1. **Application information**

In project phase 1, the application is designed to run continuously as a background process and periodically retrieves 10 stocks information including current and historical from Yahoo! Finance, parse the received responses, and store the extracted parameters into the database.

1. **Database schema design**

In our project, we chose 10 companies to collect data, including Google, Yahoo, Nvidia, Amazon, Tesla, Apple, Coca-Cola, McDonald’s, Microsoft and Sony.

We are using MongDB as the database in our project. In phase 1, we created two collections in the system, year\_info and realtime\_info, respectively. Since we collected the two collections of data slightly different, the two collections’ structures are designed respectively.

The collection in year\_info database has only one document contains six elements: time, high, low, open, close, volume. Each element here saves a large list that stores all data be required. It is the ideal structure to hold the stock data as we need to store many different numerical values all related to a single stock within one year stored as an array. Each field here is an array which stores the stock information in one year and provide the method to index specified data by array subscript.

Since realtime\_info is going to get a new data every minute, we design to create a new document every time we get new data. In this collection, each document still contains six elements, but each elements only contains one newest adta. The value of every minute is shown in a row for the collection is uploaded every minute to maintain the latest data.

The schema of two collections are shown in Fig.1.

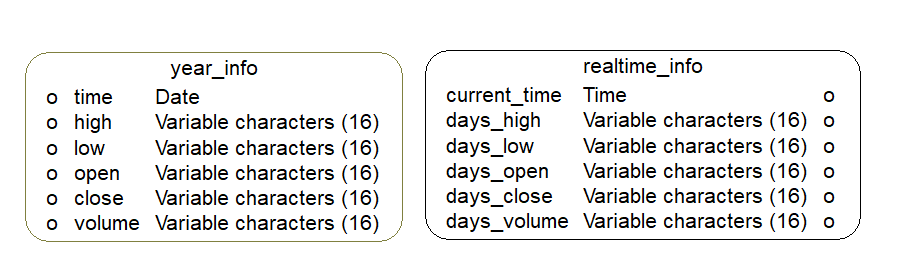


Fig.1. The collections of historical and real-time stock information. The two collections both contain time, high, low, open, close and volume about the stock.

1. **Data collection**

The second part of our code is that we try to get information and extract data. We make a url and request for data from Yahoo Finance website. As we mentioned before, there are two parts in data collection. The first is the historical data, which should contain at least data of one year. For this part, we use two timestamps to determine the time horizon so that they can get the data of the whole year. It provides a convenient way to get a whole year’s data and make it easier for users to process them. The second part is to collect real-time data. Different from the previous situation, this time we don’t use timestamp to determine. We will return and record today’s data in multiple time periods, and we record them every time within one minute. Then we write data into json file. This real-time data file can give us real-time updates for the data we need and it’s convenient for user to monitor it. To sum up, we can get two files in this project because we return and record our historical data and real-time data separately.

1. **Assignment distribution**

As for the division of work for our project, Runlin Hou and Yuxiang Song are responsible for data collection and coding. Ziheng Wang and Sichao Wang are responsible for wrapping up and writing report. We believe that each of us has equal contributions to the project.