## DATABASE INTERACTIVE PLATFORM DESIGN

# Zillow Insights Sharepoint

**Team: Claire Jiaying Wu, Jack Lok, Zimo Zhong, Chenlu Wang**

# Client Scenario

- An online real estate company
- Strong power to acquire data
- Need to manage massive data
- All the data can help Zillow to better understand the market and serve the customers
- We are hired by Zillow to design the database

**Project Objectives**

- Improve Zillow's decision-making efficiency by combing the relationship between variables, between tables, and by developing the interactive dashboard to deliver easy and quick analytical insights.
- The project results are valuable to anyone who is interested in buying or investing houses and housing agencies that want to know the market.

# A Brief View of the Original Data

## First 9 columns:

| | Unnamed: 0 | parcelid | airconditioningtypeid | architecturalstyletypeid | basementsqft | bathroomcnt | bedroomcnt | buildingclasstypeid | buildingqualitytypeid | calcul |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 541116 | 11554200 | NaN | NaN | NaN | 3.0 | 3.0 | NaN | 6.0 | |
| 1 | 1704087 | 11414286 | NaN | NaN | NaN | 3.0 | 4.0 | NaN | 3.0 | |
| 2 | 644001 | 11067227 | 1.0 | NaN | NaN | 4.0 | 4.0 | NaN | 7.0 | |
| 3 | 1338306 | 14000816 | NaN | NaN | NaN | 2.0 | 3.0 | NaN | NaN | |
| 4 | 943769 | 11580959 | NaN | NaN | NaN | 4.0 | 5.0 | NaN | 9.0 | |

## Full list of the 58 columns:

parcelid
airconditioningtypeid
architecturalstyletypeid
basementsqft
bathroomcnt
bedroomcnt
buildingclasstypeid
buildingqualitytypeid
calculatedbathnbr
decktypeid
finishedfloor1squarefeet
calculatedfinishedsquarefeet
finishedsquarefeet12
finishedsquarefeet13
finishedsquarefeet15
finishedsquarefeet50
finishedsquarefeet6
fips
fireplacecnt
fullbathcnt
garagecarcnt
garagetotalsqft
hashottuborspa
heatingorsystemtypeid
latitude
longitude
lotsizesquarefeet
poolcnt
poolsizesum
pooltypeid10
pooltypeid2
pooltypeid7
propertycountylandusecode
propertylandusetypeid
propertyzoningdesc
rawcensustractandblock
regionidcity
regionidcounty
regionidneighborhood
regionidzip
roomcnt
storytypeid
threequarterbathnbr
typeconstructiontypeid
unitcnt
yardbuildingsqft17
yardbuildingsqft26
yearbuilt
numberofstories
fireplaceflag
structuretaxvaluedollarcnt
taxvaluedollarcnt
assessmentyear
landtaxvaluedollarcnt
taxamount
taxdelinquencyflag
taxdelinquencyyear
censustractandblock

# Database Normalization Process

## Data Preprocessing

**Removed** the 21 columns with too many NAs (with over 90% missing values).

**Sampled** 1,000 of the rows for the following implementation.

## From 1NF to 3NF

**1NF: atomic & no repeating attributes - checked**

**2NF: all about the key - separated from 1 home_info table to 11 tables in total**

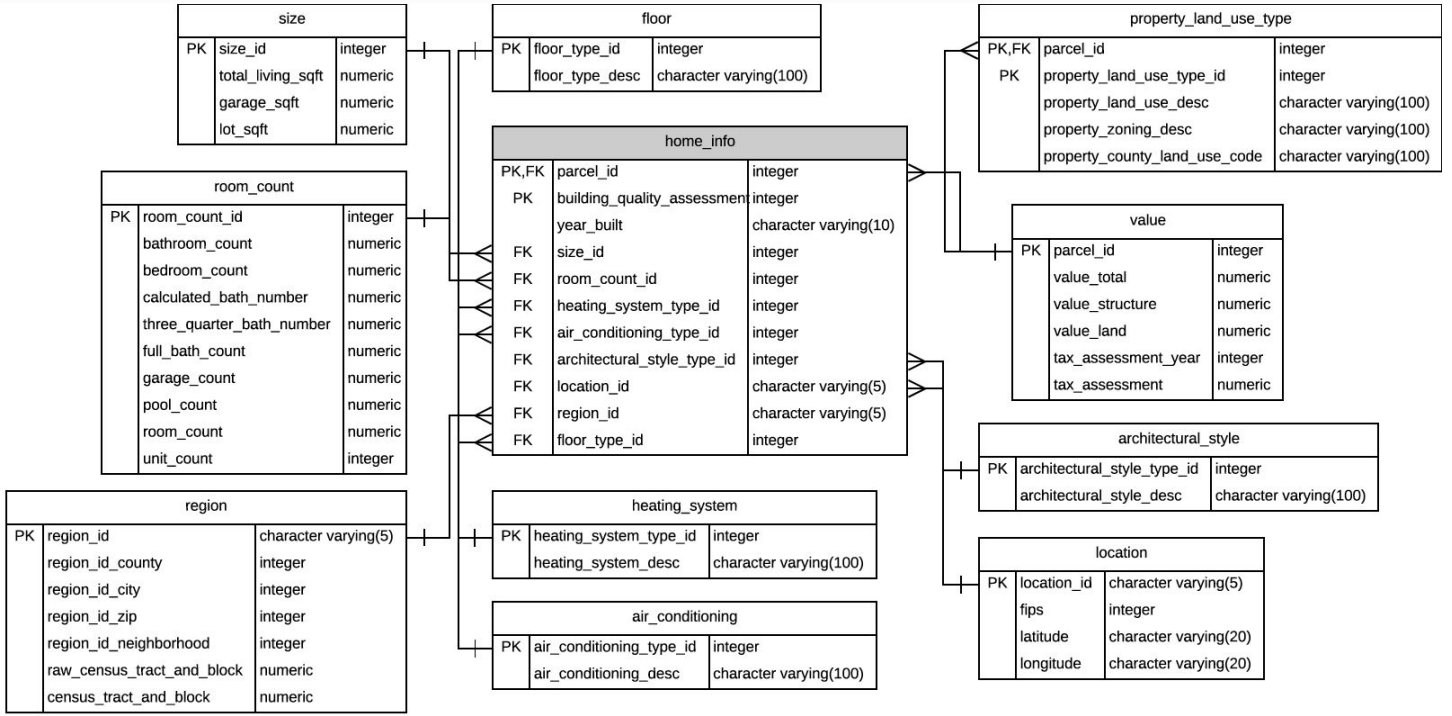
Home_info, Size, Room_count, Floor, Heating_system, Air_conditioning, Architectural_style, Location, Region, Property_land_use, Value

**3NF: only about the key - checked**

## Data Types Setting & Troubleshooting

**Verified** the data types of each attribute, orders and constraints by creating test in pgAdmin4

# Entity Relation Diagram



# ETL Processes

## Extracting

Load the properties_2017_small data to dataframe

Extract the related columns from the dataset

## Transforming

Rename the column names and check the table by head()

Add incrementing integers for id

## Loading

Write into the value table in the database

# Interacting with Zillow Data

Our client consists of two types:

- **“C” level officers**
  - non-technical background
  - visualizations: charts and graphs
  - understand findings and performances

- **Analysts**
  - technical background
  - interact with the initial queries
  - understand the details behind the tables and visualizations

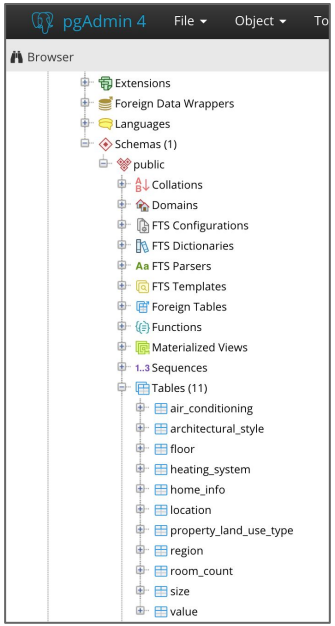
Metabase meets the needs of both groups of audience.

# Analytical Procedures

1. What's the most frequent combination of air conditioning and heating system?
2. Which year-built has the highest average home total value?
3. Among the homes from these five years, is the ratio of structure value and land value always consistent?
4. What are the top five regions (city and zip) with highest average home values per square feet in the last ten years?
5. In these cities, what are the most common architectural styles?
6. What are the average values per square feet for the properties with different architectural styles?
7. As people are most concerned about how many bedrooms, what possible choices can they have for the number of total rooms and bathrooms?
8. How are locations distributed based on FIPS?
9. What are values for the top 5 largest living square feet?
10. What is the highest number of the room count for each year_built? And does the number increases with years passing by?

# Interacting with Zillow Data


- Design schema of the database



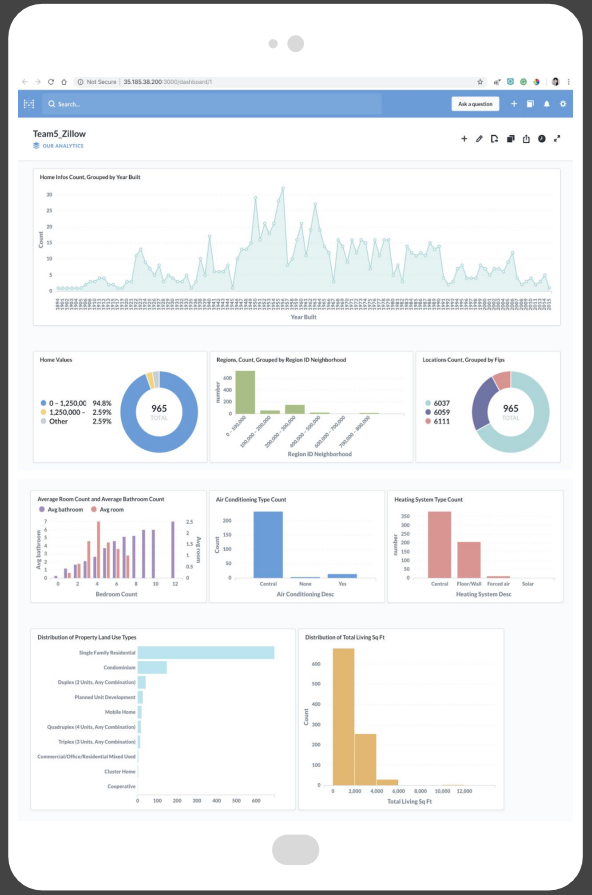

Lucidchart

- Draw the ER diagram of the database design

python Programming

- Load data into the database
- Query the database

- Create the dashboard for the presentation




# Database Interaction Demo

- Home info timeline
- Home value
- Region & Location & room count
- Air conditioning type
- Heating system type
- Property land user type
- Total living square feet

http://35.185.38.200:3000/dashboard/1

# Conclusion

- By using RDMS and ETL, we aggregated and organized raw data into an accessible format to provide useful context.
- The relational table structure in RDMS makes it possible to run queries across multiple tables at once.
- ETL streamlines the extracting, transforming, and loading of our data between applications for analysis.
- Achievement: Our client can use the insights to make decisions and drive change for business strategies.