大数据系统软件创新平台与生态建设

王建民[1,2]，王晨[1,2]，刘英博[1,2]，刘璘[1,2]
1. 清华大学软件学院，北京 100084
2. 大数据系统软件国家工程实验室，北京 100084

摘要
针对大数据系统软件开源生态发展现状、大数据系统软件技术与领域应用发展的迫切需求与工程化瓶颈，系统地介绍了大数据系统软件共性技术的研发思路以及大数据系统软件国家工程验室研制的“数为平台”的技术架构，包括大规模多源异构数据的一体化管理、异构数据的交互式分析框架、数据可视化与智能数据工程、大数据混源软件可信验证、领域大数据应用开发与运行环境等。给出了基于平台建设开放的大数据软件创新平台与生态的模式，并在工业制造业、气象环保服务领域开展应用示范。

关键词
大数据系统软件；开源软件生态；领域大数据应用；国家工程实验室

中图分类号：TP391 文献标识码：A doi: 10.11959/j.issn.2096-0271.2018054

Big data system software eco-system and platform construction

WANG Jianmin[1,2], WANG Chen[1,2], LIU Yingbo[1,2], LIU Lin[1,2]
1. School of Software, Tsinghua University, Beijing 100084, China
2. National Engineering Laboratory for Big Data System Software, Beijing 100084, China

Abstract
In view of the bottlenecks in common technology and engineering practice faced by big data system software and applications development, the key technological innovations of the National Engineering Laboratory for Big Data System Software(NEL-BDSS) were introduced systematically, including the technical architecture and domain applications of the big data system software "Tsinghua Dataway Platform", which addresses issues such as: massive multi-source heterogeneous data integration management, interactive heterogeneous data analysis framework, data visualisation and intelligent data engineering, validation and verification of hybrid source big data software, and domain-specific big data applications development and run-time environment. The big data system software eco-system, state-of-the-art big data technology and systems, domain applications, as well as future challenges were summarized systematically. The NEL-BDSS focuses on supporting demonstrative applications of industrial big data, environmental big data as well as meteorological big data.

Key words
big data system software, open source software eco-system, domain-specific big data applications, National Engineering Laboratory

1 引言

大数据系统软件连接大数据计算平台与大数据应用软件，是大数据系统的基础。大数据系统软件对分布式硬件平台资源进行统一抽象，为大数据处理提供一体化存储管理、分析框架、全生命周期数据工程、安全保障等基础共性功能，为大数据应用提供开发运行环境，是大数据产业的“核心驱动程序”。

国际大数据系统软件的发展以开源软件为主导方式，逐渐形成了以Hadoop、Spark、TensorFlow生态为代表的、支持大数据应用开发的共性基础软件与工具平台。例如以国际开源软件基金会（Apache）等社区为平台，实现了技术的快速演进，IBM等公司也以开源软件为基础，包装了各自的商业化发行版本。Hadoop 是大数据领域非常重要的开源框架，基于简单的编程模型，对大型数据集进行分布式处理。Spark是面向大规模数据处理的通用引擎，现已形成包括批处理、流式处理、统计分析、图计算等在内的数据分析软件栈。TensorFlow是将复杂的数据结构传输至人工智能神经网进行分析和处理的开源软件平台，用于语音识别或图像识别等机器学习和深度学习任务。相比于成熟的关系型数据库技术，大数据时代的数据处理技术与系统更多被认为是“负载优化系统”。以谷歌公司为例，针对其网页搜索服务的需求设计了分布式存储文件系统（GFS）、基于列族的无模式（schema-less）存储与键值索引、基于任意数据切分（MapReduce）的高容错分布式计算框架。

顺应大数据系统软件开源实践，聚焦领域化与生态化创新，是大数据系统软件的发展方向。大数据领域国家工程实验室作为一个公共的平台，通过大数据系统软件关键技术的创新和突破，促进大数据与各行业应用的深度融合，以应用带动大数据系统软件技术和产品研发，形成一批代表性示范应用案例。

2 大数据系统软件工程挑战

在“互联网+”时代，大数据应用正在从消费互联网发展到产业互联网，领域大数据应用与消费互联网应用有显著区别，数据类型以时序数据、科学数据、过程数据以及非结构化工程数据等数据类型为主，其使用者也从互联网公司的高级技术人员转变为传统领域的技术与业务人员。这对当前主要基于开源项目的大数据系统软件技术和产品在领域数据特征、多源知识融合、数据处理过程、应用开发效率、安全可控保障等方面提出了新的挑战。大数据系统软件关键技术的研发需求主要包括以下几个方面。

（1）针对领域数据特征的挑战

为应对开源数据管理引擎产品能力缺失的问题，研发面向设备高频时间序列数据、遥感数据、科学数据等领域数据的管理技术，整合多种类型异构数据管理引擎的一体化管理技术以及面向新型硬件设备的优化技术，解决“不管用”的问题。

（2）针对多源知识融合的挑战

为解决开源数据分析框架交互性与异构性支持不足的问题，研发面向人与机器、算法协同工作环境，实现差异化数据类型与负载和异构硬件设备的分析框架自动适配，降低领域人员分析应用门槛，解决“不会用”的问题。

（3）针对数据处理过程的挑战

为解决开源大数据软件栈中数据工程

工具集较为欠缺的问题，研发数据采集、数据集成、数据清理、数据可视化探索分析等丰富的数据工程支撑工具，提升数据处理的效率，解决“不够用”的问题。

（4）针对应用开发效率的挑战

为应对开源大数据系统软件产品在应用开发上的能力缺失问题，研发支撑大数据应用开发与运行一体化、智能化的管理工具，领域大数据实验床以及面向新型硬件设备的优化能力的研发和工程化，解决“不好用”的问题。

（5）针对安全可控保障的挑战

为应对开源软件代码来源（链）难以控制、版本变化迅速、质量缺乏统一标准的问题，研发面向大数据应用的混源大数据系统的可信验证体系，实现对大数据系统软件可靠性、可用性、正确性等方面的可信验证，解决“不敢用”的问题。

最终，在大规模多源异构数据一体化管理、交互式异构数据分析框架、数据可视化与智能数据工程、领域大数据应用开发运行环境、大数据混源软件可信验证等领域取得重大突破，通过以上5个方面形成大数据系统软件体系，促进中国大数据产业的发展。

3 大数据系统软件技术架构

以工业领域为例，从领域数据特征角度来看，机器设备产生的高通量时间序列数据、设计仿真阶段的非结构化工程数据等领域数据缺少有效的管理技术，同时对不同种类数据进行一体化管理的能力不足，缺少管用的大数据系统软件；从多源知识融合的角度来看，现有大数据分析软件使用门槛高、交互性支持不够，仅提供计算框架与分析类库，很难让具有机械、电子、液压、自动化等领域知识的技术人员独立使用；从数据处理过程的角度来看，由于大数据可视化、数据集成、数据清理等数据工程工具缺失，工业领域的技术人员会花费大量的时间反复尝试手动对数据进行处理，分析效果也受到了制约；从应用开发的角度来看，大数据应用系统开发涉及的系统选型、参数配置、模型选择、运行优化等工作步骤都需要基于专家知识进行反复迭代，应用开发使用门槛过高，加大了开发成本；从安全可控保障的角度来看，混源大数据系统的安全可信成为大数据开发利用亟待解决的核心问题。大数据系统软件国家工程实验室针对以上问题与需求，重点在大规模多源异构数据一体化管理、交互式异构数据分析框架、数据可视化与智能数据工程、领域大数据应用开发运行环境、大数据混源软件可信验证等领域开展研发，以形成开放的大数据系统软件体系，提升中国大数据产业的实践能力。

3.1 大规模多源异构数据一体化管理技术

多源异构数据是指数据源不同、数据结构或类型不同的数据集合。数据源不同、数据类型不同，使得这类数据集的使用变得非常复杂，因此大规模多源异构数据的一体化管理技术变得十分重要。需要突破的是针对不同类型数据的存储与查询技术，并在充分考虑多源异构数据的来源和结构随着时间的推移不断增加与变化的特点的情况下，研究如何形成可扩展的一体化数据管理系统。

现有大数据系统软件开源社区及商业产品对图数据、关系数据、键值数据等类型已经有比较成熟的技术。但是对各行业领域中广泛存在的时序数据、科学数据、过程数据以及非结构化工程数据的管理能力则有所欠缺。此外，随着计算机

硬件技术的发展，越来越多的、性能强大的新型存储硬件加入了大数据管理系统的硬件平台中，如固态硬盘（solid state drives，SSD）、非易失性随机访问存储器（non-volatile random access memory，NVRAM）等，促进了灵活、可定制的数据中心架构的出现，这对如何进行异构硬件和系统软件的有效管理和协同优化提出了新的挑战。

这项研究的重点是多种类型数据的分布式存储技术，重点开发时序数据、过程数据、科学数据、非结构化工程数据等重要领域数据的分布式存储系统；研究多源异构数据的统一存储模型、统一数据查询接口及查询协同优化技术，开发多源异构数据一体化管理系统；研究新型存储硬件与开放计算平台上的数据管理软硬件协同优化技术。

3.2 交互式异构数据分析框架技术

对大规模多源异构数据进行高效分析计算是挖掘大数据价值、发现特殊规律和洞见的关键。在分析任务中，分析师通过分析算法库运行需要的统计算法或机理模型，借助分布式或并行计算框架完成大规模模型训练，并对分析结果进行解释。上述过程不断迭代，得到最终的决策结论。在此过程中，分析师需要根据任务需求调用不同的算法库，根据不同计算硬件的编程环境手动编写、调用各种计算资源的程序，并且对黑盒的分析模型进行解释，整个分析流程的关键步骤没有完整贯通，分析门槛很高。

这项研究的重点是开发面向新型计算硬件、国产CPU整机系统、超级计算机集群的高性能计算框架，研制分布式计算与高性能计算的融合计算框架，实现多层异构软硬件计算资源的统一调度和协同优化；开发分析模型可视化工具，实现分析算法的智能调试与人机交互协同；开发分析工作流管理工具，实现大数据分析生命周期的流程化管理，提高数据分析的易用性。

3.3 数据可视化与智能数据工程技术

完整的数据分析流程除了数据分析活动外，还有大量的数据准备活动，包括数据的获取、清理、集成以及分析之前的数据探索和分析之后的结果可视化。在实践中，这些步骤需根据分析目标和结果进行反复的迭代与试错，流程中需要的工具被统称为数据工程工具集，承担这部分工作的人员主要是数据分析师和领域专家。数据工程技术在大数据时代遇到的挑战有两点：首先，现有产品针对海量异构数据的支持不足，可扩展性差，且对时间序列等新型数据的数据工程方法缺乏支持；其次，现有产品的智能化程度不高，在操作中需要大量人工参与，要求具备较丰富的领域知识，使用效率不高，过程繁复。

针对上述问题，需重点开发支持海量数据的可视化分析工具，研发面向工业、国家安全、公共服务等领域中的文本、复杂网络、时序数据、高维特征数据4类典型大数据的可视化工具。开发多源异构数据的智能数据获取、数据画像、智能数据清理、智能数据集成等智能数据工程工具，解决传统数据工程任务耗时长、成本高的问题。

3.4 大数据混源软件可信验证技术

一个具体的大数据应用解决方案往往由开源系统与闭源系统结合完成，或以黑盒的方式调用外部服务，这就形成了大数据混源系统。大数据混源系统的分布式特性、开源/闭源结合的特性导致其代码量庞大，版本匹配、接口匹配、数据安全难以控制，这就对其可信验证提出了迫切的需

求。现有的可信验证技术体系尚未综合考虑大数据混源系统的上述两种特性，例如将系统当作黑盒的基于测试用例的和基于动态测试的方法、基于代码分析的静态分析方法、基于众包数据的数据挖掘方法、基于系统设计原理的模拟仿真方法等，都难以支持大数据混源软件的可信验证。此外，大数据系统的复杂性导致其可信验证中难以避免状态空间爆炸的问题，导致验证效率不高。

这项研究的重点在于结合分布式系统建模技术和混源系统可信建模技术，研发大数据系统软件的综合建模工具，支持分布式环境下节点间的消息传递、任务的异步分发与同步汇总等分布式特征的建模，支持自主可控代码、第三方开源或闭源代码的混源特征建模，并支持两类建模技术的有效统一；研究基于高性能计算的并行可信验证算法，开发高性能可信验证系统，大幅提升软件验证的规模和效率。

3.5 清华“数为平台”技术框架

大数据系统软件国家工程实验室为实现大数据系统软件的快速、自动、智能构造和运维，突破大数据系统软件技术瓶颈，研发了大数据系统软件与工具栈——清华“数为平台”，其总体框架如**图1**所示。“数为平台”采用模型驱动的软件系统构造方法，提供了交互式的数据建模与业务建模支持，是支持可定制性和可扩展性的大数据应用服务开发平台，降低了领域大数据系统研发的代价。在以HDFS、Kafka、TensorFlow为代表的大数据系统开源软件生态基础上，重点研发了以下大数据系统软件构件。

- 可定制的数为框架（dataway framework，DWF）：提供对数据、功能、表单的底代码量可定制平台，支持移动端和PC端定制，具有灵活多样的可扩展性，能够对接主流大数据系统（Strom、Hadoop、Kafka、Spark、TensorFlow等）。从系统开发构建的角度看，可定制的数为框架是大数据软件组件集成的数据总线、控制总线和交互总线，负责数据源/数据集管理、算法管理、构件定义、构件交互、运行监控和人机交互设计组件。

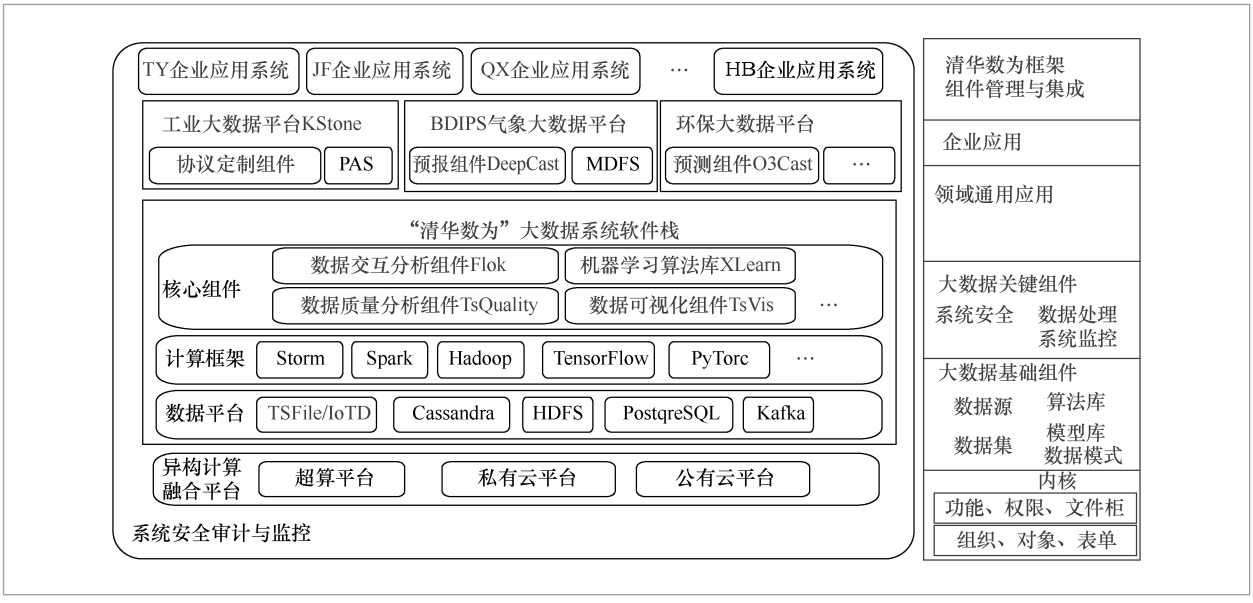


图1 大数据系统软件“数为平台”总体框架

● 多源数据和数据集管理：时序数据库TSFile/IoTDB具有多数据源接入能力，支持结构化、非结构化、时序数据数据源定制，可围绕数据源定义不同种类数据集。

● 算法库管理：内嵌数十种数据处理算法，包括数据提取、数据清洗、数据集成、统计分析、机器学习、深度学习、迁移学习等。

● 数据处理流程管理器：支持基于流程的数据处理活动选择、编排与执行，提供数据处理过程监控。

● 数据质量控制：建立数据质量分析知识库，批量分析海量数据特征，实时评估数据质量情况，并在条件具备的情况下对低质量数据给出修正建议。

● 机器学习工作坊：针对不同种类的数据集开展大规模训练任务，产生训练模型，并且可以可视化评估机器学习的过程，找到瓶颈，帮助调优。

● 模型库管理：可以管理机器学习算法产生的各种预测模型，并基于模型提供在线数据服务。

4 领域大数据应用研发——“1+X”模式

大数据技术在领域中的应用和落地涉及一个复杂的大数据全生命周期，包括大数据的采集提取清理（数据工程）、管理、分析、可视化等多个主要阶段，如**图2**所示。各阶段需要满足海量、异构、实时、人机交互等功能或性能需求。例如，大数据的采集通常是在多源异构环境中进行的，难以保证足够高的数据质量，而这种复杂性将会为大数据分析与建模带来不确定性甚至错误的结果。可见，大数据生命周期中各个阶段紧密相连、环环相扣，协作形成大数据应用解决方案。

在大数据应用开发的过程中，数据工

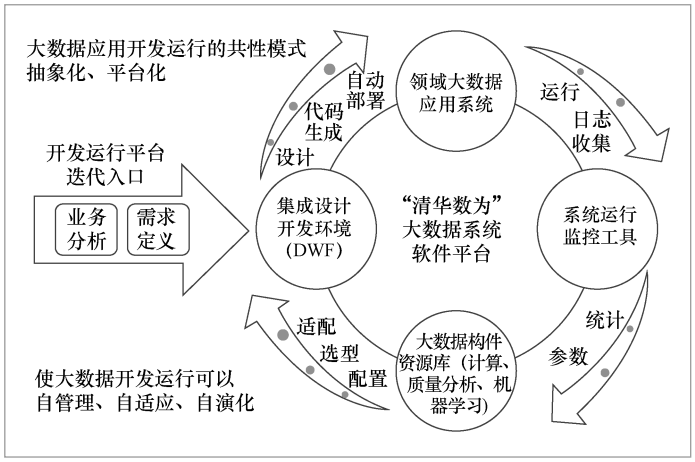


图 2　大数据应用开发的过程框架

程师需要掌握多个阶段的大数据系统软件技术，根据领域的数据类型（时间序列、流程数据、非结构化数据等）、功能需求（采集、存储、计算、分析、展现等）和性能需求（如读写时延、并发量等），在大数据软件组件中选择恰当的构件，并手动进行配置和部署，然后在此基础上进行应用开发。随着大数据系统软件的复杂性不断增加，软件发布迭代的频率也越来越高，大量的手动配置、反复试错不能满足快速交付的需求。此外，在开源大数据生态系统上进行应用开发和运行具有很高的门槛，易用性成为领域用户对大数据系统软件的迫切的需求之一。面向领域的应用还需要研制领域特定的工作环境，通过配套的软硬件资源和平台，为大数据技术提供实验、测试环境，实现大数据技术与领域知识的深度融合，为大数据技术的领域化提供有效的渠道。

为此，需要建设大数据应用的开发与运行环境，实现多样化大数据系统软件的快速、自动、智能的装配和运维，为领域专家提供“端到端”的大数据系统软件技术支持。具体说，就是领域专家只需要关注输入端的数据和需求以及输出端的结果和决策即可，两端之间的中间环节，包括数据

准备、管理、分析、可视化等，都尽可能地对领域专家隐藏。在构建大数据应用系统的过程中，领域专家无须深入了解中间环节，可将其整体视为黑盒来使用。

重点针对新型工业、国家安全、公共服务等领域大数据应用开发与运行难的问题，研制领域大数据应用开发运行环境，开发需求适配、构件选型、参数配置、自动部署、日志管理、日志分析、性能调优、故障诊断等开发运行工具，支持多样化应用需求建模与系统软件适配，自动构建领域大数据应用系统，通过日志数据的机器学习与可视化分析实现大数据应用系统智能运维和优化，提高大数据系统软件的易用性，降低大数据技术在各领域应用的门槛。

以“数为平台”为基础，大数据系统软件国家工程实验室提出了“1+X”应用架构模式，与制造、能源、农业、民生、军事等多个重点行业与多个国家级的产业优势单位合作，沉淀领域平台，构建产业应用，推动大数据技术同实体经济深度融合发展，与气象、环保、工业制造、能源等企业用户开展领域大数据平台和应用系统研发。

以气象大数据平台为例，为支撑中央和28个省的天气预报业务，需要对数值模式、卫星云图、雷达、闪电及遍布全球的地面和高空观测数据进行分析和处理。数据规模大、时效性要求高，并且持续快速增长。基于清华“数为平台”的新一代中国天气预报大数据平台突破了气象大数据分布式管理、处理、分析与可视化等一系列技术，将海量气象数据的存取速度提升百倍。在全球范围内，首次实现了全部气象数据从网络文件集群系统到大规模数据处理平台的升级换代。

在工业制造业的应用过程中，通过深入研究产业特点，基于平台的云端结合的设计思路，建设了国内价值最高的智能装备大数据平台，采集4 000余种类型的设备状态、作业操作、环境参数等实时数据；支持远程发动机运行参数调优，辅助装备远程操作；实时汇总和分析全国每台设备施工量、能耗与排放。例如：实时车辆分布及各省份排名情况；行驶、停止的车辆比例，高发故障的分布及排名情况；不同载重状态的车辆台量和占比，物流车在全国各省市的平均碳排放量和车辆的空驶情况。为生产厂商、运输企业、交管、环保、安全生产等政府部门提供实时的数据支持。物流车数据分析处理流程如**图3**所示。

5 结束语

大数据时代，软件系统与工程面临的机遇挑战体现在互为依赖的两个方面。一方面，软件系统与工程应针对大数据处理的需求，研究如何开发支持大数据处理各个环节的软件技术与系统，形成面向大数据的软件工程——面向大数据生命周期的一体化集成设计开发环境。另一方面，在软件系统与工程实施过程中，会涉及大量具有大数据特征的系统运行过程数据，因此有必要对这些多维数据进行充分的关联挖掘和机器学习，发现数据驱动的开发和运行规律，形成基于大数据的软件工程方法学，指导大数据软件系统的开发——面向软件生命周期的大数据应用系统运行分析工具。

本文分析了大数据系统软件生态、领域应用、开发运行环境现状与挑战，介绍了大规模多源异构数据一体化管理、交互式异构数据分析框架、数据可视化与智能数据工程、领域大数据应用开发运行环境、大数据混源软件可信验证等大数据系统国家工程实验室的技术研发主题，并介绍了依托工程实验室开发的大数据系统软件清华“数为平台”，该平台形成了面向大数据系统开发和应用全生命周期的技术与

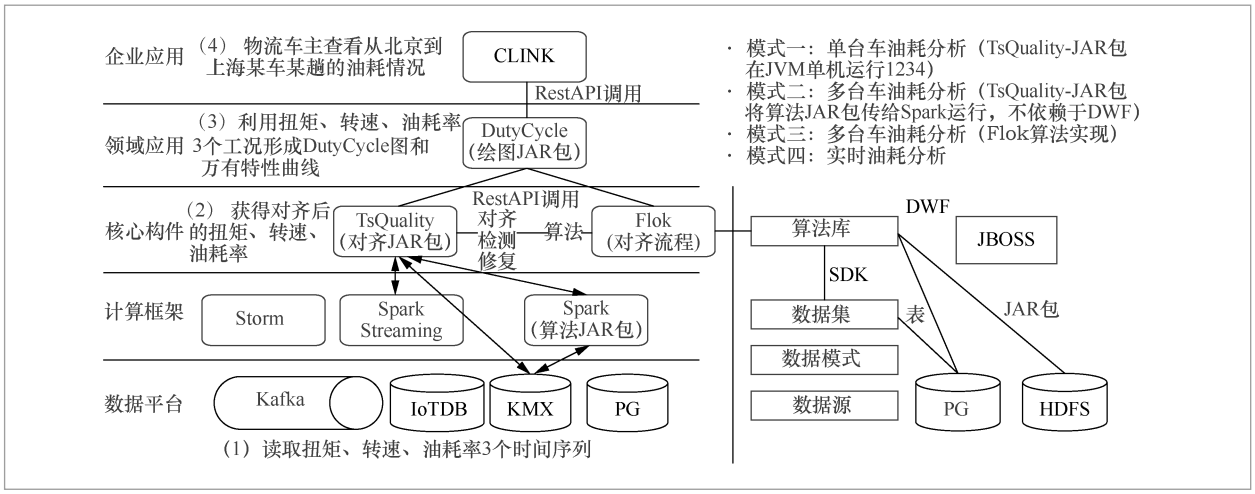


图 3　物流车数据分析处理流程

应用环路，在制造、能源、农业、民生、军事等行业已成功应用，并逐步形成大数据系统软件协同研发社区与应用创新生态环境。

参考文献:

[1] 王建民. 领域大数据应用开发与运行平台技术研究[J]. 软件学报, 2017, 28(6): 1516-1528. WANG J M. Key technologies in big data applications development and runtime support platform[J]. Journal of Software, 2017, 28(6): 1516-1528.

[2] WANG J M, JIN T, WEN L J. Process model repositories[J]. World Wide Web-Internet and Web Information Systems, 2014, 3(17): 427-454.

[3] LONG M S, WANG J M, DING G G, et al. Transfer learning with graph co-regularization[C]//The 26th AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Canada. Palo Alto: AAAI Press, 2012: 1033-1039.

[4] LONG M S, WANG J M, DING G G, et al. Adaptation regularization: a general framework for transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 5(26): 1076-1089.

[5] LONG M S, WANG J M, SUN J, et al. Domain invariant transfer kernel learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(6): 1519-1532.

[6] SONG S X, CHEN L, CHENG H. Efficient determination of distance thresholds for differential dependencies[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 9(26): 2179-2192.

[7] LIU L, ZHOU Q, LIU J L, et al. Requirements cybernetics: elicitation based on user behavioral data[J]. Journal of Systems and Software, 2016, 124(2): 187-194.

[8] LIU L, FENG L T, CAO Z Q, et al. Requirements engineering for health data analytics: challenges and possible directions[C]//The 24th IEEE International Requirements Engineering Conference, September 12-16, 2016, Beijing, China. Piscataway: IEEE Press, 2016: 266-275.

[9] LIU L, FENG L T, LI Y Q, et al. Flourishing creativity in software development via Internetware paradigm[J]. Science China: Information Science, 2016, 59(8): 080103.

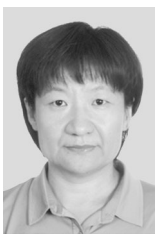
[10] YASIN A, LIU L, CAO Z Q, et al. Big data services requirements analysis[C]//Asia-Pacific Requirements Engineering Symposium 2017, November 9-10, 2017, Melaka, Malaysia. Heidelberg: Springer, 2017.

作者简介

王建民(1968-),男,清华大学软件学院教授、院长,数据科学研究院副院长,大数据系统软件国家工程实验室执行主任,2014年获国家杰出青年科学基金,国家“万人计划”科技创新领军人才,国家科技部中青年科技领军人才,入选清华大学研究生良师益友名人堂。主要从事大数据与知识工程、过程与行为数据分析、非结构化数据管理技术、产品生命周期管理技术领域的研究工作。2008年以来,发表学术论文120余篇,获得中国发明专利授权19项。1998年获国家科技进步奖二等奖,2012年获教育部科技进步奖一等奖,2014年获国家科技进步奖二等奖。国家重大科技专项“核高基”基础软件方向实施专家组成员,国家“十二五”“863”计划先进制造技术领域制造服务技术主题专家组成员,国家“十二五”“863”计划先进制造技术领域“面向制造业的核心软件开发”重大项目总体专家组组长。中国计算机学会数据库专家委员会和大数据专家委员会委员,《计算机集成制造系统》杂志主编。

王晨(1983-),男,大数据系统软件国家工程实验室、清华大学数据科学研究院工业大数据中心总工程师,四川省工业大数据创新中心总经理,北京市工业大数据创新中心总工程师,昆仑数据联合创始人,《中国制造2025》路线图(操作系统与工业软件)编写组成员。参与多项国家级、部级信息化与大数据规划指南撰写,领导实施多个制造业龙头企业工业大数据项目。曾担任IBM中国研究院主任研究员,数据管理技术研究部高级经理,IBM全球分析云研究战略负责人。在数据领域顶尖国际会议与期刊上发表20余篇论文,拥有60余项中国、美国专利(含已公开),担任多个学术会议审稿人,担任《中国计算机学会2015年中国大数据技术与产业发展报告》《工业大数据技术与应用白皮书》等多本著作编委。中国计算机学会数据库专家委员会委员、产学研工作组副组长,全国信息技术标准化技术委员会大数据标准工作组工业大数据专题组组长,广东省工业互联网专家委员会、北京市两化融合专家委员会等多地咨询专家,工业和信息化部工业强基、工业互联网等专项评审专家。

刘英博(1978-),男,清华大学软件学院副研究员,主要从事工业软件、软件工程、工业大数据平台技术、工业大数据分析方面的研究工作。2001年获得中国人民解放军国防科技大学计算机专业学士学位,2008年获得清华大学计算机软件与理论专业博士学位。博士在读期间,作为核心研发人员开发清软英泰信息技术有限公司产品全生命周期管理系统(TiPLM),获得教育部科技进步奖二等奖(第23完成人)。博士毕业后成功完成了东方电气集团东方汽轮机有限公司“5·12”大地震后PLM信息系统重建工作。作为技术负责人主持开发了复杂装备维护、维修与大修(MRO)支持平台,并在煤矿、化工、军工等行业的企业得到应用,获得教育部科技进步奖一等奖(第4完成人)。作为大数据技术负责人与三一重工股份有限公司合作完成工程机械工况数据底层管理系统的切换,实现了跨产品生命周期阶段的数据集成和大规模工程机械物联网应用,获得中国电子学会科技进步奖一等奖(第5完成人)和国家科技进步奖二等奖(第5完成人)。



刘璘(1973-),女,清华大学软件学院副研究员,主要从事软件需求工程、知识工程、医疗数据分析技术方面的研究工作。1993年毕业于吉林大学计算机专业。1999年在中国科学院数学与系统科学研究院获博士学位。2000—2004年在加拿大多伦多大学做博士后研究。现任中国计算机学会软件工程专业委员会委员,《Requirements Engineering Journal》《软件学报》编委,《International Journal of Information Systems Modelling and Design》副主编。已发表软件需求工程、信息系统工程、服务工程领域学术论文百余篇,谷歌学术他引千余次。完成国家自然科学基金、西门子海外合作、“863”计划、科技支撑项目10余项,担任国际软件与应用、需求工程,信息系统领域国际会议程序委员30余次,为多家国内外期刊担任特邀编委和审稿人。

收稿日期: 2018-08-03

大数据系统软件国家工程实验室简介

大数据系统软件国家工程实验室以创新驱动发展和自主安全可控为使命，以国家“十三五”规划为指导，面向重大领域实际需求，建设大数据系统软件技术创新研发、技术转化和应用示范平台，在大规模多源异构数据一体化管理、交互式异构数据分析框架、数据可视化与智能数据工程、领域大数据应用开发运行环境、大数据混源软件可信验证等技术研究方面进行理论和应用创新，研发一系列具有国际先进、国内领先的大数据系统软件核心技术与产品，突破创新驱动产业升级转型以及关键领域自主可控中的技术瓶颈，探索大数据系统软件应用实施工程方法，建设国际一流、开放协同的产学研用大数据系统软件技术创新团队，增强对国家重大战略方向、重点工程的大数据应用的保障能力。

大数据系统软件国家工程实验室法人单位是清华大学，参建单位为北京理工大学，共建单位共7家，分别是中国人民解放军国防科技大学（以下简称国防科技大学）、中山大学、百度公司、腾讯公司、北京大学、中国人民大学、阿里云，另外还有北京工业大数据创新中心、北京市遥感信息研究所、国家气象中心（中央气象台）、中国农业科学院农业环境与可持续发展研究所、公安部第一研究所、中国石油规划总院、公安部沈阳消防研究所、环境保护部信息中心等多家项目合作单位，在此重点介绍各建设单位。

1 清华大学

2014年4月，清华大学成立了清华数据科学研究院，对大数据人才培养、科学研究和社会服务进行了系统布局，标志着清华大学数据科学与工程学科建设工作全面启动。本工程实验室建设任务依托清华大学软件学院，协同清华数据科学研究院和信息学院大数据系统软件相关学科团队联合承担。清华大学软件学院孙家广院士及其团队在大数据系统软件技术开发与工程应用方面具有长期积累，具备大数据系统软件大规模平台运维、新型计算硬件研制与领域大数据应用基础，构成了大数据系统软件创新完整技术链，与国际大数据开源社区建立了紧密合作关系。“八五”初期即开展工程数据库技术与系统研发工作，“十二五”期间牵头完成了国家“核高基”重大专项《非结构化数据管理系统》课题，其成果LaUDMS软

件系统已经在国家气象中心、三一重工股份有限公司、金风科技股份有限公司等行业大数据领域得到实际应用。其中，在国家气象中心实施的气象大数据管理系统已成为每日中央电视台天气预报节目的后台支撑系统，被工业和信息化部选入国家“十二五”科技创新成就展；三一重工工业大数据应用中实现对Oracle的成功替换，满足全球15万辆工程机械数据实时采集与分析需求，2016年初作为北京市“互联网+制造”典型成果专报国务院，并入选2015年全球大数据峰会产业案例（中国唯一），相关成果获得2014年国家科技进步奖二等奖。

2 北京理工大学

北京理工大学在计算机相关学科拥有4个省部级重点学科，建有1个国家工程实验室、1个国家级虚拟仿真实验教学中心、4个省部级重点实验室、1个国防科技工程中心。在大数据与云计算方面，建有北京市海量语言信息处理与云计算应用工程技术研究中心，北京理工大学—IBM主机卓越中心（北京）、北京理工大学—IBM大数据分析技术中心、北京理工大学—中国轻工业联合会数据智能应用联合实验室、数据科学研究所。承担了一系列国家级科研项目，包括国家自然科学基金重点项目、国家重点研发计划项目、国家“973”计划课题、国家“863”计划课题、国家科技支撑计划、国防重点型号项目等。在大数据软件开发与运行环境方面，牵头承担了国家自然科学基金重大项目“基于互联网群体智能的软件开发方法研究”，与北京大学等单位联合承担了国家重点研发计划项目“基于大数据的软件智能开发方法和环境”，为大数据应用开发运行环境的研究积累了大量的研究成果，获得部级科技进步奖8项。

3 国防科技大学

国防科技大学计算机学院拥有计算机科学与技术领域的国防科技重点实验室、教育部工程研究中心和3个国家级重点学科，现有院士4名。近年来，在高性能计算领域形成了明显的科研优势，并进入国际领先行列。先后研制成功了20多种以天河超级计算机系统、银河系列巨型机、麒麟操作系统、微处理器、银河玉衡9108核心路由器等为代表的自主知识产权产品。在大数据方面，计算机学院基于天河二号超级计算机，设计实现了基于TCP/IP的高带宽的通信机制，在很大程度上提高了大规模二层网络环境下的网络整体性能。在可信研究方面，计算机学院长期从事可信软件和混源软件开发基础理论研究和国产基础软件的研发工作，承担多项国家自然科学基金重点项目、国家“973”计划项目、国家“863”计划项目、武器装备预先研究项目、“核高基”项目等，获国家科技进步奖二等奖1项，军队科技进步奖二等奖2项。

4 中山大学

中山大学数据科学与计算机学院现有国家工程技术研究中心1个，国家地方联合工程实验室1个，教育部重点实验室1个，教育部工程研究中心2个，广东省重点实验室3个，广东省工程技术研究中心2个。在分布式系统的通信与计算领域，提出了软硬协同的聚合通信自动触发机制，实现了聚合通信在互连网络上的数

据传输过程自动化，相关工作获国家科技进步奖一等奖1项，军队科技进步奖一等奖1项、二等奖1项以及湖南科技进步奖二等奖1项。在数据存储与管理方面，突破了全局共享并行分布式文件系统设计与实现技术，实现了基于电磁混合的多层次大规模全局并行文件系统，相关成果获军队科技进步奖一等奖1项（排名第一）。近5年共承担纵向科研项目293项，国家级项目92项，其中国家“863”计划项目1项，国家“973”计划项目1项，国家重点研发计划项目2项，国家重点研发计划课题3项。在SCI收录的杂志发表论文近500篇，其中中国科学院二区以上论文142篇，授权国家专利73项。

5 百度公司

百度公司拥有海量数据和完整、领先的大数据技术，通过平台化方式开放大数据核心能力，为传统产业提供大数据应用基础，并已在多个领域尝试开展大数据应用服务。在数据资源基础方面，百度公司已经成为中国首家拥有并运营EB级数据的公司，百度公司现阶段大规模存储系统单机群规模为15万台服务器，单机群存储容量为2 EB。大规模存储系统中在文件子系统的日读取能力为166 PB。在大数据技术方面，特别在以人工智能、深度学习为代表的大数据智能领域，已经取得了一些国际领先成果，例如百度的机器翻译技术，实现了机器翻译的真正大规模产业化应用。百度公司积极布局大数据、云计算、移动互联网等领域的核心技术和关键应用，在国家科研项目方面，承担了国家发展和改革委员会云计算重大专项、科学技术部“863”计划重点项目、“核高基”等30多项国家重大专项的科研工作，多项科技成果获得国家科技进步奖二等奖、中国电子学会科技进步奖一等奖等。

6 腾讯公司

腾讯公司在大数据上有着深厚积累和卓越表现，包括大数据量、大数据处理能力、数据挖掘分析能力、数据应用能力等。腾讯公司拥有业界覆盖最广、维度最宽、实时性最强的社交领域大数据，基于QQ、微信、视频、新闻等海量业务数据，在海量用户行为数据的采集、接入、存储、计算等方面，腾讯公司做到了高效的实时采集计算、快速的多种接入、海量最复杂的数据仓库、专业的结果展现、方便的消息推送。腾讯公司有着业界一流的大数据分析挖掘能力，并以此保证了多种数据应用的成功。腾讯公司招收各行各业中顶尖的数据分析挖掘人才，同时进行内部培养，互相促进，建立了国内第一个专业社交分析挖掘数据团队；其次是通过日常工作不断积累并形成一整套分析挖掘模型宝库，通过模型库中各种模型的使用，逐步建立了业界领先的数据分析和挖掘专业能力。

7 北京大学

北京大学于1955年建立计算数学专业，培养了我国第一批程序设计专业人员。1984年获得计算机软件博士学位授予权，1985年获计算机科学理论博士学位授予权，1996年成为国内首批获得计算机科学与技术一级学科博士学位授予权的单位之一。北京大学软件研究所已形成以中国科学院院士杨芙清教授、梅宏教授为学术带头人的结构合理稳定、学术水

平高的学术团队。在面向互联网的新型软件理论研究方向形成特色，针对云计算、大数据、移动互联网、物联网等新型应用模式的基础软件技术取得了丰富的研究成果，近年来在TSE、TOSEM、TOIS、TMC、TSC、POPL、ICSE、ASE、FSE、OOPSLA、WWW、UbiComp、IMC等国内外重要期刊、重要会议和刊物上发表论文200余篇；共获17项科研成果奖，其中牵头获国家技术发明奖二等奖1项、国家科技进步奖二等奖2项、国家自然科学二等奖2项。

8 中国人民大学

中国人民大学大数据研究以数据工程与知识工程教育部重点实验室、大数据管理与分析方法研究北京市重点实验室两个省部级重点实验室为载体，承担和参与过和大数据紧密相关的国家重点科技攻关项目、“核高基”项目、“973”计划基础研究项目、“863”计划高技术项目、国家自然科学基金项目及其他科技开发项目等数十项。科研成果曾获得国家科技进步奖二等奖、教育部科技进步奖一等奖、电子部科技进步奖特等奖、北京市科技进步奖二等奖（2次）、教育部科技进步奖二等奖、国家级优秀教材奖（2次）等多项奖励。在21世纪初，该研究团队的研究成果成功地孵化出了人大金仓国产数据库公司。

9 阿里云

阿里云自主研发了飞天大规模分布式计算系统Apsara，它是阿里云操作系统的核心组件，负责管理数据中心Linux集群的物理资源，控制分布式程序的运行，隐藏下层故障恢复和数据冗余等细节，有效地提供了弹性计算和负载均衡等服务。基于飞天大规模分布式计算系统构建的云计算基础平台，提供了弹性计算服务（elastic compute service，ECS）、负载均衡（server load balancer，SLB）服务、弹性伸缩服务（elastic scaling service，ESS）、云数据库服务（relational database service，RDS）、开放结构化数据服务（open table service，OTS）、开放缓存服务（open cache service，OCS）、阿里云对象存储服务（object storage service，OSS）、云盾安全服务、分布式拒绝服务（distributed denial of service，DDoS）防护服务、开放式数据处理服务（open data processing service，ODPS）、实时分析数据库服务（analysis database service，ADS）等。 □

大数据

BIG DATA RESEARCH

www.j-bigdataresearch.com.cn

ISSN 2096-027
CN 10-1321/G2

双月刊 • 单月15日出版 2019年共6期 • 35元/期 • 全年价210元 邮发代号：2-537

编辑部电话：010-81055490 广告部电话：010-81055448

邮局订阅

凭邮发代号，在全国各地邮局（所）订阅

发行部订阅

将订阅信息以传真或邮件方式发送至发行部，也可致电发行部提交订阅信息

电话：010-81055597

传真：010-81055464

E-mail: yrq@bjxintong.com.cn

银行汇款

户名：北京信通传媒有限责任公司

开户行：中国工商银行北京市体育馆路支行

账号：0200008109200044661

邮局汇款

北京市丰台区成寿寺路11号邮电出版大厦8层（100078）

北京信通传媒有限责任公司发行部

主管单位：中华人民共和国工业和信息化部
主办单位：人民邮电出版社

学术指导：中国计算机学会大数据专家委员会
出版单位：北京信通传媒有限责任公司

ISSN 2096-0271

邮发代号：2-537 国外代号：C9118 定价：35.00元