**组员分工：**

曹润泽：爬取**中国证券网**股票数据

王 阳：爬取**金融界**股票数据

赵 航：爬取**东方财富网**股票数据

黄恺晟：爬取**tushare**股票数据

项目地址：<https://github.com/MachineDora/BigData_stage2>

**数据源：**

本次数据集成集成了4个数据源，分别来自**中国证券网、金融界、东方财富网和tushare接口。**主要就是让各个数据源的数据相互补充，有些在一个数据源上得不到的数据可以通过另一个数据源来获得，同时，有些相互冲突的数据可以通过集成更多的数据进行约束整理。

**获取方式：**

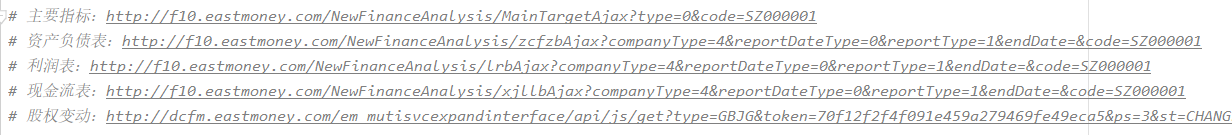
其中，中国证券网、金融界、东方财富网我们采用传统的代码爬取方法，即bs4+request框架从网页中爬取。而tushare是python自带的一个api接口，我们可以非常方便的获取股票信息，更多的是包含动态信息。

由于股票信息不只是停留在一个网页，因此中间可能包含多个子网页，所以常规的爬取可能比较慢，效率低下。（中国证券网、金融界）

同时有些实时数据在网页中无法获得，比如今日开盘这些信息等等。常规操作只能获得静态数据，比如公司名称、注册资本等等。

**关键代码：**

因此要突破限制，对于实时数据要找到其对应的api接口，这样就可以直接进行数据对接，而不需要解析html文本，从而实现更快速的数据爬取。（东方财富网）



同时，这里提供了爬取网页的另一种方法，中国证券网和金融界都是通过标签子元素来获取数据，东方财富网是通过正则表达式进行的：

