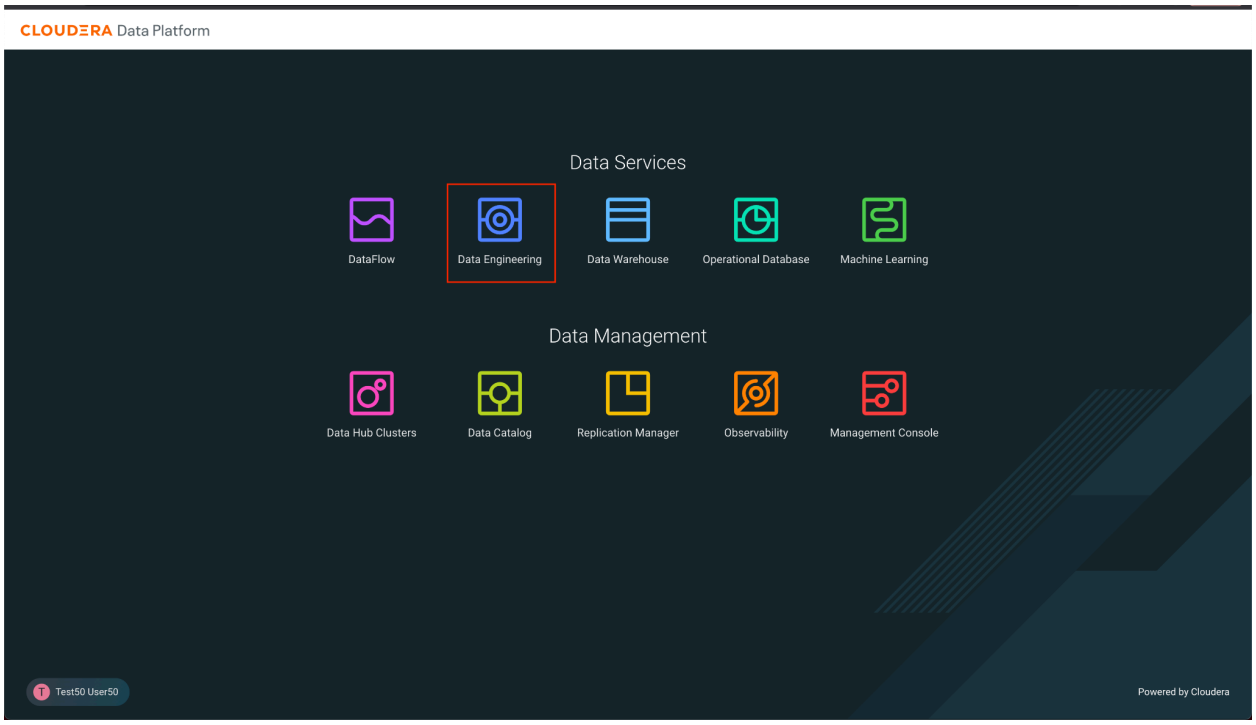
Data Lifecycle en CDP Public Cloud

Data Engineering Lab

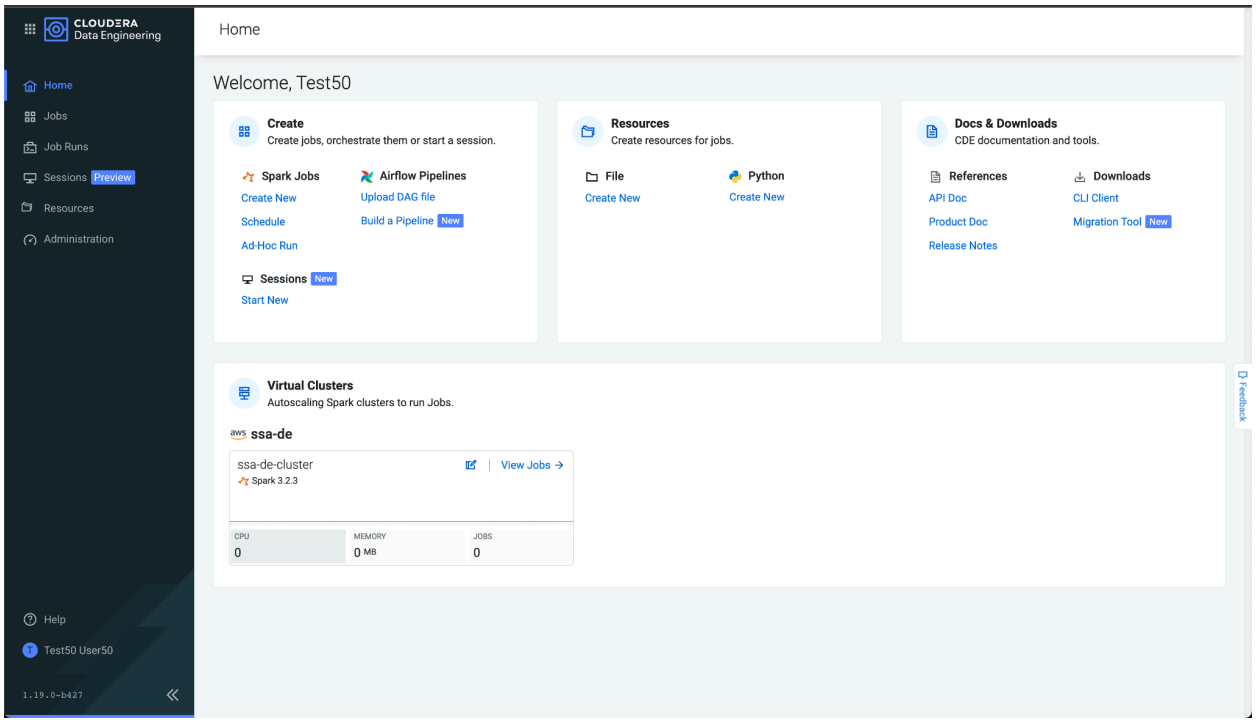
Goals:

- 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:



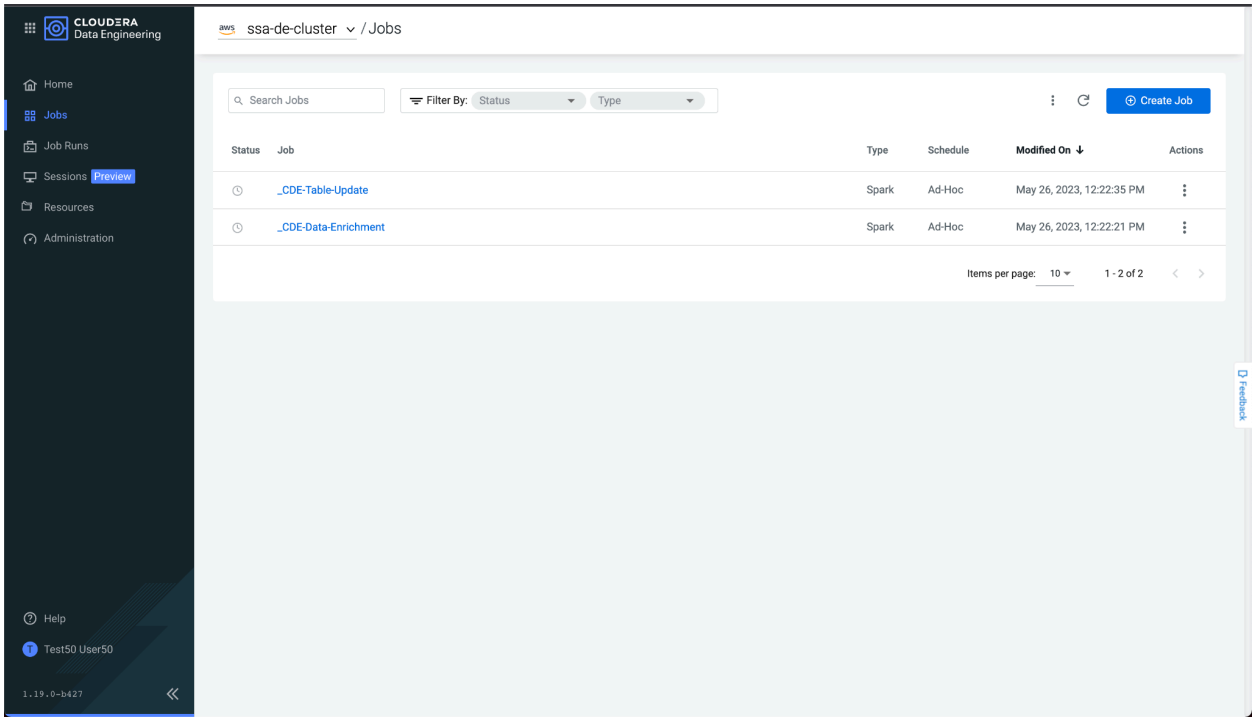
2. The Data Engineering Home shows all the actions that can be done, such as Jobs in Spark and pipelines in Airflow, Resources and useful information/documentation. Click on the option **Jobs** from the left menu to create a dataflow in Airflow.



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.

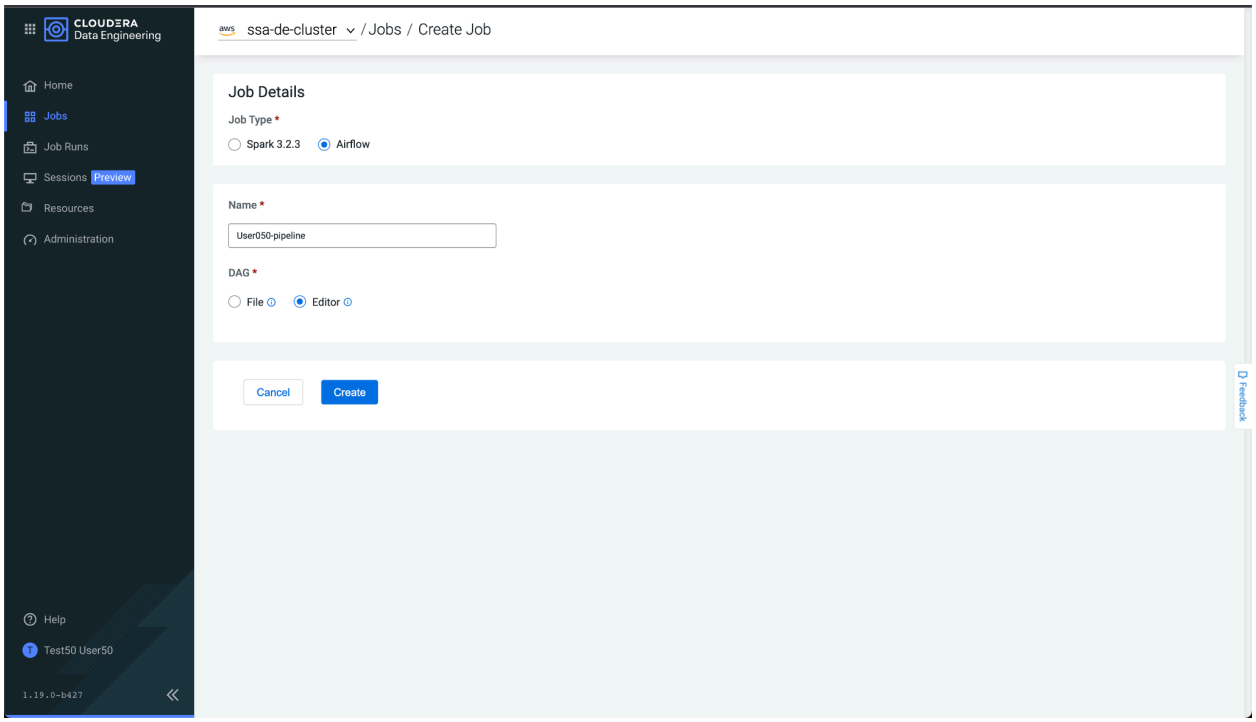
It is time to create our Job in Airflow. Click on **Create Job.**



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

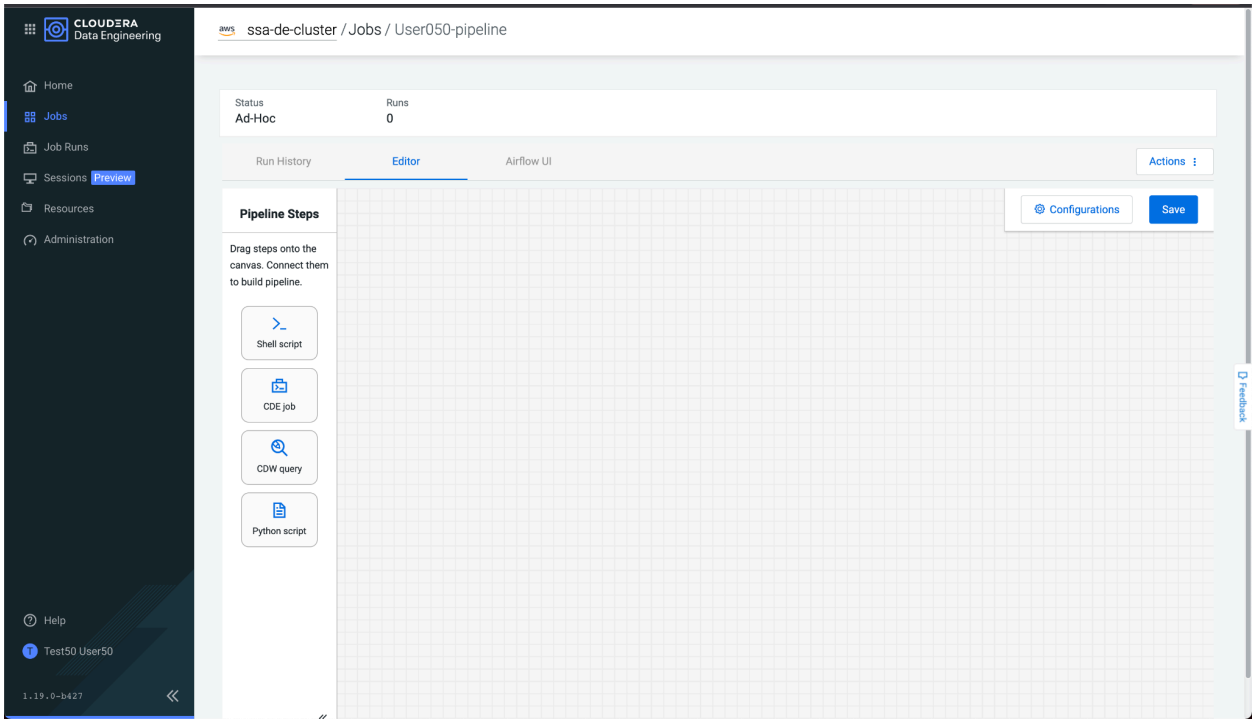
- Job Type: Airflow
- Name: Use the naming <assigned user>-pipeline. Replace <assigned user> with the user assigned to you. For example, user050
- DAG: Editor, to graphically configure the task.

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



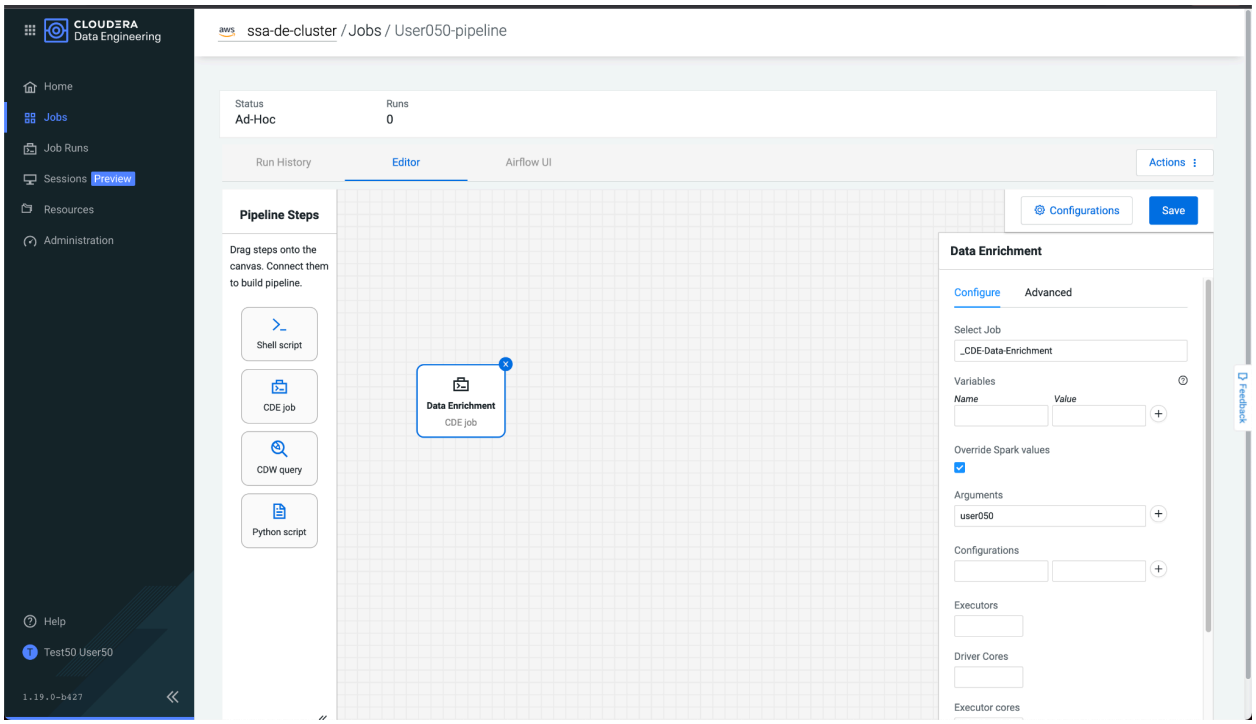
5. IMPORTANT!!!
If the editor page is NOT loading, go to the bottom of this doc for a workaround, if it does, continue as usual

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.



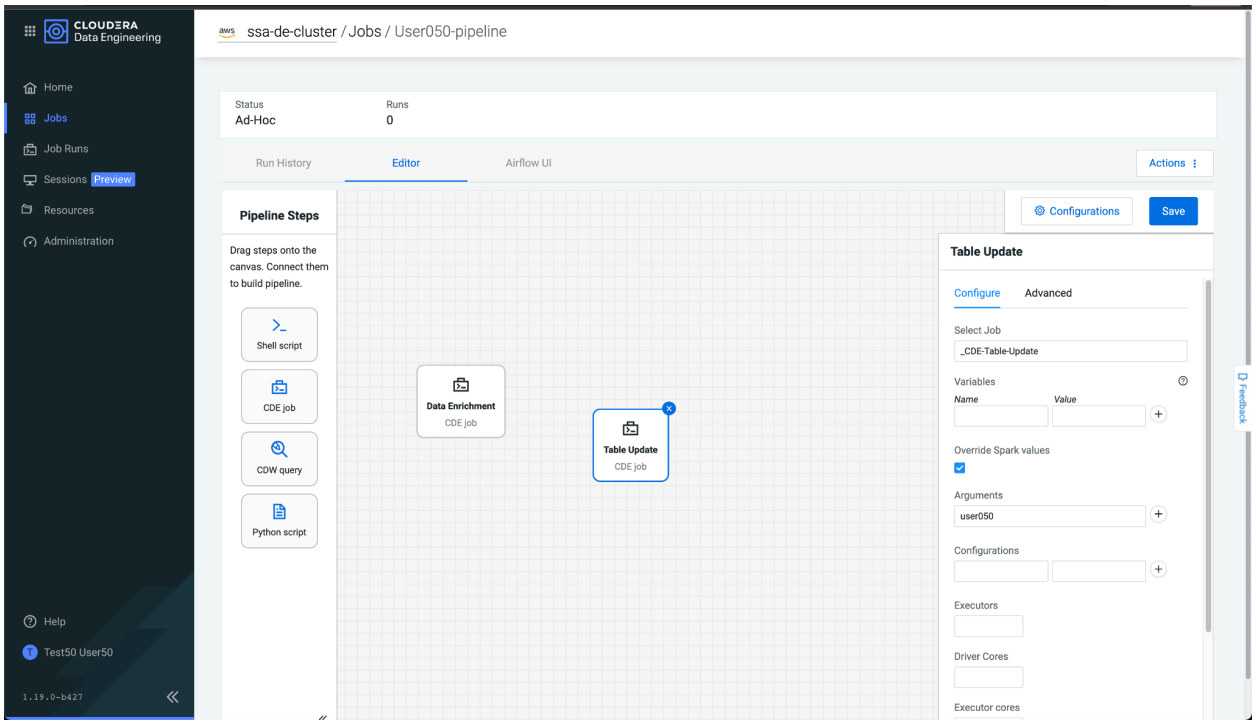
6. Let's start with the first Job. Click on the CDE Job button and drag onto the canvas, entering the following settings:

- **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

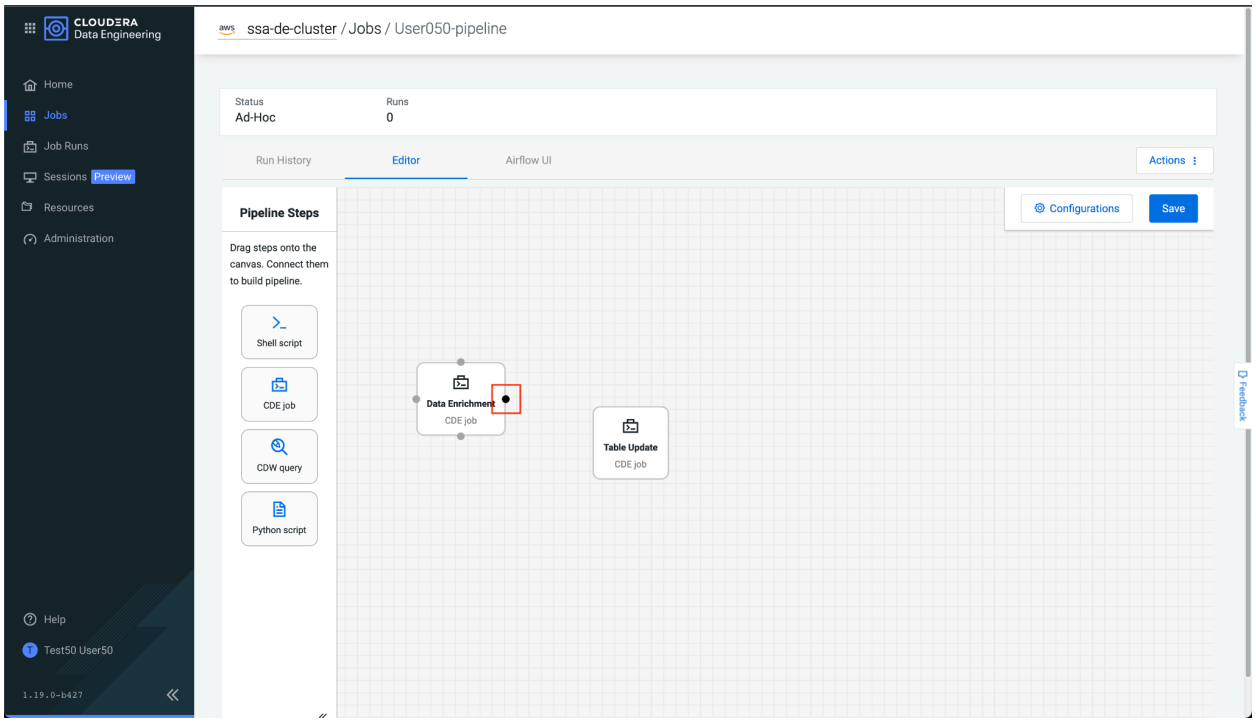


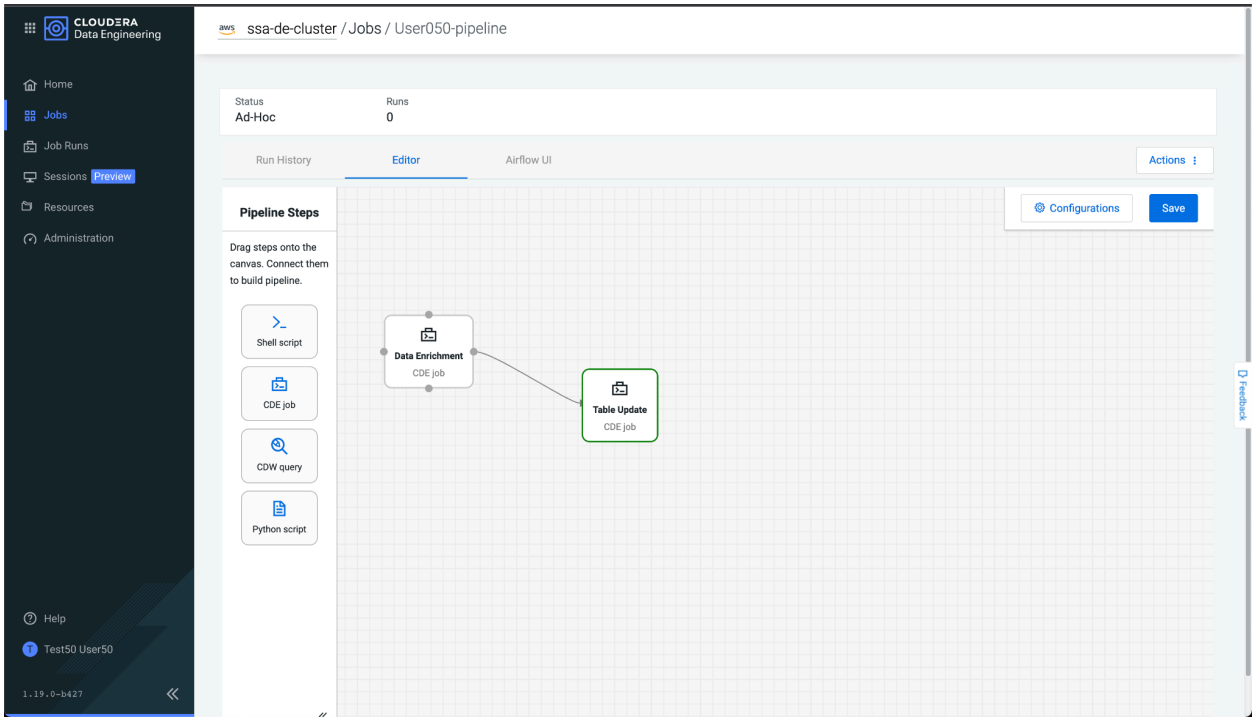
7. Configure the second Job. Click on the CDE Job button and drag onto the canvas, entering the following settings:

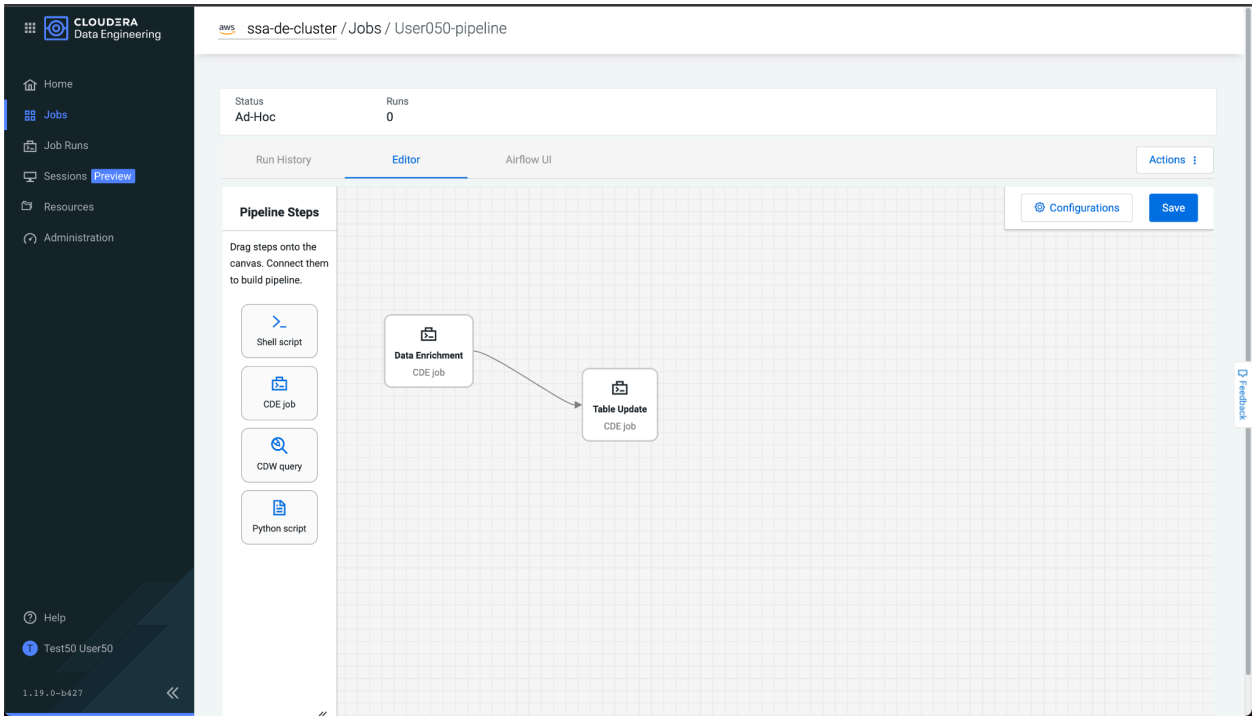
- **title/name**: Table Update
- **Select Job**: select the Job_*CDE-Table-Update*
- Check the checkbox **Override Spark values**. Additional options will appear below.
- **Arguments**: <assigned user>. Use the username assigned to you. For example, user050



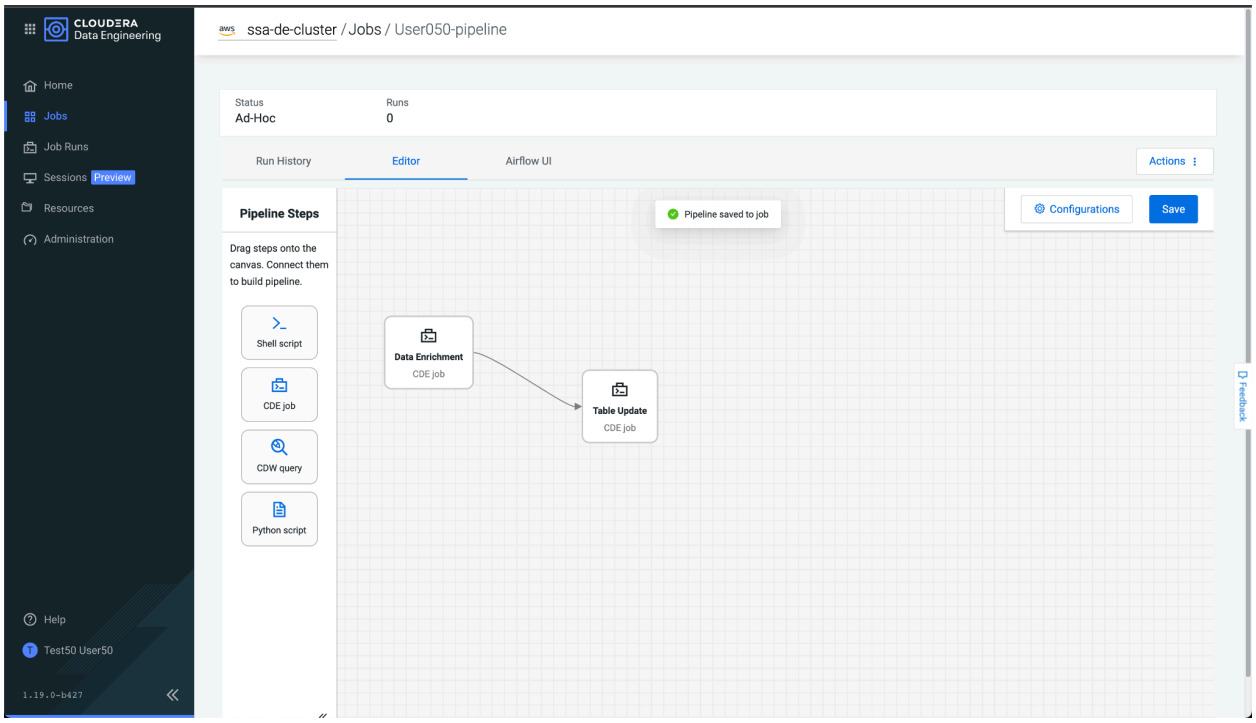
8. To set up the execution sequence, bind **Data Enrichment** with **Table Update**. For that, click on the right connector of the job of **Data Enrichment** and drag to the left connector of **Table Update**.



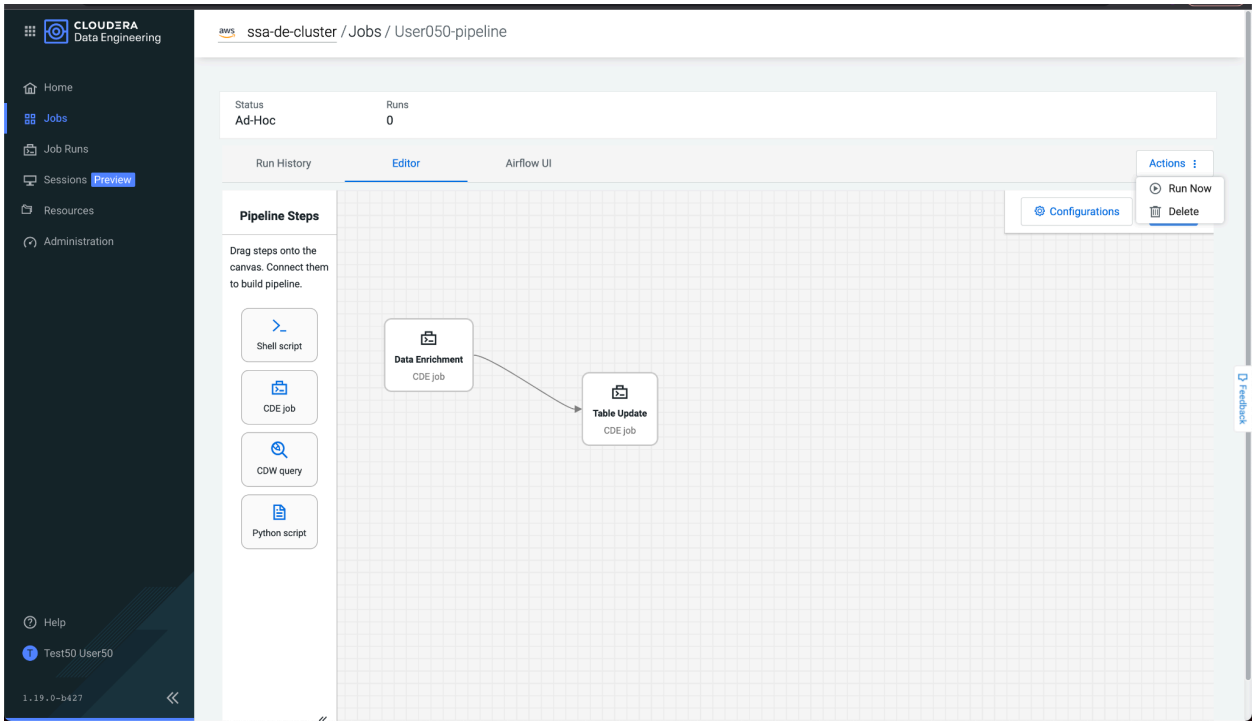




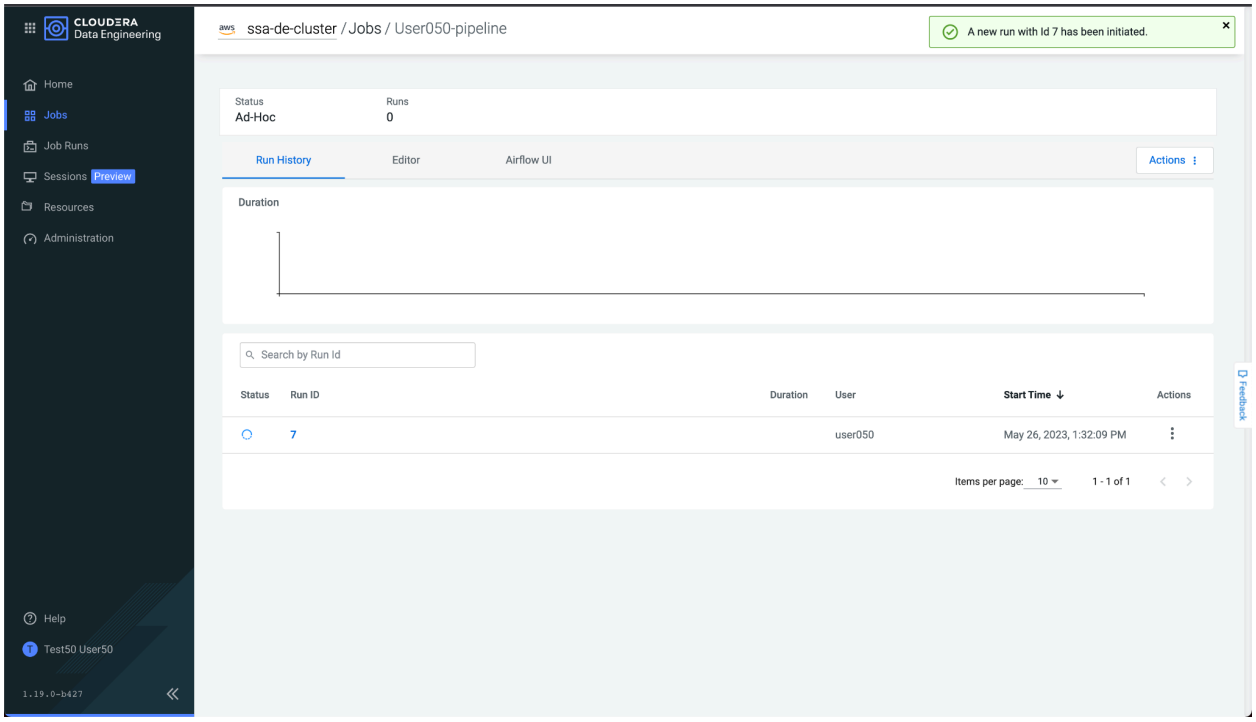
9. Once the Jobs have been joined, click on **Save** to save the settings made. You should see a message indicating **Pipeline saved to job**.



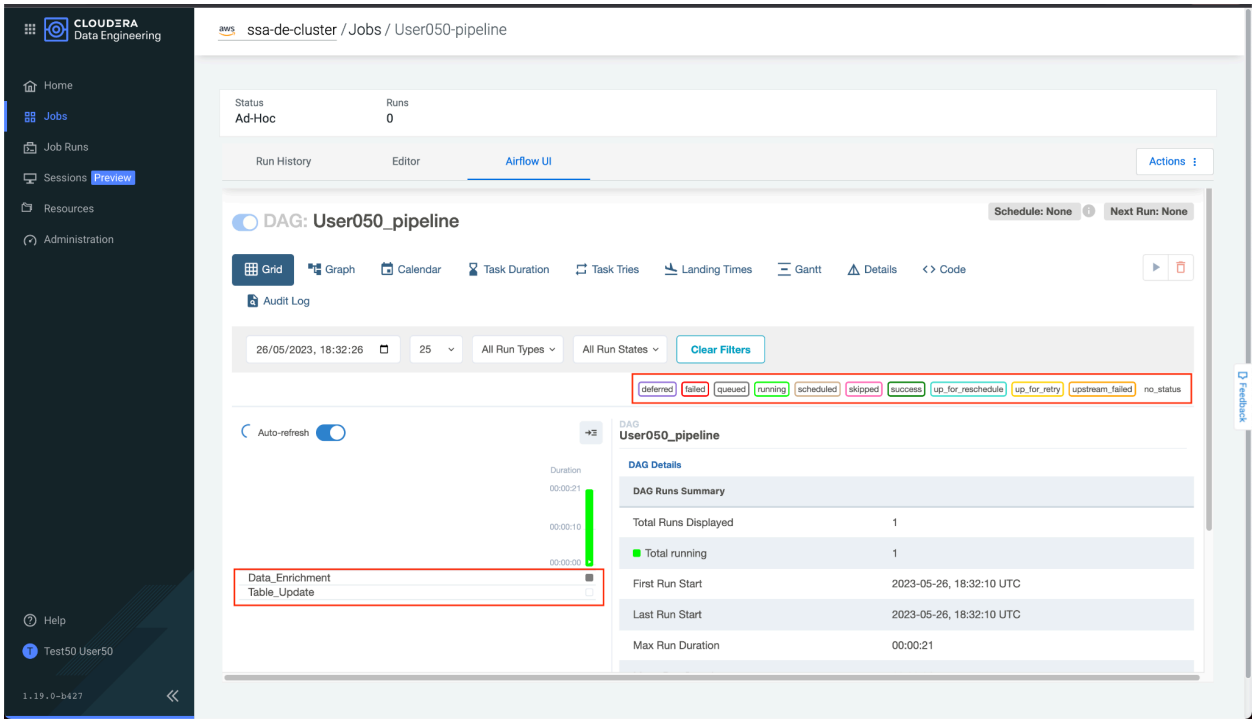
10. The time has come to run the pipeline. On the upper right side of the canvas, click **Actions -> Run Now**.



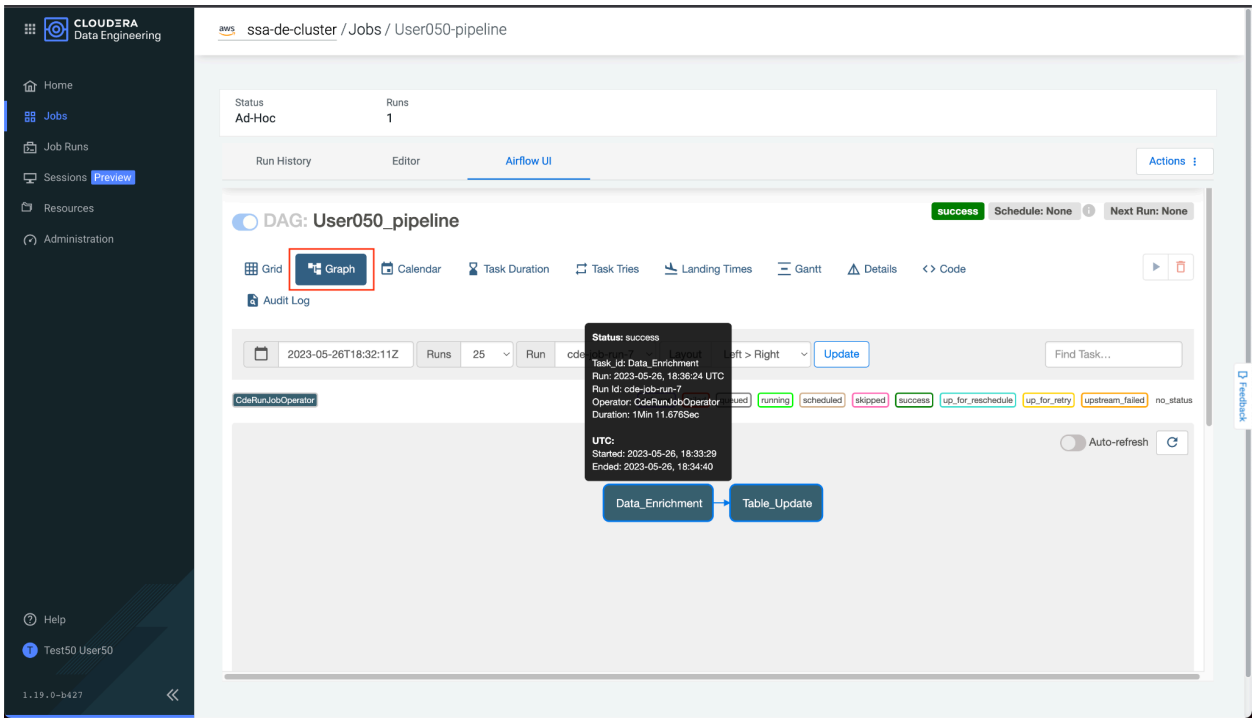
11. You should see the pipeline execution screen, indicating that the execution has been initialized.



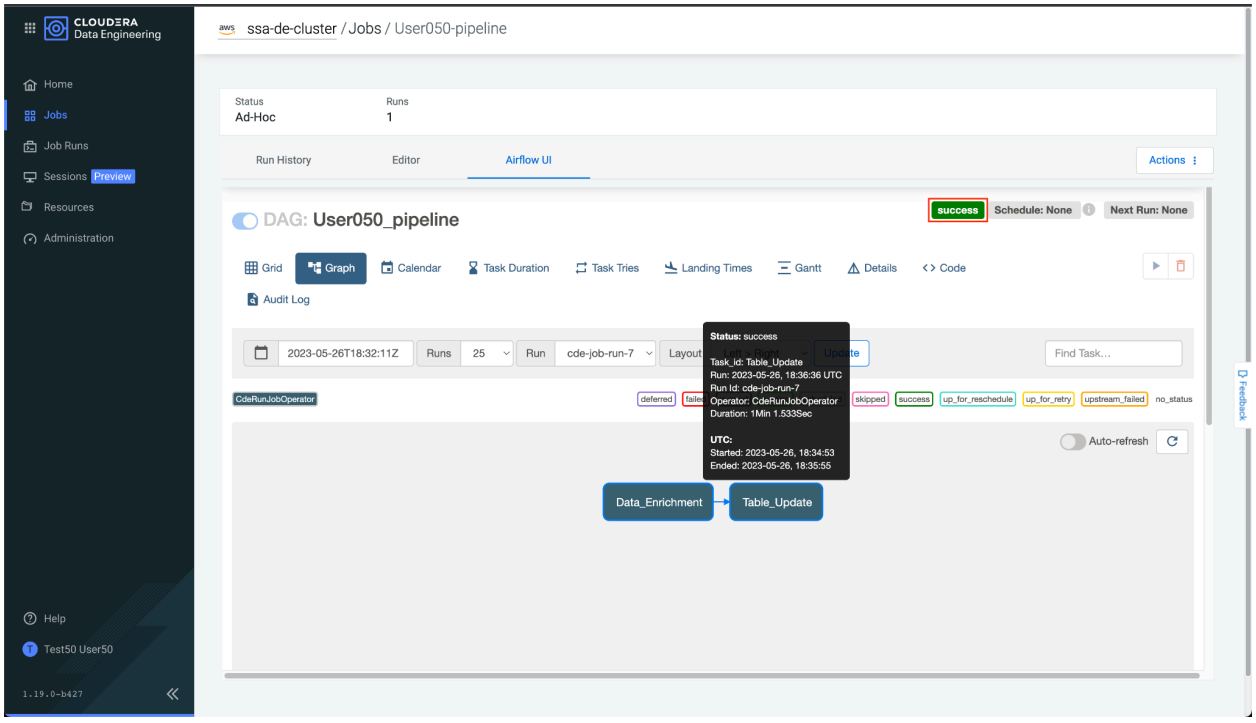
12. Click on the Airflow UI tab to see the execution detail of each step in the pipeline. The configured Data Enrichment and Table Update jobs are listed at the bottom left. The colors indicate the status of each job. Make sure the radio button **Auto-refresh** is enabled to automatically display the status of jobs.



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.



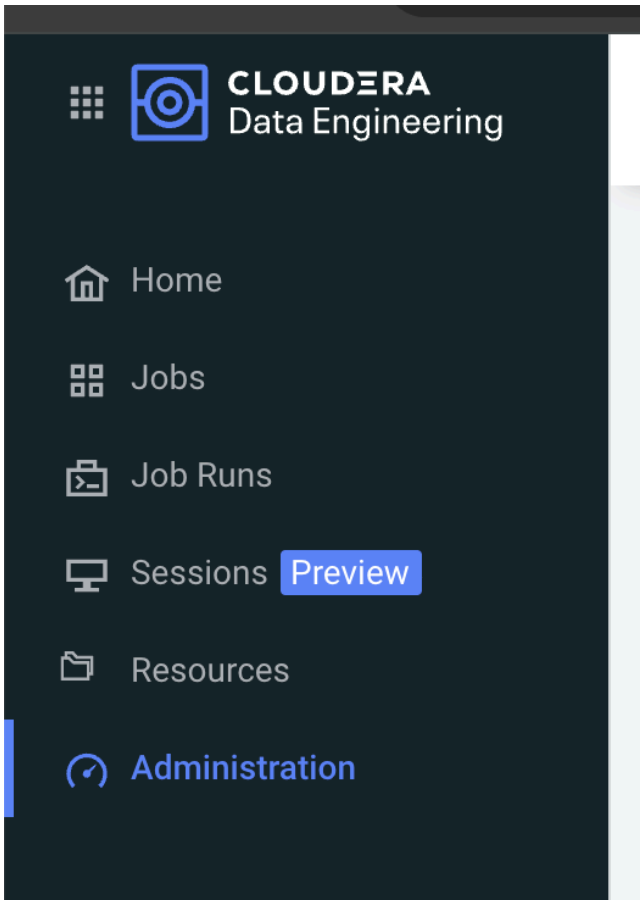
*The execution status appears next to the name of the pipeline (marked in red). If it is green and indicates **Success**, it means that the execution was successful.*



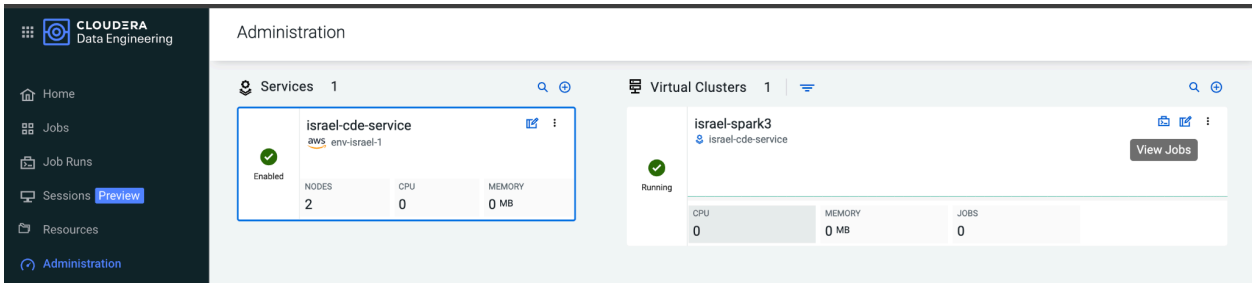
End of Lab 2

Workaround for Editor page not loading:

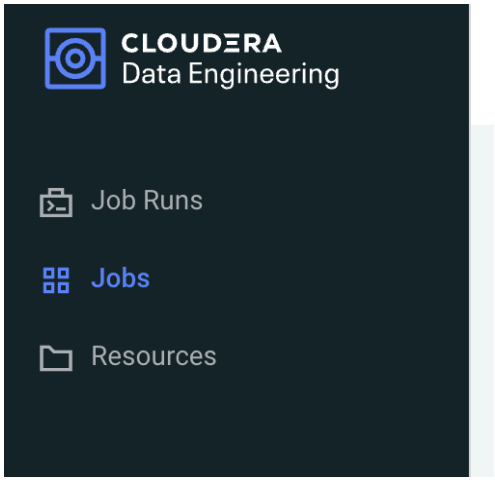
Click on Administration:



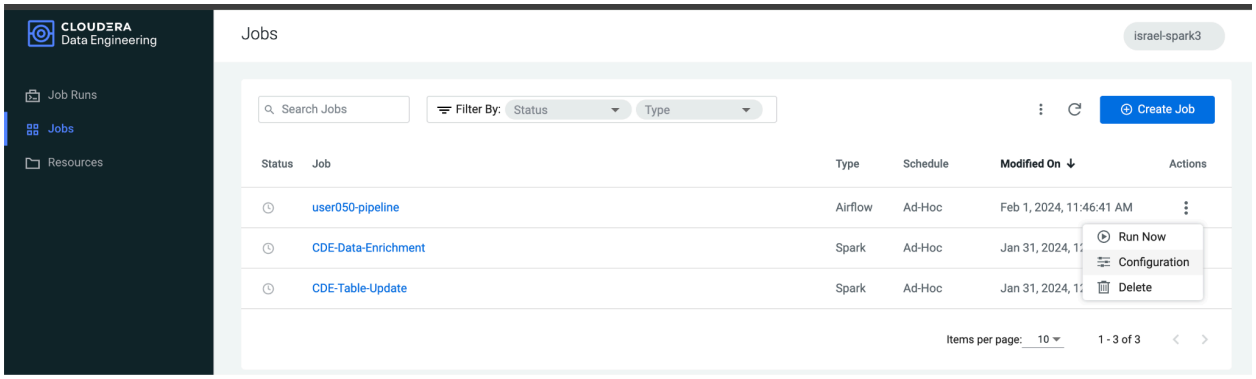
Click on the briefcase on the right side to view the spark cluster jobs:



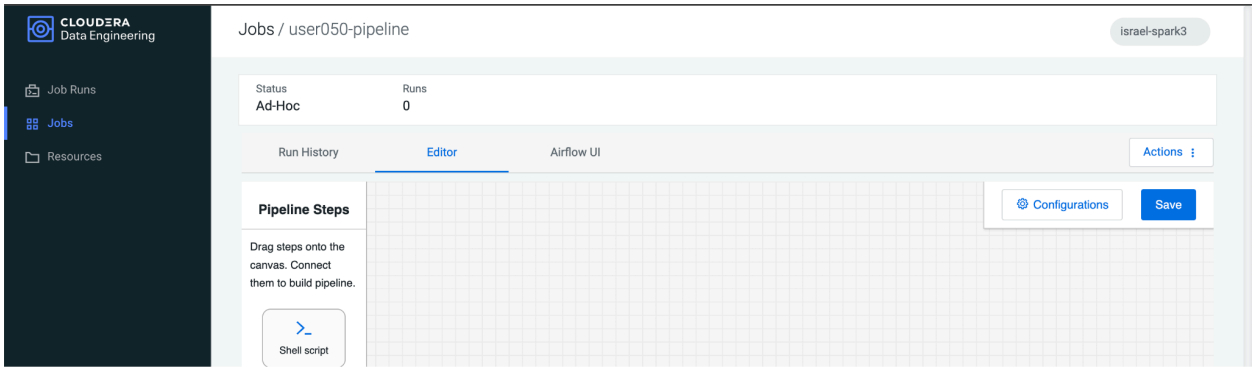
Click on Jobs:



Open Configuration for the previously created <userid>-pipeline job



Click on the Editor tab



The Editor canvas should load properly now, continue with step 6 above