**GLC SQL Database Documentation**

Last updated: February 24, 2022

By: Kalvin Wu

Navigation

[Business Case overview 1](#_Toc96607829)

[Point of Contacts 1](#_Toc96607830)

[Table/View and Script Definitions 1](#_Toc96607831)

[High Level Overview 3](#_Toc96607832)

[WBEA: L1 vs L2 Data 4](#_Toc96607833)

[Historical WBEA Data 4](#_Toc96607834)

[Silver Table 5](#_Toc96607835)

[BARR Data 5](#_Toc96607836)

[Final Views 5](#_Toc96607837)

# Business Case overview

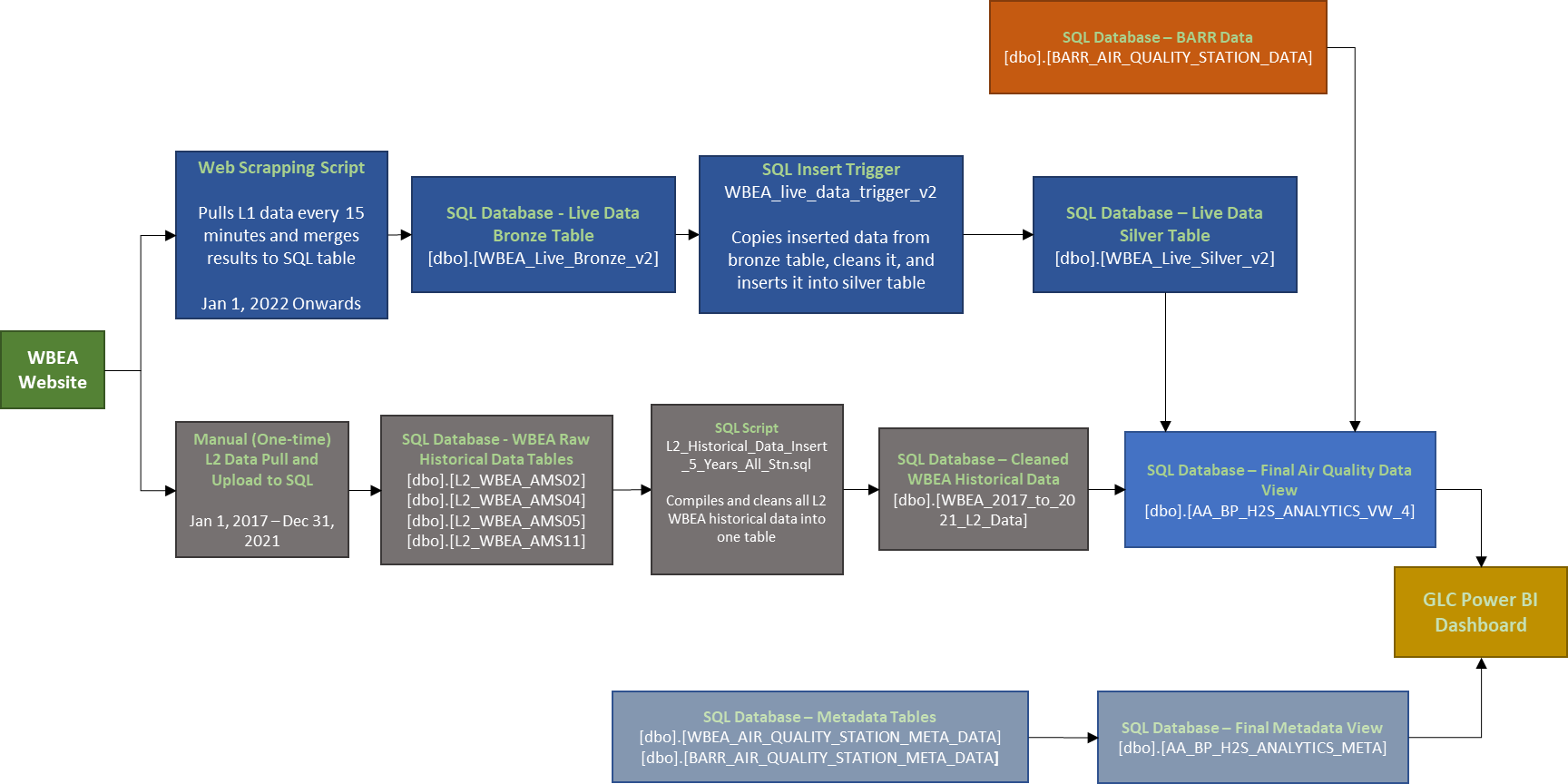
* [Please review this presentation for background and scope](https://suncor.sharepoint.com/:p:/s/GLCAnalytics/EcorLohgbyJNsUGL5LiJD_4BaEfVdAcaxMpL8f5vNDITHw?e=8vB7hM)

# Point of Contacts

* **Business**: Brian Beck, Sheri Lepoudre
* **Technical**: 
  + Data Engineering: Kalvin Wu, Jeslin Thomas
  + Visualization: Ksenia Khatova, Kalvin Wu, Rubin Rana
  + Sustainment: Konstantin Rubinstein
  + Data Scientist: Thomas Speidel
  + Advanced Analytics Manager: Brian Emmerson
* **Sustainment (EH&S)**
  + Ksenia Khatova
  + Konstantin Rubinstein
  + Chuks Nwakanma
* **WBEA**
  + info@wbea.org
  + Dianne McIsaac <dmcisaac@wbea.org>; Asad Hidayat <ahidayat@wbea.org>
* Table/View and Script Definitions

|  |  |  |
| --- | --- | --- |
| **Name(s)** | **Type** | **Description** |
| [dbo].[L2\_WBEA\_AMS02]  [dbo].[L2\_WBEA\_AMS04]  [dbo].[L2\_WBEA\_AMS05]  [dbo].[L2\_WBEA\_AMS11] | Table | Raw 5 years of historical WBEA (L2) data. Each table corresponds to a station:  AMS02 – Mildred Lake  AMS04 – Buffalo Viewpoint  AMS05 – Mannix  AMS11 – Lower Camp  Start date: Jan 1, 2017  End date: Dec 31, 2021 |
| [dbo].[WBEA\_Live\_Bronze\_v2] | Table | Raw table where live WBEA (L1) data is inserted. |
| [dbo].[WBEA\_Live\_Silver\_v2] | Table | Silver table for live WBEA (L1) data which contains only necessary columns and proper data types. |
| [dbo].[BARR\_AIR\_QUALITY\_STATION\_DATA] | Table | All data for 7 BARR stations (Suncor 1 – Suncor 7). |
| [dbo].[BARR\_AIR\_QUALITY\_STATION\_META\_DATA] | Table | BARR metadata. |
| [dbo].[WBEA\_AIR\_QUALITY\_STATION\_META\_DATA] | Table | WBEA metadata. |
| [dbo].[WBEA\_2017\_to\_2021\_L2\_Data] | Table | Processed 5 years of historical WBEA data (2017-2021) for all WBEA stations. Data is combined from:  [dbo].[L2\_WBEA\_AMS02]  [dbo].[L2\_WBEA\_AMS04]  [dbo].[L2\_WBEA\_AMS05]  [dbo].[L2\_WBEA\_AMS11] |
| [dbo].[z\_archive\_AMS02\_WBEA]  [dbo].[z\_archive\_AMS04\_WBEA]  [dbo].[z\_archive\_AMS05\_WBEA]  [dbo].[z\_archive\_AMS011\_WBEA] | Table | 20+ years of historical WBEA data.  The data was extracted from: <http://67.210.212.45/wbea.html>. However, after discussions with the WBEA, the data from this website is not completely accurate and should avoid being used. As this website is currently the only way to easily obtain 20 years of history, it will be kept in the database, but is **not currently used** for the GLC dashboard. |
| [dbo].[AA\_BP\_H2S\_ANALYTICS\_VW\_4] | View | Contains all air quality data. Connected to Power BI dashboard. |
| [dbo].[AA\_BP\_H2S\_ANALYTICS\_META] | View | Includes all metadata of WBEA and BARR stations. Meant to be used in conjunction with [dbo].[AA\_BP\_H2S\_ANALYTICS\_VW\_4] once it has been loaded into Power BI. |
| WBEA\_Insert\_Data\_Into\_Silver\_Table\_V3.sql | SQL Script | SQL Script used to fill the initial data for [dbo].[WBEA\_Live\_Silver\_v2]. It copies the desired data from [dbo].[WBEA\_Live\_Bronze\_v2], and casts to proper data type. |
| L1\_WBEA\_Bronze\_Data\_Replacement.sql | SQL Script | SQL script used to replace the data in [dbo].[WBEA\_Live\_Bronze\_v2] in the case that there is corrupted data. |
| L2\_Historical\_Data\_Insert\_5\_Years\_All\_Stn.sql | SQL Script | Consolidates 5 year WBEA L2 data from:  [dbo].[L2\_WBEA\_AMS02]  [dbo].[L2\_WBEA\_AMS04]  [dbo].[L2\_WBEA\_AMS05]  [dbo].[L2\_WBEA\_AMS11]  Into one table: [dbo].[WBEA\_2017\_to\_2021\_L2\_Data]  Data is casted to proper types and only necessary columns are selected. |

# High Level Overview



**GLC SQL Database Information**

**Server:** glcserver001.database.windows.net

**Database**: GLC

Every 15 minutes, the Python script written by Jeslin will scrape the [WBEA website](https://wbea.org/historical-monitoring-data/) for new L1 data and directly merge the data into our SQL database. The script is currently running in an Azure Machine Learning workspace in a sandbox environment. The scheduling is done using a cronjob with a compute instance. A copy of the script can be found here: <https://github.com/jjthomas22/GLC>.

The data will land in the [dbo].[WBEA\_Live\_Bronze\_v2] table. However, the numeric values in this table are still in the nvarchar data type. To take care of this, a trigger (*WBEA\_live\_data\_trigger\_v2*) was created. This trigger can be found under the [dbo].[WBEA\_Live\_Bronze\_v2] table. The trigger will cast any inserted data to its proper data type and replace any values not matching the expected data type to NULL. For example, for H2S readings, a float is expected. Some records will have "n/d" (no data) – these values will be casted to NULL which will exclude the record in future calculations. The trigger then inserts this data into the [dbo].[WBEA\_Live\_Silver\_v2] table while the raw data remains in the [dbo].[WBEA\_Live\_Bronze\_v2] table. For more details on this, refer to the following section: Silver Table.

The "live" data from [dbo].[WBEA\_Live\_Silver\_v2] is then combined with the historical L2 WBEA and BARR data in the following view: [dbo].[AA\_BP\_H2S\_ANALYTICS\_VW\_4]. This view unions all the air quality data which include:

* "Live" L1 WBEA data. (Data scraped using python script)
* Historical L2 WBEA data
* BARR data

Another view, [dbo].[AA\_BP\_H2S\_ANALYTICS\_META] unions the metadata for WBEA and BARR stations. The data from these 2 views are connected to the [Power BI dashboard](https://app.powerbi.com/groups/me/apps/654610c3-fd6f-4805-bb6c-0bc2a3016623?ctid=1aa51068-11a6-4bd2-8646-1fff31a30ffc). For more details about the views, please refer to the Final Views section.

## WBEA: L1 vs L2 Data

The WBEA website has 2 types of data, L1 and L2. Please see below for the WBEA's explanation of the difference:

**Level 1 (L1) Data**

Level 1 (L1) data are raw air quality monitoring data displayed directly from monitoring equipment output signals. L1 data is generally indicative of ambient conditions, however, it has not been reviewed, validated, nor processed for baseline drift and other fluctuations by qualified technicians. Final processed and validated data values may vary from L1 values displayed on this site. A displayed value of ‘n/d’ indicates air monitoring data is not available at the corresponding time due to scheduled maintenance and/or instrument calibration. Charted data is locked to the range of the analyzer by default, and can be unlocked by the appropriate control. Level 1 data are available from the WBEA website in near real time.

**Level 2 (L2) Data**

Level 2 (L2) data are quality controlled data. These data are processed monthly; including a data summary, validation and baseline correction of air quality measurements for analyzer drift. Data are further reviewed by WBEA as an acceptance test before generation of monthly compliance reports and distribution to various locations, including the WBEA PI database for continuous data, the WBEA website (wbea.org), the CASA data warehouse and WBEA members. This is WBEA’s final data – these data are usually available 30 days after the end of the calendar month in which they were collected.

## Historical WBEA Data

The raw historical WBEA data is stored in 4 separate tables, one table for each station:

* [dbo].[L2\_AMS02\_WBEA]
* [dbo].[L2\_AMS04\_WBEA]
* [dbo].[L2\_AMS05\_WBEA]
* [dbo].[L2\_AMS011\_WBEA]

Each table contains L2 data and 5 years of history (Jan 1, 2017 – Dec 31, 2021). The data was manually pulled from the [WBEA website](https://wbea.org/historical-monitoring-data/) as CSV files and uploaded in the SQL database.

A script was used *(L2\_Historical\_Data\_Insert\_5\_Years\_All\_Stn.sql)* to consolidate the historical data into a single table: [dbo].[WBEA\_2017\_to\_2021\_L2\_Data]. The script also joins the historical data with WBEA

station metadata.

## Script Location and Schedule

Cron is a scheduling daemon that executes tasks at specified intervals. These tasks are called cron jobs and are mostly used to automate system maintenance or administration. A bash script has been created to execute the python script every 15 minutes and python file is located in Azure ML workspace sandbox environment under folder path - Suncor EnergyInc/iaasbxarmmlwuw2001/Notebooks/Users/jjthomas/glcfinalversion.py. Log files for each script execution results are stored in a log file called intermediate.log located in the same location as script file.

## Notifications and Monitoring

Azure Application Insights allows you to monitor your application and send you alerts when it is either unavailable, experiencing failures, or suffering from performance issues. An Azure Application insight is created to monitor the website and send notifications to email address-jjthomas@suncor.com whenever the website is down. Insights monitors for application performance and availability of the website every 15 minutes.

## Version Control

WBEA script is stored at location -https://github.com/SEAdvancedAnalyticsOrg/GroundLevelConcentrations-GLC

## Silver Table

The [dbo].[WBEA\_Live\_Silver\_v2] table was created to serve as the silver table for the data from [dbo].[WBEA\_Live\_Bronze\_v2]. In the silver table, only the required columns are selected, and their data types have been corrected.

There are two steps to creating the silver table:

1. Running the SQL script (*WBEA\_Insert\_Data\_Into\_Silver\_Table\_V3.sql*) to move all existing data from the bronze/raw table to the silver table. This script is only run once to initially transfer data from the bronze table to silver table.
2. An insert trigger (*WBEA\_live\_data\_trigger*) that casts any new data entering [dbo].[WBEA\_Live\_Bronze\_v2] to the proper data type and inserts it into [dbo].[WBEA\_Live\_Silver].

\* In both steps, the main transformation is casting data to the proper data type and replacing any values not matching the expected data type to NULL. For example, for H2S readings, a float is expected. Some records will have "n/d" (no data) – these values will be casted to NULL which will exclude the record during future calculations.

## BARR Data

The BARR data is already in the proper data types and exists in the following table: [dbo].[BARR\_AIR\_QUALITY\_STATION\_DATA]. The data starts in May 2017.

The BARR metadata is contained in [dbo].[BARR\_AIR\_QUALITY\_STATION\_META\_DATA]

## Final Views

Currently there are two views that are connected to the Power BI dashboard. We separate the air quality data from the metadata to avoid mass duplication of metadata values in the query. Both views will be ingested into Power BI where a relationship between the tables will be created.

**View 1:** [dbo].[AA\_BP\_H2S\_ANALYTICS\_VW\_4]

This view compiles all the air quality data from WBEA and BARR stations. In this view, we take the union of:

1. Historical L2 WBEA data from [dbo].[WBEA\_2017\_to\_2021\_L2\_Data]
2. The live L1 WBEA data stored in [dbo].[WBEA\_Live\_Silver\_v2]
3. BARR station data from [dbo].[BARR\_AIR\_QUALITY\_STATION\_DATA]

Where required, the data is joined with WBEA/BARR metadata to include the station name in this view. This is because the "station\_name" column is used as the foreign key.

**View 2:** [dbo].[AA\_BP\_H2S\_ANALYTICS\_META]

This view unions the WBEA and BARR metadata from [dbo].[WBEA\_AIR\_QUALITY\_STATION\_META\_DATA] and [dbo].[BARR\_AIR\_QUALITY\_STATION\_META\_DATA]. The "station\_name" column is used as the primary key.