第四章数据仓库

2021年11月14日　　14:13

1. 数据仓库
    a. 定义
        - 一个面向主题的、集成的、时变的、非易失的数据集合，支持管理者的决策过程
    b. 与数据库的关系
        - 数据库的局限性
            - 传统数据库只是对已有的数据进行存取以及简单的查询统计
            - 无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势
            - 导致了“数据爆炸但知识匮乏”的现状
        - 为什么需要数据仓库
            - 提高两个系统的性能
                - 数据库是为OLTP（联机事务处理）设计的
                - 数据仓库是为OLAP（联机分析处理）设计的
            - 不同的功能和不同的数据
                - 历史数据：决策支持需要历史数据，而普通操作数据库中不会去维护
                - 数据汇总：决策支持需要讲来自异种源的数据统一
                - 数据质量：不同源数据表示、编码和格式不一致，需要有效分析后转化集成
        - 数据库用于事务处理
        - 数据仓库用于决策分析

	数据库	数据仓库
主要任务	事务处理	决策分析
内容	与事务相关的数据	与决策相关的数据
数据	当前数据	时变的数据
访问	经常是随机的读写操作	经常是只读操作
负载	事务处理量大、但每个事务涉及的记录数较少	查询量少，但每次要查询大量的记录
事务输出量	一般很少	可能非常大
停机时间	可能意味着灾难性错误	可能意味着延迟决策

2. 联机分析处理
    a. 联机事务处理（OLTP）
        - 是在网络环境下的事务处理工作，以快速的响应和频繁的数据修改为特征，使用户利用数据库能够快速地处理具体的业务
        - 要求多个查询并行，以便将每个查询分布到一个处理器上
        - 特点在于事务处理量大，但事务处理内容比较简单且重复率高

- 处理的数据是高度结构化的，涉及的事务比较简单，数据访问路径是已知的，至少是固定的
- 面对的是事务处理操作人员和低层管理人员

b. 联机分析处理（OLAP）
- 定义：是数据仓库上的分析展示工具，它建立在数据多维视图的基础上
- 主要特点：多维分析(Multi_Analysis)，这是OLAP技术的核心所在
- 特点：
  - 决策分析需要对多个关系数据库共同进行大量的综合计算才能得到结果
  - 关系数据库是二维数据（平面），多维数据库是空间立体数据
  - 基本思想是决策者从多方面和多角度以多维的形式来观察企业的状态和了解企业的变化

c. OLTP与OLAP的对比

	OLTP	**OLAP**
用户	办事员、数据库人员	知识工作（主管、分析人员）者
功能	日常操作	决策支持
数据库设计	面向应用的，事务驱动	面向主题的
数据特点	当前的, 更新的 详细的,关系型的 孤立的	历史的, 汇总的, 多维的 集成的, consolidated
特性	操作处理	信息处理
存取方式	读/写 索引	大量的扫描
工作单元	简单的事务处理	复杂的查询
记录 访问量	数十	数百万
用户数量	数以千计	数以百计
数据库规模	100MB-GB	100GB-TB

d. OLAP主要操作
- 切片和切块（slice and dice）
- 转轴（pivot）
- 上卷（roll-up）：汇总数据
- 下钻（drill-down）：上卷的逆操作

3. 数据仓库的体系结构

a. 体系结构

i. 底层：数据仓库的数据库服务器
- 关注的数据：如何从这一层提取数据来构建数据仓库

ii. 中间层：OLAP服务器
- OLAP服务器如何实施

iii. 前端客户工具层：
- 关注的问题：查询工具、报表工具、分析工具、挖掘工具等

b. 三种应用

i. 信息处理

1）查询和基本的统计分析，使用交叉表、表、图标和图进行报表处理

ii. 分析处理
    1) 多维数据分析
    2) 基本的OLAP操作
iii. 数据挖掘
    1) 从隐藏模式中发现知识
    2) 支持关联分析，构建分析性模型，分类和预测，并用可视化工具呈现挖掘的结果

c. OLAP vs 数据挖掘
    i. 观点一：两个功能不相交，OLAP是数据汇总、聚集工具，帮助简化数据分析；数据挖掘自动地发现隐藏在大量数据中的隐含模式和有趣知识，比传统的OLAP处理前进了一步
    ii. 更广泛的观点：数据挖掘包含数据描述和数据建模，而OLAP可以提供对数据仓库中数据的一般描述，基本功能是用户指导的汇总和比较，这些都是数据挖掘功能。也就是说数据挖掘的涵盖面比简单的OLAP操作宽很多

4. 云原生数据仓库