Cloudera DataFlow

Microsoft Azure account requirements for Cloudera DataFlow

Date published: 2021-04-06
Date modified: 2024-01-09

https://docs.cloudera.com/

Legal Notice

Contents

Azure requirements for DataFlow

As the administrator for your Azure environment, ensure that the environment meets the requirements listed in this topic to enable the Cloudera DataFlow experience in CDP Public Cloud.

Follow the steps to ensure that your Azure environment meets the CDP and DataFlow requirements:

Understand your Azure account requirements for CDP

- Review the *Azure subscription requirements*. The link is in the *Related information* section below.
- Verify that your Azure account for CDP has the required resources.
- Verify that you have the permissions to manage these resources.

Understand the DataFlow requirements

Networking requirements:

- Determine your networking option:
  - Use your existing VNet and subnets
  - Have CDP create a new VNet and subnets
- Review firewall exceptions for Azure AKS
- Determine your load balancer option:
  - public
  - private (internal)

Verify that the following services are available in your environment for DataFlow to use:

Azure environments used for the DataFlow service must have the following resources available in the specific Azure region where the environment is registered. Currently, there is no cross-regional support for DataFlow service.

- Azure Kubernetes Service (AKS)
- Azure Database for PostgreSQL
- Azure Data Lake Storage Gen2
- Virtual machine scale sets
- Dsv4-series
- Azure Availability Zones (optional)

Understand DataFlow role requirements

- There are two CDP user roles associated with the DataFlow service: DFAdmin and DFUser. Any CDP user with the EnvironmentAdmin (or higher) access level must assign these roles to users who require access to the DataFlow console within their environment.

Register an Azure Environment in CDP

Once you have met cloud provider requirements and have created the Azure provisioning credential for CDP, you may proceed to register an Azure Environment.

Instructions: Register an Azure environment

Use only app-based credentials

For the DataFlow service, you must only use an app-based credential, which requires the Contributor role to create a new service principal. For more information about creating an app-based credential for the environment you want to use for the DataFlow service, see Create an app-based credential. If you need to change your environment credential, see Change environment's credential. Both of these references are in the Management Console documentation.

App must have the Contributor role at the subscription level

For environments that you plan to use for the DataFlow service, you must ensure that the application you create in Azure has the built-in Contributor Azure role at the Azure subscription level. For more information, see the description of app-based credentials in Credential options on Azure.

Created Azure app must have access to the storage account used during environment registration

Ensure that the application, which the Azure app-based credentials are attached to, must have access to the ADLS Gen2 storage location that is specified when you register the Azure environment. This is the storage location specified in Step 6 in the Register an Azure environment topic. Also see ADLS Gen2 and managed identities for information about storage accounts for Azure environments. See Minimal setup for cloud storage for further details. These references are in the Management Console documentation.

Azure subscription should be in a similar region as the resources

Ensure that your Azure subscription is in a relatively similar region as the region where your resources are deployed. Particularly, be careful that the regions are governed by the same regulatory laws. For more information, see Azure region requirements in the Management Console documentation. In that topic it specifies that "CDP requires that the ADLS Gen2 storage location provided during environment registration must be in the same region as the region selected for the environment." In addition, please review Azure geographies in the Microsoft documentation.

Related Information

Azure subscription requirements

DataFlow networking in Azure

DataFlow supports different networking options depending on how you have set up your VNet and subnets. If you want DataFlow to use specific subnets, make sure that you specify them when registering a CDP environment.

Vnet and Subnet Requirements

When registering an Azure environment in CDP, you are asked to select a VNet and one or more subnets. DataFlow runs in the VNet registered in CDP as part of your Azure environment.

You have two options:

- You use your existing VNet and subnets for provisioning CDP resources.
- You let CDP create a new VNet and subnets.

Subnets for DataFlow

Learn about DataFlow subnet requirements within your VNet.

DataFlow runs in the VNet registered in CDP as part of your Azure environment. The DataFlow service requires its own subnet. DataFlow on AKS uses the Kubenet CNI plugin provided by Azure. In order to use Kubenet CNI, create multiple smaller subnets when creating an Azure environment.

Cloudera recommends the following:

- Partition the VNet with subnets that are just the right size to fit the expected maximum of nodes in the cluster.
- Use /24 CIDR for these subnets. However, if you prefer a custom range, use the following points to determine the IP addresses for the DataFlow service:
  - The DataFlow service can scale up to 50 compute nodes.
  - Each node consumes one IP address.
  - Additionally, you must allocate two IPs for the base infra nodes.

Firewall exceptions for Azure AKS

Learn about required firewall exceptions for maintenance tasks and management.

If you need to restrict egress traffic in Azure, then you must reserve a limited number of ports and addresses for cluster maintenance tasks including cluster provisioning. See *Control egress traffic for cluster nodes in Azure Kubernetes Service (AKS)* to prepare your Azure environment for AKS deployment.

Cloudera recommends you safelist the Azure portal URLs on your firewall or proxy server for management purposes. For more information, see *Safelist the Azure portal URLs on your firewall or proxy server*.

Related Information

Control egress traffic for cluster nodes in Azure Kubernetes Service (AKS)

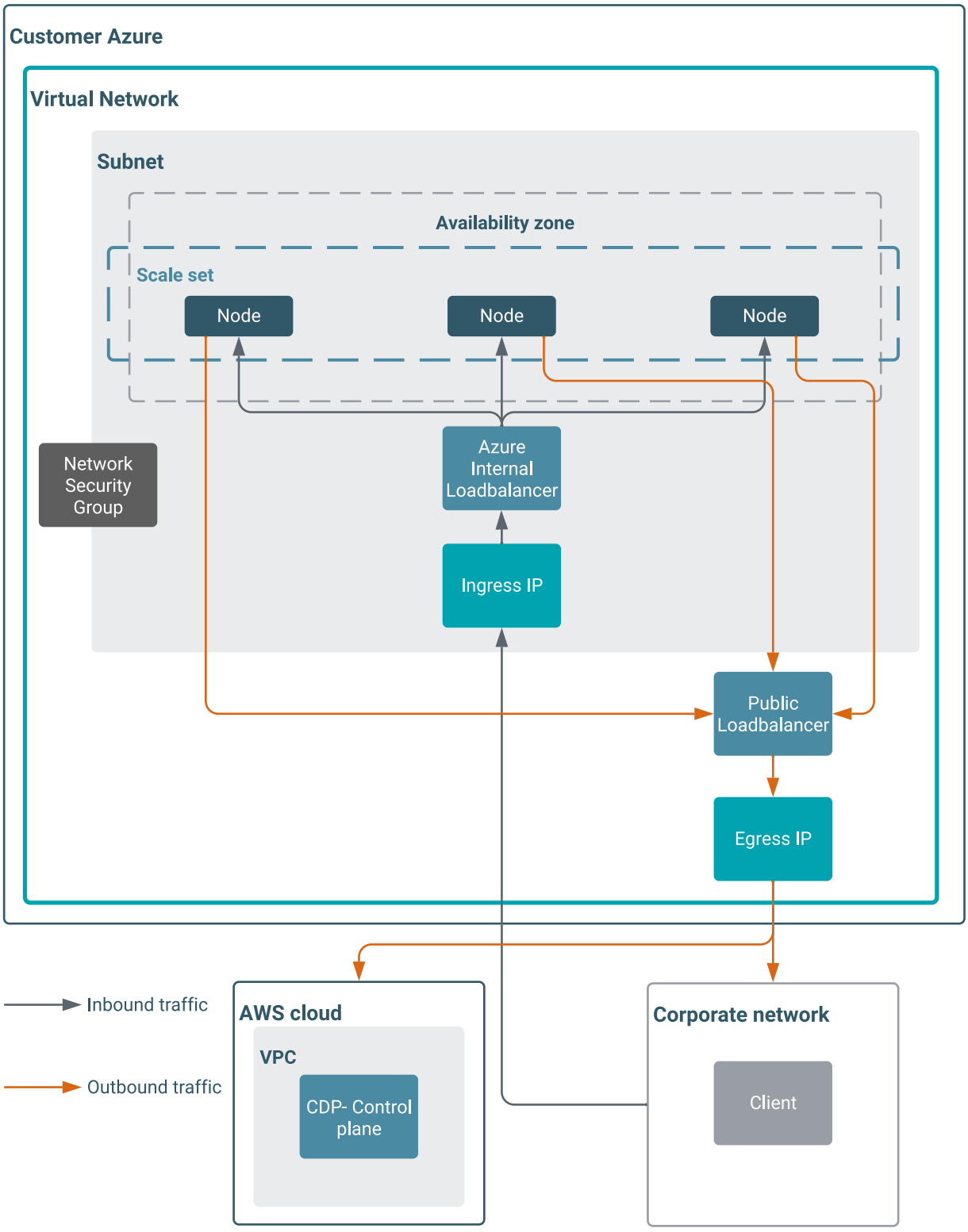
Safelist the Azure portal URLs on your firewall or proxy server

Azure load balancers in DataFlow

Learn about load balancing options in Azure environments.

Azure provides a public and a private (internal) load balancer. DataFlow uses the Standard SKU for the load balancer. You can configure DataFlow to use either private or public load balancer to allow users to connect to flow deployments. By default, DataFlow provisions a private load balancer.

The figure represents a DataFlow deployed with an internal load balacer:



User defined routing

Azure UDR (User-Defined Routing) refers to the capability provided by Microsoft Azure to define and control the flow of network traffic within a virtual network (VNet) using custom routing rules. By default, Azure virtual

networks use system-defined routes to direct network traffic between subnets and to the internet. With UDR, you can override these default routes and specify your own routing preferences.

UDRs allow you to define specific routes for traffic to follow based on various criteria, such as source or destination IP address, protocol, port, or Next Hop type. This gives you granular control over how traffic flows within your Azure infrastructure. With UDR, you can create your own custom routes to control the path of traffic within your Azure virtual network. UDR is implemented using route tables, which contain a set of rules defining the paths for network traffic. Each subnet within a VNet can be associated with a specific route table. UDR allows you to specify the next hop for a route, which can be an IP address, a virtual appliance, or the default Azure Internet Gateway. UDR enables you to segment traffic within your virtual network, directing specific types of traffic through specific network appliances or services. UDR is commonly used in conjunction with Network Virtual Appliances (NVAs), such as firewalls, load balancers, or VPN gateways. By configuring UDR, you can ensure that traffic is properly routed through these appliances.

By leveraging UDR, you can design more complex networking architectures, implement security measures, optimize performance, and meet specific requirements for your Azure deployments.

You can enable UDR when you enable DataFlow for an environment.

Limitations on Azure

This section lists some resource limits that DataFlow and Azure impose on workloads running in DataFlow workspaces.

- There is no ability to grant or revoke remote access (via Kubeconfig) to specific users. Users with the DFAdmin role in the environment can download a Kubeconfig file. The Kubeconfig file will continue to allow access even if the DFAdmin role is later revoked.
- Each DataFlow workspace requires a separate subnet. For more information on this issue, see Use kubenet networking with your own IP address ranges in Azure Kubernetes Service (AKS).
- Heavy AKS activity can cause default API rate limits to trigger, causing throttling and eventually failures for AKS clusters. For some examples, see AKS issue 1187 and AKS issue 1413.

Setting up minimum permissions

The minimum permissions for Cloudera DataFlow (CDF) on Azure govern access control between Azure resources, the Azure storage account, and CDF. The minimum permissions that allow for enabling/disabling CDF and deploying/undeploying flows can be set using a custom role.

Before you begin

- You have registered an application on the Azure Portal. For instructions, see *Create an app registration and assign a role to it*.
- You have created an app-based provisioning credential in your Azure subscription. For instructions, see *Create a provisioning credential for Azure*.

Procedure

1. Create a custom role that contains the minimum permissions.

   The following role definition outlines the minimum permissions required to create a custom role for CDF. The permissions are listed in the Actions section, so that CDF can access resources and operate correctly.

   When using the role defintiion, replace the following values:

   - *[YOUR-SUBSCRIPTION-ID]*: Your subscription ID in use.
   - *[YOUR-RESTRICTED-ROLE-NAME]*: The custom role name which is assigned to the application. For example: *Cloudera Dataflow Azure Operator for Single Resource Group*
   - *[YOUR-RESOURCE-GROUP-NAME]*: The original resource group name.

```
{
   "properties": {
       "roleName": [YOUR-RESTRICTED-ROLE-NAME],
       "description": "Custom restricted role for liftie",
       "isCustom": true,
       "assignableScopes": [
           "/subscriptions/[YOUR-SUBSCRIPTION-ID]/resourceGroups/[YOUR-
RESOURCE-GROUP-NAME]"
       ],
       "permissions": [
           {
               "actions": [
                    "Microsoft.ContainerService/managedClusters/read",
                    "Microsoft.ContainerService/managedClusters/write",
                    "Microsoft.ContainerService/managedClusters/agentPoo
ls/read",
                    "Microsoft.ContainerService/managedClusters/agentPools
/write",
                    "Microsoft.ContainerService/managedClusters/upgrade
Profiles/read",
                    "Microsoft.ContainerService/managedClusters/agentPoo
ls/delete",
                    "Microsoft.ContainerService/managedClusters/delete",
                    "Microsoft.ContainerService/managedClusters/accessP
rofiles/listCredential/action",
                    "Microsoft.ContainerService/managedClusters/agentPoo
ls/upgradeProfiles/read",
                    "Microsoft.Storage/storageAccounts/read",
                    "Microsoft.Storage/storageAccounts/write",
                    "Microsoft.ManagedIdentity/userAssignedIdentities/ass
ign/action",
                    "Microsoft.Compute/virtualMachineScaleSets/write",
                    "Microsoft.Network/virtualNetworks/subnets/join/acti
on",
                    "Microsoft.Network/virtualNetworks/subnets/read",
                    "Microsoft.Insights/diagnosticSettings/write",
                    "Microsoft.Insights/metrics/read",
                    "Microsoft.Insights/metricDefinitions/read"
                ],
               "notActions": [],
               "dataActions": [],
               "notDataActions": []
           }
       ]
   }
}
```

2. Assign the custom role to the app registration that you earlier created on the Azure Portal. For instructions, see *Create an app registration and assign a role to it*.

Related Information

Create an app registration and assign a role to it

Create a provisioning credential for Azure