数据质量人人有责，这不仅仅只是一句口号，更是数据工作者的生命线。数据质量的好坏直接决定着数据价值高低。

数据质量管理是指在数据创建、加工、使用和迁移等过程中，通过开展数据质量定义、过程控制、监测、问题分析和整改、评估与考核等一系列管理活动，提高数据质量以满足业务要求。

可按照"谁创建、谁负责；谁加工、谁负责；谁提供、谁负责"的原则界定数据质量管理责任，由数据流转环节的各责任方对管辖范围内的数据质量负责。对数据质量规则优先采取系统程序的自动化控制措施，并尽可能前移管控点，从源头上控制数据质量。

01 数据治理问题场景

在日常工作中，业务领导经常会通过报表看板等数据产品来了解各项业务的发展趋势以及KPI的达成情况。倘若某天，他打开某张核心报表，发现当日的数据一直是空白的，询问报表开发人员，开发经排查分析，发现是依赖的上游有延迟，上游数据预计要下午才能到达，导致业务领导在正常时点无法查看业务数据情况。

又或某天，业务人员点开报表发现当日AUM规模暴增，数据增长当然开心，但仔细推敲，发现这波动有点不合常理，于是通知数据负责人验证下数据是否存在异常。经过几个小时的排查分析，数据负责人报告说数据确实算的有些问题，业务业务以后对该报表数据的准确性将会打上问号。

若类似的数据问题经常出现，估计迎接你的不是美好明天，而是你的职业生涯的最后一天。

02 数据质量的重要性

数据质量为什么至关重要？因为拥有高质量的数据可以让您更好地了解任何情况，从而更精准地执行任何事情。反之亦然。

伴随着技术的进步，组织或者企业能够收集大量的数据，用好这些数据已成为焦点。然而，由于以下原因，很多组织或者企业并没有实施数据质量计划：

- 没有业务部门负责数据质量问题

- 数据质量需要跨职能合作
- 它要求组织认识到数据质量是一个重要问题
- 它需要数据质量准则
- 它需要投入财力和人力资源
- 它被认为是非常人力密集的
- 投资回报往往难以量化

看起来，挑战大于好处。

但是，数据质量务必重视，原因有三。

原因一：成本

数据质量差，是IT项目失败的主要原因，也是客户流逝背后的驱动因素之一。

原因二：合规

质量差的数据会带来重大的法律或者声誉风险。一些例子如下：

- 数据缺失导致信用风险不准确
- 信用记录不完整致使风险评估错误
- 监管违规

原因三：决策

质量好的数据意味着有准确及时的信息来管理从研发到销售的产品和服务。质量差的数据导致错误的洞察力，从而做出错误的决策。决策的错误，公司会付出沉重的代价。

在企业，数据服务的方式有报表看板、标签指标和数据接口等，而这些数据服务要想为业务带来价值的，其首要前提就是要保证数据的准确性，输出高质量的数据。

低质量的数据会误导业务做出错误的决定，致使行动方向发生偏离。特别是在数据驱动的组织中，是否有准确的、可用的高质量的数据，将直接影响领导层能否做出正确的决策和战略目标的实现。

因此需要特别重视数据的质量问题，针对数据质量进行专项治理。

03 数据质量常见问题

在前面的场景案例中，我们可以发现有如下几个数据质量问题：

1. 数据延迟，导致业务无法在正常时效内获得数据结果。
2. 数据错误，导致数据结果完全不可信，以致无法使用。
3. 数据恢复慢，问题发生后，排查分析耗时长，数据恢复时间慢。

发现滞后，数据开发晚于业务人员发现数据异常，导致影响已传导到数据应用端。

04 数据质量问题原因

那应该如何解决这些质量问题，保证数据的高质量交付呢？

首先，我们需要了解这些质量问题产生的根本原因，了解问题才能更好地去解决问题。通过对历次数据质量问题进行复盘、总结，发现质量问题主要由下面几类原因引发：

- **数据平台问题：**平台不稳定、队列资源不足等，导致作业运行延迟、报错。
- **数据开发问题：**数据开发人员的任务脚本性能太差，计算严重耗时，导致数据延迟；或是代码逻辑设计有问题，导致数据计算有误。
- **上游系统异常：**上游源系统异常，数据文件晚到，导致下游依赖作业延迟。

05 数据质量治理

出现问题不可怕，可怕的是出现问题后，我们毫无感知，不能做到“早发现、早处理、早恢复”，以致问题直接传导到业务方，影响业务的开展工作。

在大数据产品矩阵中，我们使用数据质量监控平台来支持数据质量的监控、治理。

数据质量监控平台，主要是对Hive数仓中的库表数据的质量进行监控，包括表级别和字段级别的数据进行监控，以减少或避免由数据质量引起的事故和损失。

借助数据质量监控平台，我们通过实施下面几个关键步骤来进行数据质量的治理：

（1）配置监控规则

对高价值分的作业，我们要求强制配置基础监控规则，如：主键唯一性校验、数据非空校验；

另还可以根据业务场景需要，配置对应的业务规则监控，如：字段总值环比校验、字段极值校验等，监控平台内置了约17种字段级校验规则、5种表级校验规则，供直接配置使用；

监控规则	规则描述
表非空校验	表没有记录时发送告警
表环比校验	如果环比值超出预设规则上下限值，则会发送告警
表原表对比	如果环比值超出预设规则上下限值，则会发送告警
表主键唯一	如果表根据主键值去重计数不等于表记录数，则发送告警
字段字符非空校验	如果字段出现空值，则会发送告警
字段字符长度校验	如果字段数值长度超过预定义上下限范围，则会发送告警
字段字符只允许数字校验	如果出现非数字的字符或符号，则会发送告警
字段字符非法校验	如果出现预定义禁止值，则会发送告警
字段字符空值增长率	如果空值增长率超出预定义上下值，则会发送告警
字段字符格式校验	如果字段值的格式非指定格式，则会发送告警
字段枚举允许值校验	如果字段值出现非预定义允许值，则会发送告警
字段枚举空值增长率	如果空值增长率超出预定义上下值，则会发送告警
字段枚举非空校验	如果字段出现空值，则会发送告警
字段数字非空校验	如果字段出现空值，则会发送告警
字段数字空值增长率	如果空值增长率超出预定义上下值，则会发送告警
字段数字极值校验	如果字段数值超过预定义上下限范围，则会发送告警
字段数字非法值校验	如果出现预定义禁止值，则会发送告警
字段数字环比校验	如果环比值超出预设规则上下限值，则会发送告警
字段数字均值校验	如果环比值超出预设规则上下限值，则会发送告警
字段枚举环比	如果环比值超出预设规则上下限值，则会发送告警
字段空值占比校验	如果空值的记录条数和总记录数对比，则会发送告警

除内置了丰富的校验规则，质量监控平台还支持SQL自定义监控规则，极大地满足各种数据监控场景。

（2）监控告警

当校验规则识别异常时，需要通知负责人进行跟进处理，质量监控平台支持以电话、邮件和短信等方式通知作业属主。作业属主收到告警后，需及时地处理和关闭告警，否则告警将一直挂在那，在后面的告警响应度中会被稽核到，上报其领导。

（3）全链路数据监控

根据作业的价值分级，针对高价值作业，开发人员可根据数据血缘，对上游作业依次配上监控，实现全链路的数据质量监控。

06 数据质量评价体系

在执行了一系列的举措来提高数据质量后，如何来验证数据质量的治理效果呢？

根据企业本身的数据特点，设计并构建了一个数据质量七维评价模型，如下图所示：

数据质量评价模型，分别从数据完整性、监控覆盖率、告警响应度、作业准确性、作业稳定性、作业时效性、作业性能分等七个维度来考量平台的数据质量，基于该模型，还设计了“数据质量分”这个指标，来直观地反映平台数据质量的建设水平及健康状况。

数据质量七维模型的评价视角及其计算口径：

1

数据完整性

☆ 考量数据项信息是否全面、完整、无缺失

★ 指标公式：表完整性和字段完整性的平均值

2

监控覆盖率

☆ 确保数据遵循统一的数据标准或规范要求

★ 指标公式：监控的高价值作业个数/高价值作业总个数

其中，高价值作业是指作业价值分在80分以上的作业

3

告警响应度

☆ 通过日常管理、应急响应，降低或消除问题影响，避免数据损毁、丢失

★ 指标公式：已处理告警个数(本周)/告警总个数(本周)

4

作业准确性

☆ 考量数据是否符合预设的质量要求，如唯一性约束、记录量校验等

★ 指标公式：1 - 告警作业个数(本周)/监控作业总个数

5

作业稳定性

☆ 考量作业的运行稳定性，是否经常报错，导致数据事故

★ 指标公式：1 - 错误作业个数(本周)/作业总个数

6

作业时效性

☆ 考量数据项信息可被获取和使用的时间是否满足预期要求

★ 指标公式：1 - 延迟的高价值作业个数(本周)/高价值作业总个数

其中，基准时间为作业近30天平均完成时间加30分钟，作业晚于基准即延迟

作业性能分

☆ 考量作业的执行效率和健康度，诊断作业是否倾斜等性能问题

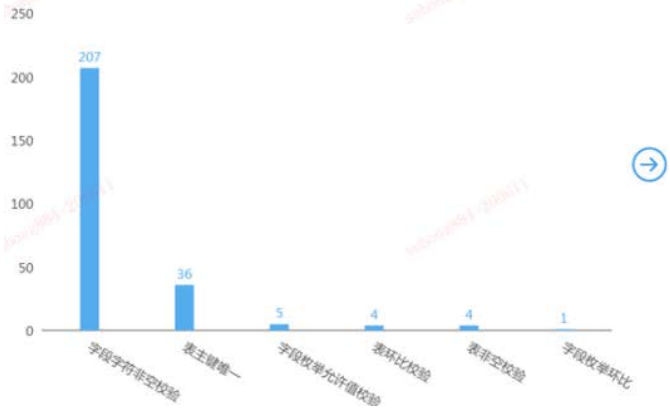
★ 指标公式：1 - 危急作业个数(本周)/作业总个数

从各质量维度的评价视角和指标公式可以发现，虽然数据质量监控的是表及字段的质量情况，但我们的质量分是设定在库这个层级。这么设计主要是为了更好地责任划分、统筹治理。
比如在银行业，每个库都有其对应的所属分层（如明细层、汇总层、应用层等），且每个库都有对应的库负责人，所以到库这个层级，我们能更好的分而治之，由库负责人对库的质量水平负责。

基于数据质量模型，我们还配套对应的数据质量监控报告。在报告中我们不仅能看到数据平台的整体质量评分，了解质量发展趋势，更能通过多维分析、单维深钻来了解平台的质量问题根源。

多维分析：详细展示七个质量维度的评分及趋势变化，每个维度下还配有TOP榜，用来展示低质量的库排名，督促库负责人进行优化、治理；

单维深钻：每一个质量维度都能从整体下钻到具体库及表，深入了解该维度质量评分低的具体原因，以便针对性地解决问题、提高质量；



告警规则（各库表明细）

告警规则	库名	表名	作业描述	
跑批超长作业	MID	BRR[redacted]_QK	BRR[redacted]TAL_Q	2
作业稳定性	MID	FS_BS[redacted]DER_INFO	FS_BSS[redacted]_SU	2
作业稳定性	MID	V_FS[redacted]TY_MQ	V_FS_[redacted]TIVT	V
作业稳定性	MID	FS_B[redacted]ETAIL	交易[redacted]交易	2
监控覆盖率	MID	FS_S[redacted]NFO	客[redacted]	H
监控覆盖率	MID	V_FS_S[redacted]RARY	FS_S0[redacted]RARY	H
监控覆盖率	MID	FS_S[redacted]NFO	机[redacted]	H
字段字符非空校验	MID	FS[redacted]DENTITY	个人[redacted]	F
告警响应度	MID	FS_P[redacted]DENTITY	个人[redacted]	F
作业准确性	MID	FS_BASE[redacted]DENTITY	个人[redacted]	F
字段字符非空校验	MID	V_FS_CU[redacted]TI	强[redacted]	F
告警响应度	MID	V_FS_CU[redacted]	[redacted]	F

1 - 25 共 9,971 条

综上，就是在数据质量治理方面的一些具体实践。数据质量治理是一个长期的、持续性的工作，不可能期望一蹴而就。

在治理过程中，需要不断优化质量短板，夯实质量基石。设目标、定责任，积极配合与行动，充分利用平台工具，共同建设一个数据乌托邦，让数据价值发挥耀眼光芒。

数据质量治理仅仅是数据治理的一个小环，而企业内部的数据质量问题并非不治之症，根据行业最佳实践开展管理体系提升，配备必要的软件，总能把这个问题解决掉，所谓企业内部的数据质量问题最终会消弭于无形。