数据存储融合

了解 IBM AutoSQL 如何跨不同数据源一站式提供通用数据访问和查询

要点

- 跨各种类型和来源的单一数据视图
- 对数据湖运行仓库级查询
- 执行分布式和虚拟化查询的速度加快 53%
- 使用广泛的数据管理解决方案，如 Db2、Netezza、Event Store、Spark 和 Hadoop

不断增长的数据量和多样性促使企业将这些数据存储在不同供应商和不同地点的无数个不同的存储库中。虽然在最适合的地方存储数据是一个不错的原则，但这类环境的碎片化性质正在造成一定的损失。74% 的潜在有价值数据未被使用。[1] 大数据的问题更加严重，其中 88% 未得到有效利用。[2]

通过提取、转换和加载 (ETL) 过程或制作数据副本以物理方式合并这些数据，已被用于尝试弥合这些孤岛之间的差距。然而，结果往往是让事情变得既复杂又耗时且成本高昂。当需要调整查询以适应不同类型的数据或其存储库时尤为如此。仅仅出于分析目的就将数据湖中的非结构化数据移动到数据仓库中，可能会消耗大量时间和预算。

IBM 通过 IBM® AutoSQL 提供了一种替代方案，这是一种通用分布式查询引擎，无需物理移动即可聚合和查询所有数据存储。该技术使用户不仅可以通过单一视图查看所有存储库中的数据，而且无需移动、复制或手动调整即可对它们进行查询。这种更简单的方法所具有的价值怎么强调都不为过。据估计，一家典型的财富 1000 强公司可因数据可访问性提高 10% 而额外获得 6500 万美元的净收入。本文深入探讨了允许融合数据查询的技术，以及可让成功锦上添花的数据和 AI 平台以及数据存储。

驱动数据存储融合的关键能力

AutoSQL 支持畅通无阻的完全数据访问

AutoSQL 作为一项关键功能，它通过利用数据虚拟化、云对象存储和自动化治理来促进跨云、数据湖、数据仓库和数据库的数据融合，从而简化了对数据的访问。AutoSQL 将这组技术从提供便利性的手段转变为不可或缺的性能助推器。它不只是通过数据虚拟化来抽象数据，还可以在混合环境中对企业数据（包括数据湖和流数据）执行仓库级查询。如果没有 AutoSQL 技术，就需要完成大量工作来准备要查询的数据，或者相反，准备查询以就地访问数据。

借助 AutoSQL 通用查询引擎，可以在所有数据存储库中运行使用 SQL 的仓库级查询，而不论数据是结构化、非结构化还是介于两者之间。
换句话说，无需手动移动数据或调整查询，因为查询将会根据数据及其位置自动调整。而且，它会以极快的速度来完成 - 与行业标准相比，执行分布式和虚拟化查询的速度加快了 53%。[3] 因此，查询 Hadoop 中的数据或流解决方案（如 Apache Parquet）中所用格式的数据，就像查询仓库中的数据一样轻松有效。据估计，到 2025 年，全球 80% 的数据将是非结构化数据，这些功能不仅是锦上添花，还是必不可少的。

当您考虑数据湖和数据仓库之间的交互时，AutoSQL 的价值就会变得尤为突出。传统上，数据湖中的非结构化数据需要移动到数据仓库中，才能利用其卓越的查询能力。如果不再需要，可以将数据保存在便宜得多的对象存储上，并如同它们就在数据仓库中一样进行查询。由于仓库存储成本大约是数据湖的十倍，因此企业既节省资金又节省时间。

AutoSQL 还可以根据公司的需求独立扩展计算和存储，这要归功于它能够利用云对象存储的能力，并且它还附有自动化的嵌入式治理功能。为了给予更多帮助，数据虚拟化现在还提供自主缓存创建功能。此功能使用推荐引擎来识别长时间运行和经常执行的查询，并提供一个潜在查询的排序列表以添加到缓存中。总的来说，通过利用 AutoSQL，这些功能有助于加强整个企业范围内目前可行的企业级查询。

通过 IBM Cloud Pak for Data 获取其他功能

其他企业功能可通过 IBM Cloud Pak® for Data 来获取，通过这个数据和 AI 平台，可使用 IBM 的 AutoSQL 和数据虚拟化功能。凭借这种方式，它们成为智能data fabric的一部分，这其中包括数据治理、数据隐私和 AI 的自动化功能。利用这种data fabric，企业不仅能够从单个自助服务点轻松移动和查询数据存储，还能够在数据传入时添加元数据，为那些没有合法需求的人屏蔽私有数据，并与相关系统建立顺畅的连接，以将数据转化为洞察。

IBM Cloud Pak for Data 也建立在 Red Hat® OpenShift® 基础之上。这意味着只要适用 OpenShift 就可适用该平台，它甚至可以在其他供应商的云端使用。这种程度的开放性有助于用户避免供应商锁定所带来的麻烦和费用。此外，在需要将平台放置在数据附近的情况下，该平台的高度容器化和移动性特征让实现这一现实比先前更简单。

如前所述，AutoSQL 功能以数据虚拟化、自动化治理和云对象存储等关键技术为基础。

这其中的每项技术对于 AutoSQL 整体功能都具有重要意义：

数据虚拟化
IBM 的数据虚拟化功能是使用数据联合构建的，且在顶部运行抽象层。通过这种方式，它可直接访问数据库、数据仓库、开源存储库（例如 Hadoop）和流数据存储，而无需移动数据。此外，即使地理位置分散或跨本地和云部署时，也可以访问数据。同时还可以跨多个供应商的产品访问数据，这有助于降低通常与供应商锁定相关的成本。

尽管有这些优势，但数据虚拟化所带来的最大节省则源自于它减少了成本高昂的 ETL 过程。通过使用数据虚拟化，ETL 请求预计将减少 25% 到 65%。在三年期间，这可以节省 90 到 240 万美元。[4] 这些节省可能是由以下几个原因而产生的。首先，由于数据驻留在一个地方，因此不再产生任何数据传输费用。这也可以帮助员工大幅提升效率，让他们能够将时间投入到更有价值的活动中去。他们可以通过单一访问点立即访问所需数据，而不必传输、复制和等待这些数据。因此，数据虚拟化为强大的数据访问奠定了基础，避免了传统上会产生的麻烦。

自动化的数据治理
IBM 的数据治理功能使企业能够发现、整理、分析、准备和共享数据。此外，数据治理还支持在整个企业内自动应用行业特定的治理规则。数据治理可确保企业的数据符合已定义的规则和流程。企业可以快速且精准地实施经过修订的法规或新法规，因而能够避免因不合规而导致的高昂罚款。

访问 IBM 的数据治理页面 →

云对象存储云对象存储 (COS) 是一种用于存储海量数据的解决方案。COS 使企业能够制定适当的数据增长和灾备策略，从而适应数据的爆炸式增长趋势。为了满足多变的数据需求，COS 为企业提供各种类别和层次，为实现数据可访问性提供了必要的灵活性和所需的性能。此外，COS 还提供了高速数据传输，可满足等待时间、带宽和成本需求。使用 AutoSQL 功能，数据也可以保存在更便宜的对象存储数据湖中，无需移动即可查询。

访问 IBM 的云对象存储页面 →

AutoSQL 功能为企业优化了数据可访问性和可用性。通用的分布式查询引擎：

- 最大限度地减少了对多个查询引擎的需求
- 最大限度地减少了数据迁移和数据复制需求
- 减少了数据仓库占用空间和相关成本
- 由于其与供应商无关的设计和独立的存储与计算扩展功能，提供了灵活性
- 嵌入了自动化治理功能

数据管理选项

无论数据存储在何处，都能够访问数据并轻松查询，这一目标还得到了一系列数据存储方案的支持，这些数据存储方案可以高效地满足数据驱动的 AI 就绪企业的需求。IBM 提供了许多能够满足这些需求的数据存储方案。

IBM DB2 - AI 数据库

IBM Db2® 数据库长期以来一直提供出色的企业级性能，最近又扩充了一些功能，这让它们既由 AI 提供支持，又为 AI 而打造。

由 AI 提供支持
机器学习查询优化
- SQL 性能随着时间的推移会受到监控，允许为特定的 SQL 语句创建和优化模型。使用更高效的访问路径，从而加快查询执行速度并减少资源消耗。

基于置信度的查询
- 根据历史查询结果的先前准确性，使用机器学习对查询结果的准确性进行评分。

自适应工作负载管理
- 通过使用机器学习，可以监控工作负载运行时，这既可用于调整正在进行的工作负载，也可用来预测利用率。观察结果表明，数据库性能提高了 30%。[5]

为 *AI* 而打造
原生图形功能
- Db2 中的多模型数据管理充分利用了图形数据库支持动态多维数据管理的能力，同时降低了拥有单独数据库的费用。

原生区块链支持
- Db2 区块链连接器将区块链数据呈现为 Db2 关系表，支持将其与 Db2 数据一起分析。

语言支持
- 支持 REST API，以及 PYTHON 和 GO 等语言、JSON 等架构及 Jupyter Notebooks 等协作开发环境。

企业级
IBM BLU Acceleration
- 结合内存计算、大规模并行处理 (MPP)、可操作的压缩、数据跳过和基于列的影子表，在不影响事务可靠性的情况下提高速度和性能。

备份和恢复
- 使用 IBM Db2 pureScale® 集群技术和分散的地理位置来避免业务中断。HADR 以及基于变更队列的复制和变更数据捕获复制功能支持所有同步模式。

安全和加密
- 与支持密钥管理互操作性协议 1.1 的集中式企业密钥管理器集成，并且可以在世界各地托管，以符合监管要求。

阅读 Db2 解决方案简介 →

IBM Db2 Warehouse

IBM Db2 Warehouse 具有许多与 Db2 数据库相同的特性，并提供旨在改进分析工作负载的额外功能：

IBM BLU Acceleration
先前描述的 BLU Acceleration 技术还将帮助加速分析工作负载。

灾备能力
云提供商的原生 Kubernetes 服务会自动检测未正常运行的计算节点，该服务会从集群中删除该节点，并从热备用池中提供一个新节点，或者及时配置一个节点。

分析能力
可以针对数据运行多种算法，包括关联规则、方差分析、k 均值、回归和朴素贝叶斯算法。同时还支持原生 Python 驱动程序，并可集成到 Jupyter Notebooks 中。

阅读数据仓库电子书 →

Netezza Performance Server

Netezza® Performance Server 来自一系列数据仓库设备，这些设备建立在以最少的努力实现极高性能的理念之上。

简便性
即取即用的性能，几乎不用建立索引或调优，因而降低了管理需求，减少了日常维护。为提供灾备能力而打造；节点故障不会导致性能显著下降。

快速
独特的非对称大规模并行处理 (AMPP) 和 IBM 享有专利的混合柱状加速辅助功能，可快速提供结果。凭借速度更快的内核和先进的 NVMe 闪存驱动器，它能够以更高的速度支持数千名用户。

智能
包含了一个涵盖 200 多个预先构建且可扩展的数据库内分析函数的库。这包括与行业标准 ESRI GIS 格式兼容的数据库内地理空间分析。

阅读 Netezza Performance Server 解决方案简介 →

数据湖和开源选项

IBM 为数据湖和开源数据管理需求提供了多种选择，对于 Hadoop 尤为如此。

Cloudera
IBM 与 Cloudera 合作，为那些寻求创建或改进其数据湖的企业提供完美的 Hadoop 实施。

Big Match
这种企业就绪技术可匹配与同一客户关联的多个碎片化或重复记录，从而提高了数据湖的性能。它使用预先配置的算法，在 Hadoop 中以原生方式对相似度进行评分并匹配记录。

MongoDB 和 PostgreSQL
IBM 提供了所需的支持，以便充分利用 MongoDB 的 JSON 文档存储和大容量数据存储以及对象关系数据库 PostgreSQL。

阅读数据湖电子书 →

IBM Db2 Event Store

Event Store 专为获取和分析高速流数据而设计。

高速获取
使用 3 节点系统每秒处理 300 万个事件，并在需要时
仅使用这 3 个节点一天内便可获取超过 2500 亿个事件。

为开源准备就绪
使用 Apache Parquet 的列式数据格式来存储数据，以实现通用访问，并可轻松集成到开源堆栈中。

新数据和现有数据
分析 Event Store 获取的数据以及历史数据，以提供更准确、更明智的洞察。

阅读 Event Store 解决方案简介 →

后续步骤

当您的数据分散到各种不兼容的存储库中时，不要无助地袖手旁观。相反，应在不移动数据的情况下聚合这些存储库并对它们进行查询，就好像它们都是单个数据仓库一样。通过节省时间和精力，使您的企业能够更快地获得更有价值的洞察力，并提升性能。

立即免费试用 IBM Cloud Pak for Data，亲眼目睹其中的不同之处。如果您有任何问题，可预约与 IBM 专家的免费会议，他们都乐于分享自己的建议。

国际商业机器中国有限公司
北京市朝 阳区 北 四环 中路27 号
盘古大观写字楼25层
邮编: 100101

美国出品
2021 年 6 月

1 IDC/Segate Rethink Data 调查，2020 年

2 西格玛研究，2020 年

3 基于内部测试

4 New Technology: The Projected Total Economic Impact of IBM Cloud Pak for Data https://www.ibm.com/downloads/cas/V5GNQKGE

5 基于 IBM 内部测试