ALY 6030 - Data Warehousing & SQL - 70526  
Module 2: Data Modeling, Normalization and Advanced SQL

Professor: Valeriy Shevchenko

CPS in Informatics

Chongfan Bao

bao.c@northeastern.edu

Northeastern University

10/05/2021

**Abstract**

In project 2, I am about to design entity relationship diagram (ERD) on MySQL workbench. Meanwhile, I am going to explore and study the data normalization rules and SQL Joins syntax to discuss business questions. All these will be conducted on the airport dataset. Bases on this dataset, I will normalize the data first into at least five tables and import data into My SQL workbench , then to research my business questions.

*Keywords:* SQL language, ERD, Database normalization rule.

**Introduction**

In this module 2, I am going to choose the airport dataset to conduct my business analysis. Meanwhile, I will also start to try to leverage data normalization rule to normalize this dataset and draw entity relationship diagram. The airport dataset is about the distribution situation of airports in different cities and states in US. There are totally 8 attributes, which includes numbers and string, and 221 records. Then, I will separate the dataset into five tables: country table as a parent table, state, city, airport and statistic as child tables. Based on these tables, I will use SQL joints syntax to conduct business questions and draw ERD in MySQL workbench. When I finish this module, I hope that I can learn how to normalize data and more familiar with this tool. Also, I hope that my problem-solve skill can be promoted.

**Data Normalization**

A good database design can avoid data redundance and enhance the capacity of computer. So, by checking the airport dataset, I find that there is some dependent relationship among some attributes and redundant data. For example, the Now York city has more than one airports: LaGuardia and John F Kennedy Intl airport. Also, a state can have many cities, a country has many states and so on. Therefore, I need to separate them to different table. Each of these table is connected by the foreigner key that I need to create and set it by myself so that we can get a relational database and refer one table to another. Following figure 1 displays my database SQL code.

From figure 1, we can know that we start to create the country table and makes countr\_code as string and primary key. I only set one attribute that is because the country name is unique and can be as a primary key. Then, I create the state table that includes one primary key state\_code and a foreigner key country\_code to connect with country table. Next, I create the city table that includes two primary key city name and state\_code to make sure the city is unique. Sometimes, some states can have the same name city. What’s more, I create airport table to storage the information of each airport, which is made of airport name, latitude, longitude. Here, I set the latitude and longitude as double characters to align with the original dataset format. Finally, I create the statistics table to store cnt date. I artificially add id attribute and make it as primary key to ensure iata\_code all the attributes are functionally dependent on solely the primary key. For all tables except for country table, I have created restrict rules on update and delete operation in order to keep data consistency and avoid mistakes when we manipulate the dateset.

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

Figure 1

From figure 2, we can know that the relationship between a pair of tables how they connect. As we see, begin with country table, there are is a one to many relationship between country table and state table, which means that there can be many states in a country. So as other tables. They connect by the foreigner key.

Diagram

Description automatically generated

Figure 2

**Business Insight**

Based on the dataset that I have created, I am going to explore three business questions by using SQL: What the airport information in New York City, such as how many airports and the location? How many airports are in the LA? What is the distribution of airports between latitude 30 and latitude 31?

From figure 3, I use the left join SQL syntax to try to select airport\_name, lat, long, city\_name, iata\_code, cnt six columns from the airport table, city table and statistic table separately. Meanwhile, I change some attributes name, like long as longitude to make people easily to read. This is to figure out how many airports in New York city and know their location. As we expect, the table displays two airports in New York City. They are John F Kennedy Intl and LaGuardia airports. We also can find exact location based on latitude and longitude data from map.

Graphical user interface, text, application, email

Description automatically generated

Figure 3

From figure 4, I attempt to leverage inner join SQL syntax to calculate the number of airports in LA. We can know that there are 6 airports in LA. Each city only has one airport. There are two international airports in LA. So, people maybe prefer to choose LA airports as their transfer station or first station to travel.

Graphical user interface, text, application

Description automatically generated

Figure 4

From figure 5, I attempt to leverage right outer join SQL syntax to explore the distribution of airports between latitude 30 and latitude 31. As we see, The Florida state has most airports in this area, totally 6 airports, which means the distribution of airport in this field is quite dense. We can also know the number of airports and their name and location in TX, MS and so on. If we want to know more information, we can add more tables.

Graphical user interface, application

Description automatically generated

Figure 5

**Conclusion**

By this project, the biggest gain is that I know how to use the SQL joint syntax and how to set primary key, foreign key and constraints by MySQL workbench. Also, I know how to draw and modify the entity relational diagram. Besides, I successfully solved the three business questions I claimed. I know that there are two airports in New York City and there are two international airports in LA. When the latitude is between 30 and 31, the Florida state has most dense the number of airports. However, even if I know how to use SQL, I still need to fix many problems and improve my skills. The first one is that when I try to set up the constraints in different table, the tool cannot work. So, I have to separate into different step. Even if we have the same data with my friend who is specialize in SQL and worked for a database company in Chine, the code is work for him but not for me. Then, he told me that maybe my version has some bug. What’s more, when I start to code, I really need to be careful. This is because even if a different sign the system will pop out mistakes. All in all, I still need to try my best to practice how to solve problems and do more practices.

**References**

Unknown column in 'field list' error on MySQL Update query.Retrieved from https://stackoverflow.com/questions/1346209/unknown-column-in-field-list-error-on-mysql-update-queryOracle Corporation, 2021. Configuring Application Character Set and Collation. Retrieved from https://dev.mysql.com/doc/refman/5.7/en/charset-applications.html

How to solve “Column 'user\_id' in field list is ambiguous” error? Retrieved from https://www.stechies.com/column-userid-error/