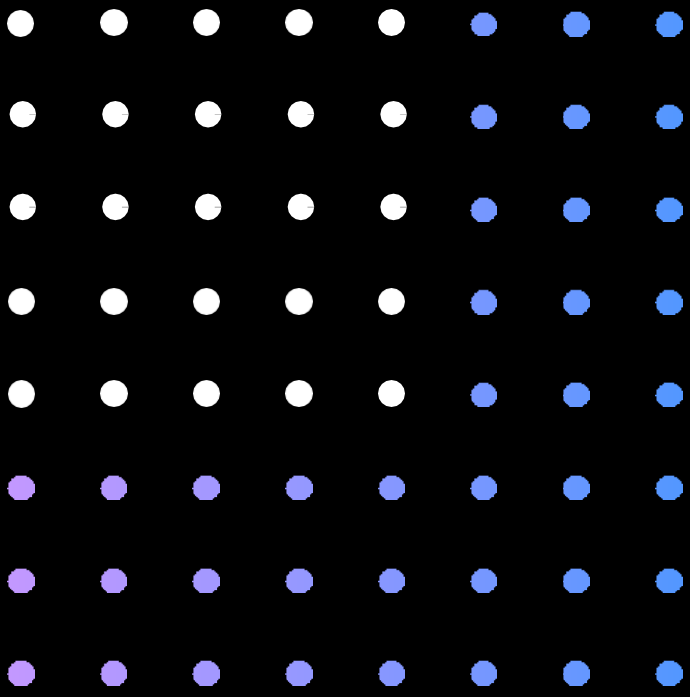
通过智能数据编目和数据湖治理，交付业务就绪数据

IBM Watson Knowledge Catalog 提供基于机器学习的数据治理平台，帮助应对数据湖挑战



IBM

目录

关键要点

- 企业构建数据湖的初衷本来是存储和分析数据，以便从中提取可信洞察，但是，真正能够获得该价值的企业却是寥寥无几。
- DataOps 能够帮助企业解决在访问、准备、整合以及向用户提供数据方面面临的大量低效挑战，同时遵守公司规定和政策法规。
- 常见的数据湖挑战包括：难以将新数据源导入数据湖且整个过程成本高昂，无法整合内外部数据集，对数据治理缺乏信心，无法使用自助式数据准备工具，以及无法找到和了解数据湖中的数据等。
- 具有编目、数据质量保证和数据发现功能的企业数据治理平台，可将失败的数据湖项目转变为真正的业务价值来源。
- IBM® Watson® Knowledge Catalog 基于 IBM Cloud Pak™ for Data，提供了用于数据发现、数据编目以及数据质量保证和数据治理的机器学习 (ML) 目录。它可以帮助数据用户快速发现、整理、分类和共享数据资产、数据集和分析模型。
- 如果企业缺乏对数据的深刻了解，那么便很难信任数据，更不用说与包括机器学习和深度学习在内的所有形式的人工智能 (AI) 一起使用这些数据。

通过 DataOps 方法应对数据湖挑战

十年前，企业开始寻找灵活的通用方法来构建能够存储所有企业数据的中央数据存储库。最终敲定了数据湖 – 一种几乎可以存储任何类型数据的通用数据存储环境。此外，数据湖还允许业务分析师和数据研究员在每个数据集的原始位置对其应用最适当的分析引擎和工具。

通常情况下，这些数据湖是使用 Apache Hadoop 和 Hadoop 分布式文件系统 (HDFS) 以及 Apache Hive 和 Apache Spark 等引擎构建的。伴随这些数据湖的增长，一系列问题逐渐浮出水面。尽管该项技术在物理上能够进行扩展，以捕获、存储和分析大量不同的结构化和非结构化数据集，但人们却很少关注如何将这些功能嵌入到业务工作流中这一实用性问题。

结果，诸如：“我们应该在数据湖中放入哪些数据？”，“谁将使用它们？”，“我们如何让用户轻松找到数据？”，“这些数据从何而来？”以及“我们如何防止数据被滥用？”等问题常常无法找到答案。在解决人员、流程和技术问题时，这些主要限制因素自然导致数据湖实施以失败告终。

今天，许多企业均已认识到了自己的失败，并且调整了数据湖实施领导团队，让数据运营 DataOps 团队担任领头人，正在第二、第三甚至第四次尝试成功实施数据湖。

本白皮书对数据湖实施项目面临的常见挑战进行了评估，并提供了诸如 DataOps 之类的新方法，旨在帮助企业将数据湖从数据沼泽转变为企业中业务就绪数据管道的核心。

DataOps 是一种协作式数据管理实践，致力于跨越整个企业来管理数据管理者和数据使用者之间的数据流，进而提高这些数据流之间的通信、集成和自动化水平。

DataOps 简介

DataOps 将 DevOps、数据管理和数据治理最佳实践引入一个通用框架中，允许协同开发和维护跨越多个利益相关方的数据流。DataOps 旨在帮助企业应对与数据访问、准备、整合和交付相关的低效挑战，同时遵守公司规定和政策法规。

业务部门、分析团队甚至运营流程均能从 DataOps 带来的增效作用中受益。

遵循这种方法需要解决决定数据湖实施成败的人员、流程和技术问题。从技术角度看，DataOps 注重使用完全集成的端到端平台来负责数据采集和整合、数据质量保证、数据治理和数据使用，从而创建受控数据湖。数据质量验证规则应作为采集过程的一部分自动运行，以便维持跨越整个企业的持续数据管道。采集流程应与作为数据管道核心的数据目录完全集成。数据使用者应能够从数据目录访问数据质量评分和数据分析结果，并相信企业在该等背景下使用相同的数据。

只有不到 29% 的企业开始从数据中获得价值 [1]，因此，立即采取行动获取竞争优势刻不容缓。现在，构建受控数据湖不仅仅是为了保护企业已在数据湖技术上投入的大量时间和资源，而且还因为别无其他选择。从实施 AI 到开展全面的分析，拥有尽量完整的数据视图至关重要，这意味着您需要一种能够在一个地方保存、分析并管理所有数据的架构。

很多情况下，受控数据湖都是满足这些需求的唯一现实选择。

当今的企业可以并且必须设法确保数据湖支持面向 DataOps 的业务就绪数据管道，以便从数据湖中提取价值

使用企业数据湖的挑战

共享数据

当企业中的某个团队获取或创建新的数据集时，很可能非常渴望了解这些数据的价值及其敏感性。例如，如果其中包含商业机密信息、个人身份信息 (PII) 或客户数据，则团队将知道应该如何使用这些信息，并将采取预防措施来确保团队中没有人会滥用它们。

他们还知道，其他潜在用户对于此类数据的价值以及相关滥用风险的理解可能有别于团队内部成员。考虑到这些风险，团队在共享数据或将数据存储在不受其控制的任何位置时都会非常谨慎。

这对数据湖来说是个坏消息。如果企业仅将数据湖视为不受控制的数据转储场，那么，他们将极不情愿将宝贵的数据存储在这里。那么，企业中的其他部门将无法从这些数据中获益，而将数据湖作为自助式存储库来共享企业数据的整个概念也会土崩瓦解。

整合数据

即使团队同意将其数据整合到数据湖中，整合过程也可能百般痛苦。数据湖的初衷是以原始格式导入数据，不像传统数据仓库那样需要复杂的提取、转换和加载 (ETL) 过程。但实际情况是，几乎所有的数据源都需要进行某种程度的预处理，然后才能用于任何有意义的分析。

因此，将新的数据源整合到数据湖中通常需要几个月的时间。此外，由于大部分的此类数据以前都保存在小型的运营孤岛而非企业级系统中，导致企业可能需要整合几十甚至数百个数据源。

这意味着在许多情况下，业务分析师或数据研究员所需的信息尚未添加到数据湖中，并且可能未来数月甚至数年都不会添加进来。
这也是阻止企业部署数据湖的重大障碍。

存储数据

过去几年中，虽然商品存储和计算资源的成本已大幅降低，但 Hadoop 集群并不是免费的。虽然将大量数据存储在数据湖中要比将其存储在高性能数据仓库设备中划算很多，但成本仍然很高。

此外，与传统上存储在数据仓库中的数据不同，存储在数据湖中的大数据的容量价值比相对较低。从数据湖中查找有价值的数据无异于大海捞针。

如果您不知道哪些数据集对您的数据研究员真正有用和有价值，那么，您可能会花费大量的资金来整合和存储毫无价值的数据 - 注定会沉入数据湖的底部、永远不会被使用的数据。

查找数据

假设您已经确定了需要存储的最有价值的数据集，已经说服了利益相关方共享这些数据，并且已将它们成功整合到您的数据湖中，此时，您仍需确保其他用户能够准确地找到、了解并使用它们。数据湖中的数据的质量是另一个挑战。您无法确定正被纳入湖中的数据的质量是高还是低。

使用企业数据湖的挑战

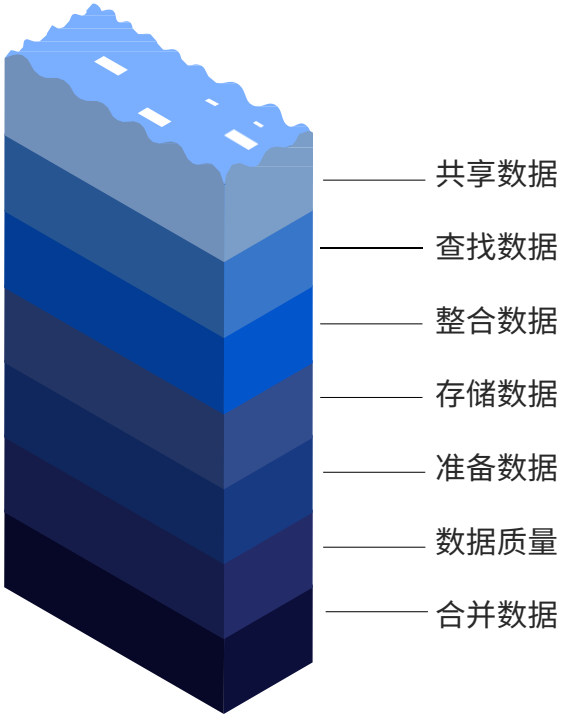


图 1 采用数据湖技术的企业可能会遇到一个或多个这些常见问题

遗憾的是，在大多数的数据湖中，了解数据质量都是一件难事。数据通常是在没有任何上下文的情况下存储的，这使得新用户很难或完全不可能在不咨询原始拥有者的情况下对其进行解码。术语经常是领域特定的，以致于企业中某个部门使用的度量标准可能会被另一个部门冠以完全不同的名称，或以不同的方式定义。对于不熟悉它们的分析师来说，造成混淆和误解的可能性非常之大，以至于许多数据集实际上毫无价值，甚至是危险的。

合并内外部数据

最后，即使是最大的数据湖也不应试图保留公司数据研究员想要使用的所有可能的数据集。例如，仅仅因为一位数据研究员希望执行地理空间分析，或者将天气数据或股票价格集成到算法中，就在数据湖中导入 Google Maps、Weather.com 或 Bloomberg 的完整副本是没有意义的。

由于您的数据湖无法容纳业务分析师开展分析所需的所有数据，因此，他们将不得不花费时间在多个应用中进行搜索。由于很大一部分有用的分析可能同时涉及到内外部数据集，这将再次增加进入障碍，从用户的角度看，这将降低数据湖的感知价值。

准备数据

许多因素导致数据准备工作面临挑战 - 从了解从何处查找数据到格式化数据。对数据用户而言，准备用于分析的数据是最低效、最耗时的任务。他们甚至将高达 80% 的时间耗费在查找、清理和格式化信息上，而不是专注于数据分析、建模以及从中获取业务影响洞察。[2]

受控数据集的有可限访问性也导致数据准备阶段过分依赖 IT。这种有限访问表明企业需要提高整个企业的自助服务能力和数据素养，以缓解这一障碍。

数据质量

将数据转储到数据湖可能导致其无法使用。由于在将数据纳入数据湖之前，没有对数据应用任何数据质量或验证规则，因此，数据湖无法提供可以信任和使用的数据。高质量的数据是决定数据决策可靠性的基本前提。数据是一项宝贵资产，必须在整个企业中进行数据管理。随着信息源的数量和种类日益增多，以及监管合规计划的针对性越来越强，通过一致的、可信赖的、可复用的方式来整合和访问不同来源的信息变得至关重要。

通过整体方法来构建受控数据湖

大多数的数据湖都将 Apache Hadoop 及其庞大的开源项目生态系统用作数据存储层和分析引擎。不出所料，围绕着 Hadoop 的开源社区已经认识到当前数据湖实施面临的问题，并且最近涌现出了旨在单独攻克各种问题的大量项目。同样，市场上还有许多旨在解决相同问题的专用工具。

因此，当您的数据湖出现问题时，您可能忍不住逐一进行补救。当数据集的数量上升到无法管理的程度时，您可能会添加一个编目工具。当用户抱怨无法找到他们所需的数据时，您可能会添加带有搜索功能的前端。当您的数据管理员不再能够跟踪数据出处或使用者时，您可能会部署数据沿袭工具和数据治理框架。

这听起来很简单，但在实践中，这种零散的方法往往以大幅增加复杂性和降低可维护性为代价，尤其是随着数据湖的规模和范围的增加。就像向数据湖中添加新的数据源会因增加 ETL 需求而增加复杂性一样，添加新工具会因增加非功能性需求而增加复杂性。

您通常会发现，每个工具都有自己的故障管理和日志记录方法，而不是使用集成的端到端平台来整合数据、对数据执行质量的操作以及对数据进行编目以供业务分析人员有效地使用。因此，故障排除和问题解决可能非常耗时。

当您以概念而非技术为重点来洞悉常见的数据湖问题时，这种零散方法的另一个更重要的缺陷就会暴露无遗。您会发现可扩展性、可查找性、整合性、数据质量和治理并不是独立的问题：它们之间有着千丝万缕的联系。解决这些问题需要采取更全面的方法。

可扩展性、可查找性、整合性、数据质量和治理并不是独立的问题：它们之间有着千丝万缕的联系。解决这些问题需要采取整体方法进行信息管理。

IBM Watson Knowledge Catalog 数据发现、数据编目和数据质量保证

基于 IBM Cloud Pak for Data 的 IBM Watson Knowledge Catalog 能够帮助数据用户快速发现、整理、分类并与其他同事共享数据资产、数据集、分析模型及其关系。它可以帮助数据治理团队定义业务词汇表、策略和规则，并为开展治理工作提供高级工作流。该目录为数据工程师、数据管理员、数据研究员和业务分析师提供了一个单一事实来源，使他们能够自助访问可以信任和放心使用的数据。

基于 IBM Cloud Pak for Data 的 IBM Watson Knowledge Catalog 等解决方案可在单一综合平台中提供解决当今主要数据湖问题所需的全部功能。该目录有助于从源头解决这些相互关联的问题：数据湖普遍无法提供有效的工具来捕获、存储和管理元数据并跟踪数据沿袭。

在许多方面，数据湖的价值都取决于它所包含的元数据，就像取决于数据本身一样。如果没有元数据来解释数据集的来源、创建者、包含的内容、谁有权使用它以及如何使用它，那么，数据本身几乎一文不值。用户将无法找到它，即使找到了，也无法理解其含义或对其给予足够的信任，或者不知道应该如何使用它。

Watson Knowledge Catalog
交付可信且有意义的数据

组织您的数据

了解

数据必须是完整的、适用的、可供随时随地访问的。发现、分类并了解所有类型的数据。

治理您的数据

信任

数据必须是安全的、整洁的、易于查找的，这样才能激发深受用户信任的自助访问。了解数据来源及其质量。

民主化您的数据

使用

数据必须能够推动自助式发现和自动决策，以推动业务发展。提供涵盖所有信息的全方位视图，并允许使用者访问该视图。

图 2IBM Watson Knowledge Catalog 针对数据发现、数据编目和数据治理提供了广泛的功能。

单一事实来源和单一访问点

基于 IBM Cloud Pak for Data 的 IBM Watson Knowledge Catalog 通过将元数据作为关键优先事项来解决这些问题。它的核心是强大的编目引擎，该引擎可对贵公司能够访问的所有数据集和分析资产进行索引，而无论数据位于何处，例如数据湖、数据仓库或事务系统，甚至一组电子表格，无论它们是结构化还是非结构化数据，也无论是存储在内部还是托管在云中。此外，该目录中还可以包括外部数据集和数据源，例如贵公司订阅的专有数据服务或开放数据 API。

除了针对您的所有数据集提供单一事实来源外，该数据目录还提供单一访问点。基于 AI 的搜索和建议功能可以帮助业务分析师、数据研究员、数据质量工程师和数据治理团队更轻松地找到资产，还能提供可用的元数据来帮助用户了解他们发现的内容并评估这些内容对他们是否有用。

内嵌的自助式数据准备功能可以加快数据转换速度，以便数据尽快在分析和 AI 应用中发挥有效的作用。业务分析师和数据研究员不必浪费时间来准备和分析数据。它还能与 IBM InfoSphere® Advanced Data Preparation 等企业级数据准备解决方案相集成，帮助确保通过该目录创建的受控数据集显示最丰富的上下文信息，从而为业务用户带来业务洞察并推动他们做出明智的行动。这种集成进一步促进了跨越数据管道的协作。

> 可扩展性、可查找性、整合性、数据质量和治理并不是独立的问题：它们之间有着千丝万缕的联系。解决这些问题需要采取整体方法进行信息管理。

该目录还能标记和分类数据集，自动跟踪数据的沿袭和使用情况，并利用内置的业务术语表来跨越所有的数据规范业务术语，从而为首席数据官 (CDO) 办公室中的数据管理员提供帮助。因此，管理员可以更轻松地了解每个数据集中包含的内容、敏感数据或 PII 的位置以及谁应有访问权限。

覆盖企业内外部多个数据源的单一目录

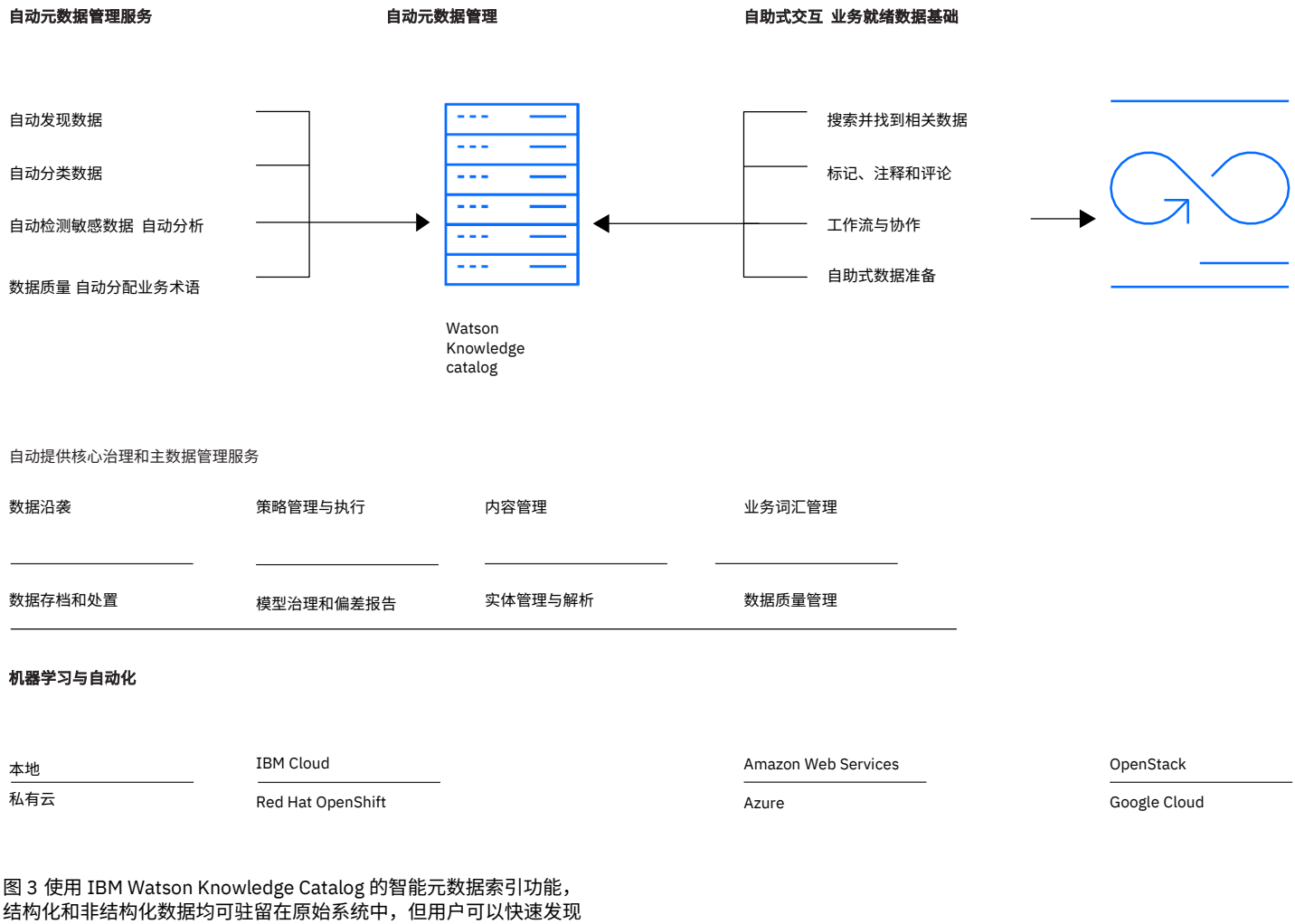


图 3 使用 IBM Watson Knowledge Catalog 的智能元数据索引功能，结构化和非结构化数据均可驻留在原始系统中，但用户可以快速发现它们，以便开展更智能的分析。

IBM Watson Knowledge Catalog 将元数据作为关键优先事项，针对贵公司可以访问的所有数据集提供单一事实来源和单一访问点

内置智能数据发现功能

为了进一步提高可查找性，该目录允许用户对数据集和分析资产进行标记和注释、丰富元数据并添加额外的上下文，以帮助同事找到他们需要的内容。该解决方案还包括内置的数据发现算法，使用机器学习自动对每个数据集的内容进行分类。通过识别常见的字段类型，如姓名、地址、邮编和社会安全号，该解决方案降低了作者手动注释数据的需求。它融合了自动化和机器学习技术，用于自动执行数据编策和元数据管理任务。借助内置的数据质量保证功能，该解决方案支持深度数据分析、数据质量保证和验证规则。自动数据运营可提供具有数据质量保证和治理功能的精心策划的数据管道，并有助于确保高质量的受控数据连续不断地流入数据湖。

您还可以通过类似方式为您的资产添加智能元数据模型，从而通过独特的方式来自动执行《通用数据保护条例》(GDPR) 和《加州消费者隐私法》(CCPA) 合规等任务。

基于 IBM Cloud Pak for Data 的 IBM Watson Knowledge Catalog 能够帮助您向几乎所有的数据用户交付可信赖的、高质量的业务就绪数据。

该解决方案的所有组件均作为微服务进行设计，遵循同一套设计原则，并使用通用方法来满足可扩展性、错误管理、安全性和日志记录等非功能性需求。

IBM Watson Knowledge Catalog 提供机器学习企业治理平台，已为大规模 AI 部署准备就绪。

IBM Watson Knowledge Catalog 提供机器学习企业治理平台，已为大规模 AI 部署准备就绪，不会像自己动手的零星方法那样产生令人困惑的错误和性能瓶颈。

IBM Watson Knowledge Catalog 有三种变体：

- 作为 IBM Cloud™ 中的软件即服务 (SaaS) 解决方案
- 作为 IBM Cloud Pak for Data 中的核心产品
- 与 IBM Watson Studio 相集成

诸如 IBM Watson Knowledge Catalog 之类的解决方案可以释放数据湖计划最初承诺的价值。具有智能编目和治理功能的 Watson Knowledge Catalog 有助于面向 AI 构建可信的受控数据湖。

面向 AI 构建受控数据湖的四个好处

1. 通过质量保证和治理来建立对数据的信任和信心

- 数据质量保证功能可帮助您提高数据湖中的数据质量，并实现高质量数据的可用性。
- 治理策略是自动设置和强制执行的，因此，每当您发现一个数据集时，便知道能否以及如何使用它。
- 就像用户添加评分、评论和其他信息来帮助他人决定某个数据集对他们是否有用一样，您也可以编策您的数据。

2. 为您的数据用户赋能

- 您的业务部门 (LOB) 工作团队愿意共享他们的数据，因为他们相信数据能够得到适当的治理和保护，不会被滥用。
- 您可以通过动态数据策略和实施来推动协作并将数据转换为可信赖的企业资产。
- 随着用户不断添加相关标签和元数据来帮助其他人从中提取价值，您的数据将变得越来越容易被发现和复用。
- 您可以通过单一界面访问企业拥有的每个数据集，无论它们存储在何处。

3. 找回被浪费的时间

- 自动数据发现功能可以帮助您减少为新数据集添加元数据所需的时间和精力。
- 自动数据编策和元数据管理功能可以帮助您减少发现元数据和分配术语所需的时间，还能减少创建业务词汇表所需的时间。
- 通过简单直观的自助式数据准备工具，您的数据用户将花费更少的时间来准备数据，从而腾出更多的时间用来获取洞察。
- 您可以将数据研究员和业务分析师从繁琐的工作中解放出来，让他们能在更短的时间内提供更准确的分析结果。
- 基于 AI 的智能搜索功能可以帮助您在几秒钟内找到所需的数据，而不是浪费数周的时间来等待另一个团队提供这些数据。

4. 管理不断增长的数据和成本

- 您可以避免将低价值数据集存入数据湖造成的开销，从而优化存储成本。
- 您还可以查看企业订阅的所有外部数据集，从而降低因订阅多余数据集而支付额外费用的风险。
- 您可以根据用户数据需求为新数据源进驻数据湖分配优先级，从而帮助您首先整合最有价值的数据源。

解锁数据的价值

无论您是 CDO 办公室和 IT 部门的工作人员，还是 LOB 数据研究员或分析师，您和您的同事都有一个共同的目标。如果您可以构建一个真正能够兑现承诺的数据湖，那么，您不仅可以让自己的工作变得更加轻松和高效，而且还能在帮助企业获得竞争优势方面发挥关键作用，成为同行业的佼佼者。

如果您可以在竞争对手仍在泥沼中挣扎的时候清理您的数据湖，定将能够创造令其望尘莫及的奇迹。真正的先发优势属于那些能够率先探索未知数据并从中解锁价值的人。

结语

您应知道所有数据的存放位置、使用者及其业务分析价值。

数据目录是 DataOps 计划取得成功的关键，因为它们可将数据治理、质量保证和主动策略管理集成在一起，帮助实现自动的、开放的元数据管理。

具有智能编目和治理功能的 IBM Watson Knowledge Catalog 能够帮助您面向 AI 建立可信的受控数据湖。该目录可将数据整合、数据质量保证和数据治理功能统统嵌入到您的数据湖环境中，从而为 DataOps 提供业务就绪数据及单一事实来源。

更多信息

了解更多信息，请访问：
ibm.com/cloud/watson-knowledge-catalog

1. 洞察系统通过人员、流程和技术推动业务行动，Forrester Research，2016 年
2. 调查显示，清理大数据是最耗时、最让人头疼的数据科学任务。Forbes，2016 年 3 月 23 日