Lab 9 - Batching en DIUs

Vereisten

Om het lab te kunnen starten is het van belang dat Lab8 is afgerond.

Doel

We hebben nu zo goed als alles behandeld rondom de standaard orchestratie in de ADF. Toch kan het voorkomen dat enkele pipelines zoveel data moeten overhalen dat ze niet heel vlot draaien. Er zijn enkele knoppen waar nog aan gedraaid kan worden om dit sneller te kunnen laten verlopen in de vorm van Batching en DIUs. Volg de opdrachten stap voor stap.

Opdracht 1 - Batching

1. Ga naar de `PL_copy_Deltaload_Training` pipeline en klik binnen de **ForEach** op de **Copy data** activiteit.

2. Ga naar de tab **Sink**. Onder **Pre-copy script** zie je de optie **Write batch size** en vul hier 1 in.

3. Klik op **Debug** en wacht tot de pipeline klaar is. Je zult zien dat het nu heel lang duurt om alles te laden omdat er 1 rij per keer wordt weggeschreven. Dit is natuurlijk niet gunstig en je wilt dit zo hoog mogelijk hebben. Normaliter bepaalt de ADF zelf hoe groot zijn Batch sizes zijn, dit is meestal tussen de 1200 en 1500 regels. Het kan zijn dat je een proces hebt, waarbij het van belang is dat alle data in 1x geladen wordt zodat er geen mismatches kunnen ontstaan. Dit is bijvoorbeeld erg fijn als je een row-based datamodel hanteert.

4. Verander de **batchsize** van 1 in iets anders, klik op **Debug** en bekijk je resultaten. Probeer enkele **batchsizes** tot het moment dat het geen verschil meer maakt.

Opdracht 2 - Data Integration Units.

1. Ga in de **Copy Tables** activiteit naar de tab **Settings**.
   Je ziet hier de optie voor **Data integration unit**, en deze staat standaard op **Auto**. Hiermee bepaalt de ADF zelf hoeveel DIUs het denkt nodig te hebben voor een bepaalde workload. Vaak is de bepaling accuraat maar...:

   - Bij **Auto** start het aantal DIU's op 4. Door dat standaard op 2 in te stellen realiseer je al een redelijke besparing.
   - Soms heb je bij voorbaat extra rekenkracht nodig, dan kun je de DIU's juist handmatig verhogen.

2. Pas de **Data integration unit** naar **2**.

3. Klik op **Debug** en wacht tot de pipeline klaar is. Bekijk de resultaten, het meeste zal klaar zijn tussen de 10 en 15 seconden.

4. Verander de **Data integration unit** van 2 in iets anders, klik op **Debug** en bekijk je resultaten. Probeer enkele **Data integration units** tot het moment dat het geen verschil meer maakt.

Wil je meer weten over de kosten die je aan ADF kwijt bent? Koen Verbeeck schreef dit handige artikel: How you can save up to 80% on Azure Data Factory pricing

Einde Lab 9

Inhoudsopgave