

DMA 漏损控制大数据处理模式

马金锋¹, 陈求稳², 徐 强¹, 饶凯锋³, 安 伟¹, 马 梅¹

(1. 中国科学院生态环境研究中心 中国科学院饮用水科学与技术重点实验室, 北京 100085;

2. 南京水利科学研究院, 江苏 南京 210029; 3. 中国科学院生态环境研究中心 环境水质学
国家重点实验室, 北京 100085)

摘 要: 管网独立计量分区(DMA)管理是目前供水管网漏损控制的主要措施之一, 日益受到自来水公司的重视。随着传感器成本的降低, 数据采集越来越密集, 产生的海量数据对传统数据处理和分析模式提出挑战, 导致供水企业对传感器投入的效益比偏低。大数据技术的发展为 DMA 有效管理提供技术支撑, 基于大数据技术指导 DMA 漏损控制决策成为必然趋势。阐述了大数据技术在 DMA 管理中的应用价值, 从数据采集、汇聚、传输、存储、计算及应用角度, 论述了适用于 DMA 数据处理和决策分析的大数据处理模式, 并实例验证了模式的可行性。该模式具有通用性, 能够实现整套模式的移植使用, 可为大中型供水企业的 DMA 数据管理及应用提供技术支持。

关键词: 供水管网; 独立计量分区; 大数据; Hadoop; HBase; MapReduce

中图分类号: TU991 **文献标识码:** A **文章编号:** 1000-4602(2019)10-0036-06

Big Data Processing Mode for District Metered Area Leakage Control

MA Jin-feng¹, CHEN Qiu-wen², XU Qiang¹, RAO Kai-feng³, AN Wei¹, MA Mei¹

(1. Key Laboratory of Drinking Water Science and Technology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China; 2. Nanjing Hydraulic Research Institute, Nanjing 210029, China; 3. State Key Laboratory of Environmental Aquatic Chemistry, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China)

Abstract: District Metered Area (DMA) management is one of the main measures for leakage control of water supply network, which has been paid rising attention by water supply enterprises. As the cost of sensors decreases, data collection becomes more and more intensive. However, the huge amount of data generated poses a challenge to traditional data storage and processing paradigm, which has resulted in low profitability of water supply enterprises against sensor input. The development of big data technology may provide a promising support for the efficient management of DMA, and the decision-making of DMA leakage control based on big data technology has become an inevitable trend. This paper explored the application value of big data technology in DMA management, discussed the big data processing mode suitable for DMA data processing and decision analysis from a perspective of data collection, transmission, storage, calculation and application, and gave an example to verify the feasibility of the mode. The model had universal property and could be used in any waterworks. And, it could provide technical support for the management and application of DMA data in large- and medium-sized water supply enterprises.

Key words: water supply pipe network; district metered area(DMA); big data; Hadoop; HBase; MapReduce

城市供水管网漏损管理,已经成为我国城市供水安全保障过程中亟待解决的问题。独立计量分区管理是实现区域计量、查漏以及漏损控制的主要技术手段。通过分区管理可以将漏损问题分解到每个DMA上,有利于管理者分析管网各个区域的水量分布及变化情况,进而通过均衡管网运行压力,达到优化水厂调度和管网运行的目的。因此,越来越多的供水企业开展DMA建设工作,积累了一批工程的实践经验^[1-3]。与此同时,研究人员也在不断探索适用于DMA漏损控制的新技术与新方法^[4-6],推动了其方法学的发展。

随着DMA监控网络建设的普及,企业获取的信息资源日益丰富。许多大型城市供水企业通过监控系统采集和存储供水管网中的管道压力、管道流量、用户用水总量和水质等数据,但是就目前的情况来看,大量的生产数据都仅限于实时监测和存储阶段,尚未发现有水司能够对海量存储数据实现有效分析挖掘^[7]。由于大量数据被当作历史数据储存,管理人员未能对这些数据进行深入的综合分析,缺乏对企业供水管网管理系统的优化和升级,导致这些数据信息的价值未被挖掘^[8]。此外,传统数据处理模式并不适合进行大数据处理分析,因而造成供水企业对传感器投入的效益比偏低。为了突破目前“数据海量、信息缺乏”的数据应用困境,充分挖掘数据中潜在的重要信息,总结归纳历史规律,评估现有状态,预测未来发展趋势,优化规划配置运营模式,进而智能辅助漏损控制决策,必须探索新的数据处理模式。

大数据是信息时代的主要技术之一,目前已经广泛应用在各行各业中。大数据技术可以对海量、高增长率和多样化的数据资产进行分析处理,进而提取出具有更强决策力、洞察力和流程优化能力的信息。大数据技术的发展为DMA数据处理提供了一种新的模式,探索适用于DMA数据处理和决策分析的数据处理模式是必然趋势。

1 大数据技术在DMA管理中的应用价值

采用长时间序列DMA数据分析甄别漏损是业内共识。分析过程中涉及到流量、压力、管材、用户数、管龄等多源数据,这些数据在甄别漏损时具有不同含义:流量分析可以反映出用水模式,压力分析可以识别可能漏损区,管材和管龄关联历史漏损可以预测漏损概率,用户数则与用水模式相关。这种蕴

含信息的多样性和数据体量大、类型多、数据密度低的特点共同决定了DMA大数据的价值发挥。在面对长时间序列DMA数据处理时,传统的数据处理模式无能为力,具体表现为存储上无法持久化,计算上无法全量化,不能充分利用数据实现预测价值进而达到资源优化配置目的,因此需要探索新的数据处理模式。

大数据技术在DMA管理中的应用价值体现在四个方面:存储价值、全量数据挖掘价值、预测价值和资源优化配置。

① 存储价值

随着DMA业务的不断增长,传统的关系型数据库已经不能满足大规模海量数据存储的需要。这主要体现在:a. 存储容量有限,当数据量达到PB级时,主流关系型数据库(如Oracle、Sql Server等)将无法负荷;b. 存储内容单一,主流关系型数据库比较适合存储结构化数据,而对于空间数据、图片、音频和视频等非结构化数据存储时效率明显下降;c. 存储模式扩展困难,传统关系型数据库多采用物理扩展模式对存储内容进行物理分割,容易造成数据分散和无法高效检索,进而导致应用程序开发复杂化。大数据技术提供分布式文件系统、分布式数据库和冗余备份机制,适合DMA结构化和非结构化数据长期持久化存储。

② 全量数据挖掘价值

传统的数据分析方法不适合处理海量、多维数据,主要体现在以下几方面:a. 不具备全量数据分析能力,采集的数据无法有效保留所有信息,必然会影 响分析结果的准确性;b. 缺乏对非结构化数据的处理能力,在面对空间数据、图片、音频、视频等非结构化数据处理时分析能力有限;c. 可扩展性差,通过提升服务器性能进行扩展运算的方式,无论是成本还是处理能力均不能满足要求。大数据技术能够提供离线计算、流式计算和内存迭代等多种分布式计算模式,支持全量数据挖掘,适用于DMA管理数据的全量分析。

③ 预测价值

预测是大数据技术应用的核心价值。传统的预测方法,由于受数据源单一(维度低)、数据量小(只能采样)、数据处理方法局限(无法处理多维度),难以捕捉数据中潜在模式或者预测精度低。大数据技术采用海量、多维数据构建完备场景库,基于全量数

据挖掘处理方法构建高性能模式识别方法,通过两者结合克服传统预测方法不足,使得发现规律、预测未来成为可能。在 DMA 管理领域,利用大数据可以分析用户的用水习惯以及区域用水量结构,预测区域用水量,可以对部分区域的用水量结构进行调整,以此来优化供水系统,提高供水企业的经济效益。此外,可以根据长时间序列历史数据,确定管网背景漏失水量与管网材质、管龄、用户数等管网特征的相关性,预测管网背景损失水量。

④ 资源优化配置。

利用大数据分析,能够总结经验、发现规律、预测趋势,预测的最终目的是为辅助决策服务。掌握的数据信息越多,决策才能更加科学、精确、合理。基于大数据预测分析的结果,进行资源优化配置,是大数据应用的落地点和真正价值。企业可以借助大数据,提升管理、决策水平,提升经济效益。在 DMA 管理领域,大数据可以辅助识别管网漏损、调整用水量结构、优化管网故障诊断与维护等,以此来保证供水管网系统管理运行的节能性和安全性,提高供水公司的整体经济效益。

2 DMA 漏损控制大数据处理模式

从数据采集、汇聚、传输、存储、计算及应用角度,按照“数据-计算-应用”的三层组织结构,提出 DMA 漏损控制大数据处理模式,具体框架如图 1 所示。



图 1 DMA 漏损控制大数据处理模式

Fig. 1 Big data processing mode of DMA leakage control

其中,数据层负责 DMA 管网漏损控制数据的采集、汇聚和存储,包括传感器硬件环境、分布式网络环境、数据传输交换体系和数据存储体系;计算层

负责 DMA 管网漏损控制大数据的处理和计算,包括 DMA 数据处理专业算法模型及适应这些算法模型的计算框架;应用层负责将计算层输出结果解释、发布、展示以及提供给第三方应用程序的集成接口,包括数据访问服务、数据可视化和复杂业务环境接口层。统一运维管理系统提供整个大数据平台的系统运维、集群部署安装和软件包管理工作。

3 可行性分析

3.1 流程分析

以北京市某已建 DMA 小区为例。该案例中,采用 CINTERION wireless modules 完成数据采集后,通过无线传输网络,经由 Kafka 缓存后存入原始数据库。以 pentaho kettle 为 ETL 工具,通过制定数据抽取规则(从原始数据集中选择流量、压力、时间字段)、数据转换规则(异常值去除、缺失值补全)和加载规则(全量加载或增量加载),依据抽取规则创建 kettle 的 transformation 变换。基于变换创建 kettle 的 job(作业),经由作业调度器定期执行作业,实现原始数据集到 HBase 目标存储数据库的迁移。

批量计算分析由 Quartz 作业调度框架统一管理。计算分析算法,比如存量漏损、新增漏损、控制评估、异常报警等,在框架内被组织成相互独立的作业。作业以 MapReduce 程序的组织形式,读取 HBase 数据,并在执行算法后将结果写回 HBase。最后,Web 服务器和报表生成器从 HBase 数据库中读取上述作业的计算结果,以页面和定制报表形式返回给终端用户。整个 DMA 漏损控制大数据处理模式流程如图 2 所示。

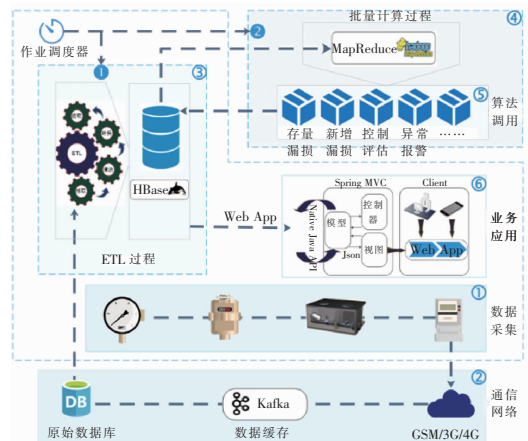


图 2 DMA 漏损控制大数据处理模式流程

Fig. 2 Flow chart of DMA leakage control big data processing mode

3.2 作业调度器

DMA漏损控制大数据处理过程的核心是Quartz作业调度器,它充当中央处理器角色。在底层数据层,调度器负责组织、调度、执行和监听ETL作业,实现原始数据集到HBase目标存储数据库的数据迁移;在核心计算层,调度器负责将各种业务分析算法,如存量漏损、新增漏损、控制评估、异常报警等各种计算需求等组织成相互独立的作业,作业依据业务需求不同,从HBase数据库中抽取不同数据执行算法。

调度器可根据业务需求进行作业的定制,比如创建、修改和删除作业,并根据实际需要,设定合适的作业执行时间,定时执行作业,监听作业状态,确保计算任务顺利完成,最后将计算结果写入HBase目标存储数据库。因此,在Quartz框架的支撑下,计算分析任务可以灵活创建、删除、修改,赋予DMA漏损控制计算分析具备可扩展性。

3.3 分布式计算

MapReduce是Hadoop生态系统中离线计算框架之一,它将计算程序分发到各个数据节点,通过在数据节点当地执行计算的方式大量降低系统I/O,非常适合于批量程序运算。

在图2框架中,MapReduce负责将数据分析方法如日常统计分析、预测算法、评估报表、分析指标等各种计算需求等组织成相互独立的计算程序(作业),同时设定各个作业的调度时间,框架密切注意调度时间,执行作业的计算过程。MapReduce的离线计算模式不需要低延迟,适合在后台处理数据。HBase的在线模式赋予数据读写访问低延迟的能力,适合在前台用户交互。两者的紧密结合赋予系统较高的可用性。

3.4 案例分析

最小夜间流量(Minimal Night Flow, MNF)是评估DMA实际漏损情况的重要指标。Xu等^[9]建立了MNF和DMA之间的多元回归模型,用于管网漏损控制的经济性评估。MNF的具体计算公式如下:

$$MNF = (a_1 L_{CI} + a_2 L_{NCI} + a_3 N_p) A^b P^c \quad (1)$$

式中 L_{CI} ——铸铁管长

L_{NCI} ——非铸铁管长

N_p ——用户数

A ——管龄

P ——凌晨02:00—05:00管网压力

a_1, a_2, a_3, b, c ——分别为需要率定的参数

以上述模型为例,该业务模型与大数据处理计算模式之间的关系见图3,其中MapReduce计算框架负责将每个DMA执行MNF模型的计算任务请求分发给各个独立的计算节点,各个计算任务之间相互独立,MapReduce计算框架监督计算任务的执行情况。

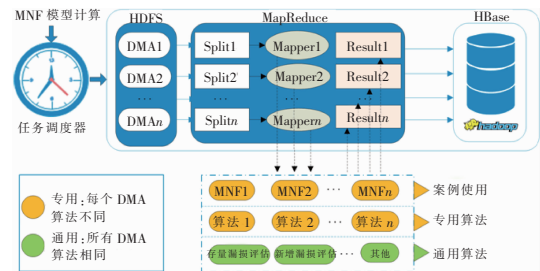


图3 MNF模型案例分析

Fig. 3 Case analysis of MNF model

相比传统的计算处理模式,该处理模式的优势在于:

① 容错性,在单个计算任务失败时,比如某个DMA的MNF模型执行任务由于机器发生故障而失败,集群框架会自动回收任务,并再分发至计算节点。

② 扩展性,一方面可以自由增加新的计算模型或者算法,如图3中可以在MNF模型的基础上扩展新的存量漏损评估算法或者新增漏损算法;另一方面,从硬件扩展角度来讲,在不影响现有集群运行的情况下可灵活增加新的计算节点。

③ 并发计算的高效性,每个计算任务占用一个独立CPU核,最大化利用多核计算资源,充分发挥硬件性能。

④ 负载均衡,确保多计算任务合理执行,提升集群计算资源的利用率。

⑤ 经济性,采用开源技术框架,具有低成本、可移植性强的优势。

3.5 集群计算环境

鉴于Hadoop平台向后兼容性,采用Hadoop 0.20.2版本,利用1台惠普服务器和4台联想PC机搭建Hadoop集群,集群的硬件配置、软件系统及集群中角色划分见表1。测试结果表明,集群提供了良好的横向扩展、负载均衡、故障修复和容错的计算任务执行环境。

表 1 DMA 漏损控制大数据集群环境配置

Tab. 1 Configuration of DMA leakage control big data cluster

机器名	硬件配置	软件系统
master	HP ProLiant DL580 g7, CPU:32 核, 2.67 GHz, 内存:64 G, 硬盘容量:900 G	操作系统:CentOS7, Hadoop 版本:0.20.2 HBase 版本:0.90.6, Zookeeper 版本:3.3.5 JDK 版本:1.7.0_76
s200, s201 s202, s203	ThinkCentre M8411t, CPU:8 核, 3.4 GHz, 内存:32 G, 硬盘容量:1 T	操作系统:CentOS7, Hadoop 版本:0.20.2 HBase 版本:0.90.6, Zookeeper 版本:3.3.5 JDK 版本:1.7.0_76

3.6 人机交互过程

图 4 为 DMA 漏损大数据处理可视化主界面, 顶部导航栏划分了平台功能模块, 包括总览概况、在线监测、设备管理、存量漏损评估、新增漏损评估、漏损控制评估、设备异常报警和工单派发。总览概况是对平台整体情况的概述, 左侧部分对小区基本信息、管网信息和漏损情况进行汇总和统计; 中部上侧以 GIS 地图表达各 DMA 小区空间位置, 中部下侧展示了背景流量统计和设备异常报警分析结果; 右侧上

部为新增漏损预警, 饼状图统计分析所有漏损情况, 列表部分显示每个小区最近一次漏损状态, 列表尾部按钮用于向下派发单任务; 右侧中部为漏损控制决策, 饼状图部分统计分析漏损控制措施, 列表部分列出所有小区最近一次采取的具体措施, 列表尾部按钮用于查看实施过程详细信息, 右侧下部为漏损控制措施效果评价, 饼状图部分统计分析所有派工完成情况, 列表部分显示派工措施完成前后最小夜间流量、压力、节水量变化。



图 4 DMA 漏损控制大数据处理可视化

Fig. 4 Visualization of DMA leakage control big data processing mode

本例测试了北京市已建 21 个 DMA 小区 3 年连续监测数据。结果表明, 所有的数据采集、入库、算法计算、查询均达到预期效果, 证明 DMA 漏损控制大数据处理模式可行, 具有很强的扩展性, 适用于 DMA 管理数据长期持久化存储和全量计算分析。

4 结语

随着供水业务的日益发展, 数据规模的不断增加给传统的关系型数据库存储和数据处理模式带来极大挑战。大数据的存储、分析、管理解决方案为解决海量数据处理提供了新的思路, 基于数据节点横向扩展的分布式存储及程序向数据迁移的分布式计

算策略为高效查询和处理海量数据提供了基础。大数据技术在 DMA 漏损控制管理中具有四个价值, 分别是存储价值、全量计算价值、预测价值和资源优化配置价值。鉴于大数据技术的应用价值和现有 DMA 管理的发展水平, 从数据采集、汇聚、传输、存储、计算及应用角度, 论述了适用于 DMA 数据处理和决策分析的大数据处理模式, 并实例验证了模式的可行性。该模式具有通用性, 可为 DMA 漏损控制海量数据持久存储、管理、分析及应用提供参考和技术支持。

基于大数据技术的 DMA 漏损控制数据处理模

式主要面向大中型供水企业,解决海量数据采集、传输、存储和计算处理难题,它是现有业务运行系统的有益补充和智能化管理水平提升的有效辅助工具。由于内在的低成本、可移植性好、扩展性强特点,适合于建设规范化、标准化的通用性大数据平台,可实现整套模式的移植使用,大幅降低开发费用,减少重复投资和资源浪费。

参考文献:

- [1] 刘锁祥,赵顺萍,曹楠,等. 供水管网漏损控制研究和实践[J]. 中国给水排水,2015,31(10):22-25.
Liu Suoxiang, Zhao Shunping, Cao Nan, *et al.* Research and practice of water loss control of water distribution network[J]. China Water & Wastewater, 2015, 31(10): 22-25 (in Chinese).
- [2] 王敏,杨志坚,李扬,等. 供水管网计量分区管理技术与应用[J]. 城镇供水,2013(2):21-25,57.
Wang Min, Yang Zhijian, Li Yang, *et al.* District metered area management technology and application[J]. City and Town Water Supply, 2013(2): 21-25, 57 (in Chinese).
- [3] 刘俊,孙佳,俞国平. 分区供水技术在漏损控制中的研究现状与进展[J]. 中国给水排水,2010,26(16):7-10,15.
Liu Jun, Sun Jia, Yu Guoping. Current situation and process of water distribution system zoning technologies for water loss control[J]. China Water & Wastewater, 2010, 26(16): 7-10, 15 (in Chinese).
- [4] 徐强,陈求稳,刘锐平,等. 基于管网水力模型的独立计量分区优化[J]. 给水排水,2008,34(3):118-120.
Xu Qiang, Chen Qiwen, Liu Ruiping, *et al.* Optimization of discrete metering area scheme by pipeline network model[J]. Water & Wastewater Engineering, 2008, 34(3): 118-120 (in Chinese).
- [5] 徐强,焦静,赵顺萍,等. 供水管网漏损评价指标对比与改进[J]. 中国给水排水,2016,32(20):14-18.
Xu Qiang, Jiao Jing, Zhao Shunping, *et al.* Comparison and improvement of assessment indexes of water loss for water distribution systems [J]. China Water & Wastewater, 2016, 32(20): 14-18 (in Chinese).
- [6] 代焕芳,刘书明. 国内外供水管网漏损评价指标初探[J]. 给水排水,2016,42(S1):258-261.
Dai Huanfang, Liu Shuming. Primary investigation on performance indicators for leakage control in water distribution systems [J]. Water & Wastewater Engineering, 2016, 42(S1): 258-261 (in Chinese).
- [7] 张维明,马名楠. 供水行业信息化管理调研报告及分析[J]. 中国给水排水,2016,32(24):54-58.
Zhang Weiming, Ma Mingnan. Investigation report and analysis on information management of water industry[J]. China Water & Wastewater, 2016, 32(24): 54-58 (in Chinese).
- [8] 信昆仑,钱昊,陶涛. 移动互联及大数据技术在供水行业管理中的应用分析[J]. 给水排水,2014,40(9):162-164.
Xin Kunlun, Qian Hao, Tao Tao. Application study of mobile internet and big data technologies in water supply industry management [J]. Water & Wastewater Engineering, 2014, 40(9): 162-164 (in Chinese).
- [9] Xu Qiang, Liu Kuo, Zhao Shunping, *et al.* Optimal control for water loss from water distribution network: A case study of Beijing [A]. Proceedings of Conference on Computing and Control for the Water Industry [C]. Sheffield, UK: CCWI, 2017.



作者简介:马金锋(1978—),男,宁夏吴忠人,博士,助理研究员,研究方向为水环境数值模拟。

E-mail: jfma@rcees.ac.cn

收稿日期:2018-08-12