Orchestrating Data Pipelines

Learning Objectives

- Introduce the meaning of orchestration
- Become familiar with Delta Live Tables
- Acknowledge Databricks Change Data Capture
- Introduce jobs in Databricks

Orchestrating data pipelines

- Execute data pipelines programmatically
- Define data pipelines scripts executions order
- Monitor data quality and establish flow control

Orchestrating data pipelines

- Delta Live Tables
- Change Data Capture
- Databricks Jobs

Delta Live Tables

- Build and manage data pipelines via a declarative approach
- Apply software engineering best practices to data orchestration
- Apply testing, error handling and monitoring to data pipelines
- Use both Python and SQL

Delta Live Tables

Data Pipelines Production Issues

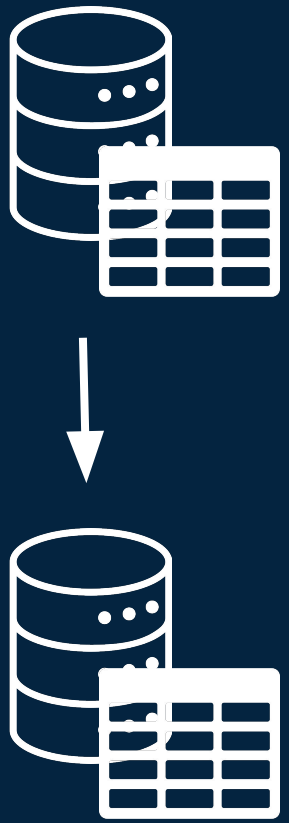
- Complex dependencies
- Batch and stream processing
- Poor data quality
- Opaque data lineage
- Difficult error handling

Delta Live Tables Solutions

- UI for dependencies
- Declarative approach
- Data quality controls
- UI for data lineage
- Declarative approach

Change Data Capture

- Spot changes applied to data in a data source
- These changes can be Insert, Update, or Delete operations
- Deliver those changes to a target storage object
- With specific metadata attributes



Change Data Capture

Data Source

artist_id	artist_name	artist_surname	operation_date	operation
1	Placido	Domingo	27-03-2023	DELETE
2	Andrea	Bocelli	28-03-2023	INSERT
2	Luciano	Pavarotti	29-03-2023	UPDATE

Target Storage Object

artist_id	artist_name	artist_surname
1	Placido	Domingo

Change Data Capture

Target Storage Object

artist_id	artist_name	artist_surname
2	Luciano	Pavarotti

Change Data Capture

Advantages

- Upsert is the default behavior
- Deletes handling is included
- Handles composite primary keys
- Supports SCD Type 1 (default) and SCD Type 2

Disadvantages

- The target table records are updated and deleted, not only appended
- The target table cannot be used as a streaming data source

Jobs

- Run non interactive code
- Exploit Databricks job clusters
- Execute notebooks and scripts programmatically
- Execute Delta Live Tables workflows