Analytics in Azure – week 5

In de vijfde week van het leertraject Analytics in Azure gaan we aan de slag met streaming data.

Omdat de week tussen week 5 en week 6 ook de laatste week van het leertraject is, is dit een soort “inhaalweek”. Er zijn daarom geen verplichte opdrachten. Wel zijn er (uiteraard) verdiepende opdrachten, voor als je nog wat meer uit streaming data wilt halen.

Het leuke aan streaming data is, dat het vaak data is die je overal vandaan kunt halen: zonnepanelen (SolarEdge), slimme thermostaten, sensors rond het huis: bijna allemaal hebben ze API’s en daarmee data die (semi)-streaming aangeleverd kan worden.

# Verdieping / verbreding

## Twitter data – revisited

Diverse mensen hebben hun tanden al stukgebeten op de Twitter Data. Belangrijkste probleem waar we in deze casus tegenaan liepen was het ontbreken van rechten op de Streaming API.

Het is echter wel een heel interessante casus om verder mee te gaan; daarom deel ik hierbij mijn eigen credentials voor het lezen van tweets.

Wanneer de basis (werkend) staat, kun je hier redelijk wat aan verdiepen en verbreden:

* De oorspronkelijke opdracht was te vinden op [analytics-in-azure/analyze-twitter-data-in-powerbi.md at main · vstrien/analytics-in-azure (github.com)](https://github.com/vstrien/analytics-in-azure/blob/main/week-5/analyze-twitter-data-in-powerbi.md)
* Je kunt de data verrijken met Cognitive Services, [bijvoorbeeld door het sentiment van tweets te analyseren en toe te voegen](https://towardsdatascience.com/using-azure-cognitive-services-for-sentiment-analysis-of-trumps-tweets-part-1-f42d68c7e40a)
* Je kunt de code in plaats van op DataBricks op je eigen laptop draaien. Dan heb je wel een beetje voorkennis van Python en Pip (modules en installatie daarvan voor Python) nodig. Dat kan sowieso geen kwaad als Data Engineer.
* Wanneer je het op je eigen laptop werkend hebt, kun je op zoek naar een voordelige PaaS-oplossing. Bijvoorbeeld [in Azure Functions](https://docs.microsoft.com/nl-nl/azure/azure-functions/create-first-function-vs-code-python).
* Eén stap verder is om het vervolgens in DevOps te kunnen uitrollen. Zorg dat je je requirements.txt op orde hebt met de benodigde libraries!
* Oja, als je dan toch met een PaaS-oplossing bezig bent, kijk dan ook even naar security. Het is niet zo'n goed idee om Twitter App-credentials in de code te hebben staan. Je kunt hier in Azure Functions bijvoorbeeld de combinatie gebruiken van [Azure Functions Environment Variables met Azure KeyVault](https://levelup.gitconnected.com/a-secure-way-to-use-credentials-and-secrets-in-azure-functions-7ec91813c807).

## Frauduleuze telefoontjes analyseren

In de volgende tutorial word je meegenomen hoe je (met behulp van een data-generator) analyses kunt doen op SIM-kaart fraude. Mét een landing in Power BI [Tutorial - Analyze fraudulent call data with Azure Stream Analytics and visualize results in Power BI dashboard | Microsoft Docs](https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-real-time-fraud-detection)

## Achtergrond: Lambda-architectuur in het Microsoft-platofrm

Blue Granite gaat nog wat dieper in op de [Lambda Architecture in Azure (bluegranite.com)](https://www.bluegranite.com/blog/exploring-the-lambda-architecture-in-azure). Niet alleen wordt de hele architectuur uitgebreider besproken. Het is een artikel van ruim twee jaar geleden, toen Synapse nog niet bestond (Synapse SQL Pools was er al wel onder de naam SQL Data Warehouse). Ook Power BI was nog een stuk beperkter.

## Power BI als serving layer

Wanneer je Power BI als serving layer gebruikt in een lambda-architectuur, wil je “live connections” combineren met datasets die je via een refresh inlaadt. Binnen Power BI kan dit met behulp van *Composite Models*. Let wel: dit gaat dus niet over streaming data, maar over een live query waarvoor je de data in je Power BI rapport niet hoeft te verversen. [Use composite models in Power BI Desktop - Power BI | Microsoft Docs](https://docs.microsoft.com/en-us/power-bi/transform-model/desktop-composite-models)

## Power BI streaming sets

Om in Power BI écht streaming data te gebruiken, zijn er drie mogelijkheden:

* Push dataset
* Streaming dataset
* PubNub dataset

Meer informatie over streaming data in Power BI vind je hier: [Real-time streaming in Power BI - Power BI | Microsoft Docs](https://docs.microsoft.com/en-us/power-bi/connect-data/service-real-time-streaming)

Dit gaat echter altijd over streaming data*sets*: de dataset zelf is dus streaming. Power BI kent ook zogenaamde streaming data*flows*. Hierbij neemt Power BI feitelijk een stukje event processing over van (bijvoorbeeld) Stream Analytics. Streaming Dataflows zijn sinds december 2021 in preview. Patrick LeBlanc (Guy in a Cube) geeft er in deze video een kort overzicht van: [Exploring Streaming Dataflows in Power BI - YouTube](https://www.youtube.com/watch?v=ILu0xh3fHvQ). Ook in de Microsoft-documentatie is hier een goed overzicht van te vinden, dat helpt om deze nieuwe optie te plaatsen: [Streaming dataflows (preview) - Power BI | Microsoft Docs](https://docs.microsoft.com/en-us/power-bi/transform-model/dataflows/dataflows-streaming)