IBM Cloud Pak for Data 上的 DataStage

多云数据平台上的自动化数据集成解决方案

IBM

目录

新型 AI 的兴起推动了数据集成战略

根据 IDC 的数据，到 2020 年，全球存储数据量将增长近 17%，达到 6.8 ZB，到 2024 年的复合年增长率将接近 18%。数据的急剧增长不仅使得提取和管理企业范围数据所需的时间和金钱增加，而且开始影响用户的工作效率和客户满意度。但是随着人工智能 (AI) 技术的兴起，出现了解决这些问题的新解决方案。AI 技术加快了数据发现的步伐，拓宽了可利用的数据范围，能够自动完成以前需要人类专业知识的任务。Gartner 甚至表示，到 2024 年底，将有 75% 的企业从试验性采用 AI 转为实施 AI，这将推动流数据和分析基础架构增长 5 倍。

也就是说，AI 仅在全部数据值得信赖、可以访问并且符合标准时才能发挥效用。随着 AI 使用的增多，数据系统中早已存在的弱点和局限性凸显出来，因此企业必须转为采用新的现代策略。此类敏捷性需要一种新的信息架构，一个允许在整个数据生命周期中进行无缝集成和操作的架构。这就是 IBM 客户正在进行现代化并从传统系统过渡到现代云端架构的原因：IBM Cloud Pak® for Data.此数据和 AI 平台改善了各种工作负载的扩展性和弹性，并降低运营成本，同时能够连接云数据仓库和实时分析应用程序。

随着 AI 的兴起，有许多因素导致数据集成工具的部署和使用方式发生重大变化。这些可能是从企业中高数据多样性到数据用户需求的任何因素。由于因素众多，公司需要采用面向流程的方法来管理 DataOps 的数据生命周期、提高业务绩效和增强竞争力。在产品和流程上采用 AI 的公司将需要高度灵活和可扩展的数据集成技术，此类技术应嵌入市场领先的数据集成工具 IBM Cloud Pak for Data 上 的 IBM® DataStage®。它能够提高企业和 IT 用户的生产效率：

- 同类最佳的并行引擎和自动工作负载平衡，可灵活扩展您的工作负载，速度比本地 DataStage 快 30%。
- 一次设计，随处运行，将数据集成带入您的数据。
- 自动化作业设计并与 Netezza®、IBM Db2® 或云数据仓库、数据虚拟化或 DataOps 服务集成。

通过将该技术与数据平台上的其他服务无缝集成，企业可以全面而自动地配置数据，同时保持所需的性能、安全性和治理。容器化架构，尤其是部署在云平台（如 IBM Cloud Pak for Data）上的容器化架构，是这种转变的关键。

使用容器作为数据集成工具

IBM Cloud Pak for Data 通过云端原生设计，统一跨整个数据和分析生命周期的市场领先服务。这包括先前由 IBM InfoSphere®Information Server 平台提供的各项功能，这些功能现在作为 IBM Cloud Pak for Data 上的 IBM Watson Knowledge Catalog 云就绪服务提供。借助 IBM Cloud Pak for Data，您可以在统一的云端原生平台上以较低的成本简化数据集成，并且借助服务中包含的自动化功能，您的组织可以近乎实时地从数据中获得商业洞察力。

过去，通常将 IBM DataStage 和 Information Server 平台部署为处理大规模企业工作负载并执行关键任务功能。为了确保无缝过渡到现代化的 AI 和云就绪架构，DataStage 和 Information Server 的现代化升级易于迁移，可轻松访问平台上的各项功能，并提供更高水平的弹性、扩展性、自动化和运营效率。

通过在 IBM Cloud Pak for Data 上部署 DataStage 和 Watson Knowledge Catalog 服务，企业可以利用打造行业领先的数据平台的所有强大功能。

针对 AI 构建

- 与 Watson Knowledge Catalog 进行在线数据质量和元数据交换，以改善数据治理。
- 与 IBM Cloud Pak for Data 上的数据科学、事件消息传递、数据虚拟化和数据仓库服务进行开箱即用集成。

AI 驱动

- 通过内置的设计加速器（例如阶段建议、模式传播和自动作业模板生成）提高用户的生产效率

IBM Cloud Pak for Data 上的 IBM DataStage 是 IBM InfoSphere DataStage 的容器化版本，基于微服务架构，已针对 Kubernetes 进行优化。通过 IBM Cloud Pak for Data，DataStage 可在世界领先的容器编排平台 Red Hat® OpenShift®上本地运行。

通过将 DataStage 功能分解为微服务而不是单体栈，您可以获得多种机会：

- 数分钟完成部署；支持标准部署和管理，同时保持按需修改参数的灵活性。
- 具有开箱即用的增强型 Kubernetes 可用性，并且支持高可用性/灾难恢复 (HADR) 自动故障切换，因此可靠性高。
- 自动化更新减轻管理负担。一键部署服务包、版本和模式。
- 通过“应用程序组”自动化管理；管理员可以使用名称空间来管理访问控制和配置选项。
- 借助平台和服务级功能在应用程序级监控和管理。
- 独立扩展微服务，以应对不断变化的需求。

容器化数据集成技术可让您将 DataStage 运行为混合云端环境（云端与非云端平台的组合）或多云环境（来自不同提供商的云端）的一部分，以对每种类型的数据使用适当的基础架构。

最近容器之所以流行，正是得益于这些优势。根据“Red Hat 2019 全球客户技术概况“，57% 的企业已经在使用容器，预计容器的使用率在接下来 2 年还会增长 89%。借助 IBM Cloud Pak for Data，您可以更轻松地访问全部 IBM 服务，以设计、部署和管理有助于您实现商业价值的高级分析。

在 IBM Cloud Pak for Data 上部署 DataStage 的五个主要好处

1. 易于在单个统一平台上启用混合云和多云

据 Gartner 调查，大多数企业使用不止一个云提供商，从数据集成的角度来看，一直以来，在不同云平台及其本地数据源之间移动数据时产生的数据延迟和数据出口成本，始终是企业需要面对的挑战。之前，组织通常必须在多个提供商之间运行单个应用程序才能执行其数据集成作业，花费比本来所需更高的时间和成本。但是现在，借助 IBM Cloud Pak for Data 上的 IBM DataStage，用户可以自由选择提供一种解决方案的任意云提供商。借助 DataStage 中一次设计、随处运行的功能，用户可以在本地一次设计其工作，将运行时移至其数据所在的位置，从而避免数据延迟和数百万美元的出口成本。无需再将数据移出存储位置。

2. 并行处理和自动工作负载平衡

借助完全云端原生架构，DataStage 可以动态扩展工作负载，并使用同类最佳的并行引擎 (PX) 对大型数据集进行优化。用户可以选择在 IBM DataStage Flow Designer 中创建并行 或 Apache Spark 作业。

此外，与传统的本地 DataStage 相比，使用 IBM Cloud Pak for Data 的 IBM DataStage 可将客户的执行时间缩短约 30%。由于自动工作负载平衡可在 OpenShift 集群中的工作节点之间分配工作负载并最大化吞吐量，因此这些性能改进在资源争用的执行窗口期间尤为明显。

3. 自动化的工作设计和研发支持节省了开发时间和成本

为了解决跨不同的操作系统管理众多容器化应用程序的挑战，组织需要一个强大的开源工具，例如 IBM Cloud Pak for Data 上提供的 Red Hat OpenShift。IBM Cloud Pak for Data 平台可帮助他们扩展和配置容器，以支持关键的 IT 计划，例如微服务和云迁移策略。DataStage 容器允许创建和自动化持续集成/连续交付 (CI/CD) 管道，用于从开发到测试再到生产的作业。它们还通过支持诸如 GitHub 之类的源代码控制工具来频繁地发布作业并将其发布到生产环境，从而帮助简化 CI/CD 管道。

IBM DataStage Flow Designer 功能丰富，包括：内置搜索、让公司快速入门的快速教程、自动元数据传播、智能调色板、建议阶段以及同时突出显示所有编译错误。开发人员在设计作业时可以使用这些功能来提高生产效率，其效率可能超出传统手工编码作业的 9 倍。与手工编码相比，使用视觉和 ML 辅助设计时，用户可以节省多达 87% 的开发成本。

许多公司在一个项目中包含数千个作业，而且他们依赖这些作业维持日常运营。重写这些作业可能会出现错误和中断，对他们来说是不可接受的。这些公司使用 IBM Cloud Pak for Data 上的 DataStage Flow Designer，可以采用任何现有的DataStage 作业，并将其呈现在瘦客户端中，因此无需重写这些作业。此外，通过使用 DataStage Flow Designer 瘦客户端，无需购买胖客户端来进行工作设计，客户可以节省数百万美元的许可成本。

除了设计和开发功能外，DataStage 还为 Amazon S3、Azure、Db2、Hive 和 Kafka 提供数百种开箱即用、预建、随时可用的连接器，并且还提供转换、编码、注释、收尾和合并等阶段。这大大减少了开发人员为分析操作而准备数据的时间。每隔几周就会增加新的运算，因此开发人员的生产效率会越来越高。

4. 与数据和 AI 服务内置集成

借助 IBM Cloud Pak for Data 上的 DataStage，可以轻松利用来自更广泛 IBM 和开源生态系统的各项功能。该平台包括许多核心服务，从数据仓库、Watson Knowledge Catalog、数据科学和数据虚拟化到事件消息传递。在 IBM Cloud Pak for Data 系统上与 Netezza 和 Db2并置，可消除网络瓶颈并支持高速数据交付。无论数据驻留在哪个云平台上，都可以使用 Snowflake 和 Amazon Redshift 的预建连接器轻松连接云数据仓库，以访问和转换数据。

为了防止数据湖因数据未使用而变成“数据沼泽“，您可以在目标环境（例如数据湖）吸收数据的同时，使用 IBM InfoSphere QualityStage® 跟踪 ETL 作业中的数据沿袭，以自动解决质量问题。您还可以提供元数据支持，以策略驱动方式访问敏感数据，并防止未经授权的用户访问您的敏感数据。这种数据质量概念可以扩展到支持整个企业数据仓库 (DWH) 的全面数据治理。

借助 IBM Cloud Pak for Data 上包含的数据虚拟化功能，业务用户可以发现数据、查询数据并尝试数据仓库流，同时还可以执行基于 SQL 的简单数据转换、运行开发和测试以及管理结构化和非结构化数据。

5. IBM Cloud Pak for Data 中 Red Hat 的价值

IBM Cloud Pak for Data 所基于的 Red Hat OpenShift 的优势，增强了混合云和多云选项。Red Hat 栈、OpenShift 和 Kubernetes 一起运行的优势尤为明显。它使您能够开发安全且可扩展的 Kubernetes 应用程序，而不会被大规模手动 Kubernetes 管理的复杂性所困扰。通过使用 Kubernetes Operators，Red Hat OpenShift 为容器栈的每个部分（操作系统、Kubernetes 和集群服务、应用程序以及持久数据存储）提供自动化的安装、升级和生命周期管理。

OpenShift 提供一个全面平台，该平台支持自动化操作，并提供对 Java、Node.js、Ruby 和 Python 等语言的开箱即用支持。OpenShift 还提供监控、身份验证和授权以及网络管理等支持服务。这些 OpenShift 功能不在开源版本的 Kubernetes 中。

此外，随附的 Kubernetes 发行版为企业级，并且可从每个版本的数百个安全性、缺陷和性能修复程序中受益。同时提供经验证的适用于 Kubernetes 的流行存储和网络插件。最后，开源 Red Hat 工具提供了其他功能选项，例如用于流数据的 Apache Spark 或用于机器学习应用程序的流行 Python 和 R 语言。附加功能可确保企业利用通过 OpenShift 平台开发、部署和运行应用程序所需的必要开源工具。当这些不同的资源全部成为 IBM Cloud Pak for Data 中单个统一平台的一部分后，集成和管理就变得比以往更加容易。

后续步骤

通过 IBM Cloud Pak for Data 进行部署时，DataStage 不仅是强大的数据集成工具，还可以大规模处理数据。它已成为基于微服务的数据平台的一部分，该平台还可以帮助您组织和分析数据，并将 AI 功能融入您的企业。

IBM Cloud Pak for Data 上的 DataStage 提供：

1. AI 功能，专为 AI 项目打造
2. 自动故障解决和备份、恢复和补丁管理等操作任务的自动化，使操作成本降低多达 50%。
3. 内置的工作负载平衡和同类最佳的并行运行时，可优化工作负载执行，与传统 DataStage 相比，工作负载执行速度提高 30%。
4. 与手工编码相比，使用可视化和 ML 辅助设计可节省 87% 的开发成本。
5. 使用一次设计，随处运行功能，可将集成工作负载引入数据中，从而节省数据移动成本。
6. 使用通用 UI 与数据科学、数据仓库和数据虚拟化服务进行预建集成。

现有客户无需购买 Windows 或 Citrix 胖客户端许可证，即可保留其在技能和资产上的投资，并节省数百万美元的许可证成本。

IBM Cloud Pak for Data 上的 DataStage 提供容器化架构、Red Hat 基础结构，数据连接性和更广泛 IBM 功能生态系统的独特组合，对于希望为未来机遇准备数据基础的企业而言，是极具吸引力的选择。

要开始使用，请免费试用 IBM Cloud Pak for Data
安排与数据集成专家的免费一对一咨询。

IBM Corporation,New Orchard Road, Armonk, NY 10504

美国印制，2020 年 10 月

7EB2XONR