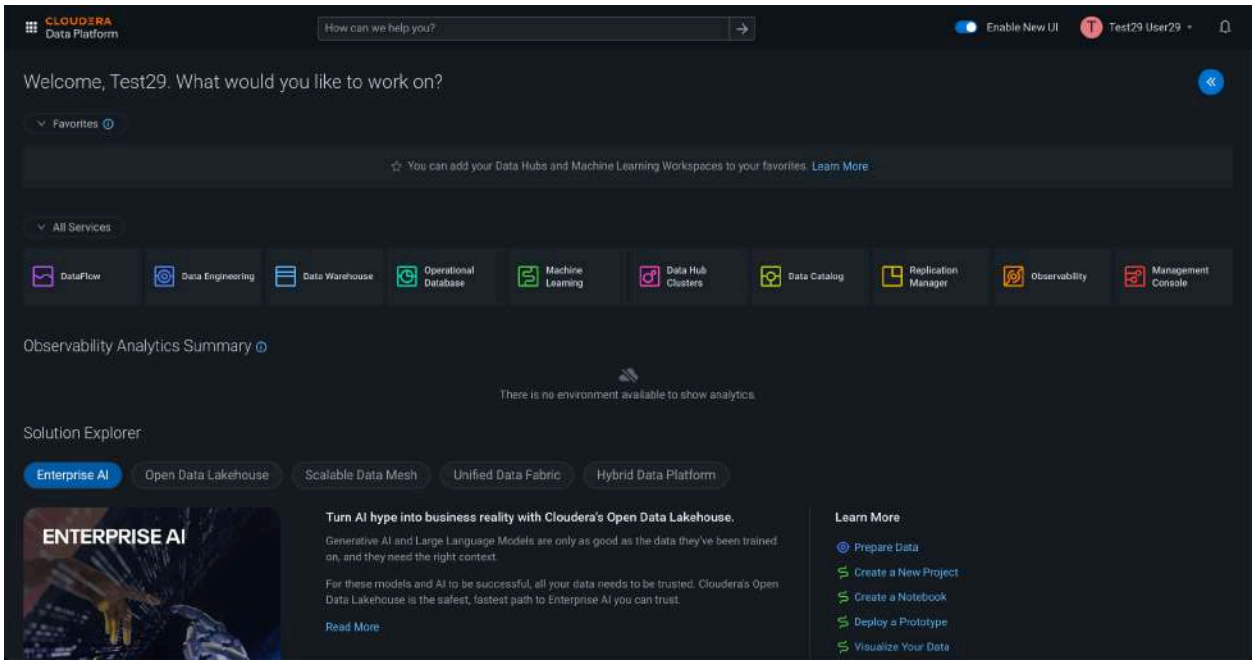
Data Lifecycle on CDP Public Cloud

Lab 002: Data Engineering Lab

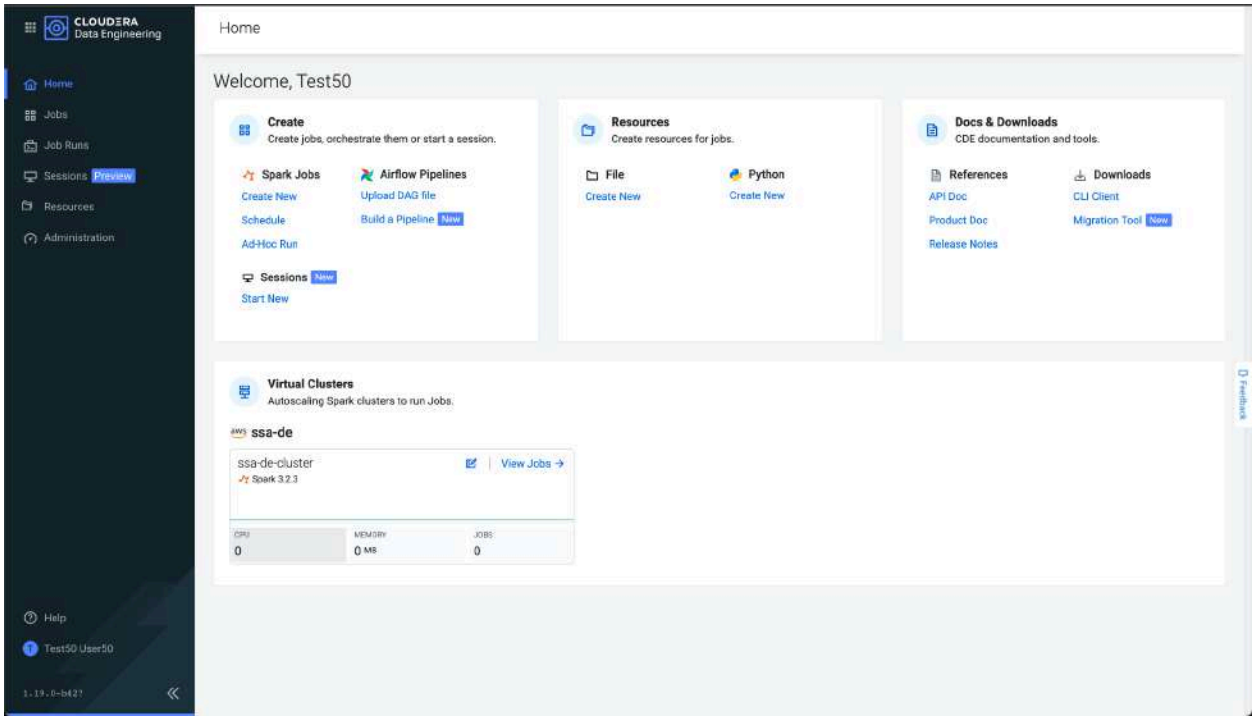
Goals:

- 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 **Data Engineering** from CDP PC Home:



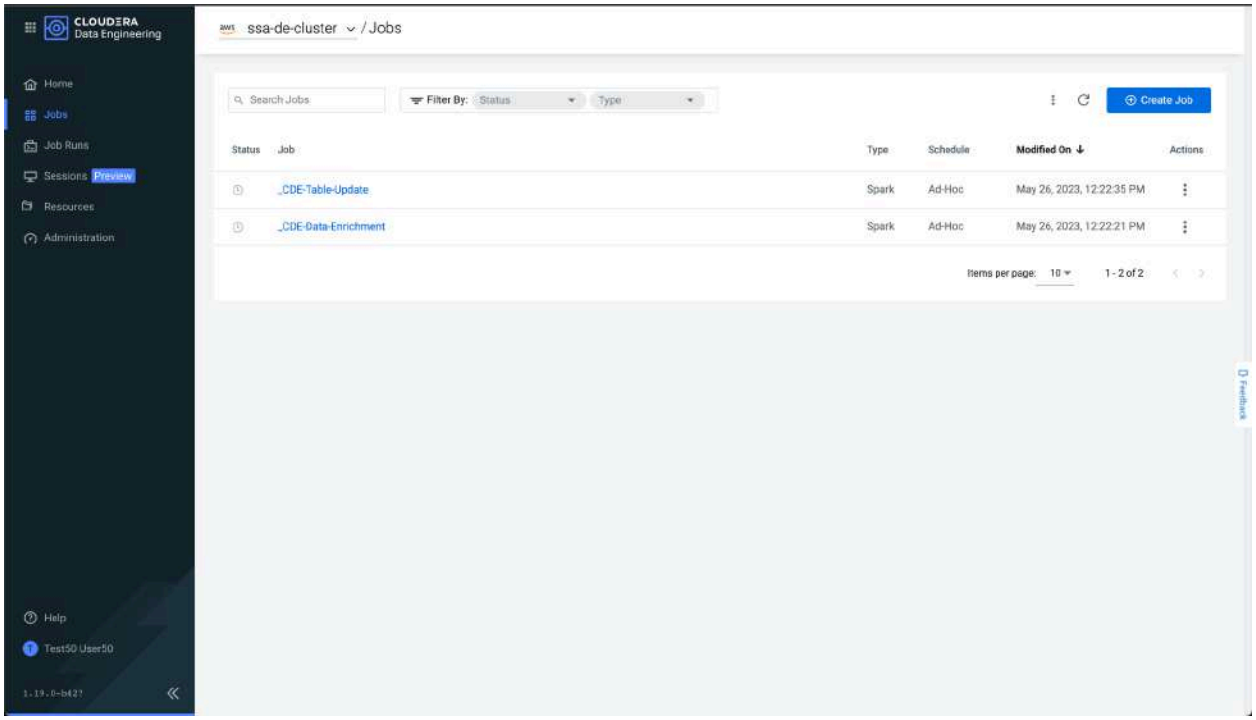
2. The Data Engineering Home shows all the actions that can be done, such as Jobs in Spark and pipelines in Airflow, Resources and useful information/documentation. Click on the option **Jobs** from the left menu to create a dataflow in Airflow.



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.

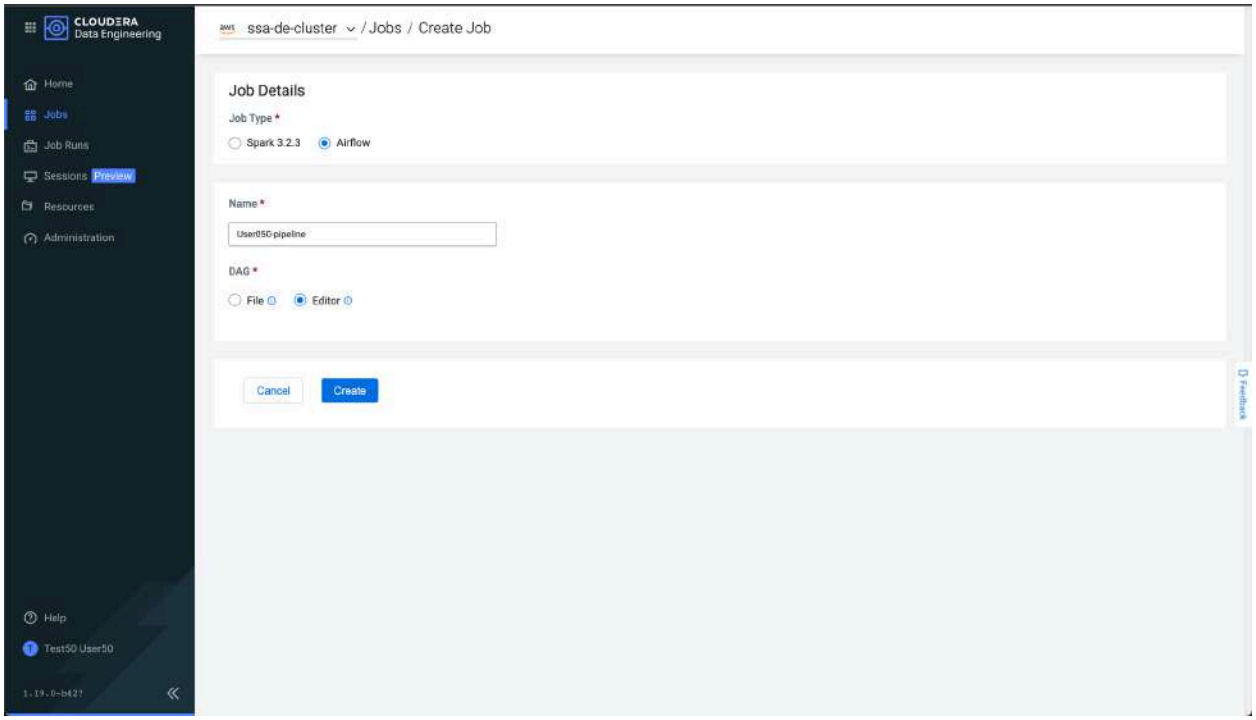
It is time to create our Job in Airflow. Click on **Create Job.**



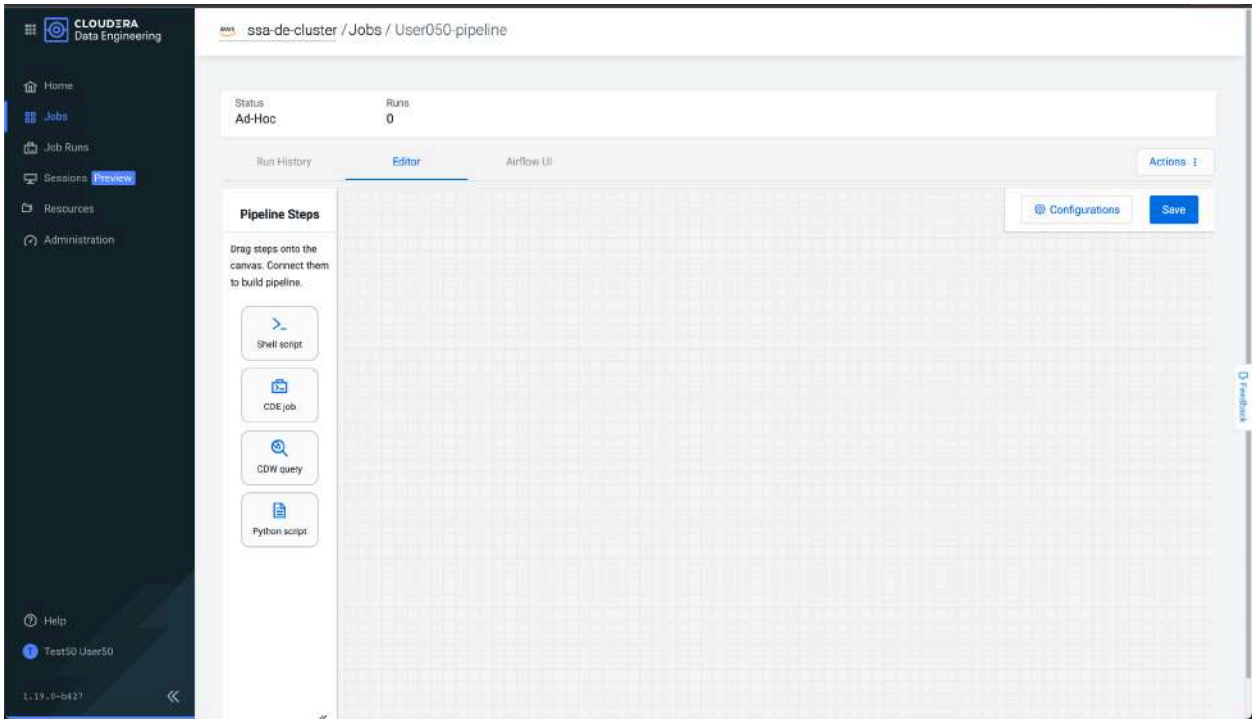
4. In the Job creation form, you must enter the following information:

- 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**.

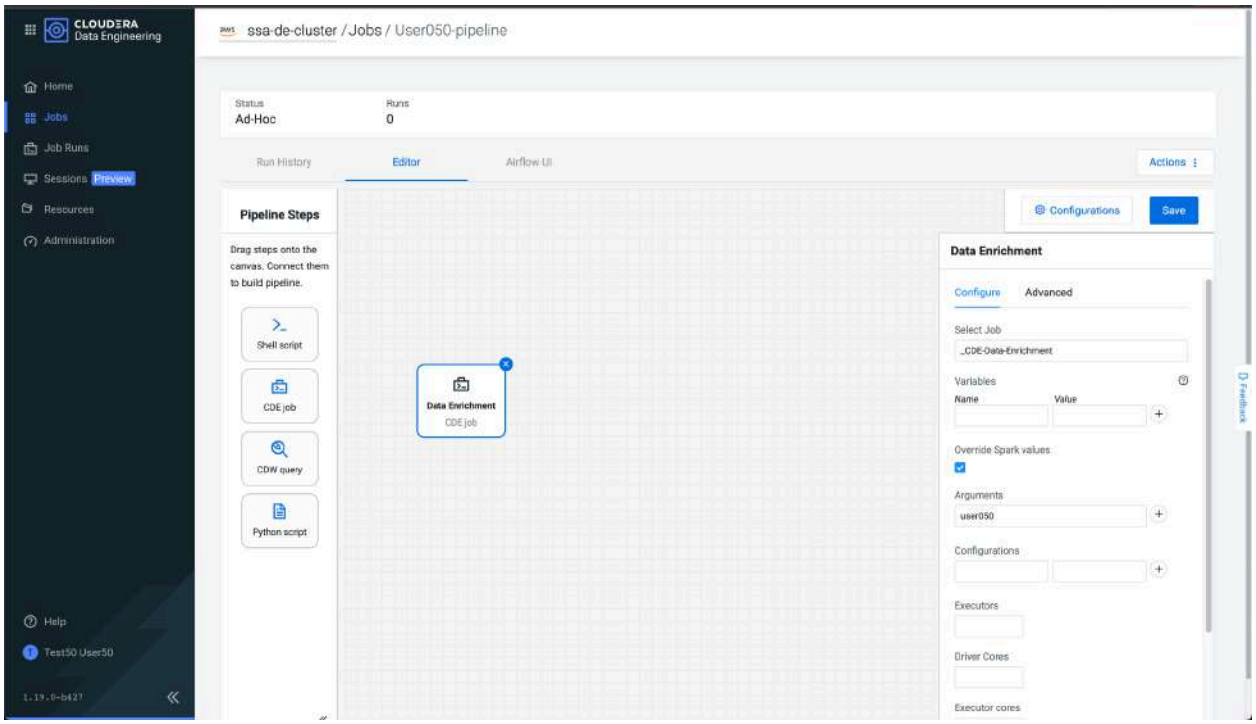


5. On the Job editing screen, select the Editor tab, and you will see the following canvas to drag the steps of the pipeline that we are going to create. In our case, we are going to create two CDE Jobs and relate them.



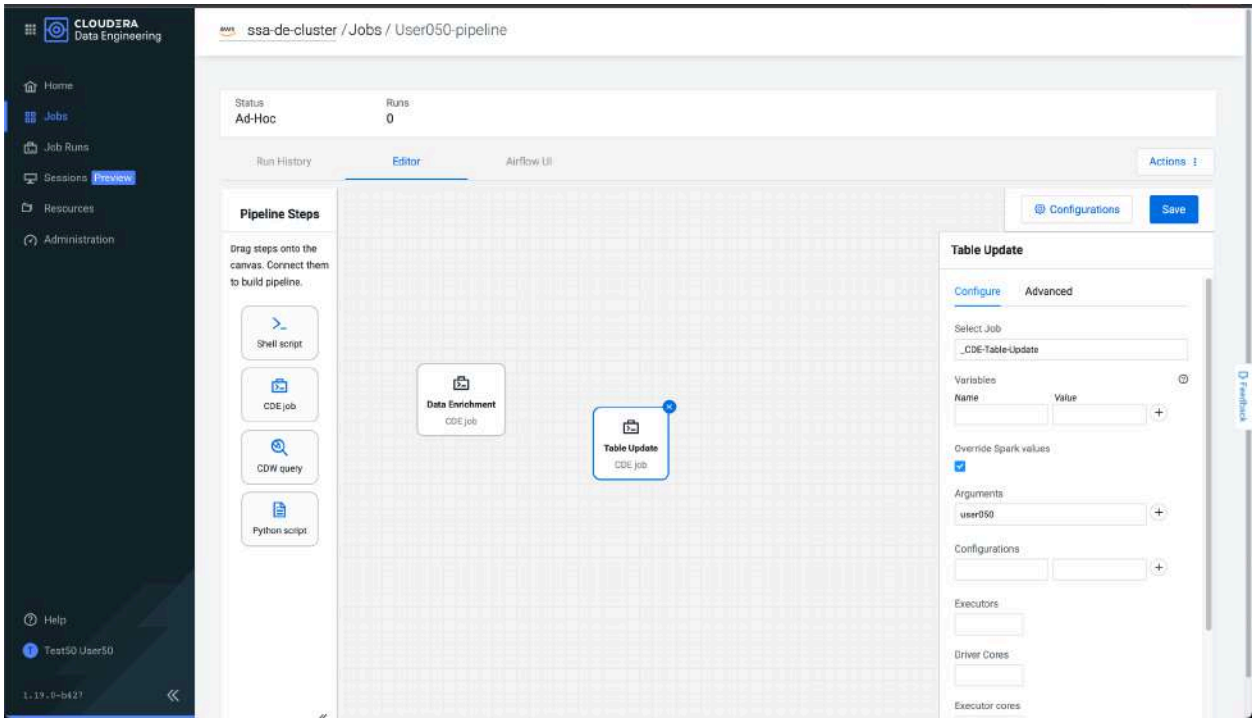
6. Let's start with the first Job. Click on the CDE Job button and drag onto the canvas, entering the following settings:

- **title/name**: Data Enrichment
- **Select Job**: select the **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

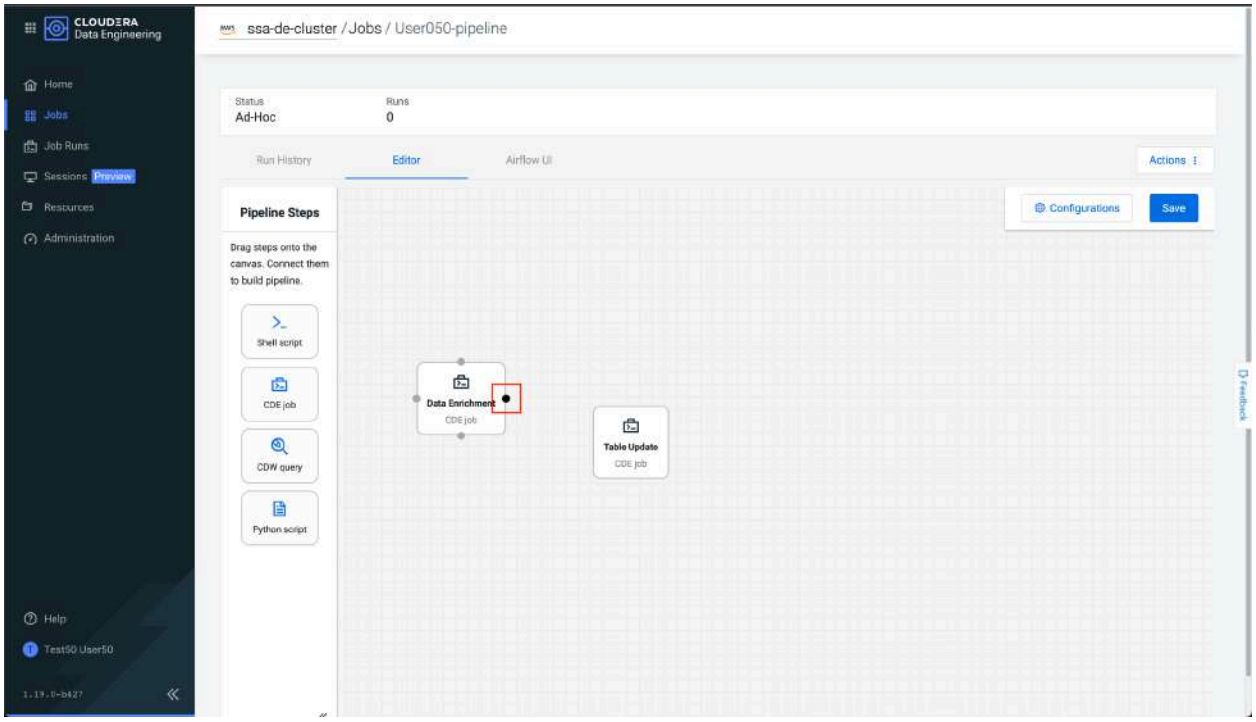


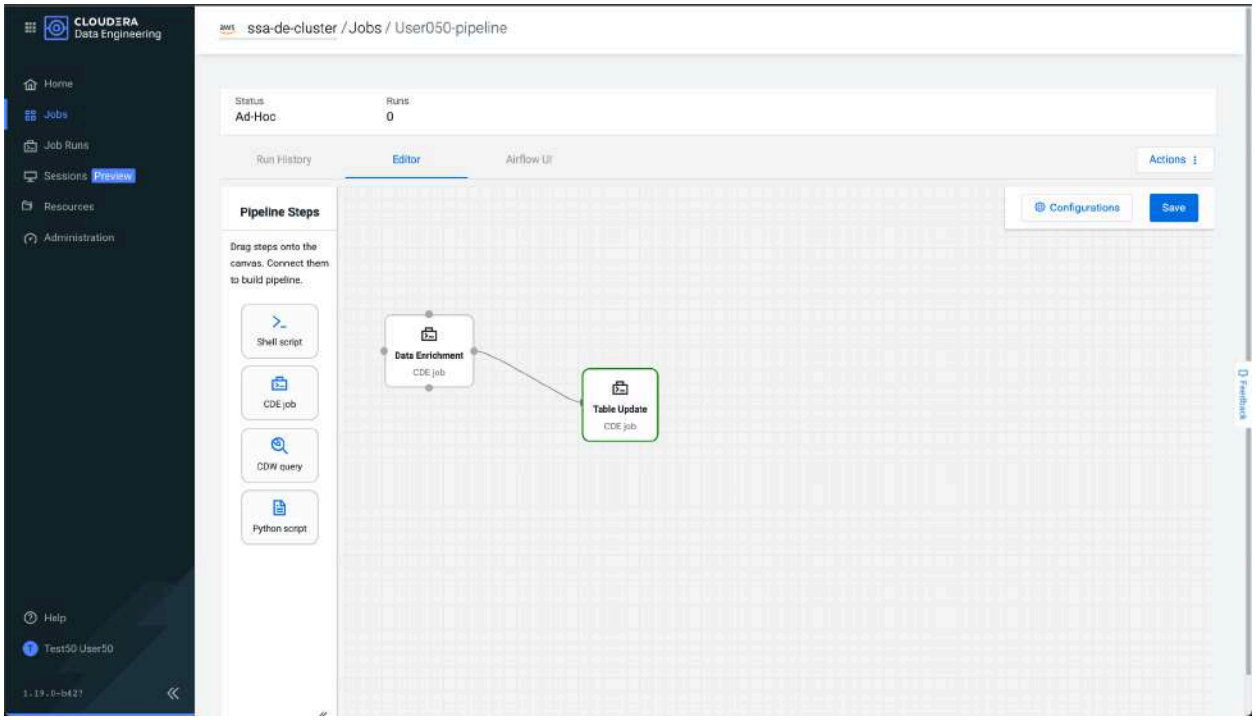
7. Configure the second Job. Click on the CDE Job button and drag onto the canvas, entering the following settings:

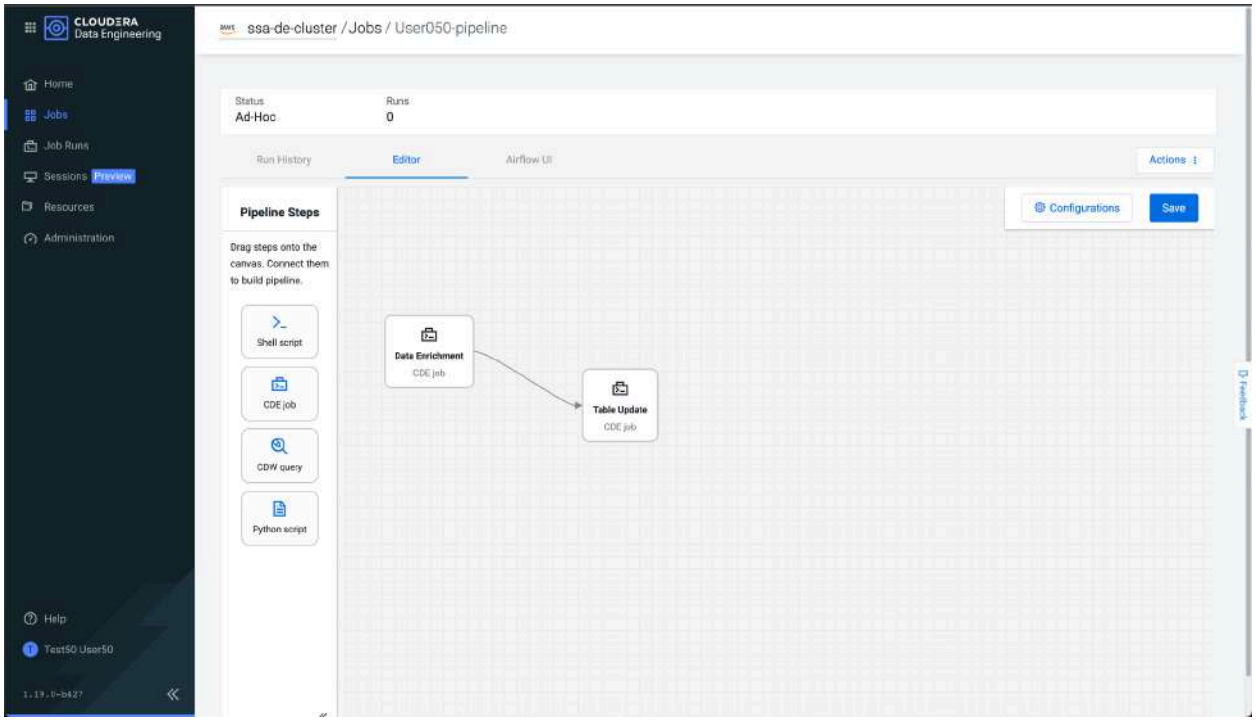
- **title/name**: Table Update
- **Select Job**: select the **CDE-Table-Update**
- Check the checkbox **Override Spark values**. Additional options will appear below.
- **Arguments**: <assigned user>. Use the username assigned to you. For example, user050



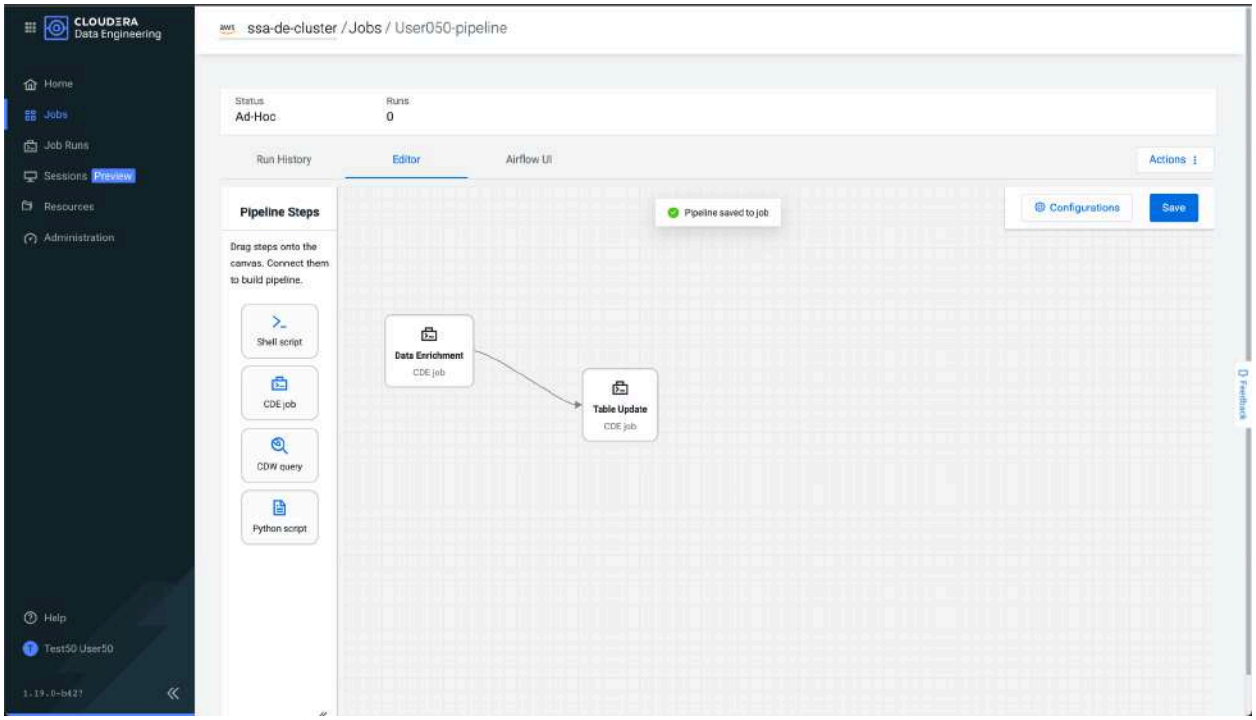
8. To set up the execution sequence, bind **Data Enrichment** with **Table Update**. For that, click on the right connector of the job of **Data Enrichment** and drag to the left connector of **Table Update**.



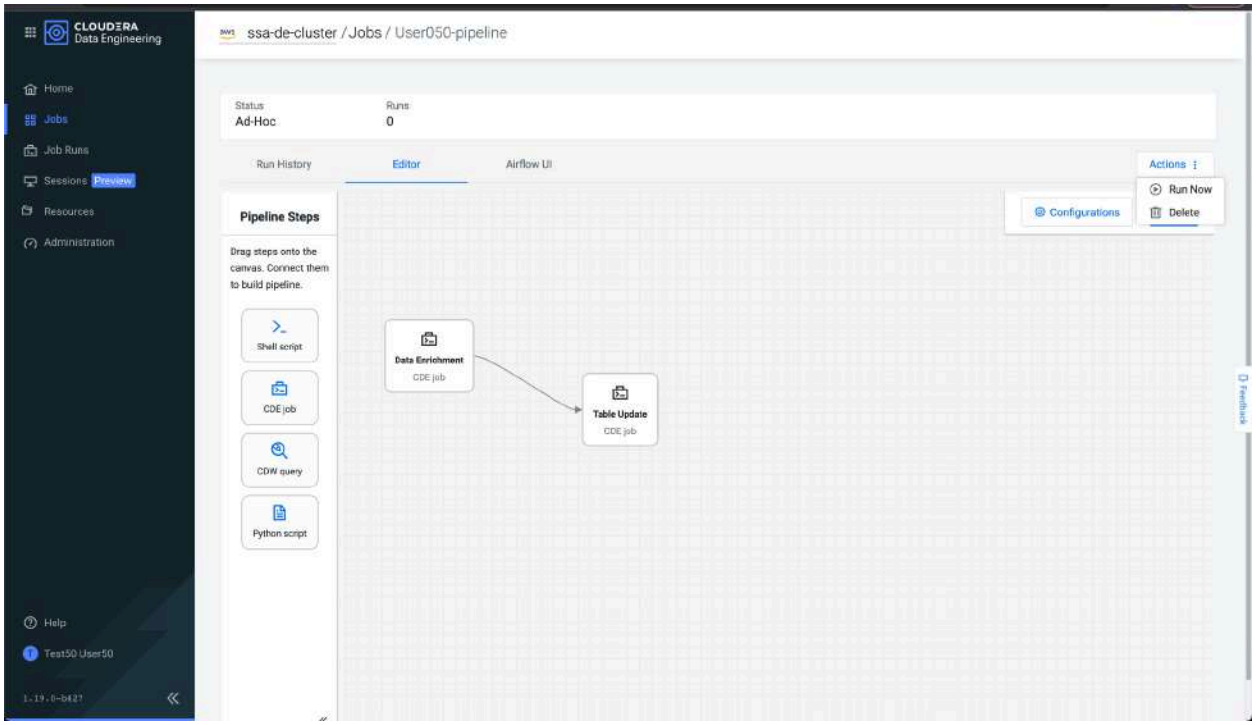




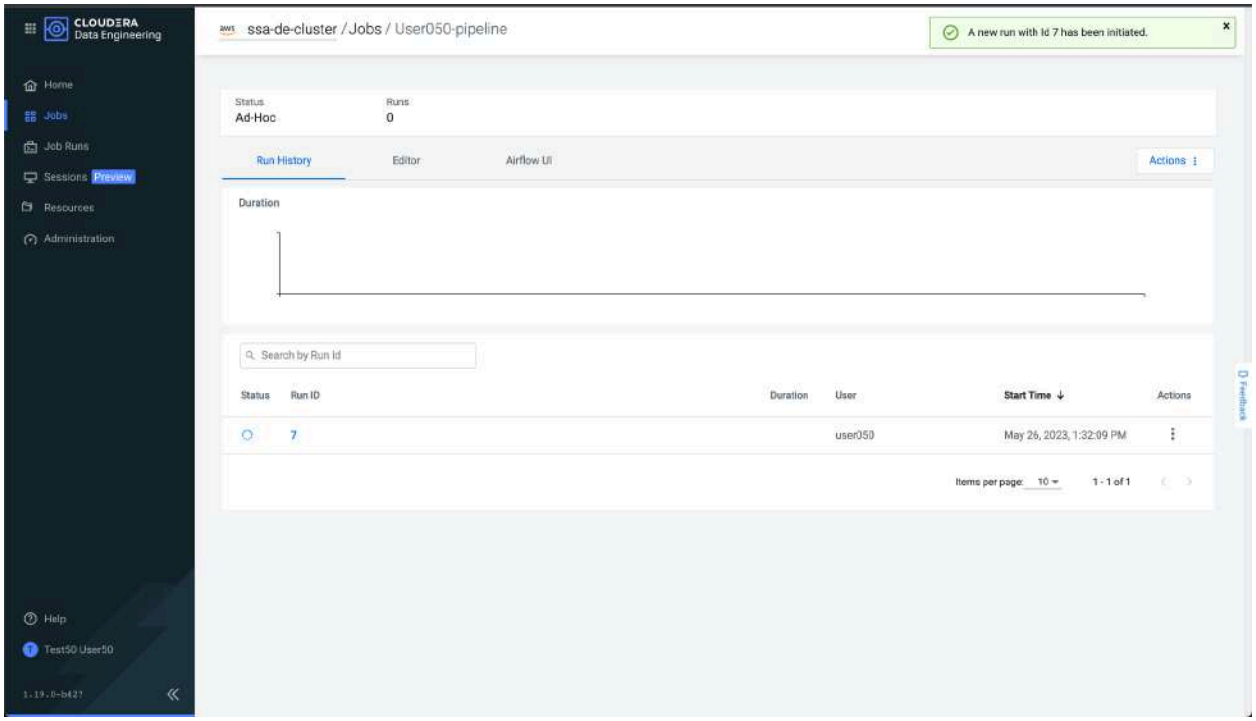
9. Once the Jobs have been joined, click on **Save** to save the settings made. You should see a message indicating **Pipeline saved to job**.



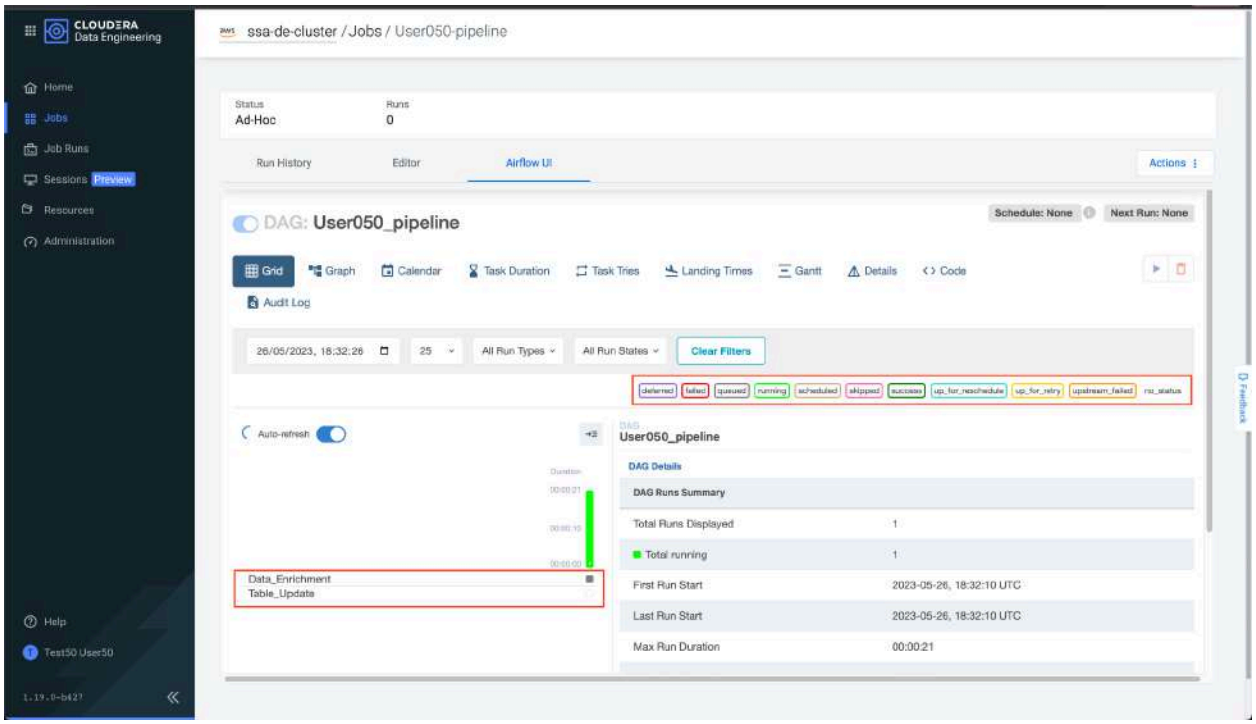
10. The time has come to run the pipeline. On the upper right side of the canvas, click **Actions -> Run Now**.



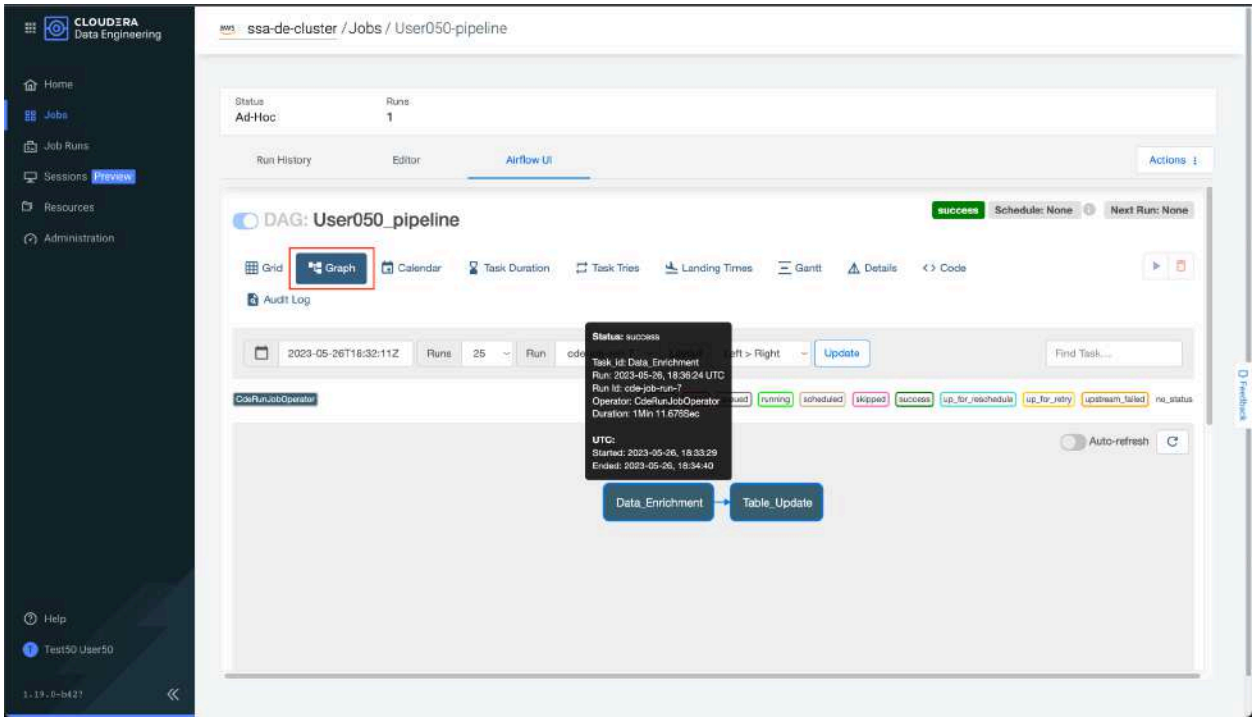
11. You should see the pipeline execution screen, indicating that the execution has been initialized.



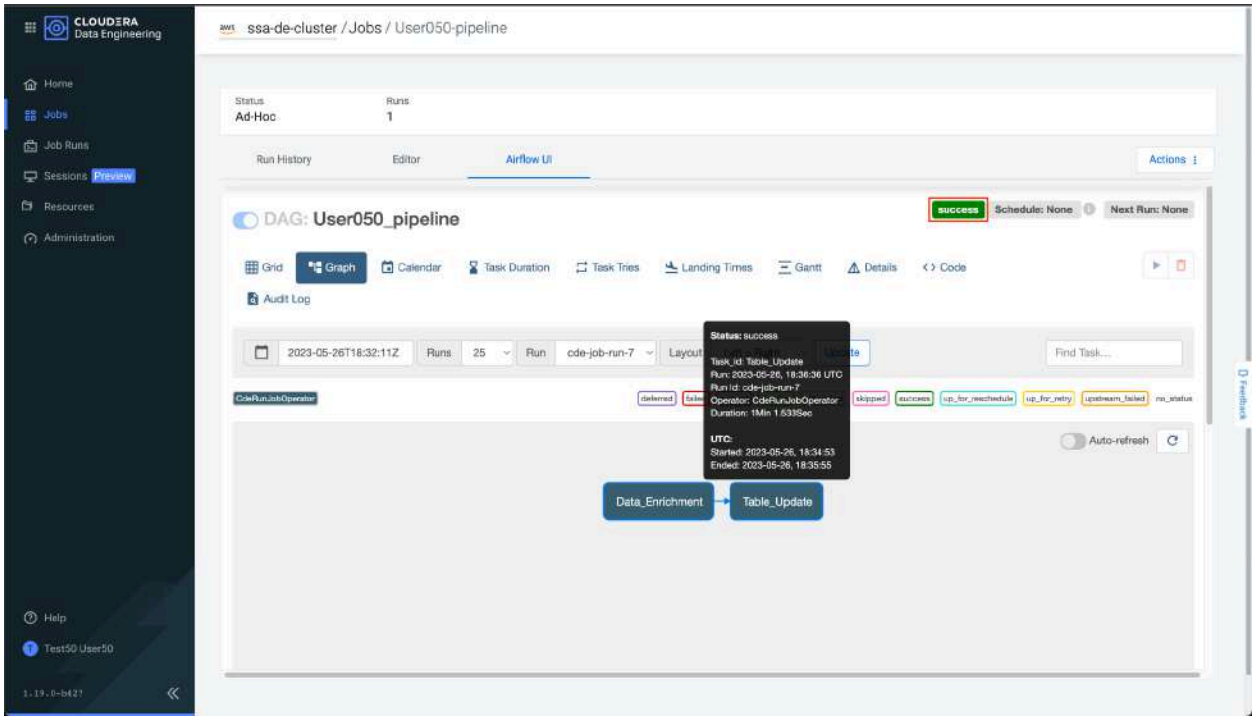
12. Click on the Airflow UI tab to see the execution detail of each step in the pipeline. The configured Data Enrichment and Table Update jobs are listed at the bottom left. The colours indicate the status of each job. Make sure the radio button **Auto-refresh** is enabled to automatically display the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


CLOUDERA
Data Engineering
Home
Jobs
Job Runs
Sessions
Preview
Resources
Administration
Help
Test50 User50
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
Find Task...
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