第三次全国工业普查数据库系统的设计和优化

周子斌　孔祥清　（国家统计局计算中心　100826）

一、前言

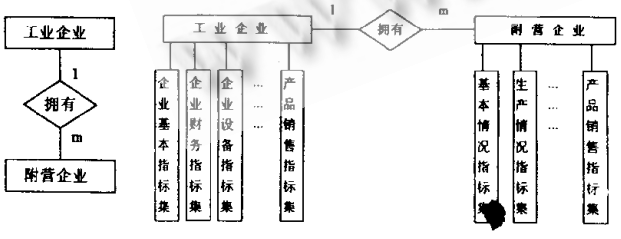
第三次全国工业普查是国务院决定建立我国周期性普查制度后的第一个大型普查，是我国国情国力的一次重大调查。工业普查指标多，数据结构复杂，且数据量大，数据库系统设计是一项非常艰难的工作。工业普查数据库系统是国内开发研制的第一个大型普查数据库应用系统，采用当前流行的 CLIENT/SERVER 结构，经过近一年的运行实践证明，系统的设计是优秀的，运行效率是高的。

本文从数据库理论出发，阐述了第三次全国工业普查数据库系统的设计方法，并结合我们选用的 DBMS (ORACLE RDBMS)以及其他软件和硬件环境，探讨了该系统进行优化的策略。

二、工业普查数据库设计

1.概念设计

一个工业企业可能有多个附营企业，而一个附营企业只能附属于一个工业企业，用 E－R 图表示工业企业和其附营企业的联系如图 1(a)所示，图中可以看到，这是一个 1:n 的联系。对于某个工业企业，又和很多其他相关的实体和普查指标相联系，因此，此联系又可以表示成图 1(b)。



(a)　　(b)

图 1

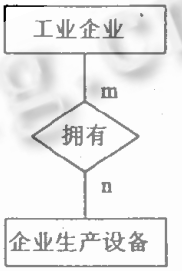


图 2

为了进一步分析工业企业中的联系，这里以工业企业及其生产设备两个实体间的联系为例来说明分析方法。作为工业企业，它进行生产必须拥有相应的生产设备，一个工业企业可以拥有多种生产设备，而每一种生产设备(比如机床)又可能为多个工业企业(比如机械厂)所使用。用 E－R 图表示这种联系如图 2 所示。由图可以看出，这是一个 m:n 的联系。用同样的方法可以分析出工业企业的其他联系。

2.逻辑设计

(1)E－R 图向关系模型转换

图 1 可以转换为两个关系：

工业企业(企业法人代码，企业名称，企业详细地址，…，资本金合计，…)

附营企业(附营企业代码，企业法人代码，附营企业名称，附营企业详细地址，……)

图 2 可以转换为关系：

工业企业(企业法人代码，企业名称，企业详细地址，…，资本金合计，…)

企业设备(企业法人代码，设备代码，设备生产年代，年末设备能力，……)

(2)数据模型的优化

①关系的垂直分割。第三次全国工业普查涉及的工业企业指标众多，如果把所有的指标都放在一个关系中，势必影响数据库的运行效率，因此，必须按指标的性

质对关系进行垂直分割。

按指标的特性，工业企业关系可分割为：

工业企业基本情况(企业法人代码，企业名称，企业详细地址，……)

工业企业财务状况(企业法人代码，资本金合计，固定资产原价，……)

……

……

……

②关系的水平分割。工业企业关系中，有些记录是每一类应用所关心的，而另一些记录则为另一类应用所关心，这种情况下，采用关系的水平分割法可以提高应用的效率。

例如，工业企业财务状况按年度可分割为：

工业企业九五年财务状况(企业法人代码，资本金合计，固定资产原价，……)

工业企业九四年财务状况(企业法人代码，资本金合计，固定资产原价，……)

③在规范化的基础上适当增加冗余。企业基本情况(企业法人代码，企业名称，企业详细地址，…，从业人数，男性，女性，……)按规范化方法，必须删除属性'从业人数'，企业基本情况才满足 3NF，但为了方便查询和统计习惯，考虑到普查数据不是经常要修改的特点，因此设计时仍然保留了这个冗余。

可以证明，除去保留的一些常用的合计数据属性，工业企业的大部分关系满足 3NF。

3.综合数据的关系分析

综合数据是由工业企业基层数据汇总而来的，客观上找不到实体，这里采用把汇总口径和分组做为虚拟实体分析的方法。结果如下：

口径(口径码，口径名)

分组(分组码，分组名)

汇总关系(口径码，分组码 1，分组码 2，…，分组码 n，指标 1，指标 2，…，指标 m)

4.工业普查数据库系统功能设计

以上从理论上探讨了工业普查数据库设计，作为一个数据库应用系统，还要进行功能模块设计。系统设计如图 3 所示，从图中可以看出，系统具有较完善的功能，是一个较完备的统计数据库应用系统。

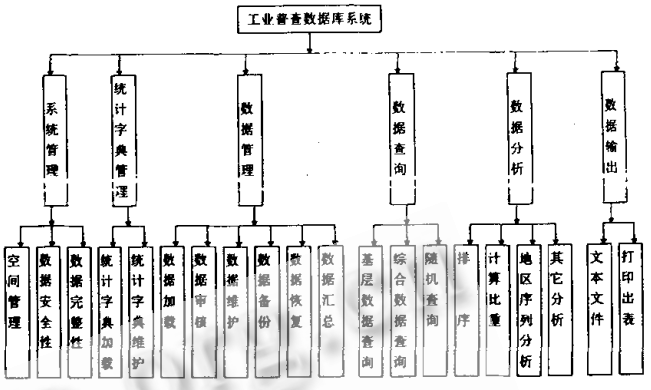


图 3

三、工业普查数据库系统的优化

数据库系统的优化涉及到系统的硬件和软件环境，和硬件相关的主要有文件 I/O、存储器、CPU 和网络等方面的优化，软件则有数据库 DBMS、数据库应用软件和操作系统软件等方面的优化。这里主要从 ORACLE 数据库系统出发，重点对应用系统的优化进行探讨。

1. I/O 的优化

数据库应用执行过程中，最重要的操作是查询和修改数据，这些操作大多由 I/O 来完成，实践证明 I/O 往往是数据库应用的瓶颈。这也是 CPU 运算速度的高速发展，而 I/O 的速度发展相对较慢的结果。提高 I/O 的效率是数据库优化的重要手段。

工业普查数据库系统中，I/O 的优化主要采用下列策略：

·建立应用系统自己的数据表空间、索引表空间、回退段表空间和临时段表空间，不让应用系统占用系统表空间。

·数据表空间和索引表空间建立在不同的盘驱上。

·建立两个以上位于不同盘驱上的回退段表空间，把回退段合理建立在各个回退段表空间，降低回退段的竞争。

·估算出数据表和索引的大小，减少扩展 extent 的数量。

·使用卸出/装入或相关工具，降低数据库表空间的碎片。

例如，工业普查数据库系统在不同的磁盘上建立了三个自己的数据表空间 GYSPACE1、GYSPACE2 和 GYSPACE3，它们分别存放统计字典、索引和普查数据。

另外，在不同的磁盘上分别建立了自己的回退段表空间、临时段表空间。这样，减少了资源竞争，提高了运行效率。

2. 利用索引技术提高查询速度

索引是大型数据库的精华，建立好的数据库索引机制，能很大地提高数据查询效率。

(1)索引字段的选取。索引字段的选取，对索引的效率至关重要，可以采用以下规则选取。

·查询时在条件中经常用到的字段

·多表连接时的连接条件字段

·相同值少的字段

·数据长度小的字段

·最大值和最小值经常被查的字段

(2)索引的创建。工业普查数据库是一个大型数据库应用系统，企业有七十多万个，最大的基层表有八百多万条记录，查询时可能还要进行多表间的连接，为了提高查询效率，系统建立了较完备的索引机制。

创建索引除了按上面的规则选取索引字段外，还遵守了以下规则：

·记录数大的表必须建索引

·常在条件中一起出现的字段适合建复合索引

·无相同值的字段应建唯一性索引

例如，在工业企业基本情况表中，企业法人代码的值是唯一的，且多个表之间的连接都可以通过它的值相等来进行，因此，以企业法人代码为索引字段，建立唯一性索引是一个很好的查询优化方法。

(3)索引的使用方法。索引的使用有很多技巧，只有充分了解索引的原理和系统使用索引的规则，正确地使用索引，才能真正提高系统的数据查询速度。不正确地使用索引，反而会降低查询的效率。

掌握好索引技术，必须了解索引规则，下面列出一些常见规则。

·条件中在 WHERE 和 AND 谓词后出现的索引字段会使用索引。

·索引字段若被一个函数引用则不会使用索引。

·复合索引的第一个字段，在条件中的 WHERE 和 AND 谓词后会使用复合索引。

·查询的记录数超过表总记录数的 15% 时不宜使用索引。

·! = 谓词不使用索引。

·IS NOT NULL 和 NOT IN 谓词不使用索引。

·同等谓词情况下系统优先使用唯一性索引。

·LIKE 谓词后匹配格式为 'X%' 时使用索引，'%X%' 时则不使用索引。

·OR 谓词一般不使用索引。

3. SQL 的优化

除了使用索引技术外，SQL 语言的优化还有很多其他技术，下面再介绍几种常用的技术。

(1)表连接的优化。在数据库中访问数据，表的连接是经常要进行的一种活动。由于各表的结构和记录数不同，加上某些字段上有可用的索引等因素，表连接的方法对数据访问的效率影响很大。表连接操作语句是比较复杂的 SQL 语句，它的优化也是比较难的。实践证明，下列方法是有效的。

·把无索引的表作为驱动表

·选取产生较少记录数的条件先连接

·尽量使用索引连接方式

例如，考察 SQL 语句：

```
select a.b01, a.b02, b.v08 from hb601 a, hb603 b
where a.b01 = b.b01 and (b.v08 < 500000 and b.v06
= 5001)
```

如果 hb601 表的 b01 字段建立了索引，且 hb603 中满足 v06 = 5001 条件的记录数较 v08 < 500000 少，则是一个好的 SQL 语句。

(2)嵌套子查询的优化。嵌套子查询是这样一种查询，该查询的结果作为其他查询或操作的条件。和表连接一样，含有嵌套子查询的数据操作是比较复杂的 SQL 语句。嵌套子查询的优化技术主要有：

·空值存在时，连接嵌套子查询的谓词 NOT EXISTS 应改为 NOT IN。

·用嵌套子查询求出数据库 ROWID 来定位主操作，比求出相关数据行来定位效率更高。

参考文献

[1] oracle7 server administrator's guilds, oracle company, 1995.

[2] 客户/服务器(client/server)结构下统计数据库的设计与实现，周子斌，中国统计，1997 - 10，P35 - 37.

(来稿时间：1998 年 2 月)