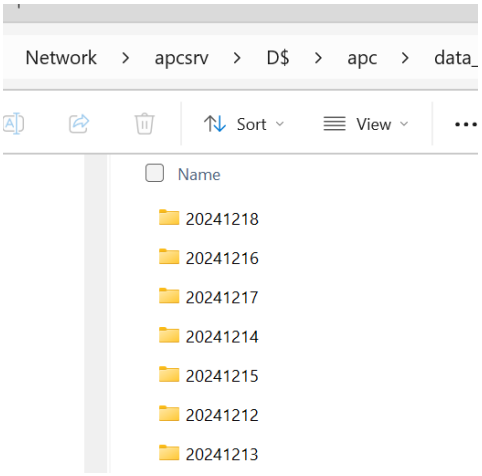
Introduction:

- The scope of Source destination delta process project is to load the raw files which are in compressed gzip format to the target location for weekly data processing.
- Load the prior 40 days bus state files from source to target location
- Azure Synapse pipeline will be triggered 2 times a week
- Email notification using Logic App when there is a pipeline failure
- Retry the data load up to 5 times if there is a data load failure

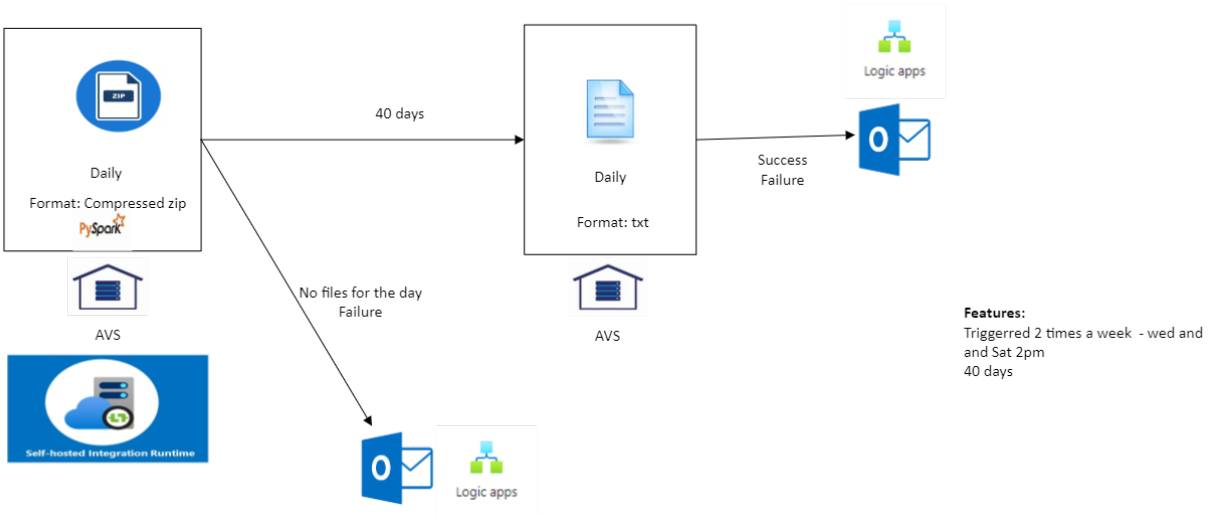
Target Folder:

Below is the target location



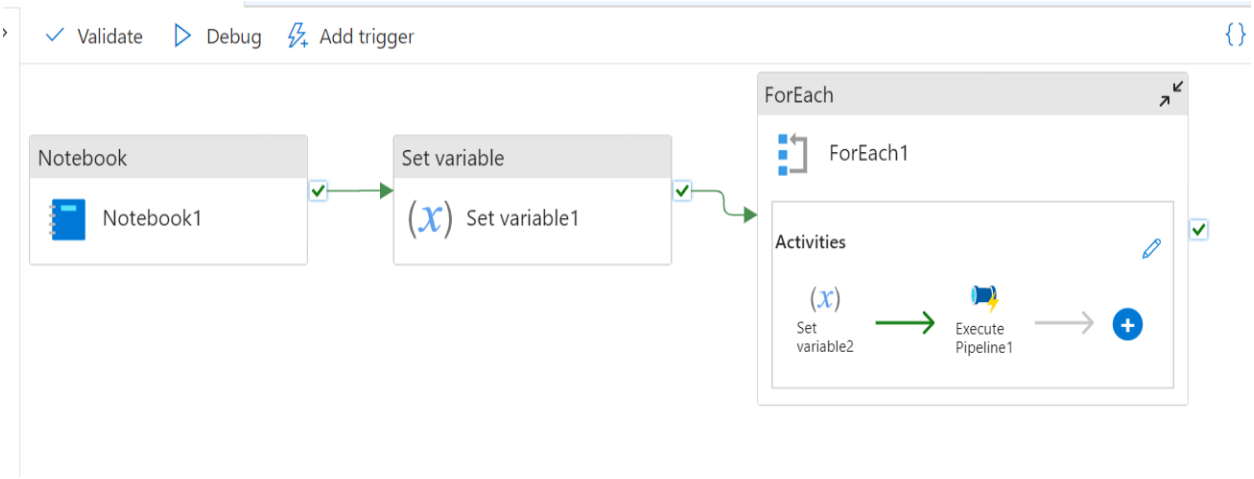
Pipeline design:

- RAW files which are in compressed gzip format to the target location.
- Azure Synapse pipeline will be triggered 2 times a week
- Wednesday and Saturday at 2pm.
- Load the prior 40 days bus state files from source to target location

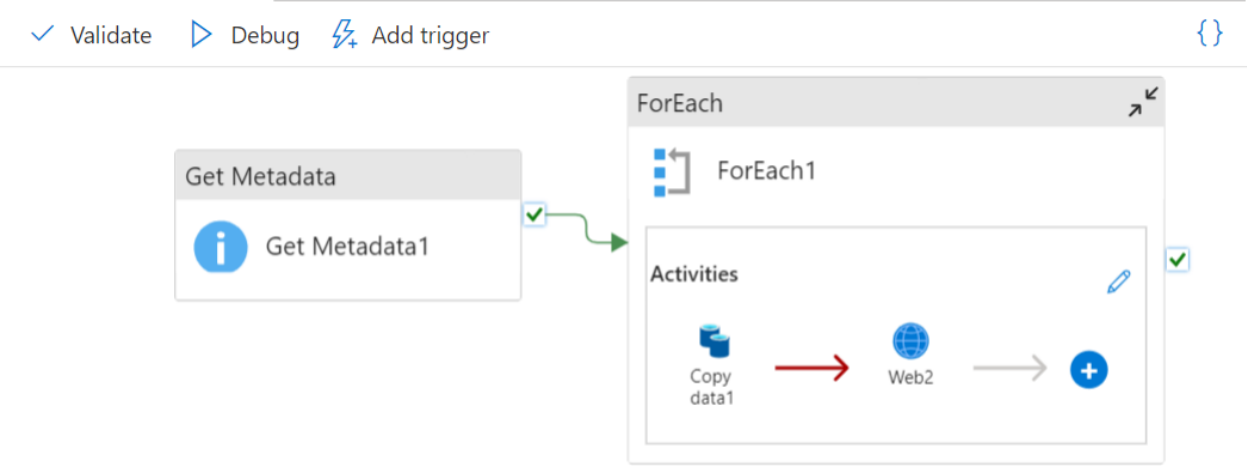


Pipeline design:

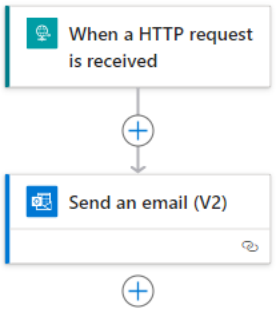
Azure Synapse pipeline to load the prior 40 days files. Pyspark is used for data processing to generate the list of prior 40 days.



Azure Synapse pipeline to uncompress and load the prior 40 days files to the target location



Email notification using Logic App when there is a pipeline failure



Key terms:

1. **Blob Storage Account:** Azure Blob Storage is Microsoft's object storage solution for the cloud. Blob Storage is optimized for storing massive amounts of unstructured data.

2. **Synapse Analytics Workspace:** A Synapse workspace is a securable collaboration boundary for doing cloud-based enterprise analytics in Azure.

3. **Spark Pool:** Apache Spark is a parallel processing framework that supports in-memory processing to boost the performance of big data analytic applications.

4. **DataFrames:** Data Frames are the distributed collections of data, organized into rows and columns.