电力企业大数据基础平台①

郑浩泉[1], 靳 丹[2], 马志程[3]

[1](国网电力科学研究院, 南京 211100)
[2](国网甘肃省电力公司信息通信公司, 兰州 730000)
[3](甘肃同兴智能科技发展有限责任公司, 兰州 730000)

摘 要: 电力企业各专业均有大数据应用需求, 而大数据解决方案和产品路线众多. 从一体化企业级信息系统的角度出发, 需要构建统一平台, 实现大数据应用服务平台化, 全面支撑各专业大数据应用需求, 避免重复建设. 文章分析了电力企业大数据应用需求和技术现状, 基于此提出了大数据基础平台的功能架构, 最终对平台的设计思想和实现思路进行了详细论述.

关键词: 智能电网; 大数据平台; 分布式技术

Big Data Platform of Electric Power Enterprise

ZHENG Hao-Quan[1], JIN Dan[2], MA Zhi-Cheng[3]

[1](State Grid Electric Power Research Institute, Nanjing 211100, China)
[2](State Grid Gansu Electric Power Company Information and Communications Branch, Lanzhou 730000, China)
[3](Gansu Tongxing Intelligent Technology Development Co. Ltd., Lanzhou 730000, China)

Abstract: In electric power, there are application requirements of big data in every specialty. At the same time, there are many big data solutions and technology roadmap. The unified platform needs to achieve for integrated enterprise information system to avoid multiple construction. In this paper, the application requirements and technique actuality have been analyzed. Then, we propose the function architecture of basic platform for big data. At last, the design theory and implementation of the platform are discussed in detail.

Key words: smart grid; big data platform; distributed technology

电力大数据主要来源于电力生产和电能使用的发电、输电、变电、配电、用电和调度各个环节, 可大致分为三类: 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、用电客户等方面数据; 三是电力企业管理数据. 电力大数据具有四个特点: (1)数据体量大: PB 级; 常规的调度自动化系统包含数十万个采集点; 配用电、数据中心将达到千万级; (2)数据类型繁多: 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; (3)价值密度低: 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[1].

电力企业大数据应用的关键不是“大”和“数据”, 其核心价值是将数据视作与人财物一样的企业核心资产, 让资产创造价值. 与传统数据挖掘分析的区别是通过采用新的采集、存储和处理技术, 实现跨业务、多类型、实时快速、灵活定制的数据关联分析, 满足企业在电网生产、经营管理、优质服务三方面的管理提升和业务创新需求. 由于涉及到的应用众多, 对计算、存储、网络等性能提出了较高要求, 因此需要构建面向电力企业应用的统一大数据处理平台[2]. 本文首先分析电力企业大数据特点以及业务应用需求, 然后结合业务应用介绍大数据关键技术, 进而提出电力企业大数据基础平台和应用框架.

① 收稿时间:2016-05-23;收到修改稿时间:2016-06-30 [doi:10.15888/j.cnki.csa.005604]

1　电力大数据应用场景分类

电力大数据应用场景存在不同维度的分类, 根据业务域不同划分为电网生产、经营管理、优质服务; 根据价值体现分为管理提升和业务创新; 根据应用层次分为战略层、管理层和操作层. 分类情况如表 1 所示[3].

表 1　应用场景分类

应用层次/业务域	电网生产	经营管理	优质服务
战略层		战略指标关联分析与战略决策优化制定.	舆情监测与分析.
管理层	1.物资库存物料需求影响因素分析 2.项目计划与物资需求的关联分析及供货周期预测	设备缺陷及其供应商评价分析	1.客服中心语音分析支撑新型客户服务业务 2.营业厅人流量统计与服务行为分析
操作层	1.用电信息采集性能提升 2.配电网低电压实时监测 3.配电设备负载估算及重过载预警 4.异常用电行为分析与窃电预警 6.电网中长期负荷预测与用电量分析 7.输变电设备状态故障识别与预测	线损计算与分析	用户能耗分析及用电方案优化

1.1 电网生产需求

电网运行生产过程中产生大量的业务数据, 急切需要提升数据存储、处理能力来满足日益增长的性能需求. 利用大数据技术可以经济的满足业务性能需要, 同时可以打破部门间的数据壁垒, 充分实现部门间的业务数据共享, 达到业务融合, 提升业务能力, 促进业务从定性到定量, 提高企业决策能力和决策效率.

典型场景有: 设备状态检修、物料库存分析、用电采集性能提升、异常用电及反窃电分析.

1.2 经营管理需求

电力企业统计分析数据分散于各专业信息系统, 各类辅助决策系统基本是按照专业条线建立, 缺乏跨部门、跨专业的经营战略分析场景, 使得企业需访问不同的专业系统才能掌握了解公司经营情况, 工作效率比较低. 利用大数据技术可以消除跨部门、跨专业壁垒, 构建比较灵活的管理风格, 实现场景模拟、可视化规划、业务全景视图展现等, 为企业领导层制定战略决策提供支撑, 提高辅助决策能力.

典型场景有: 数据驱动人财务科学配置、规划分析、预警/预判、可视化决策、战略情报分析、市场运营分析等.

1.3 优质服务需求

在电力企业生产经营活动中产生了大量的视频、客服音频等非结构化数据、日志、表计等半结构化数据和结构化数据, 迫切需要从这些数据中挖掘有潜在价值信息为客户提供优质服务, 利用大数据技术可以将这些数据进行整合处理, 通过数据内在关联, 挖掘隐含在其中的潜在价值, 及时发现客户敏感需求, 开展客户个性化服务和数据外部化服务, 提高客户服务质量和水平.

典型场景有: 情感分析、智能互动、图像识别、自动化监控、信誉评价、360 视图、语音分析等.

2　电力大数据平台功能架构

电力大数据基础平台需要为电网公司各类应用提供海量数据整合、存储、计算、分析、展现、安全等基础性支撑功能, 平台核心分布式存储与计算组件采用 Hadoop 技术体系中分布式存储(HDFS、HBase、Hive 等)、分布式计算框架(MR), 及可与 Hadoop 形成互补的 Spark 等开源产品或技术, 同时自主研发完善安全机制和运维管理功能[4]. 如图 1 所示.

2.1 数据整合

数据整合采用大数据连接器、实时消息队列、平台服务接口等多种技术手段, 导入结构化数据、非结构化数据、海量/实时数据、空间数据, 对各类数据按照统一数据规范进行标准化及关联处理后, 存储在分布式文件系统或分布式数据库中[5]. 如图 2 所示.

数据整合的组件主要采用 Hadoop 体系中的 Flume、Sqoop、Kafka 等, 这些组件作为独立的服务可以独立构成集群或组合部署.

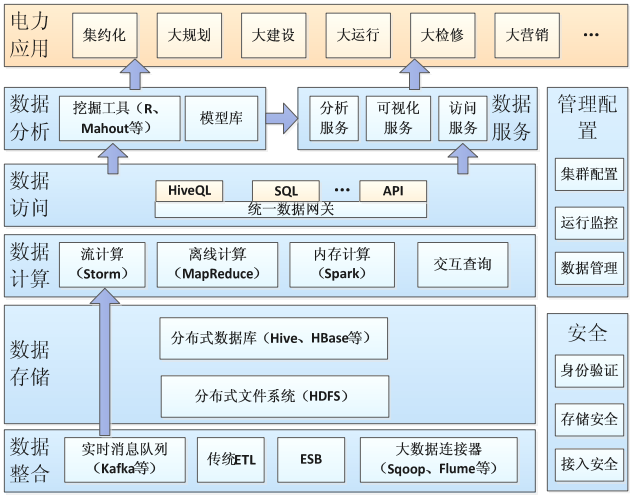

图 1 平台功能架构图

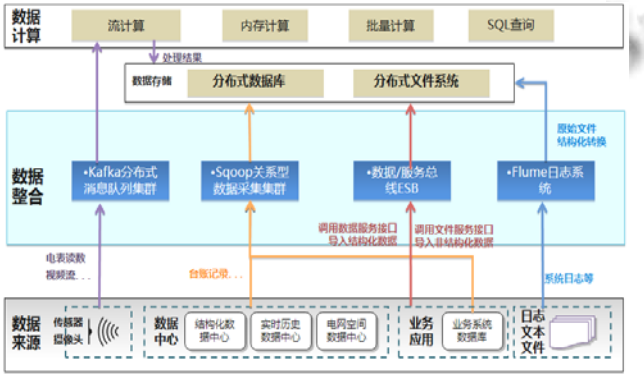

图 2 数据整合技术架构图

由于传统数据库与大数据平台分布式存储之间的差异性, 一般关系型数据难于使用传统的 ETL、OGG 等工具直接导入大数据平台, 需采用 Sqoop 等新型关系数据整合工具连接关系数据库与大数据平台.

传感器数据或来自于实时系统、实时数据库的数据由于处理的时效性要求高, 传统的数据总线或服务总线难于满足这种实时需求, 可采用实时消息队列技术或产品做为实时数据的传统通道.

业务数据也可以通过平台服务或定制的工具导入平台, 如通过文件服务接口导入业务系统文件数据, 通过数据服务接口导入结构化、半结构化数据等.

2.2 数据存储

数据存储主要面向全类型数据(结构化、半结构化、实时、非结构化)的存储、查询, 以海量规模存储、快速查询读取为特征. 在低成本硬件(X86)、磁盘的基础上, 采用包括分布式文件系统、分布式关系型数据库、NoSQL 数据库、实时数据库、内存数据库等业界典型功能系统, 支撑数据处理高级应用[6].

大数据存储的关键在于采用分布式技术和低成本存储设备, 非结构化数据中心将采用分布式存储架构进行改造, 与大数据平台融合. 数据存储模型依赖于企业级数据模型的建立和标准化. 如图 3 所示.

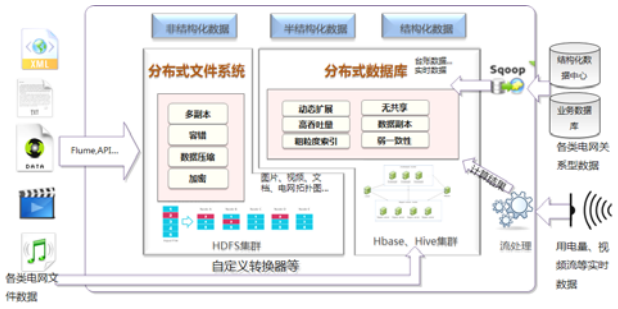

图 3 数据存储技术架构图

HDFS(Hadoop Distributed File System)是可以运行在 X86 低成本硬件上的开源分布式文件系统, 具有高吞吐量、支持大数据集、自动冗余、扩展性好等特征, 适合作为大数据平台存储的基础.

在 HDFS 之上可采用 HBase、Hive、Shark 等分布式数据库或数据仓库产品为应用系统提供面向 SQL 或类 SQL 的数据接口.

HDFS 针对小文件(小于 BlockSize, 一般为 64MB)存储提供了优化方案, 也有第三方解决方案, 具备作为非结构数据中心分布式存储的条件. 在应用 HDFS 改造非结构化数据中心时, 需针对不同大小的文件采取不同的优化策略.

2.3 数据计算

大数据平台数据计算通过流计算、内存计算、批量计算等多种分布式计算技术满足不同时效性的计算需求. 流计算面向实时处理需求, 用于在线统计分析、过滤、预警等应用, 如电表采集数据实时处理、网络状态实时分析与预警等. 内存计算面向交互性分析需求, 用于在线数据查询和分析, 便于人机交互, 如全省用电数据在线统计. 批量计算主要面向大批量数据的离线分析, 用于时效性要求较低的数据处理业务, 如历史数据报表分析[7].

数据计算的核心是分布式计算, 通过分布式计算能将一台计算机无法处理的任务分解到多个节点上. 以分布式计算为基础, 发展出可适应多种计算场景的计算框架.

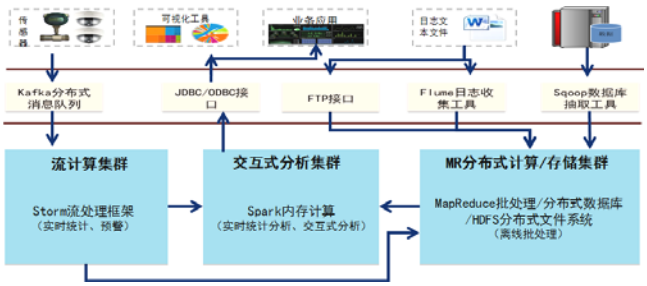

图 4 数据计算技术架构图

2.4 数据分析

平台面向电力企业各级数据分析与成果应用人员，提供多维分析、统计分析、数据挖掘分析于一体的在线和离线数据分析功能. 利用高性能计算、预测及优化分析、阈值监视分析和高维可视化技术，结合实际业务问题建立业务分析模型，通过将数据信息与分析模型相结合，对企业在经营管理和科研生产中涉及的关键要素进行分析，并借助高维可视化技术实现数据可视化展示，从而提升企业经营管控与科研生产能力. 如图 5 所示.

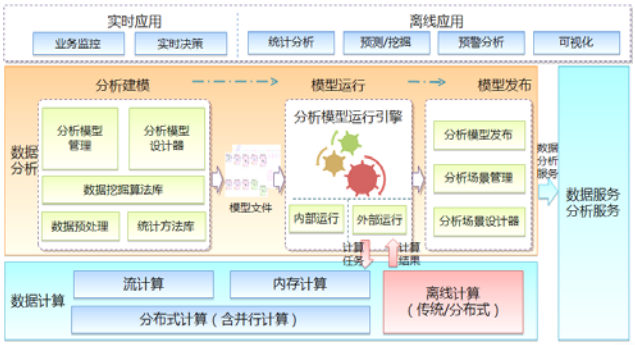

图 5 数据分析技术架构图

数据分析的关键在于支持分布式挖掘算法，提供易于使用的分析建模工具. 应对现有智能分析决策平台进行提升改造，完善分析建模、模型运行、模型发布等能力，增加对大数据分布式计算的支持，满足实时、离线应用的分析挖掘需求，为电力企业分析决策应用构建提供基础平台支撑[8].

2.5 数据服务

针对大数据应用提供必要的二次开发工具，这些工具包括平台服务接口(文件服务、数据服务、分析服务)和大数据可视化组件库. 通过平台服务接口，业务系统可以访问存储在平台中的文件、持久化数据，执行预定义的分析任务，快速构建面向大数据的数据分析功能. 如图 6 所示.

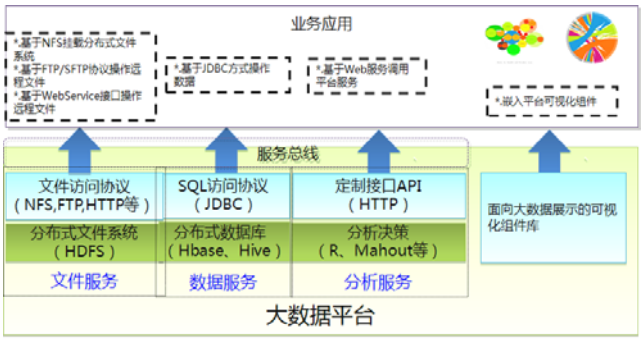

图 6 数据服务技术架构图

平台服务应尽量采用通用技术实现，如 JDBC、ODBC、FTP、HTTP、WebService 等协议，以符合业务系统开发习惯.

通过 HTTP 协议提供的服务接口，优先采用 Restful 风格的服务样式，提供比传统基于 SOAP 协议的 WebService 服务更高的性能.

2.6 管理配置

大数据平台通过专门的监控服务对集群的状态进行监控，包括服务器 CPU、内存、网络和磁盘的利用率和健康状态，以及分布式应用系统的状态，并在故障发生时提供告警功能. 管理员可通过浏览器访问集群的监控和管理界面进行日常的监控和维护，系统提供图标信息展示. 管理员可以便捷地了解到集群的计算资源是否处于空闲状态、哪些服务器的负载过高，甚至判断集群的组网及机架安排是否合理等. 管理员也可通过对各个节点的各个角色的日志信息进行检索，获得更加精确的信息.

平台集成Ganglia集群监控系统，能够从上千台服务器上收集系统信息，能够保障当集群规模扩展至上百台上千台服务器时，监控程序都能够高效地获取每台服务器的状态信息. Ganglia 由 Gmetad 与 Gmond 组成，其中集群中每一台机器上都有一个 Gmond 服务进程，Gmetad 收集所有节点 metrics 信息并在 Web 前端界面上展现出来. 集群中的每台服务器上都运行监控守护进程，守护进程能够将这些数据的精简传递，这使得 Ganglia 的运行对集群的资源消耗极少. 所有的指标数据都存储在一个 RRD(Round Robin Database)数据库中，为了防止 Gmetad 频繁写磁盘造成 I/O 瓶颈，通过 rrdcache 缓存指标数据，定量写入 RRD. 如图 7 所示.

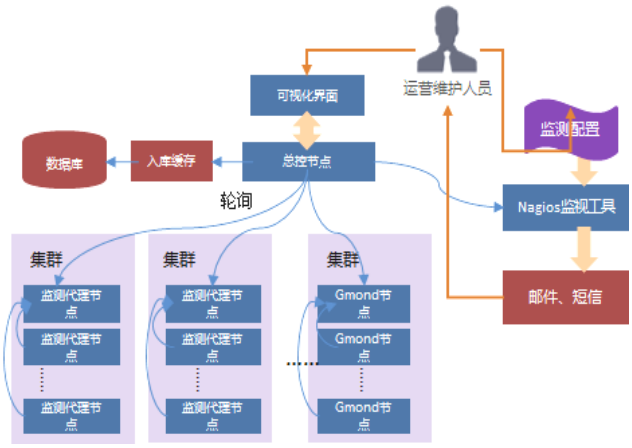

图 7 管理配置技术架构图

集成 nagios 监控告警模块, 通过 Ganglia 中的指标检测以执行报警功能. 运维人员配置监测指标策略以及指标告警阈值. 当检测到指标异常达到设定的阈值时, 平台会以邮件方式将告警信息通知到管理员, 系统会根据设定的警报级别来发送警报通知. 同时管理员可以通过 WEB 界面查看当前的网络状态、故障告警历史和日志文件等. 系统支持与第三方网管系统对接, 完成集群状态及告警信息的上报.

除上述大数据平台系统运维之外, 平台提供面向业务运维人员的管理配置功能. 业务运维人员专注于业务系统使用平台各类功能组件、资源的情况, 也可以对业务分析模型、数据整合规则等进行灵活调整.

2.7 数据安全

针对大数据存在的隐私保护、存储安全、权限控制、防泄漏等安全风险以及电力企业对数据安全的高要求, 需要在研发、集成隐私保护机制、增强分布式存储安全等功能之外, 制定大数据安全相关的标准规范, 从技术和规范两个层面确保业务系统数据在平台和应用中的安全性. 如图 8 所示.

应用系统、终端等数据源或数据消费者的接入安全可考虑传统的安全接入方案, 通过网络安全、主机安全、访问认证等技术手段加以解决.

存储安全主要采用数据加密方式保护关键业务数据或用户隐私数据. 同态加密(Fully Homomorphic Encryption)是一种可以采用的加密方法, 允许对密文进行特定的代数运算得到仍然是加密的结果, 与对明文进行同样的运算再将结果加密一样. 同态加密使得在加密的数据中进行诸如检索、比较等操作, 得出正确的结果, 而在整个处理过程中无需对数据进行解密. 存储安全的另一个方法是 Hadoop 的文件读、写、执行的 ACL 控制, 结合自定义的用户组策略实现文件权限控制.

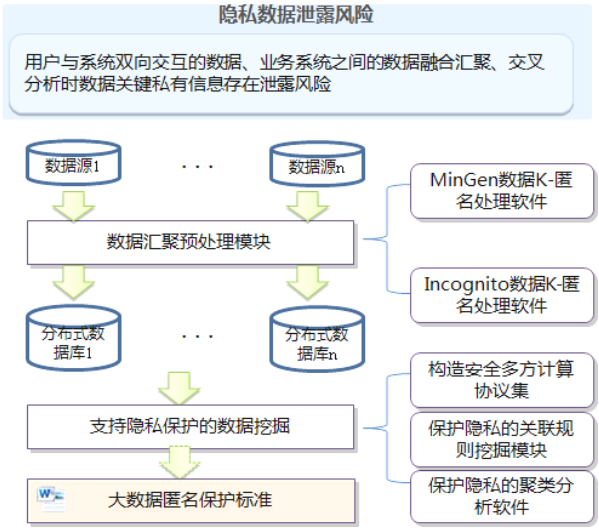

图 8 数据安全技术架构图

保护隐私的数据挖掘(Privacy-Preserving Data Mining, PPDM)能够在多方联合数据库进行挖掘, 除了挖掘结果公开之外, 不泄露各参与方的私有数据库信息. 在分析用户行为数据时, 采用保护隐私的统计分析技术可以有效避免用户数据泄露并且可以得出用户个体行为习惯的数据, 例如用户用电的规律和用电潜在需求.

由于传统的隐私保护方法并不能阻止攻击者根据类身份属性准确识别目标个体在原始数据中对应的记录, 删除了身份属性后的数据仍然可能泄露数据中个体的敏感信息. K-匿名隐私保护模型(K-anonymity privacy model)采用泛化和压缩技术对原始数据进行匿名处理以得到满足 k-anonymity 规则的匿名数据, 从而使得攻击者不能根据发布的匿名数据准确识别目标个体的对应记录.

权限管理可采用主流的用户认证授权机制对大数据用户进行权限管理, 如国家电网统一权限平台对用户进行统一的认证和授权管理.

3 电力大数据平台应用实例

目前, 大数据平台的核心组件已在国家电网电能质量系统等项目中开始应用. 经实验验证和现场应用

测试, 大数据平台在大幅度提高系统可靠性的同时, 提升系统计算效率达数百倍.

3.1 应用现状

电能质量系统的中压供电可靠性指标统计功能模块包含 14 个注册线段统计指标、16 个注册用户统计指标, 22 个停电事件统计指标, 18 个停电用户统计指标, 综合统计指标 57 个, 共 127 个指标. 指标计算共涉及 8 类数据, 包括: 中压供电系统的线段注册和用户注册数据、中压停电事件和中压跨月停电事件、中压停电线段和中压跨月停电线段、中压跨越停电用户和中压跨月停电事件.

该系统属于国网公司一级部署系统, 数据来源于国网下属各网省的业务系统, 数据体量大, 类型多, 现有数据量约 2 亿条, 每年产生的增量数据大约 5 千万条. 随着数据量的不断增加, 计算性能低下, 计算时间达到小时级, 不能满足需求.

3.2 优化方案

基于大数据平台的内存计算框架, 将系统的业务模型分解为数据对象模型和任务对象模型. 数据对象模型是指根据业务系统数据逻辑规则建立的对象模型. 任务对象模型是根据业务系统任务计算逻辑对计算任务进行封装建模. 创建过程如图 9 所示.

按照分配原则, 数据对象分布式存储到集群计算机节点(对象服务器)的内存中. 任务对象则分为两类: 一是计算任务, 它是直接根据数据对象进行计算, 二是汇总任务, 接收直接计算任务的结果进行汇总计算. 任务对象根据数据对象池所在位置进行缓存.

对象服务器缓存对象后, 将机器 IP、对象池缓存对象总数等信息以心跳方式发送到集群的管理服务器, 管理服务器收集信息后创建并维护对象索引表. 对象索引表如图 10 所示.

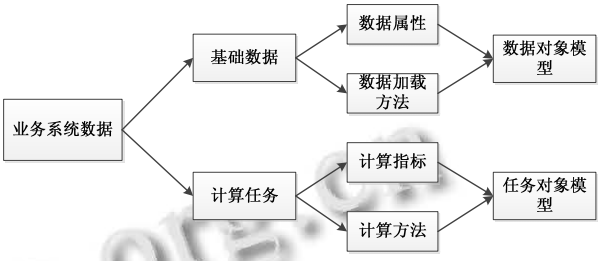

图 9　业务对象预处理过程

级别	单位	对象服务器	加载开始时间
1	区域0	对象服务器1	时间1
—2	区域1	对象服务器1	时间2
——3	区域11	对象服务器3	时间3
——3	…		
——3	…		
—2	区域2	对象服务器2	时间4
——3	…		
——3	…		

图 10　对象索引表

计算任务利用索引了解对象服务器上的数据对象分布情况, 并下发子任务对象到不同的对象服务器[9].

创建完对象索引, 整个系统就可执行任务. 完成任务的过程如图 11 所示.

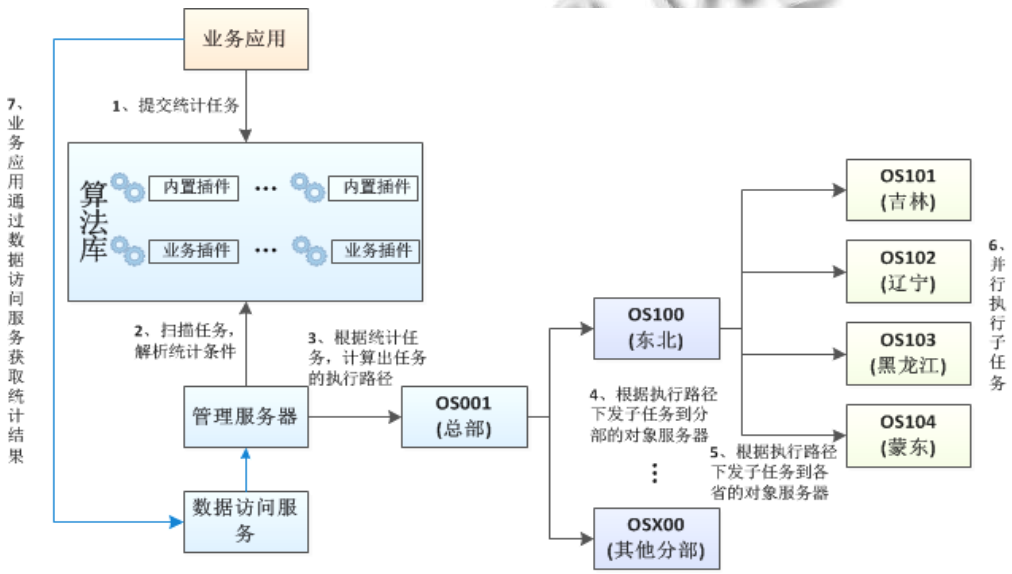

图 11　任务执行过程

图 11 中, 业务应用首先将统计任务提交到大数据平台的算法库中, 通过手动或自动(任务引擎)方式触发任务的执行. 任务执行时, 管理服务器首先扫描待执行任务, 解析出统计条件, 并据此计算出任务的执行路径, 然后将任务下发到执行路径中各对象服务器中执行, 各子任务执行完毕后, 再根据执行路径将执行结果逆向返回并逐级汇总. 在本例中, OS001、OS100、OS101、OS102、OS103、OS104 都部署了对象服务器, 分别缓存不同对象. OS001 在执行任务过程中, 将任务拆分为子任务分发给 OS100 等; OS100 在执行任务过程中, 将任务拆分成 4 个子任务, 分发给 OS101、OS102、OS103、OS104 这 4 个节点并行执行, 然后将执行结果进行逐级汇总, 最终返回给总部服务器(OS001). 业务应用则通过数据访问服务获取汇总结果.

3.3 系统部署

系统部署时, 对象服务器根据缓存的数据进行横向扩展组成集群, 如图 12 所示.

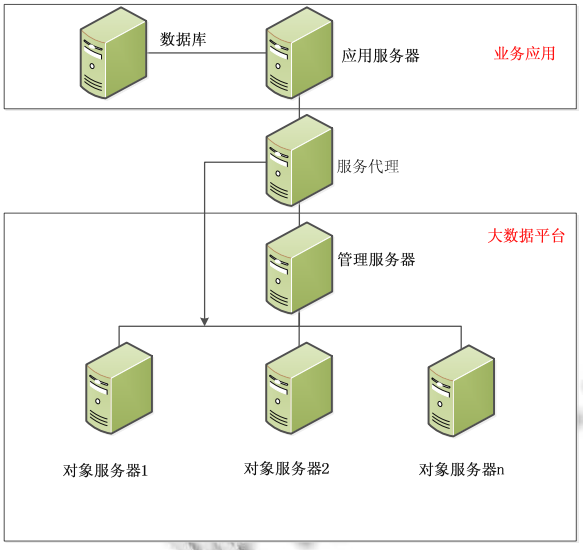

图 12 系统部署图

(1) 对象服务器是大数据平台计算框架的核心组件, 提供对象服务远程接口; 负责创建和维护对象, 加载并缓存数据; 可根据需要加载的数据规模横向扩展.

(2) 管理服务器是大数据平台的管理中心, 负责接收对象服务器的对象注册信息并缓存, 负责对象服务检索、分配和管理, 同时需要保障集群高效、持续、稳定运行.

(3) 服务代理组件部署在业务应用服务器, 为业务系统提供对象服务接口调用的本地代理组件.

3.4 运行效果

在电能质量系统中, 任务按照地区属性进行拆分, 任务拆分如下图所示, 国网任务划分为分部子任务, 分部子任务划分为省子任务, 子任务分布式并行计算, 如图 13 所示.

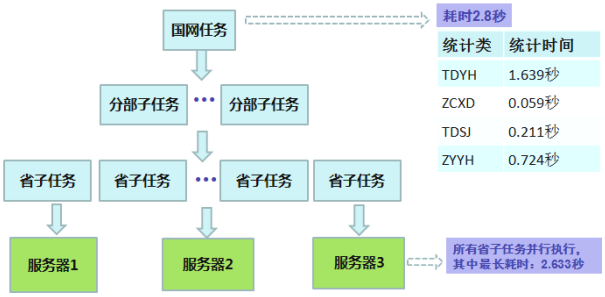

图 13 运行效果

以东北分部为例, 需要执行 TDYH、ZCXD、TDSJ、ZYYH 四个统计任务, 涉及到的数据总量为 2700 万条左右, 数据分布情况如表 2 所示, 总任务耗时 2.8 秒(含网络开销), 测试结果如表 3 所示. 相比原有技术方案(基于集中式关系库、两级数据中心)在生产环境中的平均耗时 13 分钟, 提高了 278 倍, 性能提升明显. 需要注意的是, 本测试场景的集群节点全部搭建在同一个网段, 网络开销只有 170 毫秒左右, 生产环境中还需考虑广域网的网络开销.

表 2 数据分布情况

区域	省+代码	数据量	总数
东北	吉林 012022	6977294	27133327
	辽宁 012021	7185872	
	黑龙江 12023	1885496	
	蒙东 012024	11084665	

表 3 测试结果

区域(机器名)	统计场景	完成计算(秒)	省公司(机器名)	完成计算(秒)
东北(OS100)	TDYH	1.639	吉林(OS101)	0.62
			辽宁(OS102)	1.158
			黑龙江(OS103)	0.3
			蒙东(OS104)	0.343
	ZCXD	0.059	吉林(OS101)	0.023
			辽宁(OS102)	0.028
			黑龙江	0.039

		(OS103)	
		蒙东(OS104)	0.05
TDSJ	0.211	吉林(OS101)	0.065
		辽宁(OS102)	0.138
		黑龙江(OS103)	0.144
		蒙东(OS104)	0.097
ZYYH	0.724	吉林(OS101)	0.638
		辽宁(OS102)	0.68
		黑龙江(OS103)	0.586
		蒙东(OS104)	0.522

3.5 特点与优势

通过在电能质量系统的应用, 大数据平台主要体现了如下的特点与优势:

(1) 采用两阶段的对象分布式查询方案, 统计任务按照策略拆分成查询和汇总子任务, 子任务并行执行, 大幅度提高任务执行效率.

(2) 大数据按照业务逻辑关系拆分成小数据, 任务被推送到小数据所在节点, 各节点并行查询本地数据.

(3) 采用多级查询汇总机制, 只有数据量非常小的查询结果需要通过网络传输返回到父任务所在节点, 多级汇总有效降低一级汇总带来的负载压力和网络吞吐量, 提高了汇总任务执行效率.

(4) 沉淀了一批电网模型和专用算法, 如针对结构化数据的分布式全局排序、模型自动拼接等以及针对测点数据的持续分析法、补偿分析法、滚动分析法等, 并基于大数据平台进行了并行化实现, 可作为后续电力业务大数据应用的基础工具包.

4 结语

电力企业大数据基础平台, 能够融合企业内部经营管理、电网实时运行、用户用电信息和外部社会经济、气候气象、地理空间等数据, 集成数据采集、存储、计算、分析和展现等工具组件, 为生产控制、经营管理和公共服务领域实时采集类、在线监测类、计算分析类和决策支持类应用统一提供数据服务和技术支撑, 促进电网业务创新发展.

参考文献

1 黄翔,陈志刚.智能电网大数据信息平台研究.南方能源建设,2015,2(1):17–21.

2 孟祥君,季知祥,杨祎.智能电网大数据平台及其关键技术研究.供用电,2015,3(8):19–24.

3 周爱华,戴江鹏,丁杰,饶玮,胡斌,朱力鹏.面向多源异构电网数据的获取与转换技术研究.电力信息与通信技术,2015,13(7):22–27.

4 段军红,张乃丹,赵博,闫晓斌.电力大数据基础体系架构与应用研究.电力信息与通信技术,2015,13(2):92–95.

5 胡健,袁军,王远.面向电网大数据的分布式实时数据库管理系统.电力信息与通信技术,2015,13(2):49–54.

6 王玮,刘荫,于展鹏,苏琦,周伟.电力大数据环境下大数据中心架构体系设计.电力信息与通信技术,2016,14(1):1–6.

7 Jayawardhana P, Perera D, Kumara A, Paranawithana A. Kanthaka: Big data caller detail Record (CDR) analyzer for near real time telecom promotions. 2013 4th International Conference on Intelligent Systems, Modelling and Simulation. 2013. 534–538.

8 王相伟,史玉良,张建林,梁波,程翠萍.基于 Hadoop 的用电信息大数据计算服务及应用.电网技术,2015,39(11):3128–3133.

9 Tran Q, Sato H. A solution for privacy protection in MapReduce. 2012 IEEE 36th International Conference on Computer Software and Applications. 2012. 515–520.