

Unified orchestration using data bricks

Let's explore unified orchestration with data bricks, focusing on DLT, for ETL processing pipelines and workflows for end-to-end platform wide orchestration.

Databricks has 2 main task orchestration services

- **DLT:** is a declarative ETL framework to help the team to simplify streaming and batch ETL cost-effectively. It revolutionizes ETL and real time analytics by **automating** and **scaling** streaming ingestion and transformation, providing powerful pipeline. It simplifies ETL process and allowing data engineers to focus on data transformation not operations. Define a transformation to perform on the data and let DLT do the rest.
- **Workflows:** is a fully managed cloud based general purpose task orchestration service for the entire platform. It removes operational overhead with a fully managed orchestration service, enabling you to focus on the workflows and not on managing infrastructure. Workflow is a service for data engineers, data scientists and analysts to build reliable data analytics and AI workflows using SQL notebooks and ml models. Data teams can orchestrate any combination of tasks, create, run, monitor and repair data pipelines without managing any infrastructure. It is easy to use point and click authoring for all data teams. It is important to note the DLT pipeline is a task in a workflow.

When to use a workflow vs Delta live table?

If you are orchestrating a series of dependent jobs running on schedule, contain several dependent tasks you would use workflow, making ML tasks or multiple MLflows notebooks on a schedule. If you have a series of arbitrary codes, external API calls and custom tasks such as a job which contain a jar file spark submit python script SQL tasks you would schedule it as a workflow.

But if you are building a data ingestion and transformation pipeline such as ETL jobs for batch or streaming, building data quality constraints, monitoring and logging you would use DLT.

For more complex end to end orchestration, you could add DLT pipeline as a task in a workflow and add additional task to perform other end to end orchestration.

Databricks workflows provides a collection of tools that allow you to schedule and orchestrate data processing tasks on data bricks. You see data bricks workflows to configure data bricks jobs.

Objective: use the pipeline built in the previous demonstration to create 2 tasks in a job. The pipeline can be separated into 2 notebooks for example purpose:

- Bronze
- Gold

You could have used a delta live tables pipeline for this engineering task, since DLTs can be scheduled within a workflow with additional tasks.

