

---

# Hadoop 关键技术与 Spark 内存计算框架

本课程将介绍目前大数据的核心技术和应用实例，并以实践操作和项目式教学的方式着重讲解 Hadoop 和 Spark 的基本原理和应用。

## 一、Hadoop 关键技术

学习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、MapReduce 等安装、操作和编程。其中会介绍一些 Hadoop 的应用案例，并通过一些实验初步了解 Hadoop 的操作。

### 第 1 章 Hadoop 概论

#### 1.1 缘于搜索的 Hadoop

##### 1.1.1 Hadoop 简介

##### 1.1.2 Hadoop 发展

#### 1.2 大数据、Hadoop 的关系

#### 1.3 Hadoop 设计思想与架构

##### 1.3.1 数据存储与切分

##### 1.3.2 MapReduce 模型

---

### 1.3.3 MPI 和 MapReduce

## 第 2 章 Hadoop 存储系统

### 2.1 基本概念

#### 2.1.1 NameNode

#### 2.1.2 DataNode

#### 2.1.3 客户端

#### 2.1.4 块

### 2.2 HDFS 的特性和目标

#### 2.2.1 HDFS 的特性

#### 2.2.2 HDFS 的目标

### 2.3 HDFS 架构

#### 2.3.1 Master/Slave 架构

#### 2.3.2 NameNode 和 Secondary NameNode 通信模型

#### 2.3.3 文件存取机制

### 2.4 HDFS 核心设计

### 2.5 HDFS 权限管理

---

## 第 3 章 HDFS 的使用

### 3.1 HDFS 环境准备

### 3.2 HDFS 命令的使用

### 3.3 HDFS Java API 的使用方法

## 第 4 章 MapReduce 计算框架

### 4.1 Hadoop MapReduce 简介

### 4.2 MapReduce 模型

#### 4.2.1 MapReduce 编程模型

#### 4.2.2 MapReduce 实现原理

## 第 5 章 Hadoop 命令系统

### 5.1 Hadoop 命令系统的组成

### 5.2 用户命令

## 第 6 章 Hadoop 作业调度系统

### 6.1 作业调度概述

#### 6.1.1 相关概念

---

6.1.2 作业调度流程

6.1.3 集群资源组织与管理

6.1.4 队列控制和权限管理

6.1.5 插件式调度框架

## **第 7 章 Hadoop 集群搭建**

7.1 Hadoop 版本的选择

7.2 集群基础硬件需求

7.3 安装 Hadoop

### **实验**

实验一：熟悉常用的 Linux 操作和 Hadoop 操作

实验二：熟悉常用的 HDFS 操作

实验三：熟悉常用的 HBase 操作

实验四：NoSQL 和关系数据库的操作比较

实验五：MapReduce 初级编程实践

## **二、Spark 内存计算框架**

---

介绍为什么会出现 Spark ? Spark 是什么 ? Spark 能做什么 ? 还有 Spark 安装、使用以及编程基础，并初步了解 Spark SQL 等核心技术。其中穿插一些 Spark 的典型应用案例，并通过动手实验初步体验 Spark 的应用。

## **第 8 章 Spark 概述**

8.3.1 Spark 的出现与发展

8.3.2 Spark 协议族

8.3.3 Spark 的应用及优势

## **第 9 章 Spark 原理**

9.1 Spark 工作原理

9.2 Spark 架构及运行机制

9.2.1 Spark 系统架构与节点角色

9.2.2 Spark 作业执行过程

9.2.3 应用初始化

9.2.4 构建 RDD 有向无环图

9.2.5 RDD 有向无环图拆分

---

## 第 10 章 RDD 算子

### 10.1 创建算子

#### 10.1.1 基于集合类型数据创建 RDD

#### 10.1.2 基于外部数据创建 RDD

### 10.2 transformation 变换算子

#### 10.2.1 对 Value 型 RDD 进行变换

#### 10.2.2 对 Key/ Value 型 RDD 进行变换

### 10.3 action 行动算子

#### 10.3.1 数据运算类行动算子

#### 10.3.2 存储型行动算子

## 第 11 章 安装和使用 Spark

### 11.1 安装 Spark

### 11.2 编写和运行 Spark 程序

**实验：**

---

实验 1 : Linux 系统基本命令和 Hadoop 使用方法

实验 2 : RDD 基本操作

实验 3 : 迭代式算法编程实践

实验 4 : 自定义分匙、排序、合并

实验 5 : 利用 DataFrame 实现数据库的读写

实验 6 : 利用 Spark Streaming 实现流数据处理