

# Spark 培训

## 课程定位与课程目标

Spark 是第一个脱胎于该转变的快速、通用分布式计算范式。Spark 使用函数式编程范式扩展了 MapReduce 模型以支持更多计算类型，可以涵盖广泛的工作流，这些工作流之前被实现为 Hadoop 之上的特殊系统。Spark 使用内存缓存来提升性能，因此进行交互式分析也足够快速(就如同使用 Python 解释器，与集群进行交互一样)。缓存同时提升了迭代算法的性能，这使得 Spark 非常适合数据理论任务，特别是机器学习。

本课程中，我们将首先讨论如何在本地机器上或者 EC2 的集群上设置 Spark 进行简单分析。然后，我们在入门级水平探索 Spark，了解 Spark 是什么以及它如何工作（希望可以激发更多探索）。最后两节我们开始通过命令行与 Spark 进行交互，然后演示如何用 Python 写 Spark 应用，并作为 Spark 作业提交到集群上。

**适用学员：**从事无线建设、无线规划、无线覆盖的工管、网络建设、客户经理等部门员工

## 课程设计：

课程编号：	21090203016
授课课时：	3 至 5 天
授课条件：	学员必须具有基本的 JAVA 编程知识

## 内容摘要：

### 第一章 Spark 大数据开放的技术相关

#### 1.1 什么是 Spark

#### 1.2 Spark 与 Hadoop 的区别

#### 1.3 Spark 生态

- Spark (内存计算框架)

- SparkStreaming (流式计算框架)

- Spark SQL (ad-hoc)

- Mllib (Machine Learning)

- GraphX (bagel 将被取代)

#### 1.4 安装部署

- Spark 安装简介

Spark 的源码编译

Spark Standalone 安装

Spark Standalone HA 安装

Spark 应用程序部署工具 spark-submit

## 第二章 Spark 运行架构和解析

### 2.1 Spark 的运行架构

基本术语

运行架构

Spark on Standalone 运行过程

Spark on YARN 运行过程

### 2.2 Spark 运行实例解析

#### 2.3 Spark on Standalone 实例解析

#### 2.4 Spark on YARN 实例解析

小结

## 第三章 Spark 调优

### 3.1 Spark 生态系统概述

- 回顾 Hadoop MapReduce
- Spark 运行模式
- RDD
- Spark 运行时模型简介
- 缓存策略介绍
- transformation
- action
- lineage
- 容错处理
- 宽依赖与窄依赖
- 集群配置

### 3.2 Spark 的监控

Spark UI 监控

Ganglia 监控

### 3.3 Spark 调优

## 第四章 Spark 编程模型和解析

### 4.1 Spark 的编程模型

Spark 编程模型解析

RDD 的特点、操作、依赖关系

Spark 应用程序的配置

### 4.2 Spark 编程实例解析

日志的处理

电信基站数据的处理

### 4.3 Spark 的多语言编程

Spark 的 scala 编程

Scala 基本语法

Scala 开发环境搭建

Scala 开发 Spark 应用程序

### 4.4 Spark 的 Python 编程

Python 的基本语法

Python 开发 Spark 应用程序

## 第五章 Spark Streaming 原理和实践

### 5.1 Spark Streaming 原理

Spark 流式处理架构

DStream 的特点

Dstream 的操作和 RDD 的区别

Spark Streaming 的优化

### 5.2 Spark Streaming 实例

文本实例

Window 操作

## 第六章 Spark SQL 原理和实践

### 6.1 Spark SQL 原理

Spark SQL 的 Catalyst 优化器

Spark SQL 内核

Spark SQL 和 Hive

### 6.2 Spark SQL 的实例和编程

Spark SQL 的实例操作 demo

Spark SQL 的编程

## 第七章 Spark 源码研读

### 7.1 Spark 源码研读

Spark 源码下载和研读环境搭建

### 7.2 Spark Core 介绍

SparkContext

Executor

Deploy

### 7.3 RDD 和 Storage

### 7.4 Scheduler 和 Task

### 7.5 Spark Examples 介绍

## 第八章 应用中的数据挖掘算法

### 8.1 Spark 机器学习入门

### 8.2 机器学习的原理

### 8.3 Mllib 简介

### 8.4 Mllib 的例程分析

## 第九章 大数据的 zookeeper 分布式

### 9.1 安装和配置详解

单机模式

配置文件介绍

9.2 BIN 目录介绍及 **zookeeper** 的启动

9.3 集群模式

9.4 分布式队列与设计思路

## 第十章 应用服务器 **Jboss hadoop**

10.1 服务器软硬件配置

10.2 软件需求分析

10.3 Jboss 服务器配置详解

10.4 Jboss 部署配置文件

10.5 Jboss 实例

授课语言：

中文