大数据分析平台建设与应用综述

王 强[1] 李俊杰[1] 陈小军[1] 黄哲学[1] 陈国良[2]

[1](深圳大学大数据技术与应用研究所 深圳 518060)
[2](深圳大学高性能计算研究所 深圳 518060)

摘 要 大数据分析平台是开展大数据处理与分析应用所必需的基础设施。文章基于课题组开展大数据分析平台建设的科研成果与实践经验，结合大型企业实施行业应用项目的切身感受，从大数据分析平台设计、主流热点技术、行业应用案例三个方面进行介绍。文章首先分析了大数据分析平台的主要功能和体系架构，然后介绍了大数据分析平台的关键技术，重点介绍了 Spark 技术的体系架构及核心组件，最后介绍了大数据技术在大规模制造业、零售业和智能电网三个领域的应用案例。

关键词 大数据平台；大数据分析；大数据应用；内存计算
中图分类号 TP 391.4 **文献标志码** A

Review on Construction and Application of Big Data Analytical Platform

WANG Qiang[1] LI Junjie[1] CHEN Xiaojun[1] HUANG Zhexue[1] CHEN Guoliang[2]

[1](*Big Data Institute*, *Shenzhen University*, *Shenzhen* 518060, *China*)
[2](*High Performance Computing Institute*, *Shenzhen University*, *Shenzhen* 518060, *China*)

Abstract The big data analytics platform is an indispensable infrastructure for big data processing and applications. Based on our research activities, practical experiences with big data analytics, and lessons learnt from industrial projects, this paper addressed the platform design, mainstream technologies, and industrial cases of big data analytics platforms. Firstly, the main functions and architecture of such platforms were analyzed. Then the key enabling technologies were introduced with a focus on the architecture of Spark and its core components. Finally three application case studies were presented in the areas of massive manufacture, retail, and smart grids.

Keywords big data platform; big data analytics; big data application; Spark

收稿日期：2015-12-23 **修回日期**：2015-12-27
作者简介：王强，博士后，研究方向为聚类算法和生物信息学；李俊杰，副教授，研究方向为数据挖掘与机器学习；陈小军，博士，研究方向为数据挖掘与机器学习；黄哲学(通讯作者)，特聘教授，研究方向为数据挖掘与机器学习，E-mail：zx.huang@szu.edu.cn；陈国良，教授，院士，研究方向为高性能计算。

1 引　言

当前，人类社会信息化进程正在迈向网络化信息技术普及阶段。整个社会的信息采集渠道日益丰富，信息应用广度不断拓展，信息总量呈指数级增长，以信息为核心的创新驱动力持续增强，从而带来全社会信息在类型多样性、关系复杂性、应用时效性等方面呈现出崭新的趋势和特征。这种由社会信息环境的变革而引发的社会数据环境的变革，给信息科学及相关产业发展带来了巨大的挑战和机遇。

大数据就是为有效应对“网络时代海量复杂数据带来的管理与应用难题”而产生的一种新的思维方式、技术体系和创新能力，其特有的战略意义和核心价值主要表现在以下三个方面：

第一，在战略思维层面，数据已经成为全球社会公认的创新发展要素，大数据已经从商业领域上升到国家战略层面。

自 2011 年 6 月麦肯锡公司发布了《大数据：下一个竞争、创新和生产力的前沿领域》[1]的研究报告，拉开了全球竞相发展大数据的序幕。随后，美、英、法、澳、日、韩等发达国家，以及联合国、欧盟、八国集团等国际组织，纷纷提出国家级或区域性大数据发展战略，旨在提升从大量复杂数据中获取知识和洞见的能力，进而促进政府治理效能和经济发展活力的显著提升。我国自 2012 年起，从中央部委到地方省市，连续密集地出台了十余个与大数据相关的发展规划和行动计划，特别是国务院于 2015 年 8 月出台了《促进大数据发展行动纲要》[2]，明确提出了政府率先开放政务大数据并强化与社会各方形成合力的相关任务和计划时间表，更加突显我国发展大数据的意志与决心。

第二，在信息科学与技术创新发展层面，大数据给传统的信息科学与技术体系带来了全方位的挑战，大数据科学正在加速形成以数据为核心的新的理论与技术体系。

大数据所特有的类型多样、混合异构、快速增长、体量巨大、关系复杂、高维稀疏等特性，导致传统的来源于多元统计、人工智能、机器学习、模式识别等领域的数据分析理论，以及以数据为核心的存储、索引、融合、处理、分析、应用、安全等全过程技术，亟待实现全面系统的创新与发展，不断形成和完善大数据科学与技术体系。同时，从大数据工程技术创新发展的角度，亟待将大数据相关的理论、技术成果与国际主流的大数据工程技术框架相结合，针对互联网应用的智能化和服务化的发展趋势，以及离线分析与在线分析的应用特点，围绕 Hadoop、Spark（内存计算）等当前热门主流的大数据工程技术体系，开展大数据平台开发与产业化应用，是促进大数据科技发展的另一项必要和紧迫的工作。

第三，在经济社会创新发展层面，大数据是保障我国“互联网+”和“智慧城市”战略实现的核心能力，并为推进“双创”战略提供了广阔的发展空间。

以应用为导向、以应用为引领，是大数据技术创新与发展的主要特征。当前，我国正在全力推进“互联网+”和“智慧城市”发展战略，大数据作为其中必不可少的使能性技术，将在城市虚拟空间的各种应用场景中发挥着信息整合、知识挖掘、业务协同、服务创新的作用。其中，大数据分析与应用平台更是作为大数据时代必备的基础设施：通过不断汇聚技术创新成果，为应用创新提供一站式共性基础服务，有效降低应用技术门槛，支持创业公司和创客群体在平台上开展不同领域、不同层次、不同环节的应用服务创新，加速形成以平台为核心的产业创新生态圈和产品化应用解决方案，促进大数据产业加快形成。

大数据时代，我国拥有得天独厚的发展优势。一方面，在政府大力倡导和全社会积极努力下，大数据已经成为全社会的共识，大数据所

蕴含的经济价值和创新价值已经引起社会各界的高度关注。另一方面，我国拥有海量丰富的数据资源，广阔多样的应用场景，潜力巨大的消费市场，为大数据创新与发展提供了必要条件。当务之急是如何快速有效突破数据价值挖掘的瓶颈。大数据分析与应用平台，是大数据时代必备的基础设施，也是突破当前技术瓶颈的有效突破口。开发和建设大数据分析与应用平台将带来三个方面的价值：(1)有助于不断汇集大数据技术创新成果，并用最先进的技术为用户提供一站式的应用服务；(2)有助于降低用户技术门槛，为应用开发提供共性基础设施与服务，从而加快应用创新；(3)有助于形成大数据技术产品和行业解决方案，促进我国大数据产业加快形成。

本文基于深圳大学大数据技术与应用研究所大数据分析平台课题组(以下简称“课题组”)近年来开发和建设大数据分析与应用平台的科研成果和实践经验[3,4]，同时结合课题组在人才培养、科学研究、社会服务等方面的实际感受，首先介绍了大数据平台的总体功能、体系架构及其关键技术；其次，针对当前大数据领域的前沿热点技术，重点介绍了 Spark 技术架构及其核心模块；最后，介绍了课题组已经完成的在大规模制造业、零售业和智能电网三个领域的大数据应用案例，以期为学术界和产业界提供具有一定参考借鉴价值。

2 大数据分析平台

2.1 大数据分析平台发展现状

大数据分析平台是建设和实施大数据应用所必需的基础设施，也是目前国际产业界竞相发展的前沿和热点领域。从目前全球发展现状来看，大数据分析平台建设与应用的主要力量来自于传统信息技术(Information Technology，IT)企业、新兴互联网企业、高校科研院所三大阵营，以下对其发展情况和代表成果进行概括总结。

2.1.1 传统信息技术巨头的大数据平台战略

该阵营以 IBM、ORACLE、SAP、EMC、Teradata 等传统 IT 巨头为代表，凭借长期积累的技术、产品、品牌、服务等全球领先的综合实力为基础，通过“硬件+软件+数据”整体解决方案向用户提供以平台为核心的完备的大数据基础架构与服务，同时通过密集地并购大数据分析创新型企业，以迅速增强和扩展在大数据分析领域的实力和市场份额。

国际 IT 巨头的大数据平台战略实施案例包括：

(1)IBM

①企业并购：收购了商务智能软件供应商 Congnos[5]、统计分析软件 SPSS[6]、数据库分析供应商 Netezza[7]；

②大数据管理：结合 IBM DB2 数据库，推出了支持 Apache Hadoop 的 InfoSphereBigInsights[8] 软件，支持大数据应用开发与实施；

③大数据一体机：发布了大数据一体机 Pure Data[9]，作为大数据领域的软硬件一体化解决方案。

(2)ORACLE

大数据一体机：该一体机集成了 Oracle Exalogic[10]中间件云服务器、Oracle Exadata[11]数据库云服务器和 Oracle Exalytics[12]商务智能云服务器，成为 ORACLE 企业级大数据解决方案。

(3)HP

企业并购：通过收购 Vertica 公司，推出针对大数据的 Vertica 6.1[13]数据分析平台，平台覆盖了非结构化大数据存储管理、处理分析、服务交付等全过程，成为企业级大数据应用的完整解决方案。

(4)EMC

大数据一体机：对原有的 EMC 硬件和 Greenplum 软件进行整合，推出了 Greenplum 一

体机产品[14]，平台适用于大数据分析场景，可以通过增加节点方式进行横向扩展，从而有效控制成本和性能。

整体平台解决方案厂商依靠自身原有的软件、硬件或技术优势，通过收购及整合不同公司的产品线，实现对大数据各个领域的覆盖。但是这种增量式的系统整合，只是使系统功能的体量增加。只有通过对自身产品和技术的原始创新，才能实现对大数据处理问题的彻底解决。

2.1.2 新兴互联网巨头的大数据平台战略

该阵营以 Google、Amazon、Facebook、阿里巴巴、百度、腾讯等互联网公司为代表，基于自身的应用平台、庞大用户群和海量用户信息，形成独有的互联网大数据应用生态圈，不断创新应用和商业模式，不断创造新价值。

(1) Google

①Google 提出的 GFS、MapReduce 和 BigTable 等大数据核心技术，催生了大数据处理的事实标准 Hadoop。目前，Google 通过自身开发的 Caffeine[15]平台，直接将索引放置在由 Google 开发的分布式数据库 BigTable 上；

②Google 还提供大数据虚拟服务器业务，用户可以把数据上传到 Google，Google 提供了包括 BigQuery[16]和 Google Compute Engine[17]等服务和基础设施运行用户的查询服务。

(2) Amazon

Amazon 弹性 MapReduce（Amazon Elastic MapReduce）[18]，是一项能够迅速扩展的 Web 服务，运行在亚马逊弹性计算云（Amazon EC2）和亚马逊简单存储服务（Amazon S3）上，用于满足数据密集型任务（如互联网索引、数据挖掘、日志文件分析、机器学习、金融分析、科学模拟和生物信息学研究），平台将根据用户需要立即配置和满足资源需求。

(3) Facebook

①Corona（日冕）平台[19]，可以让你在数目庞大的 Hadoop 服务器之间运行大量的任务，并且不用担心软件错误会导致整个服务器集群崩溃；

②Prism（三棱镜）[20]平台，可以自动复制数据，并在不同地点的服务器之间传输数据。这可以让 Hadoop 服务器集群运行在全球范围内的多个数据中心上，实现集群规模的灵活扩展。

(4) 阿里巴巴、百度、腾讯

①早在 2011 年，阿里巴巴就已经推出了“淘宝指数”[21]，商家可以根据以往的销售信息和“淘宝指数”进行生产、库存决策，同时，消费者也能以更优惠的价格购买商品；

②百度正开展大数据革命以应对企业时代需求，其已从数据、工具及应用三个层面布局大数据时代企业战略规划，为用户更深入地挖掘数据价值，优化营销决策；

③腾讯主要通过深入挖掘用户属性，培育社会化营销平台，利用大数据和关系链，为用户筛选、推荐最适合他的内容。

互联网公司在大数据领域的创新主要是基于自身的数据和业务需求，主要集中在搜索、个性化推荐和存储、计算等方面。但是对于“人、机、物”三元融合技术产生的多样化海量复杂数据，仍然需要新的分析平台及处理技术。

2.1.3 科研领域的大数据平台发展状况

国际顶级期刊《Nature》和《Sciences》近期针对大数据分别出版了专刊《Big Data》[22]和《Dealing with Data》[23]，从互联网技术、互联网经济学、超级计算、环境科学、生物医药等多个方面讨论了大数据处理面临的各种问题。

在国内，中国计算机学会（China Computer Federation，CCF）成立了大数据专家委员会（CCF Big Data Task Force，简称 CCF TFBD）。2012 年 10 月 19 日，中国计算机学会大数据专家委员会成立，通过竞选产生了以李国杰院士为主任的专家委员会的第一任领导班子。2012 年 11 月 30 日～12 月 1 日，中国 Hadoop 与大数据技术大会

(HBTC 2012)在北京成功举办。大会以“大数据共享与开放技术”为主题，讨论了大数据共享平台与应用、大数据的技术挑战与发展趋势。

目前，国际学术界研发的大数据平台的代表成果包括：

(1)Petuum 大数据分布式机器学习平台[24]：平台由美国卡耐基梅隆大学(CMU)邢波教授课题组针对大数据机器学习特点研发，是一个分布式机器学习框架，提供了面向超大型机器学习的通用算法和系统接口。包含数据和模型并行两套功能，平台的参数服务器为开发者提供良好的编程环境，通过共享虚拟分布内存，在编程的时候不用对每个机器进行单独通讯；平台的调度器能够对模型进行有效的分割，甚至是动态分割，然后进行任务的分布化和载量平衡。

(2)PDMiner 基于云计算的数据挖掘软件平台[25]：平台由中国科学院计算技术研究所与中国移动合作开发，集成了 ETL 组件、数据挖掘组件以及多种算法，可有效解决多种云计算数据挖掘问题。平台的挖掘效率随节点增加而增加，多个任务工作流之间互不干扰，不同节点间可同时启动，具有容错能力，架构具有开放性，算法可方便地配置加载到平台上，达到了商用软件精度，成为中国移动 数据挖掘分析支撑工具。

(3)CLAIMS 并行数据分析系统[26]：系统由华中师范大学数据科学与工程研究院研发，提供了一个基于内存(in-memory)的并行数据库系统框架，可运行在服务器集群中，提供面向关系型数据的实时数据分析。

(4)深圳大学大数据分析平台：平台由深圳大学大数据技术与应用研究所研发，也是本文主要介绍内容，详见后文。

2.2 大数据分析平台的总体功能

课题组构建的大数据分析平台的主要目标是为大数据技术研发和应用项目实施提供高效完备的开发与运行环境。为此，大数据分析平台的总体功能包括以下主要方面：

(1)云计算环境：整个平台基于云计算环境，主要包括：云存储、云资源调度与管理、云计算编程模型、云计算执行引擎等核心功能，支持对海量数据的存储、处理、建模、分析、展现等全过程的分布式并行化开发与运行；

(2)面向 SaaS 服务的开放式体系架构：整个平台采用开放式体系架构，支持插件式开发与集成，提供底层核心功能的 API 调用接口，为第三方开发提供高可扩展的平台环境，基于平台开发的应用可以 SaaS 服务形式提供给用户使用；

(3)多源异构数据集成：平台提供丰富的数据集成接口，支持与传统的关系型数据库产品以及互联网、物联网应用系统的数据采集接口的无缝集成，便于将多源异构数据导入到平台数据存储系统；

(4)海量数据云存储管理：提供 PB 级结构化和非结构化数据云存储与管理，支持高效的数据查询、索引、提取等基本数据集操作；

(5)高效数据 ETL 处理：提供分布式并行的 ETL 处理工具，全面支持数据质量问题处理；

(6)基于 WEB 的分析建模：提供基于 WEB 方式和基于工作流的数据挖掘建模系统，便于建模分析人员随时随地在线编辑和提交分析模型；

(7)离线分析与在线分析：提供以 Hadoop 为基础的离线分析环境和以 Spark 为基础的在线分析环境，满足不同应用场景下对数据分析响应效率的需求；

(8)知识库：平台提供算法库、模型库与案例库，支持用户将数据挖掘算法、分析模型及应用案例进行编辑和重用，不断积累成为用户知识库；

(9)可视化报表系统：平台提供可视化分析与报表系统，用户通过可视化分析工具、可视化引擎、报表模板等功能开展交互式可视化数据分析。

集成了上述核心功能的大数据分析平台，一方面可以有效支持科研工作者开展算法研究、模型设计、系统优化等探索性研发工作，并快速将研发成果集成到平台中，不断提升平台的技术先进性；另一方面可以有效支持企业级大数据应用系统的运营，以及第三方应用开发与扩展，促进行业应用解决方案不断成熟与完善。

2.3　大数据分析平台的体系架构

大数据分析平台的设计理念是以区域性智能数据中心和高速互联网为基础设施，以互联网服务体系为架构，以大规模海量数据存储、处理、挖掘和可视化分析等关键技术为支撑，通过多样化智能终端及互联网为用户提供数据存储、管理及分析服务。

大数据分析平台的拓扑架构如图 1 所示。区域智能数据中心提供基于云计算的大规模数据存储及数据挖掘平台，通过平台服务器对外接口提供数据存储、分析与挖掘服务。用户使用 Web 浏览器或智能终端应用程序提出数据存储和分析的服务请求，经 Web 服务器通过互联网将服务请求发送给数据中心平台服务器，平台服务器对服务请求进行解析，发送给工作流引擎调度执行，执行结果通过互联网发送给用户终端。

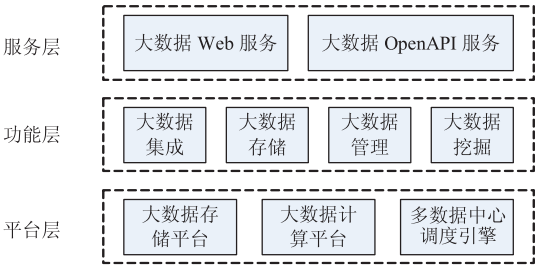

图 1　大数据分析平台体系架构

Fig. 1　Architecture of big data analytical platform

2.4　大数据分析平台的关键技术

本文提出的大数据分析平台主要包括以下关键技术：

(1) 平台层

①大数据分布式存储系统：针对数据不断增长的挑战，需要研究大规模、非结构化数据的存储问题，突破大数据的存储、管理和高效访问关键技术，当前需要构建至少 PB 级存储能力的大数据平台才能满足一般的科研和应用需求；

②分布式数据挖掘运行时系统：针对大数据挖掘算法运行的挑战，突破 MapReduce 技术的局限，研究有效支持迭代、递归、层次及集成机制的海量数据挖掘编程模型和运行时系统，构建大数据运行时系统；

③智能数据中心联合调度技术：针对大数据存储和挖掘的挑战，研究多数据中心的智能联合调度、负载均衡技术，整合多个数据中心的存储和计算资源，构建基于多智能中心的大数据服务平台。

(2) 功能层

①高可扩展性大数据挖掘算法：针对大数据挖掘的挑战，研究基于云计算的分布式大数据处理与挖掘算法，构建高可扩展的大数据处理与挖掘算法库，实现 TB 级数据的建模能力；

②大数据安全与隐私保护技术：针对数据挖掘“软件即服务”(SaaS)模式的需求，研究开发数据挖掘在云环境下的隐私保护、数据审计和节点数据挖掘技术，确保大数据挖掘过程中的数据安全，保证用户的隐私不被泄露；

③分布式工作流引擎：针对大数据挖掘分布式调度的挑战，研究基于云计算的分布式工作流调度、负载均衡技术，构建高效分布式工作流执行引擎；

④交互式可视化分析技术：针对传统分析方法交互性和可理解性不足的问题，研究启发式、人机交互、可视化数据挖掘新技术，实现大数据挖掘的高度人机交互功能。

(3) 服务层

①基于 Web 的大数据挖掘技术：突破传统的基于单机软件的数据挖掘技术，创新基于 Web 的大数据挖掘方法和流程，实现易于使用的基于

Web 的大数据挖掘技术，构建基于 Web 的大数据分析环境；

②基于 Open API 的大数据挖掘技术：突破传统的基于软件的数据挖掘技术，创新基于 Open API 的大数据挖掘方法，研究大数据挖掘开放接口、开放流程，构建基于 Open AIP 的大数据分析模式。

为广大用户提供大数据处理和分析的服务功能，大数据分析平台要突破传统的基于软件和高端服务器的数据挖掘传统技术体系，采用基于云计算的大数据存储和处理架构、分布式数据挖掘算法和基于互联网的大数据存储、处理和挖掘服务模式。实现这一目标需要做如下创新：

(1) 系统架构创新：突破传统的基于软件和高端服务器的数据挖掘技术体系，研发基于互联网和云计算的大数据存储、处理和挖掘的数据中心系统架构，支持多用户、多任务的大数据分析环境；

(2) 服务模式创新：突破传统的一次性软件销售或软件租赁的高价格解决方案，创新基于互联网的大数据存储、处理和分析服务模式，为用户提供按需、廉价的大数据存储、处理和分析服务；

(3) 使用模式创新：突破传统的使用单机软件的方式，创新基于互联网的大数据存储、管理和分析服务，提供多终端(台式机、笔记本、平板电脑、手机等)、多途径(浏览器访问、Open API 调用等)的用户使用模式。

2.5 大数据分析平台的实践案例

根据大数据分析平台的总体功能要求(详见 2.1)，课题组自主搭建了大数据分析平台，平台的硬件拓扑结构如图 2 所示。

(1) 平台的核心硬件资源配置包括：

①存储资源：2 台一体化 NAS 存储设备，数

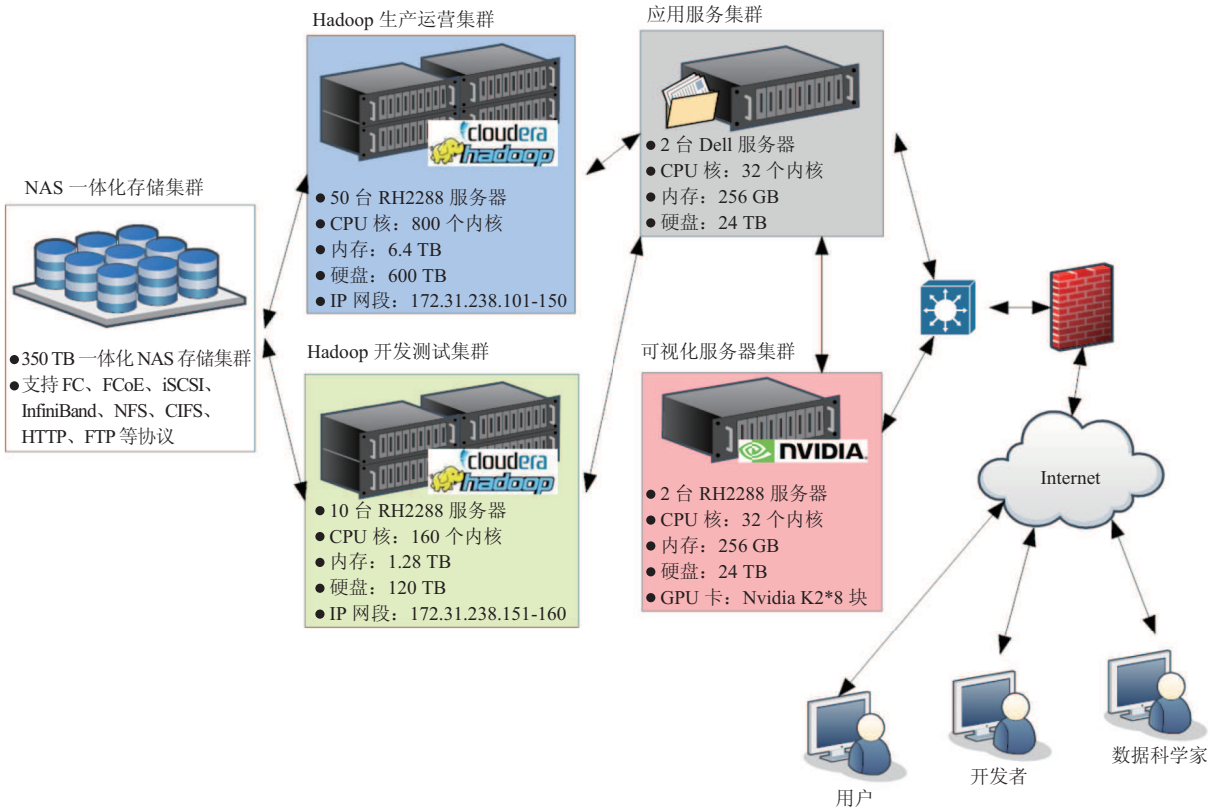

图 2 大数据分析平台拓扑结构

Fig.2 Topology of big data analytical platform

据存储总量 350 TB；

②计算资源：60 台 RH2288 服务器，CPU 内核共计 960 核，内存累积近 8 TB，硬盘存储总量超过 700 TB；

③ 网络资源：华为交换机，防火墙。

(2) 平台的核心软件资源配置包括：

① 操作系统：Ubuntu 14.04；

② Hadoop 管理软件：Cloudera 5.3；

③ 数据可视化软件：Tableau 9.0；

④ 报表系统：Fine Report；

⑤ 数据挖掘软件：Matlab，R。

为有效保障技术研发测试和应用系统运行两种场景的不同需求，大数据平台的拓扑结构从逻辑上划分为两大集群，即：基于 Hadoop 的开发测试集群和基于 Hadoop 的生产运营集群。

3　Spark 技术

Spark 是当前大数据技术的重要组成部分，近年来日益引起国际学术界的重视[27-30]。本节介绍了 Saprk平台的基本概念及关键技术。

3.1　Spark 简介

传统的 Hadoop 平台由于频繁的磁盘读写操作导致其不适合处理迭代式计算任务，同时也不适合处理对时间要求高的计算任务。为此，加州大学 Berkeley 分校研发了新一代大数据处理平台 Spark。针对迭代任务的需求，Spark 可以将数据存储在内存中以避免频繁的磁盘读写，从而提高了计算效率。Spark 主要由 Scala 编写，支持通过 Java、Scala、Python 及 R 来使用，官方测试表明其速度可以比 Hadoop 快 10～100 倍(详见 http://spark.apache.org/)。

Spark 的主要特点包括：

(1) 提供 Cache 机制来支持需要反复迭代计算或者多次数据共享，减少数据读取的 IO 开销；

(2) 提供了一套支持 DAG 图的分布式并行计算的编程框架，减少多次计算之间中间结果写到 Hdfs 的开销；

(3) 使用多线程池模型减少 task 启动开稍，shuffle 过程中避免不必要的 sort 操作并减少磁盘 IO 操作。

3.2　BDAS

BDAS 的全称为 Berkeley Data Analysis Stack，是加州大学 Berkeley 分校将基于 Spark 的整个大数据生态系统称为伯克利数据分析栈。BDAS 的体系架构如图 3 所示，其核心框架是

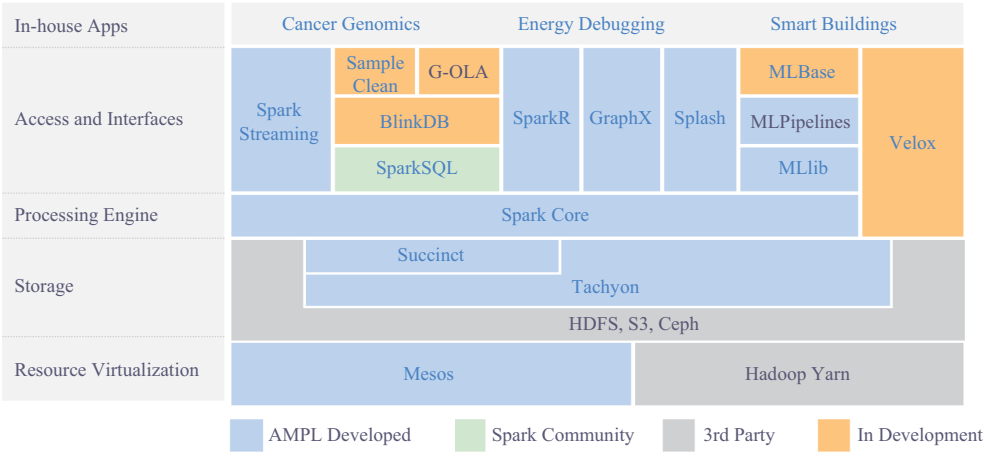

图 3　伯克利数据分析栈逻辑结构

Fig. 3　Logical architecture of Berkeley data analysis stack

Spark，同时主要包含以下基于 Spark 的大数据处理系统：

(1) Mesos：Mesos 是一个资源管理框架，提供类似于 YARN 的功能。用户可以在其中插件式地运行 Spark、MapReduce、Tez 等计算框架的任务。Mesos 会对资源和任务进行隔离，并实现高效的资源任务调度。

(2) Tachyon：Tachyon 是一个分布式内存文件系统，可以理解为内存中的 HDFS。为了提供更高的性能，将数据存储剥离 Java Heap。用户可以基于 Tachyon 实现 RDD 或者文件的跨应用共享，并提供高容错机制，保证数据的可靠性。

(3) Succinct：Succinct 支持对压缩数据不进行解压缩而直接进行搜索、范围查询及随机访问。

(4) Spark SQL：Spark SQL 提供在大数据上的 SQL 查询功能，用户可以在 Spark 上直接书写标准 SQL 语句进行查询。Spark SQL 使用 Catalyst 做查询解析和优化器，并在底层使用 Spark 作为执行引擎实现 SQL 的 Operator。

(5) Spark Streaming：Spark Streaming 通过将流数据按指定时间片累积为 RDD，然后将每个 RDD 进行批处理，进而实现大规模的流数据处理。其吞吐量能够超越现有主流流处理框架 Storm，并提供丰富的 API 用于流数据计算。

(6) GraphX：GraphX 基于 BSP 模型，在 Spark 之上封装类似 Pregel 的接口，进行大规模同步全局的图计算。

(7) BlinkDB：BlinkDB 是一个用于在海量数据上进行交互式 SQL 的近似查询引擎。它允许用户通过在查询准确性和查询响应时间之间做出权衡，完成近似查询。其数据的精度被控制在允许的误差范围内。

(8) SparkR：SparkR 是一个 R 语言包提供了一个轻量级的前端，用于从 R 语言中使用 Apache Spark。SparkR 通过 RDD 类暴露 Spark API，允许用户以交互方式在集群上从 R shell 运行 Spark 任务。

(9) Splash：Splash 是一个用于在多核集群上对随机程序进行并行执行的通用框架。

(10) MLbase：MLbase 是一个基于 Spark 的通用分布式机器学习库，由三个主要的部件组成：MLlib、MLI 和 ML Optimizer。MLbase 提供了不同抽象程度的接口，用户可以扩充自己的算法。同时，MLbase 很容易上手，不同基础的用户都可以额很方便地使用它来对大数据进行分析。

3.3 弹性分布数据集

弹性分布数据集(Resilient Distributed Datase，RDD)是 Spark 对数据的一个基本抽象，是对分布式内存的抽象使用，实现了以操作本地集合的方式来操作分布式数据集的抽象实现。RDD 可以 cache 到内存中，每次对 RDD 数据集的操作之后的结果，都可以存放到内存中，下一个操作可以直接从内存中输入，省去了 Hadoop 执行迭代计算需要的大量磁盘 IO 操作。这对于迭代运算比较常见的机器学习算法、交互式数据挖掘来说，效率提升很大。

RDD 主要具有如下特点：

(1) 它是在集群节点上的不可变的、已分区的集合对象；

(2) 通过并行转换的方式来创建(如 map、filter、join 等)；

(3) 失败自动重建；

(4) 可以控制存储级别(内存、磁盘等)来进行重用；

(5) 必须是可序列化的；

(6) 是静态类型的。

RDD 有两种计算方式：转换(Transformations)及动作(Actions)。二者的区别是，转换的返回值还是一个 RDD，而动作的返回值不是一个 RDD。转换主要包括 map、filter、groupBy 及 join 等。Transformations 操作不是马上执行的，Spark 在遇到 Transformations 操作时只会记

录需要这样的操作，并不会去执行，需要等到有 Actions 操作的时候才会真正启动计算过程进行计算。动作主要包括 count、collect 及 save 等，将返回结果或把 RDD 数据写到存储系统中。

3.4　MLbase

MLbase 主要包括三个组件：

(1) ML Optimizer：自动调度机器学习任务的执行，能解决特征选择及机器学习任务的优化搜索问题。这个模块当前还在开发中。

(2) MLI：一个高度抽象的机器学习编程抽象接口，可用于开发自己的特征提取及机器学习算法。

(3) MLlib：包含已有的基于 Spark 的机器学习算法。

给定一个机器学习任务，MLbase 中的 ML Optimizer 会选择它认为最适合的已经在内部实现好了的机器学习算法和相关参数，来处理用户输入的数据，并返回模型或别的帮助分析的结果。这样，不了解 ML 的用户也能使用 MLbase 这个工具来处理自己的数据。用户可以容易地使用 MLbase 这个工具来处理自己的数据。

Spark 将机器学习算法都分成了两个模块：(a) 训练模块：通过训练样本输出模型参数；(b) 预测模块：利用模型参数初始化，预测测试样本，输出与测值。

MLbase 提供了函数式编程语言 Scala，利用 MLlib 可以很方便地实现机器学习的常用算法。比如：如果要做分类，只需要写如下 Scala 代码：

```
1 var X = load("some_data", 2 to 10)
2 var Y = load("some_data", 1)
3 var (fn-model, summary) = doClassify(X, Y)
```

其中，X 是需要分类的数据集；Y 是从这个数据集里取的一个分类标签；doClassify() 是执行分类操作。

4　应用案例

4.1　制造业大数据——产品制造质量监测预警平台

4.1.1 应用需求与特点

2015 年 5 月，国务院出台了《中国制造 2025》发展规划。在这份被誉为中国版“工业 4.0”的规划中，明确提出了“推进信息化与工业化深度融合”、“加强质量品牌建设”等重点建设任务和发展目标。在此背景下，课题组受我国领先的通信设备制造商委托，针对海量产品生产测试数据，开展大数据分析建模和预警算法的探索性研究，构建面向产品制造质量监测预警的大数据应用平台，促进项目委托方及时、精准地发现制造质量隐患，提升产品制造质量的监控预警能力。

目前，项目委托方的部分产品采用 JDM (Join Design Manufacture) 和 ODM (Outsouce Design Manufacture) 的生产模式，由多家代工工厂生产相关产品。为保证产品来料质量、生产过程工艺质量、批次产品良品率，避免因批量产品质量问题而发生召回事件，项目委托方通过在代工工厂安装测量监测设备，对制造过程的相关质量因素进行实时检测 (所采集的数据规模如图 4 所示)，并将相关测量数据返回到项目委托方的数据中心。

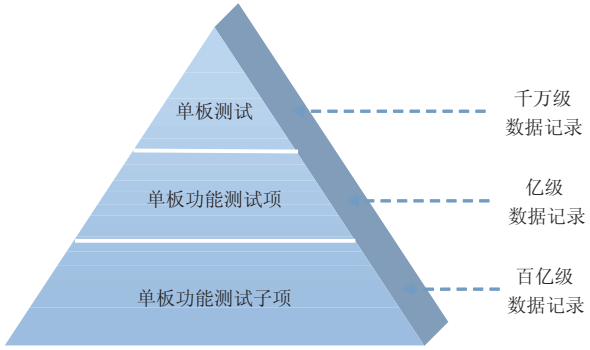

图 4　产品造质量测试数据规模

Fig. 4　Data scale level of product quality testing data

本项目的目标是针对海量的产品制造过程测

量数据，通过数据挖掘的技术手段，从人员、物料、工艺、设备、环境五大方面发现与批次产品质量相关的影响因素及其相互关系(产品制造质量因素发现与预警流程如图 5 所示)，设计相应的数据分析模型和工作流，并构建一套面向制造业海量质量检测数据分析平台。

4.1.2 应用案例设计与分析

为满足项目委托方的大数据应用需求，本项目设计并构建了如图 6 所示的大数据应用平台。该平台包括以下核心系统：面向海量数据分析的分布式文件存储系统、海量数据 ETL 引擎、流数据处理引擎、产品质量预警模型库、分析结果

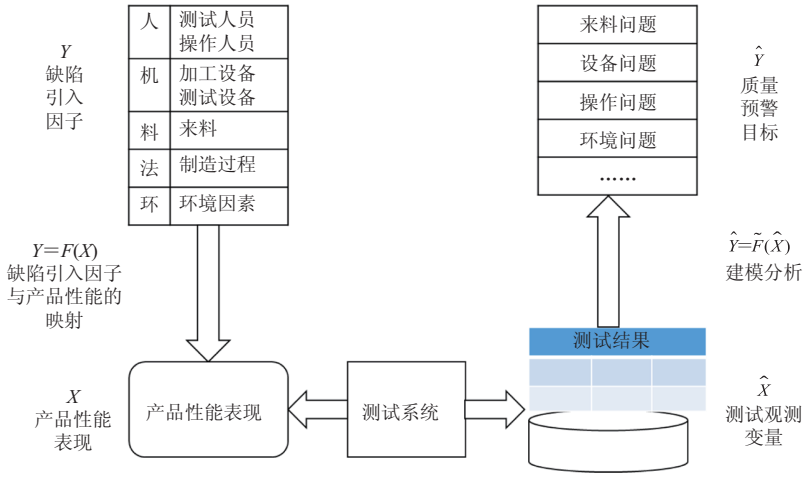

图 5 产品制造质量因素发现与预警流程

Fig. 5 Process of product quality factor detection and prediction

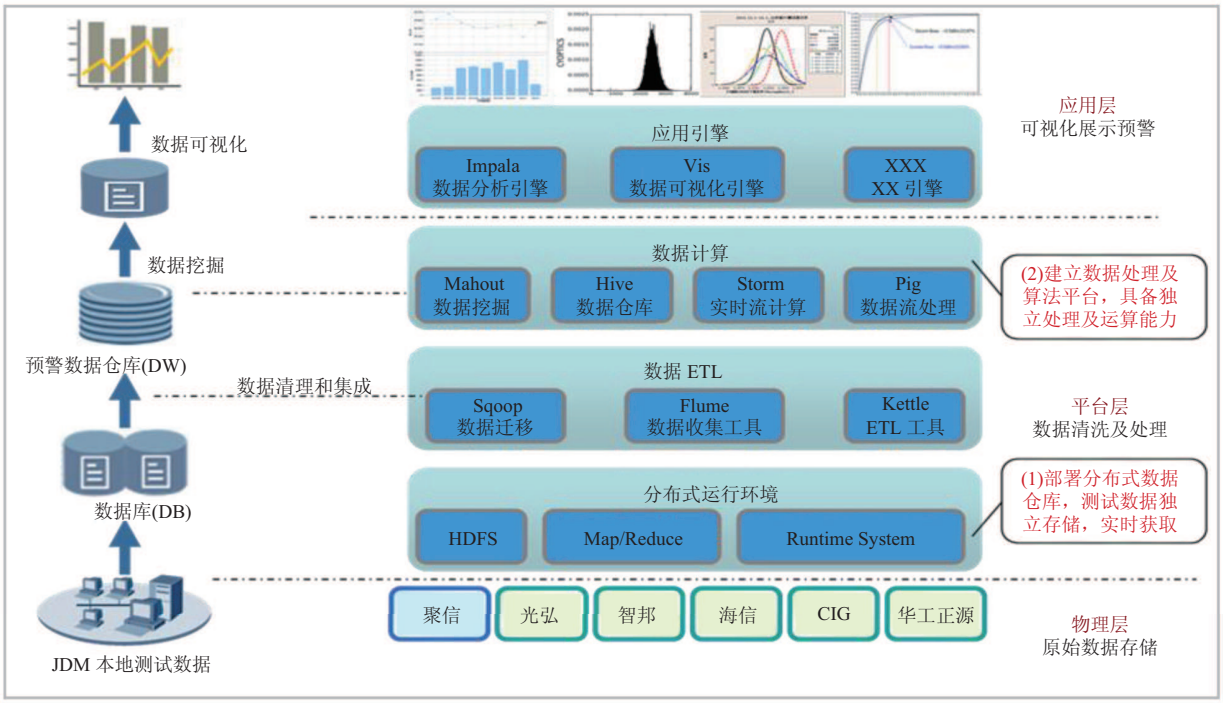

图 6 产品制造质量监测预警平台体系架构

Fig. 6 Architecture of product quality monitoring and predicting platform

可视化展现系统。

平台的技术体系与主要功能具有如下特点：

(1)平台体系架构：整个平台以当前业界成熟并广泛使用的 Hadoop 开源大数据架构作为数据存储和处理的基础架构，使用 HDFS 分布式文件系统来存储大数据，编写 Map/Reduce 分布式程序实现对大数据的处理与分析。

(2)海量异构数据存储：平台采用基于 HDFS 的 Cassandra 分布式数据库，实现对海量、异构、高速增长的数据进行存储管理。传统的数据库采用的是基于行的存储模型，而 Cassandra 采用的是基于列的存储模型，更适合高维大数据的存储和处理。由于两种存储模型不同，本项目将根据列存储模型及处理需求对现有的数据存储模型进行修改并优化其性能。同时，使用分布式数据仓库系统 Hive，设计满足多种分析需求的数据仓库系统。

(3)海量数据迁移与 ETL：针对数据仓库中典型的数据抽取、转换和加载任务，使用 Flume 系统将多种系统上的日志数据采集到 Hadoop 平台上，使用 Sqoop 大数据迁移工具将数据从现有的 Oracle 数据库迁移到 Hadoop 平台上，最后使用大数据 ETL 工具 Kettle 对存储在 Hadoop 上的大数据进行处理。

(4)数据挖掘：使用成熟的 Impala 大数据分析引擎，Storm 流数据处理系统，Vis 数据可视化引擎作为基础的分析引擎，采用 Mahout/Spark 等以及针对实际需求开发的基于 Map/Reduce 编程模型的分布式处理算法作为分析的基础算法库，以对大数据进行高效的分析处理。

在应用实施方面，首先对多种来源的大数据进行清洗处理，并整合成一个分布式数据库，以便于后续处理。接下来对数据进行清洗集成，根据业务需求设计构造数据仓库，以满足业务部门多样化的分析处理需求。最后采用数据挖掘技术对数据进行挖掘，并将挖掘结果用图表等多种可视化方式展示给用户。

更进一步，将大数据的集成、清洗、处理、挖掘、展示等环节的应用系统进行整合，构造一站式大数据应用平台。其中，分析结果将通过图标等多种可视化手段提供给用户使用，并与业务系统进行深度整合，从而满足项目委托方开展跨部门的大数据应用协同。

4.2　零售业大数据——基于产品树的购物篮分析

4.2.1 商品分类树

一个零售企业所出售的商品，通常组织成如图 7 所示的产品分类树。产品分类树的根节点为空的根节点，叶节点为具体的商品。除根节点外的其他非叶节点代表一个类别，这些类别具有层次结构，可以表达为如图 7 所示的分类树。

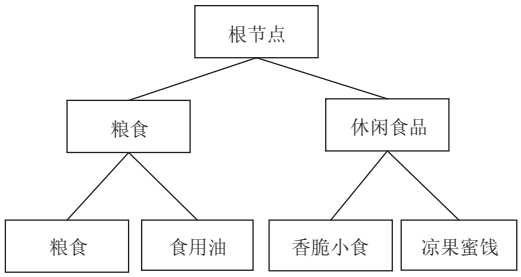

图 7　商品分类树

Fig. 7　Classification tree

4.2.2 基于商品分类树的关联分析

购物篮分析是关联规则在零售业的一个重要应用，它通过发现顾客每次放入购物篮中的商品之间联系，来分析顾客的购买行为并辅助零售企业制定营销策略。Apriori 算法[31]是一种经典的关联规则频繁项集挖掘算法，它使用一种称为逐层搜索的迭代方法来生成所有的频繁集，即用 k-项集来产生(k＋1)-项集。首先找出 1-项集的候选集，记为 C_1；然后根据最小支持度对 C_1 进行剪枝得到频繁 1-项集 L_1；再由 L_1 连接产生 2-项集的候选集 C_2，由 C_2 产生频繁 2-项集 L_2，循环下去，直到得到的 L_k 为空为止。

传统的购物篮分析方法不考虑商品的层次结构，如 Apriori 算法，通常获得的购物篮很多是

同一小类产品的组合。例如：可能得到这样的购物篮“苹果、梨子、香蕉、葡萄、牛奶”，其中“苹果、梨子、香蕉、葡萄”都属于“水果”。这种购物篮由于所包含的产品过于集中在某一小类中，应用价值不大，而真正有价值的购物篮被这种购物篮所淹没，难以被发现。

针对以上问题，课题组提出一种基于产品树的购物篮分析方法。在根据频繁集候选集 L_i 连接产生候选集 C_{i+1} 时，加入如下的约束条件：同一个购物篮中的产品属于不同的父类。在根据 C_{i+1} 生成 L_{i+1} 时，不仅考虑候选购物篮的支持度，同时也考虑候选购物篮的销售额。这样，最终得到的购物篮包含的产品分布在不同的类别中，并且销售额高。

表 1 为实验得到的购物篮分析结果示例。从表中可以看到，购物篮的组成符合我们的生活常识。例如，第一个购物篮为家居生活用品，第二个购物篮为零食小吃，并且在同一个购物篮中的商品都是属于不同的父类。这样的购物篮对企业具有更大的应用价值。

4.3 智能电网大数据——基于用电模式分析的用户分群

4.3.1 应用需求与特点

智能电网属于典型的大数据应用领域[32,33]。目前，我国电网公司不断强化在整个电网的输配电侧和用电侧安装和应用自动化数据采集装置，通过采集和分析关于电网运营和用户用电的数据，以提高电力资源的配置和使用效率。

课题组受某国家级电网公司委托，针对广东省某市多年积累的用电数据，开展“电力用户用电模式大数据分析”的应用项目。项目面临的主要应用需求和技术挑战包括：

(1) 多源异构数据融合：针对应用目标，项目将要处理和分析的数据包括用户信息、地理位置信息、电力设备信息、用电信息等具有不同来源和结构特点的海量数据，需要对这些数据进行一致性融合建模，以便为后继的建模分析提供主题数据集；

(2) 数据噪声预处理：由于采集设备和采集条件等因素的影响，在原始数据集中包含较多的缺失值、异常值等数据噪声，需要准确发现和有效处理各类噪声数据，以保障后继建模分析结果的有效性；

(3) 用电模式分析模型：需要针对电网公司经营管理需求，提炼不同应用场景下用电模式分析的具体目标，设计和构建用电模式分析模型；项目重点考虑不同行业、地域、时段下的用户用电模式的特征和差异，并据此实现用户分群和用电预测；

(4) 电力大数据应用平台：为实现电力大数据分析的自动化和规模化应用，项目将构建应用平台，实现对电力大数据采集、融合、预处理、建模、分析、报告的全流程的一站式应用。

4.3.2 应用案例设计与结果分析

根据项目委托方的应用需求和实际数据情况，课题组重点针对以下三种任务进行大数据技

表 1 基于产品树的购物篮分析实验结果示例

Table 1 Experimental results of market basket analysis based on product tree

购物篮编号	品类 1	品类 2	品类 3	品类 4	品类 5
1	洗浴用品	一次性用品	厨房保洁用品	家居清洁剂	厕用纸巾
2	口香糖	糕点	凉果蜜饯	饮用水	方便粉、面
3	巧克力	糕点	卫生巾	面包	凉果蜜饯
4	鸡蛋类	蔬菜其他	冰鲜	牛肉	鸡类
5	蔬菜其他	活鲜	牛肉	鸡类	佐餐酱菜

术的研发和实施：

(1)时间序列数据缺失值发现与处理

经过对数据的观察和探索，课题组发现在项目委托方积累的海量用电数据中，存在严重的缺失值问题，这也是传统行业大数据应用中普遍存在的问题。数据质量问题将严重影响数据分析模型的结果质量，因此，课题组首先要针对缺失值进行发现和处理。

图 8 展示了数据缺失值问题的典型情况。例如：对于 R1 和 R2 两条记录，在整个数据维度空间中(横向)，存在大量空白区域，说明记录存在严重的缺失值问题；而对于记录 A、B、C、D 四条记录，也存在较大比例的空白区域，说明记录的缺失值问题也非常明显；此外，在区域 AREA 中，存在大量空白区域，说明大多数记录都在该维度子空间中存在缺失值。因此，通过利用图 8 所示的方式，可以很明显地激发视觉识别能力，快速发现存在严重缺失值问题的记录和属性，乃至该数据集质量问题的一般性特征和规律，便于为后继制定“筛选过滤”、“推断填充”等数据预处理策略和流程提供指导依据，为接下来的建模分析提供高质量的数据。

(2)用电模式发现

经过数据预处理后，课题组针对“用电模式”分析目标进行数据挖掘建模，以期能够发现电力用户在电力使用方面是否呈现出某些共性的特征和规律。用电模式的发现，将有助于电力运营企业精准把握不同时段和时期的用电需求和用电峰值，从而有助于企业更加有效地实施电力调配，并探索实践阶梯电价以均衡峰值期间的用电压力。

经过对海量用电数据的分析挖掘，课题组发现了三种典型的用电模式，其具体特征和规律如图 9 所示：展示了电力用户在一周(周日～周六，7 天×24 小时)的用电量随时间(小时)的变化情况。其中，从图 9(a)中可以很明显地发现，

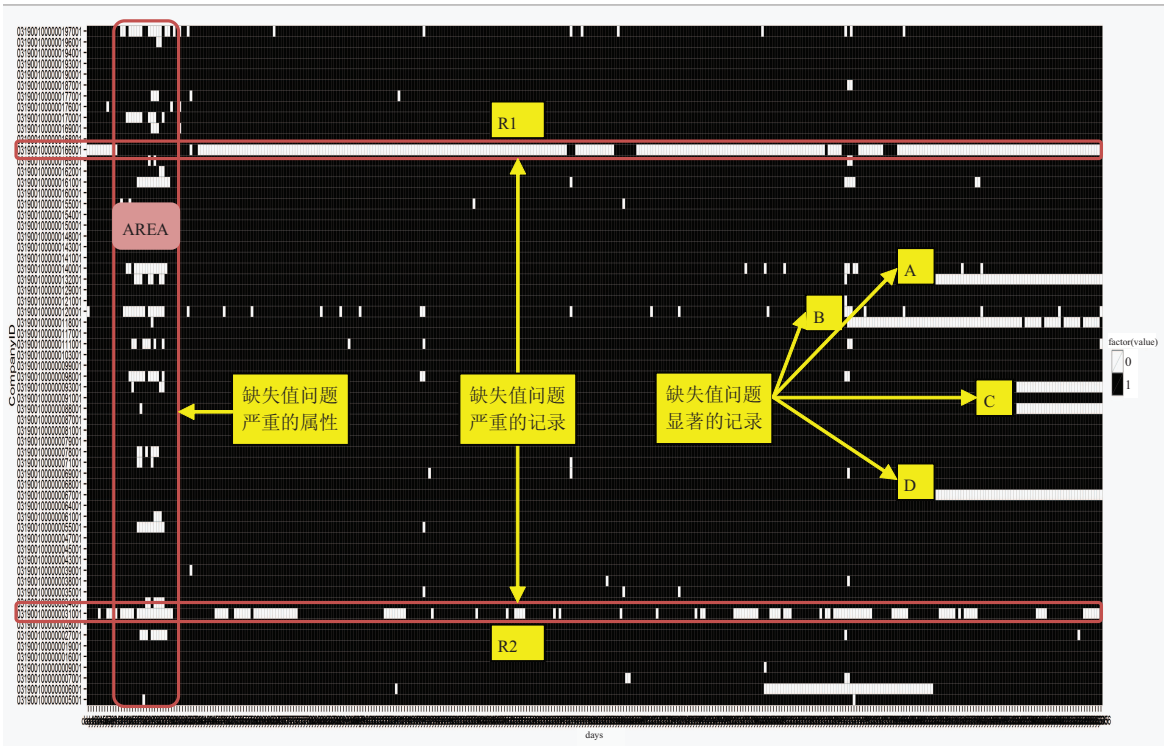

图 8　原始数据缺失值发现

Fig. 8　Missing value detction from raw data

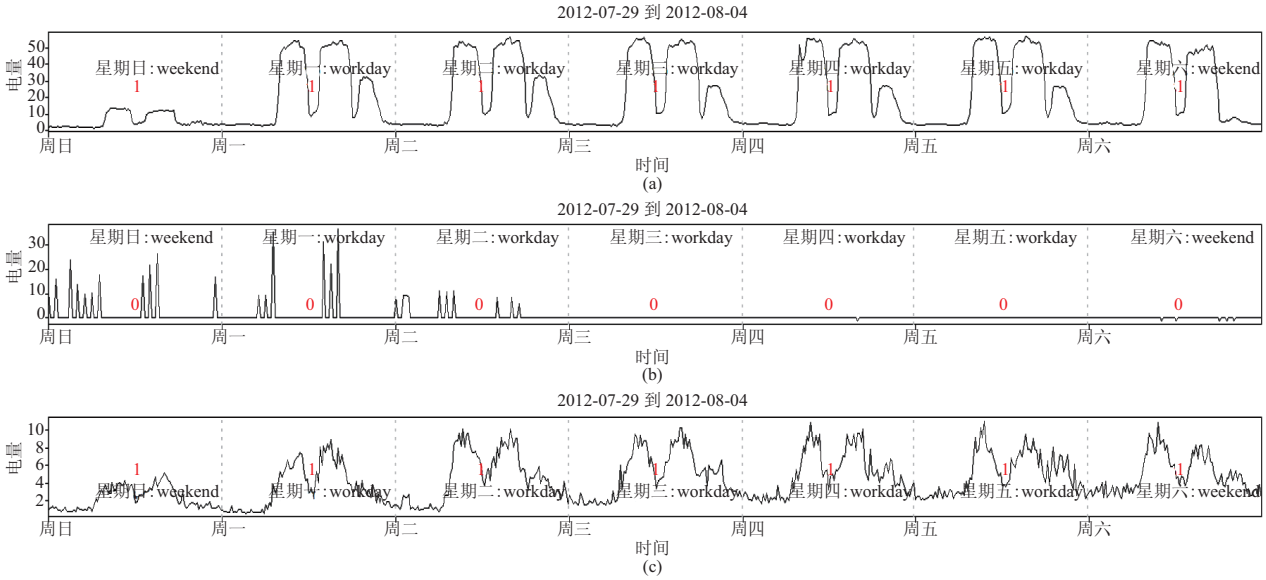

图 9 三种典型的用电模式

Fig. 9 Three typical power consumption patterns

该用电模式呈现出显著的时间特征和规律，即在周一～周六的 6 天时间中，出现三拨持续性的用电高峰，而在周日用电量不大。经过与代表性用户的实地交流验证，该用电模式正好符合企业周一～周六三班工作和周日休息的工作模式，证明所发现的用电模式符合实际应用情况。图 9(b)和图 9(c)所展示的用电模式的含义与图 9(a)相同，只是具体的模式特征存在差别。图 9(b)揭示了小型制造企业随机性用电的情况；图 9(c)揭示了企业周一～周六两班工作和周日休息的工作模式。

(3)基于用电模式的用户分群

基于上述用电模式分析的初步探索，课题组进一步研究“基于用电模式的用户分群”问题，以期待通过细化用电模式特征，找到不同模式下的用户群体。该分析结果有助于帮助电力经营企业精准把握所服务的客户的群体特征，精准预测不同时段和时期的用电需求，为制定和实施阶梯电价提供指导依据。

图 10 展现了 30 个具有典型用电模式的用户聚类(cluster)结果。从图中可以明显看出，每一个聚类所表现出的用电模式特征，以及不同模式之间的精细化差异。在此结果基础上，可进一步结合用户信息(如所属行业、所在区域等)，将可发现行业用电特征和区域用电特征等信息，有助于进一步提升电力企业对客户的服务能力和业务运营能力。

5 结论与展望

当前，大数据已经成为全社会的共识，大数据所蕴含的经济价值和创新价值已经引起社会各界的高度关注。我国拥有海量丰富的数据资源，广阔多样的应用场景，潜力巨大的消费市场，当务之急是如何快速有效突破数据价值挖掘的瓶颈。大数据分析与应用平台，是大数据时代必备的基础设施，也是突破当前技术瓶颈的有效突破口。开发和建设大数据分析与应用平台将带来三个方面的价值：第一，有助于不断汇集大数据技术创新成果，并用最先进的技术为用户提供一站式的应用服务；第二，有助于降低用户技术门槛，为应用开发提供共性基础设施与服务，从

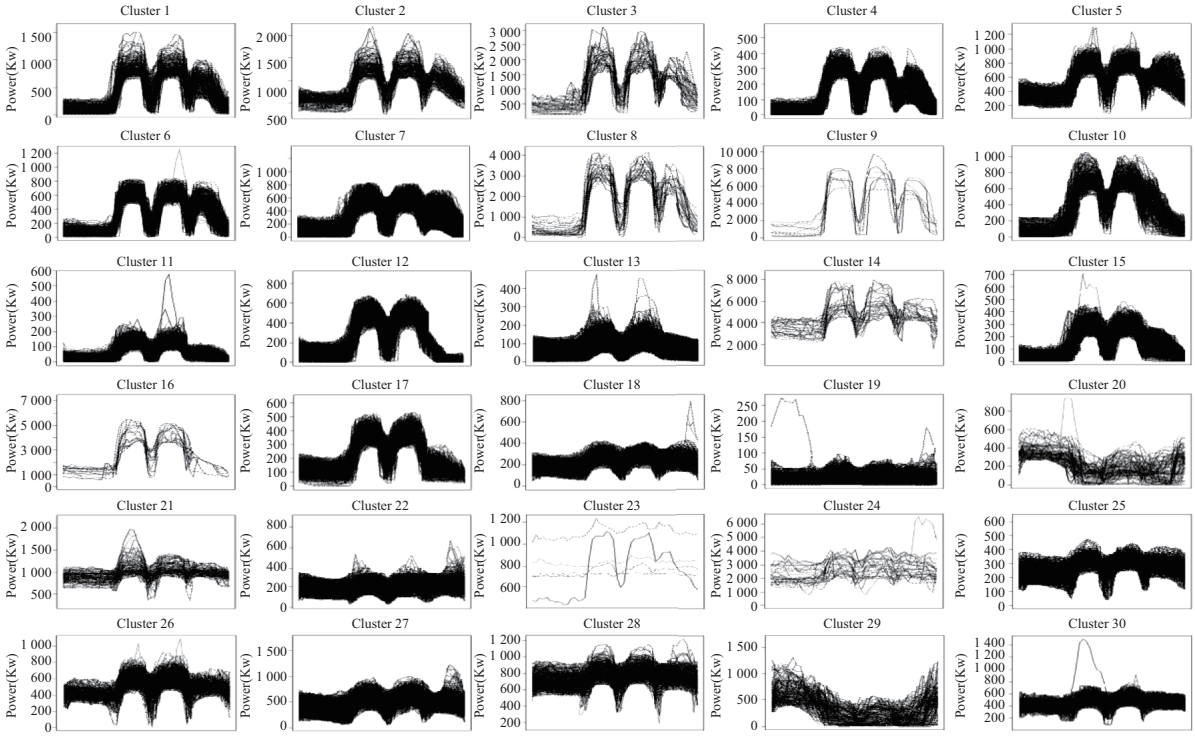

图 10　基于用电模式的用户分群

Fig. 10　User segementation based on power consumption pattern

而加快应用创新；第三，有助于形成大数据技术产品和行业解决方案，促进我国大数据产业加快形成。

参 考 文 献

[1] McKinsey Global Institute. Big data: the next frontier for innovation, competition, and productivity [DB/OL]. 2011-05[2015-12-24]. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

[2] 国务院. 促进大数据发展行动纲要 [EB/OL]. 2015-09-05[2015-12-24]. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.

[3] 黄哲学, 曹付元, 李俊杰, 等. 面向大数据的海云数据系统关键技术研究 [J]. 网络新媒体技术, 2012, 1(6): 20-26.

[4] 黄哲学, 陈小军, 李俊杰, 等. 面向服务的大数据分析平台解决方案 [J]. 科技促进发展, 2014, 10(1): 52-59.

[5] IBM. Big data analytics with IBM Cognos dynamic cubes [DB/OL]. 2012-12-07[2015-11-24]. http://www.redbooks.ibm.com/technotes/tips0942.pdf.

[6] Performing a data mining tool evaluation [DB/OL]. 2013-02-22[2015-12-24]. http://public.dhe.ibm.com/common/ssi/ecm/en/imw14300usen/IMW14300USEN.PDF.

[7] IBM. IBM Netezza analytics [EB/OL]. 2011-12-23[2014-12-24]. http://www-01.ibm.com/software/data/netezza/analytics/.

[8] IBM. What's new in IBM InfoSphere BigInsights V2.0 [EB/OL]. [2015-12-24]. http://www-01.ibm.com/software/data/infosphere/biginsights/whats_new.html.

[9] IBM. IBM PureData system for analytics N1001 [DB/OL]. 2014-12-16[2015-12-24]. http://www.smart-talk.nl/wp-content/uploads/IMD14400USEN.pdf.

[10] Oracle. Oracle exalogic elastic cloud [DB/OL]. [2015-12-24]. http://www.oracle.com/us/products/

middleware/exalogic/exalogic-elastic-cloud-x2-2-ds-1367805.pdf?ssSourceSiteId=ocomcn.

[11] Oracle. Oracle exadata database machine X2-8 [DB/OL]. [2015-12-24]. http://www.oracle.com/technetwork/server-storage/engineered-systems/exadata/dbmachine-x2-8-datasheet-173705.pdf?ssSourceSiteId=ocomcn.

[12] Oracle. Oracle exalytics in-memory machine: a brief introduction [DB/OL]. [2015-12-24]. http://www.oracle.com/us/solutions/ent-performance-bi/business-intelligence/exalytics-bi-machine/overview/exalytics-introduction-1372418.pdf.

[13] HP. HP vertica 6.1 boosts big data value [DB/OL]. [2015-12-24]. http://www.hp.com/hpinfo/newsroom/press_kits/2012/HPDiscoverFrankfurt2012/HP_Vertica_6.1_NA.pdf.

[14] EMC. EMC greenplum data computing appliance enhances EMC IT's global data warehouse [DB/OL]. [2015-12-24]. http://www.emc.com/collateral/software/white-papers/h8869-emc-greenplum-dca-oracle-gdw-wp.pdf.

[15] Martin P. Caffeine-the new google update [EB/OL]. [2015-12-24]. http://www.wearecube3.com/blog/article/caffeine-the-new-google-update/.

[16] Google. Google BigQuery-Real-time big data analytics in the cloud [DB/OL]. [2015-12-24]. https://cloud.google.com/files/BigQuery.pdf.

[17] Google. Google compute engine-computation in the cloud [DB/OL]. [2015-12-24]. https://cloud.google.com/files/GoogleComputeEngine.pdf.

[18] Amazon. [2015-12-24]. http://aws.amazon.com.

[19] Harris D. Facebook open sources Corona-a better way to do webscale Hadoop [EB/OL]. 2012-11-08[2015-12-24]. http://gigaom.com/2012/11/08/facebook-open-sources-corona-a-better-way-to-do-webscale-hadoop/.

[20] O'Dell J. Facebook's Project Prism is reimagining how big data scales [EB/OL]. 2012-08-22[2015-12-24]. http://venturebeat.com/2012/08/22/facebook-prism/.

[21] 淘宝. 淘宝指数 [EB/OL]. [2015-12-24]. http://shu.taobao.com/.

[22] Nature. Big data [J]. Nature, 2008, 455(7209): 1-136.

[23] Science. Dealing with data [J]. Science, 2011, 331(6018): 639-806.

[24] Petuum. Petuum 大数据分布式机器学习平台 [EB/OL]. 2015-02-11[2015-12-24]. http://medialab.sjtu.edu.cn/maas-blog/?p=605.

[25] 智能科学. 基于 Hadoop 的并行分布式数据挖掘平台 [EB/OL]. 2010-05-19[2015-12-24]. http://www.intsci.ac.cn/pdm/pdminer.html.

[26] 华东师范大学数据科学与工程研究院. CLAIMS: Cluster-Aware In-memory Sql query engine [EB/OL]. [2015-12-24]. http://dase.ecnu.edu.cn/index.php/system.

[27] Li M, Tan J, Wang YD, et al. SparkBench: a comprehensive benchmarking suite for in memory data analytic platform spark [C] // Proceedings of the 12th ACM International Conference on Computing Frontiers, 2015: 53.

[28] Zhu B, Mara A, Mozo A. CLUS: Parallel subspace clustering algorithm on spark [M] // New Trends in Databases and Information Systems, 2015: 175-185.

[29] Armbrust M, Das T, Davidson A, et al. Scaling spark in the real world: performance and usability [J]. Proceedings of the VLDB Endowment, 2015, 8(12): 1840-1843.

[30] Shi JW, Qiu YJ, Minhas UF, et al. Clash of the titans: MapReduce vs. Spark for large scale data analytics [J]. Proceedings of the VLDB Endowment, 2015, 8(13): 2110-2121.

[31] Agrawal R, Srikant R. Fast algorithms for mining association rules [C] // Proceedings of the 20th Very Large Data Bases, 1994: 487-499.

[32] 刘义德, 梁坚. 智能电网大数据处理技术现状与挑战 [J]. 科技创新与应用, 2015(29): 184.

[33] Song Y, Zhou G, Zhu Y. Present status and challenges of big data processing in smart grid [J]. Power System Technology, 2013, 37(4): 927-935.