

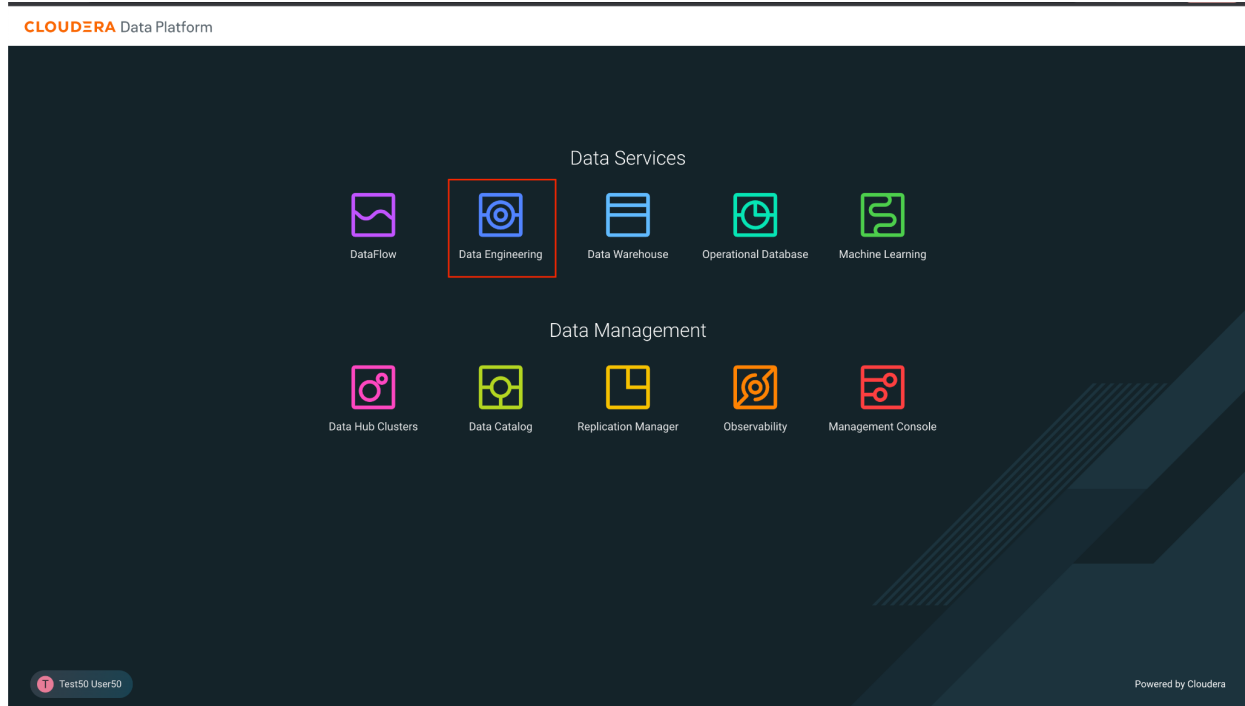
# Ciclo de vida dos dados em CDP Public Cloud

## Laboratório de Engenharia de Dados

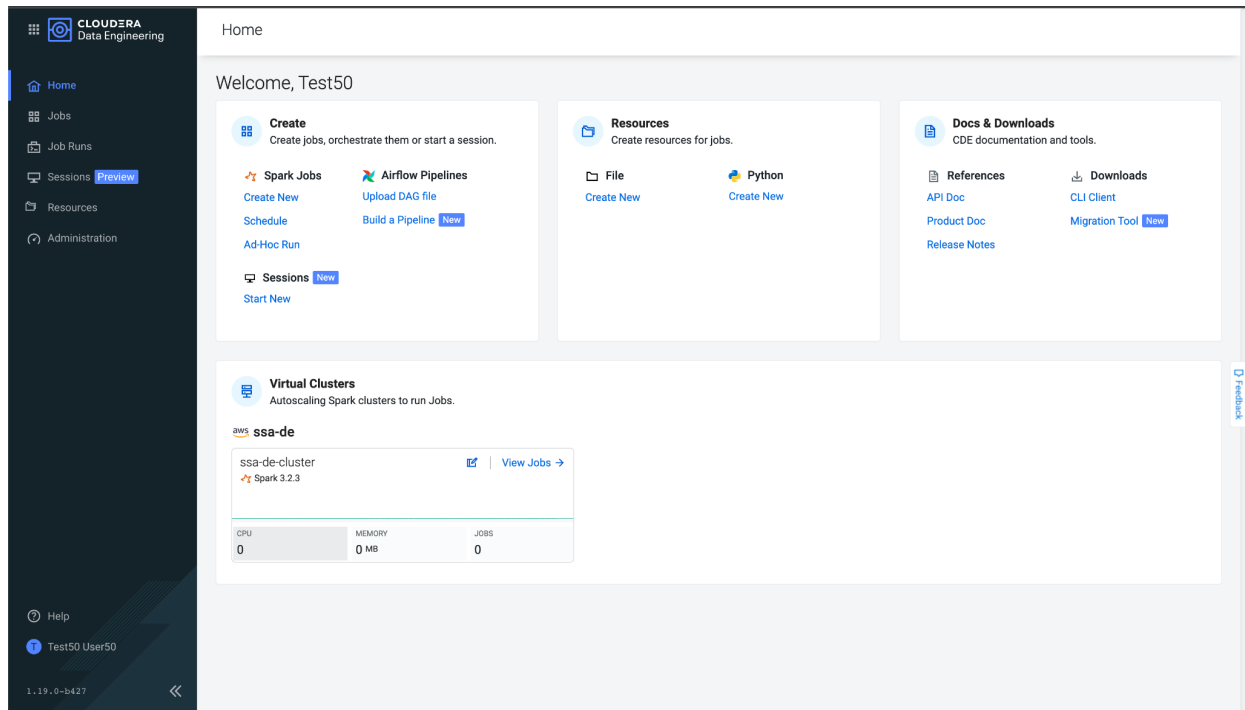
### Metas:

- Execute um processo de enriquecimento de dados
- Execute um processo para simular alterações nos dados
- Configurar a execução de um pipeline usando ferramentas de baixo código/sem código

1. Clique em DataFlow no CDP PC Home:



2. O Data Engineering Home mostra todas as ações que podem ser feitas, como Jobs no Spark e pipelines no Airflow, Recursos e informações/documentações úteis. Clique na opção **Jobs** no menu esquerdo para criar um fluxo de dados no Airflow.

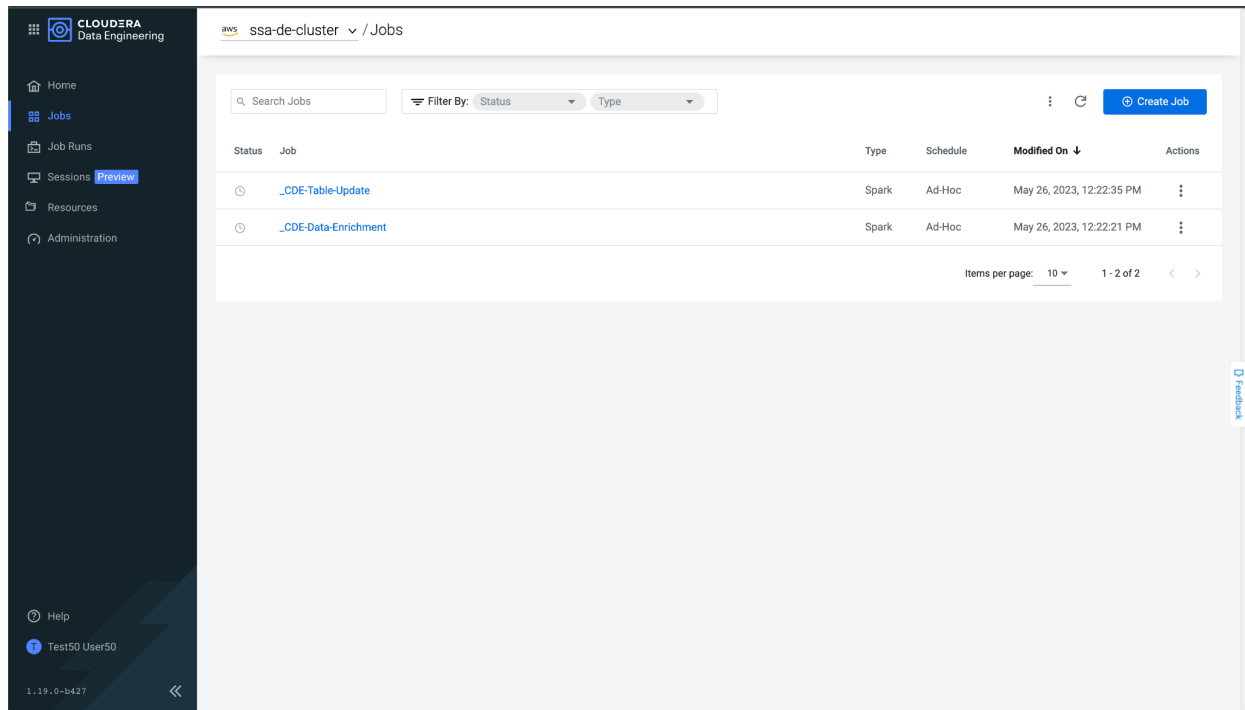


3. Aqui estão listadas as tarefas disponíveis. Para os propósitos deste workshop, dois Jobs foram configurados:

**\_CDE-Table-Update**, gera alterações aleatórias na tabela enriquecida para visualizar a funcionalidade Lakehouse Time Travel.

**\_CDE-Data-Enriquecimento**, processo no Spark (Python) para enriquecer os dados ingeridos do Kafka e salvar em uma nova tabela.

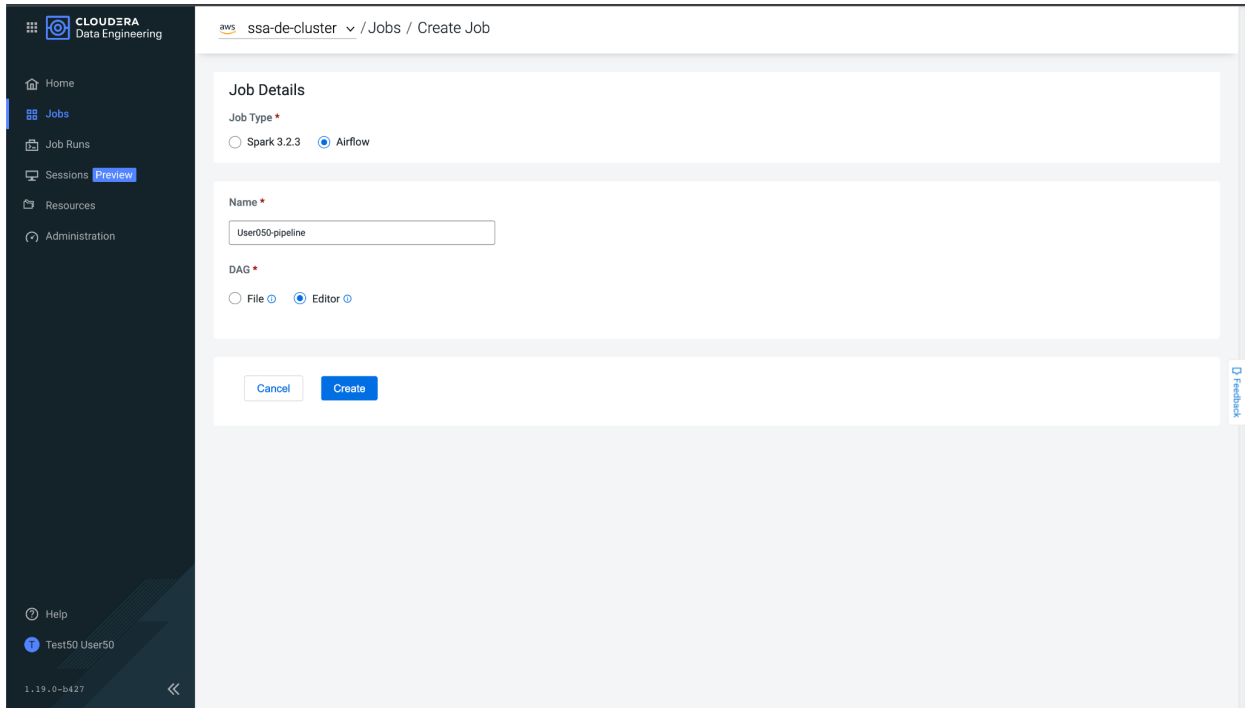
É hora de criar nosso Job no Airflow. Clique em **Create Job**.



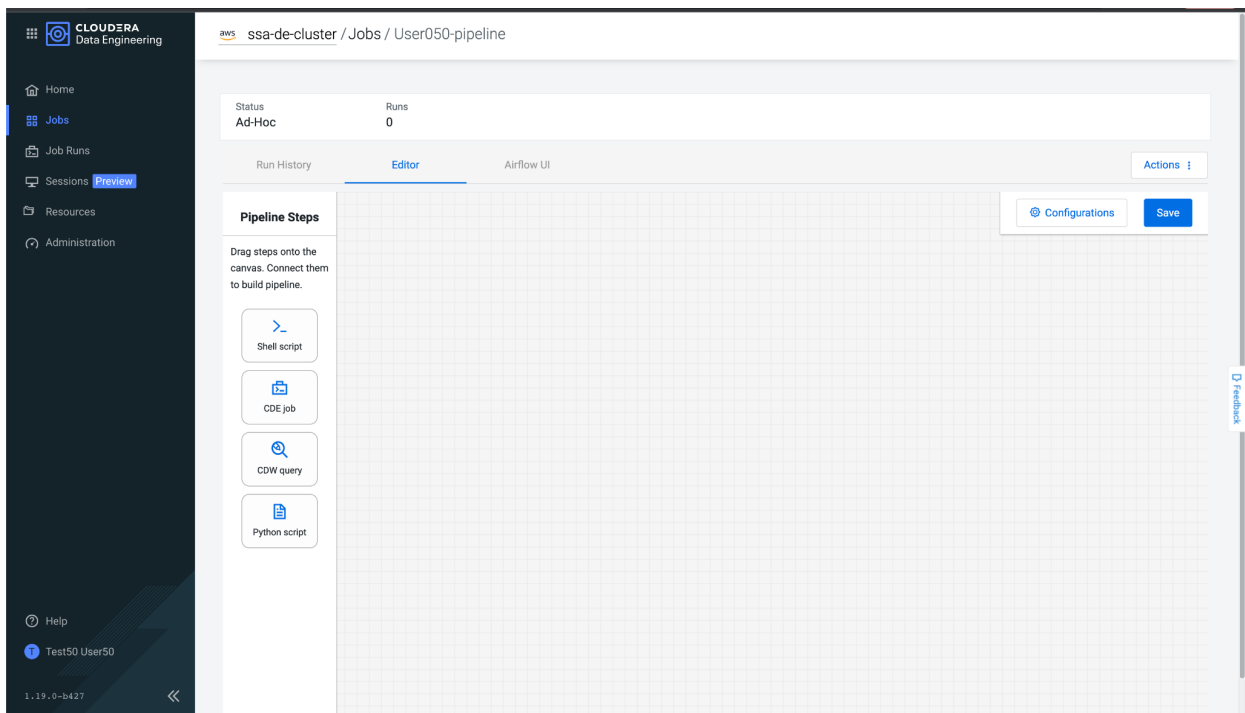
4. No formulário de criação de Job insira as seguintes informações:

- Job Type: Airflow
- Nome: Utilize o padrão de nomenclatura <usuário atribuído>-pipeline substituindo <usuário atribuído> pelo usuário atribuído a você. Por exemplo, user050-pipeline.
- DAG: selecione **Editor**, para configurar graficamente o Job.

Depois de inserir os valores corretamente, clique em **Create**.

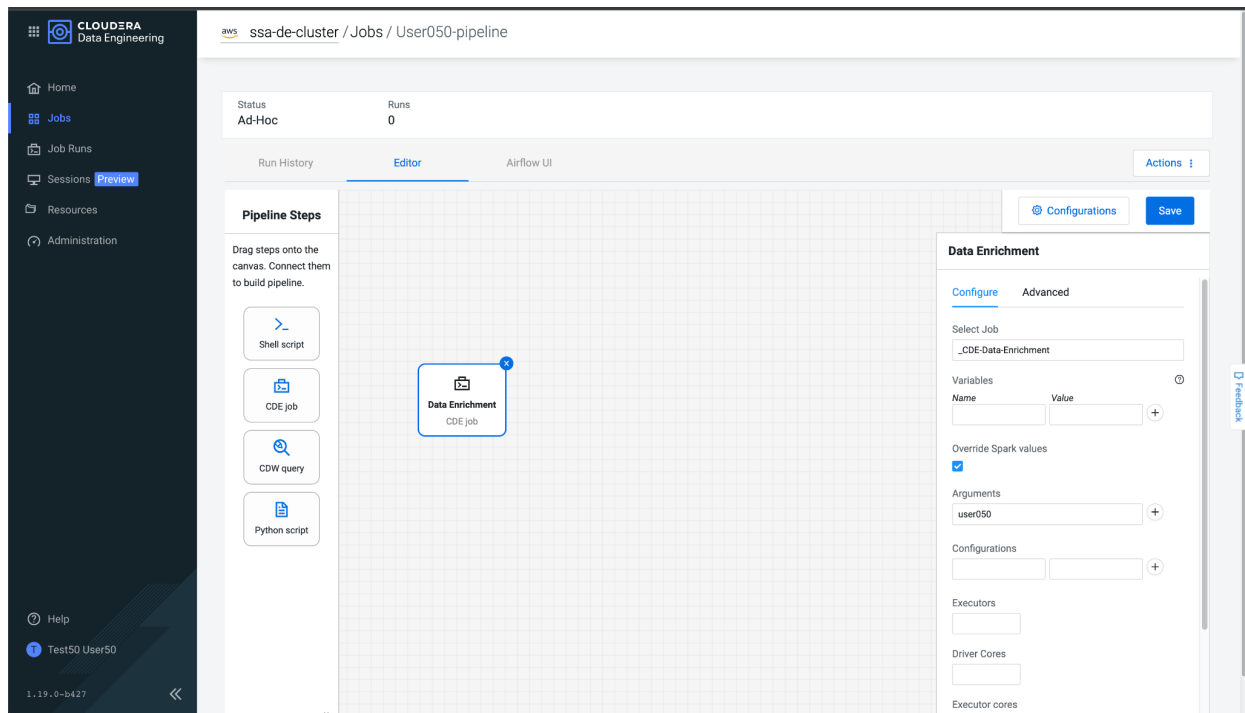


5. Na tela de edição do Job, selecione a aba Editor. Você verá a seguinte tela de canvas para arrastar os passos do pipeline que vamos criar. No nosso caso vamos criar dois Jobs CDE e relacioná-los.



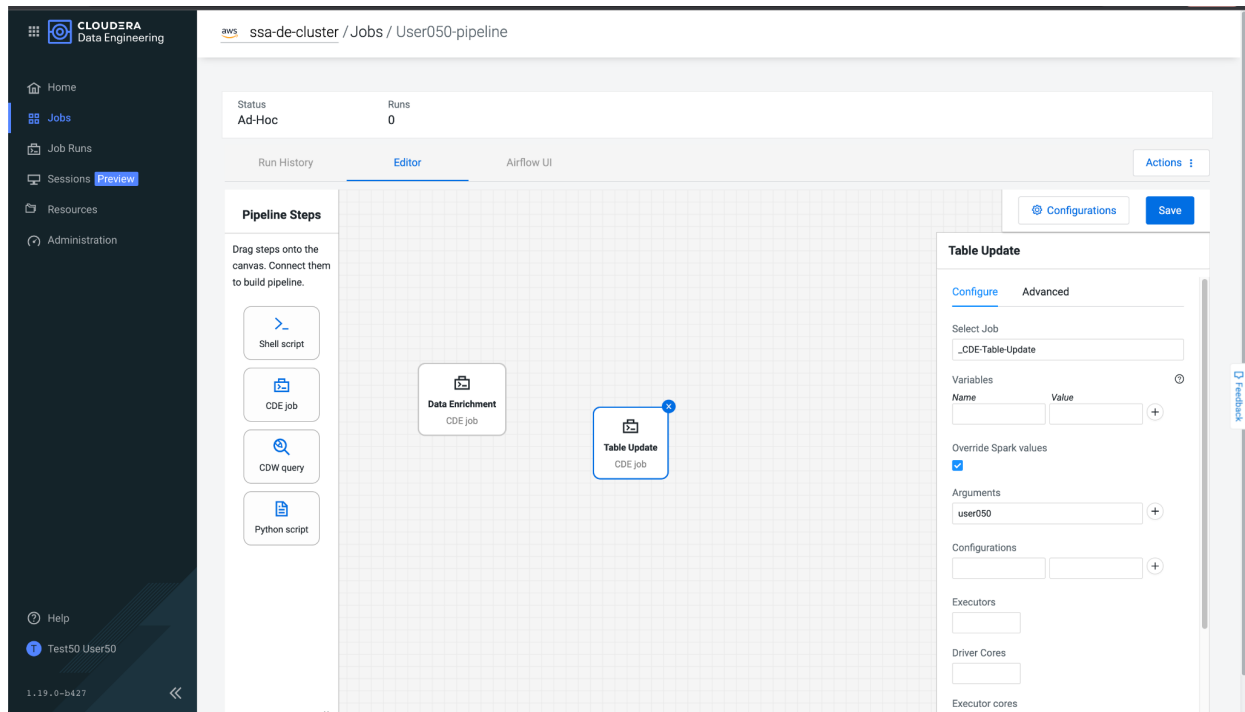
6. Vamos começar com o primeiro Job. Clique no botão CDE Job e arraste para a tela, inserindo as seguintes configurações:

- **Title:** Data Enrichment
- **Select Job:** Selecione o Job *\_CDE-Data-Enrichment*
- Marque a opção **Override Spark values**. Com isso, opções adicionais aparecerão abaixo.
- **Arguments:** Informe o usuário atribuído a você. Por exemplo, user050.

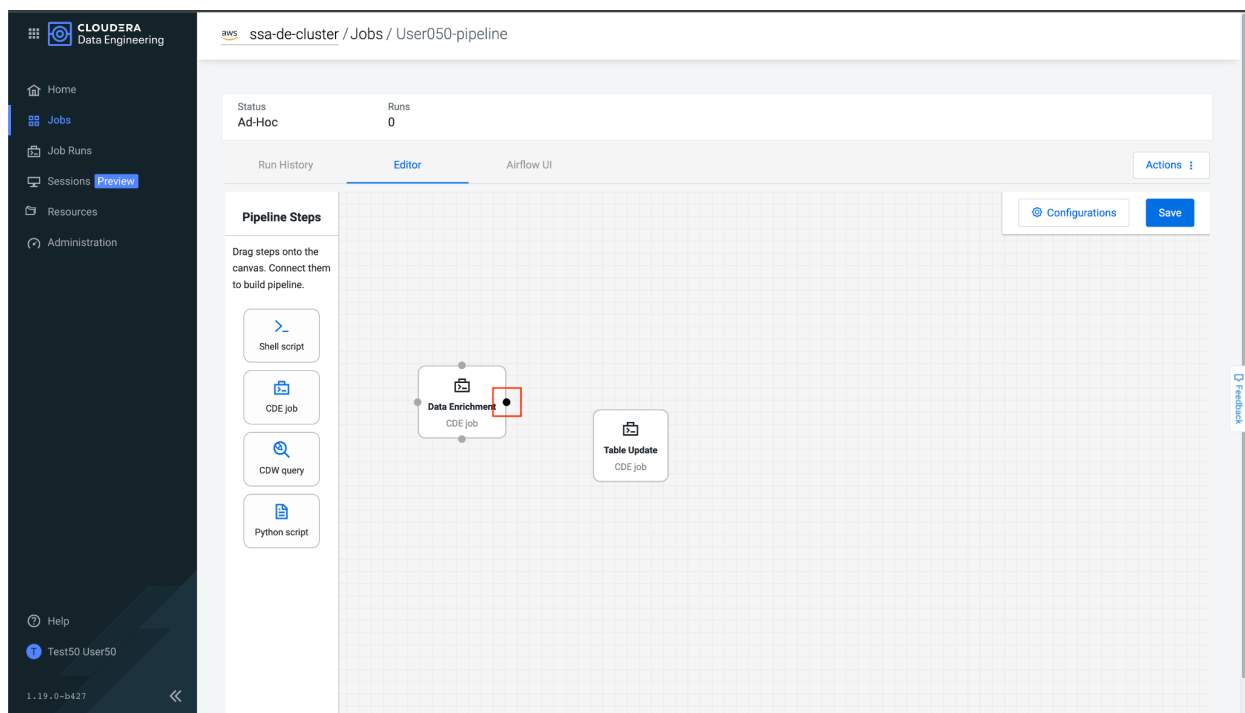


7. Configure o segundo Job clicando novamente no botão CDE Job e arrastando para a tela. Insira as seguintes configurações:

- **Title:** Table Update
- **Select Job:** Selecione o Job *\_CDE-Table-Update*
- Marque a opção **Override Spark values**. Com isso, opções adicionais aparecerão abaixo.
- **Arguments:** Informe o usuário atribuído a você. Por exemplo, user050.



8. Para configurar a sequência de execução, vincule **Data Enrichment** com **Table Update**. Para isso, clique no conector direito do Job de **Data Enrichment** e arraste para o conector esquerdo de **Table Update**.



CLUSTER: 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

Help | Test50 User50 | 1.19.0-b427

Feedback

CLUSTER: 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

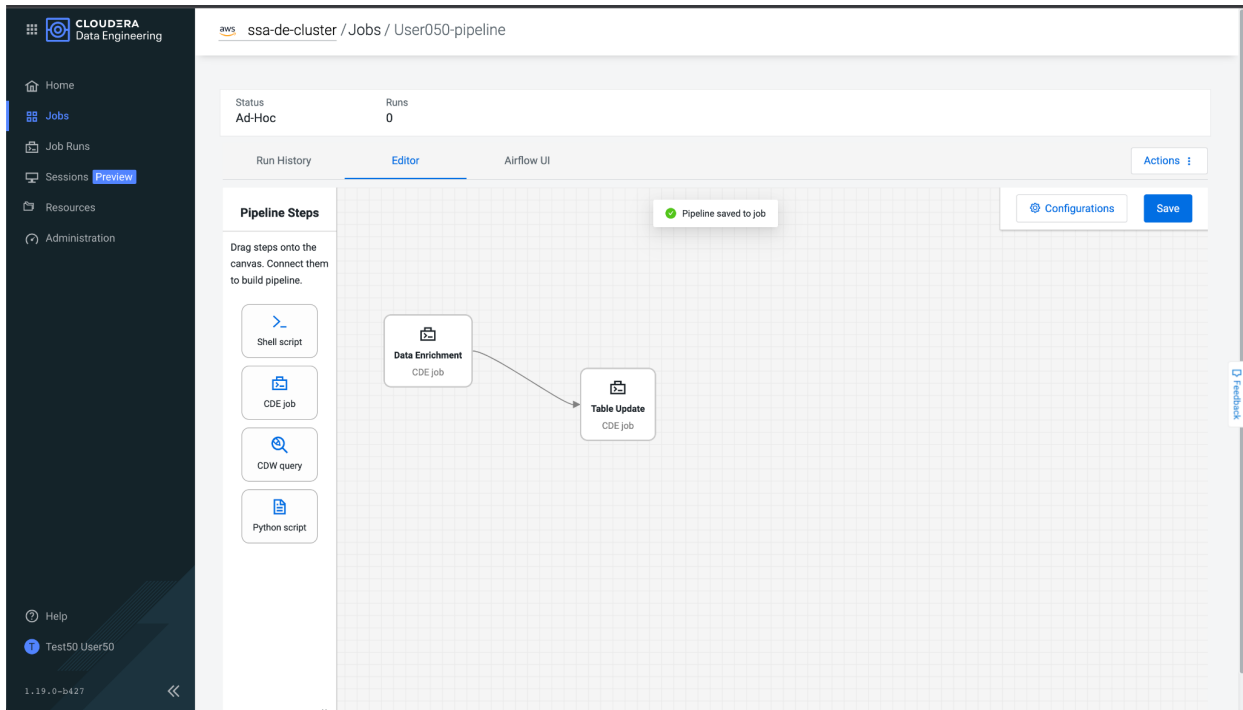
**Data Enrichment**  
CDE job

**Table Update**  
CDE job

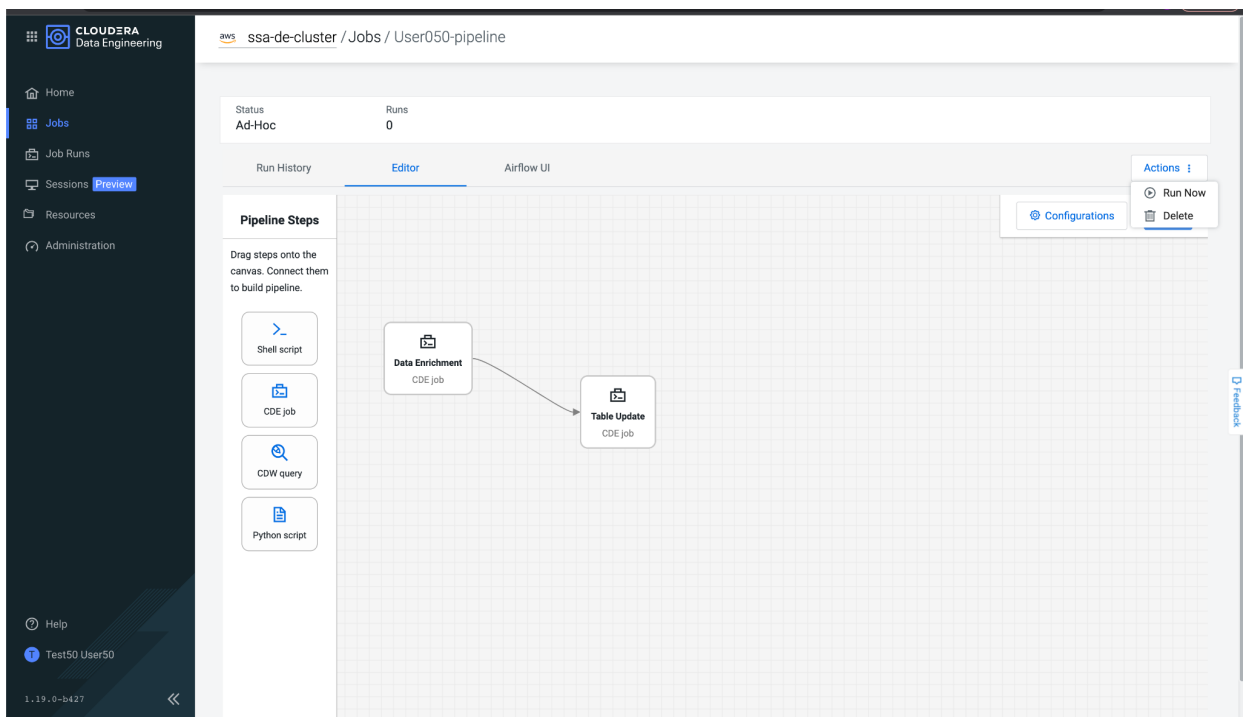
Help | Test50 User50 | 1.19.0-b427

Feedback

9. Assim que os Jobs estiverem conectados, clique em **Save** para salvar as configurações feitas. Você deve ver uma mensagem indicando **Pipeline saved to job**.



10. Chegou a hora de executar o pipeline. No canto superior direito do canvas clique em **Actions -> Run Now**.



11. Você deve ver a tela de execução do pipeline indicando que a execução foi inicializada.



CloudERA Data Engineering

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

Status: Ad-Hoc Runs: 0

Run History Editor Airflow UI

Duration

Search by Run Id

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

Items per page: 10 1 - 1 of 1

Help Test50 User50 1.19.0-b427

12. Clique na guia **Airflow UI** para ver os detalhes de execução de cada etapa do pipeline. Os Jobs de Data\_Enrichment e Table\_Update configurados são listados na parte inferior esquerda. As cores que indicam o status de cada trabalho. Certifique-se de que a opção **Auto-refresh** está habilitada para exibir automaticamente o status dos Jobs.

CloudERA Data Engineering

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

Status: Ad-Hoc Runs: 0

Run History Editor Airflow UI

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

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

Help Test50 User50 1.19.0-b427

13. Você pode ver mais informações sobre a execução clicando na opção **Graph**. Ao passar o mouse sobre o nome do job são exibidas informações específicas para cada etapa do pipeline. Certifique-se de que o status do pipeline seja Success, o que indica que todo o pipeline pôde ser executado sem problemas.

The screenshot displays the Cloudera Data Engineering interface. On the left is a dark 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 shows the pipeline's status as 'Ad-Hoc' with '1' run. Below this are tabs for 'Run History', 'Editor', and 'Airflow UI' (which is active). A 'success' status badge is visible next to the pipeline name 'DAG: User050\_pipeline'. A toolbar contains various view options: Grid, Graph (highlighted with a red box), Calendar, Task Duration, Task Tries, Landing Times, Gantt, Details, and Code. Below the toolbar is a filter bar with a date range '2023-05-26T18:32:11Z', a dropdown for 'Runs' set to '25', and a 'Run' button. A tooltip is open over the 'Data\_Enrichment' task in the DAG, displaying 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, and Duration: 1Min 11.676Sec. The DAG itself shows two tasks: 'Data\_Enrichment' and 'Table\_Update', connected by an arrow. A legend at the bottom right of the DAG shows various task statuses: failed, running, scheduled, skipped, success, up\_for\_reschedule, up\_for\_retry, upstream\_failed, and no\_status. An 'Auto-refresh' toggle is also present.

*O status de execução aparece ao lado do nome do pipeline (marcado em vermelho). Se estiver verde e indicar **Success**, significa que a execução foi bem-sucedida.*

 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