Midterm report

Group name: Simplifier

Team members (email):

Jing Wang, wangji9@oregonstate.edu

Webster Cheng, chengwe@oregonstate.edu

Xinzhuo Cheng, chengxi@oregonstate.edu

Yiting Wang, wangyit@oregonstate.edu

Yi-shan Ko, koyi@oregonstate.edu

1. What we have done?

In this project, we will complete a website for users to search the Airports, Airline, and Air flight. In the search results, if users search specific Air flight, they can see the information about this air flight, for example, the flight schedule, on-time performance, and company. Users also can know the rank of airports and airlines. Moreover, we will present a visible data about air lane and hot spot of airports' traffic. In this sprint, our team has already finished scraping the data and decompose the sources what we have scraped. Yi-shan Ko works on the historical flight schedule and the information about the airport. Yiting Wang is responsible for scraping the on-time performance of airports and the number of operation. Webster, Xinzhuo Chen and Jing Wang, they cooperate to scrape the information of Airlines include the on-time performance, number of operation, departure, destination, etc. In order to integrate the data, we made the temporary DB schema to display the data relation as figure 1 and summarize the info of tables we have as figure 2.

2. Current Problems:

Firstly, we have faced the missing data problems, when we try to insert and gather our datasets from different resources. Some data cannot be loaded by any reasons such as types of attributes and too large data. Secondly, we are not domain experts, there is no way to choose the right PK rapidly. Hence, we need time to learn the knowledge of the professional field in order to find the right PK. Finally, the inconsistent attribute types are faced when we join data from different resources. Hence, the FK selection is difficult for us to create a correct DB schema. Also, choosing appropriate column types need time to check whole the data to avoid error.

Therefore, creating relations between the data sets are quite complicated. It is a serious issue to match the column form data.

3. Future problems:

   In the future, performance issues will be our most concerned point. Because "SCHEDULE" data is too large which have more than 120,091,989 records. Also, we will find more sources for complete data set provision and integration. Hence, an efficient database will be our first priority in the future. Another issue is table schema that may change continuously that will affect the displaying website. Because we may need to alter the type and size in order to fit in different data.

   Displaying issue is concerned also. We want the interface to be as clear and convenient as possible. For example, a website with dynamic filters and multiple figures allows users to get the data only they needed. These functions would take time to do data collection as well as website design, and we may face some software technical problems.

4. A brief plan on how you plan to address the aforementioned issues.

   We will spend a few days doing data integration such as building index and FK link. These take time to organize the existing data more consistently to avoid more problems in the future. We will decompose the data again according to the type of data into small data sets. Moreover, if the information is still too large, we may remove the old data to another data sheet. In order to display the data, we will use the website framework we did before and will add appropriate UI to display data.

Figure 1 – Temp DB schema

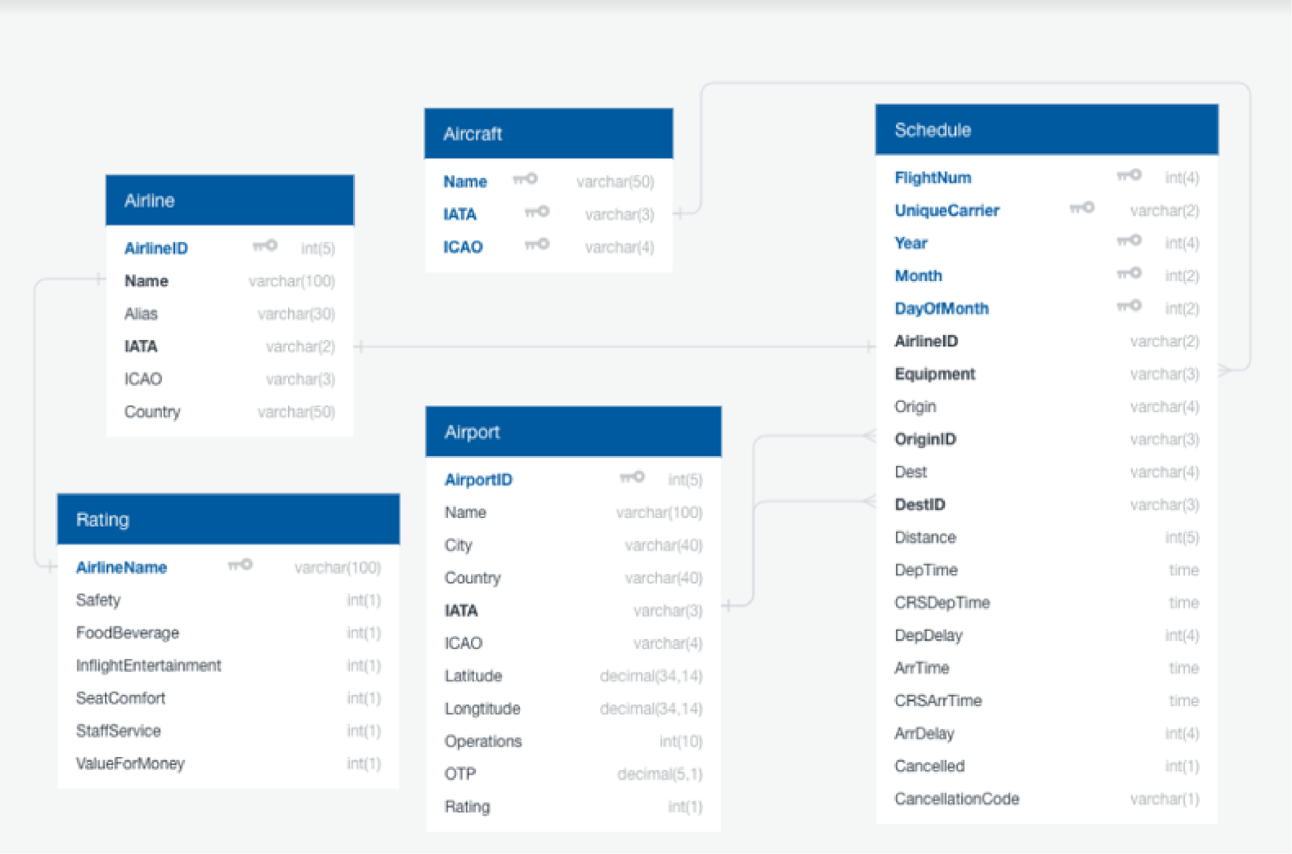

Figure2 – Table Record info

TABLE NAME	RECORDS
AIRLINE1	5882
RATEING1	364
AIRCRAFT	174
AIRPORT1	13726
AIRLINE2	21317
SCHEDULE2	100000
AIRPORT2	512
SCHEDULE1	120091989
RATING2	486
AIRPORT_TRAFIC_CAPACITY	18876