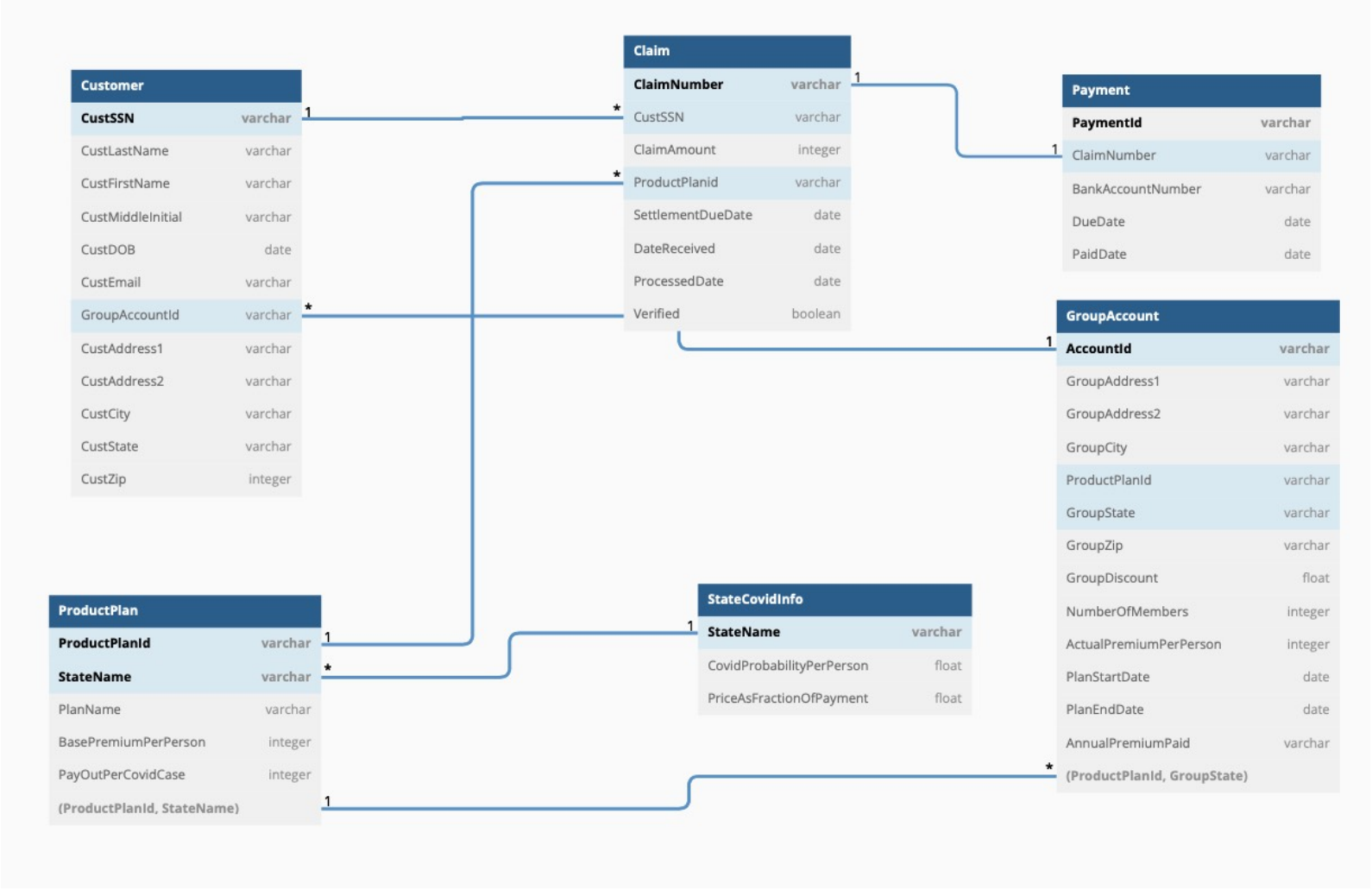
Wayne Zhu project part 2 notes
Notes on next page



It's harder than expected to find good data. Originally, I planned on creating a health insurance company, and I found unstructured data on different kind of chronic illness. Unfortunately, that design will leave no room for Machine Learning—since machine learning would rely on individual panel data. Also I will have to manually implement all the 100 or so diseases.

Under Professor Franchitti's guidance, I switched to using JHU's covid data. Link here: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

I plan on using this data to generate the state-wide annual increase in cases and run a regression model with each state as a dummy variable. My insurance company's nature also changes from a health insurance that covers healthcare to an injury insurance that pays a lump sum per each individual infected. This is so fascinating because this type of insurance companies are losing huge amount of money right now due to an unexpected increase in cases in Taiwan: https://www.wsj.com/articles/insurers-in-taiwan-pay-the-price-for-misjudging-covid-risks-11655546402

Therefore, I also make some changes in my schema. I added a StateCovidInfo table to store insights I will learn from the uncleaned data. I also removed a lot of fields from Claim table because they are no longer necessary. I spent a lot of time pondering whether state in each table should be a FK, PK or anything else, how to control redundancy and etc.

In order to leverage hybrid data, the company need to follow the following reference architecture. Of course I don't have time to describe the entire reference architecture. But basically, the analytics of uncleaned data should be as automated as possible. Therefore, the company would be able to frequently conduct analytics of uncleaned data, extract insights, and update their existing relational database accordingly.

The company also must be constantly looking for new disrupting forces like new variant, new vaccine, hospitalization rate, crowd immunity in the future. Being curious and agile is the most important aspect when it comes to this dynamic insurance category. This type of insurance is not about health, it's gambling basically.

Most tables have very few attributes as primary key. So it's in 2NF already.

We fix the payout of all plan to be 200, 500, and 1000, and premium differ for each state and each payout accordingly. As for 3NF, in ProudctPlan table, BasePremiumPerPerson is dependent on (StateName and PayOutPerCovidCase), which is a candidate key. So it's fine.

In State Covid info, PriceAsFractionOfPayment is dependent on CovidProbabilityPerPerson, but CovidProbabilityPerPerson is a candidate key also.

So it's already in 3NF.

As mentioned above, the insight will be stored in StateCovidInfo.