11 Indexing and Hashing

1. Basic Concepts  
   索引分为顺序索引(ordered indixes)和散列(has indices)索引，前者基于值得顺序；后者将值平均分布到若干bucket中，值所属的bucket由散列函数决定。  
   索引和散列的实现技术有多种，但没有哪一种是绝对最好的，每种方式有其最适合的场景，可通过这几个方面来进行评估：访问类型（能有效支持的访问类型，如特定值查找、范围查找）、访问时间、插入时间、删除时间、空间开销（索引额外占用的空间）。
2. 顺序索引  
   为了快速随机访问记录，可以使用顺序索引。用于在文件中查找记录的属性或属性集称为搜索码(search key)，每个索引结构与一个特定的搜索码关联，并按顺序存储搜索码的值，并将每个搜索码与对应的记录关联起来。  
   被所以的记录本身也可以按一定的顺序存储，如果记录按照某个搜索码指定的顺序排序，则该搜索码对应的索引称为聚集索引(clustering index)或主索引(primary index)，聚集索引的搜索码一般是主键，但不是必须这样。搜索码指定的顺序与记录的物理存储顺序不同的索引称为非聚集索引(noclustering index)或辅助索引(secondary index)。
   1. 稠密索引和稀疏索引  
      索引项由搜索码和指向具有该搜索码值的若干条记录的指针构成，指针包含block的标识和标识磁盘块内记录的块内偏移量。顺序索引又分为稠密索引(dense index)和稀疏索引(sparse index)。  
      稠密索引中，每个搜索码都有一个索引项。索引项包含搜索码值以及指向具有该搜索码值的第一条数据记录的指针。如果是稠密聚集索引，具有相同搜索码值的其余记录可以顺序地存储在第一条数据记录之后；对于非聚集索引，则需要没每条搜索码值建立索引。  
      稀疏所以中，只为搜索码的某些值建立索引项。只有索引是聚集索引时才可以使用稀疏索引，这样即使索引中没有要查找的搜索码，也可以根据索引确定搜索码的区间，进而找到要查找的记录。
   2. 多级索引  
      如果记录数很多，其索引也会变得很大。假定100条索引占用一个4K的block，那么100,000,000条记录的索引的体积会高达4G。只有将全部索引都装载到内存，才能最大化索引的速度，如果计算机内存小于4G，那么索引就无法完全放入内存中，即使内存够大，也还要供给其他应用程序使用。  
      在数据量很大、索引文件较大时可以使用多级索引，即在原先内层索引的基础上再建立若干外层索引，外层索引与内层索引之间的关系同前面索引与数据文件的关系一样。多级索引与树结构紧密相关，比如用于内存索引的二叉树。
   3. 索引的更新  
      在插入、删除数据的时候，对应的索引需要更新；而数据被修改时，其索引项不是直接修改，而是分为“删除就搜索码值，插入新搜索码值”这两步执行的，所以索引的更新只涉及插入和删除。
      1. 插入
         1. 稠密索引  
            如果插入数据的搜索码值在索引中不存在，则直接插入索引。  
            如果插入的搜索码值已经存在且指向多条数据，则在这条索引项下新增指向插入数据的指针。  
            如果插入的搜索码值已经存在但只指向第一条包含该搜索码值的记录，则仅将插入数据放到具有相同搜索码值的其他记录之后。
         2. 稀疏索引  
            假设稀疏索引为每个块保存了一个索引项。如果插入数据时，系统需要新创建一个块，则在索引中插入针对新块中第一条记录的搜索码的索引项。如果新插入的数据刷新了块中记录的最小搜索码值（按照搜索码的顺序存储），这个指向这个块的索引项需要被更新；其余情况不需要更新索引。
      2. 删除
         1. 稠密索引  
            如果待删除的记录的搜索码值是唯一的，则删除对应的索引。  
            如果待删除记录的搜索码值不唯一，且索引项指向所有具有该搜索码值的记录，则只删除索引项中指向待删除记录的指针。  
            如果待删除记录的搜索码值不唯一、索引项仅指向第一条具有该搜索码值的记录，而且第一条记录待删除，在删除这条记录的同时，将索引项指向下一条具有该搜索码值的记录。
         2. 稀疏索引  
            如果不存在包含被删除记录的索引项，则不做任何操作。  
            如果被删除的记录是具有该搜索码值的唯一记录，则索引项A指向下一个包含该搜索码值的记录，但如果下一个包含该搜索码值的记录已经有索引项，则删除索引项A。  
            如果被删除的记录并非是具有该搜索码值的唯一记录，则索引项指向下一条具有相同搜索码值的记录。