

数据质量高效校验方法研究

程 胜¹, 郜美华², 冉 瑛¹, 乔玉洁², 黄 鹏¹

(1. 中国民航信息网络股份有限公司重庆分公司, 重庆 401120;
2. 中国民航信息网络股份有限公司, 北京 101300)

摘 要: 数据质量监控是数据处理领域的重要挑战。提出了一种创新方法, 通过规则配置高效生成数据质量验证规则与告警策略, 并在单一 SQL 查询中实现多指标并发校验, 显著提升了校验效率。详细阐述了技术方案设计原理并做了性能对比实验: 数据处理速度提升 10~20 倍, 规则配置效率提升 4~5 倍, 实际应用中的资源消耗指标 (CPU、内存) 与传统解决方案基本一致, 准确率仍保持在 99.99%, 为大数据环境下的数据质量管理提供了可量化的参考实践。

关键词: 大数据; 数据治理; 数据质量; 数据清洗; 数据质量监控

中图分类号: TP311.13; F562 **文献标识码:** A **DOI:** 10.19358/j.issn.2097-1788.2025.10.012

引用格式: 程胜, 郜美华, 冉瑛, 等. 数据质量高效校验方法研究 [J]. 网络安全与数据治理, 2025, 44(10): 75-79.

Research on efficient data quality validation methods

Cheng Sheng¹, Gao Meihua², Ran Ying¹, Qiao Yujie², Huang Peng¹

(1. Travelsky Technology Limited Chongqing Branch, Chongqing 401120, China;
2. Travelsky Technology Limited, Beijing 101300, China)

Abstract: Data quality monitoring is a critical challenge in data processing. This study proposed an innovative method that efficiently generates data quality validation rules and alert strategies through rule configuration, and implements multi-indicator concurrent validation in a single SQL query, significantly improving validation efficiency. The design principle of the technical solution was elaborated in detail and performance comparison experiments were conducted. Experimental results demonstrated that the solution improves data processing speed by 10~20 times, rule configuration efficiency by 4~5 times, while maintaining resource consumption (CPU, memory) at levels comparable to traditional solutions, with an accuracy rate of 99.99%. This study provides a quantifiable reference practice for data quality management in big data environments.

Key words: big data; data governance; data quality; data cleaning; data quality monitoring

0 引言

民航局 2022 年印发《关于民航大数据建设发展的指导意见》^[1], 明确要求加强数据质量管理。航信作为民航数据服务的主要提供商, 其数据质量直接影响航空安全与运营效率。

根据国际数据治理协会统计^[2], 数据质量问题导致企业平均每年损失 15%~25% 的营收。传统数据质量工具 (如 Apache Griffin) 存在配置复杂 (平均每个规则需 4~6 步操作)、执行效率低 (单指标独立 SQL 查询) 等问题。本研究通过实证分析, 提出了一种优化方案, 经生产环境验证, 配置效率可提升 5 倍 (从平均 15 min/规则降至 3 min/规则), 执行速度提升 8~10 倍 (多指标合

并查询减少 I/O 开销)。

1 数据质量介绍

数据质量管理是数据治理核心能力之一^[2], 在众多数据管理框架、评估模型中都有体现。如国际的 DAMA 数据管理框架中, 数据质量管理是十大能力域之一; Gartner 数据管理参考框架中, 建设阶段有数据质量规范等内容; 国内的 GB/T 36073—2018 数据管理能力成熟度评估模型^[3]中, 数据质量管理作为八大过程域之一。数据质量管理能力可从数据质量需求、检查、分析和提升四个维度进行评估^[4]。面向实际业务, 由于数据可能因各种原因出现各种各样数据质量问题, 这些问题可能会导致指挥层业务决策失误、运营效率降低、客户体验下降等

后果,因此,在数据治理过程中,提高数据质量显得尤为重要。数据的质量主要有以下几个关注点^[5-6]:

(1) 准确性:数据应准确无误,避免错误输入、计算错误或不正确的转换。

(2) 完整性:所有必要的数据都应存在,没有缺失值或空洞。

(3) 一致性:数据在不同系统间的一致性,避免数据冲突或不一致的副本。

(4) 时效性:数据应是最新的,以反映当前状态。

(5) 可靠性:数据来源可信,且经过验证,可以依赖其进行决策。

(6) 可用性:数据易于访问,格式正确,便于分析报告。

评估数据质量通常包括定义一系列指标来测量上述关注点,用数据质量评价指标来描述质量问题成为定义数据质量的一个重要依据。在《DAMA 数据治理知识体系指南》中,定义了以下 6 个核心数据质量评估维度^[7]:

(1) 完备性:存储数据量与潜在数据量百分比。

(2) 唯一性:避免数据重复,确保每个实体有唯一的标识。

(3) 及时性:数据能够被及时获取并用于决策过程。

(4) 有效性:数据符合预定义的格式和范围,如日期格式正确,数值在合理区间内。

(5) 准确性:数据正确无误,没有错误或偏差。

(6) 一致性:数据在不同来源或同一数据集内没有冲突或矛盾。

此外还有完整性、数值集合、先后顺序、精确度、参照完整性、继承关系、依赖关系等非核心度量维度。

表 1 为针对代码类数据类型适用的数据质量校验规则示例^[8],还有标志类、编号类、数值类、日期时间类、文本类等类型本文不进行展开介绍。

表 1 数据质量校验规则示例

类型	质量规则名称	质量规则说明	举例
代码类	非空检查	取值不为空	性别不能为空
	值域检查	在值域范围内	性别值为“男”“女”
	准确性	代码取值逻辑正确	与业务部提供的代码取值列表是否匹配

2 技术方案设计与验证

2.1 国内外研究现状

国内外在数据质量监控领域已有较多探索。国际方面,DAMA 提出的数据质量管理框架定义了完备性、唯

一性等核心评估维度,为质量校验奠定理论基础。Apache Griffin^[9]、Deequ^[10]等开源工具支持大数据质量评估,但存在配置复杂、依赖 Hadoop 生态等问题。

国内互联网企业在数据质量监控领域进行了积极探索,各头部厂商均推出了相应的解决方案:阿里巴巴的 DataWorks、腾讯的 WeData、美团的 DataMan 等商业产品都实现了较为完善的数据质量校验功能。然而,这些产品普遍存在商业化闭源、生态绑定等问题,难以满足企业的定制化需求。

在国内开源领域,微众银行的 Qualitis^[11]作为较具代表性的数据质量工具,经过笔者团队近一年的生产环境深度使用,发现其仍存在与国外开源如 Apache Griffin 类似的局限性:

(1) 配置复杂度高:每个检查点都需要完成多步骤的复杂配置流程,大大增加了使用门槛;

(2) 系统依赖过重:不仅需要 Hadoop、Hive、Spark 等大数据组件,还依赖完整的 DataSphereStudio 技术栈,部署和维护成本高昂;

(3) 数据源支持有限:仅支持 Hive 数据源,使用 Spark 引擎,无法兼容主流关系型数据库;

(4) 执行效率低下:采用单指标单 SQL 的校验模式,导致 I/O 开销大,性能瓶颈明显。

这些局限性严重制约了数据质量监控在企业的实际落地效果,亟需更轻量化、高效率的解决方案。

2.2 整体方案

本文通过系统地梳理航信数据中所有需要进行质量检查的关键点,发现 90% 的检查点都为单表进行数据质量校验,因此可以将每个数据集中的所有指标都通过一个 SQL 进行检查数据的抓取,以便大大提高数据质量校验效率;少数需要编写联表 SQL (或跨库 SQL) 进行校验也可在同一 SQL 内检查多个指标,整体流程图如图 1 所示。

2.2.1 数据集校验目标获取

获取指标所属对象有两种方式:

(1) 若指标直接为单表中的字段,则直接获取表字段列表;

(2) 少数指标需要跨数据源或数据表进行整合,则需要自定义 SQL,通过执行自定义 SQL,获取第一条数据,获得 SQL 中的字段名作为配置指标的目标对象,对于需要跨库的联表校验场景,则需使用如 Presto、Doris 等支持联邦查询的 SQL 引擎来解决跨库问题。

2.2.2 各字段数据质量指标配置

校验字段选择:通过上一步获取指标的所属对象列表,选择一个字段,作为指标所属对象。

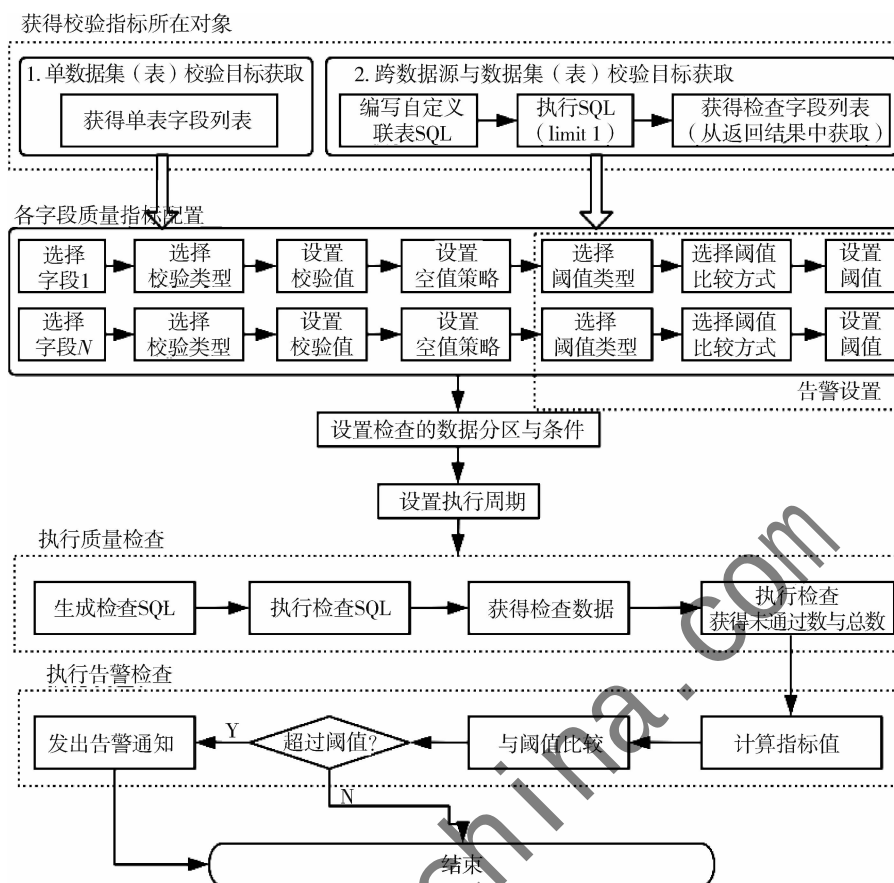


图1 数据质量校验整体流程图

校验类型选择：校验项为固定选项，选项有“日期时间校验”“字符长度校验”“行数校验”等。

校验值设置：此设置只有在校验类型为“日期时间校验”“字符长度校验”等类型时生效。

空值策略选择：值有“允许”“不允许”“允许但不统计”。若为“允许”，则当为空值时，将这条数据标记为校验成功；当值为“不允许”时，校验失败；当值为“允许但不统计”时，将这条数据不标记成功或失败，跳过此数据的统计。

2.2.3 告警设置

阈值类型选择：值有“比例”“固定值”“月波动”“周波动”“日波动”，通过阈值类型，计算出一个数据质量指标实际值，用于告警比较。

阈值比较方式选择：值有“大于”“小于”“等于”“大于等于”“小于等于”“不等于”，用于数据质量指标实际值与设置的阈值进行比较。

阈值设置：设置一个数值或比例，用于告警时与实际值做比较。

2.2.4 设置检查数据的分区与条件

此设置只对单数据集检查有效，用于指定需要校验

的数据集，此设置直接拼装为执行 SQL 的 where 条件，参考下文说明，此功能需要支持传入日期或时间变量，例如只校验昨天的数据时可以指定分区为：pt = '{yyyy-MM-dd-1}'，程序需将 {yyyy-MM-dd} 解析成今天的日期，再进行日期“-1”天，最终生成昨天的日期。

2.2.5 执行质量检查

数据质量检查步骤如下：

(1) 生成 SQL：联表 SQL 已提供了完整 SQL，因此只有单表校验需要生成 SQL，为保护数据质量校验程序，SQL 最多只查询 10 万条（此值可配置），SQL 模版为：

```
select {校验字段1}, {校验字段2}, {校验字段...}
from {表名} where {分区与条件} limit 100000;
```

(2) 执行 SQL：基于提前录入的数据库信息，通过 JDBC 执行 SQL，获得返回的数据集（数据库信息录入步骤不为本文内容，因此不做介绍）。

(3) 执行检查：拉取配置的质量校验规则，将执行 SQL 返回的数据逐行检查，获得有效数据（根据空值策略过滤）总量、不符合规则总数以及不符合规则的样例。

2.2.6 执行告警检查

计算指标值:通过质量检查计算出有效记录总数与未通过数,根据阈值类型,计算出告警指标实际值。

阈值比较与告警:将计算出的实际值与配置的阈值进行比较,若比较条件成立触发告警规则,等待所有指标完成后,将所有触发告警的指标项一次发给告警接收人。

2.3 方案优缺点

2.3.1 优点

本文方案的优点可总结如下:

(1) 方案效率高,在一个 SQL 中可同时检测多个指标;

(2) 支持数据源类型广,包括 Hive、Presto、MySQL、Oracle 等支持 JDBC 的数据库都可以进行检查,而像 Apache Griffin 与 Qualitis 都只支持 Hive;

(3) 操作简单,在一个页面中可同时配置多条校验规则,而 Apache Griffin 与 Qualitis 都是每一个指标需要进行一次配置;

(4) 若有数据标准的数据支持,则可以自动生成检查点,大大提高配置效率;

(5) 可在线查看校验不通过的数据值示例,大幅度提高找到出现数据质量原因的效率。

2.3.2 缺点

为保护检查程序不出现内存溢出,本文方案只能检查前 10 万(此值可配置)条数据(导致数据量超 10 万时,准确率无法达到 100%),对精确度高的指标,需要单独写 SQL 支持过滤掉无用数据,确保数据量在 10 万以内。

另外,若为跨数据源的校验,本方案需要 Doris、Presto 等这些有联邦查询能力的 SQL 执行引擎支持,复杂度较高。

2.4 性能对比实验

方案在航信旅客信息系统中部署后,选取生产环境中航班表(总数据量 3 亿,共 21 个检查点,每日分区数据平均 300 万)进行测试。三种方案的性能对比如表 2 所示。

表 2 三种方案性能对比(16 核 CPU/64 GB 内存)

方案	执行 耗时/s	CPU/内存	配置 耗时/min	准确率/%
Apache Griffin	743	88% /8.2 GB	15	100
Qualitis	687	68% /7.8 GB	12	100
本方案	34	77% /8.2 GB	3	99.99

如表 2 所示,本方案在 3 亿数据量、21 个检查点场景下,执行耗时从传统方案的 600 s 以上降至 34 s(效率提升 20 倍),规则配置耗时从 12~15 min 缩短至 3 min(效率提升 4 倍),平均准确率保持在 99.99%(由于每个检查点只采样前 10 万条数据,因此大于 10 万的检查点无法做到 100%);日均 300 万数据分区的处理时间从大于 10 min 压缩至 30 s 级,验证了该方案在大规模数据场景下,通过牺牲少许资源与准确率换取高效性与实用性,具有显著的性价比优势。

3 应用效果展示

3.1 单数据集校验设置

在航信某项目中,对某表的校验规则如图 2 所示。

3.2 执行结果列表

图 3 为每天的数据质量校验执行结果,展示了数据总行数、检查点通过数、未通过数、失败数(执行检查时程序出现异常);检查通过的数据条数,未通过数据条数,未通过的数据示例等。

质量概览

字段检查点配置

自由检查点配置

	字段		校验类型	校验值	空值策略	阈值设置(未通过+失败数/阈值)			#
						阈值类型	比较方式	阈值	
1	flight_no	Q	字符长度校验	4-8	允许	周波动	小于	5 %	自
2	flight_status	Q	枚举值校验	OP,CI,CL,CC,EC,OE,PC,X	不允许	比例	小于	1 %	自
3	flt_di_ind	Q	枚举值校验	D,I	允许	日波动	小于	5 %	自
4	id	Q	行数校验		不允许	月波动	小于	5 %	自
5	sch_arr_bjtime	Q	日期时间校验	yyyyMMHhmm	允许	固定值	小于	20	自

●若字段无法显示,请调整操作系统的缩放比例小于125%,或者缩放浏览器

数据分区(条件)

flight_date = '{yyyyMMdd-1}'

检查点名称: 常规字段检查 执行周期 0 15 4 ** ?

+ 自动添加检查点 + 添加检查点 自 删除 已 保存 @ 立即执行检查 sql格式化显示

图 2 数据校验规则配置

质量概览 字段检查点配置 自由检查点配置

	检查点名称	检查日期	数据量	状态	耗时	操作
>	常规字段检查	20250712 03:05:00	100000	不通过 (35/5/0)	10936ms	查看sql 效验明细
√	常规字段检查	20250711 03:05:00	100000	不通过 (35/5/0)	14508ms	查看sql 效验明细
检查点结果						
字段名	校验类型	校验值	空值策略	明细结果	阈值类型	状态
airline	字符长度效验	2-2	不允许	100000/0/0 100000/ 0 / 0	固定值	通过 (0 小于 1)
arrival_airport	字符长度效验	3-3	不允许	100000/0/0 100000/ 0 / 0	固定值	通过 (0 小于 1)
arrival_city	字符长度效验	3-3	不允许	100000/0/0 100000/ 0 / 0	固定值	通过 (0 小于 1)
arrival_terminal	正则效验	/[A-Z]*d[1,2]/	允许不统计	35952/14640/0 35843/ 1502 / 0	比例	通过 (1% 小于 5%)
arrival_time	null和空值效验		不允许	98275/17250/0 98266/ 1734 / 0	比例	通过 (2% 小于 5%)
boarding_no	数值范围效验	1 to 500	允许不统计	23054/18490/0 23337/ 1887 / 0	比例	通过 (2% 小于 25%)
book_time	日期时间效验	yyyyMMHhmm	不允许	47430/52570/0 47210/ 52790 / 0	比例	不通过 (53% 小于 10%)

图 3 执行结果列表

4 结论

本文提出的数据质量校验方案通过多指标合并 SQL 查询与规则自动化配置，在航信生产环境中实现了 20 倍执行效率提升和 4 倍配置效率优化。方案部署两年来，累计监测 509 个质量点，覆盖 65 张核心数据表，识别并修复数据质量问题 70 余次，使下游报表错误率下降 89%。相较于 Griffin 等传统工具，本方案在保持 99.99% 准确率的同时，显著降低了技术复杂度与运维成本。

本方案在批处理校验场景下取得显著成效，未来将重点突破两个方向：（1）实时数据校验：集成流式计算^[12]框架，实现航班动态等时效性数据的低延迟监控；（2）智能优化：基于机器学习自动生成质量规则^[13]，提升校验精准度，通过持续创新，构建“批流一体”^[14]、智能运维”的下一代数据质量管理体系，为民航数字化转型提供更强大的技术支撑。

参考文献

[1] 中国民用航空局. 民航大数据建设发展的指导意见 [EB/OL]. (2022-10-09) [2023-12-01]. <http://www.caac.gov.cn/XXGK/XXGK/>.

[2] 数据管理协会 (DAMA 国际). DAMA 数据管理知识体系指南 [M]. 北京: 机械工业出版社, 2020.

[3] 数据管理能力成熟度评估模型 (GB/T 36073—2018) [S]. 北京: 中国标准出版社, 2018.

[4] BATINI C, SCANNAPIECO M. Data quality: concepts, methodologies and techniques [M]. Berlin: Springer, 2016: 45-68.

[5] WANG R Y, STRONG D M. Beyond accuracy: what data quality means to data consumers [J]. Journal of Management Information Systems, 1996, 12 (4): 5-34.

[6] 信息技术 数据质量评价指标 (GB/T 36344—2018) [S]. 北京: 中国标准出版社, 2018.

[7] 郭志慧, 周傲英. 数据质量和数据清洗研究综述 [J]. 软件学报, 2002, 13 (11): 2076-2082.

[8] 吴信东. 数据治理技术 [J]. 软件学报, 2019, 30 (9): 2830-2856.

[9] Apache Griffin: data quality solution [EB/OL]. [2023-12-01]. <https://griffin.apache.org>.

[10] AWS Labs. Deequ: data quality library [EB/OL]. [2023-12-01]. <https://github.com/aws-labs/deequ>.

[11] WeBank. Qualitis: data quality platform [EB/OL]. [2023-12-01]. <https://github.com/WeBankFinTech/Qualitis>.

[12] 孙大为, 张广艳, 郑纬民. 大数据流式计算: 关键技术及系统实例 [J]. 软件学报, 2014, 25 (4): 839-862.

[13] CARBONE P, KATSIFODIMOS A, EWEN S, et al. Apache Flink: stream and batch processing in a single engine [J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 30 (4): 614-628.

[14] SCHELTER S, LANGE D, SCHMIDT P, et al. Automating large-scale data quality verification [J]. Proceedings of the VLDB Endowment, 2018, 11 (12): 1781-1794.

(收稿日期: 2025-06-24)

作者简介:

程胜 (1987-), 男, 本科, 高级工程师, 主要研究方向: 大数据与云计算。

邵美华 (1984-), 女, 硕士, 高级工程师, 主要研究方向: 数据应用工程、数据中台、大数据技术。

冉瑛 (1987-), 男, 本科, 高级工程师, 主要研究方向: 大数据、数据治理。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部

www.pcachina.com