

Production Pipelines

© Rangel

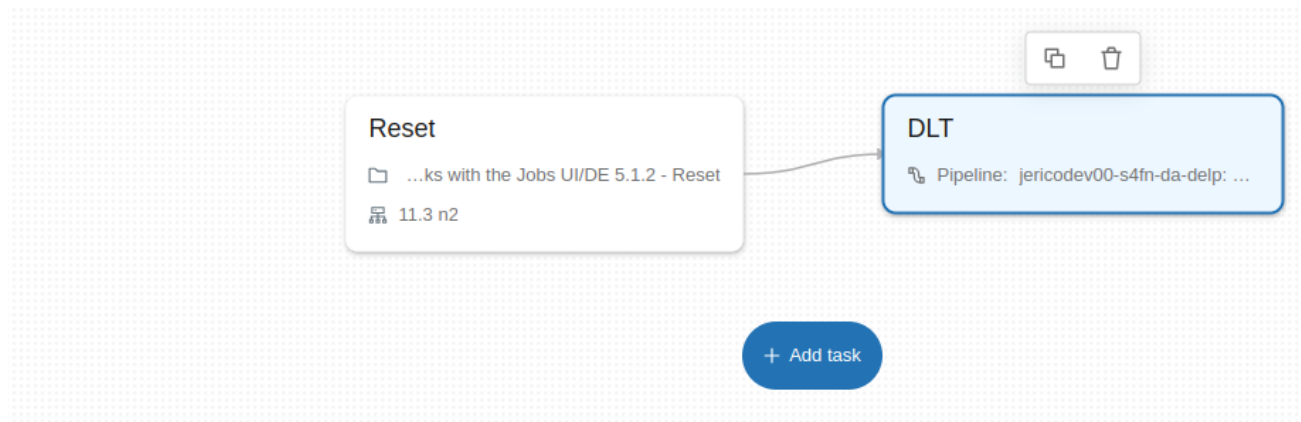
Workloads with Workflows

1. Identify benefits of using multiple tasks in Jobs.

- allows you to orchestrate processing/ingestion of data/tables
- allows isolation of each task and setting of dependencies between tasks

2. Set up a predecessor task in Jobs.

- aka dependencies
- use the **Depends on:** field to name a dependency.
- additionally can add custom run behavior using **Run if** field.



The screenshot shows a workflow in the Databricks Jobs UI. It consists of two tasks: 'Reset' and 'DLT'. The 'Reset' task is a folder task named '...ks with the Jobs UI/DE 5.1.2 - Reset' with version '11.3 n2'. The 'DLT' task is a Delta Live Tables pipeline named 'Pipeline: jericodev00-s4fn-da-delp: ...'. An arrow indicates a dependency from 'Reset' to 'DLT'. A '+ Add task' button is visible at the bottom of the workflow canvas.

Task name* ⓘ DLT

Type* Delta Live Tables pipeline ▼

Pipeline* ⓘ jericodev00-s4fn-da-delp: Pipeline Demo w/Job 🔗 ▼

☐ Trigger a full refresh on the Delta Live Tables pipeline

Depends on Reset X ▼

Run if ⓘ All succeeded ▼

Task name* ⓘ

Type*

Pipeline* ⓘ

Depends on

Run if ⓘ

All succeeded
All dependencies have executed and succeeded ✓

At least one succeeded
At least one dependency has succeeded

None failed
None of the dependencies failed and at least one was executed

All done
All dependencies have been completed

At least one failed
At least one dependency failed

All failed
All dependencies have failed

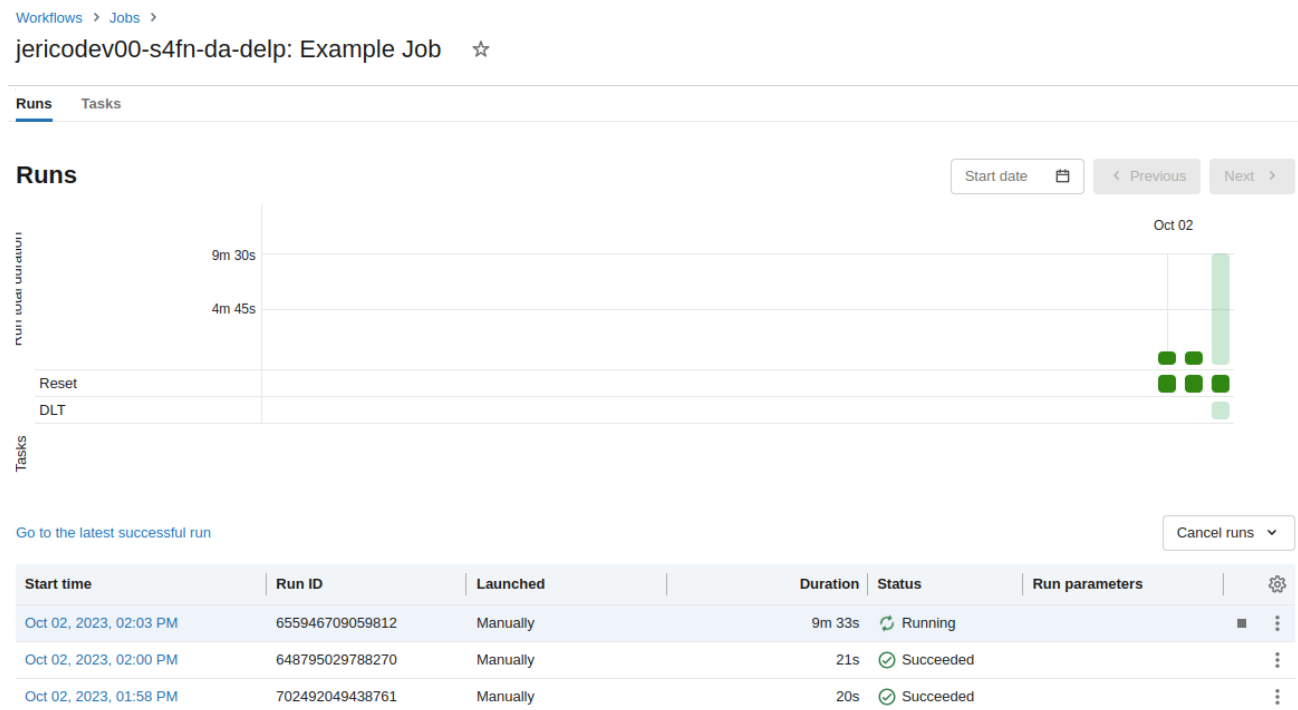
All succeeded

3. Identify a scenario in which a predecessor task should be set up.

- When you want to run another notebook first.

4. Review a task's execution history.


- Go to **Jobs** under **Workflows**



- For task specific details:







- Go to **Job runs** tab under **Workflows**
- Click on the task of interest

Reset run

Original (latest) ·  Succeeded




Task run details

Job ID	802136187570299 
Job run ID	655946709059812 
Task run ID	904331168101011 
Run as	 Jerico Dev
Launched	Manually
Started	10/02/2023, 02:03:40 PM
Ended	10/02/2023, 02:04:01 PM
Duration	20s
Queue duration 	-
Status	 Succeeded




Notebook

[/Users/jericodev00@gmail.com/data-engineer-learning-path-v1-0-2-notebooks/05 - Workflow Jobs/DE 5.1 - Scheduling Tasks with the Jobs UI/DE 5.1.2 - Reset](#) 



Compute

 11.3 n2

Driver: n2-standard-4 · Workers: n2-standard-4 · 0 workers · 11.3 LTS (includes Apache Spark 3.3.0, Scala 2.12)

[View details](#)

[Spark UI](#)

[Logs](#)

[Metrics](#)



Task values

No task values have been set for this run

5. Identify CRON as a scheduling opportunity.

- Under job details, we can schedule with CRON syntax

i

Job details

Job ID

802136187570299

Creator

Jerico Dev

Run as

i

Jerico Dev

Tags

i

+Tag

Git

Not configured

Add Git settings

Schedule

Paused - At 02:00 PM (UTC+00:00 — UTC)

Edit schedule

Resume

Delete

Schedule

Trigger Status

Active

Paused

Trigger type

Scheduled

Schedule

Every

Day

at

14

:

00

(UTC+00:00) UTC

Show cron syntax

Cancel

Save

6. Debug a failed task.
- Could not run due to GCP quota issues, but you can re-run the DAG at the task which failed, after debugging.

7. Set up a retry policy in case of failure.

Retry Policy

Jobs that fail are retried a number of times based on the following policy. You can specify a maximum number of attempts for a run and a minimal interval between attempts.

Retry at most 1 time (2 total attempts) and wait 15 mins between retries

☐ Retry on timeout

No retries

1 time (2 total attempts) ✓

2 times (3 total attempts)

3 times (4 total attempts)

4 times (5 total attempts)

5 times (6 total attempts)

6 times (7 total attempts)

7 times (8 total attempts)

Cancel

Confirm

8. Create an alert in the case of a failed task.

9. Identify that an alert can be sent via email.

Emails ⓘ

jericodev00@gmail.com

☐ Start

☐ Success

☒ Failure

☐ Duration warning

+ Add

☐ Mute notifications for skipped runs

☐ Mute notifications for canceled runs

☐ Mute notifications until the last retry

Retries ⓘ

15 min delay, at most 1x (2 total attempts) ✎

Additional Notes:

- You can share data between tasks using **task values**. [1] This is similar to XCOMs in Airflow. You can set task values inside the notebook.
- Creating an alert for a Job Task is separate from the **Alerts** tab on the left, which only alerts for queries.