Lab 7 – 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 header information explicitly – and that information cannot be re-used.

A better solution would be to convert the files into Parquet format, to add column information to each dataset permanently while benefiting from more space-efficient storage. This is the foundation 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 format, including real column names and data types.
- If ingesting a dataset fails, send an appropriate notification email.

Resources you can use

- The ten datasets are identified in a file called **TableCatalog.json**, available online at https://bit.ly/3TrPRfd.
- ADF or Synapse pipelines do not natively support sending emails – you can do so by calling an Azure Function we have prepared for you:

Function app URL	https://frameworksupportfunctions.azurewebsites.net
Function name	SendEmail
HTTP method	POST

Example body	`{`
	`  "emailRecipients": "me@mydomain.com ",`
	`  "emailSubject": "Error!",`
	`  "emailBody": "Something went wrong"`
	`}`

  The function uses **anonymous** HTTP authentication and **function key** authorisation – the function key will be provided to you directly.

What you need to do

- Produce an outline design sketch.
- Implement as much of your design as you can.
- Be prepared to present 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.