用户收藏商品次数统计案例

一、基本信息

课程名称：大数据技术概论

课程类型：☐通识教育课　　☐公共基础课　　☑专业课

☐创新创业课程　☐实验课

开课年级：大三下学期

面向专业：计算机类

教学章节：3.4

授课学时：　32

主讲教师：李乔

授课形式：在线

选用平台：泛雅超星平台

课程链接：https://mooc1.chaoxing.com/course/204854307.html

二、案例背景

课程性质：（1）能够建立对大数据知识体系的认识，了解大数据发展历程、基本概念、主要影响、应用领域、关键技术、计算模式和产业发展，并了解云计算、物联网的概念及其与大数据之间的紧密关系；（2）掌握Hadoop分布式文件系统HDFS的重要概念、体系结构、存储原理和读写过程，并熟练掌握分布式文件系统HDFS的使用方法；（3）通过对实验结果、数据分析的可视化操作，培养学生的运用课程知识能力、团队合作能力，加深学生对课程知识的理解和掌握；（4）初步建立数据分析的设计能力，为将来从事大数据维护、开发、分析

打下坚实基础。

课程标准：大数据技术入门课程，为学生搭建起通向大数据知识的桥梁和纽带，构建知识体系、阐明基本原理、引导初级实践、了解相关应用为原则，为学生在大数据领域奠定基础。

教学内容体系:讲授大数据的基本概念、大数据处理架构 Hadoop、分布式文件系统 HDFS、分布式数据库 HBase、NoSQL 数据库、云数据库、分布式并行编程模型 MapReduce、基于内存的大数据处理架构 Spark、大数据在互联网、生物医学和物流等各个领域的应用。在 Hadoop、HDFS、HBase、MapReduce、Spark 等重要章节，让学生更好地学习和掌握大数据关键技术。

学生特点：基础、年级不同，线上教学容易懈怠

教学条件：泛雅超星教学平台、网站资源下载、在线实验平台

三、案例设计思路

教学内容：现有某电商网站用户对商品的收藏数据，记录了用户收藏的商品 id 以及收藏日期，名为 buyer_favorite1。buyer_favorite1 包含：买家 id，商品 id，收藏日期这三个字段，数据以“\t”分割，样本数据及格式如下：

买家 id	商品 id	收藏日期
10181	1000481	2010-04-04 16:54:31
20001	1001597	2010-04-07 15:07:52
20001	1001560	2010-04-07 15:08:27
20042	1001368	2010-04-08 08:20:30

…

拟解决的主要问题:利用 Mapreduce 编程模型分而治之

要求统计结果数据如下:

买家 id 商品数量

10181　1

20001　2

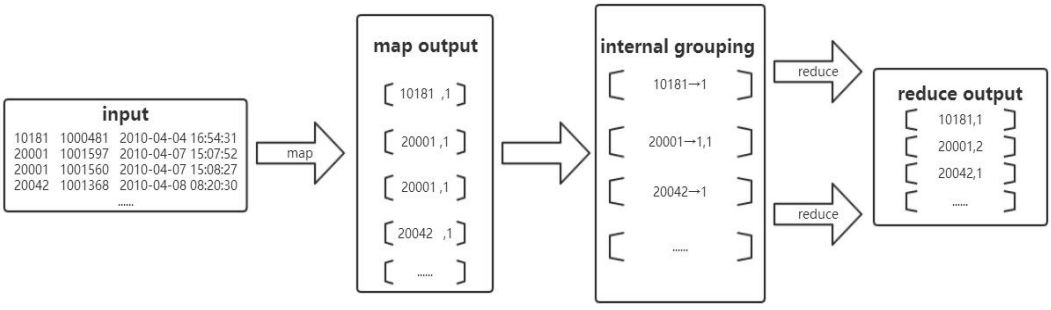
…

编程思路:

（1）map 阶段采用 Hadoop 的默认的作业输入方式，把输入的 value 用 StringTokenizer()方法截取出的买家 id 字段设置为 key，设置 value 为 1,然后直接输出<key,value>。map 输出的<key,value>先要经过 shuffle 过程把相同 key 值的所有 value 聚集起来形成<key,values>后交给 reduce 端。

（2）reduce 端接收到<key,values>之后，将输入的 key 直接复制给输出的 key,用 for 循环遍历 values 并求和，求和结果就是 key 值代表的单词出现的总次,将其设置为 value,直接输出<key,value>。

数据格式转换过程:



教学方法：利用在线平台进行讲授、阅读、视频演示、讨论、作

业实践、反馈等多维度教学

教学载体：自建虚拟机实验平台和在线实验平台相结合

教学目标： （1）准确理解 MapReduce 的设计原理 （2）学会自己编写程序进行商品数量统计（3）学习 Hadoop 运行机制和过程

四、教学目标

1.知识与能力目标

MapReduce 采用的是“分而治之”的思想，把对大规模数据集的操作，分发给一个主节点管理下的各个从节点共同完成，然后通过整合各个节点的中间结果，得到最终结果。

通过综合案例程序设计培养学生的运用课程知识能力、团队合作能力，加深学生对课程知识的理解和掌握，初步在大数据平台下的掌握对数据的清洗、建模、分析、可视化的能力，并且具备一定的算法优化与设计能力。

2.育人目标

培养学生在大数据复杂平台构建与管理、数据采集与预处理、数据分析与挖掘、数据可视化等方面的能力，提高学生的创新创业能力与就业水平。让学生能够掌握数据采集与预处理，了解大数据的存储和管理,学习应用数据处理与分析,以可视化形式展现数据分析结果,培养一批具有大数据专业素养的高级人才,满足社会对大数据人才日益旺盛的需求。

五、教学过程

教学过程:

（1）切换目录到/apps/hadoop/sbin 下，启动 hadoop。

（2）在 linux 上，创建一个目录/data/mapreduce1。

（3）切换到/data/mapreduce1 目录下，使用 wget 命令从网址下载文本文件 buyer_favorite1 作为数据源。

（4）将 linux 本地上的数据源上传到 HDFS 上的 in 目录下。若 HDFS 目录不存在，需提前创建。

（5）打开 Eclipse，新建 Java Project 项目。

（6）在项目名 mapreduce1 下，新建 package 包。

（7）在创建的包 mapreduce 下，新建类.

（8）添加项目所需依赖的 jar 包，右键单击项目名，新建一个目录 hadoop2lib，用于存放项目所需的 jar 包。

（9）编写 Java 代码，并描述其设计思路。

（10)在 WordCount 类文件中,单击右键=>Run As=>Run on Hadoop 选项，将 MapReduce 任务提交到 Hadoop 中。

（11）待执行完毕后，打开终端或使用 hadoop eclipse 插件，查看 hdfs 上，程序输出的实验结果。

教学环节：教学前期指导学生预习哪些章节的书本内容,视频讲解过程中分析理论背景知识，上课期间进行签到打卡，查看学生观看视频情况和停留时间，全程指导上机实践过程，批改在线作业，了解学生反馈。QQ 群线上实时交流，讨论问题解决方案。

教学思考：面对特殊时期的教学要求，充分利用网络教学平台数据统计的优势，动态监控学生学习情况，用数据说话，及时和辅导员交流学生学习情况；通过在线平台实验和自己搭建实验平台的不同技术方案，满足不同层次学生特点需求，补足不能面授带来的教学过程短板；用 QQ 通讯工具建群，有效解决交互实时性问题；以实际问题为导向，把 WordCount 分布式编程实例带入解决用户购物行为分析，促进学生学习的动力；在线教学的节奏感明显要比面授要快，这就需要及时得到反馈，进行实时讨论，提升解决问题能力；在线教学优势可以让学生当遇到问题后，重新反复观看教学视频，有利于实现个性化教学的目的，也有利于学生理论联系实际；通过摸索，线上教学以分配任务点方式教学有利于驱动学生完成，每次任务点讲解视频录制时间不宜太长。

六、教学效果与特色创新

特色创新：

（1）提供多种搭建实验环境方案，满足不同场景需求

（2）分析在线平台学习数据，灵活督促学生学习

（3）分层次教学，提高学生完成度和成就感

（4）把知识点赋予真实案例中，提高学习兴趣

（5）多种教学手段组合使用，促进多维度教学

教学效果：出勤率达到 95%，作业完成率 95%，视频学习率 95%

七、教学反思

上面特色创新中，在（1）中，通过下载教学网站资源自建实验平台、导入配置好虚拟机镜像系统、使用联想在线实验平台三种方案解决大数据技术实验环境问题，有利于满足不同学生需求；对于（2）利用泛雅超星平台上传教学视频、分析学生观看时长和打卡签到情况，以作业批改把握学生知识点掌握情况，通过教学平台点评和 QQ 进行指导、交流，有利于提升学习效率；对于（3）学生基础有所不同，部分同学前期学习有畏难情绪，学习有所懈怠。因此，布置作业注重层次性，尽量让每位同学都有完成度和收获；在（4）中通过案例的需求，引导学生把学习过的编程模型应用到实际场景中；在（5）中以讲授、阅读、视频演示、讨论、作业实践等多种形式较好的解决了学生学习过程中有可能的懈怠，基础薄弱，层次性问题。

特殊时期，基于本课程特点，强调本课程在未来祖国新基建和强国战略中的重要作用，引导学生树立正确人生目标、建立职业规范，极大的激发了学生的学习热情。

八、教学资源

教学资源：

（1）教材网站资源[EB/OL]. http://dblab.xmu.edu.cn/post/bigdata/

（2）在线实验平台[EB/OL]. http://www.youxuanit.com

（3）林子雨. 大数据技术原理与应用[M]. 北京: 人民邮电出版社.

（4）林子雨.大数据基础编程、实验和案例教材[M].清华大学出版社.

软件工程专业学生（谢大智）实践及作业部分内容反馈：

（1）在进行原来不符合条件但能运行的代码之后，导致之后正确的代码运行会产生如下的错误

```
Output directory hdfs://localhost:9000/mymapreduce1/out already exists
```

解决方案：将原来的该语句修改成：Path out = new Path("hdfs://localhost:9000/mymapreduce1/out0");

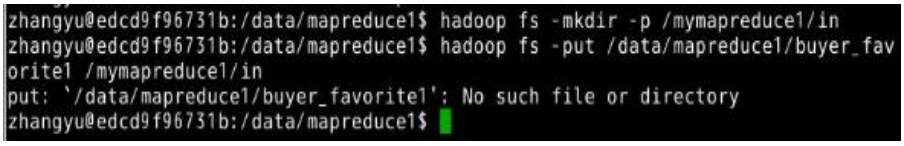
（2）如果用在线实验平台的代码来直接运行，将不符合该实验的要求，会产生如下错误的结果：

解决方案：查询相关资料对在线实验平台给出的代码进行了部分的修改。具体代码见附录。

（3）在线平台左边文档的代码直接复制到实验机的终端（给出的代码只能通过特定按钮才能复制到实验机且在实验外不可修改），不用回车键就会自动运行，导致还未来的及修改其中代码就产生了错误的结果，例如如下的代码。

```
hadoop fs -mkdir -p /mymapreduce1/in
hadoop fs -
put /data/mapreduce1/buyer_favorite1 /mymapreduce1/in
```

没按回车键自动运行



解决方案：

按上键找到上一步命令，进行部分修改再运行即可（将不存在的 buyer_favorite1 改为 input 即可）