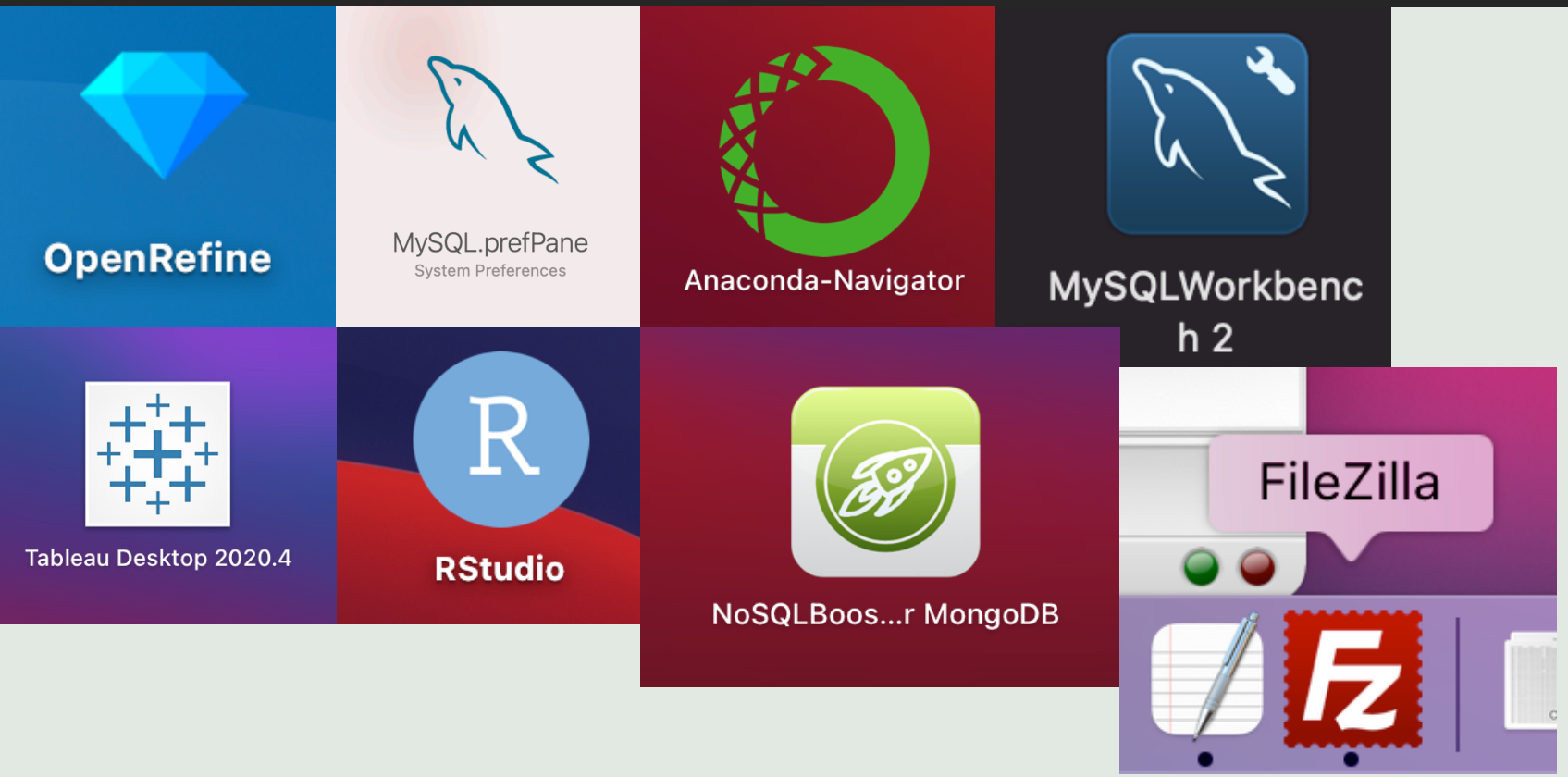
# *Data Engineering Platform*

Assignment 1

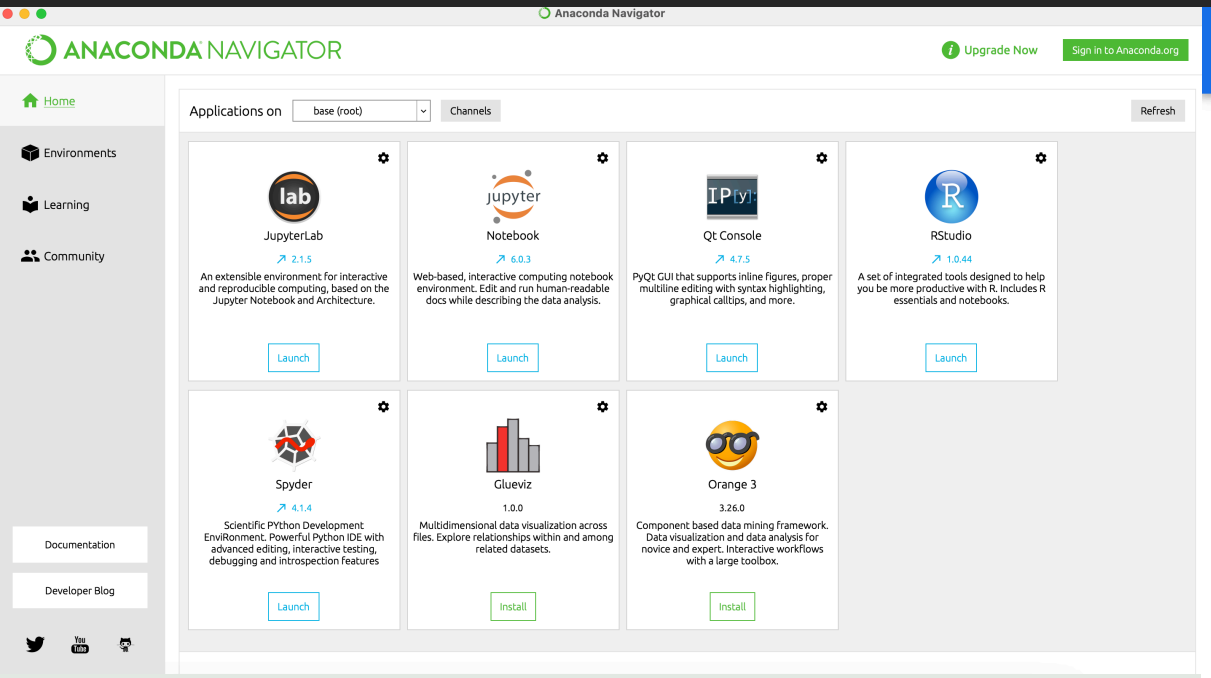
Fiona Fei

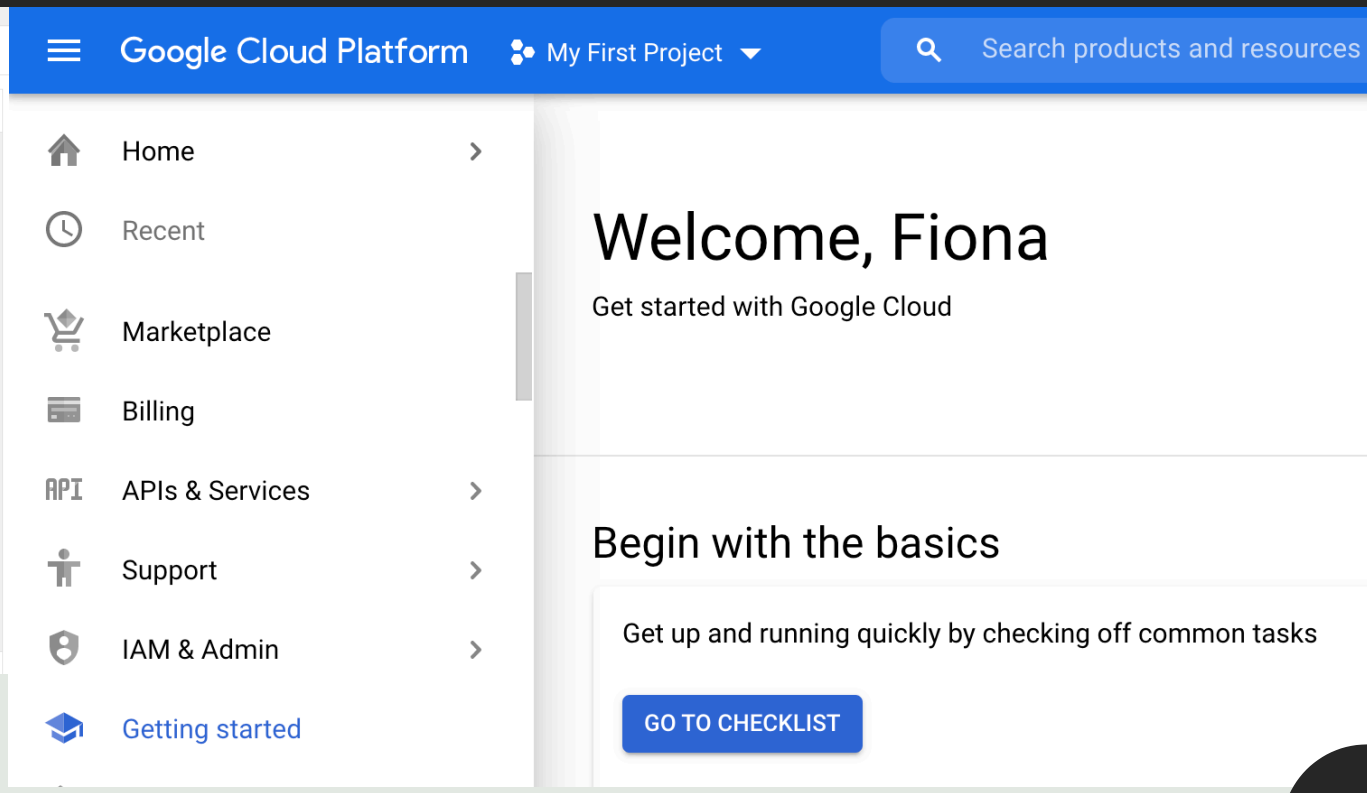
# *Part A: 1. Install Software*



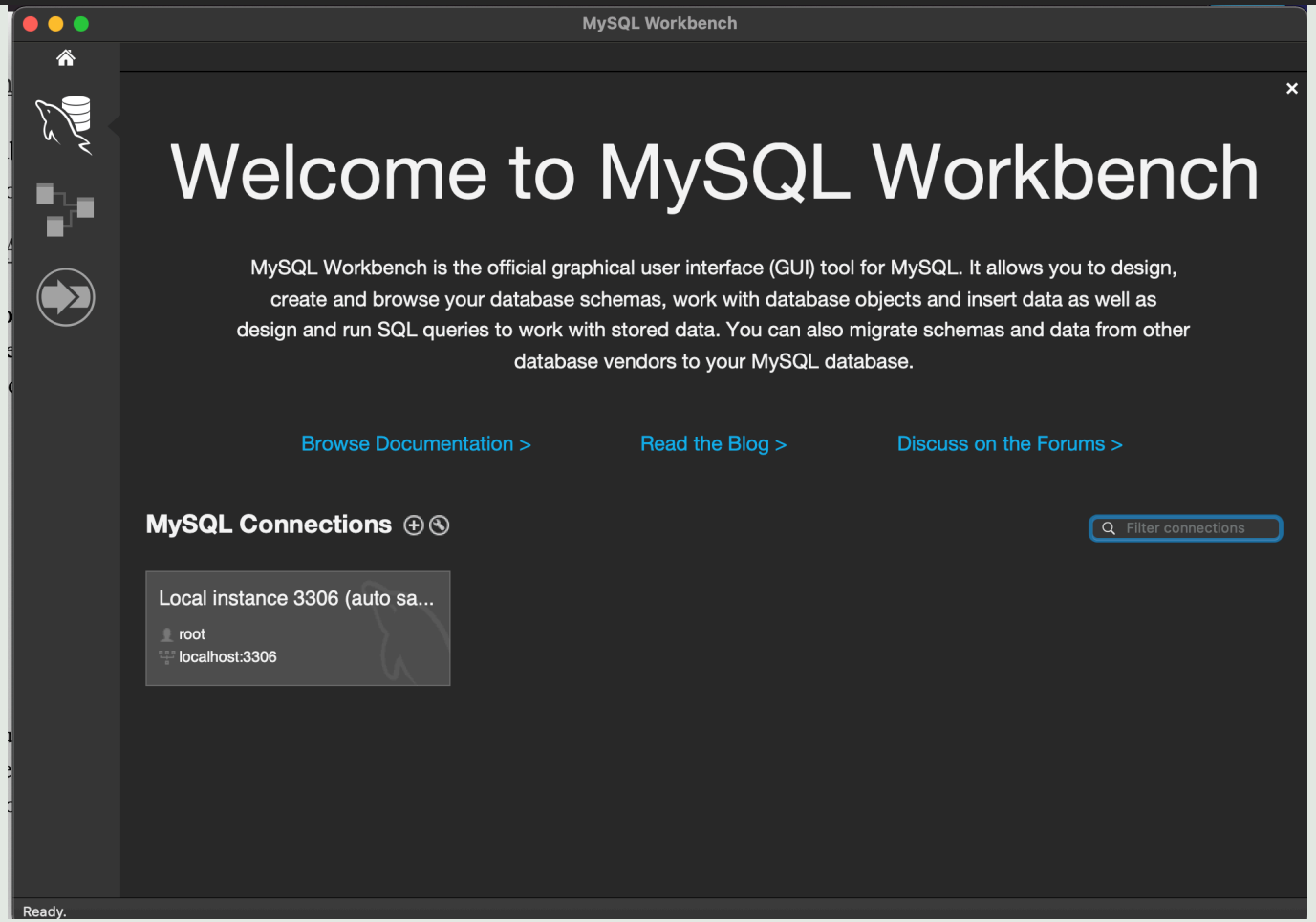

Note：
The applications are in different places on my laptop: the desktop, dock tool bar, application files, etc. I couldn't take a screenshot that contains all of them, so I took screenshots of them separately, but they are all in my laptop.

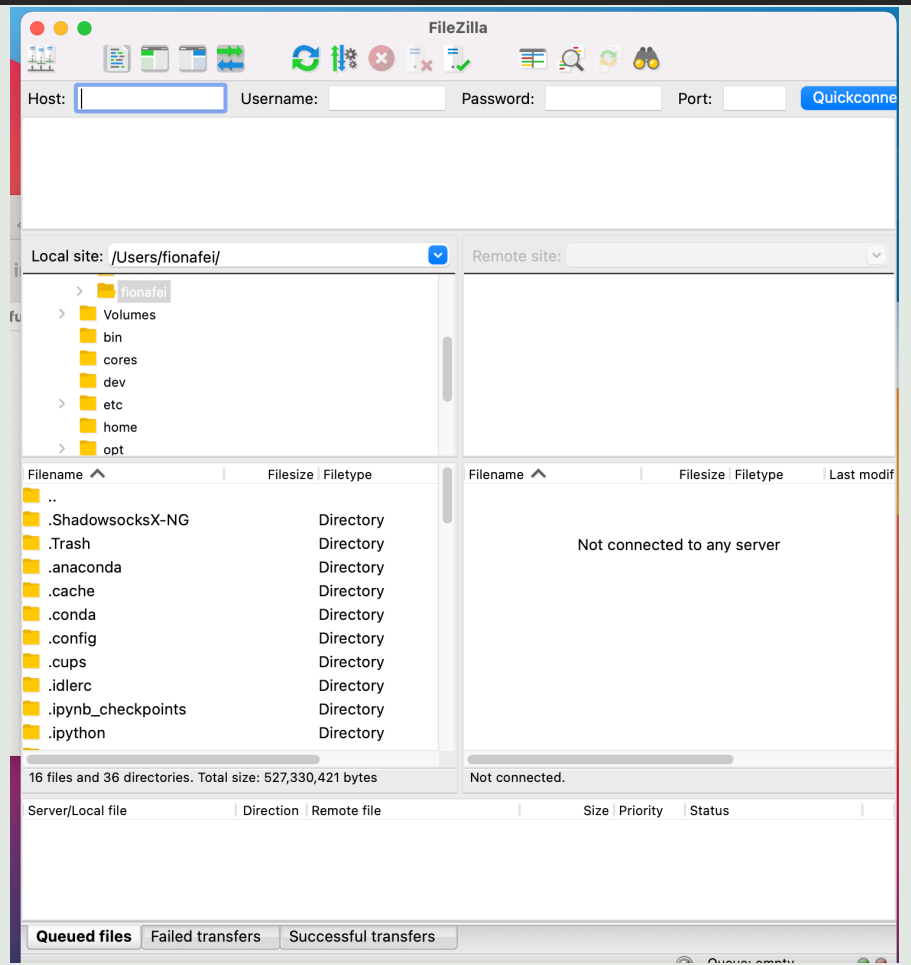
# Part A: 1. Install Software



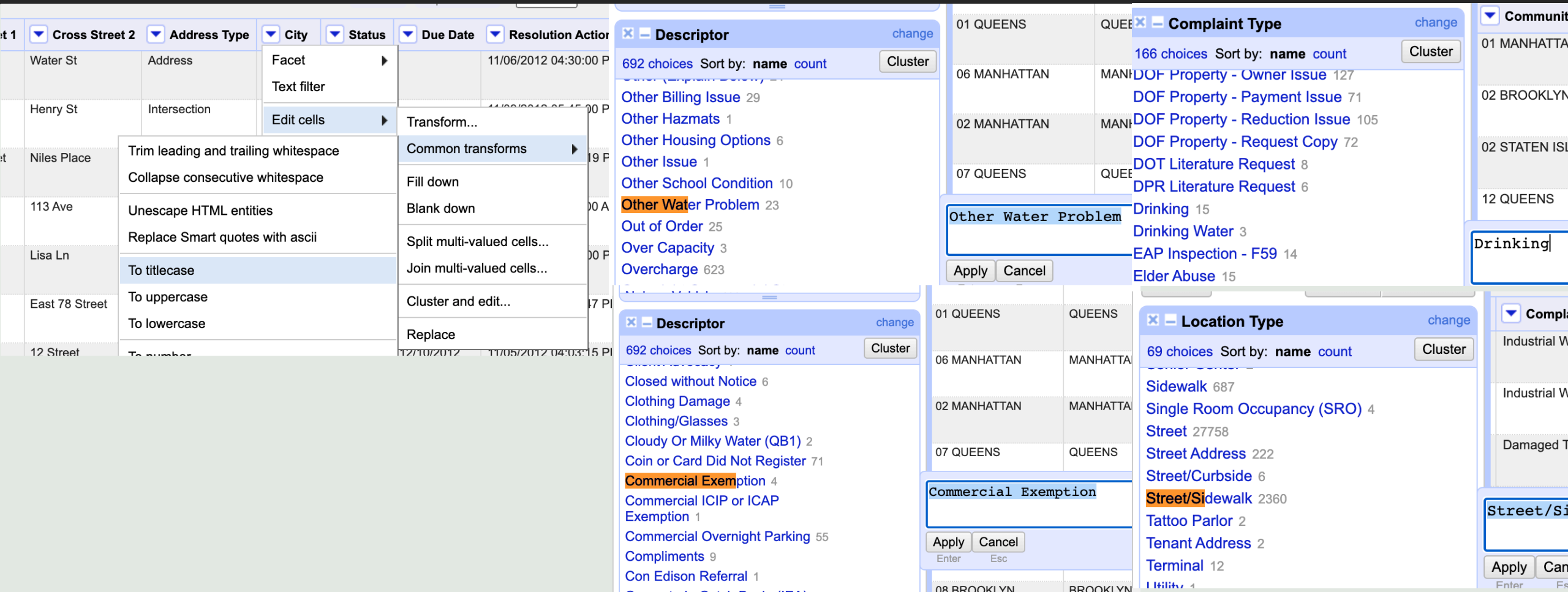



# Part A: 1. Install Software

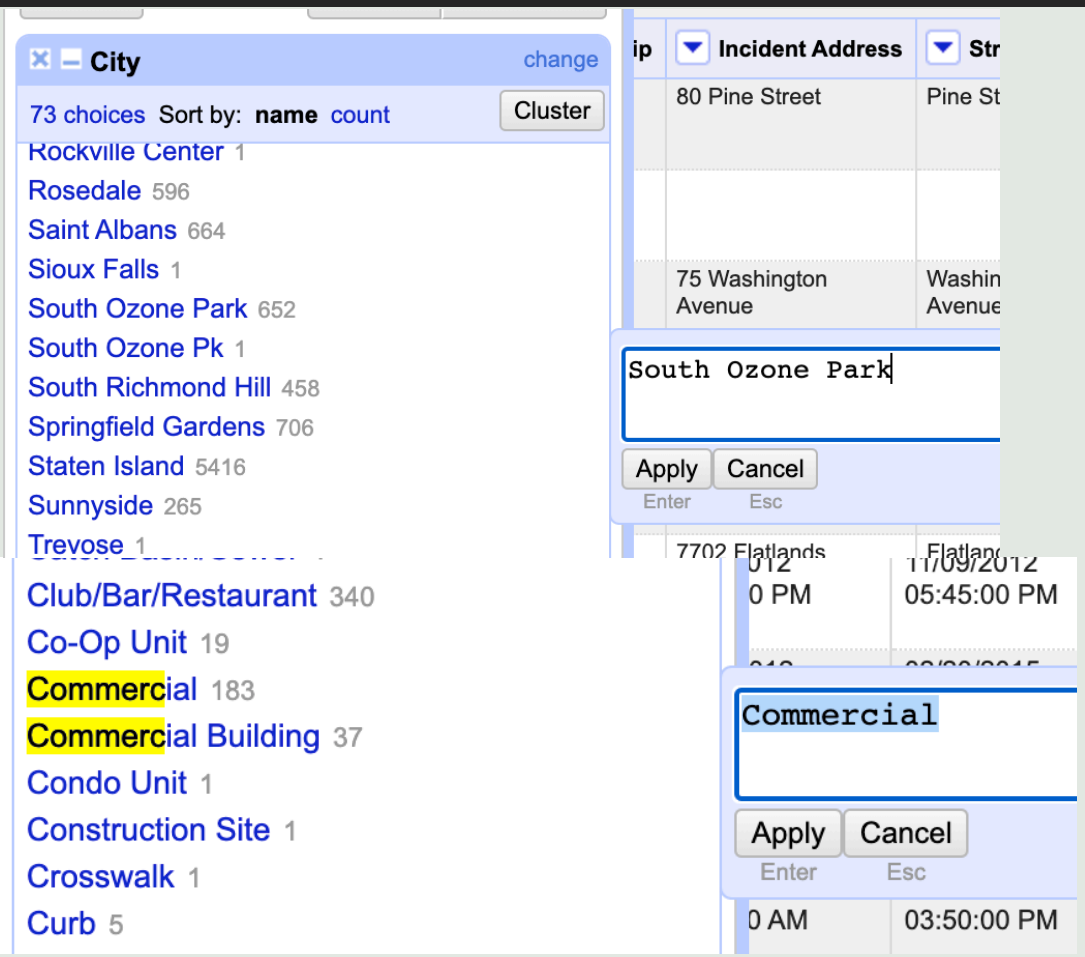


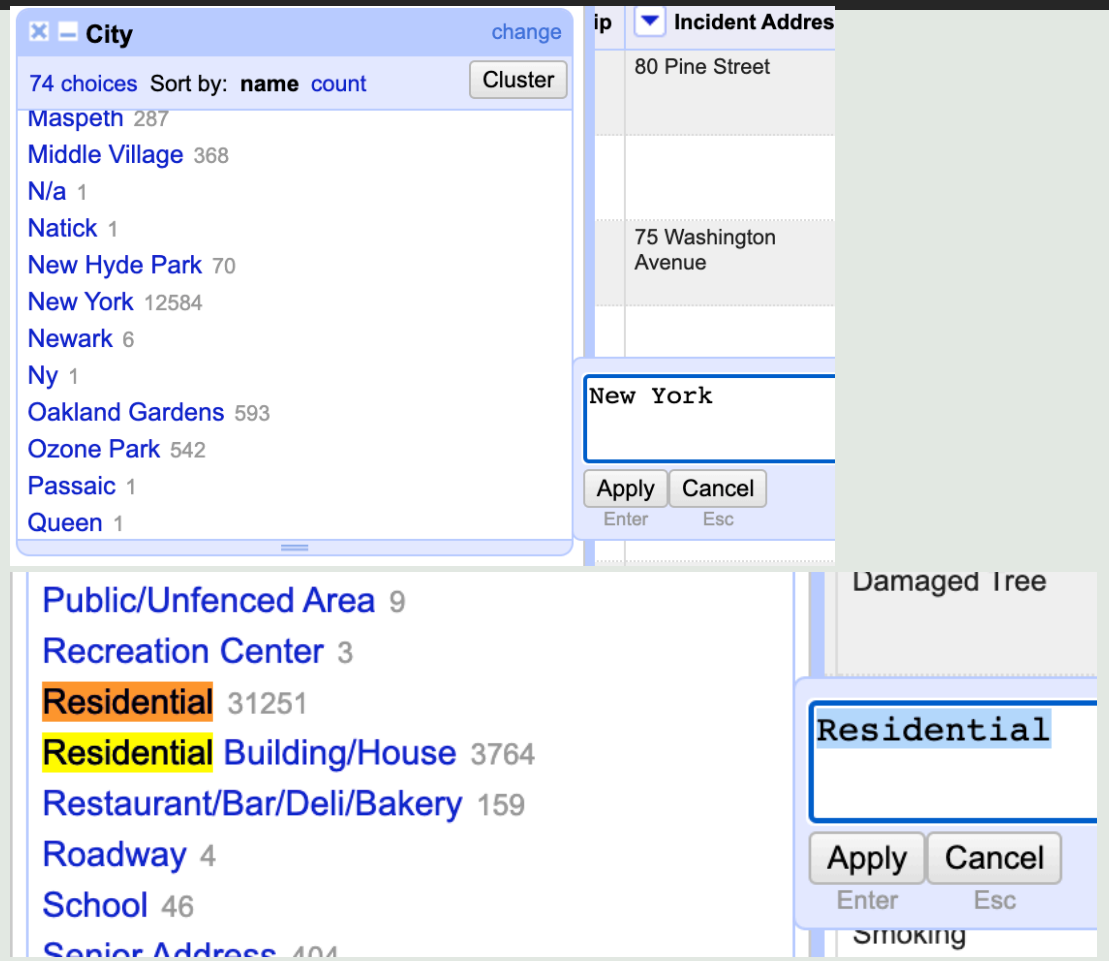



# Part A: 2. Data Preparation Step d-g

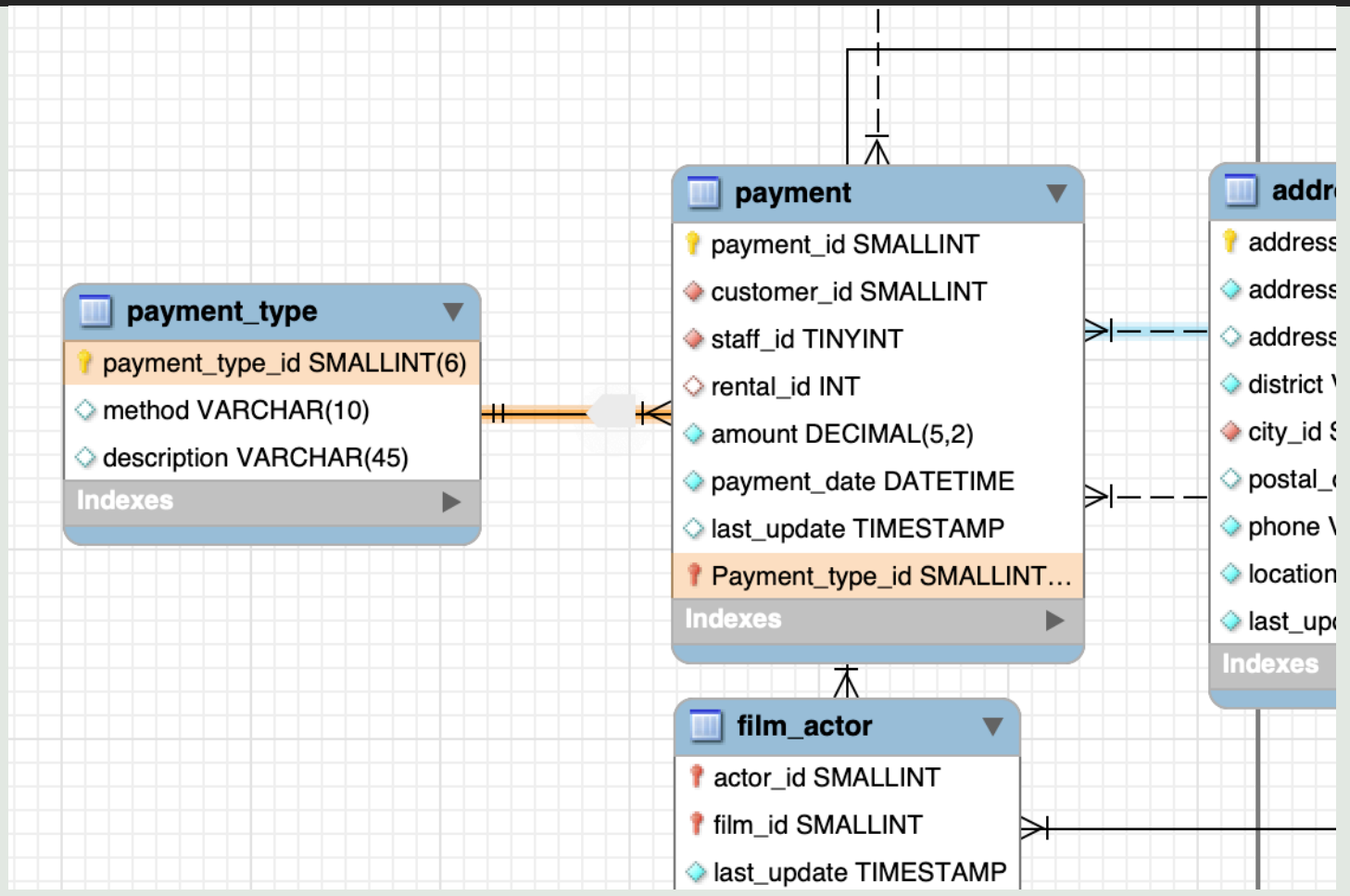


# Part A: 2. Data Preparation Step d-g



# Part B: 1.d


payment
payment_id SMALLINT
customer_id SMALLINT
staff_id TINYINT
rental_id INT
amount DECIMAL(5,2)
payment_date DATETIME
last_update TIMESTAMP
Payment_type_id SMALLINT…
Indexes
payment_type
payment_type_id SMALLINT(6)
method VARCHAR(10)
description VARCHAR(45)
Indexes
film_actor
actor_id SMALLINT
film_id SMALLINT
last_update TIMESTAMP

# *Part B: 1.e Relational Data Modeling*

| Field (Attributes) | Primary Key (Y/N) | Foreign Key (Y/N) | Related Tables(s) |
|---|---|---|---|
| Rental_id | | Y | Rental (M:N) |
| Paymentd_id | Y | Y | |
| Customer_id | N | Y | Customer (1:N) |
| Staff_id | N | Y | Staff (1:N) |

# *Part B: 2.a Examples of anomalies*

- Insertion anomalies:

  If the surgery is defined and null values are not accepted, a new patient without assigned surgery won't be able to insert into the database.

- Deletion anomalies:

  If Jay Patel quit the physician job and deleted him from the database, the Liver Transplant surgery will also be ceased to exist because it's only assigned to Jay Patel.

- Modification anomalies:

  If Robert Smith's address is incorrect and needs to be updated, it needs to be updated for two times or there will be inconsistent data in the database.

# Part B: 2.b Normalize to 3NF

Physician

| Physician Name | Physician Office |
|---|---|
| Helen Pearson | Chicago Ave, Chicago |
| Olga Kay | Clark Street, Chicago |
| Robert Smith | Madison Street, Chicago |
| Wei Jing | Adams Street, Chicago |
| Jay Patel | Monroe Street, Chicago |

Patient:

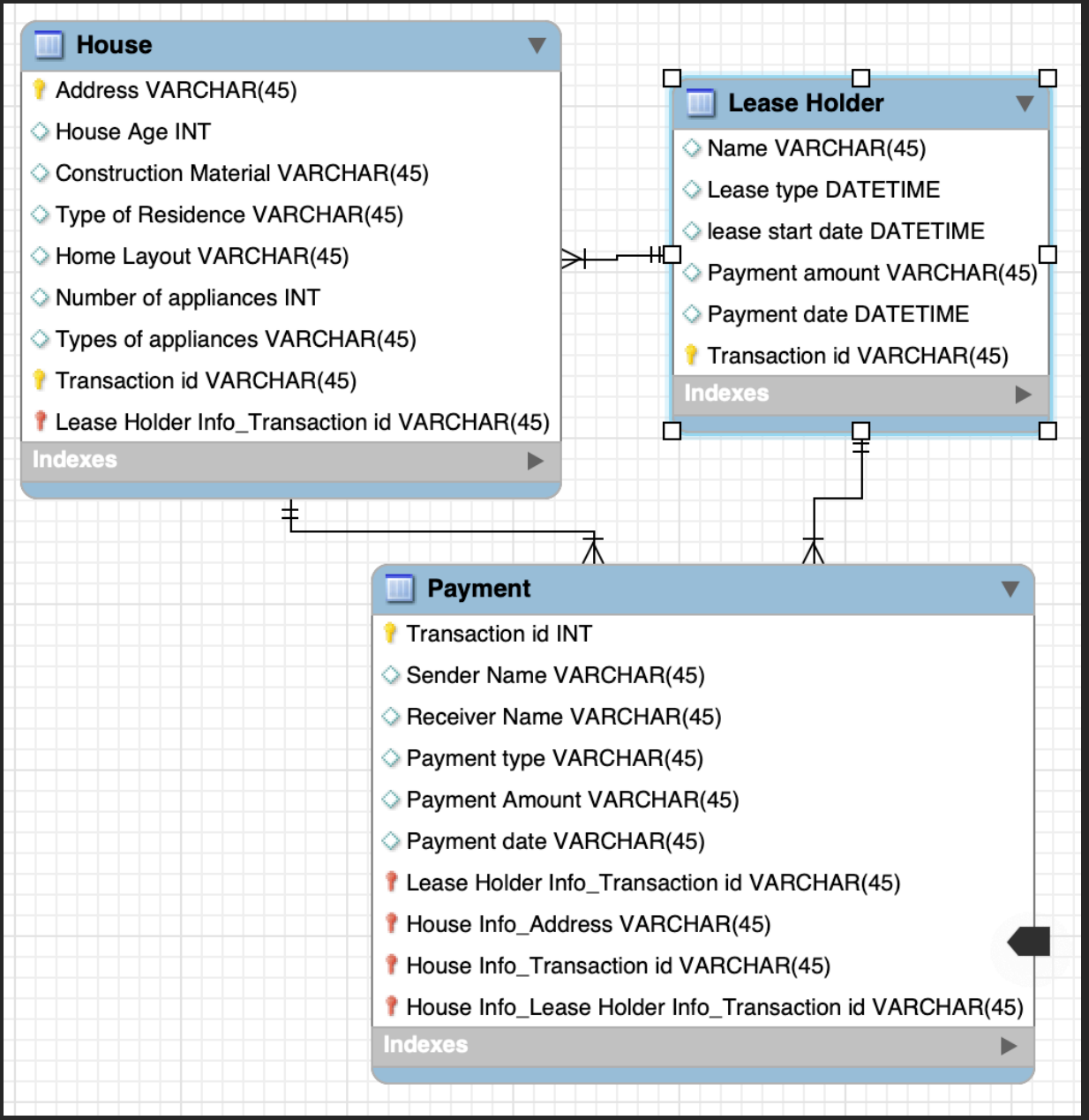
| Patient Name | Patient Address |
|---|---|
| Joe Korn | Randolph Street, Chicago |
| Gillian White | Illinois Street, Chicago |
| Jill Bell | Huron Street, Chicago |
| Mike Li | Lake Street, Chicago |
| Ian MacKay | Dearborn Street, Chicago |
| Sheela Nupur | Monroe Street, Chicago |

Appointment

| Appointment Dat | Patient Name | Physician Name | Surgery Type |
|---|---|---|---|
| 3/7/17 | Joe Korn | Helen Pearson | Tendon Repair |
| 3/22/17 | Gillian White | Helen Pearson | Skin Graft |
| 6/13/16 | Joe Korn | Olga kay | Sentinel Node Biopsy |
| 6/13/17 | Jill Bell | Robert Smith | Tendon Repair |
| 6/14/17 | Jill Bell | Robert Smith | Skin Graft |
| 6/13/17 | Mike Li | Wei Jing | Knee Arthroscopy |
| 8/15/17 | Gillian White | Jay Patel | Sentinel Node Biopsy |
| 1/4/16 | Ian MacKay | Jay Patel | Hepatic Resection |
| 1/5/18 | Ian MacKay | Jay Patel | Liver Transplant |
| 1/4/16 | Sheela Nupur | Helen Pearson | Knee Arthroscopy |
| 2/12/16 | Joe Korn | Wei Jing | Skin Graft |
| 4/15/18 | Mike Li | Wei Jing | Skin Graft |

Assumptions:

- Assume each physician has an assigned unique office
- Assume each patient have an assigned address and appointment.
- Assume each appointment has complete information about the date, patient name, physician name, and surgery type.




# *Part B: 3 Data Modeling*

- Tables(3):

  *House, Lease Holder, Payment*

- Relationships:

  *House is connected to the lease holder by sharing the same transaction id, lease holder is connected to payment by sharing the same transaction id, payment date, payment amount. House is connected to the payment sharing the same transaction id.*

# *Part B: 3. Data Modeling*

- Cardinality

    *House (N:M) Lease Holder*

    *House (1:N) Payment*

    *Lease Holder (1:N) Payment*

- Datatypes:

VARCHAR(45): Address, Construction Material, Type of Residence, Home Layout, Types of appliances, transaction id, Name, lease type, payment amount, sender name, receiver name

INT: House Age, Number of Alliances

Date: Lease start date, payment date

# *Part B: 3. Data Modeling*

Summary of Design Considerations:

This database is designed as a relational database. The database could be very large depends on the number of houses that are available and the number of customers that have the intention to purchase the houses.

The number of users and the number of houses will determine whether centralized will be sufficient or distributed, and therefore determine the critical entities and store them centrally.

# *Part B: 3. Data Modeling*

Summary of Design Considerations:

There are some fields that needs security storage such as the lease holder's name and payment amount. They should belong to high privacy components that are not openly to the public.

As for data integrity, it is necessary to restrict the users that are allowed to get access to this database because it contains the personal information of lease holders and their payments. The information should be only be accessed by limited number of users. And the 3NF format will help with keeping data private and reduce the data anomalies scenarios.