附件 1

大数据开发工程师培训内容

时间	课程模块	内容	课时
初级课程（32 课时）			
第一天	大数据清洗处理：Kettle ETL 工具	ETL 概述与 Kettle 基础操作 数据源抽取 数据预处理转换 数据迁移和装载 高级转换 任务的构建 项目实战:无人售货机 综合案例：连接 Hadoop 向 HDFS 进行 ETL 数据同步	4
	数据标注	数据标注概述 图像视频标注 文本语音标注 综合案例：基于 SnowNLP 手机商城文本案例分析	4
第二天	大数据存储及分析：Hive 数据仓库	Hive 基础入门 Hive DDL 操作 Hive DML 操作 Hive DQL 操作 开窗函数 Hive 函数 Hive 高级操作 项目实战：国内主要城市房屋出租情况统计分析 综合案例：Kettle 连接 Hive 进行 ETL 分析生成式人工智能赋能大数据分析	8
第三天	数据分析工具：Excel	数据预处理 数据分析 数据可视化	4
	前端数据可视化：Echarts	Echarts 基础 使用 Echarts 实现基本图形 Echarts 常用配置项	
	Python 数据可视化：Matplotlib	图表的常用设置 常用图表的绘制 数据统计分析案例	4
	数据可视化项目实战	Excel、Echarts、Matplotlib 可视化实战 综合案例：Hive 分析结果的数据可视化	
第四天	模拟大数据竞赛样题（初级）	综合案例讲解	8

时间	课程模块	内容	课时
中级课程（32 课时）			
第一天	大数据日志收集：Flume 数据采集工具	Flume 环境搭建 Flume 单代理流应用：实时监控文件到 HDFS Flume 多代理流应用：聚合、复制和多路复用 Flume 自定义组件应用	4
	Hadoop 批处理调度器：Azkaban	工作流调度系统概述 Azkaban 概论 Azkaban 的安装部署 Azkaban 单一任务调度管理 Azkaban 多任务调度管理 综合案例：Hive 脚本任务调度管理	4
第二天	非结构化大数据处理：Hbase 分布式数据库	HBase 简介 HBase 环境搭建 HBase 快速入门 HBase shell 操作 HBase 的 JavaAPI 数据的导入导出 项目实战：基于 HBase 的员工信息管理系统 综合案例：HBase 与 MySQL 的数据互导 综合案例：Kettle 连接 HBase 进行 ETL 过程 综合案例：HBase 与 Hive 的集成	8
第三天	BI 工具数据可视化：Tableau	常用图表的绘制 项目实战：制作参与调研的大学生信息图表 项目实战：制作物资采购分析图表 项目实战：制作电影数据分析图表 模型预测：回归分析和时间序列分析	2
	Python 前端数据可视化：PyECharts	PyEcharts 快速上手 PyEcharts 全局配置项 PyEcharts 系列配置项 PyEcharts 绘制常用图表 项目实战：平台订单数据可视化 项目实战：销售数据可视化	2
	自然语言处理	基于 n-gram 模型的中文分词 使用 LSTM 对电影评论进行情感分析 英文文本分类 LSTM 时间序列预测	4
第四天	模拟大数据竞赛样题（中级）	综合案例讲解	8

时间	课程模块	内容	课时
高级课程（40 课时）			
第一天	分布式发布订阅消息系统：Kafka	Kafka 安装部署 Kafka 基本操作 Kafka 生产者应用程序开发 Kafka 消费者基本配置以及 Java API 应用 Kafka 拦截器 Kafka Stream 项目实战：实时数据分析实现网页浏览量统计	4
	大数据流式处理框架：Flink	Flink 快速入门 Flink 批处理 API Flink 流处理 API FlinkML 应用编程 项目实战：股票行情数据统计分析项目 综合案例：Flink 与 Kafka 集成处理实时数据流	4
第二天	大数据计算引擎：Spark	Spark 简介 Spark 程序结构：核心组件和算子 弹性分布式数据集 RDD 常见类型数据的读取 Spark Streaming 实时数据流处理 Spark Mllib 机器学习 项目实战：用户在基站停留时间 综合案例：Flume 集成 Spark Streaming 综合案例：Spark Streaming 整合 Kafka	8
第三天	大数据安全之 Kerberos 认证	Kerberos 简介 Kerberos 认证原理 Kerberos 使用	2
	Hadoop 数据同步工具：Sqoop	Sqoop 环境搭建 Sqoop 数据导入与导出 综合案例：HDFS 或 Hive 数据导出到 MySQL 综合案例：HBase 数据导出到 MySQL 项目实战：购物平台用户行为数据分析	2
	大数据集群任务调度：应用容器引擎 Docker	Docker 简介 Docker 平台部署安装 常用容器命令	2
	大数据集群任务调度：Apache DolphinScheduler	DolphinScheduler 简介 DolphinScheduler 系统架构 DolphinScheduler 工作流	2

时间	课程模块	内容	课时
第四天	BI 工具数据可视化:FineBI	初识 FineBI:安装与第一个案例 选择并编辑数据 制作可视化组件 项目实战：部门考勤数据分析 项目实战：办公用品销售数据分析	2
	数据可视化大屏：Vue+Echarts	初识 Vue：基本项目搭建 综合案例：医疗健康数据可视化 综合案例：零售行业数据可视化 综合案例：工业制造行业数据可视化	2
	基于机器学习的大数据挖掘	随机森林：基于相似度的酒店推荐系统 文本处理相似度计算：商品销售额预测	4
第五天	模拟大数据竞赛样题（高级）	综合案例讲解	8