如何启动数据质量计划

IBM

数据质量挑战

若要做出明智的业务决策，为客户提供一流支持，制定有效计划并高效管理其他基本任务，组织就需要依赖于高质量数据。但随着数据来源的不断增多以及数据量的指数级增长，使得组织很难维持高质量的数据。如此一来，在某些情况下就会闹出一些尴尬的情境，例如一位 50 岁的男士收到来自医疗保健供应商的信件，要求他确认即将在近期进行的妊娠超声检查预约。不过，如果数据质量低下，还会造成严重的影响。举例来说，英国一位年轻的妈妈因患有乳腺癌而去世，其背后的原因是因为癌症诊断与后续治疗出现了延误。为什么会出现这种延误？是因为医院的病患记录中出现了错误，将原本是"1b"的病房号记录成了"16"。如此一来，导致医院的诊断书从未送到病患手中，最终造成该名病患错过了关键的治疗。[1]

如果能够解决数据质量问题，组织便可避免诸如此类的不幸结果，并改善业务成效。除了人为错误外，大多数数据质量问题还在于在整个企业范围内缺乏数据存储与识别方面的统一信息标准。如果数据来源出现不一致，就会妨碍理解关键业务实体（如相关方、产品等）之间的关系。在许多情况下，组织缺乏可靠且长期有效的关键字来检索整个企业范围内与单个相关方或产品相关的所有信息。

高质量的数据能够使战略系统将所有相关数据集成一体，进而提供组织及组织范围内相互关系的完整视图。如果在整个企业范围内缺乏高质量的数据，组织就无法指望他们在关键业务应用（如数据仓库、业务智能工具、主数据系统）的投资获得预期的回报。通过实施数据质量计划，组织将能够提升数据完整性，进而充分利用信息资产。

数据质量解决方案涉及各种工具，包括：数据标准化软件、数据匹配引擎、元数据工作台、IT/业务术语控制台及数据集成软件等等。对于许多 IT 组织而言，一次性将所有这些工具集成到完整的数据质量解决方案之中，是个极具挑战性的过程。此外，组织可能还会面临着数据质量优先级方面的问题。某个部门可能只是希望从主客户文件中擦除地址，另一个部门要合并新收购公司的财务系统，而还有一个部门则亟待解决月销售报告的可靠性问题。

解决这些问题的关键在于灵活性。就组织的信息治理战略而言，可解决许多数据质量问题的通用数据集成平台是其中的关键。相比通过购买多个非集成式单点解决方案的方式在事后解决特定问题，投资购买这种平台更为明智。

在整个信息供应链中的高效信息治理

典型的组织可能会托管数百甚至数千个不同的系统。信息可能来自于许多位置（例如事务处理系统、运营系统、文档资料库、外部信息来源等），而且格式也非常多样化（例如数据、内容、流信息等）。各种来源、各种类型的数据之间往往存在着极具意义的关系。这种信息供应链会在整个企业内流动并延伸到企业边界之外（见图 1）。与传统供应链中的实体不同，信息供应链中存在多对多的关系。举例来说，与某个个人（例如客户、员工或合作伙伴）相关的相同数据，可以来自多个来源，而且这些信息最终会出现在许多报告和应用中。分散的不同系统也可能会以不同的方式定义信息。

有鉴于这种复杂性，对信息进行集成、确保信息质量，并正确解读信息是做出合理决策的关键步骤。因此，必须将信息转化为可信资产并进行管理，以便确保在其整个生命周期中的质量。基础性系统应具有成本高效性且易于维护，而且即便它们所处理的数据以惊人的速度增长，也应能轻松应对所分配的工作负载。

有效的信息治理有助于改善企业数据的质量、可用性和完整性，并促进跨组织协作和结构化决策。它能够平衡部门性孤岛与组织利益，有助于提高组织对数据的信心，进而实现直接业务影响，例如增加收入、减少成本、降低风险等。如果数据质量不佳，就会造成多种影响，包括业务流程失效、生产效率下降、材料浪费等。如果出现信息丢失、不准确或不完整，还会导致成本增加、需要其他额外工作（例如数据捕获、信息调解等）。

高质量数据的特性

- **完整性**：所有相关数据都关联到一起。举例来说，一条完整的客户记录可能包括所有账户、地址及公司与该客户之间的关系等。
- **准确性**：如果能够确保数据准确性，就能够消除拼写错误、录入错误、随机缩写等常见的数据问题。
- **可用性**：已经找到了所需数据，而且能够根据需求使用这些数据；用户无需手动搜索信息。
- **及时性**：如果销售报告不能反映最近一个月的情况，那么它还有何价值？

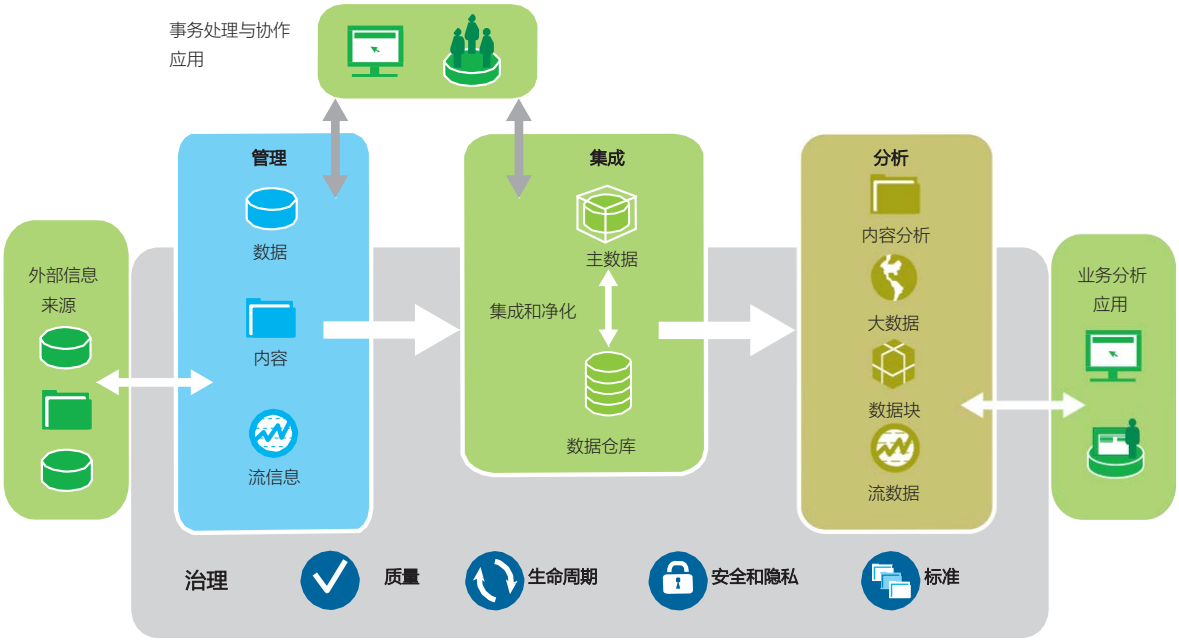

图 1：信息供应链。

一流的数据质量是成功的基础；举例来说，高质量的数据有助于我们清晰地了解客户、合作伙伴和供应商，因此说是实现业务发展还是在竞争中落败，数据质量有着非常重要的作用。

制定数据质量计划：入门

若要实现一流的信息治理，组织应根据业务目标和优先事项制定数据质量计划。并非所有的数据质量问题都会对业务成效带来相同的影响；如果尝试解决组织在数据质量方面的所有挑战，不仅任务艰巨，而且效率也将会非常低下。因此，您可以考虑以下问题，确定组织内每个潜在数据质量计划的投资回报率 (ROI)：

- 在依赖于信息的业务流程中，哪个流程是最关键的流程？
- 对此类流程而言，哪些信息最为重要？
- 如果信息质量低下，将会对此类流程的效率造成何种影响？
- 若要针对这些流程维持高质量信息，需要付出哪些成本？
- 如果针对这些流程维持高质量信息，组织将能获得哪些纯收益？

无论选择哪种方式，数据质量方面的需求都将会随着时间的流逝而发生变化；因此，重要的是要投资能够扩展且可在整个企业内充分得到利用的技术。旨在解决当前问题的单点解决方案（例如旨在改善数据准确度和统一性的地址清理解决方案）可能无法支持未来的数据质量需求。

除了要根据未来需求做好规划之外，数据质量计划还应解决两个基础性问题。首先，组织必须在数据质量的定义上达成一致。什么是"高质量"的数据？数据的出错率是否低于 1%？或者说，您的组织是否能承受 10% 的出错率？

以某家政府机构为例，他们需要确保在广泛的检查点维持高度准确的信息。一旦出现数据错误，将会造成灾难性后果。不过，如果是某个服装零售商的地址数据库，数据错误的影响就会相对小很多。无论何种情况，都必须要了解目标（这是合理分配资源并管理计划成本的关键），而且不要做出完美假设。

下一个问题与报告有关：如果确定了数据质量的定义，必须要跟踪哪些指标来确保维持质量阈值？数据质量计划应采用系统性的、业务驱动的方法来捕获并报告这些指标，而且组织必须对优先事项进行正式归档，并随着时间的流逝予以跟踪。

数据质量计划的最初重点取决于组织在这些问题上的答案。

数据质量的第一切入点：数据质量评估

无论是启动新计划，还是解决数据治理与风险减缓问题，许多组织都发现数据质量评估是个非常好的切入点；通过数据质量评估，能够确立一个基准：数据的质量如何？最佳的改善机会有哪些？

许多组织发现他们很难就整个企业内部的数据达成统一共识，尤其是当根据不断变化的需求调整系统和应用时，以及由于并立、收购等活动导致现有数据集扩大时，更是如此。业务部门和 IT 部门通过多种方式使用不同的语义和应用记录数据，比如通过不同的识别符（例如客户和账户 ID 等）、不同的格式（例如数据库中采用日期的格式存储日期数据，而在文件中采用字符串的格式存储日期数据），或者不同的值（例如在某个系统中使用"M"或"F"来表示性别，而在另一个系统中则使用"0"或"1"来表示）。此外，组织的知识管理方法会进一步加剧这种情况。

通常来说，有关数据的知识属于隐形内容，会根据人们的具体工作内容存储在人们的脑海里。或者，这些知识也可能会记录在文档中，但这些文档并不会根据最新的业务流程或系统变更进行适当更新。当个人离开组织时，很多知识就会丢失或变得支离破碎。

数据质量评估旨在针对这种复杂的情况提供相关洞察力，为后续工作确定基础性实践，并在共享式的元数据库中建立一个知识库，以便供多个项目和计划使用和复用。数据质量评估专注于业务问题，而且会根据基础数据对业务问题进行阐释。这种方式有助于凸显数据质量问题，但数据分析师或业务分析师仍旧必须审核结果，并得出结论，尤其是必须明确数据质量问题的业务影响。

如图 2 所示，数据分析师是数据质量评估的核心。分析师必须从整体上把握，包括评估的范围、目标及可交付件。如果分析师不了解组织的目标，异常或问题的识别也就无从谈起。需要分析的表格和属性数量非常巨大。简单地选择多个数据列然后进行分析，这并非分析流程的终点 - 后续的审核非常关键。

IBM® InfoSphere® Information Server 提供了诸多一流的功能，可帮助分析师学习并运用数据分析技巧。IBM 就这些功能提供的培训服务着重于下述核心分析步骤和最佳实践：

- 识别并找出有问题的数据源
- 利用自动化的数据内容驱动型功能
- 通过数据分类让分析更具针对性
- 验证数据格式和领域
- 报告并交付分析结果
- 随时间持续维持分析结果

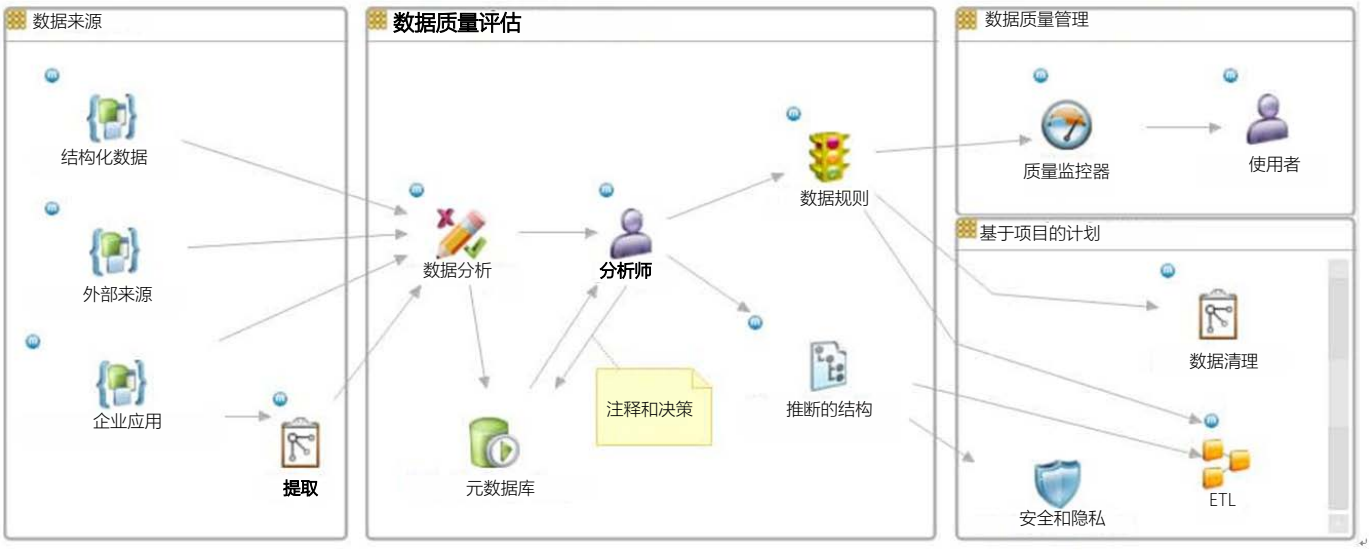

图 2：数据质量评估的格局

第二步：构建综合性的数据质量评估平台

InfoSphere Information Server 提供了数据分析和数据质量监控功能，用以创建数据质量评估。该产品通过多层架构、通用服务、共享式元数据与平行处理引擎构建了一个通用平台，该平台能够在一个安全的、基于项目的环境（见图 3）中分析广泛的数据源，处理海量数据，存储广泛的结果，并捕获分析师洞察力。

通过 InfoSphere Information Server 来执行核心分析，分析师便可在许多领域发现各种问题，例如数据违反规则、值缺失、键并不唯一或存在重复等等。分析师可以使用其他技术来专注于特定类型的条件，这些条件通常会表述为数据验证规则。这些规则可用于测试有效值组合、共识和聚合是否正确，或复杂的格式要求，同时也可用于获得整个记录或表格的综合性评估结果。此外，随着时间的流逝，还可在 InfoSphere Information Server 中报告、维持和演变其他测试。

通过在 InfoSphere Information Server 中进行核心分析所得出的共享式元数据，可直接供平台中其他功能的用户使用。数据建模人员及数据库管理人员可以使用推断出的结构及识别出的类别，建立具有正确结构的加载区域，或改进隐私和治理策略。专注于数据转换或数据清理的开发人员可以利用统计与注释功能来确保将适当的清理例行程序运用到数据之中，而且可以将通过分析生成的参考表格融入其中。之后，开发人员可以通过一个有助于快速运用规则的集成式加载区域，将分析时所用的数据验证规则直接运用到数据清理及提取、转换和加载（ETL）流程中。这种功能有助于确保：存在问题的数据条件或无效的数据条件在被加载到目标环境之前得到妥善解决。

借助 InfoSphere Information Server，企业可以首先专注于某个源系统，然后按照计划在整个企业中持续拓展基本数据质量评估，包括数据清理、信息集成与数据治理项目。企业可以将分析流程、数据验证规则和报告规划为定期执行的活动，以便实现持续的数据质量监控。InfoSphere Information Server 能够提供有关企业数据领域的洞察力，帮助企业解决在基础性数据、系统和应用的持续扩展与获得过程中所固有的挑战。

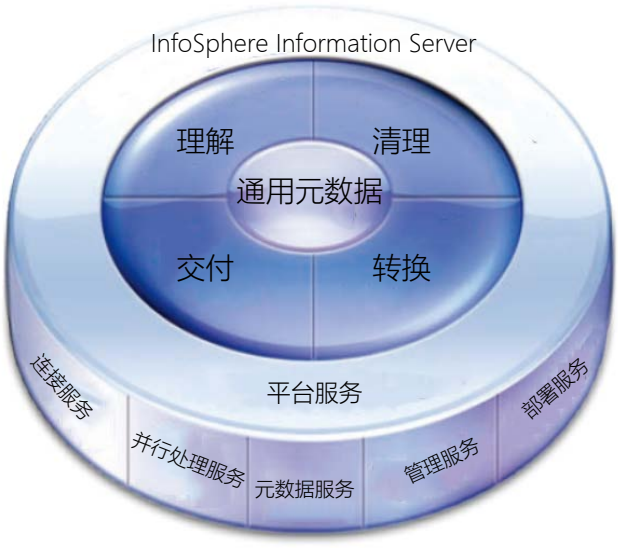

图 3：InfoSphere Information Server 基于共享的元数据、并行处理及其他服务而构建。

InfoSphere Information Server 支持的其他数据质量切入点

除了数据质量评估、信息概要与分析之外，InfoSphere Information Server 还支持完整数据质量计划的其他切入点，具体取决于组织的优先事项。

定义通用业务语言

在业务用户与 IT 用户尝试进行协作，以实现有效的信息集成时，会由于在数据理解与解释、数据重要性的确定、数据管理方面的难点而受到阻碍。在整个企业环境中，之所以会出现业务定义的不一致问题，通常是因为缺乏企业级的数据字典和管理计划。

InfoSphere Information Server 中的业务术语功能可帮助组织创建、管理和共享在整个企业范围内控制的业务术语。在业务部门与 IT 部门之间创建这种通用语言是实现技术目标与业务目标两者之间一致性的关键。除了可控制的业务术语之外，层级和分类系统还可提供额外的业务情境。

理解数据及数据关系

在实施信息治理计划或以信息为中心的项目之前，组织必须获得其数据的完整视图：所拥有的数据类型、数据所在位置、不同系统中数据之间的关联等。大多数组织的数据发现流程是手动流程，需要投入人力耗时数月才能发现业务对象、敏感数据、不同来源的数据之间的关系、转换逻辑。结果是导致整个流程非常耗时，而且很容易出现错误，不仅拖慢了价值实现速度，还可能会导致新系统中出现数据准确性问题，甚至会导致出现新系统永远无法投入运营的情况。

InfoSphere Information Server 提供了一系列功能，可帮助组织实现数据发现流程的自动化。它可以提供单来源数据概要、跨来源数据重叠分析、匹配键发现、原型设计及测试功能，进而实现数据合并、自动化转换发现。此外，它还使用启发式的高级算法实现自动化分析，相比于手动分析流程，可帮助组织节省大量的时间和成本投入。

清理、标准化并匹配信息

若要确保地址清理、记录去重等任务中的质量与一致性，组织需要部署相应的工具，以实现可靠且易于使用的标准化与匹配，以及数据集成功能，尤其是涉及到多个源和/或多个目标时，更是如此。InfoSphere Information Server 可帮助企业创建并维持准确的主数据实体视图，诸如客户、供应商、位置和产品。它还可提供一个包含有一系列下述强大、灵活的功能的开发环境：

- 为核心业务实体提供单个标准化、清理、匹配与监管规则集，这些规则可以批量执行、实时执行或作为一个 Web 服务予以执行
- 使用专门设计的概率算法对数据进行匹配，该算法旨在确保企业运行所需信息的准确性、完整性和可靠性
- 在广泛可扩展的并行平台上处理全局数据，以在要求苛刻的环境中实现最优性能
- 有助于创建并维护高质量的主数据，以便各种关键企业计划能够实现预期收益，包括主数据管理和数据治理计划
- 通过无缝的数据流集成将数据质量功能融入到数据集成情境之中
- 使用直观的"个性化"用户界面

维持数据沿袭

InfoSphere Information Server 旨在集成并扩充各个分散源系统中的信息。该解决方案采用了一个活动共享式元数据库层，可通过协作与复用原则为一系列集成活动和用户角色提供支持。这些工件包括：有关各种信息来源的技术元数据、用于描述业务涵义及信息使用情况的业务元数据，以及用于描述集成流程中所发生事件的运营元数据。

InfoSphere Information Server 平台提供了一个强大的元数据管理界面，该界面不仅支持 InfoSphere Information Server 元数据，还支持在数据集成流程中扮演着关键作用的其他元数据。该平台可在数据集成流程的整个格局中交付集中式的全局视图，同时确保对在 InfoSphere Information Server 内部和外部进行的数据转换的可视性。借助这种可视性，组织能够追溯信息来源，确保所收到信息的可靠性和可信赖性，这些对于审计或法律发现等情境尤为关键。

灵活、可扩展的数据质量解决方案

业务决策日益受到客户、合作伙伴及运营信息的影响。若要实现一流的成效，组织就必须确保决策所依赖数据的质量。通过综合性的数据质量计划，预先确定明晰的业务优先事项，有助于组织专注于投资规划。IBM InfoSphere Information Server 可提供一流的灵活性，帮助组织应对当今高优先级的数据质量问题，同时还可实现轻松扩展，以支持未来需求。对于刚刚开始实施数据质量与信息治理计划的组织而言，该产品可为其提供综合性通用数据集成平台所具有的完整灵活性。

有关更多信息

如欲了解有关数据质量及其在信息治理战略中重要作用的更多信息，请联系您的 IBM 销售代表或 IBM 业务合作伙伴，或访问以下网站：

- ibm.com/software/data/integration/capabilities/cleanse.html
- ibm.com/software/data/db2imstools/solutions/data-governance.html

IBM Corporation
Software Group
Route 100
Somers, NY 10589

美国印刷
2012 年 3 月

[1] “Mother with young son dies of cancer at 38 after hospital typing error sent urgent letters to the wrong address”. The Daily Mail. 2011 年 3 月 14 日.www.dailymail.co.uk/news/article-1366056/Mistyped-address-leaves- mother-dead-cancer-son-8-orphan.html