Published in ABN AMRO Developer Blog

ABN AMRO Follow
Aug 26, 2020 · 4 min read · Listen

Azure SQL Database AdventureWorks to Databricks Delta Migration

By Vijaybabu Nakkonda

In this article, Vijaybabu Nakkonda explains how to migrate a database platform from Azure SQL database to Databricks Delta. This use case is very interesting for teams working on migration, platform evaluation, PoC, and many more. Especially in situations with (un)foreseen data growth in terms of Volume (size), Variety (semi-structured / unstructured), and Velocity (scalability). In those cases, it is very important to evaluate and choose the best platform for migration.

In many modern data warehouse solution architectures, Microsoft recommends Azure Data Lake Store Gen 2 with Azure Databricks. This article provides step by step procedure to migrate database from Azure SQL DB to Azure Databricks Delta.

The scope of this article is limited to migrating SQL Tables to Delta. The migration of Views, Functions, Procedures, Synonyms, User Accounts, Roles is not covered. The complex topic of performance comparison between SQL DB and Delta is not covered in this article.

1. Provision Azure SQL DB

First, let's set up an Adventureworks database on Azure SQL DB. Find Powershell scripts to deploy SQL Server and Adventureworks database on Azure. In this case, Microsoft docs was used as a reference, further improved to a) Included an *if* condition, to avoid errors in case of re-run b) Added a small script to include client IP range into Azure SQL Server firewall and c) A small query on SQL database to list the number of tables available on Azure SQL DB.

The scripts are executed multiple times for testing. On average, it took 200 seconds to spin up an Azure SQL Server and create an Adventureworks database. Once the Azure SQL DB is spin up, you can perform a connection test from SQL Server Management Studio by querying few records.

2. Setup Azure Data Lake Gen2, Key Vault, Service Principle Account and Access to ADLSG2

Refer PowerShell scripts to:

- Create storage account on ADLS Gen 2 (with hierarchical namespace enabled) and create a container to hold delta table files
- Create a service principle and grant Storage Blob Data Contributor to the storage account
- Create an Azure Key Vault and securely store the service principle application id, secret, and Azure SQL DB password.

writing such an excellent article. I chose the best way of connecting: via a service principle. I struggled a little bit to create a service principle using PowerShell Az commands. This post helped me to understand the right scripts to create service principles.

3.0 Provision Azure Databricks Workspace and mount ADLSG2 container

3.1 Spin up Azure Databricks workspace

If you don't have an Azure Databricks workspace, click here. Only five parameters to configure, such as Subscription, resource group, Workspace name, Location, and Pricing Tier (For PoC purpose use 14 days Trial period). For more details on the pricing tier, follow the link here. Furthermore, there is a tab to manage networking to select Vnet for deployment. Also, a tab for Tag management, final review, and create a workspace.

3.2 Create a secret scope on Azure Databricks to connect Azure Key Vault

Creating a secret scope is basically creating a connection from Azure Databricks to Azure Key Vault. Follow this link to read more details on the secret scope. To quickly perform actions follow below instructions,

1. From your Azure Key Vault copy value of two properties DNS Name and Resource ID. A sample image is shown here
2. Navigate to your Databricks account URL https://<location>.azuredatabricks.net/?o=<your organization id>#secrets/createScope
3. Choose All users if you are in a standard pricing tier and enter a scope name. Paste DNS Name and Resource ID into the corresponding text box and save. A sample image is shown here.

4. Azure SQL DB to Azure Databricks Delta Migration

Finally, we are here to execute Databricks scripts for migration. Launch Azure Databricks, and from the workspace tab, right-click and select import menu. On Import wizard, select Import from URL and copy-paste the below URL
https://github.com/VijaybabuNakkonda/AzureSQL_to_Databricks_Delta/blob/master/notebooks/sql_to_delta.scala

The notebook contains 15 cell commands. The steps are summarized below:

1. Connect and Mount ADLS Gen2 Storage account on Azure Databricks using scoped credentials via Azure Key Vault
2. Create a connection to Azure SQL by using scoped credentials via Azure Key Vault
3. Select and query Information Schema from Azure SQL, to get base table information such as schema and table names
4. For each Schema available from SQL create the same on Databricks by executing SQL execute Create schema <schema_name>
5. For each Table exist on SQL, create spark dataframe. Read data from SQL tables and assign them to dataframes
6. Now, table data is available on spark dataframe. For each dataframe, write data to ADLS Gen2 location using delta format
7. Now, for each location from ADLS Gen2 which has been written in the previous step, Create databricks table by referring the same storage location from ADLS Gen2.

Migration is completed: Query and Play with Tables!

To drop the Azure resources provisioned for this project use scripts from here. Delete Azure Databricks manually from the portal.