1 Data Warehouse & OLAP

1.1 课本不包含内容

PPT:Lesson 4 - Data Warehouse OLAP

- P21-24: 数据立方体定义语言
- P52-53: 冰山立方体拓展 (感觉不太重要)
- P54: 数据立方体算法 (第五章内容，应该不在考试范围)
- P60-62:OLAM 简介

1.2 概念部分

书上的目录分类太详细了，遇到概念直接翻阅目录应该就能找到对应解释。

1.3 PPT 重点部分/计算部分

1.3.1 星形、雪花和事实星座 (P91-93)

1. 星形模式：由无冗余的事实表和多个维表组成，维表不会包含子维表，因此类似于以事实表为根，以维表为子节点的二层树结构。

优点: 浏览效率高，系统性能卓越。

缺点: 存在数据冗余，维护困难。

如同数据库一样，假设存在表 (name,province,city)，那么 (小明, 浙江, 杭州) 和 (小红, 浙江, 杭州) 两条数据存在冗余，但星形模式并不处理这类冗余，考虑 (name,loc_id),(loc_id,province,city) 的形式：(小明,1),(小红,1),(1, 浙江, 杭州) 的形式，尽管去除了冗余，但相比于整体的极大规模数据，有时剩下的内存微不足道，且查询时要通过连接查询，降低了系统吞吐量，因此星形模式还是非常流行的。
2. 雪花模式：由无冗余的事实表和多个维表组成，维表会包含子维表，类似于以事实表为根，以维表为子节点的多层树结构。

设计理念与星形相反，规范化维表，去除冗余，优缺点与星形倒置: 浏览效率低，性能较差，但易于维护，节省了部分空间。
3. 事实星座：通俗说就是从树变成图了，子节点直接可以相互连接。

1.3.2 度量分类 (P95-96)

用 L = [1,2,3,4,5] 和 Py 语言通俗描述
1. 分布的: 可以分段求解的聚集函数：

sum: sum(L) = sum(L[0:3])+sum(L[3:]), 因此 sum 是分布的

max: max(L) = max(max(L[0:3]),max(L[3:])), 因此 max 是分布的
2. 代数的: 可以用若干个分布的聚集函数通过代数运算得到

avg: avg = sum(L)/len(L),avg 是代数的，相反 avg 不能又前半段的平均值和后半段的平均值得到，因此不是分布的。
3. 整体的：不是分布和代数的就是整体的，就是说不能分段求解，也不能用分布组合得到，比如中位数:median([1,2,3]) = 2, median([4,5]) = 4.5, 而 L 的中位数是 3，分段求解的两个结果是 2 和 4.5 已经不存在正解本身了，因此不是分布的也不是代数的，是整体的。

1.3.3 OLAP 操作 (P96-97)

1. 上卷 (roll-up): 对数据归约，如 (杭州，100m^2)，(金华，200m^2) 对地区归约，得到 (浙江,300m^2)；特征：项个数从 2 个变成 1 个

2. 下钻 (drill-down): 与上卷相反，浙江分解成杭州和金华两项，还有一种形式：(杭州，100m^2)，(金华，200m^2) 下钻得到 (杭州，100m^2，乌龙)，(金华，200m^2，火腿)，即属性列的增加。因此特征有两种：项个数增加或每个项内维度增加。

3. 切片和切块：等价于 sql 的 where 语句，筛选功能。

4. 转轴：等价于 python.DateFrame 的 reindex，重新排列数据，除了重排以外，降维也算转轴操作，如将一个 3 层立方体转换成 3 个排列的 2D 图。

1.3.4 计算

$$\prod_{i=1}^{n}(L_i+1)$$

以 (时间 = 年份，位置 = 国家，类型) 为例，这是一个 3 维立方体，方体总数：对于每个维度只有取和不取，如时间 = {年 (取)，不取} 因此总数为 $2^n = 2^3 = 8$，当都不取的时候就是 0-D 模型，也就是 all。现在考虑时间可下钻成：年-> 季-> 月-> 日，位置可下钻成：国家-> 省份-> 城市，那么时间 = {年，季，月，日，不取}，位置 = {国家，省份，城市，不取}，类型 = {类型 (取)，不取}，所以立方体总数 = 5*4*2=40，所以 L_i 表示的就是特征可下钻的层数，+1 是附加上了不取该维，所有维度相乘得到总表达式 $\prod_{i=1}^{n}(L_i+1)$

1.4 习题

1.4.1 4.3

(a) 星形模型，雪花模型，事实星座模型
(b)

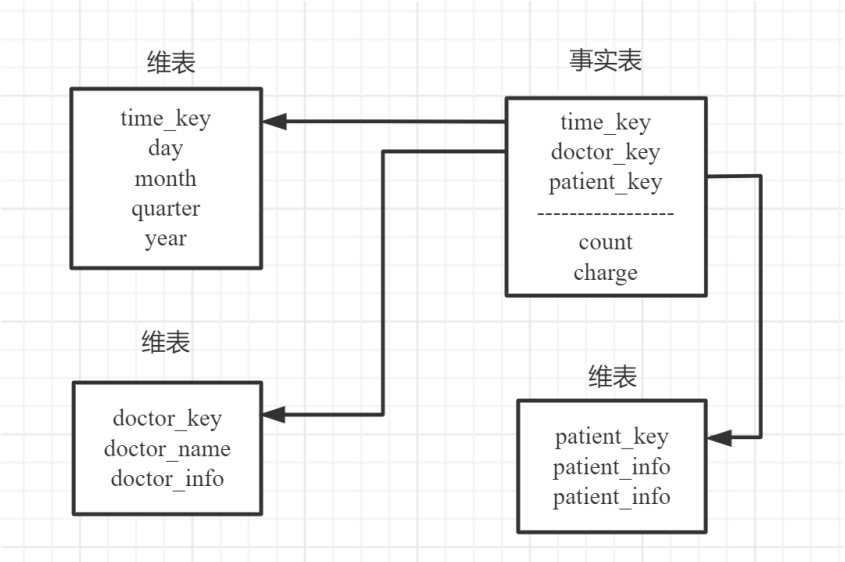

图 1: 习题 4.3

(c)

1. 上卷 day 到 year
2. 切片 year=2010
3. 上卷 patient 到 all

(d) select doctor, sum(charge) from fee where year = 2010 group by doctor;

1.4.2 4.4

(b)

1. 上卷 course 到系
2. 切片系 = CS
3. student 下钻到 student_name

(c) 5*5*5*5 = 625

1.4.3 4.5

(b)

1. 上卷 date 到 all，再下钻到 year

2. spectator 下钻到 spectator_type
3. 上卷 game 到 all
4. location 下钻到 location_name
5. 切片 year=2010 and spectator_type=" 学生" and location_name="GM"