Javier Sandoval Bustamante

Luis Alejandro Arce Sáenz

ETL Project Final Report

# Introduction

The purpose of our project was to create a database containing information of restaurants across the following four regions in Mexico in the last six months: CDMX, Guadalajara, Monterrey and Cancún. The final database should contain information of each of the most relevant consumption categories in the restaurant industry (HVAC, illumination, ventilation, kitchen, machinery and general contacts) and provide a comprehensible data frame to allow for comparisons using weather variables and commercial activity of each site.

# Extract

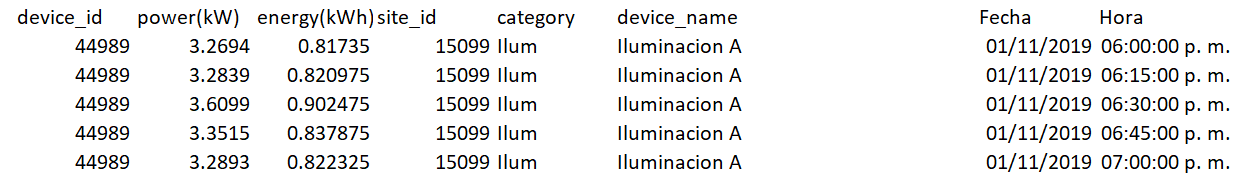
The original information was found in csv format in three different files and in JSON format from an API:

* sites.csv: This file contains sales and customers information of ten different sites, classified in the following columns:

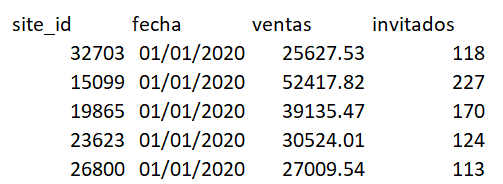


The information in this file contains additional information of each site, including their tariff category and division. The columns required for the rest of the project are the site\_id, site\_name and zone. For the purpose of this project, a sample of only four regions was required and the data had to be filtered to just the first four restaurants in the list.

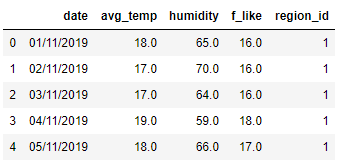
* Measurements.csv: This file contains power and energy consumption data of all the devices in each of the original ten sites. The data was collected for the last six months in intervals of 15 minutes. In addition, each of the devices is categorized into four categories: HVAC, Illumination (Ilum), Ventilation (Iny\_Ext), Machinery (Mach), Kitchen (Cocina) and electrical contacts in general (Contactos):



* ventas.csv: This file contains information of the sales and customers of each of the original ten sites. All the sales are given in Mexican pesos (MXN) and provided in daily intervals for the last six months. In addition, the data needs to be filtered to contain only the desired four sites.



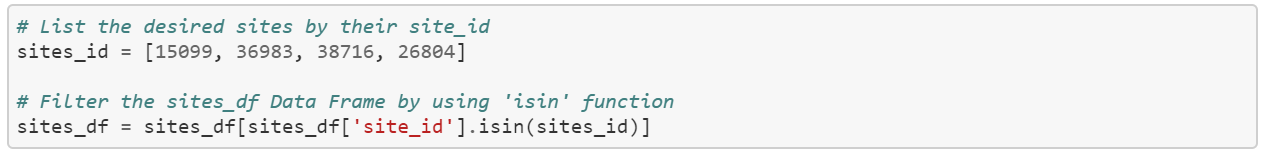
* Climate Data: The data was extracted from the Premium API of “World Weather Online”. The requested data was obtained based on the location of each restaurant (4). Once the data was obtained, all the columns were extracted from the JSON and converted into a data frame. Finally, the non-desired columns were dropped.



# Transform

The operations required to load these files into the dataset including various steps of filtering and grouping information from the raw data. In addition, some of the categories in the original tables were made in different tables to allow for maximum data normalization and primary key connections across all the data base. The most relevant steps in the process were the following:

1) Data filtering: A list of the required sites was created as a Python array using the site\_id column and all the information in the datasets including this field was filtered using our array.



2) Grouping: The data contained in the measurements.csv was grouped by four categories: zone, site\_id, date and category.



3) Column selection using Pandas: Having all the information loaded into Pandas DataFrames, it was easy to select just the required columns in each of the provided data sets. This allowed for more flexibility in the normalization process.

To enable primary keys across all the tables and to avoid using repeated fields, some of the data had to be created in new, separate tables. This was done for the regions and the consumption categories.

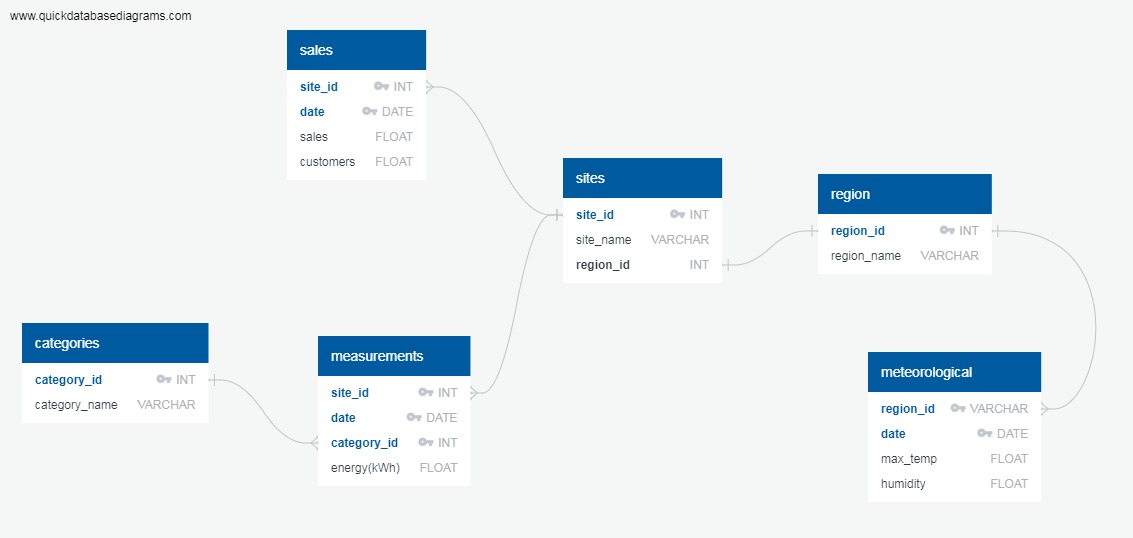




Additional steps in the transformation process included data type transformation to change string types into comprehensive date-time type.

# Load

The final database contains seven tables in total, related to each other according to the following diagram. These tables were loaded into PostgreSQL using PGAdmin.



The reason that the information was created in this format was to normalize the data as much as possible. Our final database specifies daily data across all the tables, allowing for daily categorizations of consumption categories (HVAC, illumination, kitchen, ventilation, machinery and general contacts) according to weather data and commercial activity (sales and customers).

# Conclusions

This project allowed us to have fun while learning how to transform and load data into comprehensible databases. The most challenging part was defining the database structure and formatting the original data sets to allow integration into the desired structure. The data normalization part was also challenging in that we had to extract separate tables from the original sets to