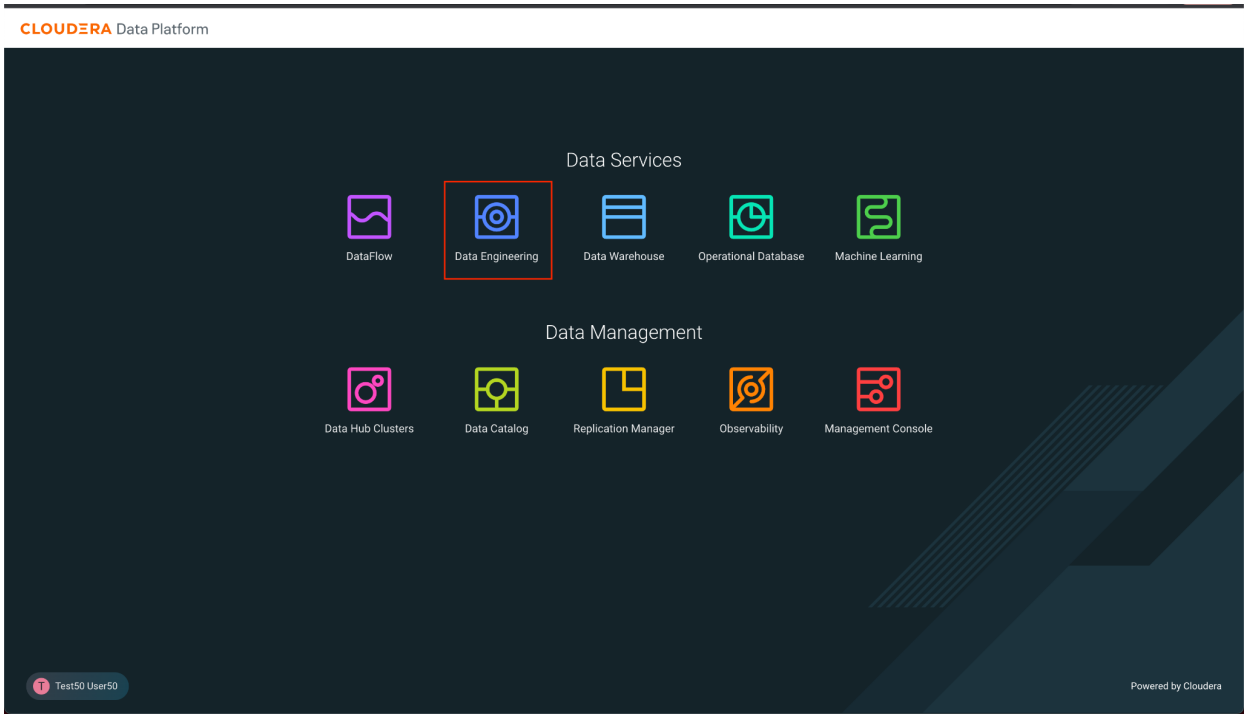
Data Lifecycle en CDP Public Cloud

Laboratorio Data Engineering

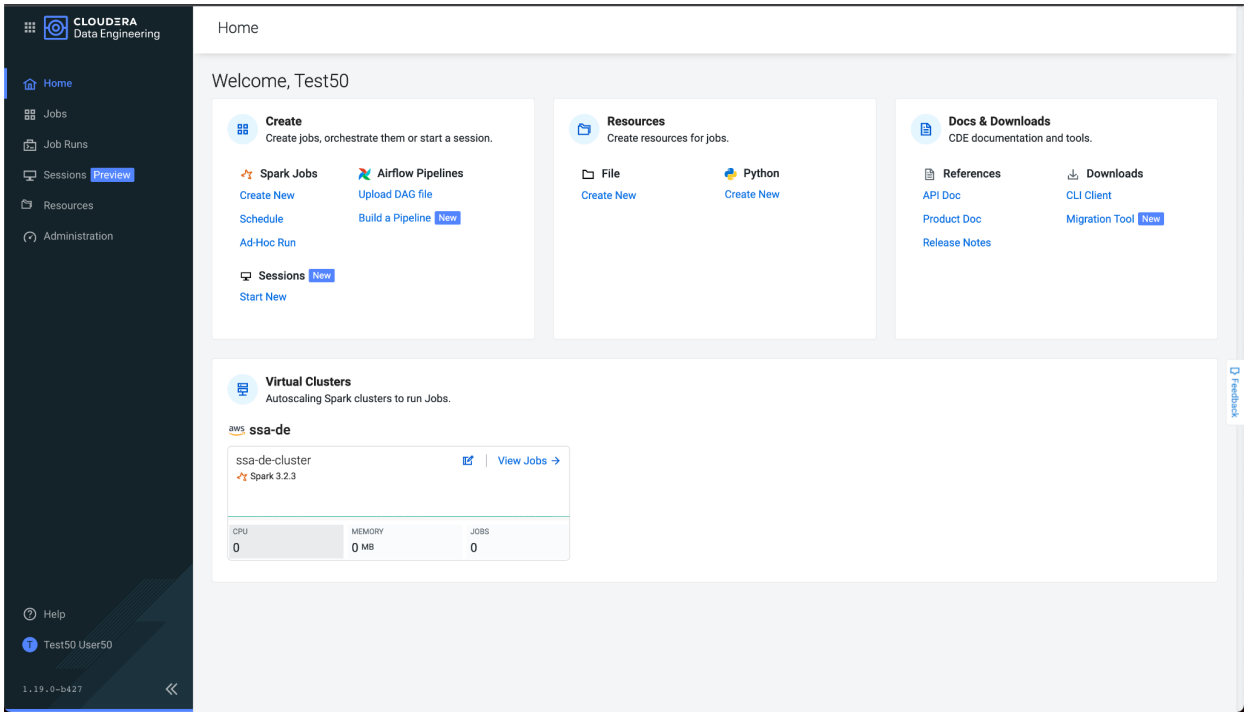
Objetivos:

- 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:



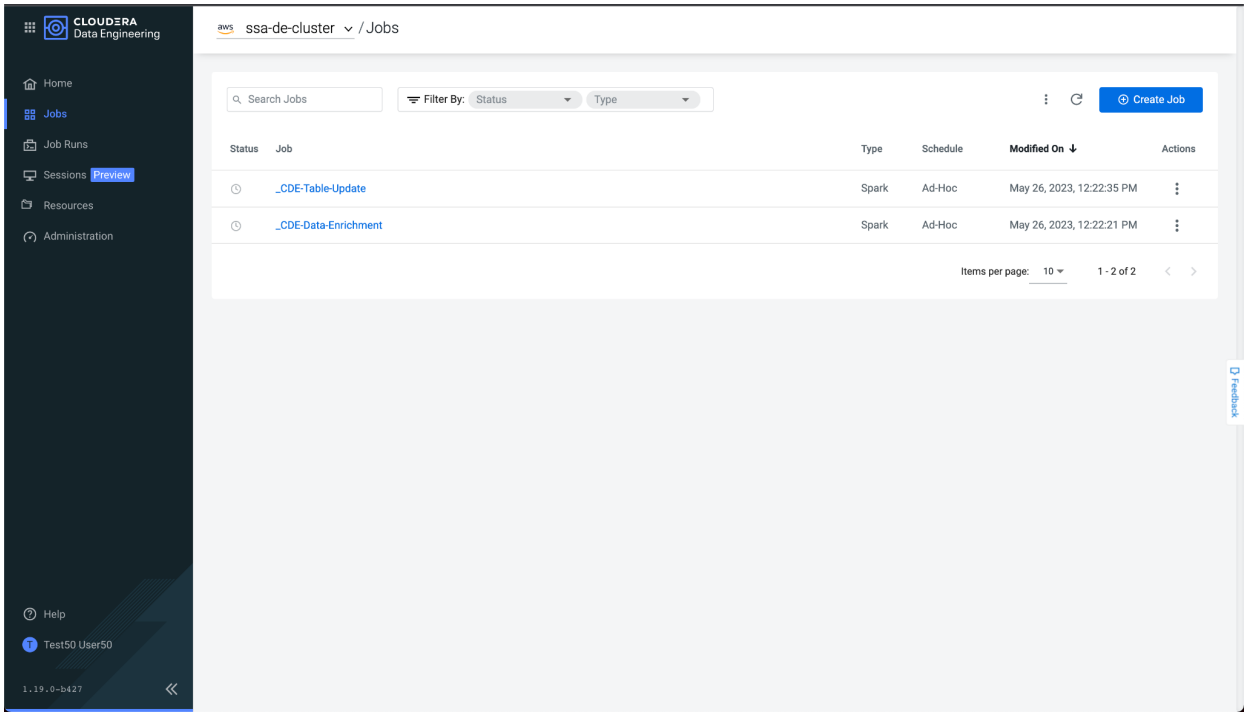
2. El Home de Data Engineering muestra todas las acciones que se pueden hacer, como Jobs en Spark y pipelines en Airflow, Resources y información/documentación útil. Hacer clic en la opción **Jobs** del menú izquierdo para crear un flujo de datos en Airflow.



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.

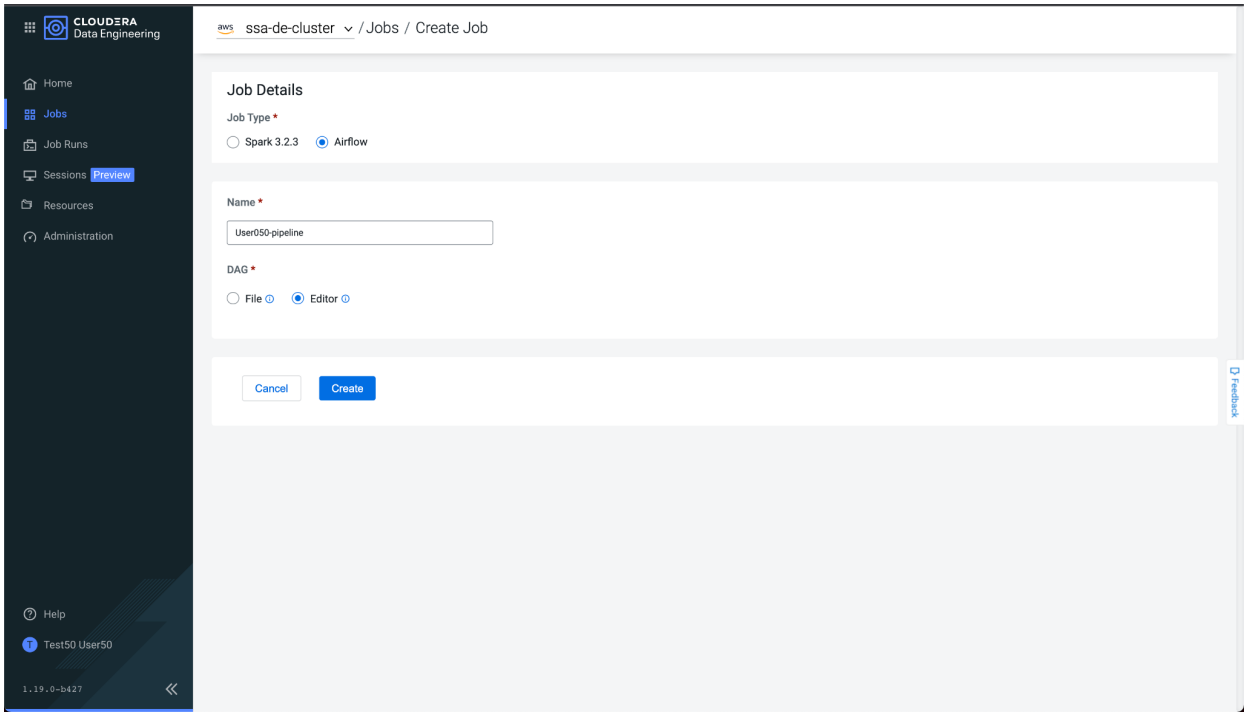
Es hora de crear nuestro Job en Airflo. Hacer clic en **Create Job.**



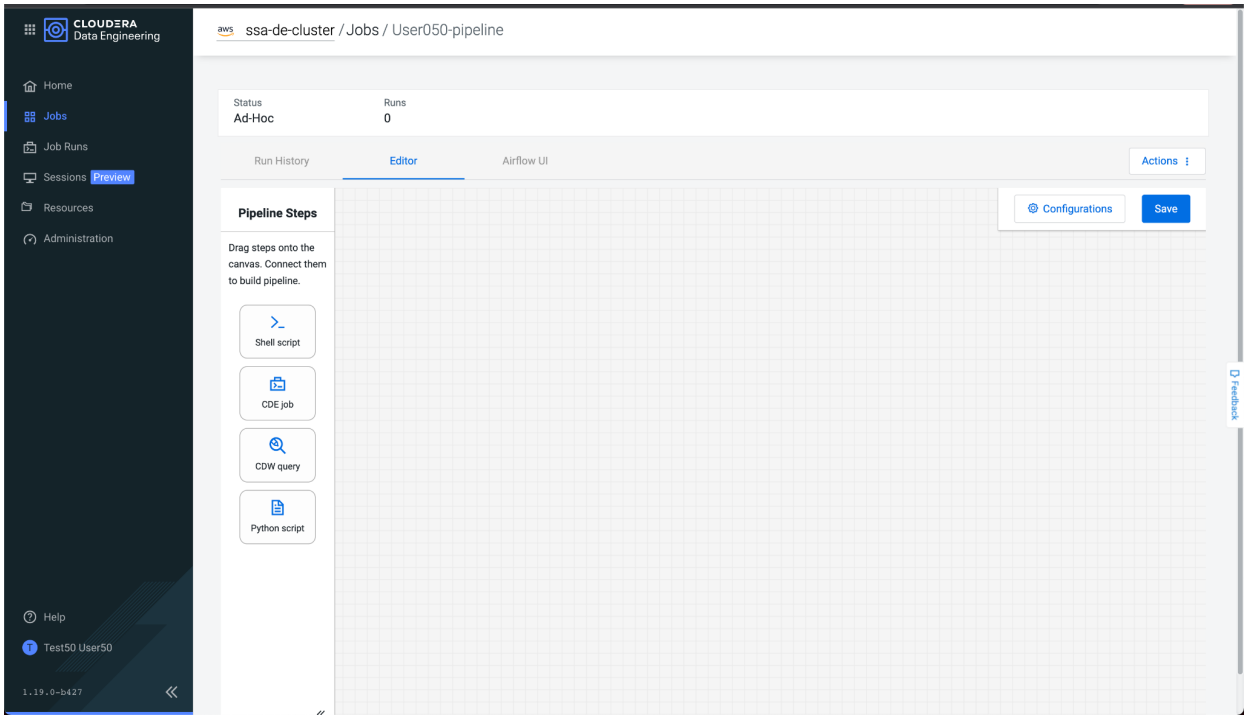
4. En el formulario de creación de Job, debe ingresar la siguiente información:

- 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**.

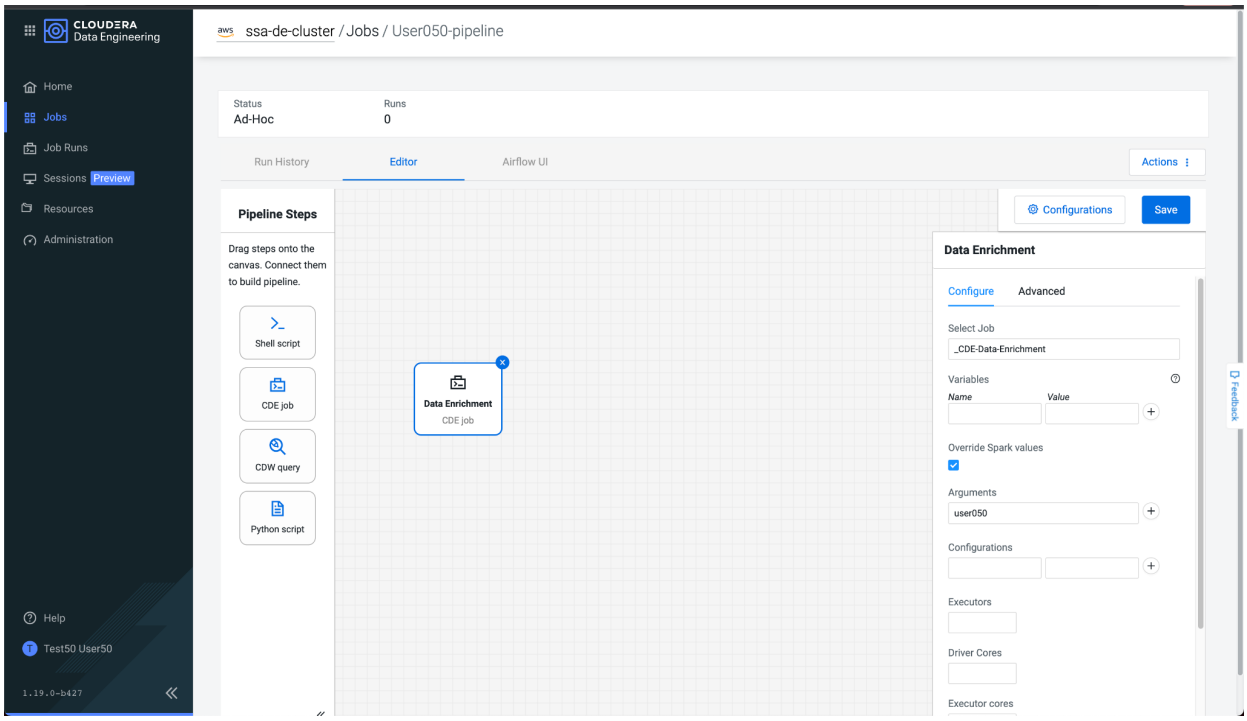


5. En la pantalla de edición del Job, seleccionar la pestaña Editor, y verás el siguiente canvas para arrastrar los pasos del pipeline que vamos a crear. En nuestro caso, vamos a crear dos tareas de CDE (CDE Jobs) y relacionarlas.



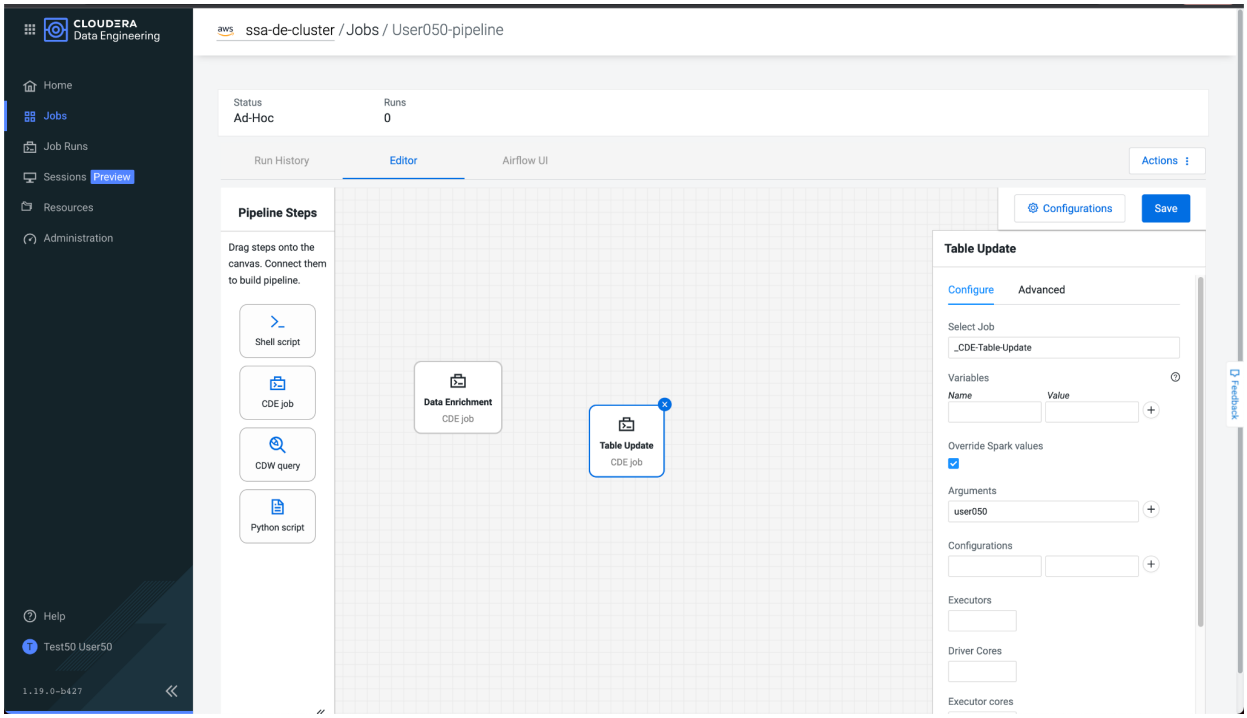
6. Empecemos con el primer Job. Hacer clic sobre el botón CDE Job y arrastrar al canvas, ingresando la siguiente configuración:

- **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

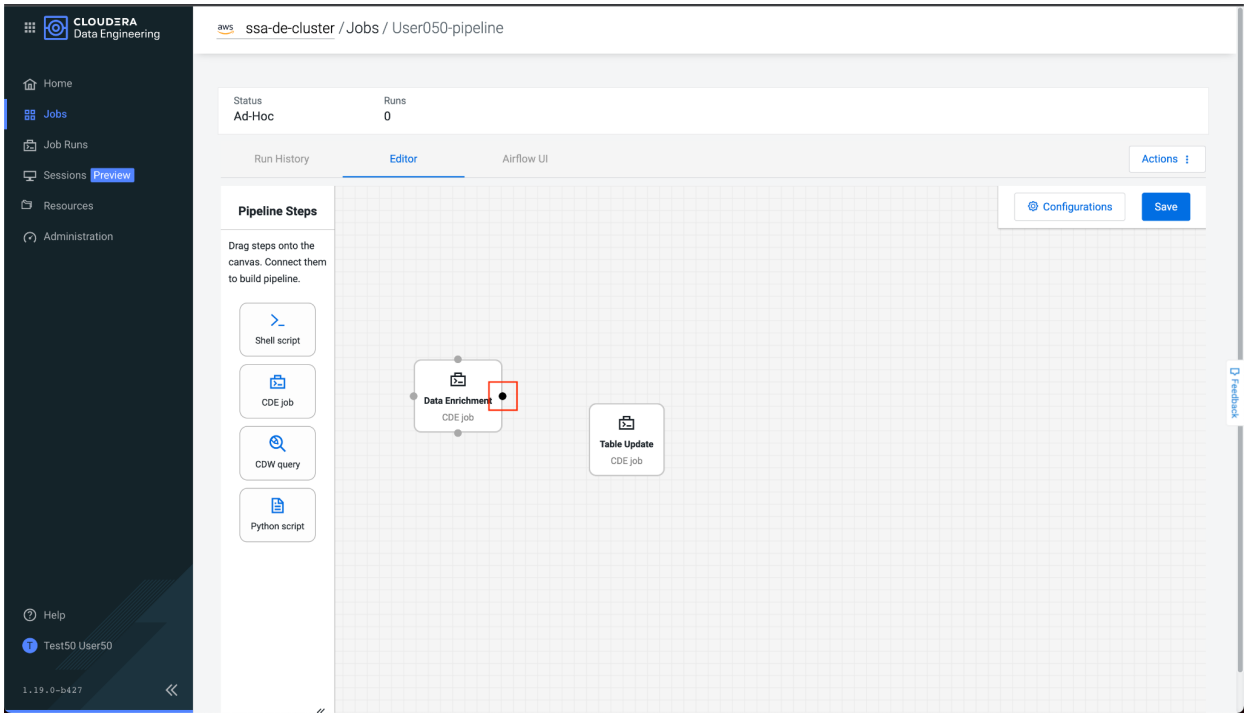


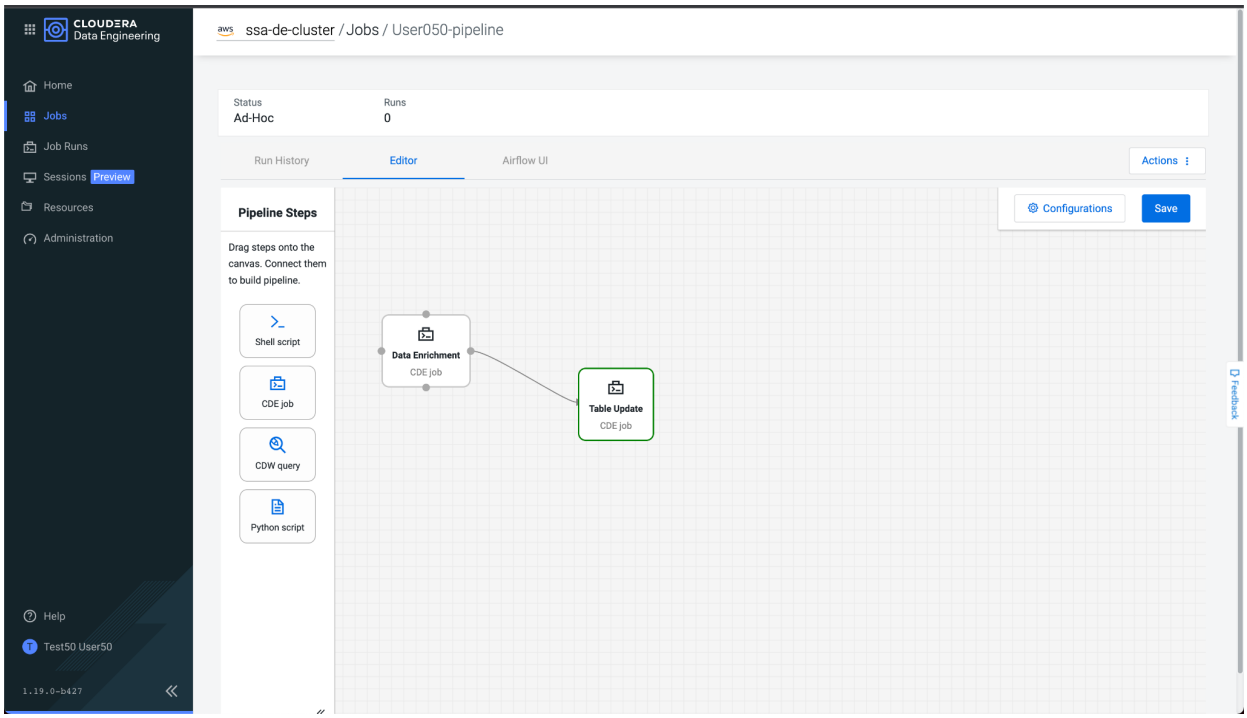
7. Configure el segundo Job. Hacer clic sobre el botón CDE Job y arrastrar al canvas, ingresando la siguiente configuración:

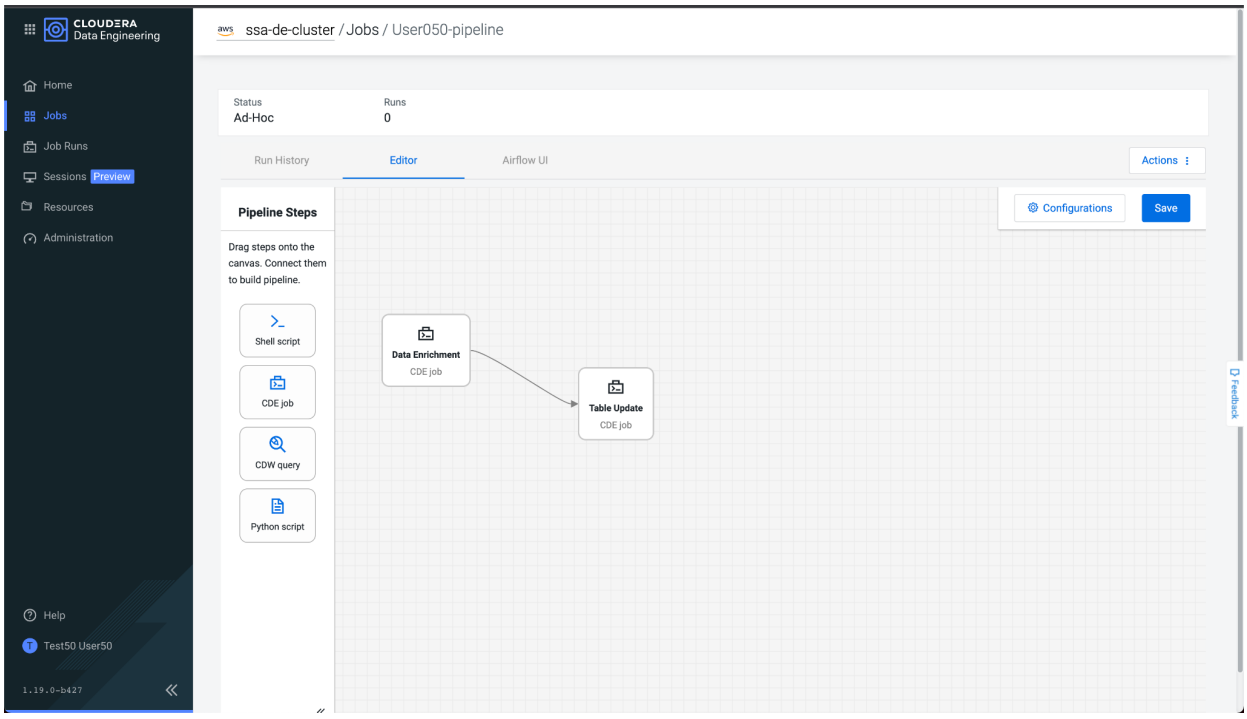
- **Título/nombre**: Table Update
- **Select Job**: seleccionar el Job *_CDE-Table-Update*
- Marcar el checkbox **Override Spark values**. Aparecerán opciones adicionales luego abajo.
- **Arguments**: <usuario asignado>. User el nombre del usuario que tienes asignado. Por ejemplo, user050



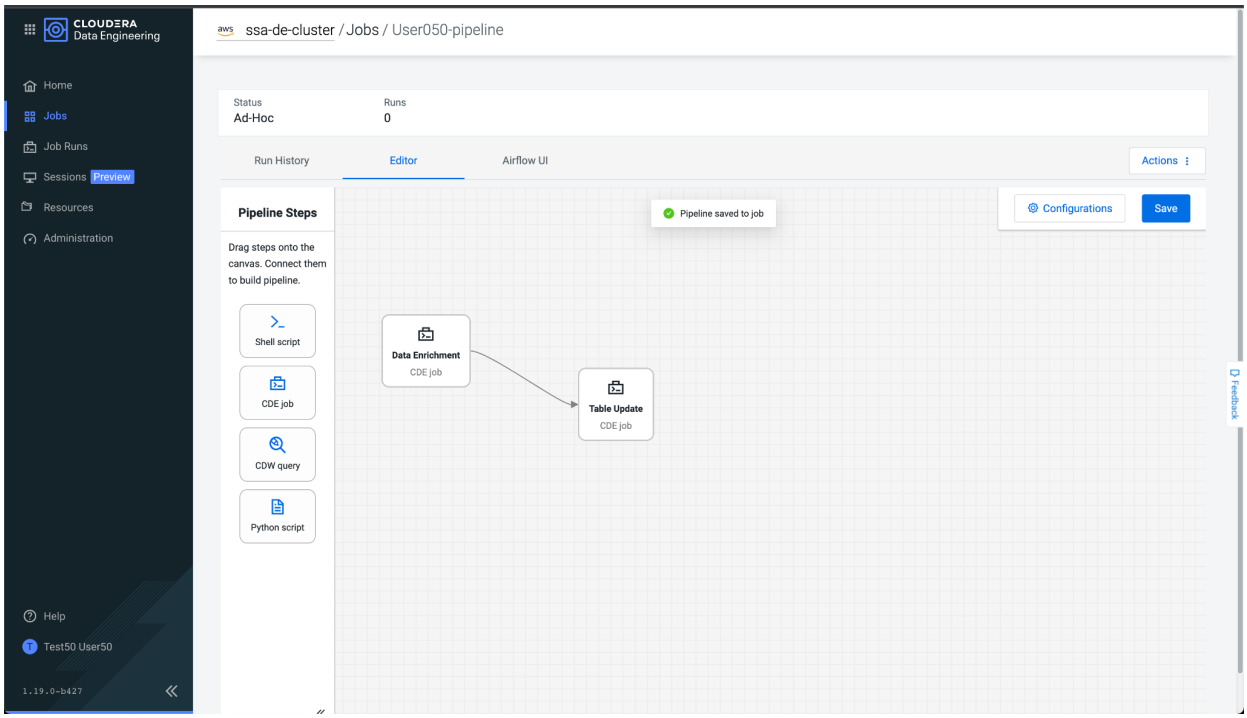
8. Para configurar la secuencia de ejecución, unir **Data Enrichment** con **Table Update**. Para eso, hacer clic en el conector derecho del job de **Data Enrichment** y arrastrar hasta el conector izquierdo de **Table Update**.



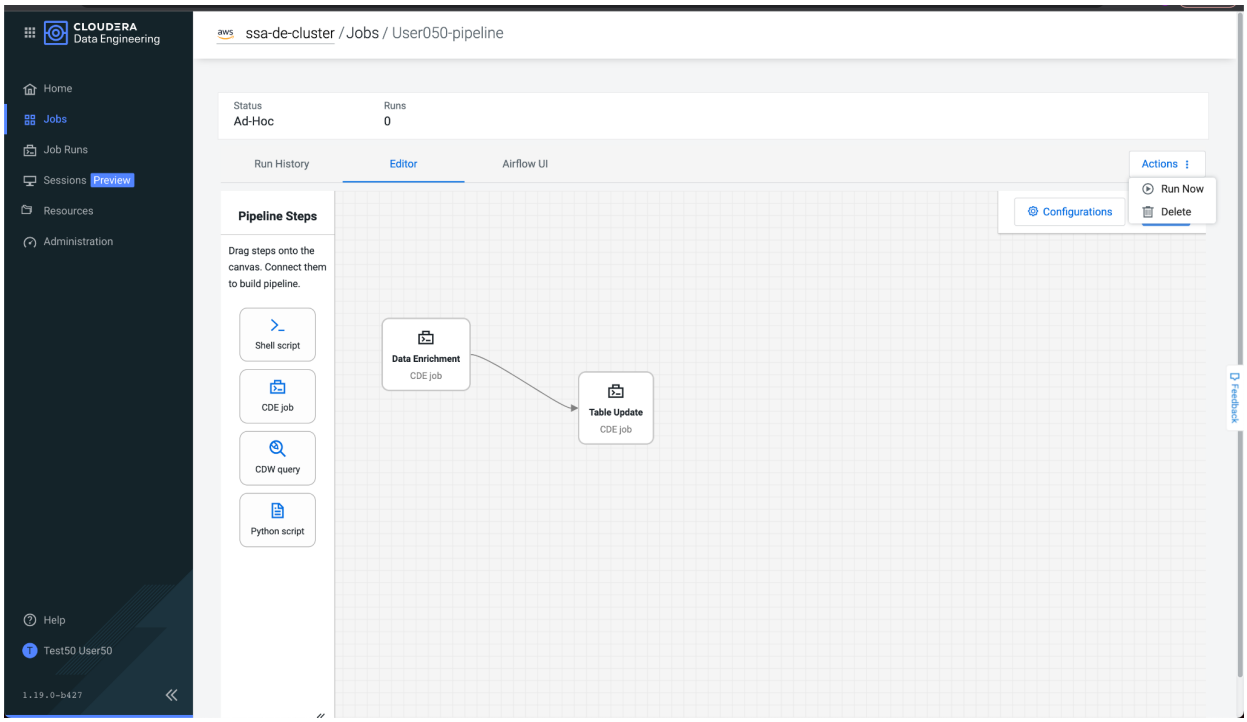




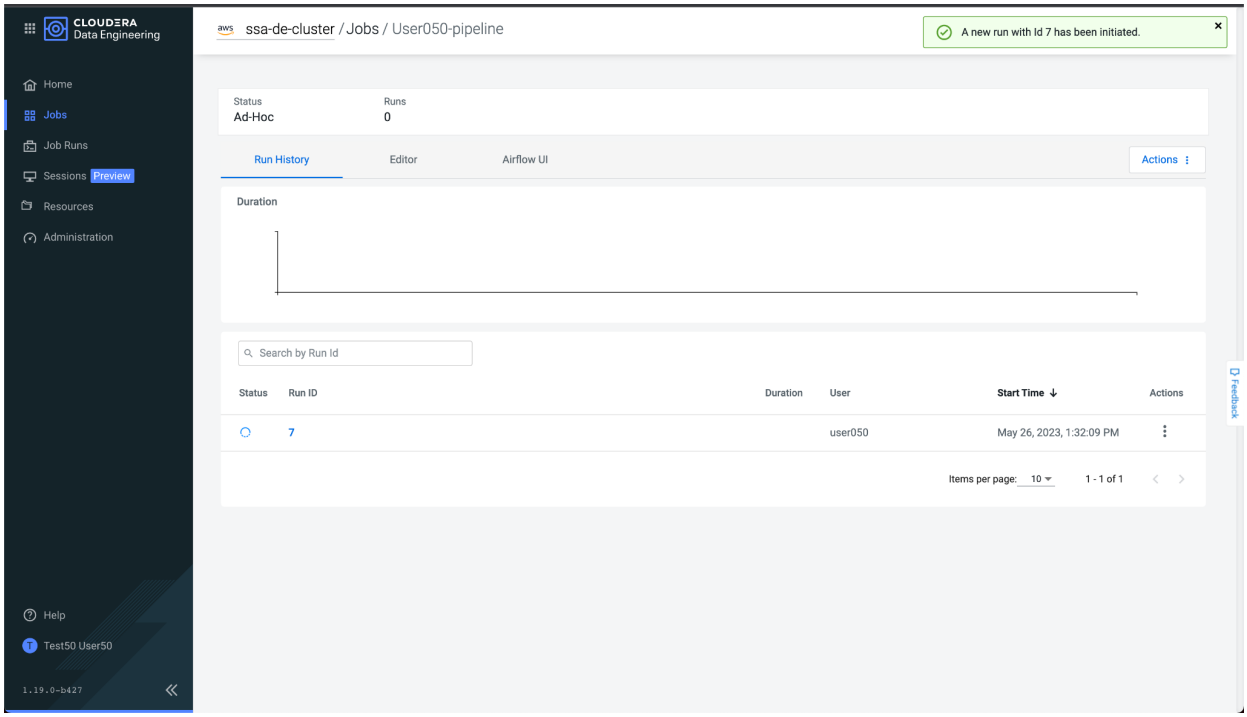
9. Una vez unidos los Jobs, hacer clic en **Save** para guardar la configuración realizada. Deberás ver un mensaje indicando **Pipeline saved to job**.



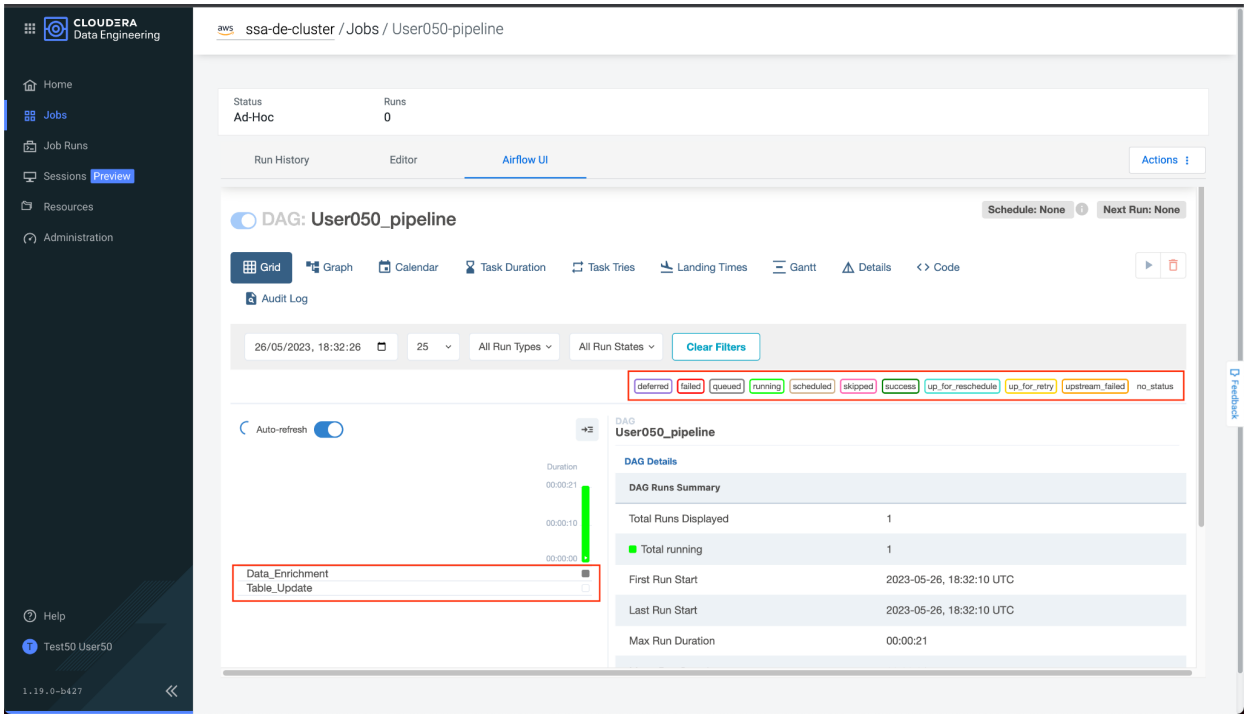
10. Ha llegado el momento de ejecutar el pipeline. En el lado superior derecho del canvas, hacer clic en **Actions -> Run Now**.



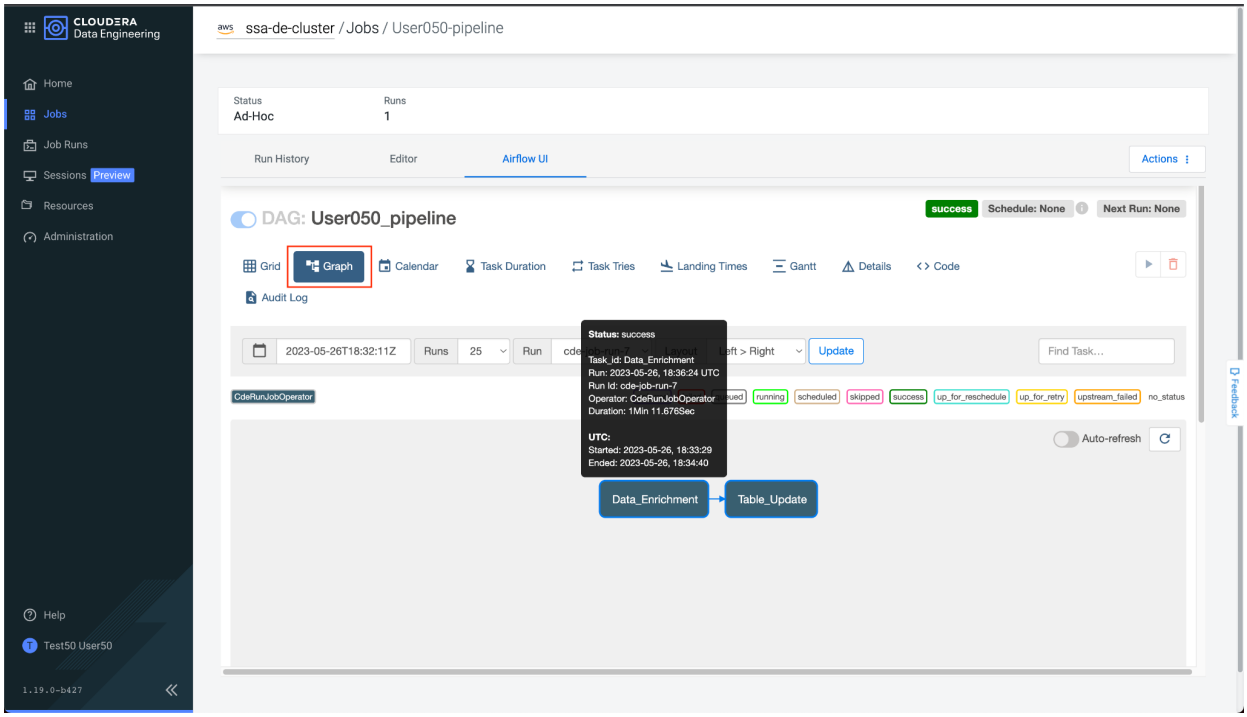
11. Deberás ver la pantalla de ejecución del pipeline, indicando que la ejecución ha sido inicializada.



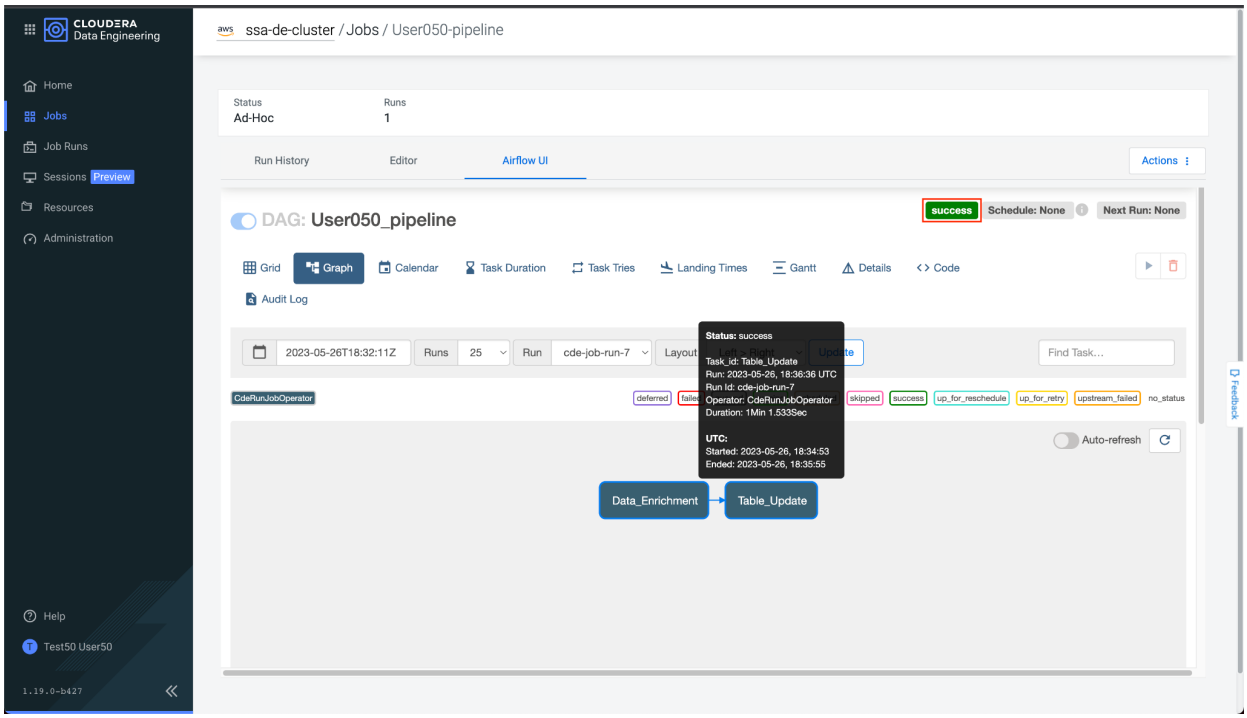
12. Hacer clic en la pestaña Airflow UI para ver el detalle de ejecución de cada paso del pipeline. En la parte inferior izquierda se listan los jobs Data Enrichment y Table Update configuradas. Los colores indicando en qué estado está cada job. Asegúrese de que el radio button **Auto-refresh** esté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



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.



*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
Update
Find Task...
CdeRunJobOperator
deferred
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
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
Auto-refresh
Data_Enrichment
Table_Update
Feedback