

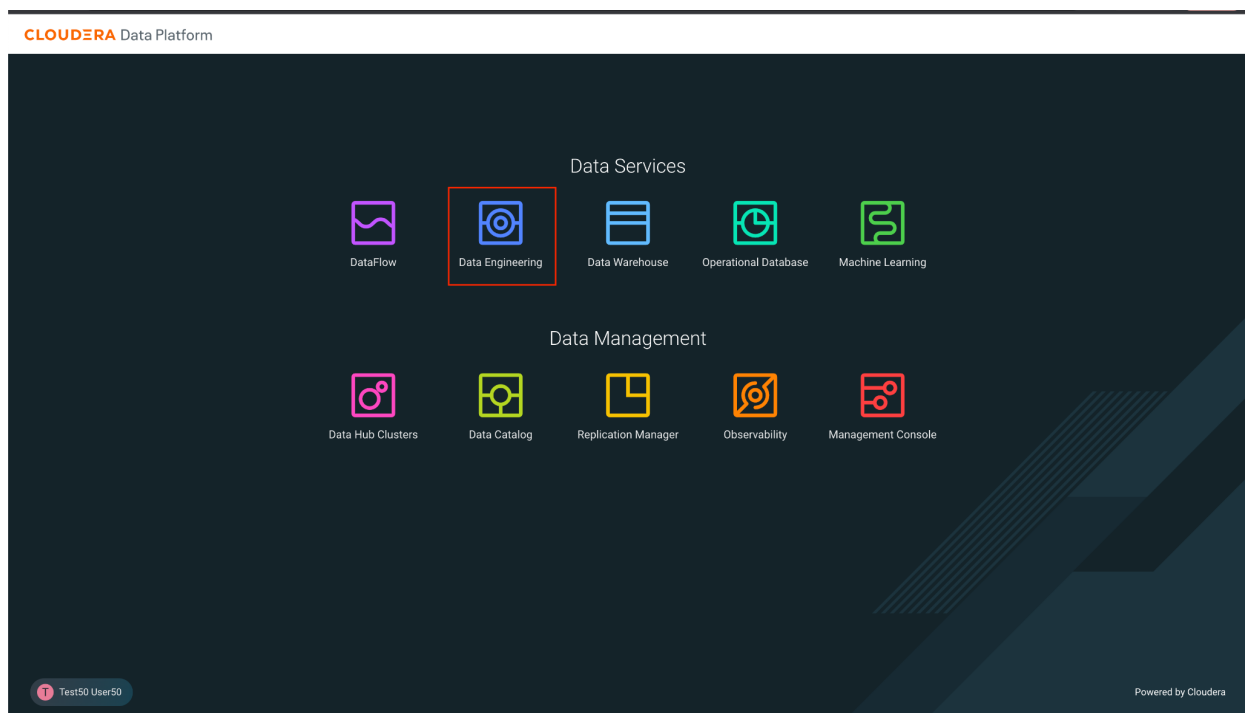
Data Lifecycle on CDP Public Cloud

Data Engineering Lab

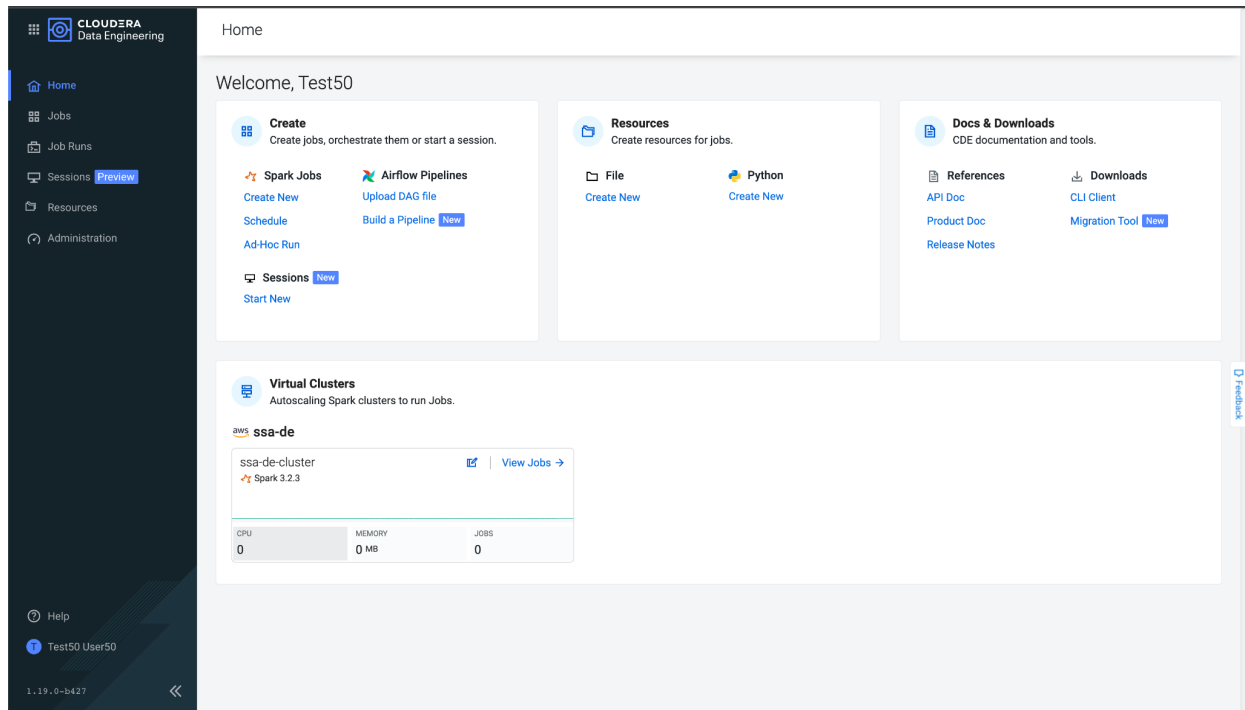
Goals:

- 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 DataFlow from CDP PC Home:



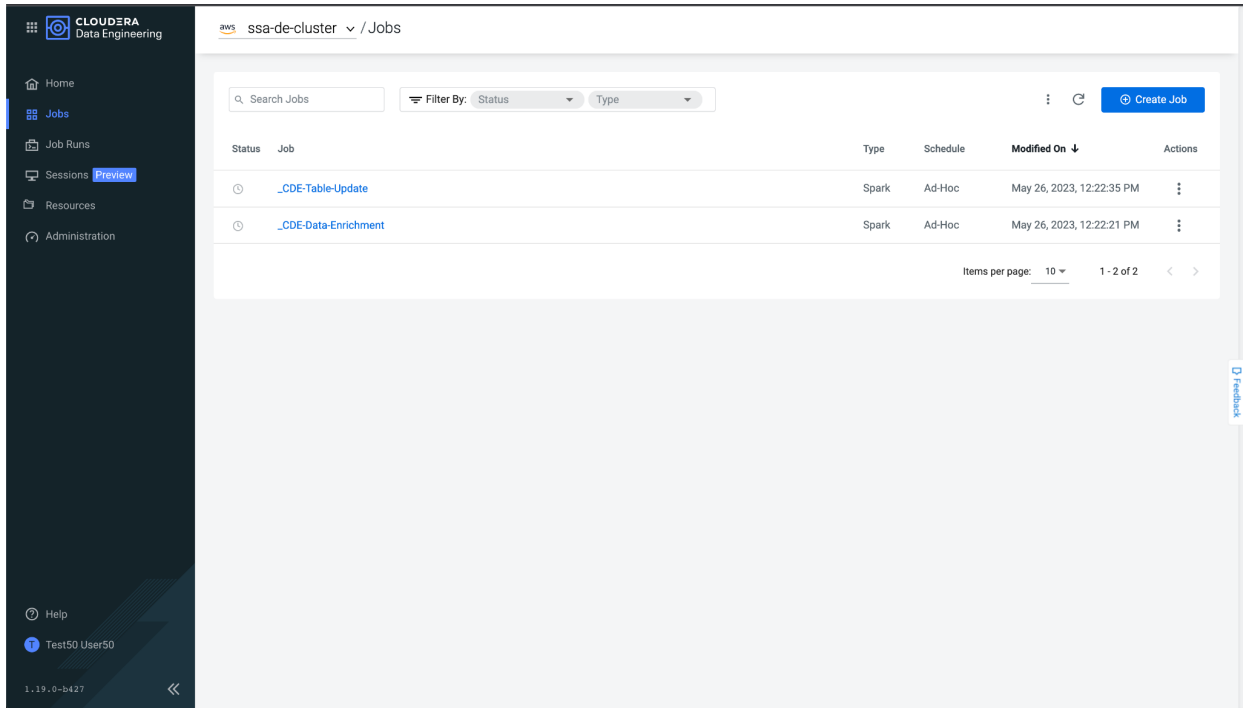
2. The Data Engineering Home shows all the actions that can be done, such as Jobs in Spark and pipelines in Airflow, Resources and useful information/documentation. Click on the option **Jobs** from the left menu to create a dataflow in Airflow.



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 Lakehouse Time Travel functionality.
- **CDE-Data-Enrichment**, process in Spark (Python) to enrich the data ingested from Kafka and save to a new table.

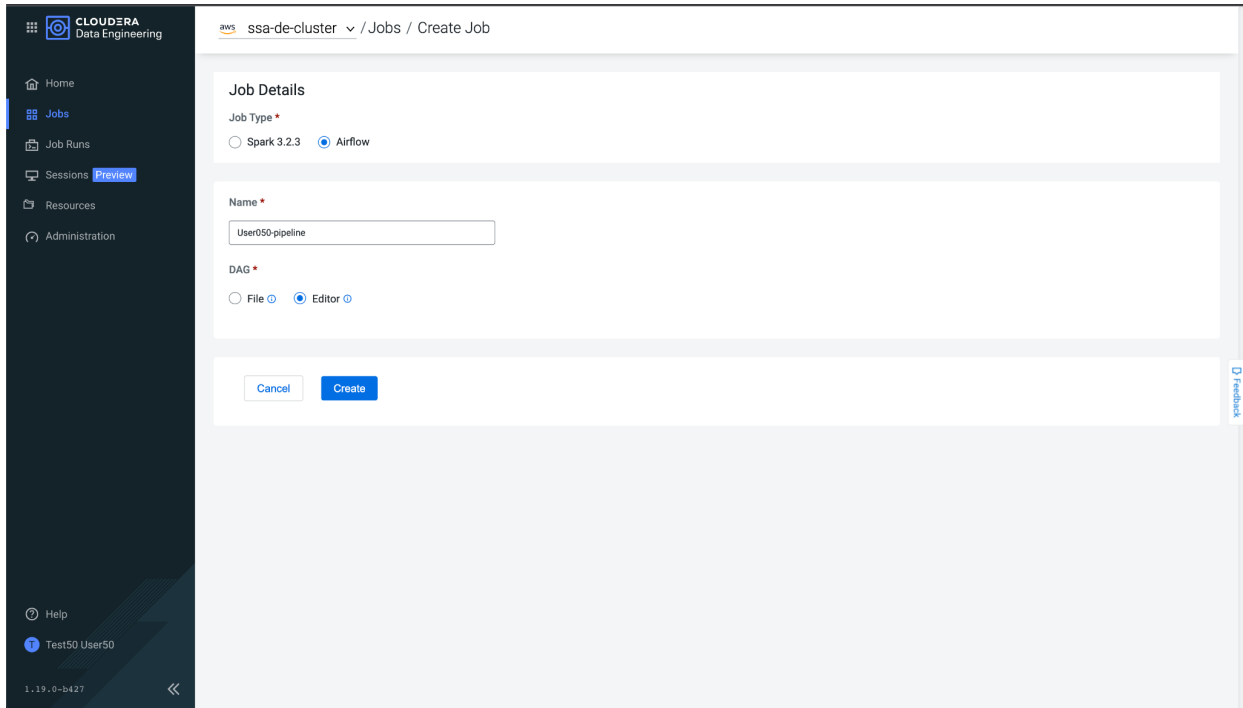
It is time to create our Job in Airflow. Click on **Create Job**.



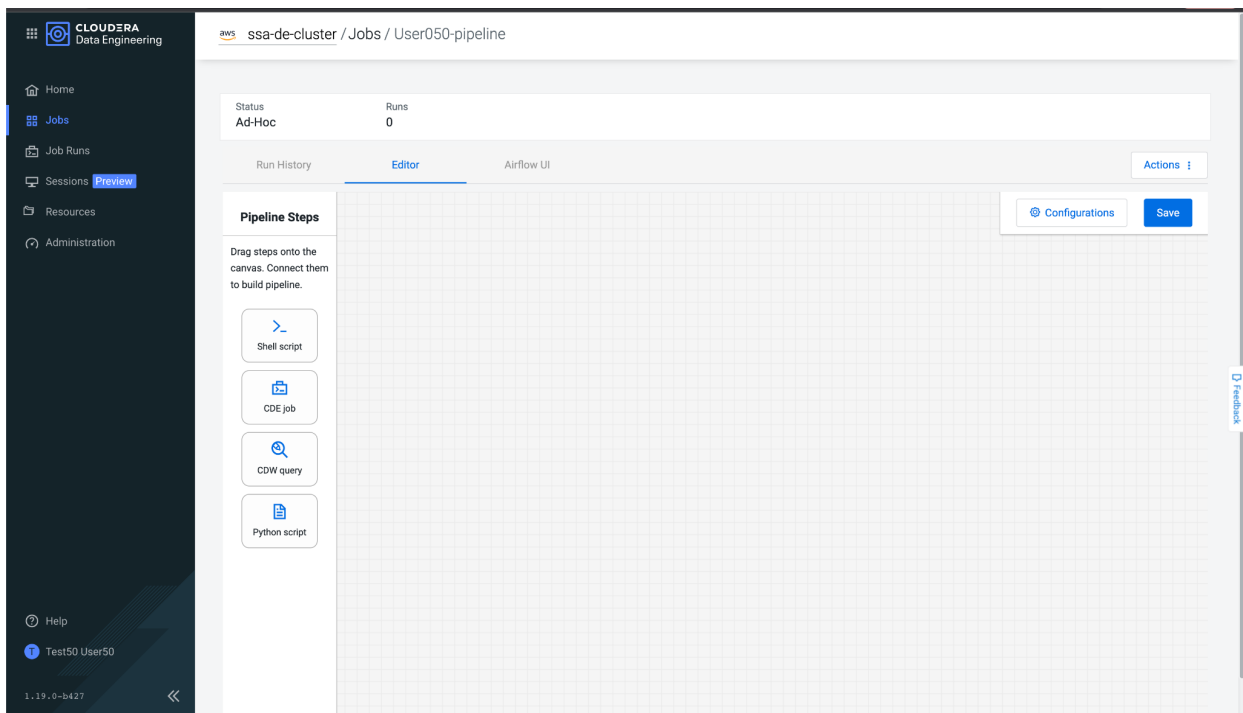
4. In the Job creation form, you must enter the following information:

- Job Type: Airflow
- Name: Use the naming <assigned user>-pipeline. Replace <assigned user> with the user assigned to you. For example, user050
- DAG: Editor, to graphically configure the task.

Once entering the values correctly, click on **Create**.

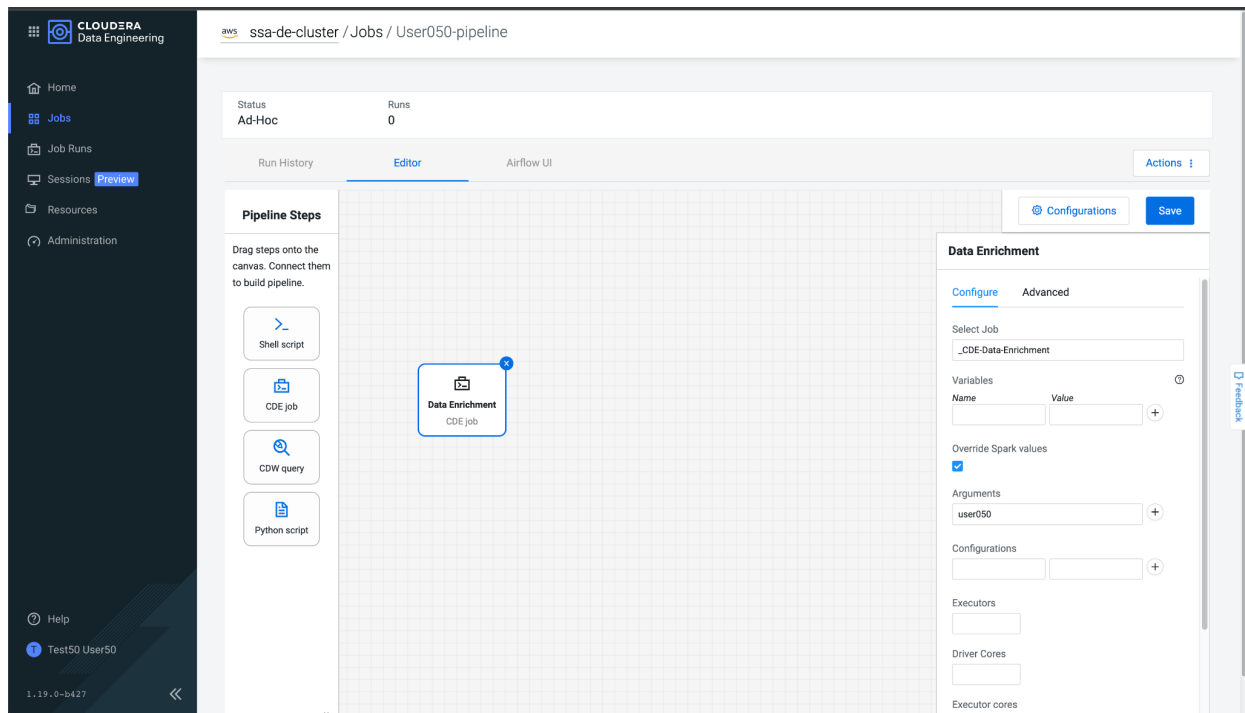


5. On the Job editing screen, select the Editor tab, and you will see the following canvas to drag the steps of the pipeline that we are going to create. In our case, we are going to create two CDE Jobs and relate them.



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

- **title/name:** Data Enrichment
- **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, user050



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

- **title/name:** Table Update
- **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, user050

The screenshot shows the Cloudera Data Engineering interface. On the left is a sidebar with navigation links: Home, Jobs, Job Runs, Sessions (with a 'Preview' button), Resources, and Administration. Below these are 'Help' and 'Test50 User50' with a version number '1.19.0-b427'. The main panel is titled 'aws ssa-de-cluster / Jobs / User050-pipeline'. It has tabs for 'Run History', 'Editor' (selected), and 'Airflow UI'. At the top right of the editor are 'Configurations' and 'Save' buttons. The 'Pipeline Steps' panel on the left lists 'Shell script', 'CDE job', 'CDW query', and 'Python script'. The canvas shows two jobs: 'Data Enrichment' and 'Table Update', both labeled 'CDE job'. The 'Table Update' job is highlighted with a blue border. On the right, the 'Table Update' configuration panel is open, showing 'Configure' and 'Advanced' tabs. The 'Configure' tab is active, displaying fields for 'Select Job' (set to '_CDE-Table-Update'), 'Variables' (Name and Value fields), 'Override Spark values' (checked), 'Arguments' (set to 'user050'), 'Configurations' (empty), 'Executors' (empty), 'Driver Cores' (empty), and 'Executor cores' (empty). A 'Feedback' button is visible on the far right edge.

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

This screenshot shows the same Cloudera Data Engineering interface, but with the 'Data Enrichment' job selected. The 'Table Update' configuration panel is no longer open. The 'Data Enrichment' job is highlighted with a red box, and its right connector is visible. The 'Table Update' job remains on the canvas. The sidebar and top navigation are identical to the previous screenshot. The status bar at the top still shows 'Ad-Hoc' status and 0 runs. The 'Feedback' button is still visible on the far right edge.

CLUSTER: aws ssa-de-cluster / Jobs / User050-pipeline

Status: Ad-Hoc | Runs: 0

Run History | Editor | Airflow UI

Actions: Configurations | Save

Pipeline Steps

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

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

Data Enrichment
CDE job

Table Update
CDE job

Feedback

Help | Test50 User50 | 1.19.0-b427

CLUSTER: aws ssa-de-cluster / Jobs / User050-pipeline

Status: Ad-Hoc | Runs: 0

Run History | Editor | Airflow UI

Actions: Configurations | Save

Pipeline Steps

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

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

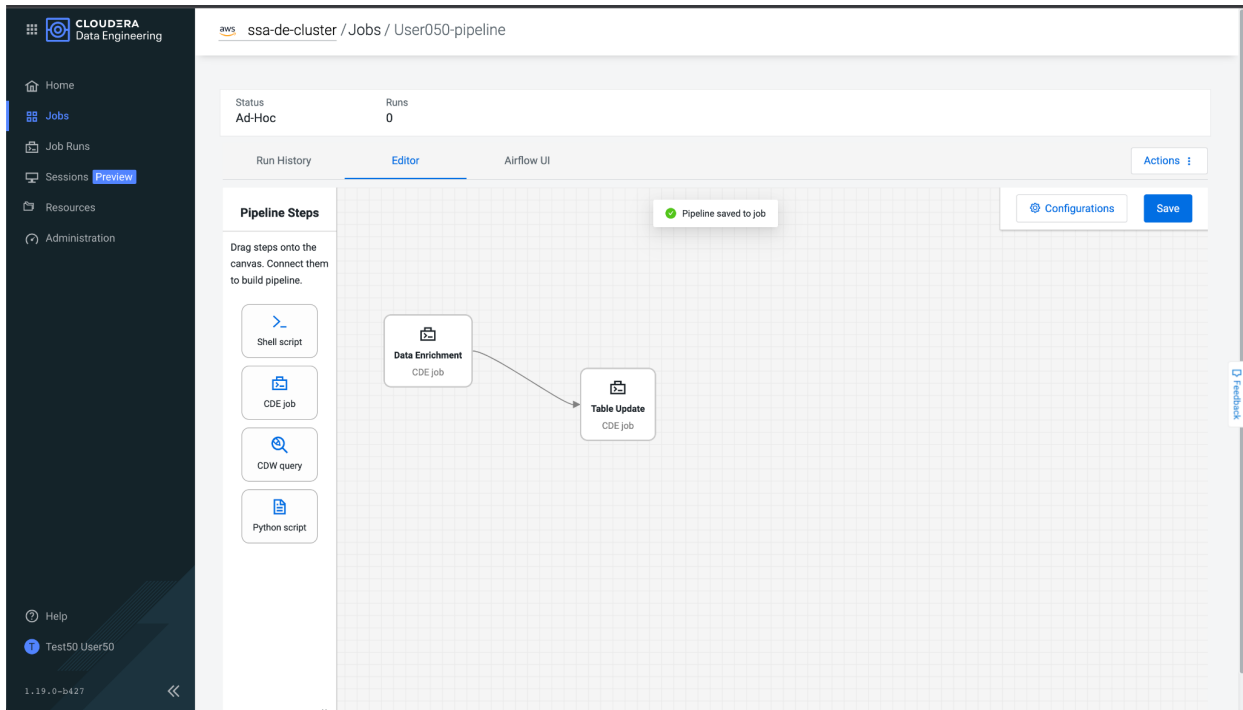
Data Enrichment
CDE job

Table Update
CDE job

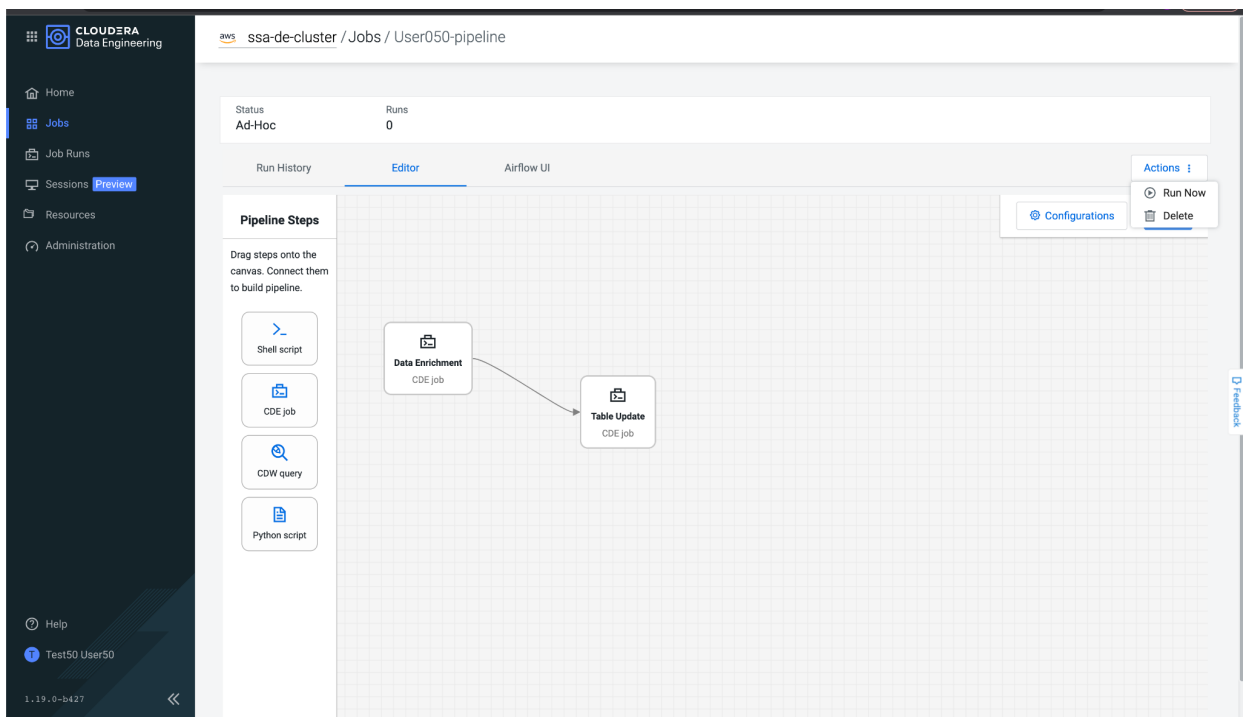
Feedback

Help | Test50 User50 | 1.19.0-b427

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**.



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.

CloudERA Data Engineering

Home Jobs Job Runs Sessions **Preview** Resources Administration

Help Test50 User50 1.19.0-b427

aws ssa-de-cluster / Jobs / User050-pipeline

Status: Ad-Hoc Runs: 0

Run History Editor Airflow UI Actions

Duration

Search by Run Id

Status	Run ID	Duration	User	Start Time ↓	Actions
Ad-Hoc	7		user050	May 26, 2023, 1:32:09 PM	

Items per page: 10 1 - 1 of 1

A new run with Id 7 has been initiated.

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.

CloudERA Data Engineering

Home Jobs Job Runs Sessions **Preview** Resources Administration

Help Test50 User50 1.19.0-b427

aws ssa-de-cluster / Jobs / User050-pipeline

Status: Ad-Hoc Runs: 0

Run History Editor **Airflow UI** Actions

DAG: User050_pipeline

Schedule: None Next Run: None

Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details <> Code

Audit Log

26/05/2023, 18:32:26 25 All Run Types All Run States Clear Filters

Auto-refresh ☒

Duration

00:00:21

00:00:10

00:00:00

Data_Enrichment Table_Update

Legend: deferred failed queued running scheduled skipped success up_for_reschedule up_for_retry upstream_failed no_status

DAG Details


DAG Runs Summary

Total Runs Displayed	1
Total running	1
First Run Start	2023-05-26, 18:32:10 UTC
Last Run Start	2023-05-26, 18:32:10 UTC
Max Run Duration	00:00:21

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 interface. On the left is a sidebar with navigation links: Home, Jobs, Job Runs, Sessions (with a 'Preview' button), Resources, and Administration. The main content area is titled 'aws ssa-de-cluster / Jobs / User050-pipeline'. It displays the pipeline's status as 'Ad-Hoc' with '1' run. Below this, there are tabs for 'Run History', 'Editor', and 'Airflow UI' (which is active). The 'Airflow UI' tab shows a DAG for 'User050_pipeline' with a 'success' status. A 'Graph' button is highlighted with a red box. A tooltip is visible over the 'Data_Enrichment' task, showing details: Task Id: Data_Enrichment, Run: 2023-05-26, 18:36:24 UTC, Run Id: cde-job-run-7, Operator: CdeRunJobOperator, Duration: 1Min 11.676Sec, and UTC start/end times. The DAG itself shows 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.*

 CLOUDERA
Data Engineering

Home

Jobs

Job Runs

Sessions Preview

Resources

Administration

Help

Test50 User50

1.19.0-b427

aws ssa-de-cluster / Jobs / User050-pipeline

Status
Ad-Hoc

Runs
1

Run History

Editor

Airflow UI

Actions

DAG: User050_pipeline

SUCCESS Schedule: None Next Run: None

Grid

Graph

Calendar

Task Duration

Task Tries

Landing Times

Gantt

Details

Code

Audit Log

2023-05-26T18:32:11Z

Runs

25

Run

cde-job-run-7

Layout

Find Task...

CdeRunJobOperator

deferred

failed

skipped

success

up_for_reschedule

up_for_retry

upstream_failed

no_status

Data_Enrichment

Table_Update

Status: success

Task_id: Table_Update

Run: 2023-05-26, 18:36:36 UTC

Run Id: cde-job-run-7

Operator: CdeRunJobOperator

Duration: 1Min 1.533Sec

UTC:

Started: 2023-05-26, 18:34:53

Ended: 2023-05-26, 18:35:55

Update

Auto-refresh