大数据系统计算技术展望

王海涛[1,2]，毛睿[1,2]，明仲[1,2]
1. 深圳大学计算机与软件学院，广东 深圳 518060；2. 大数据系统计算技术国家工程实验室，广东 深圳 518060

摘要
针对大数据系统计算技术现状与问题，系统地介绍了大数据系统计算技术国家工程实验室的大数据系统计算技术创新平台。创新平台以研制自主可控、通用性强的大数据计算平台为核心目标，解决高性能大数据计算中的共性问题，开展大数据示范性应用研发。创新平台的建设将提升国家大数据创新能力，满足我国大数据技术与应用重大需求，促进我国大数据技术服务产业发展壮大，助力我国大数据“双创”生态环境成熟完善。

关键词
大数据系统计算平台；大数据共性技术；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.2096-0271.2018022

Perspective on big data system computing technology

WANG Haitao[1,2], MAO Rui[1,2], MING Zhong[1,2]
1. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
2. National Engineering Laboratory for Big Data System Computing Technology, Shenzhen 518060, China

Abstract
The state of the art of big data system computing technology was elaborated and the big data system computing innovation platform at the National Engineering Laboratory for Big Data System Computing Technology was introduced. The platform strives to develop China-made general-purpose big data computing system, to conquer common problems in high performance big data computing, and to develop applications. The construction of the platform will boost national big data innovation capability, fulfill the demand of big data technology and application in China, promote the big data industry, and help the construction of big data “double-innovation” eco-system.

Key words
big data system computing platform, big data common technology, big data application, National Engineering Laboratory

1 引言

大数据是新一代信息技术的核心方面和竞争前沿，也是制约大数据产业快速发展的关键瓶颈。大数据技术创新能力已经成为后信息时代衡量国家竞争力的重要指标。与传统信息产业的发展过程相似，大数据必将逐渐形成一个相对独立、体系完善的产业形态，完成传统信息产业的升级换代。互联网和云计算的发展过程与趋势已经证明，大数据未来的产业形态将是以服务为核心的新型产业形态，大数据产业体系的各个环节将提供极为丰富的服务。

大数据是国家、社会和产业在后信息时代的战略性资源，以大数据为核心支撑的新一代信息技术与应用（如互联网+、物联网、智慧城市、智能制造等）利用大数据资源的手段和工具，为社会提供信息服务，其最终目的是利用大数据解决科学研究、社会管理、产业发展等一系列实际问题，从而在战略决策、运营管理、终端服务等不同层面和环节提升效能与效益，形成新的核心竞争力。当前，全社会数据产生越来越快、积累越来越多，大数据资源越来越丰富，而现有的信息技术已经跟不上数据的发展，特别是对大数据的处理、分析与应用已经成为全球性问题，引起了各国政府和产业界的高度重视。

大规模且高复杂性的大数据，其处理时间、响应速度等都有明确且具体的要求，这对计算平台的架构、计算模型的框架、共性技术等提出了更高的要求。传统的以计算速度为优先的设计理念已经不能满足当前大数据时代的处理需求，新计算平台的研发、框架设计和共性技术开发等需要兼顾效率与效能的双重标准，同时兼顾大数据类型多、变化快、价值稀疏的特性。

2 大数据系统计算技术现状与问题

大数据计算平台是大数据的硬件与系统基础，对大数据的所有分析与处理都需要在高性能的计算平台上进行；共性技术是大数据分析与处理的知识与技术基础，所有的大数据系统都涉及数据采集、传输、存储、处理和分析过程中的多项共性的技术；典型的应用可以用来验证计算平台和共性技术的可行性与执行效率，并为相近应用的研发提供借鉴。

经过近几年的快速发展，大数据已经形成从数据采集、数据处理到数据分析的完整产业，为社会经济的发展提供有力的数据支持。然而技术的发展赶不上数据数量和形式的变化速度，这需要大数据研发和从业人员不断努力来适应新情况。下面从计算平台、共性技术和应用3个方面对大数据的技术现状及存在的问题进行阐述。

2.1 大数据计算平台现状及存在的问题

大数据计算平台是大数据产品体系的核心龙头。大数据计算平台以云计算为基础环境、以服务模式为总体架构，覆盖大数据应用全过程，支持多源异构海量数据的采集、存储、集成、处理、分析、可视化展现、交互式应用，涉及企业大数据产品体系的各个层面，为各层产品实现提供关键技术支撑。

当前，面向信息服务的大数据计算平台还远不成熟，平台优化和节能等关键技术有待突破，面临的问题主要包括以下几个方面。

- 平台研发需要交叉学科知识。由于大数据计算平台环境复杂，要提高大数据计算平台的处理能力、优化平台性能和降低能耗、提高平台的安全性和隐私保护，

需要多学科、多种专业背景的科研人员协同创新、联合攻关，才能实现技术突破。平台研发需要交叉学科知识。

● 平台研发人才缺乏。由于大数据应用在许多行业属于新生事物，虽然很多企业管理者认识到了大数据的潜在价值，但还处于观望状态，不愿冒风险做领先者；多数企业缺少大数据技术和应用人才，也缺少大数据应用的技术平台，无法开展实际应用项目；目前大数据技术还很不成熟，很多是开源软件，实施应用项目成本高、风险大。企业、科研机构和大学协同创新是突破行业示范应用障碍的有效途径。

● 平台框架需要突破。大数据计算平台的处理和分析能力主要依赖分布式计算框架来完成。分布式计算框架不仅要提供高效的计算模型、简单的编程接口，还要具备容错能力和高扩展性。大数据计算平台的计算框架主要分为批处理和流处理两种。批处理计算框架主要针对静态数据的离线计算，吞吐量好，但是不能保证实时性；流处理计算框架主要针对动态数据的在线实时计算，时效性好，但是难以获取数据全貌。

● 技术环境不统一。目前大数据最前沿的计算框架和衍生的生态圈都分布在以GitHub为主的开源社区内，形成了大数据基础架构和大数据分析处理技术群。大数据的技术环境为用户提供了丰富的技术选择，但也给用户掌握和使用多样性的大数据技术造成障碍，且提高了应用成本。

2.2 大数据共性技术现状及存在的问题

根据处理流程，大数据技术可以分为基础架构支持、数据采集、数据存储、数据计算以及展现与交互几类。具体来说，基础架构涉及的技术包括云计算平台、云存储、虚拟化、网络以及资源监控等；数据采集涉及的技术包括数据总线、抽取—转换—加载（extract-transform-load，ETL）工具等；数据存储涉及的技术包括分布式文件系统、关系型数据库、NoSQL数据库、内存数据库等；数据计算涉及的技术包括数据查询、统计与分析、数据预测与挖掘、图谱挖掘、商业智能（business intelligence，BI）等；展现与交互技术包括图形与报表、可视化工具、增强现实技术等。此外，大数据技术还包括大数据的安全与隐私保护等相关技术。

当前，大数据的存储、处理、分析、展示等关键技术有待突破，大数据处理和分析能力、算法的数据处理能力远远达不到实际应用要求。现有的信息技术不足以解决大数据的问题，面对丰富和复杂的大数据应用场景，需要多学科、多领域的交叉合作才能解决。这给传统信息技术带来了挑战，主要表现在：多源多模大数据的获取和融合；对数据质量和低质量数据的容忍；面向大数据算法及理论的进一步研究；对现有可扩展存储和处理的计算架构更新设计；对大数据并行处理及分析挖掘技术适应性设计等。

2.3 大数据应用的现状及存在的问题

随着国家政策扶持力度的不断加大，产业资本的持续投入，国内外发展大数据的积极性都很高，行业应用得到快速推广。当前，大数据的应用给企业不仅带来了技术和应用模式上的突破，还为商业模式的创新以及企业的转型发展带来了驱动力。对公共服务机构来说，挖掘大数据的潜在价值对解决城市发展问题、完善社会机制、更好地服务于市民有着重大的意义。

虽然大数据应用逐步深入人们的生活，但限于目前大数据技术仍不成熟，因此与实际问题解决需求存在一定的差异，主

要表现如下。

● 时空大数据面临海量数据管理、时空数据融合、时空大数据理论框架和时空推理与数据挖掘的深度结合的挑战；针对时空大数据模式发现与价值提炼的问题，揭示大尺度事件的演化推理机理，突破大尺度安全时间的理解和局限的挑战。

● 媒体大数据的包容量大、源头多且成分迥异的特性，给媒体大数据计算理论和关联算法的研究带来了巨大的挑战。此外，目前无论是科学研究还是新技术开发，都没有解决计算机按内容自动搜索视频媒体的可行性方案。

● 为了有效地利用物联网海量多样化的大数据，挖掘其价值，一方面需要对网络进行优化，使其具备未来物联网需具备的承载大数据的能力；另一方面，对用户的服务进行分析挖掘，充分实现物联网的潜在价值。

● 物流大数据应用领域面临使用大数据技术推动电子商务与现代物流的协同发展问题，是物流大数据应用的一大挑战。

3 大数据系统计算技术创新平台

在国际社会和我国政府将大数据上升为国家战略的背景下、在我国大数据计算平台和共性技术研究不充分的情况下，国家发展和改革委员会于2017年正式批复由深圳大学牵头建设大数据系统计算技术国家工程实验室（以下简称国家工程实验室）。国家工程实验室面向国家实施大数据战略的需求，着重解决我国大数据计算平台安全可信性差、软件通用性低的突出问题；采用国产高性能处理器互联芯片和计算平台，实现大数据智能诊断和自动纠错等可靠性信息高速转发，提升大数据计算平台的安全控制能力，满足国家加密算法的要求；以研制自主可控、通用性强的大数据计算平台为核心目标，解决高性能大数据计算中的共性问题，并以此为基础，着力开展大数据示范性应用研发。各技术方向的架构和关系如**图1**所示。

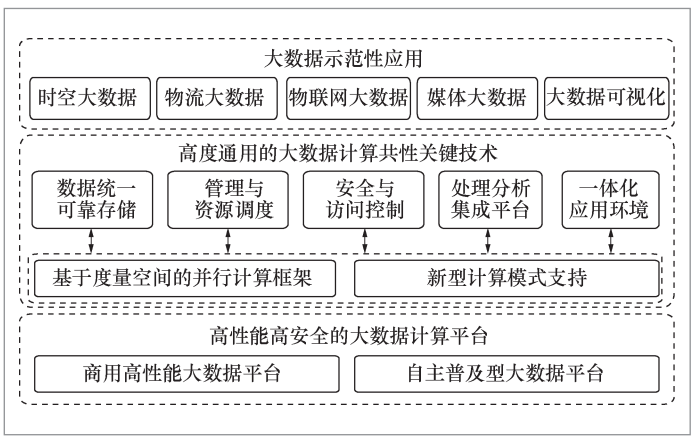


图1 国家工程实验室研究方向

国家工程实验室计划突破的3个主要技术方向包括高性能高安全的大数据计算平台研发、高度通用的大数据计算共性关键技术研究、大数据示范性应用研发，有针对性地解决目前我国大数据计算平台面临的安全性差、通用性低等突出问题。采用全国产化、面向特定工程领域的芯片及软件技术，实现完全自主可控的高性能大数据计算平台。将在该自主可控的硬件计算平台上，研发可以适用于大数据计算共性关键技术的通用系统软件。还将面向大数据计算中的示范性应用，开展应用软件技术的研发，实现“硬件—系统软件—应用软件”的完整的大数据计算生态环境。

（1）高性能高安全的大数据计算平台研发

随着国家、国防信息化建设的发展，构建自主可控的大数据计算平台是国家战略安全的至关重要的问题。目前我国绝大多数的大数据计算平台都基于国外的软硬件系统。大数据计算平台安全可

信性差和软件通用性低等问题，成为大数据行业最为突出的核心问题。采用国外的基础技术平台，不仅在核心技术上受制于人，而且对我国重要数据信息的安全构成了直接威胁。国家工程实验室将从“硬件—系统软件—应用”3个层次，全面构建国产化高可靠性、高性能大数据平台。该平台的建设以研制基于国产化处理器的大数据计算硬件系统为核心方向。采用创新的RAS 2.0（增强的可靠性（reliability）、可用性（availability）及可服务性（serviceability））和处理器互联技术，提供超越现有标准开放平台的单机可靠性和计算性能。采用基于固件、不依赖操作系统（operatin system，OS）的故障响应和诊断机制，能够全面、自动收集故障信息并进行分析，实现主动预警、隔离、替换失效部件，并率先支持CPU、内存等核心部件的在线维护，面向计划内零停机维护设计，保障重要部门和企业的关键业务连续运行。

（2）高度通用的大数据计算共性关键技术研究

国家工程实验室将开展面向高度通用的大数据计算共性关键技术的研究。大数据计算在数据量、数据类型方面与传统高性能计算有显著不同，对数据的实时响应、安全性等方面也有特殊的要求。国家工程实验室拟在通用大数据并行计算框架、一体化平台集成、数据存储、资源调度、数据安全、一体化应用环境6个方面开展研究。基于度量空间的大数据抽象框架，利用度量空间将大数据复杂繁多的数据类型和多样化的距离抽象为统一的数据类型和距离函数接口，并通过支撑点选取算法选取若干支撑点，以数据到支撑点的距离为坐标，从而将数据从无坐标的度量空间映射到有坐标的多维空间，使众多数学工具得以使用。

（3）大数据示范性应用研发

研发特定领域的大数据示范性应用及其共性技术是国家工程实验室建设的重要目标之一。国家工程实验室建设期间将从“智慧城市”“智能制造”“智慧政府”及“智慧服务”等领域的关键应用技术出发，开展五大示范性应用及共性技术的开发，包括多维时空大数据分析在智慧城市中的示范应用、物流大数据示范应用、媒体大数据分析的示范应用、物联网大数据示范应用以及大数据可视化分析技术的示范应用。通过示范应用的开发及共性技术的研究，建立大数据“产学研用”协同创新生态链，促进大众创业和万众创新。

4 基于大数据系统计算技术平台打造大数据生态创新基地

国家工程实验室的总体功能定位概括为我国大数据领域领先的“4个基地”，分别介绍如下。

（1）大数据技术创新与计算平台基地

国家工程实验室特别注重大数据计算共性关键技术的源头创新能力，大数据计算平台的基础支撑能力以及对前沿技术成果和重大项目方案的集成、测试、验证的能力，保障产业界对先进、成熟的技术的应用需求。

（2）人才培养与交流基地

国家工程实验室以重大需求引领技术创新和多学科交叉为人才培养特色，以校企联合及国际化为人才培养途径，着力培养兼具理论探索和实践开发的复合型人才；同时，通过与中国港澳地区及欧美知名高校开展联合研究、特色专业、访问学者等活动，实现国际高端人才的常态化交流与合作，发展成为区域性大数据高端人才培养与国际交流基地。

(3)产业孵化与服务基地

国家工程实验室针对深圳、香港及珠三角地区新一代信息技术领域创客群体对大数据技术与服务的共性需求，提供大数据创新创业的技术、平台、人才等公共技术服务支撑，促进创业项目加速成熟。

(4)大数据协同创新基地

国家工程实验室广泛对接政府和社会资源，成为以政府开放数据为源头、以社会应用需求为引领、以大数据平台为核心、以广东及香港高校人才团队为支撑、以社会资本为纽带、以龙头企业和产业园区为出口的大数据上下游链条协同合作与发展的枢纽，共同建设良好的融合大数据领域“政产学研资介用”多元创新资源的产业生态圈。

国家工程实验室以高校为建设主体，实现大数据领域的共性核心技术创新、核心平台建设、高端人才培养、公共技术服务，符合大数据领域技术驱动的发展要求。同时，高校作为国家公立科研机构，能够从根本上保障协同创新的公立性和持续性。

5 大数据系统计算技术“产学研”合作及行业应用示范平台

为有效地支撑大数据产业发展与应用推广，充分发挥国家工程实验室建设成果对各行业大数据应用的基础共性需求的支撑能力，实验室将建设大数据“产学研”合作及行业应用示范平台，平台提供的核心子系统与功能包括以下几个方面。

(1)数据资源目录系统与服务

平台将提供数据资源目录服务。一方面，将已经聚集的数据资源处理成标准数据集，授权第三方合作伙伴使用；另一方面，接收外部数据资源汇聚，将所用数据资源经过处理后的标准数据集以数据资源目录的形式提供给授权用户访问、查询、分级使用，同时提供基于数据资源开展项目的清单，便于数据资源价值理解与横向对比，促进产业生态发展。

(2)行业大数据应用案例库及服务

为促进行业大数据应用普及推广，不断强化行业大数据应用的标准化与产品化发展，平台提供行业大数据应用案例库及相关服务，将依托平台构建的各个行业大数据应用案例进行统一管理，并对外授权服务。同时，借鉴开源软件管理与发展模式，任何基于本平台提供的案例进行定制开发的案例有义务提交到本平台，明确不同案例版本的所有权，促进合作方之间的共赢发展。

(3)资源管理与共享服务

为有效支撑不同用户及案例的开发与运行，平台将提供资源管理与共享服务，进行用户及案例管理，平台将根据用户案例的开发与运行所需资源进行动态分配与管控，实现在IaaS层、PaaS层的平台资源共享。

6 强化国家大数据创新能力

大数据系统计算技术创新平台处在国际大数据竞相发展的起步阶段和国际大数据的技术体系与产业格局尚未完善成熟的初级阶段。瞄准国际大数据技术与产业发展趋势，针对国家大数据发展战略需求，依托区域大数据创新创业优越环境，发挥建设单位的创新成果基础与互补资源优势，集中围绕大数据处理与分析的主题，以建设大数据计算平台与大数据共性技术为核心任务，以研发和促进大数据重大行业应用示范为导向，以构建

多学科、多领域交叉融合的创新载体为手段，全面系统地突破大数据领域的技术与应用问题，通过自主创新的大数据成果，支持和驱动相关产业转型升级和双创战略深化实现。

首先，创新平台将强化国家大数据创新能力建设。大数据是新一代信息技术的核心方面和竞争前沿，也是制约大数据产业快速发展的关键瓶颈。大数据技术创新能力已经成为后信息时代衡量国家竞争力的重要指标。与传统信息技术创新相比，大数据技术创新对科技人才、科研载体、合作模式、资源要素等提出了更高的要求，迫切需要能够有效组织、融合、协同、发展多元化创新主体和资源的新型、高水平的大数据创新载体，承担起满足国家大数据需求的创新、组织、交流、合作的职责。通过本项目研发出自主可控的大数据计算平台和一系列大数据共性技术，将成为国家大数据创新体系的重要组成部分和推进抓手。

其次，创新平台将满足我国大数据技术与应用的重大需求。大数据系统技术按技术创新平台建设的直接产出就是系统全面地提供大数据技术成果，具体包括时空大数据、物流大数据、媒体大数据、物联网大数据等示范性大数据应用的相关机器学习理论、分析模型与算法以及应用相关工具、系统与平台、行业大数据应用解决方案等。这些技术成果具有极强的适用性和推广性，可进一步面向具体应用场景与需求进行第三方定制开发，应用到更广泛的行业、领域、环节、场景，突破制约我国大数据应用发展的瓶颈。

再次，创新平台将促进我国大数据技术服务产业发展壮大。与传统信息产业的发展过程相似，大数据必将逐渐形成一个相对独立、体系完善的产业形态，成为传统信息产业的升级换代。互联网和云计算的发展过程与趋势已经证明，大数据未来的产业形态将是以服务为核心的新型产业形态，大数据产业体系的各个环节将形成极为丰富的服务。建设从计算平台、共性技术到示范应用，全面覆盖大数据技术体系中主要环节的核心技术创新与突破，相应的技术成果将以技术、专利、软件、系统、平台、方案、培训、咨询等形式为社会提供服务，促进我国大数据技术服务产业发展壮大。

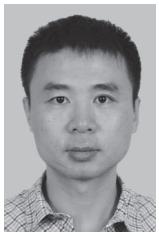
最后，创新平台将助力我国大数据“双创”生态环境成熟完善。当前我国“双创”战略激发了众多创客团队和创业孵化器的快速发展，大量与信息技术和服务相关的创业项目，特别是在互联网、移动互联网、物联网、智能硬件等领域的创业项目，都离不开大数据技术的支持。大数据系统计算技术创新平台研制大数据计算平台，研究大数据共性技术，具备开展大数据应用的全部核心功能与组件，并通过服务的方式一站式提供给用户，因此将成为“双创”战略的重要支撑，降低创客群体使用大数据技术的门槛，有助于创客群体之间的交流合作，形成以平台为核心的大数据相关创新创业生态圈。

7 结束语

本文从大数据计算平台、共性技术和大数据应用的现状及存在的问题出发，介绍了大数据系统计算技术国家工程实验室的大数据系统计算技术创新平台。创新平台以研制高性能高安全的计算平台、高度通用的大数据共性技术为基础，进行大数据示范应用的研发。依托创新平台打造大数据生态创新基地，完善大数据“产学研”一体的综合生态圈；建设大数据“产学研”合作及行业应用示范平台，改善大数据人

才培养的环境，促进大数据研究领域科研成果的转化。创新平台的建设将提升国家大数据创新能力，满足我国大数据技术与应用的重大需求，促进我国大数据技术服务产业发展壮大，助力我国大数据“双创”生态环境的成熟完善。

作者简介



王海涛（1977-），男，大数据系统计算技术国家工程实验室办公室主任、深圳大学计算机与软件学院助理研究员，主要研究方向为数据挖掘与大数据计算平台。

毛睿（1975-），男，博士，大数据系统计算技术国家工程实验室教授、副主任，主要研究方向为通用大数据管理分析方法和高性能计算。1997年和2000年在中国科学技术大学分别获得计算机科学与技术学院学士和硕士学位，2006年和2007年在美国德克萨斯大学奥斯汀分校分别获统计学硕士和计算机科学博士学位。2007—2010年在甲骨文美国公司从事数据库研发工作；2010年加入深圳大学，现任深圳大学计算机与软件学院副院长、广东省国产高性能数据计算系统工程技术研究中心主任及广东省普及型高性能计算机重点实验室常务副主任；中国计算机学会高级会员、大数据专家委员会和数据库专业委员会常务委员、广东省计算机学会云计算专业委员会主任、CCF YOCSEF（深圳）主席（2016—2017年）、2016年中国数据库学术会议程序委员会共同主席、2017年中国计算机学会大数据学术会议程序委员会共同副主席。提出了面对多样性挑战的大数据泛构理念，建立了基于度量空间的通用大数据管理分析理论框架，获得SISAP2010和BIBE2003国际会议最佳论文奖。

明仲（1967-），男，博士，大数据系统计算技术国家工程实验室执行主任，深圳大学计算机与软件学院教授、常务副院长。现为“鹏城学者计划”特聘教授、深圳大学二级教授、华东理工大学兼职教授、深圳大学及中国科学院深圳先进技术研究院博士生导师、教育部大学计算机基础教学指导委员会委员、第八届中国计算机学会（CCF）理事、广东省省部院物联网产学研联盟副理事长、CCF YOCSEF（北京）AC荣誉委员，曾任CCF YOCSEF（广州）主席，为广东省“千百十工程”省级培养对象、深圳大学计算机软件与理论学科带头人。获广东省优秀教学成果奖一等奖2次、二等奖2次。2009年被评为深圳市优秀教师。“云数据交换和服务调度平台及应用”项目获得2013年度广东省科学技术奖一等奖（排名第一），“基于数据集成和信息共享的ESB集成平台软件”项目获2009年度深圳市科技创新奖（排名第一），并获得2010年度的广东省科学技术奖三等奖（排名第一）。共发表学术论文96篇。

收稿日期：2017-12-15