课程大纲

课程名称	课程内容	课时
第一阶段	**开发环境搭建，Java 基本语法，方法，数组，类与对象，面向对象特点，抽象类与接口，包机制，开发工具，异常，常用类，集合框架，多线程，IO 流**	**20 天**
Java 基础	从企业项目开发角度重新诠释讲解 Java 语言。在教学过程中特别突出 Java 语言的跨平台原理，Java 语言的陷阱以及注意事项等，以大量实例分析着重介绍常用类库与程序结构，字符串、数组、方法、面向对象、抽象类与接口、集合框架、IO 流、递归和位运算。强化学员 Java 编程的代码能力和编码调试能力。使得学员具有扎实的 Java 语言开发功底	
第二阶段	**Linux 的规则与安装，Linux 文件、目录与磁盘格式，shell 与 shellscript，Linux 使用者管理，Linux 系统管理员**	**10 天**
Linux 基础	Linux 是什么，如何学习 Linux，主机规划与磁盘分区，安装 CentOS5.x，Linux 的文件权限与目录配置，Linux 文件与目录管理，Linux 磁盘与文件系统管理，文件与文件系统的压缩与打包，vim 程序编辑器，认识与学习 bash，正则表达式与文件格式化处理，shellscript，Linux 账号管理与 ACL 权限设	

	置，磁盘配额(Quota)与高级文件系统管理，例行性工作(crontab)，程序管理与 SELinux 初探，认识系统服务(daemons)，认识与分析日志文件，启动流程、模块管理与 Loader，系统设置工具(网络与打印机)与硬件检测，RPM、SRPM 与 YUM 功能，Linux 备份策略	
第三阶段	**Hadoop 基础，数据分析引擎 Pig，数据仓库 Hive，关于 HBase，关于 ZooKeeper，关于 Sqoop，Mahout**	**20 天**
Hadoop 基础	数据的存储与分析，相较于其他系统的优势，Hadoop 发展简史，Apache Hadoop 和 Hadoop 生态系统，Hadoop 的发行版本，气象数据集，使用 Hadoop 来分析数据，横向扩展，HadoopStreaming，HadoopPipes，HDFS 的设计，HDFS 的概念，命令行接口，Hadoop 文件系统，Java 接口，数据流，通过 Flume 和 Sqoop 导入数据，通过 distcp 并行复制，Hadoop 存档，数据完整性压缩，序列化 Writable 集合，序列化框架 Avro，基于文件的数据结构，用于配置的 API，配置开发环境，用 MRUnit 来写单元测试，本地运行测试数据，在集群上运行，作业调优，MapReduce 的工作流，剖析 MapReduce 作业运行机制，任务失败，作业的调度，shuffle 和排序，	

	任务的执行，MapReduce 的类型，MapReduce 输入格式，MapReduce 输出格式，计数器，排序，连接，边数据分布，MapReduce 库类，集群规范，集群的构建和安装，SSH 配置，Hadoop 配置，YARN 配置，安全性，利用基准评测程序测试 Hadoop 集群，云端的 Hadoop，HDFS HDFS 监控，HDFS 维护	
Pig 数据分析	安装与运行 Pig， Grunt PigLatin 语言 用户自定义函数 数据处理操作 Pig 实战	
数据仓库 Hive	安装 Hive 运行 Hive Hive 与传统数据库相比 HiveQL hive 表 查询数据 用户定义函数	
关于 HBase	HBase 基础 概念	

	安装 Thrift 客户端 HBase 和 RDBMS 的比较	
关于 ZooKeeper	安装和运行 ZooKeeper ZooKeeper 服务 使用 ZooKeeper 来构建应用 生产环境中的 ZooKeeper	
关于 Sqoop	获取 Sqoop Sqoop 连接器 一个导入的例子 生成代码 深入了解数据库导入 使用导入的数据 导入大对象执行导出 深入了解导出功能	
Mahout	了解 mahout 了解常用的算法 协同过滤算法 taste 了解聚类算法 了解分类算法	
第四阶段	**python 简介，python 基本语法和数据类型，Python 中的函数，面向对象的 python，Python 中的正则**	**10 天**

	表达式了解，Python 语言的发展，Python 常用类库，Python 爬虫采集网络数据	
Python 基础	python 数据类型和核心语法 Python 中元组、列表、字典使用 Python 中函数的定义和调用方法 Python 的开发规范 Python 类的声明和使用 Python 的正则表达式操作	
数据采集	爬虫基础 爬虫的原理 数据集成 数据清洗 数据转换 掌握数据的展现	
第五阶段	**Scala 语言，Scala 的基础，Scala 的常用数据类型，Scala 的高阶函数，Scala 的隐式转换，Scala 实战实现 Akka，Spark 数据分析，Spark 下载与入门，RDD 编程，spark 键值对操作，数据读取与保存，Spark 编程进阶，在集群上运行 Spark，Spark 调优与调试**	**20 天**
Scala	Scala 的特性以及它的编程思想 Scala 的基本语法以及与 Java 的异同	

	Scala 的常用数据类型 Scala 的常用数据类型如数组，元祖等 Scala 的类和对象 Scala 的匹配模式和样本类 Scala 的高阶函数如柯里化 Scala 的隐式转换 Akka 的原理和实现	
Spark 数据分析	Spark 是什么 一个大一统的软件栈 Spark 的用户和用途 Spark 简史 Spark 的版本和发布 Spark 的存储层次	
Spark 下载 与入门	下载 Spark Spark 中 Python 和 Scala 的 shell Spark 核心概念简介 独立应用	
RDD 编程	RDD 基础 创建 RDD RDD 操作 向 Spark 传递函数 常见的转化操作和行动操作	

	持久化(缓存)	
数据读取 与保存	文件格式 文件系统 SparkSQL 中的结构化数据 数据库	
Spark 编程进阶	简介 累加器 广播变量 基于分区进行操作 与外部程序间的管道 数值 RDD 的操作	
在集群上运行 Spark	Spark 运行时架构使用 spark-submit 部署应用 打包代码与依赖 Spark 应用内与应用间调度 集群管理器 选择合适的集群管理器	
Spark 调优与调试	使用 SparkConf 配置 Spark Spark 执行的组成部分：作业、任务和步骤 查找信息 Spark 网页用户界面 驱动器进程和执行器进程的日志 关键性能考量	

SparkSQL	连接 SparkSQL 在应用中使用 SparkSQL 读取和存储数据 JDBC/ODBC 服务器 用户自定义函数 SparkSQL 性能	
Spark Streaming	一个简单的例子 架构与抽象 转化操作 输出操作 输入源 核心数据源 24/7 不间断运行 Streaming 用户界面 性能考量	
基于 MLlib 的机器学习	系统要求 机器学习基础 数据类型 算法 提示与性能考量 流水线 API	
第六阶段	**Kylin 与 Nifi**	**10 天**

Kylin 概述	背景和历史、Kylin 的工作原理、维度和度量简介、Cube 和 Cuboid、工作原理。、Kylin 的技术架构、Kylin 的主要特点、标准 SQL 接口、支持超大数据集、可伸缩性和高吞吐率、BI 及可视化工具集成、与其他开源产品比较	
Kylin 快速入门	核心概念、数据仓库、OLAP 与 BI、维度和度量、事实表和维度表、Cube、Cuboid 和 Cube Segment、在 Hive 中准备数据、星形模型、维度表的设计、Hive 表分区、了解维度的基数、SampleData、设计 Cube、导入 Hive 表定义、创建数据模型、创建 CubP、构建 Cube、全量构建和增量构建、历史数据刷新、合并、查询 CubP、SQL 参考	
增量构建	为什么要增量构建、设计增量 Cube、设计增量 Cube 的前提、增量 Cube 的创建、触发增量构建、Web GUI 触发、构建相关的 Rest API、管理 Cube 碎片、合并 Segment、自动合并、保留 Segment、数据持续更新	
流式构建	为什么要流式构建、准备流式数据、数据格式、消息队列、创建 Schema、设计流式 Cube、创建 Model、创建 Cube、流式构建原理、触发流式构建、单次触发、自动化多次触发、出错处理	

查询和可视化	Web GUI、查询、显示结果、Rest API、查询认证、查询请求参数、返回结果、ODBC、JDBC、获得驱动包、认证、URL 格式、获取元数据信息、通过 Tableau 访问 Kylin、连接 Kylin 数据源、设计数据模型、通过 Live 方式连接、白定义 SQL、可视化、发布到 Tableau Server、Zeppelin 集成、Zeppelin 架构简介、Kylin Interpreter 的工作原理、如何使用 Zeppelin 访问 Kylin、小结	
Cube 优化	Cuboid 剪枝优化、维度的诅咒、检查 Cuboid 数量、检查 Cube 大小、空间与时间的平衡、剪枝优化的工具、使用衍生维度、使用聚合组、并发粒度优化、Rowkeys 优化、编码、按维度分片、调整 Rowkeys 顺序、其他优化、降低度量精度、及时清理无用的 Segment	
应用案例分析	基本多维分析、数据集、数据导入、创建数据模型、创建 Cube、构建 Cube、SQL 查询、流式分析、Kafka 数据源、创建数据表、创建数据模型、创建 Cube、构建 Cube、SQL 查询	
扩展 Kylin	可扩展式架构、工作原理、三大主要接口、计算引擎扩展、Engine Factory、MR Batch Cubing Engine、Batch Cubing Job Builder、IMRInput、IMROutput、数据源扩展、存储扩展、聚合类型扩展、聚合的 JSON	

	定义，聚合类型工厂、聚合类型的实现、维度编码扩展、维度编码的 JSON 定义、维度编码工厂、维度编码的实现	
Kylin 的企业级功能	身份验证、自定义验证、LDAP 验证、单点登录、授权	
运维管理	安装和配置、必备条件、快速启动 Kylin、配置 Kylin、企业部署、监控和诊断、日志、任务报警、诊断工具、日常维护、基本运维、元数据备份、元数据恢复、系统升级、垃圾清理、常见问题和修复、获得社区帮助	
Kylin 的未来	Apache 开源社区简介、大规模流式构建、拥抱 Spark 技术栈、更快的存储和查询、前端展现及与 BI 工具的整合、高级 OLAP 函数、展望	
Apache Nifi	Nifi 是什么 Nifi 的核心概念 Nifi 架构 Nifi 的性能和特性 Nifi 关键特性	
Nifi 入门	术语介绍 Nifi 用户界面 多用户认证登录界面 自定义属性	

	控制服务 报告任务	
Nifi 的表达式语言	Nifi 表达的结构 数据类型和函数	
Nifi 系统管理员	安全配置 用户认证 多用户认证 加密解密配置 在配置文件对密码进行加密 Nifi 集群配置 状态管理 Kerberos 服务 系统属性 声明管理	
第七阶段	**云计算 openstack 平台，就业指导，职业规划**	**5 天**
大数据云计算融合生产环境讲解	Openstack 云平台介绍，Openstack 手动部署，Openstack 自动部署，生产环境中的集群配置以及容灾策略讲解，CDH 集群安装与部署，Hadoop 集群管理。	
就业指导与职业规划	就业老师根据学员意向及学习情况，对学员进行就业指导和职业规划。包括面试技巧，笔试题精讲，模拟面试，职业生涯规划等，择优推荐用友企业	