Researchers recently deployed several instruments to continuously monitor stream flow and water quality at two sites in the Little Bear River, Cache Valley, Utah.  The instruments report measurements in numerous [datalogger and metadata files](https://sjsu.instructure.com/courses/1488212/files/69178200/download) that the researchers must now regularly retrieve, organize, quality-control/quality-check, aggregate, and manipulate before they can perform further analysis.

* Design a data model to represent the data and its associated metadata
* Use MySQL Workbench to create an entity relationship diagram of your data model
* Upload a one-page briefing report along with a full-page entity-relationship diagram that shows your logical model design

1. Describe the entities and relationships that you have included in your data model.
   * For the model to store the water quality monitoring data for two locations, we have identified the entities and structured them in a snowflake schema as below:

Diagram

Description automatically generated

* + The fact table, dimension tables and sub-dimension table are defined in the snowflake schema
  + We have identified following entities:
    1. Data Values:
       - The Data Values table records the water quality monitoring data for respective site location.
       - The attributes for Data Values are: ValueID, SiteID, SensorID, VariableID, DataValue, LocalDateTime, UTCOffset, DateTimeUTC, OffsetValue, and OffsetTypeID
       - The primary key for this table is the Value ID.
       - This table has foreign keys from the following tables:
         * SiteID in Sites table
         * SensorID in Sensors table
         * VariableID in Variables table
         * OffsetTypeID in OffsetTypes table
    2. Sensors:
       - The Sensors table stores the information on the water quality sensor used at the monitoring site for measuring different variables.
       - The attributes for the sensors table are: SensorID, SensorName, SampleTime, UoM, VariableName, and ColumnName.
       - The primary key for this table is the SensorID.
    3. Sites:
       - The Sites table stores the site details where the water quality monitoring is conducted.
       - The attributes for the sites table are: SiteID, SiteCode, SiteName, Latitude, Longitude, Elevation, State, County, and Comments.
       - The primary key for this table is the SiteID.
    4. Variables:
       - The Variables table stores the information on the variable (such as Water temperature) that is measured and the column (such as LevelTemp\_Avg) details on the measurement activity.
       - The attributes for the variables table are: VariableID, VariableCode, VariableName, ColumnName, VariableUnitsID, SampleMedium, ValueType, TimeSupport, TimeUnitsID, DataType, and NoDataValue.
       - The primary key of this table is the VariableID.
       - This table has foreign key from the following tables:
         * UnitsID in Units table
    5. Offset Type:
       - The table stores the details of the offset location details of the site location.
       - The attributes for the Offset Type table are: OffsetTypeID, OffsetUnitsID, and OffsetDescription.
       - The primary key of this table is the OffsetTypeID
       - This table has foreign key from the following tables:
         * UnitsID in Units table
    6. Units:
       - The Units table stores all the unit formats, and the data type details.
       - The attributes for the Units table are: UnitsID, UnitsName, UnitsType, and UnitsAbbreviation.
       - The primary key for the Units table is the UnitsID

1. Explain how you will structure the metadata to avoid repetition.
   * In the datalogger we have two raw files for two sites with the same metadata. As the record values overlap in both the files, we will need additional fields to distinguish whether the record is from the sites Little Bear River at Mendon Road or Little Bear River Near Wellsville.
   * After the data transformation, when we extract the information in the datavalues table, we will have unique records for each site, variable measured, and the sensor used for the measurement. This will help in avoiding repetition.
2. Overview the software technology, file formats, etc. you will use to organize the data and implement your data model.

We made use of the following to implement the data model :

1. Data extraction from datalogger as .dat file format.
2. We converted the .dat files that were collected from different site locations to CSV to understand the given data and data transformation.
3. MySQL Workbench - We used MySQL Workbench software to create a database, define schema and to generate entity relationship (ER) diagrams.
4. SQL - We used SQL queries to define the tables

Overall, we have made use of the relational database management system (DBMS) to organize and implement our data model.

1. Describe how you could make it easier to get data into and out of your data model.
   * Data Input:
     1. Create a data logging and monitoring system in the cloud.
     2. Store the measurement data into the cloud.
     3. Retrieve the data using MySQL.
   * Data Output:
     1. Link to a data visualization tool (such as Tableau, Power BI, etc.) to filter and extract data.
     2. We can then generate weekly/monthly reports and also have a real-time monitoring dashboard of the water quality measures.
2. Specify whether your data model design will facilitate querying and retrieval of subsets of data.
   * Yes, with the use of the snowflake data model, we will be able to query and retrieve the subset of the data.
3. Provide an entity relationship diagram that shows the entities needed to describe the data, their attributes, and the relationships between them.
   * In Solution I. we covered the entity and relationships identified in the model.
   * Relationships between the entities:
     1. Sites and Datavalues

Diagram

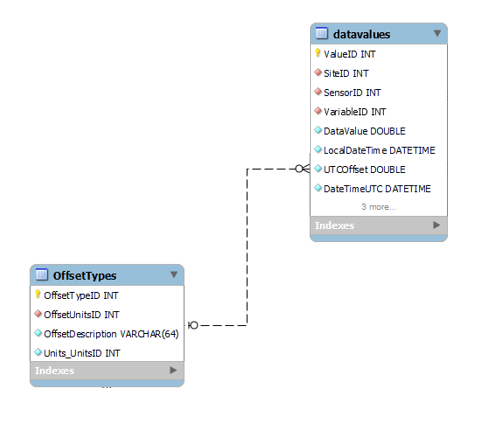
Description automatically generated

* Left to Right: A site has 0 or more datavalues of data
* Right to Left: A datavalues is measured at 1 and only 1 site
  + 1. Datavalues and Sensors

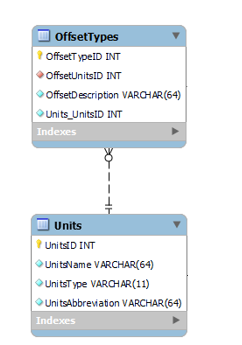
Diagram

Description automatically generated

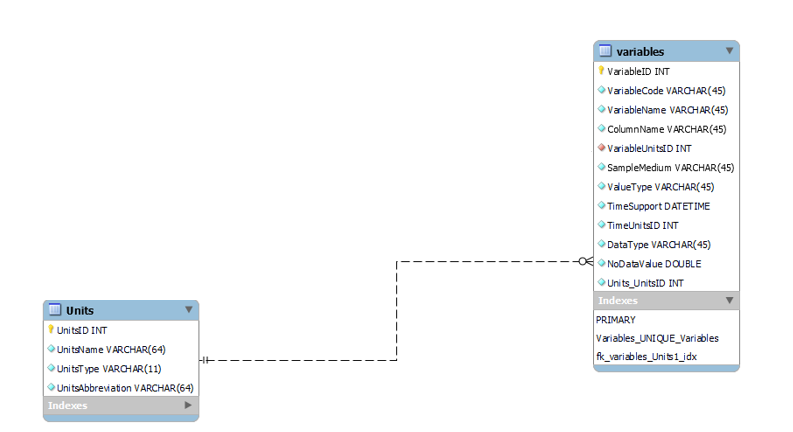
* Left to Right: A datavalue is measured by 1 and only 1 sensor
* Right to Left: A sensor can be used for recording 1 or more datavalues
  + 1. OffsetTypes and Datavalues



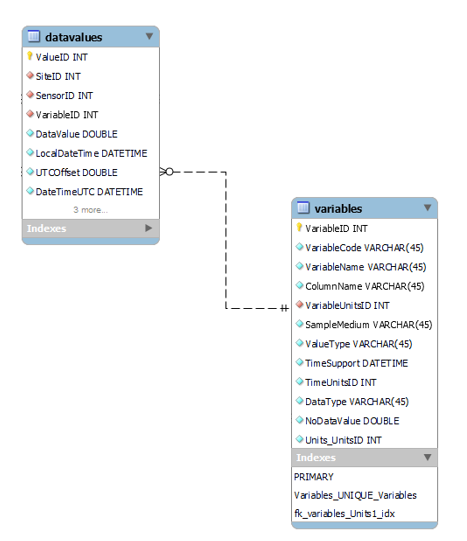
* Left to Right: An Offset Type has 0 or more datavalues of data
* Right to Left: A datavalues has 0 or 1 offset type
  + 1. OffsetTypes and Units



* Left to Right: An offset types can have 1 and only 1unit
* Right to Left: A unit has 0 or more offset types
  + 1. Units and Variables



* Left to Right: A unit can have 0 or more variable units
* Right to Left: A variable unit can have 1 and only 1 unit
  + 1. Datavalues and Variables



* Left to Right: A datavalues can have 1 and only 1 variable
* Right to Left: A variable can have 0 or more datavalues
  + The ER diagram is as mentioned below:

