数据集市≠数据仓库

NCR公司可扩展数据仓库解决方案小组　　王闯舟　编译

我们知道，决策支持系统 (DSS)主要有两种实现方式 ，即建立一个数据集市或者一个数据仓库。到底哪一种更能满足决策支持的要求并且适合企业今后的发展 ，是近两年来学术界和有关供应商激烈争论的一个话题。

在数据集市领域 ，主要的供应商和拥护者以美国红砖 (Red Brick) 公司为代表，其总裁 Ralph Kimball 在 1997 年 12 月的一篇论文中提出 ," 数据仓库只不过是一些数据集市的集合而已 "。认为企业多建立一些数据集市 ，将来自然就形成了数据仓库。而业界公认的数据仓库之父 Bill Inmon 在今年 1 月立即撰文反驳 ，旗帜鲜明地指出 ," 你可以在大海中捕到很多的小鱼并堆积起来 ，但它们仍然不是鲸 "。在 5 月份的《数据管理综述》(DataManagement Review) 中,Bill Inmon 又发表了 " 数据集市不等于数据仓库 " 的论文，进一步阐述两者在本质上的区别以及各自的适用场合 ，本文就是根据这篇论文的主要内容编译而成的。

问题的提出

现在，各企业 IT 部门的经理所面临的最主要问题之一是先建立数据仓库还是先建立数据集市。长期以来，数据集市供应商们不断地给他们灌输这样的观念，即建立数据仓库比较复杂，投资过大，设计与开发周期太长，难以集成和管理企业范围内的各种源数据；并认为，基于数据仓库的 DSS 投资方案难以得到企业管理层的批准。数据集市供应商们给业界描绘了一幅数据仓库前景暗淡的图画，这完全是出于自身的目的，是不正确的。

数据集市供应商们把数据仓库当成其增加营业收入的绊脚石，自然要避开和攻击数据仓库。事实上，他们在销售时强调数据集市的建设周期短，是以企业信息系统结构的长期规划为代价的。

持数据集市主张的人认为，决策支持系统的成功实现，除了数据仓库以外，还有更简便、更有效的其它途径。方法之一就是建立多个数据集市，当它们增加得足够大时，那就是所谓的数据仓库了。这些人声称，建立数据集市要快得多也便宜得多，因为当考虑建立一个数据集市时，不必考虑各部门之间的区别，也不必设立部门之间协调的规则，更不存在结构设计上的长期规划问题。

不幸的是，这种方法虽然避免了建立数据仓库存在的部门协调与规划上的问题，却完全偏离了数据仓库的要点。当企业的信息结构完全由数据集市构成时，其整个组织将变得更加混乱。因为在建立决策支持系统以前，我们可能只是原来的生产系统有些凌乱，现在的状况则可能是凌乱的生产系统再加上杂乱的数据集市。由于企业内所有的决策支持系统均是数据集市，相互之间没有集成，其结果可想而知——没有集成的决策支持系统就像没有骨骼的人体一样，是没有实用价值的。

方式的改变

早期，数据集市供应商们宣称数据集市和数据仓库是相同的系统，试图通过这种偷梁换柱的方式来进入数据仓库市场。在各种展示会期间，他们不遗余力地进行着各种宣传，从而混淆了数据集市与数据仓库的概念。

由于这种错误概念的传播，使一些客户建立了数据集市而非真正的数据仓库。但随着时间的推移，数据集市结构上的缺陷开始暴露出来，主要体现在以下几点：

1) 各数据集市之间对详细数据和历史数据的存储存在大量冗余；

2) 同一个问题在不同数据集市的查询结果可能不一致甚至相互矛盾；

3) 各数据集市之间以及与源生产系统之间难以管理。

总之，业界已经普遍认同，一个没有数据仓库而建立的决策支持系统是很难达到预期效果的。

大量事实表明，为了处理决策支持方面的需求，建立数据集市不是正确的途径。在这种情况下，数据集市供应商们及其代言人稍微改变了一些原来的说法，向客户承诺成功实施决策支持系统的新方式。和原来不同的是，他们现在宣称，数据仓库只不过是多个数据集市的集成而已。这从另外一方面混淆了数据仓库与数据集市的概念。事实上，这样的论断是矛盾的。因为数据集市的实质就是每个部门拥有自己的数据，最终用户各自负责自己的业务，相互之间没有关系，各集市之间没必要也没办法相互集成。

为了理解为什么数据集市不能转变为数据仓库，我们首先必须搞清楚两者的定义。

框架的不同

1. 什么是数据集市

一般说来，一个数据集市是按照某一特定部门的决策支持需求而组织起来的、针对一组主题的应用系统。例如，财务部拥有自己的数据集市，用来进行财务方面的报表和

分析，市场推广部、销售部等也拥有各自专用的数据集市，用来为本部门的决策支持提供辅助手段。

这些部门数据集市之间相似之处很少，但最严重的缺点是，每个部门独立拥有自己的硬件平台、软件平台、数据和应用程序。这种关系使得部门之间没有任何约束，而许多数据在整个企业内原本应该是相互制约、相互协调的。这种独立最终导致了不一致性。

由于每个部门有自己特定的需求，因此他们对数据集市的期望也不一样。一般说来，数据集市中数据库的设计采用星形连接(Star-Join)的结构，这种结构对部门用户而言是最优的，但对企业范围而言则不然。为了提高星形连接的性能，必须事先收集齐该部门业务用户的需求。数据集市中包含的历史数据不很全，其详细程度也不够，数据选取的基本原则是能满足本部门的需求。数据集市大都采用多维数据库技术，这种技术对数据的分析而言也许是最优的，但肯定不适合于大量数据的存储，因为多维数据库的数据冗余度很高。为了提高速度，对数据集市中的数据一般都建立大量的索引。换言之，数据集市中往往靠对数据的预处理来换取运行时的高速度，当业务部门提出新的问题时，如果不在原来设计的范围内，则需要数据库管理员对数据库作许多调整和优化处理。

业界有两种数据集市，即从属数据集市和独立数据集市。前者的数据来源于中央的数据仓库，后者的数据则直接来源于源应用环境。所有的从属数据集市都从属于同一个数据仓库，各子系统的数据均能保持一致，因此这种数据集市的结构是可行的。而每个独立数据集市都从各源生产系统中单独提取数据，无法保证数据的一致性；从长远来看，这种结构是不稳定也是不可行的。图 1 清楚地说明了两者在结构上的区别。遗憾的是，独立数据集市的这些问题在开始往往反映不出来，企业只有在建立了多个独立数据

集市之后才能认识到其缺点。数据集市供应商们所大力宣传的其实正是这种独立数据集市，因此在本文的讨论中，我们所指的数据集市也是独立数据集市。

@@0489400.JPG 图 1@@

2. 什么是数据仓库

数据仓库与数据集市之间具有很大的差异。数据仓库是基于整个企业的数据模型建立的，它面向企业范围内的主题。一般来讲，数据仓库是由一个中央的协调组织（例如传统的 IT 部门）来建立和管理。数据仓库完全是整个企业共同努力的结果。

某个部门的主题与企业的主题之间可能存在也可能不存在关联。数据仓库中存储整个企业内非常详细的数据，相对而言，数据集市中数据的详细程度要低一些，相反，它包含了许多概要和累加数据。数据仓库的数据模型一般是规范的，比较多的是符合第三范式。其数据的结构和内容反映的不是某个特定部门的特殊要求，它代表的是整个企业对于数据的需求。数据仓库中的数据量与数据集市差别很大，因此，数据仓库中的索引很少。这和传统的 OLTP数据库有很大的区别。 数据仓库中包含有相对稳定的历史数据，所有数据都是从许多操作数据源中经一定的业务规则转换并集中进来的。简而言之，在数据仓库与数据集市中，无论是数据的结构还是其内容都存在着显著的差别。图 2 形象地说明了这种区别。左边的数据集市是星形连接结构，而右边的数据仓库是正则结构，各实体之间通过外键 (Foreign Key) 连接。

@@0489401.JPG 图 2@@

由于数据仓库中的数据是详细的、集成的和历史的，其中的数据量一般都很大，而且随着时间的推移，增长速度也非常快。因此，建立数据仓库最好是分步进行，否则建设周期将非常长。即使从最早的文献来看，学术界就几乎公认建立数据仓库必须使最终用户能尽快看到具体、明确的结果。直到现在，有关的专栏作者和咨询顾问们还是一致认

为数据仓库的建设速度必须很快 ，尽量避免冗长、庞大的投资行为。当然 ，这并不意味着数据仓库的投资小 ，正确的理解是，数据仓库一般是从小处着手 ， 取得一定成效后再逐步完善。世界上许多成功的 1000GB(指用户数据量而非数据库大小) 级以上的数据仓库在开始时的规模都不大 ，这就是所谓的 " 全盘考虑，逐步完善 " 的思想。

图 3 给出了建立数据仓库的正确途径。从图中可以看出 ，数据仓库的建设是分步进行的，每步都能取得阶段性的成果 ，不需要等到二、三年后才能访问数据仓库中的信息。

@@0489402.JPG 图 3@@

目前，数据集市的理论是 ，先建立一个或多个数据集市 ，然后把它们集成起来 ，当它们增长到一定规模时就变成了数据仓库。遗憾的是 ，这种理论在很多方面都站不住脚 ：

1) 数据集市是设计用来满足部门需求的 ，各部门的目标可能差别很大 ，这也是为什么企业内各部门拥有结构和特征都不同的数据集市的原因。数据仓库则是设计用来满足企业综合需求的。一个设计方案可以是对一个特定部门最优的 ，也可以是对一个企业最优的，但不可能对两者均是最优方案。针对企业的设计目标和针对部门的差别很大。

2) 数据集市与数据仓库中数据的详细程度也完全不同。数据集市中包含有许多概要和累计数据 ，而数据仓库中则包含有大量的详细数据。显然 ，你可以从详细数据中计算出概要和累加数据 ，但反之则不行。对业务分析而言 ，详细数据在很多场合都非常重要。

综上所述，我们可以归纳出以下要点 ：

· 数据集市和数据仓库中的数据模型不同 ，前者一般采用星形连接结构 ，后者则用第三范式为主；

· 数据集市中的历史数据信息量比数据仓库少很多 ；

· 数据集市中的主题和数据仓库中的主题关联并不很多　　；

· 数据集市中的关系与数据仓库中的关系不同　　；

· 数据集市中的查询类型与数据仓库中的查询类型差别很大　　；

· 数据集市中的用户类型　（ 较低层次) 和数据仓库中的用户类型　（ 较高层次) 差别很大；

· 数据集市的主要结构与数据仓库的主要结构具有显著的区别。

小结

数据集市与数据仓库应用环境的差别很大　，如果认为一个数据集市在增长到一定程度时能转换成数据仓库　，那无异于说小草可以长成橡树。虽然这两种绿色植物在生长的某个阶段具有一些相同的特征　，但这并不能遮盖两者的区别。这毕竟是现实世界　，数据集市与数据仓库的道理是一样的。