**Azure Data LakeHouse in an Hour Links and Code Snippets**

*Created by Greg Beaumont:* [*https://linkedin*](https://linkedin)*.com/in/gregbeaumont*

*This Document is a collection of links and code snippets from the Workshop that users can cut and paste for use in Azure Data Factory. Here is a link to the Virtual On-Demand Workshop -* [*https://youtu.be/G1IzJG0N8Z0*](https://youtu.be/G1IzJG0N8Z0)

* **Intro**
  + Link to get an Azure Account – <https://azure.microsoft.com/en-us/free/>
  + Information about CDC National Notifiable Diseases Surveillance System (NNDSS) Data – <https://data.cdc.gov/NNDSS/NNDSS-Weekly-Data/x9gk-5huc>
* **Step 1 – Create Resource Group, Azure Data Lake, and Azure Data Factory (no info needed)**
* **Step 2 – Rapid Prototype of the Data Model**
  + Link to get the CDC data for profiling – <https://data.cdc.gov/api/views/x9gk-5huc/rows.csv?accessType=DOWNLOAD>
* **Step 3 – Define the Datasets in Azure Data Factory**
  + Create Source Dataset for CDC Data
    - Link to web CSV file – Link to get the CDC data for profiling – <https://data.cdc.gov/api/views/x9gk-5huc/rows.csv?accessType=DOWNLOAD>
  + Create Source Dataset for Date table
    - Link to GitHub site Date table – <https://raw.githubusercontent.com/gregbeaumont/PowerPopHealth/main/Date%20%26%20Time%20Scripts/Date.csv>
* **Step 4 – Populate the Bronze Layer of the Data LakeHouse (no info meeded)**
* **Step 5 – Populate the Silver Layer of the Data LakeHouse**
  + Data Flow *BronzeDatetoSilver* – The Step “AssignDataTypes” uses either **toInteger(ColumnName)** or **toDate(Column Name)** to specify the data types for the parquet file.
  + Data Flow *BronzeCDCtoSilver* – The Step “NewColumns” has code snippets to break out Latitude and Longitude, add CreatedDate, and add YearWeekKey:
    - Longitude - replace(split(geocode,' ')[2],"(","")
    - Latitude - replace(split(geocode,' ')[3],")","")
    - CreatedDate – currentDate()
    - YearWeekKey - concat(MMWRYEAR, case(toInteger(MMWRWEEK) >= 10,MMWRWEEK,concat(toString(0),MMWRWEEK)))
  + Data Flow *BronzeCDCtoSilver* – The Step “DataTypes” uses **toInteger(ColumnName)** to specify data types for the parquet file.
* **Step 6 – Populate the Gold Layer of the Data LakeHouse**
  + The Dataset BronzeCDCWeeklySnapshots uses Dynamic Content for the Filename so that it is different every time the weekly update runs - @concat('CDC\_NNDSS\_',formatDateTime(convertTimeZone(utcnow(),'UTC','Eastern Standard Time'),'yyyy-MM-ddTHHmmss'), '.csv')