Lab 9 - Batching and DIUs

Requirements

In order to start the lab, it's important that Lab 8 is completed.

Objective

We've now covered just about everything regarding standard orchestration in the ADF. However, there may be situations where certain pipelines have to pull so much data that they don't run very smoothly. There are a few buttons that can still be tweaked to speed this up in the form of Batching and DIUs. Follow the assignments step by step.

Assignment 1 - Batching

1. Go to the `PL_copy_Deltaload_Training` pipeline and click on the **Copy data** activity within the **ForEach**.
2. Go to the **Sink** tab. Under **Pre-copy script**, you'll see the option **Write batch size**. Fill in 1 here.
3. Click on **Debug** and wait until the pipeline is done. You'll see that it now takes a long time to load everything because only one row at a time is written. This, of course, is not favorable, and you want this to be as high as possible. Normally, the ADF itself determines how large its batch sizes are, usually between 1200 and 1500 lines. It could be that you have a process where it is important that all data is loaded at once so that no mismatches can occur. This is especially nice if you use a row-based data model.
4. Change the **batch size** from 1 to something else, click on **Debug**, and check your results. Try a few **batch sizes** until it makes no difference anymore.

Assignment 2 - Data Integration Units.

1. In the **Copy Tables** activity, go to the **Settings** tab. Here you see the option for **Data integration unit**, which is set to **Auto** by default. With this, the ADF determines itself how many DIUs it thinks it needs for a certain workload. Often, this determination is accurate, but...:
   - With **Auto**, the number of DIUs starts at 4. By defaulting this to 2, you can already achieve considerable savings.
   - Sometimes you need extra computing power in advance, then you can manually increase the DIUs.
2. Change the **Data integration unit** to **2**.
3. Click on **Debug** and wait until the pipeline is done. Check the results, most will be ready between 10 and 15 seconds.
4. Change the **Data integration unit** from 2 to something else, click on **Debug**, and check your results. Try a few **Data integration units** until it makes no difference anymore.

Do you want to know more about the costs you incur with ADF? Koen Verbeeck wrote this helpful article: How you can save up to 80% on Azure Data Factory pricing

End of Lab 9

Table of Contents