电力大数据应用现状及前景

文/南瑞埃森哲信息技术研究中心　张　沛

20世纪90年代，“大数据”概念被提出，最初只是对一些在一定时间内无法用传统方法进行抓取、管理和处理的数据统称。2009年以来，互联网数据以每年50%的速度增长，整个互联网领域开始意识到大数据时代来临。

2011年5月，麦肯锡全球研究机构发布了《大数据：创新、竞争和生产力的下一个前沿领域》报告，大数据概念在计算机行业迅速火热起来。2012年1月，瑞士达沃斯论坛上《大数据，大影响》(Big Data，Big Impact)报告称，数据如同货币或黄金一样，已成为一种新经济资产类别。显然，人们对大数据的关注度正在逐步提升。

电力大数据的概念和特征

当前，大数据概念已基本达成共识，但未形成统一定义。麦肯锡认为，大数据是“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”；Gartner认为，大数据是“需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产”。

2013年3月，中国电机工程学会针对目前电力企业和电力行业数据的状况，发布了《中国电力大数据发展白皮书》。电力行业的信息时代正处于关键转折点，随着智能变电站系统、现场移动检修系统、测控一体化系统、GIS系统和智能表计等的建设，以往数据类型较为单一、增长较为缓慢的情况将发生转变，将逐渐步入到由复杂及异构数据源广泛存在和驱动的时代。

电力大数据是指通过传感器、智能设备、视频监控设备、音频通信设备和移动中断等各种数据采集渠道收集到的，结构化、半结构化和非结构化的海量业务数据的集合。

电力大数据的特征。数据量在TB PB以上；在速度上，持续实时产生数据，要求即时处理。如调度部门的大部分数据及营销用采数据都是时时数据流；在价值方面，通过数据挖掘等一系列手段，使电力企业实现业务趋势预测和分析决策。

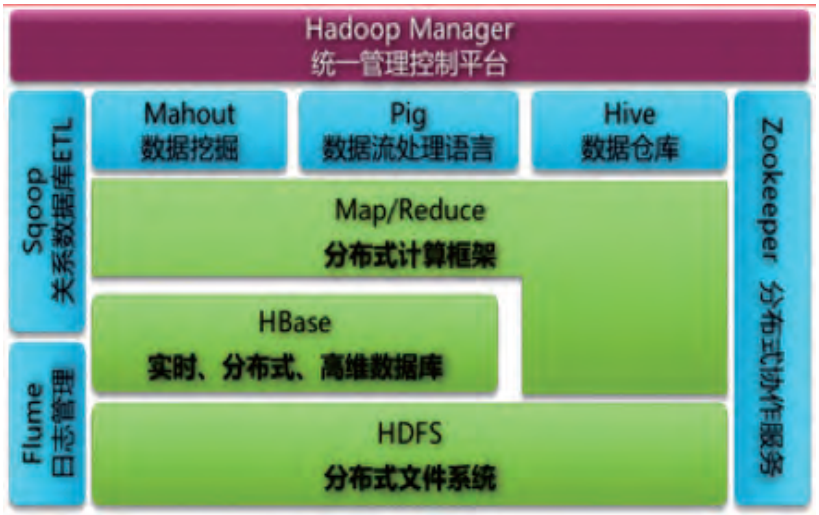
大数据的关键技术

Hadoop致力于“大数据”处理的关键技术架构。Hadoop结构中HDFS作为分布式文件系统负责非结构化数据存储；Hbase作为数据库软件对实时、分布式和高维数据进行存储和管理；Map/Reduce是分布式计算框架；Mahout、Pig和Hive等是封装挖掘、流处理等算法的处理工具。主要采用Java编程语言，可运行于Linux、Mac OS/X和Solaris等系统环境。

Hadoop特点是能够轻松扩展到PB级别的数据存储，处理规模；带有容错功能的并行处理架构；基于普通的X86平台硬件架构，硬件成本低廉；用内置格式存储/处理数据；基于开源项目，并且传统厂商也日益重视对其的支持，已成为重要的并行处理架构标准之一。

在Hadoop出现之前，处理大数据需要在Pig上进行，而利用Hadoop插挂硬件的结构可以轻松达到Pig的规模，这对于电力企业而言，无疑在很大程度上降低了成本。但是，Hadoop并不是解决大数据问题的唯一方式。业界大数据厂商目前的技术思路多为将原有的数据库产品同Hadoop、内存数据库等大数据技术产品相结合，形成统一的数据处理解决方案。具体表现为使用数据仓库处理结构化数据，大数据分析平台处理异构数据挖掘，从而提升系统整体效能。研发重点在数据集成管理和数据分析处理方面，具体表现为数据仓库产品与大数据平台产品的交互，以及面向大数据应用进行产品定制化研发及性能提升。

如Microsoft公司大数据连接器技术、NoSQL数据库技术、流计算、内存数据处理、数据处理与硬件协同技术，可将结构化数据和非数据进行转换，这种转换可以处理数据。ORACLE公司大数据连接器技术、数据批处理与流计算、NoSQL数据库技术、MapReduce并行框架、数据处理与硬件协同，能够提供硬件的完整解决方案。SAP公司大数据连接器技术、内存计算技术、行存储和列存储的混合模式、数据流式计算、NoSQL数据库技术、数据压缩与虚拟建模、数据处理与硬件协同。TERADATA大数据连接器技术、SQL/MR技术、NoSQL数据库技术、数据挖掘技术集成、数据处理与硬件协同，统一数据架构平台针对不同数据有不同的存储方式。



电力行业数据管理分存储层、整合层、计算层和应用层四个层次。

存储层的关键技术有列数据库、键值数据库、内存数据库和HDFS。技术特征是以列相关存储架构进行数据存储的数据库，主要适合于批量数据处理和即席查询；基于Key/Value存储架构，优势在于简单、易部署；基于内存体系结构设计，提供数据共享和高速计算；有效整合分布式环境下的存储资源，提供统一存储服务管理。

整合层的关键技术有流数据总线、大数据连接器、PIG和HIVE。技术特征是提供分布式、可靠和高可用的海量流数据聚合服务，支持定制各类流数据发送方；关系型数据库与分布式数据库的交互接口，提供单一可靠的数据视图；提供描述数据流的类SQL语言，为海量数据并行计算提供简单的操作和编程接口；是一个数据仓库框架，在HDFS之上提供类SQL处理语言，适合数据仓库的统计分析。

计算层的关键技术有并行数据处理和流式计算。技术特征是提供分布式计算框架，采用简化的模型，更易于编程人员理解；适用于海量流式数据的处理和实时响应。

应用层则的关键技术产有自然语言处理、图像识别和数据挖掘。技术特征是研制能有效地实现自然语言通信的计算机系统；

对图像、视频数据进行特征检索和运动跟踪，能够有效辅助人工监控作业；从数据库的大量数据中揭示出隐含的、先前未知的并有潜在价值的信息。提供分布式计算框架，采用简化的模型，更易于编程人员理解；适用于海量流式数据的处理和实时响应。

国内电力企业数据现状和需求

国内电力企业数据存储现状。随着智能电网、三集五大两中心业务建设推进，产生了大量的数据。目前主要通过关系型数据来存储、计算。GB级电网规划信息数据和计划类数据；每月约100GB结构化、300GB非结构化建设管理数据；每个网省每年约7亿条/6T电网电压、电流等电网运行数据；生产设备状态数据量大，视频数据尤其巨大，达到PB级；每年新增约90TB营销用电采集数据；：3TB客户档案及交易数据，日增4GB客服音频数据；运监接入明细数据后，网省结构化、非结构化和空间地理信息达到TB级。

目前业务数据主要采用关系数据库，如ORACLE、达梦、金仓进行存储，不能满足智能电网条件下大量传感器对数据I/O吞吐量的要求。对此，采用分布式存储架构，适应数据量由TB级向PB级发展，实现数据高性能读写、存储和高可扩展性。

从数据类型上看，除传统的结构化数据外，还产生了系统日志、表计等半结构化数据和视频监测、客服音频等非结构化数据，对于这些非结构化数据，多数保存在本地系统中，且不能被检索分析，缺乏对其进行数据管理的手段。目前对半结构化、音频、视频和图像数据大多采用文件存储，未实现对文档、音频及视频等非结构化文件的检索，以及对混合数据类型的分析挖掘。因此，需要采用自然语言识别、视频分析等技术实现对非结构化数据向结构化数据的转换，实现对多类型数据的全景分析。

从数据处理速度上看，目前分析功能和辅助决策功能大体上能满足业务需要，但随着数据量的不断增加，一些分析性能问题也逐渐显现。

采集及分析用电信息目前只能采集96点数据，每天入库一次；对用电信息数据分析只能实现按天统计。采用队列对数据进行缓存，分批写入数据；采用离线分析模式，后台存储过程离线运行。对此，需要采用分布式存储及并行计算技术提高数据入库及分析速度。

生产管理对设备只能判断当前状态信息，无法实现大范围、长周期设备状态分析及统计。在关系数据库中对设备信息采用基础数据加规则匹配模式进行数据分析。因此，需要采用分布式存储，流式计算等技术对设备状态及视频流式监测数据进行分析。

从从数据价值挖掘上看，对数据利用的手段还主要停留在基于报表的统计分析，缺乏对数据进行挖掘和探索的高级分析手段，制约了从数字化向智能化的发展。目前的普遍情况是数据主要停留在对结构化数据进行指标的统计分析阶段，对单业务的分析较好，对跨业务的分析较弱，对数据挖掘和探索方面还停留在浅层学习阶段，数据资产价值体现上整体停留在粗放型阶段。面对这些问题，企业应提供多样化统计分析和数据挖掘手段，增强关联度和预测性分析，发现数据潜藏价值，服务公司战略决策、业务应用、管理模式创新。

大数据应用现状和在电力行业应用场景

大数据应用涉及到电力企业的各个业务领域。在规划领域，通过对用电采集大数据的分析，利用数据挖掘技术，更准确地掌握用电负荷分布和变化规律，提高中长期负荷预测准确度。在建设业务中，通过对现场照片进行批量比对分析，利用分布式存储、并行计算、模式识别等技术，掌握施工现场的安全隐患，或者对安全整改措施落实情况进行核查。在运行领域，利用机器学习、模式识别等多维分析预测技术，分析新能源出力与风速/光照/温度等气象因素的关联关系，更准确地对新能源发电能力进行预测和管理。在检修领域，通过研究消缺、检修、运行工况、气象条件等因素对设备状态的影响，以及设备运行风险水平，利用并行计算等技术实现检修策略优化，指导状态检修的深入开展……

1.国外电力行业大数据应用经验

法国电力公司基于大数据的用电采集应用。法国目前已安装3 500万个智能电表，智能电表采集的主要是个体家庭的用电负荷数据。以每个电表每10 min抄表一次计算，3 500万智能电表每年产生1.8万亿次抄表记录和600 TB压缩前数据，电表产生的数据量5～10年内达到PB级。

针对这一情况，法国电力公司的研发部门成立了Big Data项目组，对数据进行挖掘分析，从而实现负荷曲线数据能够进行高速处理，短期用户用电趋势能被预测。通过借助大数据技术研

究海量数据的处理架构，形成能够支撑在规定延迟内的复杂、并行处理能力；可以在不同尺度上进行处理，某些应用实现了实时处理；实现了电网调度等高级应用（电网状态监测、电网自动愈合）；对电网调度进行局部优化；用电需求侧管理，实现了实时电价及电网的可再生能源的接入。

丹麦维斯塔斯基于大数据的数据实时处理平台的经验值得借鉴。该企业在全球65个国家，安装了43 000台风力发电机。这些风力发电机的位置选择直接关系到发电能力和投资回报，因此安装位置时要考虑温度、风向、风力和湿度等因素。

在过去十年中，丹麦维斯塔斯已安装的风力发电机及其他收集到的环境信息，遍及全世界，已累积2.6 PB气象数据。面对这一庞大数据，通过现有方案无法及时处理。

丹麦维斯塔斯采用了IBM的BigInsights大数据平台解决了海量数据分析与处理问题，优化风力涡轮机配置方案，从而实现最高效的能量输出。该项目对天气建模以优化风力发电机的放置，最大限度提高发电量并延长设备使用寿命；将确定风力发电机的位置所需的时间从几周缩短为几小时；纳入2.5 PB的结构化和半结构化信息流，预计数据量将增长到6 PB。

2.电力大数据应用场景

目前，电力大数据应用场景主要在以下方面：

规划，提升负荷预测能力。通过对用采大数据的分析，利用数据挖掘技术，更准确地掌握用电负荷分布和变化规律，提高中长期负荷预测准确度。

建设，提升现场安全管理能力。对现场照片进行批量比对分析，利用分布式存储、并行计算、模式识别等技术，掌握施工现场的安全隐患，或者对安全整改措施落实情况进行核查。

运行，提升新能源调度管理能力。利用机器学习、模式识别等多维分析预测技术，分析新能源出力与风速/光照/温度等气象因素的关联关系，更准确地对新能源发电能力进行预测和管理。

检修，提升状态检修管理能力。研究消缺、检修、运行工况、气象条件等因素对设备状态的影响，以及设备运行风险水平，利用并行计算等技术实现检修策略优化，指导状态检修的深入开展。

营销。提升用电行为分析能力。扩展用电采集的范围和频次，利用聚类模型等挖掘手段，开展用电行为特征深入分析，并实施区别化的用户管理策略。

运监。提升业务关联分析能力。利用流式计算、可视化和并行处理等技术，实现全方位在线监测、分析、计算，通过聚类和模式识别技术解决跨业务关联分析，数据因子分析、数据诊断规则和算法提高数据质量监控和治理。

客服，提升服务效率。对客服录音进行实时监管，利用模式识别、机器学习等技术，对热点问题的服务资源进行优化分配，提升交互水平。

从需求、技术、实施、环境和价值五个成熟度对以上五大两中心业务进行评估，得出检修、营销、客服中心和运监中心的综合评分较高。

结束语

电力企业已经进入大数据时代，电力大数据利用大数据存储、整合、计算和应用四类核心技术，驱动电力公司信息技术平台和业务应用的升级与改造，扩展对数据的容纳和处理能力，填补在非结构化数据分析与利用、海量数据挖掘等领域的空白，提升电力公司在数据资源价值挖掘的整体水平，从而促进业务管理向着更精细，更协同，更敏捷，更高效的方向发展。

目前，电力大数据仍处于前期研究阶段，需要电力企业、生产厂商、学术组织和研究机构共同致力大数据关键技术及在电力行业的应用研究和开发。我们坚信，大数据将为电力行业带来显著的价值，甚至其发展模式将由业务驱动向数据驱动转变。EA

（本文转载自《动力与电气工程师》）