大数据课程体系

课程说明

本系列课程适合所有对大数据开发有兴趣的人员，从 JavaSE 到大数据生态圈高端课程的开发。课程内容量很大，有一定的难度和深度，让大家在 4 个月的时间内，技术能够有很大程度的提升。

课程路线

第一阶段 JavaSE+MySql+Linux

Java 语言入门 → OOP 编程 → Java 常用 Api、集合 → IO 流 →
Java 实用技术 → Mysql 数据库 → 阶段项目实战 → Linux 基础 →
shell 编程

第二阶段 Hadoop 与生态系统

Hadoop → MapReduce → Avro → Hive →
HBase → Zookeeper → Flume → Kafka →
Sqoop → Pig

第三阶段 Storm 流式计算

Storm

第四阶段 Spark 与生态系统

scala → spark → Mahout → R 语言

第五阶段 项目实战、技术综合运用

云盘存储系统 → 商品推荐系统 → 就业指导

JavaSE

这个阶段是 Java 基础课程，帮大家打好编程基础。

大数据开发需要编程语言的基础，因为大数据的开发是基于一些常用的高级语言，Java 较多不论是 hadoop，还是数据挖掘，都需要有高级编程语言的基础，因此，Java 基础课程是学习大数据的基础。

一、Java 语言入门

1) 介绍计算机的基本使用和 DOS 常用命令，认识软件开发
2) Java 概述和开发环境
3) 了解 Java 基础语法、数据类型、运算符
4) 流程控制语句、函数、数组

二、OOP 编程

1) 面向对象编程(OOP)、类和对象
2) 封装、继承、多态三大特性
3) 抽象类与接口、匿名内部类、帮助文档的生成及使用

三、Java 常用 Api、集合

1) 集成开发工具(Eclipse)安装使用
2) API 常用类：Object、String、StringBuffer、Integer、Date 等
3) 常见排序、查找算法
4) 单列集合 Collection 体系
5) 双列集合 Map 体系

四、IO 流

1) 异常体系 Exception、Error
2) File 类及编程技巧递归
3) 常用 IO 流及编码表
4) NIO

五、Java 实用技术

1) 多线程：多线程实现、同步机制、线程间通信问题等
2) 网络编程：网络编程三要素 IP、端口号、协议 Socket 编程之 UDP 协议和 TCP 协议的实现

3) 反射、内省

4) 设计模式：工厂模式之简单工厂模式、工厂方法模式、单例模式装饰模式适配器模式等

阶段培训目标

掌握常见的数据结构和实用算法培养良好的企业级编程习惯。掌握面向对象的基本原则以及在编程实践中的意义掌握 Java 面向对象编程基本实现原理。熟练掌握 JDK 核心 API 编程技术理解API设计原则具备熟练的阅读API文档的能力为后续的课程学习打下坚实的语言基础。熟练掌握 JavaSE 核心内容，特别是 IO 和多线程初步具备面向对象设计和编程的能力掌握基本的 JVM 优化策略。

数据库 MySql

1) MySql 数据库的应用

2) 触发器、存储过程序列、索引、视图对象

3) JDBC 技术：JDBC 基础 Statement、PreparedStatement、ResultSet 结果集对象等

4) 数据库连接池技术

5) ORM 对象关系映射基本思想。

阶段培训目标

熟练掌握 SQL 语句掌握一定的数据库查询技巧及 SQL 语句优化技巧。掌握 MySql 体系结构及核心编程技术。

阶段项目:实时聊天软件

运用前面学习的知识，综合运用 Java 基础、Swing、集合、IO、数据库、Socket 编程，编写一个实时聊天软件。

Linux&VMware 基础

这章是 Linux 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，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。

Linux 的介绍，Linux 的安装：VMware Workstation 虚拟软件安装过程、Ubuntu 虚拟机安装过程

1) Linux 的常用命令：介绍、使和练习
2) nano 编辑器：nano 编辑器的介绍、nano 托使用和常用快捷键
3) Linux 用户和组账户管理：用户的管理、组管理
4) Linux 系统文件权限管理：文件权限介绍、文件权限的操作
5) apt 命令，源修改
6) Linux 网络管理：hosts,hostname,ifconfig 等
7) Linux 系统进程管理常用命令 ps、pkill、top、htop 等的使用
8) Shell 编程：Shell 的介绍、Shell 脚本的编写

Hadoop 课程

此部分带领大家了解 Hadoop 在大数据中的用途，在 Linux 环境中搭建 JDK、SSH 和 Hadoop 的环境等

深入剖析 Hadoop 文件系统架构，让你清楚知晓大数据存储的机制

MR 是大数据常用的计算框架，是大数据工程师应该熟练掌握的，带领体验 MR 案例，并分析 MR 的作业流程和工作机制等

任何程序的开发，少不了程序调优，Hadoop 也是如此，本章节还将带领大家学习常用的调优方法。

一、Hadoop 介绍和环境搭建

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 独立模式安装和测试

7) Hadoop 的集群结构

8) Hadoop 伪分布的详细安装步骤

9) 通过命令行和浏览器观察 Hadoop

10) Hadoop 启动脚本分析

11) Hadoop 完全分布式环境搭建

12) Hadoop 安全模式、回收站介绍

二、HDFS 体系结构和 Shell 以及 Java 操作

1) HDFS 底层工作原理

2) HDFSdatanode,namenode 详解

3) 单点故障（SPOF）和高可用（HA）

4) 通过 API 访问 HDFS

5) 常用压缩算法介绍和安装使用

6) Maven 介绍和安装，eclipse 中试用 Maven，搭建 Maven 本地仓库

三、详细讲解 Mapreduce

1) Mapreduce 四个阶段介绍

2) Job、Task 介绍

3) 默认工作机制

4) 创建 MR 应用开发，获取年度的最高气温

5) 在 Windows 上运行 MR 作业，

6) Mapper、Reducer

7) InputSplit 和 OutputSplit

8) Shuffle：Sort，Partitioner，Group,Combiner

9) 通过计数器调试程序

10) 在 Windows 安装 Hadoop

11) 在 eclipse 安装 hadoop 插件，访问 hadoop 资源

12) 在 eclipse 中编写 ant 脚本

13) YARN 调度框架事件分发机制

14) 远程调试资源管理器

15) Hadoop 的底层 google ProtoBuf 的协议分析

16) Hadoop 底层 IPC 原理和 RPC

四、HA

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

数据序列化系统 Avro

Avro 是一个数据串行化系统，提供了丰富的数据结构，紧凑、快速、二进制数据格式，存储持久化数据的容器文件，远程过程调用 RPC，动态语言的简单集成。代码生成不需要读写数据文件，也不需要实现 RPC 协议。

1) Avro 简介

2) Avro 环境搭建

3) 数据类型和模式

4) 使用方式

5) 使用 AVRO-tools 工具生成源代码

6) 在 MR 中使用 avro 串行化

数据仓库 Hive

Hive 可以将 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，不必开发专门的 MapReduce 应用，十分适合数据仓库的统计分析。

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 体系结构简介

3) Hive 客户端简介

4) Hive 集群

5) HiveQL 定义

6) HiveQL 与 SQL 的比较

7) 数据类型

8) 配置 Hive 使用 MySql 数据库

9) Hive 管理表、外部表、临时表、分区表和桶表

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) Hive join、union、View、Index 演示

14) Operators 和 functions 与 CLI 客户端演示

15) 安全、锁

16) 用户自定义函数（UDF 和 UDAF）的开发与演示

17) Hive 压缩和优化

分布式数据库 HBase

HBase 是 Apache 软件基金会 Hadoop 项目的一部分，运行于 HDFS 文件系统之上，为 Hadoop 提供类似于 BigTable 规模的服务，它可以容错地存储海量稀疏的数据。

这章将带领大家学习非常常用的 HBase，包括 HBase 环境搭建，常用 CRUD 操作，常用 API 以及 HBase 的调优等。

1) HBase 简介

2) HBase 与 RDBMS 的对比

3) HBase 安装：本地模式、为分布式模式、完全分布式模式

4) HBase Shell 体验

5) 数据模型

6) 系统架构

7) HBase 核心术语介绍

8) 通过 API 操作 HBase

9) 表的设计

10) HBase 自定义协处理器

11) HBase 上的 MapReduce

12) 集群的搭建过程讲解

13) 集群的监控

14) 集群的管理

15) Hbase 表级优化

16) Hbase 写数据优化

17) Hbase 读数据优化

18) 使用 Kundera ORM 操纵 hbase

Zookeeper 开发

Zookeeper 曾是 Hadoop 的子项目，现为顶级项目，在分布式集群（Hadoop 生态圈）中的地位越来越突出， Zookeeper 是协同服务、为分布式应用提供服务的。

学好 Zookeeper，对后面学习其他技术至关重要。

1) Zookeeper 简介

2) Zookeeper 组件

3) Zookeeper 名字空间等级

4) ZNode 类型

5) Zookeeper 的工作流程，leader select 过程

6) 搭建 Zookeeper 为分布搭建、集群搭建

7) Zookeeper Cli

8) 使用 Zookeeper 的客户端 API 连接 Zookeeper

9) Zookeeper rmi 高可用分布式集群开发

10) Netty 异步 io 通信框架

11) Zookeeper 实现 netty 分布式架构的高可用

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 优点

3) flume 架构:水槽

4) flume 安装与测试

5) flume 部署方式

6) flume source 相关配置及测试

7) flume sink 相关配置及测试

8) flume 源代码分析

9) flume selector 相关配置与案例分析

10) flume Sink Processors 相关配置和案例分析

11) flume Interceptors 相关配置和案例分析

12) flume AVROClient 开发

13) flume 和 kafka 的整合

Kafka 分布式队列系统课程

Kafka 是一种高吞吐量的分布式发布订阅消息系统，它可以处理消费者规模的网站中的所有动作流数据。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) 消息压缩

6) Kafka 集群镜像

7) Kafka 的存储策略

8) Kafka 分区特点

9) Kafka 的发布与订阅

10) Zookeeper 协调管理

11) Java 编程操作 Kafka

12) scala 编程操作 kafka

13) flume 和 Kafka 的整合

14) Kafka 和 storm 的整合

Sqoop 数据迁移

Sqoop 是一个用来将 Hadoop 和关系型数据库中的数据相互转移的工具，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

   a) DBMS hdfs

   b) DBMS hive

   c) DBMS hbase

4) Sqoop-export

5) Job 创建、查看、执行和删除

Pig 开发

Pig 是分析大数据平台，使用表达式(脚本)语言。也是 mr 抽象，和 hadoop 配合使用 Pig

拉丁用于编写高级分析程序，提供了各种操作可用开发人员进行读写操作使用 pig 拉丁语言编写脚本，脚本内部转换成 mr taskpig 引擎将脚本转换 mr job。

1) Pig 特点
2) Pig 架构
3) Pig 组件：parser、优化器、编译器、执行引擎
4) Pig Data ModelL: field、tuple、bag、relation、Map
5) 安装 Pig
6) Pig 运行模式：local、hdfs
7) Pig 执行模式：交互、批处理、嵌入式
8) Shell 命令：kill、exec、run、clear 等
9) 诊断操作符：Dump、Describe、Explanation、Illustration
10) Pig 操作：group、cogroup、join、union、split 切割、过滤、函数等
11) load() / storage：PigStorage、TextLoder、BinStorage、Handling Compression 等

Storm 实时数据处理

Storm 是开源分布式的、容错的实时计算系统。可以实时可靠地处理无线数据流。可以使用任何语言开发。Storm 适用于实时分析、在线机器学习、分布式 RPC、ETL、高效运算，在多节点集群中每秒可处理上百万条记录。

本章节的学习，让大家全面掌握 Storm 内部机制和原理，通过项目实战，让大家拥有完整项目开发思路和架构设计。

1） Storm 的基本概念
2） Storm 的应用场景
3） Storm 和 Hadoop 的对比
4） Storm 优势
5） Storm 集群的安装的 linux 环境准备
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) 在单节点集群上部署 topology

13) Storm 程序本地模式 debug、Storm 程序远程 debug

14) Storm 事物处理

15) Storm 消息可靠性及容错原理

16) 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

17) Storm Trident 概念

18) Trident state 原理

19) Trident 开发实例

20) Storm DRPC(分布式远程调用)

21) Storm DRPC 实战讲解

22) Storm 和 Hadoop 2.x 的整合：Storm on Yarn

Scala 语言编程

Scala 是一门多范式的编程语言，一种类似 java 的编程语言，设计初衷是实现可伸缩的语言、并集成面向对象编程和函数式编程的各种特性。

在此部分内，将学习语言规则与简单直接的应用， 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。

1) Scala 介绍

2) Scala 与 Java 比较

3) Scala 解释器、变量、常用数据类型等

4) Scala 的条件表达式、输入输出、循环等控制结构

5) Scala 的函数、默认参数、变长参数等

6) Scala 的数组、变长数组、多维数组等

7) Scala 的映射、元组等操作

8) Scala 的类，包括 bean 属性、辅助构造器、主构造器等

9) Scala 的对象、单例对象、伴生对象、扩展类、apply 方法等

10) Scala 的包、引入、继承等概念

11) Scala 文件操作和正则表达式

12) Scala 串行化

13) Scala 的特质

14) Scala 的操作符

15) Scala 的高阶函数

16) Scala 的集合

17) Scala 数据库连接

Spark 大数据编程

Spark 是类 Hadoop MapReduce 的通用并行框架，中间输出结果可以保存在内存中，适用于数据挖掘与机器学习等需要迭代的 MapReduce 的算法，启用了内存分布数据集，优化迭代工作负载。

本章节内容全面涵盖了 Spark 生态系统的概述及其编程模型，深入内核的研究，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

Mahout

主要用于创建可伸缩机器学习算法。

1) Mahout 特性

2) 机器学习介绍
3) 实现机器学习的方式
4) 使用 Mahout 实现推荐功能
5) Mahout 推荐引擎
6) 构成推荐引擎的组件
7) 使用 Mahout 构建推荐器
8) 通过 Eclipse 创建 Mahout 项目，实现推荐功能
9) 聚类：聚类的过程、复制文件到 hdfs、从 input 数据准备序列文件、运行任何可以使用的聚类算法
10) 分类算法、分类过程

R 语言

1) R 语言介绍、下载和安装
2) R 语言包、库
3) R 批处理
4) 数据集、对象、向量、标量、矩阵、数组、数据框、因子、列表、加载 xlsx 文件；
5) plot 制图、修改图形属性、颜色、案例
6) 文本大小的参数、字体、图形大小和边界大小
7) legend 图例、条形图、饼图、点图
8) 基本统计分析
9) 定义函数、使用内置汽车数据集

大型企业项目实战

项目一：国内某电视台卫视节目 HDFS 的云盘存储系统

国内某电视台卫视节目云盘存储系统，基于 Hadoop HDFS 分布式存储，实现对文件的浏览、上传、下载、删除功能，

系统支持多种文件格式，文件大小支持几十 K 到几十 M，甚至上百 M。

视频存储容量为每天 10 小时有效视频文件,每小时的视频大小为 1g（高清视频），每周七天，存储近 10 年的数据。

总容量评估为 ：10 x 365 x 10 x 1g = 35tb 字节数。

基于 HDFS 的云盘系统可以把独立的服务器磁盘或磁盘阵列统一为有机整体，由 Hadoop HDFS 全局维护数据的存储与备份，

以存储海量数据，对外部系统提供一致的文件下载服务。

基于 HDFS 的云盘系统可以将数据冗余存储，保证了数据的安全存储与备份，并使整个存储的水平扩展非常容易。

namenode 节点使用 QJM 实现高可用集群，支持自动+手动两种容灾方式。

为避免工作人员因专业性强导致集群资源分布不均，根据需求设定空间配额和目录配额进行约束管理。

为防止管理员对资源目录进行快速备份和后期恢复工作，支持快照功能，且可以设定快照数量。

为防止管理员操作不当，误删除重要数据，集群支持回收站机制，并设有告警和提示功能。

云盘存储系统支持存储节点的热伸缩，保证数据高可用性。

项目二：国内某前三甲著名电商的商品推荐系统

国内某前三甲著名电商的商品推荐系统，项目又名--“猜你喜欢”。

项目采用 MapReduce 计算模型结合 mahout 机器学习实现用户相似度、商品关联度和降维分析等

协同过滤算法。

数据直接来自企业在线系统的生产数据，具有权威性和真实性，数据量在 tb 级以上。

利用该系统，直接促成商业交易额提升 25%。