ALY 6030 - Big Data and Data Management- 70890  
Module 1: Discovery on Demand computing and Jupyter notebook Lab

Professor: VJanos Mako

CPS in Informatics

Chongfan Bao

bao.c@northeastern.edu

Northeastern University

11/05/2021

**Introduction**

In the first week, I will select the TechCrunch dataset regarding start-up company raising funds, which includes 9 attributes and 1460 records. Based on this dataset, I start to learn to use the MySQL server and MySQL workbench to build a database and blank table to restore the dataset. This is in order to further learn fundamental database knowledge and how to use these tools and SQL syntax formally. Meanwhile, I claim some business questions, such as which state or company raises the most money? Then, I am going to try to use SQL language to solve the business questions I claimed before. After practicing this module, I hope that I can be more familiar with how to use this tool and SQL syntax. The most important is that I can fast abstract useful information by SQL to solve real-world problems straightforwardly and manage the databases.

**Tool’s installment**

This picture 1 shows the information about the MySQL server included in the version, and this is the first step to install the software.

Graphical user interface, text, application, email

Description automatically generated

Figure 1

This picture shows the MySQL server license agreement included the copyright and responsibility statement and so on. Also, this is the second step to install the software and click continue icon to get into agree pane and click agree icon to next step.

Graphical user interface, text, application, chat or text message, email

Description automatically generated

Figure 2

Picture 3 shows MySQL server installation. In this step, you can choose the local where to install and which software you need to install. Usually, we set it by default and hit the install icon to the next step. We select the preference pane means that we can open MySQL from MAC system preference pane directly and don’t need to use the terminal.

Graphical user interface, application

Description automatically generated

Figure 3

Picture 4 shows MySQL server installation officially started after inputting the passwords. It will take few minutes to install the packages that we selected before and put them in the /usr folder.

Graphical user interface, application, email

Description automatically generated

Figure 4

Based on pictures 5 and 6, I know that we need to encrypt the MySQL server as a root user. This is to intensify the security of the database.

Graphical user interface, text, application, email

Description automatically generated

Figure 5

Graphical user interface, text, application, email

Description automatically generated

Figure 6

According to picture 7, I know that we have successfully installed the MySQL server.

Graphical user interface, text, application, email

Description automatically generated

Figure 7

When I install this software, to be honest, I met a lot of troubles. I installed it when I took the ITC 6000 class. However, according to the professor’s requirements, I have to re-install this MySQL server. When I installed the latest version, I found that I couldn’t open it by the system preference pane, but I can open it by the terminal. It’s weird. Lately, I realized that it could be a problem with the version, so I chose an old version, and the problem is gone. All in all, I try many methods before finding this solution.

From picture 8, I am ready to install MySQL workbench that is a GUI tool to manipulate and manage the database intuitively. After I downloaded the package and double-clicked the file, it would show as following. Then, I just need to drag it to the application folder. The system will automatically install this software, I don’t need to set anything. With this software, I can connect with the MySQL database server and extract data from it. I can also build my own database on the server using MySQL workbench.

Graphical user interface, application

Description automatically generated

Figure 8

**Data Selection and Build Database**

I will select the TechCrunch dataset regarding star-up company raising funds, which includes 9 attributes like company, category, city, raised amount, and 1460 records. I noticed that the dataset is made of int and string. It also includes null values and no primary key and foreigner key, which means when I create a new schema in MySQL, I don’t need to set any primary or foreigner key and allow the attribute to have null values.

From picture 9, I start to create a database named ‘DB1\_bao’ using MySQL workbench. For all processes, I will use SQL syntax. After I click the yellow flash icon to run this code, I will get a new database located on the left pane under the Schemas menu. Meanwhile, people can notice that I create the tables created in the database used utf8 and utf8\_bin for any character columns. Normally, for applications that store data using the default MySQL character set and collation (latin1, latin1\_swedish\_ci), no special configuration should be needed (Oracle Corporation,2021).

Graphical user interface, text, application

Description automatically generated

Figure 9

In figure 10, I start to create a blank table to fill in the data that I chose before. According to the SQL syntax and original data attributes, people notice that I set the ‘numEmps’ and ‘raisedAmt’ as INT, other attributes as VARCHAR which can store 45 characters that are enough to store all data in this case. In the future, we also can change the length of VARCHAR. Also, I make the rule that some attributes can be null values and others are not be null. One more thing that people need to notice is the ‘fundedDate’ cannot be DATETIME attributes. Otherwise, people cannot import the raw data into the blank table.

Graphical user interface, text, application, email

Description automatically generated

Figure 10

In figure 10, I start to import the raw into the table using MySQL workbench. The mouse right-clicks the Tech\_table and chooses the file pathway to import the data.

Graphical user interface, text, application, email

Description automatically generated

Figure 11

From figure 12, we need to choose the using existing table and truncate table before import so that we can delete the existing data and keep the original structure of the table to avoid data redundancy.

Graphical user interface, application

Description automatically generated

Figure 12

From figure 12, people can notice how long the data and how many records are imported. Also, based on the number of records, we can know if the data is imported successfully. If it succeeds, we can check it by MySQL syntax in MySQL workbench and conduct data analysis and manipulation.

Graphical user interface, text, application, email

Description automatically generated

Figure 13

**Business Insight**

Based on the dataset, I will discuss three business questions by using SQL language: What is the situation of the round of raising money? Which tools are more popular to raise money in the market? How is the distribution of company raising money in MA state?

From figure 14, we can know the number, amount, maximize, minimize and average of each round of raising money. A round gets the most money, which is 126,095,000 dollars and total 169 funds. The largest fund is300,000,000 dollars, which happened in the c round. The smallest fund is 10,000 dollars, which happened in the seed round. In the d round, the company gets the highest average money, which is Another interesting thing that I found is that even if some companies are in debt, they also can raise money. I guess that the investor believes this company has great potential. The purpose of the table is to roughly understand the average funds in the different round so that the start-up companies can maximize the raising money. Also, the company can analyze the reasons why they cannot get more money compared with average raising money.

Graphical user interface, application

Description automatically generated

Figure 14

From figure 15, we can know that the most common way to raise money is from the website. There are almost 90% of start-up companies getting money from the web. So, for investors and start-up companies, the best way to find a good start-up company and raise money is to search for the website.

Graphical user interface, text, application, email

Description automatically generated

Figure 15

From figure 16, we can know that the situation of raising money in MA state. The Boston Power raised 45,000,000 dollars in c round in January 2008 by website. Based on the table, we can know any company’s fundraising in MA. If we want to know other states, we just need to change the IN syntax.

Graphical user interface, application, email

Description automatically generated

Figure 16

**Conclusion**

By this project, the biggest gain is that I know how to create databases tables and manipulate the dataset by SQL syntax. I successfully solved the three business questions I claimed. I know that website is the most popular way to raising money. Normally, a round gets the most money and Boston Power raised 45,000,000 dollars in c round in January 2008 by the website. 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 I don’t know why when I imported raw data into my blank table, it only imported part of data about 425 records. I have asked classmates and googled and found that they leveraged the same method. They can import all data. Nevertheless, I just can get part of the data. Besides, the system didn’t show any mistakes. What’s more, I find that as I leverage the BETWEEN syntax, there are not any outcomes and mistakes in the action output pane. All are weird. I still need to work on this in the future.

Additionally, I also find this MySQL workbench is not good as I thought. For example, in figure 14, if I want to show the raising money situation of each company in different round. I cannot use the SQL syntax to present it. Or the tools themselves haven’t this function. Also, the MySQL workbench is not easy to check mistake reasons under the action output pane. Sometimes, you need to write a lot of code to visualize data. If I use Microsoft Excel, I just need to click the mouse. But it is good for dealing with massive data and faster than Microsoft Excel.

**References**

How To Use MySql On Mac.Retrieved from <https://www.dev2qa.com/how-to-use-mysql-on-mac/#:~:text=You%20can%20run%20the%20command,enter%20the%20MySQL%20interactive%20console.&text=Enter%20password%3A,Welcome%20to%20the%20MySQL%20monitor>.

Oracle Corporation, 2021. Configuring Application Character Set and Collation. Retrieved from https://dev.mysql.com/doc/refman/5.7/en/charset-applications.html

SQL WHERE IN. Retrieved from https://www.dofactory.com/sql/where-in

Structured Query Language (SQL). Retrieved from https://www.geeksforgeeks.org/structured-query-language/