

智能电网数据处理关键技术研究及应用^①

任 萱¹, 程伟华², 陆 琳³

¹(江苏省电力公司 镇江供电公司, 镇江 212002)

²(江苏电力信息技术有限公司, 南京 210024)

摘 要: 随着配电网规模和数据采集终端数量的增大, 数据规模呈现出爆炸式增长态势. 能否高效的处理这些电网数据, 已经成为电力部门面临的巨大挑战之一. 从技术的角度提出适合智能电网数据信息处理的整体框架, 该架构分为四个层次: 数据采集、数据存储、数据处理和信息展示. 在此基础上, 详细阐述各个层次功能的实现技术与效果. 最终实现了智能电网海量数据处理在江苏电力公司的应用.

关键词: 智能电网; 数据处理; 关键技术

Key Technology of the Smart Grid Data Processing Research and Application

REN Xuan¹, CHEN Wei-Hua², LU Lin³

¹(Zhenjiang Power Supply Company of Jiangsu Electric Power Company, Zhenjiang 212002, China)

²(Jiangsu Electric Power Information Technology Co. Ltd, Nanjing 210024, China)

Abstract: Because of the large volume of data acquisition terminal and the rising of the network scale in the smart grid, data size is growing explosively. How to deal with the data efficiently, has become one of huge challenges that the power sector faces. This article proposes overall framework from the Angle of technology which is suitable for the development of data information processing of smart grid. The architecture is divided into four aspects: data acquisition, data storage, data processing and data visualization. Based on that, the paper introduces the techniques and effect of all aspects. Finally the massive data processing in application of smart grid has been achieved in Jiangsu Electric Power Company.

Key words: smart grid; data processing; Key technology

智能电网(Smart Grid)自 2003 年由美国提出以来, 得到大量企业和研究机构的广泛关注, 智能电网技术不断得到丰富和完善. 目前, 虽然未对智能电网形成统一的定义, 但是, 有一点已经达成共识, 即: 智能电网试图利用先进的数据处理和通信技术对传统电网的各个环节升级改造, 使之更加安全、可靠、环保、经济、高效, 实现电网与用户之间的电能和信息的双向流动^[1,2].

从技术角度看, 传感器网络和云计算技术的快速发展为智能电网数据采集和处理提供了有力的前端装置和存储计算能力支持. 依托这些先进的信息技术来处理智能电网中的数据, 为网络集中监控、信息资源

共享、分析决策和指挥调度提供一体化平台成为技术发展融合的主流趋势. 本文基于云计算的强大存储和计算能力, 对智能电网海量数据挖掘和分析的关键瓶颈技术展开攻关.

1 系统概述

智能电网数据处理平台存储了三类重要数据: 业务数据、系统数据以及实时数据. 综合利用 SQL 和 NoSQL 技术对这些数据进行存储, 并构建分布式索引, 在物理存储的基础上, 构建语境信息立方体, 在此基础上, 研发智能电网大屏展示系统, 将实时监控数据、数据分析结果展示给电网管理人员. 系统总体架构如

① 收稿时间:2015-03-17;收到修改稿时间:2015-05-25

图 1 所示.

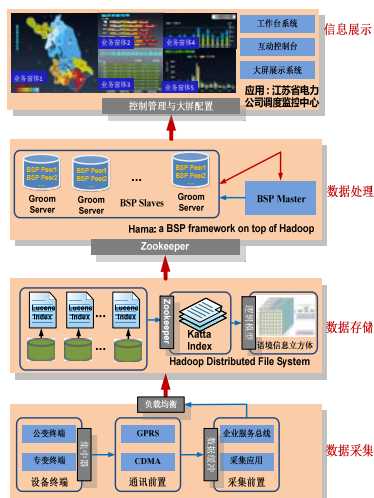


图 1 系统总体架构

2 智能电网数据采集及存储

针对利用感知式系统获取数据逐渐流行、且对数据采集实时性要求高的现状,采用多节点同步采样及基于 FPGA 的多点实时数据采集技术,实现了纳米级采样^[3]。在此基础上,针对关系型数据主要通过关系型数据库(RDBMS)进行管理,非关系型数据通过 NoSQL 云数据库解决 NoSQL 数据存储管理,同通过已有的 BSP 框架解决关系型数据和非/半结构化数据融合的问题^[4-6]。在物理存储的基础上,构建语境信息立方体,灵活融合时效性、空间位置特性、语义特征等多维度属性,为数据查询和处理提供了逻辑模型。

2.1 智能电网数据采集

智能电网数据采集主要集中于发电、输电、变电、配电和用电五大模块的数据, 本文主要集成江苏电力公司现有的五大模块的数据采集技术. 数据采集技术主要由终端设备层、通信前置层和采集前置层构成, 其中终端设备主要为五大模块的采集设备, 通信前置层主要用于将终端设备采集到数据传输到采集前置层的通信接入设备、采集服务器、企业服务总线等, 具体流程如图 2 所示.

(1) 数据采集过程与方法

1) 终端设备层: 主要负责采集各设备的相关信息, 主要包括配电终端、厂站计量、负荷管理、公变计量、低压集抄等终端设备, 主要采用 FPGA(现场可编程门阵列)实时采集技术, 该技术首先建立由多个过程量采集控制终端、基于 FPGA 的通信控制装置和一个或多个

个上层数据处理装置组成的实时数据采集控制系统:接着通信控制装置将采集控制装置的实时采样数据通过以太网将数据同步发送到上层若干个数据处理设备,而数据处理设备的下行控制命令经过通信控制装置分发到各个过程量采集控制终端:

2) 通信前置层: 通信前置系统是监控和数据采集系统的重要组成部分。该层的主要任务是通信模块的主要任务是当用电现场服务与管理系统和远程终端有数据交互时, 利用 CDMA、GPRS, 负责通信信道的管理、数据上行链路和下行链路的维护, 提供快速、准确的数据转发。

3) 采集前置层: 通信前置系统对远程终端上传的数据经过规约解释后, 需要将数据存储在采集前置层。假如当一些终端在上传数据的过程中, 数据库服务器出现故障, 传统的方法是这些数据作废, 等数据库服务器修复后再次提交。这种方法造成了很大的浪费和损失。本文提出了一种改进的方法, 在通信前置层内建立一个临时远程工作站, 用于保存 48 小时内所有终端上传的数据。当采集前置层出现故障时, 终端上传的数据暂时存放在远程工作站, 等采集前置层修复后, 通信前置系统将终端数据上传到数据库服务器。

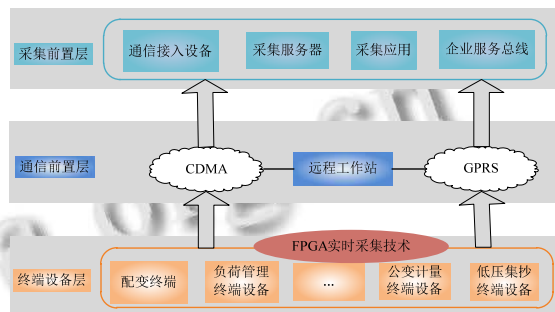


图2 数据采集实现流程图

(2) 应用效果

以江苏电力公司为依托,以低压电力用户为例,实现 475 万户低压电力用户的用电信息采集系统建设,实现用户用电信息的全面准确采集,全面支持预付费业务,并建立了 5 套不同的采集方案:

1) 载波转 485: 适用于城镇集中居住区, 集中表箱方式安装表计. 集中器和采集器之间低压窄带载波方式通信, 不需要布线, 工程施工简单快速, 但载波通信成功率不高, 维护量大, 在用于预付费控制时存在风险, 具体实施方式如图 3 所示.

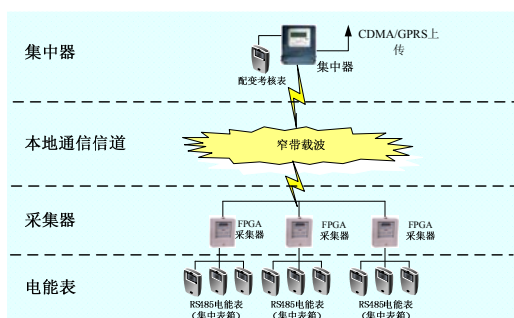


图3 载波转485采集方式

2) 载波和485混合: 适用于县城及城郊居民区, 既具有集中表箱方式安装的表计, 也具有独立安装的表计。对集中表箱安装的表计仍采用“集中器+采集器+RS485电表”的方式建设, 对独立安装的表计更换为载波电表, 直接连接到集中器, 即“集中器+采集器+RS485智能电表+载波电表”方式, 具体实施方式如图4所示。载波方案不需要布线, 工程施工简单快速, 但载波通信成功率不高, 维护量大, 在用于预付费控制时存在风险。

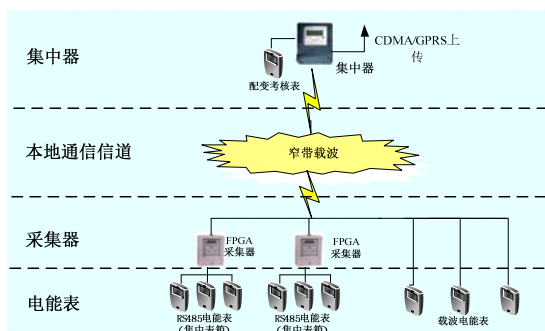


图4 载波和485混合采集方式

3) 全载波: 适用于农村用户, 多采用独立安装表计, 安装载波电表, 具体采集方式如图5所示。载波方案不需要布线, 工程施工简单快速, 但载波通信成功率不高, 维护量大, 在用于预付费控制时存在风险。

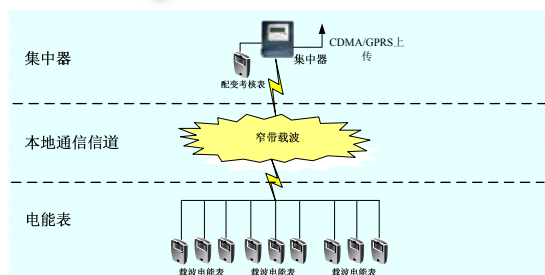


图5 全载波采集方式

4) 全485: 适用于多层、小高层和高层居住区, 具体实施方式如图6所示。多层居住区采用集中表箱安装表计, 施放RS485电缆直接连接表箱和集中器, 在楼栋到配变间电缆施工需要局部开挖电缆沟, 工程量较大。小高层和高层住宅有电缆沟道, 便于施放RS485电缆直接连接到配变侧的集中器。

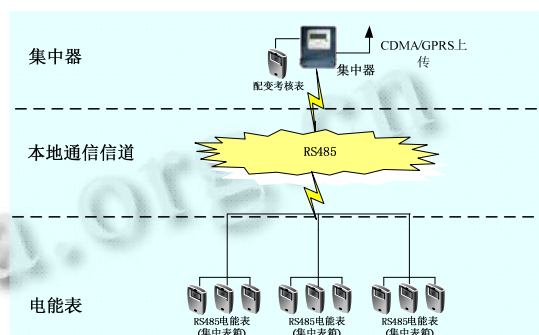


图6 全485采集方式

5) 楼栋集中: 适用于多层和小高层居住区, 具体实施方式如图7所示。在条件许可的多层和小高层居住区尽可能不使用载波, 而是采用RS485有线通信方式来提高通信成功率, 实现楼道口局部集中直接上传的方案, 以减少电缆施工量。楼道集中的集中器为简易集中器, 无载波信道。配变侧的集中器采集不能在楼道集中的分散电表, 同时完成配变关口采集。

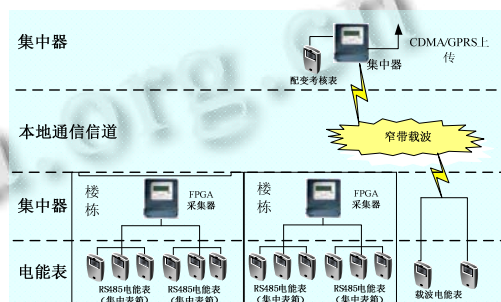


图7 楼栋集中采集方式

2.2 智能电网数据云数据库存储

关系模型和NoSQL技术的数据模型均存在不足, 所以本文的核心工作是解决智能电网关系型数据和非关系型数据存储问题, 即主要基于关系型数据模型的可扩展性、NoSQL数据的存储模型以及基于BSP的智能电网数据计算模型来完成智能电网数据云数据库存储管理, 具体流程如图8所示。

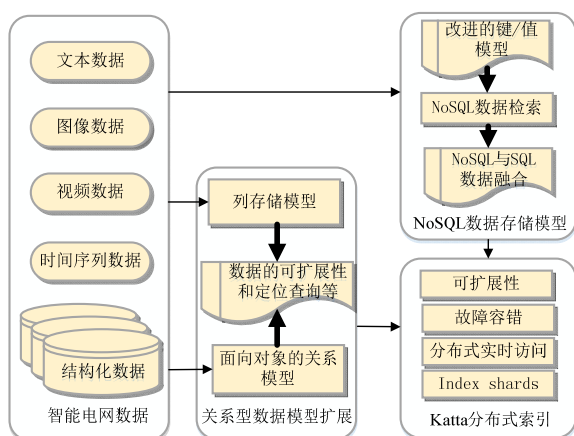


图8 智能电网数据云数据库存储流程

(1) 关系型数据模型的可扩展性模型实现

关系模型对结构化数据存储和查询是最有利的，故核心是解决其扩展性问题。一种思路是将其扩展至面向对象的关系模型，这可以给系统带来更好的扩展性和增量性。在扩展关系模型以处理非结构化数据方面，列存储模型是一个候选。它可实现数据压缩和定位查询，对于拓展关系模型以处理高维稀疏数据是有利的，特别适于插入和删除操作较少的数据分析任务。

(2) NoSQL 数据存储模型

NoSQL 技术通常具有较好的扩展性，且提供了不同类型的数据模型，对文本等多媒体存储提供了很好的支持。其问题在于对查询支持较弱。本文主要通过改进简单的键/值模型，对“值”进行类型定义并设计更为复杂和精巧的数据结构和索引结构，以利于特定类型数据的复杂检索。本文的NoSQL数据通过分层思想来实现关系型数据和非关系型数据的融合，其中下层融合RDBMS和多种NoSQL数据库，以支持包括文本数据在内的异构数据的事务和查询处理；上层基于BSP和Map Reduce计算模型，提供数据共性分析计算服务模块，为应用层实施进一步的数据分析奠定基础。

(3) Katta 分布式索引

Katta 分布式索引与Lucene类似同样是以文件夹的形式建立索引^[7]，该文件夹中包含一套所谓的Index Shards(文件形式)，这些子文件包含了Lucene索引。Index shards能够很简单的用Lucene的Index Writer创建。创建一个Katta的索引相当于把多个Lucene索引拷贝到一个文件夹下，因此Katta索引可以用Hadoop Map Reduce创建。

(4) 数据分析与结果

为了验证智能电网云数据存储方案的有效性，基于江苏电力公司拥有先进的云计算研发实验室环境包括完整的计算机网络环境，20台刀片服务器(2路4核CPU 4GB内存和相应的操作系统环境Oracle、HBase等大型数据库系统环境各种开发工具套件和丰富的组件库门类齐全的技术文档资料库，在此基础上构建云存储原型系统对提出的数据分布策略进行了仿真执行和分析并记录了任务执行的数据传输次数和时间开销。

测试数据包括50,438个文档、14,099张图片、182个视频配置库及数据库文件共计1.1TB。将测试数据分别划分为20、40、60、80、100及120个数据集，每个数据集集中的大文件以文件切片的方式最大切片为64MB进行分布式存储，实验任务集是参数相同的测试流程随机分布于10个数据中心，且与数据集间的调用关系也是随机，传输次数和时间开销的统计结果如表1所示。

表1 数据传输次数和时间开销的统计结果

数据集数目	传输次数	时间开销/h
20	37	0.126
40	142	0.203
60	226	0.374
80	291	0.552
100	322	0.591
120	335	0.638

表1给出的测试结果是测试流程的统计均值，从统计结果可以看出，数据的传输次数和时间开销随着数据集数据的增多而增加，从实验的数据规模来看，两者的增加率逐渐呈现下降趋势。

2.3 语境信息立方体构建

对目前智能电网中主要的信息产生来源进行分析，对各种传感器进行归类，为对其产生的数据添加时间、空间以及语义信息。选取具有代表性的信息来源，建立智能电网数据源的语义库。基于该语义信息，对数据源进行深入分析，包括数据源本身的内容特性，其所含有的子数据源，以及数据源之间的关系，建立数据源的“空间坐标”。对内容相似的数据源，进行语义抽取，基于所获取的语义信息，将数据源进行分类归纳并组织成多个细节层次。通过多层次细节信息源管理，可以得到数据的“空间坐标”。另一方面，通过对所获取数据的时间信息进行分析，发掘数据源之间的

时效性联系并建立所的数据的时间坐标. 此外, 智能电网中每个传感器所获取的数据都具有起独特的物理意义, 因此需要对所设计的物理器件进行分析与建模, 以获取相关元数据. 基于“时空坐标”以及智能电网中传感器的元数据, 建立语义信息立方体对数据源进行综合组织与管理. 语境立方体构建模型如图 9 所示.

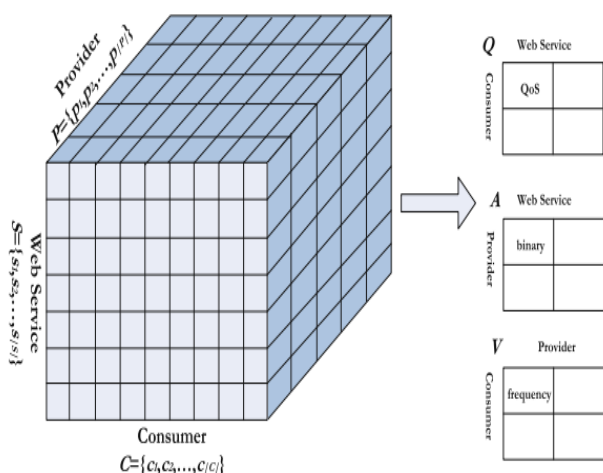


图 9 语境立方体构建模型

为验证智能电网语境数据立方体的有效性, 基于江苏电力公司的数据作为实验数据, 维数固定为 6 维, 各维基数统一为 100, 各点所用业务数据元组个数分别为 200k、400k、600k、800k 和 1M 个, 并验证了立方体的构建时间. 在上述实验中, 主要是验证两种数据立方体构建时间的有效性, 同时将它们与 Dwarf Cube 进行比较. 从图 10 可以看出, 对于数据集 1, 在不同的基本业务元组个数情况下, 本文的语境立方体 (Sets Operation) 的构建时间均明显低于基于 Dwarf Cube 以及 Ibc Tree.

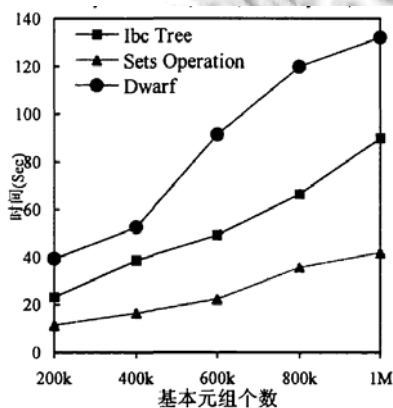


图 10 语境立方体构建时间的实验结果

3 智能电网数据处理

3.1 基于 BSP 的智能电网数据计算模型设计

BSP 模型在迭代计算和依赖性强的任务并行计算上具有优势. 但其将数据存放于内存, 面向数据分析时容易出现存储空间不足; 且未提供容错管理机制, 不利于数据计算时的子任务并行计算^[8-10]. 因此将 BSP 扩展到支持磁盘存储数据, 并增强其容错管理能力.

1) 针对数据类型, 选择基于 BSP 模型的开源系统作为扩展 BSP 模型的起点, 如: 针对大规模图处理的 Google Pregel 系统, 或支持矩阵处理和图处理的 Hama 系统. 目前 Pregel 和 Hama 都直接从分布式文件系统读取数据, 拟结合在云数据库管理系统和公共计算服务模块的研究, 研制新的 Pregel 和 Hama 的 I/O 操作 API 接口, 使其与云数据库无缝融合. BSP 模型中一个超级步和整体代价估算:

$$\begin{cases} T_k = \max_{1 \leq i \leq N} \{w_i\} + \max_{1 \leq i \leq N} \{g h_i\} + L \\ T_{BSP} = \sum_{k=1}^S w_k + g \sum_{k=1}^S h_k + SL \end{cases} \quad (1)$$

式中, w_i 为子任务计算时间, h_i 为任务发送或接收的最大通信包数(称为 h -relation), g 是带宽的倒数, L 是障碍同步时间. 因此, 超级步的长度必须能够充分的适应任意的 h -relation, 才能准确估算出任务代价. 本文通过研究并行粒度和 h -relation 的优化确定算法来解决.

2) 目前, Pregel 和 Hama 系统在容错管理上均存在不足, 而其他基于 BSP 模型的系统也存在类似问题. 数据计算将导致子任务数量急剧增多, 而由于云计算环境资源动态性和异构性强, 由资源负载过重或故障导致计算任务出错或终止的可能性将大大增加. 因此, 研究 BSP 数据计算模型中容错管理机制, 提升 BSP 计算模型的可靠性成为不可忽视的问题. 本文首先通过构建计算任务监控机制, 定时构建检查点, 对其计算状态进行保存; 一旦计算任务终止, 通过重调度功能, 基于最近检查点状态继续计算, 从而实现软件级容错. 同时, 考虑云环境虚拟机宕机问题, 通过定期保存虚拟机状态, 研究虚拟机热迁移方法, 提供虚拟机级的容错.

3.2 语境信息立方体构建

针对 Map Reduce 面向数据处理时的固有弱点, 聚焦于面向数据的 Map Reduce 存储优化与数据类型支持、Map Reduce 数据计算中通信开销优化、Map

Reduce 任务调度优化三个方面^[11,12].

1) Map Reduce 没有提供任何数据模式, 仅仅利用键/值对存储半/无结构化数据, 这极大限制了 Map Reduce 在结构化数据上的应用. 本文基于数据模式的研究, 结合应用需求, 研究 Map Reduce 数据模式, 以扩展 Map Reduce 所支持的数据类型. 由于数据移动是影响数据计算的关键因素, 基于广泛使用的 Hadoop 系统, 通过研究 Map Reduce 存储优化来提升其计算效率. 具体而言, 目前 Hadoop 仅根据机架位置放置副本, 本文综合考虑可靠性、写入带宽和读取带宽、数据流行度等因素研究数据块及其副本的放置优化策略. Map Reduce 本身不提供索引机制, 它依赖于云数据库索引的支持. 因此, 结合云数据库索引的研究, 使 Map Reduce 能利用索引机制提升效率.

2) 在数据计算中, Map 任务可能产生大量的中间结果, 其中键相同的数据将在 Shuffle 阶段发送给同一个 Reduce 节点. Shuffle 阶段需要交互、共享网络带宽, 将成为整个计算过程的瓶颈. 拟通过对 Map 及其 Combiner 函数的精心设计, 使其产生尽量少或排好序的中间结果, 以减轻 Shuffle 阶段的负担; 同时, 通过对数据勾勒、压缩感知、核心集等数据摘要的研究, 压缩中间结果传输量. 拟从 Shuffle 与 Reduce 交互角度, 研究 Shuffle 阶段的高效哈希函数和排序算法, 提升 Shuffle 映射速度.

3) Map 和 Reduce 任务粒度越细, 每个任务计算速度固然越快, 但随之而来的任务启动开销也越大. 假设 M 个 Map 任务和 R 个 Reduce 任务, Master 节点需要执行 $O(M+R)$ 次调度, 在内存中保存 $O(MR)$ 个状态. 事实上, 调度是个多目标优化问题, 拟融合输入数据读取代价、启动开销、Worker 数量对优化问题建模, 进而提出启发式优化算法. 由于 Map Reduce 计算过程是阻塞操作, 一个阶段所有任务完成才能进入到下一个阶段, 这就要求调度优化时考虑负载均衡性. 同时, 对于某阶段的“落伍”任务, 需通过任务备份和迁移技术, 将“落伍”任务重调度到性能好的节点, 从而提升整个阶段的执行效率.

4 智能电网信息展示

在信息展示阶段采用的是基于云平台为视频切换数据矩阵计算而提出的一套高分辨率大屏显示方案. 即由矩阵和拼接控制器组合的传统可视化方案, 改进

到由多个云计算节点分区域控制大屏显示, 多个计算节点协同工作, 以点对点的高清信号、将显示场景以一幅完整窗体显示.

4.1 研究方法

1) 视频输入模块

采用分立三极管组成的同相放大电路形式和集成电路芯片对输入的模拟视频信号进行隔离、缓冲和放大处理, 目的是将视频信号转换为符合切换要求的信号.

2) 云计算节点分区域控制

作为数字显示拼接墙系统的信息处理和控制中心, 它不但要接受从系统外部输入的各式各样的信号源(如来自江苏电力公司各工作部和其它子系统的大量数据流), 并且还要对这些信号源进行处理, 最后送到大屏幕拼接显示墙上进行显示, 为每个显示单元提供高分辨率的驱动信号, 并使图像拼接成一个超高分辨率的整体桌面.

云计算节点分区域控制是整个大屏幕显示的集中管理控制中心, 担负着开关投影机引擎、对信号源管理、窗口显示管理、显示模式管理、显示预案管理等功能. 通过云计算节点的 RS232 接口与显示单元的 RS232 接口通讯, 通过安装在控云计算控制节点上的 VScreen 大屏幕系统控制软件, 对显示单元内的投影机的各种参数进行调整和设置, 对显示墙的亮度、对比度、色彩等进行统一调整.

3) 点对点多频拼接模块

通过云计算主控节点控制的拼接控制器将全屏幕作为统一的逻辑屏, 实现全屏显示和分辨率的叠加功能, 满足系统高分辨率应用程序的显示.

云计算其他节点主要将来自用户的计算机网络的重点工作站信号、远程重要目标数字 IP 等信号, 接入多屏拼接控制器的 VGA 输入端口, 作为独立的、非网络连接的计算机信号, 以开窗口方式在大屏幕拼墙上实时显示. 系统可选配置支持 VGA 信号的直通显示, VGA 窗口的动态视频速度全实时. 窗口在全屏任意位置, 进行放大、缩小、移动、叠加显示, 显示位置不受物理屏边界影响.

另外, 外部视频输入信号可以通过视频矩阵进行输入输出的切换控制, 外部 RGB 输入信号可以通过 RGB 矩阵进行输入输出的切换控制. 可以对包括各种信号源、矩阵、摄像机云台及镜头控制器、多功能设备等外围设备进行统一管理和联动, 所有的操作都可以

通过控制软件通过人性化的图形界面进行统一控制。

4.2 研究结果

1) 具体实现方式: 智能电网可视化模块在软件上的实施方式如图 11 所示, 主要包括四层, 分别为数据层、基础架构层、应用服务层以及表现层。数据层来自云存储的业务数据、系统数据和实时数据; 而基础架构服务层用于权限管理、日志管理、数据访问控制以及规则的制定; 应用服务层主要基于云平台实现服务器的指令管理、同步管理和系统管理等功能; 表现层主要用户智能电网可视化管理。



图 11 可视化在软件上的实施方式

2) 多屏管理: 系统能够实现信号共享, 多屏联动控制, 支持大屏幕分区管理, 支持多人同时对大屏幕进行调度和管理。三台主机分别控制三个区域的显示器, 实现信号同步输出, 大屏多屏分区展示图如图 12 所示。

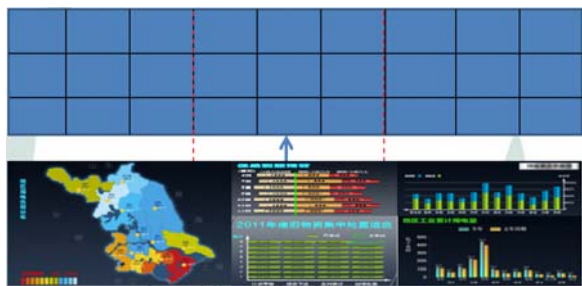


图 12 大屏多屏分区展示图

3) 系统组件扩展: 智能电网可视化模块的组件样式丰富、内容多样, 是可视化系统中展现、配置的最小单元, 丰富的组件库也是完美展示的重要前提, 其组建库图如图 13 所示。使用分布式控制器可以轻松实现组件扩展, 无需更换硬件, 没有浪费, 而传统的处理器几乎要全部更换硬件, 成本巨大。

5 结语

本文主要阐述了基于云计算的智能电网数据处理的 4 个关键技术及其实现效果, 即数据采集技术、存储技术、处理技术与展示技术。在下一步的研究工作中, 将针对智能电网大规模数据处理与分析算法开展进一步研究, 以期促进我国电力行业乃至其他行业在海量多源异构数据上的深度分析和利用, 构建企业级智能应用。



图 13 智能电网可视化组件库图

参考文献

- 曹军威, 万宇鑫, 涂国煜, 等. 智能电网信息系统体系结构研究. 计算机学报, 2013, 36(1): 143-167.
- 白红伟, 马志伟, 宋亚奇. 基于云计算的智能电网状态监测数据的处理. 华东电力, 2011, 39(9): 1485-1488.
- 宗劲冲. RFID 技术在智能电网数据采集中的研究. 保定: 华北电力大学, 2012.
- Codd EF. A relational model of data for large shared data banks. Communications of the ACM, 1970, 13(6): 377-387.
- Stonebraker M. SQL databases v. NoSQL databases. Communications of the ACM, 2010, 53(4): 10-11.
- Chang F, Dean J, Ghemawat S, et al. Bigtable: A distributed storage system for structured data. ACM Trans. on Computer Systems (TOCS), 2008, 26(2): 4.
- 陈宜明. 分布式索引在大规模视觉搜索中的研究与应用. 沈阳: 中国科学院研究生院(沈阳计算技术研究所), 2012.
- Ekanayake J, Li H, Zhang B, et al. Twister: a runtime for iterative mapreduce. Proc. of the 19th ACM International Symposium on High Performance Distributed Computing. ACM, 2010: 810-818.
- Malewicz G, Austern MH, Bik AJC, et al. Pregel: a system for large-scale graph processing. Proc. of the 2010 int. Conf. on Management of Data. ACM, 2010: 135-146.
- 高峰. 基于 BSP 模型的大图处理系统数据划分模块的设计与实现. 沈阳: 吉林大学, 2012.
- Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Communications of the ACM, 2008, 51(1): 107-113.
- Lee KH, Lee YJ, Choi H, et al. Parallel data processing with MapReduce: a survey. ACM SIGMOD Record, 2012, 40(4): 11-20.