大数据集成与 Hadoop

可最大限度降低Hadoop计划风险并提高ROI的最佳实践

IBM

简介

Apache Hadoop技术通过支持新的流程和架构，不断改进大数据措施的经济性和活力，这样不仅有助于削减成本、增加收益，而且还能树立竞争优势。Hadoop是一个开源软件项目，支持在多个商业服务器群集间分散处理和存储大型数据集，并可根据需求变化从单一服务器扩展到数以千计的服务器。主要的Hadoop组件包括Hadoop Distributed File System（用于存储大型文件）和Hadoop分布式并行处理框架（称为MapReduce）。

但是，Hadoop基础架构本身并没有提供完整的大数据集成解决方案，摆在人们面前的既有挑战，也有机遇，只有处理好这些问题，才能安享各项优势，最大限度提高投资回报率（ROI）。

大数据集成对于Hadoop措施的重要性

Hadoop的迅速崛起推动企业在如何抽取、管理、转换、存储和分析大数据方面实现了范式转变。无论是要更深入的分析，还是希望获得更出色的洞察、新产品、新服务以及更高的服务水平，都可以通过这项技术一一实现，从而大幅降低成本并创造新的收入。

依靠收集、移动、转换、清除、集成、治理、探索以及分析多种不同来源的大量不同类型的数据来实现大数据与Hadoop项目。实现所有这些目标需要运用富有弹性的端到端信息集成解决方案，该解决方案不仅可实现大规模扩展，还能提供支持Hadoop项目所需的基础架构、功能、流程和行为准则。

“在很大程度上，80%的大数据项目开发精力用于数据集成，只有20%的精力投入到数据分析中。”

—Intel Corporation，“使用 Apache Hadoop 抽取、转换和加载大数据”[1]

有效的大数据集成解决方案可实现简便性、高速度、可扩展性、功能和治理，从Hadoop沼泽中生成可使用的数据。没有有效的集成，势必形成“垃圾进垃圾出”的情况—这不是出色的受信任数据使用方法，更谈不上准确完整的洞察或转型成果。

随着Hadoop市场的不断发展，顶级技术分析师一致认为，Hadoop 基础架构本身并非完整或有效的大数据集成解决方案（请阅读此报告，其中对Hadoop为何并非数据集成平台进行了讨论）。更加糟糕的是，一些Hadoop软件供应商利用炒作、神话、误导或矛盾信息来渗透市场。

为彻底切断这种误导，并开发适合您的Hadoop大数据项目的采用计划，必须遵循最佳实践方法，充分考虑各种新兴技术、可扩展性需求以及当前的资源和技能水平。面临的挑战：创建最佳的大数据集成方法和架构，同时避免各种实施缺陷。

海量数据可扩展性：总体要求

如果您的大数据集成解决方案无法支持海量数据可扩展性，那么很可能无法达到预期的效果。为发挥大数据措施的整体业务价值，对于大部分Hadoop项目的大数据集成而言，海量数据可扩展性是必不可少的。海量数据可扩展性意味着对处理的数据量、处理吞吐量以及使用的处理器和处理节点数量全无限制。只需添加更多的硬件，即可处理更多的数据，实现更高的处理吞吐量。添加硬件资源的同时，无需修改即可运行相同的应用程序并且性能也会随之提高（参见图1）。

关键成功因素：避免炒作，分辨是非

在这些新兴的Hadoop市场阶段，请仔细分辨听到的所有说明Hadoop卓尔不群的言论。充分使用Hadoop的神话与现实之间存在巨大的反差，这在大数据集成方面表现尤为突出。很多业界传言称，任何不可扩展的抽取、转换和加载（ETL）工具搭配Hadoop后都会得到高性能、高度可扩展的数据集成平台。

事实上，MapReduce的设计宗旨并非是对海量数据进行高性能处理，而是为了实现细粒度的容错。这种差异可能会使整体性能和有效性降低一个数量级乃至更多。

Hadoop Yet Another Resource Negotiator(YARN)纳入了MapReduce的资源管理功能，并将它们内置其中，这样需要在Hadoop群集间动态执行的其他应用即可使用它们。结果是，这种方法可将大规模可扩展数据集成引擎作为本机 Hadoop应用程序来实现，而且不会影响MapReduce的性能。希望在Hadoop上实现可扩展性和有效性的所有企业技术都需要采用YARN，并将其作为产品路线图的一部分。

开始集成之旅以前，请务必了解MapReduce的性能限制，以及数据集成供应商在解决这类问题方面的差异。请在“Themis: An I/O-Efficient MapReduce”一文中了解更多信息，文中对该主题进行了详细讨论：http://bit.ly/1v2UXAT

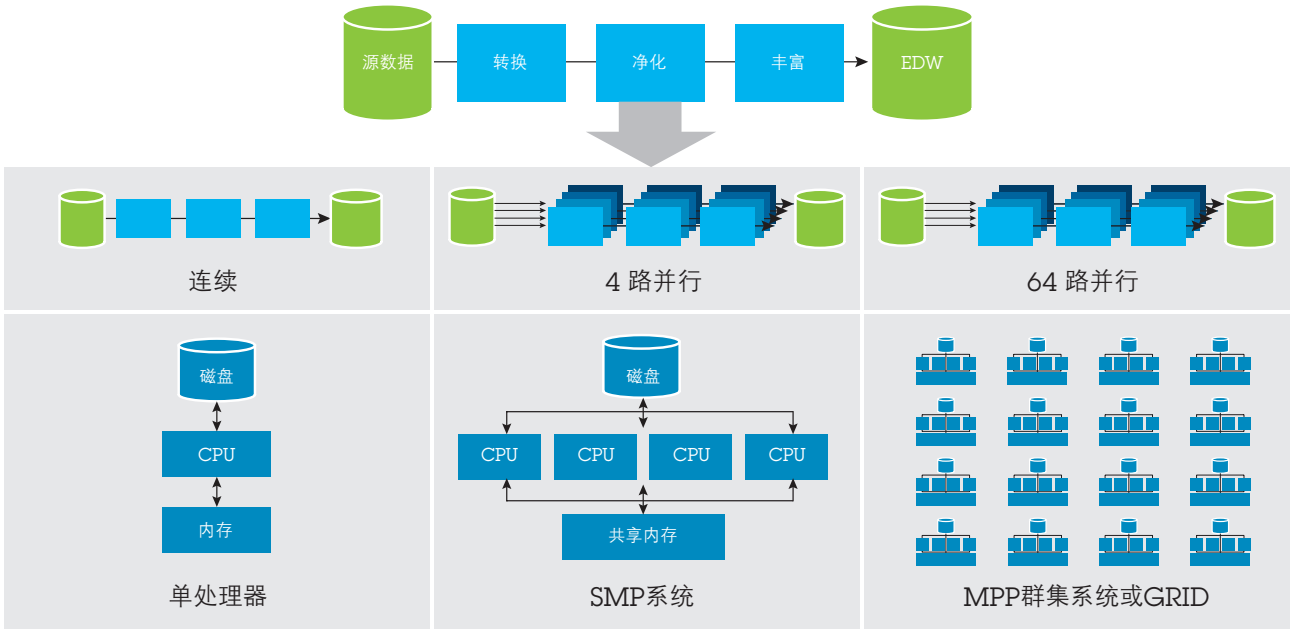


图1. 海量数据可扩展性是一项大数据集成的强制要求。在大数据时代，企业必须支持MPP群集系统才能实现扩展。

关键成功因素：大数据集成平台必须支持全部三个维度的可扩展性

- **线性数据可扩展性：硬件和软件系统通过线性增加硬件资源来线性提高处理吞吐量。例如，如果在50个处理器上运行4小时可以处理200GB数据，在100个处理器上运行4小时可以处理400GB数据，以此类推，则说明应用程序可以实现线性数据可扩展性。**
- **应用程序纵向扩展：衡量软件在一个对称多处理器(SMP)系统中的多个处理器间实现线性数据可扩展性的有效程度。**
- **应用程序横向扩展：确定软件在非共享架构的多个 SMP 节点间实现线性数据可扩展性的有效程度。**

支持海量数据可扩展性的需求并非只与Hadoop基础架构的出现有关。多年来，领先的数据仓库供应商（如IBM和Teradata）和领先的数据集成平台（如IBM® InfoSphere® Information Server）纷纷提供可支持海量数据可扩展性的非共享大规模并行软件平台，有些企业采用此做法已有近20年。

久而久之，这些供应商陆续集中关注4个常见的软件架构特征，以便为实现海量数据可扩展性提供支持，如图2所示。

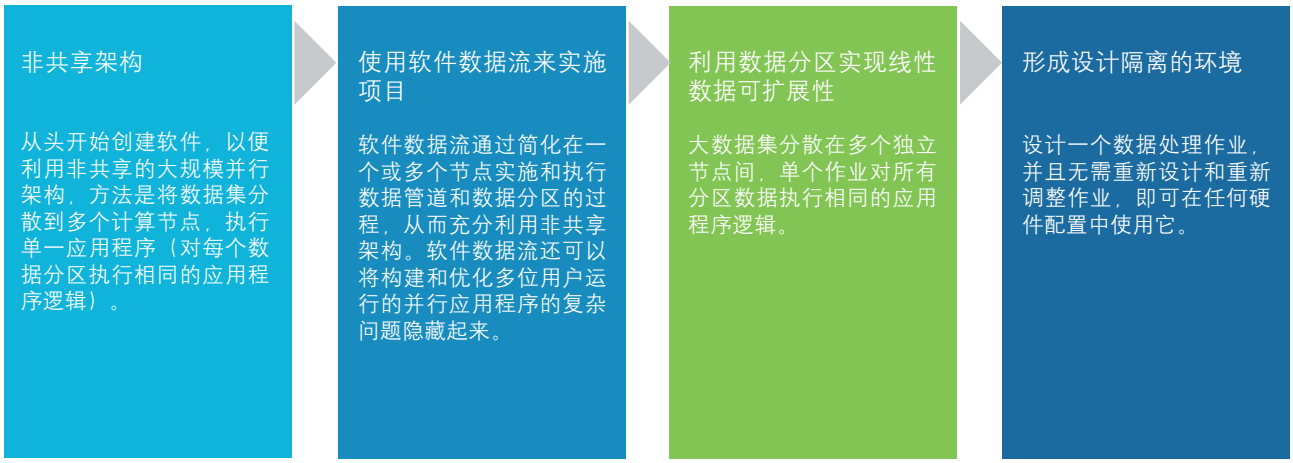


图2. 海量数据可扩展性的4大特征。

大部分商业数据集成软件平台在设计时从未考虑过支持海量数据可扩展性，这意味着在设计之初，并未考虑利用非共享大规模并行架构。它们依靠共享的内存多线程，而非软件数据流。

此外，有些供应商不支持将大数据集分散在多个节点间，无法对独立数据分区并行运行单一数据集成作业，也无法实现设计一次作业，无需重新设计和重新调整作业即可在任何硬件配置中使用它。这些功能对于通过提升效率来降低成本至关重要。没有它们，该平台将无法处理大量的大数据。

InfoSphere Information Server数据集成产品组合支持4大海量数据可扩展性架构特征。请在Forrester报告“Measuring The Total Economic Impact Of IBM InfoSphere Information Server”中了解更多信息： http://ibm.co/UX1RqB

优化大数据集成工作负载：一种平衡的方法

由于几乎所有Hadoop大数据用例和场景都需要首先进行大数据集成，所以企业必须确定如何优化整个企业的此类工作负载。一个Hadoop与大数据集成的重要用例是将大型ETL工作负载从企业数据仓库（EDW）卸载下来，以便降低成本并改善查询服务水平协议（SLA）。该用例会引发以下问题：

- 企业是否应卸载EDW中的所有ETL工作负载？
- 是否应将所有大数据集成工作负载都推送到Hadoop？
- 在没有并行关系数据库管理系统（RDBMS）和Hadoop的情况下，大数据集成工作负载在ETL网格中发挥怎样的持续作用？

这些问题的正确答案取决于企业独特的大数据需求。企业可以选择并行RDBMS、Hadoop和可扩展的ETL网格来运行大数据集成工作负载。但无论选择哪种方法，信息基础架构都必须满足一个常见的要求：全面支持大规模可扩展处理。

某些数据集成操作在RDBMS引擎内外的运行效率较高。同样，并非所有数据集成操作均适用于Hadoop环境。设计精妙的架构必须足够灵活，可以充分利用系统中每个环境的优势（参见图3）。

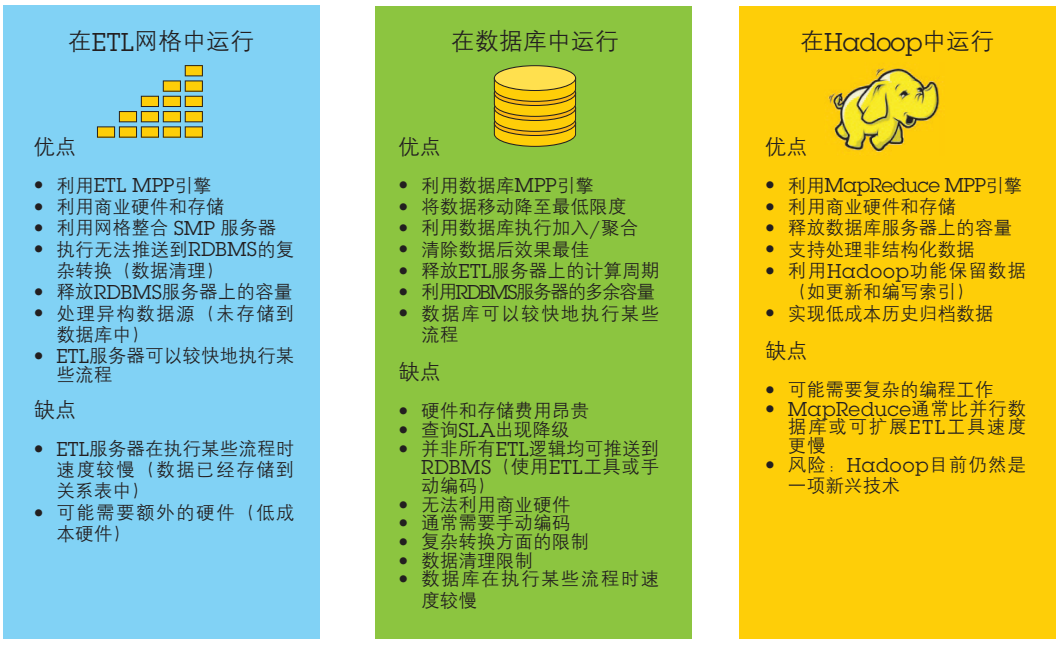


图3. 大数据集成需要一种可利用任何环境优势的平衡方法。

以下是优化大数据集成工作负载时需要遵循的三大重要指导原则：

1. **将大数据集成处理推向数据，而不是将数据推向处理：**指定可在RDBMS、Hadoop和ETL网格中执行的适当流程。

2. **避免手动编码：**手动编码费用昂贵，而且无法有效适应快速频繁的调整。另外，手动编码不支持自动收集对数据治理至关重要的设计和操作元数据。

3. **不要为RDBMS、Hadoop和ETL网格创建单独的集成开发环境：**这种做法没有任何实际意义，而且支持费用非常昂贵。您应该能够构建一次作业，然后即可在三个环境中的任意一个环境内运行它。

最适合Hadoop的流程

Hadoop 平台由以下两个主要组件构成：分布式容错文件系统（称为Hadoop Distributed File System (HDFS)）和并行处理框架（称为MapReduce）。

HDFS平台十分适合处理大型顺序操作，其中的数据读取“切片”通常为64MB或128MB。通常情况下，除非应用程序加载数据来管理相关任务，否则不会对HDFS文件进行分区或排序。即使应用程序可以对生成的数据切片进行分区和排序，也无法保证数据切片在HDFS系统中的位置正确。这意味着，无法在该环境中有效管理数据搭配工作。数据搭配（Data collocation）至关重要，因为它可确保将联接（join）键相同的数据整合到相同的节点，因此该流程不仅性能高，而且很准确。

虽然有很多方法可以应对数据并置支持缺乏的问题，但费用往往十分昂贵－通常需要额外的应用程序处理和/或重建工作。另外，HDFS文件不可更改（只读），处理HDFS文件类似于运行全表扫描，往往需要处理全部数据。对于像联接两个超大表这样的操作应该发出危险信号，因为没有将数据并置到同一Hadoop节点。

MapReduce V1是一个并行处理框架，并非用于高性能处理大型ETL工作负载。默认情况下，可在映射之间重新划分或重新并置数据，并减少处理阶段的时间。为加快恢复操作，可以先将数据保存到运行映射操作的节点，再进行随机选择和发送以减少操作。

MapReduce包含多种设施，可将较小的引用数据结构迁移至各映射节点，以便执行某些验证和增强操作。因此，会将整个引用文件迁移至各映射节点，这使其更适合较小的引用数据结构。如果进行手动编码，必须考虑这些处理流，因此最好采用一些工具来生成代码，从而将数据集成逻辑下推到MapReduce（也称为ETL pushdown）。

在Hadoop中使用ETL pushdown处理方法（无论采用哪种工具进行推送）可能会导致一种情形：必须在ETL引擎（而非MapReduce）中继续运行数据集成处理的重要部分。采用这种做法有以下几个原因：

- 较为复杂的逻辑无法推送到MapReduce
- MapReduce具有很大的性能局限性
- 通常数据按随机顺序方式存储到HDFS中

所有这些因素表明，在Hadoop环境中执行大数据集成需要以下三个组件来实现高性能的工作负载处理：

1)Hadoop发行版

2)非共享大规模可扩展ETL平台（如IBM InfoSphere Information Server提供的平台）

3)MapReduce ETL pushdown功能

需要同时具备全部三大组件，因为如果不进行手动编码，大部分数据集成逻辑将无法推送到MapReduce，因为MapReduce存在很多已知的性能限制。

关键成功因素：考虑数据集成工作负载处理速度

InfoSphere Information Server非共享大规模并行架构已针对高性能、高效处理大型数据集成工作负载进行了优化。IBM InfoSphere DataStage®-InfoSphere Information Server的一部分，运用高性能并行框架集成多个系统的数据，该框架处理典型数据集成工作负载的速度比MapReduce高10到15倍。[2]

InfoSphere DataStage还对Hadoop环境进行了均衡优化。均衡优化可生成Jaql代码，以便在MapReduce环境中本机运行它。Jaql自带优化器，该优化器会分析所生成的代码，并将其优化到map组件和reduce组件中。这样可自动执行传统的复杂开发任务，并让开发人员不必再为MapReduce架构而担忧。

InfoSphere DataStage可直接在Hadoop节点上运行，而不必像一些供应商实施计划要求的那样在单独的配置节点上运行。在与IBM General Parallel File System (GPFS™)-FPO搭配使用时，该功能有助于降低网络流量，这样即可在Hadoop环境中提供符合POSIX要求的存储子系统。POSIX文件系统允许ETL作业直接访问Hadoop中存储的数据，而无需使用HDFS接口。该环境支持将ETL工作负载迁移到运行Hadoop的硬件环境，从而帮助将处理工作移到数据存储位置，并充分利用Hadoop和ETL处理硬件。

资源管理系统（如IBM Platform™ Symphony）还可用于管理Hadoop环境内外的数据集成工作负载。

这意味着，虽然InfoSphere DataStage与数据可能不在同一个节点上运行，但却在同一个高速背板上运行，因而无需将数据移出Hadoop环境，也无需在速度较低的网络连接之间移动数据。

支持Hadoop的ETL可扩展性要求：许多Hadoop软件供应商纷纷宣扬一种理念：任何不可扩展的ETL工具与MapReduce pushdown集成后均可提供出色的性能，并实现应用程序横向扩展以执行大数据集成，但这种说法显然不真实。

没有非共享、大规模可扩展ETL引擎（如InfoSphere DataStage），企业势必会遇到功能和性能限制。越来越多的企业意识到，不可扩展的ETL工具与MapReduce pushdown之争无法在Hadoop中提供所需的性能水平。因此他们争相与IBM合作解决这个问题，因为IBM大数据集成解决方案以其独有的方式支持大数据集成的大规模数据可扩展性要求。

以下是依赖ETL pushdown会造成的一些累积负面影响：

- ETL包含大部分EDW工作负载。由于相关成本的影响，对于运行ETL的工作负载而言，EDW是一种非常昂贵的平台。
- ETL工作负载会导致查询SLA降级，最终需要您额外投资购买昂贵的EDW容量。
- 数据被转储到EDW之前未清理数据，一旦进入EDW环境将永远无法进行清理工作，继而导致数据质量较差。
- 企业持续严重依赖手动编码SQL脚本来执行数据转换。
- 添加新数据源或修改现有ETL脚本较为昂贵并且需要很长的时间，限制了快速响应最新需求的能力。
- 数据转换相对简单，因为无法使用ETL工具将较为复杂的逻辑推送到RDBMS。
- 数据质量受到影响。
- 关键任务（如数据剖析）无法实现自动化-在很多情况下根本无法执行。
- 未实施有效的数据治理（数据管理、数据沿袭、影响分析），因而响应法规要求变得更加困难且非常昂贵，对关键业务数据的信心更无从谈起。

相反，采用海量可扩展数据集成平台来优化大数据集成工作负载的企业，则可最大限度降低潜在的负面影响，更有效地通过大数据实现业务转型。

大数据集成最佳实践

决定采用Hadoop实施大数据措施后，如何在保护自己免受Hadoop可变性影响的同时实施大数据集成项目？

在与Hadoop技术的大量早期采用者共事的过程中，IBM总结了5个基础大数据集成最佳实践。这5个原则体现了成功实施大数据集成措施的最佳方法：

1. 避免出于任何目的在任何位置进行手动编码
2. 整个企业采用一个数据集成和治理平台
3. 可在需要运行海量可扩展数据集成的任何位置提供该功能
4. 在企业间实施世界级数据治理
5. 在企业间实施强大的管理和操作控制

最佳实践1：避免出于任何目的在任何位置进行手动编码

在过去的二十年中，大型企业认识到使用商业数据集成工具替换手动编码具有很多优势。手动代码与数据集成工具之争早已平息，很多技术分析师纷纷总结采用世界级数据集成软件将会实现的巨大ROI优势[3]。

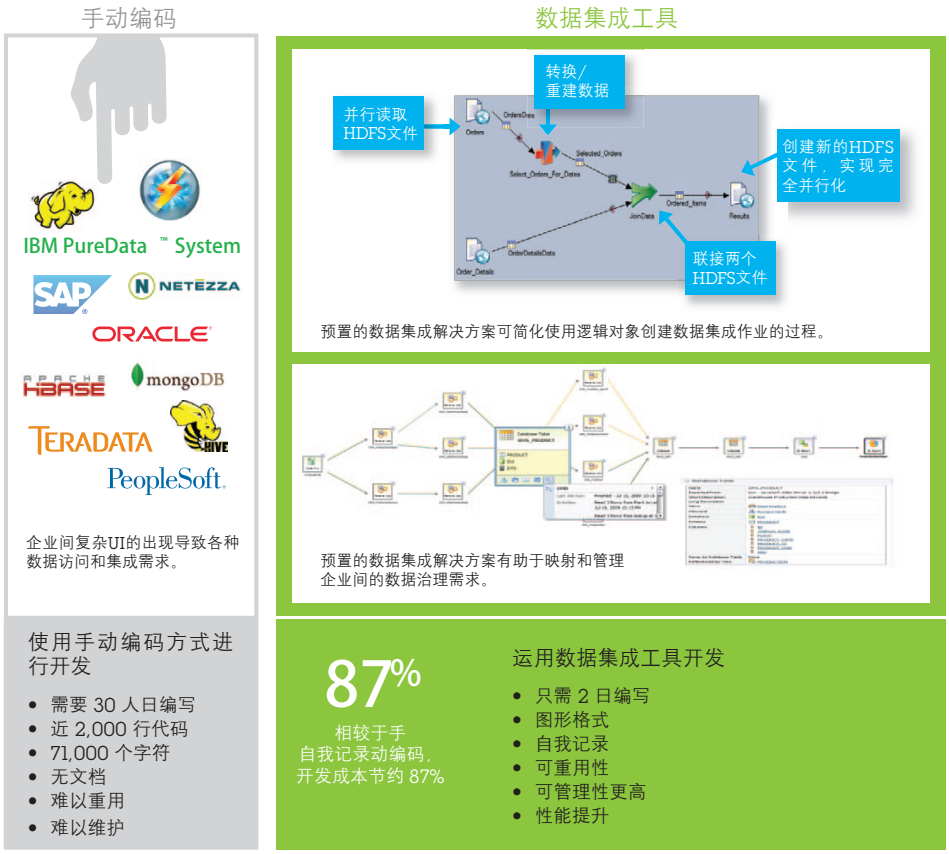
"如有疑问，请尽可能使用更高级的工具。"

— "Large-Scale ETL With Hadoop"，Eric Sammer（Cloudera 首席解决方案架构师）于 Strata+Hadoop World 2012 期间所做的演示[4]

第一项最佳实践是随时随地避免在大数据集成的各个层面采用手动编码。相反，利用商业数据集成软件提供的图形用户界面提供活动支持，如：

- 在企业中实施数据访问和移动
- 数据集成逻辑
- 通过各种逻辑对象组装数据集成作业
- 组装更大的工作流
- 数据治理
- 运营和行政管理

通过采用这项最佳实践，企业就能利用商业数据集成软件久经考验的生产、成本、价值实现时间以及强大的运营和行政控制优势，同时避免手动编码带来的负面影响（参见图4）。



手动编码和工具成果来源：IBM制药客户示例

图4. 数据集成软件提供多个GUI来支持各种活动。这些GUI取代了复杂的手动编码，为企业节约了大量的开发成本。

最佳实践2：整个企业采用一个数据集成和治理平台

过度依赖向RDBMS推送ETL（由于缺乏可扩展数据集成软件工具）会妨碍很多企业替换SQL脚本手动编码，更不要说在企业中建立有效的数据治理机制。然而，他们意识到将大型ETL工作负载从RDBMS迁移至Hadoop将会节约巨额成本。尽管如此，从RDBMS中的ETL手动编码环境迁移至ETL和Hadoop的新手动编码环境只会使高昂的成本和冗长的供货周期问题雪上加霜。

部署单一数据集成平台后，可通过以下功能为企业转型创造机遇：

- **一次构建作业，随时随地运行**-无需修改，即可在企业中的任何平台上运行该作业
- **访问、移动和加载数据**-在企业内的各种来源和目标之间均可实现这些工作
- **支持各种数据集成范式**，包括批量处理、联盟、更改数据捕获、为数据集成任务启用SOA、与事务完整性实时集成和/或企业用户自助数据集成

另外，还可以建立世界级的数据治理工作，包括数据管理、数据沿袭和跨工具影响分析。

最佳实践3：可在需要运行海量可扩展数据集成的任何位置提供该功能

Hadoop能以极低的成本对数据集成工作负载实施大规模分布式处理。但是，客户需要的是海量可扩展数据集成解决方案，从而实现Hadoop可以提供的各种潜在优势。

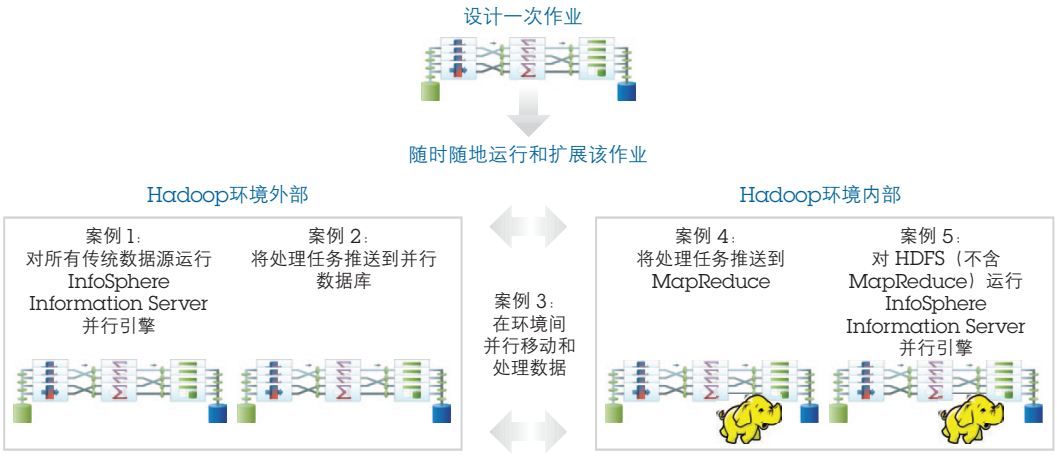


图5. 可扩展大数据集成必须适用于任何环境。

运行数据集成工作负载的场景可能包括：

- 并行RDBMS
- 不含RDBMS或Hadoop的网格
- Hadoop（包含或不含MapReduce pushdown）中
- Hadoop环境内外之间，在一端抽取数据卷，动态处理和转换记录，然后在另一端加载记录

为了实现成功和可持续发展并保持较低的成本，一项有效的大数据集成解决方案必须灵活支持上述各种场景。根据 IBM 与大数据客户的合作经验，InfoSphere Information Server是目前支持全部上述场景（包括向MapReduce推送数据集成逻辑）的唯一商业数据集成软件平台。

业界流传着很多有关在Hadoop中为大数据集成运行ETL工具的神话。流行的说法似乎是，组合使用任意不可扩展的ETL 工具与Hadoop均可提供全部所需的海量可扩展数据集成处理。事实上，MapReduce在处理大规模数据集成工作负载方面有着很多限制：

- 并非所有数据集成逻辑均可使用ETL工具推送到MapReduce。根据与广大客户的合作经验，IBM估计约有半数的数据集成逻辑无法推送到MapReduce。
- 用户不得不通过繁复的手动编码在Hadoop中运行较为复杂的数据集成逻辑，或者限制流程在MapReduce中运行相对简单的转换。
- MapReduce在处理大型数据集成工作负载方面具有多种已知的性能限制，因为其目的在于牺牲高性能处理来支持细粒度容错。

最佳实践4：在企业间实施世界级数据治理

绝大部分大型企业发现，在企业中建立数据治理机制即便是可行的，也会十分困难。造成这种局面的原因很多。例如，企业用户使用自己熟悉的业务术语来管理数据。时至今日，仍未出台任何机制来定义、控制和管理此类业务术语并将其与IT资产联系起来。

此外，无论是企业用户还是IT人员均高度信任其数据，但可能连数据出处和/或历史都含糊不清。根本不存在通过数据沿袭和跨工具影响分析等功能创建和管理数据治理的技术，并且手动方法会导致异常的复杂。行业法规要求只会进一步加大治理管理工作的复杂度。最后，严重依赖手动编码进行数据集成导致难以在整个企业中实现数据治理。

建立世界级数据治理机制至关重要，并为所有关键数据资产（包括Hadoop环境，但不仅限于此）创建完全受治理的数据生命周期。以下是创建全面数据生命周期的建议步骤：

- **查找**：利用条款、标记和集合来查找接受治理和监管的数据源
- **监管**：为相关资产添加标记、条款和自定义属性
- **收集**：通过收集来捕获资产，并开展具体的分析或治理工作
- **协作**：共享其他内容管理和治理集合
- **治理**：创建并引用信息治理策略和规则；应用数据质量、屏蔽、归档和清除操作
- **卸载**：单击HDFS来复制数据并执行分析，以便强化仓库
- **分析**：分析已卸载的数据
- **重用和信任**：了解如何通过沿袭功能运用数据进行分析和报告

通过部署全面的数据治理计划，您可以构建环境来帮助确保所有Hadoop数据具有出色的品质、安全可靠且适合使用目的。这可以帮助企业用户回答以下问题：

- 我理解这些数据的内容和意义吗？
- 我能衡量这些信息的质量吗？
- 报告中的数据来自何处？
- 这对Hadoop内部数据有着怎样的影响？
- 数据在抵达Hadoop数据湖之前存储在哪里？

最佳实践5：在企业间实施强大的管理和操作控制

采用Hadoop开展大数据集成的企业势必期望实现强大的大型机级治理和操作管理，包括：

- 操作平台界面，在操作数据集成应用程序的各方人员（开发人员和其他利益干系人）监控运行时环境时，快速回答他们的提问
- 工作负载管理，为共享服务环境中的某些项目分配资源优先级，在繁忙系统上对工作负载进行排队
- 性能分析，深入了解资源使用情况，辨别瓶颈并确定何时系统可能需要更多的资源
- 构建工作流，其中包括通过Oozie直接按作业序列定义的基于Hadoop的活动，以及其他数据集成活动

大数据集成的行政管理必须包括：

- 基于Web的集成式安装程序，用于执行所有功能
- 高可用性配置，用于满足全天候需求
- 灵活的部署选项，用于部署新实例或展开经过优化的专家硬件系统上的现有实例
- 集中实现身份验证、授权和会话管理
- 审核安全相关事件的日志记录，推动满足《萨班斯奥克斯利法案》合规性要求
- 实验室认证，针对各种Hadoop发行版

大数据集成最佳实践为成功奠定了坚实的基础

企业正在纷纷转向大数据措施，期望帮助自己削减成本、提高收益并实现先发优势。Hadoop技术支持新的流程和架构，有助于推动业务转型，但必须先行解决所面临的某些大数据挑战并把握相关机遇才能实现各项目标。

IBM建议构建一个大数据集成架构，该架构足够灵活，可充分利用RDBMS、ETL网格和Hadoop环境的优势。用户应能够构建一次集成工作流，即可在上述三个环境中的任意一个环境中运行该工作流。

本文列出的5个大数据集成最佳实践体现了筹备项目并实现成功的最佳方法。遵循这些原则有助于企业尽量降低Hadoop项目的风险和成本，同时最大限度提高ROI。

更多信息

如需有关大数据集成最佳实践和IBM集成解决方案的更多信息，请联系您的IBM代表或IBM业务合作伙伴，或者访问：**ibm.com**/software/data/integration

此外，IBM Global Financing可帮助您以最经济高效的战略性方式获得您的业务所需的软件功能。我们将与信用合格的客户展开合作，定制一个财务解决方案来满足您的业务目标，实现有效的现金管理，以及改善您的总体拥有成本。IBM Global Financing是您进行关键IT投资和向前推进您业务的最智慧选择。有关更多信息，请访问：**ibm.com**/financing

国际商业机器中国有限公司
北京市朝阳区北四环中路27号
盘古大观写字楼
邮编: 100101

在中国印刷

2014年12月

[1] Intel Corporation。"使用Apache Hadoop抽取、转换和加载大数据。"2013年7月。http://intel.ly/UX1Umk

[2] 测量结果由IBM现场进行客户部署时生成。

[3] International Technology Group。"企业数据集成战略业务案例：IBM InfoSphere Information Server与开源工具比较。"2013年2月。ibm.com/common/ssi/cgi-bin/ ssialias?infotype=PM&subtype=XB&htmlfid=IME14019USEN

[4] "Large-Scale ETL With Hadoop"，Eric Sammer（Cloudera首席解决方案架构师）于Strata+Hadoop World 2012期间所做的演示。www.cloudera.com/content/cloudera/en/resources/library/hadoopworld/strata-hadoop-world-2012-large-scale-etl-with-hadoop.html

IMW14791-CNZH-00