菜鸟窝大数据全能实战班课程目录

报名大数据课程送 价值1898元的 **Java Web工程师课程，** 点我报名

报名咨询:微信 phoebe2016524

一.Linux 操作系统

在企业中无一例外的是使用 Linux 来搭建或部署项目，在平常我们也经常在Linux环境下进行开发。进入大数据领域就要打好 Linux 基础，以便更好地学习Hadoop，Kafka，Spark，Storm，Redis 等众多课程。

1.linux来源与发展概况

2.Linux目前的发行版

3.Linux系统安装与流程说明

4.Linux目录结构介绍

5.Linux目录完整参数列表说明

6.Linux过滤目录输出列表

7.Linux常用文件命令：创建文件、复制文件、链接文件、重命名文件、删除文件

8.Linux常用目录命令：创建目录、删除目录

9.Linux查询文件统计信息

10.Linux查看文件类型

11.Linux查看整个文件、查看部分文件

12.Linux文件权限说明及权限修改

13.Linux压缩解压缩数据文件

14.Liunx检测磁盘空间

15.Linux用户与用户组，添加新用户，删除新用户，添加用户组和修改组

16.Linux sudo权限说明

17.Linux环境配置讲解、全局环境变量和局部环境变量、删除环境变量、设置PATH

18.Linux vim编辑器基础，编辑数据，复制和粘贴，查找和替换

19.Linux 进程： 探查进程，实时检测进程，结束进程

20.Linux Shell脚本编程基础

21.Linux shell结构化指令，for循环，if else语句详解

22.Linux 符号，输入输出重定向，管道符号说明

23.Linux shell脚本任务调度

实战项目:编写一个每天清理个人目录下.log后缀文件的脚本，清理到自定义回收站，每天2点执行

二.Hadoop生态圈（离线计算）

Hadoop是一种分析和处理大数据的软件平台，是Appach的一个用Java语言所实现的开源软件的加框，在大量计算机组成的集群当中实现了对于海量的数据进行的分布式计算。在本章节中不仅将用到前面的 Linux 知识，而且会对 hadoop 的架构有深入的理解，并未你以后架构大数据项目打下坚实基础

hadoop

1.hadoop 介绍，发展简史，诞生来由

2.hadoop 生态圈体系结构，组件说明

3.hadoop 伪分布式环境搭建及完全分布式环境说明

HDFS

1.HDFS分布式文件系统说明

2.HDFS block概念

3.HDFS namenode ,datanode 详解

4.HDFS HA 详解

5.HDFS命令行接口，读取数据

6.HDFS命令行接口，写入数据

7.HDFS命令行接口，删除数据

8.HDFS命令行接口，distcp跨集群分布式拷贝数据

9.HDFS压缩和分片

10.HDFS文件格式：textfile，sequencefile，rcfile，orcfile，parquet

11.HDFS各类文件格式比较

YARN

1.经典的Mapreduce 1结构弊端

2.Mapreduce 2 中YARN的引入

3.YARN的核心结构说明

4.YARN的工作机制

5.YARN的架构剖析

6.YARN 内置调度器：公平调度和容量调度

7.YARN上任务的执行环境

8.YARN上任务的推测执行机制

9.YARN上任务的JVM重用

Mapreduce

1.MapReduce整体流程说明

2.MapReduce目录输入，多目录输入，inputformat子类介绍

3.Mapreduce map 过程

4.Mapreduce combine过程

5.Mapreduce reduce 过程

6.Mapreduce结果输出，outputformat子类介绍

7.MapReduce世界的helloword之wordcount操作演练

8.MapReduce Wordcount 项目打包，运行

9.MapReduce 内置计数器含义讲解

10.Mapreduce 实例讲解之全排序

11.Mapreduce 实例讲解之部分排序

12.Mapreduce 实例讲解之 join map端连接

13.Mapreduce 实例讲解之 join reduce端连接

14.Mapreduce 实例讲解之 大矩阵相乘

15.Mapreduce自定义的format

16.Mapreduce MRUnit单元测试使用

Hadoop Streaming

1.hadoop streaming引入的目的

2.Hadoop streaming机制讲解

3.使用Python编写 hadoop streaming

4.使用Shell 编写 hadoop streaming

实战项目：分别使用原生mr和python版本的hadoop streaming，编写数据清洗，数据处理，压缩逻辑，结果dump到hdfs

三.Hive（数据仓库）

是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的SQL查询功能，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 。
Hive是工作中最常用到的部分，也是面试的重点

1.Hive 简介

2.Hive Hbase Pig三者的不同点

3.Hive 系统架构

4.Hive 安装搭建与常用参数配置

5.Hive shell命令使用

6.Hive 数据库数据表操作

7.Hive 数据导出

8.Hive 数据加载

9.Hive 外部表与分区表讲解

10.HiveQL 常用语句

11.HiveServer2讲解

12.Hive 函数介绍

13.Hive 分析函数与窗口函数

14.Hive 自定义UDF / UDAF函数

15.Hive 优化和安全

实战项目：结合flume等框架的日志处理

四.HBase（分布式数据库）

HBase是一个开源的非关系型分布式数据库（NoSQL），它参考了谷歌的BigTable建模，实现的编程语言为 Java。HBase在列上实现了BigTable论文提到的压缩算法、内存操作和布隆过滤器。HBase的表能够作为MapReduce任务的输入和输出，可以通过Java API来访问数据。也可以通过REST、Avro或者Thrift的API来访问。

虽然最近性能有了显著的提升，HBase 还不能直接取代SQL数据库。如今，它已经应用于多个数据驱动型网站，包括 Facebook的消息平台

1.HBase 综合概述

2.HBase 数据库特点

3.HBase 搭建

4.HBase Shell 操作讲解

5.HBase Java API 讲解

6.HBase 协处理器使用

7.HBase 与Mapreduce集成使用讲解

8.HBase backup master讲解

9.HBase 数据模型讲解

10.HBase 数据库数据存储与读取思想讲解

11.HBase 数据在线备份思路讲解

12.HBase 数据迁移与导入方案讲解

13.Region 寻址方式

14.HBase 二级索引构建方案

15.HBase RowKey设计原则

16.HBase 性能调优

五.Redis（缓存数据库系统）

Redis是的一款内存高速缓存数据库。Redis全称为：Remote Dictionary Server（远程数据服务），该软件使用C语言编写，支持网络、可基于内存亦可持久化的日志型、Key-Value数据库，并提供多种语言的API

1.Redis 简介

2.Redis 特性

3.Redis 应用场景

4.Redis 字符串类型 / 散列类型 / 列表类型 / 集合类型

5.Redis 的事务

6.Redis 的访问

7.Redis 的管道(pipeline)

8.Redis 的持久化(AOF+RDB)

9.Redis 的主从复制

10.Redis 的调优

11.Redis 的sentinel

六.mongoDB（分布式文件存储的数据库系统）

MongoDB 是一个介于关系数据库和非关系数据库之间的产品，是非关系数据库当中功能最丰富，最像关系数据库的。他支持的数据结构非常松散，是类似json的bson格式，因此可以存储比较复杂的数据类型。Mongo最大的特点是他支持的查询语言非常强大，其语法有点类似于面向对象的查询语言，几乎可以实现类似关系数据库单表查询的绝大部分功能，而且还支持对数据建立索引

1.mongoDB 简介

2.mongoDB 与Mysql 对比

3.mongoDB 的相关名词

4.mongoDB Shell

5.mongoDB 自定义Shell脚本

6.mongoDB 安全机制

7.mongoDB 设计应用

8.mongoDB 与MapReduce 集成

9.mongoDB 高级操作

七.Zookeeper

ZooKeeper是一个分布式的，开放源码的分布式应用程序协调服务，是Google的Chubby一个开源的实现，是Hadoop和Hbase的重要组件。它是一个为分布式应用提供一致性服务的软件，提供的功能包括：配置维护、域名服务、分布式同步、组服务等。

1.分布式协调技术起源概述

2.分布式协调技术架构原理分析

3.Zookeeper java客户端使用演示

4.Zookeeper环境搭建注意事项

5.Zookeeper常见命令使用和API精讲

6.Zookeeper通信协议介绍

7.Zookeeper请求处理过程分析

8.Zookeeper数据存储和选举机制分析

9.Zookeeper配置管理实战/监测连接数

实战项目：演示Zookeeper分布锁的使用，在互联网电商企业中，分布式开发如何和Zookeeper结合进行分布式缓存、锁的处理。

八.Sqoop(数据迁移工具)

sqoop 主要用于在Hadoop(Hive)与传统的数据库(mysql、postgresql...)间进行数据的传递，可

以将一个关系型数据库中的数据导进到Hadoop的HDFS中，也可以将HDFS的数据导进到关系型数据库中。

1.Sqoop框架介绍

2.Sqoop框架原理分析

3.Sqoop框架安装步骤演示

4.Sqoop1和Sqoop2分析对比

5.Sqoop深入了解数据库导入原理

6.Sqoop导出数据原理分析

7.Sqoop 设置存储格式与使用压缩

8.Sqoop导入数据到hdfs分析实战

9.Sqoop 增量导入功能代码实现

10.Sqoop RDBMS与Hive的操作演示

备注：sqoop主要是源码分析和API使用，考虑到中小型公司使用频次，将会在项目中演示用法。

九.Flume(分布式日志收集系统)

Flume是Cloudera提供的一个高可用的，高可靠的，分布式的海量日志采集、聚合和传输的系统，Flume支持在日志系统中定制各类数据发送方，用于收集数据；同时，Flume提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力。

1.Flume框架原理和应用场景分析

2.Flume框架使用场景分析

3.Flume概述以及原理解析

4.Flume中Event的概念和Socket的关联

5.Flume运行机制分析

6.NetCat Source源码分析

7.Flume agent原理说明和shell配置

实战项目：Flume和HDFS的结合使用，在HDFS中采集数据，收集Hive运行的目录到hdfs文件系统。

十.Oozie(工作流程调度管理系统)

Oozie是Yahoo针对Apache Hadoop开发的一个开源工作流引擎。用于管理和协调运行在Hadoop平台上（包括：HDFS、Pig和MapReduce）的Jobs。Oozie是专为雅虎的全球大规模复杂工作流程和数据管道而设计

1.Oozie 综合概述

2.Oozie 架构简析

3.Oozie 搭建部署

4.Oozie 管理界面的使用

5.Oozie Helloworld

6.Oozie Cli的使用

7.Oozie Job配置

8.Oozie 流程处理文件

9.Oozie hDPL语言定义节点

十一.Scala 课程

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

1.scala 环境配置

2.scala 体系结构

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 的特质

12.scala 的操作符

13.scala异常处理

十二.Kafka(流处理平台)

Kafka 是在大数据流处理场景中经常使用的分布式消息系统，配合 Spark 内存计算框架， 是流处理场景中的黄金组合。本课程以实战的方式学习 Kafka 分布式消息系统，包括 Kafka 的安装配置、Producer API 的使用、Consumer API 的使用以及与第三方框架(Flume、 Spark Streaming)的集成开发。每个知识点的学习，都有编程实战和操作实战，用眼见为 实的方式学习抽象的理论概念

1.Kafka 入门

2.Kafka 集群搭建理论与实践

3.Kafka Topic 实战

4.Kafka 开发 Producer 理论与实践

5.Kafka 开发 consumer 理论与实践

6.Kafka 发送和接收结构化数据

7.Kafka 发送和接收非结构化数据

8.Kafka 整合 Flume 框架

9.spark 读取 kafka 数据

十三.Spark Core

Spark 内存计算框架，是当前最流行的大数据计算框架，Spark 已经成为大数据开发人员以 及数据科学家的必备工具。

本课程主要学习 Spark Core 的内容。包括 Spark 集群安装、Spark 开发环境搭建，Spark Core 编程模型、Spark 程序运行原理、Spark 性能调优等

1.Spark 的起源及其哲学思想

2.Spark 集群的安装、启动、测试

3.Spark 基本架构及 API 介绍

4.Spark 开发环境搭建并开发运行 wordCount 程序(Scala、 Java)

5.wordCount 程序的集群部署及 Spark UI 简介

6.Spark 计算框架的核心抽象--RDD(理论及入门)

7.Spark RDD创建实战(Scala、 Java)

8.Spark RDD 操作--transformation 算子实战(Scala、 Java)

9.Spark RDD 操作--action 算子实战(Scala、 Java)

10.Spark RDD计算结果保存实战(Scala、 Java)

11.Spark RDD 缓存及持久化实战(Scala、 Java)

12.Spark 分布式共享变量实战--累加器和广播变量(Scala、 Java)

13.Spark 程序集群部署方式实战

14.Spark 程序运行流程分析

15.Spark 程序的监控和调试

16.Spark 内核解读

17.Spark 性能调优(shuffle)

18.Spark Core 数据分析实战

十四.Spark SQL

本课程将深入浅出学习 Spark 的结构化 API(DataFrame、Dataset 和 SQL)。SparkSQL 是在大数据项目中，Spark 开发工程师经常使用的 Spark 模块，除了深入讲解 SparkSQL 本身的每个知识点、SparkSQL 性能调优，还会涉及到 HDFS、Hive、HBase、MongoDB、 Oracle、MySQL 等第三方数据存储框架。每个知识点都以代码实战的方式讲解，知其然，更知其所以然。

1.Spark SQL 背景介绍(1 小时)

2.SparkSQL、 DataFrame、 Dataset 之间的关系

3.SparkSQL 概述

4.SparkSQL 数据类型

5.SparkSQL join 操作实战

6.SparkSQL 读写数据实战

7.SparkSQL 操作 Hive 中的数据(

8.SparkSQL 调优

9.SparkSQL 数据分析案例实战

十五.Spark Streaming(流处理平台)

Spark streaming是Spark核心API的一个扩展，它对实时流式数据的处理具有可扩展性、高吞吐量、可容错性等特点。我们可以从kafka、flume、Twitter、 ZeroMQ、Kinesis等源获取数据，也可以通过由高阶函数map、reduce、join、window等组成的复杂算法计算出数据。最后，处理后的数据可以推送到文件系统、数据库、实时仪表盘中。事实上，你可以将处理后的数据应用到Spark的机器学习算法、 图处理算法中去

1.Spark Streaming 框架机制

2.Spark Streaming 时间和窗口的概念

3.Spark Streaming DStream和RDD的关系

4.Spark Streaming 性能调优

实战项目:读取kafka数据做聚合处理，条件过滤后写入HDFS

十六.Storm（分布式实时数据计算系统）

Storm是一个开源的分布式实时计算系统，可以简单、可靠的处理大量的数据流。

而且支持水平扩展，具有高容错性，保证每个消息都会得到处理。

Storm处理速度很快（在一个小集群中，每个结点每秒可以处理数以百万计的消息）。

Storm的部署和运维都很便捷，更为重要的是可以使用任意编程语言来开发应用。

1.Storm 简介

2.Storm 原理和概念

3.Storm 与 Hadoop 的对比

4.Storm 环境搭建

5.Storm API 入门

6.Storm Spout

7.Storm Grouping策略及并发度

8.Storm 优化引入zoolkeeper锁控制线程

9.Storm 去重模式

10.Storm shell脚本开发

11.Storm 批处理事务

12.Storm 普通事务分区事务

13.Storm 按天计算

14.Storm 不透明分区事务

15.Storm 事务

16.Storm Trident

十七.数据可视化

数据可视化部分课程内容大部分是前端的内容，本章节所涉及部分基本都是一些JS框架，用于做数据展示。每个框架涉及到很多知识点，这里不再细化。

1.Tableau

2.Echarts

3.D3.js

4.Vue.js

5.Datav

商业项目实战

1.企业电商用户Session日志检测系统

项目来源： 一线电商线上系统

方向： 电商大数据

项目介绍：

在电商项目中，为了收集更多的数据需要通过客户端、PC网页进行相关埋点统计，为了支撑运营团队进一步的进行产品运营策略，需要对公司产品销售数据、网站uv/pv进行数据分析。本项目从电商企业实战出发，进行总结和提升。包含3块子项目，销售数据分析统计，Storm架构代码实战、Cloudera Manager实战。

项目实操：

第1-2天：Cloudera Manager环境部署、Storm部署、项目概述、Hadoop、HBase、Zookeeper环境搭建。

第3天：Kafka实战、API编写，Storm和Kafka集成测试

第4-5天：HBase实战和ECharts使用

第6-7天： 电商销售数据展示、HBase数据存储，job restart等

第8天： JStome介绍、项目总结

技术热点：

1.Cloudera Manager实战，多服务器管理

2.Storm实战讲解

3.Kafka-Storm实战

4.前端技术学习 ECharts等

5.HBase项目使用

6.Zookeeper使用

2.城市交通车辆分析系统

项目来源： 公安大数据项目

方向： 交通大数据

项目介绍：

基于大数据的城市交通车辆分析项目，采用Spark+Hive+HDFS+Mysql大数据架构，对城市卡口数据进行分析，主要分析模型包括：套牌车分析、同行车辆分析、首次入城车辆分析等。本项目中涉及到的大数据应用业务场景：离线数据分析、实时数据分析（流处理）、大数据可视化展示。

项目实操：

第1天：大数据项目业务场景介绍，车辆分析项目环境（Spark+Hive+HDFS+Mysql）搭建，分析项目中的数据流向。

第2天：套牌车分析模型的业务规则介绍，套牌车分析---编程实战。

第3天：同行车辆分析模型的业务规则介绍，同行车辆分析-编程实战。

第4天：首次入城车辆分析的业务规则介绍，首次入城车辆分析–编程实战，实现实时监控首次入城车辆的功能。

第5天：项目分析结果的可视化展示

第6天：项目总结及项目优化

技术热点：

涉及到的大数据相关技术及组件：SparkSQL、Spark Streaming、Hive、HDFS

3.百度ProtoBuffer格式日志结构化打点处理系统

项目来源： 百度大数据项目

方向： 日志大数据

项目介绍：

数据处理的痛点就是日志打点格式复杂，一切特殊化处理使下游数据处理系统处理逻辑繁杂冗乱。打点业务方的每一次格式变动都有可能造成数据处理程序漏掉相关的日志，日志打点规范化是数据团队的保证数据质量的痛点，规范化的日志结构，不仅利于各团队数据有效沟通，还使数据处理程序更加健壮。

为此设计一套结构化打点方案及相应的数据处理pipeline是对海量数据获取和分析的完美解决方案。

本项目针对百度内部使用的方案进行详细解析和复现。

功能描述：

项目概述，难点描述

profobuffer格式系统化介绍

业务方pb格式打点数据模拟

构建通用pb格式数据处理平台

基于spark和hadoop streaming实现平台代码细节

技术热点：

spark rdd，Hadoop steaming，protobuffer，linux，hdfs，mapreduce