2017 年 8 月

数据仓库设备和崭新的分析世界

Claudia Imhoff 博士

赞助方：

IBM

目录

简介

毋庸置疑，在 25-30 年的发展历程中，分析架构一直在不断演进。自从数据仓库这个简单的单一数据存储库概念出现以来，近期的众多创新对分析架构产生了重大影响。首先，数据仓库设备 (DWA) 以及 NoSQL 革命、自助分析和其他趋势的兴起对传统架构造成巨大的冲击。其次，数据科学、实时运营分析和自助服务需求的出现无疑对分析架构产生了显著影响。

仅凭单一数据仓库存储库已无法满足全部分析需求。 扩展型数据仓库架构 (XDW) 这种全新的架构理念逐步取代了单一存储库概念。扩展型数据仓库不仅适应新的数据形式和数据量要求、满足不同子环境的不同分析需求，而且支持采用层出不穷的创新技术。

在本文中，我们将重点介绍数据仓库设备及其推出多年以来的演变过程。同时，还会介绍扩展型数据仓库架构，说明采用这种架构不仅需要维护数据仓库，还要新增组件和功能以扩展分析功能。另外，这一部分还会讨论设备在扩展型数据仓库架构中的适当用法。在本文的剩余部分中，我们还会介绍实施数据仓库设备的优势、选择注意事项，以及数据仓库设备的未来发展趋势。

数据仓库设备是什么?

早在 2000 年初，术语 “数据仓库设备” 开始流传开来[1]。当时，人们将它定义为:

“一种集 CPU、内存、存储、操作系统 (OS) 和 RDBMS 软件于一体的交钥匙式、完全集成的堆栈，专为处理数据仓库和商业智能工作负载而构建和优化。” [2]

根本上而言，它是一种嵌入式专用高级分析平台，旨在帮助分析企业满足乃至超越业务需求。究其根本，数据仓库设备是由一家数据仓库和分析专业供应商提供的综合软件和硬件包。

1 来源：https://en.wikipedia.org/wiki/Data_warehouse_appliance
2 来源：http://www.infostor.com/index/articles/display/293088/articles/infostor/top-news/introducing-data-warehouse-appliances.html

起初，数据仓库设备是一种本地“黑盒子”，用户只需加载数据，并附加人们所钟爱的商业智能技术即可。 这种设备专为执行高性能分析而设计，通过预配置的易用系统实现。老实说，这种黑盒子的特质确实令 IT 界深感不安，稍后会详细说明。

此类设备逐渐演变成为专业集成系统，不仅内置分析功能，还能提供精简用户体验。部署方便，无需调试，维护需求微乎其微。采用最新数据库功能，如内存功能、列式存储、MPP 架构、数据忽略和数据压缩，在此仅举几例。

而今，现代数据仓库设备提供了容器化软件环境，终结了过往的“黑盒子”时代，演变成为一种私有云计算平台。能够充当私有云平台的数据仓库设备具备以下优势：

- 提供数据科学家所需的灵活数据湖计算平台及其他灵活自助访问解决方案；以及
- 与数据仓库内部的分析应用和开源工具无缝整合

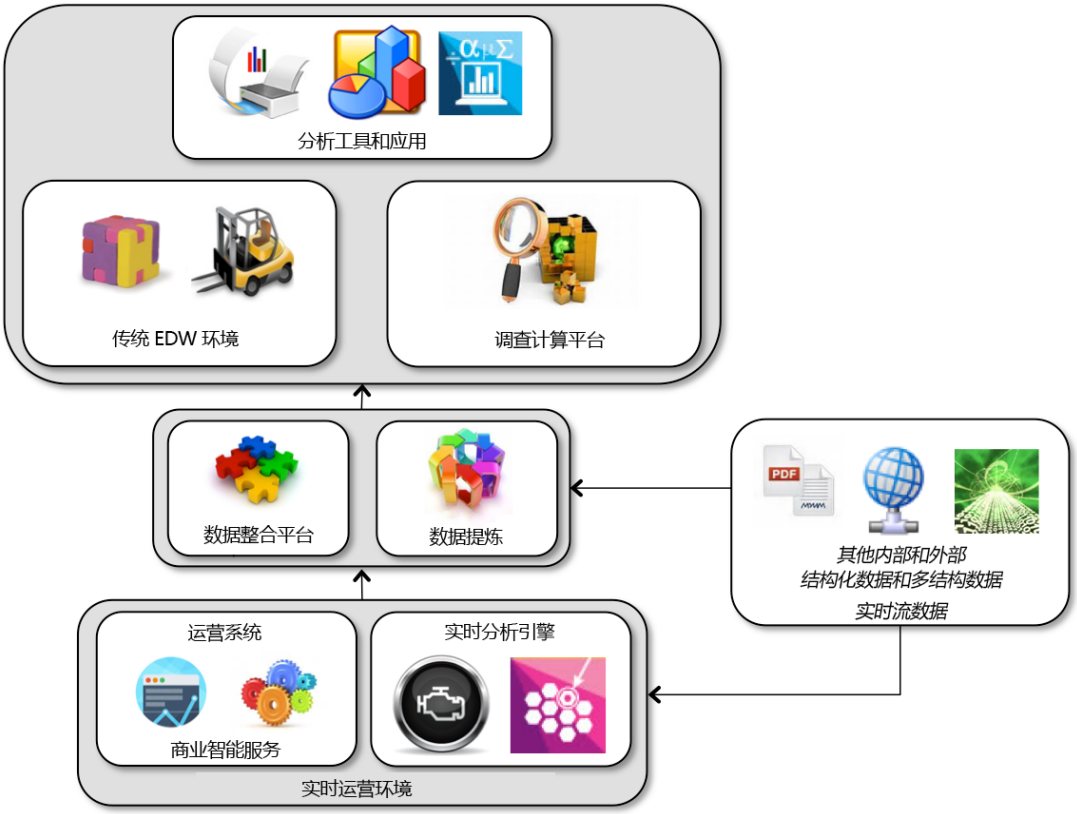
扩展型数据仓库：新型分析架构

近年来，总有文章、研究论文和演示文稿一再宣称：“数据仓库已经退出历史舞台。”这只不过是一种浮夸的营销手法，完全是为了吸引公众眼球。遗憾的是，数据仓库依然活跃在市场；更可惜的是，这种说法完全错误。

数据仓库不仅没有消亡，而且发展势头良好。当然，数据仓库曾经的*唯一*分析来源的头衔不复存在。这不禁令我们产生几个疑问：现代分析架构还需要哪些其他组件？增设组件后，新架构将会变成什么样子？

答案是扩展型数据仓库 (XDW)，它是一种综合分析架构，其中涵盖大量分析源，包括饱受诟病的数据仓库（见图 1）。

图 1：扩展型数据仓库架构



扩展型数据仓库架构还有很多其他曾用名，比如逻辑分析架构、混合环境或改进的数据湖环境。

扩展型数据仓库架构组件

首先来看看传统企业数据仓库 (EDW)，回答下面这个问题：“在新的分析环境下，企业数据仓库是否仍占据一席之地”？答案无疑是肯定的 – 至少在可预见的分析技术发展前景下如此。不过，它的角色发生了变化。企业数据仓库成为成熟的标准*生产*报告、历史数据比较和分析来源。数据仓库依然是执行关键或敏感分析所需的可靠、一致、集成的高质量数据的最佳来源，根据财务、合规或监管要求执行的分析尤其如此。与此同时，也是标准仪表板组件（想想企业 KPI）和标准度量指标（如运营、营销、销售及其他部门采用的盈利能力度量标准）的重要来源。没有什么比这一主力工具及其相关

数据整合与质量 / 分析流程（数据整合平台）更有利于实现这些意义重大且值得信赖的生产分析交付成果。

但必须承认，企业数据仓库也存在一定的局限性 – 在处理不寻常的数据类型和数量、执行实验或调查分析以及开展实时分析时表现尤为明显。按照这些新要求，人们需要从传统企业数据仓库中迁出部分分析操作，将它们转移到新的分析架构组件。

第一个新组件是调查计算平台，也有人称之为数据湖。该组件用于*探索*大数据源及执行专用分析，如数据挖掘、因果分析、“假设”探索、模式分析，以及一般性计划外数据调查。一些企业可能将调查计算平台用作简单的实验沙盒；另一些企业则将它实施为完整的分析平台，或者用作数据提炼扩展（如下所述）。在该组件的帮助下，企业能够以惊人的性能自由分析大量数据并利用它们开展试验。企业数据仓库可以运用这些活动的输出，因为企业数据仓库不仅可以作为运营环境的实时分析引擎，还适用于业务部门应用（嵌入式商业智能）。

下一个组件是数据管理领域的新成员 – 数据提炼。 这个组件用于从新的和不常见的大数据源（如传感器、社交媒体、IoT 和 RFID 标签）批量和 / 或近实时地提取原始详细数据。接着，这些源将数据加载到关系型或非关系型外包数据存储系统。数据提炼 – 类似于炼油厂 – *提取原始（大）数据*，精炼出有用又易用的信息，再将信息分发至其他组件（主要是调查计算平台）。 支持数据准的技术十分适合这个全新的数据管理领域。

最后一个数据仓库架构扩展通过在运营环境中添加实时 (RT) 分析平台支持运营智能。这个组件用于开发和 / 或部署实时分析应用，或者*流式分析*应用，例如欺诈检测、Web 事件分析、流量优化和风险分析。嵌入实时分析平台的模型和规则很可能在企业数据仓库、调查计算组件或实时分析平台本身开发而来，需要紧密整合并确保与这些组件实现数据自由流动。

数据仓库设备的适用场景有哪些?

扩展型数据仓库架构是一种逻辑架构；实际实施方法取决于构建过程中所使用的资源。例如，团队可能使用数据仓库设备作为企业数据仓库，使用 Hadoop 作为调查计算平台，同时使用复杂事件处理或事件流处理产品执行流式分析。

令人欣慰的是，现代数据仓库设备可以处理多个扩展型数据仓库架构组件。数据仓库设备（如 IBM Integrated Analytics System、PureData for Analytics 及其在数据存储、性能和可扩展性方面开展的一系列创新）十分适合企业数据仓库和调查计算平台。另外，还很适合执行低延迟运营分析。在云实现过程中，数据存储弹性意味着它们可以支持具有不同存储需求的企业数据仓库和调查计算平台。

要执行流分析或真正的实时数据分析，就需要一种截然不同的架构，有别于企业数据仓库和调查计算平台的先存储后分析架构。实时流引擎首先分析数据流，而后可能会存储数据。 因此，数据仓库设备可能不适合执行此类分析。

确定在扩展型数据仓库架构的哪个位置使用哪项技术，取决于大量因素。以下是需要考虑的几个方面：

1. 通常，首先要考虑总拥有成本。倘若企业预算紧张且技术资源有限，应考虑采用数据仓库设备。“外包”分析环境构建和维护工作听起来很吸引人。

2. 预期分析所需的数据量是仅次于成本的第二大考量因素。选择技术时必须平衡数据存储易用性与最大数据可扩展性和性能。同样，云端数据仓库设备产品具有极大的优势，可以根据数据和工作负载要求快速轻松地扩展或收缩数据存储。选择本地版本时，必须着眼于未来的新兴数据需求，因此要为迎接日后需求做好准备。

3. 必须考虑实际分析类型。要利用环境执行简单的描述性分析，比如报告和比较分析？还是利用环境执行更复杂精密的分析，比如预测和规范分析？或者利用环境支持生产分析及数据科学家提出的出乎意料的意识流分析？又或者执行流式分析和传统历史数据分析？分析环境可能必须支持一种、两种或全部三种分析功能。

4. 最后，确定适当的技术时，务必考虑基础数据管理功能复杂性。现代数据仓库设备简化了数据采集流程。 需要通过内存优化数据库或等效数据库实现查询性能和低维护目标（无需优化）。必须具备简洁易用的管理 UI，而且能够处理各类数据并适应多种数据量。最后，应具备内置分析（数据库内）算法和模型，如线性回归和 k 均值聚类，以及地理空间扩展。

除了作为企业数据仓库和调查计算平台存储库以外，数据仓库设备还适用于其他用例，包括自助数据访问、动态工作负载管理、存档数据查询功能（特别是与云或 Hadoop 数据存储结合使用时）和数据准备（例如，作为数据湖组件的一部分）。

数据仓库设备的优势

我们已经讨论了数据仓库设备的一些优势，下面再罗列几项更重要的优势：

- 扩展型数据仓库架构这类分析架构十分复杂，其中包含大量可移动的部件。如果实施者决定使用多家不同供应商的技术，或者技术供应商对企业缺乏了解，情况势必更为复杂。数据仓库设备是一项重大突破，因为它消除了采购实体面临的所有复杂问题。传统硬件和软件组合往往涉及多家供应商，而数据仓库设备可将供应商减少为一家，即一家供应商负责整个堆栈。核心宗旨：一次呼叫，一个联络点。

- 数据仓库设备预配置的功能是另外一项重大优势。究其根本，数据仓库设备旨在提供自我管理、自我调整、即插即用的数据库系统，同时确保系统可以经济高效地进行模块化扩展。操作极为简便，无需配置，而且具备线性可扩展性。部署方式灵活，支持本地部署和云端部署。您将借此获得一个非常适合各类企业和各种分析需求的环境。

- 数据仓库设备环境的另一大优势在于，托管数据将得到非凡的安全保护。例如，IBM Integrated Analytics System 和 PureData for Analytics 自动加密静态数据和传输中的数据，实施数据库活动监控，通过数据库访问控制进行用户授权的和部署加固（防火墙后），消除端口扫描及其他网络安全威胁。

- 倘若需要执行快速分析，但出于监管或隐私 / 安全顾虑无法实施云计算，那么本地数据仓库设备无疑是一项不错的解决方案。本地版本具有人们期望从数据仓库设备中获得的所有优势 - 易于安装、无需调优、即插即用功能、可扩展性和优良性能。

然而，还有一项优势并不那么广为人知。最近提出的构想是，这类设备还能用于其他用途。具体来说，是将数据仓库设备用作其他非分析应用的“容器化”环境。数据仓库设备环境可能不只适合处理分析工作负载；事实上，也可以在数据仓库设备上加载其他侧重运营（而非分析）性质的应用。例如，除用于满足分析需求以外，您还可以使用数据仓库设备运行 CRM 系统或办公工具。

数据仓库设备选择注意事项

数据仓库设备确实具有显著优势，但某些企业在选用时也不免要考虑适用性问题。

- 如前所述，一些 IT 组织对黑盒技术持谨慎态度。对于很多 DBA 及其他数据库管理人员而言，无法调整或配置环境中的任何元素堪称魔咒。一旦出现

问题，他们该如何　应对？另外，可能还会担心作业速度放缓或者内容丢失。为保证妥善运行设备，基本原则在于供应商具备深厚的专业知识，对数据仓库的各个方面了如指掌。因此，他们负责配置硬件和软件，确保完美处理提交的数据和查询。

- 虽然数据仓库设备确实扫除了硬件和软件设置和配置方面的诸多技术障碍，但实施者在整合及清理数据过程中的日常繁重工作却丝毫没有减轻。有些时候，只需装入即可立即执行，特别是处理近似或实验性质的分析。而有些时候，则必须对数据进行整理并证明数据准确性。数据仓库设备只不过是一种数据存储库，因此要确保数据达到目标用户所期望的适当整合度和准确度。

- 最后，在采购数据仓库设备技术时，应确保供应商在部署选项层面提供最大的灵活度。数据仓库设备技术应支持“一次设计，随处部署”的理念。团队可以从本地数据仓库设备入手，而后再确定云版本是否更适合企业使用。如果未来可能采用数据仓库设备，那么无论在何处（云端或本地）部署，启用同类技术环境的供应商都是最佳选择。这样，采购企业不仅轻松完成行政管理工作（团队不需要利用不同技能管理不同系统），还能更方便地管理工作负载(使用同一个引擎,便于迁移工作负载,而且无需调优或配置)。

数据仓库设备的未来前景

未来几年，数据仓库设备将在分析环境中占有一席之地。鉴于总体 TCO 和用电量下降，而且云平台简便易用，人们有充分的理由采用数据仓库设备。

数据仓库设备灵活、可扩展而且十分安全，因而能够为扩展型数据仓库架构的各类重要分析组件提供良好的支持。毋庸置疑，数据仓库设备不仅可以支持传统企业数据仓库及其生产分析，还能满足调查计算平台的实验性和探索性大数据要求。

鉴于数据仓库设备快速装入即可立即处理海量数据、记录性能卓越，而且支持处理混合工作负载，因而堪称通过简短查询执行复杂高级分析的理想环境。 因此，数据仓库设备环境不仅适宜传统商业智能报告和多维分析，还支持企业用户和传统业务分析师及先进的数据科学专业人员执行自助分析。