**Walkthrough: Building an advanced solution with SQL Data Warehouse**

This walkthrough provides guidance on how to build advanced analytics solutions on SQL Data Warehouse. A SQL programmer should be able to complete this walkthrough almost entirely by executing the provided Powershell Scripts and SQL commands.

This walkthrough is built on a public data, [NYC Taxi Trip data](http://chriswhong.com/open-data/foil_nyc_taxi/) in 2013. This data is about 20GB of compressed CSV files (~48GB uncompressed), comprising more than 173 million individual trips and the fares paid for each trip. Each trip record includes the pickup and drop-off location and time, anonymized (driver's) license number (aka hack\_license) and medallion (taxi’s unique id) number, trip distance, etc. Each fare record includes payment related information like payment type, total amount of payment, tip amount, etc.

Building your first advanced analytics solutions on SQL Server: step by step tutorial

A typical end to end solution of advanced analytics starts from getting the data, then exploring the data, creating features, training the model, and ends at deploying the model in production. With the following steps, you should be able to build an end to end solution on the example data set.

*Step 1: Provision a SQL Data Warehouse in Azure cloud*

*Step 2: Create Azure blob storage account*

*Step 3: Download the sample scripts to a local file directory*

*Step 4: Execute a single PowerShell script to:*

* *Download AzCopy*
* *Copy data from public blob to your private blob storage account with AzCopy*
* *Load data from your SQL DW*
  + *Create external tables for NYC taxi dataset on the blob storage account*
  + *Create tables on SQL DW to store NYC taxi dataset*
  + *Import the NYC taxi dataset from external tables into SQL DW tables*

*Step 4: Basic data exploration*

**Pre-Requisites**

The walkthrough is intended for users familiar with fundamental database operations like creating databases, tables, importing data into tables and querying the tables using SQL. We will provide the SQL scripts to execute. To use this walkthrough you must have SQL Data Warehouse.

### Step 1: Provision a SQL Data Warehouse in Azure cloud

Follow the documentation at <https://azure.microsoft.com/en-us/documentation/articles/sql-data-warehouse-get-started-provision/> to provision a SQL Data Warehouse instance. Make sure that you keep down the following the SQL Data Warehouse credential which will be used to make SQL DW connections.

* Server Name
* SQLDW (Database) Name
* User Name
* Password

### Step 2: Create Azure blob storage account

The data used in this walkthrough is shared in a public blob storage container in Azure in a .csv format. You can check out the data at [this link](https://public.blob.core.windows.net/nyctaxidataset/).

Follow the documentation at <https://azure.microsoft.com/en-us/documentation/articles/storage-create-storage-account/> to create your own Azure storage account, if you don’t already have one. Please keep down the following storage account credential. The data will be copied from the public blob storage container to a container in your own storage account.

* Storage Account Name
* Storage Account Key
* Container Name (which you want the data to be stored)

### Step 3: Download the sample scripts to a local file directory

Open a **Windows PowerShell** command console. Run the following PowerShell commands to download the sample dataset that is used in this walkthrough, and the example SQL script files that we share with you on Github to a local directory. You can change the value parsed to parameter *DestDir* to any local directory. If *DestDir* does not exist, it will be created by the PowerShell script.

**Note:** You might need to **Run as Administrator** when executing the following PowerShell scripts if your *DestDir* needs Administrator privilege to create or write.

*$source = ‘https://raw.githubusercontent.com/Azure/Azure-MachineLearning-DataScience/master/Misc/SQLDW/Download\_Scripts\_SQLDW\_Walkthrough.ps1’*

*$ps1\_dest = “$pwd\Download\_Scripts\_SQLDW\_Walkthrough.ps1”*

*$wc = New-Object System.Net.WebClient*

*$wc.DownloadFile($source, $ps1\_dest)*

*.\Download\_Scripts\_SQLDW\_Walkthrough.ps1 –DestDir ‘C:\tempSQLDW’*

### Step 4: Execute a single PowerShell script to

* *Download AzCopy*
* *Copy data from public blob to your private blob storage account with AzCopy*
* *Load data from private blob storage account to your SQL DW*
  + *Create external tables for NYC taxi dataset on the blob storage account*
  + *Create tables on SQL DW to store NYC taxi dataset*
  + *Import the NYC taxi dataset from external tables into SQL DW tables*

Open a **Windows PowerShell** command console. Run the following PowerShell commands to import data to SQL DW.

*./SQLDW\_Data\_Import.ps1*

Input your credential when prompted. Please be patient in this step. The process of copying data from public blob to your private storage account could take about 15 minutes and the process of loading data from your storage account to SQL DW could takes about 20 minutes.

Step 5. Explore the data and engineer features in SQL DW

After the data is imported to SQL DW, follow sections in documentation <https://azure.microsoft.com/en-us/documentation/articles/machine-learning-data-science-process-sql-walkthrough/> to explore the dataset and build machine learning models.