项目名称：多源异构海量数据集成平台及应用

一、提名意见

提名意见：

我单位认真审阅了该项目提名书及附件材料，确认全部材料真实有效，相关内容符合国家科技奖的提名要求。

该项目主要围绕国家重大战略技术需求，在国家 863 计划、国家自然科学基金和产学研合作等项目资助下，发明了面向多源异构海量数据多样性集成管理需求的聚合优化计算引擎、基于对象代理的多源异构海量数据动态柔性集成机制、基于人机交互的多源异构海量数据复合错误清洗方法和在线-离线相结合的多源异构海量数据融合处理查询分析技术等，解决了多源异构海量数据集成中的结果可用和处理高效两大根本问题，实现了重大技术创新与突破；自主研制了多源异构海量数据集成平台的核心技术，主要技术指标达到或超过国际同类产品先进水平，有力推动了我国信息产业的自主研发能力进步，带动了相关产业链的发展。在东软集团股份有限公司、上海宝信软件股份有限公司、国家海洋信息中心和广东昆仑信息科技有限公司等 10 余家企事业单位进行了应用，取得突出的经济效益和社会效益，近三年累计新增销售额 8.35 亿元，新增利润 1.33 亿元。该项目共获得授权发明专利 61 项，登记软件著作权 71 项，发表学术论文 72 篇，学术成果得到了高度评价，产生了积极的学术影响。对照国家技术发明奖授奖条件，

提名该项目为国家技术发明奖 二 等奖。

二、项目简介

多源异构海量数据集成通过对海量远程源端的异构数据进行采集、集成与清洗，支持高效的查询与分析服务，是大数据采集领域的核心关键技术。源端不仅模式易变、模态多样、错误共生，而且高度自治、质量参差不齐、访问模式各异，为数据集成、错误清洗与查询分析等带来巨大挑战。开展多源异构海量数据集成平台技术研发，实现源头创新，已成为掌握大数据采集核心技术、发展自主可控的战略新兴信息产业的必然选择。

本项目在国家 863 计划、国家自然科学基金和产学研合作等项目的持续支持下，面向多源异构海量数据的多样性集成需求，在聚合计算引擎、动态数据集成、复合错误清洗、融合查询分析等四方面实现了重大技术创新与突破，自主研制了多源异构海量数据集成管理与分析的核心技术，主要技术指标达到或超过国际同类产品先进水平。主要技术发明点如下：

(1) 发明了基于消息的任务交互模型和基于障栅的迭代处理机制，突破了 MapReduce 任务独立的理想并行计算模型的性能制约，实现了数据依赖多样性查询和作业模式多样性查询的聚合优化处理；发明了规模感知的弹性分布式文件系统，实现了规模多样性文件的统一存储。

(2) 发明了基于对象代理的多源异构海量数据集成机制，提出了基于语义知识的海量异构模式集成方法，突破数据源模式多变、结构灵活、源端伸缩带来的挑战，实现了多源异构海量数据的动态集成。

(3) 发明了基于人机交互和分布式的多模态数据清洗算法，提出了共生错误最优化清洗流程，突破模态多样、错误共生、数据高熵的桎梏，实现了多源异构海量数据的快速有效清洗，准确率达 90%以上，破解了复合错误修正的难题。

(4) 针对查询分析中的鲜效权衡问题（结果时鲜与查询效率间的矛盾），发明了基于多目标优化的在线-离线协同调度策略和基于鲜效协同保障的多源结果渐进融合方法，有效解决了源端差异导致的性能劣化问题；发明了最小化冗余访问的在线-离线协同优化机制和基于分布式 ELM 的分析模型训练方法。实现了百万量级数据源查询分析的鲜效权衡。

该项目获得授权发明专利 61 项，登记软件著作权 71 项，发表论文 72 篇。核心技术应用于东软集团、上海宝信、国家海洋信息中心等 10 余家企事业单位的产品和业务化运行系统中，支撑了系统集成商、政府部门、企事业单位等 100 余家大型单位的关键业务系统，在我国“数字医疗”、“数字水资源”和“数字国土资源”等领域起到了不可替代的作用。其中，东软集团依托项目整体技术研制的“数字医疗”系列产品，已服务国内 1700 多家医疗机构，市场份额全国第一；基于对象代理模型的动态柔性集成技术应用于国家海洋信息中心的“908 专项”数据集成交换体系，填补了国内空白。近三年，累计新增销售额 8.35 亿元，新增利润 1.33 亿元。曾获得 2016 年度“教育部科技进步一等奖”、2014 年度“辽宁省科技进步一等奖”和 2011 年度“教育部科技进步二等奖”。

三、客观评价

1．成果获奖

(1) 项目完成人王国仁的“海量异构数据集成管理与分析技术及应用”项目获得 2016 年教育部科学技术进步一等奖（排名第一）。

(2) 项目完成人王国仁的“非结构化数据管理关键技术及应用”项目获得 2014 年辽宁省科学技术进步一等奖（排名第一）。

(3) 项目完成人彭智勇的“面向大众的城市交通按需服务的关键技术及其应用平台”项目获得 2011 年教育部科学技术进步二等奖（排名第三）。

(4) 技术发明成果第 1 项中“基于消息管道的任务交互计算模型”获得了数据库领域重要学术会议 DASFAA 2012 的“Best Paper Award 1st Runner-up”。

(5) 技术发明成果第 3 项中“基于概率图模型的实体识别方法”获得了数据库领域重要学术会议 WISE 2013 的“Best Challenge Paper Award”。

(6) 东软集团依托本项目整体技术研发的“RealSight 大数据高级分析应用平台”荣获 2016 中国软件行业“创新产品奖”。

2. 项目验收意见

(1) 科技部高技术研究发展中心 2015 年 9 月 25 日组织专家在沈阳组织验收了 863 计划课题“海量不确定异构数据的集成管理与分析技术/2012AA01100”，验收意见：“…探索了海量不确定异构数据存储、清洗与集成管理和分析等关键技术；…研制了海量不确定异构数据集成管理和分析工具软件原型。…开展了典型应用示范验证，…发表论文 35 篇，申请发明专利 8 项，获得软件著作权 3 项。…已完成合同规定的主要研究任务及其指标…一致同意通过验收。”

(2) 国家自然科学基金委员会信息科学部 2015 年 4 月 24 日发布资助项目准予结题通知：“李战怀 同志：您承担的国家自然科学基金项目：(数据密集型计算环境下的数据管理方法与技术)，批准号：(61033007) 按有关规定已审核完毕，准予结题。”该项目发表论文 160 篇，其中 SIGMOD、VLDB、ICDE、 TODS、TKDE、VLDBJ 等顶级会议和期刊论文 12 篇；申请发明专利 17 项，登记软件著作权 26 项。

(3) 国家自然科学基金委员会信息科学部 2006 年 7 月 15 日发布资助项目准予结题通知：“彭智勇 同志：您承担的国家自然科学基金项目：(基于对象代理模型的网上异构多信息源集成系统研究)，批准号：(60273072) 按有关规定已审核完毕，准予结题。”该项目发表论文 17 篇，包括 TKDE 和计算机学报等国内外顶级期刊；获得软件著作权 2 项。

(4) 国家自然科学基金委员会信息科学部 2014 年 4 月 25 日发布资助项目准予结题通知：“王国仁 同志：您承担的国家自然科学基金项目：(不确定数据管理的理论与关键技术)，批准号：(60933001) 按有关规定已审核完毕，准予结题。”该项目发表论文 98 篇，其中 VLDB、TKDE 等顶级会议和期刊论文 18 篇；申请发明专利 12 项，其中授权 6 项。

3. 检测评价

数据库领域重要学术会议 WISE 2013 Challenge“Entity Linking Track (T1)”使用 2011 年 Wikipedia 实体名对文章中出现的命名实体和具体类别进行了标注，并提供了考虑同义词问题的实体识别性能评测工具。项目研制的“基于统一框架的多模态数据清洗”（属于**技术发明点三**）中的实体识别工具参加了该竞赛的“recall of proper nouns plus detailed categories”评测环节，取得了 47.5%的成绩，在 6 个代表队提交的 13 份结果中名列第一。

4. 学术评价

项目获得授权发明专利 61 项，软件著作权 71 项，发表学术论文 72 篇。

IET 和 BCS Fellow、北达科他州立大学的 Samee U. Khan 教授将课题组提出的“基于消息管道的任务交互计算模型”（属于**技术发明点一**）作为“Most influential articles contributed to the improvement in MapReduce framework”之一，给予了高度评价，认为：ComMapReduce 框架在保留 MapReduc 框架优势的同时，全面提升了处理效率。原文如下：“Compared with the original MapReduce framework in all metric without affecting the existing characteristics of the former, ComMapReduce is deemed better.”(MapReduce: Review and Open Challenges. Scientometrics, 2016, 109(1): 389-422)。

COAST 是德国国家信息研究中心开发的面向对象的同步协同工作工具集。它的一个重要特点是共享文档能够以不同粒度和方式，动态地呈现给协同工作成员，提高了协同工作中文档共享的柔软性。他们正是采用对象代理模型（属于**技术发明点二**）实现这一功能，即用对象表示文档，通过代理对象实现文档和用户的动态绑定。该论文被 Google Scholar 引用 231 次。(Designing Object-Oriented Synchronous Groupware with COAST, CSCW, 1996: 30-38)。

如何根据用户的需要对大量的半结构化数据进行结构化处理是当前的重要研究问题之一。日本关西大学上岛绅一教授提出了一个层次构造图模型。该模型通过定义视点，允许用户以不同角度观察和集合半结构化数据，具有相当的柔性。该项工作发表在日本计算机权威学术期刊《情报处理学会论文杂志》上，文中对层次构造图模型和对象代理模型进行了详细比较，认为：对象代理模型（属于**技术发明点二**）也是表现数据多面性的有效方法，但它主要是针对地理数据，而层次构造图模型主要用于半结构化数据。(階層構造グラフを用いた半構造化データの構造化手法, 日本情報処理学会論文誌, 1998, 39(4): 857-867)

IEEE Fellow、湖南大学“千人计划”李克勤教授将课题组提出的分布式极限学习机技术（属于**技术发明点四**）作为并行 ELM 的典型成果之一，给予了高度评价，并进行了系统评测与性能对比分析，认为提出的分布式极限学习机 ELM*仅次于他们新提出的基于 Spark 平台的 SELM，在 MapReduce 框架下具有最优的分布式训练性能。(A Parallel Multiclassification Algorithm for Big Data Using an Extreme Learning Machine, TNNLS, 2018, 29(6): 2337-2351)

四、推广应用情况

本项目首次提出了基于消息的任务交互计算模型，突破了 MapReduce 任务独立理想并行计算模型的性能制约，建立了面向多源异构海量数据多样性集成管理需求的聚合优化计算引擎。发明了基于对象代理的多源异构海量数据动态柔性集成机制、基于人机交互的多源异构海量数据复合错误清洗方法和在线-离线相结合的多源异构海量数据融合处理查询分析技术。

技术发明成果得到了国内外同行和评测机构的广泛关注，同行专家认为分布式 ELM 分析模型训练方法、基于人机交互的复合错误清洗算法、基于概率图模型的实体识别技术等成果处于国际领先水平，解决了数据集成和大数据采集领域的多项核心技术，显著增强我国数据集成和大数据采集核心技术的国际竞争力。自主研制的更新迁移模式演进机制、基于分布式 ELM 的分析模型训练方法、多源结果渐进融合技术等已与国内外多家机构开展合作，发展前景广阔，具有潜在的巨大经济和社会效益。主要知识产权中的“一种面向大数据的数据清洗系统及方法”、“一种海量微博数据的分布式分类装置及方法”、“具有增减量功能的海量微博数据分布式分类装置及方法”和“面向大规模不确定物流网络的需求概率查询方法”等 10 余项授权发明专利在沈阳东深、扬州恒隆、沈阳希艾和沈阳智慧时代四家公司许可实施，使用费合计 1160.00 万元，如表 1 所示。

表 1 专利实施许可情况表

序号	单位名称	专利实施许可数量	许可方式	许可期限	使用费 (万元)
1	沈阳东深科技有限公司	5	普通许可	5 年	260.00
2	扬州恒隆软件有限公司	5	普通许可	5 年	300.00
3	沈阳希艾科技咨询有限公司	6	普通许可	5 年	300.00
4	沈阳智慧时代科技有限公司	6	普通许可	5 年	300.00
	合计	——	——	——	1160.00

本项目的技术发明成果已经被广泛应用，成功应用于“数字海洋”、“数字医疗”和“数字水资源”等领域，支撑了东软集团、上海宝信、国家海洋信息中心、广东昆仑、扬州恒隆、深圳同立方、武汉图信、哈尔滨乐辰、黑龙江亦轩和宁夏理工学院等企事业单位的关键业务系统。

聚合优化计算引擎在上海宝信的“国控上海水资源一期工程、国控浙江水资源一期应用开发、江苏中小水库防汛预警”等项目中应用，突破了多年来一直困扰的数据依赖处理、迭代作业支持、海量小文件存储等方面的性能瓶颈。

对象代理柔性集成应用于国家海洋信息中心的“908 专项”，实现了 11 个省市、3 个分局、4 个中心的 9 类基础数据的无缝集成，海洋数据动态集成与交换体系填补了国内空白；应用于深圳同立方和武汉图信，攻克了模式多变、结构灵活、源端伸缩等数据集成难题。人机交互错误清洗在国家海洋信息中心的“908 专项”中应用，实现了复合错误的快速有效清洗，准确率达 95%以上；在哈尔滨乐辰、黑龙江亦轩和宁夏理工学院应用，解决了各系统中广泛存在的数据重复和不一致问题；在线-离线查询分析在东深科技和希艾咨询的培养和发展中，发挥了十分重要的支撑作用，为提高自主创新能力、建设创新型社会提供手段和保障，为产业结构优化升级发挥支撑和引领作用。应用情况如表 2 所示。

表 2 主要应用单位情况表

序号	单位名称	应用的技术	应用对象及规模	应用起止时间	单位联系人/电话
1	东软集团股份有限公司	第 1-4 项发明	国内 1700 多家医疗机构	2012 至今	窦丽莉/ 024-83665402
2	上海宝信软件股份有限公司	第 1 项发明	国控上海和浙江水资源一期工程等项目	2012 至今	嵇晓/ 021-50801155
3	国家海洋信息中心	第 2、3 项发明	11 个省市、3 个分局、4 个中心	2012 至今	张峰/ 022-24010888
4	广东昆仑信息科技有限公司	第 1 项发明	韶关市齿轮厂、丹霞冶炼厂等单位	2013 至今	黄健美/ 13602915079
5	扬州恒隆软件有限公司	第 1 项发明	秦皇岛、通辽、喀什、亳州等地区	2012 至今	王牧人/ 0514-86448889
6	深圳市同立方科技有限公司	第 2 项发明	发展到杭州、武汉两家分公司，上百员工	2015 至今	丁丁/ 0755-28245090
7	武汉图信科技有限公司	第 2 项发明	依托武汉 •中国光谷，服务湖北省大型企业	2012 至今	宋伟/ 027-68775717
8	哈尔滨乐辰科技有限公司	第 3 项发明	基于云计算平台的大数据健康服务系统	2014 至今	赵岩/ 15048139853
9	黑龙江亦轩科技有限公司	第 3 项发明	提升公司工业信息化系统的产业化能力	2014 至今	姜波/ 18845151169
10	沈阳东深科技有限公司	第 4 项发明	服务省内 36 所高校，覆盖 34 万大学生	2012 至今	刘洪伟/ 024-83769398
11	沈阳希艾科技咨询有限公司	第 2、4 项发明	辽宁情报所	2016 至今	吴浩/ 024-83186098
12	宁夏理工学院	第 3 项发明	学校教学、科研、后勤等部门	2012 至今	王巨轮/ 15226229901

五、主要知识产权和标准规范等目录（不超过 10 件）

知识产权（标准）类别	知识产权（标准）具体名称	国家（地区）	授权号（标准编号）	授权（标准实施）日期	证书编号（标准批准发布部门）	权利人（标准起草单位）	发明人（标准起草人）	发明专利（标准）有效状态
发明专利	一种面向对象代理数据库的虚属性查询优化方法	中国	201310139781.8	2016-06-08	2108322	武汉大学	彭智勇、王梁、 付祖发、彭煜玮	有效
发明专利	一种面向大数据的数据清洗系统及方法	中国	201410483041.0	2017-07-18	2554986	东北大学	王国仁、信俊昌、聂铁铮、赵相国、邓诗卓、季航旭、侯喆、梁帅	有效
发明专利	一种海量微博数据的分布式分类装置及方法	中国	201210583886.8	2015-10-28	1823542	东北大学	王国仁、信俊昌、聂铁铮、赵相国、丁琳琳	有效
发明专利	面向云资源调度的热点移除方法	中国	201310323538.1	2016-03-30	2007492	西北工业大学	刘文洁、李战怀、潘巍、张晓	有效
发明专利	一种云环境下轻量级的细粒度访问控制方法	中国	201310138434.3	2015-09-06	1786767	武汉大学	彭智勇、程芳权、王书林、宋伟	有效
发明专利	Web 数据管理系统	中国	201010140168.4	2012-02-08	908682	武汉大学	彭智勇	有效
发明专利	基于频繁关联标签序列的 XML 结构相似度度量方法	中国	201110398187.1	2013-04-24	1183176	西北工业大学	张利军、李战怀、陈群、李霞	有效
发明专利	具有增减量功能的海量微博数据分布式分类装置及方法	中国	201310732005.9	2017-01-18	2349161	东北大学	王国仁、信俊昌、聂铁铮、赵相国、丁琳琳	有效
发明专利	面向大规模不确定物流网络的需求概率查询方法	中国	201210248045.1	2013-06-05	1588321	东北大学	王国仁、袁野、孙永佼、赵相国、韩东红、王斌	有效
发明专利	基于特征分布信息的文本分类特征筛选方法	中国	201310050583.4	2016-02-10	1947316	西北工业大学	李思男、李战怀、李宁	有效

六、主要完成人情况表

姓　　名	王国仁	排　　名	1
行政职务	无	技术职称	教授
工作单位	北京理工大学		
完成单位	东北大学		

对本项目技术创造性贡献：

本项目总负责人和总体技术路线制订者，提出了聚合优化计算引擎、动态柔性集成机制、复合错误清洗方法和融合处理查询分析技术，并深入合作单位和应用单位组织技术实施与协调。对第 1、2、3 和 4 项技术发明均做出了创造性贡献。

在本项目研发工作中投入的工作量占本人工作量的 80%，获得授权发明专利 16 项。

六、主要完成人情况表

姓　　名	李战怀	排　　名	2
行政职务	无	技术职称	教授
工作单位	西北工业大学		
完成单位	西北工业大学		

对本项目技术创造性贡献：

本项目总体技术路线制订者，设计了面向多源异构海量数据多样性集成管理需求的聚合优化计算引擎、实现了鲜效协同保障的多源结果渐进融合技术，并深入合作单位和应用单位组织技术实施与协调。对第 1 和 4 项技术发明做出了创造性贡献。

在本项目研发工作中投入的工作量占本人工作量的 70%，获得授权发明专利 13 项。

六、主要完成人情况表

姓　　名	彭智勇	排　　名	3
行政职务	大数据研究院副院长	技术职称	教授
工作单位	武汉大学		
完成单位	武汉大学		

对本项目技术创造性贡献：

重点开展基于对象代理的多源异构海量数据动态集成机制等关键技术的研究，提出了基于对象代理的柔性集成机制，设计了基于语义知识的异构模式集成方法，发明了基于更新迁移的模式演进方法。对第 2 项技术发明均做出了创造性贡献。

在本项目研发工作中投入的工作量占本人工作量的 80%，获得授权发明专利 26 项。

六、主要完成人情况表

姓　　名	王宏志	排　　名	4
行政职务	无	技术职称	教授
工作单位	哈尔滨工业大学		
完成单位	哈尔滨工业大学		

对本项目技术创造性贡献：

重点开展弹性文件管理、人机交互复合错误清洗、模式集成、数据源选择等关键技术的研究，提出基于人机交互的分布式大数据清洗方法和基于多目标优化的数据源调度策略，设计了弹性文件管理系统，发明了语义知识的异构模式集成方法。对第 1、2、3 和 4 项技术发明均做出了创造性贡献。

在本项目研发工作中投入的工作量占本人工作量的 80%。

六、主要完成人情况表

姓　　名	信俊昌	排　　名	5
行政职务	无	技术职称	教授
工作单位	东北大学		
完成单位	东北大学		

对本项目技术创造性贡献：

重点开展聚合优化计算引擎、复合错误清洗方法和融合处理查询分析等关键技术的研究，提出了基于消息的任务交互计算模型，设计了基于概率图的大数据清洗方法，发明了基于分布式 ELM 模型训练方法。对第 1、3 和 4 项技术发明均做出了创造性贡献。

在本项目研发工作中投入的工作量占本人工作量的 80%，获得授权发明专利 7 项。

六、主要完成人情况表

姓　　名	闻英友	排　　名	6
行政职务	东软研究院院长	技术职称	教授
工作单位	东北大学		
完成单位	东北大学		

对本项目技术创造性贡献：

重点开展基于人机交互的多源异构海量数据复合错误清洗方法中多模态数据清洗关键技术的研究，发明了一种垃圾语音信息的检测方法和装置。对第 3 项技术发明做出了创造性贡献。主要发明成果已集成于东软集团大数据系列平台产品并实现了市场推广应用。

在本项目研发工作中投入的工作量占本人工作量的 50%，获得授权发明专利 4 项。

七、完成人合作关系说明

2003 年以来，我们团队与西北工业大学李战怀教授团队、武汉大学彭智勇教授团队和哈尔滨工业大学李建中教授团队建立了长期且稳定的合作关系，开展多源异构海量数据集成技术研究。

2009 年至 2013 年，王国仁教授团队 (王国仁和信俊昌等人) 和李建中教授团队 (王宏志等人) 合作承担了国家自然科学基金重点项目“不确定数据管理的理论与关键技术”的研究工作；2012 年至 2014 年，王国仁教授团队(王国仁和信俊昌等人)、李战怀教授团队 (刘文洁等人)、彭智勇教授团队 (彭煜玮等人)和李建中教授团队 (王宏志等人) 合作承担了国家高技术研究发展计划 (863 计划) 课题“海量异构数据集成管理与分析技术及应用/2012AA011004””的研究工作。王国仁教授团队(王国仁、信俊昌和聂铁铮等人)、李战怀教授团队 (刘文洁、潘巍等人)、彭智勇教授团队 (彭煜玮、刘斌等人)和李建中教授团队 (王宏志等人)一起获得了 2016 年度教育部科技进步一等奖。王国仁教授团队 (王国仁和信俊昌等人)与闻英友等人联合东软研究院，研发了多源异构海量数据集成平台，并在东软集团、上海宝信、国家海洋信息中心、广东昆仑、扬州恒隆、深圳同立方、武汉图信、哈尔滨乐辰、黑龙江亦轩和宁夏理工学院等企事业单位等多家企事业单位进行了应用。

通过十余年合作研究，取得了以下创新性成果：(1) 发明了基于消息的任务交互模型和基于障栅的迭代处理机制，突破了 MapReduce 任务独立理想并行计算模型的性能制约，实现了数据依赖多样性查询和作业模式多样性查询的聚合优化处理；发明了规模感知的弹性分布式文件系统，实现了规模多样性文件的统一存储。(2) 发明了基于对象代理的多源异构海量数据集成机制，提出了基于语义知识的海量异构模式集成方法，突破数据源模式多变、结构灵活、源端伸缩带来的挑战，实现了多源异构海量数据的动态集成。(3) 发明了基于人机交互和分布式计算的多模态数据清洗算法，提出了共生错误最优化清洗流程，突破了模态多样、错误共生、数据高熵的桎梏，实现了多源异构海量数据的快速有效清洗，准确率可达 90%以上，破解了复合错误修正的难题。(4) 针对查询分析中的鲜效权衡问题（结果时鲜与查询效率间的矛盾），发明了基于多目标优化的在线-离线协同调度策略和基于鲜效协同保障的多源结果渐进融合方法，有效解决了源端差异导致的性能劣化问题；发明了最小化冗余访问的在线-离线协同优化机制和基于分布式 ELM 的分析模型训练方法。实现了百万量级数据源查询分析的鲜效权衡。

综上所述，王国仁、李战怀、彭智勇、王宏志、信俊昌和闻英友在多源异构海量数据集成方向形成了稳定的合作关系。

承诺：本人作为项目第一完成人，对本项目完成人合作关系及上述内容的真实性负责，特此声明。

第一完成人签名：