SearchBI 电子书 June 2012

企业数据集成实践手册

数据集成的处理方法与应用工具有很多种，企业的选择方案也不在少数。但无论如何，归根结底还在于“低成本高效率”，本手册为这一目标的实现提供了相关指导。

- *数据集成的五大最佳实践*
- *数据集成案例分享*
- *如何选择合适的数据集成工具*

TechTarget 中国

企业数据集成实践手册

我们今天谈到的数据集成早已不是几年前或十几年前数据集成的概念，它现在与商务智能（BI）相结合、与“大数据”密不可分，因此，我们需要重新认识并操控它。在本技术手册中，TechTarget 数据库与 SearchBI 网站编辑为您解析了数据集成的概念与技巧，奉上了其发展趋势与成功案例，并提供了极具参考价值的应用实践。

数据集成技巧

数据集成是一个很繁杂的工作流程，DBA 必须根据实际的操作需求使用合适的工具才能对数据进行全方位的掌控。随着商务智能（BI）的需求和大数据的出现，数据集成注定有一个全新的走向，也面临着前所未有的挑战。

- CRM 与 BI 系统的集成
- 选择 ETL 工具时的三个注意事项
- 理解数据集成的发展趋势
- 数据集成面临的挑战

数据集成案例

在数据集成的实际操作过程中，任何软件的使用都存在其利与弊。对于企业来讲，怎样挑选适合自己的一款软件并成功实施解决方案显得尤为重要，而想要做到这一点，就少不了对前人经验的借鉴。

- 医疗培训行业数据集成案例分析
- Dell Boomi 数据集成工具案例分析

- 云服务与数据集成联合解决方案案例

数据集成最佳实践

目前，越来越多的数据管理员都把数据集成项目看成一项巨大的挑战。大家都知道，不同的企业有不同的架构和需求，涉及到具体细节自然就会产生许多种不同情况。然而，总有那么一些点是在大多数情况下不容忽视的，我们以下作出了总结。

- 数据集成的五大最佳实践
- 企业数据库标准化最佳实践
- 数据集成工具选型最佳实践

实现 CRM 与 BI 最大程度上的集成

在大多数企业中，将 CRM 应用与商业智能(BI)系统进行关联，可以满足客户需求并确保其持续的忠诚度。当然，假如你有足够的行业专长并对你正介入的市场有一个相当不错的了解，那这是可以实现的。CRM 应用向 BI 系统提供数据。将 CRM 应用程序连接到 BI 产生可操作的洞察力；没有 BI，孤立的 CRM 应用程序仍然是一个毫无意义的业务磨盘。

例如，可以考虑建立一个银行的汽车保险产品的目标客户群。这个目标群来自 CRM 应用中的数据，与之对应的划分：20 岁出头每月收入 3000 元打算购买价值 5000 员摩托车的客户;29 岁或以上的客户，他们可以针对性地推销汽车保险;或者，企业经营者可能需要为他们的业务运输车辆购买保险。向这些客户推销汽车保险，要有与其相适应的特别计划，而这只能通过 BI 系统对 CRM 信息进行客户分析。

成功需要良好的数据

一旦 CRM 应用与 BI 相一致，它就成为商业战略的一部分。每份报表都内嵌智能，提供的输入数据是干净并适当聚合的。如果 CRM 应用程序的数据没有优化到 BI 系统所要求的水平，它可能会带来一个问题，即企业会认为 BI 系统是没有效能的。

要避免这种情况，可以采取严厉和不妥协的程序来聚合 CRM 应用中的多个数据库。除非把这一要求作为 BI 成功的强制性前提，填充 CRM 应用程序数据的部门是不会情愿去做的。

在企业里，发起更好的数据实践，开始收集数据的过程。许多企业要求新客户填写多页数据表，但是由于太乏味，总是填写不正确或部分完成。相反，数据可以使用创新手法在一段时间内逐步收集。例如，每两至三个月，呼叫中心人员可以电话回访，检查客户是否对提供的服务感到满意，同时可以获得更多的信息。

评估你的需求

下一步是定义 CRM – BI 扩展规模在哪里。不要被最新的 BI 技术所左右。专注于什么有助于企业。预测分析是一个安全的赌注，因为它促进行为智能，并有助于按需营销、向上销售和交叉销售。例如，如果通过 CRM 应用可知，客户 X 在每月第一周收到 30000 卢比的工资，预测分析应基于 X 先生花费最多的日期历史信息进行计算。利用这些预测数据开展交叉销售。

请记住，实时仪表板和 CRM 分析是不同的。自动分析包括预编程可操作技术。如果客户有一定额度的资金投资于一个互惠基金，预设分析会自动显示其他的客户可以投资的同类产品。实时智能使得信息更加市场化，使得 BI 更敏捷，有助于更快地响应市场趋势。这些市场趋势可通过收集来自 CRM 应用的数据加以察觉。

计划和发展

CRM 应用与 BI 系统的集成协作是应该有计划的。最常见的错误：整合两个应用的人认为 CRM 应用是业务驱动的，而将 BI 归属于单独的 IT 范畴。让两个队伍协调一致。关键是 IT 部门要理解业务部门希望从 BI 部署获得什么，适当的部署映射。不要管成本压力，不要急于实施 BI。坚持长期展望，随着时间的推移按计划扩展。

来自不同行业细分市场的组织将有不同的需求，供应商必须把重点放在这些需求上。如果一个供应商告诉你，他们能做一切，那么就让他们证明给你看。提出你所有的问题，不满足于使人不满意的答复。一旦开始部署，你不希望看到意外的弹出窗口。

CRM 应用将通过 BI 让你可以对客户进行 360 度观察。使用这种视图，明智地引导你的企业。不要鲁莽从事，因为你的手段而“错过目标客户”。请记住，克制是智慧的一部分。

选择 ETL 工具时的三个注意事项

在数据库管理技术中，提取、转换、加载(ETL)操作扮演了一个非常重要的角色。根据实际的操作需求，DBA 可以通过 ETL 手段对客户数据有一个全方位的掌控。有一些人认为，ETL 只是简单地将数据从多个源系统中提取出来，然后在加载到数据仓库中进行转换和集成。但是在实际操作当中，ETL 要比想象的复杂许多，因此 DBA 需要对它有一个熟练的掌握。本文就将介绍关于 ETL 的几点注意事项，希望引起您的足够重视。

良好的 ETL 中断重启功能

试想这样一个情况，你需要对 19 个数据加载进行转换，而由于某些原因在进行到第 9 个的时候发生了中断，那么再进行转换的时候你肯定不希望重头再来一遍。所以当遭遇操作中断的时候，能够从中断点继续进行操作的功能是十分必要的。如果 ETL 操作受阻，报表将得不到及时的更新，导致的结果就是管理人员只能从陈旧的数据中做出决策，想必这是所有人都不愿意看到的。

要解决上述问题，你需要建立一个“记录点”机制。如果任务被迫中断，你可以在记录点上继续完成任务，这有点像过关游戏中的“checkpoint”。因此，在选择 ETL 解决方案的时候，这样的功能应该是最优先考虑的选项之一。

另外，你还可以利用 C 语言等编写一个中断处理程序，这个程序将存储 ETL

操作的进程，它会记录故障点，然后再任务重新开启之前寻找到正确的位置。一个重要的准则，就是数据移动的速度究竟有多快。在这一点上，当评估 ETL 工具的时候，还需要考虑性能级别和重启功能。

管理快速变更的数据集

为能够顺利运行 ETL 操作，你所选择的工具应该拥有以下几个功能：

- 处理海量数据；能够将数据以最快地速度从一个地方转移到另一个地方。
- 实时监测交易的变更，并对数据进行同步。
- 能够处理多种数据类型，包括文本、非结构化数据等。
- 利用多处理进行分布式操作以及并行处理。

任何一款自动化 ETL 工具都必须能够提供最低级别的块复制功能，并拥有非常好的快速变更数据集管理特性。

针对大数据，为 Hadoop/Hive/PIG 架构建立一个沙箱。你需要有一个轮廓清晰的策略，在这基础上，新一代的大数据架构能够同之前的系统并存。你还需要对团队进行大数据技术培训，以应对新的 ETL 挑战。或者直接招募新的技术人员，对大数据处理有相关经验的员工，也可以免去一些培训的繁琐任务。

将数据加载到个体数据集市

在没有一个集中化的数据库情况下，拥有数据模板是非常重要的。它们是标

准化的接口，每一个个体或者部门数据集市都能够填充。确保你的 ETL 工具有这样的功能，能够扩展到一个数据仓库平台，将信息从一个数据集市流动到下一个。

理解关键的数据集成趋势与商业驱动因素

在近十年间，特别是经历了这次金融衰退之后，企业想要做出正确的业务决策就必须需要大量的数据。然而，拥有数据但不进行整合，业务质量也会持续下降。数据集成不但可以带动销售和盈利，而且还能提供透明性、隐私与安全性。

信息的需求在不断增长，数据集成的道路也在延伸。在本文中，我们将介绍重要的数据集成趋势。

业务需要越来越多的数据

业务对信息的需求量正处在空前的高度。它们需要准确的实时信息才可以高效地运行、增长。数据量的增长同时也提升了整合的复杂程度。

一些趋势带动了呈指数增长的数据：

- 公司不断生成更多的内部数据。例如：市场部会从 Web 分析器等客户接触点收集越来越多的数据。而跨国公司需要将许多国家的数据收集过来进行整合、分析和管理。
- 合作伙伴与供应商的外部沟通越来越多。沟通多了，数据的传输也就跟着多了起来。存货级别、发货日期、产品描述等等信息，每个公司都需要最及时的信息以便他们同客户共享。

- 数据的结构已经发生了变化，涉及到更多的非结构化数据，比如电子表格和网页等。非结构化的数据来自于企业的各个部门。虽然生成容易但是整合起来非常困难。在过去这部分数据往往被忽略，但是现在发现它的价值非常高，同样需要进行整合。
- 实时数据的需求提升。随着 blackberry 以及 iPhone 等移动设备的普及，人们希望得到更多更及时的信息。

理解数据集成的价值

数据需要集成才可以变得更有价值。这个是显而易见的，企业在近几年才意识到这一点。他们走了不少的弯路：多年以来，企业一直在使用 spreadmarts(Excel 等桌面数据库)来充当数据集市的角色，这不仅不能交付他们需要的信息，还造成了更多的问题和麻烦。

这些 spreadmarts 并不能为企业提供准确的数据，在进行业务决策时往往会冒很大的风险。使用它们是非常昂贵的，因为创建与维护它们需要专业人士花上大把时间，而他们应该做的是分析数据而不是收集数据。

仅仅搞清楚了 spreadmarts 也不能解决业务存在的问题。你需要制定一个有条不紊的计划，以保存业务信息价值的方式来替换 spreadmarts，同时还要达到

最高的收益率。目前许多企业已经开始着手进行这一项目，希望在进行数据整合的同时能够真正地吧数据融入到整个业务决策过程中。

数据集成在不断发展

数据集成已经超越了数据仓库和 ETL。尽管数据集成的基本任务是数据收集、转换并加入目标位置，听起来蛮像 ETL，但是最新的数据集成趋势以及工具所提供的技术已经超越了基础的 ETL。这些技术可以将数据转化成全面的、一致的、干净的信息。这些工具支持数据迁移、应用整合、数据分析、主数据管理以及业务处理等功能。

这些工具可以使企业了解源系统状态，进行数据清洗、确保一致性并管理所有过程，包括错误处理和监控。在过去，IT 部门都需要手动才能将这些功能添加到数据整合过程中。往往会由于时间与经验的不足导致失败，而最新的工具都预安装了这些功能。

在过去，ETL 被限制在夜间执行。数据集成套件现在包括了企业信息集成、企业应用集成以及面向服务的架构同 ETL 一起提供批量数据集成、能够与应用程序或实时 BI 进行交互。由于业务对实时信息的需求在增加，数据集成恰好可以解决这一问题。

手动编码的习惯很难打破

尽管数据集成工具的应用已经越来越广泛，但是在 IT 技术领域还有一个争议：手动编码还是使用 ETL 工具。企业数据仓库的标准是使用 ETL 工具，而像数据集市、cube 等下游应用往往还是靠手动编码。结果就是 IT 不能像企业要求的那样响应，spreadmarts 这样的工具也应运而生。

手动编码的应用程序往往不规范，难以更新与修改。而且现在还有一大批好用的工具帮你完成这一工作，只是会花些钱，而有的工具还是免费的。因此使用预安装的工具进行数据转换可以节省大量的 IT 时间与资源，好过从零开始对它们进行编码。

数据集成面临的挑战

BI 系统以及后端支持的数据仓库好坏取决于进入其中的数据质量。如果没有正确地处理 BI 集成过程，那么终端用户，甚至整个组织都可能会有麻烦。

据 Garter 公司的数据管理分析师 Ted Friedman 说，随着 BI 工具在组织中越来越流行，它对业务运营的成功也越来越关键，确保你有设计良好、执行很好的 BI 数据集成过程是最最重要的。

Friedman 说，Gartner 将与 BI 相关的数据集成挑战看做是 BI 和分析项目成功的拖累，这是项目彻底失败的最大原因。随着组织要管理的数据越来越多越来越复杂，数据种类和数据源也更多了，现在又加入了大数据，很多时间和精力都要花在为 BI 应用匹配、清洗和准备数据上。这是个讨厌的难题，尤其是当需要集成遗留系统的时候，为了揭示数据就不得不先了解旧系统。

另外一个复杂的因素是随着业务用户需要更快地访问 BI 数据，数据集成技术世界正在发生变化。

ETL 仍然是 BI 数据集成的最佳选择吗？

传统 BI 数据集成技术中使用最多的是抽取、转化和加载(ETL)软件，它从源系统中用批处理方式抽取数据。Friedman 说，新的数据集成技术比 ETL 工具需要更短的延迟。如变更数据捕获(CDC)软件和其它实时数据集成工具让你将新的

或修改的信息以实时、近实时的方式推送到数据仓库和 BI 系统中，这对类似欺骗检测这样的任务尤其有用。它是细粒度形式的流数据而不是像 ETL 那样采用的大批量数据。

另外一个选择是：联邦和虚拟数据集成交付方法，这种方法不需要将数据从源系统中移出来，而是从多个数据源中创建数据的统一视图让 BI 使用。用数据虚拟化工具，集成的数据不会到处都有。实时地抓取数据并将它们 Join 在一起，让它们看起来对于应用而言就像位于某处的一个数据库一样。

Fredman 认为，尽管出现了这种新的数据集成和交付工具，但如果认为 ETL 软件不再有价值了，也是不对的。 “ETL 仍然有用，”他说：“我们认为总是有地方需要用 ETL 的方式进行处理，因为不是所有的数据都能或者应该实时交付。”

的确，当许多组织仍然能从批处理方法中获得他们需要的数据时，数据集成供应商正在大力推广 BI 数据集成 实时选项。实时集成花很多成本，要求组织过去一直在做的东西都要发生改变，所以这需要是一个比较强的业务需求。

Intelligent Solutions 咨询公司负责 BI 解决方案的 Claudia Imhoff 表示赞同，她认为 ETL 还有一个角色——它是数据集成的搬运工，ETL 比它的新竞争者更灵活快速部署，更适合按时给操作 BI 应用的业务用户交付数据。

实时并不总是正确但更加真实

位于南非开普敦的 9Sight 咨询创始人 Barry Devlin 承认 BI 的实时数据集

成常常是不太必要的，但是 BI 和分析应用正在日益朝那个方向变化。“我认为人们之所以对它感兴趣是想看看它是如何运作的。”他说。

Devlin 举了一个美国保险行业的用户案例，来自汽车的实时数据，包括刹车和速度数据、行驶时间和其它信息正在通过移动网络传输给保险公司的业务用户，以确保保险人修改保费或者甚至可以在飞机上提供折扣。

正如 Friedman 说的，人们对获取和分析大数据的关注日益增加，这些大数据包括 Web 服务器日志、社会媒体数据和其他形式的非结构化信息，这给许多组织的 BI 数据集成过程增加了另外一层复杂性。

James Kobielus 曾是 Forrester 公司的分析师，他说非结构化数据“跟 BI 和分析所用的结构化数据一样关键”。甚至那些还在计划或正准备开始实施大数据分析程序的公司也应该前瞻性地确保能提前准备好应对数据集成挑战。他强调，“你需要事先做好准备，如果有来自社交媒体的大数据量输入需求，还应该早点做好预算和增加人员”。

医疗培训行业数据集成案例分析

美国 UMA 医学院(Ultimate Medical Academy)的 CIO 认为，在非常紧迫的时间内实施数据集成软件，成功的秘密就是直奔主题，避免繁琐的修饰，必要时还可以寻求外部经验的支持。

总部位于美国佛罗里达州坦帕市的 UMA 医学院是一家盈利性公司，提供在线或者课堂式培训，在全球有大约九千名学生。UMA 培训学生，使他们成为护理技师、医疗账单和编码专家、医药办公室经理、放射技师等等。

Pervasive 软件的利与弊

Pervasive 软件公司成立于美国德州奥斯汀，他们提供价格很有吸引力的数据集成工具，支持批量和近乎实时的数据交付方法，这方面表现非常出色，但是 Gartner 公司 IT 分析组最近的一份魔力象限报告中认为，该供应商确实还有一些提升空间。

Gartner 公司的这份数据集成工具魔力象限报告发布于去年 10 月份，报告发现除了固定价格的优势，Pervasive 公司还提供高度可扩展的产品，提供了强大的映射和转换功能。

但是，Pervasive 公司的一些客户经常说该公司的软件有许多 bug，而且抱怨他们通常必须等下一版本发布才能修复这些 bug。另外，Pervasive 公司也不

提供对数据联邦功能的支持。

从该报道还可以看出，客户们经常提到的其它缺点还包括元数据和建模功能欠缺，相比于较大的竞争对手，Pervasive 的产品还是稍显不足。在 Pervasive 公司发布的第十版本中，已经针对上述的问题进行了一些改进。”

UMA 公司 CIO Sam Collier 透露，该校维护了比较有限的项目范围，去年邀请了一些重要的供应商做支持，因为公司需要马上购买和启动一套数据集成软件平台。UMA 需要数据集成工具把一个新学习管理系统中的信息与内部开发的应用做信息合并，项目团队有大约一个月时间来完成这项工作。

Collier 说：“我们一直在持续投资来增强我们对学生支持的功能，改善学生学习效果。更及时地获得这些数据并与内部应用程序相整合是关键组成部分。”

学习管理应用程序加强了学校的在线培训可能，该应用是由 Blackboard 公司提供的。该校选择 Blackboard 公司的托管版本，运行在 Oracle 数据库上。而 UMA 的学生信息系统和其它 nebulizer 系统是运行在微软公司的 SQL Server 数据库上。

UMA 的 IT 团队对于 Pervasive 软件公司的数据集成平台有一些体验，他们很快就判断该平台可以让 Oracle 与 SQL Server 数据库完美结合。

Collier 说：“我们没有花太多时间考察其它不同产品，在我们看了 Pervasive 公司的产品以后，发现它实现的功能正是我们需要用的，我们做了一个

快速试用之后，就确定它基本上能满足我们的需求。”

Pervasive 产品有几款方案可供选择，但是时间很短，Collier 不想让项目扩展超出控制。该团队决定只关注完成整合 Blackboard 和内部应用程序必要的基本特性和功能。

该团队还获得了来自 Pervasive 软件公司支持团队的实施帮助，大部分是通过电话或者在线方式进行的。Collier 说，审查和购买 Pervasive 公司产品花了十天时间，测试运行通过又用了两周时间。

Collier 说：“我们没有试图完成一切，而只是实现范围很窄的从一组 Oracle 数据库表到一组 SQL Server 数据库表的转换，关注点控制在较小的范围内就不会导致范围蔓延。我们还借助了 Pervasive 公司的专业服务，从根本上为我们加速了初始整合的实施和执行。”

虽然 UMA 数据集成软件实施完成的相对较快，但也并不是说没有问题。团队运行软件时面临的问题之一集中在大数据集上，UMA 计划从 Blackboard 应用程序中把这些数据集抓取过来。

Collier 解释说，除了提供托管虚拟教室的平台，Blackboard 应用还跟踪了与学生学习活动有关的数据点。例如，系统维护了每次学生登入在线课程的记录，或者完成具体相关课程任务的记录。UMA 的教职人员可以使用这些信息跟踪学生的进度，弄清楚哪个学生需要哪些额外的注意。

但是在实施过程期间，从 Blackboard 提取出来的数据集非常庞大，花了很长时间进行处理。Collier 和他的团队不得不与 Pervasive 公司一起研究解决方案，否则整个系统性能将持续下降。

Collier 说："我们不得不花几天时间与 Pervasive 公司工程师们一起研究，让他们研究数据并且与我们一起协作。最大的问题实际上是如何能让查询能高效地返回结果。"

此外，Collier 的团队还与 Blackboard 员工密切合作，完善了查询功能，使查询能够高效运行。

他说："我们让 Blackboard 提供必要的服务器配置变更和优化，使得查询能在他们那一段高效执行。这让我们可以在合理的等待时间内能得到查询结果。"

项目成果

自从运行新托管 Blackboard 应用程序和 Pervasive 数据集成平台之后，在没有 Oracle 数据库专家的情况下，UMA 的 IT 团队已经可以把 Oracle 数据和他们的 SQL Server 基础设施完美结合了。

该系统还提供给学生服务团队以更大的灵活性，因为来自 Blackboard 应用的学生活动数据能以比以前更快的速度提交过来。Collier 表示，在以往的学习管理系统中，学生活动数据是通过平面文件分一天几次传送过来的，这种处理混乱而且难以管理。

现在，Pervasive 公司的软件可以让来自 Blackboard 公司 Oracle 数据库的数据近乎实时地复制到 UMA 的 SQL Server 部署的环境中。结果，现在的数据比过去更及时，学校员工可以更容易地获取需要的报表。

Collier 说：“这使我们能更高效地获取学生参与信息，我们可以提前获得这些信息，只有很少的延时。IT 架构的变革改善了公司的业务流程，使得我们可以为学生提供更好的服务。”

Dell Boomi 数据集成工具案例分析

从 Siebel 系统迁移到 SalesForce.com 需要将现有的数据迁移并集成进中，Panasas 公司向我们介绍，Dell 的 Boomi 提供了比 Cast Iron 系统更好用的工具。

Panasas 公司最早是在大约 3 年前开始使用 Boomi Atomsphere 这个云服务平台的，那是在 2010 年 Dell 计算机公司收购 Boomi 和 IBM 收购 Cast Iron 之前。

“我们很容易在 Boomi 环境中积累经验并建造经过概念验证的数据集成，”Panasas 公司信息系统总监 John Lake 说：“我们跟 Cast Iron 开始有一些初步的关于成本、价格以及何时完成部分功能并上线运行的接洽，但是正当我们跟 Cast Iron 沟通的同时，也开始着手用 Boomi 构建系统。我认为快速实施能力给我们带来了竞争差异。”

去年 4 月，TechTarget 还采访了两个 Boomi 以前的用户，他们最终发现自己的 SaaS 集成需求太复杂太耗费时间。但 Lake 说他在 Boomi 方面的经验起了积极的作用。

“我们在三天之内就完成了第一个集成”，Lake 说：“我们特别喜欢的一点是它内置的 Connector。我们只需要选中连接器并拖拽进工作流中即可。”

数据集成工具帮助主流 CRM 升级

Panasas 成立于 2004 年，是一个存储硬件供应商，该公司的专长是高性能计算。其客户包括石油天燃气公司，制造企业，研究院所和其他有着快速存储性能需求的公司。该公司常常与按照客户需求建造 Panases 存储应用的签约第三方厂商一起工作。

Lake 首次加入 Panasas 是在 2008 年，他进公司不久就被分派了重整整个组织的应用架构的任务。其中的第一步，公司想将整个 CRM 系统从 Oracle – Siebel7 迁移进 SalesForce.com 中。

“我们内部的销售团队正在使用和权衡 Salesforce.com，考虑到领导管理，领导资格和类似的方面，最后决定是将所有功能集成进同一个 CRM 平台”，Lake 说：“所以，我们调研了许多不同的供应商，最后决定选型 Salesforce.”

Lake 说 CRM 升级项目最大的挑战就是要维护几个点与外部合同制造商的集成需求。这些集成要能支持客户定单、发票和状态报告的自由交换——而且还要保护好数据。

“我们知道，除了要集成 CRM，还要集成许多其他系统，如 ERP。我们需要有某种工具来构建和运行这些集成任务，” Lake 说：“我们也知道我们并不想用客户的方法重做一遍，如写脚本，运行 Cron 作业等等类似的工作。”

最初是 Saleforce.com 的销售代表首先建议使用 Boomi 的。Lake 说 Boomi 系统相对容易实施，而且可以支持 Panasas 与 Salesforce.com 和外部合同厂商

之间的双向连接。据 Dell 表示，Boomi 现在还能支持 Panasas 查询 Salesforce.com 准备发送给厂商的订单;也支持当厂商完成了一个产品并且发布之后，或者收到新订单时，由 Panasas 更新 Saleforce.com。

Dell 计划收购 Boomi 时，Lake 起初最关心的问题是，他能从数据集成供应商处得到的支持和获得的个人关注要减少了。但实事上这种担心并没有发生，客户与 Boomi 支持团队沟通仍然是相当畅通。

Boomi 的 CTO Rick Nucci 表示，Dell 收购 Boomi 并决定将其做为云战略的重要一环。收购后，在保持 Boomi 在设计和开发独立性的同时，两家公司一直在努力做好在支持、销售和财务运营方面的整合。

“我们作为独立的业务单元高效地运营着，” Nucci 说：“但是年末的时候我们的规模就将翻番，Dell 正支持我们做我们想做的事情。”

Nucci 透露 Boomi 投资项目里，最近的追加投资项目包括：处理超大数据集的新功能;允许用户在云或中间件环境中与内部和外部系统连接的连接器功能等。

数据集成项目需要全局考虑

Lake 表示，做数据集成项目一定要有全局考虑——尤其是当与外部供应商集成的时候。他说：“对于你正在做的事情一定要有更广泛的观点，你需要端到端的思考，而不只是点到点。”

如果头脑里没有一个大图，当项目范围扩展时，无疑会带来兼容性问题。让第三方业务合作伙伴在每一步中都了解需求也很重要，因为从长远看很少有需求是静态不变的。

变化是不可避免的。当从一点迁移到另一点时，也许要增加一个字段，可能要改变字段的格式，或者你可能要按照这种方式引入新的集成点。 当处理第三方的事情时，你要跟他们有一个良好的、开放的、清晰的沟通，不仅讨论今天的需求是什么，而且还要对未来有什么需求给予足够的提前关注。

案例分析：云服务与数据集成的结合

目前，许多中等规模的企业已经发起了向云计算的进军，他们需要一种数据集成解决方案来将基于云服务的数据和按需定制的系统连接起来。这一目标不仅对分离的数据有意义，而且能实时获得访问，来创建新的服务。

以 AWPRx 为例来说，它是佛罗里达州 Altamont Springs 的一家提供软件即服务应用的供应商，管理着工人补偿处方的制药豁免。在不到一年之前，该公司通过手工编码增强了一套数据集成解决方案，把公司使用 10 年之久的 Oracle J2EE 数据中心链接到了 Salesforce.com 公司的云计算空间。

然而，几个月前，CEO Jay Roy 实施了亚马逊 Elastic C2 存储解决方案，其中采用了 Jitterbit 公司的新版集成软件，该软件是为在 Salesforce.com 公司的平台运行而设计的。Roy 说，APWRx 现在完全处在云服务中，支持实时审批处方。

Roy 说："Salesforce 是这一切发生的地方，但是 Jitterbit 公司支持了 Salesforce 与世界各地进行交流。我们的客户一直在与我们交流。我们的工作是控制我们客户的流程，以便（受伤的工人们）能得到必要的医疗服务。"

它的工作流程是这样的：一名受伤工人带着他的处方去药店，药店登录到 AWPRx 来核对共付额并判断该处方是否已经被授权，还是需要先进行审批。AWPRx 的商业客户向该公司提供关于新伤者享有医疗福利资格的信息；Jitterbit

获取到这些信息并把它导入到 Salesforce，在 Salesforce 中该信息对网络中的六万家药店都是可用的。

Roy 说：“Jitterbit 公司允许我们知道哪些受伤工人享有医疗福利，而哪些没有，还可以使我们的客户可以接到药店活动的通知。在工人们的公司，主要的作用是对欺诈的管理和遏制。”

他补充说，如果能够从业务方面整合数据，知道工人们在哪里受伤对药店网络很关键。工伤补偿是一种反应性的业务模式；45%的受伤工人在通知发生之前会寻求治疗。他说，到了医生的办公室工人们才能知道这是因工作而导致的受伤，有可能符合他们工人们的薪酬制度。Roy 说：“他们已经寻求治疗，每个人都在做出反应。这就是我这里混乱的源头。”

迈向数据集成的第一步可能会比较慢

有关从云服务到按需系统之间数据集成解决方案的需求，在信息管理协会（SIM）波士顿分会中的参会者来说已经非常明显了，该会议有一场云计算专题。Michael Draper 是云服务供应商 Pega 系统公司的“平台即服务运营”全球总监，他认为，虽然云服务使得中等市场规模的公司可以在几个小时内就能获得服务，但是要把数据迁入云服务可能会比人们想象的要慢很多。

Draper 说：“一些云服务供应商不会允许你连接租用线路，因此你不得不使用公共线路。从现场向云迁移数据时，你可能花费几个小时甚至几天。”

Draper 在 SIM 分享了内容为“联邦快递将拿到一件存储设备，并把它（物理上）交付给云服务供应商，这样你就可以实现即插即在云中”的报告。出席本次专题会议的人很多，达到了 450 名成员，波士顿 SIM 分会是其他分会人数的两倍，大家都对云计算的人工传递网络（Sneaker net）付之一笑，Draper 对此的反应是“在与数据中心集成方面，业界仍然有很长的路要走。”

新产品易用性宣传

然而，数据集成工具的价格在下跌，易用性功能特性在增加，这使得它们对于中型企业更具吸引力。专家们说，总部位于加利福尼亚州奥克兰的 Jitterbit 属于一种新工具，可以使集成数据和业务流程更加直观化和图形化，不需要额外进行编码。Cast Iron 系统公司在今年早些时候被 IBM 收购之前，一直是 Jitterbit 的主要竞争对手。

CloudSwitch 公司有一种服务，支持 IT 部门从按需定制的系统传输数据到云服务，无需重新设计应用程序架构，或者改变系统管理工具。Iron Mountain 增加了它的数据管理能力，在收购了 Mimosa 系统公司之后也包括了数据集成功能；而

Pervasive 软件公司拥有它的 Pervasive 数据整合器，用于云服务或者按需系统中的应用开发。

Julie Hunt 是德克萨斯州圣马科斯的独立市场研究员和分析师，按照他的观点，他认为也有供应商专门为企业级大公司设计知名的集成工具，比如 Informatics 公司，但是小一点的公司可能不知道存在几十种集成工具，这些工具可以帮助实现从云服务向按需系统衔接的各种局部问题。

Hunt 说："不管采纳了哪种集成工具和方法，数据仍然是关键。"要理解该公司的数据问题，不只要关心近期项目或流程，而且要有长期规划并考虑企业的成功发展，这对理解哪款工具最好这一问题是最本质的。她说："直到最近，大部分公司已经采用数据集成解决方案来解决棘手问题，而这正是关注战略规划和利用的时候。"

有了 Jitterbit，AWPRx 已经削减了它集成预算的 80%以上，而且，通过把应用程序迁离它自己的 Oracle J2EE 数据中心，转向它云计算合作伙伴的服务器，彻底节约了内部培训，支持和开发成本。该公司共有 35 名员工，其中 IT 部门有 9 个人；三分之一的人整个工作时间都在 Salesforce；一个人做 Jitterbit 工作，调整流程，两个人专门处理亚马逊环境；而且他们三个人兼职做客户支持工作。

下一步是什么呢？Roy 说：“下一步是尽力以更直接的方式处理受伤工人。我们的工作是满足客户的需求，但是同时，我们的客户需要找到帮助（他们自己）需求的途径。我们希望工人们能通过获得的信息参与更多，而不是放弃信息。”

数据集成的五大最佳实践

根据 TechTarget 的 2011 年读者调查，随着“大数据”和越来越多的公司对实时商业智能(BI)需求的不断上升，大约 40%的数据管理员认为数据集成项目将成为他们今年面对的最高挑战。

为了帮助大家解决这一难题，TechTarget 记者采访了两位资深业内人士：数据虚拟化和云集成公司 Queplix 的首席技术官 Steve Yaskin，以及 Gartner 分析公司的研究主任 Bill Gassman，我们一同探讨确保整合成功的最佳途径有哪些。以下是他们提供的数据集成的五大最佳实践：

1.确保集成的数据是清晰可见的。电气和电子工程师学会(IEEE)最近的一项研究发现，典型的数据集成项目会将超过 40%的时间花在发现上，即清查组织内数据存储的过程上。

这就是为什么 Yaskin 会认为：考虑使用自动化的数据发现和数据字典管理工具是一个好主意，如 Queplix 和其他厂商提供的工具。Yaskin 说，最重要的就是在开始任何数据集成项目计划之前，获得数据源的可视性。

“在很多时候，我们看到客户，尤其是大型企业，他们有像 SAP 或 Oracle 那样大的系统，而且这些系统已经使用多年。随着时间的推移，业务拥有者对这些数

据来源的背后究竟是怎么回事，已经没有完整的清晰理解，” Yaskin 说：“获得这些对象的即时可视性，是第一个大步。”

2.用流程控制数据质量。软件供应商越来越多地把数据集成功能捆绑进数据质量工具。Gartner 公司的分析师预测，数据集成和数据质量工具的市场最终会合并成一个，这一趋势反映了确保数据质量流程与数据集成计划进行合并的重要性。

尽管客户对捆绑集成和质量工具的需求不断增长，许多公司仍然否认他们的信息产品销售惨淡。

“在他们能够以一种有目的的方法将一个或多个系统的数据集成起来之前，他们仍将不得不对数据进行清洗，” Yaskin 说：“数据处于什么状态。数据之间存在什么偏差?什么是空值?什么是日期范围?什么是数据启发?”

3.形成一个数据集成卓越中心。根据 Gartner 公司的调查，推出数据集成计划的公司可以获得巨大的收益，即通过一个小团队的共同努力推动项目前进。分析师表示，在理想情况下，这个小团队的成员将包括：来自技术方面的人员，知道如何使用集成工具;业务用户，基于新集成的数据运行分析报告;和公司决策者，即信息消费者。

“有组织战略或某种能力中心去真正保证战略的延续，确保技能的传递，以及确保雇员离职后的连续性，” 他说。

4.与云集成公司打交道时，要保留对信息的控制权。像 Queplix 这样的云数据集成和数据虚拟化厂商随处可见。他们包括 Dell Boomi，Composite Software，Informatica，还有长长的一个清单。正在考虑与这些公司进行合作的企业应该仔细再仔细地检查他们的数据管理策略。

“有一点是要确保的：你必须知道谁拥有你的数据，” Gassman 说：“因为如果你把数据转交给第三方，接下来出现第三方倒闭或者你想更换供应商，这时你就需要确保这种改变是很容易做到的。”

5.寻找适合企业用户的工具。数据集成问题不只是 IT 部门的问题。不断增长的普通企业用户通过所谓的自我服务的商务智能(BI)和分析应用程序卷入进来。组织是否正在寻求通过数据虚拟化或更传统的数据集成方式来组合信息，专家们说，寻求通过可视化的用户界面和拖拽模板来帮助企业用户的工具总是好的。

“这样一来，每当业务部门需要一种新的组合对象并在该对象上运行 BI 工具或运行报表时，他们就不再需要 IT 部门的帮助，” Yaskin 说。

企业数据库标准化最佳实践

现如今有许多企业计划将他们的 IT 应用整合到一个单一的数据库标准之上，因为这样做会降低 IT 复杂度和成本并简化操作流程。但是专家指出，尽管数据库标准化项目会给企业带来诸如清晰的数据能见度，但是在作出决定之前，企业还有许多的因素需要考虑。从零做起是容易的，但是大多数企业的 IT 系统都是多样化的，进行数据库标准化整合将面临难以想象的困难。

Ventana 研究机构的副总裁 David Menninger 表示，将所有应用全部迁移到一个数据库平台，从理论上来说是个不错的想法，但是实际情况也许并不像我们想象的那样。当企业至少使用两种数据库平台时，你才有和厂商谈判的余地，尽管购买一种数据库平台在成本上会节省一些，但是这就像是训练赛马一样，通常在练习时你会让其它的马做陪练，因为在有压力的情况下，才能让你的赛马跑的更快，在进行项目实施时也是一样。

数据库标准化 Vs.数据联邦

目前，企业正在面临着高速的业务和数据增长挑战，独立分散的 IT 系统和应用遍布企业的每一个角落。因此选择一个单独的数据库标准可以很好的解决这些挑战，并获得一个一致的、准确的、集中化的信息管理能力。而专家认为数据库标准化是一个不错的方法，但并不是唯一的方法。

数据联邦和数据虚拟化技术可以使用元数据层和逻辑层将不同的信息从分散的系统中抽取出来，并整合到一个集中的视图当中，同样可以帮助企业很好地解决业务挑战。

在数据库标准化整合与数据联邦技术之间做出选择，很大程度上是取决于业务需求以及现有的应用系统，而专家认为二者的适当结合是往往是最佳实践。数据库标准化的支持者认为，更多的整合可以带来更可靠的信息，而当业务变更时，更少的数据库会更易于管理。然而现在大多数企业出于某种原因，都有多种数据库来支撑不同的系统和应用。举个例子，一个大量需求的应用可能会绑定在一个特定的数据库平台或者业务单元之上，这样的情况下进行数据库整合造成的结果就是更高的成本和更长的业务中断时间。

当然进行数据库整合的企业还会遭遇技术难题，某电子商务网站的高级 DBA Andrew Kerber 表示：“你的企业在处理数据时是不是需要通过因特网?如果是的话，那么你就需要进行合理的配置才能获取正确的信息。你的企业是否有足够的设备来存储所以数据?如果你计划购买新的服务器，你必须确保它能够处理好工作负载。业务人员不可能给你这些问题的答案，所以最好的方法是自己去尝试。”

与此同时，数据联邦软件同数据库标准化相比，有着成本更小、耗时更少的优势。但是它需要你花更多的精力去同几家厂商打交道，而对于实时变更的业务需求，数据联邦的表现就不如前者更加灵活。Menninger 说：“力争标准化是一

个不错的目标，但是我认为企业的观点应该是不求最好只求更好。这意味着你可以尽量进行数据库标准化，但是同时在充分考虑业务需求的情况下，也可以尝试一些更合适的技术。”

无论企业选择数据库整合、数据联邦还是其他相结合的整合方式，Enterprise Applications Consulting 的创始人 Joshua Greenbaum 认为最重要的是让各种利益相关人员参与到整个项目流程中。

Greenbaum 表示无论做出怎样的技术迁移，各个部门的利益相关者都可以帮助项目发展，确保满足所有的需求，让业务连续良性的运转下去。

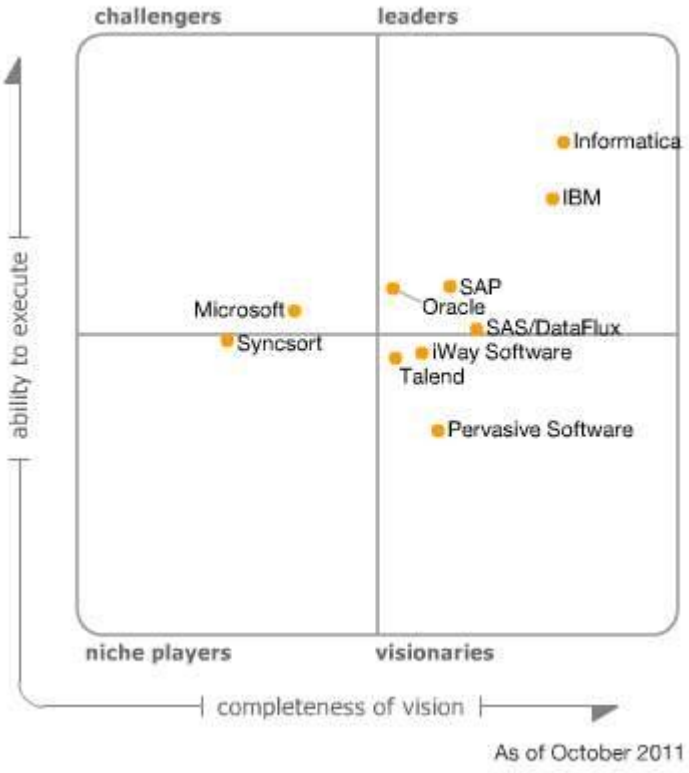
Gartner 解读数据集成工具选型最佳实践

分析机构 Gartner 近期发布了数据集成工具的魔力象限报告，分析师在报告中指出，由于主数据管理(MDM)、云计算、商业智能以及大数据分析的需求不断上升，数据集成软件市场将在未来一段时间内迎来高速的发展。

根据报告显示，截止到 2010 年 12 月，数据集成工具市场的总价值超过了 16.3 亿美元，同比 2009 年增长了超过 20%。Gartner 预测到 2011 年年底，这一数目将进一步提升 15%。

Gartner 信息管理分析师 Ted Friedman 表示：“我们看到数据集成软件市场保持了增长的势头，这来自于几个方面。越来越多的企业希望通过战略性手段来进行信息管理，因此他们正在需求更加流行更加强大的方式移动数据。”

Gartner 数据集成工具魔力象限是一份年度报告，其中将入选的厂商分为四个象限：领导者、挑战者、特定领域者以及有远见者。评选的标准涵盖了“前瞻性”以及“执行力”等因素。



Informatica 和 IBM 两家公司保持了领导者的位置，而其他位于领导者象限的厂商还包括 SAP AG、甲骨文和 SAS 的子公司 DataFlux。Gartner 将微软评为唯一的挑战者象限，而 iWay 软件、Talend 和 Pervasive 软件公司被评为有远见者象限。另外特定领域者只有一家公司——Syncsort，该公司成立于 1968 年，最早专注于大型机市场，此后又将业务扩展至分布式系统市场。

根据 Syncsort 公司全球销售总监 Josh Rogers 的介绍，尽管目前许多大型厂商都将数据质量以及 MDM 等功能添加到他们的数据集成软件产品当中，但 Syncsort 目前还是只专注于为客户提供性价比较高的 ETL 工具。他表示：“Syncsort 为用户提供了一系列的 ETL 工具，而这正是目前市场中其他厂商所不

具备的。因此我们还将继续在这方面开发新的引擎。虽然我们也在努力扩展产品的功能集，但可以肯定的是，我们不会像 Informatica 或者 IBM 的产品一样包含太多的功能。”

新挑战驱动数据集成工具快速发展

对数据集成工具需求的不断提升，主要源自于企业已经逐渐意识到战略性数据管理的重要性，特别是考虑到目前商业环境的快节奏变化。而这样的现状促使越来越多的公司开始实施主数据管理项目，以保证重要信息的一致性，如产品数据、客户数据等。Friedman 表示，数据集成工具的发展动力，还包括高效的信息共享。企业需要在部门之间、外部商业伙伴之间进行快速准确的信息共享。

与此同时，许多公司还在考虑使用基于云服务的应用，一方面是降低成本，另外一方面是减少内部系统管理的复杂度，从而将更多的精力放在提升核心竞争力之上。

Friedman 表示：“当企业将更多的应用迁移到云中，就需要对云中以及企业内部的数据进行同步。因此不能否认的是，云计算的普及也是驱动数据集成工具发展的一个动力。”

大数据是另外一个因素，大数据分析、大数据存储等在过去的一年中已经无数次被提及。要做到这一点，你需要对不同类型的数据进行收集、存储与管理，因此数据集成工具在这方面也将发挥作用。

Friedman 提醒用户，在考虑选择新的数据集成工具的时候，有几点注意事项需要牢记。首先要考虑的就是功能性，选择那些能够提供更多数据集成技术的厂商，因为业务的需求是时刻在变的。

在 Gartner 魔力象限报告中，Friedman 将 Syncsort 公司评定为特定领域者，因为他们只专注于 ETL 技术。而像 IBM 这样的厂商，他们的数据集成工具包含了 ETL、数据复制、变更数据捕获、数据联邦等功能。除非只想要 ETL 工具，否则企业在考虑数据集成工具时，还是要选择功能性相对完整的产品。

Friedman 还建议企业在选择数据集成工具的时候，考虑数据质量问题。对此他表示：“如果没有良好的数据质量功能作保障，企业是不可能成功实现数据集成任务的。因为数据集成并不是意味着将数据从一个地方移动到另一个地方，你还需要保证你所移动的数据是准确无误的。”

最后，Friedman 建议企业充分考虑主数据管理功能，因为它能够为你的数据源提供良好的透明度与能见度，你可能把握数据的流向以及传输的方式。这一点对于业务人员是非常重要的。

我们的编辑团队

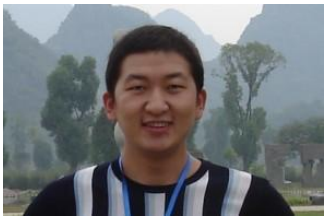
您若有何意见与建议，欢迎与我们的编辑联系。

诚挚感谢以下人员热情参与 TechTarget《商务智能电子书》的内容编辑工作！

诚邀更多的 BI 专业人士加入我们的内容建设团队！

曾少宁

TechTarget中国特邀技术编辑。软件工程硕士学学位，4年以上软件开发经验，熟悉Oracle、Java以及Linux等领域，曾经任职于juniper等著名企业，目前从事计算机教学工作。



冯昀晖

TechTarget中国特邀技术编辑。资深软件工程师，有超过7年的政府和企业信息化软件解决方案经验，熟悉SQL Server、Oracle等数据库技术，爱好阅读、健身和中国象棋。

孙瑞

TechTarget 中国高级网站编辑，四年网络媒体从业经验。负责“TT 数据库”和“SearchBI”网站的内容建设，熟悉数据库以及商业智能等企业信息化领域，拥有计算机学士学位。