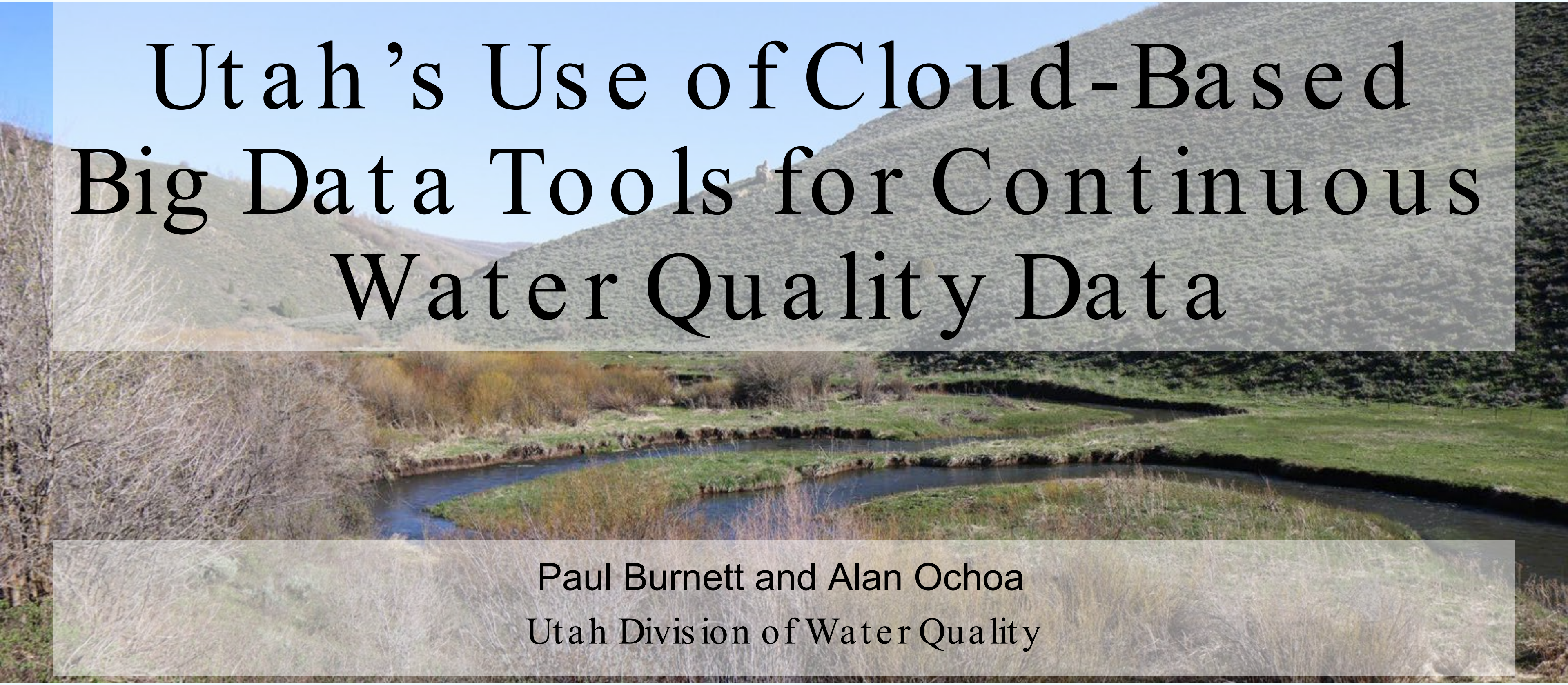


Utah's Use of Cloud-Based Big Data Tools for Continuous Water Quality Data

Paul Burnett and Alan Ochoa

Utah Division of Water Quality

UTAH DEPARTMENT of ENVIRONMENTAL QUALITY
WATER QUALITY

Utah Division of Water Quality Continuous Water Quality DB Team

Alex Heppner

Ben Holcomb

Alan Ochoa

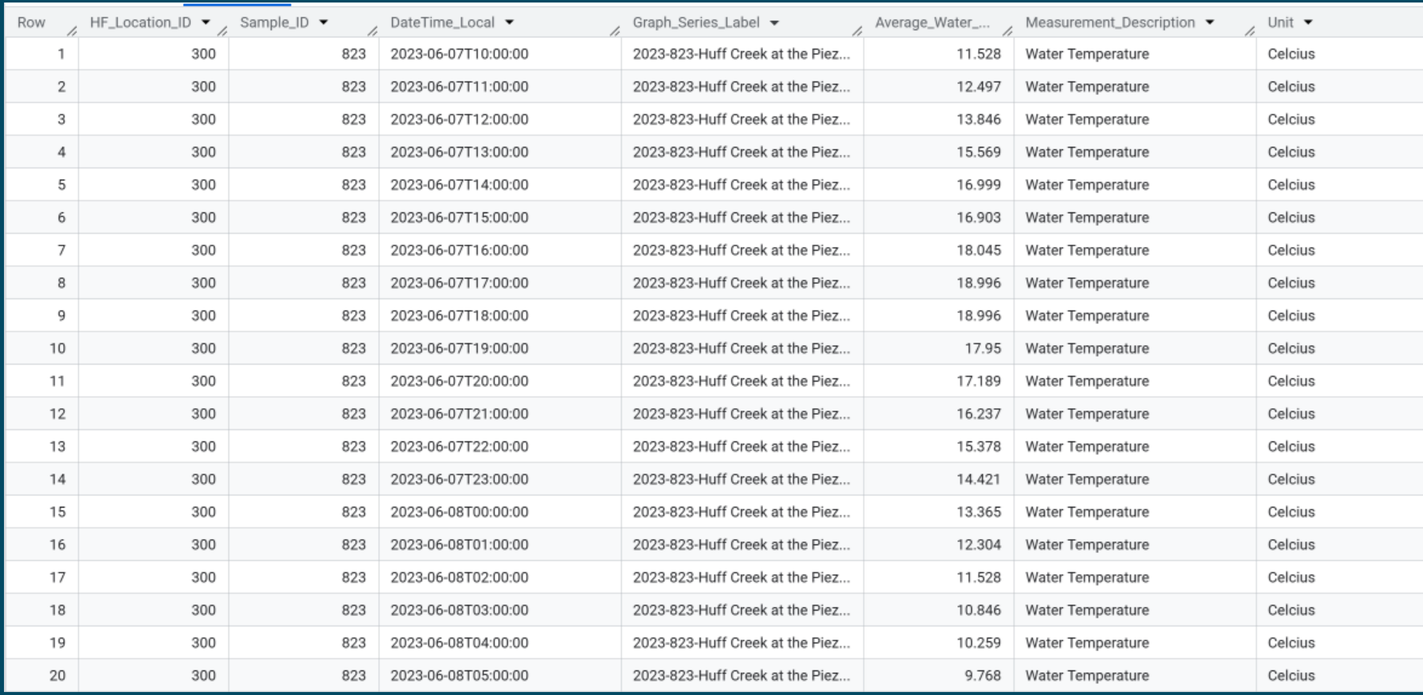
Ryan Parker

Marshall Baillie

Julian Carroll

Toby Hooker

Paul Burnett



Row	HF_Location_ID	Sample_ID	DateTime_Local	Graph_Series_Label	Average_Water_...	Measurement_Description	Unit
1	300	823	2023-06-07T10:00:00	2023-823-Huff Creek at the Piez...	11.528	Water Temperature	Celcius
2	300	823	2023-06-07T11:00:00	2023-823-Huff Creek at the Piez...	12.497	Water Temperature	Celcius
3	300	823	2023-06-07T12:00:00	2023-823-Huff Creek at the Piez...	13.846	Water Temperature	Celcius
4	300	823	2023-06-07T13:00:00	2023-823-Huff Creek at the Piez...	15.569	Water Temperature	Celcius
5	300	823	2023-06-07T14:00:00	2023-823-Huff Creek at the Piez...	16.999	Water Temperature	Celcius
6	300	823	2023-06-07T15:00:00	2023-823-Huff Creek at the Piez...	16.903	Water Temperature	Celcius
7	300	823	2023-06-07T16:00:00	2023-823-Huff Creek at the Piez...	18.045	Water Temperature	Celcius
8	300	823	2023-06-07T17:00:00	2023-823-Huff Creek at the Piez...	18.996	Water Temperature	Celcius
9	300	823	2023-06-07T18:00:00	2023-823-Huff Creek at the Piez...	18.996	Water Temperature	Celcius
10	300	823	2023-06-07T19:00:00	2023-823-Huff Creek at the Piez...	17.95	Water Temperature	Celcius
11	300	823	2023-06-07T20:00:00	2023-823-Huff Creek at the Piez...	17.189	Water Temperature	Celcius
12	300	823	2023-06-07T21:00:00	2023-823-Huff Creek at the Piez...	16.237	Water Temperature	Celcius
13	300	823	2023-06-07T22:00:00	2023-823-Huff Creek at the Piez...	15.378	Water Temperature	Celcius
14	300	823	2023-06-07T23:00:00	2023-823-Huff Creek at the Piez...	14.421	Water Temperature	Celcius
15	300	823	2023-06-08T00:00:00	2023-823-Huff Creek at the Piez...	13.365	Water Temperature	Celcius
16	300	823	2023-06-08T01:00:00	2023-823-Huff Creek at the Piez...	12.304	Water Temperature	Celcius
17	300	823	2023-06-08T02:00:00	2023-823-Huff Creek at the Piez...	11.528	Water Temperature	Celcius
18	300	823	2023-06-08T03:00:00	2023-823-Huff Creek at the Piez...	10.846	Water Temperature	Celcius
19	300	823	2023-06-08T04:00:00	2023-823-Huff Creek at the Piez...	10.259	Water Temperature	Celcius
20	300	823	2023-06-08T05:00:00	2023-823-Huff Creek at the Piez...	9.768	Water Temperature	Celcius

Project Overview

- Goal: Centralize and manage continuous water quality data from multiple monitoring sites and sources throughout Utah.
- Scope:
  - Multiple sensor types (e.g. lake, stream, piezometers)
  - Multiple parameters (e.g. water temp, ph, etc)
  - Real-time or near-real-time data streams
  - Bulk data ingestion from deployed loggers
  - Allow multi-platform access for analysis
- Users: Program staff, researchers, and partner agencies, general public.

Constraints

Out-of-the-box solutions are too expensive

No dedicated staff for database development or management

Existing system pulling data from streaming buoys was manual (monthly)

No existing framework for logger data (~1200 datasets and counting)

System Architecture

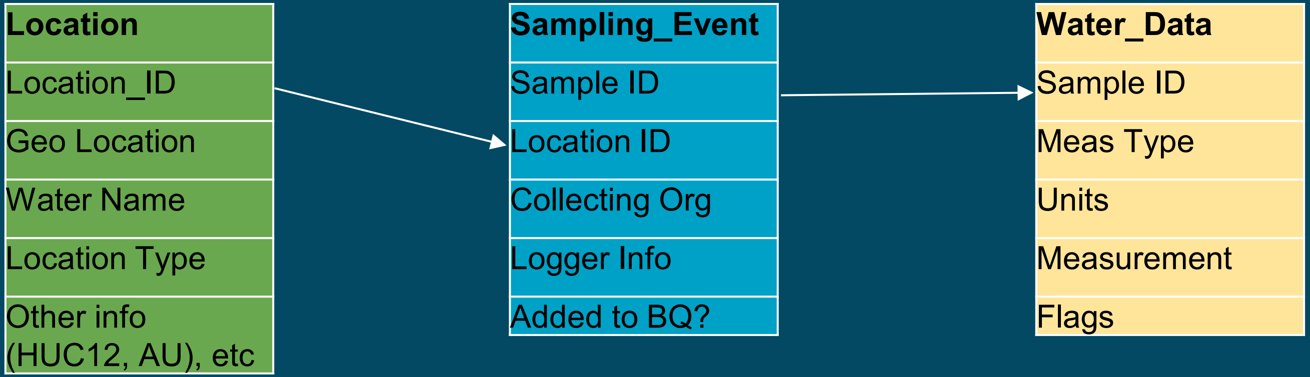
Data streams

- Deployed standalone loggers
- Streaming data from buoys
- Streaming USGS data

Leveraging Google Sheets

- Location Table
- Samping Event Table
- Bulk Data Entry (template)
- Google forms for basic data entry
- Apps Scripts for automation

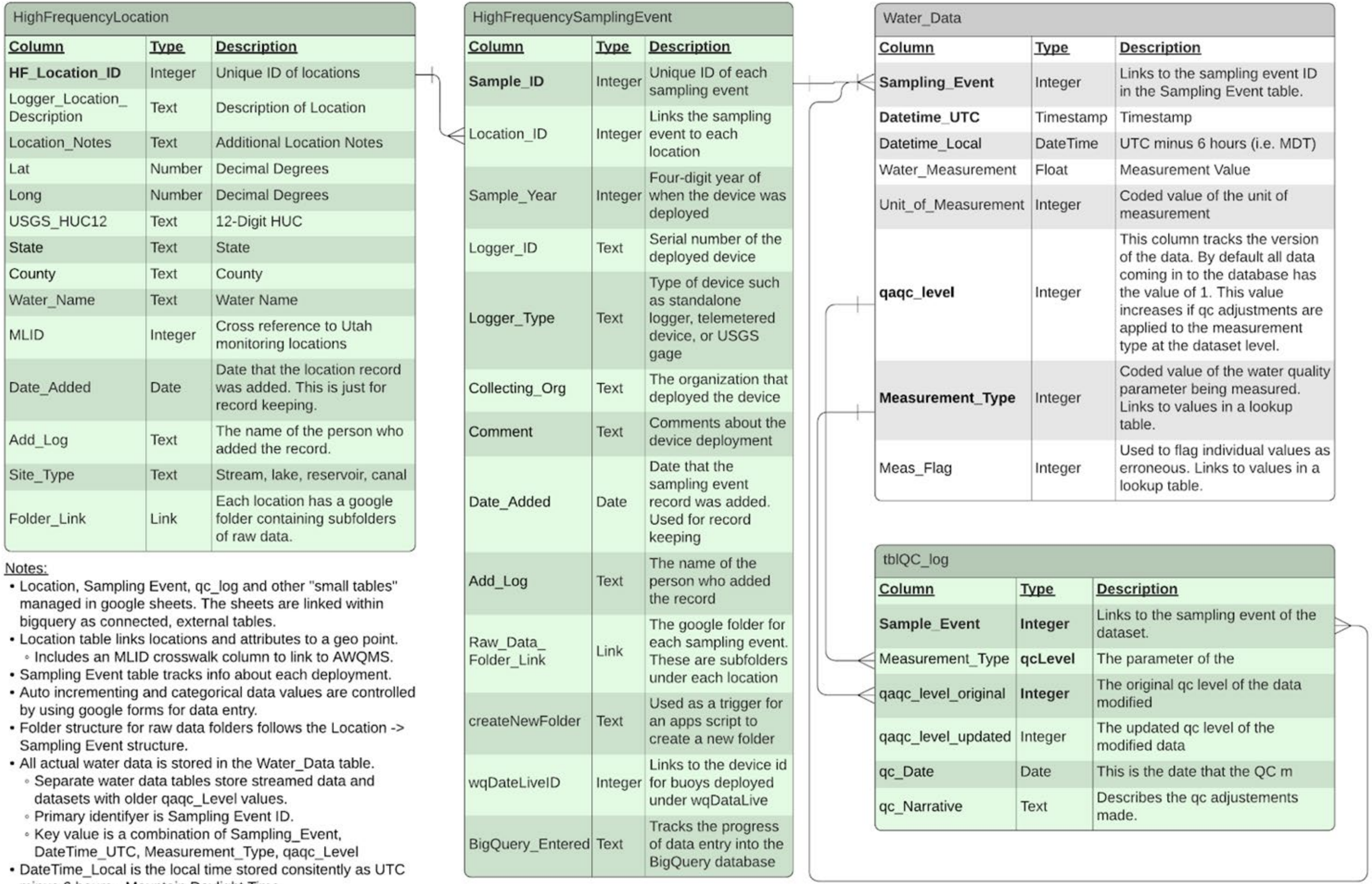
- Bigquery for data warehousing
- Cloud Functions for automation
- Scheduled queries for:
  - Automated data wrangling
  - Denormalization for visualization
- Looker Studio for visualization
- Google Shared Drive - Working Folders for Raw Data



Location
Location_ID
Geo Location
Water Name
Location Type
Other info (HUC12, AU), etc

Sampling_Event
Sample ID
Location ID
Collecting Org
Logger Info
Added to BQ?

Water_Data
Sample ID
Meas Type
Units
Measurement
Flags



HighFrequencyLocation

Column	Type	Description
HF_Location_ID	Integer	Unique ID of locations
Logger_Location_ Description	Text	Description of Location
Location_Notes	Text	Additional Location Notes
Lat	Number	Decimal Degrees
Long	Number	Decimal Degrees
USGS_HUC12	Text	12-Digit HUC
State	Text	State
County	Text	County
Water_Name	Text	Water Name
MLID	Integer	Cross reference to Utah monitoring locations
Date_Added	Date	Date that the location record was added. This is just for record keeping.
Add_Log	Text	The name of the person who added the record.
Site_Type	Text	Stream, lake, reservoir, canal
Folder_Link	Link	Each location has a google folder containing subfolders of raw data.

HighFrequencySamplingEvent

Column	Type	Description
Sample_ID	Integer	Unique ID of each sampling event
Location_ID	Integer	Links the sampling event to each location
Sample_Year	Integer	Four-digit year of when the device was deployed
Logger_ID	Text	Serial number of the deployed device
Logger_Type	Text	Type of device such as standalone logger, telemetered device, or USGS gage
Collecting_Org	Text	The organization that deployed the device
Comment	Text	Comments about the device deployment
Date_Added	Date	Date that the sampling event record was added. Used for record keeping
Add_Log	Text	The name of the person who added the record
Raw_Data_ Folder_Link	Link	The google folder for each sampling event. These are subfolders under each location
createNewFolder	Text	Used as a trigger for an apps script to create a new folder
wqDateLiveID	Integer	Links to the device id for buoys deployed under wqDataLive
BigQuery_Entered	Text	Tracks the progress of data entry into the BigQuery database

Water_Data

Column	Type	Description
Sampling_Event	Integer	Links to the sampling event ID in the Sampling Event table.
Datetime_UTC	Timestamp	Timestamp
Datetime_Local	DateTime	UTC minus 6 hours (i.e. MDT)
Water_Measurement	Float	Measurement Value
Unit_of_Measurement	Integer	Coded value of the unit of measurement
qaqc_level	Integer	This column tracks the version of the data. By default all data coming in to the database has the value of 1. This value increases if qc adjustments are applied to the measurement type at the dataset level.
Measurement_Type	Integer	Coded value of the water quality parameter being measured. Links to values in a lookup table.
Meas_Flag	Integer	Used to flag individual values as erroneous. Links to values in a lookup table.

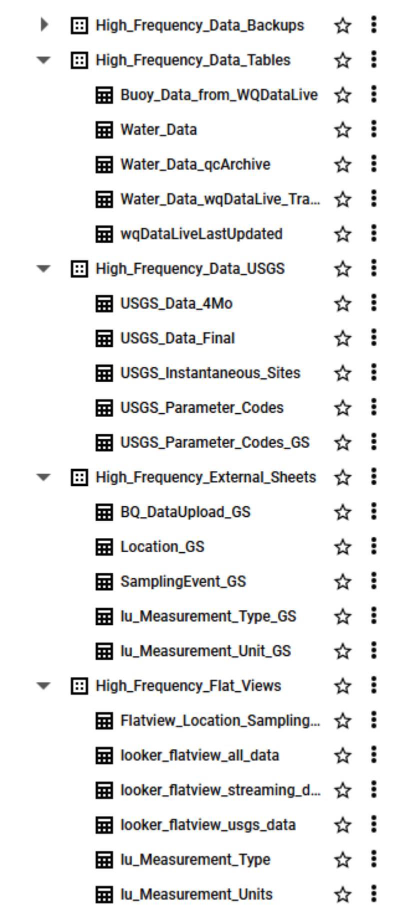
tblQC_log

Column	Type	Description
Sample_Event	**Integer**	Links to the sampling event of the dataset.
Measurement_Type	**qcLevel**	The parameter of the
qaqc_level_original	**Integer**	The original qc level of the data modified
qaqc_level_updated	Integer	The updated qc level of the modified data
qc_Date	Date	This is the date that the QC m
qc_Narrative	Text	Describes the qc adjustements made.

Notes:

- Location, Sampling Event, qc_log and other "small tables" managed in google sheets. The sheets are linked within bigquery as connected, external tables.
- Location table links locations and attributes to a geo point.
  - Includes an MLID crosswalk column to link to AWQMS.
- Sampling Event table tracks info about each deployment.
- Auto incrementing and categorical data values are controlled by using google forms for data entry.
- Folder structure for raw data folders follows the Location -> Sampling Event structure.
- All actual water data is stored in the Water_Data table.
  - Separate water data tables store streamed data and datasets with older qaqc_Level values.
  - Primary identifyer is Sampling Event ID.
  - Key value is a combination of Sampling_Event, DateTime_UTC, Measurement_Type, qaqc_Level
- DateTime_Local is the local time stored consitently as UTC minus 6 hours - Mountain Daylight Time.
- QC is applied to a Meas Type for entire sampling event.
  - Only increase qc level if measurement values are changed.
  - Adding qc flag would not constitute a new qc level.

Bigquery for Big Tables



- High_Frequency_Data_Backups
- High_Frequency_Data_Tables
  - Buoy_Data_from_WQDataLive
  - Water_Data
  - Water_Data_qcArchive
  - Water_Data_wqDataLive_Tra...
  - wqDataLiveLastUpdated
- High_Frequency_Data_USGS
  - USGS_Data_4Mo
  - USGS_Data_Final
  - USGS_Instantaneous_Sites
  - USGS_Parameter_Codes
  - USGS_Parameter_Codes_GS
- High_Frequency_External_Sheets
  - BQ_DataUpload_GS
  - Location_GS
  - SamplingEvent_GS
  - lu_Measurement_Type_GS
  - lu_Measurement_Unit_GS
- High_Frequency_Flat_Views
  - Flatview_Location_Sampling...
  - looker_flatview_all_data
  - looker_flatview_streaming_d...
  - looker_flatview_usgs_data
  - lu_Measurement_Type
  - lu_Measurement_Units

Water Data Tables

- Data from dataloggers and streamed data
- QC archived data

USGS Data tables

- Final data table (data older than 4 months)
- 4 month data (provisional data)
- Sites and parameter codes

External sheets

- linked to google sheets

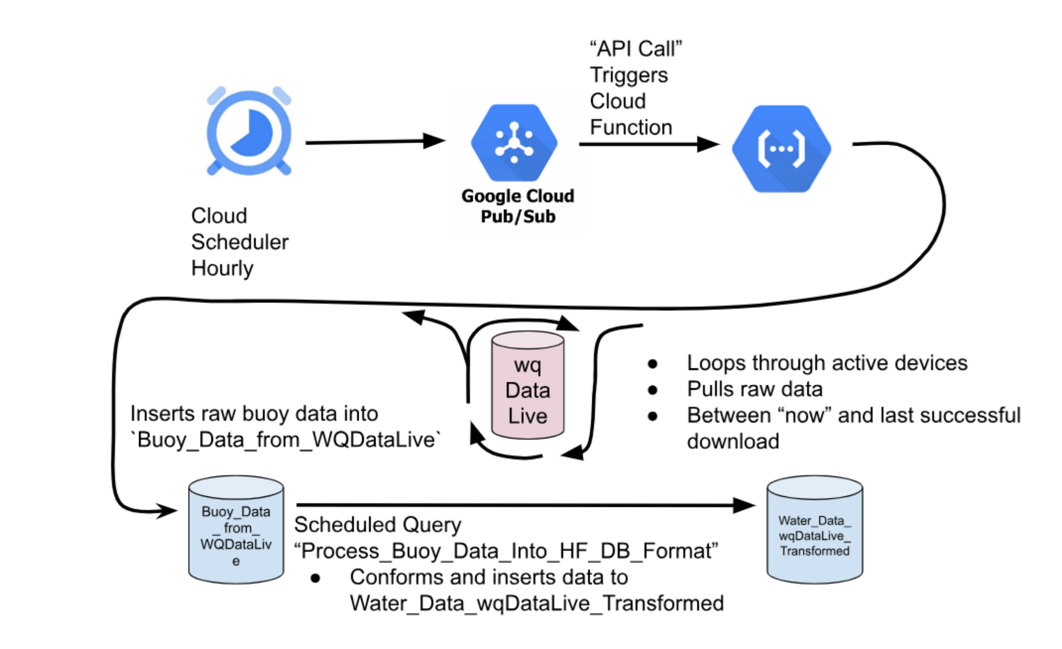
Flat Views

- Data flattened for dashboards
- Pulled from data tables hourly/daily with scheduled queries

Cloud Functions for Automation

Lightweight, event-driven functions that run backend code in response to events without needing to manage servers.

- Using for streaming buoy data from wqDataLive
- Streaming USGS data
- Testing of streaming microcontroller data (soon)



Cloud Scheduler Hourly

Google Cloud Pub/Sub

“API Call” Triggers Cloud Function

- Loops through active devices
- Pulls raw data
- Between “now” and last successful download

wq Data Live

Inserts raw buoy data into \`Buoy_Data_from_WQDataLive\`

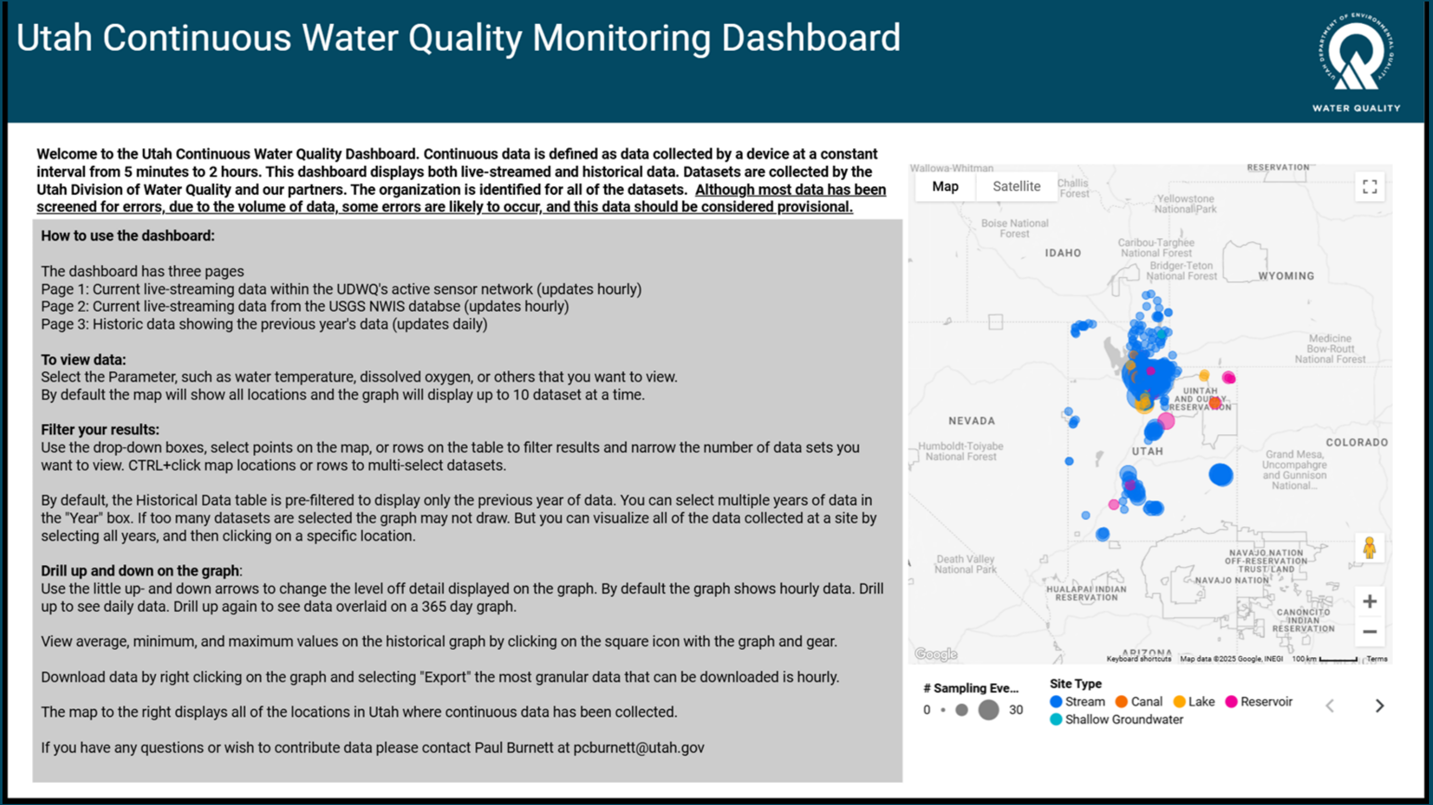
Buoy_Data_from_WQDataLive

Scheduled Query “Process_Buoy_Data_Into_HF_DB_Format”

- Conforms and inserts data to Water_Data_wqDataLive_Transformed

Water_Data_wqDataLive_Transformed

Public Data Accessibility



Utah Continuous Water Quality Monitoring Dashboard

Welcome to the Utah Continuous Water Quality Dashboard. Continuous data is defined as data collected by a device at a constant interval from 5 minutes to 2 hours. This dashboard displays both live-streamed and historical data. Datasets are collected by the Utah Division of Water Quality and our partners. The organization is identified for all of the datasets. Although most data has been screened for errors, due to the volume of data, some errors are likely to occur, and this data should be considered provisional.

How to use the dashboard:

The dashboard has three pages
Page 1: Current live-streaming data within the UDWQ's active sensor network (updates hourly)
Page 2: Current live-streaming data from the USGS NWIS databse (updates hourly)
Page 3: Historic data showing the previous year's data (updates daily)

To view data:
Select the Parameter, such as water temperature, dissolved oxygen, or others that you want to view.
By default the map will show all locations and the graph will display up to 10 dataset at a time.

Filter your results:
Use the drop-down boxes, select points on the map, or rows on the table to filter results and narrow the number of data sets you want to view. CTRL+click map locations or rows to multi-select datasets.

By default, the Historical Data table is pre-filtered to display only the previous year of data. You can select multiple years of data in the "Year" box. If too many datasets are selected the graph may not draw. But you can visualize all of the data collected at a site by selecting all years, and then clicking on a specific location.

Drill up and down on the graph:
Use the little up- and down arrows to change the level off detail displayed on the graph. By default the graph shows hourly data. Drill up to see daily data. Drill up again to see data overlaid on a 365 day graph.

View average, minimum, and maximum values on the historical graph by clicking on the square icon with the graph and gear.

Download data by right clicking on the graph and selecting "Export" the most granular data that can be downloaded is hourly.

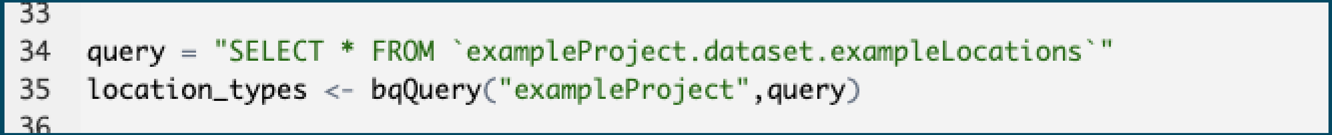
The map to the right displays all of the locations in Utah where continuous data has been collected.

If you have any questions or wish to contribute data please contact Paul Burnett at pcburnett@utah.gov

Map
Satellite
IDAHO
WYOMING
NEVADA
UTAH
COLORADO
ARIZONA
Yellowstone National Park
Boise National Forest
Caribou-Targhee National Forest
Bridger-Teton National Forest
Medicine Bow-Routt National Forest
UINTAH AND OURAY RESERVATION
Humboldt-Toiyabe National Forest
Grand Mesa, Uncompahgre and Gunnison National...
Death Valley National Park
HUALAPAI INDIAN RESERVATION
NAVAJO NATION OFF-RESERVATION TRUST LAND
NAVAJO NATION
CANONCITO INDIAN RESERVATION
Google
Map data ©2025 Google, INEGI
100 km
# Sampling Eve...
0
30
Site Type
● Stream
● Canal
● Lake
● Reservoir
● Shallow Groundwater

Seamless Access to Continuous Data

- Pull millions of rows with just a few lines of R code:

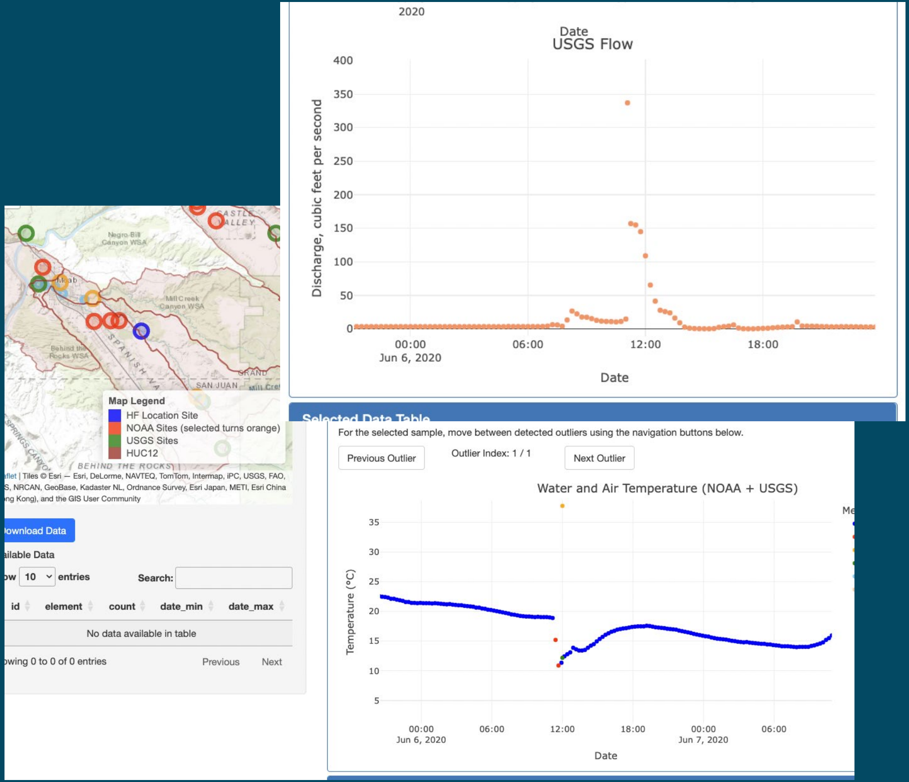


```
query = "SELECT * FROM `exampleProject.dataset.exampleLocations`"
location_types <- bqQuery("exampleProject",query)
```

- Access data from:
  - Deployed loggers
  - USGS sites
  - NOAA weather data
- Fast, centralized access supports efficient workflows

Example Workflow : QC Data

- Workflow:
  - Pull data from DB into RStudio
  - Load into dashboard for visualization
  - Overlay continuous data with contextual data
  - Identify and flag issues (e.g., out - of-water sensors)
  - Push QC flags back into BigQuery
- Visual checks improve confidence in the QA process



Map Legend

- 🟦 HF Location Site
- 🟥 NOAA Sites (selected turns orange)
- 🟩 USGS Sites
- 🟫 HUC12

aflet | Tiles © Esri — Esri, DeLorme, NAVTEQ, TomTom, Intermap, iPC, USGS, FAO, S, NRCAN, GeoBase, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China ng Kong), and the GIS User Community

ownload Data

ailable Data

ow 10 entries Search:

id	element	count	date_min	date_max
No data available in table				

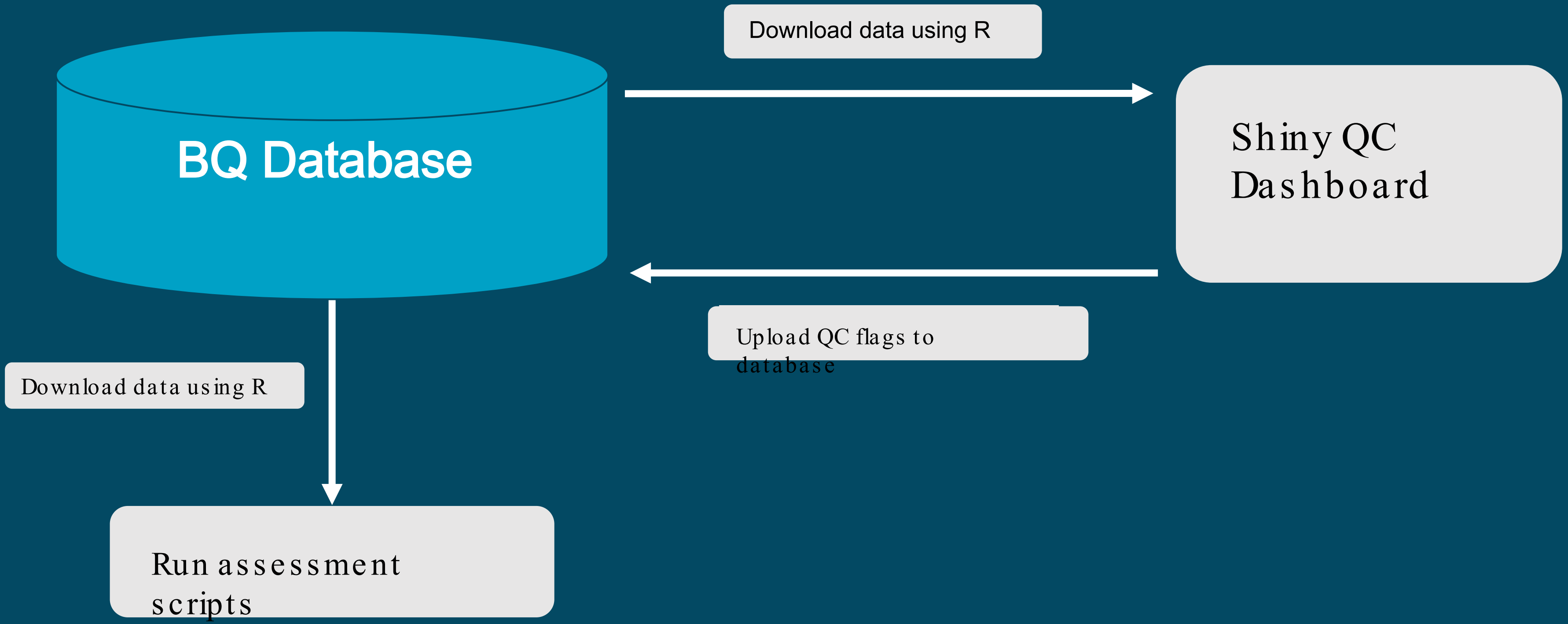
owing 0 to 0 of 0 entries Previous Next

Selected Data Table

For the selected sample, move between detected outliers using the navigation buttons below.

Previous Outlier Outlier Index: 1 / 1 Next Outlier

Turning QC Data into Assessment



Closing the Loop

- BigQuery Database
    1. Fast access
    2. Flexible
    3. Can be complex
- Takeaway: Cloud tools enable access to high-resolution water quality data
- Current Data Volumes:
    1. 1800 datasets (including ~340 USGS sites)
    2. Water Data: 29M rows
    3. Streaming Data: 1M rows
    4. USGS Data: 18M rows
- Looker studio dashboards can be developed with minimal code
- Monthly Cost: ($25-30) - January 1-June 2, 2025 = $127.05