

1.数据仓库概念

- 数据仓库,中文简称数仓,英文叫做 Data WareHouse ,简称DW
- 数据仓库是面向分析的集成化数据平台,分析的结果给企业提供决策支持
- 数据仓库本身不生产数据; 数据仓库本身也不消费数据;
- 数据仓库--数据源: 来自企业的业务数据
 - 业务数据库
 - 爬虫爬取的数据
 - 日志数据(点击流)
 - 平面文件(文本文件,excel文件)
- 数据处理--进行分层:
 - 第三层 ODS层(operational data store)或者贴源层、缓冲层、备份层
 - 临时数据存储中转,和数据源进行解耦,数据之间往往彼此差异较大,一般不用于直接分析
 - 第二层 DW层(data waerhouse)数据仓库层
 - 其数据来自ODS经过层层ETL数据往往是干净规则的,基于主题聚集的,甚至还有一些模型数据
 - 第一层 App层(应用层)
 - 最终消费数据的应用 DV DM ad hoc
 - 第零层 报表层,看板层,可视化层,展示层
 - 分层好处:
 - 解耦合
 - 空间换时间,提高最终应用层使用数据的效率
- 数据仓库的作用:
 - 辅助业务决策,业务流程优化
- 企业中一般先有数据库,然后有数据仓库,可以没有数据仓库,但是不能没有数据库

2.数据仓库的主要特征

- 数据仓库是面向主题、集成性、非易失性和时变性的数据集合,用以支持管理决策
 - 面向主题: -- 抽象的概念,数据的综合体,分析主题和业务需求相关
 - 主题是一个抽象的概念,是较高层次上企业信息系统中的数据综合、归类并进行分析利用的抽象
 - 集成性: -- 把数据从外部数据源经过ETL填充到数仓主题下面
 - 主题相关的数据通常会分布在多个操作型系统中,彼此分散、独立、异构.需要集成到数仓主题下
 - 非易失性
 - 数据仓库是分析数据的平台,而不是创造数据的平台
 - 时变性 -- 数仓的数据在时间维度成批次更新变化
 - 数据仓库的数据需要随着时间更新,以适应决策的需要

3.数据仓库与数据库的区别

- 联机事务处理 OLTP(On-Line Transaction Processing) == 数据库
 - 联机事务处理系统,面向业务面向事务,支持事务
 - 焦点: 当下 ;任务: 读写操作; 响应时间 毫秒 ; 数据量: 小数据,MB ,GB
 - 目的: 面向应用,面向业务,支撑事务
 - 基于事物的业务逻辑的处理
 - 事务操作: 增删改操作, 非事务操作: 查询操作
 - 快速捕获业务数据,存储,事务机制 , 不能重复,必须结构化
- 联机分析处理 OLAP(On-Line-Analytical processing)==数据仓库
 - 联机分析处理系统,面相分析支持分析
 - 焦点: 主要面向过去,面向历史,时时数仓除外 ; 任务: 大量读而很少写操作 ; 响应时间: 秒、分钟、小时或者天 ; 数据量: 大数据,TP,PB
 - 面向主体,面向分析,支撑分析决策
 - 基于分析的分析历史数据,对历史数据进行分析
 - 离线的、批的、实时的,不需要考虑结构化,可以重复

4.Apache Hive

4.1 Hive的概念

- Hive是Facebook开源出来的,后来贡献给Apache宗旨: 提高分析数据的能力降低数据的开发成本
- Hive是基于Hadoop的一个数据仓库工具,用于分析数据的
 - 具备存储数据的能力
 - 具备分析数据的能力
 - Hive使用Hadoop HDFS作为存储系统
 - Hive使用Hadoop MapReduce来分析数据
- Hive的作用: 可以将结构化的数据文件映射为一张数据表,并提供类SQL查询功能,sql转换成mr来分析
 - 结构化数据: 具有schema约束的数据,便于程序解读解析
 - 映射表示的就是一种对应关系
 - 映射成为表之后,提供了类SQL查询分析功能
 - sql叫做声明式编程,程序员不用关系,利于数据分析

Hive 数据仓库工具, 将HDFS上的结构化的文件映射到数据库表中的工具;
Hive 翻译工具, 将 SQL 翻译成 MR / Spark翻译软件。

Hive 是如何翻译的?

user.txt 文件是文本文件, 每行之间用 /r/n 回车每列之间用, 逗号 分割。

需求: 如何计算年龄大于 25 岁的平均毕业薪资?

方案:

1.使用MR直接读取数据, mapper然后reducer写出到 文件系统。

2.使用 Hive 来解决, 将文本文件加载到 Hive 的表中, 然后通过类SQL实现查询 HiveQL?

Hive 如何实现读取 HDFS 然后加载到表中

2.1 创建数据表

2.2 将每行数据进行切分 terminated by ','

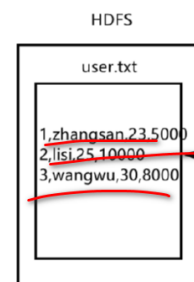
2.3 将每个字段和数据字段映射到一起

数据库名
数据表名
数据表字段
字段类型
数据的路径
创建的时间
创建的权限
数据的大小
数据表的分区

元数据

SQL:
select name,avg(salary) as avgSalary
from t_user
where age > 25
group by name

hdfs dfs -put localfile /destfile



映射



元数据 metadata =>
存储到 metastore
mysql

t_user
id 姓名 年龄 收入
1 zhangsan 23 5000
2 lisi

select * from t_user

hive 就是翻译软件, 将SQL转换成 mapreduce程序

1,zhangsan,23,5000
2,lisi,25,10000

4.2 Hive架构组件

Hive数据仓库的架构

1.用户通过 JDBC / ODBC 提交SQL 到 Hive Thrift server

2. Hive Driver 驱动程序

2.1 解析器, 将SQL语法转换成抽象语法树

2.2 编译器, 将抽象语法树转换成逻辑执行计划

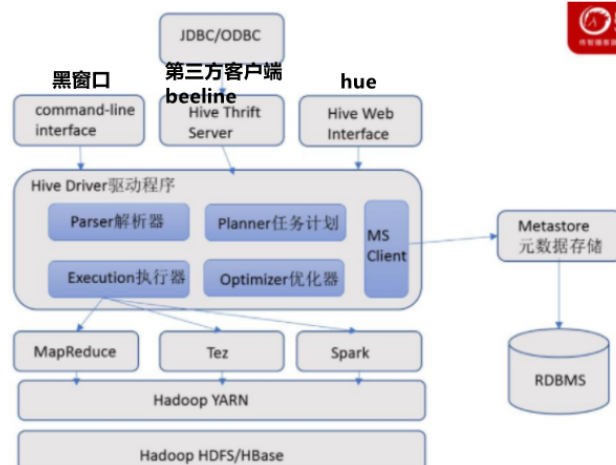
2.3 优化器, 将逻辑执行计划转换成优化的逻辑执行计划

2.4 执行器, 将优化的逻辑执行计划翻译成物理执行计划

通过是 MR的模板生成执行代码

3. 将生成的 MR 的代码提交给 YARN 集群执行

4. MR 程序会读取 HDFS 上的数据进行任务的执行



- 客户端用户结接口
 - 所谓的客户端指的是给用户一种方式编写Hive SQL
 - 目前常见的客户端: CLI(命令行接口 shell)、web UI、JDBC|ODBC
- Hive Driver驱动程序
 - hive的核心, 完成从接受HQL到编译成为MR程序的过程

- sql解释 编译 校验 优化 制定计划
- meatdta
 - 元数据存储.描述性数据
 - 对于hive来说,元数据指的是表和文件之间的映射关系
- Hadoop
 - HDFS 存储文件
 - MapReduce计算数据
 - Yarn 程序运行的资源分配
- Hive 不是分布是软件,只需要在一台机器上部署Hive服务即可; Hive的分部处理能力是借于Hadoop完成的.HDFS分布式存储,MapReduce分布式计算

4.3Hive和MySQL的区别

- 从外表、形式模型、语法各层面上看,hive和数据库(MySQL)很类似
- 底层应用场景是完全不一样的
- hive属于olap系统,是面向分析的侧重于数据分析(select)
- 数据库属于oltp系统,是面向事务的,侧重于数据时间交互(crud)
- Hive绝不是大型数据库,也不是为了要取代MySQL这样的数据库

5.Hive的安装部署模式

- 三种模式
 - 内嵌模式--metastore存储在Derby,不需要配置启动
 - 1.元数据存储在内置的derby
 - 2.不需要单独配置metastore也不需要单独启动metastore服务
 - 适合测试体验,实际生产中没人用,适合单机单人使用
 - 本地模式-- metastore存储在MySQL,不需要配置启动
 - 元数据使用外置的RDBMS,常见使用最多的是MySQL
 - 不需要单独配置metastore,也不需要单独启动metastore服务
 - 远程模式--metastore存储在MySQL,需要单独配置、单独启动
 - 1.元数据使用外置的RDBMS,常见使用最多的是MySQL
 - 2.metastore服务单独配置,单独手动启动,全部唯一
 - 这样的话各个客户端只能通过这一个metastore服务访问Hive,企业生产环境中使用的模式,支持多客户端远程并发操作访问Hive