数据治理的本质及实践

> 近三年，随着阿里数据中台战略的提出，以及各种数据应用场景的成功落地，企业和政府对自身数据资产的价值也前所未有的重视起来。但是，数据资产的价值发掘依赖于有序、完整和高质量的数据，数据治理则是保障数据质量和实现数据价值的基础，它包含一整套构建核心数据资产的方法论、规章制度和实施工具。
>
> 本篇文章就结合龙石数据的理论研究和实践经验，从以下方面展开描述，帮助大家揭开数据治理的面纱。

目 录

（1）什么是数据治理?

我们认为，数据治理是指从使用零散数据变为使用统一数据、从具有很少或没有组织流程到企业范围内的综合数据管控、从数据混乱状况到数据井井有条的一个过程。

所以，数据治理强调的是一个过程，是一个从混乱到有序的过程。从范围来讲，数据治理涵盖了从前端业务系统、后端业务数据库再到业务终端的数据分析，从源头到终端再回到源头，形成的一个闭环负反馈系统。从目的来讲，数据治理就是要对数据的获取、处理和使用进行监督管理。

具体一点来讲，数据治理就是以服务组织战略目标为基本原则，通过组织成员的协同努力，流程制度的制定，以及数据资产的梳理、采集清洗、结构化存储、可视化管理和多维度分析，实现数据资产价值获取、业务模式创新和经营风险控制的过程。

所以，数据治理是一个过程，是逐步实现数据价值的过程，也正是因为这个过程特性，我们认为，**数据治理是一个持续性的服务，而不是一个有着明确范围的一锤子买卖。**

（2）为什么要实施数据治理?

当前，企业变革已经成为企业适应剧烈变化的市场环境、实现长期发展的必经之路。然而，**过去为组织带来工作效率提升的烟囱式的孤岛式的业务系统已经成为组织变革重组的阻力**，这也是从数据层面打通各个组织单元、实现业务单元快速重组的最根本的需求来源。

并且，在互联网的冲击下，各行各业都在寻求业务模式的创新，实现业务自动化向业务持续优化的转变，以求在竞争中找到一片蓝海。而组织要能实现业务模式的创新，第一步就是对自身的洞察，以及提升组织的运行效率，方能在互联网背景中立于不败之地。

所以，以下就是企业实施数据治理的根本原因：

1、经过 30 年的信息化建设，企业和政府部门都围绕着业务需求建设了众多的业务系统，从而导致数据的种类和数量大增，看似积累了众多的数据资产，实则在需要使用时，困难重重。

2、因为各个业务系统的建设都是围绕着业务需求来建设的，当业务环境发生变化时，原来的业务系统不能互联互通，不能满足跨部门、跨职能、跨组织的协作需求。

3、各个业务系统所产生的海量数据以复杂而分散的形式存储，导致数据之间的不一致和冲突等质量问题，从而导致数据在应用过程中的无所适从，难以实现数据的深度利用，从而难以实现业务模式创新和经营风险控制。

（3）数据治理的目标是什么？

数据治理本身不是目的，它只是实现组织战略目标的一个手段而已，例如基于需求的智能制造、智慧市场监督管理、融合市民服务、一网通办等。

从组织职能和体量大小方面来看，不同类型组织的数据治理目标大不相同，例如：

- 集团企业总部和政府大数据管理局的目标是：制定数据政策、保障数据安全、促进数据在组织内无障碍共享，其重点目标是推进和保障数据战略的顺利实施。
- 企业和政府业务部门的目标是：通过提升信息管理能力，提升组织精细化管理水平，提高业务运营效率，增强组织决策能力和核心竞争力，从而为实现组织战略目标提供能力支撑，其重点目标是数据价值获取、业务模式创新和经营风险控制。

（4）当前数据治理存在哪些问题和困难？

数据治理不只是技术问题，更是一个管理问题。例如大家常见的项目管理系统只是一个工具，如何让项目管理工具与项目管理思想相匹配才是项目管理系统实施过程中的最大挑战，也才能发挥最大的效果。数据治理也是同样的道理。

组织信息化建设正从以应用为中心向以数据为中心转变的关键时期，组织也逐步认识到数据的巨大价值，但低质量的数据和复杂的数据应用手段，让数据价值发掘的效果大大降低，甚至，会让组织决策层丧失数字化转型的信心。

那么，如果在项目实施的初期能识别出影响项目实施效果的困难，并找到相应解决办法，就显得异常重要。以下是龙石数据在工作中总结的最常见的数据治理问题：

1、跨组织的沟通协调问题。数据治理是一个组织的全局性项目，需要 IT 部门与业务部门的倾力合作和支持，需要各个部门站在组织战略目标和组织长远发展的视角来看待数据治理。因此，数据治理项目需要得到组织高层的支持，在条件允许的情况下，成立以组织高层牵头的虚拟项目小组，会让数据治理项目事半功倍。

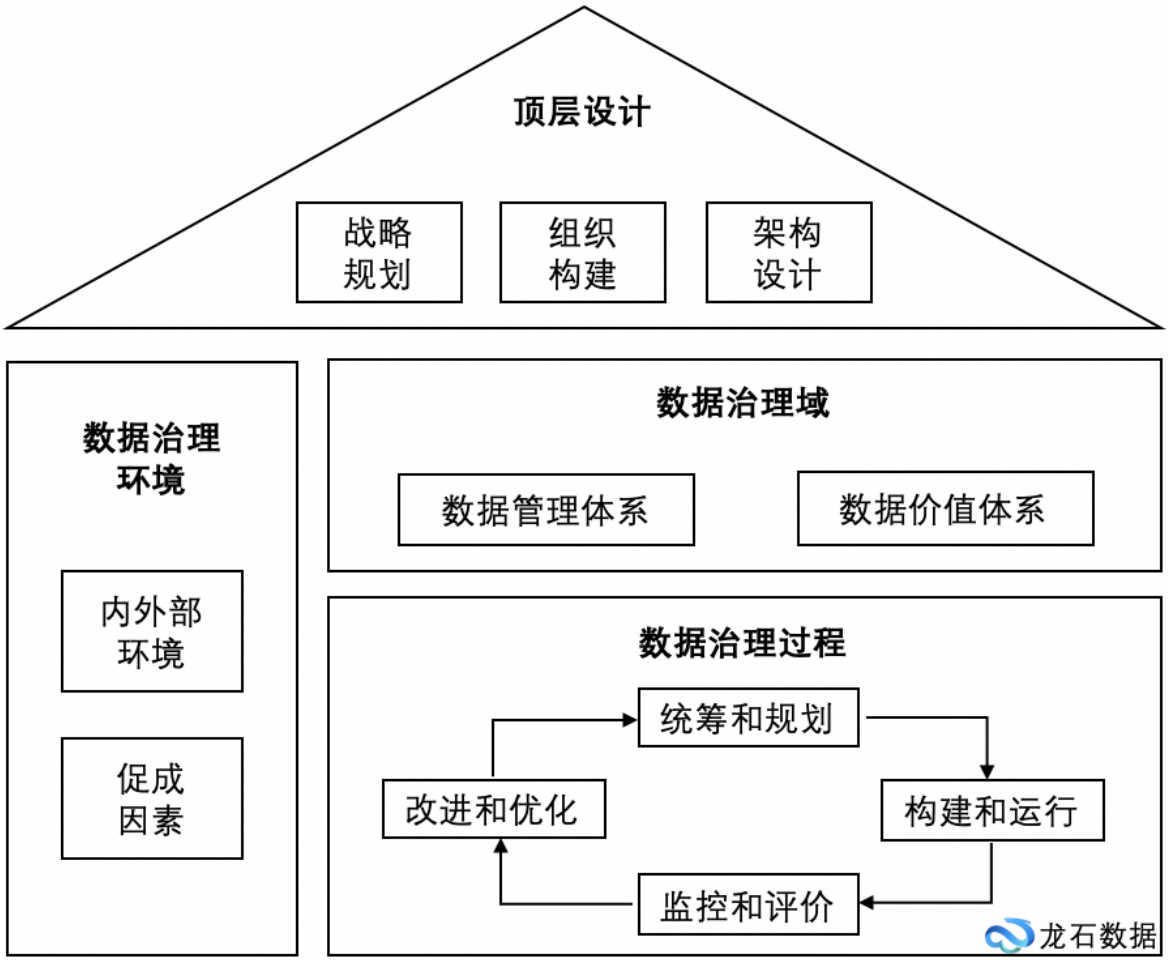
2、投资决策的困难。组织的投资决策以能够产生可预期的建设成效为前提，但往往综合性的数据治理的成效并不能立马体现，它更像一个基础设施，是以支撑组织战略和长期发展为目标，所以，导致此类项目无法界定明确的边界和目标，从而难以作出明确的投资决策。面对此类情况，我们的经验是采用“大平台 + 小目标”的实施方案。“大平台”指的是数据治理的支撑平台。“小目标”指的是利用基础支撑平台和一小部分数据实现某一个具体业务目标。“大平台 + 小目标”方案的优势是能够快速实现可评估的工作成效，难点就在于基础支撑平台要能够对未来的综合治理提供足够的支撑能力，避免重头再来。以智慧市场监管为例，全部的数据包含企业法人监管、质量监督检查、食品监管、药品监管、特种设备监管、主题检查和执法等等，实施综合性的数据治理项目，则可以对企业法人实现全维度的分析和预警，而“大平台 + 小目标”的实施方案则可以实现诸如食品药品安全监管这些主题性的建设目标。

3、工作的持续推进。数据治理是以支撑组织战略和长远发展为目标，应当不断吸收新的数据来源，持续追踪数据问题并不断改进，所以数据治理工作不应当是一锤子买卖，应当建立长效的数据改进机制，并在有条件的情况下，尽量自建数据治理团队。

4、技术选型。前几年，随着大数据的发展，各种名词层出不穷，令人眼花缭乱，例如：数据仓库、ETL、元数据、主数据、血缘追踪、资源目录、结构化非结构化、Hadoop、Spark、联机事务处理（OLTP）、联机分析处理（OLAP）、商业智能（BI），等等。这里面有针对传统数据库的，有针对大数据数据库的，再加上组织对自身数据资产情况没有一个清晰的认识，这也就导致了数据治理的技术选型困难。而当下，基于传统关系型数据库仍然符合绝大多数据企业的业务需求，为避免误解，以下内容主要针对的是传统关系型数据库数据治理的介绍。

（5）数据治理包含哪些内容?

从我们龙石数据的实践经验来看，相对于国际组织和国际企业发布的数据治理框架，以下国家标准 GB/T 34960 发布的数据治理框架比较符合我国企业和政府的组织现状，更加全面地和精炼地描述了数据治理的工作内容，包含顶层设计、数据治理环境、数据治理域和数据治理过程。



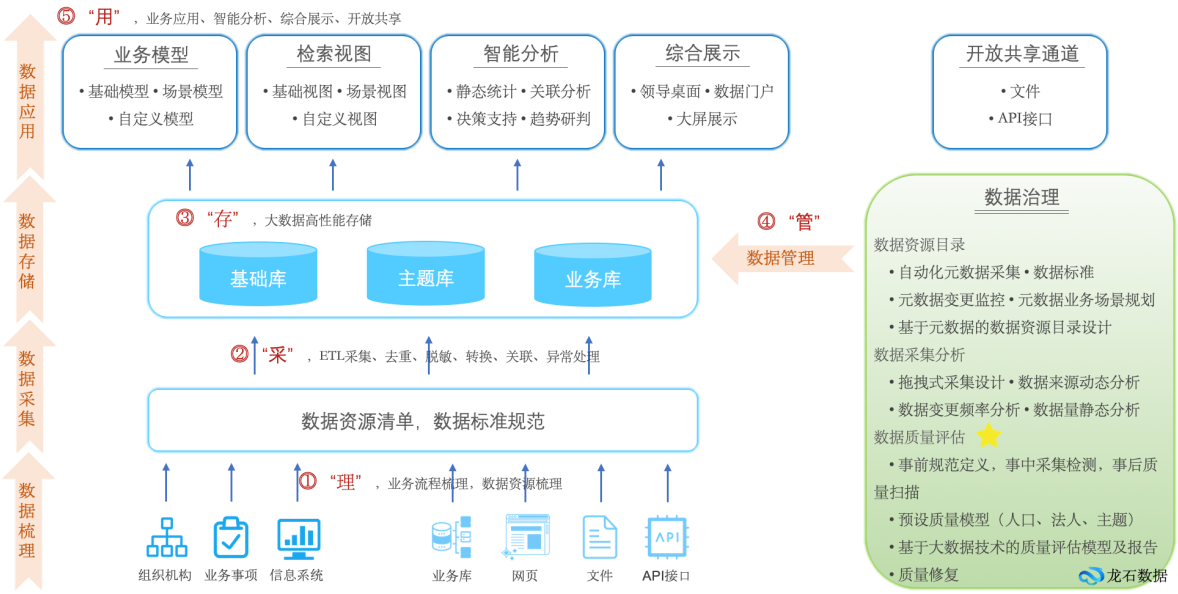
图：国标 GB/T 34960 的数据治理框架

1. 顶层设计是数据治理实施的基础，是根据据组织当前的业务现状、信息化现状和数据现状，设定组织机构的职权利，并定义符合组织战略目标的数据治理目标和可行的行动路径。
2. 数据治理环境是数据治理成功实施的保障，指的是分析领导层、管理层、执行层等等利益相关方的需求，识别项目支持力量和阻力，制定相关制度以确保项目的顺利推进。
3. 数据治理域是数据治理的相关管理制度，是指制定数据质量、数据安全、数据管理体系等相关标准制度，并基于数据价值目标构建数据共享体系、数据服务体系和数据分析体系。
4. 数据治理过程就是一个 PDCA（plan-do-check-act）的过程，是数据治理的实际落地过程，包含确定数据治理目标，制定数据治理计划，执行业务梳理、设计数据架构、数据采集清洗、存储核心数据、实施元数据管理和血缘追踪，并检查治理结果与治理目标的匹配程度。

GB/T 34960.5-2018 的详细信息请参考： http://www.gb688.cn/bzgk/gb/newGbInfo?hcno=F3B2108863A2292F5AF0FA645CEE047F

（6）数据治理需要哪些工具?

从技术实施角度看，数据治理包含“理”“采”“存”“管”“用”这五个步骤，即业务和数据资源梳理、数据采集清洗、数据库设计和存储、数据管理、数据使用。



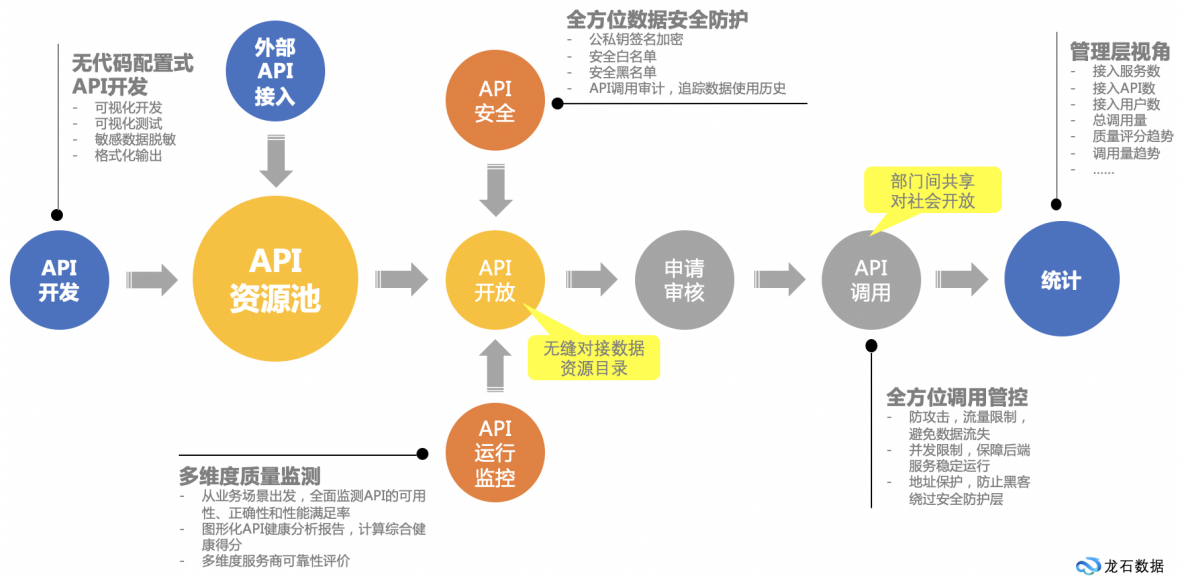
- **数据资源梳理**：数据治理的第一个步骤是从业务的视角厘清组织的数据资源环境和数据资源清单，包含组织机构、业务事项、信息系统，以及以数据库、网页、文件和 API 接口形式存在的数据项资源，本步骤的输出物为分门别类的数据资源清单。
- **数据采集清洗**：通过可视化的 ETL 工具（例如阿里的 DataX，Pentaho Data Integration）将数据从来源端经过抽取 (extract)、转换 (transform)、加载 (load) 至目的端的过程，目的是将散落和零乱的数据集中存储起来。
- **基础库主题库建设**：一般情况下，可以将数据分为基础数据、业务主题数据和分析数据。基础数据一般指的是核心实体数据，或称主数据，例如智慧城市中的人口、法人、地理信息、信用、电子证照等数据。主题数据一般指的是某个业务主题数据，例如市场监督管理局的食品监管、质量监督检查、企业综合监管等数据。而分析数据指的是基于业务主题数据综合分析而得的分析结果数据，例如市场监督管理局的企业综合评价、产业区域分布、高危企业分布等。那么基础库和主题库的建设就是在对业务理解的基础上，基于易存储、易管理、易使用的原则抽像数据存储结构，说白了，就是基于一定的原则设计数据库表结构，然后再根据数据资源清单设计数据采集清洗流程，将整洁干净的数据存储到数据库或数据仓库中。
- **元数据管理**：元数据管理是对基础库和主题库中的数据项属性的管理，同时，将数据项的业务含义与数据项进行了关联，便于业务人员也能够理解数据库中的数据字段含义，并且，元数据是后面提到的自动化数据共享、数据交换和商业智能（BI）的基础。需要注意的是，元数据管理一般是对基础库和主题库中（即核心数据资产）的数据项属性的管理，而数据资源清单是对各类数据来源的数据项的管理。
- **血缘追踪**：数据被业务场景使用时，发现数据错误，数据治理团队需要快速定位数据来源，修复数据错误。那么数据治理团队需要知道业务团队的数据来自于哪个核心库，核心库的数据又来自于哪个数据源头。我们的实践是在元数据和数据资源清单之间建立关联关系，且业务团队使用的数据项由元数据组合配置而来，这样，就建立了数据使用场景与数据源头之间的血缘关系。

- **数据资源目录**：数据资源目录一般应用于数据共享的场景，例如政府部门之间的数据共享，数据资源目录是基于业务场景和行业规范而创建，同时依托于元数据和基础库主题而实现自动化的数据申请和使用。

- **质量管理**：数据价值的成功发掘必须依托于高质量的数据，唯有准确、完整、一致的数据才有使用价值。因此，需要从多维度来分析数据的质量，例如：偏移量、非空检查、值域检查、规范性检查、重复性检查、关联关系检查、离群值检查、波动检查等等。需要注意的是，优秀的数据质量模

型的设计必须依赖于对业务的深刻理解，在技术上也推荐使用大数据相关技术来保障检测性能和降低对业务系统的性能影响，例如 Hadoop，MapReduce，HBase 等。

- **商业智能（BI）**：数据治理的目的是使用，对于一个大型的数据仓库来说，数据使用的场景和需求是多变的，那么可以使用 BI 类的产品快速获取需要的数据，并分析形成报表，比较知名的产品有 Microsoft Power BI，QlikView，Tableau，帆软等。
- **数据共享交换**：数据共享包括组织内部和组织之间的数据共享，共享方式也分为库表、文件和 API 接口三种共享方式，库表共享比较直接粗暴，文件共享方式通过 ETL 工具做一个反向的数据交换也就可以实现。我们比较推荐的是 API 接口共享方式，在这种方式下，能够让中心数据仓库保留数据所有权，把数据使用权通过 API 接口的形式进行了转移。API 接口共享可以使用 API 网关实现，常见的功能是自动化的接口生成、申请审核、限流、限并发、多用户隔离、调用统计、调用审计、黑白名单、调用监控、质量监控等等。



作者介绍

苏槐，微信号 Sulaohuai，现服务于龙石数据，曾就职于神州数码、Oracle、新加坡电信等企业。擅长容器技术、微服务架构、数据治理及技术管理。