智能电网
大数据云计算技术研究

周国亮 宋亚奇 朱永利 王桂兰 萨初日拉 著

清华大学出版社

智能电网
大数据云计算技术研究

周国亮 宋亚奇 朱永利 王桂兰 萨初日拉 著

清华大学出版社
北京

内 容 简 介

本书针对智能电网大数据云计算技术展开研究，通过提高数据利用率，探索数据驱动的电力系统各种挑战解决方案。重点研究了批处理、流式计算和内存计算等在智能电网大数据分析处理中的应用，探讨了电力系统全景实时大数据体系架构。

本书在深入分析智能电网大数据特点的基础上，开展了云计算与电力系统深度融合的研究工作，是云计算在电力系统中落地的尝试。本书可以作为电气工程类、计算机科学与技术类专业的研究生课外参考读物，也可作为相关专业工程技术人员、教师及科技工作者的参考。

图书在版编目(CIP)数据

智能电网大数据云计算技术研究/周国亮等著. --北京：清华大学出版社，2016
ISBN 978-7-302-43489-4

Ⅰ. ①智… Ⅱ. ①周… Ⅲ. ①互联网络－应用－智能控制－电网－研究 Ⅳ. ①TM76

中国版本图书馆 CIP 数据核字(2016)第 078303 号

责任编辑：张　弛
封面设计：何凤霞
责任校对：袁　芳
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京嘉实印刷有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：10.25　　**字　　数**：247 千字
版　　次：2016 年 9 月第 1 版　　**印　　次**：2016 年 9 月第 1 次印刷
定　　价：39.00 元

产品编号：069110-01

FOREWORD 前言

近年来，随着智能电网研究和建设的不断推进、各种分布式可再生能源的大规模应用等，电网规模不断扩大、结构日趋复杂，对电网监测的广度不断扩大、深度不断加强，而对大规模电网的全方位、多尺度感知将使电力系统运行和监测过程中收集的数据量呈指数级增长，并逐渐形成了电力大数据。通常情况下，电力大数据具有如下特点。

（1）体量巨大。从TB级别跃升到PB级别，广域向量测量系统（Wide Area Measurement System，WAMS）和数据采集与监视控制系统（Supervisory Control and Data Acquisition，SCADA）遥测点增加、采集频率提高，每年将产生超过TB级的数据；用电信息采集系统成为世界上最大的自动化计量系统，连接着以亿计算的用户，随着采集频率提高，数据量将以PB计算。

（2）多源异构特性。电力系统本身结构复杂、规模庞大，采集的数据具有明显的多源异构特性。同时也需要对电网系统外数据（气象、地理、环境等）与内部数据进行关联分析。比如与电力设备状态评估相关的数据来源广、种类多，包括在线监测实时数据、设备台账信息、预试数据以及音视频和气候环境等非结构化数据。

（3）生成速度快。由于电力系统的特殊性，在某些场景下数据采样频率很高，数据生成速度非常快。分布式能源随气候环境动态变化，要求快速准确预测变化，需要对设备和环境实时监控。在SCADA调度系统中，每分钟产生的数据量也将达到GB级。

（4）价值密度低。以视频数据为例，连续不间断监控过程中，可能有用的数据仅仅有1～2s。在基于经验和人工的传统输变电设备状态监测评估中，只对少量异常数据关注、处理和采用，而丢弃所谓“正常数据”，然而大量的正常数据也可能成为故障分析判断的重要依据。

基于电力大数据的特点，深入探讨数据驱动的电力系统各项技术，提高数据利用率是当前智能电网建设过程中必须要面对的问题。云计算技术是处理分析大数据的有效方式，具有良好的可扩展性和容错机制，在商业互联网领域应用广泛。随着电力大数据的形成，探讨基于云计算技术的电力大数据分析处理，进而对提高电力系统的整体安全性和可靠性具有重要的研究价值。然而，由于电力系统运行模式与商业互联网企业相比具有自身的特点和性能要求，当云计算应用于生产运行数据时具有很大的挑战性，本文正是基于此背景展开相关研究和探讨。

当前云计算技术包括批处理计算、流式计算和内存计算等三种方式，分别适应于不同的

数据处理类型。其中，批处理适合离线静态数据分析，对运行时间要求较低；而流式计算面对流动的实时数据动态计算，对时限性要求较高；内存计算是针对批处理频繁磁盘操作性能瓶颈，适应需要多次迭代的机器学习类算法。另外，针对电力大数据的多源异构特点，需要开展数据融合分析技术，为用户提供统一综合的查询视图。综上所述，本书主要开展了如下几项研究工作。

（1）智能电网大数据批处理技术模式研究与应用（第1章）

利用收集的静态离线电力大数据，开展了电力短期负荷预测、绝缘子泄漏电流数据高效存储和聚类划分及局部放电信号处理的并行EEMD算法研究。主要采用的技术是Hadoop的MapReduce技术，通过并行计算和大数据的支撑，提高数据处理的效率。

（2）智能电网大数据高性能内存计算技术研究（第3章）

针对电力系统中部分应用需要高性能计算分析的特点，开展内存计算技术在电力系统中的应用研究。重点对监测数据尤其是在线监测领域，监测数据速度快的特点，研究高效的内存聚类算法。另外，对电力数据高级分析的OLAP技术开展了基于内存计算的并行方体计算技术研究。

（3）智能电网大数据的多源数据融合分析技术（第4章）

针对智能电网大数据多源异构的特点，研究数据融合技术，为用户提供统一的查询或显示视图。重点研究内容包括多通道数据融合特征提取和多数据源的连接技术，并结合Hadoop和Spark研究并行计算模式和高速融合算法。

（4）智能电网大数据的流式计算研究（第5章）

针对智能电网大数据速度快，多种场景下表现出流数据的特征，研究流式计算技术动态处理数据，应用在在线监测和实时用电信息采集系统中。研究了监测流数据实时过滤和检测技术，通过设计基于工作流的处理模型，快速实现对大规模流数据的监测，实时发现异常，提高监测效率。针对大规模用电信息采集系统，设计流式计算的聚类算法，完成数据聚类划分，实时发现用电行为异常，从而提升数据分析技术与生产系统的融合水平。

（5）全景实时分析平台关键技术探索（第6章）

智能电网最终目标的实现需要借助全景实时分析平台的支撑，将大数据平台变为智能电网的神经中枢，促进大数据与电力系统的深度融合，从数据驱动的角度研究当前电力系统面临的挑战问题。探讨基于内存计算、实时流式大数据处理技术、大规模并行计算技术及列存储等在电力大数据实时分析中的应用，满足电力系统对大数据分析的时效性需求；结合主流开源大数据处理技术，设计了电力大数据分析平台的分层体系架构，为电力系统的高效运行提供保证。

（6）能源互联网中的大数据（第7章）

能源互联网是实现广域范围内的能源分享和支持分布式清洁能源为主体，将涵盖大规模的分布式能源、交通网络、天然气网络等，而支撑能源互联网高效运行和实时能源分享的核心技术应包括大数据分析技术。探讨了通过大数据提高能源互联网的安全等级和高效实时调度技术。

本书由国网冀北电力有限公司技能培训中心周国亮统稿，并负责第1、2、5、6、7、8章的编写，华北电力大学宋亚奇和王桂兰分别负责第3、4章的编写，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河北省自然科学基金(F2014502069)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,没有你们认真负责的工作,此书不可能完成。感谢本书引用中所涉及的各位学者、专家。本文引用了多位学者的研究文献,如果没有各位学者研究成果的帮助和启发,我们将很难完成本书的内容。感谢为本书出版做出辛勤工作的出版社同志,没有你们的专业劳动,展现在读者面前的内容会很凌乱,很难示人。感谢阅读此书的每一位读者,是你们的阅读才使我们的工作有意义,谢谢你们。

由于我们的学术水平、工程经验有限,对所研究内容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和指正!我们的邮箱是 yu_bing_2000@163.com,谢谢!

周国亮

2016 年 3 月 28 日

CONTENTS 目录

第1章

智能电网大数据现状及挑战

1.1 智能电网

智能电网通俗地讲是指电网的智能化或智能电力，也被称为“电网2.0”，它是建立在集成的、高速双向通信网络的基础上，通过先进的传感和测量技术、先进的设备技术、先进的控制方法以及先进的决策支持系统技术的应用，实现电网的可靠、安全、经济、高效、环境友好和使用安全的目标，其主要特征包括自愈、激励，用户抵御攻击、提供满足高要求的用户电能质量、支持大规模分布式等各种不同发电形式的接入、启动电力市场以及资产的优化高效运行等。建设智能电网将有效促进世界经济社会发展，并更好地应对全球气候变化和能源危机，对促进世界经济社会可持续发展具有重要作用。智能电网的作用主要表现在如下5个方面。

(1) 促进清洁、可再生能源的开发利用，减少温室气体排放，推动低碳经济社会发展。

(2) 优化能源结构，实现多种能源形式的互补，确保能源供应的安全稳定，减少对化石能源的依赖程度。

(3) 有效提高能源输送和使用效率，增强电网运行的安全性、可靠性和灵活性，促进在更大范围内的能源动态平衡。

(4) 推动相关领域的技术创新，促进装备制造和信息通信等行业的技术升级，扩大就业，促进社会经济可持续发展。

(5) 实现电网与用户的双向互动，创新电力服务的传统模式，为用户提供更加优质、便捷的服务，提高人民生活质量。

随着智能电网的发展，电网功能逐步扩展到促进能源资源优化配置、保障电力系统安全稳定运行、提供多元开放的电力服务、推动战略性新兴产业发展等多个方面。作为我国重要的能源输送和配置平台，智能电网从投资建设到生产运营的全过程都将为国民经济发展、能源生产和利用、环境保护等方面带来巨大效益，具体表现在如下几个方面。

(1) 在电力系统方面：可以节约系统有效装机容量；降低系统总发电燃料费用；提高

电网设备利用效率，减少建设投资；提升电网输送效率，降低线损。

（2）在用电客户方面：可以实现双向互动，提供便捷服务；提高终端能源利用效率，节约电量消费；提高供电可靠性，改善电能质量。

（3）在节能与环境方面：可以提高能源利用效率，带来节能减排效益；促进清洁能源开发，实现替代减排效益；提升土地资源整体利用率，节约土地占用。

（4）其他方面：可以带动经济发展，拉动就业；保障能源供应安全；变输煤为输电，提高能源转换效率，减少交通运输压力。

智能电网的核心技术包括如下几个方面。

（1）发电领域：主要包括大规模可再生能源、分布式能源、光伏发电等电源的接入和协调运行技术。

（2）输电领域：主要包括大电网规划技术、电力电子技术、输电线路运行维护技术、输电线路状态检修技术和设备全寿命周期管理技术等。

（3）调度领域：主要包括大电网安全稳定分析与控制技术、经济运行技术、综合预警和辅助决策技术、安全防御技术等。

（4）变电领域：主要包括变电站信息采集技术、智能传感技术、实时监测与状态诊断技术、自适应保护技术、广域保护技术、智能电力设备技术等。

（5）配电领域：主要包括配电网安全经济运行与控制、电能质量控制、智能配电设备研究、大规模储能、电动汽车变电站等技术。

（6）用电领域：主要包括高级量测技术、双向互动营销技术、用户储能技术、用户仿真技术等。

综上所述，智能电网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。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，以及累积的海量历史多源异构数据快速分析。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，因此需要相应的存储和快速处理技术作为支撑。

1.2 云计算与大数据

大数据可以通俗地理解为无法在一定时间内用传统数据库软件工具对其内容进行抓取、管理和处理的数据集合。根据国际数据公司（International Data Corporation，IDC）的监测统计，即使在遭遇金融危机的2009年，全球信息量也比2008年增长了62%，达到80万PB（1PB=2^{20}GB），到2011年全球数据总量已经达到1.8ZB（1ZB=2^{40}GB），并且以每两年翻一番的速度飞速增长，预计到2020年全球数据量总量将达到40ZB。在数据规模急剧增长的同时，数据类型也越来越复杂，包括结构化数据、半结构化数据、非结构化数据等多种类型，其中采用传统数据处理手段难以处理的非结构化数据已接近数据总量的75%。

鉴于大数据分析技术在经济、社会的应用和潜在的巨大影响，很多国家都将大数据视作战略资源，并将大数据应用提升为国家战略。2012年3月，美国奥巴马政府宣布推出“大数据的研究和发展计划”。2012年9月，日本总务省发布2013年行动计划，提出以复苏日本

为目的推进“活跃在ICT领域的日本”ICT综合战，明确提出“通过大数据和开放数据开创新市场”。2013年2月，法国政府发布了《数字化路线图》，列出了5项将会大力支持的战略性高新技术，“大数据”就是其中一项。2013年1月，中国国家能源局下发了《关于数据中心建设布局的指导意见》，国家发展和改革委员会与中科院正式启动基础研究“大数据服务平台应用示范”项目；2013年3月，中国电机工程学会发布了《中国电力大数据发展白皮书》；2013年年初，贵州省发布《贵州“云计算”战略规划》；2013年10月，中国国内领先水平的大规模云计算数据中心、云计算研发应用示范基地——中国电信云计算贵州信息园在贵阳正式开工建设。这些实例进一步说明了大数据应用的重要性，未来大数据可能成为国家创新能力和竞争力的重要体现。

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。淘宝公司通过对大量交易数据变化分析，可以提前6个月预测全球经济发展趋势。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》，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。

目前，大数据应用已在社会经济活动方面展示出巨大的价值和潜力，在电力行业也有成功的应用范例。丹麦的维斯塔斯风力技术集团（Vestas Wind System A/S），通过在世界上最大的超级计算机上部署国际商业机器公司（International Business Machines Corporation，IBM）大数据解决方案，得以通过分析包括PB量级气象报告、潮汐相位、地理空间、卫星图像等结构化及非结构化的海量数据，从而优化风力涡轮机布局，提高风电发电效率。这些以前需要数周时间完成的分析工作现在只需不到1小时即可完成。美国的Space-Time公司2011年利用大数据可视化技术为美国加州独立系统运营商设计了一套实时监控电力传输系统能源基础设施的可视化软件Space-Time Insight，该系统可实时监测25000km的输电线路状况，可根据发生问题的严重性和临近地区的反应及时做出决策，保障电网的安全运行。中国国家电网所属的国家冀北电力有限公司，正在使用智慧风能解决方案来整合可再生能源并入所属电网，通过使用IBM风力预测技术，张北项目的第一阶段目标，旨在增加10%的可再生能源的整合发电量。通过分析提供所需的信息，将使能源电力公司得以减少风能并网的限制，进而更有效地使用已产出的能源，强化电网的运行。这种大数据的应用实践对中国电力大数据分析展示乃至整个能源相关行业都具有巨大的参考价值。应对大数据处理分析的有效技术方式是云计算技术。

云计算（Cloud Computing）是基于互联网的计算存储服务的增加、使用和交付模式，通常涉及通过互联网提供动态易扩展且通常是虚拟化的资源，是应对当前大数据挑战的有效方式。云是对网络或互联网的一种比喻说法。过去在图中往往用云表示电信网，后来也用于表示互联网和底层基础设施的抽象。云计算可以让用户体验每秒10万亿次的运算能力，拥有这么强大的计算能力可以模拟核爆炸、预测气候变化和市场发展趋势。用户通过计算机、笔记本、手机等方式接入数据中心，按自己的需求进行运算。

现阶段广为接受的云计算定义是美国国家标准与技术研究院（National Institute of Standards and Technology，NIST）提出的：云计算是一种按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问，进入可配置的计算资源共享池（资源包括网络、服务器、存储、应用软件、服务等），这些资源能够被快速提供，只需投入很少的管理工作，或与服务供

应商进行很少的交互。

当前，被普遍接受的云计算特点如下所述。

(1) 超大规模

“云”具有相当的规模，Google 云计算已经拥有 100 多万台服务器，Amazon、IBM、微软、Yahoo 等的“云”均拥有几十万台服务器。企业私有云一般拥有数百上千台服务器。“云”能赋予用户前所未有的计算能力。

(2) 虚拟化

云计算支持用户在任意位置、使用各种终端获取应用服务。所请求的资源来自“云”，而不是固定的有形的实体。应用在“云”中某处运行，但实际上用户无须了解、也不用担心应用运行的具体位置。只需要一台笔记本或者一个手机，就可以通过网络服务实现我们需要的一切，甚至包括超级计算这样的任务。

(3) 高可靠性

“云”使用了数据多副本容错、计算节点同构可互换等措施保障服务的高可靠性，使用云计算比使用本地计算机可靠。

(4) 通用性

云计算不针对特定的应用，在“云”的支撑下可以构造出千变万化的应用，同一个“云”可以同时支撑不同的应用运行。

(5) 高可扩展性

“云”的规模可以动态伸缩，满足应用和用户规模增长的需要。

(6) 按需服务

“云”是一个庞大的资源池，你按需购买；云可以像自来水、电、煤气那样计费。

(7) 极其廉价

由于“云”的特殊容错措施，可以采用极其廉价的节点构成云，“云”的自动化集中式管理使大量企业无须负担日益高昂的数据中心管理成本，“云”的通用性使资源的利用率较之传统系统大幅提升，因此用户可以充分享受“云”的低成本优势，经常只要花费几百美元、几天时间就能完成以前需要数万美元、数月时间才能完成的任务。云计算可以彻底改变人们未来的生活，但同时也要重视环境问题，这样才能真正为人类进步做贡献，而不是简单的技术提升。

(8) 潜在的危险性

云计算服务除了提供计算服务外，还必然提供了存储服务。但是云计算服务当前垄断在私人机构(企业)手中，而他们仅能够提供商业信用。对于政府机构、商业机构(特别像银行这样持有敏感数据的商业机构)选择云计算服务应保持足够的警惕。一旦商业用户大规模使用私人机构提供的云计算服务，无论其技术优势有多强，都不可避免地让这些私人机构以“数据(信息)”的重要性挟制整个社会。对于信息社会而言，“信息”是至关重要的。另一方面，云计算中的数据对于数据所有者以外的其他用户云计算用户是保密的，但是对于提供云计算的商业机构而言确实毫无秘密可言。所有这些潜在的危险，是商业机构和政府机构选择云计算服务、特别是国外机构提供的云计算服务时，不得不考虑的一个重要前提。

在智能电网运行过程中，大数据产生于整个系统的各个环节。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。通过对数据分析可以更好地理解电力客户的用电行为、合理地设计电力需求响应系统和进行短期负荷预测等，从而有利于电网的规划和运行。

在智能电网中，随着高压、特高压电网及配电自动化建设的不断推进，智能化设备及系统应用数量不断增长，电网设备的部署结构与产生的数据日益复杂庞大。一方面，设备的自身状态和外部环境都会影响系统的运行，迫切需要对输变电设备负载能力、运行状态进行动态评估，以降低故障发生概率及相关风险，减少设备运行维护成本，提高设备净资产收益率；另一方面，由于智能输变电设备数量的不断增长，电网中获取与传输的各类数据也在发生几何级数的增长。这些数据不仅包括设备异常时出现的各类故障信号，运行过程中设备的各类状态信息，同时还包含了大量的相关数据，如地理信息、气象、视频图像、设备台账、实验数据与文档等。如何将这些多源异构高维的数据资源进行统一的收集、过滤与处理，并对现有的设备状态检测方案进行优化成为新的挑战。此外，基于因果关系的传统设备状态评价方法着眼点为单一设备和少量异常数据，难以实现对大量“数据资产”的综合有效利用以及面向整个电网的准确状态评估和风险预测。

鉴于大数据在电力系统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。本节试图将智能电网中大数据的研究和应用现状及挑战进行综述，并给出了智能电网大数据存储与分析系统的一种可选的框架。

1.3 智能电网大数据及特点

1.3.1 智能电网中的大数据

电网业务数据大致分为三类：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，随着信息技术发展，这部分数据增长很快。但由于数据库存储容量的限制，数据会定期更新，一般只存储最新的数据。相对于结构化数据而言，不方便用数据库二维逻辑表表现的数据即称为非结构化数据，主要是包括视频监控、图形图像处理等产生的数据等。这部分数据增长非常迅速，据 IDC 的一项调查报告中指出，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％。在电力系统中，非结构化数据占智能电网数据很大比重，这部分数据增长速度也很快，对电网数据中心的存储压力很大。

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。数据依据时限要求不同可以采取不同的处理方式，比如实时数据采用流式内存计算方式，而准实时数据可以采用批处理方式。

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的实时状态数据及时获取和快速分析处理，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。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，特高频检测需要吉赫兹的采样率。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) 为了真实而完整地记录生产运行的每一个细节，完整地反映生产运行过程，要求达到“实时变化采样”，实现对设备的全生命周期管理和实时状态评估。

同时，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，海量的过程数据被保存。这些数据中蕴藏丰富的信息，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。基于数据驱动的故障诊断方法被提出，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，需要对分布广泛和大量的分布式能源进行实时监测和控制。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。

(2) 输变电侧：在 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。相量监测只是智能电网监控的一小部分，电网中还包括其他大量需要高采样监测的设备。

(3) 用电侧：为了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，这些电表可以每隔五分钟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。国家电网公司也建成了包含上亿用户的自动化采集系统。

电动汽车的无序充放电行为会对电网运行带来麻烦，如果能合理地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汽车电池的充放电状态进行监测，并合理制定调度规则，而这也会产生大数据，需要大数据处理技术的支撑。

1.3.2 智能电网中大数据的特点

与互联网中的大数据相似，智能电网中的大数据也具备“4V”特征，即规模大(Volume)、类型多(Variety)、价值密度低(Value)和变化快(Velocity)。

1. 规模大

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3～4s计算，每年产生1.03TB的数据（1.03TB＝12字节/帧×0.3帧/s×10000遥测点×86400秒/天×365天）；广域向量测量系统（Wide Area Measurement System，WAMS）10000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。某省级电力公司已有数字化变电站178座，每天产生的数据量约为700TB。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. 类型多

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百千赫兹，甚至吉赫兹。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

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一两秒。在输变电设备状态监测中也存在同样问题，所采集的绝大部分数据都是正常数据，只有极少量的是异常数据，而异常数据是状态检修的最重要的依据。以视频数据和高压设备放电波形为例，连续不间断监测过程中，基于经验和人工的传统输变电设备状态监测评估方法可能只对小段时间（如1～2s）的数据和少量异常数据予以关注、处理和分析，而丢弃大量所谓的“正常数据”，对这些正常数据的深度挖掘也可能为故障分析提供重要的线索和依据。

4. 变化快

在几分之一秒内对大量数据进行分析，以支持决策制定。这种在线的流数据分析与挖掘同传统的数据挖掘技术有着本质的不同。输变电设备状态评估和风险预测要求对大量数据进行及时分析并做出判断，以支持生产调度决策的制定，对在线状态数据的处理性能要求远高于离线数据。尤其对输变电设备状态监测系统，在极端（天气、故障发展阶段等）情况下，大规模报警数据会在短时间内以“井喷式”的方式产生与传输，并要求实时处理，将对信息处理系统的吞吐技术提出挑战。SCADA采集了大量的电压、电流、开关状态等电网稳态数据。常规SCADA系统10000个遥测点，按采样间隔3～4s计算，每年产生1.03TB数据（1.03TB＝12字节/帧×0.3帧/s×10000遥测点×86400s/天×365天），目前三华的数据量每日65.3万条，7.58GB；WAMS 10000个遥测点，采样率100次/s，按上述公式计算，每年产生495TB的数据。

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。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