Lab 8 – Mini-project

Background

In lab 3 you ingested several datasets from the AdventureWorks GitHub repository “as-is” – the files you copied to the data lake were tab-separated and contained no column headers or types. A result of this was that in lab 4 you had to add column metadata to the data flow yourself, in a way that cannot be reused elsewhere.

A better solution would be to convert the files into Parquet or Delta format, persisting metadata permanently in each dataset while benefiting from more space-efficient storage. Doing this is the purpose of this mini-project.

The requirement

Your task is to:

- Extract ten specific AdventureWorks datasets from the GitHub repository.
- Store each dataset in the data lake, in Parquet or Delta format, including real column names and data types.
- If ingesting a dataset fails, send an appropriate notification.

The ten datasets are identified in a file called **TableCatalog.json**, available online at https://tinyurl.com/38334wsc.

What you need to do

- Produce an outline design sketch.
- Implement as much of your design as you can.
- Be prepared to talk about your design and the pipeline(s) you’ve built – if you haven’t had time to get everything working, you’ll be able to use the design sketch to explain where you were going.