

1. 考虑在公共属性 a 上连接关系 R(a,b)和 S(a,c,d)，假设表上没有可用的索引来加速连接算法。

缓冲区中有 B=75 页

表 R 包含 M= 2400 个页面，每个页面包含 80 个元组

表 S 包含 N = 1200 个页面，每个页面包含 100 个元组

请回答以下关于计算连接的 I/O 开销的问题。您可以假设最简单的开销模型，即每次只读写一个页面。你还可以假设你需要一个缓冲块来保存演变中的输出块，以及一个输入块来保存内部关系的当前输入块。你可以忽略写最终结果的成本。

A . 以 S 为外部关系，R 为内部关系的散列连接。您可以忽略递归分区和部分填充的块。
划分阶段的成本是多少？

$$2 \times (M + N) = 2 \times (2,400 + 1,200) = 2 \times 3,600 = 7,200$$

B . 以 R 为外部关系，S 为内部关系的块嵌套循环连接：

$$M + \lceil \frac{M}{B-2} \rceil \times N = 2,400 + \lceil \frac{2,400}{73} \rceil \times 1,200 = 2,400 + 39,600 = 42,000$$

C . 以 S 为外部关系，R 为内部关系的块嵌套循环连接：

$$N + \lceil \frac{N}{B-2} \rceil \times M = 1,200 + \lceil \frac{1,200}{73} \rceil \times 2,400 = 1,200 + 40,800 = 42,000$$

2. 设关系 R(X,Y)和 S(Y,Z)，R 共有 1000 个元组，S 共有 1500 个元组，每个块中可容纳 20 个 R 元组或 50 个 S 元组。S 中 Y 不同值的个数为 20。

(1) 若在 S.Y 上建有聚簇索引，估计 R 和 S 基于索引连接的 IO 代价。

$$\text{IO 代价为 } 1000/20 + 1000 * \text{ceil}(1500/50/20) = 2050$$

(2) 若在 S.Y 上建有非聚簇索引，估计 R 和 S 基于索引连接的 IO 代价。

$$\text{IO 代价为 } 1000/20 + 1000 * 1500/20 = 75050$$

3. 已知 2 个关系 R(A,B)和 S(B,C)，其主键分别为 R.A 和 S.B。R 有 40000 个元组，S 有 60000 个元组，一块中可以容纳 20 个 R 元组或 30 个 S 元组。设 2 个关系均采用聚簇存储，且每个关系中的元组均已按照其主键值递增排序。现在要执行自然连接操作 $R \bowtie S$ 。设缓冲区中可用内存页数为 $M = 41$ 。回答下列问题：

(1) 采用嵌套循环连接算法执行 $R \bowtie S$ 分别需要进行多少次 I/O？给出具体分析过程。

答： R 做外关系，用 40 个块存，S 做内关系 用 1 个块存：

$$\text{I/O} = 50 + (2000 * 2000) / 40 = 100050$$

(2) 采用归并连接算法执行 $R \bowtie S$ 分别需要进行多少次 I/O？给出具体分析过程。
根据题意，外存中存储 R 的结果需要占用 $40000/20=2000$ 个块。

外存中存储 S 的结果需要占用 $60000/30=2000$ 个块。

算法第一阶段 I/O 代价分析：因为 R 中元组未按属性 R.B 进行排序，所以首先需要对 R 建立归并段，需要执行 $2B(R)=4000$ 次 IO。因为 S 中的元组已经按照属性 S.B 进行排序，所以无需对 S 建立归并段。因此算法第一阶段需要执行 $B(S)=2000$ 次 IO。

算法第二阶段的 IO 代价分析：因为 $M-1=40$ ，所以 R 的归并段数量为 $B(R)/40=50$ 。算法第二阶段可以将 R 的全部归并段与 S 进行归并，需要进行 $B(R)+B(S)=4000$ 次 IO。因此，该算法总共执行 6000 次 IO。

- (3) 设 R.B 是关系 R 的外键，参照 S.B。如果 $R \bowtie S$ 的结果中元组的平均大小是 R 中元组平均大小的 1.2 倍， $R \bowtie S$ 的结果中元组的平均大小是 S 中元组平均大小的 2 倍，那么在外存中存储 $R \bowtie S$ 的结果需要占用多少个块（页）？给出具体分析过程。

因为 R.B 是关系 R 的外键，所以 $R \bowtie S$ 的结果中包含的元组数一定与 R 的元组数相同。

因为 $R \bowtie S$ 的结果中元组的平均大小是 R 的 1.2 倍，则 $1.2 * (40000/20) = 2400$ 个块。

4. 设教学管理数据库有如下 3 个关系模式：

S(S#, SNAME, AGE, SEX)

C(C#, CNAME, TEACHER)

SC(S#, C#, GRADE)

其中 S 为学生信息表、SC 为选课表、C 为课程信息表；S#、C#分别为 S、C 表的主码，(S#, C#)是 SC 表的主码，也分别是参照 S、C 表的外码

用户有一查询语句：

Select SNAME

From S, SC, C

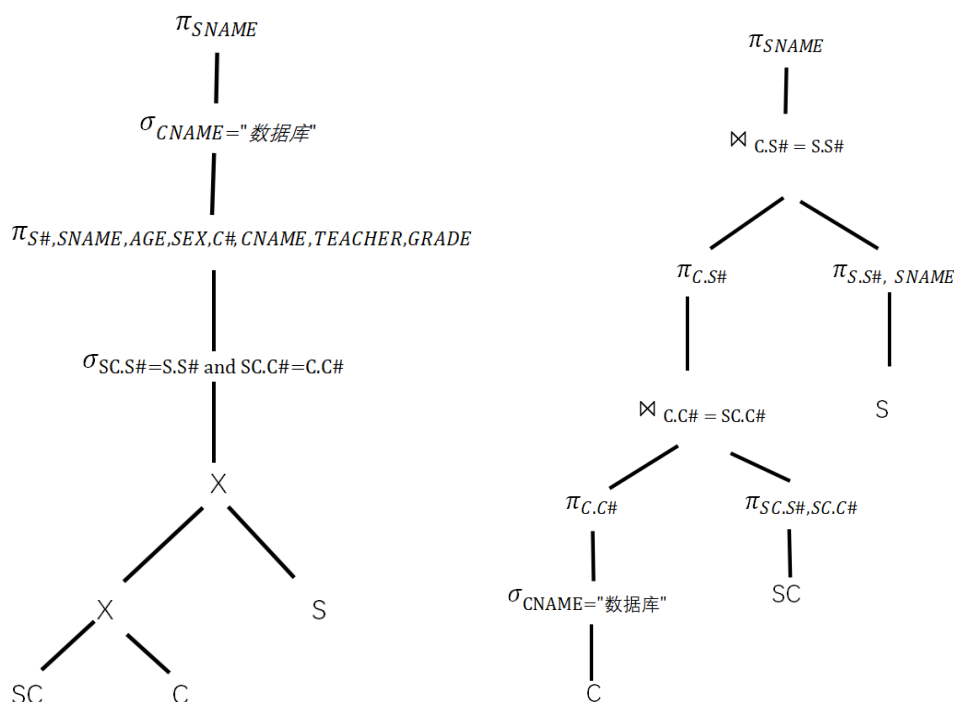
Where SC.S#=S.S# and SC.C#=C.C# and CNAME= “数据库”

检索选学“数据库”课程的学生的姓名。

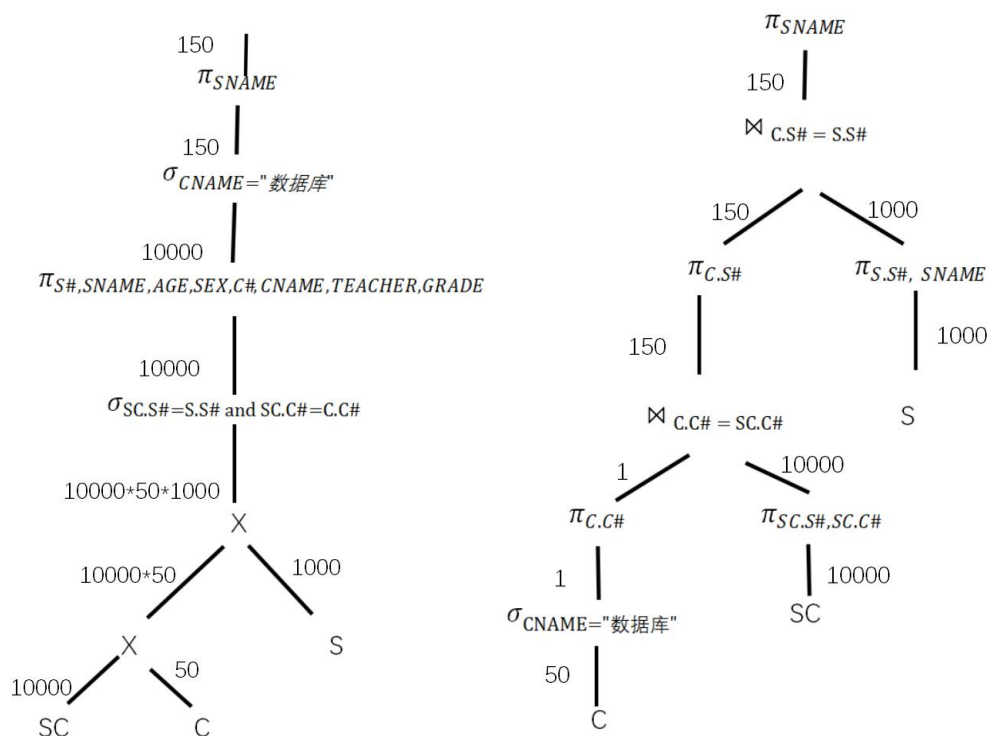
- (1) 写出以上 SQL 语句所对应的关系代数表达式。

$$\Pi_{SNAME}(\sigma_{CNAME="数据库" \wedge SC.S#=S.S# \wedge SC.C#=C.C#}(SC \times C \times S))$$

- (2) 画出上述关系代数表达式所对应的查询计划树。使用启发式查询优化算法，对以上查询计划树进行优化，并画出优化后的查询计划树。



(3) 设 SC 表有 10000 条元组，C 表有 50 条元组，S 表中有 1000 条元组，SC 中满足选修数据库课程的元组数为 150，计算优化前与优化后的查询计划中每一步所产生的中间结果大小

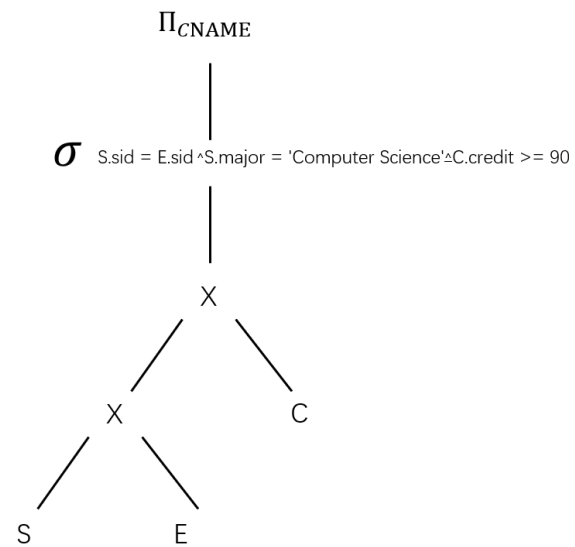


5. 给定以下关系模式，

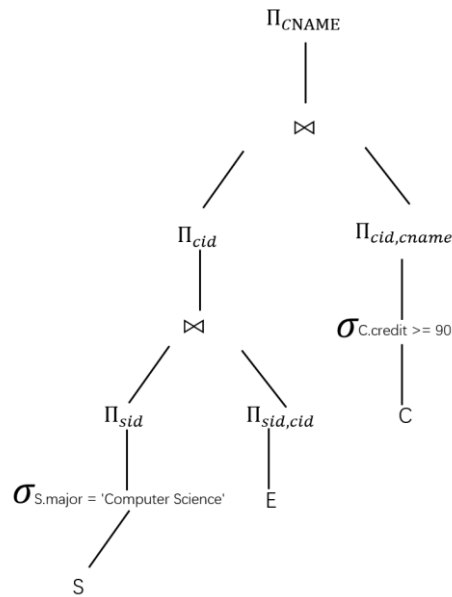
Student (sid, sname, major)
Course (cid, cname, credit)
Enrollment (sid, cid, grade)

(1) 考虑以下的 SQL 查询语句，绘制其查询计划树。

```
SELECT C.name  
FROM Student S, Course C, Enrollment E  
WHERE S.sid = E.sid  
AND S.major = 'Computer Science'  
AND C.credit >= 90;
```



(2) 假设在 Student.major 和 Enrollment.sid 上建有索引，绘制优化后的查询计划树。



6. 已知一个关系数据库的模式如下：

关系 $B(\underline{bno}, bname, author)$ 为图书表，其中 bno 为书号， $bname$ 为书名， $author$ 为作者；

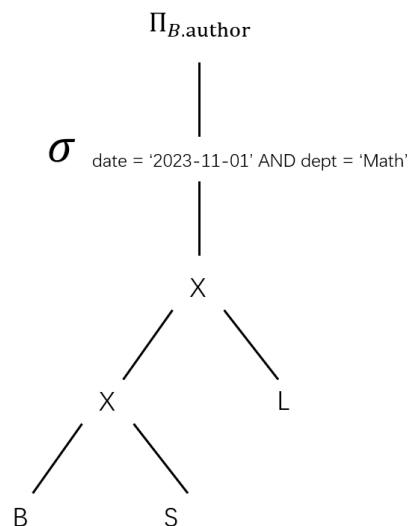
关系 $S(\underline{sno}, sname, dept)$ 为学生表，其中 sno 为学号， $sname$ 为姓名， $dept$ 为学生所在系；

关系 $L(\underline{sno}, bno, date)$ 为借书表，其中 sno 为学号， bno 为书号， $date$ 为借书时间。

回答下列问题：

(1) 绘制下面的 SQL 查询语句的逻辑查询计划树。

```
SELECT author FROM B NATURAL JOIN S NATURAL JOIN L
WHERE date = '2023-11-01' AND dept = 'Math';
```



(2) 使用启发式查询优化方法对上面的逻辑查询计划树进行优化，绘制优化后得到的逻辑查询计划树，具体说明你进行这些优化的理由。

将选择操作 $\sigma_{dept='Math'}$ 和 $\sigma_{date='2023-11-01'}$ 以及投影操作 $\Pi_{b.author}$ 进行下推可以尽早减少中间结果的大小。

