Spark课程实验案例：Spark+Kafka构建实时分析Dashboard（2022年9月V2.0版本）

林子雨 2022-09-20 (updated: 2022-09-21) 22868

中国高校大数据课程
公 共 服 务 平 台

免费为全国高校提供大数据教学案例

建设者：厦门大学计算机科学系 林子雨 博士/副教授

(E-mail: ziyulin@xmu.edu.cn, 个人主页：https://dblab.xmu.edu.cn/linziyu)

（版权声明：本平台所有资源有版权，请勿用于商业用途）

（未经授权，其他网站请勿转载）

扫一扫手机访问本主页

案例简介 | 案例目的 | 软件工具 | 案例任务 | 实验步骤 | 开发团队 | 版本历史 | 联系人

相关其他案例推荐：Spark课程综合实验案例：淘宝双11数据分析与预测



点击这里观看厦门大学林子雨老师主讲《大数据技术原理与应用》课程视频

教师风采



林文水

最新内容 更多

林子雨老师应邀为水利部小浪底...
林子雨老师应邀为福建农信系统...
林子雨应邀为江西省程序设计类...
林子雨老师应邀为龙岩烟草员工...
实验室拍摄2025届毕业生合影
林子雨老师应邀为瑞安科级干部...
实验室举办大数据百家讲坛第13...
林子雨老师应邀为泉州经贸学院...
林子雨老师应邀为广西中烟员工...
林子雨编著Python系列教材（4...

热门内容

《大数据技术原理与应用（第
大数据学习路线图
林子雨编著《大数据技术原理与...
林子雨编著《大数据技术原理与...
林子雨
重磅：大数据课程实验案例：网...
林子雨编著《大数据技术原理与...
Spark课程综合实验案例：淘宝...
林子雨主讲入门级大数据在线课程
《Spark编程基础(Scala版)》教...

案例简介

大数据课程实验案例：Spark+Kafka构建实时分析Dashboard案例，由厦门大学数据库实验室团队开发，旨在满足全国高校大数据教学对实验案例的迫切需求。本案例涉及数据预处理、消息队列发送和接收消息、数据实时处理、数据实时推送和实时展示等数据处理全流程所涉及的各种典型操作，涵盖Linux、Spark、Kafka、Flask、Flask-SocketIO、Highcharts.js、sockert.io.js、PyCharm等系统和软件的安装和使用方法。案例适合高校（高职）大数据教学，可以作为学生学习大数据课程后的综合实践案例。通过本案例，将有助于学生综合运用大数据课程知识以及各种工具软件，实现数据全流程操作。各个高校可以根据自己教学实际需求，对本案例进行补充完善。

案例目的

1. 熟悉Linux系统、Spark、Kafka、Flask、Flask-SocketIO、Highcharts.js、socket.io.js等系统和软件的安装和使用；
2. 了解Spark+Kafka实时处理数据的基本流程；
3. 熟悉Python操作Kafka的方法；
4. 熟悉Structured Streaming程序编写；
5. 熟悉Spark操作Kafka的方法；
6. 熟悉Python构建Web程序；
7. 熟悉SocketIO实时推送消息；
8. 熟悉Highcharts.js展示数据。

适用对象

1. 高校（高职）教师、学生
2. 大数据学习者

时间安排

本案例可以作为《大数据处理技术Spark》课程在学期结束后的“大作业”，或者可以作为学生暑期或寒假大数据实习实践基础案例，完成本案例预计耗时7天。

预备知识

需要案例使用者，已经学习过大数据相关课程（比如入门级课程《大数据技术原理与应用》），了解Spark Streaming的基本概念与原理、Linux操作系统、Kafka消息队列发送和接收消息操作、Python构建Web应用，Socket.IO框架的应用以及前端JavaScript基础用法。

不过，由于本案例提供了全部操作细节，包括每个命令和运行结果，所以，即使没有相关背景知识，也可以按照操作说明顺利完成全部实验。

硬件要求

本案例可以在单机上完成，也可以在集群环境下完成。

单机上完成本案例实验时，建议计算机硬件配置为：500GB以上硬盘，8GB以上内存。

软件工具

本案例所涉及的系统及软件：

1. Linux系统
2. Spark
3. Kafka
4. Flask
5. Flask-SocketIO
6. Highcharts.js
7. Socket.io.js
8. PyCharm

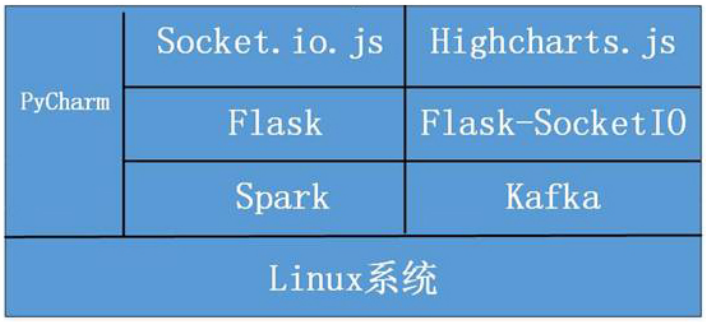


图 案例所涉及软件总体概览图

数据集

淘宝购物行为数据集 (5000万条记录，数据有偏移，不是真实的淘宝购物交易数据，但是不影响学习)

案例任务

1. 安装Linux操作系统
2. 安装Spark
3. 安装Kafka
4. 安装Python依赖库
5. 安装PyCharm
6. 对文本文件形式的原始数据集进行预处理
7. 将预处理后的数据发送至Kafka
8. Spark从Kafka获取数据，实时处理，结果发送至Kafka
9. Flask构建的Web程序从Kafka获取处理后的数据
10. Flask-SockerIO实时推送数据至客户端
11. 客户端Socket.io.js实时获取数据
12. 客户端Highcharts.js实时展示数据



图 案例所涉及操作总体概览图

实验步骤

步骤零：案例介绍	查看实验指南
步骤一：实验环境准备	查看实验指南
步骤二：数据处理和Python操作Kafka	查看实验指南
步骤三：Structured Streaming实时处理数据	查看Scala版实验指南 查看Python版实验指南
步骤四：结果展示	查看实验指南

每个实验步骤所需要的知识储备、训练技能和任务清单如下：

步骤零：案例介绍

所需知识储备	无
训练技能	了解本案例整体框架和所用到的工具
任务清单	无

步骤一：实验环境准备

所需知识储备	Linux系统命令使用、了解如何安装Python库
训练技能	熟悉Linux基本操作、Pycharm的安装、Spark安装，Kafka安装，PyCharm安装
任务清单	1. 安装Linux系统；2. Spark安装；3.Kafka安装；4. Python安装；5. Python依赖库；6. PyCharm安装

步骤二：数据处理和Python操作Kafka

所需知识储备	简单使用Python，了解Kafka的使用
训练技能	Python基本使用，Python操作Kafka代码库kafka-python使用
任务清单	1. 利用Python预处理数据；2. Python操作Kafka

步骤三：Spark Streaming实时处理数据

所需知识储备	使用Scala编写Spark Streaming程序，Kafka原理
训练技能	编写Spark Streaming程序，熟悉Spark操作Kafka
任务清单	1. Spark Streaming实时处理Kafka数据；2. 将处理后的结果发送给Kafka；

步骤四：结果展示

所需知识储备	了解Flask创建web程序，了解如何在html编写js代码
训练技能	利用Flask创建web程序，利用Flask-SocketIO实现实时推送数据，利用socket.io.js实现实时接收数据，hightlights.js展现数据
任务清单	1. 利用Flask-SocketIO实时推送数据；2. socket.io.js实时获取数据；3. highlights.js展示数据

开发团队

为了解决高校大数据教学需要综合实验案例的迫切需求，2017年2月，厦门大学数据库实验室组建了由林子雨老师和罗道文（厦大数据库实验室2014级研究生）、阮榕城（厦大数据库实验室2015级研究生）、薛倩（厦大数据库实验室2015级研究生）、魏亮（厦大数据库实验室2016级研究生）、曾冠华（厦大数据库实验室2016级研究生）同学组成的案例开发小组，通过大量调研学习网络资料和相关案例，开发了本教学案例。期间，多次举行小组会议，讨论案例制作思路和技术细节。最终，经过近1个月的团队努力，于2017年4月23日顺利完成案例第1版的开发并上线发布。2022年9月，由刘浩然同学（厦大数据库实验室2022级研究生）对案例进行升级改版，生成V2.0版本。



本案例主创人员 罗道文



图（摄影 夏小云） 大数据案例开发团队合影

（人员从左到右名单：魏亮、阮榕城、林子雨、薛倩、曾冠华）

版本历史

2017年4月23日，发布案例V1.0版本。

2022年9月，发布案例V2.0版本。

联系人

本案例相关事宜，欢迎联系厦门大学数据库实验室林子雨老师：E-mail: ziyulin@xmu.edu.cn

< 实验室欢度2022年教师节 林子雨老师负责的项目获得“2022年福建省... >

© 2014 厦门大学数据库实验室, All Rights Reserved.