**Data cleansing and data prep with Azure Synapse Analytics data flows and pipelines**

*Tutorial 001 June 2021*

Contents

[Overview 3](#_Toc74308893)

[Step 1: Taxi data 3](#_Toc74308894)

[Step 2: Acquire data for your Synapse workspace 3](#_Toc74308895)

[Step 3: Create a new pipeline 5](#_Toc74308896)

[Step 4: Create a new data flow 7](#_Toc74308897)

[Step 5: Explore the data 8](#_Toc74308898)

[Step 6: Clean and prep the data 10](#_Toc74308899)

[Step 7: Land the data in Synapse SQL Pool 11](#_Toc74308900)

[Step 8: Debug & schedule your pipeline 12](#_Toc74308901)

# Overview

This tutorial will guide you through building a data flow solution in Azure Data Factory intended to analyze the taxi data sample trip data to prepare it for downstream analytics by data scientists and for reporting. You’ll learn how to copy the data from Azure storage accounts outside of your Synapse workspace into the ADLS Gen2 storage that is part of your workspace, explore the data, interactively analyze and clean the data, and load the results into a SQL pool. You’ll also use Synapse pipelines to orchestrate the Copy Activity and Data Flow with the Synapse scheduler and monitor the progress of the pipeline.

# Step 1: Taxi data

The data that you will use for this scenario can be downloaded here:

<https://github.com/kromerm/adfdataflowdocs/blob/master/sampledata/trip_data_1.csv>

<https://github.com/kromerm/adfdataflowdocs/blob/master/sampledata/trip_fare_1.csv>

For this tutorial, you should store those files in a Blob Store account. We will then move these files into your workspace ADLS Gen2 account below.

# Step 2: Acquire data for your Synapse workspace

1. Go to Data > Connect to external data

Graphical user interface, application

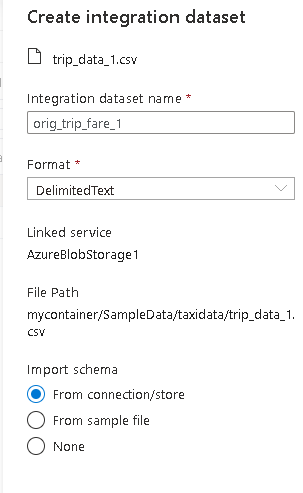
Description automatically generated

1. Select Azure Blob Store and connect to the location where you uploaded the taxi data from step 1.
2. In the storage explorer, find the taxi CSV files in your blob store from step 1.
3. Right-click and select “New integration dataset”

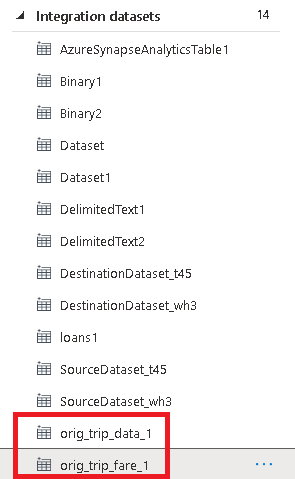
Graphical user interface, application

Description automatically generated

1. Name each dataset as “orig\_trip\_data\_1” and “orig\_trip\_fare\_1” respectively
2. Set the format to Delimited Text and Import schema to “From connection/store”



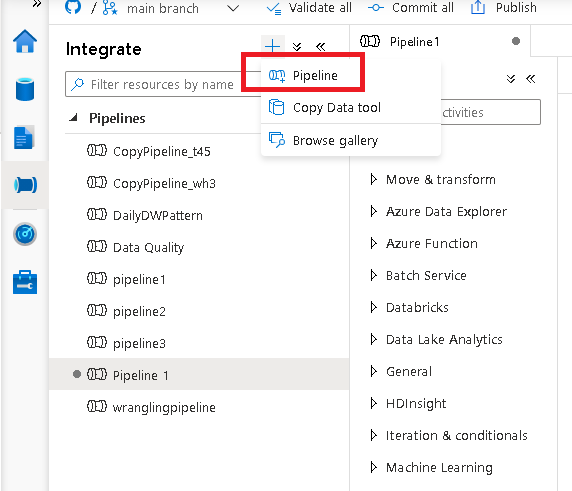
1. In your new dataset configuration, be sure to set “First row as header”.
2. Repeat these steps for each of the 2 taxi CSV data files
3. Under “Integration datasets” in the Data section of the Synapse studio UI, you should now see your 2 new datasets under “Integration datasets”.

# Step 3: Create a new pipeline

Next, create a new Synapse pipeline. We’ll use this pipeline as the primary place to build our Copy activity for moving data into our workspace as well as for analyzing and cleaning data in data flows.

1. Go to pipelines > New Pipeline



1. Name the pipeline “TaxiPipeline”
2. Add a Copy activity
3. For your source, choose the orig\_taxi\_data\_1 dataset

Graphical user interface, text, application, email

Description automatically generated

1. For the sink, select New dataset and choose ADLS Gen2 as the dataset type and for the linked service, choose the default linked service for your workspace.

Graphical user interface, text, application, chat or text message

Description automatically generated

1. We’re going to create datasets that defines a folder in your workspace’s data lake for a location to store the files along with the representative file names. Call the new datasets as “trip\_data\_1” and “trip\_fare\_1”, respectively.

Graphical user interface, text, application, email

Description automatically generated

1. For the folder name, type “SampleData”, choose “First row as header”, and “From connection/store” for Import schema.
2. Your copy pipeline is now ready to move the files into your workspace’s data lake.
3. Click on “Debug” to execute your pipeline in debug mode. This will move the taxi data files into your lake account.
4. Repeat step 3 for the 2nd file: taxi\_fare\_1.csv

Graphical user interface, text, application, email

Description automatically generated

1. You should now see the 2 files in your ADLS Gen2 storage explorer under SampleData

# Step 4: Create a new data flow

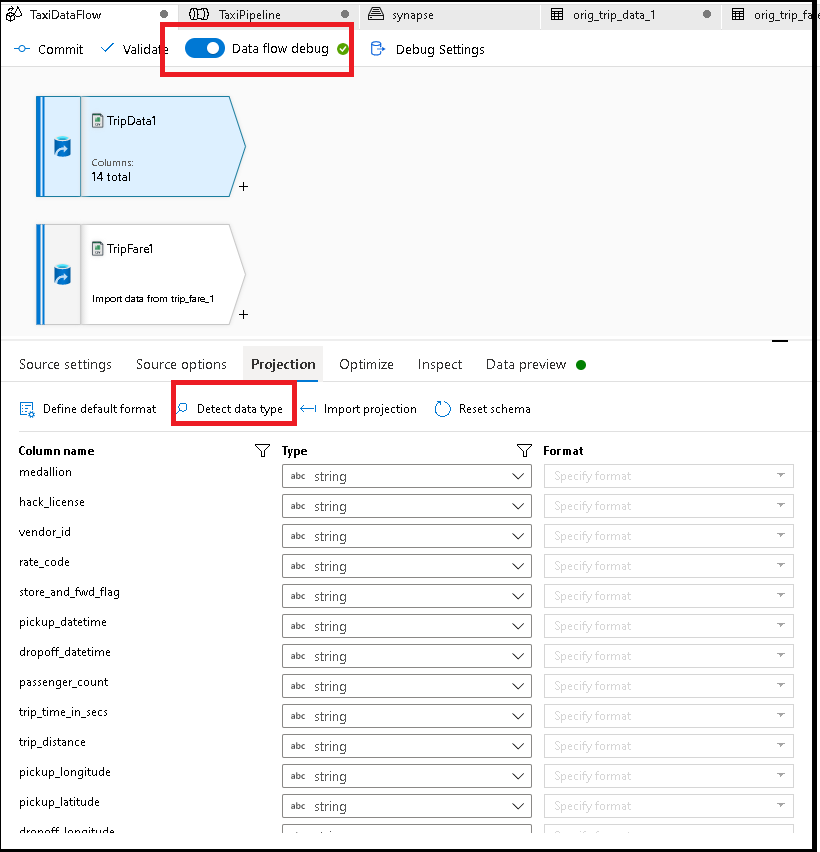
1. From the storage explorer, select “trip\_data\_1.csv”
2. Right-click and select “New data flow”
3. Name the data flow as “TaxiDataFlow”
4. Rename the sources as TripData1 and TripFare1 respectively to look similar to this:

Graphical user interface

Description automatically generated

# Step 5: Explore the data

Now that we have created a data flow, we are going to explore and analyze the data, then prepare it for downstream processing by data scientists and business analysts. Once you are inside the data flow with your sources defined, you can start a Debug session to open a Synapse Spark environment so that you can interactively work with the data.



1. Turn on Data flow debug and click “Detect data type” on each of the 2 sources from the Projection tab.
2. Data flows will sample the data sources and infer the data types, which will appear in the Projection tab.
3. Explore the data by clicking on “Data preview” on each of the sources.
4. On the trip data source, click on “rate\_code”, then “Statistics”

Table

Description automatically generated

1. We need to map values 1 and 2 as enumerations. We’ll do that in the next step.
2. On the trip fare source, click on “vendor\_id”, then “Statistics”

Table

Description automatically generated

1. Let’s take out the value “CMT” and only keep the “VTS” records. We’ll do that in the next step.
2. Continue exploring the data. You should observe that medallion and hack\_license will make a primary key / foreign key for a Join, which will do in the next step.

# Step 6: Clean and prep the data

1. First, let’s join the data. Click on the TripData1 source transformation and add a Join transformation.
2. In the Join, select TripFare1 for right stream, inner join, and join on medallion and hack\_license fields.

Graphical user interface, application

Description automatically generated

1. After the Join, add a Filter transformation. Here you will filter out the vendor\_id CMT with this expression: vendor\_id != 'CMT'
2. Next, add a Derived Column. This is where you will create a new column that is an enumeration of the rate\_code field. Name the new column as ‘rate\_code\_enum’ and the formula to use is: iif (rate\_code == 1, 'Basic','Standard').
3. We want to see the data in preview for exploration sorted by fare\_amount descending. So add a Sort transformation to do that.
4. In the Sort transformation, click “Data preview”, then “Refresh”. Your results should look similar to this:

Graphical user interface, table

Description automatically generated

1. Make sure to commit or publish your changes

# Step 7: Land the data in Synapse SQL Pool

1. Now you will create a new SQL table on the fly in your data flow to store your results in the Synapse SQL pool. Add a Sink transformation to the end of your data flow.
2. Create a new dataset and call it “sqltaxi”
3. For the Linked Service, choose the Synapse Analytics that points to your workspace.
4. For the Table Name, click “Edit” and type in dbo.taxidemo1
5. For DBName, enter the name of your SQL database
6. It should look similar to this:

Graphical user interface, text, application

Description automatically generated

1. Click Data Preview in your sink transformation to confirm that the data looks correct.
2. Your completed data flow should look like this:

Graphical user interface, diagram

Description automatically generated with medium confidence

# Step 8: Debug & schedule your pipeline

1. Now you are ready to test your pipeline end-to-end in debug mode. Open your existing pipeline from step 3.
2. Click on the Develop category on the left side of the Synapse Studio UI and find the TaxiDataFlow.
3. Drag and drop it into the pipeline canvas.
4. Connect it to your existing Copy activity, after the Copy:

A picture containing diagram

Description automatically generated

1. Next, click Debug to test it out.
2. You can click on the eye glasses icon in the Output tab of the pipeline to view details of each activity execution during the pipeline run.
3. After completion of the pipeline, you should see your data by clicking on the Data category in the Synapse Studio explorer on the left-hand side.

Table

Description automatically generated

1. Find the dbo.taxidemo1 table, click “Select Top 100 rows”.
2. Once you have validated your data, publish your work.
3. Now the pipeline is ready to be scheduled.
4. From the pipeline designer window, click Add Trigger > New/Edit.

A picture containing graphical user interface

Description automatically generated

1. You will be prompted to create a new trigger.