BIG DATA I FINAL PROJECT PART 1

November 3rd, 2019

Group 16
Yanghe Liu, Qingyue Su,
Xiaofan Sun, Hsiang Yau Tsai,
Pengcheng Xu

We use five Python modules to:

1, create a database named “db_consumer_panel” in mysql server from Python;

2, import data from csv files into Python pandas dataframes;

3, populate the “db_consumer_panel” tables in mysql server by importing data from the pandas dataframes.

The five modules are: mysql.connector, csv, os, sqlalchemy and pandas.

In the beginning, we create two variables in Python. They are DB_NAME, which is a string and equals to the name of the database we would like to name, and TABLES, which is a dictionary. TABLES takes names of tables as keys and mysql queries to create variables in one table as values.

Then we enable connect and connect functionality from mysql.connector to connect to mysql server from Python. After that, we define a function create_database which takes cursor as variable to create a database named using the variable “DB_NAME.” Then we run a for loop over TABLES, and in each loop, cursor.execute function takes keys from TABLES to create tables in mysql server.

After that, we import data from csv files to Python and store them as pandas dataframes by using pandas.read_csv(file_path) function and drop unnecessary columns and duplicated data entries in the dataframes.

Finally, we use engine function from sqlalchemy module to create an engine to import data from pandas to mysql server. pandas.to_sql function takes name of the table in mysql database, a corresponding dataframe in pandas, con=engine, index=False to populate the table in mysql database using data from pandas. In addition, we define chunksize = 1000 for each importation so that the importation process can be faster.