CS502大数据工程师直通车课程大纲（第九版）

资深大数据工程师带你从Data Infrastructure 和 大数据分析应用两个角度双管齐下，全面提升背景，赢取心仪offer.

课时安排

【第一节课】（课程主页免费注册）

2018年9月8日 7:00 pm（PST）
2018年9月8日 10:00 pm（EST）
2018年9月9日 10:00 am（北京时间）

【课程安排】

课程长度： 12周

开课时间： 9/14/2018 - 12/6/2018 (美国时间)

课程时长： 8小时/周

授课语言： 中文

课程名称	美西时间	美东时间	北京时间
理论学习	周五 7:00 pm - 9:00 pm	周五 10:00 pm - 12:00 am	周六 10:00 am - 12:00 pm
项目实战 I	周六 7:00 pm - 9:00 pm	周六 10:00 pm - 12:00 am	周日 10:00 am - 12:00 pm
项目实战 II	周一 7:00 pm - 9:00 pm	周一 10:00 pm - 12:00 am	周二 10:00 am - 12:00 pm
作业讲解与面试指导	周四 7:00 pm - 9:00 pm	周四 10:00 pm - 12:00 am	周五 10:00 am - 12:00 pm

课程负责人

课程组老师：Davy

- ❏ Career Consultation 项目负责人
- ❏ 擅长面试准备与交流技巧
- ❏ 辅导上百名学员找准求职方向

联系方式

微信账号： adadazz

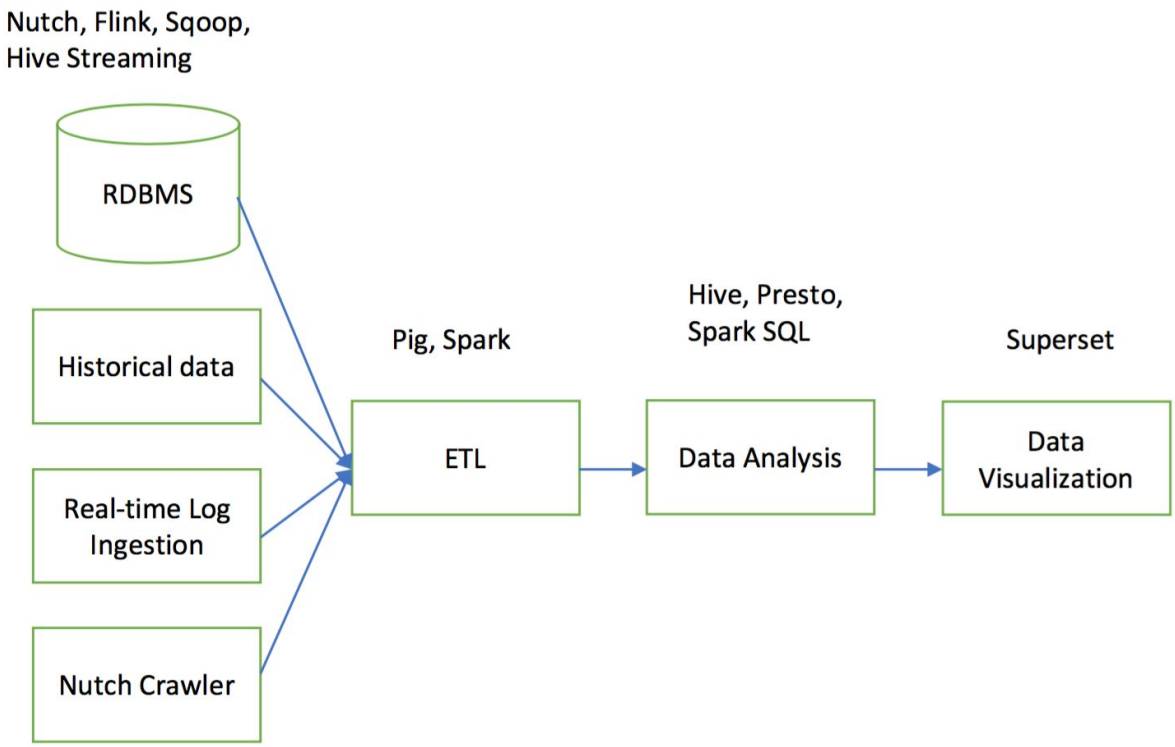
电子邮件： davy@bittiger.io

Project 1: Real-time Log Analysis System

【项目介绍】

本项目我们将着重训练同学们对Hadoop Ecosystem以及Apache系列软件的应用及开发能力。我们将从Hadoop生态系统的主要项目介绍开始，从使用Nutch搭建分布式爬虫开始入手，逐渐建立对Hadoop的更深层理解，能清楚理解Hadoop能解决什么问题。同时课程将教会学生在面对大规模且实时的数据时，如何熟练使用多种大数据处理分析工具，从不同数据源(包括爬虫捕捉的数据，实时导入的数据，历史数据，关系型数据库)中采集数据，并实现ETL，数据分析和数据可视化等分布加工及处理。

【项目图示】



【学习成果】

1. 了解Hadoop生态系统，熟悉并掌握Hadoop及MapReduce的原理、构成以及基本操作和编程
2. 学习并掌握基于Nutch的网络爬虫的搭建，具备使用AWS将其运行的能力，并能够针对其源代码级故障进行分析

3. 了解Apache开源社区以及开源软件开发流程，对Apache主要项目，如Spark，Hive，Flink，Presto等从原理到实战全方面的学习及掌握
4. 具备对实时大规模数据源的处理能力，能够根据数据源的不同类型，选择并熟练应用Storm、Flink、Sqoop、Oozie、Hive Streaming等常见大数据工具
5. 深入理解并应用Hive、Presto和Spark SQL对大规模数据进行分析处理，并应用Superset生成数据图表，完成数据可视化

Week 1 课程安排

【理论理解】

课程内容
理解Hadoop原理和构成
深入理解HDFS
深入理解Yarn
深入理解MapReduce
了解Hadoop生态系统

【项目实战】

课程内容
Hadoop系统搭建
HDFS基本操作
MapReduce编程
使用Nutch搭建网络爬虫
Nutch源代码级故障分析
使用AWS运行Nutch分布式爬虫

Week 2 课程安排

【理论理解】

课程内容
大规模数据处理简介
Apache和开源软件开发
Pig功能与内部原理
Spark功能与内部原理

【项目实战】

课程内容
用Hadoop进行数据分析
用Pig进行ETL
用Spark进行ETL
用Sqoop从关系型数据库导入数据
用Oozie协调工作流

Week 3 课程安排

【理论理解】

课程内容
Hive功能与内部原理
Presto简介
Spark SQL功能与原理

【项目实战】

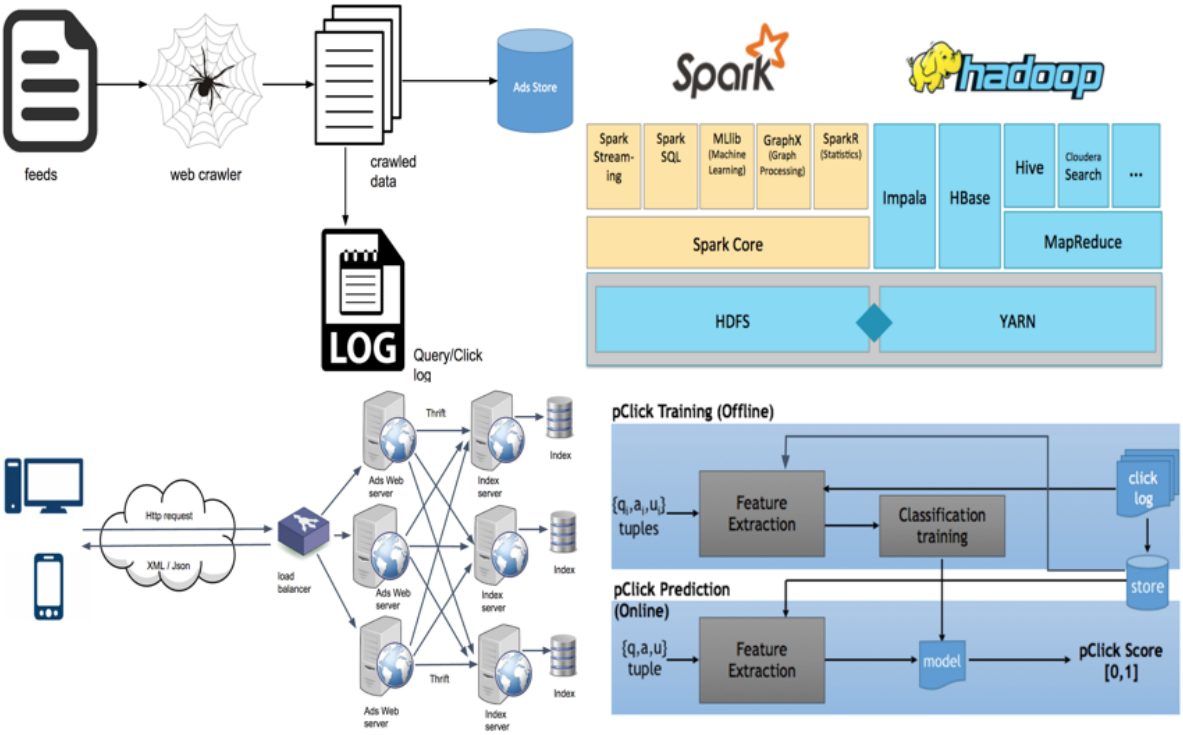
课程内容
用Hive进行数据分析
用Presto进行数据分析
用Spark SQL进行数据分析
用Flink导入实时数据
用Superset生成数据图表

Project 2: 搜索广告平台

【项目介绍】

在本阶段中，通过高强度的实战操作，带领学生逐步了解搜索广告的基本流程和数据结构，在对搜索广告实现原理的基本了解下，指导同学们实现电商数据爬虫，模拟搜索日志数据的生成，并引入信息检索在广告中的应用，实现一个基本的搜索广告服务器 v1.0。在此基础上，通过对Spark MLlib和Spark MapReduce的学习和应用，优化原有搜索广告服务器，在对Query Understanding和排序/定价/位置分配几大核心算法的研究中，进一步探索返回广告的广度，相关度，从理论和实战中深刻理解Google、Facebook等互联网巨头是如何利用广告系统知识、分布式系统优化，机器学习和大数据处理等技术搭建电商搜索广告系统，实现广告服务器后端，健全的大数据处理pipeline，和机器学习离线训练与线上预测系统。

【项目图示】



【学习成果】

1. 深入理解搜索广告的业务流程，数据结构以及query understanding，广告排序、定价、位置等搜索广告的核心算法
2. 熟练掌握使用消息队列实现稳定、可扩展的大规模数据爬虫（half million级别）

3. 理解信息检索在搜索广告中的应用，能够构建广告倒排索引和正向索引并利用Java, gRPC, memcached, mySQL 搭建分布式广告后端系统
4. 理解并应用Spark MapReduce做特征提取，用Spark MLlib实现 query understanding
5. 灵活运用machine learning 改进广告排序算法，并预测pClick及广告相关度

Week 4 课程安排

【理论理解】

课程内容	课程要点
搜索广告的业务流程	• 搜索广告概况 • 搜索广告的数据结构 • 搜索广告后台的业务流 • 爬虫基本原理以及实现爬虫要解决的难点
数据准备	• 设计电商数据爬虫 • 模拟搜索日志

【项目实战】

课程内容
配置开发环境 • Java + IntelliJ • MemCache • MySQL • Spark
实现稳定，可扩展的大规模电商数据爬虫 • Java + Jsoup
用reverse engineering生成大量模拟搜索日志 • Python + pipeline

Week 5 课程安排

【理论理解】

课程内容	课程要点
信息检索的基础	• 信息检索在搜索引擎中的应用 • 倒排表 • 分词
信息检索在广告中的应用	• 用户查询的预处理 • 建立广告关键字倒排表 • 用倒排表选择广告 • 计算相关度
网络服务的理论基础	• HTTP • Java Servlet
设计广告服务器	
设计索引服务器	
MapReduce的理论基础	

【项目实战】

课程内容
建立广告数据索引 • 用MemCache实现倒排表 • 用MySQL实现前向索引 • 用JDBC连接MySQL
搭建广告后台服务 • 配置本地web服务器运行环境 • gRPC实现分布式索引服务器 • 用Java servlet 开发广告服务器 v1.0 • 整合分布式索引服务器返回的结果并实现广告业务逻辑

○ 过滤广告 ○ 广告多样化 ○ 根据相关度排序

Week 6 课程安排

【理论理解】

课程内容	课程要点
Machine Learning 入门	• Gradient descent • Linear regression • Neural network • Classification
为什么需要 Query understanding	• 展示上一个版本的缺陷
什么是 Query understanding	• Query rewrite • Query intent extraction
如何实现 Query understanding	• Word2Vec • Page Rank
Spark MapReduce 入门	• Spark Context • Spark Shell • Spark RDD

【项目实战】

课程内容
Spark MapReduce练习
Query rewrite • 用spark MLlib 实现Word2Vector model • 用Word2Vector model实现rewritten query并应用到广告服务器
实现广告服务器 v2.0：用extended query查询广告索引，比较返回广告的广度的变化

Week 7 课程安排

【理论理解】

课程内容	课程要点
改进广告排序算法	• 广告排序算法公式 • 什么是pClick • 如何用machine learning 预测pClick • 如何用machine learning 预测广告相关度
广告定价算法	
广告位置算法	

【项目实战】

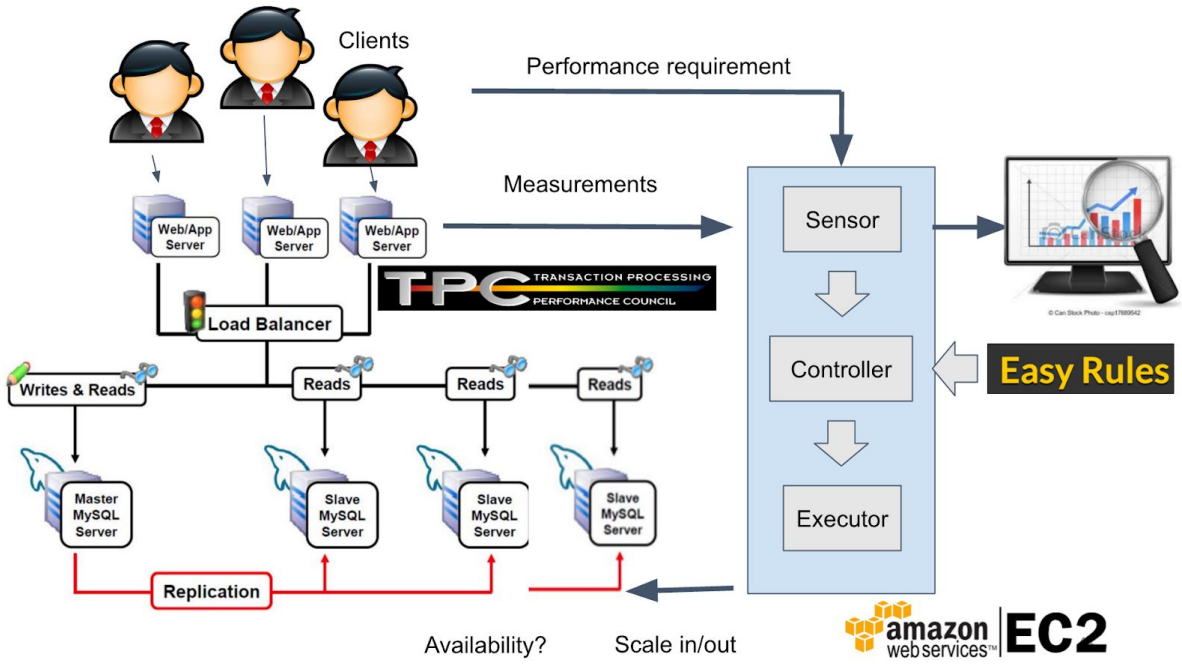
课程内容
pClick 预测 • 用Spark map reduce 实现pClick 特征工程并实现特征提取pipeline • 用Spark MLlib 预测pClick
实现广告服务器 v3.0 • 实现改进后的排序算法 • 跟v2.0的结果比较

Project 3: ElasticDB - 高可用性、可动态扩展的分布式数据库系统

【项目介绍】

本阶段课程将通过对现今最大电商平台的研究与分析，深入了解并从多个角度分析其workload，了解如何设计aws，如何设计load balancer，如何设计分布式数据系统等。课程将带领同学们理论上从TPCW Benchmark，分布式系统设计思路原则CAP，负载均衡的原理和MySQL replication 工作原理入手，根据电商平台的用户需求，研究分布式系统性能的控制与设计，测试与监控，最终实现高可靠性、可动态扩展的分布式数据库系统，达到实时处理经典OLTP的查询请求，实时动态显示资源使用和系统性能，以最优的经济动态增加/减少系统处理能力保证系统性能的最优等功能。

【项目图示】



【学习成果】

1. 理解电商平台例如Amazon.com的主要功能
2. 理解为什么有秒杀访问量
3. 理解云计算平台的经济学原理
4. 理解分布式系统性能、可用性、可扩展性
5. 达到熟练掌握设计、分析、部署、测试分布式系统的目的

6. 实现高可用性的分布式数据库系统
7. 实现动态扩展的分布式数据库系统
8. 动态控制分布式系统性能
9. 掌握CAP原理在分布式系统中的设计和实际运用

Week 8 课程安排

【理论理解】

课程内容	课程要点
电商平台的那些事	● 电商平台的昨天今天与明天 ● 电商平台上面的workload分析 ○ 浏览Browsing ○ 订单Ordering ○ 从数据读写角度来分析workload ● OLTP和OLAP对比
TPCW benchmark 介绍与部署	● TPCW benchmark介绍 ○ 为什么要用TPCW benchmark ○ 经典的多层体系结构 ■ HTTP 与 WWW ■ 前端与后端的通讯 ○ Client Emulator
分布式系统设计思路原则CAP	
负载均衡的原理和使用	● 常用负载均衡的方法 ● 分布式数据库负载均衡的特殊性
MySQL replication 工作原理	● Master-Slave的架构 ● 在CAP原理下，如何让MySQL牺牲C来换取AP ● MySQL replication可扩展性实现原理

【项目实战】

课程内容

在AWS云计算平台上部署TPCW benchmark ● 运行TPCW benchmark ● 运行Client Emulator
实现Load balancer ● 下载，安装，设置，部署MySQL replication ● Routing不同的查询到MySQL replication ● 学习Load balancer部分代码

Week 9 课程安排

【理论理解】

课程内容	课程要点
数据库系统的可用性	● 可用性定义 ● 如何解决可用性的问题
数据库系统的可扩展性	● 可扩展性定义以及Scale in/out，Scale up/down区别 ● 如何解决可扩展性的问题
电商平台的用户需求	● Service Level Agreement是什么 ● Availability的SLA ● Performance的SLA
控制系统设计	● 系统建模 ● 控制系统的三大组成 ○ 检测装置 ○ 控制装置 ○ 执行装置 ● 自动控制原理在计算机系统中的使用
控制系统和TPCW平台的融合来控制分布式系统性能	● 收集检测分布式数据库查询响应时间 ● 基于EasyRules的java规则引擎 ○ 规则的设置 ○ 规则的触发 ● 如何做到 0 down time的scale in/out ● 执行MySQL replication的scale in/out

系统瓶颈分析	• 什么是瓶颈 • 数据库系统和瓶颈之间的一般关系 • CPU, memory, IO bound查询
分布式数据库系统实际测试	• 系统动态访问量和时间之间的关系 • 分布式数据库系统在访问量变化时动态调整资源
分布式数据库系统性能和资源监控	• 监控数据的存储与访问 • 基于CanvasJS的动态曲线绘制 • 基于浏览器的动态曲线展示
项目总结和方向扩展	• 新的架构 • 模型的准确性 • 控制器的使用 • 系统响应时间和稳定性

【项目实战】

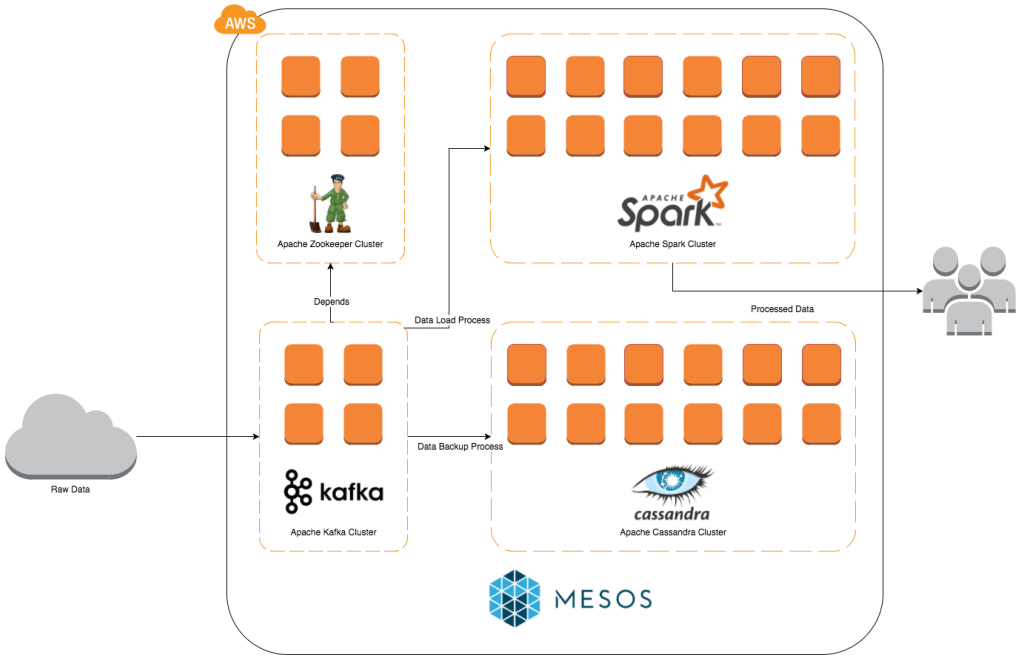
课程内容
实现高可用性和可扩展性 • 演示在某个mysql server停用的时候系统可用性的变化，以及如何检测某个mysql server的运行情况。通过添加新的mysql server来实现高可用性。 • 演示在用户访问量增加的情况下系统响应时间的增加；以及用户访问量减少的情况下系统响应时间的减少，明确通过系统的可扩展性达到稳定系统响应时间的目的 • 基于ECA（事件-条件-动作）框架来实现可扩展性控制器的检测，控制，执行
实现高可用性和可扩展性数据库的系统动态展示 • 收集分析系统状态常用工具dstats应用 • 创建系统状态数据库 • 基于CanvasJS的系统状态实时显示

Project 4: 基于通用大数据处理平台的数字货币分析系统

【项目介绍】

本阶段项目将会从最基础的大数据框架出发，分析它们的优势劣势，学习当前业界最火的系统架构，并将其应用到我们的项目当中，从而构建出一个高性能的基于流数据处理平台的实时数字货币（BTC，BCH，ETH，LTC）分析系统。课程将带领学员从理论到实战，逐步深化对分布式系统的理解，学习高性能数据pipeline搭建的设计理念，并在实践中，亲身体会如何基于Kafka实现数据采集读取层，如何基于Cassandra实现数据持久层，以及如何基于Spark实现流数据的分析。学员将使用AWS，搭建起属于自己的云服务并使用Docker技术，简单快速拥有属于自己大数据平台，经历一个完整的流数据处理平台的搭建过程，并在这个过程中，深究其内部原理，对于每一个技术栈做到知其然，还知其所以然，而不再是简单的利用API的调用。

【项目图示】



【学习成果】

1. 学习、理解并掌握大数据框架的学习路线和方法；理解并灵活运用SMACK架构，解决大数据实际问题
2. 熟悉掌握AWS原理及常用工具，比如EC2，EMR等，并学会如何在AWS上部署Zookeeper，Kafka，Cassandra；

3. 深化对分布式分布式系统(Distributed System)的特性和实现原理的理解，对分布式消息队列(Distributed Message Queue)、分布式数据库(Distributed Database)、分布式批处理(Distributed Batch Processing)以及分布式调度(Distributed Scheduling)从理论架构到实际应用的全方面掌握
4. 熟悉掌握通过命令行操作Kafka，Zookeeper，Cassandra及Spark的能力，并能够搭建完整的流数据处理平台
5. 了解Kafka/Cassandra/Mesos等主流大数据工具的常见Failure Case，并学习掌握如何进行Failure Recovery

Week 10 课程安排

【理论理解】

课程内容	课程要点
Kafka介绍	• Kafka架构 • 内部实现原理 • Kafka事故情景及灾难恢复 • 业内常见应用案例
Zookeeper介绍	• Zookeeper架构 • 内部实现原理 • Zookeeper事故情景及灾难恢复 • 业内常见应用案例

【项目实战】

课程内容
Docker介绍 • Docker架构 • 常用Docker命令
通过命令行操作Zookeeper
通过命令行操作Kafka
Kafka API
基于Kafka实现数据传输层

Week 11 课程安排

【理论理解】

课程内容	课程要点
NoSql 数据库介绍	• NoSql 数据库特点及介绍
Cassandra介绍	• Cassandra架构 • 内部实现原理 • Cassandra事故情景及灾难恢复 • 业内常见应用案例

【项目实战】

课程内容
通过命令行操作Cassandra
Cassandra API
基于Cassandra实现数据持久层

Week 12 课程安排

【理论理解】

课程内容	课程要点
Spark介绍	• Spark架构 • 内部实现原理 • Spark常用API • 业内常见应用案例
Mesos介绍	• Mesos架构 • 内部实现原理 • Mesos事故情景及灾难恢复 • 业内常见应用案例

【项目实战】

课程内容
通过命令行使用Spark实现批数据处理
通过命令行使用Spark实现流数据处理
基于Spark实现数据处理层

备注：课程大纲仅供参考，请以老师实际上课为准。