# Azure data factory project.

## Introduction

In this project we will be building a data platform for reporting and prediction of the covid-19 outbreak.

* Data sources: European Center for Disease prevention and Control (Eurostat)
* Data flows: azure data factory
* Data transformation: HDInsight and databricks
* Ingestion of transformed data: Data Lake
* Necessary data for reporting the data trends: SQL Data warehouse
* We will orchestrate all of these pipelines using Azure data factory.

## Project overview:

**Our build data lake:**

Data Lake to be built with the following data, to aid Data Scientists to predict the spread of the virus/mortality

* Confirmed cases
* Mortality
* Hospitalization/ ICU Cases
* Testing Numbers
* Country’s population by age group

**Our build data warehouse:**

Data Warehouse to be built with the following data

* Confirmed cases
* Mortality
* Hospitalization/ ICU Cases
* Testing Numbers

**The data sources:**

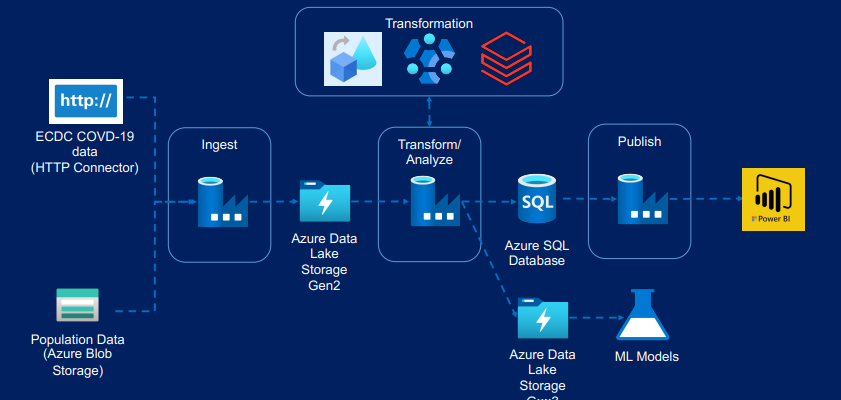
ECDC website:

* Confirmed cases
* Mortality
* Hospitalization / ICU Cases
* Testing Numbers

Eurostat Website:

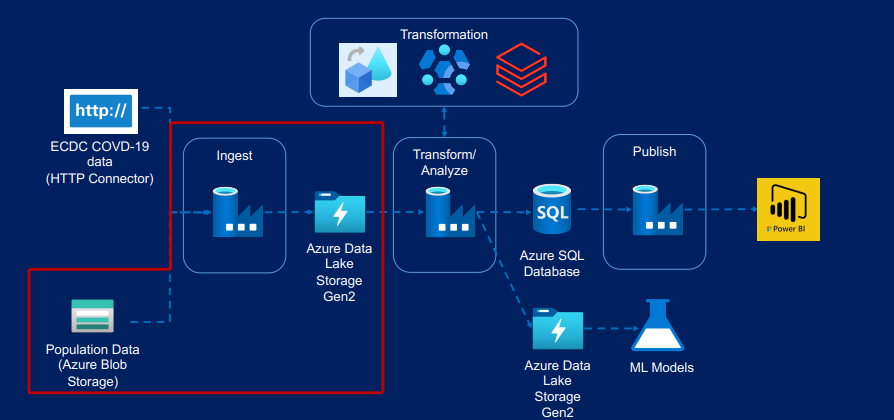
* Population by age

# Solution Architecture



# Step 1 – Data ingestion

Overview

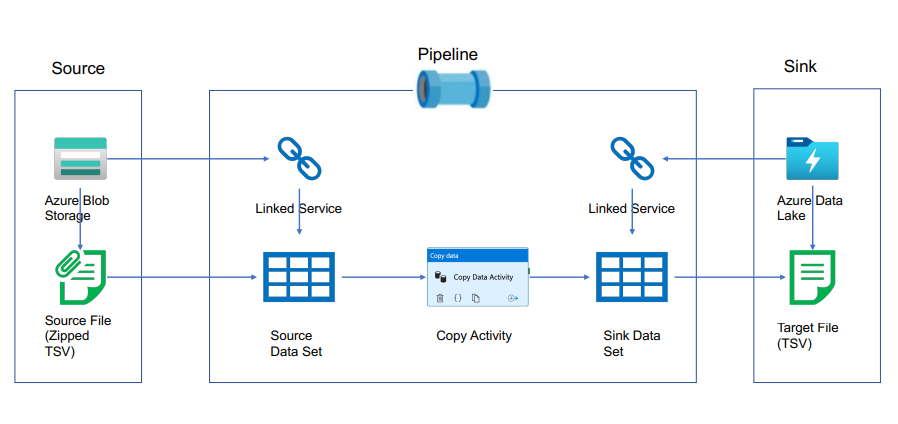


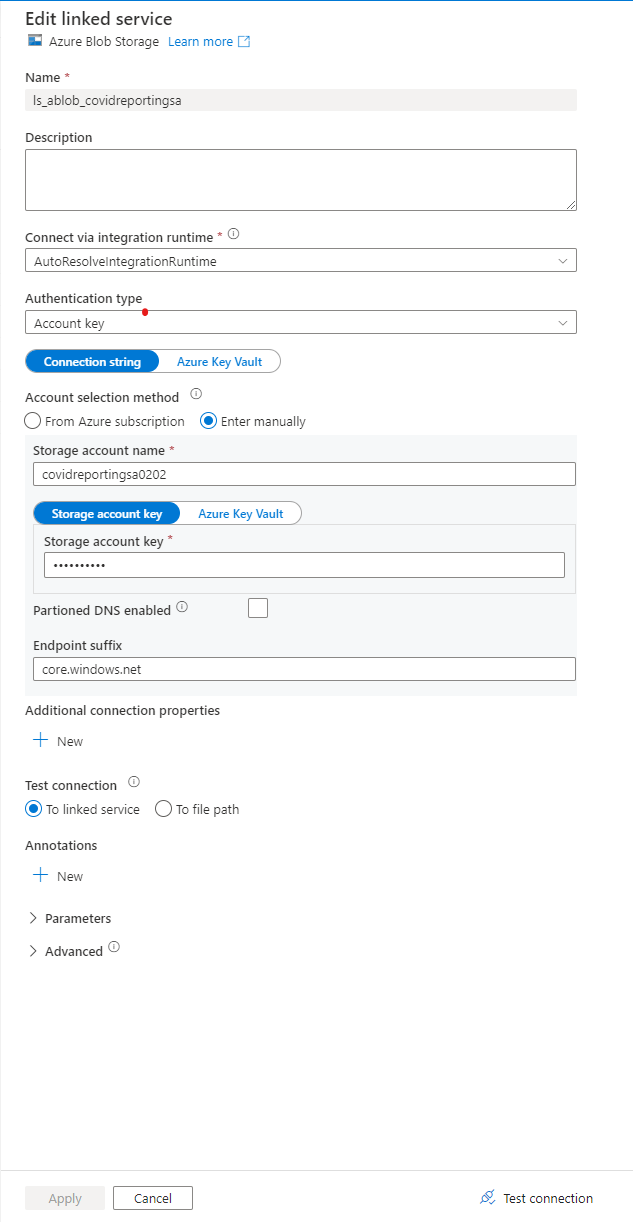
Worked with:

* Copy Activity
* Linked Services
* Datasets
* Validation Activity
* If Condition Activity
* Get Metadata Activity
* Web Activity
* Delete Activity
* Trigger

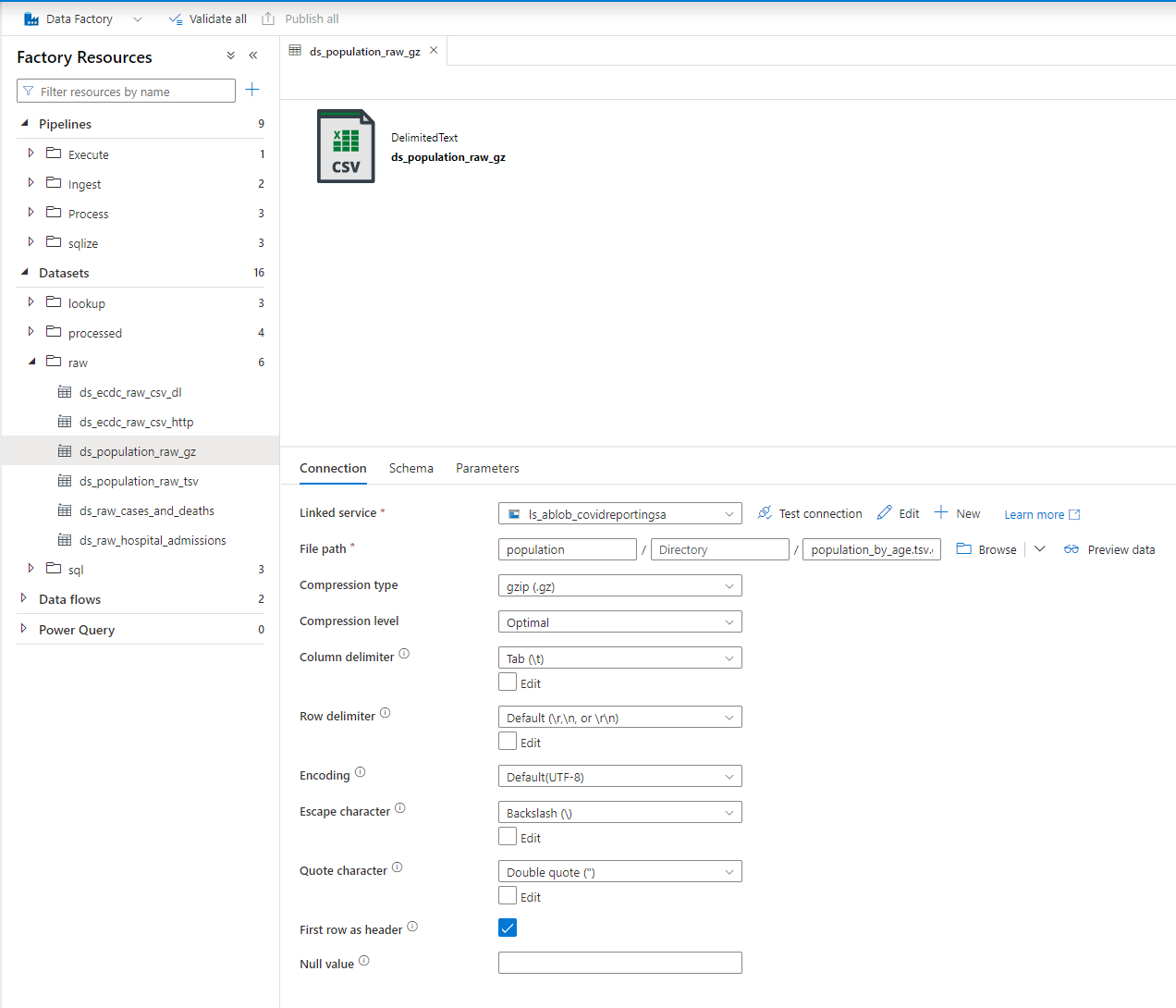
## Data ingestion with Azure blob storage

Copy activity overview

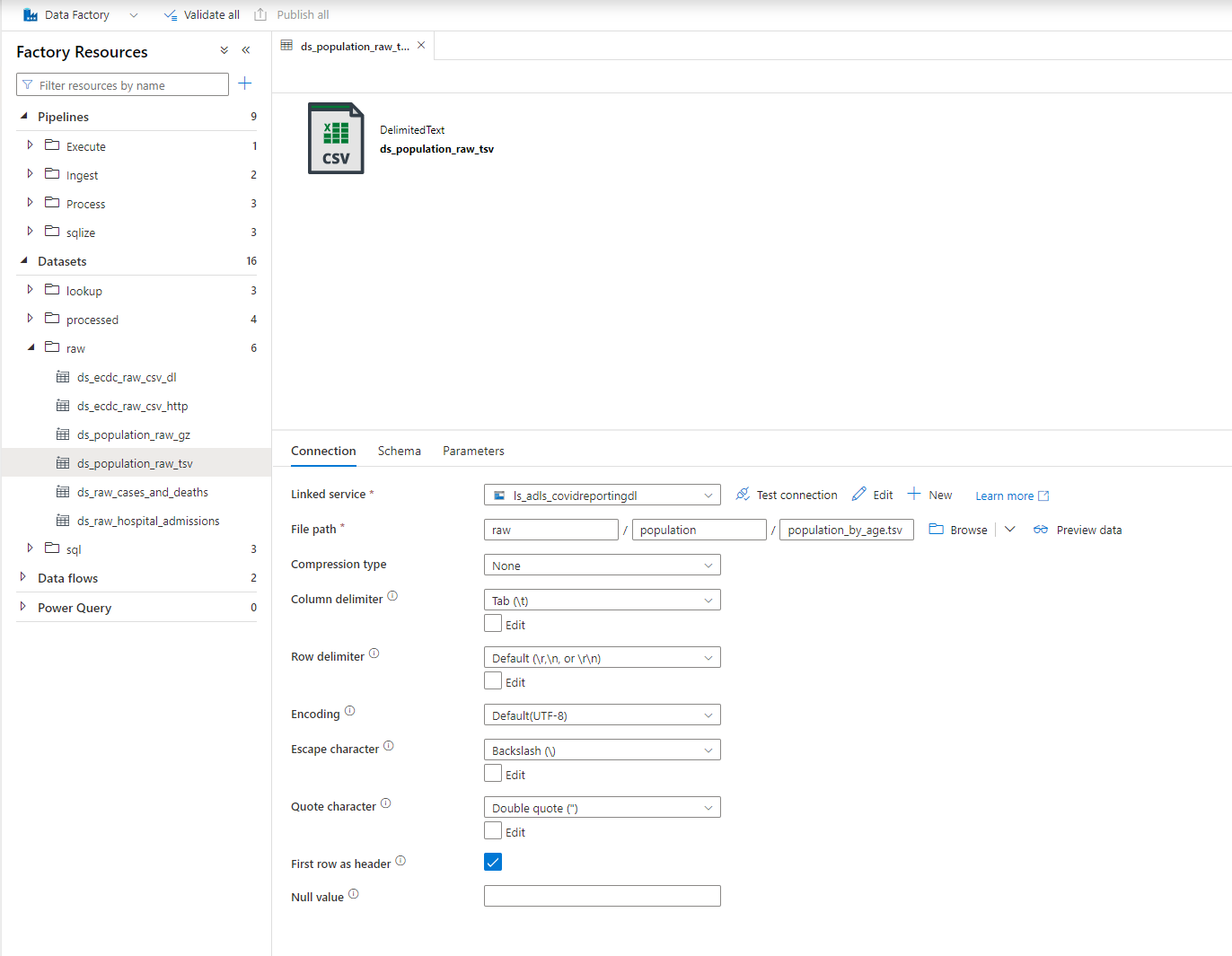
Linked service



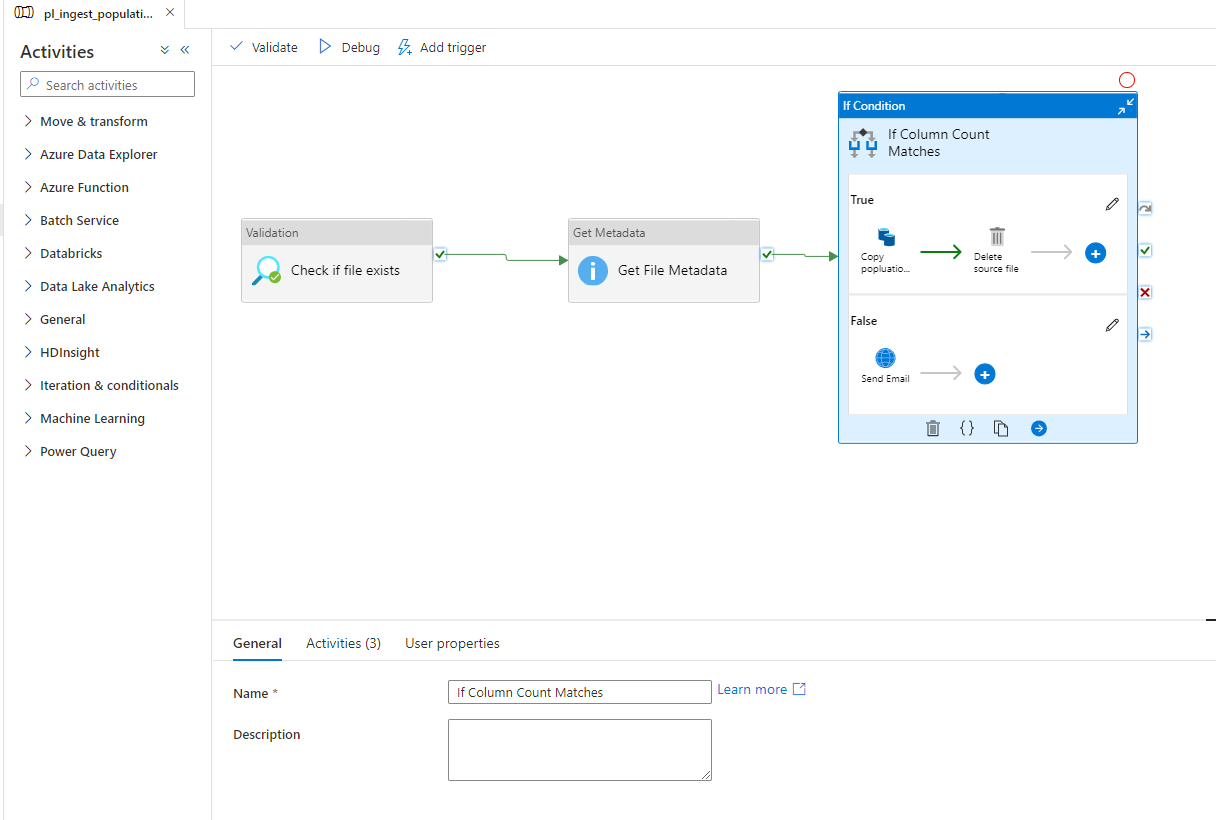
Blob data set (source)



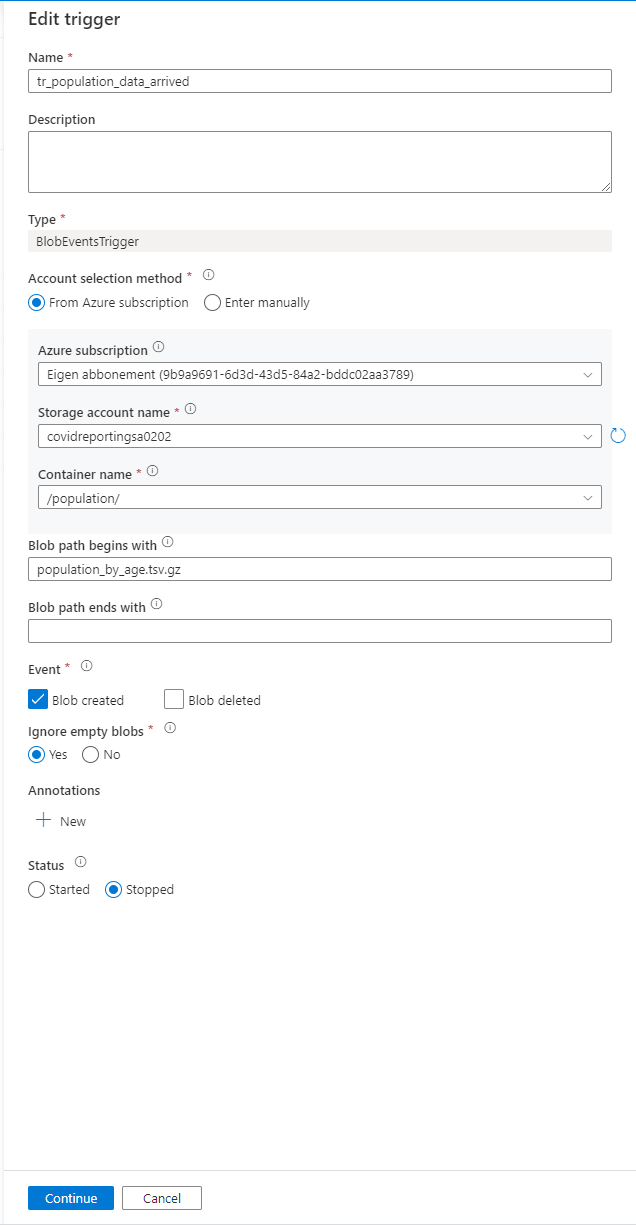
Data Lake data set (sink)



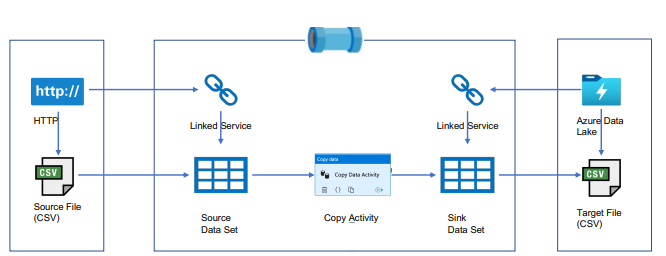
Creating an ingestion pipeline



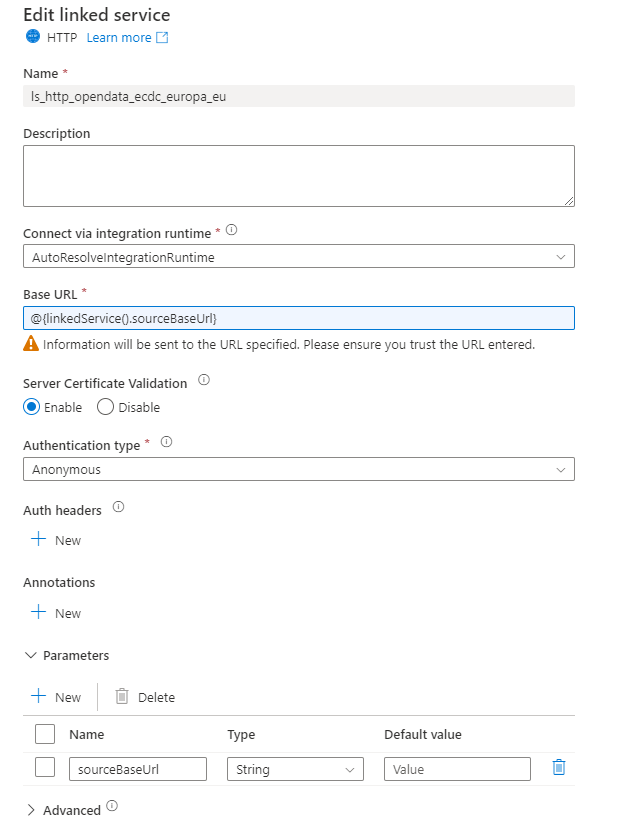
Trigger (ingestion) pipeline



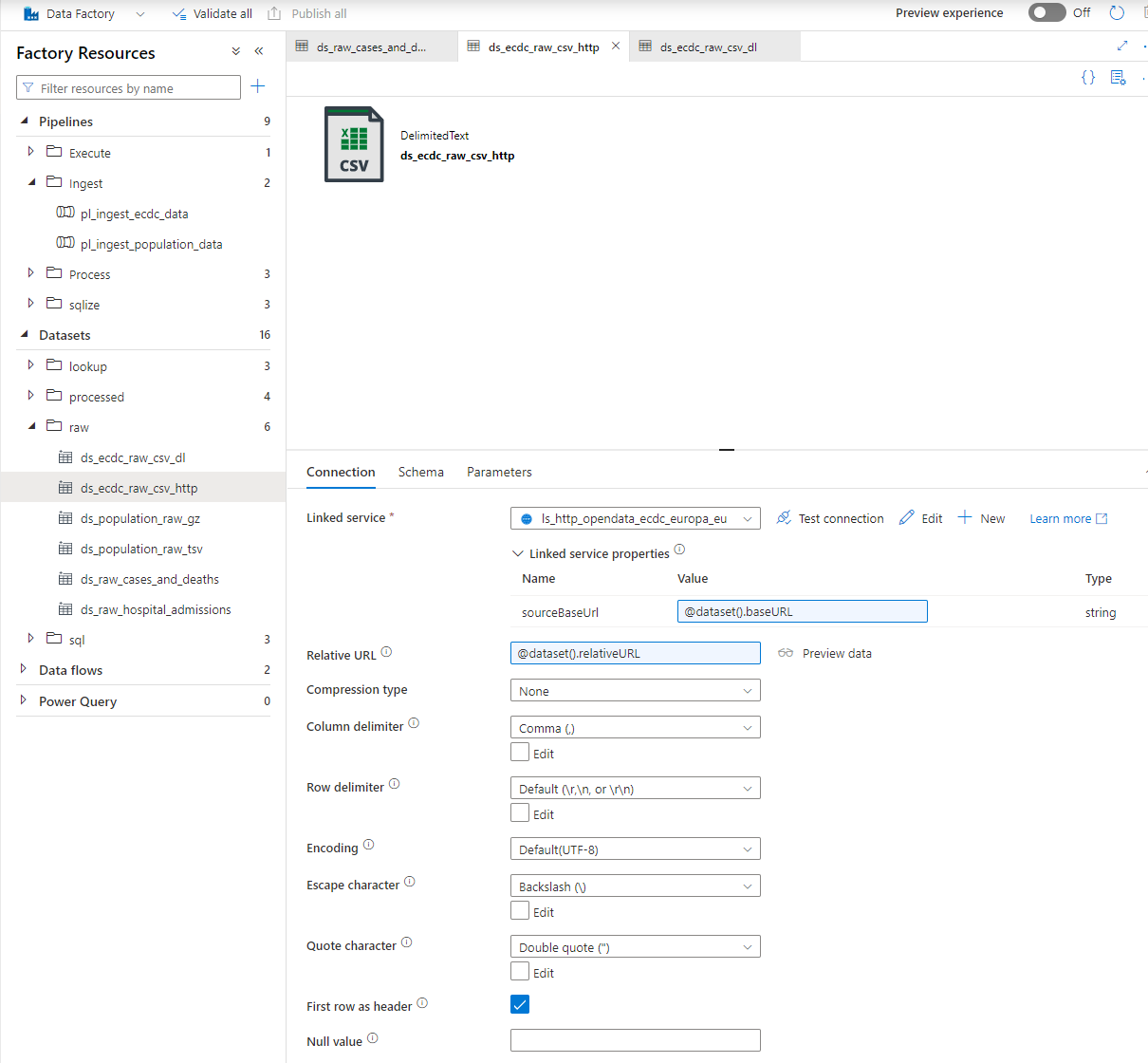
Data ingestion from HTTP:

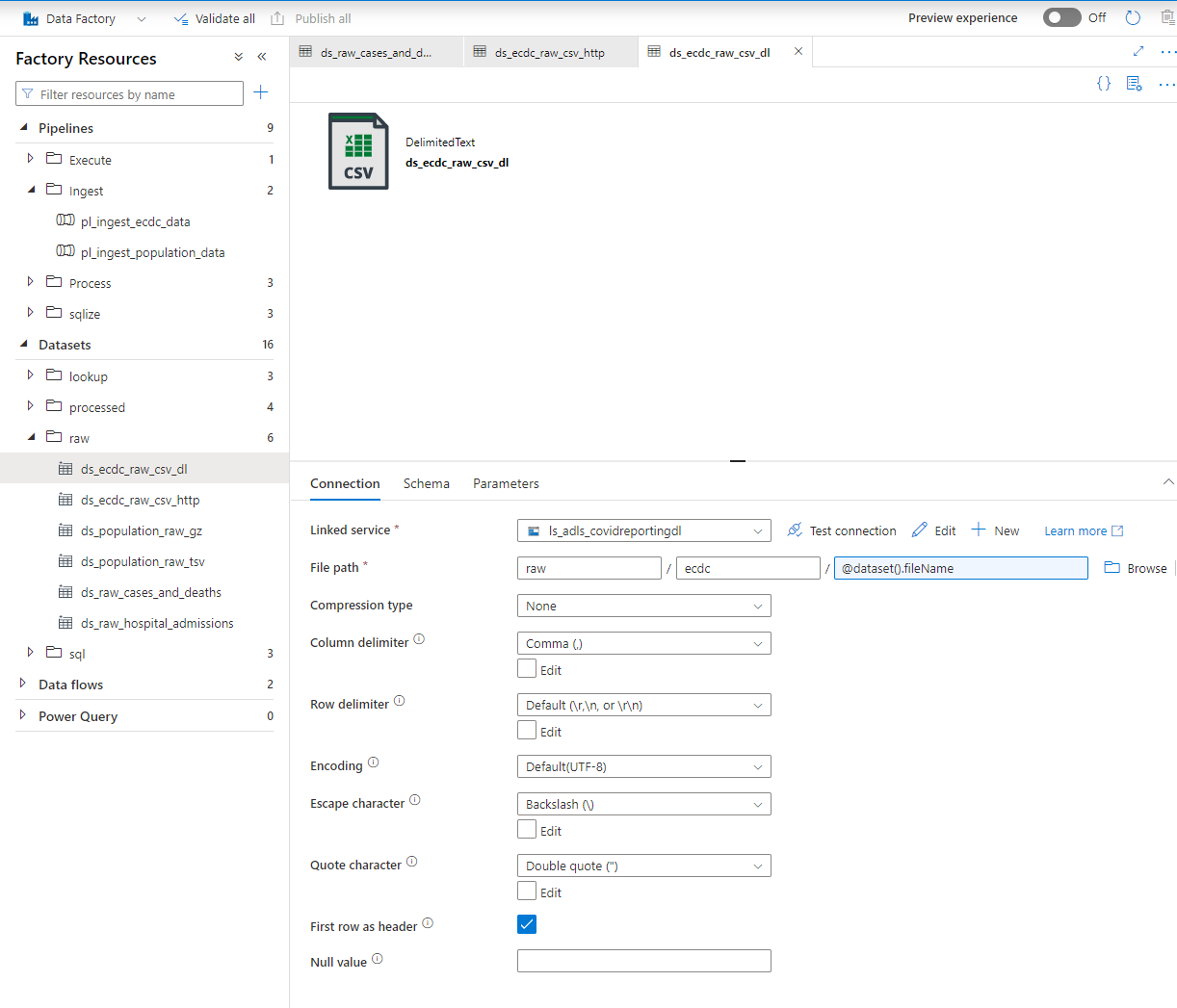
Overview copy activity from HTTP:

Step 1 - Created linked services:

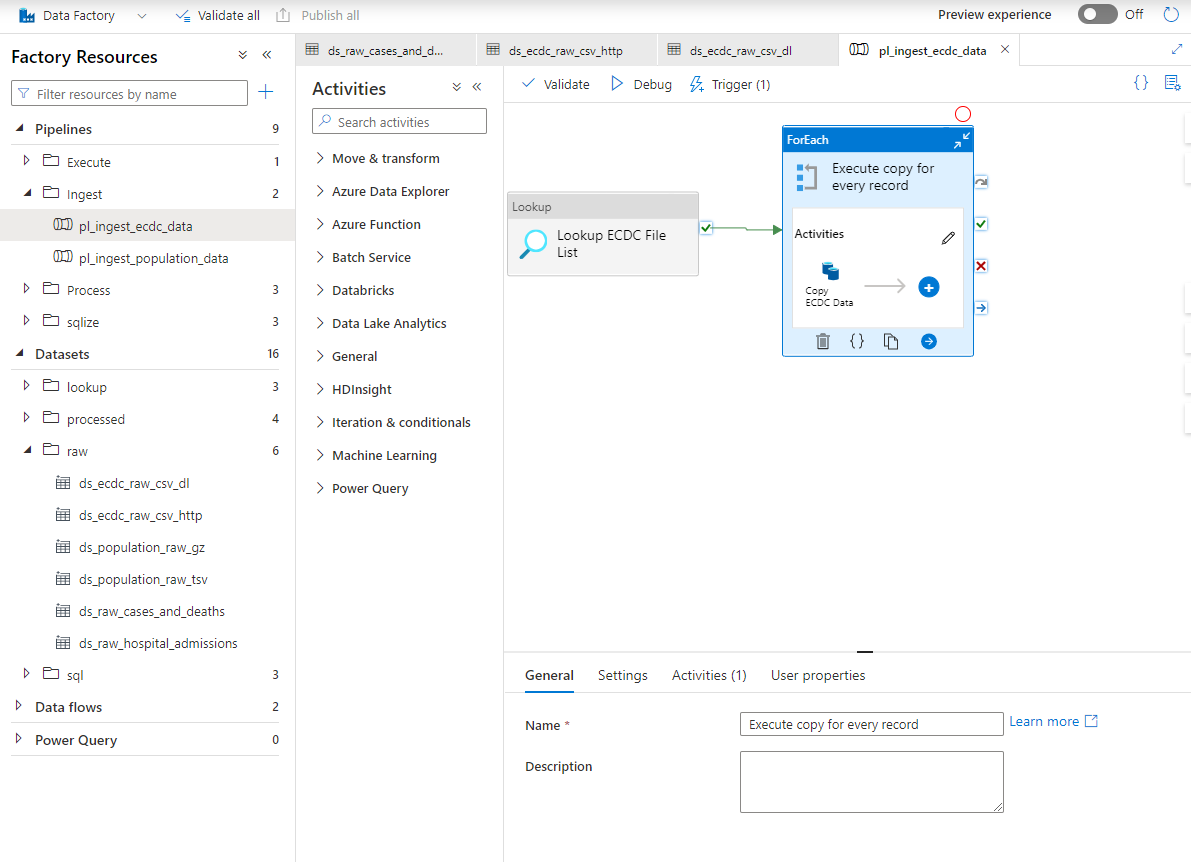


Step 2 - Create data set (http - source)

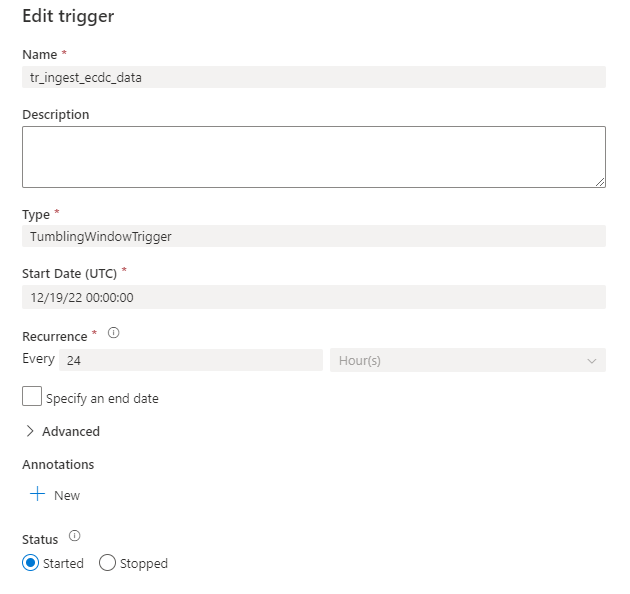


Step 3 - Create data set (DL sink) 

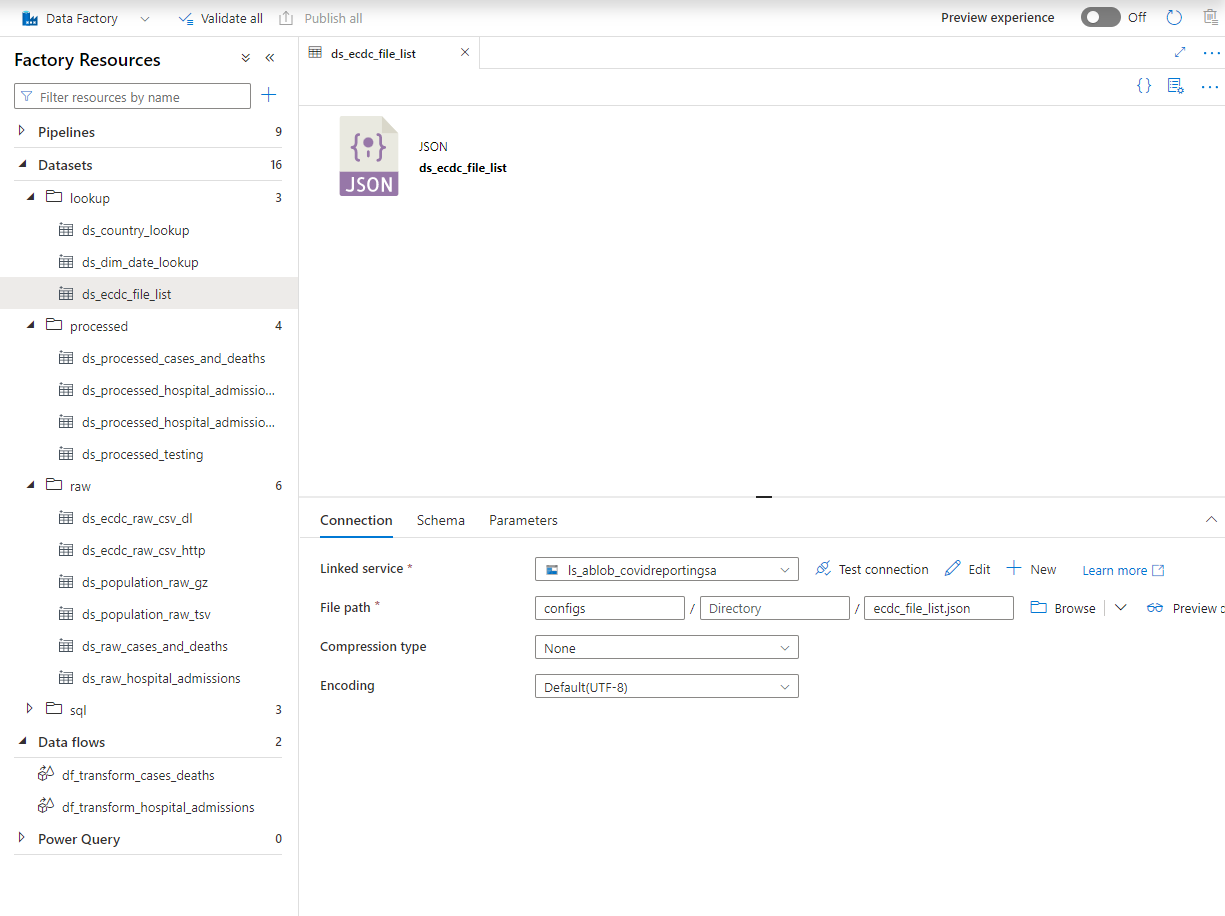
Step 4 - Create pipeline for data ingestion



Step 5 - Create trigger for pipeline



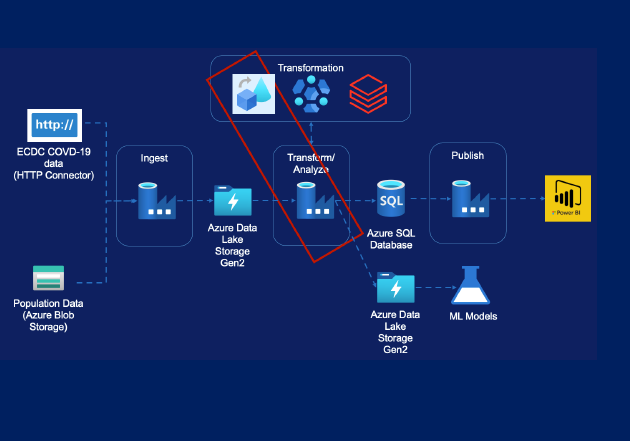
Step 6 - Ecdc file list data set



# Data transformation

Data flow

Step 0 – Overview transformation in data flow



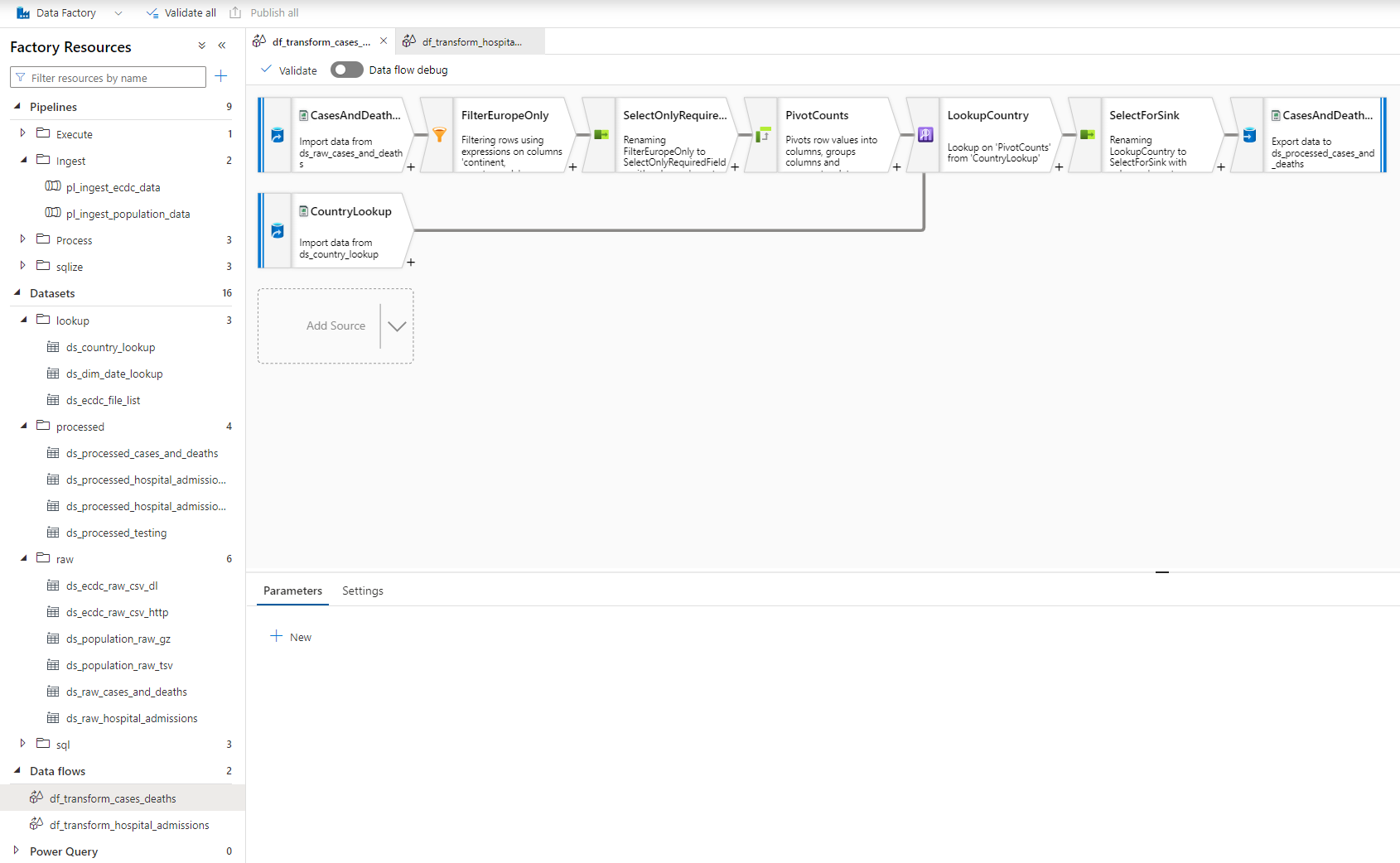
(1) Steps taken:

* Source Transformation
* Filter Transformation
* Select Transformation
* Pivot Transformation
* Lookup Transformation
* Sink Transformation
* Create Pipeline

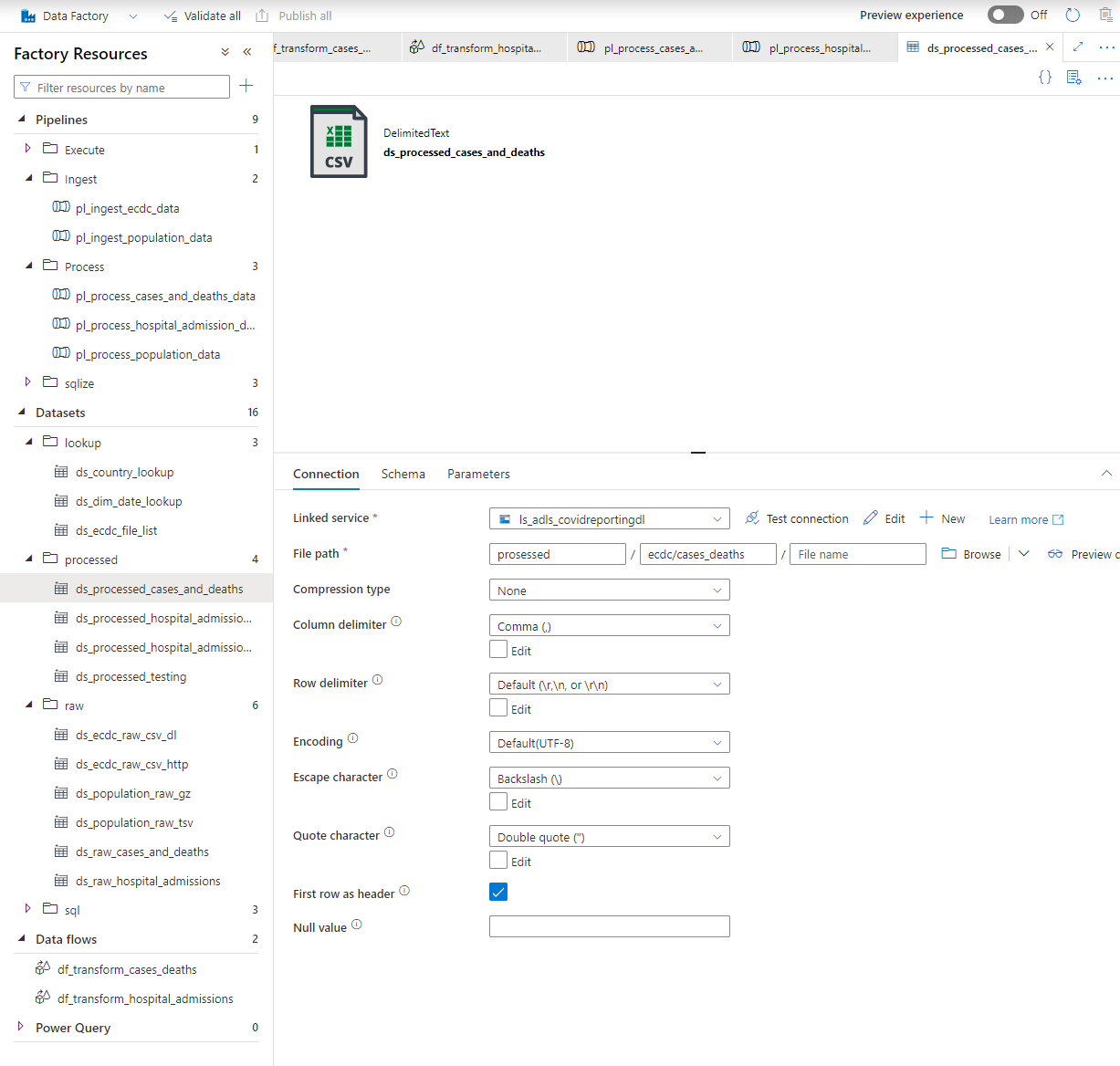
(2) Steps taken:

* Source Transformation
* Select Transformation
* Lookup Transformation
* Pivot Transformation
* Sink Transformation
* Conditional Split Transformation
* Derived Column Transformation
* Aggregate Transformation
* Sort Transformation
* Join Transformation
* Select Transformation
* Create Pipeline

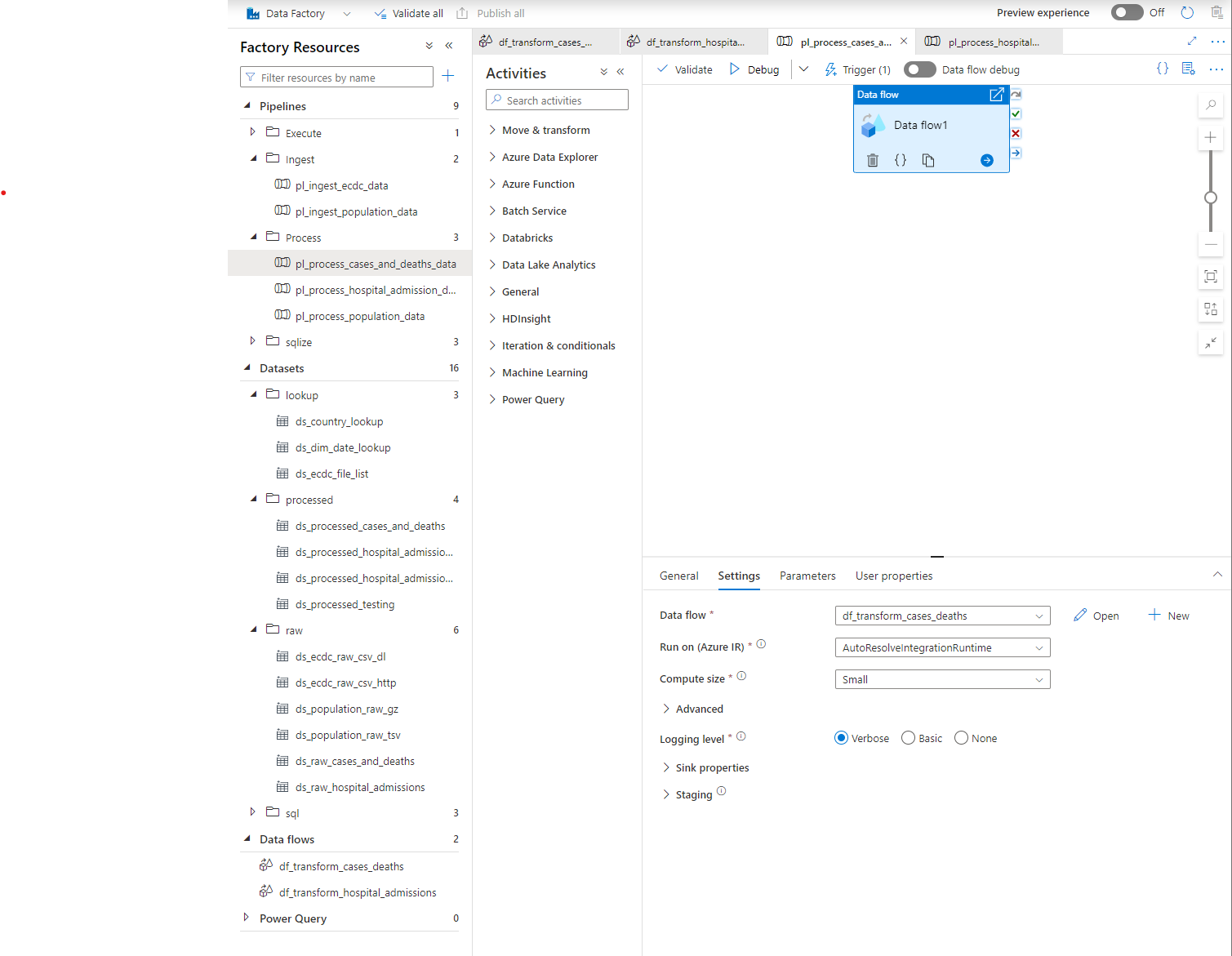
(1) Step 1 - Create transformations with data flow



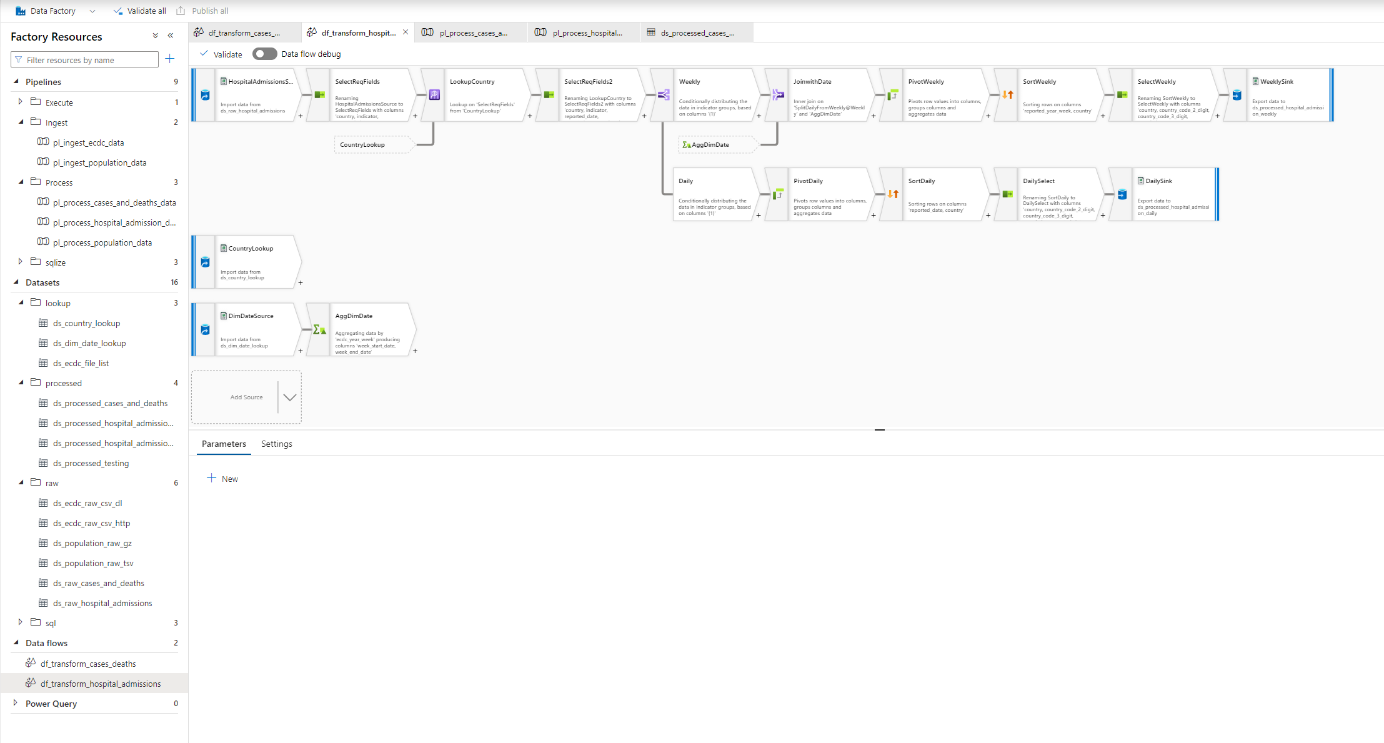
(1) Step 2 - Create data set for processed data



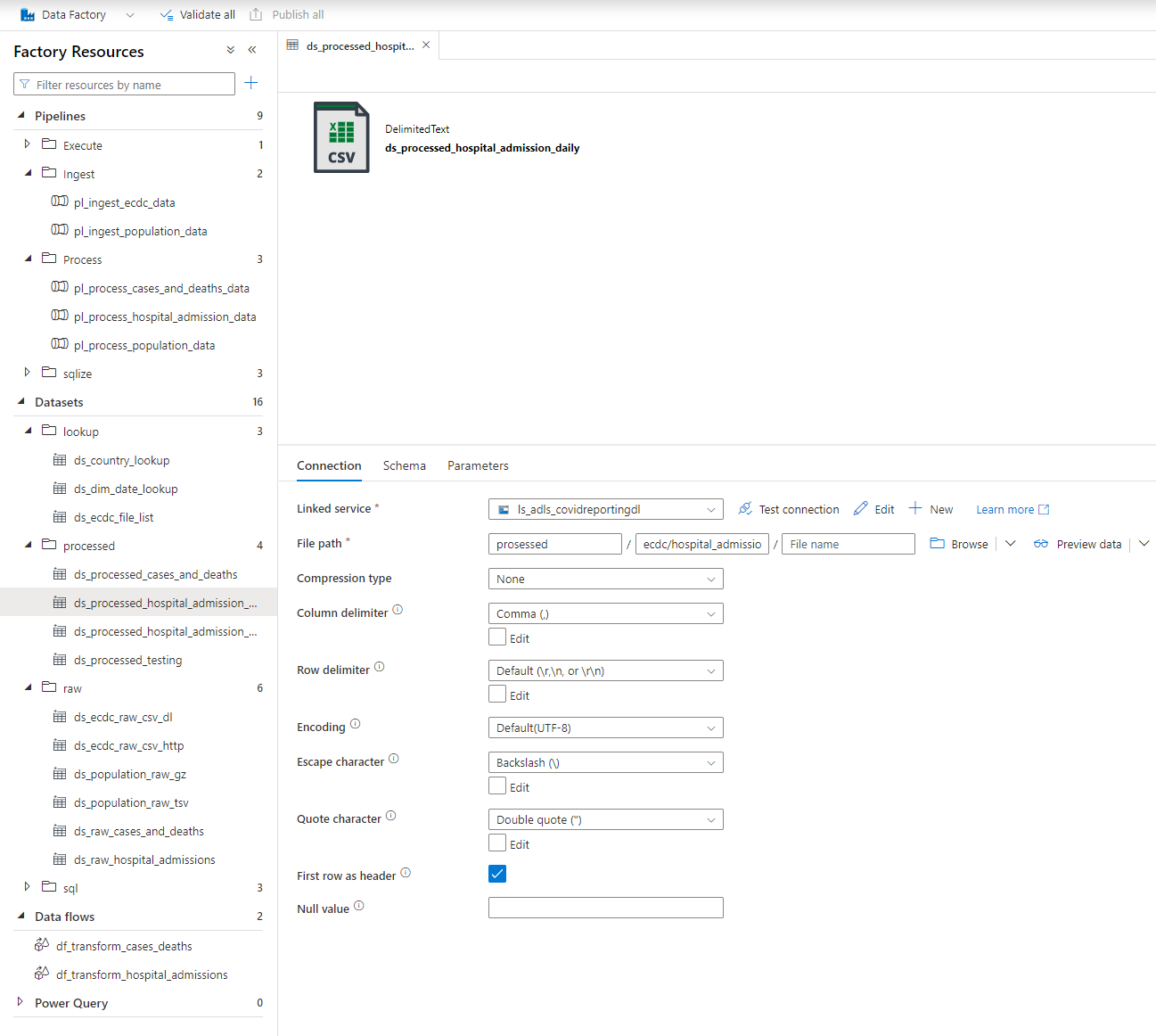
(1) Step 3 – Create pipeline



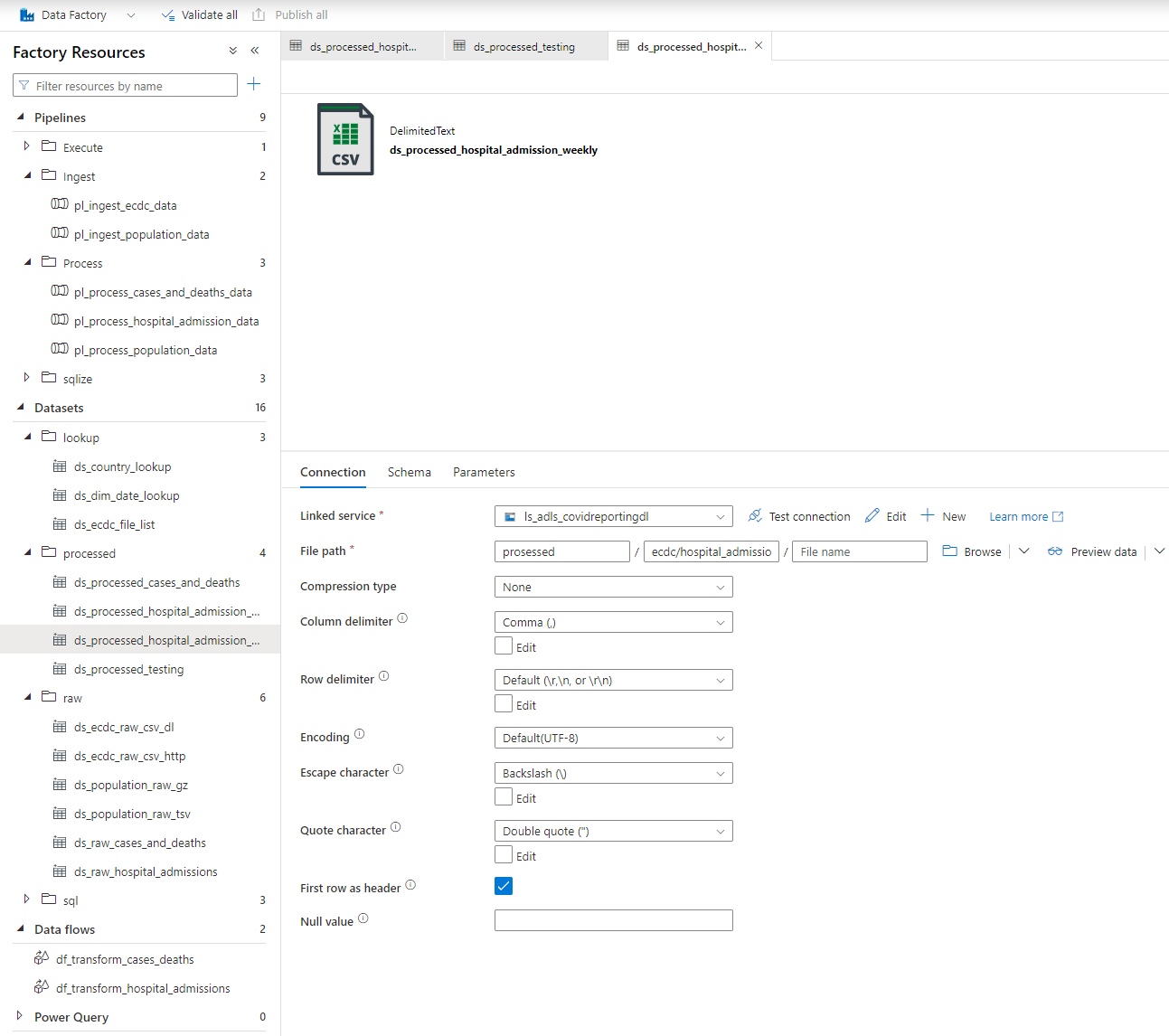
(2) Step 4 - Create transformations with data flow



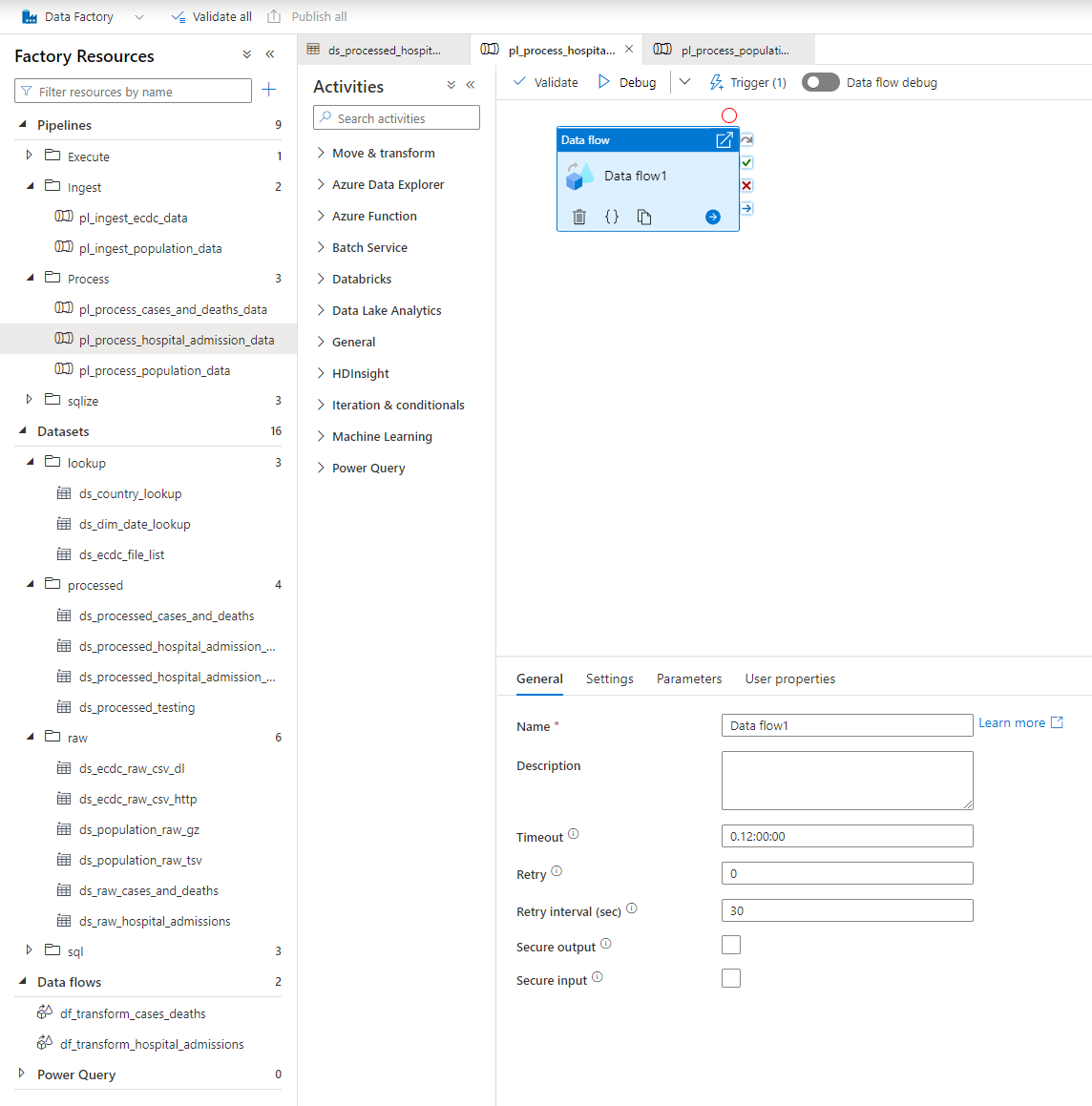
Step 5 - Create data set for processed data (daily)



Step 6 - Create data set for processed data (weekly)

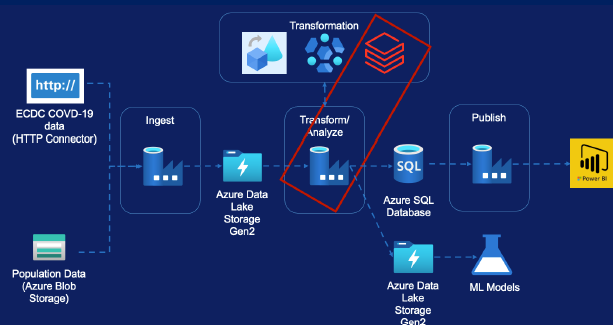


Step 7 - Create pipeline



Databricks

Step 0 - Databricks overview



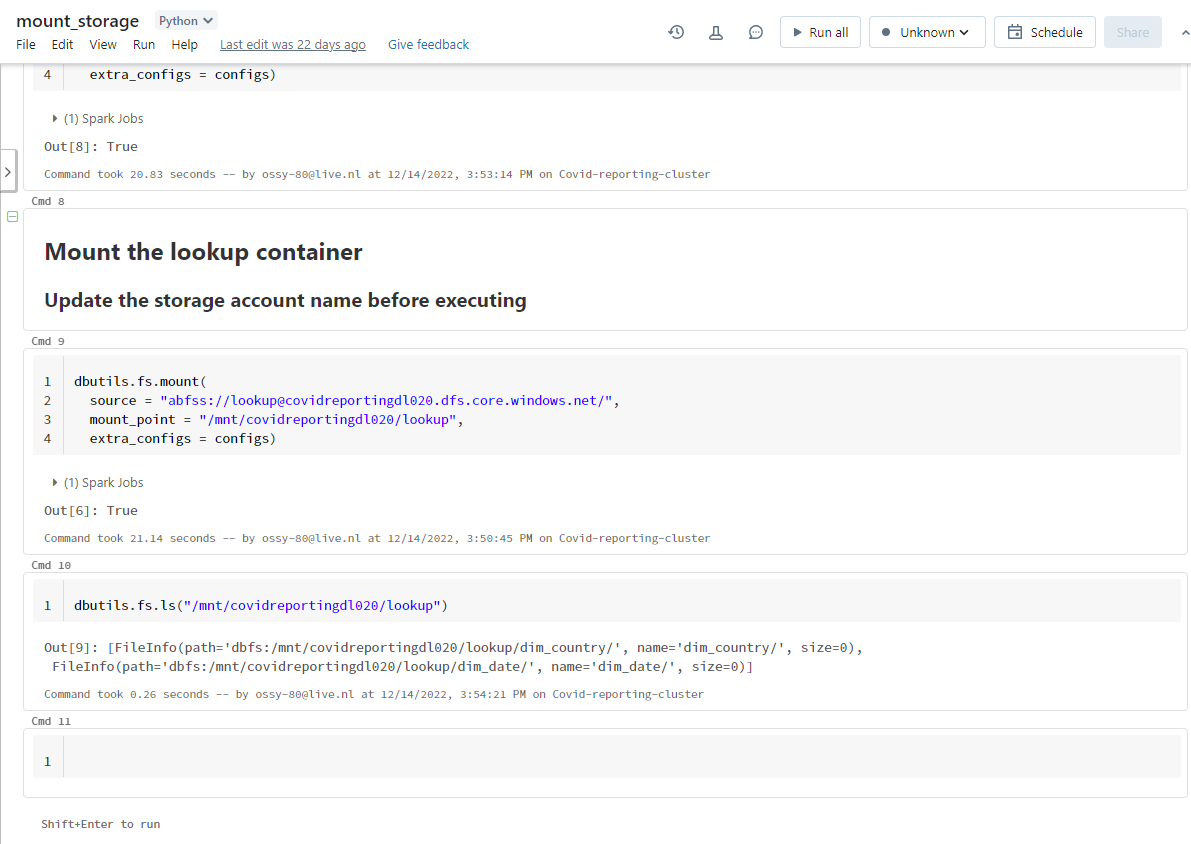
Steps taken:

* Create Databricks Service
* Create Databricks Cluster
* Mount Storage Accounts
* Transformation requirements
* Creating Pipeline

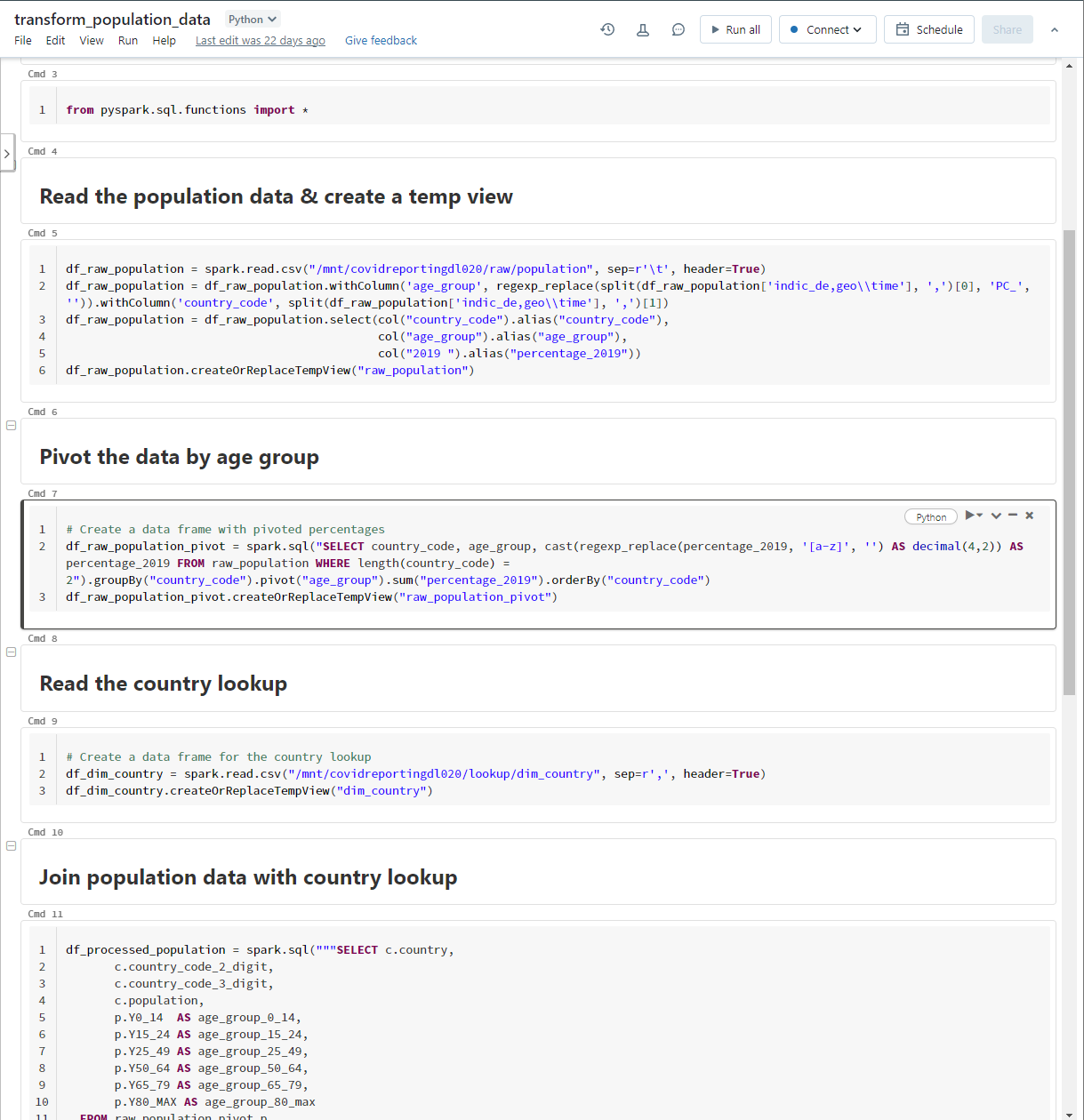
Step 1 – Mount cluster 1/2



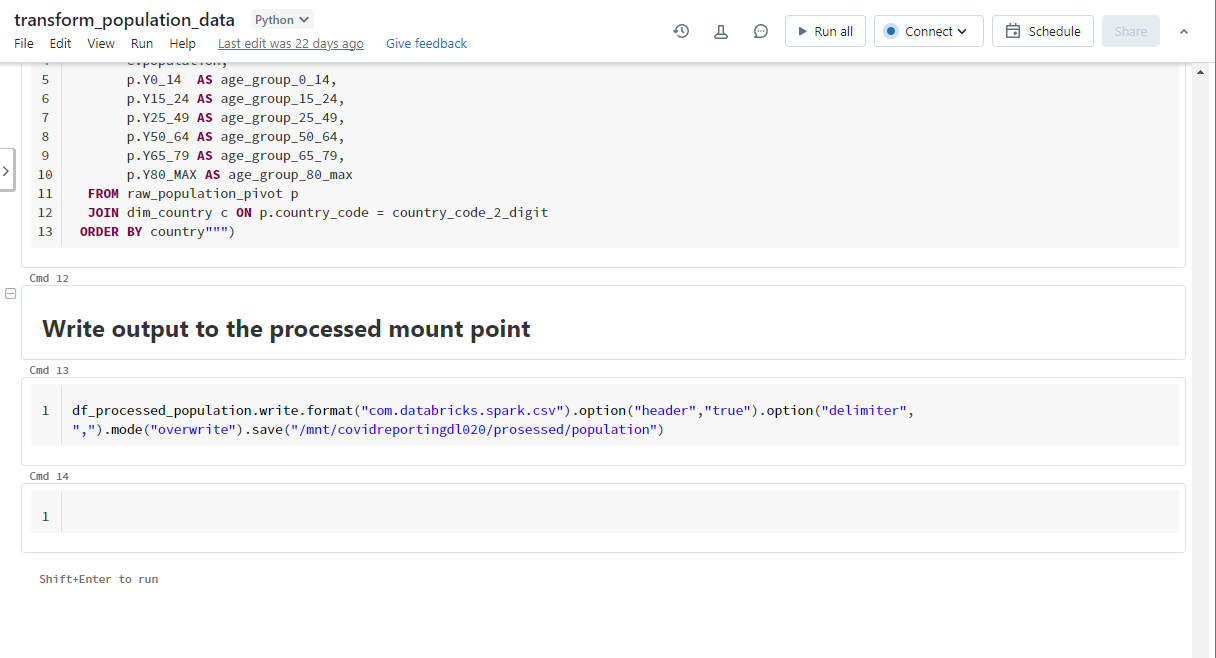
Step 1 – Mount cluster 2/2



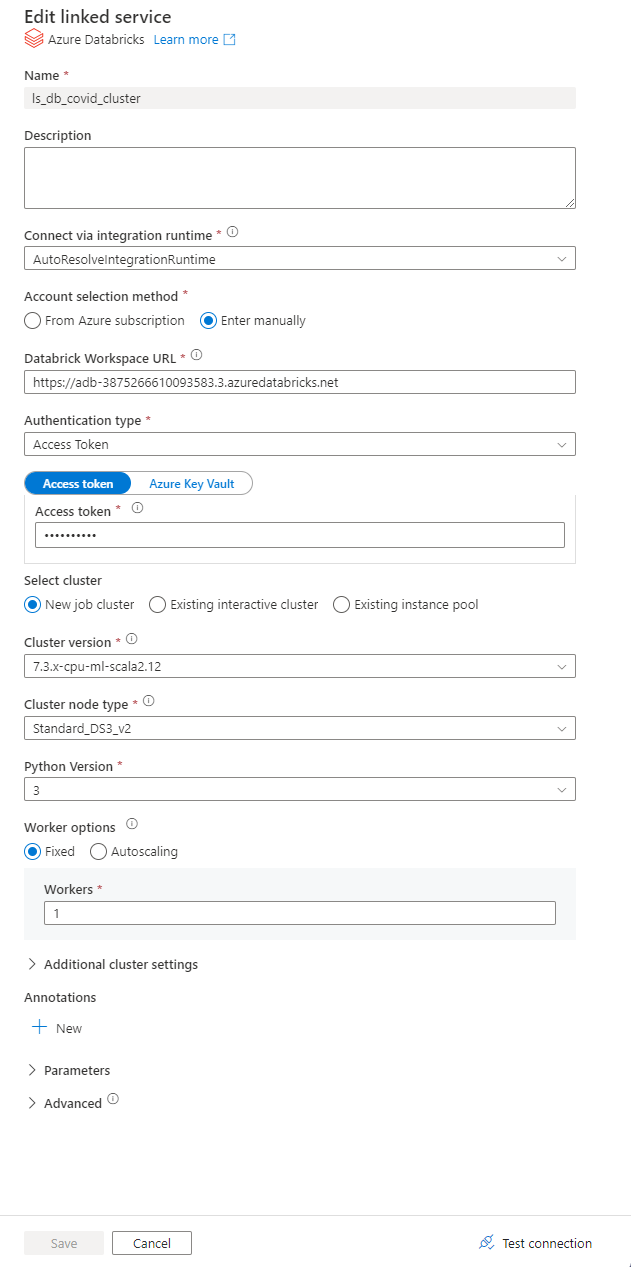
Step 2 – transform data ½



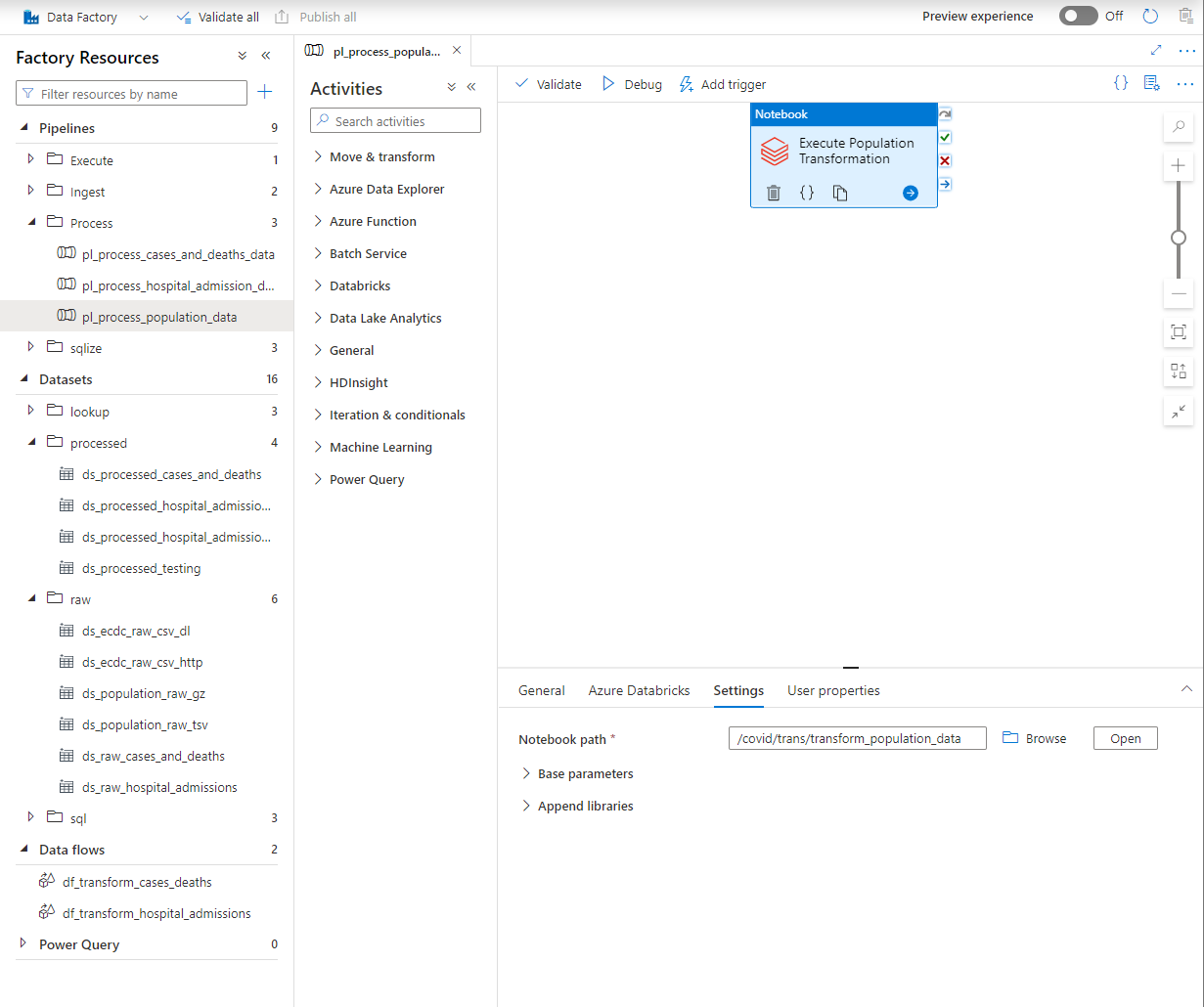
Step 2 – transform data 2/2



Step 3 – Create linked service

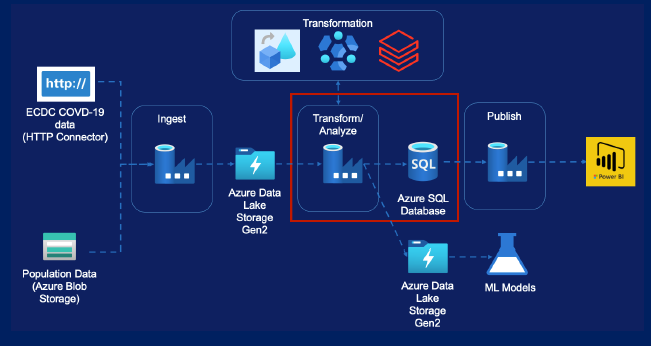


Step 4 – Create pipeline



# Load into database

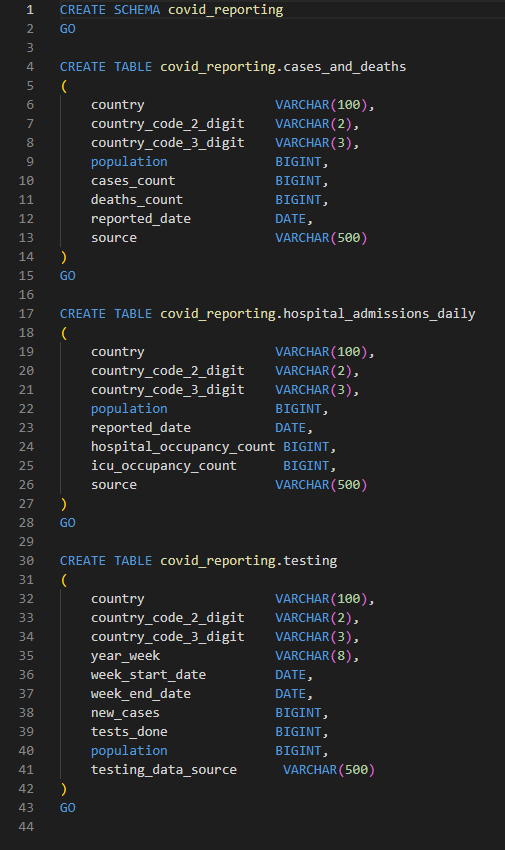
Step 0 – Overview SQL database



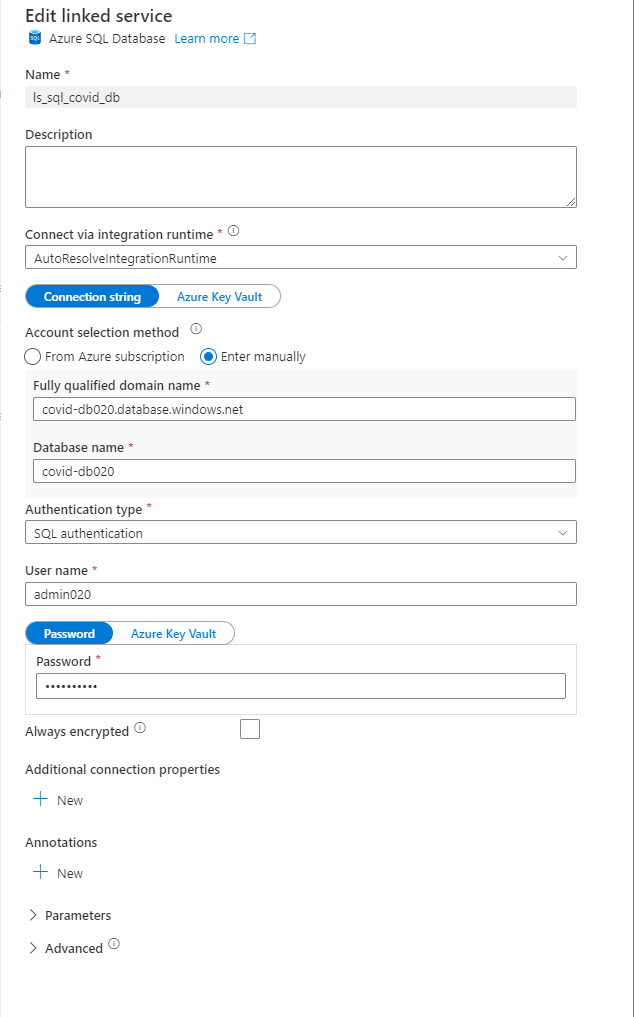
Steps taken:

* Copy Cases & Deaths data
* Copy Hospital Admissions data
* Copy testing data

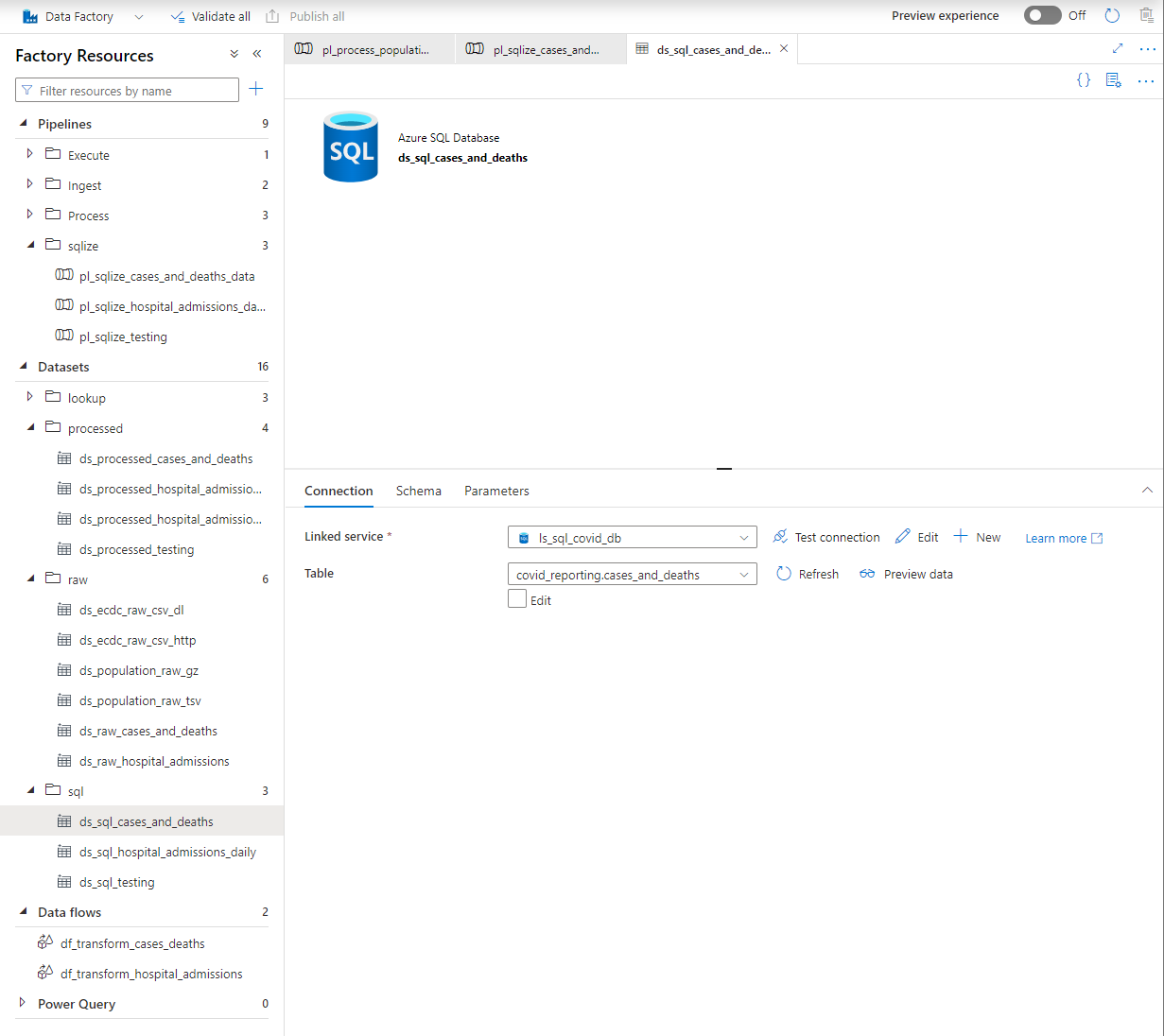
Step 1 – Create SQL script



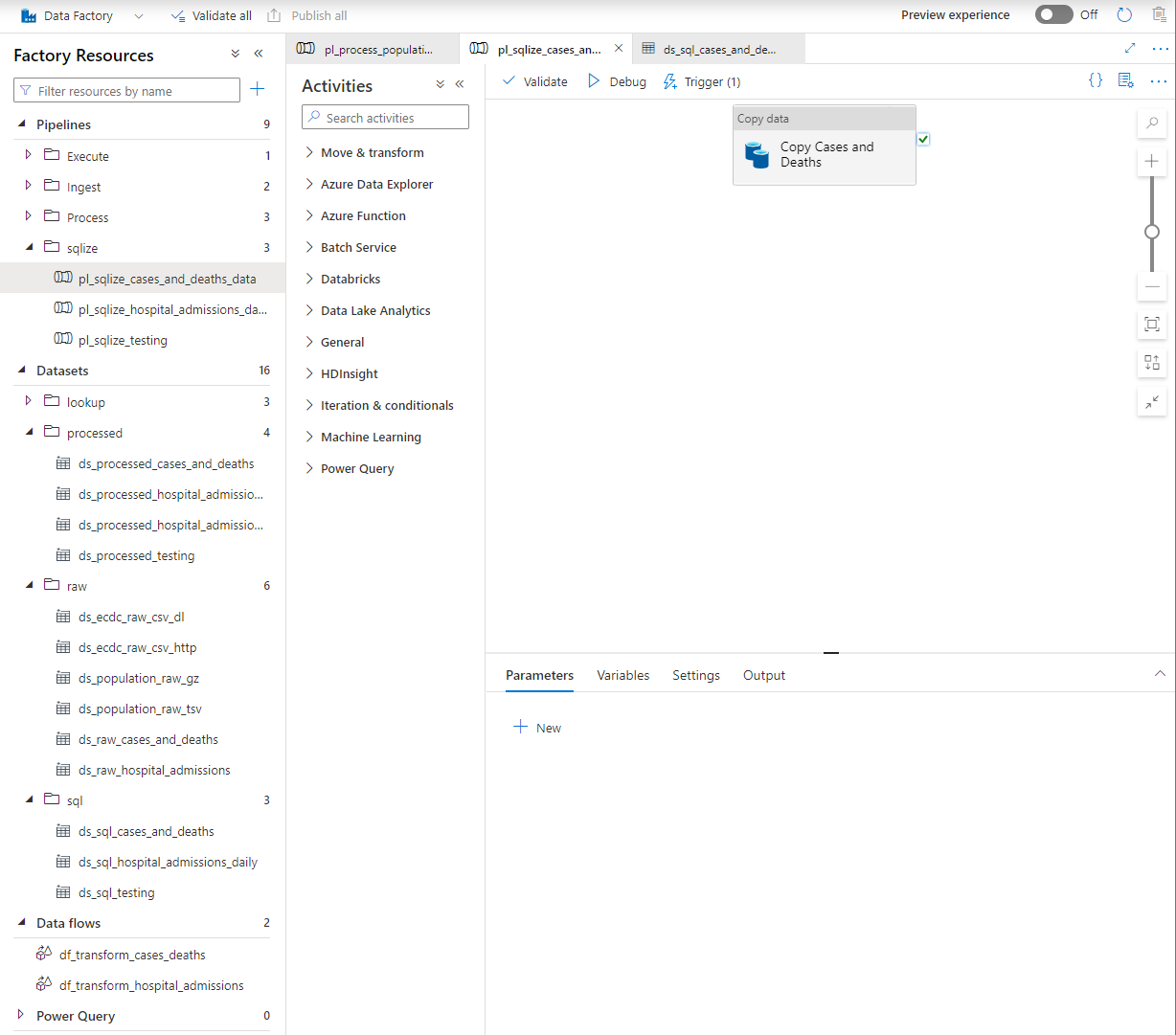
Step 2 – Create linked service



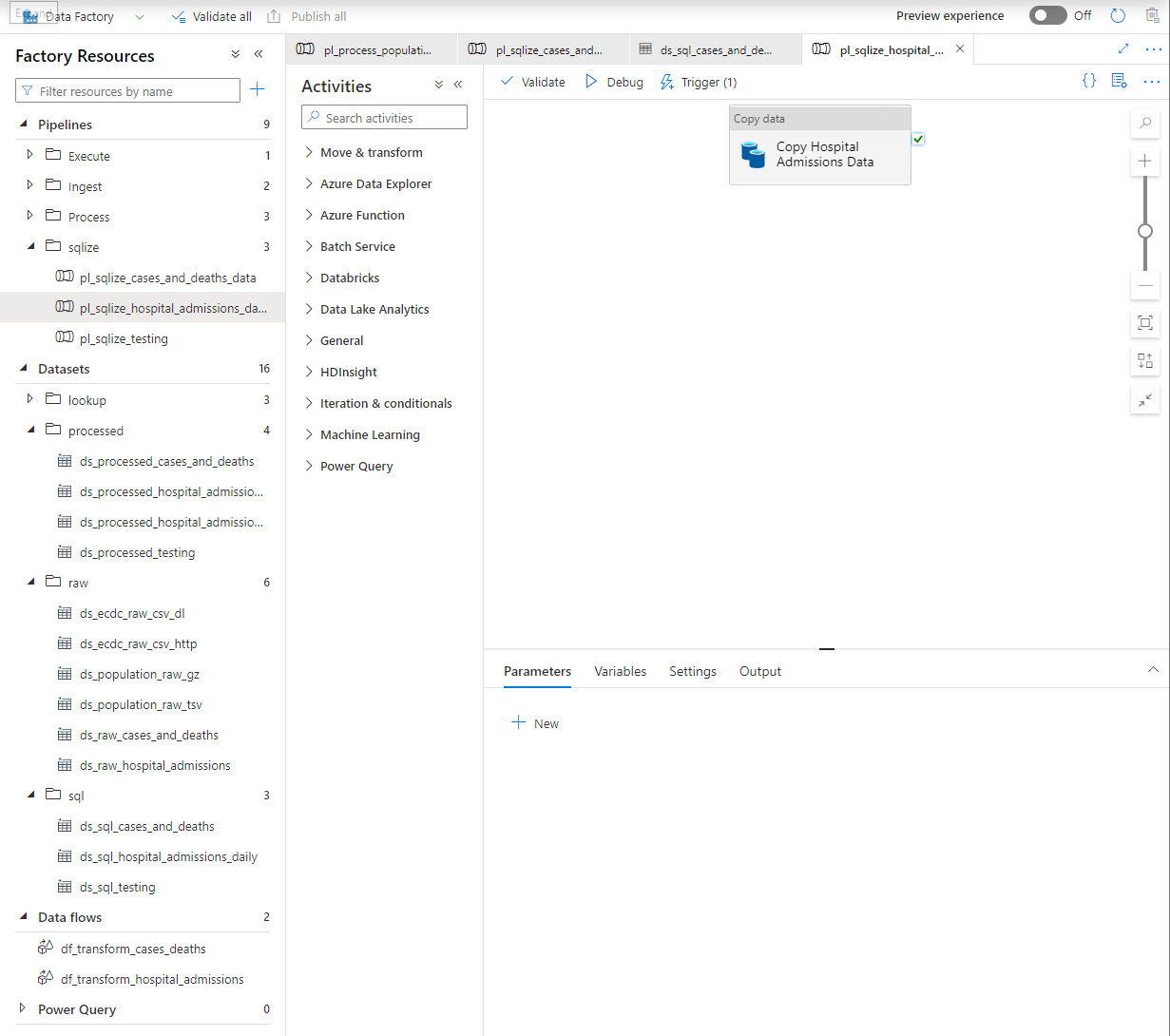
Step 3 - Create sink dataset



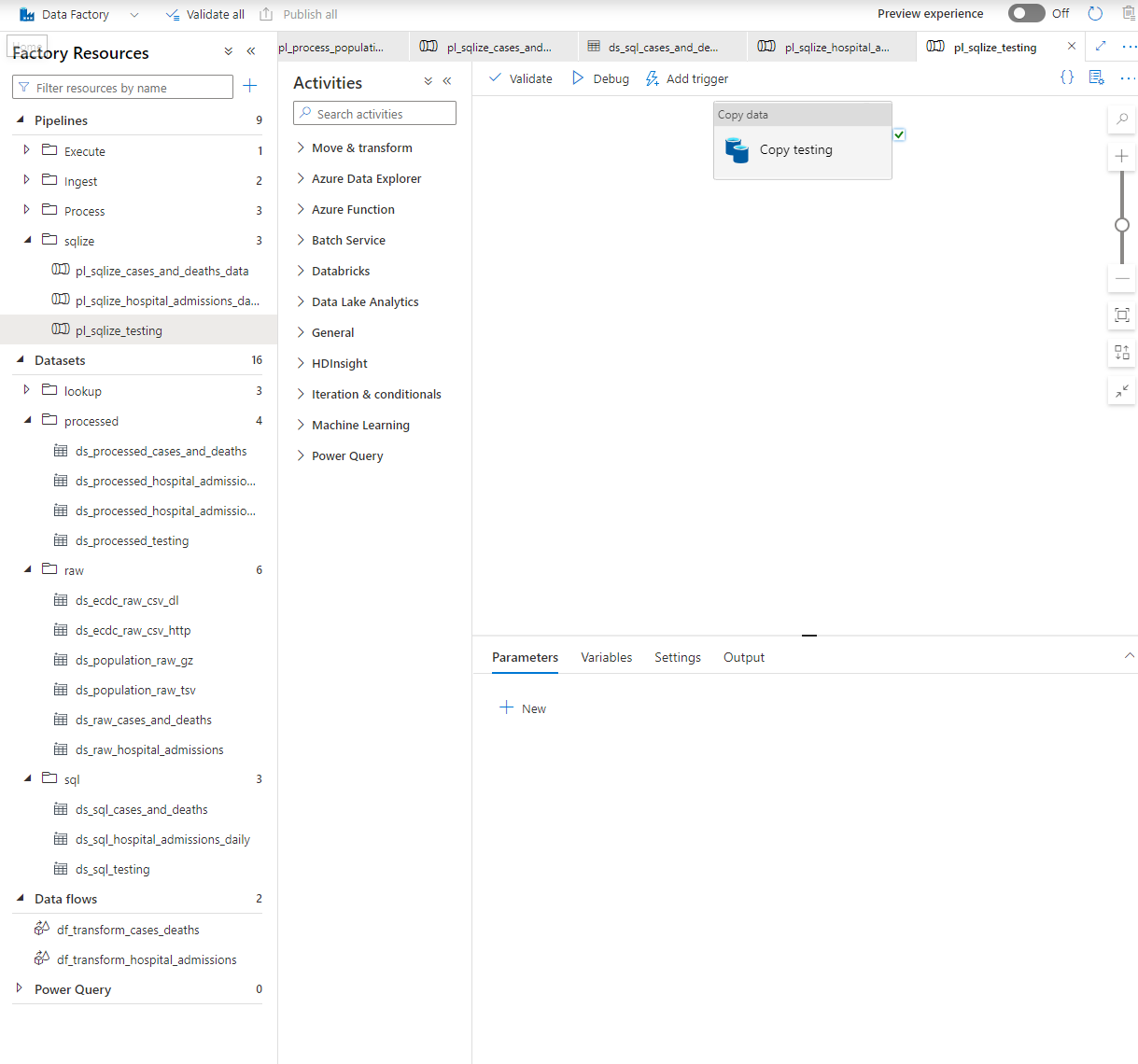
Step 4 – Create pipeline



Step 5 – Create pipeline hospital admission data

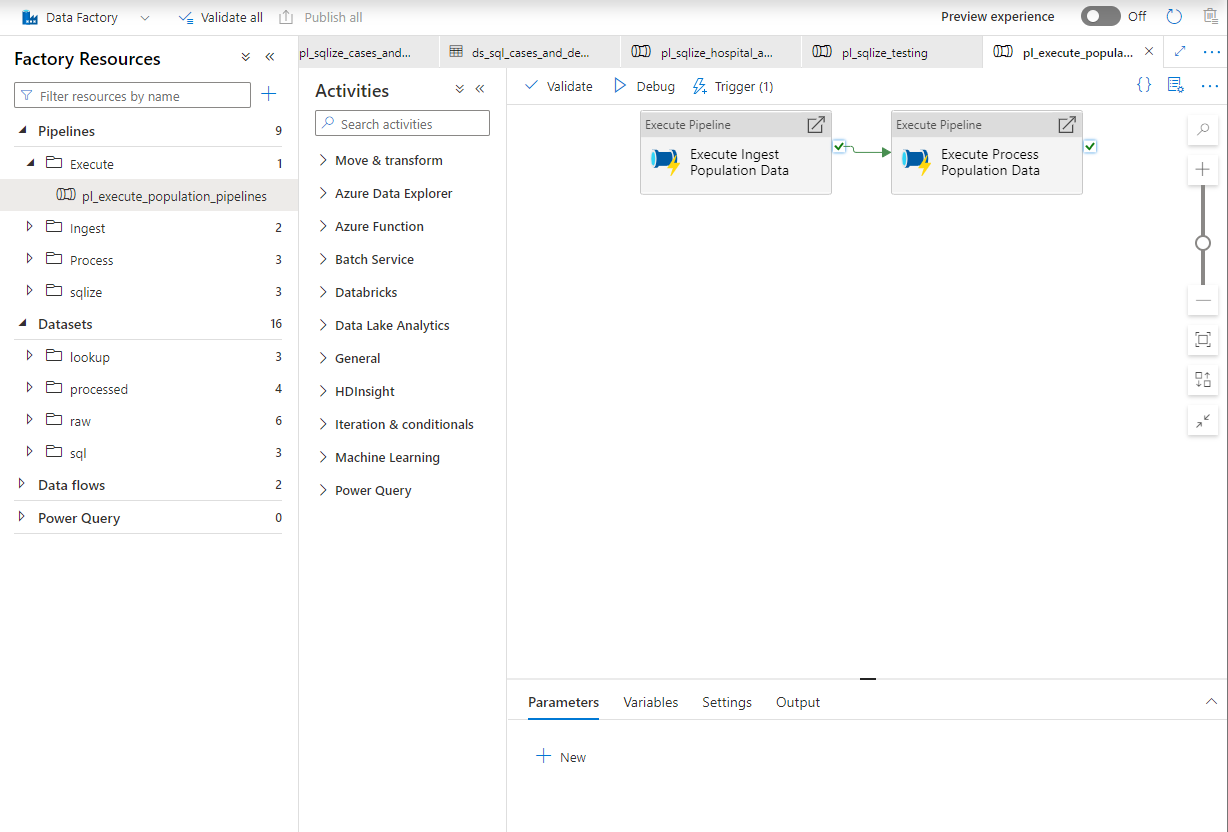


Step 6 - Create pipeline testing data

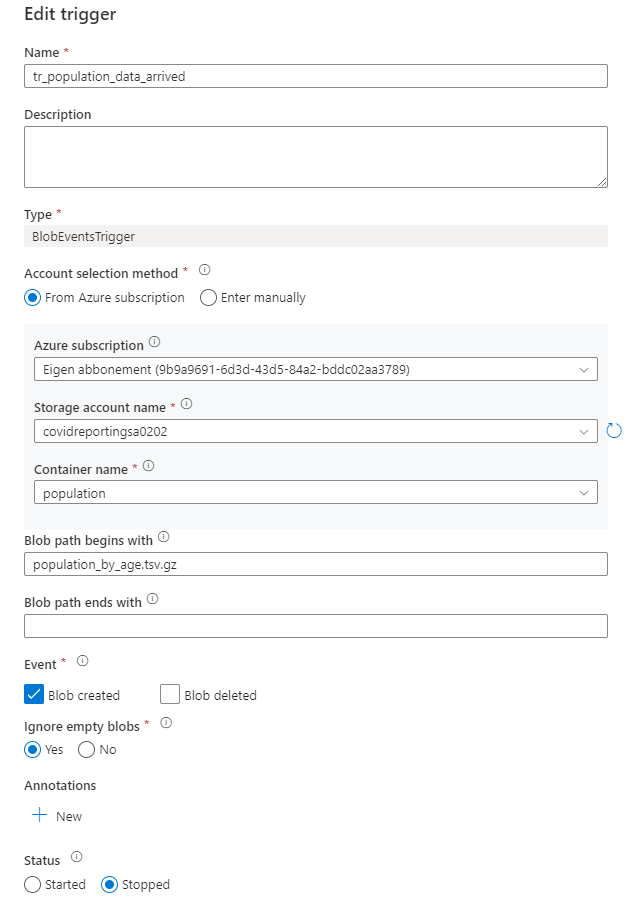


# Data orchestration – Making pipelines production ready

Step 1 – Build pipeline

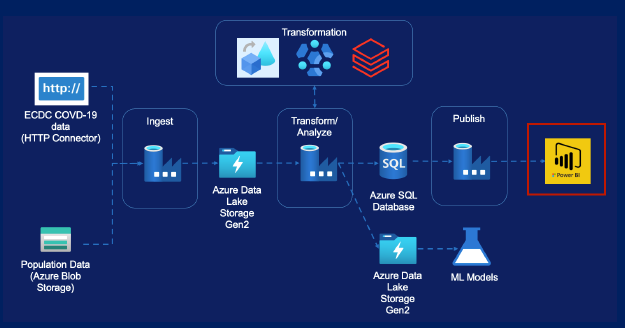


Step 2 – Create trigger

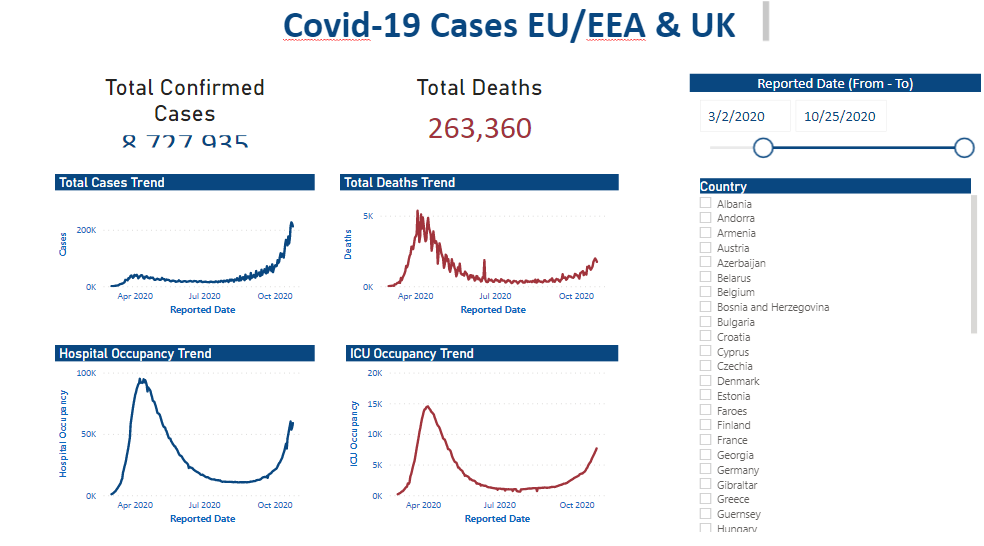


# Data reporting

Step 0 – Reporting overview in Power BI



Step 1 – Create report 1 / 2

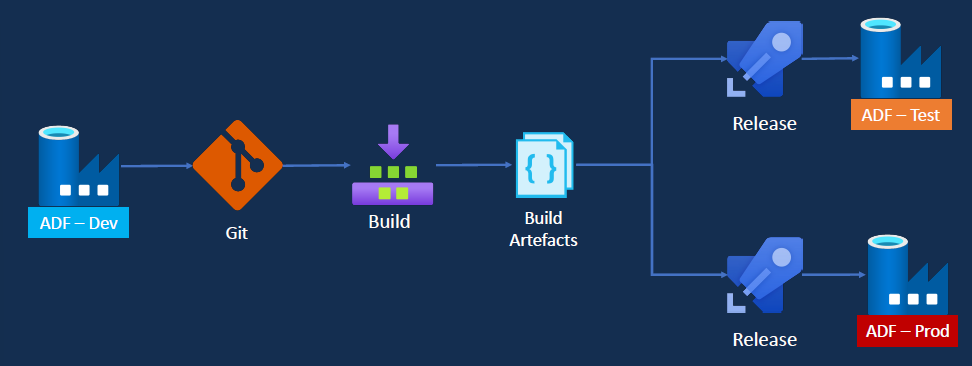


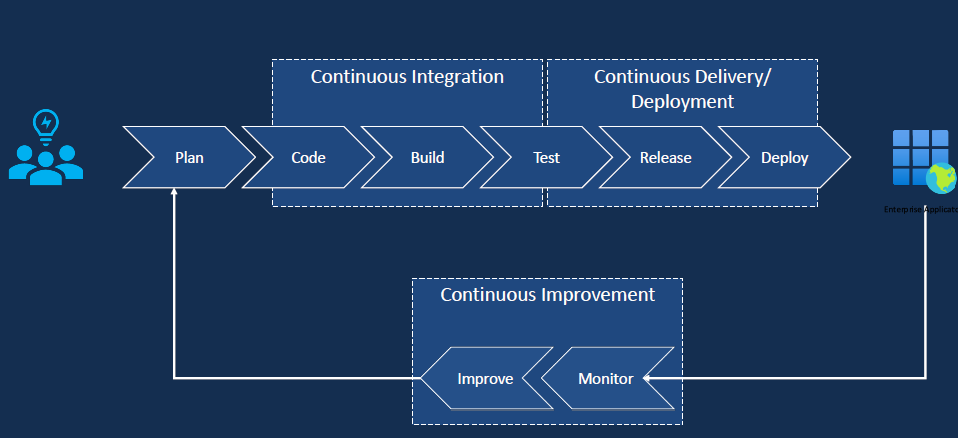
Step 1 – Create report 2 / 2



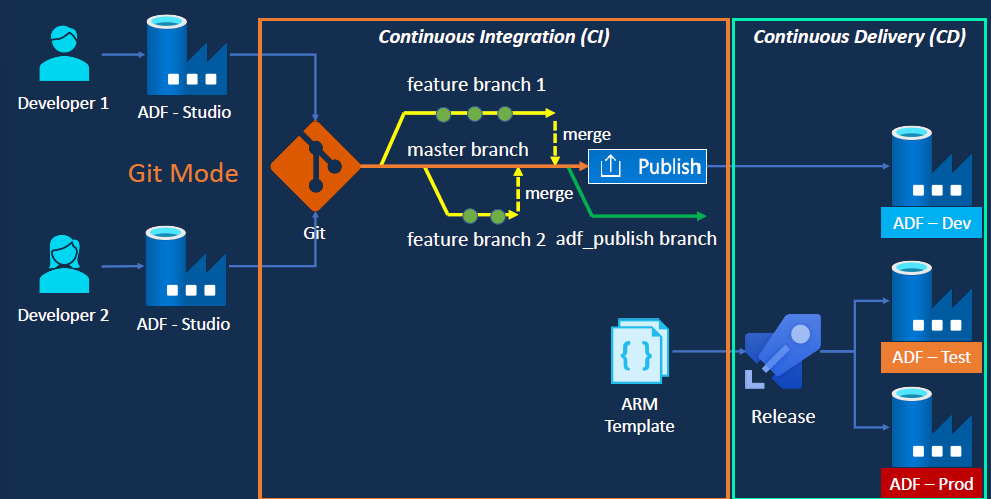
CI / CD

Step 0 – CI / CD overview

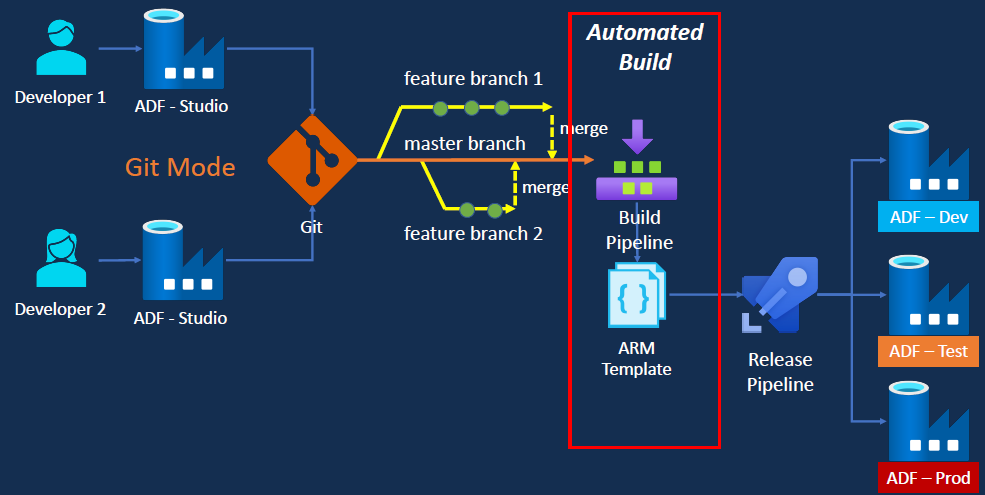




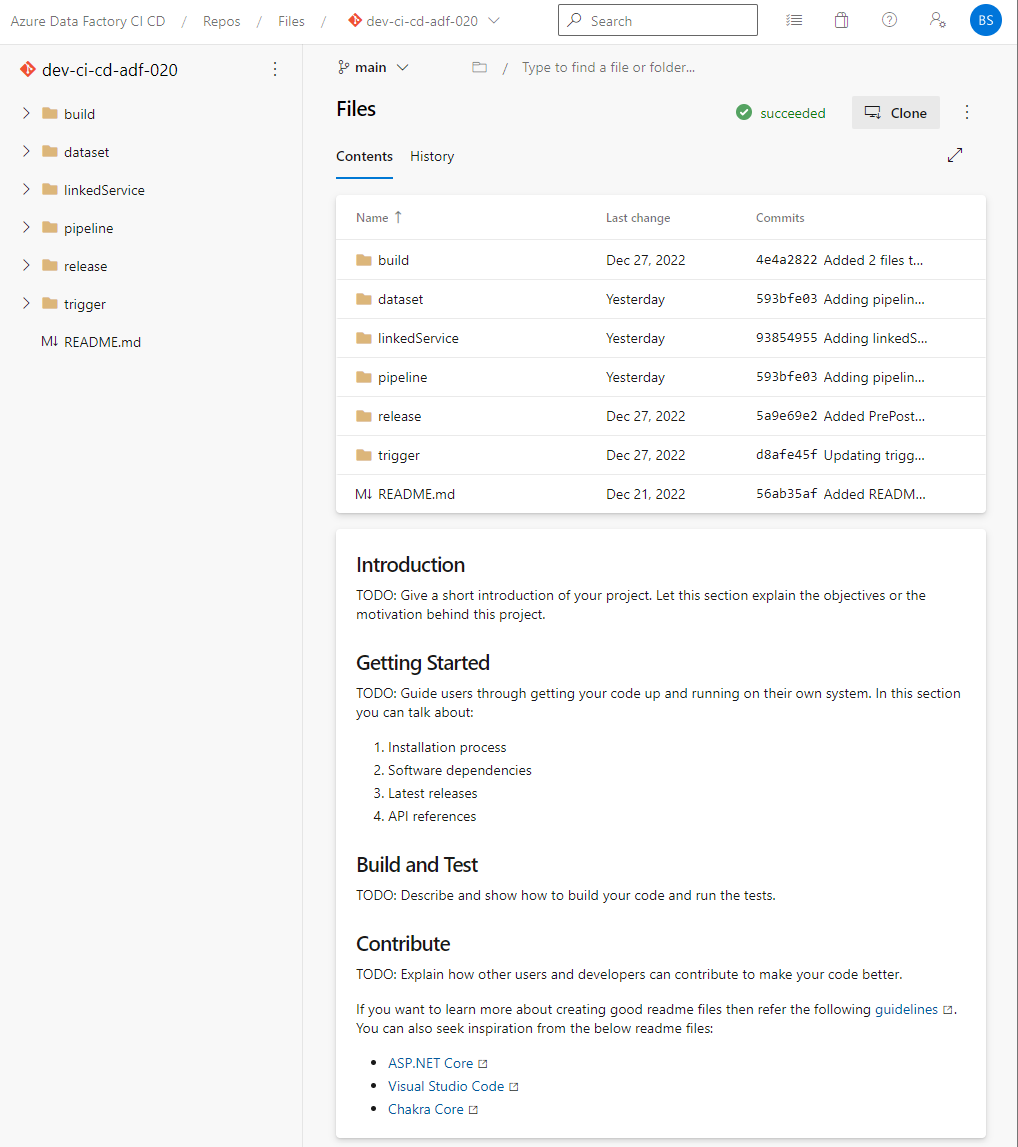
Step 0 – Git configuration within DevOps option 1



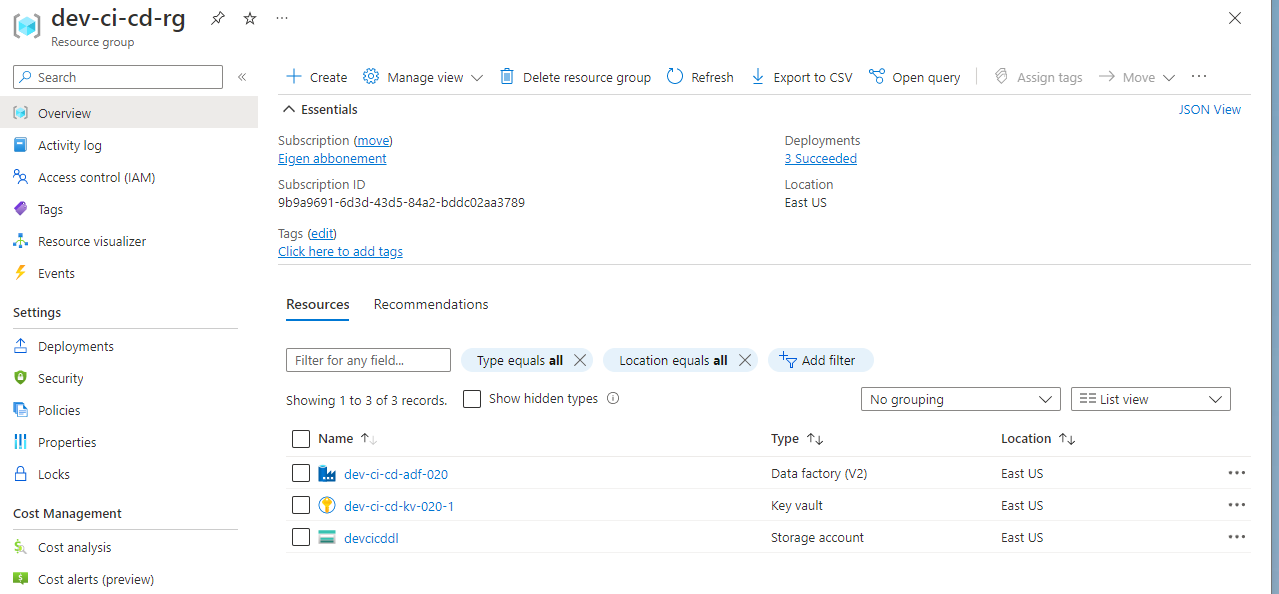
Step 0 – Git configuration within DevOps option 2



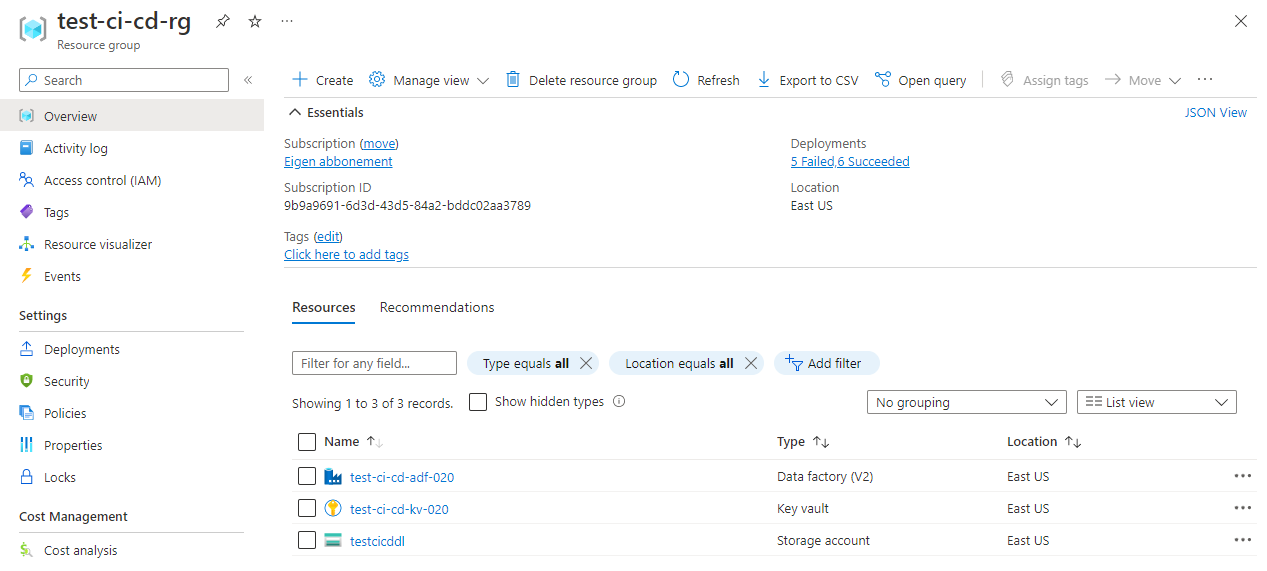
Step 1 – Create git repo



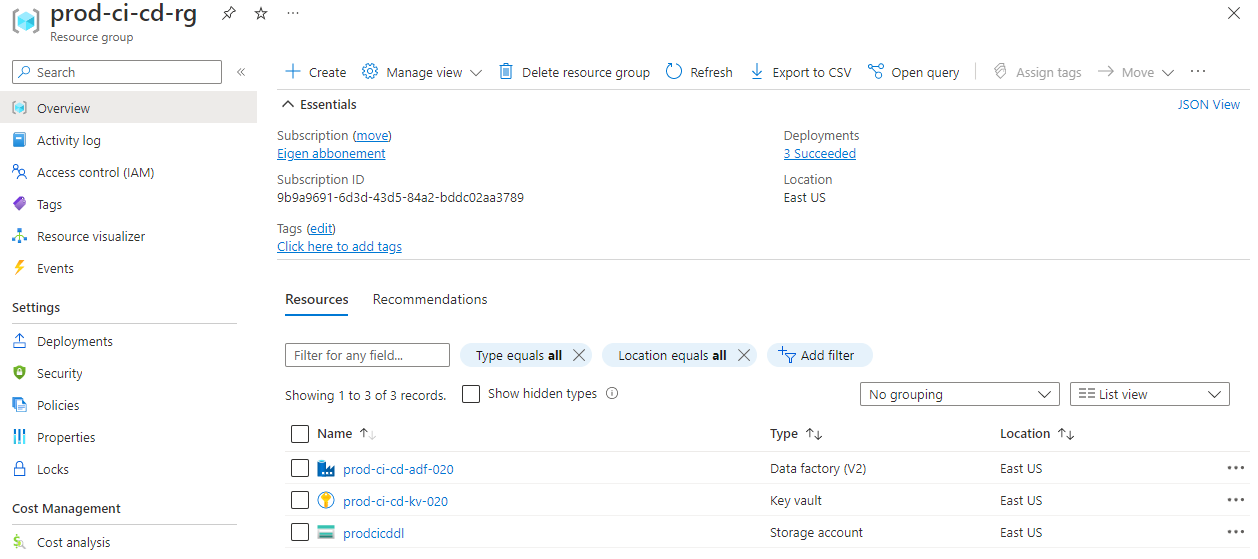
Step 2 – Create Tools for azure (DEV)



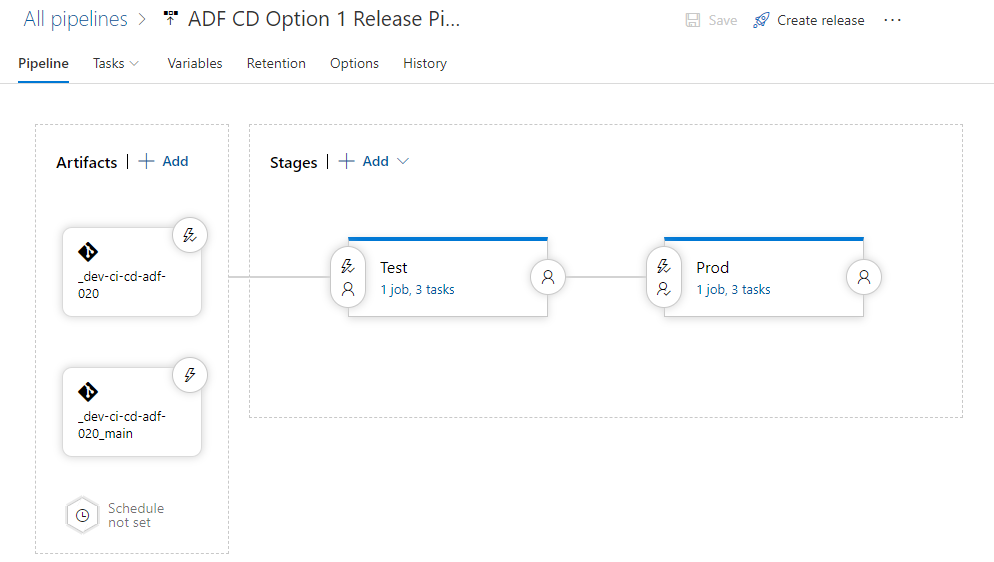
Step 2 – Create Tools for azure (TEST)



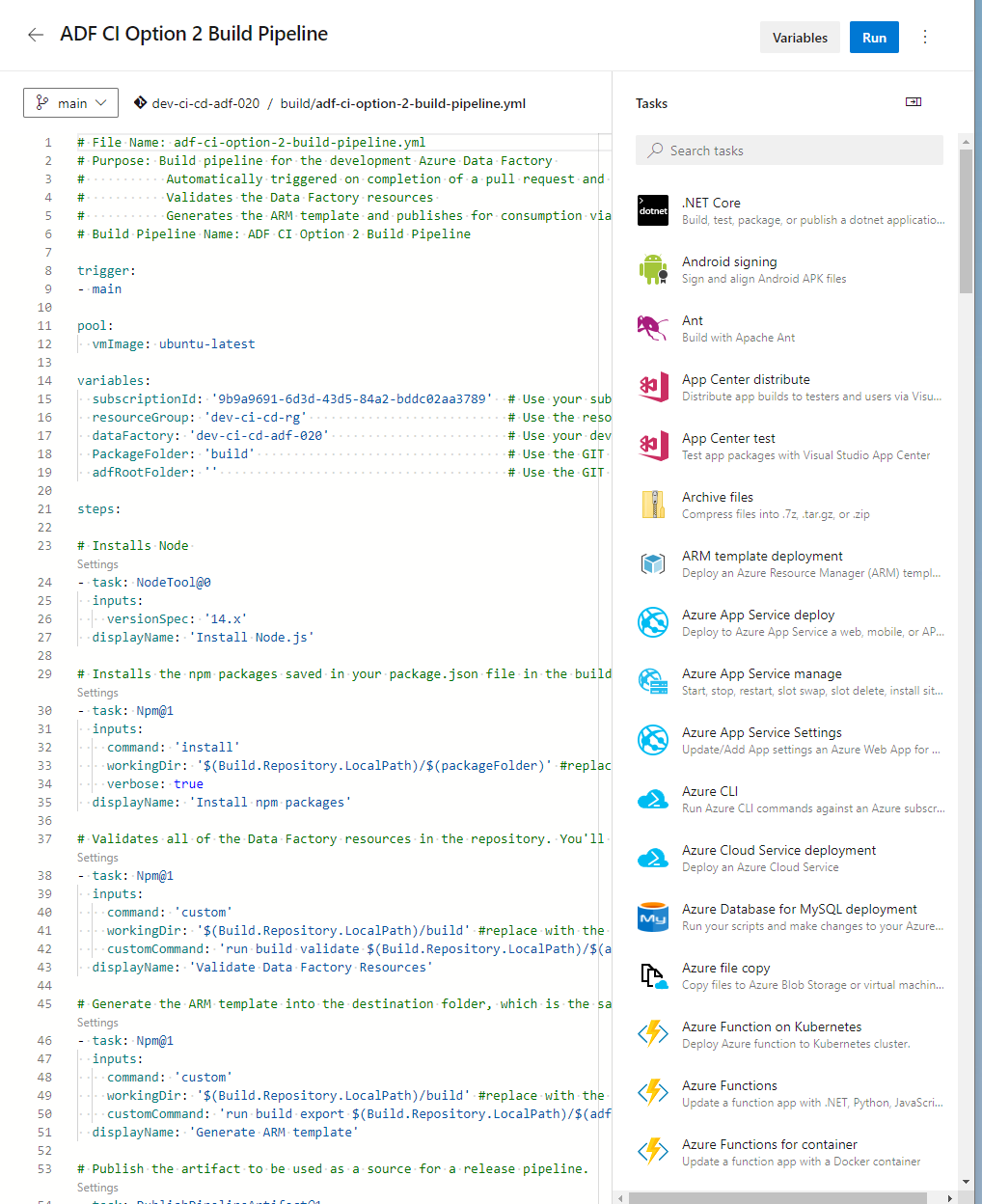
Step 2 – Create Tools for azure (PROD)



Step 3 - Create release pipeline option 1



Step 4 – Create build pipeline option 2 – 1 / 2



Step 5 – Create release pipeline option 2

