**ETL PROJECT - GROUP 2**

## **Project Title**

Houston Real Estate Investment Database

## **Team Members**

* Jonathan Antia Barrero
* Adriana Ovalle
* Ellen Rud

## **Project Description/ Outline**

To do the ETL (Extract, Transform and Load) of the data needed to run an application for Houston Real Estate Investment Dashboard.

## **Project Plan**

As the data used in this project is well structured and contains well defined, related entities, we chose to use a SQL object-oriented database management system.

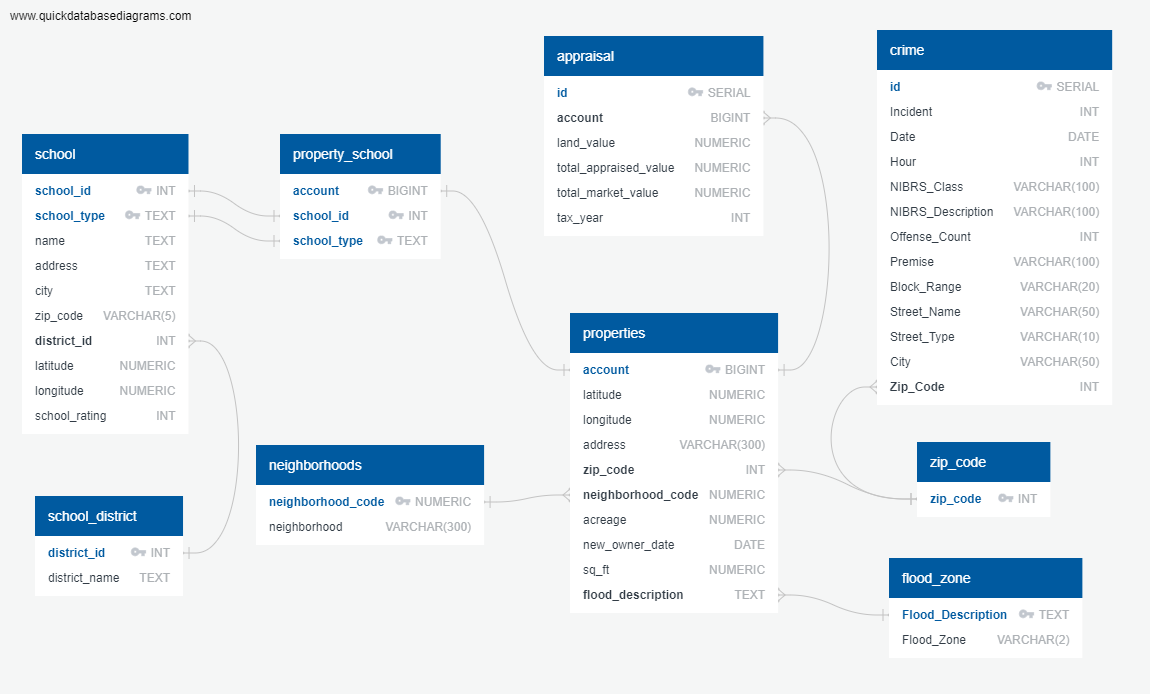
1. Data to be stored in a relational database – SQL Database.
2. Create an ERD Diagram
3. Extract HCAD Data, transform it, and load it in PostGres
4. Extract School Ratings Data, transform it, and load it in PostGres
5. Extract Flood Data, transform it, and load it in PostGres
6. Extract Crime Data, transform it, and load it in PostGres

## **Data Sources**

* School Ratings:
  + Source: <https://texasschoolguide.org/school-rankings/>
  + Download: CSV file
* National Flood Data:
  + Source: <http://nationalflooddata.com/flood/floodapi/>
  + Web request API using street addresses from HCAD dataset
* Harris County Appraisal District (HCAD):
  + Source: <https://pdata.hcad.org/download/2019>
  + Download: ZIP folders comprised of text files and csv files
* Crime:
  + Source: <https://www.houstontx.gov/police/cs/Monthly_Crime_Data_by_Street_and_Police_Beat.htm>
  + Download: CSV file

**Files location:** <https://drive.google.com/drive/folders/15sKnJGCKU_VsG290i97kzsAnyc1tFNNT>

## **Entity-Relation Diagram**



## **Data Transformation**

### **School Data:**

The goal of the transformation of schools data is to 1) create an entity table for all schools, 2) create an entity table for all school districts, and 3) use the school IDs and school type to classify the residential property data per school zone. In the interest of time, the classification was based on the scikit learn nearest neighbor algorithm using the Euclidean distance between each residence and all elementary, middle, and high schools.

#### Transformation Steps:

* Slice the dataframe to show only schools in Harris County
* Drop any rows that do not contain a grade for a school
* Remove any rows where school type is listed as unavailable
* Locate rows missing latitude and longitude coordinates and retrieve the coordinates for those rows using google’s geocode api
* Limit the data to only the columns needed
* Drop rows associated with charter schools, as we only need public schools
* Rename columns as needed
* Convert TEA grades into a numeric school ratings column
* Add unique ids to each district
* **Create a separate table for all districts and a separate table for schools**. This will achieve database normalization.

### **HCAD Data:**

The data obtained form HCAD includes multiple datasets: real\_acct\_2018, real\_acct\_2019, building\_res, neighborhood codes, where the real\_acct files contain the appraisal values for the years 2018 and 2019.

#### Transformation Steps Real\_Acct:

* Reduce the real\_acct\_{year} files to only relevant fields and merge
* Filter the state class to only single-family homes (class = “A1”)
* Filter the dataframe to only the zip codes being used
* Rename and reformat column names as need

#### Transformation Steps Building\_Res:

* Reduce the dataset to only show relevant fields
* Rename the fields to match those in the real\_acct dataset
* Group the table by account number and sum the square footage for each property, as some properties contain multiple buildings under the same account (ex. Garage apartment, main house, etc).
* Merge the building\_res dataset with the real\_acct dataset

#### Transformation Neighborhoods:

* Cull neighborhood columns to show only relevant fields

#### Final Transformation:

* **Create separate entities for 1) Properties, 2) Neighborhoods, and 3) Appraisal**

### **Flood Zone, Schools, and Properties:**

The flood zone dataset contains latitude, longitude, and street addresses, which represent the connecting points between the school data (with lat/long) and the HCAD properties table (with street addresses). The end goal is to have each property in the HCAD dataset zoned to the nearest elementary, middle, and high schools. To achieve this goal we created a transfer table between the school table and the properties table using the following transformations:

#### Transformation Steps:

* First, all the files from the flood api responses were appended together into one flood zone dataset
* The flood zone dataset was classified with school zones using scikit learn’s nearest neighbor algorithm.
* Rename columns as needed
* The school classification was then transferred to the HCAD dataset by merging with the classified flood zone dataset.
* From this merged dataset we extracted a table called property\_school, which serves as a **transfer table between schools and properties**
* The flood zone dataset was reduced into a **singular entity consisting of flood zone and description**

### **Crime:**

#### Transformation Steps:

* Removing any rows where zip code is shown as NaN
* Reformatting the zip code to a format consistent with other datasets
* **This data forms a separate entity called crimes**
* To connect this table with the properties table we created a **transfer table that contains all the zip codes used in the study**

## **Data Loading Schema**

Sequence of table creation:

1. Zip\_code
2. Flood\_zone
3. School\_district
4. Neighborhoods
5. School
6. Crime
7. Properties
8. Property\_school
9. Appraisal