**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:
* Xin Yang:
* Ke Xia:
* Yiyuan Fu:
* Zhuohang Li:

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.

1. **Implementation**

The module can be divided into two main stages: data fetching and data storing.

**Data Fetching-Python Web Crawler**

We’ve developed two separate sets of python code to collect both real-time and historical stock information. One direct way of doing involves two python packages: Selenium and BeautifulSoup4. The main idea is to

1. open financial website (Yahoo-finance is chosen for this project)
2. use Selenium to control browser driver and mimic mouse operation to scroll down and show the full page.
3. use BeautifulSoup4 to resolve HTML to soup package to get data.

Another way is to call functions in alpha vantage API to get real-time and historical data. The program is designed to run continuously to collect data every minute and store into local database.

**Data Fetching-** **Realtime Stock Data Getter**

The API from Yahoo finance is officially announced closed sine middle of 2017, so we cannot just send a request to yahoo to get the real-time data of a specific stock. The good news is sending a http get request to yahoo will still get information although the JSON data returned is really such a mess. I tried to find the useful information among thousands of paths, and get what I want.

**Data Storing**

MySQL is used to build a local relational database to store data collected in the previous stage. The entity-relationship and UML are as shown in the Figure #1:

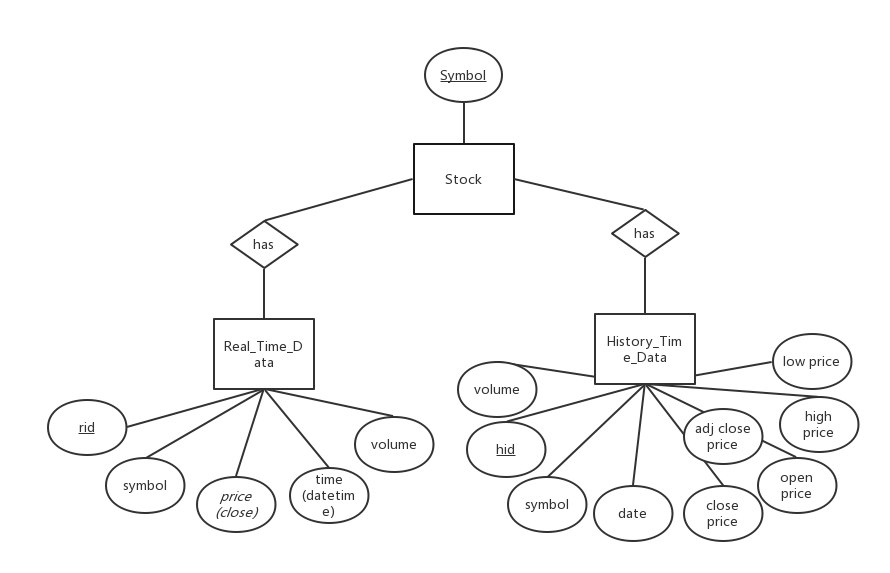


Figure 1: E – R Diagram

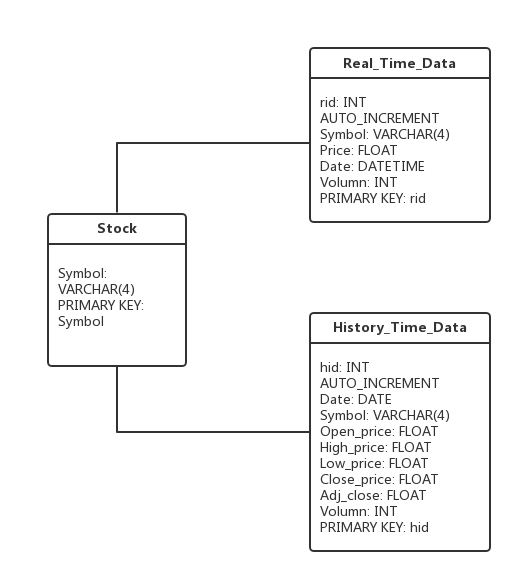


Figure 2: UML Diagram

In total, three tables are created to store data 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”.

Library “MySQLdb” is used to do SQL operations in python.