W205 Final Project Proposal
October 11th, 2016

`where(ii)`

By Melanie Costello, Lisa Barcelo, and Priya Gupta

INTRO

In today's job market, new graduates in competitive fields like healthcare and technology are faced with more job opportunities than ever before. They are increasingly relocating to cities that allow them to maximize their income and enjoy excellent quality of life. As part of your job search, **`where(ii)`** can help you figure out where to look. **`where(ii)`** aggregates home rental prices, crime statistics, and income data geographically in order to give you the best options available.

DATA

Our data comes from a variety of sources. We are utilizing data from Zillow to gauge housing prices to determine the average of cost of housing to income. We will use this income-to-housing-cost ratio to determine which area has the highest ratio for the user, and will use this as a benchmark in estimating the likelihood of a fit. Additionally, we will use household income data from Zillow as a baseline to compare a user's income to the median household income. We will incorporate crime data from the department of justice in order to further assess living conditions in a given metropolitan area.

Finally, we will use the income by occupation for each region in the US Census to provide an estimate income for the user. This will give us a consensus on the income to forecast/expect from the area the user would like to check. We will need to do some extensive ETL to unite the data sources. We will rely on Zillow's existing categorization of geo-footprints and make the US Census data conform. We will also have to add new columns derived from both the Zillow and Census data in order to meet the needs of our analysis.

TECHNOLOGY

We will import and store our data files in a Hadoop file system as part of the DDL process. During the ETL phase, we will then clean the files and organize them into tables using Hive and SQL. When considering the end user's experience, we decided an interactive bash script will guide the user through the appropriate selections in order to receive the optimal output. That program will be executed via the AWS web app, which will allow a user to access the program through a browser.

ROAD MAP

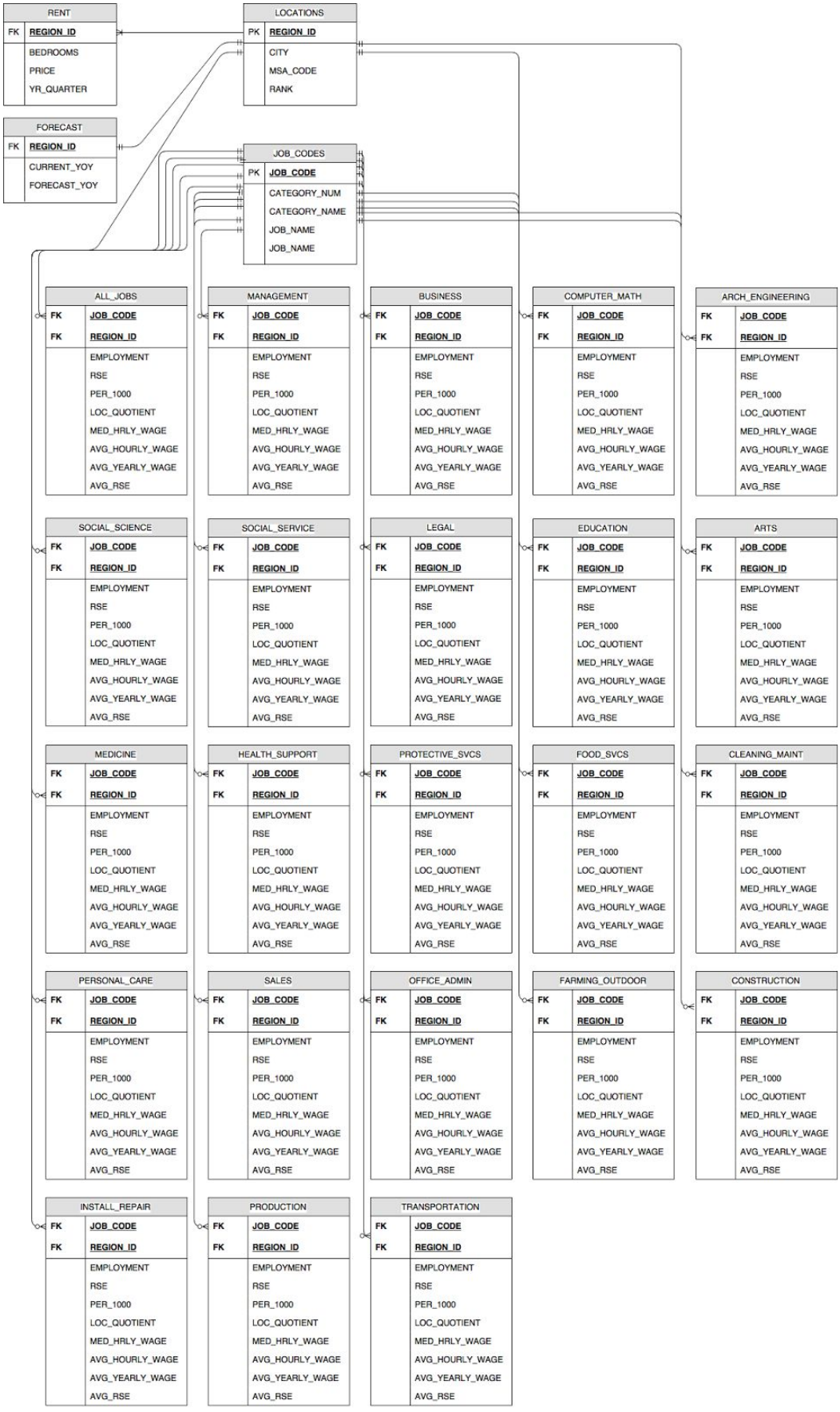
1. Create Business Proposal and Plan -> October 11
2. Gather and finalize data and data sources -> October 17
3. Finalize/narrow down formulas needed to calculate results -> October 17
4. Start EC2 Instance and load filed into hadoop file system. Create scripts to update instance as data is updated -> October 22
5. Create DDL scripts for tables -> October 29
6. Implement calculated fields into queryable tables -> October 20
7. Finalize ETL -> November 14
8. Build web application on top of storage system -> November 14
9. Use case testing and modification phase -> November 30

ENTITY RELATIONSHIP DIAGRAM

We have put together a preliminary entity relationship diagram that is subject to change. The core tables are LOCATIONS and JOB_CODES.

From LOCATIONS, we can branch out to the RENT table which will contain data on rent prices for 1, 2, 3, 4 and 5+ bedroom dwellings from 2010 through June 2016. We can also branch from LOCATIONS to the FORECAST table. The FORECAST table will contain two key data points - CURRENT_YOY (the percentage change in rent from last year to this year) and FORECAST_YOY (forecasted percentage change in rent from 2016-2017). This will allow us to add some nuance to the results returned from queries. For example, we will be able to warn users that rent in a certain location is rising quickly.

The JOB_CODES table branches to 23 similarly-structured tables that are specific to the broad job categories provided in the Census data. We split the data into these category tables to avoid having one monstrous job information table. We know that the majority of queries will be returning only a small portion of the data. Because of this, we hope that this design will allow for continued additions of data, without compromising query performance.


RENT
FK REGION_ID
BEDROOMS
PRICE
YR_QUARTER
LOCATIONS
PK REGION_ID
CITY
MSA_CODE
RANK
FORECAST
FK REGION_ID
CURRENT_YOY
FORECAST_YOY
JOB_CODES
PK JOB_CODE
CATEGORY_NUM
CATEGORY_NAME
JOB_NAME
JOB_NAME
ALL_JOBS
MANAGEMENT
BUSINESS
COMPUTER_MATH
ARCH_ENGINEERING
SOCIAL_SCIENCE
SOCIAL_SERVICE
LEGAL
EDUCATION
ARTS
MEDICINE
HEALTH_SUPPORT
PROTECTIVE_SVCS
FOOD_SVCS
CLEANING_MAINT
PERSONAL_CARE
SALES
OFFICE_ADMIN
FARMING_OUTDOOR
CONSTRUCTION
INSTALL_REPAIR
PRODUCTION
TRANSPORTATION
FK JOB_CODE
FK REGION_ID
EMPLOYMENT
RSE
PER_1000
LOC_QUOTIENT
MED_HRLY_WAGE
AVG_HOURLY_WAGE
AVG_YEARLY_WAGE
AVG_RSE

CHALLENGES AND CONSIDERATIONS

Initially, we have scoped the project to include data on the 50 largest metro areas in the country, based on the Zillow rankings. However, we may be able to add volume to the project by expanding the data to include more cities. Both Zillow and the Census website have additional data that would support this expansion.

Additionally, our intent is to use data on rental prices, but we could incorporate data on home sale prices as well. That addition would pose some challenges. First, we would like queries to return the cost of housing per month. If we incorporate home purchase prices, we would need to compute a monthly mortgage estimate. This would require acquiring data on interest rates, which can change daily, and we'd also have to make some assumptions about down payments that would be made.

We considered the issue of using 2015 census data regarding occupation and income relationships, as opposed to more "current" sites like Indeed and Glassdoor. While the census data is slightly "behind," we felt the data quality would be significantly better--while the census relies on self-reporting data, the sample set is large enough so that outliers would be minimized. Additionally, some current job/income sites prohibit scraping per their API terms of service. The census data is plentiful, available, and standardized, and ultimately our preferred choice. The fact that it is already organized into easily recognizable job categories is also an advantage, as the user will be able to get their occupation as narrow or as broad as they wish.

Our income/occupation data set will be considerably large as we will be downloading information for all metro areas in the US, but we will further add to our data by including crime data for those same areas.