IBM DataStage

使用 IBM Cloud Pak for Data DataStage 实时
为 AI 交付业务就绪数据

通过数据集成交付业务就绪数据

当今的数字企业正在创建和使用前所未有的数据。这包括跨多个系统和存储库存储的有关客户、交易和员工的数据。这些数据存储分布在各种多云、混合云环境和数据湖中，因此组织正在寻找方法将这些不同的数据源和环境联合起来，以**使用 AI** 获得更快的见解，向客户提供差异化和个性化的体验。根据 Forrester 调查，数据科学家将 80% 的时间花在为 **AI 计划**准备和管理数据上。这些结果再加上 IBM 的一项调查 - 91% 的组织没有有效使用其数据 - 意味着企业正在费劲地从各个数据孤岛中交付价值。用于从大量数据中实时访问所需数据并交付业务就绪数据的架构技术、实践和工具，被称为数据集成。利用灵活且可扩展的数据集成技术，企业可以通过提取、转换和加载 (ETL) 多个数据源上的数据，执行对下一个最佳优惠的分析、客户流失检测和分析、供应链预测以及即时欺诈检测。

对于疲于管理跨多个云端或数据湖的数据，并希望缩短构建和更新 AI 模型和应用程序所需时间的 CXO、企业架构师或运营领导者来说，IBM® InfoSphere™ DataStage 是他们的理想选择，它是一款超越 ETL 交付可信业务就绪数据功能的市场领先数据集成解决方案，可提供可扩展的多云数据集成和交付解决方案，以确保实时使用可信的业务就绪信息。DataStage 中的关键功能包括多云运行时支持 - 只需使用设计一次，即可在任何云上运行，同时能够通过自动工作负载平衡和低延迟并行引擎来扩展工作负载。此外，它还通过内置复制技术来实现实时数据交付，通过持续集成和持续交付 (CI/CD) 支持来减少 DevOps 的时间和成本，通过自主集成设计和验证规则来快速构建 AI 模型，并通过使用在线数据质量来自动检测和解决数据问题。

DataStage 是 IBM DataOps 功能的一部分，用于实现连续的高质量数据以支持 AI，并在适当的时间从任何数据源向适当的人员提供自动的自助数据管道。IBM InfoSphere DataStage 适用于本地、IBM Cloud 和超融合平台（例如 IBM® Cloud Pak™ for Data），可部署在任何地方。IBM® Cloud Pak™ for Data 是一个完全集成的数据和 AI 平台，基于 Red Hat® OpenShift® 而构建，提供完全云原生的 DataStage 架构，可随您的业务发展而扩展。它还为组织提供一个支持多种数据交付方式的平台，包括数据集成、数据复制和数据虚拟化，而 CDC 则在发生基于日志的更改时捕获这些更改，并使用基于 Kafka 的消息队列将信息传递到云端和数据湖上的目标数据库。

设计一次，即可在任何云上运行

根据 IDC 调查，90% 的企业客户在使用多个云。利用多云数据集成，用户可以将设计与运行时分开 - 您可以只设计一次 ETL 作业，然后通过容器将运行时组件部署到任何云环境上，以减少因处理大量数据而导致的延迟。您可以在本地创建和测试作业，然后在云环境（例如利用云上 Azure 数据湖的 Microsoft Azure 实例）中运行它。作业参数及其值通过 Kafka 消息传递到 DataStage 的远程实例。

多云数据集成具有以下优点：

- 能够跨本地和云环境集成数据
- 自动化作业设计体验，可简化设计过程
- 远程作业执行，可最大程度地减少移出数据的成本
- 满足地缘政治要求
- 减少了处理大型数据集的延迟，因为数据保留在现有位置，不需要移动

自动工作负载平衡和并行处理

借助完全云原生的架构，您可以使用 DataStage 的本地容器或共享容器来动态扩展工作负载，并使用同类最佳并行引擎 (PX) 来优化大型数据集。用户可以选择在 IBM DataStage Flow Designer 中创建并行、串行或 Apache Spark 作业。

您可以在两个运行时引擎上运行 DataStage Flow Designer 作业：

- 作业类型为并行或串行的作业只能在并行引擎上运行。通常，资源密集型作业在并行引擎上运行，因此，使用并行处理完成复杂作业的平均时间为两分钟。
- 作业类型为 Spark 的作业只能在 Spark 引擎上运行。

实时数据交付

DataStage 完全内置部署为容器的更改数据捕获 (CDC) 技术，支持实时捕获，可提供最佳的数据集成和数据复制功能。DataStage 允许对大型数据集进行复杂转换，而 CDC 在发生基于日志的更改时捕获这些更改，使用复杂转换对其进行转换，然后使用基于 Kafka 的消息队列将其传递到云端和数据湖上的目标数据库。DataStage 还允许将基于批处理和基于偶数的批量数据转换作业馈送到数据仓库。

通过 CI/CD 支持来减少 DevOps 的时间和成本

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 Cloud Pak for Data 上提供的 Red Hat OpenShift。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/持续交付 (CI/CD) 管道，以支持从开发到测试和生产的作业流程，并通过支持 GitHub 等源代码控制工具来支持 CI/CD 管道，从而频繁地发布作业并将其投入生产。

自主集成设计助力 AI

通过自动发现和分类资产、生成基于内置自定义转换和质量规则的集成流以及检测和保护敏感信息，以更快的速度规模化收集和集成支持 AI 的数据。

通过自动化作业设计
快速实现价值

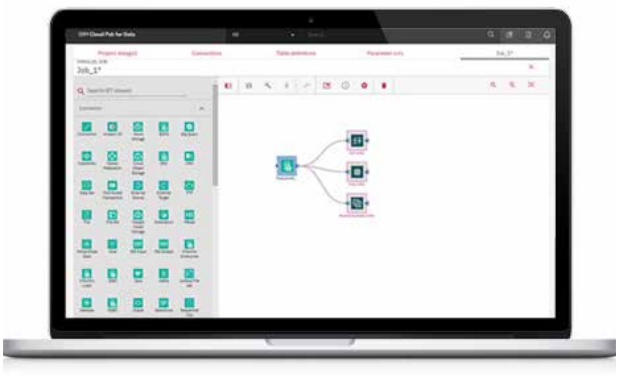

图 1具有自动设计功能的 DataStage Flow Designer

IBM DataStage Flow Designer 是 DataStage 中基于 Web 的 UI，具有机器学习 (ML) 功能，可帮助用户（甚至非技术用户）在作业中构建流程和阶段。

DataStage Flow Designer 具有以下优点：

- 向后兼容。不需要迁移作业。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。迁移可能会出现错误和中断，对他们来说是不可接受的。这些公司可以使用任何现有的 DataStage 作业并将其呈现在 IBM DataStage Flow Designer 中，因此不需要将这些作业迁移到新位置。
- 提高开发人员的生产效率。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。
- 广泛的运算符和连通性。除了设计和开发功能外，DataStage 还提供数百种现成的预建、即用型运算符。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符，因此开发人员的生产效率会越来越高。

动态数据质量和安全性
确保可信的数据交付

DataStage 在数据集成方面提供独特的用户体验，它使用 DataStage Flow Designer 在将数据传递到目标环境（例如数据湖）时运行数据验证、标准化和匹配规则，以防止出现质量问题和潜在安全问题（例如允许未经授权的用户访问您的敏感数据）。这种数据质量概念也可以扩展到支持整个数据仓库 (DWH) 的全面数据治理。

总结

DataStage 的优点：

- 通过内置的自动工作负载平衡、并行性和可扩展性，实现一次设计，随时随地运行
- 实时或使用基于批处理的交付方式捕获更新
- 内置弹性、易操作性和 CI/CD
- 为 AI 而优化的数据集成
- 使用 ML 功能的自动化作业设计
- 动态数据质量和数据安全性确保可信的数据交付

IBM 跨混合多云环境、本地、超融合系统（例如 IBM Cloud Pak for data）或任何所选云平台提供广泛的数据集成功能。这些不同功能促成了一个灵活且可扩展的数据集成解决方案，可在他们选择的部署模型上快速访问大量适合 AI 的高质量数据。

免费观看引导演示，以详细了解
IBM InfoSphere DataStage

为何选择 IBM?
IBM DataOps 的功能有助于创建业务就绪的分析基础：通过提供市场

领先的技术，并结合使用人工智能自动化、注入式治理和强大的知识目录，在整个企业中运营连续的高质量数据。提高数据质量，以便在适当的时间从任何来源向适当的人员提供高效的自助数据管道。

要了解有关 DataOps 的更多信息，请访问
ibm.com/dataops
要了解有关 IBM InfoSphere DataStage 的更多信息，请访问
ibm.com/products/infosphere-datastage
访问大数据和分析中心：
ibmbigdatahub.com

IBM Corporation
New Orchard Road, Armonk, NY 10504
美国印制
2020 年 4 月