How to enable Cloudera AI Inference Service (CAII) on Sandbox and Workshop tenants

- About this Page
- Instructions
  - 1. Create Environment and Datalake
  - 2. Initialize Compute Clusters
  - 3. Create New Compute Cluster for CAII Service
  - 4. Create CAI Workbench and Model Registry
  - 5. Create the CAII Service App
    - 5.a Check CAII Service App Status
  - 6. Deploy a Model Endpoint
  - 7. Cleanup of CAII and supporting Services
    - 7.a Delete the CAII Service
    - 7.b Delete the CAI Workbench and Model Registry
    - 7.c Delete the CAII Compute Cluster
    - 7.d Destroy the Environment, Datalake, and Cloud Prereqs

About this Page

The CAII service has multiple dependencies and so deploying it is a multi step process. This guide will walk through the steps to set up the environment/datalake, enable compute clusters for the environment, create a compute cluster for CAII, deploy a CAI Workbench, deploy a model registry, and finally deploy the CAII service and create a model endpoint.

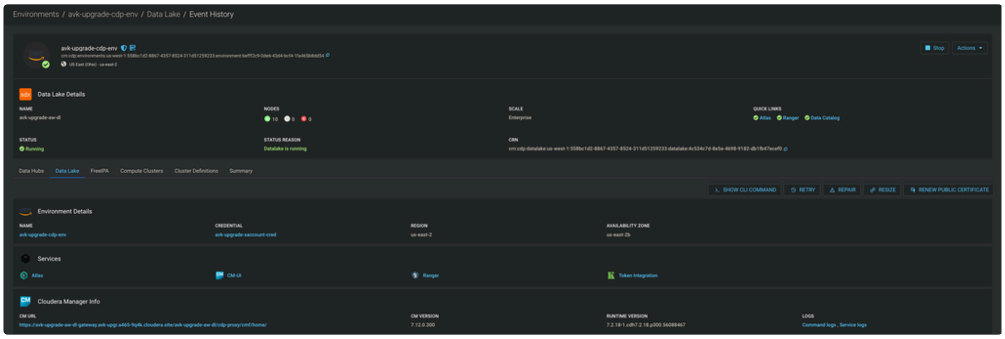
Documentation on CAII Service can be found here: Cloudera AI Inference service Overview

⚠ ***Please note:*** using the CAII service requires significant resources - 4 total kubernetes clusters plus GPUs for model serving - be mindful of the cost associated with deployment and utilize autoscaling and appropriate CPU and GPU node shapes for a given demo/workshop/poc to help limit cost. Please also stop/terminate the environment when not in use

Instructions

1. Create Environment and Datalake

It's recommended that you use terraform automation for speed and simplicity. The terraform quickstart can be found here: GitHub - cloudera-labs/cdp-tf-quickstarts



avk-upgrade-cdp-env
crn:cdp:environments:us-west-1:558bc1d2-8867-4357-8524-311d51259233:environment:6efff2c9-0de6-43d4-bcf4-1fa465b8dd54
US East (Ohio) - us-east-2

Data Lake Details

NAME: avk-upgrade-aw-dl
NODES: 10 0 0
SCALE: Enterprise
QUICK LINKS: Atlas Ranger Data Catalog
STATUS: Running
STATUS REASON: Datalake is running
CRN: crn:cdp:datalake:us-west-1:558bc1d2-8867-4357-8524-311d51259233:datalake:4c534c7d-8e5e-4698-9182-db1fb47ecef0

Environment Details

NAME: avk-upgrade-cdp-env
CREDENTIAL: avk-upgrade-xaccount-cred
REGION: us-east-2
AVAILABILITY ZONE: us-east-2b

Services

Atlas CM-UI Ranger Token Integration

Cloudera Manager Info

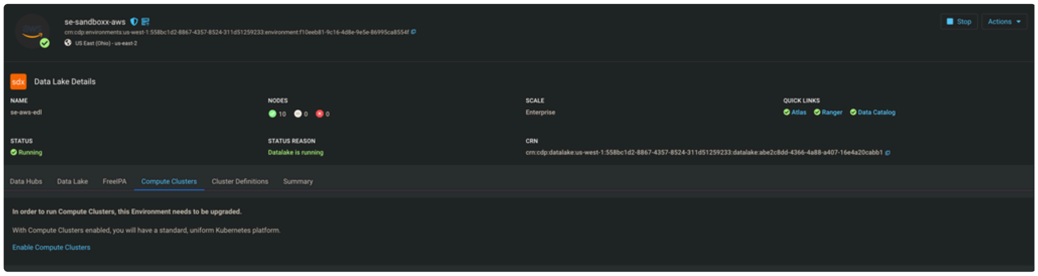
CM URL: https://avk-upgrade-aw-dl-gateway.avk-upgr.a465-9q4k.cloudera.site/avk-upgrade-aw-dl/cdp-proxy/cmf/home/
CM VERSION: 7.12.0.300
RUNTIME VERSION: 7.2.18-1.cdh7.2.18.p300.56088467
LOGS: Command logs , Service logs

2. Initialize Compute Clusters

Once the environment is available you have to initialize the environment to use compute clusters. This operation will migrate the environment from v1 to v2 and create a default compute cluster (aka containerized datalake).

Option 1: UI

Navigate to the environment you deployed in the previous step and go to the `Compute Clusters` tab



crn:cdp:environments:us-west-1:558bc1d2-8867-4357-8524-311d51259233:environment:f10eeb81-9c16-4d8e-9e5e-86995ca8554f

US East (Ohio) - us-east-2

NAME: se-aws-edl

NODES: ● 10 ● 0 ● 0

SCALE: Enterprise

QUICK LINKS: ✓ Atlas ✓ Ranger ✓ Data Catalog

STATUS: ✓ Running

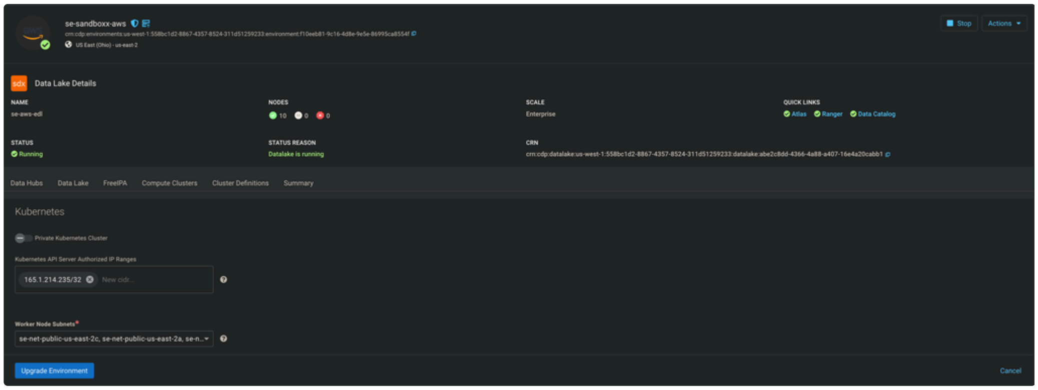
STATUS REASON: Datalake is running

CRN: crn:cdp:datalake:us-west-1:558bc1d2-8867-4357-8524-311d51259233:datalake:abe2c8dd-4366-4a88-a407-16e4a20cabb1

In order to run Compute Clusters, this Environment needs to be upgraded.

With Compute Clusters enabled, you will have a standard, uniform Kubernetes platform.

Select `Enable Compute Clusters` on the screen. You will be prompted for a few options. We recommend keeping the `Private Kubernetes Cluster` option disabled. We also recommend using public subnets in the `Worker Node Subnets` field. Last, for the `Kubernetes API Server Authorized IP Ranges` enter the cidr for the VPN endpoint that you are connected to. If you are unsure of your endpoint IP you can search `My IP Address` on Chrome and it will show you the IP. Just add a `/32` to the end and it will work. Note that this IP whitelist is for access to the kubernetes cluster for administrative purposes. Once everything is set select `Upgrade Environment`



NAME: se-aws-edl

NODES: ● 10 ● 0 ● 0

SCALE: Enterprise

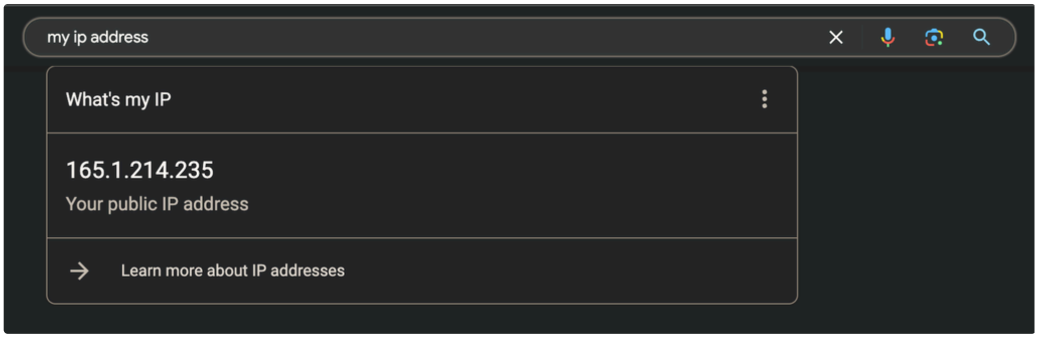
STATUS: ✓ Running

STATUS REASON: Datalake is running

Private Kubernetes Cluster: ○ off

Kubernetes API Server Authorized IP Ranges: 165.1.214.235/32

Worker Node Subnets*: se-net-public-us-east-2c, se-net-public-us-east-2a, se-n...



my ip address

165.1.214.235

Your public IP address

Option 2: CLI

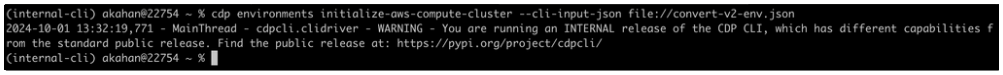
Run the below command from the CDP CLI. This command is available in the public cli release - more details can be found here [initialize-aws-compute-cluster — CDP CLI 0.9.137 documentation]

```
cdp environments initialize-aws-compute-cluster --cli-input-json file://convert-v2-env.json
```

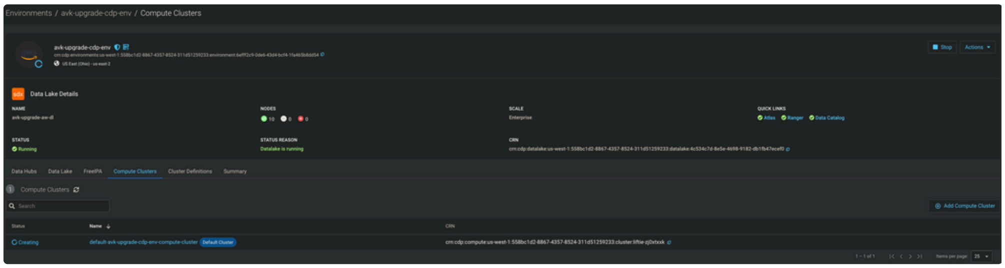
Example convert-v2-env.json:

```
{
    "environmentName": "avk-upgrade-cdp-env",
    "computeClusterConfiguration": {
        "privateCluster": false,
        "kubeApiAuthorizedIpRanges": [
            "165.1.214.236/16",
            "134.238.0.0/16",
            "208.127.0.0/16"
        ],
        "workerNodeSubnets": [
            "subnet-054399ed657492b02",
            "subnet-094a93473bca14029",
            "subnet-025cf012f047c7392"
        ]
    }
}
```

Note: Make sure you use the correct environmentName and workerNodeSunets for your environment created in step 1. Also make sure to add the IP for the VPN endpoint you are connected to in `kubeApiAuthorizedIpRanges` so that you can connect to the kubernetes cluster.



```
(internal-cli) akahan@22754 ~ % cdp environments initialize-aws-compute-cluster --cli-input-json file://convert-v2-env.json
2024-10-01 13:32:19,771 - MainThread - cdpcli.clidriver - WARNING - You are running an INTERNAL release of the CDP CLI, which has different capabilities f
rom the standard public release. Find the public release at: https://pypi.org/project/cdpcli/
(internal-cli) akahan@22754 ~ %
```



Environments / avk-upgrade-cdp-env / Compute Clusters

avk-upgrade-cdp-env

crn:cdp:environments:us-west-1:558bc1d2-8867-4357-8524-311d51259233:environment:6efff2c9-0de6-43d4-bcf4-1fa465b8dd54

US East (Ohio) - us-east-2

Data Lake Details

NAME: avk-upgrade-aw-dl

NODES: 10 0 0

SCALE: Enterprise

QUICK LINKS: Atlas Ranger Data Catalog

STATUS: Running

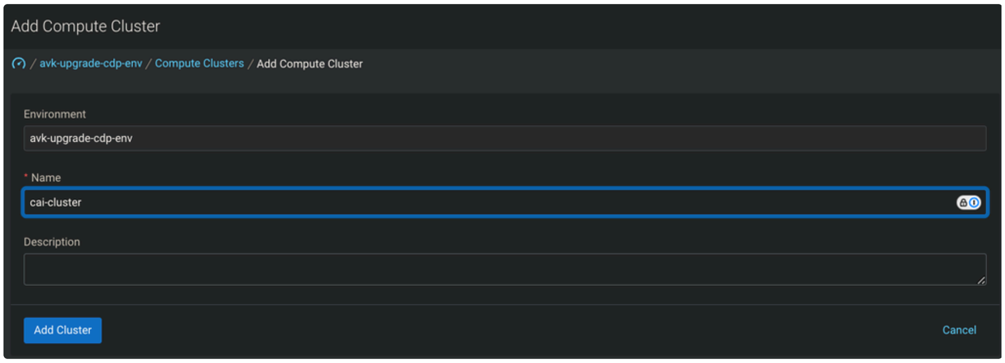
STATUS REASON: Datalake is running

CRN: crn:cdp:datalake:us-west-1:558bc1d2-8867-4357-8524-311d51259233:datalake:4c534c7d-8e5e-4698-9182-db1fb47ecef0

Status	Name	CRN
Creating	default-avk-upgrade-cdp-env-compute-cluster Default Cluster	crn:cdp:compute:us-west-1:558bc1d2-8867-4357-8524-311d51259233:cluster:liftie-zj0xtxxk

3. Create New Compute Cluster for CAII Service

The CAII Service requires a dedicated compute cluster. This can be created directly from the UI.



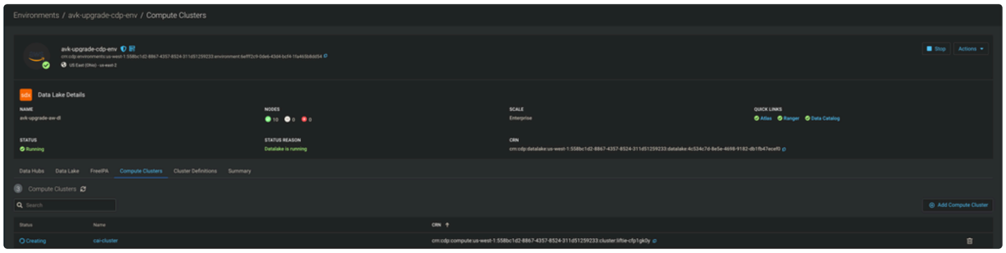
Add Compute Cluster

/ avk-upgrade-cdp-env / Compute Clusters / Add Compute Cluster

Environment: avk-upgrade-cdp-env

* Name: cai-cluster

Description:

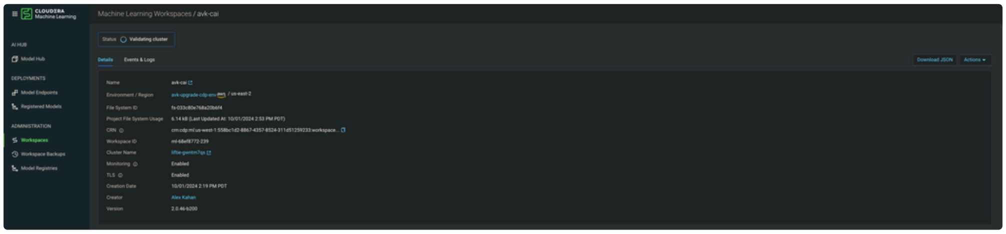


NAME: avk-upgrade-aw-dl
NODES: 10 0 0
SCALE: Enterprise
QUICK LINKS: Atlas Ranger Data Catalog
STATUS: Running
STATUS REASON: Datalake is running
CRN: crn:cdp:datalake:us-west-1:558bc1d2-8867-4357-8524-311d51259233:datalake:4c534c7d-8e5e-4698-9182-db1fb47ecef0

Status	Name	CRN
Creating	cai-cluster	crn:cdp:compute:us-west-1:558bc1d2-8867-4357-8524-311d51259233:cluster:liftie-cfp1gk0y

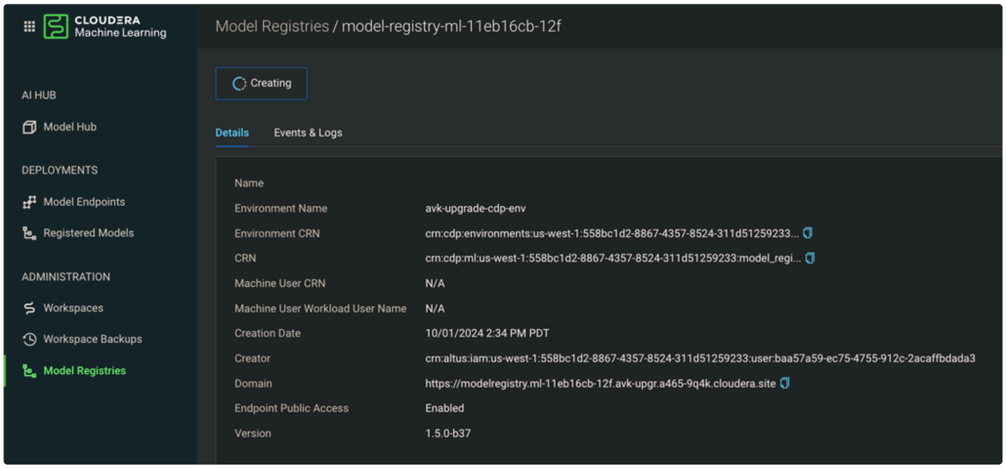
4. Create CAI Workbench and Model Registry

CAII also requires a CAI Workbench and Model Registry to use. These can be created in parallel to the compute cluster.



Status: Validating cluster

Name	avk-cai
Environment / Region	avk-upgrade-cdp-env aws / us-east-2
File System ID	fs-033c80e768a20b6f4
Project File System Usage	6.14 kB (Last Updated At: 10/01/2024 2:53 PM PDT)
CRN	crn:cdp:ml:us-west-1:558bc1d2-8867-4357-8524-311d51259233:workspace...
Workspace ID	ml-68ef8772-239
Cluster Name	liftie-gwntm7qs
Monitoring	Enabled
TLS	Enabled
Creation Date	10/01/2024 2:19 PM PDT
Creator	Alex Kahan
Version	2.0.46-b200



Creating

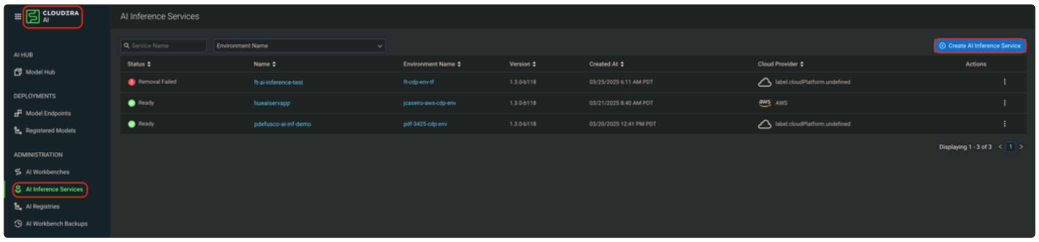
Name	
Environment Name	avk-upgrade-cdp-env
Environment CRN	crn:cdp:environments:us-west-1:558bc1d2-8867-4357-8524-311d51259233...
CRN	crn:cdp:ml:us-west-1:558bc1d2-8867-4357-8524-311d51259233:model_regi...
Machine User CRN	N/A
Machine User Workload User Name	N/A
Creation Date	10/01/2024 2:34 PM PDT
Creator	crn:altus:iam:us-west-1:558bc1d2-8867-4357-8524-311d51259233:user:baa57a59-ec75-4755-912c-2acaffbdada3
Domain	https://modelregistry.ml-11eb16cb-12f.avk-upgr.a465-9q4k.cloudera.site
Endpoint Public Access	Enabled
Version	1.5.0-b37

5. Create the CAII Service App

Once the Workbench, Registry, and Compute Cluster are ready you can deploy the CAII service app. This can currently be done through the UI or CLI

Option 1: UI

Go to `Cloudera AI` → `AI Inference Services` and select `Create New AI Inference Service`



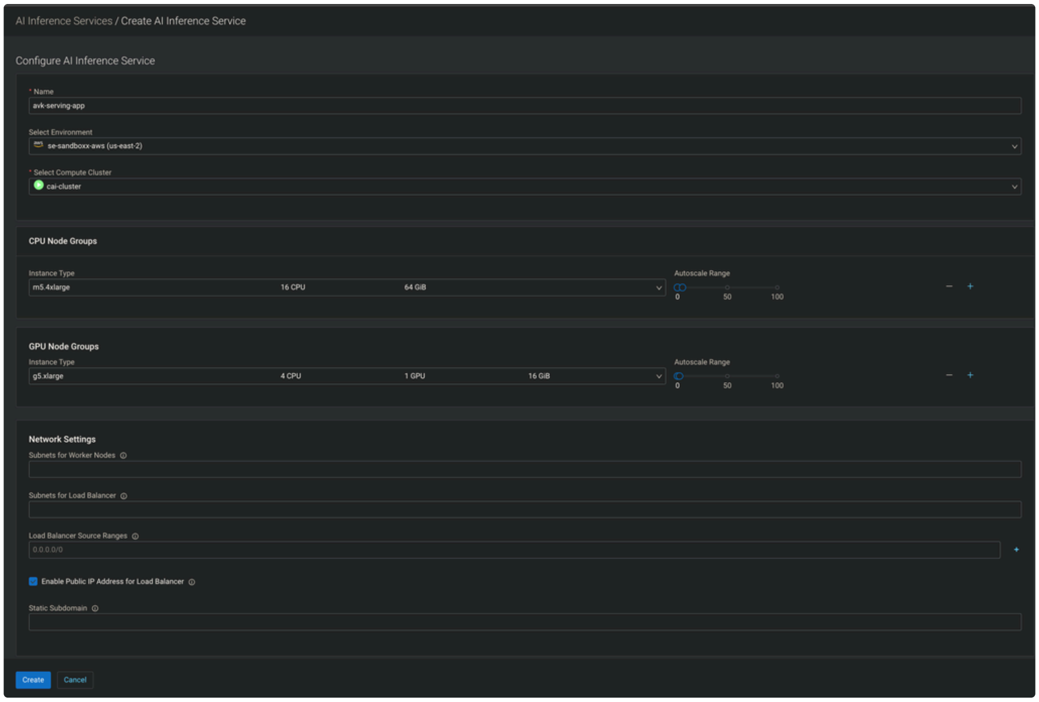
Status	Name	Environment Name	Version	Created At	Cloud Provider	Actions
Removal Failed	ft-ai-inference-test	ft-cdp-env-tf	1.3.0-b118	03/25/2025 6:11 AM PDT	label.cloudPlatform.undefined	⋮
Ready	hueaiservapp	jcaseiro-aws-cdp-env	1.3.0-b118	03/21/2025 8:40 AM PDT	AWS	⋮
Ready	pdefusco-ai-inf-demo	pdf-3425-cdp-env	1.3.0-b118	03/20/2025 12:41 PM PDT	label.cloudPlatform.undefined	⋮

Displaying 1 - 3 of 3

In the configuration page enter a name for the Inference Service, select the correct environment and compute cluster created in Step 3, select the CPU and GPU instance shapes and select `Enable Public IP Address for Load Balancer`

Note: To create a model serving app you must have the EnvironmentAdmin and MLAdmin roles on the environment

Note: We recommend using the g5 gpu shape for the CAII service



AI Inference Services / Create AI Inference Service

Configure AI Inference Service

* Name: avk-serving-app

Select Environment: se-sandboxx-aws (us-east-2)

* Select Compute Cluster: cai-cluster

CPU Node Groups

Instance Type: m5.4xlarge 16 CPU 64 GiB

Autoscale Range: 0 50 100

GPU Node Groups

Instance Type: g5.xlarge 4 CPU 1 GPU 16 GiB

Autoscale Range: 0 50 100

Network Settings

Subnets for Worker Nodes:

Subnets for Load Balancer:

Load Balancer Source Ranges: 0.0.0.0/0

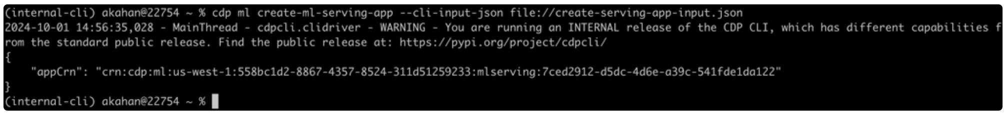
☑ Enable Public IP Address for Load Balancer

Static Subdomain:

Option 2: CLI

Run the below command from the CLI:

```
cdp ml create-ml-serving-app --cli-input-json file://create-serving-app-input.json
```



```
(internal-cli) akahan@22754 ~ % cdp ml create-ml-serving-app --cli-input-json file://create-serving-app-input.json
2024-10-01 14:56:35,028 - MainThread - cdpcli.clidriver - WARNING - You are running an INTERNAL release of the CDP CLI, which has different capabilities f
rom the standard public release. Find the public release at: https://pypi.org/project/cdpcli/
{
    "appCrn": "crn:cdp:ml:us-west-1:558bc1d2-8867-4357-8524-311d51259233:mlserving:7ced2912-d5dc-4d6e-a39c-541fde1da122"
}
(internal-cli) akahan@22754 ~ %
```

Example create-serving-app-input.json:

```
{
    "appName": "avk-serving-app",
    "environmentCrn": "crn:cdp:environments:us-west-1:558bc1d2-8867-4357-8524-311d51259233:environment:6efff2c9-0de6-43d4-bcf4-1fa465b8dd54",
    "clusterCrn": "crn:cdp:compute:us-west-1:558bc1d2-8867-4357-8524-311d51259233:cluster:liftie-cfp1gk0y",
    "provisionK8sRequest": {
        "instanceGroups": [
            {
                "instanceType": "m5.4xlarge",
                "instanceTier": "",
                "instanceCount": 1,
                "name": "",
                "ingressRules": [
                    ""
                ],
                "rootVolume": {
                    "size": 256
                },
                "autoscaling": {
                    "minInstances": 0,
                    "maxInstances": 2,
                    "enabled": true
                }
            },
            {
                "instanceType": "g5.12xlarge",
                "instanceCount": 1,
                "rootVolume": {
                "size": 256
                },
                "autoscaling": {
                    "minInstances": 0,
                    "maxInstances": 1,
                    "enabled": true
                }
            }
        ],
        "environmentCrn": "crn:cdp:environments:us-west-1:558bc1d2-8867-4357-8524-311d51259233:environment:6efff2c9-0de6-43d4-bcf4-1fa465b8dd54",
        "tags": [
            {
                "key": "owner",
                "value": "akahan"
            }
        ],
        "network": {
            "plugin": "",
            "topology": {
                "subnets": [
                    ""
                ]
            }
        }
    },
    "usePublicLoadBalancer": true,
    "skipValidation": true,
    "loadBalancerIPWhitelists": [
        ""
    ],
    "subnetsForLoadBalancers": [
        ""
    ],
    "staticSubdomain": "avk-serving"
}
```

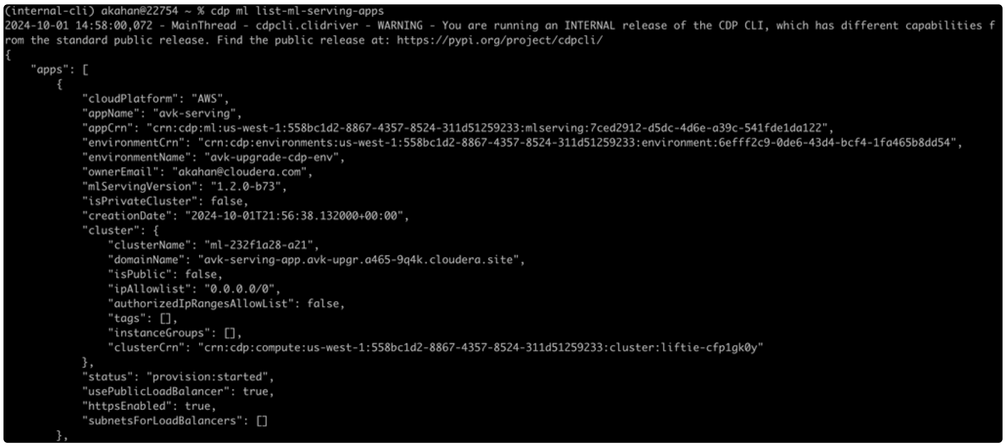
Note: Make sure to update the environmentCrn and clusterCrn with the environment created in step 1 and the compute cluster created in step 3
Note: To create a model serving app you must have the EnvironmentAdmin and MLAdmin roles on the environment

Note: We recommend using the g5 gpu shape for the CAII service

5.a Check CAII Service App Status

You can check the status of the CAII Service app install using command:

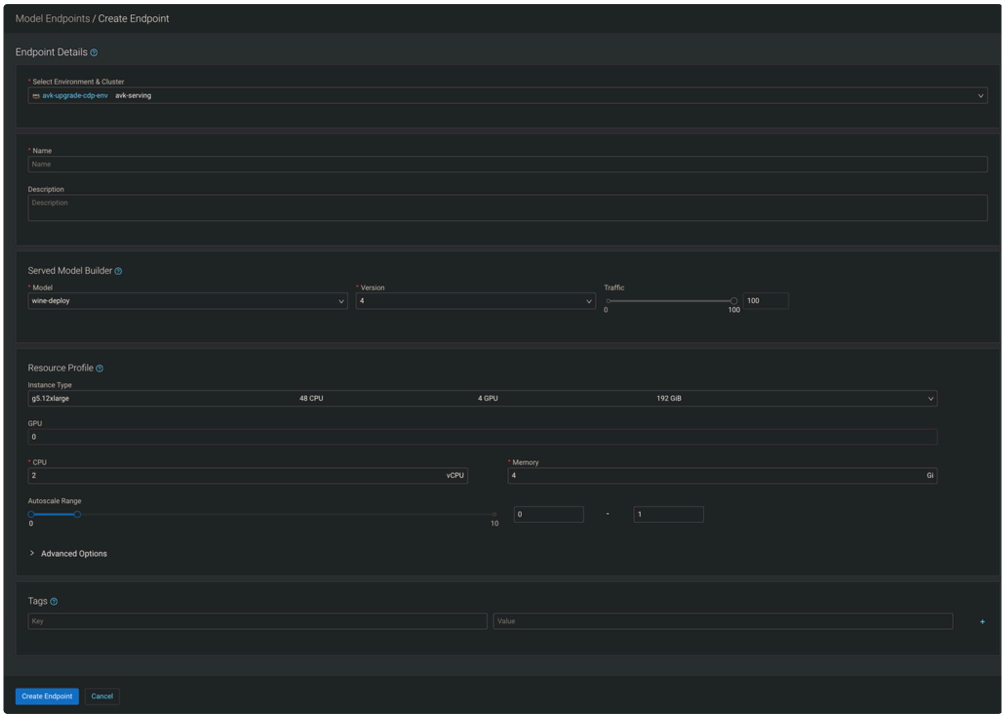
```
cdp ml list-ml-serving-apps
```



```
(internal-cli) akahan@22754 ~ % cdp ml list-ml-serving-apps
2024-10-01 14:58:00,072 - MainThread - cdpcli.clidriver - WARNING - You are running an INTERNAL release of the CDP CLI, which has different capabilities f
rom the standard public release. Find the public release at: https://pypi.org/project/cdpcli/
{
    "apps": [
        {
            "cloudPlatform": "AWS",
            "appName": "avk-serving",
            "appCrn": "crn:cdp:ml:us-west-1:558bc1d2-8867-4357-8524-311d51259233:mlserving:7ced2912-d5dc-4d6e-a39c-541fde1da122",
            "environmentCrn": "crn:cdp:environments:us-west-1:558bc1d2-8867-4357-8524-311d51259233:environment:6efff2c9-0de6-43d4-bcf4-1fa465b8dd54",
            "environmentName": "avk-upgrade-cdp-env",
            "ownerEmail": "akahan@cloudera.com",
            "mlServingVersion": "1.2.0-b73",
            "isPrivateCluster": false,
            "creationDate": "2024-10-01T21:56:38.132000+00:00",
            "cluster": {
                "clusterName": "ml-232f1a28-a21",
                "domainName": "avk-serving-app.avk-upgr.a465-9q4k.cloudera.site",
                "isPublic": false,
                "ipAllowlist": "0.0.0.0/0",
                "authorizedIpRangesAllowList": false,
                "tags": [],
                "instanceGroups": [],
                "clusterCrn": "crn:cdp:compute:us-west-1:558bc1d2-8867-4357-8524-311d51259233:cluster:liftie-cfp1gk0y"
            },
            "status": "provision:started",
            "usePublicLoadBalancer": true,
            "httpsEnabled": true,
            "subnetsForLoadBalancers": []
        },
```

6. Deploy a Model Endpoint

When the serving app is complete, you can create an endpoint from a model imported from the Model Hub or from a model created in the CAI Workbench. The below is an example deployment of the wine-deploy model.



Model Endpoints / Create Endpoint

Endpoint Details

* Select Environment & Cluster: avk-upgrade-cdp-env avk-serving

* Name: Name

Description: Description

Served Model Builder

* Model: wine-deploy

* Version: 4

Traffic: 0 – 100: 100

Resource Profile

Instance Type: g5.12xlarge 48 CPU 4 GPU 192 GiB

GPU: 0

* CPU: 2 vCPU

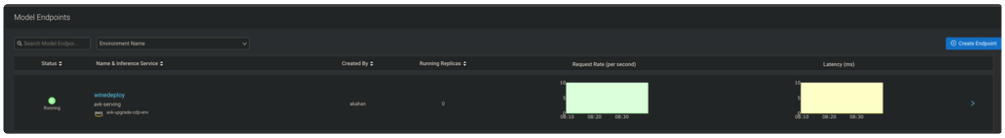
* Memory: 4 Gi

Autoscale Range: 0 – 10: 0 - 1

Advanced Options

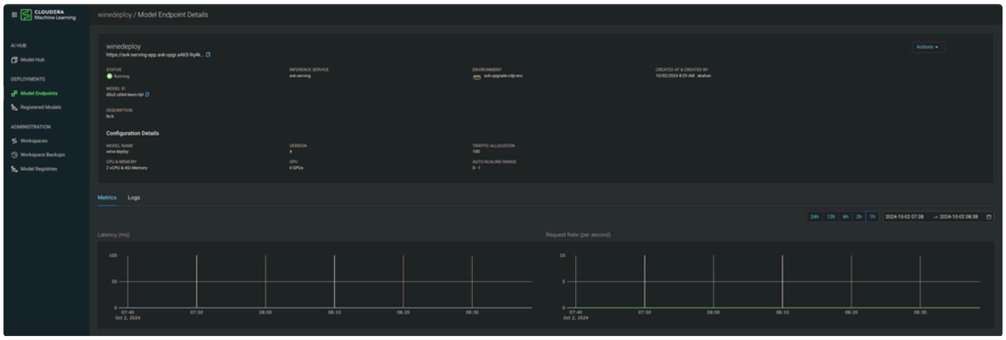
Tags

Key Value



Model Endpoints

Status	Name & Inference Service	Created By	Running Replicas	Request Rate (per second)	Latency (ms)
Running	winedeploy avk-serving avk-upgrade-cdp-env	akahan	0		



winedeploy / Model Endpoint Details

winedeploy
https://avk-serving-app.avk-upgr.a465-9q4k...

STATUS: Running
INFERENCE SERVICE: avk-serving
ENVIRONMENT: avk-upgrade-cdp-env
CREATED AT & CREATED BY: 10/02/2024 8:29 AM akahan
MODEL ID: d0u2-u06d-4esn-rlpl
DESCRIPTION: N/A

Configuration Details

MODEL NAME: wine-deploy
VERSION: 4
TRAFFIC ALLOCATION: 100
CPU & MEMORY: 2 vCPU & 4Gi Memory
GPU: 0 GPUs
AUTO-SCALING RANGE: 0 - 1

2024-10-02 07:38 → 2024-10-02 08:38

Note: To create a model endpoint you must have the EnvironmentUser and MLUser roles on the environment you are using

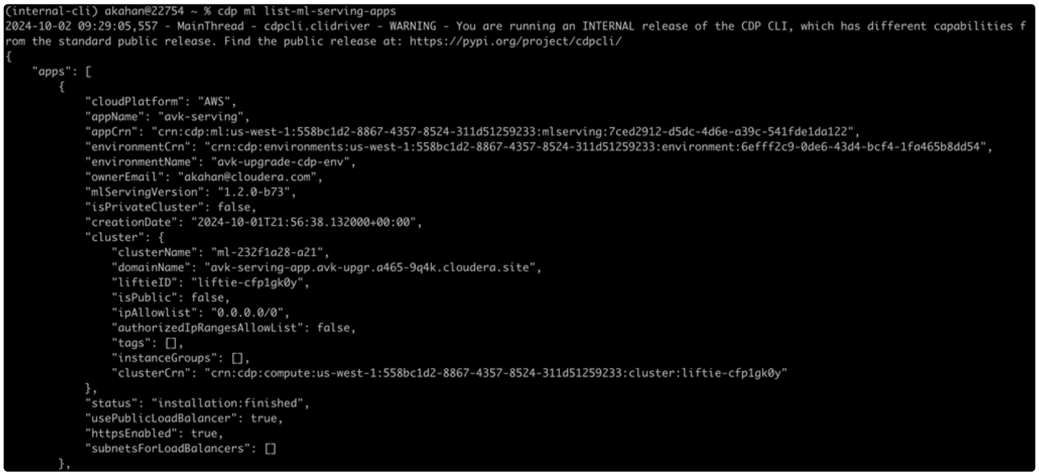
7. Cleanup of CAII and supporting Services

To delete the CAII service and supporting resources you must delete all of the CAI resources first before you can terminate the environment/datalake. Failure to do so will result in orphaned cloud resources and continue to accrue cost after the environment is gone.

7.a Delete the CAII Service

To delete the CAII service you first need to get the service CRN. This can currently only be done with the cli. You can do this with the command:

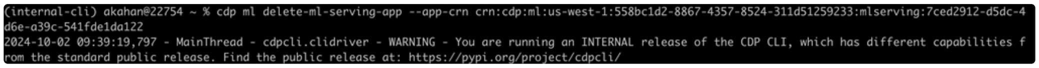
```
cdp ml list-ml-serving-apps
```



```
(internal-cli) akahan@22754 ~ % cdp ml list-ml-serving-apps
2024-10-02 09:29:05,557 - MainThread - cdpcli.clidriver - WARNING - You are running an INTERNAL release of the CDP CLI, which has different capabilities f
rom the standard public release. Find the public release at: https://pypi.org/project/cdpcli/
{
    "apps": [
        {
            "cloudPlatform": "AWS",
            "appName": "avk-serving",
            "appCrn": "crn:cdp:ml:us-west-1:558bc1d2-8867-4357-8524-311d51259233:mlserving:7ced2912-d5dc-4d6e-a39c-541fde1da122",
            "environmentCrn": "crn:cdp:environments:us-west-1:558bc1d2-8867-4357-8524-311d51259233:environment:6efff2c9-0de6-43d4-bcf4-1fa465b8dd54",
            "environmentName": "avk-upgrade-cdp-env",
            "ownerEmail": "akahan@cloudera.com",
            "mlServingVersion": "1.2.0-b73",
            "isPrivateCluster": false,
            "creationDate": "2024-10-01T21:56:38.132000+00:00",
            "cluster": {
                "clusterName": "ml-232f1a28-a21",
                "domainName": "avk-serving-app.avk-upgr.a465-9q4k.cloudera.site",
                "liftieID": "liftie-cfp1gk0y",
                "isPublic": false,
                "ipAllowlist": "0.0.0.0/0",
                "authorizedIpRangesAllowList": false,
                "tags": [],
                "instanceGroups": [],
                "clusterCrn": "crn:cdp:compute:us-west-1:558bc1d2-8867-4357-8524-311d51259233:cluster:liftie-cfp1gk0y"
            },
            "status": "installation:finished",
            "usePublicLoadBalancer": true,
            "httpsEnabled": true,
            "subnetsForLoadBalancers": []
        },
```

Then you can delete the serving app with the command:

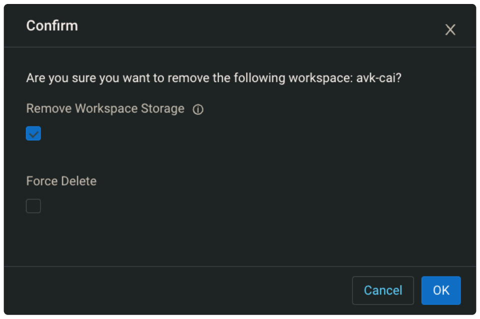
```
cdp ml delete-ml-serving-app --app-crn <serving app crn>
```



```
(internal-cli) akahan@22754 ~ % cdp ml delete-ml-serving-app --app-crn crn:cdp:ml:us-west-1:558bc1d2-8867-4357-8524-311d51259233:mlserving:7ced2912-d5dc-4
d6e-a39c-541fde1da122
2024-10-02 09:39:19,797 - MainThread - cdpcli.clidriver - WARNING - You are running an INTERNAL release of the CDP CLI, which has different capabilities f
rom the standard public release. Find the public release at: https://pypi.org/project/cdpcli/
```

7.b Delete the CAI Workbench and Model Registry

While the CAII service app is deleting you can also delete the Workbench and Model Registry. This can be done through the UI.

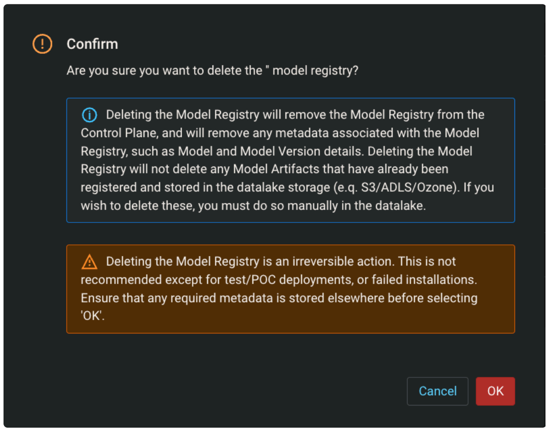


Confirm

Are you sure you want to remove the following workspace: avk-cai?

Remove Workspace Storage ☑

Force Delete ☐



Confirm

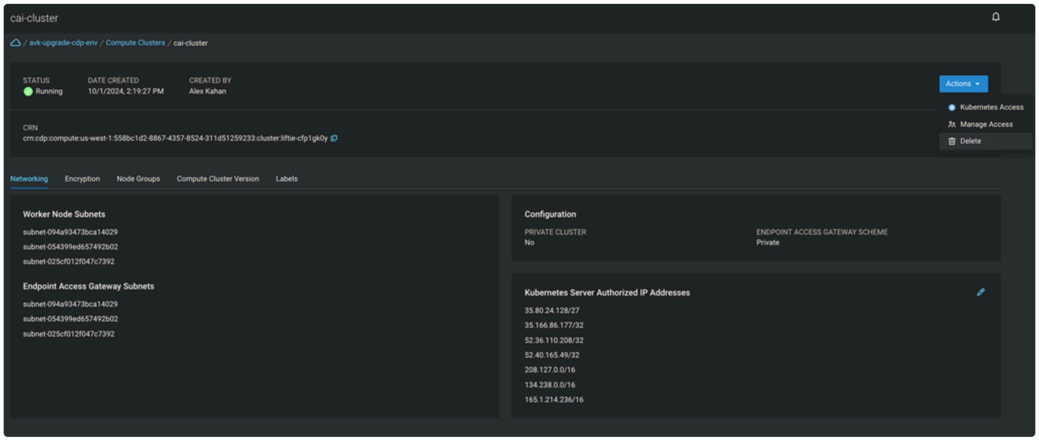
Are you sure you want to delete the " model registry?

Deleting the Model Registry will remove the Model Registry from the Control Plane, and will remove any metadata associated with the Model Registry, such as Model and Model Version details. Deleting the Model Registry will not delete any Model Artifacts that have already been registered and stored in the datalake storage (e.q. S3/ADLS/Ozone). If you wish to delete these, you must do so manually in the datalake.

Deleting the Model Registry is an irreversible action. This is not recommended except for test/POC deployments, or failed installations. Ensure that any required metadata is stored elsewhere before selecting 'OK'.

7.c Delete the CAII Compute Cluster

When the CAII service app is deleted you then have to do delete the compute cluster that was created in step 3. This can be done through the UI



cai-cluster

/ avk-upgrade-cdp-env / Compute Clusters / cai-cluster

STATUS: Running
DATE CREATED: 10/1/2024, 2:19:27 PM
CREATED BY: Alex Kahan

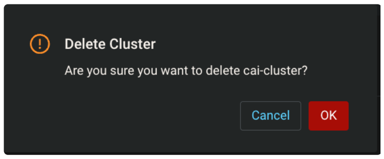
CRN
crn:cdp:compute:us-west-1:558bc1d2-8867-4357-8524-311d51259233:cluster:liftie-cfp1gk0y

Worker Node Subnets
subnet-094a93473bca14029
subnet-054399ed657492b02
subnet-025cf012f047c7392

Endpoint Access Gateway Subnets
subnet-094a93473bca14029
subnet-054399ed657492b02
subnet-025cf012f047c7392

Configuration
PRIVATE CLUSTER: No
ENDPOINT ACCESS GATEWAY SCHEME: Private

Kubernetes Server Authorized IP Addresses
35.80.24.128/27
35.166.86.177/32
52.36.110.208/32
52.40.165.49/32
208.127.0.0/16
134.238.0.0/16
165.1.214.236/16



Delete Cluster

Are you sure you want to delete cai-cluster?

7.d Destroy the Environment, Datalake, and Cloud Prereqs

The last step is to delete the environment, datalake, and cloud resources. If you used terraform to create the environment you should also use terraform to destroy it. This will ensure that all resources on the cloud side are deleted. If you created the environment any other way, make sure that all resources are deleted on the cloud side.

> ℹ Highlight important information in a panel like this one. To edit this panel's color or style, select one of the options in the menu.

📋 Related articles

- How to Enable Dataviz On Base With Knox SSO
- How To Connect Octopai To Hive/Impala on Base
- How to Configure Octopai Client on Platform9
- How to Configure Octopai Client on GCP
- How to Configure Octopai Client on AWS