发电企业设备域数据质量管理方法及流程设计

马勇

（华能山东发电有限公司 山东省济南市 250014）

摘要：随着数据资源在企业数字化转型中越来越重要，企业管理对数据管理的需求也在不断增加，而如何提高企业数据质量是一个关键问题。为应对发电企业在数据质量管理方面所面临的挑战，本文提出了一个基于分布式技术进行数据质量管理的创新解决方案。而对集中式数据质量系统的性能缺陷这部分问题，我们研究了数据质量系统的特点，利用国内外大数据相关技术，然后提出了基于 Hadoop 分布式数据处理框架的解决方案。

关键词：发电企业；数据质量，管理方法；流程设计

引言

在大数据时代背景下，随着现代信息技术的发展企业管理更加精细化，因为对数据的要求更高。在新的时代背景下企业的核心竞争力高低很大程度上取决于数据资源处理能力的强弱，这种能力对企业业务管理有着重要影响。数据资源是企业发展和未来规划的重要参考依据，事关企业的生存问题。电力行业的竞争日益激烈，为更好地适应未来电力行业发展的趋势，规避未知风险，数据管理的相关研究早已成为电力行业研究的热点同时取得一系列研究成果，基本的数据管理系统已经初见雏形。因数据管理人才缺失等许多方面引起的的数据问题明显减少了很多，但也会遇到一些问题：数据涉及范围广，涉及多单位的生产、安全、人力资源、财务等业务领域；校验规则繁多，各业务域都有相应的数据质量校验规则，涵盖统一性、一致性、准确性等类型的质量检查点；数据量大，经统计，全公司设备域主数据已超 200 万条，每月围绕主数据所发生的业务数据则是几何级数增长，但数据质量未作评定；数 据 校 验 耗 时 长，以最重要的量测数据为 例，当校验数据量达 到一定数量级时， 原基于集中式数据存储 和计算架构的数据质量管理系统由于数据读写和 数据统计操作瓶颈，完成数据质量校验和问题分析 需耗费超长以上的系统运行时间，不满足数据应用及时性需求；如需要在短时间内执行大量的校验规则，对磁盘读写 性能、CPU 和内存性能要求较高。

1 数据质量管理系统的设计思路

1.1 数据清洗

数据清洗的第一步是“数据清洗”，可以分为 7 个部分。分别为数据子集的选择、数据列命名、删除重复值、缺失值处理问题、一致化处理、数据清洗排序、异常值处理。其中缺失值处理是最为复杂的环节，需要人工手动补充，计算机系统进行缺失值删除，平均值替代和统计值换算。

在数据列命名中若同时出现两个或多个含义相同的列名时需要对其中的一个或者多个列名进行重新命名，减少因列名问题导致的分析结果偏差现象；原始数据中可能会存在数值重复或缺失现象，此时需要对重复的数据值进行删除，只保留第一条原始数据即可。对于缺失的数据（即数据集中存在无数据的单元格）可人工进行补录，通过函数模型查找缺失值，Ctrl 选中相关数据，在公式框输入相关数据。

最基本的数据清除可以采取的清理办法有以下几种：忽略元组、使用属性的平均值办法、中间值、最大值、最小值、使用整体变量提取空缺值还有其他复杂的概率统计函数值等。比如固定资产在进行数据清除的过程之中。因为之前的数据不清，数据不完整，有缺陷。所以这时候可以运用其他的方法，利用到固定资产、平均值、中间值、最大值这样的概率进行统计分析。最后确定一个比较相对完整的累积折旧值，这样它可以用来补全最基本的信息。设备设施，它们主要的数据有时候比较困难的去限定界限，可以根据安装的位置，功能的特点，电压的等级，设备设施变动的方向去进行分辨，根据分类中相同的特点进行别类整理，考察他们之间的属性来平滑属性的数据。

1.2 构建模型

该部分选择 Vlookup 进行函数构建和数据透视表，最终描述统计分析。

（1）数据清理过程的纠错管理

设备设施账单以及资金卡片完整性、精确性的检验除了实行单位全部检查系统上线所需要的静止数据、动态数据之外，在数据导入以及检测过程中，对信息系统创造的接口配置经过开发改造成自动检测、数据清理、检验并处理重合的信息化方法。

（2）自动检测属性错误的方法

人工检验设施设备账单和资金卡片合集的错误，需要花费许多的人力，物力和时间，并且容易发生错误，需要经过高效率的方式自动检验数据汇集的错误，采取的方式自动检验数据汇集的错误，采取的方式主要部分有:根据统计的方式、聚集方式、关联规定方法等。这里经过维系之前的数据和使用的集合函数(Sum，Count，Min，Max 等) 规定的象征等方法来结束属性出错的检测内容。

2 数据质量管理的目标

2.1 确保数据准确性

发电企业数据治理工作中，包含着电网数据、用电数据、设备数据、监测数据等多种数据类型，数据质量管理成效对电力系统的运行效果具有显著影响。发电企业在进行设备域数据质量管理时，应当首要明确数据准确性的管理目标，强调关键性数据的精准，旨在借助准确的数据内容，分析和反馈实际运行情况。企业应选择高精度的数据采集设备、尽可能避免数据传输中的外界因素干扰或人为操作失误对数据准确性的影响，通过数据点校验、异常检测等手段，确保数据与实际情况充分符合，从而促进数据准确性的提升。

2.2 保障数据完整性

保障数据完整性目标是指发电企业在进行设备域数据质量管理时，应当包含全部所需信息，防止出现数据缺失或数据遗漏的现象，通过细致严谨的数据质量管理措施，促进发电企业设备域数据可靠性提高。保障数据完整性旨在实现实体不缺失、属性不缺失、记录不缺失和字段值不缺失四个方面，负责人员可以通过完整性校验等方式优化管理手段，从而达到保障数据完整性的管理目的。

2.3 加强数据一致性

发电企业设备与数据质量管理工作中，不同数据源的数据信息需要进行数据整合，由于数据源之间的格式差异和结构差异，容易造成数据冗余、数据冲突甚至数据矛盾的现象。企业数据质量管理应当通过技术手段统一数据的格式、定义和规范，消除数据中的异构性，整合多种数据源的数据信息，从而形成统一的数据视图，为企业决策与优化调整提供可靠的数据支撑，确保数据在不同系统和应用中保持一致。

2.4 提高数据时效性

数据时效性，是指数据的及时传输与更新。发电企业传统数据质量管理手段存在运行速率差的问题，管理人员难以快速获取电力系统运行状况的准确信息，当出现运行故障时难以及时作出决策，阻碍了电力系统的稳定运行。发电企业设备域数据质量管理工作应当通过高质量技术手段，通过数据传输优化、数据更新周期控制等措施促进数据时效性的提高，为电力系统的实时性监测提供有效保障。

3 数据质量管理的实践

3.1 设计思路

在 x86 服务器的基础上，可使用比较便宜的服务器建立一些集群；具有良好的延展性，每当业务增长的时候，就需要处理非常多的数据，可以水平扩大和增加比之前多的节点。

3.2 分布式数据质量管理系统

此系统采用 J2EE 这样的结构进行开发，把它们划分成了三个层次。J2EE 是一种构建分布式、可扩展和可靠的企业级应用，有利于落实企业级分布式应用平台的解决方案，在发电企业数据质量管理实践中存在重要价值。J2EE 的核心理念在于将应用程序拆分成多个可重复利用的组件，基于开发模型与多种技术规范进行应用处理。应用 J2EE 构建分布式数据质量管理系统，能够提高分布式数据质量管理系统的可扩展性、灵活性和安全性，通过多层架构支持，借助极强的分布式计算能力，促使各个独立结构都能自主进行演化升级，从而达到提高发电企业数据质量管理效益的核心目的。

3.2.1 用户交互层

用户交互层这个界面有许多功能，可以解决许多问题，比如有关系统安装、 数据方面管理规则、结果报告单打出等。 主要的功能包括以下几方面。

（1）数据属性管理：对数据库里的数据进行管理，包括一些数据的名称、每一段的名称、数据的属性等信息。

（2）模型管理：对于被校验库表之间的联系进行管理。

（3）规则配置：管理质量校验规则，包括规则名称、规则 描述以及校验脚本。

（4）策略管理：对于检查数据的具体操作时间和正确参考进行管理。

（5）报表管理：对数据检查生成的报告加以管理，包括导出、报告、筛查等功能。

（6）平台管理：对于组织机构、用户信息、系统日志的管理等。 用户交互层的作用用户怎样进去此系统的一个界面，在有关系统设计方面的问题，考虑了此界面对于用户的实用性和便利性，对于降低系统的反应时间

有很大的帮助，极大的提高了用户的体验感；在数据检查方面的工作有很大的帮助，系统有自动检查数据错误的功能，对于 AJAX 技术的应用，我们就可以提前加载出元数据，在用户操作流畅度这方面有了质的飞跃，提升了系统加载相关数据的效率；激发规则参数潜在功能。在此基础上借助功能的筛选作用筛查出类似的规则进行规则参数配置进而降低系统规则数量缩短规则反应时间。

3.2.2 数据处理层

数据处理层是整个系统的核心，负责规则执行、缺陷数据查询分析等任务，主要功能包括以下几方面。

（1）策略执行任务调度： 进行控制执行策略的任务调度，按一定的周期开始执行策略。

（2）丢失数据分析：负责丢失或者残缺明细进行分析，并结算出数据质量报告。

（3）规则执行引擎：负责执行规则脚本，记录执行日志。

（4）缺陷明细查询和导出：从 Hadoop 集群查到的数据的接口由用户交互层进行提供。也是数据处理层的重要组成部分，数据质量管理系统依靠的是规则的执行引擎，这对于验证的结果有非常大的影响。对于此规则有以下几个要素需要考虑。性能是第一考虑因素。规则执行引擎也就是一种嵌入在应用程序中的软件，而开源 Kettle 则被作为错误数据进而提取引擎。Kettle 可以把一个表裂解并将它提取为很多条线路的几段，所以可以解决大部分数据提取的问题。此外，使用缓存机制，缓存出来的结果可以被优先处理，用这个方法可以提高效率。

4 流程设计

4.1 数据查询

实现功能包括但不限于:系统支持实时查询数据资产的所有信息，包括申请、审批、明细属性、统计信息、使用评价、变更历史、分发历史等内容:系统实现标准快捷查询、高级查询、递进查询等功能，支持基于语法分析和词法分析等的全文检索功能。

4.2 数据分发

实现功能包括但不限于:支持数据资产实时发布、任务发布、模板导出等功能，支持分发规则管理，分发任务管理、分发计划管理、分发历史管理和分发监控等。

4.3 数据安全

实现功能包括但不限于:支持对系统菜单、参数、日志、消息和备份进行控制和管理;应满足灵活管理、分层级严谨授权原则，支持角色权限控制，并对系统状态进行监控和异常管理等;对于一些被保护的数据，平台支持用户根据业务需要自己进行脱敏规则，对它们进行进行脱敏处理，提高对五于敏感隐私数据保护的可靠性。

4.4 工作流管理

提供完善的流程管理功能，支持系统管理员自己进行对于各个层次分级的流程，包括对流程实例进行新增、编辑、删除、发布、停用等，有关流程表单和流程图的配置;流程表单需要支持见到就可以摸到的控件拖拉;支持自由拖拉画出流程图的审批步骤;流程中可设置节点关联审批人/角色，节点可同时关联多个审批人/角色，支持流程一键导出为独立文件，并支持流程文件的导入，实现在线快速的流程发布，即发布即可用。

5 结束语

为了解决发电企业设备域数据质量管理难点，本文提出了分布式数据质量管理系统的解决方案，首先对分布式存储和计算的核心技术进行研究，建立大数据存储和计算的平台，并根据业务数据的不同，设置了大量的数据质量校验规则。数据质量验证所需时间大大缩短，有效提高了系统的处理和分析效率，保证了公司设备域数据应用的可靠性、及时性，助力发电企业电力系统的安全稳定运行。

参考文献：

[1]袁满,杜杨杨.领域数据质量知识建模方法研究[J/OL].现代情报:1-11[2023-02-09].

[2]袁燕,卢建军,熊莺.以“首页关键指标缺陷率”为抓手，提升病案首页数据质量[J].现代医院,2022,22(12):1866-1868+1873.

[3]黄俊超,胡勇,徐启丰,赵春林.航材管理信息系统数据质量评价方法[J].军事交通学报,2022,1(11):28-33.

[4]周林兴,林凯.大数据时代档案数据质量控制：现状、机制与优化路径（摘编）[J].中国档案,2022(10):76.

[5]冯雨晴,谭雅文.央行征信系统数据质量管理问题探讨[J].征信,2022,40(10):35-38.

作者简介：马勇，男 1973.10 山东日照，单位 华能山东发电有限公司，学位 本科 研究方向 数字化转型、信息化应用、大数据分析，邮编：250014

受华能集团总部科技项目 HNKJ21-HF292 技术研究资助