现代分析平台的构成要素

作者：Dan Kogan、Jen Underwood

什么是“现代分析”，我们为什么需要现代分析？

我们生活在一个令人振奋的时代，其中伴随着不断加速的创新、日益激烈的全球竞争，以及前所未有的打破常规和重新创造的机遇。飞速发展的数字技术、无处不在的互联网连接和大规模普及的智能设备 - 所有这一切共同创造了海量数据，能够快速、准确理解这些数据的任何人都将获得有竞争力的优势，这势必开启第四次工业革命。

现代技术的潜能不容忽视。现实世界与虚拟世界之间的严格分界逐渐消失，市场进入屏障逐渐减小，创造着全新的商业模式。各行各业都已准备好突破，即使十年前的突破性行业也是如此。想想众筹融资对小企业贷款的影响，线上零售对广阔的零售市场的影响，或是订阅媒体服务对传统有线电视的影响。在您身边，这种转变无处不在，并且正在加速进行：众多的新技术正在推动人们走向令人振奋而又未知的未来，比如自动化过程、信使机器人、人工智能，不一而足。

若要详细了解商业智能领域最受欢迎术语的变化，请参阅**“定义分析”**，此白皮书中着重介绍与分析相关的当红关键词。

但是，所有这些趋势都有一条共通的线索交织其中：海量数据。

如果说数据是原材料，那么分析就是将数据提炼成有用的信息的过程，这一过程最终将为企业带来巨大的竞争优势。现在，数据变得空前重要。为顺应行业转变而寻求发展，企业必须依赖于准确且具有时效性的数据，快速作出明智决策。但传统分析技术不仅速度慢，而且繁琐，无法适应数据量的增加和形式的改变。如今，似乎每周都会出现一种新型数据库，每套新设备都在产生数据，每个数据库都是采用全新技术构建的。分析却显然没能跟上步伐。

若要详细了解 IT 与业务部门的合作对实现现代企业分析方法的重要作用，请参阅我们的白皮书：**How to Build a Culture of Analytics（如何建立分析文化）。**

为了在当今的数字时代创造数据驱动文化，以便随时准备好准确快速地应对当今的业务挑战，组织要投资的不仅是新技术，还有交付信息的新途径，包括引领变革所需的人员。这种同步进行的文化转变是 IT 与业务关系的根本改变：IT 与业务是合作伙伴，他们合作收集和挖掘数据，并按需提炼数据、提供适当信息。只有 IT 与业务部门共同努力，组织才能将自助式分析的空想变为现实。

> 在向自助式分析转变的过程中，IT 组织可以通过发挥领头作用来确保大规模的管控和安全性。而通过让业务部门获得数据推动力及灵活性，IT 将成为业务部门信任的合作伙伴。
>
> - DOMINO’S PIZZA 有限公司 CIO，COLIN REES

本白皮书将探索现代分析平台的构成要素，业务和 IT 可共同使用这样的平台，为整个公司提供数据、实现价值、作出决策。这不仅涉及当代新式工具，还涉及几十年来作为商业智能基础的传统工具。我们将介绍每种构成要素在更大规模地将数据转化为见解的流程中有何作用，您是否需要这些工具来捕获和报告数据，或者是否需要利用这些新技术来共享交互式见解。我们还将探讨 Tableau 为何既是现代分析平台的基础，又是创造新式数据驱动分析文化的催化剂。

目录

1. 当今的 3 大数据挑战

数据无处不在，并且时刻都在大量生成。例如，现在已经有智能牙刷，能在您刷牙时记录刷牙时长及其内置部件的状况。它还能将这些信息发送给您的牙医。现在，一个简单的任务就能生成数千数据点。如果现在有数百个智能牙刷，那么单是这一个小型产业就能创造海量数据，而这种规模的数据几年前根本不存在。再加上事件日志、API、社交媒体、网站跟踪和大量的其他互联网技术，数据时刻都在呈爆炸式增长。

这种现代生态系统带来了三项业务挑战：

1. 数据无处不在
2. 每个人都需要数据
3. 数据瞬息万变

数据无处不在

图 1 Tableau 连接到位于各处的数据

以前，大多数组织都在本地保存数据。对于在看似设计合理的数据仓库中创建和存储的所有数据，组织需要努力进行控制。如果有任何未捕获到的数据，则表示此数据不重要，可能根本不值得捕获。

若要详细了解向云端迁移的当前趋势，请仔细阅读我们的**云数据简报**。

在当代，这种心态可能使企业走到尽头，因为现在网站、移动设备和云应用程序都在组织外部生成数据，Google Analytics、Splunk、ServiceNow 和 Salesforce 就是其中几个例子。这种趋势只会加快，第三方提供商将在云中生成越来越多的有用数据。一些组织还将本地基础结构也移到云端，再度强化这种云端优势。

每个人都需要数据

过去 20 年内，我们见证了市场向数字化经营和云端的转变。这是现代分析变革的一个重要方面。另一个是借助自助式商业智能向数据驱动型文化的转变。分析文化已经渗透到当今最具创新力的公司，其中提出问题的业务用户往往也能自己发现答案。因此，企业能够将数据这一"数字黄金"快速提炼成信息。为充分形成分析文化，组织必须整合其最重要的资产 - 员工和数据，使每位员工都能访问适当的数据，并鼓励他们进行探索和协作。

通过现代分析方法，IT 和业务部门可以实现合作。IT 提供集中式的环境，业务用户能在其中找到可信任的数据和内容，这样每个人都可以安全地使用数据，提出问题、进行试验并快速作出决策。这是一种自下而上的方法，需要领域专家创建元数据、业务规则和报告模式，以便提供流畅的敏捷性和快速的持续改进。

> **我们最开始使用 Tableau 时，考虑的只是仪表板和报告。我们从未想到 Tableau 会让整个组织脱胎换骨。它所带来的不仅仅是一种解决方案或技术，而是改变了在数据方面的文化。**
>
> - 全球商业智能总监 ASHISH BRAGANZA

数据瞬息万变

我们都知道，唯一不变的就是变化。现代分析平台优先考虑灵活性：即跨平台移动数据、按需调整基础结构、利用新数据类型和实现新用例的能力。此外，世界上几乎每天都在发布新的数据分析技术，例如机器学习、语音助手和自然语言查询等。至少就目前而言，其中某些技术似乎只是不切实际的空想，但是，新方法和新技术必将走向成熟，向您的公司和客户证明其价值。

在数据快速演变的大环境中，无论是对于您不断扩展的基础结构需求，还是对于新技术，灵活性都占据首要地位。灵活性对于形成和维持明显竞争优势至关重要。考虑将来仍适用的分析架构时，请勿局限于某家供应商的专利架构，因为这会在将来严重限制您的敏捷性。

2. 现代分析平台的构成要素

当今业务面临的这三项挑战并不像表面看起来那样难以解决。如果说数据是当今不断发展的商业环境中的共同线索，那么现代分析平台就是释放数据潜能的关键。但现代分析平台并不是一个单独的基础结构：它包含多个独立的构成要素。其中一些要素是商业智能的传统组成部分，只是进入了新时代（如数据仓库）。其他则是一些全新的概念，它们一诞生便革新了企业进行数据分析的方法（如可视化分析）。结合这两种构成要素构建的分析平台，能帮助任何组织应对当今的业务挑战。

现代分析平台可以归结为两个独立部分：

- **IT 支持**，包括收集、处理和准备数据
- **创作和使用**，包括分析数据并与利益相关者分享见解。

以前，这两部分合并成一个处理过程，仅由 IT 负责。我们现在将第一部分 - 创建和处理数据源 - 视为 IT 支持。第二部分 - 分析和交付 - 仍由 IT 提供实现能力，但由业务用户自己负责实现。

结合这两个部分，使业务部门和 IT 之间形成真正的合作关系，希望快速作出数据驱动型决策的组织应以这种现代方式运营。这种方式有时称为双模商业智能，既保留传统商业智能和运营报告的优势，又采用现代分析的自助式机制。

在 IT 与业务部门的这种关系中，IT 负责设计数据架构，并进行适当的数据安全保护和访问控制。业务领域专家根据需要创建自己所需的分析资产。这样，IT 让每个人都能高效解答自己遇到的重要问题，而业务用户能在出现问题时即时解答，就让组织更加敏捷，随时准备好解决现代业务环境中的挑战。

我们将详细介绍组成现代分析平台每一部分的各种构成要素、每种要素的主要趋势以及需要牢记的重要概念。若要深入了解特定组成部分（包括提供这些组成部分的市场领先的供应商）及其对您是否适用，请参阅附录中每个构成要素的对应部分。

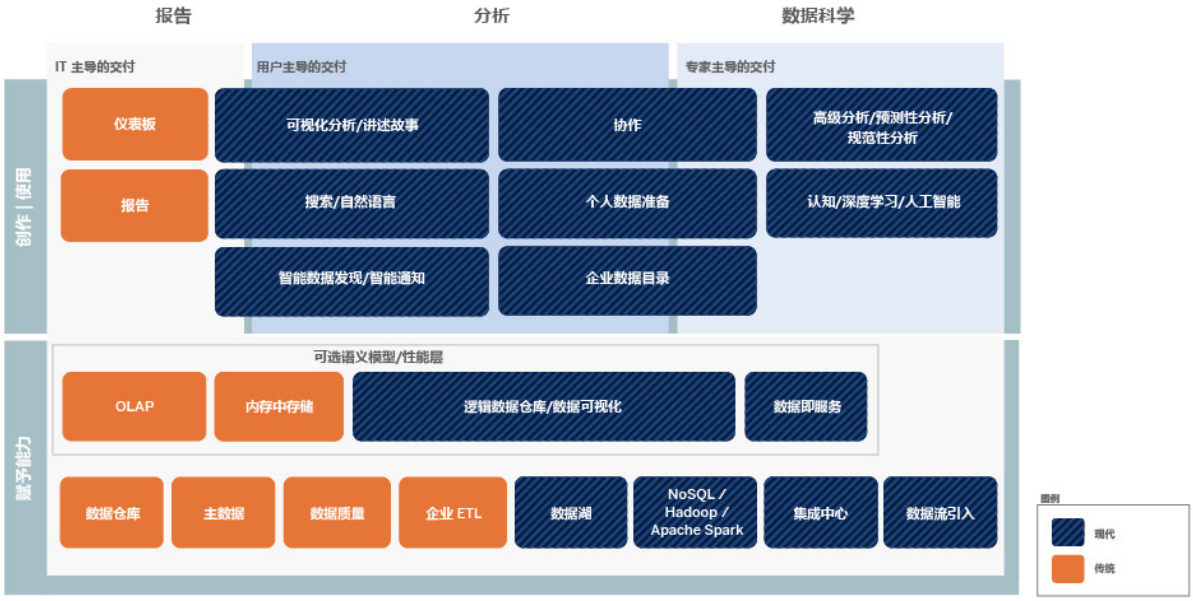


图 2 基本构成要素

IT 支持

与传统 IT 主导的商业智能不同，当代最具效率的 IT 组织专注于协调、组织和统一数据，提供分析型数据源，供用户和专家创作和使用。这一角色的职能很显著，但仍不应过分强调。收集数据、管理数据源以及处理数据以供他人使用，这一直在商业智能中占据重要位置，今后也仍然会是现代分析平台的核心。如果没有可提炼的原材料，又何谈发现见解呢？

现代分析平台的不同之处在于业务部门与 IT 之间的合作关系。当业务用户获得工具用于自助分析数据时，他们可以随时随地解答疑问，并且知道自己可以信赖数据。这样便能形成准确、灵活的报告和仪表板。IT 不再需要处理仪表板和更改请求，终于可以优先考虑数据本身：保障数据管控和安全，确保数据准确，建立收集、处理和存储数据的最有效途径。

现在就是优先考虑数据的最佳时机。您的企业（无论规模大小）一定已经在收集数据，但很可能只对其中的一小部分进行了分析 - 其余的则是暗数据。可以收集数据的位置不计其数，并且越来越多的工具进入市场，来帮助您尽可能多地收集数据。当前，您会发现有多种技术可用于处理各种特性的数据，例如数据量大、数据位置多和数据源类型各异。每个组织都是独一无二的，您应该着眼于现在和未来，基于适用性对组件进行优先级排序。

以下是几点注意事项。

有关特定技术（如流数据和数据即服务）以及特定供应商选择的详细信息，请阅读下面的附录。

数据库和数据仓库

几十年来，数据库和数据仓库一直是商业智能的基础。现代分析平台继续利用其中某些数据库和数据仓库，而另一些则逐渐变得无关紧要。

一些最早期的数据库采用 OLAP（在线分析处理）。此项技术旨在应对数据库技术速度过慢的问题，并且通过聚合和缓存加快可预测查询的响应时间。但随着公司问题变得越来越复杂并且越来越难以预测，OLAP 便无法跟上步伐，往往要求同时构建多个全新聚合。它与日益改进的数据库技术也越来越脱节。

当今的数据库采用先进的计算技术，例如内存中技术和大规模并行处理 (MPP) 技术。这使数据库能够提供极其快速的性能和线性扩展，同时优化数据存储、硬件内存使用情况，有时甚至能提供内置计算和数据科学功能。

此外，云技术的诞生为数据库技术注入了新的生命力，所提供的功能是本地存储完全无法比拟的。其中包括无需借助硬件即可启动的能力，随着业务需求更改灵活扩展的能力，以及无需组建团队即可管理基础结构的能力。

现代分析架构将始终包含数据库和数据仓库，它们将继续发挥重要作用，在整个企业中提供受管控且维度一致的准确数据，以实现自助式报告。即使是采用其他技术（如 Hadoop 和数据湖）的公司，往往也会保留关系数据库作为混合数据源的一部分。

NoSQL、非结构化数据和数据湖

当数据来自可预测源或采用可预测格式时，数据库和数据仓库可提供强大的分析支持。当然，并非所有数据都是可预测的。在现代分析架构中，NoSQL 数据库日益成为每个组织的必备工具，因为它能快速加载来自任何源的数据，包括不具有清晰结构或格式的数据源。NoSQL 数据库（有时称为"non-SQL"或"not-only-SQL"）可提供传统关系数据库无法提供的交替型数据存储，包括列、文档、键值和图形存储类型。

大数据和数据湖概念与非结构化数据相关。任何地方都能生成数据，有时是在难以预料的位置，因此收集所有数据并将其整理成可用格式可能非常棘手。经过技术发展，分析工具可以连接到原有原始数据，而无需强制数据首先适应某种格式。

其中一种技术称为数据湖，它是一种存储库，能够存储大量保留其原有格式、结构或其他特征的数据。因此，可使用经过优化的处理机制（如 API 或类 SQL 语言），随时随地转换数据以进行分析，而无需预先将所有数据处理成某种特定格式。

这些工具通常用于以下用例的相关项目中：物联网、数据科学、流数据，以及数据创建量和创建位置都不可预测的其他无结构分析用例。

如需与 NoSQL、Hadoop 和数据湖相关的数据技术列表，请参阅附录。

平面文件

Excel 和 CSV 文件的辉煌时代还远远没有结束。无论是小型还是大型组织，这些看似凭空出现的平面文件将会继续存在，并且可能永远存在。事实上，平面文件比以往的分布更广。以前，它们只存在于个人的物理计算机中。而现在，它们位于 Google Drive 或 Dropbox 等云端存储系统中。第三方组织可以在数据即服务中生成平面文件。平面文件创建速度快，因此可用作各种数据字段的图例、额外客户研究或用于增补现有数据集的小块信息。

此外，应在合适时机对平面文件采用适当安全措施。在必要时，尤其是一次性使用的情况下，鼓励使用平面文件。如果特定文件的使用量增加，则应用适当的安全协议，确保文件安全并且仅由正确的用户访问。

创作和使用

将业务用户引入商业智能平台是现代商业智能的独有标志。以前，决策者需要向 IT 索要报告并且要等待几天，最终只能收到过时的报告，而且并不能准确解答决策者的问题。这样的日子已经一去不复返了。现在，有疑问的决策者能够利用工具自己解答问

题。由于 IT 使整个组织都能信任可用数据，因此业务用户无需编程即可按需作出明智的数据驱动型决策。

实际分析工具是其中发挥关键作用的构成要素，并且可以立即开始使用。在确定与您的组织最相关的构成要素之前，就应该开始连接您位于各处的数据，以便在准备构建整个架构时验证并快速浏览您的数据源。

尽管存在各种用于分析数据的组件，但我们相信可视化分析始终是核心，利用这一方法，组织中的任何人，无论其是否具有编程经验，都可以直接连接到数据源并从中获得见解。如果将这类工具交到教师、医生和销售人员等业务用户手中，其组织将从变更请求的艰难工作中解放，像充分润滑的机器一样顺畅运行。

商业智能的新特性是更注重如何与他人共享见解。用户不再局限于仪表板和报告：他们可以生成完全交互式的应用程序，以及交织了数据、文本和图片甚至针对移动设备优化的查看体验的长篇文章。

此外，随着企业及其各部门的壮大，他们会借助智能生产力工具，快速共享信息、发现数据源、与仪表板保持同步以及跟踪最重要的指标。

在此部分中，我们介绍了最先进的现代分析工具中可用于增强数据交付的组成部分。若要了解更多信息，如讲述故事和通知，请参阅附录。

可视化分析

人类的视觉系统是世界上最强大的工具之一。如今，它终于成为不可或缺的数据分析方法。可视化分析以大脑每天使用的模式识别为基础，能够通过类似方式揭示数据内的模式，例如上行和下行的趋势、活动中的不规则尖峰或异常的特定记录。

传统电子表格要求按行和列分析数据，选择要共享的子集，然后创建图表。通过麻烦的向导或基于文本的命令，这些图表有时能解答问题，有时会引出全新的问题，但始终不能用于进一步分析。相反，可视化分析提供精美的直观分析体验，通过简单的拖放操作将可视化加入分析过程，而不仅仅提供图表作为最终结果 - 获得见解的过程与答案同样有价值。

可视化分析不只是出色的可视化工具。它还是一种语言，可用于计算、分组和假设条件，因而不需要编程辅助即可合并数据、发现异常以及扩充数据。

传统 BI 和报告

传统 BI、仪表板和报告今天仍然占有一席之地，只是创建方式有所不同。许多静态报告（如管理层仪表板或财务审计）需要具备技术开发技能才能创建，不仅要提早提出分析查询，往往还需要更改基础数据模型。所有这些流程可能需要几天、几星期甚至几个月才能完成。

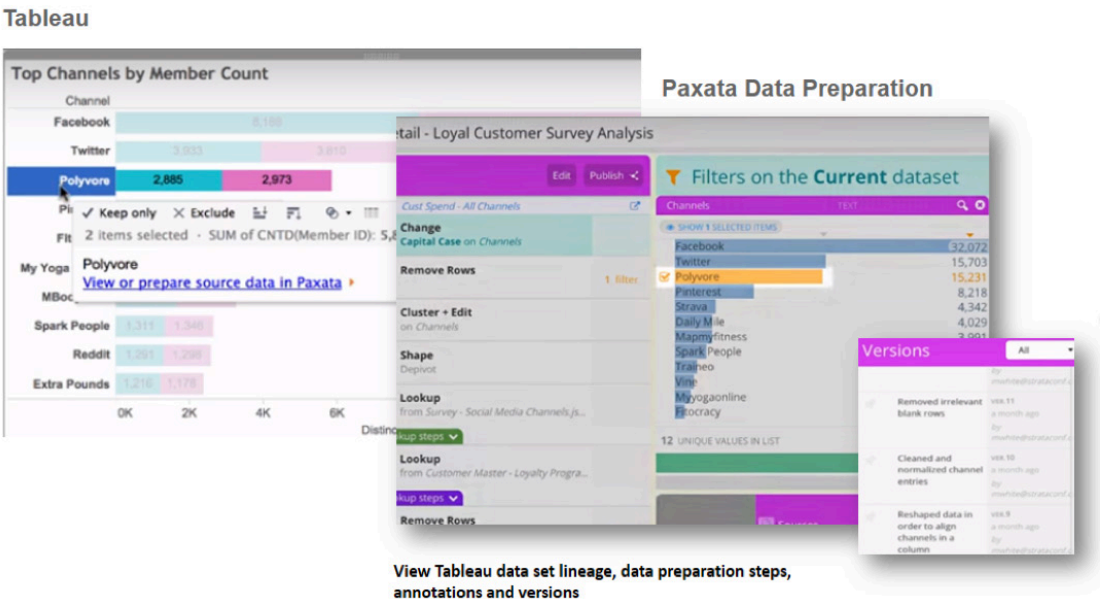
在现代分析平台中，由于所解答问题的本质，这类仪表板和报告中很多从临时问题开始，由 IT 和数据管理者进行强化和验证，并最终替代传统静态报告。通过这种经过更新的流程，在问题发生演变、更改以及引出全新问题的过程中，业务用户能在挖掘数据时利用自身领域专长，最终找到正确答案。即使对传统报告的需求仍然存在，现代分析仍将凭借其灵活性逐渐取代传统工具。

个人数据准备工具

与 ETL 不同，数据准备工具是轻量级应用程序，旨在帮助非 IT 用户有效地精确操纵数据。其构建原则是与可视化分析工具同样简单易用、高速和敏捷，让业务用户每天都能合并数据集、自动联接、重命名字段以及为准备分析而对数据进行其他改进。

这些改进应稍后进行，因为在开始使用数据之前，通常难以确定需要如何调整。但在组织的生命周期中，要解答疑问，大部分时间

根据《哈佛商业评论》上的**“您的数据策略是什么？”**一文，“分析人员 80% 的时间都花在发现和准备数据上”。



都用于将数据处理为适当形式。个人数据准备工具能够大大降低此类时间投入，但却并不总是需要 IT 构建的预定义报告语义层。

图 3 Paxata 的可视化分析与个人数据准备集成

高级分析

随着公司转向使用统计数据、预测算法和机器学习来最大限度提升超大数据集的价值，高级分析已成为现代分析架构的重要组成部分。

过去，只有经过培训的数据科学家才能使用 R、Python、SPSS 或 SAS 等程序实现高级分析。而现在，经过改进，可视化分析扩展了高级分析的适用范围，使所有自助式分析用户都能将其作为内置功能使用。如今，盒须图、树地图和基本预测建模技巧等功能已非常常见，单击鼠标即可使用。

现在仍然存在许多专用统计分析工具的用例。一些组织正在积极创作能够持续演变的算法，以便为客户提供"下一关注点"，或者编写函数来确定客户信用卡受到欺诈的时间，他们希望研究出特定的专用工具，以完善其核心可视化分析工具体系。部署此类工具需要进行专门培训，并且要求具备特定于工具的编程经验，这些可能需要几个月才能实现。

共享与协作

在现代 BI 平台中，共享、协作处理和交流见解是一项重要功能。从评估环境到排列出下一项最佳做法，见解的影响在协作过程中逐渐达到最大化。现代分析平台借鉴领先的生产力和门户应用，具备论坛、注释、评论、收藏、点赞和其他社交概念。支持直接在工具内部交流从分析平台获得的见解，使分析流变得更加简单，并且能鼓励用户深入探索和讨论其他有价值的发现。通过嵌入，还可以将协作扩展到外部应用程序和门户。

嵌入式分析

"流"是一个最强大但却经常被忽略的业务概念。与其使业务用户脱离其标准处理流程以从数据中获得答案，不如将见解无缝插入已创建的流中并将其合并为现成流程。

利用现代分析，您将发现数据和仪表板已直接嵌入公司门户和其他应用程序，或已与生产力工具集成。最好的分析平台会利用成熟的 API、软件开发人员套件和灵活的交付机制支持所有这些方案，使您能轻松在工具之间切换，甚至能将所有工具合并到一个门户中。

流还可以延伸到实际位置。如今的工作人员经常不在办公室内，无法越过防火墙访问内部资源。现代分析支持通过任何设备从任何位置访问数据。这意味着，销售人员可通过移动设备作出明智的数据驱动型决策，而无需使用笔记本电脑。工程总监也可以在现场访问重要信息，而无需通过 VPN 进入公司网络。移动和云技术已经彻底改变了企业的运营方式，真正的现代分析平台也必须帮助企业利用这些优势。

3. 综合利用

只有具备这些基本构成要素，现代分析平台才能帮助企业应对任何未来挑战。这一平台结合业务部门和 IT 之间真正的合作关系，让公司内的每个人都信心十足，因为他们知道，无论要作出任何决策，都有可帮助决策的工具，并且自己的决策以可信赖的数据为基础。

组建完整的现代分析平台可能本身就是一项挑战。好消息是，您并不需要提前构建整个生态系统。事实上，在当今时代，企业不这样做时才能取得最大的成功。他们反而应从小处着手，并逐步转变，最终确定接下来的业务投入方向。将技术解决方案推广到更大的部门前，公司可以进行测试。开始时，您不必集成整个策略中的每一个组成部分。例如，完成整个数据仓库前，您可以使用可视化分析工具发现数据管道中的漏洞。这能帮助您通过分析创造直接价值，找到数据中的漏洞和错误，最终构建出更准确、功能更完善的数据仓库。

关键在于使用可实现这种逐步更改的工具，而现代分析平台恰好由一系列可逐一整合的构成要素组成，因此它不仅能增强组织的可访问性和敏捷性，还使组织能够从各种类型的数据源中获得见解。这正是世界一流的分析领导企业当前的工作方向，他们战略性地利用可视化分析，并结合其他最佳大数据分析、物联网和数据科学解决方案。

例如，Netflix 构建了综合性大数据平台和数据湖，来支持他们在运营过程中生成的大量数据。Tableau 是其中的关键组成部分，帮助该组织将 S3、EMR 和 Spark 独立工具合并到整合的分析平台中，对其业务提供支持。

无论您处在数字业务转变过程的哪一阶段，都应立即开始使用您当前拥有的数据。为应对下一大型市场变动，企业必须加快行动。使用现代分析工具，让企业能够作出数据驱动型决策，最终成为变革者。

4. 附录

现在有许多不同的分析技术和解决方案选择，每一种都具有其独特用途和优势。在本附录中，我们将介绍概述部分未提及的技术和解决方案，并逐一介绍其中的可选工具。

流数据摄取

社交网络、智能电表、家用自动化设备、电子游戏和 IoT 传感器等联网设备和应用不断生成流数据。通常，这种数据通过半结构化数据管道进行收集。虽然可对流数据应用实时分析和预测算法，但通常使用 Lambda 架构将流数据按原始格式路由并保存在数据湖（如 Hadoop）中，用于分析。

Lambda 架构是一种数据处理架构，旨在利用批处理和流处理方法处理大量数据。其设计解决了在延迟、吞吐量和容错方面的挑战。

现在有多种可用于流式数据的工具，包括 Amazon Kinesis、Storm、Flume、Kafka 和 Informatica Vibe Data Stream。

集成中心协调

中心辐射型集成模式是一种易于理解且广泛使用的数据架构设计。中心能够分离位于任何位置的数据源，并通过减少要管理的点到点接口数来提升集成灵活性。集成中心的发布/订阅功能可促进数据重复使用，并提供集中控制，以实现数据优化、标准化和管控。有序组织的数据移动管道中涵盖任意位置的数据源，而集中式管理能提高此类管道的透明度。

新一代数据集成中心使自助式分析用户也能使用传统功能。任何人都可以向现代集成中心数据源进行发布和订阅，基本无需 IT 参与。数据使用者可以利用经认证的数据，深入了解线性集成过程。现代数据集成中心还具有其他优势，包括无缝数据质量功能、快速数据源引入和及时交付小规模或大规模数据集。

Informatica 和 Cisco 在数据集成中心技术方面占据市场领先地位。利用 Tableau 与 Informatica 的深度集成，您可以将数百种不同的数据源合并到 Tableau 数据提取，在 Tableau 数据服务器上保存并随时更新，供组织中的任何用户使用。

详述非结构化数据、NoSQL 和数据湖

数据湖通过更快、更灵活的数据集成和存储，满足现代大数据分析的要求，可供任何人以各种方式快速分析原始数据。数据湖不会取代数据仓库。

在现代"摄取并加载"的设计模式中，任何规模或形态的原始数据最终通常都归于数据湖。数据湖是一种存储库，能够以原有格式（结构化、半结构化和无结构）存储大量数据。数据湖通过 API 或类 SQL 语言提供经过优化的处理机制，可通过"读取模式"功能转换原始数据。

Hadoop 恢复能力强且成本低，从第一代 Hadoop 分布式文件系统 (HDFS) 起就用于数据湖；尽管如此，它并不是唯一的数据湖实现选项。Amazon Web Services Simple Storage Service (S3) 和 NoSQL 数据库等对象存储具有灵活架构，也可用作数据湖。Tableau 现在支持 Amazon Athena 数据服务连接到 Amazon S3，并提供可直接连接到 NoSQL 数据库的多种工具。

在现代分析架构中，NoSQL 凭借可从任意位置快速加载数据的优势和无架构数据库概念，正逐渐成为行业标准。NoSQL、non-SQL 或 not-only-SQL 数据库可提供另外的数据存储类型。常见 NoSQL 存储类型包括列、文档、键值和图形。

常用于 Tableau 的 NoSQL 数据库的示例包括但不限于 MongoDB、Datastax 和 MarkLogic。

尽管 Hadoop 常用作大数据平台，但它并不是数据库。Hadoop 是一个开源软件框架，用于在商用硬件的群集上存储数据和运行应用程序。它能大量存储任意类型的数据，具备执行大型处理和处理大量并行任务或作业的能力。

在现代分析架构中，Hadoop 提供低成本存储和数据存档，可将陈旧的历史数据从数据仓库移入在线冷存储。它还可用于 IoT、数据科学和其他无结构的分析用例。

在 Hadoop 框架中，用于加载、组织和查询数据的相关技术包括但不限于以下各项：

- **Apache Spark** – 开源群集计算框架，具有高性能的内存中分析和数量不断增加的相关项目
- **Apache Impala** – 用于 Apache Hadoop 的开源分析型 MPP 数据库。 这是 Tableau 的成功 Hadoop 相关项目中最常用的数据连接
- **Apache Presto** – 开源分布式 SQL 查询引擎，用于跨所有规模的数据集运行交互式查询。Tableau 在版本 10 中增加了 Presto 支持
- **MapReduce** – 并行处理软件框架，可接受输入并将其划分为较小问题，然后分发到各个工作节点
- **Hive** – 既是数据仓库，又是类 SQL 查询语言。 Hive 2.0 还包含长久守护进程（Live Long and Process，LLAP），可显著提高 Hive 查询性能。
- **Hadoop 分布式文件系统 (HDFS)** – 无需预先整理即可跨多台计算机存储数据的可扩展系统
- **YARN** –（Yet Another Resource Negotiator，又一资源协调者）为在 Hadoop 上运行的进程提供资源管理
- **Ambari** – Web 接口，用于管理 Hadoop 服务和组件
- **Cassandra** – 分布式数据库系统
- **Flume** – 用于将数据流式处理成 HDFS 的软件
- **Hbase** – 在 Hadoop 上运行的非关系数据库
- **HCatalog** – 表和存储管理层
- **Oozie** – Hadoop 作业计划程序
- **Pig** – 用于操纵 HDFS 中所存储数据的平台
- **Solr** – 可扩展搜索工具

• **Sqoop** – 在 Hadoop 和关系数据库之间移动数据

• **Zookeeper** – 用于协调分布式进程的应用程序

值得注意的是，在过去的两年中，Apache Spark 已经从 Hadoop 生态系统的组成部分变身，转而成为诸多企业理想之选的独立大数据分析平台。与 Hadoop 相比，Spark 的数据处理速度大幅提升。Spark 本身拥有许多相关项目，包括核心 Apache Spark 运行时 Spark SQL、Spark Streaming、MLlib、ML 和 GraphX。它是现在最大的大数据开源项目，拥有来自 250 多家组织的 1,000 多名数据贡献者。

Tableau 在特定于大数据的分析连接和可视化数据分析方面处于市场领导地位。行业一流的大数据分析程序都在通过 Cloudera、Spark SQL、Amazon EMR、Hortonworks、Microsoft HDInsight/数据湖和 MapR 使用 Tableau。通过这些公开支持的技术或相关驱动程序，可将其他许多大数据技术连接到 Tableau。

数据即服务

在这个数字化时代，数据就是金钱，也是任何人都可以使用的产品。客户数据、财务数据、营销数据、天气数据、地理数据和人口统计数据，都已在数据市场和交易平台中作为服务供买家购买。

数据即服务采用灵活的服务导向架构 (SOA) 模式，以便通过云端传递数据。此方法能提供超高敏捷性，因为 SOA 架构非常简单。目前可见 ISV、CRM 和 ERP 提供标准的数据即服务 REST API，用于实现集成或外部报告方案。

通过 Tableau 的 Web 数据连接器 SDK，可以连接到现有连接器外部的数据。通过 HTTP（包括内部 Web 服务、JSON 数据和 REST API），自助式分析用户几乎可以连接到任何可访问的数据。

逻辑数据仓库

分析行业领先组织通过 Cisco 和 Denodo 等供应商提供的数据可视化技术，为位于各处的数据提供灵活、逻辑清晰、维度统一的视图。对于分析用户，逻辑数据仓库与关系数据仓库的表征和行为都很相似。Tableau 用户可以使用现成的 ODBC 驱动程序连接到这些工具。

数据可视化的一项重要功能是，优化对各种数据源和 REST API 的远程分布式异构查询。逻辑数据仓库还能充当语义层，可在数据源更改时为报告应用程序提供缓冲。逻辑数据仓库通常用于企业数据目录。

主数据管理

分析结果的准确性取决于所用数据的质量，只有当数据是准确的，使用者才能作出正确决策。越来越多领域专家自下而上地创建数据，可见业界再次开始注重传统数据质量和主数据管理，因为它们能够确保报告数据源的时效性、整洁性、一致性和准确性。

备受喜爱的主数据管理产品/服务包括（但不限于）Informatica、IBM 和 Stibo。Tableau 中常用的几种数据质量解决方案有 Trillium、Informatica Data Quality、Talend Data Quality 6.0 和 Tamr Eisenhower。

企业数据目录

企业数据目录是另一项新兴技术。利用企业数据目录，自主报告用户能从经认可的数据源中轻松找到适当数据，并据此作出决策。企业数据目录存在于可视化分析解决方案中，也可用作与 Tableau 无缝集成的独立服务。

通过扫描所摄取的数据源，企业数据目录将由表、视图和存储过程中的元数据填充。数据目录能自动发现新数据源、进行智能数据分类和跨数据源条目映射，主要充当数据源和常用数据定义的企业业务术语表。领域专家通过添加注释、版本和文档，可进一步强化目录数据源的上下文关联性。

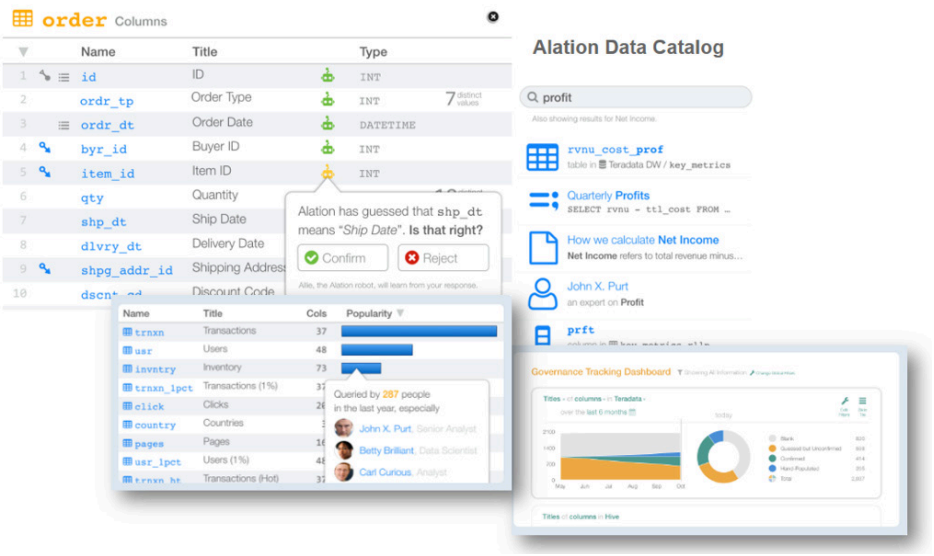


图 4 Alation 的企业数据目录

数据目录解决方案能促进数据整合和现有数据的有效重用。它们还能提供优越的数据沿袭，以及更高水平的数据管控、保护、记录和审核。

提供与 Tableau 良好集成的丰富数据目录的供应商包括 Alation、Collibra、Attivio、Informatica 和 Waterline。

机器学习

认知、深度学习和人工智能进一步利用高级分析，从现有数据和模式进行推理，基于现有知识库得出结论，然后将此结论放回知识库，形成永久连续的自主学习循环。

使用此类型的分析，通常需要在报告或集成应用程序中嵌入 API 以查看输出。Tableau 现可用于对 CognitiveCode、Digital Reasoning 和其他供应商的输出进行可视化。

在 Tableau 10.3 中，我们首次推出了 建议表和智能联接，可缩短连接和准备数据的时间。这些建议由机器学习提供支持，会随数据库使用频率的增加而改进。

自然语言

自然语言和语音查询进一步利用数据故事讲述，以更灵活的新方法为每个人提供数据发现能力。通过为重要发现自动添加上下文说明、请求预测或分析大量文本文档，自然语言增强了分析在任何平台上的可用性。

现在，Tableau 可视化分析可与先进的自然语言生成 (NLG) 解决方案集成，包括 Yseop、Narrative Science 和 Automated Insight 等。由于此类技术已基本解释了 Tableau 可视化的上下文，因此集成往往在自然语言工具内进行，或通过 JavaScript 作为扩展实现。此外，收购 ClearGraph 后，我们将直接在 Tableau 中提供更智能的数据发现和分析功能，进而简化通过自然语言与数据进行交互的过程。

建议的数据发现

智能数据发现采用机器学习算法，利用针对特定目的格式化的数据，提供更深层的分析（“未来趋势”）和规范（“优化方法”）功能。随着可视化分析的发展，新的自动化见解和建议也将会增加。这些功能已在 2016 年 Tableau 全球用户大会的未来蓝图主题演讲中进行了展示。高级分析功能集成了 R、Python、API 和分析数据库功能，您可在 Tableau 中进行可视化，探索其效果。

搜索

利用现代分析架构，无论数据位于何处，用户和专家都可以像使用 Google 一样轻松地搜索和查找。分析搜索引擎索引技术不对数据进行建模，而是根据字段名称、数据类型和机器学习智能自动关联不同的数据源。一直以来，动态搜索建议都是基于历史查询和报告的使用情况生成的。但最近，随着语音技术（如 Siri 和 Alexa）的出现，开始出现与分析搜索相结合的语音查询功能。2016 年 Tableau 全球用户大会上，来自 Automated Insight 的年度 Hackathon 的获胜者将 Alexa 语音控制与 Tableau 集成，这就是一个非常棒的例子。

通知

现代分析架构中包含可配置的智能数据驱动型通知提醒，可持续监视海量数据中的有价值信号。没人能够全天候监视每一项重要价值。正因如此，自动化和通知已成为现代分析必备组件中的出色资产。

某些工具能以固定间隔提供快照；其他工具则能真正跟踪日志以了解数字是否已超过特定阈值。这两种工具都有其存在的原因。一些仪表板可提供有用信息，您需要每天查看。另一些仪表板则是重要行动的基础，但每天查看仪表板却不能获得任何可行性见解，这无异于浪费时间。

在 Tableau 中，您可以借助 Tableau Server 数据驱动型通知功能，随时把握业务动向。只需为自己或整个团队选择接收电子邮件通知的阈值即可。

讲述故事

有时候，见解或"是什么"并不够。数据中隐藏的"为什么"也很重要。销售额为什么增加了？网站流量为什么激增？我们为什么要辛辛苦苦地储备医疗用品？

各企业长期努力将分析与其他沟通形式（文本、图片或视频）相结合，力求解决此类问题。分析师使用 Powerpoint 制作幻灯片，编写 PDF 格式的长篇报告，更麻烦的是，甚至需要打印出一页又一页文档并把它们装订在一起。

当今的现代分析工具利用这些讲述故事概念的最突出优势，并将其实际集成为一流的功能。利用这些功能，您可以构建交互式仪表板，发送可在添加新数据时在后台自动更新的特定数据快照，甚至创建结合了交互式图表、文本和图像的报告。讲述故事能够说明对数据所进行的分析，而不仅仅是提供数字。

Tableau 极其重视提供选择和开放标准的重要性。我们在研究和开发方面进行了大量投入，目的是使分析更快速、更简单，当然，这也需要与我们的合作伙伴生态系统共同创新。这能确保在分析环境发展和新技术投入市场的过程中，分析领导企业始终能将 Tableau 与其当前和未来的数据技术相集成。

关于 Tableau

Tableau 致力于帮助人们查看并理解数据，不论数据规模有多大、来自什么渠道或存储在何种数据库中，用户都能轻松驾驭。从 PC 到 iPad，您可以使用各种设备对自己的数据实现快速连接、混合和可视化，感受行云流水般的使用体验。用户无需掌握编程技能，就能创建和发布可自动更新数据的营销仪表板，与同事、团队、高层领导、合作伙伴或客户分享实时的分析洞见。立即免费试用！

其他资源

How to Build a Culture of Analytics（如何建立分析文化）

定义分析

Approach to Analytics（分析的方法）（解决方案页面）

云数据简报

使用 Tableau 进行高级分析