六大真实企业大数据项目，助你成为大型项目（数据量大，并发高）的顶级架构师！

课程内容涵盖：高并发架构，hadoop,storm,spark,openstack等热门技术！

官网 http://www.bjsxt.com/html/cloud/

云计算大数据实战班

2016

# 目录

[目录 1](#_Toc450741449)

[课程说明（必读） 3](#_Toc450741450)

[课程线路 3](#_Toc450741451)

[Linux大纲 4](#_Toc450741452)

[大型网站高并发处理 5](#_Toc450741453)

[Hadoop离线计算大纲 6](#_Toc450741454)

[分布式数据库Hbase 8](#_Toc450741455)

[数据仓库Hive 9](#_Toc450741456)

[Lucene课程 10](#_Toc450741457)

[Solr/solrcloud课程 11](#_Toc450741458)

[elasticsearch分布式搜索 12](#_Toc450741459)

[CM+CDH集群管理课程 13](#_Toc450741460)

[Impala课程 13](#_Toc450741461)

[Oozie课程 14](#_Toc450741462)

[数据迁移工具Sqoop 14](#_Toc450741463)

[Flume分布式日志框架 15](#_Toc450741464)

[Zookeeper 开发 16](#_Toc450741465)

[R语言机器学习 16](#_Toc450741466)

[Mahout机器学习 17](#_Toc450741467)

[Redis缓存数据库 19](#_Toc450741468)

[Kafka分布式队列系统课程 20](#_Toc450741469)

[Storm实时数据处理 20](#_Toc450741470)

[Python课程 22](#_Toc450741471)

[Scala课程 23](#_Toc450741472)

[Spark core大数据编程 24](#_Toc450741473)

[Spark SQL 26](#_Toc450741474)

[Spark Streaming实时计算 26](#_Toc450741475)

[Spark MLlib机器学习 27](#_Toc450741476)

[Sparkt GraphX 图计算 28](#_Toc450741477)

[Tachyon 内存分布式文件系统 29](#_Toc450741478)

[Docker 课程 29](#_Toc450741479)

[虚拟化KVM 30](#_Toc450741480)

[云平台OpenStack 31](#_Toc450741481)

[大型企业项目实战 31](#_Toc450741482)

[项目一某大型电商日志分析和订单管理 31](#_Toc450741483)

[项目二某运营商基站保障平台 33](#_Toc450741484)

[项目三微博营销数据挖掘项目 33](#_Toc450741485)

[项目四智慧城市大数据分析项目 34](#_Toc450741486)

[项目五手机软件推荐系统项目 34](#_Toc450741487)

[项目六网络流量异常检测项目 35](#_Toc450741488)

[毕业学员就业薪资 35](#_Toc450741489)

[就业前景调查 40](#_Toc450741490)

[当前教室一览 49](#_Toc450741491)

[联系方式 52](#_Toc450741492)

# 课程说明（必读）

本系列课程适合有一点编程基础的人员学习（比如java,python,c/c++），最好是java编程人员，特别是从事过j2ee开发的人员。因为本套课程属于高端课程，所以报名过程中我们会有对报名人员进行简单技术评估，希望大家不要嫌麻烦。因为这样是对您最好的负责，也能更好保障您能学会并消化我们的课程（有一定难度和深度）内容，另外我们也可以给你一些比较好的复习和学习建议。学习完本套课程，可以帮助你成为大型项目架构师，特别是数据量大，并发量高的大型项目架构师，当然也能很大程度上提高你的薪资待遇。

# 课程线路

**六大真实企业大数据项目，助你成为大型项目（数据量大，并发高）的顶级架构师！**

**课程内容涵盖：**

高并发架构，hadoop,storm,spark,机器学习，docker,openstack等热门技术



# Linux大纲

这章是基础课程，帮大家进入大数据领域打好Linux基础，以便更好地学习Hadoop，NoSQL，Oracle，MySQL，Spark，Storm等众多课程。因为企业中无一例外的是使用Linux来搭建或部署项目。

1. Linux的介绍，Linux的安装：VMware Workstation虚拟软件安装过程、CentOS虚拟机安装过程
2. 了解机架服务器，采用真实机架服务器部署linux
3. Linux的常用命令：常用命令的介绍、常用命令的使用和练习
4. Linux系统进程管理基本原理及相关管理工具如ps、pkill、top、htop等的使用；
5. Linux启动流程，运行级别详解，chkconfig详解
6. VI、VIM编辑器：VI、VIM编辑器的介绍、VI、VIM扥使用和常用快捷键
7. Linux用户和组账户管理：用户的管理、组管理
8. Linux磁盘管理，lvm逻辑卷，nfs详解
9. Linux系统文件权限管理：文件权限介绍、文件权限的操作
10. Linux的RPM软件包管理：RPM包的介绍、RPM安装、卸载等操作
11. yum命令，yum源搭建
12. Linux网络：Linux网络的介绍、Linux网络的配置和维护
13. Shell编程：Shell的介绍、Shell脚本的编写
14. Linux上常见软件的安装：安装JDK、安装Tomcat、安装mysql,web项目部署

# 大型网站高并发处理

通过本章的学习大家将会了解大数据的源头，数据从何而来，继而更好的了解大数据。并且通过学习如果处理大型网站高并发问题反向更深入的学习了Linux，同时站在了更高的角度去触探了架构。

1. 第四层负载均衡
   1. Lvs负载均衡
      1. 负载算法，NAT模式，直接路由模式（DR），隧道模式（TUN）
   2. Haproxy
2. 第七层负载均衡
3. Nginx/tengine
4. Tomcat、jvm优化提高并发量
5. 缓存优化
   1. Java缓存框架
      1. Oscache，ehcache
   2. 缓存数据库
      1. Redis，Memcached
6. Lvs+nginx+tomcat+redis|memcache构建二层负载均衡千万并发处理
7. Fastdfs小文件独立存储管理

# Hadoop离线计算大纲

1. 初识hadoop

听过大数据，必听过hadoop，此部分带领大家了解hadoop的用途，在大数据中的用途，以及快速搭建一个hadoop的实验环境，在本过程中不仅将用到前面的Linux知识，而且会对hadoop的架构有深入的理解，并未你以后架构大数据项目打下坚实基础。

1. Hadoop生态环境介绍
2. Hadoop云计算中的位置和关系
3. 国内外Hadoop应用案例介绍
4. Hadoop 概念、版本、历史
5. Hadoop 核心组成介绍及hdfs、mapreduce 体系结构
6. Hadoop 的集群结构
7. Hadoop 伪分布的详细安装步骤
8. 通过命令行和浏览器观察hadoop
9. HDFS体系结构和shell以及java操作

详细剖析HDFS，从知晓原理到开发网盘的项目让大家打好学习大数据的基础，大数据之于分布式，分布式学习从学习分布式文件系统（HDFS）开始。

1. HDFS底层工作原理
2. HDFS datanode,namenode详解
3. Hdfs shell
4. Hdfs java api
5. 详细讲解Mapreduce

Mapreduce可以说是任何一家大数据公司都会用到的计算框架，也是每个大数据工程师应该熟练掌握的，此处的学习除了老师详细的讲解理论外，会通过大量的案例让大家彻底掌握。

1. Mapreduce四个阶段介绍
2. Writable
3. InputSplit和OutputSplit
4. Maptask
5. Shuffle：Sort，Partitioner，Group,Combiner
6. Reducer
7. Mapreduce案例案例
8. 二次排序
9. 倒排序索引
10. 最优路径
11. 电信数据挖掘之-----移动轨迹预测分析（中国棱镜计划）
12. 社交好友推荐算法
13. 互联网精准广告推送 算法
14. 阿里巴巴天池大数据竞赛 《天猫推荐算法》案例
15. Mapreduce实战pagerank算法
16. Hadoop2.x集群搭建

前面带领大家开发了大量的MapReduce程序，此部分将带来大家让开发的程序运行在分布式集群中，并且运行在健壮高可用的集群中。

1. Hadoop2.x集群结构体系介绍
2. Hadoop2.x集群搭建
3. NameNode的高可用性（HA）
4. HDFS Federation
5. ResourceManager 的高可用性（HA）
6. Hadoop集群常见问题和解决方法
7. Hadoop集群管理

# 分布式数据库Hbase

大数据中使用Hbase的案例多的举不胜举，也可凸显大家学习的必要性。及时工作多年的大数据工程师Hbase的优化也是需要好好学习的重点。

1. HBase定义
2. HBase与RDBMS的对比
3. 数据模型
4. 系统架构
5. HBase上的MapReduce
6. 表的设计
7. 集群的搭建过程讲解
8. 集群的监控
9. 集群的管理
10. HBase Shell以及演示
11. Hbase 树形表设计
12. Hbase 一对多 和 多对多 表设计
13. Hbase 微博 案例
14. Hbase 订单案例
15. Hbase表级优化
16. Hbase 写数据优化
17. Hbase 读数据优化

# 数据仓库Hive

Hive是使用sql进行计算的hadoop框架，工作中最常用到的部分，也是面试的重点，此部分大家将从方方面面来学习Hive的应用，任何细节都将给大家涉及到。

1. 数据仓库基础知识
2. Hive定义
3. Hive体系结构简介
4. Hive集群
5. 客户端简介
6. HiveQL定义
7. HiveQL与SQL的比较
8. 数据类型
9. 外部表和分区表
10. ddl与CLI客户端演示
11. dml与CLI客户端演示
12. select与CLI客户端演示
13. Operators 和 functions与CLI客户端演示
14. Hive server2 与jdbc
15. 用户自定义函数（UDF 和 UDAF）的开发与演示
16. Hive 优化

# Lucene课程

在大数据里面文本数据的搜索是很重要的一块，特别是里面的分词技术，是后面机器学习里面文本挖掘的基石，我们需要深入学习java领域里面的搜索核心技术lucene，同时也可以了解到百度 google这样的搜索系统是怎么架构实现的。

1. Lucene介绍
2. Lucene 倒排索引原理
3. 建索引 IndexWriter
4. 搜索 IndexSearcher
5. Query
6. Sort和 过滤 （filter）
7. 索引优化和高亮

# Solr/solrcloud课程

接着前面lucene技术搜索，如果把lucene技术比如为发动机，那solr就是一两成型的汽车了。学习完solr可以帮助你在企业里面快速的架构搜索系统。首先Solr是基于Lucene做的，Lucene是一套信息检索工具包，但并不包含搜索引擎系统，它包含了索引结构、读写索引工具、相关性工具、排序等功能，因此在使用Lucene时你仍需要关注搜索引擎系统，例如数据获取、解析、分词等方面的东西。而Solr的目标是打造一款企业级的搜索引擎系统，因此它更接近于我们认识到的搜索引擎系统，它是一个搜索引擎服务，通过各种API可以让你的应用使用搜索服务，而不需要将搜索逻辑耦合在应用中。而且Solr可以根据配置文件定义数据解析的方式，更像是一个搜索框架，它也支持主从、热换库等操作。还添加了飘红、facet等搜索引擎常见功能的支持。

1. 什么是solr
2. 为什么工程中要使用solr
3. Solr的原理
4. 如何在tomcat中运行solr
5. 如何利用solr进行索引与搜索
6. solr的各种查询
7. solr的Filter
8. solr的排序
9. solr的高亮
10. solr的某个域统计
11. solr的范围统计
12. solrcloud集群搜索详解
13. solr+hbase大型搜索系统架构

# elasticsearch分布式搜索

Elastic Search 是一个基于Lucene构建的开源，分布式，RESTful搜索引擎。设计用于云计算中，能够达到实时搜索，稳定，可靠，快速，安装使用方便。支持通过HTTP使用JSON进行数据索引。

1. elasticsearch简介
2. elasticsearch和solr的对比
3. elasticsearch安装部署
4. elasticsearch service wrapper启动插件
5. 使用curl操作elasticsearch索引库
6. elasticsearch DSL查询
7. elasticsearch批量查询meet
8. elasticsearch批量操作bulk
9. elasticsearch插件介绍
10. elasticsearch配置文件详解
11. java操作elasticsearch
12. elasticsearch的分页查询
13. elasticsearch中文分词工具的集成
14. elasticsearch优化
15. elasticsearch集群部署
16. elasticsearch+hbase大型搜索系统架构

# CM+CDH集群管理课程

Cloudera 的定位在于 Bringing Big Data to the Enterprise with Hadoop

Cloudera为了让Hadoop的配置标准化，可以帮助企业安装，配置，运行hadoop以达到大规模企业数据的处理和分析。由cloudera公司开发的集群web管理工具cloudera manager(简称CM)和CDH目前在企业中使用的比重很大，掌握CM+CDH集群管理和使用 不仅简化了集群安装、配置、调优等工作，而且对任务监控、集群预警、快速定位问题都有很大的帮助。

1. CM + CDH集群的安装
2. 基于CM主机及各种服务组件的管理
3. CDH集群的配置和参数调优
4. CDH集群HA配置及集群升级
5. CM的监控管理
6. 集群管理的注意事项
7. HUE实战详解

# Impala课程

Impala是Cloudera公司参照 Google Dreme系统进行设计并主导开发的新型查询系统，它提供复杂SQL语义，能查询存储在Hadoop的HDFS和HBase中的PB级大数据。

1. Impala介绍和架构
2. Impala实战安装，架构，外部shell
3. Impala内部shell,存储分区，SQL
4. Impala SQL、hbase整合，JDBC、性能优化
5. Impala配置及其调优
6. Impala项目应用

# Oozie课程

oozie是Hadoop生态系统中比较重要的框架，在系统中扮演的角色是工作流引擎服务器，用于运行Hadoop Map/Reduce任务工作流（包括MR、Hive、Pig、Sqoop等），企业中多由于整个项目流程任务调度。

1. Oozie入门介绍
2. Oozie安装配置及其简单操作
3. hPDL语言学习及流程定义
4. oozie工作流配置及元数据库定义
5. oozie定时任务调度和oozie API操作

# 数据迁移工具Sqoop

sqoop适用于关系型数据库和HDFS分布式数据系统之间进行数据转换，在企业中，是构建数据仓库的一大工具。

1. 介绍 和 配置Sqoop
2. Sqoop shell使用
3. Sqoop-import
   1. DBMS 🡪 hdfs
   2. DBMS 🡪 hive
   3. DBMS 🡪 hbase
4. Sqoop-export

# Flume分布式日志框架

Flume最早是Cloudera提供的日志收集系统，目前是Apache下的一个孵化项目，Flume支持在日志系统中定制各类数据发送方，用于收集数据。大家学习完此节后不但可以掌握Flume的使用，而且可以进行对于Flume的开发。

1. flume简介-基础知识
2. flume安装与测试
3. flume部署方式
4. flume source相关配置及测试
5. flume sink相关配置及测试
6. flume selector 相关配置与案例分析
7. flume Sink Processors相关配置和案例分析
8. flume Interceptors相关配置和案例分析
9. flume AVRO Client开发
10. flume 和kafka 的整合

# Zookeeper 开发

  Zookeeper在分布式集群（Hadoop生态圈）中的地位越来越突出，对分布式应用的开发也提供了极大便利，这也是这里我们带领大家深入学习 Zookeeper的原因。本课程主要内容包括Zookeeper深入、客户端开发（Java编程，案例开发）、日常运维、Web界面监控。大家这里学好Zookeeper，对后面学习其他技术至关重要。

1. zookeeper架构
2. zookeeper实战环境
3. zookeeper内部算法详解
4. Zookeeper java api开发
5. Zookeeper rmi高可用分布式集群开发
6. Zookeeper实现SOA高可用架构框架
7. Netty 异步io通信框架
8. Zookeeper实现netty分布式架构的高可用
9. Zookeeper分布式锁实现

# R语言机器学习

R本身是一款十分优秀的数据分析和数据可视化软件，同时作为第一代机器学习的工具，其中包括大量用于机器学习的添加包。此部分带领大家学习R语言更是带领大家进入机器学习的领域，机器学习算法为主线的同时，通过案例学习将会让大家对内容脉络掌握的更加清晰。

1. R语言介绍，基本函数，数据类型
2. 线性回归
3. 朴素贝叶斯分类
4. 决策树分类
5. k均值聚类
   1. 离群点检测
6. 关联规则探索
7. 神经网络

# Mahout机器学习

Mahout提供一些可扩展的机器学习领域经典算法的实现，很多公司会使用Mahout方便快捷地创建智能应用程序。Mahout包含许多实现，包括聚类、分类、推荐过滤、频繁子项挖掘。Mahout通过使用 Apache Hadoop，可以有效地扩展到云中。被业界奉为第二代机器学习工具。此部分过后大家不仅会学习到mahout的组件而且会有项目让大家真正把它应用到工作中。

1. 介绍为什么使用它,它的前景
2. 简单介绍Mahout
3. 简单介绍机器学习
4. 实例演示Mahout单机推荐程序
5. 配置安装（hadoop2.x版本的）编译安装步骤说明
6. 命令行中测试运行协同过滤概念
7. 推荐
8. 讲解基于用户的协同过滤
9. 讲解基于物品的协同过滤
10. 分类
11. 分类概念
12. 分类的应用及Mahout分类优势
13. 分类和聚类、推荐的区别
14. 分类工作原理
15. 分类中概念术语
16. 分类项目工作流
17. 如何定义预测变量
18. 线性分类器的介绍，及贝叶斯分类器
19. 决策树分类器的介绍，及随机森林分类器
20. 如何使用贝叶斯分类器和随机森林分类器的代码展示
21. 聚类
22. 聚类概念
23. 聚类步骤流程
24. 聚类中的距离测度
25. 讲解K-means聚类
26. K-means聚类算法展示
27. 聚类其他算法
28. 介绍TF-IDF
29. 归一化
30. 微博聚类案例

# Redis缓存数据库

1. redis特点、与其他数据库的比较
2. 如何安装redis
3. 如何使用命令行客户端
4. redis的字符串类型
5. redis的散列类型
6. redis的列表类型
7. redis的集合类型
8. 如何使用java访问redis【a.python访问redis,scala访问redis】
9. redis的事务(transaction)
10. redis的管道(pipeline)
11. redis持久化(AOF+RDB)
12. redis优化
13. redis的主从复制
14. redis的sentinel高可用
15. twemproxy,codis实战
16. redis3.x集群安装配置

# Kafka分布式队列系统课程

Kafka是当下流行的队列，可以说是从数据采集到大数据计算承上启下的重要环节，大家在此部分将会详细学习它的架构，kafka在大家大数据的项目中几乎都会涉及到。

1. kafka是什么
2. kafka体系结构
3. kafka配置详解
4. kafka的安装
5. kafka的存储策略
6. kafka分区特点
7. kafka的发布与订阅
8. zookeeper协调管理
9. java编程操作kafka
10. scala编程操作kafka
11. flume 和kafka 的整合
12. Kafka 和storm 的整合

# Storm实时数据处理

本部分学习过后，大家将全面掌握Storm内部机制和原理，通过大量项目实战，让大家拥有完整项目开发思路和架构设计，掌握从数据采集到实时计算到数据存储再到前台展示，所有工作一个人搞定！譬如可以一个人搞定淘宝双11大屏幕项目！不光从项目的开发的层次去实现，并可以从架构的层次站在架构师的角度去完成一个项目。

1. Storm的基本概念
2. Storm的应用场景
3. Storm和Hadoop的对比
4. Storm集群的安装的linux环境准备
5. zookeeper集群搭建
6. Storm集群搭建
7. Storm配置文件配置项讲解
8. 集群搭建常见问题解决
9. Storm常用组件和编程API：Topology、 Spout、Bolt
10. Storm分组策略(stream groupings)
11. 使用Strom开发一个WordCount例子
12. Storm程序本地模式debug、Storm程序远程debug
13. Storm事物处理
14. Storm消息可靠性及容错原理
15. Storm结合消息队列Kafka：消息队列基本概念(Producer、Consumer、Topic、Broker等)、消息队列Kafka使用场景、Storm结合Kafka编程API
16. Storm Trident概念
17. Trident state 原理
18. Trident开发实例
19. Storm DRPC(分布式远程调用)介绍
20. Storm DRPC实战讲解
21. Storm和Hadoop 2.x的整合：Storm on Yarn
22. 淘宝核心架构套件
    1. Jstorm+rocket详解

Storm开发实战： flume+Kafka+Storm+Hbase+redis项目实战，以及多个案例

# Python课程

Python语言的部分大家在学习后可以完全掌握Python的精髓，并通过这部分的学习给大家打好一个基础，在其他计算框架中多语言的使用上都会涉及到Python这门流行的语言。同时课程里会通过机器学习的案例让大家学习Python的同时去更好的理解机器学习。

1. 介绍Python以及特点
2. Python的安装
3. Python基本操作（注释、逻辑、字符串使用等）
4. Python数据结构（元组、列表、字典）
5. 使用Python进行批量重命名小例子
6. Python常见内建函数
7. 更多Python函数及使用常见技巧
8. 异常
9. Python函数的参数讲解
10. Python模块的导入
11. Python中的类与继承
12. 网络爬虫案例
13. 数据库连接，以及pip安装模块
14. Mongodb基础入门
15. 讲解如何连接mongodb
16. Python的机器学习案例

# Scala课程

在此部分内，将更注重scala的各种语言规则与简单直接的应用，而不在于其是如何具体实现，通过学习本课程能具备初步的Scala语言实际编程能力。本部分课程也可以视为大家下面学习Spark课程的铺垫，供大家扫盲熟悉Scala，提前进行热身运动。

1. scala解释器、变量、常用数据类型等
2. scala的条件表达式、输入输出、循环等控制结构
3. scala的函数、默认参数、变长参数等
4. scala的数组、变长数组、多维数组等
5. scala的映射、元组等操作
6. scala的类，包括bean属性、辅助构造器、主构造器等
7. scala的对象、单例对象、伴生对象、扩展类、apply方法等
8. scala的包、引入、继承等概念
9. scala的特质
10. scala的操作符
11. scala的高阶函数
12. scala的集合
13. scala数据库连接

# Spark core大数据编程

本部分内容全面涵盖了Spark生态系统的概述及其编程模型，深入内核的研究，Spark on Yarn,Spark Streaming流式计算原理与实践，Spark SQL,Spark的多语言编程以及SparkR的原理和运行。不仅面向项目开发人员，甚至对于研究Spark的学员，此部分都是非常有学习指引意义的课程。

1. Spark介绍
2. Spark应用场景
3. Spark和Hadoop MR、Storm的比较和优势
4. RDD
5. Transformation
6. Action
7. Spark计算PageRank
8. Lineage
9. Spark模型简介
10. Spark缓存策略和容错处理
11. 宽依赖与窄依赖
12. Spark配置讲解
13. Spark集群搭建
14. 集群搭建常见问题解决
15. Spark原理核心组件和常用RDD
16. 数据本地性
17. 任务调度
18. DAGScheduler
19. TaskScheduler
20. Spark源码解读
21. 性能调优
22. Spark和Hadoop2.x整合：Spark on Yarn原理
23. Spark Core核心编程
24. RDD内核架构概览
25. RDD的不同数据来源的创建方式详解
26. RDD的操作算子综述与本质分析（转换算子、行动算子）
27. 常用操作算子的案例实战
28. RDD持久化实战以及Checkpoint
29. RDD共享变量以及累加器的使用实战
30. RDD简单排序功能（优化之前WordCount程序）以及二次排序的实战
31. Spark实战Top N功能详解
32. Spark任务调度流程整体架构分析详解
33. 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）
34. Spark执行任务相关原理以及源码分析（TaskScheduler、Executor、Task、Shuffle）
35. Spark实战之PageRank
36. 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

# 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

前面课程大家已经掌握第一代机器学习工具R，而后又学习了第二代机器学习工具Mahout，这里大家将会学习第三代机器学习工具MLlib，大家不仅将会了解MLlib的组件及其调用，而且会通过Spark的项目深入了解MLlib的现实使用。通过此部分大家也可以看出课程不仅着眼于现在，更是着眼于大家的未来在行业中的发展。

1. 介绍
   1. Spark MLlib组件介绍
   2. 基本数据类型
2. 回归算法
   1. 广义线性模型
   2. 逻辑回归
3. 分类算法
   1. 朴素贝叶斯
   2. 决策树
   3. 随机森林
4. 第四章 推荐系统
   1. 协同过滤
5. 第五章 聚类
   1. Kmeans
   2. Sparse kmeans
   3. Kmeans++
   4. Kmeans II
   5. Streaming kmeans
   6. Gaussian Mixture Model

# Sparkt GraphX 图计算

* 1. 二分图
  2. 概述
  3. 构造图
  4. 属性图
  5. PageRank

# 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

# Docker 课程

Docker 是一个开源的应用容器引擎，让开发者可以打包他们的应用以及依赖包到一个可移植的容器中，然后发布到任何流行的 Linux 机器上，也可以实现虚拟化。容器是完全使用沙箱机制，相互之间不会有任何接口（类似 iPhone 的 app）。几乎没有性能开销,可以很容易地在机器和数据中心中运行。最重要的是,他们不依赖于任何语言、框架包括系统。

1. 基本介绍
2. vm docker 对比
3. docker基本架构介绍
4. unfs cgroup namespace
5. 进程虚拟化 轻量级虚拟化
6. docker 安装
7. docker 镜像制作
8. docker 常用命令
9. docker 镜像迁移
10. docker pipework【i.openvswitch】
11. docker weave

# 虚拟化KVM

云计算算平台，比如openstack，cloudstack 底层采用的技术都是虚拟化方案，现在以kvm市场占有率最高，我们要深入的去学习这些原生的虚拟化，才能深入了解和架构openstack这样的云计算的平台，也才更有能力去开发自己的云计算平台。

1. 虚拟化介绍，虚拟化适用场景等等
2. Qemu Libvirt & KVM
3. 安装KVM, Qemu, Libvirt
4. QEMU-KVM: 安装第一个能上网的虚拟机
5. Kvm虚拟机 nat,网桥基本原理
6. kvm虚拟机克隆
7. kvm虚拟机vnc配置
8. kvm虚拟机扩展磁盘空间
9. Kvm快照
10. Kvm 迁移
11. Java,python,c语言编程控制kvm
12. 构建自己的虚拟云平台

# 云平台OpenStack

  在实战中学习，课程绝不纸上谈兵，通过搭建和调试一个真实的openstack平台，深入浅出， 详细讲解openstack的各个组成模块：keystone, glance, nova, cinder, neutron, horizen。课程中遇到的各种实际问题，不仅演示了如何解决，更是要教会大家学会去找到解决问题的方法。难点问题全面讲解。在云计算的各种技术当中，网络部分是最难，也是最复杂多样的。课程中针对虚拟网络进行了详细的讲解，包括基本原理，以及实际环境搭建，问题的跟踪和解决。讲师拥有丰富的移动集团工作经验，负责云平台的各方面工作，讲课内容完全贴近企业需求，绝不纸上谈兵。

1. openstack介绍和模块基本原理分析
2. openstack多节点安装部署【a.采用centos6.x系统】
3. Keystone基本原理
4. glance
5. Cinder
6. Swift
7. Neutron
8. Openstack api 二次开发

# 大型企业项目实战

## 项目一某大型电商日志分析和订单管理

架构技术：CDH,flume、sqoop、HDFS,Mapreduce，hive，hbase，zookeeper，mysql，oozie等等）。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。电商日志分析包括：pv、uv，跳出率，二跳率、广告转化率等。订单模块有：商家排名，历史订单查询，订单报表统计等。

* 1. Web项目和云计算项目的整合
  2. Flume通过avro实时收集web项目中的日志
  3. 数据的ETL
  4. Hive 批量 sql执行
  5. Hive 自定义函数
  6. Hive和hbase整合。
  7. Hbase 数据支持 sql查询分析
  8. Mapreduce数据挖掘
  9. Hbase dao处理
  10. Sqoop 在项目中的使用。
  11. Mapreduce 定时调用和监控
  12. Flume通过avro实时收集web项目中的日志
  13. Hive批量sql执行
  14. Hive自定义函数
  15. HBase数据支持sql查询分析
  16. MapReduce数据挖掘
  17. Mapreduce定时调用和监控
  18. IP流量分析
  19. 会话分析
  20. 浏览器pv分析
  21. 地域信息分析
  22. 订单分析
  23. 事件分析
  24. MR,hive,impala任务oozie部署

## 项目二某运营商基站保障平台

一个市级运营商公司，每天的产生海量话务数据（一线城市更高），通过大数实时分析，监控每个基站的掉话率，基站通话总数，基站掉话总数，基站告警，3g/4g上网流量实时监控。对以上维度进行实时分析以达到对基站工作情况的监控。

技术架构：Storm+hbase+kafka+flume+echarts

1. flume实时采集日志
2. kafka缓冲队列
3. storm实时处理
4. Hbase dao存储处理结果
5. 前端Web实时展示报表

## 项目三微博营销数据挖掘项目

使用数据来自微博平台，项目目标通过机器学习所学知识挖掘目标客户群体，找到代言人进行微博营销广告投放。

1. 分布式平台 Hadoop，MapReduce
2. 数据采集 Flume
3. 数据清洗 ETL
4. 数据库 Hbase，Redis
5. 机器学习 Mahout

## 项目四智慧城市大数据分析项目

城市中每时每刻都会产生海量数据，应用数据挖掘、机器学习和可视化技术，分析出的数据可以改进城市规划，缓解交通拥堵，抓捕罪犯。项目会使用真实的数据。

涉及到所学知识如下：

1. 实时流处理 Kafka，Spark Streaming
2. 分布式运算 Hadoop，Spark
3. 数据库 Hbase，Redis
4. 机器学习 Spark Mllib
5. 前台web展示数据 Struts2，echart

## 项目五手机软件推荐系统项目

使用数据来自某互联网平台手机助手，项目目标通过机器学习所学知识挖掘平台手机用户喜好，给用户准确推荐手机软件，类似360手机助手、华为手机助手、百度手机助手推荐功能。

1. 分布式平台 Hadoop，Spark
2. 数据清洗 Hive
3. 数据分析 R RStudio
4. 推荐服务 Dubbox
5. 规则过滤 Drools
6. 机器学习 MLlib

## 项目六网络流量异常检测项目

项目目标通过机器学习所学知识检测出异常，包括检测欺诈，网络攻击，服务器传感器故障灯（本项目用户现在热门的电商网站的流量分析检测，比如京东，天猫，淘宝等）。项目功能应用于各大互联网平台中，各大互联网平台均需要网络安全予以重视和预防以及检测。

1. 数据存储 Hadoop
2. 数据准备 Spark
3. 数据分析 R RStudio
4. 机器学习 MLlib
5. 数据评价及调参

# 毕业学员就业薪资

P.S. 由于现在公司内部同事之间薪资待遇都是保密的，避免就业信息给已就业学员带来不必要的麻烦，公布的就业信息会有部分信息暂时隐藏。

薪资为北上广深一线城市薪资, 非一线城市不计入薪资。

下列表格内待遇没特殊说明都有五险一金，这是现在公司基本福利。有多家公司是说明当时这位同学拿了好几家公司的Offer。

2015年部分学员就业薪资一览表格：

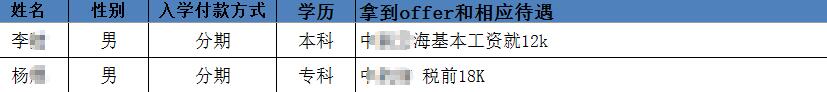




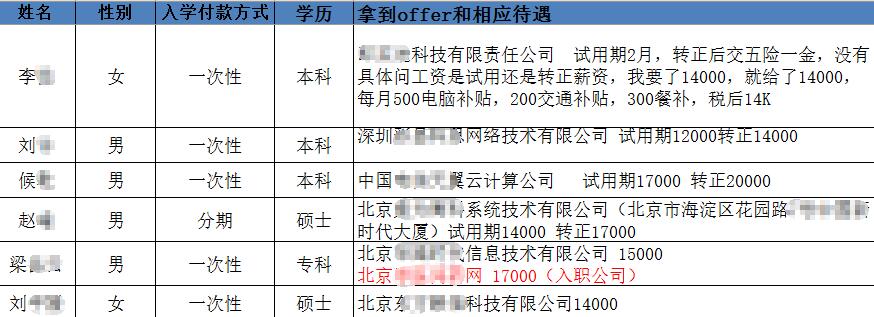












# 就业前景调查





































# 当前教室一览







# 联系方式

北京尚学堂官网：http://www.bjsxt.com/html/cloud/

QQ讨论群：172599077,156927834