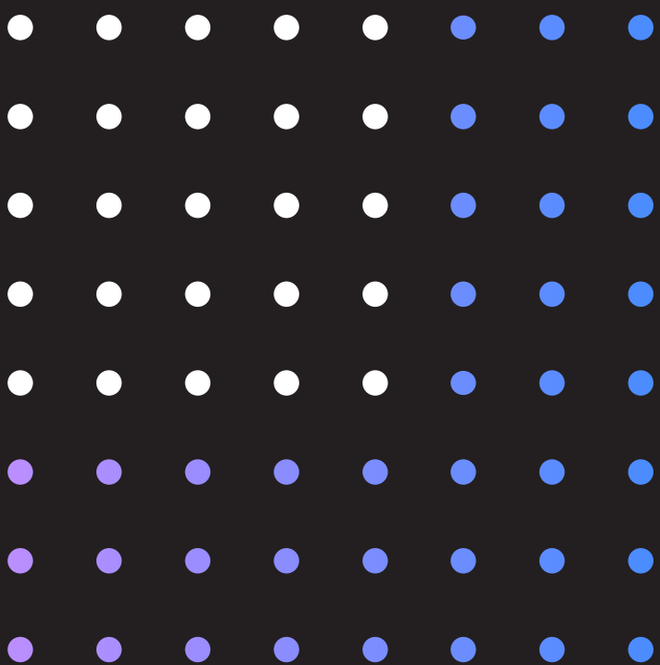
通过智能数据编目和数据湖治理提供直接用于业务的数据

IBM Watson Knowledge Catalog 提供一个机器学习驱动的数据治理平台，有助于解决数据湖挑战



IBM

目录

要点

- 许多组织建立了数据湖，用于存储和分析数据以获取可靠的见解，但很少有组织能够从其数据湖中看到所期望的价值。
- DataOps 可解决组织在访问、准备、集成和向客户提供数据方面效率低下的挑战，同时遵循公司和监管政策。
- 常见的数据湖挑战包括将新数据源导入数据湖的难度和成本；无法集成内部和外部数据集；对数据治理缺乏信心；无法使用自助数据准备工具；以及无法找到和理解数据湖中的数据。
- 具有编目、数据质量管理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 由 IBM Cloud Pak™ for Data 驱动，提供用于数据发现、数据编目、数据质量和治理的机器学习 (ML) 目录。它可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。
- 如果组织对其数据缺乏深入了解，将很难信任这些信息并将其用于各种形式的人工智能 (AI)，包括 ML 和深度学习。

使用 DataOps 方法解决数据湖挑战

十年前，许多企业开始寻找一种灵活、通用的方法来构建可存储所有企业数据的中央数据存储库。解决方案就是数据湖 - 一种通用的数据存储环境，几乎可以存储任何类型的数据。它还允许业务分析师和数据科学家将最合适的分析引擎和工具应用于每个数据集（在其原始位置）。

通常，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。随着这些数据湖开始增长，一系列问题浮出水面。尽管该技术在物理上能够扩展以捕获、存储和分析大量不同的结构化和非结构化数据集合，但对于如何将这些功能嵌入业务工作流以提高实用价值，却很少得到关注。

到 2022 年，80% 以上的数据湖项目将无法实现价值，因为发现、清点和整理数据将成为分析和数据科学取得成功的最大障碍。[1]因此，诸如以下问题往往难以解答：“我们应该在数据湖中放入哪些数据？”、“谁将使用它？”、“如何让它易于查找？”、“这些数据来自何处？”以及“如何防止数据被滥用？”。这些在有效解决人员、流程和技术问题方面的严重限制，导致了数据湖实施的失败。

如今，许多组织已经意识到项目失败，于是变更了负责数据湖实施的领导团队，并且正在发起第二次、第三次甚至第四次成功实施数据湖的尝试 - 这次是以数据运作 DataOps 为主导的尝试。

本白皮书将评价数据湖面临的常见挑战，并提供诸如 DataOps 之类的新方法，以帮助组织将数据湖从数据沼泽转变为适合业务数据通道的核心资产。

DataOps 是一种协作式数据管理实践，致力于改善整个组织中数据管理者与数据使用者之间的数据流通信、集成和自动化。

DataOps 简介

DataOps 将 DevOps、数据管理和数据治理的最佳实践植入公共框架，以协作的方式开发并保持多个利益相关方之间的数据流。DataOps 旨在解决组织在遵照企业和监管政策访问、准备、集成数据并向用户提供数据时效率低下的挑战。这些效率就存在于业务部门、分析团队甚至运营流程中。

根据此方法，数据湖实施能否成功，取决于能否解决人员、流程和技术方面的问题。从技术方面看，DataOps 强调必须使用完全集成的端到端平台来吸入和集成数据、管理数据质量、治理数据和控制数据使用，创建一个有效治理的数据湖。数据质量验证规则应作为数据吸入过程的一部分自动运行，以维持整个企业的连续数据管道。数据吸入过程应与数据目录完全集成，成为数据管道的核心。数据用户应该能够查看数据质量分数和数据目录的数据归档结果，并且信任组织在环境下使用相同的数据。

数据的增长超过组织从中获取价值的能力。当问到组织在使用洞察力系统时遇到的最大挑战是什么时，他们的回应如下：1) 40% 是合并现有业务流程来获取数据进行分析，2) 39% 是获取、收集、管理和治理不断增长的数据。[2] 今天，不仅要保护在数据湖技术上付出的重大时间和资源投资，更重要的是，数据湖没有替代解决方案。从实施 AI 到执行全面的分析，全面了解尽可能多的数据非常关键，这意味着需要能够在一个位置保存、分析和治理所有数据的架构。在许多情况下，经过治理的数据湖是满足这些要求的唯一现实选择。

当今的企业能够 - 并且必须 - 找到一种方法，通过确保其数据湖支持 DataOps 的适合业务数据通道来从数据湖挖取价值。

使用企业数据湖面临的挑战

共享数据
企业内的团队在获取或创建新的数据集时，很可能对数据的价值及其相关敏感性有强烈的意识。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，则团队会了解这些信息的正确使用方式，并采取预防措施以确保团队中没有人滥用它们。

他们还会意识到，团队外部的其他潜在数据用户可能对数据的价值或滥用数据的相关风险没有同等的理解。这些风险自然会使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极为谨慎。

但这对于数据湖来说不是件好事。如果企业将数据湖看成只是一个不受控制的数据堆放场，他们就会很不愿意将其有价值的数据交给它。造成的结果是，企业的其他部门将无法从这些数据中受益，将数据湖用作共享企业数据自助仓库的整体概念也就崩塌了。

集成数据
即使团队同意将其数据集成到数据湖中，也可能是一个痛苦的过程。数据湖的初衷是以原始格式导入数据，而不需要传统数据仓库复杂的提取、转换和加载 (ETL) 过程。但现实情况是，几乎所有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。

因此将新的数据源集成到数据湖中往往需要几个月的时间。并且，由于许多数据以前存储在各个小型操作孤岛中，而不是在企业系统中，因此可能总共要集成数十个甚至数百个数据来源。

这意味着在许多情况下，业务分析师或数据科学家所需的信息尚未添加到数据湖中，并且可能几个月甚至几年都没有添加。同样，这可能也是影响数据湖采用的重大障碍。

存储数据
尽管在过去几年中，商品存储和计算资源的成本已大大降低，但是 Hadoop 集群并非免费的。将大量数据存储在数据湖中，比存储在高性能数据仓库设备中具有更高的成本效益，但成本仍然不菲。

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价值/体积比相对较低。您可能需要存储海量数据，才能找到几条高价值信息，如同大海捞针。

如果您不知道哪些数据集是对您的数据科学家真正有用和有价值的，那么您可能会投入大量资金来集成和存储那些注定要沉入数据湖底而永无出水之日的数据。

查找数据
即使您已经确定需要存储的最重要数据集，已经说服利益相关人分享它们，并且已经成功地将它们集成到您的数据湖中，您仍然需要将其组织成供其他用户查找、理解并正确使用。数据湖中数据的质量是另一项挑战。您不确定数据质量是高还是低，只是将其馈入数据湖中。

使用企业数据湖面临的挑战

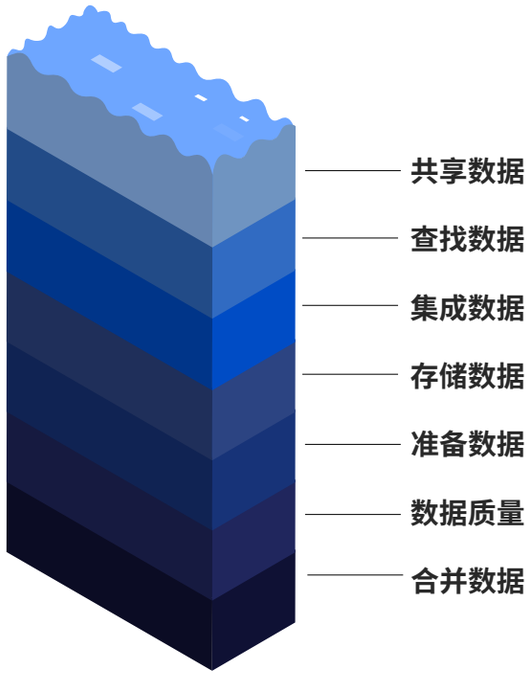


图 1已经采用数据湖技术的企业可能遇到其中一个或多个共同的问题。

遗憾的是，在大多数的数据湖中，质量控制目标并不容易实现。数据在存储时通常没有上下文，因此新用户若不咨询数据的原始拥有者，就难以甚至无法解码数据。术语通常是各领域特定的，因此在一个业务领域使用的指标，在另一个领域可能是完全不同的名称，或者定义方式略有差别。对于不熟悉它们的分析员来说，很可能会产生混淆和误解，这样很多数据就没什么价值，甚至带来危险的后果。

合并内部和外部数据
最后一点，不要妄想保存公司数据科学家需要的全部数据，即使是最大的数据湖也不能做到。例如，如果只是因为数据科学家想执行地理空间分析，或者想将天气数据或股票价格集成到某种算法中，就在公司数据湖中导入 Google Maps、Weather.com® 或 Bloomberg 的完整复制品，这种做法毫无意义。

由于数据湖没有业务分析师需要的所有数据，因此他们只能去多个应用程序中搜索。在有用的分析中，很大比例涉及到内外数据的组合，从而进一步加大了数据治理的难度，从用户的角度来看，则降低了数据湖的感知价值。

准备数据
数据准备的挑战来自多个方面 - 从了解数据查找位置到格式化数据。对于数据用户来说，准备用于分析的数据是一项最低效、最耗时的任务。数据用户的大多数时间用于查找、清理和格式化信息，而不是数据分析、建模和推导影响业务的洞察力。

对治理的数据集访问受限，也导致在数据准备阶段过多依赖 IT。访问受限意味着需要改进整个企业的自助服务功能和数据认知技能，以缓解此难题。

数据质量
将数据一股脑倒入数据湖没有用。在数据馈入数据湖之前不实施数据质量或验证规则，就无法提供值得信赖和使用的数据。在确定决策所用数据的可靠性时，高质量数据是必不可少的特性。数据是宝贵的资产，必须妥善管理其在组织中的流动。随着信息来源越来越多，越来越杂，监管越来越严，必须以一致、可信、可重复使用的方式集成和访问这些分散来源的信息。

以整体性方法构建治理的数据湖

大部分数据湖的数据存储层和分析引擎采用 Apache Hadoop 及其广泛的开源项目生态系统。不出所料，Hadoop 开源社区已经认识到数据湖实施当前面临的问题，最近也兴起了许多旨在分别解决各种问题的项目。类似地，市场上有许多用于解决这些问题的专有工具，

试图逐一解决数据湖产生的问题。当数据量增长太快而难以管理时，加入编目工具。当用户抱怨找不到需要的数据时，用搜索功能锁定前端。当数据管家不再跟踪数据来源或使用者时，部署数据沿袭工具和数据治理框架。

听起来简单，但真正实践殊为不易，这种零碎的方法通常很复杂，而且管理难度大，特别是在数据湖的规模和范围不断扩大时。同样，新的数据源加入数据湖也会增大 ETL 要求的复杂性，新工具的加入则会增大数据湖非功能性要求的复杂性。

如果不使用能够集成数据的集成式端到端平台，而是由业务分析师集成数据、对数据执行质量操作以及对数据编目，您通常会发现，每个工具都有自己的失败管理方法和日志记录方法。结果，排除故障和解决问题非常耗时。

另外，当您不是技术角度、而更多从概念角度来看数据湖常见的问题时，会暴露零碎方法更大的缺陷。主要观点是：可扩展性、可检索性、集成、数据质量和治理不是孤立的问题，而是紧密相关的。解决它们需要更加整体性的方法。

可扩展性、可检索性、集成、数据质量和治理不是孤立的问题，而是紧密相关的。解决它们需要整体性的信息管理方法。

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

IBM Watson Knowledge Catalog 由 IBM Cloud Pak for Data 支持，可帮助数据用户快速发现、组织、分类及共享数据资产、数据集、分析模型及其与组织其他成员的关系。它可帮助数据治理团队定义业务词汇、政策和规则，为治理提供高级工作流程。目录可用作单一数据源，供数据工程师、数据管家、数据科学家及业务分析师自助访问他们信赖和放心的数据。

诸如 IBM Cloud Pak for Data 支持的 IBM Watson Knowledge Catalog 这样的解决方案，可以提供所需的全部功能，在单一、全面的平台上解决当今数据湖的主要问题。该目录可帮助解决这些相关问题的根源：数据湖广泛失效，无法提供有效的工具来获取、存储及管理元数据以及跟踪数据的关联。

在许多方面，数据湖的价值取决于其包含的元数据，元数据的重要性一点也不亚于数据本身。如果没有元数据来说明数据来自哪里、由谁创建、包含什么、谁可以使用以及是如何使用的，那么数据本身无实际价值。用户找不到数据，即使找到了，也不能理解它、信任它或知道如何使用它。

Watson Knowledge Catalog
提供可信且有意义的数据

组织数据

了解

数据必须完整、适用且可随地访问。发现、分类并理解所有类型的数据。

治理数据

信任

数据必须安全、清洁且易于查找，以促进可信的自助访问。了解数据来源及其质量。

普及数据

使用

能够推动自助发现和自动化决策以发展业务。向需要信息的用户提供所有信息并允许他们访问。

图 2IBM Watson Knowledge Catalog 提供广泛的数据发现、数据编目和数据治理等功能。

单一数据源和单一访问点

IBM Cloud Pak for Data 支持的 IBM Watson Knowledge Catalog 将元数据视为重中之重，着力解决这些问题。其核心是强大的编目引擎，对您可以访问的所有数据集和分析资产编制索引，无论数据是位于数据湖、数据仓库或交易系统中，还是位于一系列电子表格中。无论它们是结构化数据还是非结构化数据，是存储在本地还是托管于云中。此外，目录还可包含外部数据和来源，例如公司订阅的专有数据服务，或者开放式数据 API。

就像为所有数据提供单一数据源一样，数据目录还提供单一访问点。AI 支持的搜索和建议功能帮助业务分析师、数据科学家、数据质量工程师及数据治理团队更轻松地查找资产，同时提供可用的元数据帮助用户理解找到的数据，并评估对他们是否有用。

嵌入式自助数据准备功能可加速数据的转换，以提高在分析和 AI 应用中的使用效率。业务分析师和数据科学家不必将时间浪费在数据的准备和分析上。集成企业数据准备解决方案，如 IBM® InfoSphere® Advanced Data Preparation，帮助确保通过目录创建、经过治理的数据显示给最相关的用户，以促进业务洞察力和业务用户的行动。此集成可增强跨数据管道的协作。

可扩展性、可检索性、集成、数据质量和治理不是孤立的问题，而是紧密相关的。解决它们需要整体性的信息管理方法。

目录还可帮助首席数据官 (CDO) 办公室的数据管家对数据标记和分类，自动跟踪其关联和使用，并且利用内置业务术语表标准化不同数据之间的业务术语。因此，管家更容易了解每个数据集包含什么内容，敏感数据或 PII 在哪里，应允许谁访问数据。

涵盖组织内外多个数据源的统一目录

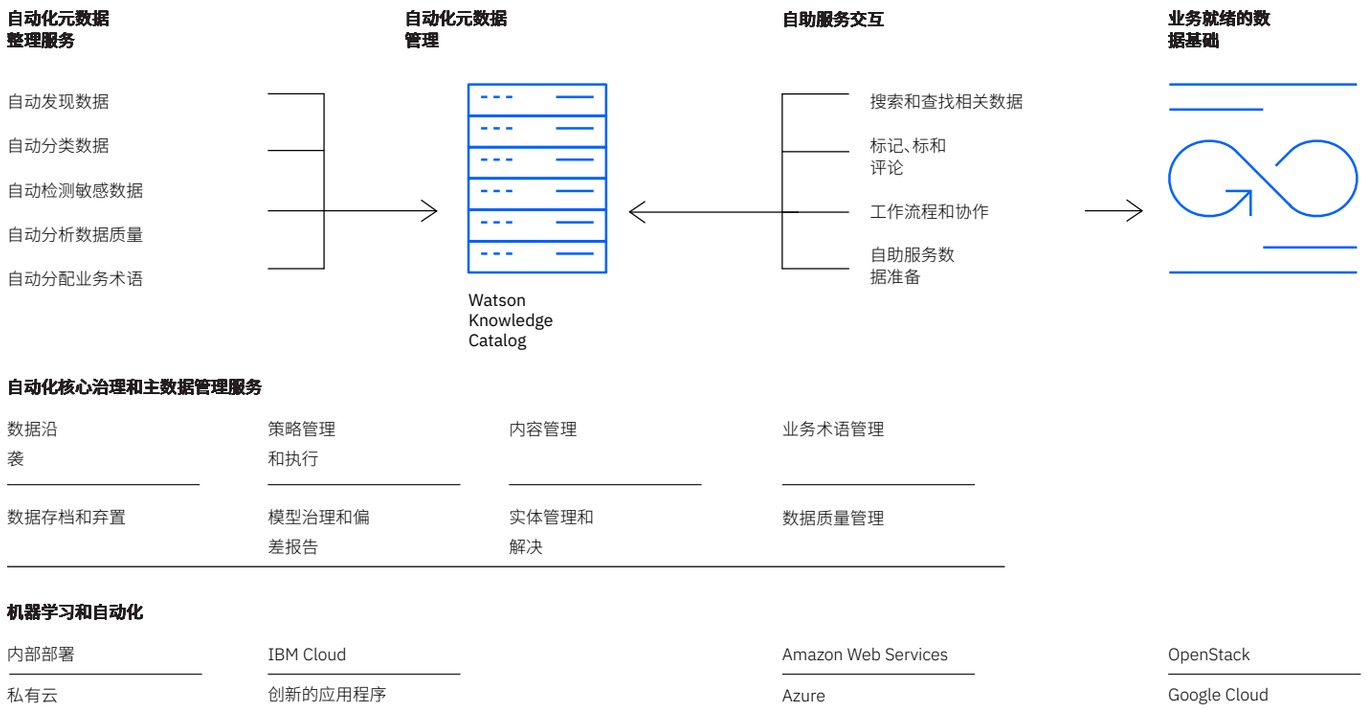


图 3利用 IBM Watson Knowledge Catalog 智能元数据索引，结构化和非结构化数据都可置于原系统中，但用户可以快速找到它们进行更智能的分析。

IBM Watson Knowledge Catalog 将元数据视为重中之重，为企业可以访问的所有数据集提供单一数据源和单一访问点。

内置智能数据发现

为进一步改进搜索性，目录允许用户标记并加注数据集和分析资产，丰富元数据并添加额外上下文，来帮助同事查找需要的数据。解决方案还内置发现算法，利用 ML 自动对每个数据集的内容分类。通过识别公共字段类型，如姓名、地址、邮编和社会保障号，解决方案可减少作者手动标注数据的工作。它融合自动化和 ML，自动进行数据整理和元数据管理。通过内置的数据质量功能支持深度数据归档、数据质量和验证规则。

自动化数据操作提供整理的数据管道，包含数据质量控制和治理功能，帮助确保经过治理的高质量数据连续流入数据湖。

类似地，在添加资产的智能元数据模型后，能以独特的方式自动遵守法律，如《通用数据保护条例》(General Data Protection Regulation, GDPR) 和《加利福尼亚消费者隐私法》(California Consumer Privacy Act, CCPA)。

IBM Cloud Pak for Data 支持的 IBM Watson Knowledge Catalog 向几乎所有数据用户提供可信、优质、可直接用于业务的数据。

解决方案的所有部分都已设计成微服务，采用一组同样的设计原则和共同方法满足非功能要求，如扩展性、错误管理、安全性和日志记录。

IBM Watson Knowledge Catalog 提供 ML 企业治理平台 - 可直接大规模用于 AI。

IBM Watson Knowledge Catalog 不像自成一套的零碎方法那样出现混淆错误和性能瓶颈，而是提供 ML 企业治理平台，可直接大规模用于 AI。

IBM Watson Knowledge Catalog 以三种形式提供：

- IBM Cloud™ 中的软件即服务 (SaaS)
- 内置于 IBM Cloud Pak for Data
- 与 IBM Watson Studio 集成

像 IBM Watson Knowledge Catalog 这样的解决方案可以实现数据湖计划最初承诺的价值。内置智能编目和治理功能的 Watson Knowledge Catalog 可帮助为 AI 构建可信和经过治理的数据湖。

为 IA 构建经过治理的数据湖有四大优点

1.通过质量和治理建立对数据的信任和信心

- 数据质量功能帮助改进数据质量，在数据库中注入高质量的数据。
- 自动设定并实施治理政策 - 这样在查找数据集时，就会清楚您能否使用数据以及如何使用。
- 您可以将数据整理为用户添加评分、评论及其他信息，以帮助其他人确定数据对其是否有用。

2.支持数据用户

- 业务 (LOB) 团队很愿意分享数据，因为他们相信数据会得到正确的治理和防滥用保护。
- 您可以推动团队协作，通过动态的数据政策和执行将数据转变为可信的企业资产。
- 随着时间的推移，您的数据会变得更加容易搜索和重复利用，因为用户会添加相关的标记和元数据，帮助其他人从中获取价值。
- 从单一界面访问组织拥有的所有数据，无论其存储在何地。

3.节省时间

- 自动数据发现可减少为新数据集添加元数据的工作。
- 自动数据整理和元数据管理可减少发现元数据和分配术语的时间，同时减少业务术语创建时间。
- 通过简便、直观的自助数据准备工具，数据用户在准备数据用时更少，而有更多时间探寻洞察力。
- 数据科学家和业务分析师获得了解放，能在更短的时间内提供更好的分析。
- AI 支持的智能搜索可帮助您在数秒内找到所需的数据，而无需等待另一个团队在几周后提供数据。

4.管理不断扩展的数据和成本

- 您可以优化存储成本，消除向数据湖注入低价值数据所产生的费用。
- 您还可以查看组织订阅的所有外部数据，避免订阅超过需求。
- 您可以根据用户对数据的需求排定新数据源注入数据湖的优先级，帮助优先集成最有价值的数据源。

释放数据的价值

无论是在 CDO 办公室、IT 部门工作还是作为 LOB 数据科学家或分析师，您与同事都有一个共同的目标。如果您能构建一个真正兑现承诺的数据湖，不仅能够提高您的工作简便性和效率，也会大幅提供企业的竞争力，打造无可匹敌的竞争优势。

如果您的数据湖清澈明净，而竞争对手仍在泥淖中挣扎，您就会打开更多可能性，而竞争对手只能梦想。真正的先发优势等待着第一个从以前未开发的数据中挖掘价值的人。

结论

了解所有数据在哪里、谁在使用它、它对您的业务分析有何价值。

DataOps 计划的关键在于数据目录，因为它们可以集成数据治理、质量和主动式策略管理，提供自动化的开放式元数据管理。

内置智能编目和治理功能的 Watson Knowledge Catalog 可帮助为 AI 构建可信和治理的数据湖。目录将数据集成、数据质量和治理嵌入数据湖环境，帮助为 DataOps 提供可直接用于业务的数据 - 以及单一数据源。

了解更多信息

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

IBM Corporation
New Orchard Road
Armonk,NY 10504

美国印刷，2019 年 10 月 IBM、IBM 徽标、**ibm.com**、IBM Cloud、IBM Cloud Pak、IBM Watson 和 InfoSphere 是 International Business Machines Corp. 在世界多个司法管辖区注册的商标。

1.Augmented Data Catalogs:Now an Enterprise Must-Have for Data and Analytics
（增强的数据目录：当今企业数据和分析必备）
Leaders–Gartner，2019 年 9 月
2.The Forrester Wave:Machine Learning Data Catalogs
（The Forrester Wave：机器学习数据目录），2018 年第 2 季度

ASW12449-CNZH-03