Project Phase 1 - Data Collection Module

Design Report

Group 5

Song Yang
Xin Yang
Ke Xia
Yiyuan Fu
Zhuohang Li

ECE 16:332:568 -
Software Engineering Web Application

Mar. 1st, 2018

Individual Contributions

- Song Yang (21.5%):

  Designed and implemented the real-time data getter using a single thread and insert fetched data into the local database.

- Xin Yang (21.5%):

  Designed the historical data Python crawler and connected the python program to database.

- Ke Xia (21.5%):

  Designed and built the database. Encapsulated functions to access and manipulate database in python.

- Yiyuan Fu (14%):

  Test API and programs, find problems in finished individual parts. Try to find data from Yahoo.

- Zhuohang Li (21.5%):

  Implement program to collect both real-time and historical stock data and save to csv files using API.

1. Introduction

This phase is part of the ECE 568 project to develop a stock-forecasting website. For this module, we are aiming to develop a program that can run continuously to retrieve stock information from financial website and stores the extracted data into a local database. We'll be using python to develop the crawler code and MySQL for the relational database. The information of following 10 stocks will be used for demonstration: GOOG, AABA, CSCO, T, WMT, NOK, NFLX, APA, NKE, GE.

2. Implementation

The implementation contains two main stages: data fetching and data storage.

Data Fetching

We've developed two python programs to collect real-time and historical stock information separately.

- Python Web Crawler

Thanks to two useful libraries imported in this part: BeautifulSoup and Selenium. We came up with the idea of retrieving data from Yahoo Finance using Python spider.

Main procedures:

1. As Yahoo Finance only displays limited data unless scroll down to the bottom, Selenium will utilize browser driver to mimic the scrolling down operation for the complete page access.
2. BeautifulSoup4 will then resolve the HTML file into "soup" data structure for easier data processing.

- Real-time Data Crawler

The API service from Yahoo finance has been officially suspended since mid-2017, so we have to find other methods. Our result is finding the "HTTP GET" request sent to Yahoo and resolve what we need from returned JSON data.

Another way is calling functions in "alpha vantage" API to get both real-time and historical data. This program is designed to run continuously and store the data real-time.

Data Storage

We use MySQL to store the collected data locally in this early stage. In the future phases, we might transplant data into a remote server. The entity-relationship diagram and UML are as below:

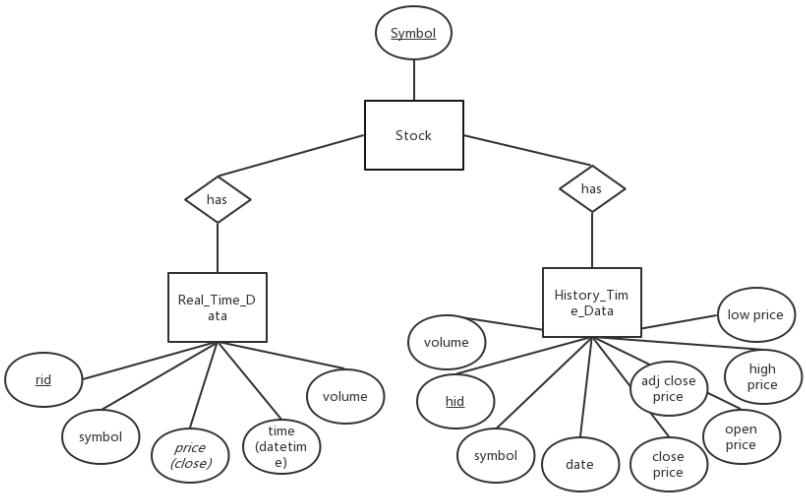

Figure 1: E – R Diagram

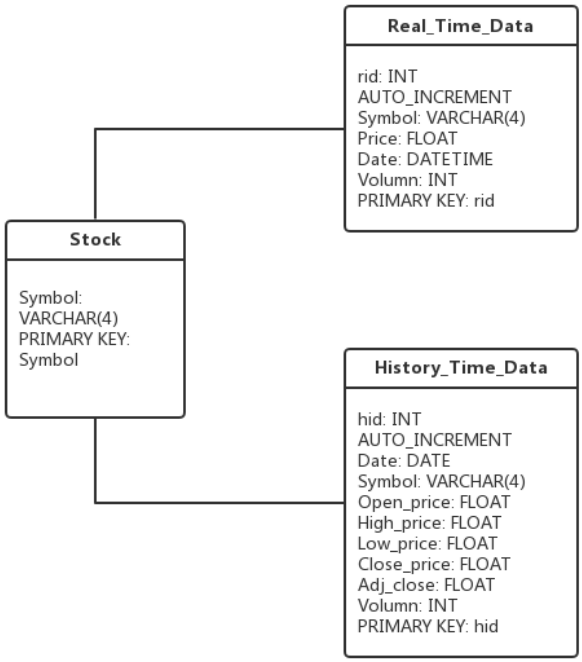

Figure 2: UML Diagram

Three tables are designed in database “stockDB”. Table “stock” stores the basic information of each stock, table “Real_Time_Data” stores the real-time data for each stock, including symbol, close price, date(date and time) and volume, and table “History_Time_Data” stores the history time data, including date, symbol, open price, high price, low price, close price, adjusted price and volume for each stock. “Symbol” in table “Real_Time_Data” and “History_Time_Data” are the foreign keys which reference to “symbol” in table “stock”.

Since it's not necessary to keep the stock information in this phase, we only implemented the "Real_Time_Data" table and the "History_Time_Data" table.
Python library "MySQLdb" is imported to connect to MySQL server and execute SQL sentences in this project.