科学通报 2014 年 第 59 卷 第 36 期：3547 ~ 3553
专题：大型数据发展战略研究 进 展 www.scichina.com csb.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

面向互联网金融行业的大数据资源服务平台

蒋昌俊, 丁志军, 王俊丽*, 闫春钢

同济大学嵌入式系统与服务计算教育部重点实验室, 上海 200092
* 联系人, E-mail: junliwang@tongji.edu.cn

2014-02-26 收稿, 2014-05-09 接受, 2014-07-18 网络版发表
国家自然科学基金可信软件重大研究计划集成项目(91218301)和港澳合作基金(2013DFM10100)资助

摘要 在大数据时代, 数据已经渗透至各个行业, 并且呈现出数量大、动态性、类型复杂等显著特征, 尤其是互联网金融等为代表的典型行业. 本文简要阐述了大数据的研究现状与重大意义, 探讨了大型数据资源服务平台架构及其 3 个主要组成部分: 数据资源识别和获取、数据资源存储和分析、服务支撑平台, 并介绍了项目组在面向可信网络金融交易的大型数据分析研究与应用方面所开展的工作. 具体来讲, 围绕软件行为认证等关键技术, 研究并开发了以行为认证为核心的可信网络金融交易系统, 支持在线交易过程中产生的用户行为数据与软件行为数据的实时监控和动态展示.

关键词
大数据
互联网金融
可信网络交易
行为认证

受全球信息化、人类社会发展和需求多样性、云计算和物联网等信息技术发展的推动, 全球数据增长超越了历史上任何一个时期, 据 IDC 研究报告[1]中指出, 2011 年全球数据总量为 1.8 ZB, 预计到 2020 年将增至 35.2 ZB, 年均增长率超过 40%. 《福布斯》分析指出全球 90%的数据都是在过去 2 年中生成的. 其中, 信息爆炸式地增长最为典型的当属互联网行业, 而且这些信息和数据包括不同数据类型(结构化数据、半结构化数据和非结构化数据). 据统计, 全球每个月发布 10 亿条 Twitter 信息和 300 亿条 Facebook 信息. 而且现在越来越多的新兴科学研究领域完全建立在大量数据的基础上, 比如系统生物学、宏生态学、基因组学、脑科学等. 除此之外, 全世界有着无数的传感器, 随时测量和传递着有关位置、运动、温度、湿度等变化, 产生了海量的数据信息. 因此, 大数据已经不同程度地渗透到工业、科技、交通、电力、医疗、金融、社保、国防、公共安全等人类社会的各个行业领域和部门. 作为新一轮科技和产业竞争的战略制高点, 大数据将推动整个信息产业的创新发展, 促进社会生产力的发展, 改善人们的生活和工作方式, 成为推动世界经济增长和社会发展的重要动力[2,3].

早在 1980 年美国社会思想家托夫勒在 *The Third Vave*[4]中就预言, “如果说 IBM 的主机拉开了信息化革命的大幕, 那么大数据则是第 3 次浪潮的华彩乐章”. “大数据”一词首次被正式提出是在 2011 年麦肯锡全球研究院发布的研究报告[5]中, 这份报告从经济角度讲解了处理这些数据能够释放出的潜在价值, 引发全球对大数据的关注. 当今数据正以前所未有的速度在不断地增长和累积, 但是人类对这些数据的利用率却很低. 学术界、工业界甚至政府机构都已经开始密切关注大数据问题, 并对其产生浓厚的兴趣. 英国 *Nature* 杂志 2008 年“大数据”专刊[6]集中报道了大数据所带来的技术挑战及未来的发展方向,

引用格式: 蒋昌俊, 丁志军, 王俊丽, 等. 面向互联网金融行业的大数据资源服务平台. 科学通报, 2014, 59: 3547–3553
英文版见: Jiang C J, Ding Z J, Wang J L, et al. Big data resource service platform for the internet financial industry. Chin Sci Bull, 2014, 59: 5051–5058, doi: 10.1007/s11434-014-0570-5

标志着大数据分析与处理已经成为科学研究、商业活动、日常生活中的一个核心问题, 成为计算机科学研究最重要的内容之一. *Science* 杂志 2011 年的"数据处理"专刊 [7]主要围绕科学研究中大数据的问题展开讨论, 阐明大数据对科学研究的重要性. 微软研究院出版的 *The Fourth Paradigm* [8]一书中, 图灵奖获得者、著名数据库专家 Jim Gray 博士揭示了在海量数据和无处不在的网络上发展起来的与实验科学、理论推演、计算机仿真这 3 种科研范式相辅相成的科学研究第四范式——数据密集型科学发现. 最初的科学研究是以实验物理学为代表的实验科学; 随后出现了运用了各种定律和定理, 比如开普勒定律、牛顿运动定律等的理论科学; 而对于许多问题, 理论分析方法变得非常复杂以至于难以解决, 人们开始借助计算机仿真的方式来模拟现实世界, 例如模拟神舟飞船从发射到返回各个阶段的飞行状态, 在这一阶段数据主要体现在计算机的输入和输出; 当前, 大数据的重要性正在不断凸显, 数据已成为科学研究甚至是产业的源泉, 因此以数据为中心, 包括数据的识别与获取、数据的存储与分析、数据的交易与决策等主要内容的数据驱动式的研究方式正成为一种新型的科学研究思路.

基于以上考虑, 本文将面向可信网络金融交易典型行业, 提出大型数据资源服务架构, 并介绍项目组所开展的大型数据存储与分析研究与应用方面的相关工作.

1 大型数据资源服务架构

大数据技术及相应的基础研究已经成为科技界的研究热点, 大数据研究作为一个横跨信息科学、社会科学、网络科学、系统科学、心理学、经济学等诸多领域的新兴交叉方向正在逐步形成. 尽管大数据中几乎包含了所有我们需要的信息, 但是由于大数据在数量、类型、动态特征等方面已大大超出了人类的认知, 如何高效处理这么多的动态信息成为一个公认的难题. 最近几年来研究者们已经提出了一些创新的方法来构建大数据平台, 这些研究推动了大数据相关技术的发展和创新. Google 针对大数据问题提出了具有代表性的技术: Google 文件系统(GFS)和 MapReduce 处理模型[9,10]. 有研究显示, Hadoop 和 HDFS 已经发展成为大数据分析的主要平台[11~13]. Garlasu 等人[14]提出了采用网格体系结构的方式来管理大数据的框架. Wu 等人[15]提出了包括数据的访问和计算、数据隐私和领域知识及大数据挖掘算法 3 个层次的大数据处理框架. 目前已有的这些大数据分析平台的研究工作, 侧重于大数据管理、处理、分析和可视化这几个部分中的一个或两个方面. 但是, 随着大数据爆炸式增长、多样化趋势等特征越来越显著, 现有的方法本质上缺少对数据整体上的考虑, 无法刻画和度量数据资源的总体分布和数据成分等特征.

基于这样的考虑, 我们指出大数据分析的首要任务是通过数据"勘探"的方法, 形成大数据资源宏观上的认识[16,17]. 为此, 我们提出了一个基于索引网络的大型数据资源服务框架, 其中包括 3 个主要部分: 数据资源识别和获取、数据资源存储和分析、构建服务支撑平台, 如图 1 所示.

1.1 数据资源识别和获取

大型数据资源通常是分散的、异构的, 而且由于数据量非常之大, 数据完全获取的方式显然是不可能, 需要通过抽样的方法, 获取少量有效样本以统计出总体的分布. 因此, 在数据资源识别和获取这一层次, 一方面, 将通过探讨所访问的互联网资源的类型、数据成分、网络接口限制等特点, 正确分析这些因素对于数据获取和分析的影响, 建立符合大规模网络数据资源特性的统计模型. 另外一方面, 将在综合考虑各种网络限制的基础上, 通过数据资源勘探和探索等方法, 引入拒绝抽样等技术确保样本单元的独立性.

1.2 数据资源存储和分析

目前海量异构数据一般采用分布式存储技术, 如 GFS 和 HDFS, 但它们仍不能解决数据的爆炸性增长带来的存储问题, 静态的存储方案并不能满足数据的动态演化所带来的挑战. 因此, 在数据资源存储和分析这一层次, 需要根据特定的数据资源建立相应的分析和存储方法, 一个良好的存储机制可以从多样化的方面支持资源分析. 而资源分析的目的是为了提取数据资源之间的关联. 其中, 复杂数据分析方法有助于从多个数据源推断出的聚集的分析结果, 侧重于结构化数据的数值型统计分析, 而针对非结构化和半结构化数据, 为了得到更有价值的数据信息分析结果, 需要借助于机器学习等语义分析技术, 获取数据资源之间的语义和逻辑关系.

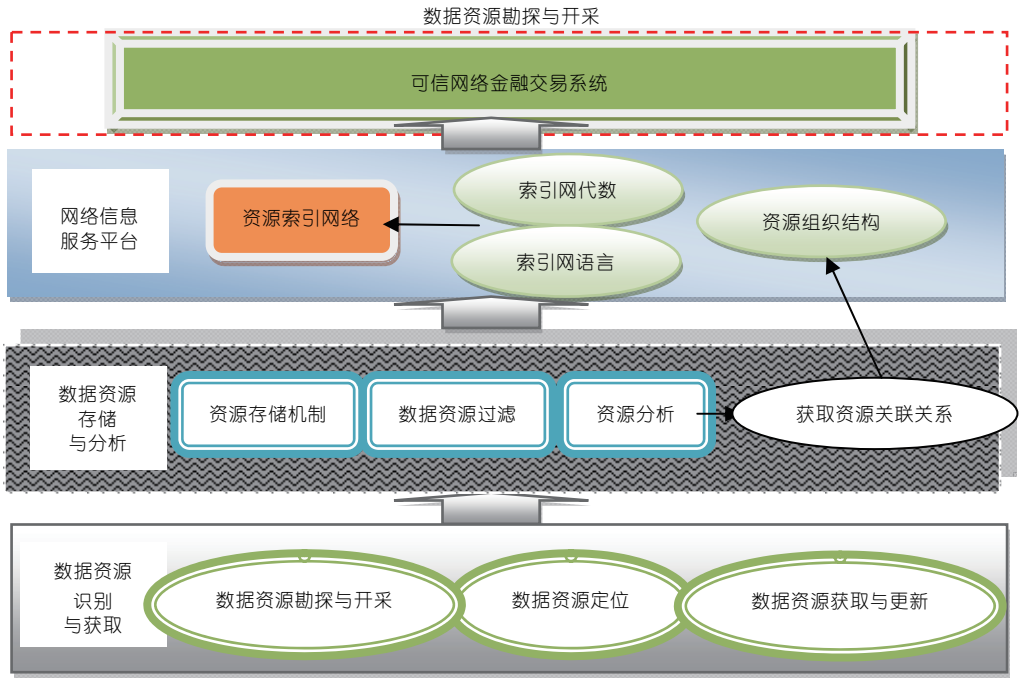

图 1 (网络版彩色)大型数据资源服务架构

1.3 网络信息服务平台

网页是目前互联网服务中最基本的资源，在信息呈现、支持应用程序和提供服务等方面发挥主导作用. 每天都有众多的网页加入到互联网中，其中大部分是冗余的、无序的. 因而从互联网上查找所需的服务资源是非常有挑战性的. 为此，我们之前已经在 Web 超链分析领域进行了深入的研究，并将网页之间的超链接视作现实世界的客观关系，并在此基础上，提出建立基于网页的分类和超链接分析的索引网络模型[18]，并给出了其代数运算的定义. 索引网络支持根据具体要求获取服务资源，以及寻找它们之间的语义关联，能产生更丰富的知识和有价值的信息服务. 文献[19]对这一原型系统进行了更深入的探讨.

在此基础上，面向典型的数据驱动行业，开发了相应的服务应用系统. 本文接下来将重点介绍项目组在可信网络金融交易系统研究与开发过程中开展的数据勘探、分析等方面的工作.

2 可信网络金融交易系统

金融业是大数据的重要产生者，交易、报价、业绩报告、消费者研究报告、官方统计数据公报、调查、新闻报道无一不是数据来源. 据中国人民银行支付结算司的《2013 年第一季度支付体系运行总体情况》，第一季度，全国共发生电子支付 56.12 亿笔业务，金额 217.59 万亿元，同比分别增长 29.99%和 26.20%，其中网络移动支付 1.98 亿笔，金额 1.10 万亿元，同比分别增长 139.30%和 206.46%，移动支付增长速度位于各类支付业务之首，这与互联网金融的快速发展有相当大的关联. 金融业也高度依赖信息技术，应用大数据方法与技术收集、处理、分析金融数据，并对数据进行挖掘提取，寻找其中有价值的信息，并将这些信息转化为知识，可以帮助企业做出及时准确的决策. 阿里巴巴集团是互联网金融企业的代表，数据显示，支付宝 2013 年双 11 全天交易额达 350.19 亿元，相当于 9 月份中国社会零售总额的一半，其年交易额过万亿元，用户从搜索到浏览、支付，每一个节点都将产生大量数据. 淘宝首席商业智能官车品觉表示：“阿里集团目前拥有的大数据达到 30 PB，目前有 800 名员工从事大数据相关的工作”，大数据可以帮助他们分析历史数据，寻找其中的金融创新机会.

互联网金融环境中，数据作为金融核心资产，具有相当大的价值，但同时它又存在着巨大的安全隐患，金融行业不能容忍任何安全问题，一旦出现问题，必然会对企业和个人造成巨大的损失[20]. 针对网络金融信息安全问题，项目组研究并开发了以行为认

证为核心的可信网络金融交易系统, 围绕软件行为认证等关键技术, 搭建了行为认证平台体系.

2.1 行为认证平台体系

在认证中心搭建过程中, 我们通过在用户安全客户端以及在电商网站和支付平台部署行为监控器, 形成网络交易可信认证系统平台, 并制定网络交易可信认证的认证协议. 在网络交易可信认证系统中, 认证中心主要负责管理用户行为和软件行为证书, 同时能够实时认证软件及用户行为的可信性.

网络交易可信认证中心底层支持多种操作系统, 具有良好的跨平台能力. 系统之上的支撑技术为上层的应用开发提供了良好的支持. 在支撑技术之上设计通信管理模块、证书管理模块和数据库管理模块; 通信管理模块能够针对本系统特定需求对网络通信功能进行封装, 为上层提供数据交换等通信服务; 证书管理模块对软件行为证书、用户行为证书以及数字证书进行统一的管理, 包括证书的搜索、更新、发布等操作; 数据库管理模块负责更新和维护数据库, 提高数据访问效率. 在基础管理模块之上, 就是网络交易可信认证系统的第四方认证域, 其主要功能是监控和认证网络交易过程, 对交易三方进行数字认证、通过用户行为证书验证用户身份的可信性、通过软件行为证书验证交易三方的网络交易行为的可信性. 网络交易可信认证中心架构如图 2 所示.

网络交易可信认证中心的认证协议流程如下: 当网络交易发生时, 用户通过登录安全客户端, 上传数字证书进行数字认证, 电商和第三方支付也同时上传其数字证书进行相应的数字认证. 当数字认证通过后, 用户通过用户行为证书下载模块下载行为证书, 三方正式进入交易流程. 在交易过程中, 安全客户端通过用户行为采集模块实时采集用户行为, 并交给用户行为认证模块, 根据从第四方认证中心下载的该用户行为证书认证用户当前访问行为的可信性. 如果认证通过, 那么继续采集用户的访问行为, 进行认证; 若认证不通过, 则将详细认证结果上传至认证中心, 由认证中心进行审查、判定. 同时, 通过软件行为采集模块实时采集客户端软件行为, 并由通信交互模块上传至认证中心. 而电商和第三方支付也同样通过软件行为监控模块实时采集其软件行

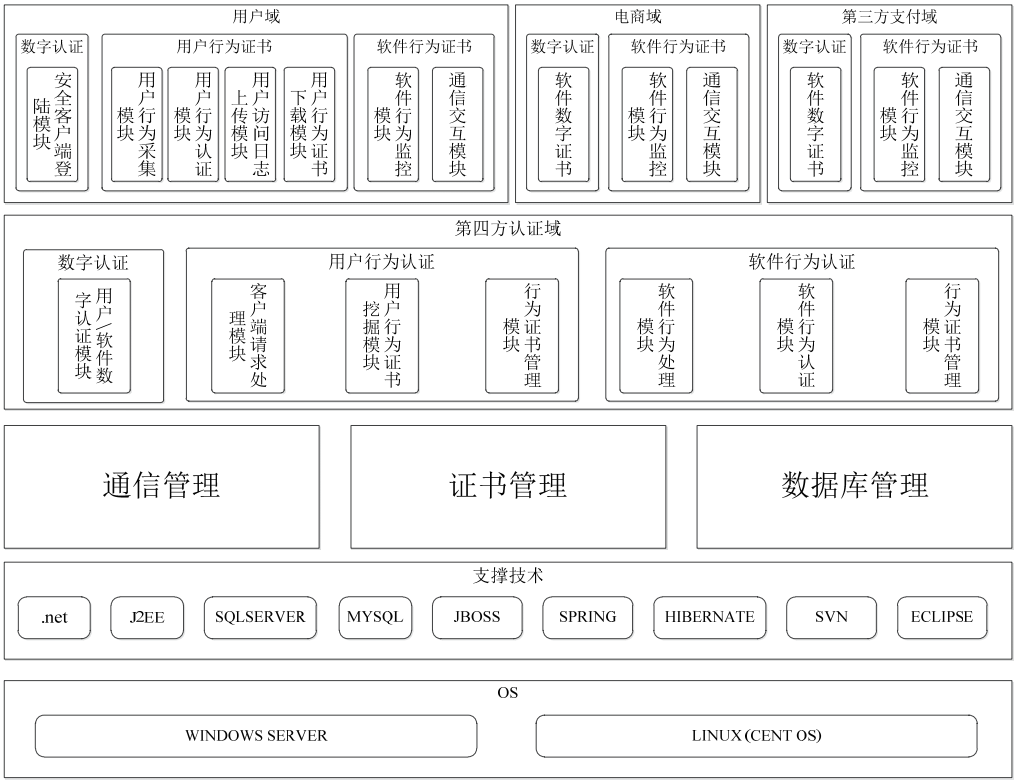

图 2 网络交易可信认证中心架构图

为, 并由通信交互模块上传至认证中心. 如果软件行为认证通过, 则认证中心发回反馈信息, 继续进行交易流程, 同时三方软件行为监控继续进行实时采集; 若认证不通过, 则由认证中心广播通知交易三方交易流程出现异常, 并终止交易. 当交易完成后, 安全客户端由用户访问日志上传新的访问日志至认证中心, 当认证中心收到新的访问日志后, 发回反馈信息, 用户退出安全客户端. 接着, 认证中心通过证书管理模块调用用户行为证书挖掘模块对新的用户访问日志进行挖掘, 更新该用户的行为证书. 当一个新的电商或第三方支付平台加入, 则首先对其进行审核, 通过后颁发数字证书; 接着通过分析其网站源码, 挖掘出其相应的软件行为证书, 上传至认证中心, 由行为证书管理模块统一进行管理. 网络交易可信认证中心认证流程如图 3 所示.

2.2 软件行为认证

根据用户、电子商务网站、第三方支付平台在正确交易流程下的三方通信数据包, 由专业人员刻画三方正常合法交互行为, 形成软件行为模型, 构建软件行为证书. 软件行为分析整体框架如图 4 所示.

软件行为证书将三方之间的交易信息交互过程抽象成 Petri 网, 将三方每次执行一步作为一个变迁, 例如修改数据库、修改订单状态等; 将三方特定的行为理解为触发条件并抽象为库所, 如订单消息、状态消息等和单击购买按钮行为等; 同时, 规定一个变迁中每个输入库所必须有且唯有一个 token, 变迁才有资格被触发. 在软件行为证书构建完成后, 三方身份判别由软件行为监控验证系统来实现. 软件行为监控验证系统由三方软件行为监控器和软件行为实时验证系统 2 个部分组成. 三方软件行为监控器主要监

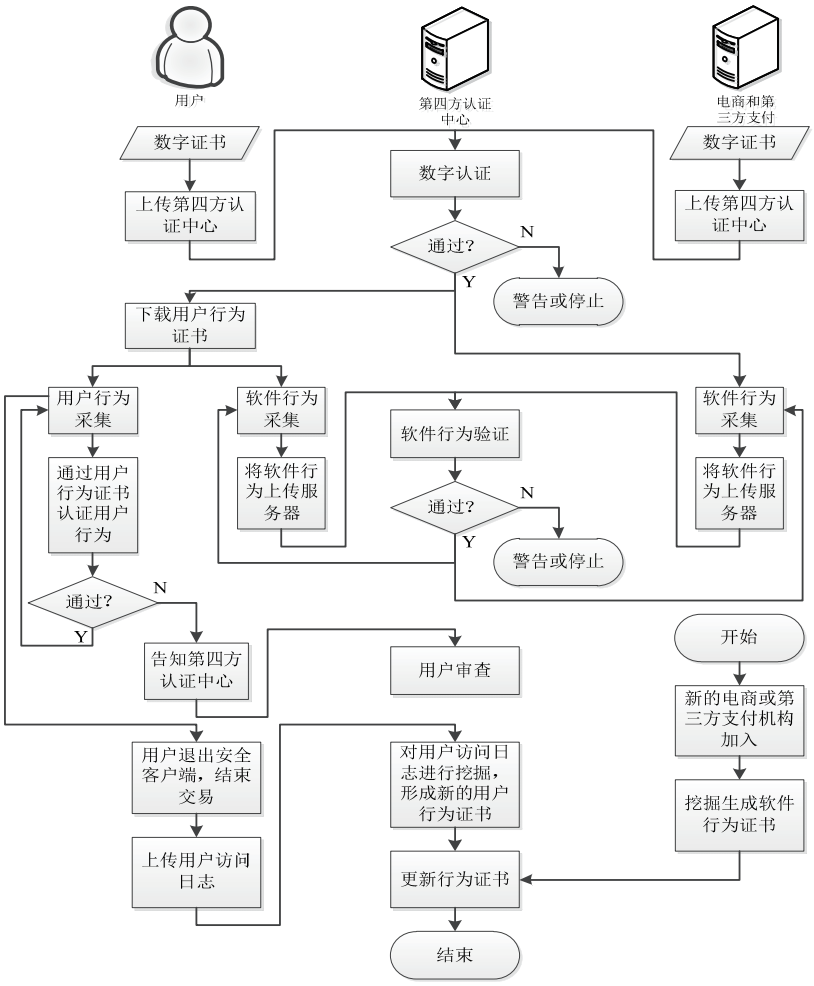

图 3 网络交易可信认证中心认证流程

控三方交易交互数据包并提取必要信息(URL 地址、参数等), 将关键信息以数据包的形式发送给软件行为实时验证系统. 软件行为实时验证系统在接收三方监控器分别提交的交易交互信息数据包后, 提取并整合其中的关键序列与信息, 并将此序列信息设置为交互序列与软件行为模型进行实时对比, 一旦发生乱序, 如假冒身份等非法行为则进行警报并关闭交易[21].

2.3 可信认证在线监控可视化呈现

可信认证中心监控中心属于可信网络交易软件系统试验环境与示范应用项目下, 用于监控用户、商家和第三方支付公司在进行在线交易行为时产生的用户行为数据与软件行为数据, 并采用多种类、多维度的表格与图表的方式直观动态地展现过程中产生的数据. 监控中心作为直观动态展现以上数据的平台, 目前主要分为 3 个部分, 每个部分又分别由 3 个屏幕组成, 共 9 个屏幕组成. 3 个部分分别为平台软件行为监控、平台交易数据监控和平台用户行为监控. 软件行为监控分屏显示了包括购物者、电商、第三方支付平台三方的软件行为监控日志. 平台交易数据监控为模拟经过四方认证平台的实时交易模拟数据, 具体包含了滚动展现的交易日志, 全国交易数据的分布以及平台实时的交易额与交易笔数数据. 用户行为监控以单用户与多用户的用户行为浏览日志与评分, 以及包含频繁访问类和访问时间段在内的多维度的用户浏览习惯展现.

第一部分为平台软件行为监控, 其主要监控包含了电商、第三方支付以及用户的软件行为监控, 监控系统通过滚动列表的方式, 展示软件行为的日志, 并高亮显示异常交易, 以此帮助业务人员分析异常报警. 部分界面如图 S1 所示.

第二部分是平台用户行为监控可视化, 这部分是对平台用户行为习惯监控数据的可视化, 其子部分包含了多维度的用户网络行为信息, 如用户的上网时间段的分布、用户访问的网站类的成分等, 通过多维度信息展现用户的行为习惯. 部分界面如图 S2 所示.

最后一部分为平台交易数据可视化, 用于展示经过平台的交易数据, 其数据可以通过实时数据服务从受监控的外部电商平台获取, 包括全国交易量统计, 实时交易量监控等信息, 部分界面如图 S3 所示.

3 结束语

当今人类社会的各个行业, 如工业、科技、交通、医疗、金融等领域和部门都产生了大量的数据信息, 这些大数据已成为一种资源, 几乎包含了所有我们需要的信息, 蕴含着巨大价值. 但正是由于这些大数据的广度和容量, 以及这些数据的多源异构的本质对数据收集、存储和处理, 特别是数据分析与计算带来了非常大的困难. 大数据分析与处理已经成为科

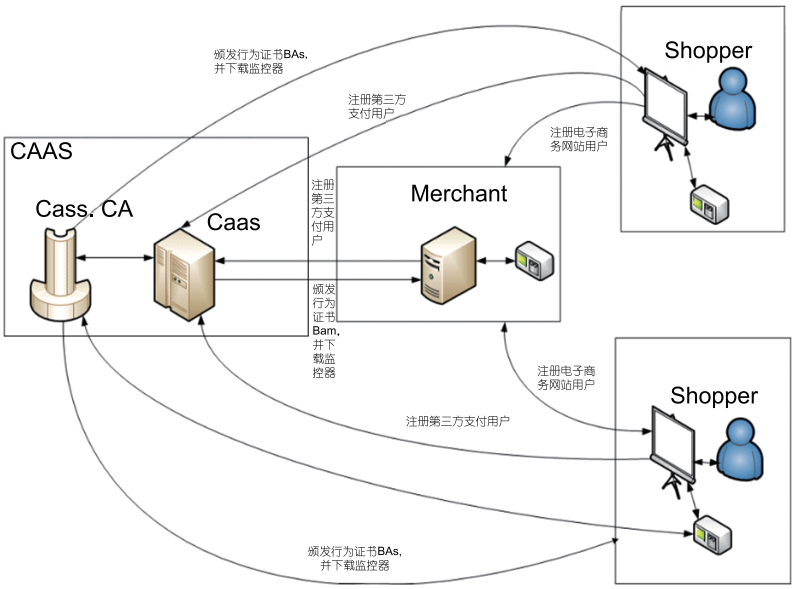

图 4 (网络版彩色)软件行为分析整体框架

学研究、商业活动、日常生活中的一个核心问题. 本文中我们以典型的数据驱动行业(网络金融行业)为背景, 介绍了项目组前期在数据勘探、分析等方面开展的工作.

在互联网金融环境中, 数据作为金融核心资产, 仅阿里巴巴集团就拥有 PB 量级大数据, 具有相当大的价值. 但同时它又存在着巨大的安全隐患, 针对这一问题, 项目组研究并开发了以行为认证为核心的可信网络金融交易系统, 围绕软件行为认证等关键技术, 搭建了行为认证平台体系.

参考文献

1 IDC Digital Universe Study: Extracting Value from Chaos, sponsored by EMC. 2011
2 李国杰. 大数据研究的科学价值. 中国计算机学会通讯, 2012, 8: 8–15
3 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. 中国科学院院刊, 2012, 6: 647–657
4 Toffler A. The Third Wave. New York: William Morrow and Company, 1980
5 Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/Insights/MGI/Research/Technology and Innovation/Big_data_The_next_frontier_for_innovation
6 Nature. Big Data. [20130320]. http://www.nature.com/news/specials/bigdata/ index.html
7 Science. Special Online Collection: Dealing with Data. [20130320]. http://www.sciencemag. org/site/special/data/. 2011.2
8 Hey T, Tansley S, Tolle S. The fourth paradigm: Data-intensive scientific discovery. Microsoft Res, 2009, Redmond, Wasington
9 Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Sixth symposium on operating system design and implementation, San Francisco, CA. 2004
10 Hall A, Bachmann O, Bussow R, et al. Processing a trillion cells per mouse click. Proc VLDB Endow, 2012, 5: 1436–1446
11 Olofson V, Eastwood M. Big Data: What it is and why you should care, White Paper, IDC. 2012
12 Ferguson M. Architecting a big data platform for analytics. A Whitepaper Prepared for IBM, 2012
13 Borkar V R, Carey M J, Li C. Big data platforms: What's next? XRDS, 2012, 19: 44–49
14 Garlasu D, Sandulescu V, Halcu I, et al. A big data implementation based on Grid computing. Roedunet International Conference (RoEduNet), 2013: 1–4
15 Wu X D, Zhu X Q, Wu G Q, et al. Data mining with big data. IEEE Trans Knowl Data En, 2014, 26: 97–107
16 蒋昌俊. 大数据的勘探与分析的若干思考. 国家自然科学基金委双清论坛报告, 2013
17 蒋昌俊. 互联网非合作环境下大数据的探析问题. 中国科学院科学与技术前沿论坛报告, 2013
18 Jiang C J, Ding Z J, Wang P W. An Indexing Network Model for information services and its applications. In: Proceedings of the 6th IEEE International Conference on Service Oriented Computing and Applications, 2013. 290–297
19 Deng X D, Jiang M, Sun H C, et al. A novel information search and recommendation services platform based on an Indexing Network. In: Proceedings of the 6th IEEE International Conference on Service Oriented Computing and Applications, 2013. 194–197
20 Du H, Wang J, Liu Y N. A time sequence protocol to achieve the effect of fair exchange without trusted third party. Chin Sci Bull, 2014, 59: 699–702
21 Yu W Y, Yan C G, Ding Z J, et al. Modeling and monitoring of online shopping business processes based on system behavior patterns. J Comput Infor Syst, 2013, 9:1–8

补充材料

图 S1 软件行为监控

图 S2 用户行为监控

图 S3 交易量监控

本文以上补充材料见网络版 csb.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.