驯服数据巨龙

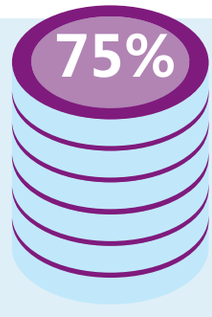
如何采集任意来源的数据、随时随地监管数据并利用数据造福每个人

数据已经改变

在企业数字化转型的推动下，数据量呈现指数级增长。据 IDC 研究发现，地球上每个人每秒平均产生 1.7 MB 数据。扩展连接、将服务配置到数字和移动平台、依托客户关系启用用户生成的内容 - 所有这一切都促使数据点不断增加，企业亟需控制这一增长局面，以便在新环境中处理、管理、分析数据及做出决策。

数据量不断增长，本身不足为奇 - 这是数十年来数据管理面临的永恒主题。当前的大数据时代之所以挑战不断升级，其根本原因在于数据源种类持续增加。内部生成的数据不再仅限于传统的产品、客户、销售和交易结构化数据，还包括面向客户的员工创建的笔记、客户发送的文本和消息，以及在内部和外部品牌平台中通过各数字接触点的持续点击流创建的图像和视频。

新型商业智能和预测性分析实践（如数据科学团队的出现）进一步扩大了需要发现和分析的数据集范围。IBM 意识到外部数据存在巨大价值，于是在近期收购了 The Weather Company（该公司运营美国第四大最常用移动应用），并支持通过 IBM Cloud 访问该公司的数据。IBM 还与地理数据提供商 Mapbox 及消费者数据代理 Acxiom 达成合作意向，意味着这些精心策划的第三方数据集现已能够与 IBM 解决方案的第一方客户数据轻松整合。



预计到 2019 年，75% 的分析解决方案将包含十个乃至更多第二方合作伙伴或第三方提供商外部数据源（来源：Gartner 报告）

企业现在产生的数据比以往任何时候都多，使用数据的速度也显著提升，目的是为了做出有效决策以及提供实时客户支持。从数据管理流程和技术的角度而言，这意味着亟需做出重大调整，确保满足这些新需求。

IBM

新型数据基础架构

数据仓库是一种结构完善、稳定成熟的环境。在数据仓库中，企业可以存储各种关键业务数据，包括交易和财务信息、客户记录、互动信息和人口统计数据。无论面向整个企业还是在各个运营部门单独创建，数据仓库都可以提供核心业务流程，支持报告和分析，而且允许企业通过针对各个部门创建的不同视图来了解其当前状态。

然而，此类数据基础架构并未针对实时、高容量非结构化数据流进行优化。配置新的数据源（特别是第三方数据源）可能需要很长一段准备时间。同样，支持数据科学（通常包括采用探索性方式发现和分析大量数据）可能对企业数据仓库产生极大的颠覆性效应。企业数据仓库已经过优化，支持定义明确的可重复性连续业务流程。

因此，一种新型数据基础架构方法应运而生，该方法采用分布式存储，以及基于云的或商用硬件计算解决方案，而且通常运行开源应用。此类解决方案的一大魅力（特别是对于数据研究员而言）在于，能够针对时间有限或用途受限的项目创建解决方案，并在达成目的后关闭解决方案。

但这些解决方案也存在缺点，那就是面临“影子 IT”运营风险 - 技术项目不受企业框架约束 ,包括项目控制、治理和监督。分析结构化和非结构化组合数据的能力也很有限，难以发掘数据驱动洞察或运行实验环境下创建的各种模型。

企业存储所创建数据的 80%，但仅有 0.5% 的数据得到分析（来源：MIT）

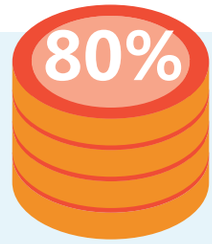
为解决这个问题，必须开创一种全新的数据架构，融合两种环境的优势功能，保留现有传统系统，同时扩展可供访问的数据范围。人们将其称之为数据湖，数据湖可接受各种来源的数据流，而后将数据导入通用平台以供使用。数据以未经优化的原始状态存储，根据需要查找、处理、优化和提取。

IBM 更进一步，根据数据湖中的数据来源对数据进行编目和分类，确保全面监管存储的数据以及在使用中的数据。这样可以为 IT 人员和数据研究员带来很多重要优势，因为这不仅可以灵活满足项目需求，还能降低成本、保持关键业务数据的配置灵活性，以及防止无管制数据环境和使用方法突然出现。结果是生成值得信赖的数据集（开放、受监控且经过维护），避免陷入数据沼泽（不受控制且不可预测）。

业务用例

每一家企业创建数据湖都有自己的独特原因，综合反映自身的市场地位、技术采用、成熟度和资源状态。但不同的部署项目存在一些常见的驱动因素，如下：

移动应用 – 鉴于传统数据基础架构专为进行批处理（而非实时决策和验证）而设计，通过移动设备为用户提供服务及交易支持对传统数据基础架构是一项极大的挑战。在数据湖中，您可以整合上下文信息（如设备 ID 和位置）与结构化数据（帐号、密码），通过安全、灵活、可靠而又可复制的方式推动提供移动应用服务。



数据研究员通常花费 80% 的时间和精力准备数据（来源：Forrester Research）

预测性分析 – 基于倾向模型做出决策，比如，下一步最佳行动或产品建议，需要融合客户特定数据（采购历史记录、偏好）、更广泛的上下文（采购现状、产品供应水平）乃至外部参考数据（天气、假期日历）精准提供个性化的相关通信。

欺诈检测 – 越来越多的网络犯罪分子使用有效的客户凭据访问和盗用帐户、进入业务系统或提取高价值机密信息。新一代欺诈检测不仅可以检查这些凭证是否存在报告的盗窃或丢失现象，还能检查可能表现出不一致或不寻常维度的社交和数字数据，如通过未知 IP 地址访问或异地访问。通过尽可能广泛的数据集（最好来自受监管的环境）实时实现这一目标，可以提高发现欺诈的概率并减少误报。

数据增强 – 除了企业生成的各类客户互动和交易视图以外，还可以引入外部数据源了解客户对整个产品组合的看法，从而加强优势。使用这些数据源配置数据湖，将支持对数据和数据源本身进行动态更改。IBM 收购了 The Weather Channel 并与大量第三方数据所有者（如 Mapbox 和 Acxiom）建立合作关系，将经过验证和精心策划的信息预先整合至分析和决策环境（如数据湖）。

数据科学 – 尽可能探索最大的数据集，融合内部数据源与外部数据源，从全新的视角深入理解客户行为、营销机会、产品功能和服务主张，发现并总结重要模式。如果数据经过预先整理、编目和监管，不仅可以缩短探索数据前的发现及整理时间（可能占数据研究员工作时间的 80%），还能缩短价值实现时间。

利用数据盈利 – 扩展数据可用性是一个全新的重要维度，对于营销工作尤为如此，目的在于加强相关性以增加客户互动，同时发掘转售增强数据、派生变量或针对性媒体商机的潜在机会，从而通过上述两种方式实现附加价值。

克服障碍

数据湖很可能掀起重大变革，进而支持新流程并推进价值驱动活动。与此同时，由于数据湖相关技术尚不成熟，很可能对现有实践带来一定的冲击。某些职能部门认为自己是特定类型数据的所有者，很可能不愿合并数据建立更大、更全面的资源体系。为强制瓦解这些数据孤岛，赢得高层管理者的支持显得至关重要。

美国企业每年因数据质量问题导致的损失高达 6000 亿美元（来源：Data Warehousing Institute）

亟需对数据仓库应用治理策略，务必确保数据湖成为值得信赖的数据源，而不是结构混乱的“着陆区”，即随意存储数据却未考量数据有效性、价值或保存期限（衰减或更改速度，更未考虑预期使用期限）。由业务、法务与合规、IT 及分析部门代表组成的跨职能团队可以制定可持续发展策略和数据定义，从而避免未来可能出现的种种问题。

若能达成一致定义并通过元数据管理层进行部署，将能够长期保持数据湖中数据的一致性和可用性。首先，该流程可以突出整个企业的显著差异，例如，如何定义客户或销售、使用哪些单元记录库存水平、地址或电话号码格式等。IBM 通过信息治理解决方案提供流程支持，该解决方案要求数据湖中的所有源遵循一致的规则和控制策略。这种方法的一项重要特征在于，它不仅可以确保核心企业数据仓库和扩展型数据湖采用相同的标准和控制策略，还能监管本地系统和云系统。

另外，还必须应用访问控制和监控策略，确保全面了解新增强的数据集的使用情况。某些从业者可能会企图过度使用许可访问权限，比如，尽管隐藏数据和化名数据足以满足需求，但数据研究员仍然希望使用完整的个人数据集。通过这种方式确保有效监督，同样有助于企业履行合规义务。

畅享数据优势

驯服数据巨龙将为整个企业带来巨大的优势，包括提高生产力以及改善销售和营销成果。鉴于数据湖项目的潜在规模和复杂性，重要的是要确保尽可能面向更多领域应用指标，从而全面了解投资回报。值得注意的优势包括：

降低数据仓库成本 – 在数据湖中，企业可以将目前存储在数据仓库中的大量非核心数据元素卸载到低成本分布式存储中。

数据可访问性提高 10% 将能够为一家典型的《财富》1000 强企业额外创造超过 6500 万美元的净收入（来源：Baseline）

降低系统整合成本 – 恶意 IT 项目往往会掩盖外部顾问的真实成本，因为这部分费用由部门预算支付，而不是由中央 IT 预算支付。

降低外部分析咨询需求 – 面向数据科学团队提供数据和技术资源可以消除对外部专家的依赖，从而深化洞察及推进预测性分析。

提高分析团队工作效率 – 典型数据湖体验旨在加速模型和洞察输出，从而推动提高营销和销售业绩。

降低数据质量成本 – 从重复记录，到无法送达的地址，数据质量差将会导致大量隐藏成本，数据湖应避免此类问题。

后续步骤

IBM 在数据基础架构管理、高级分析、数据转换、质量管理、治理和数据保护领域拥有深厚的专业经验。您可以通过多种参与方法实现这些目标，比如，技术验证、概念验证、数据实验室和现场项目交付。

要了解有关成功数据湖所带来的优势和商机的更多信息，请访问：**ibm.biz/data_lake**

IBM United Kingdom
PO Box 41
North Harbour
Portsmouth
Hampshire
PO6 3AU

英国出品
2016 年 9 月