Lab 4 - Datasets

Requirements

In order to start the lab, it is important that Lab3 has been completed.

Objective

Now that the Linked Services have been created, ADF can access specific data such as a table in a database, a .csv file on a storage account, and more. To specify what you want, you need to create a Dataset. This is what we will do in the tasks below.

Task 1 - Source Database

The first *dataset* we connect is a table that lives within our source database.

1. Click on the **Pencil** (Author) on the left. On the left, you see a list of categories such as: Pipelines, Datasets, Data flows, and Power Query.
   Today, we focus on **Pipelines** and **Datasets**.
2. Next to **Datasets**, you currently see a 0. When you hover over the **Datasets** box with your mouse, you see an option with **3 dots** (Datasets Actions) appear on the right. Click on **Dataset Actions** and then click on **New Dataset**.
3. A screen similar to the **Linked Services** will appear. Search for **SQL**. Double click on **Azure SQL Databases**.
4. Give the Dataset a clear name. The recommended format is to start with `DS_`, the type of dataset, possibly the *schema* within which the table is located, the table name, and ending with _environment.
   - Practical example: `DS_sql_dwh_dimdatum_acc`
   - Training example: `DS_asql_SalesLT_Address_training`
5. At **Linked Services**, you choose the Linked Service that refers to the source database (`LS_sqldb_source`).
6. The IR is automatically applied from the Linked Service. The option to select a **Table name** should now also have appeared, click on it and choose **SalesLT.Address**. Complete the creation by clicking **OK** at the bottom of the page.
7. Once the **Dataset** has been created, you will enter the dataset overview screen. Click on the magnifying glass (**Preview Data**) to preview the data.
8. Click on the **Schema** tab. Here you see the columns from the selected table and the corresponding datatypes.
9. Repeat Task 1, but this time for the **sqldb-target** Database for the tables **Address**, **ProductCategoryDiscount**, and **SalesPersonal**.

Task 2 - Storage Account / File system

1. Click on the **Dataset Actions** and then click on **New Dataset**.
2. Search for **storage**. Click on **Azure Blob Storage**.
3. Choose **DelimitedText** (csv).

> ### Which file format

> You will see several common file formats here:
>
> - Excel
> - Json
> - XML
> - DelimitedText (csv)
>
> For Cloud Data Platforms, the **Parquet** format is also often used. Parquet is very compact in storage, optimized for analyses (Column-based instead of Row-based) and contains datatypes (unlike CSV files, where commas, dots, list separators, string delimiters, and date notations often lead to confusion - not to mention encoding).
>
> For now, we will use CSV here - a major advantage of it for now is that it is human-readable, so you can see what is happening.

4. Give the Dataset a clear name.
5. At **Linked Services** choose the **storage account**.
6. The option to specify a path will appear. Click on the white folder (**Browse**). Then choose the **data** folder and the file named **ProductCategoryDiscount.csv**.
7. Click **OK** and then again **OK** to complete the Dataset.
8. Click on **Preview data**, you will see that the data does not yet look very cool. To adjust this, we still need to make 2 changes.
9. Choose the **Semicolon (😉** options for **Column delimiter**. and check **First row as header**. When you now click on **Preview data** again, it should be in a table with columns.
10. Repeat Task 2, but now choose the **File system** connector and choose the .csv file named **SalesPersonal.csv**.
11. Click on the **Blue button** with the text **Publish all** and then on the **Publish** button.

Table of Contents