TDWI RESEARCH

TDWI 核对清单报告

云端数据仓库

作者：David Loshin

赞助方：

IBM

tdwi.org

tdwi

2015 年 7 月

TDWI 核对清单报告

云端数据仓库

作者：David Loshin

目录

Advancing all things data.

555 S Renton Village Place, Ste.700
Renton, WA 98057-3295

电话 425.277.9126
传真 425.687.2842
邮箱 info@tdwi.org

tdwi.org

序言

企业对于分析技术的热情持续高涨，中小型企业越来越希望采用商业智能 (BI)、报告和分析战略。过去，大型企业一直热衷投资硬件、软件和专业技能来打造企业数据仓库环境。然而，小型企业往往预算不足、缺乏技术娴熟的专业人员，或者难以下定决心设计、构建和支持必要的专用平台。

为扩展企业环境以提高数据仓库和商业智能能力，企业必需：

- 降低分析环境实例化的复杂性
- 降低进入成本门槛
- 使用自助分析工具快速访问数据，让业务用户不再依赖信息技术 (IT) 部门

令人欣慰的是，随着数据专业人员在实施通用设计模式和整合模型方面积累的经验日益丰富，数据仓库行业不断成熟，企业在端到端商业智能和分析框架体系架构方面普遍达成共识。

综合运用原型架构与日益精密的软件工具可以应对这一挑战。这样可以简化启动专用数据分析系统的过程，降低专业技能需求。随着标准化数据仓库和商业智能架构的融合、数据整合与数据发现技术的进步，以及基于云的成本效用计算模型的出现，供应商能够开发出基于云的数据仓库系统。

在这种范式的辅助下，供应商可以将工具套件实例化为外包平台，为广大企业社区实现分析战略目标创造大好机会。依靠服务提供商来实例化平台，可以简化企业利用分析技术优势的过程。运用基于云的数据仓库可以消除配置和管理平台所需的资本硬件投资和 IT 支持人员配置。这样，数据科学家、分析师和架构师就可以将精力集中在分析模型上 ,从而提高业务绩效。

通过本核对清单报告，我们希望分享最佳实践，帮助读者充分利用基于云的数据仓库解决方案。

第一章

运用数据仓库平台开展分析

对于不同的信息使用者而言，数据仓库概念可能会产生不同的影响。实际上，数据仓库涵盖的功能越来越多广泛。这些功能包括：定期发布且经运营经理审核的预定义报告；业务分析师为应对特定业务挑战而开展的即席查询和交互式深入分析；以及统计人员和数据科学家所使用的复杂预测性分析模型。实际上，这些不同类型的应用必定在数据可访问性、计算和复杂算法方面具有不同的功能和系统要求。

若能部署有效策略，数据仓库从业者将可以确定资源要求及所需的实用程序，从而快速满足每一位客户的需求。这表明，企业需要重新审视企业数据仓库，将它视为一个整体系统，旨在持续支持混合工作负载，而不是仅仅运用数据仓库平台达成特定的业务目的。例如，常规、定期和固定报告可能适用于成熟的多维星形架构数据集市，而预测建模应用可能更适合支持处理海量数据及执行并行计算的专业高性能分析设备。

基于云的数据仓库有助于促进报告与分析生态系统的持续融合。基于云的方法可以提高部署灵活性。当合作伙伴或服务提供商管理云数据仓库并帮助实例化模式和数据加载时，将可保证利益相关者专注于分析和结果，而不必浪费精力构建系统。

云部署项目往往要求十分严格。例如，短期项目、季节性分析、短期相关分析、限制性自助报告系统，甚至设计新的报告和分析原型。在此类项目中，采用基于云的数据仓库具有一定的价值，因为无需设计、开发及部署平台和数据管理框架。这种数据仓库不仅可以降低启动成本，还可以加速分析，减少乃至消除持续维护成本。

第二章

利用成本模型确定云商业智能的适用场景

在享受报告与分析环境带来的优势之前，数据仓库管理成本是必须克服的一大障碍。然而，大部分数据从业者大都不了解在整个系统生命周期中，数据仓库运营总成本包含哪些变量，这包括：

- **采购成本**，与评估和采购硬件、存储、软件和网络连接相关的成本
- **部署成本**，如项目规划、项目监督和管理、系统设计、开发、配置、测试和实施
- **数据开发和管理成本**，包括数据提取、数据整合应用设计和开发，以及数据仓库架构设计和实现
- **商机成本**，因系统运行延迟影响业务开展而产生的成本
- **运维成本**，涵盖电力、冷却、空间和通信成本
- **经常性成本**，如软件许可维护、系统升级以及数据归档、数据备份、恢复和灾难规划费用

不同企业对不同类别成本的容忍度可能有所不同。成熟企业可能更倾向于开展基础架构资本投资，因为他们深知由此获得的优势远远超过启动成本。小型企业或新兴企业或许尚不具备充足资本来支付长期经常性费用，因而可能希望实现短期收益。

开发成本模型，平衡主要开支会对价值实现时间产生影响。运用成本模型来确定最适合采用基于云的数据仓库的场景。有些时候，利用平台完成其他企业任务，即可将系统采购和管理成本分摊至多个项目。另外，外包系统的敏捷性也可能会带来回报 — 如果可以提前六个月使用基于云的系统增加收入，那么增加的收入或许可以抵消系统采购资本投资。

第三章

通过简化部署流程缩短价值实现时间

基于云的数据仓库和商业智能会大大简化部署流程。首先，大部分基础架构工作已然完成 — 服务提供商接下来将选择硬件平台和数据库管理系统。选择过程和平台管理对客户基本透明。

其次，客户将受益于服务提供商在综合利用各种工具来支持整体流程方面的丰富经验，包括数据提取、分析、转换、加载、报告和查询。利用提供商的数据整合、数据传输和展示经验可以简化数据开发工作。第三，许多云端数据仓库供应商通过整合更多复杂功能（如数据发现和可视化工具）以及整合预测和规范建模工具（如 R 建模语言中提供的工具），提供增值服务。

卸下底层基础架构工程任务后，客户将可以专心研究数据分析。实现快速部署流程标准化后，客户就可以满足信息需求。此方法至少应涵盖以下任务：

- **业务目标**：阐明企业利用数据进行报告和分析的目标，并向特定的用户社区展示这些数据集
- **数据需求评估**：确定填充数据仓库所需的数据集
- **信息建模**：考虑如何在数据仓库中组织和表达数据
- **数据整合**：开发并实施将所需数据迁入云平台的流程
- **基于规则的转换**：利用数据准备工具实现标准化信息的参数化转换
- **业务驱动分析**：确定要执行的分析类型，建立分析功能，实现预期项目成果

幸运的是，服务提供商执行团队可在机械层面支持上述众多任务，例如，数据建模、数据整合及配置基于规则的转换引擎。因此，采用标准化流程部署云商业智能 / 分析项目不仅有助于提高敏捷性，还能提高分析结果的可访问性。

第四章

寻找具有集成分析功能的云端系统

由于近年来商业智能、决策支持和决策分析方法日趋成熟，业务数据使用者的精明程度也随着技术的发展不断提升。虽然一些云端数据仓库提供商专注于开展直接报告和维度分析，但另一些企业则在快速整合预测性分析和规范性分析功能，其中包括：

- **聚类**，算法试图根据实体（如客户）特征和行为对其进行分组
- **分段**，根据事先创建的聚类模型区分实体（如供应商）的方法
- **分类**，采用迭代算法将个人分配至预定义类，如“最佳客户”、“良好客户”、“中等客户”和“不受欢迎的客户”
- **决策树**，面临多种决策选择时，用于进行分类或选择最优方案的标准
- **关联分析**，迭代审查数据集事件之间的关系，进而揭示关联，展现潜在商机

过去，很多功能需要单独的高级分析计算平台，但现在这些功能在架构创新中得到了越来越多的支持，如：

- 数据仓库设备，专门用于支持混合工作负载的平台，包括传统的查询和报告以及更高级的分析。
- 数据库内分析，数据库管理系统供应商通过设计将数据挖掘算法集成到更传统的 SQL 样式界面中，以便分析技术嵌入更熟悉的查询格式。
- 内存计算，数据库供应商优化自身的存储组织模式及数据访问模型，将最常用的数据（倘若无法存储全部数据）存储到内存而非磁盘。这样可以显著加快传统分析和高级分析的速度。

寻找一家适当的提供商，既要保证云服务支持预处理数据并将数据加载到数据仓库，又要确保在环境中提供广泛的分析功能。此外，服务提供商的产品和服务还应适应创新设计，以满足客户的特定需求。

第五章

确保云平台满足一贯的性能需求

任何托管应用都面临一大风险：提供商依靠虚拟化环境部署应用。这样或许可以降低客户的总体运营成本。但是，企业可能随时需要在不同的底层硬件上重新部署应用，而且可能与其他应用共存，运行这些应用可能会对客户的应用性能造成影响。

在大多数企业中，倘若无法面向所有数据使用者快速交付报告和执行结果，势必影响应用采用，继而影响应用的成功推广。请记住，您可能无法在虚拟化环境中保持一致的性能。如果企业需要保证性能可预测，请明确指定性能标准及可接受程度，并向外包服务提供商候选方传达相关目标。评估提供商的方法，以便保证或提升性能。提出以下几个问题：

- 云端数据仓库供应商是否提供性能基准，准确反映应用的预期运行效果？
- 供应商是否提供在“裸机”云平台（而非虚拟化平台）上部署项目的选项？
- 可否使用架构增强功能配置平台，以便加速执行查询及呈现结果，比如使用柱状数据对齐或内存处理功能？

联合供应商，共同确保满足您的性能要求。此外，确保明确定义报告及解决性能缺陷的协议。

第六章

主动管理数据连接性和可集成性

如果考虑采用基于云的商业智能和分析技术，务必认识到这需要轻松迁移数据，以便在云环境中进行报告和分析。尽管填充小型数据集市所需的数据量似乎并不那么令人生畏，但仍需意识到数据连接与整合期望及相关成本远不止数据迁移和加载那么简单。为此，还必需考虑理解、准备及整合各种数据源等复杂问题。此类数据源可能包括平面文件数据、使用 SQL 访问的关系数据库管理系统数据、新 NoSQL 环境中管理的数据、地理空间数据以及 Hadoop 上的 HDFS 文件等。

制定计划，主动管理数据连接与整合。倘若不考虑以下因素，您的计划将不够完善：

- 各数据源与云端数据仓库之间的**网络连接**。也就是指贵企业的环境关联，以及其他 SaaS 和云系统管理的数据源的可访问性。
- 备用**数据迁移**方式，如果数据仓库数量超出标准网络连接容量，可能需要提高带宽来加快连接速度。
- **数据归档和分析**，评估潜在异常，揭示与后续数据转换相关的嵌入式结构、元数据和业务规则。
- 将**数据标准化和转换**业务规则作为持续数据准备的一部分。
- 通过**复制和更改数据捕获**降低刷新整个数据仓库的相关开支。
- 采用**数据压缩**作为备选方案，缩短将任意来源的数据迁移到云仓库所需的时间。

用户迫切需要寻找更多不同来源，发掘的数据量也随之不断增加，因而需要实施更复杂的整合。寻找云端数据仓库供应商提供工具，特别是支持数据分析和发现、压缩、传输、数据准备及高效数据加载的服务。

第七章

满足安全和数据保护要求

使用托管或云端数据仓库还面临另外一项重大风险：很可能违反企业的数据安全策略或监管指令。根据传统思维，人们可能想当然地认为保障访问安全性与数据保护存在一定的不确定性 ,原因有两个:首先 ,在某些情况下，多租户架构允许在同一环境中运行多个客户应用，因而人们担心发生跨应用边界的数据泄露。其次，虚拟平台存储可能分散于多个物理机，因而很可能引发恐慌，导致人们对迁移应用后删除“残余”数据的能力表示担忧。

显而易见，企业必须执行尽职调查，评估安全与数据隐私保护需求，确保供应商有能力满足这些需求。云端数据仓库供应商可能提供以下方法：

- 用户身份验证和用户授权，防止未授权的数据访问
- 细粒度的数据访问控制，防止暴露受保护的数据属性
- 数据屏蔽，防止显示受保护的数据属性
- 数据加密，很可能适用于“静态”数据、存储数据，还有经受访问及交付至用户门户的“动态”数据
- 数据擦除，用于完全覆盖硬盘驱动器，以防止恶意恢复

随着供应商在识别和解决现有及潜在安全漏洞方面越来越主动 ,将云系统视为数据保护风险的观念也逐渐消退。尽管如此，还有另外一种消除数据暴露恐惧的方法，这种方法与此前就确保可预测性能提出的建议不谋而合：寻找提供基于“裸机”云平台（而非虚拟化平台）部署项目选项的合适供应商。隔离应用将可避免现存的虚拟化和多租户风险。

后记

选择供应商并与之建立良好关系

此清单中的建议为确定云端数据仓库是否适合贵企业提供了相关参考。一旦决定将数据仓库和商业智能应用迁移至云服务提供商序，务必确保寻找适当的服务提供商。概括而言，我们提到的用于评估云端数据仓库服务的部分标准侧重于：供应商的产品对您的报告和分析程序带来的补充效果，如下所示：

- 降低总体开发和运营成本
- 缩短价值实现时间
- 减轻对内部 IT 资源的依赖
- 简化数据提取、整合和加载
- 通过提高易用性为您的数据使用者社区提供支持
- 支持您的弹性和可扩展性需求
- 通过容错和外包故障转移实现业务连续性
- 建立系统安全性和隐私信息保护信任

一经确定供应商，我们建议您与可信的云端数据仓库供应商建立良好的合作关系，这一点非常重要，原因有以下三点：

- **环境可持续发展：**值得信赖的合作伙伴不仅可以确保环境能够满足数据仓库生命周期各阶段的所有业务分析需求，还能满足项目生命周期内不断增长的弹性和可扩展性、安全性和总体性能需求。
- **响应能力：**高价值服务提供商可以证明自身值得信赖，有能力及时可靠地解决出现的任何问题。
- **参与度：**适当的提供商不但可以帮助您快速部署系统，还能*联合*您及您的数据使用者推动报告、商业智能和分析程序不断成熟。

云端数据仓库供应商可以利用自身的客户实施经验，根据客户的短期、中期和长期战略综合调整。

关于我们的赞助方

ibm.com

IBM Cloud Data Services 为开发人员提供一整套丰富的综合数据服务，广泛涵盖内容、数据和分析。Cloud Data Services 产品可缩短上市时间、延长正常运行时间，为 Web 和移动应用开发人员创造更高的价值。有关 IBM Cloud Data Services 如何改革开发人员服务创建和交付模式的信息，请关注我们的 Twitter @getdashDB 和 @cloudant，以及访问 www.dashdb.com 和 www.cloudant.com。

关于作者

David Loshin 现任 Knowledge Integrity, Inc. (www.knowledge-integrity.com) 总裁，他是公认的思想领袖，兼任 TDWI 讲师，同时也是数据管理和商业智能领域的知名专家顾问。David 是商业智能最佳实践领域的高产作者，发表过大量数据管理书籍和论文，包括 *数据质量改进从业者指南*，另外在 www.dataqualitybook.com 上也发表了很多其他内容。David 经常应邀前往各类会议、网络研讨会及赞助网站和频道发表演讲。David 的联系方式为 loshin@knowledge-integrity.com。

关于 TDWI RESEARCH

TDWI Research 面向全球数据专业人士提供研究结果和建议。TDWI Research 专注商业智能、数据仓库和分析问题，携手行业思想领袖和从业人员全面深入地了解部署使用商业智能、数据仓库和分析解决方案所面临的业务和技术挑战。TDWI Research 针对用户和供应商机构提供深入的研究报告、评论、咨询服务、主题会议及战略规划服务。

关于 TDWI 核对清单报告

TDWI 核对清单报告概括阐述了特定商业智能、数据仓库或相关数据管理学科项目的成功因素。企业可以在启动项目前按照本概述进行组织，或者确定当前项目目标和改进领域。