DSCI551 Project Proposal - Housing Database

Changxun Li

September 18, 2022

1 Team Member and Responsibilities

1.1 Team Member

I will be doing this project alone, and the only team member is "Changxun Li".

1.2 Responsibilities

Since this is a solo project, I will be responsible for the entirety of the project including research and task implementation.

2 Project Description and Data Used

2.1 Description and Background

As graduate students get ready to graduate from universities, their next step is likely to enter the industry, making the best of our young adult lives, and potentially create families. It is not far for many to start considering purchasing an property to better accommodate themselves and the important people in their lives. Thus it is crucial that there is somewhere that graduate students can investigate the ever changing property market, since the economy is as unstable as ever and there are many disinformation out there on the market. So it is important for students to observe current listings and make up their own mind, since only they know what suit them the best.

This project creates a database where students can easily browse through current property listings, students can look around at random listings if they choose since they are all listed in a database and free for all to access. But if choose, students will also be able to look up property listings based on certain conditions they desire to better match their needs and desires, since everyone has different needs in their lives. Students will be able to do searches on one or multiple queries, some examples such as price range of the property, number of bedrooms, number of bathrooms, or total land area etc. This database will help students to get a better understanding of the US housing market and make better purchasing decisions.

2.2 Data Set

The data set that will be used for this project is the "USA Real Estate Dataset" on Kaggle, https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset. The data was scraped from realtor.com, https://www.realtor.com/, which is the second most visited real estate listings website in the United States as of 2021 with more than 100 million active monthly active users. Which means that this data set is very diverse, incusive, and great for our purposes, since we are looking to publish this database to many students of different needs. This data set was updated last in August 2022, as of writing this project proposal, so it is very relevant and up to date.

This data set has more than a million entries, with each entry being a property listing, once again very diverse and can help students find exactly what they need. The data set has 12 columns or variables, being status (status of the listing), price, bed (bedroom count), bath (bathroom count), acre_lot, full_address, street, city, state, zip_code, house_size (house size in square feet), and sold_date. All of these variables can be potentially used for search queries, and make student's searches more specific and accurate.

3 Task Implementation

3.1 Task 1 - Building EDFS

I will be using tables to simulate a EDFS, the tables will be the directory structure of the file system. We would have one table for storing metadata of files, one table for the directory structure of the file system, one table for storing the partition information about files, and then other tables to store the actual data that we use for extracting information.

The table for storing metadata of files would have the id of all the files, and file names. The table for the directory structure of the file system would have the parent and child pairs of the files, so we know where to look for other tables. The table for storing the partition information about files would store the location of the tables where actual data are stored, this is needed because we need to partition the large data set into small partitions. We would partition the entire data set into partitions by their state or city, then store each partition in their own table.

3.2 Task 2 - Implement PMR for Search and Analytic

For the search function based on the user input, we would find the locations of all the partitions that could contain results that the user desires. Then we would find the results from each of the possible partitions, and ultimately combine all of them to create the final result to show the user. We would have search possibility on almost all of our variables, since they can all be used to search for specific properties in the real world.

For the analytic function it is similar to that of the search function, we would find the analytic results from each partition and combine them all at the end. Once again we would have analytic on most of our variables, since they can often be useful to get a better understanding of the housing market and make things easier for the students, which was our original goal of this project.

3.3 Task 3 - App for Search and Analytic

For the app interface I will be using either jupyter notebook or a local browser based interface for the user to interact with, the interface would be simple and clean to look at with clear labeling of where everything is for easy access. The interface would have the ability for the user to use the basic EDFS commands, to browse through the EDFS file structure interactively with UI, and for the user to use the search/analytic functions with proper explanations on how each function works. For the search/analytic functions, the user would either choose from a list of all possible options, or have the ability for the user to enter their custom queries.

4 Project Timeline

Below is a table of the project time line, with this proposal is due on 09/19, the midterm report is due on 10/31, and the final delivery being due on 11/28. I spaced the three tasks' completion times between the proposal due date and the final delivery date, with Task 1 of establishing the database having the most time since it is the fundamental of this project, and Task 2/3 taking the same amount of time as they should take less time after the database is setup completely. Finally we would also have the peer evaluations due on 12/02.

Date	Task
09/19	Proposal
10/19	Task 1 finish
10/31	Midterm report
11/02	Task 2 finish
11/16	Task 3 finish
11/28	Final report/project video
12/02	Peer evaluations