LAB

02 – Cloudera Data Engineering

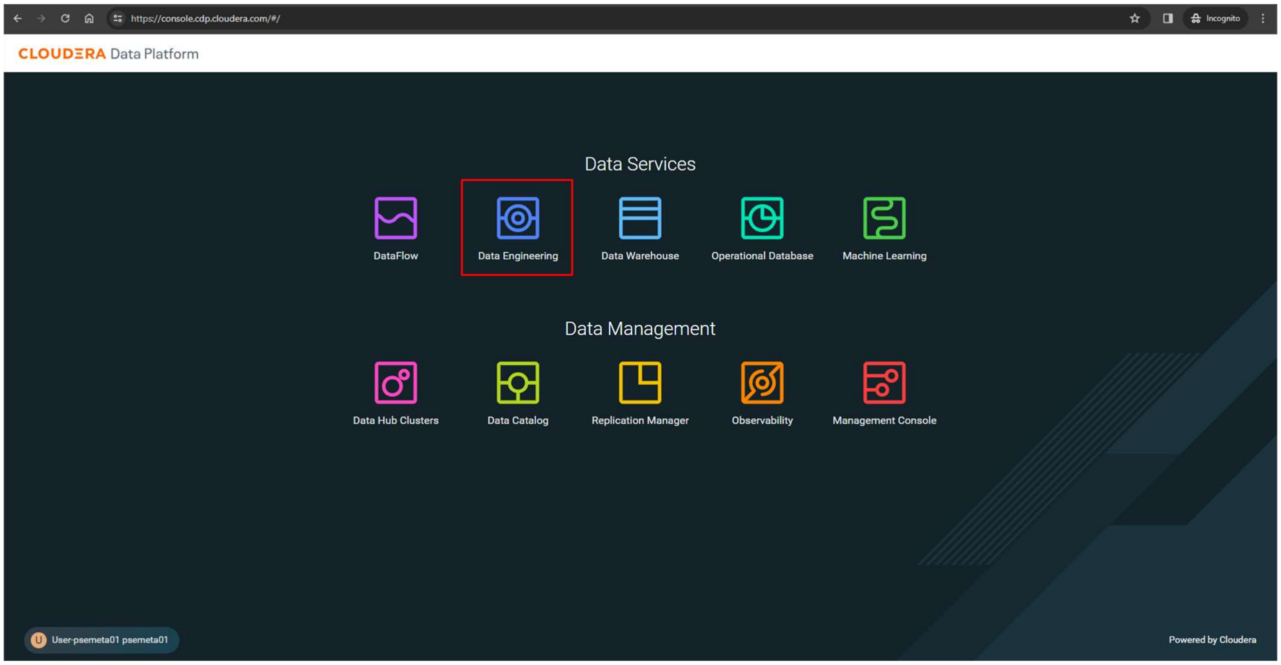
Data Lifecycle on CDP Public Cloud

Data Engineering Lab

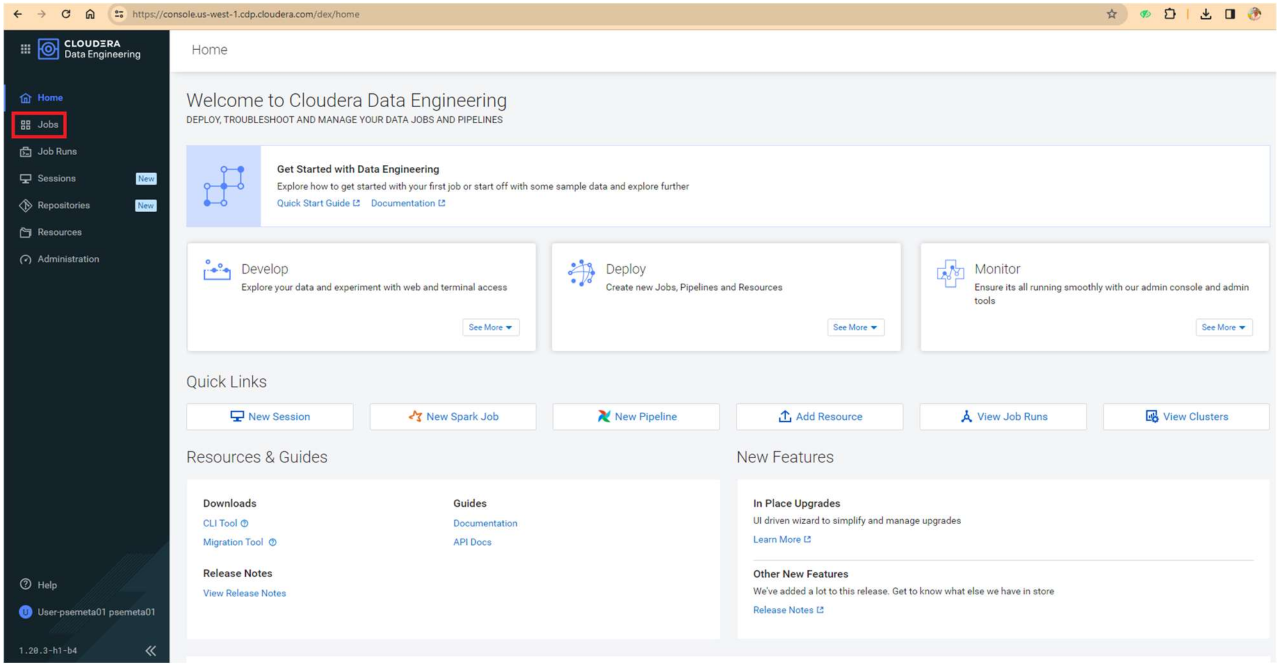
Goals:

- Run a data enrichment process
- Run a process to simulate changes to the data
- Configure the execution of a pipeline using low-code/no-code tools

1. Click on **Data Engineering** from CDP PC Home:



2. The Data Engineering Home shows all the actions that can be done, such as Jobs in Spark and pipelines in Airflow, Resources and useful information/documentation. Click on the option**Jobs** from the left menu to create a dataflow in Airflow.



Welcome to Cloudera Data Engineering

DEPLOY, TROUBLESHOOT AND MANAGE YOUR DATA JOBS AND PIPELINES

Get Started with Data Engineering

Explore how to get started with your first job or start off with some sample data and explore further

Quick Start Guide Documentation

Develop

Explore your data and experiment with web and terminal access

Deploy

Create new Jobs, Pipelines and Resources

Monitor

Ensure its all running smoothly with our admin console and admin tools

Quick Links

Resources & Guides

Downloads

CLI Tool

Migration Tool

Guides

Documentation

API Docs

Release Notes

View Release Notes

New Features

In Place Upgrades

UI driven wizard to simplify and manage upgrades

Learn More

Other New Features

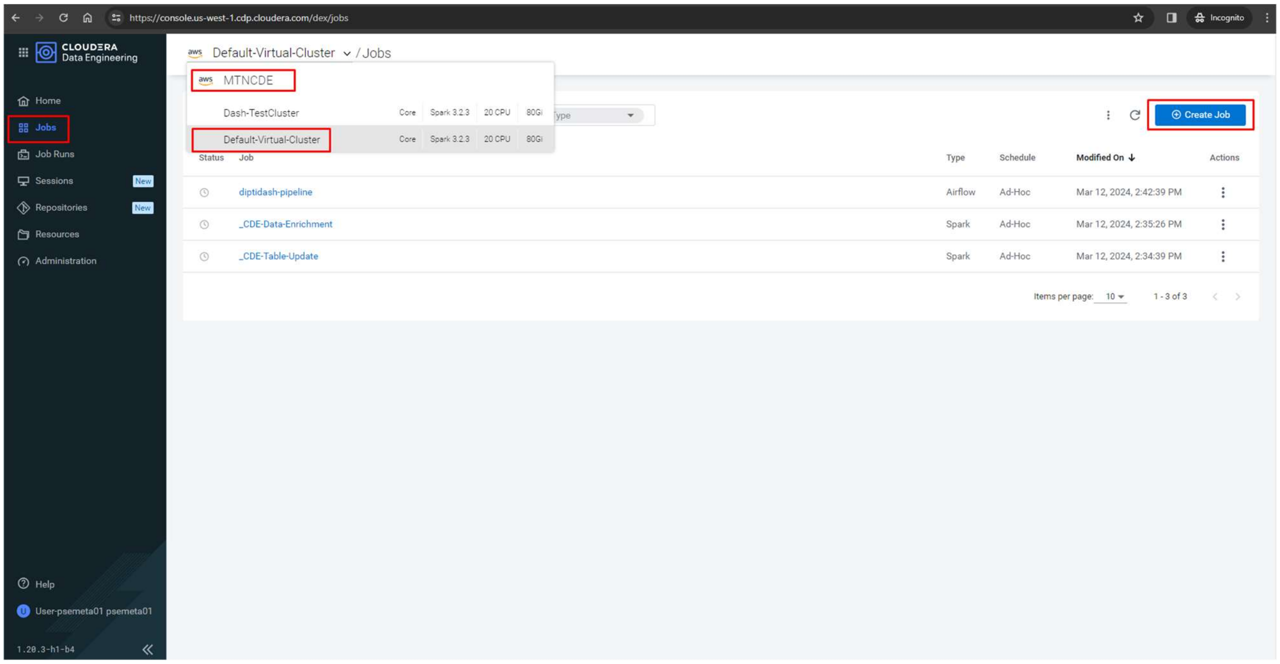
We've added a lot to this release. Get to know what else we have in store

Release Notes

3. Here the available tasks are listed. For the purposes of this workshop, two Jobs have been configured:

- **_CDE-Table-Update**, generate random changes and enrich table to visualize LakehouseTime Travel functionality.
- **_CDE-Data-Enrichment**, process in Spark (Python) to enrich the data ingested fromKafka and save to a new table.

It is time to create our Job in Airflow. Click on **Create Job.**

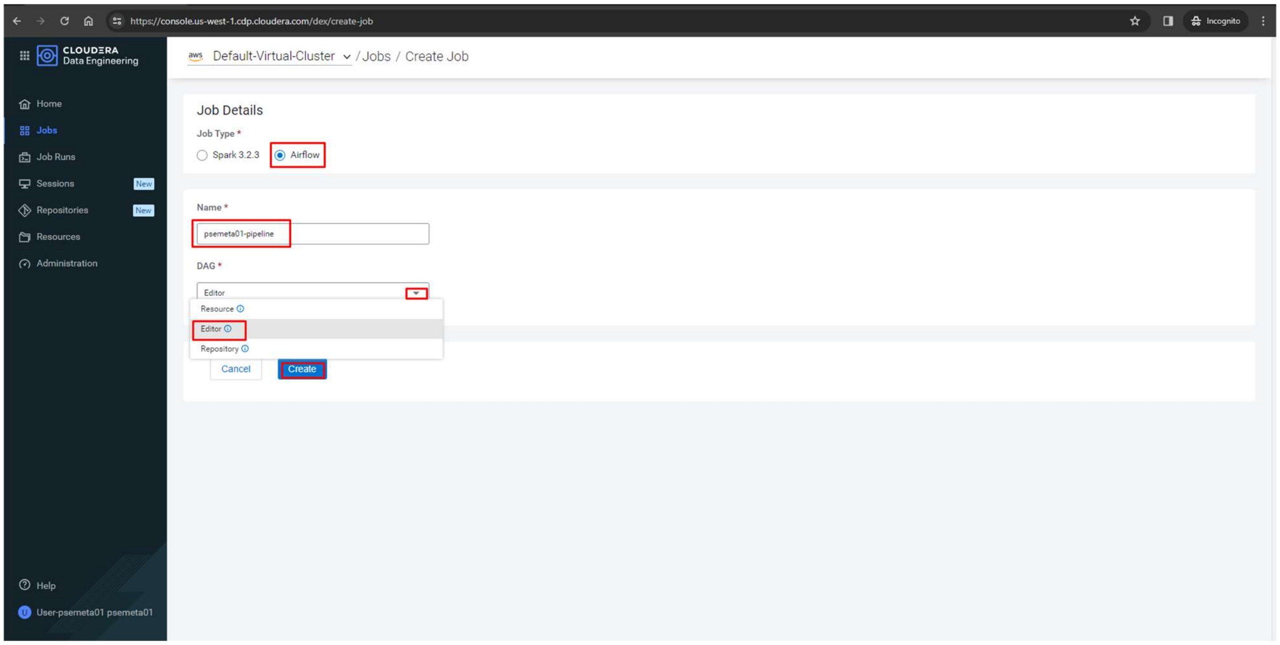


Dash-TestCluster	Core	Spark 3.2.3	20 CPU	80Gi
Default-Virtual-Cluster	Core	Spark 3.2.3	20 CPU	80Gi

Status	Job	Type	Schedule	Modified On ↓	Actions
	diptidash-pipeline	Airflow	Ad-Hoc	Mar 12, 2024, 2:42:39 PM	⋮
	_CDE-Data-Enrichment	Spark	Ad-Hoc	Mar 12, 2024, 2:35:26 PM	⋮
	_CDE-Table-Update	Spark	Ad-Hoc	Mar 12, 2024, 2:34:39 PM	⋮

4. In the Job creation form, you must enter the following information:

- Job Type: **Airflow**
- Name: Use the naming <assigned user>-pipeline. Replace <assigned user> with theuser assigned to you. For example, ***psemeta01***
- DAG***:*** **Editor*,*** to graphically configure the task. Then, click **Create**.


https://console.us-west-1.cdp.cloudera.com/dex/create-job
Incognito
CLOUDERA Data Engineering
Default-Virtual-Cluster / Jobs / Create Job
Home
Jobs
Job Runs
Sessions New
Repositories New
Resources
Administration
Help
User-psemeta01 psemeta01

Job Details

Job Type *

○ Spark 3.2.3 ● Airflow

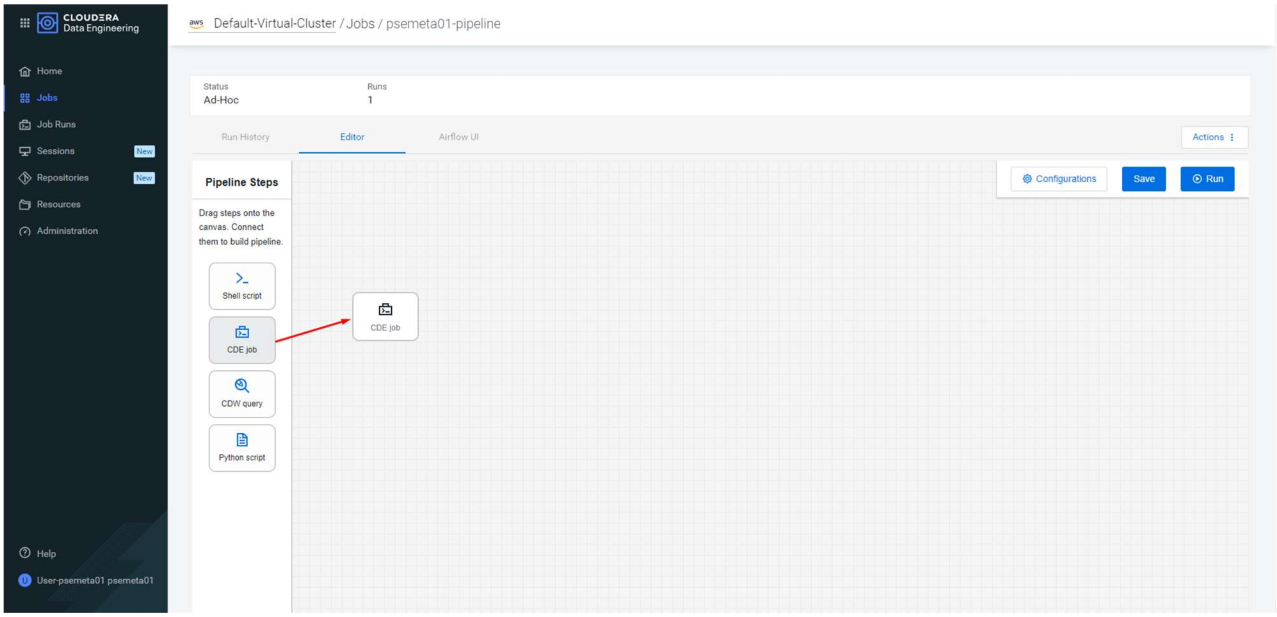
Name *

psemeta01-pipeline

DAG *

Editor

5. On the Job editing screen, select the Editor tab, and you will see the following canvas to drag the steps of the pipeline that we are going to create. In our case, we are going to create two CDE Jobs and relate them. Drage the **CDE job** into the canvas.



Status: Ad-Hoc

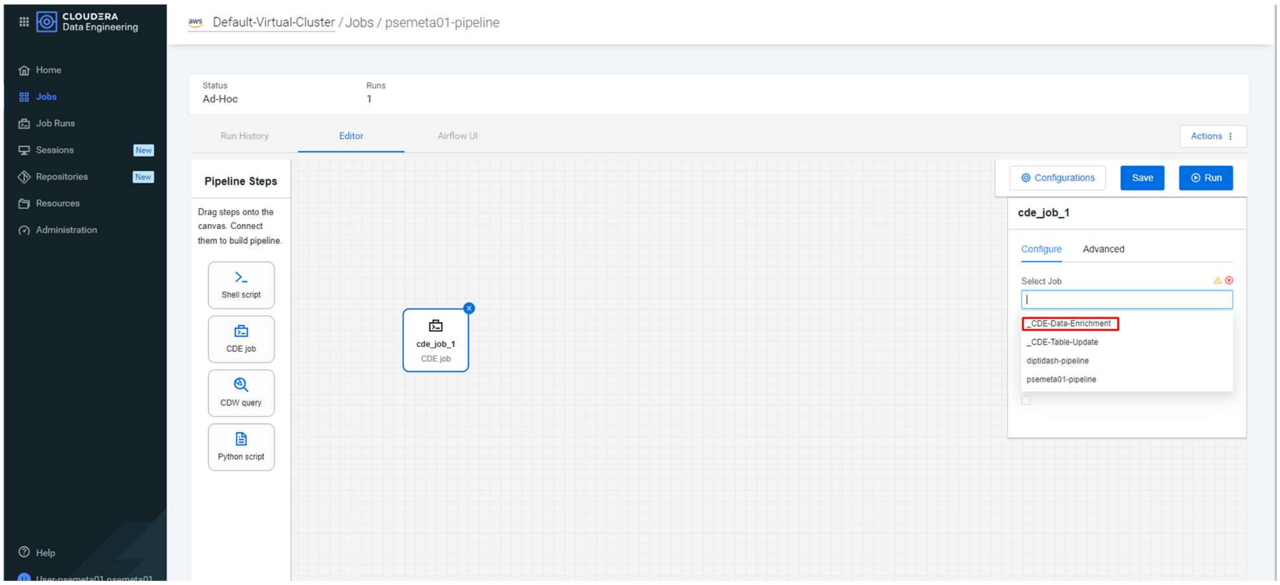
Runs: 1

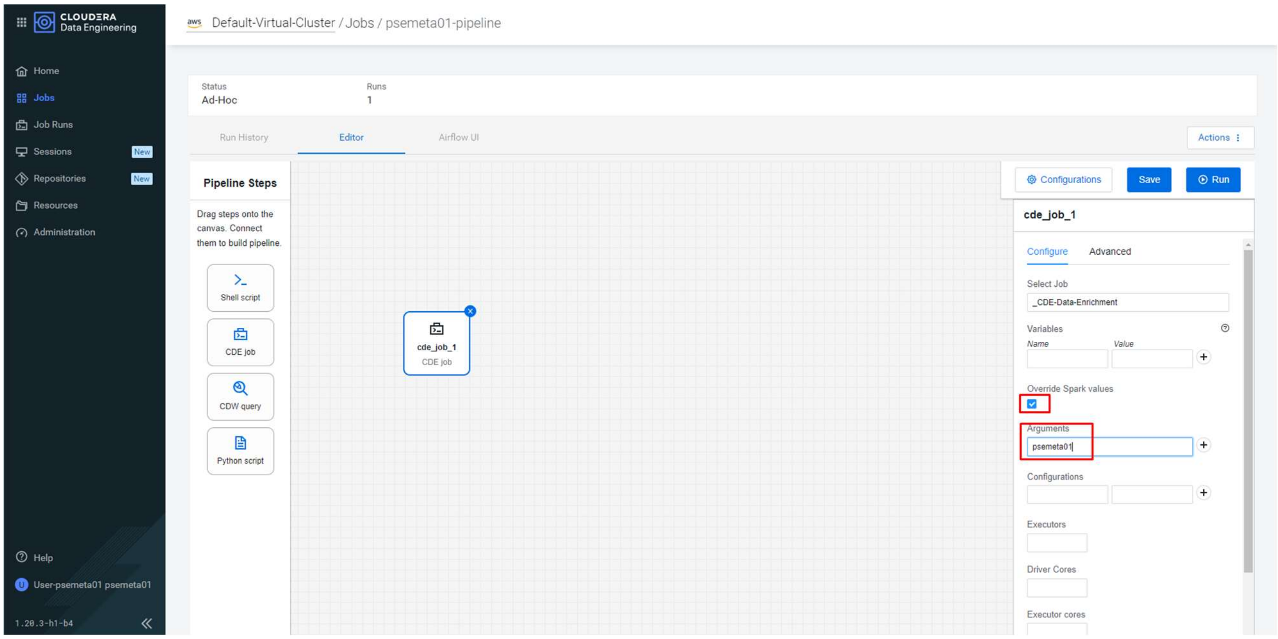
Pipeline Steps

Drag steps onto the canvas. Connect them to build pipeline.

6. Let's start with the first Job. Click on the CDE Job button and drag onto the canvas, entering the following settings:

- **Select Job**: select the Job ***_CDE-Data-Enrichment***
- Check the checkbox **Override Spark values**. Additional options will appear below.
- **Arguments**: <assigned user>. Use the username assigned to you. For example, ***psemeta01***





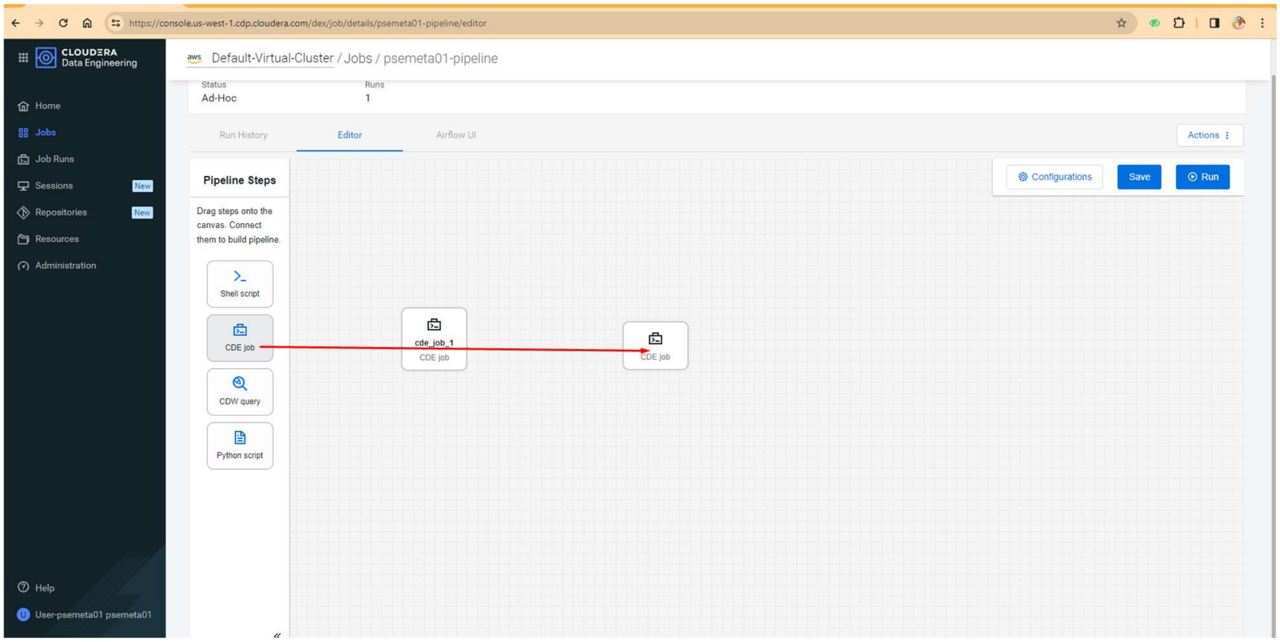
Select Job: _CDE-Data-Enrichment

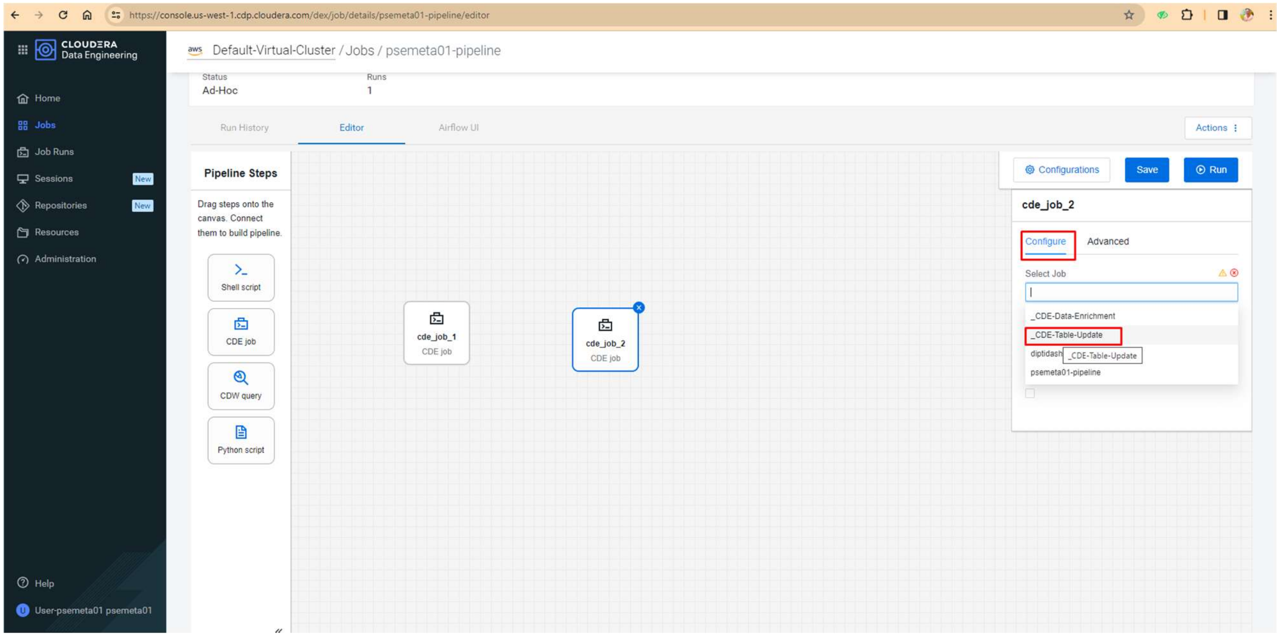
Override Spark values: ☑

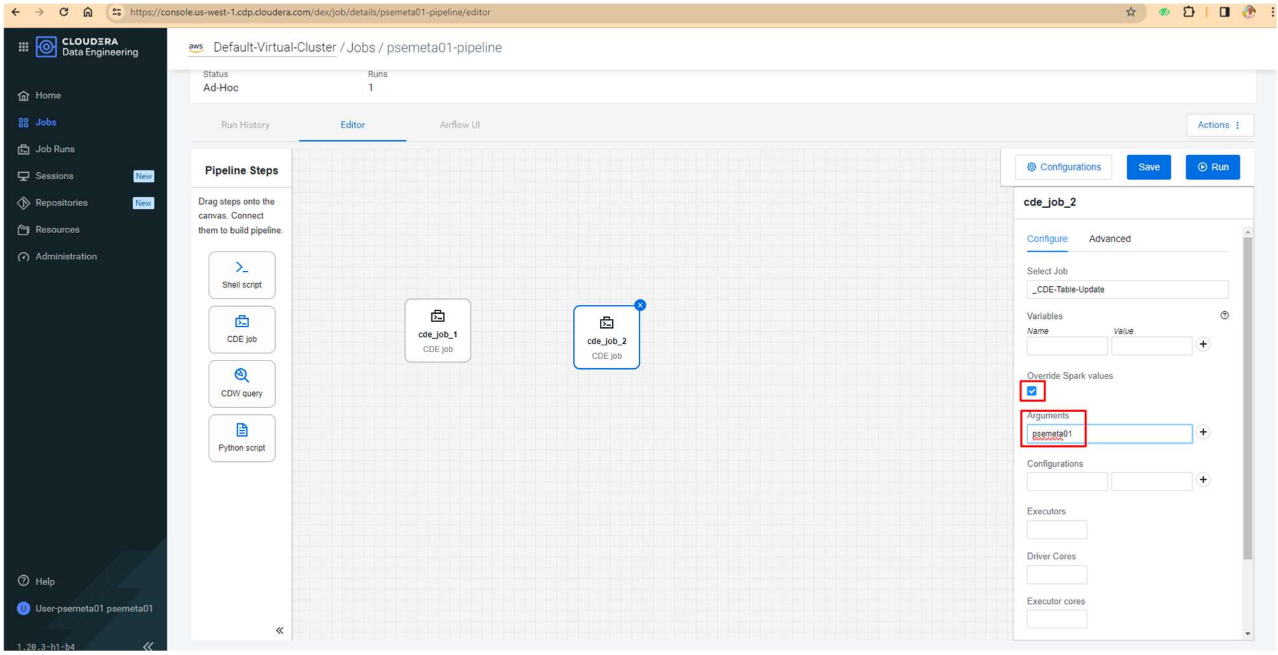
Arguments: psemeta01

7. Configure the second Job. Click on the CDE Job button and drag onto the canvas, entering the following settings:

- **Select Job**: select the Job ***_CDE-Table-Update***
- Check the checkbox **Override Spark values**. Additional options will appear below.
- **Arguments**: <assigned user>. Use the username assigned to you. For example, ***psemeta01***


https://console.us-west-1.cdp.cloudera.com/dex/job/details/psemeta01-pipeline/editor
CLOUDERA Data Engineering
Home
Jobs
Job Runs
Sessions New
Repositories New
Resources
Administration
Help
User-psemeta01 psemeta01
aws Default-Virtual-Cluster / Jobs / psemeta01-pipeline
Status Ad-Hoc
Runs 1
Run History
Editor
Airflow UI
Actions
Configurations
Save
Run
Pipeline Steps
Drag steps onto the canvas. Connect them to build pipeline.
Shell script
CDE job
CDW query
Python script
cde_job_1 CDE job
CDE job


https://console.us-west-1.cdp.cloudera.com/dex/job/details/psemeta01-pipeline/editor
CLOUDERA Data Engineering
Home
Jobs
Job Runs
Sessions New
Repositories New
Resources
Administration
Help
User-psemeta01 psemeta01
aws Default-Virtual-Cluster / Jobs / psemeta01-pipeline
Status Ad-Hoc
Runs 1
Run History
Editor
Airflow UI
Actions
Configurations
Save
Run
Pipeline Steps
Drag steps onto the canvas. Connect them to build pipeline.
Shell script
CDE job
CDW query
Python script
cde_job_1 CDE job
cde_job_2 CDE job
cde_job_2
Configure
Advanced
Select Job
_CDE-Data-Enrichment
_CDE-Table-Update
diptidash
_CDE-Table-Update
psemeta01-pipeline



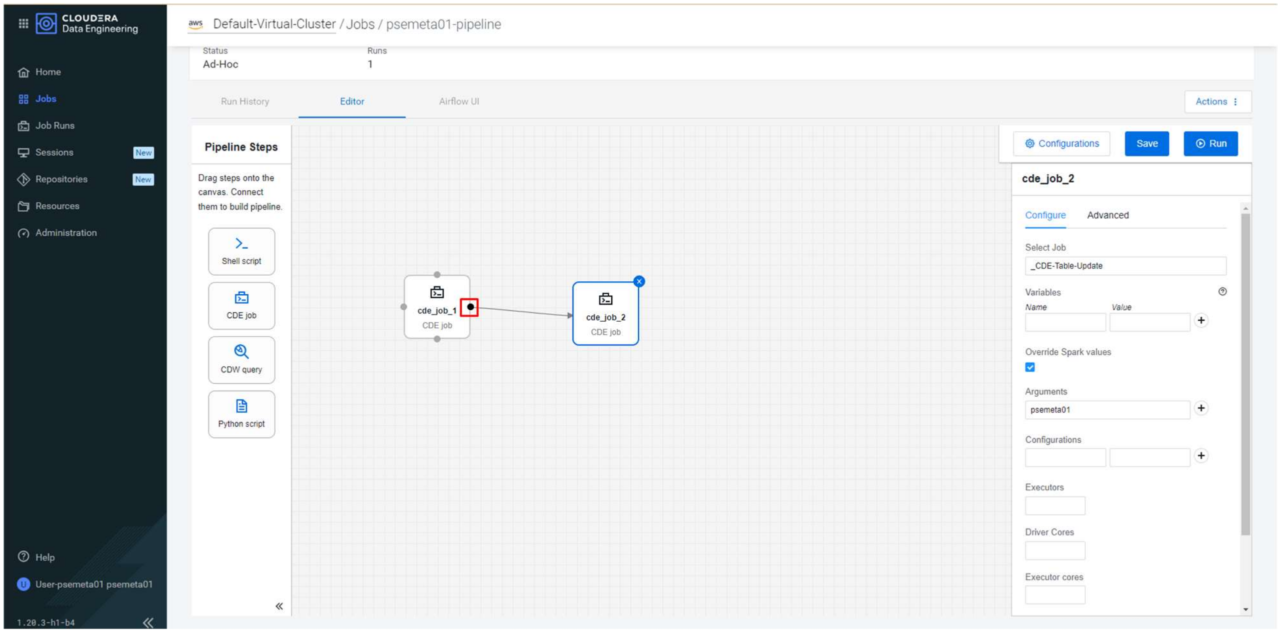
Status: Ad-Hoc
Runs: 1

Select Job: _CDE-Table-Update

Override Spark values ☑

Arguments: psemeta01

8. To set up the execution sequence, bind **cde_job_1** with **cde_job_2**. For that, click on the right connector of the job of **cde_job_1** and drag to the left connector of **cde_job_2**.



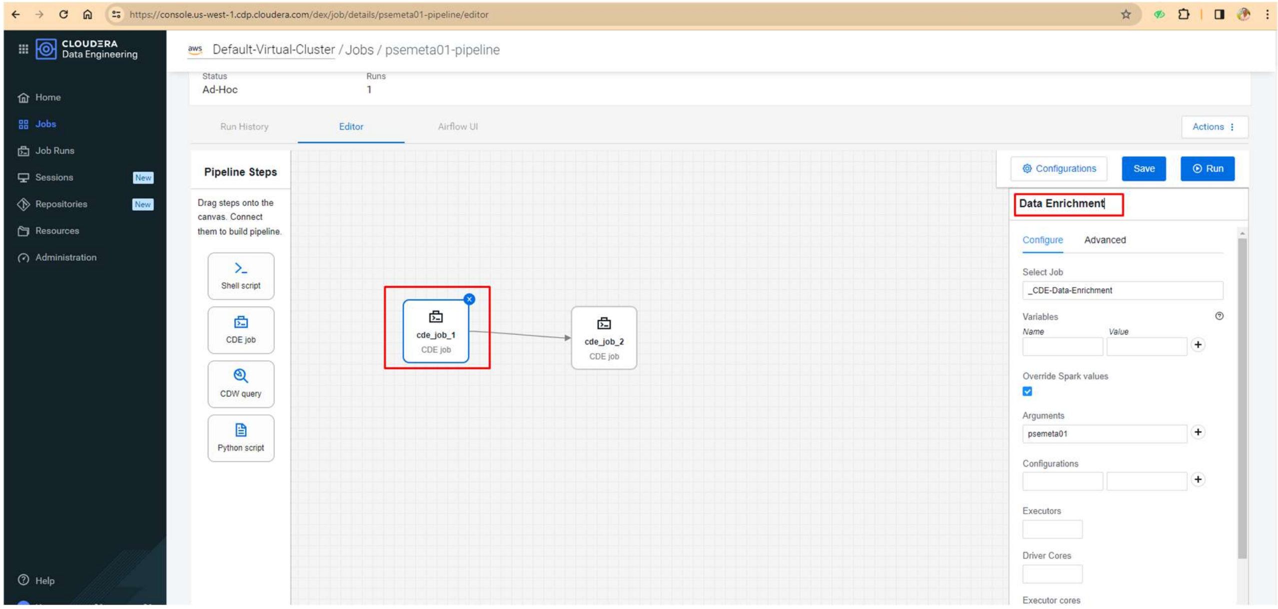
Status: Ad-Hoc
Runs: 1

Select Job: _CDE-Table-Update

Override Spark values ☑

Arguments: psemeta01

Once the Jobs are linked let’s rename the jobs. Click on **cde_job_1** and then rename it as **Data Enrichment.**



https://console.us-west-1.cdp.cloudera.com/dex/job/details/psemeta01-pipeline/editor

Default-Virtual-Cluster / Jobs / psemeta01-pipeline

Status: Ad-Hoc
Runs: 1

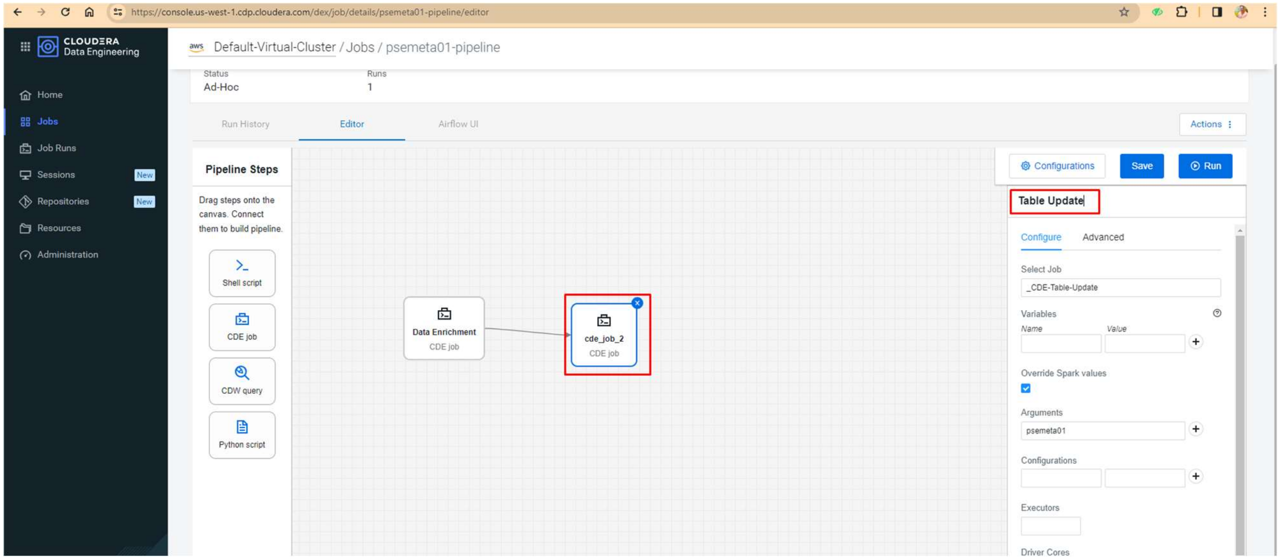
Data Enrichment

Select Job: _CDE-Data-Enrichment

Override Spark values: ☑

Arguments: psemeta01

Click on **cde_job_2** and then rename it as **Table Update.**



https://console.us-west-1.cdp.cloudera.com/dex/job/details/psemeta01-pipeline/editor

Default-Virtual-Cluster / Jobs / psemeta01-pipeline

Status: Ad-Hoc
Runs: 1

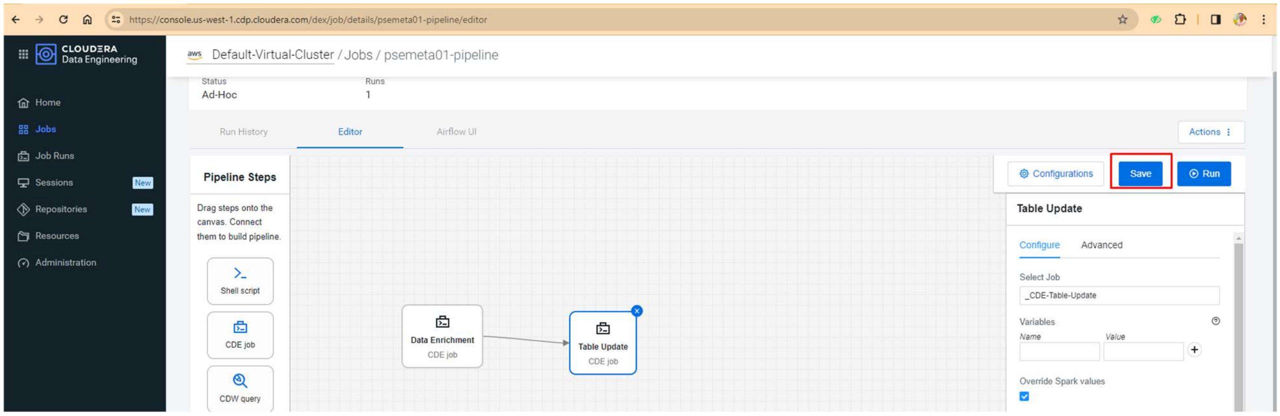
Table Update

Select Job: _CDE-Table-Update

Override Spark values: ☑

Arguments: psemeta01

9. Once the Jobs have been joined, click on **Save** to save the settings made. You should see a message indicating **Pipeline saved to job**.



https://console.us-west-1.cdp.cloudera.com/dex/job/details/psemeta01-pipeline/editor

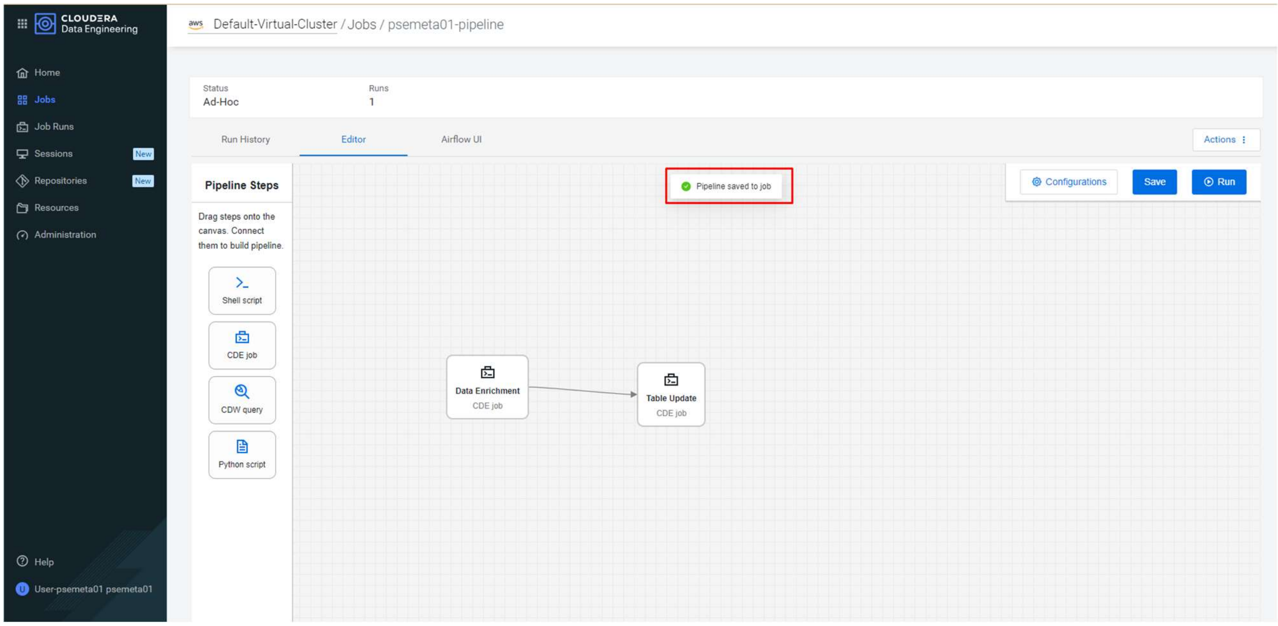
Default-Virtual-Cluster / Jobs / psemeta01-pipeline

Status: Ad-Hoc
Runs: 1

Table Update

Select Job: _CDE-Table-Update

Override Spark values ☑

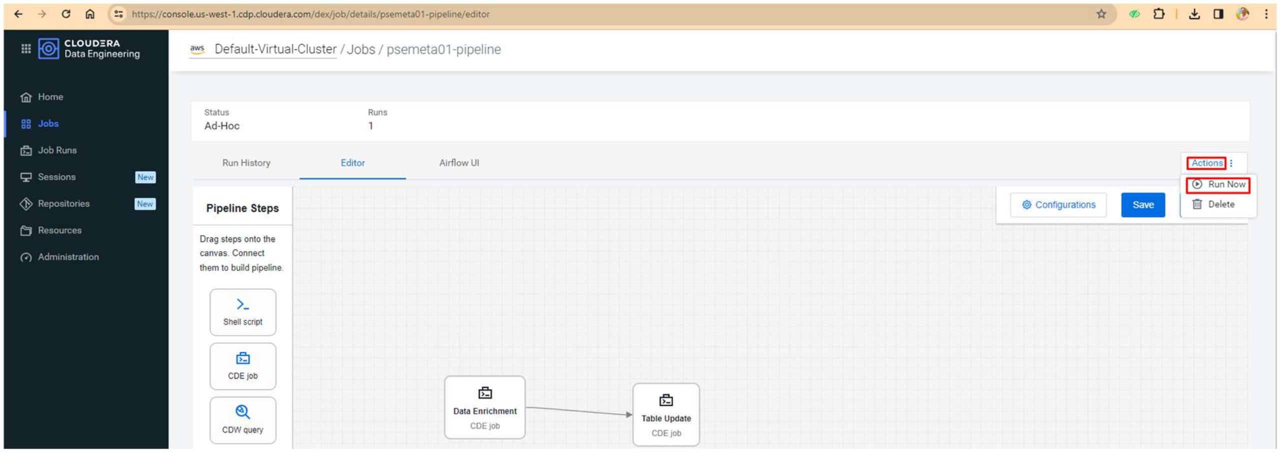


Default-Virtual-Cluster / Jobs / psemeta01-pipeline

Status: Ad-Hoc
Runs: 1

Pipeline saved to job

10. The time has come to run the pipeline. On the upper right side of the canvas, click **Actions -> Run Now**.

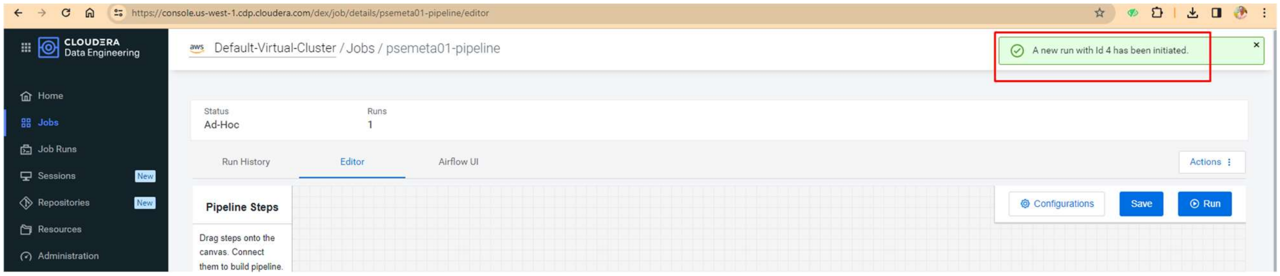


Status: Ad-Hoc

Runs: 1

https://console.us-west-1.cdp.cloudera.com/dex/job/details/psemeta01-pipeline/editor
CLOUDERA Data Engineering
Home
Jobs
Job Runs
Sessions New
Repositories New
Resources
Administration
aws Default-Virtual-Cluster / Jobs / psemeta01-pipeline
Run History
Editor
Airflow UI
Actions
Run Now
Delete
Configurations
Save
Pipeline Steps
Drag steps onto the canvas. Connect them to build pipeline.
Shell script
CDE job
CDW query
Data Enrichment
CDE job
Table Update
CDE job

11. You should see the pipeline execution screen, indicating that the execution has been initialized.



aws Default-Virtual-Cluster / Jobs / psemeta01-pipeline

A new run with Id 4 has been initiated.

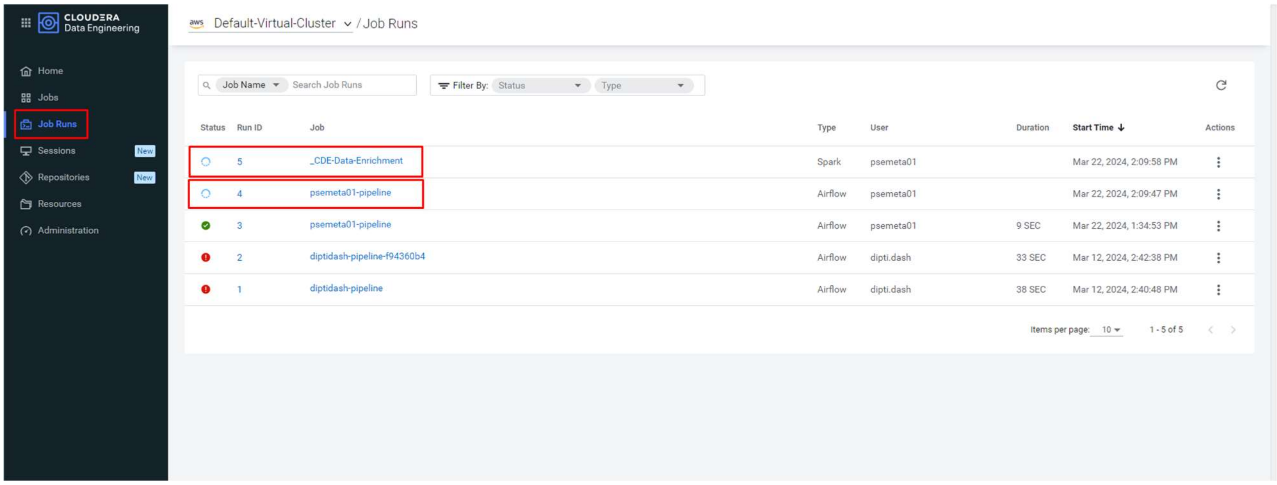
Status: Ad-Hoc

Runs: 1

Pipeline Steps

Drag steps onto the canvas. Connect them to build pipeline.

Also, on the Job Runs tab you can see the pipeline and the very first job of the pipeline getting started.

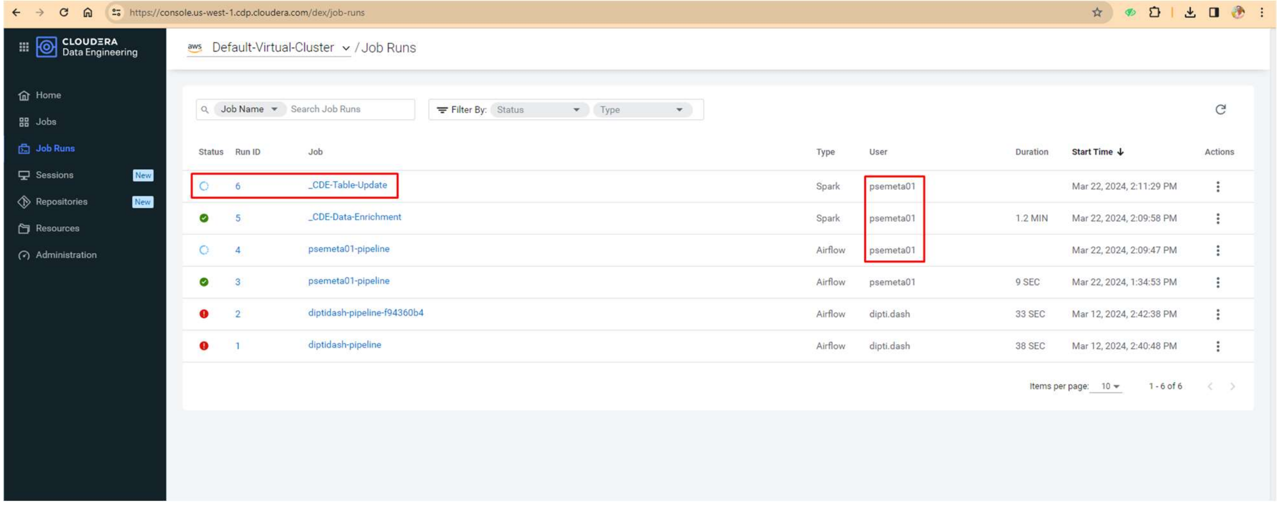


aws Default-Virtual-Cluster / Job Runs

Status	Run ID	Job	Type	User	Duration	Start Time ↓	Actions
○	5	_CDE-Data-Enrichment	Spark	psemeta01		Mar 22, 2024, 2:09:58 PM	⋮
○	4	psemeta01-pipeline	Airflow	psemeta01		Mar 22, 2024, 2:09:47 PM	⋮
✓	3	psemeta01-pipeline	Airflow	psemeta01	9 SEC	Mar 22, 2024, 1:34:53 PM	⋮
!	2	diptidash-pipeline-f94360b4	Airflow	dipti.dash	33 SEC	Mar 12, 2024, 2:42:38 PM	⋮
!	1	diptidash-pipeline	Airflow	dipti.dash	38 SEC	Mar 12, 2024, 2:40:48 PM	⋮

Items per page: 10 1 - 5 of 5

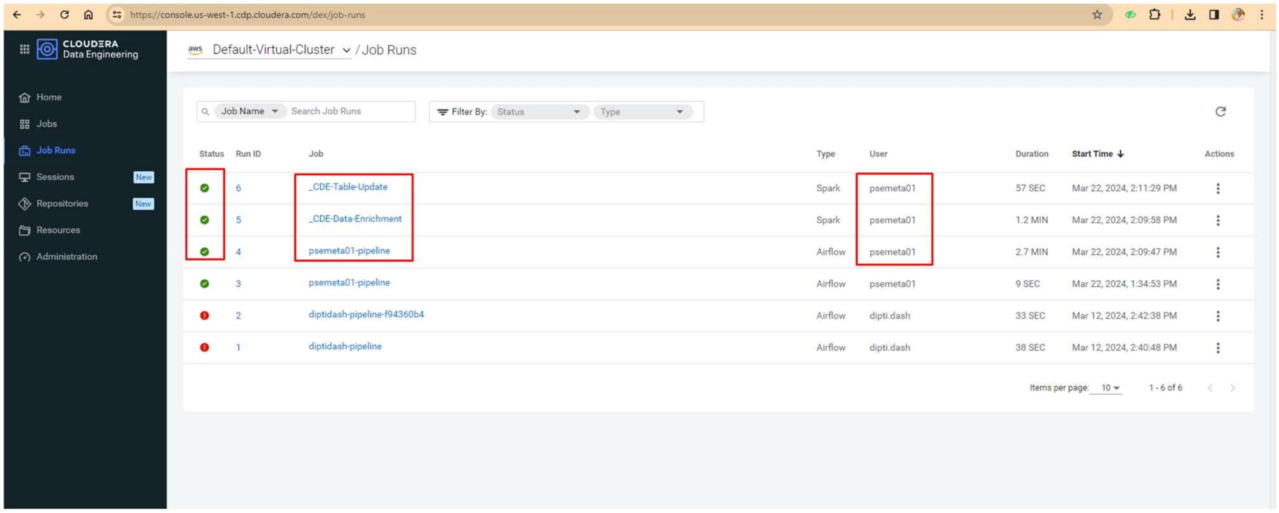
After some time the second job starts and then the pipeline and the two jobs are completed.



aws Default-Virtual-Cluster / Job Runs

Status	Run ID	Job	Type	User	Duration	Start Time ↓	Actions
○	6	_CDE-Table-Update	Spark	psemeta01		Mar 22, 2024, 2:11:29 PM	⋮
✓	5	_CDE-Data-Enrichment	Spark	psemeta01	1.2 MIN	Mar 22, 2024, 2:09:58 PM	⋮
○	4	psemeta01-pipeline	Airflow	psemeta01		Mar 22, 2024, 2:09:47 PM	⋮
✓	3	psemeta01-pipeline	Airflow	psemeta01	9 SEC	Mar 22, 2024, 1:34:53 PM	⋮
!	2	diptidash-pipeline-f94360b4	Airflow	dipti.dash	33 SEC	Mar 12, 2024, 2:42:38 PM	⋮
!	1	diptidash-pipeline	Airflow	dipti.dash	38 SEC	Mar 12, 2024, 2:40:48 PM	⋮

Items per page: 10 1 - 6 of 6

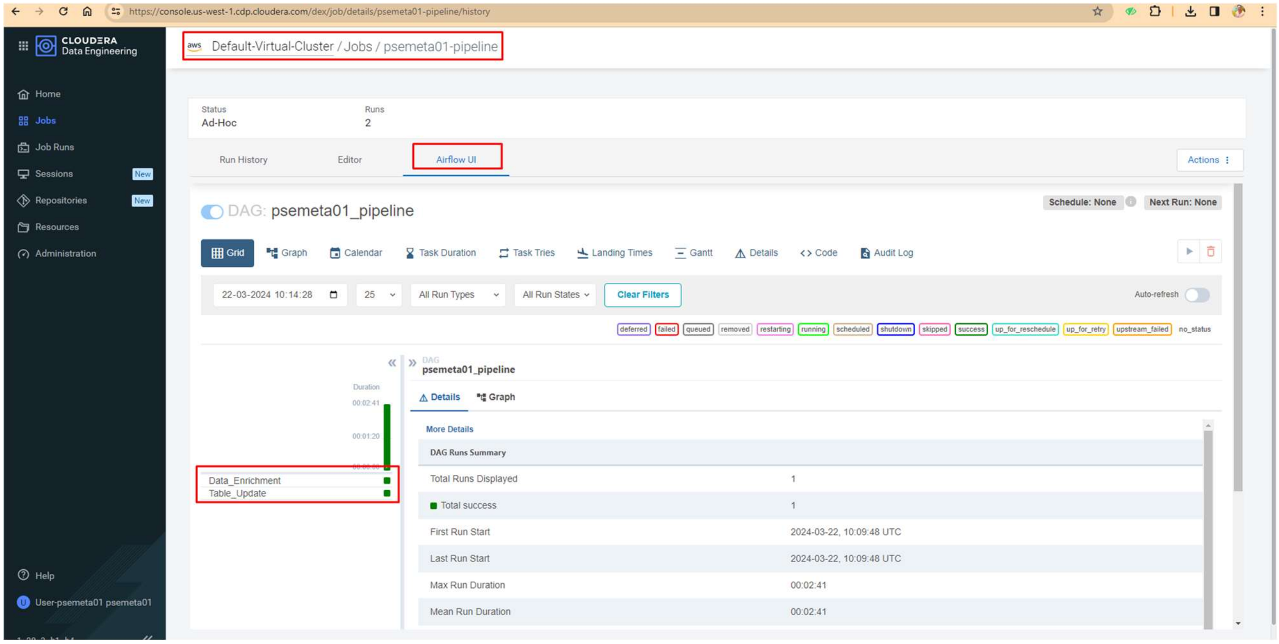


https://console.us-west-1.cdp.cloudera.com/dex/job-runs
CLOUDERA Data Engineering
Home
Jobs
Job Runs
Sessions New
Repositories New
Resources
Administration
Default-Virtual-Cluster / Job Runs
Job Name
Search Job Runs
Filter By: Status
Type

Status	Run ID	Job	Type	User	Duration	Start Time	Actions
✓	6	_CDE-Table-Update	Spark	psemeta01	57 SEC	Mar 22, 2024, 2:11:29 PM	⋮
✓	5	_CDE-Data-Enrichment	Spark	psemeta01	1.2 MIN	Mar 22, 2024, 2:09:58 PM	⋮
✓	4	psemeta01-pipeline	Airflow	psemeta01	2.7 MIN	Mar 22, 2024, 2:09:47 PM	⋮
✓	3	psemeta01-pipeline	Airflow	psemeta01	9 SEC	Mar 22, 2024, 1:34:53 PM	⋮
!	2	diptidash-pipeline-f94360b4	Airflow	dipti.dash	33 SEC	Mar 12, 2024, 2:42:38 PM	⋮
!	1	diptidash-pipeline	Airflow	dipti.dash	38 SEC	Mar 12, 2024, 2:40:48 PM	⋮

Items per page: 10 1 - 6 of 6

12. Click on the Airflow UI tab to see the execution detail of each step in the pipeline. The configured Data Enrichment and Table Update jobs are listed at the bottom left. The colours indicate the status of each job. Make sure the radio button **Auto-refresh** is enabled to automatically display the status of jobs.



Status: Ad-Hoc
Runs: 2

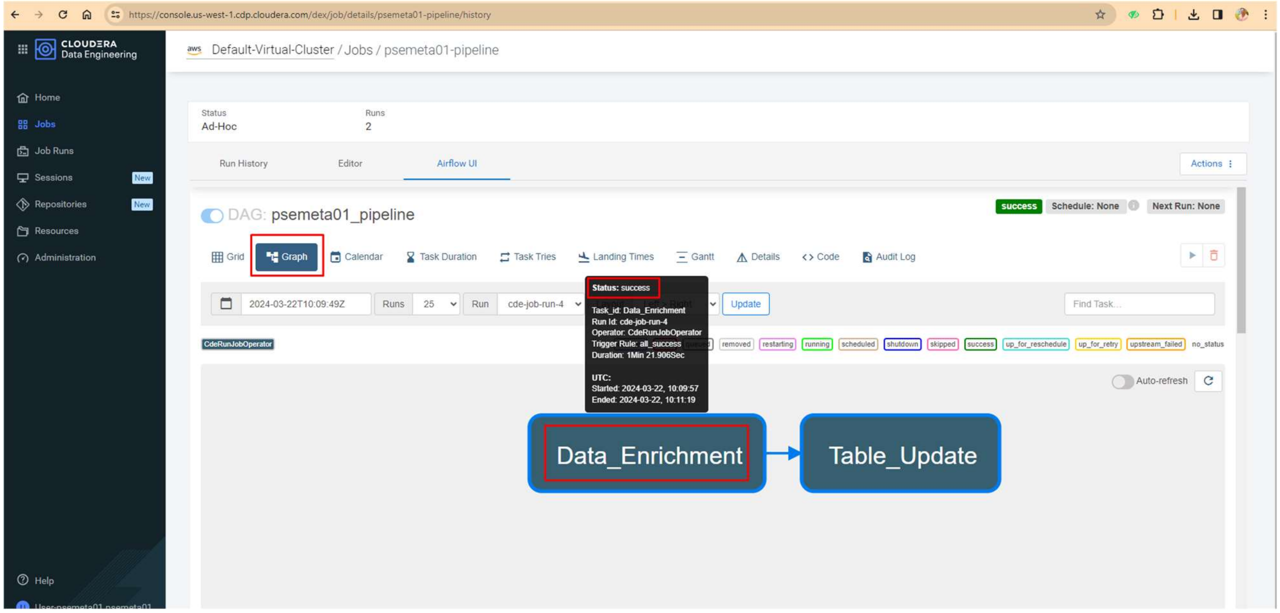
22-03-2024 10:14:28 | 25 | All Run Types | All Run States | Auto-refresh

Data_Enrichment 🟩
Table_Update 🟩

psemeta01_pipeline

DAG Runs Summary	
Total Runs Displayed	1
🟩 Total success	1
First Run Start	2024-03-22, 10:09:48 UTC
Last Run Start	2024-03-22, 10:09:48 UTC
Max Run Duration	00:02:41
Mean Run Duration	00:02:41

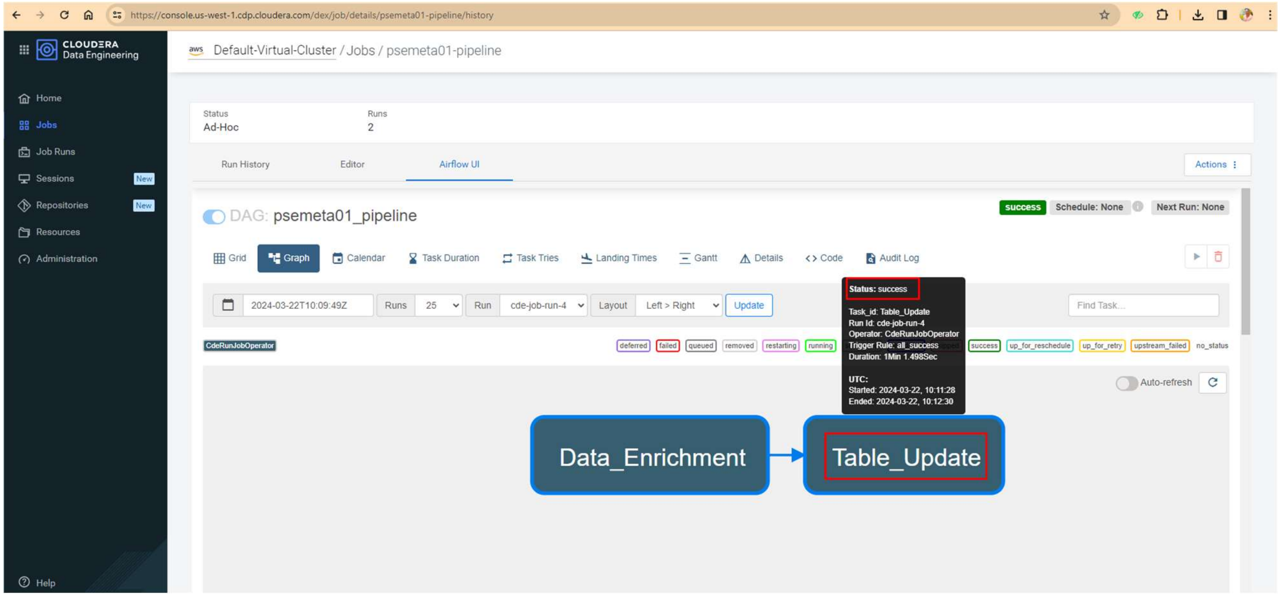
13. You can see more information about the execution by clicking on the view **Graph**. Hovering the mouse over the Job name displays specific information for each step in the pipeline. Make sure the pipeline status is Success, which indicates that the entire pipeline was able to run without issue.



Status: success

Task_id: Data_Enrichment
Run Id: cde-job-run-4
Operator: CdeRunJobOperator
Trigger Rule: all_success
Duration: 1Min 21.906Sec

UTC:
Started: 2024-03-22, 10:09:57
Ended: 2024-03-22, 10:11:19

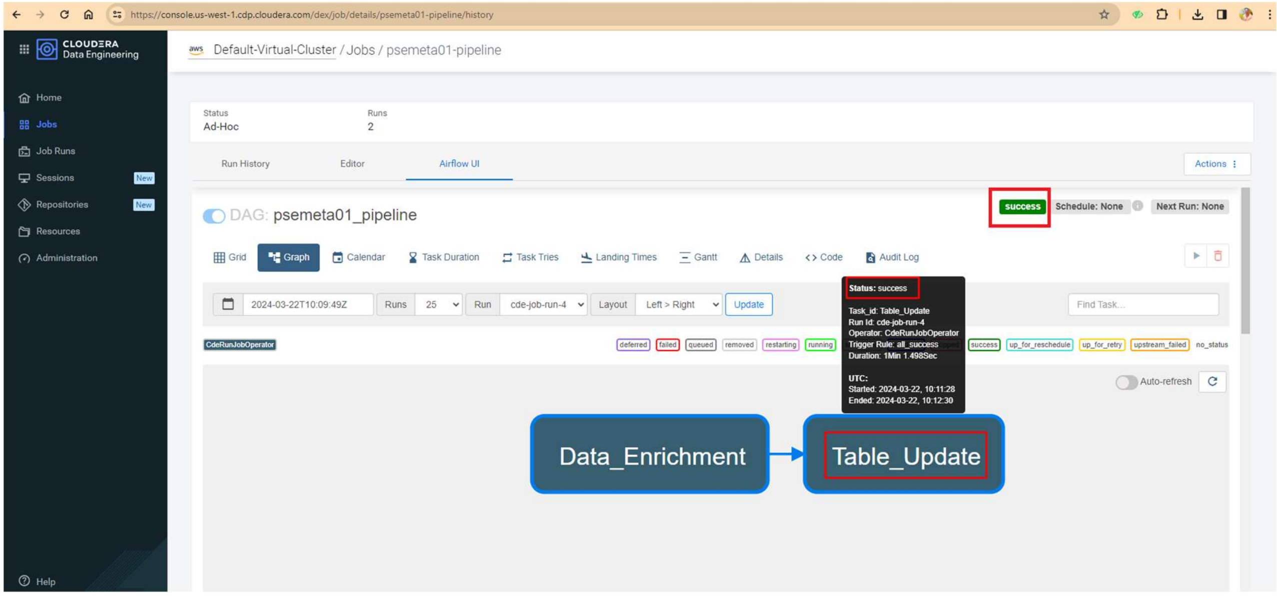


Status: success

Task_id: Table_Update
Run Id: cde-job-run-4
Operator: CdeRunJobOperator
Trigger Rule: all_success
Duration: 1Min 1.498Sec

UTC:
Started: 2024-03-22, 10:11:28
Ended: 2024-03-22, 10:12:30

*The execution status appears next to the name of the pipeline (marked in red). If it is green and indicates **Success**, it means that the execution was successful.*



https://console.us-west-1.cdp.cloudera.com/dex/job/details/psemeta01-pipeline/history

Default-Virtual-Cluster / Jobs / psemeta01-pipeline

Status
Ad-Hoc

Runs
2

DAG: psemeta01_pipeline

success Schedule: None Next Run: None

2024-03-22T10:09:49Z Runs 25 Run cde-job-run-4 Layout Left > Right Update

Status: success

Task_id: Table_Update
Run Id: cde-job-run-4
Operator: CdeRunJobOperator
Trigger Rule: all_success
Duration: 1Min 1.498Sec

UTC:
Started: 2024-03-22, 10:11:28
Ended: 2024-03-22, 10:12:30