数字化转型时代的数据仓库

作者：Neil Raden，Hired Brains

Pivotal

HIRED BRAINS
Consultants & Industry Analysts

白皮书

目录

概要

观点：过去几年间大数据风起云涌，Hadoop 和 Spark，作为大规模数据管理和分析平台，获得了最多关注。而基于关系型数据库技术的数据仓库，则由于各种原因，经受审视，如同或将不再需要。但最近用户开始意识到，数据仓库使命的重要性，一如既往。过往推定的数据仓库在成本和运维方面的缺陷，则可通过云计算与开源软件的结合，利用同传统大数据项目相同的经济元素—即廉价的纵向扩展和横向扩展，予以克服。

实时（技术角度而言，“近实时”）收集数据并作出响应，显然在取代过往的分析节奏，这是一种更具互动性、临时性、甚至思考性（“我看到的这些分析的背后，发生了什么？”）的分析。这种对非常低时延分析的需求，若你愿意，可称为“即刻法”，驱动着大规模并行处理 (MPP) 关系型数据库技术，和分布式处理模型（如 Hadoop 和 Spark）的发展。数字化转型，将分析融入运营，大量依赖即刻法。

在 Hadoop 之前，数据仓库是唯一用于分析的数据汇集点，集成和治理了多数内部数据源。一些数据仓库的部署中，也集成了第三方数据，特别是金融服务数据、制药数据、零售数据，甚至有关经济和人口普查信息的政府数据。但数据仓库主要用于企业领域建模，并从“防火墙后”各个系统（如 CRM 和 ERP 系统）中获取数据。其他科技都不具备如此有效的集成技术。数据仓库不仅只存储数据，还被用于支撑非常多种应用，包括将数据分发至其他系统，其中主要是分析和 BI 工具。

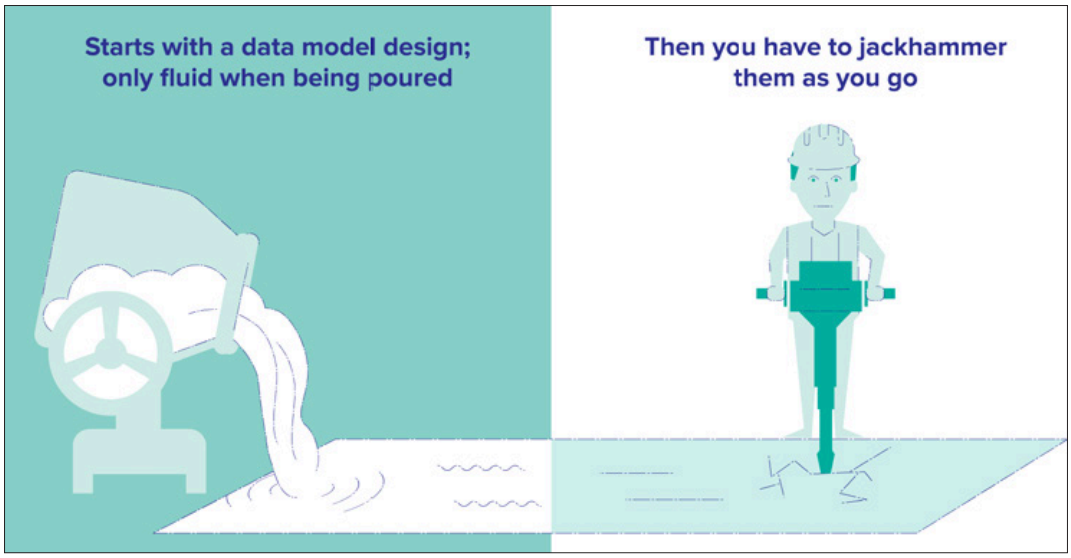
与关系型数据库过去所做的一样，Hadoop 和 Spark（严格来说不是 Hadoop，而是一个独立的 Apache 项目）已发展了新的数据处理特性，包括大规模并行处理能力、几乎无限制的可扩展性，以及与大型云数据存储无缝交互，并将数据存储和处理切分至不同逻辑位置的功能。但最重要的是，数据仓库延续的使命，依然是提供其他平台所不可能提供的业务价值。

本白皮书中，我们将说明进化过的数据仓库实现，将继续在企业中扮演关键角色，实际上具备辅助数字化转型的独特业务价值。

1 即刻法 (Instantology) 是我们发明的词汇，在此用于描述当前对实时流式数据和实时决策制定的关注。

数据仓库的使命，过去和现在

为了评估基于关系型数据库技术的大规模数据仓库是否仍然能够发挥作用，首先我们必须搞清楚什么是数据仓库，即数据仓库的整体概念是什么。数据仓库，这个名字本身，一直有些误导。举个类似例子，通用汽车公司在墨西哥销售雪佛兰诺瓦 (Nova) 时察觉，西班牙语中“诺瓦 (Nova)”的 意思是“不行”，但发现时为时已晚。与此类似，数据仓库是最不幸的科技词汇之一，因为它造成的印象如下： 存储信息的固定区域，且除此之外几无其它： 仅设计建造阶段具备流动性 ，一旦功成难于改变。



相反，一个有用的数据仓库，不仅是一个非常动态的区域，甚至可能根本不是一个单一的物理“区域”。但因为它与数据存储关联，随着新技术平台的兴起，尤其是 Hadoop 和 Spark ，一些质疑随之而来：数据仓库是否已过时。毕竟，数据仓库构建于昂贵的专有平台之上，无法灵活地扩展至 PB 级数据。而最糟糕的是，它采用“写时模式 (schema on write)”而非“读时模式 (schema on read)”，这是一种与大数据“可转换为多种形式”潮流相悖的方案。对于所有这些一直存在争议的问题，我们将在本白皮书中深入探讨。

数据仓库曾是一种相当“前卫”的技术，但随后，由于上述原因，有些人认为它有些“过时”。因其性质使然，数据仓库被认为不能很好地参与到各公司进行中的数字化转型中。但事实胜于雄辩。

数字化转型和“即刻法”的兴起

曾经，拥有一个由文员输入新信息或修改信息的工资系统，并无不妥。或许工资报告可以每月提供一次。销售代表可以将整理好的订单副本传真给制造部门，让制造部门了解需要什么材料，并将订单传真给供应商。以前，银行可以每年进行一次压力测试。所有这些运营时延都曾是可接受的，因为竞争对手、合作伙伴、监管结构和客户都理解延迟不可避免。以前打电话给 Dinimo 披萨，从条目有限的菜单上订餐，也曾是可接受的。但现在，您可通过智能手机自选材搭配披萨，并确切知道何时取餐或何时送达。这就是您的业务，已被数字化转型“包围”的一个事例。过去，在您面临交通拥堵时只能枯坐，不知还需要多久才能到达目的地；现在，您可获取实时交通信息，收到绕行路线和预计到达时间。这就是一个依靠全新技术解决亘古难题的事例。这些既是内部流程数字化转型的事例，也是全新应用的事例。正是数字化转型，使得我们可在家工作。

数字化转型不是一个新词。上世纪 90 年代，它被用于描述 ERP 和 CRM 系统热潮。今天，它的含义未变，只是技术已变。数字化转型，是一个随应用不同而变化的含义广范的词汇。但整体而言，它意味着随着组织、个人和政府的活动，越来越多地发生计算机和网络上。事务处理方式在改变，尤其是事物间连接方式在不断改变。因此，在这个数据四处流动的新世界中，数据不断产生也不断被利用，那么我们该在何时稍事停留，查看发生了什么呢？过去，也许 5 到 10 年前，答案显然是分析工具，如 BI 或 Excel，可将数据从数据仓库提取出来。但现在事态已变：“即刻法”已兴起。分析解决方案，主要是 ETL、数据仓库和 BI 工具，无法应对数字化转型伴生的数据流转速度和容量要求。

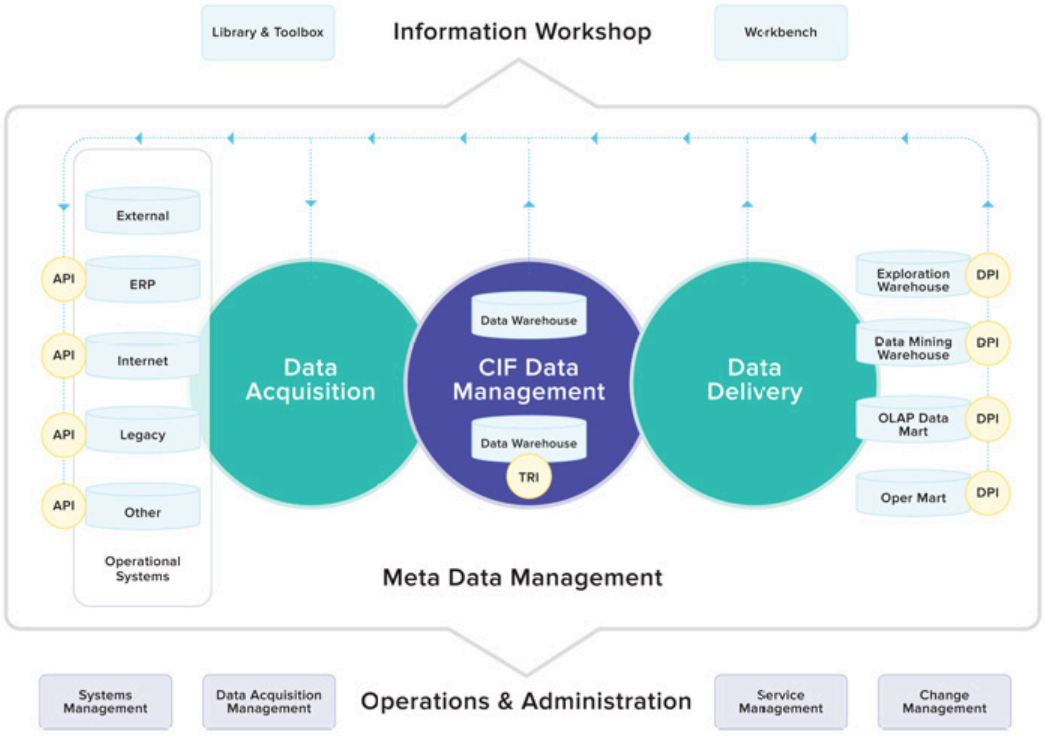
回顾历史

数据仓库一词，时间悠久，历经更迭。首次更迭时，人们（错误地）认为，可创建一个包含集成很多内部运营系统数据的、整洁、治理良好的数据库。这样做的主要动机，是让（IT 部门内部）开发人员更轻松地获取报告数据，进而减少 IT 部门的报告请求积压。这些早期尝试，发生在八十年代末至九十年代中，大多失败。原因多种多样，但最常见的包括：

- 误认为数据仓库开发前，需先建立“企业数据模型”。
- 无法就多源数据的通用语义模型达成共识，经常无休止地讨论。
- 应用的数据库设计，适用于事务 / 操作型系统，但完全不堪用于快速、批量加载和不可预知的即席查询的系统。
- 可用技术，包括数据库技术、计算、内存、存储和网络被不可预见需求压垮，且相当昂贵。
- 数据库规模庞大，压垮其构建和维护资源 。

前两个问题容易解决，其成因只是当时关系型数据库设计领域盛行的一种观点，即某些“规则”是不可侵犯的，这造成了很大阻力。但不久之后，许多组织意识到，数据仓库应是按业务领域精心设计的一系列模型。其负面影响是，造就了一些非常复杂的模型，及太多的伴生数据流，消耗了刷新时间（及数据新鲜度），并造成极度僵化，即便对模型进行简单的更改，都是困难且耗时的。

企业数据仓库主要参考架构之一，是企业信息工厂 (Corporate Information Factory)。如您所见，它包含至少七种不同的物理模式和复杂的单向数据流。奇怪的是，它没有规定数据的实际表现形式或使用。今天对数据仓库的许多批评，都是基于这些早期概念，实际上这些概念已得到精炼和简化。



显然，数据库技术需得到增强，才可应对数据仓库工作负载，其使用特征与 OLTP 系统 完全不同。现有供应商在这方面的工作相当糟糕，因为他们未能真正弄明白，如何构建优化器、负载平衡器、工作负载管理器等，以便在同一个软件中同时处理分析型应用和操作型应用。

最终，这些问题得到了解决。基于更好的设计和方法论的本地部署数据仓库，现在大行其道，但科技再一次影响到这些数据仓库。Hadoop 提出了一个严肃的问题， 是否还需要这些数据仓库，或至少它们该如何演进。当然，Hadoop 和 Spark 是可以加速某些数据密集型工作负载的宝贵技术。但一些重要应用，像为满足法定和监管要求的外部报告、合同合规、安全和人力资源，不应由分析师个人使用自助工具和 SQL“引擎”，对未认证数据源加工完成。**这些始终是数据仓库的职责所在。**

现在数据仓库应该提供什么

过去人们常用“应对稀缺”的观念设计数据仓库，即必须找到节省方案。因为一切都太昂贵了，这种考量必须放在首位。如今，200 节点和 2000 多核心构成的群集很常见。这意味着，必须具备无缝集成 Hadoop、云数据存储库和其他新兴数据平台的能力。访问这些数据源以满足查询的能力，现在是一项关键功能。抽取无法解决这个问题。由此引出另一个多年前已出现而现今有了实际意义的词汇：“虚拟数据仓库”。

虚拟数据仓库

人类未能改变任何物理定律以容纳大数据。一个数据仓库应无缝连接许多数据源（例如 Hadoop 和云数据存储库）以读取数据，来处理针对未经过整理和治理的非结构化数据的查询。例如，一个数据仓库的模式 (schema)，可包含实际存于数据库内的数据，但当中间件连接外部数据源时，一个针对此模式 (schema) 的查询仍可被执行。该数据仓库可提供全部所需服务，例如：

- 支持所有类型的数据局域化（本地磁盘、Hadoop、私有和公有云数据）。
- 数据库内高级分析。
- 处理空间、时间序列和 / 或文本原生数据类型的能力。
- 运行新的分析工作负载（包括机器学习、地理空间、图形和文本分析）的能力。
- 部署不受限制，可以部署在本地、私有云和公有云。
- 针对大数据的查询优化。
- 复杂查询队列。
- 基于模型而不仅是分片 / 区的大规模并行处理。
- 工作负载管理。
- 负载平衡。
- 扩展至数千并发查询。
- ANSI SQL 全集及扩展。

虚拟数据仓库方案的一大优势，在于查询对下游进程 (如 BI 应用) 是透明的。它为数据定位提供了灵活性，因为对数据仓库的查询是针对抽象对象的。数据可被重定向至逻辑位置，而无需作出更改。

据来自 BitYota 的 Poulomi Damany 所言 [2]，新数据仓库必须支持异构数据发现以供分析，而无需先对数据进行标准化、清洗和建模 。分析人员需能够用多元价值、多条件和多范围对所有这些非标准化、未定义的数据，提出智慧型问题，以理解这些数据的真实性质和价值。只有这样，才能设计模型 / 模式以生成 KPI 和相关报告。

2 http://www.enterpriseappstoday.com/business-intelligence/5-ways-data-warehousing-is-changing.html

这并非全部。为实现高效，数据仓库应能够支持对多态数据的查询优化，同时将数据结构分析延至运行时。（若解决方案能应对受限的 I/O 和片状网络，会额外加分。）

换言之，数据仓库既需修正其原始概念，也需摄取乍看之下似乎属于 Hadoop/Spark 世界的数据。数据的逻辑位置和处理将会有些模糊，具体设计将取决于您的特定需求。您可能发现，在数据仓库中存储似乎本应属 Hadoop 的数据，是一个不错的设计方案。如果文件中数据是“热的”，意味着必须及时满足频繁访问所需数据，这种方案可能合适。其他数据仓库部署杂糅了从未使用的数据，则可将其移到 Hadoop 或云存储库中的“冷”存储中。所有这些最佳实践才开始崭露头角。

进化后的虚拟数据仓库的一些其他功能包括：

- MPP 数据仓库能在本地、公有或私有云上无缝运行，完成相比先前设计大幅扩展的任务。
- 主要基于开源项目，背后有强大社区支持。
- 支持数据科学计算和保存以及数据科学模型的发布。
- 数据库内分析和数据科学库。替代方案是 Hadoop 或云存储库上运行机器学习算法，但需将结果迁至另一个平台进行进一步的分析和演示（可视化、场景规划的维度模型等）
- 支持对多态数据的基于成本式查询优化，同时延迟数据结构分析至运行时。

现在何处适用数字化转型？有哪些用例？

数字化转型并非仅仅事关“利用科技”的又一词汇。关键词是“转型”。数字科技如何改变我们的生活方式、组织运作的方式，甚至如何让教育机构重塑自我？请容我指出：拥有大量数据，包括流式和存好的数据，并不能帮您了解您做得如何，什么有效，什么无效以及采取何种应对措施。人们面临的各种问题（查询）很复杂，而一个指向一大堆无序数据的简单 SQL 引擎，将是无济于事的。产品演示通常展现非常简单的分析，但实践中，问题可以非常复杂：欺诈管理、风险分析、定价模型、优化库存、评估程序有效性等。

制定决策：Hadoop/Spark 还是关系型数据库？

CIO 们首次听闻 Hadoop，通常是通过在地下室做网页分析的草台班子，真正引起他们注意的是成本。起初，实施 Hadoop 比实施关系型数据库要便宜很多，原因有两个：物理硬件是“便宜货”，意味着只需把大量最便宜的服务器和存储捆绑在一起，而软件则是开源或“免费”的。关系型数据库通常部署在最佳的专用设备上，着眼于大规模的性能。但两件事颠倒了这种核算。首先，Hadoop 上的企业应用，开始需要更高质量的硬件，包括最新的多核处理器，甚至弃用现有轴式旋转磁盘，改用固态盘。还有，Hadoop 发行商，开始提供包含许可费和支持服务的“企业版”软件。其次，与此同时，数据库供应商开始将其产品服务迁移至云，有些甚至开放源码，于是关系型数据库和 Hadoop 的成本已趋于接近。结果是，从整体拥有成本 (TCO) 角度考量，如今 Hadoop/Spark 和数据仓库之间的成本差距已无关紧要。

正如我们前面提到的，“数据仓库”这个术语给人一种印象：它只是一个存储大量数据的地方，那么问题来了，为什么不在 Hadoop 中存储数据呢？事实上，数据仓库还执行许多更重要的功能。它是集成且治理好了的单一数据源，既包括历史数据也包括当前数据。它支持，并时常提供数据科学家使用的复杂分析和模型处理。因其关系型特性，它在原子级别存储和处理数据。Hadoop 文件系统 (HDFS)，和云数据存储库将数据存储为文件，因此对这些文件的内容进行操作，需执行更多分解工作，进而将结果传递给其他模块。

现今数据科学和人工智能获得许多关注，同时组织依然必须为满足监管和法规要求、准备报税、确保遵守很多地方、国家和国际机构要求，而生成清晰报告。Hadoop 的文件结构因其缺少成熟关系型数据库的许多特性，不支持这种“单一版本事实”。我们需多维细评 SQL 引擎，以及”SQL on Hadoop”解决方案作为数据仓库的理想替代物。SQL 处理仅是关系型数据库的一个方面。

要回答“为什么不在 Hadoop 中存储数据仓库数据”的问题， 部分答案如下：

	HADOOP	数据仓库
查询方式	批量	交互
数据库模式	发展中	高级
查询优化器	无	强大
最大数据	PB	PB
最多处理器	无限	无限
成本	增加中	减少中

为分析工作优化过的大规模可扩展关系型数据库软件，现在有了基于开源的产品，可在本地、私有云、公有云或混合云上运行。Hadoop 的成本优势已然下降，因为企业不满足于使用低端廉价硬件，而这恰是造就 Hadoop 价格优势的因素之一，并且 Hadoop 软件和分销商支持服务的成本超出“简单下载开源版本”（当然，这个免费）相当多。

数据仓库的实际应用

人们对数据仓库的普遍认知是摄取数据，以及或通过“抽取、转换、加载” (ETL) 模式，预先完成清理，或存入数据仓库中并在其中清理，后者称为“抽取、加载、转换” (ELT) 模式。不论哪种方式，数据仓库均定位为查询所用的数据源（包括元数据）。但很多案例中，并非单向流。数据仓库可用来创建反向生成结果的模型。例如，在一家人寿保险公司，数据仓库治理着来自二十个不同内部系统的源数据。从数据仓库中提取数据供给估值系统，该系统基于利率假设、死亡率和失效数，生成大量现金流信息。在同一案例中，源自数百次估值运算的精细时间序列现金流经分析后，用于生成偿付能力报告及其他法规要求的报告。

Pivotal Greenplum 案例研究：金融服务和保险

全球性金融服务公司

一家美国最大的全类别服务投资银行，为机构提供融资和财务咨询等服务，包括兼并和收购咨询、重组、房地产和项目财务以及企业贷款。它还设有股票和固定收益部门。作为一个组织，关注风险是持之以恒的过程。为不同类型风险建模，引出各种复杂的系统输入、建模场景和输出。

计算风险会生成数十亿个需整合的数据点， 而其每天的数据量每天可能超过 1 TB。所有这些数据都必须存储 10 年，且收到监管方调查 24 小时内需可恢复。风险报告既须准确，也要及时。若银行向监管方提供了不准确报告，则监管方可作出直接影响业务的决策。例如，美联储可根据一家银行的风险，控制其股票回购。

对任何风险汇总和报告系统而言，稳定性是关键需求。系统不能在高压阶段崩溃，需能够处理极度波动，有时是正常交易量的两到四倍。若此期间银行的风险报告系统崩溃，则可危及银行一整年的利润。

由于老系统的故障、近 400 亿行的数据量及每天数千份的复杂报告，现存系统无法可靠地满足监管方和内部分析的需求。因此，客户决定更新架构以降低基础架构成本、更好地满足既定 SLA、更高效率地满足监管义务。

在实施 Pivotal Greenplum 数据库后，他们能捕获每笔交易，每笔交易在数据库内累积 400 亿行数据，且分析师可分析不同场景，最终做出更好的知情决策。在先前的系统中，只有 60% 能满足 SLA 的需求。新系统则几近完美。

使用 Greenplum，此银行可捕获其多达数百的所有交易柜台的每笔交易 。他们能整合并对数十亿行数据进行计算，每天生产 2000 亿场景结果（输出）。该银行的风险管理人员切实受益。IT 能更快地提供报告给风险管理人员：日报所需的小时数减少，周报所需的天数减少。Greenplum 还支持数千使用各种工具的风险用户，这些工具包括 PL 语言、Excel、Tableau 和基于 SQL 的专有报告工具。

虽然监管方只要求每年进行一次压力测试，但该银行希望每天进行压力测试，以最大限度利用其营运资本。若无每日压力测试，便会趋于更加保守，“钱扔桌上”未赢回报。依赖更好地风险理解和风险管理，该银行可通过减少抵御不可预见风险的预留金，将更多资金用于投资，进而增收。

人寿保险公司

人寿保险公司的监管方主要关注偿付能力。由于寿险公司通常签订几十年后才到期的合约，监管方仔细监督保险公司业务实践、投资组合和投资行为，至关重要。 这家人寿保险公司由于几点原因，引起了监管方关注。

偿付能力研究的基础是投资组合估值。多数情况下，估值在轻度汇总级进行，而非每个保单，例如对 10 岁年龄组，或其它变量，进行十年级别汇总。但监管机构要求这家公司全面逐一估值—不仅对每份保单，而且分析每份保单保险范围，例如死亡抚恤金、意外死亡抚恤金和残疾豁免补偿。此外， 作为每年提供估值和所有相关研究结果的替代，监管方要求每季度提供上述结果，直到来年底，共约 18 个月时间。若不遵守，监管方将撤销其出售新保单的许可，而这对保险公司相当于死刑判决。

专门负责满足这些监管要求的精算部门，遇到了几个问题。当前流程高度依赖手工操作，从 20 多个源系统提取数据，其中很多系统的数据质量很差。该部门精算师费力地使用 COBOL、Fortran、APL、Easytrieve 等过时工具获取数据，然后尝试手工整理为电子表格。完成一次非逐一估值要耗费五个月，似乎不可能找到一种方案可实现 3 个月生成一次全面逐一估值。此类挑战本不应由精通精算科学而非数据转换的精算师来面对，故导致士气低下。该部门人员流动率达到近 50%，让问题进一步恶化。

幸运的是，首席精算师是一位有远见、有魅力的领导者。他向寻常的大型咨询公司寻求建议，同时也雇用了一家因擅长解决此类问题而声誉极佳的小型精干公司。在不到六个月时间里，他们构建并实现了一个 ETL/ 数据仓库 /BI 平台。他们将数据仓库与第三方估值工具紧密集成，实现了：

- 全面逐一估值
- 多司法管辖区整合
- 估值与场景测试整合
- 估值周期从每年变为每季
- 实际周期间隔从五个月缩短为两周

所有关于经验、保险到期、利率风险等的高级研究，都可轻松使用 BI 工具生成，因为关系型数据库模式本为极端灵活性而设计。一年期间，该公司能够运行数百种估值，并将详细的现金流发送回数据仓库，在那里可以轻松执行混合情景和假设 (what-if) 分析。对数据信心的提升，促使释放了 2000 万美元储备金，同时更出色的过程研究进一步提升了利润。其他优势还包括降低了精算部门的人员流动率，并大大改善了所有财务精算工作的流程。

紧急关头过后，该组织寻求应用技术和方法的其它方式。在实施阶段偏安配角的 IT 部门，接手了数据仓库，并决意整合其他“数字化转型”方面的驱动力。他们决定将该数据仓库改建为他们的“标准”数据库， 这暂时扰动了所有上游和下游工作进程，但一年内他们就实现了一个与数据仓库肩并肩的 Hadoop 系统。精算师可访问先前数据仓库中不具备的大量外部数据，能够在产品设计、定价和承保方面取得长足进展。

数字化转型在两方面让该公司受益。其一，通过满足监管方要求，真正挽救了公司。意料外的受益，是减少了精算部门的人员流动率。其二，建立了数据仓库同 Hadoop 装备协同工作的环境，该环境中的数据仓库流程推动了新产品和新方法的发展。

结论

数据仓库并未消亡。事实上，它比以往更加重要。过去，组织级的工作，包括开展营销并评估成效，分类账月结分析，或评估正式员工对比外包人员成本，虽不高效但都完全可通过导出一些数据至电子表格来完成。数据仓库和 BI，大大改变了这些流程。今天扩展中的数字化转型，不仅产生了大幅增加的数据，而且这些数据需得到快速分析。基于云的大规模并行和可扩展关系型数据库，专为数据仓库任务设计，继续扮演着关键和不可或缺的角色。若不能跟踪和度量，转型就无法成功。Hadoop/Spark 栈，无法满足这些需求。这是数据仓库的领域。