Recognizing Value in an Open-Architecture Digital Water System for Research Projects: Managing Time Series and Big Data

By Shawn Dent and Pierre Mishra

Like much of the water and wastewater industry, research projects are experiencing a digital transformation. The management and visualization of "Big Data," specifically extensive amounts of time series data, can be a significant challenge. To meet this challenge, Carollo has developed an open-architecture, cloud-based solution to help researchers better collect, manage, analyze, visualize and integrate raw data to make data driven inquiries and decisions.

A Digital Water System (DWS) is a combination of software, database, web applications, and data pipelines forming an integrated system for organizing, processing and visualizing data. A DWS may manage many different types of data across an organization, however, this paper will focus on the management of the extensive amount of time series data that can be generated by any water/wastewater treatment research project.

Introduction

Many water and wastewater treatment research projects require the collection of significant amounts of data. In many cases, a Supervisory Control and Data Acquisition (SCADA) system is used to collect streaming data such as flows, concentrations, temperature, pH, etc. that are measured with meters and sensors. Researchers use databases to manage these continuous data. Data for other parameters that cannot be measured continuously are collected through grab samples, which are then processed in a laboratory and the measurements are returned to the researcher on a consistent basis. If the grab sampling data can also be integrated into the database with the SCADA data, a full picture of the monitored facilities can come into focus.

The volume of data can grow significantly as a research project progresses. Although spreadsheets have historically been used to manage datasets, they are limited in both functionality and data size. The use of spreadsheets for collection, management, analysis and visualization adds a significant amount of time to any researcher's budget and can distract from the ultimate purpose of the research project.

To overcome these limitations, a DWS can be developed and deployed to efficiently extract, transform and load (ETL) data into a cloud database, which can then be connected to industry standard dashboards. This approach not only facilitates data management, but also allows researchers to analyze and visualize the data across any time interval (e.g., minute, hours, days) and time period (e.g., June through September).

Approach to DWS Development

A DWS is generally defined as a combination of software, database and web applications to form an integrated system for organizing, processing and visualizing planning, operational and management of water-related data to help make data-driven decisions. Many research projects, such as development and testing of new treatment processes at a pilot level, may only be operated for a limited time period (e.g., months). With limited research budgets, these projects may not allow comprehensive data management software to be purchased and used, thus spreadsheets become the only answer to manage time series data. However, once the project is started, it can become apparent that a scalable system is needed to efficiently and effectively turn the raw data into information.

Existing Solutions

Desktop and cloud-based solutions do exist as commercial-off-the-shelf (COTS) software solutions specifically purpose-built for time series data and have been available for many years. However, these software packages can be expensive and beyond the budget of many research projects. These products have historically been desktop solutions that are charged per seat, but many vendors are now changing to the software-as-a-service (SaaS), or subscription, model.

The SaaS model does provide many advantages to a desktop solution. They are easier to maintain and update, they are potentially less costly if a monthly subscription can be used that ends after a year, and they usually provide 24/7 technical support. However, any COTS will have the limitation of analyses that are built into the software. If the software cannot readily be customized and configured (for example, directly using Python code), the researcher may then have to export certain data to be analyzed in a spreadsheet or an on-premises database such as one of the many versions of SQL database such as Microsoft SQL Server, PostgreSQL, MySQL and SQLite.

Selected Solution

To overcome many of the disadvantages of spreadsheets and COTS software, Carollo has developed an open-architecture structure using standard Microsoft cloud resources that can provide researchers (or entire utilities for that matter) a system to manage extensive amounts of data using a cloud database/tool, a dashboard and low-code programming. By not relying on a COTS solution, the researcher can store, manage, scale and share their raw data, QA/QC data, and analyzed time series data with a dashboard(s) through a web browser. Similar architecture can also be developed on other leading cloud providers such as Amazon Web Services (AWS) and Google Cloud Platform (GCP).

The cloud database (e.g., Azure SQL) acts as a central location (data warehouse) to compile select data from each individual on-premises database. This can be accomplished in many ways, but the use of Azure SQL and other Azure tools (e.g., Azure Blob Storage, Azure Databricks, Azure Data Factory) in the cloud now provides a very efficient way to create a data processing pipeline to:

- ingest select data from multiple on-premises systems and other data sources
- transform and blend that data together
- analyze the data
- create star-schema data models to visualize through dynamic dashboards
- help develop new insights

Using Azure cloud data tools coupled with Power BI Pro is an extremely efficient, inexpensive and flexible way to process and share data. With the simple use of a web browser to access these cloud dashboards, researchers have the data and information in one location so they can blend, analyze and extract data to make it easier to develop informed results.

To define the key issues that need to be incorporated into the planning, design and construction of a DWS, design and performance criteria were developed (***Table 1***). The goal is to develop a DWS that adheres to as many of these requirements as possible.

The DWS can be described using an organized interconnected system of six elements that are common to any DWS. These elements include collection, management, analytics, visualization, integration and security. Whether a project is small or large, straightforward or complex, these

continued on page 44

continued from page 43

Table 1. General DWS Requirements.

Requirement	Description of System Requirements
Maximize use of Non Proprietary Applications	Limit custom programmed apps. Thoroughly document any programmed apps (e.g., custom APIs).
Maximize No Code/Low Code	Use COTS software as much as possible (e.g., drag and drop features of Power BI).
Use Standard Technology	As much as possible, use technology that is "standard" within the industry (e.g., Microsoft Azure, PowerBI).
Provide Transferability	DWS should be easily transferable from developers' cloud to client's cloud at the end of project.
Maximize Accessibility	Dashboards, editing applications, online GIS, etc. should be easily accessible to the client and project teams using a standard web browser (e.g., use of a landing page and URLs to access dashboards).
Provide Scalability	Database used for DWS should be easily scalable to accommodate future data types, quantities and qualities.
Maximize Efficiency	System must work quickly and serve data to multiple users (with little lag) no matter the size of the database.
Provide Flexibility/ Interoperability	System can easily be connected to outside software systems through I/O exchange (such as COTS or custom programmed apps).
Provide Mobile Options	Dashboards and apps work on desktop computers as well as mobile devices with minimal reworking.
Maximize Security	Minimize threats in the system that could potentially cause data breaches, especially when data is exchanged with outside applications.
Start with Low-Cost Solutions	Start with inexpensive solutions and move to higher cost COTS and custom software when prudent.

six elements are critical to help define detailed data framework components that make up successful digital water research projects.

Structuring the DWS with Azure cloud (database and tools) as the "hub" and the other applications as the "spokes" provides for maximum flexibility and scalability. Purpose-built software used in a research project (e.g., SCADA, web Data, IoT and LIMS) move data in one direction to the database hub. Data leaving the hub is mainly connected to visualization applications (e.g., dashboards, reporting software).

The two-way data exchange with the hub is included for analytical models (e.g., statistics, models, AI/ML). This two-way data exchange is an important feature because an analytic engine, model, or algorithm can be separated but integrated in this space to connect directly with the database. By reading input/output (I/O) data, and performing complex computations outside the database, the two-way data exchange provides extensive flexibility in using any model with this DWS framework. Any other analytic engine (algorithm) where data can be input from the cloud and output returned to the cloud can work. It also allows other specific data (e.g., SCADA data) to be pushed to the models through the database and could provide for near real-time analysis and projections. ***Figure 1*** illustrates one example of how these systems can be integrated to form a DWS for research projects.

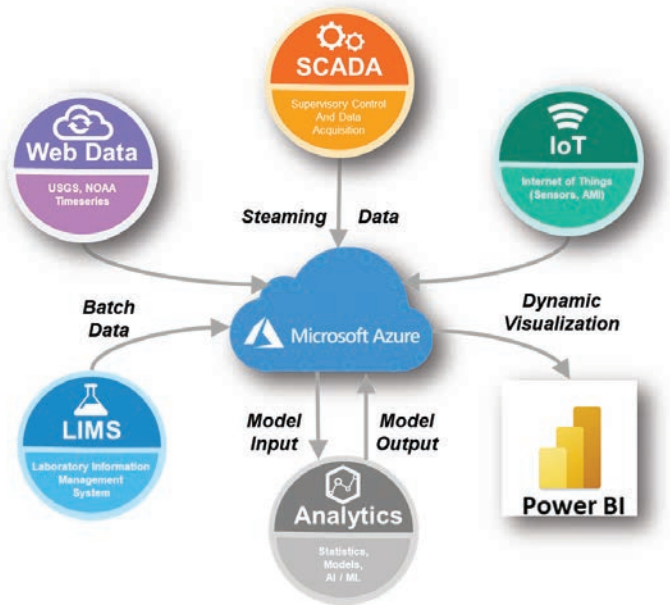

Figure 1. DWS interconnected system and dataflows. *Carollo Engineers*

Results

A "use case" in this DWS refers to an individual application that was developed to fulfill specific needs, such as management of multiple sources of data for a research project. Defining this use case helps develop data pipelines, dashboards and a data warehouse that serves specific data to each dashboard. Raw data was stored in Azure Blob Storage, which is used as a data lake to dump all types of structured and unstructured data in their original formats. Data was transformed and analyzed using low-code ETL tools, such as Azure Data Factory or Python/R scripts in Azure Databricks, where low-code ETL was not sufficient due to complex logic and advanced mathematical needs. Data was cleaned and transformed into star-schema tables and was stored in Azure SQL as the backend database. Power BI Pro was used to develop dashboards that connect to this database. This whole data pipeline was orchestrated and scheduled using Azure Data Factory. The data pipeline and dashboards were both developed in or published to the cloud.

The dashboard provides research personnel with better consolidated and reported information, allowing them to see historical trends as well as data in near real time. The researcher chooses the minimum time interval of the time series data (e.g., one hour). Lab results will be included as they are completed (e.g., one day). The SCADA data and lab data can then be integrated so that researchers can examine any parameters measured or sampled at the pilot facility in one tool.

Figure 2 illustrates two tabs within a dashboard used to analyze and visualize data coming from a SCADA system and lab results. The data for this dashboard includes hourly time series data measured at three treatment facilities for over two years, for 34 different parameters. The hourly data are automatically averaged from one-minute data that is measured by the SCADA system (and stored in the SCADA historian database). The SCADA data are then automatically averaged into daily, weekly, monthly and yearly trends. In this case, water quality, flows and precipitation are included in this one dashboard.

The top graphic in Figure 2 shows three measured time series of hourly data for the month of January 2023 measured by a SCADA system and automatically transferred to the DWS every night. The top trend illustrates dissolved oxygen, the middle is flow at the outfall, and the bottom is precipitation. Using this one screen, the user can select any time period between January 2022 and present date (the data are automatically updated every night from the SCADA and LIMS). The user can then select which treatment facility (or train) they want to examine and select a water quality parameter and a flow. These historical trends will help the researcher identify historical patterns in their data and analyze specific variables that will assist in the research objectives.

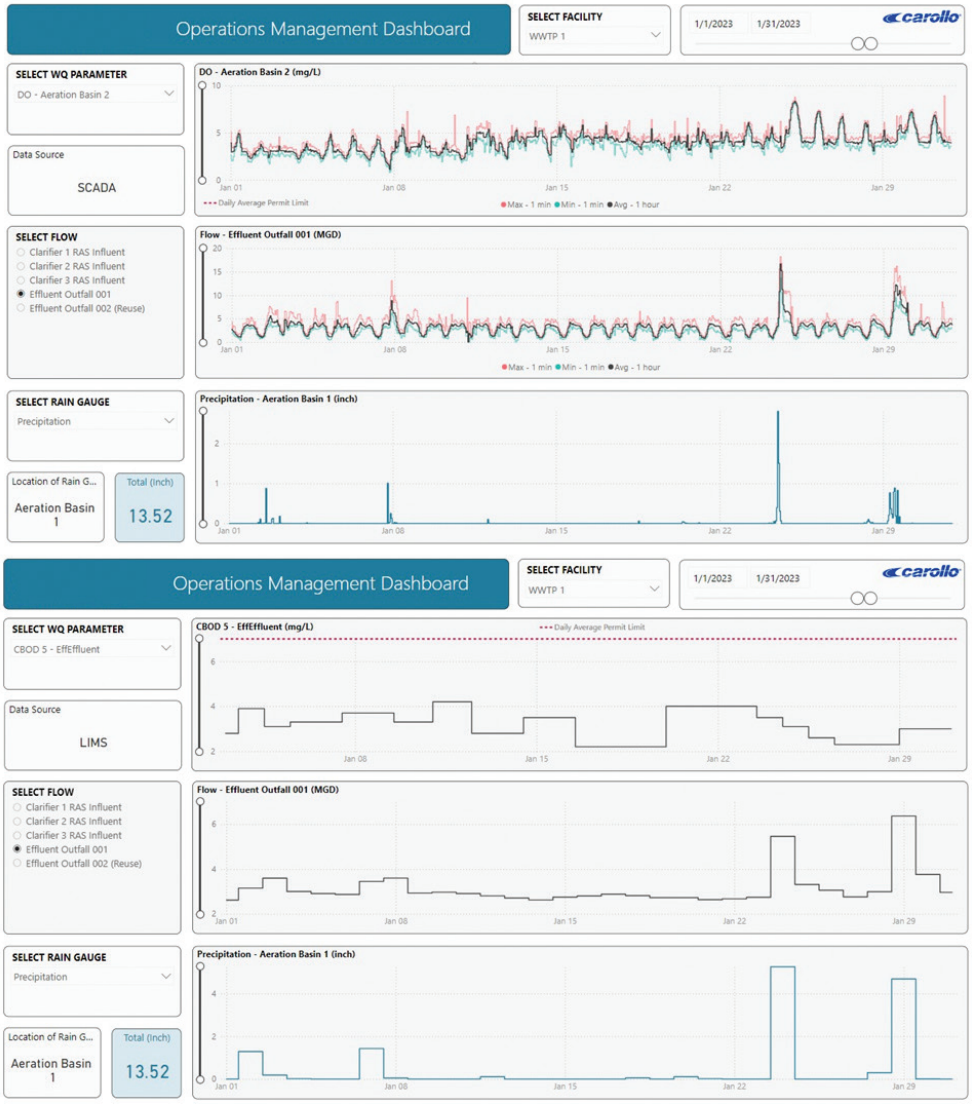

Figure 2. Flow and water quality dashboard examples. *Carollo Engineers*

The bottom graphic in Figure 2 is the same dashboard but is shown in another tab where the data are averaged to daily intervals. The top graph has been changed from dissolved oxygen (DO) to carbonaceous biochemical oxygen demand (CBOD) since CBOD is sampled every two to three days. Any parameters that are measured at the lab are transferred to the same database and therefore can be displayed in the same dashboard along with SCADA data. By visualizing both SCADA measurements (online sensors) and lab measurements on a daily basis, the trends can be further analyzed at any scale. The dashboard also automatically calculates daily, weekly, monthly and yearly averages each night from the hourly data.

The two tabs of the dashboard shown in Figure 2 are just one example of how big data can be managed and visualized with a DWS. An Azure cloud-based solution can be a very effective tool to process, store, analyze and visualize the extensive amounts of time series data that are generated by many research projects. An open architecture system can provide the researcher and colleagues with a more efficient and cost-effective tool to analyze time series data to better draw conclusions about pilot treatment systems.

Shawn Dent, PE, is the principal technology lead for digital water and vice president with Carollo Engineers in Boston, MA, and may be reached at sdent@carollo.com. Pierre Mishra is an analytics and data engineer for the Digital Water Group with Carollo Engineers in Los Angeles, CA, and may be reached at pmishra@carollo.com.