

Workflows Hands-on

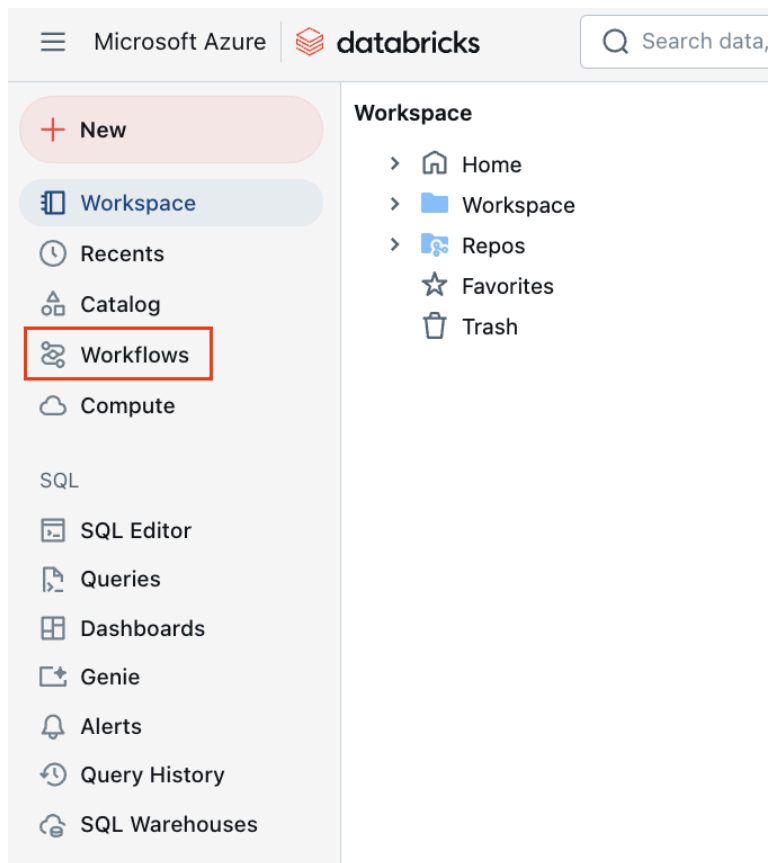
01. Workflow simples

Nesse exercicio vamos orquestrar os notebooks criados anteriormente:

- 01 - Bronze Layer (autoloader - csv)
- 01Lab - Bronze Layer (autoloader - json)
- 02 - Silver Layer
- 03 - Gold Layer (Join)

Tudo o que estiver imprimindo (display) vocês podem comentar.

1. Clique em **Workflows**



2. Clique em **Create Job**

Microsoft Azure databricks

Search data, notebooks, recents, and more... ⌘ + P

field-eng-east

Workflows

Jobs Job runs Pipelines

Filter jobs

Owned by me Accessible by me Favorites

Send feedback Load tutorial Create job

Name	Tags	Cre...	Trigger	Recent runs
Gmarket for/each condition		Ins		⊗ ⊖ ⊕ ⊕ ▶ ⋮
Starbucks_Workflow		Ins		— — — — ⊕ ▶ ⋮
gamer-lifecycle	CME solacc_testing	Dur		— — — — ⊖ ⊗ ▶ ⋮
sadasd		Ale	Paused - Scheduling	— — — — — ▶ ⋮
Simple Job	test test2	Dar		⊗ ⊗ ⊕ ⊕ ⊕ ▶ ⋮
(andres_urrutia) mfg_llm_demo		Anc		— — — — — ▶ ⋮
(andrew_kraemer) mfg_llm_demo		Anc		— — — — — ▶ ⋮
(dave_geyer) mfg_llm_demo	2024-12-30	Dav		▶ ⊖ ⊖ ⊖ ⊖ ▶ ⋮
(jai_karve) mfg_llm_demo		Jai		— — — — — ▶ ⋮
(paul_mrcek) mfg_llm_demo		Pat		— — — — ⊗ ▶ ⋮

3. No topo abaixo de [Workflows](#) > [Jobs](#) > edite o nome do seu Workflow, exemplo: clientes_workflow_SEUNOME

Microsoft Azure databricks

Search data, notebook

Workflows > Jobs >

clientes_workflow_anasanchez ☆

Runs Tasks

4. Preencha os seguintes campos dentro da Task:
- Task name: Clientes Bronze
 - Type: Notebook
 - Path: Caminho para o notebook 01 - Bronze Layer (autoloader - csv)

The screenshot shows the Databricks Jobs interface. On the left is a sidebar with navigation options: New, Workspace, Recents, Catalog, Workflows (selected), Compute, SQL, SQL Editor, Queries, Dashboards, Genie, Alerts, Query History, SQL Warehouses, Data Engineering, Job Runs, Data Ingestion, Pipelines, and Machine Learning. The main header shows 'Microsoft Azure' and 'databricks' with a search bar. The breadcrumb is 'Workflows > Jobs >'. The job title is 'New Job 2024-09-05 20:46:58'. Below the title are tabs for 'Runs' and 'Tasks'. The 'Tasks' tab is active, showing a canvas with a task named 'Clientes_Bronze' (path: ...1 - Bronze Layer (autoloader - csv)). A '+ Add task' button is visible on the canvas. On the right side of the canvas are zoom controls. Below the canvas is a form to configure the task:

Task name*	Clientes_Bronze
Type*	Notebook
Source*	Workspace
Path*	...1 - Engenharia de Dados/GUIDE/01 - Bronze Layer (autoloader - csv)
Compute*	Serverless
Dependent libraries	+ Add

4. Agora, clique em **+ Add Task > Notebook** e preencha os seguintes campos dentro da Task:

- Task name: Clientes Prata
- Type: Notebook
- Path: Caminho para o notebook 02 - Silver Layer
- Depends on: Clientes_Bronze

Seu workflow ficará da seguinte forma:

clientes_workflow_anasanchez ☆

Runs Tasks



5. No menu lateral, abaixo de **Job Details**, em **Tags** preencha:

- MODO : TREINAMENTO
- AMBIENTE: PD-MDL
- SIGLA: SLN

Tags



Key* ⓘ

Value ⓘ

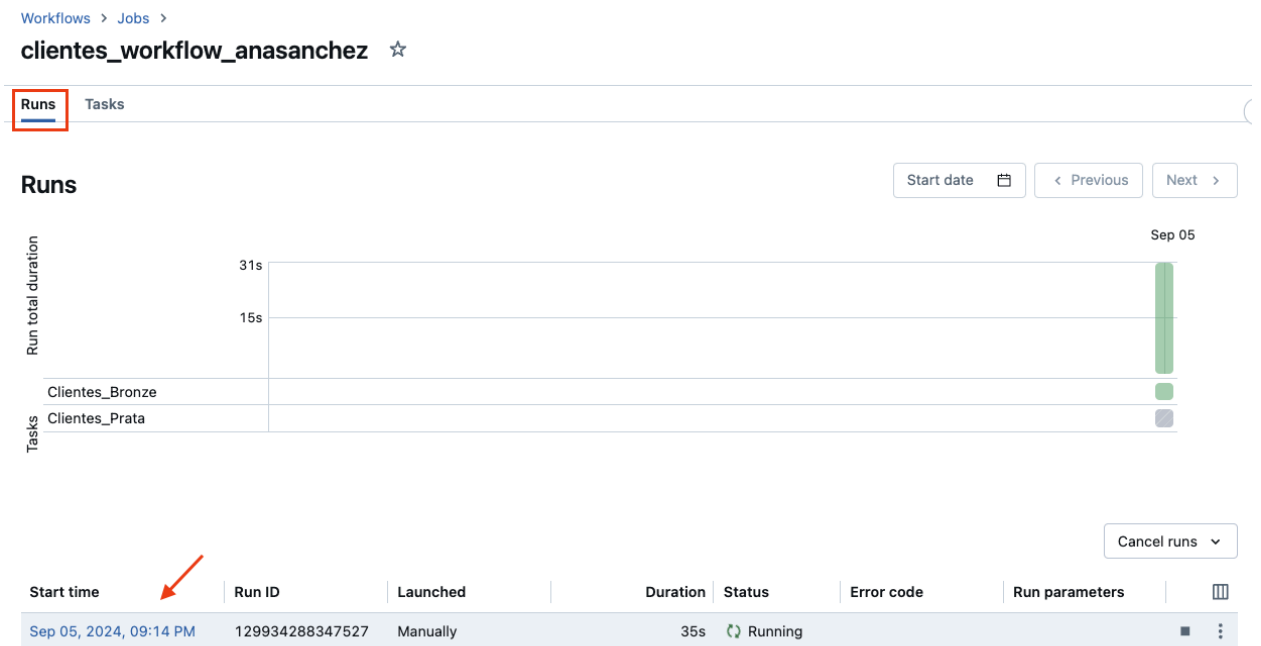
MODO	TREINAMENTO	
AMBIENTE	PD-MDL	
SIGLA	SLN	
Key	Value (optional)	

6. Agora, acesse em outra aba o **SQL Editor**. E conte quantas linhas temos na tabela.

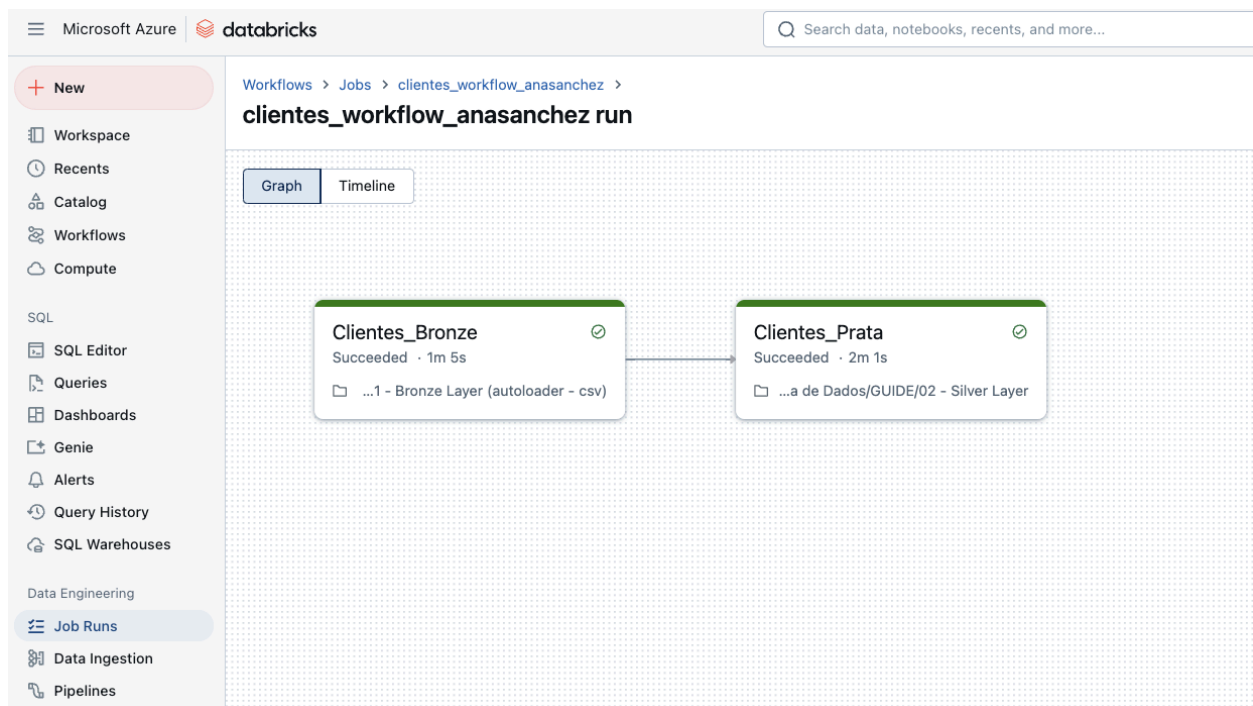
7. Volte para o workspace, e dentro da pasta **Workflows** execute o notebook **01 - workflows_ingerir novos clientes**.

8. Agora, volte para o seu Workflow e clique em **Run now**.

9. Clique em **Runs** para acompanhar a execução de seu workflow



10. Para ver o detalhe de execução de cada etapa, clique no link do horário de execução e verifique se seu job executou com sucesso.



11. Volte para o **SQL Editor**. E verifique quantas linhas temos na tabela agora.

02. Workflow - Trigger Mode: File Arrival

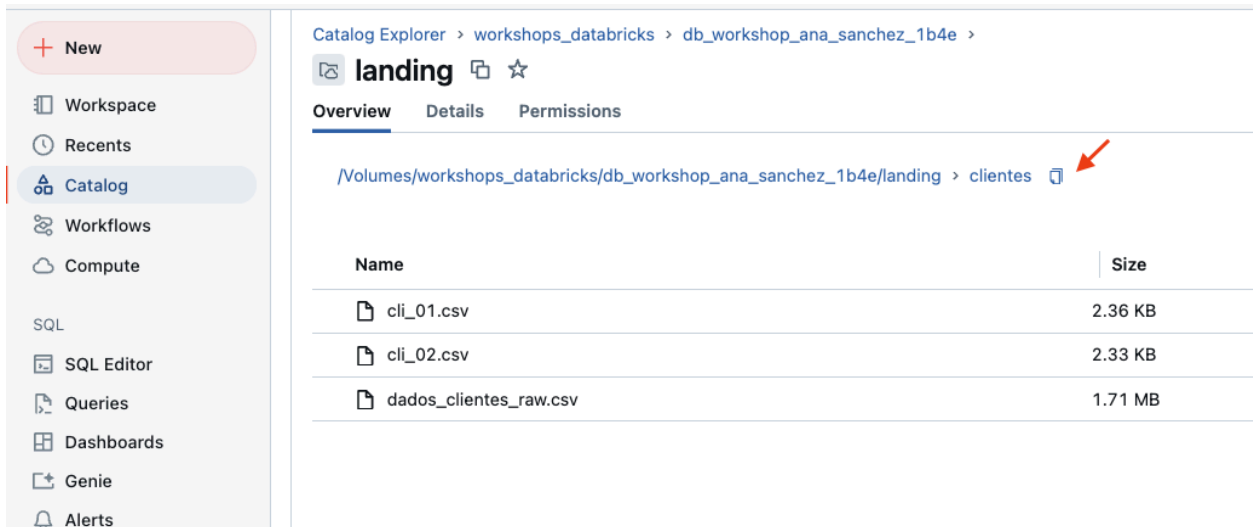
1. Volte para a página inicial dos workflows e selecione o seu workflow e no menu lateral direito, clique em **Add Trigger**.

Schedules & Triggers




None

Add trigger

2. Selecione o **Trigger Type** como **File Arrival** e em **Storage Location** preencha com **/Volumes/workshops_databricks/db_workshop_SEU_SCHEMA/landing/clientes/**. Você consegue encontrar essa informação através do Catalog acessando o diretório e copiando o caminho:



The screenshot shows the Databricks Catalog Explorer interface. On the left is a sidebar with navigation options: New, Workspace, Recents, Catalog (selected), Workflows, Compute, SQL, SQL Editor, Queries, Dashboards, Genie, and Alerts. The main panel displays the path `Catalog Explorer > workshops_databricks > db_workshop_ana_sanchez_1b4e > landing`. Below the path, there are tabs for Overview, Details, and Permissions. The Overview tab is active, showing a table of files in the `clientes` directory. A red arrow points to the `clientes` directory link in the path.

Name	Size
 cli_01.csv	2.36 KB
 cli_02.csv	2.33 KB
 dados_clientes_raw.csv	1.71 MB

E clique em **Save**.

Schedules & Triggers



Trigger Status

- ☒ Active
☐ Paused

Trigger type

File arrival



This job does not have any failure notifications. Consider adding email or webhook notifications to alert if trigger evaluation fails.

File arrival triggers monitor cloud storage paths of up to 10,000 files for new files. These paths are either Unity Catalog volumes or external locations managed through Unity Catalog.

Storage location ⓘ

/Volumes/workshops_databricks/db_workshop_ana_sanchez_1b4e/landing/clientes/

Advanced



Test connection

Cancel

Save

- Agora, em Job notifications, clique em **Edit notifications** > **+Add notification**.



Job notifications ⓘ

No notifications

Edit notifications

- Preencha com seu email e selecione o modo **Start** para te avisar quando iniciar a execução do job.

Job notifications



Supported destinations: Email, Microsoft Teams, PagerDuty, Slack, Webhook

Destination	Start	Success	Failure	Duration warning	Streaming backlog	
ana.sanchez@databricks.com	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

☐ Mute notifications for skipped runs

☐ Mute notifications for canceled runs

+ Add notification

Cancel


Save

- Volte para o workspace, e dentro da pasta **Workflows** execute o notebook **02 - workflows_ingirir novos clientes**.
- Abra seu e-mail para ver a notificação chegando de que seu job foi iniciado.




[984752964297111] Azure Databricks run 307167668394861 of clientes_workflow_anasanchez has started

External

Inbox x

 Microsoft Azure <azure-noreply@microsoft.com> to me

Thu, Sep 5,

 Microsoft Azure

Azure Databricks

Your Azure Databricks job run has started









Run details

Workspace	field-eng-east (984752964297111)
Job	clientes_workflow_anasanchez (855599052474471)
Job run	307167668394861
Started at	2024-09-06 02:29:51 UTC
Launched	By file arrival trigger

View run in Databricks >

Go to the latest successful run

Cancel runs

Start time	Run ID	Launched	Duration	Status	Error code	Run parameters	
Sep 05, 2024, 11:29 PM	307167668394861	By file arrival	1m 39s	 Running			
Sep 05, 2024, 11:13 PM	473836482803619	Manually	3m 7s	 Succeeded			
Sep 05, 2024, 11:09 PM	60813499028222	Manually	3m 23s	 Succeeded			
Sep 05, 2024, 11:03 PM	673756622602899	Manually	3m 56s	 Succeeded			

03. Workflow - Trigger Mode: File Arrival

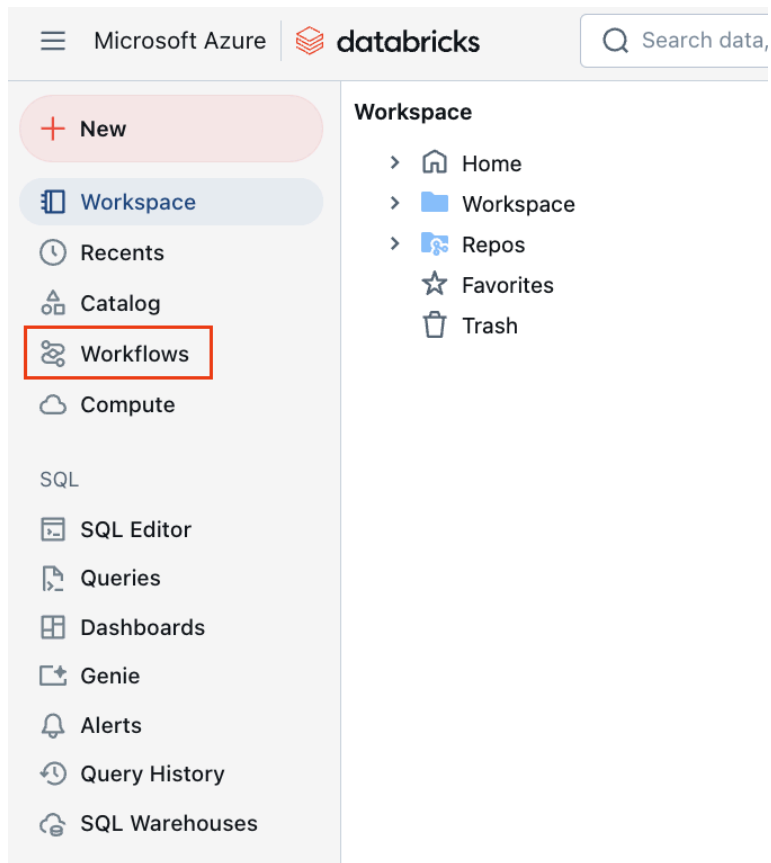
Nesse exercicios vamos orquestrar os notebooks criados anteriormente:

- 03.1 - Listar diretorios e tabelas para realizar ingestão
- 03.2 - Carregar Tabelas

Tudo o que estiver imprimindo (display) vocês podem comentar.

- Antes de criar o wokrflow, vamos adicionar alguns novos arquivos para criar várias tabelas de uma vez. E para isso, execute o seguinte notebook **workflows/03 - Carregar Dados para o Workflow em lote**

2. Clique em **Workflows**



3. Clique em **Create Job**

Microsoft Azure

databricks

Search data, notebooks, recents, and more...

field-eng-east

New

Workspace

Recents

Catalog

Workflows

Compute

SQL

SQL Editor

Queries

Dashboards

Genie

Alerts

Query History

SQL Warehouses

Data Engineering

Job Runs

Data Ingestion

Workflows

Jobs

Job runs

Pipelines

Filter jobs

Owned by me

Accessible by me

Favorites

Send feedback

Load tutorial

Create job

Name	Tags	Cre...	Trigger	Recent runs
Gmarket for/each condition		Ins		⊗ ⊖ ⊕ ⊕ ▶ ⋮
Starbucks_Workflow		Ins		— — — — ⊕ ▶ ⋮
gamer-lifecycle	CME solacc_testing	Dur		— — — — ⊖ ⊗ ▶ ⋮
sadasd		Ale	Paused - Scheduling	— — — — — ▶ ⋮
Simple Job	test test2	Dar		⊗ ⊗ ⊕ ⊕ ⊕ ▶ ⋮
(andres_urrutia) mfg_llm_demo		Anc		— — — — — ▶ ⋮
(andrew_kraemer) mfg_llm_demo		Anc		— — — — — ▶ ⋮
(dave_geyer) mfg_llm_demo	2024-12-30	Dav		▶ ⊖ ⊖ ⊖ ⊖ ▶ ⋮
(jai_karve) mfg_llm_demo		Jai		— — — — — ▶ ⋮
(paul_mracek) mfg_llm_demo		Pat		— — — — ⊗ ▶ ⋮

4. No topo abaixo de [Workflows](#) > [Jobs](#) > edite o nome do seu Workflow, exemplo: **workflow_tabelas_em_lote_SEUNOME**

[Workflows](#) > [Jobs](#) >

workflow_tabelas_em_lote_ana_sanchez ☆

Runs **Tasks**

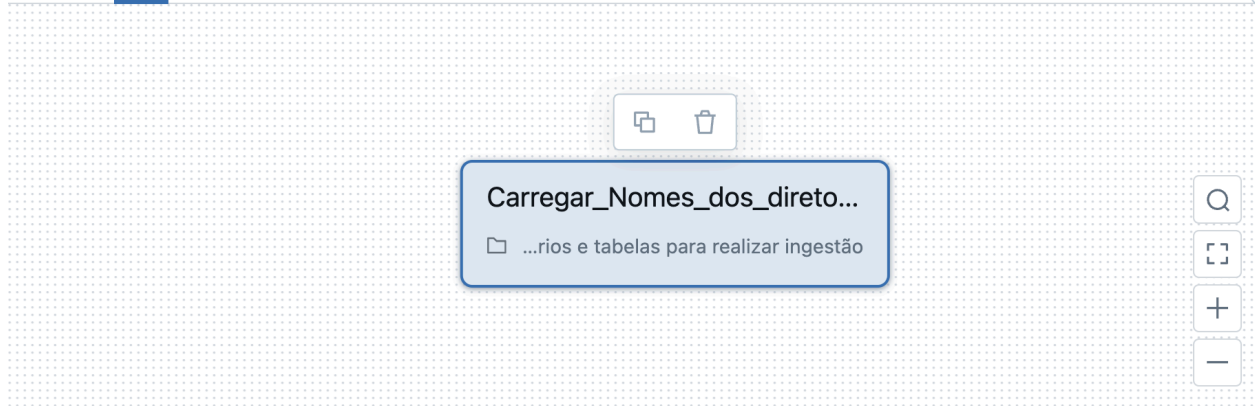
5. Preencha os seguintes campos dentro da Task e depois clique em **Add**:

- Task name: Carregar_Nomes_dos_diretorios
- Type: Notebook
- Path: Caminho para o notebook **03.1 - Listar diretorios e tabelas para realizar ingestão**

Workflows > Jobs >

workflow_tabelas_em_lote_ana_sanchez ☆

Runs **Tasks**



Task name* ⓘ	Carregar_Nomes_dos_diretorios
Type*	Notebook ▼
Source* ⓘ	Workspace ▼
Path* ⓘ	...04 - Workflows/03.1 - Listar diretorios e tabelas para realizar ingestão 🔗 ▼ 📄

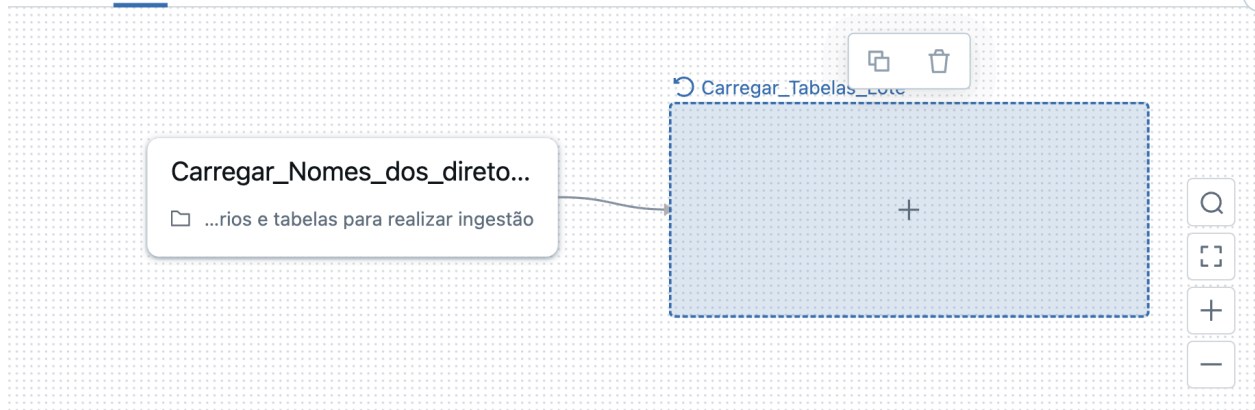
4. Agora, clique em **+ Add Task > For Each** e preencha os seguintes campos dentro da Task e depois clique em **Add a task to a loop over**:

- Task name: Carregar_Tabelas_Lote
- Type: For Each
- Inputs: `{{tasks.Carregar_Tabelas_Lote.values.tables}}`
- Concurrency: 5
- Depends on: Carregar_Nomes_dos_diretorios

Seu workflow ficará da seguinte forma:

workflow_tabelas_em_lote_ana_sanchez ☆

Runs Tasks



Task name* ⓘ	<input type="text" value="Carregar_Tabelas_Lote"/>
Type*	<div>For each ▾</div>
Inputs* ⓘ	<div>{{tasks.Carregar_Nomes_dos_diretorios.values.my_value}} {}</div>
Concurrency (optional) ⓘ	<input type="text" value="5"/>
Depends on	<div>Carregar_Nomes_dos_diretorios × ▾</div>
Run if dependencies ⓘ	<div>All succeeded ▾</div>
Notifications ⓘ	<div>+ Add</div>

Cancel

Add a task to loop over

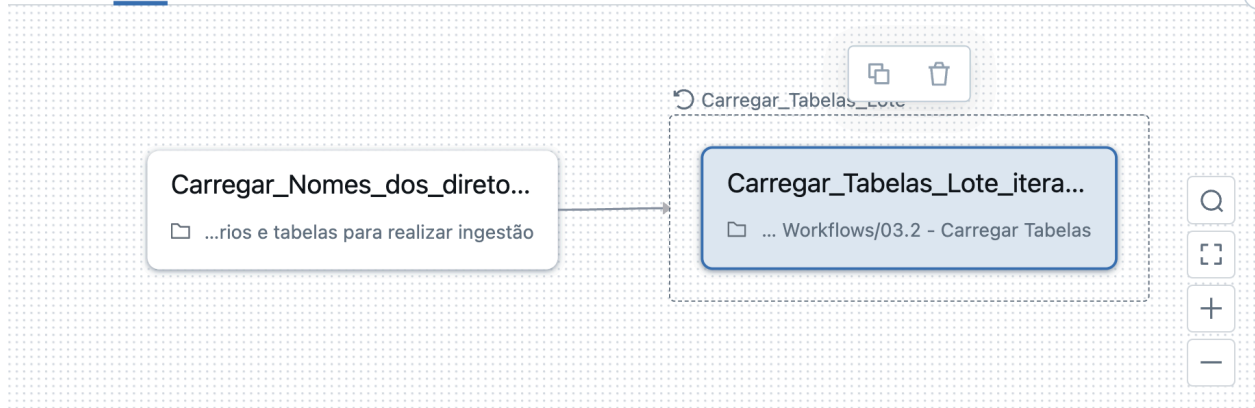
5. Agora, preencha a iteração e depois clique em **Create Taks**.

- Task name: Carregar_Tabelas_Lote_Iteration
- Type: Notebook
- Path: Caminho para o notebook **03.2 - Carregar Tabelas**
- Parameter:
 - table_name = {{input.table_name}}
 - volume_folder = {{input.volume_folder}}
 - volume_folder_files = {{input.volume_folder_files}}
 - tipo = {{input.tipo}}
 - catalogo = {{input.catalogo}}
 - schema = {{input.schema}}

Seu workflow ficará da seguinte forma:

workflow_tabelas_em_lote_ana_sanchez ☆

Runs Tasks



Task name* ⓘ	Carregar_Tabelas_Lote_iteration
Type*	Notebook
Source* ⓘ	Workspace
Path* ⓘ	... Engenharia de Dados/GUIDE/04 - Workflows/03.2 - Carregar Tabelas 🔗
Compute* ⓘ	Serverless
Dependent libraries ⓘ	+ Add
Parameters ⓘ	<div><div>table_name</div><div>{{input.table_name}}</div><div>{}</div><div>×</div></div>

Cancel

Create task

volume_folder	{{input.volume_folder}}	{}	×
volume_folder_files	{{input.volume_folder_files}}	{}	×
tipo	{{input.table_name}}	{}	×
catalogo	{{input.catalogo}}	{}	×
schema	{{input.schema}}	{}	×

6. No menu lateral, abaixo de **Job Details**, em **Tags** preencha:

- MODO : TREINAMENTO
- AMBIENTE: PD-MDL
- SIGLA: SLN

Tags

✕

Key* ⓘ

Value ⓘ

MODO

TREINAMENTO

✕

AMBIENTE

PD-MDL

✕

SIGLA

SLN

✕

Key

Value (optional)

7. Em seu Workflow e clique em **Run now**.
8. Clique em **Runs** para acompanhar a execução de seu workflow

Workflows > Jobs >

workflow_tabelas_em_lote_ana_sanchez ☆

RunsTasks

Runs

Start date📅

< Previous

Next >

Run total duration

23s

11s

Sep 11

Carregar_Nomes_dos_diretorios

Carregar_Tabelas_Lote

Cancel runs ▾

Start timeRun IDLaunchedDurationStatusError codeRun paramet...📄

Sep 11, 2024, 0...

280064550...

Manually

24s

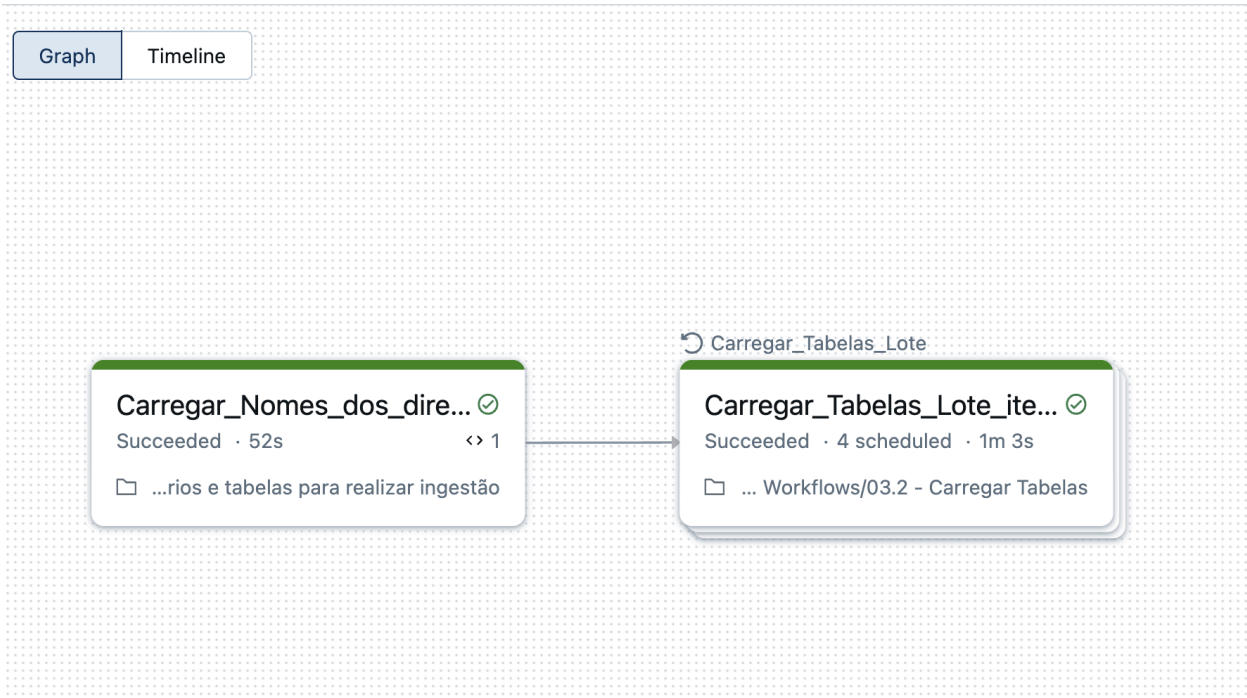
🔄 Runn...

■⋮

12. Para ver o detalhe de execução de cada etapa, clique no link do horário de execução e verifique se seu job executou com sucesso.

[Workflows](#) > [Jobs](#) > [workflow_tabelas_em_lote_ana_sanchez](#) >

workflow_tabelas_em_lote_ana_sanchez run



13. E clique em **Carregar_Tabelas_Lote_iteration** para verificar.