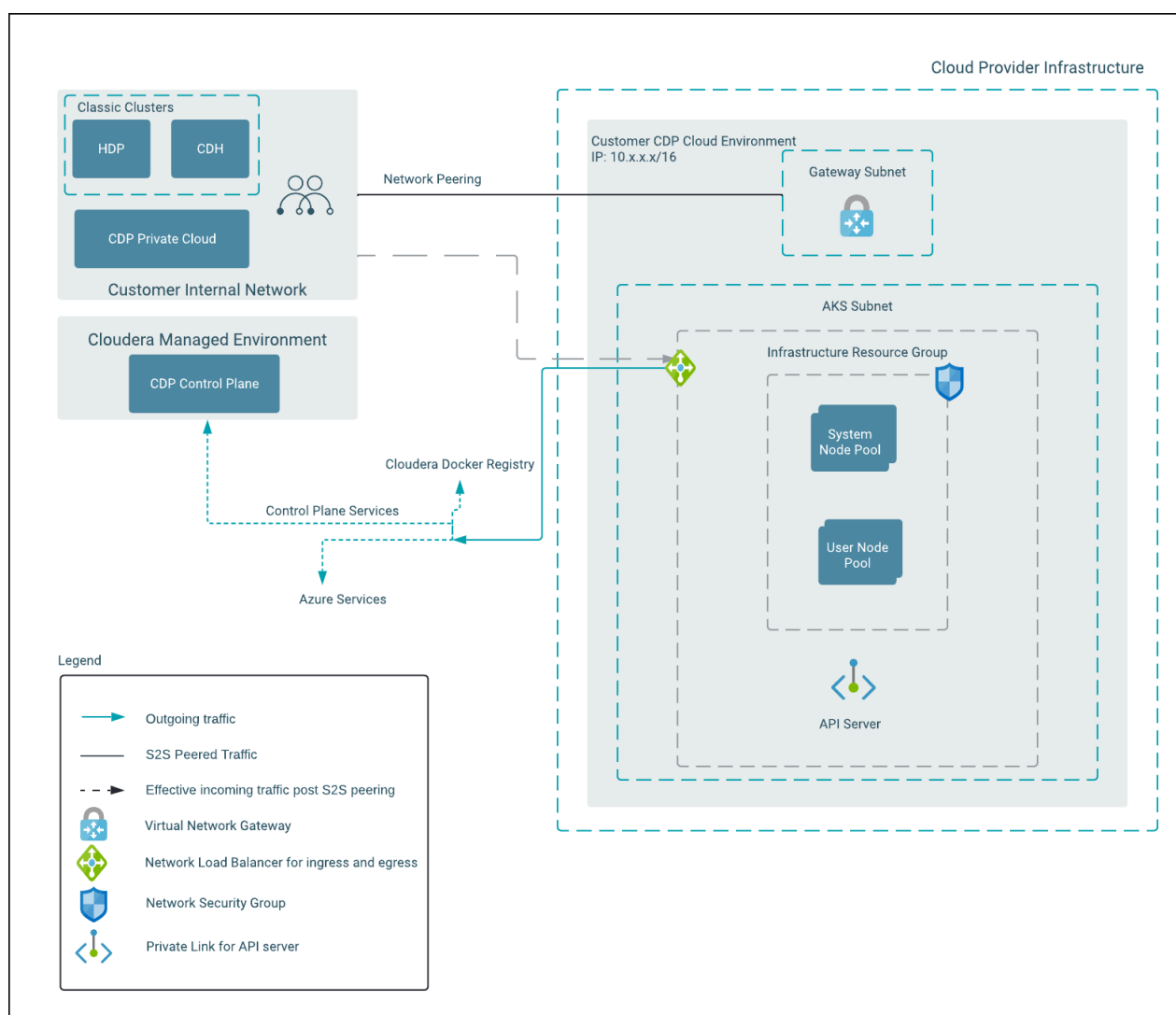
This network architecture only covers the fully private networks and assumes unrestricted outbound access.

The cloudera-deploy tool, which is released along with this document can be used to automatically set up a model of this reference architecture, which can then be reviewed for security and compliance purposes.

Architecture diagrams

This topic includes diagrams illustrating the various elements of the network architecture in the customer's cloud account into which CDP environments with Data Lakes, Data Hubs, and data services will be launched.

Cloudera recommends that customers configure their cloud networks as described in this chapter and illustrated in the following diagrams. This will help onboarding Data Lakes, Data Hubs, and data services smoothly. The following diagram illustrates the configuration for a fully private network that can be configured by the customer. This configuration can be provided by the CDP admins when they are setting up CDP environments and workloads which will get launched into this configuration.



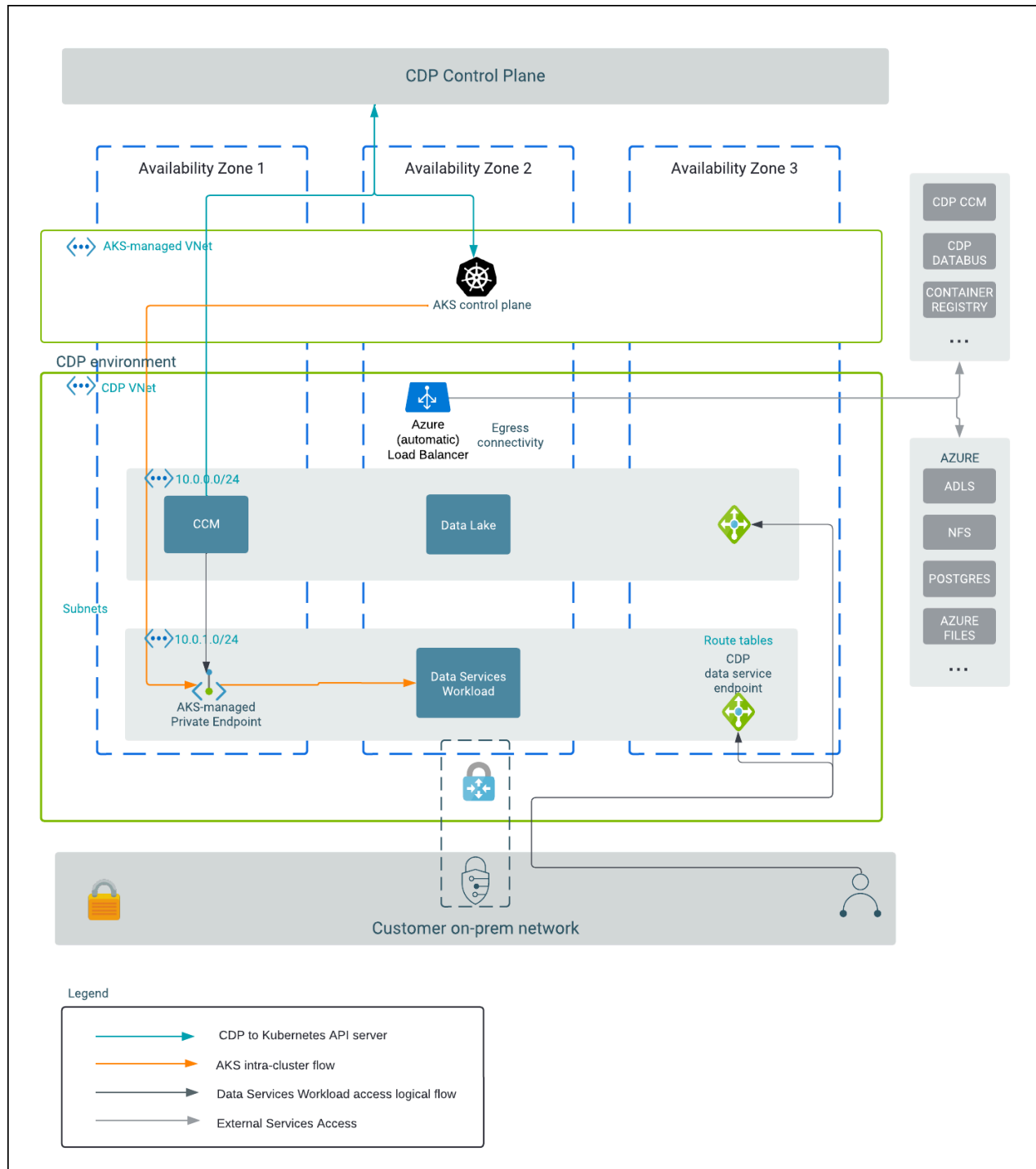
Note the following points about this architecture:

- The configuration is a fully private network configuration - that is, the workloads are launched on nodes that do not have public IP addresses.
- Workloads connect to the CDP Control Plane over a fixed IP and port range.

- For users to be able to connect from the customer on-prem network to the CDP workloads in the private subnet, some network connectivity setup is required. In this case, a customer's network peered to Azure VNet via Azure VPN gateway is shown.

Some of the CDP data services are based on Azure AKS clusters. Azure's AKS manages the Kubernetes control plane nodes while the worker nodes that make up the CDP workload cluster get provisioned in the customer's VPC. The AKS control plane has an API endpoint for administrative purposes which is commonly referred to as API server address. The data service itself is accessible through a service endpoint ELB.

This is illustrated in the following diagram:



As can be seen in the above diagram, CDP workloads have dependencies on some Azure cloud services such as ADLS, Azure Postgres and so on. A full list of these services, described in the context of each workload is specified in “Azure outbound network access destinations”.

In this example of a fully private deployment with an existing network of your own configuration, you must configure your own egress connectivity through a [NAT gateway setup](#) or [user-defined routing](#). For more information see [VNet and subnets](#) and [Azure Load Balancers in Data Lakes and Data Hubs](#).

In the chapters that follow, we detail the elements of this architecture, including specifying the configuration and options in each of the components.

Related Information

[Azure outbound network access destinations](#)

Component description

This section includes an overview of the VNet, subnets, gateways and route tables, and security groups required for CDP Public Cloud for Azure.

VNet

An Azure Virtual Network (VNet) is needed for deploying CDP workloads into the customer’s cloud account. VNet is similar to a traditional network that you would operate in your own data center, but brings with it additional benefits of Azure's infrastructure such as scale, availability, and isolation.

Cloudera recommends that the VNet used for CDP is configured with the properties specified below:

- The CIDR block for the VNet should be sufficiently large for supporting all the Data Hubs and data services that you intend to run. Refer to “VNet and subnet planning” to understand how to compute the CIDR block range.
- In addition, you may want to ensure that the CIDR ranges assigned to the VNet do not overlap with any of your on-premise network CIDR ranges, as this may be a requirement for setting up connectivity from your on-premise network to the subnets.

Related Information

[VNet and subnet planning](#)

Subnets

A subnet is a partition of the virtual network in which CDP workloads are launched.

It is recommended that the subnets be configured with the following properties:

- The CIDR block for the subnet should be sufficiently large for supporting the CDP workload targeted to run inside it. Refer to “VNet and subnet planning” to understand how to compute the CIDR block range.
- Several CDP data services run on Kubernetes and use Kubenet CNI plugin for networking. Azure's Kubenet CNI plugin requires that the subnet is not shared between multiple Kubernetes clusters as it adds custom routes (see “Bring your own subnet and route table with kubenet”). Therefore, as many subnets as the expected number of workloads need to be created.
- In addition, you may want to ensure that the CIDR ranges assigned to the subnets do not overlap with any of your on-premise network CIDR ranges, as this may be a requirement for setting up connectivity from your on-prem network to the subnets.
- A subnet can be associated with a Network Security Group (NSG). However, since Cloudera works with a fully private network configuration where communication is always initiated from VMs within the subnets, an NSG at subnet level is generally not useful for this configuration.

Related Information

[VNet and subnet planning](#)

[Bring your own subnet and route table with kubenet](#)

Gateways and route tables

This topic covers recommended gateway and route table configurations for CDP Public Cloud for Azure.

Connectivity from Control Plane to CDP workloads

- As described in the “Subnets” section above, each CDP data service workload requires its own subnet and a non-shared route table associated with it (see “Bring your own subnet and route table with kubernetes”).
- As described in the “Taxonomy of network architectures”, nodes in the CDP workloads need to connect to the CDP Control Plane over the internet to establish a “tunnel” over which the CDP Control Plane can send instructions to the workloads.
- Private AKS cluster is a feature that lets CDP access a Kubernetes workload cluster over a private IP address (see “Enabling a private CDW environment in Azure Kubernetes Service”). When it is enabled for a CDW workload, CDW requires the user to have already set up internet connectivity for the subnet (see “Outbound type of userDefinedRouting”).
- CDP data services such as Datalake and CML create a public load balancer for internet connectivity.
- If a firewall is configured, the destinations described in “Azure outbound network access destinations” need to be allowed for the CDP workloads to work.

Connectivity from customer on-prem to CDP workloads

- As described in the “Use cases” section (see “CDP Public Cloud reference network architecture for Azure”>“Use Cases”), data consumers need to access data processing or consumption services in the CDP workloads.
- Given that these are created with private IP addresses in private subnets, the customers need to arrange for access to these addresses from their on-prem or corporate networks in specific ways.
- There are several possible solutions for achieving this, but one that is depicted in the architectural diagram, uses “Azure VPN Gateway”.
- Each workload provides an option to connect to the workload endpoint over a public or a private IP. It's recommended that the public IPs are disabled and that users rely on the corporate network connectivity for accessing the workloads with private IPs.

Related Information

[Subnets](#)

[Bring your own subnet and route table with kubernetes](#)

[Taxonomy of network architectures](#)

[Enabling a private CDW environment in Azure Kubernetes Service](#)

[Outbound type of userDefinedRouting](#)

[Azure outbound network access destinations](#)

[CDP Public Cloud reference network architecture for Azure](#)

[Azure VPN Gateway](#)

Private endpoints

CDP workloads can be configured to access the Azure resources over a private IP (which is called a private endpoint) or over a public IP (which is called a service endpoint).

The private endpoint setup requires a private DNS zone which the VNet is linked to and resolves the Azure resources to private IP addresses. CDP supports a private endpoint configuration for Azure postgres only. CDP admin can choose to either create the private DNS zone and link it to the VNet as described in “Bringing your own private DNS”, or let the CDP create them when provided with necessary configuration described in “Using CDP-managed private DNS”.

CDP supports only service endpoint configuration for other Azure resources(such as Microsoft Storage). The subnets need to be enabled to support the service endpoints. See “Service endpoint for Azure Postgres” for detailed steps.

Related Information

[Bringing your own private DNS](#)

[Using CDP-managed private DNS](#)
[Service endpoint for Azure Postgres](#)

Security groups

During the specification of a VNet to CDP, the CDP admin specifies security groups that will be associated with all the CDP workloads launched within that VNet. These security groups will be used in allowing the incoming traffic to the hosts.

Security groups for Data Lakes

During the specification of a VNet to CDP, the CDP admin can either let CDP create the required security groups, taking a list of IP address CIDRs as input; or create them in Azure and then provide them to CDP.

When getting started with CDP, the CDP admin can let CDP create security groups, taking a list of IP address CIDRs as input. These will be used in allowing the incoming traffic to the hosts. The list of CIDR ranges should correspond to the address ranges from which the CDP workloads will be accessed. In a VPN-peered VNet, this would also include address ranges from customer's on-prem network. This model is useful for initial testing given the ease of set up.

Alternatively, the CDP admin can create security groups on their own and select them during the setup of the VNet and other network configuration. This model is better for production workloads, as it allows for greater control in the hands of the CDP admin. However, note that the CDP admin must ensure that the rules meet the requirements described below.

For a fully private network, network security groups should be configured according to the types of access requirements needed by the different services in the workloads. The "Network security groups" section includes all the details of the necessary rules that need to be configured.

Note that for a fully private network, even specifying an open access here (such as 0.0.0.0/0) is restrictive because these services are deployed in a private subnet without a public IP address and hence do not have an incoming route from the Internet. However, the list of CIDR ranges may be useful to restrict which private subnets of the customer's on-prem network can access the services.

Rules for AKS based workloads are described separately in the following section.

Related Information

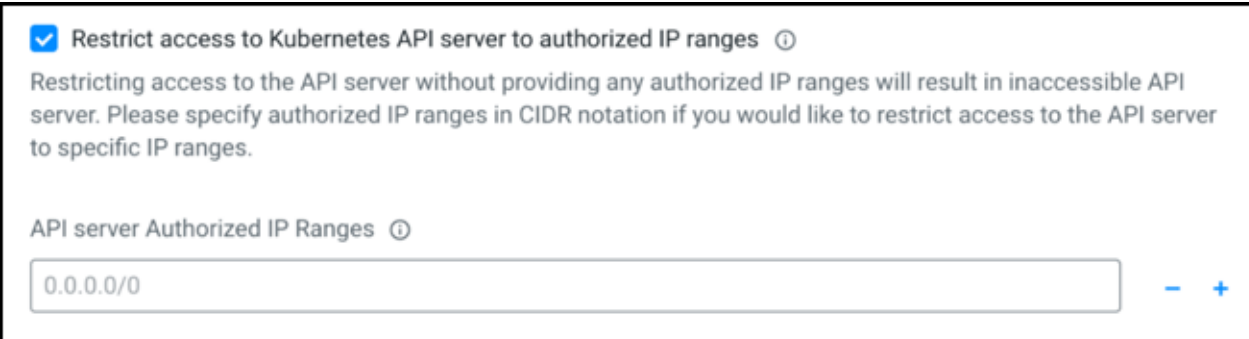
[Network security groups](#)

Additional rules for AKS-based workloads

At the time of enabling a CDP data service, the CDP admin can specify a list of CIDR ranges that will be used in allowing the incoming traffic to the workload Load Balancer.

This list of CIDR ranges should correspond to the address ranges from which the CDP data service workloads will be accessed. In a VPN peered VNet, this would also include address ranges from customer's on-prem network. In a fully private network setup, 0.0.0.0/0 implies access only within the VNet and the peered VPN network which is still restrictive.

CDW data service currently supports fully private AKS clusters, which means the Kubernetes API server will have a private IP address. CML uses a public endpoint by default for all AKS cluster control planes. It is highly recommended to provide a list of outbound public CIDR ranges at the time of provisioning a data service to restrict access to the AKS clusters. In addition, the data service adds CDP public CIDR range to allow access between workloads and CDP Control Plane. An example configuration section for a data service looks like below:



☒ Restrict access to Kubernetes API server to authorized IP ranges ⓘ

Restricting access to the API server without providing any authorized IP ranges will result in inaccessible API server. Please specify authorized IP ranges in CIDR notation if you would like to restrict access to the API server to specific IP ranges.

API server Authorized IP Ranges ⓘ

0.0.0.0/0 - +

Specific guidelines for restricting access to Kubernetes API server and workloads are detailed in “Restricting access for CDP services that create their own security groups on Azure” by each data service.

Within the AKS cluster, security groups are defined to facilitate AKS control plane-pod communication, inter-pod and inter-worker node communication as well as workload communication through Load Balancers.

Related Information

[Restricting access for CDP services that create their own security groups on Azure](#)

Outbound connectivity requirements

Outbound traffic from the worker nodes is unrestricted and is targeted at other Azure services and CDP services. The comprehensive list of services that get accessed from a CDP environment can be found in “Azure outbound network access destinations”.

Related Information

[Azure outbound network access destinations](#)

Domain names for the endpoints

The previous sections dealt with how connectivity is established to the workload infrastructure. This section deals with “addressability”.

The workloads launched by CDP contain a few services that need to be accessed by the CDP admins and data consumers. These include services such as Cloudera Manager; metadata services such as the Hive Metastore, Atlas, or Ranger; and data processing or consumption services such as Oozie server, Hue, and so on. Given the nature of the cloud infrastructure, the IP addresses for the nodes running these services may change (for example if the infrastructure is restarted or repaired). However, these should have statically addressable DNS names so that users can access them with the same names.

In order to help with this, CDP assigns DNS names to these nodes. These naming schemes have the following properties:

- The DNS name is of the following format for each Data Lake node and the Data Lake cluster endpoint:

```
<CLUSTER_NAME>-{<HOST_GROUP><i>}.<ENVIRONMENT-IDENTIFIER>.<CUSTOMER_IDENTIFI-  
TIFIER>.cloudera.site
```

An example could be:

```
my-dataeng-master0.my-envir.aaaa-1234.cloudera.site
```

This name has the following components:

- The base domain is cloudera.site. This is a publicly registered “DNS suffix”. It is also a registered Route53 hosted zone in a Cloudera owned AWS account.
- The <CUSTOMER_IDENTIFIER> is unique to a customer account on CDP made of alphanumeric characters and a “-” (dash).
- The <ENVIRONMENT_IDENTIFIER> is generated based on the environment name and is truncated to 8 characters.
- The <CLUSTER_NAME> is the cluster name given to the Data Lake or Data Hub. It is appended with a <HOST_GROUP> name such as “gateway”, “master”, “worker”, and so on, depending on the role that the node plays in the cluster. If there are more than one of these nodes playing the same role, they are appended with a serial number <i>.
- The DNS name of the endpoints of the CDP data services is of the following format:
 - For a Virtual Warehouse in CDW, it is:

```
<VIRTUAL_WAREHOUSE_NAME>.dw-<CDW_ENVIRONMENT_IDENTIFIER>.<CUSTOMER_IDENT-  
IFIER>.cloudera.site
```

- <VIRTUAL_WAREHOUSE_NAME> is the name of the Virtual Warehouse created. There could be multiple virtual warehouses for a given CDP environment.
- <CDW_ENVIRONMENT_IDENTIFIER> is the identifier for the CDP environment.
- For a Session Terminal in CML workspace, it is:

```
<TTY_SESSION_ID>.<CML_WORKSPACE_ID>.<ENVIRONMENT_IDENTIFIER>.<CUSTOMER_I-  
DENTIFIER>.cloudera.site
```

- <TTY_SESSION_ID> is the ID of the CML Terminal Session ID.
- <CML_WORKSPACE_ID> is the ID of the CML workspace created.
- The <ENVIRONMENT_IDENTIFIER> is generated based on the environment name and is truncated to 8 characters. If the 8th character is a “-” (dash), then it is truncated to 7 characters instead.
- For all the CDP data services listed above, the common portions of the DNS include:
 - The base domain is cloudera.site. This is a publicly registered “DNS Suffix”. It is also a registered Route53 hosted zone in a Cloudera owned AWS account.
 - The <CUSTOMER_IDENTIFIER> is unique to a customer account on CDP made of alphanumeric characters and a “-” (dash).
- The length of the DNS name is restricted to 64 characters due to some limitations with Hue workloads.
- These names are stored as A records in the Route53 hosted zone in the Cloudera managed CDP Control Plane AWS account. Hence, you can resolve these names from any location outside of the VPC. However, note that they would still resolve to private IP addresses and hence are constrained by the connectivity setup described in preceding sections.
- Within a CDP environment, the DNS resolution happens differently. Every CDP environment has a DNS server that is played by a component called FreeIPA. This server is seeded with the hostnames of the nodes of all workload clusters in the environment. Every node in a Data Lake, Data Hub and data service is configured to look up the FreeIPA DNS service for name resolution within the cluster.
- FreeIPA nodes are internal to the CDP environment both for resolving and use. They don't have public DNS records.

Related Information[DNS Suffix](#)**DNS**

Azure-provided DNS is recommended for the VNet. This is required to enable private AKS clusters and private access to Postgres databases among others, as described in “Private endpoints” above.