Partner Summit 2022 - Workshop Student Guide

This document guides students through the Hands on lab for Partner Summit 2022. It will take you step by step to completing the Prerequisites and deliver this demo.

Introduction

The purpose of this repository is to enable the easy and quick setup of the Partner Summit workshop. Cloudera Data Platform (CDP) has been built from the ground up to support hybrid, multi-cloud data management in support of a Data Fabric architecture. This worshop provide an introduction to CDP, with a focus on the data management capabilities that enable the Data Fabric and Data Lakehouse.

Overview

In this exercise, we will work get stock data from Alpha Vantage, offers free stock APIs in JSON and CSV formats for realtime and historical stock market data,

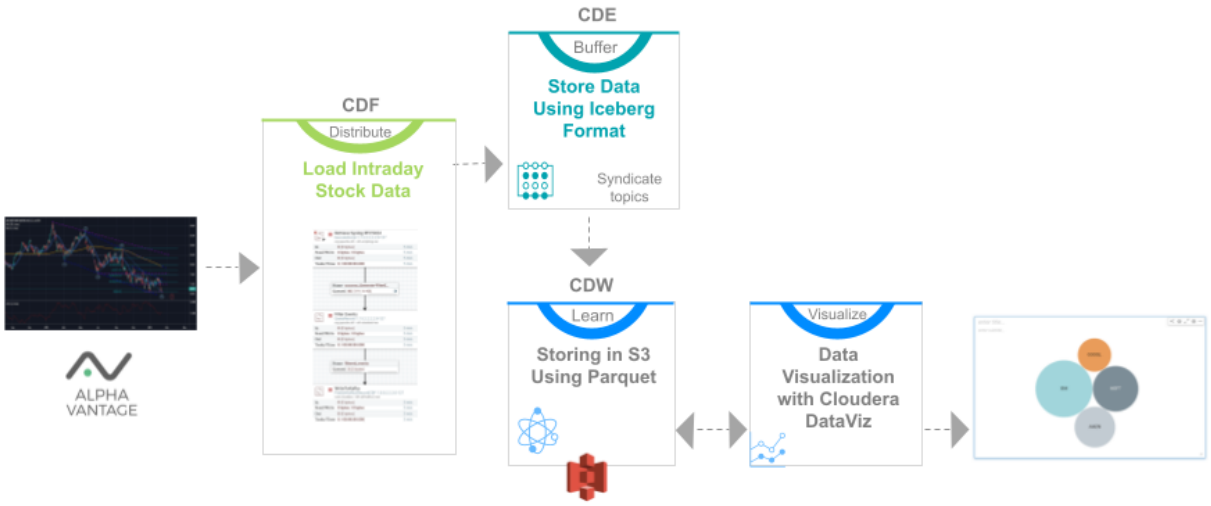
- Data ingestion and streaming—provided by ***Cloudera Data Flow (CDF)*** and ***Cloudera Data Engineering (CDE)**.
- Global data access, data processing and persistence—provided by ***Cloudera Data Hub (CDH)***.
- Data visualization with ***CDP Data Visualization***.

Cloudera DataFlow (CDF) is a scalable, real-time streaming analytics platform that ingests, curates, and analyzes data for key insights and immediate actionable intelligence. CDF's Flow Management is powered by Apache NiFi, a no-code data ingestion and management solution. Apache NiFi is a very mature open source solution meant for large scale, high velocity enterprise data ingestion use cases.

Cloudera Data Engineering (CDE) is a serverless service for Cloudera Data Platform that allows you to submit batch jobs to auto-scaling virtual clusters. CDE enables you to spend more time on

your applications, and less time on infrastructure. CDE allows you to create, manage, and schedule Apache Spark jobs without the overhead of creating and maintaining Spark clusters. With Cloudera Data Engineering, you define virtual clusters with a range of CPU and memory resources, and the cluster scales up and down as needed to run your Spark workloads, helping to control your cloud costs.

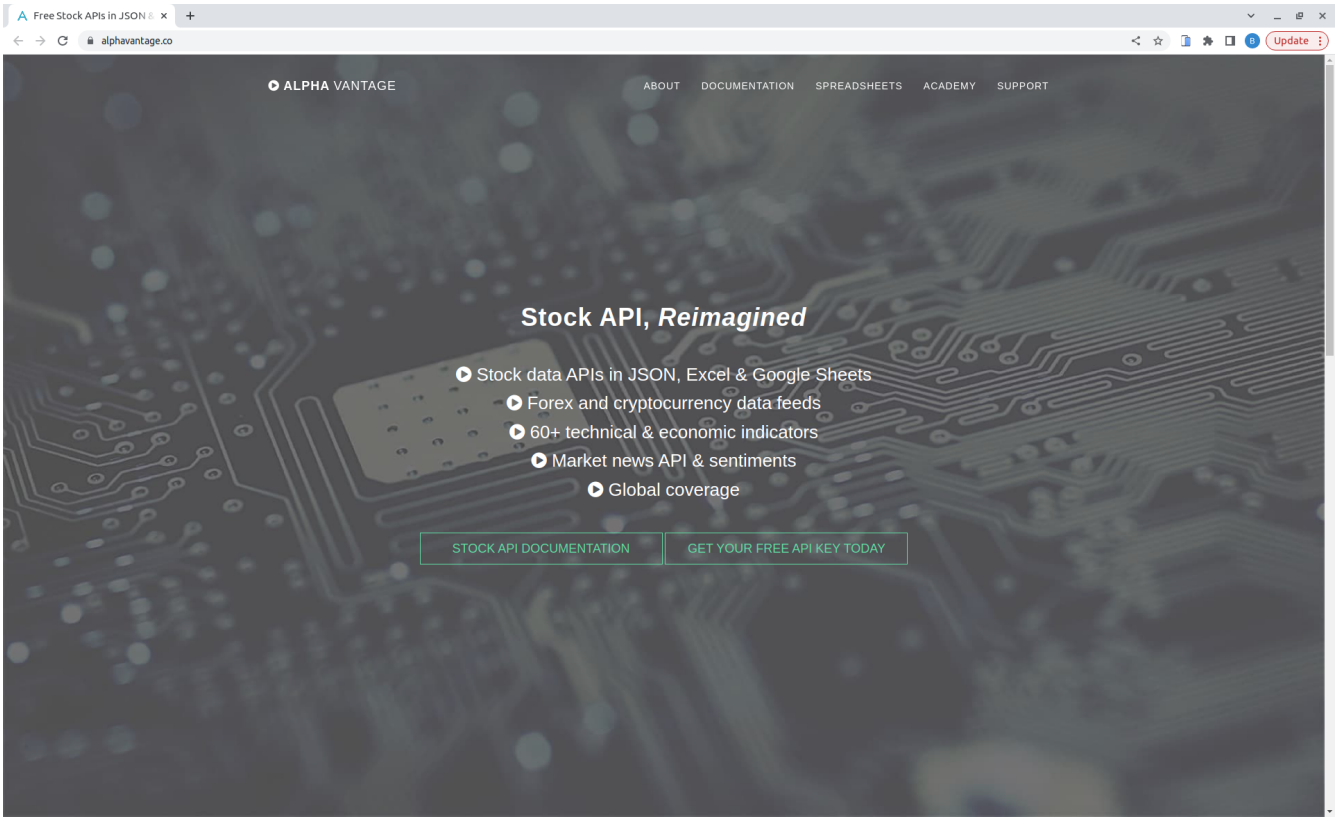
CDP Data Visualization enables data engineers, business analysts, and data scientists to quickly and easily explore data, collaborate, and share insights across the data lifecycle—from data ingest to data insights and beyond. Delivered natively as part of Cloudera Data Platform (CDP), Data Visualization delivers a consistent and easy to use data visualization experience with intuitive and accessible drag-and-drop dashboards and custom application creation.

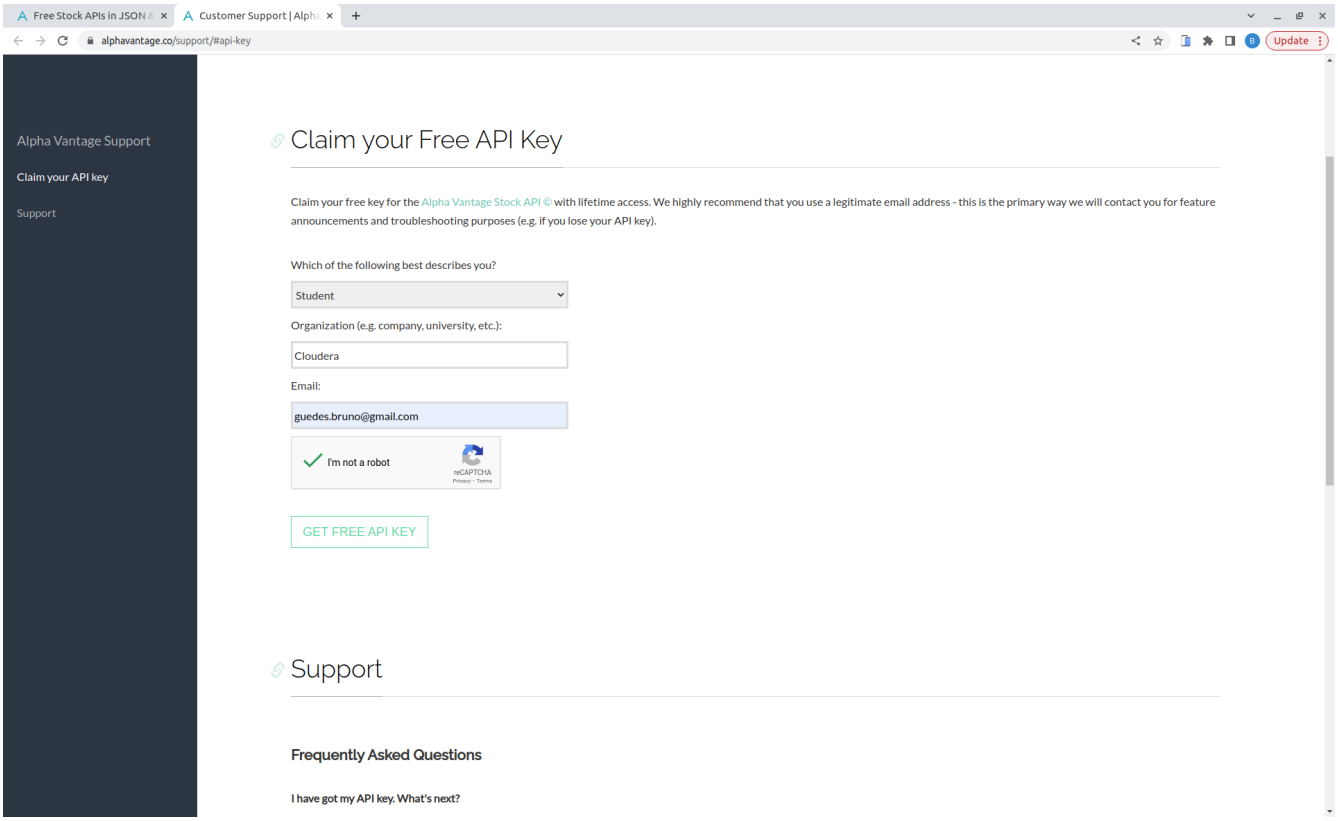


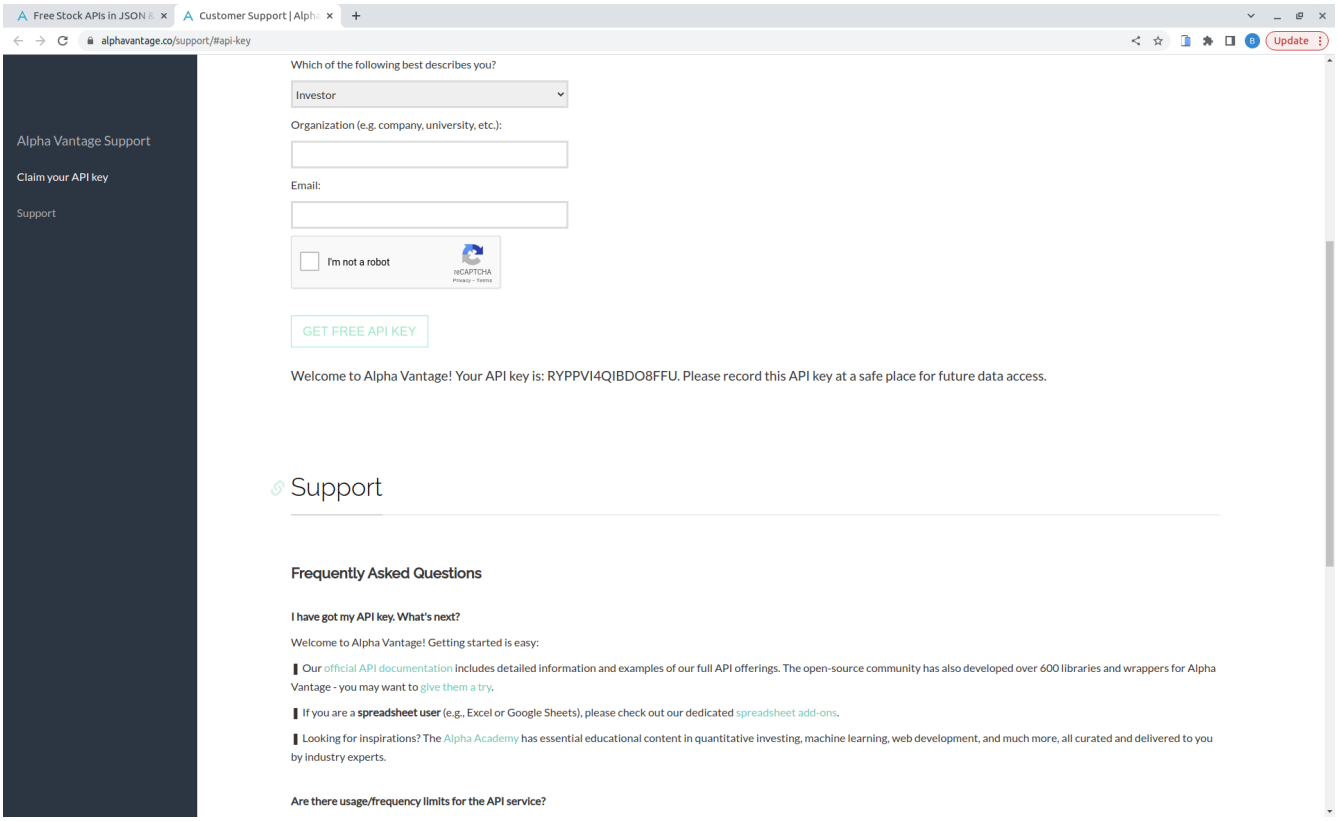
Pre-requisites

1. Laptop with a supported OS (Windows 7 not supported) or Macbook.
2. A modern browser - Google Chrome (IE, Firefox, Safari not supported).
3. Wifi Internet connection.

Step 1: Get Alpha Vantage Key

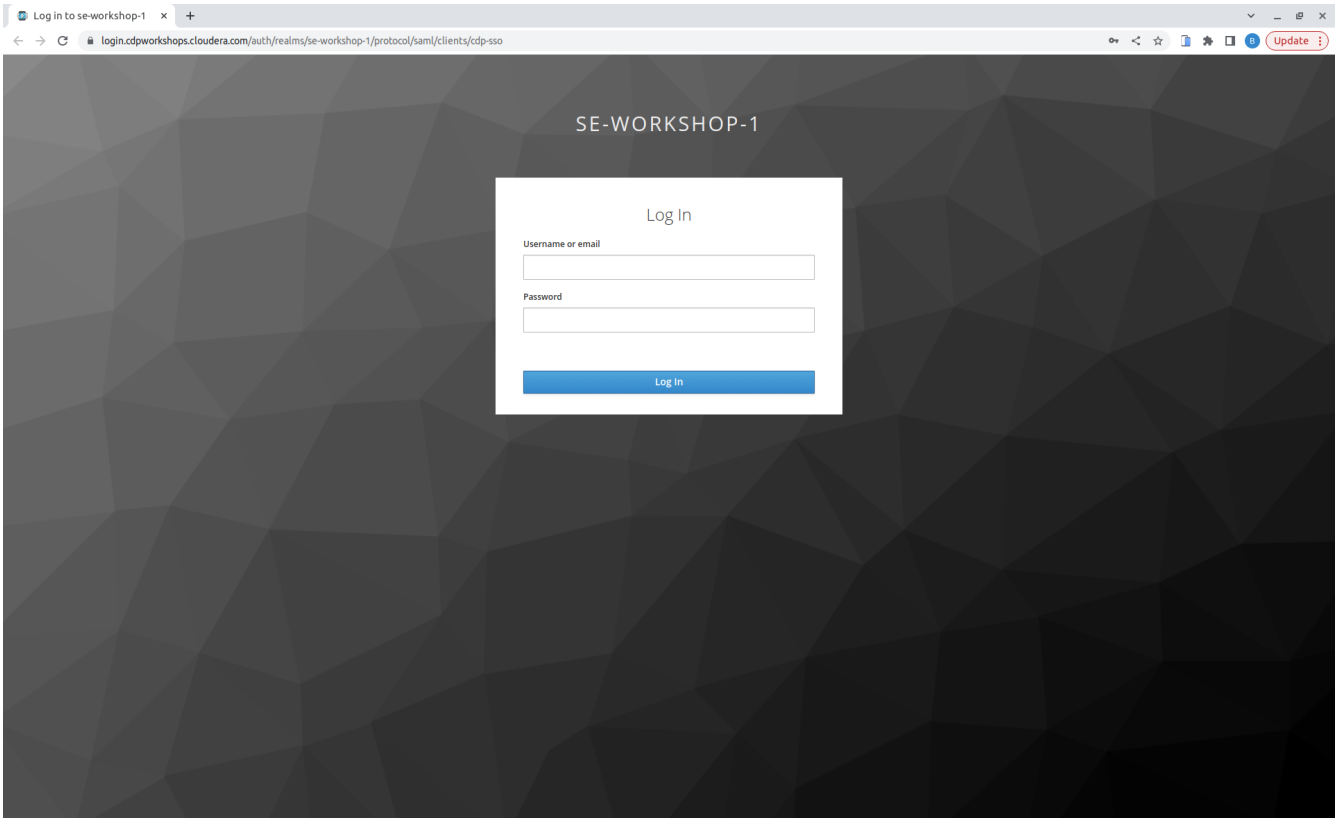

Free Stock APIs in JSON
alphavantage.co
Update
ALPHA VANTAGE
ABOUT
DOCUMENTATION
SPREADSHEETS
ACADEMY
SUPPORT
Stock API, Reimagined
Stock data APIs in JSON, Excel & Google Sheets
Forex and cryptocurrency data feeds
60+ technical & economic indicators
Market news API & sentiments
Global coverage
STOCK API DOCUMENTATION
GET YOUR FREE API KEY TODAY


Free Stock APIs in JSON
Customer Support | Alpha
alphavantage.co/support/#api-key
Update
Alpha Vantage Support
Claim your API key
Support
Claim your Free API Key
Claim your free key for the Alpha Vantage Stock API © with lifetime access. We highly recommend that you use a legitimate email address - this is the primary way we will contact you for feature announcements and troubleshooting purposes (e.g. if you lose your API key).
Which of the following best describes you?
Student
Organization (e.g. company, university, etc.):
Cloudera
Email:
guedes.bruno@gmail.com
I'm not a robot
reCAPTCHA
GET FREE API KEY
Support
Frequently Asked Questions
I have got my API key. What's next?

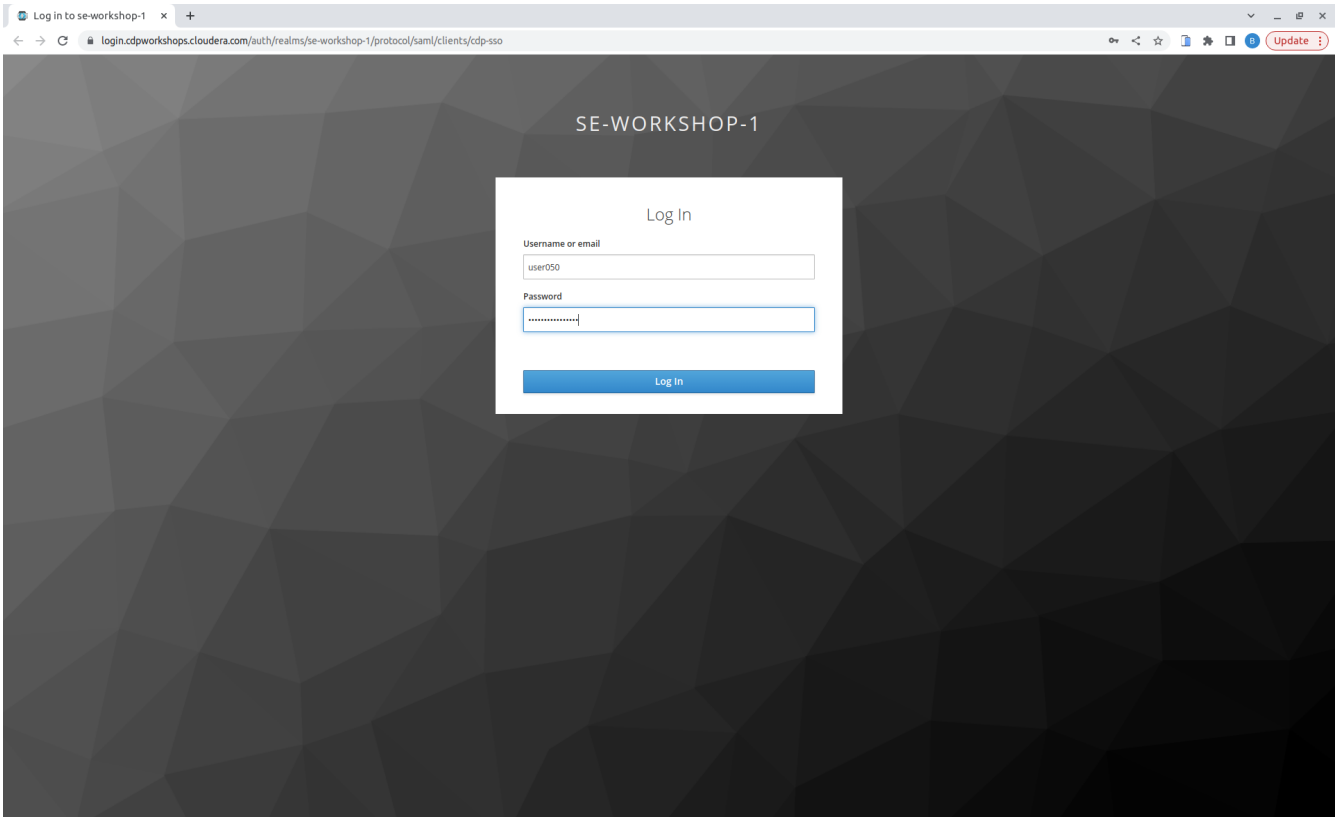


Step 2: Access CDP Public Cloud Portal

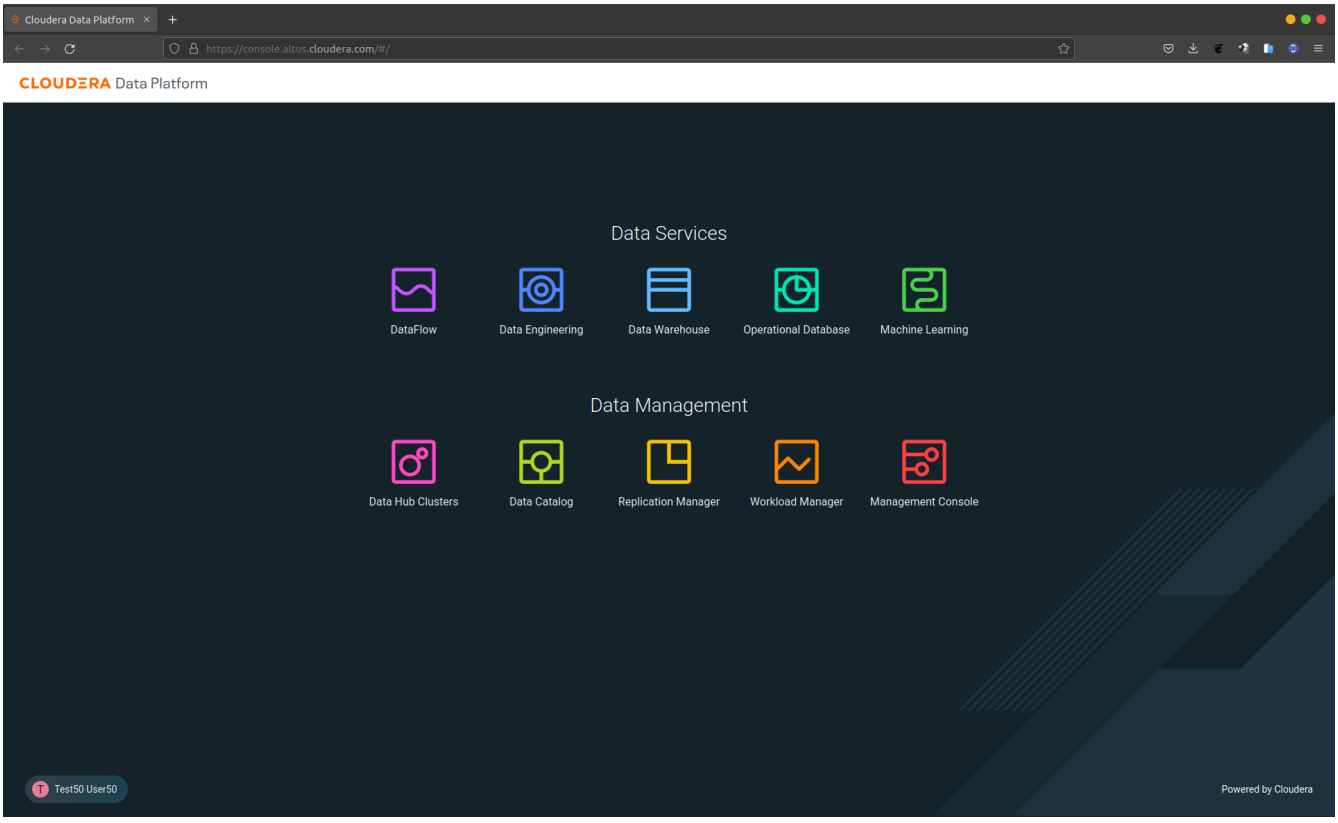
Please use the login url Workshop login



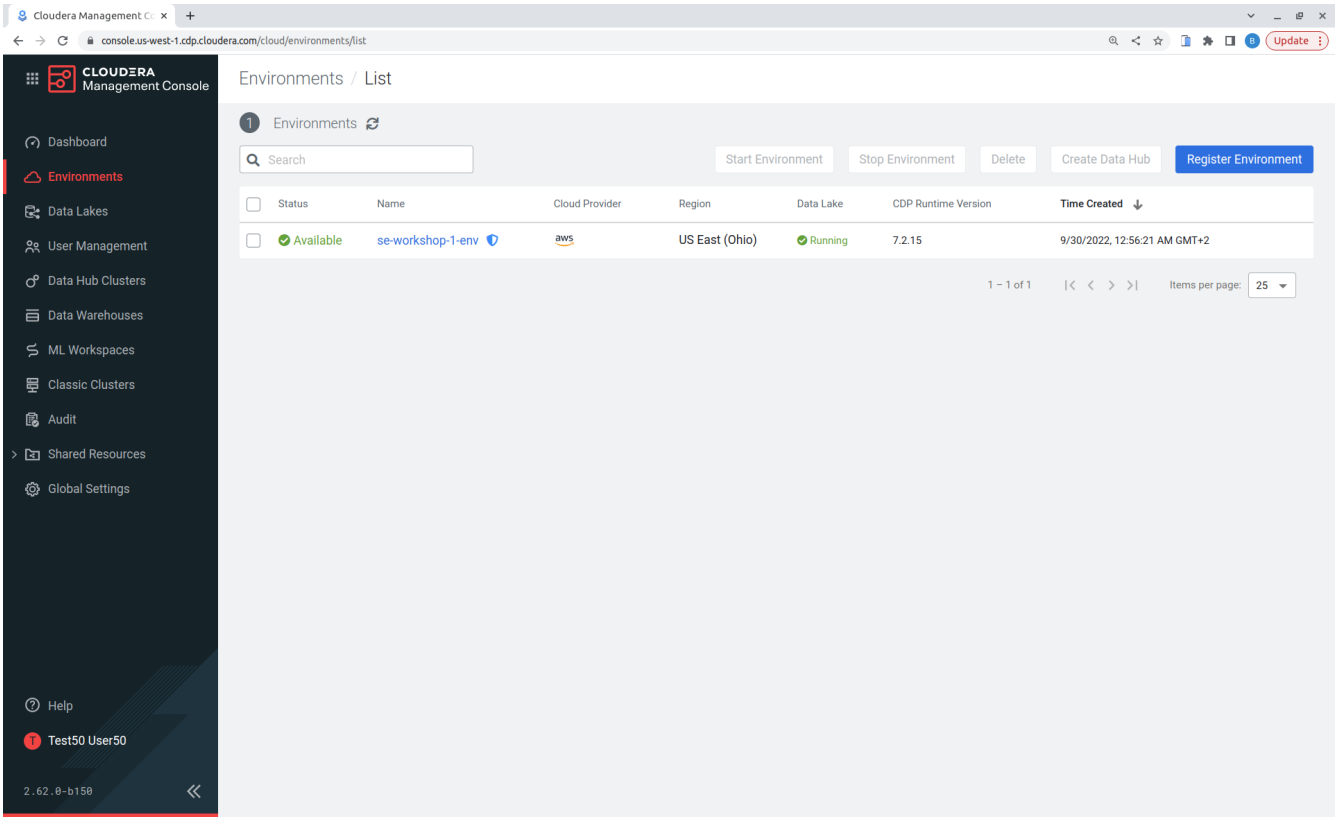
Enter the username and password shared by your instructor.

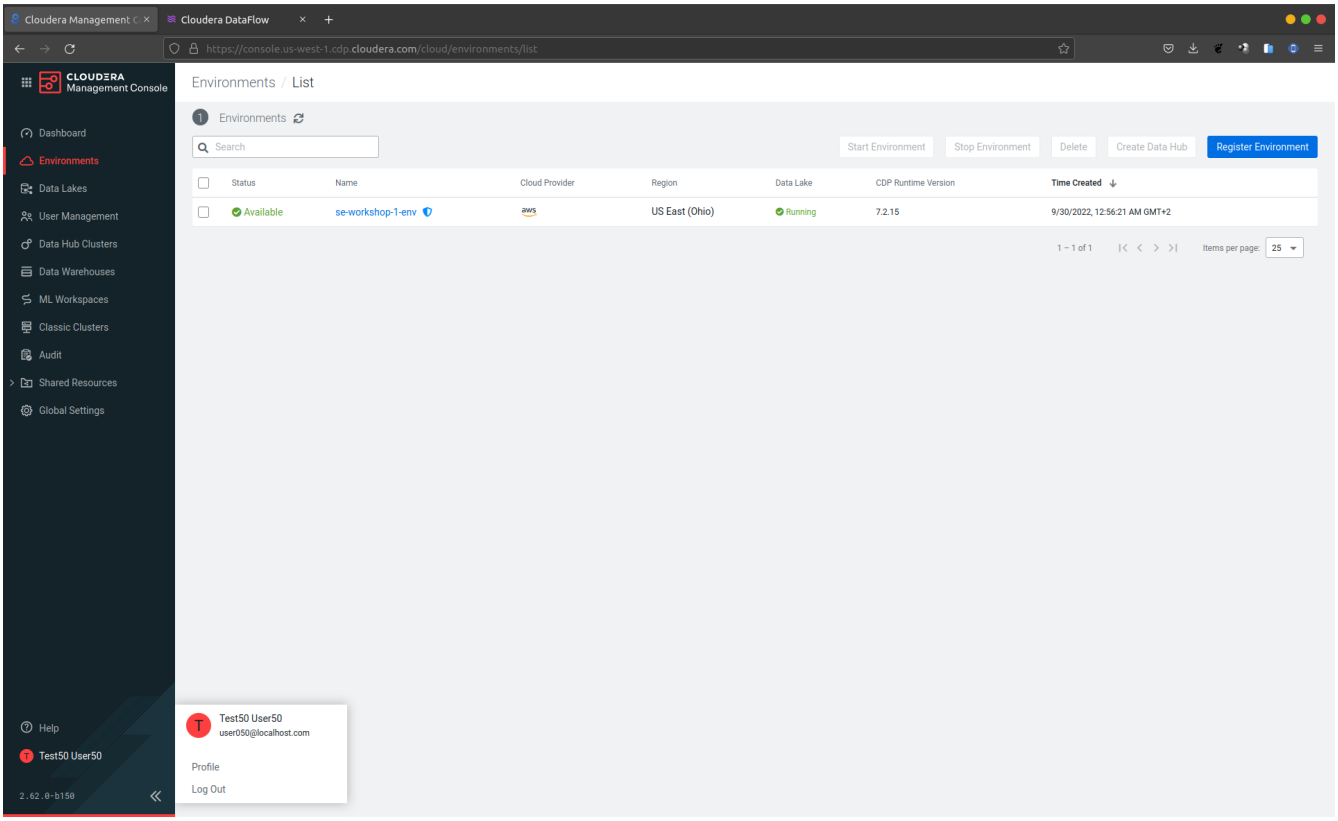


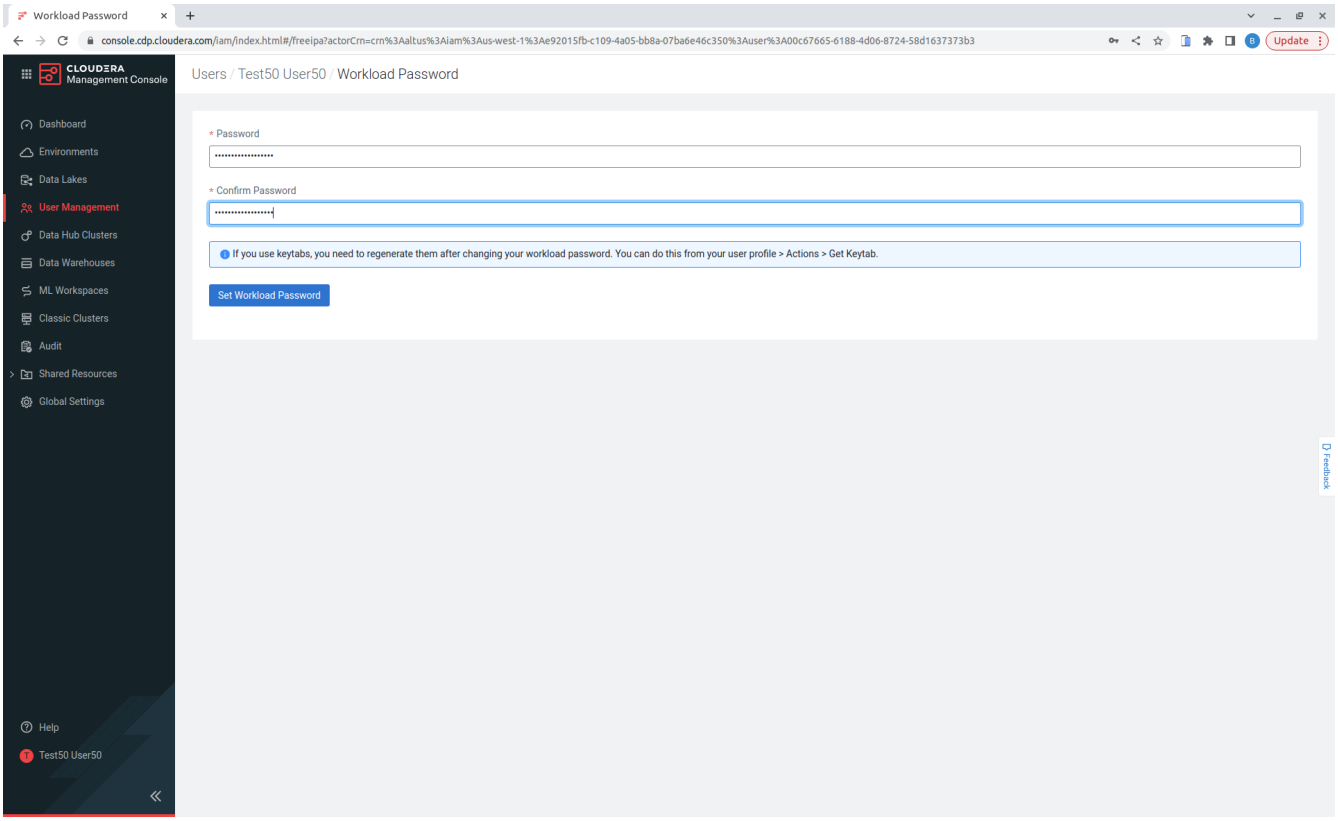
You should be able to get the following home page of CDP Public Cloud.

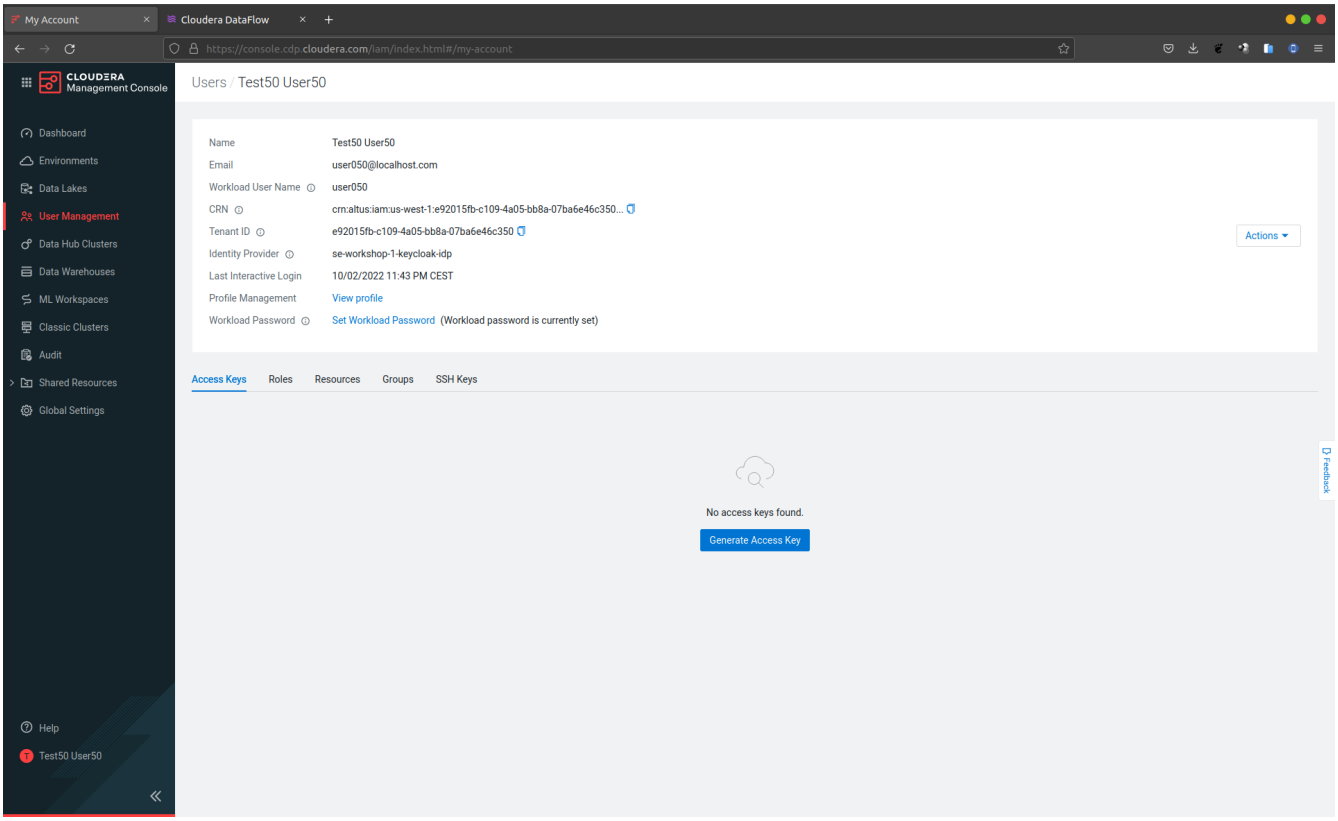


Step 3: Define Workload Password

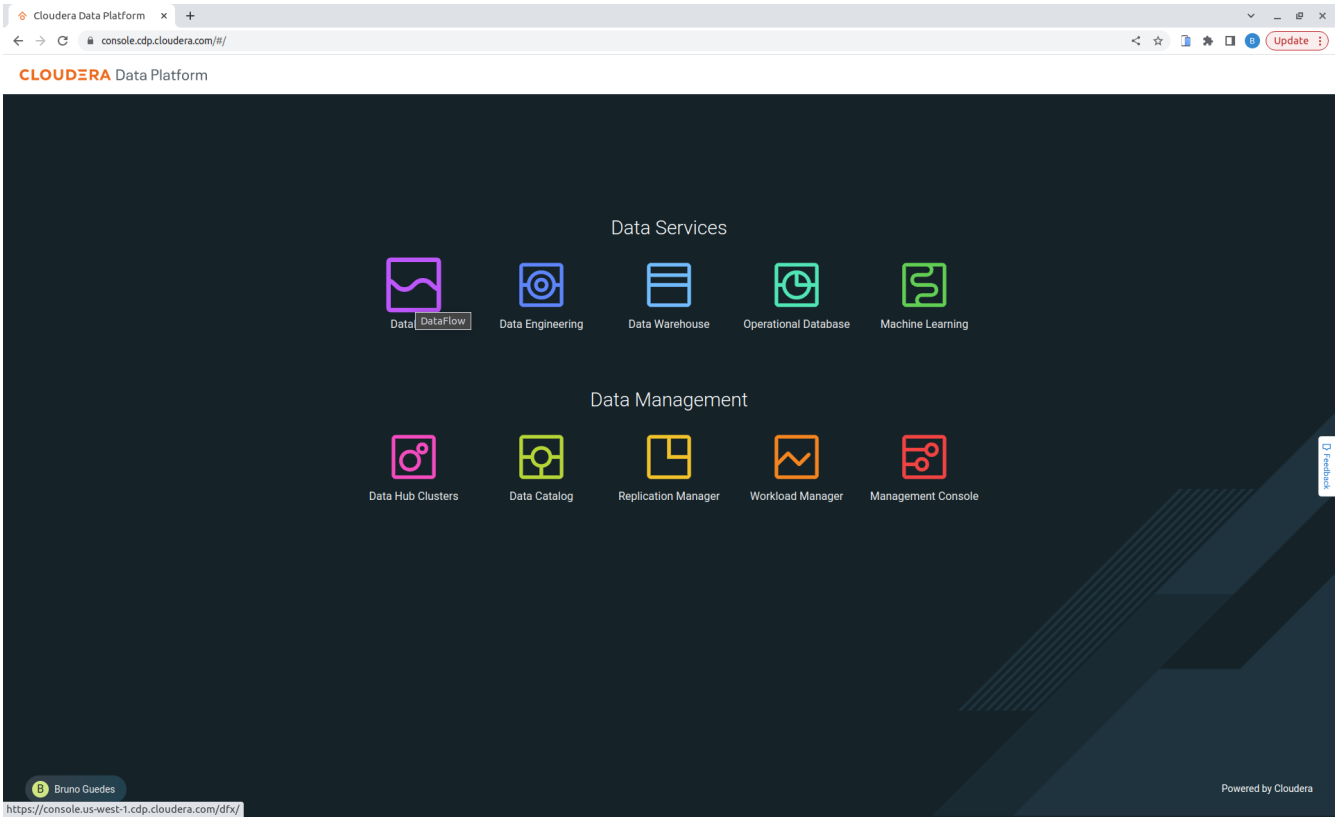

Cloudera Management C
console.us-west-1.cdp.cloudera.com/cloud/environments/list
Update
CLOUDERA
Management Console
Environments / List
Dashboard
Environments
Data Lakes
User Management
Data Hub Clusters
Data Warehouses
ML Workspaces
Classic Clusters
Audit
Shared Resources
Global Settings
Help
Test50 User50
2.62.0-b150
Environments
Search
Start Environment
Stop Environment
Delete
Create Data Hub
Register Environment
Status
Name
Cloud Provider
Region
Data Lake
CDP Runtime Version
Time Created
Available
se-workshop-1-env
aws
US East (Ohio)
Running
7.2.15
9/30/2022, 12:56:21 AM GMT+2
1 – 1 of 1
Items per page:
25


Cloudera Management C
Cloudera DataFlow
https://console.us-west-1.cdp.cloudera.com/cloud/environments/list
CLOUDERA
Management Console
Environments / List
Dashboard
Environments
Data Lakes
User Management
Data Hub Clusters
Data Warehouses
ML Workspaces
Classic Clusters
Audit
Shared Resources
Global Settings
Help
Test50 User50
2.62.0-b150
Test50 User50
user050@localhost.com
Profile
Log Out
Environments
Search
Start Environment
Stop Environment
Delete
Create Data Hub
Register Environment
Status
Name
Cloud Provider
Region
Data Lake
CDP Runtime Version
Time Created
Available
se-workshop-1-env
aws
US East (Ohio)
Running
7.2.15
9/30/2022, 12:56:21 AM GMT+2
1 – 1 of 1
Items per page:
25

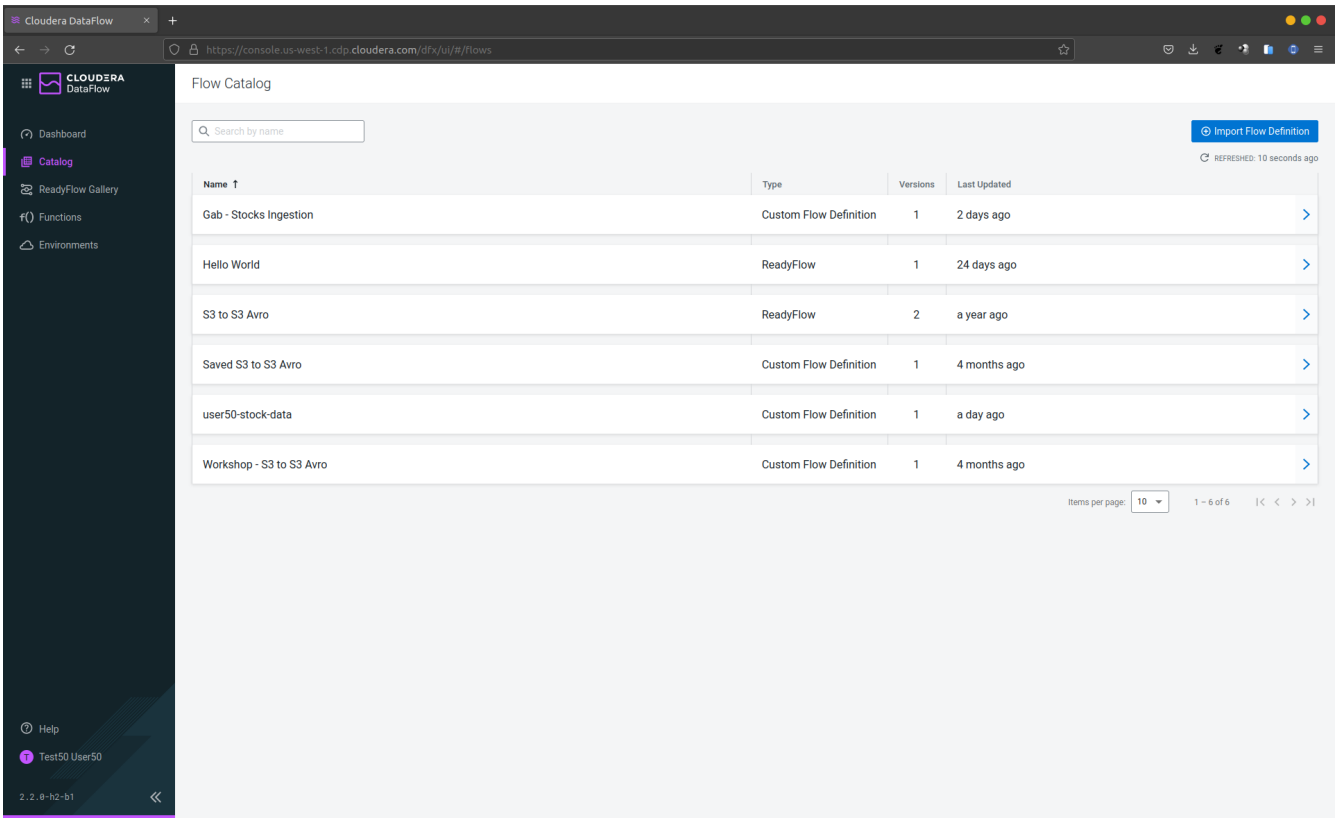


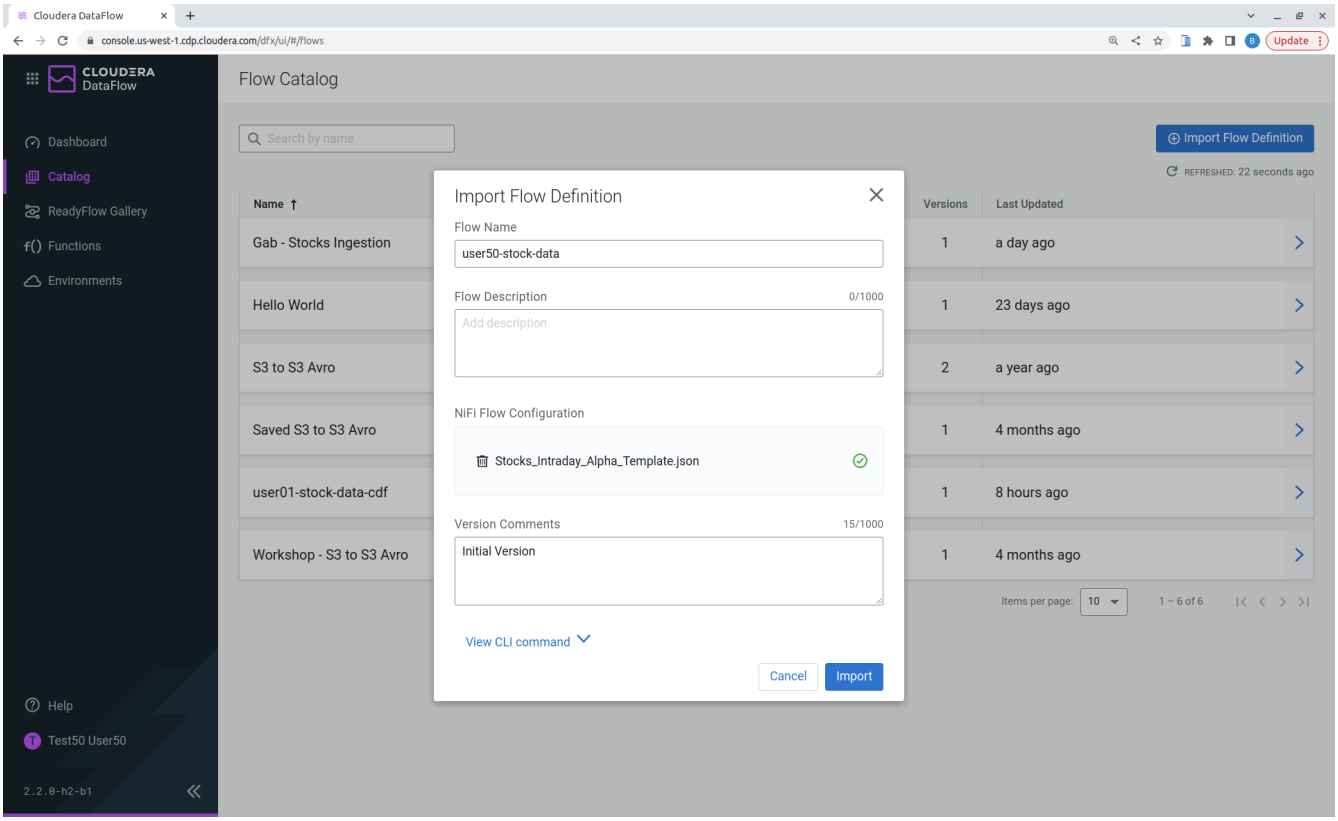


Step 4: Create the flow to ingest stock data via API to Object Storage

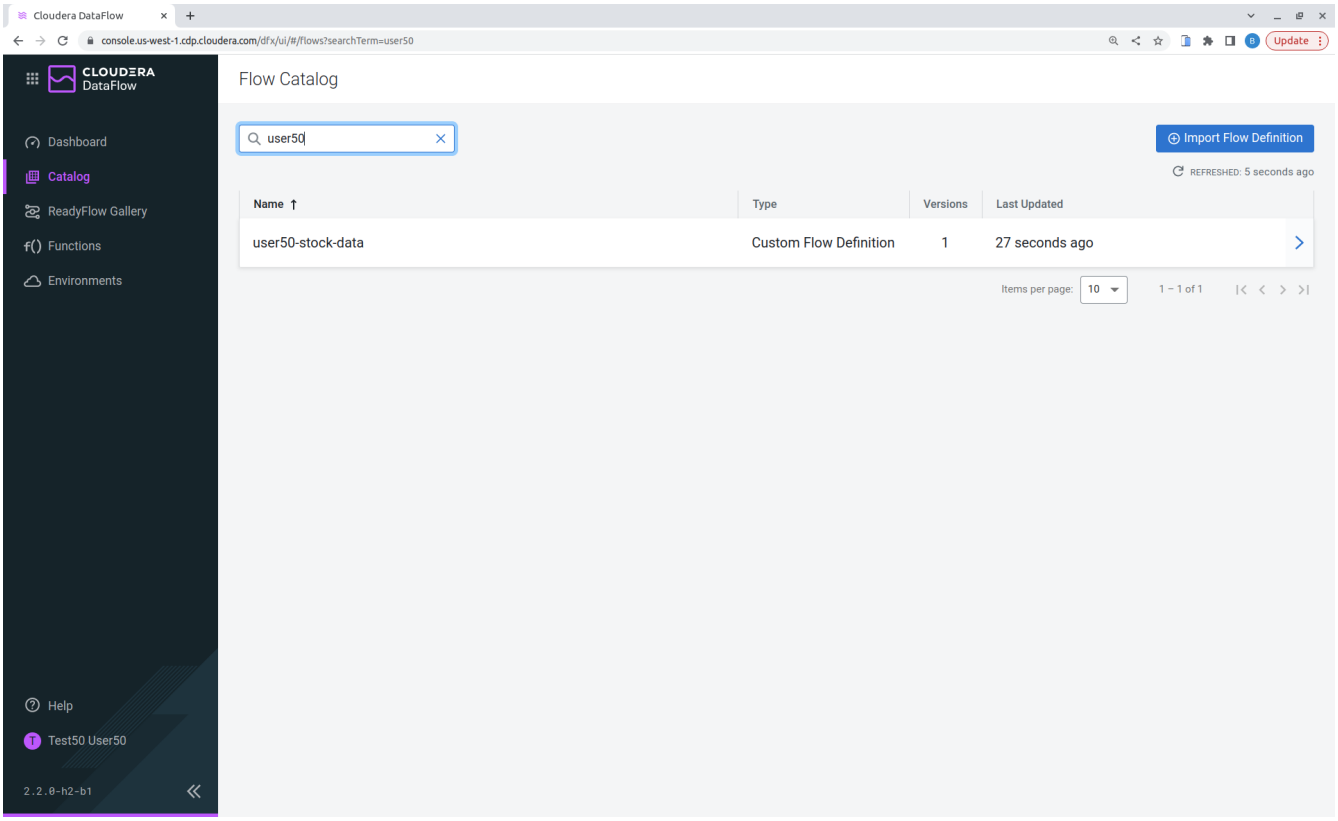


Create a new CDF Catalog

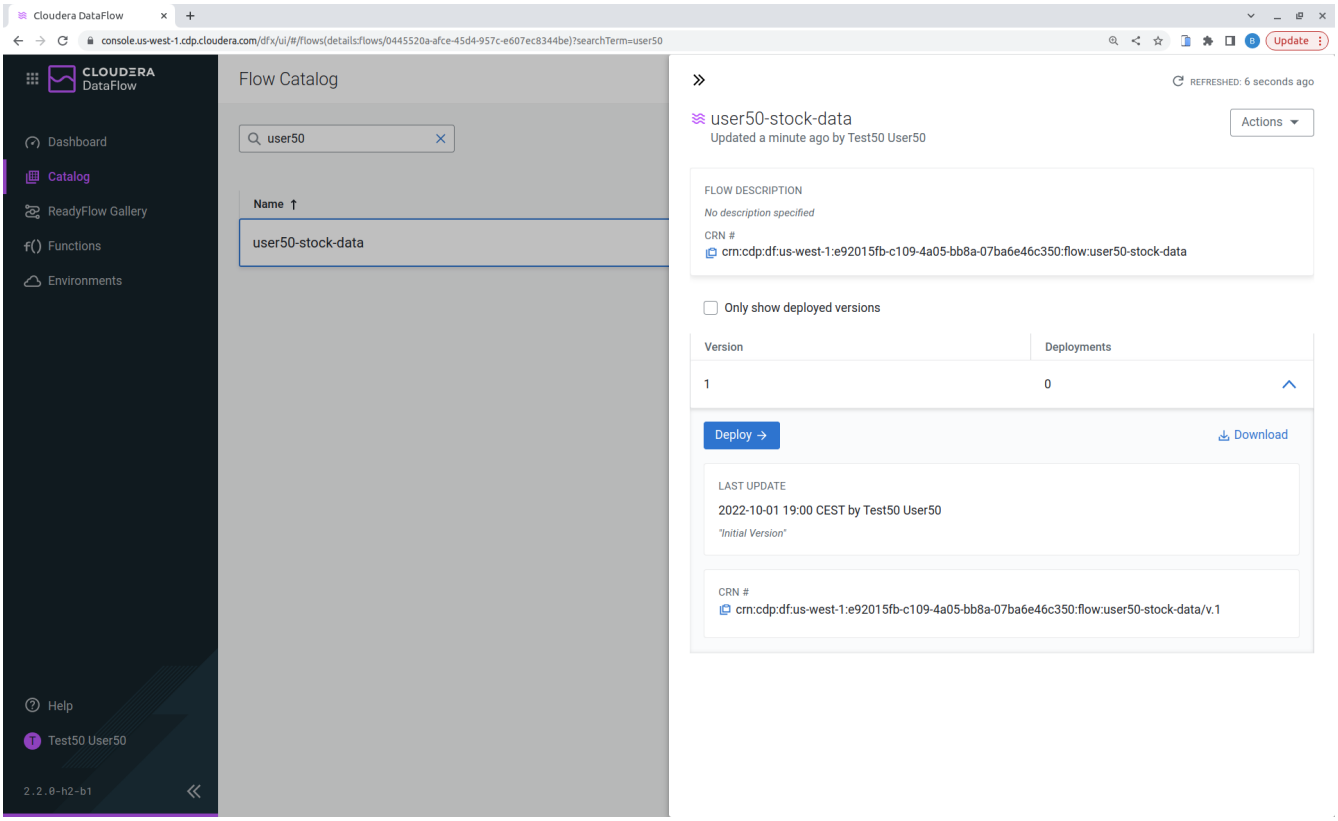


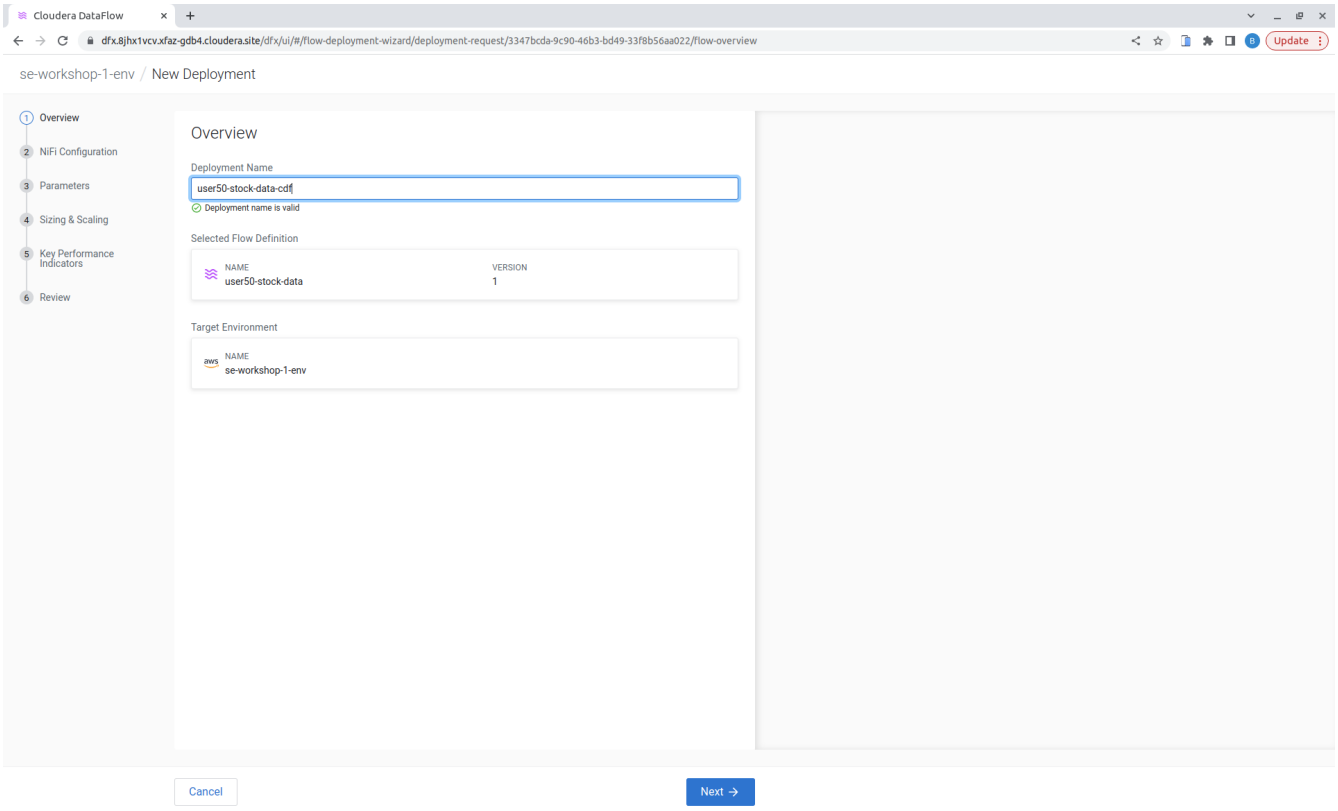


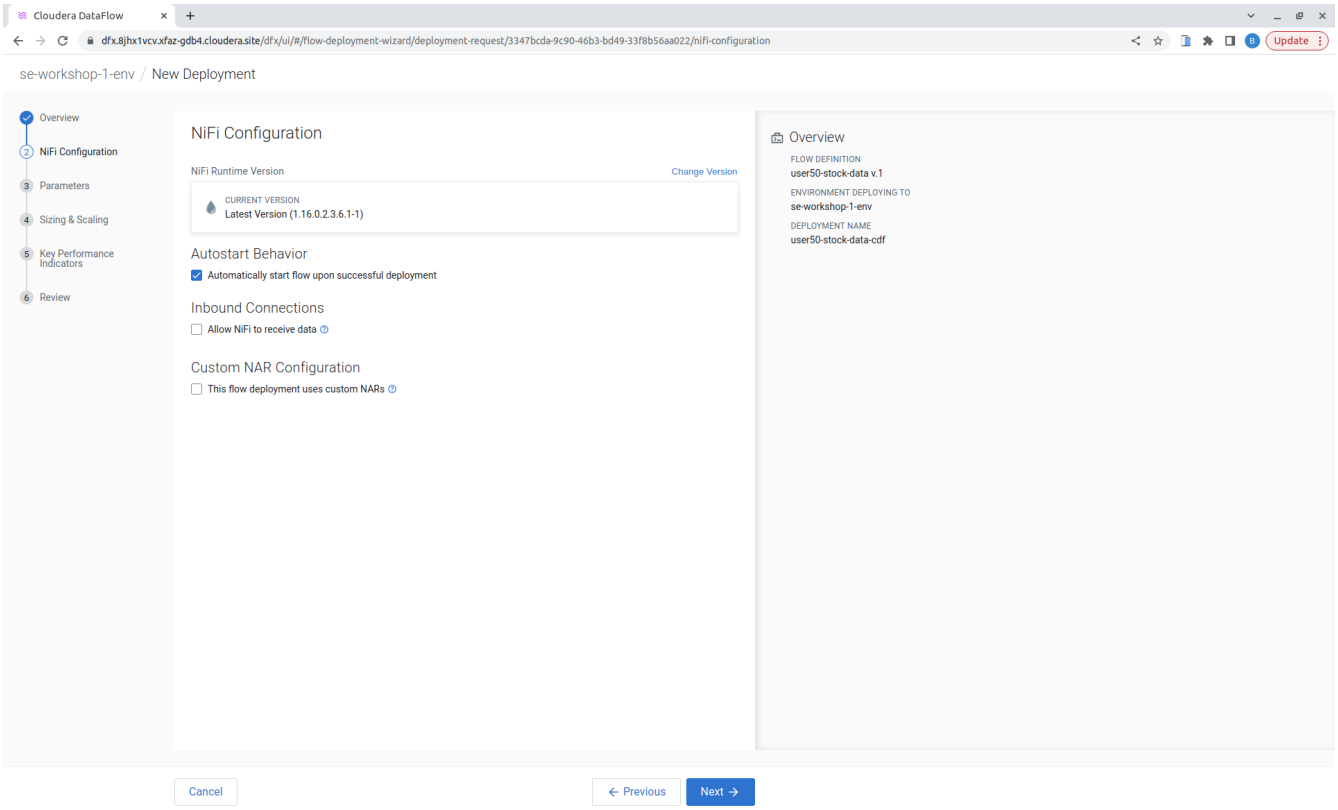
Stocks_Intraday_Alpha_Template.json

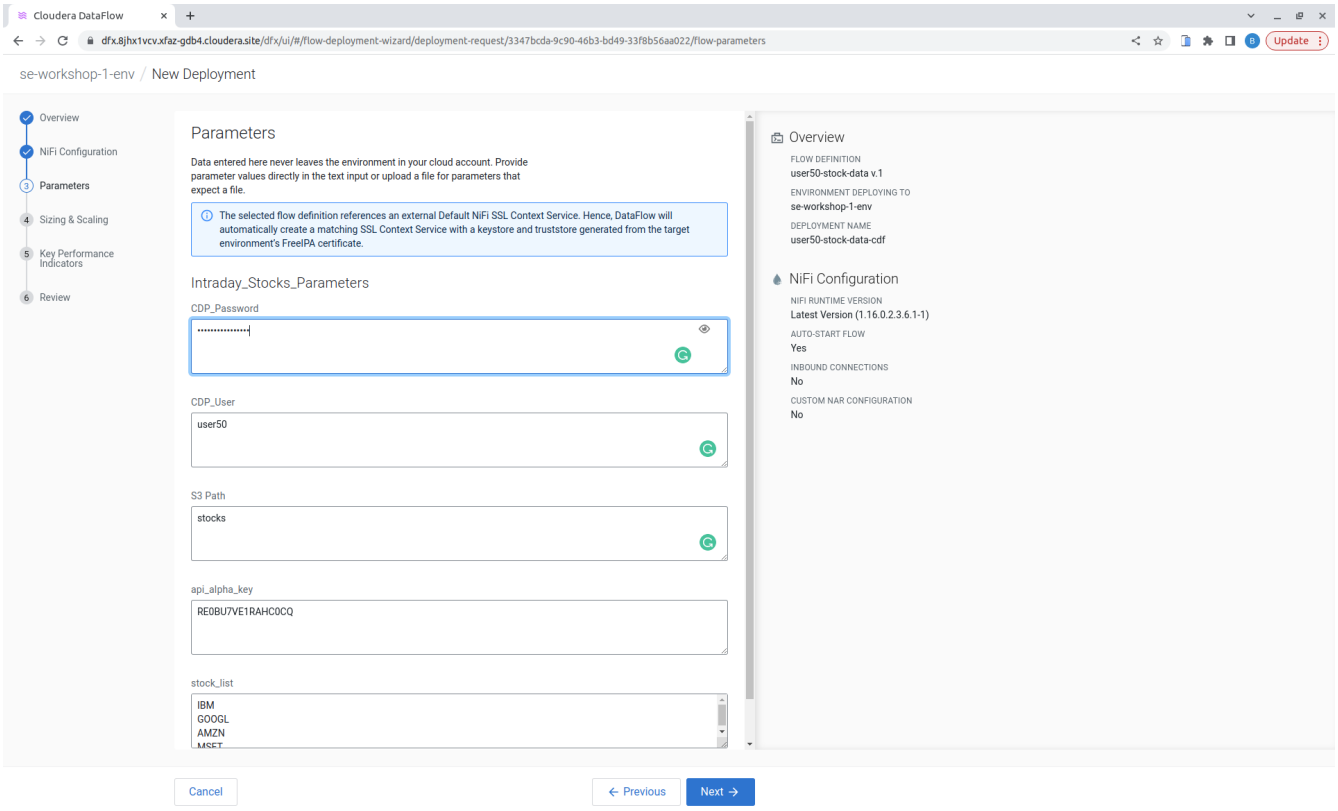


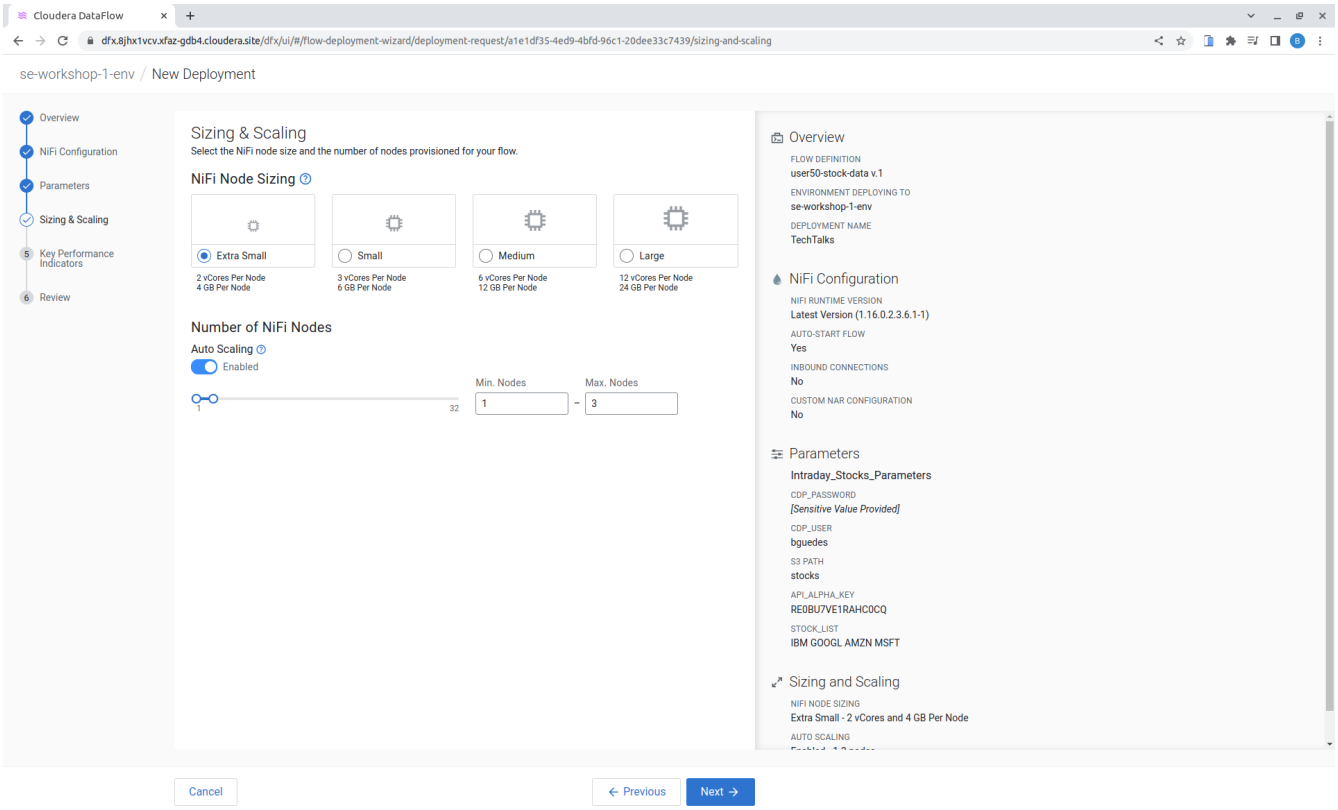
Deploy DataFlow

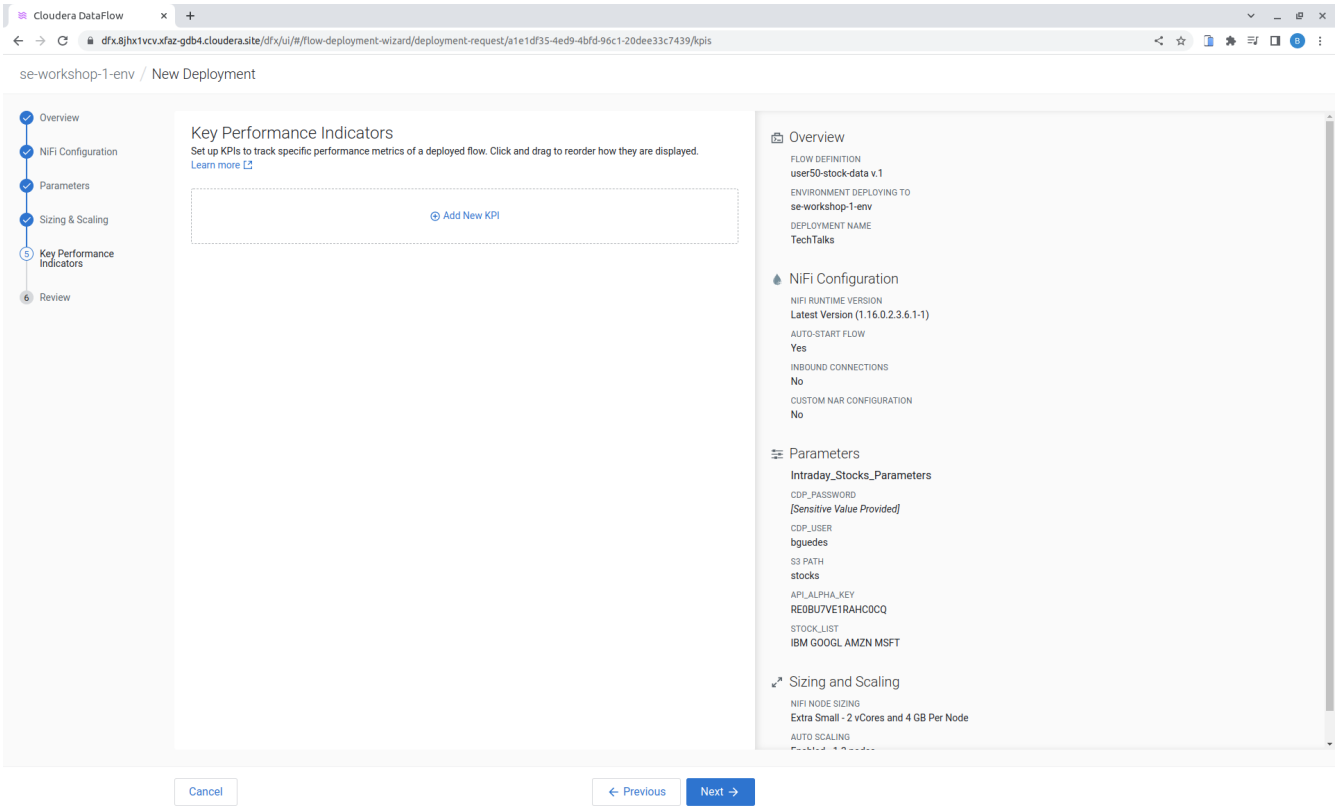

Cloudera DataFlow
console.us-west-1.cdp.cloudera.com/dfx/ui/#/flows(details:flows/0445520a-afce-45d4-957c-e607ec8344be)?searchTerm=user50
Update
CLOUDERA DataFlow
Flow Catalog
Dashboard
Catalog
ReadyFlow Gallery
Functions
Environments
user50
Name
user50-stock-data
Help
Test50 User50
2.2.0-h2-b1
REFRESHED: 6 seconds ago
user50-stock-data
Updated a minute ago by Test50 User50
Actions
FLOW DESCRIPTION
No description specified
CRN #
crn:cdp:df:us-west-1:e92015fb-c109-4a05-bb8a-07ba6e46c350:flow:user50-stock-data
Only show deployed versions
Version
Deployments
1
0
Deploy
Download
LAST UPDATE
2022-10-01 19:00 CEST by Test50 User50
"Initial Version"
CRN #
crn:cdp:df:us-west-1:e92015fb-c109-4a05-bb8a-07ba6e46c350:flow:user50-stock-data/v.1

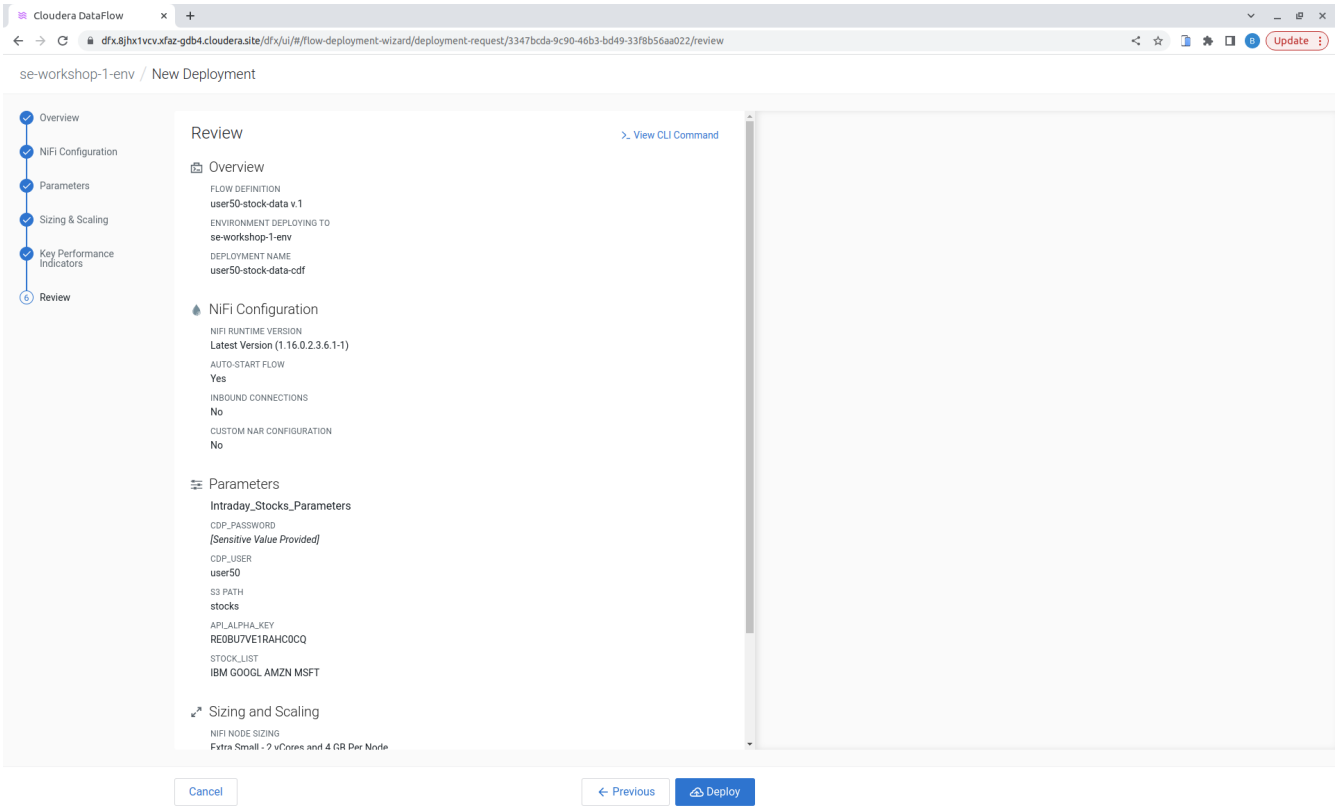

Cloudera DataFlow
dfx.8jhx1vcv.xfaz-gdb4.cloudera.site/dfx/ui/#/flow-deployment-wizard/deployment-request/3347bcda-9c90-46b3-bd49-33f8b56aa022/flow-overview
Update
se-workshop-1-env / New Deployment
1 Overview
2 NiFi Configuration
3 Parameters
4 Sizing & Scaling
5 Key Performance Indicators
6 Review
Overview
Deployment Name
user50-stock-data-cdf
Deployment name is valid
Selected Flow Definition
NAME
user50-stock-data
VERSION
1
Target Environment
NAME
se-workshop-1-env
Cancel
Next

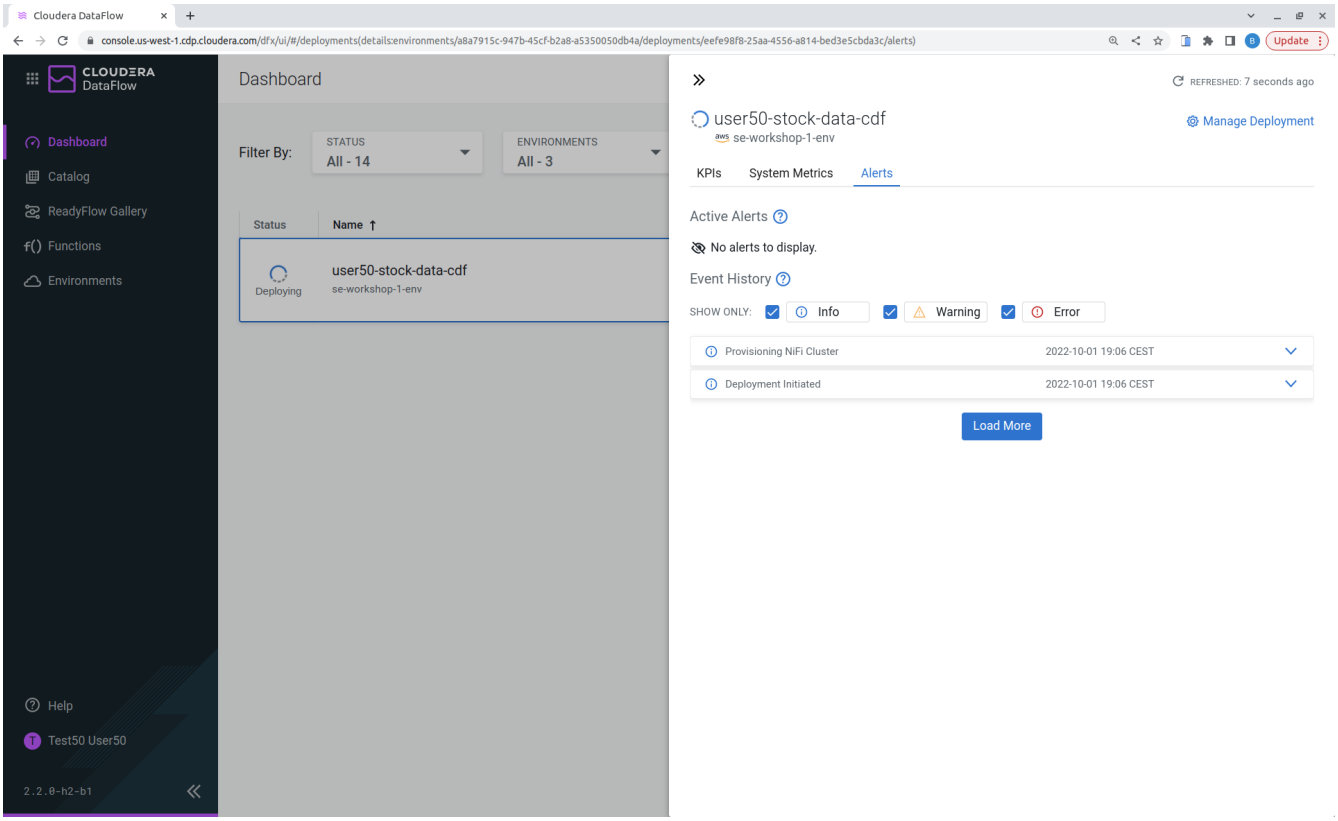

Cloudera DataFlow
dfx.8jhx1vcv.xfaz-gdb4.cloudera.site/dfx/ui/#/flow-deployment-wizard/deployment-request/3347bcda-9c90-46b3-bd49-33f8b56aa022/nifi-configuration
Update
se-workshop-1-env / New Deployment
Overview
NiFi Configuration
Parameters
Sizing & Scaling
Key Performance Indicators
Review
NiFi Configuration
NiFi Runtime Version
Change Version
CURRENT VERSION
Latest Version (1.16.0.2.3.6.1-1)
Autostart Behavior
Automatically start flow upon successful deployment
Inbound Connections
Allow NiFi to receive data
Custom NAR Configuration
This flow deployment uses custom NARs
Overview
FLOW DEFINITION
user50-stock-data v.1
ENVIRONMENT DEPLOYING TO
se-workshop-1-env
DEPLOYMENT NAME
user50-stock-data-cdf
Cancel
← Previous
Next →

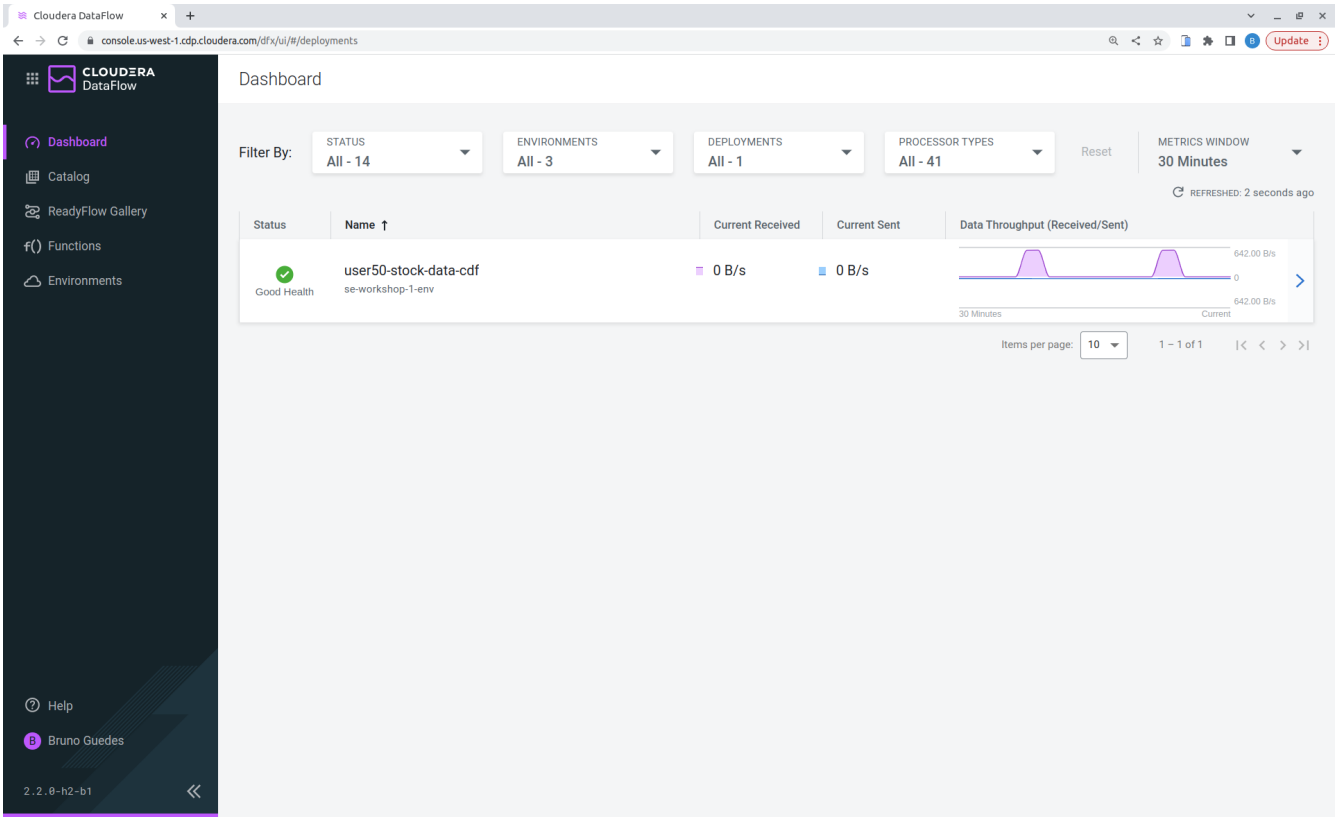

Cloudera DataFlow
dfx.8jhx1vcv.xfaz-gdb4.cloudera.site/dfx/ui/#/flow-deployment-wizard/deployment-request/3347bcda-9c90-46b3-bd49-33f8b56aa022/flow-parameters
Update
se-workshop-1-env / New Deployment
Overview
NiFi Configuration
Parameters
Sizing & Scaling
Key Performance Indicators
Review
Parameters
Data entered here never leaves the environment in your cloud account. Provide parameter values directly in the text input or upload a file for parameters that expect a file.
The selected flow definition references an external Default NiFi SSL Context Service. Hence, DataFlow will automatically create a matching SSL Context Service with a keystore and truststore generated from the target environment's FreeIPA certificate.
Intraday_Stocks_Parameters
CDP_Password
CDP_User
user50
S3 Path
stocks
api_alpha_key
RE0BU7VE1RAHC0CQ
stock_list
IBM
GOOGL
AMZN
Overview
FLOW DEFINITION
user50-stock-data v.1
ENVIRONMENT DEPLOYING TO
se-workshop-1-env
DEPLOYMENT NAME
user50-stock-data-cdf
NiFi Configuration
NIFI RUNTIME VERSION
Latest Version (1.16.0.2.3.6.1-1)
AUTO-START FLOW
Yes
INBOUND CONNECTIONS
No
CUSTOM NAR CONFIGURATION
No
Cancel
← Previous
Next →


Cloudera DataFlow
dfx.8jhx1vcv.xfaz-gdb4.cloudera.site/dfx/ui/#/flow-deployment-wizard/deployment-request/a1e1df35-4ed9-4bfd-96c1-20dee33c7439/sizing-and-scaling
se-workshop-1-env / New Deployment
Overview
NiFi Configuration
Parameters
Sizing & Scaling
Key Performance Indicators
Review
Sizing & Scaling
Select the NiFi node size and the number of nodes provisioned for your flow.
NiFi Node Sizing
Extra Small
2 vCores Per Node
4 GB Per Node
Small
3 vCores Per Node
6 GB Per Node
Medium
6 vCores Per Node
12 GB Per Node
Large
12 vCores Per Node
24 GB Per Node
Number of NiFi Nodes
Auto Scaling
Enabled
Min. Nodes
Max. Nodes
1
32
1
3
Overview
FLOW DEFINITION
user50-stock-data v.1
ENVIRONMENT DEPLOYING TO
se-workshop-1-env
DEPLOYMENT NAME
TechTalks
NiFi Configuration
NIFI RUNTIME VERSION
Latest Version (1.16.0.2.3.6.1-1)
AUTO-START FLOW
Yes
INBOUND CONNECTIONS
No
CUSTOM NAR CONFIGURATION
No
Parameters
Intraday_Stocks_Parameters
CDP_PASSWORD
[Sensitive Value Provided]
CDP_USER
bguedes
S3 PATH
stocks
API_ALPHA_KEY
RE0BU7VE1RAHC0CQ
STOCK_LIST
IBM GOOGL AMZN MSFT
Sizing and Scaling
NIFI NODE SIZING
Extra Small - 2 vCores and 4 GB Per Node
AUTO SCALING
Cancel
Previous
Next

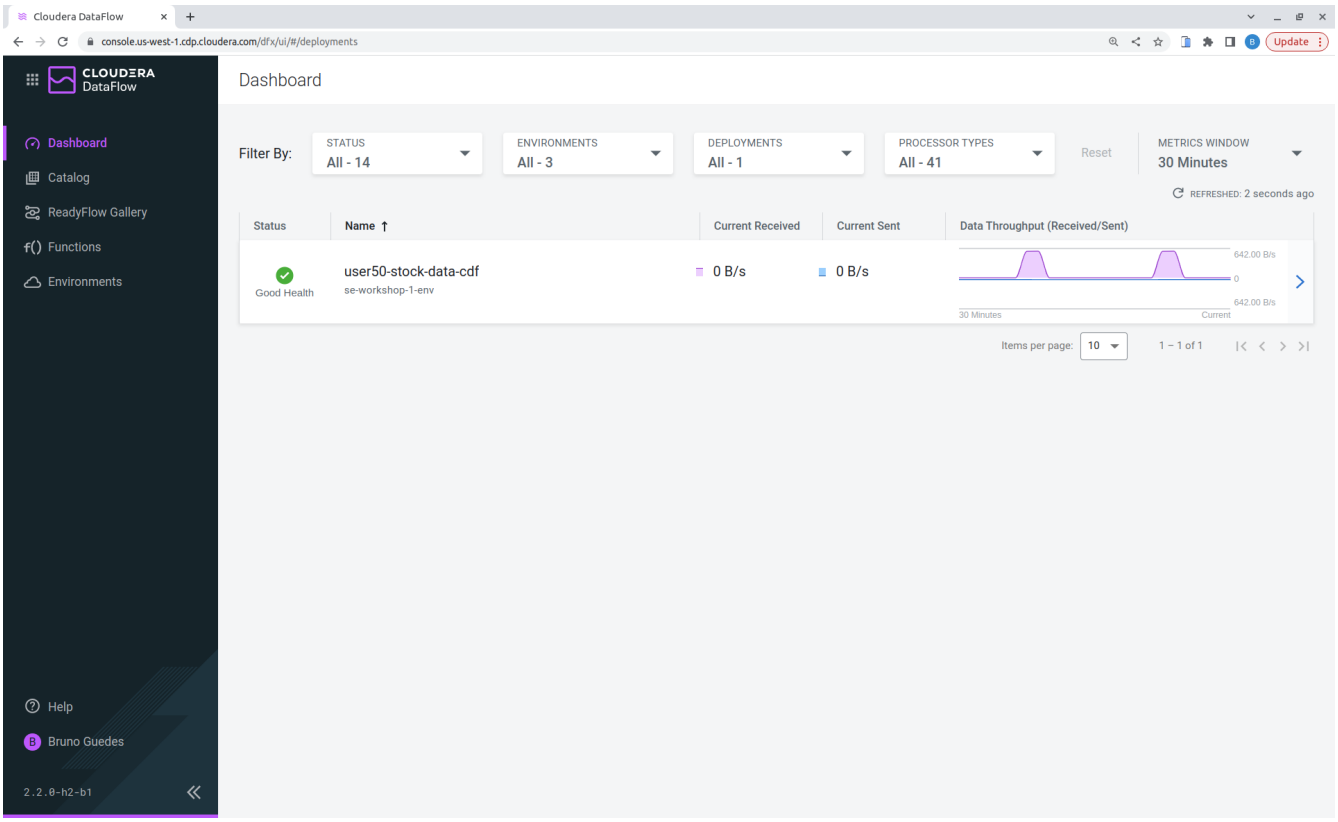

Cloudera DataFlow
dfx.8jhx1vcv.xfaz-gdb4.cloudera.site/dfx/ui/#/flow-deployment-wizard/deployment-request/a1e1df35-4ed9-4bfd-96c1-20dee33c7439/kpis
se-workshop-1-env / New Deployment
Overview
NiFi Configuration
Parameters
Sizing & Scaling
Key Performance Indicators
Review
Key Performance Indicators
Set up KPIs to track specific performance metrics of a deployed flow. Click and drag to reorder how they are displayed.
Learn more
Add New KPI
Overview
FLOW DEFINITION
user50-stock-data v.1
ENVIRONMENT DEPLOYING TO
se-workshop-1-env
DEPLOYMENT NAME
TechTalks
NiFi Configuration
NIFI RUNTIME VERSION
Latest Version (1.16.0.2.3.6.1-1)
AUTO-START FLOW
Yes
INBOUND CONNECTIONS
No
CUSTOM NAR CONFIGURATION
No
Parameters
Intraday_Stocks_Parameters
CDP_PASSWORD
[Sensitive Value Provided]
CDP_USER
bguedes
S3 PATH
stocks
API_ALPHA_KEY
RE0BU7VE1RAHC0CQ
STOCK_LIST
IBM GOOGL AMZN MSFT
Sizing and Scaling
NIFI NODE SIZING
Extra Small - 2 vCores and 4 GB Per Node
AUTO SCALING
Cancel
Previous
Next

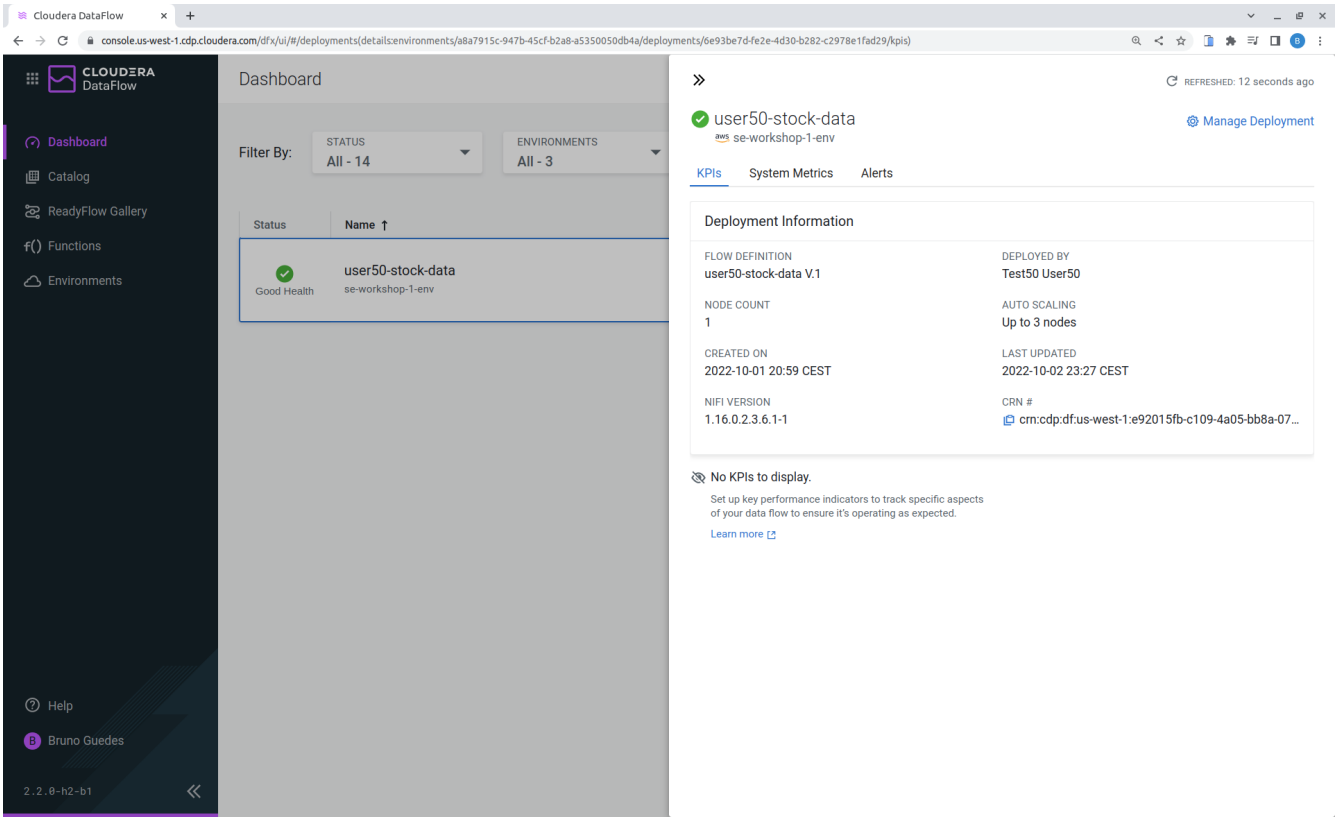

Cloudera DataFlow
dfx.8jhx1vcv.xfaz-gdb4.cloudera.site/dfx/ui/#/flow-deployment-wizard/deployment-request/3347bcda-9c90-46b3-bd49-33f8b56aa022/review
Update
se-workshop-1-env / New Deployment
Overview
NiFi Configuration
Parameters
Sizing & Scaling
Key Performance Indicators
Review
Review
View CLI Command
Overview
FLOW DEFINITION
user50-stock-data v.1
ENVIRONMENT DEPLOYING TO
se-workshop-1-env
DEPLOYMENT NAME
user50-stock-data-cdf
NiFi Configuration
NIFI RUNTIME VERSION
Latest Version (1.16.0.2.3.6.1-1)
AUTO-START FLOW
Yes
INBOUND CONNECTIONS
No
CUSTOM NAR CONFIGURATION
No
Parameters
Intraday_Stocks_Parameters
CDP_PASSWORD
[Sensitive Value Provided]
CDP_USER
user50
S3 PATH
stocks
API_ALPHA_KEY
RE0BU7VE1RAHC0CQ
STOCK_LIST
IBM GOOGL AMZN MSFT
Sizing and Scaling
NIFI NODE SIZING
Extra Small - 2 vCores and 4 GB Per Node
Cancel
Previous
Deploy

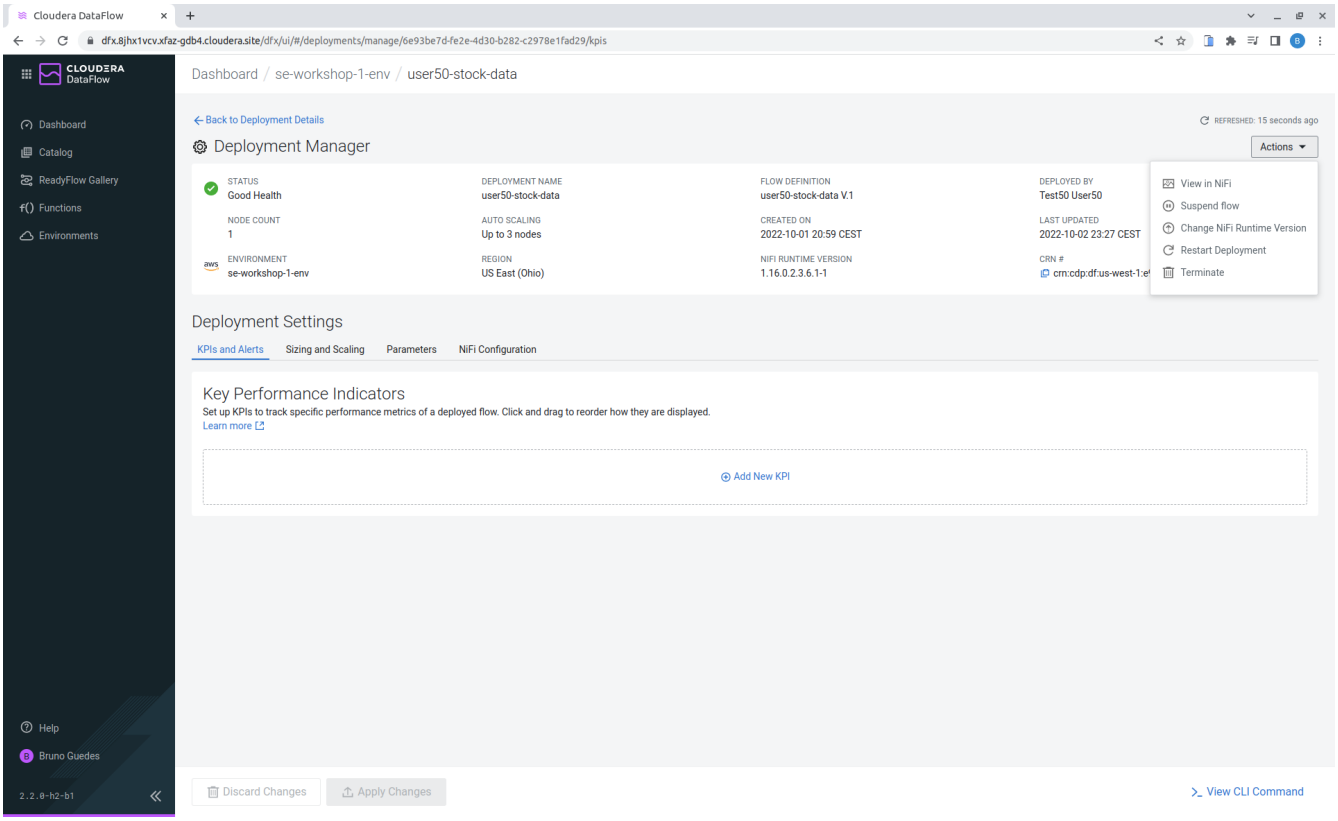

Cloudera DataFlow
console.us-west-1.cdp.cloudera.com/dfx/ui/#/deployments/details:environments/a8a7915c-947b-45cf-b2a8-a5350050db4a/deployments/eefe98f8-25aa-4556-a814-bed3e5cbda3c/alerts
Update
CLOUDERA DataFlow
Dashboard
Catalog
ReadyFlow Gallery
Functions
Environments
Help
Test50 User50
2.2.0-h2-b1
Dashboard
Filter By:
STATUS
All - 14
ENVIRONMENTS
All - 3
Status
Name
Deploying
user50-stock-data-cdf
se-workshop-1-env
REFRESHED: 7 seconds ago
user50-stock-data-cdf
se-workshop-1-env
Manage Deployment
KPIs
System Metrics
Alerts
Active Alerts
No alerts to display.
Event History
SHOW ONLY:
Info
Warning
Error
Provisioning NiFi Cluster
2022-10-01 19:06 CEST
Deployment Initiated
2022-10-01 19:06 CEST
Load More

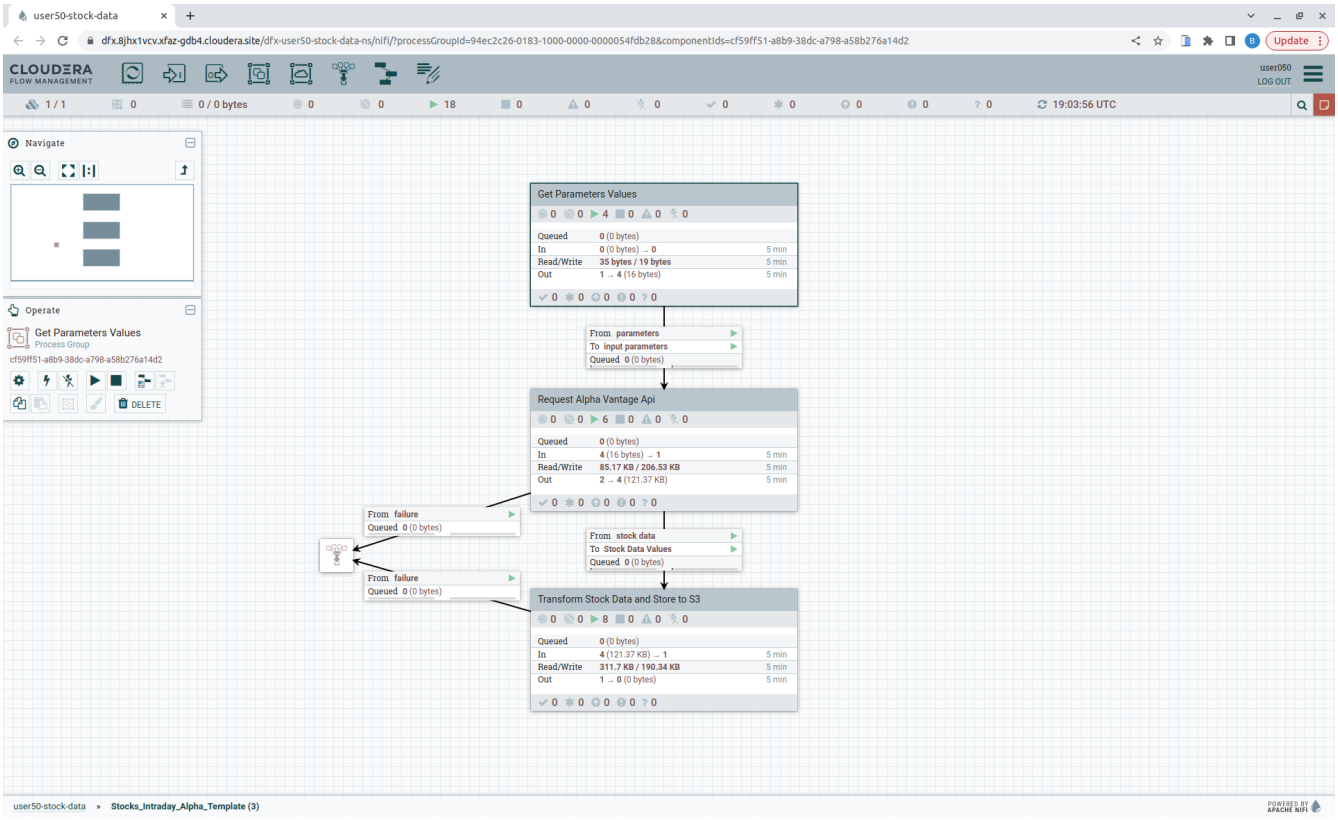


View Nifi DataFlow




Cloudera DataFlow
CLOUDERA DataFlow
Dashboard
Catalog
ReadyFlow Gallery
Functions
Environments
Help
Bruno Guedes
2.2.0-h2-b1
Filter By:
STATUS
All - 14
ENVIRONMENTS
All - 3
Status
Name
Good Health
user50-stock-data
se-workshop-1-env
REFRESHED: 12 seconds ago
user50-stock-data
se-workshop-1-env
Manage Deployment
KPIs
System Metrics
Alerts
Deployment Information
FLOW DEFINITION
user50-stock-data V.1
DEPLOYED BY
Test50 User50
NODE COUNT
1
AUTO SCALING
Up to 3 nodes
CREATED ON
2022-10-01 20:59 CEST
LAST UPDATED
2022-10-02 23:27 CEST
NIFI VERSION
1.16.0.2.3.6.1-1
CRN #
crn:cdp:df:us-west-1:e92015fb-c109-4a05-bb8a-07...
No KPIs to display.
Set up key performance indicators to track specific aspects of your data flow to ensure it's operating as expected.
Learn more


Cloudera DataFlow
dfx.8jhx1vcv.xfaz-gdb4.cloudera.site/dfx/ui/#/deployments/manage/6e93be7d-fe2e-4d30-b282-c2978e1fad29/kpis
CLOUDERA DataFlow
Dashboard
Catalog
ReadyFlow Gallery
Functions
Environments
Help
Bruno Guedes
2.2.0-h2-b1
Dashboard / se-workshop-1-env / user50-stock-data
Back to Deployment Details
REFRESHED: 15 seconds ago
Deployment Manager
Actions
STATUS
Good Health
DEPLOYMENT NAME
user50-stock-data
FLOW DEFINITION
user50-stock-data V.1
DEPLOYED BY
Test50 User50
NODE COUNT
1
AUTO SCALING
Up to 3 nodes
CREATED ON
2022-10-01 20:59 CEST
LAST UPDATED
2022-10-02 23:27 CEST
ENVIRONMENT
se-workshop-1-env
REGION
US East (Ohio)
NIFI RUNTIME VERSION
1.16.0.2.3.6.1-1
CRN #
View in NiFi
Suspend flow
Change NiFi Runtime Version
Restart Deployment
Terminate
Deployment Settings
KPIs and Alerts
Sizing and Scaling
Parameters
NiFi Configuration
Key Performance Indicators
Set up KPIs to track specific performance metrics of a deployed flow. Click and drag to reorder how they are displayed.
Learn more
Add New KPI
Discard Changes
Apply Changes
View CLI Command



Create Iceberg Table

```
CREATE DATABASE stocks;

CREATE TABLE IF NOT EXISTS stocks.stock_intraday_1min (
  interv STRING,
  output_size STRING,
  time_zone STRING,
  open DECIMAL(8,4),
  high DECIMAL(8,4),
  low DECIMAL(8,4),
  close DECIMAL(8,4),
  volume BIGINT)
PARTITIONED BY (
  ticker STRING,
  last_refreshed string,
  refreshed_at string)
STORED AS iceberg;
```

Step 5: Process and Ingest Iceberg using CDE

Step 6: Query Iceberg Tables in Hue and Cloudera Data Visualization

```
DESCRIBE HISTORY stocks.stock_intraday_1min;
```

```
SELECT count(*), ticker
FROM stocks.stock_intraday_1min
FOR SYSTEM_VERSION AS OF <snapshotid>
GROUP BY ticker;
```

```
SELECT count(*), ticker
FROM stocks.stock_intraday_1min
GROUP BY ticker;
```