企业级 Spark 1.x

从菜鸟到高手精品进阶课程

Spark 市场需求

拉手招聘网站【Spark 工程师】薪资待遇

Spark开发工程师 [北京]
月薪：10k-20k 经验： 1-3年 最低学历： 本科
职位诱惑：核心部门，待遇丰厚，气氛融洽
2天前发布
58同城 投个简历
领域： 生活服务,分类信息 创始人： 姚劲波
阶段： 上市公司 规模： 2000人以上
绩效奖金 注重实力 免费班车

spark大数据分析工程师 [北京]
月薪：15k-30k 经验： 不限 最低学历： 不限
职位诱惑：双休、弹性工作、年假、前景绝对很好
2015-04-16
CelLoud 投个简历
领域： 健康医疗
阶段： 初创型(未融资) 规模： 15-50人
技能培训 带薪年假 管理规范

Spark/Hadoop开发工程师 [杭州]
月薪：10k-20k 经验： 1-3年 最低学历： 大专
职位诱惑：五险一金、餐补、节假日福利、outing
2015-04-22
IN—我的生活IN记 投个简历
领域： 电子商务,社会化营销 创始人： 清水
阶段： 成长型(A轮) 规模： 50-150人
节日礼物 股票期权 年度旅游

hadoop/spark平台开发工程师 [南京]
月薪：10k-20k 经验： 3-5年 最低学历： 本科
职位诱惑：高薪 前沿技术 团队氛围
2015-04-24
苏宁 投个简历
领域： O2O 创始人： 张近东
阶段： 上市公司 规模： 2000人以上
技能培训 节日礼物 免费班车

【Spark 开发工程师】职位描述或者职位职责

58同城在线收入研发部招聘

Spark开发工程师

10k-20k 北京 经验1-3年 本科及以上 全职

职位诱惑：核心部门，待遇丰厚，气氛融洽

发布时间：2天前发布

职位描述

职位描述：
- 负责互联网广告策略、海量广告数据相关的Spark开发，偏mllib/streaming方向
- 基于Spark生态系统的广告算法平台开发
- 基于Spark进行二次开发，有机会深入学习Spark底层架构
- 配合算法工程师/数据挖掘工程师进行机器学习相关算法的开发

职位要求：- 本科及以上学历，计算机相关专业
- 扎实的Java语言基础；熟悉Scala优先，熟悉Python/Shell加分
- 熟悉Spark/Hadoop生态系统；有Spark开发经验优先；有Storm经验亦可
- 对机器学习有基本了解，了解分类、聚类，了解LR、SVM等
- 对大数据、机器学习有强烈的兴趣；对技术创新有热情和激情；具有良好的自学能力
- 有互联网从业经验，有大数据处理经验优先
- 良好的编码规范
- 能够阅读英文技术文档
- 有开源社区代码贡献的加分

大数据Spark高级工程师

北京数字家园科技有限公司

45-55万

收藏

北京-朝阳区 | 全日制统招本科 | 4年以上

发布时间：昨天

我感兴趣 请联系我

年底双薪 股票期权 带薪年假 定期体检 弹性工作

节日礼物 免费班车 领导好 扁平管理 五险一金 团队聚餐

基本信息：

工作地点： 北京-朝阳区　　所属部门：平台

汇报对象：　　下属人数： 0人

职位描述：

工作职责：

1.负责构建Spark/HDFS大数据处理架构；

2.负责基于Spark技术的海量数据的自动化分析处理和统计工作；

3.基于Spark框架大数据架构的设计、开发和维护；

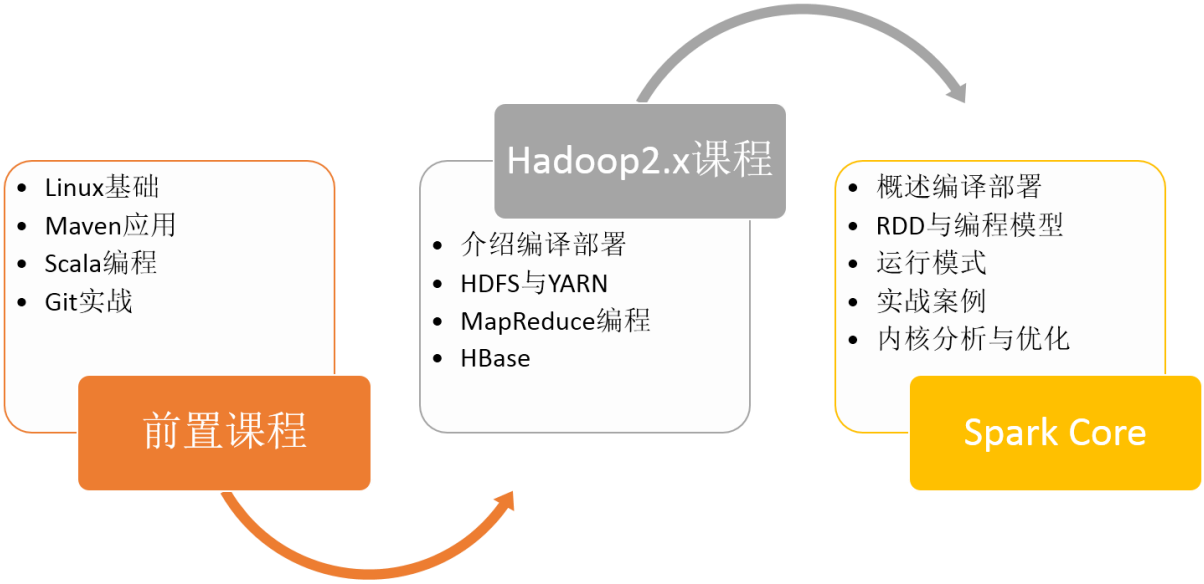
4.根据相关需求使用Spark Streaming、SQL进行数据处理、查询和统计等工作；

5.负责基于Spark MLLib和GraphX进行机器自动学习的设计和编程，实现大数据的深层次挖掘和准确业务推荐/营销。

课程优势


云帆大数据推出的国内首套企业级Spark 1.x实战课程
内容全部来自企业实战经验，全是干货，杜绝理论派
手把手教学，从零基础讲起，让你轻松转行大数据
学完本课程至少拥有两年以上Hadoop/Spark工作经验
为就业学员提供免费的企业级大数据项目的咨询与指导

课程总览


Linux基础
Maven应用
Scala编程
Git实战
前置课程
Hadoop2.x课程
介绍编译部署
HDFS与YARN
MapReduce编程
HBase
概述编译部署
RDD与编程模型
运行模式
实战案例
内核分析与优化
Spark Core


Spark Streaming
概述原理
DStream
编程模型
实战案例
优化与HA
Spark SQL
Hive和Shark
DataFrame
External Data Sources
Hive on Spark
案例实战
GraphX与MLlib
GraphX设计与RDD
基本GraphX操作
案例分析实现
机器学习算法
推荐算法案例
高级应用一
Tachyon
HistoryServer
JobServer
HBase on Spark
Phoenix on Spark
高级应用二
Redis
Zeppeline
Stream SQL
BlinkDB
实战应用
案例赏析
项目实战
如何贡献Spark PR
CM5.x和CDH5.x
就业指导

课程特色

实战性	实时性	服务性
• 以企业项目为指导	• 网络实时在线授课	• 课后专业技术辅导
• 企业一线人员授课	• 实时动态技术分享	• 定期在线实时答疑
• 从基础到应用实战	• 课程及时升级更新	• 免费就业咨询指导
• 授人以鱼授人以渔	• 密切关注技术动态	• 一次交费终身学习
• 阅读官网研习源码	• 紧跟企业场景需求	• 入职学员圆桌交流

教学大纲

开课

网络课程开学典礼，与大家进行互动交流，彼此了解，为后面讲师的授课、师生的互动做好前战准备。此外，对本套**课程大纲进行深入浅出的分析讲解**，让大家清楚的明白课程的内容，课程的主线脉络，**如何的把握重点**，哪些方面细节需要注意。

➢ **课程大纲讲解**

- 【九大阶段】总体内容说明
- 重点内容纲要讲解
- 如何进行各个阶段的学习
- 学员学习要求

➢ **学员互动交流**

- 自我介绍
- 目前从事什么工作
- 是否有一定的 Hadoop 和 Spark 基础
- 讲师答疑总结

第一阶段【前置课程】

本阶段是**整套课程的基石**，为大家从 Linux 系统的认识、安装、命令开始讲起，让大家快速进入企业大数据部署开发环境中。本阶段中对 Maven 和 Git 的讲解，使大家对企业中项目管理和版本管理有一定的认识，能够基本进行使用。Scala 编程，是目前大数据处理框架中使用较多的一个编程语言，掌握好基本的语法和使用，有助于我们对本套课程 Spark 的学习和使用。

《Linux 基础》

通过对 Linux 系统的介绍、安装、简单的命令讲解，让没有接触过 Linux 的学员轻松上手。**Linux 系统下常见的各种命令的功能使用讲解**，为学员后面对 Hadoop 2.x 和 Spark 1.x 等大数据处理框架环境搭建部署打下扎实的基础。使用“连接 Linux 系统的四大远程工具”，进一步方便学员的学习和提升实际操作能力，更加专注于 Hadoop 2.x 和 Spark 1.x 本身的使用。

➢ Linux 系统安装

- Linux 系统常见版本
- VMWare 10 虚拟软件安装
- Cents OS 6.4 版本系统虚拟机安装
- 文件权限管理
- 系统常用设置管理

➢ Linux 常用命令

- 文件和目录的创建、删除、移动、拷贝、重命名
- 文件编辑器 VI（VIM）常见命令
- 用户和组的创建、删除、更改和授权命令
- 文件权限和授权管理命令

➢ **Linux 网络管理**

- 配置 IP 地址（虚拟机中网络选择方式）
- 设置主机名称
- 关闭防火墙
- 禁用 SELinux
- 配置 Linux 启动 Level
- SSH 无密钥登陆

➢ **远程连接 Linux 方式**

- 远程命令行工具 SecureCRT
- 远程 FTP 连接工具 FileZilla
- 远程编辑工具 NotePad++或者 UE
- 远程桌面工具 Xmanager 4.x

➢ **Linux 常见软件安装**

- RPM 方式软件的安装、卸载
- JDK 安装、设置环境变量
- Eclipse 安装
- Tomcat 安装
- IDEA 安装
- MySQL 安装

《Maven 应用》

目前几乎所有的开源项目框架都使用 Maven 进行管理，在实际的企业开发中 90%以上的企业项目都也都使用 Maven 管理，主要方便管理依赖包、管理目录结构和有助于团队成员之间的交流。通过本节 Maven 相关的学习，大家可以轻松容易上手，为后面课程中 Hadoop 2.x、Spark 1.x 等编译打好基础，也为在 Linux 下使用 Eclipse 开发做好准备。

- **Maven 基本使用**

  - Maven 介绍、安装（Windows 环境）
  - 常用命令（clean、test、package、install、compile 等）
  - 命令行创建工程（Java 工程）

- **与 Eclipse、IDEA 集成**

  - 与 Eclipse 开发工具集成、配置
  - Eclipse 下 Maven 项目创建开发测试
  - 与 IDEA 集成

- **Maven 高级应用**

  - 常用插件使用（编译插件、测试插件、Jar 插件、Jetty 插件等）
  - 自动化部署
  - 仓库配置

《Scala 编程》

Scala 语言一种非常简洁优雅的编程语言，使用一种面向对象+函数式的编程语言，目前很多大数据处理框架都是使用 Scala 进行编写的项目，如 Spark、Flink、

Kafka 等，所以学好 Scala 是必须的，尤其是 Scala 中【高阶函数】是框架中常用的，为后面更好的学好用好 Spark 等框架打下扎实的根底。

- **Scala 入门**

  - Scala 环境搭建
  - 值与变量
  - 常用数据类型
  - 函数定义和使用
  - lazy
  - 条件表达式、循环与高级循环
  - 默认参数、带名参数、变长参数
  - 数组
  - 异常处理

- **Scala 面向对象**

  - 类、属性、主构造器、辅助构造器
  - 继承、方法重写、字段重写
  - object
  - apply
  - 抽象类
  - trait
  - 文件访问

- **Scala 函数式编程**

  - 集合
  - 序列
  - 可变列表与不可变列表
  - 集合操作
  - case class

- 模式匹配

➢ Scala 高级编程

- 泛型
- 隐式转换

《Git 应用》

通过本周的学习，学员将能巩固 Linux 的基本概念及常规操作、系统检查方法；加强对 Java 基本代码实现方法的理解，掌握常用的问题排查方式。通过几个简单的实践体验项目，了解通过 Git 如何进行基本的代码提交及项目协作。

➢ Git 基本使用

- 什么是 GIT
- 为什么要使用 GIT
- GIT 安装和配置
- GIT 基本流程
  - 初始化仓库
  - 添加文件
  - 添加版本
  - 推送变更
  - 创建管理分支
  - 撤销改动
- GIT 常用命令

➢ Git 实际应用

- 获取 Hadoop 2.x master branch 源码
- 分支 branch 切换

- 创建本地分支 branch
- 提交到本地仓库 repo

第二阶段【Hadoop 2.x 课程】

本阶段课程是整套【Spark 1.x】课程的基石：其一，Spark 设计是建立在 Hadoop 2.x 之上的（处理的数据存储在 HDFS；程序可运行在 YARN 集群上）；其二，Spark 中的核心思想与 MapReduce 是相通的；其三，Spark 可以很好地与 HDFS、HBase 集成。此外 Hadoop 2.x 的编译、环境搭建、HDFS Shell 使用，YARN 集群资源管理与任务监控，对于 Spark 来说非常重要，必须要掌握的。

《Apache Hadoop 2.x 编译部署》

了解 Hadoop 2.x 生态系统，如何进行编译，为后面大数据框架编译做好准备，打好基础。搭建 Hadoop 2.x 为分布式环境，熟悉 HDFS Shell 命令使用，了解 MapReduce 程序如何运行在 YARN 上以及如何进行监控。

➢ Hadoop 2.x 概述及编译

- Hadoop 2.x 概述（四大模块功能）
- Hadoop 2.x 生态系统
- 编译 Hadoop 2.x

➢ Hadoop 2.x 安装部署

- 伪分布式环境的安装部署
- 启动 HDFS、YARN
- 测试（创建文件、运行 MapReduce）
- 分布式安装部署

《HDFS 与 YARN》

了解 HDFS 设计架构，以及 Hadoop 2.x 中 HDFS HA 设计，以便深入学习后续课程。掌握应用 Application 如何运行在 YARN 之上，为后续 Spark 运行在 YARN 之上流程分析做好准备。

- **HDFS 架构**

  - HDFS 设计构架
  - 启动过程
  - HDFS HA 架构原理

- **HDFS 交互**

  - HDFS Shell 常见命令（put、ls、rm、mkdir、text 等）
  - Java API 使用

- **YARN 架构原理**

  - YARN 设计架构
  - 应用程序如何在 YARN 上运行
  - YARN 监控程序运行
  - 配置日志聚集

《分布式计算框架 MapReduce》

通过讲解 MapReduce 编程，分析执行流程，深层次掌握分布式数据处理框架处理数据的思想。**MapReduce Join 的讲解，为后续 Hive 和 Spark SQL 中的 Join 操作做好准备。**

- **MapReduce 入门**

  - 编程模型（map、reduce）
  - 编写 wordcount 程序
  - 测试与监控

- **MapReduce 执行流程**

  - 运行流程（input、map、shuffle、reduce、output）
  - MapReduce Shuffle 过程

- **MapReduce 编程**

  - 数据类型
  - MapReduce 编程模板
  - MapReduce 调优
  - MapReduce Join

《分布式 NoSQL 数据库 HBase》

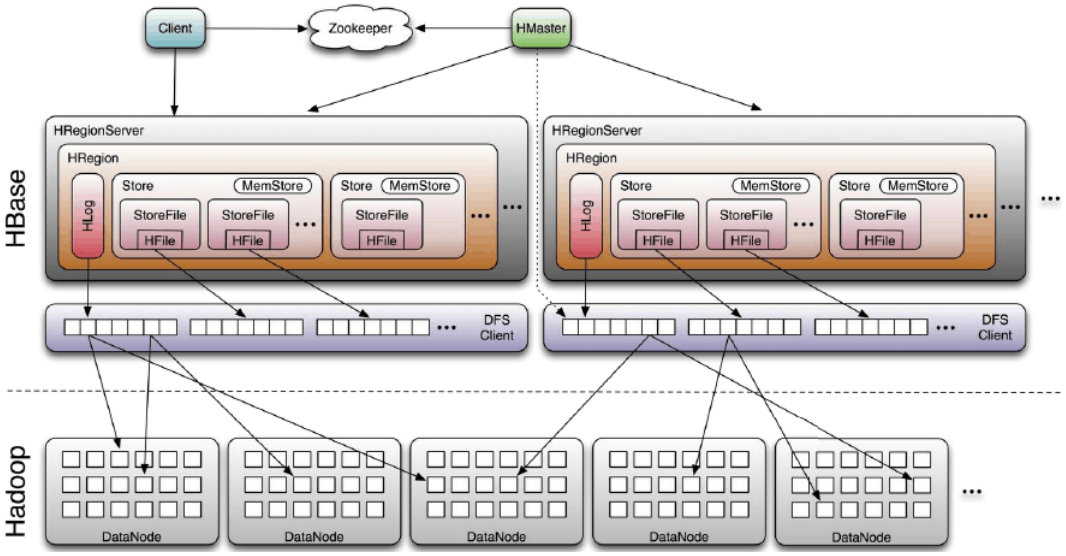
HBase 是 Hadoop 的分布式数据库，也是一种 NoSQL 数据，使用非常广泛，并且很好的众多大数据处理框架集成。此外，Zookeeper 是分布式的协作服务框架，目前大多数分布式框架都使用，必须要掌握其设计原理和集群部署。

- **Zookeeper**

  - Zookeeper 体系结构
  - Zookeeper 集群安装
  - Shell 操作

- **HBase 入门**

  - HBase 概述、架构设计



- HBase 安装部署
- HBase Shell 使用

➢ **HBase 进阶**

- Java API 使用
- 表的设计
- HBase 架构图深入剖析

第三阶段【Spark Core】

本阶段课程是 **Spark 核心和基石**，主要讲解了 Spark 1.x **功能特性、编译部署、运行模式、RDD、内核解析**，一步步深入浅出的进行理论结合实际的讲解操作。

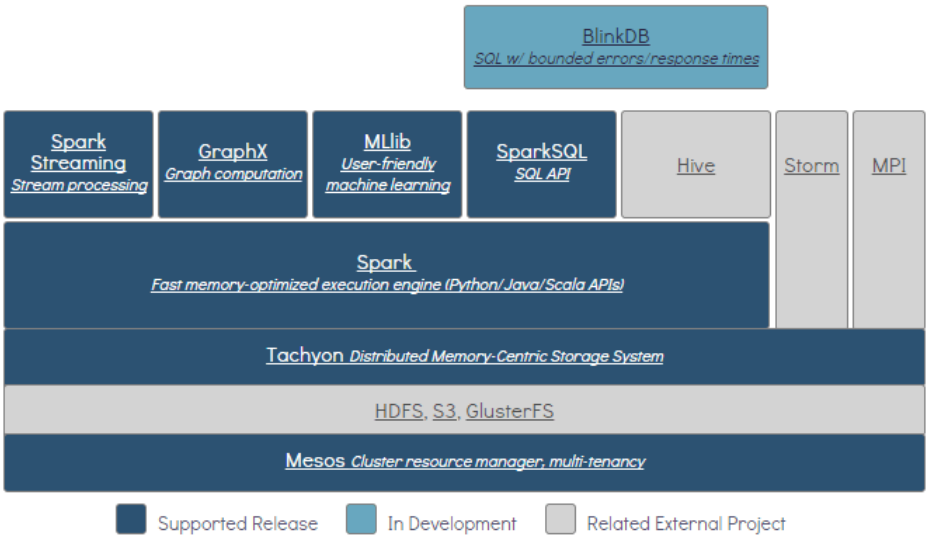
《Spark 概述、编译与环境搭建》

通过对比性的讲解 Spark 在处理大数据方面的优势，让大家明白为什么众多公司在调研测试 Spark。从 Spark 源码编译开始，逐步深层次环境部署和如何使用。

➢ **Spark 1.x 概述**

- Spark 1.x 概述及特点

- **BDAS: Berkeley Data Analytics Stack**



- 与 Hadoop 生态系统对比
- 与 MapReduce 对比
- Spark 1.x 发展与前景

➢ Spark 1.x 编译

- 使用 Maven 编译
- 使用 SBT 编译
- **使用 make-distribution.sh 编译**
- 导入源码到 IDEA 并编译

➢ Spark 1.x 环境搭建

- **Spark 1.x 环境搭建**
- Spark 运行模式（Local、Standalone、YARN 等）
- **spark-shell 与 spark-submit 的使用**
- **Spark 开发环境搭建(IDEA)**
- Standalone HA 搭建

《Spark Core》

Spark 基础的三大核心：**RDD、编程模型和内核解析，是重中之重**。

➢ RDD

- RDD 概述及特性
- RDD 常用操作（transformation、action）
- RDD 缓存策略
- Lineage 与 Checkpoint
- Broadcast Variables、Accumulators
- Dependency、Stage
- 常用类型 RDD 源码分析

➢ Spark 1.x 编程模型

- 编程模型详解
- 使用 IDEA 开发 Spark 应用程序（Maven 构建项目）
- **应用程序远程调试**
- 实战案例（日志分析统计）

➢ Spark 1.x 内核分析

- 核心组件概述（Application、Driver Program、Worker、Executor 等）
- DAGScheduler 和 TaskScheduler 详解
- Standalone 模式启动流程剖析
- Application 提交过程剖析
- Standalone 模式容错分析

➢ Spark 1.x 优化策略

- 序列化
- mapPartition 的使用

- reduceByKey 与 groupByKey 注意事项
- 合理设置 partition 的数量
- AKKA 参数设置及 GC 参数设置
- 慎用 collect
- spark.local.dir 的使用

第四阶段【Spark Steaming】

随着大数据的发展，人们对大数据的处理要求也越来越高，原有的批处理框架 MapReduce 适合离线计算，却无法满足实时性要求较高的业务，如实时推荐、用户行为分析等。 Spark Streaming 是建立在 Spark 上的实时计算框架，通过它提供的丰富的 API、基于内存的高速执行引擎，用户可以结合流式、批处理和交互试查询应用。

《Spark Streaming 基础》

讲解 Spark Streaming 的基本知识和常用操作，初步了解流式计算的原理和基本使用。

➢ Spark Streaming 概述及原理

- Spark Streaming 概述
- Spark Streaming 原理
- DStream 讲解

➢ Spark Streaming 常用操作

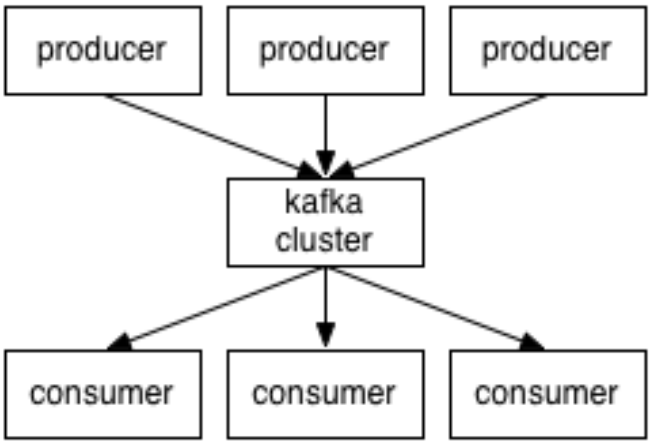
- Input DStreams and Receivers
- Transformations on DStreams
- Output Operations on Dstreams
- Caching / Persistence
- Checkpointing

《Kafka 应用》

Kafka 是一个高吞吐量的分布式消息系统，作为多种类型的数据管道（data pipeline）和消息系统使用，与 Spark Streaming 或者 Storm 等流式计算相集成使用，更好地处理数据。

- **Kafka 概述**

  - Kafka 综合概述
  - Kafka 设计架构



  - Kafka 核心概念（Topic、Producer、Consumer、Distribution）
  - Kafka 应用场景

- **Kafka 环境搭建及使用**

  - Kafka 环境部署
  - Kafka 应用

《Spark Streaming 实战案例》

结合实际应用案例，集成各种数据源，讲解 SparkStreaming 的用途。

- **Spark Streaming 处理各种数据源**

- 案例一：处理 Socket 数据源
- 案例二：处理 HDFS 数据源
- 案例三：处理 Kafka 数据源

➢ **Spark Streaming 处理 stateful 和 window 案例**

- 案例一：处理 stateful
- 案例二：处理 window

➢ 日志分析

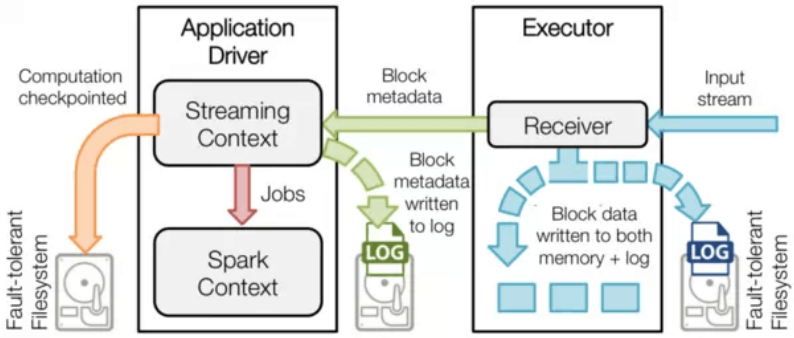
- The average, min, and max content size of responses returned from the server;
- A count of response code's returned;
- All IPAddresses that have accessed this server more than N times;
- The top endpoints requested by count

《Spark Streaming 高级特性》

Spark Streaming HA 实现和在实际生产环境中的优化策略。

➢ **Spark Streaming HA**

- 产生背景
- HA 实现原理剖析



- 如何在生产环境中配置 Streaming HA 环境

➢ Spark Streaming 优化策略

- 设置合适的并行度
- 设置合适的 batch size
- 内存优化
- 设置合理的存储级别
- 序列化

第五阶段【Spark SQL】

SQL on Hadoop 三大框架：Hive、Spark SQL、Impala，是目前大多数公司用于处理存储在 HDFS 上数据的框架，由于其使用类似于 SQL 语句方式，对数据进行查询分析，方便开发人员上手使用，尤其是熟悉 SQL 或者数据库 DBA 来说。

《分布式数据仓库 Hive》

Hive 是基于 Hadoop 的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供 sql 查询功能，Hive 将 sql 语句转换为 MapReduce 任务进行运行。 其优点是学习成本低，可以通过类 SQL 语句快速实现简单的 MapReduce 统计，也支持实现自己的 UDF 函数来完成比较复杂的业务逻辑的实现，非常适合数据仓库的统计分析。

➢ Hive 概述、环境搭建及入门

- Hive 是什么、Hive 与 Hadoop 的关系、Hive 体系架构
- Hive 与 RDBMS 的区别、Hive 实用场景以及优缺点
- Hive 环境搭建
- Hive 元数据存储、Hive 数据存储
- Hive shell 常用操作
  - hive -e
  - hive -f
  - hive -v
  - hive -i
  - hive -S

➢ Hive 常见表操作

- Hive 内部表、外部表、分区表(静态分区、动态分区)常用操作
  - 表创建
  - 数据加载
  - 数据导出
  - 内/外部表的区别以及各自在生产中的适用场景
- Hive 常用查询操作
  - select
  - where
  - distinct
  - join
  - group by
  - union
  - case when then
  - IN/NOT IN/EXISTS/NOT EXISTS
- Hive 排序
  - order by
  - sort by

- distribute by
- cluster by

- Hive 复合数据类型
  - Array
  - Map
  - Struct
- HiveServer2/beeline 使用
- Java 操作 Hive
- Hive 内置函数以及 UDF 编程
- Hive 窗口和分析函数
  - row_number
  - rank
  - dense_rank
- Hive 虚拟列以及在项目中的使用
  - INPUT__FILE__NAME
  - BLOCK__OFFSET__INSIDE__FILE

➢ Hive 常用存储格式与压缩格式

- 存储格式：TextFile/Sequenece File/RCFile/ORCFile/Parquet
- 压缩格式：gzip/bzip/lzo/snappy

➢ Hive 常用优化策略

- 合理设置 Mapper/Reducer 个数
- 合理利用压缩技术以及分布式缓存
- 充分利用多个 job 之间的共用的中间结果集
- 执行计划分析
- 数据倾斜分析及常用解决方案
- MapJoin
- 并行执行
- JVM 重用

- 分区的合理使用

➢ 案例实战

- 电商行业用户行为分析
- 电商行业商品评分统计分析

《Shark》

为 Spark 设计并开源的一款数据仓库系统，提供了分布式 SQL 查询引擎，并能够兼容 Hive；通过将 HQL 翻译成 Spark 作业并提交到 Spark 集群上运行。

➢ **Shark 概述及后续发展方向**

- Shark 产生原因
- Shark 体系架构
- Shark vs Hive
- Shark 项目终止以及后续发展的方向

《Spark SQL》

Spark SQL 的愿景：写更少的代码，读更少的数据，将优化的操作交由底层优化器去执行；DataFrame 提供了丰富多样的外部数据源支持。

➢ **Spark SQL 概述及入门**

- Spark SQL 概述
- Spark SQL 体系架构
- Catalyst 引擎
- SqlContext 与 HiveContext
- spark-sql、thriftserver/beeline

➢ **DataFrame**

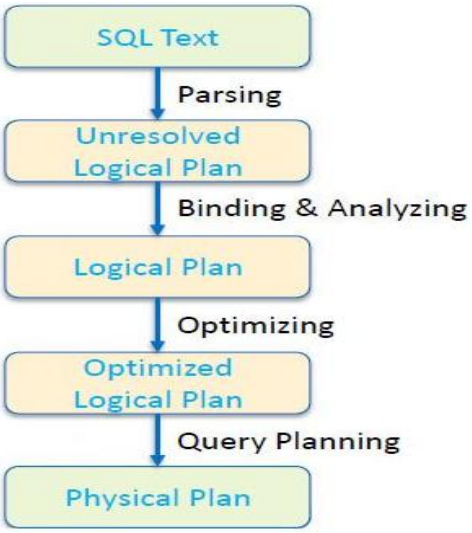
- project
- filter/where
- alias
- join
- group by
- agg
- 查看执行计划

➢ External Data Source

- RelationProvider
- BaseRelation
- 根据数据源类型选择实现不同级别的 buildScan
- JDBC 外部数据源(关系型数据库)实现过程分析及使用

➢ SparkSQL 执行流程源码剖析

- Spark SQL 执行流程分析



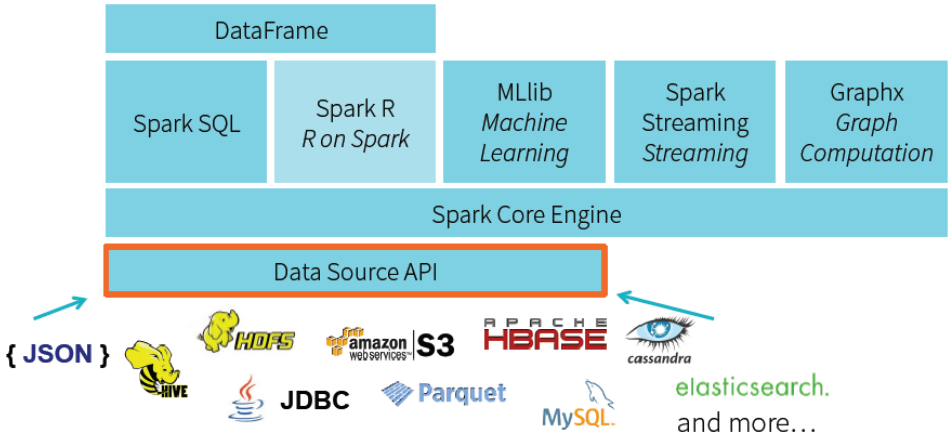
- SqlParse 分析
- Analyze Logic Plan 分析
- Optimize Logic Plan 分析

- Physical Plan 分析
- Execute 分析

《Spark SQL 案例实战》

结合具体案例，主要针对【结构化数据】进行分析统计讲解，进一步体会 Spark SQL 使用的优势。

➢ DataFrame 结合 External Data Source 案例



- 处理 json 文件
- 处理 Parquet 文件
- 处理 Hive 表数据
- 处理关系型数据库表数据
- 不同数据源格式之间的数据互操作
- 保存 DataFrame 到 jdbc 表

➢ 日志分析统计案例

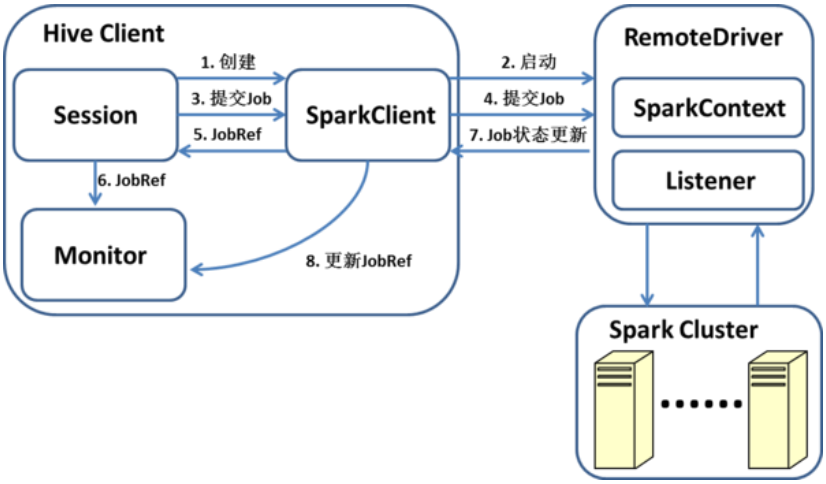
- 使用 DataFrame 完成日志分析
- 使用 Streaming 和 DataFrame 完成日志分析

《Hive on Spark》

Hive 是基于 Hadoop 平台的数据仓库，已经成为 Hadoop 事实上的 SQL 引擎标准；Hive 拥有更为广泛的用户基础以及对 SQL 语法更全面的支持；Hive 最初的计算引擎为 MapReduce，受限于其自身的 MapReduce 模式在性能上难以得到提升；Hive on Spark 目的是把 Spark 作为 Hive 的一个计算引擎，将 Hive 的查询作为 Spark 的任务提交到 Spark 集群上进行计算，利用 Spark 的特性提高 Hive 查询性能。

- **概述、架构及部署**

  - 产生背景
  - 设计原理及运行架构



  - 编译及部署
  - 与 Spark SQL 对比

- **案例实战**

  - 电商行业用户行为分析
  - 电商行业商品评分统计分析

第六阶段【GraphX 与 MLlib】

Spark GraphX 是一个分布式图处理框架，Spark GraphX 基于 Spark 平台提供对图计算和图挖掘简洁易用的而丰富多彩的接口，极大的方便了大家对分布式图处理的需求，同时还提供的 使用 Pregel 的接口，极大方便了使用。大家都知道，社交网络中人与人之间有很多关系链，例如 Twitter、Facebook、微博、微信，这些都是大数据产生的地方，都需要图计算，现在的图处理基本都是分布式的图处理，而并非单机处理，Spark GraphX 由于底层是基于 Spark 来处理的，所以天然就是一个分布式的图处理系统。

《GraphX 设计与 RDD》

由图的概念出发讲解图形计算，目前存在哪些图计算框架，GraphX 优势在哪里，在 Spark 中，核心的 RDD 以及基本计算操作。

➢ GraphX 设计

- 图的定义与应用
- GraphX 设计核心原理

➢ GraphX RDD 与基本操作

- Graph Operator（类 Graph 和 GraphOps）
- VertexRDD、EdgeRDD、EdgeTriplet
- 构建 Graph
- GraphX 五大操作
  - 转换操作（mapVertices，mapEdges，mapTriplets）
  - 结构操作（reverse，subgraph，mask，groupEdges）
  - 连接操作（joinVertices，outerJoinVertices）
  - 聚合操作（mapReduceTriplets，maxInDegree，collectNeighbors）
  - 缓存操作（cache，unpersistVertices）
- GraphX 架构

➢ **Graphx 案例**

- 图例演示（基本使用）
- PageRank 概念
- GraphX 中如何调用 PageRank（深度解析）
- 最短路径实现

《GraphX 项目实战》

通过具体的【关系】类型项目，由企业的实际需求考虑，讲解 GraphX 使用，包括如何构建图，如何使用 grepel 接口，如何编码、测试、运行、监控程序。

➢ **GraphX 实现【关系】查找**

- 业务需求分析
- 如何构建图 Graph
- 查找最短路径
- 【关系】查找与过滤

➢ **鉴赏淘宝对 GraphX 使用**

- 淘宝如何使用 GraphX
- 淘宝的“图流合璧”（Spark Streaming 与 GraphX 集成）

《Spark MLlib 入门》

Machine Learning 机器学习是指一套工具、方法或程式，使到我们可以从现实世界的海量数据里提炼出有价值的知识，规则和模式，然后将它们反哺给前台应用系统，进行预测，推荐等能产生直接经济价值的场景，给用户带来“机器具备人类般高智能”的震撼性体验。基于 MapReduce 的机器学习库 Mahout，在进行数据挖掘时，显得稍微吃力，然而由于 Spark RDD 保持在内存中，适合迭代计算的，更加方便的进行数据挖掘，Spark MLlib 就是一个机器学习库。

- **Machine Learning 机器学习**

  - 机器学习定义
  - 分类、聚类、协同过滤
  - 常用算法

- **Spark MLlib**

  - 什么是 Spark MLlib
  - Spark MLlib 设计架构
  - Spark MLlib 运行结构

- **MLlib 案例**

  - K-Means 算法介绍
  - K-Means 实例应用
  - 协调过滤算法介绍
  - 基于 Item 方式协调过滤实现
  - 决策树与组合学习

第七阶段【高级应用一】

讲解一些与 Spark 集成的大数据处理框架，包括内存文件系统 Tachyon、Spark 应用历史服务器 HistoryServer、Spark as a Service、以及 HBase 和 Phoenix on Spark 的集成，在企业实战中通常集成使用。每个框架的集成使用，均通过企业实际案例 Demo 进行讲解，让大家既复习巩固前面 Spark 知识，又清晰如何与其他框架集成使用，满足业务功能需求。

《Tachyon》

分布式内存文件存储系统，是底层的分布式文件存储（HDFS/S3 等）和上层的各种计算框架之间的一种中间件。让不同的 Job 或者框架共享数据，绕过 HDFS，以更快地速度执行；避免任务失败时的数据重算；让系统避免多次 GC 带来的开销。

- **Tachyon 概述与环境搭建**

  - 产生背景
  - 设计架构
  - 容错机制
  - 编译、环境搭建
  - Tachyon Shell

- **Tachyon 案例**

  - Java 操作 Tachyon 文件系统
  - MapReduce on Tachyon
  - Spark on Tachyon

- **Tachyon 核心源码分析**

  - Format
  - TachyonMaster
  - TachyonWorker
  - UnderFileSystem

《HistoryServer》

通过配置 HistoryServer 可以在 Application 执行的过程中记录下日志事件信息，那么在 Application 执行结束后，可以重新渲染生成 UI 界面展现出该 Application 在执行过程中的运行时信息。

- **HistoryServer 概述及使用**

  - 产生背景
  - 配置与使用
  - Jetty 入门
  - Jetty 在 HistoryServer 中的使用

- **HistoryServer 源码剖析**

  - Spark Application 运行时记录日志信息
  - HistoryServer 访问 Spark Application 运行时信息

《JobServer》

提供了一个 RESTful 接口来提交和管理 Apache Spark job、jars 及 job contexts，即 Spark as a Service。

- **JobServer 架构与使用**

  - RESTful 风格
  - Job Server 是什么
  - Job Server 架构
  - Job Server 部署
  - Job Server 案例

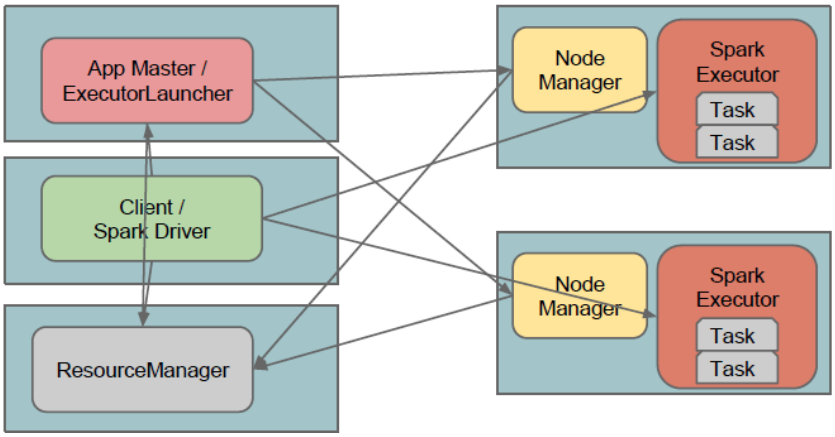
《Spark on YARN》

通过 YARN 进行共享数据，资源统一调度，进而提高集群资源的利用率，与其他框架共享集群资源。

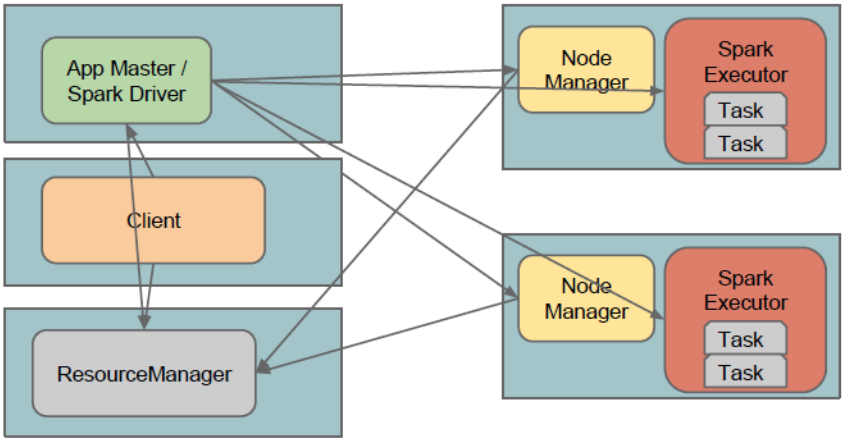
- **Spark on YARN 原理及使用**

  - Spark on YARN 原理

- yarn-client 与 yarn-cluster 的区别
  - yarn-client 调度流程



  - yarn-cluster 调度流程



- Spark on YARN 案例
- YARN 集群分配资源 Spark Application

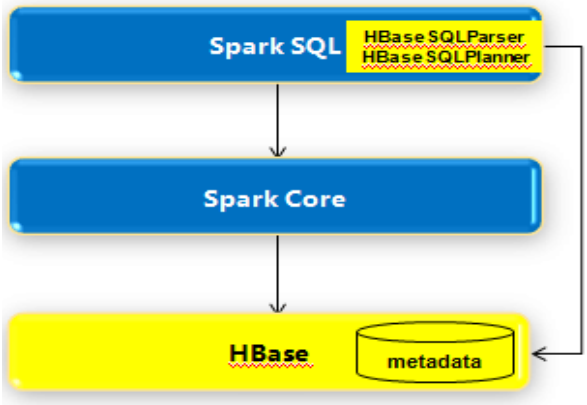
《HBase on Spark》

HBase 是一个 NoSQL 数据库，可存储海量数据，可以用 HBase shell、Java api 等方式进行操作；HBase 虽然是一个数据库，但是它提供的查询功能对比关系型数据库却是非常的薄弱；现在 Spark 想要操作 HBase 表，只能通过 TableInputFormat

等接口的方式进行访问；操作起来很不方便；**HBase on Spark 采用 Spark SQL 中的外部数据源实现了对 HBase 表的操作，直接通过类 SQL 的查询即可完成对 HBase 中的表数据进行操作**，而不再需要面向 HBase API 进行编程；

- **概述、原理及部署**

  - 实现原理剖析



  - 源码编译
  - 环境部署

- **案例实战**

  - Spark 应用程序如何访问 HBase 表
  - 通过 HBase on Spark 访问 HBase 表

《Phoenix on Spark》

Phoenix 是构建在 HBase 之上的一个 SQL 中间层，是 HBase on SQL 的一个解决方案，可以直接通过 SQL 的方式访问 HBase 中的表数据。

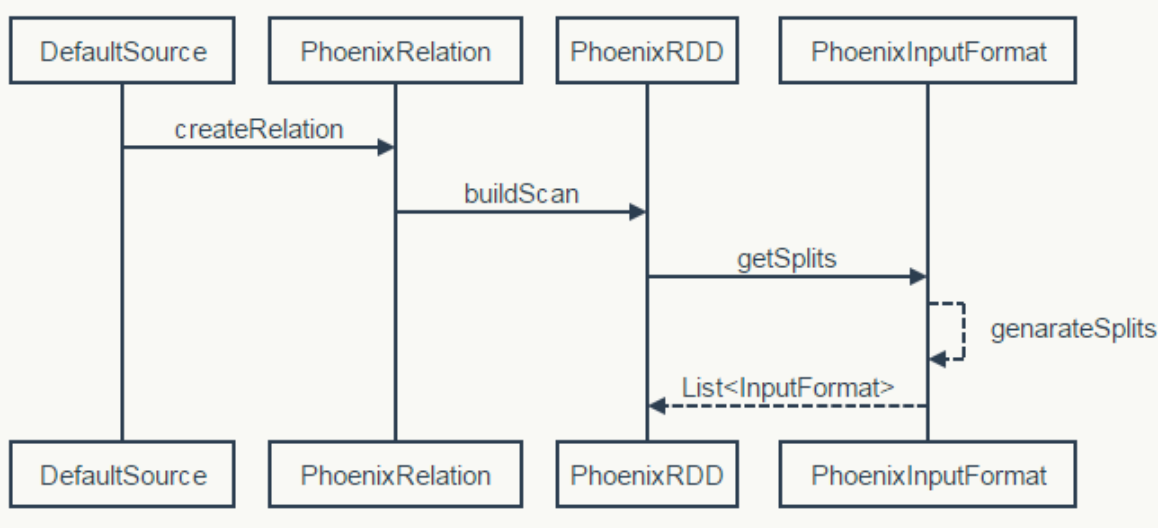
Phoenix on Spark 采用 Spark SQL 中的外部数据源实现了对 Phoenix 表的操作，**允许 Spark 通过 DataFrame 的方式对 Phoenix 表进行操作**。

- **Phoenix 入门**

- 概述、编译及部署
- 基本使用

➢ Phoenix on Spark

- 设计原理剖析



- 编译及部署

➢ 案例实战

- 通过官方的 Phoenix External Data Source API 访问 Phoenix 表
- 动手开发实现简易版本 External Data Source 访问 Phoenix 表

第八模块【高级应用二】

目前大多数电商与互联网公司，都会使用 NoSQL 类型数据库，进行数据的存储和使用，其中使用最多是 Redis 数据库，是一种基于内存的。在实际的企业中，往往通过 MapReduce 或者 Spark 处理好的数据，存储到 Redis 中，以便前端可以快速的获取数据，在页面进行展示。目前 Redis 3.0 正式版本已发布，支持集群部署，其性能和高可用性更加受到公司的亲昵。此外，围绕 Spark 发展的其他大数据处理框架不断

增多，进一步满足不同类型的业务场景，如可视图化展现的交互式数据分析工具 Zeppeline、类似 SQL 方式的 Streaming 类型实时数据分析 StreamSQL 和基于 Spark 之上的数据实时 SQL 查询框架 BlinkDB，这些框架都在都是未来与 Spark 共同完成大数据的分析和展示，在此我们从官方 Documents 和源码学习，进行调研式讲解，给大家不仅授之以鱼，还授之以渔。

《Redis》

Redis 是一个高性能的基于内存的 key-value 数据库，和 Memcached 类似，但是解决了断电后数据完全丢失的情况；支持多种不同的数据存储结构；使用 Spark Streaming 结合 Redis 完成实时流处理计算场景；

- **Redis 基础**
  - Redis 概述及特点
  - 单机版部署
  - Redis 常用数据结构的用法：字符串、散列、列表、集合、有序集合、HyperLogLog
  - Java 操作 Redis
- **Redis 高级**
  - 流水线、事务
  - 持久化：RDB、AOF
  - 主从复制、Sentinel、twemproxy
  - Redis 3.x 集群介绍及部署
- **实战案例：【电商 XXXX 网热门/人气推荐】**
  - 数据处理流向图：

- 【访问】实时统计
  - 业务描述：实时收集用户访问的各个商品信息，进行快速分析，给用户做出精准推荐热门或者人气高的相关商品。
  - 业务实现：Spark Streaming 将数据按照每秒进行统计，结果存入 Redis，然后通过 Redis API 取出最新的 N 条数据，前台结合 Ajax 技术定时更新数据，渲染页面。
- 【访问】实时报表
  - 业务描述：统计过去十分钟热门或者人气高的相关商品。
  - 业务实现：使用 Spark Streaming 的 window 函数，结果存入 Redis 库中，然后通过 Redis API 取出最新的 N 条数据，前台结合 Ajax 技术定时更新数据，渲染页面。

《Zeppelin》

基于 Web 的记事本，能够进行交互式数据分析的工具；支持将查询统计结果以可视化的方式展现。支持 Spark、Hive、Markdown、Shell 等语言。

➢ Zeppeline 使用

- 概述
  - 数据采集、数据分析、可视化和协作

- 编译及部署
- 案例实战
  - Zeppelin 处理 hive 表数据
  - Zeppelin 处理 Spark 应用程序

《StreamSQL》

使用类似关系型数据库 SQL 的方式对 Streaming 实时数据进行处理，大大降低 Streaming 的学习门槛，提高开发效率；

➢ StreamSQL 使用

- 概述
- 系统架构
- 编译及部署
- 案例实战
  - 使用 StreamSQL 处理 Kafka 数据
  - 使用 StreamSQL 关联处理多种不同数据源的关联操作

《BlinkDB》

用于在海量数据上运行交互式 SQL 查询的大规模并行查询引擎。它允许用户通过权衡数据精度来提升查询响应时间，其数据的精度被控制在允许的误差范围内

➢ BlinkDB 使用

- 综合概述
- 系统架构
- 编译及部署
- 案例

第九阶段【实战应用】

本阶段为整套课程**综合实战应用**，如何 Hadoop 2.x 和 Spark1.x 构建的大数据平台，对企业实际的需求进行深入浅出的剖析。首先，通过对***大型互联网公司 Hadoop /Spark 大数据平台架构和优化***进行赏析，学习如何依据需求构建大数据平台；其次，依据企业需求进行分析，从数据处理流程和技术架构两个角度设计，具体功能实现；再次，讲解企业中使用最多的 Hadoop 2.x 发型版本 **CDH 5.x，如何使用 CM 5.x 进行集群安装部署监控**；最后，从面试官的角度讲解**如何编写简历、如何面试**。

《Spark 案例赏析》

通过赏析大型互联网公司 Spark 大数据平台的技术架构和应用场景，进一步深入理解 Spark 用途。

- **腾讯在 Spark 上的应用和实践优化**

  - Spark 版本与集群
  - 腾讯使用 Spark 三大 Case
  - 具体业务赏析

- **Spark 平台在电信运营商的应用实践**

  - 业务分析
  - Spark Streaming 应用

- **基于 Spark on Yarn 的淘宝数据挖掘平台**

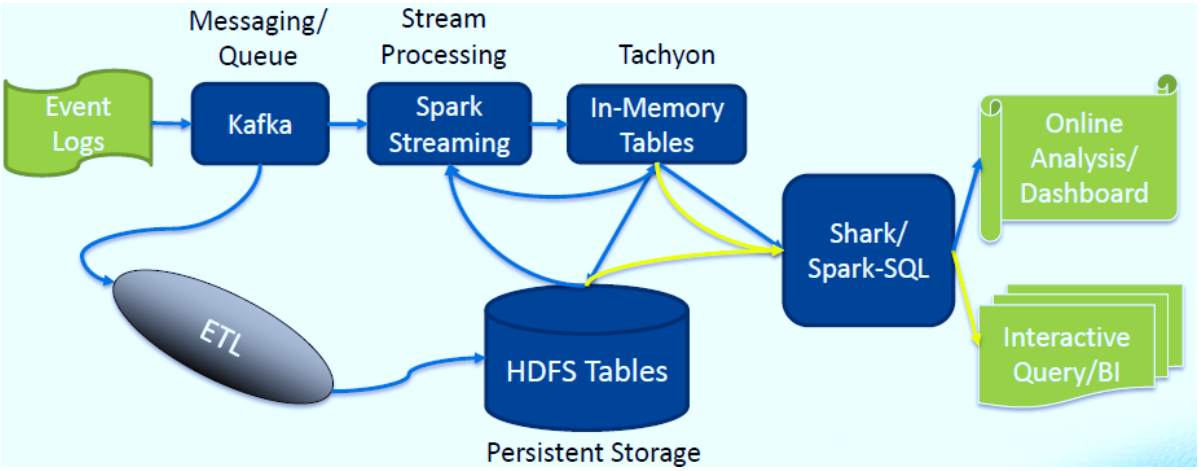
- Spark on Yarn 数据平台架构
- 淘宝推荐系统架构
- 数据挖掘 Case

《Spark 项目实战》

通过实际的项目，融会贯通整个 Spark 课程的核心知识；从数据处理流程和平台技术架构两个角度，进行具体功能分析和编码实现；通过测试，对程序和各个框架进行性能参数方面的调优，使得学员对 Spark 数据分析平台更深的认识。

➢ 项目概述

- 项目整体功能描述
- 项目技术架构



➢ Spark 数据分析平台

- 大数据平台安装部署
- 大数据平台 benchmark
- 大数据平台各个组件的基本性能调优
  - JVM GC 调优
  - 集群资源分配（YARN 管理）
  - 数据压缩
  - 序列化

- Streaming 并行度

➢ **项目业务功能实现**

- 抽取特定模型的业务进行需求分析、功能设计、代码编写
- Spark 集成 Redis、HBase 等实现
- 实时的进行统计模型业务分析实现
- 业务应用拓展
- 项目优化与总结

《如何贡献 Spark PR》

如何贡献自己的代码到 Spark 社区，成为 Spark 的 Contributor。

➢ **如何提交 Spark PR**

- fork 项目
- 创建本地仓库、加入并下载到自己的 github 代码库
- JIRA 创建 Issue
- 修改代码
- 创建 PR
- 提交 Branch

《CM 5.x 和 CDH 5.x》

Apache Hadoop 是一个开源项目，很多公司对其进行商业化，将各个大数据处理框架集成到一个组件中，其中以 Cloudera 公司的发行版本使用最广。目前众多中小型公司和一些大公司一直使用 Cloudera 发布的各个框架的版本，针对 Hadoop 2.x 来说，Cloudera 发布了 CDH 4 和 CDH 5 两大版本，目前企业中使用最多的是 CDH 5，此外 Cloudera 也发布了针对大数据集群的可视化安装部署监控的工具 Cloudera

Manager，极大方便运维工程师对集群的运维工作。本章节主要**通过实际的操作和讲解，让大家掌握 CM 5 离线安装和 CDH 5 的安装部署以及如何配置优化集群**。

- **初识 Cloudera**

  - Cloudera 公司发展
  - Cloudera Manager 版本和功能
  - CDH 版本发展（CDH 3、CDH 4 和 CDH 5）

- **Cloudera Manager 安装**

  - CM 5.x 几种安装方式
  - CM 5.x 离线安装
  - Cloudera Manager Service 安装部署

- **CDH 5.x 安装部署**

  - 使用 CM 5.x 通过 Parcels 方式安装 CDH 5.x
  - 对 HDFS、YARN 进行优化配置
  - 安装部署 CDH 5.x 中 Hive、HBase、Spark 等组件
  - 如何使用 CM 5.x 监控集群
  - CM 5.x 常见问题说明

《Spark 方向就业指导》

从 Spark/Hadoop 工程师招聘需求，逐步分析市场所需要的人才，编写【面试简历】，把握与面试官交流的重点，准确应答大数据 Spark/Hadoop 项目或技术方面面试题，最后如何选择公司和快速进入工作状态。

- **Spark 招聘需求**

  - 分析 Spark 岗位招聘需求
  - 编写简历（项目经验）

- 面试交流中的“四要四不要”原则

➢ **大数据 Spark/Hadoop 面试题讲解**

- Hadoop 2.x 相关面试题
- Spark Core 相关面试
- SQL 相关面试（Hive 和 SQL on Spark）
- 集群部署优化方面

➢ **Spark 职位与薪资**

- 公司选择（大数据职位）
- 待遇薪资
- 如何快速进入工作岗位

课程试听

试听内容大纲如下：

01 Spark 1.x 是什么以及发展状况.wmv

02 Spark 生态栈以及与 MapReduce 对比讲解.wmv

03 如何学习 Spark、几种编译方式和使用 Maven 编译讲解.wmv

04 Spark 安装部署步骤以及完成准备工作（安装 JDK、Scala、Hadoop 2.x）.wmv

05 如何使用【打包】方式编译 Spark 以及 Local 运行模式讲解(结合案例说明).wmv

06 Spark Standalone 安装部署和编写运行 WordCount 程序.wmv

试听视频讲义软件下载地址如下：

链接：http://pan.baidu.com/s/1kTkU02V 密码：h53n

课程学费

本期课程价格为：6999 元

备注说明：

1）本期课程报名前十名学员，打九五折优惠，价格为：6599 元

2）如果是《企业级 Spark 1.x 应用开发课程》学员，优惠 700 元，价格为：6299 元

3）如果是《企业级 Hadoop 2.x 应用开发课程》和《企业级 Hadoop 2.x 项目实战课程》学员，优惠 500 元，价格为：6499 元

4）三人组团，每人优惠 300 元，价格为：6699 元

授课时间

开班时间：

2015 年 5 月 18 日

学习周期：

自开课之日起，约 3 个自然月

温馨提醒：

如果本期学不会，也可以下期免费学习

付款方式

统一使用支付宝转账方式：

账号：18611869824

姓名：陈凯

温馨提示：

付款前，请与咨询顾问洽谈，确认付款账户正确性；付款后，请把付款结果截图给咨询顾问，确保对方一定收到。

云帆大数据（www.cloudyhadoop.com）

2015 年 5 月 4 日 星期一