设计篇
第六章 物理数据库设计

主讲：程思瑶
海量数据计算研究中心

物理数据库设计

- 设计任务
  - 在逻辑数据库设计基础上，为每个关系模式选择合适的存储结构和存取方法，使得数据库上的事务能够高效率的运行
- 设计步骤
  - 分析影响物理数据库设计的因素
  - 为关系模式选择存取方法
  - 设计关系、索引等数据库文件的物理存储结构

物理数据库设计步骤

- 物理数据库设计步骤
  - 分析影响物理数据库设计的因素
  - 为关系模式选择存取方法
  - 设计关系、索引等数据库文件的物理存储结构

影响物理数据库设计的因素

- 对于数据库**查询事务**，需得到如下信息：
  - (1) 查询的关系；
  - (2) 查询条件所涉及的属性；
  - (3) 连接条件所涉及的属性；
  - (4) 查询的投影属性。
- 对于数据**更新事务**，需得到如下信息：
  - (1) 被更新的关系；
  - (2) 每个关系上的更新操作的类型；
  - (3) 删除和修改操作条件所涉及的属性；
  - (4) 修改操作要改变的属性值。
- 了解每个事务在各关系上运行的**频率**
- 了解每个事务的**时间约束**

例如，关系R更新频率很高，则R上的索引等要尽可能少

上述信息是我们确定关系的存取方法的依据

物理数据库设计步骤

- 物理数据库设计步骤
  - 分析影响物理数据库设计的因素
  - 为关系模式选择存取方法
  - 设计关系、索引等数据库文件的物理存储结构

为关系模式选择存取方法

- 常用的存取方法可以分为三类：
  - 聚集方法
  - 索引方法
  - HASH方法

为关系模式选择存取方法

- 聚集方法
  - 把经常进行连接操作的多个关系的记录以连接属性为中心分类存储，从而提高连接操作的效率。
  - 即参加一个连接的所有关系中具有相同连接属性值的记录被物理地存储在一起。
  - 一个物理数据库可以有多个聚集存储
  - 但一个关系只能加入一个聚集存储

为关系模式选择存取方法

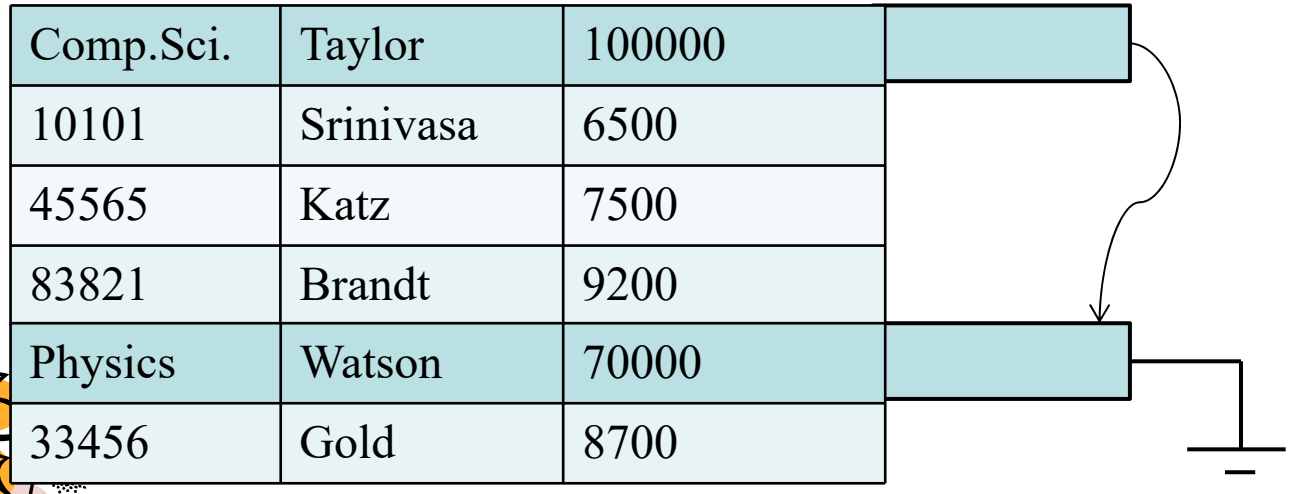
- 聚集存取方法，例：

dept_name	*building*	*budget*
Comp.Sci. Physics	Taylor Watson	100000 70000

department关系

ID	*name*	*dept_name*	*salary*
10101 33456 45565 83821	Srinivasa Gold Katz Brandt	Comp.Sci. Physics Comp.Sci. Comp.Sci.	6500 8700 7500 9200

instructor关系



带指针的多表聚簇文件结构

为关系模式选择存取方法

- 聚集存取方法的选择
  - 首先需要确定聚集关系组，即：
    - 确定需要多少个聚集存储
    - 每个聚集存储中包括哪些关系
  - 然后确定优化的聚集方案

为关系模式选择存取方法

- 聚集存取方法的选择
  - 确定聚集关系组
    - 经常在一起进行连接操作的关系可以作为聚集关系组，连接属性作为聚集键；
    - 如果一个关系的一组属性经常出现在相等比较条件中，则该单个关系可作为聚集关系组，这组属性作为聚集键；
    - 如果一个关系的一个(或一组)属性上的实例值重复率很高，则此单个关系可作为聚集关系组，这组属性作为聚集键。
    - 取消候选聚集关系组中不必要的关系，规则如下：
      - 从聚集组中删除经常进行全关系扫描的关系；
      - 从聚集组中删除更新操作远大于连接操作的关系

为关系模式选择存取方法

- 聚集存取方法的选择
  - 确定优化的聚集方案
    - 一个关系可能有多种聚集存储方式

$$cost(C) = \sum_{i=1}^{n} f_i \, cost(T_i)$$

为关系模式选择存取方法

- 索引存取方法的选择
  - 根据在R上事务T_1、T_2、…、T_k的信息确定候选索引，规则如下：
    - 如果一个(或一组)属性经常在查询操作条件中出现，则考虑在这个(或这组)属性上建立索引;
    - 如果一个属性经常作为最大值和最小值等聚集函数的参数，则考虑在这个属性上建立索引;
    - 如果一个(或一组)属性经常在连接操作的连接条件中出现，则考虑在这个(或这组)属性上建立索引;
    - 如果一个(或一组)属性经常作为投影属性使用，则考虑在这个(或这组)属性上建立索引;

为关系模式选择存取方法

- 索引存取方法的选择
  - 一个关系上可以建立多个索引
  - 优化配置索引
    - 不加索引？
    - 一个索引
    - 两个索引？
    - …

最小化Cost(R)

$$cost(R) = \sum_{i=1}^{n} f_i \, cost(T_i)$$

其中，(1) Cost(T_i)是事务T_i的代价
(2) f_i是T_i发生的频率

为关系模式选择存取方法

- HASH存取方法的选择
  - 有些数据库管理系统提供了HASH存取方法
  - 选择HASH存取方法的规则：
    - 如果一个关系的属性主要出现在相等连接操作条件中，或主要出现在相等比较选择条件中，而且满足下列两个条件之一，则此关系可以选择HASH存取方法：
      - (1) 如果一个关系的大小可预知，而且不变；
      - (2) 如果关系的大小动态改变，而且数据库管理系统提供了动态HASH存取方法。

物理数据库设计步骤

- 物理数据库设计步骤
  - 分析影响物理数据库设计的因素
  - 为关系模式选择存取方法
  - 设计关系、索引等数据库文件的物理存储结构

物理存储结构设计

- 确定如何在磁盘存储器上存储关系、索引和聚集，使得空间利用率最大化，数据操作引起的系统开销最小化

- 与具体数据库管理系统相关

小结

- 物理数据库设计的步骤
  - 关系存取方法
    - 索引
    - 聚集
    - HASH
  - 物理存储设计
- 本章重点
  - 掌握数据库物理存储结构与存取方法的设计

Now let's go to

Next Chapter