

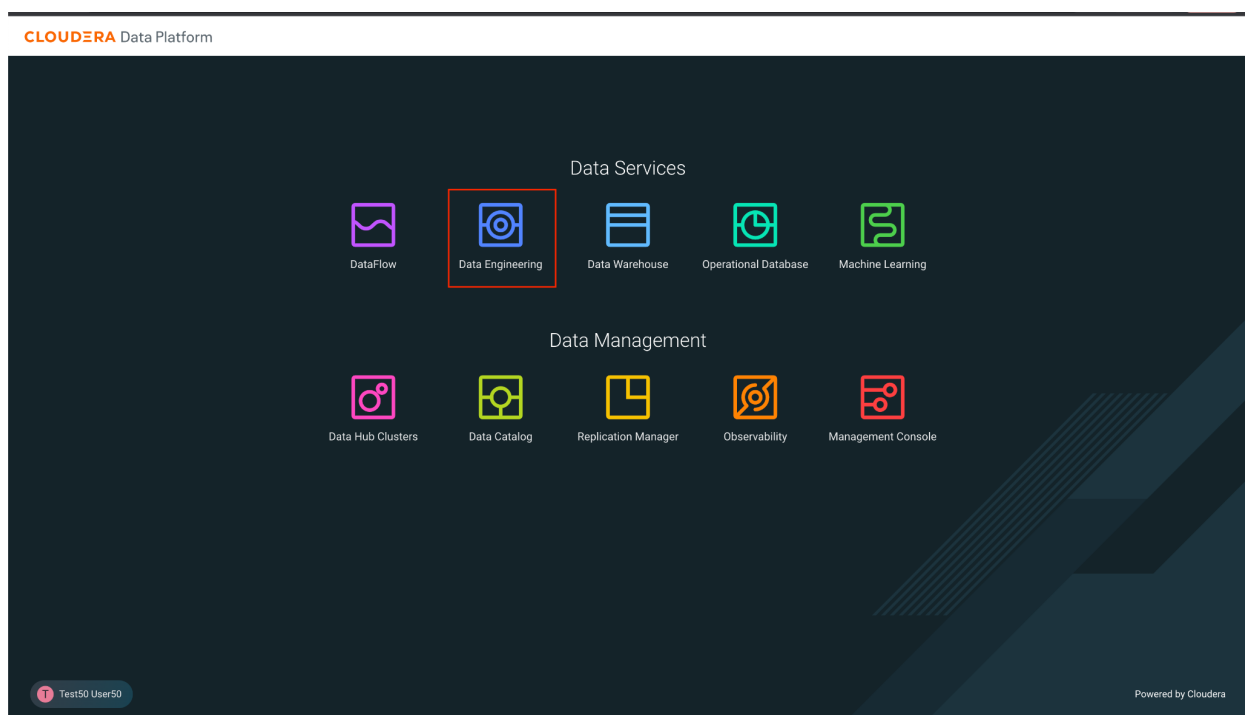
# Data Lifecycle en CDP Public Cloud

## Laboratorio Data Engineering

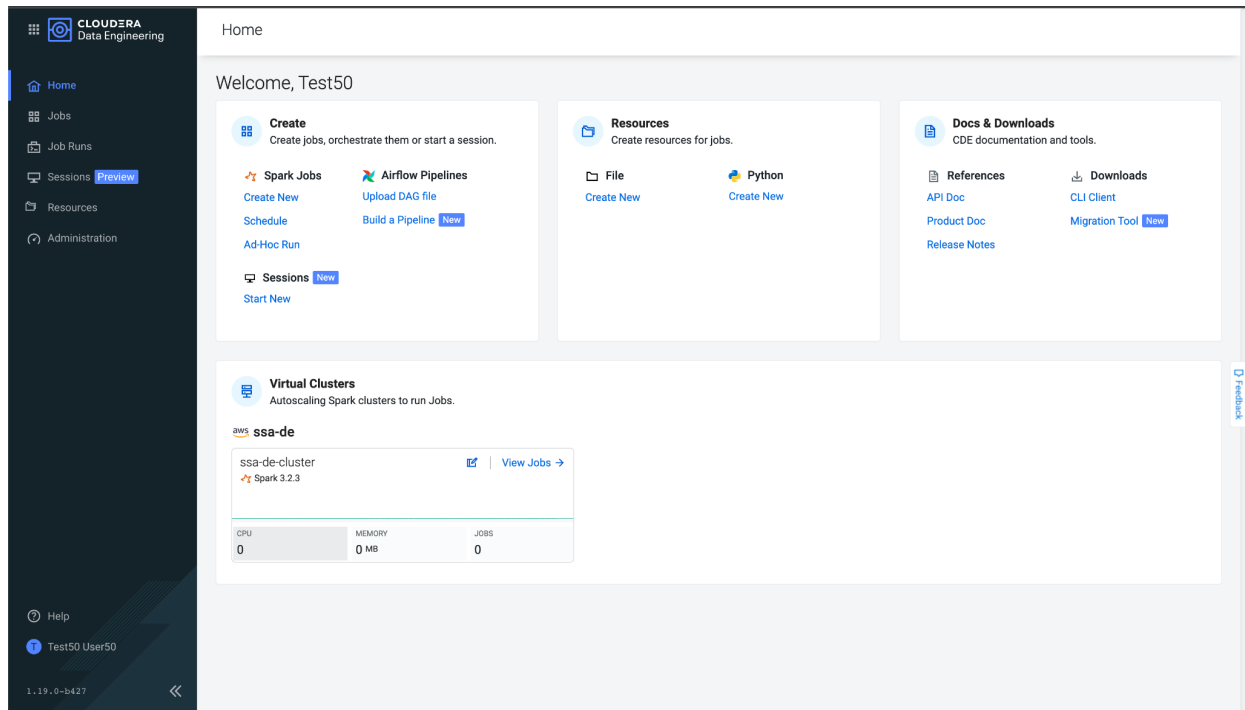
### Objetivos:

- Ejecutar un proceso de enriquecimiento de datos
- Ejecutar un proceso para simular cambios en los datos
- Configurar la ejecución de un pipeline utilizando herramientas low-code/no-code

1. Hacer clic en DataFlow desde el Home de CDP PC:



2. El Home de Data Engineering muestra todas las acciones que se pueden hacer, como Jobs en Spark y pipelines en Airflow, Resources y información/documentación útil. Hacer clic en la opción **Jobs** del menú izquierdo para crear un flujo de datos en Airflow.

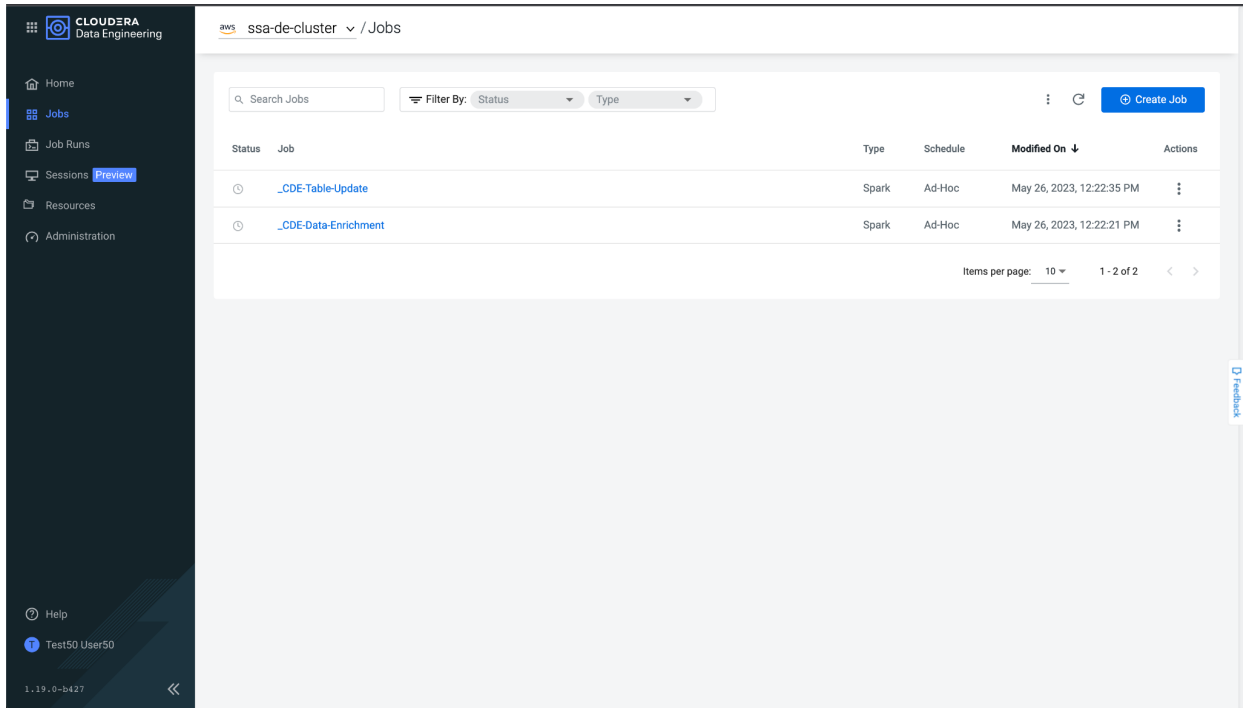


3. Aquí se listan las tareas disponibles. Para propósitos de este taller, se han configurado dos Jobs:

**\_CDE-Table-Update**, genera cambios aleatorios en tabla enriquecida para visualizar la funcionalidad de Time Travel del Lakehouse.

**\_CDE-Data-Enrichment**, proceso en Spark (Python) para enriquecer los datos ingestados desde Kafka y guardar en una nueva tabla.

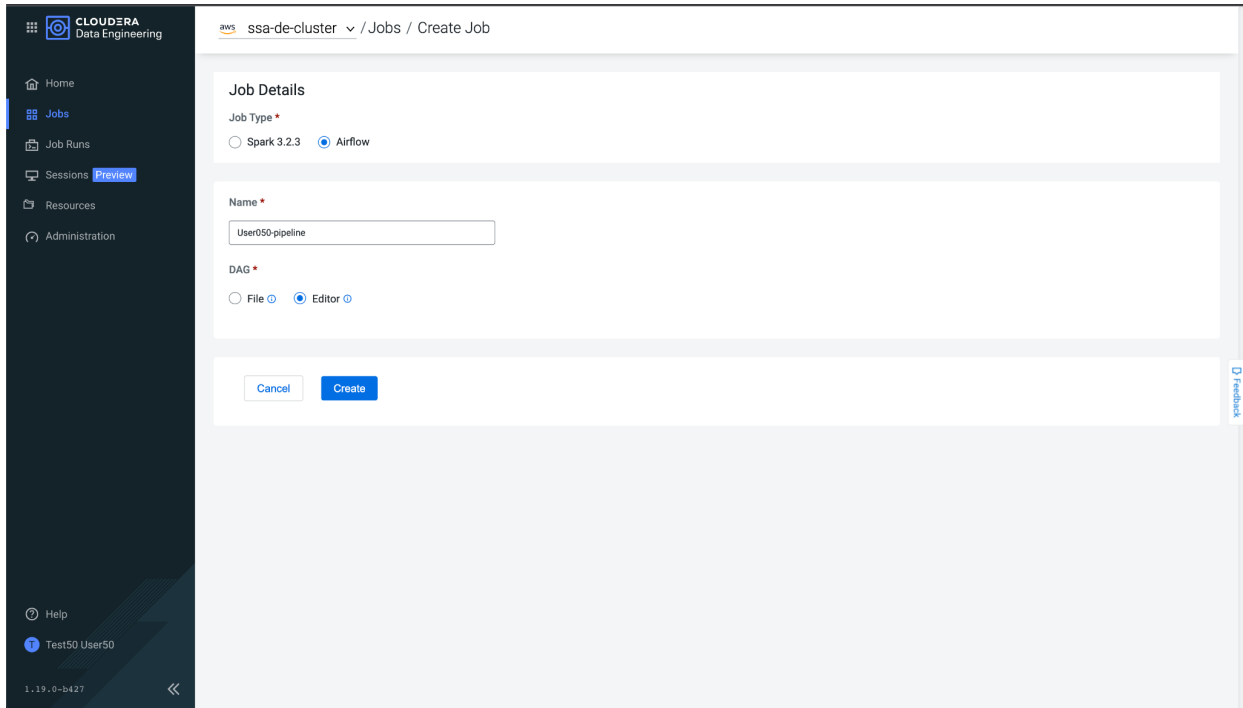
Es hora de crear nuestro Job en Airflo. Hacer clic en **Create Job**.



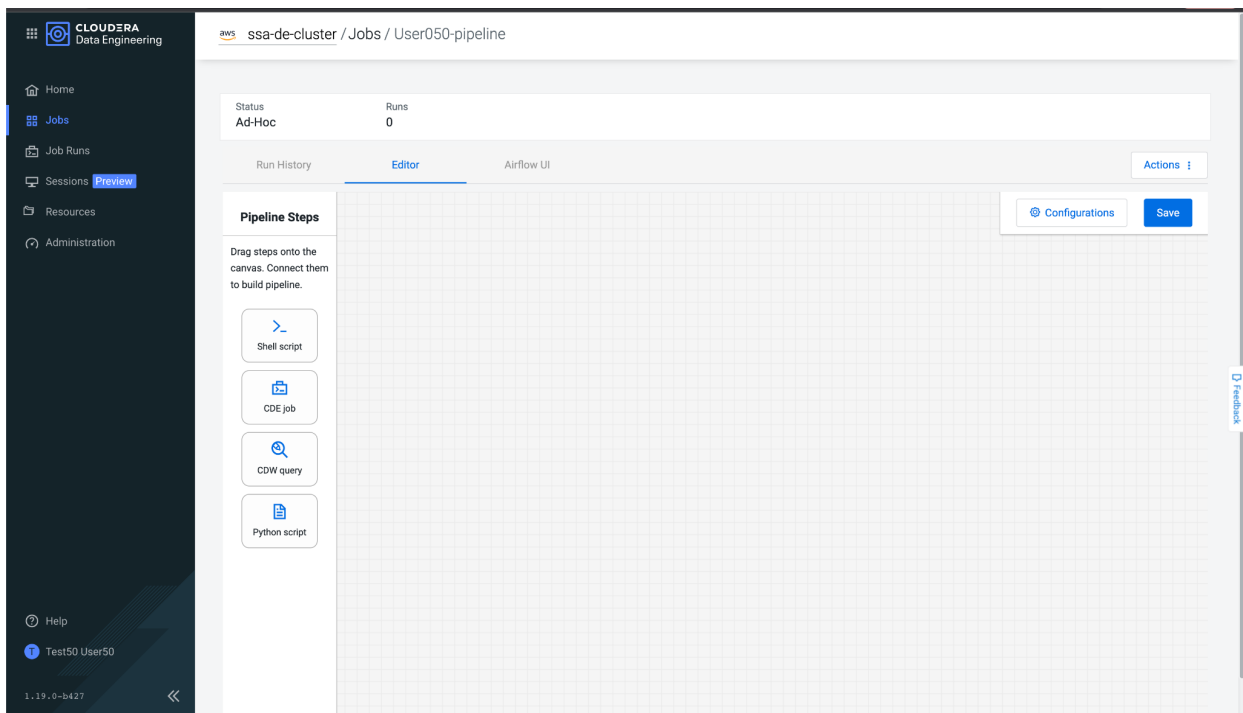
4. En el formulario de creación de Job, debe ingresar la siguiente información:

- Job Type: Airflow
- Name: utilizar la nomenclatura <usuario asignado>-pipeline. Reemplazar <usuario asignado> por el usuario que tienes asignado. Por ejemplo, user050
- DAG: Editor, para configurar gráficamente la tarea.

Una vez ingresando los valores correctamente, hacer clic en **Create**.

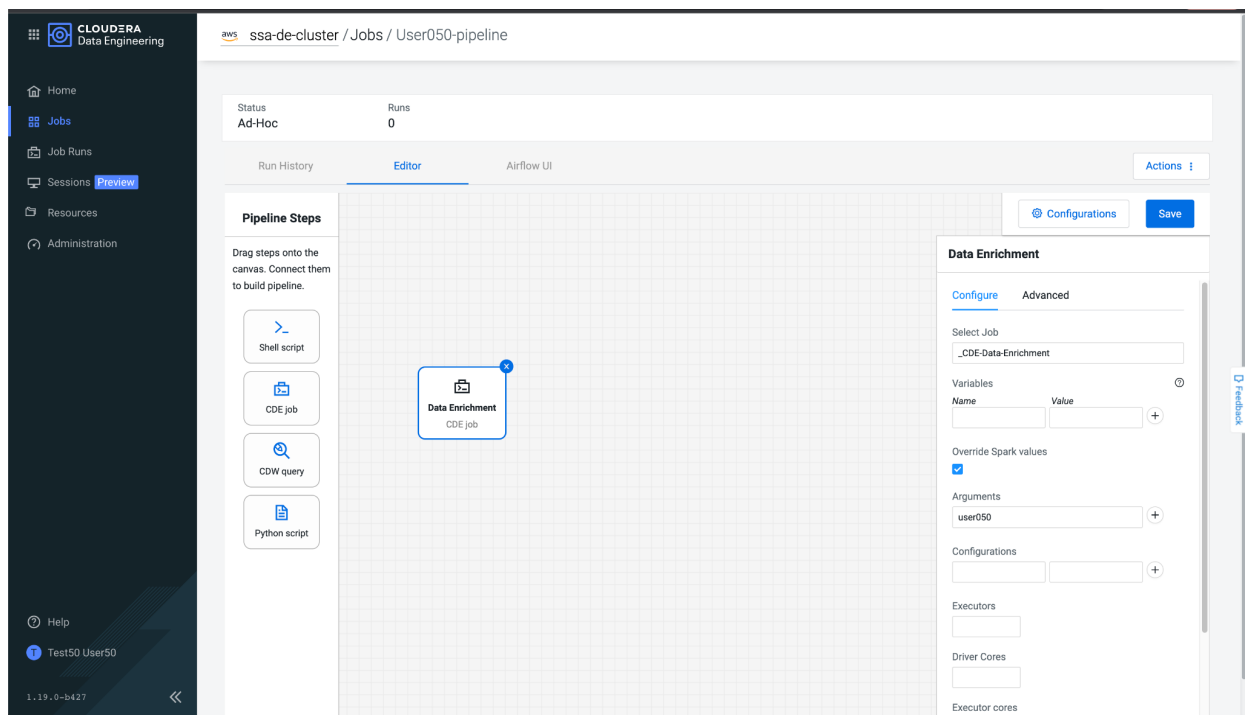


5. En la pantalla de edición del Job, seleccionar la pestaña Editor, y verás el siguiente canvas para arrastrar los pasos del pipeline que vamos a crear. En nuestro caso, vamos a crear dos tareas de CDE (CDE Jobs) y relacionarlas.



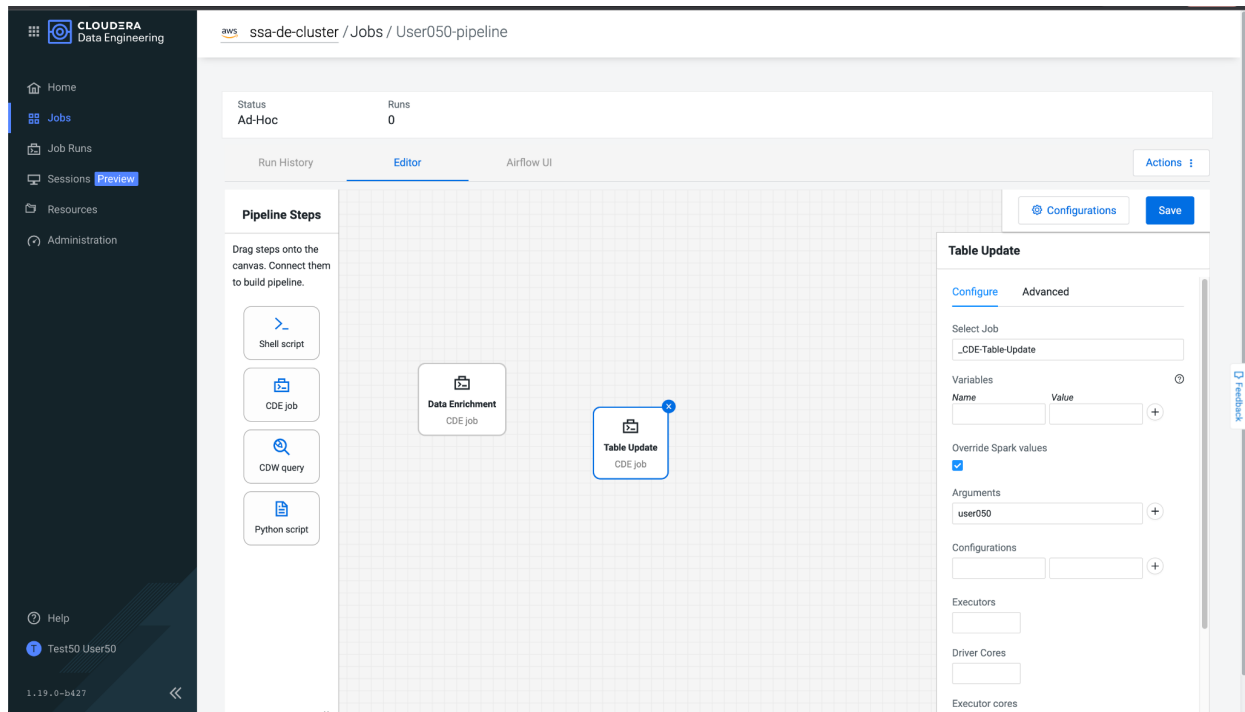
6. Empecemos con el primer Job. Hacer clic sobre el botón CDE Job y arrastrar al canvas, ingresando la siguiente configuración:

- **Título/nombre:** Data Enrichment
- **Select Job:** seleccionar el Job *\_CDE-Data-Enrichment*
- Marcar el checkbox **Override Spark values**. Aparecerán opciones adicionales luego abajo.
- **Arguments:** <usuario asignado>. User el nombre del usuario que tienes asignado. Por ejemplo, user050

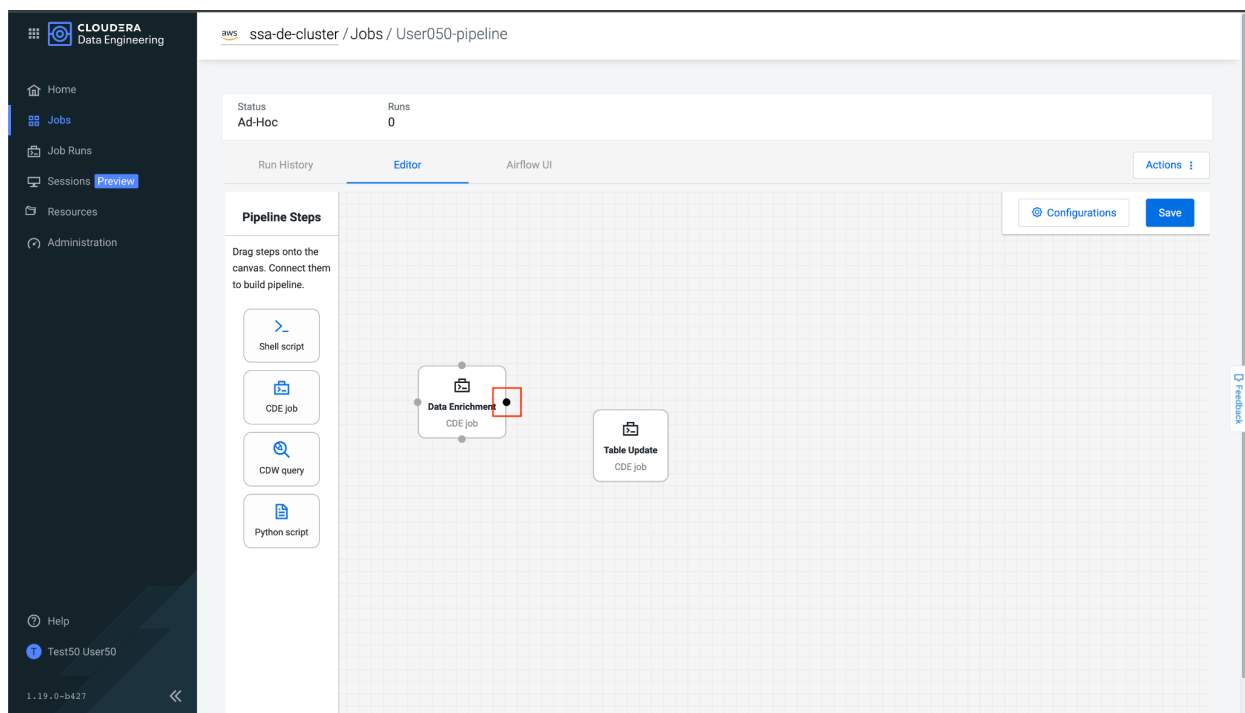


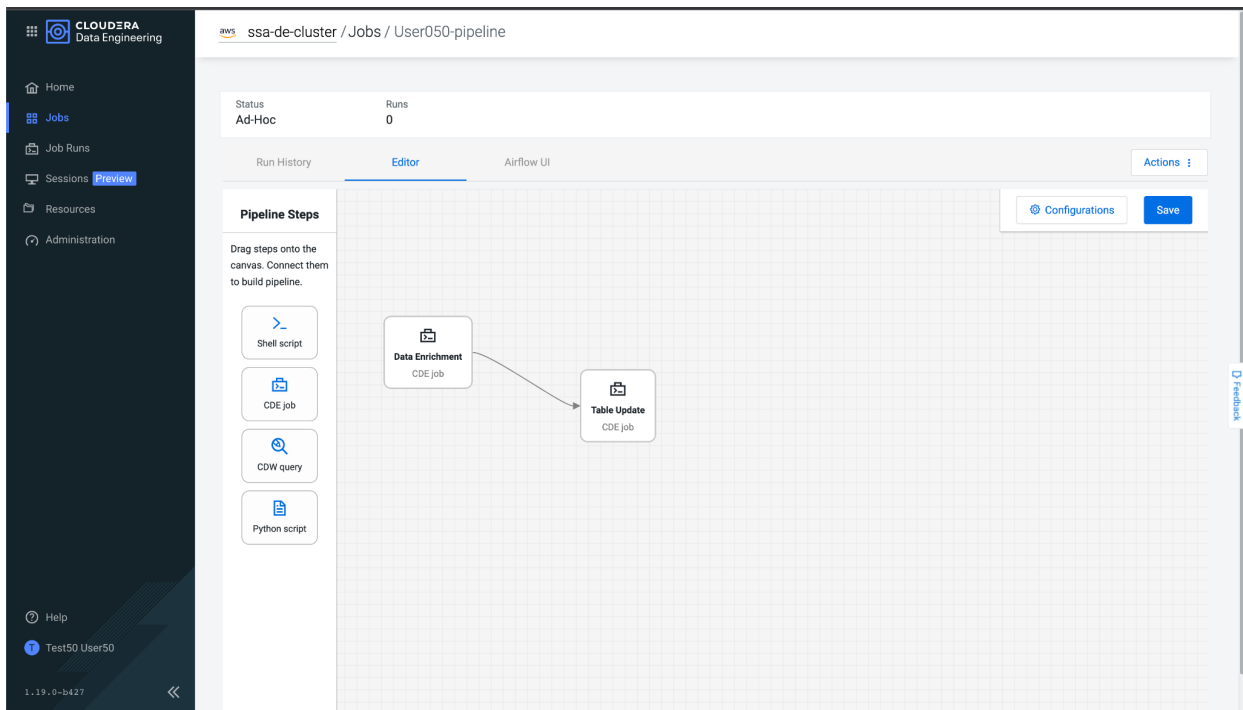
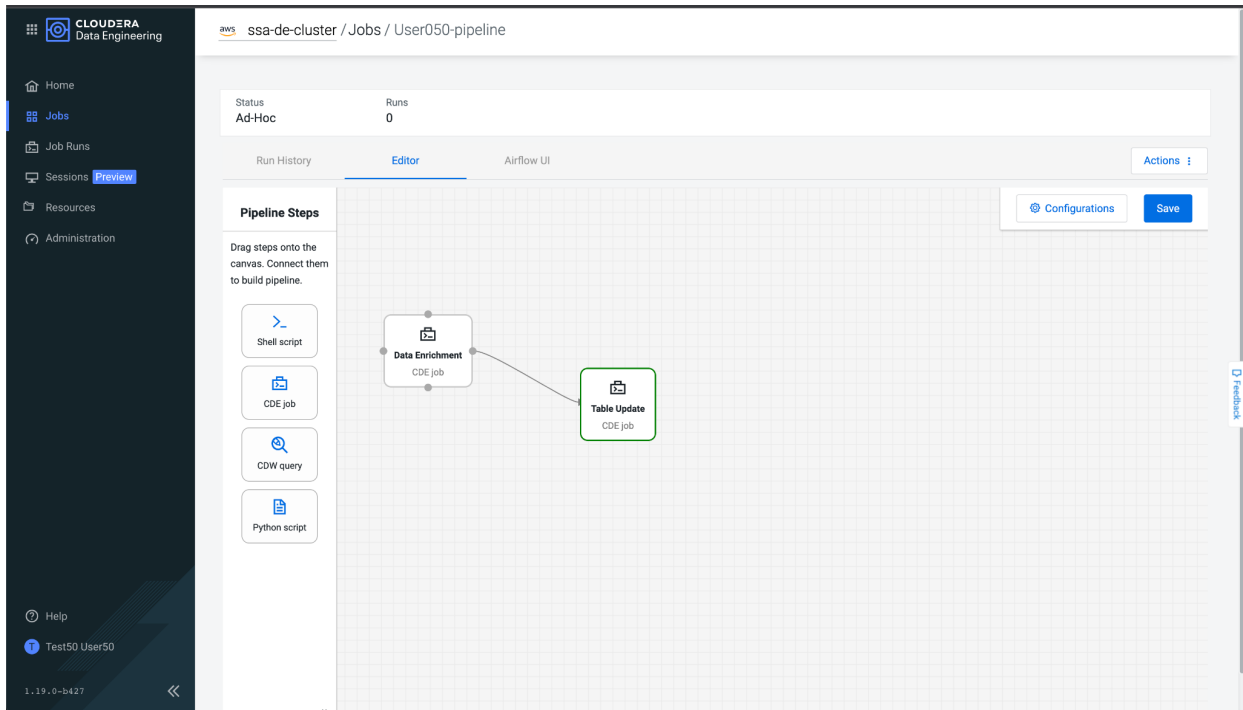
7. Configure el segundo Job. Hacer clic sobre el botón CDE Job y arrastrar al canvas, ingresando la siguiente configuración:

- **Título/nombre:** Table Update
- **Select Job:** seleccionar el Job *\_CDE-Table-Update*
- Marcar el checkbox **Override Spark values**. Aparecerán opciones adicionales luego abajo.
- **Arguments:** <usuario asignado>. User el nombre del usuario que tienes asignado. Por ejemplo, user050

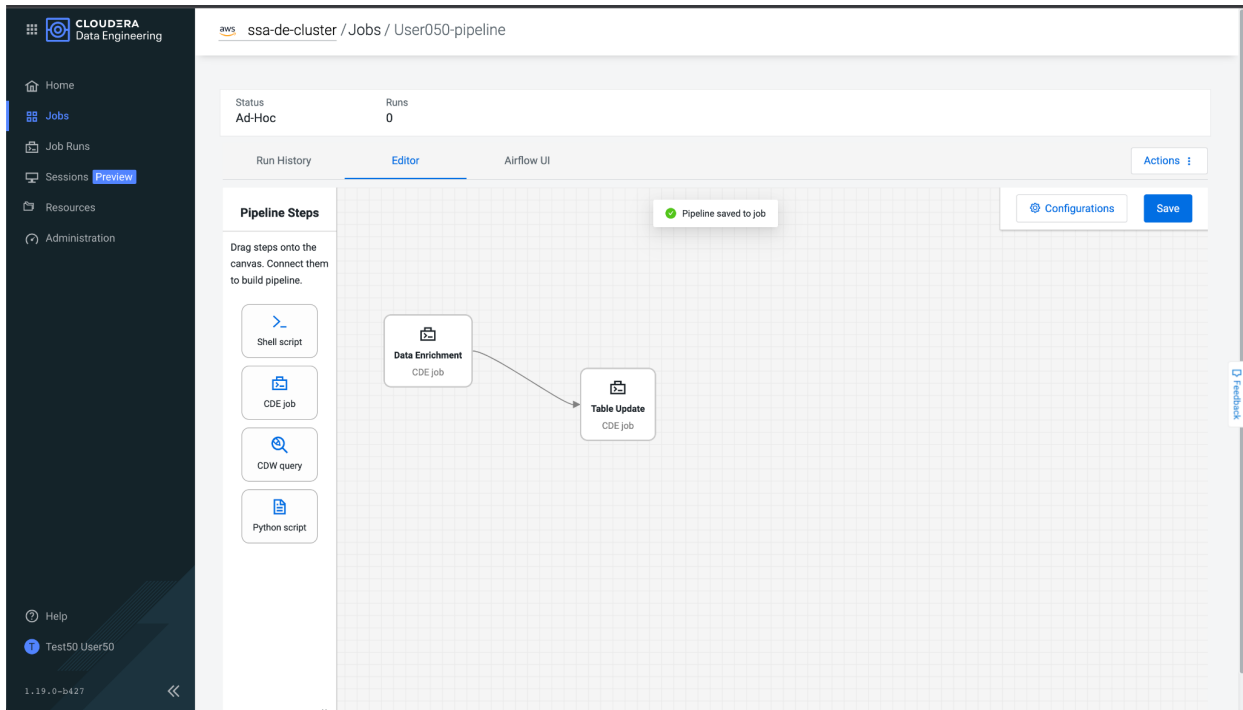


8. Para configurar la secuencia de ejecución, unir **Data Enrichment** con **Table Update**. Para eso, hacer clic en el conector derecho del job de **Data Enrichment** y arrastrar hasta el conector izquierdo de **Table Update**.

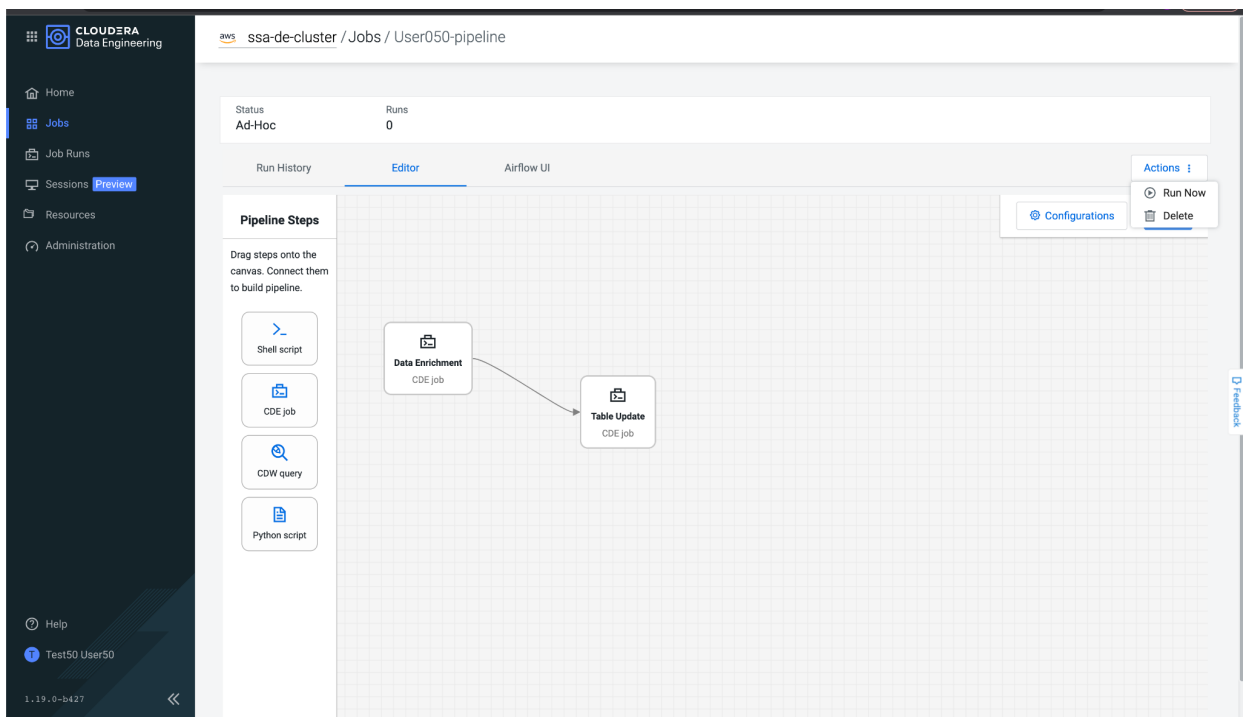




9. Una vez unidos los Jobs, hacer clic en **Save** para guardar la configuración realizada. Deberás ver un mensaje indicando **Pipeline saved to job.**



10. Ha llegado el momento de ejecutar el pipeline. En el lado superior derecho del canvas, hacer clic en **Actions -> Run Now**.



11. Deberás ver la pantalla de ejecución del pipeline, indicando que la ejecución ha sido inicializada.



The screenshot shows the Cloudera Data Engineering interface. On the left is a sidebar with navigation links: Home, Jobs, Job Runs, Sessions (Preview), Resources, and Administration. The main content area is titled 'aws ssa-de-cluster / Jobs / User050-pipeline'. At the top right, a green notification box states 'A new run with Id 7 has been initiated.' Below this, the 'Run History' tab is selected, showing a table with columns: Status, Run ID, Duration, User, Start Time, and Actions. The table contains one entry with Run ID 7, User user050, and Start Time May 26, 2023, 1:32:09 PM. The 'Airflow UI' tab is also visible.

12. Hacer clic en la pestaña Airflow UI para ver el detalle de ejecución de cada paso del pipeline. En la parte inferior izquierda se listan los jobs Data Enrichment y Table Update configuradas. Los colores indicando en qué estado está cada job. Asegúrese de que el radio button **Auto-refresh** esté habilitado para mostrar automáticamente el estado de los jobs.

The screenshot shows the Cloudera Data Engineering interface with the 'Airflow UI' tab selected. The main content area is titled 'aws ssa-de-cluster / Jobs / User050-pipeline'. Below the 'Run History' and 'Editor' tabs, the 'Airflow UI' tab is active, displaying the DAG diagram for 'User050\_pipeline'. The DAG diagram shows two tasks: 'Data\_Enrichment' and 'Table\_Update'. The 'Auto-refresh' button is highlighted. Below the DAG diagram, there is a table of job states with columns: Status, Run ID, Duration, User, Start Time, and Actions. The table contains one entry with Run ID 7, User user050, and Start Time May 26, 2023, 1:32:09 PM. The 'Auto-refresh' button is highlighted.


13. Podrás ver más información de la ejecución haciendo clic en la vista **Graph**. Al pasar el mouse sobre el nombre del Job, se despliega información específica de cada paso del pipeline. Asegúrese que el estado del pipeline sea exitoso (Success), lo que indica que todo el pipeline pudo ejecutarse sin problemas.

The screenshot displays the Cloudera Data Engineering interface for the 'User050\_pipeline'. The left sidebar shows navigation options: Home, Jobs, Job Runs, Sessions (Preview), Resources, and Administration. The main panel shows the pipeline details for 'User050\_pipeline' with a status of 'success' (highlighted in red). The 'Graph' view is selected, showing a DAG with two tasks: 'Data\_Enrichment' and 'Table\_Update'. A tooltip for the 'Data\_Enrichment' task provides the following information:

- Status: success
- Task Id: Data\_Enrichment
- Run: 2023-05-26, 18:36:24 UTC
- Run Id: cde-job-run-7
- Operator: CdeRunJobOperator
- Duration: 1Min 11.676Sec
- UTC: Started: 2023-05-26, 18:33:29; Ended: 2023-05-26, 18:34:40

The DAG diagram shows the flow from 'Data\_Enrichment' to 'Table\_Update'. The interface also includes a 'Run History' section with a table of runs, a 'Calendar' view, and a 'Task Duration' chart. The 'Update' button is visible next to the 'Data\_Enrichment' task.

Al lado del nombre del pipeline aparece el estado de ejecución (marcado en rojo). Si está en verde e indica **Success**, significa que la ejecución fue correcta.

 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

Refresh