DOI:10.3969/j.issn.1009-6868.2016.03.013
网络出版地址:http://www.cnki.net/kcms/detail/34.1228.TN.20160304.1338.002.html

大数据技术发展趋势及灯塔大数据行业应用平台

Big Data Technology Development Trend and DENGTA Application Platform

中图分类号:TP393 文献标志码:A 文章编号:1009-6868 (2016) 03-0057-005

摘要: 指出大数据发展的趋势:混合数据存储是大数据技术的基础;融合数据库架构是大数据发展的趋势;异构数据关联是大数据平台的关键;行业知识库是产业互联网应用发展的要素;深度标签是大数据挖掘的核心技术之一。介绍了中国电信灯塔大数据行业应用平台的架构,及所采用的关键技术和行业应用,认为该平台的使用可以充分发挥运营商数据与外部数据相结合的优势,加速产业升级和商业模式创新。

关键词: 大数据;趋势;灯塔;应用

Abstract: In this paper, trends in big data technology are discussed. Mixed data storage is the foundation of big data technology; database schema integration is the trend of the development of big data; heterogeneous data association is key to big data platform; industry knowledge database is the key elements of the application and development of the Internet industry; depth labels is one of the core technologies of data mining. Then, the Dengta big data industry application platform of China Telecom is introduced. This platform can be fully combined with operator data and external data in order to accelerate industrial upgrading and innovation of business model.

Key words: big data; trend; Dengta; application

王若倪/WANG Ruoni
赵慧玲/ZHAO Huiling
(中国电信股份有限公司北京研究院,北京 100035)
(China Telecom Beijing Research Institute, Beijing 10035, China)

大数据是信息时代技术创新的产物,大数据与云计算、物联网等新技术相结合,正日益深刻地改变着人们的生产生活方式。大数据产业的出现和发展是现代信息技术与互联网时代海量信息发展到一定阶段的必然结果,必将对当今社会的信息技术、商业模式和相关的法律法规产生深刻影响。大数据经历了基础理论研究和产业应用探索,与行业应用结合已成为大数据发展的新机遇。

1 大数据技术发展趋势

(1)混合数据存储是大数据技术的基础

在大数据环境下,数据量达到了PB级甚至EB级。大数据存储一方面需要提供超大容量的存储空间,另一方面需要支持对海量数据的智能检索和分析。为了兼容各种类型的大数据应用,大数据存储需要提供混合的数据存储模型,支持文件、对象、键值、块等多种访问接口,作为大数据技术的基础[1-2]。

(2)融合数据库架构是大数据发展的趋势

随着大数据业务的发展,除了面向强关系型的结构化查询语言(SQL)数据库之外,面向各类应用的接口灵活、功能丰富且高效的NoSQL数据库也得到了蓬勃发展。在应用类型多样、数据种类繁多的大数据平台中,融合关系型数据库、列数据库、内存数据库、图数据库等多种数据库的混合数据库架构,能够满足多种场景下的数据处理需求,是大数据发展的必然趋势。

(3)异构数据关联是大数据平台的关键

当前,各行业、企业、系统、平台都累积了海量的数据,这些数据结构不同且相对独立,在没有建立起关联关系的情况下,难以展现出数据的优势。将这些多源异构数据进行关联和融合,挖掘数据之间的相关性,能够为数据分析奠定坚实的基础,最大限度地发挥数据价值,是大数据平台的关键所在。

(4)行业知识库是产业互联网发展的要素

随着“互联网+”战略的实施,各产业尤其是传统产业,纷纷进行互联网化转型。在“互联网+”的浪潮下,面向多个行业,深挖行业知识详情,构建行业知识库,形成完整的行业知识体系,能有效推动数据应用与价值

收稿时间:2016-02-14
网络出版时间:2016-03-04

落地，是产业互联网发展的关键。

（5）深度标签是大数据挖掘的核心技术之一

数据挖掘越来越多地应用到各个行业应用领域，使用数据挖掘技术而打造用户深度标签，已经逐渐成为大数据挖掘的热点。通过针对大数据场景的数据挖掘，深入分析用户行为，打造多层次、多角度的用户深度标签。深度标签是大数据挖掘的核心技术之一，它使得大数据应用更加精准，业务能够更加贴近用户，更好地满足用户的需求[3]。

2 灯塔大数据行业应用平台总体架构

在大数据的发展浪潮下，中国电信股份有限公司北京研究院通过大数据技术创新，自主研发了业内领先的灯塔大数据行业应用平台。灯塔大数据行业应用平台深入研究大数据平台技术和应用技术，为满足顶层大数据应用需求，自主开发大数据能力，实现电信数据与外部数据相融合的大数据分析挖掘，打造了ID关联模型、用户深度标签、行业知识库、分布式爬虫、数据可视化等平台即服务（PaaS）层能力，并以标准化应用程序编程接口（API）的形式支持顶层数据的相关应用，打造了市场研究、泛义征信、地理洞察等三大领域的6款大数据应用。

灯塔大数据行业应用平台技术架构如图1所示，其底层平台基于开源技术搭建，融合了离线批处理、内存计算、流计算等多种计算模型，以及关系型数据库、列数据库、内存数据库、图数据库等多种数据库模型，向上提供计算和存储能力；在大数据开放能力层，研发了ID图谱、用户标签等多种大数据分析挖掘技术，并结合第三方的地理信息系统（GIS）等能力，面向多个行业领域，向应用层以API的形式提供多种数据服务。

3 灯塔大数据行业应用平台关键技术

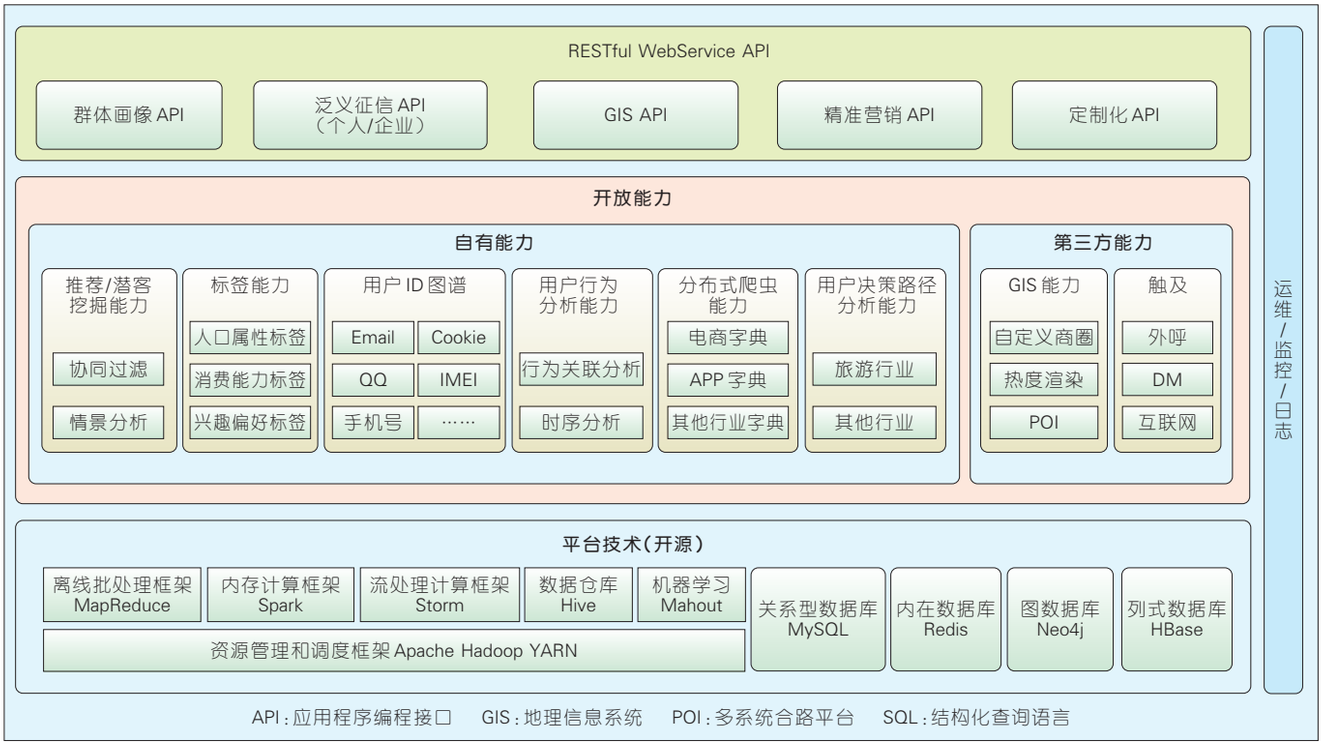
灯塔大数据行业应用平台主要研发了混合数据库、ID关联模型、用户深度标签、行业知识库、统一数据采集与存储等几项关键技术。

3.1 混合数据库

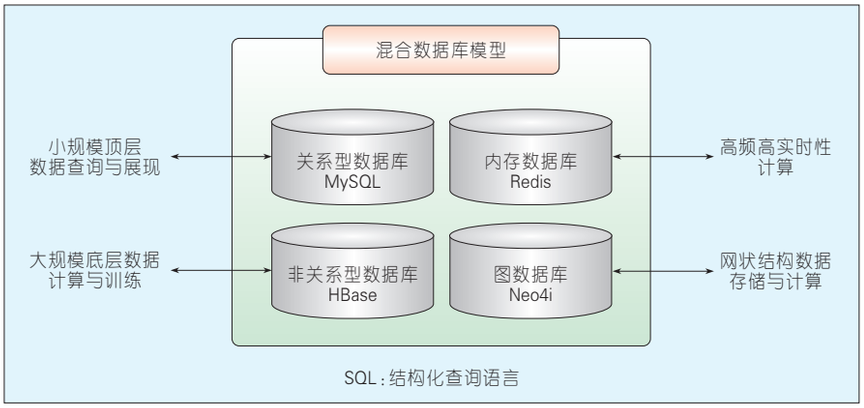
融合关系型数据库、列数据库、内存数据库、图数据库，并提出面向不同存储过程和计算需求的混合数据库模型，可以满足多种场景下的数据处理需求，解决单一数据库模型无法满足大规模数据训练、高频高实时性计算、网状结构计算等不同场景下的数据处理问题。

如图2所示，海量数据计算使用非关系型数据库（NoSQL）来支持；网状结构数据的机器学习训练依靠图数据库（Neo4j）来支持；高频高实时性计算对接内存数据库（Redis）；小规模顶层数据查询与展现对接关系型数据库（SQL）。具体来说，包含4点内容：

（1）能够实现有一定实时性需求的、传统千万级及以下的数据查询与展现业务，并基于传统关系型数据库



▲图1 灯塔大数据行业应用平台



▲图2 混合数据库模型

MySQL来构建。通过加载数据预读取算法，MySQL的单机处理能力可以达到秒级访问5 000万条多维数据的水平，能够满足一般的数据查询业务需求。

（2）对于千万级以上的数据查询业务，已超出单台MySQL的支持水平，更适宜转化成离线查询业务，直接使用非关系型数据库HBase来支持。此时数据查询的范围可扩展至数十亿甚至上百亿，系统仍可平稳输出查询结果，前提是付出分布式离线计算的延时代价。

（3）对于在深度包检测技术（DPI）数据的K-V查询过程中需同步完成标签数据在灯塔本地服务器的ETL工作的场景，任何传统磁盘输入输出（IO）基本都无法支持该高频数据存取操作，则借助内存数据库Redis来完成。Redis可在典型的单台计算资源下支持100毫秒级的数据ETL操作，并且可以与K-V查询进行无缝衔接，轻松应对每日2亿条标签数据入库。

（4）对于图状数据结构，如灯塔平台中典型的ID知识体系，则适合从边和节点的角度进行数据存储、表达和计算，无论行数据库还是列数据库都不再适合，因此采用图数据库Neo4j来支持。

目前，灯塔大数据行业应用平台支持1 000万级多维数据的秒级查询展现，10亿级多维数据的24 h内基础演算，100毫秒级的数据流处理，并可秒级完成10亿级边、1 000万级节点的子图查询运算。

3.2 ID关联模型

基于图计算技术构建ID关联模型，采用图数据库进行数据存储和模型计算，实现DPI数据内的多种用户ID关联，解决了电信数据与外部数据有效关联和拼接的问题。ID关联模型建立设备标识—场景的图模型，通过图数据库、图计算得到隐性变量用户唯一标识，打通用户各个设备，实现全面的用户画像。

ID关联模型对内实现数据融合，将DPI数据内的多种用户ID关联，实现多场景、多屏幕信息打通，从而实现更全面和精准的用户描述；对外实现数据开放，借助从DPI中挖掘出的外部ID，实现运营商数据与外部数据的打通，从而打破了电信数据开放的壁垒。

目前，灯塔大数据行业应用平台已积累超过100类ID数据，ID总量超5 000万。

3.3 用户深度标签

根据用户上网行为、使用机器学习和模式识别等算法，如树状增强型朴素贝叶斯（TAN）分类算法等，推断用户的性别、年龄等基础人口属性，并打造消费偏好、消费能力等其他深度标签，用于支持用户行为分析的大数据应用。

目前，灯塔大数据行业应用平台已构建超过10个行业的总计6 000余类用户深度标签。

3.4 行业知识库

通过整合数据采集、数据存储、数据形式化、数据表达等环节，打造完善的行业知识库，为运营商网络大数据的解析提供必要的支持。其中，行业知识库的构建包含以下环节：

（1）基于分布式爬虫进行数据采集。如图3所示，分布式爬虫DTSpider基于开源技术WebMagic与内存数据库技术Redis而研发，搭建在云主机上，提供行业知识库数据采集解决方案。

（2）面向垂直行业构建知识体系。如图4所示，行业知识库面向如电商、新闻、影视等不同的垂直行业，分别构建树状知识体系，能够直接对接标签能力应用。例如，电商行业的树状知识体系，可按照商品类别进行构建，如图书、服饰、运动健康等。

（3）深挖垂直行业知识详情。基于从页面抓取的标题和正文，经自然语言处理得到知识详情，例如电商库存量单位（SKU）名称、价格、参数、评论等。

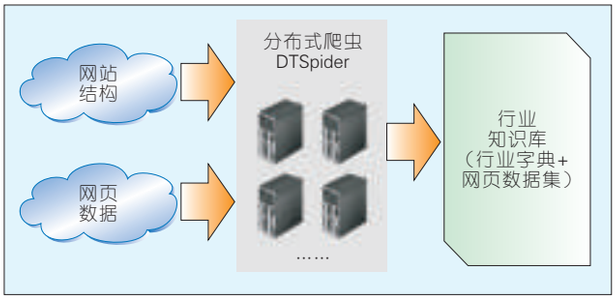
目前，灯塔大数据行业应用平台的行业知识库整体字典规模超过2亿，其中电商和视频分别占1.2亿和6 000万。

3.5 统一数据采集与存储

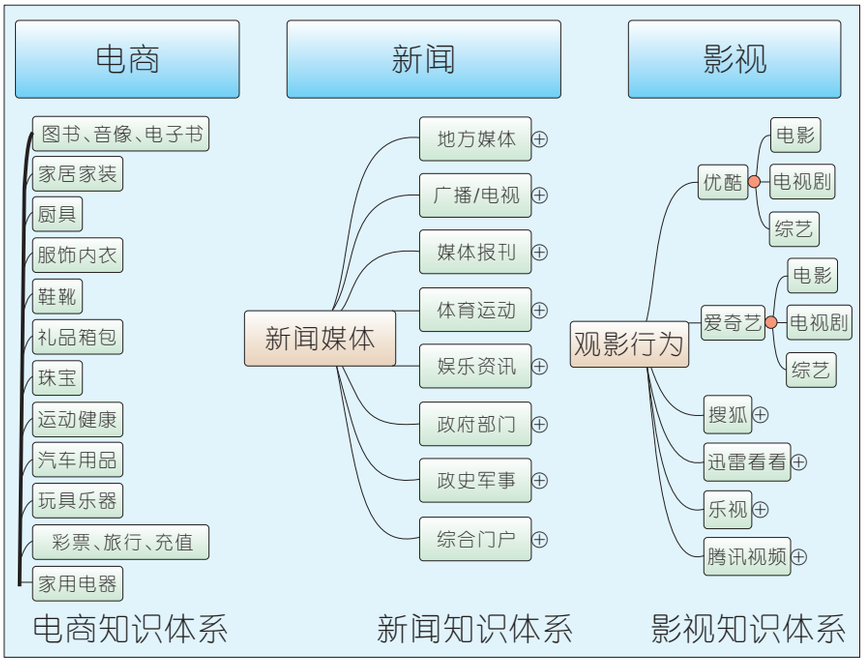
面向电信管道数据、互联网公开数据和企业自有数据等多种数据类型，分别构建数据采集能力，并定义了统一的数据采集接口与存储接口，解决了多源异构数据的采集与存储的相关问题。

（1）电信网络大数据采集

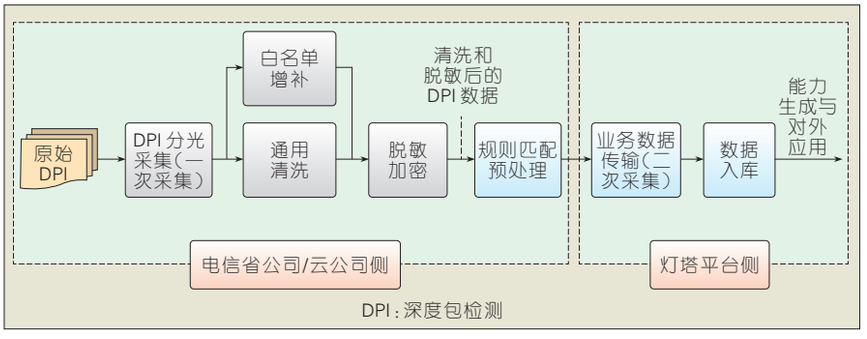
电信网络大数据采集包含以下环节：DPI分光采集、数据清洗、数据脱敏、规则匹配预处理、业务数据传输、数据入库等环节，如图5所示。



◀图3
分布式爬虫DTSpider



▲图4 行业知识体系



▲图5 电信网络大数据采集流程

在基层分局进行一次采集与清洗，在业务平台进行二次采集与存储。具体来说，在基层分局分光设备采集（一次采集）得到固网/移动DPI数据，然后采用通用清洗规则与白名单规则相结合的方式，过滤掉DPI数据中无效及重复请求，并保证各类业务的数据需求。经过规则匹配预处理，从DPI中抽离并编码得到业务所需的数据，以标签形式传输（二次采集）并入库至业务平台，提供给PaaS层的生成数据能力，最终对接软件即服务（SaaS）层的数据应用。

根据生产平台数据接口差异以及顶层业务类型差异，电信网络大数据的二次采集可采用实时或离线模式。如图6所示，实时流处理模式是通过K-V查询接口，以流处理模式，逐条传输、ETL、融合并入库至业务平台。离线批处理模式是通过安全文件传送协议（SFTP）传输接口，将数据离线批量采集至业务平台缓存中，再进行批量抽取、加载、转换（ETL）、融合并入库至业务平台。

（2）互联网大数据采集

互联网大数据采集通过分布式爬虫DTSpider进行。DTSpider支持节点动态接入，有效提升爬取效率，避免IP封锁，具有良好的稳定性和可扩展性。

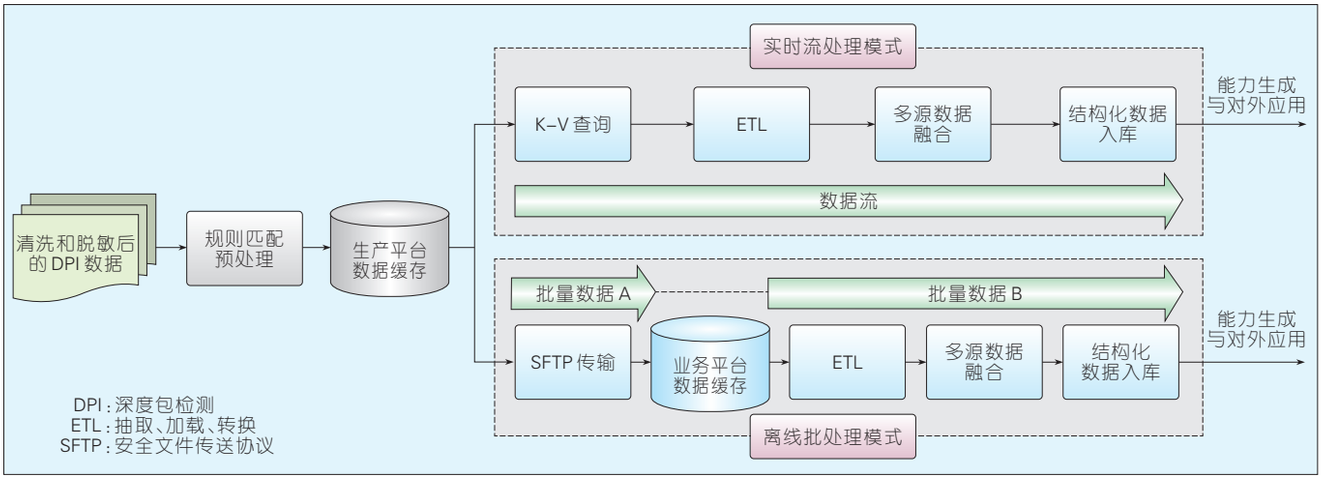
（3）企业自有数据接入

基于ID图谱，可对企业的客户管理系统（CRM）数据进行导入与融合。订单及其他业务数据，也可导入并可对接灯塔平台主体数据，支持大数据分析。

目前，灯塔大数据行业应用平台已采集并汇聚电信数据600多亿条，外部数据5亿条。

4 灯塔大数据的行业应用

在混合数据库、ID关联模型、用户深度标签、行业知识库、统一数据采集与存储等几项关键技术的支持之上，灯塔大数据行业应用平台打造了市场研究、泛义征信、地理洞察等三大领域的6款大数据应用，实现了数据产品及服务的规范化、流程化，探索出大数据价值落地的商业模式。其中，在市场研究领域，基于灯塔平台行业知识库、深度标签等数据能力，我们研发了零售研究、消费者研究、决策路径分析等方面的大数据应用；在泛义征信领域，基于灯塔平台ID图谱、深度标签等数据能力，我们研发了用户画像等技术，应用于人力资源、企业征信等场合；在地理洞察领域，基于灯塔平台ID图谱、深度



▲图6 电信网络大数据二次采集模式

标签等数据能力，结合第三方GIS能力，我们打造了人群流量监测、迁徙分析、店铺选址等应用。

（1）灯塔在线零研

灯塔在线零研基于电信管道数据，打造电商分析能力，提供在线零售研究业务，数据更新频率最快可达T+1，支持联机分析处理（OLAP）查询，分析维度多达20个。

（2）灯塔消费者洞察

与合作伙伴共同研发的灯塔消费者洞察应用，可以实现电商内容监测、论坛内容监测、用户多维画像等功能，支持基础人口属性和互联网行为画像。

（3）灯塔大数据招聘

与在线人力资源行业相结合，提供求职人员的个人画像新型简历，包括量化的行为偏好、性格特征、个人优势数据，覆盖消费能力、学习指数、作息指数、勤奋程度、运动指数等多种维度，从而基于用户的互联网行为为企业提供客观的招聘参考

（4）灯塔背景调查

将网络行为报告与第三方个人数据相结合，研发并上线新型在线背景调查产品，打造更加高效、完善的背景调查体系。

（5）灯塔在线人口普查

灯塔在线人口普查基于地理位置及互联网行为数据，为客户提供基础人口普查、人口迁徙分析和互联网偏好分析等服务。

（6）灯塔慧选址

灯塔慧选址结合灯塔标签数据和线下位置数据，能够为客户提供在线选址、运营分析等功能。

除了以上6种应用之外，灯塔大数据行业应用平台还紧跟市场趋势及热点事件，产出10多份高质量数据分析报告，例如“2015年一季度奶粉市场研究报告”、“2015抗战胜利日大阅兵互联网分析”、“2015双十一未消费人群报告”等，并通过移动互联网进行传播，覆盖近万互联网受众，吸引了大数据行业关注。

5 结束语

作为快速发展的新兴产业，大数据已经上升到国家战略层面，成为整个社会最有价值的资产。大数据已经渗透到各个行业领域，其行业应用具有广阔的发展空间。

灯塔大数据行业应用平台立足自主研发，深入研究大数据底层平台能力及数据分析挖掘能力，充分发挥运营商数据与外部数据相结合的优势，加速产业升级和商业模式创新。灯塔大数据旨在充分发挥数据价值，通过技术创新和应用创新共同驱动，与行业合作伙伴共同打造大数据行业应用生态圈。未来，灯塔大数据平台将面向房地产、汽车、金融等行业领域打造更多的行业应用产品并提供服务。

参考文献

[1] 赵慧玲, 杨明川, 孙静博. 大数据技术发展及其应用[J]. 中国电信建设, 2015, 27(4):36–38
[2] 张引, 陈敏, 廖小飞. 大数据应用的现状与展现[J]. 计算机研究与发展, 2013(S2): 216–233
[3] ZHAO H L, XIE Y P, SHI F. Network Function Virtualization Technology: Progress and Standardization [J]. ZTE Communications, 2014, 12(2): 03–07. DOI: 10.3969/j.issn.1673–5188.2014.02.001

作者简介

王若倪，中国电信股份有限公司北京研究院工程师；主要研究领域为大数据技术和应用。

赵慧玲，中国电信股份有限公司北京研究院总工，教授级高工，中国通信学会常务理事，信息通信网络技术专业委员会主任委员，中国通信学会北京通信学会副理事长，中国通信标准协会网络与交换技术工作委员会主席，中国SDN/NFV产业联盟技术委员会副主任；主要从事宽带网络和下一代网络的技术研究以及通信网络发展战略研究等工作。