第五次作业答案:

习题 7.3（1）:

7.3 考虑关系模式 Executives(*ename, title, dname, address*{假定这些属性均为字串且等长})，关系实例包含 10,000 个页，可用缓存页数为 10。

（1）对于查询 SELECT E.title, E.ename FROM Executives E WHERE E.title='EFO'，假设有 10%的元组满足选择条件。

（a） 假定 title 上有 B+树索引，试针对索引是聚集/非聚集两种情况，说明最好计划的代价；

（b） 假定 ename 上有聚集 B+树索引，试说明最好计划的代价；

（c） 假定<ename,title>上有聚集 B+树索引，试说明最好计划的代价。

【解答】

假设单字段搜索键的索引项大小＝元组大小*0.25，B+树索引叶节点总数＝10000/4=2500；双字段组合键索引项大小=元组大小*0.5，B+树索引叶节点总数＝5000。

（1）

① 利用 title 上的非聚集 B+树索引，代价:= 2 + 2500*0.1+ （元组总数）*0.1

最好计划是文件扫描，代价：10,000 次 I/Os

若有 title 上的聚集 B+树索引可用，使用它应是最好计划，

代价=2 + 2500*0.1+ （数据文件总页数）*0.1 ＝2+250+1000=1252

② 这时索引不匹配选择条件，基本没有什么作用，最好计划是文件扫描，代价＝10,000I/Os

③ 虽然索引不匹配选择条件，但仍可用只扫描索引来实现赋值，这时需要扫描所有的页节点。代价＝组合键 B+树索引叶节点总数＝5000.

习题 7.4:

7.4 考虑针对关系 R(A,B,…)和 S(C,D,…)的连接查询 $\pi_{A,B,C,D}(R \bowtie_{R.A=S.C} S)$。假设基于排序实现带删除重复的投影，且有足够智能在排序的初始阶段就删除所有不用的属性，并在排序后通过流水线删除重复元组。已知：

- 页大小为 1K。
- R 共有 10 个页，每个元组为 300B(字节)；S 共有 100 个页，每个元组为 500 字节；
- S 中的每个元组都能在 R 中找到匹配连接元组；
- 属性 A,B,C,D 总大小为 450 字节，其中，属性 A 与 B 的合计大小为 200 字节。

（1） 估算最后输出结果的大小。

（2） 假定只有基于页的嵌套循环连接可用，分别针对有 3 个及 11 个可用的缓存页情况，计算以下赋值方案的总代价：

① “先投影后连接”；

② “先连接后投影”；

③ “先连接后投影，且连接结果通过流水线传递给投影”。

（3） 假定只有基于块的嵌套循环连接可用，重新计算（2）中的各种情况。

【解答】

（1）从已知信息，R 有 30 个元组，每页可存储 3 个 R 的元组；S 有 200 个元组，每页可存储 2 个 S 元组。由于每个 S 元组恰好与 R 的一个元组连接，连接产生的元组总数等于 S 的元组数（200）、连接投影后的元组大小等于 450，因此，最后输出的结果大小＝200×450＝90,000 字节，约 100 个页。

（2）

① 先投影后连接：根据题意，采用基于排序的删除重复投影算法，我们必须先排序作为连接内层的关系 S(含 C,D 两个字段，共有 100 个页)，

1）若只有 3 个缓存页，

排序删除重复投影 S 的代价为 $2*100*\log_2 100=1400$。假设删除重复可减少 1/10 元组，即可剩下 180 个元组，每个元组大小 250 字节，每页可存储 4 个元组。排序投影后的 S 关系大小为 50 页。

针对 R 的排序删除重复投影做法类似，R（含 A,B 两字段，共有 10 个页，30 个元组），排序代价为 $3*10*\log_2 10$＝120（R 作为外层，最后 1 次写出可以不算）。假设删除重复后只剩 9/10 的元组，即 27 个元组（每个元组大小 200 字节），1 个页存储 5 个元组，总共有 6 个页。

采用基于页的嵌套循环连接算法的代价＝6 ＋ 6×50＝306；

执行该查询赋值计划的总代价＝1400＋120＋306＝1826 次 I/Os

2）若只有 11 个缓存页，

排序删除重复投影 S 的代价为 $2*100*\log_{10}100=400$。……。

针对 R 的排序删除重复投影做法类似，R（含 A,B 两字段，共有 10 个页，30 个元组），排序代价为 $3*10*\log_{10}10$＝30……。

采用基于页的嵌套循环连接算法的代价＝6 ＋ 6×50＝306；

执行该查询赋值计划的总代价＝400＋30＋306＝736 次 I/Os

② 先连接后投影，基于页嵌套循环连接算法的代价＝10 + 10*100 = 1010 I/0s，结果关系的元组数为 200，每个元组大小 300＋500＝800 字节，共有 200 个页。

1）若只用 3 个缓存页，来排序删除重复结果关系。第一个阶段直接投影掉不用字段，产生只有 450 字节元组， 2 个元组/每页。这样，该阶段读入为 200 页，写出为 100 页（分 33 个子表、每个子表大小为 1 页）。这些子表需要另外进行 $\log_2 33=6$ 个归并阶段。因此，排序删除投影的总代价＝200＋2×6×100＝1400，加上连接时的 1010 次 I/Os，赋值的总代价为 2410 次 I/Os

2）若有 3 个缓存页，来排序删除重复结果关系。第一个阶段直接投影掉不用字段，产生只有 450 字节元组， 2 个元组/每页。这样，该阶段读入为 200 页，写出为 100

页（分 10 个子表、每个子表大小为 11 页）。这些子表需要另外进行 $\log_2 10=4$ 个归并阶段。因此，排序删除投影的总代价＝200＋2×4×100＝1000，加上连接时的 1010 次 I/Os，赋值的总代价为 2010 次 I/Os

③ “先连接后投影，且连接结果通过流水线传递给投影”。这意味着可不计算排序删除重复投影的代价，赋值的总代价只有 1010 次 I/Os。

习题 7.5（2）

7.5 考虑如下关系模式：

Emp(*eid:* integer, *sal:* integer, *age:* real, ***did:*** integer);

Dept(***did*: integer, *projid:* integer**, *budget:* real, *status:* char(10));

Proj(***projid***: integer, *code:* integer, *report:* varchar);

已知：Emp 元组总数 20,000，每个元组 20 字节；Dept 元组总数 5,000，每个元组 40 字节，每个 *did* 值平均约对应有 10%的元组；Proj 元组总数 1,000，其元组平均长度为 2,000 字节。另假定文件系统支持的页大小为 4000 字节，有 10 个可用的缓存页；另外，如果你认为必要，还可自己增设一些其它额外假定。

（2） 假定部门 *budgets* 值均匀分布在 0～100,000 范围，考虑如下查询：

```
SELECT   E.eid, D.did, P.projid
FROM     Emp E, Dept D, Proj P
WHERE    E.eid=D.did AND D.projid=P.projid
         AND D.budget>20,000 AND E.sal=50,000
```

① 列出本查询优化中，有含 1 个关系、2 个关系和 3 个关系的子计划。

② 给出本查询的最优赋值方案及其估计代价。

【解答】

（2）

① 含 1 个关系的子计划：E.sal 上的聚集索引;文件扫描 Dept;文件扫描 Proj。

含 2 个关系的子计划：

1） 以 E.sal 上的聚集索引扫描为左关系；用 did 上的索引探测 Dept,并应用 budget 选择过滤 Dept 元组。

2） 应用 budget 选择条件，扫描 Dept 作为左关系，探测 Proj。

3） 扫描 Proj 作为左关系；探测连接 Dept（下推选择 σ budget>20,000）；

含 3 个关系的子计划：

1） 先连接 Emp 和 Dept，探测 Proj；

2） 先连接 Dept 和 Proj， 探测 emp。

② 最优方案是：利用 E.sal 上的索引删除大部分元组，基于 D.did 探测 Dept，并同时处理 σ budget>20,000；结果元组再流水线方式传送给下级基于 Proj.projid 的索引连接。具体代价估计略。