Capstone Development (DS-5999) Course

Comprehensive Drinking Water Database

Table of contents

I. Executive Summary

When multiple sets of data need to be linked together, a database is widely used. Instead of having several data sheets that are difficult to deal with, databases are searchable and sortable for the essential data. The DWJL proposed the idea of a **Comprehensive Drinking Water Database** to contribute to the environmental justice assessment of drinking water quality in the United States. The team gathers data on water quality from a variety of websites and manually combines it. This is a time-consuming operation and hence they needed a database solution with an easy-to-use interface. This database will be used by the research team to answer issues concerning drinking water justice at the national, state, county, and community water system levels. This is important because it will enable the team to answer research questions or easily query filters, and aggregate drinking water data to respond to environmental concerns at various geographic scales. Hence, my capstone project aims to create and deploy a Database **(CDWD)** that directly contributes to the environmental justice assessment of drinking water quality in the United States which the DWJL is conducting. The database will have the capacity to query the data and find answers to research questions associated with drinking water via filters and various options related to assessing drinking water quality. Understanding different geographical scales and units of analysis are the greatest challenge. Also, one of the issues during the initial phase of developing the database is figuring out how to handle and align multiple datasets to aggregate information.

II. Background

Information about safe water is stored in SDWIS, the EPA's data system. The Enforcement and Compliance History Online(ECHO) system incorporates Public Water Systems data from EPA's Safe water Information System(SDWIS) database. The latest SDWIS download has **11 data** files. The first data file has information about the Public Water Systems which includes variables like epa region and population served count. The second data file has site visits which has information about all the visit information including reason, id, date, first and last reported date. The third data file has event milestones which has the identifier, code and the details about milestone events. The fourth data file has facilities which, when used with the PWSID, uniquely identifies a water system facility. The character codes identifying the type of water sources are ground water, surface water and also a variable that indicates whether the water system source is being treated or not. The fifth data file has LCR which has result sign code which indicates if the sample result was below the minimum detection limit or equal to the value reported. The sixth data file has PN Violations which has the violation code, compliance start and end data and the contaminant code. The seventh data file has ANSI information which has the name of the area associated with the ANSI entity and state code. The eighth data file has reference codes which have mapping values to the values and code from other data sheets. The ninth data file has service areas which have the service area type codes. The values of codes and description matches with the reference code value data file and also has a value which indicates if the area is the primary service area served by the water system.

The most important data file is violations and enforcements which has information about the different violation categories, violation measures that checks if the standards are set at a level sufficient to protect the public health. It also has the current state of violation which is identified as resolved, archived and addressed. The violation originator code indicates who the violation was initiated by. There are also codes for rule groups and enforcement details which uniquely identifies the multiple occurrences of an enforcement action. All the data files have the first and last reported date of the particular violation.

The purpose of my capstone as a database project is to reflect the data utilized by the EPA. This is mainly to enable the DWJL research team and collaborators to answer research questions or easily query filters, and aggregate drinking water data to respond to environmental concerns at various geographic scales.

Unit of Analysis Levels :
National | State | County | CWS Levels

III. Data

A. About the Data

The dataset files from SDWIS are refreshed quarterly and were published in July'21 to reflect the latest SDWIS release. The most recent data about the facility evaluations, violations, and enforcement are contained in these files. The dataset download comes organized under a folder named SDWA Dataset on the echo website. Each zip file in this folder contains comma-separated (CSV) files. The folder has 11 files that provide national data for key EPA/State Drinking Water Dashboard metrics, which are some of the SDWIS data elements most commonly used in the enforcement and compliance program. There are repeating data fields like PWSID where a handler can have multiple values of repeating fields. Basic information about each public water system includes the system's name, and ID number. By using the data, evaluation of public water systems such as drinking water violations, and source water can be attained. The main goal is to create an accessible relational database by combining all the separate csv files and making it available for all the members who are interested in drinking water variables.

B. Data Structure

Below is the SDWA Data Download File Structure from the echo epa website.

Data	**File name**
Public Water Systems	SDWA_PUB_WATER_SYSTEMS.csv
Site Visits	SDWA_SITE_VISITS.csv
Events milestones	SDWA_EVENTS_MILESTONES.csv
Facilities	SDWA_FACILITIES.csv
LCR	SDWA_LCR_SAMPLES.csv
PN Violations	SDWA_PN_VIOLATION_ASSOC.csv
ANSI	SDWA_REF_ANSI_AREAS.csv
Reference Codes	SDWA_REF_CODE_VALUES.csv
Service Areas	SDWA_SERVICE_AREAS.csv
Violations and enforcement	SDWA_VIOLATIONS_ENFORCEMENT.csv
Geographic areas	SDWA_GROGRAPHIC_AREAS.csv

C. Data Processing

The data from the download is already accessible and usable for public usage and research purposes. Not much data processing can be done because the end product of my capstone is a fully-functioning database with multiple options for the users to dynamically interact with it. Although standardization of date format was done as date and time are frequently used data types. Certain fields having combinations of values and units were separated for better data storage in tables while creating the database. Inconsistencies in values like format and irrelevant values like hash were removed. There was no outliers or bias in the dataset.

IV. Methodology

Initially, I created a simple application that allowed people to view, interact with, and filter data through a user interface. However, this application loaded the full dataset into memory and handled each user request, which was inefficient and took a long time to complete.As a result, I switched from a filesystem-based to a database-based strategy. Even after I created a shiny database-driven application, there were a few flaws. With a shiny application I was not able to do server-side pagination. This program works with a lot of data, and users frequently need to view reports containing thousands of rows. It would be helpful to have server-side pagination for such queries. As a result, in order to create a comprehensive database tool that provides a positive user experience and allows users to efficiently interact, change, filter, and examine data, I chose to follow the approaches listed below.

a. Develop a database with great performance and low latency.
b. Create a backend that can handle user requests, connect to a data source, and return data in the format requested by the user.
c. Create a user interface that offers a consistent user experience.

a. Database Design

Databases and tables were created for each entity in normalized form. Each csv dataset represents a particular entity and in its normalized form except violation_enforcement dataset. It is the combined table of violation and enforcement entity. It is the mega table with more than 13 million entries. Hence this mega table was normalized and decomposed into two tables named violation and decomposition. Normalization helps to reduce duplicated data from each table, resulting in a proper database architecture and great efficiency while searching the database for information.

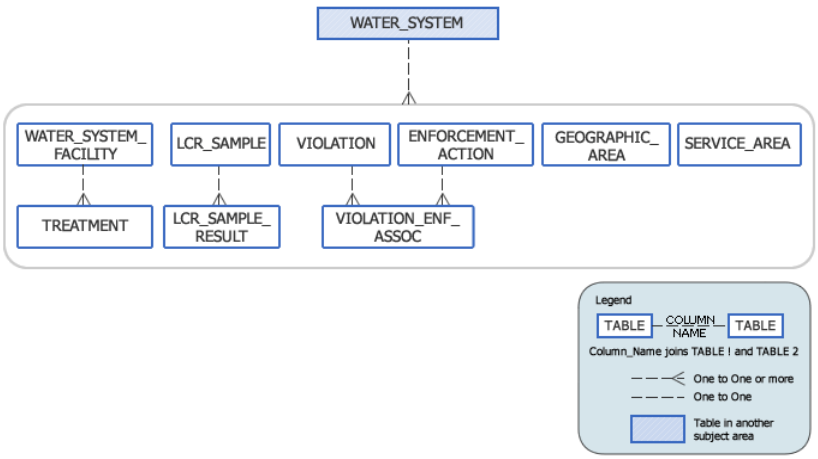


Fig 1 : SDWIS Model

Then, to prevent missing values and wrong values from being put in certain columns, I added the corresponding constraints, which will check the values being inserted into the table and ensure the database's accuracy and reliability. Restrictions were added to the phone number field, email field, zip code, state code, and a few other fields to ensure that they match the corresponding pattern of their attributes, and that incorrect values are not entered in those fields. A primary key was added to each table, which is a database's bare minimum of attributes that uniquely identify rows. Each normalized dataset was sorted by primary key and entered it into the database tables that corresponded to it. Presorting aids in the fast insertion of data into the table. Indexes were added for the columns that may be utilized by users to filter the tables to make data search more efficient and optimized. Filtering queries take a long time to get the relevant data from the database without an index. Filtering is optimized and latency is reduced after the indexes are added, and data is returned in seconds.

b. Backend Development

Node js was used to create the backend, which is responsible for managing the requests made by the user through the UI. Different REST API endpoints are built for various backend functionalities. Each endpoint listens for a specific HTTP request, manages it, and uses the HTTP protocol to respond with the appropriate data. The backend application reads the incoming HTTP request and connects to the database. It then constructs the optimized SQL query for fetching the data. The server-side pagination helps not to overflow the front-end application with huge data, and then responds with the required data in JSON format using the HTTP protocol.

c. Frontend Design

The angular framework was used to create the frontend application, which is the user interface used by end users to review each data table and filter, sort, group, and aggregate them by different columns. After the user filters, groups, and aggregates the dataset, the front-end

application sends a request to the backend application's corresponding endpoint, which returns the required data in the required format.

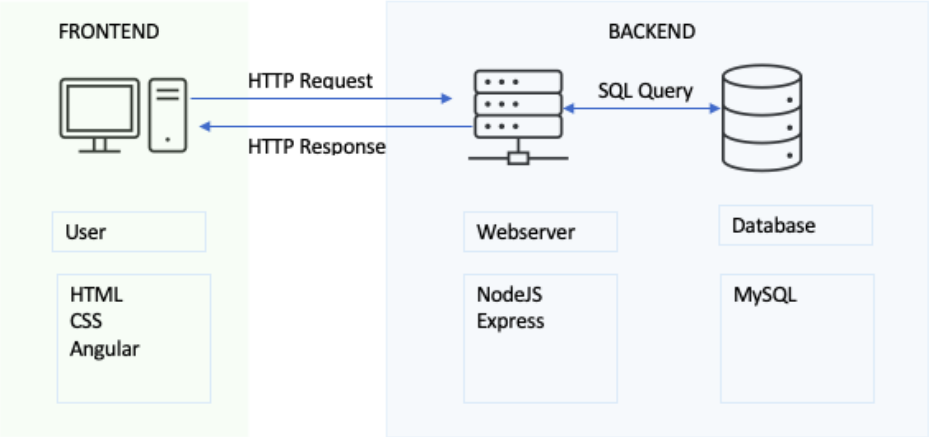
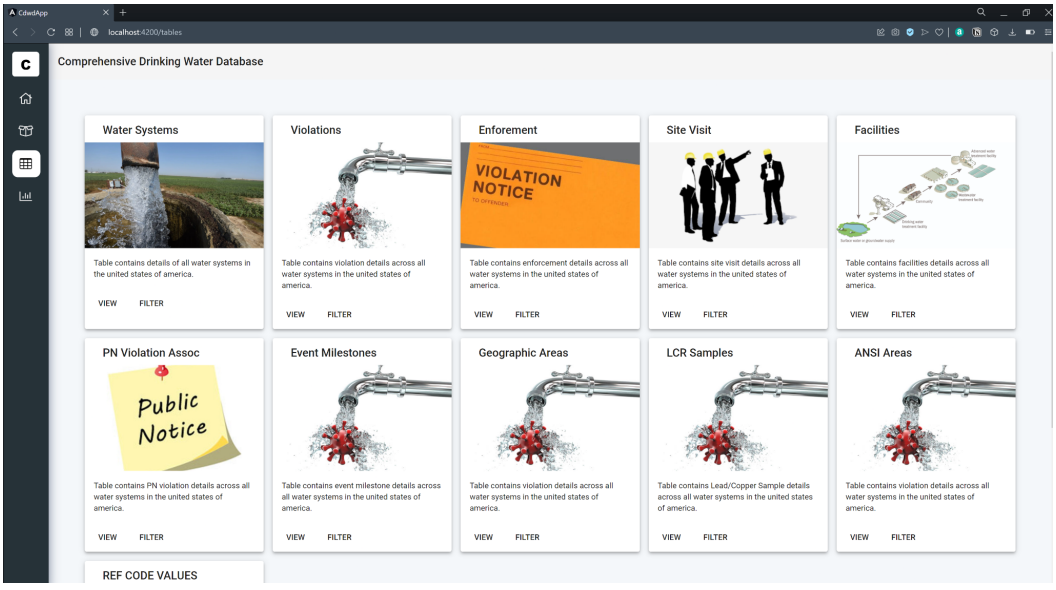


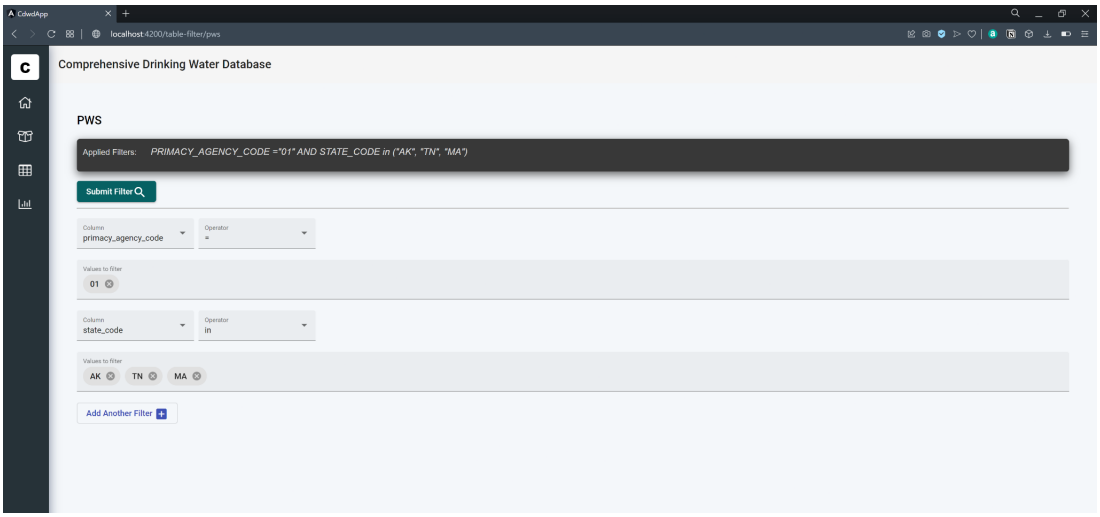
Fig 2 : Architecture of the application

The data is rendered as tables or charts by the front-end application, depending on the user's preference. This allows the user to create any type of report they want in any format they choose. The users can choose which columns should be included in the final report and download the report. As a result, users have complete control over their data and a simple way to generate reports from numerous tables using this application.

V. Results

The Comprehensive Drinking Water Database is delivered as a fully functional database that includes both front-end and back-end features. Users can interact with data by filtering, grouping, aggregating any columns of their choice dynamically, and generating/downloading the report in the required format as a CSV. The front end will look like this :





VI. Conclusion and Next Steps

My capstone project's major purpose was to assist the DWJL and be able to manage massive amounts of data in a way that allows users to properly analyze and draw conclusions from it. The process of creating this database tool involved data extraction, data storage, data cleaning, data preprocessing, data manipulation, data filtering, aggregation, and grouping. The team at DWJL can interact with real-time data through UI, query it with the help of UI elements, and then generate dynamic data reports in the required format which was extracted from different data sources and then transformed and loaded into UI. This database will be hosted for the research group to use in evaluating various drinking water quality measures. It will be maintained by an undergraduate computer science major who is working as a research assistant with DWJL. He will work on updating or making modifications to the database if any additional modifications are made in future. In addition, spatial data will be merged to visualize the violations on a map.

VII. References

ECHO Video Tutorials. https://echo.epa.gov/help/tutorials

Ground Water and Drinking Water. https://www.epa.gov/ground-water-and-drinking-water

SDWIS MODEL | US EPA. https://www.epa.gov/enviro/sdwis-model

SDWA Data Download Summary.
https://echo.epa.gov/tools/data-downloads/sdwa-download-summary