Azure data Factory

Sigma Data Academy 2021

Waarom ELT

Tegenwoordig kunnen we ook al een hoop data ophelen in PowerBI. Waarom zouden we dit dat ook nog met data factory moeten doen?

> Waarom zouden we dit allemaal doen?

- Minimalen belasting op bron (operationel system)

ELT biedt het voordeel dat we bronnen minimaal **belasten** voor BI. Als we 5 BI reports hebben dat allemaal **dezelfde** bron onsluiten dan wordt dat voor een bron systeem extra belastend. Als de dan ook nog een een veel voudt aan gebruikers zijn dan wordt het voor het bron stysteem te belasten.d

- Performance

Door dat we met ADF een keer de bron belasten en daarna zo klaar zetten voor front-end rapporates en dashboard, worden deze ook vele malen sneller.

- Organisatie breed

Door business rulles op een plek uit te voeren hoevan we dat in de front-end niet meer

Azure Dafa Factory (ADF)

- Interface
  - > **URL**: https://adf.azure.com/

Demo laten zien hoe Azure data facotry er uit ziet!

- Documentatie
  - > **Docs**: https://docs.microsoft.com/en-us/azure/data-factory

> - Git_repository
>   - *(ADF kan tegenwoordig de code opslaan in een git repository, daarvoor > hebben we een repository nodig in Gibhub of Azure Devops. Maar er een aan > voor deze training.)*

SQL configuratie

Set je zelf even als AD Admin voor je SQL Server. zo kun je makkelijke inloggen met management studio en hoe je niet de sa user te gebruiken.

Ook even firewall settings aanpassen. client ip toevoegenen Allow All Auzre Services.

KeyVault Access policies

Voeg je eigen account (emailadres) even toe aan de access policies. Daarmee kun je de secrets uitlezen van keyvault.

Setup Srouce control

Open je Azure data factory en gaan naar Manage.
Onder het kopje Sources control vind je Git configuratie. Klik op configure en configureer koppeling met de repository dat je hiervoor gemaakt hebt.

https://docs.microsoft.com/en-us/azure/data-factory/concepts-linked-services

Een Linked Services is als het waren je verbinding met een resources. Een aantal voorbeelde van resources: Bron systemen, KeyVault, Databases, webpages, api, datawarehouse.

- Type verbinding
- Authentication
- Authorization

We hebben een aadige keuze aan type verbindingen. voorbeeld zijn http, ftp, blob, sql, keyvault. Het is hierbij belangrijk om te weten of je ook toegang hebt om de betreffende verbinding op te kunnen zetten. Bij SQL Server gebruiken we in deze training bijvoorbeeld het SA Account, in de praktijk zul je een specifieke account gebruiken ter behoeven van ADF. Als je eenmaal verbinding kunt maken is het ook ban belang dat je ook daat werkelijk dat geen mag uitlezen wat nodig is voor je Data platform onsluiting.

Het aanmaken van dee linked services kan via het aanmaken van een data set. Dat is goed mogelijk, maar in een project zullen lindedservices vaak een vast structuur

Setup Links Services

> **Opdracht**: Maken van een Linked Services naar:
>
> - StorageAccount Containers stg & dwh
> - Key Vault
> - Azure SQL Database awlt & dwh

data movement activities, data transformation activities and control activities.

Data moment kan een copy data activiteit hebben dat een simple kopie maakt van een een bron naar doel.

Data transformation kunnen het opschonen zijn van een data set. Hiervoor is wel processing powernodig dus deze acties zullen meer kosten mee brengen.

Control heeft je de mogelijkheid om volgoordelijkheid te bepalen, afhankelijkheden zichtbaar te maken en geeft je de mogelijkheid om een goed Orchestratie van je complete data flow.

Vanuit de interface kunnen van elke pipeline de onderliggende json ophalen. Deze json file wordt ook weg geschreven naar je git repository dat je eerder hebt gekoppeld.

Een activiteit kan de volgende statussen hebben.

- succeeded
- failed
- completed

Data Set

```
* > **Docs**: https://docs.microsoft.com/en-us/azure/data-factory/concepts-datasets-linked-services
```

- Wat is en een data set
  Een dataset is een benomende view dat verwijsed of refereerd naar de data wat gebruikt kan worden als input of output van een activiteit.
- Kolommen & Data types
  Met een dataset definieer je ook wat voor bestandstype het betreft. Ook kan er al op dit moment een Definitie gemaakt wordt hoe een data set uit ziet. Welke namen hebben columns en welke data type betrefd het. Dit is echter geen verplichting dit kan namelijk ook later in een data flow worden opgelost.
- Maken van een Data set
  [sql_AWLT_Customer] op basis van de tabel customer uit de database AWLT
- Parameters

Opdrachten

> Bron > Datalake (CopyDate)

Integration runtime

- Waar draait het dan
  - Best effort voor Locatie
- Drie spaken
  - Azure
  - Self-hosted
  - Azure-SSIS

https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

Execution & Triggers

- Onces
- Schedule trigger
- Tumbling window trigger
- Event-based trigger

> **Docs**: https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

Monitoring/ Logging

Recap Module 1

- Wat is er allemaal besproken.
- Zijn er nog vragen?