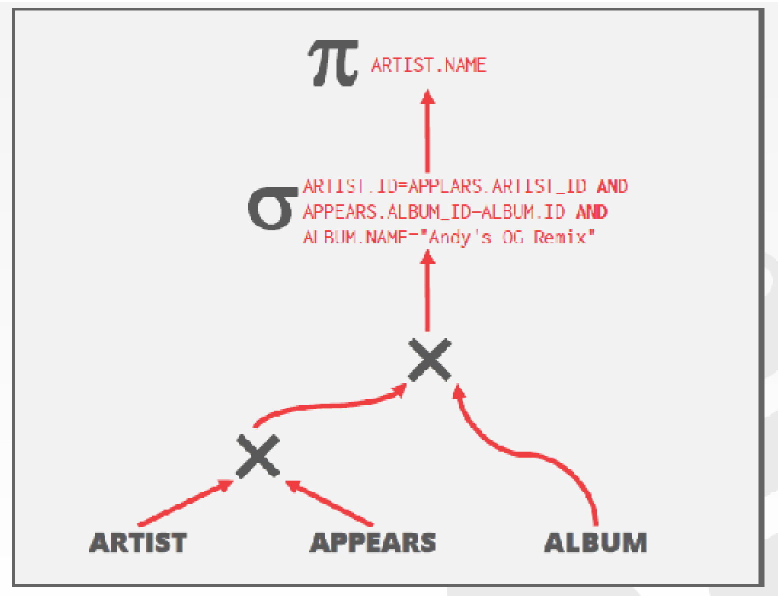
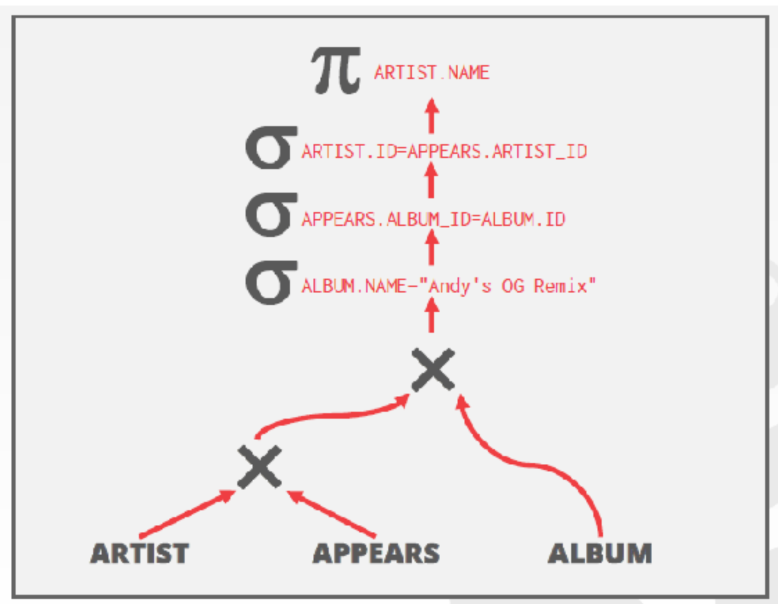
10 查询优化

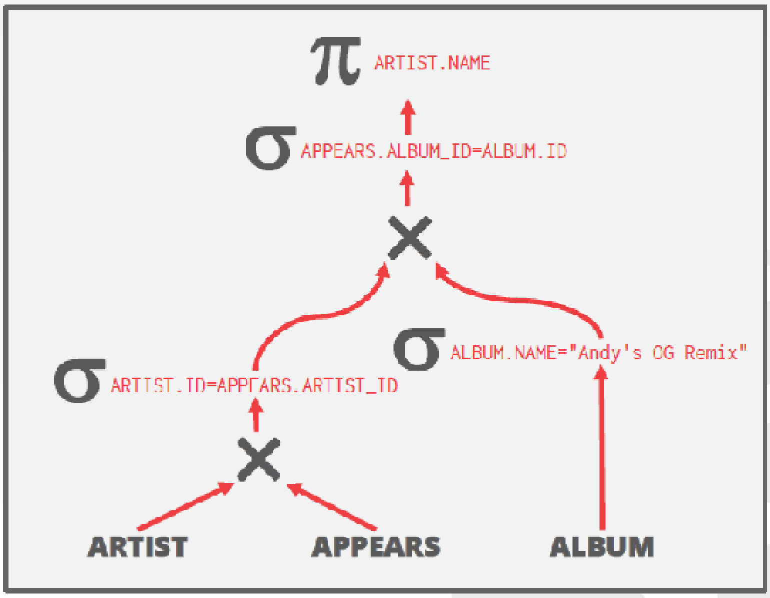
- 概述
  - 查询重写——关系代数等价
    - 执行计划：确切地定义了每个运算应使用的算法，以及运算之间的执行应该如何协调。
  - 启发式
    - 尽早执行select（减少元组数量）
    - 尽早执行projection
    - 在其他类似操作之前执行最严格的选择和连接操作（即具有最小的结果大小）。
    - 在左深连接树中，每个连接的右侧输入是一个关系，而不是中间连接的结果。
    - 基本步骤

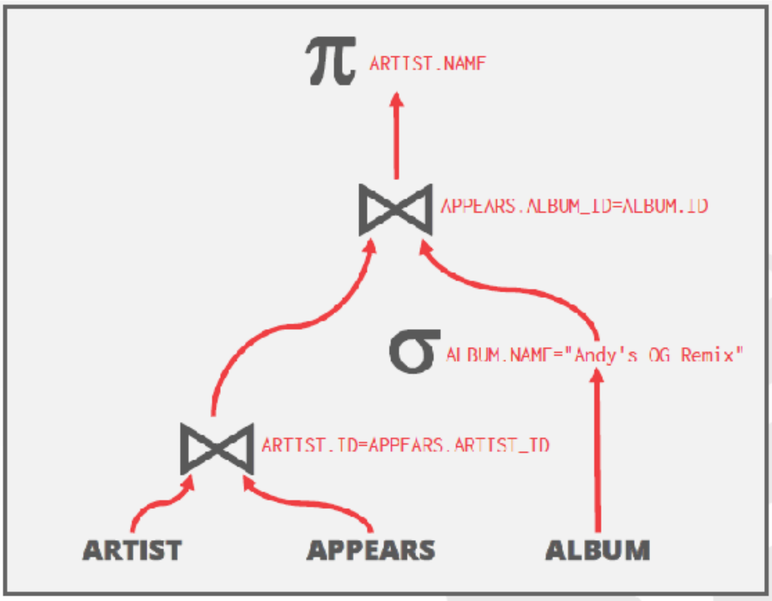
      - 拆分连接谓词：and分开

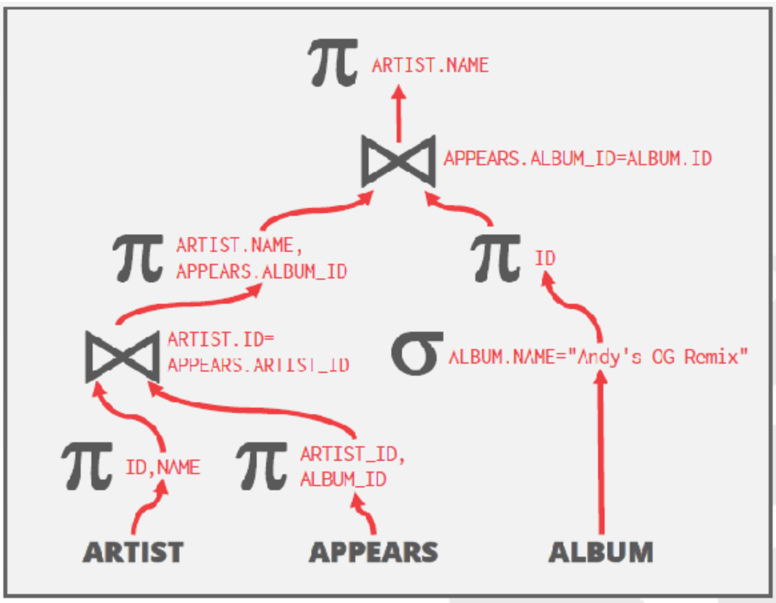
- 谓词下推：将谓词移动到最低适用点



- 用连接谓词替换笛卡尔积为内部连接



- 投影下推：中断之前消除冗余



- 基于成本的搜索
  - 查询执行计划
    - 产生逻辑上与给定表达式等价的表达式
    - 对所产生的表达式以不同方式注释，产生不同的查询计划
    - 估计每个执行计划的代价，选择估计代价最小的一个
- 关系表达式的转换
  - 如果两个关系表达式在每一个有效数据库实例（满足所有完整性约束）中都会产生相同的元组集，则我们称它们是等价的。
    - 顺序不重要
  - 等价规则：两种不同形式的表达式是等价的。
    - 合取选择运算可分解为单个选择运算的序列

      $$\sigma_{\theta_1 \wedge \theta_2}(E) \equiv \sigma_{\theta_1}(\sigma_{\theta_2}(E))$$

    - 选择运算满足交换律

      $$\sigma_{\theta_1}(\sigma_{\theta_2}(E)) \equiv \sigma_{\theta_2}(\sigma_{\theta_1}(E))$$

    - 一系列投影运算中只有最后一个运算是必须的

      $$\Pi_{L_1}(\Pi_{L_2}(\ldots(\Pi_{L_n}(E))\ldots)) \equiv \Pi_{L_1}(E)\text{，其中 } L_1 \subseteq L_2 \ldots \subseteq L_n$$

    - 选择操作可以与笛卡尔积以及θ连接相结合

      $$\sigma_{\theta}(E_1 \times E_2) \equiv E_1 \bowtie_{\theta} E_2$$

      $$\sigma_{\theta_1}(E_1 \bowtie_{\theta_2} E_2) \equiv E_1 \bowtie_{\theta_1 \wedge \theta_2} E_2$$

    - θ连接运算满足交换律

      $$E_1 \bowtie_{\theta} E_2 \equiv E_2 \bowtie_{\theta} E_1$$

    - 自然连接的结合率

      $$(E_1 \bowtie E_2) \bowtie E_3 \equiv E_1 \bowtie (E_2 \bowtie E_3)$$

    - θ连接的结合律，θ_2只涉及关系2和3

      $$(E_1 \bowtie_{\theta_1} E_2) \bowtie_{\theta_2 \wedge \theta_3} E_3 \equiv E_1 \bowtie_{\theta_1 \wedge \theta_3} (E_2 \bowtie_{\theta_2} E_3)$$

    - 选择运算在以下两个条件对θ连接运算具有分配律
      - 当选择条件θ_0中的所有属性只涉及参与连接运算的表达式之一(如E_1)

        $$\sigma_{\theta_0}(E_1 \bowtie_{\theta} E_2) \equiv (\sigma_{\theta_0}(E_1)) \bowtie_{\theta} E_2$$

      - 当选择条件θ_1只涉及E_1的属性，当选择条件θ_2只涉及E_2的属性

$$\sigma_{\theta_1 \wedge \theta_2}(E_1 \bowtie_\theta E_2) \equiv (\sigma_{\theta_1}(E_1)) \bowtie_\theta (\sigma_{\theta_2}(E_2))$$

- 投影运算在以下两个条件对θ连接运算具有分配律
  - 令L_1和L_2分别代表E_1和E_2的属性。假设连接条件只涉及他们并集的属性

    θ involves only attributes from $L_1 \cup L_2$:

    $$\Pi_{L_1 \cup L_2}(E_1 \bowtie_\theta E_2) \equiv \Pi_{L_1}(E_1) \bowtie_\theta \Pi_{L_2}(E_2)$$

  - 考虑连接 $E_1 \bowtie_\theta E_2$。令 L_1、L_2 分别代表 E_1、E_2 的属性集；令 L_3 是 E_1 中出现在连接条件 θ 中但不在 $L_1 \cup L_2$ 中的属性；令 L_4 是 E_2 中出现在连接条件 θ 中但不在 $L_1 \cup L_2$ 中的属性。那么：

    $$\Pi_{L_1 \cup L_2}(E_1 \bowtie_\theta E_2) = \Pi_{L_1 \cup L_2}((\Pi_{L_1 \cup L_3}(E_1)) \bowtie_\theta (\Pi_{L_2 \cup L_4}(E_2)))$$

- 9. 集合的并与交满足交换律。

  $$E_1 \cup E_2 = E_2 \cup E_1$$
  $$E_1 \cap E_2 = E_2 \cap E_1$$

  集合的差运算不满足交换律。

  10. 集合的并与交满足结合律。

  $$(E_1 \cup E_2) \cup E_3 = E_1 \cup (E_2 \cup E_3)$$
  $$(E_1 \cap E_2) \cap E_3 = E_1 \cap (E_2 \cap E_3)$$

  11. 选择运算对并、交、差运算具有分配律：

  $$\sigma_P(E_1 - E_2) = \sigma_P(E_1) - \sigma_P(E_2)$$

  类似地，上述等价规则将"－"替换成∪或∩时也成立。进一步有：

  $$\sigma_P(E_1 - E_2) = \sigma_P(E_1) - E_2$$

  上述等价规则将"－"替换成∩时也成立，但将"－"替换成∪时不成立。

  12. 投影运算对并运算具有分配律。

  $$\Pi_L(E_1 \cup E_2) = (\Pi_L(E_1)) \cup (\Pi_L(E_2))$$

- <ppt19书364>

- 表达式结果集统计大小的估计
  - 目录信息
    - n_r: 关系r中的元组数
    - V(A, r)：关系r中属性A出现的非重复值个数。如果A是关系r的主码，则它等于n_r
    - 假设取值均匀分布，每个之以同样的概率出现
  - **选择运算结果**大小的估计
    - $\sigma_{A=a}$： 选择基数 $SC(a, r) = \frac{n_r}{V(A,r)}$：假设关系r的一些记录的属性A中的取值为a
    - $\sigma_{A \leq v}$：属性A的最小值min(A,r)和最大值max(A,r)可存储到目录中。对满足条件$A \leq a$的记录数进行下列估计：若$a < min(A, r)$，则为0；若$a \geq max(A, r)$，则为n_r；否则满足条件的元组的估计数量为$n_r \times \frac{a - min(A,r)}{max(A,r) - min(A,r)}$
    - 合取：$\sigma_{\theta_1 \wedge \theta_2 \cdots \wedge \theta_n}(r)$的元组数量为$n_r \times \frac{s_1 \times s_2 \times \cdots \times s_n}{n_r^n}$

- 析取

  $$n_r * \left(1-(1-\frac{s_1}{n_r})*(1-\frac{s_2}{n_r})*\ldots*(1-\frac{s_n}{n_r})\right)$$

- **连接结果**大小估计
  - 关系R ∩ S = 空集，笛卡尔积
  - 交集是R的码，则小于S的元组数，若正好是S中参照R的外码，则等于n_s
  - 既不是R的码也不是S的码={A}
    - $\frac{n_r * n_s}{V(A,r)}$和$\frac{n_r * n_s}{V(A,s)}$最小的
- 投影结果：$V(A,r)$