SearchBI 电子书

Sep. 2012

SQL Server 2012

商务智能手册

SQL Server 2012 延伸自助服务 BI 概念，加强了易于管理性，重点关注可视化和数据探索功能，也为 IT 人员带来了全新的挑战。本技术手册将对其进行全面分析，并为读者支招。

- *解读 SQL Server 2012 中的最新 BI 功能*
- *微软 Power View：让 BI 报表变得更简单*
- *详细解读微软 HadoopOnAzure 的大数据处理功能*

TechTarget 中国

SQL Server 2012 商务智能手册

SQL Server 2012 自线上发布以来，获得了多方关注，其可视化工具 Power View 成为亮点，在数据质量和管理方面添加了多项全新功能，引起了很大的讨论。在本技术手册中，TechTarget 数据库与 SearchBI 网站编辑将为您解读 SQL Server 2012 的理念，分析其可视化与集成服务功能，并提供一些在大数据领域的应用见解。

SQL Server 2012 商务智能概览

如果 SQL Server 2008 R2 的重点是让商务智能(BI)的使用者像使用自助服务一样便捷，那么 SQL Server 2012 则是让自助服务 BI 这一概念延伸至让 IT 人员更容易进行管理。SQL Server 2012 线上发布以后，满足了两个极端的使用者。

- 解读 SQL Server 2012 中的最新 BI 功能
- 扩展 SQL Server BI 系统的三种方式
- SQL Server BI 系统硬件选择三大妙招

Power View 数据可视化

在去年十一月举行的 SQL Server 专家会议(PASS)上，微软公司推出了 Crescent 项目，即 SQL Server 2012 的 BI 核心。Power View 扩展了产品性能，提供给业务用户一个基于 Web 的环境来构建临时报表，提供了丰富的数据可视化搭配。

- 玩转自助式 BI：SQL Server Crescent
- 微软 Power View：让 BI 报表变得更简单
- 微软 Power View：丰富 BI 报表展现方式

集成服务（SSIS）

当微软公司发布 SQL Server 2012 与 SQL Server 集成服务（SSIS）更新版本的时候，BI 领域的人们都开心得不能自已。更新后的 SSIS 增加了很多新功能，使 BI 开发变得更加简单，大大提高了开发人员的开发效率。

- 针对 BI 开发人员的 SSIS 重要新特性
- SSIS 2012 新特性：连接管理器还有更多
- 更多 SSIS 技巧：使用参数与撤销操作

大数据

SQL Server 2012 中添加的一些新功能对于目前愈演愈烈的大数据形势非常有用，包括基于列的查询，基于 Excel 的分析和报表功能的增强，和高可用性方面的改进等。这样一来，SQL Server 2012 的 Apache Hadoop 集成自然成为了企业 DBA 的必修课。

- SQL Server 2012：大数据大问题
- 解读微软大数据组件 SQL Server for Hadoop 连接器
- SQL Server Hadoop：开拓大数据新疆域
- 详细解读微软 HadoopOnAzure 的大数据处理功能

解读 SQL Server 2012 中的最新 BI 功能

如果 SQL Server 2008 R2 的重点是让商务智能(BI)的使用者像使用自助服务一样便捷，那么 SQL Server 2012 则是让自助服务 BI 这一概念延伸至让 IT 人员更容易进行管理。

事实上，SQL Server 2008 R2 的某些特性表现出商务智能的现代化：适度简化先进的 BI 功能，使得像混搭数据的能力、创建关系和维度这样的过程，对任何 Excel 用户而言是友好和熟悉的。

“如果你了解一下 SQL Server 2008 R2 及其围绕管理、自助服务 BI 的主题 ，那么 2012 版本就是这一战略的延续，” 微软产品管理总监 Herain Oberoi 说：“所有这一切都是为了帮助最终用户和 IT 人员。”

在本月初进行了一次“SQL Server 2012 线上发布”，以满足两个极端的使用者。对于最终用户，专家们说这个版本的亮点是可视化工具 Power View(原先项目代号为 Crescent)，而数据质量服务、分析服务和主数据管理这些新功能旨在让 IT 人员对自助服务 BI 的管理和维护获得更多的控制并提供更好的性能。

虽然观察家们说在 SQL Server 2008 R2 中没有太突出的 BI 功能，但还是有很多人认为 SQL Server 2012 的新添特性使其成为了一个更引人注目的 BI 集成平台。考虑微软在.NET 和 Transact-SQL 等开发语言上的实力，SQL Server 开发人员能够在数据质量或主数据管理等关键领域中创建一套自己的功能集，但

他们需要专业知识和技能。

“如果说失去这些功能太震撼了，因为总会有些人会以自己的方式去构建解决方案，” 微软最有价值专家和 MarkTab 咨询公司的总裁 Mark Tabladillo 说。这可能适合于拥有许多 SQL Server 开发人员的大公司，但对于只有一两个身兼数据库开发人员和数据库管理员两职的人员的小公司而言，它不是最好的选择。微软提供一个包装好的应用程序，而不是提供工具包级别的功能。

Power View 的自我发现

SQL Server 2012 关注的 BI 重点将是 Power View，这个给微软平台带来高度关注的可视化和数据探索功能。当 SQL Server 2008 R2 发布时，它提供了基于 Excel 的 PowerPivot，这一工具让用户可以方便地连接不同的数据源和对数据进行混搭，而 Power View 是建立在这个前提下的，它允许用户使用拖拽式界面进行数据浏览去发现他们希望得到的东西。

“如果你想使用自我发现功能，你需要使用这个工具去发现数据的模式和趋势，”微软的 SQL Server 技术专家 Mark Kromer 说：“例如，用户分析公司的销售数据，发现了一个不应该踏入的特殊区域。Power View 让用户对数据进行切片和切块，找出产品销售或客户特征之类的东西，使他们能够确定他们是否在这个特定地区向错误的客户推销错误的产品。”

SQL Server 2012 新的数据质量服务(DQS)是另一个针对企业用户和 IT 管

理人员的特性。虽然以前的 SQL Server 版本都具备一些数据质量的能力，作为集成服务的一部分;但是即将发布的 DQS 是一个专门的数据质量服务器和客户端，针对个体用户可以运行在独立模式，也可以和 SQL Server 集成服务(SSIS)一起使用提供企业级的数据分析、清洁和匹配。“数据质量服务将允许你控制来自不同数据源的数据质量，而不是依靠编写自定义代码，” Marco Russo 说，他是专注于 BI 和微软技术的顾问、作家和讲师。他运营一个基于 SQL Server 的专注于 BI 的网站 SQLBI.com：“DQS 提供一套标准化工具，可以更快的获得完成工作所需的数据。”

主数据服务，这是 SQL Server 2012 改进后，与 SSIS 和 DQS 协同工作以帮助 IT 人员获得主数据结构的功能，提供对象映射、引用数据、元数据管理和维度层次管理。另一个为 IT 人员和最终用户设计的增强特性是 BI 语义模型(BISM)，一个通用元数据层，微软说这将涵盖所有的 BI 用户体验。它使传统的多维模型与关系模型并存于称为表格模式的格式中。

“现在 IT 人员可以将最终用户创建的 BI 报表和仪表板无缝地过渡到分析服务，” Oberoi 说：“你拥有所有这些最终用户在 Excel 中建立的数据模型，并将其发布在 SharePoint 之上。这将有助于这些解决方案的维护工作的规范化和管理。”

扩展 SQL Server BI 系统的三种方式

SQL Server BI 系统可能占用大量计算机处理能力，而且大多数组织也都会部署尽可能多的处理能力来使 BI 系统正常运行。不幸的是，商业智能的用户想要的更多——更多报表、更多仪表盘、更多数据。

所有这些都意味着你将最终需要更多的计算能力。换句话说，你将需要扩展你的 BI 系统使其有更多的处理能力。有两种广泛使用的扩容技术：纵向扩容和横向扩容。

横向扩容（Scale Out）

横向扩容意思是增加更多的服务器来支持整个系统。通常，你可以通过选择具体的 BI 服务，把一些服务转移到其它计算机上来实现这一点。例如，可能你的 BI 系统服务于组织内不同的用户;这样扩展以后每类用户群都能受益，因为各自 BI 任务都有自己的专用计算机来处理了。

或许你的 BI 系统由许多不同的组件组成，可以分离到不同的服务器上。实际上这很大程度上依赖于系统如何架构，以及用户如何使用它。一定要记得：横向扩容可能不会一帆风顺。它总会涉及到重新架构，这对于某些系统来说可能是不现实的。

纵向扩容（Scale Up）

这就是为什么纵向扩容常常是许多组织首选的原因。纵向扩容可以简单地理解为把 BI 系统迁移到更大的，具有更高处理能力的服务器上，或者对现有服务器进行升级。当你进行纵向扩容时，要注意下面一些通用原则：

BI 系统严重依赖于内存分析，如果有更多的内存和更高的处理器能力，它的获益最大。一般来说，你可以给系统的 RAM 越多，效果越好。至于处理器，越多通常会越快。换句话说，四个处理器插槽支持四个处理器比运行更小数量的插槽和千兆赫处理器运行起来会更快。

不幸的是，服务器很少有空间容纳更多处理器，也很少能支持更快的处理器。处理器升级通常会整个换新的服务器，那样价格会很贵。可以从考虑从内存升级开始，因为服务器在添加更多内存以后通常会更灵活。

严重依赖于数据仓库的 BI 系统最容易受到磁盘性能的影响。获得更快速的磁盘驱动(也就是说，使用有更高转速的磁盘)将很有效果，但通常不是最有效的方式。相反，请试试使用更多磁盘。磁盘速度通常会归结为磁盘位能被旋转盘片以多快的速度读取，盘片数越多(也就是阵列中的物理驱动数越多)，通常就意味着更多的容量，表现出来就更快。

传统磁盘固然会遭遇瓶颈，所以你也可以考虑固态硬盘(SSD)缓存系统。我曾见过最有效率的是在你的数据库服务器上以软件分块的形式运行，用固态硬盘结合 SAS 或者 PCI Express 扩展卡(范围在 150GB 到 300GB)。我见证了这种

简单的升级就带来了相同负载下三倍的性能提升，而且这次升级并不昂贵，大约每台数据库服务器 8000 美元。

在 BI 系统中，更大通常意味着更好

不管你拥有的是哪种 BI 系统——数据仓库，内存分析或者是兼而有之的混合系统，更多的计算资源通常会带来更好的性能和处理更大负载的能力，所以我们需要更多处理器，更多磁盘和更多内存。

如果获得更多资源的开销赶上采购一台新的整机服务器，那就买一台大的吧。要确保你能获得大量内存，获得尽可能多你能给新服务器买得起的处理器内核。在这方面投资将会获得服务器长期持久的汇报，纵向扩容有满足更高需求的能力。

SQL Server BI 系统硬件选择三大妙招

在中小企业的商业智能系统中，有大部分是基于 SQL Server 平台并且是预构建的，这意味着系统性能在很大程度上依赖于他们的数据库架构，然而就是因为这样，它们往往是很难进行更改的。

这样方式的好处是你不需要再为复杂的数据仓库设计而烦恼，并且能够省下一大笔资金。然而缺点也是显而易见的，性能方面你没有办法去把握。这时候你该怎么办?我们可以动一动硬件的主意，让你的 SQL Server BI 系统能够征服更多的挑战。

单独的服务器来运行 BI 系统

最坏的计划，就是你打算让 SQL Server BI 系统运行在已经运行了其他产品的服务器上。而想要购买一个大型的、性能超级的(不算存储设备)服务器就要花去你 3 万美金。如果这么大的开销是你无法接受的，那么一个折中的方案就是听取 BI 厂商的建议，选择性价比最高的服务器。然而无论你怎么选择硬件产品，记住一定要单独使用一个服务器来跑你的 BI 系统。

我们知道，大多数 BI 系统都包含了多个组件：一个数据库引擎、一个分析引擎、一个 Web 服务器等等。如果在此基础之上再添加其他的应用，那么你还期待你的 BI 性能能够好到哪里去呢?

有一些 BI 系统允许你将一些组件运行到多个服务器上，如果你的 BI 性能是最关键的因素，那么这样的产品就是你所需要的。更多的服务器来运行同样的系统，那么性能比然会更好一些。

RAM！RAM！

购买一个新的服务器，最多能有三分之一的钱都是花在了 RAM 上，或者内存上。在这方面最好不要投机取巧，SQL Server BI 系统中的数据库引擎需要内存，但是需要的方式非常不好。缩减四分之一的内存对于系统来说就是降低了一半的性能，而这就取决于你的数据库引擎、BI 平台以及其他涉及到的组件。

如果你接受不了将你的 BI 服务器添加更多内存，那么至少你要购买一款能够支持更多内存扩展的服务器。而且现在越来越多的中小企业 BI 系统都倾向去使用内存分析引擎，过去传统的数据库和数据仓库已经显得有些过时了。内存分析顾名思义就是在服务器的内存中，进行快速的实时分析操作。这表示在未来的企业 BI 系统中，你的服务器肯定需要更多的内存。

硬盘：更快更多

建立在数据仓库上的 SQL Server BI 系统需要它们的硬盘性能是卓越的，那么这意味着你需要的是更多的硬盘而不是更大的硬盘。这与内存恰好相反，如果你想要 1TB 的硬盘空间，你肯定不想让它在单一的 1TB 硬盘上。理想中，每块硬盘

大小为 100 GB 左右是最合适的。这是因为硬盘在性能方面有它们的物理限制，所以把数据放到更多的硬盘上对于性能只有好处没有坏处。

在选择硬盘时，你应该注意的是硬盘转速和平均搜索时间，当然更高的转速和更短的搜索时间就意味着更好的性能，不要被厂商的宣传所迷惑，比如数据传输速度就是不需要考虑的。因为更高的转速和更短的搜索时间通常就意味着硬盘的数据传输速度更快。

玩转自助式 BI：SQL Server Crescent

在去年十一月举行的 SQL Server 专家会议(PASS)上，微软公司推出了 Crescent 项目，它是一款新的商业智能(BI)工具，专为使非技术用户能更容易地创建视觉效果良好的报表而设计。它是下一代 SQL Server BI 核心。

几个月后，我们仍然没有得到更多其它详细信息——Denali 的第一社区技术预览(CTP)没有出现 Crescent，CTP2 只是针对微软 MVP 的(MVP 必须签署相关保密协议)，而 CTP3 仍然在建设中。幸运的是，上月底在亚特兰大举行的 TechEd 北美大会上，微软公司在几场培训会上展示了其 Crescent 产品。从我见到的情况来看，我认为大家的等待是值得的。

首先，我们来讨论一下 Crescent 背后的设计目标。正如其口号所说的“面向大众的 BI”，微软希望提供这样一款工具，它能易于使用并且能为数据视图提供强大的实用的视觉效果。在设计界面时，该公司设立了非常雄心勃勃的目标：所有的功能都只要通过一次或两次点击就能完成。

一旦你用上了这款工具，你就会发现微软公司确实做到了。你打开一个模型，Crescent 会读取该模型的元数据，分析其中的关系并给你展现一些查看数据的选项，比如查看表和字段域。例如，Crescent 可以检测父子关系，你可以有几种途径钻取到子记录。

让我们来看看技术细节。考虑到 Crescent 是 PowerPivot 的下一代产品，

或者正像微软公司给它的称呼，它是 PowerPivot 和 Excel 之间的连接点。因为它设计的就是要在 SharePoint 中运行的，但 SharePoint 企业版价格昂贵，而且安装也不简单，许多人把它看做是一种潜在的障碍。但是既然微软公司使用 SharePoint 来把各种 BI 服务连接到一起，你迟早不得不接受这一点——SharePoint 就在那儿。

有了 Crescent，你就有了一个直观的报表设计器，可以构建和预览报表。一旦你部署了某个报表，用户可以在浏览器中查看它，也可以运行在应用框架 Silverlight 中，Silverlight 是微软公司针对 Flash 的替代品。

至于数据源，我们有几种选择。你可以在已经上传到 SharePoint 的 PowerPivot 模型中访问数据，或者也可以使用分析服务中的 BI 语义模型 (BISM)。(BISM 是一种可以使在 Crescent 中构建 BI 模型更容易的技术。)这两种模型提供了对数据最快速的访问。你还可以直接查询 SQL Server 表。但是如果你想这样做的话，微软公司推荐你使用列存储索引，它是微软公司采用 VertiPaq 技术实现的超级搜索功能，它可以保证查询的快速响应时间。

微软公司把 Crescent 称为“交互式数据研究和可视化展现体验”。在看了 Redmond 程序经理们在 TechEd 技术大会上的演示之后，我非常赞同这种提法。构建报表和分析数据看起来如此的有趣。除了所有常见的 BI 工具，你还能获得一个数据模型，会展示给你度量和可用域。你可以快速修改图表类型，也就是你查

看数据的方式。你可以使用与在 Excel 中相同的图表——柱状图，条形图，稀疏图以及其它。要加上过滤功能也很容易：传统过滤功能中你只能“全选”或者选择单独的条件，但是现在你可以在你的过滤器中输入过滤条件，省得在几百个项目中滚动查找。

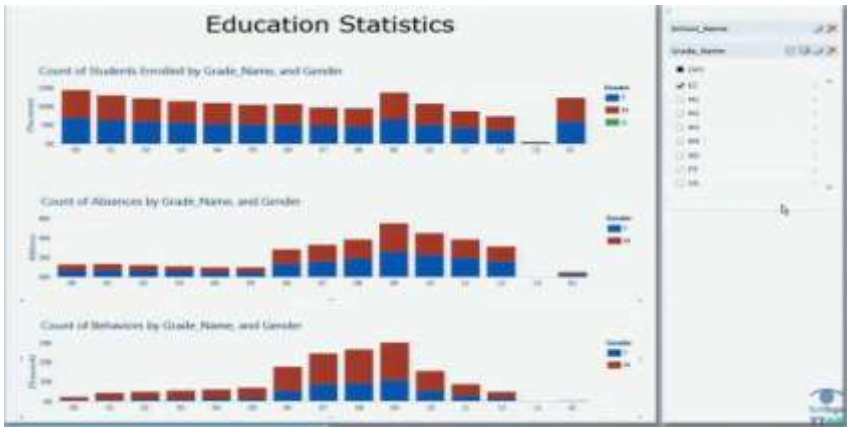


图 1。该图展示的是 Crescent 的报表设计器。该报表展示了三栏图标，每一栏展示了一种不同的度量值，还有过滤器。该图也按性别做了区分。

Crescent 还有几个令人印象深刻的视觉功能。你可以设置父记录按区块显示图像。例如，你可以滚动你的类别显示为区块。当你将鼠标悬浮到它们之上时，Crescent 会在明细区域显示分类明细，如图 2 所示。

图 2。该图展示了 Crescent 把分类按区块显示的效果。如果某个分类有图像，它会显示为图表形式。当你选择目录时，Cerscent 会自动显示分类细节。

下面将介绍 Crescent 的另一个优秀功能：当你浏览图表时，你可以点击图例条中的一个，Crescent 会自动按选中项进行过滤并高亮显示明细。不过，最吸引人的特性可能要数图表的动画展示了。把图表类型修改为散点图，然后设置值播放轴，可以设置一个时间度量，比如一年或者一个季度。在 Crescent 播放动画时，这些散列点会在图上移动，这样你就可以直观地看到所有季度或者年的取值段趋势。

虽然 Crescent 令人印象深刻，但它的定位并不是替代 Visual Studio 的报表构建器或者报表设计器，而是对它们的补充。Crescent 可以让你不费吹灰之力创建交互式的可视化报表;所有的工作都交给 Crescent 做了。但是轻松易用的代价就是你不能对最终报表的布局有完全绝对的控制了，不可能像 Visual Studio 和设计器那样的灵活。因此，你可能不得不等待将来的版本提供更多查看数据的方式或者对整体布局的更多控制了。如果你需要在你的报表中做复杂运算的话，你还是最好使用成熟的，经验证好用的工具吧。

总体而言，Crescent 的预展示证明了其承诺，微软公司似乎实现了它的目标：以互动的方式提供简单快速的方法显示数据，同时帮助实现了“面向大众的 BI”。

微软 Power View：让 BI 报表变得更简单

随着 SQL Server 2012 的发布， 新工具 Power View 也受到了极大关注，它是 SharePoint Server 2010 的一个报表插件。Power View 扩展了微软公司的 SQL Server 商业智能（BI）产品，提供给业务用户一个基于 Web 的环境来构建临时报表，提供了丰富的数据可视化搭配。要构建 Silverlight 应用框架，Power View 支持用户创建并与多个数据视图交互以方便信息共享、决策和数据分析。

与 Report Builder 和 Report Designer 这类 SQL Server 报表工具不同，在 Power View 中创建报表与在 Excel 中创建数据透视表和数据透视图类似。Power View 用户不必在视图和设计模式之间来回转换来查看他们修改后的结果，他们操作的数据会及时更新。通过几次点击，他们就可以创建多个报表，并为目标数据提供独特的视图。

Power View 数据模型

要使用 Power View，用户必须从 SharePoint 站点运行应用程序。也就意味着要支持 Power View，必须同时运行 SQL Server 和 SharePoint Server。此外，必须设置至少一个数据模型，以提供报表必须的数据。数据模型隐藏了源数据，这样用户无需知道数据内部结构、托管数据的服务器和与数据使用相关的安全问题就可以使用数据。用户看到的只是数据实体的一个简单列表，如图 1。

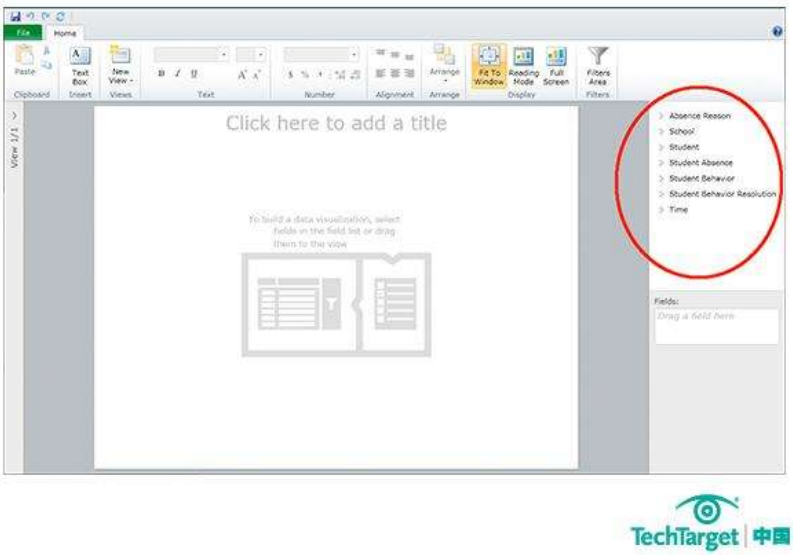


图 1：数据模型对于 Power View 用户是自动可用的。

该图展示了第一次运行时 Power View 设计窗口显示的样子。数据模型位于右上角的面板中。要注意，默认情况下，只列出模型实体名。但是用户可以扩展每个实体以便于访问其字段。用户访问数据无需采取其它步骤。本质上，数据模型为基于该模型的所有报表提供了基本构件块。

用户总是会从数据模型启动 Power View。该模型可以被创建到两种位置：SharePoint Server 文档库或者 PowerPivot 库（一种 SharePoint 文档库专门类型）。你可以在 Excel 中的 PowerPivot 或者在 SQL Server 数据工具中创建数据模型。模型所基于的数据可以来自 SQL Server、DB2、OData、Oracle 或 Teradata 等。一旦你设置好了必要的数据模型，用户就可以在 Power View 中创建报表了。

Power View 可视化

Power View 最大的优势点在于其 Silverlight 界面。要创建报表，用户要使用支持 Silverlight 的浏览器连接到 SharePoint 站点，从文档库或者 PowerPivot 库找到一个数据模型，然后从该模型运行 Power View。当 Power View 窗口出现以后，它们会切换到设计模型(参见图 1)并选择他们想在报表中包括的字段。从数据模型实体列表中展开实体，选择字段，然后点击紧挨着字段名的复选框。用户还可以把字段名拖拽到字段列表中或者拖动到设计界面。

例如，图 2 展示了我创建的一个报表，是基于“Contoso 学校”的（微软公司的一个示例学校区域，您可以在 MSDNblog 访问它以及其它样例模型）示例数据模型做的。

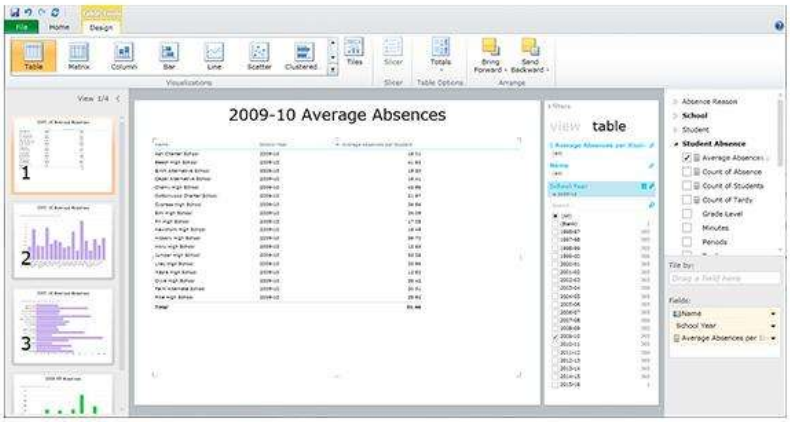


图 2：该数据模型支持你在 Power View 中创建基于表格形式的报表。

该报表是一个基本表，使用了下面三个字段：

- 名称（在学校实体中）
- 学年（在时间实体中）
- 每个学生的平均缺勤数（在学生缺勤实体中）

我把这三个字段添加到了表中，然后用学年字段过滤，这样就能只查看 2009 年到 2010 学年的信息。然后重新调整该表和列（简单的拖拽操作）形成信息的最佳展示。剩下的最后一步就是添加一个标题。

这就是我创建这个报表需要做的所有工作。如果这不是示例数据模型，而且我拥有对 SharePoint 站点的访问权限，我可以把报表保存并与其它业务用户共享。但是要记住，Power View 报表被保存为 RDLX 文件，该文件只能在 Power View 中查看。该报表与在 Report Builder 或者 Report Designer 中创建的报表不兼容，那些工具保存的是 RDL 文件。

微软 Power View：丰富 BI 报表展现方式

作为一款 BI 工具，Power View 使得在数据模型中创建数据的不同视图变得非常容易。例如，图 1 展示了我创建的一个报表，与我在前面的报表中使用的字段相同。我只是把原报表复制了一下，把表格展示改为了列图表展示。Power View 保留了我的字段选择，过滤行和标题。整个过程只花了三十秒时间。

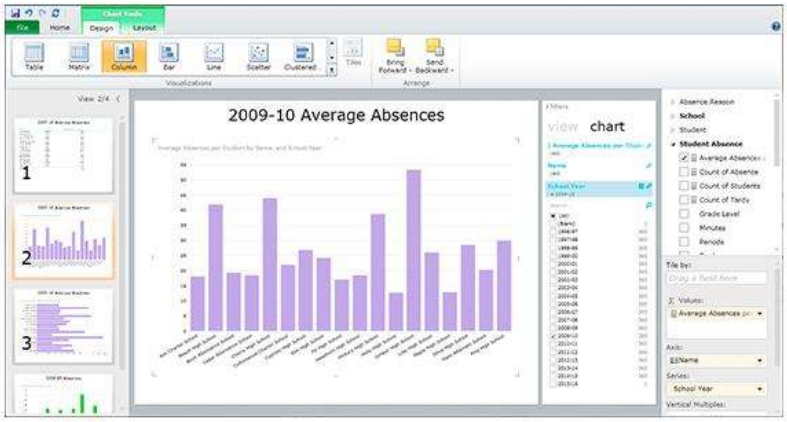


图 1：您可以基于原表创建列柱状图。

我按照同样的流程创建了报表，如图 2，复制第二个报表把列图表改为柱状图。Power View 使得替换展现效果变得非常容易，你可以尝试许多选项，选择最适合的配置。

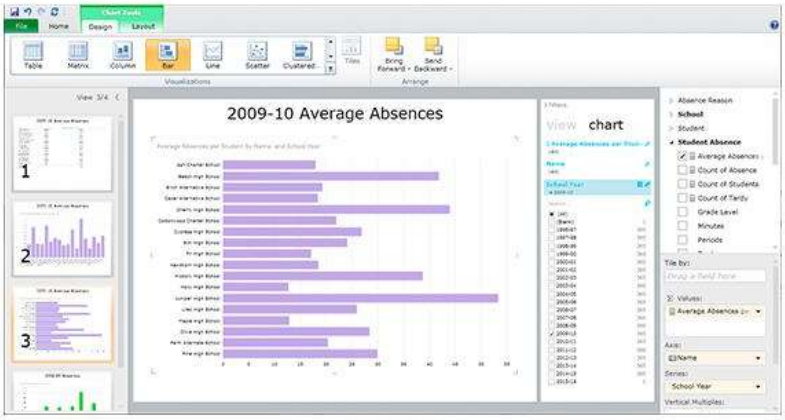


图 2：列图表可以很容易转换为柱状图。

前面我创建的这三个报表都是基于相同的数据产生的，但是它们是以不同的方式展现的。我可以很容易地修改数据。在图 3 所示的报表中，我又复制了前面的报表，同时把柱状图改回了列图表。然后，我修改了过滤条件，学年字段只包括“2008-2009”学年，而不是“2009-2010”学年。我还从图表中去掉了每个学生的平均缺勤数，增加了缺勤数合计字段。

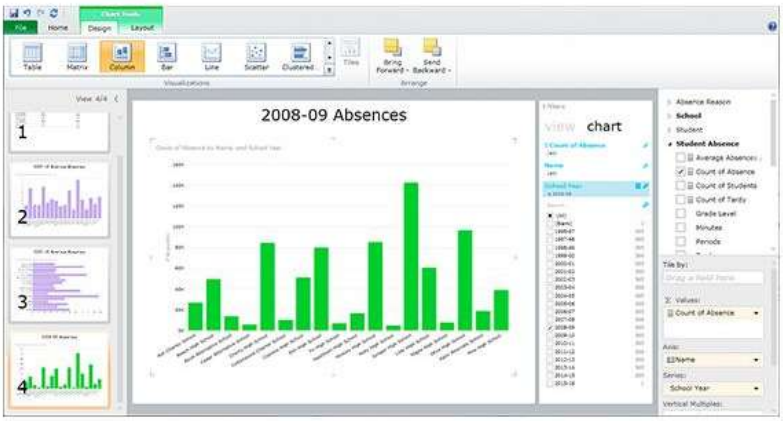


图 3：在 Power View 中你可以很容易地修改显示字段和过滤条件。

选择 Power View 需要考虑的事

Power View 使得创建报表更容易了，可以给数据实现各种视图。我前面展示的那些报表只是可以展示给客户可视效果的一小部分。除了可以用图表和表格显示数据，Power View 还支持诸如排序、数据高亮显示、交互式气泡图，交互式幻灯片展现，给具体细节视图突出显示或特写等。最突出的优势在于没有设计时和运行时的差异。你在创建报表时看到的就是其他人将在查看报表时看到的，包括实际数据都一样。

然而，Power View 也不是没有局限性。首先，它同时需要 SQL Server 和 SharePoint Server。如果你的组织只部署了 SQL Server，而没有部署 SharePoint，而且你又没有考虑增加另一个应用程序的开销，那么你的运气就不太好了。没有这两款产品，就没有 Power View。即使你愿意部署这两套产品，安装流程也会比较复杂。这个过程绝对不是开箱即用的操作。Power View 还依赖于微软的 Silverlight，也就是说 iPad 用户不能使用该应用产品。但是，如果这些限制对你来说不是问题，你就会获得 Power View 的全部优势，你的业务用户可以简单快速地分析和传递数据。

针对 BI 开发人员的 SSIS 重要新特性

开发人员总是在寻求使工作变得更轻松的方法。这一点对于商业智能（BI）社区的所有人都一样。因此，当微软公司发布 SQL Server 2012 与 SQL Server 集成服务（SSIS）更新版本的时候，BI 领域的人们都开心得不能自已。更新后的 SSIS 为保证开发人员的效率，增加了很多新功能。而其中五项功能将使 BI 开发变得更简单，大大提高开发人员的开发效率。

项目连接管理器

自从 SQL Server 2005 发布以来，SSIS 就能支持范围很广的连接管理器。这些管理器支持你访问来自多种数据源的数据，比如文本文件、关系数据库、分析服务数据库等。然而，你总是必须在程序包环境内创建那些连接管理器，也就是说，只有这一个程序包可以使用它们。即使多个程序包需要相同的连接管理器，你也必须为每个包重新创建一遍；不管有多频繁，都需要重复劳动。

在 SSIS 2012 中，一切都改变了。现在，你可以在一个项目中创建对所有程序包可用的项目连接管理器。创建项目连接管理器的过程与创建程序包连接管理器的过程类似，不同之处在于你创建它们的位置。对于项目连接管理器，应该使用“解决方案资源管理器”中的“连接管理器”节点，而不是“SSIS 设计器”中的“连接管理器”窗口。

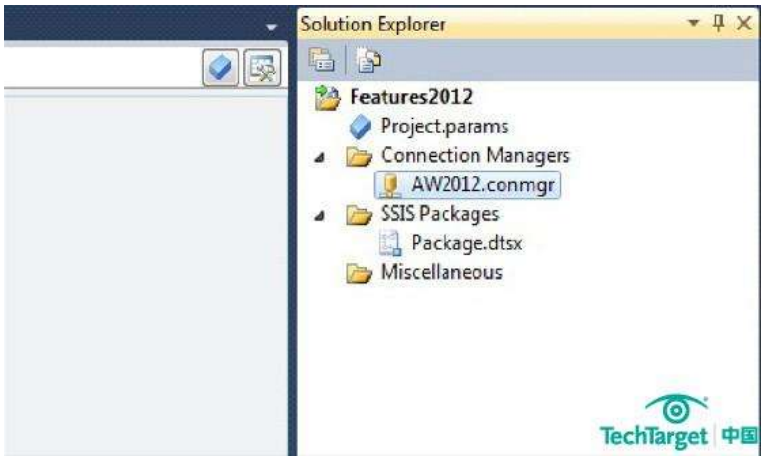


图 1.创建项目连接管理器

要创建项目连接管理器，请在解决方案资源管理器中的“连接管理器”节点上单击右键，然后点击“新建连接管理器”，如图 1 所示。这个操作会运行“添加 SSIS 连接管理器”对话框。在该对话框中，你可以为具体的数据源类型创建项目连接管理器，与创建程序包连接管理器时的操作一样。

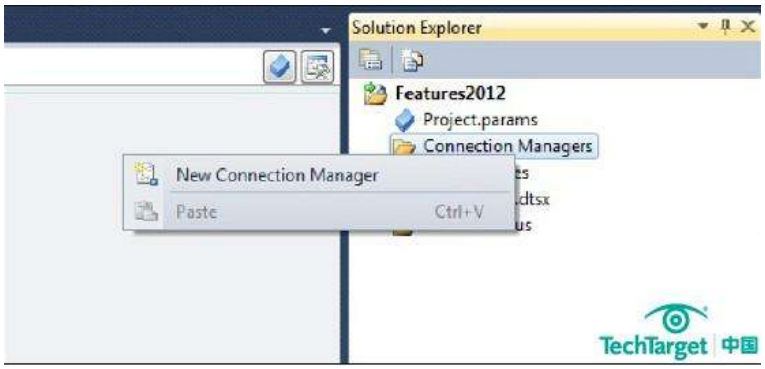


图 2.解决方案资源管理器中的连接管理器节点

创建了项目连接管理器之后，它会显示在“解决方案资源管理器”中的“连接管理器”节点下，如图 2 所示。该图中显示的连接管理器名称是“AW2012”。SSIS 自动给这个名称后面添加“.conmgr”文件扩展名。

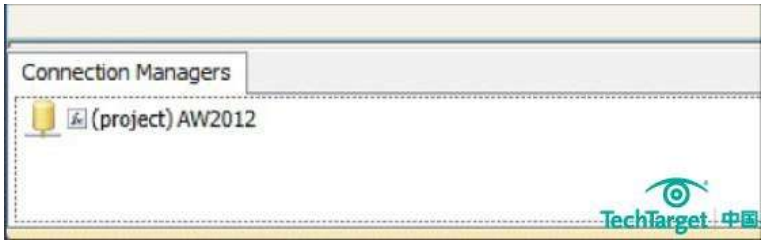


图 3.连接管理器窗体中的项目连接管理器

一旦你创建了项目连接管理器，就可以在项目中的任何程序包中使用它。事实上，每个包的项目连接管理器还会显示在“SSIS 设计器”中的“连接管理器”窗体中，不过连接管理器名称会多了前缀“（project）”，去掉了文件扩展名。例如，“解决方案资源管理器”中的“AW2012.conmgr”连接管理器在“连接管理器”窗体中显示为“（project）AW2012”，如图 3 所示。

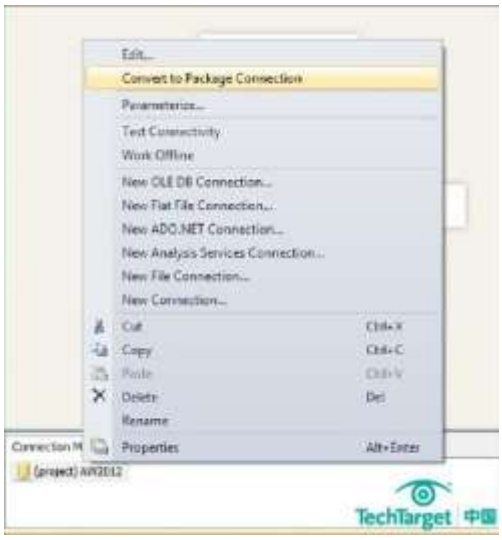


图 4.把项目连接管理器转换为包连接管理器

你还可以把项目连接管理器转换为程序包连接管理器。要实现这一点，请在“连接管理器”窗口中右键单击连接管理器，然后点击“转换为包连接”菜单，如图 4 所示。这个操作会从“解决方案资源管理器”中删除该连接管理器，并删除“连接管理器”窗体连接管理器名称中的“（project）”字样。

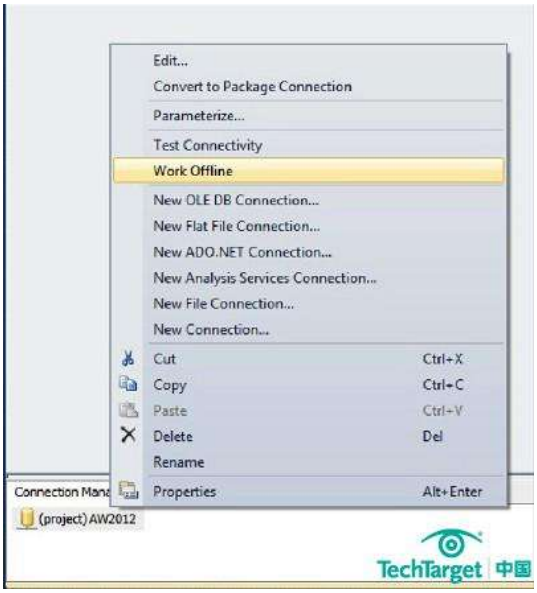
一旦你将项目连接管理器转换为包连接管理器，就只能在当前包中使用连接管理器了。如果还有其它程序包在使用该连接管理器，它们就不能运行了。然而，你还可以把包连接管理器转换为项目连接管理器。在连接管理器窗口中，右键单击该连接管理器，然后点击“转换为包连接”菜单，该连接管理器就对你项目中的所有包可用了。

SSIS 2012 新特性：连接管理器还有更多

SSIS 2012 有一个新特性就是，它支持关闭连接管理器的功能，这一点使开发人员的生活更轻松了。默认情况下，当你打开程序包时，SSIS 会检查所有数据源，确保外部元数据是有效的。如果你的包中包括了许多网络资源连接，或者数据源访问比较慢（或者临时不可用），那打开包的这个过程就会消耗非常多的时间；因此，如果你需要经常重新打开包的话，这会把你搞得很沮丧。

离线连接管理器

你可以通过在个体组件上设置“DelayValidation”属性来解决这个问题；但是如果你发现你必须经常重设这个属性，而且有许多这类组件需要操作，你就需要花大量时间处理这个问题，而不能专心开发你需要的程序包了。



配置连接管理器为离线工作

然而，SSIS 2012 使你有了更简便的办法。现在只要通过几次点击，你就可以指定连接管理器的离线或者在线状态。要设置连接管理器为离线工作状态，请在“连接管理器”窗体中该组件上右击，然后点击“离线工作”菜单，如图 1 所示。

当你设置连接管理器为离线工作时，SSIS 就会在“连接管理器”窗体中的连接管理器名称前面增加一个图标（红色的向下箭头）。图 2 展示了“AW2012 ”连接管理器设置为离线工作之后的样子。

连接管理器设置为离线工作

一旦你设置连接管理器为离线工作，使用这个连接的组件将不再验证元数据，直到连接管理器回到在线状态才会验证。事实上，任何引用该连接管理器的任务或者数据流组件，都会标记上一个红色叉号，表示该组件不能获取连接。

如果想让连接管理器回到联机状态，只需要右键单击该连接管理器，然后再次点击“离线工作”菜单取消该选项选中状态即可。你还可以不用做任何操作，当你关闭并重新打开程序包时，所有连接管理器都会被验证并设置为联机工作状态。

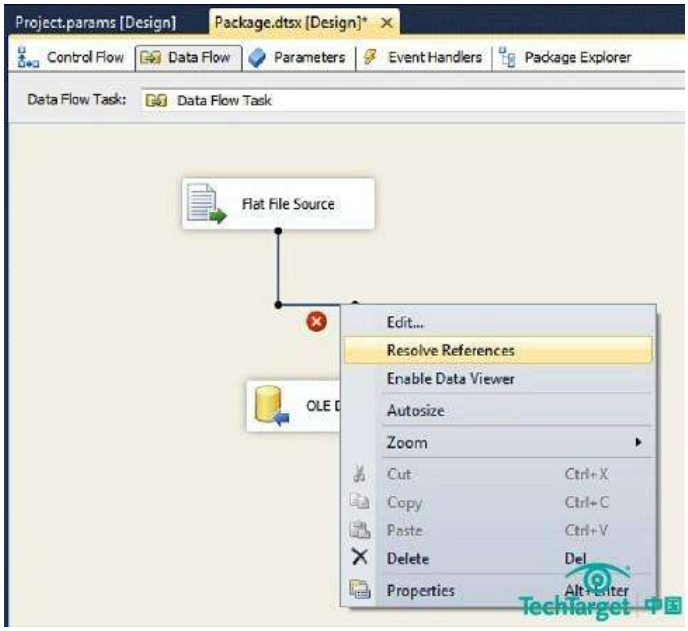
既然重新打开包会设置所有连接管理器为联机工作的话，看起来使用离线工作选项似乎没有什么优势了。“DelayValidaton ” 属性的设置在你关闭并重新打开包之后仍然存在。与该属性不同的是，SSIS 支持在打开包之前会将所有连接管理器设置为离线工作。在项目打开包没打开的情况下，在 SSIS 菜单中选择“离线工作”选项，然后再打开你的包，所有连接管理器将被设置为离线工作状态，这样就避免了初始化的验证过程。

列映射解析

想象一下这样的场景：你创建了一个 SSIS 包，配置复杂的数据流，然后发现在你的数据源那边有一个字段名变了。在过去，SSIS 对这类干扰处理得不是很好，尤其是涉及数据流的数据路径变化时。这种变更情况下重新分配映射列经常会导致个别组件的重新配置或者重新实现。SSIS 2012 解决了这个问题，它支持轻而易举地更新列引用。

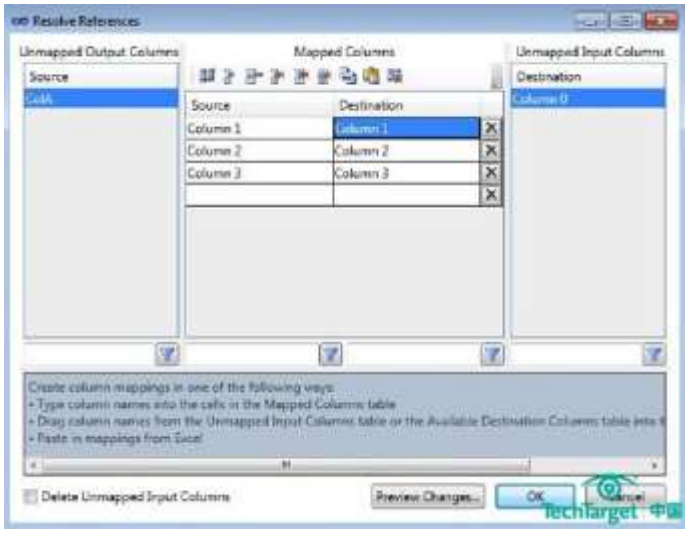
在数据路径列映射中的一个错误

我们来看一个例子，理解一下它是如何工作的。图 3 展示了一个数据流，其中包含文本文件数据源和“OLE DB”目标。在该程序包创建之后，数据源中的一个列名改变了，这就导致上图看到的红色叉号“X”指向数据路径。



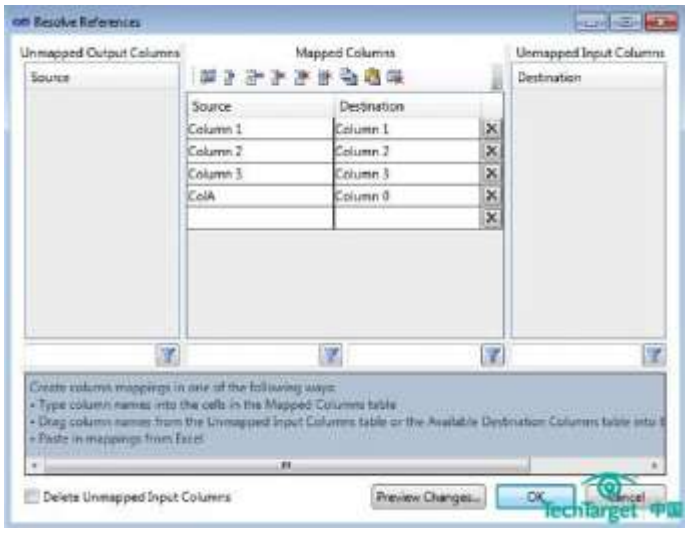
解决数据路径中的列映射问题

如果你在数据路径上单击右键，就会发现一个新选项菜单“重新映射引用”的出现，如图 4 所示。点击该菜单，你就可以修改数据流中的输入和输出引用，让它指定为特定的数据路径。



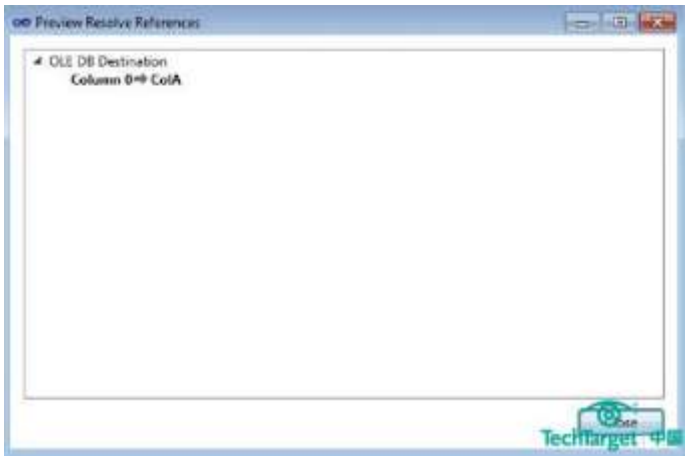
数据路径中未映射的列

点击该菜单项运行“调整引用”对话框，如图 5 所示。你可以看到三个列映射到“Column 1, Column 2 和 Column 3 ”，这些信息都显示在对话框中间的“映射列”表格中。请注意，源列表中包含有“ColA”列。这就是名称变化的列，原来是“ Column 0”，改成了“ColA”。还要注意的是“目标”列表中包含有“ Column 0”列。原来的源列和目标列共享使用这个名称，现在变成彼此映射了。



映射与数据路径相关联的列

要映射“ColA ”到“Column 0”列，你可以把每个列拖拽到“映射列”表格合适的行中，这样他们就手动映射上了，如图 6 所示。



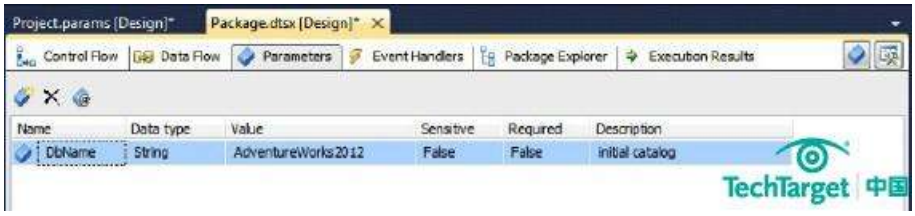
预览调整后的列引用

一旦你给“已映射列”表格中增加了列，就可以通过点击“预览变更”按钮来验证你的变更。这个操作会运行“预览调整引用”对话框，如图 7 所示。如你所见，指向“ OLE DB”目标的“Column 0” ，现在映射到“ColA ”列了。

那些经历过数据路径引用映射调整的人们将会对此有特殊的理解，他们会认为这个新功能可以节约很多时间。只需要几次简单的点击，你就可以重新映射列，马上就能回到正常业务。

更多 SSIS 技巧：使用参数与撤销操作

SSIS 2012 中另一项方便好用的新特性就是创建参数的功能，它可以让你在执行时很容易地传递属性值给程序包。例如，如果你想在运行程序包时传递 SQL Server 实例名称，只需要创建一个参数，并在执行包时提供实例名称就可以了。

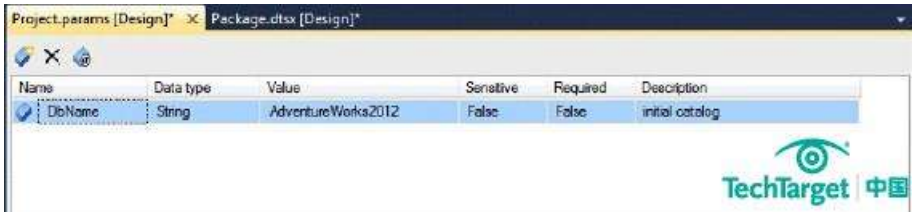


创建程序包参数

程序包参数和项目参数

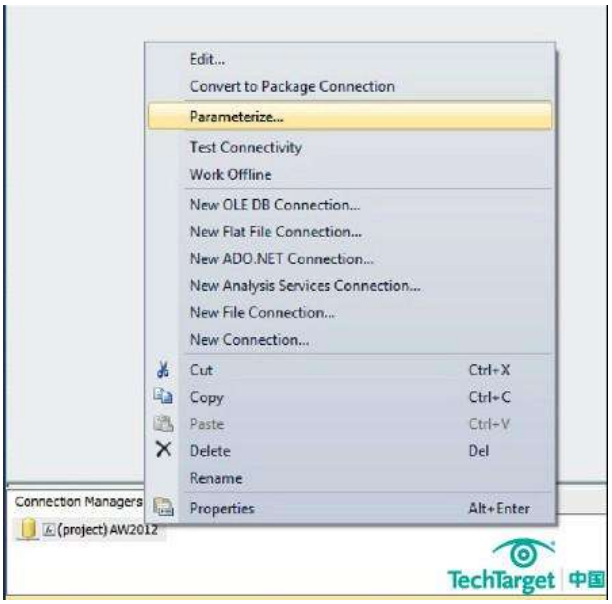
你还可以选择在程序包级别或者项目级别创建参数。要创建程序包参数，请在 SSIS 设计器中找到“参数”标签页，点击“添加”按钮，然后在空行中输入创建参数需要的信息。例如，图 1 展示了名为“DbName”的参数，它用来传递数据库名称给程序包，在其执行时传递即可。

请注意，在提供参数名称的同时，你必须指定它的数据类型、初始值和说明信息。另外两项设置属性是“是否敏感信息”和“是否必须信息”，这两项的设置要看你的个人需求了。“是否敏感”属性决定了该值是否要被加密，而“是否必须”属性决定了该值在包执行时是否必须提供。这两项设置的默认状态都是未选。



创建项目参数

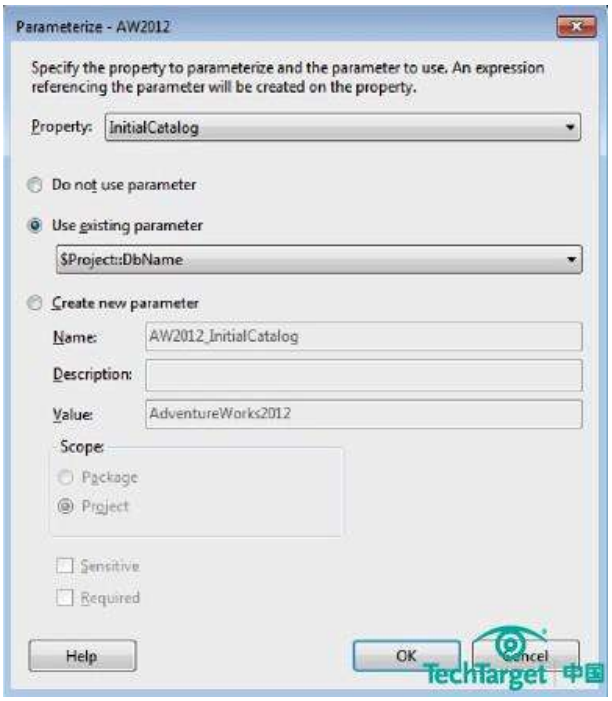
创建项目参数也很容易。请在“对象资源管理器”中双击“Project.params ”节点，此时“Project.params ”标签页运行，如图 2 所示。请注意，该标签页列出的参数与上面的程序包参数有同名的。因为项目参数和包参数彼此是分离的，而且在不同的作用域生效，所以你可以使用相同的名称。在你后续引用参数时，你可以通过项目名称和包名称前缀对它们加以区分。



参数化 SSIS 组件

这就是创建程序包参数和项目参数的全部内容。毫无悬念的是，你只能在它们创建的程序包中使用包参数，但是你可以在项目中的任何包中使用项目参数。

在创建参数之后，接下来的一步就是把它与组件中的具体属性相关联。你可以在要关联的组件上单击右键，然后点击“参数化”，如图 3 所示。在本例中，选择的组件是“AS2012”项目连接管理器。

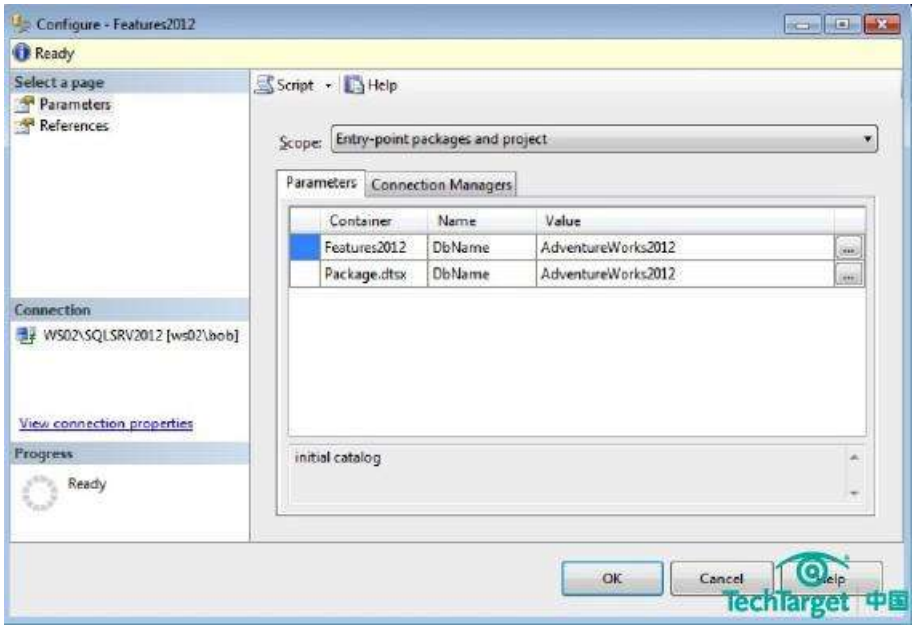


给组件属性设置参数

当你点击“参数化”菜单时，会弹出“参数化”对话框，如图 4 所示。在这里，你可以选择组件的某个属性，并选择你想与该属性关联的参数。在本例中，选择的是“InitialCatalog ”属性，项目参数“DbName ”被分配给了该属性。

属性与变量关联之后，你就可以正常编译和部署你的程序包了，与其它程序包没什么不同。然后，在你运行包时要提供参数值。SSIS 提供了几种方式供你选择。例如，如果你从 SQL Server Management Studio 中运行包，就可以从

“对象资源浏览器”提供参数值。采用这种方式，请先连接到你部署包的 SSIS 实例，然后打开“集成服务目录”节点。看到你的项目以后，单击右键，然后点击配置。此时会出现“配置”对话框，如图 5 所示。



在 SQL Server Management Studio 中设置参数值

请注意，在本例中，对话框为项目“Features2012”和程序包“Package.dtsx”都显示“DbName”参数。但是，如果你选择从程序包本身运行这个“配置”对话框，那么就只会显示程序包级别的参数了。

如图 5 所示，每个参数在创建时都分配了初始值。你可以通过点击参数右边的“浏览”按钮（省略号按钮）改写提供新值。然后，当你运行该包时，它就会使用该值。这里的亮点就是你可以在每次运行包时提供不同的值（也可以是相同的值）。

撤销与重做操作

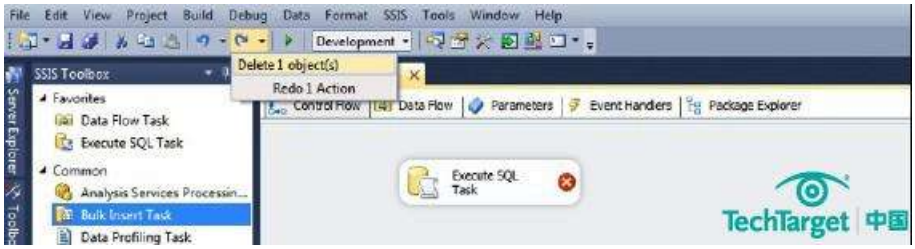
在 SSIS 2012 所有新增特性中，最值得开发人员大声欢呼的就要数操作的“撤销”和“重做”功能了。过去，如果你不小心删除了配置复杂的组件，后来发现还需要使用该组件，就得必须从头再配置一遍。而现在，“撤销”和“重做”可实现多达 20 次的动作处理。该操作在“控制流”、“数据流”、“参数”和“事件处理器”标签页都可使用，在“变量”窗口也可以。另外，不仅工具栏上添加了“撤销”和“重做”按钮，“编辑”菜单下也有了“撤销”和“重做”菜单。

撤销一个或多个操作

要撤销单个操作，只需要点击“撤销”按钮或者在“编辑”菜单下选择“撤销”菜单就可以了。如果你想撤销多个步骤，点击“撤销”按钮旁边的下拉箭头，把你想撤销的动作高亮选中，然后点击“回车”即可，如图 6 所示。

“重做”按钮的用法与“撤销”按钮类似。要重做单个动作，只需要点击“重做”按钮或者选择“编辑”菜单下的“重做”菜单即可。如果你想重做多个动作，请点击“重做”按钮旁边的下拉箭头，高亮选中你想重做的动作，然后点

击“回车”完成，如图 7 所示。



重做一个或多个操作

在 SSIS 中添加撤销和重做功能看起来似乎不是什么大事，但是与 SSIS 的其它更新功能相比，它们给你节约的时间更多。只要试一试，我保证你就再也离不开它们了。

SQL Server 集成服务 2012 版

学习了本系列关于 SSIS 新特性的文章，你以后的工作定会变得更加轻松。这些新特性已经无缝地集成到 SSIS 中，很快会变成一种操作习惯。此外，SSIS 2012 还增强了许多其它功能，帮助改善服务器环境、流程部署、内存管理、数据质量和专门的 SSIS 组件。尽管仍有很多内容没有提到，但本系列文章对你通往 SQL Server 2012 的殿堂之路是个不错的开始。

SQL Server 2012：大数据大问题

对 Frank Rietta 而言，微软拥抱开源项目是 SQL Server 2012 最好的一部分。

Frank Rietta 正在谈论与 SQL Server 2012 的 Apache Hadoop 集成。他是一个软件开发人员、亚特兰大市 Rietta 公司的总裁、先进技术发展中心(ATDC)的公司会员，ATDC 负责运营乔治亚理工学院的孵化计划。

“即便对于小型创业公司来说，大数据也变得越来越重要，”他说：“微软与开源项目的合作令人耳目一新。”

另一种重复劳动是不必要的，呈现与本身不兼容的竞争。

Rietta 补充说，SQL Server 2012 中添加的一些新功能看起来非常有用，包括基于列的查询，基于 Excel 的分析和报表功能的增强，和高可用性方面的改进。但他认为，它们还没有足够令人信服促使他从现有的开源产品迁移至微软平台。所以他不打算升级。

大数据是 SQL Server 2012 的一个大特性

与某些用户和顾问的讨论，揭示出 SQL Server 超越竞争对手的普遍原因，即易用性和低成本。SQL Server 2012 在三月份进行了虚拟发布，4 月 1 日起正式供货。大数据毫无疑问是许多 IT 部门所牵挂和好奇的。

Count Sanjay Bhatia 是其中之一。Bhatia 是 Izenda 公司的首席执行官和创

始人，该公司从事定制报表业务已有十年历史。他们使用 SQL Server 存储其所有数据，其 90%的客户使用 SQL Server 数据库。Bhatia 喜欢 Hadoop 功能，因为它与他所称之的“SQL 核心能力”相集成。

“或许你有大量的非结构化数据，具备类似 Hadoop 的查询功能可以使你能够访问海量用户数据，”他说。

Bhatia 补充说，当你为即将发布的 Windows Server 8 考虑 SQL 2012 时，业务部门可以“按一个钮”，而事实上是运行一个应用程序。例如，如果你需要某个应用程序，在一个私有云中它能被自动安装。然后，你的业务部门付费使用;或者如果没有达到指定服务级别协议，它就退回费用。这样就可以取代你进行软件安装的大量工作，所以可以先尝试一下，他解释说。

另一位爱好者是 Jordan Hudgens，德克萨斯州米德兰市 MCW 服务公司的高级软件工程师。该公司专门为油田行业开发定制软件。他的公司拥有一个异构数据库环境，包括基于微软 SQL Server 2008 的.NET 应用程序和基于 MySQL 的 PHP 应用。他们使用 Rackspace 和 LiquidWeb 的托管服务器，以及其他基于云的解决方案，现在正处于迁移到 SQL Server 2012 的过程中。

Hudgens 说 MCW 服务公司使用 SQL Server 2012 最大的功能是其即将上线的即席分页查询。

“通过更有效地筛选查询结果，从而可以减少滞后时间，给用户提供更有效

的数据，”他说：“特别是，新功能内置于 OFFSET、SELECT 和 FETCH 命令中，这将有助于为我们的富数据应用程序返回更准确的结果。”

Hudgens 觉得这些 SQL Server 2008 R2 命令“有点麻烦”，它们将会在 SQL Server 2012 中更加精简和高效。

解读微软大数据组件 SQL Server for Hadoop 连接器

最近“大数据”在我们网站的话题中占据了很重的份量。就在上个月，微软公司发布了两个基于开源分布式计算框架 Hadoop 的用于大数据处理的社区技术预览版连接器组件，一个用于 SQL Server，另一个用于 SQL Server 并行数据仓库(PDW)。

在本月的 SQL Server 访谈栏目中，微软数据库平台专家 Mark Kromer 向我们介绍了微软公司的大数据策略。Kromer 还提及了微软对访问关系型数据的开放数据库互连标准 OBDC 的支持，其变化会给开发者带来什么好处和挑战，以及是如何推动该公司进入云世界的。

微软希望向客户提供的 SQL Server for Hadoop 连接器组件具备什么样的大数据处理能力？

Mark Kromer：其中一个使用场景：我非常熟悉这些适配器，它们适用于使用 Hadoop 进行大数据处理的需求，或者数据存储在一个向外扩展的文件系统中。企业可以使用 SQL Server PDW 和 SQL Server BI 提供大数据的分析处理能力，以充分利用 SQL Server 的投资。这些连接器都是双向的，允许你在 SQL Server 和 HDFS(Hadoop 分布式文件系统)之间相互传输数据，便于你迁移大量的 SQL Server 数据，也就是说，从一个大型的分布式 PDW 数据仓库把数据迁移到

Hadoop，同样可以使用 SQL Server BI 功能在 SQL Server 中分析 Hadoop 的数据。

作为拥有海量数据的微软客户，需要面对什么样的数据处理挑战？

Kromer：企业有大数据的需求，比如搜索引擎或者大型社交网站都需要非常非常快地处理超大数据集。在这种情况下，它使用像 Hadoop 和 MapReduce 之类的分布式 NoSQL 工具可能是有好处的，在这些工具中数据库模式被最小化成经典的 SQL 构造，如 ACID(原子性、一致性、隔离性、持久性)和参照完整性，保留加快和方便数据访问的特性。微软支持他的客户们使用这些连接器来解决大数据需求。围绕分布式处理和大数据，微软研究院和 Windows Azure 设立了一些非常令人兴奋的项目。由微软出版的微软观察家 Andrew Brust 撰写的白皮书中，谈到如何在 Windows Azure 中使用现有的功能，例如 Azure Table Storage、以键值对形式使用 Lite 模式存储结构化数据进行便捷快速地访问。

微软把 Hadoop 连接器的发布称作大数据之旅的“第一步”。那么下一步呢？

Kromer：尽管有了这些“试用”的连接器组件，发表路线图的评论还为时过早。一旦我们从 SQL Server 社区看到更多的反馈和对 Hadoop 和 SQL Server 的测试结果，那么我们就可以对企业有什么样的需求有一个更清晰的了解。这种反馈将有助于确定下一步应该做成什么样子。虽然 Hadoop 和 MapReduce 是目前非

常流行的满足企业对大数据的要求，微软继续加大对大数据和分布式编程的投资。SQL Server PDW 是微软第一个完全意义上的分布式数据库，用于内部部署的数据仓库。SQL Azure 很快就会推出 SQL Federations，这一特性允许分发联机事务处理 OLTP 数据库的工作负载，也可以使用此功能来分发非结构化大数据和相关的数据库模式。沿着相同的路线，在分布式计算方面，Windows HPC(高性能计算)小组刚刚发布用于处理大数据集的 LINQ to HPC，在 HPC 群集节点间分发 LINQ 操作。

SQL Server Hadoop：开拓大数据新疆域

在大数据的背景下，微软似乎并不像其他数据库厂商一样在高调宣传他们的大数据产品或解决方案。而在应对大数据挑战方面，倒是一些互联网巨头走在最前面，比如 Google 和 Yahoo，前者每天都要处理 20 PB 的数据量，其中一大部分是基于文档的索引文件。当然，如此界定大数据是不准确的，它并不仅限于索引，企业中的电子邮件、文档、Web 服务器日志、社交网络信息以及其他所有非结构化的数据库都是构成大数据的一部分。

为了应对这些数据的挑战，像 Autodesk、IBM 、Facebook 当然还包括 Google 和 Yahoo，都毫无例外地部署了 Apache Hadoop 开源平台。微软也注意到这一趋势，所以在他们的数据库平台中添加了 Hadoop 连接器。该连接器可以让企业将海量的数据在 Hadoop 集群和 SQL Server 2008 R2、并行数据仓库以及最新的 SQL Server 2012(Denali)之间进行自由的移动。由于连接器可以让数据双向移动，所以用户不仅可以利用 SQL Server 所提供的强大的存储以及数据处理功能，还可以用 Hadoop 来管理海量的非结构化数据集。

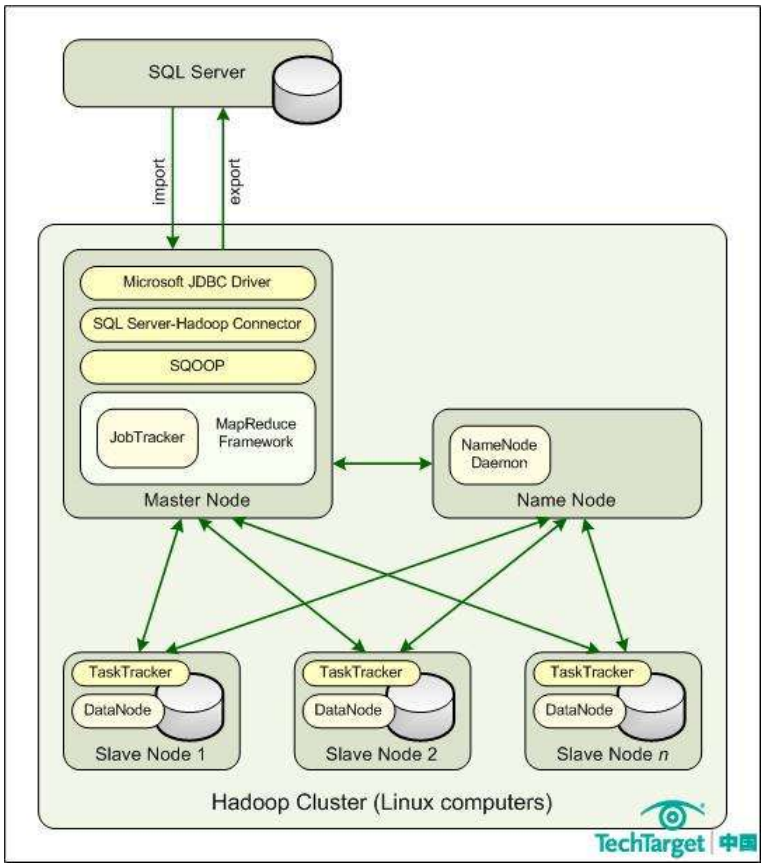
但是传统的微软用户对于 SQL Server Hadoop 连接器还比较陌生，使用起来会很不习惯。该连接器是一个部署在 Linux 环境中的命令行工具，在本文中，我们就将为您具体讲解一下 SQL Server Hadoop 连接器的工作原理。

Apache Hadoop 集群

Hadoop 是一个主-从架构，部署在 Linux 主机的集群中。想要处理海量数据，Hadoop 环境中必须包含一下组件：

- 主节点管理从节点，主要涉及处理、管理和访问数据文件。当外部应用对 Hadoop 环境发送作业请求时，主节点还要作为主接入点。
- 命名节点运行 NameNode 后台程序，管理 Hadoop 分布式文件系统(HDFS)的命名空间并控制数据文件的访问。该节点支持以下操作，如打开、关闭、重命名以及界定如何映射数据块。在小型环境中，命名节点可以同主节点部署在同一台服务器上。
- 每一个从节点都运行 DataNode 后台程序，管理数据文件的存储并处理文件的读写请求。从节点由标准硬件组成，该硬件相对便宜，随时可用。可以在上千台计算机上运行并行操作。

下图给出了 Hadoop 环境中各个组件的相互关系。注意主节点运行 JobTracker 程序，每个从节点运行 TaskTracker 程序。JobTracker 用来处理客户端应用的请求，并将其分配到不同的 TaskTracker 实例上。当它从 JobTracker 那里接收到指令之后，TaskTracker 将同 DataNode 程序一同运行分配到的任务，并处理每个操作阶段中的数据移动。



你必须将 SQL Server Hadoop 连接器部署在 Hadoop 集群之内

MapReduce 框架

再如上图所示，主节点支持 MapReduce 框架，这一技术是依赖于 Hadoop 环境之上的。事实上，你可以把 Hadoop 想象成一个 MapReduce 框架，而这个框架中会有 JobTracker 和 TaskTracker 来扮演关键的角色。

MapReduce 将大型的数据集打散成小型的、可管理的数据块，并分布到上千台主机当中。它还包含一系列的机制，可以用来运行大量的并行操作，搜索 PB 级

别的数据，管理复杂的客户端请求并对数据进行深度的分析。此外，MapReduce 还提供负载平衡以及容错功能，保证操作能够迅速并准确地完成。

MapReduce 和 HDFS 架构是紧密结合在一起的，后者将每个文件存储为数据块的序列。数据块是跨集群复制的，除了最后的数据块，文件中的其他数据块大小都相同。每一个从节点的 DataNode 程序会同 HDFS 一起创建、删除并复制数据块。然而，一个 HDFS 文件只可以被写一次。

SQL Server Hadoop 连接器

用户需要将 SQL Server Hadoop 连接器部署到 Hadoop 集群的主节点上。主节点还需要安装 Sqoop 和微软的 Java 数据库连接驱动。Sqoop 是一个开源命令行工具，用来从关系型数据库导入数据，并使用 Hadoop MapReduce 框架进行数据转换，然后将数据重新导回数据库当中。

当 SQL Server Hadoop 连接器部署完毕之后，你可以使用 Sqoop 来导入导出 SQL Server 数据。注意，Sqoop 和连接器是在一个 Hadoop 的集中视图下进行操作的，这意味着当你使用 Sqoop 导入数据的时候是从 SQL Server 数据库检索数据并添加到 Hadoop 环境中，而相反地，导出数据是指从 Hadoop 中检索数据并发送到 SQL Server 数据库当中。

Sqoop 导入导出的数据支持一些存储类型：

- 文本文件：基础的文本文件，用逗号等相隔;
- 序列文件：二进制文件，包含序列化记录数据;
- Hive 表：Hive 数据仓库中的表，这是针对 Hadoop 构建的一种特殊的数据仓库架构。

总体来说，SQL Server 和 Hadoop 环境(MapReduce 和 HDFS)能够让用户处理海量的非结构化数据，并将这部分数据整合到一个结构化的环境中，进行报表制作以及 BI 分析。

微软大数据策略才刚刚开始

SQL Server Hadoop 连接器在微软大数据之路上算是迈出了重要的一步。但与此同时，由于 Hadoop、Linux 和 Sqoop 都是开源技术，这意味着微软要对开源世界大规模地敞开胸怀。其实微软的计划并不只如此，在今年年底，他们还将推出一个类似于 Hadoop 的解决方案，并以服务的形式运行在 Windows Azure 云平台上。

在明年，微软还计划推出针对 Windows Server 平台的类似服务。不能否认，SQL Server Hadoop 连接器对于微软来说意义重大，用户可以在 SQL Server 环境中处理大数据挑战，相信在未来他们还会带给我们更多的惊喜。

详细解读微软 HadoopOnAzure 的大数据处理功能

在大数据技术中，Apache Hadoop 和 MapReduce 是最受用户关注的。但管理 Hadoop 分布式文件系统，或用 Java 编写执行 MapReduce 任务则不是简单的事。那么 Apache Hive 也许能帮助您解决这一难题。

Hive 数据仓库工具也是 Apache Foundation 的一个项目，同时是 Hadoop 生态系统的关键组件之一，它提供了基于语境的查询语句，即 Hive 查询语句。这套语句可以将 SQL 类查询自动翻译成 MapReduce 工作指令。

在 BI 领域，包括 IBM DB2、Oracle 和 SQL Server 在内的关系型数据库一直处于统治地位。这使得 SQL 成为商业智能的首选语言，大部分数据分析专家都掌握了较强较全面的 SQL 技能。同样道理，做数据分析的专家对 Excel，Pivot 表格和图表等工具更熟悉。

先让我们来看一个端到端的 BI 项目在 Windows Azure 系统中是如何运行的。在 Excel 图表中显示美国航空公司的航班正点到达的数据，数据量很大，整个过程不需要编写任何程序代码。

Windows Azure CTP 上的 Apache Hadoop

去年年底，微软 SQL Server 研发团队宣布了 Windows Azure 平台上的 Apache Hadoop 功能，即 HadoopOnAzure。微软方面称这将简化 Hadoop 的

使用和设置，利用 Hive 来从 Hadoop 集群中提取非结构化的数据，并在 Excel 工具中进行分析，同时增强了 Windows Azure 的弹性。

HadoopOnAzure 的社区预览版还处于未公开状态，用户需要在 Microsoft Connect 上填写一个简单的问卷调查来获得邀请码，并通过 Windows Live ID 登陆。输入唯一的 DNS 用户名，选择初始 Hadoop 集群大小，提供一个集群登录名和密码，点击 Request Cluster 按钮。（见图 1）

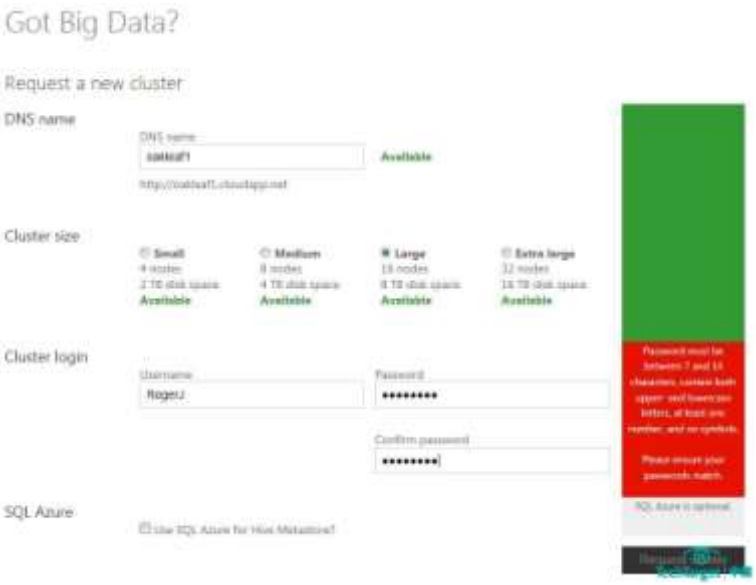


图 1 用户只需简单的操作即可修改一个集群设置

开通并设置集群需要 15 到 30 分钟时间。HadoopOnAzure 社区预览版的资源是免费的，但开通集群 24 小时之内，你需要在最后 6 小时的时候更新一下订阅。此后的使用过程中，证书需要每天都更新一次。

用户如要使用 Windows Azure 的 blob 持久性数据存储，那么就需要一个 Windows Azure 的订阅和一个存储账户。否则当集群超时，所有存储在 HDFS 上的数据都将丢失。如果没有订阅，用户还可以申请注册免费试用三个月的 Windows Azure 账号，这个账号赠送每位用户 20GB 存储空间、上百万次存储传输以及 20GB 的外网带宽。

向 SQL Azure blob 中填充大数据

这个 Apache Hive 项目从美国联邦航空管理局（FAA）提取数据，收集了 2011 年后 5 个月到 2012 年 1 月共 6 个月以来航班正点到达的信息及延误信息。6 页文本资料子集包涵 FAA 文件栏，栏下有 50 万行数据每页 25 MB。

用户需要将数据上传到一个 blob container 的文件夹中，这样 Hive 可搜索到这些数据。关于如何创建 Azure blob 源数据的详细步骤，可以参考我之前的一篇文章。文章还提到了数据文件以及如何用 Windows Live SkyDrive 账号下载数据，最后怎样将数据上传到 Windows Azure blob 的具体步骤。

集群配置完成后，将弹出 Elastic MapReduce 门户登录页面和集群、账户管理对话框。（见图 2）

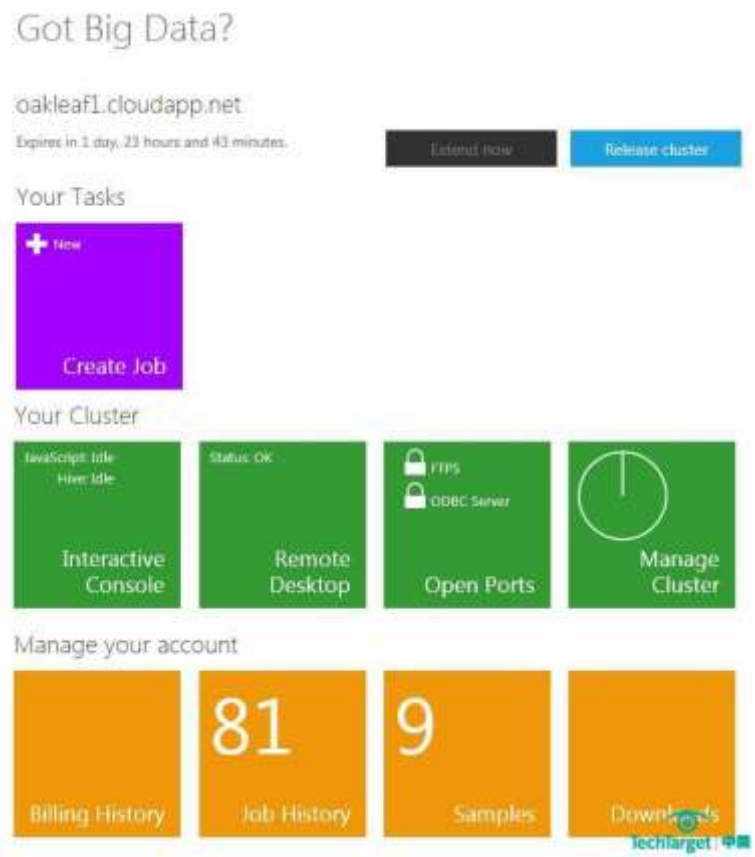


图 2：HadoopOnAzure 的 MapReduce 控制板页面特性和功能

复制 Windows Azure 管理门户的 Primary Access Key 保存到剪切板，点击 Manage Cluster，打开页面然后点击 Set Up ASV（Azure 存储库），将 Windows 存储账户作为 Hive 表的数据源。此外，用户还可以使用 Amazon S3 或 Windows Azure Dataplace DataMarket 中的数据作为 Hive 表的数据源。

输入你的存储账号，在 Passkey 框中粘贴 Primary Access Key，点击保 Save Settings，Hive 即可成功登录数据库访问 blob。如果证书获得认证，用户将收到短信通知 Azure 账号设置成功。

与 HDFS 不同，即便是最简单的 KV（Key-Value）数据，Hive 表都需要有 schema。要从非 HDFS 或外部制表符号数据中生成一个 Hive 表，给其列命名并定义数据类型的话，用户就需要运行 CREATE EXTERNAL TABLE 语句，如下面的 HiveQL 所示：

```
CREATE EXTERNAL TABLE flightdata_asv (
year INT,
month INT,
day INT,
carrier STRING,
origin STRING,
dest STRING,
depdelay INT,
arrdelay INT
)
COMMENT 'FAA on-time data'
ROW FORMAT DELIMITED FIELDS TERMINATED by '9'
STORED AS TEXTFILE
LOCATION 'asv://aircarrier/flightdata';
```

Apache Hive 的数据类型相对较少，并且不支持日期或时间字段，好在源数据 *.csv 对应的整数字段如年，月和日数值正好有利于数据的维护。出发（depdelay）和到达（arrdelay）的延误时间值将以分钟的形式呈现。

执行动态 HiveQL 语句，可以点击 Elastic MapReduce 的 Interactive Console，然后点击 Hive 按钮打开动态 Hive 页面，页面顶部出现只读文本框，点击下方文本框为说明。（见图 3）

图 3：Hive 图表选项列表包括新图表标题，列单元格显示某个选定图表字段名。点击 › ›键在单元格中插入选定的条目

下载并安装 Apache Hive ODBC 驱动及 Excel 插件

返回 Elastic MapReduce 主页面，点击 Downloads 面板。找到与用户 Excel 版本对应的安装链接，然后点击 Run，打开警告对话框。点击 More Options，出现 Run Anyway 选项，点击开始安装，打开 ODBC 驱动启动 Hive 设置对话。在 I Accept 框中打钩。

点击 Install 开始安装驱动，完成后点击 Finish 退出安装。下一步，打开 Excel，点击 Data 标签，确认 Hive Pane 图标存在，点击图标，工作表右面出现 Hive Query 仪表盘。安装插件会放置一个 Hive Pane 图标到目录的 Hive Data 部分。

返回 EMR 控制主页面，点击 Open Ports 打开 Configure Ports 页面，点击 ODBC Server，往右拖动，打开 TCP port 10000。

执行交互式 Apache Hive 查询

返回 Excel，点击 Hive Pane 图标，打开 Hive Query 任务框，点击 Enter Cluster Details 来打开 ODBC Hive Setup 对话框，输入一个描述及 DNS 主机名称，接受 TCP 端口。下一步，选择 Username/Password authentication，输入你的 Elastic MapReduce 门户实例的用户名及密码。（见图 4）

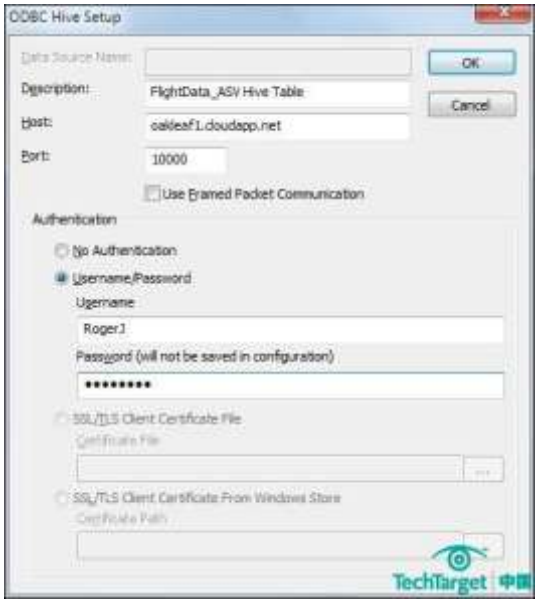


图 4：每个链接，机场，TCP 端口和集群用户名密码都有其对应的具体名称

若 ODBC Hive 对话框中设置的 Hive 选项正确，那么当用户打开 Select or Enter Hive Connection 时，输入的名称会作为描述字段弹出。打开 Select the Hive Object to Query 列表，选择 flightdata_asv 生成 Columns 列表。

如要执行一个聚合查询来显示延误的平均时间，可以勾选 carrier 和 arrdelay 栏，打开 arrdelay 字段的函数列表，然后双击列表中的 avg，将其添加到 HiveQL 语句当中（见图 5）。

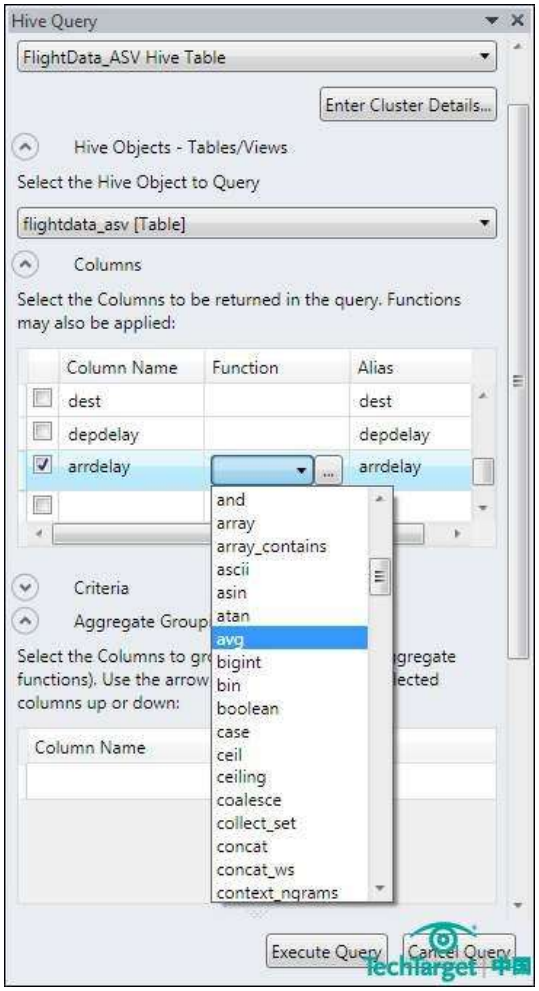


图 5：选择 avg HiveQL 进行聚合查询并双击，HiveQL 功能比大多数的 SQL 更丰富一些

下一步，划去 Limit Results 勾选框，打开 Aggregate Grouping 列表，选择 carrier 列。

在 avg()中输入 arrdelay，如 avg(arrdelay)，这可以消除查询语句设计流程的缺陷，点击 Execute Query 得出查询结果。（见图 6）

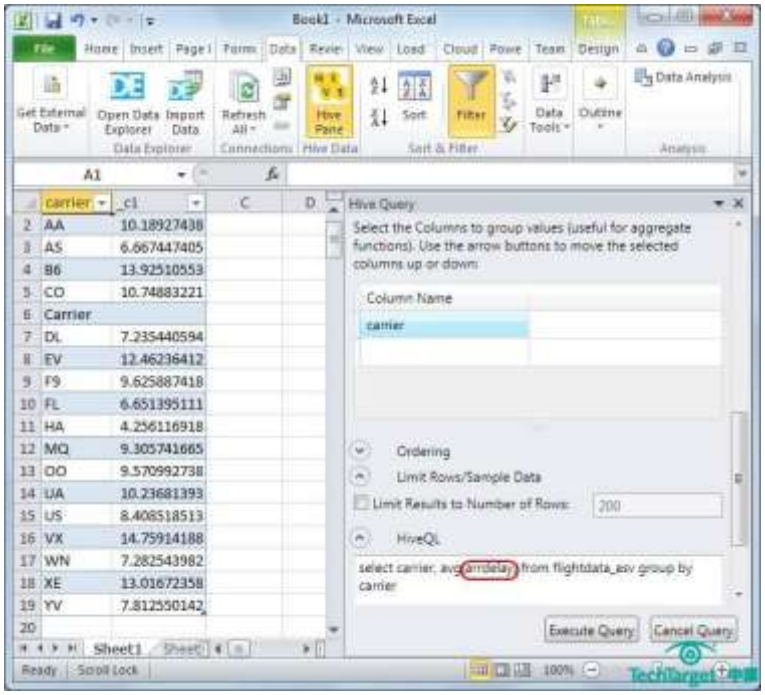


图 6：这是 HiveQL 查询执行后的结果，B6 好 F9 是 FAA 专用的两个字节代码，B6 指代 Jet Blue，F9 指代 Frontier Airlines

删除错误的 Carrier 条目，这可能是由于每列的首标发生错误，导致信息被留在了文档中，结果出现在查询结果里。保留一位小数，关掉任务框，将信息添加到工作表，添加标题，X 轴标题和数据标签。（见图 7）

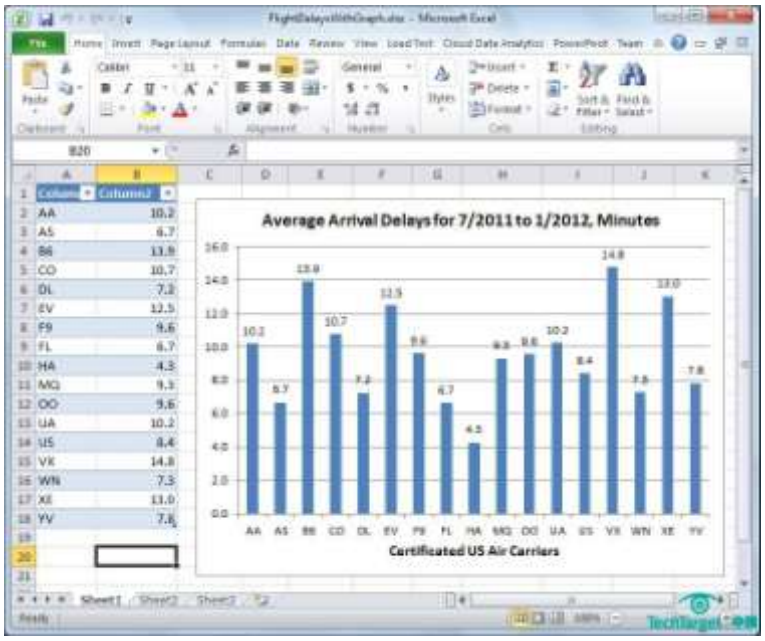


图 7：Excel 表单从图 6 的数据中生成得来

文章给出的例子阐述了运行 HadoopOnAzure CTP 的简单流程。微软代码为“Cloud Numerics”的项目也有相似的功能，但是它需要我们在 Visual Studio 10 以上版本的环境中进行操作。HadoopOnAzure 能够将表格数据直接传送到 Excel 当中，以便做进一步的分析。此外，交互式 Hive、Hive ODBC 数据源以及对应的 Excel 插件，都使得 HadoopOnAzure 成为大数据处理的理想平台。

我们的编辑团队

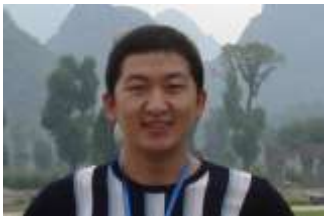
您若有何意见与建议，欢迎与我们的编辑联系。

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。