Big Data HW2

Report on Logical Data Model Design

Explain how you will structure the metadata to avoid repetition

Metadata : For better database management and to avoid data redundancy, the tables are structured in a normalized form. It also makes efficient use of primary key and foreign key relationships.

Site - Contains information related to sites like Site name, latitude, longitude, state, county, comments and elevation with 'site_id' being the primary key.

Variable information- Contains information related to variables or parameters to be monitored like Variable name, Sample medium, Units, Time support, Value type, data type and no data value with variable_id being the primary key

Methods - Contains information regarding various methods that are used to measure variables like method name attributes which contains information like Batt_Volt_Min, Level_ft_Avg, Temp_degC_Avg and so on. Here, 'method_id' is the primary key and 'variable_id' is the foreign key.

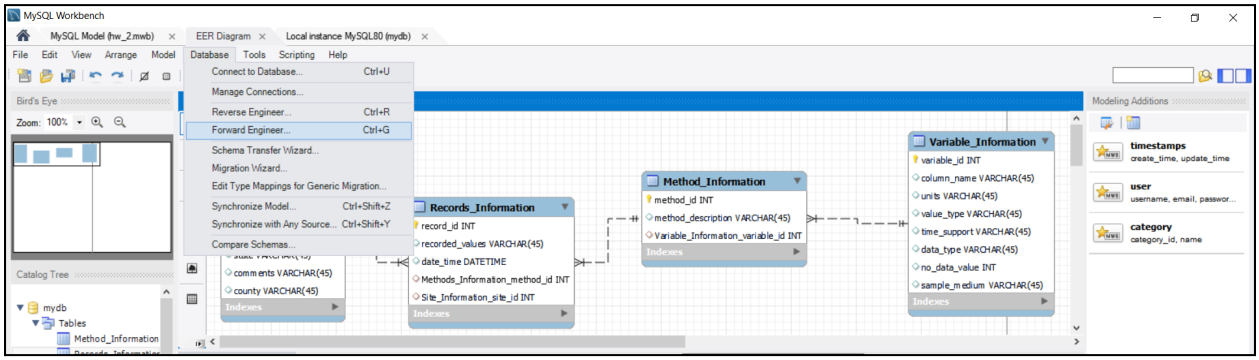
Recorded values – Recorded values contain all the readings measured for various sites, variables and methods along with the timestamp. It contains record_id as the primary key, and site_id, variable_id, and method_id as foriegn keys.

Overview the software technology, file formats, etc. you will use to organize the data and implement your data model.

Data can be categorized by using comma delimited text file (CSV) file format that could be previewed using Microsoft Excel. These csv files can be imported directly into MySQL workbench by creating a new database and respective tables under it. We are using the MySQL workbench tool, to make Entity-Relationship diagrams for the database design. Data can also be imported by using Amazon S3 bucket followed by Amazon Redshift clusters to draw the database tables.

Describe how you could make it easier to get data into and out of your data model

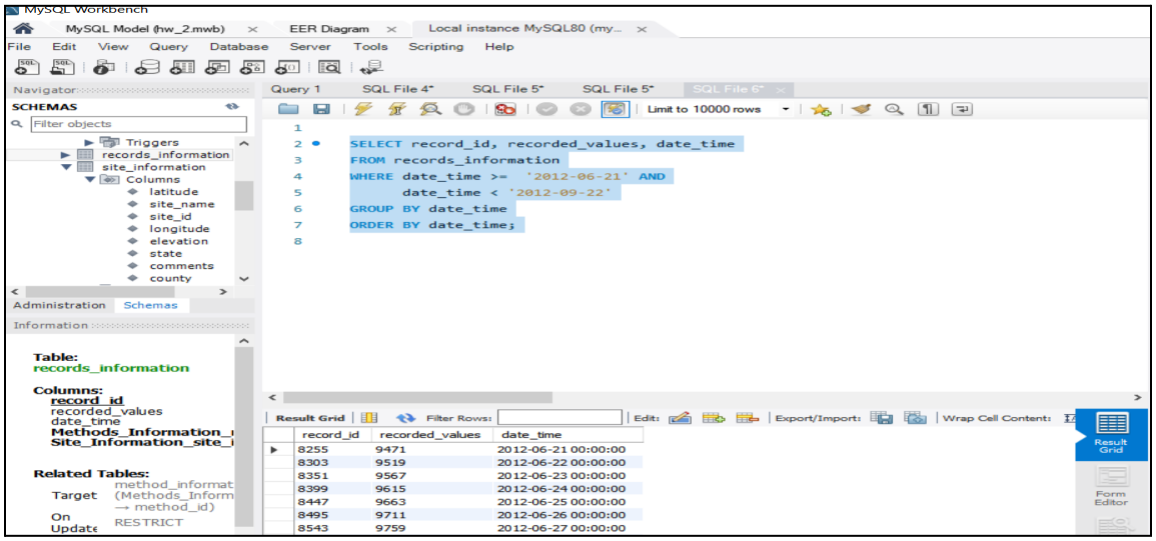
By using the Forward Engineer Function (we first created the ER diagram and then did the forward engineer process to enter the data into tables). In mysql workbench, we can easily import and export the data by connecting the ER diagram to the database. After importing the data into the table, we can query the data using DQL (data query languages).



Also, we can use COPY command in Amazon Redshift, data which was present in the CSV file can be directly copied in respective tables. After this the data can be inserted into the tables and by using SELECT queries, we can easily retrieve the data.

Specify whether your data model design will facilitate querying and retrieval of subsets of data.

We use the below query to retrieve the data from the records_information table.



Describe the entities and relationships that you have included in your data model.

Entities:

- Site Information
- Variable information
- Method Information
- Recorded Information

Relationships:

The relationship between entities is mentioned below .

- Each site has multiple recorded values.
- Each Method has multiple recorded values.
- Each Variable has multiple methods.

Provide an entity-relationship diagram that shows the entities needed to describe the data, their attributes, and the relationships between them.