交通大数据：一种基于微服务的敏捷处理架构设计

杜圣东，杨燕，滕飞
西南交通大学信息科学与技术学院，四川 成都 610031

摘要
面对智慧交通广泛的大数据应用场景和技术需求，一般大数据系统难以适应多种处理情况并做出快速响应。针对这一问题，首次提出了敏捷大数据方法论，对其概念、处理流程、核心原则与关键技术等进行了研究和探索。基于数据科学迭代性本质，设计了面向微服务的敏捷大数据架构，对交通大数据微服务化、交通大数据融合等关键环节进行了详细设计和论述。敏捷大数据架构的提出为交通大数据环境下的高效、灵活数据挖掘和机器学习提供了新思路、新方法。

关键词
敏捷大数据；微服务；大数据架构；容器；大数据融合

中图分类号：TP274 文献标识码：A doi: 10.11959/j.issn.2096-0271.2017030

Traffic big data: an agile architecture design based on micro service

DU Shengdong, YANG Yan, TENG Fei
School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China

Abstract
Faced with a wide range of intelligent transportation application scenarios and technical requirements, the general big data system is difficult to adapt to a variety of processing and make a quick response. In order to solve this problem, the methodology of agile big data for the first time was put forward. Based on the iterative nature of data science, the agile big data architecture based on micro service was designed, and the key points of the traffic oriented micro service and data fusion technology were discussed. The agile big data architecture provides new idea and method for efficient and flexible data mining and machine learning under the environment of traffic big data.

Key words
agile big data, micro service, big data architecture, container, big data fusion

1 引言

近年来，随着大数据、云计算、人工智能、物联网等前沿信息技术的高速发展，人们已然迈入大数据时代[1]。数据被喻为大数据时代的石油，如何拥有继陆权、海权、空权之后的数权，已经上升为各国的重大科技战略。随着我国高铁、地铁、轻轨和高速公路等交通基础设施建设和城市化进程的日益提速，交通拥堵、排放污染、能源消耗等问题日益严重，交通事故、交通紧急事件频繁发生，交通管控风险加大，这些都是智慧交通和城市计算[2]亟待解决的关键问题。由于现代交通涉及经济、环境、社会和安全等多个方面，大数据时代的到来为综合交通一体化、智能化管理提供了新的机遇和广阔的应用前景，同时也对交通数据感知、交通大数据管理和时空大数据机器学习等方面的关键技术和架构设计提出了新的挑战。当前的交通规划设计、运营管控、安全维护等受到跨地域、跨行业等限制，各类管理主体的分割造成了交通数据的条块分割和碎片化，使得传统信息技术难以支撑交通大数据的综合管理和挖掘分析。所以通过不断获取、整合和分析交通大数据解决上述挑战，通过大数据机器学习和城市计算技术提升智慧交通管理水平，应用潜力巨大；另外，交通海量数据的多源异构、时空关联、多模态、动态性等特点，大数据的集成性、可预测性等技术特征，更是凸显了大数据技术在智慧交通中的深层应用价值。

随着交通大数据的急速增长，各类数据在时空关联、跨模态和动态处理方面都呈现出不断增长的复杂性，一般信息技术和传统数据挖掘技术越来越难以处理。如何有效支持交通大数据敏捷处理，是交通大数据应用要解决的关键问题。近年来，以微服务、构件化和数据融合等为代表的大数据架构和关键技术研究备受关注。针对软件架构的复用性，杨芙清院士等人[3]很早就开创性地提出了构件化软件设计理论与方法，对软件复用技术作了全面综述及关键技术分析，为近年来的服务化和容器化技术打下了坚实的基础；王珊等人[4]对大数据架构设计的挑战和现状做了分析和展望，为了设计适合大数据分析的数据仓库架构，探讨了大数据平台需要具备的重要特性，并对当前的主流大数据框架进行了深入研究，指出了各自的优势及不足。孟小峰等人[5]阐述了大数据处理的基本框架，并就云计算技术对大数据时代数据管理产生的作用进行分析。Chen H M等人[6]提出了以架构为中心的敏捷大数据开发方法，初步引入了传统软件工程中的敏捷开发思想。在交通大数据处理方面，郑宇[2]提出了城市计算的定义、框架和主要研究问题，并概述了城市计算的典型应用和所需的技术。在微服务方面[7]，亚马逊、谷歌、FaceBook、阿里巴巴等企业进行了大量研究和应用，通过微服务技术将大型、复杂的应用程序构建为一组相对独立和协同联系的微小服务，以提高大型复杂系统的稳定性、敏捷性和研发维护效率。从上述分析可以看出，虽然已有针对大数据架构的探讨和主流大数据框架比较研究，但在大数据架构设计的方法论和关键技术支撑体系，特别是结合微服务、敏捷方法论的大数据架构研究方面，暂时还是空白。

本文探讨了交通大数据应用面临的关键问题和技术挑战，阐述了大数据技术在智慧交通中的关键地位和相互之间的关联关系。从4个层面系统分析了交通大数据处理涉及的共性关键技术。由于交通大数据涉及技术面十分广泛，大数据计算模式和处理数据模态的不同，决定了其采取的

技术也会不同，面对交通大数据分析广泛的应用场景和特定技术需求，一般大数据分析系统难以适应所有的处理情况，其他领域的大数据应用也面临着同样的难题。因此，提出了敏捷大数据的方法论，并对其概念、原理、处理原则与关键支撑技术进行了深入研究和分析。并基于大数据多粒度信息融合、计算微服务化、云计算和容器等技术，从采集、存储、计算和应用4个层面设计了面向智慧交通的敏捷大数据处理架构，为交通大数据应用研究提供了新的思路、方法和技术。

2 交通大数据应用挑战

面对海量的多源异构交通大数据，如何从中根据用户需求提取有效特征，并挖掘出有价值的信息进行决策支持，是智慧交通要解决的关键问题。通过对交通大数据的分析处理和预测预警，可以辅助交通管理者制定出更好的解决方案，提升交通监控、预测和安全保障水平，并提高交通网络的运行效率，这是智慧交通的核心目标。大数据技术和智慧交通是互为支撑的关系（如**图1**所示），交通大数据系统在智慧交通应用的中心地位凸显。首先，为了解决数据孤岛问题，大数据分析和决策需要基于统一的交通大数据中心，交通环境中的人、路、车和相关时空感知数据，通过采集、抽取、汇聚进入大数据中心进行分布式存储，各类管理信息系统也需要数据集成，为大数据挖掘分析提供数据源；其次，大数据分析技术除了提供独立的智慧交通决策支持服务之外，还要能对所有相关交通管理信息系统和对外信息发布方面提供数据分析服务，业务系统数据库为大数据系统提供数据来源，大数据系统为业务系统提供海量数据处理架构下的汇总统计、探索分析、多维分析和决策指导等大数据分析支持。**图1**是交通大数据在智慧交通应用中的定位和相互关系示意。

交通大数据分析作为智慧交通的关键支撑技术，其应用面临着巨大挑战。大数据技术的关键词是预测、迭代和优化，技术模型和系统设计不同于传统信息化领域，特别是大数据很多技术采用的是全新技术范型，核心技术的突破短期仍受限于理论、技术成熟度和研发能力，具有复杂性和不确定性。另外，根据交通大数据的特点，需对数据采集、多模态数据融合、数据质量监控、隐私信息清洗和机器学习模型的有效性、稳定性等方面，进行持续深入研究和验证。一般来讲，交通大数据应

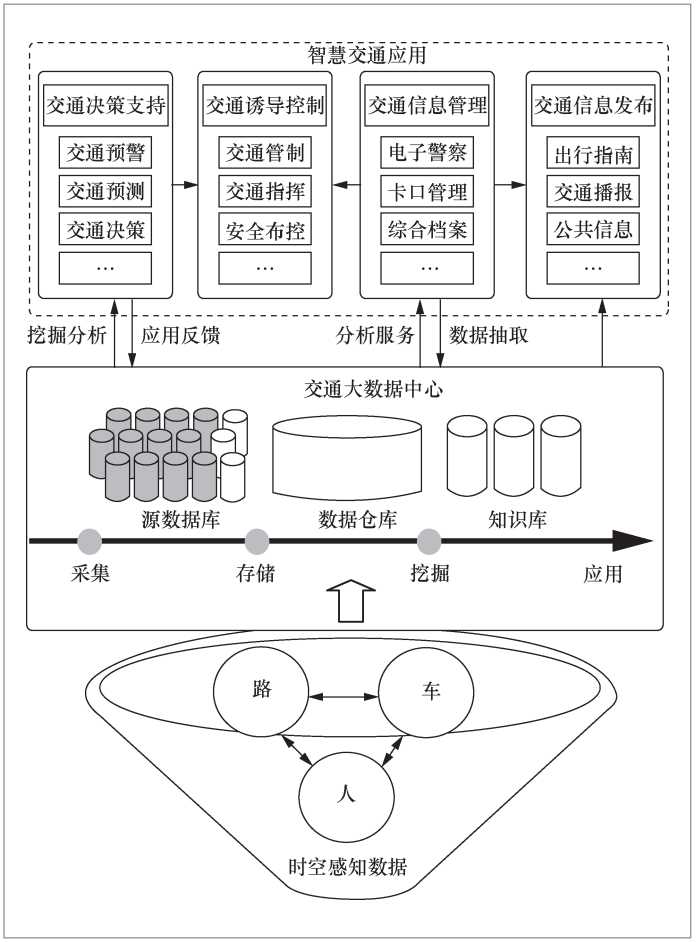


图1 交通大数据在智慧交通应用中的定位与相互关系示意

用主要面临如下挑战。

（1）数据孤岛和集成共享问题

全面、及时、准确地采集和获取交通各类源数据是交通大数据分析的基础。传统交通规划设计、运营管控的跨地域、跨行业等限制问题，造成了各类管理主体的分割，从而导致了各类交通信息的条块分割和碎片化，交通数据碎片化地分散在不同的信息系统中，数据孤岛问题严重。而交通大数据的集成性、可预测性必须基于海量数据的集成和融合，才能实现智能化、一体化的智慧交通管理，所以解决数据孤岛和实现数据集成共享是交通大数据应用面临的首要挑战。

（2）多源异构和时空关联问题

交通大数据横跨多个行业领域，结构化和半结构化、非结构化数据混合并存，不同系统或平台的数据呈离散化、碎片化分布，面对交通大数据多源异构、时空关联复杂耦合和相互影响等特点，传统数据挖掘与机器学习技术难以直接应用，这给交通大数据的处理带来了巨大挑战。如何针对交通大数据的时空关联性、互补性及其动态变化规律进行深入解析，并在此基础上构建合理有效的机器学习模型，是交通大数据关联分析要解决的关键问题。

（3）多模态数据和多任务处理问题

交通大数据随着时间和空间的动态变化常常呈现出多模态和动态性特点，由于多模态交通大数据处理固有的复杂性、相关性、动态性等问题，需分析从单模态数据机器学习到多模态关联机器学习的高阶相关性原理，并且研究交通大数据中不同层次之间信息的关联集成与融合机理。结合多模态数据源的互补信息，分析数据源之间的各类特征关联关系和语义关系，构建面向多源异构数据的多粒度信息融合模型，建立统一数据融合层和标准数据集来处理和共享交通大数据，并设计基于多模态交通大数据处理的跨模态机器学习和多任务集成学习方法，是交通大数据建模要解决的关键问题。

3 交通大数据关键技术

要实现交通大数据分析目标，需明确所需数据源和关键技术，针对分析目标，进行交通多个维度数据（人、车、路、时间、空间等）的采集和存储。通过交互式分析技术、可视化分析技术和数据挖掘、机器学习或深度学习技术，对海量交通数据进行大规模建模计算，最后对分析结果进行展示评价，并在实际系统中进行应用评估和反馈、优化，从而科学指导和辅助交通管理决策支持。由于大数据技术生态体系庞杂[8]，基础技术覆盖数据采集、数据预处理、数据脱敏、分布式计算、NoSQL数据存储、多模式（离线、实时、流、内存）计算、多模态（图片、视频、音频、网页、传感器）数据处理、并行计算、数据仓库、数据挖掘、机器学习、深度学习、数据可视化等各个层级，应用场景更是涉及交通流分析、交通拥堵分析、污染排放分析、交通风险分析、交通诱导分析、交通轨迹分析等各个方面。要支撑基于大数据技术的智慧交通管理，需对交通大数据的基础性共性关键技术进行深入研究，包括应用环节、方法原理和关键技术等方面。本文从4个层面对交通大数据分析的共性关键技术进行研究。

（1）交通大数据采集与集成融合技术

随着城市交通、铁路、航空、水运等各领域物联网、车联网传感采集和智能终端设备的普及应用，交通数据采集的范围、广度和深度进一步加强，为实现跨行业与跨地域的智慧交通大数据分析提供了海量

的全样本数据。类型繁多、结构多样、体量巨大的交通数据决定了其采集的复杂性，需研究大数据条件下的跨行业、跨地域的交通动、静态数据采集，数据交换共享以及预处理关键技术，主要支撑对静态、动态两大类交通数据的采集、抽取和融合集成，包括公路、铁路、航空、水运等多行业的设施（如路网）、设备（如车辆）、人员、安全保障等基础数据，还有综合交通网络中人、车、路、环境等相互间作用的关系数据，如速度、密度、流量、位置、票价等。需要研究移动式动态交通数据采集技术、固定式交通数据采集技术、基于Hadoop的关系数据抽取技术、大规模交通数据预处理技术等，为智慧交通的时空感知和综合管控提供数据支撑。

（2）交通大数据云存储与资源管理调度技术

面对现代综合交通的信息化、物联网化发展，其数据呈现几何增长的趋势，数据库技术已经不能仅依靠单台或几台机器进行纵向扩展（scale up）的升级来满足存储需求。必须采用基于大数据架构的分布式云存储技术，基于Hadoop的分布式文件系统（Hadoop distributed file system，HDFS），高可用数据存储，结合HBase、Redis等面向列的数据存储模型，解决海量数据存储扩展的问题。需针对具有时空特性的地理空间数据，如全球定位、卫星图像、地理轨迹等多种形式的交通大数据，分析各数据实体之间的物理和逻辑关系，设计相关分布式存储索引。另外，云存储作为交通海量数据处理的基础设施，面临节点庞杂、负载不均衡、数据实时性响应等问题，需对云集群的资源调度、集群性能、负载均衡等关键技术进行深入研究。

（3）交通大数据挖掘建模与机器学习技术

针对交通大数据多源、异构、海量等特征，传统的数据挖掘、处理、分析方式已经无法满足大数据的要求。以数据计算效率和计算复杂性理论为基础，研究基于云计算和分布式集群的大数据多模式计算（包括批处理、实时流处理、内存处理等）和多模态计算（包括图像、文本、视频、传感器数据等）方法。围绕交通大数据分析目标，对需要用到的数据挖掘和机器学习（如预测、聚类、分类、相关性模型）模型和算法展开应用研究，针对超大规模数据量级，可利用深度学习和分布式集群等前沿大数据技术进行挖掘分析。利用大数据机器学习技术，融合碎片化、离散化数据，提升交通管理主动性、及时性和预见性。

（4）交通大数据可视化与辅助决策支持技术

交通大数据应用要落地，对挖掘分析结果的直观理解和形成辅助决策知识很关键。基于大数据分析的综合交通决策支持以交通管理、运筹控制等为基础，以大数据挖掘、可视化、仿真等技术为手段，以先进的人机交互方式辅助管理者进行科学决策。交通大数据具有多源异构、多维度、时空动态及海量的特征，这对现有可视化技术提出更高的要求，需基于地理信息系统（geographic information system，GIS）/全球定位系统（global positioning system，GPS）/卫星图像的大规模交通空间路网可视化、动态数据流可视化等关键技术，研究大数据可视化互动操作（钻取、多层次、多视角）。针对交通大数据分析结果的数据形态和逻辑关系，进行知识库建设，并设计基于大数据分析的智慧交通决策支持系统。

4 敏捷大数据方法论及其关键技术

根据上述交通大数据共性关键技术的

分析，可以看到，交通大数据的复杂性和分析目标的多样性，对传统的数据挖掘分析模式和大数据技术架构提出了挑战。例如，针对海量交通历史静态数据，需采用离线批处理技术，而动态实时流交通数据则需要流式计算框架进行处理。另外，对文本、图像、视频、传感器等多模态数据需采用不同的机器学习模型进行处理，如何进行跨模态的融合计算分析也是研究热点和难点。针对交通大数据分析需求的变更和扩展，大数据系统如何进行快速响应和功能、模型的扩充和调整，是交通大数据架构设计要解决的关键问题。针对这一问题，本文首次提出了敏捷大数据方法论，并基于信息融合、微服务化、容器等技术设计了具有快速响应、快速迭代、持续优化的敏捷大数据处理架构，为构建交通大数据分析平台提供有力支撑。

4.1 敏捷大数据方法论

4.1.1 大数据要具有敏捷性的原因

大数据分析的核心目标是：面向过去，发现数据规律，归纳已知；面向未来，挖掘数据趋势，预测未知。通过大数据分析提高对事物的理解和决策处置能力，最终实现智能化。不管是商业智能、机器智能、人工智能，还是智能客服、智能问答、智能推荐、智慧医疗、智慧交通等相关技术和系统，其本质都是朝着这一目标演进。随着云计算平台和开源大数据系统（如Hadoop、Spark、Storm等）的高速发展，获得大数据基础设施建设相关技术和支持越来越容易。同时，移动互联网和物联网技术具备的全面数据采集能力客观上促进了大数据的积累和爆发。但是，大数据应用落地还面临如下几个关键瓶颈。

- 信息技术（information technology，IT）向数据技术（data technology，DT）转变，使得传统软件技术架构面临挑战。现阶段各种大数据相关的开源技术和系统百花齐放，大数据技术生态体系庞杂，如云计算、机器学习、微服务、容器、分布式存储、分布式集群、深度学习等。如何在掌握有限技术的条件下（或已有技术不够成熟的条件下），快速敏捷地进行大数据研究和落地应用，需要从技术架构角度进行深入探讨和研究。
- 传统商业智能（business intelligence，BI）应用的失败教训太多，项目周期漫长，考验客户耐性，应用成本高，最终成果是昂贵的豪华报表难以达到预期目标，其失败的本质原因是缺乏敏捷性。大数据分析如何从BI项目中总结失败教训和获得经验，大数据应用如何与传统BI系统融合，传统企业的大数据应用如何与发源于互联网公司的主流大数据技术框架进行有机统一等问题，也需要从大数据架构层面进行研究和解决。
- 大数据应用标准化问题。特别是针对如交通大数据这样的多源异构、动态性、关联性等特点，对大数据分析流程和应用进行标准化、敏捷化管理，对离线分析、在线分析、实时分析、内存分析等计算框架进行快速切换处理，对图像、文本、视频、音频、网页、关系数据库等多源异构数据进行多模态融合处理，对大数据分析结果的应用效果进行快速量化与评价等问题，也需要敏捷架构的支持。

从大数据应用过程和特点来看，数据科学的本质是迭代，就好比婴儿的学习一样，输入—回应—反馈—学习—再输入，持续训练和学习才会产生智能，大数据分析系统是一样的道理，自适应优化和持续改进是大数据系统的必备特征。大数据时代，各种新兴技术和计算框架层出不穷，加之分析需求的不断变化，如何使大数据

架构能灵活调整以适应业务变化需求，并跟上技术快速迭代更新的步伐，是大数据应用要解决的关键问题，这是本文提出敏捷大数据架构的本质原因，相比传统信息处理架构，大数据架构需要极强的灵活性、可扩展性，或者说敏捷性。

4.1.2 敏捷大数据定义与处理流程

什么是敏捷大数据？本文给出一个初步定义："敏捷大数据是基于数据科学的迭代性本质，利用高效构件化和服务化技术，对大数据架构和关键组件等进行服务化设计，实现多粒度数据的融合处理和计算模型的标准化配置和管理，从而能根据特定分析需求快速搭建原型系统，快速迭代大数据分析结果，快速将原型转化为生产系统，提升大数据分析效率和数据决策价值"。敏捷大数据的核心要素遵从SFV（small-fast-validation）原则：从小的业务分析目标切入，快速出原型，快速迭代，证明有效之后再扩张。根据SFV原则，笔者对传统的跨行业数据挖掘标准流程（cross-industry standard process for data mining，CRISP-DM）[9]进行了改进，提出了基于微服务的敏捷大数据处理流程（agile big data processing based on micro-services），如图2所示。

敏捷大数据处理流程相比传统数据挖掘流程，从两个层面进行了扩展：首先是架构上采用基于容器的微服务技术进行支撑；其次，针对传统数据挖掘模型、系统和现实决策反馈的脱节问题，根据数据科学迭代性本质特点，对模型系统和现实决策反馈两个过程分别进行自适应迭代升级和决策优化支持。通过这种扩展，使得敏捷大数据方法论与传统数据挖掘技术以及新兴的主流大数据框架能在架构和流程上进行互补和融合。

4.2 敏捷大数据关键技术初探

（1）多粒度信息融合技术

针对大数据呈现的关联性、动态性、多源异构性等特点，如何对其进行标准化的融合处理，是敏捷大数据要解决的首要问题。与一般的大数据融合目标不同[10]，本文主要从构建标准数据集的角度进行大数据多粒度融合，通过构建统一数据单元（unified data unit，UDU）来支持多模态特征融合和多种类型、结构数据集的封装融合。将多源异构数据进行抽取、融合、集

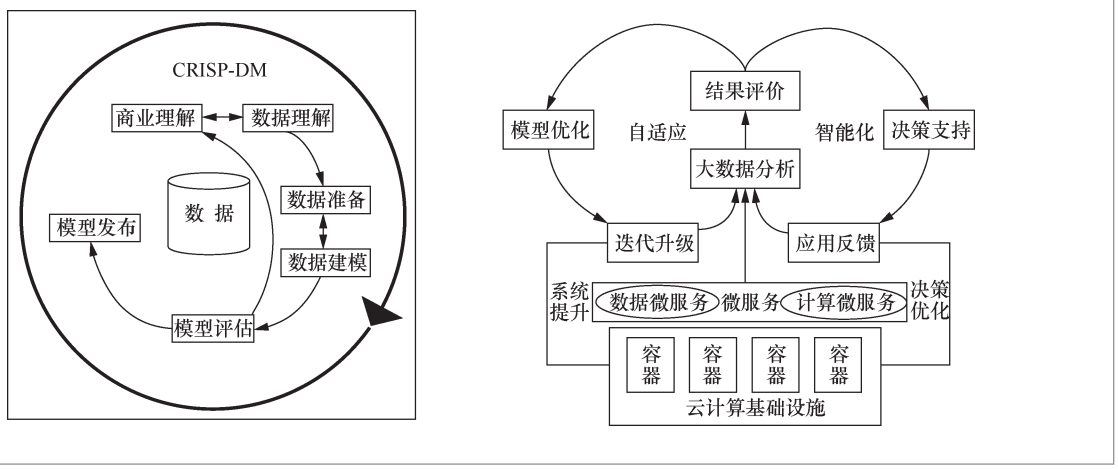


图 2 CRISP-DM 流程与基于微服务的敏捷大数据处理流程

成，形成支持不同计算模型处理的UDU数据集，是多粒度信息融合技术的核心目标。统一数据单元是独立、灵活的实体数据集，可随数据源和分析需求的变化进行快速重组、调整和更新。通过信息融合形成的UDU标准数据集，是进行敏捷大数据处理的基础。

（2）微服务与构件化技术

由于数据科学的迭代性本质，大数据分析即服务成为研究热点[11]，而微服务与容器技术能有效支持本文提出的敏捷大数据SFV核心原则。近年来，微服务成为互联网和大数据企业的研究和设计热点，诸如谷歌、亚马逊、Facebook、百度、京东、携程等公司都在采用微服务理论和技术进行产品的设计、研发和部署。Netflix公司的Adrian Cockcroft将微服务称为“细化面向服务的架构（service oriented architecture，SOA）”，并认为这是一套具备开创意义的新型架构。敏捷开发方法创始人Martin Fowler在他的《Micro services》一文中给出了微服务的定义①，概括来说，微服务设计思想是一种使用若干小服务开发庞大应用的方法，每个服务运行在自己的进程中，通过轻量级的通信机制进行信息交互，每个微服务的粒度基于业务能力大小进行构建，并可以由不同的程序语言实现，构建的服务链能够通过容器等技术进行自动化部署。

从上述定义可以看到，去中心化、原子化、独立自治、快速组合、自动部署等特性是微服务技术的核心要素，中心思想是将一个单体应用架构打散，把原来庞大的应用层功能切分成粒度较小的微服务模块，数据库也按微服务功能单元进行相应的拆分（如**图3**所示），采用基于云的容器技术单独运行这些服务模块，通过网络和轻量级通信机制将这些分解的服务模块协同连接起来，形成微服务簇和网络，完成大型复杂任务，这种通过将复杂系统切分成若干小的微服务模块的方式，其分布式、低耦合架构能极大地适应大数据分布式处理特性。

（3）容器技术

微服务技术采用类似搭积木的构建方法，使得服务之间不相互影响，而且同一个系统的微服务可以采用不同的开发语言和数据库技术。但是面对大型企业动辄数十、上百个信息系统，如何通过跨物理、虚拟、公共和私有云环境实现一致性、互操作性和可移植性，对IT基础设施来讲是极大的挑战，所以容器技术应运而生。容器最早由Docker公司提出并应用于自家的PaaS云服务平台，近年来获得广泛认可，很多大型企业开始将单体应用系统微服务化，并部署在容器中。容器基于操作系统之上能实现相比传统虚拟化技术（如VMware）更轻量级的虚拟化[12]，而且完全使用沙箱机制，相互之间没有接口。Hadoop的子系统Kubernetes已经能支持基于云计算和Docker容器技术的微服务开发和部署[13]，容器技术与云计算的天然结合及其快速的研发、部署、维护优势，对于微服务和敏捷大数据架构的设计和实现具有重要支撑作用。

5 面向交通的敏捷大数据架构设计

如何根据不同的交通大数据分析需求设计灵活的处理架构进行支持，或者说大数据架构能否从采集、存储、计算、应用多个层面，满足不同分析需求的变更或扩张，是交通大数据分析要解决的关键问题。基于敏捷大数据方法论及其关键技术的分析和研究，从数据采集集成、大规模数据存储和数据融合、多模式/多模态计算微服务、数据应用4个层面进行了敏捷

① http://martinfowler.com/articles/microservices.html

大数据架构设计。通过基于多粒度大数据集成融合构建统一数据单元，形成标准数据集，通过基于微服务的计算模型抽象和汇聚层处理，实现交通大数据挖掘的敏捷化、服务化。对大数据主流处理框架（如Hadoop、Spark、Storm等）进行统一配置管理，基于即插即用的构件化和服务化设计，各层子系统和组件可根据分析目标进行快速选型、灵活配置，构建原型和迭代升级（如**图4**中根据两条虚线不同的设计路径，可以快速配置实现历史数据库数据的批处理分析，或公网采集数据的流处理分析），总体架构设计如**图4**所示。

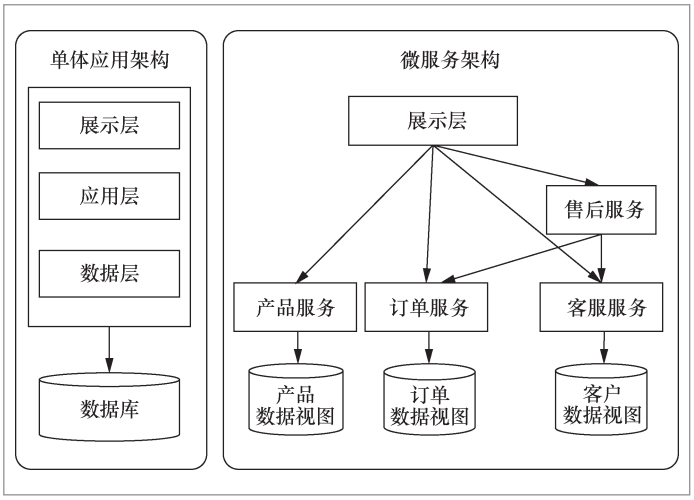


图 3　传统单体应用架构与微服务架构的区别

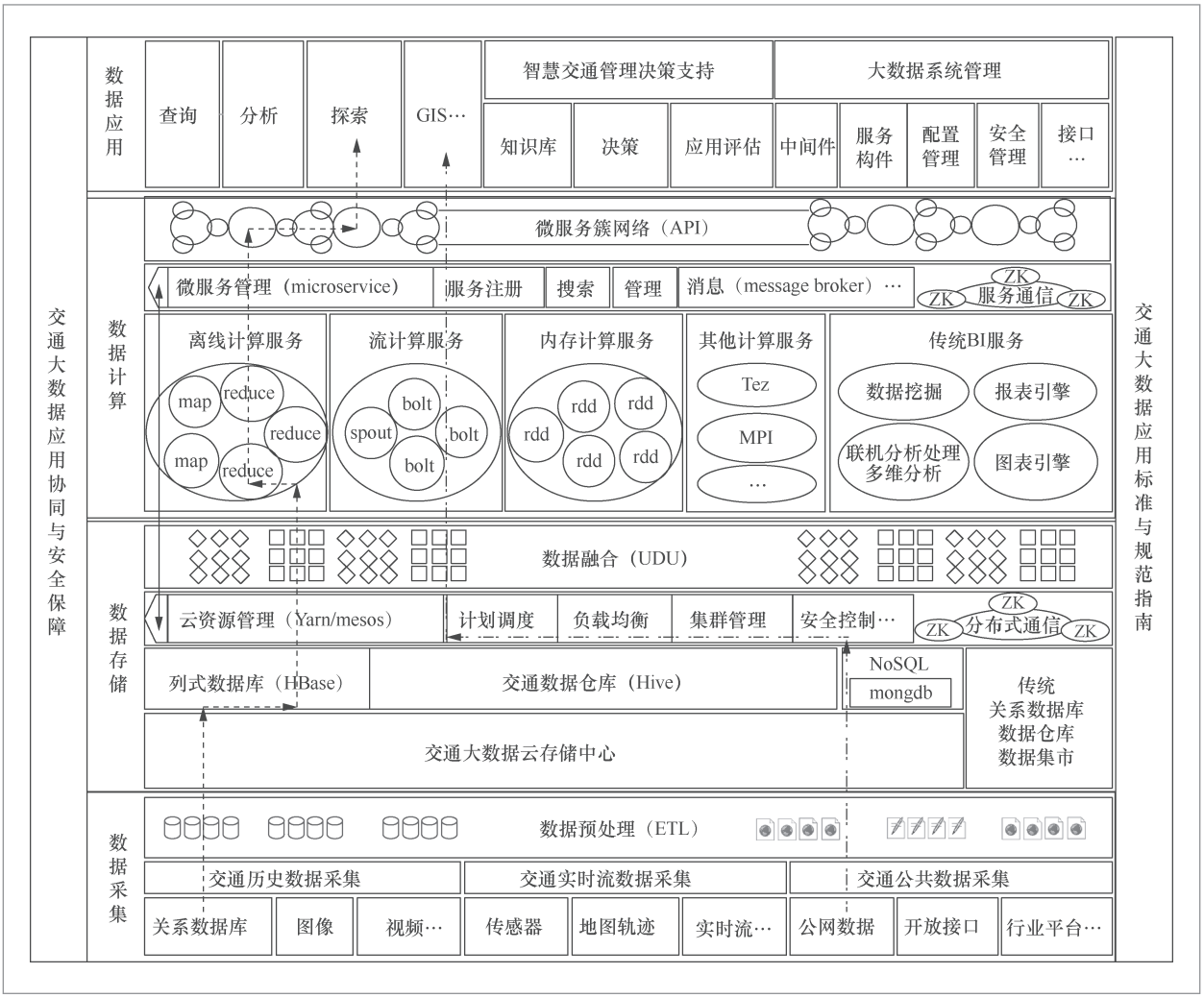


图 4　面向交通的敏捷大数据总体架构设计

● 数据采集层：分3个层面的采集技术进行支持，一是传统业务系统数据库和半结构化、结构化数据的采集和集成，如采用Sqoop技术进行关系数据库和Hadoop系统之间的数据抽取和交换；二是交通实时流数据的采集，包括实时传感器数据、定位轨迹数据和其他实时流数据；三是交通公共数据的采集，包括公网的数据爬取、开放平台的数据接口、行业公共数据库的数据交换等。对采集到的数据需进行提取、转换和加载（extract-transform-load，ETL）处理，包括数据抽取、转换、清洗和隐私脱敏等预处理工作，预处理集成后的数据进入交通大数据云存储中心。

● 数据存储层：交通领域数据规模巨大，数据存储层需设计基于云计算的分布式云存储系统，以支持海量数据的存储扩展。提供基于云的列式存储、NoSQL存储或数据仓库存储能力；根据业务需求和快速配置，可切换相应的分布式存储模式，还可根据需要对传统BI系统的数据仓库和数据集市进行集成。利用Hadoop集群提供PB级存储能力扩展，同时Hadoop Yarn和Spark Mesos等集群资源管理框架可支持多种存储模式和计算模式（主要对存储和计算两个层面资源进行管理，如**图4**中双向箭头所示）的云资源调度。在此基础上，对各类存储数据进行多粒度信息融合，构建统一数据处理单元，为计算服务层提供标准化的分析数据集。

● 数据计算层：针对交通大数据多源、异构、海量等特征，传统的计算模型难以直接处理。数据计算层需满足主流大数据处理框架的各种计算模型与方法实现，如基于云计算并行框架，实现基于Hadoop批处理、Storm流处理、Spark内存处理的高效数据挖掘与机器学习。数据计算层采用基于统一数据处理单元和计算模式、模型微服务化的大数据分析框架，通过构建多种微服务簇网络（主要分为计算微服务、数据微服务和流程微服务，涵盖存储和计算两个层面，如**图4**中双向箭头所示），为应用层提供支持MapReduce、Storm、Spark等多种计算模式下的多种数据挖掘模型与方法（如分类、聚类、序列等）；根据大数据分析需求和数据特征，可基于组件配置和服务治理技术进行各类服务的快速切换和灵活管理。

● 数据应用层：数据应用层首先要满足智慧交通各类大数据分析需求，包括基本的可视化与查询、展示、探索等，分析结果能结合知识库应用于决策支持。另外，在大数据系统本身管理方面，针对构件化和微服务设计这个特点，需对相关中间件进行设计，实现服务治理、组件配置、安全、接口等功能，以支撑存储层和计算层各类微服务的敏捷管理。

5.1 交通大数据多粒度信息融合设计

针对机器学习各类模型的数据适配特点，本文提出了一种标准数据集定义：统一数据单元对交通大数据多个层次和粒度的信息进行融合处理。特别是对多模态数据，通过统一数据单元设计方法能实现数据的存储优化和机器学习模型数据输入的标准化，以统一数据单元作为敏捷大数据系统的基本数据组织和处理单元，能提高大数据分析模型和算法的适应性和敏捷性，提升大数据处理能力。多粒度信息融合设计如**图5**所示。

设计和构造大数据统一数据单元，需进行如下3个环节的处理。

● 特征抽取：对交通领域的结构化数据、半结构化数据和非结构化数据进行数据集成和特征抽取，抽取出数据中的各类不同特征，包括时间特征、空间特征或其他全局特征等，实现对数据相关的位置属

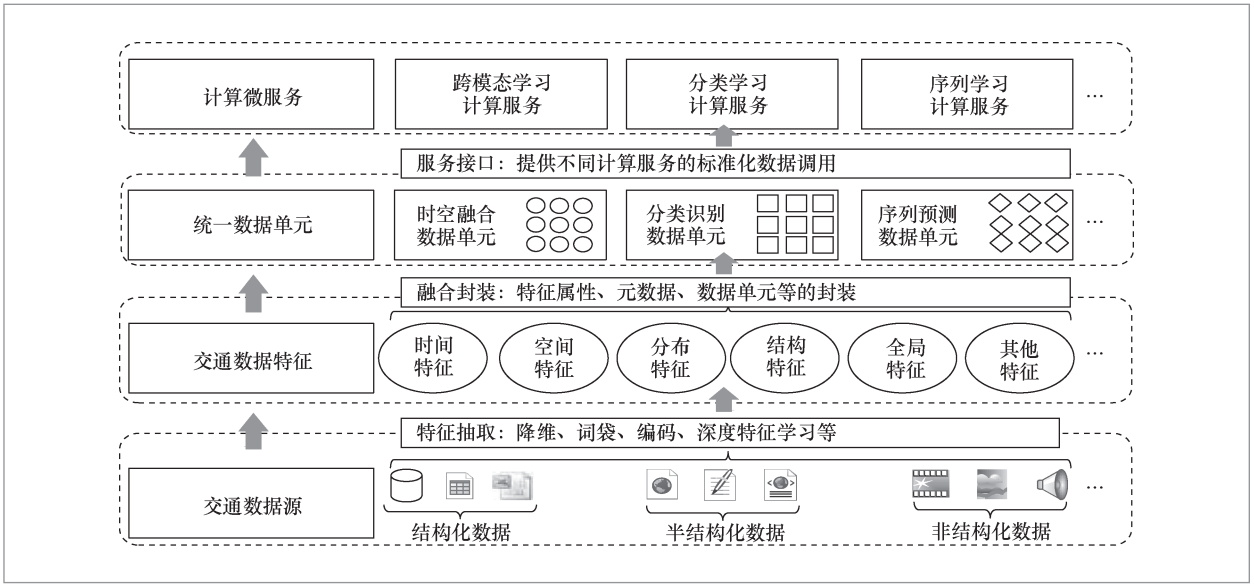


图 5 交通大数据多粒度信息融合设计

性、时间空间关联属性和其他观测属性等的特征描述。

● 融合封装：抽取出来的各类数据特征，或者初步预处理后的数据，根据不同的计算模型数据处理特点和要求，封装成结构和格式统一的数据处理单元，形成标准分析数据集，为上层的挖掘计算服务提供快速数据适配。本文通过元数据定义方法和JSON数据交换技术，实现不同种类的统一数据单元定义，对每类统一数据单元进行基本信息和各类属性定义和描述，包括标识ID、基本属性、语义属性、结构属性等内容。

● 服务接口：封装好的统一数据单元数据集，针对不同的挖掘计算服务模型实现快速数据适配，设计统一的数据单元调用接口，通过接口定义和参数设置对封装数据单元进行解析，并对数据集各类属性特征、结构信息等进行提取。

5.2 交通大数据计算微服务化设计

实现基于统一数据单元的标准数据单元的敏捷大数据挖掘分析，还有一个关键环节就是对多种计算模式框架和不同计算模型的微服务化设计，包括计算框架、模型和算法的微服务化，数据获取调用的微服化和分析流程的微服务化3个层面。其核心是计算微服务，如MapReduce批处理服务、Storm流处理服务、Spark内存处理服务，每种大的计算框架微服务下包含具体挖掘模型等小粒度的计算微服务支持（如分类模型服务、序列模型服务）等，交通大数据微服务化架构设计如图6所示。

大数据微服务化的核心理念是一个服务只专注做好一类或一个分析，服务的粒度和分析功能大小、边界要匹配，服务方（计算微服务）和消费方（基于微服务的分析功能）要能解耦，即调整或升级一个微服务不能影响其他微服务。由于微服务详细技术涉及面比较广，包括微服务的注册、定位、发现和搜索（基于分布式一致算法Paxos和Zookeeper框架等）；微服务的轻量级通信机制，如表述性状态传递（representational state transfer，REST）、远程过程调用协议（remote

procedure call protocol, RPC)、进程间通信(inter-process communication, IPC)等; 微服务的容错处理(熔断、限流、负载均衡等); 微服务依赖的容器技术; 服务的自动测试和部署等方面。由于篇幅原因和架构设计主题, 各方面的技术细节这里不做赘述, 在交通大数据敏捷处理总体架构和数据融合统一数据单元基础上, 进行了交通大数据计算的微服务设计(如**图6**所示), 主要设计内容包括如下几个方面。

(1)交通大数据分析业务抽象和微服务划分

按多模式计算框架划分, 有MapReduce批处理框架、Storm流式处理框架、Spark内存处理框架、图计算框架等; 按挖掘模型划分, 有分类、聚类、序列、多模态、多任务等不同种类计算模型。针对交通大数据各类分析需求和数据处理特点, 进行大数据业务分析和抽象建模, 并选择相应的计算模型和计算框架进行支撑, 再决定需要哪些微服务, 并实现微服务的划分和组合, 设定好微服务网络的总体设计目标, 并通过统一的微服务接口(microservices API gateway)进行调用。

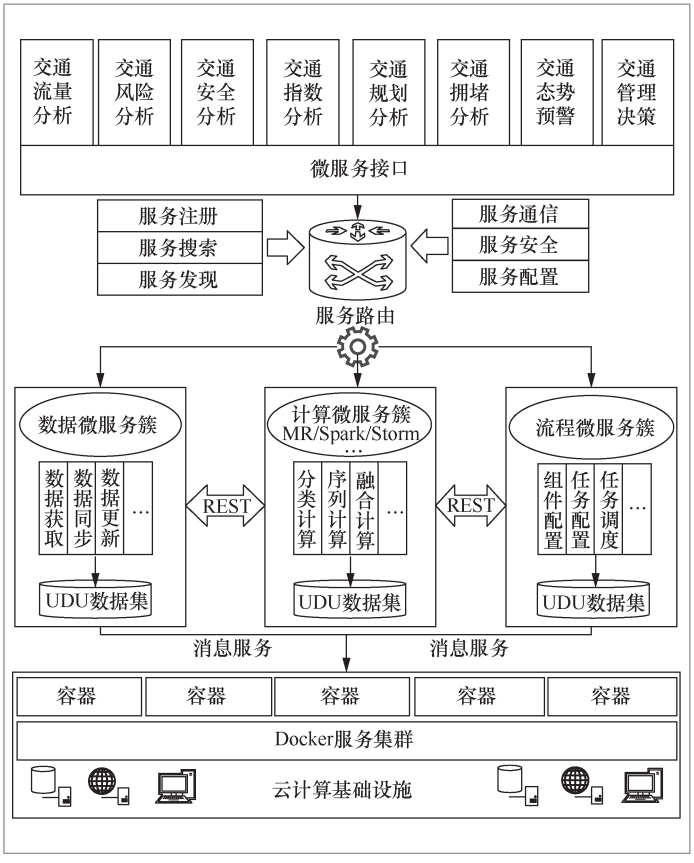


图 6 交通大数据微服务化设计

(2)微服务簇设计及契约接口定义

针对大数据处理特点, 服务层中的微服务分3类进行规划设计, 数据微服务簇负责从UDU标准数据集中进行数据获取、数据同步和更新等操作; 计算微服务簇是敏捷大数据处理的核心, 按多模式计算框架和多类计算模型两个维度进行挖掘分析服务的统筹设计; 流程微服务簇负责数据微服务、计算微服务的协同处理, 同时对系统组件的配置管理和调度进行支持。各类微服务通过REST、RPC等轻量级通信机制和MessageBroker等消息服务进行交互和联系[14], 构建微服务簇网络, 并通过服务路由进行统一管理和调度。

(3)微服务治理和容器部署

由各类微服务簇连接成的微服务网络, 其高效协调工作离不开微服务治理技术和容器管理技术。通过服务路由和服务治理负责各种大小微服务的注册、搜索、发现、通信和统一配置, 最后基于云计算和容器技术进行微服务的自动部署和动态管理。

6 设计评述与应用实践

本文提出的敏捷大数据架构在一定程度上实现了大数据融合处理和挖掘计算的服务化、标准化和流程化。基于统一数据单元和计算、流程微服务设计思想, 对微

服务的构建和管理进行快速实现，通过各种微服务的划分、组合、编排和动态配置，能构建模块化、可配置、可复用、可扩展的敏捷大数据分析系统。相比传统大数据处理架构，敏捷大数据架构在如下几个方面具有明显优势（见**表1**）。

从上述各项特性的比较分析可以看到，敏捷大数据架构除了支持大规模数据增长之外，更重要的是能适应大数据分析业务的扩展和变化，由于采用容器和服务化技术，具备高容错性和稳定性，能支持大数据分析的失败处理和自动恢复等，能在海量数据条件下快速完成多种计算模型和分析处理，能基于数据融合单元和计算服务化技术，支持多模态计算和多模式计算任务处理，能在有限的成本条件下提高大数据分析效率。

在实践方面，结合正在规划建设的综合交通大数据分析云服务平台，以本文提出的敏捷大数据架构设计作为关键技术选型和技术路线实现的指导思想，进行了初步应用。综合交通大数据分析云服务平台的主要目标是通过大数据技术解决交通资源的供需智能匹配和瓶颈预测分析问题。应用多粒度信息融合和多模态计算微服务技术，对交通大数据进行集成、融合和挖掘；采用定量分析和定性分析相结合的机器学习进行供求配置预测，为智能出行推荐、交通枢纽瓶颈分析、多模接驳换乘、实时交通管控等智慧交通关键环节提供大数据分析和决策支持。由于综合交通大数据多源异构、时空关联和动态处理等特点，传统大数据架构面临扩展性、兼容性、稳定性诸多问题。基于敏捷大数据方法论，设计了面向智慧交通的具有构件化、云服务化、容器化特性的敏捷大数据架构，为综合交通大数据分析云服务平台的研发提供了切实参考和应用指导，并在一定程度上提高了开发效率，控制了技术风险。

7 结束语

本文论述了交通大数据在智慧交通应用中的核心地位与相互关系，探讨了交通大数据关键技术和面临的挑战。针对交通大数据特点及其分析瓶颈问题，首次提出了敏捷大数据方法论及其关键技术和核心原则，通过分析敏捷大数据的设计缘由，基于传统数据挖掘流程改进，设计了面向微服务的敏捷大数据处理流程，并对其关键支撑技术进行了初步研究和探

表1 敏捷大数据架构与传统大数据架构比较分析

特性	传统大数据架构	敏捷大数据架构
扩展性	主要支持数据存储扩展	存储扩展之外，能支持计算框架和分析模型扩展
容错性	单体应用高耦合，难恢复	微服务自治隔离，出错不会相互影响扩散，容易恢复
开放性	接口难以适应业务变化	服务化、标准化的轻量级接口，能快速适应业务变化
兼容性	难以兼容历史遗留系统	容器和微服务支持多种技术栈的快速集成和前后向兼容
数据融合	难以处理，需针对性开发	通过统一数据单元的灵活构建，支持多模态数据处理
多模式计算	难以处理，需定制开发	各种计算框架的服务化，方便技术选型和扩展
维护管理	单体应用维护难度大	通过服务自治，容器隔离运行，维护升级快捷
成本和效率	周期长、投入大、效率低	快速、敏捷、低成本
技术风险	新技术泛型下风险较大	基于敏捷架构，快速迭代试错，能有效控制技术风险

索。构建了基于微服务和多粒度信息融合技术的敏捷大数据架构，并对交通大数据微服务化、交通大数据融合等关键技术环节进行了详细设计和论述。敏捷大数据架构的提出是基于数据科学迭代性本质的，为交通大数据环境下的高效、灵活机器学习和知识发现提供了新方法、新思路和新的技术架构，从而最大程度地发挥交通大数据的应用价值，相比传统大数据处理技术和架构，本设计的重要意义和参考价值不言而喻。当然，敏捷大数据作为一个新的涵盖多种前沿信息技术的跨领域研究问题，还需要在设计方法、关键技术和系统架构等方面进行深入研究和应用实践。

参考文献:

[1] 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(6): 647-657.
LI G J, CHENG X Q. Research status and scientific thinking of big data[J]. Bulletin of the Chinese Academy of Sciences, 2012(6): 647-657.

[2] 郑宇. 城市计算概述[J]. 武汉大学学报(信息科学版), 2015(1): 1-13.
ZHENG Y. Introduction to urban computing[J]. Geomatics and Information Science of Wuhan University, 2015(1): 1-13.

[3] 杨芙清, 梅宏, 李克勤. 软件复用与软件构件技术[J]. 电子学报, 1999(2): 69-76.
YANG F Q, MEI H, LI K Q. Software reuse and software component technology[J]. Acta Electronica Sinica, 1999(2): 69-76.

[4] 王珊, 王会举, 覃雄派, 等. 架构大数据: 挑战、现状与展望[J]. 计算机学报, 2011(10): 1741-1752.
WANG S, WANG H J, QIN X P, et al. Architecting big data: challenges, studies and forecasts[J]. Chinese Journal of Computers, 2011(10): 1741-1752.

[5] 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.
MENG X F, CI X. Big data management: concepts, techniques and challenges[J]. Journal of Computer Research and Development, 2013, 50(1): 146-169.

[6] CHEN H M, KAZMAN R, HAZIYEV S. Agile big data analytics development: an architecture-centric approach[C]// Hawaii International Conference on System Sciences, Jan 5-8, 2016, Koloa, HI, USA. New Jersey: IEEE Press, 2016: 5378-5387.

[7] NAMIOT D, SNEPS-SNEPPE M. On micro-services architecture[J]. International Journal of Open Information Technologies, 2014, 2(9): 24-27.

[8] 程学旗, 靳小龙, 王元卓, 等. 大数据系统和分析技术综述[J]. 软件学报, 2014(9): 1889-1908.
CHENG X Q, JIN X L, WANG Y Z, et al. Survey on big data system and analytic technology[J]. Journal of Software, 2014(9): 1889-1908.

[9] WIRTH R, HIPP J. CRISP-DM: towards a standard process model for data mining[C]// The Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, April 18-20, 2000, Kyoto, Japan. [S.l.:s.n.], 2000: 1-11.

[10] 孟小峰, 杜治娟. 大数据融合研究: 问题与挑战[J]. 计算机研究与发展, 2016(2): 231-246.
MENG X F, DU Z J. Research on the big data fusion: issues and challenges[J]. Journal of Computer Research and Development, 2016(2): 231-246.

[11] ZHENG Z, ZHU J, LYU M R. Service-generated big data and big data-as-a-service: an overview[C]//IEEE International Congress on Big Data, June 27- July 2, 2013, Santa Clara, CA, USA. New Jersey: IEEE Press, 2013: 403-410.

[12] FELTER W, FERREIRA A, RAJAMONY R,

et al. An updated performance comparison of virtual machines and Linux containers[J]. Lecture Notes in Computer Science, 2014(1140): 438-453.

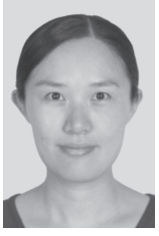
[13] BREWER E A. Kubernetes and the path to cloud native[C]// The Sixth ACM Symposium on Cloud Computing, August 27-29, 2015, Kohala Coast, Hawaii, USA. New York: ACM Press, 2015: 167-167.

[14] LI L, TANG T, WU C. A REST service framework for fine-grained resource management in container-based cloud[C]// IEEE International Conference on Cloud Computing, June 27-July 2, 2015, New York, NY, USA. New Jersey: IEEE Press, 2015: 645-652.

作者简介

杜圣东(1981-),男,西南交通大学信息科学与技术学院讲师,中国计算机学会(CCF)和国际计算机学会(ACM)会员,主要研究方向为数据挖掘、机器学习、交通大数据、医疗大数据等。

杨燕(1964-),女,西南交通大学信息科学与技术学院教授、博士生导师,主要研究方向为大数据分析与挖掘、人工智能、教学管理等。



滕飞(1984-),女,博士,西南交通大学信息科学与技术学院讲师,CCF和ACM会员,主要研究方向为云计算、工业大数据等。

收稿日期: 2017-03-27

通信作者: 杨燕, yyang@swjtu.edu.cn

基金项目:国家重点研发计划基金资助项目(No. 2016YFC0802209);国家科技支撑计划基金资助项目(No. 2015BAH19F02)

Foundation Items: The National Key Research and Development Program of China(No. 2016YFC0802209), The National Science and Technology Support Program of China(No. 2015BAH19F02)