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

*Tutorial 004 August 2021*

Contents

[Overview 3](#_Toc80861038)

[Prerequisites 3](#_Toc80861039)

[Step 1: Acquire taxi data 3](#_Toc80861040)

[Step 2: Upload data to your Azure Synapse workspace 3](#_Toc80861041)

[Step 3: Explore the data 8](#_Toc80861042)

[Step 4: Clean and prep the data 10](#_Toc80861043)

[Step 5: Land the data in Synapse tables 11](#_Toc80861044)

[Step 5b (Optional): Land data in SQL Pool 12](#_Toc80861045)

[Step 6: Create a new pipeline 13](#_Toc80861046)

[Step 7: Schedule your pipeline 16](#_Toc80861047)

# Overview

This tutorial will guide you through building a data flow solution in Azure Synapse Analytics 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 table. 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.

# Prerequisites

Please follow [this video link](https://azure.microsoft.com/en-us/resources/videos/creating-azure-synapse-workspace/) to learn how to set-up your Synapse workspace before progressing through this tutorial. Provisioning a Synapse workspace is required to work with Synapse Data Integration.

# Step 1: Acquire taxi data

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

<https://aka.ms/tripdata>

https://aka.ms/tripfare

Click on the Download link and use “save as” for each of the files above.

Graphical user interface, application

Description automatically generated

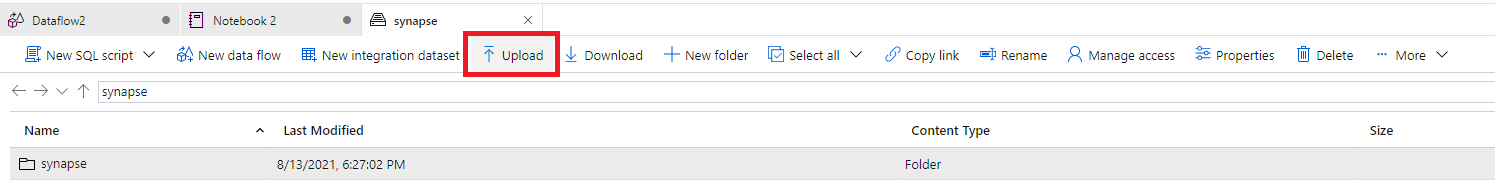
# Step 2: Upload data to your Azure Synapse workspace

1. Open Azure Synapse Analytics studio from your web browser
2. Go to Data > Linked > Azure Data Lake Storage Gen2

Graphical user interface, text, application, email

Description automatically generated

1. Click on the (Primary) container
2. Create a New Folder and call the folder “taxidata”
3. Double-click into the new folder “taxidata”
4. Click on Upload



1. Click the folder icon and choose the 2 trip files from the download above

Graphical user interface, application

Description automatically generated

1. You will now see the 2 Taxi Trip delimited text files (CSVs) in your Synapse ADLS Gen2 folder
2. Let’s now create a new data flow with this data
3. First, right-click on trip\_data\_1.csv and select “New data flow”

Graphical user interface, application

Description automatically generated

1. This will create both a new data flow as well as an integration dataset for the trip data CSV file
2. Name the data flow as “taxiDataflow” and name the Integration Dataset as “trip\_data\_1”
3. Set the format to Delimited Text and import schema from connection/store as below

Graphical user interface, text, application, email

Description automatically generated

1. This will create a new data flow and will drop you into the data flow design surface
2. For Source1, rename it to “TripData1”
3. Click Open on the Dataset property and set the First Row as Header checkbox
4. Click Schema > Import Schema from Connection

Graphical user interface, application

Description automatically generated

1. Navigate back to the data flow design tab
2. Now, add another Source to the data flow graph designer and call it “TripFare1”
3. Click New Dataset

Graphical user interface, application

Description automatically generated

1. Pick Azure Data Lake Gen2, Delimited Text and select the linked service that points to your ADLS Gen2 for your Synapse workspace
2. In your new dataset configuration, be sure to set “First row as header” and select the trip fare csv file
3. Name this new dataset “trip\_fare\_1”
4. Under “Integration datasets” in the Data section of the Synapse studio UI, you should now see your 2 new datasets under “Integration datasets”

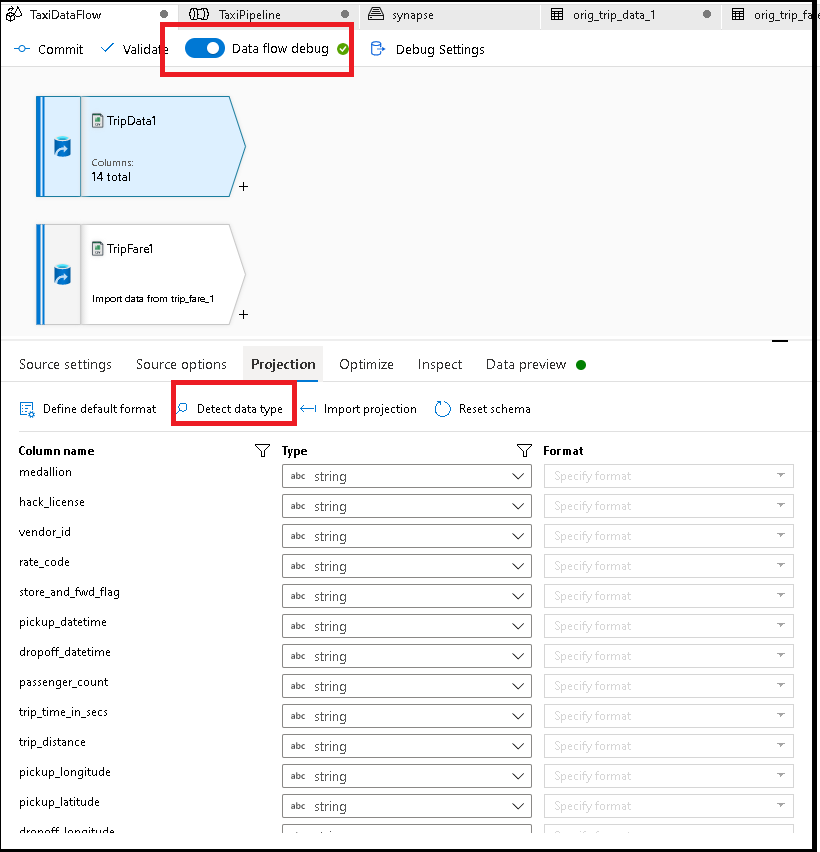
NOTE: If you have enabled VNET for your Synapse Workspace, you may have to enable the Interactive Authoring option in your Azure Integration Runtime in order to access the data files.

Graphical user interface, text, application

Description automatically generated

# Step 3: 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 4: 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 sort on the fare\_amount column.
4. In the Sort transformation, click “Data preview”, then “Refresh”. Your results should look similar to this:

Graphical user interface, table

Description automatically generated

# Step 5: Land the data in Synapse tables

Now you will create a new SQL table on the fly in your data flow to store your results as a Synapse table. This table should be considered the superset of all of the transformed data which could be used downstream by business reporting and business intelligence tools.

1. Add a Sink transformation to the end of your data flow
2. Select Workspace DB for the sink type and select your default Spark database
3. Enter “taxidata” for a new table name

Graphical user interface, text, application, email

Description automatically generated

1. Click on the sink settings and select “Recreate table” so that Synapse can create a new table called “taxidata” for you.
2. Click on the Mapping section in the sink and turn off auto mapping
3. Remove the first occurrence of the Medallion column
4. Click Data Preview in your sink transformation to confirm that the data looks correct.
5. If you’d like to perform the optional Sort transformation, you can add it next to the Derived Column
6. Sorting is generally used in big data analytics in this way for data exploration and understanding your data during data flow designing. You can remove it when you are finished previewing the data.
7. Your completed data flow should look something like this:

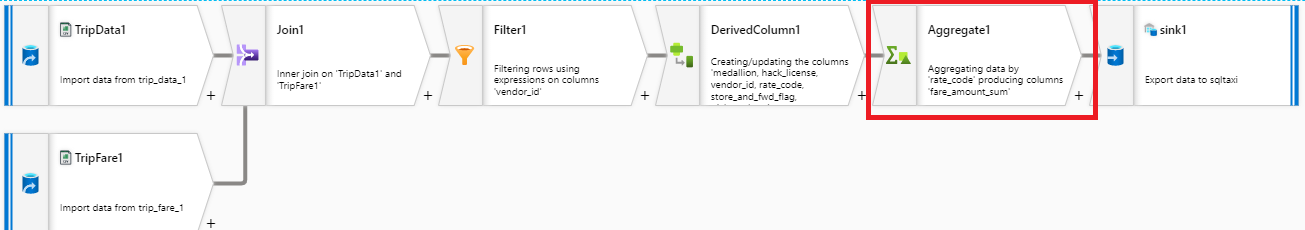
Graphical user interface, diagram

Description automatically generated with medium confidence

# Step 5b (Optional): Land data in SQL Pool

Now you will create a new SQL table on the fly in your data flow to store your results in the Synapse SQL pool. In this scenario, we’re going to create a table in the SQL dedicate pool for data science that is a set of aggregated values for fare amount across rate codes.

1. Add an Aggregate transformation after the Derived Column



1. In the Group By section, choose rate\_code
2. In the Aggregates section, create a new column called “SumOfFareAmount”
3. For the aggregate expression use this expression: sum({ fare\_amount})
4. Add a Sink transformation to the end of your data flow
5. Create a new Azure Synapse Analytics dataset and call it “sqltaxi”
6. For the Linked Service, choose the Synapse Analytics that points to your workspace.
7. For the Table Name, click “Edit” and type in dbo.taxidemo
8. Click “None” for import schema since we will generate a new table on the fly

Graphical user interface, text, application, email

Description automatically generated

1. For DBName, enter the name of your SQL database. In my case, I called my SQL Pool as “taxidata”

Graphical user interface, application

Description automatically generated

1. Click on the Mapping tab and switch off auto mapping
2. Rename the rate\_code column to rateCode

Graphical user interface, application

Description automatically generated

# Step 6: Create a new pipeline

Now that we’ve completed designing and unit testing the data flow logical graph, we can now add it to a Synapse integration pipeline so that we can test it end-to-end and schedule the data flow as part of a pipeline.

1. Click on the Integrate section of the Synapse studio

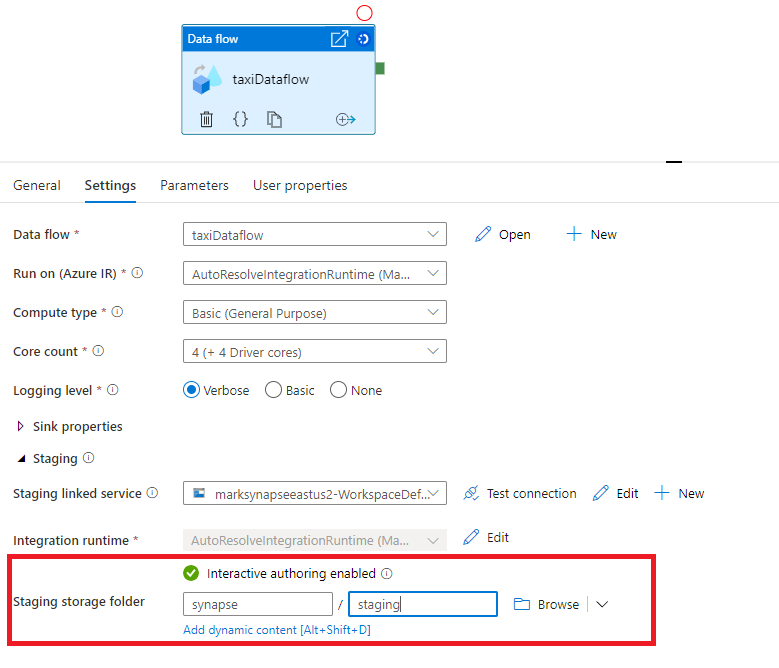


1. Go to pipelines > New Pipeline

Graphical user interface, text, application

Description automatically generated

1. Name the pipeline “TaxiPipeline”
2. Click on the Develop section of the studio
3. Open the Data Flows category and drag in the taxi dataflow to your pipeline design surface
4. If you are landing your data into the SQL Pool for the Sink, you will need to define a staging location in the data flow activity inside the pipeline



1. Click on “Debug” to execute your pipeline in debug mode. This will execute your data flow logic and load the results into your Synapse table.
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.
4. Find the “taxidata” Spark table that you just created, click “Select Top 100 rows”.

Graphical user interface, application

Description automatically generated

# Step 7: Schedule your pipeline

1. Once you have validated your data in the above steps, you can now publish your work by clicking the publish button.
2. Now the pipeline is ready to be scheduled.
3. From the pipeline designer window, click Add Trigger > New/Edit.
4. You will be prompted to create a new trigger.