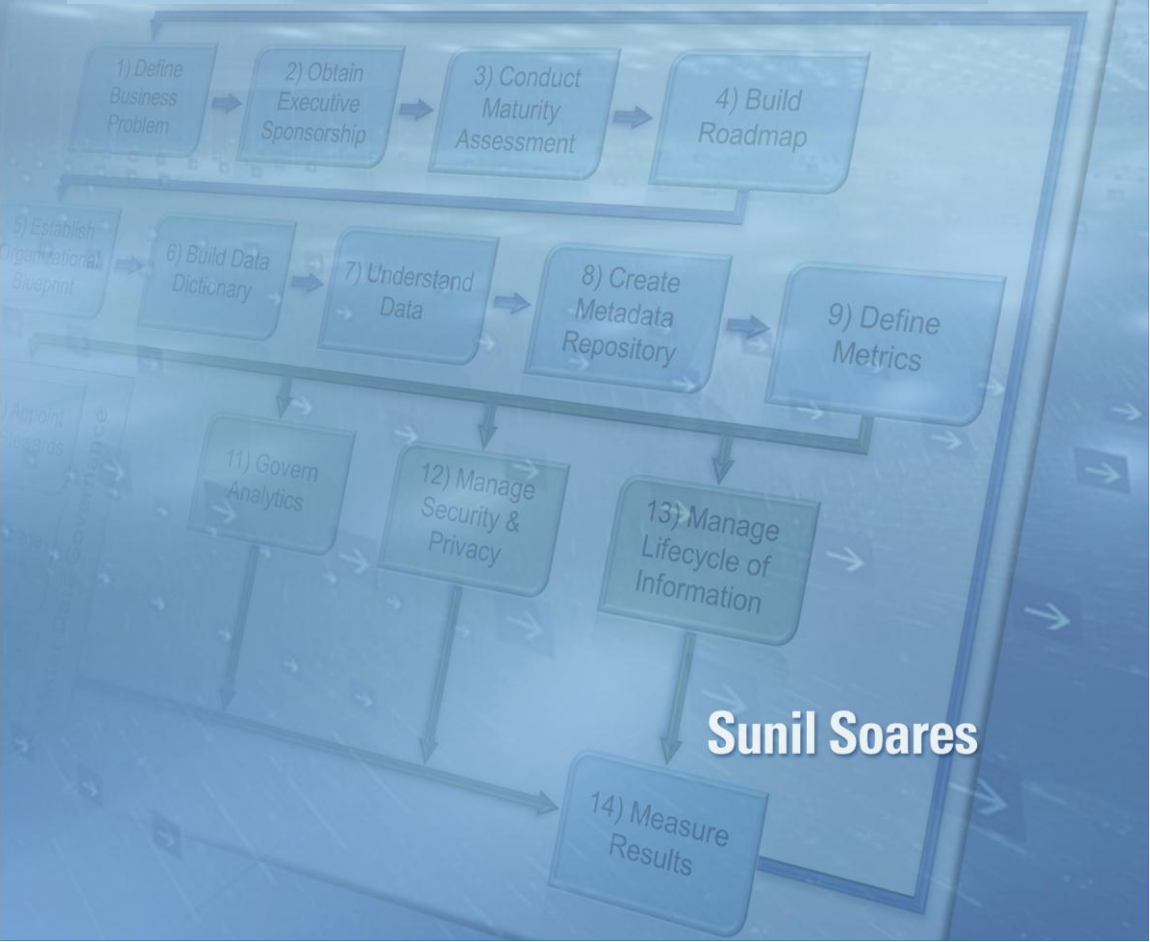
IBM 数据治理统一流程

使用 IBM 软件和最佳实践提升业务价值



Sunil Soares

对《IBM 数据治理统一流程》的赞美

数据治理是主数据管理 (MDM) 项目取得成功的关键，无论是在最初还是在项目实施过程中。在 2010 到 2011 年间，全球拥有 5000 名员工的企业将越来越多地要求，“没有预备的数据治理框架，将不会为 MDM 计划提供资助”。

“实现治理，尽早行动”是聪明的企业和解决方案架构师的战斗口号，他们担负着设定其企业数据整合计划的范围和方向的重任。理解范围、种类和整合挑战本身就极具挑战。而负责定义和执行 MDM 计划的业务和 IT 领导层需要理解和理顺众多的数据治理选项。通过阅读和应用 Sunil Soares 在这篇不可或缺的企业级数据治理介绍中收集的这些优秀的最佳实践、加速器和模型，解决方案架构师、数据治理主管、MDM 计划经理都可以获得通往“实现治理，尽早行动”的捷径。

Aaron Zornes
首席研究官，The MDM Institute
Conference 大会主席，“MDM 和数据治理峰会”全球大会系列

Air Products 花了 18 个月时间实现了一个重要的数据治理计划。这是一趟艰难，但非常激动人心且令人愉悦的旅程。该计划的成功涉及到大量步骤，其中大部分都是我们不曾熟悉的新步骤。数据管理小组和供应商提供了众多建议，但是，每个问题都可通过多种方式解决，该计划必须针对每个公司的需要和条件而调整。正如 Sunil Soares 所详细介绍的，IBM 数据治理统一流程比我们目前为止所见的任何流程的完备程度高出一个数量级。很明显，作者掌握了许多类型的公司的丰富的实践经验，为此学科提供了一种新颖、实用的方法。尽管本书的一个用途是推荐技术方法，但它的重要价值在于它认识到首先要解决的核心问题基本上与技术无关，而关系到人员和流程。本书以一种与供应商独立的方式陈列这些问题。

Tony Harris
企业数据架构师
Air Products

数据治理是一个关键流程，但在危机出现之前很少有人意识到这一点，这些危机包括报告不会交叉统计，高级管理人员不同意关键措施的状态，安全破坏和合规性问题在最糟糕的时刻暴露给了最危险的人，以及多个相互冲突的信息副本的成本摆脱了控制。如果我们可以相信，组织可在仪表板全变为红色之前解决数据治理可以预防的问题，不是很好吗？

Sunil Soares 相信他们可以。在这场优秀的讨论中，他向读者介绍了实现此目标的战略、流程和重要工具。自然，其中的示例为 IBM 工具，这并不奇怪。但这不是推销员的一种宣传腔调，它是成功解决可能毁灭性的问题的指南。阅读之后一定会有所收获。

Merv Adrian
IT Market Strategy

业务世界是一个动态且富有挑战的领域。业务领导们始终面临着调整和应用他们的专业技能和经验来影响市场形势，以保持业务成功的挑战。信息时代在千禧年拉开了帷幕，这以信息和数据在推动业务运营、促进战略决策制定和鼓励组织利用数据实现市场竞争优势的过程中的战略作用为标志。个人、社区、企业、政府和社会机构最容易受到信息时代的影响。大部分商业成功案例都离不开创造性、战略性地利用企业中存在的数据。在信息时代，数据是一项战略性企业资产，必须像这样进行管理。

数据治理准则概括了如何实现该管理学科，确保组织继续享受其数据带来的好处。Sunil Soares，IBM 团队的一名重要成员，在本书中详细介绍了数据治理的准则。他专业地组织了内容，其中包括从他与全球众多业务合作伙伴之间的服务活动中收集的示例所支撑的理念。

此作品是一份可供业务领导在其数据治理实现中采用的实用手册。正如俗语所说，“当其他某人完善了模型设计，为什么还要重复发明车轮呢？”IBM 团队已定义了数据治理的模型。通过本书，业务领导可获取并采用它来确保其实现的成功。

我很荣幸能够在我组织内的数据治理项目上与作者直接合作。Sunil 是此领域的专家，他为此主题付出了大量心血，拥有该主题的丰富知识，并且对数据治理充满着激情。他忙于以专业但简单的方式为组织提供指导和建议，这使他广受客户喜欢。我与 Sunil 打交道的时间对于我和我的组织都具有宝贵的价值。作者讨人喜欢的品质是本书的一大闪光点，也是本书吸引读者的地方。我相信您也一定会喜欢并赞赏本书，因为其中包含众多宝贵的见解。Sunil，谢谢您通过您的专业技能为作为数据治理从业者的我们注入了力量。

如果您正在参与数据治理项目，祝您通过努力取得巨大成功。

Komalin Chetty
数据治理支持者
Telkom South Africa

IBM 数据治理统一流程

使用 IBM 软件和最佳实践提升业务价值

Sunil Soares

MC Press Online, LLC
Ketchum, ID 83340

IBM 数据治理统一流程
Sunil Soares

第一版
第一次印刷——2010 年 9 月

MC Press Online, LLC
Corporate Offices
P.O. Box 4886
Ketchum, ID 83340-4886 USA
有关销售和/或客户服务的信息，请联系：
MC Press
P.O. Box 4300
Big Sandy, TX 75755-4300 USA
有关权限或特殊订购的信息，请联系：
mcbooks@mcpressonline.com

ISBN：978-158347-360-3

特别的感谢 Maya Soares、Lizzie Soares、Helena Soares、Cecilia Soares 和 Hubert Soares 在本书创作期间给予的巨大支持。

致谢

确保正确对待数据治理等复杂的主题，离不开 IBM 内外众多思想领袖的参与。本书的创作是一项团队工作，需要感谢的人很多。

感谢 Arvind Krishna、Bob Keseley、Dave Laverty、Inhi Cho、Paraic Sweeney、Piyush Gupta、Tom Inman 和 Mike Nolan 对作为 IBM 的关键计划的数据治理的大力支持。

特别感谢 Steve Adler，他通过过去 5 年执掌 IBM 数据治理委员会期间的强大领导力，大力推动将数据治理合并到一个拥有新兴的从业者团体的独立学科中。

感谢许多人对本书做出的贡献：

- Ken Bisconti、Craig Rhinehart、Laurence Leong 和 Paula Fricker 在信息生命周期治理领域执行了市场领先的活动
- David Corrigan 和 Ian Stahl，他们提供了宝贵的主数据治理见解
- Michael Dziekan，他拥有商业智能能力中心的丰富经验
- Farnaz Erfan、Beate Porst 和 Steven Totman，他们在业务术语和元数据方面做出了卓越的贡献
- Todd Goldman 和 Alex Gorelik 提供了数据查询方面的内容
- Brett Gow 从从业者角度提供了数据治理方面的见解

- Bill Mathews 在建议 IBM 保险行业客户和分享数据治理最佳实践方面提供了丰富经验
- Marty Moseley 提供了主数据治理方面的整体视角，他还完成了本书附录 E 中的数据治理声明
- Eric Naiburg 在与 IBM 软件相关的所有主题上提供了重要信息
- Phil Neray 和 Brian Roosevelt 提供了与数据安全和合规性相关的主题的卓越见解
- Arvind Sathi 提供了他在典型行业的数据治理主题相关经验
- Helena Soares 对手稿进行了严谨细致的编辑
- Wayne Wilczynski 在为 IBM 银行和金融市场客户提供建议以及分享真实世界中所用的最佳实践方面提供了丰富经验

另外，感谢来自 IBM 的 Michael Curry、Glenn Hintze、Jan Shauer、Steven Stansel 和 Susan Visser，来自 Air Products 的 Tony Harris，来自 Chevron 的 Sebastian Gass，以及来自 KeyCorp 的 Michael O’Connor 在本书创作期间提供的建议和见解。

目录

Arvind Krishna 作的序言

IBM 自 2005 年 IBM 数据治理委员会成立开始就走在信息治理运动的前列。我们与全球行业领先的公司紧密合作，解决与治理相关的最大挑战。信息治理拥有其合规性和风险方面的根源，但在过去几年来，我们发现了一种向利用它来创建价值和减轻风险方面的转变。

组织的信息治理方法和对技术的采用存在不同的成熟度水平。对一家企业而言最佳的方法可能不适合另一家企业。组织结构、角色、基本功能是成功的等式中的重要部分。

每家公司都有多个可优化的信息供应链。大部分组织如今面临的挑战是，他们无法识别自己的信息供应链，很少管理它们来挖掘整个组织中一致且广泛存在的新业务洞察。

IBM 创造了一种完善的信息治理方法，为行业提供了最强大的产品、服务和最佳实践组合来解决每个组织的需要。本书提供了一组实用的详细步骤和子步骤来实现信息治理计划，以及 IBM 软件所提供的相关信息。

Arvind Krishna
总经理，
Information Management
IBM

Michael Schroeck 作的序言

如今，全球的组织都认识到他们的信息资产的重要性和价值。与此同时，高层管理人员并未充分利用这些信息，原因在于缺乏准确性、一致性、相关性和及时性。结果，信息治理被推到了前线，许多公司正在尽力研究如何有效设计和实现信息治理计划。本书提供了答案，介绍了一个成熟、完善且实用的企业信息治理方法。，提供了答案。对于经验丰富的信息治理专家，以及不熟悉此领域的新手，它都是一本必读的著作。

仅通过应用后文中介绍的准则，组织就可以真正最大化其信息的价值，这是成为“智慧”公司的必备条件。

Michael Schroeck
合作伙伴兼全球领导，
BAO 分析解决方案团队和能力中心，
IBM Global Services

Steve Adler 作的简介

最近，我向银行申请了一笔租车贷款。在线申请流程非常好，为我节省了通过其他方式将花在分支办公室或电话上的许多时间。我在 60 秒内就得到了一份接受报价和利率。几分钟后，我打电话给银行要求完成申请流程。不幸的是我犯了一个错误，将贷款分类为再筹款，实际上应为租赁买断。一旦提交，我就无法在线更改表格，呼叫中心代表也无法这么做。所以，我必须取消申请并在做一遍。再一次地，我花两分钟填好了表格并获得了限额，并再一次打电话告诉呼叫中心代表。

几天后，我进入支行完成了交易。支行经理彬彬有礼并且很热情，在 30 个不同的表格上签名后，我为我的车筹得了租车款。3 天过后，租赁公司打电话说申请中缺少两个必须转发给银行的表格，我呼叫了银行，但代表也毫无线索。“一定是租赁公司出错了，”他们说。我又呼叫租赁公司，如此往返了几次，最后才同意去银行重新签署表格。然后，我开始从银行获得电子邮件，通知我最初的在线报价已获得批准并等待我进一步操作，但我可以确定银行以告诉我它已取消了该报价。

这种时常出现的数据治理问题困扰着每家企业。我看到了许多糟糕的情形。大部分人只是认为它们是错误，但它们可能导致业务损失、更高的风险，还一定会导致额外的成本。无论您是否拥有正式的数据治理计划，您的组织都会碰到类似这样的数据治理问题，其他许多组织也是如此。您知道它，您的客户也知道它。

一旦认识到这个问题，选择就非常简单了：您可以处理它，或者忽略它。如果您正在阅读本书，则表明您已决定不忽略它。很好。

您接下来的决定是如何处理它。错误是人生活的一部分。您的业务中出错是因为您的业务是由人在运作。数据本身不会出错。您需要首先更改运作您业务的人对待数据的方式、他们对数据执行的操作，以及他们建立使用数据的业务的方式。为此，您需要一个系统，一个可帮助将人们集中起来相互协调、协作和沟通的数据治理计划。

本书提供的一些重要工具可帮助您踏上通往构建数据治理计划的正确道路，它们可以修复您的组织每天产生的简单和复杂的错误和遗漏。

购买了本书，您就已经做出了最重要的决策。现在完整阅读本书并启动您的计划，因为时间不会停止。在您阅读本书的短短的时间内，您组织某个地方的某个人就已经丢弃了一些表格、错误输入了新帐户的编码，或者向客户发送了重复的帐单。

时不我待！！！

Steve Adler

IBM 数据治理委员会主席

1

数据治理简介

数据治理是一门将数据视为一项企业资产的学科。它涉及到以企业资产的形式对数据进行优化、保护和利用的决策权利。它涉及到对组织内的人员、流程、技术和策略的编排，以从企业数据获取最优的价值。从一开始，数据治理就在协调不同的、孤立的且常常冲突的策略（可能导致数据异常）的过程中扮演着重要角色。

非常类似于客户关系管理 (CRM) 诞生之初，组织开始任命全职或兼职数据治理负责人。与任何新兴学科一样，数据治理有许多定义，但市场已开始围绕将数据视为资产的定义进行具体化。

传统的会计规则不允许公司在他们的财务负债表上将数据视为财务资产，除非它是从外部实体购买的。尽管存在这种保守的会计方法，但企业现在认识到他们的数据应该视为类似于工厂和设备的资产。

将数据视为战略性企业资产，意味着组织需要建立其现有数据的清单，就像建立物理资产的清单一样。典型的组织拥有与其客户、供应商和产品相关的过量的信息。这样的组织甚至可能不知道所有这些数据位于何处。

这可能具有挑战，尤其是对于*个人可识别信息* (*PII*)。组织需要防御其财务、企业资源规划和人力资源应用程序中的关键业务数据受到未授权更改，因为这可能影响到其财务报告的完整性，以及日常业务决策的质量和可靠性。他们必须也防御信用卡号和 PII 数据等敏感的客户信息，以及客户名单、产品设置和专用算法等知识产权受到内部和外部威胁的破坏。最后，组织需要从他们的数据获取最大的价值，推动改进的风险管理和客户中心性等计划的实施。

数据曾经是组织最大的价值来源和最大的风险来源。糟糕的数据管理常常意味着糟糕的业务决策和提供给违规和盗窃更大的暴露面。举例而言，美国的 Sarbanes-Oxley、类似的 European Sarbanes- Oxley 以及 Japanese Financial Instruments and Exchange Law (J-SOX) 等制度在受限的访问与恰当的数据使用之间指定了一个平衡点，这个平衡点由规则、策略和制度来控制。另一方面，利用规则、可信的数据的能力可帮助组织提供更好的服务，提升客户忠诚度，花更少的工作来遵守制度和进行报告，以及提升创新能力。

组织必须考虑其非结构化数据的业务价值。这种非结构化数据常常称为*内容*，需要像结构化数据一样进行治理。

非结构化数据治理的一个不错示例是设置记录管理策略。许多公司都被要求将电子和纸张记录保留一段给定的时间。他们需要在法律查询过程中迅速且经济高效地生成这些效率。他们还需要遵守为特定的文档类型既定的保留计划。一些组织使用词汇“信息治理”来定义此计划。尽管我们交替使用词汇“数据”和“信息”，但我们将在本书中坚持使用更常见的词汇“数据治理”。

以下是组织可通过治理其数据而获得的一些收益：

- 改进用户对报告的信任级别
- 确保数据在来自组织不同部分的多个报告上的一致性

- 确保恰当地保护企业信息，以满足审计者和监管者的需求
- 改进客户的洞察水平，推动营销计划的实施
- 直接影响组织最关注的 3 个因素：提高收入、降低成本和减少风险

由 Steve Adler 于 2004 年 11 月创立，IBM® 数据治理委员会是一个面向数据治理领导、信息治理领导、首席数据官、企业数据架构师、首席信息安全架构师、首席风险官、首席合规官和首席隐私官等从业者的领导论坛。该委员会关注与组织如何有效地以企业资产的形式治理数据相关的问题。它专注于信息、业务流程和信息对组织的价值之间的关系。

依据 IBM 数据治理委员会的 Adler 在白皮书《IBM 数据治理成熟度模型：为有效的数据治理建立路线图》中介绍的发现，以下是如今最重大的数据治理挑战：

- 不一致的数据治理可能导致业务目标与 IT 计划之间脱节。
- 治理策略未链接到结构化的需求收集和报告。
- 未从生命周期角度解决常见的数据存储库、策略、标准和计算流程中的风险。
- 元数据和业务术语库未用于弥合全球化企业中多个应用程序之间的语义区别。
- 如今很少存在能链接安全、隐私和合规性的数据资产价值评估技术。
- 控件和架构在建模长期后果之前就已部署。
- 跨不同数据领域和组织边界的治理可能难以实现。
- 需要治理的准确内容常常不明确。

- 数据治理包含战略和战术元素，它们常常未明确定义。

数据治理关乎决策权力和对人类行为的影响。本书是一位从业者的指南，基于与实现类似计划的组织打交道的真实经验。它重点介绍了 IBM 软件工具和最佳实践支持数据治理流程的具体区域。

2

IBM 数据治理统一流程

投入精力实施完善的企业数据治理计划的收益丰富多样，实现强大的数据治理的挑战也是如此。

许多企业已要求获得一个列出了实现数据治理计划的步骤的流程手册。显然，每个企业将以不同方式实现数据治理，这主要是因为他们具有不同的业务目标。一些企业可能专注于数据质量，而其他企业专注于客户中心性，还有一些企业专注于确保敏感客户数据的隐私。一些组织将接受一种正式的数据治理计划，而其他组织希望实现更加轻量型和战术性的方案。

且不说这些细节，每个组织应该执行一些步骤来治理自己的数据。图 2.1 中所示的 IBM 数据治理统一流程列出了这 14 个主要步骤（10 个必需步骤和 4 个可选专题），以及支持有效的数据治理计划的相关 IBM 软件工具和最佳实践。

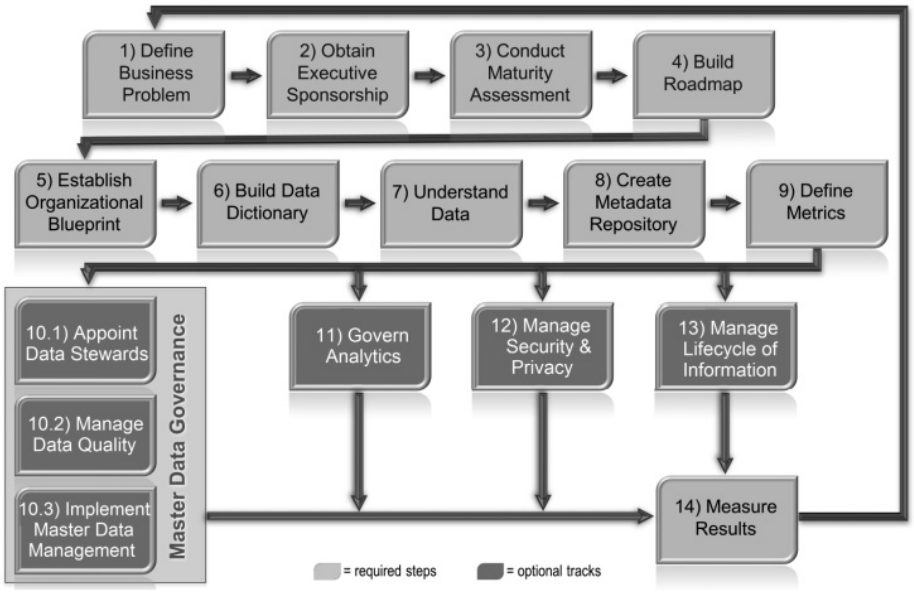


图 2.1：IBM 数据治理统一流程概述。

10 个必需步骤是为有效的企业治理计划奠定基础所不可或缺的。企业然后将选择从 4 个可选专题（也就是主数据治理、分析治理、安全和隐私，以及信息生命周期治理）中选择一个或多个。最后，需要定期度量数据治理统一流程，将结果传送给管理层支持者。

让我们更详细地分析一下图中的步骤：

1. *定义业务问题。*

   数据治理计划失败的主要原因是，它们无法识别实际的业务问题。组织亟需围绕一个特定的业务问题（比如失败的审计、数据破坏或出于风险管理用途对改进的数据质量的需要）定义数据治理计划的初始范围。一旦数据治理计划开始解决已识别的问题，业务职能部门将支持它将范围扩展到更多区域。

2. *获取高层支持。*

   得到关键 IT 和业务高层对数据治理计划的支持很重要。获得此支持的最佳方式是以业务案例和“快捷区域”的形式建立价值。例如，业务案例可以专注于一家人和名称匹配，改进数据的质量以支持客户中心性计划。

   与任何重要的计划一样，组织需要任命数据治理的整体负责人。组织在过去将首席信息安全官视为数据治理的负责人。但是，今天，数据治理的责任常常在 CIO 的办公室内履行，在商业智能或数据架构区域。数据治理领导职责可能属于首席风险官，尤其是在银行。越来越多的企业正在以全职形式安排数据治理角色，使用“数据照管人”（表明将数据视为企业资产的重要性）等头衔。无论头衔是什么分配给此角色的职责必须在高层评分中足够高，以确保数据治理计划能促进有意义的变化。

3. *执行成熟度评估。*

   每个组织需要对其数据治理成熟度执行一项评估，最好每年执行一次。IBM 数据治理委员会基于 11 种类别（比如“数据风险管理和合规性”、“价值创建”和“照管”）开发了一种成熟度模型（将在第 5 章中探讨）。数据治理组织需要评估组织当前的成熟度水平（当前状态）和想要的未来成熟度水平（未来状态），这通常在 12 到 18 个月后。这段时间必须长到足够生成结果，短到确保关键利益相关者的持续支持。

4. *创建路线图。*

   数据治理组织需要开发一个路线图来填补 11 个数据治理成熟度类别的当前状态与想要的未来状态之间的空白。例如, 数据治理组织可以检查“照管”的成熟度空白，确定企业需要任命数据照管人来专门负责

目标主题区域，比如客户、供应商和产品。数据治理计划也需要包含“快捷区域”——计划可带来近期业务价值的区域。

5. *建立组织蓝图。*

   数据治理组织需要建立一种章程来治理其操作，确保它拥有足够的成熟度来在关键形势下担当决胜者。数据治理组织最好在一种 3 层格式下操作。顶层是数据治理委员会，它由依靠数据作为企业资产的关键职能和业务领导组成。中间层是数据治理工作组，它由经常会面的中层经理组成。最后一层由数据照管社区组成，它负责每天的数据质量。

6. *创建数据字典。*

   业务词汇的有效管理可帮助确保相同的描述性语言适用于整个组织。数据字典或业务术语库是一个存储库，包含关键词汇的定义。它用于在组织的技术和业务端之间实现一致性和达成一致。例如，“客户”的定义是什么？客户是某个进行购买的人还是某个考虑购买的人？前员工是否仍然分类为“员工”？词汇“合作伙伴”和“经销商”是否同义？这些问题可通过创建一个通用的数据字典来回答。一旦实现，数据字典可应用到整个组织，确保业务词汇通过元数据与技术词汇相关联，而且组织拥有单一、共同的理解。

7. *理解数据。*

   有人曾经说过，“您无法控制您还未理解的东西。”如今很少有应用程序是独立存在的。它们由系统和“系统的系统”组成，包含散落在企业各个角落但整合或至少相互关联的应用程序和数据库。关系数据库模型实际上使情况更糟了，它使业务实体的存储分散化。但是所有一切是如何关联的？数据治理团队需要发现整个企业中关键的数据关系。

数据查询可能包括简单但难以发现的关系，以及企业 IT 系统内的敏感数据的位置。

8. *创建元数据存储库。*

   元数据是关于数据的数据。它是有关任何数据工件（比如其技术名称、业务名称、位置、被认为的重要性和与企业中其他数据工件的关系）的特征的信息。在查询阶段，数据治理计划将从数据字典生成大量业务元数据和大量技术元数据。此元数据需要存储在一个存储库中，所以它可以在多个项目之间共享和利用。

9. *定义度量指标。*

   数据治理需要拥有可靠的度量指标来度量和跟踪进度。数据治理团队必须认识到当您度量某个东西时，性能就会改进。因此，数据治理团队必须挑选一些关键性能指标 (KPI) 来度量计划的持续性能。例如，一家银行将希望评估行业的整体信贷风险。在这种情况下，数据治理计划可以选择空的标准行业分类 (SIC) 代码的百分比作为 KPI，跟踪风险管理信息的质量。

这些是前 9 个必需的步骤。最后一个必需步骤将在本章后面介绍。企业还需要在 4 个可选的数据治理专题（主数据治理、分析治理、安全和隐私，以及信息生命周期治理）中至少选择一个。

让我们选择主数据治理可选专题，分析一下它的必需子步骤的应用。一家组织需要确保业务问题（比如客户中心性）得到了明确传达，确定了业务和 IT 部门中的高层支持者。组织将执行一个简短的数据治理成熟度评估并定义一个路线图。需要有某种级别的数据治理组织来协调业务和 IT，确保近期收益。“客户”等业务词汇需要明确定义，尤其是如果“客户”是一个主数据领域。数据治理组织需要理解现有的数据源和关键的数据元素。业务定义和来自查询过程的技术元数据需要捕获到元数据存储库中。最后，数据治理组织需要建立 KPIs，

比如客户重复率的减少，以确保主数据治理计划的持续性能。

对必需步骤的重视水平将因为数据治理选择的可选专题不同而不同。举例而言，让我们回顾一下可以基于所选的一个或多个可选专题，如何应用第 7 步（“理解数据”）。主数据治理专题将涉及到理解促进源到目标的映射的关键数据元素。分析治理专题将涉及到理解关键报告和关键数据元素之间的关系。安全和隐私专题将涉及到理解敏感数据的位置。最后，信息生命周期治理专题将使企业能够理解业务对象（比如客户）的位置，作为一个存档项目的前身。

我们将在后续章节中更详细探讨这些主题，所以我们仅将提供一些示例问题和本章剩余内容的可能重点区域。这里是 IBM 数据治理统一流程中的可选专题的简短描述：

10. *治理主数据。*

    企业内最有价值的信息（与客户、产品、材料、供应商和帐户相关的关键业务数据）统称为*主数据*。尽管它很重要，主数据常常是重复的并分散在整个企业的各种业务流程、系统和应用程序中。治理主数据是一种持续的实践，其中业务领导为实现业务目标而定义准则、策略、流程、业务规则和度量指标，管理他们的主数据的质量。

    与主数据相关的挑战可能困扰着大部分组织，但并不总是可以轻松获得合适的业务支持水平来修复问题的根源。因此，论证对主数据计划的投资的合理性很重要，例如，考虑一个类似银行的组织，它将多封邮件发送到同一个家庭。此银行可以通过清理其客户数据来创建“家庭”单一视图，从而建立快速的投资回报。基本而言，大部分数据治理计划会处理围绕数据照管、数据质量、主数据和合规性的问题。

11. *治理分析。*

企业已投入了巨额资金建立数据仓库来获取竞争洞察。但是，这些投资并不总是得到了结果，导致企业越来越多地审查其对分析的投资。我们将“分析治理”专题定义为设置更好地协调业务用户与对分析基础架构的投资的策略和过程。数据治理组织需要询问以下问题：

- 我们的数据在每个业务区域有多少用户？
- 我们在每个业务区域创建了多少份报告？
- 用户是否从这些报告获得了价值？
- 我们每月执行了多少报告？
- 生成一份新报告需要多长时间？
- 生成一份新报告有哪些成本？
- 我们能否培训用户来生成他们自己的报告？

许多组织将希望设立一个商业智能能力中心 (BICC) 来培训用户，传播商业智能，以及开发报告。

12. *管理安全和隐私。*

数据治理领导，尤其是向首席信息安全官报告的领导，常常必须处理围绕数据安全和隐私的问题。一些常见的数据安全和隐私挑战包括：

- 我们的敏感数据位于何处？
- 组织是否已在非生产环境（开发、测试和培训环境）中屏蔽了它的敏感数据以符合隐私制度？
- 是否已有数据库审计控件来阻止特权用户（比如 DBA）访问隐私数据，比如员工工资和客户名单？

13. *治理信息生命周期。*

非结构化内容占典型企业中的数据的 80% 以上。随着组织从数据治理转向信息治理，他们开始考虑这种非结构化内容的治理。

信息的生命周期始于数据创建，结束于它从生产环境删除和不复存在。数据治理组织必须处理以下与信息生命周期相关的问题：

- 我们与数字化纸张文档相关的策略是什么？
- 我们针对纸张文档、电子文档和电子邮件的记录管理策略是什么？（换句话说，我们将哪些文档保留为记录？保留多长时间？）
- 我们如何归档结构化数据以减少存储成本和改善性能？
- 我们如何将结构化和非结构化数据结合到一个通用的策略和管理框架下？

在这些可选的专题之后，在数据治理统一流程的末尾还有一个必须步骤：

14. *度量结果。*

数据治理组织必须通过不断监控度量指标来确保持续改进。在第 9 步中，数据治理团队设置度量指标。在此步骤中，数据治理团队依据这些度量指标向来自 IT 和业务部门的高层利益相关者报告进度。

整个数据治理统一流程需要以持续循环的形式操作。该流程需要度量结果并循环回到高层支持者，以获得数据治理计划的持续支持。

3

第 1 步：定义业务问题

在我们花大量时间探讨最佳实践之前，值得检查一下许多数据治理计划失败的关键原因。大部分具有停滞的数据治理计划的组织识别了以下症状：

- “业务部门没有在数据治理中看到任何价值。”
- “业务部门认为应该由 IT 负责管理数据。”
- “业务部门专注于近期目标，数据治理被视为一个长期计划。”
- “CIO 削减了我们的数据治理部门的资金。”
- “业务部门为数据照管人重新分配了其他职责。”

归根到底，数据治理计划失败的根源是缺乏与业务价值的链接。在本质上，IT 在没有合适的业务支持的情况下治理数据。治理数据不是 IT 的职责。而 IT 是保管者，支持、实现和提供必要的功能来度量和跟踪业务部门所利用的数据。

业务价值的推动因素因行业、公司不同而不同。在业务角度上讲，形成数据治理价值的因素有太多要讲的内容，这超出了度量和管理 DBA 和商业智能分析师感兴趣的事务的范畴。一旦您使用业务高层所使用的语言形成了价值主张，您将吸引有权利大展拳脚的高层领导从更出色的数据治理实现改进的业务收益。

使用后面的示例作为起点。当然，确保您考虑了您自己的行业和组织的独特环境。

银行

在银行业内，首席风险官是作为数据治理计划的关键业务支持者的角色出现的。信贷风险是数据治理改善决策制定质量的一个出色示例。考虑一家无法无法轻松量化对方风险的商业贷款公司的示例。该贷款公司无法轻松量化在不同国家拥有独立的信贷业务的多家子公司的企业实体的整体风险。信贷风险组织通常将使用电子表格计算对方的风险，这个流程既耗时又容易出错。信贷风险高级副总裁支持 IBM InfoSphere™ Master Data Management 中包含的一个解决方案以及 D&B 的企业层次结构。因此，贷款公司能够以更快的周转时间制定更好的信用决策。多家银行现在安排了全职数据照管人角色来专门负责管理企业和法律层次结构。

首席风险官还必须确保他们的报告可以信任，始终询问问题“报告是否包含值得信赖的信息？”大型银行的一位数据治理从业者将问题简单地描述如下：

> *我们的首席风险官担忧监管者希望理解我们报告的数据的来源。没有合适的元数据和数据血统，我们无法证明报告中的特定字段源于一个特定的数据集市，该数据集市又进而来自企业数据仓库，最终来自一组后端数据源，以及证明它们之间的所有数据转换。*

安全和隐私也是银行业内的重要驱动因素，其中使用制度来保护个人可识别信息 (PII)，比如

加拿大的个人信息保护和电子文档法案 (Personal Information Protection and Electronic Documents Act, PIPEDA)。除了定义敏感数据，一家大型金融机构的数据治理组织还建立了一条策略，要求所有访问敏感数据字段的应用程序需要经过首席隐私官或她的代表的批准。

保险

保险业越来越多地对保险客户和经纪人的单一视图施予高度关注，推进以客户为中心的计划，比如交叉销售和向上销售营销活动。例如，一家多险种保险公司希望向其寿险保险客户销售汽车保险和房屋所有者保险。

另一个示例是 Solvency II，它是一组更新的制度要求，适用于在欧盟内运作的保险公司。Solvency II 将于 2012 年晚期生效，已成为在欧洲运作的保险公司的一大关注区域。Solvency II 的目标是减少保险公司将无法满足索赔的风险。

Solvency II 有时称为保险公司的“Basel II”。银行在为其 Basel II 计算提供质量数据来源方面面临着诸多挑战，预计保险公司也将如此。Committee of European Insurance and Occupational Pensions Supervisors (CEIOPS) 建议 Solvency II 数据的质量应该基于适当性、完备性和准确性来评估。因此，欧洲保险公司需要高度重视构建专门解决 Solvency II 需求的数据治理计划。例如，数据治理委员会可以建立一项策略，要求恰当地建立企业层次结构来准确量化集团和对立方的风险。

零售

最先进的零售商开始部署客户中心性计划，包括基于客户的希望和需要提供产品以及甚至完整的零售体验的忠诚度计划。在零售商利用大量数据来划分客户时，数据治理尤其重要。零售商也可以利用数据治理最佳实践来减少成本。例如，零售商可能寻求减少将多个目录邮寄给相同家庭的开支，方法是匹配具有相同地址的客户。

类似地，财务和供应链团队将关注“供应商单一视图”，以降低采购成本和优化制造折扣。事实上，一些零售商已开发了他们的重要供应商在多个部门和产品线的单一开支视图。

数据治理组织可以围绕所有客户所需的邮寄地址数据格式建立策略。它也可以编写业务规则，基于特定条件（比如名称和地址）唯一地标识“客户”。它可以围绕识别属于相同家庭的客户而建立策略。最后，可以合理地要求数据治理组织编写一条策略和规程，规定在添加新供应商之前必须搜索供应商名称以最小化重复。

政府

一些州和地方政府已经开始向居民提供其所有服务的单一视图。目的在于允许社会工作者查看居民的家族历史、财务信息、工作背景和参与食品券和公共医疗等计划的资格。

政府内的单一视图项目带来了一些有趣的数据治理问题。例如，数据治理组织必须建立与客户信息文件中丢失或不完整的流浪人口地址信息相关的策略。此外，子福利机构的数据治理组织必须建立策略来跟踪未出生的婴儿，他们没有姓名、地址或社会安全号码。数据治理组织必须建立策略来解决数据异常，最大化居民数据共享和服务质量，同时最小化可能在无数政府流程中导致糟糕结果的数据错误。需要建立围绕标准命名约定的策略，处理丢失的数据的规则，以及在发现潜在欺诈时的通知规则。

政府机构和部门拥有大量关于其居民（包括孩子）的数据。任何人都不希望让这些敏感数据落入坏人的受众。因此，数据治理组织应该建立围绕敏感数据定义和基于“需要知道”的准则而限制访问的机制来建立策略。

医疗

医疗服务付费者和提供者行业必须遵守 United States Health Insurance Portability and Accountability Act (HIPAA) 等制度，HIPAA 保护*受保护的健康信息 (PHI)* 的安全和隐私。数据治理在识别 PHI 数据并设置策略来确保该数据的安全和隐私的过程中扮演着重要角色。例如, 数据治理组织必须管理策略和规程，以确保患者记录没有混杂在一起，阻止添加重复的患者记录。必须针对命名约定以及添加新患者记录的最低数据需求建立策略。这些策略处理生和死的问题，数据治理组织是唯一有权编组必要的资源来保证一致的数据质量水平，保障患者安全和隐私的组织。

数据治理还在确保健康计划拥有其成员、提供者和代理的单一视图的过程中扮演着关键角色。健康几乎目前拥有大量孤立的系统，在这些系统中常常会将新成员信息输入到 10 来个不同的屏幕上。结果，成员信息不一致且分散在健康计划系统的各个角落。数据治理组织需要围绕整个企业中的“成员”的定义设置策略。它还需要提供单一机制来在所有孤立系统中更新对姓名和地址等数据的更改。采用恰当的数据治理策略实现，健康计划可减少管理这些零散系统的成本和工作。

电信

许多电信服务提供商（“telcos”）通过并购得到了发展。在美国，telcos 已演化为由 Public Utilities Commission (PUC) 授权的垄断组织。随着时间的推移，独立的、每个州独有的组织合并到了代表几个州的组合的组织中。例如，Pacific Northwest Bell 从两个州（华盛顿和俄勒冈）组建。在与 AT&T 脱离之后，这些 Bell 经营的公司经过合并，形成了 Regional Bell Operating Companies (RBOC)。例如, Pacific Northwest Bell 于 Mountain Bell 和 Northwestern Bell 合并，形成了 US WEST。最终，BellSouth、Ameritech 和 Pacific Bell 等地区公司被国家运营商收购。

州 PUC 管理着在每个州提供的产品。每个 RBOC 投资了自己的订购和结算系统，对 Universal Service Order Codes (USOC) 和 Field Identifiers (FID) 中编码的产品数据具有具体的标准和格式。要向企业提供标准产品，产品经理必须遍历 3 种级别的数据标准：企业、地区和州。此外，订购和结算系统拥有自己的产品数据和规则标准。

所有这些复杂性的影响可在一些问题中看出，比如对呼叫中心代表的较长培训时间，以及推出新产品、程序包和促销活动的较长的前期准备时间（因为每项产品更改必须传达到每个州、地区或应用程序）。随着 telcos 开始重新设计其订购和结算系统来提供更加灵活的产品推出环境，这些产品代码和规则必须放在所有地区、州和应用程序所通用的数据治理流程下。带给组织的收益非常巨大，标准产品提供了更好的品牌形象、更快的上市时间、对呼叫中心代表更短的培训时间，以及更简单的自助服务 Web 界面。

telcos 还越来越关注他们的数据增长。随着价格的总体下降，数据通信正在快速发展，而收入上没有任何相应的增长。与此同时，在数据保留方面还存在一些重要的制度约束。结果，CIO 看到失去控制的成本吞噬着 IT 预算的剩余部分。许多 telcos 正在多级存储环境中采用数据归档来减少总体存储成本，将在线存储替换为离线磁盘或磁带存储。但是，在分散化的信息环境中一致地应用归档策略非常困难。归档策略最好在业务对象级别上采用。例如，由于存在法律争议，可能与特定客户需要关联的数据存储在主要存储中。此数据必须跨所有业务对象（客户数据、订单数据、库存数据、网络事件等）进行管理。在零散化的环境中，没有数据治理，很难（甚至不可能）隔离订购和结算系统或数据仓库中的特定表和列，以及将它们与特定业务对象相关联。数据治理提供了建立企业级业务对象定义的准则，业务对象分散在各个业务职能部门、地理区域和收购的实体中。

最后，典型的 telco 拥有以产品服务（比如陆上运输、DSL 和无线服务）为导向的结算系统和用户数据。因此，很难运行为使用多于一项服务的客户提供折扣的营销活动。

随着 telcos 的运作变得更加以客户为中心，而更少以产品为中心，数据治理组织需要建立策略来匹配多个系统中的用户名称。

具有大型 ERP 实现的制造商

具有大型企业资源规划 (ERP) 实现的制造商，比如 SAP 或 Oracle，面临着围绕数据治理的重大挑战 。甚至最广泛的 ERP 实现也不会涵盖整个企业。结果，企业继续保留他们 ERP 环境外的大量数据。许多公司具有针对不同业务部门、职能部门和地理区域的多个 ERP 实例。基本而言，企业数据存在零散化的趋势。

具有大型 ERP 实例的企业需要考虑数据治理的许多方面：

- *数据质量*——数据质量指的是拥有“满足用途”的数据。每个数据字段不需要是完整或准确的，它只需要在使用它的上下文内是准确的。较差的数据质量是任何大型 ERP 实现失败的主要原因。老练的从业者都同意，数据整合占据了典型 ERP 实现的成本的大约 40%。

  因为在将数据加载到 ERP 应用程序中之后就极难删除，所以必须特别关注添加到新环境的数据的质量。我们常常看到一些组织花费了数百万美元实现一次性的数据质量计划，但没有持续实施他们的数据质量工作。这些组织的数据质量会不断倒退。组织需要设定计划来持续监控数据质量，确保质量标准得到满足和延续。

- *主数据管理*——ERP 实现需要一个记录系统 (SOR) 来存储客户、供应商、物料清单、产品和会计科目表等关键实体。企业应该任命能够确保数据“适合其用途”的数据照管人。例如, 客户数据需要解决销售、客户服务和营销的需要。供应商数据需要为采购和供应链区域服务，而产品数据是研发组织关注的对象。数据治理

组织必须编写策略来最大化共享数据在整个企业的重用。例如，它必须建立一条策略，要求在整个企业使用某种地理编码格式，确保营销和 ERP 应用程序可无缝地协同工作。

- *信息生命周期治理*——存储的成本是企业寻求归档数据的重要推动因素，无论该数据是结构化的、非结构化的（比如供应商发票）还是二者的组合。此外，归档显著改善了应用程序的性能，减少了生产环境中的数据量，进而加快了数据访问速度。
- *安全和隐私*——企业需要遵守隐私制度，屏蔽敏感数据，比如国家标识符和员工工资。这需要在开发、测试和培训等非生产环境以及生产环境中执行。此需求在数据在向外发布（比如发布给外包商）时甚至更加重要。
- *元数据管理*——拥有数据字典或业务术语库来确保业务词汇得到了业务和 IT 的正确解释，这很重要/例如，词汇“到岸成本”可能标识“商店的到岸成本”、“港口的到岸成本”或“经销商处的到岸成本”。所有这些定义都可能是正确的，所以提供上下文内的定义很重要。一个合理的元数据层对于实现*数据血统*（从报告一直到来源对数据进行跟踪的能力）也很重要。

4

第 2 步：获取高层支持

与任何项目一样，获得恰当级别的高层管理人员对数据治理计划的支持很重要。IBM 数据治理统一流程的第 2 步解决此问题，需要问以下典型问题：

- 数据治理应该归 IT 还是业务部门“负责”？
- 我们如何利用现有的基层数据治理计划？
- 何种级别的组织应该参与数据治理？

以下是与获取高层支持相关联的子步骤：

2.1 创建虚拟数据治理工作组。

2.2 获取 IT 和业务部门内高级管理人员的支持。

2.3 识别数据治理的负责人。让我们更详细地分析一下每个子步骤。

2.1 创建虚拟数据治理工作组

与任何重要的尝试一样，数据治理很少始于一个自顶向下的计划。它由组织内一些关注更好的数据管理方式的具有类似意向的个人发起。例如，大型制造商的数据治理组织由来自数据架构、风险管理、记录管理、商业智能、数据照管和财务小组的一群人发起。这些人向组织的不同部门报告，但他们都在尽力解决类似的数据问题。

数据架构团队希望更好地协调 IT 与业务。风险管理想要围绕数据隐私的更好的策略。记录管理团队希望为电子文档建立保留策略。商业智能和财务团队处理企业报告中的数据质量问题。最后，首席数据照管人希望确保业务为整个照管计划提供了正确的支持级别。这个团队开始两周会一次面。随着时间的推移，他们建立了一个数据治理工作组。

2.2 获取 IT 和业务部门内高级管理人员的支持

企业治理提供了一种有用的类推法来在数据治理流程中尽早并频繁地解释参与的利益相关者的重要性。在企业治理的顶点是董事会，它代表着股东们的利益。董事会负责设置策略来确保企业高层官员执行合适的治理。类似地，数据治理只有在流程吸引 IT 和业务部门正确的利益相关者时才会成功。

识别利益相关者的流程非常重要。作为一般规则，任何依靠数据来实现有效性能的职能部门人员都是利益相关者。在大部分组织中，IT 组织将参与进来，尤其如果它包含专注于数据架构和商业智能的团队。首席营销官常常因“共同参与”整个企业的客户数据管理而首当其冲。类似地，财务常常是关键的利益相关者。首席信息安全官 (CISO) 也可以参与进来，以及来自关键业务部门的代表。最后，一些关键的职能部门

将希望参与数据治理流程。这些职能部门因行业不同而不同，它们包含银行中的营销、财务和风险管理，保险公司内的保险精算师、保险签署和索赔，以及制造业内的供应链。

数据治理流程需要通过展示合理的数据治理的收益来吸引这些利益相关者。介绍度量指标的第 11 章将更详细地探讨此主题。

2.3 识别数据治理的负责人。

我们可以将此子步骤放在第5步“建立组织蓝图”中，而不放在第2步中的这里。但是，我们希望强调在数据治理流程中尽早确定负责人的重要性。与任何重要的内容一样，职权明确是确保成功的数据治理计划实现的关键。

因为数据是业务的命脉，负责数据治理可能伴随着组织内的政治问题。在许多情形下，负责数据治理的可能是对该主题最有激情或拥有先动优势的人。尽管如此，可通过多种方式确定数据治理的合适负责人：

- *按组织*——数据治理可以组织为一种企业智能，或者可以由特定的业务经理、多个业务经理负责，或者同时符合这二种情形。数据治理作为企业职能的优势是，数据治理计划将在整个企业中保持一定的一致性。缺点在于，数据治理可能被视为离业务经理的需要太远。

  让业务经理负责数据治理可改善该计划与业务紧密链接的特征。但是，存在多个业务经理开发他们自己的数据治理计划的风险。此方法还使得更难实现企业级计划，比如主数据管理。作为折衷，某种混合方法（同时企业和业务经理都会参与）适用于许多组织。

- *按职能部门*——在大部分组织中，一些关键职能部门可能会负责数据治理：
  - » *安全*——在早些年，数据治理主要用于确保安全和隐私，首席信息安全观是该计划的主要支持者。在过去几年来，数据治理计划由 CISO 负责的情形越来越少，但安全和隐私仍然是该计划的重要部分。
  - » *风险*——首席风险官逐渐成为一些数据治理计划的关键支持者，尤其是在银行内。全球金融危机已让银行相信，他们需要值得信赖的数据来支持合理的风险管理。
  - » *营销*——首席营销官继续在寻找外部和内部数据的新来源，以获得客户行为和竞争情报的更透彻洞察。许多行业的全球化在访问和控制理解新市场和国外竞争威胁所需的不同的数据（比如货币、语言和关税）方面带来了新的挑战。
  - » *其他职能领域*——这常常因行业不同而不同。例如, 钻探公司的钻探部门和副总裁将负责管理石油和天然气公司内的油井数据。
  - » *信息技术*——在过去几年来，首席信息官 (CIO) 和 IT 组织越来越多地开始负责数据治理。此方法的不足在于，业务（而不是 IT）处于定义数据治理规则和策略的最佳位置。

以下是可能在 IT 部门领导数据治理的角色：

- 企业数据架构师
- 经理、计划经理、总监或信息管理副总裁
- 经理、计划经理、总监或商业智能副总裁

企业数据仓库可能将成为数据治理问题的避雷针，因为存在与数据质量相关联的所有挑战和缺乏值得信赖的信息。回想一下二十世纪90年代末期失败的许多仓库项目。它们为什么失败？数据存储到了仓库中，报告也可用，但是，数据的质量常常很差，所以依靠数据的业务部门拥有糟糕的体验，因此产生了负面印象。这就是“无用输入，无用输出”的一个经典示例。

进一步讲，新数据仓库项目正寻求通过实现治理流程来改进数据质量，将以前的体验从消息更改为积极。商业智能团队的一种常见抱怨是，多个提供给高级管理层的报告具有不一致的数据。”尽管如此，尽早、在源头治理数据比在晚期、在数据仓库中治理数据要有效得多。

许多组织拥有也履行其他职能的兼职数据治理所有者，包括企业数据架构、风险管理和企业安全。但是，全职数据治理职位越来越多，组织认识到了数据作为企业资产的价值。附录C给出了数据治理官的一个工作职位示例。该工作职位专注于数据质量、数据照管、度量指标、报告结果和协调业务。

5

第 3 步：执行成熟度评估

本章首先开始对成熟度模型进行一般性讨论。由 Software Engineering Institute (SEI) 在 1984 年开发，容量成熟度模型 (Capability Maturity Model, CMM) 是一种用于开发和完善组织的软件开发流程的方法。CMM 描述了一种 5 级毕业路径，如图 5.1 所示。此路径提供了一个确定操作优先级的框架、

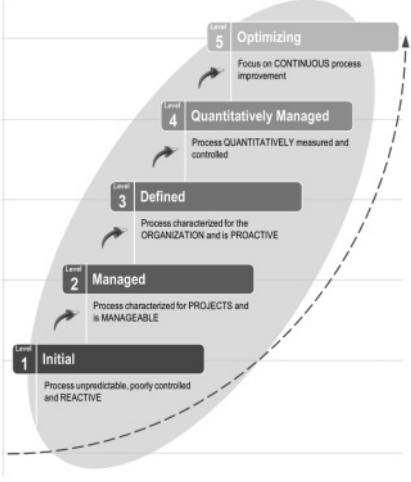


图 5.1：容量成熟度模型。

一个起点、一种通用语言和一种度量进度的方法。最终，这个结构化的元素集合提供了一系列通往想要的最终成熟度状态的稳定、可度量的进度。

在成熟度级别 1（初始），流程通常是临时的，环境也不稳定。成功反映组织内个人的能力，而不是成熟流程的使用。尽管处于级别 1 的组织常常会生成有效的产品和服务，但他们常常会超出预算和项目时间表。

在成熟度级别 2（管理），成功是可重复的，但流程可能无法为组织内所有的项目而重复。基本的项目管理有助于跟踪成本和时间表，而流程学科有助于确保保留了现有的实践。当这些实践就绪之后，项目就会依据它们所备案的计划执行和管理。但是，仍然存在超出成本和预计时间的风险。

在成熟度级别 3（定义），组织的标准流程集用于在整个组织中建立一致性。对组织的标准流程集中的项目标准、流程描述和规程进行调整，以适合特定的项目或组织部门。

在成熟度级别 4（定量管理），组织设置流程和维护的数量质量目标。所选的子流程对整体流程性能具有重大贡献，使用统计技术和其他量化技术来控制。

最后，在成熟度级别 5（优化），量化的流程改进目标被明确地建立并继续修订以反映不断变化的业务目标，以及用作管理流程改进的条件。

IBM 数据治理成熟度模型是向前的重要一步，因为它有助于指导其他利益相关者如何帮助提高战略效率。基于 IBM 数据治理委员会的成员的输入而开发，该成熟度模型定义了整个组织中需要参与业务治理数据（例如，敏感的客户信息或财务细节）的治理和度量的人员范围。

IBM 数据治理成熟度模型基于 11 个数据治理成熟度类别来度量数据治理能力，如图 5.2 所示。

图 5.2：IBM 数据治理成熟度模型。

1. *数据风险管理和合规*是一种方法，识别、定性、量化、避免、接受、减轻或转移风险。
2. *价值创建*是一个流程，定性和量化数据资产来使业务能够最大化数据资产所创建的价值。
3. *组织结构和感知*指业务和 IT 之间的相互责任水平，识别受托责任以在不同的管理级别治理数据。
4. *照管*是一种质量控制学科，旨在确保为资产管理、风险减轻和组织控制而对数据进行照管。
5. *策略*是想要的组织行为的书面表达。
6. *数据质量管理*指度量、改进和验证生产、测试和归档数据的质量和完整性的方法。
7. *信息生命周期管理*是一种系统、基于策略的信息收集、使用、保留和删除方法。

8. *信息安全和隐私*指组织用于减轻风险和保护数据资产的策略、实践和控制。
9. *数据架构*是结构化和非结构化数据系统和应用程序的架构设计，实现数据针对合适用户的可用性和分配。
10. *分类和元数据*指用于创建业务和 IT 词汇、数据模型和存储库的通用语义定义的方法和工具。
11. *审计信息日志和报告*指监控和度量数据价值、风险和数据治理有效性的组织流程。

这 11 个数据治理类别可分为 4 个相互关联的组：

- *成果*是数据治理计划的预期结果。
  这些结果可能专注于减少风险和提高价值，而后者是由减少成本和提高收入所推动的。
- *促成因素*包括组织结构和感知、策略和管理。
- *核心学科*包括数据质量管理、信息生命周期管理以及信息安全和隐私。
- *支持性学科*包括数据架构、分类和元数据，以及信息日志和报告。

这些类别存在着紧密的联系。例如，一个组织可能希望关注作为其数据治理计划成果的价值创建，以对现有客户的交叉销售和向上销售为基础。该企业将希望部署数据照管人角色来改进其客户数据的质量。该企业也将希望通过“客户”的 SOR 实现一个企业主数据管理计划。最后，该企业将需要设立一个数据治理组织来推动计划实施，设置围绕客户属性定义和跨组织边界的客户数据共享的策略。

只有首先承认您具有问题，才可以开始解决该问题。开始数据治理的最佳方式是沿以下路径执行评估：

- 当前状态：我们目前处于什么位置？
- 未来状态：我们希望在未来处于什么位置？
- 路线图：我们需要哪些人员、流程、技术和策略接话来填补当前和未来状态之间的差距？

图 5.3 使用 IBM 数据治理成熟度模型的框架提供了当前和未来状态之间的差距示例。成熟度级别直接对应到容量成熟度模型。执行数据治理成熟度评估的最佳方式是与来自 IT 和业务部门的正确参与者举行一场研讨会。

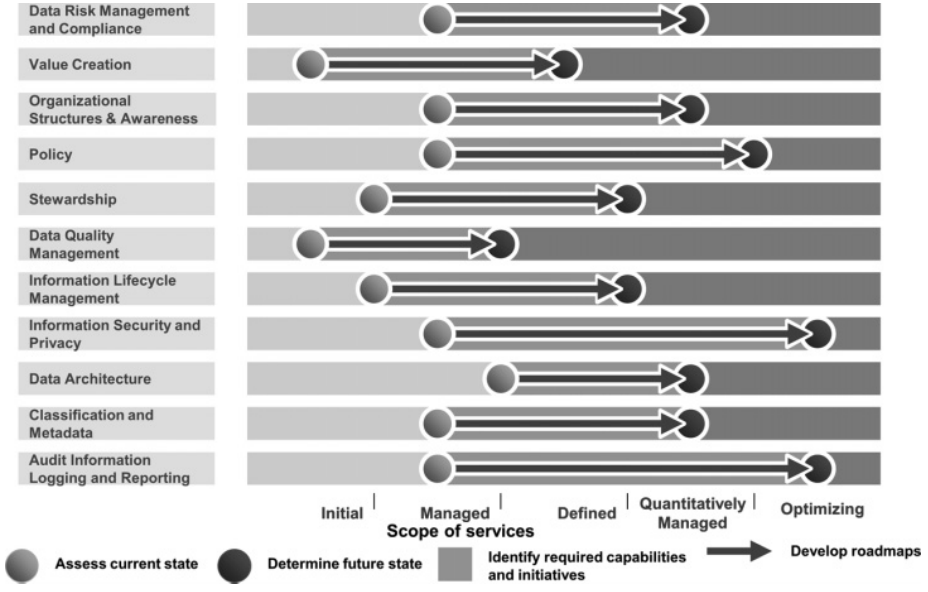


图 5.3：示例数据治理成熟度评估。

以下是执行数据治理成熟度评估的过程中所涉及到的子步骤：

3.1 定义评估的组织范围。

3.2 定义想要的未来的数据治理状态的时间范围。

3.3 定义要评估的数据治理类别。

3.4 确定业务和 IT 部门中正确的研讨会参与者。

3.5 执行数据治理成熟度评估研讨会。

3.6 与高层管理人员沟通评估结果。本章剩余部分提供这些子步骤的详细讨论。

3.1 定义评估的组织范围

除了在非常小的公司，否则不太可能对整个企业提供一个评分。甚至对于小型公司，可能也有一个具体的部门或业务职能将从数据治理计划获得最高价值。因此，您可能决定仅对一个地理区域、业务部门或职能部门（比如供应链或销售）执行初始数据治理成熟度评估。在一天结束后，这只是向关键利益相关者的一种数据治理计划“内部推销”。

3.2 定义想要的未来的数据治理状态的时间范围

定义您希望在哪个时间范围内更改数据治理成熟度评估，这很重要。这个时间范围应该不会短到很难生成有意义的更改，也不会长到由于缺乏切实的结果而使组织分散注意力。

大部分组织倾向于挑选 12 到 18 个月的时间段。例如, 一家银行可能决定专门关注数据治理成熟度评估的风险职能。做出了该决策之后，该银行可以明确地确定它需要在 18 个月时间段内对数据治理成熟度评估进行想要的更改。

因此，该银行然后可能必须决定如何在 18 个月内，将风险职能的分类和元数据能力评估值从“1”提升到“3”。从价值创建角度讲，风险团队将希望能够利用改进的元数据功能向关键制度性报告论证数据血统。

3.3 定义要评估的数据治理类别

依赖于您组织内对数据治理的渴求，您可能将决定仅从 IBM 数据治理成熟度模型类别子集开始。例如，您可能决定今关注您企业内一个部门。因此，您可能决定安全和隐私能力不属于评估范围，因为该职能由企业处理。或者，可能事实证明您的数据治理计划需要更多地关注结构化数据，所以任何围绕记录管理和非结构化内容的探讨都将超出信息生命周期管理能力的范围。

3.4 确定业务和 IT 部门中正确的研讨会参与者

业务和 IT 的良好结合是执行合理的数据治理成熟度评估的必要条件。没有正确的职能和部门列表，但您需要确保通过正确的参与者来最大化获取对研讨会中任何建议的适当支持的机会。

典型的 IT 参与者可能包括信息管理团队、商业智能和数据仓库领导、企业数据架构师、记录管理团队，以及安全和隐私专业人员。业务参与者可能包括来自销售、财务、营销、风险和其他依靠数据来实现有效性能的相关工作职能或部门的代表。这个参与者群体的典型职责包括设置策略、性能分析、生成报告、开发模型、设计业务流程和管理数据照管。

3.5 执行数据治理成熟度评估研讨会

确定了数据治理成熟度评估的范围之后，是时候举行研讨会了。研讨会的持续时间在两天到几周内，具体取决于组织的需要。在许多情况下，组织可能决定将关键利益相关者集中在一起举行为期一两天的研讨会，后跟一系列对关键利益相关者的访谈。

3.6 与高层管理人员沟通评估结果

完成数据治理成熟度评估之后，将结果与关键 IT 和业务部门利益相关者共享很重要。这样，您可以开始在组织上就关键问题达成一致，比如组织一致性、元数据和数据质量缺乏。您也可以帮助高级领导开始就潜在的后续步骤达成一致或安排负责人的过程。我们将在下一章介绍创建路线图的流程。

附录 D 包含一个示例数据治理成熟度评估调查问卷。

6

第 4 步：创建路线图

3 个子任务可方便数据治理路线图的开发：

4.1 总结数据治理成熟度评估的结果。

4.2 列出填补评估中强调的差距所需的关键人员、流程和技术计划。

4.3 基于关键计划的优先级创建路线图。

4.1 总结数据治理成熟度评估的结果

完成数据治理成熟度的评估后，您将得到所评估的每个类别的 3 个数据点：

- 当前状态评估（低 = 1，高 = 5）
- 想要的未来状态评估（低 = 1，高 = 5）
- 当前状态与想要的未来状态之间的偏差

一定要认识到，评分“1”在本质上并不差，评分“5”也不一定很好。数据治理组织必须与 IT 和业务部门利益相关者合作，（最好）开发一个业务案例来确定提高对想要的未来状态中给定类别的评分是否可行。

4.2 列出填补评估中强调的差距所需的关键人员、流程和技术计划

图 6.1 显示了一家建立了数据治理计划，专注于客户中心性，以营销部门为关键支持者的银行示例。该银行希望通过增加其向每个零售客户销售的产品数量，提高收入。该银行的营销部门认为其零售客户通常在该银行只有一个帐户——获取存款帐户或贷款帐户。

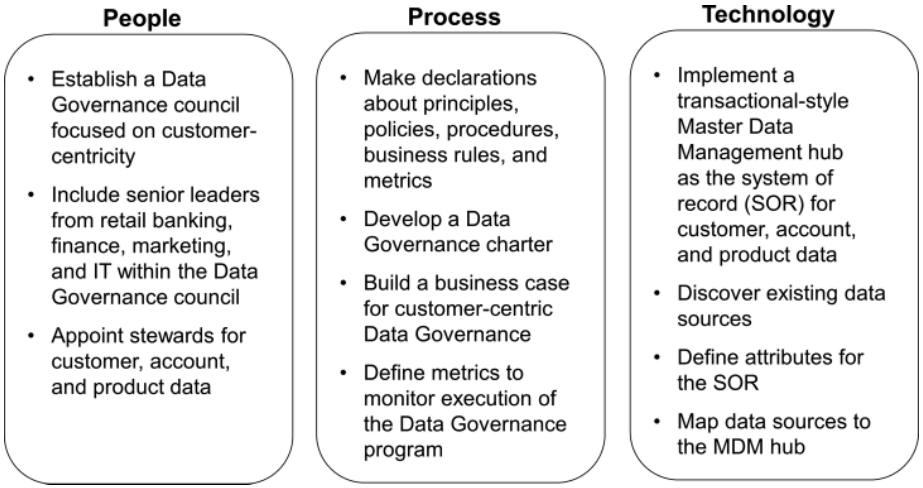


图 6.1：银行的关键数据治理计划列表。

该银行决定围绕数据治理实现一系列人员、流程和技术计划。

在人员方面，该银行需要建立将客户中心性作为近期业务目标的数据治理委员会。该银行将客户、帐户和产品优先确定为关键数据领域。相应地，数据治理委员会需要包含来自银行、财务和营销领域的成员作为关键业务支持者，当然还有 IT。

该银行还需要在零售银行中任命照管人来监督这些数据领域，持续管理数据质量。

在流程方面，数据治理组织需要对原则、策略、规程、业务规则和度量指标进行声明。（请参阅附录 E 了解有关数据治理声明的更多信息。）它必须采用一个关注客户中心性的章程，必须建立关键度量指标，比如每个客户购买的产品数量。它还需要开发一个业务案例来论证整个计划的合理性。

最后，在技术方面，数据治理计划需要监督主数据管理 (MDM) 中心的实现，该中心将作为客户、帐户和产品数据的记录系统 (SOR)。数据治理计划需要查询现有的数据源，定义主数据的属性，以及将源系统的数据模型映射到 MDM 中心。

4.3 基于关键计划的优先级创建路线图

图 6.2 显示了数据治理计划路线图，利用一个 18 个月的时间段来论证初始结果。银行的数据治理流程

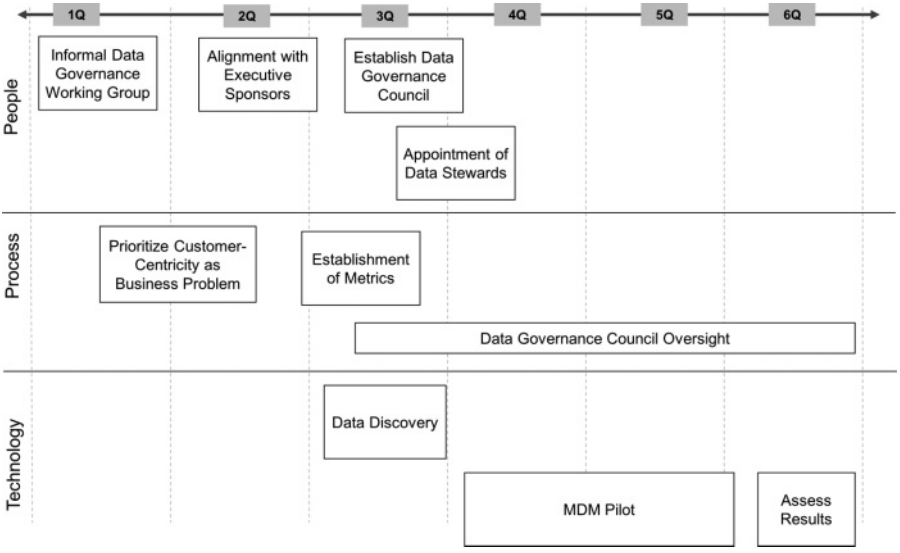


图 6.2：银行数据治理计划的示例路线图。

从非正式的数据治理工作组的组建开始，该工作组由来自关键职能区域（比如零售银行、财务、营销和 IT）的中层经理组成。考虑到客户数据分散在多个数据源，营销和零售银行团队将把客户中心性优先设置为数据治理计划的关键业务问题。该银行必须实际地花费两个季度来适当完成高级管理人员协调过程。

在接下来的两个季度，数据治理委员会将正式建立，将任命数据照管人，协商关键度量指标。与此同时，数据治理团队将查询现有数据源，就 SOR 的属性达成一致。MDM 试点将在第四季度开始并持续另外两个季度。

最后，该银行应该能够评估计划在 18 个月内的结果。整个流程将由数据治理委员会监督。

7

第 5 步：建立组织蓝图

SIBM 数据治理统一流程的第 5 步定义组织数据治理计划的最佳方式，以实现最大成果。以下是此不走的关键子步骤：

5.1 定义数据治理章程。

5.2 定义数据治理的组织结构。

5.3 建立数据治理委员会。

5.4 建立数据治理工作组。

5.5 确定数据照管人。

5.6 举行数据治理委员会和工作组定期会议。

这些子步骤中的每一步将在下面的一节中探讨。

5.1 定义数据治理章程

数据治理章程类似于企业的公司条例。该章程阐明计划的主要目标及其关键利益相关者，依据角色和职责、决策权利和成功度量标准。附录 B 提供了一个示例数据治理章程。

5.2 定义数据治理的组织结构

数据治理的最优组织结构是一种 3 层结构。数据治理委员会，位于组织的顶点，包含高层利益相关者。下面的一层是数据治理工作组，由负责定期治理数据的成员组成。最后，数据照管社区负责每天实际处理数据。

5.3 建立数据治理委员会

数据治理委员会由计划的高层支持者组成。委员会定义数据治理愿景和目标，在组织内跨业务和 IT 进行协调，设置数据治理计划的总体方向，在发生策略分歧时进行协调。

取决于预计的计划成果，数据治理委员会将由首席信息官、信息管理副总裁、首席信息安全官或首席风险官主持。此委员会也将包含来自财务、法律、HR 团队的职能部门代表，以及来自各种将数据视为企业资产的业务线的代表。这些高层管理人员是数据治理计划的所有拥护者，会确保在整个组织内获得支持。

5.4 建立数据治理工作组

数据治理工作组是组织内委员会下面的下一个级别。工作组每天运行数据治理计划。

它还负责监督数据照管社区。

数据治理工作组由数据治理领导主持。而这位领导也可能在数据架构、信息安全或风险小组内拥有其他角色，许多组织现在都任命了全职经理和数据治理总监。

5.5 确定数据照管人

数据照管人理想情况下向业务部门报告，履行数据的保管职责。数据照管人每天解决具体的问题和担忧，定义组织内外的数据。（此主题将在第 12 章更详细地介绍。）

5.6 举行数据治理委员会和工作组定期会议

数据治理委员会举行会议来设置数据治理策略，跟踪数据治理计划的性能。该委员会（包括高层领导）定期会面，但不一定是经常会面。典型的委员会会议安排为每月或每季度举行一次，持续一两个小时。

数据治理会员会会议议程中的示例主题包括：

- 检查数据治理计分卡。（此主题将在第 11 章更详细介绍。）
- 签署记录管理战略，包括文档分类、保留时间表和电子查询 (eDiscovery)。
- 与首席信息安全官联合签署查询和保护隐藏的 PII 的策略。
- 协商客户和产品数据的总体高层支持者。

数据治理工作组包含中层管理人员。它更频繁地举行会议，通常为每两周一次。工作组会议可能持续三四个消息，具体取决于具体计划的紧急程度。

以下是数据治理工作组会议议程中的一些示例主题：

- 就“客户”的 SOR 的属性达成一致。
- 就在两个部门有能力更新相同属性时的流程达成一致。
- 创建业务规则来匹配、合并和链接相关的客户记录。
- 联合法律部门一起检查 eDiscovery 流程。

图 7.1 描述了制造商的一个示例数据治理组织。这家制造商采用了一种三层数据治理组织，包含一个委员会、一个工作组和数据照管人。数据治理委员会由首席信息官主持。该委员会也包含关键职能部门利益相关者，比如首席财务官、首席风险官和供应链小组高级副总裁。最后，关键业务部门的领导也是委员会的成员。

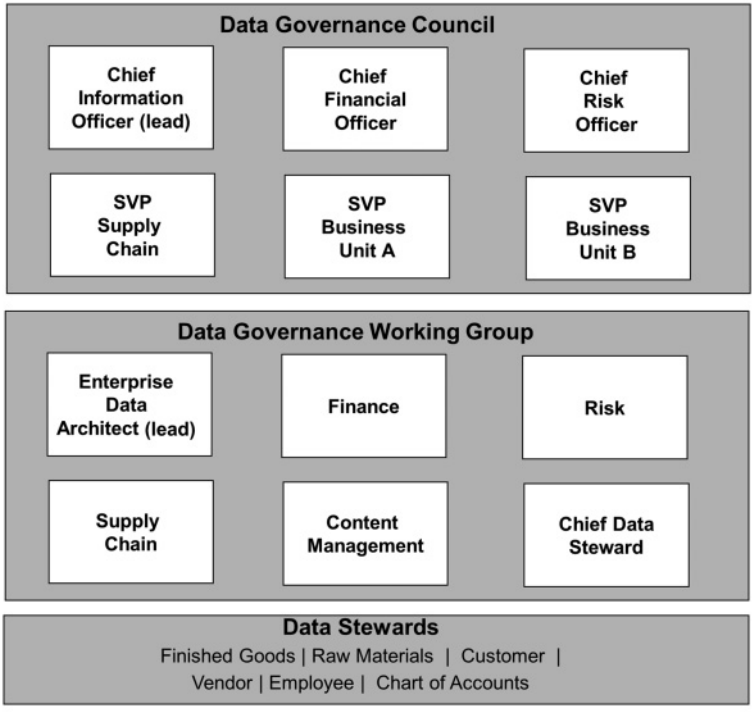


图 7.1：一家制造商的示例数据治理组织。

在接下来的一层，数据治理工作组由企业数据架构师主持，他设置小组的议程并主持会议。工作组还包含来自财务、风险和供应链区域的利益相关者。考虑到非结构化数据治理的重要性，工作组也包含内容管理区域的成员。首席数据照管人（监督数据照管计划）也是数据治理工作组的成员。

数据照管社区由负责关键主题区域的数据照管人组成。成品、原材料和供应商数据的照管人向供应链小组报告。客户数据照管人向销售部门报告。员工数据照管人向人力资源部门报告。最后，会计科目数据照管人向财务部门报告。

图 7.2 描述了一个中型银行的一个示例数据治理组织。此银行也采用了一个三层数据治理组织，包含一个委员会、一个工作组和数据照管人。

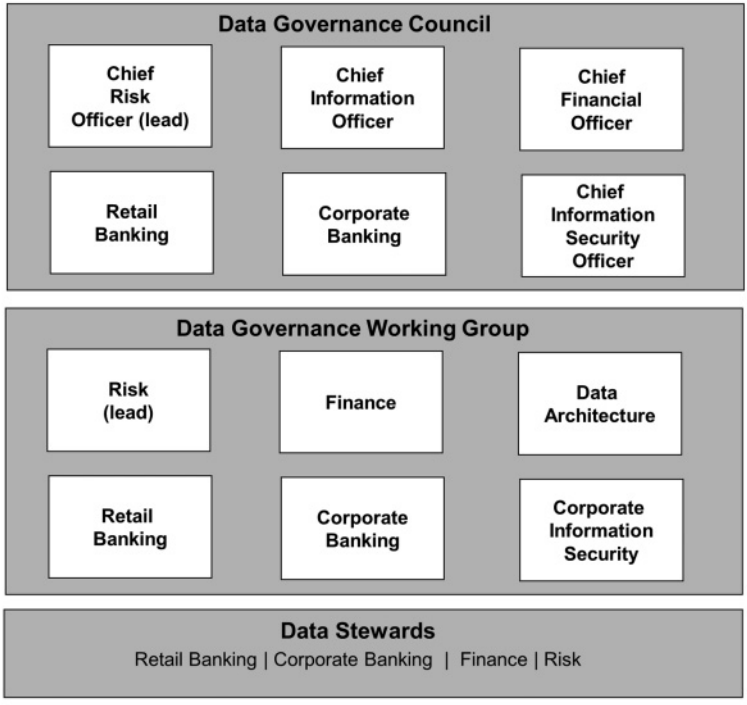


图 7.2：一家中型银行的示例数据治理组织。

该银行的数据治理委员会由首席风险官主持，因为它密切关注改进风险数据的质量。数据治理委员会也包含关键职能部门利益相关者，比如首席信息官、熟悉财务官和首席信息安全官。最后，零售银行和企业银行部门领导也是委员会的成员。

在下面一个级别，数据治理工作组由风险部门副总裁主持，他设定小组的议程并主持会议。工作组还包含来自财务、数据架构和安全区域的利益相关者。该工作组也包含来自零售银行和企业银行的中层代表。

该银行采用了一种由组织和职能区域混合的数据照管形式。该银行在零售银行、商业银行、财务和风险区域拥有数据照管人，他们以兼职形式执行质量控制。

在零售银行内，数据照管人负责识别在多项产品中（比如活期账户、贷款账户和信用卡）客户与银行的关系。这些关系不仅需要在个人级别上识别，还需要在家庭级别上识别。例如，如果 Smith 先生和 Smith 太太都在该银行拥有帐户，需要在单一视图中采集他们的数据。

在企业银行中，数据照管人负责维护法律层次结构。他们将使用 D&B D-U-N-S 编号等工具，这些编号是用于识别全球数百万企业的唯一的 9 位数序列。D&B D-U-N-S 编号将使数据照管人能够确保相同企业集团内的两家公司将包含在相同的法律层次结构中。财务数据照管人负责确保财务数据的质量。类似地，风险数据照管人负责确保用于风险计算的数据的质量。

8

第 6 步：创建数据字典

组织内的一个部门称为“收入”，而另一个部门称为“销售”。两个部门是否指同一种活动？一个词语谈论的是“顾客”，另一个谈论的是“用户”或“客户”。它们是两种不同的分类，还是相同分类的不同词汇？

业务元数据与业务词汇的定义相关联。业务元数据对信息的最终用户至关重要。它使这些最终用户能够确信他们依靠来制定业务决策的数据是完全是他们所需的数据。业务元数据的有效管理科确保在整个组织应用相同的描述性语言。

数据字典（或*业务术语库*）是一个存储库，包含将业务和 IT 的共同定义集中在一起的关键词汇的定义。组织部署数据字典来确保业务词汇在上下文内良好定义。但是，要生效，数据字典需要填入已由相关业务区域达成一致的业务词汇。

许多组织受到了整个企业内的业务词汇的不一致性的困扰。这种不一致性可能由并购等事件或定义常见业务词汇表的系统、孤立方法所导致。根源常常在于缺乏有效的数据治理和照管计划。

数据字典的创建涉及到 8 个关键的子步骤：

6.1 选择一个数据领域。

6.2 安排数据照管人来维护关键业务词汇。

6.3 识别关键数据元素。

6.4 从现有的词汇术语表创建数据字典。

6.5 填充数据字典。

6.6 链接业务词汇与技术工件。

6.7 支持数据治理审计、报告和日志需求。

6.8 整合数据字典与应用程序环境。本章剩余部分更详细介绍

这些子步骤。

6.1 选择一个数据领域

组织需要挑选一个数据领域，比如风险或财务。最佳实践是挑选拥有大量围绕数据定义的问题，愿意使用围绕数据字典开发的数据治理计划的领域。

6.2 安排数据照管人来维护关键业务词汇

选择了数据领域之后，数据治理组织需要确保向与词汇相关的小组分配了照管人。这些照管人将负责定义的持续维护，

确保业务和 IT 之间恰当地协调一致。图 8.1 显示了 IBM InfoSphere Business Glossary 内的一个数据照管示例。在此示例中，安排了 Scott Montgomery 担任照管人。我们可以查看他管理的所有资产，以及他所属的组织和联系人信息。

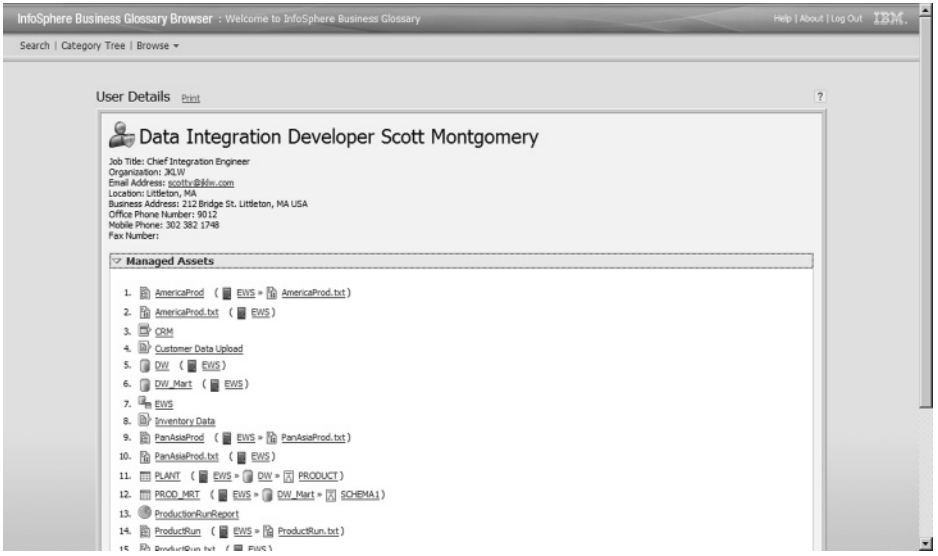


图 8.1：IBM InfoSphere Business Glossary 内的数据照管示例。

6.3 识别关键数据元素

对于关键数据元素的识别，让我们看看一个与每个组织相关的示例。虚构的公司 ABC 有一份报告表明它拥有 250,000 个客户。但是，问题在于如何定义词汇“客户”？营销部门可能将它定义为包含潜在客户。销售部门可能将它定义为有机会参与 CRM 系统的各方。财务部门可能仅希望包含在过去 12 个月内购买了产品的各方。最后，您将具有多个子公司的跨国公司视为一个还是多个客户？

图 8.2 显示了已由 IBM InfoSphere Business Glossary 内的数据照管人记录的“高价值客户”的定义。该术语库不仅显示词汇的当前定义，还显示以前的定义和更改该定义的人的姓名。

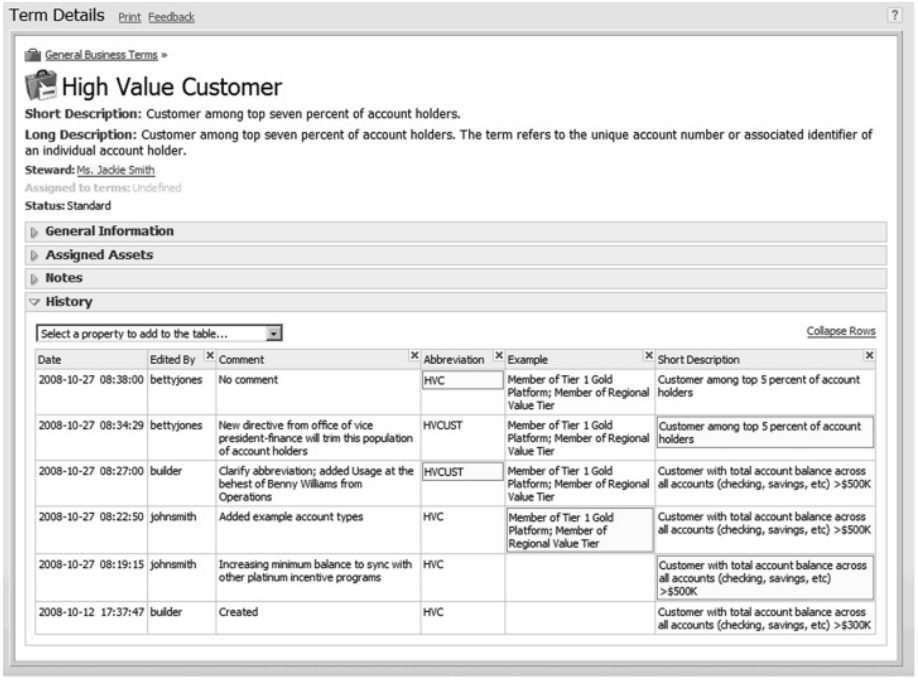


图 8.2：IBM InfoSphere Business Glossary 显示了词汇“高价值客户”的当前和以前的定义。

在该示例中可以看到，词汇“高价值客户”指前 7% 的帐户所有者中的任何客户。但是，该词汇在过去拥有多个其他定义，包括任何在所有帐户（活期、储蓄等）上具有超过 30 万美元以上余额的客户。

一家欧洲电信服务提供商具有此挑战的第一手经验，组织内的不同职能区域无法就词汇“活跃用户”的一致定义协商一致。例如，结算部门将活跃用户定义为在过去 30 天内收到一个帐单的人。网络部门将活跃用户定义为在过去 30 天内使用了运营商的网络的人。这两个定义之间的关键区别关系到签署了服务的用户，而不是更换了 SIM 卡并且长期在运营商网络外部漫游的用户。

数据治理团队在此情况下能够建立一致的定义集合。它让人们摆脱他们的孤岛，以共享信息和看到更广泛的业务概念和驱动因素。数据治理组织然后也在组织的其他部门内重用这些共同定义。

6.4 从现有的词汇术语表创建数据字典

IBM InfoSphere Business Glossary 提供了一种易于使用、基于 Web 的用户界面来创建、管理和共享一个受控的词汇表。存在着一些针对电信、金融服务、零售、保险和医疗行业的 IBM InfoSphere Business Glossary 包。这些包为启动术语库项目提供了丰富的行业内容。组织然后可以自定义这些术语库来满足其具体需要。

6.5 填充数据字典

下一个子步骤是使用协商一致的业务词汇填充数据字典。IBM InfoSphere Business Glossary 为业务词汇表及其规则和关系的定义、管理、搜索和浏览提供了一种基于 Web 的门户。从 IBM InfoSphere Business Glossary 的初始页面，业务用户可使用类别树或搜索功能来查找信息。类别可基于“财务”、“人力资源”和“产品”等业务分类模式或基于地域和位置来定义。

图 8.3 显示了 IBM InfoSphere Business Glossary 内的词汇的类别树视图的一个示例。在此示例中，词汇“高价值客户”列在“客户交互分析”之下。

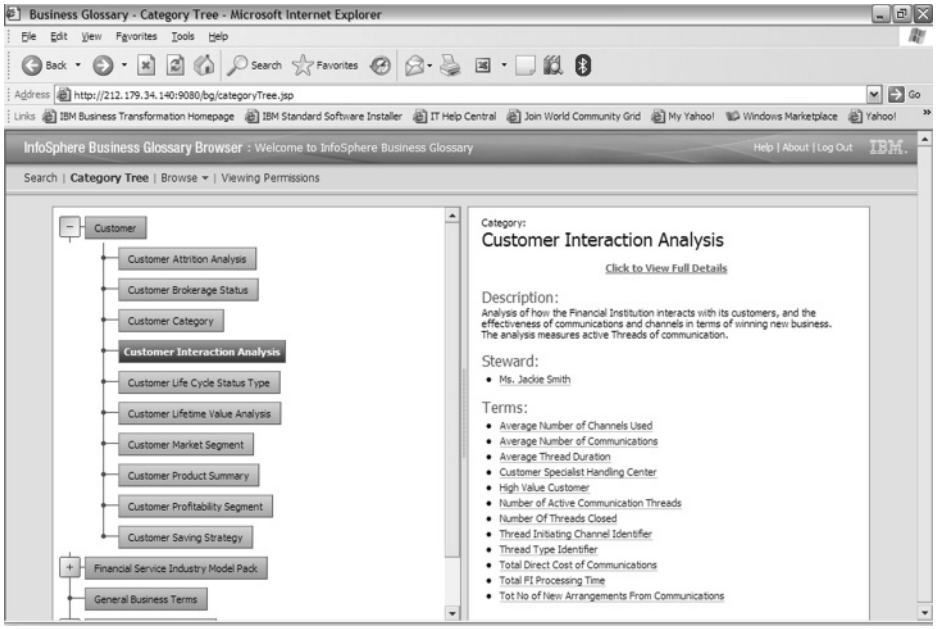


图 8.3：IBM InfoSphere Business Glossary 的类别树。

业务用户也可以使用 IBM InfoSphere Business Glossary 浏览器搜索和浏览术语库中定义的含义。一个示例如图 8.4 所示。

图 8.4：IBM InfoSphere Business Glossary 浏览器。

6.6 链接业务词汇与技术工件

如白皮书 *The Business Value of a Business Glossary*（IBM Software Group 和 Lowell Fryman，2008 年 10 月）中所述，就业务定义达成一致后，数据架构师需要在词汇和技术工件（比如数据库表和列）之间建立链接。例如，数据架构师可以将词汇“供应商”链接到数据库中的 SUPPLIER 表，将词汇“供应商编号”链接到 SUPPLIER 表中的 SUPP_NUM 列。这些链接有助于在业务和 IT 之间建立双向通信，促进有效的数据治理。

业务用户可从一个词汇下钻来查找技术数据源。与此同时，处理数据源或 ETL 工作或创建业务报告的技术用户可以理解所使用数据的业务上下文。

6.7 支持数据治理审计、报告和日志需求

业务术语总是容易变化。今天的“高价值客户”的定义可能在明天又不同。随着业务需求不断演化，一个词汇的可接受定义也会变化。能够了解更改历史、更改的内容和执行更改的人，这与更改本身一样重要。

图 8.2 提供了一个具有历史的定义示例。这样的历史对于数据治理协议至关重要，因为它增加了信息的可信度和易理解性。某个定义更改的原因可能影响我们报告它的方式或我们收集它的支持数据的方式。（此方面也由第 5 章中介绍的 IBM 数据治理成熟度模型中的“审计信息、日志和报告”类别解决。）记录系统还可以确保遵守 Sarbanes- Oxley Act 和 Basel II 等制度。

6.8 整合数据字典与应用程序环境

业务理解至关重要，而时间压力常常使业务用户无法利用可用的资源。例如，您可能阅读了一封电子邮件或白皮书，遇到了一个含义模糊的词汇或短语。您不确定它是如何在您组织内定义或使用的。您知道您只需

打开 Web 浏览器就可在您公司的在线术语库中找到它，但这可能需要您暂停当前的任务。因此，您会推迟到以后查找该词汇。而在这之后，您可能已忘记它，失去了一些重要信息。

如果您可以理解从您所在位置获取该信息，而不会丢失上下文该多好？IBM InfoSphere 的 Business Glossary Anywhere 整合了用户桌面上的任何应用程序。所以，在您位于 Microsoft® Excel®、IBM Cognos®、电子邮件应用程序、用户手册、在线表格或其他任何地方时，您只需突出显示一个词汇并右键单击，它的定义就会立即弹出来。例如，在图 8.5 中，用户右键单击业务词汇“GL Account Number”即可获得它的定义。

图 8.5：使用 IBM InfoSphere 的 Business Glossary Anywhere 查找 Microsoft Word 中的定义的示例。

在第 6 步结束时，您应该有一个强大的数据字典，它跨业务和 IT 对齐关键词汇，链接到技术工件，并与应用程序环境整合。

9

第 7 步：理解数据

正如有人曾经说过，“您无法控制您不理解的东西。”也就是说，在治理数据之前，您需要知道您拥有哪些数据，它们位于何处，以及它们在系统之间如何关联。诚然，数据查询时任何以信息为中心的项目的一个关键预备活动，这些项目包括归档、数据隐私、主数据管理 (MDM)、数据仓库、数据连接和应用程序整合等。对于大部分组织，数据查询和分析流程高度手动，需要数个月的人工干预来查询业务对象、敏感数据、跨源数据关系和转换逻辑。结果是一个耗时且容易出错的流程，减缓了价值实现时间。

数据治理组织需要及时理解数据，以推动从更广泛的以信息为中心的计划获取业务价值。以下是“理解数据”步骤的子步骤：

7.1 理解范围内的每个数据源。

7.1.1 执行列和表级别分析。

7.1.2 通过逆向工程主-外键关系查询遗留模式。

7.1.3 识别每个来源中的关键数据元素的位置。

7.1.4 识别每个来源中的敏感数据的位置。

7.2 理解来源之间的关系。

7.2.1 理解关键数据元素在各个数据源之间的数据重叠情况.

7.2.2 发现来源之间的数据血统和复杂转换逻辑。

7.2.3 发现数据不一致性和异常。

在本章剩余部分中，将更详细地介绍这些子步骤。

7.1 理解范围内的每个数据源

查询流程中的初始步骤是理解您以信息为中心的项目中包含的每个数据源。

7.1.1 执行列和表级别分析

数据查询包括列分析和主-外键分析。列分析包括有关数据源中每一列的基本统计数据。IBM InfoSphere Discovery 自动生成统计数据，比如隐含的数据类型、模式频率、值频率、长度频率、比例、格式、基数、空计数、最小值、最大值、长度和精度。

7.1.2 通过逆向工程主-外键关系查询遗留模式

IBM InfoSphere Discovery 在您尝试确定具有超过 20 个表的数据集的实体关系 (ER) 图时，以及在备案糟糕的数据集上执行数据分析时很有用。它获取所有表中的所有值，并基于统计分析和具有专利的算法，基于实际数据值的分析自动生成一个 ER 图。图 9.1 显示了一个相关示例。

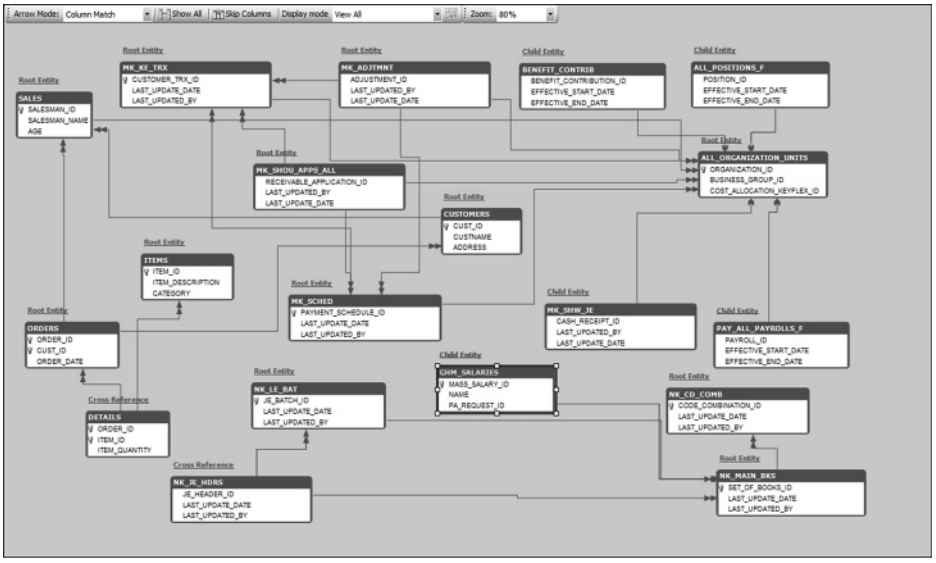


图 9.1：IBM InfoSphere Discovery 的主-外键分析示例。

数据分析师可在这个 ER 图中检查每个自动发现的键的统计信息，考虑备用键，以及修改结果。IBM InfoSphere Discovery 也可发现组合键。ER 图可导出到 IBM InfoSphere Data Architect 或 CA ERwin® Data Modeler 等数据建模工具中。

最后，IBM InfoSphere Discovery 使用主-外键关系将相关的表集中到业务实体中，比如客户、订单、材料和供应商。这些业务对象定义对 MDM 非常有用。它们也可供 IBM InfoSphere Optim™ 用于归档数据，确保数据隐私，以及创建一致的示例数据集用于测试数据管理。

7.1.3 识别每个来源中的关键数据元素的位置

关键数据元素 (CDE) 是在您将数据迁移或整合到一个新应用程序、MDM 中心或数据仓库时，希望包含在目标模式中的属性。标记 CDE 是跨所有数据源构造一个统一模式的不错起点。

一家公司的数据治理委员会要求数据建模人员“理解所有数据并标记所有属性。”数据建模人员在 3 个月之后回来了，表示存在的数据太多了，无法在合理的时间段内完成该任务。

数据治理委员会然后要求团队仅关注财务数据。该团队花了一年时间标记 40,000 个财务数据属性，向业务部门提供的有意义的返回结果很有限。数据治理委员会最后要求数据建模人员仅关注发现与其客户相关的、受到隐私制度约束的敏感数据。这个不错的示例说明了数据治理委员会提前识别 CDE 的重要性。存在的数据太多，确定数据治理计划的优先级也很重要。

IBM InfoSphere Discovery 等产品对此任务也非常有帮助，因为它们执行了两个有用的功能。首先，IBM InfoSphere Discovery 执行一种自动化的重叠分析，所以它可迅速确定哪些属性与其他数据源中的其他属性重复。这在 40,000 个属性的示例中一定非常有帮助！第二，通过整合上一章中探讨的 IBM InfoSphere Business Glossary 产品，数据分析师可使用一个业务术语库词汇标记任何属性或重叠属性组。这使业务术语库词汇可以连接到这些词汇在整个数据范围内的实际实例。此功能在建立完整的文档来将技术人员对数据的理解与业务人员对数据的理解链接起来非常有帮助。

7.1.4 识别每个来源中的敏感数据的位置

数据照管人常常被要求确保他们的数据是安全和私有的。但是，某些包含个人可识别信息 (PII) 的数据字段可能未受到必要的隐私保护。例如，一家客户的社会安全号码 (SSN) 可能位于一个名为“EMP_NUM”的字段中，SSN 的最后 4 位可能是另一个名为“PIN”的字段的一部分。结果，仅查看列标题并不足够。

IBM InfoSphere Discovery 查看实际数据。它可使一个表中的 EMP_NUM 字段实际与另一个表中的 SSN 相关联，也与 PIN 列相关联。获取之后，这些物理属性即可链接到术语库中相应的业务词汇。

7.2 理解来源之间的关系

不仅理解数据在数据库中的存在方式很重要，理解在移动和从一个来源传输到另一个来源时的数据血统联系也很重要。理解数据血统对于由商用 ETL 工具移动的数据是一个相对简单的任务。但是，在由遗留、硬编码的程序（其中文档是受限的或不存在）移动数据理解数据时怎么做？

IBM InfoSphere Discovery 通过提供一种集中、准确的方式来查询、备案和理解复杂、异构的数据来源之间的数据关系（包括转换逻辑），实现数据治理。

IBM InfoSphere Discovery 还包含一个 Unified Schema Builder 功能集。这是一个用于分析多个数据源和将这些来源组合的原型设计为一个整合的目标（比如 MDM 中心、新应用程序或企业数据仓库）中的工作台。在您必须编写 ETL 代码或配置 MDM 中心之前，Unified Schema Builder 帮助构建统一的数据-表模式原型，将已知的关键数据元素考虑在内并建议基于统计信息的匹配和冲突解决规则。

7.2.1 理解关键数据元素在各个数据源之间的数据重叠情况

标记 CDE 和执行重叠分析将有助于识别以下条件：

- 包含大部分 CDE 的数据源，它们是构造统一的模式来组合所有来源的不错起点
- 未重叠的数据源
- 包含其他数据源的数据源
- 重叠的数据源之间的一致性水平

使用 IBM InfoSphere Discovery，可同时在多个数据源上执行重叠分析。所有列会迅速与所有其他重叠的列进行快速对比，然后以电子表格形式显示，以供查看、排序和过滤。

图 9.2 显示了 3 个数据源（Regional_Branch、Community 和 CRM）中的 CDE 和重叠摘要。

图 9.2：IBM InfoSphere Discovery 的重叠和 CDE 摘要。

7.2.2 发现来源之间的数据血统和复杂转换逻辑

IBM InfoSphere Discovery 包含一个 Transformation Analyzer 组件，自动化复杂、跨来源转换以及业务规则（比如子字符串、串联、交叉引用、聚合、条件语句和两个结构化数据集之间的算术等式）的发现。图 9.3 演示了 IBM InfoSphere Discovery 自动将应用程序 1 中 Product Sales 表中的列映射到应用程序 2 中 Product Sales 表中的列所涉及的步骤。IBM InfoSphere Discovery 读取实际的数据值（不仅是列名称等元数据）以识别这些数据关系。

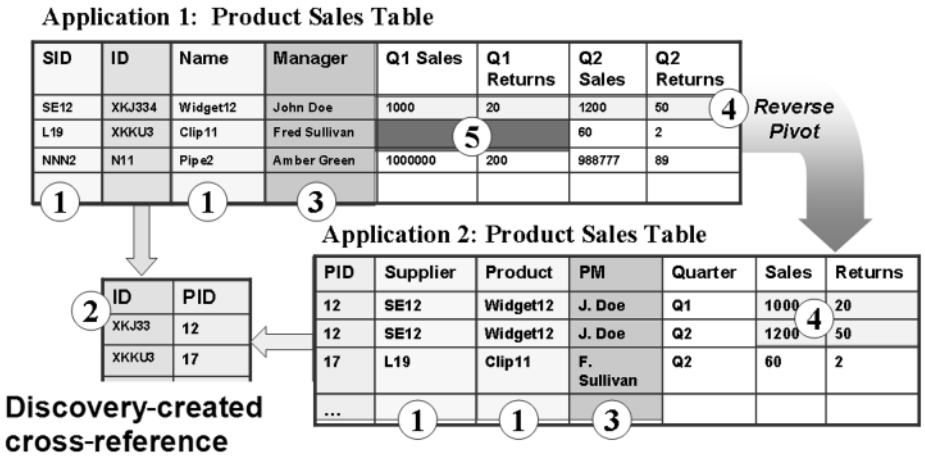


图 9.3：IBM InfoSphere Discovery 的 Transformation Analyzer 组件。

图中标有编号的元素表示发现过程的步骤：

1. 首先，IBM InfoSphere Discovery 发现用于在两个数据集之间对齐行的匹配键。在本例中，该软件发现包含提供者 ID 和产品名称的自然键与两个表相关联。这个键存储在应用程序 1 中的 SID 和 Name 列中和应用程序 2 中的 Supplier 和 Product 列中。列名称“SID”和“Supplier”自己无法表示一种逻辑关系，“Name”和“Product”也是如此。体仅通过读取数据值，而不是元数据，IBM InfoSphere Discovery 就可能发现此关系。

2. 在两个表（应用程序 1 中的 ID 和应用程序 2 中的 PID）中的主键之间创建一个交叉引用表。IBM InfoSphere Discovery 使用步骤 1 中的自然键来交叉引用主键。

3. IBM InfoSphere Discovery 发现应用程序 1 中的 PM 列由应用程序 2 中的 Manager 列、一个句点、空格和 Manager 列中的第二个标志组成。

4. 应用程序 1 中的 Q1Sales、Q1Returns、Q2Sales 和 Q2Returns 列中的值在

应用程序 2 中经过了行列变换（转换为行）。IBM InfoSphere Discovery表为变换的列生成一个独立的映射，创建一行（比如 Q1Sales 和 Q1Returns）。

5. 最后，IBM InfoSphere Discovery 在 Q1Sales 列上发现一个过滤器：只有具有非空 Q1Sales 的行在应用程序 2 中拥有相应的行。

7.2.3 发现数据不一致性和异常

因为 IBM InfoSphere Discovery 计算实际的数据值来发现转换，所以此方法还会识别可能倒是收入损失、客户不满意和罚金的不一致性。在图 9.4 中的真实示例中，该软件自动发现 AGE 列（显示保险应用中的驾驶员年龄）通过一个条件语句与第二个应用中的 Youthful_Driver 列相关联。但是，不是所有数据行都遵守所发现的规则，也就是 Youthful_Driver 列在 AGE 列小于或等于 25 时应该设置为“Y”。

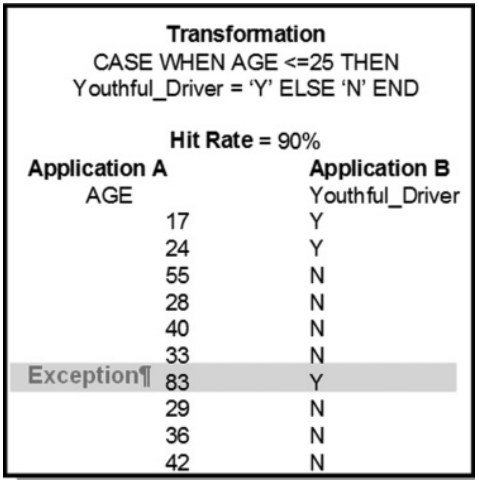


图 9.4：IBM InfoSphere Discovery 发现保险商数据内隐藏的关系。

在该示例中，一位 83 岁的驾驶员在 Youthful_Driver 列中具有值“Y”。这行数据自动标记为不遵守所发现的规则。数据照管人现在可以分析该驾驶员是否实际为 83 岁，或者是否有某种人为改动导致了违背业务规则。

10

第 8 步：
创建元数据存储库

元数据是“关于数据的数据”。它是与任何数据工件（比如其名称、位置、认识到的重要性、质量或对企业的价值，以及与企业认为值得管理的其他数据工件的关系）相关的信息。元数据形成 IT 对业务和信息基础架构如何满足业务需要的认知。尽管关注单一数据资产的元数据很重要，但它不允许设定与数据质量、货币或与整个企业的诚信相关的假设。理解数据在不同系统中流动以及它的使用的更大场景需要一种整体方法，通常称为*元数据管理*。

关于企业来源和流程的元数据可充实它们的上下文和含义，如果没有它，这些信息资产可能无法识别、不受信任以及甚至不适用。如果我们不知道在业务报告中看到的信息是如何集中在一起，如何才能信任它？如果没有任何与质量标准相关的业务规则，我们如何识别糟糕的数据质量。正是这些以及许多其他问题，使关于信息资产的合适元数据定义变得至关重要。无论它是源表和列、数据模型、ETL（提取、转换和加载）流程还是目标系统，我们都需要知道：

- 谁创建了它们？
- 它们是何时创建的？
- 它们设计来做什么？
- 它们曾经更改过吗？
- 如果它们更改，更改是否对其他信息资产有任何影响？
- 它们的质量标准是什么？

与创建元数据存储库相关的关键子步骤如下：

8.1 合并来自数据字典的业务元数据和来自发现流程的技术元数据。

8.2 确保合适的数据血统。

8.3 执行影响分析。

8.4 管理操作元数据。

下面将更详细探讨这些子步骤。

8.1 合并业务和技术元数据

数据治理计划将从数据字典生成大量业务元数据并在发现阶段生成技术元数据。尽管技术元数据对于为 IT 人员配备支持有效的业务应用程序和企业资源的工具至关重要，但将它链接到业务元数据会在业务和技术团队之间产生隔阂。IT 人员需要保持与业务的联系，理解业务语言，支持与业务目标一致的基础架构。

业务和技术元数据需要存储在 IBM InfoSphere Metadata Workbench 等存储库中，以便它可跨多个项目使用。这样，当技术用户查看表、数据转换流程和将“高价值客户”数据集中在一起的起源时，他们可全面理解这些客户是谁，它们是如何定义的，以及哪些业务度量指标控制它们的状态。IBM InfoSphere Metadata Workbench 创建整个数据整合流程的一个集中、整体的视图。

8.2 确保合适的数据血统

财务欺诈漏洞是组织的风险缓解的关键业务驱动因素。我们中的许多人都还记得最终导致美国政府建立 Sarbanes-Oxley Act (SOX) 的臭名昭著的财政丑闻。SOX 的目标包括建立和实现措施来在公司的会计实践中建立信任。

风险缓解和合规性需求影响着组织管理它们的信息的方式。财务报告中采用一种数据欺诈保护或者甚至欺诈检测机制是，证明数据的来源，它流经何处，以及它在企业中传输时经过了何种转变。但是，传统的基于项目的数据整合实践或业务部门的合并所导致的工具激增，创造了一种似乎无法实现的导航和整合做法。使该流程更加透明和有效，需要合并或整合工具集和一种元数据驱动的方法来为建立一种通用的基础答案。

数据血统为数据在整合流程中的移动提供了一种审计线索。数据血统流程的结果是对“此数据来自何处？”、“此数据去向何处？”和“在此过程中它发生了什么？”等基本问题的回答。图 10.1 提供了 IBM InfoSphere Metadata Workbench 内一种数据血统报告的示例。IBM InfoSphere Metadata Workbench 利用数据血统扩展程序来合并非

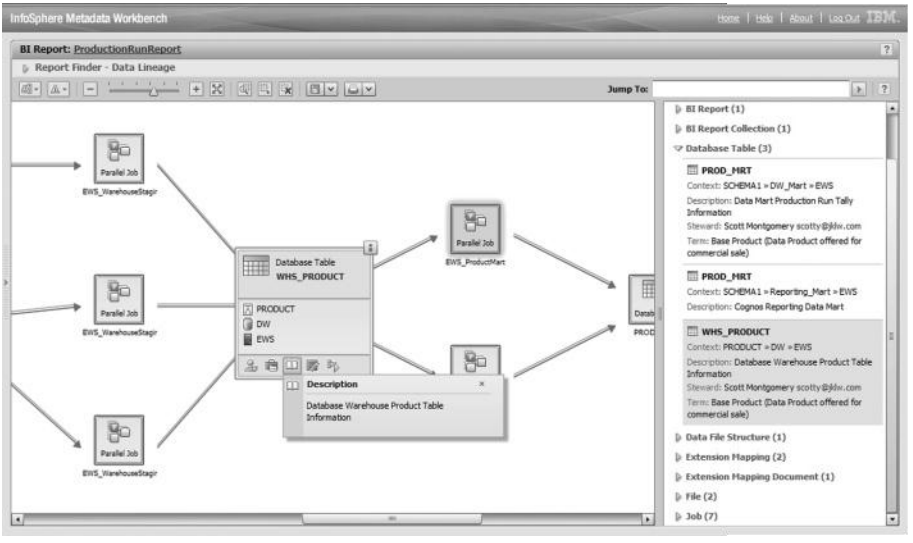


图 10.1：来自 IBM InfoSphere Metadata Workbench 的一份数据血统报告。

IBM InfoSphere 数据整合流程，比如存储过程、COBOL 程序和第三方 ETL 流程。

8.3 执行影响分析

数据治理计划需要确保用户能够检查与一个对象相关的所有关系，进而提供在创建任何更改之前评估和减轻风险的能力。理解对一种数据工件的更改对其他数据工件有何影响的能力称为*影响分析*。考虑到在开发生命周期中不可避免地会引入更改，影响分析允许公司更有效地治理数据。

图 10.2 给出了一个服务器的图形化依赖关系分析报告，列出了如果服务器必须离线以进行维护，会影响到哪些数据库、作业或商业智能 (BI) 报告。在传统上，收集此类信息将需要利用多个用户和多个工具集来评估潜在风险。

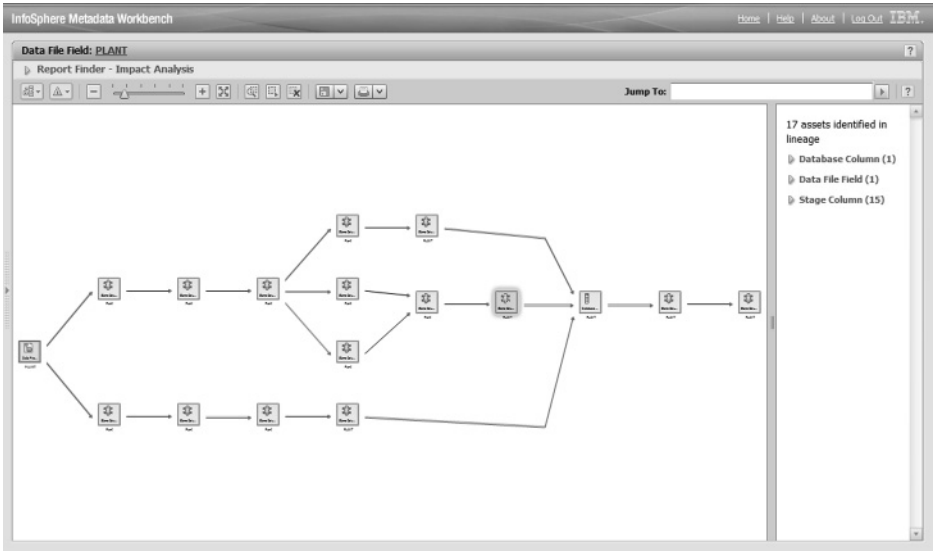


图 10.2：来自 IBM InfoSphere Metadata Workbench 的一份影响分析报告。

8.4 管理操作元数据

操作元数据填补了应该发生的事件与实际发生的事件之间的差距。企业可能有一个商业智能环境，在 IBM InfoSphere DataStage 等 ETL 工具内包含数千个作业，还包含以批量模式执行的转换。从数据治理角度，亟需及时知道任何作业是否在流程中的某个位置失败，或者某些数据行是否丢失。

一些操作元数据示例包括：

- 运行的作业是否失败或遇到了警告
- 读取、写入或引用了哪些数据库表或文件
- 读取、写入或引用了多少行
- 作业何时开始和结束
- 使用的阶段和链接
- 作业所属的项目
- 运行作业的计算机
- 作业所使用的任何运行时参数
- 在作业运行期间发生的事件，包括在作业链接上写入和读取的行数
- 作业的调用 ID

总结来讲，本章探讨了作为有效的数据治理计划的基础，合理的元数据战略的重要性。元数据存储库将支持数据血统、影响分析和操作元数据的分析。

11

第 9 步：定义度量指标

数据治理倾向于集中在人员和流程上，二者都是无形的。因此，拥有协商一致的度量指标或关键绩效指标 (KPI) 集对于度量和监控数据治理计划的进度很重要。

本章介绍确定 KPI 来度量和监控您组织的数据治理计划绩效的过程。一定要认识到这些 KPI 需要针对您的组织及其人员、流程和数据进行调整。定期度量这些 KPI 并向数据治理委员会和高级管理人员报告结果，这很重要。业务驱动的和技术数据治理 KPI 需要每 1 到 3 个月度量和跟踪一次。数据治理成熟度评估具有定性的形式，通常应该每年跟踪一次。

IBM 在 Cognos 中开发了一个数据治理计分卡，以帮助公司管理其数据治理计划的绩效，本章稍后将会探讨。

以下是“定义度量指标”步骤只能怪包含的子步骤：

9.1 理解业务的整体 KPI。

9.2 定义数据治理的业务驱动 KPI。

9.3 定义数据治理的技术 KPI。

9.4 建立数据治理成熟度评估的仪表板。让我们更详细地分析一下每个子步骤。

9.1 理解业务的整体 KPI

每个业务部门都将有一个 KPI 层次结构来运行其业务。IBM 出版了一部由 Roland Mosimann、Patrick Mosimann 和 Meg Dussault 编写的著作，名为 *The Performance Manager: Proven Strategies for Turning Information into Higher Business Performance*（Cognos, Inc.，2007 年）。该书按工作职能（比如销售、营销、财务和风险）描述了业务 KPI 层次结构。理解这些度量指标很重要，因为数据治理的重要目标是提高产生这些 KPI 的数据的可信程度。

9.2 定义数据治理的业务驱动 KPI

数据治理计划需要行医一组有针对性的 KPI，以提高业务 KPI 的可信程度。例如，在一家银行，风险团队将希望按行业度量银行的整体暴露面，避免在严峻的经济下滑形势下或特定的行业趋势下的过量收入损失。该银行的按行业划分的总体风险是一个业务 KPI。但是，银行从多个源系统获取行业数据并发现标准行业分类 (SIC) 代码在多种情况下是空的。结果，行业风险将错误地计算。数据治理 KPI 将为包含空 SIC 代码的记录的百分比。数据治理计划应该使用此度量指标每月跟踪绩效并向风险团队报告进度。

考虑另一个来自保险行业的示例。从操作角度讲（为了遵守制度，比如欧洲的 Solvency），一家生命或财产和意外保险商将希望限制过量的地域风险，避免由飓风或地震等事件导致惨重的损失。

该保险公司在地域上的总体风险是一个业务 KPI。但是，该保险公司发现它拥有的几个源系统包含与邮政编码相关的不完整的保险客户数据。数据治理 KPI 将是包含空邮政编码的保险客户百分比。此 KPI 将提高保险公司的 Solvency II 灾难风险计算的可信程度。

9.3 定义数据治理的技术 KPI

技术 KPI 度量数据治理的技术方面的进度。以下是数据治理技术 KPI 的一些示例：

- *元数据*——示例 KPI 包含已备案数据流的数量、已监控数据流的数量、具有填充到数据字典内的协商一致的定义的业务词汇百分比（按领域划分，比如风险或财务），以及“孤立资产”的数量。

  孤立资产可能源于导入过多或不完整的元数据，或者删除资产的身份层次结构中的一个或多个对象。例如，如果您导入一个数据库而不指定主机，数据库将成为元数据存储库中的孤立资产。孤立资产可是物理数据资源 (PDR) 或商业智能 (BI) 资产。可能孤立的 PDR 资产包括数据库、数据文件、模式、存储过程和数据集合。可能孤立的 BI 资产包括 BI 集合、BI 多维数据集和报告查询。

- *内容管理*——示例 KPI 包括已数字化并经过了记录管理的纸张、电子和电子邮件文档的百分比，按业务部门划分。其他 KPI 包括用于内容分析的核心业务文档（比如保险索赔或银行贷款文档）的百分比，周转一个 eDiscovery 请求的平均时间（以小时计算），以及填充到企业元数据存储库中的非结构化元数据的百分比。

- *归档*——示例 KPI 包括总存储空间（以 GB 计算）、总存储成本、平均应用程序响应时间和周转审计查询的平均时间（以天计算）。

- *商业智能能力中心*——示例 KPI 包括每月每个业务区域的用户、报告和报告执行数量。
- *安全和隐私*——示例 KPI 包括针对 Sarbanes-Oxley、Payment Card Industry Data Security Standard、United States Health Insurance Portability and Accountability Act 和 European Data Protection Directive 等制度的失败审计数量。
- *数据库审计*——示例 KPI 包括在过去 12 个月内执行来测试敏感数据的漏洞的测试数量，发现的数据库漏洞异常数量，每天用于协调实际的数据库更改与批准的数据库更改请求的工时数，对敏感数据的未授权更改的数量，以及生产数据库中的 SQL 错误数量。

9.4 建立数据治理成熟度评估的仪表板

KPI 需要基于对组织成熟度的定性评估而开发。这些 KPI 一般应该至少每 12 个月重复度量一次。它们在 1 到 5 的量表上度量 11 个 IBM 数据治理成熟度模型类别中的每一个的实际评分、目标评分和偏差。作为参考，下面列出 11 个类别：

1. 数据风险管理和合规性
2. 价值创建
3. 组织结构和感知
4. 策略
5. 数据照管
6. 数据质量管理
7. 信息生命周期管理
8. 安全和隐私

9. 数据架构
10. 分类和元数据
11. 审计信息、日志和报告

图 11.1 显示了一个度量数据治理成熟度评估结果的 IBM Cognos 计分卡。

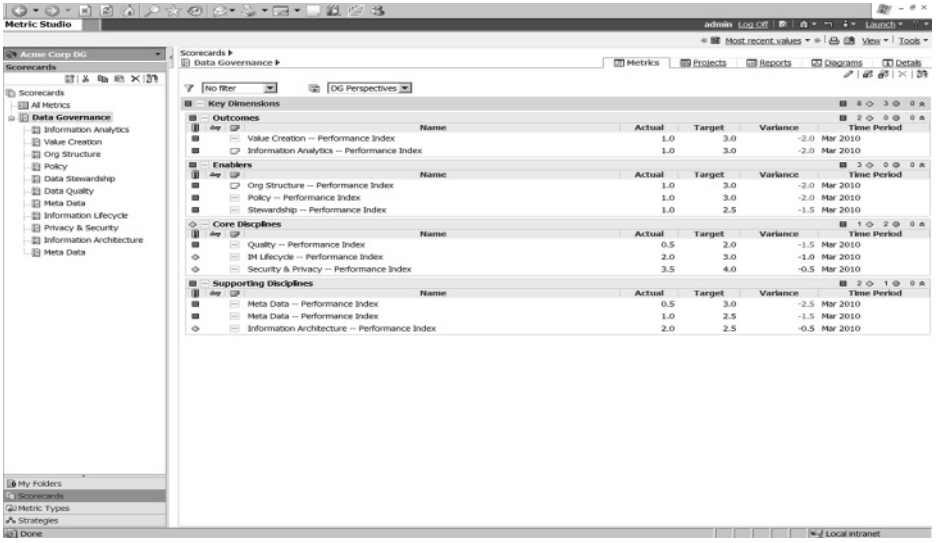


图 11.1：数据治理成熟度评估的 IBM Cognos 计分卡。

在上图中，数据治理成熟度评估的类别基于组织的需要进行了细微调整。该组织选择了 IBM 数据治理成熟度模型类别的一个子集。由于需要密切关注商业智能，它还添加了一个围绕“信息分析”的自定义类别。此外，它将元数据分解为两个类别：技术和业务元数据。最后，为了强调非结构化数据的重要性，组织将“数据架构”类别的名称更改为了“信息架构”。

12

第 10.1 步：任命数据照管人

主数据治理是一种持续的实践，其中业务领导定义准则、策略、流程、业务规则和度量指标，通过管理其主数据的质量来实现业务目标。数据治理委员会编组人员、组织、资源、优先级和技术以实现这些策略声明。数据治理委员会然后监控和度量针对这些目标的进度，确保主数据满足质量目标。

接下来的 3 章将介绍数据照管、数据质量和主数据管理 (MDM)，它们是主数据治理的关键组件。

数据照管可能是企业数据治理计划的第一颗种子开始发芽的地方。数据照管是一种质量控制学科，旨在确保对信息进行保管，解决业务需要。组织任命理解业务的数据照管人来确保信息“适合其用途”。数据照管人不是数据的所有者，它们是负责改进作为企业资产的数据质量的保管人。

以下第 10.1 的子步骤，本章剩余部分将更详细探讨这些子步骤。

10.1.1 任命首席数据照管人.

10.1.2 确定数据照管计划的配置
（比如由 IT 系统、组织或主题区域执行）。

10.1.3 确定每个数据领域的高层支持者。

10.1.4 招聘每个数据领域的数据照管人。

10.1.5 授权数据治理委员会监督数据照管计划。

10.1.1 任命首席数据照管人

在理想情况下，数据照管人应该向业务部门报告。因为数据照管人将向多个部门或职能区域报告，所以组织应该任命一位首席数据照管人，以确保跨各种照管角色的一致性。首席数据照管人需要是数据治理工作组的成员，必须在整体数据治理计划中发挥积极作用。

10.1.2 确定数据照管计划的配置

可采用多种方式配置数据照管计划。这些方式可以成熟度模型的形式表示，如图 12.1 中所示。

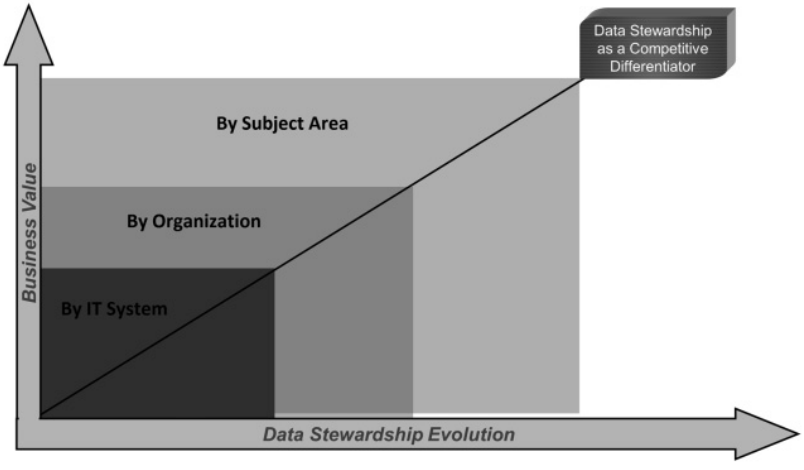


图 12.1：IBM 数据照管成熟度模型。

初级成熟度级别：按 IT 系统调整的数据照管

在最常见的模型中，数据照管人分配用于管理给定的 IT 系统或应用程序中的数据。例如，可以分配数据照管人来清理客户信息文件 (CIF) 中的客户数据。在此场景中，数据照管人将最可能位于 IT 部门内。

这种数据照管模型最容易实现，因为它不需要业务在数据治理计划中担任任何有意义的职责。在另一方面，以 IT 为中心的模型也具有一些不足。数据质量需要一种业务透视图来确保信息适合其用途。因为业务部门在数据照管计划中的参与有限，所以 IT 照管数据的能力将受到约束。另外，典型的组织拥有多哦 CIF 或产品和商品层次结构。结果，没有大量手动干预，将你很难回答“我们整个企业的最重要客户是谁？”和“我们最畅销的产品有哪些？”等问题。

中级成熟度级别：按组织调整的数据照管

一些企业将选择按工作职能（比如风险、营销和销售）或按业务线部署数据照管人。这些数据照管人通常向业务部门报告，负责相应领域内的所有数据。此方法的优势是数据照管与业务之间存在链接。缺点是组织会继续分散地处理企业数据，比如客户、商品、供应商和产品信息。

高级成熟度级别：按主题区域调整的数据治理

成熟的组织将联合部署一个数据照管计划和一个主数据管理计划。这些组织认识到需要使用单个记录系统来记录关键的主数据实体，比如客户、商品、供应商和产品，将这些数据实体视为企业资产。

此方法使组织能够最灵活地解决客户中心性等企业计划。但是，此方法也最难实现。各个业务经理必须将对其数据的控制转交给 MDM 中心。

要确保此方法成功，数据照管人需要积极参与定义关键数据元素或属性，其中两位或多位业务经理拥有关于如何接受添加和更新的有争议的规则。数据照管人还需要在两个或更多业务经理发生冲突时，拥有明确定义的角色。例如，业务经理 A 无法改写与业务经理 B 也存在关系的客户的地址。如果业务经理 A 拥有一位共同客户的地址更新，它将处于一种挂起状态。更新必须经过业务经理 B 批准，才会被接受。

一种分阶段企业照管方法（一次处理一个数据领域）也是一种最佳实践。与在所有情况下一样，在设计您的数据照管计划之前去那个考虑您自己的独特情形。

10.1.3 确定每个数据领域的高层支持者

组织需要识别数据照管的高层支持者。这些支持者将拥有数据的最高职权，但他们将最可能将每天的活动委托给其他某个人。高层支持者的选择受数据照管计划的目标配置驱动。

如果数据照管计划由 IT 系统调整，高层支持者将位于 IT 部门内，可能是特定 IT 系统的所有者。如果数据照管计划由组织调整，高层支持者可能位于 IT 或业务部门或同时位于两个部门，但在所有情况下都将由组织调整。最后，如果数据照管计划由主题区域调整，高层支持者将可能位于销售、HR、财务和供应链等职能部门内，具体取决于相关联的数据领域，比如客户、员工、财务数据、供应商和产品。

以下是高层支持者的一些职责：

- 拥有一个领域内的数据质量的最高处理职责
- 确保一个领域内所有敏感数据（比如 PII 和 PHI）的安全和隐私
- 任命数据照管人执行处理一个领域内的数据质量、安全和隐私问题的日常职责
- 建立和监控与一个领域内的数据治理进度相关的度量指标
- 在业务规则冲突的情形下与其他高层支持者协作，确保企业继续从其数据获取最大价值

10.1.4 招聘每个数据领域的数据照管人

前面已经提到，高层数据照管支持者将最可能将日常责任委托给某个级别较低的人，这个人能够将足够的时间投入到任务中。执行这些日常任务的个人称为“数据照管人”。许多组织内的数据照管人常常依据现状进行操作。，通过花大量时间解决数据质量问题，他们成为了数据照管人，即使他们不知道他们在具有该能力。有时，使用或支持创建数据的关键应用程序的业务用户或分析师也是不错的数据照管人。例如，保险公司内某个处理承保或保险系统的人可能是不错的数据照管人，因为这个人理解数据的业务用途。

许多数据照管人倾向于担任多个头衔。但是，当一个组织拥有正确级别的业务支持时，将有一股自然的力量推动他们将更多时间花在其照管职责上。

10.1.5 授权数据治理委员会监督数据照管计划

当数据照管计划成熟时，数据照管人应该向业务部门报告。这时，确保对所有数据照管人存在一定级别的监督，确保角色和职责的一致性，开发一种社区认知，这很重要。

数据治理委员会的理想定位是监督数据照管计划，在整个组织内实现一致的执行并与业务部门紧密链接。委员会可通过一致地关注跟踪数据照管计划绩效的 KPI，实现一种学科认知。第 11 章中对度量指标的探讨将更详细地介绍此主题。

13

第 10.2 步：管理数据质量

典型的组织拥有大量与其分散在所有运营系统中的客户、产品和供应商相关的信息。没有合适的监督，此数据的质量将不断下降。数据质量管理是一门学科，包含度量、改进和验证组织数据质量和完整性的方法。数据质量包括数据标准化、匹配、寿命和持续的质量监控。

数据治理组织需要建立策略来识别高价值数据属性，建立机制来度量对数据质量的不断改进。以下是与这个“管理数据质量”步骤相关的子步骤：

10.2.1 建立数据质量策略, 包括高价值数据属性的识别。

10.2.2 设置数据质量基准。

10.2.3 创建业务案例。

10.2.4 清理数据。

10.2.5 持续监控数据质量。下面更详细地探讨这些子步骤。

10.2.1 建立数据质量策略

每个业务部门都拥有对其运作至关重要的数据。这个子步骤与关于数据治理度量指标的第 9 步紧密链接。在数据治理组织识别了业务驱动的数据治理 KPI 之后，很容易确定具有最高价值的数据属性。例如，标准行业分类 (SIC) 代码将是希望按行业评估其整体风险的银行的高价值数据属性。类似地，材料所有者的再订购水平将是希望严格管理其供应链的制造商的高价值数据属性。

还需要围绕遵守业务规则的其他策略。例如，一家制造商可能明确要求，对于需要再订购规划的材料，再订购字段不得为空。

也需要围绕可接受的数据质量水平的策略。例如，无法递送的邮寄地址是影响邮寄成本的一个关键的质量问题。但是，数据治理组织可以明确规定 1% 的糟糕数据质量是可接受的，只要无法递送的邮寄地址在该阈值之下，就无需进一步措施。最后，数据治理组织需要定义处理数据质量解决方式的策略和规程。

10.2.2 设置数据质量基准

数据必须具有合适的质量，才能解决业务的需要。可通过多种方式评估数据集的质量：

- *有效性*——数据值具有可接受的格式。例如，员工编号为6位文字数字字符。
- *唯一性*——数据字段中没有重复的值。
- *完备性*——数据字段中没有空值。例如，邮政编码应该始终填入到地址表中。
- *一致性*——数据属性与可能基于该属性本身或多个属性的业务规则一致。例如，一条业务规则可能检查出生年份是否早于 1900 年 1 月 1 日或保险单有效期是否早于保险单生成日期。

- *及时性*——数据属性表示没有过时的信息。例如，没有客户合同拥有已过期的有效期。
- *准确性*——数据属性是准确的。例如，员工工作代码是准确的，可确保员工不会受到错误类型的培训。
- *符合业务规则*——数据属性或数据属性组合遵守指定的业务规则。（介绍度量指标的第 11 章将更详细地介绍此主题。）

这个子步骤与第 9 章中介绍的“理解数据”步骤紧密相关。IBM InfoSphere Information Analyzer 提供了一种自动方式来设定数据质量基准。

10.2.3 创建业务案例

数据治理组织识别了高价值数据属性并设定数据质量基准之后，它就有了足够的信息来创建业务案例。图 13.1 给出了一个虚构的业务案例，通过在客户数据库内匹配重复值，以及确定多个人是否真正属于同一家人，从而改进数据质量。营销

A. 营销列表中的客户总数	950,000
B. 个人方面匹配值数量	40,000
C. 在一个家庭中计算了两次的其他重复的个人	50,000
D. 重复匹配的总数	90,000
E. 每年每位客户的营销邮寄次数	2
F. 每次邮寄的成本	$3.25
G. 总计可避免的重复邮寄成本 (DxExF)	$585,000
H. 每位客户每年的呼出电话销售次数	4
I. 每次呼出电话销售的成本	$1.50
J. 总计可避免的呼出电话销售成本 (DxHxI)	$540,000
K. 总计可避免的重复匹配成本 (G+J)	$1,125,000
L. 实现数据质量工具的成本	$500,000
M. 全职客户数据照管人每年成本	$200,000
N. 数据质量解决方案总成本 (L+M)	$700,000
O. 回收期	7.5 months

图 13.1：一个针对营销部门的虚构数据质量业务案例。

部门在邮寄产品目录和向客户电话销售上花了数百万美元。结果，通过删除重复值而对客户列表的任何精减将直接转换为底线利润。

另一个具有主要数据质量含义的场景是企业资源规划 (ERP) 实现。一个典型的 ERP 实现可能涉及到数据从数十个（可能是数百个）遗留应用程序向目标系统的转移。观察得到的证据表明，超过 40% 的 ERP 项目成本与数据整合相关。此外，数据质量问题已被视为 ERP 项目失败的主导原因之一。正确加载数据是一回事，“正确加载正确的数据”又是另一回事。因此，任何 ERP 项目都需要关注迁移到目标应用程序中的数据质量。

10.2.4 清理数据

IBM Initiate Master Data Service 是一个 MDM 系统，包含匹配和链接功能。IBM InfoSphere QualityStage 帮助组织清理数据和管理数据质量。这些工具方便了高质量主数据的创建和维护，可匹配姓名、地址、电话号码、电子邮件地址和生日等数据。

图 13.2 演示了数据照管人如何利用匹配引擎来标准化部件数据。输入文件包含非标准格式的部件数据，但匹配引擎能够基于汇编指令、数量、类型、部件、大小、度量单位和 SKU 而输出标准化格式的数据。

输入文件

WING ASSY DRILL 4 HOLE USE 5J868A HEXBOLT 1/4 INCH WING ASSEMBLY, USE 5J868-A HEX BOLT .25”- DRILL FOUR HOLES USE 4 5J868A BOLTS (HEX .25) - DRILL HOLES FOR EACH ON WING ASSEM RUDER, TAP 6 WHOLES, SECURE W/KL2301 RIVETS (10 CM)

结果文件

Assembly Instruction	Qty	Type	Part	Size	Measure	SKU
WING DRILL	4	HOLES	HEXBOLT	.25	INCH	5J868A
WING DRILL	4	HOLES	HEXBOLT	.25	INCH	5J868A
WING DRILL	4	HOLES	HEXBOLT	.25	INCH	5J868A
RUDDER DRILL	6	HOLES	RIVET	10	CM	KL2301

图 13.2：一个部件标准化示例（基于白皮书 IBM InfoSphere Information Server: Cleansing Data and Managing Data Quality，*IBM，2006 年）。*

10.2.5 持续监控数据质量

清理数据之后，您需要确保它保留了高质量。数据常常在清理之后，回到了一种低质量状态。然后需要大量工作来再次清理所有数据。

数据质量流程需要考虑多个步骤。一些步骤是自动化的，一些不是。数据质量始于改进流程和培训数据条目和输入。您需要让输入数据的人保持一致，在最终提交之前验证数据。

我们都知道，培训和流程充其量将提供 50% 的数据质量改进。这就是为什么您还需要采样和探查数据来实现持续改进。IBM InfoSphere Information Analyzer 探查数据以提供数据质量评估、分析和持续监控，确保信息持续保持高质量。

14

第 10.3 步：实现主数据管理

要满足收入增长、成本减少和风险管理等基本战略目标，组织需要控制常常禁锢在业务部门中的孤岛内的数据。此信息的最有价值的部分（关于客户、产品、材料、供应商和帐户的关键业务数据）常常称为主数据。尽管它很重要，主数据常常是重复的，分散在整个企业的业务流程、系统和应用程序中。组织现在认识到了主数据的战略价值。他们正在开发长期的主数据管理 (MDM) 操作计划，利用此信息来促进企业成功。

*主数据领域*指一种特定的信息类别，比如客户、产品、材料、供应商或帐户。每个数据领域具有需要“适合其用途”的特定属性。例如，一个电话号码是客户数据领域的重要属性，因为企业拥有有效的联系人数据以满足需要，这至关重要。

主数据领域之间存在着许多关系，它们代表着真正的理解。例如，银行拥有给定客户的所有帐户和产品的链接视图很重要，这样它就可以理解

方便服务、附加产品销售和可移植性分析的总体关系。客户、帐户和产品代表着拥有关系的主数据领域。

认识到不是主数据的内容很有用。主数据不是仅供一个应用程序使用。它是高价值数据，不是低价值数据。最后，它通常不是很少更新的数据。

没有合适的监督，MDM 计划无可避免地会导致组织冲突，因为各个业务经理会认为他们必须将对其数据的控制权转交给 MDM 中心。我们将此场景称为“主动 MDM”，这正是主数据治理至关重要的原因。

以下是与实现 MDM 相关的子步骤：

10.3.1 识别业务问题。

10.3.2 定义主数据主题区域。

10.3.3 识别使用数据的系统和业务流程。

10.3.4 识别当前的数据源。

10.3.5 定义记录系统的数据属性。

10.3.6 为每个记录系统任命数据照管人。

10.3.7 建立主数据治理策略。

10.3.8 为手动干预和监控实现数据照管控制台。

10.3.9 管理潜在的重叠任务。

10.3.10 匹配来自相同来源或多个来源的可疑重复内容，创建一个新主记录。

10.3.11 链接来自多个来源的相关记录。

10.3.12 检查唯一标识符是否重复。

10.3.13 管理关系。

10.3.14 管理层次结构。

10.3.15 管理分组。

10.3.16 构建主数据管理解决方案。

本章剩余部分更详细地介绍其中每个子步骤。

10.3.1 识别业务问题

识别业务问题与 IBM 数据治理统一流程的第一步“定义业务问题”紧密关联。MDM 计划需要足够巧妙，以确保快速获得回报。大部分 MDM 计划倾向于关注提高收入、减少成本和管理风险的关键目标。收入增长可能专注于客户中心性，成本减少可以确定供应商效率，而风险管理可能希望改进向关键对立方暴露的总体信用风险的计算。

10.3.2 定义主数据主题区域

尽管主数据实体的优先级因行业不同而不同，但有一些常见的线索。关键主数据实体可能是客户、供应商、代理、位置、产品、材料、员工和财务数据。目前为止，客户数据是所管理的最常见的主数据实体，因为客户是任何业务的命脉。对于大部分企业，HR 部门将为员工主数据提供支持，财务部门将希望确保一致的会计客户表，以方便顺利的财务整合。

制造商倾向于关注客户、供应商、产品、商品和资产主数据。客户数据的典型支持者包括销售、营销和客户服务区域。著名工业产品公司的 CEO 在他在另一家公司的对立方告诉他“我们是您的前 5 大客户之一”（这是他自己的团队无法轻松验证的）时，将支持实施一个 MDM 计划。供应链和工程小徐对产品和材料主数据具有很高的兴趣，典型的制造商在多个层次结构中拥有成百上千个 SKU。最后，供应链、采购和采取区域对供应商主数据具有极高兴趣。

保险公司关注保险客户和经纪人主数据，具有来自营销、保险单管理和经销部门的业务支持。银行关注客户、帐户和产品主数据，受风险管理、营销、客户服务、零售银行和企业银行部门支持。

电信服务提供商（“telcos”）拥有多个专为各项产品定制的系统，比如铜线型陆线、无线和 DSL。结果，telcos 需要与其用户的总体关系的单一视图，以方便流失管理和交叉销售额外的产品等计划的实施。其他主数据实体包括产品、资产、客户使用、供应商和关税。

零售商可能对客户、产品和供应商主数据非常感兴趣。许多零售商的营销部门正使客户主数据成为其客户中心性和忠诚计划的基础。这些组织的供应链和促销团队对有效管理产品主数据非常感兴趣，因为他们已拥有跨多个层次结构管理成百上千个 SKU 的众多人员。例如，一家大型零售商的供应链和财务在理解了每个供应商在所有产品和业务部门上的总体花费时，就可以与他们的提供商协商主要折扣，有效使用制造商的退款。

10.3.3 识别使用数据的系统和业务流程

理解哪些系统和业务流程在使用数据很重要。新 MDM 中心也将需要支持这些系统和业务流程。

10.3.4 识别当前的数据源

识别数据的当前来源和与该数据关联的业务规则很重要。您可能已在 IBM 数据治理统一流程的早期解决（至少部分解决）了此步骤。

10.3.5 定义记录系统的数据属性

再一次，您可能已在 IBM 数据治理统一流程的早期解决（至少部分解决）了此步骤。就记录系统 (SOR) 的数据属性达成一致，这很重要。例如，“客户”的数据属性可能包括姓、名、电话号码、社会安全号码或国家 ID、街道地址、城市、州和邮政编码。IBM InfoSphere MDM 和 IBM Initiate 都提供了一个既定的数据模型，该模型对着手

数据治理组织需要在 SOR 的关键数据属性由多于一个业务流程更新时进行协调。让我们看看一个来自金融服务行业的示例。一家金融服务公司拥有一家银行也一个生命保险公司，正在

实现一个集中的 MDM 计划来更好地管理客户关系。

数据治理团队就“客户”数据的 21 个属性达成了一致。数据治理策略是，客户可以通过呼叫其银行或生命保险公司来更改他或她的履历，将这些更改传播到整个企业。这个流程创造了一些数据治理问题。银行将接受对生日的更改并进行有限备案，而由于会影响到生命保险费用，生命保险部门将需要支持文档。银行相信生命保险流程将减缓他们的流程，生命保险公司认为银行的流程风险太大。

在经历了太多绝望和多次内部会议之后，数据治理委员会商定了一种折衷方法。银行的客户服务代表将有一个标记了还拥有生命保险的客户的屏幕。如果客户没有生命保险，银行将接受对生日的更改，而不备案。如果银行客户还拥有生命保险，银行会要求客户提交支持文档。

10.3.6 为每个记录系统任命数据照管人

组织需要为每个 SOR 任命数据照管人。这些照管人需要位于业务部门中，并向其各自数据领域的高层支持者报告。

数据照管人必须拥有数据在企业日常操作中的使用方式的足够知识。例如，客户数据照管人将位于销售、客户服务或营销部门内。类似地，产品数据照管人将位于供应链、工程或研发部门内。介绍数据照管的第 12 章更详细地介绍了此主题。

10.3.7 建立主数据治理策略

数据治理计划需要建立围绕 MDM 的策略。这些策略的示例包括：

- 数据匹配规则
- 自动匹配与手动干预的规则

- 数据验证规则
- 治理对关键数据（比如生命保险公司示例中的生日）的更改的规则
- 命名约定
- 将特定属性识别为敏感数据

这些策略的实现在后面的子步骤中解决。另请参阅附录 E 了解数据治理声明的示例。

10.3.8 为手动干预和监控实现数据照管控制台

不考虑数据匹配过程中的自动化水平，数据照管人的一定程度的手动干预始终是必要的。IBM InfoSphere 主数据管理服务器和 IBM Initiate 通过数据照管控制台提高了数据照管人的效率。这些控制台分别称为 IBM InfoSphere Master Data Management Data Stewardship User Interface (DSUI) 和 IBM Initiate Inspector。图 14.1 提供了 DSUI 的一个屏幕截图。我们将在本章剩余部分使用词汇“IBM MDM”来涵盖 IBM InfoSphere Master Data Management Server、IBM InfoSphere Master Data Management Server for Product Information Management 和 IBM Initiate。

图 14.1：IBM InfoSphere MDM Data Stewardship User Interface，一个供数据照管人使用的控制台。

10.3.9 管理潜在的重叠任务

当使用与记录中已有的数据差别巨大的信息更新记录时，可能会发生重叠。折衷情形通常被视为要解决的最紧迫的任务。例如，考虑图 14.2 中所演示的情形。数据照管人检查过去属于 Jane Lewis 的记录。但是，在 2006 年 8 月 24 日，该记录被更新。它现在看起来属于一个名为 Linda Xiang 的女士。Linda Xiang 和 Jane Lewis 很明显不是同一个人，但当您查看记录的结构时，可以看到 Linda Xiang 的数据曾经保存在 Jane Lewis 数据值上。原因可能是很常见的数据录入错误，其中 Jane Lewis 的记录在客户服务代表开始渐入时在屏幕上打开，该代表没有意识到他或她键入在了其他某人的数据上。

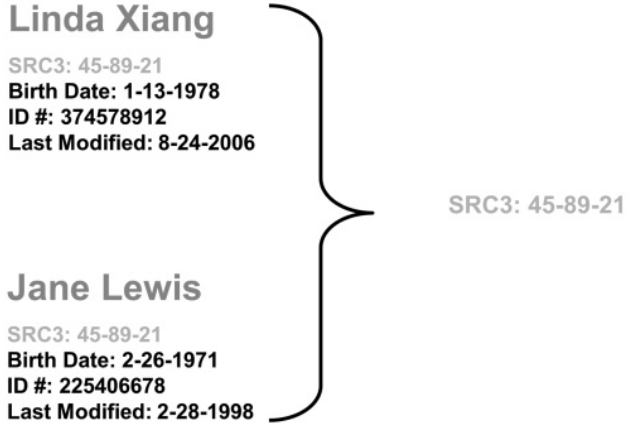


图 14.2：潜在的重叠任务。

还存在一些情形，其中此场景可能完全有效。例如，很明显可能由于结婚、离婚、搬家或电话号码更改，一个人的数据发生重大变化而导致被标志为潜在的重叠任务。无论在哪种情况下，都不会影响数据照管人调查、反复检查和证实用户查看的是正确的人的数据。

由于存在紧迫的特征，潜在的重叠任务通常是数据照管人要解决的最高优先级任务。如果某些数据元素（比如姓名、身份号码或电话号码）完全不同，IBM MDM 会将这些记录标记为潜在的重叠任务。

10.3.10 匹配可疑重复内容以创建新主记录

让我们考虑另一个示例，一家多险种保险公司的数据照管人能够利用 IBM Initiate Master Data Service 或 IBM InfoSphere Quality Stage 等匹配引擎，使用来自所有来源的最佳数据创建客户配置文件。该保险公司在生命保险、家庭保险和汽车保险领域拥有多个客户信息文件 (CIF)，希望使用来自所有这些险种的最佳数据创建客户配置文件。

如表 14.1 中所示，该保险公司从来自生命、家庭和汽车保险的 3 个帐户开始，这 3 个帐户具有类似的名称，稍微不同的地址和电话号码，生日是空的或者不一致。数据照管人需要评估这些帐户是否需要链接起来。如果链接了这些帐户，数据照管人需要使用来自整个企业的最佳信息创建客户配置文件。

表 14.1：经典的帐户到客户转换 – 帐户视图

来源	旧有键	姓名	地址	电话	生日
生命	70328574	John Smith Jr.	10 Main St Boston MA 02110	781-259-9945	02/05/1940
家庭	80328575	Mr. John Smith	10 Main St Unit 10 Boston MA 02111	617-259-9000	
汽车	90238495	J. Smyth	Main St Bostan Mass 02110	781-295-9945	02/05/1941

基于名称和街道地址的类似性，匹配引擎能够确定生命和家庭保险单需要链接起来，如表14.2 所示。对于具有汽车保险的“J. Smyth”是否与“John Smith”是同一个人，还存在疑问。

表 14.2：客户视图

来源	旧有键	姓名	地址	电话	生日	客户 ID
生命	70328574	John Smith Jr.	10 Main St Boston MA 02110	781-259-9945	02/05/1940	0001
家庭	80328575	Mr. John Smith	10 Main St Unit 10 Boston MA 02111	617-259-9000		0001
汽车	90238495	J. Smyth	Main St Bostan Mass 02110	781-295-9945	02/05/1941	

经过进一步检查，数据照管人确定地址和名称的相似程度足以将它们与同一个人关联。现在，数据照管人如何从多个不一致的地址、电话号码和生日中进行选择？幸运的是，这正是数据寿命规则应用的地方。数据治理寿命规则规定，生命保险是生日的最佳来源，因为该信息决定了保险费用。类似地，家庭保险是地址信息的最佳来源，因为该数据直接与受保的实体绑定。在此基础上，数据照管人能够使用来自整个企业的最佳信息编写客户配置文件，如表14.3 所示。

表 14.3：客户配置文件

来源	姓名	地址	电话	生日	客户 ID
客户配置文件	Mr. John Smith Jr.	10 Main St Unit 10 Boston MA 02111	617-259-9000	02/05/1940	0001

数据照管控制台允许照管人基于特定的条件搜索可疑的重复值。数据照管人通过对比每个可疑值的属性，然后基于每个可疑值的最佳属性压缩各方，确定各方是否是一个匹配值。数据照管控制台然后使用合并的属性创建新的一方，该方的所有现有记录被动地呈现。

10.3.11 链接来自多个来源的相关记录

子任务 10.3.11 与注册表或后面将探讨的虚拟的 MDM 架构方法密切相关。对于重叠，数据照管人验证已存在的记录。对于可疑的重复值，数据照管人删除额外的记录。在此子任务中，数据照管人在系统之间链接记录。

IBM MDM 可实现注册表样式的主数据管理，链接来自多个很可能表示同一个人的来源的记录。但是，记录可能没有足够的相同数据以便自动链接。例如，图 14.3 中的场景显示了针对“Ken Richardson”、“Kenneth Richardson”和“Len K. Richardson”的 3 条记录。当您首先查看这 3 条记录时，他们看起来是同一个人。但是，更详细地查看名称，就可以发现一些区别。

首先，名字不同：“Ken”、“Kenneth”和“Len K”。第二，一条记录具有用户 ID“46”，而其他两个记录具有 ID“64”。这种印刷上的移位很常见。IBM MDM 将这些记录分组为一个任务，以便数据照管人可确认它们是同一个人，然后链接它们。

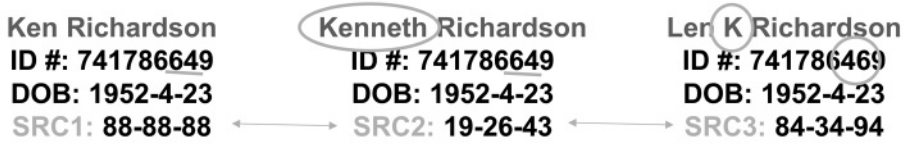


图 14.3：链接来自多个来源的相关记录。

10.3.12 检查唯一标识符是否重复

当来自相同来源的两条记录看起来使用了相同的唯一标识符（比如社会安全号码、护照比那好或驾驶员的执照编号）时，需要检查标识符任务。一些组织（比如执法和国土安全部的机构）可能从数据照管角度将此情形视为高优先级。

图 14.4 显示了两条明显不是同一个人，但具有相同的键标识符编号的人员记录。当记录包含相同的唯标识符，但具有其他反映两个不同个人的属性时，IBM MDM 将记录标记为手动检查。

图 14.4：检查标识符任务。

10.3.13 管理关系

Joe 和 Mary 都拥有相同银行的帐户，他们已结为夫妻。通过查看他们的帐户分组，数据照管人可看到 Joe 的名下有两个帐户（一个汽车贷款和一个信用卡）。Mary 的名下拥有 3 个帐户（一个活期、一个信用卡和一个汽车贷款）。他们还

拥有 3 个联合帐户（抵押贷款、活期帐户和储蓄帐户）。通过使用关系，数据照管人可开发 3 个帐户所有者的完整视图——Joe 的帐户、Mary 的帐户和他们的联合帐户——以建立整个家庭的信用风险，如图 14.5 中所示。

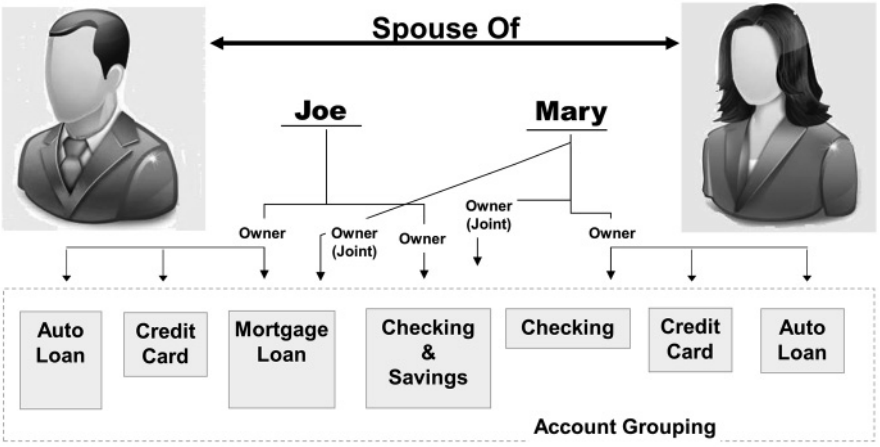


图 14.5：管理关系以确定家庭信用风险。

10.3.14 管理层次结构

数据照管控制台应该支持一名照管人添加或编辑层次结构。例如，涵盖 Jaguar 的销售组织将希望确保它的客户有资格遵守与 Tata 相同的合同条款，Tata 是新的父公司。因此，客户数据照管人将需要更改 Jaguar 的法律层次结构并将它包含在 Tata 下。

10.3.15 管理分组

数据照管控制台应该支持照管人添加或编辑分组，向分组添加一个相关方。例如，数据照管人可能决定将 John Smith 添加到富人市场区域。

10.3.16 构建 MDM 解决方案

最后，数据治理组织需要提炼业务需求，以确定 MDM 解决方案的合适的架构。有多种 MDM 架构方法：

- *事务架构*——此方法通常以一个面向服务架构 (SOA) 为基础，与现有的业务流程紧密整合。对一个源系统中的属性的更改首先会传播到集中的客户信息文件，然后传播到其他源系统。
- *注册表架构*——数据保留在原地，MDM 中心仅拥有几个系统中的源数据的指针。此方法非常适合医疗和执法等情形，在这些情形下有一些具体的制度不允许创建事务中心。
- *分析架构*——主数据被清理并转移到一个中央存储库，以回答“我们具有最大利润的客户是谁？”和“我们的顶级供应商有哪些？”等问题。此方法不会尝试更改源系统中的数据，相反，它仅将 MDM 中心用于分析用途。
- *混合架构*——此方法将其他 3 种架构的各种元素结合在一起。

IBM MDM 可解决所有这些样式的主数据管理。

15

第 11 步：治理分析

IBM 的 Michael Dziekan 是一位商业智能能力中心的一位长期的从业者。本章中的许多概念都基于他执行的 IBM Cognos 客户工作。

许多企业疲于应对对其分析环境的治理。各个部门创建使用不一致的数据创建了自己的报告，IT 并不总是知道来自仓库的数据是如何使用的，使用了哪些报告。企业正在开始实现商业智能能力中心 (BICC) 来解决这些挑战。

以下是与分析治理相关联的子步骤：

11.1 定义 BICC 的目标。

11.2 准备 BICC 的业务案例。

11.3 确定 BICC 的组织结构。

11.4 协商 BICC 的关键功能。

11.1 定义 BICC 的目标

尽管技术总是具有转型业务的潜力，但它这么做的能力常常受到组织内部的采用阻碍的危害。组织复杂性和“简介需要”的紧急性导致了商业智能(BI)、绩效管理和数据仓库解决方案的烟囱式实现。此情形已导致整个企业中 IB 解决方案的管理、交付和履行中的技能零散化和总体不一致性。

让我们看看一个例子，一家著名银行在设置其报告环境方面面临着巨大挑战。银行内的最终用户胡乱拼凑了数千个 Microsoft Access® 数据库，因为企业数据仓库被认为非常不灵活且昂贵，生成新报告要花太长的时间。数据治理组织委派了一个 BICC 来专门处理最终用户培训，减少生成新报告的成本和周转时间。

随着 BI 变得越来越战略性，组织创建了 IT 和 BI 用户工作组（现在常称为 BICC）来应对。BICC 是一种组织结构，通过相互关联的学科、知识领域、经验和技能将人们进行分组，以在整个组织推广专家经验。BICC 也称为卓越中心 (COE)、能力中心或知识中心。

BICC 可通过以下不同方式提供帮助：

- 通过一致的技能、标准和最佳实践集提供 BI 功能。
- 通过开发实现可重复且成功的 BI 部署，以对整个企业或部门而只是一个项目有意义的方式关注人员、技术和流程。

如果 BI 意欲扩展战术部署以成为一种更宽泛的解决方案，需要一种托管和可预测的方法。BICC 对于 BI 的战略部署至关重要。它通过以下方式，提高了以更低成本取得成功的可能性：

- 促进最终用户采用和消除业务与 IT 之间的隔阂
- 合并最佳实践功能和服务，支持来自其他部署的快速且可重复的成功

- 集中化能力和操作效率，最大化技术资源和资产的使用
- 确保完整的 BI 生命周期和“单一事实版本”在整个企业的更高和更快采用，改善用户满意度和自助服务
- 通过注册、指南和识别利用 BI 的新机会的能力来执行 BI 标准，导致针对战略目标调整技术和澄清未来协调的 BI 的愿景
- 向关键利益相关者介绍采用 BI 的优势

11.2 准备 BICC 的业务案例

投资回报 (ROI) 对于获得高级管理层的支持至关重要。BICC 可能需要对人员和技术的一些提前投资，所以在流程中尽早建立“硬钱”ROI 很重要。

通常，一个 BICC 将通过集中化服务器等基础架构和标准化商业智能、绩效管理、分析和数据管理工具和流程，提高 IT 效率。业务用户常常会订阅 BICC 的服务，与独立、孤立的实现相比具有显著的成本节省。共享服务中心也应该提供一个中央人才池来培训和支持业务用户。此流程将提高最终用户采用率和业务人员的自助服务水平。通过为 BICC 功能创建一个通用的位置，IT 利用规模经济及通用的培训和支持计划。

11.3 确定 BICC 的组织结构

BICC 结构因组织的需要及其成熟度水平不同而不同，如图 15.1 中所示。BICC 可能是一个仅支持 IT的计划，旨在关注对确保一致的企业BI战略所必需的系统知识进行合并。BICC 也可以由业务经理进行组织，专注于功能业务技能和得到业务高管支持的能力。一些 BICC 在企业总部办公室级别上进行集中化，

而其他 BICC 是由业务和 IT 人员组成的分散的地区、部门和职能团队网络。

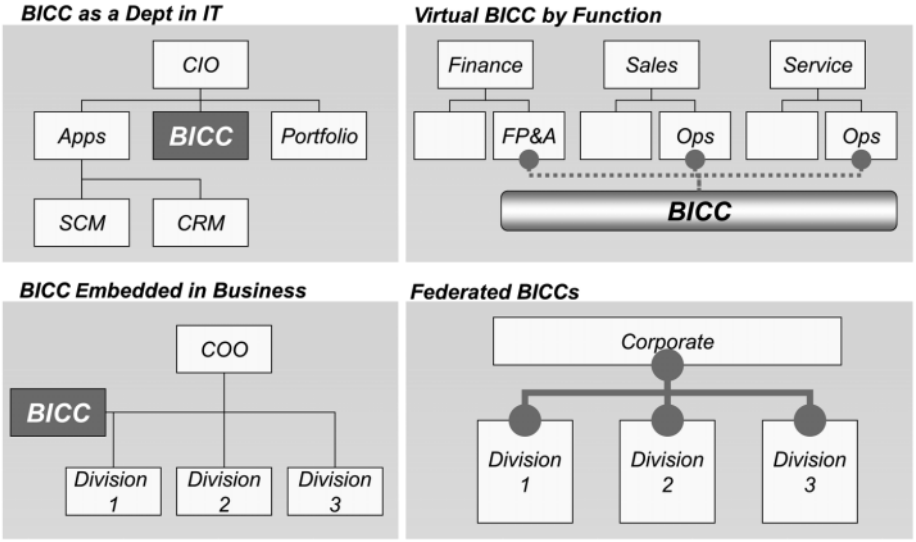


图 15.1：BICC 组织结构。

BICC 的设计可以集中化或分散化，基于全职员工或一个虚拟的社区技能集。它的组织依赖于它负责的功能和它寻求解决的问题。

11.4 协商 BICC 的关键功能

图 15.2 描述了 BICC 的典型功能：

- *建议和顾问*——BICC 为业务部门的一个功能区域提供了建议、指南、指导和内部顾问，以便项目团队可独立地解决其 BI 需要。
- *社区服务*——BICC 负责设计和构建 BI 内容，比如常见的报告和数据包，以供更广泛的业务社区使用。
- *沟通和推介*——BICC 沟通和推广 BI 计划的状态、进度、成就和成功，以促进文化转变。

- *企业技术架构*——BICC 构建和支持为业务的 BI 需要提供支持的技术基础架构。
- *支持*——BICC 向业务提供了一种 BI 服务台功能。
- *培训*——BICC 培训和教育业务用户各种 BI 技术。
- *IT 治理调整*——BICC 协调更广泛的 IT 治理流程和筹划委员会，比如项目和变更管理、产品组合管理、供应商管理和许可证管理。
- *数据治理调整*——BICC 与整个组织的现有数据治理计划进行互动。它可能位于“信息供应链”的接收端，需要值得信赖的数据。例如，在发现来自其组织不同部分的报告包含不一致的数据时，一家著名保险公司的 CEO 授权建立了一个新的数据治理计划。在另一种情形下，一家政府机构中的 BICC 部门发现它的下游分析师无法信任来自他们的 SAP 财务和核算系统的数据，但这些系统的 IT 负责人认为数据的质量“还不错”。

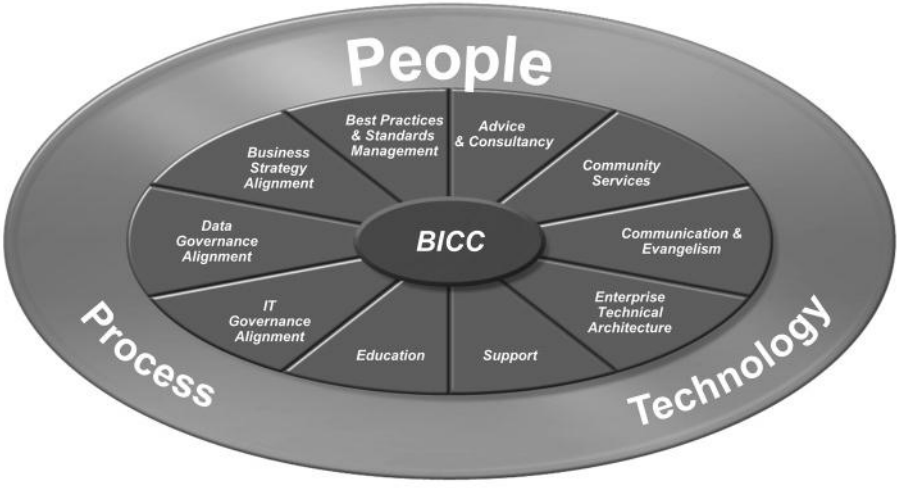


图 15.2：BICC 的典型功能。

- *业务战略调整*——BICC 与企业业务战略保持一致，以确保与技术相关的计划满足最终的需要和业务优先级。
- *最佳实践和标准管理*——BICC 提供了以后总明确的流程和存储库来在整个企业批准和共享 BI 最佳实践和标准。

16

第 12 步：管理安全和隐私

本章介绍一家组织用于减轻风险和保护数据资产的策略、实践和控制，包含来自 IBM InfoSphere Guardium 营销团队的内容。

在越来越多的企业中，数据治理组织通过设置安全和隐私战略，连同首席信息安全官 (CISO) 来掌控。数据治理安全和隐私有多种驱动因素。在此类别构建竞争计划的最大业务驱动因素已经是满足高成本的制度需求。制度需求可能特定于一个行业，或者它们可能跨越行业边界线以影响广泛的业务领域。美国众多的制度需求中的一个就是 Sarbanes-Oxley Act。此法案包含针对高层管理人员控制财务数据以实现最高完整性级别的规定。

以下是与管理安全和隐私相关联的子步骤：

12.1 与关键利益相关者协调一致。

12.2 收集企业安全架构蓝图。

12.3 加强数据库变更控制。

12.4 自动化合规性工作流程。

12.5 定义敏感数据。

12.6 发现敏感数据。

12.7 分类和标记敏感数据。

12.8 加密敏感数据。

12.9 保护非生产环境中的敏感数据。

12.10 监控应用程序中的欺诈。

12.11 预防计算机攻击。

12.12 编校非结构化文档中的敏感信息。本章剩余部分更详细介绍这些子步骤。

12.1 与关键利益相关者协调一致

这个子步骤与 IBM 数据治理统一流程的第 2 步“获取高层支持”紧密相关。以下是与专注于安全和隐私的有效数据治理计划相关的关键利益相关者：

- *CISO* 是一位重要的支持者，设定整个企业的总体安全和隐私策略。
- *首席风险官*不一定是安全组织或 IT 机构的一部分，但可能与首席财务官之间存在一个报告链。因为合规计划可能是数据安全和隐私项目的重要驱动因素，所以数据治理计划可通过改进合规性报告的准确性和降低成本，可通过首席风险官建立价值。
- *企业架构*包含首席架构师或某个为首席架构师工作的人，比如企业安全 IT 架构师。随着业务负责人部署应用程序，安全 IT 架构师在确定如何在企业的整体架构内设计、部署和实施数据治理安全和隐私计划方面扮演着关键角色。安全 IT 架构师依靠功能和非功能需求来建立应用程序开发人员需要遵守的蓝图和“路标式”标准。

- *业务支持*或业务支持的缺乏已困扰许多安全计划多年。许多业务高管对信息安全组织存在负面的看法，直到“数据破坏”等不幸事故迫使执行激进、迅速且高成本的更改。数据治理安全和隐私计划应该与业务人员合作，降低与制度需求相关的成本。数据治理安全和隐私计划的成熟度也可能基于反映动态业务条件的能力进行评估，而维持扁平的成本曲线。

当开发和实现合适的安全和隐私策略时，数据治理团队需要更加受限的信息访问带给安全和风险组织的利益与业务部门更轻松访问信息的需要。一个极端示例可能是一个完全锁定的系统。显然，系统将是高度安全的，但它将不会解决业务的需要。

12.2 收集企业安全和隐私架构蓝图

下一步是列出所有相关安全控制的清单，建立一个安全性蓝图架构。安全性蓝图架构应该用作企业 IT 和业务的参考工件。安全性蓝图架构有许多方面，类似于搭建一座房子并确定其房间的数量和位置。但是，您应该仅关注对您计划的成熟度最重要的关键区域。本章将关注安全和隐私的以数据为中心的方面。

12.3 加强数据库变更控制

依据Ron Ben-Natan编写的白皮书 *Data Security, Governance, and Privacy: Protecting the Core of Your Business*（Guardium，2006 年），大部分组织都拥有正式的策略来控制数据管理员、服务台成员和外包人员等特权用户如何和何时

能够访问数据库系统。但是，组织并不总是拥有有效的机制来监控、控制和审计这些特权用户的操作。更糟的是，由于特权用户常常共享用于访问数据库系统的凭证，很难实现责任性。

监控特权用户可在以下方面为数据治理保驾护航：

- *数据隐私*——监控可确保只有授权的应用程序和用户在查看敏感数据。
- *数据变更控制*——监控可确保关键的数据库结构和值没有在企业变更控制规程外部更改。
- *防御外部攻击*——成功、有针对性的攻击常常会导致攻击者获得特权用户访问权限。例如，在您查看用户的位置等其他标识信息之前，Uzbekistan 的一名外部人员可能看起来像内部人员，因为他拥有经过验证的访问权限。

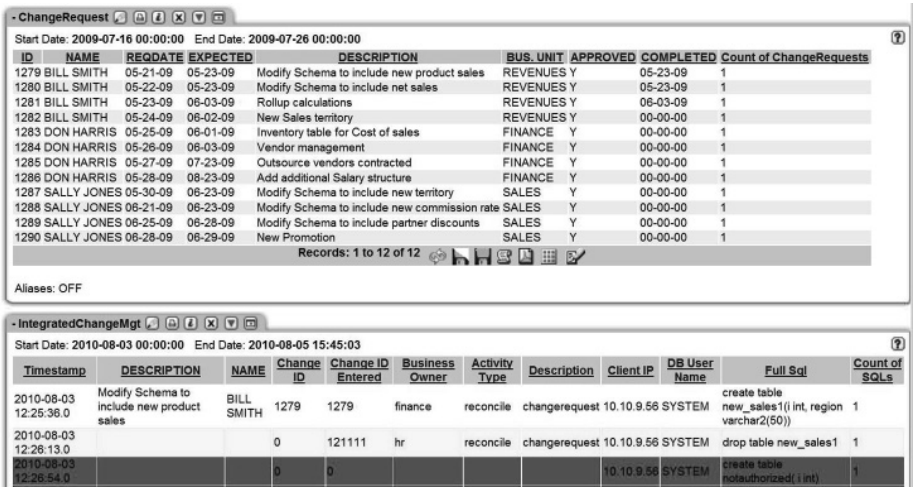
组织将希望跟踪对以下方面的所有更改：

- *数据库结构*，比如表、触发器和存储过程。例如，组织将希望检测关键表中影响业务决策质量的意外删除或插入。
- *关键数据值*，比如影响财务交易完整性的数据。
- *安全和访问控制对象*，比如用户、角色和权限。例如，一个外包合同工可能创建一个具有对关键数据库的无限访问权的新用户帐户，然后删除整个帐户，消除她的所有行为痕迹。
- *数据库配置文件*和其他外部对象，比如环境/注册表变量、配置文件（比如 NAMES.ORA）、shell 脚本、OS 文件和可执行文件，比如 Java™ 程序。

IBM InfoSphere Guardium Database Activity Monitor 提供了一个解决方案，它创建所有数据库活动的持续、详细的审计线索，包括

每个事务的“人员”、“对象”、“时间”、“位置”和“方式”。此审计线索会实时进行分析和过滤，以识别未授权的可疑活动。为了实施职权分离，所有审计数据存储在所监控数据库外部的一个安全、防篡改的存储库中。

IBM InfoSphere Guardium Database Activity Monitor 的解决方案对数据库性能具有极低的影响，不需要对数据库或应用程序执行任何更改。IBM InfoSphere Guardium Database Activity Monitor 还使组织能够自动化以下流耗时的流程：跟踪所有观察到的数据库更改，依据现有更改票证系统（比如 BMC Remedy ）和自定义变更管理应用程序中授权的工作订单对它们进行调节。例如，一家大型金融机构使用 IBM InfoSphere Guardium Database Activity Monitor 设立了一个自动化更改调节流程。图 16.1 显示了一个结果示例。在以前，该机构的 DBA 每天会花 1 小时以上的时间，使用电子表格依据经过批准的更改票证请求手动调节实际的数据库更改。



- ChangeRequest

Start Date: 2009-07-16 00:00:00 End Date: 2009-07-26 00:00:00

ID	NAME	REQDATE	EXPECTED	DESCRIPTION	BUS. UNIT	APPROVED	COMPLETED	Count of ChangeRequests
1279	BILL SMITH	05-21-09	05-23-09	Modify Schema to include new product sales	REVENUES	Y	05-23-09	1
1280	BILL SMITH	05-22-09	05-23-09	Modify Schema to include net sales	REVENUES	Y	05-23-09	1
1281	BILL SMITH	05-23-09	06-03-09	Rollup calculations	REVENUES	Y	06-03-09	1
1282	BILL SMITH	05-24-09	06-02-09	New Sales territory	REVENUES	Y	00-00-00	1
1283	DON HARRIS	05-25-09	06-01-09	Inventory table for Cost of sales	FINANCE	Y	00-00-00	1
1284	DON HARRIS	05-26-09	06-03-09	Vendor management	FINANCE	Y	00-00-00	1
1285	DON HARRIS	05-27-09	07-23-09	Outsource vendors contracted	FINANCE	Y	00-00-00	1
1286	DON HARRIS	05-28-09	08-23-09	Add additional Salary structure	FINANCE	Y	00-00-00	1
1287	SALLY JONES	05-30-09	06-23-09	Modify Schema to include new territory	SALES	Y	00-00-00	1
1288	SALLY JONES	06-21-09	06-23-09	Modify Schema to include new commission rate	SALES	Y	00-00-00	1
1289	SALLY JONES	06-25-09	06-28-09	Modify Schema to include partner discounts	SALES	Y	00-00-00	1
1290	SALLY JONES	06-28-09	06-29-09	New Promotion	SALES	Y	00-00-00	1

Records: 1 to 12 of 12

Aliases: OFF

- IntegratedChangeMgt

Start Date: 2010-08-03 00:00:00 End Date: 2010-08-05 15:45:03

Timestamp	DESCRIPTION	NAME	Change ID	Change ID Entered	Business Owner	Activity Type	Description	Client IP	DB User Name	Full Sql	Count of SQLs
2010-08-03 12:25:36.0	Modify Schema to include new product sales	BILL SMITH	1279	1279	finance	reconcile	changerequest	10.10.9.56	SYSTEM	create table new_sales1(i int, region varchar2(50))	1
2010-08-03 12:26:13.0			0	121111	hr	reconcile	changerequest	10.10.9.56	SYSTEM	drop table new_sales1	1
2010-08-03 12:26:54.0			0	0				10.10.9.56	SYSTEM	create table notauthorized(i int)	1

图 16.1：IBM InfoSphere Guardium 提供了更改自动调节。

12.4 自动化合规性工作流程

一些组织认为，时常检查他们的日志会对于通过审计就足够了。但是，审计人员希望知道 3 件无法由日志解决方案全面解决的事情：

- 您在实际保护您的数据。
- 您在监控数据库访问，可证明合规性。
- 您实现了一个正式的监督流程。

IBM InfoSphere Guardium 按计划的时间表自动生成合规性报告，并将它们分发给利益相关者供电子审批。这些报告（包括上报和签署报告）使组织能够为审计用途证明存在监督流程。

12.5 定义敏感数据

政府要求对个人可识别信息 (PII) 进行控制，如果忽略了控制，可能导致身份欺诈和缺乏对政府或企业发放的凭证的信任。欧盟建立了个人数据保护指令作为控制其成员国家的隐私保护框架。全球许多其他国家具有类似的制度。例如，美国的 Congress，制定了 Health Insurance Portability and Accountability Act of 1996 (HIPAA)，其中包括针对受保护健康信息 (PHI) 的隐私的规定。此外，行业联盟正在开发每个领域具体的治理标准，比如 Payment Card Industry Data Security Standard (PCI DSS)。

依据来自 Centers for Disease Control and Prevention (CDC) 的指南，PHI 是通过电子媒体或任何其他形式或媒介传输或维护的个人可识别健康信息。此信息与以下一个或多个方面相关：

- 个人过去、现在或未来的身体或精神健康情况或条件
- 个人健康护理服务
- 个人健康护理服务付费

如果信息表示（或提供了合理的证据证明它可用于表示）个人，它会被视为 PHI。

Payment Card Industry Security Standards Council 定义了 PCI DSS 来保护敏感的卡持有者信息。PCI DSS 适用于所有存储、传输或处理持卡人数据的金融机构、商家或服务提供商。

依据 PCI DSS，敏感的持卡人数据包括以下信息类型：

- 主要帐户编号 (PAN)
- 持卡人姓名
- 服务代码，磁条上一个 3 或 4 位的编号，指定磁条读取事务的接受需求和限制
- 有效期
- 完整的磁条数据
- 卡验证值或代码，印刷在卡上的 3 位值（对于 American Express 为 4 位值），将每张信用卡与信用卡帐号相绑定
- PIN 和 PIN 数据块

出于 PCI DSS 的用途，持卡者姓名和服务代码仅在它们与 PAN 一起存储时才被视为敏感数据。组织需要拥有数据治理策略来发现敏感的持卡者数据，屏蔽此数据以预防未授权使用。

最后，专用数据也被视为是敏感的。例如, 食品配方、内部财务报告和制造流程的知识产权应该被视为敏感数据，因为它们对组织最终的业务成功至关重要。

12.6 发现敏感数据

这个子步骤基于来自 IBM InfoSphere Optim 白皮书的内容，与第 7 步“理解数据”紧密关联。一些敏感数据很容易找到。例如，名为“credit_card_num”的列中的信用卡编号很容易识别。但是，大部分应用程序数据库更加复杂。敏感数据有时与其他数据元素组合在一起，或者埋藏在文本或注释字段中。主题专家有时能够提供有用的见解，但前提是他们全面理解该系统。

图 16.2 演示了一个示例。表 A 在“PHONE”列包含电话号码。但在表 B 中，电话号码在“TRANSACTION_NUMBER”列中一个包含时间、电话号码和日期的组合字段中模糊化了。

两个实例都表示必须保护的机密信息。尽管数据分析师可清楚识别表 A 中的电话号码，但他们可能在表 B 中忽略它。私有数据的每一次缺失都代表着组织的一个风险。

图 16.2：隐藏在组合字段中的机密信息为组织带来了一个隐私风险。

IBM InfoSphere Discovery 使组织能够识别整个环境中的机密数据实例，无论数据是明显可见还是从视图中模糊化。IBM InfoSphere Discovery 检查多个来源的数据值，以确定可能隐藏敏感内容的复杂规则和转换。正如示例中所述，它可找到较大字段中包含的或分散在多列中的机密数据项。

12.7 分类和标记敏感数据

发现敏感数据后，您必须使用“Privacy-Restricted”或“Regulated Record”等元数据分类对它进行标记。此方法使组织能够在具有类似属性的项上实现一致的访问策略和审计流程。IBM InfoSphere Guardium 向对象组自动分配可自定义、细粒度的访问策略，规定谁有权访问它们，从哪些应用程序和位置，在什么时刻，使用哪些 SQL 命令等来进行访问。

12.8 加密敏感数据

加密用于以不可读的形式呈现敏感数据，使攻击者无法获得对数据的未授权访问。数据治理组织需要

与 CISO 紧密合作解决传输中的数据和静止数据的加密。加密传输中的数据可确保攻击者无法在网络层窃听，在将数据发送到数据库客户端时获得它的访问权。静止数据的加密可确保攻击者无法提取数据，甚至通过盗窃包含数据库的实际存储媒体，比如服务器硬盘或备份磁带。

12.9 保护非生产环境中的敏感数据

嵌入到测试和培训环境中的敏感数据代表着组织的一种潜在风险。通常，活动的生态系统（可能包含机密数据）可克隆到测试或培训环境。开发人员和质量保证测试人员发现很容易处理活动的数据，因为它会生成每个人都可理解的测试结果。但是，非生产环境实际需要活动的数据吗？答案是否定的。使用逼真的数据对于测试固然重要，但活动的数据值不是绝对必要的。去标识化或屏蔽生产数据的能力提供了一种最佳实践方法来保护敏感数据，同时支持测试流程。

数据屏蔽是将机密数据元素（比如贸易秘密和 PII）系统化地转换为逼真但虚构的值。数据屏蔽表示一种简单概念。但它在技术上难以执行。找到和屏蔽数据是解决方案的一部分。但是，还有一种附带影响。您需要将屏蔽的数据元素传递到数据库中所有相关表或传递到其他数据库来维护参照完整性的能力。例如，假设屏蔽的数据元素（比如电话号码）是数据库表关系中的一个主键或外键。这个屏蔽的数据值必须传播到数据库中所有相关的表或跨越数据源。如果数据是另一行的数据的一部分，它必须也使用相同数据更新。

IBM InfoSphere Optim Data Privacy Solution 可应用各种成熟的数据转换技术，使用在上下文内准确、逼真的数据来屏蔽敏感、实际的数据。用户可屏蔽一个数据库或多个相关系统中的数据。这些能力使得很容易去标识化许多类型的敏感信息，比如生日、银行帐号、街道地址和国家标识符，比如加拿大的社会保险编号或意大利的

Codice Fiscale。图 16.3 提供了 IBM InfoSphere Optim Data Privacy Solution 的数据屏蔽功能的一个示例。

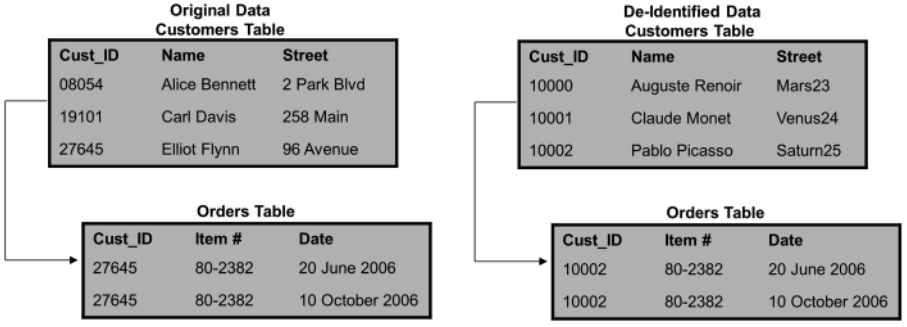


图 16.3：IBM InfoSphere Optim Data Privacy Solution 的数据屏蔽。

12.10 监控应用程序中的欺诈

多级企业应用程序（比如 Oracle 和 SAP）以及甚至依靠应用服务器（比如 IBM WebSphere）的应用程序使用一种称为*连接池*的优化机制在数据库连接级别上屏蔽最终用户的身份。使用池连接，应用程序将一些仅由一般服务帐户名称所标识的数据库连接中的所有用户流量聚合在一起。结果，组织发现很难将特定的数据库事务与特定的应用程序最终用户相关联。

应用层监控的主要用途是检测通过企业应用程序而不是通过对数据库的直接访问而发生的欺诈和其他合法访问权滥用行为。IBM InfoSphere Guardium Database Activity Monitor 通过在网络和 OS 级别上从数据库外部观察应用程序与数据库服务器之间的交互，监控应用程序用户 ID。

12.11 预防计算机攻击

边缘防御（比如防火墙和反病毒系统）不再足以防御有目的性的计算机罪犯，这些罪犯使用复杂的技术渗透到后端数据库中。

SQL 注入是攻击者用于攻击 Web 应用程序漏洞的代码注入技术的一个示例。组织需要在以下区域创建和

执行实时、前瞻性的策略（参阅附录 E 获取数据治理声明集合示例）：

- *访问策略*通过连续对比所有数据库活动与正常行为基准，识别反常的行为。例如，SQL 注入攻击通常会表现出不符合标准应用程序特征的数据库访问模式，比如通过攻击他们存储盗窃来的数据或恶意软件的地方来创建新表。
- *异常策略*基于可定义的阈值，比如过量的失败登录或 SQL 错误。SQL 错误可能表明，攻击者正在通过使用不同的参数试验 SQL 命令（比如“Credit_Card_Num”或“CC_Num”）来“寻找”关键表的名称。异常策略也可能基于来自数据库的特定 SQL 错误代码，比如“ORA-00903: Invalid table name”或“ORA-00942: Table or view does not exist”。这些错误代码可能表明存在攻击行为。
- *排除策略*检查使数据库出现特定的数据值模式的数据，比如信用卡编号或可能表明存在破坏的大量返回记录。
- *预先配置的策略签名*识别对未修补的漏洞或系统功能（比如具有已知漏洞的系统存储过程或没有禁用的默认系统帐户）的攻击尝试。

IBM InfoSphere Guardium Database Activity Monitor 提供了可全面自定义的策略违规响应，可能包括：

- SNMP 和 SMTP 真实警报
- 自动终止，比如从数据库系统注销帐户或 VPN 连接关闭
- 可在违反策略（比如外包的 DBA 尝试查看或更改敏感的表）时立即终止会话的阻塞，基于主机的代理
- 将策略违规转发到企业级安全信息和事件管理 (SIEM) 系统，比如 IBM Tivoli SIEM

最后, IBM InfoSphere Optim pureQuery 在设计时识别 SQL 注入威胁，以确保它们从不会带到生产环境中。

12.12 编校非结构化文档中的敏感信息

依据白皮书 *IBM Optim Data Redaction: Reconciling Openness with Privacy*（Joshua Fox 和 Michael Pelts，IBM，2010 年），编校是从信息来源删除敏感内容的过程。编校通常可在纸张文档上通过黑色记号笔或白色的涂改液或在数字文档上使用类似的电子工具来完成。许多类型的文档都需要编校，比如课税扣押权证、房契、出生证明、医院出院小结和患者病史。

数据治理组织在处理编校解决方案时需要平衡开放和隐私的双重目标。例如，制度通常指定具有有效业务目的的人才可以查看某些实体。因此，医生可以查看患者的医疗信息，但无法查看敏感的财务信息，而开票员正好相反。类似地，对于 eDiscovery，United States Federal Rules of Civil Procedure 要求诉讼当事人的律师可以查看特权客户-律师信息，而对方的辩护律师则不能。在一些情形下，法官可以看到所有形式的信息。

IBM InfoSphere Optim Data Redaction 使用测试提取技术来实现已由数据治理计划建立的编校策略。

17

第 13 步：治理信息生命周期

信息生命周期治理指的是一种系统、基于策略的信息架构、分类、收集、使用、存档、保留和删除方法。因为信息生命周期治理是一个专门的能力区域，所以本章打算仅为数据治理从业者提供简短的概述。

IBM Information Lifecycle Governance 是一个全面的合规性平台，使组织能够控制和管理其信息的寿命。从数据治理的角度讲，它可帮助组织解决以下挑战，如图 17.1 中所示：

- *内容评估*——解决未进行管理的“野外的数据”，有助于评估和决定管理、信任和利用哪些信息。
- *内容收集和归档*——管理激增的信息量和类型。
- *高级分类*——减少最终用户的负担并提高分类信息的能力。
- *记录管理*——执行保留和处理策略，自信地公开信息。

- *eDiscovery 搜索和分析*——迅速、经济高效地响应 eDiscovery、审计和内部调查请求。

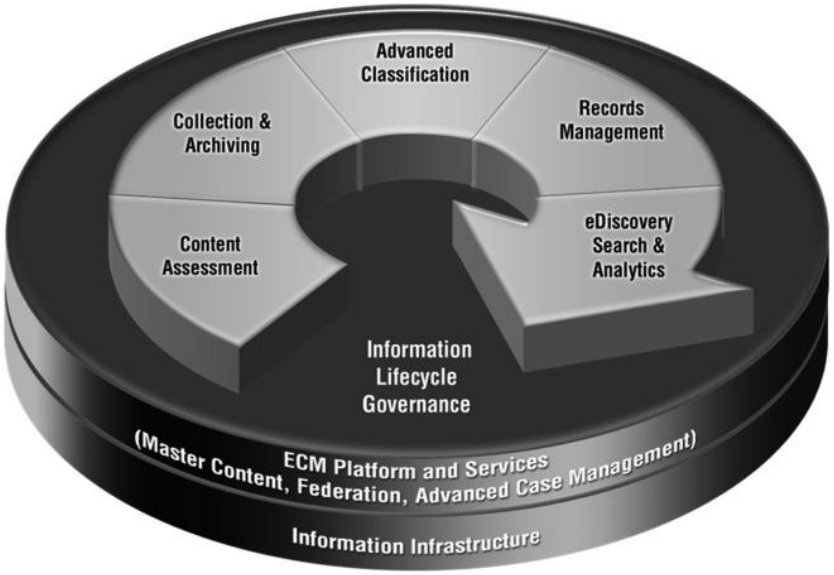


图 17.1：IBM Information Lifecycle Governance 模型。

以下是与治理信息生命周期相关联的子步骤：

13.1 建立信息架构。

13.2 建立数据库大小和存储架构基准。

13.3 发现业务对象。

13.4 分类数据和定义服务水平。

13.5 归档数据和非结构化内容。

13.6 建立管理测试数据的策略。

13.7 定义电子文档法律查询策略。

13.8 分析内容。

下面更详细地介绍这些子步骤。

13.1 建立信息架构

数据治理团队需要确保组织为信息架构设定了标准。更重要地，数据治理委员会需要拥有实施架构标准的权利。信息架构在提高整体 IT 效率上发挥着重要作用。例如，在组织寻求减少许可证、软件维护和支持成本时，工具的标准化和遗留应用程序的退役至关重要。对于像一个县级医疗系统内跨医院和诊所的实验室测试代码这样的实体，标准的命名约定也很重要。

13.2 建立数据库大小和存储架构基准

获取对哪些区域积累了大部分信息的全面理解，使组织能够应用最有效的信息生命周期治理战略。以下信息来自 IBM 白皮书 *Control Application Data Growth Before It Controls Your Business*（2009 年 9 月）。

数据重复对统计数据的增长具有显著影响。组织常常克隆或复制生产数据库来支持其他功能，或用于应用程序开发和测试。它们还维护着关键数据的多个备份副本，或者实现镜像的数据库来防御数据丢失。最后，灾难恢复计划需要数据复制，以将关键数据存储在一个替代位置。

所有这种重复创造了所谓的“数据倍乘器效应”。因为数据是重复的，所以存储和维护成本会成比例增长。图 17.2 提供了一个生产数据库的示例，

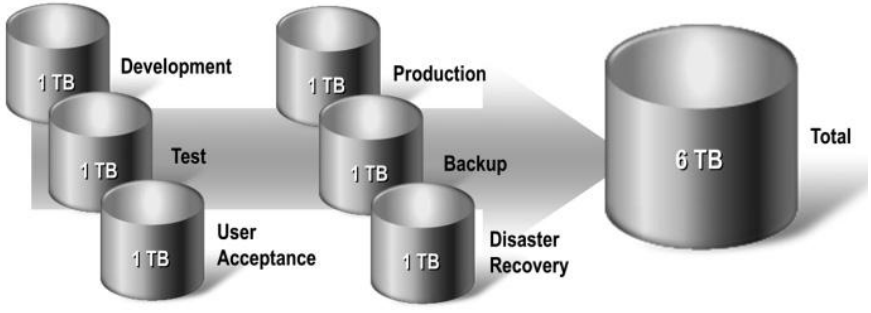


图 17.2：实际数据负担等于生产数据库的大小加上所有复制的克隆。

其中包含 1 TB 数据。当复制该数据库以用于备份、灾难恢复、开发、测试和用户验收时，总数据负担将从 1 TB 增长到 6 TB。

13.3 发现业务对象

如果不理解数据，您就无法治理它，所以首先使用数据发现备案您的现有数据范围很重要。数据发现可分析数据值和模式，识别将不同数据元素链接到逻辑信息单元或*业务对象*（比如客户、患者或发票）中的关系。

这些业务对象为归档提供了重要输入。没有识别数据关系和定义业务对象的自动化流程，组织可能要花几个月时间来执行手动分析，无法保证完备性或准确性。IBM InfoSphere Discovery 自动识别关系并定义业务对象。它已在第 9 章“理解数据”中详细介绍。

13.4 分类数据和定义服务水平

典型企业中的非结构化内容量每年都在以令人惊讶的速度增长。从数据治理的角度讲，将这些海量信息进行编目，以便可通过内容管理系统有效管理它，这非常重要。法律发现是文档和电子邮件分类重要性的一个不错的业务示例。

考虑专注于诉讼的企业。它需要确保潜在的相关文档和电子邮件得到自动分类并受到记录管理系统的控制。该企业需要为每个文档分配合适的保留和处理规则。要控制存储和法律审查成本，该企业需要过滤掉不相关的数据，比如公司公告、实事通讯、个人电子邮件和与正在处理的法律案例没有关系的个人文档。记录管理团队首先得到法律团队提供了的一组关键词，随后不断细化这些关键词以确保分类系统最高效地工作。

IBM InfoSphere Classification Module 支持企业通过分析文档和电子邮件的全部文本，创建新分类或向现有分类添加内容。

13.5 归档数据和非结构化内容

让我们考虑一个来自电信行业的示例。典型的大型电信公司会定期生成大量数据。例如，一家电信公司每年将生成数百万条呼叫细节记录 (CDR)，这些记录需要存储和分析。CDR 存储逐渐变得单调和昂贵。一家大型典型公司发现，它拥有接近 4 TB 的 CDR 数据，该数据在过去两年几乎翻了一翻。该电信公司有多个 CDR 数据版本在归档、数据仓库中以及用于分析。结算应用程序拥有 12 个数据库克隆。还有其他用于灾难恢复的副本。该电信公司甚至发现它拥有的大量磁带中包含它从不知道存在的数据。

数据治理计划被要求围绕存档、保留和删除设置一条策略，以减少存储成本而不对业务造成负面影响为目标。一些策略问题包括“我们需要在仓库中维护几个月的 CDR 数据？”和“对于 3 个月以前的数据，我们能否仅在仓库中保留摘要数据？”

从数据治理的角度讲，组织需要同时归档结构化和非结构化内容，以减少存储成本，改善系统性能，确保遵守制度需求。具体来讲，电子邮件和其他文档形式的非结构化内容占典型企业中内容的 80% 以上。诚然，一些组织已认识到这一模式的转变，并重新将他们的计划标榜为“信息治理”。此内容需要归档以减少存储成本。

IBM InfoSphere Content Collector 是一个归档解决方案，旨在归档 Lotus® Domino®、Microsoft Exchange、Microsoft SharePoint® 和 Windows® 文件系统中的内容。此外，IBM InfoSphere Content Collector 允许使用 IBM InfoSphere Enterprise Records 将内容动态地声明为记录。最后，它利用来自 IBM InfoSphere Classification Module 的元数据制定动态的分类决策。

依据业务价值来存储归档的数据，是整合的数据管理战略的一个逻辑组成部分。一种 3 层分类战略是解决该问题的一种有用方式。当前的事务在高速的主要存储中进行维护。报告数据转移到中层存储。参考数据保留在一个安全的 Write Once, Read Many (WORM) 设备中，使它在应该提出审计请求时可用。这种分层存储方法和归档战略是减少成本和最大化业务价值的一种不错方式。IBM InfoSphere Optim Data Growth 提供了成熟的数据库归档功能，支持组织将历史数据与当前数据分开，安全且经济高效地存储它，同时维持统一的访问。

13.6 建立管理测试数据的策略

依据白皮书 *Enterprise Strategies to Improve Application Testing*（IBM，2008 年 4 月），仅为了测试用途而克隆包含数百个不相关的表的整个生产数据库通常不切实际。首先，存在针对测试配置全新的数据库环境的容量、成本和时间问题，第二，存在质量问题，当处理大型测试数据库时，开发人员可能发现难以跟踪和验证特定的测试案例。

以下是有效测试数据管理的一些需求：

- *创建逼真的数据。*创建更小、逼真且准确反映应用生产数据的数据子集很重要。
- *保留测试数据的参照完整性。*数据子集需要考虑在数据库和应用程序中实施的参照完整性。通常，应用程序执行的参照完整性更加复杂。例如，应用程序可能包含使用兼容但不同的数据类型、组合和部分列的关系，以及数据驱动的关系。
- *执行错误和边界条件。*从生产数据库创建逼真的相关测试数据子集是一个不错的开始。但是，有时必须编辑数据以执行特定的错误条件，或验证特定的处理功能。

- *屏蔽和转换测试数据。*随着对数据隐私的关注越来越高，转换和去标识化开发和测试环境中的敏感数据对于避免数据破坏和严重的惩罚至关重要。
- *对比之前和之后的测试数据。*对比一系列测试之前和之后的测试数据对应用程序的整体质量至关重要。此过程涉及到将每次测试迭代与基准测试进行对别，识别可能未检测到的问题，尤其是当测试可能影响数百或数千个表时。

IBM InfoSphere Optim Test Data Management Solution 简化测试环境的创建和管理，设置数据子集并迁移数据以构建逼真且具有正确大小的数据库，自动化测试结果对比，消除维护多个数据库克隆的开支和工作。

13.7 定义电子文档法律查询策略

查询还是和解？这是全球的企业律师都在问的问题，尤其是如果他们参与到美国联邦法庭系统和修订的 Federal Rules of Civil Procedure (FRCP) 中的诉讼。要进行电子查询的电子存储信息 (ESI) 量通常非常巨大，以至于和解诉讼常常比执行进行诉讼所必要的大量查询流程要省钱得多。事实上，在某些公司，防御诉讼的内部 IT 和外部成本可能超出 100 万美元。由于制裁、罚金和对企业声誉的损害，不遵守 FRCP 驱动的需求的成本很容易变得很高。组织需要自动化的工具来获取对案例相关内容的访问和早期洞察，以在其碰面商谈时间内形成 eDiscovery 计划。

IBM InfoSphere eDiscovery Manager 和 IBM InfoSphere eDiscovery Analyzer 可减少 eDiscovery 成本。例如, 当收到查询请求时，组织可使用 IBM InfoSphere eDiscovery Manager 执行关键词或日期范围搜索，收集可能相关的 ESI。

单独这一步可能在包含 100 万项的归档文件中识别出 10 万个可能相关的内容片断。

更进一步，IBM InfoSphere eDiscovery Analyzer 可迅速识别和标记响应迟钝的内容，将与案例相关的内容池潜在地缩小 10 到 15%。因为每封电子邮件的外部分析需要花大约 1 美元费用，而且因为许多组织管理着数百个有效案例，迅速筛除“噪音”可帮助组织显著减少 eDiscovery 审查成本。

13.8 分析内容

内容分析是一个新兴的分析领域，使公司能够释放非结构化内容中包含的洞察。这种非结构化内容可能包括表单、文档、数据库中的注释字段、网页、客户信件和其他未存储在结构化数据字段中的信息。内容分析能够访问、排序和分析内容，然后将它与结构化数据和其他现有的信息资源和应用程序相组合，以供报告和分析。

内容分析是商业智能的一种自然扩展。许多组织已使用商业智能来指定“数据驱动的决策”。这种决策过程基于从过去的事务记录和通常位于数据仓库中的其他非结构化信息收集的洞察。组织可使用内容技术为这些商业智能方法提供补充，这些技术可用于分析非结构化内容中的趋势。例如，组织可分析内容来解决关键业务问题，比如：

- 基于保险索赔表格识别欺诈性的索赔。
- 基于对呼叫中心记录文本的分析来度量和监控客户服务度量指标。
- 基于保修记录的分析来计划产品发布优先级。
- 基于竞争对手备案文件和得失数据中的文本的分析来开发成功的竞争销售战略。

IBM Cognos Content Analytics 是一个为组织提供工具来释放非结构化内容中包含的业务洞察的解决方案。

18

第 14 步：度量结果

许多数据治理计划仅仅由于已存在独立治理数据的计划而失败。IBM 数据治理统一流程中的最后一步是依据预先定义的 KPI 集度量结果，确保计划继续催生业务价值。这些结果需要定期传达给数据治理委员会和高层管理人员。进度度量指标将确保对数据治理计划的持续支持和资金支助。

粽子，本书旨在为组织应该如何实现一个数据治理计划提供一个模板。它的目的在于从数据治理中消除一些猜测，使实现成功的计划更加容易。

尽管数据治理从不能完全自动化，但 IBM 拥有软件工具和测试实践来简化总体流程，正如您在本书中所看到的。

附录 A

数据治理统一流程中的步骤和子步骤

1. 定义业务问题

2. 获取高层支持

2.1 创建虚拟数据治理工作组

2.2 获取 IT 和业务部门内高级管理人员的支持

2.3 识别数据治理的负责人

3. 执行成熟度评估

3.1 定义评估的组织范围

3.2 定义想要的数据治理未来状态的时间范围

3.3 定义要评估的数据治理类别

3.4 确定业务和 IT 部门中正确的研讨会参与者

3.5 执行数据治理成熟度评估研讨会

3.6 与高层管理人员沟通评估结果

4. 创建路线图

4.1 总结数据治理成熟度评估的结果

4.2 列出填补评估中强调的差距所需的关键人员、流程和技术计划

4.3 基于关键计划的优先级创建路线图

5. 建立组织蓝图

5.1 定义数据治理章程

5.2 定义数据治理的组织结构

5.3 建立数据治理委员会

5.4 建立数据治理工作组

5.5 确定数据监管人

5.6 举行数据治理委员会和工作组定期会议

6. 创建数据字典

6.1 选择一个数据领域

6.2 安排数据照管人来维护关键业务词汇

6.3 识别关键数据元素

6.4 从现有的词汇术语表创建数据字典

6.5 填充数据字典

6.6 链接业务词汇与技术工件

6.7 支持数据治理审计、报告和日志需求

6.8 整合数据字典与应用程序环境

7. 理解数据

7.1 理解范围内的每个数据源

7.1.1 执行列和表级别分析

7.1.2 通过逆向工程主-外键关系查询遗留模式

7.1.3 识别每个来源中的关键数据元素的位置

7.1.4 识别每个来源中的敏感数据的位置

7.2 理解来源之间的关系

7.2.1 理解关键数据元素在各个数据源之间的数据重叠情况

7.2.2 发现来源之间的数据连接和复杂转换逻辑

7.2.3 发现数据不一致性和异常

8. 创建元数据存储库

8.1 合并来自数据字典的业务元数据和来自发现流程的技术元数据

8.2 确保合适的数据血统

8.3 执行影响分析

8.4 管理操作元数据

9. 定义度量指标

9.1 理解业务的整体关键绩效指标 (KPI)

9.2 定义数据治理的业务驱动 KPI

9.3 定义数据治理的技术 KPI

9.4 建立数据治理成熟度评估的仪表板

10. ***可选专题:*主数据治理**

10.1 任命数据照管人

10.1.1 任命首席数据照管人

10.1.2 确定数据照管计划的配置
（比如由 IT 系统、组织或主题区域执行）

10.1.3 确定每个数据领域的高层支持者

10.1.4 招聘每个数据领域的数据照管人

10.1.5 授权数据治理委员会监督数据照管计划

10.2 管理数据质量

10.2.1 建立数据质量策略, 包括高价值数据属性的识别

10.2.2 设置数据质量基准

10.2.3 创建业务案例

10.2.4 清理数据

10.2.5 持续监控数据质量

10.3 实现主数据管理

10.3.1 识别业务问题

10.3.2 定义主数据主题区域

10.3.3 识别使用数据的系统和业务流程

10.3.4 识别当前的数据源

10.3.5 定义记录系统的数据属性

10.3.6 为每个记录系统任命数据照管人

10.3.7 建立主数据治理策略

10.3.8 为手动干预和监控实现数据照管控制台

10.3.9 管理潜在的重叠任务

10.3.10 匹配来自相同来源或多个来源的可疑重复内容, 创建一个新主记录

10.3.11 链接来自多个来源的相关记录

10.3.12 检查唯一标识符是否重复

10.3.13 管理关系

10.3.14 管理层次结构

10.3.15 管理分组

10.3.16 构建主数据管理解决方案

11. ***可选专题:*治理分析**

11.1 定义 BICC 的目标

11.2 准备 BICC 的业务案例

11.3 确定 BICC 的组织结构

11.4 协商 BICC 的关键功能

12. ***可选专题:*管理安全和隐私**

12.1 与关键利益相关者协调一致

12.2 收集企业安全架构蓝图

12.3 加强数据库变更控制

12.4 自动化合规性工作流程

12.5 定义敏感数据

12.6 发现敏感数据

12.7 分类和标记敏感数据

12.8 加密敏感数据

12.9 保护非生产环境中的敏感数据

12.10 监控应用程序中的欺诈

12.11 预防计算机攻击

12.12 编校非结构化文档中的敏感信息。

13. ***可选专题:*治理生命周期信息**

13.1 建立信息架构

附录 B

示例数据治理章程（针对一家制造公司）

数据治理的定义

数据治理是对组织内的人员、流程、技术和策略进行编排，以企业资产的形式利用、优化和最大化数据的过程。就像我们的董事会治理企业以最大化股东价值一样，数据治理计划旨在最大化数据对关键 IT 和业务部门利益相关者的价值。

业务目标

组织拥有一个稳定的 SAP 环境，但不是我们的所有数据都位于 SAP 中。数据治理的范围应该不断扩大，只要该计划能够证明成功。但是，在接下来的 12 个月内，数据治理计划将重点关注对以下业务目标的支持：

1. 客户中心性
2. 供应商管理
3. 供应链优化
4. 财务数据质量
5. 电子文档和电子邮件生命周期的管理

高层支持者

首席信息官将充当计划的总体高层支持者，与业务经理和关键职能区域紧密合作。

组织

数据治理计划的中心角色将是数据治理主管。

数据治理委员会将对计划拥有最高监督职责。委员会将由首席信息官主持，将包含营销部门高级副总裁、供应链部门高级副总裁、首席财务官和总顾问。委员会将每月举行一次会议，至少在最初是这样。数据治理主管将负责连同首席信息官和其他成员一起设置数据治理委员会的议程。

数据治理工作组将由数据治理主管主持。工作组将包含来自营销、供应链、财务、法律、数据架构和内容管理部门的成员。这些成员将入则对各自职能区域内的数据相关问题进行日常处理。数据治理工作组将每周举行一次会议，至少在最初是这样。

度量指标

数据治理工作组负责建立关键度量指标计分卡来监控数据治理计划的性能。数据治理主管负责定期向数据治理委员会报告这些度量指标。

数据照管社区

数据照管人拥有对其领域内的数据的保管责任。关键数据领域的高层支持者如下所示：

- 客户数据：营销部门高级副总裁
- 供应商和材料数据：供应链部门高级副总裁
- 财务数据：首席财务官

高层支持者将任命数据照管人来负责每天的数据质量。数据治理主管将负责监督，确保数据照管计划在整个组织内一致地实施。数据照管人将负责每月定义、收集和报告与他们各自的数据领域相关的关键度量指标。

附录 C

示例工作描述(针对一个数据治理官)

目标

- 担当将数据视为企业资产的中心角色，类似于任何其他物理资产。

负责的业务

- 确保数据治理计划与关键业务优先级（比如企业营销部门所支持的客户中心性计划）协调一致。
- 改进客户数据的质量，基于信息来优化决策的有效性。
- 推动将数据的所有权交给业务部门。

度量指标

- 定义关键绩效指标 (KPI) 来监控和跟踪数据治理结果。
- 定期向数据治理委员会及 IT 和业务部门内的关键高层利益相关者报告结果。

组织

- 促进整个组织的对数据治理计划收益的可视化和认知。
- 设置数据治理委员会定期会议的议程，主持会议，以及管理结果。
- 确保关键 IT 和业务部门利益相关者始终参与到数据治理计划中，确保数据治理委员会将精力集中在业务所面临的正确战略问题上。
- 督促数据治理工作组定期举行会议，确保它的活动与数据治理委员会的活动保持一致。
- 督促数据照管人向业务部门报告。确保数据照管人一致地操作，业务可不断看到计划中的价值。

附录 D

示例数据治理成熟度评估调查问卷

此附录提供一个示例问题集来评估数据治理计划的成熟度级别。这些问题应该用于在 1（最低）到 5（最高）的量表上对计划进行评分。

以下是每个成熟度级别的指导准则，基于软件工程院所开发的 Capability Maturity Model (CMM)：

- *成熟度级别 1*（初始）——流程通常是临时的，环境也不稳定。成功反映组织内个人的能力，而不是成熟流程的使用。尽管处于级别 1 的组织常常会生成有效的产品和服务，但他们常常会超出预算和项目时间表。
- *成熟度级别 2*（管理）——成功是可重复的，但流程可能无法为组织内所有的项目而重复。基本的项目管理有助于跟踪成本和时间表，而流程学科有助于确保保留了现有的实践。当这些实践就绪之后，项目就会依据它们所备案的计划执行和管理。但是，仍然存在超出成本和预计时间的风险。

- *成熟度级别 3*（定义）——组织的标准流程集用于在整个组织中建立一致性。对组织的标准流程集中的项目标准、流程描述和规程进行调整，以适合特定的项目或组织部门。
- *成熟度级别 4*（定量管理）——组织设置流程和维护的数量质量目标。所选的子流程对整体流程性能具有重大贡献，使用统计技术和其他量化技术来控制。
- *成熟度级别 5*（优化）——量化的流程改进目标被明确地建立并继续修订以反映不断变化的业务目标，以及用作管理流程改进的条件。

以下是用于评估数据治理成熟度的示例问题列表。

1. 数据风险管理和合规性

- 数据治理计划与组织的整体风险管理框架的绑定程度有多高？
- 风险管理是否是数据治理委员会中的关键利益相关者？
- 您是否对数据治理计划如何改进风险管理整体效率执行了评估？
- 风险管理组织是否同意此评估？
- 您是否定义了一组度量指标来从风险管理角度监控数据治理计划的性能？
- 您是否有依据这些度量指标定期跟踪、分析并向数据治理委员会报告的流程？

2. 价值创建

- 您是否已确定数据治理计划的关键业务利益相关者？
- 您是否确定了数据治理计划的关键业务收益？
- 您是否就数据治理的业务收益获得了关键业务利益相关者的签字认可？
- 您是否开发了一个业务案例来支持特定的数据治理计划？
- 您是否定义了关键的业务驱动度量指标来监控数据治理计划的性能？
- 您是否有依据这些度量指标定期跟踪、分析并向数据治理委员会报告的流程？

3. 组织结构和感知

- 您的高层管理人员对将数据视为企业资产的支持程度如何？
- 您的整个企业中对将数据视为企业资产的认知水平如何？
- 您是否有包含来自业务部门的高层参与者的数据治理委员会？
- 您是否有包含来自业务部门的高层参与者的数据治理工作组？
- 您是否有全职的数据治理官？
- 您是否建立了一个数据治理章程，拥有在高层领导和数据治理委员会之间达成一致且明确定义的目标？

4. 策略

数据治理策略在企业将如何管理其数据方面提供了总体方向。数据治理策略示例包括：

- *多险种保险公司的主数据管理*——保险客户数据归企业所有，而不是归生命、财产和意外以及退役服务等的各个业务经理所有（多险种保险公司）。
- *制造商的数据照管*——数据照管人将将客户、产品和供应商数据作为核心主题区域重点对待。客户、产品和供应商区域的数据照管人将分别向销售、研发和供应链组报告。数据照管人将负责确保其各自主题区域内的数据质量。
- *元数据*——数据治理办公室将维护关键业务词汇的一个数据字典。数据照管人将负责确保其主题区域内的数据定义的正确性。
- *隐私*——数据治理办公室将维护所有包含个人可识别信息 (PII) 的数据库字段的记录。对访问这些字段的所有新数据库请求将经过首席隐私官或她委派的人批准。DBA 等特权用户不能访问 PII。此策略也适用于拥有数据库访问权限的顾问和外包人员。
- *记录管理*——将由记录管理团队开发一个记录管理策略，涵盖所有类型的文档，包括纸张、电子和电子邮件。此策略将决定文档类型的主要列表，设置每个文档类型的保留时间表，以及就用于实现该战略的工具达成一致。

以下是评估围绕策略的数据治理成熟度水平的总体问题：

- 数据治理计划参与策略设置的程度如何？
- 数据治理策略是否已备案？

- 数据治理委员会审查和更新数据治理策略的频率是多少?
- 数据治理计划参与策略执行的程度如何?

5. 照管

- 组织是否拥有数据照管人?
- 数据照管人如何协调一致(按 IT 系统、组织或主题区域)?
- 哪些 IT 系统与数据照管计划协调一致(CRM、ERP、财务、其他)?
- 哪些业务组织与数据照管计划协调一致(销售、营销、财务、风险、其他)?
- 哪些主题区域由照管计划管理(客户、代理、供应商、会计科目表、员工、位置、产品、材料、其他)?
- 每个数据领域是否拥有负责使数据“适合其业务用途”的高层支持者?
- 每个主题区域是否有企业级数据照管计划?
- 数据照管人是否负责定义其各自数据领域的属性(例如,客户数据照管人是否负责词汇“客户”的定义?)
- 数据照管人是否负责定义和监控与其领域内的数据质量相关的关键度量指标?捕获这些指标的频率是多少?
- 数据照管计划是否在整个组织内是公开的并得到支持?
- 业务部门是否认识到了数据照管计划的价值?
- 是否有负责数据照管计划在整个组织的总体一致性的首席数据照管人?

6. 数据质量管理

- 组织是否拥有已由数据治理委员会达成一致的标准数据质量度量指标集合？
- 您是否备案了您系统中的数据质量问题？
- IT 和业务之间是否就您系统中的数据质量问题大体达成了一致？
- IT 和业务之间是否就糟糕的数据质量的根源达成了一致？
- 您是否开发了业务案例来解决数据质量问题？
- 您是否使用任何数据质量工具？
- 您收集数据质量度量指标的频率是多少？
- 分析并向数据治理委员会报告这些度量指标的流程是什么？
- 基于数据治理委员会的反馈对数据质量度量指标采取更正措施的流程是什么？

7. 信息生命周期管理

- 您在多大程度上围绕将数字化的纸张文档类型设置了策略？
- 您在多大程度上围绕纸张文档、电子文档和电子邮件的保留和删除设置了策略？
- 您在多大程度上围绕电子信息（无论是结构化的还是非结构化的）的归档设置了策略？
- 您是否在关键价值驱动因素（比如改进的系统性能和更低的存储成本）方面建立了归档业务案例？
- 您在多大程度上自动化了 eDiscovery 的流程？
- 您在多大程度上拥有针对适应性需要而调节的自动化内容收集和分类规则？

- What percentage of your core business content (例如, lending documents, for a bank) is being leveraged for analytics?

8. 安全和隐私

- 数据治理计划是否设置了安全和隐私策略？
- 首席信息安全官是否是数据治理计划的关键支持者？
- 您的组织是否受到任何隐私制度（比如 PCI DSS 或 HIPAA）约束？
- 您是否未能通过任何隐私审计？
- 您是否加密了您系统中的任何 PII 或受保护的健康信息 (PHI)？
- 您是否在开发或测试系统中使用了未加密的 PII 或 PHI 数据？
- 您是否有可以访问未加密的 PII 或 PHI 数据的数据库管理员、合同工和其他第三方？
- 您是否在监控拥有超级用户特权的用户（比如数据库管理员）对 PII 和 PHI 数据的访问？

9. 数据架构

- 数据治理委员会是否设置了数据架构标准（比如数据库、报告工具、分析工具、ETL 工具、主数据管理、内容管理工具和记录管理工具）？
- 是否有对已由数据治理委员会建立的数据架构标准实施合规的流程？
- 您是否有合理化数据架构的业务案例？
- 您是否在记录系统中识别了特定的主题区域，比如客户、供应商和产品？

10. 分类和元数据

- 您是否有一个针对关键业务词汇的字典？
- 数据字典涵盖了哪些主题区域（比如客户、产品和供应商）？
- 数据字典涵盖了哪些业务组织（财务、风险、营销、销售、其他）？
- 在数据字典中填入的各个主题区域或业务组织的关键业务词汇占多大比例？
- 您是否就数据字典中的词汇与关键业务区域达成一致？
- 您是否拥有词汇“元数据”的标准定义？
- 您是否拥有元数据架构师？
- 您是否拥有技术元数据存储库？
- 您的技术元数据存储库是否有助于影响分析？（例如，如果丢失了一行或一个表，对数据架构有何影响？）
- 您的技术元数据存储库是否支持一直到源系统的数据血统？
- 您是否在捕获关键操作元数据度量指标？
- 您是否开发了业务案例来支持技术、业务和操作元数据？

11. 审计信息日志和报告

- 您在多大程度上拥有战略来减少进程外数据库修改的数量？
- 您在多大程度上拥有合适的内部控件来确保为财务和制度用途而验证报告？
- 您在多大程度上拥有监控超级用户（比如 DBA）所做变更的能力？
- 您在多大程度上拥有审计对关键数据的所有变更的能力？

附录 E

示例数据治理声明

作者：Marty Moseley

Marty Moseley 是 IBM Information Agenda 医疗转型部门的首席架构师，是 Initiate Systems, Inc 的前首席技术官。

声明是数据治理计划的主要交付结果。它们基于常常难以在所有所有影响的相关方（通常是为数据治理委员会服务的业务领导）之间达成一致的共同愿景和范围。它们使业务部门能够调整其操作来满足不同的业务目的和目标，这些目的和目标在数据治理中始终是最重要的。它们还为负责运营业务的人员提供了具体的方向和指南。一些声明由数据照管人编写，一些由业务分析师编写，另一些由架构师编写。解决数据的跨职能共享和管理的所有声明都应由数据治理委员会批准或签认。

要使声明生效，它们应该是“智慧 (SMART)”的：

- 具体声明“谁”、“什么”、“何时”、“何处”、“如何”和“为什么”
- 可度量和可审计
- 可操作
- 逼真且与业务目标相关
- 确凿的——不是抽象的、理论上的、模糊的或含糊的

数据治理计划的声明可分为几类：原则、策略、规程、业务规则和度量指标。

原则

原则是对为什么某类数据的质量对组织至关重要的最高级声明。它们解决该类数据对组织的价值或实现概述的一定质量水平的价值。它们可以确定组织的使命、使行为更符合预期，确定从最终结果受益的人，因此它们最可能是公司的使命陈述或用途陈述。原则通常将通过回答以下问题，声明监督的广度、范围和细节：

- 为什么执行此数据治理计划？
- 此计划背后有哪些业务需要、风险、目标、成本和机会？
- 为什么需要执行以下策略和操作？

原则由数据治理委员会编写、批准和签认。它们在最高组织级别（而不是最基层的操作视角）上反映数据的价值和重要性。

以下是一些原则示例：

- 建立数据的综合关键性。

  综合数据管理原则：质量、安全和保护

  关键数据的质量——患者和医生数据最明显——对 <组织> 的有效医疗成果和持续成功至关重要。我们的患者和员工的安全将在入院、诊断、治疗和后续护理的每个阶段受到保护。将实施持续的实践来确保 <组织> 所需的越来越高的质量水平。将对必要工作进行调整，以确保患者的数据受到保护，<组织> 保持最优秀的员工随时可提供服务。
  经过细微的措辞更改，可将这段话用于商业组织的客户数据或政府机构的居民数据。

- 解决数据治理委员会的必要性。

  <组织> 的监督和表示广度

  对数据治理委员会的成功至关重要的是〈组织〉目的和目标的公正、平衡和跨组织的表示。我们将竭尽全力确保在整个〈组织〉平衡目的、目标、风险、问题、机会和优先级。数据治理委员会成员将表示他们具体的组织，但始终有一个针对 <组织> 最关键的优先级的整体视图。

- 解决主数据领域的关键性。

  患者数据原则

  高质量患者数据对于实现 <组织> 的使命、目的和目标绝对不可或缺。这样，患者数据可视为高度共享的企业资产。我们将竭尽全力确保我们代表患者采集和管理的数据具有最高质量，

可在每个护理点使用，根据我们的患者和 <组织> 的利益受到了保护。

再一次，只需很少的措辞更改，相同的话语就可用于商业组织的客户数据或公共机构的居民数据。

策略

对必须完成何种操作来实现原则的目标进行可度量的陈述的策略声明。策略不规定它如何实现，这由规程和业务规则声明来指定。策略可能指内部或外部标准、制度、规则或指导原则（比如隐私指导原则），可能需要遵守 SOX 和 HIPAA 等法规。策略在一个原则内进行以下陈述：

- 必须为每种类型的数据实现何种质量水平？
- 谁拥有该责任？
- 如何处理异常和争议？

策略必须由数据照管人、数据治理委员会的成员或各种附属协会、专题小组、董事会、工作组、项目团队等的成员编写。它们需要进行重要协商，以确保范围和细节是正确的。

以下是不同类型的数据策略示例：

- 定义策略
  - *“居民是一个独立实体，他（在过去、现在和未来）与一个国家拥有直接或间接（继承）的居民关系。”*
- 内容、结构/模式和语义策略
  - *“每个选民记录，如果它是一个组织，必须拥有有效的办公地址和有效的寄送地址，另外，至少一个公司与一个个人拥有联系人关系。”*
  - *“每个用户记录在最低限度上必须拥有一个企业标识符，至少一个用户姓名，一个帐单地址，一个电话号码和一个电子邮件地址。”*

- *“邮政地址必须符合标准的 XYZ。”*

- 整合策略
  - *“每个交换居民数据的系统必须将它的本地模式转换为 INDIV.XML schema v3.5.1.2a。”*
- 安全策略
  - *“财产物料清单在一定的机密安全级别上进行处理。”*
- 责任和纠正策略
  - *“土地管理局负责与位置和所有附属实体的结构和内容相关的所有决策。”*

规程

规程，也称为“流程”和“实践”，是对如何实现策略的目标的声明。它们定义执行来实现策略规定（目标或需求）的任务。规程声明定义谁如何、何时和在何处执行策略。

以下是规程解决的一些主题：

- 谁做什么，他们何时必须这么做
- 谁参与了进来，具有何种能力（他们进接受通知、咨询还是是批准主体）
- 执行某些操作的合适的步骤序列
- 如何处理异常
- 如何沟通操作

业务规则

业务规则是对如何计算一部分数据来查看它是否满足策略的质量目标的具体规范。它们提供了如何对待策略中的具体数据异常的详细信息，描述了以下方面：

- 哪些数据元素是关键的
- 哪些值和关系是有效（或无效）的
- 如何确定一个值是否正确或允许
- 在一个值位于被认为可接受的范围外时如何操作

业务规则可调用参考数据来检查可允许的值，或者它们可调用探查引擎来检查一组数据的有效性。它们通常由数据照管人、编程人员/分析师和业务分析师编写。它们非常详细，远远超出了组成数据治理委员会的高级管理人员的视野。

以下是可能在主数据管理活动期间编写的业务规则示例：

- 唯一性规则：“一个<*实体*>*在特定上下文内由……唯一标识。*”
- 领域完整性规则：“此字段仅拥有<*这些*>*允许的*<*值/范围/格式/掩码*>。”
- 语义规则：含义、准确性、一致性、有效性、有用性。
- 流行性规则：“*此数据仅在*<*日期-时间戳*>*之前是良好的。*”
- 使用规则：谁请求了此数据。
- 参照完整性规则：必须包含所引用数据的键、依赖的数据必须在……之前/之后恰当地创建。
- 安全规则：允许谁出于何种用途查看/使用哪些记录和字段，谁可以与谁共享数据。
- 结构规则：“*对象必须遵守*<*模式定义*>”、“*用户将按此顺序获取*<*字段*>*……*”
- 集合规则：一次可使用多少个记录。
- 容量约束：“*用户一次仅能获得* <*n*> *条记录。*”
- 性能约束：“*用户将在* <*nnnn*> *秒内获得数据。*”

度量指标

本书正文中已介绍，度量指标是对应该度量哪些内容来确保成功实现一组业务目标的声明。它们是所采集的与转换相关的关键数据元素，提供了数据质量的洞察。度量指标也可用于度量在实现策略声明（支持原则）所规定的业务目标过程中的进度。

IBM 数据治理统一流程

使用 IBM 软件和最佳实践提升业务价值

数据治理已成为一种行业时髦用语，但它对不同的人具有不同的含义。数据治理的定义类似于盲人摸象的故事。取决于他们的背景和业务条件，从业者倾向于将数据治理与一个或多个元数据、业务术语库、主数据治理、分析治理、安全和隐私以及信息生命周期治理等同起来。所有这些定义都是正确的，但它们不完整。数据治理是一门将数据视为企业资产的学科。它涉及到运用决策权利来以企业资产的形式优化、保护和利用数据。

一些数据治理计划以成熟度评估开始和结束。如果您已发现您相对不成熟，接下来应该做什么？实现数据治理计划需要哪些步骤？本书介绍了一个适合数据治理从业者的统一流程。它以 IBM 产品、服务和数百次客户服务活动的最佳实践为基础，提供了对实现数据治理的 14 个步骤和将近 100 个子步骤的严密解释。

任何已在其组织内实现或希望实现数据治理计划的从业者都应阅读本书。其中包含来自不同行业、工作职能和地域的示例。而且无论您组织内的数据治理成熟度水平如何，本书都有适合您的内容。

Sunil Soares 是 IBM Software Group 内的一个数据治理主管。他与超过 200 位行业顾问的团队合作，帮助 IBM 的重要客户构建他们的数据治理计划。Sunil 已帮助多个行业的 100 多个客户（包括银行、保险、生命科学、制造、医疗、零售、电信和政府）评估他们的数据治理成熟度和确定促进其发展的合适流程和工具。

在出任目前的职位之前，Sunil 是 IBM Software Group 的 Worldwide Channels and Alliances for InfoSphere 主管，他在这里与众多合作伙伴就其数据治理实践进行合作。在加入 IBM 之前，他在纽约的 Booz Allen & Hamilton 的 Financial Services Strategy Consulting Practice 工作。Sunil 拥有芝加哥大学布斯商学院金融与营销专业的 MBA 学位。他生活在新泽西。

MC Press Online, LLC
P.O. Box 4886
Ketchum, ID 83340-4886

定价：$24.95 US/$27.95 CN

ISBN 978-1-58347-360-3
52495
9 781583 473603