Data Movement: Data Lake Store to Azure SQL DW

Updated on: 1/31/2017

# Introduction

You will learn how to setup a recurring job to run and how to copy the output of that job in a recurring format from the Data Lake Store to SQL DW. This is a common pattern employed to move transformed data to an database for reporting/analytics on aggregated data scenarios.

The input data for this lab is here:

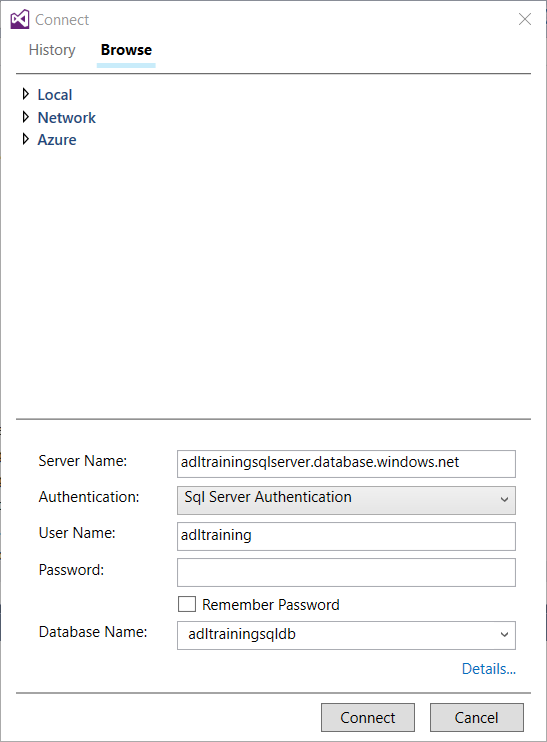
adl://**adltrainingsampledata**.azuredatalakestore.net/GHData/ProjectMembers\_large.csv

# Prerequisites

For this lab, you will need:

* Access to a Data Lake Store account that you can write to
* Access to a Data Lake Analytics account that you can submit jobs to
* Access to the **adltrainingsampledata** Data Lake Store account that you can read from
  + This means be a member of the ADLTrainingUsers security group.
* Access to a Data Factory account where you can author
* Access to an Azure SQL DW (or Azure SQL DB if you prefer)

# Create Table in SQL DW

* Connect to your SQL DW using Visual Studio
  + In Visual Studio, open Server Explorer
  + Navigate to **Azure > SQL Databases**
  + Find your database, right-click on it and select **Open in SQL Server Object Explorer**
  + You’ll see this dialog
  + 
  + The fields in this dialog are already filled in for you
  + All you have to do is enter the correct authentication information
  + If this is the first time you are connecting to this Azure DQL DB, you may be asked to enable a firewall rule – go ahead and do it
  + The SQL Server Object Explorer window will open
* Run this query to create a table called **ProjectMembers\_large**
* Right click on your database in SQL Server object Explorer and select new Query
* Then paste in the following script to create the table

CREATE TABLE ProjectMembers\_large

(

repo\_id int,

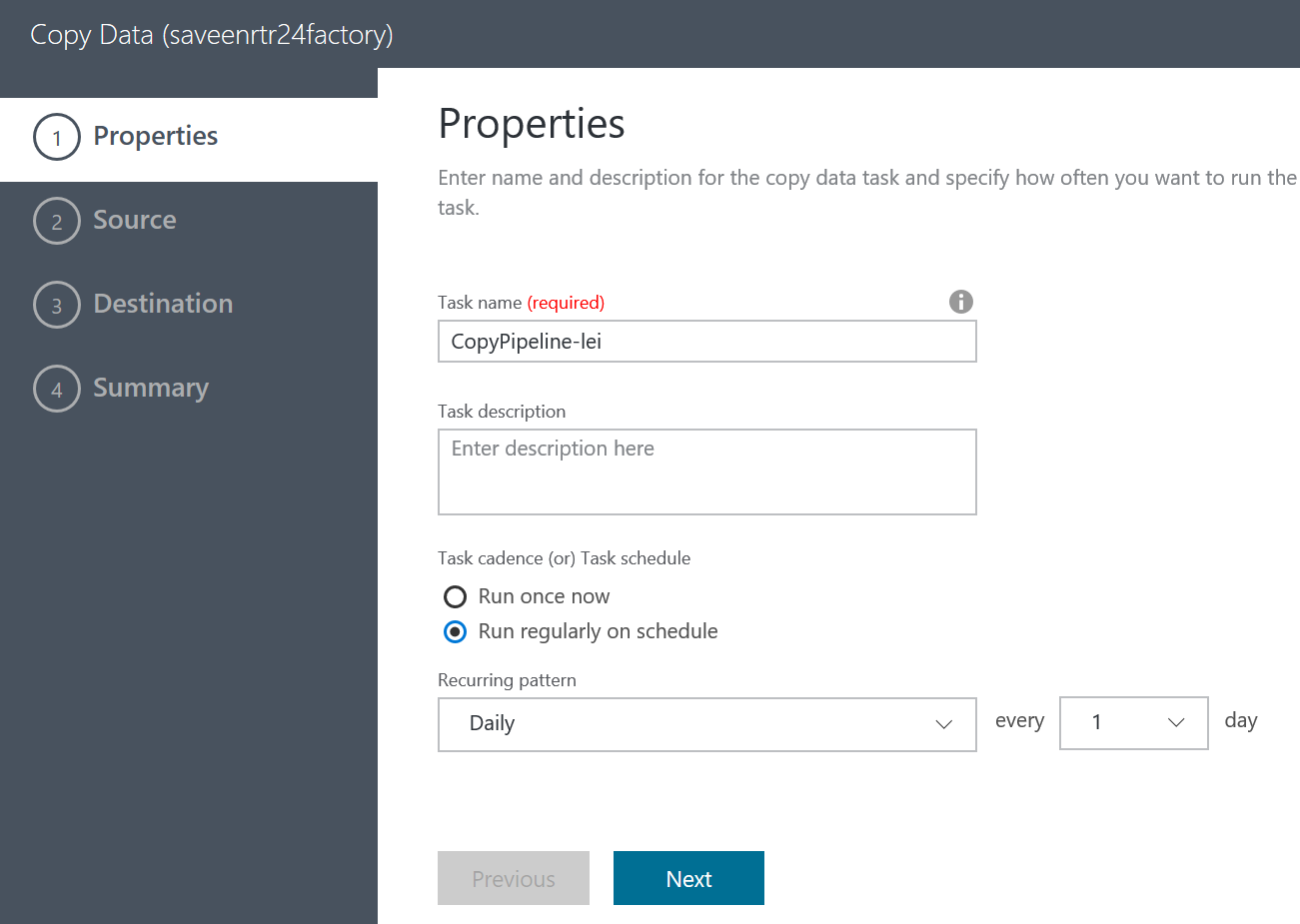
user\_id int,

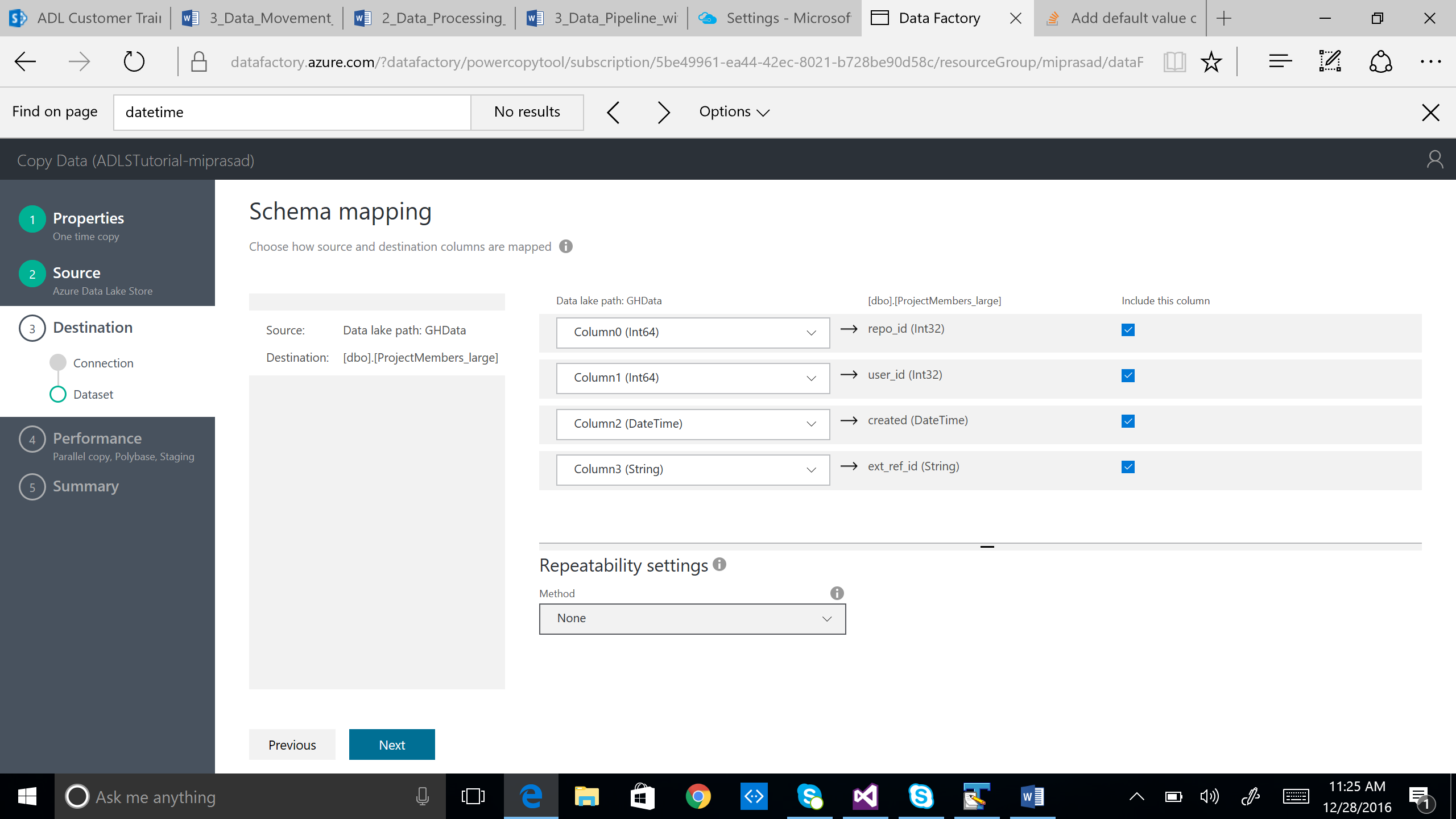
created datetime,

ext\_ref\_id varchar(255)

);

# Create a copy pipeline in ADF

* Open your ADF account
* click on **Copy data (PREVIEW)**
* Setup properties for the ADF Task
* 
  + **Name** for the Task(Activity) -> call it any thing you want
  + Set **Recurrence** to Daily
* Select the **Source** data store to be Data Lake Store
  + Set the **Authentication type** to **OAuth**
  + For the Subscription choose **ADLTrainingMS**
  + For the store account name select **adltrainingsampledata**
  + Click **Next**
  + Browse to the GHData folder
  + Select the file ProjectMembers\_large.csv
  + Click **Choose**
  + Click **Next**
* Select the output data store to be Azure SQL DW
  + Fill in the connection info for the Azure SQL DW you want to use
  + Select the table created in Step #1 and the columns will be mapped automatically



* Create the pipeline and Deploy it

# Verify that data was copied

* Run this query in the Azure SQL DB/DW

SELECT TOP 3 \* FROM [dbo].[ProjectMembers\_large]