第10章　数据仓库的设计复查要目

在操作型环境中确保质量的最有效的方法之一是设计复查。通过设计复查可以检测到各种错误，并在编码之前更正这些错误。在开发生命周期的早期费点功夫找到错误，能得到很大的好处。

在操作型环境中，设计复查通常是在一个应用的物理设计完成以后进行的。操作型设计复查所围绕的中心问题的类型有以下这些：

- 事务处理性能。
- 批窗口是否适当。
- 系统可用性。
- 容量。
- 项目准备的充分性。
- 用户需要的满足程度。

如果在操作型环境中我们正确地进行了设计复查，就可以节约可观的资源，并且大大增加用户对系统的满意度。更重要的是，当设计复查正确地实施以后，一旦开始系统的编写，系统的主要代码就用不着推倒重写了。

与在操作型环境中一样，设计复查在数据仓库环境中也是适用的，并有几个附带条件。

一个附带条件是系统是在数据仓库环境中以循环重复的方式开发起来的，在这种开发方式下，需求表现为开发过程的一个部分。典型的操作型环境是在严格定义的 SDLC(系统开发生命周期)下建立的，而数据仓库环境下的系统并不是按 SDLC建立的。操作型环境和数据仓库环境下的开发过程的其他区别如下：

- 操作型环境中的开发是按一次一个应用问题进行的，数据仓库环境下的系统是按一次一个主题领域进行的。
- 在操作型环境中，有一套稳定的需求，构成操作型环境下设计和开发的基础。而数据仓库环境下，在DSS开发的开端，人们对处理需求很少有一个稳定的认识。
- 在操作型环境中，事务响应时间是主要的而且是极其重要的问题。而在数据仓库环境中，事务响应时间基本上不算是个问题。
- 在操作型环境中，来自不同系统的输入通常来自企业的外部数据源，最常见的是通过与外部代理的交互获取数据。在数据仓库环境中，数据通常来自企业内部的各个系统，而各个系统的数据是由很多种类的现有数据源集成而来的。
- 在操作型环境中，数据几乎都是当前值 (也就是说，数据在用的那一刻是准确的)。而在数据仓库环境中，数据是随时间变化的 (也就是说，数据与某个时刻相关)。

这样，在操作型环境和数据仓库环境之间就存在一些根本的区别，这些区别在进行设计复查的过程中就可以体现出来。

10.1 进行设计复查所涉及的问题

在数据仓库环境中，一旦一个主要的主题领域设计好以后，并准备加入到数据仓库环境中时，就应开始做设计复查。但并不是每建一个新的数据库就需要做设计复查，相反，当整个新的主题领域加入数据库中时，就有必要进行设计复查。

10.1.1 谁负责设计复查

设计复查时的参加者包括那些与所复查的 DSS主题领域有关的开发人员、操作人员或使用人员。

通常情况下，包括如下人员：

- 数据管理人员(DA)
- 数据库管理员(DBA)
- 程序员
- DSS分析人员
- 除DSS分析人员外的最终用户
- 操作人员
- 系统支持人员
- 管理人员

在这组人员中，最重要的参与者是最终用户和 DSS分析人员。

在同一房间与同一时间，有所有这些不同人员聚在一起，这有一个显著的好处，那就是有机会加强沟通，消除不同认识。在日常环境中，最终用户将问题告诉联络者，联络者转达给设计者，设计者又通知程序员，在此过程中很有可能造成误传和误解。而当所有这些不同类别的人员聚在一块时，就有了直接交流的机会，这对一个正在被复查的系统是非常有益的。

10.1.2 有哪些议事日程

对数据仓库环境进行复查的主题可以是任何可能会导致失败的设计、开发、项目管理或者应用问题。简言之，任何有碍成功的障碍在设计复查过程中都会涉及到。通常，一个主题越有争议，在复查期间就应更加重视它。

复查过程的基本问题将在本章后面部分中讨论。

10.1.3 结果

数据仓库设计复查能产生三种结果：

- 对问题管理的评价和对进一步行动的建议。
- 有关系统在设计中的位置以及复查时间的文档。
- 一个“行动要目”表，阐明作为复查过程的结果的特定目标和行动。

10.1.4 复查管理

复查过程由两个人领导，一个督导人和一个记录员，督导人绝对不能是要复查的项目的管理者或开发者。在有些情况下，若督导人是项目领导，从许多角度而言，复查的目标都有

可能会失败。

要进行一个成功的复查，督导人不能参与项目，这点必须强制执行，这有以下的理由：

- 作为一个局外人，督导人会用新的眼光，从外部角度观察系统。这种新鲜的眼光经常能揭示出那些与系统的设计和开发很接近的人所不能发现的重要的见识。
- 作为一个局外人，督导人能建设性地提出批评。与开发工作很密切的人员给出的批评往往具有个人观点，并可能使设计复查局限于一个非常低的水平之上。

10.1.5 典型的数据仓库设计复查

1. 复查过程中遗漏了谁？是不是有应该参加的小组被遗漏了？以下这些小组成员出席了吗？

- DA。
- DBA。
- 编程人员。
- DSS分析员。
- 最终用户。
- 操作人员。
- 系统编程人员。
- 审计人员。
- 管理人员。

各小组的正式代表是谁？

解答：不管其他任何因素，合适的人员合适地参加了设计复查对于复查的成功是至关重要的。最重要的参加者是DSS分析员或最终用户，管理人员或许参加或许不参加，随他们自己。

2. 最终用户需求都已完全预见到了吗？如果是，达到了什么程度？设计复查中最终用户代表是否同意已做好的有关需求的表述？

解答：从理论上讲，DSS环境可以在无需与最终用户交互的情况建立起来，也就是无需对最终用户的需求做出什么预测。然而，如果需要修改数据仓库环境中数据的粒度，或者在数据仓库的顶层，需要建立EIS或人工智能处理过程，那么对需求做一些预见是很有益的尝试。通常，甚至在预测了DSS需求时，最终用户的参与程度是非常低的，最终结果也是非常粗略的。进而，不应该花费大量的时间去预测用户的需求。

3. 数据仓库环境中的数据仓库的多少内容已经建立好了？

- 哪些主题？
- 有哪些细节？哪些综合？
- 按字节算、按行算、按磁道/柱面算有多少数据？
- 有多少处理量？
- 独立于被复查项目，有哪些增长模式？

解答：数据仓库环境的当前状态对被复查的开发项目有很大的影响。刚开始的开发工作应在有限的范围内进行，并且应在边试边改的基础上开展。在这个阶段，不应有很关键的处理或数据。另外，应该预计到一定量的快速反馈和重复开发。

此后的数据仓库开发工作出错的机会就会少些。

4. 已经从数据模型中找出了多少主要主题？有多少是正在实现的？有多少是已全面实现的？有多少是由正在被复查的项目来实现的？有多少是能在不远的将来得到实现的？

解答：通常，数据仓库环境一次只实现一个主题，开始少数几个主题应该被当作实验一样。到了后来，早期开发工作中所学到的教训，就能在主题的实现中应用。

5. 数据仓库环境之外是否存在重要的DSS处理(也就是数据仓库)？如果存在的话，有没有可能产生重复或冲突？对数据仓库环境外的 DSS数据和处理的迁移方案是什么？对于将不可避免地要出现的迁移，最终用户能理解吗？在什么时间范围内做迁移工作？

解答：在正常情况下，在数据仓库环境中，只有部分数据仓库，而其他部分的数据在数据仓库环境之外的话，这将是一个重大的错误。只有在一些最例外情况下，才允许存在一个“分裂”方案(其中的情况之一是分布式数据仓库环境)。

如果数据仓库环境确实有部分实际上处于数据仓库环境之外，就应该有种方案将 DSS体系中的那部分搬回到数据仓库环境中。

6. 已经找出的主要主题是否都已经划分到更低的细节级？

- 各个键码是否已标识？
- 各个属性是否已经标识？
- 键码和属性是否已组合起来？
- 不同数据分组之间的关系是否已经标识？
- 每一组的时间变化是否已经标识？

解答：需要有一个数据模型作为数据仓库的智能中心。这种数据模型在正常情况下有三个层次：一个标识实体和关系的高层模型；一个标识键码、属性和关系的中层模型；以及一个可以在此进行数据库设计的低层模型。然而，在开始建立 DSS环境时，并不是所有的数据都需要模型化到最低层，但至少高层模型应该建好。

7. 问题6中所讨论的设计是否要周期性地复查？(多长时间一次？非正式地还是正式地？)作为复查的结果，会出现什么变化？最终用户的反馈如何传递给开发者的？

解答：经常需要修改数据模型以反映企业的业务变化。通常情况下，这些变化具有逐渐增长的特点，革命式的变化是不常见的。这种变化对已有数据仓库数据和计划中的数据仓库数据所造成的影响应该做出评估。

8. 是否已确定操作型环境的记录系统？

- 每一个属性的数据源确定没有？
- 是否已经确定某一个或另外一个属性会成为数据源的条件？
- 如果某一个属性没有数据源，是否确定了它的默认值？
- 数据仓库环境中那些数据属性的属性值的公用度量确定没有？
- 数据仓库环境中那些数据属性的共同的编码结构确定没有？
- 数据仓库环境中的公共键码结构确定没有？记录系统的键码在哪些地方不符合 DSS的键码结构的条件？确定转换途径没有？

解答：数据模型建立好以后，记录系统就定好了。记录系统通常存在于操作型环境中，记录系统代表了支持数据模型的现存数据中最好的数据源。集成问题在定义记录系统时是一个非常重要的因素。

9. 从操作型记录系统中抽取数据到数据仓库环境的过程的频率确定没有？当前抽取过程如何从上次的抽取过程中识别出操作型数据的变化？

■ 通过查看时戳数据？
■ 通过改变操作型应用代码？
■ 通过查看日志文件？或是某个审计文件？
■ 通过查看某个差异文件？
■ 通过比较“前”映像和“后”映像？

解答：抽取过程的频率之所为成为一个问题，是因为刷新中所需要的资源、刷新过程的复杂性、以及数据及时刷新的需要等原因造成的。数据仓库的可用性常常与数据仓库是以什么频率刷新的有关。

从技术角度而言，最复杂的问题之一是在抽取过程判定应该扫描哪些数据。在有些情况下，需要从一个环境中传到下一个环境中的操作型数据是相当明确的。在另外一些情况下，根本就无法知道应该对哪些数据进行检查，并将其作为载入数据仓库环境的候选数据。

10. DSS环境中通常包含多少数据量？如果数据量很大的话，那么

■ 是否应指定多重粒度级？
■ 是否应该对数据进行压缩？
■ 是否应进行定期数据清理？

解答：除了抽取过程所处理的大量数据外，设计者自己需要考虑数据仓库环境中实际的数据量。对数据仓库环境中数据量的分析，直接地导致数据仓库环境中数据的粒度问题，并可能会导致多重粒度级的出现。

11. 在执行抽取过程以创建数据仓库环境的时候，哪些数据将被滤出操作型环境？

解答：所有的操作型数据都传送到 DSS环境中是很少见的。几乎每一操作型环境都包含有只与操作型环境相关的数据。这些数据不应该进入到数据仓库环境中去。

12. 采用什么软件来给数据仓库环境提供数据？

■ 这种软件已被彻底卸出了吗？
■ 存在或可能会存在什么瓶颈？
■ 接口是单向的还是双向的？
■ 需要什么技术支持？
■ 该软件能传送多大容量的数据？
■ 需要对软件做什么样的监控？
■ 对软件需要周期性地做什么变更？
■ 这种变更会伴有什么中断？
■ 安装这种软件需要多少时间？
■ 谁负责这种软件？
■ 这种软件何时能充分投入使用？

解答：数据仓库环境能够处理大量的不同类型的软件接口。然而，不应低估中断时间和基础构造所需时间量。 DSS体系结构设计者不能想当然地认为，将数据仓库环境与其他环境连接起来必定是直截了当的与容易的。

13. 在数据仓库环境外对DSS部门的和个体的处理传送数据需要什么软件 /接口？

■ 是否已经彻底地测试接口？
■ 可能存在什么瓶颈？
■ 接口是单向的还是双向的？
■ 需要什么技术支持？
■ 经过接口的预期数据流量是多大？
■ 接口需要什么样的监控？
■ 将对接口做哪些变更？
■ 对接口做变更后可能会产生什么中断？
■ 安装接口需要多长时间？
■ 谁负责这个接口？
■ 接口什么时候才能投入全面的应用？

14. 在数据仓库环境中将使用什么样的数据物理组织？数据能直接存取吗？能进行顺序存取？能简单且廉价地创建索引吗？

解答：设计者应该复查数据仓库环境的物理配置以确保足够的可用空间，并保证一旦数据到了这种环境中，就能以一种应答的方式去操纵它。

15. 数据仓库环境建好以后，往其中增加更多的存储空间容易不容易？在数据仓库环境中重新组织数据难不难？

解答：所有数据仓库都不是静态的，没有数据仓库能在设计的初始阶段就能完全地进行规格说明。在数据仓库环境的整个生命期，做一些设计上的修改是完全正常的。建立一个数据仓库环境时，如果在中间过程要么不能做任何修改，要么很难进行一些修改，则这个设计必定是一个失败的设计。

16. 数据仓库环境中的数据需要经常重新组织 (就是说增加、减少或扩大列，或修改键码等)的可能性有多大？这些重新组织工作对数据仓库正在进行的处理有什么影响？

解答：给定数据仓库环境中所能找到的大量数据，重新组织这些数据并不是件容易的事。另外，有了存档数据后，过了一定时间以后重新组织数据在逻辑上几乎是不可能的了。

17. 期望的数据仓库环境的性能水平如何？是否正式或非正式地拟定了 DSS服务级的协议？

解答：除非正式拟定了一个 DSS服务级协议，否则不可能度量出性能指标是否已经达到要求。这个DSS服务级协议应该涵盖DSS性能水平及停机时间。典型的 DSS服务级协议会阐明以下内容：

■ 每个数据单元在高峰时刻的平均性能。
■ 每个数据单元在非高峰时刻的平均性能。
■ 每个数据单元在高峰时刻的最坏性能。
■ 每个数据单元在非高峰时刻的最坏性能。
■ 系统可用性标准。

DSS环境的一个难题是性能度量，不像操作型环境，可以用绝对标准来度量它的性能，DSS处理性能度量与下列内容有关：

■ 单个请求代表多少处理。
■ 当前正在并发进行的处理有多少。

■ 在执行时刻有多少用户在系统中。

18. 期望的可用性水平有多高？是否已对数据仓库环境正式或非正式地拟定了可用性协议。

解答：(见问题17的解答。)

19. 数据仓库环境中的数据是如何索引或存取的？

■ 任何表是否会有超过四个的索引？

■ 任何表是不是会被哈希存储？

■ 任何表是否仅有主键索引？

■ 为了维护索引需要些什么开销？

■ 为了最初装入索引需要什么开销？

■ 索引的使用频率如何？

■ 为了服务于更广泛的应用，索引是否能或是否应该被改变？

解答：数据仓库环境中的数据要求能被高效而灵活地存取。不幸的是，数据仓库处理所具有的启发式特性使得对索引的需求具有不可预测性。造成的结果是，我们不能想当然地去存取数据仓库环境中数据。通常，采用多层方法去管理对数据仓库的数据的存取是最理想的：

■ 哈希键或主键应该满足多数存取。

■ 二级索引应满足其他大多数存取模式。

■ 临时索引应该满足不常见的存取。

■ 对数据仓库数据的一个子集的抽取和顺序索引，应该满足不频繁或者生命期内只出现一次的数据存取操作。

在任何情况下，数据仓库环境中的数据不应该按太大的分区存放，以免使它们无法自由地进行索引。

20. 预期的数据仓库环境中处理量的大小如何？高峰期如何？日平均量的情况如何呢？峰值处理率又是如何？

解答：不但需要预计数据仓库环境中的数据量，而且也应该预计到数据处理量。

21. 数据仓库环境中的数据应有怎么样的粒度级？

■ 高粒度级？

■ 低粒度级？

■ 多重粒度级？

■ 要不要做滚动式持续汇总？

■ 是否有真实档案级的数据？

■ 是否有真实样本级的数据？

解答：显然，在数据仓库环境中，最重要的设计问题是数据的粒度和采用多重粒度级的可能性。简言之，如果数据仓库环境的粒度级已经正确地取好了，那么所有其他问题就变得简单明了了；如果数据仓库环境中的粒度级没有正确地设计好，那么所有其他问题将会变得复杂而沉重。

22. 数据仓库环境中有什么数据清除标准？数据是真的被清除，还是压缩好放到其他地方？有什么法定需求？有什么核查要求？

解答：即使DSS环境中的数据是存好档的，而且必然地具有很低的存取可能性，这些数

据还是具有某种存取可能性(否则它就不应被存储)。当存取的可能性减低到0(或接近0)时，数据就应该被清除了。如果数据量是数据仓库环境中一个极严重的问题的话，将不再有用的数据清除出去就成为数据仓库环境的比较重要的方面之一。

23. 为满足以下两项需要，各自需要的总数据处理能力是多少？

■ 为了最初实现？

■ 为了成熟时期的数据仓库环境？

解答：假设无法将所需要的能力需求规划到最后一位，对需要的能力至少估计一下也是有益的，以免造成实际需要和可用能力之间的不匹配。

24. 在数据仓库环境中，能够识别出主题领域之间的哪些关系？这些关系的实现：

■ 能不能使外键码得到不断的刷新？

■ 能不能利用人工关系？

建立和维护数据仓库环境中的关系时需要哪些开销？

解答：数据仓库设计者要做的最重要的设计决策之一，就是该如何实现数据仓库环境中的数据之间的关系。在数据仓库中，数据关系几乎不可能以与数据原来在操作型环境中的关系相同的方式来实现。

25. 数据仓库环境内部的各个数据结构是否利用了以下各项技术：

■ 数据阵列？

■ 选择性的数据冗余？

■ 数据表的合并？

■ 导出数据的通用单元的创建？

解答：尽管在数据仓库环境中操作型性能并不算是个什么问题，但性能仍然毕竟还算是一个问题。如果前面所列的这些设计技术能够减少 IO总量，设计者就应考虑采用这些技术。这些技术是典型的物理反正规化技术。因为在数据仓库环境中并不修改数据，所以，对于哪些事能做，哪些事不能做，并没有什么限制。

决定该采用哪些技术的因素包括如下几条：

■ 数据出现的可预测性。

■ 数据存取模式的可预测性。

■ 收集数据人工关系的必要性。

26. 数据仓库中的数据恢复需要多长时间？执行一次完整的数据仓库数据库恢复工作的操作准备好没有？是部分性地恢复？这些操作会不会周期性地执行恢复工作，以使它们在需要恢复时就已经准备好了？准备的程度是在下面的哪一级体现的：

■ 系统支持？

■ 应用编程？

■ DBA？

■ DA？

对于每类可能出现的问题，问题由谁负责是否已经明确？

解答：就像在操作型系统中一样，设计者必须为在恢复期间出现的中断做好准备工作。恢复的频率、对系统进行备份所需的时间、以及在中断期间可能会产生的多米诺骨牌效应，都需要认真加以考虑。

使用说明书是否已经准备、测试和编写好？这些使用说明书是否得到经常的更新？

27. 在什么级别对重新组织/重新构造进行准备：

■ 各种操作？

■ 系统支持？

■ 应用编程？

■ DBA？

■ DA？

是否编写了说明书和建立了过程，并做了测试？是及时更新的吗？它们能一直得到及时更新吗？

解答：(见问题26的解答。)

28. 在什么级别对数据库表的装载进行准备：

■ 各种操作？

■ 系统支持？

■ 应用编程？

■ DBA？

■ DA？

是否编写了说明书和建立了过程，并做了测试？是及时更新的吗？它们能一直得到及时更新吗？

解答：装载所需的时间和资源可能是相当大的，应该谨慎地做这个估计，这个估计需要在开发生命周期的早期进行。

29. 在什么级别对数据库索引装载进行准备：

■ 各种操作？

■ 系统支持？

■ 应用编程？

■ DBA？

■ DA？

解答：(见问题28的解答。)

30. 如果对数据仓库环境中的某项数据的精确性产生过争议的话，该如何解决这种争议？数据仓库环境中的每个数据单元的所有权 (或至少数据出处)定好没有？如果有需要的话，能不能建立数据的所有权？谁负责处理所有权问题？有关所有权问题，谁拥有最终的决定权？

解答：在数据仓库环境中，数据的所有权或管理权是数据仓库环境成功与否的基本因素。有时数据库的内容不可避免地会出现问题。对这种可能性，设计者应该提前计划好。

31. 一旦数据放到数据仓库环境后，该如何修改数据？修改的频率如何？应该对修改进行监控吗？如果存在一种定期出现修改的模式，在数据源层次上 (也就是操作环境下)的修改该如何进行？

解答：有时偶尔地或不定期地，需对数据仓库环境下的数据做一些修改。如果这些修改具有一定的模式，那么DSS分析人员就需要调查一下操作型系统中是否存在什么问题。

32. 公共汇总数据是否要与正常的原始 DSS数据分开存放？有多少公共汇总数据？是否应该存储用于创建公共汇总数据所需要的算法？

解答：即使数据仓库环境包含原始数据，在数据仓库环境中存在公共汇总数据也是很正常的。设计者应该准备一些逻辑空间来存放这些数据。

33. 需对数据仓库环境中的数据库采用哪些安全措施？如何实施安全措施？

解答：数据访问成为一个问题，特别是在形势很明显地表明细节数据变成汇总数据或聚集数据时。设计者应该预计到安全需求，并为之准备好数据仓库环境。

34. 有什么审查需求？怎样才能达到审查需求？

解答：通常，系统审计可以在数据仓库层做，但这几乎总是个错误。相反，在记录系统层做细节记录的审查是最好的。

35. 是否采用数据压缩？是否考虑到压缩 /解压缩数据的开销？有什么开销？通过 DASD压缩/解压缩数据能节省什么？

解答：一方面，对数据进行压缩或编码能节省大量的空间。而另一方面，数据压缩和编码都需要CPU时间，因为访问数据时需要解压缩或解码。设计者应该对这些问题做充分的研究，并在设计中做一个审慎的折衷方案。

36. 需要对数据进行编码吗？考虑到编码 /解码的开销没有？实际上有哪些开销？

解答：(请参看问题35的解答。)

37. 数据仓库环境中应该存储元数据吗？

解答：作为一条法则，元数据需要与任何档案数据一起存储。分析人员使用档案数据解决问题的时候，如果他不知道他所分析的数据域的内容的含义时，绝对是很难办的。在将数据存档时，把数据语义与其存放在一起，就可以缓解前面的问题。随着时间的过去，数据仓库环境中的数据的内容和结构发生些变化是绝对正常的。设计者必须确保系统能始终跟踪随着时间变化的数据定义。

38. 参照表是否应该存放在数据仓库环境中？

解答：(请参看问题37的解答。)

39. 数据仓库环境中需要维护哪些目录 /字典？谁负责维护？如何保持更新？是为谁准备的？

解答：不但随时跟踪数据定义是一个问题，而且跟踪数据仓库环境中的当前数据变化也很重要。

40. 数据仓库环境中允许进行数据更新 (与装载和访问数据相反)吗？(为什么？更新多少？在什么情况下？是不是仅仅限于异常的情况下？)

解答：在数据仓库环境下，在正常情况下如果任何更新操作都可以做的话，设计者就应该探讨一下其中的原因了。唯一的一种会出现的更新，应该发生在出异常情况时，并且只能对一小部分数据进行更新。除此以外的任何更新都会严重地危及数据仓库环境的功效。

在做更新的时候(如果确实要做的话)，它们应在一个专用窗口中执行，应该在系统中没有其他处理，且在处理器有空闲时间的时候进行。

41. 从操作型环境中取数据到数据仓库环境时，有什么样的时间迟延？这个时间迟延会不会少于24个小时？如果会的话，为什么？在什么情况下会这样？数据从操作型环境传送到数据仓库环境的过程是一个“ 推送 ”过程还是一个“ 拖拉 ”过程？

解答：从策略上讲，任何少于 24小时的时延都是有疑问的。通常，如果需要有少于 24小时的时延，就表明开发者在将操作型需求构建到数据仓库中。流向数据仓库的数据流应该总

是一个拖拉过程，也就是说数据只是在需要的时候被拖拉进数据仓库，而不是在系统可用的时候推送到数据仓库环境中。

42. 应该做哪些有关数据仓库的活动日志？谁将会访问这些日志？

解答：大部分DSS处理不需要日志。如果需要做大量的日志，通常情况下表明人们对当前数据仓库环境中正在发生的处理的类型缺乏足够的理解。

43. 除了公共汇总数据以外，还有其他的数据从部门层或个体层流向数据仓库环境中吗？如果有，就描述这些数据。

解答：只有在很少的情况下，公共汇总数据来自于其他数据源，而不是来自部门层或个体层处理。如果有许多公共汇总数据是来自于其他数据源的话，则分析者就应该找一下原因了。

44. 有什么样的外部数据(即不是由一个企业内部的数据源和系统产生的数据)进入数据仓库环境？这些数据要不要加上特别标记？它们的数据源要与数据存储在一起吗？这些外部数据进入系统的频率如何？有多少外部数据进入？是否需要非结构化的格式？如果发现外部数据不准确会出现什么情况？

解答：除了公司的操作型系统外，即使是允许一些外部数据源合理地存在，但当有许多数据从外部进入时，分析人员就应该找一下原因。在外部数据的内容和可用性规则方面，它所能具有的灵活性都不可避免地要少得多，虽然外部数据也是一个不可忽视的重要的数据源。

45. 存在什么样的环境工具能帮助部门和个体用户查找数据仓库环境中的数据？

解答：数据仓库的一个主要的特点就是易于数据存取。而数据的可存取性问题第一步就是这些数据的初始位置。

46. 有人尝试将操作型处理和DSS处理同时混在一台机器上吗？(为什么？有多少处理？有多少数据？)

解答：出于许多方面的考虑，同时将操作型处理和 DSS处理混合在同一台机器上是没有什么意义的。只有在数据量和处理量都很小的时候，才有可能进行混合。但这些条件不是使数据仓库环境达到最大成本效益的条件。 (请参阅《Data Architecture —— The Information Paradigm》一书， Wiley, 1992年出版，该书对此问题有更深的探讨。)

47. 有多少数据会从数据仓库层流回到操作层？按照什么频率？数据量多大？对响应时间有什么限制？回流的数据是汇总性数据还是单个数据单元？

解答：通常，数据从操作型处理流向仓库层处理，再流向部门层处理，再流向个体层处理。但也有一些值得注意的例外情况，只要没有太多的数据回流，并且只要回流是以有序的方式执行的，通常情况下就不会出现问题。然而，如果回流时所涉及的数据量很大，那么就出示红牌了。

48. 会出现多少针对数据仓库环境的重复性处理？导出数据的预先计算和存储能节省处理时间吗？

解答：对于数据仓库环境而言，具有一定量的重复性处理绝对是正常的。如果只有重复性处理要做，或相反根本没有重复性处理要做，设计者就都应找一下原因。

49. 主要主题该如何划分？(按年？按地域？按功能单元？按生产线？)对数据的分割的精细程度如何？

解答：对于数据仓库环境所固有的数据量以及数据的不可预测的用途，必须强制性地把

数据仓库的数据在物理上分割为小单元，以便能对它们进行独立的管理。我们要面对的设计问题不是是否应该进行分割，相反，问题是分割该如何完成。一般地说，分割是在应用层而不是在系统层进行的。

对分割策略进行复查的时候，应注意以下问题：

- 当前数据量。
- 未来数据量。
- 数据的当前用途。
- 数据的未来用途。
- 仓库中其他数据的分割。
- 其他数据的用途。
- 数据结构的易变性。

50. 要创建稀疏索引吗？它们有用吗？

解答：在合适的地方创建的稀疏索引能够节省大量的处理。同时，稀疏索引创建和维护需要相当数量的开销。数据仓库环境的设计者应考虑好它们的使用。

51. 要创建什么样的临时索引？它们要保留多长时间？它们会有多大？

解答：(参看问题50的解答，它也适合于临时索引。)

52. 部门层和个体层会有什么文档？有关数据仓库环境和部门环境之间的接口、部门环境和个体环境之间的接口、数据仓库环境和个体环境之间的接口有些什么文档？

解答：在部门和个体环境下，它们的处理的特点是形式自由，也就不太可能会有太多的可用文档。然而，有关各个环境之间的关系的文档对数据的一致性是很重要的。

53. 用户要为部门层处理、个体层处理付费吗？谁承担数据仓库处理的费用？

解答：有一点是很重要的，就是用户必须有自己的预算，并且必须为所使用的资源付费。可以预见，处理如果变成免费的，肯定产生大量的对资源的滥用。收费系统能够在资源的使用方面给用户灌输一种责任感。

54. 如果数据仓库环境必须是分布式的，仓库的公用部分确定没有？它们是如何管理的？

解答：在分布式数据仓库环境中，一些数据必然地会受到严密的控制。这些数据需要由设计者标识出来而由元数据控制存放位置。

10.2 小结

设计复查是一个重要的质量保证环节，它可以大大地提高用户的满意程度，并减少开发和维护费用。在建立数据仓库之前，彻底地对数据仓库环境的许多方面进行复查，是一种很有效的实际工作。