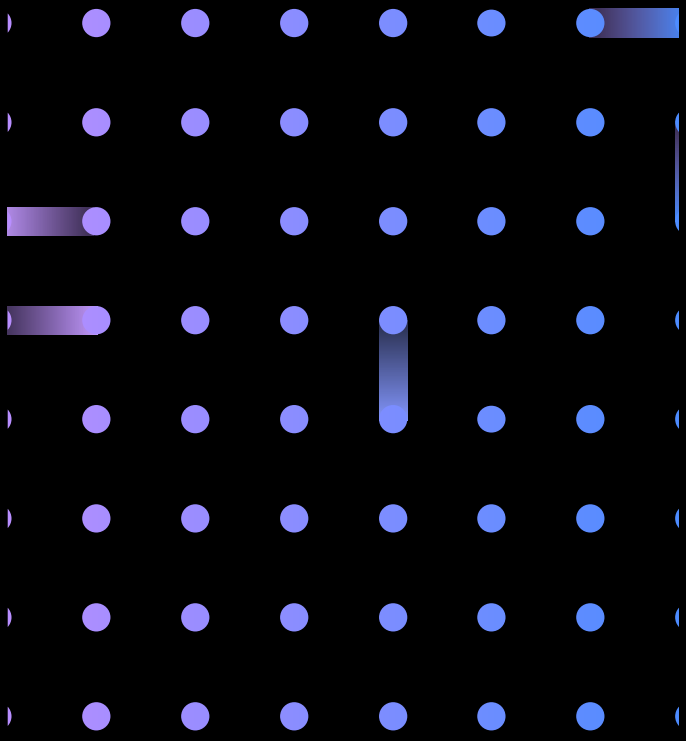
使用 DataOps 快速交付业务就绪数据

IBM DataOps 方法与实践简介



IBM

目录

亮点

- DataOps 是人员、流程和技术的有机结合，用于快速向数据公民提供可信的高质量数据
- 在自动化的推动下，DataOps 可解决与数据访问、准备、集成和交付方面效率低下相关的挑战。
- IBM DataOps 实践利用高度自动化对以下功能产生重大且可衡量的影响：数据整理服务、元数据管理、数据治理、主数据管理和自助交互
- IBM 通过规范性方法、人工智能 (AI) 自动化和 IBM DataOps Center of Excellence，提供通往 DataOps 实践之路。
- DataOps 研讨会是 DataOps 路线图的组成部分，可帮助组织评估其 DataOps 成熟度和规划试点项目的执行方案

简介

数据是支持创新和保持竞争优势的燃料。它是推动分析和了解商业趋势与机遇的关键因素。以新的方式发掘数据的价值甚至可以加速组织的 AI 之旅。

然而，当与数据相关的项目未能实现承诺的投资回报 (ROI) 时，利益相关者要刨根究底。根据 Experian 的《2019 年全球数据管理研究》报告，89% 的企业表示他们在管理数据方面存在挑战。这些挑战包括获取洞见的延迟和对底层数据缺乏信任。[1]

了解组织的业务目标对于为分析和 AI 制定有效的数据策略至关重要。任何商业模型要产生效果，必须满足客户的需求。成功依赖于使用集成的业务就绪数据管道来简化数据运营，能够在任何时间点提供完整和一致的业务视图。

企业对于更快取得成果的期望与日俱增。世界各地的企业都在寻找提高运营效率和有效性的方法，以实现最佳决策，特别是其组织内存在许多数据孤岛的企业。这两个因素促使企业领导者寻求能够在单个框架内解决其严峻挑战的新方法。

对于在数据运营方面寻求转型的组织来说，自动化技术可以带来竞争优势。当可信的业务就绪数据有助于推动组织的差异化见解和卓越运营时，数据就变得有价值了。

本白皮书重在阐述 DataOps 方法、实践和路线图的好处。

定义 DataOps

数据运营 (DataOps) 是人员、流程和技术的有机结合，用于快速向数据公民提供可信的高质量数据。该实践的重点是支持整个组织的协作，以推动规模化推动敏捷性、速度和新的数据计划。DataOps 旨在利用自动化的力量，解决与数据访问、准备、集成和交付方面效率低下相关的挑战。

DataOps 的潜在好处包括显著提高向个人提供信息和数据的能力，以及改进流程以提高效率和实现优化。包含 AI 数据主导计划的自动化数据运营有助于推动以下成果：

- 提供集成的业务就绪数据，以规模化推动分析和 AI
- 实现运营效率
- 满足数据隐私和合规性要求

89%

的企业在管理数据方面存在挑战。[1] 了解组织的业务目标对于为分析和 AI 制定有效的数据策略至关重要。

DataOps 不是 DevOps

大多数组织已在其开发领域实施了某种程度的 DevOps。DevOps 实践的广泛熟悉度和命名规范的相似性，不由得让人将其与新兴的 DataOps 实践进行比较。虽然两者都是推动运营最佳实践的方法，但它们在组织中各有其独特的位置。

在下表中，查看两种实践在组织目标和收益方面的比较。

比较目标

	DataOps	DevOps
主要目标	快速使用业务就绪、可信赖、高质量的数据。	应用程序和软件开发
转型目标	– 通过让所有数据公民能够自助访问可信的高质量数据，推动业务的持续快速创新 – 通过自动化数据治理、集成和避免监管问题，实现数据的连续交付 – 通过监控和优化数据管道，提供不断向所有数据公民学习的反馈回路	– 通过支持整个价值链的协作开发和测试，加快思想的持续创新 – 通过自动化软件交付流程并消除浪费，实现这些创新的持续交付，同时还有助于解决监管问题 – 通过监控和优化软件驱动的创新，提供不断向客户学习的反馈回路
效率目标	– 通过在 IT 系统支持、运营和业务之间建立更紧密的联系，纠正人员和目标的偏差 – 通过在整个数据交付周期中引入自动化，加快变更的交付并提高交付质量 – 通过使用结果驱动优化，提高对元数据和数据之真实价值的洞察	– 通过在开发人员、运营和业务之间建立更紧密的联系，纠正人员和目标的偏差 – 通过在整个开发周期中引入自动化，加快变更的交付并消除其中的错误 – 通过使用客户反馈驱动优化，提高对应用程序之真实价值的洞察

DataOps 是人员、流程和技术的有机结合，要履行 DataOps 实践的承诺，需要跨所有职能部门进行深度协作。它需要重点培养数据管理实践和流程，以提高分析的速度和准确性。

人员和流程

DataOps 利用自动化技术支持高生产力团队，以帮助提高项目产出和交付速度方面的效率。但是，要体验这些好处，内部文化需要向真正的数据驱动演变。随着越来越多的业务部门需要并希望管理数据以实现背景洞察，现在正是采取以下措施的时候：

- 提高流向组织的数据的质量和速度。
- 获得领导层的承诺，以便在整个企业中支持和践行数据驱动的愿景。

这种类型的变革始于理解企业的真正目标。*数据如何为影响客户的决策和服务提供信息？数据如何帮助保持市场竞争优势？数据可以帮助我们解决哪些财务优先事项？*

DataOps 领导者应定义所有数据公民在推动文化和 DataOps 实践向前发展中所扮演的角色。每个组织都有其独特的需求，IT、数据科学和业务领域的利益相关者需要增加价值来推动企业的成功。此外，利用现有的数据治理委员会和从长期数据治理计划中获得的经验教训有助于建立这种文化和承诺，因为治理是支持 DataOps 所需的驱动力之一。

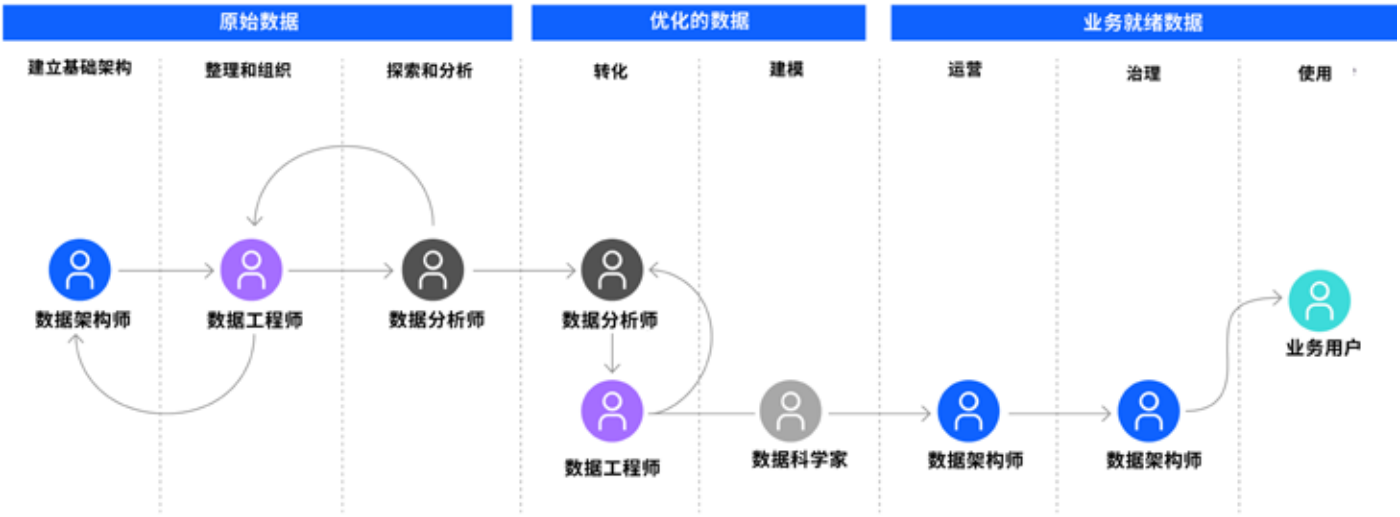


图 1：基于角色的 DataOps 工作流程示例

技术

DataOps 的核心是组织的信息架构。*您了解您的数据吗?您信任您的数据吗?您能够快速检测到错误吗?您能够在不“破坏”整个数据管道的情况下进行增量更改吗?* 要回答这些问题,第一步是清点您的数据治理以及数据集成工具和实践。工具对于支持任何依赖自动化的实践都是必需的。

在考虑使用工具来支持组织内的 DataOps 实践时,请考虑以下五个关键领域的自动化如何转换数据管道:

1. 数据整理服务
2. 元数据管理
3. 数据治理
4. 主数据管理
5. 自助交互

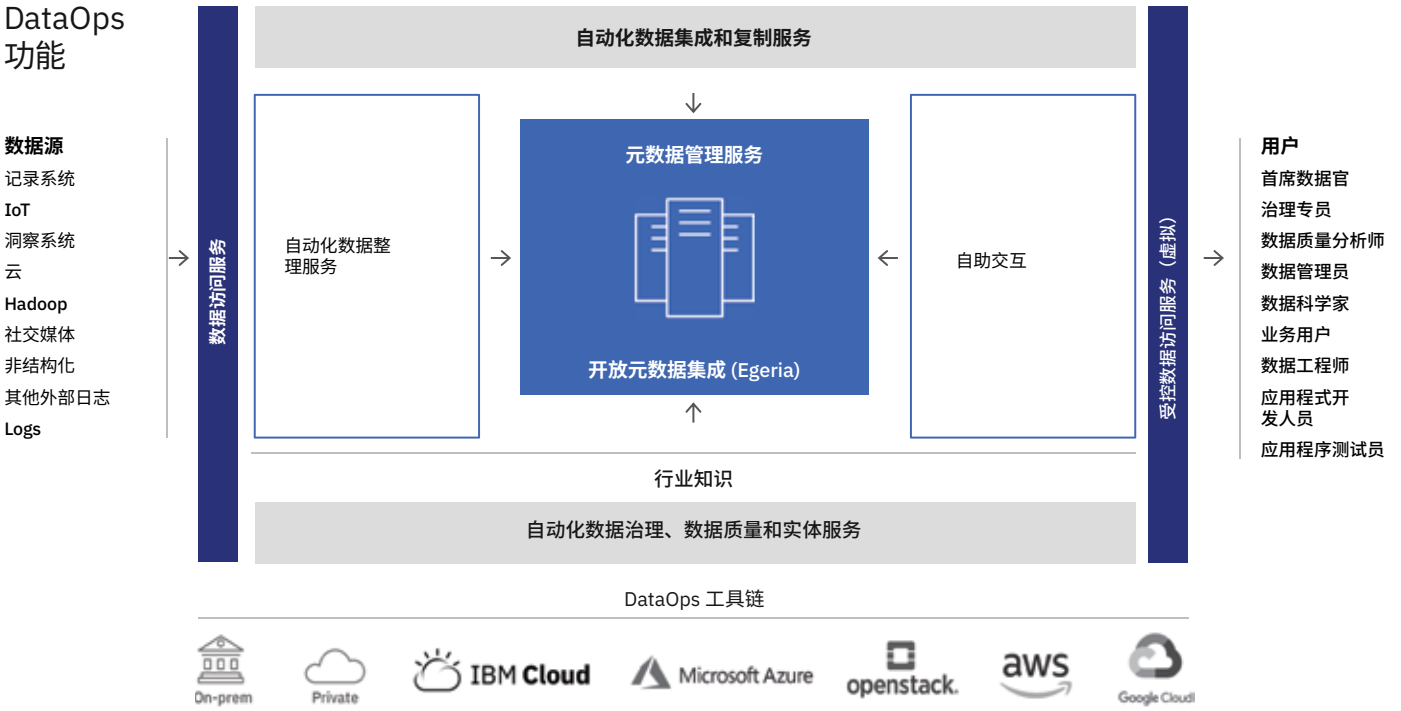


图 2:剖析支持 DataOps 的信息架构

业务就绪数据的交付包括所有这些方面,任何 DataOps 实践都必须包括涵盖所有 5 个方面的整体方法。对于数据管道的要素顾此失彼的组织不太可能从实施 DataOps 实践中获益。技术对话和实施不应与有关人员和流程的持续规划隔离开来。工具有助于支持和践行文化。

底线

当组织仍在为定义数据管理员的角色或创建数据验证规则等基本问题而挣扎时,DataOps 似乎令人望而生畏。但实际上,DataOps 实践为组织在其数字转型计划中所经历的许多失败提供了解决方案。

组织认识到,最普遍的失败例子发生在其数据湖中。许多组织正在进行第二次、第三次或第四次尝试,以寻求技术上的成功,并且重振了领导团队,以实施所需的文化变革。*但是,为什么这些数据湖实施在过去失败了呢?*

因为许多此类项目的重点只是将未清理和未治理的数据扔到数据湖中。在有效解决人员、流程和技术问题方面的局限性,也极有可能导致失败。

IBM DataOps 计划

向 DataOps 转变是客观事实。根据最近的一项调查，73% 的公司计划投资 DataOps。[2] IBM 通过规范性方法、领先的技术和 IBM DataOps Center of Excellence (CoE)，提供通往 DataOps 实践之路。在 CoE 中，IBM 专家与组织合作，根据业务目标定制方法，并确定合适的试点项目，为利益相关者创造价值。

IBM DataOps能力通过提供业界领先的技术来帮助交付业务就绪数据，提供业界领先的技术，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。它可以提高效率、数据质量、可查找性，并嵌入管理规则，以便在适当的时间从几乎任何来源向适当的人员提供自助数据管道。

从有助于治理数据湖的解决方案到开发应用程序并帮助确保合规性，IBM DataOps 可帮助组织展示数据在优化决策和时间方面的价值。当组织能够了解、信任在云端和任何关键环境中的数据并使用这些数据推动价值时，交付高质量的企业数据以实现 AI 是指日可待的。

IBM Cloud Garage 方法符合 DataOps

IBM Cloud Garage 方法是一套使业务、开发和运营能够持续设计、交付和验证新功能的方法。其实践、架构和工具链涵盖了整个产品生命周期，从开始到捕获和响应客户反馈与市场变化。开放式工具链架构旨在简化 IBM Cloud™ 平台服务（例如持续交付 (CD)）与开源和领先的第三方工具的结合，以形成符合 DataOps 实践的集成工具链。这些模式可以作为模板在团队之间共享，以促进整个组织中 DataOps 的成功采用。

为了成功实施 DataOps 实践，IBM 已确定 DataOps 生命周期的六个阶段以及必要的文化考虑因素。这些阶段基于向 IBM 转型之旅中内部采用 DataOps 的过程。

IBM Cloud Garage 方法将这六个阶段描述为：

- **思考。**功能的概念化、完善和优先级排序
- **代码。**功能的生成、增强、优化和测试
- **交付。**产品的自动化生产和交付
- **运行。**运行所需的服务、选项和功能
- **管理。**产品的持续监控、支持和恢复
- **学习。**基于实验结果的持续学习和反馈

图 3IBM Cloud Garage 方法的六个阶段

思考：持续评估您的 DataOps 成熟度并使其与业务目标保持一致

DataOps 可以转换为现有组织和已建立的流程。DataOps 的目的是自动化许多现有的手动任务，并简化数据管道创建过程。无论是开始实施还是维持基本的 DataOps 实践，评估团队快速交付业务就绪数据以及制定与创造业务价值相符的改进计划的能力都非常重要。

DataOps 的成功始于通过捕获元数据并为数据类分配策略、评估和评分数据质量以及利用集成数据的工具（而不是电子表格、部落知识或手工编码）对数据资产进行分类。确定团队的成熟度级别后，目标应该是尽可能改进多个 DataOps 方面的功能。

DataOps 团队应该专注于使必要数据的交付符合它能为业务带来的价值。提出问题：*如果能更快地交付这些信息，可以省多少钱或赚多少钱？*

代码：使用版本控制系统 - 源代码控制管理

数据管道是负责将原始内容转换为有用信息的源代码。这种管道是数据分析的核心，可以端到端自动生成可重复使用的源代码。与分析相关的不同文件、配置和参数分布在组织内的不同位置和环境中，没有任何管理控制，这导致部署不一致。GitHub 等版本控制工具有助于存储和管理对代码和配置的所有更改。集中式存储库还可以帮助企业每次都能从不同环境中获得一致且可靠的信息，并且具有可应对任何事件或灾难的可靠恢复能力。版本控制还有助于团队并发进行开发工作，并通过使用分支和合并提高交付管道的敏捷性。

为了确保数据分析管道能够正常运行，必须对其进行测试。通过持续集成/持续开发 (CI/CD)，并辅以参数化，可以完全自动化地进行部署和测试。必须在数据分析管道的每个阶段对输入、输出和业务逻辑进行测试，并在发布之前检查其准确性或潜在偏差以及错误或警告，以确保质量稳定一致。手动测试在高性能组织中没有立足之地，因为它容易出错，且费时费力。稳健的自动化测试套件是实现 CI/CD 的关键要素，也是随需应变经济的基本要求。

交付： DataOps 过程和工作流自动化 - 数据技术

DataOps 方法要取得成功，自动化是必不可少的，并且需要具备运行时灵活性的数据分析管道。交付可信数据的关键要求是一个受控、一致的数据管道，它依赖于使用元数据和数据采样技术的数据整理、数据提取、目录和分类。

用于交付可信和受控数据的可重复、稳健的数据管道需要一种机制来执行以下活动：

- 一致地定义和实施数据治理和数据隐私策略。
- 支持高效的数据移动。
- 启动补救工具或采用特定于行业的最佳实践和带有预定义词汇表的模板。

此过程可以在不同平台上一致地部署受控数据管道，而无需更改任何源代码或配置，并提供完全受控和可信的数据。DataOps 过程还需要辅以适当的补救工具，以支持异常处理和管理。对任何更改的回溯和可审计性是这些受控数据管道的基本要求。

IBM 提供新的创新功能，包括嵌入式机器学习 (ML)、AI 自动化、注入式治理和强大的数据目录，可在整个企业中运营连续的高质量数据。DataOps 的效率取决于用于数据管道的数据技术组件的高度自动化。

- IBM Cloud Pak for Data® 包含 IBM Watson® Knowledge Catalog (WKC)，能够以高效、稳健、自动化和可重复的方式满足这些要求。
- IBM Cloud Pak for Data Server 可以满足数据管道中数据移动、发布和使用的需求，同时有助于确保数据质量和策略实施。利用高效的源代码控制管理，它可以在 CI/CD 管道中实现自动化并高效地执行。
- IBM Cloud Pak for Data 的 IBM Watson Knowledge Catalog 中内置的 ML 补充了自动化过程，并在每次迭代时对其进行优化，以完善稳健的管道。
- IBM Cloud™ DevOps Insights 有助于为数据管道提供运营见解和可视化。它基于高度自动化以及与 IBM Cloud Pak for Data 的定制集成，有助于实施持续监控的安全和质量措施，检测任何意外变化并生成运营统计信息。
- Apache Airflow 和 NiFi 有助于工作流设计及其编排。
- 通过 REST 端点使用高度自动化以及参数化，可以帮助动态选择特定的数据集或环境，在不影响管道源代码的情况下改变行为，并满足数据分析专业人员的日常需求。

运行：持续集成和部署

持续集成

数据管道工程师或所有者可以随时对管道进行更新或更改，并将其作为开发分支或私有分支中的私有副本保存在版本控制系统中。多个工程师可以并行工作，并同时将更改交付给开发部门或分公司，从而将生产效率提高数倍。当管道更改完成并在分支中进行测试后，可以将源代码合并到主代码库或主干中，然后交付给生产线。如果合并的代码行不通，数据管道总是可以退回到管道源代码的前一个有效版本。分支和合并功能允许数据分析团队运行自己的测试、进行更改、检验风险以及试验更改并放弃被证明不成功的更改。

持续部署

数据分析专业人士要求将管道使用的相关数据与源代码的私有副本和执行这些管道的环境分离开来。直接在生产数据库

或生产环境中工作往往效率不高，并且容易导致冲突。为了减少冲突和依赖，数据管道需要：

- 高效的源代码控制管理
- 灵活的环境部署选项可用性
- 测试数据行为

Jenkins Pipeline 是补充 IBM 交付管道的工具。Red Hat® OpenShift® 提供了一个可重复、一致的部署平台，可用于验证在运行时提供不同值的数据管道并发实例。

管理：使用一致的频繁部署

数据分析专业人员希望避免部署会破坏生产环境中当前数据管道的更改。两个关键工作流可以解决这种情况：

- **价值管道。**通过让数据流入生产环境，立即为组织创造价值。
- **创新管道。**指引正在开发中的新分析的未来方向，并将其添加到生产管道中。

这两个管道在生产中相交，使 DataOps 组织能够掌控生产数据的编排和新功能的部署，同时保持完美的质量。数据管道质量控制，例如监控数据和新开发管道的统计过程控制，使开发团队可以放心部署，而无需担心破坏生产系统。使用敏捷开发和 DevOps，最大程度地提高新分析的速度。它有助于最大程度地减少将业务需求转化为分析思想并将其发布为可重复和可重用的生产过程所需的时间和精力。

学习：通信和流程管理

高效和自动化的通知是 DataOps 方法中通信和补救流程的核心。当对任何源代码进行更改时，或者当管道被触发、失败、完成或部署时，都会发出通知。在发生故障时，故障信息可能与通知一起推送以帮助解决问题。修复后流程可以自动触发，以验证管道，将其部署到下一阶段，并使用最新信息和数据质量更新仪表板。Slack、Apache Kafka、PagerDuty 和 Trello 等工具通常作为 DataOps 工具链的一部分，用于支持利益相关者之间的通信、协作、反馈捕获和共享。

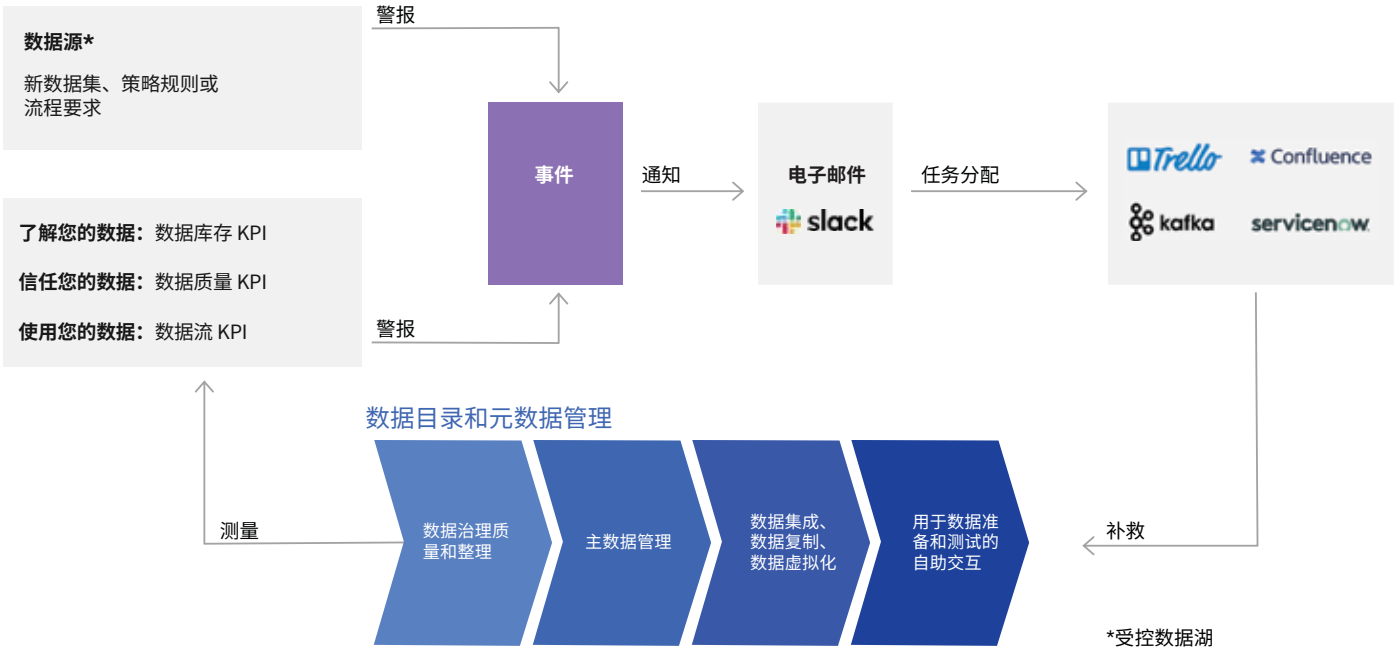


图 4：受控数据湖环境中的通信和流程管理可视化

成功 DataOps 实践的影响

通过采用 DataOps 实践，一家零售商对其各个数据管道进行了改进，使数据更改在整个组织中得到应用所需的时间不到两分钟，而以前需要三周时间。结果，零售商利用业务就绪数据在不到一天的时间内完成了客户亲和度分析，而这一过程以前需要 20 天。此外，报告库存状况所需的时间减少到了原来的六分之一。

成功 DataOps 实践的标准包括：

1. **建立数据办公室。**此流程包括：明确定义将数据作为资源提供给数据库的过程中各个角色的职责范围，确定关键执行干系人，以及了解数据管道中每个干系人对协作运营和文化的承诺。
2. **与业务目标保持一致。**为了保持竞争力，市场需要对新机会作出快速反应，而这只能通过以信息为导向和数据主导的方法来实现。简而言之，除非业务与数据交付之间有良好的沟通，否则领导层会认为其组织不会蓬勃发展。
3. **成功扩展数据。**在每一项数据主导的计划中，领导者需要确保所产生的数据可以使用和不断重用，并且每次使用都可以增加价值。只有数据被集中共享、可搜索并与业务语言保持一致时，才能实现此结果。

结论

成功部署 DataOps 的组织知道他们可以访问哪些数据资产，相信数据的含义和质量，并充分发挥其数据的潜力。当可信的业务就绪数据有助于推动差异化见解、卓越运营、协作和竞争优势时，数据就体现了价值。

建立 DataOps 实践需要：

- *通过运行试点项目来了解其组织的独特能力和挑战*
- *利用试点项目的成功来扩展和发展 DataOps 技能和组织*
- *招募更多的团队参与 DataOps 实践，以促进其成功*
- *分享经验教训并开始构建 DataOps CoE*

进行下一步，安排您自己的 IBM DataOps Garage 研讨会，并联系 dataops@us.ibm.com 以加快您的业务就绪数据获取之路。

如果组织已经在研究数据目录、数据湖或主数据计划，他们会发现采用 DataOps 正当其时。要了解有关 DataOps 支持市场领先技术的更多信息，请访问 **ibm.com/DataOps**。

1 2019 Global data management research: Taking control in the digital age.” Experian, 2019.

2 Jarah Euston, "The DataOps Trend is Real: 73% of Companies Plan to Invest in DataOps to Manage Data Teams in 2018," Nexla

10028810-CNZH-01

附录：DataOps 试点计划模板

项目名称：
日期：
部门或单位：
试点计划负责人：

姓名	职责	电子邮件	电话

企业范围内扩展的干系人：

姓名	职责	电子邮件	电话

问题陈述：

根本原因清单：

挑战	是否适用?是/否	附加说明

成功指标：

开始日期：

冲刺结束日期：

评估：

实施	当前状态	期望的冲刺结果	实现期望结果的行动步骤
数据资产提取、自动发现和分类			
数据质量评估和补救			
商业术语指定			
数据隐私、法规遵从和公司政策定义与执行			
数据使用者需求定义和请求处理			
数据请求通信和通知，包括异常和错误处理与补救			
将整理好的数据发布到目录			
数据沿袭和报告			
协作、反馈和审计			

实施审计期间的提问示例：

数据资产提取、自动发现和分类

- 我们是否执行高容量、低延迟的复制以支持业务连续性？
- 我们是否使用先进的流式分析方法进行实时、低延迟的分析？
- 我们是否可以轻松连接到任何数据源并执行复杂的数据转换和集成？
- 我们是否可以提供来自社交媒体、天气数据或其他公有云数据源的数据？
- 我们的数据使用者是否可以从任何桌面应用程序实时访问我们的元数据存储库？
- 我们的数据使用者是否可以实时访问我们的数据目录，以便在发现与其工作相关的数据集的过程中获得自助服务和帮助？
- 我们是否使用数据分析工具来理解数据、验证数据值、列和表关系以及查找和分析异常情况？
- 我们的业务规则管理是否与我们的元数据管理基础架构相融合？