传智播客

《Hadoop 大数据技术原理与应用》

教学设计

课程名称：Hadoop 大数据技术原理与应用

授课年级：20xx 年级

授课学期：20xx 学年第一学期

教师姓名：某某老师

2019 年 9 月 1 日

课题名称	第 4 章 MapReduce 分布式计算框架	计划学时	8 课时
内容分析	MapReduce 是 Hadoop 系统核心组件之一，它是一种可用于大数据并行处理的计算模型、框架和平台，主要解决海量数据的计算，是目前分布式计算模型中应用较为广泛的一种，本章通过对 MapReduce 原理、编程模型及案例进行深入讲解。		
教学目标及基本要求	1、 理解 MapReduce 的核心思想 2、 掌握 MapReduce 的编程模型 3、 掌握 MapReduce 的工作原理 4、 掌握 MapReduce 常见编程组件的使用		
教学重点	1、 MapReduce 的编程模型 2、 MapReduce 的工作原理 3、 MapReduce 常见编程组件的使用		
教学难点	1、 MapReduce 的工作原理 2、 MapReduce 常见编程组件的使用		
教学方式	教师课堂教学要以讲演法讲授为主，并结合多媒体进行教学		
教学过程	**第一课时** **（MapReduce 核心思想、MapReduce 编程模型、MapReduce 编程实例——词频统计）** **一、回顾第三章内容，讲解 MapReduce 概述** **1. 回顾第一章学习知识，引出本节主题** 带领学生回顾第三章 HDFS 相关的知识，由于 MapReduce 是 Hadoop 系统的另一个核心组件，它是一种可用于大数据并行处理的计算模型、框架和平台，主要解决海量数据的计算，是目前分布式计算模型中应用较为广泛的一种。因此，本章将针对 MapReduce 分布式计算框架进行详细讲解。 **2. 明确学习目标** （1）理解 MapReduce 核心思想 （2）掌握 MapReduce 编程模型 （3）理解 MapReduce 编程实例——词频统计 **二、进行重点知识讲解** **1. MapReduce 核心思想** 教师可以参考课件讲解 MapReduce 核心思想。 （1） 介绍 MapReduce 的核心思想是分而治之； （2） 通过生活的例子介绍分而治之的思想，便于学生更好的理解 MapReduce 核心思想； （3） 接着介绍 MapReduce 程序的 Map 阶段和 Reduce 阶段的工作。 **2. MapReduce 编程模型**		

教师可以参考课件讲解 MapReduce 编程模型。

（1） 先通过一张图来介绍 MapReduce 的 Map 阶段和 Reduce 阶段的模型；

（2） 接着介绍 MapReduce 程序的实现，实现的过程是通过 map()和 reduce()函数来完成的；

（3） 最后对 MapReduce 的编程模型进行详细说明。

3．MapReduce 编程实例—词频统计

教师可以参考课件以讲演法的方式讲解 MapReduce 编程实例—词频统计。借助 MapReduce 编程思想，来实现词频统计功能。

（1） MapReduce 通过默认组件 TextInputFormat 将的待处理的数据文件的每一行数据转变成<Key，Value>键值对；

（2） 调用 Map()方法，将每一行的单词进行切割计数；

（3） 调用 Reduce()方法将单词汇总和排序；

（4） MapReduce 通过默认组件 TextOutputFormat 组件将数据输出到结果文件中。

三、归纳总结，随堂练习，布置作业

对课堂上讲解的知识点进行总结，使用高校教辅平台中的随堂练习题巩固本节课的知识点。

第二课时

（MapReduce 工作过程、MapTask 和 ReduceTask 以及 Shuffle 工作原理）

一、回顾上节课内容，讲述 HDFS 的架构和原理

1．对上节课留的作业进行答疑

2．回顾上节课内容，引出本节课的主题

通过上节课的学习，学生了解到 MapReduce 框架主要是由 Map 和 Reduce 两个阶段来实现计算的，那么这两个阶段的内部是如何协同工作的呢？本节课将针对 MapReduce 工作原理进行详细讲解。

3．明确学习目标

（1）掌握 MapReduce 工作过程

（2）掌握 MapTask 工作原理

（3）掌握 ReduceTask 工作原理

（4）掌握 Shuffle 工作原理

二、进行重点知识的讲解

1．MapReduce 工作过程

教师可以参考课件来讲述 MapReduce 工作过程。

（1） 先通过一张图来介绍 MapReduce 的工作过程，工作过程大致分为五步；

（2） 详细介绍 MapReduce 工作工程的每一个步骤，先进行分片和格式化数据源操作，接着执行 MapTask、执行 Shuffle 过程、执行 ReduceTask 操作，最后写入文件中。

2．MapTask 工作原理

教师可以参考课件来讲述 MapTask 工作原理。先介绍 MapTask 作为 MapReduce 工作过程的前半部分，它主要经历了 5 个阶段，分别是 Read 阶段、Map 阶段、Collect 阶段、

Spill 阶段和 Combiner 阶段；接着根据 MapReduce 工作原理图详细介绍这五个阶段。

3. ReduceTask 工作原理

教师可以参考课件来讲述 ReduceTask 这个原理。先介绍 ReduceTask 作为 MapReduce 工作过程的后半部分，它主要经历了 5 个阶段，，分别是 Copy 阶段、Merge 阶段、Sort 阶段、Reduce 阶段和 Write 阶段；接着根据 ReduceTask 工作原理图详细介绍这五个阶段。

4. Shuffle 工作原理

教师可以参考课件来讲述 Shuffle 工作原理。先介绍 Shuffle 是 MapReduce 的核心，它用来确保每个 reducer 的输入都是按键排序的，Shuffle 的性能高低直接决定了整个 MapReduce 程序的性能高低；介绍在 Map 和 Reduce 阶段都涉及到了 Shuffle 机制，根据 Shuffle 的过程图来介绍在 Map 和 Reduce 阶段 Shuffle 机制的影响。

三、归纳总结，随堂练习，布置作业

对课堂上讲解的知识点进行总结，使用高校教辅平台中的随堂练习题巩固本节课的知识点。

第三课时

（介绍 InputFormat、Mapper、Reducer、Partitioner、Combiner 及 OutputFormat 组件）

一、回顾上节课内容，讲解 MapReduce 编程组件

1. 对上节课留的作业进行答疑

2. 回顾上节课内容，引出本节课的主题

上一节中，主要讲解了 MapReduce 工作原理，让我们明白了 MapReduce 分布式计算框架的底层是如何协调工作的。本节将针对 MapReduce 编程组件进行详细讲解。

3. 明确学习目标

（1）掌握 InputFormat 组件

（2）掌握 Mapper 组件

（3）掌握 Reducer 组件

（4）掌握 Partitioner 组件

（5）掌握 Combiner 组件

（6）掌握 OutputFormat 组件

二、进行重点知识的讲解

1. InputFormat 组件

教师可以参考课件以讲演法的方式讲述 InputFormat 组件。

（1） 先介绍 InputFormat 组件提供的两个功能，即数据切分和为 Maper 提供输入数据；

（2） 通过源代码介绍 InputFormat 接口定义的两个方法，即 getSplits()和 createRecordReader()，其中 getSplits()方法负责将文件切分为多个分片(split)，

createRecordReader()方法负责创建 RecordReader 对象，用来从分片中读取数据。

2. Mapper 组件

教师可以参考课件以讲演法的方式讲述 Mapper 组件。

（1） 先介绍 Mapper 组件；
（2） 通过源代码介绍继承 Mapper 类并重写 map()方法，实现词频统计。

3. Reducer 组件

教师可以参考课件以讲演法的方式讲述 Reducer 组件。

（1） 先介绍 Reducer 组件及其作用；
（2） 通过源代码介绍 Reducer 类及其内部定义 run()、setup()、reduce()及 cleanup()方法的使用。

4. Partitioner 组件

教师可以参考课件以讲演法的方式讲述 Partitioner 组件。先介绍 Partitioner 组件的作用，接着通过源代码介绍 Hadoop 自带了一个默认的分区类 HashPartitioner，它继承了 Partitioner 类，重写 getPartition()方法，在方法中通过调用 hash 函数对文件数量进行分区，获得一个非负整数的 hash 码，然后用当前作业的 reduce 节点数进行取模运算，从而实现数据均匀分布在 ReduceTask 的目的。

5. Combiner 组件

教师可以参考课件以讲演法的方式讲述 Combiner 组件。先介绍 Combiner 组件的作用，接着通过源代码介绍如果想自定义 Combiner，则需要继承 Reducer 类，并重写 reduce()方法，从而可以将 Key 相同的单词进行汇总。

6. OutputFormat 组件

教师可以参考课件以讲演法的方式讲述 OutputFormat 组件。先介绍 OutputFormat 组件的作用，接着通过源代码介绍 OutputFormat 类中定义的三个方法，即 getRecordWriter()方法用于返回一个 RecordWriter 的实例，checkOutputSpecs()方法用于检测任务输出规范是否有效，getOutputCommiter()方法来负责输出被正确提交。

三、归纳总结，随堂练习，布置作业

对课堂上讲解的知识点进行总结，使用高校教辅平台中的随堂练习题巩固本节课的知识点。

第四课时

（MapReduce 运行模式、MapReduce 性能优化策略）

一、回顾上节课内容，讲解 MapReduce 的运行模式和性能优化策略

1. 对上节课留的作业进行答疑。

2. 回顾上节课的内容，引出本节课的主题。

通过上节课的学习，我们对 MapReduce 的编程组件有了更深入的了解，这有助于我们更好的掌握 MapReduce 分布式计算框架去处理数据。当数据量很大时，针对在部署方法上存在 MapReduce 程序执行效率的问题，采取基于参数优化的方法，来进行调参数从而提高 MapReduce 程序的执行效率。本节课将针对 MapReduce 的运行模式和性能优化策

略进行详细讲解。

3．明确学习目标

（1）了解 MapReduce 运行模式

（2）理解 MapReduce 性能优化策略

二、进行重点知识的讲解

1．MapReduce 运行模式

教师可以参考课件讲述 MapReduce 运行模式。先介绍 MapReduce 的运行模式有两种，分别是本地运行模式和集群运行模式，再介绍这两种运行模式的区别。

2．MapReduce 性能优化策略

教师可以参考课件讲述 MapReduce 性能优化策略。先介绍 MapReduce 性能优化的重要性，再依次介绍从数据输入、Map 阶段、Reduce 阶段、Shuffle 阶段以及其他调优属性五个方面对 MapReduce 进行性能优化。

三、归纳总结，随堂练习，布置作业

对课堂上讲解的知识点进行总结，使用高校教辅平台中的随堂练习题巩固本节课的知识点。

第五课时

（MapReduce 经典案例——倒排索引）

一、回顾上节课内容，讲解 MapReduce 经典案例——倒排索引

1．对上节课留的作业进行答疑。

2．回顾上节课的内容，引出本节课的主题。

通过上节课的学习，我们对 MapReduce 运行模式和性能优化策略有了一定的认识。本节课将针对 MapReduce 经典案例——倒排索引进行详细讲解。

3．明确学习目标

熟悉 MapReduce 经典案例——倒排索引的实现流程

二、进行重点知识的讲解

MapReduce 经典案例——倒序索引

教师可以参考课件以讲演法的方式讲述 MapReduce 经典案例——倒序索引。

（1） 先介绍倒排索引的定义和作用；

（2） 分析案例的需求；

（3） 通过代码实现倒排索引功能。

三、归纳总结，随堂练习，布置作业

1．对课堂上讲解的知识点进行总结，使用高校教辅平台中的随堂练习题巩固本节课的知识点。

2．让学生自己按照步骤实现倒排索引的功能，以此来巩固本节的学习内容。

第六课时

（MapReduce 经典案例——数据去重）

一、回顾上节课内容，讲解 MapReduce 经典案例——数据去重

1．对上节课留的作业进行答疑。

2．回顾上节课的内容，引出本节课的主题。

通过上节课的学习，我们熟悉了 MapReduce 经典案例——倒序索引的具体实现流程。本节课将针对 MapReduce 经典案例——数据去重进行详细讲解。

3．明确学习目标

熟悉 MapReduce 经典案例——数据去重的实现流程

二、进行重点知识的讲解

MapReduce 经典案例——数据去重

教师可以参考课件以讲演法的方式讲述 MapReduce 经典案例——数据去重。

（1） 先介绍数据去重的作用；

（2） 分析案例的需求；

（3） 通过代码实现数据去重的功能。

三、归纳总结，随堂练习，布置作业

1．对课堂上讲解的知识点进行总结，使用高校教辅平台中的随堂练习题巩固本节课的知识点。

2．让学生自己按照步骤实现数据去重的功能，以此来巩固本节的学习内容。

第七课时

（MapReduce 经典案例——TopN）

一、回顾上节课内容，讲解 MapReduce 经典案例——TopN

1．对上节课留的作业进行答疑。

2．回顾上节课的内容，引出本节课的主题。

通过上节课的学习，我们熟悉了 MapReduce 经典案例——TopN 的具体实现流程。本节课将针对 MapReduce 经典案例——TopN 进行详细讲解。

3．明确学习目标

熟悉 MapReduce 经典案例——TopN 的实现流程

二、进行重点知识的讲解

MapReduce 经典案例——TopN

教师可以参考课件以讲演法的方式讲述 MapReduce 经典案例——TopN。

（1） 先介绍 TopN 的作用；

（2） 分析案例的需求；

（3） 通过代码实现 TopN 的功能。

	三、归纳总结，随堂练习，布置作业 1．对课堂上讲解的知识点进行总结，使用高校教辅平台中的随堂练习题巩固本节课的知识点。 2．让学生自己按照步骤实现 TopN 的功能，以此来巩固本节的学习内容。 **第八课时** **（上机练习）** **上机一：使用 MapReduce 实现倒排索引** 请按照教材中 4.6.2 小节的案例，独立完成。 **上机二：使用 MapReduce 实现数据去重** 请按照教材中 4.7.2 小节的案例，独立完成。 **上机三：使用 MapReduce 实现 TopN** 请按照教材中 4.8.2 小节的案例，独立完成。
思考题和习题	
教学后记	