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）

SparkSteaming（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

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