

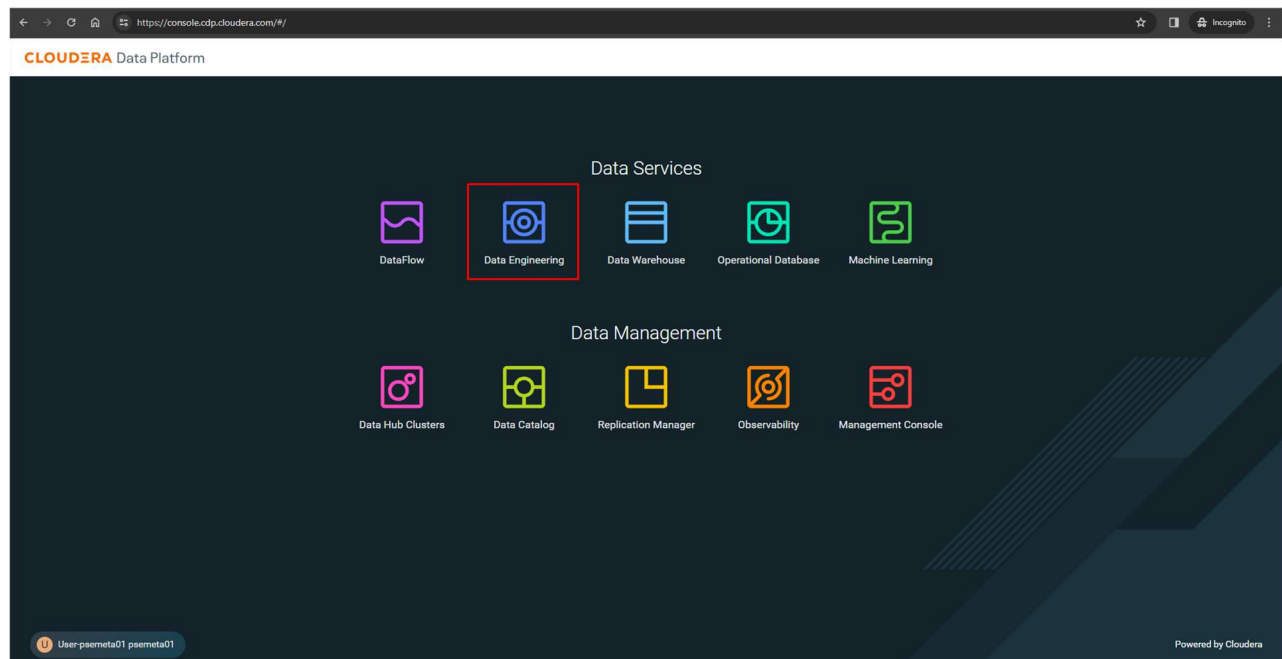
Data Lifecycle on CDP Public Cloud

Data Engineering Lab

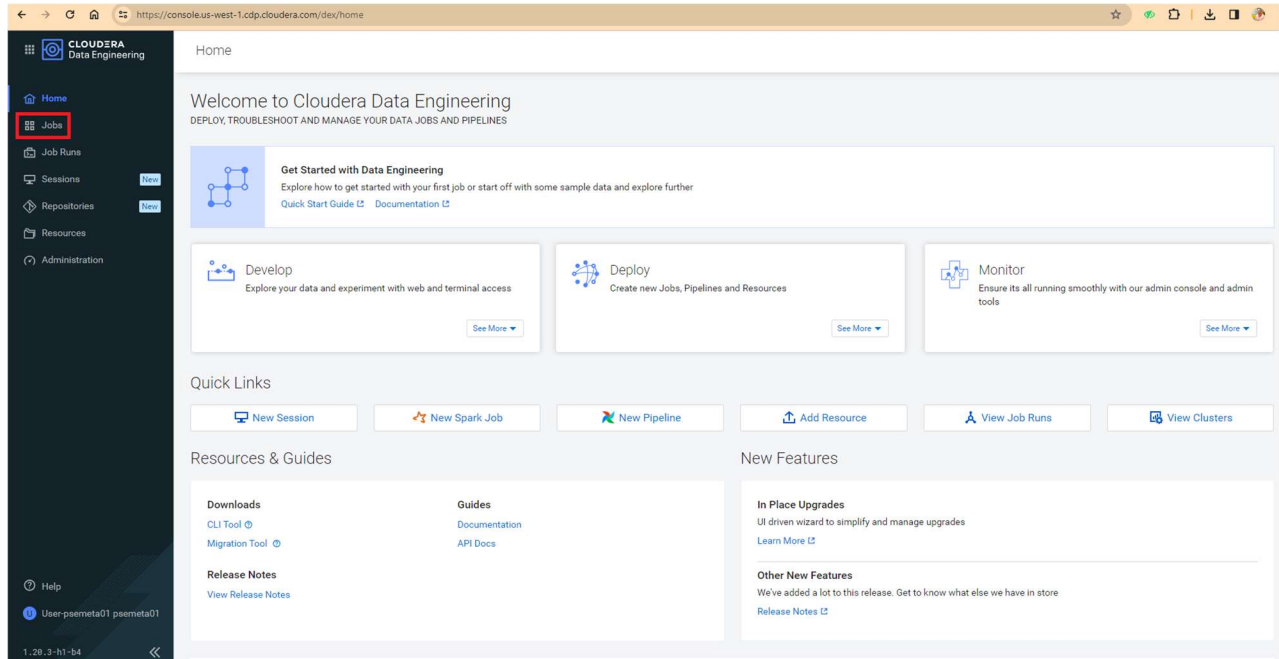
Goals:

- 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:



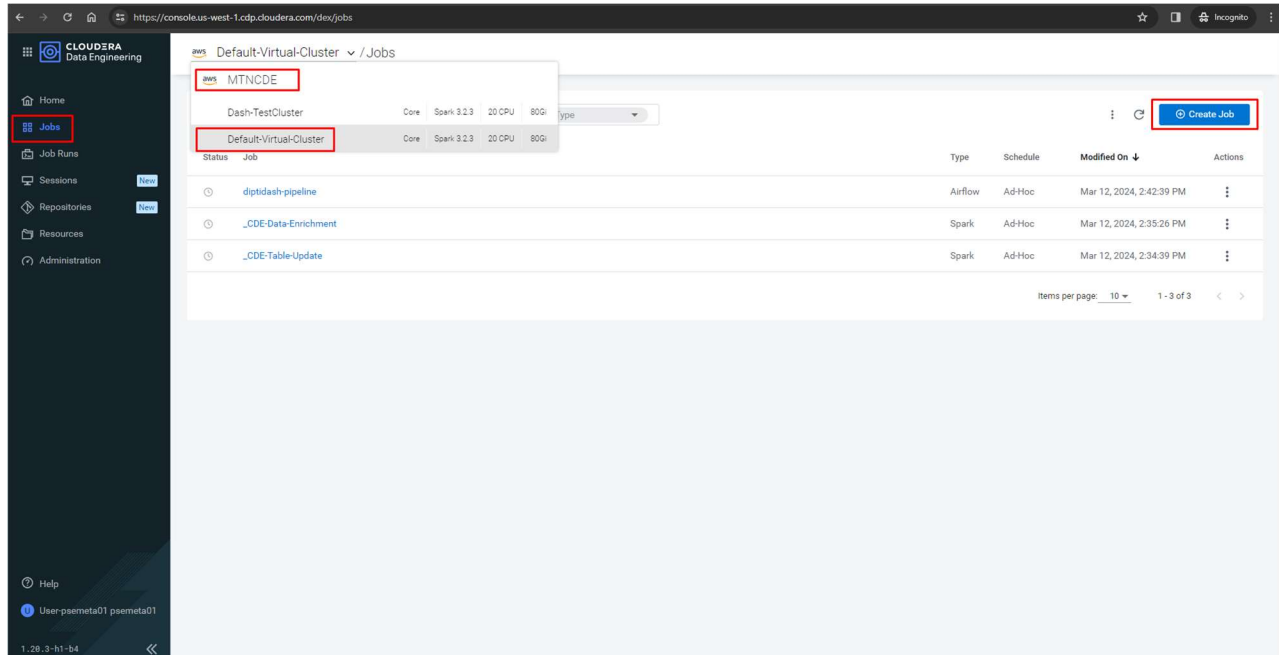
2. The Data Engineering Home shows all the actions that can be done, such as Jobs in Spark and pipelines in Airflow, Resources and useful information/documentation. Click on the option **Jobs** from the left menu to create a dataflow in Airflow.



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

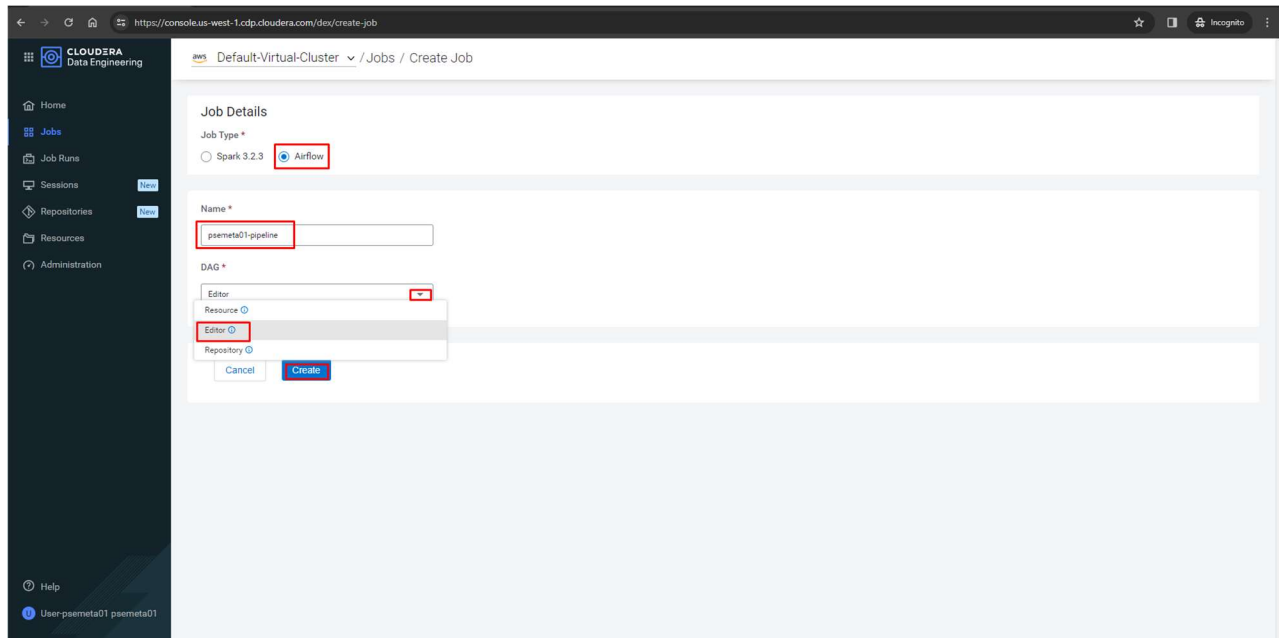
- **_CDE-Table-Update**, generate random changes and enrich table to visualize LakehouseTime Travel functionality.
- **_CDE-Data-Enrichment**, process in Spark (Python) to enrich the data ingested fromKafka and save to a new table.

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

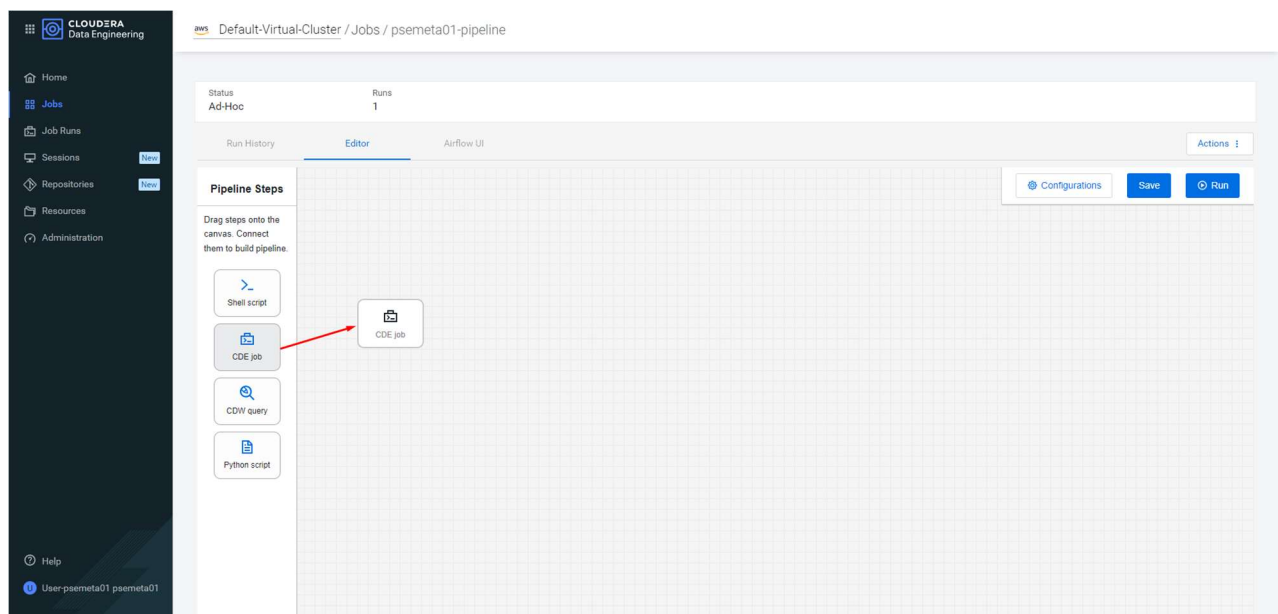


4. In the Job creation form, you must enter the following information:

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

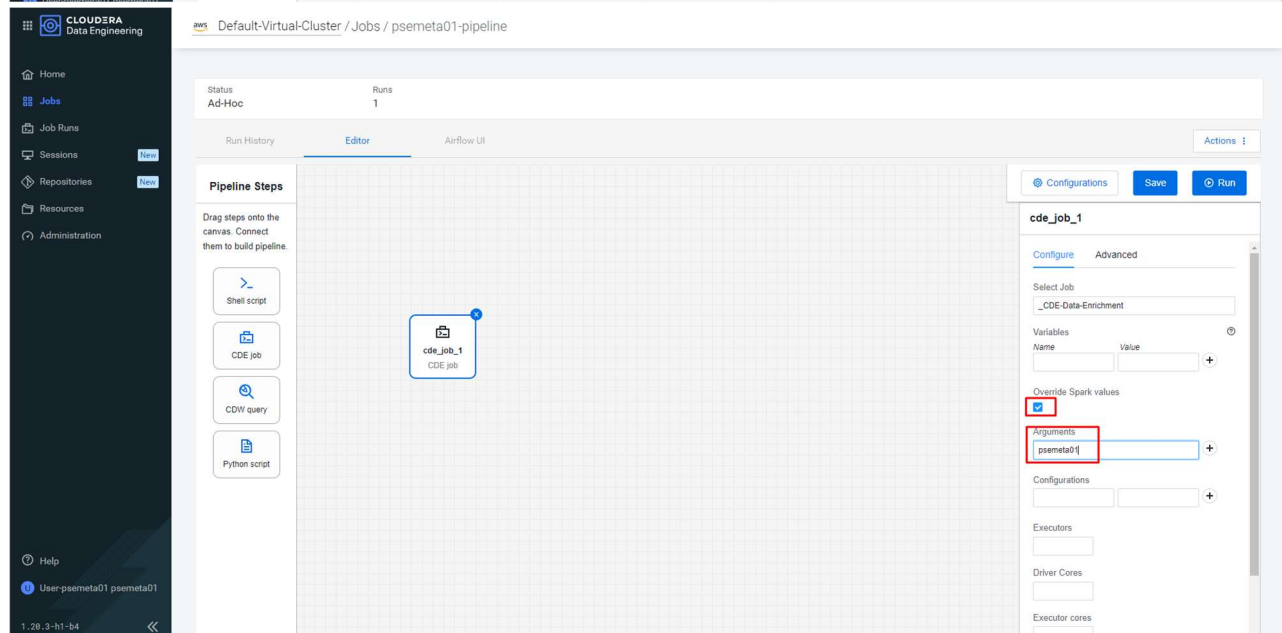
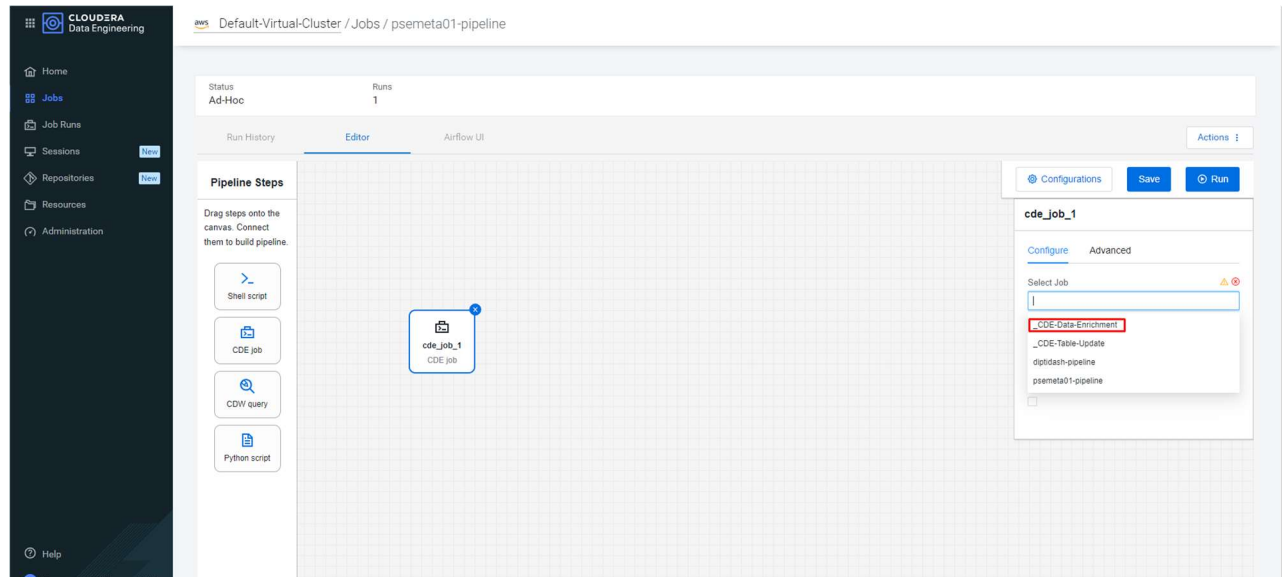


5. On the Job editing screen, select the Editor tab, and you will see the following canvas to drag the steps of the pipeline that we are going to create. In our case, we are going to create two CDE Jobs and relate them. Drag the **CDE job** into the canvas.



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**



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**

CloudERA Data Engineering console showing the pipeline editor for 'psemeta01-pipeline'.

URL: <https://console.us-west-1.cdp.cloudera.com/dev/job/details/psemeta01-pipeline/editor>

Page Header: aws Default-Virtual-Cluster / Jobs / psemeta01-pipeline

Left Sidebar:

- Home
- Jobs
- Job Runs
- Sessions
- Repositories
- Resources
- Administration
- Help
- User: psemeta01 psemeta01

Main Canvas:

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

Pipeline Steps:

- Shell script
- CDE job
- CDW query
- Python script

Canvas Content:

The canvas shows a single 'cde_job_1' step connected to another 'CDE job' step via a red line.

Right Panel:

Configurations Save Run

CloudERA Data Engineering console showing the pipeline editor for 'psemeta01-pipeline'.

URL: <https://console.us-west-1.cdp.cloudera.com/dev/job/details/psemeta01-pipeline/editor>

Page Header: aws Default-Virtual-Cluster / Jobs / psemeta01-pipeline

Left Sidebar:

- Home
- Jobs
- Job Runs
- Sessions
- Repositories
- Resources
- Administration
- Help
- User: psemeta01 psemeta01

Main Canvas:

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

Pipeline Steps:

- Shell script
- CDE job
- CDW query
- Python script

Canvas Content:

The canvas shows two 'CDE job' steps: 'cde_job_1' and 'cde_job_2'.

Right Panel:

Configurations Save Run

cde_job_2 Configuration:

Configure Advanced

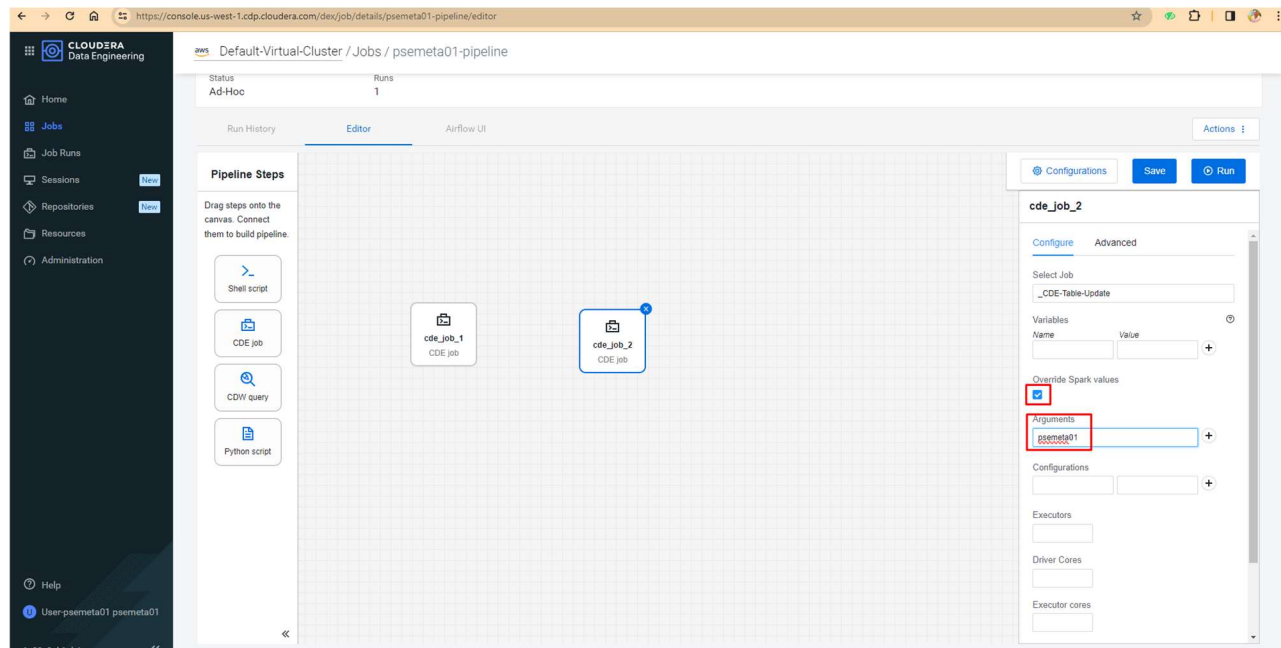
Select Job

...CDE-Data-Enrichment

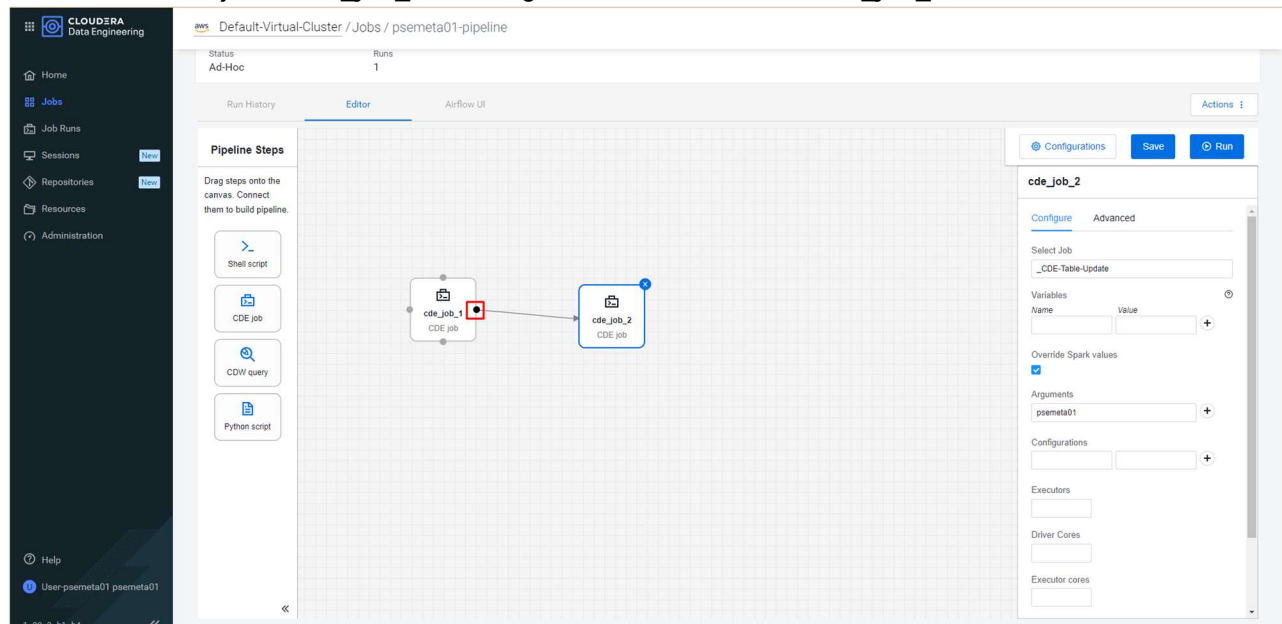
_CDE-Table-Update

dpidsat _CDE-Table-Update

psemeta01-pipeline



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**.



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/dev/job/details/psemeta01-pipeline/editor

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

Status: Ad-Hoc Runs: 1

Run History Editor Airflow UI

Pipeline Steps

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

- Shell script
- CDE job
- CDW query
- Python script

Data Enrichment

Configure Advanced

Select Job: _CDE-Data-Enrichment

Variables:

Name	Value
cde_job_1	

Override Spark values: ☒

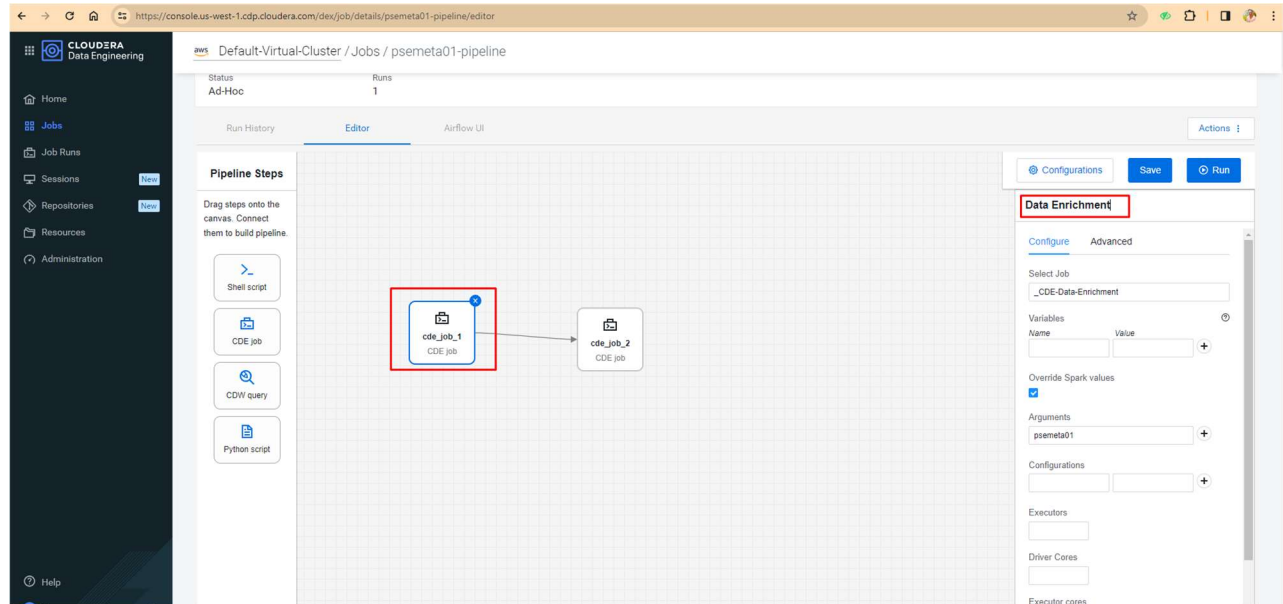
Arguments: psemeta01

Configurations:

Executors:

Driver Cores:

Executor cores:



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

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

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

Status: Ad-Hoc Runs: 1

Run History Editor Airflow UI

Pipeline Steps

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

- Shell script
- CDE job
- CDW query
- Python script

Table Update

Configure Advanced

Select Job: _CDE-Table-Update

Variables:

Name	Value
cde_job_2	

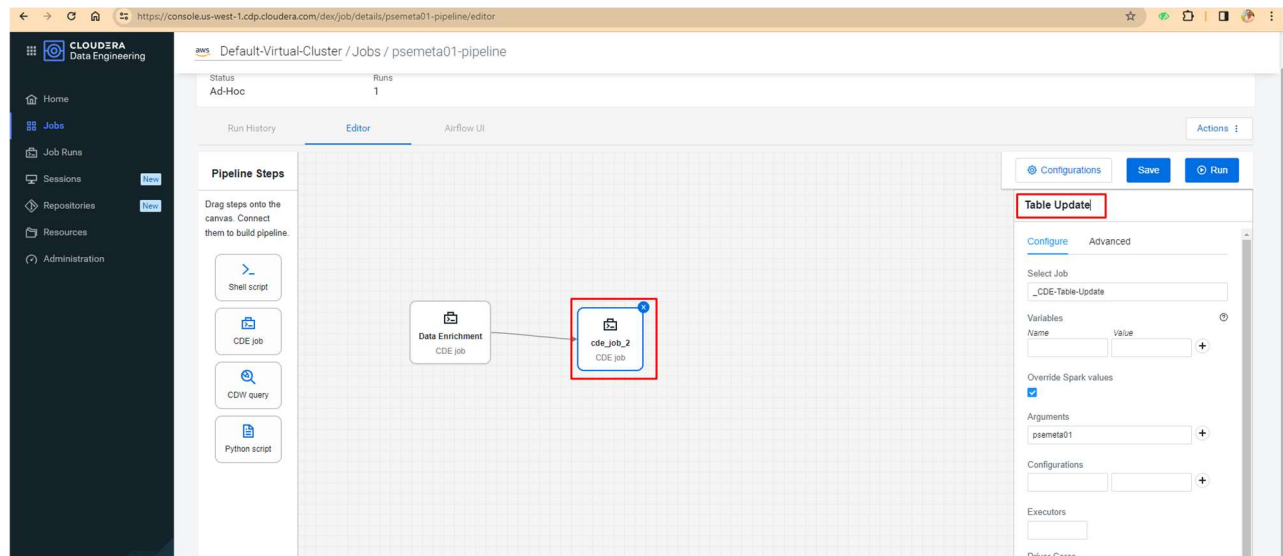
Override Spark values: ☒

Arguments: psemeta01

Configurations:

Executors:

Driver Cores:

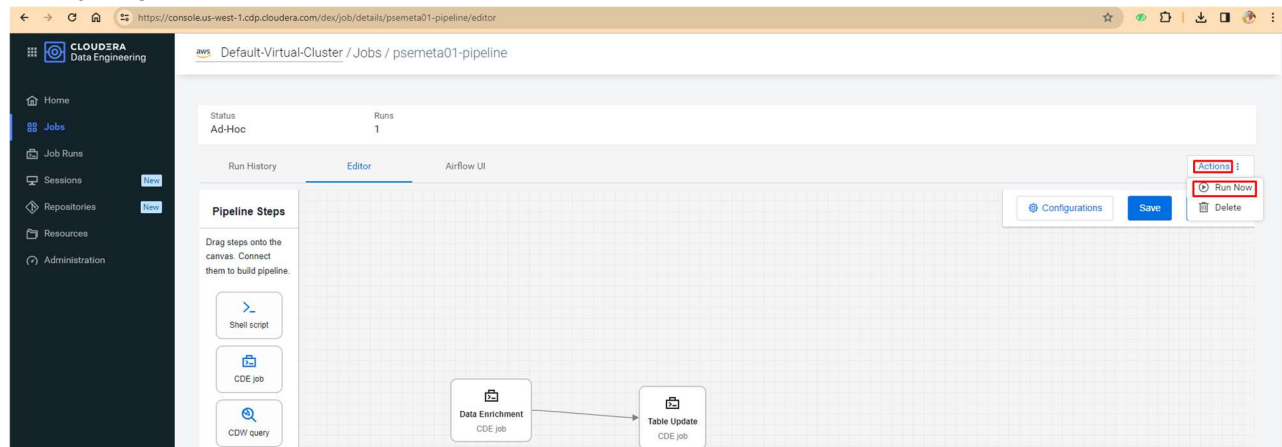


9. Once the Jobs have been joined, click on **Save** to save the settings made. You should see a message indicating **Pipeline saved to job**.

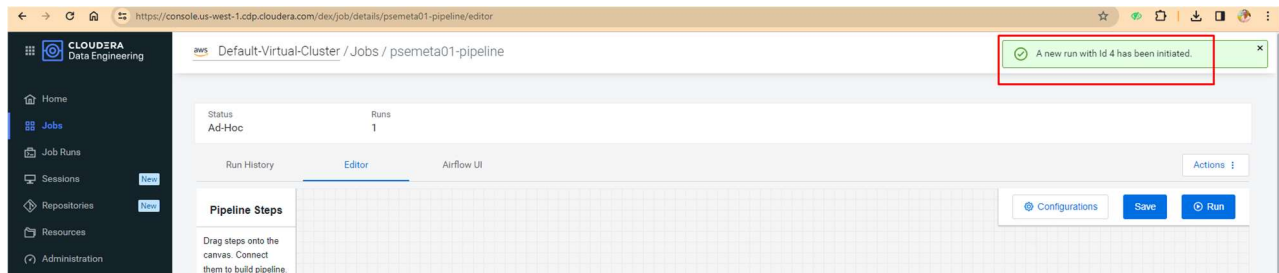
The screenshot shows the Cloudera Data Engineering console interface. The left sidebar contains navigation links: Home, Jobs, Job Runs, Sessions, Repositories, Resources, and Administration. The main area displays the 'Default-Virtual-Cluster / Jobs / psemeta01-pipeline' editor. The pipeline canvas shows two steps: 'Data Enrichment' and 'Table Update', connected by an arrow. The 'Table Update' step is highlighted with a red box. On the right, the 'Table Update' configuration panel is visible, with the 'Save' button highlighted by a red box. The status bar at the top indicates 'Status: Ad-Hoc' and 'Runs: 1'.

This screenshot shows the same Cloudera Data Engineering console interface after the pipeline has been saved. A green message box with the text 'Pipeline saved to job' is displayed in the center of the pipeline canvas. The pipeline steps, 'Data Enrichment' and 'Table Update', remain the same. The 'Save' button in the top right corner is still visible. The status bar at the top continues to show 'Status: Ad-Hoc' and 'Runs: 1'.

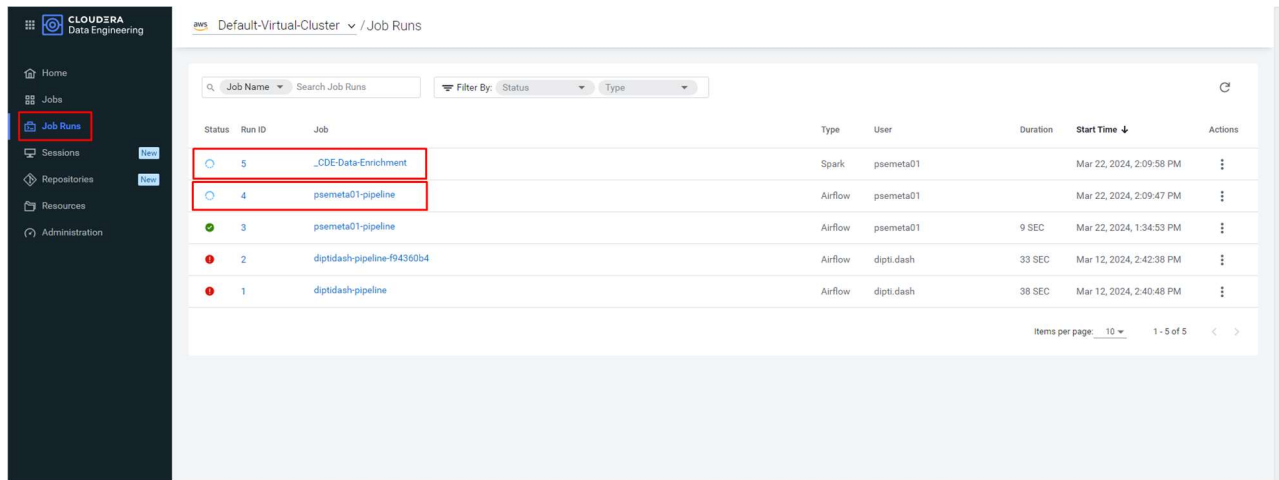
10. The time has come to run the pipeline. On the upper right side of the canvas, click **Actions** -> **Run Now**.



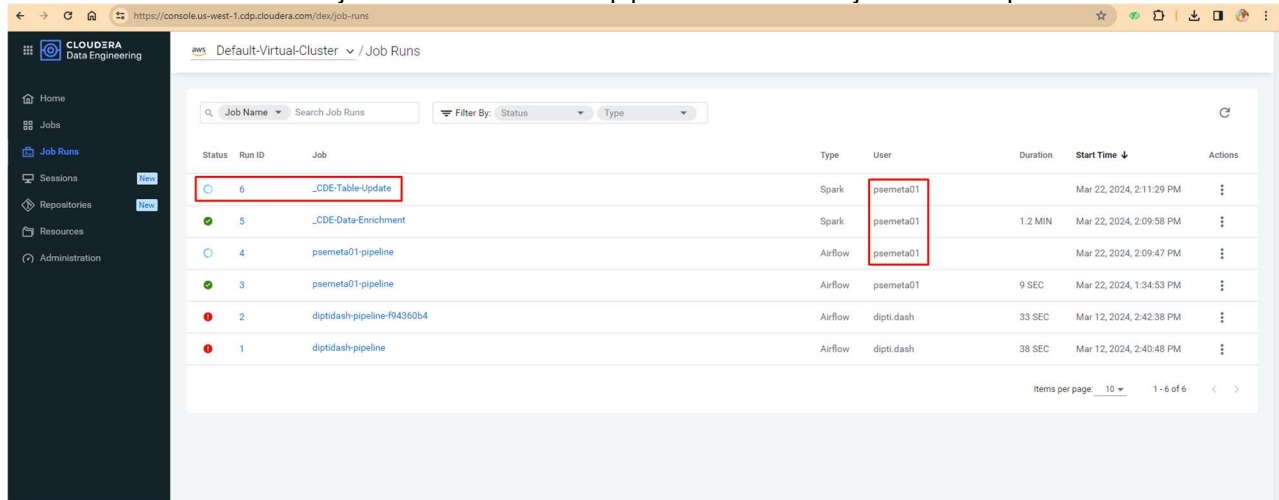
11. You should see the pipeline execution screen, indicating that the execution has been initialized.



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



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



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

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.

Default-Virtual-Cluster / Jobs / psemeta01-pipeline

Status: Ad-Hoc | Runs: 2

Run History | Editor | **Airflow UI** | Actions

DAG: psemeta01_pipeline | Schedule: None | Next Run: None

Grid | Graph | Calendar | Task Duration | Task Times | Landing Times | Gantt | Details | Code | Audit Log

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

Task Status Legend: deferred, failed, queued, removed, restarting, **running**, scheduled, shutdown, stopped, success, up_for_reschedule, up_for_retry, upstream_failed, no_status

Task List (Left):

- Data_Enrichment (Green)
- Table_Update (Green)

More Details (Right):

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.

The screenshot shows the Cloudera Data Engineering console interface. The left sidebar contains navigation links: Home, Jobs, Job Runs, Sessions, Repositories, Resources, and Administration. The main panel displays the 'psemeta01_pipeline' DAG in the 'Graph' view. The pipeline status is 'success'. A tooltip for the 'Data_Enrichment' task is visible, showing its status as 'success' and execution details. The DAG consists of two tasks: 'Data_Enrichment' and 'Table_Update'.

The screenshot shows the Cloudera Data Engineering console interface. The left sidebar contains navigation links: Home, Jobs, Job Runs, Sessions, Repositories, Resources, and Administration. The main panel displays the 'psemeta01_pipeline' DAG in the 'Graph' view. The pipeline status is 'success'. A tooltip for the 'Table_Update' task is visible, showing its status as 'success' and execution details. The DAG consists of two tasks: 'Data_Enrichment' and 'Table_Update'.

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/dev/job/details/psemeta01-pipeline/history

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

Status: Ad-Hoc Runs: 2

Run History Editor Airflow UI Actions

DAG: psemeta01_pipeline

Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details Code Audit Log

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

CdeRunJobOperator

deferred failed queued removed restarting running

success up_for_reschedule up_for_retry upstream_failed no_status

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

Auto-refresh

Data_Enrichment → Table_Update

success

Schedule: None Next Run: None

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

