David Sacco

Fairfield University

Capstone Project

Business Analytics Capstone Project Proposal

22 May 2019

**Overview**

I would like to create a database warehouse that would house public mortgage data from Fannie Mae and Freddie Mac. If this is successful, I will then create an R Shiny application that will sit on top of the warehouse, that will then be given to my professional advisor, Robert Sacco, who will then determine if there is a use for it in his line of work. Robert Sacco is an Executive Director at JP Morgan Chase for the Mortgage Banking division. He will be my main contact for any questions related to the dataset that I will be using.

The dataset that I intend to use has 56,000 rows and almost 80 columns. Professionals, such as Robert, must sift through excel sheets that have enormous amounts of data, and this project may help with speeding up this manual spreadsheet analysis with a few clicks on an R Shiny application.

**Project Specifics**

The dataset that I will be using for the mortgage data will be cut down to 16 attributes, as the team that uses this data only uses 16 of the 80 attributes. The years that this dataset will span from is 2014 to 2017, so show the relationship of the attributes over time, as that is a crucial aspect to the team’s visuals.

As for the database warehouse, I am going to be hosting it on the cloud, using AWS RDS feature. I will be using a t2.micro database, that will use PostgreSQL as the object-relational database system. This database and database instance size are both free up to 750 hours a month. It will also allow me to just remotely connect to the database instead of hosting the database on a local PC that I own and must go through the manual set up of the database.

The reason that I cannot create a database is due to each row containing data for one loan from a specific bank. There is not a feasible way to split up the attributes in the data set to create a normalized database. All of the attributes that are being used must be used in order to determine the unique identifier for the loan.

I am not going to be using MySQL, as we have used MySQL in class, and for the purpose of this project, I would like to get familiar with different database system syntaxes, as PostgreSQL and MySQL have minor differences in their syntax. I will attempt to add an SSL layer to the database, as that will keep the data secure from the database to the R Shiny app.

For the R Shiny app, the reason that I am going to be using R instead of Python is because R has a larger community and has been around for a longer time than Python. R, to me, is a more statistically and visually powerful program than Python, as Python is very good for programming and statistics, R is more heavily focused on the statistical spectrum.

I do have a template report that I have at my disposal for creating the R Shiny app for visuals, I cannot share that though, as there is information on there that cannot be released to the public, as it is for internal use only.

For the R Shiny app, I would like to generate a set of visuals, and hopefully create a way to generate and download a report that is something like what I have from JP Morgan Chase. There is a solid community of developers and programmers that have very detailed forums/discussion boards about creating an R Shiny app, so I plan on using them.

**Goals**

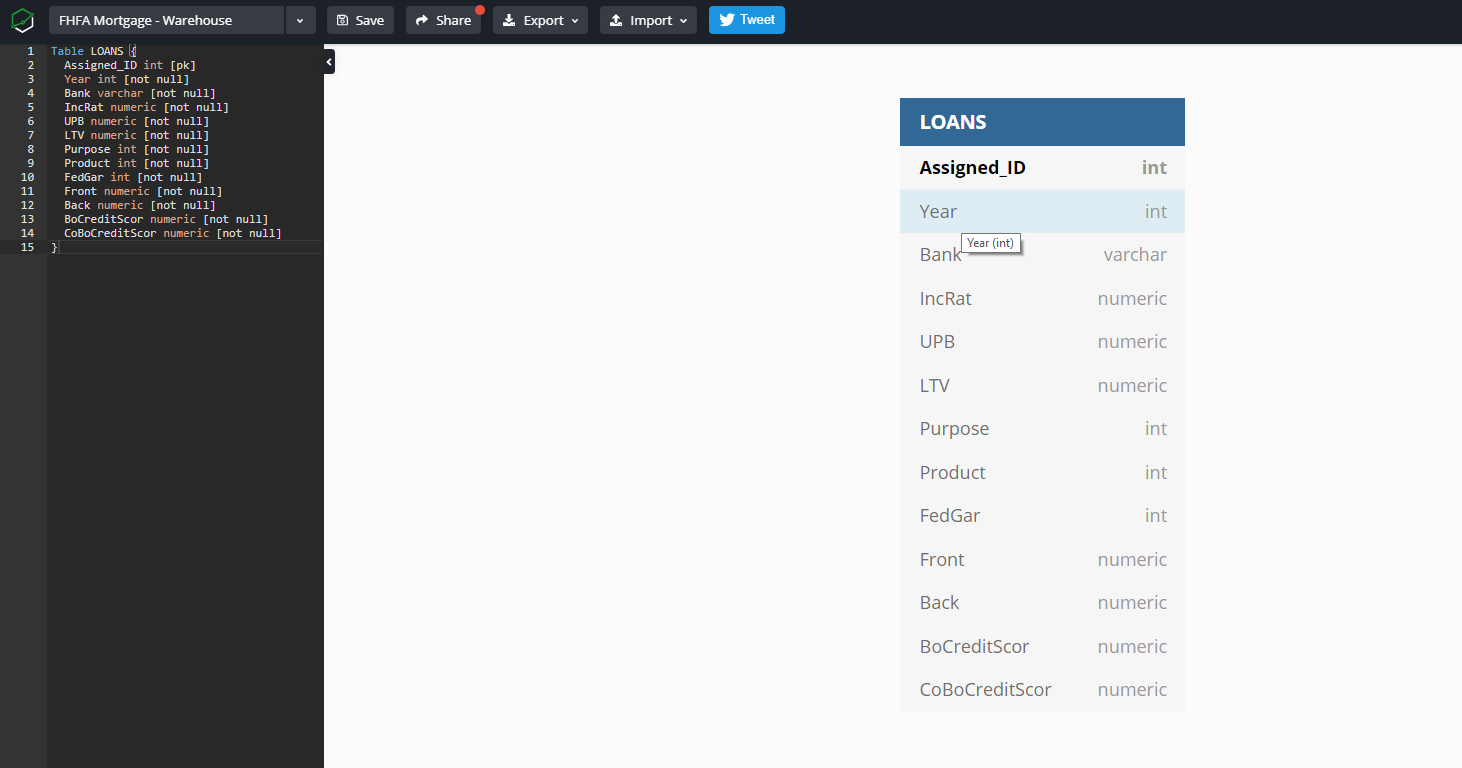
1. Create a fully functional database using Oracle as the infrastructure, possibly PostgreSQL, also may use AWS for this project
2. Create an R Shiny application that will both visualize and export data that is useful to the end user

**Specifications**

This project will begin once I have the set criteria from Rob where he will give me the key indicators that he and his team look at while analyzing the data, and I will then begin to develop the R Shiny apps look and functionality. He will also tell me which data can be left out, as it may not all be useful to himself and his team.

**Milestones/Deadlines**

**ERD** – June 7th



**R Shiny App Format** – June 8th

**Database Warehouse Completion** – June 10th

**R Shiny App Completion** – June 17th

**Additional Comments**

For the milestones/deadlines, I would like both the ERD and R Shiny App layout be completed by the same day, as those will serve as my backbone for creating my project. The ERD to me is the most challenging part to make sure that the database is normalized, and the R Shiny App’s layout will determine what needs to be added and how it must be created.

I would like the database to be completed before the R Shiny App, as I will have to test if the results that are being displayed from the app are accurate. Being completed with the project a week before the presentation will allow me to refine and tune my project if I find any issues with it or if I would like to add on to it with some backend scripting for data verification/quality.

I have left out the Python aspect of the project, as creating the database and the R Shiny app will be enough for the project. It will be enough work setting up both of those, and if I have time, I may do some Python at the end if I want to.