# 股票大数据分析平台的构建

**作者**：吴光磊，王惊涛，杨磊，冯永

**单位**：重庆大学，计算机学院

**案例版权**：该案例归重庆大学计算机学院所有

**涉及的知识点**：Hadoop 、Spark、 线性回归

**案例来源及案例真实性情况**：该案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。

**摘要** 随着大数据时代的到来，人们逐渐认识到了数据的重要性。数据不仅是一种资源，更是一种财富。在大数据应用领域中，金融数据分析被视为一个很有前景的方向。股票分析一直是金融领域一个很热门的话题，而且涉及多个领域的知识。在此之前，人们更多的是采用基本分析，即通过宏观及微观的经济政策、本行业领域的发展状况、投资者的行为态度、反映企业自身发展状况的指标等方面来预测股票今后走势。随着大数据相关技术的发展，在海量的股票历史数据中发现规律进而预测股票走势成为一个很热门的研究课题。本案例选取中国股票大数据分析平台的构建为主题，使用源于财经网站的股票数据作为数据集，结合数据挖掘、机器学习等相关方法进行股票数据可视化及走势预测。本案例核心内容有：（1）使用Hadoop 和Spark搭建服务平台；（2）股票数据预处理及可视化；（3）使用机器学习算法进行股票走势预测。

**关键词**： Hadoop ，Spark，机器学习，股票数据可视化，走势预测

## 1 引言

该教学案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。该案例以中国股票大数据分析平台的构建为具体问题，需引导学生进行的主要内容有：（1）掌握并使用Hadoop 和Spark搭建服务平台；（2）掌握股票数据预处理及可视化方法；（3）掌握常用机器学习算法进行股票走势预测。

## 2 背景介绍

随着大数据时代的到来，人们逐渐认识到了数据的重要性。数据不仅是一种资源，更是一种财富。在大数据应用领域中，金融数据分析被视为一个很有前景的方向。股票分析一直是金融领域一个很热门的话题，而且涉及多个领域的知识。在此之前，人们更多的是采用基本分析，即通过宏观及微观的经济政策、本行业领域的发展状况、投资者的行为态度、反映企业自身发展状况的指标等方面来预测股票今后走势。随着大数据相关技术的发展，在海量的股票历史数据中发现规律进而预测股票走势成为一个很热门的研究课题。股票市场发展到目前为止产生了海量的数据，包括股票行情数据以及股票交易数据。股票金融分析与当今火热的大数据有很高的契合度。在第一届的CCF大数据学术会议中，大会委员会曾预测接下来几年内大数据应用前景比较光明的领域分别是互联网、金融、健康医疗、政府管理及安全等。IBM高级工程师曾预测金融投资领域将会是大数据应用最有前景的领域之一。

利用大数据分析技术去探索、挖掘大量股票数据背后蕴含的价值信息，有着广阔的前景。无论是工业界还是学术界均仍需大量的相关人才，对专业硕士进行股票大数据分析的指导是有必要的。在本案例中，我们选取中国股票大数据分析平台的构建作为主要研究问题，借助大数据分析与处理、数据挖掘、机器学习的相关方法，以满足培养具备创新能力与实践能力的人才的需求。

## 3 内容

该案例的主要内容主要分为三个小节，分别为服务平台的搭建、股票数据预处理及可视化以及股票走势预测。

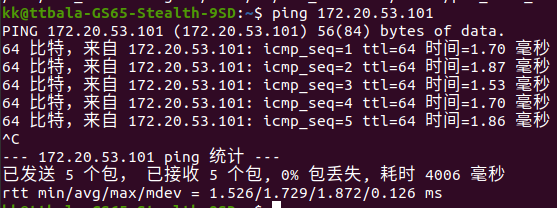
3.1 服务平台的搭建

3.1.1 完全分布式Hadoop集群搭建

本文使用Hadoop进行完全分布式集群搭建，共2台主机，worker1（172.20.53.166）、worker2（172.20.53.101）。在安装Hadoop前需完成两台设备之间相互SSH免密登录授权。之后进行Hadoop的安装与部署。

具体实现步骤：

①实现集群中的机器之间ping通

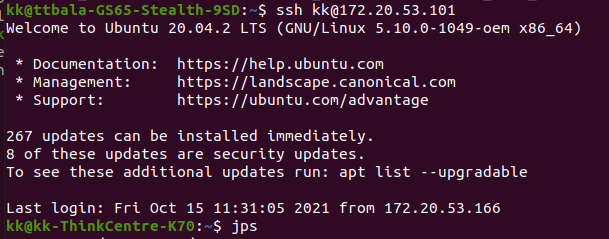


②配置SSH实现免密登录

在主机上生成密钥与公钥。



测试免密登录： ssh 用户名@机器名



③修改hadoop配置文件

修改core-site.xml文件：



修改hdfs-site.xml文件：



修改yarn-site.xml文件：



修改worker文件：

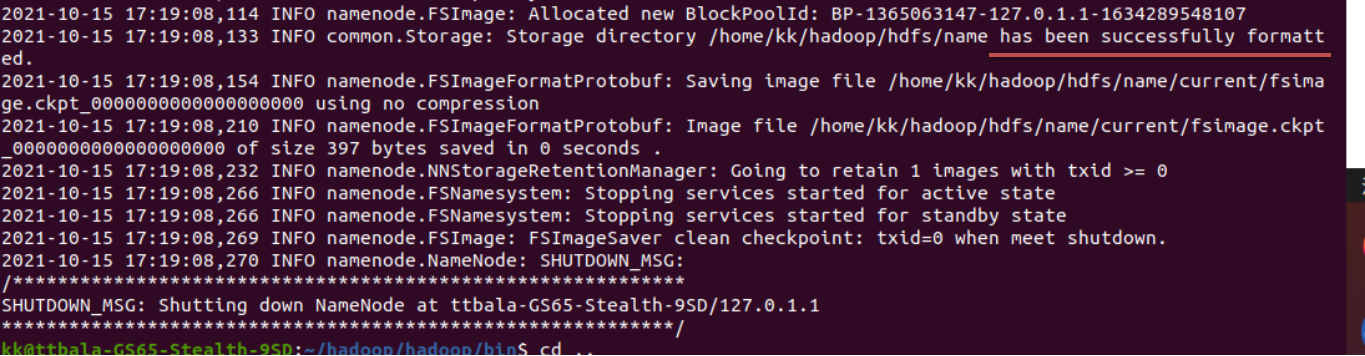


利用scp命令从主机向从机分发配置好的hadoop，虚拟机的话可以利用镜像文件，这样的话集群中的电脑均配置好了hadoop。

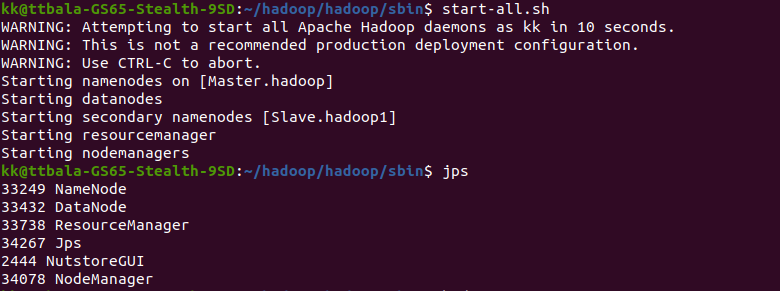


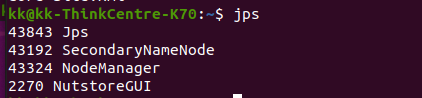
④对Hadoop集群进行验证：

格式化HDFS文件系统：hadoop namenode -format

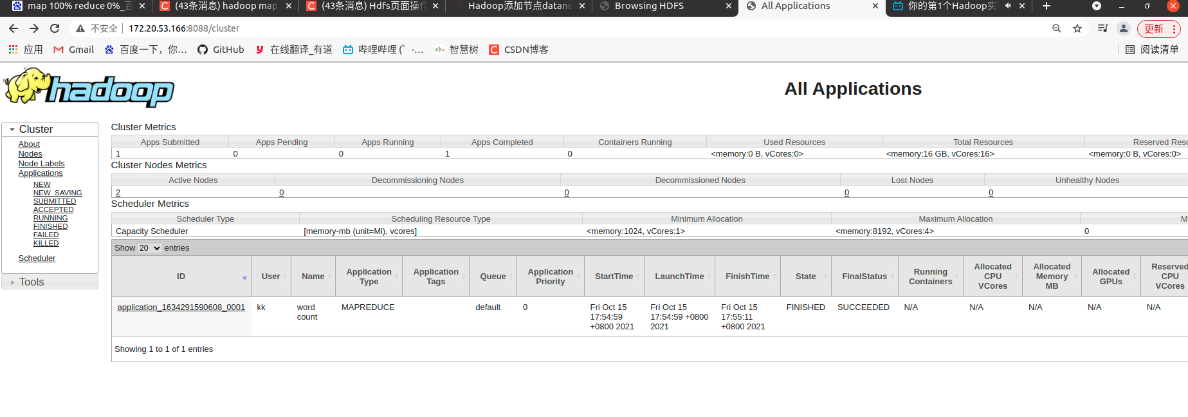


启动hadoop：在bin下start-all.sh，分别查看jps





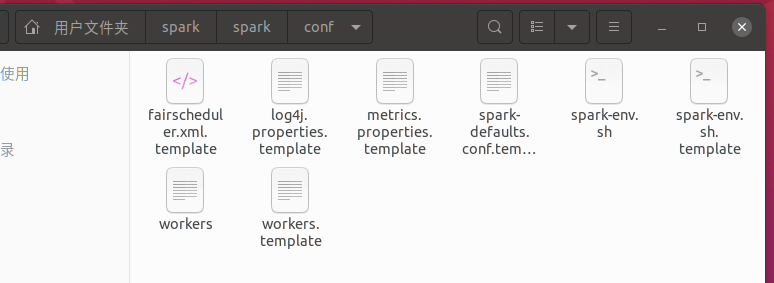
进入8088端口查看状态：



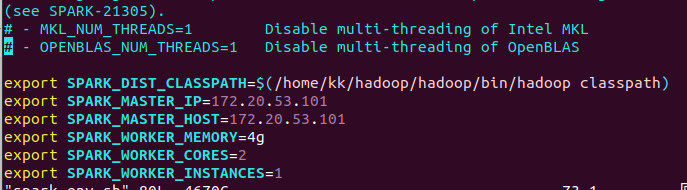
3.1.2 Spark集群搭建

①在~/spark目录下，解压已下载的spark压缩包。

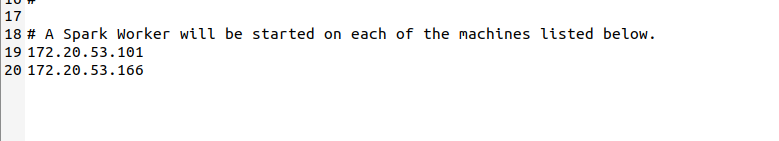
②使用命令cd spark/spark/conf，在该目录下，看到很多文件都是以template结尾的，这是因为spark给我们提供的是模板配置文件，我们可以先拷贝一份，然后将.template给去掉，变成真正的配置文件后再编辑。



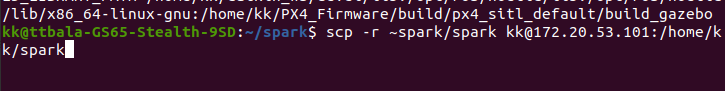
③配置spark-env.sh，该文件包含spark的各种运行环境

****

④配置workers文件

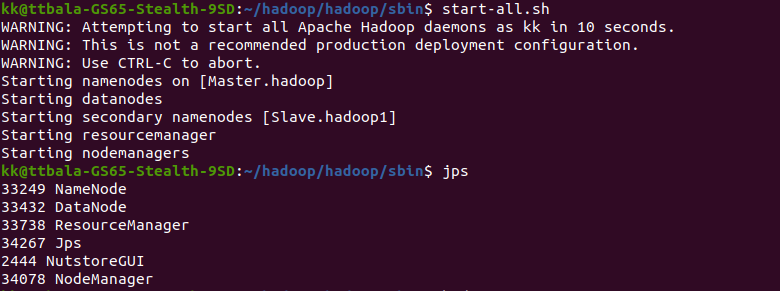


⑤将配置好的spark文件分发给集群中的其他电脑

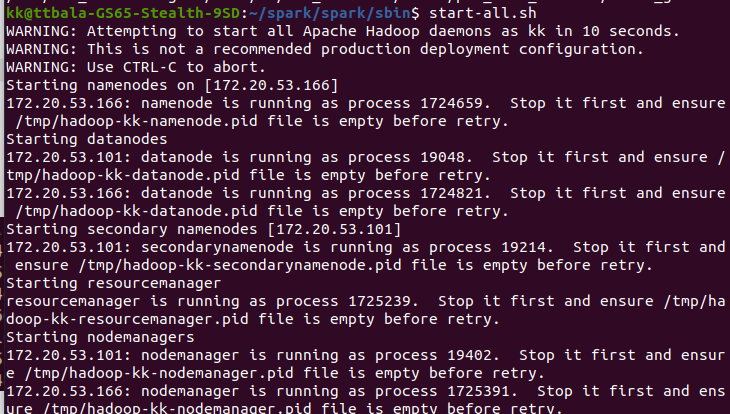


⑥启动Spark集群

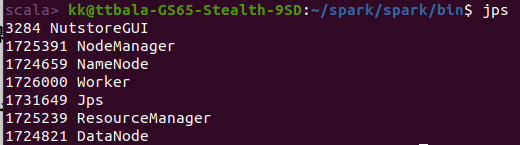
启动 hadoop 文件管理系统 HDFS以及启动 hadoop 任务管理器 YARN。



启动Spark集群：



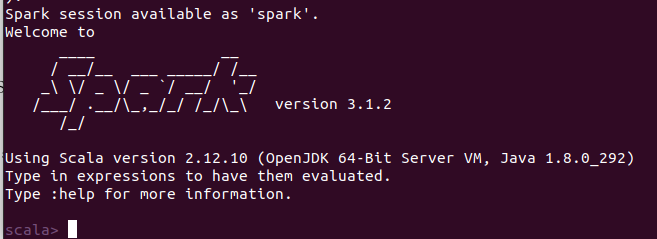
使用jps查看Spark集群信息：



进入Spark管理页面查看spark集群状态：



运行 spark-shell，可以进入 Spark 的 shell 控制台

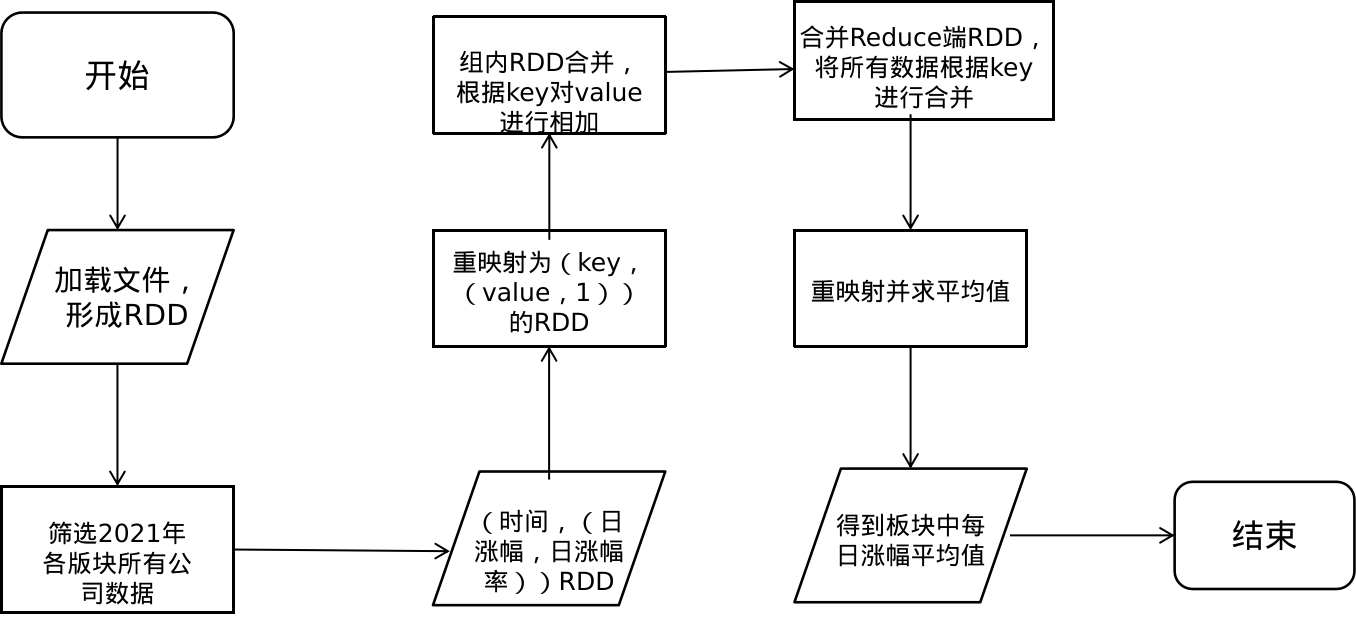


停止集群运行：停止集群时，运行sbin/stop-all.sh停止Spark集群，运行sbin/stop-dfs.sh来关闭hadoop 文件管理系统 HDFS，最后运行sbin/stop-yarn.sh来关闭hadoop 任务管理器 YARN。

3.2 股票数据预处理及可视化

3.2.1 统计各版块中股票的每日涨跌幅的数据清洗

为详细地了解每个版块所具有的增长潜力，我们需要统计每个版块的每日平均涨幅，涨跌的次数，故而需要进行数据清洗，该功能的具体实现过程如下：



首先我们会从HDFS文件系统中加载我们需要进行处理的文件，形成RDD，使得数据实现并行化。然后按照需求我们只会统计2021年的数据，所以我们设计了handleLine函数进行年份的筛选，返回以（时间，（日涨幅，日涨幅率））RDD，并重映射为（key，（value，1））的RDD以方便进行数据的统计。并且，在组内进行第一次合并，根据key对value值进行相加；最后进行在reduce端的组间大合并，把相同的key的数据拉取到一个节点上，然后按组合并。最后，我们进行重映射并求平均值，我们便得到了该板块的每日涨幅平均值、涨跌幅天数了。

其中关键的RDD设计体现如下：

rdd2 = rdd1.combineByKey(createCombiner, mergeValue, mergeGroup)

combineByKey算子中的三个lambda表达式均由我们根据需求进行设计，其中第一个以每一个组的第一个key开始，将其value进行初始化；第二个在同组内若是key已经出现过，便将value值按我们定义的方法进行合并；第三个reduce端组间大聚合，把相同的key的数据拉取到一个节点上，然后按组合并。

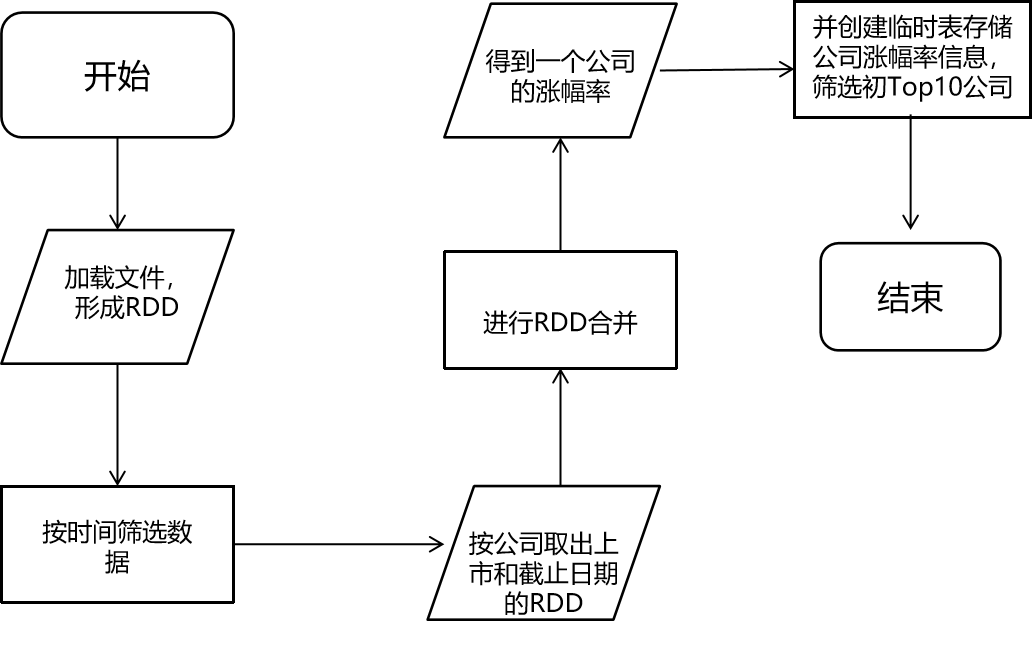
下面是我们的具体代码的实现：





3.2.2统计各板块中涨跌幅top10的股票的数据清洗

为详细地了解每个版块所中涨跌最明显的股票是那几只，我们需要列举出每个版块的涨跌top10股票，该功能的具体实现过程如下：



同样，我们会从HDFS文件系统中加载文件，形成RDD，使得数据实现并行化。然后按照需求进行数据的筛选，返回按公司取出上市和截止日期的RDD。并且，根据公司的key对value值进行相加，得到一个公司的涨跌率；最后创建临时表对该公司的涨跌率信息进行存储，直到最后所有的公司信息都清洗完，进行排序就可以得到top10的股票信息。

下面是我们代码的具体实现：



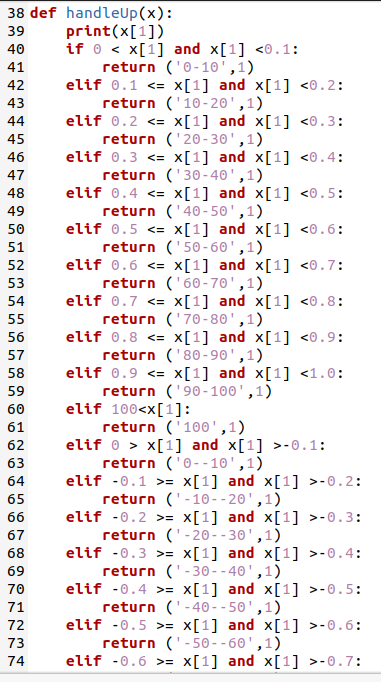
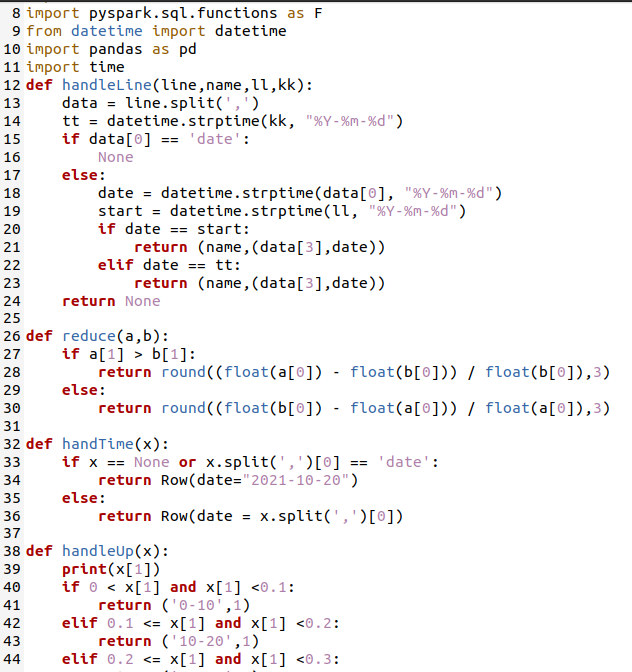
3.2.3统计各版块股票涨跌幅度占公司分布的数据清洗

为详细地了解每个版块各个涨跌比例公司的分布情况，我们需要例统计出涨跌率在－100%~100%中每个阶段的具体公司分布数量，该功能的具体实现过程如下：



同样，我们会从HDFS文件系统中加载文件，形成RDD，使得数据实现并行化。然后按照时间进行数据的筛选，按公司取出上市和截止日期的RDD。并且，根据公司的key对value值进行相加，得到每个公司的涨跌率，再根据涨跌幅重映射为（等级，1）的RDD，最后进行合并，我们便得到了该板块中个股票的涨幅率占比的统计信息，最后将相关信息存入数据库就可以了。

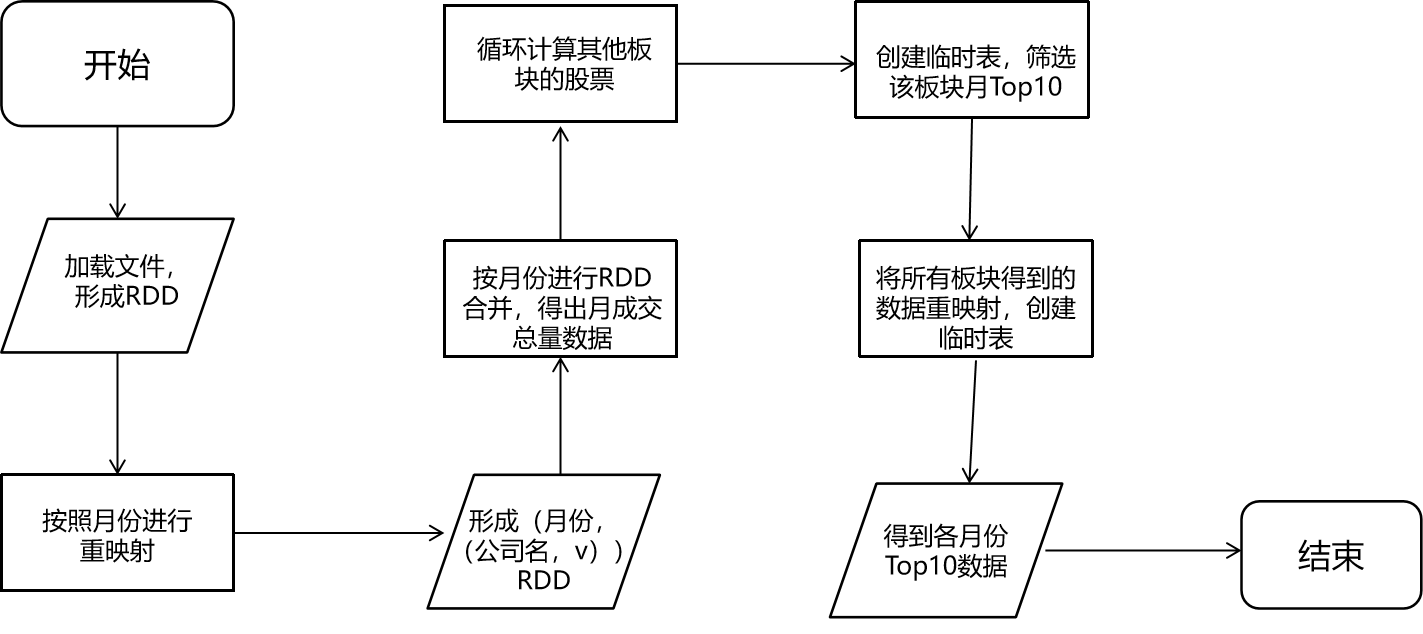
下面是我们代码的具体实现：





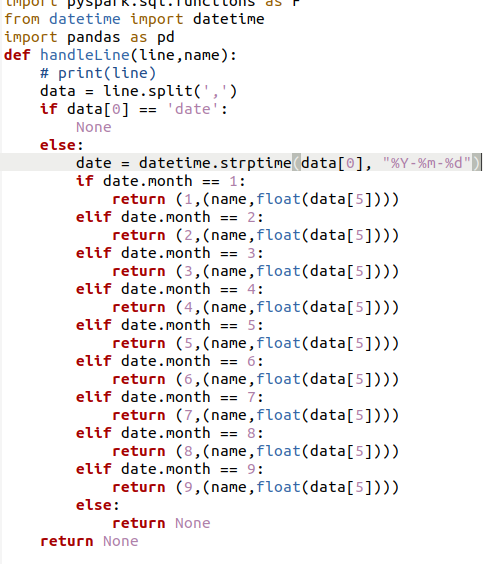
3.2.4统计各板块中月成交量top10的股票的数据清洗

为详细地了解每个版块哪几支股票交易得更为火爆，我们需要例统计出每个版块中交易量top10的股票，该功能的具体实现过程如下：



同样，我们会从HDFS文件系统中加载文件，形成RDD，使得数据实现并行化。然后按照月份对数据进行重映射，形成（月份，（公司名，value））的RDD。并且，进行合并相加，得到每个公司的月成交量数据，循环的计算出其他版块的所有股票信息，创建临时表进行存储，最后将数据进行重映射就可以得到各月份交易量top10的数据了。

下面是我们代码的具体实现：



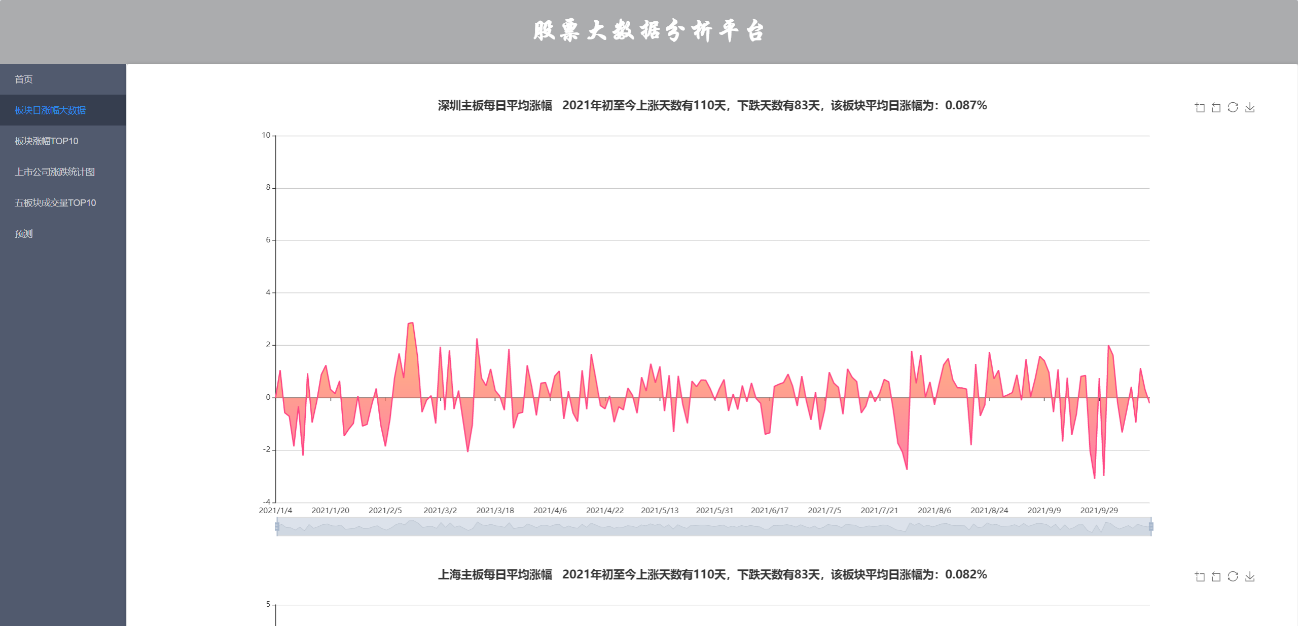


3.2.5页面展示

首先，首页展示了股票的基本信息，我们可以简单的查看最近一天所有股票的交易信息。



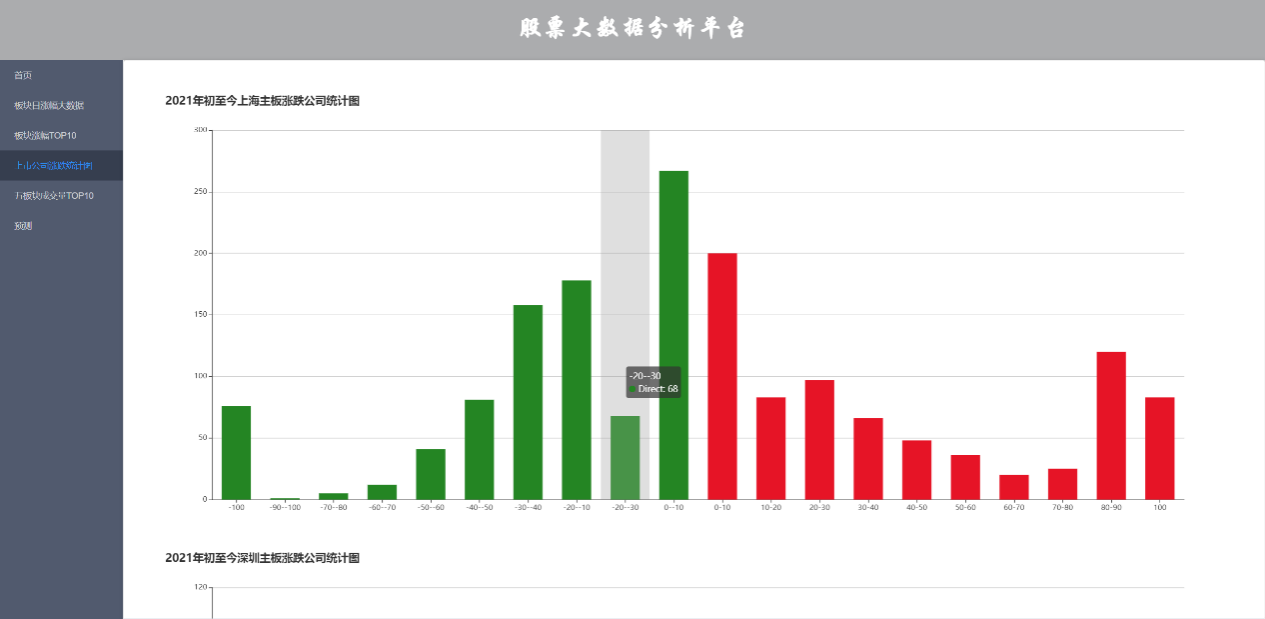
在第二个页面，统计了五个板块股票的每日涨跌幅，以及2021年中上涨的天数，下跌的天数和该板块平均日涨幅，这样我们就可以观察出哪个板块更有增长优势。此时展示的是上海主板的信息，下滑可以看到其他板块的信息。



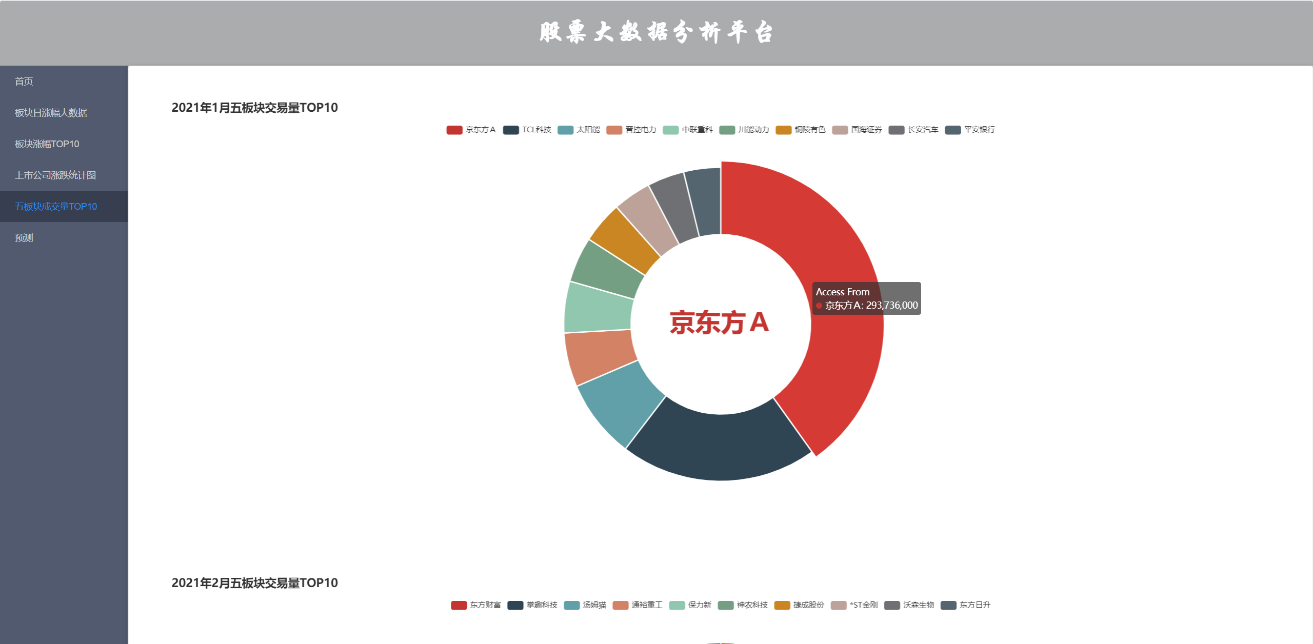
在第三个页面，统计各板块中涨跌幅top10的股票，红色的数据代表了该板块中涨幅top10的股票，绿色的数据代表了该板块中跌幅top10的股票，我们可以更直观地从图中看出每个板块中涨跌幅最显著的股票是哪几支。



在第四个页面，统计了各板块涨跌公司的分布，红色的数据代表了该板块中涨幅在这个范围内的股票有多少，绿色的数据代表了该板块中跌幅在这个范围内的公司有多少，我们可以更直观地从图中看出各板块的涨跌幅公司个数的分布。



在第五个页面，统计各板块中月成交量top10的股票：可以更直观地看出该板块中哪几支股票交易最为火爆。

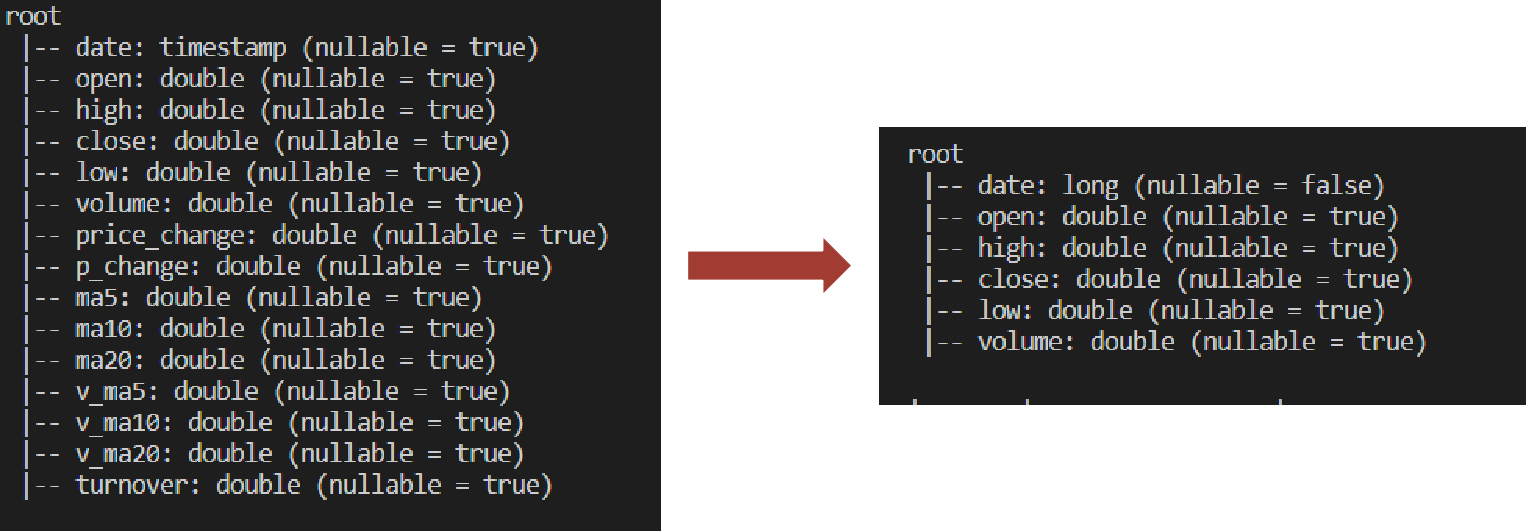


3.3 股票走势预测

3.3.1 预测算法

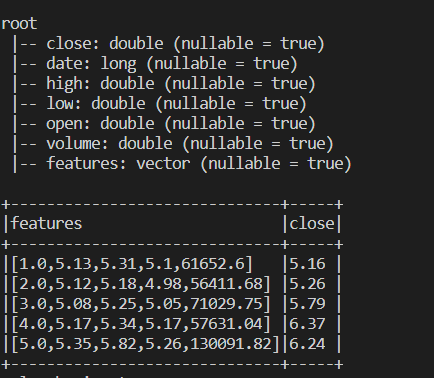
在我们的项目中，我们需要预测的目标是：未来10天内股票的close（闭盘价），根据Spark的分布式机器学习库中为我们提供的众多算法，我们选择了建模迅速，对简单的关系很有效，有利于决策分析的线性回归算法进行预测。

首先，我们根据需要对股票数据进行简单的处理，我们只需要右边的参数进行预测。

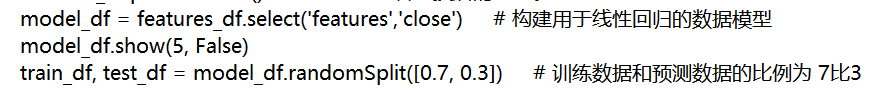


其次，根据多元线性回归方程的表达式，我们构建了如下的特征向量矩阵：

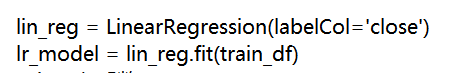
gif



其中每个特征向量为：[date,open,high,low,volume]，close为我们的预测值。同时，我们将数据集按7：3的比例划分为训练集和测试集：

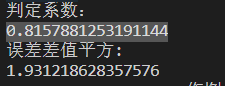


对训练集进行的训练，我们可以得出如下的相关系数：

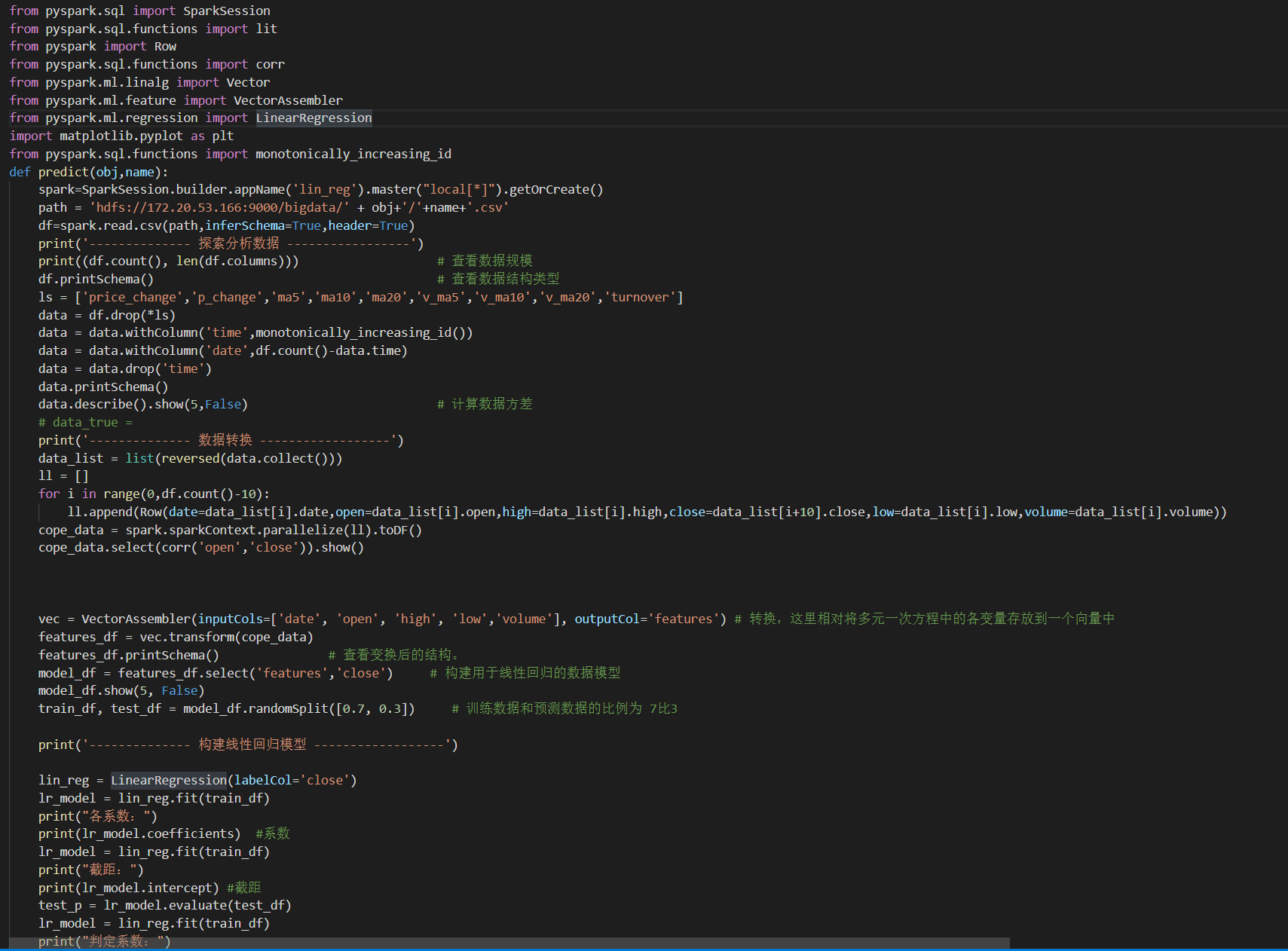


09[OKU4ARN{ZD$WDE`F]N{3

用训练好的模型来预测测试集中的数据，可以得出一些评价指标：

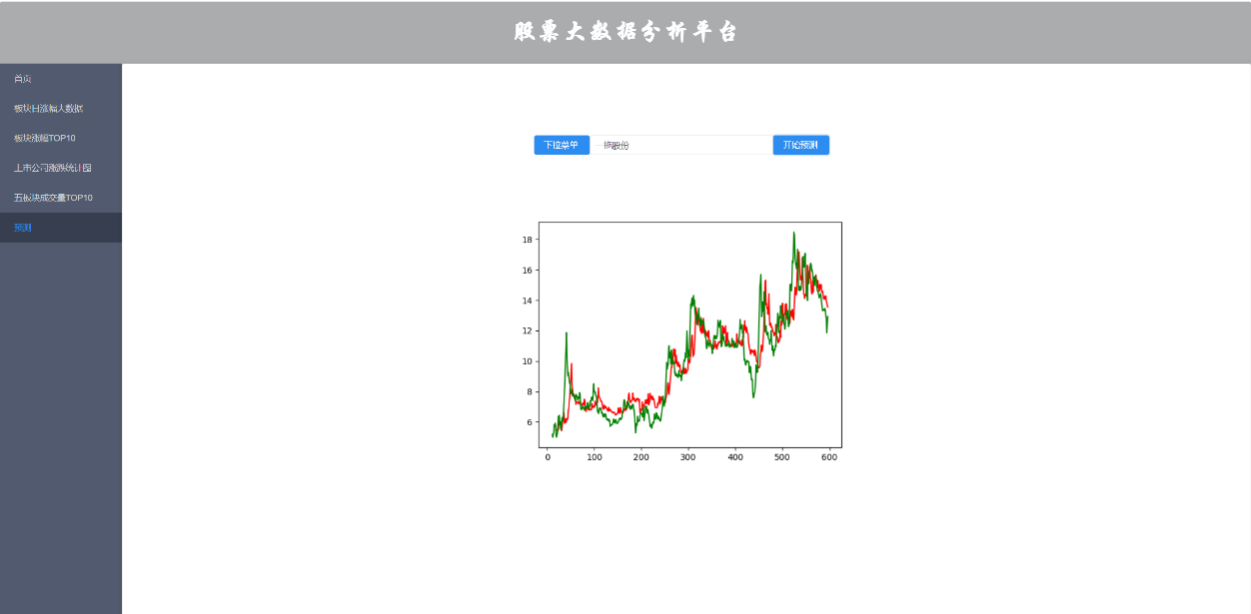


下面是我们的具体代码设计：



3.3.2 预测结果

根据用户填入的参数，向后端发起请求，后端调用汉书将这个任务提交到spark集群中进行模型的训练和预测，然后将预测结果返回给前端。



## 4 小结

本案例选取中国股票大数据分析平台的构建为主题，使用源于财经网站的股票数据作为数据集，结合数据挖掘、机器学习等相关方法进行股票数据可视化及走势预测。本案例核心内容有：（1）使用Hadoop 和Spark搭建服务平台；（2）股票数据预处理及可视化；（3）使用机器学习算法进行股票走势预测。该案例主题新颖，结合了工业界现实需求与大数据分析与挖掘的多种理论与技术，可以充分增强学生的实践能力与理论基础。另外，本案例的内容仅为指导性的过程，在实际教学中，可保持基本研究内容不变，鼓励学生引入其它的数据预处理、数据挖掘、机器学习方法完成任务。

## 附录

1. 本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及到数据预处理以及多种机器学习算法，建议使用python语言进行编写，推荐的工具包有Tushare（财经数据接口包）、scikit-learn（机器学习库）。

3. 本案例参考文献如下：

[1] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.

[2] 陈文颖. 基于Hadoop的股票数据聚类分析[D].浙江大学,2018.