Milestone 2

Team Name

Urban Analysts

Team Members

Angelina Paredes
Bryton LeValley
Taryn Burns
Melanie Bell

1. **Datasets and tools**

    a. Data model: what's the data model you will be using to represent the dataset? Why?

    We decided to use a relational data model due to the structured nature of our data set. A relational model appeared to be the best suited data model for the data set that we selected.

    b. Datasets statistics: how large is the dataset you will be dealing with? Report the following statistics for your datasets:

        i. If you are using relational data model: how many tuples? How many attributes? Any data constraints (FDs?) What's the physical storage size (in KB/MB/GB).

        The dataset consists of 2,804,813 tuples, and five attributes. The attributes consist of date, time, sensor, status, and activity. The physical storage size of the file is 100.5 MB.
        There is a two candidate keys for the dataset, date/time and sensor type. Each sensor activation time is recorded, the sensor type isn't unique but by adding date/time as a key we can generate a unique identifier for each tuple. date/time and sensor name functionally determines all attributes in the dataset. Another functional dependency is sensor name determines the type attribute.

    c. In data processing, you may need to develop a parser to transform the raw data into the format/tools you are using. Briefly describe the functions of the parser you implemented.

    The function of the parser we implemented is to extract the data from a text file and import it into a dataframe. An additional function of the parser is to separate the tuples by the location of the sensor. The parser pulls the data from the file line by line and removes the blank space used in the text file to separate data points. It then organizes the data points based upon which room the sensor is located in.

d. You should have successfully loaded the data into the tools you use. What's the estimated loading time, if you have the result?
    i. The data took 9.4 seconds to separate into individual csv files, the time it took to load into postgres was not recorded

2. **Data structure and auxiliary structure.**
    a. What data structure you developed of your own to represent the data, if any? For example, if you are using graphs, put up pseudo code that represent the data. Do you use any types of indexes? If so, briefly describe the indexes you developed for fast access the data.
        i. We utilized an index to see all the types of sensors used in the home such as motion, light, door etc.
        ii. Our data is structured in relations that contain the attributes Date/Time, Sensor Name, Message, and Type. Each relation represents a room of sensors. Every value in relation is a string.

3. **What's your plan for Milestone 3? In Milestone 3, you will be describing the detailed algorithm for your problem. Give a brief description about how you plan to design the algorithm. Provide any related work you plan to read if any.**

In Milestone 3, we will be planning out and explaining our algorithm and problem statement. Additionally, we will be starting out initial experimental plan. We plan to read related work surrounding the idea of health prediction and behavioral monitoring using smart home sensors. This reading will help us to understand the impacts this kind of analysis can have on people, as well as the algorithms that have been successful in making these predictions.