EDU
CSDN学院 IT实战派

CSDN 学院大数据生态体系课程大纲

课程优势：

1. 技术新。课程首推业界最先进技术标准，引领行业技术更新：Hadoop3.0、Hive2.0、HBase1.0、Storm1.0、Spark2.2等。
2. 完整项目贯穿教学体系。理论与实践并重，模块以真实企业级项目贯穿始终。多个大数据项目，覆盖多个行业领域，对于学员历练非常难得。
3. 课程体系完善。课程与项目，知识点与案例穿插巩固，理论与实践强有力的结合，易于消化。
4. 项目实战营模式：每阶段完成后，课程提供独立项目，由学员自行研发，结合老师辅导，真正做到活学活用。
5. 讲师团队实力强。课程体系讲师既有专业技术讲师，更有企业一线开发工程师。动态把握行业实用性技术、流行技术、先进技术。做到了知识点有效性和学习效率的大幅度提高。
6. 实战场景。内容涵盖大数据离线分析应用，实时计算应用，数据仓库应用等多种业务场景。
7. 业界首推的云实验室实战：提供独家大数据云实验室环境，通过实验室感受 TB 级数据分析效果。

8. 真实还原大数据应用开发场景：持续 ETL 流程，定时 ETL 流程，多业务并行计算，业务协同管理等场景，做到真实演练。
9. 企业级大数据平台演练：课程涵盖 Cloudera 大数据平台，CDH 企业级开源生态。让学员真正体会企业级大数据软件平台。

模块一： Linux 实践与 Java 基础				
时间	**章**	**节**	**案例/作业/练习**	**学习目标**
第一周 **时长：6 小时**	**第一章 Linux基础实践**	1.Linux 系统概述 2.CentOS 系统安装及配置 3.Linux 常用命令 4.Linux 网络基础	**Linux CentOS 安装** 所需技能点： 1.Linux 基础 2. CentOS 系统安装及相关配置	1.掌握 Linux 基本使用方法 2.掌握 Linux 文件操作 3.掌握 Linux 用户管理 4.掌握 Linux 权限管理
	第二章 Linux 用户管理	1. Linux 用户与组 2. 用户的管理与维护	**Linux 操作文件系统** 所需技能点： 1. Linux 文件与目录管理	

批注 [yz1]：对应 PPT 一级标题
批注 [yz2]：对应 PPT 二级标题
批注 [yz3]：对应 PPT 学习目标

	第三章 Linux 文件系统	1.文件属性与访问权限 2.文件权限与属主修改 3.基本磁盘管理 4.LVM 高级磁盘管理	2. Linux 终端常用命令 3. Linux 防火墙操作 4. vi 文本编辑器 5. 磁盘管理	
第二周 **时长：6 小时**	**第四章 Java SE 基础**	1. Java 发展史和开发环境搭建 2. Java 基本语法 3. Java 异常处理 4. Java 集合框架	**Java 实现文件操作操作** 所需技能点： 1.Java 类的概念 2.Java File IO	1.掌握 Java 原理与架构 2.熟练 Java 操作 3.熟练 Java 类，基本语法 4.掌握 Java 多线程

	第五章 JavaSE 高级应用	1.Java IO 流与文件操作 2.字节流的包装和链接 3.字符流的包装和链接 4. 对象的序列化 5.线程的创建、运行和结束 6.线程控制	3.Java 基本语法 **Java 实现数据库操作** 所需技能点： 1. JDBC 2. 集合框架 **Java 多线程文件操作** 所需技能点： 1.Java 多线程 2.Java 类和对象	5.掌握 Java 操作数据库方法
	第六章 Java 操作数据库	1.Linux 系统下安装 MySQL 数据库 2.MySQL 数据库操作 3.事务与隔离级别 4.JDBC 操作 MySQL 数据库		

模块二：Hadoop 生态体系

EDU
CSDN学院 IT实战派

时间	章	节	案例/作业/练习	学习目标
第一周 **时长：7 小时**	**第一章** **Hadoop 2.X 入门**	1. 大数据行业现状分析与最新行业动态 2. Hadoop 的起源与简史 3. Hadoop 2.X 生态体系简介：HDFS，MapReduce，Hive 等 4. Hadoop 的发行版本 5. Hadoop 3.0 新特性介绍 6. Hadoop 在互联网公司的应用案例解析 • 互联网企业遇到大数据的问题	**大数据当前应用现状** 所需技能点： 1.大数据行业现状 2.大数据核心能力 **某企业微博数据分析平台搭建：** 所需技能点： 1.Linux 基础 2.Hadoop 伪分布式开发环境搭建 3.SSH 免密访问	1. 掌握大数据的概念与适用领域 2. 熟悉 SSH 的安装与无密码访问 3. 熟练掌握 Hadoop 的安装方法，配置文件 4. 了解Hadoop不同的发行版本间的区别 5. 熟悉 Hadoop 3.0 新特性 熟悉 Hadoop 系统体系结构架构

批注 [yz4]：对应 PPT 一级标题

批注 [yz5]：对应 PPT 二级标题

批注 [yz6]：对应 PPT 学习目标

批注 [yz7]：建议“使用 XX 实现 XX”格式

		• 案例分享：12306 大数据实践 • 案例分享：淘宝飞天平台实践 • 案例分享：微信红包大数据分析 • 案例分享：春晚大数据案例分析 • Hadoop 如何和传统 IT 系统配合完成原来不可能的任务		

		7. Hadoop 2.X 安装部署的三种模式：集群，伪分布式，Local ● Hadoop 伪分布式开发环境搭建		

	第二章 大数据文件系统之HDFS	1.HDFS 原理与架构说明 2.HDF 的 namenode 和 datanode 3.HDFS 缓存机制(Cache) 4.HDFS 快照(Snapshot) 5.HDFS 命令行操作 6.HDFS 的 Java API 编程 7.HDFS Web HDFS API 编程实践	**某企业微博数据分析平台数据存储模块设计与实现：** 所需技能点： 1.HDFS Hadoop 启动与使用 2.使用 Unix/Linux 工具来分析数据 3.HDFS 命令 4. Java 语言 5. HDFS Java API 6. Eclipse 工具	1.掌握 HDFS 原理与架构 2.熟练 HDFS 操作 3.熟练 Java 操作 HDFS 4.掌握 HDFS 数据操作时读取与写入的执行流程

第二周 **时长：4 小时**	**第三章 MapReduce 实战及原理**	1.MapReduce 概要介绍 2.MapReduce 示例运行与解析 3.搭建 Eclipse Hadoop 开发环境 4.WordCount 案例实践 5.Yarn 原理及架构	**某企业微博数据分析平台数据分析模块设计与实现：** 所需技能点： 1.Java 语言 2.MapReduce 原理 3.MapReduce 的输入与输出 4.Eclipse Hadoop 开发环境 **MapReduce 分析微博声量分析与单元测试** 所需技能点： 1.数据 Map 和 Reduce 的过程 2.MapReduce 编程 3.Eclipse Hadoop 开发环境	1.熟练掌握数据 Map 和 Reduce 的过程，输入与输出 2.掌握 MapReduce 原理 3.掌握 MapReduce 开发技术 4.掌握 Yarn 架构与原理 5.了解 Reduce 的优化，了解 combine 函数 6.掌握 Hadoop Eclipse 开发插件的安装与配置方法

	第四章 Hadoop 高级编程	1.shuffle 过程:Map 端，Reduce 端排序 2.使用 MRUnit 进行单元测试过程 3.MapReduce 实现数据去重 4.MapReduce 实现数据排序 5.MapReduce 实现倒排索引 6.MapReduce 数据压缩：Snappy,Gzip,LZO 7.MapReduce Partitioner,Combiner 实现及应用	**使用 MapReduce 实现微博情感分析与单元测试** 所需技能点： 1. MapReduce 原理、编程 2. MapReduce 输入输出 3. Eclipse Hadoop 开发环境 **使用 MapReduce 实现微博发帖用户归属地分析与单元测试** 所需技能点： 1. MapReduce 原理、编程 2. MapReduce 输入输出 3. Eclipse Hadoop 开发环境	1.理解 shuffle 过程 2.掌握 MapReduce 单元测试过程 3.灵活运用 MapReduce 完成各种需求的开发

	第五章 Pig 数据仓库技术(选修)	1. Pig 架构与原理简介 2. Pig 与 MapReduce 的不同及关系 3. Pig 的安装与配置 4. Pig 的内置函数 5. 使用 Pig Latin 语句分析数据	**某企业微博数据分析平台数据分析模块设计与实现：** **Pig 环境搭建** 所需技能点： 1. Pig 安装 2. Pig 配置文件 3. Linux 环境 **使用 Pig 实现微博声量情况分析** 所需技能点： 1. Pig Latin 编辑器 2. Pig Latin 脚本 3. Pig 操作技术	1.了解 Pig 的功能与适用环境 2.掌握 Pig 的安装与配置方法能 3.熟练掌握 Pig Latin 中的结构、表达式、语句、类型、模式、函数 4.掌握 Pig 的自定义函数技术

第三周 **时长：7 小时**	**第六章 Hive数据仓库技术**	1.Hive 数据仓库概要介绍 2.Hive 安装与配置 3.Hive 数据模型 4.操作 Hive 5.用户自定义函数:udf,udtf,udaf 6.Hive2.0 新特性 7.Hive2.0 存储过程：HPL/SQL 实践 8.Hive Index 使用说明 9.Hive Update,Delete 操作说明 10.Hive OrcFile,Parquet 文件格式实践 11.Hive 数据压缩及解决数据倾斜问题	**某企业微博数据分析平台数据仓库模块设计与实现及可视化：** **Hive 安装部署** 所需技能点： 1. Hive 的安装 2. Hive 配置文件 3. MySql 数据库 4. 依赖组件的安装 **Hive 命令行操作** 所需技能点： 1. Hive 启动、访问 2. Hive 表 3. Hive 基本操作技术	1.了解 Hive 的特性与适用环境 2.掌握 Hive 的安装与配置方法 3.掌握 Hive 的数据存储特性 4.了解 Hive 的 Metastore 数据库 5.了解 Hive 数据模型 6.掌握查询中的连接、子查询与视图 7.熟悉用户自定义函数

			Hive 实现微博情感分析/发帖/用户归属地分析/声量分析/用户活跃度分析/KOL 分析及可视化 所需技能点： 1. Hive 数据模型 2. Hive 数据迁移 3. Hive 操作技术 4. Hive HQL DDL 5. Hive Query 分析使用	

第四周 **时长：8 小时**	**第七章 分布式数据库 HBase**	1. HBase 基础概念与数据模型 2.HBase 的系统架构 3.HBase 的安装与配置 4.HBase Shell 5.HBase 的 Java 编程接口 6.Zookeeper 简介 7.ZooKeeper 实现高可用 8.HBase 协处理器原理 9.HBase 二级索引 10.HBase1.X 新的架构等 11.HBase 与 Hive 协同工作 12.HBase Phoenix 操作	**某企业微博数据分析平台** **NoSQL 模块设计与实现：** **HBase 安装与高可用安装配置** 所需技能点： 1. HBase 配置文件 2. Zookeeper 安装配置 3. HBase 架构 **Java 编程完成 HBase 增删改查** 所需技能点： 1. HBase 数据结构 2. JDBC 编程 3. HBase 访问	1. 了解 Hbase 的基础概念 2. 掌握 HBase 数据结构 3. 掌握 HBase 的操作技术 4. 了解 Zookeeper 的作用 5. 掌握 Zookeeper 的安装配置方法，掌握 HBase 高可用安装配置 6. 掌握 HBase 编程技术

		13.HBase 表结构设计案例	**HBase 实现用户微博内容存储及查询：** 所需技能点： 1. HBase 操作技术 2. HBase 列族，数据结构 3. HBase 编程技术 **HBase 实现用户微博统计结果存储与查询：** 所需技能点： 1. HBase 操作技术 2. HBase 列族，数据结构 3. HBase 编程技术	

第五周： **时长：7 小时**	**第八章 数据迁移工具 Sqoop**	1.数据传输工具 Sqoop 的功能与适用环境 2.Sqoop 的安装与配置 3.Sqoop 连接器 4.Sqoop 数据导入示例解析 5.Sqoop 生成的辅助代码	**某企业微博数据分析平台数据导入 ETL 模块设计与实现：** **Sqoop 环境搭建** 所需技能点： 1. Sqoop 安装配置 2. Linux 环境 **Sqoop 从 HDFS 数据导入与导出** 所需技能点： 1. HDFS 操作技术 2. Sqoop 命令 3. 关系型数据库操作技术	1.了解 Sqoop 的功能与适用环境 2.熟练掌握 Sqoop 的安装配置步骤 3.了解 Sqoop 的连接器 4.掌握利用 Sqoop 将 RDBMS 中的数据导入 HDFS 的方法与详细步骤 5.了解导入构成中生成的辅助代码功能

			Sqoop 从 Hive 数据导入与导出 所需技能点： 1. Hive 数据库操作技术 2. Sqoop 命令 3.关系型数据库操作技术	

	第九章 日志收集工具 Flume	1.Flume 核心概念 2.Flume 基础结构与三大组件 3.Fme 的特性:Event 等 4.Flume 的分布式安装与配置 5.启动 Flume 集群节点 6.Flume 的命令行工具 7.Flume 的 Source 和 Slink 8.数据流过程解析 9.flume 拦截器：自定义拦截器，Regex 拦截器等 10.flume 自定义 Source、Sink 和 Interceptor	**Flume 与 Shell 配合使用** 所需技能点： 1. Flume 的命令行工具 2. Shell 命令 3. Flume 的特性 4. 数据流过程 **Flume 多种 Source、Sink 说明** 所需技能点： 1. Flume 核心概念 2. Flume 基础结构与三大组件 3. Flume 多种 Source 4. Flume 多种 Sink	1.熟悉 Flume 的体系结构与特征，掌握其三大组成部分 2.掌握 Flume 的分布式安装、配置与启动方法 3.了解 Flume 集群的动态配置 4.熟悉 Flume 的命令行工具 5.掌握 Flume 的各种类型 Source 和 Slink 及数据采集流通的过程

第一阶段实战项目		
项目名称	**项目需求**	**其他**
某企业微博社交数据分析平台 **时长：6 小时**	项目背景： 随着移动互联网的迅猛发展，微博作为一种新型的社交媒体和信息交流平台，由于其写作简短、发布便捷、交互实时等特点深受大众欢迎。越来越多的网民通过微博分享自己的日常生活、心情感悟以及表达个人情感，并且利用关注和其他交互功能建立自己的朋友圈。充分利用微博平台，深入分析和挖掘用户的情感信息和社会关系，对于舆情监控、市场营销、政府决策以及个性化推荐等具有重要意义。 本项目基于微博数据，深入挖掘微博所包含的大数据意义。 项目难点： 1.多个 MapReduce 任务运行的时候，如何保证执行效率 2.HDFS 文件夹规划的一般方法，如何划分权限，避免出现用户使用问题 3.微博数据采集以后如何平均分配到各个机器上 4.微博数据放置到 HDFS 上如何解决乱码 5.大量的微博数据如何解决占用存储空间问题 6.Hive 数据倾斜优化 7.HBase 表结构设计	

项目知识要点:

HDFS 知识要点：

1. 如何设计存储目录
2. 如何定期清理 HDFS 数据
3. 如何定期实现脚本化数据存储

MapReduce 知识要点：

1. MapReduce 如何实现数据压缩
2. MapReduce 能够处理的压缩格式
3. 如何用 Python 编写 MapReduce 定时任务处理数据
4. 大数据量情况下 MapReduce 代码优化
5. MapReduce 程序与脚本程序协同化工作

Hive 知识要点：

1. Hive 数据模型设计
2. Hive 动态数据仓库设计
3. Hive SQL 脚本编写

HBase 知识要点：

1. HBase 数据模型设计
2. HBase 与 Hive 如何协同
3. HBase 表结构优化

4. HBase 程序与脚本程序协同化工作

Sqoop 知识要点：

1. 如何实现与 MySQL 数据库对接
2. 大数据量情况下 Hbase 代码优化

项目技术选型：

HDFS＋MapReduce＋Yarn+＋Hive+ HBase+Sqoop+Flume +Zookeeper

项目业务实现：

1. 项目需求 1: 数据目录规划与环境部署
2. 项目需求 2: 微博数据 HDFS 存储设计
3. 项目需求 3: 微博数据导入 HDFS
4. 项目需求 4: 微博数据 MapReduce 分析模块，包含 combiner 等使用
5. 项目需求 5: 微博情感分析及单元测试编写
6. 项目需求 7: 微博声量趋势分析及单元测试

	7. 项目需求 8: 微博用户活跃度分析及单元测试 8. 项目需求 9: 微博用户 KOL 分析及单元测试 9. 项目需求 10: 项目测试	
某企业电商 **数据运营分析平台** **时长：6 小时**	项目背景: 随着移动互联网、物联网、云计算等新兴信息技术在社会、经济各个领域的不断应用，全球数据量正呈现出前所未有的爆发式增长态势。与此同时，数据类型及来源的多样性、数据产生与分析的实时性、数据的低价值密度等复杂特征日益显著，标志着“大数据”时代的到来。大数据同云计算、物联网一样，是信息技术领域的重大技术变革，成为一种重要的新型战略资源，受到各国政府的高度重视。美国将大数据上升为国家战略，英国开展了“数据权”运动，欧盟提出了开放数据战略。电商作为中国崛起的业务，需要通过大数据技术进行深入的挖掘与分析使用。电商数据运营包括，流量经营，用户运营，内容运营，产品运营等。 项目难点: 1.多个 MapReduce 任务运行的时候，如何保证执行效率 2.HDFS 文件夹规划的一般方法，如何划分权限，避免出现用户使用问题 3.电商数据采集以后如何平均分配到各个机器上 4.电商数据放置到 HDFS 上如何解决乱码 5.大量的电商数据如何解决占用存储空间问题 6.Hive 数据倾斜优化 7.HBase 表结构设计	

项目知识要点:

HDFS 知识要点：

1. 如何设计存储目录
2. 如何定期清理 HDFS 数据
3. 如何定期实现脚本化数据存储

MapReduce 知识要点：

1. MapReduce 如何实现数据压缩
2. MapReduce 能够处理的压缩格式
3. 如何用 Python 编写 MapReduce 定时任务处理数据
4. 大数据量情况下 MapReduce 代码优化
5. MapReduce 程序与脚本程序协同化工作

Hive 知识要点：

1. Hive 数据模型设计
2. Hive 动态数据仓库设计
3. Hive SQL 脚本编写

HBase 知识要点：

1. HBase 数据模型设计
2. HBase 与 Hive 如何协同
3. HBase 表结构优化

	4. HBase 程序与脚本程序协同化工作 Sqoop 知识要点： 1. 如何实现与 MySQL 数据库对接 2. 大数据量情况下 Hbase 代码优化 项目技术选型： HDFS＋MapReduce＋Yarn+＋Hive+ HBase+Sqoop+Flume +Zookeeper 项目业务实现： 项目需求 1: 电商运营大数据平台平台搭建 项目需求 2: 电商运营大数据平台 ETL 脚本设计及存储设计：Sqoop+Flume 项目需求 3: 电商运营大数据平台分析模块 - 使用 HBase 进行用户画像实现 项目需求 4: 电商运营大数据平台数据仓库模块 - 流量分布分析实现及单元测试 项目需求 5: 电商运营大数据平台数据仓库模块 – 用户订单情况分析及单元测试 项目需求 6: 电商运营大数据平台数据仓库模块 – 用户订单情况分析及单元测试 项目需求 7: 电商运营大数据平台模块 - 使用 HBase 实现网页 URL，PV 等查询	

	项目需求 8: 项目测试	

EDU
CSDN学院 IT实战派

模块三：Spark 核心架构				
时间	章	节	案例/作业/练习	学习目标
第一周 时长：7 小时	第一章 Scala 基础课程	1. Scala 入门 2. Scala 基础语法 3. Scala 函数 4. 面向对象	**Scala 研发环境搭建案例** 所需技能点： 1. Scala 安装 2. Scala 相关配置 3. 操作系统平台特性	1. 掌握 Scala 环境搭建与开发； 2. 掌握 Scala 编程语法与使用； 3. 掌握 Scala 面向对象编程基本方式； 4. 掌握 Scala 面向对象编程基本方法；

批注 [yz8]：对应 PPT 一级标题

批注 [yz9]：对应 PPT 二级标题

批注 [yz10]：对应 PPT 学习目标

	第二章 Scala 编程实战 (选修)	1. 函数式编程 2. 函数式编程之集合操作 3. 模式匹配 4. 类型参数		1. 掌握 Scala 函数式编程基本方法； 2. 掌握 Scala 集合操作方法 3. 掌握 Scala 类型参数技术

第二周 **时长：7 小时**	**第三章 Spark 原理**	1. Spark 2.X 入门 与 Spark1.X 对比 2. Spark 2.X 运行机制 a) 基本术语 b) 运行架构 c) Spark On Standalnoe 部署与实例分析 d) Spark on YARN 实例解析 3. Spark 2.X 原理分析	**某金融企业用户交易行为分析大数据平台搭建：** 所需技能点： 1. Hadoop 原理及开发环境搭建 2. Spark RDD 原理 3. Spark RDD transformation 和	1. 掌握 Spark 基本原理与架构 2. 掌握 Spark 基本的部署方式

	第四章 Spark2.X 算子及高级应用	1. Spark 编程模型解析 2. RDD 的特点、操作、依赖关系 3. Spark 应用程序的配置 4. Spark 2.X Shell 基本使用 5. Spark 2.X submit 基本使用 6. Spark 2.X 的算子 7. Spark 2.X Cache 机制 8. Spark 2.X 宽依赖与窄依赖 9. Spark 2.X 数据持久化机制 10. Spark 2.X 参数配置调优说明	action 案例实战 4. Spark Shell 使用 5. Spark Submit 使用 **某金融企业用户交易行为分析大数据平台：Spark 数据清洗模块开发** 所需技能点： 1. HDFS 基本使用 2. Spark Shell 使用 3. Spark Submit 使用 4. Sqoop 使用 5. Flume 使用	1. 掌握 Spark RDD 开发技巧，掌握 Spark 核心编程技术 2. 掌握 Spark-Shell 及 Submit 使用 3. 掌握 Spark 算子基本原理

模块四：Spark SQL 2.X				
时间	**章**	**节**	**实战案例**	**学习目标**
第一周 **时长：8 小时**	**第一章 Spark SQL 核心编程**	1. Spark SQL 发展历史过程 2. Spark SQL 环境搭建：Metastore 3. Spark Session 介绍 4. 使用编程方式将 RDD 转换为 DataFrame 5. Parquet 数据源之使用编程方式加载数据 6. Spark SQL JDBC 服务介绍	**某金融企业用户交易行为分析大数据平台：数据仓库模块开发：用户交易类型分析，用户消费趋势分析，用户交易时间段分析及可视化** 所需技能点： 1. 使用编程方式将 RDD 转换为 DataFrame 2. JSON 数据源复杂综合案例实战 3. JDBC 数据源复杂综合案例实战	1. 掌握 Spark SQL 环境搭建 2. 掌握 Spark SQL 核心原理

		7. Spark SQL MetaStore Server 构建 8. 深度了解 Spark SQL 运行计划与调优 9. Spark SQL DataSets API 实践完成数据分析	4. UDF 自定义函数实战 **某金融企业用户交易行为分析大数据平台：** **数据仓库模块：金融交易欺诈分析** 所需技能点： 1. 使用编程方式将 RDD 转换为 DataFrame 2. JSON 数据源复杂综合案例实战 3. JDBC 数据源复杂综合案例实战 **4.** UDF 自定义函数实战	

	第二章 Spark SQL 综合实战	1.JSON 数据源复杂综合案例实战 2.Hive 数据源复杂综合案例实战 3.JDBC 数据源复杂综合案例实战		1.掌握 SparkSQL 加载各种数据源的方式 2.掌握 Spark SQL JDBC 以及常见的优化策略
	第三章 用户自定义函数	1.UDF 自定义函数实战 2.UDAF 自定义聚合函数实战 3.Spark SQL 常见优化策略		1. 掌握 Spark SQL UDF, UDAF 编写方法 2. 掌握 Spark SQL 基本优化方法

模块五：流计算引擎：Spark Streaming 2.X + Storm 1.0 + Kafka				
时间	章	节	实战案例	学习目标

第一周 **时长：6 小时**	**第一章 Spark Streaming 实时计算**	1. Spark Streaming：大数据实时计算介绍 2. Spark Streaming：DStream 基本工作原理 3. Spark Streaming：StreamingContext 详解技能点 4. Spark Streaming：输入 DStream 和 Receiver 详解 5. Spark Streaming：DStream 的 transformation 操作概览	**某金融企业用户交易行为分析大数据平台：实时分析部分：用户来源分析，用户交易欺诈检测，异常客户甄别** 所需技能点： 1. Spark Streaming：DStream 的 transformation 操作概览 2. Spark Streaming：updateStateByKey	1. 掌握 Spark Streaming 核心原理 2. 掌握 Spark Streeaming Window，Update State 使用方法 3. 掌握 Storm 核心使用方法与原理

		6. Spark Streaming：updateStateByKey 以及基于缓存的实时 wordcount 程序 7. Spark Streaming 与 Spark SQL 协同工作	3. Spark Streaming：Window 操作	
	第二章 Spark Streaming 数据存储与调优	1. Spark Streaming：缓存与持久化机制 2. Spark Streaming：Checkpoint 机制 3. Spark Streaming：部署、升级和监控应用程序 4. Spark Streaming：容错机制以及事务语义详解		1. 掌握 Spark Streaming 与 Spark SQL 整合使用方案 2. 掌握 Spark Streaming 与 Storm 的区别，可以合理的的选择方案

第二周:6 小时	**第三章 Kafka 核心技术**	1. Kafka 的概念与功能 2. 常见消息系统对比分析 3. Kafka 的 Topics 和 Logs 4. Kafka 的分布式环境搭建 5. 消息生产者、消费者以及消息发布的不同模式 6. Kafka 的命令行工具	**某金融企业用户交易行为分析大数据平台：日志消息队列部分：使用 Kafka 实现企业数据存储** 所需技能点： 1. Kafka 高可以用环境原理与搭建 2. Kafka 的命令行工具	1. 掌握 Kakfa 核心原理 2. 掌握 Kafka Topics 等核心概念 3. 掌握 Kafka HA 高可用原理及配置方案

	第四章 Kafka 编程实战技术	1.搭建 Kafka 开发环境 2.开发 Kafka 的消息发送和接收组件代码 3.Kafka 的数据持久化 4.Kafka 性能优化 5.Kafka 消息和日志 6.Kafka + Spark Streaming 构建日志分析能力 7.Kafka 高可用实现原理：数据高可用保证	**Kafka与Spark Streaming实现用户交易趋势分析** 所需技能点： 1. Kafka + Spark Streaming 构建日志分析能力	1. 掌握 Kafka 常见的优化方案 2. 掌握 kafka 与 Spark Streaming 协同工作的方案

第三周：7 小时	**第五章 Storm 深度解析**	1. Storm 1.0 的安装与配置与入门 2. Storm 1.0 与 Zookeeper 的协作 3. Storm 1.0 的事务工作原理及事务 API 4. Storm 1.0 Trident 实现原理与 API 使用 5. Storm 1.0 Trident Spout 实践 6. Storm 1.0 Trident State 实现 7. Storm 1.0 Trident RAS API 8. Storm1.0 与 Spark 2.XStreaming 对比	**某金融企业用户交易行为分析大数据平台：实时分析部分 Storm 实现方案：用户来源分析，用户交易欺诈检测** 所需技能点： 1. Kafka 高可以用环境原理与搭建 2. Storm 的运行模式：Nimbus,Supervisor 3. Storm 工程的结构与构建方式:Toplogy,Spout,Bolt 4.Storm 拓扑结构与组件协作	1. 掌握 Storm 核心原理 2. 掌握 Storm Bolt 等核心概念 3. 掌握 Storm API HA 高可用原理及配置方案 4. 掌握 Storm 常见的优化方案 5. 掌握 kafka 与 Storm 协同工作的方案

模块六：基于 Spark 的大数据挖掘分析：Spark Mllib				
时间	章	节	实战案例	学习目标

第一周 **时长：7 小时**	**第一章：基于 Spark 的大数据挖掘大数据挖掘 (5 小时)**	1. 机器学习理论基础 a) 机器学习需要解决的问题 b) 机器学习与 SQL 的不同 c) 机器学习常见的算法 i. 分类，聚类，回归，关联 d) 机器学习项目的实施流程 e) 大数据时代的机器学习的特征 f) 人工智能时代的深度学习技术介绍：原理，TensorFlow 等 2. 大数据挖掘工具介绍：Mahout，Spark MLlib 3. Spark 2.X MLlib 原理，安装及配置。	**某金融企业用户交易行为分析** **大数据平台：算法模型部分：用户性别预测，用户聚类分析** 所需技能点： 1. 机器学习理论基础 2. 大数据挖掘工具介绍：Mahout，Spark MLlib 3. Spark 2.X MLlib 原理 4. 分类算法实现 5. 聚类算法实现	1. 掌握机器学习理论与开发方法 2. 掌握分类算法应用场景 3. 掌握聚类算法应用场景 4. 掌握关联规则算法应用场景 5. 掌握回归分析算法应用场景 6. 熟练掌握 Spark Mllib 与 Spark Ml 开发方法 7. 熟练掌握基于 Spark 的算法上线流程与开发方 8. 能胜任在企业中对于机器学习相关需求的理解与实现

		4. Spark 2.X ML 原理 5. Spark ML 与 Spark Mllib 的区别 6. 使用 Spark MLlib 开发基础机器学习算法上线流程 7. Spark ML 与 Spark DataFrame 协同工作 8. 分类算法 Spark Mllib & Spark ML 实现 a) 决策树算法实现 b) 贝叶斯算法实现 c) 逻辑回归实现		

		9. 聚类算法 Spark Mllib & Spark ML 实现 10.回归分析 Spark MLlib 与 Spark Ml 实现 a) 线性回归实现 b) 时间序列实现 11.关联规则 Spark 实现 a) FPGrowth 算法实现购物篮分析 12.Spark MLlib 其它类库实现：推荐系统等 13.Spark MLlib 过程中常见的问题		

	第二章：Spark GraphX 实践 (2 小时)(选修)	1. 分布式图计算框架的目的及图原理 a) 图存储模式 b) 图基本概念 c) 常见的图计算框架：Spark GraphX, GraphLab 2. Spark GraphX 是什么 a) Spark GraphX 是什么 b) Spark GraphX 架构 c) Pregel 模式介绍 d) Spark GraphX 基础概念：Vertex，Edge	**GraphX 实战案例：社群发现，PageRank 实现关系分析** 所需技能点： 1. Spark GraphX 基础发现 2. Spark GraphX 开发实践	

		e) Spark GraphX 应用场景案例：阿里，JD 等 3. Spark GraphX 开发实践 a) 图算法工具包介绍 b) Table Operators c) Graph Operators d) 案例：GraphX 实现 Page Rank 算法 e) 案例：GraphX 实现社群发现		

模块七：大数据 MPP 数据库最佳实践

EDU
CSDN学院 IT实战派

时间	章	节	实战案例	学习目标
第一周 **时长：7 小时**	**第一章 大数据 MPP 数据库 Impala (5 小时)**	1. Impala 原理与架构：Impalad， 2. Impala 与 Hive 的不同 3. Impala 安装部署 4. Impala Shell 使用 5. 采样 Impala 数据分析 a) 基础语法 b) 数据类型 c) 过滤，排序 and Limit Results 6. 链接和组队数据	**某金融企业用户交易行为分析大数据平台：MPP 部分：使用 Impala 实现交易分析实时数据数据仓库实现** 所需技能点： 1. Hue 实现 Hive 2. Hue 实现操作 HBase	1. 掌握 Impala 核心原理应用场景 2. 掌握操作 Impala 数据建模方法 3. 掌握 Impala 操作 Hive 方法 4. 掌握 Impala 基础调优方法使用 5. 掌握 Impala 链接与组队数据

		7. Impala 调优分析		
	第二章大数据 MPP 工具 Presto (选修)(2 小时)	1. Presto 技术原理与架构 2. presto 架构 3. presto 低延迟原理 4. presto 存储插件 5. presto 执行过程 6. presto 引擎对比 7. presto 演示	**Azkaban 操作 Hive，MapReduce 案例：建表，导数据，SQL 分析** 所需技能点： 1. Azkaban 技术原理与架构 2. Azkaban 部署实施 Azkaban 实现调度 Hadoop job，Spark job	1. 掌握 Presto 核心原理 2. 掌握 Presto 基本操作数据的方式 3. 掌握 Presto 优缺点及常见的应用场景

EDU
CSDN学院 IT实战派

模块八：大数据应用调度工具使用及企业平台实战				
时间	**章**	**节**	**实战案例**	**学习目标**
第一周 **时长：7 小时**	**第一章 Hadoop 高级客户端工具 Hue**	1. Hue 原理与架构 2. Hue 安装部署 3. Hue 实现操作 HDFS 4. Hue 实现操作 Hive 5. Hue 实现操作 HBase 6. Hue 实现 WorkFlow	**某金融企业用户交易行为分析大数据平台：ETL 配置部分：使用 Hue 平台，实现测试的测试与 ETL 流程的配置开发** 所需技能点： 3. Hue 实现 Hive 4. Hue 实现操作 HBase	1. 掌握 Hue 核心原理 2. 掌握 Hue 操作 HDFS 方法 3. 掌握 Hue 操作 Hive 方法 4. 掌握 Hue 操作 HBase 方法

	第二章 Hadoop 高级调度工具：Azkaban	1. Azkaban 技术原理与架构 2. Azkaban 部署实施 3. Azkaban 实现调度 Hadoop job，Spark job	**某金融企业用户交易行为分析大数据平台：ETL 配置部分：使用 Azkaban 平台，实现测试的测试与 ETL 流程的配置开发** 所需技能点： 3. Azkaban 技术原理与架构 4. Azkaban 部署实施 Azkaban 实现调度 Hadoop job，Spark job	1. 掌握 Azkban 核心原理 2. 掌握 Azkaban 操作 Mapreduce 方法 3. 掌握 Azkaban 操作 Hive 方法
第二周 时长：7 小时	**第三章 Hadoop 企业级平台**	1. Hadoop 企业级平台介绍：互联网公司与传统行业的不同，华为，星环，Cloudera 等公司介绍	**使用 Cloudera 进行集群搭建与使用** 所需技能点： 1. Hadoop 企业级平台 Cloudera 介绍与搭建	1. 掌握企业级大数据平台使用 2. 了解大数据企业与平台 3. 熟悉企业级大数据平台

		2. Hadoop 企业级平台 Cloudera 介绍与搭建 3. Cloudera 架构介绍 4. Cloudera 基本使用 5. Hadoop 企业级平台 HDP 介绍 6. 企业级架构大数据方案说明	2. Cloudera 基本使用	架构方案

第二阶段实战项目		
项目名称	**项目需求**	**其他**

某金融企业用户交易行为分析大数据平台 **时长：6 小时**	项目背景： 利用大数据技术的部署，未来企业客户将拥有一个全面的实时分析平台，可洞察重要的业务事项，如客户流失预测、产品建议和欺诈警示。其中，客户流失预测是业务最为敏感的领域。几乎任何公司都面临客户流失的问题，尤其对于电信公司，保险公司，信用卡公司，有线电视，这类依赖周期性循环消费业务模型的公司。因此，客户流失管理已成为一个重要的竞争武器。本项目提供基于客户关系管理系统，金融信用系统，银行业务数据，现金交易系统的历史数据。通过整个多个交易系统的用户及交易数据实现用户购买行为预测。通过多个交易系统的用户及交易数据实现用户对汽车保险购买意向的预测。 项目难点： 1.多个 Spark 任务运行的时候，如何保证执行效率 2.HDFS 文件夹规划的一般方法，如何划分权限，避免出现用户使用问题 3.如何优化 Spark 程序使其更好的完成训练 4.Spark SQL 与 Hive 协同工作使用方式 5.大量的交易数据如何解决占用存储空间问题 项目知识要点： HDFS 知识要点： 1. 如何设计存储目录 2. 如何定期清理 HDFS 数据 3. 如何定期实现脚本化数据存储	

	Spark 知识要点： 1. Spark 读取 HDFS Block 数据的方式与方法 2. Spark 框架 Exectuor 设计方法 3. Spark 参数优化分析 4. Spark 与 Yarn 协同工作 5. Spark 与 Yarn 资源情况下工作方式 Spark SQL 知识要点： 1. Spark SQL 读取 Hive 数据 2. Spark SQL 定时执行脚本设计 3. Spark SQL 数据倾斜优化 4. Spark SQL 数据仓库构建 Hive 知识要点： 1. Hive 数据模型设计 2. Hive 动态数据仓库设计 3. Hive SQL 脚本编写 Sqoop 知识要点： 3. 如何实现与 MySQL 数据库对接 4. 大数据量情况下 Hbase 代码优化	

	项目技术选型： HDFS＋Yarn+＋Hive+ HBase+Sqoop+Flume +Zookeeper＋MySQL+Spark SQL 2.X+Spark Streaming 项目业务实现： 项目需求 1: 某金融交易大数据分析平台搭建与架构设计 项目需求 2: 某金融企业交易大数据平台 ETL 脚本设计及存储设计：Sqoop+Flume 项目需求 3: 某金融企业交易大数据平台分析模块 - 使用 Spark SQL 进行交易数据分析：用户交易类型，用户消费趋势分析，用户交易时间段分析 项目需求 4: 某金融企业交易大数据平台分析模块 - 金融交易欺诈分析 项目需求 5: 某金融企业交易大数据平台分析模块 - 使用 MySQL 进行结果存储 项目需求 6: 项目测试	

某企业运营数据实时指挥室 **时长：6 小时**	项目背景： 大数据时代，企业需求实时监控企业运营状态。通常，我们将构建运营数据实时指挥室。某企业是一家全国性质的连锁企业，希望通过大数据技术，实时监控各个门店经营及运转情况。 项目难点： 1.多个 Storm 任务运行的时候，如何保证执行效率 2.HDFS 文件夹规划的一般方法，如何划分权限，避免出现用户使用问题 3.如何优化 Storm 程序使其更好的完成训练 4.Spark SQL 与 Storm 协同工作使用方式 项目知识要点： HDFS 知识要点： 1. 如何设计存储目录 2. 如何定期清理 HDFS 数据 3. 如何定期实现脚本化数据存储 Spark 知识要点： 1. Spark 读取 HDFS Block 数据的方式与方法 2. Spark 框架 Exectuor 设计方法 3. Spark 参数优化分析 4. Spark 与 Yarn 协同工作	

	5. Spark 与 Yarn 资源情况下工作方式 Spark SQL 知识要点： 1. Spark SQL 读取 Hive 数据 2. Spark SQL 定时执行脚本设计 3. Spark SQL 数据倾斜优化 4. Spark SQL 数据仓库构建 Hive 知识要点： 1. Hive 数据模型设计 2. Hive 动态数据仓库设计 3. Hive SQL 脚本编写 Sqoop 知识要点： 1. 如何实现与 MySQL 数据库对接 2. 大数据量情况下 Hbase 代码优化 Storm 知识要点 1. Storm 与 Kafka 整合要点 2. Storm 优化 3. Storm 容错 项目技术选型：	

	HDFS＋Yarn+＋Hive+ HBase+Sqoop+Flume +Zookeeper＋MySQL＋Storm 1.0 +Spark SQL 2.X 项目业务实现： 项目需求 1: 某企业实时运营指挥室大数据分析平台搭建与架构设计 项目需求 2: 某企业实时运营指挥室大数据平台 ETL 脚本设计及存储设计：Sqoop+Flume 项目需求 3: 某企业实时运营指挥室数据平台实时分析模块 - 使用 Spark SQL 进行交易数据分析：用户交易类型，用户消费趋势分析，用户交易时间段分析，用户采购商品分析 项目需求 4: 某企业实时运营指挥室大数据平台分析模块 - 使用 MySQL 进行结果存储 项目需求 5: 项目测试	

EDU
CSDN学院 IT实战派

结业实战项目：某运营商 O 域用户行为分析项目		
项目名称	**项目需求**	**其他**
某欧洲运营商 O 域用户行为分析项目	项目背景： 利用某年，某欧洲城市的电信运营商的用户活动日志数据，包括：电话呼叫 GEO 记录，呼叫详细记录（CDR 日志），SMS 日志，上网流量记录，呼叫记录，本地呼叫，长途呼叫； 采用详细的日志分析精准分析移动运营商的用户行为。包括用户流量运营分析，用户画像平台，用户精准营销，实时推荐等场景。 项目难点： 1.多种技术混合结构，Spark，Hadoop 如何进行合理规划 2.大规模数据分析，如何设计存储 3.如何优化 Spark 程序使其更好的完成训练 4.实时计算场景如何实现 5.大量的交易数据如何解决占用存储空间问题 项目知识要点： HDFS 知识要点： 1. 如何设计存储目录	

时长：12 小时	2. 如何定期清理 HDFS 数据 3. 如何定期实现脚本化数据存储 Spark 知识要点： 1. Spark 读取 HDFS Block 数据的方式与方法 2. Spark 框架 Exectuor 设计方法 3. Spark 参数优化分析 4. Spark 与 Yarn 协同工作 5. Spark 与 Yarn 资源情况下工作方式 Spark SQL 知识要点： 5. Spark SQL 读取 Hive 数据 1. Spark SQL 定时执行脚本设计 2. Spark SQL 数据倾斜优化 3. Spark SQL 数据仓库构建 Hive 知识要点： 1. Hive 数据模型设计 2. Hive 动态数据仓库设计 3. Hive SQL 脚本编写 Sqoop 知识要点： 3. 如何实现与 MySQL 数据库对接	

	1. 大数据量情况下 Hbase 代码优化 项目技术选型: HDFS＋Yarn+＋Hive+ HBase+Sqoop+Flume +Zookeeper＋MySQL＋Spark Streaming+ Spark SQL+Spark 项目业务实现: 项目需求 1: 运营商 O 域大数据分析平台搭建与架构设计 项目需求 2: 运营商 O 域大数据分析平台 ETL 脚本设计及存储设计：Sqoop+Flume ＋Spark Core 项目需求 3: 运营商 O 域大数据分析平台分析模块 - 使用 Spark SQL 进行交易数据分析：用户上网时段分布及变化，用户终端品牌型号特征及变化，用户活跃度区域分布及变化，用户上网流量统计及变化，用户上网流量统计及变化，用户区域分布特征及数量变化，用户画像 项目需求 4: 运营商 O 域大数据分析平台实时模块 - 基于用户画像的实时营销引擎实现	

	项目需求 5: 运营商 O 域大数据分析平台结果展示模块 - 使用 MySQL 进行结果存储 项目需求 6: 项目测试	

课程概览与课时分配:

模块一：Linux 实践与 Java 基础

周期：2 周

课时：12 小时

模块二：Hadoop 生态体系(HDFS+MapReduce+Pig(选修)+HBase+Zookeeper+Hive+Sqoop+Flume)

周期：5 周

课时：33 小时

模块三：Spark 核心架构：Scala+Spark 核心

周期：2 周

课时：14 小时

模块四：Spark SQL 2.X

周期：1 周

课时：8 小时

模块五：流计算引擎：Spark Streaming 2.X + Storm 1.0 + Kafka

周期：3 周

课时：19 小时

模块六：基于 Spark 的大数据挖掘分析：Spark Mllib(选修))

周期：1 周

课时：7 小时

模块七：大数据 MPP 数据库最佳实践：Impala+Presto(选修)

周期：1 周

课时：7 小时

模块八：大数据应用调度工具使用及企业平台实战（Hue+Azakban+Cloudera）

周期：2 周

课时：14 小时

大数据项目

项目一：某企业微博社交数据分析平台（6 小时）(贯穿项目)

项目二：某企业电商数据运营分析平台（6 小时）（Hadoop 学习完以后，综合练习)

项目三：某金融企业用户交易行为分析大数据平台（6 小时）(贯穿项目)

项目四：某企业运营数据实时指挥室（6 小时）(Spark 学习完以后，综合练习)

结业项目：某运营商 O 域用户行为分析项目（12 小时）