

智能电网大数据处理技术现状与挑战

宋亚奇¹, 周国亮¹, 朱永利²

(1. 华北电力大学 控制与计算机学院, 河北省 保定市 071003;

2. 新能源电力系统国家重点实验室(华北电力大学), 北京市 昌平区 102206)

Present Status and Challenges of Big Data Processing in Smart Grid

SONG Yaqi¹, ZHOU Guoliang¹, ZHU Yongli²

(1. Department of Computer Science, North China Electric Power University, Baoding 071003, Hebei Province, China;

2. State Key Laboratory of Alternate Electrical Power System With Renewable Energy Sources
(North China Electric Power University), Changping District, Beijing 102206, China)

ABSTRACT: Smart grid operation needs panoramic state data, and during the operation, maintenance and management of smart grid massive heterogeneous and multi-state data, namely the big data, are generated. At present, how to store the big data efficiently, reliably and cheaply and access and analyze them rapidly are important research topics. Firstly, the source of the big data generated in various process of smart grid, such as power generation, transmission, transformation and power utilization, and the features of the big data are analyzed; secondly, the existing big data processing techniques adopted in the fields of business, Internet and industrial monitoring are summarized, and the advantages and disadvantages of these techniques in coping with the construction of smart grid and big data processing are analyzed in detail; finally, in aspects of big data storage, real-time data processing, fusion of heterogeneous multi-data sources and visualization of big data, the chance and challenge brought by smart grid big data are expounded.

KEY WORDS: smart grid; big data; cloud computing; parallel database

摘要: 智能电网需要全景的状态数据。电网运行、检修和管理过程中会产生海量异构、多态的数据, 也即大数据。如何对它们进行高效、可靠、低廉地存储, 并快速访问和分析, 是当前重要的研究课题。文章首先分析了发电、输变电以及用电各个环节中大数据的产生来源和特点; 其后, 综述了目前在商业、互联网和工业监测领域已有的大数据处理技术, 并详细分析这些技术在应对智能电网建设和大数据处理方面的优势和不足。最后, 从大数据存储、实时数据处理、异构多数据源融合以及大数据可视化 4 个方面论述了智能电网大数据带来的机遇和挑战。

关键词: 智能电网; 大数据; 云计算; 并行数据库

基金项目: 国家自然科学基金项目(61074078)。

Project Supported by National Natural Science Foundation of China (61074078).

0 引言

近年来, 随着全球能源问题日益严峻, 世界各国都开展了智能电网的研究工作^[1]。智能电网的最终目标是建设成为覆盖电力系统整个生产过程, 包括发电、输电、变电、配电、用电及调度等多个环节的全景实时系统^[2]。而支撑智能电网安全、自愈、绿色、坚强及可靠运行的基础是电网全景实时数据采集、传输和存储, 以及累积的海量多源数据快速分析。因而随着智能电网建设的不断深入和推进, 电网运行和设备检/监测产生的数据量呈指数级增长, 逐渐构成了当今信息学界所关注的大数据, 这需要相应的存储和快速处理技术作为支撑。

由于云计算平台的广泛应用, 积累了海量、多源异构数据, 这急需人们研究这种大数据的分析技术和理论。目前, 大数据已成为学术界和产业界共同关注的研究主题^[3], 在很多领域获得了应用, 具有广阔的应用前景。仅 2009 年, 谷歌公司通过大数据业务对美国经济的贡献就为 540 亿 USD, 而这只是大数据所蕴含的巨大经济效益的冰山一角^[4]。淘宝公司通过对大量交易数据的变化分析, 可以提前 6 个月预测全球经济发展趋势。IBM 公司利用多达 4 PB 的气候、环境历史数据, 设计风机选址模型, 确定风机安装的最佳位置, 从而提高风机生产效率和延长使用寿命^[5]。

2011 年 5 月, 麦肯锡公司发布了关于大数据的调研报告《大数据: 下一个前沿, 竞争力、创新力和生产力》^[6], 文中充分阐明了大数据研究的地位以及将会给社会带来的价值, 大数据研究已成为社会发展和技术进步的迫切需要。

在智能电网系统中,大数据产生于整个系统的各个环节。比如在用侧,随着大量智能电表及智能终端的安装部署,电力公司和用户之间的交互行为迅猛增长,电力公司可以每隔一段时间获取用户的用电信息,从而收集了比以往粒度更细的海量电力消费数据,构成智能电网中用户侧大数据^[7]。通过对数据进行分析可以更好地理解电力客户的用电行为^[8]、合理地设计电力需求响应系统^[9]和短期负荷预测系统^[10]等。

鉴于大数据在电网中出现的场合越来越多,有必要对目前的应用现状和将来的挑战进行总结,为大数据技术在智能电网建设中的应用提供有益的参考。本文试图对智能电网中大数据的研究和应用现状及挑战进行综述,并给出智能电网大数据存储与分析系统的一种可选的框架。

1 智能电网大数据及其特点

1.1 智能电网中的大数据

电网业务数据大致分为3类:一是电网运行和设备检测或监测数据;二是电力企业营销数据,如交易电价、售电量、用电客户等方面的数据;三是电力企业管理数据。

根据数据的内在结构,这些数据可以进一步细分为结构化数据和非结构化数据。结构化数据主要包括存储在关系数据库中的数据,目前电力系统中的大部分数据是这种形式,随着信息技术的发展,这部分数据增长很快。相对于结构化数据而言,不方便用数据库二维逻辑表来表现的数据即称为非结构化数据,主要包括视频监控、图形图像处理等产生的数据。这部分数据增长非常迅速,互联网数据中心(Internet data center, IDC)的一项调查报告指出:企业中80%的数据都是非结构化数据,这些数据每年都按指数增长60%^[11]。在电力系统中,非结构化数据占到了智能电网数据的很大比重。

结构化数据根据处理时限要求又可以划分为实时数据和准实时数据,比如电网调度、控制需要的数据是实时数据,需要快速而准确地处理;而大量的状态监测数据对实时性要求相对较低,可以作为准实时数据处理。

智能电网与传统电网存在很大的不同,具有更高的智能化水平,而实现智能化的前提是大量实时状态数据的获取,目前智能电网中的大数据主要是因为以下几个方面:

1) 为了准确实时获取设备的运行状态信息,

采集点越来越多,常规的调度自动化系统含数十万个采集点,配用电、数据中心将达到百万甚至千万级^[12]。需要监测的设备数量巨大,每个设备都装有若干传感器,监测装置通过适当的通信通道把这些传感器连接在一起,由变电站的数据收集服务器按照统一的通信标准上传到数据中心,这实际上构成了一个物联网。而物联网的后端采用云计算平台已被认为是未来的发展趋势。智能电网设备物联网同云计算平台的基础设施层互联,进行数据交换。

2) 为了捕获各种状态信息,满足上层应用系统的需求,设备的采样频率越来越高。比如在输变电设备状态监测系统中,为了能对绝缘放电等状态进行诊断,信号的采样频率必须在200 kHz以上,特高频检测需要GHz的采样率。这样,对于一个智能电网设备监测平台来说,需存储的监测或检测的数据量十分庞大。

3) 为真实完整记录生产运行的每个细节,完整反映生产运行过程,要求达到“实时变化采样”^[13]。

在智能电网中,大数据产生于电力系统的各个环节,包括:

1) 发电侧。随着大型发电厂数字化建设的发展^[14],海量的过程数据被保存下来。这些数据中蕴藏着丰富的信息,对于分析生产运行状态、提供控制和优化策略、故障诊断以及知识发现和数据挖掘具有重要意义^[15]。基于数据驱动的故障诊断方法被提出^[16],利用海量的过程数据,解决以前基于分析的模型方法和基于定性经验知识的监控方法所不能解决的生产过程和设备的故障诊断、优化配置和评价的问题。另外,为及时准确掌握分布式电源的设备及运行状态,需要对大量的分布式能源进行实时监测和控制^[17]。为支持风机选址优化,所采集的用于建模的天气数据每天以80%的速度增长^[5]。

2) 输变电侧。2006年美国能源部和联邦能源委员会建议安装同步相量监测系统(synchrophasor-based transmission monitoring systems)。目前,美国的100个相位测量装置(phasor measurement unit, PMU)一天收集62亿个数据点,数据量约为60 GB,而如果监测装置增加到1000套,每天采集的数据点为415亿个,数据量达到402 GB^[18]。相量监测只是智能电网监控的一小部分。

3) 用电侧。为准确获取用户的用电数据,电力公司部署了大量的具有双向通信能力的智能电表,这些电表可以每隔5 min的频率向电网发送实

时用电信息。美国太平洋天然气电力公司(Pacific Gas & Electric)每个月从 900 万个智能电表收集超过 3 TB 的数据^[19]。电动汽车的无序充放电行为会对电网运行带来麻烦,如果能合理安排电动汽车的充放电时间,则会对电网带来好处,变害为利,而前提是对基数很大的电动机车电池的充放电状态进行监测,也会产生大数据。

1.2 智能电网中大数据的特点

智能电网中的大数据具备“4V”特征,即规模大(volume)、类型多(variety)、价值密度低(value)和变化快(velocity)。

1) 数据体量巨大。从 TB 级别,跃升到 PB 级别。常规 SCADA 系统 10000 个遥测点,按采样间隔 3~4 s 计算,每年产生 1.03 TB 数据($1.03 \text{ TB} = 12 \text{ 字节/帧} \times 0.3 \text{ 帧/s} \times 10000 \text{ 遥测点} \times 86400 \text{ s/天} \times 365 \text{ 天}$);广域相量测量系统(WAMS)10 000 个遥测点,采样率可以达到 100 次/s,按上述公式计算,则每年产生 495 TB 的数据^[13]。

2) 数据类型繁多。电网数据广域分布、种类众多,包括实时数据、历史数据、文本数据、多媒体数据、时间序列数据等各类结构化、半结构化数据以及非结构化数据,各类数据查询与处理的频度和性能要求也不尽相同。比如,电力设备状态监测数据中的油色谱数据 0.5 h 采样一次,而绝缘放电数据的采样速率高达几百 kHz,甚至 GHz。

3) 价值密度低。以视频为例,连续不间断监控过程中,可能有用的数据仅仅有 1~2 s。在输变电设备状态监测中存在同样问题,所采集的绝大部分数据都是正常数据,只有极少量的异常数据,而异常数据是状态检修的最重要依据。

4) 处理速度快。在几分之一秒内对大量数据进行分析,以支持决策制定。对在线状态数据的处理性能要求远高于离线数据。这种在线的流数据分析与挖掘同传统数据挖掘技术有本质的不同^[20]。

另外,智能电网中的数据处理,对数据质量有一定的要求,可以考虑为各类智能电网数据引入一个新的属性:数据的真实性。数据的真实性是指与特定类型数据相关的可靠性级别^[5]。高质量数据对于数据分析结果的正确性有重要影响。然而即使最好的数据清洗方法也无法去除某些数据固有的不可预测性。承认不确定性需求,并将数据的真实性作为智能电网大数据的一个维度是可行的。

智能电网中汹涌而来的大数据,为智能电网建设

带来了新的挑战和机遇。国网信通公司成立了大数据团队应对智能电网建设中的大数据挑战问题^[21]。IBM 收集并建模大数据,服务于智能电表分析、基于决策的运维、基于天气数据的风机选址、分配负荷预测与调度等各类能源行业与公用事业^[5]。

2 大数据处理技术

2.1 大数据处理的价值和复杂性

近年来,大数据已经成为科技界和产业界共同关注的热点。2012 年 3 月,美国政府宣布投资 2 亿 USD 启动“大数据研究和发展计划”。美国政府认为大数据是“未来的新石油”,将“大数据研究”上升为国家意志,对未来的科技与经济发展必将带来深远影响。一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分,对数据的占有和控制也将成为国家间和企业间新的争夺焦点^[3]。

目前全球数据的存储和处理能力已远落后于数据的增长幅度。例如,淘宝网每日新增的交易数据达 10 TB;eBay 分析平台日处理数据量高达 100 PB,超过了美国纳斯达克交易所全天的数据处理量;沃尔玛是最早利用大数据分析并因此受益的企业之一,曾创造了“啤酒与尿布”的经典商业案例。现在沃尔玛每小时处理 100 万件交易,将有大约 2.5 PB 的数据存入数据库,此数据量是美国国会图书馆的 167 倍;微软花了 20 a,耗费数百万美元完成的 Office 拼写检查功能,谷歌公司则利用大数据统计分析直接实现。

与大数据在商业及互联网领域的广泛研究和应用相比,大数据在智能电网建设的研究中还有待进一步加强。由于云计算平台具有存储量大、廉价、可靠性高、可扩展性强等优势,但在实时性方面难以保证,故它不适合于作为电网调度自动化系统的主系统,但可用于调度自动化系统的后台,也可用于智能电网数据中心(营销、管理和设备状态监测)。云平台环境下的通用大数据处理和展现工具正在不断涌现,为减少软件开发工作带来了好处。然而,数据挖掘通常是与具体应用对象相关的,大数据挖掘是一个不小的挑战。如故障录波数据初次筛选^[22]等一些基于聚类方法的应用,在面对海量数据时,传统聚类算法在普通计算系统上无法完成。此外,在数据处理面临规模化挑战的同时,数据处理需求的多样化逐渐显现。相比支撑单业务类型的数据处理业务,公共数据处理平台需要处理的大数据涉及在线/离线、线性/非线性、流数据和图数据等多种

复杂混合计算方式。下面对目前主流的大数据处理技术进行综述,并指出在应对智能电网大数据时这些技术的局限性,探讨可能的解决方案。

2.2 并行数据库

关系数据库(如 Oracle 等)主要存储结构化数据,提供便捷的数据查询分析能力、按照严格规则快速处理事务(transaction)的能力、多用户并发访问能力以及数据安全性的保证。通过 SQL 查询语言及强大的数据分析能力以及较高的程序与数据独立性等优点获得了广泛应用。

然而随着智能电网建设的加速,数据已远远超出关系型数据库的管理范畴,地理信息系统以及图片、音视频等各种非结构化数据逐渐成为需要存储和处理的海量数据的重要组成部分。面向结构化数据存储的关系型数据库已不能满足智能电网大数据快速访问、大规模数据分析的需求。主要表现在:

1) 数据存储容量有限。关系数据库可以有效处理 TB 级的数据,当数据量达到 PB 级时,目前主流数据库很难处理。为了回避此问题,目前电力企业采用先从“生数据”中提取“熟数据”的存储方式,这样虽然可以减少网络传输和数据库存储的数据量,但不可避免损失“生数据”中隐藏的重要特征量信息,如绝缘的放电频谱。

2) 关系模型束缚对海量数据的快速访问能力。关系模型是一种按内容访问的模型^[23]。即在传统的关系型数据库中,根据列的值来定位相应的行。这种访问模型,会在数据访问过程中引入耗时的输入输出,从而影响快速访问的能力。虽然,传统的数据库系统可以通过分区的技术(水平分区和垂直分区)来减少查询过程中数据输入输出的次数以缩减响应时间,提高数据处理能力,但是在海量数据的规模下,这种分区所带来的性能改善并不显著。

3) 缺乏对非结构化数据的处理能力。传统的关系型数据库对数据的处理只局限于某些数据类型,比如数字、字符、字符串等,对非结构化数据(图片、音频等)的支持较差。然而随着用户应用需求的提高、硬件技术的发展和互联网上多媒体交流方式的推广,用户对多媒体处理的要求从简单的存储上升为识别、检索和深入加工,面对日益增长的处理庞大的声音、图像、视频、E-mail 等复杂数据类型的需求,传统数据库已显得力不从心。

4) 扩展性差。在海量规模下,传统数据库一个致命弱点,就是其可扩展性(scalability)差。通常

解决数据库扩展性问题有 2 种方式:向上扩展(scale up)和向外扩展(scale out)。面对海量数据处理,通过提升服务器性能进行 scale up 的方式在成本及处理能力方面均不能满足要求,唯一可行的方法就是进行 scale out。关系数据库管理系统 scale out 的方法是通过数据库的垂直和水平切割将整个数据库部署到一个集群上,这种方法的优点在于可以采用 RDBMS 这种成熟技术,但缺点在于它是针对特定应用的,应用不同的话切割方法也不一样^[24]。

2.3 云计算技术

大数据技术的需求是伴随着云计算平台的出现而出现的,故有必要介绍一下云计算技术。实际上目前云计算技术是大数据存储与处理技术的重要组成部分。由于大数据的数据量和分布式的特点,使得传统的数据管理技术难以胜任这种海量数据。

云计算的核心是海量数据存储和数据并行处理技术。其核心思想包括分布式文件系统(distributed file system, DFS)和 MapReduce 技术,主要思路由 Google 公司提出。

DFS 有着高容错性的特点,并且是为部署在价格低廉的硬件上而设计的,而且它为应用程序提供高吞吐量的数据访问,适合那些有着超大数据集(large data set)的程序。Hadoop 提供了 DFS 的一种开源实现(HDFS),该分布式文件系统放宽了 POSIX 的要求,可以实现流的形式访问文件系统中的数据(streaming access),并具有高可靠性、高可扩展性以及负载均衡等能力。

MapReduce^[25]是 2004 年由谷歌公司提出的一个用来进行并行处理和生成大数据集的并行编程模型。Hadoop 包括了 MapReduce 的开源实现^[26],是引起关注的大数据处理技术之一。为使 MapReduce 并行编程模型更易使用,出现了多种大数据处理高级查询语言,如 Facebook 的 Hive^[27]、雅虎的 Pig^[28]、谷歌的 Sawzall^[29]等。这些高层查询语言通过解析器将查询语句解析为一系列 MapReduce 作业,在分布式文件系统上并行执行。与基本 MapReduce 系统相比,高层查询语言更适于用户进行大规模数据的并行处理^[30]。MapReduce 及高级查询语言在应用中也暴露出实时性和效率方面的不足,因此有很多研究针对它们进行优化。Cloudera 发布了实时查询开源项目 Impala 1.0 beta 版,实测表明比原来基于 MapReduce 的 Hive SQL 查询速度提升 3~90 倍^[31]。Mahout 是 Apache 开发的基于 MapReduce 的并行

数据挖掘项目,相对传统数据挖掘算法,性能大幅提升^[32]。

2.4 云计算在智能电网中的应用

智能电网中数据量最大的应属于电力设备状态监测数据。状态监测数据不仅包括在线的状态监测数据(时序数据和视频),还包括设备基本信息、实验数据、缺陷记录等,数据量极大,可靠性要求高,实时性要求比企业管理数据要高。

云计算技术在国内电力行业中的应用研究还处于探索阶段,研究内容主要集中在系统构想、实现思路和前景展望等方面。文献[33]针对智能电网状态监测的特点,结合 Hadoop,借助虚拟化技术、分布式冗余存储以及基于列存储的数据管理模式来存储和管理数据,以保证电网海量状态数据的可靠和高效管理,目前还只是一个框架。为了解决电力系统灾备中心资源利用率低,灾备业务流程复杂等一系列问题,文献[34]设计了云计算资源管理平台框架和部分模块,其目标是实现电力企业 ERP 数据的备份,但尚未实现。文献[35]初步设计了电力系统仿真云计算中心的系统架构及其所属的层次:基础设施云、数据管理云、仿真计算云等。文献[36]探讨了未来智能电网控制中心面临的挑战,提出物联网和云计算技术结合是新型控制中心的技术支撑。笔者课题组在实验室中搭建了 Hadoop 云计算平台,设计实现了基于 Hadoop 的电力设备状态监测存储系统^[37],对动态时序数据、静态数据以及视频数据进行了存储、关键字查询与并行处理方面的研究,并对系统进行了测试,验证了云计算平台高可靠性、良好的可扩展性和数据并行访问的性能优势。

在国外,云计算应用目前已用于海量数据的存储和简单处理,已有实现并运行的实际系统。文献[38]分析了电力系统中不同用户的实时查询需求,设计了用于实时数据流管理的智能电网数据云模型,特别适合处理智能电网中产生的海量流式数据,同时基于该模型实现了一个实时数据的智能测量与管理系统。Cloudera 公司设计并实施了基于 Hadoop 平台的智能电网在田纳西河流域管理局(Tennessee Valley Authority, TVA)上的项目^[39],帮助美国电网管理了数百 TB 的 PMU 数据,突显了 Hadoop 高可靠性以及价格低廉方面的优势;另外, TVA 在该项目基础上开发了 superPDC,并通过 openPDC 项目将其开源,此工作将有利于推动量测数据的大规模分析处理,并可为电网其他时序数据的处理提供通用

平台。日本 Kyushu 电力公司使用 Hadoop 云计算平台对海量的电力系统用户消费数据进行快速并行分析^[40],并在该平台基础上开发了各类分布式的批处理应用软件,提高了数据处理的速度和效率。

文献[41]对云计算平台应用于智能电网进行了详细的分析,得出的结论是:现有云计算平台可以满足智能电网监控软件运行的可靠性和可扩展性,但实时性、一致性、数据隐私和安全等方面的要求尚不能满足,有待进一步研究。

3 智能电网大数据的机遇与挑战

3.1 大数据传输及存储技术

随着智能电网建设的逐步推进,在电力系统各个环节的运行数据及设备状态在线监测数据被记录下来,由此产生的海量数据传输和存储问题不仅对监控装置造成极大的负担,而且也制约着电力系统智能化的跨越式发展^[42]。

通过数据压缩可以有效减少网络数据传输量,提高存储效率。因此数据压缩技术获得了广泛关注,文献[43]探讨了基于提升格式的故障暂态过程信号实时数据的压缩和重构算法,利用线性整数变换小波双正交滤波器组合哈夫曼编码方法对电力系统的实时数据进行压缩和解压缩。文献[44]研究了基于二维提升小波的火电厂周期性数据压缩算法。文献[45]研究了电力系统稳态数据参数化压缩算法。在输电线路状态监测系统中,为了发现绝缘子放电,泄漏电流的采样频率比较高,数据量大。目前该类系统普遍采用无线通信方式,网络带宽有限,因此需要进行数据压缩。文献[46]提出自适应多级树集合分裂(set partitioning in hierarchical trees, SPIHT)算法,该算法可以根据小波系数集合的显著性自适应地进行集合划分,尤其适合压缩泄漏电流这类高噪声信号。数据压缩一方面减少了存储空间,另一方面压缩和解压缩造成大量 CPU 资源的耗费。在数据到达监控中心后需要对数据进行解压缩,需要合适的计算与存储平台。

在数据存储方面,智能电网中的海量数据可以利用分布式文件系统来存储,比如利用 Hadoop 的 HDFS 等存储系统,然而这些系统虽然可以存储大数据,但很难满足电力系统的实时性要求^[47]。因此必须对系统中的大数据根据性能和分析要求进行分类存储:对性能要求非常高的实时数据采用实时数据库系统;对核心业务数据使用传统的并行数据仓库系统;对大量的历史和非结构化数据采用分布

式文件系统。本文提出为智能电网中的大数据构建多级存储系统,如图1所示。需要指出的是,鉴于目前云平台接收智能电网监测数据的实时性不能保证,可以在图1的数据接入与信息集成前面设置若干前置机,负责实时接收通信网中送来的报警信息或监测数据,并在云平台不能响应时负责暂存。

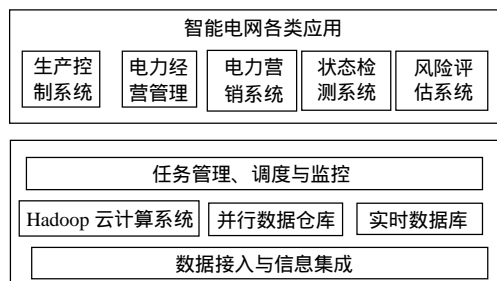


图1 智能电网大数据多级存储系统
Fig. 1 Multi-level storage system for big data from smart grid

另外,智能电网中的数据格式与传统商业数据具有很大的不同,拥有自己的特点。比如在故障录波及输变电设备状态监测中,波形数据较多,而波形数据与传统商业数据具有本质的不同,具有数据生成速度快等特点。因此需要研究面向智能电网大数据存储的格式,从而有利于后续的数据分析和计算。

智能电网环境下各类数据异构,不能用已有的简单数据结构来描述,而计算机算法在处理复杂结构数据方面相对低效,但处理同质的数据则非常高效^[48]。因此,如何将数据组织成合理的同质结构,是大数据存储处理中的一个重要问题。另外,智能电网中存在大量的非结构化和半结构化数据,如何将这些数据转化为一个结构化的格式,是一项重大挑战。

3.2 实时数据处理技术

3.2.1 数据处理的时效性

对大数据而言,数据处理速度十分重要。一般情况下,数据规模越大,分析处理的时间就会越长。传统的数据存储方案是为一定大小的数据量而设计的,在其设计范围内处理速度可能非常快,但不能适应大数据的要求。未来智能电网环境下,从发电、输变电环节,到用电环节,都需要实时数据处理。目前的云计算系统可以提供快速的服务,但有可能会受到短暂的网络拥塞,甚至是单台服务器故障的影响,而不能保证响应时间^[41]。

基于内存的数据库越来越受到关注。内存数据库就是将数据放在内存中直接操作的数据库。相对于磁盘,内存的数据读写速度要高出几个数量级,

将数据保存在内存中比从磁盘上访问能够极大地提高应用的性能。目前电力系统已经开始使用内存数据库,以提高实时性。例如,针对去年我国部分地区出现用电荒,而另一部分地区呈现电能过剩的状态,SAP推出了基于HANA内存数据库的智能电表分析解决方案^[49],希望能够将智能电网涉及的环节和电力大用户的数据进行集成和整合分析,以实现各地电能消费情况的分析,以做好相应的预防措施。

在大数据集中进行关键字的查询也是一个重要挑战。通过对整个数据集进行扫描来找到符合要求的记录的方法显然不可行,即使通过类似MapReduce这样的并行处理技术加快扫描,也不是很合理。而通过事先为数据建立索引结构来帮助查找是一种比较快速同时节省系统资源的方法。目前一般索引结构的设计仅支持一些简单数据类型,大数据则要求为复杂结构数据建立合适的索引结构^[50],这也是一个巨大的挑战。例如,物联网采集的多维数据,其数据量不断增长,同时对查询时限有要求,需要不断更新索引结构,索引的设计就非常具有挑战性。下面分别从发电、输变电和用电环节分析智能电网大数据在数据处理方面带来的挑战。

3.2.2 发电环节

发电企业的特点是生产过程连续、自动化程度高,要求全过程的实时监控、高速的实时数据处理、长期的历史数据存储以及生产信息的集成与共享。有研究表明,正常运行的SCADA系统接收到监测数据延时如果超过50ms,就会导致错误的控制策略^[51-52]。还有研究表明,SCADA系统在使用Internet环境下最普遍的TCP/IP协议时出现故障^[52-53],主要原因是TCP协议在进行流量控制和数据纠错,而造成数据延迟。未来的智能电网解决方案将需要实时响应,即使出现节点故障的情况。目前的关系数据库系统和云计算系统被设计为是处理永久、稳定的数据。关系数据库强调维护数据的完整性、一致性;云计算系统强调可靠性和可扩展性,但很难顾及有关数据及其处理的定时限制,不能满足工业生产管理实时应用的需要。

3.2.3 输变电环节

状态监测对数据存储与处理平台的性能或实时性具有较高的要求,而云计算技术虽然可以有效地处理大数据,但需要进一步提升云平台对海量监测数据的存取性能,以满足实时性的要求。以往的大规模停电事故^[54],最初是由一些环境因素引起

的,比如大风导致的线路跳闸等。现有 SCADA 系统的监控范围仅限于系统的主参数,对构成系统的各重要设备的健康状况的信息缺失,致使运行人员在事故面前难以做出正确的处理。未来智能电网要求具有故障自愈功能,其 SCADA 系统须拥有全网的监测数据,需要将电力设备的状态数据纳入其中,这对平台的实时处理提出了更高的要求。

新型绿色能源发电功率的不稳定造成电网的波动,对整个电网调度形成很大的压力。目前电网调度与控制模型不能够处理这种大量的小型发电系统产生的波动和不可预知的行为。最新的研究表明^[41],为支持这种情况,需要创建一种新型的电网状态监控系统,能够更加细粒度地跟踪电网实时状态。因此未来的 SCADA 系统需要实时处理比目前多几个数量级的监控数据。

3.2.4 用电环节

未来智能电网环境下,家庭可能配备多种电能、电量监测设备,用以实现低成本的用电,并与电网的负载相匹配。例如,电热水器可能会选择夜间这种用电量低谷时段运行;空调会根据用户舒适度、电价以及电网负荷等参数实时自动调整。某种程度上,我们可以认为 SCADA 系统进入了普通家庭,用电环节的实时数据处理变得越来越重要。

3.3 异构多数据源处理技术

3.3.1 异构信息的整合

未来智能电网要求贯通发电、输电、变电、配电、用电、调度等多个环节,实现信息的全面采集、流畅传输和高效处理,支撑电力流、信息流、业务流的高度一体化。因此,首要功能是实现大规模多源异构信息的整合,为智能电网提供资源集约化配置的数据中心。针对海量异构数据,如何构建一个模型来对其进行规范表达,如何基于该模型来实现数据融合,以及对其进行有效的存储和高效查询是亟需解决的问题。

电网各信息系统大多是基于本业务或本部门的需求,存在不同的平台、应用系统和数据格式,导致信息与资源分散,异构性严重,横向不能共享,上下级间纵向贯通困难^[55],例如:电力系统中存在监控、能量管理、配电管理、市场运营等各类信息系统,大多相互独立,数据信息不能共享。使用云平台实现各独立系统的集成,可实现这些分散孤立系统之间的信息互通。

另外,智能电网的基础设施规模庞大,数量众

多且分布在不同地点。例如:国家电网公司的信息化平台在公司总部与各个网省公司建立 2 级数据中心,实现公司总部、网省公司、地市县公司的 3 层应用。如何有效管理这些基础设施、减少数据中心的运营成本是一个巨大的挑战。

3.3.2 各类电网数据的高效管理

在智能电网异构多源信息融合和管理中,建立类似 IEC 61850 或 IEC 61970 的信息互操作模型是很有必要的。由于智能电网中的数据类型比 IEC 61850 所涉及的类型要多,所以应用多层知识结构和语义的方法、建立面向领域的分析模型与基于语义的服务模型是一种可选的方法。综合运用统计学习、支持向量机、相关向量机和关联规则挖掘等理论,研究异构数据融合与挖掘的集成方案以及实时挖掘算法。由于设备状态的劣化是一个由量变到质变的过程,像多年积累的油色谱这样的时序数据的挖掘更有意义,目前这种大数据挖掘虽有一些研究成果,但实用化程度不高。

3.4 大数据可视化分析技术

面对海量的智能电网数据,如何在有限的屏幕空间下,以一种直观、容易理解的方式展现给用户,是一项非常有挑战性的工作^[56]。可视化方法已被证明为一种解决大规模数据分析的有效方法,并在实践中得到广泛应用^[57]。智能电网各类应用产生的大规模数据集,其中包含高精度、高分辨率数据,时变数据和多变量数据等。一个典型的数据集可达 TB 数量级。如何从这些庞大复杂的数据中快速而有效地提取有用的信息,成为智能电网应用中的一个关键技术难点。可视化通过一系列复杂的算法将数据绘制成高精度、高分辨率的图片,并提供交互工具,有效利用人的视觉系统,并允许实时改变数据处理的算法参数,对数据进行观察和定性及定量分析^[58]。

这方面的挑战主要包括可视化算法的可扩展性、并行图像合成算法、重要信息的提取和显示等方面^[59]。

4 结论

未来的智能电网将是依托大数据处理分析技术的全景实时电网。云计算为这种异构且多样化的数据提供了存储和分析的平台。平台运行一段后必然产生大数据,云平台和大数据分析将会为电力设备的状态检修、电网自愈、孤立信息系统的互通提供支持,并成为重要的候选方案,具有低成本、好的系统扩展性(存储容量无限)、高可靠性、并行

分析等优势,在国际上已有几例系统投入实际运行,但在实时性、数据一致性、隐私性和安全性方面仍有不少的挑战,需要找出相应的解决方法。大数据的处理技术还很欠缺,有待人们去探索。

参考文献

- [1] Xi Fang, Satyajayant Misra, Guoliang Xue, et al. Smart Grid, the new and improved power grid: a survey[J]. IEEE Communications Surveys and Tutorials (COMST), 2012, 14(4): 944-980.
- [2] 张文亮, 汤广福, 查鲲鹏, 等. 先进电力电子技术在智能电网中的应用[J]. 中国电机工程学报, 2010, 30(4): 1-7.
Zhang Wenliang, Tang Guangfu, Zha Kunpeng, et al. Application of advanced power electronics in smart grid[J]. Proceedings of the CSEE, 2010, 30(4): 1-7(in Chinese).
- [3] 李国杰. 大数据研究的科学价值[J]. 中国计算机学会通讯, 2012, 8(9): 8-15.
Li Guojie. The scientific value of big data[J]. Research Communications of The CCF, 2012, 8(9): 8-15(in Chinese).
- [4] Divyakant Agrawal, Philip Bernstein, Elisa Bertino, et al. Challenges and opportunities with big data[J]. Proceedings of the VLDB Endowment, 2012, 5(12): 2032-2033.
- [5] IBM Corporation Software Group. IBM big data overview for energy and utilities[EB/OL]. 2011-06[2012]. <http://www-01.ibm.com/software/tivoli/solutions/industry/energy-utilities/>.
- [6] McKinsey Global Institute. Big data: the next frontier for innovation, competition, and productivity[R]. 2011.
- [7] Peijian Wang. D-pro: dynamic data center operations with demand-responsive electricity prices in smart grid[J]. IEEE Transactions on Smart Grid, 2012, 3(4): 1743-1754.
- [8] 周晖, 钮文洁, 王毅. 从缴费行为分析电力客户的信用度[J]. 电力需求侧管理, 2006, 8(6): 12-16.
Zhou Hui, Niu Wenjie, Wang Yi. Analysis of clients' credit based on their paying behaviors[J]. Power Demand-Side Management, 2006, 8(6): 12-16(in Chinese).
- [9] Conejo A J, Morales J M, Baringo L. Real-time demand response model[J]. IEEE Transactions on Smart Grid, 2010, 1(3): 236-242.
- [10] 牛东晓, 谷志红, 邢棉, 等. 基于数据挖掘的 SVM 短期负荷预测方法研究[J]. 中国电机工程学报, 2006, 26(18): 6-12.
Niu Dongxiao, Gu Zhihong, Xing Mian, et al. Study on forecasting approach to short-term load of SVM based on data mining[J]. Proceedings of the CSEE, 2006, 26(18): 6-12(in Chinese).
- [11] 谢华成, 陈向东. 面向云存储的非结构化数据存取[J]. 计算机应用, 2012, 32(7): 1924-1928.
Xie Huacheng, Chen Xiangdong. Cloud storage-oriented unstructured data storage[J]. Journal of Computer Applications, 2012, 32(7): 1924-1928(in Chinese).
- [12] 李锋, 谢俊, 兰金波, 等. 智能变电站继电保护配置的展望和探讨[J]. 电力自动化设备, 2012, 32(2): 122-126.
Li Feng, Xie Jun, Lan Jinbo, et al. Prospect and discussion of relay system configuration for intelligent substation[J]. Electric Power Automation Equipment, 2012, 32(2): 122-126(in Chinese).
- [13] 江苏瑞中数据股份有限公司. 海迅实时数据库助力智能电网建设[EB/OL]. 2011-05[2013-02]. <http://hvinc.chinapower.com.cn/memberscenter/sitebuild4/content.asp>.
- [14] 侯子良, 潘钢. 建设数字化电厂示范工程加快火电厂信息化进程[J]. 中国电力, 2005, 38(2): 78-80.
Hou Ziliang, Pan Gang. Constructing demonstration projects of digitized power plant to speed up the informatization process in fossil-fired power plants[J]. Electric Power, 2005, 38(2): 78-80(in Chinese).
- [15] 李晗, 萧德云. 基于数据驱动的故障诊断方法综述[J]. 控制与决策, 2011, 26(1): 1-16.
Li Han, Xiao Deyun. Survey on data driven fault diagnosis methods[J]. Control and Decision, 2011, 26(1): 1-16(in Chinese).
- [16] 周东华, 胡艳艳. 动态系统的故障诊断技术[J]. 自动化学报, 2009, 35(6): 748-758.
Zhou Donghua, Hu Yanyan. Fault diagnosis techniques for dynamics system[J]. Acta Automatica Sinica, 2009, 35(6): 748-758(in Chinese).
- [17] Pregelj A, Begovic M, Rohatgi A. Quantitative techniques for analysis of large data set in renewable distributed generation[J]. IEEE Trans on Power Systems, 2004, 19(3): 1277-1285.
- [18] Versant. NoSQL and the smart grid big data challenge[EB/OL]. 2012-08[2013-02]. <http://www.greentechmedia.com/articles/read/versant-nosql-and-the-smart-grid-big-data-challenge/>.
- [19] David Kligman. PG&E's Austin kicks off conference on dealing with smart grid data[EB/OL]. 2012-08[2013-02]. <http://www.pgecurrents.com/2012/08/14/pg-topic-is-dealing-with-data-that-comes-with-smart-grid/>.
- [20] 金澈清, 钱卫宁, 周傲英. 流数据分析与管理综述[J]. 软件学报, 2004, 5(8): 1172-1181.
Jin Cheqing, Qian Weining, Zhou Aoying. Analysis and management of streaming data: a survey[J]. Journal of Software, 2004, 5(8): 1172-1181 (in Chinese).
- [21] 国网信通有限公司. 信通公司举办大数据开启智能电网新时代研讨会[EB/OL]. 2012-07[2013-02]. <http://www.sgit.sgcc.com.cn/newzxzx/gsxw/07/277345.shtml>.
- [22] 张广斌, 束洪春, 于继来. 利用广义电流模量的行波实测数半监督聚类筛选[J]. 中国电机工程学报, 2012, 32(10): 150-158.
Zhang Guangbin, Shu Hongchun, Yu Jilai. Travelling wave field data contingency screening based on semi-supervised clustering using generalized current modal component[J]. Proceedings of the CSEE, 2012, 32(10): 150-158(in Chinese).
- [23] Codd E F. A relational model of data for large shared data banks[J]. Communications of the ACM, 1970, 13(6): 377-387.
- [24] Roland Bouman. Database sharding at Netlog with MySQL and PHP[EB/OL]. 2009-02[2013-02]. <http://www.jurriaanpersyn.com/archives/2009/02/12/database-sharding-at-netlog-with-mysql-and-php/>.
- [25] Jeffrey Dean, Sanjay Ghemawat. MapReduce: simplified data processing on large clusters[C]//OSDI'04: Sixth Symposium on Operating System Design and Implementation. San Francisco, California: USENIX Association Berkeley, 2004: 137-150.
- [26] Apache. Apache Hadoop core[EB/OL]. 2012-08[2013-02]. <http://hadoop.apache.org/core/>.
- [27] Thusoo A, Sarma J, Jain N, et al. Hive: a warehousing solution over map-reduce framework[C]//Proc of the 35th Int Conf on Very Large Data Bases (VLDB). Lyon, France: VLDB, 2009: 1626-1629.
- [28] Christopher Olston, Benjamin Reed, Utkarsh Srivastava. Pig latin: a not-so-foreign language for data processing[C]//Proceedings of the 2008 ACM SIGMOD international conference on Management of data. Vancouver, Canada: ACM, 2008: 1099-1110.
- [29] Rob Pike, Sean Dorward, Robert Griesemer, et al. Interpreting the data: parallel analysis with Sawzall[J]. Scientific Programming, 2005, 13(4): 277-298.
- [30] 王鹏, 孟丹, 詹剑锋, 等. 数据密集型计算编程模型研究进展[J]. 计算机研究与发展, 2010, 47(11): 1993-2002.
Wang Peng, Meng Dan, Zhan Jianfeng, et al. Review of programming models for data-intensive computing[J]. Journal of Computer Research and Development, 2010, 47(11): 1993-2002(in Chinese).
- [31] Marcel Kornacker, Justin Erickson. Cloudera Impala: real-time queries in Apache Hadoop for real[EB/OL]. 2012-10

- [2013-02]. <http://blog.cloudera.com/blog/2012/10/cloudera-impala-real-time-queries-in-apache-hadoop-for-real/>.
- [32] Apache. What is Apache Mahout[EB/OL]. 2011-05[2013-02]. <http://mahout.apache.org/>.
- [33] 王德文, 宋亚奇, 朱永利. 基于云计算的智能电网信息平台[J]. 电力系统自动化, 2010, 34(22): 7-12.
Wang Dewen, Song Yaqi, Zhu Yongli. Information platform of smart grid based on cloud computing[J]. Automation of Electric Power Systems, 2010, 34(22): 7-12(in Chinese).
- [34] 朱征, 顾中坚, 吴金龙, 等. 云计算在电力系统数据灾备业务中的应用研究[J]. 电网技术, 2012, 36(9): 43-50.
Zhu Zheng, Gu Zhongjian, Wu Jinlong, et al. Application of cloud computing in electric power system data recovery[J]. Power System Technology, 2012, 36(9): 43-50(in Chinese).
- [35] 沐连顺, 崔立忠, 安宁. 电力系统云计算中心的研究与实践[J]. 电网技术, 2011, 35(6): 170-175.
Mu Lianshun, Cui Lizhong, An Ning. Research and practice of cloud computing center for power system[J]. Power System Technology, 2011, 35(6): 170-175(in Chinese).
- [36] 王广辉, 李保卫, 胡泽春, 等. 未来智能电网控制中心面临的挑战和形态演变[J]. 电网技术, 2011, 35(8): 1-5.
Wang Guanghui, Li Baowei, Hu Zechun, et al. Challenges and future evolution of control center under smart grid environment[J]. Power System Technology, 2011, 35(8): 1-5(in Chinese).
- [37] 刘树仁, 宋亚奇, 朱永利, 等. 基于 Hadoop 的智能电网状态监测数据存储研究[J]. 计算机科学, 2013, 40(1): 81-84.
Liu Shuren, Song Yaqi, Zhu Yongli, et al. Research on data storage for smart grid condition monitoring using Hadoop[J]. Computer Science, 2013, 40(1): 81-84(in Chinese).
- [38] Rusitschka S, Eger K, Gerdes C. Smart grid data cloud: a model for utilizing cloud computing in the smart grid domain[C]//Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference. Gaithersburg, MD: IEEE, 2010: 483-488.
- [39] Christophe Bisciglia. The smart grid: Hadoop at the Tennessee Valley Authority(TVA)[EB/OL]. 2009-06[2013-02]. <http://www.cloudera.com/blog/2009/06/smart-grid-hadoop-tennessee-valley-authority-tva/>.
- [40] Kawasoe S, Igarashi Y, Shibayama K, et al. Examples of distributed information platforms constructed by power utilities in Japan[C]//CIGRE 2012. Paris, France: CIGRE, 2012: 108-113.
- [41] Kenneth P Birman, Lakshmi Ganesh, Robbert van Renesse. Running smart grid control software on cloud computing architectures[C]//Workshop on Computational Needs for the Next Generation Electric Grid, Cornell University. Ithaca, NY: DOE, 2011: 1-28.
- [42] 张保会. 加强继电保护与紧急控制系统的研究提高互联电网安全防护能力[J]. 中国电机工程学报, 2004, 24(7): 1-6.
Zhang Baohui. Strengthen the protection relay and urgency control systems to improve the capability of security in the interconnected power network[J]. Proceedings of the CSEE, 2004, 24(7): 1-6(in Chinese).
- [43] 闫常友, 杨奇逊, 刘万顺. 基于提升格式的实时数据压缩和重构算法[J]. 中国电机工程学报, 2005, 25(9): 6-10.
Yan Changyou, Yang Qixun, Liu Wanshun. A real-time data compression & reconstruction method based on lifting scheme[J]. Proceedings of the CSEE, 2005, 25(9): 6-10(in Chinese).
- [44] 鲍文, 周瑞, 刘金福. 基于二维提升小波的火电厂周期性数据压缩算法[J]. 中国电机工程学报, 2007, 27(29): 96-101.
Bao Wen, Zhou Rui, Liu Jinfu. A periodical data compression method based on 2-D lifting wavelet transform in thermal power plant[J]. Proceedings of the CSEE, 2007, 27(29): 96-101(in Chinese).
- [45] 张斌, 张东来. 电力系统稳态数据参数化压缩算法[J]. 中国电机工程学报, 2011, 31(1): 72-79.
Zhang Bin, Zhang Donglai. Parametric compression algorithm for power system steady data[J]. Proceedings of the CSEE, 2011, 31(1): 72-79(in Chinese).
- [46] 朱永利, 翟学明, 姜小磊. 绝缘子泄漏电流的自适应 SPIHT 数据压缩[J]. 电工技术学报, 2011, 26(12): 190-196.
Zhu Yongli, Zhai Xueming, Jiang Xiaolei. Adaptive SPIHT algorithm for data compression of insulator leakage currents[J]. Transactions of China Electrotechnical Society, 2011, 26(12): 190-196(in Chinese).
- [47] Stonebraker M, Abadi D J, Madden S, et al. MapReduce and parallel DBMSs: friends or foes?[J]. Communications of the ACM, 2010, 53(1): 64-71.
- [48] 周晓方, 陆嘉恒, 李翠平, 等. 从数据管理视角看大数据挑战[J]. 中国计算机学会通讯, 2012, 8(9): 16-20.
- [49] 丁慧茹. SAP 推 HANA 电力行业应用智能电表分析提升服务[EB/OL]. 2011-12[2013-02]. <http://cio.zdnet.com.cn/cio/2011/1220/2070971.shtml>.
- [50] Cooper B F, Neal Sample, Franklin M J, et al. A fast index for semistructured data[C]//Proceedings of the 27th VLDB Conference. Roma, Italy: VLDB, 2001: 341-350.
- [51] Chi Ho, Robbert van Renesse, Mark Bickford, et al. Nysiad: practical protocol transformation to tolerate byzantine failures[C]//USENIX Symposium on Networked System Design and Implementation (NSDI 08). San Francisco, CA: USENIX, 2008: 175-188.
- [52] Hopkinson Ken M, Giovanini Renan, Wang Xiaoru, et al. EPOCHS: integrated cots software for agent-based electric power and communication simulation[C]//WSC 2003. New Orleans, Louisiana, USA: IEEE, 2003, 2: 1158-1166.
- [53] Junqueira F P, Reed B C. The life and times of a Zookeeper[C]//SPAA '09. New York, USA: ACM, 2009: 46-46.
- [54] Wikipedia. Northeast blackout of 2003[EB/OL]. 2003-12[2013-02]. http://en.wikipedia.org/wiki/Northeast_Blackout_of_2003.
- [55] 张文亮, 刘壮志, 王明俊. 智能电网的研究进展及发展趋势[J]. 电网技术, 2009, 33(13): 1-11.
Zhang Wenliang, Liu Zhuangzhi, Wang Mingjun. Research status and development trend of smart grid[J]. Power System Technology, 2009, 33(13): 1-11(in Chinese).
- [56] Wong P C, Shen H W, Chen C, et al. Top ten interaction challenges in extreme-scale visual analytics[J]. Computer Graphics and Applications, 2012, 32(4): 63-67.
- [57] 袁晓如, 张昕, 肖何, 等. 可视化研究前沿及展望[J]. 科研信息化技术与应用, 2011, 2(4): 3-13.
- [58] Wong P C, Thomas J. Visual analytics[J]. IEEE Computer Graphics and Applications, 2004, 24(5): 20-21.
- [59] Thomas J J, Cook K A. Illuminating the path: the research and development agenda for visual analytics[M]. IEEE Computer Society, 2005: 28-32.



宋亚奇

收稿日期: 2012-11-28。

作者简介:

宋亚奇(1979), 男, 博士研究生, 主要研究方向为电力信息智能处理、云计算, E-mail: bdsyq@163.com;

周国亮(1978), 男, 博士后, 副教授, 主要研究方向为智能电网、大数据处理;

朱永利(1963), 男, 博士, 教授, 博士生导师, 主要研究方向为人工智能及应用、网络化监控与电力系统自动化。

(责任编辑 李兰欣)