Spark 线上班大纲

开班起始时间

2016 年 5 月 22 日

网址

http://www.bjsxt.com/html/cloud/

本课程招生对象

本课程是零基础 spark 班，scala 会从零开始讲，适合有一点计算机编程基础的学生。如果以前接触过 java 或 c 语言，了解关系型数据熟悉 sql 语句学起来效果会更好。

培训方式

在线直播互动教学

QQ 群在线答疑

直播平台：YY 直播或腾讯课堂（如果有效果更好的直播平台，根据实际情况调整）

培训课程安排

直播课程安排：

每周 5 个晚上直播（留 2 个晚上自己复习，消化知识）

晚上时间安排：20:30-10:00

直播一共持续 8 周（如果有国家法定节假日，课程往后合理顺延）

直播课程结束后发放一个录制详细的大型电商大数据分析项目视频（视频采用加密方式，一机一码方式发给学生）

课程内容涵盖 spark 所有体系：

Scala

Spark core

Spark Streaming

Flume

Kafka

Spark SQL

Spark MLlib

Spark GraphX

学费

第一期优惠价 6980（原价 8980）

第二期涨价到 8980

大型电商数据分析平台真是项目(使用技术 :Spark 、Flume、HDFS、Yarn、Kafka、Zookeeper、Hive、MySQL)

本部分统一发送加密的视频自主学习，老师线上答疑

在实战中学习，技术点非常多，怎么样实际运用这些点是我们在自学过程中体验不到的。

电商数据分析包括：区域海量热门商品统计、页面转化率计算、用户访问分析、点击广告流量统计分析。以及性能优化、数据倾斜、故障解决相关处理。

Scala 编程零基础实战

1. 快速上手实战（基础语法、条件控制与循环）
2. Scala doc 文档使用
3. 打牢基础实战（函数入门、数据结构）
4. 面向对象编程之类、继承实战
5. 面向对象编程之对象、特质实战
6. 面向函数编程详解
7. 面向函数编程之集合操作
8. 高级特性实战（模式匹配、类型参数、隐式转换）
9. 常用注解实战
10. Actor 并发模型应用开发实战

从 Scala 到开发 Spark 程序再到集群中运行以及运维

1. 集群环境搭建实战（Centos 集群搭建、hadoop 集群搭建、spark 集群搭建）
2. 启动脚本和配置详解
3. Spark 工作原理初探与 RDD 详解
4. 实战详解 Spark-shell 交互式、Scala IDE 和 IntelliJ IDEA 的开发使用
5. 实战 Scala、Java、Python 编程语言 API 来进行 WordCount 程序开发
6. 对 WordCount 程序的详细讲解
7. 实战 Local、Standalone 集群、Yarn 集群模式下运行 Spark 程序
8. Spark-submit 脚本的详解
9. Spark Web UI 以及作业监控
10. Zookeeper 的安装以及利用 zookeeper 实现 spark 集群 HA 高可用

Spark Core 核心编程

1. RDD 内核架构概览
2. RDD 的不同数据来源的创建方式详解
3. RDD 的操作算子综述与本质分析（转换算子、行动算子）
4. 常用操作算子的案例实战
5. RDD 持久化实战以及 Checkpoint
6. RDD 共享变量以及累加器的使用实战
7. RDD 简单排序功能（优化之前 WordCount 程序）以及二次排序的实战
8. Spark 实战 Top N 功能详解

9. Spark 任务调度流程整体架构分析详解
10. Spark 任务划分流程整体架构分析详解（宽依赖与窄依赖、DAGScheduler 源码分析）
11. Spark 执行任务相关原理以及源码分析（TaskScheduler、Executor、Task、Shuffle）
12. Spark 实战之 PageRank
13. 性能优化与调优的分析

Spark SQL

1. Spark RDD 应用 SQL 实战
2. RDD 转化为 DataFrame 数据框的方式详解
3. Spark DataFrame 数据框操作实战
4. 加载和保存数据操作（load 与 save）
5. JSON 数据源实战案例
6. JDBC 数据源实战案例
7. Hive 数据源实战案例
8. Parquets 数据源实战加载数据、自动分区推断、合并元数据
9. 内置函数的实战案例
10. 开窗函数的实战案例
11. Spark SQL UDF 自定义函数实战
12. Spark SQL UDAF 自定义聚合函数实战
13. Spark SQL 工作原理详解以及 Spark SQL 的源码分析
14. Hive on spark 深度解密

Flume 实战

1. Flume 安装及配置测试实战
2. 与 Spark 整合接收 Flume 实时数据流实战

Kafka 实战

1. Kafka 分布式消息队列特点以及架构分析
2. Kafka 配置安装以及测试实战
3. Kafka 元数据以及实际数据存储详解
4. Kafka 的 API 使用实战
5. Kafka 和 spark 的整合实战案例

Spark Streaming 实时计算

1. Spark Streaming 和 Storm 对比讲解
2. Spark Streaming 本质原理分析
3. Wordcount 程序的实时版本开发
4. Spark Streaming 和 Spark Core 里面 context 的不同
5. 输入 DStream 和 Receiver 的讲解
6. 不同输入源（Kafka、HDFS）的 DStream 操作实战
7. 基于 DStream 的 window 滑动窗口实战案例

8. 基于 DStream 的 updateStateByKey 实战案例
9. 基于 DStream 的 transform 实战案例
10. DStream 的输出存储操作以及核心函数 foreachRDD 实战
11. Spark Streaming 的持久化实战以及 Checkpoint
12. 与 Spark SQL 结合使用实战案例
13. 架构原理分析与性能优化

Spark MLlib 机器学习

1. 线性回归
2. 线性分类
3. 逻辑回归
4. SVM 支持向量机
5. 推荐引擎之协同过滤
6. SVD 分解技术
7. K-means 聚类分析
8. 主成分分析之 PCA 降维技术
9. 文本处理技术之 tf-idf
10. 神经网络
11. 词向量
12. 机器学习三板斧—流程处理框架

Spark GraphX 图计算

1. PageRank 实战以及架构原理详解
2. Table 算子操作实战
3. Graph 算子操作实战
4. 常见算法解析和实战
5. 社交网络应用

Tachyon 内存分布式文件系统

1. Tachyon 带来的好处以及特性详解
2. Tachyon 架构原理分析
3. Tachyon 的安装部署实战
4. Tachyon 命令行操作实战
5. 整合 Spark 以 Tachyon 为输入输出源的实战
6. 整合 Spark 以 Tachyon 作为持久化 RDD 的实战

大型电商数据分析平台真是项目

使用技术：**Spark（spark core,spark streaming,spark sql,spark mlib） 、Flume、Kafka,hive Zookeeper,Hadoop,mysql**

本部分统一发送加密的视频自主学习，老师线上答疑

在实战中学习，技术点非常多，怎么样实际运用这些点是我们在自学过程中体验不到的。

电商数据分析包括：区域海量热门商品统计、页面转化率计算、用户访问分析、点击广告流量统计分析。以及性能优化、数据倾斜、故障解决相关处理。

1. 需求分析讲解
2. 数据来源日志采集实战
3. 海量商品热查询实战
4. 关联表查询实战
5. 自定义函数实战
6. 开窗函数实战
7. 存储海量结果入数据库实战
8. 页面 PV、UV 计算实战
9. 页面转化率计算分析实战
10. 随机采样用户行为实战
11. 用户行为聚合分析实战
12. Top N 活跃用户分析实战
13. Top N 畅销商品分析实战
14. 二次排序的使用实战
15. 按时间按区域统计广告流量实战
16. 项目的高可用实战
17. 性能优化实战
18. 解决数据清晰问题实战
19. 一些项目中故障处理实战