Lab 5 - Pipelines

Vereisten

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

Doel

Het wordt nu tijd dat we data gaan verplaatsen van punt A naar punt B. Dit doen we door een pipeline aan te maken met een copy activiteit. De activiteit zorgt ervoor dat er een Source en Sink (Destination / Target) aan elkaar gekoppeld kunnen worden en dat er letterlijk een pump & dump plaats kan vinden. Volg de opdrachten stap voor stap.

Opdracht 1 - Database pipelines

1. Naast **Pipeline** zie je op dit moment een 0 staan. Wanneer je met je muis op het vak van **Pipeline** gaat staan zie je een optie met **3 bolletjes** (Pipeline Actions) verschijnen aan de rechterkant. Klik de **Pipeline Actions** aan en klik vervolgens op **New Pipeline**.

2. Geef de Pipeline een duidelijke naam. Het aangeraden format is om te beginnen met `PL_`, het soort activiteit, (schema), de tabel/bestands naam, bron (source), doel (sink) en eindigend met `_omgeving`. Heb je een pipeline die meerdere pipelines orchistreerd kan je het format globaler houden.

    - Praktijkvoorbeeld: `PL_copy_visits_clubmanager_to_datalake_prd`
    - Trainingsvoorbeeld: `PL_copy_address_training`

3. Aan de linkerkant zien we een lijst met de categorieen van de **Activities**. Klik op **Move & transform**. 2 opties zullen verschijnen **Copy data** en **Data flow**. De Data flow houden we buiten beschouwing voor vandaag. Klik en sleep de **Copy data** naar het canvas in het midden van het scherm.

4. Geef de Activiteit een duidelijke naam.

5. Klik op de tab **Source**. Er wordt gevraagd om een **Source dataset** op te geven. Klik deze aan en kies de Dataset voor **Address** vanuit de **sqldb-source**.

6. Klik op de tab **Sink**. Er wordt gevraagd om een **Sink dataset** op te geven. Klik deze aan en kies de Dataset voor **Address** vanuit de **sqldb-target**.

7. Verschillende opties zullen verschijnen, waaronder ook de optie voor een **Pre-copy script**. Hier kan je SQL-code uitvoeren voordat de Copy activiteit data gaat verplaatsen. Gezien we de pipeline meerdere keren willen kunnen draaien zonder dubbele data te krijgen kan je hier het volgende invullen/ plakken:

    ```
    Truncate table [Stg].[Address]
    ```

8. Klik op de tab **Mapping**. Je zult een knop zien met **Import schemas**, klik hierop.

    > Weet je nog dat bij Datasets de kolommen en datatypes kunnen komen te staan? Door dit proces te draaien worden de kolommen die matchen aan elkaar gekoppeld. Hiermee weet je

> zeker dat de data in de juiste kolom terecht komt. Dit is handig voor een tabel waarbij er een 1 op 1 mapping is. Doe je meerdere tabellen tegelijk dan zijn er andere opties.

9. Doe stap 1 t/m 8 opnieuw maar nu ook voor **ProductCategoryDiscount** en **SalesPersonal**. Hiervoor kan je de volgende **Pre-copy scripts** gebruiken:
    - `Truncate table [Stg].[ProductCategoryDiscount]`
    - `Truncate table [Stg].[SalesPersonal]`
10. Wanneer alle 3 de pipelines zijn aangemaakt. Maak een nieuwe pipeline aan genaamd: `PL_copy_Master_Training`.
11. Onder de tab van **Activities** is er een optie genaamd **General**, welke een **Execute Pipeline** activiteit bevat. Sleep er 3 naar het canvas.
12. Hernoem elke pipeline 1 voor 1 naar de 3 pipelines die je hiervoor hebt aangemaakt voor **Address**, **SalesPersonal** en **ProductCategoryDiscount**. Mocht de naam te lang zijn voor wat mag, maak hem voor nu wat korter.
13. Ga per pipeline naar de tab **Settings** en kies de bijbehorende pipeline. Doe dit voor alle 3 de pipelines.
14. Op dit moment zouden alle 3 de pipelines parallel lopen, wat makkelijk zou moeten kunnen gezien er geen afhankelijkheid van elkaar is. Ondanks dat gaan we ze sequentieel maken. Klik op 1 van de 3 pipelines. Je zult aan de rechterzijde van het blokje van de pipeline **vier vierkantjes met symbolen** zien. Klik hier op en een lijst met de volgende opties komt naar voren:

    On Skip = Wanneer de pipeline wordt overgeslagen ga door naar de volgende.

    On Success = Wanneer de pipeline succesvol heeft gedraaid ga door naar de volgende.

    On Failure = Wanneer de pipeline faalt ga door naar de volgende.

    On Completion = Wanneer de pipeline klaar is, ongeacht sucess of falen ga door naar de volgende.

    Klik en sleep het **groene blokje** naar 1 van de andere pipelines en doe dat vervolgens nog één keer voor een andere pipeline. Je zou nu alle 3 de pipelines aan elkaar verbonden hebben met 2 **groene pijlen**.
15. Klik op de **Blauwe knop** met de tekst **Publish all** en vervolgens op de knop **Publish**. Door te publishen komen de andere aanpassingen **Live** te staan, en kan het gebruikt worden.
16. Hoera! je eerste pipelines klaar. Nu willen we de pipeline nog draaien, dit kan op verschillende manieren:
    - In het scherm van de pipeline zelf zie je een `Play knop` met de tekst **Debug**. Dit zorgt ervoor dat je de pipeline draait zoals je hem nu hebt gemaakt. Ook als je nog niet hebt opgeslagen of gepubliceerd, wordt de pipeline uitgevoerd zoals je deze nu in je scherm ziet.
    - Naast **Debug** zien we een **Bliksemschicht** met de tekst **Add trigger**. Als je deze aanklikt krijg je de optie voor **Trigger now**, hiermee draai je de pipeline zoals deze gepubliceerd is. Klik **Trigger now** aan en een optie zou verschijnen om parameter waarde in te vullen, gezien deze er niet zijn kunnen we op **OK** klikken.

17. Wacht tot je de melding rechtsboven in beeld krijgt met dat de pipeline succesvol heeft gedraait. Draai de pipeline hierna nog eens via de **Debug knop**. Je zult zien dat de informatie over het draaien van de pipeline onder in beeld verschijnt.

Opdracht 2 - Monitoring

In het onderdeel "monitoring" kun je niet alleen bekijken hoe eerdere pipelines gedraaid hebben, maar je kunt ook notificaties uitsturen wanneer er aan bepaalde voorwaarden voldaan wordt.

1. Klik aan de linkerkant op het metertje (**Monitor**). Je komt nu meteen bij **Pipeline runs** uit, en zal in de horizontale navigatie balk 2 opties zien in de vorm van **Triggered** en **Debug**. In beide tabs zou zowel de `PL_copy_Master` pipeline moeten staan als de bijhorende onderliggende pipelines.

2. Klik de één van de onderliggende pipelines aan, in 1 van de 2 tabbladen. Net als bij het draaien van de Debug variant zie je een regel met informatie over de gedraaide pipeline. Houd je muis op de naam van de activity onderin het scherm, er verschijnen nu 2 opties: **Input**, **Output** en **Details**.

3. Klik op **Input**, je ziet nu een stuk JSON code waaruit te lezen is welke kolom uit de source, naar welke kolom in de sink is gegaan. Hierin kan je ook informatie zien als je specifieke data d.m.v. een query ophaalt, parameters, variabelen en meer. Sluit de **Input Tab** af door op het **Kruisje** te klikken.

4. Klik op **Output**, ook hier zie je een stuk JSON code. De **Output** bevat informatie over het draaien, zoals: Hoelang duurde het, hoeveel rijen zijn gelezen en hoeveel zijn overgehaald en meer. Sluit de **Output Tab** af door op het **Kruisje** te klikken.

5. Klik op **Details**, je ziet een visuele weergaven van de **Output**. Sluit de **Details** af door op het **Kruisje** te klikken.

6. Aan de linkerkant zien we **Notifications** met de optie **Alerts & metrics**. Klik deze aan.

7. In de horizontale navigatiebalk zien we de optie **New alert rule**. Klik deze aan.

8. We gaan een Alert Rule maken die een notificatie stuurt wanneer de pipeline een fout heeft. Geef de **Alert rule name** een duidelijke naam die de lading dekt (bijv. `Alert on error`).

9. Bij **Severity** zijn er meerdere opties mogelijk, namelijk:

   Sev 0 = Critical

   Sev 1 = Error

   Sev 2 = Warning

   Sev 3 = Informational

   Sev 4 = Verbose

   Voor ons doeleinde kiezen we **Sev0**.

10. Klik bij **Targer criteria** op het **Add criteria**. Een lange lijst met opties zal verschijnen voor verschillende soorten metrics waarover gerapporteerd kunnen worden. Kies voor de **Succeeded pipeline runs metrics** en klik op **Continue**.

11. Klik bij **Values** de optie bij **Name** aan en kies de `PL_copy_Master` pipeline.
12. De andere settings kunnen blijven zoals ze zijn. Klik vervolgens op **Add criteria**.
13. Klik bij **Configure Email/SMS/Push/Voice notification** op **Configure notification**.
14. Een nieuwe **Action group** zal aangemaakt moeten worden. Dit is een groep waarin mensen geplaatst kunnen worden om genotificeerd te worden over de door jouw aangemaakte regel. Vul bij **Action group name** een duidelijke naam in en geeft bij **Short name** een herkenbare afkorting van de groepsnaam.
15. Klik bij **Notifications** op **Add notification** en geeft de **Action name** een duidelijke naam. Kies vervolgens bij **Select which notifications you'd like to receive** de optie **Email** en vul hier een e-mailadres is waar je nu toegang tot hebt. Andere opties mogen ook zodat je deze kan uitproberen. Wanneer je alles hebt toegevoegd dat je wilt, klik je op **Add notification**.
16. Klik vervolgens op **Add action group**. Gaat dit fout, laat het weten aan de trainer.
17. Klik op **Create alert rule**
18. Ga terug naar **Pipeline runs** en de tab **Triggered**, houd je muis op de naam van de `PL_copy_Master`. Er verschijnt een `Play knop met pijltjes` (rerun) klik deze aan. Wacht tot pipeline weer klaar is, na iets meer dan een minuut zou je een mail en/of andere notificaties dienen te ontvangen.

Opdracht 3 - Parameters en Variablen

Met behulp van parameters kun je je pipeline meer dynamisch maken. Bijvoorbeeld door alleen de data op te halen die gewijzigd is na een bepaalde datum/tijd.

1. Klik links op het **Potloodje** (Author) en ga vervolgens terug naar de pipeline voor **Address**.
2. In de balk onderin zie je de tab **Parameters**, klik deze aan als je hier niet al opzit.
3. Klik op **New**, een nieuwe parameter wordt aangemaakt. Vul bij **Name** het volgende in: **ModifiedDate**. De **Type** kan op **String** blijven staan.
4. Klik op het blokje voor de **Copy data**. Klik vervolgens op de tab **Source** en kies bij **Use query** de optie **Query**.
5. Er verschijnt nu een Query veld, klik deze aan. Onder het veld verschijnt de optie **Add dynamic content** klik deze aan.
6. Type of plak de volgende query in het veld:

```
SELECT * FROM [SalesLT].[Address] WHERE ModifiedDate >=
'@{formatDateTime(pipeline().parameters.ModifiedDate,'yyyy-MM-dd')}'
```

7. Wanneer je nu op **Preview data** klikt, krijg je de vraag in een waarde in te vullen. Vul hier **1900-01-01** in om mee te testen. Klik vervolgens op **OK**.

8. Ga nu naar de `PL_copy_Master` en klik de "execute pipeline activity" voor **Address** aan. In de tab **Settings** zal je zien dat er gevraagd wordt om een **Value** voor de parameter **ModifiedDate**. We gaan deze niet handmatig vullen, maar met behulp van een variabele.

9. Klik op het canvas en vervolgens op de tab **Variables**. maak een nieuwe variable aan door op **New** te klikken.

10. Noem de variabele `FilterDate`

11. Uit de lijst met **Activities**, klik op de optie **General**. Klik en sleep **Set variable** op het canvas.

12. Verbind het **groene blokje** met de **Address** pipeline.

13. Klik op het **Set variable** blokje en geef deze een duidelijke naam, bijvoorbeeld "Stel ModifiedDate in".

14. Ga naar de **Settings** tab en kies **FilterDate**. Het is nu mogelijk om een waarde te plaatsen. Vul hier het volgende in: **2007-01-01**.

15. Klik vervolgens weer op de **Address** pipeline en vervolgens op de tab **Settings**.

16. Klik het invul veld bij **Value** aan en klik vervolgens op **Add dynamic content**.

17. Ga in het nieuwe scherm naar de tabl **Variables**. Klik op de variabele **FilterDate** en vervolgens op **OK**.

18. Klik op de **Blauwe knop** met de tekst **Publish all** en vervolgens op de knop **Publish**.

19. Klik op **Add trigger** en vervolgens **Trigger now** en gevolgt bij **OK**.

20. Klik aan de linkerkant op **Monitor** (het metertje). Ga naar **Pipeline runs** indien deze niet meteen opent. Je ziet nu nieuwe pipelines draaien en bij de pipeline van **Address** zou je nu een **[@]** moet zien staan onder de kolom **Parameters**. Klik deze aan, je zou de waarde moeten zien die in je variable had gestopt.

Inhoudsopgave