

面向企业数据中台的数据质量监控技术研究

郑子怡, 杨 静

汉江师范学院数学与计算机科学学院, 湖北 十堰

收稿日期: 2023年11月29日; 录用日期: 2023年12月29日; 发布日期: 2024年1月10日

摘 要

由于上市公司数据量的飞速增长, 数据质量问题逐渐成为各企业所面临的重大挑战。隐含的数据质量问题会导致数据处理分析结果的不准确、决策误判等诸多关键问题。企业数据中台的数据统计是企业进行管理决策的重要基础, 提升数据的使用价值, 对企业的日常经营、精准营销、管理决策、风险管理等领域发挥积极有效作用。本项目对企业中台数据质量监控技术展开深入研究、进一步提高企业的数据质量、促使企业信息化管理工作节约成本、提高企业行使权力效率有着重要的理论及现实意义。

关键词

数据中台, 大数据, 数据监控平台, 数据清洗

Research on Data Quality Monitoring Technology for Enterprise Data Center

Ziyi Zheng, Jing Yang

School of Maths and Computer Science, Hanjiang Normal University, Shiyan Hubei

Received: Nov. 29th, 2023; accepted: Dec. 29th, 2023; published: Jan. 10th, 2024

Abstract

Due to the rapid growth of data volume of listed companies, data quality has gradually become a major challenge faced by enterprises. Implicit data quality problems can lead to many key problems such as inaccurate data processing and analysis results and misjudgment of decision making. Data statistics in the enterprise data center is an important basis for enterprises to make management decisions, improve the use value of data, and play a positive and effective role in the daily operation of enterprises, precision marketing, management decision-making, risk management and other fields. This project has important theoretical and practical significance to carry out in-depth research on the data quality monitoring technology of the enterprise middle desk, fur-

ther improve the data quality of the enterprise, promote the enterprise information management to save costs, and improve the efficiency of the enterprise to exercise power.

Keywords

Data Center, Big Data, Data Monitoring Platform, Data Cleaning

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

在互联网技术日趋成熟的时代背景下,产生了海量的业务数据,而数据已经成为互联网企业非常依赖的新型重要资产,为了有效的管理和使用这些数据,数据质量监控也就不可或缺。基于大数据的各类应用的基础是拥有高质量的数据资源,而数据质量的监控与治理正是数据质量持续提升的基础性保障,故从全球的范围而言,不断加强数据质量监控已正式成为企业持续提升管理能力的关键性任务。因此,针对各大企业来说,进一步提高企业的数据信息质量至关重要。本文旨在研究面向企业数据中台的数据质量监控技术,以提高企业数据的质量和可靠性。

2. 相关研究综述

2019年6月28日,在大阪举行的二十国集团(G20)峰会上,在谈到有关数字经济的议题上,G20提出要打造开放、包容和非歧视的数字经济,致力于实现基于信任的数据自由流动和跨境数据流动,推进以人民为中心、赋能和可持续的数字化转型,在此期间,中国首次将数字经济纳入G20议程,同时提出《二十国集团数字创新合作行动计划》,旨在推动数字技术创新应用,实现创新成果普惠共享,在数字产业化、产业数字化方面推进国际合作,释放数字经济推动全球增长的潜力[1]。而数字经济的发展离不开高质量的数据资源,故加强数据质量监控必不可少。

企业启动中台战略的目的是为了提高效率和解决业务匹配度问题,最终达到降本增效,让一切业务数据化,一切数据业务化。数据业务化的关键要素之一就是要先确保数据质量,才能有效的发挥出数据的价值。数据本身是不会撒谎,而收集样本的维度、数据量会撒谎,毕竟收集数据的质量良莠不齐,然而,数据质量又常常反应的是数据的“适用性”,即数据满足使用时需要的合适程度。故从企业长远发展来看,进行数据质量监控已是大势所趋。

据市场调查网发布的《中国数据中台标杆企业市场发展形势现状及行业前景预测研究报告》显示,我国数据中台行业目前仍处于发展中期,市场规模持续攀升,在2020年国内数据中台市场规模约为70亿元,未来随着参与布局数据中台的企业增长,市场规模仍旧呈现增长趋势,预计到2023年达到180亿元。

目前,国内外对于数据质量监控技术的研究主要集中在以下几个方面:数据清洗、数据验证、数据标准化和数据安全。其中,数据清洗是提高数据质量的重要手段之一。数据验证是一个过程,通过多种方法检查数据是否符合我们预期的规范和标准。总的来说,数据验证和标准化都是为了提高数据的可靠性和准确性,从而更好地支持决策和数据分析。数据安全则涉及到数据库安全、数据备份与恢复、数据防泄漏、隐私计算等方面。

本项目主要针对企业数据在生成、采集、传输、使用及保障等阶段中可能出现的风险问题,拟定数据中台质量监控技术研究方案。为顺利实现对企业业务发展、运营情况、经营考评、财务管理等核心资源尽早提供更多最有效的数据信息维护[2]。

3. 面向企业数据中台的数据质量监控技术研究

(一) 如何提升数据中台的数据信息质量

提升数据中台的数据信息质量可以从以下几个方面来进行:

- 1) 建立完善的数据质量监控机制,对数据中台的数据质量进行实时监测。设定明确的目标和标准、构建合理的管理流程、设计有效的检查方案、对数据质量的定义、评估、监控和改进等,以确保数据中台的数据质量符合业务需求[3]。通过及时发现和解决数据质量问题,能够提高数据可靠性,降低风险,并为企业的决策提供更加准确的数据支持。
- 2) 提升数据安全性、优化数据处理流程。通过数据加密、访问控制、安全审计等方式提升数据安全性,防止数据泄露和被攻击,同时避免因流程问题导致的错误和损失。
- 3) 定期对数据中台的数据质量进行评估、开展数据治理培训,并且引入先进的数据质量管理技术,如人工智能、机器学习等。根据测评结果进行改进和优化,以提高数据中台的数据质量(见下图 1)。只有全面管理和监控数据中台的数据质量,才能更好地满足业务需求并推动企业的发展。

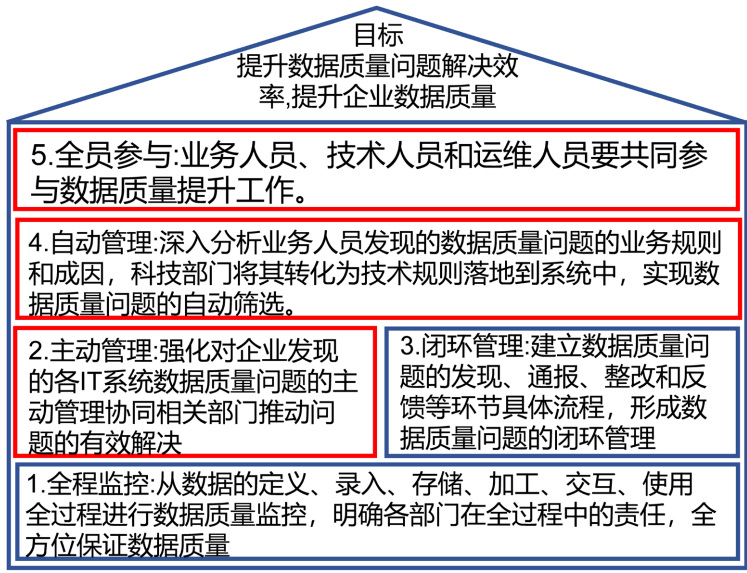


Figure 1. Methods of data quality in indigenous enterprises
图 1. 提升企业数据质量问题的方法

(二) 预期规划

- 1) 初步开发一套适用于企业数据中台的数据质量监控系统。通过该系统,能够尽量有效地监控数据中台所持有的数据质量,提高数据使用的准确性和价值。
- 2) 制定一套科学、合理的数据质量评估指标。通过这些评估指标,为企业数据质量管理服务。数据中台的建设必然会帮助企业建设数据标准,包括数据建设规范和数据消费规范[4]。针对数据质量监控过程中可能存在的风险,建立相应的风险识别和规避机制,对数据安全进行保障。
- 3) 数据质量监控系统和数据质量评估指标,旨在确保数据的准确性、完整性和一致性,以及满足企

业的特定需求和标准。通过实时监控数据质量,能够及时发现和解决潜在的数据质量问题。可以提升企业的数据使用的准确性和价值,帮助企业提高竞争力。

(三) 从系统实现的角度上,要解决三个关键技术。

1) 检核脚本的自动生成。数据质量检核实际上是按照脚本执行并筛选出有问题的数据。随着数据质量度量规则的增多,通过人为手工编写脚本的方式就无法应对快速增加的度量规则,通常一个中等规模的金融企业,就具备上千条度量规则。因此通过配置的方式,利用脚本生成引擎自动生成检核脚本,是数据质量工具必须具备的功能[5]。

2) 多线程检核架构。检核脚本的执行时间是影响能够及时查看到数据质量问题的另一个关键因素。在脚本执行过程中,需要采用多线程并发来执行保证在较短的时间内检核出有问题的数据。

3) 数据质量报告。数据质量报告是对企业数据质量情况的总结和分析,需要能够从不同维度系统、部门、检核类别等维度生成固定数据质量报告。还需要支持按照选择的数据质量规则,时间等条件,来生成个性化的数据质量报告。

(四) 数据质量监控技术的实施步骤

1) 数据源分析是数据质量管理的第一步,需要对数据源进行全面深入的分析,了解数据的来源、格式、特点等关键信息。针对不同的数据源,需要采用不同的数据处理方法和技术,以确保数据的正确性和完整性。

2) 数据预处理是数据质量管理的第二步,在获取到原始数据后,需要进行一系列的数据清洗、转换和整合工作。数据清洗的主要任务是提高数据的可用性。数据转换是将不同格式的数据转换为统一的标准格式,以方便后续的数据整合和分析。数据整合是把在不同数据源的数据收集、整理、清洗[6],转换后加载到一个新的数据源,为数据消费者提供统一数据视图的数据集成方式[7]。

3) 数据质量评估是数据质量管理的第三步,在数据预处理之后对数据进行全面的质量评估。通过采用统计分析、模式识别、关联规则挖掘等多种方法和技术。

4) 实现数据质量监控技术是数据质量管理的第四步。通过实时监测数据的质量,及时发现和解决潜在的数据质量问题,保障数据的准确性和可靠性[8]。在实现过程中,需要运用异常检测、时间序列分析、聚类分析等多种技术手段,对数据进行全面监控和管理。

4. 实验特色及结果分析

(一) 实验特色分析:

1) 全面覆盖:数据中台的数据质量监控技术可以全面覆盖数据的整个生命周期,包括数据的产生、存储、处理、分析和应用等。这种全面的监控可以确保数据的质量和准确性,及时发现和解决潜在的数据质量问题[9]。

2) 实时监控:数据中台的数据质量监控技术可以实时对数据质量进行监测和控制。这种实时的监控可以及时发现和解决数据质量问题,避免数据出现错误和造成损失。通过实时监控,企业可以更好地掌控数据的质量。

3) 智能分析:数据中台的数据质量监控技术可以利用人工智能和机器学习等技术进行智能分析。这些智能分析方法可以提高数据质量监控的效率和准确性。通过智能分析,企业可以更加及时地发现 and 解决潜在的数据质量问题,提高数据的可靠性和可信度[10]。

4) 自动化修正:数据中台的数据质量监控技术可以自动化修正数据质量问题,包括数据的清洗、整理、标准化和校验等操作。这种自动化修正可以快速纠正数据质量问题,提高数据质量管理的效率。

面向企业数据中台的数据质量监控技术具有全面覆盖、实时监控、基于规则的监控、可视化界面、

智能分析、自动化修正和安全性保障等特色和创新。这些技术可以提高数据中台的数据质量和管理效率,更好地满足业务需求和发展需要。

(二) 结果分析:

为了验证面向企业数据中台的数据质量监控技术的可行性和有效性,我们进行了一系列实验。实验结果表明,该技术可以有效地提高企业数据的质量和可靠性,降低了数据治理的成本和风险。同时,该技术还可以帮助企业及时发现和解决数据质量问题,提高企业的竞争力和创新能力。

5. 结论与展望

本文研究了面向企业数据中台的数据质量监控技术,以提高企业数据的质量和可靠性。通过实验验证,该技术可以有效地提高企业数据的质量和可靠性,降低数据治理的成本和风险。未来,我们将继续深入研究该技术,探索更加高效、智能的数据质量监控方法和技术。尝试开发一套适用于企业数据中台的数据质量监控系统,业务覆盖数据采集、数据清洗、数据存储、数据计算等环节,有效监控数据中台所持有的数据质量,提高数据使用的准确性和价值,以满足企业的实际需求。

基金项目

汉江师范学院 2023 年省级大学生创新训练计划项目资助,项目编号为 S202310518050。

参考文献

[1] 笪志刚, 赵新月. 数字贸易对东北亚区域经济的促进研究[J]. 东北亚经济研究, 2023, 7(2): 48-59.

[2] 徐彩云. 企业数据质量管控关键技术问题研究[J]. 数字技术与应用, 2018, 36(6): 69-70.

[3] 宋峰光. 山东省农村信用社基于 ERP 系统的企业级总账系统设计与实现[D]: [硕士学位论文]. 济南: 山东大学, 2013.

[4] 曹荣. 传统型 RT 批发市场的数字化转型[D]: [硕士学位论文]. 南京: 南京邮电大学, 2021.

[5] 苏建华. W 供电公司海量实时/历史数据管理信息系统的设计与实现[D]: [硕士学位论文]. 镇江: 江苏大学, 2014.

[6] 宋剑. 智慧校园大数据分析决策平台研究[D]: [硕士学位论文]. 阜新: 辽宁工程技术大学, 2019.

[7] 王江. S 公司 ERP 运行诊断分析及提升对策研究[D]: [硕士学位论文]. 成都: 西南石油大学, 2013.

[8] 汤天宇. 计算机技术在计量检测体系中的应用分析[J]. 大众标准化, 2023(11): 189-191.

[9] 徐鹏飞. 水质监测中存在的问题及对策探讨[J]. 科海故事博览, 2023(24): 49-51.

[10] 吴信东, 应泽宇. 数据中台框架与实践[J]. 大数据, 2023, 9(6): 137-159.