厦门大学计算机科学系研究生课程

大数据处理技术 Spark

（2016-2017 学年春季学期）

期末作业说明

主讲教师：林子雨
授课地点：厦门大学海韵教学楼 107 教室
二零一七年五月一日

目录

大数据处理技术 Spark

2016-2017 学年春季学期期末作业说明

主讲教师：林子雨　ziyulin@xmu.edu.cn

一、 作业题目

基于 Spark 的数据处理与分析

二、 作业目的

综合运用大数据处理技术 Spark、Hadoop 及数据可视化，对数据进行存储、处理和分析。

三、 作业性质

必做。作为评定期末总成绩的重要参考。

四、 作业考核方法

根据自己能力，任意选择基础作业或者高级作业两者中的一种，作业成绩评定方法如下：

- 不按时交作业、所提交的作业无法打开或抄袭他人作业：零分
- 基础作业评分范围：0-60 分
- 高级作业评分范围：0-100 分

温馨提示：作业必须自己独立完成，不得抄袭他人作业，否则，期末总成绩不及格。

五、 提交日期与方式

1、必须于 2017 年 5 月 26 日（厦门大学教学周第 15 周周五）晚 24 时之前提交；

2、提交的内容为压缩文件 RAR 文件，里面所包含的 WORD 文档，一定要转换成 WORD2003 格式，即.doc 格式，最后把压缩包文件发送到班级助教阮榕城同学邮箱：ruanrc@qq.com；如果是团队完成作业，只需要组长负责提交作业；如果是个人独立完成作业，则只需要自己负责提交作业；

3、文件名命名为“姓名学号.rar”，例如“王小明 23020091152890.rar”；

4、文件夹中应该包含实验报告 WORD 文档、小组成员名单 TXT 文件（包含学号、姓名、手机、E-mail，请标明组长）、软件版本号 TXT 文件（包含作业中用到的所有软件和编程语言的名称和版本号信息）以及其他有必要提交的文档。

六、 作业工具和环境要求

（1） 必须在 Linux 系统下完成作业。

（2） 可以任意选择自己喜欢的开发工具，比如 Eclipse、IntelliJ IDEA 等。

（3） 相关软件的版本要求如下：

- Linux: Ubuntu16.04
- MySQL: 5.7.16
- Hadoop: 2.7.1
- HBase:1.1.5
- Hive:1.2.1
- Sqoop:1.4.6
- R:3.2.3
- Eclipse:3.8
- Spark2.1.0

注意，如果同学使用了其他软件，请一定在软件版本号 TXT 文件中明确列出。

七、 作业内容和要求

（一）基础作业

基础作业是基础等级的作业，只要具备 Spark 课程基础知识即可顺利完成作业内容。作业具体要求如下：

（1）1 个人独立完成作业。

（2）根据本学期的 Spark 课程内容，设计完成三个编程题目，每个题目类似于课程讲义 PPT“第 5 章 Spark 编程基础”的“5.6 综合案例”，设定一个目标任务（比如求 TOP N 个值），根据若干个输入文件，编写 Spark 应用程序，对数据进行处理，实现目标任务。数据必须存放在分布式文件系统 HDFS 中（可以采用伪分布式），不能使用本地文件。涉及的编程知识可以涉及 RDD、Spark SQL、Spark Streaming 和 Spark MLlib 中的任意一个或多个。

（3）必须撰写实验报告 WORD 文档，里面包含实验题目、实验代码、实验过程说明（包含必要的屏幕截图）。实验过程描述，没有固定的格式，但是，必须要能够让老师在批改作业时，根据同学的实验过程说明能够在老师的计算机上顺利运行实验代码。可以参考《Spark 入门教程》的写作风格。

（4）可以借鉴网络代码，但是，不能直接复制粘贴，必须经过自己理解消化后写成自己的代码。

（5）不能直接使用课堂讲义 PPT 中的实例，也不能直接使用厦门大学数据库实验室网站的林子雨编著《Spark 入门教程》的实例。

（6）不同同学之间的作业不能雷同，如果批改作业过程发现作业明显雷同，需要同学给出合理解释，如果无法解释，则雷同双方作业都判定为零分。

（二）高级作业

高级作业是具有一定难度的作业，不仅需要具备 Spark 课程基础知识，还需要具备数据预处理、数据可视化、数据存储等多方面的知识。作业具体要求如下：

（1）可以 1 个同学独立完成作业，也可以 1 到 3 个同学组队完成作业。

（2）请阅读厦门大学数据库实验室建设的大数据课程公共服务平台上的《Spark 课程实验案例：Spark+Kafka 构建实时分析 Dashboard》（地址：
http://dblab.xmu.edu.cn/post/8274/）。

（3）利用上面的参考案例，设计一个新的实时分析 Dashboard 案例。可以采用参考案例中的数据集，或者也可以自己寻找数据集。

（4）新的案例，需要像参考案例一样，包含数据数据预处理、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。

（5）如果同学作业完全采用参考案例中的相同组件（Linux、Spark、Kafka、Flask、Flask-SocketIO、Highcharts.js、socket.io.js），由于参考案例采用的是 Python 编程，则需要同学自己设计的案例不能采用 Python 语言，而是需要采用 Java 语言或 Scala 语言，或者其他编程语言，也就是说，用其他编程语言去实现参考案例，也可以作为作业提交。如果同学作业采用了与参考案例不相同的组件，比如，把 Kafka、Flask、Flask-SocketIO、Highcharts.js、socket.io.js 中的一部分或全部组件用其他软件工具去实现，则也可以采用 Python、Java 等各种编程语言。总之，同学不能把参考案例原方不动地简单复制一遍。

（6）必须撰写实验报告 WORD 文档，里面包含实验题目、实验代码、实验过程说明（包含

必要的屏幕截图）。实验过程描述，没有固定的格式，但是，必须要能够让老师在批改作业时，根据同学的实验过程说明能够在老师的计算机上顺利运行实验代码。可以参考《Spark 课程实验案例：Spark+Kafka 构建实时分析 Dashboard》的写作风格。

（7）可以借鉴网络代码，但是，不能直接复制粘贴，必须经过自己理解消化后写成自己的代码。

（8）不能直接使用课堂讲义 PPT 中的实例，也不能直接使用厦门大学数据库实验室网站的林子雨编著《Spark 入门教程》的实例。

（9）不同小组之间的作业不能雷同，如果批改作业过程发现作业明显雷同，需要同学给出合理解释，如果无法解释，则雷同双方作业都判定为零分。

八、 参考资料

（1） 厦门大学林子雨编著《Spark 入门教程》http://dblab.xmu.edu.cn/blog/spark/

（2） 厦门大学数据库实验室制作《Spark 课程实验案例：Spark+Kafka 构建实时分析 Dashboard》（地址：http://dblab.xmu.edu.cn/post/8274/）。

（3） 厦门大学林子雨主讲《大数据处理技术 Spark》2017 班级主页，里面包含讲义 PPT。班级主页地址：http://dblab.xmu.edu.cn/post/7659/。

（4） 厦门大学林子雨主讲《大数据技术原理与应用》在线课程视频 http://dblab.xmu.edu.cn/post/bigdata-online-course/

（5） 厦门大学数据库实验室编写《大数据软件安装和编程实践指南》，详细介绍如何安装运行各种大数据软件以及如何进行初级编程实践，包括 Hadoop、HDFS、HBase、MapReduce、Spark、MongoDB 等安装、操作、编程指南。在线访问网址：http://dblab.xmu.edu.cn/post/5663/。

九、 附录 1:教师介绍

林子雨(1978－),男,博士,厦门大学计算机科学系 助理教授,主要研究领域为数据库,数据仓库,数据挖掘,大数据
主讲课程：大数据处理技术
办公地点：厦门大学海韵园科研 2 号楼
E-mail: ziyulin@xmu.edu.cn

林子雨，男，1978 年出生，博士（毕业于北京大学），现为厦门大学计算机科学系助理教授（讲师），曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员，荣获“2016 中国大数据创新百人”称号。中国高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013 年度厦门大学奖教金获得者。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括 1 项国家自然科学青年基金项目(No.61303004)、1 项福建省自然科学青年基金项目(No.2013J05099)和 1 项中央高校基本科研业务费项目(No.2011121049)；作为课题负责人主持的教学项目包括 1 项福建省教改课题和 1 项教育部产学合作育人项目。同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015 泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009 年至今，“数字教师”大平台累计向网络免费发布超过 100 万字高价值的研究和教学资料，累计网络访问量超过 100 万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过 50 万次。具有丰富的政府和企业信息化培训经验，厦门大学管理学院 EDP 中心、浙江大学管理学院 EDP 中心、厦门大学继续教育学院、泉州市科技培训中心特邀培训讲师，曾给中国移动通信集团公司、福州马尾区政府、福建龙岩卷烟厂、福建省物联网科学研究院、石狮市物流协会、厦门市物流协会、浙江省中小企业家、四川泸州企业家、江苏沛县企业家等开展信息化培训，累计培训人数达 3000 人以上。