

基于大数据技术的运营监测（控）领域数据质量管控研究

针对电力行业数据量大、种类繁多、跨专业等特点而引起的数据质量不高的现状，以数据中心数据为研究对象，探索揭示数据质量问题产生的原因，以元数据、数据质量控制、数据挖掘等理论为基础，并对数据加工，再加工的流转过程中的数据质量检查、数据质量控制及数据质量评估等进行深入研究，健全数据管理机制，从制度与流程、组织与人员、技术与工具等方面支撑数据资源的统一管理，确保企业级数据资源集中融合、规范有序，构建电力大数据质量评价指标体系，研究电力大数据质量规则库的构建与关键技术，提出适合于电力大数据数据质量控制与评估体系模型和数据质量问题分析与处理措施，在此基础上开发面向电力大数据处理的原型系统并进行验证，进而补充和完善技术研究成果，为提升公司数据可用性和运营效率提供支撑。

张自强¹ 高建勇¹ 李志² 张驰^{1/1} 1.国家电网甘肃省电力公司
2.安徽南瑞继远软件有限公司



张自强/高级工程师

0 引言

信息化时代，各企业面临着数据量的大规模快速增长，例如，根据IDC最近报告到2020年，全球数据量规模将扩大近50倍。目前，大数据规模还是一个快速变化的指标，单一数据集规模从几十太字节到数拍字节不等范围波动。数据多样性的增加主要是由于半结构化、非结构化数据，以及包括互联网日志、互联网搜索、手机通话记录及物联网等数据类型造成^[1-2]。

随着数据的大爆炸，数据被快速地创建和移动，对数据实时处理的要求也越来越高。在高速互联网时代，通过基

于实现软件性能优化的服务器和高速电脑处理器，创建实时数据流已成为大势所趋。企业不但需要掌握如何快速创建数据，而且需要知晓如何快速处理、分析并将分析结果返回给用户，以满足用户的实时需求。

大数据具有多层结构的特点，这也就意味着大数据将会以多变的形式和类型呈现。与传统的业务数据相比，大数据还具有不规则和模糊不清的特性，因此也就造成传统的应用软件很难甚至无法对大数据进行分析。随着时间的演变，传统业务数据已具有了标准的格式呈现方式，能够被标准的商务智能软件识别。目前，企业面临的挑战就是对各种形式呈现的复杂数据进行处理并从中挖掘价值。

在电力行业，随着坚强智能电网的迅速发展和要求，信息通信技术也与电网生产、企业管理以前所未有的广度和深度快速融合，信息通信系统已经逐步成为智能电网的“中枢神经”，并支撑新一代电网生产运营和管理发展。目前，公司已初步建成了国内领先、国际一流的信息集成平台。随着三地集中式数据中心投运，一级部署业务应用范围延伸及扩展，以及结构化、半结构化和非结构化数据中心的上线运行，电网业务数据从数据总量和类别上都已初具规模。伴随着智能电表的逐步普及，电网业务数据将从时效性层面进一步丰富和拓展。海量、实时的电网业务数据中已经凸显了大数据的“量类时”特性，电力大数据分析迫在眉睫。

关键词/Keywords

电力大数据 ·
数据质量 ·
评价指标体系 ·

但是电力行业数据质量及数据共享利用水平仍有待进一步提高,一是数据对公司分析决策支持深度较低,数据存在一源头多、统计口径不一致的现象;二是数据对公司生产经营管理水平的支持度有待进一步提升,数据质量差异化,部分数据无业务系统支撑,缺乏统一的公司数据标准和认责机制^[3];三是一线人员手工录入数据量依然繁重,数据重复录入率较高;四是数据质量管控水平相对滞后,管控手段局部化,没有形成全面的、完备的数据质量管控体系,从而制约了公司数据资产价值的最大发挥。因此,迫切需要围绕企业数据全生命周期,实现对电力企业数据的全过程管控,强化数据基础,提升数据质量,确保数据的准确、及时、有效和可信。

基于大数据技术,开展运营监(测)控领域数据质量管理研究将紧密围绕公司推进数据管理体制和工作机制创新要求,依托运营监测(控)中心业务系统应用现状,借鉴公司数据治理经验,研究运营监控中心的元数据标准^[4-5]、数据质量管理机制^[6]、基于大数据背景下的运营监测(控)中心数据质量评价体系^[7]、数据质量规则库^[8]、数据质量问题分析及处理措施,旨在提供运营监测(控)中心数据流转过程中的质量检查、质量控制和质量评估,解决运营监测(控)中心信息支撑系统建设与应用所面临的数据质量问题,为提高数据的可用性和运营效率提供支撑。

1 元数据模型、数据质量管控机制、数据质量评价体系、数据质量规则库、数据质量问题分析与处理措施研究与应用

1.1 元数据模型

为了统一元数据管理策略,支持集中式的元数据存储,需要建立遵循CWM元数据模型标准的元数据模型库,包含了各类有代表性的运监中心元数据的有效描述,研究内容主要包括以下几个方面:

1) 分布式文件系统元数据服务模型,即集中式元数据服务模型、分布式元数据模型和无元数据服务模型。

2) 360°全景视图表示法研究。

3) 数据质量元模型研究,包括数据质量规则

元模型、数据质量流程元模型、数据质量评价结果元模型、数据质量分析元模型和数据质量辅助元模型。

1.2 数据质量管理机制

结合电力行业数据数量大、种类繁多及跨专业环境下的数据应用需求,研究制定公司电力大数据环境下的数据管理目标,建立数据质量管控机制,对数据进行分类、分级,依据数据来源与数据类型选择不同的数据管理策略和控制策略。研究全面的数据准入机制及关键技术、数据质量接入规范、数据质量评估机制。

1) 研究数据接入检查机制,主要包括数据身份认证、数据安全性检查、数据规范性检查等。数据安全性检查主要检查数据来源,数据规范检查主要检查数据结构、数据规模、数据类型等。

2) 研究数据接入机制,主要包括数据安全管控机制、数据分类、分级评定、数据接入应急机制、数据测试管控机制、数据回退机制。

3) 研究数据质量评估机制,主要包括数据质量考核评价体系、数据质量公布机制及奖惩机制,保障数据质量提升工作的有效落实。

4) 研究数据接入关键技术,主要包括数据接入异常检查技术、数据库网关技术、数据流量控制技术、数据复制技术研究、数据等级保护技术等。研究电力行业数据数量大、种类繁多环境下数据接入的数据ETL、数据接入服务等关键技术。

5) 为了有效推进公司数据共享和数据融合,建立长效数据质量评价机制,开展常态化数据质量评价工作。

1.3 数据质量评价体系

针对电力行业数据数量大、种类繁多、跨专业等特点和开展数据质量管理所面临的问题与挑战,研究面向电力数据量大、种类庞杂及横跨专业多的数据质量管理规范,构建以业务系统为依托的数据质量管控体系、数据质量评估体系,设计面向电力数据量大、种类庞杂及横跨专业多的元数据模型、数据质量监控规则及体系、数据质量评估指标,通过运用标准化的数据质量管理规范,实时监控,在线评估,加强数据质量事中管控,事后评估,问题分析及整改,提高公司决

策分析依据的正确性和实用性。研究内容主要包括以下几个方面:

1) 设计业务数据指标体系, 形成数据标准规范。

2) 构建数据质量评价体系, 主要包括数据质量元素和数据质量约束规则的定义及分类、数据质量评估指标的建立及算法、数据质量分析评估体系结构和流程。

3) 数据质量评价模型的研究, 主要包括基于上下文的数据质量评价模型研究、元组粒度的数据质量模型研究、数据项粒度的数据质量评价模型研究等。

1.4 数据质量规则库

1) 研究设计面向电力大数据数据质量规则库, 包括规则的可配置和专家经验知识库的建立。

2) 研究基于元数据的数据质量约束规则。

3) 研究数据质量规则的自动识别和动态探查技术, 清洗脏数据。

1.5 数据质量问题分析与处理措施

数据质量问题分析与处理措施研究主要包括以下几个方面:

1) 研究设计面向电力大数据背景下的数据质量问题分析的流程、制定合理数据质量问题分析与处理策略。

2) 研究设计面向电力大数据背景下的数据整改跟踪流程, 实现及时驱动业务系统或相关人员整改数据异常事件, 发起整改流程, 全程跟踪整改情况, 并对整改结果执行在线验证, 从而全面提升整改效果。

3) 研究制定电力大数据质量报告发布标准与流程, 根据数据实时监控结果、质量评价结果、数据质量改进措施、改进结果等自动生成质量管理报告, 方便管理层快速了解数据质量管理状况。

2 结束语

针对电力行业数据数量大、种类繁多、跨多专业等特点而引起的数据质量整体不高的现状, 以数据中心数据为研究对象, 探索揭示数据质量

问题产生的原因, 以元数据、数据质量控制、数据挖掘等理论为基础, 并对数据加工及再加工的流程过程中的质量检查、质量控制和质量评估等进行深入研究, 健全数据管理机制, 从制度与流程、组织与人员、技术与工具等方面支撑数据资源的统一管理, 确保企业级数据资源集中融合、规范有序, 构建电力大数据质量评价指标体系, 研究电力大数据质量规则库的构建与关键技术, 提出适合于电力大数据数据质量控制与评估体系模型和数据质量问题分析与处理措施, 在此基础上开发面向电力大数据处理的原型模型并进行论证, 进而补充、完善及丰富技术研究成果, 为提升公司数据可用性和运营效率提供支撑。

参考文献

- [1] 李皎. 大数据时代到来对电力行业发展提出新要求[J]. 华北电业, 2012 (4): 82-83.
- [2] 孟小峰, 慈祥. 大数据管理: 概念、技术与挑战[J]. 计算机研究与发展, 2013, 50 (1): 146-149.
- [3] 刘军, 吕俊峰. 大数据时代及数据挖掘的应用[N]. 国家电网报, 2012-05-15(10).
- [4] 李建花, 曲守宁. 面向信息集成的元数据模型的研究与应用[J]. 计算机与数字工程, 2010, 38 (12): 149-152.
- [5] 袁满, 陈永芳. 元数据驱动的数据中心字典管理元模型研究[J]. 长江大学学报, 2011, 8 (1): 92-94.
- [6] 刘景义, 张晖, 赵伟涛, 等. 中国石油A1A2系统数据质量管理体系建设实践[J]. 信息技术与信息化, 2011(3): 61-66.
- [7] 袁满, 张雪. 一种基于规则的数据质量评价模型[J]. 计算机技术与发展, 2013, 23(3):81-89.
- [8] 焦殿科. 基于蚂蚁算法的突发数据准入优化及仿真[J]. 计算机工程与应用, 2008, 44(4):119-121.

(收稿日期: 2015-03-30) EA

作者简介

张自强 (1975-), 男, 高级工程师, 从事电力公司运营监测工作, 国网甘肃省电力公司。

高建勇 (1976-), 男, 硕士, 工程师, 从事电力公司运营监测工作, 国网甘肃省电力公司。