

基于大数据的数据服务应用研究

陈 光

(江苏自动化研究所,江苏 连云港 222006)

摘 要:目前各行业信息系统中数据积累规模迅速增长,传统的数据存储、处理以及应用服务难以适应持续增长的数据应用需求,各级部门无法有效跨单位获取重要数据信息,进而不能帮助上级机关做出重要辅助决策。对此,提出一种适用于行业数据应用的大数据架构设计,具有并行、分布、稳定、高效等技术特点。研究大数据存储与处理技术、大数据查询与分析技术以及大数据可视化技术,建立数据分类目录体系标准与信息交换共享机制,确保多数据采集渠道的大规模数据能够有效整合、有序组织,综合运用数据统计、数据分析、数据挖掘等方法,提炼能够保障服务于指挥决策支持信息知识并形成可视化平台,满足不同指挥层级应用人员便捷、及时地掌握应用信息的保障需求,提升应用行动的快速反应能力。

关键词:大数据;云计算;数据采集;数据分析;数据挖掘;辅助决策

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2018)08-0129-06

doi:10.3969/j.issn.1673-629X.2018.08.027

Research on Data Service Based on Big Data

CHEN Guang

(Jiangsu Automation Research Institute, Lianyungang 222006, China)

Abstract: With the rapid growth of the data accumulation in the information system of the industry, the traditional data storage, processing and application services are difficult to adapt to the data application demand of continuous growth, and the departments cannot effectively obtain important data information across the units, which cannot help the higher authorities to make important auxiliary decision. For this, we present a large data architecture design for industrial data application, which is parallel, distributed, stable and efficient. We study large data storage and processing technology, large data query and analysis technology and large data visualization technology, and establish system standards and information exchange sharing mechanism to ensure that large data collection channels of large-scale data can be effectively integrated, orderly organization. The data statistics, data analysis, data mining and other methods are used to extract knowledge to support command and decision information and form visualization platform, which meets the needs of grasping the application information conveniently and timely in different command level application personnel and enhances rapid response capability of application action.

Key words: big data; cloud computing; data collection; data analysis; data mining; auxiliary decision

0 引 言

智能电网是大数据最重要的应用领域之一,支撑智能电网绿色安全、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,以及累积的海量多源数据快速分析,证实了大数据在大规模数据集处理、应用方面的优势。而目前随着智能电网的快速发展,智能电表的大量部署和传感技术的广泛应用,电力工业产生了大量结构多样、来源复杂的数据,如何存储和应用这些数据,是电力公司面临的难题^[1]。

(1)难以适应灵活多变的应用需求。

现有的服务方式主要面向指挥所内使用,模式固

定、手段单一,缺少灵活、精确的数据组织与服务保障,对机动环境下应用人员信息支持不足,难以适应灵活多变的应用需求。

(2)决策支持信息获取能力不足。

现有的数据应用手段尚停留在数据查询、统计层面,对于数据缺乏有针对性、深层次的分析、信息提炼手段,应用数据汇集的价值未能充分发挥,对于应用人员更为关注的决策支持信息获取能力不足。

(3)不便于掌控全局应用信息。

现有的数据展示手段单一,主要依赖图、表、文字,缺少综合性、多维的展示方式,不便于应用人员直观快

速地掌握全局应用信息。

应用数据建设将以服务应用、保障指挥为核心,能够整合、积累来自业务保障部门等各种渠道获取的相关数据资源,具备从大规模的应用数据资源中获取有价值的决策支持信息,同时形成面向应用指挥的统一数据视图,满足不同层级应用人员随遇接入、按需服务保障需求,发挥数据最大的应用效能^[2]。

(1)多层次决策支持信息需求,能够根据应用需求,提供动态的应用数据服务,并且能够保障指挥过程中在任何时间、任何地点都能享受权限范围内的应用信息服务;

(2)应用数据按需服务保障需求,能够从多渠道获取的数据资源中分析、挖掘价值信息,为不同指挥层级应用人员提供不同粒度的决策支持,快速提供其所需的应用信息响应支持;

(3)跨部门应用数据整合利用需求,能够科学划分应用数据资源目录,规范化应用数据采集渠道,有效

整合各部门与应用相关的数据资源,有序组织、标准化整编各类数据资源,保障应用数据灵活、充分利用。

(4)常态化应用数据管控需求,能够支持应用数据采集、存储、分析、服务等过程的运维管理,扩充应用数据容灾备份手段,加强数据安全防护与密码保密手段,辅助应用数据全生命周期管控,同时建立信息使用评估反馈机制,全面把握不同层级应用人员的信息需求。

1 需求架构

针对应用过程中对数据的建设需求,构建数据中心。一方面整合应用、各业务保障部门的应用相关数据资源,形成统一的应用数据视图;另一方面,实施数据深层次分析利用,为本级、各区域提供大规模应用数据的按需保障支持。同时加装应用数据库提供其应用数据应用支持,如图 1 所示。

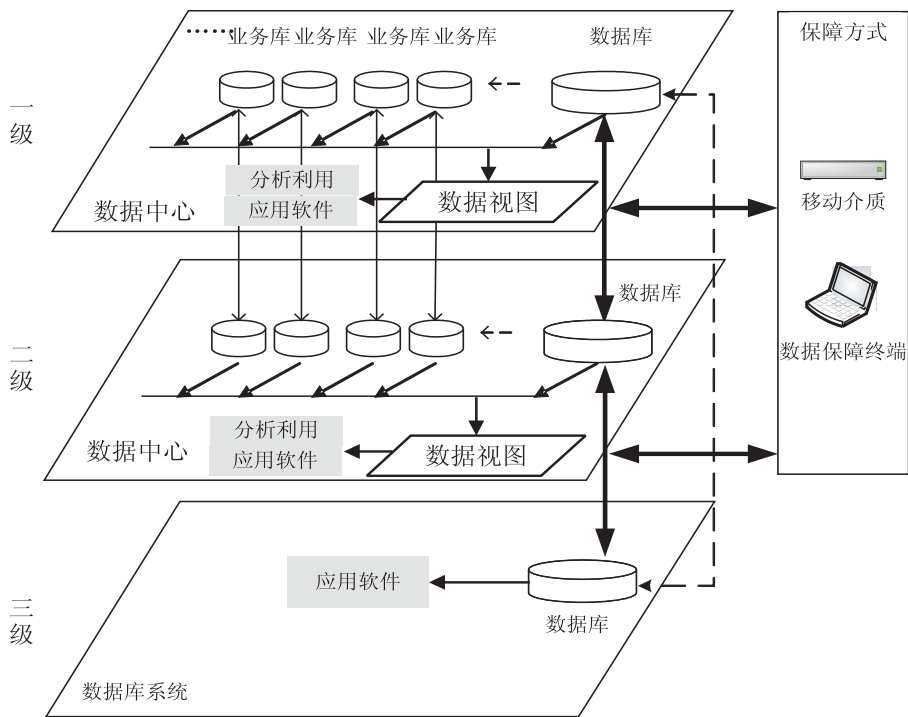


图 1 需求架构设计

2 大数据支撑环境

大数据应用支撑环境是基于大数据业务全生命周期,提供大数据业务一站式支撑服务,包括大数据的数据规划、采集与预处理、存储与管理、分析与挖掘以及应用与服务等,实现了大规模数据资源的有机整合和充分利用,如图 2 所示。

3 技术架构

针对智能电网等领域应用数据具有的容量大、非

结构化、多样性、冗余度高等特点,以及实际业务中快速开展大数据分析处理的应用需求,提出支持横向扩展,具有并行、分布、高效等特点,且面向服务并支持大数据全生命周期活动的平台体系架构,从而解决大数据的服务、共享、分析和整合等集成性问题^[3-4]。然而,从海量数据中“提纯”出有用的信息,这对网络架构和数据处理能力而言是一个巨大的挑战。从三个方面描述大数据的技术体系,包括支持大数据实施的云计算平台、支撑上层业务应用的大数据应用支撑以及业务应用三个层次。其典型的技术架构如图 3 所示。

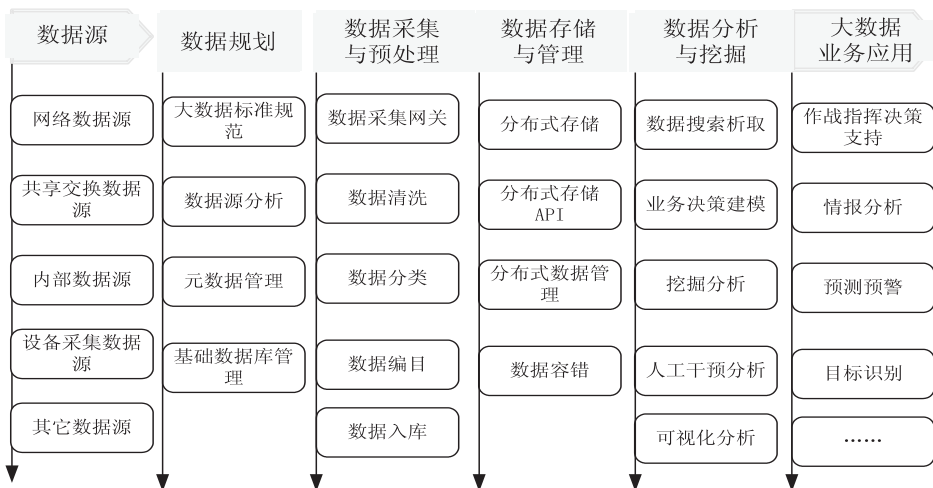


图 2 大数据业务生命周期流程

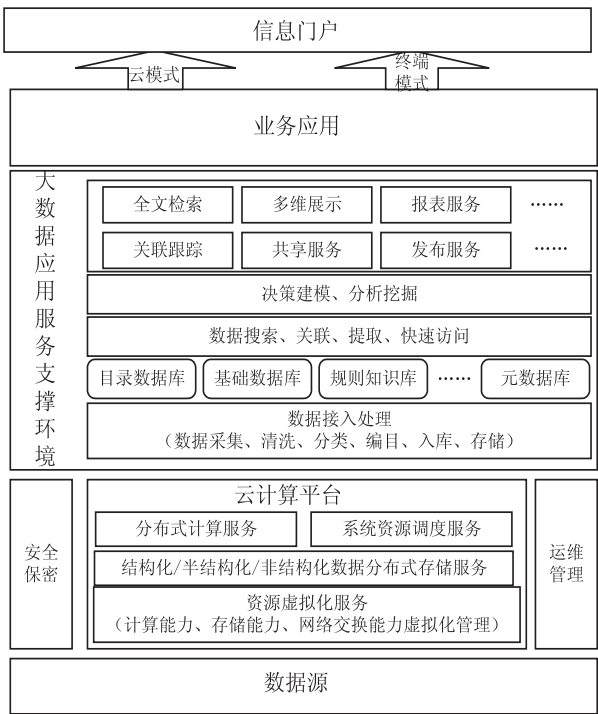


图 3 技术架构

大数据技术架构在逻辑上表现为一种层次架构,自上而下包括大数据业务应用层、大数据应用服务支撑、云计算平台以及数据资源层。而相关的标准、规范和安全机制贯穿所有层次。该架构主要包括:

- (1)数据源层,涵盖各类结构化、半结构化、非结构化数据。
- (2)云计算平台,作为大数据的技术支撑基础,为大数据应用提供大规模数据的计算能力、存储能力以及交换能力,同时确保大数据的应用安全和数据安全,面对大容量数据处理、存储等需求,实施资源虚拟化管理,提供易扩展的计算能力、存储能力、网络交换能力的虚拟化管理服务,统一进行系统资源调度^[5-6]。
- (3)大数据应用服务支撑,依托云计算平台提供的海量数据计算、存储与交换能力,面向业务应用提供

大规模数据的采集、处理、存储、管理、分析与挖掘、应用服务等支持,主要包括:

- ①对全面采集的各类结构化、非结构化数据进行清洗、分类、编目、入库、存储等数据接入处理;
 - ②构建基础数据库、规则知识库、元数据及索引库等数据库,提供接入处理以及决策建模、分析挖掘等生成的数据分类存储支持;
 - ③根据应用服务、决策建模、分析挖掘等需求,提供各类数据搜索、关联、提取以及快速访问支持;
 - ④面向信息的各类应用需求,进行决策建模、分析挖掘,提炼各类规则知识、规律信息;
 - ⑤统一为不同类型用户等提供多样化的应用服务,包括全文检索、多维展示、信息关联跟踪等等。
- (4)大数据业务应用,在大数据应用服务支撑环境支持下,面向具体业务需求,提取所需数据集,展开分析、提炼工作,发现其中价值信息,提供信息融合、目标识别、目标动向预测等多样化的信息应用。

(5)信息门户层,为不同类型用户提供“云模式”和“云+端模式”两种使用方式,即只需通过登录信息门户,即可访问各个系统,有效地支撑完成各类业务应用。

4 应用开发架构设计原则

- 大数据应用架构需要满足业务需求^[7]:
- (1)能够满足基于大容量、多类型、快速流通的数据环境下的大数据处理需求,支持大数据的采集、存储、处理和分析;
 - (2)满足大数据应用在可用性、可靠性、可扩展性、容错性和安全性等方面的要求;
 - (3)满足基于原始技术和格式实现数据分析和整合的基本要求,支持对复杂的原始格式数据进行整合分析的能力。

目前,各个大数据应用开发的架构基本都是基于

Apache 开源的大数据平台,涉及的架构参考模型如图 4 所示。大数据的处理流程主要是通过分布式文件处理系统来实现的,使用的主流技术主要是 Hadoop +

MapReduce,其中以 Hadoop 的分布式文件处理系统 (HDFS)作为大数据存储框架,MapReduce 作为大数据处理框架^[8-9]。

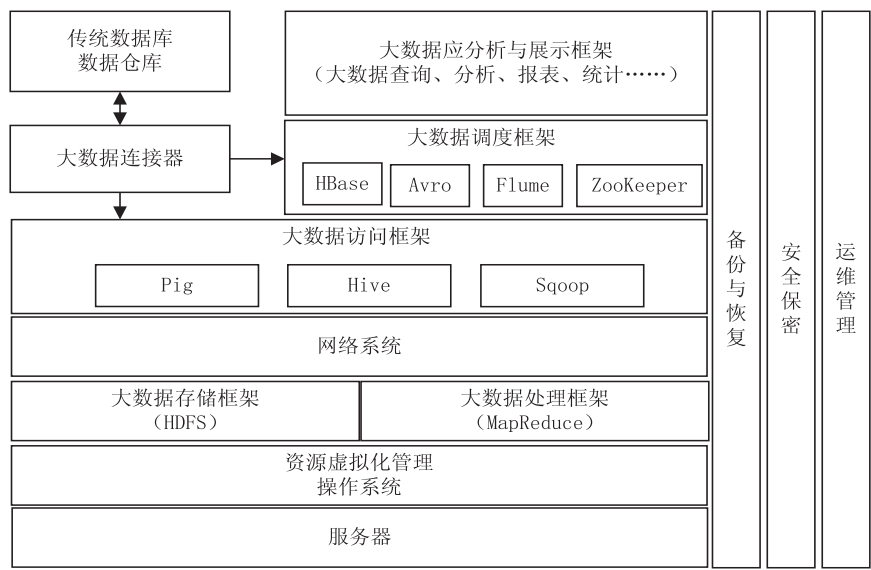


图 4 大数据开发框架

(1) 大数据存储框架。

HDFS: Hadoop 的分布式文件系统,用于存储非结构化数据,运行于大规模服务器组成的集群系统之上,采用元数据集中管理与数据块分散存储相结合的模式,并通过数据复制实现高度容错,在架构上通常在服务器、操作系统或虚拟机之上。

(2) 大数据处理框架。

MapReduce: 分布式并行计算框架,使得应用程序能够运行于大规模集群系统之上,并以可靠容错的方式并行处理 TB 级以上的数据集。

(3) 大数据访问框架。

大数据访问框架是实现传统的关系数据库和 Hadoop 的访问,主流技术包括 Pig、Hive、Sqoop 等。

Pig: 是基于 Hadoop 的并行计算,提供一种类 SQL 的数据分析,具备将类 SQL 的数据分析请求转化为一系列经过优化处理的 MapReduce 运算,常用方法包括分组、过滤、合并等^[10]。

Hive: 数据仓库工具,是 MapReduce 实现的用来查询分析结构化数据的中间件,提供类 SQL 的查询语言支持^[11]。

Sqoop: 用于在 Hadoop 与传统的数据库之间进行数据传递,即能将存储于 HDFS 的数据与关系型数据库的数据进行互导。

(4) 大数据调度框架。

大数据调度框架实现对大数据的组织和调度,为大数据分析做好准备,主流技术包括 HBase、Avro、Flume、ZooKeeper、Oozie 等^[12]。

HBase: 基于列存储的非关系型数据库,可直接运

行于 Hadoop 之上,提供大规模数据的实时读取和写入随机存取。可以存储结构化或非结构化数据。

Avro: 数据序列化格式与传输工具。

ZooKeeper: 分布式锁工具,用于分布式应用中高性能的协同服务^[13]。

Flume: 提供可靠的分布式流收集服务。

Oozie: 基于服务器的工作流引擎,用于调度和运行 Hadoop 作业的工作流。

此外,还有集成开发环境和集成应用程序环境,前者提供源代码编辑器、编译器、自动化系统构建工具、调试工具等,后者提供基于图形用户界面组装完整的应用程序。

(5) 大数据分析与展示框架。

大数据分析与展示框架通过结合使用智能分析与展现工具实现大规模数据的分析和可视化,主流技术包括 Mahout、Hama 等。

Mahout: 提供分布式机器学习和数据挖掘算法库。

Hama: 提供基于 BSP 的超大规模科学计算框架。

(6) 大数据连接器。

大数据分析与展示需要与传统的关系数据库、数据仓库连接,主流的技术是 ETL,为关系型数据库与 HDFS 的数据交互提供专门接口,同时提供对元数据、数据质量控制等可视化支持^[14]。

(7) 大数据安全、运维框架。

大数据安全、运维框架提供大数据治理、安全性以及日常的维护管理支持,主流技术包括如 Ambari、Chukwa 等。

Ambari: 提供 Hadoop 管理工具,快捷地监控、部

署、管理集群。

Chukwa:提供大规模分布式集群的数据收集管理。

嵌入 Hadoop 管理:支持 Hadoop 运行管理,包括日志审计、文件系统检查、数据节点块验证、性能监控、元数据备份等等。

GangliaContext:提供差大集群的开源分布式监控系统。

5 关键技术

5.1 大数据存储与处理技术

目前,在主流的大数据存储与处理平台领域仍然以 Hadoop 技术为主,提供对大数据分布式存储、计算、快速访问等支持,便于用户快捷地处理大规模数据。

Hadoop 的特点在于能够存储并管理 PB 级数据,支持处理非结构化复杂类型数据,同时由于采用分布式体系架构,Hadoop 具有很好的容错性和扩展性。

Hadoop 主要提供分布式存储和计算平台,包括分布式文件系统(HDFS)、分布式计算框架(MapReduce)和分布式数据库(NoSQL)。

(1)HDFS。

一个分布式文件系统(非结构化数据存储),隐藏下层负载均衡,冗余复制等细节,对上层程序提供一个统一的文件系统 API 接口。HDFS 针对海量数据特点做了特别优化,包括:超大文件的访问,读操作比例远超过写操作,PC 机极易发生故障造成节点失效,等等。

(2)MapReduce。

分布式存储运算可以抽象为 MapReduce 操作。Reduce 把 Key/Value 合成最终输出 Output。Map 是把输入 Input 分解成中间的 Key/Value 对,下层设施把 Map 和 Reduce 操作分布在集群上运行,并把结果存储在 HDFS 文件系统中。

(3)NoSQL。

NoSQL 是按列存储的、多维表结构的实时分布式数据库,可以提供大数据量结构化和非结构化数据的高速读写操作,为高速在线数据服务而设计。

5.2 大数据查询与分析技术

对于 Hadoop 存储的数据,无法通过 SQL 来查询使用。但是为了便于 SQL 使用人员能够通过 SQL 语言操作和分析大数据,SQL on Hadoop 技术因运而生,典型技术是 Hive 以及在 Hive 基础上扩展形成的 PostgreSQL、DRAWN Scale、Salesforce Phoenix 等等。

Hive 是基于 Hadoop 的大数据分布式数据仓库引擎,可以将数据存放在分布式文件系统 HDFS 或分布式数据库中,并使用 SQL 语言进行海量数据统计、查询和分析操作。

5.3 大数据分析可视化技术

大数据对数据分析和可视化提出了更高的要求,包括:要求数据分析从传统联机分析处理和报表向数据发现转变;要求从结构化数据向结构化、非结构化混合数据类型转变;要求支持 PB 以上的大数据进行分析;要求支持对关系型、非关系型、多结构化、机器生成的数据分析;要求支持重组数据成为新的复杂结构并进行分析和可视化,如图分析、时间/路径分析;要求大数据分析支持更快、更适应迭代的分析。

因此,在大数据环境下,通过传统的数据采样、OLAP 方式的数据分析,已经难以满足大数据分析的需要。必须要能对全量的数据集进行多样化的数据分析、挖掘,同时以适合的展示方式提供用户使用。

(1)大数据挖掘和高级分析。

大数据挖掘的主要任务就是从大量数据中发现模式,并且根据挖掘任务不同分为多种类型,包括:

关联分析:用于发现不同数据集之间的关联规则,包括关联、相关性、因果结构或频繁出现的模式。

分类分析:根据数据的特征为每个类别建立一个模型,根据数据属性将数据分配到不同组中。

聚类分析:按照某种相近程度度量方法将数据分成互不相同的分组。

序列分析:用于分析数据中某类与时间相关的数据,并挖掘时序模式、周期性、趋势和偏离等。

偏差检测分析:用于检测并解释数据分类的偏差,即数据集中显著不同于其他数据的对象。

预测模型分析:从数据集中已知的数据推测未知的数据集中某些属性值的分布。

模式相似性挖掘:用于在时间数据库或空间数据库中搜索相似模式时,从所有对象中找出用户定义范围内的对象,或找出所有元素对中两者距离小于用户定义的距离范围的元素对。

此外,Mahout 提供了基于 Hadoop 的分布式机器学习和数据挖掘库,便于大数据挖掘以及高级分析实施。

(2)非结构化数据分析。

非结构化数据包括文本、Web 页面、语音、视频、图像等,因此,在大数据环境下,需要具备对大规模的非结构化数据进行分析、利用,包括:文本挖掘、Web 数据挖掘、语音识别、图像识别与分析、地理空间分析。

(3)实时预测分析。

描述性分析主要是帮助用户了解过去发生的事情,而预测性分析专注于正要发生的事情,并进一步预测将来可能发生什么。

(4)大数据可视化。

大数据可视化主要包括可视化报表和可视化分

析。其中,可视化报表主要使用图、表描述业务情况,主要工具有仪表盘、报告、基于 Excel 分析、维度分析等等。而可视化分析是辅助用户可视化地探索数据来发现新的洞察,能够帮助用户按照思维的速度可视化地过滤、比较和关联数据,从而更快地分析、更好地决策和更有效地展现和理解数据。

主要工具包括:入门级工具-Excel、在线数据可视化工具、互动图形用户界面控制、基于地理信息系统工具、可视化设计工具、可视化分析工具。

6 结束语

目前随着电力信息化的迅速推进以及智能变电站、实时监测系统、现场移动检修系统、测控一体化系统等一大批服务于各个专业的信息管理系统的建设和应用,数据的种类和规模同步快速增长,这些多源异构数据共同构成了智能电网大数据。未来将大数据技术适用在智能电网等信息领域,提供重要指挥决策信息可视化平台,辅助指挥人员提升指挥能力。

(1)以应用需求为牵引,持续深化应用信息顶层规划。

从不同指挥层级应用人员的应用需求出发,加强应用信息顶层规划,持续支持与应用需求相关数据采集渠道接入,建立应用数据分类目录体系标准与信息交换共享机制,确保多数据采集渠道的大规模应用数据能够有效整合、有序组织,统一为应用信息服务提供数据资源调配与使用服务。

(2)以应用为核心,紧密围绕应用保障知识化需要。

针对现有应用数据基础性数据多、决策支持性数据少等问题,在应用信息顶层规划与数据有效整合的支持下,综合运用数据统计、数据分析、数据挖掘等方法,从多数据采集渠道获取的大规模应用数据中,提炼能够保障服务于应用指挥的决策支持信息知识,提升应用行动快速反应能力。

(3)以信息用户为中心,重点突出多层级信息服务能力。

面对复杂、动态、不确定的应用环境,能够支持便携式终端快速接入、准确获取信息,同时能够面向不同

应用人员提供个性化的信息服务,支持根据不同应用人员重点关注内容,灵活组织、多视角全方位展示信息内容,满足不同指挥层级应用人员便捷、及时地掌握应用信息的保障需求。

参考文献:

- [1] 史英杰,孟小峰.云数据管理系统中查询技术研究综述[J].计算机学报,2013,36(2):209-225.
- [2] 金培权,郝行军,岳丽华.面向新型存储的大数据存储架构与核心算法综述[J].计算机工程与科学,2013,35(10):12-24.
- [3] 孟小峰,慈 祥.大数据管理:概念、技术与挑战[J].计算机研究与发展,2013,50(1):146-169.
- [4] 李国杰,程学旗.大数据研究:未来科技及经济社会发展的重大战略领域大数据的研究现状与科学思考[J].中国科学院院刊,2012,27(6):647-657.
- [5] HOOD L, FLORES M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory[J]. New Biotechnology, 2012, 29(6): 613-624.
- [6] KHOURY M J, GWINN M L, GLASGOW R E, et al. A population approach to precision medicine [J]. American Journal of Preventive Medicine, 2012, 42(6): 639-645.
- [7] TAUBES G. Epidemiology faces its limits [J]. Science, 1995, 269(5221): 164-169.
- [8] LOOS R J, SCHADT E E. This I believe: gaining new insights through integrating "old" data [J]. Frontiers in Genetics, 2012, 3: 137.
- [9] 余 侃.云计算时代的数据中心建设与发展[J].信息通信, 2011(6): 100-102.
- [10] 靳强勇,李冠宇,张 俊.异构数据集成技术的发展和现状[J].计算机工程与应用, 2002, 38(11): 112-114.
- [11] 冯启蒙,王振辉,王振铎.基于数据库的XML存储技术设计和实现[J].计算机系统应用, 2006, 15(9): 32-34.
- [12] 郑 重,黄天纳,刘运成,等.军队数字化卫勤建设系列讲座(3) 建设军队数字化卫勤数据中心的构想[J].人民军医, 2011, 54(10): 931-932.
- [13] 张新兰,曲 江.企业数据仓库系统 in 管理决策中的应用[J].中国管理信息化, 2006, 9(10): 16-19.
- [14] SPAETH D A. Representing text as data: the analysis of historical sources in XML [J]. Historical Methods, 2004, 37(2): 73-86.