*CEE 6110 Assignment #4 database implementation and loading data - resubmission*

*Karun Joseph*

*A02240287*

Database implementation and loading Logan River data

Introduction

Multiple aquatic monitoring sites in the Logan River watershed GAMUT network continuously stream sensor data. A database management system would allow these observational data and additional third party data to be stored and managed effectively. A physical data model provides a degree of structure that would allow researchers to easily retrieve such data and facilitate scalable software development. This reports identifies and recommends a database management system with data model, their physical implementation and methods to load watershed data.

Methods

Sample Logan River aquatic monitoring site data and metadata were obtained from the GAMUT network. A logical data model was first built which identified entities, attributes and their relationships. CUAHSI Hydrologic Information Systems’ Observations Data Model (ODM) (Horsburgh et al. 2008), a relational based data model tailored for environmental and water resources data, was the chosen data model. MySQL Workbench, an open source relational Structured Query Language (SQL) based database management system, physically implemented the ODM data model by defining the respective entities, attributes and relationships. An SQL script created a database based on the ODM model with an Entity Relationship (ER) diagram shown in Fig. 1. Logan River’s table based data such as ‘sites’, ‘methods’ and ‘sources’ were loaded into the ODM database using MySQL’s ‘Load Data Local Infile’ function. Since the data values of measured variables were not in the right format to be loaded into the database’s ‘datavalues’ table, Python’s Pandas package was used to reformat and remove data inconsistencies. Temperature, specific conductance and pH data values were loaded as an example. Fig. 2, 3 and 4 shows time series of the measured temperature, specific conductance and pH and Fig. 5 a box/ whisker plot of measured specific conductance and pH respectively. A USGS stream gage on Logan River’s stream flow data was also loaded into the database and visualized, to demonstrate the ease by which third party sources’ data could be integrated into the model (Fig. 6).

Results

ODM database was able to capture all of the data and metadata characteristics of the given Logan River watershed data. Data reformatting for both the Logan River and USGS streamflow data was a required step. ODM Tools Python software was an effective means to perform basic analysis and visualization. A Python wrapper to interact with ODM data and treat them as Python objects for advanced analysis, could be a valuable addition.

Conclusions

ODM based relational database is an effective framework to organize and store Logan River sensor data and metadata. A standard data model like ODM has advantages of being an established model, along with their supporting software tools. However, like every other relational database models, ODM may prove to be too structured and opinionated as the project progresses. This report recommends the combined use of MySQL workbench, ODM and ODM Tools Python for Logan River watershed GAMUT network’s data management.

References

Horsburgh, Jeffery S., David G. Tarboton, David R. Maidment, and Ilya Zaslavsky. 2008. “A Relational Model for Environmental and Water Resources Data.” *Water Resources Research* 44 (5): W05406. doi:10.1029/2007WR006392.

Appendix A: Figures

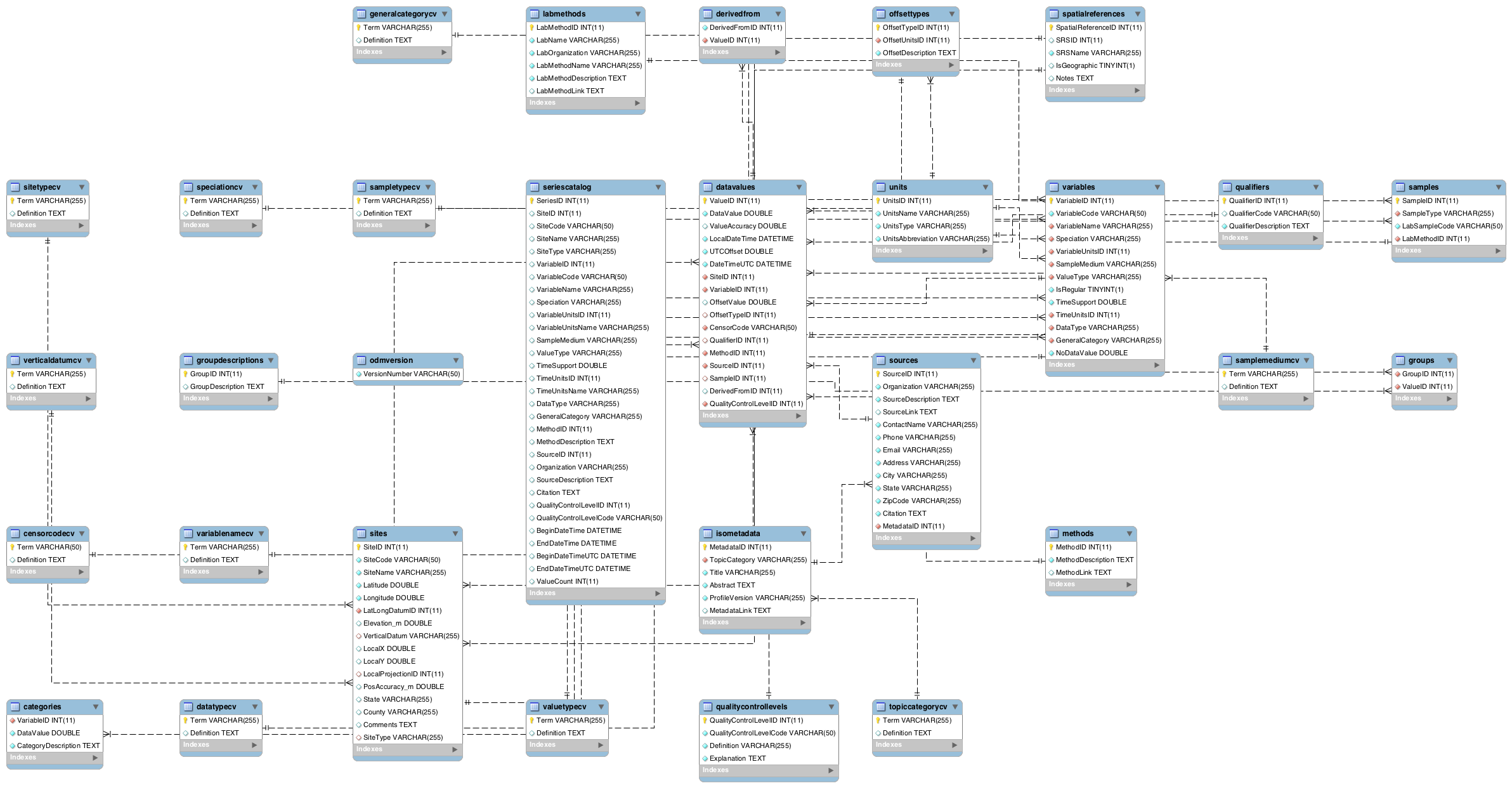


Fig 1: Entity Relationship (ER) diagram of ODM model created in MySQL Workbench

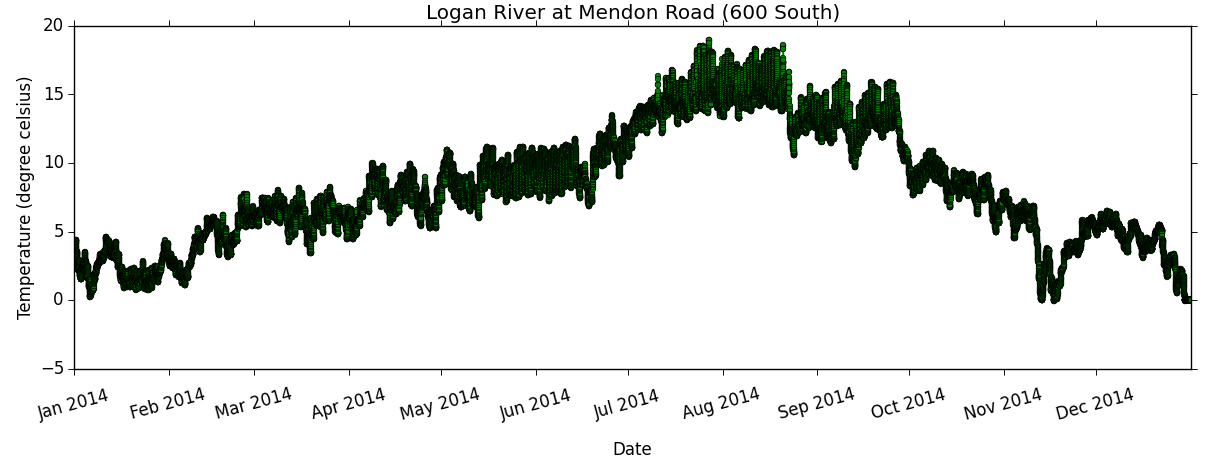


Fig 2: Measured temperature at Logan River site

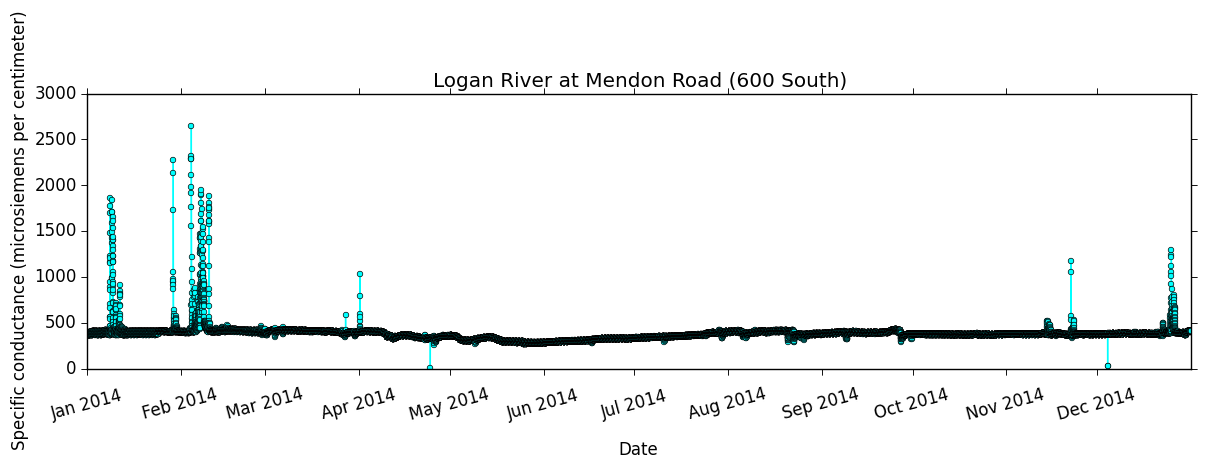


Fig. 3: Measured specific conductance at Logan River site

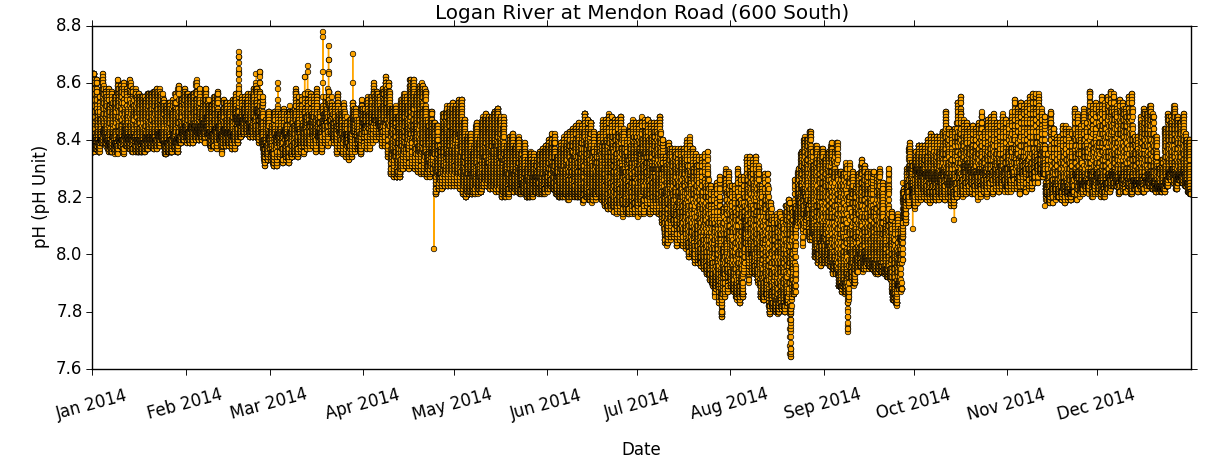


Fig. 4: Measured pH at Logan River site

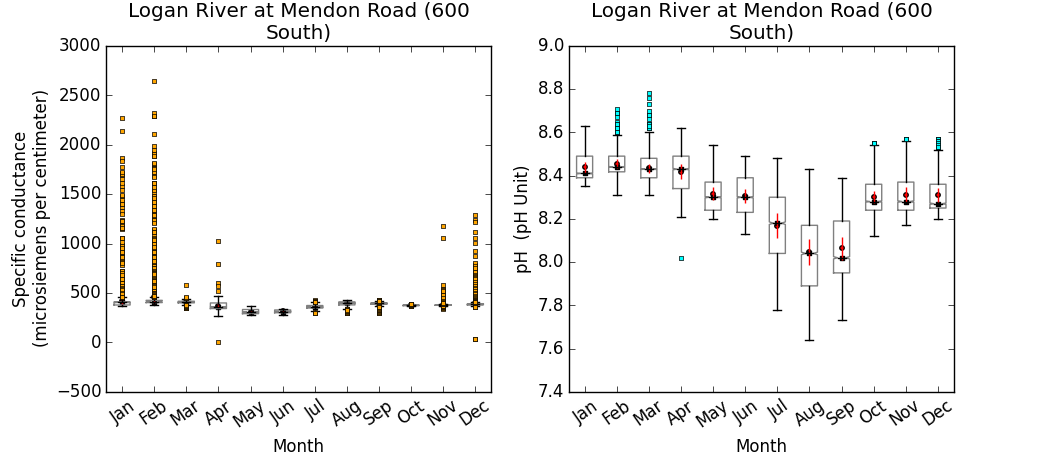


Fig. 5: Box/ whisker plots of measured specific conductance and pH at Logan River site

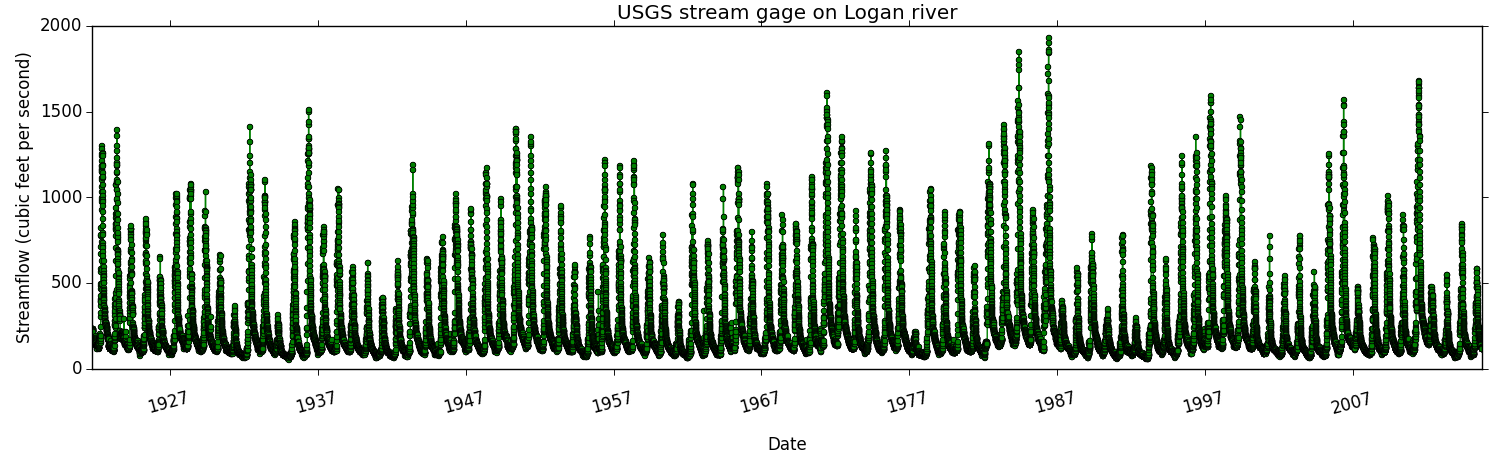


Fig. 6: Stream flow measured at Logan River site ‘10109001’by USGS