

# A Practical Guide To Analyze Data from the SRBC Water Quality Monitoring

Eswara Sravani Munnangi, Joss C.Steward and Jeonghwa Lee

**Abstract**—Today more data will be gathered and processed than ever before. This is the age of Big Data. However, with huge amounts of data comes huge performance issues. Although many performance problems can be solved simply by throwing more computing power at them, we wish to solve these problems through software optimization instead of hardware improvements. We propose a research project focused on allowing rapid access to large amounts of data on subpar hardware. Our project will focus on the e-client storage and rapid retrieval of infrequently accessed data.

**Website**—<http://www.srbc.cs.ship.edu>

## INTRODUCTION

The sample data for this project will come from the real world. We have decided to use water-quality information gathered by the Susquehanna River Basin Commission and made available to the public through their web site. At the current time, the Susquehanna River Basin Commission (SRBC) provides free access to water quality data gathered prior to December 31, 2011. For data gathered after Presumably, the SRBC does not have the resources to monitor all the stations all the time. Each sample contains the following information: Water Temperature, Specific Conductivity, PH level, Turbidity, and Dissolved Oxygen Content. Each Sample also includes metadata consisting of the time stamp and station name. As far as sheer scale goes, January 1, 2012 there is a dollar 250 fee per quarter<sup>1</sup>. In practice, this means we will be using data gathered prior to December 31, 2011 as our sample data for the foreseeable future.

As far as sheer scale goes, January 1, 2012 there is a dollar 250 fee per quarter<sup>1</sup>. In practice, this means we will be using data gathered prior to December 31, 2011 as our sample data for the foreseeable future. A quick glance at the data gathered in Q1 of 2010 reveals that the SRBC has been gathering water-quality data from 60 stations. The sample rate seems to vary between stations from once every 5 seconds to once every 4 hours. This is not actually an overwhelming amount of data. Assuming one sample every 5 seconds, each station would produce approximately 6,307,200 samples per year. Given 60 stations, that would be 378,432,000 samples per year. This may seem like a lot, but given that the largest SQL Server databases in 2008 were exceeding 1 petabyte in size<sup>2</sup> adding fewer than 400 million rows per year is not actually that much data. However, it should be obvious that not every organization capable of producing data on this scale can afford to run server farms powerful enough to process and store this much data. Ideally, we will be able to store all of the data collected and still be able to access it in reasonable amounts of time on relatively low end hardware. To accurately gauge performance, we will store the data in a indexed SQL Server database and compare the resulting size and performance to any other data storage schemes we come up with. Other possibilities include storing infrequently accessed data in segmented compressed files and storing statistics such as min, max, standard deviation, and etc. on each segment of data in a SQL Server Database. The overall performance may be satisfactory using existing software. If that is the case, we will still attempt to achieve a performance gain through the use of special techniques. Our hope is that this study will aid in the storage and analysis of large amounts of data on slower, more power e-client hardware than is currently required.

## OVERVIEW

The Susquehanna River Basin Commission (SRBC) initiated the establishment of the Remote Water Quality Monitoring Network (RWQMN) in January 2010. The main objective of this Monitoring Network (RWQMN) is to continuously monitor the conditions of rivers and streams located in Northern Tier Pennsylvania and Southern Tier of New York., and also produces reports the water quality conditions, which helps the agency officials to track the quality changes and to take measures.

---

<sup>1</sup>Manuscript received June xx, xxxx; revised June xx, xxxx. This work was supported in part by xxx.

- Eswara Sravani Munnangi is with the Department of Computer Science and Engineering, Shippensburg University, Shippensburg, PA 17257, (e-mail: em8678@ship.edu).
- Joss C. Steward is with the Department of Computer Science and Engineering, Shippensburg University, Shippensburg, PA 17257, (e-mail: js0289@cs.ship.edu)
- Jeonghwa Lee is with the Department of Computer Science and Engineering, Shippensburg University, Shippensburg, PA 17257; To whom correspondence should be addressed. <http://www.cs.ship.edu/~jlee> (e-mail: jlee@ship.edu).

## Stations

The stations are operating in areas where drilling for natural gas is most active, as well as other locations where no drilling activities are planned so SRBC can collect control-data. A contribution from East Resources provided the initial funding for the project. In 2010, the New York State Energy Research and Development Authority provided funding for the expansion of the network into the New York portion of the basin. SRBC is covering the ongoing maintenance costs.

The monitoring network provides constant data collection with instruments sensitive enough to detect subtle changes in water quality on a frequency that will allow background conditions and any changes to them to be documented throughout the year. The following five water quality parameters are being measured at each station:

- Water Temperature
- pH
- Specific Conductance
- Dissolved Oxygen
- Turbidity

## Water Temperature

Water temperature is very important to fish and other aquatic life, as well as for swimmers, fishermen, and industries. Temperature affects the ability of water to hold oxygen, which can affect respiration and an organism's ability to resist certain pollutants. Human induced changes to water temperatures can have a great effect on stream ecosystems. The loss of riparian tree cover can raise stream temperatures through exposure to the sun. In addition, a lot of water in the Susquehanna Watershed is used for cooling purposes in power plants that generate electricity, resulting in warmer water releases back to the environment. The temperature of the released water can affect downstream habitats.

## pH

pH is a measure of water's acidity/basicity. The range goes from 0 - 14, with 7 being neutral. A pH of less than 7 indicates acidity, whereas a pH of greater than 7 indicates basicity. pH is really a measure of the relative amount of free hydrogen and hydroxyl ions in the water. Water that has more free hydrogen ions is acidic, whereas water that has more free hydroxyl ions is basic. Since pH can be affected by chemicals in the water, pH is an important indicator of water that is changing chemically. pH is reported in "logarithmic units," like the Richter scale which measures earthquakes. Each number represents a 10-fold change in the acidity/basicity of the water. Water with a pH of 5 is ten times more acidic than water having a pH of six.

Pollution can change a stream's pH, which in turn can harm animals and plants living in the water. For instance, water coming out of an abandoned coal mine can have a PH of 2, which is very acidic and would definitely affect any fish's health. By using the logarithm scale, this mine-drainage water would be 100,000 times more acidic than neutral water – so stay out of abandoned mines.

## Specific Conductance

Specific conductance is a measure of the ability of water to conduct an electrical current. It is highly dependent on the amount of dissolved solids (such as salt) in the water. Pure water, such as distilled water, will have a very low specific conductance, and sea water will have a high specific conductance. Rainwater often dissolves airborne gasses and airborne dust while it is in the air, and thus often has a higher specific conductance than distilled water. Specific conductance is an important water-quality measurement because it gives a good idea of the amount of dissolved material in the water.

High specific conductance indicates high dissolved-solids concentration; dissolved solids can affect aquatic life, as well as the suitability of water for domestic, industrial, and agricultural uses. At higher levels, drinking water may have an unpleasant taste or odor or may even cause gastrointestinal distress. Additionally, high dissolved-solids concentration can cause deterioration of plumbing fixtures and appliances. Relatively expensive water-treatment processes, such as reverse osmosis, are needed to remove excessive dissolved solids from water.

## Dissolved Oxygen

A relative measure of amount of oxygen that is dissolved in given medium. Oxygen dissolved in the lakes, rivers, and oceans is crucial for the aquatic organisms and creatures living in it. As the amount of dissolved oxygen drops below normal levels in water bodies, the water quality is harmed and creatures begin to die off. Indeed, a water body can "die", a process called eutrophication.

A small amount of oxygen, up to about ten molecules of oxygen per million of water, is actually dissolved in water. This dissolved oxygen is breathed by fish and zooplankton and is needed by them to survive.

Rapidly moving water, such as in a mountain stream or large river, tends to contain a lot of dissolved oxygen, while stagnant water contains little. Bacteria in water can consume oxygen as organic matter decays. Thus, excess organic material in our lakes and rivers can cause an oxygen-deficient situation to occur. Aquatic life can have a hard time in stagnant water that has a lot of rotting, organic material in it, especially in summer, when dissolved-oxygen levels are at a seasonal low.

## Turbidity

Turbidity is the amount of particulate matter that is suspended in water. Turbidity monitors measure the scattering effect that suspended solids have on light: the higher the intensity of scattered light, the higher the turbidity. Material that causes water to be turbid includes:

- clay
- silt
- finely divided organic and inorganic matter
- soluble colored organic compounds
- plankton
- microscopic organisms

Turbidity makes the water cloudy or opaque, and is reported in nephelometric turbidity units (NTU). During periods of low flow (base flow), many rivers are a clear green color, and turbidities are low, usually less than 10 NTU. During a rainstorm, particles from the surrounding land are washed into the river making the water a muddy brown color, indicating water that has higher turbidity values. Also, during high flows, water velocities are faster and water volumes are higher, which can more easily stir up and suspend material from the stream bed, causing higher turbidities.

The picture with the three glass vials shows turbidity standards of 5, 50, and 500 NTUs. Once the meter is calibrated to correctly read these standards, the turbidity of a water sample can be taken. Turbidity can be measured in the laboratory and also on-site in the river. A handheld turbidity meter (left-side picture) measures turbidity of a water sample. The meter is calibrated using standard samples from the meter manufacturer. The picture with the three glass vials shows turbidity standards of 5, 50, and 500 NTUs. Once the meter is calibrated to correctly read these standards, the turbidity of a water sample can be taken. State-of-the-art turbidity meters (left-side picture) are beginning to be installed in rivers to provide an instantaneous turbidity reading. The right-side picture shows a close-up of the meter. The large tube is the turbidity sensor; it reads turbidity in the river by shining a light into the water and reading how much light is reflected back to the sensor. The smaller tube contains a conductivity sensor to measure electrical conductance of the water, which is strongly influenced by dissolved solids (the two holes) and a temperature gauge (the metal rod).

# DataBase Design

The database is fairly simple, consisting of only two tables. One table is used for the Water Quality Data and the other one is used for the Station metadata. This was done to prevent repetition of the station metadata in the Water Quality Data table.

## 0.1 WaterQualityData Table

The Water Quality Data table contains quantitative data on the river water as measured at each station. Each sample includes a measurement time and a station ID keyed to the Station Metadata table. This table is indexed on StationID and SampleTime to provide rapid data retrieval. StationID provides the primary index. This was done based on the assumption that user data access would primarily request data grouped by station. SampleTime is the secondary index. This was done based on the assumption that SampleTime would typically provide the sort order for requested data.

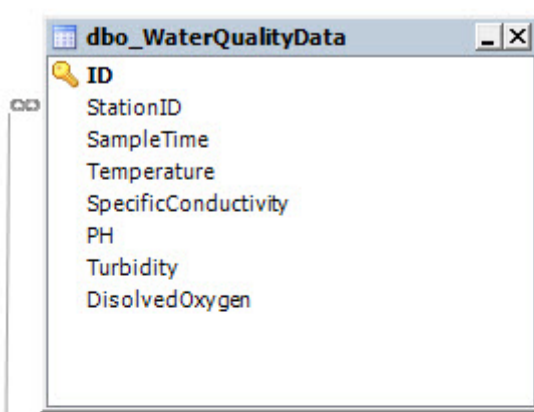
Column Name	Data Type	Description
ID	Bigint	A autoincrementing field that provides a unique ID for each record
StationID	Int	Stores the station ID
SampleTime	Date Time	Stores the time the measurement was taken
Temperature	Float	Store the temperate value
SpecificConductivity	Float	Stores the specific conductivity value
pH	Float	Stores the pH value
Turbidity	Float	Stores the turbidity value
DissolvedOxygen	Float	Stores the Dissolved Oxygen Value

## 0.2 StationMetadata Table

The Station Metadata table contains general metadata on each water sampling station, including the name and coordinates of each station.

Column Name	Data Type	Description
ID	Bigint	A autoincrementing field that provides a unique ID for each record
StationName	char	Stores the station name
SampleLocation	Geography	Stores Latitude and Longitude of the station

## 0.3 Schema Diagram



# GRAPHICAL USER INTERFACE (GUI):

The tool is very user friendly and intuitive and uses a GUI interface implemented in Aspx to communicate with the user. Various features are self explanatory. Forms are easy to fill in and components can be added, removed and updated very easily through a single dialog box.

List boxes are used to display all the components at once so that user can see all the components of a particular type at once. One can just select the component and modify and remove the component.



[Home](#) [SRBC](#)

## SUSQUEHANNA RIVER BASIN COMMISSION

The screenshot shows a web application interface for the Susquehanna River Basin Commission. On the left, there are two panels: 'Stations' and 'Time Span'. The 'Stations' panel contains a list of seven creeks with checkboxes: Apalachin Creek, Baldwin Creek, Blockhouse Creek, Bobs Creek, Bowman Creek, Canacadea Creek, and Cherry Valley. The 'Time Span' panel includes text boxes for 'Start Date' and 'End Date', a 'Histogram Time Step' dropdown menu set to 'Minutes', and a 'GRAPH' button. The main area on the right has a tabbed interface with tabs for 'Temperature', 'PH', 'Specific Conductivity', 'Turbidity', and 'Dissolved Oxygen'. The 'Temperature' tab is selected, showing a large empty graph area. A vertical scrollbar is visible on the right side of the graph area.

### 1. Stations Panel



This panel is an Checkbox list which is used to select multiple Stations to monitor the data.

### 2. Time Span Panel



This panel is used to select the date range for the graph to be plotted.

Start Date textbox and End Date text box is an Ajax Calender popout textbox which enables us to select the date from the calendar in the format MM/DD/YYYY

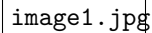


Histogram Time step is a parameter which will enable us to plot the graph by grouping the data specified in the time step intervals.

A placeholder box for image6.jpg.

Time step can be selected from any of the parameters from the dropdown list.

### 3.Graph Panel

A placeholder box for image1.jpg.

This panel is an Ajax Tabbed Container which allows to display only one panel at a time. Monitoring Parameters are listed for tabs.

A placeholder box for image4.jpg.

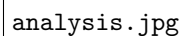
Monitoring Parameters can be any of the following

- Water Temperature
- pH
- Specific Conductance
- Turbidity
- Dissolved Oxygen

Expalanation about the parameters is detailed in section Overview.

### 4. Analysis Panel

This panel is populated with the statistics data once graph is generated for better analysis.

A placeholder box for analysis.jpg.

**A snapshot of the analysis for two stations:**

A placeholder box for image7.jpg.

## Formulae:

Formulas used to generate Statistical Analysis Data:

- *Minimum:* The smallest value in the given range of specified data between start Date and End Date.
- *Maximum:* The largest values in the given range of data specified between Start Date and End Date.
- *Mean:* The mean or average when the context is clear is the sum of a collection of numbers divided by the number of numbers in the collection.
- *Standard Deviation:* The standard deviation (SD) (represented by the Greek letter sigma  $\sigma$ ) measures the amount of variation or dispersion from the average. A low standard deviation indicates that the data points tend to be very close to the mean (also called expected value); a high standard deviation indicates that the data points are spread out over a large range of values.
- *Median:* The median is the numerical value separating the higher half of a data sample, from the lower half. The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one
- *Q1:* The first quartile (Q1) is defined as the middle number between the smallest number and the median of the data set.
- *Q2:* The second quartile (Q2) is the median of the data.
- *Q3:* The third quartile (Q3) is the middle value between the median and the highest value of the data set.