数据仓库最佳实践手册

数据仓库最佳实践手册

数据仓库的职责就是把数据整合到一起，然后将它们转换成可供分析的信息。在评估与创建数据仓库的时候，我们会遇到各种各样的问题，而伴随着 IT 技术的发展，新的产品与概念也进入了我们的视野。为此，在本次的 TechTarget 数据库技术手册中，我们就将为您带来一些有关数据仓库的最佳实践，其中包括数据仓库的评估、数据仓库管理技巧、数据仓库实施案例等内容，希望能对 DBA 的工作带来一定的帮助。

评估与选型

在实施一个数据仓库项目之前，我们需要对不同的数据仓库产品进行评估与选型，以便找到最适合企业的产品。在本部分中，我们就将对数据仓库的评估与选型提供一些建议与技巧。

- 评估数据仓库软件的五个步骤
- 如何选择数据仓库与数据集市
- 合理选择数据仓库工作负载管理工具
- 工作负载管理对于数据仓库 ROI 很关键

数据仓库管理与策略

在本部分中，我们将了解一些关于日常数据仓库管理的一些最佳实践，其中包括理解数据仓库架构的区别、不同类型的数据仓库、数据仓库性能优化以及在大数据背景下，企业如何调整数据仓库以及 BI 的战略。

- 逻辑数据仓库和物理数据仓库的区别
- 数据仓库性能优化的六个技巧

- 如何规划一个高效的 BI 数据仓库项目
- 如何应对数据仓库架构挑战
- 考虑两种数据仓库架构共存的可行性
- 大数据背景下的数据仓库最佳实践
- 大数据蔓延 企业需重新定位数据仓库策略

数据仓库集成设备（Data Warehouse Appliance）

最近几年中，数据仓库设备（Data Warehouse Appliance）在细分市场中得到了长足又快速的发展，越来越多的厂商开始将他们的软件与硬件集成起来，并与商业服务器硬件提供商结成合作伙伴关系，于是用户开始逐渐适应了这种即插即用的部署方式。

- 数据仓库设备趋势：聚焦 BI，细分市场
- 数据仓库集成设备选型指导
- 部署数据仓库系统需要避免的三大问题
- 分析数据仓库设备优势与局限性

数据仓库实施案例分享

在本部分中，我们将与您一同分享几个数据仓库项目的实施案例，其中涉及到医疗行业以及大型娱乐行业，希望能够通过具体的案例，DBA 能从中学习到一些有用的实践经验。

- 医疗行业面对的数据仓库挑战
- 迪斯尼乐园诠释数据仓库最佳实践

评估数据仓库软件的五个步骤

1、确定你需要的数据仓库软件。核心工具包括：数据库;ETL;商业智能(BI)。我还建议选择一款数据建模工具。有些公司可能还需要数据清理软件，但是要注意的是大部分数据质量都能在你写的 ETL 代码中体现。

2、检查已有软件，或者检查你的数据库软件标准。你是否在应用中使用了其它的工具，比如一些 BI 工具？要提醒一句，一定要确保你对已有软件的使用得当，至少你要有能力去操作和管理，用户也能使用得顺手。

3、制定一个名单。有几种方法可以制定名单：看看业内分析报告(推荐 Forrester 或 Gartner 的);查看已绑定的软件，比如与数据库或 BI 工具绑定的 ETL 工具等;向同行咨询;经常读一读相关专栏作家的文章。

4、建立你的评估标准。如果你是 TWDI 的会员，那么你就可以找到制定标准所需要的清单。尽量避免制定功能冗长的清单，因此这样缺乏针对性。标准应该紧密联系自身的需求，而不是去追求功能最强大的软件。你选择的工具应该是成熟技术的产物，它能满足大多数人的标准，因此每个软件类别都应该有不止一种的选择。

5、开始对数据仓库软件进行评估。评估过程尽量简短，当然如果你能将它组织成 POC 模型并在项目中使用那就更好了。

(作者：Rick Sherman 译者：孙瑞 来源：TT 中国)

原文标题：评估数据仓库软件的五个步骤

链接：http://www.searchdatabase.com.cn/showcontent_23815.htm

如何选择数据仓库与数据集市

数据仓库的职责就是把数据整合到一起，然后将它们转换成可供分析的信息。在创建数据仓库的时候，我们面临着以下四个问题：

- 数据是不兼容格式的
- 数据质量比较低
- 散布在不同的系统之中
- 数据的结构不利于分析

想要解决这些问题，我们可以使用数据抽取、转换和加载（ETL）工具将数据移动到数据仓库中。抽取和转换处理可以解决不兼容和质量问题。另外，将数据全部存放在单一的数据仓库中，这就解决了数据过于分散的问题。

我们可以构建数据仓库来进行分析（就像 Taradata）,这种情况我们不需要再创建数据集市了。然而，数据集市通常包含数据仓库中的数据子集，它被组织起来提供特定的分析视图并交付给企业中的工作组或个人。例如，数据集市中的市场数据可以提供给广告部，员工数据可以提供给人力资源部。

因此，在有没有数据仓库的前提下，你都可以创建数据集市。你可以把特定系统中的数据，比如金融系统，放到数据集市中，然后将数据结构化以便分析使用。一旦弄清楚数据仓库和数据集市都是设计用来干什么的，相信你可以很容易地进行选择。

（作者：Mark Whitehorn 译者：孙瑞 来源：TT 中国）

原文标题：如何选择数据仓库与数据集市

链接：http://www.searchdatabase.com.cn/showcontent_23351.htm

合理选择数据仓库工作负载管理工具

从来没有人说过管理数据仓库是一项容易的工作。

从实时加载数据，生成多种日报表到返回结果供深入研究的专业分析查询，今天的数据仓库被要求处理更复杂的任务，要比以前处理更多用户的更多数据。

但是我们只有这么多磁盘空间和 CPU 可用来分配，也就是说瓶颈仍会出现，性能会慢下来并最终停止运转。

Chris Stewart 是圣地亚哥 Premier 公司的高级架构师，他说：“那是用户在他们的桌面能感触到的事物之一，很明显这是非常重要的。”Stewart 说 Premier 公司一直在努力保持能跟得上加诸于数据仓库日益增长的需求。

Stewart 说：“人们期望在任何时候都能(从数据仓库中)得到非常快速的响应。”

那么，像 Stewart 这类希望维持企业数据仓库运转的数据库管理员和架构师们，如何才能照顾到方方面面?那就是工作负载管理该出现的地方，他现在的出现比以往任何时候都重要。

Forrester 的分析师 Jim Kobielus 说：“目前，存在许多不同的工作负载需要被企业数据仓库处理。工作负载管理工具是 DBA 们每日每夜都要用的，这样才能保持一切运转良好。”

工作负载管理涉及一系列内建于数据仓库的功能，这些功能支持管理员们优先处理特定类型的任务(会考虑是谁发起的请求，以及该任务的时间敏感性如何)，然后给任务分配可用于完成这些请求的有限的计算机处理能力。

工作负载管理实际上是被 DBA 们通过监视工具来执行的，这些工具支持他们跟踪数据仓库如何执行任务，并依据跟踪信息做相应调整。例如：如果 CEO 的查询由于消耗资源巨大而被挂起了，DBA 会推迟其他任务，让 CEO 的查询请求先通过。

Kobielus 说：“我们可以把工作负载管理工具看作是一个交通警察。”没有它，一切都会陷入混乱。

十年前，大部分数据仓库只被预期产生少量报表，只是每天或者甚至每周以批量形式加载数据，有时也处理少部分用户发起的偶尔性的临时查询。

现在，除了更常见的批量加载工作之外，更先进的数据集成工具已经使得用数据仓库近乎实时地整合数据成为可能。由于有更多易用的前端商业智能工具，更多的用户可以访问数据仓库。同时，进出数据仓库的数据量在冲天猛涨，而且并没有显示出来减缓的迹象。

Gartner 的分析师 Donald Feinberg 认为，这样一来工作负载管理就变得非常重要了。实际上，在他最新针对数据仓库的“魔力象限”报告中，Gartner 公司声称，混合性的工作负载性能确实是数据仓库中唯一最重要的性能问题，“将来也会是这样。”

Feinberg 说，对于客户来说，好消息是几乎所有数据仓库供应商都具备一些负载管理能力，尽管其中一些比另外一些更优秀。他选出 Teradata 公司作为数据仓库供应商中具备最先进，最全面负载管理工具的供应商，尽管他表示 Oracle 公司，IBM 公司和微软公司的产品也差的不太远。

然而，具备最先进工作负载管理能力的供应商也是最大的企业客户，这一点并不奇怪。工作负载管理在大型的集中式企业数据仓库环境中是最关键的，这种数据仓库会被整个大公司的用户访问。

在这种情况下，数据仓库被期望处理更多流量和查询，比多个部门在更分布式的环境中分设数据仓库和数据集市需要处理的负载量要多得多。

Randy Lea 是 Teradata 公司产品和服务市场的负责人，她说：“你的数据越集中，负载管理对你的环境就越关键。

按照 Lea 的观点，Teradata 公司对负载管理采取了双管齐下的做法。首先，该供应商的 Teradata 活动系统管理(TASM)应用程序支持客户实现基于预定义业务规则的负载优先次序管理。

例如：许多 Teradata 公司的客户给公司高层主管的查询比其他员工的查询分配更高的优先级。其他人在特定的时间给特定的部门分配更多的带宽，这取决于使用模式。例如：如果市场部门在上午八点到十点之间要做大量复杂查询，我们就可以给那两个小时里分配更高的优先级。

另外，数据仓库工作可以基于访问数据量大小和复杂度分配优先级。例如，需要在庞大的数据集上运行的更大的查询可以被设置避开使用高峰，这样可以避免使整个系统性能变慢。

Lea 说：“不是每一个查询都是一样重要的。”

Teradata 活动系统管理还支持 DBA 们快速对优先级计划做出变更，以便在出现瓶颈时可以迅速调整优先级策略。

但是，Kobielus 说，即便是更小的数据仓库供应商也提供工作负载管理功能。Aster Data 公司就是一个例子，该公司“有特殊强大的工作负载管理功能来管理 MapReduce 任务”。

Phil Francisco 是一家数据仓库设备供应商的副总裁，该公司总部位于美国马萨诸塞州马尔堡。按照他的观点，Netezza 公司还提供完整功能的负载管理功能。

除了预先设置和邻近备用设备工作负载优化的能力，Netezza 公司还给客户提供对 CPU 执行小型分割分配的机会，这可以被用来快速处理他们遇上的简短的临时查询。

Francisco 解释道，正如 Netezza 公司给这种功能起的名字“小查询偏爱”，它“允许更小的查询”快速进出系统，即便我可能在机器上还运行着其他东西。这有点像食品杂货店购买货物较少的快速结账通道。

（作者：Jeff Kelly 译者：冯昀晖 来源：TT 中国）

原文标题：合理选择数据仓库工作负载管理工具

链接：http://www.searchdatabase.com.cn/showcontent_33778.htm

工作负载管理对于数据仓库 ROI 很关键

虽然最近几年工作负载管理(WLM)的重要性在逐渐增长，但不是所有人都认为数据仓库供应商的工作负载管理能力能跟得上增长步伐。

Datelligence GmbH 公司的咨询顾问 Silvio Schurig 在一次电子邮件采访中表示："工作负载管理是一种综合型的话题，我认为(只靠)工具不能完成出色的工作。"

Schurig 在部署数据仓库方面有丰富的经验，实践证明他在以前用过的工作负载管理工具多少有点不太灵活。他希望看到数据仓库供应商们把他们的工作负载管理工具设计的更简单，需要的编码更少。

Schurig 说："没有哪一款工具能够以可伸缩的方式(不需要，或者在一定程度上不需要大量编程和脚本编写工作，只通过配置参数实现)管理和处理工作负载。起码我没遇到过。"

然而，像 Netezza 公司这样的供应商不同意这种说法。一家供应商的市场副总裁 Phil Francisco 认为 Netezza 公司的数据仓库应用包含有工作负载管理功能，DBA 们可以通过直观的图形用户界面与之进行交互。他说，DBA 们还可以在 10 分钟之内就设置好业务规则，定义用户组。

Merv Adrian 是一名数据仓库分析师，他为 IT 市场策略提供咨询。按照他的意见，一些更小的，更新出现的数据仓库供应商确实缺乏先进的工作负载管理功能。他们依靠原始的速度来赢得客户。

Adrian 说："新兴的分析数据库管理系统供应商通常通过概念性验证来证明自己的实力，即关注于运行一两项任务并与现存系统做同样的任务进行比较。他们通常会灵巧地取胜，但是这未必能证明工作负载管理的能力。"

当然，在一定程度上，在考虑工作负载管理时对于新生的数据仓库供应商不能度量是可以理解的。Adrian 说：“他们(新生数据库供应商)还没有足够长的时间来积累丰富的，复杂的并发工作流。”

但是，那并不意味着他们最终将不必开发健壮的工作负载管理功能，如果他们想在炙手可热的数据仓库市场上竞争的话。

Adrian 说：“即便是更小的数据仓库，也经常会遇到大量并发工作，比如：在临时查询提交时运行大型报表，还有备份任务。当这些任务开始堆积时，情况就不一样了。”

Forrester 的一名分析师 Jim Kobielus 认为，改进数据仓库性能是实际性的目标，但是选择最理想的工作负载管理工具也是使你公司在数据仓库投资上花的钱物有所值的关键。

Kobielus 解释道，例如：工作负载管理工具越简单，越有效，运行数据仓库需要的 DBA 就越少。更少的 DBA 意味着更低的工资成本和可以更快地从你的数据仓库获得投资回报率(ROI)。

他说：“为保证一切正常运行，工作负载管理是 DBA 们每天必须进行的工作。那就是从你的数据仓库获得短期回报背后的秘密。”

Randy Lea 是 Teradata 公司产品和服务市场的负责人，他也同意这一点。他说，工作负载管理工具可以帮助你从数据仓库中获得最大的性能，而且你从其中挖掘出的性能越多，你需要花钱升级的时间就推迟的越久。

Lea 说，大部分 Teradata 公司的客户都能利用上他们数据仓库能力的 85%到 100%，这要感谢供应商的工作负载管理功能。他说：“这给我们的客户带来了巨大的益处。”

Gartner 公司分析师 Donald Feinberg 是这么说的：“这不只意味着你可以以最快的速度查询，以最快的速度加载数据，而且你可以以最少的投资得到需要的工作负载。”

尽管工作负载管理在技术层面是非常简单的，但是当真正要决定工作负载管理工具所遵从的业务规则时，就不那么简单了。

谁的查询在任何既定条件下都是优先的?销售部门该分配多少 CPU?什么时间会运行大批量任务?所有这些问题都需要在应用工作负载管理工具参数和规则之前明确。

Lea 说：“关于如何最优化地利用你环境的资源确实是一项业务决策。你带入系统的用户越多，你面临的冲突就越多。”

他补充道：“这会变的有点政治性。”

Netezza 公司的 Francisco 说，考虑选择数据仓库的公司还应该理解，工作负载管理工具不是一项一次就能完成的任务，不是在部署开始一次性就能设置好的。

他说，更多的用户和更多任务不可避免地会在数据仓库中堆积，需要对工作负载管理进行持续的监视和调整。

Francisco 说：“这是一项永远没有结束的任务。你总是可以对它做出改进。”

（作者：Jeff Kelly 译者：冯昀晖 来源：TT 中国）

原文标题：工作负载管理对于数据仓库 ROI 很关键

链接：http://www.searchdatabase.com.cn/showcontent_35758.htm

逻辑数据仓库和物理数据仓库的区别

典型的数据库设计可以分为以下各三个不同的层次：

- 用户定义
- 逻辑设计
- 物理设计

这种划分的做法在数据库开发很早的时期就出现了。这三个层次第一次出现是在 1975 年 ANSI/SPARC 数据库管理系统研究组发表的临时文件上描述的。

毫无疑问，记住 ANSI 代表美国国家标准协会和 SPARC 代表标准计划与需求委员会并不重要。该委员会认识到数据库设计的基本问题是缺乏沟通。

需要数据库的用户对于他们想要的东西在头脑中一般都有一个模型。

用户不会去考虑正式意义上的数据库，而是趋向于考虑他们希望显示在屏幕上的有关信息，这些信息是他们完成工作所必须的。“我想能输入我所有待售商品的明细”。他们还会从他们想要的功能出发进行考虑。“我还想能管理客户发给我的订单”。

然而，数据库设计者们（DBDs）是从本质上考虑数据库结构。关系数据库设计者们容易考虑表、列（字段）、行（记录）、主键、完整性约束、簇索引和非簇索引。

当这样两种个人（客户和数据库设计者）谈论数据库时，问题就出来了。他们对答如流，交流的严丝合缝，但说的却不是一回事。下面的对话（纯属虚构）就暴露了这样的问题。

客户：“你好，我们需要一个数据库，用来存储我们的房地产业务信息。”

数据库设计者：“好的，你打算使用什么类型的表（tables）?”

客户：“噢，不，不是房子的内容，只是财产本身。”

（数据库设计者讲的“table”是指数据库中的表，而客户理解为房子里的桌子。）

数据库设计者：“你的表中需要字段（field）吗?”

客户：“不，不是那块地上的所有房子。但是新系统必须能告诉我们哪些房子是在财产目录（index）中的。”

（数据库设计者讲的“field”是指数据库“字段”，而客户理解的意思是“土地”。）

数据库设计者：“采用聚簇索引还是非聚簇索引?”

（客户讲的“Index”指的是“目录”，而数据库设计者理解的是数据库中的‘索引’）

逻辑模型概述

逻辑模型专注于把用户对数据库的视图正式化，把它从相对非结构化的状态转化为用户需求的明确描述。一旦逻辑模型形成了，把它映射为数据库设计人员期望产生的物理模型就相对容易了。逻辑模型可以以多种方式构建，但是最通用的方法之一叫做 ER（实体关系）建模。实体关系（ER）模型的名称来源于它确实记录了用户需求中可以确认的实体和这些尸体之间存在的关系。

在称之为业务需求分析的过程中，业务分析师（BA）与用户交谈并理解用户头脑中存在的用户模型。在用户的配合下，它被正式化为实体关系模型，实际上就形成了逻辑模型。

很重要的一点是，要认识到逻辑模型是完全基于用户需求的。逻辑模型中没有来自数据库设计者的输入。实际上，在这个阶段，不仅不必要决定数据库将运行于哪种数据库引擎，而且也没必要选择哪种数据库模型（关系模型，层次模型等）。

一旦逻辑建模完成了，就可以移交给数据库设计者。在这个时候，要决定的一件事是选择哪种数据库模型，或者（更普遍的做法是）将使用哪种数据库引擎。在逻辑模型中，数据库设计者获得了对于业务需求易于理解的，正式的描述（实体，关系等。），这些描述可以相对容易地映射到他们理解的世界（表，表关系等）。数据库设计者还会增加大量用户不关心的细节（比如，数据类型，主键，索引等。）。逻辑模型逐渐转化成了数据库设计者们一开始就想要的物理模型。

（作者：Mark Whitehorn 译者：冯昀晖 来源：TT 中国）

原文标题：逻辑数据仓库和物理数据仓库的区别

链接：http://www.searchdatabase.com.cn/showcontent_28340.htm

数据仓库性能优化的六个技巧

1.在考虑性能优化之前，需要先找出现存的瓶颈。如果你的查询性能 CPU 负载比较高，那么买更快的磁盘就完全是浪费钱。

2.了解你的数据库引擎。当用户还不了解自己数据库输入和输出的时候，往往就要开始危及性能了。例如，我见过用户使用 SQL 的 Insert 语句加载数据，而不是用更高效的批量加载工具。

我也见过用户用 delete * 来删空一个表。这与 Drop 和 Create 相比是非常慢的，甚至比 Truncate 还要慢很多。当然，也许你的数据库软件并不支持 Truncate——这就是为什么你需要了解它是如何工作的。如果你还没有一个好的 DBA，也许单单从性能优化的角度看就值得聘用一个。

3.就查询而言，考虑将数据做成 MOLAP 立方体结构(例如，多维在线分析处理)并事先做好聚集，这可以让查询性能有很大提升。虽然会占用更多的磁盘空间和数据处理时间，但在性能方面会有很大提升。

4.考虑使用固态硬盘。这对磁盘速度问题会有意想不到的成本节约。最近，我正在跟一个有 35GB 的 OLAP Cube 的客户一起工作，性能很慢，聚集时间和查询速度都很慢。我推荐他们试试 SSD。果然，在测试工作台上就有了一台崭新的 70GB SSD PC。

这台机器是配置合理，RDBMS 安装在硬盘上， OLAP cube 创建在 SSD 上。Cube 聚集得更快了，查询性能的改进就更显著了。有些查询比相同级别的聚集快 20 倍。与性能的提高相比，SSD 的成本是完全是微不足道的。

5.考虑在笔记本上做 SSD。(我自己也这么做了，我现在正在用的这台笔记本上有 250G)，但这也是对磁盘集中的令人难以置信的应用。如果可能，在内存中执行抽取、转

换和加载(ETL)处理。对于执行时间很长的 ETL 操作，可以用磁盘缓存(以防处理失败)，缓存到 SSD 而不是硬盘。

6.为分析的目的，而不是事务的目的对分析结构做索引。听起来很显然，但是，我见过无数的关系型 OLAP 星型模式，其中的索引策略都是根据事务型系统的长期经验而做的，很少有是根据分析型系统的经验。

例如，默认情况下，许多关系型数据库引擎都会对表的主键做索引。这在事务型结构中很普及，许多查询都会用到这个索引——但是很少有人知道，在星型模式中，很少会用主键索引进行常规的单行查询。另一方面，维表中所有的分析列都很有可能被查询，也常常会用到索引。

(作者：Mark Whitehorn 译者：包春霞 来源：TT 中国)

原文标题：数据仓库性能优化的六个技巧

链接：http://www.searchdatabase.com.cn/showcontent_49141.htm

如何规划一个高效的 BI 数据仓库项目

Circor International Inc.曾经有一段向外开拓市场的历史。这家位于麻省的制造商生产阀门、导管和其他工业产品，它在过去的 20 多年中收购了 10 多家公司，以扩展市场并为客户提供更优质的服务。

但是，现在这家中型制造商的关注点已经转移到内部。它正在建立它的第一个商业智能(BI)数据仓库，以便在某种程度上理解如何将单独的实体一起绑定到更加高效的整体上。

“这其中的主要动因是如何理解所有的这些业务，这样它就可以使过程无缝地执行，并让它们变得更加高效，”Mercury Software 咨询公司的 CEO Vinay Balasubramanian 说道，该公司现在正在帮助 Circor 建立它的数据仓库。

与产品数据仓库不同的是，数据仓库技术是 BI 和分析报告的基础。管理员通过分析来自生产系统收集的数据来识别新兴市场变化趋势或者可能实现的生产效率。

例如，如果跨地理区域的多个机构购买某些同类型的原材料，那么一个数据仓库可以帮助管理员决定是否合并多个来源或者商议总额折扣。

Circor 并不是唯一认识到数据仓库重要性的公司。“数据仓库对于现代企业是至关重要的，”McKnight Consulting Group 总裁 William McKnight 说道。

规划一个 BI 数据仓库项目

虽然数据仓库已经克服了价格昂贵和难以启动的问题，但是它们仍然需要重要的前期规划，以及认真选择正确的硬件和软件。

关键的第一步是要定义仓库将要解决的总体业务目标，Balasubramanian 说：“大多数公司在他们了解业务问题之前就错误地购买了技术，这往往在之后给他们带来很大的麻烦。”

类似地，组织不应该将仓库完全看作是一个 IT 项目。通常需要一个由技术和业务领导所组成的跨功能团队来建立一个满足业务用户需求的解决方案。

当他们从市场上选择多种硬件和软件时，制造商也可以使用这些定义来建立协议需求(RFP)。从正面看，产品过剩意味着可能已经有一个解决方案的可以满足任意制造商的需求。但是问题是对所有选择进行分析以得到一个可管理列表是很费时间的。

咨询师建议客户避免卷入供应商的功能竞争中，相反应该关注他们的制造技术，以及这些技术是如何服务小型、中型或者大型公司的。当然，价格也是很重要的因素。“如果是一个产值 500 万美元的制造公司，而建立数据仓库需要花费 1000 万美元，那么这个投资回报显然是不合理，”Balasubramanian 说道：“然而，如果您投入了 50 万美元，而数据仓库能够节省 300 万或者 400 万美元，那么 ROI 就肯定是合理的。”

新的数据仓库技术

数据仓库技术正在多个方面得到发展，如可能降低成本和加快实现速度。而走在发展前列的是数据仓库设备和云计算。

设备将必要的硬件、软件、操作系统和存储资源整合到一个预制的包中。分析师表示这种水平的整合可以减少相当大的项目开支。根据所使用的设备，公司也可以通过开源软件和数据库节省成本，这些开源产品可以替代高价的名牌产品。但是，这里有一个折衷的问题：因为设备供应商一般只在其中 1 或 2 个技术领域中较为擅长，因此并非所有设备组件都将是同行中最好的。

尽管如此，根据 Gartner 最近的报告显示，数据仓库设备在去年出现了增长高峰。随着市场的升温，Gartner 分析师建议潜在的买家采用客户参考和经过概念验证的项目来指导他们的购买决定。

有些行业观察者预测在一个设备中实现的数据仓库趋势将会更多地受到基于云的按需计算服务和 Software as a Service (SaaS)解决方案的影响。但是，目前有些制造商还没有做好这种转变的准备。“虽然我们了解这方面的信息，但是‘我们也不想成为试验品’”Balasubramanian 说道。

建立需求驱动的预测

通过部署一个正常运行的数据仓库，制造商就拥有了一个进入特别重要分析领域的机会：按需预测。需求信号仓库(DSR)是一个大型消费产品公司长期使用的工具，它依靠一个中央数据库收集日常销售数据和相关信息，以便帮助管理员分析最新客户的购买模式，这通常发生在详细的 SKU 级别。用户反馈 DSR 具有很多优点，包括减小库存损耗，提高销售预测和减少库存量。新的基于 SaaS 的 DSR 推动了最大型公司的 DSR 应用。

但是 Balasubramanian 提出了最后一个警告。需求数据并不总是能够适合作为传统数据仓库基顾的关系数据库的记录和字段。他警告说，“有很多市场需求有很多未结构化的数据，而且很难将它[与传统的数据库]无缝隙地整合到一起”。

(作者：Alan Joch 译者：曾少宁 来源：TT 中国)

原文标题：如何规划一个高效的 BI 数据仓库项目

链接：http://www.searchdatabase.com.cn/showcontent_48782.htm

如何应对数据仓库架构挑战

本周在美国芝加哥举行的 TDWI 大会上，演讲者发表了关于如何处理问题数据仓库架构，以及防止像数据源和信息管理孤岛这类问题扩大化的建议。

参加本次会议的技术专家们有许多关于数据仓库方面的问题。

Kevin Najimi 是一家金融服务公司的高级集成开发者，他期望能寻找到关于如何评估甚至重新架构建立数据仓库实施的创新想法。

Najimi 说：“我们正处在重新思考我们的架构，我认为业界以及公司最大的挑战之一就是数据量，还有第三方应用和第三方数据源的迅速扩展。”

Kenneth Jarvis 是 HCA 公司一名应用程序开发人员和数据仓库团队领导，他也出席了会议，并时刻关注阻碍数据仓库成功的这种“经典”障碍。

他说：“我们从不同的系统中获取数据，并努力使他们对客户展现起来显得像是同源的。这一直是过去面临的困难，估计将来也还要面对这一问题。”

Jarvis 还说，另一个挑战是要跟得上 HCA 公司业务部门采用的各种信息管理实践。

他说：“我们是一家很大的公司，每个人都有自己做事的规矩和方式，从公司的一部分到另一部分可能会有一些做事方式不一致的地方。”

Suvendu Datta 是一家保险公司的信息管理专家，他在寻求如何发起数据仓库策略的一些建议。像许多保险公司机构一样，Datta 的公司也在同时处理流入的电子信息和严格的数据保留规则。

Datta 希望他公司的新数据仓库项目能使事情变得更加容易些。

他说：“主要的挑战在于架构，因为我们的复杂性很高。我们与精算师一起工作，他们构想了各种各样的指标。我们拥有庞大的数据量，还有大约十年的历史数据。”

Datta 的公司运行着 90 多个自构建和商业应用程序，其中有许多重复信息。他说，清理以前的数据质量问题给数据仓库的实施又带来了另一个挑战。

他说：“我们现在有一些孤岛，甚至到处都是孤岛。你可以在许多地方找到相同的信息。”

Datta 说 TDWI 会议的演讲者肯定了他的一次只对一个系统或者业务单元推行数据仓库项目的计划，同时要记住保证企业范围内目标一致。他说，数据仓库项目最开始将首先集中在金融信息方面。

他说：“基本上，我们的计划是选择一个项目，在新的数据仓库架构上创建它，然后开始在它顶部慢慢构建。”

颠覆数据仓库架构的经典规则

Evan Levy 是 Baseline 咨询公司的联合创始人，该公司是一家数据管理咨询公司，最近被 DataFlux 公司收购了。他也参加了本地 TDWI 会议，并提出了一些不同寻常的建议。

Levy 说，与会者应该从另外一个角度来看待数据仓库方面大家广为接受的一些经典原则，如果有必要的话我们可以考虑实施替代方法，只要对我们有意义就行。

Levy 告诉其他人至少要重新考虑“人们不应该对运营系统使用第三范式”这一观点。

他说：“我们的目标是发起质疑，假定你可以或者不可以用二维或者第三范式做某些事情。我们不是在挽救生命，每个人都对我们工作的方式太严肃了，好像只有这一种方式能完成工作似的。”

Levy 认为关于数据仓库架构的决策应该基于个别组织的具体需求决定，而不一定选择最流行的方法。

在做架构决策时，Levy 说 IT 专业人士应该着重考虑对他们的环境有意义的部分，他们支持多少用户，以及整体复杂性和复杂程度，还有预算限制。

他说：“有一件事情要记住，我们不是因为喜欢看漂亮的图画才构建架构，我们构建架构是用来支持一组商业前提和需求的。”

(作者：Mark Brunelli 译者：冯昀晖 来源：TT 中国)

原文标题：如何应对数据仓库架构挑战

链接：http://www.searchdatabase.com.cn/showcontent_49430.htm

考虑两种数据仓库架构共存的可行性

在 Google 上搜索“Inmon 和 Kimball”，你会和轻松地找到这两个名字的概念，它们是两种著名的数据仓库架构方式。然而在这信息的海洋中，你会发现几乎所有的内容几乎都能得出一个结论，那就是要在 Bill Inmon 和 Ralph Kimball 两者之间选择其一。

但是，“数据仓库之父”Bill Inmon 却告诉我们，在一定的环境这下，其实这两种架构完全可以共存，且协作良好。对此，来自电力公司的 BI 系统架构师 Bill Harrison 表示：“两种架构都有各自的应用，Inmon 的标准化数据模型对于集中式数据仓库来说是非常好的，但当你设计数据集市的时候，Kimball 则更佳。所以没有理由我们不能两者都使用。”

回顾 Inmon 和 Kimball 的历史

在 TechTarget 数据库编辑的一篇博客中，我们对 Inmon 和 Kimball 的两种数据仓库架构进行了对比，那么这里我们就来回顾一下他们两个人的历史。Bill Inmon 和 Ralph Kimball 早在上世纪 90 年达初就发表了各自的方法论，然而其中有一点是二者相同的：他们都希望帮助企业实现高效的信息管理并做出更好的商业决策。只是二者实现的方法不同而已。

Inmon 以“数据仓库之父”著称，他的方法就是构建企业数据仓库——集中式的关系型数据库管理系统，它能够为用户提供访问高质量、高集成、标准化数据的能力。有趣的是，Inmon 是第一个承认他的方法比较昂贵，且短期内不易看到 ROI 的人，但同时 Inmon 也强调他的方法拥有长期持久的投资回报率。

Kimball 被人们称为“商业智能之父”，他开创了数据集市的概念，这是一个小型的信息库，设计用来企业内特定部门的需求，比如财务、人力资源、销售等。虽然 Kimball

的空间模型能够带来快速的投资回报率，但是专家认为在数据集市的海洋中想要维护数据质量是一件棘手的事。

两种数据仓库架构在过去的 20 年中都经历了不断地演变过程，Inmon 架构目前包含了文本数据仓库，而 Kimball 现在则更加关注数据一致性。近期，TechTarget 网站对 Bill Inmon 进行了专访，他相信两种架构完全可以协作共生，甚至可以起到意想不到的“化学反应”。

“Kimball 架构应对数据集市是非常好的，部门级数据管理可以通过构建数据集市的方式进行，”Inmon 说：“但是企业同样需要通过集中整合的数据仓库来对数据集市进行掌控，这可以用到经典的 Inmon 模式的数据仓库。”

Inmon 与 Kimball 共存的证明

在 TechTarget 本周的报道中，我们分享了一个数据集成建模工具的实施案例。来自美国的 Omaha 电力公司实施了 Inmon 模式的数据仓库和 Kimball 模式的数据集市，他们的 BI 系统架构师 Bill Harrison 认为网络中报道 Inmon 同 Kimball 不能共存的文章是一派胡言。

“这样的说法已经存在不是一年两年了，网上的报道更多的是在炒作，吸引人的眼球，” Harrison 说：“我认为 Inmon 和 Kimball 不能共存的说法太荒诞了，两种架构应该是互补的，我们完全可以一起使用。”

Harrison 表示，在构建数据集市时选择 Kimball 是有道理的，因为这种架构理解起来相对简单，而且能够提供过硬的性能和快速的回报。这个架构设计的初衷，就是以最快的速度将数据传递给业务用户，大多数的源系统并不是这样设计的，相反地，它们是设计用来更快地从业务人员那里获得数据。

举例来说，Omaha 电力公司运行了 Oracle PeopleSoft 软件来帮助企业跟踪特定区域内的客户。PeopleSoft 应用中的信息表格主要是用来让用户更轻松地将信息填进去。

Harrison 对此解释说：“那么如何将信息从系统中传递出去呢?你需要不同的设计，数据集市和 Kimball 模式可以解决这个问题，它对数据完全地重新进行设计，然后将部分数据‘去标准化’，使得它们可以适应更快的报表和查询需求。”

另一方面，Harrison 表示 Inmon 架构在公司内部负责设计并维护集中式企业数据仓库，同样运行的非常好。理解并遵循 Inmon 理论是非常重要的，在创建一个企业数据仓库时，它强调了数据标准化。无论是否服务于 Kimball 模式的数据集市，Harrison 认为这都是需要注意的地方。

(作者：Mark Brunelli 译者：孙瑞 来源：TT 中国)

原文标题：考虑两种数据仓库架构共存的可行性

链接：http://www.searchdatabase.com.cn/showcontent_53187.htm

大数据背景下的数据仓库最佳实践

Wayne Eckerson 是 TechTarget 业务应用程序和架构媒体部门的研究主管，他总结说：如果想要成功处理“大数据”，您需要正确的文化、人员、数据和工具。只要将所有这些元素整合在一起，就能够形成一个数据仓库最佳实践计划。

根据 Eckerson 和其他分析师的看法，实现这个过程需要精心计划，以及清晰理解大数据管理技术和过程所带来的潜在机会和挑战。

首先，Eckerson 说，“您的企业最高领导必须愿意”购买所需技术，并决心培养面向分析的文化来保证公司将会使用这些信息，最重要的是“不会回归使用电子表格”进行数据分析。当组织寻求应对大数据存储和管理挑战时，他们需要制造转向使用更有针对性数据仓库平台的可能性。Eckerson 说，这些产品能够实现比通用数据库高“一个数量级”的处理性能。

然而咨询公司 Winter Corp 的总裁 Richard Winter 警告说，Hadoop 和 MapReduce 等新兴技术并非是所有大数据管理问题的解决方法。Winter 说，企业必须要谨慎，不能“急于将婴儿放进浴池里。有些人认为他们现在就能够使用 Hadoop 做任何事，而停止购买传统数据仓库技术——但是，对于大多数企业而言，这是错误的。”

Winter 建议先寻找独立应用程序，评估最适合大数据应用的平台。他说，要考虑两个关键因素：数据保存时间有多长，以及数据使用方式是什么。核心事务数据属于数据仓库，它可以基于长期使用和价值进行系统管理。另一方面，点击流数据、反映客户情感的社交网络内容和其他非结构化数据可能适合保存在一个 Hadoop 集群中，特别是那些保存时间不如事务数据长的信息。组织中数据的广泛访问方式也会影响技术平台的选择。

按照 Forrester 和 Gartner 的定义，容量并不是大数据的唯一特点;这两家公司都重视这样一些属性，如种类和可变性。但是 Forrester 分析师 James Kobielus 说，在实践

中，准备一个数据仓库来处理大数据本质仍然是关于可扩展性的问题。此外，他提出了三个关于数据仓库最佳实践的技巧，目的是帮助组织交付更强大和更具扩展性的系统。

大数据决策点：纵向扩展还是横向扩展？

首先，要考虑数据仓库架构升级和可能的构建并行性。Kobielus 说，可能的步骤包括基于共享内存的对称多处理器纵向扩展数据仓库服务器节点，或者使用服务器集群或无共享的大规模并行处理系统进行横向扩展。将 MPP 安装分成中心、分段和查询层是另一种方法。但如果不注意底层技术架构而草率地实施这种改变，很可能会产生不良结果。例如，单核 CPU 可能无法满足 MPP 需求，而一般必须增加存储 I/O 带宽才能够支持增长的处理能力。

其次，企业考虑在硬件和软件能够解决具体性能问题或缺陷时，采用数据仓库集成设备。第三，他建议公司对数据仓库的数据管理和存储分层进行优化，以提高性能。这可能包括压缩数据来提高效率，优化数据库模式，联合与分区，以及使用非传统数据库技术，如“特殊用途的”列式或者内存数据库。

Lyndsay Wise 是咨询公司 WiseAnalytics 的总裁和创始人。她指出，大数据项目通常的最终目标通常与传统数据仓库项目相同——例如，提供能够帮助业务用户确定客户购买模式或协助欺骗识别的信息。它们面临的挑战也是类似的：“我们处理的数据包含许多不同的细微差别，但是结构往往取决于完整性和数据质量问题，或者受到数据管理和数据监管的影响。”

但是，Wise 补充道，这些挑战的难度可能由于管理的数据量和复杂性而进一步提升，特别是当大数据项目需要从多个数据源获取信息时。结果，将大数据整合到一个数据仓库过程的公司需要认真评估他们的能力。Wise 说：“组织希望表明他们拥有最好的 IT 人员，但是除非他们的 DBA 和开发人员精通于数据仓库和专业的大数据技术，否则购买外部服务更有利于真正实现一个强大的平台。”

Wise 指出，对于大数据，从分析角度确定希望实现的目标，以及预先决定所需要的信息与影响整合的问题类型，都是至关重要的。“一定要理解各方面的相互关系。”

（作者：Alan R. Earls 译者：曾少宁 来源：TT 中国）

原文标题：大数据背景下的数据仓库最佳实践

链接：http://www.searchdatabase.com.cn/showcontent_54708.htm

大数据蔓延 企业需重新定位数据仓库策略

在计算机技术出现之前，人们一直苦恼于没有足够能力去处理大量的信息，IT 技术给了我们自动化的系统和工具，从此我们可以去存储并分析大量信息。但是现在的交易系统已经发展到非常繁杂的阶段，包括互联网、传感器、移动设备在内的数据源每天都会产生各种各样的数据，有时这些数据就如同洪水猛兽般，吞没了不知多少公司的 IT 系统，多年前搭建起来的数据仓库架构已经无法再应对海量数据的压力。

企业的数据仓库团队正面临着巨大的挑战，管理信息海啸或者我们称之为“大数据”，需要技术人员平衡已有的系统和新近的工具以及技术。来自 Forrester 研究机构的分析师 Brian Hopkins 认为，要想处理好大数据问题，企业至少应该对旧有的传统数据仓库系统重新审视。他举例说，使用传统的中心辐射型连接进行数据集市分割，并将某种形式的大数据集成到一个集中式数据仓库系统中是非常具有挑战性的。目前的数据仓库系统主要是应对结构化数据，但是所谓的大数据则更多的是指那些非结构化数据或者半结构化数据。

Hopkins 表示：“从某种意义上说，大数据已经颠覆了传统数据仓库的设想。数据仓库以及商业智能环境的主要目的就是为了能够回答业务用户提出的具体问题，它包括对数据进行清洗，通过 ETL 过程将数据最终导入到报表中进行分析。因此可以说这样的方式，企业中只有 5%的可用数据得到了充分的利用。然而更糟的是，有时甚至还要远远低于这个百分比。

新的数据仓库策略

相比之下，大数据策略往往将精力放在一个更宽泛的信息范围之内。几年前那种大数据库、要求统一的概念已经逐渐消失了。另一方面，目标数据存储以及所谓的分析沙箱(Analytic Sandbox)是大数据环境下非常常见的，它们的复杂程度会对 IT 技术人员以及数据仓库团队造成很大困难。

TechTarget 业务应用分析总监 Wayne Eckerson 表示：“随着大数据概念的提出，相信会有更多新鲜的模式出现，有些甚至是与我们传统概念截然相反的。但是分析大数据一定是非常困难的，因为量太大，而且成本颇高。这也就是为什么有很多企业在寻求更新的技术，比如开源 Hadoop MapReduce。”

来自一家咨询机构的顾问 Richard Winter 认为，大数据为企业提供了更多的机遇，这些企业善于从数据中洞察业务趋势，这在以前是不可能出现的。他举例说，一家“智慧”的空气净化机企业开发出治疗哮喘的方法，它们的产品内置有无线接收发送设备，能将病人数据、时间以及地点传送到数据库中。通过信息整合，空气净化机甚至可以提醒患者附近是否有潜在的过敏源。这样的设备还可以帮助医疗研究机构更好地分析哮喘症状。

数据量带来的挑战

当上面提到的产品获得广泛应用之后，来自数据的压力就会随之而来。制造商需要对这些数据进行存储，处理并最终提供给系统进行分析。

数据仓库专业人士还需要理解大数据并不只是在数据量这一个层面上，TDWI 机构的数据分析师 Philip Russom 表示，其他的大数据属性同样非常重要，比如它的变化多样，包括 Web 点击流数据、呼叫细节记录、销售网点数据、社交媒体文本等。

Philip Russom 认为，当企业面临管理大数据平台时往往会像车灯前举足无措的小鹿(焦虑不安)，他们能够了解到大数据的潜在价值，但是又苦于管理的复杂程度，特别是大多数数据不能够通过传统数据仓库来处理的时候。

为了避免上述的情况，专家建议可以试着将大数据管理的目光放到更小的范围内，比如客户行为这样高回报率的领域，然后可以利用传统数据仓库和新技术新工具的混搭来完成最终的目的。企业关注的话题不一定要大而全，小而精反而更好。

(作者：Alan R. Earls 译者：孙瑞 来源：TT 中国)

原文标题：大数据蔓延 企业需重新定位数据仓库策略

链接：http://www.searchdatabase.com.cn/showcontent_51589.htm

医疗行业面对的数据仓库挑战

从理解业务需求到拆除文化藩篱，USF 医疗中心的 IT 部门面临一些重大的数据仓库挑战。

总部位于 Tampa 的 USF 医疗中心包括 USF 医学学院、公共健康、护理、理疗各部门和即将开放的药房。它还拥有超过 300 名执业医师教师、一项医疗实践计划、一家诊所、一家研究机构，还与当地几家教学医院保持合作伙伴关系。

USF 医疗中心的助理 CIO 和应用开发主管 Sidney Fernandes 说，早在 2006 年，USF 医疗中心的 IT 团队就接到任务：为财务报告及其他商业智能(BI)相关的业务创建单一的真实数据源，即一个数据仓库。但该团队在实现这一目标之前必须先克服一连串障碍。

作为第一步，团队需要识别、评估并找出如何操作即将投入的数据仓库任务的复杂性。然后，他们必须评估、选择合适的供应商，让系统运行并取得成果。

USF 医疗中心的主管们要快速、方便的访问财务信息和其他报告以用来评价全体教员和分配资源。这将要求 USF 医疗中心使用与临床收益、研究收益、教育收益、开支及员工工资单相关的多数据源系统的数据。

Fernandes 说，数据仓库实施的源系统包括两款金融应用系统即 Oracle PeopleSoft 财务系统和 CODA 财务管理软件;两款人力资源应用系统，即 Oracle PeopleSoft 和 Cyborg;一个电子病历档案库;通用电气业务管理系统;流动外科支持系统;图片存档系统;一个放射信息系统;以及各种电子表格。

“我们希望获取所有源系统并进行集成，从而使决策者能完全了解一名研究员或医生的生产，他们申请的拨款、发表的文章、诊断病人的多少以及 USF 为他们的支出，”他解释说：“这是数据仓库建设的起源。”

商业用户面对的数据仓库挑战

USF 医疗中心的 IT 团队面对的两大障碍包括如何访问到各种信息源和从商业用户那里获得数据仓库系统的需求。

“我们必须克服一些行政障碍使人们能够信任我们，向我们提供数据从而将数据载入数据仓库”，USF 医疗中心的数据仓库架构师和应用开发助理 Swapna Chackravarthy 如是说：“由于这些源系统是如此不同，以至于他们很难集成为一个统一的数据视图。”

Chackravarthy 和 Fernandes 与部门主管协商获权访问源系统后，他们遇到更多的问题，因为 USF 医疗中心各部门从来没有标准化文件命名约定和其他成规。

“我们发现，各部门的代码没有标准化且各源系统之间差异很大，”Chackravarthy 说：“我们必须通过建立一种体系架构封装这些不同的数据结构来克服这些问题。这是我们制作集成报表的唯一方法，当然这也需要应用一些现存的为此目的的内置程序。”

同时，已从商业用户那里获得制作实用 BI 报表的系统需求。IT 专业人士曾报道商业用户在技术立场上通常不知道他们想获得什么，但 Fernandes 认为这是胡扯。

“业务用户始终知道自己想要得到什么，”他说：“他们只是不知道如何操作。”

Fernandes 说，任何适当的数据仓库项目的关键需求应着眼于企业用户的痛点。这涉及到确定在他们的工作中哪些任务是过于耗时的，哪些任务只是有些繁琐。

“每个商业用户都知道他们的痛点，” 他说：“给他们五分钟就能告诉你痛苦所在。”

USF 医疗中心选择 BusinessObjects 而放弃 Cognos

从商业角度获得系统需求之后，USF 医疗中心考虑了许多软件供应商。看起来最有可能满足 USF 医疗中心需求的两家公司是 BusinessObjects 和 Cognos。

Fernandes 说，这两种产品在特性和功能上都很适合。但最终选择了 BusinessObjects，主要是因为 Fernandes 发现它在应用业务规则上更灵活。

“Cognos 使用大量的直接 cube，所以我们不得不制定更加严格的业务规则，”他说：“BusinessObjects 则更强调整体构建，所以我们不用去做重构多维数据集之类的事情。而我们的数据和指标之类的定义随时间是变化的，因此 BusinessObjects 让我们感到更加合适。”

Chackravarthy 补充说，她相信 BusinessObjects 在即时报表方面提供更好的能力。

“从最终用户的接口角度来看 ad hoc 报表制作以及基本的钻取、切片和切块操作，我们发现 BusinessObjects 更加直观和更易于使用，”她说。

Fernandes 补充说，数据仓库本身系统驻留在 Oracle 10g 上。

克服数据仓库的挑战：来自专业人士的建议

Fernandes 说，USF 医疗中心目前建立的新数据仓库已上线运行，并以时间片段的形式提供必要的报表。

“利用数据仓库，数据直接展示出来正如在源系统中的一样”他说：“这直接导致业务流程的改变，因为 CFO 几乎实时访问薪酬、进度和研究指标或者临床指标之类的事项。”

Fernandes 说，考虑构建数据仓库项目的组织在投资任何一项新技术之前都应该获得对预设范围的详细了解。

“请确保你有一个数据架构的计划，并确保该计划不是一成不变的，”他说：“因为数据要告诉你很多业务范围内的流程可能没有想到的事情。”

(作者：Mark Brunelli 译者：宋广磊 来源：TT 中国)

原文标题：医疗行业面对的数据仓库挑战

链接：http://www.searchdatabase.com.cn/showcontent_42262.htm

迪斯尼乐园诠释数据仓库最佳实践

在实施了一个数据仓库的内部项目之后，儿童梦想缔造者迪斯尼乐园从无数销售经营中获得了“可付诸行动的智能”。

该项目在去年已经基本完成。它的目的是将 Disney 的全球商品经营管理集中到一个 ERP 系统上，即知名的 SIMBA（辛巴），或者称为单一整合商品业务应用程序。

迪斯尼乐园的数据仓库和分析主管 Juan Gorricho 表示：“我们的项目名称本身就是很有意思的”。 Gorricho 在纽约最近举办的 2011 年企业架构和数据仓库峰会上谈到了 SIMBA 项目。

Gorricho 表示，Disney 每年都有 10 亿美元商品销售收入，而建立一个 ERP 系统可以处理这些信息是极具挑战性的。这就是 Disney 首先关注位于美国的主题公园和水上公园的项目的原因所在。但是，最终这个系统将会包括东京、香港、巴黎以及最近在上海新建的公园。

最新的集中式 ERP 系统是设计用来处理商品管理、存货管理和相关业务过程的。但是 Disney 也希望平衡财务和业务智能(BI)报告和业务分析系统，这意味着建立一个新的数据仓库。

“从项目的角度看，数据仓库的目的是为了支持大部分和关键业务过程的分析报告，”Gorricho 说道：“这些数据可以带来很多价值。”

最新的集中式 ERP、数据仓库和分析系统正帮助 Disney 更好地管理存货、分析销售额和预报特定领域的商品需求。Disney 在该项目中所使用的一些产品包括 SAS 分析软件和 Teradata 数据仓库技术。

Gorricho 表示，这个项目代表了 Disney 的几个 “第一”：第一次使用一个商品企业数据仓库来制作财务报表。结果，Disney 团队通过某些特定数据仓库最佳实践获得了回报，Gorricho 向与会人员分享了该个经历。

从业务方面着手处理数据仓库项目

在过去，Disney 的 IT 部门主要负责发布 BI 报告和业务用户分析。但是 Disney 希望改变现状，由业务用户来主要负责创建他们自己的报告。为了实现这个目标，公司决定对结构进行一些修改，以便强调技术单元对于业务需求的新关注点。

在开始 SIMBA 项目之前，Disney 将它的 IT 部门的名称变更为“全球业务技术占略”单元——从业务的角度来说，这为 IT 着手处理数据仓库和分析项目提供了一个平台，Gorricho 说道。

Gorricho 表示，当对数据仓库、BI 和分析报告功能以及任何相关的业务过程进行建模时，对于技术员工而言，有几个问题是很重要的：业务尝试实现什么目标?业务用户打算如何使用数据仓库中的数据?从数据仓库的角度看，有哪些方式可以使业务更容易执行?

“之前，业务只是一个远程客户，而我们只是为他们填写[订单]，”Gorricho 说道：“他们现在看起来类似于伙伴。”

只使用最重要的数据

当 Disney 的技术员工开始建立数据仓库时，首要的策略是从全国引进所有的商品数据，包括 450 个数据库表和“我无法计算的数据兆字节，”Gorricho 说道。

Gorricho 解释道，很快人们就会发现如此大量的信息将很难管理，因此 Disney 采用了一个新策略：专注于将帮助业务人员实现他们目标的与商品有关的数据。在 Disney 中，这意味着使用所需的信息来实现财务、预测、存货和其他业务报告。

“将所有的信息都保存到数据仓库是毫无意义的，”他说道。

建立相互制衡关系

对于业务用户而言，完全信任存储在数据仓库中的信息是很重要的。这就是 Disney 创建一个复杂的制衡和审计控制系统以便确保数据仓库中的数据与财务系统记录的数据相匹配的原因所在。

Gorricho 表示，差异总是会不停地出现，因此一定准备好快速地修复问题。唯一的希望是，这些错误一般能提供了足够的机会让业务和 IT 一起创建恰当和可持续的解决方案。

保持 BI 和分析报告的简单性

希望一般的业务用户使用自助 BI 功能的组织最好要确保它们足够简单易用。Gorricho 表示，这意味着要保证报告是面向业务用户的需求、要容易使用，而且最重要的是要容易生成。

技术人员需要完全地理解哪些业务用户拥有特定的数据和报告，并且他们必须能够从长期业务需求和“可维护性”方面来考虑系统的定制。

“总之，我们建立了一些使业务从此能够很容易管理的简单解决方案，”他说道。

（作者：Mark Brunelli 译者：曾少宁 来源：TT 中国）

原文标题：迪斯尼乐园诠释数据仓库最佳实践

链接：http://www.searchdatabase.com.cn/showcontent_48332.htm

数据仓库设备趋势：聚焦 BI，细分市场

在数据仓库领域，服务提供商一贯在性能、存储以及分析功能上面下很大功夫。在功能不断完善而价格持续走低的今天，企业在选择数据仓库系统时往往能够以较低的价格获得更多的实惠。

随着数据仓库设备(data warehouse appliance)的出现，企业在进行采购时又多出一个选项。因此，用户更加倾向于购买集成的设备。数据仓库厂商也将更多的精力放在了如何为客户提供能高的性价比上。

最近几年中，数据仓库设备在缝隙市场中得到了长足又快速的发展，越来越多的厂商开始将他们的软件与硬件集成起来，并与商业服务器硬件提供商结成合作伙伴关系，于是用户开始逐渐适应了这种即插即用的部署方式。

虽然在使用数据仓库设备的时候，想要获得最佳的性能还需要一些微调，但是快速部署以及更低成本的优势还是为企业开发更多 BI 应用铺平了道路。而目前部门级别的数据仓库技术需求以及敏捷 BI 部署方式也是集成设备越来越流行的原因之一。

随着数据仓库环境逐渐成熟，企业开始寻求更多的 BI 应用，比如预测分析等，这也直接影响了数据仓库设备在市场中的地位。目前设备中经常用到的列式数据库以及大规模并行处理(MPP)技术让高级分析应用不再是一纸空谈，除此之外还包括内存分析、数据库内分析等等都是高端 BI 的必须品。

除此之外，业界的并购热潮也在继续，传统的数据仓库以及 BI 软件巨头都在寻求更好的技术来完善自身的产品线以实现集成设备的需求。在 2010 年中就不乏大的收购，如甲骨文收购 Sun、EMC 收购 Greenplum、IBM 收购了 Netezza，目前市场还处在不稳定的阶段，大型的并购也许还将继续下去，但与之同时到来的则是更多的研发力度，因此我们有理由相信数据仓库设备在未来还会得到更长足的发展。

在业内大型并购的同时，对于购买者还提出了一系列的问题。尽管站在市场发展的角度，厂商并购绝对是一件好事，但现实是企业数据仓库管理员能否在第一时间适应产品或者技术的变化。

任何一个并购都需要一定的时间来完成，于是交易过程将在用户中导致许多不确定性，比如之前的技术该如何发展?是被彻底改造还是保持原样?比如大家都熟悉的 MySQL。之前单独部署的模式是否会随着并购而变成完全集成模式，然后技术支持以及维护的工作如何进行都是比较棘手的问题。从根本上说，想要为客户带来信心的唯一途径就是不断加强技术研发的力度。

数据仓库设备逐渐进入 BI 市场并带来巨大影响是一个不争的事实，尽管并不是所有的企业都要使用集成设备，但是很大一部分都已经进入了评估的阶段，这些用户希望能够弄清楚集成设备究竟能在数据管理架构中带来多大的价值。由于数据仓库设备的先天优势，它完全能够满足大型企业的部门级需求，比如销售、市场以及供应链管理等。

总体来说，数据仓库市场无论是在产品还是流行度上还会有一个更大的发展，而随着大型并购带来更大的竞争，预计下一步厂商还将开发更好的系统来满足细分的行业需求。

(作者：Lyndsay Wise 译者：孙瑞 来源：TT 中国)

原文标题：数据仓库设备趋势：聚焦 BI，细分市场

链接：http://www.searchdatabase.com.cn/showcontent_50374.htm

数据仓库集成设备选型指导

选择合适的数据仓库技术从来都不简单。无论是传统的企业数据仓库软件还是目前非常流行的数据仓库集成设备，如何选择合适的方案来满足企业的业务需求是每一位管理者都要面对的，这其中不仅涉及到一些步骤和注意事项，同样存在着许多挑战。

目前介绍数据仓库设备的资料有很多，比如媒体报道的文章、厂商提供的白皮书、产品的 Demo 等等。但有些时候，太多的信息会对产品本身造成一定的影响，用户如何才能从众多的数据中找到他们想要的点?这些产品都有什么样的区别?它们能不能真正解决企业中存在的数据仓库问题?

在进行数据仓库设备的选型时，非常重要也是最核心的一点就是区分它们之间的相同点与不同点，特别是每一个厂商都会提到“查询性能更好更快、海量存储能力、支持高级分析”这样的功能特性。但是随着技术的不断发展，许多厂商也能提供类似特性的产品，还能够帮助客户进行定制以满足特定的需求。所以我认为真正能够区分这些设备产品的因素包括：是否有内存分析功能?是传统的行式数据库还是列式数据库?最为重要的，厂商能够提供怎样的增值服务?

随着集成设备市场逐渐走向成熟，许多小型的独立技术提供商都遭遇了并购的命运，这些技术产品纷纷融入到更广泛的数据仓库以及商业智能产品线中，因此大部分的价值都源自于厂商服务以及整体的客户体验。技术支持、灵活的许可政策、数据集成能力以及未来的扩展性将成为数据仓库设备竞争力的主要来源，而不是客户首次投入的成本。

因此在本文中，我们将为您介绍数据仓库集成设备选项的一些准则和经验：

你目前的 IT 环境以及内部标准。在进行数据仓库集成设备选型的时候，企业应该首先审视一下自身的内部环境，比如大型企业都有自己的软硬件标准，在进行集成或者添加新硬件的时候，很可能将违背这些标准，因此在进行选型的时候需要考虑更多的定义方式。

业务需求与项目范围。通过鉴定使用数据仓库设备的初衷以及关键目的，企业可以更好地理解什么样的设备能够满足什么样的需求。是主要进行高级分析还是用来存储海量数据进行基本的 BI 查询?这决定了选择什么样的数据库。举例来说，列式数据库在分析性能方面就有比较大的优势。其次还需要考虑你的源系统数量。

管理员知识储备以及 IT 资源。数据仓库设备越来越靠近主流，究其原因我们不难发现目前多数企业都已经跨入到 TB 基本的数据量，同时他们需要更高级的分析应用，而企业往往不希望进行复杂的数据仓库搭建工作。尽管集成的设备在进行部署时也需要根据需求来调整，但是通常它们的部署周期要远远低于传统数据仓库系统。即便如此，一些基本的需求我们还要考虑进来，一般来说，无论什么样的设备产品，都需要有 3-5 名管理员进行维护，这些管理员还需要有一定的知识储备。因此企业要么招聘该领域的技术人才，要么需要进行内部的技术培训才能驾驭集成设备。

未来增长因素。在进行数据仓库设备选项时，你还需要一些前瞻性的考虑，数据增长是一个重要元素，在未来的发展中是否会添加新的数据源?业务会扩展到怎样的规模?是否会调用更多的历史数据?企业需要区分不同的设备：一些厂商支持多种存储级别，一些厂商则需要客户自己添加新的硬件或者升级设备。在传统数据仓库系统中，针对数据增长的灵活性更高一些，而在选择设备产品时，你一定要将技术与业务增长评估紧密结合在一起。

总而言之，数据仓库设备市场存在一定的特殊性，企业在进行选型评估时需要考虑到这一点。每一个企业在选择合适的数据仓库设备时都会面临不一样的挑战，你需要认清关键的业务目标并对比每一个厂商产品技术的特点。

(作者：Lyndsay Wise 译者：孙瑞 来源：TT 中国)

原文标题：数据仓库集成设备选型指导

链接：http://www.searchdatabase.com.cn/showcontent_50425.htm

部署数据仓库系统需要避免的三大问题

企业在决定购买数据仓库集成设备，在对要实施的产品做好了选型之后，项目管理阶段就开始了，其目标就是成功地交付能满足特定业务和技术需求的数据仓库系统。但是，在达成目标的路上还有着大量的挑战。

记住，IT 项目通常有很高的失败率，部署数据仓库设备的公司应该确保制定了比较详细的项目计划。另外优先级对设备部署中遇到的其它问题也会变得很敏感，所以你可以从中学到很多。通过灵活计划和识别潜在的障碍阻力，将会避免遇到更多的困难。

审视在实施数据仓库设备项目中可能遇到的问题，为了确保项目成功，技术和业务挑战都是需要考虑的。

业务挑战。有时部署设备的时候，由于数据仓库项目的详细技术需求，企业可能会忽视了业务因素。除了识别数据仓库设备一些仓促的业务问题和需求以外，实施者应查看数据仓库中与数据有关的业务规则，试图计算项目的投资回报和新架构/扩展架构的总拥有成本。ROI 和 TCO 图有助于判断项目并评估其总价值。

不幸的是，一些企业不能展示数据仓库在除了数据清洗及其所提供的优点之外的潜在价值。问题是即便最初的设备部署得到审批，如果没有对业务有直接的看得见的价值，将来的扩展也很可能不被优先考虑，例如，成本节约或增加赢利等。

技术挑战。在数据仓库实施期间，还会遇到一些常见的技术挑战，这些挑战是与项目有关的。数据仓库团队应该了解正在实施的其他 IT 项目，以及这些项目如何影响数据源——例如，是否有什么变更会产生源数据或任何有关的业务规则。这些因素会影响项目时间，还会在正确的数据质量和完整性保证方面影响部署设备的能力。

识别集成的需求——包括与组织内现有 IT 架构有关的数据方面和系统层面。建立内部硬件和软件标准能减少集成挑战，流程将指导需要的工作。当然，数据库就基本技术而言对项目团队永远是个挑战——数据仓库设备项目也不例外。

数据准备。实施应用的时候，初始的数据装载和表连接预示了项目的成功和失败，这是开发数据质量规则的过程。做数据验证如果确实合理的控制，即使是反复数据清洗也不能解决质量问题。“垃圾进，垃圾出”这句话用在数据仓库设备项目上很适用：商务智能和高级分析要求很高的数据质量以便进行有效地知识发现。有些公司将重点放在确保数据完整性或者开发业务过程支持这种完整性，他们的部署更有可能成功。

无论潜在的问题是业务问题还是技术问题，评估每个问题是怎么影响实施过程的，又是怎么影响数据仓库系统环境结果的，这很重要。开发一个详细的部署计划和检查列表，然后照着这个列表，确保所有需要做的任务都如期完成，这能帮助一个组织在数据仓库应用项目中避免问题和错误。此外，应该有足够的持续性计划使数据仓库团队处理任何项目计划和实际部署之间的差异。

应用供应商极力宣扬将数据仓库软件和硬件绑定比传统的数据仓库更容易部署。数据仓库项目跟任何 IT 项目一样，还是要注意一些问题和障碍——但是如果做好了详细计划，并且多加小心就可以很好地开展项目而避免许多难题。

（作者：Lyndsay Wise 译者：包春霞 来源：TT 中国）

原文标题：部署数据仓库系统需要避免的三大问题

链接：http://www.searchdatabase.com.cn/showcontent_51029.htm

分析数据仓库设备优势与局限性

随着数据仓库设备(Data Warehouse Appliance)的出现，商业智能以及高级分析应用的潜能也被激发出来，许多企业将利用数据仓库新技术的优势提升自身竞争力。但是并不是说集成设备就是万能的，它还存在着许多的局限性，根据一个企业的业务目标和可用资源，有些时候数据仓库设备并不一定是最好的选择。

集成设备将企业的数据仓库硬件软件整合在一起，能够提供同传统数据仓库系统一样的优势。无论企业选择什么样的方式，他们都需要从多个数据源对数据进行整合，并且需要对数据质量、整合流程进行管理，同时支持 BI 和分析功能，只有这样才能从数据中获取更多的洞察力。

在许多情况下，当选择部署数据仓库设备时，有些企业往往希望提升查询性能、扩充存储空间并获得更多的分析功能。而另外还有些企业将其视为通向 BI 和分析的一个捷径，因为集成设备同传统数据仓库系统相比在部署周期上要短很多。

集成设备和传统数据仓库系统在结构上存在许多相似之处，但是传统的做法中包含了更多的软件和硬件，并需要技术人员对数据仓库架构进行设计和开发;而使用集成设备时，厂商将提供一个或多个服务器，这些服务器都是针对数据仓库软件进行过充分优化的，技术人员无需自行设计。

当我们要在二者之间选择其一的时候，中小企业或者针对部门级的用户往往会选择集成的设备，因为其部署周期上的优势。而随着“大数据分析”成为许多企业的首要任务，数据仓库设备往往会被视为传统系统的扩展，专门用来处理海量信息。同时伴随厂商和企业用户对数据仓库设备的愈加青睐，未来整合的系统必将在数据仓库和 BI 建设中发挥更多的作用。

因此，正确理解数据仓库设备的优势是非常重要的，它能够指导企业是否应该选择集成的系统。举例来说，那些希望解决具体业务问题的用户，比如增加客户数据洞察力，提升客户满意度，数据仓库设备的快速部署优势以及单独管理的特点将为这些用户提供更多的价值，因此再选择传统的数据仓库构建模式显然是不合适的。

尽管数据仓库设备能够满足许多不同的业务需求，但是有些情况下企业还是需要选择传统的数据仓库软件。无论从成本还是技术角度来说，对于那些在数据仓库和 BI 系统相对完善成熟的企业，将集成设备视为扩展组件就不太可行了。而且并不是所有的企业都对新的技术做好了充分的准备，比如列式数据库和内存分析等功能，这需要一定的技术知识储备才能够正常驾驭。

事实上，数据仓库设备存在着一定的局限性，而且企业在进行选型评估的时候一定要牢记这些点：比如拿服务器来说，尽管一些集成设备厂商提供了不同的硬件产品选择，但是大部分还是集中在一两个品牌上，这是否符合以往的选择?企业应三思而后行。

除此之外，新技术人才的招聘，内部员工培训过程都应该计算在投入成本之中，有些时候花的人力物力会与构建新系统差不多。存储空间也是需要考虑的问题，不适合集成设备的业务由于数据存储瓶颈也将到时总体性能下降的问题(当然对于传统数据仓库系统来说也是如此)。

由于媒体和厂商的信息轰炸，想要得到数据仓库设备的相关信息并不是一件困难的事。但是认清它的正负两方面作用时最为重要的，这能够帮助你决定是否部署新的设备还是自己搭建新的系统。

(作者：Lyndsay Wise 译者：孙瑞 来源：TT 中国)

原文标题：分析数据仓库设备优势与局限性

链接：http://www.searchdatabase.com.cn/showcontent_50485.htm