

As we know Parquet Apache Parquet is designed for columnar storage, which is more efficiency to compare with CSV files during to query data.

Let's use this article to introduce how to Azure Data Factory to convert CSV file to Parquet.

Step1: Create linked service for CSV and Parquet files from Azure Data Lake Gen2.

In this sample, linked service is shared since the both CSV and Parquet files will store one Azure Data Lake Gen 2 account.

The screenshot shows the 'Create linked service' form in Azure Data Factory. The form includes the following fields and options:

- Name ***: A text input field containing 'ADLSGen2linkedservice'.
- Description**: A text area for providing a description.
- Connect via integration runtime ***: A dropdown menu showing 'AutoResolveIntegrationRuntime'.
- Authentication method**: A dropdown menu showing 'Account key'.
- Account selection method**: Two radio buttons, 'From Azure subscription' (unselected) and 'Enter manually' (selected).
- URL ***: A text input field containing 'https://myadfs.dfs.core.windows.net'.
- Storage account key**: A section with two tabs, 'Storage account key' (selected) and 'Azure Key Vault'. The 'Storage account key' tab contains a text input field with masked characters '*****'.
- Test connection**: Two radio buttons, 'To linked service' (selected) and 'To file path' (unselected).
- Annotations**: A section with a '+ New' button.
- Advanced**: A section with a '► Advanced ⓘ' link.

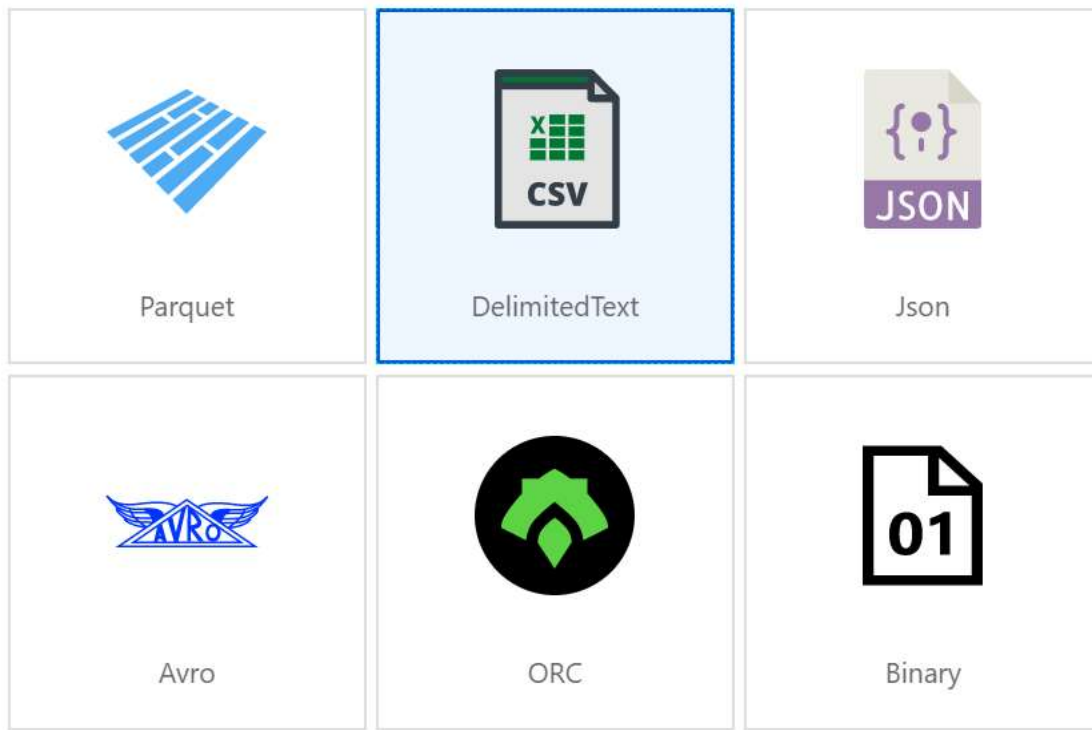
At the bottom of the form, there are three buttons: 'Apply', 'Test connection' (with a test icon), and 'Cancel'.

Step 2: Create Datasets for both CSV file and Parquet file.

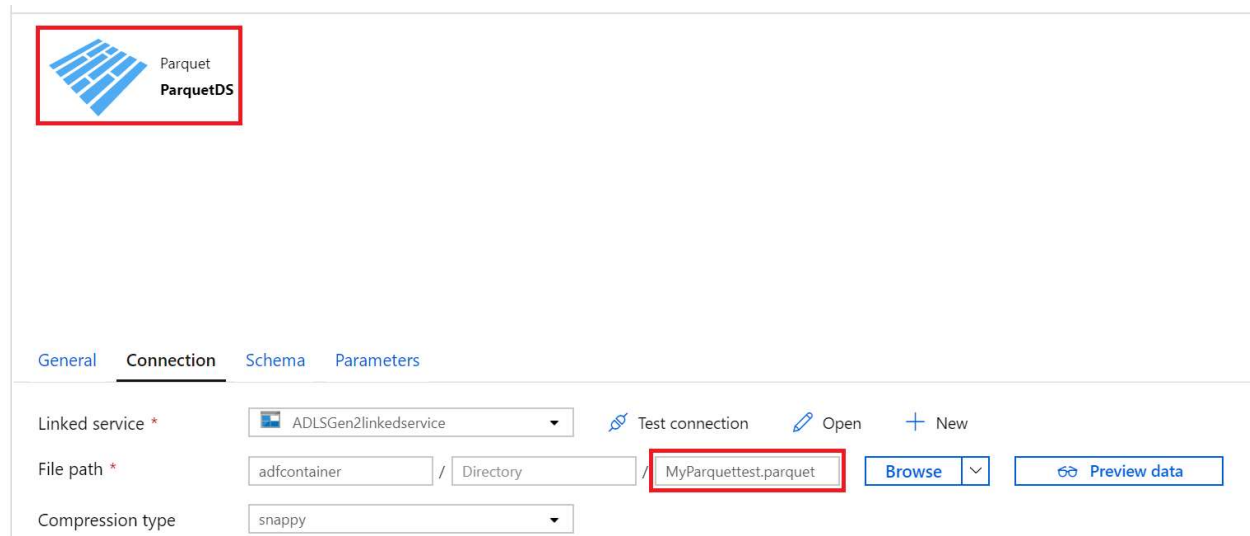
Please note to create one dataset for CSV file, and another dataset for Parquet file, both datasets will be used in next step.

Select format

Choose the format type of your data



Choose CSV format in dataset (Parquet is similar)



View the Parquet dataset, it will show data format is Parquet

Step 3: Create a copy activity in pipeline, then set source to CSV dataset, and sink to Parquet dataset.

General

Source

Sink

Mapping

Settings

User properties

Source dataset *

ADLSGen2CSVTestDS

Open

New

Preview data

Recursively

☒

Wildcard folder path

Please note that this will overwrite the folder path defined in dataset

General

Source

Sink

Mapping

Settings

User properties

Sink dataset *

ParquetDS

Open

New

Step 4: Of course, the last one step is to run the pipeline, Cheers!

Post step: No need other tools, it's simple to use ADF to view Parquet data format output.

Preview data

Linked service: ADLSGen2linkedservice

Object: MyParquetttest.parquet

Col1	Val1
1	2
2	3
3	4
4	5
5	6
6	7
7	8
8	9
9	10