智能配用电大数据存储架构设计

葛磊蛟[1],王守相[1],瞿海妮[2]
(1. 天津大学 智能电网教育部重点实验室,天津 300072;2. 国网上海市电力公司,上海 200122)

摘要:针对智能配用电数据海量、高维度、多源异构等特点,提出一种大数据存储的三层管理框架设计方案。首先,对智能配用电数据组成进行分类划分。然后,提出智能配用电数据存储的资源层、元数据层和实际数据层的三层管理框架,在资源管理层,应用虚拟化技术、负载均衡和资源调度完成对物理资源的管理;在元数据管理层,使用数据预处理技术对智能配用电的元数据进行分类,采用中间件技术完成XML元数据仓库管理;在实际数据管理层,利用NoSQL技术,通过Map和Reduce的有效结合,完成海量数据的分层分区快速存储。最后,在实验室环境下对该设计方案进行初步的应用,验证了所设计方案能够对结构化、半结构化和非结构化数据进行统一存储,可为智能配用电大数据技术的应用提供基础支撑作用。

关键词:大数据;海量数据;框架设计;智能配用电;数据处理;存储
中图分类号:TM 721 **文献标识码**:A DOI:10.16081/j.issn.1006-6047.2016.06.029

0 引言

大数据技术正成为各行各业的应用热点,智能配用电(SPDU)作为实现电网资源优化配置、电能合理利用、节能降耗和能效提升的重要手段,不仅含有量大面广的电力设备和应用系统,还包括支撑这些设备和系统正常运行的多种途径、多类型输入输出海量数据,并且这些数据容量大、采集频率高、来源渠道(电网企业、电力用户、社会资源等)多样、类型复杂而价值密度低。

配用电数据之所以称为大数据,是因为其符合大数据的特征,一是数据来源广,即多源,既包括多达十几个的配用电数据采集系统,也包括气象数据和经济社会数据采集;二是异构,即为多种类型的数据,既有文本数据,也有多媒体数据,既有时间序列数据,也有经济社会指标数据;三是数据量大,仅上海市某个区的配用电数据年存储量就达到几TByte。配用电大数据的多源和异构的特性给数据集成和存储处理带来技术挑战,整合存储配用电大数据的目的和用途是:实现配用电结构化、非结构化数据的集成,满足大数据集成和存储的高容量和高速度要求,提高数据集成和存储的效率;通过对智能配用电大数据的存储和分析,实现基于大数据的融合多源高维时序数据的用电预测、节电、网架优化、错峰调度等多种电力应用。

具体而言,智能配用电数据不仅包括来自配电自动化系统DAS(Distribution Automation System)、地理信息系统(GIS)、配电SCADA/EMS(能量管理系统)、用电信息采集系统等应用系统的配用电运行、计量计费等结构化数据,也包括电力用户、电网企业等监控平台中的Web页面等半结构化数据,还有95598等电力客户的文本、音频、图片等非结构化数据,这些结构化和非结构化混合数据在网省公司年存储量达到TByte级,甚至PByte级,从存储的经济性角度考虑需要应用大数据存储技术;此外,智能配用电涉及多应用系统、多业务主体、多用户对象等,随着采集频度大幅增强和采集类型多样化,致使数据的维度越来越大,如果依然按照以往分别采集-传输-集中存储的方式,将使数据冗余大、重复存储多、系统资源利用率低等问题更加突出[1-3]。因此,有必要对高维度、非结构化和结构化混合、点多面广、多源异构的智能配用电大数据存储技术进行研究。

近年来,国内外学者在智能电网的大数据方面进行了一些研究,在智能配用电大数据应用需求分析方面,文献[4]从配电网负荷预测、运行状态评估与预警、电能质量监测和评估、基于配电网数据融合的停电优化等多方面进行了大数据的应用阐述;文献[5]从用户用电行为和负荷预测2个应用场景,提出了大数据的研究思路和方法;文献[6]阐述了大数据在智能配电网中的应用所涉及的大数据存储与处理以及大数据解析等关键技术;文献[7]从能量数据平台(EDP)和需求侧管理系统(DROMS)的应用角度,分析了大数据技术在配电网数据分析中的应用。在智能配用电大数据的基本概念[8]、应用架构[9-10]、技术现状和特点[11-13]、关键技术[14-15]等诸多方面,国内外学者也均有研究。针对电力系统特定应用场景,在大数据存储方面也有一些学者进行了探讨,文献[16]针对输变电设备状态监测大数据,基于Hadoop云计算试验平台进行了数据分布策略、数据块尺寸

收稿日期:2016-03-07;**修回日期**:2016-04-19
基金项目:国家高技术研究发展计划(863计划)资助项目(2015AA05203)
Project supported by the National High Technology Research and Development Program of China (863 Program) (2015AA05203)

调优、集群网络拓扑规划 3 个方面的存储优化研究;文献[17]面向用电信息采集系统大数据,提出了 Hadoop 集群的大数据处理架构和计算服务架构。

本文针对智能配用电数据具有海量、高维度、结构化与非结构化混杂、多源异构等特点,从资源管理、元数据管理和实际数据管理 3 个方面阐述了智能配用电数据存储技术框架设计方案,为配用电大数据的应用奠定良好的基础。

1　智能配用电大数据组成

智能配用电海量数据依据数据类型主要分为结构化数据、半结构化和非结构化数据;依据数据来源可分为电网内部数据和电网外部数据,且这些数据一般以信息集成化平台的方式呈现。其中,电网内部数据主要包括配电自动化、GIS、SCADA/EMS、用电信息采集系统、营销管理系统以及 95598 等电网数据;电网外部数据主要包括政府、电力用户、第三方机构等 3 个方面的数据。政府数据主要包括能耗监测系统、能源公共服务平台、智慧城市监控系统等监测数据;电力用户数据主要包括分布式电源 EMS、微电网 EMS、家庭 EMS、楼宇 EMS、企业 EMS 等用户数据;第三方机构数据主要包括气象监测系统、交通监控系统、医务信息平台等平台数据。综上所述,智能配用电数据组成情况如表 1 所示。

智能配用电的数据涉及电网、用户、政府和第三方机构等多个主体,各参与主体由于所聚焦的业务重心、工作流程和关注重点不完全一致,数据呈现以下几个特点。

(1) 结构化、半结构化与非结构化共存。智能配电网相关的传统业务涉及状态估计、潮流计算、短路计算等稳态分析,主要由结构化数据支撑。随着分布式电源、微电网、电动汽车接入智能配电网,电网与用户的双向互动化,以及区域负荷预测、第三方机构的客户在线认证、客户日志信息分析等高级应用业务分析的发展,智能配用电的基础分析数据逐渐包括了越来越多的 Web 页面、文本、视频、声音等半结构化和非结构化数据。另外,电网企业、用户、政府等特定应用网站的 Web 类半结构化数据也快速增长,这就形成了结构化、半结构化和非结构化数据混存的情形。

(2) 数据维度相当大。一方面,智能配用电数据所属的多个主体之间,原则上相对独立,数据采集和存储的时序性基本无法保证一致,致使数据的维度较大;另一方面,电力企业的用电信息、营销、客服等各个业务也是相对独立运行,数据采集过程中虽然通过全球 GPRS 进行了时钟同步,但是每一个独立系统所选择的数据采集时间较难保证一致性,也使其数据维度大幅增加。

(3) 数据颗粒度混杂。

a. 不同应用系统由于自身业务特色需要,数据的颗粒度需求不同,例如:配电自动化的实时调度数据是 s 级或者 min 级;用电信息采集系统数据一般是 1 d 级;95598 客户信息数据中,一些运行稳定的用户可能是月度级或者年度级。

b. 视频、音频、文本等非结构化数据格式不同,存储空间范围和元数据的内存划分尺度不同。

c. 相同数据格式的结构化数据,不同用户的数据的容量大小和属性不同。

d. 不同数据格式和不同数据属性组成混合文

表 1 智能配用电数据组成

Table 1 Components of SPDU data

数据来源			数据类型	
			结构化数据	半结构化和非结构化数据
电网内部数据		配电自动化系统 GIS SCADA/EMS 用电信息采集系统 营销管理系统 95598	累计有功电量、累计无功电量、功率因数、负荷总量、电能质量、可靠性、电压、电流、频率等	企业管理 Web 页面数据、客户文本信息、客户报修音频、现场施工视频、设备维修图片等
电网外部数据	政府	能耗监测系统 能源公共服务平台 智慧城市监控系统	全社会用电总量、用油总量、用气总量等	企业 Web 数据、企业信息文本、能源政策 Web 页面等
	电力用户	企业 EMS 楼宇 EMS 家庭 EMS 微电网 EMS 分布式电源 EMS	日用电量、月用电量、年用电量、电价信息等	设备安防监控视频、用户运维文本、用户门户网站 Web 页面数据等
	第三方机构	气象监测系统 交通监控系统 医务信息平台	日天气预报、路灯用能情况、医院用电信息等	电动汽车充电监控视频、医务信息发布 Web 页面等

件属性。

e. 半结构化数据的来源渠道多样,时间尺度、采集精度、频率差异性较大。

智能配用电数据具有海量、数据更新速度极快、分布地域广泛等特点。当前非结构化数据管理技术、大数据存储和分析技术均处于研究阶段,数据挖掘分析技术手段还不够完善,大数据快速分析算法还不够成熟。且当前硬盘、磁带、磁盘阵列等IT信息存储物理设备在缓存容量、硬盘容量和处理器速度,以及性价比、异构兼容性等方面也正在发展之中。因此,构建一种快速有效的智能配用电数据存储技术解决方案是十分必要的。

2 智能配用电大数据存储技术框架

智能配用电数据一方面具有用户种类复杂、点多面广、类型多样、海量、难以快速发现有价值信息和规律性等典型大数据特点,另一方面具有很多内在的数据规律,符合大数据的特征信息,具备很大的挖掘空间。为此,依据数据存储所涉及的存储介质、映射地址和物理空间,将智能配用电数据按照资源管理、元数据管理和实际数据管理3个方面进行框架设计,如图1所示,下面将逐一进行详细阐述。

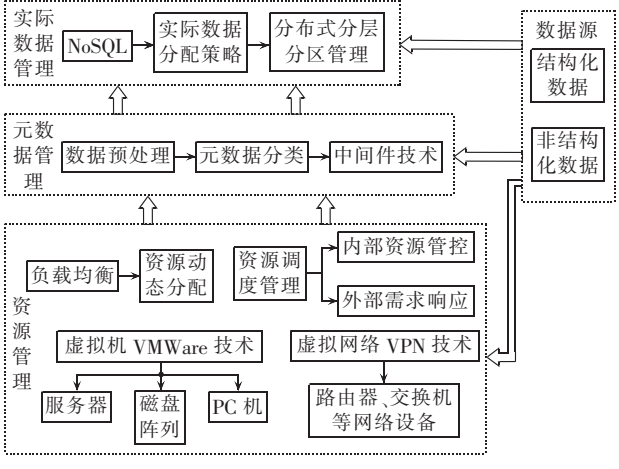

图1 智能配用电数据存储技术架构

Fig.1 Storage framework of SPDU data

其中,资源管理主要实现智能配用电大数据计算资源的虚拟化、负载均衡和资源调度管理;元数据管理利用数据预处理技术,结合元数据分类原则,通过中间件技术实现元数据(实际数据映射地址)管理;实际数据管理采用NoSQL技术,依据实际数据分配策略,实现实际数据的分层分区存储。

2.1 资源管理

智能配用电数据存储的物理资源具备多方面异构。一方面,其不仅包括硬件异构,即国-网-省-市-县多级电网企业、政府机构、电力用户等配置的大型服务器、刀片机、普通PC机等硬件资源,具有不同层次、品牌、性能配置混杂的特点,也包括软件异构,主要有Linux、Windows等不同操作系统和Oracle、SQL Service、MySQL等不同数据库,以及多参与主体根据业务所需所设计的应用平台。另一方面,智能配用电涉及的电网企业、电力用户、政府及第三方机构等多方主体均各自具有多个应用系统。在单一主体内部的多个应用系统之间不仅所选择的数据存储软、硬件平台不一致,而且存在数据重复采集和存储现象,即使近年来数据总线技术、共享内存技术等得到了较好的应用,但是数据的冗余依然较大。同时,主体与主体之间的数据存储由于涉及数据隐私、组织管理、经济能力等诸多方面,所选用的存储形式有直接附加存储DAS(Direct Attached Storage)、网络附加存储NAS(Network Attached Storage)、存储域网络SAN(Storage Area Network)等多种形式,也存在较大的兼容性问题。

为此,针对多源异构的智能配用电IT资源,资源管理的工作原理是:首先,通过虚拟化技术在智能配用电所涉及的国-省-市-县的硬件平台上构建Master/Slave集群的逻辑结构;其次,通过负载均衡完成智能配用电数据存储资源动态分配;最后通过资源调度,实现电力系统智能配用电的存储资源高效运转。

2.1.1 虚拟化技术

虚拟化技术是一种使用户不受物理资源架构、地域和形态所限的计算资源管理技术,有硬件虚拟化、虚拟内存、桌面虚拟化、服务虚拟化等多种形式[18-20]。本文主要采用虚拟机VM(Virtual Machine)和虚拟网络2种技术,即在智能配用电中,利用虚拟机技术将现有的配电自动化主站平台虚拟成为智能配用电的大数据资源层的主节点,将其他应用平台,空闲的办公、区/市/县配用电相关的前置子系统以及电网企业外网的政府、电力用户与第三方机构等IT资源虚拟化为从节点;利用虚拟网络技术将用于连接各个物理设备的路由器、交换机等网络设备虚拟化,从而使数据存储资源对使用者或者管理者保持逻辑完整性,组成一个智能配用电大数据中心集群。

(1) 虚拟机技术。

虚拟机是一种真实计算环境的抽象和模拟,主流的产品有Oracle公司的VM VirtualBox、EMC公司的VMware和微软的Virtual PC等。它通过虚拟机监视器VMM(Virtual Machine Moniter)为系统平台中的每个虚拟机设定一组数据结构(虚拟处理器的全套寄存器、物理内存的使用情况、虚拟设备的状态等),完成管理虚拟机的状态。

智能配用电中的电网企业、电力用户、政府等多主体的各个应用平台,分别在其应用系统中安装相

应的虚拟机，其部署过程如下：

a. 依据电网内、外部的数据归属权不同，预设不同的虚拟机镜像；

b. 按照安装向导，创立虚拟机；

c. 在虚拟机中，设定节点属性；

d. 依据网络划分规则，在虚拟机中设置网络参数。

(2) 网络虚拟技术。

网络虚拟技术是将不同网络的硬件和软件资源，通过虚拟化的技术手段，结合成一个加密连接的具有类似局域网功能的虚拟专用网络，形成在物理连接上分散而逻辑上连续的分区分段网络。

智能配用电的网络相关设备不仅包括电力企业所构建的光纤、以太网、无线 GPRS 网、微功率 230 MHz 无线等，还包括政府、电力用户和第三方机构所选用的移动互联网、万维网等，其网络虚拟的工作原则如下。

a. 电网企业内部按照计量、营销、检修、客服分为四大部分；电网企业外部按照政府、电力用户和第三方机构分为三大部分。

b. 电网企业内、外部之间采用网络层地址的网络虚拟化技术。

c. 电网企业内部的网络，在各个部分之间采用 IP 广播组虚拟化，在每一个部分内部采用 MAC 地址或者交换端口号。

d. 电网外部网络之间采用 IP 广播组或者网络层地址，电网外部网络的每一个部分内部采用 MAC 地址或者交换端口号。

2.1.2　负载均衡模型

经过虚拟化后，智能配用电的电网内、外部的计算资源构成了一个完整的资源池，为保障接入的每一个 IT 计算资源能够高效、可靠运行，需要进行合理优化，为此对整个资源池构建数学模型。

假设智能配用电系统有 M 个存储节点 A_1、A_2、…、A_M，一定时间内有 N 个存储任务 B_1、B_2、…、B_N，且每个存储任务至少分配 1 个存储节点，则每个节点的任务量总数为：

$$C_i=\sum_{j=1}^{N}D_{ij}\quad i=1,2,\cdots,M \tag{1}$$

其中，C_i 为存储节点 A_i 的总任务量；D_{ij} 表示任务 j 是否在节点 i 执行，表达式如式(2)所示。

$$D_{ij}=\begin{cases}1 & \text{任务 } j \text{ 在节点 } A_i \text{ 执行}\\ 0 & \text{任务 } j \text{ 不在节点 } A_i \text{ 执行}\end{cases} \tag{2}$$

同时，设定一个时间参数 t_{ij} 来表示任务 j 在节点 A_i 上的执行时间，可表示为：

$$t_{ij}=\begin{cases}t_1+t_2+t_3 & \text{任务 } j \text{ 在节点 } A_i \text{ 执行}\\ 0 & \text{任务 } j \text{ 不在节点 } A_i \text{ 执行}\end{cases} \tag{3}$$

其中，t_1 为任务处理时间；t_2 为等待队列时间；t_3 为进程阻塞时间。

因此，智能配用电资源池的负载均衡模型为一个优化问题，即在最短的处理时间内，实现所有的存储任务，优化模型如式(4)所示。

$$\begin{cases}\min \sum_{i=1}^{M}\sum_{j=1}^{N}C_iD_{ij}t_{ij}\\ \text{s.t.}\quad \sum_{i=1}^{M}C_i=N\\ \qquad t_{ij}\geqslant 0\end{cases} \tag{4}$$

其中，D_{ij}、t_{ij} 为离散变量。

不难发现，智能配用电负载均衡模型是一个非线性的整数规划模型，在实际的运行过程中，依据实际的任务量、节点数等具体数据信息，采用遗传算法、蚂蚁算法、BP 神经网络算法等进行求解，从而获得最优解，实现资源池的优化利用。

2.1.3　资源调度

智能配用电的 IT 资源经过虚拟化和负载均衡优化，已达成了资源的基本分配，但要实现智能配用电存储资源的高效运转，资源调度非常必要，依据实际的功能需求，设计其组成模块，如图 2 所示。

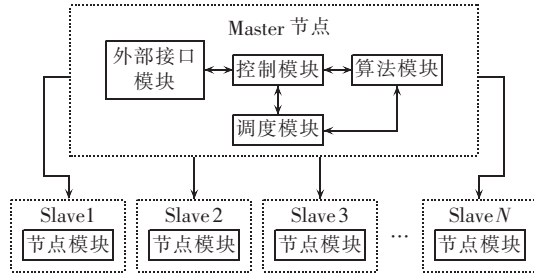

图 2 资源调度管理的功能模块

Fig.2 Functional module of resource scheduling management

图 2 所示模块的工作原理是：控制模块通过实时监控系统状态，并负责整个资源池系统的调度管理，它不仅与外部接口模块进行信息交互，也与算法模块进行实时的数据互动，同时根据任务需求对调度模块进行管理；调度模块主要是响应控制模块的指令，并下发给节点模块，同时将相关的变动信息实时互动给算法模块；算法模块负责负载均衡算法的在线求解；外部接口模块负责与资源池外部的信息交换；节点模块负责实际的调度指令执行。

依据上述的资源调度原则，充分结合智能配用电的应用特征，在充分利用网省公司现有的配电自动化主站、GIS 主站、用电信息采集系统主站等数据中心软、硬件 IT 资源的前提条件下，以现有的配电自动化主站平台构建智能配用电的大数据资源管理的主节点；其他网省公司应用平台均作为从节点，并将空闲的办公、区/市/县配用电相关的前置子系统，以及电网企业外网的政府与第三方机构、电力用户等的 IT 资源，也作为计算层的从节点，利用电力业

务所构建的光纤、以太网、无线 GPRS 网、微功率 230 MHz 无线等网络作为连接纽带,与用户、第三方机构进行友好互联互通,从而组成智能配用电大数据资源管理中心。

资源管理中心利用静态和动态相结合的优化调度策略,如图 3 所示:电网企业内网采用静态优化调度策略,即带权重的轮循算法,图 3 中小圆圈内的数字代表每一个 Slave 节点所占的缺省权重系数,依次循环利用内网的 IT 资源;电网企业外网采用动态优化调度策略,即最快响应速度算法,图 3 中椭圆内的时间代表每一个 Slave 节点所缺省的响应时间,根据时间的长短,依次利用外网的 IT 资源。

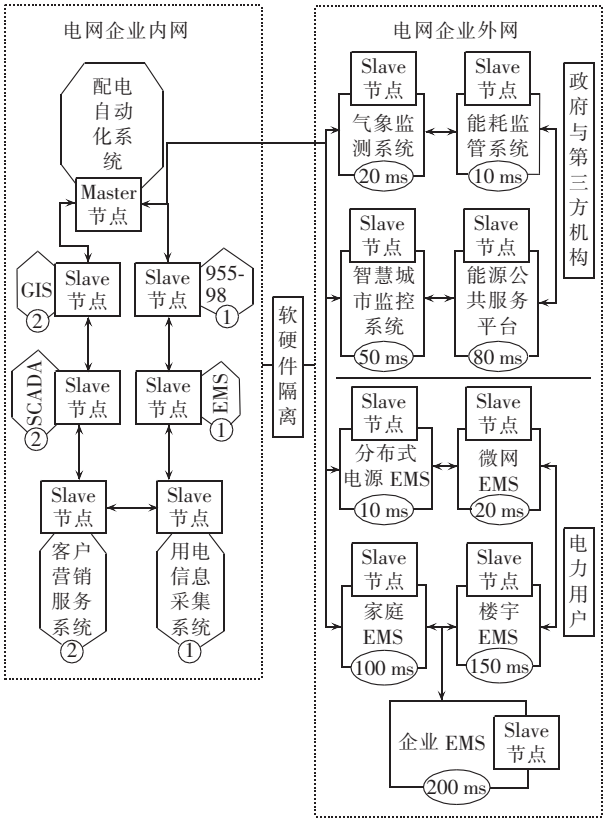

图 3 智能配用电数据存储资源管理优化调度策略
Fig.3 Optimal scheduling strategy for data storage and resource management of SPDU

2.2 元数据管理

多样性的海量智能配用电数据存储之前,如何根据实际数据大小情况,快速判断、辨析、分配指定的存储空间,形成有效的地址映射(元数据)是数据存储的第一步,因此,形成标准化的系统可识别的元数据,并在此基础上,进行元数据管理是大数据存储至关重要的一步。

元数据管理主要采用数据预处理和数据中间件(简称中间件)2 种技术。其中,数据预处理主要负责结构化和非结构化数据的辨析[21-25],生成 XML(eXtensible Markup Language)格式的元数据;中间件主要完成元数据仓库管理。

2.2.1 数据预处理

智能配用电数据预处理策略是:对结构化数据和非结构化数据进行数据筛选、数据变换和数据标准化,最终实现将结构化和非结构化数据的元数据以 XML 格式存入 Master 节点中,从而完成数据预处理工作。其过程如图 4 所示。

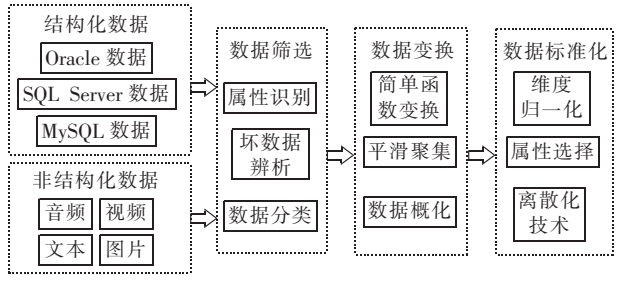

图 4 智能配用电数据预处理
Fig.4 Preprocessing of SPDU data

a. 数据筛选。数据筛选是对实际数据进行初步遴选,包括属性识别、坏数据辨析和数据分类三部分。

b. 数据变换。数据变换是根据数据筛选的结论,生成 XML 格式初始元数据的过程。一般根据不同数据决定选用不同的数据变换方法,主要包括简单函数变换、平滑聚集和数据概化。

c. 数据归一化。数据归一化是将变换后的元数据集合以标准化的统一 XML 格式表示,主要的技术有维度归一化、属性选择和离散化技术等。

智能配用电数据通过数据预处理后,结构化和非结构化数据的元数据形成统一标准 XML 格式,为有效识别元数据类型,以及创建元数据仓库提供基础准备,依据数据所属电网内、外部原则分别进行命名分类:电网内部的元数据,按照电网电压等级,分为 110 kV 电压等级元数据、35 kV 电压等级元数据、10 kV 电压等级元数据、0.4 kV 电压等级元数据 4 个部分;电网外部的元数据,分为政府元数据、电力用户元数据和第三方机构元数据 3 个部分。

数据进行初步的分类完毕后,依据分类的规则分别提交给相对应的中间件,从而形成 XML 元数据仓库,有利于发现相同或者相近元数据,方便进行合并、组合或者重新排序等,达成对元数据的降维存储。

2.2.2 中间件技术

中间件技术是一种在不同技术之间共享资源的系统软件或服务程序的统称,有终端仿真、数据访问、消息等多种不同的应用形式。本文采用的中间件是数据访问中间件技术,主要负责元数据的关联、整合、协同、互动和按需服务等,实现 XML 元数据仓库的管理,可分为配置、关联、开关、读、写、查询、删除和迁移等多个子模块。中间件数据处理流程见图 5。

中间件的工作原理如下:

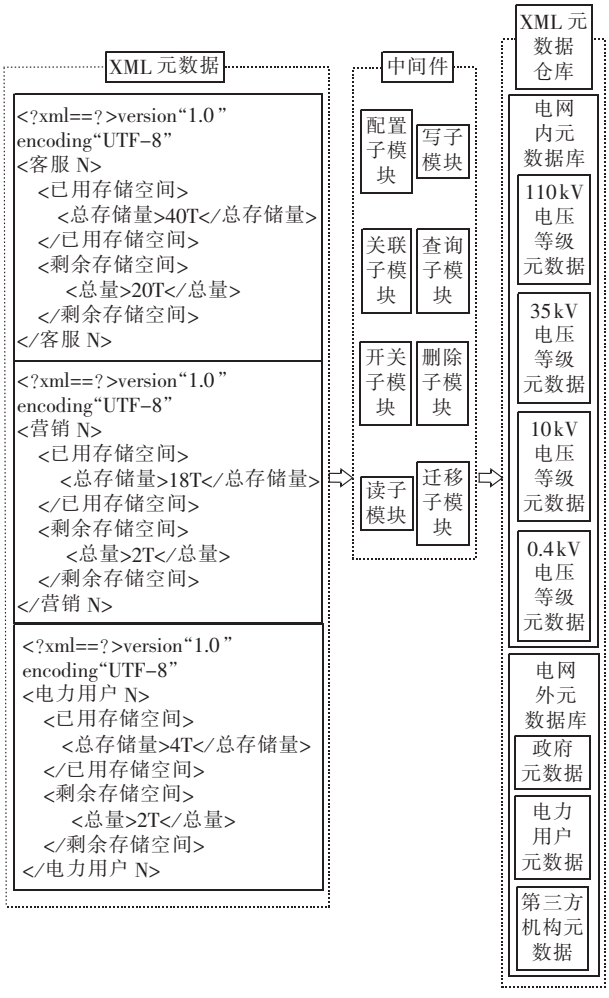

图 5 中间件数据处理流程

Fig.5 Flowchart of middleware data processing

a. 将经过数据预处理所形成的 XML 文件提交至中间件后,依据分类原则先对 XML 文件进行形式识别,从而确定 XML 文件所属元数据仓库的所属分类;

b. 根据所属分类,中间件从元数据仓库查询相关元数据文件,然后反馈与该元数据文件相关的具体信息;

c. 中间件根据反馈信息,对 XML 元数据文件进行删除、增加、迁移等处理,并依据时序降序或者升序排列,对相近或相关的元数据进行合并、组合等降维,直至形成数据仓库。

2.3 实际数据管理

高维度、多源异构的智能配用电实际数据,在元数据(实际数据存储的逻辑地址)形成的基础上,依据其映射关系,采用 NoSQL 技术对实际数据进行分布式存储管理。

2.3.1 NoSQL

NoSQL 是一种非关系型、分布式、不提供 ACID 的数据库设计模式[26-28]。针对关系型数据库在处理密集型数据时所面临的灵活性差、扩展性差、性能差等实际问题,NoSQL 从简化、高效的目标出发,采用以下核心技术。

a. 简单数据模型。NoSQL 采用简单数据模型,即一条数据记录对应唯一的键值,不支持外键和跨记录的关系。

b. 元数据和应用数据的分离。NoSQL 数据管理系统包括元数据和实际数据 2 种数据。其中,元数据是用于系统检索、管理的数据;实际数据是用户的原始数据。

c. 弱一致性。NoSQL 通过对实际数据保存多个副本来保持数据的一致性,且弱一致性模型常采用最终一致性和时间轴一致性技术。

NoSQL 数据库的工作原理是:一方面,运行于 Master 节点的 NameNode 进程对 2.2 节中所形成的智能配用电元数据仓库进行操作;运行于 Slave 节点上的 DataNode 进程将智能配用电的任意一个大文件按照缺省的 64 MByte 大小数据块进行分割,并存储在分区分层的多个不同数据 Slave 节点上。另一方面,按照 Map/Reduce 的映射关系进行数据分散和合并,从而完成数据存储。NoSQL 的数据处理工作流程如图 6 所示。

2.3.2 分层分区管理

图 7 为 NoSQL 对智能配用电实际数据的分层分区原则,具体描述如下:

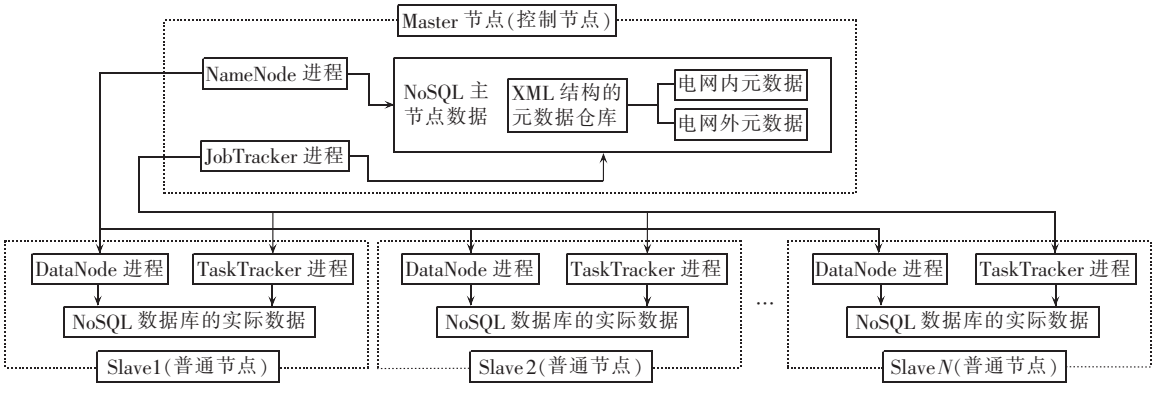

图 6 NoSQL 的数据处理工作流程

Fig.6 Flowchart of NoSQL data processing

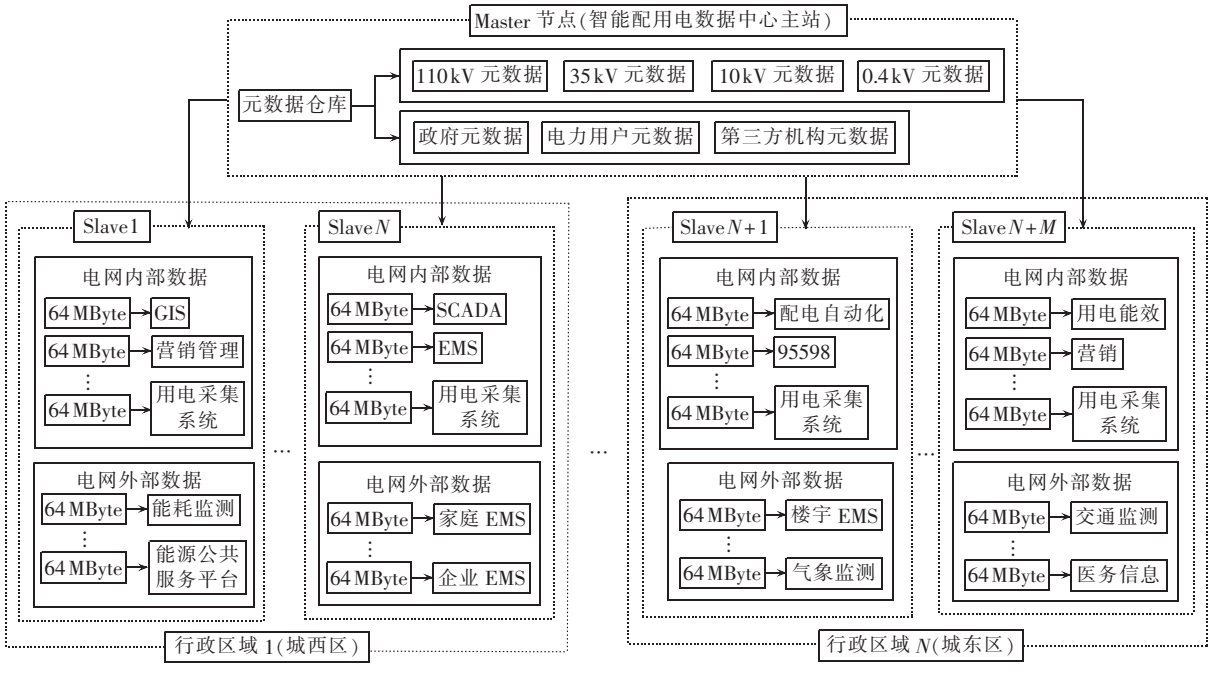

图 7 数据分层分区处理原则

Fig.7 Principle of data processing in layers and divisions

a. Master 节点上 NameNode 进程负责对元数据仓库存储管理;

b. 依据智能配用电的地域行政进行分区,例如行政区域 1、…、行政区域 N;

c. 任意一个行政区域内的计算资源从 Slave 1 依次进行排序标号,存储实际数据;

d. 任意一个 Slave 节点依据电网内、电网外进行数据分层,然后按照 64 MByte 的存储空间进行分片,进行顺序存储。

智能配用电实际数据存储策略如下。

a. 电网企业内、外网的实际数据完全分层存储,即电网企业内部数据仅仅存储于电网内,电网企业外部数据存储在相应的电网外。

b. 将配电自动化、GIS、SCADA/EMS、用电信息采集系统、客户营销服务系统以及 95598 等内网数据按照电压等级属性进行分类,即电网内部数据分为 110 kV、35 kV、10 kV、0.4 kV 数据分别进行分类存储;电网企业外部数据依据政府、第三方机构、电力用户等不同对象分别进行存储。

c. 最小路径分配存储策略,即以 Master 节点为出发点,先从元数据仓库中查询到所属类别的对应 XML 数据表;然后从元数据 XML 表中分配足够的存储空间给实际数据,若同类元数据 XML 表中剩余存储空间不足时,以相邻最近为原则以续存方式给实际数据分配所缺额的存储空间;最后依据分配指令对实际数据进行存储操作。

依据本文所提的存储框架设计方案,在实验室 3 台 Lenovo 的普通 PC 机平台上,进行了初步实验,硬件平台的配置如表 2 所示。

表 2 硬件平台的参数

Table 2 Parameters of hardware platform

电脑型号	CPU 型号	内存/GByte	硬盘/GByte	已投入使用年限/a
联想 H3050	I3-4160	4	500	2
扬天 T4900C	I5-4590	4	1 000	1
联想 G5050	I5-4460	4	1 000	3

在以上的不同配置、不同使用年限的异构硬件平台上,先用 VMware 虚拟机进行虚拟化,并应用 C++ 语言开发的调度管理程序进行资源化管理;然后,应用商用的甲骨文 Fusion Middleware 中间件软件,对用电信息采集系统结构化数据、智能微电网的 Web 网页等半结构化数据、实验室智能家居的图片等非结构化数据进行元数据仓库管理;最后应用商用 NoSQL 数据库,将以上实际混合数据存储于这 3 台 PC 机中,完成了数据的基本存储。但随着上海 863 科技项目智能配用电大数据的充实,框架方案有待进一步地完善研究。

3 结论与展望

智能配用电数据,具有海量、高维度、存储和查询维护难等特点,在存储架构方面,本文按照资源管理、元数据管理和实际数据管理的思路在一定程度上解决了海量智能配用电大数据的存储问题,本文设计方案从先进理念上对智能配用电的大数据存储进行了前期框架设计,虽在实验室进行了基本存储的实现和简单验证,但受限于大数据系统构建的复杂性和长期性,还需要在未来工程的实践中加以验

证和完善。对实践中所发现的问题开展深入研究和探索,是未来工作的重点。

另外,由于智能配用电数据关系企业的用能特点、居民用户的生活习惯涉及一定的商业机密或者个人隐私,对于数据的安全性和网络共享,以及如何合理地挖掘利用这些大数据,也是需要重点关注的研究内容。

参考文献:

[1] 王守相,王成山. 现代配电系统分析[M]. 2 版. 北京:高等教育出版社,2014:142-148.

[2] 何颖鹏. 非结构化数据统一存储平台的设计与实现[D]. 杭州:浙江大学,2013.
HE Yingpeng. Design and implementation of unstructured data unified storage platform[D]. Hangzhou:Zhejiang University,2013.

[3] 余斌. 海量非结构化数据分布式分析与检索[D]. 杭州:浙江大学,2012.
YU Bin. Distributed analysis and retrieval of massive unstructured data[D]. Hangzhou:Zhejiang University,2012.

[4] 刘科研,盛万兴,张东霞,等. 智能配电网大数据应用需求和场景分析研究[J]. 中国电机工程学报,2015,35(2):287-293.
LIU Keyan,SHENG Wanxing,ZHANG Dongxia,et al. Big data application requirements and scenario analysis in smart distribution network[J]. Proceedings of the CSEE,2015,35(2):287-293.

[5] 王继业,季知祥,史梦洁,等. 智能配用电大数据需求分析和应用研究[J]. 中国电机工程学报,2015,35(8):1829-1836.
WANG Jiye,JI Zhixiang,SHI Mengjie,et al. Scenario analysis and application research on big data in smart power distribution and consumption systems[J]. Proceedings of the CSEE,2015,35(8):1829-1836.

[6] 赵腾,张焰,张东霞. 智能配电网大数据应用技术与前景分析[J]. 电网技术,2014,38(12):3305-3312.
ZHAO Teng,ZHANG Yan,ZHANG Dongxia. Application technology of big data in smart distribution grid and its prospect analysis[J]. Power System Technology,2014,38(12):3305-3312.

[7] 王璟,杨德昌,李锰,等. 配电网大数据技术分析与典型应用案例[J]. 电网技术,2015,39(11):3114-3121.
WANG Jing,YANG Dechang,LI Meng,et al. Analysis of big data technology in power distribution system and typical applications[J]. Power System Technology,2015,39(11):3114-3121.

[8] 张东霞,苗新,刘丽平,等. 智能电网大数据技术发展研究[J]. 中国电机工程学报,2015,35(1):2-12.
ZHANG Dongxia,MIAO Xin,LIU Liping,et al. Research on development strategy for smart grid big data[J]. Proceedings of the CSEE,2015,35(1):2-12.

[9] 刘道新,胡航海,张健,等. 大数据全生命周期中关键问题研究及应用[J]. 中国电机工程学报,2015,35(1):23-28.
LIU Daoxin,HU Hanghai,ZHANG Jian,et al. Research on key issues of big data lifecycle and its application[J]. Proceedings of the CSEE,2015,35(1):23-28.

[10] 宋亚奇,周国亮,朱永利. 智能电网大数据处理技术现状与挑战[J]. 电网技术,2013,37(4):927-935.
SONG Yaqi,ZHOU Guoliang,ZHU Yongli. Present status and challenges of big data processing in smart grid[J]. Power System Technology,2013,37(4):927-935.

[11] LABRINIDIS A,JAGADISH H V. Challenges and opportunities with big data[J]. Proceedings of the VLDB Endowment,2012,5(12):2032-2033.

[12] 田世明,杨增辉,时志雄,等. 智能配用电大数据关键技术研究[J]. 供用电,2015(8):12-18.
TIAN Shiming,YANG Zenghui,SHI Zhixiong,et al. Research on the key technology of big data for smart power distribution and utilization[J]. Distribution & Utilization,2015(8):12-18.

[13] AAMIR M,UQAILI M A,AMIR S,et al. Framework for analysis of power system operation in smart cities[J]. Wireless Personal Communications,2014,76(3):399-408.

[14] YU Y X,ZENG Y,LIU H,et al. Challenges and R&D opportunities of smart distribution grids in China[J]. Science China:Technological Sciences,2014,57(8):1588-1593.

[15] 彭小圣,邓迪元,程时杰,等. 面向智能电网应用的电力大数据关键技术[J]. 中国电机工程学报,2015,35(3):503-511.
PENG Xiaosheng,DENG Diyuan,CHENG Shijie,et al. Key technologies of electric power big data and its application prospects in smart grid[J]. Proceedings of the CSEE,2015,35(3):503-511.

[16] 宋亚奇,周国亮,朱永利,等. 云平台下输变电设备状态监测大数据存储优化与并行处理[J]. 中国电机工程学报,2015,35(2):255-266.
SONG Yaqi,ZHOU Guoliang,ZHU Yongli,et al. Storage optimization and parallel processing of condition monitoring big data of transmission and transforming equipment based on cloud platform[J]. Proceedings of the CSEE,2015,35(2):255-266.

[17] 王相伟,史玉良,张建林,等. 基于 Hadoop 的用电信息大数据计算服务及应用[J]. 电网技术,2015,39(11):3128-3133.
WANG Xiangwei,SHI Yuliang,ZHANG Jianlin,et al. Computation services and applications of electricity big data based on Hadoop[J]. Power System Technology,2015,39(11):3128-3133.

[18] 韩璞,袁世通. 基于大数据和双量子粒子群算法的多变量系统辨识[J]. 中国电机工程学报,2014,34(32):5779-5787.
HAN Pu,YUAN Shitong. Multivariable system identification based on double quantum particle swarm optimization and big data[J]. Proceedings of the CSEE,2014,34(32):5779-5787.

[19] 张建伟,潘秀琴. 基于量子优化的云服务器负载均衡算法研究[J]. 计算机应用研究,2015,32(10):3128-3130.
ZHANG Jianwei,PAN Xiuqin. Cloud server load balancing algorithm based on quantum optimization[J]. Computer Application Research,2015,32(10):3128-3130.

[20] 曲朝阳,陈帅,杨帆,等. 基于云计算技术的电力大数据预处理属性简约方法[J]. 电力系统自动化,2014,38(8):67-71.
QU Zhaoyang,CHEN Shuai,Yang Fan,et al. An attribute reducing method for electric power big data preprocessing based on cloud computing technology[J]. Automation of Electric Power Systems,2014,38(8):67-71.

[21] 菅志刚,金旭. 数据挖掘中数据预处理的研究与实现[J]. 计算机应用研究,2004,21(7):117-118,157.
JIAN Zhigang,JIN Xu. Research on data preprocess in data mining and its application[J]. Computer Application Research,2004,21(7):117-118,157.

[22] 欧阳柳波,李学勇,杨贯中,等. 基于近似匹配模型的 XML 元数

据检索[J]. 计算机应用,2005,25(4):820-823,826.
OUYANG Liubo,LI Xueyong,YANG Guanzhong,et al. XML metadata retrieval based on approximately matching model[J]. Computer Application,2005,25(4):820-823,826.
[23] BOU-HARB E,FACHKHA C,POURZANDI M,et al. Communication security for smart grid distribution networks[J]. IEEE Communications Magazine,2013,51(1):41-49.
[24] SATHYANARAYANA B R,HEYDT G T. Sensitivity-based pricing and optimal storage utilization in distribution systems[J]. IEEE Transactions on Power Delivery,2013,28(2):1073-1082.
[25] 李滨,杜孟远,韦维,等. 基于准实时数据的智能配电网理论线损计算[J]. 电力自动化设备,2014,34(11):122-128,148.
LI Bin,DU Mengyuan,WEI Wei,et al. Calculation of theoretical line loss based on quasi real-time data of smart distribution network[J]. Electric Power Automation Equipment,2014,34(11):122-128,148.
[26] 徐青山,王文帝,林章岁,等. 面向行业大数据特征挖掘的电力经理指数指标体系的建立与应用[J]. 电力自动化设备,2015,35(7):15-21.
XU Qingshan,WANG Wendi,LIN Zhangsui,et al. Establishment and application of EMI indicator system orienting to massive industrial data mining[J]. Electric Power Automation Equipment,2015,35(7):15-21.
[27] 王德文,肖磊,肖凯. 智能变电站海量在线监测数据处理方法[J]. 电力自动化设备,2013,33(8):142-146,156.
WANG Dewen,XIAO Lei,XIAO Kai. Processing of massive online monitoring data in smart substation[J]. Electric Power Automation Equipment,2013,33(8):142-146,156.
[28] 宋亚奇,刘叔仁,朱永利,等. 电力设备状态高速采样数据的云存储技术研究[J]. 电力自动化设备,2013,33(10):150-156.
SONG Yaqi,LIU Shuren,ZHU Yongli,et al. Cloud storage of power equipment state data sa mpled with high speed[J]. Electric Power Automation Equipment,2013,33(10):150-156.

作者简介:


葛磊蛟

葛磊蛟(1984—),男,湖北咸宁人,博士研究生,研究方向为智能配用电和不确定性仿真(**E-mail**:legendglj99@tju.edu.cn);

王守相(1973—),男,山东潍坊人,教授,博士研究生导师,博士,研究方向为智能配用电、分布式能源(**E-mail**:sxwang@tju.edu.cn);

瞿海妮(1981—),女,湖北宜昌人,高级工程师,博士,研究方向为智能电网技术、数据挖掘与人工智能(**E-mail**:239255951@qq.com)。

Design of storage framework for big data of SPDU

GE Leijiao[1],WANG Shouxiang[1],QU Haini[2]
(1. Key Laboratory of Smart Grid of Ministry of Education,Tianjin University,Tianjin 300072,China;
2. State Grid Shanghai Electric Power Company,Shanghai 200122,China)

Abstract: A three-layer management framework is designed for the storage of massive,highly dimensional and heterogeneous data of SPDU(Smart Power Distribution and Utilization). The data components of SPDU are classified and a hierarchical management framework with resource,metadata and actual-data layers is proposed for the storage of SPDU data. The virtualization technology,load balancing and resource scheduling are applied in the resource management layer to realize the management of physical resources. The data preprocessing and middleware technologies are applied in the metadata management layer to respectively classify the metadata of SPDU and manage the warehouse of XML metadata. The NoSQL technology is applied in the actual data management layer to fast store the massive data into different layers and divisions through the effective combination of Map and Reduce. The proposed design scheme is preliminarily applied in the laboratory environment to verify its unified storage of structuralized,semi-structuralized and non-structuralized data,which provides a basic support for the application of big data technology to SPDU.

Key words: big data; massive data; framework design; smart power distribution and utilization; data processing; storage