# 西京学院《大数据技术原理及应用》课程教学大纲

## 课程基本信息

（一）课程中文名称：大数据技术原理及应用 代码：08091040121

（二）课程英文名称：Principles and Applications of Big Data Technology

（三）课程性质：专业核心课程

（四）课程学分：3学分（理论2学分 实验1学分）

（五）适用专业：数据科学与大数据技术

（六）开课单位：信息工程学院

## 课程教学目标

（一）课程定位

《大数据技术原理及应用》是学习大数据技术的入门课程，是计算机科学与技术、数据科学与大数据技术、人工智能等专业的一门专业核心课。该课程着重讲述大数据技术的知识体系，阐明其基本原理，引导学生进行初级实践和了解相关应用，培养学生运用大数据平台搭建、大数据收集与预处理、大数据迁移、大数据存储与管理、大数据处理与分析、数据可视化等方法和技术工具完成大数据应用的能力，在整个大数据课程体系中起着承前启后重要作用。该课程由理论教学和实验教学两个部分组成，其先修课程包括面向对象程序设计、Python语言程序设计、操作系统、数据库原理等，后续课程包括数据收集与处理、大数据存储与管理、数据挖掘、大数据技术课程设计、Spark大数据技术、数据可视化技术等。

（二）课程目标

《大数据技术原理及应用》围绕“构建知识体系、阐明基本原理、引导初级实践、了解相关应用”的指导思想，着重讲述大数据技术的基本原理、基本方法、技术工具和相关应用，帮助学生构建完整的大数据技术知识体系，使学生了解大数据应用的整个生命周期，能够熟练使用Hadoop、Spark等主流大数据平台及其生态系统中的技术工具完成基本的大数据应用闭环操作，达到知行合一、以用促学的目的。

经过学习该课程，使学生达到以下目标：

1. 理解大数据基本概念，了解大数据与云计算、物联网、人工智能、5G的关系，了解当前大数据的岗位及需求、学习路线；

2. 认识大数据处理平台Hadoop，熟练掌握全分布模式Hadoop集群的部署基本过程；

3. 理解分布式文件系统HDFS的体系架构、文件存储机制和数据读写过程，熟练掌握HDFS Web UI的使用、HDFS Shell常用命令的使用和HDFS Java API编程，了解HDFS NameNode HA高可用机制、HDFS NameNode Federation联邦机制和HDFS Snapshots快照机制原理和配置；

4. 理解分布式计算框架MapReduce的编程思想和作业执行流程，熟练掌握MapReduce简单编程，了解当前其他主流的大数据计算框架如Spark、Storm、Flink；

5. 理解YARN体系架构和工作原理，掌握YARN Web UI、YARN Shell的基本使用，了解ResourceManager Restart自动重启机制、ResourceManager HA高可用机制和YARN Federation联邦机制的原理与配置；

6. 理解分布式协调框架ZooKeeper的工作原理和应用场景，熟练掌握ZooKeeper集群的部署、ZooKeeper四字命令的使用和ZooKeeper Shell常用命令的使用，了解ZooKeeper Java API编程；

7. 理解NoSQL数据库与关系数据库的区别、分布式数据库HBase的数据模型、体系架构和运行机制，熟练掌握HBase集群的部署、HBase Shell常用命令的使用，了解HBase Web UI、HBase Java API编程、在HBase中使用MapReduce和HBase性能优化；

8. 理解Hive与传统关系数据库相比的特征、体系架构、数据类型、数据模型，掌握Hive函数的使用，熟练掌握Hive的部署、Hive Shell命令和HiveQL的使用，了解Hive Java API编程和优化策略；

9. 理解数据迁移工具Sqoop、日志采集系统Flume、分布式发布/订阅消息系统Kafka和ETL工具Kettle的作用、架构和工作原理，掌握Sqoop、Flume、Kafka和Kettle的安装和基本使用；

10. 了解数据可视化的概念、作用和意义，理解常用的可视化图表类型，了解当前主流的数据可视化工具。

## 课程教学内容

**1.理论教学**

**第一章 大数据概述**

内容与目标：

理解大数据的内涵与特征；了解大数据的影响、应用、关键技术；了解大数据与云计算、物联网、人工智能、5G的关系；了解当前大数据的岗位及需求、学习路线。

1.1 大数据内涵与特征

1.2 大数据影响

1.4 大数据应用

1.5 大数据关键技术

1.6 大数据与云计算、物联网、人工智能的关系

1.7 大数据岗位介绍

1.8 大数据学习路线

重点难点：重点为理解大数据的内涵和特征。

学习要求：课后借阅书籍《大数据时代:生活、工作与思维的大变革》。

思考题（案例）：大数据对思维方式的影响。

作业：

（1）完成在线测试。

（2）课后查阅资料，预习大数据处理平台Hadoop。

**第二章 初识Hadoop**

内容与目标：

了解大数据处理平台Hadoop的来源、发展史、版本、生态系统和应用现状；理解Hadoop的体系架构和运行模式；熟练掌握全分布模式Hadoop集群的部署过程。

2.1 Hadoop概述

2.2 Hadoop生态系统

2.3 Hadoop体系架构

2.4 Hadoop应用现状

2.5 部署Hadoop

重点难点：重点为理解Hadoop的体系架构和运行模式；难点为熟练掌握全分布模式Hadoop集群的部署。

学习要求：课前查阅资料，预习大数据处理平台Hadoop；课后参照实验指导书，自行实践部署Hadoop。

思考题（案例）：部署Hadoop单机模式、伪分布模式、全分布模式的异同点。

作业：

（1）完成在线测试。

（2）课后参照实验指导书，自行实践部署Hadoop。

（3）课后查阅资料，试读Google论文《The Google File System》，预习分布式文件系统HDFS。

**第三章 分布式文件系统HDFS**

内容与目标：

了解HDFS的来源、功能、优缺点；理解HDFS的体系架构、文件存储机制和数据读写过程；熟练掌握HDFS Web UI、HDFS Shell常用命令的使用；掌握HDFS常用Java API的编程；了解HDFS NameNode HA高可用机制和HDFS NameNode Federation联邦机制的原理与部署，了解HDFS Snapshots快照的基本使用。

3.1 HDFS简介

3.2 HDFS体系架构

3.3 HDFS文件存储机制

3.4 HDFS数据读写过程

3.5 实战HDFS

3.5.1 HDFS Web

3.5.2 HDFS Shell

3.5.3 HDFS Java API编程

3.6 HDFS高可靠机制

重点难点：重点为理解HDFS的体系架构、文件存储机制和数据读写过程，熟练掌握HDFS Web UI、HDFS Shell常用命令的使用，掌握HDFS常用Java API的编程；难点为理解HDFS的文件存储机制，掌握HDFS常用Java API的编程。

学习要求：课前查阅资料，预习Hadoop分布式文件系统HDFS；课后参照实验指导书，自行实践HDFS Web UI的基本使用，HDFS Shell常用命令的使用，HDFS常用Java API的编程。

思考题（案例）：分布式文件系统的设计需求和设计目标。

作业：

（1）完成在线测试。

（2）课后参照实验指导书，自行实践HDFS Web、HDFS Shell、HDFS Java API。

（3）课后查阅资料，试读Google论文《MapReduce: Simplified Data Processing on Large Clusters》，预习分布式计算框架MapReduce。

**第四章 分布式计算框架MapReduce**

内容与目标：

了解MapReduce的来源；理解MapReduce计算框架思想、MapReduce体系架构；深刻理解MapReduce作业的执行流程；理解MapReduce的数据类型和Shuffle机制；了解MapReduce Web UI、MapReduce Shell的基本使用；掌握MapReduce简单编程和MapReduce自定义组件；了解MapReduce调优技术和当前其他主流的大数据计算框架如内存计算框架Spark的生态系统、体系架构、部署方式，以及流计算框架Storm、Flink等的设计思想、体系架构等。

4.1 MapReduce简介

4.2 第一个MapReduce案例：WordCount

4.3 MapReduce作业执行流程

4.4 MapReduce数据类型与格式

4.5 Shuffle机制

4.6 在MapReduce中自定义组件

4.7 实战MapReduce

4.7.1 MapReduce Web UI

4.7.2 MapReduce Shell

4.7.3 MapReduce Java API编程

4.8 MapReduce调优

4.9 其它主流计算框架

重点难点：重点为理解MapReduce计算框架思想、MapReduce体系架构，深刻理解MapReduce作业的执行流程，掌握MapReduce基本编程；难点为深刻理解MapReduce作业的执行流程，理解Shuffle机制，掌握MapReduce基本编程。

学习要求：课前查阅资料，预习Hadoop分布式计算框架MapReduce；课后参照实验指导书，自行实践MapReduce Web UI、MapReduce Shell的基本使用，MapReduce基本编程。

思考题（案例）：MapReduce可用于对数据进行排序，一种想法是利用MapReduce的自动排序功能，即默认情况下，Reduce的输出结果是有序的，则可以只使用一个Reducer来对数据进行处理、输出，则结果就是有序的了。但这样的排序过程无法充分利用MapReduce的分布式优点。试设计一个基于MapReduce的排序算法，假设数据均位于[1,100]，Reducer数量为4，升序输出或降序输出均可。

作业：

（1）完成在线测试。

（2）课后参照实验指导书，自行实践MapReduce Web UI、MapReduce Shell，并尝试简单的MapReduce编程。

（3）课后查阅资料，预习统一资源管理框架YARN。

**第五章 统一资源管理和调度框架YARN**

内容与目标：

了解MapReduce 1.0存在的问题及YARN产生背景；理解YARN体系架构和工作流程；掌握YARN Web UI、YARN Shell的基本使用；了解ResourceManager Restart、ResourceManager HA、YARN Federation的原理与配置；了解其它统一资源管理调度框架。

5.1 初识YARN

5.1.1 MapReduce 1.0存在的问题

5.1.2 YARN简介

5.1.3 YARN发展目标

5.2 YARN体系架构

5.3 YARN工作流程

5.4 实战YARN

5.4.1 YARN Web UI

5.4.2 YARN Shell

5.4.3 YARN Java API编程

5.5 YARN新特性

5.5.1 ResourceManager Restart自动重启机制

5.5.2 ResourceManager HA高可用机制

5.5.3 YARN Federation联邦机制

重点难点：重点为理解YARN体系架构和工作原理，掌握YARN Web UI、YARN Shell的基本使用；难点为了解ResourceManager自动重启、ResourceManager高可用机制和YARN联邦机制的原理与配置，。

学习要求：课前查阅资料，预习YARN；课后参照实验指导书，自行实践YARN Web UI、YARN Shell的基本使用。

思考题（案例）：完成YARN ResourceManager高可用机制环境的搭建。

作业：

（1）完成在线测试。

（2）课后参照实验指导书，自行实践YARN Web UI、YARN Shell，再次尝试MapReduce编程。

（3）课后查阅资料，预习分布式协调框架ZooKeeper。

**第六章 分布式协调框架ZooKeeper**

内容与目标：

了解ZooKeeper的设计目标、数据模型；理解ZooKeeper的工作原理；了解ZooKeeper的典型应用场景；熟练掌握ZooKeeper集群的部署、ZooKeeper Shell常用命令的使用；掌握ZooKeeper四字命令的使用；了解ZooKeeper编程。

6.1 初识ZooKeeper

6.2 ZooKeeper工作原理

6.3 ZooKeeper典型应用场景

6.4 部署ZooKeeper集群

6.5 实战ZooKeeper

6.5.1 ZooKeeper 四字命令

6.5.2 ZooKeeper Shell

6.5.3 ZooKeeper Java API编程

重点难点：重点为理解ZooKeeper的执行原理，熟练掌握ZooKeeper集群的部署、ZooKeeper Shell常用命令的使用；难点为了解ZooKeeper编程。

学习要求：课前查阅资料，预习分布式协调框架ZooKeeper；课后参照实验指导书，自行实践ZooKeeper集群的部署、ZooKeeper Shell常用命令的使用。

思考题（案例）：假设有两个线程，两个线程要同时到MySQL中更新一条数据，对数据库中的数据进行累加更新。由于在分布式环境下，这两个线程可能存在于不同机器上的不同JVM进程中，所以这两个线程的关系就是跨主机和跨进程的，使用Java中的synchronized锁是无法实现的。思考如何使用ZooKeeper实现分布式锁？

作业：

（1）完成在线测试。

（2）课后参照实验指导书，自行实践ZooKeeper集群的部署，ZooKeeper Shell常见命令的使用。

（3）课后查阅资料，试读Google论文《Bigtable: A Distributed Storage System for Structured Data》，预习分布式数据库HBase。

**第七章 分布式数据库HBase**

内容与目标：

理解NoSQL数据库与关系数据库的区别、NoSQL的四大类型；了解HBase的发展历程、功能和特点；理解HBase的数据模型、体系架构、运行机制；熟练掌握HBase集群的部署、HBase Shell常用命令的使用；了解HBase Web UI的使用、HBase Java API简单编程；了解HBase性能优化策略。

7.1 NoSQL简介

7.2 初识HBase

7.3 HBase数据模型

7.4 HBase体系架构

7.5 HBase运行机制

7.6 部署HBase集群

7.7 实战HBase

7.7.1 HBase Web UI

7.7.2 HBase Shell

7.7.3 HBase Java API编程

7.8 HBase性能优化

重点难点：重点为理解HBase的数据模型、体系架构、运行机制，熟练掌握HBase集群的部署、HBase Shell常用命令的使用；难点为熟练掌握HBase Shell常见命令的使用，了解HBase Java API的简单编程。

学习要求：课前查阅资料，预习分布式数据库HBase；课后参照实验指导书，自行实践HBase集群的部署、HBase Shell常用命令的使用。

思考题（案例）：当一台Region服务器意外终止时，Master如何发现这种意外终止情况？为了恢复这台发生意外的Region服务器上的Region，Master应该做出哪些处理?

作业：

（1）完成在线测试。

（2）课后参照实验指导书，自行实践HBase集群的部署、HBase Shell常用命令的使用。

（3）课后查阅资料，预习数据仓库Hive。

**第八章 数据仓库Hive**

内容与目标：

了解Hive的产生背景和设计目的；理解Hive的体系架构和执行流程；熟练掌握Hive表、分区、分桶的定义、应用场景和使用方法，熟练掌握Hive的部署和连接方式包括CLI、HiveServer2/beeline等；熟练掌握Hive基本操作包括Hive DDL、Hive DML的书写和Hive Shell命令的使用；了解Hive优化策略。

8.1 初识Hive

8.2 Hive体系架构和执行流程

8.3 Hive数据存储模型

8.4 Hive函数

8.5 部署Hive

8.6 实战Hive

8.6.1 HiveQL之DDL

8.6.2 HiveQL之DML

8.6.3 Hive Shell

8.6.4 Hive API

8.7 Hive优化策略

重点难点：重点为理解Hive的体系架构和执行流程，熟练掌握Hive表、分区、分桶的定义、应用场景和使用方法，熟练掌握Hive的部署和连接方式，熟练掌握Hive基本操作包括Hive DDL、Hive DML的书写和Hive Shell命令的使用；难点为Hive函数、Hive API和优化策略。

学习要求：课前查阅资料，预习数据仓库Hive；课后参照实验指导书，自行实践Hive的部署、Hive表的DDL和DML操作、Hive Shell命令的使用。

思考题（案例）：使用Hive完成用户访问量TOP5统计。数据字段格式为：url、session\_id、referer、ip、end\_user\_id、city\_id，分隔符为制表符。

作业：

（1）完成在线测试。

（2）课后参照实验指导书，自行实践Hive集群的部署、Hive表的DDL和DML操作、Hive内置函数的使用、Hive Shell命令的使用。

（3）课后查阅资料，预习数据迁移和采集。

**第九章 大数据迁移和采集工具**

内容与目标：

了解Sqoop的产生背景；理解Sqoop的作用、架构、工作原理；掌握Sqoop的安装；掌握使用Sqoop完成数据从关系型数据库到Hadoop生态系统的导入、导出；理解Flume的作用、架构、工作原理；了解Flume的部署类型；掌握Flume的安装和简单使用；理解Kafka的作用、架构、工作原理；掌握Kafka的部署和简单使用；理解Kettle的作用、架构、工作原理；掌握Kettle的安装和简单使用。

9.1 数据迁移框架Sqoop

9.1.1 初识Sqoop

9.1.2 Sqoop架构与工作原理

9.1.3 安装Sqoop

9.1.4 使用Sqoop

9.2 日志采集系统Flume

9.2.1 初识Flume

9.2.2 Flume架构与工作原理

9.2.3 安装Flume

9.2.4 Flume简单使用

9.3 分布式发布/订阅消息系统Kafka

9.3.1 初识Kafka

9.3.2 Kafka架构与工作原理

9.3.3 部署Kafka集群

9.3.4 Kafka简单使用

9.4 ETL工具Kettle

9.4.1 初识Kettle

9.4.2 Kettle架构与工作原理

9.4.3 安装Kettle

9.4.4 使用Kettle

重点难点：重点为理解Sqoop、Flume、Kafka和Kettle的作用、架构、工作原理，掌握Sqoop、Flume、Kafka和Kettle的安装和简单使用。

学习要求：课前查阅资料，预习Sqoop、Flume、Kafka和Kettle；课后参照实验指导书，自行实践Sqoop、Flume、Kafka和Kettle的安装和简单使用。

思考题（案例）：

（1）使用Sqoop以增量的方式导入数据，要求只导入符合某指定条件的数据到HDFS。

（2）Flume的部署类型包括单一流程、多代理流程、流的合并、多路复用流，参考官网尝试实践各种类型的配置。

（3）在Kafka在0.8以前的版本中，一旦某一个Broker宕机，则其上所有的Partition数据都不可被消费，这与Kafka数据持久性及Delivery Guarantee的设计目标相悖。随着集群规模的增加，整个集群中出现该类异常的几率大大增加，因此在实际生产系统中，如何解决这种问题？

作业：

（1）完成在线测试。

（2）课后参照实验指导书，自行实践Sqoop、Flume、Kafka和Kettle的安装和简单使用。

（3）课后查阅资料，预习数据可视化。

**第十章 数据可视化**

内容与目标：

了解数据可视化的概念、作用和过程；理解常用的数据可视化图形；了解数据可视化工具；了解数据可视化的典型案例。

10.1 数据可视化概述

10.2 数据可视化作用与意义

10.3 数据可视化图表类型

10.4 数据可视化工具

10.5 可视化典型案例

重点难点：重点为理解常用的数据可视化图表类型；难点为了解数据可视化工具。

学习要求：课前查阅资料，预习数据可视化；课后自行实践几种数据可视工具的简单使用。

思考题（案例）：查阅资料，找出几个数据可视化的有趣案例。

作业：完成在线测试。

**2.其它教学环节（如实习、实验、上机等）**

**实验一 部署全分布模式Hadoop集群**

任务：在Linux下部署全分布模式Hadoop集群。

场所：计算机实验室。

内容：

（1）规划部署。

（2）准备机器。

（3）准备软件环境：配置静态IP；修改主机名；编辑域名映射；安装和配置Java；安装和配置SSH免密登录。

（4）下载和安装Hadoop。

（5）配置全分布模式Hadoop集群。

（6）关闭防火墙。

（7）格式化文件系统。

（8）启动和验证Hadoop。

（9）关闭Hadoop。

教学方法：演示法，实验教学法，任务驱动法，小组讨论。

基本要求：

（1）理解Hadoop的体系架构、运行模式。

（2）熟练掌握Linux基本命令。

（3）掌握静态IP地址的配置、主机名和域名映射的修改。

（4）掌握Linux环境下Java的安装、环境变量的配置、Java基本命令的使用。

（5）理解为何需要配置SSH免密登录，掌握Linux环境下SSH的安装、免密登录的配置。

（6）熟练掌握在Linux环境下如何部署全分布模式Hadoop集群。

重点难点：重点为熟练掌握在Linux环境下如何部署全分布模式Hadoop集群；难点为掌握Linux环境下SSH免密登录的配置，熟练掌握在Linux环境下如何部署全分布模式Hadoop集群。

说明：注意用电安全。

**实验二 实战HDFS**

任务：使用HDFS Web和HDFS Shell常用命令，编写HDFS Java API常规操作程序如创建目录、创建文件、重命名、上传本地文件到HDFS、查看文件位置等。

场所：计算机实验室。

内容：

（1）查看HDFS Web界面。

（2）练习HDFS Shell文件级命令。

（3）练习HDFS Shell系统级命令。

（4）搭建HDFS项目开发环境。

（5）使用HDFS Java API编写HDFS文件操作程序，完成上传本地文件到HDFS的功能，采用本地执行和集群执行的两种执行方式测试，观察结果。

（6）使用HDFS Java API编写HDFS文件操作程序，完成查看上一步所上传的文件在HDFS集群的位置的功能，采用本地执行和集群执行的两种执行方式测试，观察结果。

教学方法：演示法，实验教学法，任务驱动法，小组讨论。

基本要求：

（1）理解HDFS体系架构。

（2）理解HDFS文件存储机制，读取和写入HDFS文件的工作流程。

（3）熟练掌握HDFS Web和HDFS Shell常用命令的使用。

（4）掌握HDFS项目开发环境的搭建。

（5）掌握使用HDFS Java API编写HDFS文件操作程序。

重点难点：重点为熟练掌握HDFS Web和HDFS Shell常用命令的使用；难点为掌握使用HDFS Java API编写HDFS文件操作程序。

说明：注意用电安全。

**实验三 MapReduce编程之英文词频统计**

任务：基于MapReduce计算框架，使用Java语言编写MapReduce应用程序WordCount，用于实现统计多个文件中英文单词出现的频次。

场所：计算机实验室。

内容：

（1）使用Java语言编写MapReduce应用程序WordCount，用于实现统计多个文件中英文单词出现的频次，并在Hadoop集群上运行该MR-App，查看运行结果。

（2）查看MapReduce Web UI界面。

（3）练习MapReduce Shell常用命令。

教学方法：演示法，实验教学法，任务驱动法，小组讨论。

基本要求：

（1）理解MapReduce计算框架思想、MapReduce体系架构。

（2）深刻理解MapReduce作业的执行流程。

（3）了解MapReduce Web UI的使用。

（4）了解MapReduce Shell常用命令的使用。

（5）理解MapReduce的数据类型，了解MapReduce应用程序编写步骤，编写简单的MapReduce应用程序，熟练掌握在Hadoop集群上运行MapReduce应用程序并查看运行结果。

重点难点：重点为理解MapReduce计算框架思想、MapReduce体系架构，深刻理解MapReduce作业的执行流程，熟练掌握在Hadoop集群上运行MR-App并查看运行结果；难点为深刻理解MapReduce作业的执行流程，理解MapReduce的数据类型，编写简单的MapReduce应用程序。

说明：注意用电安全。

**实验四 MapReduce编程之排序**

任务：基于MapReduce计算框架，使用Java语言编写MapReduce应用程序Sort，用于实现对输入文件中的数据进行排序。

场所：计算机实验室。

内容：

（1）使用Java语言编写MapReduce应用程序Sort，用于实现对输入文件中的数据进行排序，输入文件中的每行内容均为一个数字，即一个数据，要求在每行输出两个间隔的数字，其中，第一个代表原始数据在原始数据集中的排位，第二个代表原始数据。

（2）在Hadoop集群上运行该MapReduce应用程序Sort，并查看运行结果。

教学方法：演示法，实验教学法，任务驱动法，小组讨论。

基本要求：

（1）继续深刻理解MapReduce作业的执行流程。

（2）掌握MapReduce应用程序的编写。

重点难点：重点为掌握MapReduce应用程序的编写；难点为掌握MapReduce应用程序的编写。

说明：注意用电安全。

**实验五 部署ZooKeeper集群和实战ZooKeeper**

任务：部署ZooKeeper集群，练习ZooKeeper Shell常用命令，进行简单ZooKeeper编程。

场所：计算机实验室。

内容：

（1）部署ZooKeeper集群。

（2）练习ZooKeeper Shell常用命令。

（3）实战ZooKeeper编程，实现对zookeeper文件系统的增删改查。

教学方法：演示法，实验教学法，任务驱动法，小组讨论。

基本要求：

（1）理解ZooKeeper执行原理。

（2）熟练掌握ZooKeeper集群的部署。

（3）熟练掌握ZooKeeper Shell常用命令的使用。

（4）了解ZooKeeper Java API，进行简单的ZooKeeper编程。

重点难点：重点为熟练掌握ZooKeeper集群的部署、ZooKeeper Shell常用命令的使用；难点为了解ZooKeeper Java API，进行简单的ZooKeeper编程。

说明：注意用电安全。

**实验六 部署HBase集群和实战HBase**

任务：部署HBase集群，查看HBase Web UI界面，练习HBase Shell常用命令，进行简单HBase编程。

场所：计算机实验室。

内容：

（1）部署全分布模式HBase集群。

（2）查看HBase Web UI界面。

（3）练习HBase Shell常用命令。

（4）实战HBase编程。

教学方法：演示法，实验教学法，任务驱动法，小组讨论。

基本要求：

（1）理解HBase数据模型。

（2）理解HBase体系架构和运行机制。

（3）熟练掌握HBase集群的部署。

（4）了解HBase Web UI的使用。

（5）熟练掌握HBase Shell常用命令的使用。

（6）了解HBase Java API，进行简单的HBase编程。

重点难点：重点为熟练掌握HBase集群的部署、HBase Shell常用命令的使用；难点为了解HBase Java API，进行简单的HBase编程。

说明：注意用电安全。

**实验七 部署Hive和实战Hive**

任务：部署Hive，练习HiveQL和Hive Shell，进行简单Hive编程。

场所：计算机实验室。

内容：

（1）部署Hive。

（2）配置和查看HWI。

（3）练习Hive DDL、Hive DML。

（4）练习Hive Shell命令。

（5）实战Hive编程。

教学方法：演示法，实验教学法，任务驱动法，小组讨论。

基本要求：

（1）理解Hive体系架构和执行流程。

（2）了解Hive的部署模式，熟练掌握本地模式Hive的部署。

（3）熟练掌握Hive的连接方式包括CLI、HiveServer2/beeline、HWI。

（4）掌握HWI的配置和使用。

（5）熟练掌握Hive表的DDL、DML操作和Hive Shell命令的使用。

（6）了解Hive Java API，进行简单的Hive编程。

重点难点：重点为熟练掌握本地模式Hive的部署，熟练掌握Hive的连接方式包括CLI、HiveServer2/beeline、HWI，熟练掌握Hive表的DDL、DML操作和Hive Shell命令的使用；难点为了解Hive Java API，进行简单的Hive编程。

说明：注意用电安全。

**实验八 安装和简单使用Sqoop、Flume、Kafka和Kettle**

任务：安装Sqoop，练习Sqoop基本命令，使用Sqoop导入导出数据；安装Flume和简单使用；安装Kafka和简单使用；安装Kettle和简单使用。

场所：计算机实验室。

内容：

（1）安装Sqoop，练习Sqoop基本命令的使用，使用Sqoop导入MySQL数据到HDFS，使用Sqoop导出HDFS数据到MySQL。

（2）安装Flume和简单使用。

（3）部署Kafka集群和简单使用。

（4）安装Kettle和简单使用。

教学方法：演示法，实验教学法，任务驱动法，小组讨论。

基本要求：

（1）理解Sqoop的作用、架构、工作原理，掌握Sqoop的安装，掌握使用Sqoop完成数据从关系型数据库到Hadoop生态系统的导入、导出。

（2）理解Flume的作用、架构、工作原理，了解Flume的安装类型，掌握Flume的安装和简单使用。

（3）理解Kafka的作用、架构、工作原理，掌握Kafka集群的部署和简单使用。

（4）理解Kettle的作用、架构、工作原理，掌握Kettle的安装和简单使用。

重点难点：重点为掌握Sqoop的安装，掌握使用Sqoop完成数据从关系型数据库到Hadoop生态系统的导入、导出，掌握Flume的安装和简单使用，掌握Kafka集群的部署和简单使用，掌握Kettle的安装和简单使用；无难点。

说明：注意用电安全。

## 教学内容学时分配

|  |  |  |  |
| --- | --- | --- | --- |
| **教学环节**  **教学时数**  **课程内容** | **理论**  **（学时）** | **实验/上机**  **（学时）** | **小计**  **（学时）** |
| 第一章 大数据概述 | 2 | 0 | 2 |
| 第二章 初识Hadoop | 2 | 2 | 4 |
| 第三章 分布式文件系统HDFS | 4 | 2 | 6 |
| 第四章 分布式计算框架MapReduce | 6 | 4 | 10 |
| 第五章 统一资源管理和调度框架YARN | 2 | 0 | 2 |
| 第六章 分布式协调框架ZooKeeper | 4 | 2 | 6 |
| 第七章 分布式数据库HBase | 4 | 2 | 6 |
| 第八章 数据仓库Hive | 4 | 2 | 6 |
| 第九章 大数据迁移和采集工具 | 2 | 2 | 4 |
| 第十章 数据可视化 | 2 | 0 | 2 |
| **合 计** | **32** | **16** | **48** |

课程教学学时分配表

## 教学参考资料

教材：

[1] 徐鲁辉. 《Hadoop大数据原理与应用》[M]. 西安:西安电子科技大学出版社, 2020年1月.

[2] 徐鲁辉. 《Hadoop大数据原理与应用实验教程》[M]. 西安:西安电子科技大学出版社, 2020年2月.

参考资料：

[1] 林子雨. 《大数据技术原理与应用（第2版）》[M]. 北京:人民邮电出版社, 2017年1月.

[2] 维克托·迈尔-舍恩伯格, 肯尼思·库克耶. 盛杨燕等译. 大数据时代:生活、工作与思维的大变革[M]. 杭州:浙江人民出版社, 2013年1月.

[3] Tom White. 《Hadoop权威指南（第4版修订版）》[M]. 北京:清华大学出版社, 2017年7月.

[4] 蔡斌. 《Hadoop技术内幕:深入解析Hadoop Common和HDFS架构设计与实现原理》[M]. 北京:机械工业出版社, 2013年5月.

[5] 董西成. 《Hadoop技术内幕：深入解析MapReduce架构设计与实现原理》[M]. 北京:机械工业出版社, 2013年5月.

[6] 董西成. 《Hadoop技术内幕：深入解析YARN架构设计与实现原理》[M]. 北京:机械工业出版社, 2014年1月.

[7] Bill Chambers, Matei Zaharia. 《Spark权威指南（影印版）》[M]. 南京:东南大学出版社, 2018年11月.

[8] 薛志东. 《大数据技术基础》[M]. 北京:人民邮电出版社, 2018年8月.

[9] 王宏志. 《Hadoop集群程序设计与开发》[M]. 北京:人民邮电出版社, 2018年8月.

[10] 余明辉. 《Hadoop大数据开发基础》[M]. 北京:人民邮电出版社,2018年2月.

[11] 林子雨. 《Spark编程基础》[M]. 北京:人民邮电出版社,2018年7月.

[12] 肖芳. 《Spark大数据技术与应用》[M]. 北京:人民邮电出版社,2018年2月.

[13] 肖睿. 《基于Hadoop与Spark的大数据开发实战》[M]. 北京:人民邮电出版社,2018年4月.

[14] Apache Hadoop Project. https://hadoop.apache.org.

[15] Apache Spark Project. https://spark.apache.org.

[16] The Hadoop Ecosystem Table. https://hadoopecosystemtable.github.io

[17] 高校大数据课程公共服务平台. http://dblab.xmu.edu.cn/post/bigdata-teaching-platform.

[18] 林子雨. 中国大学MOOC《大数据技术原理与应用》. https://www.icourse163.org/course/XMU-1002335004.

## 课程考核

（一）考核方式：考试。

（二）考核方案：平时考核30%，实验考核20%，期末考核50%。把握整个教学过程的各个环节，实施多元化考核，从学生的课堂出勤、课堂表现、课下作业、在线测试、实验完成情况及完成质量，对学生进行全方位的考核。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **考核项目** | **考核主要内容及其知识点** | **考核方式** | **考核时间** | **所占权重** |
| 平时考核 | 到课率，参与态度，作业完成情况 | 课堂考勤，课堂表现，纸质作业，在线测试 | 第1-12教学周 | 30% |
| 实验考核 | 实验态度，各实验重要技术的掌握情况 | 实验报告 | 第1-12教学周 | 20% |
| 期末考核 | 全部知识考点的掌握情况 | 在线理论考核+设计制作（大作业） | 第18教学周 | 50% |

（三）考核标准：闭卷考试，考核标准按照试卷答案及评分标准。