湖南工业大学研究生课程教学大纲

课程编号：00812011

编写人：万烂军　　编写日期：2022-02-20

<table>
<tr><td>课程中文名称</td><td colspan="8">数据科学与工程</td></tr>
<tr><td>课程英文名称</td><td colspan="8">Data Science and Engineering</td></tr>
<tr><td>开课学期</td><td colspan="2">春季</td><td colspan="2">开课单位</td><td colspan="3">计算机学院</td></tr>
<tr><td>课程类别</td><td colspan="8">专业课（电子信息专业）
选修课（计算机科学与技术专业）</td></tr>
<tr><td>主讲教师</td><td>万烂军</td><td colspan="2">职称</td><td colspan="2">副教授</td><td>联系电话</td><td colspan="2">13762336664</td></tr>
<tr><td>备讲教师</td><td></td><td colspan="2">职称</td><td colspan="2"></td><td>联系电话</td><td colspan="2"></td></tr>
<tr><td rowspan="2">总学时</td><td rowspan="2">32</td><td rowspan="2">其中</td><td colspan="3">讲　课</td><td colspan="2">32</td></tr>
<tr><td colspan="3">实践课时</td><td colspan="2">0</td></tr>
<tr><td>学分</td><td>2</td><td>教学方式</td><td colspan="6">讲授</td></tr>
<tr><td rowspan="2">面向专业</td><td colspan="5" rowspan="2">电子信息专业
计算机科学与技术专业</td><td>考核方式</td><td colspan="2">☐考试
☑考查</td></tr>
<tr><td colspan="8"></td></tr>
<tr><td>预修课程</td><td colspan="8">数理统计</td></tr>
</table>

课程内容 ：

1 数据科学概论 (6 学时)

1.1 大数据的相关概念 (2 学时)

1.2 大数据的关键技术 (2 学时)

1.3 大数据的处理架构 (2 学时)

基本内容：大数据的发展历程、基本概念、主要影响、应用领域；大数据的关键技术和计算模式；大数据与云计算、物联网的关系；大数据处理架构 Hadoop 的生态系统及其各个组件。

2 分布式文件系统 HDFS (8 学时)

2.1 HDFS 的相关概念 (2 学时)

2.2 HDFS 的体系结构 (4 学时)

2.3 HDFS 的存储原理 (2 学时)

基本内容：分布式文件系统的结构；HDFS 的块、名称节点、数据节点、第二名称节点；HDFS 的体系结构设计；HDFS 命名空间管理；HDFS 中 HA 架构和联邦架构的设计；HDFS 的存储原理。

3 分布式并行编程模型 MapReduce (10 学时)

3.1 MapReduce 概述 (2 学时)

3.2 MapReduce 的体系结构 (4 学时)

3.3 MapReduce 的工作流程 (4 学时)

基本内容：分布式并行编程的基本概念；MapReduce 编程模型的基本概念；Map 函数和 Reduce 函数；MapReduce 的体系结构及其缺陷；新一代资源管理调度框架 YARN 的设计思路与体系结构；MapReduce 的各个执行阶段；Map 端的 Shuffle 过程； Reduce 端的 Shuffle 过程。

4 分布式并行计算框架 Spark (8 学时)

4.1 Spark 的相关概念 (2 学时)

4.2 Spark 的运行架构 (4 学时)

4.3 Spark RDD 的基本操作 (2 学时)

基本内容：Spark 的基本概念；Spark 的生态系统；Spark 的环境搭建和使用方法；Spark 运行架构的基本概念；Spark 运行架构设计；Spark 运行基本流程；Spark RDD 的设计与运行原理；RDD 常见的转换操作和行动操作。

课程内容英文简介：

The main teaching contents of 《Data Science and Engineering》include: the basic concepts of big data, the key technologies of big data, and the big data processing architecture; the basic concepts of HDFS, the architecture of HDFS, and the storage principle of HDFS; the overview of MapReduce, the architecture of MapReduce, and the workflow of MapReduce; the basic concepts of Spark, the operation architecture of Spark, and the basic operations of Spark RDD.

课程教学目标及重点、难点：

课程教学目标：

目标 1：能够理解大数据的基本概念、关键技术和处理架构，能够掌握分布式文件存储与分布式并行处理的体系架构及其运行原理，并将大数据的基本理论知识应用在课题研究中。

目标 2：能够掌握分布式文件系统 HDFS、分布式并行编程模型 MapReduce 和分布式并行计算框架 Spark 的基本操作，并运用其来解决大数据存储与处理问题。

课程教学重点：HDFS 的体系结构、MapReduce 的体系结构、MapReduce 的工作流程、Spark 的运行架构、Spark RDD 的基本操作

课程教学难点：HDFS 的体系结构、MapReduce 的工作流程

教学要求：

主要通过课堂教授和课堂讨论等方式进行教学，且选择实际案例进行分析。

教材及主要参考书：

[1] 林子雨. 大数据技术原理与应用(第 3 版)[M]. 北京:人民邮电出版社, 2021.

[2] 石川, 王啸, 胡琳梅. 数据科学导论[M]. 北京:清华大学出版社, 2021.

[3] 林子雨. 大数据基础编程、实验和案例教程(第 2 版)[M]. 北京:清华大学出版社, 2020.

[4] 林子雨. Spark 编程基础(Scala 版)[M]. 北京:人民邮电出版社, 2018.

[5] Tom White. Hadoop 权威指南(第 5 版)[M]. 北京:清华大学出版社, 2015.

[6] 王晓华. Spark MLlib 机器学习实践(第 2 版)[M]. 北京:清华大学出版社, 2017.

大作业：

结合自己的研究方向，撰写一篇与“数据科学与工程”相关的研究综述，要求 3000 字以上。

备注：

1.请任课教师将此电子文档及纸质文档交到相关学院汇总审批后统一于指定日期前（由研究生秘书）交到研究生处。

2.除格式中个别已规定字体外其它采用小四号宋体