电力大数据多元数据采集监视技术研究与应用

孙 超,常夏勤,王永贵,胡剑锋
(南京南瑞继保电气有限公司,江苏 南京 211102)

摘 要:随着智能电网的蓬勃发展,电力系统中产生的数据也迅速增长,大数据技术在电网安全稳定运行的决策过程中起到的作用越发重要,然而各类数据的采集与集成的复杂度高稳定性差,缺乏有效的监视手段,成为制约大数据技术应用的一个主要障碍。针对这种情况,研究面向电力大数据采集环境的监视技术,通过分布式采集代理架构实现对多源异构系统数据接入程序日志和运行信息的采集,通过弹性消息队列以及跨安全区和跨级传输代理实现海量监视数据的可靠安全传输,通过智能索引技术实现对数据库海量数据的快速检索和分析,为运维人员提供了良好的数据接入监视管理手段。该系统已经在南方电网公司广州供电局承接的863课题"基于大数据分析的城市电网设备状态评估系统开发与应用"中应用,实现了对课题示范工程接入的22个业务系统采集过程的监视。投运以来运行效果良好,能够快速、及时、准确地提供系统运行状态数据和告警,明显提高了运维人员的工作效率。

关键词:电力大数据;多源数据;分布式系统;智能索引;数据监视
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2020)07-0180-06
doi:10.3969/j.issn.1673-629X.2020.07.038

Research and Application of Multi-source Data Acquisition Monitoring Technology for Power Big Data

SUN Chao,CHANG Xia-qin,WANG Yong-gui,HU Jian-feng
(NR Electric Engineering Co.,Ltd.,Nanjing 211102,China)

Abstract:With the rapid development of smart grid,the data generated in power system also increases rapidly. Big data technology plays an increasingly important role in the decision-making process of safe and stable operation of power grid. But the high complexity,poor stability and lack of effective monitoring means of data acquisition and integration have become a major obstacle to the application of large data technology. In view of this situation,the monitoring technology for large data acquisition environment of electric power is studied. The data access program log and operation information of multi-source heterogeneous system are collected by distributed acquisition agent architecture,elastic message queue with cross-security zone and cross-level transmission agent are used to achieve reliable and secure transmission of massive monitoring data,and with intelligent indexing technology,the rapid retrieval and analysis of the massive data in the database is realized,which provides a means of data access monitoring and management for operation and maintenance staff. The system has been applied in 863 project "Development and application of equipment status assessment system for city power grid based on big data analysis" undertaken by Guangzhou Power Supply Bureau of China Southern Power Grid Corporation. It has achieved the monitoring of the acquisition process of 22 business systems connected to the demonstration project. Since operation,the system has worked well and can provide its operation status data and alarm quickly,accurately and just in time,obviously improving the efficiency of operation and maintenance staff.

Key words:big data for power system;multi-source data;distributed system;intelligent index;data monitoring

0 引 言

随着智能电网的蓬勃发展,新技术在电力系统中的应用越来越广泛,从发电侧的光伏、风电等新能源厂站,到输配电环节的特高压交直流电网、柔性直流配电网,再到电动汽车为代表的各种智能用电设备,电力系统中涉及的设备越来越复杂,产生的数据呈爆炸式增长,对电网的安全运行风险的管控难度与日俱增,这些变化促使了近年来对电力大数据分析方面的研究不断深入。然而,电力大数据在多源数据接入方面仍存在以下问题:(1)现有电力系统的信息化和自动化系统

收稿日期:2019-08-10 **修回日期**:2019-12-11
基金项目:国家高技术研究发展计划(863计划)(2015AA050201)
作者简介:孙 超(1978-),男,硕士,高工,研究方向为分布式系统。

建设周期跨度大,采用的技术复杂。同时,涉及通讯、自动化、调度、营销、气象等多个专业,缺乏统一的数据采集监视方法;(2)电力系统产生的数据中存在海量毫秒和秒级实时数据,可靠性要求高,缺乏有效的监视手段;(3)现有大数据平台的监视侧重于平台内部的资源调度、数据存储、计算分析等方面,对外部数据接入方面的监视功能较弱,无法满足运维需求。

如何结合电力大数据的特点解决数据采集中的监视运维问题,成为关系到电力大数据能否实用化的重要制约因素。

1　电力大数据采集的场景分析

1.1　多专业数据混合采集监视

电力系统涉及的专业方向较多,比如设备运维、通讯、IT、计量、安监、市场交易等[1];每个专业都有多种数据源,且交互方式繁杂,比如 WebService、电力专用规约、特殊文件格式等;数据种类繁多[2],比如实时数据、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]。电力数据产生的速率跨度大[4],比如毫秒级广域向量测量实时数据,秒级的稳态监视数据,分钟级的微气象数据[5],小时级的操作票流转数据和更长时间周期的设备实验数据等[6]。因此,开展电力大数据分析的前提是多源异构数据混合采集。当前电力大数据混合采集多采用针对已有系统接口开发代理采集程序[7],每个采集程序都会产生大量的日志和状态信息记录工作状态和数据接入的情况,需要有效的监视手段[8]。

1.2　跨区跨级数据采集监视

电力系统的数据源多分布于不同的物理位置,从电能产生到消费中间会经过发电厂、输电网、配电网和用户多个环节,每个环节都会产生需要采集的数据;电网公司是分级管理的集团,分为网、省、地、县四级电网调度管理机构,每一层级都会产生大量的电网运行监控类的数据,因此采集程序为分布式部署方式,对应的采集监视模块也要适应分布式和远距离通讯的要求。

电力系统属于关系到国家安全的重要基础设施,其网络安全防护要求非常高[9],各类应用按照安全区隔离部署的方式运行[10],在不同安全区之间有物理隔离网闸实现数据通讯的单向传输[11],同时,由于业务数据的传输占用了大量带宽,采集监视系统在尽量节省带宽不影响业务数据传输的前提下实现跨物理隔离网闸的可靠传输[12]。

1.3　对数据采集监视的要求

综合以上电力大数据采集的特点,数据采集监视需要满足以下的要求:

(1)采用分布式架构,能够根据业务需求横向扩展。

由于采集的数据源众多且分布广,因此采集监视系统需采用分布式架构,贴近采集程序或装置侧部署,前端就地处理数据将结果传输回后端,减少网络占用,提高整体性能。同时,能够适应接入量的不断增加,支持前端数据收集、数据传输汇集、后端数据处理的横向可扩展性,避免由于某个环节的性能瓶颈影响整体运行稳定性。

(2)采用插件化方式,满足各类系统接入的需求。

接入的数据源接口方式众多,需要采用插件化方式实现灵活定制开发数据收集前端程序。同时插件实现外部数据格式到内部格式的转换,实现内部统一的数据交互。

2　总体架构

面向电力大数据的数据采集监视系统由采集对象层、采集代理层、数据汇集层、数据处理层、数据存储层和前端展示层组成,如图 1 所示。

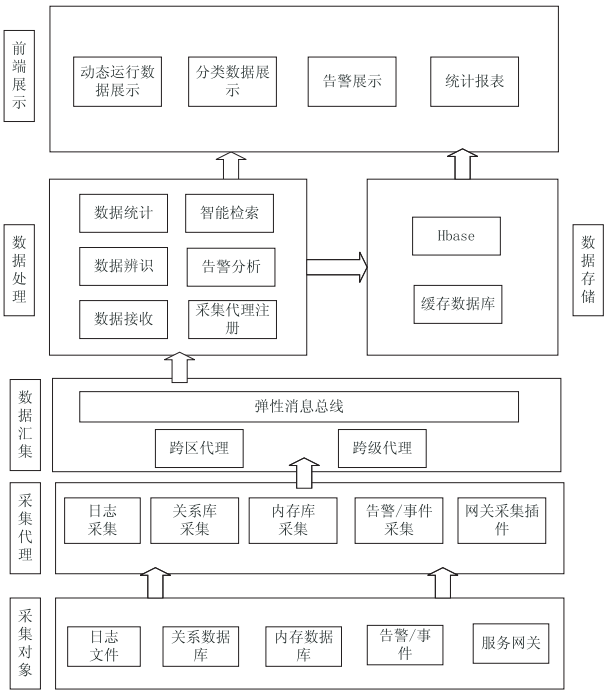

图 1　面向电力大数据的数据采集监视系统架构

采集对象层由各种多源异构数据接入程序或工具产生的状态数据组成,典型的是程序或系统服务的日志文件、关系数据库中有关数据集成情况的记录;在电力系统中由于有大量秒级、毫秒级数据采集,因此有些采集统计或者运行状态数据存储在内存数据库中以提高性能,同时还会产生实时告警或事件;对于采用微服务或 WebService 方式交互数据的情况,服务网关是重要的采集监视对象。

采集代理层是针对各类采集对象,采用插件架构的代理程序,有部分代理是独立于采集对象部署,有的

需要嵌入采集对象,比如服务网关上的服务调用情况和流量监控就需要采用服务网关插件的方式嵌入网关内部才能获取完整的信息。

数据汇集层主要包括弹性消息队列作为采集数据的传输管道,提供海量数据的吞吐能力;面向电力行业的跨级和跨区采集监视需求,设计跨区和跨级转发代理功能,实现数据可靠跨区跨级传输。

数据处理层主要实现对采集代理的注册管理、数据的后端接收、数据辨识、告警分析、数据统计,以及基于灵活索引的数据智能检索。同时,将处理后的数据入库。

数据存储层包括保存所有数据的 HBase 数据库和存放经常访问数据的缓存数据库。

前端展示层为人机交互界面,提供动态运行数据的展示、分类统计数据展示、告警展示和统计报表功能。

3 关键技术

3.1 采集代理

采集代理分为两类,即独立代理和嵌入代理。其中独立代理采用文件、SQL 等松耦合方式与采集对象交互,可以独立于采集对象运行环境部署;嵌入代理集成在采集对象运行框架中或利用其内部 API 交互,需在采集对象运行环境中部署运行。

3.1.1 独立代理

独立代理采用 Flume 架构实现操作系统日志、应用程序日志和中间件日志文件的采集,以及关系型数据库中数据监控记录的采集。它充分利用 Flume 的日志采集和数据格式化处理能力,少量代码即可实现数据采集功能。如图 2 所示,独立代理首先通过各类数据收集模块获取对应数据源的数据信息并缓存在内存缓存管道中,然后通过传输模块串接另外一个代理将缓存数据发送到弹性消息总线上,数据接收和处理模块从弹性消息总线获取数据。

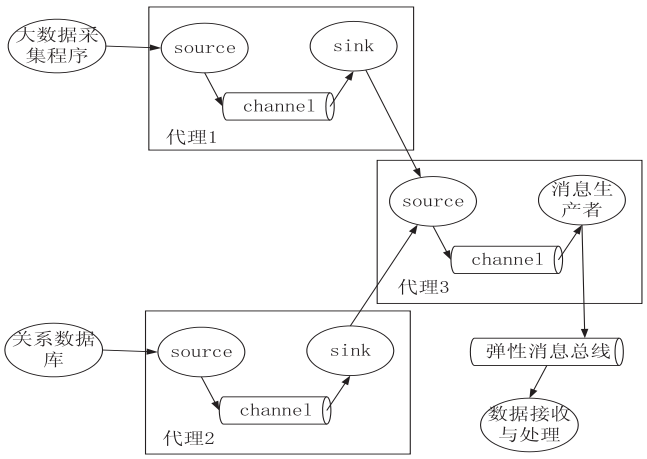

图 2 独立代理数据交互图

3.1.2 嵌入代理

嵌入代理采用插件化方式运行于采集对象运行框架内,比如通过定制的网关过滤器插件统计网关交互的信息,包括服务调用频度、交互数据量、服务响应周期等。网关过滤插件的运行机制如图 3 所示,在“pre”过滤器中记录服务调用信息、调用时间、输入的数据流量;“post”过滤器中记录收集服务调用响应周期、返回的输出数据流量,并计算统计指标;“error”过滤器中记录服务调用错误情况。

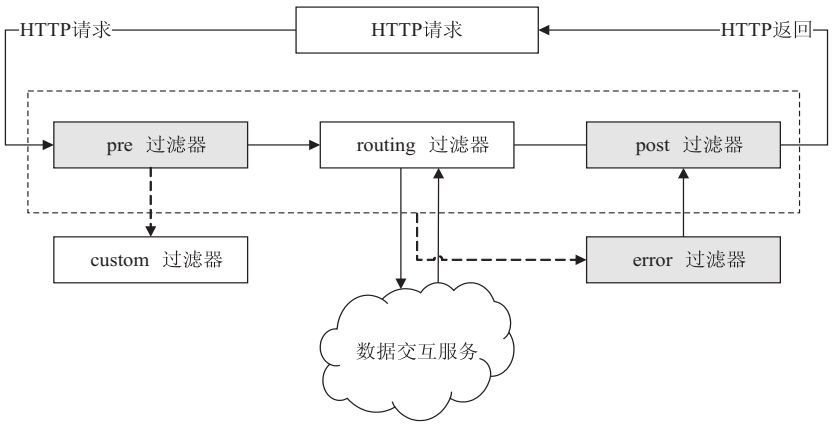

图 3 网关过滤器插件机制

嵌入代理调用采集对象系统私有 API 接口的方式获取数据,比如调用电网调度自动化系统内存数据库接口获取实时数据采集通道的运行状态、数据采集的流量等信息。

嵌入代理收集到数据后直接调用弹性消息总线接口将数据传输给数据接收和处理模块。

3.2　数据汇集

监视数据的汇集层由弹性消息总线,跨区代理和跨级代理组成。

3.2.1　弹性消息总线

前端代理程序收集的数据通过弹性消息总线转发给后端的处理模块,由于涉及的业务系统较多且部署分散,因此需要消息总线具备高并发吞吐量、弹性扩展、支持各种编程语言接口、可靠传输机制等特性。对主流的三种消息总线的比较如表 1 所示。

表 1　主流消息总线比较

比较项	ActiveMQ	RabbitMQ	Kafka
吞吐量	低	高	非常高
可用性	主从	主从	分布式
负载均衡	支持	支持	支持
消息延迟	秒	毫秒	毫秒
编程接口	丰富	AMQP 客户端	丰富
可靠传输	较好	好	好

结合电力大数据多元数据采集监视对消息总线的要求以及主流消息总线的特性比较,选择 Kafka 作为数据汇集的弹性消息总线。它具有分布式多路并发弹性扩展部署的特点,具备消息时延低、可靠性高和编程接口丰富等方面的特性,能够满足前端代理数据汇集与后端数据接收与处理的需求。同时,采用 Flume 对接 Kafka 的方式可以将数据自动导入到 Hbase 数据库,提高开发和运维效率。

3.2.2　跨区代理

电力系统安全防护规范规定业务数据从安全 I/II 区进入安全 III 区需要经过隔离装置(网闸),它具有单向通讯特点,反向只能传输 1 字节报文用于状态确认,而弹性消息总线基于标准 TCP 双向通讯无法穿越隔离装置,需要通过跨区转发代理实现报文转发。跨区代理分为部署在安全 I/II 区的内网跨隔离装置代理和部署在安全 III 区的外网跨隔离装置代理。安全 I/II 区的弹性消息总线数据由内网隔离装置代理的消息消费者模块接收并转入缓存由压缩模块将报文压缩以减少对隔离装置带宽的占用,然后再由转发程序通过并发的多个单向 TCP 链路发送给安全 III 区的外网跨隔离装置代理的转发接收程序,再经过解压和缓存队列交互最终由消息生产者将消息报文发送给安全 III 区的弹性消息总线。两种代理均具备运行状态监视模块,负责转发链路的监视和运行状态告警。跨区转发过程如图 4 所示。

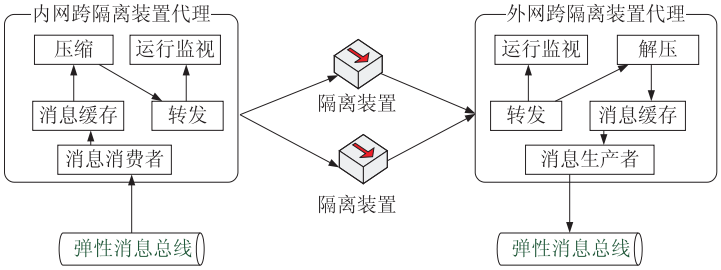

图 4　跨区代理转发机制

3.2.3　跨级代理

电力系统的数据采集分布在网、省、地、县多级电网调控管理机构,要实现集中的采集监视必须在各级调控机构部署采集代理,这些代理采集的数据通过远程网络传输到上级的管理单位,而远程网络存在带宽波动、时延长、时断时续等特点,因此需要跨级数据转发代理提高远程传输的可靠性。跨级代理程序采用滑动窗口技术实现批量数据转发和断点续传,发送端启动滑动窗口后将窗口内的报文开始编号和窗口大小作为窗口信息发送给接收者,接收代理准备好后就按序发送窗口报文,最后一条报文带有窗口发送完毕标识,接收代理检查收到的报文序号和数量并发送窗口整体确认报文,发送端收到确认后将窗口在报文队列中向前移动,已经发送过的数据删除,交互过程如图 5 所示。

如果在跨级传输过程中出现丢包,则窗口整体确认报文中发送丢包前的最后一帧报文编号,发送端重发滑动窗口中此编号后的报文,实现断点续传。

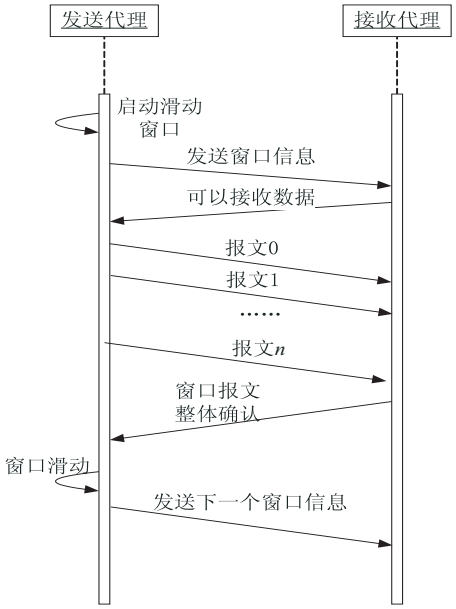

图 5 跨级代理窗口滑动传输机制

3.3 数据处理

数据处理层获取弹性消息总线中的数据并辨识数据类型后分类处理,比如对告警数据按照告警规则分类入库和发送短信或电话语音告警;对流量数据入库后按数据源、数据性质等规则分类统计;对需要高速检索的数据建立索引等。数据处理的基础是数据接收和分类,数据处理高效的关键是智能索引。

3.3.1 数据接收与分类

数据接收模块从弹性消息队列中获取的报文分为两部分:报文头和报文体,如表 2 和表 3 所示(报文头和报文体均采用 JSON 格式)。在报文头定义中“datapedigree”是数据族谱信息,用于数据分类,每个数据源的采集对象不同,族谱的层级可能不一样,因此采用数组方式,比如营销业务的电量计量数据源中智能电表实时数据的族谱定义为[“营销”,“计量”,“智能电表实时采集量”]。有了族谱定义数据处理模块可以将数据存入 HBase 数据库的对应列簇中,并将数据族谱作为标签,后续的智能索引中针对此列簇建立索引,实现分类统计和动态展示。

表 2 报文头各属性定义及 body 部分定义

header 报文头				body 消息体
sourcetype	sourcename	msgtype	datapedigree	自定义内容
数据源类型	数据源名	消息类型	数据族谱	由各采集端定义

电量计量数据流量消息的 body 消息体定义如下,采用 JSON 格式。

表 3 网关流量消息体定义

Key	Value 格式	说明
id	Chart[32]	对象编号
name	Char[128]	对象名称
time	longlong	采集时间
flowtype	Octet	流量类型
flowrate	Double	数据流量
node_id	Chart[32]	采集节点的 ID

3.3.2 智能索引

统计和动态展示模块需要快速地检索 HBase 数据库中的数据,然而 HBase 数据库需通过行键访问[13],这种访问方式与常规检索通过关键字对数据列查询的模式不一致[14]。为了解决这个问题,系统采用基于 ElasticSearch 组件的关键字检索为分类统计和用户界面展示等模块提供服务[15],通过 ElasticSearch 将 HBase 数据库中经常被访问的列按照数据标签建立倒排索引,并封装数据检索微服务。统计程序和界面实时检索调用微服务,后者通过输入的关键字调用 ElasticSearch 检索功能获取行键后再访问 Hbase 数据库获取相关数据返回调用者,调用过程如图 6 所示。

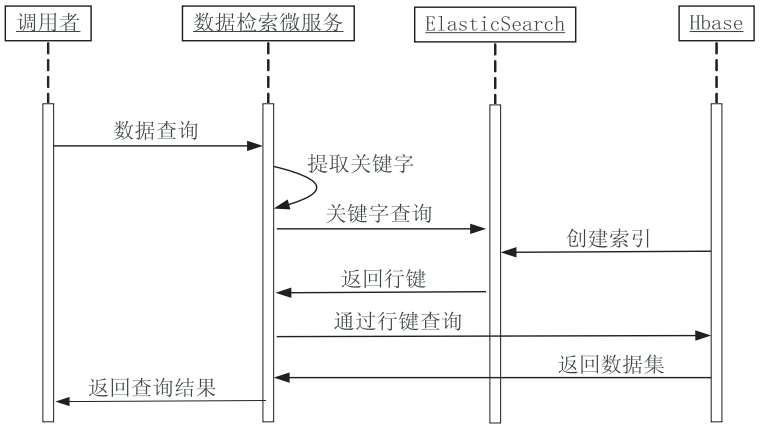

图 6 数据检索交互过程

数据采集监视获取的海量日志和运行记录在 ElasticSearch 上建立倒排组合索引,包括 sourcetype、sourcename、datapedigree、id、time 列。数据检索微服务实现根据关键字和检索时间段查询。在查询请求中分析出的多个关键字通过 ElasticSearch 查询符合关键字的数据,同时叠加检索时间查询,包括绝对时间查询即

通过查询请求中的起止时间查询和根据一个具体时间点及其向后或向前相对时间段实现时间区间查询。

4　应用案例

电力大数据多元数据采集监视技术已经在南方电网公司广州供电局承接的 863 课题“基于大数据分析的城市电网设备状态评估系统开发与应用”中应用,采集监视集群部署在 6 台服务器节点上,单台服务器配置 2 颗 12 核 Xeon E7 2.1 GHz 的 CPU、内存 64 G,安装 CentOS6.7 操作系统。

系统接入 22 个采集源,源端系统涉及电力设备数据、电网运行数据、管理信息、环境气象数据四大类 239 子类的数据。通过定制开发代理程序实现对 sqoop 脚本和 Kettle ETL 工具的批处理采集的信息监视,以及对电力规约如 IEC61850 的采集信息监视;通过开发部署于 Ngix 和 Zuul 网关中的嵌入式代理实现对 WebService 和 Restful 的数据集成模块的信息监视,系统具备不间断连续运行能力,现场无故障运行超过一年。

系统动态地统计各类数据的实时流量信息,对实时数据的采集监视周期达到 5 秒钟,系统一天采集的日志量在 10 GB 以上,数据库入库记录上亿条;传统架构下对如此大数据量的信息进行检索将耗时数分钟甚至无法返回,此系统采用智能索引技术可在 3 ~ 5 秒钟内返回结果。以往大数据系统的故障排查需要运维人员查看各个处理环节的日志分析原因,一个问题需要数小时才能定位,此系统收集各节点的日志信息,并进行快速的检索和分析,将最关键的信息呈献给运维人员,可将故障定位时间缩短到半小时以内,大幅提高运维效率。

5　结束语

通过分析电力大数据多源数据分布式采集监视的应用场景,提出一套适用于电力系统多专业广域大数据采集监视的技术方案,运用 Flume、Kafka、ElasticSearch 等大数据技术,适应电力系统实时数据接入、跨区和跨级传输的特点,实现基于数据分类标签的智能索引和分析,为运维人员提供高效的系统监控手段,推进电力大数据技术的实用化。基于此技术的系统已经在客户现场投运,运行效果良好。在今后的工作中,将进一步研究人工智能技术与数据监视的结合,利用机器学习技术对系统运行状态进行分析,实现系统态势智能感知、无人自动巡航和故障自主处置。

参考文献:

[1] 李子乾,王乐之,张云志,等. 电网大规模数据仓库的数据接入研究与设计[J]. 计算机应用与软件,2018,35(8):180-185.

[2] 甘似禹,车品觉,杨天顺,等. 大数据治理体系[J]. 计算机应用与软件,2018,35(6):1-8.

[3] 闫龙川,李雅西,李斌臣,等. 电力大数据面临的机遇与挑战[J]. 电力信息化,2013,11(4):1-4.

[4] 张　沛,杨华飞,许元斌. 电力大数据及其在电网公司的应用[J]. 中国电机工程学报,2014,34:85-92.

[5] 商　皓,雷　明,马海超,等. 电网供应链大数据应用规划方法研究[J]. 中国电力,2017,50(6):69-74.

[6] 冷喜武,陈国平,白静洁,等. 智能电网监控运行大数据分析系统总体设计[J]. 电力系统自动化,2018,42(12):160-166.

[7] 蒋湘涛,贺建飚,李　楠. 电力信息采集的通用型通信规约解析系统研究与设计[J]. 电力系统保护与控制,2012,40(9):118-122.

[8] 谢大为,杨晓忠. 调度自动化系统中远动技术网络化的实现[J]. 电网技术,2004,28(8):34-37.

[9] OLINER A,GANAPATHI A,XU W. Advances and challenges in log analysis[J]. Communications of the ACM,2012,55(2):55-61.

[10] KIM Y,HUH E. A rule-based data grouping method for personalized log analysis system in big data computing[C]// Fourth edition of the international conference on the innovative computing technology (INTECH 2014). Luton:IEEE,2014:109-114.

[11] 李兴梅,张名杨. 电力系统二次安防综合措施的研究[J]. 网络安全技术与应用,2017(2):119.

[12] 郑子淮,裘雨音,陈琭草,等. 电力监控系统二次安防的防护策略[J]. 自动化应用,2017(8):121-122.

[13] WANG J,WANG W,CHEN R. Distributed data streams processing based on Flume/Kafka/Spark[C]//3rd international conference on mechatronics and industrial informatics. Zhuhai:Atlantis Press,2015:167-171.

[14] 李祥池. 基于 ELK 和 Spark Streaming 的日志分析系统设计与实现[J]. 电子科学技术,2015,2(6):674-678.

[15] Monitoring of IaaS and scientific applications on the cloud using the Elasticsearch ecosystem[C]//16th international workshop on advanced computing and analysis techniques in physics research (ACAT2014). Prague: IOP Publishing,2014:608-713.