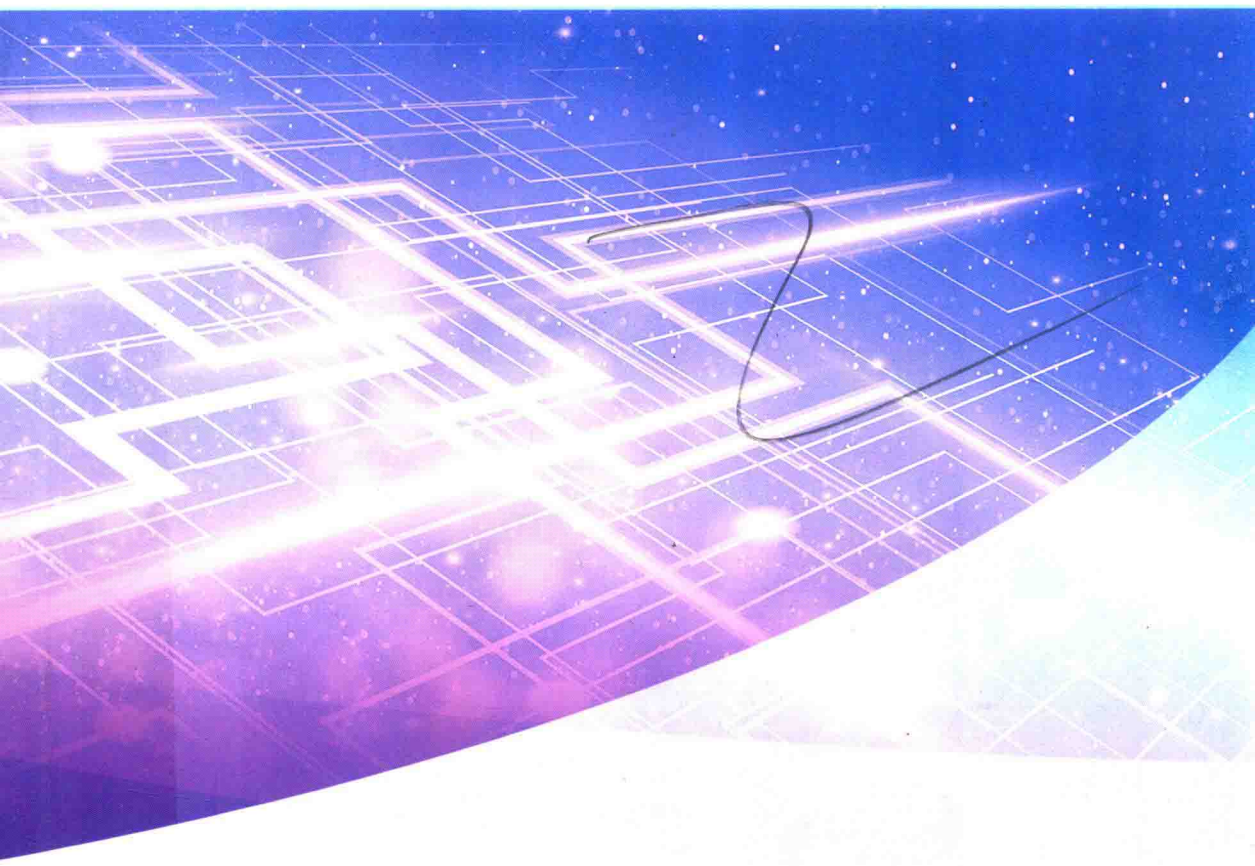
电力设备监测大数据分析方法

宋亚奇　李　莉　朱永利　编著



中国电力出版社
CHINA ELECTRIC POWER PRESS

电力设备监测大数据分析方法

宋亚奇　李　莉　朱永利　编著

中国电力出版社
CHINA ELECTRIC POWER PRESS

内 容 提 要

本书针对智能电网环境下电力设备监测大数据的存储、处理和分析方法展开研究，主要内容涉及利用云计算和大数据处理技术（Hadoop、MaxCompute、Spark）研究电力设备监测大数据的存储方法、数据在分布式平台下的分布策略、波形信号的并行分析和特征提取方法、多源数据的并行关联查询和分析方法、监测数据的并行聚类方法、短时高并发报警数据的实时模式识别、监测数据流式处理方法等方面。

本书可以作为普通高校电气工程类、计算机和电子信息类研究生教材和参考读物，也可作为云计算、大数据等相关专业研究人员、工程技术人员和教师的参考用书。

图书在版编目（CIP）数据

电力设备监测大数据分析方法 / 宋亚奇，李莉，朱永利编著. —北京：中国电力出版社，2018.10

ISBN 978-7-5198-2236-1

Ⅰ. ①电… Ⅱ. ①宋… ②李… ③朱… Ⅲ. ①电力设备-监测-数据处理 Ⅳ. ①TM4

中国版本图书馆 CIP 数据核字（2018）第 155824 号

出版发行：中国电力出版社
地　　址：北京市东城区北京站西街 19 号（邮政编码 100005）
网　　址：http://www.cepp.sgcc.com.cn
责任编辑：陈　丽（010-63412348）
责任校对：黄　蓓　常燕昆
装帧设计：郝晓燕
责任印制：石　雷

印　　刷：北京雁林吉兆印刷有限公司
版　　次：2018 年 10 月第一版
印　　次：2018 年 10 月北京第一次印刷
开　　本：710 毫米×1000 毫米　16 开本
印　　张：12
字　　数：195 千字
印　　数：0001—1000 册
定　　价：60.00 元

前　言

随着电网规模迅速增长，电网结构越来越复杂，信息化与电力生产深度融合，智能化电力一次设备和常规电力设备的在线监测都得到了较大发展并成为趋势，监测数据变得日益庞大，设备中进行获取与传输的监测数据成几何级增长。电力设备在线监测系统在数据存储、查询和数据分析等方面面临巨大的技术挑战。如何对电力设备监测大数据进行高效、可靠的存储，并快速访问和分析，是当前电力信息处理领域和大数据处理领域重要的研究课题。电力设备监测大数据的特点和所面临的技术挑战包括：

（1）电力设备状态监测数据的规模非常巨大，从 TB 级别往 PB 级别发展。现阶段输变电设备以及输电线路的在线监测系统（局部放电、油色谱、线路覆冰监测、绝缘子泄漏电流、电站视频监测等）涉及的监测点数量多、数据采样率高、数据类型多样（结构化数据、非结构化数据），监测数据体量巨大。在线监测系统的计算处理速度及响应时间受限于硬件性能，在发生电网故障情况下，短时间内大量数据若得不到及时处理，可能面临信息延迟甚至丢失的风险。

（2）处理速度快。对海量的输变电设备监测历史数据进行离线分析处理的过程包括数据清洗、格式转换、信号去噪、特征提取、模式识别等，任何一个环节处理速度慢，都会成为应用系统的性能瓶颈。数据处理平台要能够提供并行化、高吞吐量、批处理的能力。除历史数据的离线分析处理外，其他的一些应用场景，包括：Ad Hoc 数据分析查询、监测大数据流式处理等，都对系统的数据处理速度提出挑战。

（3）数据存储与处理平台的架构。如何根据输变电设备监测大数据的特点和应用需求，选择、组合、合理利用现有大数据技术（Hadoop、Spark、多核计算、云计算等）构建高可靠性及高可用性的分布式存储与计算平台，并利

用并行计算技术（MapReduce、MR^2、MPI 等），满足海量历史数据查询分析、数据挖掘、在线服务等各类计算任务性能需求，助力电力大数据价值释放极具挑战性。

（4）多源异构数据的关联分析。在输变电设备监测大数据应用中，需要对多源监测数据进行关联分析，也需要对气象、环境等电网系统外数据与监测数据进行关联分析。由于关联分析涉及的数据体量巨大，传统的基于关系数据库和数据仓库的表连接查询、表关联分析方法，以及传统的基于单机环境的统计分析算法、模式识别算法在执行效率方面无法满足“数据密集型”大数据应用系统的性能要求。这种需求对数据的存储模式、数据的分布策略以及算法执行的性能提出挑战。

（5）时空属性。监测数据采样值具有时间属性，监测装置节点具有地理位置属性（空间属性）。对监测数据的查询不仅局限于按照设备关键字、采集时间进行查询，还可以基于更加复杂的条件约束进行多条件查询。例如，根据用户指定的某个地理区域（经纬度范围），查询区域内监测装置在指定时间范围里的监测数据，绘制趋势曲线，完成统计分析等。

（6）价值密度低。以电站视频监测数据为例，连续监测的视频流中，有用的数据可能仅有几秒钟。传统的电力设备状态评估方法中，只对异常数据关注、处理和采用，而丢弃所谓“正常数据”。然而大量的正常数据，或者介于正常和异常之间的临界状态的数据，也可能成为故障分析判断的重要依据。

面对上述挑战，常规的数据存储与管理方法大都构建在大型服务器、磁盘阵列（存储硬件）以及关系数据库系统（数据管理软件）上，系统扩展性差、访问性能低下、成本高，在存储和处理监测大数据时遇到了极大的困难。

本书基于云平台和 Hadoop、Spark、MaxCompute 等大数据处理技术，对电力设备监测大数据的存储模式、数据分布策略、波形信号并行分析、特征提取、多源数据关联查询、并行聚类划分以及报警数据的实时模式识别等方面进行了研究，并取得了一系列的创新性成果。

本书由华北电力大学宋亚奇、李莉统稿和编写，华北电力大学朱永利教授对全书进行了审阅。

本书的研究工作得到了国家自然科学基金项目（51677072）以及中央高校基本科研业务费专项资金资助项目（2016MS116，2016MS117）的资助。在这里，谨对所有给予我们指导、关心和帮助过的单位和个人表达最诚挚的谢意。

由于学术水平和工程经验有限，对所研究内容和把握能力还存在不足和欠缺，书中不足之处在所难免，恳请各位专家和读者批评指正！

作　者

2018 年 5 月

目 录

第一章　电力设备监测大数据的特点和所面临的技术挑战

第一节　电力设备监测大数据的特点

一、智能电网与监测大数据

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。

由于物联网和无线传感技术的广泛应用，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 亿美元，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 4PB 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，大数据产生于系统的各个环节。比如在用电侧，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，有必要对智能电网中的大数据的特点进行归纳，本章根据业务领域、数据结构、数据的来源，对数据特点进行了分类总结。电网业务数据大致分为三类：① 电网运行和设备检测或监测数据；② 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；③ 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。

根据处理时限要求，结构化数据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 200kHz 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。

（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 60GB，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 402GB[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 5min 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 万个智能电表中收集超过 3TB 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。

书中内容主要针对输变电环节中的设备监测数据展开，发电侧和用电侧的大数据应用请参考相关论文和书目。

二、电力设备监测的发展现状

目前国内电力设备的状态监测尚处于起步阶段，按设备分类构成各个单一的监测系统，彼此相互独立，形成信息孤岛。一个电力企业常常拥有多个不同的状态监测系统需要维护，服务器等硬件重复配置。这种状况不符合电力企业向统一数据平台整合的趋势。国家电网有限公司在“十一五”期间实施了“SG186 工程”，目标是建设统一的数据中心系统，将原来分散、孤立的数据资源集中存储、统一管理，建立完善、统一的报表与指标体系规范，有效改善指标多人维护、多重上报的问题，为各应用系统提供数据层集中服务的数据环境[20]。目前建设的这种系统的数据主要是从电力企业已有的业务系统（如生产 MIS 和营销系统）中抽取而来，采用 oracle 关系数据库，存储的数据主要是生产、营销和设备等的静态数据，一些电力系统的动态数据，如故障录波、设备绝缘状态信号和电能质量记录数据均还未接入，且这种海量的时序动态数据直接存入数据中心的关系数据库会占用过多的存储。因此，有必要研究动态时序数据的高效存储方法，为电网设备状态的在线监测系统以及下一代数据中心存储电网设备的动态信号提供理论支持和技术储备。

三、电力设备监测数据的特点

电力设备监测数据具备大数据所普遍具有的“4V”特征，即体量大（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 10 000 个遥测点，按采样间隔 3～4s 计算，每年产生 1.03TB 的数据（1.03TB=12 字节/帧×0.3 帧/s×10 000 遥测点×86 400s/天×365 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 495TB 的数据。目前正在发展的直升机和无人机巡线技术所产生的红外、紫外视频信息，每年作业采集的数据量达 40TB。一个省级电力公司已有数字化变电站可达 200 座左右，每天产生的监测数据量可达百 TB。随着监测系统规模的扩大，以及数据采样频率的提高，数据量还将成倍

增加。若同时考虑环境、气象、地理信息等，则数据量更为庞大。

（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半个小时采样一次，而绝缘放电数据的采样速率高达几百 kHz，甚至 GHz。随着状态监测技术的发展和智能化设备类型与数量的增加，音视频等非结构化数据在数据中的占比进一步加大。此外，大数据应用过程中还存在着对电网系统运行环境相关数据（气象、地理、环境等）的大量关联分析需求，而这些都直接导致了数据类型的增加以及状态评估应用领域数据的复杂度。

（3）价值密度低。由于监测数据的第一个特征——“体量巨大”，导致了数据集肯定是有价值的，但是一个“大数据集”的价值有可能与一个“小数据集”的价值相当，因此该特点被称为价值密度低。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 1～2s。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）变化快。这个特点有 2 层含义：① 监测数据产生的速度很快，如前所述，由于采样率很高所致；② 对不断到达的监测数据，要求在短时间内对其进行数据加工和分析，也就是要求处理的速度快。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[21]。监测数据在流式计算的场景下，数据的价值会随着时间的推移而逐渐降低。

另外，电网中对监测数据的处理，对数据质量也会有一定的要求，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。

上述电网中监测数据的特点，给智能电网建设，尤其是输变电设备监测系统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[22]。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]。

第二节　电力设备监测数据存储和数据处理所面临的技术挑战

一、电力设备监测大数据技术挑战分析

目前电网规模增长迅速，电网结构也越来越复杂，信息化与电力生产深度融合，智能化电力一次设备和常规电力设备的在线监测都得到了较大发展并成为趋势，监测数据变得日益庞大，设备中进行获取与传输的监测数据成几何级增长。输变电设备在线监测系统在数据存储、查询和数据分析等方面面临巨大的技术挑战。如何对输变电设备监测大数据进行高效、可靠地存储，并快速访问和分析，是当前电力信息处理领域和大数据处理领域重要的研究课题。对电力设备监测大数据的特点和所面临的技术挑战分析如下。

（1）电力设备状态监测数据的规模非常巨大。从 TB 级别往 PB 级别发展。现阶段输变电设备以及输电线路的在线监测系统（局部放电、油色谱、线路覆冰监测、绝缘子泄漏电流、电站视频监测等）涉及的监测点数量多、数据采样率高、数据类型多样（结构化数据、非结构化数据），监测数据体量巨大。以一个省级电网公司为例，按照 10 000 套终端，每套终端每 1min 采集一次数据计算，每天产生数据总量约 2150GB，每年产生数据达到 760TB。目前，正在发展的直升机巡线所产生的红外、紫外视频信息，每年作业采集的数据量达 TB 级别。随着监测系统规模的升级，监测数据的体量还将成倍增长。在线监测系统的计算处理速度及响应时间受限于硬件性能，在发生电网故障情况下，短时间内大量数据若得不到及时处理，可能面临信息延迟甚至丢失的风险[23]。

（2）处理速度快。对海量的输变电设备监测历史数据进行离线分析处理的过程包括数据清洗、格式转换、信号去噪、特征提取、模式识别等，任何一个环节处理速度慢，都会成为应用系统的性能瓶颈。以利用 EMD 进行局部放

电信号分解为例，在单机环境下，对长度为 5000 点的局放信号完成 EMD 分解，大约需要 1min；如果对海量历史数据执行串行处理，速度将极其缓慢。因此，数据处理平台要能够提供并行化、高吞吐量、批处理的能力。除历史数据的离线分析处理外，其他的一些应用场景，包括：Ad Hoc 数据分析查询[24]、海量数据的在线服务[25]、监测大数据流式处理[26]等，都对系统的数据处理速度提出挑战。

（3）数据存储与处理平台的架构。如何根据输变电设备监测大数据的特点和应用需求，选择、组合、合理利用现有大数据技术（Hadoop、Spark、多核计算、云计算等）构建高可靠性及高可用性的分布式存储与计算平台，并利用并行计算技术（MapReduce、MR^2、MPI 等），满足海量历史数据查询分析、数据挖掘、在线服务等各类计算任务性能需求，助力电力大数据价值释放极具挑战性。

（4）多源异构数据的关联分析。在输变电设备监测大数据应用中，需要对多源监测数据进行关联分析，也需要对气象、环境等电网系统外数据与监测数据进行关联分析。由于关联分析涉及的数据体量巨大，传统的基于关系数据库和数据仓库的表连接查询、表关联分析方法，以及传统的基于单机环境的统计分析算法、模式识别算法在执行效率方面无法满足“数据密集型”大数据应用系统的性能要求。这种需求对数据的存储模式、数据的分布策略以及算法执行的性能提出挑战。

（5）时空属性。监测数据采样值具有时间属性，监测装置节点具有地理位置属性（空间属性）。对监测数据的查询不仅局限于按照设备关键字、采集时间进行查询，还可以基于更加复杂的条件约束进行多条件查询。例如，根据用户指定的某个地理区域（经纬度范围），查询区域内监测装置在指定时间范围里的监测数据，绘制趋势曲线，完成统计分析等。

（6）价值密度低。以电站视频监测数据为例，连续监测的视频流中，有用的数据可能仅有几秒钟。传统的电力设备状态评估方法中，只对异常数据关注、处理和采用，而丢弃所谓“正常数据”。然而大量的正常数据，或者介于正常和异常之间的临界状态的数据，也可能成为故障分析判断的重要依据。

面对上述挑战，常规的数据存储与管理方法大都构建在大型服务器、磁盘阵列（存储硬件）以及关系数据库系统（数据管理软件）上，系统扩展性差、

访问性能低下、成本高，在存储和处理监测大数据时遇到了极大的困难。

鉴于高速光纤数据网和无线传输已在电力行业广泛普及，在下一代电力设备远程监测系统中，监测装置的数据处理和分析的大部分工作应当上移至监测中心，这样一方面可降低监测装置的资源配置，另一方面便于监测数据处理和分析软件的更新。下一代电力设备远程监测系统需要获取和传输数据的主流应当是原始监测数据，不仅包括设备异常时出现的各类异常报警数据和定时监测的数据，还应该有某些重要参数的连续监测数据，如发电机的振动信号、变压器和 GIS 设备的放电数据、以及一些设备的视频数据等。

综上所述，电力设备在线监测数据具备了大数据所拥有的体量大、类型多、变化快（动态）和价值密度低（大量数据涉及正常状态，有用数据少）的种种特征，适用于新兴的大数据存储与处理技术。

二、云计算技术在电力系统中的应用现状与问题分析

云计算作为一种新兴的计算模式，将数据存储和处理任务分布在由大量服务器所构成的资源池上，根据用户需求提供存储空间、计算能力以及信息服务[27]。云计算通过虚拟化、海量分布式数据存储、并行编程模型等技术，可以有效地解决海量数据的存储和大数据的并行计算问题。目前，云计算正在向行业应用发展。《中国云计算产业发展白皮书》[28]中指出，在未来几年，教育、医疗、电信、金融、政府、石油以及电力行业都将成为云计算应用的重点。

在众多云计算技术中，Apache Hadoop 项目[29]包含的 Hadoop 分布式文件系统[30]（Hadoop Distributed File System，HDFS）和并行编程框架 Hadoop MapReduce[31]专长于大数据的分布式存储和并行处理，适合运行“数据密集型”应用程序，目前已应用于 Facebook、雅虎等互联网公司的大数据处理中[32~34]，这为解决电力设备监测大数据存储与处理提供了一种新的思路，其优势和技术特点主要包括：

（1）Hadoop 非常适合对实时性要求不高的历史数据进行批量分析和计算。Hadoop 是典型的，具有代表性的大数据批处理系统，其 HDFS 文件系统提供了高可靠性和可方便横向扩展的存储能力，适合海量历史数据的可靠存储；Hadoop 提供的 MapReduce 并行技术适合对存储在 HDFS 上的历史数据进行的批量分析，如：数据清洗、格式转化、信号去噪、特征提取、模式识

别等。

（2）MapReduce 相对传统并行计算框架，如 MPI 等，简单易用，屏蔽了大量底层通信细节，使用户可以专注于系统业务逻辑开发。

（3）Hadoop 提供了完整的生态系统，为系统开发提供了多层次的支撑。Hadoop 提供的 HBase[35]非关系型数据库，适合存储结构化、半结构化以及非结构化数据，并提供在线查询的低延迟性能，非常适合电力设备监测数据（采样数据，时序波形信号）的存储和在线查询。在 MapReduce 上层，提供了 Hive[36]、Pig[37]等高级查询分析工具，支持使用类 SQL 语言进行历史数据的查询分析，比使用 MapReduce 编程更简洁。

虽然 Hadoop 是较为通用的平台，但在应用于电力设备监测系统时，仍有许多具体的应用问题需要考虑，包括：

（1）HDFS 存储数据时所采用的机架感知策略仅从提高可靠性的角度对多数据副本进行随机分布。在进行多源监测数据关联分析时，将相关的数据聚集在一起会引起数据节点间大量的通信，导致计算任务执行缓慢。站在应用层的角度考虑，监测数据之间可能具有较强的相关性，相关的数据会在同一个计算任务中使用，比如同一条输电线路上导线两端的张力、三相的电流值，具有较强的相关性。如果能够根据数据间的相关性设计数据分布策略，就有可能减少数据使用时在节点间的迁移，从而有助于提升计算性能。

（2）Hadoop 的分布式结构化数据表 HBase 采用“key-value”模式，按照主关键字对数据进行分布组织和查询处理。这种方法无法有效地支持多条件查询处理[38]。

（3）MapReduce 编程框架提供了简易、方便的并行程序开发接口，其并行模式是“数据并行”，而并非“功能并行”，需要根据具体计算任务的特点，分析其是否适合采用 MapReduce 实现并行。

（4）系统的可靠性、可用性和维护。自建的 Hadoop 平台大都构建在局域网内，且没有进行 Web Service 的封装，不能通过 Internet 访问；没有专人维护，停电、服务器宕机、硬盘故障、交换机宕机等各类硬件故障都会导致系统不可用。虽然 Hadoop 默认采用 3 副本策略进行数据备份，但自建系统规模较小，所有服务器均在同一个机架下，可靠性会大打折扣。

（5）Hadoop 只擅长对海量历史数据的批量分析。批处理任务执行过程中