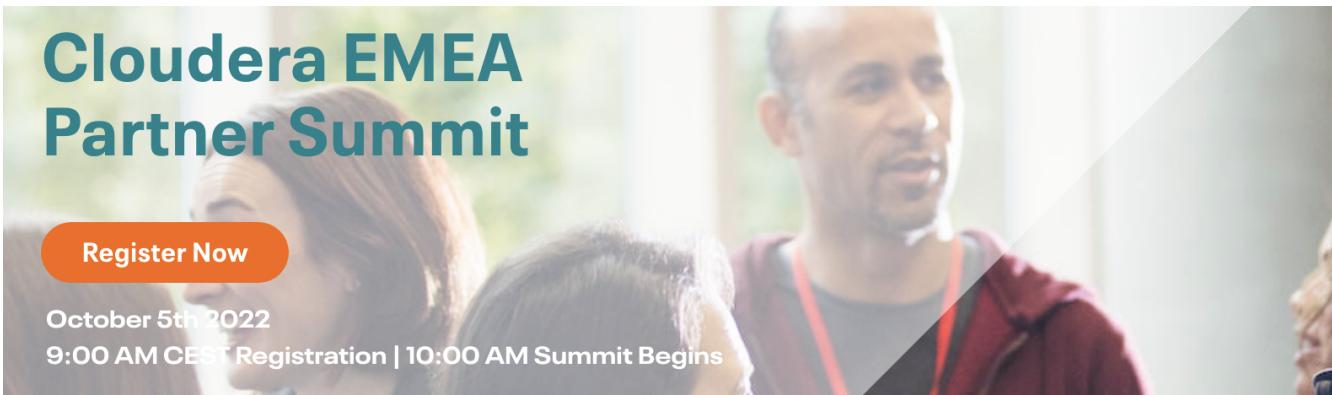


Partner Summit 2022 - Workshop

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 workshop 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,

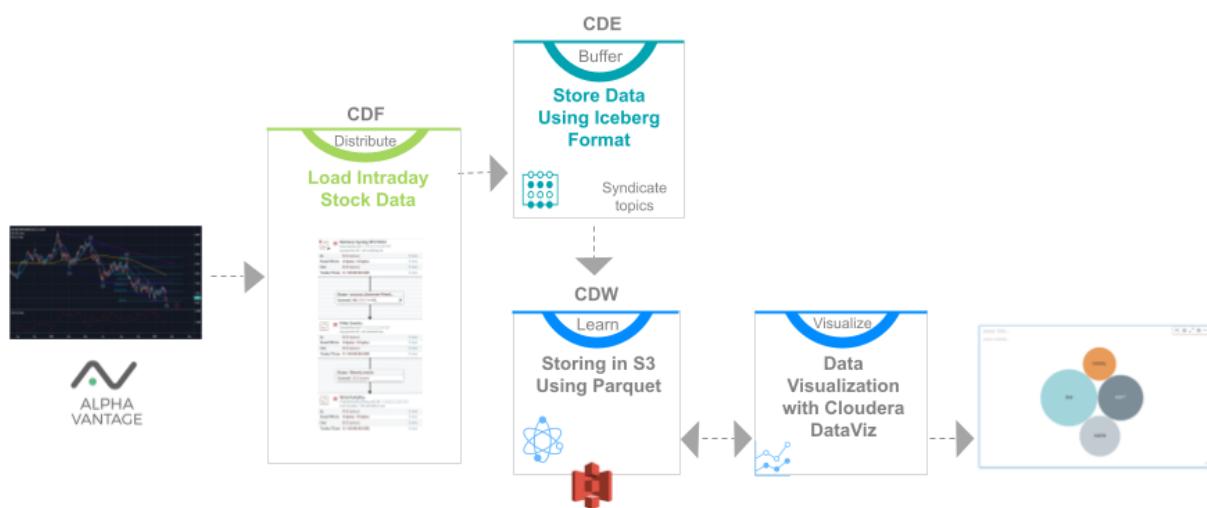
- Data ingestion and streaming—provided by ***Cloudera Data Flow (CDF)*** and ****Cloudera Data Engineering (CDE)***.
- Global data access, data processing and persistence—provided by ***Cloudera Data Hub (CDH)***.
- Data visualization with ***CDP Data Visualization***.

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.

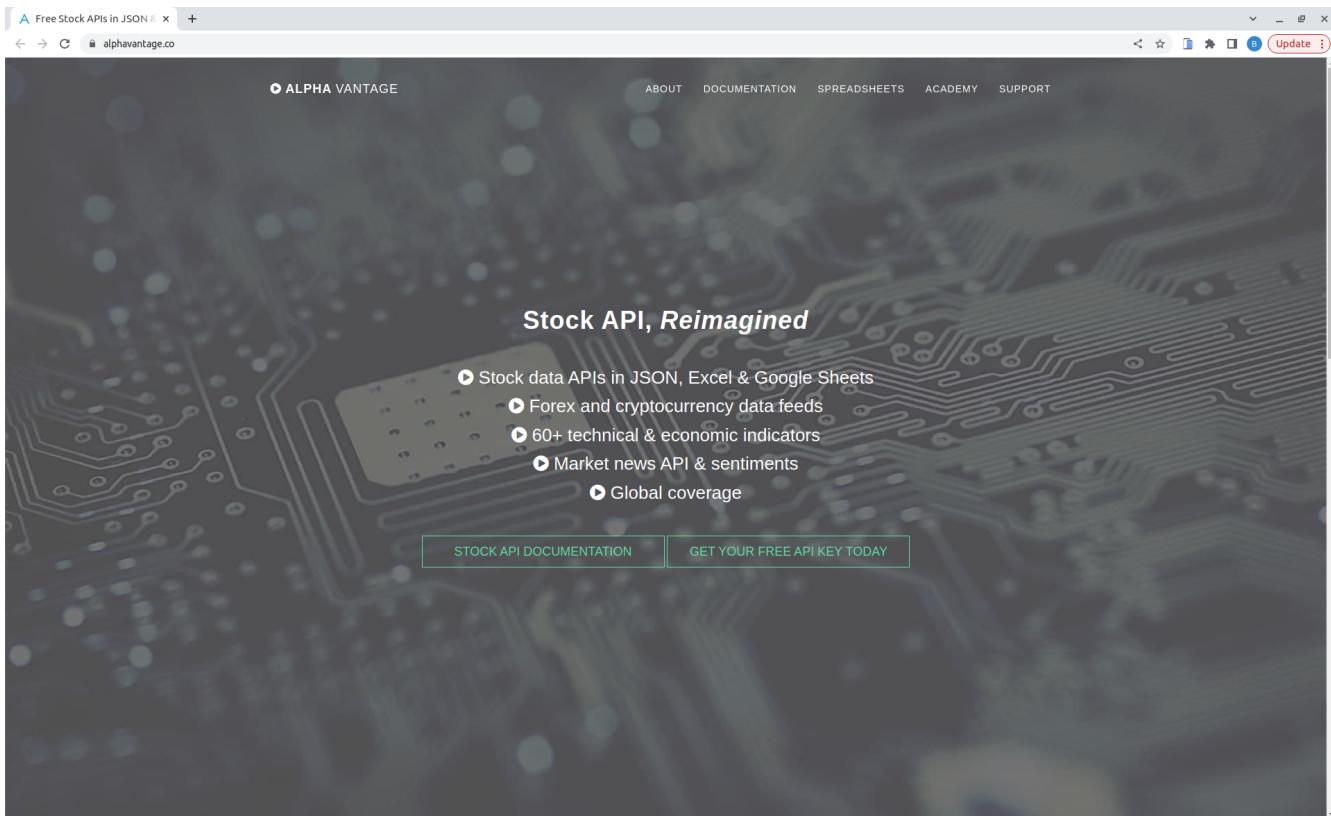
CDP Data Visualization enables data engineers, business analysts, and data scientists to quickly and easily explore data, collaborate, and share insights across the data lifecycle—from data ingest to data insights and beyond. Delivered natively as part of Cloudera Data Platform (CDP), Data Visualization 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



The screenshot shows a support page titled 'Claim your Free API Key'. On the left, there's a sidebar with links for 'Alpha Vantage Support', 'Claim your API key', and 'Support'. The main content area starts with a heading 'Claim your Free API Key' with a small gear icon. Below it is a paragraph of text: '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).'. There are several input fields: a dropdown menu set to 'Student', a text input for 'Organization (e.g. company, university, etc.)' containing 'Cloudera', and an email input field containing 'guedes.bruno@gmail.com'. Below these is a reCAPTCHA checkbox labeled 'I'm not a robot'. A large green button at the bottom right says 'GET FREE API KEY'.

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

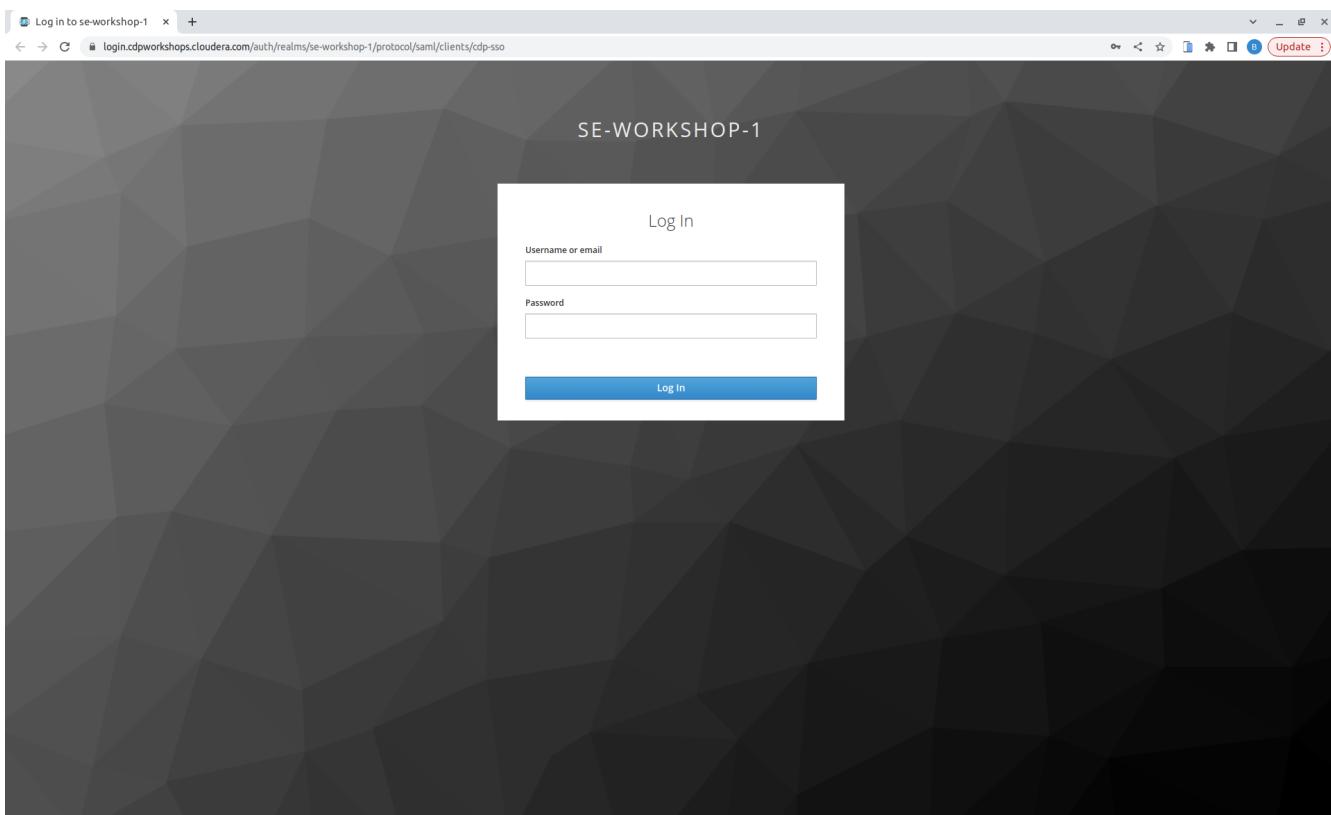
Frequently Asked Questions

I have got my API key. What's next?

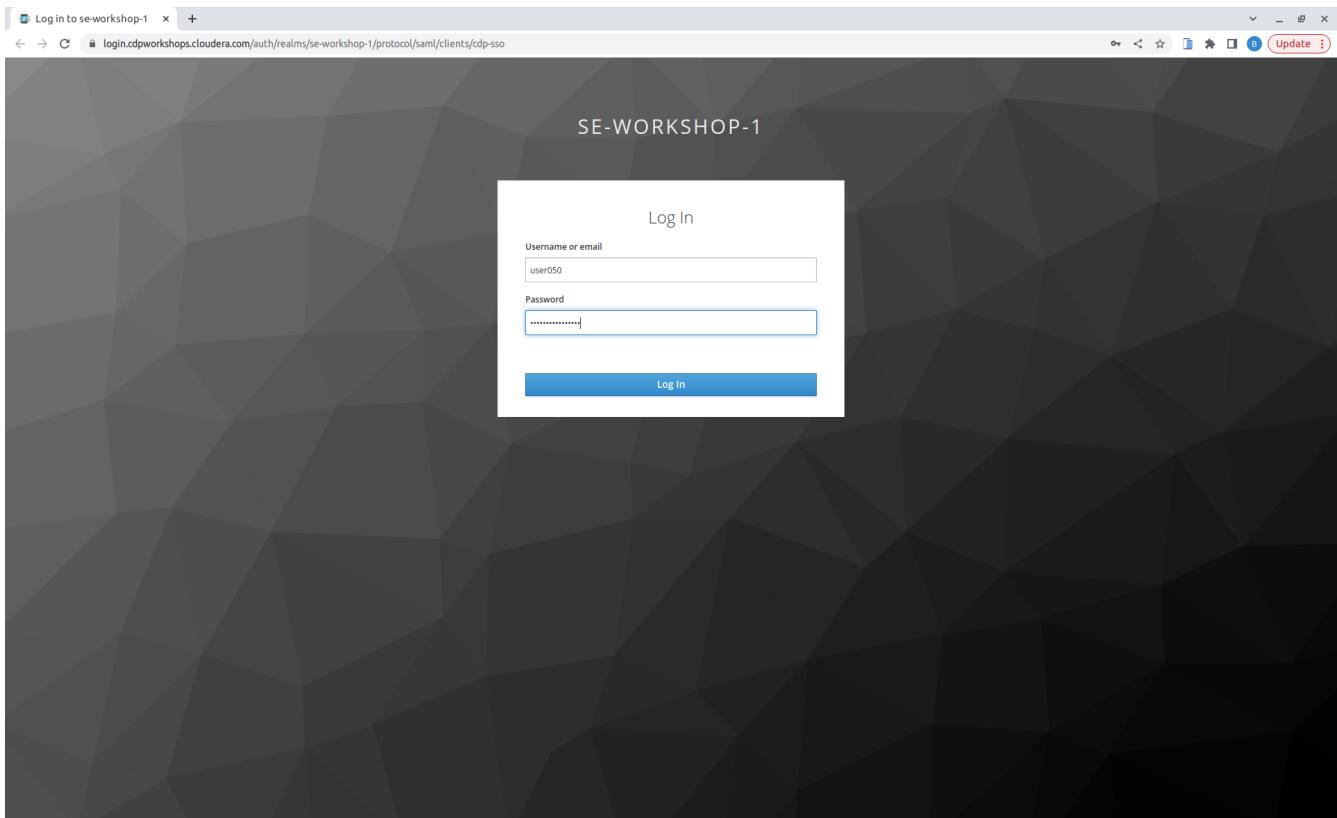
The screenshot shows a web browser window with two tabs open: "Free Stock APIs in JSON" and "Customer Support | Alpha Vantage". The main content is a form titled "Alpha Vantage Support" for claiming an API key. It asks for user information: "Which of the following best describes you?" (Investor), "Organization (e.g. company, university, etc.)", "Email:", and a reCAPTCHA field. A green "GET FREE API KEY" button is at the bottom. Below the form, a message says "Welcome to Alpha Vantage! Your API key is: RYPPVI4QIBDO8FFU. Please record this API key at a safe place for future data access." On the left sidebar, there are links for "Claim your API key" and "Support".

Step 2: Access CDP Public Cloud Portal

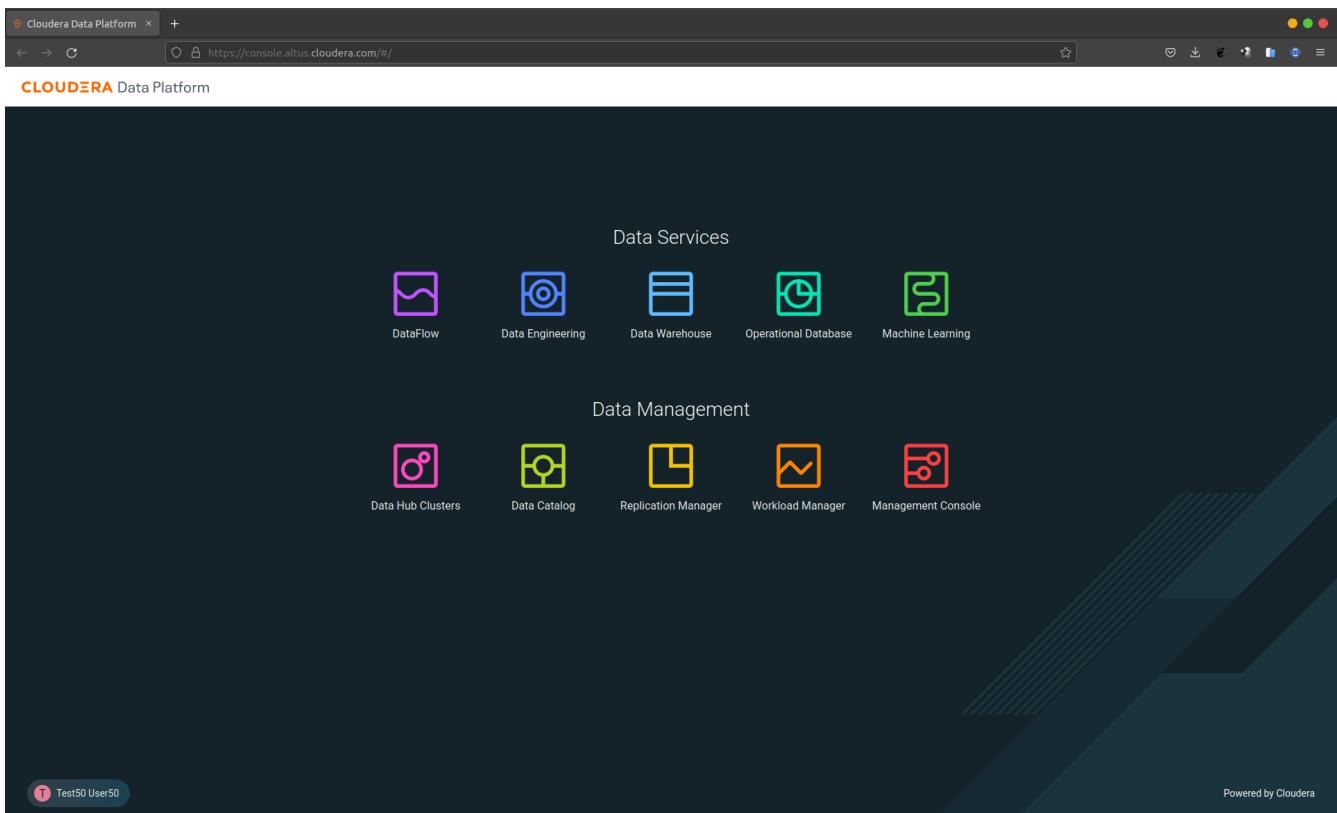
Please use the login url [Workshop login](#)



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

Cloud Management C x +

console.us-west-1.cdp.cloudera.com/cloud/environments/list

CLOUDERA Management Console

Environments / List

1 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

Help Test50 User50 2.62.0-b150

Cloud Management C x +

Cloud DataFlow

https://console.us-west-1.cdp.cloudera.com/cloud/environments/list

CLOUDERA Management Console

Environments / List

1 Environments

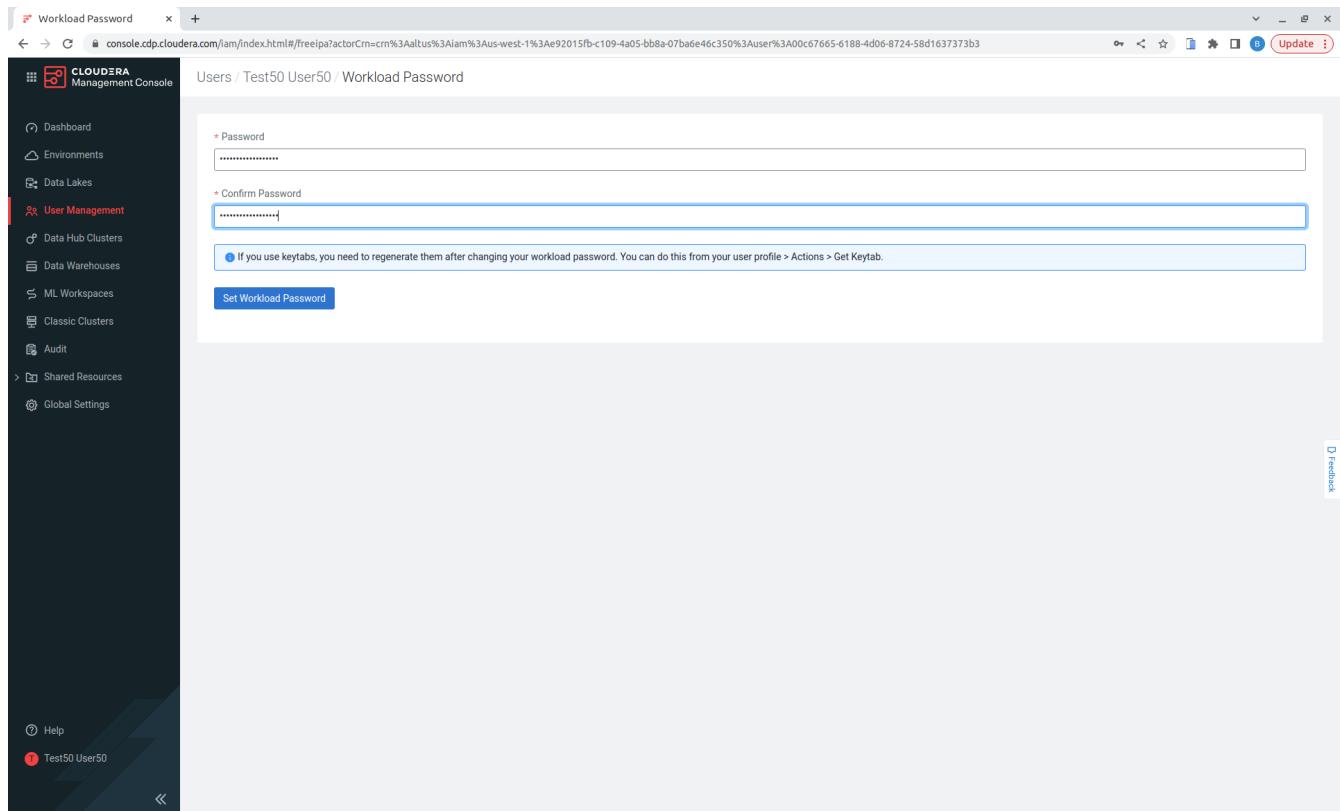
Search Start Environment Stop Environment Delete Create Data Hub Register Environment

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

Test50 User50 user050@localhost.com

Profile Log Out



The screenshot shows the user profile page for 'Test50 User50' in the Cloudera Management Console. The left sidebar is identical to the previous screenshot. The main content area displays various user details: Name (Test50 User50), Email (user050@localhost.com), Workload User Name (user050), CRN (cm:altus:iam:us-west-1:e92015fb-c109-4a05-bb8a-07ba6e46c350...), Tenant ID (e92015fb-c109-4a05-bb8a-07ba6e46c350), Identity Provider (se-workshop-1-keycloak-idp), Last Interactive Login (10/02/2022 11:43 PM CEST), and Profile Management (View profile). Below this, there is a section for 'Workload Password' with a link to 'Set Workload Password' (Workload password is currently set). At the bottom, there are tabs for 'Access Keys', 'Roles', 'Resources', 'Groups', and 'SSH Keys'. The 'Access Keys' tab is selected, showing a message 'No access keys found.' and a blue 'Generate Access Key' button. The top right corner of the browser window has an 'Actions' dropdown menu.

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

The screenshot shows the Cloudera Data Platform dashboard. At the top, there's a header bar with the title "Cloudera Data Platform" and a URL "console.cdp.cloudera.com/#/". Below the header is a dark-themed dashboard area. In the center, there are two main sections: "Data Services" and "Data Management".

Data Services:

- DataFlow (Icon: purple square with a wavy line)
- Data Engineering (Icon: blue square with a circular arrow)
- Data Warehouse (Icon: white square with a blue grid)
- Operational Database (Icon: green square with a blue circle)
- Machine Learning (Icon: orange square with a blue "S")

Data Management:

- Data Hub Clusters (Icon: blue square with a circular arrow)
- Data Catalog (Icon: blue square with a magnifying glass)
- Replication Manager (Icon: yellow square with a blue L-shape)
- Workload Manager (Icon: orange square with a blue line graph)
- Management Console (Icon: blue square with a circular arrow)

On the right side of the dashboard, there's a vertical sidebar with the text "D. Freitas" and a "Logout" button. At the bottom left, there's a user profile for "Bruno Guedes" and a link to "https://console.us-west-1.cdp.cloudera.com/dfx/". On the bottom right, it says "Powered by Cloudera".

Create a new CDF Catalog

The screenshot shows the Cloudera DataFlow interface. The left sidebar has navigation links: Dashboard, Catalog (which is selected), ReadyFlow Gallery, Functions, Environments, Help, and a Test50 User50 entry.

The main content area is titled "Flow Catalog" and contains a table of flow definitions:

Name	Type	Versions	Last Updated	Action
Gab - Stocks Ingestion	Custom Flow Definition	1	2 days ago	>
Hello World	ReadyFlow	1	24 days ago	>
S3 to S3 Avro	ReadyFlow	2	a year ago	>
Saved S3 to S3 Avro	Custom Flow Definition	1	4 months ago	>
user50-stock-data	Custom Flow Definition	1	a day ago	>
Workshop - S3 to S3 Avro	Custom Flow Definition	1	4 months ago	>

At the bottom of the interface, there are buttons for "Items per page" (set to 10), "1 = 6 of 6", and navigation arrows (<<, <, >, >>).

The screenshot shows the Cloudera DataFlow interface with the 'Catalog' tab selected. A modal dialog titled 'Import Flow Definition' is open. In the 'Flow Name' field, 'user50-stock-data' is entered. The 'Flow Description' field contains the placeholder 'Add description'. Under 'NiFi Flow Configuration', a file named 'Stocks_Intraday_Alpha_Template.json' is selected. In the 'Version Comments' field, 'Initial Version' is typed. At the bottom right of the dialog are 'Cancel' and 'Import' buttons. The background shows a list of existing flows: 'Gab - Stocks Ingestion', 'Hello World', 'S3 to S3 Avro', 'Saved S3 to S3 Avro', 'user01-stock-data-cdf', and 'Workshop - S3 to S3 Avro'. To the right of the dialog is a table of flow versions with columns 'Versions', 'Last Updated', and a 'More' column.

Stocks_Intraday_Alpha_Template.json

The screenshot shows the Cloudera DataFlow interface with the 'Catalog' tab selected. A search bar at the top contains 'user50'. The search results table has columns 'Name', 'Type', 'Versions', and 'Last Updated'. One result is listed: 'user50-stock-data' (Custom Flow Definition, 1 version, last updated 27 seconds ago). The background shows the same list of flows as the previous screenshot.

Deploy DataFlow

Cloudera DataFlow

console.us-west-1.cdp.cloudera.com/dfx/ui/#/flows/detailsflows/045520a-alce-45d4-957c-e607ec8344be?searchTerm=user50

CLOUDERA DataFlow

Flow Catalog

user50

Name ↑

user50-stock-data

»

REFRESHED: 6 seconds ago

Actions

user50-stock-data

Updated a minute ago by Test50 User50

FLOW DESCRIPTION
No description specified

CRN #
crm: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 #
crm:cdp:df:us-west-1:e92015fb-c109-4a05-bb8a-07ba6e46c350:flow:user50-stock-data/v.1

Help

Test50 User50

2.2.0-h2-b1

This screenshot shows the Cloudera DataFlow interface. On the left, there's a sidebar with links like Dashboard, Catalog, ReadyFlow Gallery, Functions, Environments, Help, and a user profile for 'Test50 User50'. The main area is titled 'Flow Catalog' and has a search bar with 'user50'. A list of flows is shown, with 'user50-stock-data' selected. To the right, detailed information about this flow is displayed, including its CRN, version history (version 1 with 0 deployments), last update (2022-10-01 19:00 CEST), and its specific CRN.

Cloudera DataFlow

dfx.8jhx1vcv.xfa-z-gdb4.cloudera.site/dfx/ui/#/flow-deployment-wizard/deployment-request/3347bcda-9c90-46b3-bd49-33f8b56aa022/flow-overview

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

Selected Flow Definition:

NAME	VERSION
user50-stock-data	1

Target Environment:

NAME
aws
se-workshop-1-env

Cancel Next →

This screenshot shows the 'New Deployment' wizard in the Cloudera DataFlow interface. It's currently on the 'Overview' step (step 1). The deployment name is set to 'user50-stock-data-cdf'. Under 'Selected Flow Definition', it shows a single flow named 'user50-stock-data' at version 1. In the 'Target Environment' section, it lists 'aws' and 'se-workshop-1-env'. At the bottom, there are 'Cancel' and 'Next →' buttons.

Cloudera DataFlow Update

se-workshop-1-env / New Deployment

NiFi Configuration

- ① Overview
- ② **NiFi Configuration**
- ③ Parameters
- ④ Sizing & Scaling
- ⑤ Key Performance Indicators
- ⑥ Review

NiFi Runtime Version

CURRENT VERSION Latest Version (1.16.0.2.3.6.1-1)

Change Version

Autostart Behavior

Automatically start flow upon successful deployment

Inbound Connections

Allow NiFi to receive data

Custom NAR Configuration

This flow deployment uses custom NARs

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 Update

se-workshop-1-env / New Deployment

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: G

CDP_User: user50 G

S3 Path: stocks G

api_alpha_key: RE0BU7VE1RAHC0CQ G

stock_list: IBM
GOOGL
AMZN
MSFT

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.xfa-z-gdb4.cloudera.site](#) / #/flow-deployment-wizard/deployment-request/a1e1df35-4ed9-4bf9-96c1-20dee33c7439/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

<input checked="" type="radio"/> Extra Small 2 vCores Per Node 4 GB Per Node	<input type="radio"/> Small 3 vCores Per Node 6 GB Per Node	<input type="radio"/> Medium 6 vCores Per Node 12 GB Per Node	<input type="radio"/> Large 12 vCores Per Node 24 GB Per Node

Number of NiFi Nodes

Auto Scaling

Min. Nodes: Max. Nodes:

Parameters

Intraday_Stocks_Parameters

- COP_PASSWORD *[Sensitive Value Provided]*
- COP_USER *bguedes*
- S3 PATH *stocks*
- APIALPHA_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.xfa-z-gdb4.cloudera.site](#) / #/flow-deployment-wizard/deployment-request/a1e1df35-4ed9-4bf9-96c1-20dee33c7439/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 ↗](#)

Parameters

Intraday_Stocks_Parameters

- COP_PASSWORD *[Sensitive Value Provided]*
- COP_USER *bguedes*
- S3 PATH *stocks*
- APIALPHA_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 Update

dfx.8jhxcv.xfaclaz.cloudera.site/dfx/ui/#/flow-deployment-wizard/deployment-request/3347bcda-9c90-46b3-bd49-33f8b56aa022/review

se-workshop-1-env / New Deployment

Overview

NiFi Configuration

Parameters

Sizing & Scaling

Key Performance Indicators

Review

Review

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

APL_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 Deploy

Cloudera DataFlow Update

console.us-west-1.cdp.cloudera.com/dfe/u/#/deployments/details/environments/a8a7915c-947b-45cf-b2a8-a5350050db4a/deployments/eefe98f8-25aa-4556-a814-bed3e5cbda3c/alerts

REFRESHED: 7 seconds ago

Dashboard

Filter By: STATUS All - 14 ENVIRONMENTS All - 3

Status	Name ↑
Deploying	user50-stock-data-cdf se-workshop-1-env

user50-stock-data-cdf
aws 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

Help Test50 User50
2.2.0-h2-b1

Screenshot of the Cloudera DataFlow Dashboard showing a single deployment named "user50-stock-data-cdf" in "Good Health". The dashboard includes filters for Status, Environments, Deployments, Processor Types, and Metrics Window, along with a table and a throughput chart.

Dashboard

Filter By: STATUS All - 14 ENVIRONMENTS All - 3 DEPLOYMENTS All - 1 PROCESSOR TYPES All - 41 METRICS WINDOW 30 Minutes

Status	Name ↑	Current Received	Current Sent	Data Throughput (Received/Sent)
Good Health	user50-stock-data-cdf se-workshop-1-env	0 B/s	0 B/s	642.00 B/s

REFRESHED: 2 seconds ago

Items per page: 10 | 1 - 1 of 1 | < < > >|

Help Bruno Guedes 2.2.0-h2-b1

View Nifi DataFlow

Screenshot of the Cloudera DataFlow Dashboard, identical to the one above, showing a single deployment named "user50-stock-data-cdf" in "Good Health". The dashboard includes filters for Status, Environments, Deployments, Processor Types, and Metrics Window, along with a table and a throughput chart.

Dashboard

Filter By: STATUS All - 14 ENVIRONMENTS All - 3 DEPLOYMENTS All - 1 PROCESSOR TYPES All - 41 METRICS WINDOW 30 Minutes

Status	Name ↑	Current Received	Current Sent	Data Throughput (Received/Sent)
Good Health	user50-stock-data-cdf se-workshop-1-env	0 B/s	0 B/s	642.00 B/s

REFRESHED: 2 seconds ago

Items per page: 10 | 1 - 1 of 1 | < < > >|

Help Bruno Guedes 2.2.0-h2-b1

Cloudera DataFlow REFRESHED: 12 seconds ago

console.us-west-1.cdp.cloudera.com/dfx/ui/#/deployments/detailsenvironments/a8a7915c-947b-45cf-b2a8-a5350050db4a/deployments/6e93be7d-fe2e-4d30-b282-c2978e1fad29/kpis

Dashboard

Filter By: STATUS All - 14 ENVIRONMENTS All - 3

Status	Name ↑
✓ Good Health	user50-stock-data se-workshop-1-env

✓ user50-stock-data aws se-workshop-1-env Manage Deployment

KPIs System Metrics Alerts

Deployment Information

FLOW DEFINITION	DEPLOYED BY
user50-stock-data V.1	Test50 User50
NODE COUNT	AUTO SCALING
1	Up to 3 nodes
CREATED ON	LAST UPDATED
2022-10-01 20:59 CEST	2022-10-02 23:27 CEST
NIFI VERSION	CRN #
1.16.0.2.3.6.1-1	crm: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 REFRESHED: 15 seconds ago

dfx.8jhx1vcv.xfa2-gdb4.cloudera.site/dfx/ui/#/deployments/manage/6e93be7d-fe2e-4d30-b282-c2978e1fad29/kpis

Dashboard / se-workshop-1-env / user50-stock-data

◀ Back to Deployment Details

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	aws se-workshop-1-env	REGION	US East (Ohio)	NIFI RUNTIME VERSION	1.16.0.2.3.6.1-1	CRN #	crm:cdp:df:us-west-1:e92015fb-c109-4a05-bb8a-07...

Deployment Settings

KPIs and Alerts Sizing and Scaling Parameters NiFi Configuration

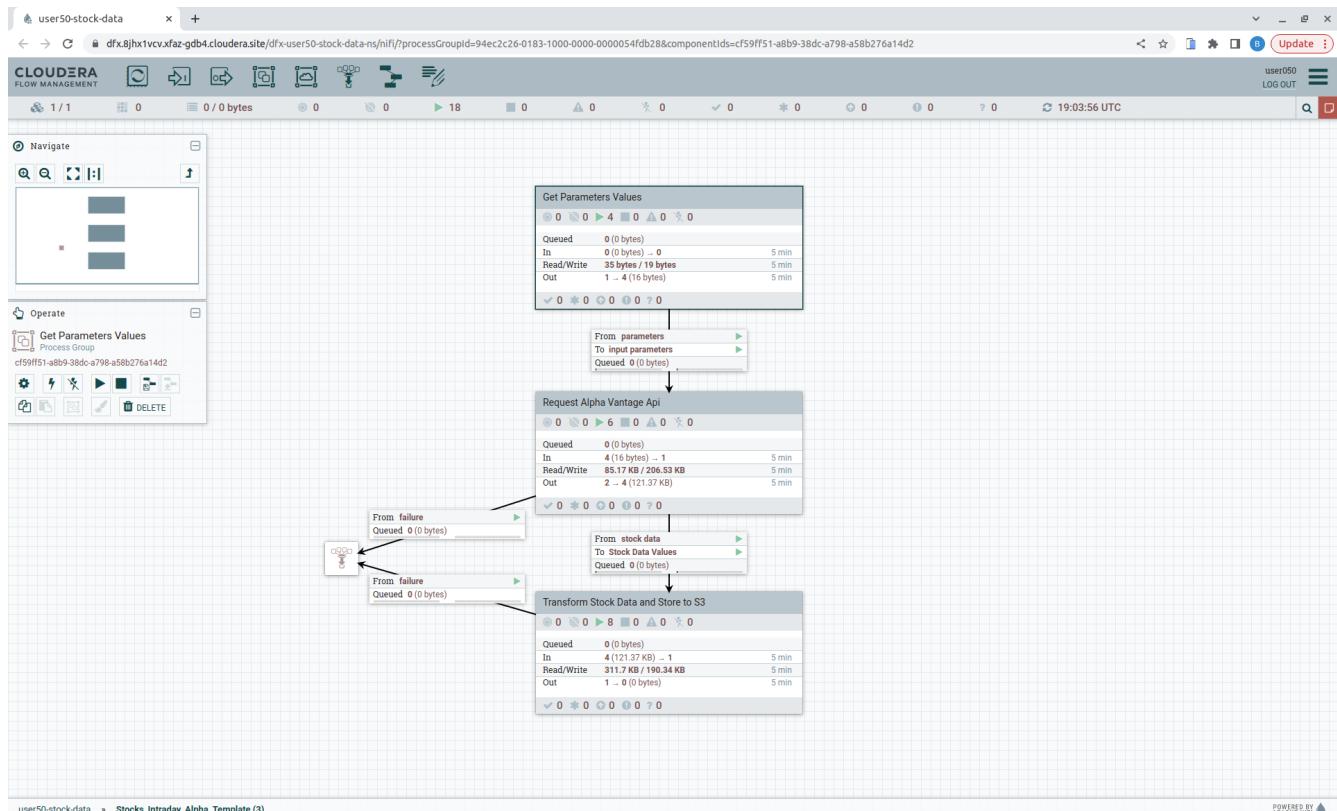
Key Performance Indicators

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;
```