Lab4 - Datasets

Vereisten

Om de lab te kunnen starten is het van belang dat Lab3 is afgerond.

Doel

Nu de Linked Services aangemaakt zijn kunnen wij 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. Volg de opdrachten stap voor stap.

Opdracht 1 - Source Database

1. Klik links op het **Potloodje** (Author). Aan de linkerkant zie je een lijst met categorien zoals: Pipelines, Datasets, Data flows en Power Query. De laatste vallen buiten beschouwing voor deze training.

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** (Dataset 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 typen dataset, (schema) ,de tabel/bestands naam en eindigend met _omgeving.

   Praktijkvoorbeeld: DS_sql_dwh_dimdatum_acc

   Trainingsvoorbeeld: DS_asql_SalesLT_Address_training

5. Bij **Linked Services** kies de **sqldb_source** database.

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 **Address** en 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. Een aantal veelvoorkomende bestands formaten zullen voorbijkomen zoals Excel, Json, XML en DelimitedText (csv). Maar ook enkele Big Data formaten welke ieder opzich zelf weer voordelen met zich mee brengen. Het **Parquet** is de populairste op dit moment. Dit komt door het feit dat een Parquet bestand zeer klein is door de compressie (~1/10 van een .csv) en Column-based is i.p.v. Row-based zoals een .csv dat is. Dit houd in dat je een enkele kolom uit het bestand kan kiezen voor het laden i.p.v. allemaal zoals bij Row-based. Voorals nog kiezen we de optie **DelimitedText**.

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 **blauwe 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, maar kies nu voor de **File system** connector en kies het .csv bestand genaamd **SalesPersonal.csv**.

11. Klik op de **Blauze knop** met de tekst **Publish all** en vervolgens op de knop **Publish**.

Einde Lab4