《企业级 Hive 实战课程》

Hive 市场需求

Hive 课程大纲

第一阶段【分布式数据仓库 Hive】

Hive 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 SQL 查询功能，Hive 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 语句快速实现统计查询，也支持实现自己的 UDF 函数来完成比较复杂的业务逻辑，非常适合数据仓库的统计分析。

➢ Hive 概述、环境搭建及入门

- Hive 是什么、Hive 与 Hadoop 的关系、Hive 体系架构
- Hive 与 RDBMS 的区别、Hive 实用场景以及优缺点
- Hive 环境搭建
- Hive 元数据存储、Hive 数据存储
- Hive shell 常用操作
  - hive -e
  - hive -f
  - hive -v
  - hive -i
  - hive -S

➢ Hive 常见表操作

- Hive 内部表、外部表、分区表（静态分区、动态分区）常用操作
  - 表创建
  - 数据加载
  - 数据导出
  - 内/外部表的区别以及各自在生产中的适用场景
- Hive 常用查询操作
  - SELECT
  - WHERE
  - DISTINCT
  - JOIN
  - GROUP BY
  - UNION
  - CASE WHEN THEN
  - IN/NOT IN/EXISTS/NOT EXISTS

- Hive 排序
  - ORDER BY
  - SORT BY
  - DISTRIBUTE BY
  - CLUSTER BY
- Hive 复合数据类型
  - ARRAY
  - MAP
  - STRUCT
- Hive 索引
- Hive 编程
  - HiveServer2/beeline 使用
  - Java 操作 Hive
  - Hive 内置函数以及 UDF 编程
- Hive 窗口和分析函数
  - SUM/AVG/MIN/MAX
  - NTILE/ROW_NUMBER/RANK/DENSE_RANK
  - CUME_DIST/PERCENT_RANK
  - LAG/LEAD/FIRST_VALUE/LAST_VALUE
  - GROUPING SETS,GROUPING_ID,CUBE,ROLLUP
- Hive 虚拟列以及在项目中的使用
  - INPUT__FILE__NAME
  - BLOCK__OFFSET__INSIDE__FILE

➢ Hive 常用存储格式与压缩格式

- 行式存储 VS 列式存储
- 存储格式：TextFile/ SequenceFile /RCFile/ORCFile/Parquet
- 压缩格式：gzip/bzip/snappy 等
- 如何在项目中选择合适的存储格式以及压缩格式

第二阶段【Hive 实战开发】

依据企业中常见的【日志文件】分析，使用 Hive 进行数据处理，把握如何设计表，运行 HiveQL 语句时，出现数据倾斜等问题时调优，以及如何修复 Hive Bug 和提交。

- **Hive 常用优化策略**

  - 并行执行
  - JVM 重用
  - 合理设置 Mapper/Reducer 个数
  - 合理利用压缩技术以及分布式缓存
  - 充分利用多个 job 之间的共用的中间结果集
  - 执行计划深入剖析
  - 深入剖析常用的几种 Join：Reduce Join/Map Join/SMB Join 工作原理以及各自的使用场景
  - PPD：Predicate Pushdown
  - 数据倾斜分析及常用解决方案
  - 分区的合理使用

- **Hive 实战**

本部分包含两个 Hive 的实战案例，涉及到 Hive 相关的绝大部分知识点，**由于项目业务数据的保密性，暂不对外公布案例描述。**

  - 项目实战一
  - 项目实战二

- **Hive 高级**

本部分将重点介绍**在工作中遇到的各种真实的 Hive 相关的问题，如何分析、定位以及解决/修复这些问题，包括分析执行计划、源码 bug 修复等。**

  - **如何分析、跟踪、解决/修复在生产环境中遇到的问题**
  - **修复问题后 Hive 源码编译等相关环节**

- Hive 的元数据表结构详解
- Hive 执行流程源码分析
- Hive 如何 Debug 执行

第三阶段【Spark SQL】

2014 年 4 月 Spark 发布了 1.0 版本，该版本中包含了 SparkSQL 模块，它是 Spark 的核心组件之一。Spark SQL 是一个用于处理结构化数据的 Spark 组件，Spark SQL 作为 Shark 的继任者，其主要功能之一就是方便用户访问和操作已经存在的 hive 表数据。由于 Spark 是基于内存的计算框架，使用 Spark SQL 之后可以将原有的 hive 脚本直接跑在 Spark 之上，大大提高运行效率。

➢ Shark

本部分将讨论 Shark，它作为 Spark 设计并开源的一款数据仓库系统，提供了分布式 SQL 查询引擎，并能够兼容 Hive；通过将 HQL 转换成 Spark 作业并提交到 Spark 集群上运行。

- Shark 产生原因
- Shark 体系架构
- Shark VS Hive
- Shark 项目终止以及后续发展的方向

➢ Spark SQL

SparkSQL 的愿景：写更少的代码，读更少的数据，将优化的操作交由底层优化器去执行；DataFrame 提供了丰富多样的外部数据源支持。

- Spark SQL 愿景
- Spark SQL 概述

- Spark SQL 体系架构
- Spark 环境搭建
- SqlContext 与 HiveContext
- spark-sql、thriftserver/beeline
- Spark SQL 常用操作

➢ Spark SQL 案例实战

结合具体案例，主要针对【日志类型 】数据进行分析统计讲解，进一步体会 Spark SQL 使用的优势

- 将 Hive 部分的实战案例运行在 Spark SQL 之上
- 日志分析统计案例
- SparkSQL 综合案例，暂不对外提供项目描述。

第四阶段【Hive on Spark】

Hive 是基于 Hadoop 平台的数据仓库，已经成为 Hadoop 事实上的 SQL 引擎标准；Hive 拥有更为广泛的用户基础以及对 SQL 语法更全面的支持；Hive 最初的计算引擎为 MapReduce，自身性能的提升受限于 MapReduce 的计算框架；Hive on Spark 目的是将 Spark 作为 Hive 的另外的计算引擎，把 Hive 的查询作为 Spark 的任务提交到 Spark 集群上进行计算，利用 Spark 的特性提高 Hive 查询性能。

➢ 背景及环境搭建

- 产生背景
- 设计原理及运行架构
- 编译及部署

➢ 实战操作

- Hive&Spark SQL 对比

- 将 Hive 部分的实战案例运行在 Hive on Spark 之上

第五阶段【Hive 方向就业指导】

在实际面试中，如何编写 Hive 方面的项目简历、与面试官描述企业大数据平台中的数据仓库的设计与使用开发的问题解决。

➢ 大数据 Hadoop 工程师简历编写

- 企业项目描述
- 大数据平台的架构（主要针对数据仓库）
- 项目任务（工作职责）

➢ 如何把握 Hive 使用

- 常见的 Hive 问题解决
  - Hive 架构
  - 数据倾斜、数据压缩、文件格式
  - 项目中表的常见设计（分区表、桶表、外部表）
  - 内存溢出的分析
- Hive 源码修改
  - 实际发行源码 Bug，如何修改
  - SQL on Hadoop 中的各个框架的比较

课程试听

试听内容大纲如下：

试听视频讲义软件下载地址如下：

链接： http://pan.baidu.com/s/1dD2L5LZ

密码：dthm

课程学费

本期课程价格为：1200 元

备注说明：

1）如果报名学员为云帆内部学员（已报名云帆大数据 Hadoop 和 Spark 中的任何一门课程）一律半价，价格为 600 元。

2）其他学员，三人组团报名，优惠价格为每人 1000 元。

授课时间

开班时间：

自报名日起招满 20 人即开课

学习周期：

自开课之日起，**约 5 周**

付款方式

统一使用支付宝转账方式：

账号：yunfanwl@sina.com

姓名：邹建明

温馨提示：

付款前，请与咨询顾问洽谈，确认付款账户正确性；付款后，请把付款结果截图给咨询顾问，确保对方一定收到。

云帆大数据（www.cloudyhadoop.com）

2015 年 7 月 20 日 星期一