IBM Unified Governance & Integration

白皮书

认知型数据治理

借助机器学习技术，查找并使用可管控的数据

Jo Ramos

IBM Analytics 杰出工程师兼总监

Rakesh Ranjan

IBM Analytics 项目总监兼数据科学家



目录

简介

本白皮书旨在讨论当企业使用 IBM Unified Governance & Integration 平台实施元数据发现和业务术语分配工作时，机器学习和深度学习技术如何发挥作用、产生影响。

我们先介绍一组定义，确保大家对该主题形成一致的理解：

定义

主数据：描述企业核心实体的一组一致而统一的标识符和扩展属性，实体可包括现有或潜在客户、产品、服务、员工、供应商、提供商、层次结构和会计科目表等

机器学习：一种人工智能 (AI) 应用，使系统无需显式编程，就能够自动学习，随着经验的积累而不断改进

数据治理：针对企业中数据的可用性、相关性、易用性、完整性和安全性的整体管理

合规性：企业遵守与其业务相关的法律、法规、准则和规范

大多数企业都花费大量的时间和精力，处理杂乱无章、疏于整合的数据。他们的员工要么无法找到合适的数据，要么不信任所找到的数据。最重要的是，各式各样的行业法规制约了自助服务和数据民主化进程。因此，企业尝试通过各种劳动密集型任务（包括编写自定义的程序，开发全局替换功能等）以修复数据，这严重影响数据分析师和数据科学家的生产力。

大企业尤其如此，多年的并购云集了各色系统和数据库，导致数据环境极其复杂。虽然维护这些遗留数据环境已令企业感到疲惫不堪，但新数据仍以无法想象的速度不断产生。为解决这一问题，某些企业开始尝试使用主数据管理工具，通过统一各类数据源，形成关键业务实体的单一视图。

一些供应商提供的工具借助基于规则的引擎将各种数据源统一到他们的产品中，以解决这个问题。虽然这些规则很容易实施和理解，但基于规则的引擎的可扩展性并不是很理想。因此，大型企业面对海量数据的处理和各种不同系统数据交互，纷纷开始使用机器学习技术取代规则引擎。

事实证明，机器学习非常强大，能够帮助企业实现各种分析目标，例如预测客户流失，或检测在线信用卡交易中的欺诈行为。虽然确定数据相似性或者统一数据可能并不是机器学习最令人兴奋的应用领域，但对 IBM 客户而言却是最有益、最具财务价值的功能之一。

管理主数据的好处

构建数据目录可能非常耗时耗力，这也正是如此众多的企业放弃创建和更新组织良好的数据目录的原因。他们还面临其他挑战，例如：

- 实现业务定义标准化，建立业务术语表
- 对所有数据源编目，通过明确的业务描述进行更新
- 将业务术语和所有数据源中的数据字段对接

构建强大数据目录并不只是需要时间。为持续执行这项任务而聘请领域专家的费用也极其高昂。而人工智能和机器学习技术可在这个领域一展身手。IBM Unified Governance & Integration 可借助机器学习和神经网络，确定多条数据记录与同一实体的匹配概率，即使它们看起来并不相同。这可以帮助 IBM 客户分析主数据的数据质量以及业务术语关系，轻松化解他们在这方面的难题。过去需要几个月才能完成的项目现在只需几周便可完成。

虽然机器学习可自动执行任务，但执行过程中总是需要人类介入，就像任何其他人工智能或机器学习应用一样。通过反馈学习，如果匹配的置信度得分低于某个阈值，系统就会使用该工作流程，将候选数据记录提交给人类专家做出判断。这些专家只需处理整个数据集中具有弱匹配特征的一个小子集即可，从而大大提高工作效率。

这项活动带来的效益是巨大的 — 对任何组织的数据管理员和数据科学家来说都是如此。例如，一位新入职的数据科学家接受了开发机器学习模型的任务，用于检测特定产品或服务的客户流失情况。虽然此人知道自己需要完成的目标，但并不知道可以使用哪些数据集来开始任务。借助基于机器学习的 IBM 数据治理技术，这位数据科学家就能够轻松搜索“客户保留”等业务术语，从而获得所有相关实体的图形视图。然后，就可通过深入分析来了解数据的质量和真实性。

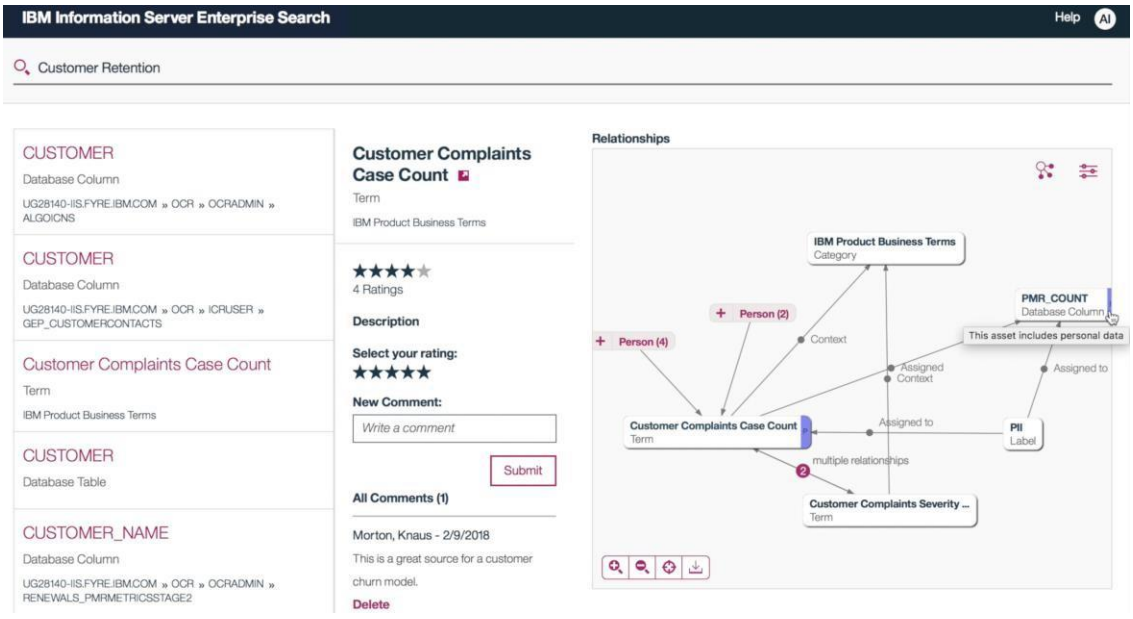

图 1：IBM Information Server Enterprise Search 显示了业务术语和技术元数据关系

上图显示了使用机器学习模型进行自动化数据分类和术语分配的结果。分类法是通过分组和分类来了解世界的一种方式。许多组织都使用社保号码 (SSN) 跟踪具有各类投资产品的客户，但 SSN 的形式可能多种多样，例如税务识别号码或员工识别号码。若使用基于规则的传统引擎，就很难知道这三个不同术语指的是同一个实体。此外，同一个术语在同一个组织中也可能具有多个不同的含义。机器学习模型为训练系统提供了新方法，当系统通过数据来描述某个领域时，这种模型有助于识别这些关系。

自动数据分类

使用机器学习进行数据分类是涉及到聚类、排序和分类的三步过程。聚类旨在从数据中找到类似实体。用于比较数据的各种特征称为*特性*，如物体的形状和大小等。两个数据值之间相同或接近相同的特性数量越多，则可认为这两个数据值越相似。可通过几种可用的统计方法来执行聚类活动。有些方法要求用户预先指定特性，另一些技术则可通过比较不同的数据值来开发特性。

下图显示了 IBM Unified Governance and Integration 软件组合及其提供的核心服务。其中一些服务尚处于 Beta 测试阶段，或将在今年晚些时候推出。

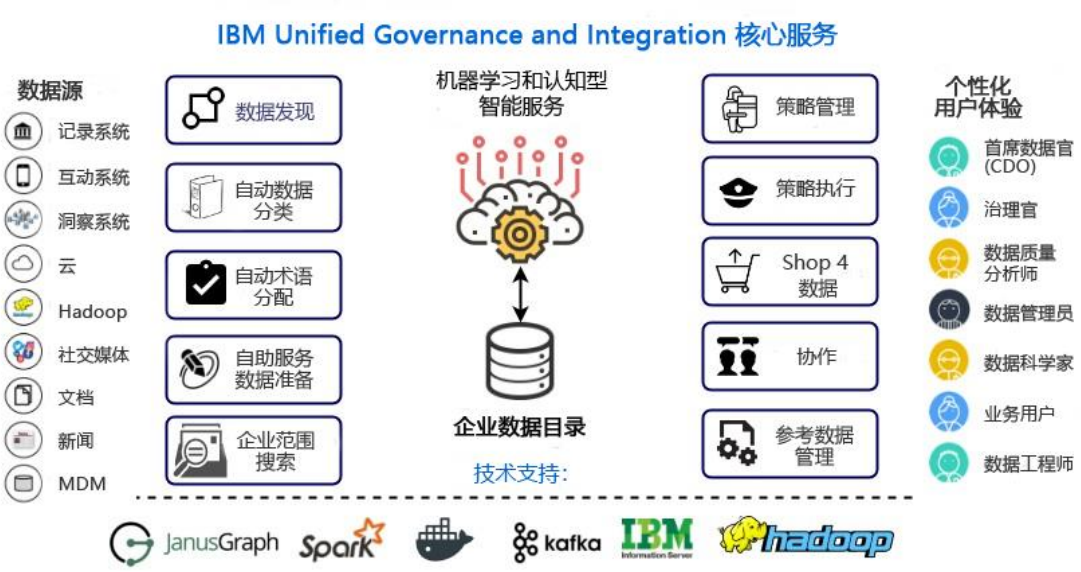

***图 2**：IBM Information Server 11.7 提供的核心服务*

IBM Unified Governance and Integration 软件组合专注于为各种用户角色提供个性化的用户体验和预期结果。例如，由于数据孤岛的存在以及缺乏中央数据目录，经常导致业务用户难以找到所需的数据，因此需求助于 IT 部门。即使业务用户找到了所需的数据，也会因为以下原因而难以理解：

- 数据源没有标签，导致用户只能自己去猜测每个数据字段的上下文和含义
- 数据源中的模式和数据字段使用难以理解的缩写或首字母缩略词，尤其是来自原有系统的数据源
- 数据字段在多个数据源中使用不同的标签重复标记，因此需要对业务术语进行标准化处理

IBM Unified Governance & Integration 支持企业中的业务和技术用户轻松识别数据，还帮助他们找到被组织中其他用户认可的有用的高质量数据。

反馈学习在发现过程中的重要性

许多组织的业务术语和技术元数据之间没有明确的关系。IBM Unified Governance & Integration 可通过面向行业的数据集和标签对模型进行预先训练，以便在用户将业务术语表加载至 IBM Governance Catalog 时立即提供候选的匹配项。对于每个候选术语，用户都会获得相关的置信度分数，用于帮助他们选择正确的术语。如果匹配度达到阈值 (80%)，该术语将被自动分配给元数据，否则，系统将其视为知识输入，用来重新训练机器学习模型。这种动态反馈循环有助于提高数据管理员的工作效率，在随后的元数据发现工作中获得更准确的结果。

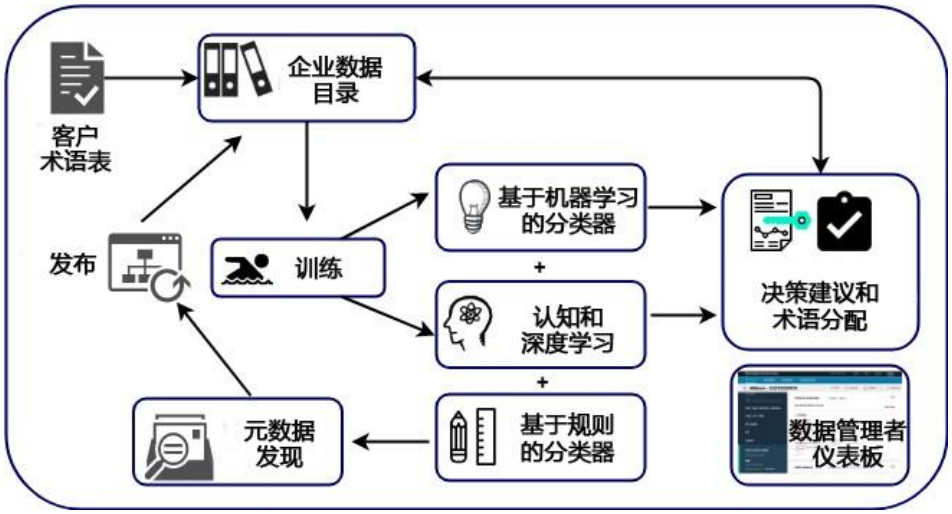

***图 3**：IBM Information Server 11.7 — 基于机器学习的元数据发现流程*

传统的元数据匹配和分配方法是基于规则的。虽然机器学习模型可以利用模糊数据集更好地完成这项工作，但这并不旨在取代现有的应用规则，尤其是在事实证明，现有应用规则和正则表达式能够胜任此项工作的情况下。IBM Unified Governance & Integration 使用机器学习，作为对现有规则的补充，并提供优化的结果，帮助系统从客户的数据领域中进行学习。

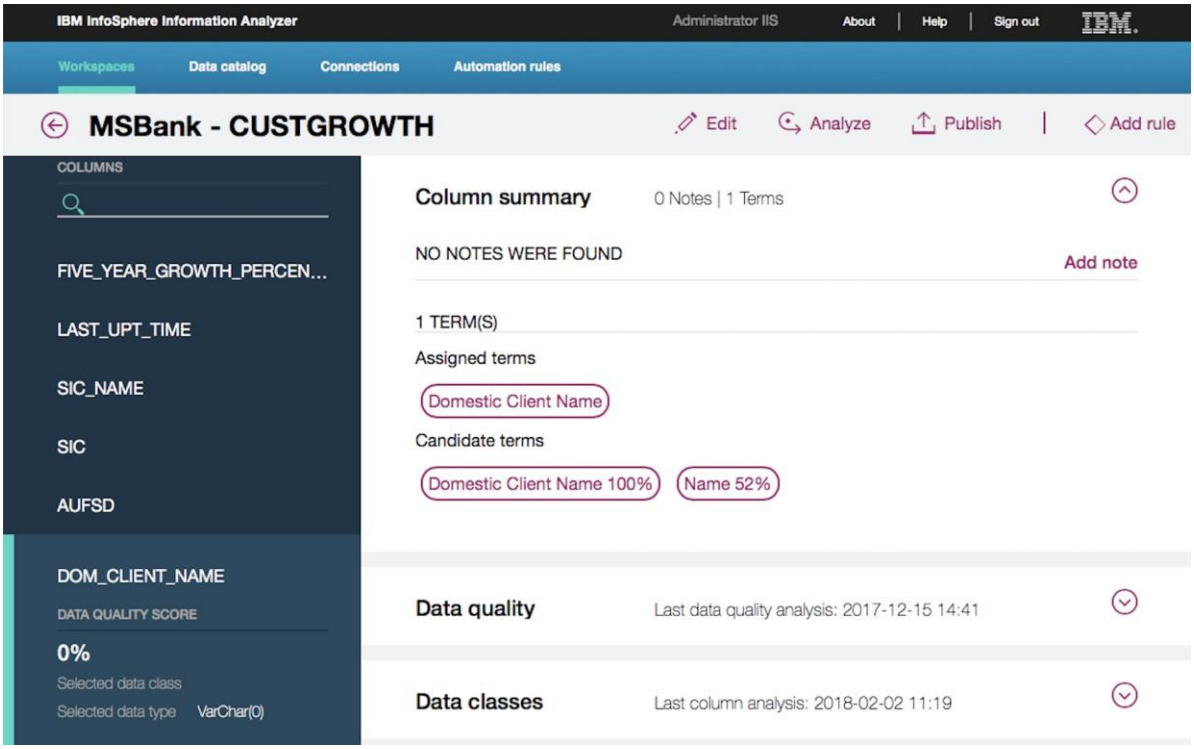

***图 4**：使用机器学习的自动化业务术语分配*

将法规应用于业务数据

当企业尝试查找、理解和使用内部信息时，他们会使用多种方法来组织数据。他们会将数据从互动系统实际转移到数据仓库或 数据集市等洞察系统，或将图形数据库直接连接到实例数据。他们会构建数据湖，为可能感兴趣的所有信息创建“着陆区”，以供数据科学家从中捕获有用的信息。他们会在数据湖的边缘创建沙箱，以供其他人对进入数据湖的数据进行筛选，开始“淘金”。最后也是最重要的一点，企业需了解他们手头现有哪些数据、它们有何含义、来自何处、经历了哪些操作处理（也称为数据沿袭）以及哪些用户可以使用这些数据。

为了帮助企业确定他们现有的数据，IBM 开发了自动发现和分类产品 — 如 StoredIQ 和 Information Analyzer。这些产品使用机器学习模型来读取客户数据、了解模式并从中学习。针对可管控的数据湖应用法规方面所存在的主要问题，IBM Unified Governance & Integration 能够进行解决，它将 IBM 行业模型与神经网络模型相结合，将各种法规悉数纳入到 IBM Governance Catalog 中，然后再将法规术语和业务术语对应起来。

我们以《通用数据保护条例》(GDPR) 为例。我们需要实施四个流程，才能将 GDPR 术语与业务术语关联，以帮助实现隐私合规：

1. 必须从 GDPR 文档（各部分、各条款）中手动提取支持性的内容术语
2. 必须为主要类别创建层次结构
3. 必须由领域专家手动将支持性的内容术语与业务术语匹配

4. 必须将这些支持性的术语映射至业务数据模型

IBM Unified Governance & Integration 使用机器学习来创建神经网络模型，根据对类似法规的处理经验来解释该项法规。这不仅能从原始文档中提取支持性的内容术语，而且还支持您创建可轻松地集成到 IBM Governance Catalog 之中的结构良好的分类表。

结语

目前，面对全球监管趋势以及逐渐兴起的混合云环境（公有、私有及混合云），企业必须提高对主数据实体的管理效率，使用人工智能和机器学习来理解法规及各类变更对其产生的业务影响。他们需要借助强大的解决方案来扩展业务流程，顺利完成数据管理、数据发现、数据质量和数据治理之旅。IBM Unified Governance and Integration 解决方案旨在帮助客户轻松高效地实现这一目标。

观看本次网络直播，了解如何使用 IBM Information Server 的认知数据治理功能，从数据中加速获取洞察。

治理数据湖

将数据整合、数据质量和可用性保障机制嵌入到数据湖环境中，以便加速数据探索和洞察创建流程，同时避免形成数据沼泽。

为企业数据仓库减负

纳入数据整合、数据质量和治理机制，以便将 EDW 数据和 ETL 工作负载转移至数据湖或 Hadoop，保持可信、清洁的数据，供分析之用。

备战 GDPR

通过专注执行保护个人数据和达成一致意见等关键要素，加速完成《通用数据保护条例》(GDPR) 合规准备工作。

实现信息驱动的洞察

提供全方位、高价值的可信数据实体，为每一个数据用户和业务部门主管提供强大支持，推动形成业务洞察与智能。