# 第 20 讲模拟练习题解析

**2001**、已知内存共有 8 块，若要排序有 70 块的数据集，应如何组织，才能使磁盘读写次数最少。下列方案中磁盘读写次数最少的方案是_____。

正确答案：A。解析：少量内存排序大规模数据，首先是要划分子集合并进行子集合排序。划分原则是子集合块数<=可用内存块数，然后将其装入内存并进行排序后再写回磁盘。此一步骤四个方案都满足要求，且磁盘读写次数都是 70*2=140 次(读一次，写一次)。关键是多路归并的磁盘读写次数的差异。方案 I，先做三路归并(3 个子集合*(8 块子集合+8 块子集合+6 块子集合)*2 次=44 次—因有一个集合为 6 块)，再做 7 路归并(70*2=140 次)，所以总的磁盘读写次数为 140+140+44=324 次。方案 II，先做五路归并(5 个子集合*7 块每个子集合*2 次=70 次，再做六路归并(70*2=140 次)，所以总的磁盘读写次数为 140+140+70=350 次。方案 III，先做七路归并(7 个子集合*8 块每个子集合*2 次=112 次，再做三路归并(70*2=140 次)，所以总的磁盘读写次数为 140+140+112=392 次。方案 IV，先做五路归并(5 个子集合*8 块每个子集合*2 次=80 次，再做五路归并(70*2=140 次)，所以总的磁盘读写次数为 140+140+80=360 次。通过比较：选项 A 方案的磁盘读写次数最少。

**2002**、已知内存共有 100 块，若要排序有 10000 块的数据集，则下列说法正确的是_____。

正确答案：B。解析：100 块内存，每个子集合 100 块，10000 块数据集需要划分为 100 个子集合。100 块内存，留出一块作为输出块，则最多可进行 99 路归并。因此在进行最终排序前，需要先做一个 2 路归并，即将 2 个子集合归并成 1 个有序集合，然后再做 99 路归并。因此，“该数据集不能在两趟内实现排序”，需要“外加一个 2 路归并”，因此总的次数应为 10000*4+100（一百块每个子集合）*2（两个子集合）*2（读一次写一次）=40400 次。

**2003**、已知内存共有 8 块，若要排序有 100 块的数据集，则给定多路归并算法如下：(1)以 8 块为一个单位划分子集合，每个子集合进行内排序并存储，形成 13 个已排序子集合(含一个仅有 4 块的子集合)；(2)接着在 13 个子集合中任选 7 个子集合（包含仅有 4 块的子集合）进行一个七路归并，形成一个已排序子集合；(3)再将剩余 6 个子集合与刚才归并后形成的子集合，进行一个七路归并，形成最终的已排序集合。问：这个方案的磁盘读写次数是_____。

正确答案：C。解析：8 块内存，每个子集合 8 块，100 块数据集需要划分为 13 个子集合（其中一个仅有 4 块），读一次写一次，子集合划分并排序需要 100*2 次磁盘读写。8 块内存，留出一块作为输出块，则最多可进行 7 路归并，第一个七路归并的次数为 52（六个 8 块子

集合+一个 4 块子集合）*2（读一次写一次）=104 次。最后一个七路归并的次数为 100*2（读一次写一次）=200 次，因此，总的次数为 200+104+200=504 次。

**2004**、关于基于排序的两趟算法，下列说法不正确的是_____。

正确答案：D。解析：选项 A 说法是正确的。选项 B 说法是正确的，归并过程中，重复元组会按归并次序出现，只要在归并过程中将重复元组去掉，即完成去重复操作。选项 C 说法是正确的，归并过程中，相同分组的元组会按归并次序出现，只要在归并过程中将同一分组的相关元组进行聚集计算即可完成分组聚集计算操作。选项 D 说法是不正确的，按该选项说法是不能完成集合并操作的，集合并操作的关键是在归并过程中是否存在 R 与 S 相同的元组，相同的元组只保留一个。应该将 R 与 S 同时进行归并，并区分是 R 的元组还是 S 的元组，然后判断 R 的元组和 S 的元组是否相同，只保留一个。

**2005**、已知关系 R 和 S。关系占用的磁盘块数 B(R)=1000，B(S)=1000，已知可用内存页数 M=40。采用基于排序的算法，下列说法不正确的是_____。

正确答案：C。解析：关系 R 需划分为 25 个子集合，S 需划分为 25 个子集合，子集合排序并存储，此为第一趟。在第二趟归并过程中至少需要 50 个内存块，而目前只有 40 个内存块，一次一趟归并不能够完成。因此：选项 A 说法是不正确的，集合并操作需要去重复，而 R 的排序和 S 的排序都不能在一趟内完成。选项 B 说法是不正确的，集合并操作需要去重复，由于不能同时进行 R 和 S 的归并排序(需要先将 R 的 12 个子集合归并成 1 个子集合，使 R 有 14 个子集合，这样才能同时进行 R 和 S 的归并)，即不能在两趟内完成。选项 C 说法是正确的，包的并不需要去重复，所以一趟内可以完成。选项 D 说法是不正确的，包的并不需要去重复，所以一趟内可以完成，不必用两趟。

**2006**、已知关系 R 和 S。关系占用的磁盘块数 B(R)=1000，B(S)=500，已知可用内存页数 M=50。采用基于排序的算法，下列说法正确的是_____。

正确答案：B。解析：关系 R 需划分为 20 个子集合，S 需划分为 10 个子集合，子集合排序并存储，此为第一趟。在第二趟归并过程中需要 30 个内存块，目前有 50 个内存块，一次一趟归并能够完成。因此：选项 A 说法是不正确的，集合并操作需要去重复，而 R 的排序和 S 的排序都不能在一趟内完成。选项 B 说法是正确的，集合并操作需要去重复，而 R 和 S 的同时归并是能够在两趟内完成的。选项 C 说法是不正确的，集合交操作需要比较是否相同，而 R 和 S 的同时归并能够在两趟内完成但一趟完成不了。选项 D 说法是不正确的，包的并不需要去重复，所以一趟内可以完成，不必用两趟。

**2007**、关于基于散列的两趟算法，下列说法不正确的是_____。

正确答案：B。解析：选项 A 说法是正确的。选项 B 说法是不正确的，如果选择与第一趟相同的散列函数，则相当于每有散列，因为同一散列子表俱有相同的散列值。选项 C 说法是正确的。选项 D 说法是正确的。

**2008**、基于散列的两趟算法和基于排序的两趟算法，其中第一趟都是划分子表，都要求子表的存储块数要小于可用内存数，以便子表可以一次性装入内存进行处理。关于划分子表，下列说法正确的是_____。

正确答案：C。解析：基于排序的算法总是可以均匀地划分子表(即每个子表的大小都一样，除最后一块外)，它是先划分子表，再一个一个将其装载入内存进行排序，然后再存回磁盘，所以总是均匀是可以做到的；基于散列的算法不能保证总是均匀地划分子表，它依赖于散列函数的选择以及主文件数据的分布，通常情况下可以做到准均匀的分布，但不能保证总是均匀分布。故选项 C 的说法是正确的。

**2009**、关于 R 与 S 的并、交、差运算的基于散列的两趟算法，其中第一趟都是划分子表，都要求子表的存储块数要小于可用内存块数，以便子表可以一次性装入内存进行处理。关于划分子表，下列说法正确的是_____。

正确答案：A。解析：做 R 与 S 的并、交、差等运算，需要进行“两个关系中元组是否相同”的比较，因此 基于散列的两趟算法在划分子表时必须保证的一个特性是：R 的元组 a，和 S 的元组 b，如果相同，则必须存在于各自的具有相同散列值的子表中。为保证这一特性就必须采用相同的散列函数，散列 R 和 S，选项 A 说法是正确的。

**2010**、关于基于散列的两趟算法和基于排序的两趟算法的基本思想，下列说法正确的是_____。

正确答案：B。解析：选项 B 说法是正确的。排序算法是先划分子表，独立处理子表（第一趟），然后再对各子表进行关联性处理（第二趟）；散列算法是先从关联性角度处理，形成子表（第一趟），然后再独立处理每一个子表（第二趟）。

**2011**、关于连接运算 R (JOIN on R.A=S.B) S 的基于散列的两趟算法，下列说法正确的是_____。

正确答案：C。解析：选项 C 说法是不正确的，其对应的正确的说法是“散列过程中，R 必须以 A 属性值作为散列函数的键值，S 必须以 B 属性值作为相同散列函数的键值”。

**2012**、关于基于散列的两趟算法，下列说法正确的是_____。

正确答案：C。解析：选项 C 说法是正确的。第一趟散列的目的是使数据子集具有某一种特

性(如具有相同的散列值)，以便于将“大规模数据全集上的操作”等价地 转换为“(数据子集上操作)的简单并集”。而第二趟散列的目的是提高数据处理的速度，散列到不同内存块中，使得比较时快速地和少量内存块中的数据进行比较。