大数据质量稽核的监控实施方法

Monitoring Method of Big Data Quality Audit

张　恒，曹丽娟，程新洲，徐乐西（中国联通研究院，北京 100048）
Zhang Heng，Cao Lijuan，Cheng Xinzhou，Xu Lexi（China Unicom Research Institute，Beijing 100048，China）

摘　要：
数据质量的好坏直接关系到信息准确度和业务有效性。只有解决数据质量问题并保证数据资产的科学性，才能有效反映企业运营和市场事实。科学有效的数据能够让企业经营决策可靠精确。大数据时代，数据将会呈爆炸式增长，海量的数据一方面为运营商带来大量信息资产，另一方面无用数据、干扰数据也呈指数级增长。因此，围绕企业大数据的生命周期管理，实现数据的全过程质量监控非常重要。

关键词：
大数据；稽核；规则配置；质量评分
doi：10.12045/j.issn.1007-3043.2021.11.002
文章编号：1007-3043（2021）11-0006-05
中图分类号：TN915
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
The quality of data is directly related to the accuracy of information and the effectiveness of business. Only by solving the problem of data quality and ensuring the scientificity of data assets can data effectively reflect the facts of enterprise operation and market. Scientific and effective data can make business decisions reliable and accurate. In the era of big data，data will grow explosively. On the one hand，massive data will bring a lot of information assets to operators. On the other hand，useless data and interference data will also increase exponentially. Therefore，Therefore，it is very important to realize the whole process quality monitoring of data around the life cycle management of enterprise big data.

Keywords：
Big data；Audit；Rule configuration；Quality score

引用格式：张恒，曹丽娟，程新洲，等. 大数据质量稽核的监控实施方法[J]. 邮电设计技术，2021（11）：6-10.

1 概述

大数据平台可用性是一项重要的平台运行指标，一个优秀的大数据平台，首先能够让操作者快速发现和理解数据，最终实现数据的高效应用。因此在整个过程中，平台中数据获取后的质量管控非常重要，只有对采集的数据进行严格的分析治理和质量管控，发现并完善数据的质量问题，才能解决用户对数据可用性的疑虑，保证后期业务的准确性和有效性。

数据质量管理主要依靠管理制度和事后稽核。在平台建设过程中，设计者通过改变模型管理和数据开发的模式，将后向管理变更为前向管理，从数据源头保障数据质量。

数据质量稽核从流程上可以分为以下3个层级。

a）元数据管理：最基础性的管理机制，可以识别、评价、追踪资源，达到有效管理。

b）数据的标准化管理：建立标准化体系，保证数据的统一运营和维护。

c）数据质量稽核：实现数据的深度质量检查，打造优质数据资产。

2 元数据检查

元数据管理应具备对元数据本身质量进行检查的功能，保证元数据自身的数据质量。元数据质量检

基金项目：工业和信息化部大数据产业发展试点示范项目（5G大数据跨行业异构融合创新应用试点示范）
收稿日期：2021-09-16

查包含但不限于以下内容：元数据一致性、元数据关系的健全性、元数据属性的填充率、元数据名称重复性和元数据关键属性值的唯一性。大数据平台一般会提供专门的界面进行元数据质量管控和呈现检查结果。

a）平台将提供在开发阶段定义好对象的元数据质量规则，并要求开发者在开发过程中按照规则录入元数据信息，并由系统进行统一检查。

b）平台提供元数据质量检查机制，及时发现、报告和处理元数据的数据质量问题。检查包括自动检查和人工检查2种方式。

c）平台提供可视化元数据血缘分析图，可进行影响分析、血缘分析，同时可以在血缘分析图中修改元数据信息，增加质量规则。

d）对于一些必须手工维护的元数据可通过开发维护人员进行手工维护、审批、发布。同时检查所提供的元数据与生产环境上元数据的一致性，形成元数据质量报告，产生手工维护的任务单，以确保元数据质量和可用性。

3 数据的标准化管理

数据标准是大数据平台数据治理的基础性工作，是数据治理建设中的首要环节，为大数据平台提供统一的数据标准定义和平台逻辑模型，是大数据平台进行数据治理的依据和根本，同时也是衡量大数据平台数据资产运营和管理的评估依据，最终能实现对大数据平台全网数据的统一运营管理。

平台通过建立统一的数据标准，结合制度约束、系统控制等手段，实现大数据平台中数据的完整性、有效性、一致性、规范性、开放性和共享性管理，提高大数据平台的数据治理水平。

数据资产标准化主要包括以下内容。

a）标准化的命名规则：数据的名称、编码、层级、层的属性名称等协调一致，统一管理，改变各源系统不规范的命名方式，避免同名不同意，同意不同名的现象。

b）统一数据扩展规则：对指标代码、元数据、子类等扩展要素的扩展规则进行统一限定，保证后续数据的持续规范管理。

c）标准化规范执行：平台通过对数据资产产生过程的监控（包括命名规范、信息完整性、合理性、基础信息完整性等以及存储周期、数据安全敏感信息和加密信息、权限赋权）以确保数据满足整体规划要求。

4 数据质量稽核规则体系

数据质量体系需要通过实践和规划的相互促进，不断完善改进，为此，需要确保数据架构合理，条理清晰，过程可控，知识积累传承，并通过监控和审计不断促进质量水平的持续提升。

数据质量管理是对采集入库的数据进行全面质量管理。开发者制定相应的技术手段和组织、流程、评价考核规则，通过平台操作，及时发现并解决数据质量问题，提升数据的完整性、及时性、准确性及一致性，提升业务价值。

数据质量规则配置如下。

a）提供便捷的IDE界面，可通过图形化或者标准SQL的方式，实现对数据质量规则的配置，允许对校验规则进行维护、优化等处理。

b）根据预先定义的质量规则，在应用运行时进行自动化监控。

c）提供元数据质量检查机制，及时发现、报告和处理元数据的数据质量问题。

d）提供问题定位分析，对问题的节点进行回溯，定位问题可能原因，分析其处理路径上可能存在的问题；提供问题影响分析，能对问题的严重性、影响面做出判断，并对重要问题提前进行预警。

数据质量监控功能设计如下。

a）提供数据映射分析，以拓扑图的形式对各类数据实体、数据处理过程元数据进行分层次的图形化展现，满足开发、运维或者业务上不同应用场景的图形查询和辅助分析需要。

b）根据预先定义的质量规则，在应用运行时进行自动化监控。

c）对数据采集层数据质量进行监控，主要包括文件接口、数据库接口、采集接口监控。

d）对数据处理过程进行监控，主要包括数据处理任务执行的情况，包括是否按时调度，是否成功等状态消息。

e）定期提供数据质量监控报告，根据系统健康状态按模板生成文本、图形等结果信息。

5 数据质量稽核的整体过程

数据质量体系需要通过实践和规划的相互促进，不断完善改进，为此，需要确保数据架构合理，条理清

晰，过程可控，知识积累传承，并通过监控和审计不断促进质量水平的持续提升。

设定稽核规则：通过不同的内置规则，可以对数据进行一定的计算处理，如空值、去重、最大、最小等，从而对数据有个直观的认识，发现数据缺陷，具体操作如图1所示。

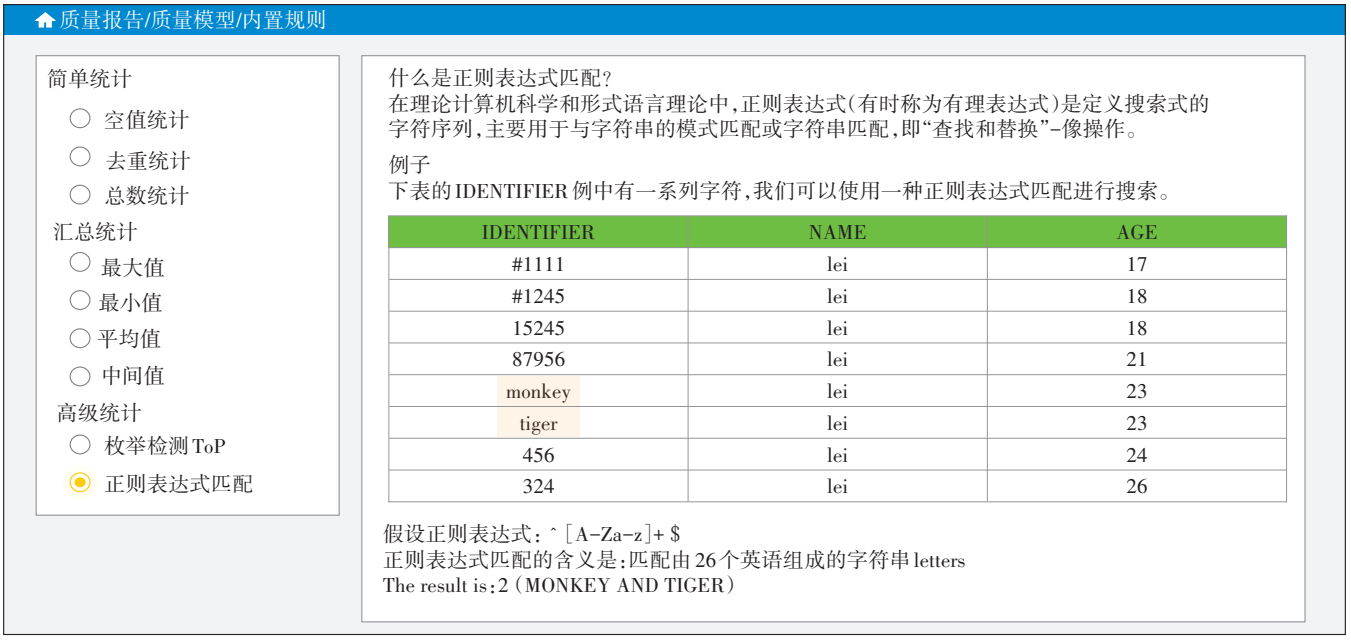

图1　规则设置

创建稽核模型：通过流程化的操作，操作者首先确定数据来源，根据不同数据源和目标，进行分区配置，从而建立对应的数据稽核模型，具体如图2所示。

稽核任务的创建：平台在使用过程中，操作者首先选择要进行稽核的数据时间、范围等要素，再加载对应的数据稽核模型，从而完成任务创建。

5.1 数据精度

数据精度决定后期业务分析的准确性，在平台使用分析中，操作者一般通过对比目标值与来源的真实情况来进行分析评估，流程如下。

图2　数据配置

a）选择用于比较的源数据和目标数据的集合和字段。

b）将目标字段与源字段进行关系映射。

c）将源数据集和目标数据集进行分区配置。

d）对分析模型进行配置，包括名称、参数、阈值等。

5.2 数据剖析

数据剖析是检查现有数据集中可用数据，同时收集相关数据的统计信息的过程，主要包括以下内容，具体如图3所示。

质量报告/质量任务/离线任务详情

查询

任务名称	最近一次运行时间	得分	最近平均得分（近10次）	计算公式	详情
唯一性-测试任务01	2021/10/14 18:00	0	0	（1-异常数/总数）×100	
hive22-去重统计-任务	2021/10/14 18:00	0	0	（1-异常数/总数）×100	
hive总数-准确性-任务test	2021/10/14 18:00	0	0	（1-异常数/总数）×100	
差异性任务	2021/10/14 18:00	100	100	（1-异常数/总数）×100	
数据完整性校验job	2021/10/14 18:00	100	100	（1-异常数/总数）×100	
唯一性校验任务-1	2021/10/14 18:00	0	0	（1-异常数/总数）×100	
及时性校验任务-1	2021/10/14 18:00	0	0	（1-（阈值-延迟时间）×100 ms；延迟时间≥阈值，得分为0	
数据差异性校验job任务id_age_desc	2021/10/14 18:00	100	100	（1-异常数/总数）×100	
手机号与地市差异性校验job	2021/10/14 18:00	94.17	94.73	（1-异常数/总数）×100	
手机号不为11位异常记录校验	2021/10/14 18:00	13.59	13.59	（1-异常数/总数）×100	

图3　数据结果详情

a）选择需要进行剖析的目标数据集和字段。

b）定义将应用于所选字段的语法检查逻辑。

c）将目标数据集进行分区配置。

d）对分析模型进行配置，包括名称、参数、阈值等。

5.3 数据模型

在所有分析中，数据模型的建立是最重要的一环，不同的数据模型可以分析不同的数据质量。数据模型可以根据分析需求和数据类型，从5个维度进行设计，首先建立相应的模型，其次定义模型详细的源、目标、以及映射关系等的属性，最终在任务中可调用该模型进行数据任务的设定。本文以平台中的“数据准确性校验模型1”为例进行说明。

5.3.1 模型信息

模型信息是对模型的基本情况的展示，包括类型、源、源分区、源条件、目标、目标分区、目标条件和责任人等信息，能够清晰地呈现该模型的属性，以方便使用，具体如图4所示。

基本信息

模型名称	数据准确性校验模型-1
模型描述	描述
模型类型	accuracy
源	demo_src
源分区大小	1 hour
源条件	dt=20210409 AND hour=#HH#
目标	demo_tgt
目标分区大小	1 hour
目标条件	dt=20210409 AND hour=#HH#
责任人	admin

图4　模型信息说明

5.3.2 模型精确度计算映射信息

source.id=target.id AND source.age=target.age AND source.desc=target.desc

该映射代表了源数据各字段与目标字段的一致性对比结果，其中source和target代表了不同的数据源，id、age、desc代表要对比的具体字段。

精度计算公式如下：

$$\text{准确率}(\%) = \frac{\text{demo\_tgt 和 demo\_src之间的匹配记录总数}}{\text{default.demo\_tgt记录总数}} \times 100\%$$

该公式分子代表了2个数据源（demo_tgt和demo_src）的匹配结果为一致的数量，分母代表了demo_tgt中的匹配的数据数量，demo_tgt和demo_src代表了2个对比数据源。该计算公式可以计算出稽查数据的准确率。

通过建立任务可以将模型应用到不同数据的稽核中，给出数据准确性的结果，图5是针对2021年10月28号到10月29号入库的2批不同数据的稽核结果，横坐标是以小时为单位，可以看出不同时间对应数据的稽核结果，以方便使用人员针对问题进行后期处理。

图5为2类数据准确性校验模型的处理结果，从

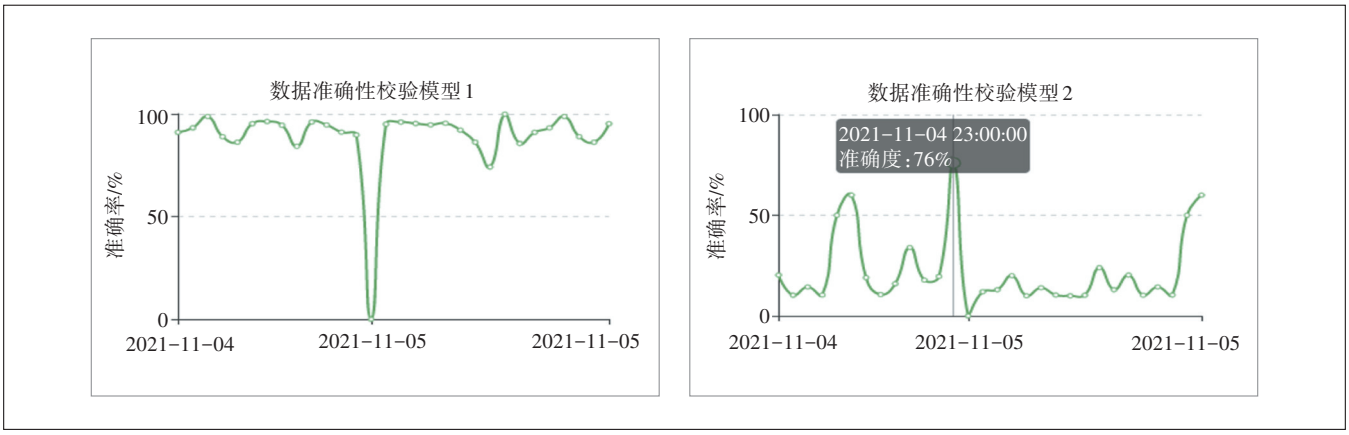

图5　准确度检验结果

图5可以看出随着数据的不断采集和入库，数据稽核任务以小时为周期持续性进行，因此，数据使用者可以实时查看数据准确性，以方便及时发现处理数据问题。

5.4 数据质量评分

评分是以分值来展示数据的质量，从而形成直观的数据质量感受，评估一般包括以下3个层面。

a）质量评分=参与评分的各质量维度评分总和/参与评估维度项。

b）某个维度质量评分=参与该维度评估的任务的评分总和/参与该维度评估的任务的总数。

c）任务某个维度评分=该维度下参与评估各项规则得分之和。

5级维度数据质量说明如下。

a）准确性：度量数据是否与指定的目标值匹配，如金额的校验，校验成功的记录与总记录数的比值。

b）完整性：度量数据是否缺失，包括记录数缺失、字段缺失，属性缺失。

c）差异性：度量数据记录是否重复，属性是否重复；常见度量为hive表主键值是否重复。

d）及时性：度量数据达到指定目标的时效性。

e）有效性：度量数据是否符合约定的类型、格式和数据范围等规则。

6 结束语

大数据是未来数字化的重要能力。数据质量是保证业务顺利执行的重要要素，因此数据质量稽核非常重要。数据稽核包括普通的数据完整度、完善性等核查，同时根据不同数据的来源及业务特点，可以建立不同的特征模型进行针对性的稽核，这样可以保证数据业务特征的准确性。因此未来数据稽核更关键的是针对性的稽核，尤其是针对不同业务特点的定制化数据稽核，它是保证数据质量的关键方法。

参考文献：

[1] 胡尚华. 稽查信息化的定位与策略思考[J]. 经济研究参考，2017(70)：58-60，94.

[2] 徐启建. 基于Spark的交通监控目标大数据分析系统的设计与实现[D]. 北京：北京邮电大学，2018.

[3] 解铁铮. 电信服务开通系统大数据分析子系统的设计与实现[D]. 北京：中国科学院大学，2017.

作者简介：

张恒，高级工程师，硕士，主要从事通信大数据分析及行业的研究工作；曹丽娟，工程师，主要从事大数据算法研究及行业应用研究工作；程新洲，教授级高级工程师，主要从事通信大数据分析及架构的研究工作；徐乐西，教授级高级工程师，主要从事大数据算法研究及行业应用研究工作。