大数据开发工程师

1 大数据技术概论

1 大数据技术的起源和发展
1.1 课程大纲
1.2 大数据技术的起源
1.3 互联网泡沫：大数据技术的发端
1.4 重识大数据技术

2 大数据技术与相关领域的关系
2.1 大数据与云计算
2.2 大数据与区块链
2.3 大数据与人工智能

3 大数据管理技术概述
3.1 大数据管理技术概述
3.2 大数据存储技术
3.3 大数据事务处理技术
3.4 大数据查询处理技术
3.5 人机交互技术

4 大数据应用
4.1 “双十一”与海量支付 4.2 商品推荐：亚马逊的秘密武器 4.3 流立方与金融反欺诈 4.4 关联分析与投资组合 4.5 群组分析：洞悉人们的行为趋势

2 数据平台综述

1 问题回顾
1.1 课程介绍
1.2 数据管理技术的演化
1.3 学习的方法论

2 数据平台设计理念
2.1 分布式系统可扩展性
2.2 分布式系统CAP理论
2.3 用分布式理论扩展关系数据库
2.4 BASE原则和NoSQL系统
2.5 小结

3 简单说一说选型
3.1粗识大数据平台

3 数据存储：HDFS

1 基础架构
1.1 课程介绍
1.2 背景
1.3 HDFS架构
1.4 HDFS读写
1.5 副本放置策略

2 部署配置
2.1 部署安装
2.2 部署实操

3 管理使用
3.1 HDFS管理与使用

4 高级内容
4.1 HDFS高可用
4.2 HDFS联邦
4.3 HDFS安全
4.4 压缩与分片

5 异常处理
5.1 异常处理

4 日志解析及计算：MR

1 MapReduce的基本原理和运行流程
1.1 MR的应用场景
1.2 MR的原理和运行流程
1.3 编写一个MR程序

2 MR编程实战
2.1 Hadoop的IO模型
2.2 完整编写Map和Reduce
2.3 灵活使用Configuration
2.4 精准控制Shuffle过程
2.5 MR程序的输入
2.6 MR程序的输出
2.7 简单好用的计数器

3 案例实操
3.1 MR实现关联操作

4 MR性能调优
4.1 MR参数调优
4.2 数据倾斜

5 数据获取和预处理：Flume

1 日志及日志收集系统
1.1 课程介绍
1.2 日志及日志收集系统

2 Flume设计原理
2.1 Flume Agent组成
2.2 Flume支持的组件类型
2.3 Flume基本配置

3 Flume安装部署
3.1 Flume-ng部署

4 Flume配置示例
4.1Flume配置示例

5 实战
5.1Flume高级配置
5.2构建复杂日志收集系统

6 结构化查询：Hive

1 从MR到Hive
1.1 Hive解决了什么问题
1.2 Hive擅长什么

2 Hive 系统介绍
2.1 Hive结构与数据仓库
2.2 数据模型与元数据

3 Hive的安装及调试
3.1 Hive安装与配置
3.2 创建和管理Hive中的数据库

4 Hive查询语法
4.1 写一个基本的查询语句
4.2 子查询和关联表操作
4.3 使用简单函数
4.4 使用聚合函数
4.5 利用正则表达式精确提取信息
4.6 窗口函数的使用
4.7 “行转列”与“列转行”
4.8 用户自定义函数(UDF)的使用

5 案例
5.1 Hive优化案例

7 数据获取和预处理：Sqoop

1 来自于业务系统的数据
1.1 课程介绍
1.2 业务系统数据
1.3 数据同步与传统数仓

2 Sqoop功能与架构
2.1 sqoop功能与架构
2.2 数据划分

3 sqoop安装及配置
3.1 java, hadoop-client等基础依赖安装
3.2 sqoop服务安装

4 sqoop语法介绍
4.1 语法分析

5 案例
5.1案例

8 大数据调度框架：Azkaban

1 任务调度基本概念
1.1 课程介绍
1.2 调度系统背景知识

2 Azkaban系统介绍
2.1 架构组件和任务流程讲解

3 Azkaban的安装和配置
3.1 代码下载、编译、部署
3.2 插件的安装：hadoopJava、Spark等

4 Azkaban工作流调度实战
4.1 具体任务编写要点 和 DAG设计
4.2 不同调度参数详解

5 Azkaban进阶
5.1 如何实现web高可用
5.2 如何提高任务可用性
5.3 如何增加新的插件类型

9 Scala编程基础

1 Scala实战入门
1.1 安装Scala开发环境
1.2 Scala常用类型介绍
1.3 值与变量的声明
1.4 Scala函数与方法的定义和使用
1.5 默认参数、带名参数及变长参数
1.6 动手编写条件表达式
1.7 循环表达式与For循环的使用
1.8 异常处理

2 Scala面向对象入门实战
2.1 类的定义：属性与方法
2.2 不同的构造
2.3 object对象
2.4 apply方法
2.5 方法重写与字段重写
2.6 抽象类

2.7 trait
2.8 case class
2.9 模式匹配

3 Scala集合类详解
3.1 集合
3.2 序列
3.3 可变列表与不可变列表
3.4 集合操作

4 Scala高级特性实战
4.1 隐式转换
4.2 隐式参数
4.3 隐式类

10 Spark框架教学

1 spark基础
1.1 Spark概述
1.2 Spark安装
1.3 什么是RDD?
1.4 RDD的创建方式
1.5 RDD基本操作

2 df与ds的基础
2.1 DataSet与DataFrame概述
2.2 DataSet的创建方式
2.3 DataSet基本操作
2.4 DataFrame的创建方式
2.5 DataFrame基本操作

3 SparkSQL
3.1 SparkSQL前世今生
3.2 SparkSQL使用
3.3 UDF开发
3.4 SparkSql调优

4 Spark调优
4.1 共享变量（广播变量，累加变量）
4.2 持久化
4.3 使用高性能的算子
4.4 其他

11 大作业：网站分析大数据框架调度

实战大作业：网站分析大数据框架调度作业

【基于网易云私有集群环境，使用网易提供的脱敏数据库及日志】

第一步：学员通过flume将日志同步到hdfs，mr解析日志到hive表

第二步：数据库通过sqoop同步到hive表，按照给定的统计口径，将结果同步到mysql数据库或者hdfs文件系统

第三步：通过azkaban配置任务依赖，至少保证3天的稳定运行

第四步：梳理设计文档并将代码打包上传提供评审