Lab 4 - Datasets

Vereisten

Om het lab te kunnen starten is het van belang dat Lab2 is afgerond.

Doel

Nu de Linked Services aangemaakt zijn kan ADF bij specifieke data zoals een tabel in een database, een .csv bestand op een storage account en meer. Om te specificeren wat je wilt hebben dien je een Dataset aan te maken. Dit gaan we in onderstaande opdrachten doen.

Opdracht 1 - Source Database

De eerste *dataset* die we aankoppelen is een tabel die binnen onze brondatabase leeft.

1. Klik links op het **Potloodje** (Author). Aan de linkerkant zie je een lijst met categorien zoals: Pipelines, Datasets, Data flows en Power Query.
   Vandaag leggen we de focus op **Pipelines** en **Datasets**.
2. Naast **Datasets** zie je op dit moment een 0 staan, wanneer je met jouw muis op het vak van **Datasets** gaat staan zie je een optie met **3 bolletjes** (Datasets Actions) verschijnen aan de rechterkant. Klik de **Dataset Actions** aan en klik vervolgens op **New Dataset**.
3. Een vergelijkbaar scherm als bij de **Linked Services** zal verschijnen. Zoek naar **SQL**. Dubbelklik de **Azure SQL Databases** aan.
4. Geef de Dataset een duidelijke naam. Het aangeraden format is om te beginnen met `DS_`, het type dataset, eventueel het *schema* waarbinnen de tabel zich bevindt, de tabelnaam en eindigend met _omgeving.
   - Praktijkvoorbeeld: `DS_sql_dwh_dimdatum_acc`
   - Trainingsvoorbeeld: `DS_asql_SalesLT_Address_training`
5. Bij **Linked Services** kies je de Linked Service die verwijst naar de brondatabase (`LS_sqldb_source`).
6. De IR wordt automatisch toegepast vanuit de Linked Service. De optie om een **Table name** te selecteren zal nu ook verschenen zijn, klik hierop en kies voor **SalesLT.Address**. Voltooi het aanmaken door onderaan de pagina op **OK** te klikken.
7. Wanneer de **Dataset** is aangemaakt kom je in het overzichtscherm van de dataset. Klik op het brilletje (**Preview Data**) om een voorbeeld van de data te zien.
8. Klik op de tab **Schema**. Je ziet hier de kolommen uit de geselecteerde tabel en de bijhorende datatypes.
9. Doe Opdracht 1 nogmaals, maar nu voor de **sqldb-target** Database voor de tabbellen **Address**, **ProductCategoryDiscount** en **SalesPersonal**.

Opdracht 2 - Storage Account / File system

1. Klik de **Dataset Actions** aan en klik vervolgens op **New Dataset**.
2. Zoek naar **storage**. Klik de **Azure Blob Storage** aan.
3. Kies voor **DelimitedText** (csv).

> ### Welk bestandsformaat

> Je ziet hier een aantal veelvoorkomende bestandsformaten:
>
> - Excel
> - Json
> - XML
> - DelimitedText (csv)
>
> Voor Cloud Dataplatforms wordt daarnaast het **Parquet**-formaat veel gebruikt. Parquet is zeer compact in de opslag, geoptimaliseerd voor analyses (Column-based is i.p.v. Row-based) en bevat datatypes (in tegenstelling tot CSV-bestanden, waar komma's, punten, lijstscheidingstekens, string delimiters en datumnotaties nogal eens tot verwarring leiden - om maar niet te spreken over encoding).
>
> Voor nu gebruiken we hier even CSV - groot voordeel daarvan voor nu is dat het door mensen leesbaar is, zodat je kunt inzien wat er gebeurt.

4. Geef de Dataset een duidelijke naam.
5. Bij **Linked Services** kies het **storage account**.
6. De optie om een pad op te geven zal verschijnen. Klik op het witte mapje (**Browse**). Kies vervolgens de map **data** en het bestand genaamd **ProductCategoryDiscount.csv**.
7. Klik op **OK** en vervolgens nog een keer op **OK** om de Dataset te voltooien.
8. Klik op **Preview data**, je zult zien dat de data er nog niet erg gaaf uitziet. Om dit aan te passen dienen we nog 2 aanpassingen te verrichten.
9. Kies bij **Column delimiter** voor de opties **Semicolon (**😉. en vink aan **First row as header**. Wanneer je nu weer op **Preview data** klikt zou het in een tabel moeten zijn met kolommen.
10. Doe Opdracht 2 nogmaals, kies nu het .csv bestand **SalesPersonal.csv**.
11. Klik op de **Blauwe knop** met de tekst **Publish all** en vervolgens op de knop **Publish**.

Inhoudsopgave