# 上海市科技成果转化和产业化
# 项目申报书

**（V1.0 版）**

项目名称 支持数据交易的大数据试验场关键技术与工具集研制

所属指南（目录名称）_____

推荐渠道_____

推荐单位_____（盖章）

推荐联系人_____联系电话_____

开始日期_____

结束日期_____

项目申报单位_____（盖章）

项目责任人_____

手机_____电子邮件_____

20　　年　　月　　日订

# 填 写 说 明

一、本提纲供编写上海市科技成果转化和产业化项目建议书使用。

二、项目责任人应根据本提纲要求，逐项认真编写，表达要明确严谨，字迹要清楚易辨。外来语同时用原文和中文表达。

三、申请项目资助经费在 20 万元人民币及以下时，毋须填写表 2 至表 3。

所有项目无须填报《表 5 实验动物使用情况表》

若项目涉及国际合作事项，则必须填写《表 6 国际合作基本信息表》

所有项目必须填写《表 7 知识产权基本情况表》

四、纸版材料**请使用 A4 纸双面印刷,不要采用胶圈、文件夹等带有突出棱边的装订方式，请采用普通纸质材料作为封面。**

五、报送市科委书面材料一式一份（特殊情况，另定）和电子文本一份。电子文本通过网络递交。项目申报人必须确保书面材料和电子文本的一致性。

六、本项目需要保密的，密级由项目推荐单位和申报单位提出建议。没有提出保密要求的，一般按非保密项目处理。

七、本提纲解释权归上海市科学技术委员会。

# 推荐单位声明

1、本单位认为该项目符合上海市科技成果转化和产业化项目库申报要求，推荐其申报；

2、本单位所推荐的材料不存在任何违反《中华人民共和国保守国家秘密法》和《科学技术保密规定》等相关法律法规的情况；

3、本单位所推荐的材料不存在侵犯他人知识产权或剽窃的情形。

如有不符，愿意承担相关责任。

推荐单位（盖章）：

年　　月　　日

# 目录

第 1 章　趋势判断和需求分析 ..... 4
1.1.　国内外现状、水平和发展趋势 ..... 5
1.1.1.国外现状、水平和发展趋势 ..... 5
1.1.2.国内现状、水平和发展趋势 ..... 6
1.2.　经济建设和社会发展需求 ..... 7
1.2.1.大数据发展提出的数据流通需求 ..... 7
1.2.2.支持数据交易的大数据试验场的提出 ..... 8
1.2.3.大数据试验场建设对上海科技创新中心的意义 ..... **错误!未定义书签。**
1.2.4.试验场建设对我国大数据研究与发展的意义 ..... **错误!未定义书签。**
1.3.　科学技术价值、特色和创新点 ..... 10
第 2 章　研究内容和技术关键 ..... 12
2.1.　项目目标 ..... 12
2.2.　创新点 ..... 12
2.3.　主要研究内容 ..... 13
2.3.1.研究内容一：探索性大数据分析与价值评估系统 ..... 14
2.3.2.研究内容二：数据质量评估与修复系统 ..... 17
2.3.3.研究内容三：大数据试验沙箱 ..... 18
2.3.4.研究内容四：公平可信数据交易保障工具集 ..... 20
2.3.5.研究内容五：交易数据管理和访问控制平台 ..... 23
2.3.6.研究内容六：大数据试验过程协同管理平台 ..... 25
2.3.7.研究内容七：大数据试验场管理运维平台 ..... 26
2.3.8.研究内容八：大数据众创空间 ..... 28
2.4.　关键技术 ..... 29

2.4.1. 交互式探索性分析 ..... 29
2.4.2. 大数据试验场中的数据推荐。 ..... 30
2.4.3. 应用适配的大数据试验沙箱软硬件集群自动配置技术 ..... 30
2.4.4. 支持多试验沙箱实例的全局优化的自适应资源调度 ..... 34
2.4.5. 基于 FCM 方法的多维可扩展数据质量度量 ..... 30
2.4.6. 可配置的数据质量修复融合 ..... 31
2.4.7. 电子交易多方全流程风险控制 ..... 35
2.4.8. 基于数据目录和血缘追溯的数据管控 ..... 37
2.4.9. 海量异构节点动态监控 ..... 39
2.4.10. 多属主、多粒度数据访问控制 ..... 39
2.4.11. 试验全流程安全审计监管 ..... 36
2.4.12. 多用户多模态数据接口与共享机制 ..... 41
第 3 章 执行年限和计划进度 ..... 42
第 4 章 工作条件和环境保障 ..... 45
4.1. 项目申请单位情况 ..... 45
4.2. 已经具备的实验条件 ..... 51
4.3. 项目组织机制设计 ..... 53
4.4. 产学研结合加快工作进展的设想 ..... 55
第 5 章 成果形式和考核指标 ..... 56
第 6 章 预期效果和风险分析 ..... 58
6.1. 项目成果对社会发展所起的作用 ..... 58
6.2. 经济效益和产业化前景 ..... 59
6.3. 对环境影响程度及资源综合利用情况 ..... 59
6.4. 风险分析 ..... 59
6.4.1. 技术风险 ..... 59

6.4.2. 市场风险 ..... 61

# 第1章　趋势判断和需求分析

大数据的重要性已是全球各国政府的共识。美国白宫发布的《2014 年大数据白皮书》中提到：“大数据的爆发带给政府更大的权利，为社会创造出极大的资源，如果在这一时期实施正确的发展战略，将给美国以前进的动力，使美国继续保持长期以来形成的国际竞争力。”今天的美国，从政府到企业，从医疗、教育等公共服务部门到商业、科技领域，大数据技术正在催生各个领域的变革力量，整个社会也在不遗余力地主动进行大数据技术的发展与应用。2010 年 11 月欧盟通信委员会向欧洲议会提交了“开放数据：创新、增长和透明治理的引擎”的报告，报告以开放数据为核心，制定了应对大数据挑战的战略。2011 年 11 月报告被欧盟数字议程采纳，12 月 12 日正式推进这一战略。2015 年 9 月，国务院印发的《促进大数据发展行动纲要》（以下简称“《纲要》”）5 日对外公开。《纲要》提出未来 5 至 10 年我国大数据发展和应用应实现的目标，包括 2017 年底前形成跨部门数据资源共享共用格局；2018 年底前建成国家政府数据统一开放平台。大数据时代，无论是国家、社团还是机构、企业，其竞争力将主要取决于拥有的数据规模以及对数据的分析、运用的能力。因此，推动大数据的发展具有重大的战略意义。

数据的开放、共享、流通是当前发展大数据的首要问题：一方面数据的需求者不知道数据特别是合适的数据在哪里，也不知已知数据的真实性和真实价值，更缺乏有效的数据价值发掘技术和平台；另一方面，数据的拥有者有着重重顾虑，担忧一旦数据被其他公司使用后带来的各种风险和价值损失。这需要通过技术手段打破两者的藩篱，降低数据拥有者在数据共享过程中的风险，并提高数据价值的可见程度，从而发现数据价值，促进数据流通。数据交易是数据流通的新型手段，人们可以利用数据交易中心作为数据需求方和供应方的中介服务，发布数据目录，实现数据共享。2014 年 12 月 31 日，贵州在贵阳成立第一所以大数据为命名的交易所；2016 年 3 月 29 日，浙江省批准筹建大数据交易中心；2016 年 4 月 1 日，上海市在静安区挂牌成立大数据交易中心。预计到今年底各地政府推动成立的数据交易中心将达到 15 家至 20 家。

然而数据对于用户的价值评估以及公平交易的安全保障需要合适的技术平台。因此，有必要研制支持数据交易的大数据试验场作为大数据关键基础设施。

支持数据交易的大数据试验场建设可以构建公平可信安全的计算环境，发掘数据价值，促进数据流通。通过大数据试验场提供系列工具，进行数据质量分析和价值评估，促成数据交易，从而将分散沉淀在数据拥有者手里的数据通过大数据试验场共享出来，发掘这些数据的价值，支持上海数据交易中心的数据交易工作，促进大数据技术和产业的快速发展的。研究成果有望推广到全国各个数据交易市场，形成广泛影响力。因此启动支持数据交易的大数据试验场关键技术和工具集研制具有十分重要的应用价值和时代意义。

## 1.1. 国内外现状、水平和发展趋势

### 1.1.1. 国外现状、水平和发展趋势

在数据流通和交易方面，欧美发到国家尤其是美国已经走在了前面，数据中介通过政府、公开和行业渠道，从数据源头处收集各类信息，进而向用户直接交互数据产品和服务。其中，数据源头、数据中介和最终用户构成了数据流通和交易的主体。

数据源头和中介环节构成了大数据资源的供给端。譬如 Twitter 将自身数据授权给公司 Gnip、DataSift 和 NTT DATA 进行售卖；Acxiom 等公司通过各种手段收集、汇聚关于企业和个人的信息；Sermo.com 和 Inrix 等公司则通过网络和传感器直接从公众采集数据，获得传统上单个企业难以采集的海量、实时数据。数据市场的另一端是数据需求方，包括各类数据分析服务商和行业用户，涉及政府决策、公共服务、交通物流、医疗健康、健康、人力资源、广告营销等领域。

国外的数据供给端和需求端可以概括如下特征：(1)数据中介大多与采集和聚合为主；（2）集市类的形态逐渐弱化，相关平台都最终在数据类型上有所侧重，不再以“综合类”为主要卖点。(3)用户端需求强烈、应用广泛。在引入外部数据支撑自身业务的典型案例中，Rentrak 公司基于机顶盒数据，检测各种屏幕上的媒体消费情况，为影视制作公司和广告公

司提供咨询服务；Carolinas HealthCare System 公司采集 200 多万客户的消费数据，识别其中高风险的患者；SAP 公司从运营商处手机智能手机使用信息和位置信息，并销售给市场营销机构等等。

从世界各国的实践来看，建立统一的数据开放共享平台，并集中开放可加工的数据集和工具集已经成为了一个通行做法，如美国的 data.gov 网站、新加坡的 data.gov.sg 网站、印度的 data.gov.in 网站、西班牙的 datos.gob 网站等。Google 公司内部的数据共享平台推动了 Google 公司数据产品的创新。本项目拟研制的支持数据交易的大数据试验场将成为数据开放共享和利用的高效平台，从而促进数据流通。

### 1.1.2. 国内现状、水平和发展趋势

国内大数据应用得到社会各界的广泛重视。随着数据治理理念的影响逐步加大，我国的大数据开放共享平台的进程开始逐渐加快。2011-2013 年陆续上线了国家数据（data.stats.gov.cn）、北京市政务数据资源网（BjData.gov.cn）和上海市政府数据资源网（DataShanghai.gov.cn）等大数据开放共享平台。据“开放知识基金会”发布的《2013 年开放政府数据普查》结果，在被普查的全球 70 个国家和地区政府中，我国综合排名第 35 位，这与我国经济大国和数据大国的身份极不匹配。为此，我国政府工作报告中也多次提出发展大数据，并启动了一批相关科学研究计划。2015 年 8 月，国务院印发了《大数据行动纲要》，强调要大力推动政府部门数据共享，稳步推动公共数据资源开放，统筹规划大数据基础设施建设，支持宏观调控科学化，推动政府治理精准化，推进商业服务便捷化，促进安全保障高效化，加快民生服务普惠化，明确了大数据领域的十大工程建设。同时，上海《关于加快建设具有全球影响力的科技创新中心的意见》明确指出要“实施一批重大战略项目，布局一批重大基础工程”，其中就包括大数据和云计算等。

虽然近年来，我国在数据的收集、整理等方面取得了长足的进步，但我们必须清楚地认识到，我们在可控的数据开放/流通、高效的数据利用等方面仍需坚持不懈的努力，这样才可促进“数据——技术——应用”的高速迭代发展。

2014 年起，各地开始建设大数据交易场所。2014 年 12 月 31 日，贵州在贵阳成立第一所以大数据为命名的交易所；2016 年 3 月 29 日，浙江省批准筹建大数据交易中心；2016 年 4 月 1 日，上海市在静安区挂牌成立大数据交易中心。这些交易场所的筹建为数据流通奠定了良好的场所基础，但也对如何保障和进行数据交易提出了严峻挑战。

国内数据流通环节上，普遍存在数据源活性不够、应用覆盖面较窄等问题，也就是我国大数据产业发展尚处于非常初级的阶段。在这个阶段，大数据价值发现和实现链条缺位，缺少像大数据试验场这样的基础设施帮助数据拥有者和最终用户发现并认可数据价值；此外，数据交易双方普遍对数据流通存在各种恐惧心理：包括担忧虚假数据、数据泄密、隐私泄露、数据价值流失等。

## 1.2. 经济建设和社会发展需求

### 1.2.1. 大数据发展提出的数据流通需求

大数据的开放流通对许多行业的发展都有着强力的推动作用，尤其是在社会发展、人民福祉相关的行业，数据的使用可以很大地推动对现状的了解及研究，进而做出预测及防控。以交通数据为例，随着各类交通大数据被采集并保存，以数据为驱动的交通管理智能化已经成为智能交通系统的核心。数据被视为与资本、能源同等重要的生产要素。高质量的交通数据为解决交通拥堵、改善交通服务、监控道路环境等问题提供了新的方案，是各政府与企业的重要财富并引起广泛重视。例如，大规模 GPS 轨迹数据中蕴含了群体对象的泛在移动模式与规律，有助于理解交通演化的内在机理；手机的蜂窝定位数据中包含了人群的分布、移动和相关行为等信息；通过交通卡数据可以分析公共交通流量等特征，引导公交线路、班次的优化；通过位置服务网站和社交媒体数据分析，我们可以了解用户的出行意图，实现合理的出行推荐；大型活动、事故会对交通造成影响，可以通过相关数据分析评估它们所造成的影响及其演化；同时，结合气象数据的交通分析使我们能够了解不同天气条件下的交通规律，实现更加精准的预测与导航。不难看出，交通数据具有多样、海量等特征，具有重要的分析

和应用价值。在智能交通系统构建中，人们期望能够汇总这些数据并使之高效地被处理，从中挖掘交通模式、车流规律、拥堵演化等关键知识，从而优化车辆导航、行程推荐、城市规划等业务应用。

然而在很多领域，尤其是工业领域，由于数据一直以来都存放在工业的企业内部没有得到共享，因此数据价值没能得到有效的体现。值得注意的是，工业是整个国家实体经济中的主要组成部分，同样地根据麦肯锡大数据研究报告显示，涉及工业的数据总量占整个社会大数据中的绝大部分。通过数据流通，可以使数据拥有者通过数据分享获得拥有数据的价值回报；另一方面，可以使数据获取者有更多的机会获得有价值数据。因此，如何促进这些领域的数据流通已经是大数据技术和产业发展提出的亟待解决的问题。

### 1.2.2. 支持数据交易的大数据试验场的提出

作为大数据产业的两大支撑基础设施，大数据试验场和大数据交易中心存在着各自亟待解决的问题：大数据试验场迫切需要大量的数据资源与杀手级应用以为大数据试验场的技术选型和核心技术攻关指明方向；大数据交易中心则迫切需要大数据试验场提供技术支撑，提供各类数据资源，包括真实数据、样本数据、仿真数据，保障数据交易前、中、后中对数据的质量分析、价值评估、交易风险控制、数据访问控制、隐私保护等。为此，适时启动支持数据交易的大数据试验场建设，解决其中的关键技术问题，形成相关工具集具有十分重要的应用价值和时代意义。

支持数据交易的大数据试验场是支撑用户进行基于数据组织、分析、探索及其系统架构方面试验，促进数据开放共享的公共平台。其目的是保障数据交易，促进数据流通，帮助交易双方发现数据，发掘数据价值。试验场将为从事大数据交易的企业、个人和科研机构提供大数据处理的模拟环境，以支撑用户从平台、数据、数据分析方法等方面对大数据的处理、应用和分析系统进行展开实验，保障数据交易。

简单地说，支持数据交易的大数据试验场是拥有大量、各类数据及相应的分析计算能力，能够在线开展各种研究和试验，面向全球大数据技术研究和人才培养、科技与工程创新

的开放性、领先性的重大基础设施，是上海科技创新中心建设与发展的重要基础，将成为国家大数据战略的重要组成部分。同时，支持数据交易的大数据试验场还将为上海市政府治理模式创新、民生服务创新和创新产业发展提供试验和推演场所。到 2016 年底，拟建设的支持数据交易的大数据试验场将具有处理 5PB 规模数据的能力：

（1） 支持数据交易的大数据试验场的数据储备功能可以有效解决数据门槛问题。支持数据交易的大数据试验场将构建一个公共的数据储备社区与环境：一方面，可用于政府部门及国有大型企事业单位的部分业务数据的备份与储备。按照数据开放与信息共享的原则，国有及全民共有的数据是全社会的共有资源，应该在政府监管下进行有效地管理和共享。同时，大数据试验场还可以通过专业技术搜集、整理 Web 上的公共数据资源。最重要的是，试验场可以采用数据交易获取、购买大量数据资源等用于科学研究、试验开发、大众创新等工作。

（2） 支持数据交易的大数据试验场建设为解决大数据带来的技术挑战提供了探索和示范。大数据试验场采用先进的软硬件大数据分析处理平台，并前瞻性地运用大数据技术（引领未来 5 年的大数据相关技术）搭建新型大数据试验平台，研究大数据计算框架、分析平台和应用工具，试验验证新型大数据产品。

（3） 支持数据交易的大数据试验场建设为推动大数据交易中心提供了数据、技术和运维技术基础，这将极大地推动数据开放共享，促进数据流通，提高数据利用效率，促进上海科技创新中心的建设。

大数据试验场将在数据层面通过元数据、真实数据、样本数据、仿真数据、结果数据等，为大数据交易提供数据支撑。此外，大数据试验场中探索性大数据分析与价值评估系统、数据质量评估与修复系统、大数据试验沙箱、公平可信数据交易保障工具集、交易数据管理和访问控制平台、大数据试验过程协同管理平台、大数据试验场管理运维平台，为大数据处理分析平台提供交易沙箱，支撑数据交易前的数据测试和质量验证、大数据交易中的分析处理和大数据交易后的结果验证和风险管控。这些技术和平台的聚集效应为上海大数据交易中心建设提供了坚实的技术支撑。

## 1.3. 科学技术价值、特色和创新点

通过本项目的研究开发，研制支持数据交易的大数据试验场，其科学技术价值体现在：

(1) 构建支持数据交易的大数据试验基础设施，为上海科创中心的功能性平台打下技术和系统基础。通过支持数据交易的大数据试验场建设，上海在国内乃至全世界形成公开服务的大数据试验平台。

(2) 提供创新性数据价值发现的试验环境，为大数据创新创业提供双创空间。通过支持数据交易的大数据试验场建设，为大数据探索、大数据价值评估、大数据分析提供技术和系统平台，从而支持基于大数据的创新创业。

(3) 研制安全可信公平的数据流通技术，为大数据交易提供技术支持。通过支持数据交易的大数据试验场建设，研制安全可信公平的数据交易技术和规范、数据访问和权属管控，保障大数据交易。

项目研制支持数据交易的大数据试验场，厘清并解决数据质量评估与修复、探索性大数据分析与价值评估、公平可信数据交易与交易审计、及交易试验沙箱等关键技术问题，形成关键工具集，支持数据交易，促进数据流通。项目主要创新点包括：

(1) 提出基于融合后数据的数据分析方法，解决了试验场大数据探索性分析的价值评估问题，达到了辅助用户进行数据选取的目的。

(2) 提出基于 FCM（因子准则测量）方法，维度可剪裁与扩展的数据质量度量模型与评估指标体系，解决多样化数据质量评估需求，建立可定制质量度量模型与评估指标体系。

(3) 提出基于质量规则和管道过滤架构的多算法数据修复融合方法，解决目前单一质量指标方法难以解决的质量修复问题，提高数据流通及应用价值。

(4) 提出基于异构应用模糊适配和共享状态全局调度的软硬件集群自适应配置技术，为多用户构建大数据试验沙箱专用空间，支持隔离和高效的大数据交易与数据分析试

验。

(5) 提出基于区块链技术的去中心化公平交换技术，解决公平交换对可信第三方的依赖，为数据交易及试验提供可信机制和保障。

(6) 提出基于可信审计监管机制，形成了试验虚拟环境的可信初始化以及审计数据的可信生成方法，建立可信的、具备安全隔离和全流程可追溯可取证能力的大数据试验沙箱虚拟环境，支持公平可信的交易和试验。

(7) 提出基于数据资源血缘图谱的大数据试验场数据权属管理方法，解决大数据试验过程中衍生数据的结果控制难题。

(8) 提出面向大数据试验的平台即服务的系统集成技术，解决试验场内异构工具集与系统间集成问题，简化大数据试验构建。

# 第2章　研究内容和技术关键

## 2.1. 项目目标

本项目拟构建支持数据交易的大数据试验基础设施，提供创新性数据价值发现的试验环境，形成支撑性的试验场基础平台，支持上海数据交易中心的交易活动。项目拟研制一系列创新性大数据试验关键技术，形成相应的系统及工具集，用于发现和挖掘数据价值、提供安全可信公平数据流通的保障技术，形成相关的技术规范为数据交易前、交易中、交易后各环节提供技术指导，支持数据交易，促进数据流通。项目拟探索支持创新性大数据试验的基础设施构建方法，为大数据基础设施建设形成良好的技术借鉴。

## 2.2. 创新点

项目研制支持数据交易的大数据试验场，厘清并解决数据质量评估与修复、探索性大数据分析与价值评估、公平可信数据交易与交易审计、及交易试验沙箱等关键技术问题，形成关键工具集，支持数据交易，促进数据流通。项目主要创新点包括：

(1) 提出基于融合后数据的数据分析方法，解决了试验场大数据探索性分析的价值评估问题，达到了辅助用户进行数据选取的目的。

(2) 提出基于 FCM（因子准则测量）方法，维度可剪裁与扩展的数据质量度量模型与评估指标体系，解决多样化数据质量评估需求，建立可定制质量度量模型与评估指标体系。

(3) 提出基于质量规则和管道过滤架构的多算法数据修复融合方法，解决目前单一质量指标方法难以解决的质量修复问题，提高数据流通及应用价值。

(4) 提出基于异构应用模糊适配和共享状态全局调度的软硬件集群自适应配置技术，为多用户构建大数据试验沙箱专用空间，支持隔离和高效的大数据交易与数据分析试验。

(5) 提出基于区块链技术的去中心化公平交换技术，解决公平交换对可信第三方的依赖，

为数据交易及试验提供可信机制和保障。

(6) 提出基于可信审计监管机制，形成了试验虚拟环境的可信初始化以及审计数据的可信生成方法，建立可信的、具备安全隔离和全流程可追溯可取证能力的大数据试验沙箱虚拟环境，支持公平可信的交易和试验。

(7) 提出基于数据资源血缘图谱的大数据试验场数据权属管理方法，解决大数据试验过程中衍生数据的结果控制难题。

(8) 提出面向大数据试验的平台即服务的系统集成技术，解决试验场内异构工具集与系统间集成问题，简化大数据试验构建。

## 2.3. 主要研究内容

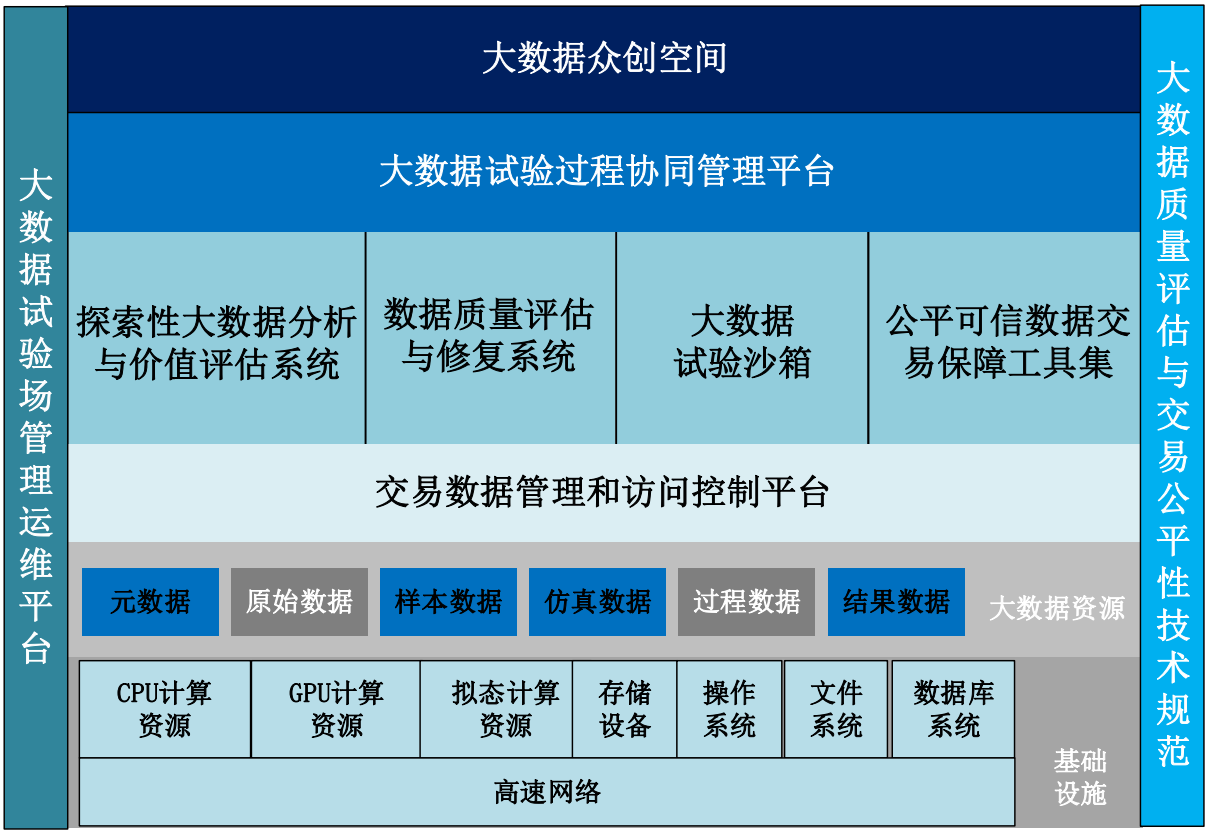


图 1 支持数据交易的大数据试验场研究内容

为保障数据交易，促进数据流通，本项目拟研制支持数据交易的大数据试验场，解决大数据试验场建设中的关键技术问题，建设面向创新性大数据试验的核心能力。项目拟重点对探索性大数据价值分析、数据质量及价值评价、数据管理与访问控制、电子交易风险

分析与控制等关键技术进行研发，研制相关系统及工具集，满足包括上海大数据交易中心在内的各类数据交易场所对于数据加工处理环节的技术和系统需求。研制的大数据试验场提供基础设施及管理系统，既可以实现数据交易前的数据检索、数据质量分析、数据试用，帮助用户发现并初步体验目标数据集；也可以为数据交易提供安全的数据分析加工环境，支持实现公平的数据交易；最后可以为数据交易后的结果验证、合同履行验证提供技术支撑。

项目研究内容针对大数据试验场的软硬件基础设施、大规模数据管理、核心分析处理引擎、协同创新、试验场运维保障等层面的需求，开展以下八方面的研究内容：(1) 探索性大数据分析与价值评估系统：支持数据交易前的探索性大数据分析与价值评估，帮助数据用户发掘数据价值；(2) 数据质量评估与修复系统:支持数据交易前的数据质量评估，并根据预设规则执行数据修复；（3）大数据试验沙箱：支持数据交易的沙箱模式，在保护数据安全性的基础上，构建动态可重构数据分析处理环境，帮助数据用户确定交易数据分析处理技术平台选型；（4）公平可信数据交易保障工具集：支持公平可信数据交易的实现；（5）交易数据管理和访问控制平台：保护交易数据在数据处理过程中的安全性、并根据数据血缘关系管控大数据试验中的衍生数据；（6）大数据试验过程协同管理平台：实现大数据试验过程中各个工具间的协同，支持实现数据交易；（7）大数据试验场管理运维平台：管理大数据试验过程中的软硬件；（8）大数据众创空间：支持数据用户创新创业，发掘数据价值。

各研究内容的层次关系如图 1 所示。

### 2.3.1. 研究内容一：探索性大数据分析与价值评估系统

如何让用户在大数据中发现有价值的数据是数据交易流通过程中的一个重要问题。试验场可以为数据用户提供两种数据发掘服务模式：用户在试验场中选取合适的算法和工具进行数据分析，或者选取多属主数据或使用系统推荐的数据，并在融合后的数据上进行交互式分析,以评价数据的价值，达到数据选取、维度选取、价值评估等，以最大化数据价值。

传统的技术方法和手段在解决上述问题时，在能力方面尚显不足：大数据试验场中存放的数据规模庞大、结构各异、内容多样，这需要新的数据分析处理手段；其次，传统的分析方法和手段会导致多次“数据选取-数据分析-结果整理”过程的迭代。为了减少迭代次数，需要探索式的交互，提供数据在选择试探方向上的选择；最后，传统的数据价值评估往往由领域专家对自身领域的数据做出评价，难以评价多领域数据融合后的价值，需使用探索式分析的方法进行价值评估。为此，本研究内容拟：

(1) 研究试验场数据发布模板设计、管理、检索和共享方法。大数据试验场中数据通过发布（即数据拥有方提出请求）方可进行交易。为此，拟面向交易过程，研究数据发布模板，设计面向数据开放和请求的发布模板设计、管理、检索和共享方法。针对拥有者未公开发布的数据，为数据使用的请求者提供匿名检索功能，检索到数据后，数据使用的请求者可定向传送交易请求。

(2) 研究统一的数据元数据表达规则描述语言和元数据管理方法。通过对数据自身的元数据和数据间关联的信息进行统一的元数据建模，用于索引存放于试验场中规模庞大、结构繁杂，内容多样的数据。这些元数据被抽象成不确定维度的多维结构。传统的元数据表达规则描述方法在针对此场景时，往往忽略了数据的维度刻画、隐形关系的提取，这使得本研究内容中的统一元数据表达和管理显得尤为重要。

(3) 研究面向大数据多维表达的数据组织和管理方法。经过建模的数据在不同粒度上具有不同的表现，而被建模的数据体积庞大且查询语言具有灵活的过滤功能。因此，需要研制面向试验场数据多维表达的数据组织和管理的算法集合，特别是研究适用于不确定维度的数据组织和存储管理方法。针对高维的非固定维度的数据，研究其高效的数据存储结构，研究数据的多粒度存储、索引和访问等方法。这些方法将被支持实现数据选择、维度选择。并且，这些方法对外表现是透明的。

(4) 研究探索性分析的数据选取技术和选取模型导入接口。研究试验场中数据粒度空间上的粒度组合的选择方法。首先，研究如何根据分析任务获得与其它颗粒具有明显差异的粒度组合。拟利用多种抽样技术获得抽样数据，并对抽样数据进行初步的聚

类，根据聚类结果组合的最小覆盖与最大包含思想生成每个类数据的描述，从而可以获取粒度组合之间的关系。然后，根据关系以及粒度的层次特点，找到结果较好且代表性强的若干粒度组合。拟设计有效性、代表性和多样性等不同角度进行数据选取的方法。为设计探索性分析的数据选取方法，拟研究粒度组合选择的计算复杂性问题，利用近似算法和流水调度等来研究粒度组合选择的优化问题，并优化探索性分析的数据选取方法。

(5) 研究探索性分析的维度获取、管理和选择方法。试验场中数据的使用者在进行了初步分析后，可能对当前的数据分析粒度组合进行细微调整。此时，使用者可以获取存放在管理器中的可选维度，并根据获取到的维度进行进一步的维度选择。从另一个角度，维度管理需为进一步分析上卷或下钻的操作提供统计数据，并由用户选择方向或根据用户指定的最优化分析目标，进行上卷或下钻的维度选择。

(6) 研究探索性分析的数据分析工具推荐技术。拟对用户的分析目标和数据进行评估，并推荐适合的分析方法和工具。

(7) 研究试验场数据推荐技术。大数据试验场中的数据拥有者、数据使用者、多粒度的数据间形成复杂的交互网络。针对该网络，利用数据发布信息、数据的粒度信息、数据的使用信息等，进行详尽的分析。根据上述分析结果，为用户的数据选择、维度选择以及进一步的数据使用提供推荐。

(8) 研究试验场数据的价值评估方法。拟从两个角度进行试验场中价值评估，一是评估固有数据的价值，二是评估数据融合后的价值。虽然，目前有第三方机构进行面向数据交易的数据价值评估，但这些评估往往在领域专家的指导下进行，尚无法直接应用于大规模数据的价值评估；更进一步，尚没有自动评价融合数据价值的方法出现。拟研究提供支持外部评估模型接入的接口，以便于使用新的、更有效的数据价值评估模型；针对众创空间等应用，研究面向融合后数据的价值评估手段，以向众创空间中的诸多企业提供更具价值的数据推荐。

基于上述研究内容，我们将研制一套大数据试验场探索性分析系统，包含试验场数据融

合和元数据管理、试验场数据的探索性分析、试验场数据的价值评估系列工具，实现探索性数据分析的数据选取、数据融合、工具选取、数据价值评估、数据推荐等功能。

### 2.3.2. 研究内容二：数据质量评估与修复系统

在大数据试验场中，数据提供方的行业多种多样，数据的来源、产生过程等各不相同，因而获取的数据可能存在各种的质量问题,包括数据缺失、数据值不正确、同一数据对象在不同来源定义和描述不同等上的不准确性、不一致性等问题。这些质量问题对数据价值有着重要影响，例如，影响数据分析结果的准确性，影响数据的最终定价，导致数据难以甚至不可应用等，给包含数据估值、数据交易等的数据流通过程、体现数据价值带来严重障碍。在大数据时代，我们尚缺乏系统化的数据质量评估方法、数据质量修复方法以及相关的系统与技术规范。因此，本研究内容将以数据交易这一应用场景为引导，研究数据质量保证技术与规范、研制数据质量评估与修复系统，其主要研究内容包括：

(1) 研究数据质量评估指标体系和方法。数据质量评估，需要针对特定的数据对象，构建数据质量评估指标体系，计算出各类评估数据质量的关键特性指标，从而对数据质量的“好”与“坏”的程度给出定性和定量评价。迄今为止，软件质量评估技术已研究和应用多年，形成了相对成熟的评估方法和标准；但对于数据，仍然缺乏系统化的评估方法，用来指导数据质量的评估。因此，本研究内容将借鉴 ISO/IEC 25024 系统与软件工程，数据质量评估的国际标准，研究多层次数据质量评估模型，数据质量评估过程框架：定义数据质量评估的主要任务及相互关系、质量评估的里程碑及确认方法，数据质量评估指标体系：质量评估不同层次的指标及相互约束关系，数据质量评估方法（质量度量的量化模型），形成数据质量评估规范。

(2) 研究数据质量修复技术。在数据质量评估的基础上，针对数据质量各个层面的问题，例如数据缺失、数据重复、异常数据值、数据之间关系不正确等，研究数据质量修复方法。在数据管理技术的发展过程中，针对不同的数据质量问题，已经有大量研究工作，如数据去重方法、异常值检测方法、缺失值插入方法等。但是，这些方法

都是通过单一质量指标入手进行质量修复，难以应用于大数据试验场数据的数据来源多样化、应用需求多元化。因此，我们不仅需要针对单个数据质量指标的修复方法，还需要针对具体的一组数据对象，利用、选择和综合多个单个质量指标的修复方法，从整体上修复数据对象的数据质量。本研究内容将研究数据质量的单质量指标修复方法，包括基于模式的数据质量修复方法和基于实例的数据质量修复方法，分别用于修复数据模式和数据实例值的相关质量问题；研究可融合多种算法的数据质量综合修复方法，支持修复策略的定义、满足的修复策略的算法选择和融合等；

基于上述研究内容，研制数据质量评估和修复系统，主要功能包括质量建模、质量评估数据管理、质量报告自动生成、质量统计和分析、质量跟踪方法、基于模式和基于实例数据质量修复的算法库、修复策略定义工具、多算法融合方法等。

### 2.3.3. 研究内容三：大数据试验沙箱

面对多个用户不同数据大小、不同性能和代价要求的大数据交易和处理请求，系统需要适配各种应用特征，动态高效的完成多种计算集群（Hadoop、Spark）的自动快速配置，形成每个用户的专用空间——称之为大数据试验沙箱，在该相对隔离的沙箱内进行安全可信和高效的大数据分析处理。因此，大数据试验沙箱是在大数据试验场共享物理和虚拟计算/存储资源基础上，承载特定大数据试验的，与其它计算资源相对安全隔离的自完整的独立虚拟环境专用空间。本研究内容围绕支持构建大数据交易沙箱，通过动态软硬件环境自动配置及协调机制，在基于虚拟化技术构建的共享 IT 基础设施上建立起高效、可信、具有安全隔离机制的大数据试验沙箱，提供大数据处理软硬件专用环境供用户使用，保证试验顺利进行，并防止敏感信息从大数据试验沙箱泄露。

本研究内容拟针对现有的大数据集群资源管理系统和应用需求特点，展开研发大数据集群资源配置和调度的关键方法与技术，实现高效全局的集群配置管理，以支持大数据试验沙箱的构建：

(1) 研究应用适配的大数据试验沙箱软硬件集群自动配置技术。系统在运行试验过程中

将搜集不同类型典型应用作业运行中各阶段的动态数据，包括数据规模和节点规模，节点配置等静态信息，运行中的网络吞吐量、磁盘读写速度、CPU 和内存占用率和最终运行时间等动态信息，首先分析大数据应用特征，形成典型聚类，例如 CPU 密集型，磁盘 I/O 密集型，网络密集型等多种类型的应用。在此基础上设计高效的系统部件性能评估模型，该模型根据已有的数据规模、集群规模和类型，与执行时间的关系，进行分析挖掘，合理估计一定规模集群处理一定大小数据处理的执行时间，评估不同规模和资源分布的集群对各种 CPU 密集型、I/O 密集型等多种典型应用运行时间的影响。在此基础上，面对多用户的请求，包括数据大小，处理时间要求和所付代价等参数，设计合理规模的 Hadoop、Spark 等集群资源自动化快速部署方法，高性价比地满足用户的安全数据交易和处理需求，以支持实现交易试验沙箱。

(2) 研究支持多试验沙箱实例的全局优化的自适应资源调度机制。面对多个用户构建多个交易试验沙箱实例的并发处理请求，系统在做资源自动化快速配置时，需要通盘考虑全局资源的调度，高效利用资源池中的计算和存储资源，例如以互补的方式将 CPU 密集型的任务和 I/O 密集型的任务放在同一个物理机上可以达到资源高效利用的目标。大数据集群的资源调度系统是对底层硬件的一种抽象，屏蔽了硬件的异构性（目前，各系统主要是对 CPU、内存、IO、磁盘进行抽象），对上层各种应用提供资源的统一调度。而在大数据试验场中需要部署和运行 Hadoop/Spark/Storm 多个计算框架，从而为每个大数据应用选择最优的计算框架。为此，本研究内容拟研究针对并行构建多个大数据交易沙箱实例的需求试验场资源的统一调度机制。资源调度的最终目的是将用户任务分配到合适的资源上，使得在满足用户需求的前提下，任务完成时间尽量小，且资源利用率尽量高，并实现资源公平分配，充分高效利用集群资源，有效提高系统吞吐量。

基于上述研究内容，研制大数据试验沙箱，支持应用适配的软硬件集群自动配置，并实现全局优化的自适应资源调度，支持多用户高效构建交易试验沙箱，在沙箱内进行安全可信和高效的大数据交易和分析处理。

### 2.3.4. 研究内容四：公平可信数据交易保障工具集

电子数据作为一种特殊的高价值标的，具有可信度低、易复制窃取、易伪造篡改、交易和试验过程难溯源、难审计验证等弱点。涉及电子数据交易的数据买方、数据卖方、交易中介、以及审计方在交易前、交易中和交易后都需要公平可信的环境和机制解决各环节的风险控制问题以保障各自的权益。大数据试验场需要提供一系列工具，为公平可信数据交易提供试验环境。在数据交易前，需要解决可信交易试验环境的建立、交付和接入问题、数据完整性的验证问题；在数据交易中，需要解决交易各方高效可靠的相互信任问题、合同的公平签署和数据的公平交易问题、交易的公开透明见证公证问题、可信的全流程审计记录生成问题；在数据交易后，需要解决数据试验结果正确性验证问题、交易全流程审计记录的可追溯、可分析、可取证问题。

目前技术方案在解决上述问题时多存在效率低下、依赖可信第三方、技术分散且缺乏系统整合等方面的缺陷。为了克服这些缺陷，本项目拟：

(1) 研究支持电子数据交易的公平机制。在电子数据交易中，考虑如下场景： A 公司欲向 B 公司购买某数据，双方对此次交易达成共识，于是通过网络进行电子合同的签订。当 A 公司把签名文件发送给 B 公司后，此时 B 公司可能反悔，于是 B 公司就拒绝将本公司的签名文件发送给 A 公司。对 A 而言，诚实遵守协议规定，却又遭受巨大损失。产生此类不公平事件的主要原因在于网络消息传输的异步性，必然存在参与交易的某一方首先发送自己的签名，可是签名发送之后，就可能面临无法接收到对方签名的风险。

要防止此类不公平事件的发生，必须制定安全有效的公平交换协议，用于保证网络环境下信息交换的安全性和公平性，使得交换的主体以公平的方式交换信息。这样，要么任何一方都可以得到对方的信息，要么双方都得不到对方的信息。可以说，如果交易能正常进行，协议保证双方都能得到各自所需的信息；如果协议异常终止，协议应当保证通信双方都处于同等地位，任何一方都不占任何优势。

本研究内容将研发高效的电子数据公平交换方法。研究将区块链技术与交易公平技术相结合，研发出同时保障交易信任机制且去中心化的交易公平机制，并制定数据交易公平性技术规范。

(2) 研究基于区块链技术的交易总账管理方法。电子数据交易往往在网络虚拟环境下运行，交易的参与方有各自的权益诉求但又往往并不互信，并且参与数据交易的各方也存在信息的不对称性问题。因此，电子数据的交易需要建立一个令交易各方互信、公平、透明的机制和环境。

区块链就是电子交易各方信任机制建设的一个解决方案。其本质是一串使用密码学方法相关联产生的数据块，用于验证其信息的有效性（防伪）和生成下一个区块（用于溯源）。区块链包含一个分布式数据库。在基于区块链技术的数据交易系统中，参与整个系统中的每个节点之间进行数据交换是无需互相信任的，整个系统的运作规则是公开透明的，因此在系统指定的规则范围和时间范围内，节点之间是不能也无法欺骗其它节点。另外，基于区块链技术的数据交易系统也将打破传统商业模式的信息不对称性。

(3) 研究数据完整性和数据试验结果验证方法与机制。数据卖方将数据托管在交易中介，但是交易中介由于技术的原因或成本的原因，可能对数据卖方托管的数据故意删减更改，造成损害。因此，本研究内容拟研发相应的高效实用机制以允许数据卖方可以随时高效地检查确认其托管在交易中介数据的完整性，并且数据中介可以随时高效地向数据卖方和托管方出具数据完整性证明，从而有效解决数据卖方/托管方和数据交易中介或数据试验场之间在这个环节的公平可信问题。

数据买方或使用方向数据试验场上传软件代码或操作指令对试验场的数据进行操作处理。由于数据由试验场管理，且用户处理数据的软件也是上传至大数据试验场并由试验场代为运行，用户会对试验场返回的数据结果存疑和不信任（即：是否对约定的数据正确运行了用户的软件代码）。因此，本项目拟研发相应的高效实用机制来使得数据使用方确信其得到的数据试验结果是正确可信的，从而解决数据买

方/使用方和数据试验方之间的公平可信问题。

(4) 研究大数据试验沙箱的安全初始化技术。本研究内容将研究大数据试验沙箱初始化技术，实现交易相关的各类试验沙箱的安全初始化和可信交付。试验正式启动之前，该技术采用虚拟化环境的可信验证方法对试验沙箱的关键部件进行完整性校验，确保试验所用的可信虚拟空间被合法建立，将各类审计数据采集模块植入虚拟空间。在试验沙箱的交付环节，该技术利用安全自动初始化方法对试验场安全机制进行初始配置（密钥设定、口令设定、初始安全策略设定等），覆盖缺省配置，并使用可信安全通道将初始认证因子交付试验参与方，使试验沙箱的初始控制权由平台运营方安全地转移到试验参与方，实现规范、可追溯的“交钥匙”流程，防止平台运营方的恶意人员劫持、复制对试验沙箱的初始控制权。在试验沙箱被安全激活后，该工具使用安全远程接入通道构建方法，为试验参与方提供远程接入试验沙箱的途径。

(5) 研究针对试验行为的多层次可信审计数据采集技术，拟解决交易过程中可信审计数据生成问题。各类审计数据采集模块在试验前被植入试验场基础平台以及大试验沙箱虚拟空间，采用网络数据截获、日志抽取、状态轮询等机制从网络、系统、存储、应用等多环节获取数据，筛选与交易流程相关的数据。通过采用高速签名和完整性保护方法，形成层次的可信的、可相互映证的多层次原始审计数据。通过审计监控数据关联分析，从形成的可信的原始审计数据中，根据数据中存在的关联性，针对试验全流程关键操作进行抽取和规整，建立针对试验全流程事件的记录视图，关联涉及相关试验的所有事件的审计记录，并提供面向试验流程关键操作的查询和统计。通过审计数据采集系统的自完整校验，可对各类数据采集模块的运行情况进行自完整校验，及时发现植入试验沙箱虚拟空间的数据采集模块被旁路、卸载、篡改和失控的情况。

基于上述研究内容，本项目拟研制公平可信数据交易保障工具集，包括电子数据公平交换系统工具、基于区块链技术的交易总账工具、研究数据完整性和数据试验结果验证工具、大数据试验沙箱的安全初始化工具、针对试验行为的多层次可信审计数据采集工具。

### 2.3.5. 研究内容五：交易数据管理和访问控制平台

大数据试验场将面向多行业的海量用户提供数据访问、交易及分析试验等服务，产生大量的数据交易、分析和敏感数据访问任务。不同数据量级、数据安全等级的任务带来了大量的数据处理、重构、分发等数据行为，这对数据资源的全生命周期管理与资源统一调配提出了极高的要求；大批量、高并发的数据访问任务给大数据试验场的数据访问执行效率、任务敏感性判别、数据安全保护机制带来了挑战；用户在获取并利用数据之后，可能产生如科研成果外泄、商业信息泄露、不正当商业行为等不可预期的后果，如何在数据服务完成后对数据权属进行有效管控是大数据试验场所需面对的严峻问题。

本研究内容将围绕着数据统一管理、数据血缘追溯管理、敏感数据访问控制等内容，贯穿大数据采集、处理、治理、分析利用、开放、交易等数据生命环节，研究并形成交易数据管理和访问控制平台，从而实现交易数据的统一管理、合理调配、高效利用与血缘追溯管理。

(1) 研究大数据试验场元数据管理方法。大数据试验场所汇聚的数据来源多样、结构不一、难以整合与管理，为了对数据资产进行统一的管理，本研究内容将研究面向结构化、非结构化数据的元数据管理技术。通过研究多源异构数据描述方法，建立多源异构数据元模型，构建大数据试验场元数据库，研究面向多源异构数据的元数据采集、维护、查询、分析、地图、容错技术，以及元数据目录管理与关联的方法，形成形成元数据管理工具，为大数据试验场数据资源管理与利用提供支撑。

(2) 研究大数据试验场数据目录管理方法。大数据试验场中的数据资源是跨行业、跨领域、动态增长的，而大数据试验场数据资源的访问与管理面临着数据难以统一管理、用户难以精准检索、更新难以同步知晓等问题，我们将研究大数据试验场数据目录管理技术，研究多源异构数据资产统一描述、管控、展示的方法，通过对数据的业务属性、数据量级等信息进行描述，根据业务逻辑对数据资产的进行目录分层，实现数据资源的分级分类开放；对数据的安全等级进行限定，建立可供开放、浏览的数据资产统一视图，形成大数据试验场数据资源目录；研究基于数据资源更新的同

步数据挂接技术，实现数据目录的同步更新；研究数据目录多级检索技术，为数据的快速检索、调阅、申请、交易提供支撑。

(3) 研究数据血缘管理与追溯方法。传统的大数据平台无法在数据使用者获取数据服务之后进行数据权属的管控，这将产生如数据滥用、隐私外泄等问题，我们将通过对交易数据资产历代数据源的计算、查询，发现该交易数据的初始来源，实现血缘追溯，并通过对该数据源的数据权属管控，实现对交易数据资产的管理。首先，本研究内容拟采用元数据管理技术，建立多源异构数据元模型，通过对数据获取、存储、重构、分发、交易等数据生命变迁周期的管理、查询，发现数据资产血统并构建形成血缘图谱模型；其次，本研究内容通过对交易数据资产的数据血统分析，基于历代数据源变迁的路径，进行面向大数据试验场交易数据查询求逆方法研究，实现基于数据血统的交易数据溯源；

(4) 研究分析大数据试验场中敏感数据所面临的安全威胁。理顺数据交易过程中的数据生命周期，在不同的环节，包括数据开放、数据清洗、数据共享、数据使用、数据聚合等环节，分析敏感数据可能遭到的安全威胁。初步的分析包括数据的污染、数据窃取、访问控制的扩权攻击、数据权益的侵占等问题。通过本研究内容，将建立大数据数据交易时的敏感数据安全威胁树。

(5) 设计支持数据交易的访问控制语言及规范。设计的语言需要有效表达对研究内容一形成的威胁的对策。我们拟扩展 XACML，一种当前最为流行的基于属性的访问控制策略语言，为大规模数据交易引入相关属性：包括对多重数据权属的支持；对时间维度、空间维度、交易环节维度、交易粒度维度等多重控制维度的广泛支持；对数据流通管控的支持、对数据血缘表达的支持等。

(6) 研究高效数据访问控制算法和可扩展访问控制执行框架。大规模的数据交易对访问控制机制的功能和性能有着较高的要求。尤其是在细粒度数据访问控制时，大规模的数据处理将对系统产生极大的性能负担，为此，我们将研制高性能的访问控制决策算法；为实现对数据流通的控制，我们将研究可扩展的数据访问控制策略执行框

架。此外，我们仍旧需要考虑不同控制粒度，不同控制层面的访问控制执行框架，在应用层、操作系统层和数据库/数据源层实现对数据的有效安全管控。

(7) 研究基于数据血缘图谱的数据交易管控方法。通过对数据血统模型、查询求逆方法、数据权限管控技术的研究，形成大数据试验场的数据血缘追溯管理方法。通过对用户数据资产的血统分析、查询求逆，发现其初始数据源，并对该初始数据源权限加以管控，实现对已交易数据资产的有效管理。

基于上述研究内容，通过对多源异构交易数据的数据管理和数据访问控制进行研究，研制交易数据管理和访问控制平台，制定支持数据交易的数据访问控制语言规范，支持交易数据从初始数据源获取、存储，到数据重构、分发、交易的全生命周期，提供面向多源异构交易数据的元数据管理功能，实现对元数据的采集、维护、查询、分析、地图、容错，以及元数据目录的管理与关联；提供数据目录管理功能，形成基于数据业务逻辑与安全等级的数据资产统一视图；建立血缘追溯图谱，实现对数据资产的寻逆溯源；提供数据访问控制与安全授权功能，实现面向敏感交易数据的对外安全授权与用户访问权限控制。

### 2.3.6. 研究内容六：大数据试验过程协同管理平台

大数据试验场中汇聚了海量行业数据、大数据工具集、存储计算资源，公众群体、政府机构、科研机构、行业企业、创新创业人群等用户将使用数据资产、大数据工具集、存储计算资源等进行计算分析。由于大数据工具集的设计开发来源多样，传统的 Hadoop、Spark 等平台无法支持这些大数据工具集的不同接口、平台兼容性要求；海量的数据分析与交易试验任务产生大量对数据资源、试验及交易工具调度的请求，这些数据分析与交易试验任务的轻重缓急、对资源数量与时间的占用各不相同，这给协同调配资源、提升大数据试验场的资源利用效率带来了巨大的挑战。

本研究内容拟通过研究大数据工具接口规范、面向多租户和多工具的协同等技术，形成统一的接口规范，研发大数据试验过程协同管理平台，解决服务难以统一标准、资源难以统一协调等问题，并与试验沙箱形成对接，为大数据试验场运营、服务提供支撑：

(1) 研究并建立大数据工具接口规范。大数据试验过程协同管理平台实现的关键研究内容是数据质量工具、数据处理工具、分析挖掘工具、算法工具等大数据工具集的接口研究。工具接口的研究即通过对平台接口规范的研究与定义，面向各类工具提供标准的平台接口规范，实现大数据试验场中的大数据工具集与平台的无缝对接。本研究内容将依托行业专家经验，研究面向大数据试验场工具集的 API 接口规范、消息接口规范、数据接口规范等规范，建立大数据试验过程协同管理平台工具接口规范体系，实现各类工具的统一接入、管理，为平台运营方、数据使用方对工具的高效调用提供支撑；同时还将研究面向第三方工具的开放接口规范，实现用户第三方工具的接入利用，提高大数据试验场的对外服务质量。

(2) 研究面向多租户和多工具的协同方法。研究面向多租户和多工具的协同的目的是为了协调不同租户计算任务中的系统资源、数据资源、工具集，实现资源的合理分配，解决资源冲突和资源不合理利用的问题。通过采用流程化的任务处理机制，结合运维平台对资源利用的监控，实现计算任务的合理安排和资源的平衡使用，通过任务优化实现计算任务的高效处理。

基于上述研究内容，研制大数据试验过程协同管理平台。通过对工具接口规范以及面向多租户和多工具协同的研究，结合对平台资源协调管理的要求，研制形成大数据试验过程协同管理平台。面向多租户、多工具的计算任务，对加工、算法工具、存储、计算资源等资源进行协同管理、调配，通过对任务的调度、执行，面向用户提供数据服务，实现用户对大数据试验场数据资产的高效检索、查阅、交易；同时还提供处理工具、分析工具等工具的目录维护与调用服务、存储服务、计算服务。

### 2.3.7. 研究内容七：大数据试验场管理运维平台

大数据试验场运营方作为海量数据资源以及数据利用环境的管理、运营方，承载着数据存储、计算、分析挖掘、交易环境管理等职责。由于大数据试验场拥有大量的主机、网络、数据库、试验沙箱等异构软硬件资源，并需要在计算任务的分配和运行过程中进行动态调整，

传统的固化监控无法满足动态监控的现实需求，需要研究面向海量节点和资源动态变化的运维监控技术。

本研究内容将研究海量异构节点动态监控相关技术，形成大数据试验场管理运维平台，对异构软硬件资源进行动态监控、管理、运维，提升软硬件资源运行效率，识别闲置资源，降低运维管理成本，提高运维整体服务水平，实现绿色节能：

(1) 研究分析大数据试验场面临的海量节点运维问题。大数据试验场面对着海量数据资源的使用，将搭建海量节点集群环境。在运营过程中不同计算框架与随需变化的应用场景对基础设施资源将产生不同的需求，这些计算节点需要根据需求的变化进行动态调整。因此我们需要研究海量异构节点动态监控技术，采用灵活的监控策略，对海量、动态调整的节点进行监控，自动识别闲置资源。

(2) 研究海量异构节点动态监控技术。海量异构节点动态监控技术旨在海量异构节点的环境下对接入节点组合运行、动态调整等行为产生的大量信息进行实时采集，并基于节点的动态调整采用灵活的监控策略，对节点运行过程中发生的异常现象进行告警提醒，并对海量节点中的闲置资源进行标注、提醒。研究异构节点监控接入技术，通过标准协议接入、日志分析与模拟控制实现对不同类型节点的监控；研究海量异构节点的的实时采集技术并基于实时监控信息进行节点冗余事件的关联特征分析，实现对海量异构节点状态的高效监控、闲置提醒、异常告警。

基于上述研究内容，研制大数据试验场管理运维平台。通过对海量异构节点的实时采集和动态监控技术的研究，形成大数据试验场管理运维平台。平台提供资源状态管理功能，面向主机、网络、数据库等异构软硬件资源进行运维数据的自动采集与状态的动态更新，实现对资源状态的实时监控；提供运维监控管理功能，实现面向大数据试验场软硬件资源的统一监控和监控告警；提供运维服务管理功能，实现对运维服务流程、运维故障事件的高效管理；提供统计管理功能，提高大数据试验场设备资源的管理效率。研制的大数据试验场管理运维平台将对大数据试验场软硬件设备环境进行有效地管理、使其更加适应业务持续变化的需求，不断提高运维质量，实现高效运维，提升运维服务满意度。

### 2.3.8. 研究内容八：大数据众创空间

大数据试验场运营需要行业数据作为基础支撑，实现不同类型的、多源异构数据的数据交易及数据分析试验。大数据众创空间平台围绕大数据产业生态链，汇聚各领域数据样本，打破大数据信息交流阻碍，探索生态环境建设，支持创新创业，探索众创空间服务模式。以重点领域产业调研为基础，依托上海大数据联盟聚集效应，构建数据目录资源库、样本数据资源库、数据加工工具库和政策资源库的“四库”，基于试验场相关技术实现多源数据采集、存储、加工处理、服务对接。构建面向多种机构、不同创客等创新创业团队的合作交流平台，提供整合性大数据政策、产业资讯、各类服务信息，提供创业发展的资源支持，推动大数据产业生态建设和大数据团队创新创业。

(1) 研究多渠道数据汇聚机制。由于标准缺乏、信息不对称等原因，数据拥有方对通过公共服务平台共享数据顾虑重重，梳理大数据应用现状及发展趋势，关注行业企业大数据重点应用及热点领域，描述分析具有典型意义的应用案例，从数据来源、数据分析、数据价值等方面展示大数据产业发展现状，探讨大数据研究面临的科学问题和技术挑战，为大数据试验场建设提供数据资源方、应用需求方等产业布局信息。

跨行业数据融合应用必须以多源数据共享为基础，拟发挥上海大数据联盟聚集效应，汇聚成员单位所在行业样本数据，逐步拓展平台数据资源，通过平台建立数据拥有方与需求方有效沟通渠道，探索数据跨界融合应用潜在需求，形成数据促进平台、平台推动数据良性发展通道，为大数据试验场建立数据源、数据交互、数据应用的实际运作基础。

(2) 研究多层次数据共享机制。数据资源存储必须以可信云平台为基础，确保数据安全独立的前提下，实现多层次数据共享。基于试验场软硬件基础设施环境，研究面向云平台服务、海量数据存储、大数据共享和分析处理的高性能资源升级方案，提供应用示范基础环境资源，从硬件设施、系统环境到应用平台建立信任关系，实现可信数据存储、流通、共享、交易和使用，建立大数据试验场窗口示范环境。

研究支持多层次数据共享机制，为数据来源方提供安全、稳定、高速、便捷的数据传输通道和存储机制，提供数据验证服务，确保数据有效性与一致性；为应用需求方提供直观、可靠的样本数据展现方式和访问渠道；为基于样本数据开展工具研发团队提供安全、稳定的测试与应用环境。样本数据拟通过两种方式实现共享：一是平台提供数据算法链接，推送至数据拥有方；二是样本数据存储在云服务平台，对大数据试验场其他功能模块提供数据服务支撑。

基于上述研究，研制大数据众创空间平台，建立数据用户多渠道沟通交流。依托上海大数据联盟资源，通过社交媒体（如微信公众平台）建立多方沟通交流平台，促进数据融合应用。基于可信云平台实现代表性行业企业、科研院所样本数据配置与共享，开发满足不同数据源要求的数据共享接口，提出多领域数据共享与互通解决方案，为科研创新、团队创业提供服务支撑，对多个数据模型进行管理和分布执行，实现元数据查询、试用数据管理、应用运算执行、隐私安全保护等方面功能。在平台上，通过样本数据全生命周期应用示范，探索行业通用数据标准中的数据需求定义、数据清单产生、数据标准制定方法和流程，通过平台建设，聚合科研、产业、政府机构等各方力量，促进和规范大数据交易。试点特定行业标准体系，推动标准体系和交易规范建设，实现资源聚集，辅助产业发展，满足多样化的数据流通及应用需求，为大数据试验场研制提供探索与支持。

## 2.4. 关键技术

### 2.4.1. 交互式探索性分析

试验场大数据探索性分析技术不仅包括试验场数据的组织、存储、计算和分析，还包括了数据的选取、工具的选取、价值的评估等。在数据组织建模和组织过程中，需要充分考虑数据融合的粒度感知问题，即建立大数据中的实体及属性到多粒度层次上的映射。现实数据的多样性决定了映射会具有模糊特性。建立多源、异构以及碎片化的大数据具有多粒度的统一描述后，试验场大数据在粒度层次上具有多层次、模糊的特点，且面临多个粒度共存与数

据场中的使用问题。多粒度层次在不同维度上形成巨大的搜索空间，而数据选取、维度选取等任务需要在这个巨大的空间中进行搜索。我们可以利用工具交互式获得用户的需求（这些需求会缩减搜索的空间范围），但巨大的搜索空间会给数据选取、数据分析带来巨大的挑战。另一个现实问题是，试验场中的同一份数据可能在不同粒度上进行了存储，在方法上需根据分析任务选取在哪些颗粒上进行分析，将不断缩减搜索空间，也为分析带来挑战。

交互式探索性分析这一关键技术的研究将解决研究内容一（探索性大数据分析与价值评估系统）的关键技术问题。

### 2.4.2. 大数据试验场中的数据推荐

如何对大数据试验场中的数据和数据关联方式进行推荐是本项目的关键技术问题，对于发掘数据价值具有十分重要的研究意义和应用价值。在大数据试验场中，数据拥有者、数据使用者、多粒度的数据间形成复杂的交互网络。针对这一复杂的交互网络，利用数据发布信息、数据粒度信息、数据使用信息等，进行详尽的分析。对数据的类型、粒度上的价值等进行深入分析，形成价值模型，并利用该价值模型对数据、工具、潜在使用者、潜在使用方式等进行全方位的推荐，从而发掘数据价值。

大数据试验场中的数据推荐这一关键技术的研究将解决研究内容一（探索性大数据分析与价值评估系统）中的关键技术问题。

### 2.4.3. 基于 FCM 方法的多维可扩展数据质量度量

对于数据质量的度量，目前没有形成一个权威性的数据质量标准模型或参考模型。大多数据质量的研究都是针对很单一的问题进行的，解决系统中的比较重要的质量指标，如一致性问题、完整性问题、重复性问题等，比较系统的研究也只是提出了质量建模的观点，并给出了建模的步骤，但没有提出数据质量模型的体系结构，没有形成系统化的数据质量评估指标。尽管在数据建模理论中，对参照完整性、一致性等指标的定义已经非常统一和严格，但这些指标只是数据质量复杂的指标中的一小部分。可见，从不同的角度对数据质量衡量的指

标是不一样的。 存在以上问题的主要原因是，当前的研究大多是针对数据库的某一个或几个方面的质量需求进行研究的，提出的质量描述形成的是单一的质量模型，不能构成完整的质量体系。这样单一的质量模型难以满足数据质量方面的需求。

为此，我们拟基于 ISO/IEC 25024 的相关标准为基础，根据实际应用需求不同，基于 FCM（因子准则测量）方法，多维度制定可扩展评估指标体系，建立数据质量度量模型。

在 FCM 方法实施过程中，我们采用基于复合因子评估矩阵的数据质量质量评估方法，其中复合因子矩阵描述了多因子对数据质量的内在约束，这些约束来源于数据本身的内在信息、数据应用需求及特点等方面。该复合因子矩阵的构建可以先分别研究每个因子与数据评估指标之间的因子评估矩阵，选择构建因子评估矩阵的方法时应充分考虑每个因子的特点，例如考虑数据内在信息因子的影响就不能选用具有主观因素的方法。然后综合各个因子评估矩阵得到复合因子评估矩阵，最后通过建立评估映射关系获取数据质量的评估结果。该质量评估方法既可以改善客观因素的片面性，又可以考虑决策者的主观需求，可以充分挖掘满足应用需求下的数据质量的潜在信息，更好地提高数据质量评估的可信性与适应性。

通过 FCM 方法的应用，我们解决了多样化数据质量评估体系的需求，建立维度可剪裁与扩展的数据质量度量模型与评估指标体系，从而解决研究内容二（数据质量评估与修复系统）中的关键技术问题。

### 2.4.4. 可配置的数据质量修复融合

针对单一指标的质量修复方法难以解决数据交易及数据试验背景下的数据来源多样化和应用需求多元化问题，着重研究可配置的数据质量修复融合方法，在数据质量评估结果的基础上，统一定义数据质量修复策略，针对不同的质量问题，可自适应、动态组合多种质量修复方法，对数据质量进行综合修复。该关键技术主要包括：

1）数据修复策略定义语言

针对数据评估报告提出的各种质量问题，在分析不同质量修复算法特点的基础上，研究基于 XML 的数据质量修复策略定义语言，形成可灵活定义和配置不同数据质量

修复算法之间相互协同的配置文件。

2）数据质量修复融合架构

研究可动态组合不同种类数据质量修复算法的灵活架构。主要包括：通过对各种算法的抽象和封装，实现修复算法的模块化；建立数据质量修复的管道过滤器体系结构，将各种算法以及算法之间的接口转换作为相对独立、可复用的对象，实现算法模块的动态组合。

3）数据质量修复的多算法融合方法

通过基于启发式和基于规则的方法，根据不同质量修复算法的特点、基于 XML 的数据修复策略配置文件，自动或半自动选择不同种类的质量修复算法，实现可自适应的多算法融合，综合修复数据评估报告提出的质量问题。

可配置的数据质量修复融合这一关键技术的研究将解决研究内容二（数据质量评估与修复系统）中的关键技术问题

### 2.4.5. 应用适配的大数据试验沙箱软硬件集群自动配置技术

在大数据交易试验沙箱内需要部署与大数据交易和处理试验相关的多种资源，主要包括数据分析工具集、集群计算框架 Hadoop/Spark/Storm 和集群计算节点、存储节点和虚拟子网的综合体。面对多个用户的数据处理请求，配置大数据交易沙箱的目标必须实现 CPU、FPGA、内存、拟态计算资源、存储和网络资源的平衡，适配各种应用特征，并满足处理时间和安全隔离的要求。因此，选择多大规模的 Hadoop/Spark/Storm 集群快速自动配置构建大数据交易沙箱，包括虚拟计算节点或物理计算节点的数目，节点的 CPU、内存和硬盘配置，存储的配置，节点之间的网络带宽和虚拟子网配置以及选取的工具，在此配置下高性价比地完成数据处理是本项目的一个关键问题。

首先我们研制大数据应用特征的提取聚类技术并完成应用程序行为的建模技术。不同的大数据应用通常具有不同的资源使用特征。如 CPU 密集型、I/O 密集型或网络流量密集型。在传统的计算环境中，由于购买的硬件服务器其配置不能动态调整，经常会观测到不同资源

的使用不均衡，间接导致了资源浪费。在虚拟和物理混合的大数据计算环境下，由于虚拟资源分配可以细粒度的调整，通过对应用资源使用特征进行建模，可以更有效地进行集群的合理配置。同时，当前的大数据应用一般来说都是由多个组件组成，比如典型的 MapReduce 应用一般通常有 Map，Shuffle，和 Reduce 等的工作组件，这些组件的工作负载通常有着内在的联系，在大数据处理环境下采用合适的数学模型对多计算组件应用的资源需求进行分析与建模，可以形成指导性的集群资源配置方案，以达到消除性能瓶颈、保证系统的 SLA、提高资源利用率的目的。主要关键点包括：（1）应用的自动资源使用特征提取和聚类。（2）多组件大数据应用的服务模型研究。对大数据应用程序行为的建模，我们将结合数据挖掘和模式匹配方法来进行。对大规模的系统监控数据，需要使用数据挖掘的方法，从中提取出有效的信息。在大数据应用程序的建模过程中，我们重点采用模式匹配的方法，对其特征进行建模，在建模的过程中进行模式匹配。

在多层次的大数据应用程序体系结构中，应用由不同的工作组件所构成，而不同应用的组件之间又因为资源的共享而形成了有向图的关系，所以，我们主要采用图论方法，对应用程序组件构成的有向图进行分析，自动化生成上游组件如 Map 与下游组件如 Reduce 的关系，从而能够分析最上游的负载特征与最下游的资源情况之间的对应关系。

为了解决资源的自动配置问题，在应用特征聚类和行为建模技术研制基础上，我们研究多个层次的大数据性能预测和评估模型。现在有许多性能预测和评估模型方面的研究，可以使用机器学习的方法、花费模型等来预测作业的完成时间。有学者建立了较为完备的线性数学模型，我们在此基础上，利用历史作业执行时间的数据，选取合适的特征值，建立简单线性和局部线性加权的混合模型，来预测作业的执行时间。利用已有的历史作业信息，将输入数据大小，作业的资源配置和执行完成时间等信息设置为特征值来建立通用的 MapReduce 和 Spark 任务的执行时间模型。在解决负载与资源的拟合关系时，我们主要采用回归分析法来进行。大数据应用的特征是各不相同的，应用程序之间也有关联性，需要采用一种统一的方法就应用的负载特征与底层的资源使用情况进行拟合，得出系统普适的模型，来预测未来资源的使用情况。对简单的应用，我们主要采用线性回归和渐进式线性回归并结合反馈算法来

进行。对于应用的耦合度较高的系统，我们研制采用更高层次的回归分析算法来进行建模。在建模过程中，结合自适应和反馈的过程，达到建模的自动化与实时性的要求。

总体上，在给出处理数据大小，执行时间要求的前提下，并结合应用的资源使用特征，快速推算需要的资源配置情况，并根据大数据交易试验的约定，自动选择和部署相关的试验工具，进行相关的适配，验证所有工具的完整性，确保经过认证和允许的工具/代码进入沙箱运行，限定经过验证的工具的操作行为，实现有序的工具和资源整合配置，最终构建安全高效的大数据交易沙箱，解决研究内容三（大数据试验沙箱）中的关键技术问题。

### 2.4.6. 支持多试验沙箱实例的全局优化的自适应资源调度

面对多个用户的并行构建多个大数据交易沙箱实例的要求，系统需要进行全局优化的资源调度，达到保证任务质量和资源高效利用的双重目标。近年来，有一些同时考虑计算资源和数据资源综合调度的系统，典型代表有 Apache YARN(Yet Another Resource Negotiator)，Facebook CoronaL 和 Berkeley Mesos. 这些开源的系统都是为了解决编程模型和计算框架多样化环境下，不同框架间的资源隔离和共享问题，我们的大数据试验场是一个共享集群——多个不同交易沙箱部署和运行在同一个试验场集群上。集群共享可以提高资源利用率，并降低系统硬件成本和运维成本。资源调度是根据一定的资源使用规则，在不同资源使用者之间进行共享集群全局资源调整的过程。不同的计算任务对应着不同的资源使用者，每个计算任务在集群节点上对应于一个或多个进程(或者线程)。资源调度的目的是将多个用户的试验沙箱分配到合适的资源上，使得在满足用户需求的前提下，任务完成时间尽量小，且资源利用率尽量高。

我们首先研究提出了全局优化的自适应资源调度模型和技术。资源调度最终要实现时间跨度、服务质量、负载均衡、经济原则最优的目标。资源管理调度模型按照调度实体之间的关系可以分为统一资源调度模型和多资源调度协作调度模型. 按照资源的组织调度形式可分为统一集中调度模型、层次调度模型和非集中调度模型。我们研究提出新型的基于共享状态的全局优化调度模型(Shared State Scheduler)，共享状态调度是将双层调度中的中央式资

源调度模块简化成一种持久化的共享数据，这里的“共享数据”实际上就是整个集群的实时资源使用信息. 一旦引入共享数据后，共享数据的并发访问方式成为全局优化调度的核心模型。

在全局优化的资源调度模型研究基础上我们重点研究攻关大数据作业的自适应调度机制。在确定了集群资源分配之后，分布式共享集群一般采用 FIFO(First In First Out)的简单作业调度机制，为了克服单队列 FIFO 调度器的不足，多种类型的多用户多队列作业调度器相继出现。这些调度策略允许管理员按照应用需求对用户或者应用程序分组，并为不同的分组分配不同的资源量，同时通过添加各种约束防止单个用户或应用程序独占资源，进而满足多样化的 QoS 需求。当前主要有两种多用户作业调度器的设计思路：第一种是在一个物理集群上虚拟多个子集群，典型的代表是 HOD(Hadoop On Demand)调度器；另一种是扩展传统调度策略，使之支持多队列多用户，不同的队列拥有不同的资源量，可以运行不同的应用程序，典型的代表是 Yahoo 的 Capacity Scheduler 和 Facebook 的 Fair Scheduler。基于 Capacity Scheduler 和 Fair Scheduler 的混合调度机制，我们提出自适应混合作业调度器(Adaptive Hybrid Scheduler)，是以满足用户期望作业运行时间为目标的调度器。该调度器根据每个作业会被分解成多个任务的事实，通过已经运行完成的任务的运行时间估算剩余任务的运行时间，进而使得该调度器能根据作业的进度和剩余时间动态地为作业分配资源，以期望作业在规定时间内完成，从而解决研究内容三(大数据试验沙箱)中的关键技术问题。

### 2.4.7. 电子交易多方全流程风险控制

针对电子数据交易及试验参与多方在交易前、交易中和交易后的公平可信权益诉求，将密码技术与系统技术创新整合，建立电子数据安全交易多方全过程风险控制系统。

数据交易前的关键技术包括数据持有证明技术。数据持有证明技术基于密码技术和可累积的消息认证码技术对用户的数据进行组织和管理，建立相应的文件系统，允许数据卖方和托管方可以随时高效地维护并检查确认其托管在试验场数据的完整性，为数据卖方与数据交易中介方之间的公平可信提供保障。

数据交易中的关键技术包括区块链技术和公平交换技术。区块链技术包含一个去中心化的分布式数据库系统，其本质是一串使用密码学方法相关联产生的数据块，用于验证其信息的有效性（防伪）和生成下一个区块（用于溯源）；任何数字资产的认证，记录，登记，注册，存储，交易，支付，流通，一个账本统统解决。公平交换技术保证要么任何一方都可以得到对方的信息，要么双方都得不到对方的信息，并且没有任何其他方能获得关于协议双方交换内容的任何信息。但是，目前的去中心化的区块链技术没有考虑数据的公平交换，而实用公平交换技术往往依赖可信第三方。本项目将区块链技术与交易公平技术相结合，以同时保障交易信任机制以及去中心化的交易公平机制，形成一个整合的电子数据公平可信交换系统，为数据交易过程中的公平可信提供保障。

数据交易后的关键技术包括动态简洁知识证明技术。在大数据试验环境下，数据使用方通常将其数据,记为 x，托管存放于数据试验场，然后上传数据操作软件和指令，记为 f，对数据进行操作和试验,并得到试验和处理结果 y。正确的数据处理结果应该是 y=f(x)。但是，数据试验方由于技术或成本原因可能返回的是 y’=f’(x’)。基于动态简洁知识证明技术的数据试验结果验证工具在返回试验结果 y 的同时提供一个关于 y=f(x)的证明Π。为了减轻用户的计算负担，要求验证Π正确性所需的时间远远小于计算 y=f(x)的时间。这使得数据用户以很小的代价就能确信其得到的数据试验结果是正确可信的，从而为数据使用方和数据试验方之间的公平可信提供保障。

电子交易多方全流程风险控制这一关键技术的研究将解决研究内容四（公平可信数据交易保障工具集）中的关键技术问题。

### 2.4.8. 试验全流程安全审计监管

大数据试验场需要为大数据交易相关的试验提供全流程安全审计监管机制，在试验正式启动前，安全可信地将安全审计工具部署到试验沙箱中，并将试验沙箱初始控制权交付到试验参与方；在试验过程中，持续生成多方可信的原始审计数据，为试验全流程的回溯与取证服务提供支持。为实现以上需求，需要研究试验全流程安全审计监管关键技术，重点突破以

下两个机制：

（1） 大试验沙箱虚拟空间初始镜像安全初始化机制：在大数据试验沙箱虚拟环境初始化环节，采用虚拟化环境上的可信验证方法对试验虚拟空间初始镜像进行可信性校验，并植入抗卸载、抗旁路的可信审计数据采集工具；基于规范化声明实现独立于运营方的初始安全因素生成及自动化配置；采用可信通道实现初始安全认证信息的打包交付，并为试验参与方提供安全远程接入。该机制实现了试验虚拟空间初始控制权的可信转移，防止平台运营方的恶意人员劫持、复制对交易试验沙箱的初始控制权。

（2） 虚拟化环境中的多层次数据高速采集及签名机制：通过已在虚拟化环境中植入可信审计数据采集工具，从各层面采集原始数据；采用虚拟网络系统全流量数据捕获与即时排重与筛选方法实现高速数据过滤；采用高速数据标注与签名方法，形成多层次的可信原始审计记录；通过面向试验全流程的关键操作还原和关联抽取方法，在离散的、不可直接理解和分析的原始审计记录中形成可以进行回溯和分析的、针对试验全流程关键环节的事件记录。该机制解决了试验全流程可信审计数据的生成问题，为实现试验全流程的回溯和保护隐私的取证服务提供有力支持。

试验全流程安全审计监管这一关键技术的研究将解决研究内容四（公平可信数据交易保障工具集）中的关键技术问题。

### 2.4.9. 基于数据目录和血缘追溯的数据管控

在数据交易的过程中，数据资源被重复使用，分析结果被再利用的情况将普遍存在，在某一个数据资源发生变化（如：失效、禁用、权限变更、隐私泄露等）时，将会涉及到一系列的查找和追溯问题，通过交易日志追溯将是一个复杂和漫长的过程，且识别困难、容易发生遗漏。我们通过建立大数据试验场数据目录、数据交易指纹、数据资源血缘图谱的方式，提供一种快速便捷、无遗漏的管理与追踪方式，实现对数据交易的管理。

- 大数据试验场数据目录

针对数据资源的行业、业务属性等不同特征，进行分级分类，构造可灵活定义的大数据

试验场数据目录树，数据资源将依据属性挂接到不同节点中，将大数据试验场的各数据源与数据目录树中的各节点进行关联，将更新的数据自动匹配到目录节点中，实现数据目录的自动化扩展，根据数据安全等级设置开放级别，面向不同权限的用户进行开放。用户可通过节点、数据表等不同粒度进行数据资源的精准检索。

- 数据交易指纹

针对每一次数据交易，记录交易所涉及的数据资源、交易时间、关键数据集、分析结果集等数据交易资源，通过 Hash 算法产生数据交易指纹，将数据交易指纹和数据交易资源进行分离存储，通过数据交易指纹可以唯一追查到确定的数据交易。存储方式可以采用集中存储或分布式存储（如：区块链等）。

- 数据资源血缘图谱模型

对数据资源目录中的数据集和数据项、数据交易产生的数据集和数据项等元数据，根据时间序列和关联关系，建立数据血统模型，形成父-子关系的树状血缘图谱，图谱中任一节点均可追溯其亲代和子代，从而识别任一数据资源的产生和被利用的数据链，在数据资源发生变化是能够实时识别其影响的深度和广度，从而加强对数据资源使用的管理。

- 数据交易血统溯源

通过建立数据资源血缘图谱，将数据交易指纹关联到图谱中，形成多层次的网状结构，任一数据资源均可检索到其本身和子代所关联的数据交易，任一数据交易均可检索到利用其结果集的所有数据交易，使得所有交易不再是孤立的存在，从而实现全面的管理。

- 基于血缘追溯的权限管控

通过数据资源血缘图谱，在任一数据资源节点建立细粒度的权限控制，从而使得其子代继承其权限和属性，在数据权限实时管控引擎中对数据交易资源访问进行精细的权限控制，避免因数据源头追溯不当而引起的安全问题。

基于数据目录和血缘追溯的数据管控这一关键技术的研究将解决研究内容五（交易数据管理和访问控制平台）中的关键技术问题。

### 2.4.10. 多属主、多粒度数据访问控制

数据访问控制利用预定义的柔性安全策略，自动保障大数据试验过程中的交易数据的安全，因而是保障大数据试验场安全的核心关键技术之一。针对交易数据权属复杂、控制粒度需求变化多样，提出的多属主、多粒度访问控制机制通过访问控制策略语言表达交易数据所有者(们)对数据访问控制意图。与传统的企业数据安全和个人隐私保护方法不同点在于，当支持多个交易数据拥有者后，其策略的管理和执行需要通过对多个管理者的协同完成。这种协同需要设计侵入式协同机制，比如必须有多个管理者完成确认后，交易数据才可以被选取和分析，也可以是非侵入式的，比如只需要对交易数据访问进行日志，并通知多个管理者。该访问控制语言将扩展自基于属性的访问控制模型，支持多种数据类型、多种交易方式和多种权属模式的数据访问控制。

围绕多属主、多粒度访问控制机制的效率优化问题，项目将在访问控制决策引擎中采用多种缓存机制，包括策略缓存、属性缓存、访问令牌缓存方式，优化决策算法，降低访问控制，尤其是细粒度访问控制对系统性能的消耗。

为支持包括数据层、系统层和交易层的复杂访问控制，项目将采用一种语言，多种执行引擎的方式设计部署实施。利用访问控制策略语言表达数据交易试验中的数据保护意图，通过自动化工具，针对不同的系统层面编译到高效策略执行代码，通过可扩展的策略实施引擎，实现多属主、多粒度访问控制。

多属主、多粒度数据访问控制这一关键技术的研究将解决研究内容五（交易数据管理和访问控制平台）的通用数据访问控制和基于数据血缘的数据权属管理等关键问题。

### 2.4.11. 面向大数据试验平台即服务的系统集成

大数据试验场中各项研究内容产生的工具集具有结构多样、平台异构等挑战，如何方便地将这些工具集成起来，与大数据试验场中的基础软硬件一起构成一个统一的计算环境，满足大数据试验的需求，是本项目的一个难点技术问题。

面向大数据试验的平台即服务的系统集成首先研究面向多租户和多工具的协同等技术，为大数据试验平台中部署的数据质量工具、数据处理工具、分析挖掘工具、算法工具等大数据工具集形成统一的接口技术规范，并通过系统集成技术实现大数据试验场中的大数据工具集与平台的无缝对接。

针对大数据试验场中的多租户现象，设计不同租户计算任务中的系统资源、数据资源、工具集的分配机制，实现高效计算资源分配，从而实现平台即服务的系统集成。特别是，利用基于规则的方法解决资源冲突问题；通过采用流程化的任务处理机制，结合运维平台对资源利用进行监控，检测资源使用的不合理利用问题，实现计算任务的合理安排和资源的平衡使用。进一步通过任务优化实现大数据试验场中计算任务的快速部署和高效处理。

面向大数据试验的平台即服务的系统集成这一关键的研究将解决研究内容六（大数据试验过程协同管理平台）的关键技术问题。

### 2.4.12. 海量异构节点动态监控

大数据试验场中，拥有大量的异构软硬件资源，包括主机、网络、数据库、虚拟机、大数据存储以及相关的应用系统和工具集，这些资源根据不同的大数据分析计算需求，随时可以组成不同的计算节点，完成计算任务。这些节点的组合运行、动态调整产生大量的监控信息，一方面需要接入这些异构节点，实时处理海量的运行监控事件，另一方面需要面对节点的动态调整采用灵活的监控策略。

- 异构节点监控接入

针对不同类型的节点资源，采用不同的监控接入方式：

1) 标准协议接入：对那些提供开发协议或监控接口的软硬件设备，我们遵循这些开发协议（SNMP、SYSLOG、SMTP、SSH 等），实现服务等对产品进行监控。
2) 日志分析与模拟控制：对提供开放日志的软硬件设备，我们采用日志分析的方式进行监控；对既不支持标准协议、又不提供开放日志的产品，我们采取模拟控制、网络数据分析等多种手段，对产品进行监控。

- 海量监控数据实时采集分析

采用分布式监控代理，通过异步和流式处理对监控数据进行过滤和归并，根据目标节点信息和监控规则，对来自各种节点资源的告警与监控日志等的大量监控信息进行校验和合并，从监控类型、空间和时间特征 3 个方面对冗余事件关联特征进行系统的分析，从而提取出准确的、精简的监控事件，使得海量的报警信息能够快速处理。通过分布式监控和集中分析，使得在海量的监控信息中识别出关键信息，并进行预警和报警。

- 自适应动态监控策略

针对每一项数据交易所涉及的节点，自动建立监控策略，并根据不同类型的数据计算设定各项监控参数，在交易运行过程中的弹性资源自动调整监控的对象，自动识别闲置资源，自适应资源的变化。通过监控视图建立全局的监控态势，报警并隔离发生故障的资源，并对资源的使用进行预警预测，评估资源故障对相关资源和交易的影响，保障数据交易的正常运行。

海量异构节点动态监控这一关键技术的研究将解决研究内容七（大数据试验场运维平台）中的关键技术问题。

### 2.4.13. 多用户多模态数据接口与共享机制

针对数据融合应用需求，实现多源异构数据共享，设计开发第三方数据交换公共认证数据 API 接口，在改进传统 XML 数据交换模式的基础上，拟采用双 XML 关联配置模式和基于 Web Services 数据交换解决方案，设计数据交换规约配置框架，探索数据交换过程同步性与一致性问题，给出多级耦合关系的数据交换处理机制，保障关联数据交换的完整性，支持平台用户间多源异构信息交互，探索具有一定通用性的数据标准。基于数据交易接口定制及商业化数据服务模式，以标准化的通讯协议针对特定领域开发友好便捷的数据共享接口，为不同数据源的展示、交互与共享创造便捷条件。通过远程过程调用（RPC）、标准查询语言（SQL）、文件传输、信息交付等步骤实现应用开发方与数据资源方之间的数据通信，通过开发数据 API 接口，定义系统接口标准，满足开放式 API 功能要求，共享数据缓

存器、数据库结构、文件框架，以标准化的通讯协议针对全平台提供友好便捷的数据共享接口。

多用户多模态数据接口与共享机制这一关键技术的研究将解决研究内容八（大数据众创空间）中的关键技术问题。

# 第3章　执行年限和计划进度

项目执行期为2016年7月至2018年6月。项目计划进度如下：

■ **2016年，第3季度**

**时间**：2016年7月－2016年9月

**任务：完成大数据试验场和大数据交易中心的技术调研，形成支持数据交易的大数据试验场技术路线图，开展初步的技术和方法预研。**

**阶段成果：**

1. 大数据试验场技术调研报告。

2. 大数据交易中心技术调研报告。

■ **2016年，第4季度**

**时间**：2016年10月－2016年12月

**任务：确定技术和系统平台；对探索性大数据分析与价值评估系统、数据质量评估与修复系统、大数据试验沙箱、公平可信数据交易保障工具集、交易数据管理和访问控制平台、大数据试验过程协同管理平台、大数据试验场管理运维平台进行关键技术攻关，设计关键算法。**

**阶段成果：**

1）录用发表高水平论文5篇。

2）系统类平台调研报告。

■ **2017年，第1季度**

**时间**：2017年1月－2017年3月

**任务：对探索性大数据分析与价值评估系统、数据质量评估与修复系统、大数据试验沙箱、公平可信数据交易保障工具集、交易数据管理和访问控制平台、大数据试验过程协同管理平台、大数据试验场管理运维平台进行深入研究，构建原型系统，完成探索性大数据分析与价值评估系统。**

**阶段成果：**

1）探索性大数据分析与价值评估系统。

2）申请发明专利 5 项。

3）录用发表高水平论文 5 篇。

■ **2017 年，第 2 季度**

**时间：**2017 年 4 月－2017 年 6 月

**任务：采购支持数据交易的大数据试验场相关硬件。优化数据质量评估与修复系统、大数据试验沙箱、公平可信数据交易保障工具集、交易数据管理和访问控制平台、大数据试验过程协同管理平台、大数据试验场管理运维平台。启动研制大数据众创空间平台的原型系统。完成数据质量评估与修复系统。**

**阶段成果：**

1）**数据质量评估与修复系统**。

2）软件著作版权 3 项。

3）申请发明专利 5 项。

4）录用发表高水平论文 5 篇。

■ **2017 年，第 3 季度**

**时间：**2017 年 7 月－2017 年 9 月

**任务：完善大数据试验沙箱、公平可信数据交易保障工具集、交易数据管理和访问控制平台、大数据试验过程协同管理平台、大数据试验场管理运维平台。研制并优化大数据众创空间平台。完成交易数据管理和访问控制平台的研制。**

**阶段成果：**

1）交易数据管理和访问控制平台。

2）软件著作版权 3 项。

3）申请发明专利 5 项。

4）录用发表高水平论文 5 篇。

■ **2017 年，第 4 季度**

**时间**：2017 年 10 月－2017 年 12 月

**任务：搭建大数据试验场；完善和部署探索性大数据分析与价值评估系统、数据质量评估与修复系统、大数据试验沙箱、公平可信数据交易保障工具集、交易数据管理和访问控制平台、大数据试验过程协同管理平台、大数据试验场管理运维平台、大数据众创空间平台，根据部署情况优化相关系统和平台。调研大数据质量评估与交易公平性技术规范。完成大数据试验过程协同管理平台、大数据试验场管理运维平台、大数据众创空间平台的研制。**

**阶段成果：**

1）适合大数据试验的基础设施。

2）大数据试验过程协同管理平台。

3）大数据试验场管理运维平台。

4）大数据试验沙箱。

5）大数据众创空间平台。

6）软件著作版权 3 项。

7）申请发明专利 5 项。

8）录用发表高水平论文 10 篇。

9）培养硕士/博士 10 名。

■ **2018 年，第 1 季度**

**时间**：2018 年 1 月－2017 年 3 月

**任务： 完成大数据质量评估与交易公平性技术规范标准研制，实现对数据交易的示范应用支持。完成公平可信数据交易保障工具集。**

**阶段成果：**

1）公平可信数据交易保障工具集。

2）软件著作版权 3 项。

3）录用发表高水平论文 10 篇。

4）形成 2 份技术规范。

■ **2018 年，第 2 季度**

**时间：**2018 年 3 月－2017 年 6 月

**任务： 完成支持数据交易的大数据试验场研制及项目验收。**

**阶段成果：**

1）形成 2 份技术规范。

2）软件著作版权 3 项。

3）录用发表高水平论文 10 篇。

4）形成 100 种典型大数据试验数据集。

5）支持数据交易的大数据试验场原型。

6）培养硕士/博士 20 名。

# 第4章 工作条件和环境保障

## 4.1. 项目申请单位情况

<u>复旦大学</u>

复旦大学始创于 1905 年，原名复旦公学，是中国第一所由国人自主创办的高等院校。上海医科大学前身是 1927 年创办的国立第四中山大学医学院，是国人自主创办的第一所高等医科院校。2000 年，复旦大学与上海医科大学合并，组建新的复旦大学，进一步拓宽学科格局，增强办学实力，已经发展成为一所拥有哲学、经济学、法学、教育学、文学、历史学、理学、工学、医学、管理学、艺术学等 11 个学科门类的综合性研究型大学。项目组主

要成员来自于复旦大学计算机科学技术学院和大数据学院。复旦大学计算机学科有近 60 年的历史，始于 1956 年自主建造国内第一台电子模拟计算机。1975 年，复旦大学成立计算机科学系。2008 年，学校整合校内所有计算机学科力量，成立计算机科学技术学院（以下简称学院）。2002 年成立的国家级示范性软件学院及 2011 年成立的国家保密学院，现均依托学院开展办学。学院师资力量雄厚，拥有一支学科结构合理、富有学术活力的教学科研队伍，现有在职教职工 163 人，其中专任教师 107 人。有教授 30 人，研究员 4 人，副教授 48 人，其他副高级职务 10 人，其中博士生导师 40 人。有国家“千人计划”2 人（复旦大学特聘教授），上海市东方学者讲座教授 2 人，上海“千人计划”讲座教授 2 人。近年来，学院承担了国家重大专项、973 计划、863 计划、支撑计划、自然科学基金重点项目及上海市重大科技攻关计划等课题，年均到款科研经费超过五千万元，在高质量学术论文方面取得了显著进步，并连续获得省部级以上科技成果奖励。新成立的大数据学院学以计算机科学、数学和统计学为基础，与经济金融、生命科学、医疗卫生和社会管理等众多学科领域进行深度交叉，开展科学与研究和人才培育，有效推动大数据学科和相关学科的发展，直接面向产业需求建立跨学科、跨领域的研发团队，集聚产业创新人才，着力创造具有重大市场应用价值的科技成果和应用基础研究成果。

复旦大学拥有良好的研发大数据关键技术与应用的基础。复旦大学校园网四校区主干以 10G 环状连接，每个校区三层架构，主干 10G，1G 到楼宇，100M/1G 到桌面。本部至江湾校区、本部至张江校区能够提供专用的光缆 1 对；本部-枫林校区光缆已经用尽，可用 MPLS 做 VPN。光华楼至逸夫楼、逸夫楼至大数据学院可提供专用的光缆 1 对；枫林新建大楼光缆还未部署，今后可预留；张江计算机楼光缆资源不足，需自行投放。科技网宝山机房到复旦大学主节点（在光华楼）之间能够提供至少 1 对物理光缆，连接大数据试验场与复旦大学。宝山机房拥有 1000 多个机柜，供电和空调等足以满足大数据试验场建设的需要。

项目组依托的上海市数据科学重点实验室（Shanghai Key Laboratory of Data Science）是数据科学领域首个政府支持的重点实验室，2013 年 9 月 6 日由上海市科委批准筹建。实验室总体目标是发展成为国际数据科学研究的重要研究场所和数据科学人才培养基地，引领

数据科学研究。实验室前身是复旦大学数据科学研究中心，成立于 2007 年，是国内首个致力于数据科学理论、方法和技术研究的机构，发表了一批高质量的论文。实验室也重视技术应用，涉及金融、智能交通、医疗健康、智慧城市等多领域的大数据分析。

**万达信息股份有限公司**

万达信息股份有限公司是国内最早专业从事城市信息化领域服务的企业之一。公司以行业应用软件（Software）、专业 IT 服务（Service）和软硬件系统集成（System Integration）为主营业务，在智慧健康、智慧社保、智慧教育、平安城市、文化旅游、电子政务、市场监管、智慧环保、国土资源、金融保险、智慧科技等多个行业，形成了一体化的应用解决方案。构建了以信息网络与资源为基础，服务政府——智慧城市管理与运行、服务个人——民生与服务、服务企业——产业与经济的价值体系。截止目前，公司各类解决方案产品已经在上海、浙江、江苏、黑龙江、四川、广东、山东、内蒙古等多个省市的智慧城市领域得到了成功应用。

公司拥有国家计算机信息系统集成壹级和 CMMI5 两项业内最高资质，目前已通过 CMMI V1.3 五级评估，达到国际最高标准。是国家规划布局内重点软件企业、国家发改委高新技术产业化示范工程企业、2009、2013、2014 中国十大创新软件企业、国家第四批创新型试点企业、中国软件和信息服务十大领军企业、全国第一批通过 ITSS 符合性评估企业。被认定为“ 国家级企业技术中心”、“国家级技术创新示范企业”，是“国家卫生信息共享技术及应用工程技术研究中心”的依托单位， 公司拥有共计 500 余项自主知识产权的软件产品和软件著作权、14 项专利；承担多项国家各类标准及指南和上海市及其他地方各类标准的制定工作；曾获得 5 项上海市科技进步一等奖、1 项教育部科技进步奖一等奖，参与的医联工程项目获 2013 年国家科学技术进步奖二等奖。

根据万达信息的总体战略规划，基于公司目前已有的智慧城市综合服务领域多年的基础和积累，公司重点在自主创新、技术和产品研发、成果转化、科技人才培养等方面的进行了重点投入，构建起了可持续发展的创新技术研发体系，努力提升企业技术创新能力。凭借公司高效地资源整合能力和市场能力，企业研发成果能够及时有效的转化为知识产权或产业化

成果。在技术转移工作的过程中，努力进行运行模式的创新，加快良性循环自我发展的能力形成，提高了企业创新能力并带动新一轮技术发展。

围绕着应用系统研发这个主线，以技术研发攻关机制为基础，配合各类支撑要素、管理规范和技术规范，稳步推进研究开发工作。其中，公司具备了技术支撑要素主要包括人才的培养和培训、产学研联盟机制、与外部厂商和高校研究机构建立的合作关系，市场营销体系的建立和通过高质量的行业咨询服务发现客户和市场机会；管理规范主要有通过规范的项目管理和架构设计保障项目质量和项目成果的可复用性，以软件资产复用机制为基础，落实规范化应用平台的研发管理流程和方法；技术规范主要是遵循的国际国内相关标准、参与制定的多项国家相关标准，形成的万达信息软件开发方法论。

凭借自身行业优势和核心技术竞争力，万达信息是中国智慧健康产业联盟骨干企业、上海大数据产业技术创新战略联盟首届理事会理事长单位、是上海市大数据交易中心的发起人和股东单位、上海市云海联盟副理事长单位、上海市物联网产业联盟的理事单位等，和国内多所高校保持了良好的产学研合作关系，承担了多项国家重大专项核高基专项课题、科技部科技支撑课题、863 计划课题等，在智慧城市、云计算、物联网、移动互联、SOA 技术架构、业务基础平台、分布式计算等方面均具有丰富科研成果积累。

在大数据方面，积极牵头大数据领域的技术攻关和平台研发，先后承担了科技部 863 面向生物大数据开发与利用技术课题 2 项、科技部“十二五”重大科技支撑面向养老服务业项目 1 项，市科委医疗大数据 3 项、工商大数据 1 项、大数据管理平台研究 1 项，主要研究大数据挖掘的技术平台、模型与算法。通过搭建大数据挖掘环境，形成较为成熟的平台、软件与大数据挖掘方法论，已在医疗卫生等万达众多优势领域中使用，促进大数据技术成果惠及民众；通过研究和突破面向大数据分析的计算框架、数据获取、数据存储、数据分析和数据展示等关键技术，并结合具体的应用场景，研制面向不同行业的大数据分析平台，将大数据分析与行业应用紧密的联系在一起，使数据得到综合利用和科学分析，提高行业服务水平。在将大数据分析技术落到实处的同时，扩大智慧城市建设成果的经济和社会效益。

近年来，公司也在积极调整业务体型以适应信息化产业的互联网行业特征，成立内部创

新实验室、创新项目团队、创新项目激励体制，围绕互联网医疗、健康管理、医药电商、医保控费、大数据分析、大数据管理、云支撑平台等典型业务和关键技术领域开展攻关和重点研发。公司现已从课题研究、项目实施、运营探索、联合攻关、原型研发、信息消费、信息惠民等层面，开展包括平安城市、坚强电网、居家养老、智慧园区、金融等领域的创新业务开拓，并尝试互联网平台自运营服务，极大的拓展和丰富了公司的产业链角色，也为公司拓展了更多的产业合作伙伴，获得了政府及业界的普遍认可。

**上海超级计算中心**

上海超级计算中心成立于 2000 年 12 月，是 2000 年上海市一号工程----上海信息港主体工程之一，由上海市政府投资建设，是国内第一个面向社会开放，资源共享、设施一流、功能齐全的高性能计算公共服务平台，上海重要的信息技术基础设施，目前运营着“魔方”（曙光 5000A）、“蜂鸟”等超级计算机，同时配备丰富的科学和工程计算软件，致力于为国家科技进步和企业创新提供高端计算服务。

多年来，上海超级计算中心立足上海，面向全国，为来自工程科研院所和多所知名大学的超过 400 家用户，提供了随需应变的高性能计算资源、技术支持以及高级技术咨询服务，支持了一大批国家和地方政府的重大科学研究、工程和企业新产品研发，在汽车、航空、钢铁、核能、市政工程、新材料、生物制药、天文、物理、化学等多个领域取得了大批重大成果。

随着大数据时代的来临，上海“创新驱动发展、经济转型升级”，中国 （上海）自由贸易实验区、上海“四个中心” 和智慧城市建设等发展目标给上海超级计算中心的转型发展带来机遇和挑战，在中心新一轮的建设中，中心将在高性能计算公共服务平台的基础上，大力拓展大数据和云计算服务应用领域，形成三大业务并驾齐驱。未来中心将在业务方向、运营方式、建设模式、体制机制和人才结构等五个方面实现“五大转型”，并建设成为国内一流、世界知名的信息化运营管理中心，形成大数据应用与产业发展生态环境；信息化资源集聚与相关产业孵化的“航空母舰”；新一代信息技术的研究创新、应用服务、产业孵化、人才培养的重要基地。

上海产业技术研究院

上海产业技术研究院成立于 2012 年 8 月，以推进产业发展为根本任务，面向国家战略和上海战略性新兴产业发展，以产业共性技术研发与服务平台为载体，将企业、高校、科研院所、金融投资以及各类技术创新平台等有机组织起来，实施集成创新，实现产业链上下游的高效联动、协同发展，推进产业技术创新和产业化。

上海产业技术研究院承担并完成了“面向科技与工程的大数据共享服务平台”项目，该项目结合生物医学、工程创新、科技服务和实验动物领域的数据应用和共性需求，联合产学研合作机构，研发大数据应用架构、数据共享模型、集成框架和技术架构，搭建具有数据汇聚、分析、挖掘和应用集成等功能的“服务/交易/协同”大数据共享平台，实现数据供给方、加工方和需求方之间的价值和成果转化，形成了 30 个服务案例。在平台功能建设的基础上，开展面向领域的内容建设工作，为未来更多行业或领域大数据平台建设的标准化和规范化运行奠定了基础。为了推进大数据产业的的标准建设，成立了上海产业技术研究院大数据标准专家委员会，主要进行数据质量与安全的标准研究。与韩国数据振兴院合作，成立了中韩数据工程中心，在数据的质量检测、访问控制、数据漏洞扫描、加密等方面形成了 5 个软件产品。

上海科技网络通信有限公司

上海科技网络通信有限公司前身是上海市人民政府科技网，是一个大型城域宽带网络运营商（NSP），公司于 2000 年 11 月在徐汇区注册成立、于 2012 年 5 月公司注册变更到宝山区，注册地址为宝山区长江西路 255 号，工作场所面积约 16000 平方米。由云赛信息(集团)有限公司（持股比例为 80%）和上海科技投资公司（持股比例为 20%）共同投资组建，注册资金 2 亿元人民币，为国有全资企业。上海科技网自 1995 年开始建设，是“九五”期间上海市政府三项标志工程之一。1997 年上海科技网建成覆盖全上海地区的 ATM 光纤宽带网，是国内最早建设的城域宽带网络平台。2002 年，公司已建成了覆盖全市的高效、可靠的 IP 宽带城域网。公司注重技术开发工作，培养了一支稳定的、较高素质的技术团队，现有员工 123 名，工程技术人员 82 名（其中高级工程师 5 人，工程师等专业技术人员 82 人）。一直

以来，公司技术部门持续跟踪 IDC、云计算、物联网等技术发展，并与高等院校、研究院所进行产、学、研的密切合作。

上海科技网络通信有限公司已获得 1 项上海市科学技术进步奖二等奖， 获批并实施 5 项国家级专项，获批并实施 4 项市级重大专项，已获得 4 项软件著作权；承担大量国家科技支撑计划、国家 863 专项、国家发改委专项类国家级重大项目，承担大量市经信委软件和集成电路专项、市科委重大课题类上海市级项目。上海科技网于 2008 年被认定为高新技术企业、并于 2011 年及 2014 年连续通过复审；上海科技网已获得 ISO/IEC 20000-1：2005 和 ISO/IEC 27001：2005 认证。通过科技研发和项目实施,丰富和优化了数据中心产品线、云计算服务产品线和增值服务产品线、网络接入产品线，形成基于两大数据中心、一个云计算平台、一个城域网资源的 ICT 多产品综合运营服务能力。

## 4.2. 已经具备的实验条件

项目承担单位由复旦大学、万达信息股份有限公司、上海超级计算中心、上海产业技术研究院、上海科技网络通信有限公司组成，在大数据研究和基础信息基础设施建设方面研究和技术基础雄厚。

复旦大学计算机科学技术学院建设有云计算机房，与上海市数据科学重点实验室（Shanghai Key Laboratory of Data Science）联合的计算资源可供本项目技术的预研和原型系统搭建。项目组依托的上海市数据科学重点实验室（Shanghai Key Laboratory of Data Science）是数据科学领域首个政府支持的重点实验室，2013 年 9 月 6 日由上海市科委批准筹建。实验室总体目标是发展成为国际数据科学研究的重要研究场所和数据科学人才培养基地，引领数据科学研究。实验室已经具备：联想 RD640 服务器（英特尔至强 E5-2609v2*2/192GB 内存/600GB SAS，10K RPM*3/R700RAID 卡/）12 台，联想 RD450 服务器（英特尔至强 E5-2620v2*2/256GB 内存/600GB 10K SAS*3/RAID）2 台；戴尔刀片服务器 1 套（含 DELL PowerEdge M1000e 模块化刀片式盘柜 10U 机箱 1 个、DELL PowerEdge M620 刀片服务器（英特尔至强 E5-2609v2 *2/128GB 内存/1TB 7.2K RPM 近线 SAS 6Gbps 2.5 英寸热插拔

硬盘*2/H310 RAID 控制器/Broadcom 57810-k 双端口 10Gb KR Blade 网络子卡）9 台、Dell Force10 MXL 10/40 GbEDCB 刀片式交换机 1 台）。DELL PowerVault MD3600i（双控）iSCSI 存储设备（4TB 近线 SAS6Gbps，3.5 英寸，7.2K RPM 硬盘热插拔*12）6 台；DELL PowerVault MD3800i（双控）iSCSI 存储设备（4TB 近线 SAS 6Gbps，3.5 英寸，7.2K RPM 硬盘热插拔*12）1 台；DELL PowerVault MD1200（双控）存储扩展盘柜（4TB 近线 SAS 6Gbps，3.5 英寸，7.2K RPM 硬盘热插拔*12）14 台；DELL PowerVault MD3600i（单控）iSCSI 存储设备（4TB 近线 SAS 6Gbps，3.5 英寸，7.2K RPM 硬盘热插拔*12）1 台；DELL PowerVault MD3200i 存储设备（4TB 近线 SAS 6Gbps，3.5 英寸，7.2KRPM 硬盘热插拔*3）2 台；PowerVault MD3200 直连存储设备（300GB 近线 SAS 6Gbps，3.5 英寸，15KRPM 硬盘热插拔*5），共计 1081TB 的数据存储能力。戴尔万兆以太网交换机(Dell N4064,48 个 10GBASE-T 端口和 2 个 40GbEQSFP+接口)，思科 C3750X 三层千兆交换机(WS-C3750X-24T-S 24 口三层交换机)3 台，思科 C3560X 千兆二层交换机（WS-C3560X-24T-L 24 口二层交换机）1 台，思科 C2960G 千兆二层交换机（WS-C2960G-24TC-L）1 台。30 个独用的互联网 IP 地址。

合作单位上海科技网宝山机房有上万平米的机房，到复旦大学主节点（在光华楼）之间能够提供至少 1 对物理光缆，连接大数据试验场与复旦大学。宝山机房拥有 1000 多个机柜，供电和空调等足以满足支持数据交易的大数据试验场的建设需要。

合作单位万达信息是上海地区大数据技术研究及行业应用的先行者，在大数据技术研究及行业综合应用方面处于领先优势：基于在大数据技术研究及行业综合应用方面处于领先优势，万达信息成为上海大数据产业技术创新战略联盟的首届理事长，是上海市大数据交易中心的发起人和股东单位，作为行业的领军企业，在社会保障、卫生服务、平安城市、市场监管、环境保护、物流管理、文化教育等优势领域中已经开始尝试使用大数据技术来满足用户多样化、需求个性化的要求。重点选取医疗卫生、食品安全、终身教育、智慧交通、公共安全、科技服务等具有大数据基础的领域，探索交互共享、一体化的服务模式，建设大数据公共服务平台，促进大数据技术成果惠及民众。在大数据技术、产品与应用方面承担了一系列国家和地方的重要科研项目，为本项目的执行提供技术支撑：在大数据技术研究与应用方面，

先后承担了国家高技术研究发展计划（863 课题）《基于区域医疗与健康大数据处理分析与应用研究》，上海市科委重点项目《智慧城市领域大数据分析关键技术研究和应用》、《基于区域医疗大数据的数据挖掘技术与示范应用》等 4 项；在大数据产品研发与应用方面，承担了上海市科委科技创新行动计划《面向行业应用的大数据管理系统研发和示范》、《用于市场主体信用领域的大数据融合分析技术平台研发及应用示范》等 6 项，以及上海张江国家自主创新示范区专项发展资金重点项目《大数据开发利用支撑平台研发和应用示范》。

合作单位上海超级计算中心正在大力拓展大数据和云计算服务应用领域，规划建设具有自主知识产权的公共云服务平台、大数据公共处理分析平台，已部署测试与生产环境用于应用开发与移植，提供对外服务，将为大数据双创服务应用示范提供可靠的软硬件环境支撑，基于十多年高性能计算资源管理、运维和服务经验，确保提供长期、专业的后续服务。

此外，上海超级计算中心是上海大数据联盟秘书长单位。联盟正积极落实市相关产业发展需求，承担产业对接各项活动，形成数据资源集聚、品牌效应彰显、技术合作深入的共享服务平台，推动大数据产业发展应用。

## 4.3. 项目组织机制设计

本项目由产学研三方联合申报，项目负责人从事科学研究和项目组织管理工作多年，具有研究的持续性，产学研合作的成员主要由年富力强的、在相关领域已有丰富研究成果的教授、技术总监、项目经理、副教授、工程师和研究生组成。在项目开展前期，主要由项目负责人提出研究方案，并安排项目组教师和研究生分小组展开模型和算法研究，在主要模型和算法完成后，安排公司研发团队开发软件平台系统实现主要模型和算法，安排上海超算中心团队进行系统测试和测评，并部署在大数据试验场环境进行试用，同时由大数据联盟进行众创空间推广，项目各个阶段申请研究成果专利和软件著作权，并发表相关论文。

为保证项目的顺利实施，拟建立项目工程组织，管理项目的整体规划和具体实施。

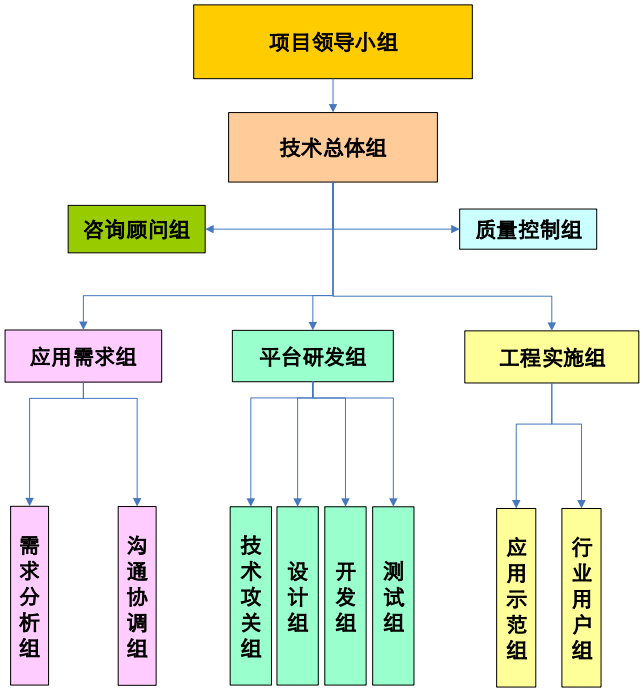


图 2 项目管理组织结构图

项目领导小组：本项目将建立以项目负责人牵头，项目主要参与单位负责人共同组成的项目管理小组，负责对总项目以及子项目进展监管、控制及管理。

技术总体组：以复旦大学、万达信息、超算中心等多方技术专家为主组成，负责项目总体技术路线、设计、标准规范等。

咨询顾问组：聘请来自高校以及业界内的专家组成，对项目的具体技术路线、应用实施、总体设计、标准规范进行指导和提供咨询，规避可能存在的风险，提出解决思路。

质量控制组：对整个项目执行 ISO9000 质量管理体系的落实、实施和监督。

应用需求组：需求分析组负责开展项目的前期需求调研工作。

沟通协调组：对项目的实施起协调推广工作。

平台研发组：关键技术攻关组进行关键技术的实现设计；设计组进行总体设计；开发组进行大数据试验场多个平台关键技术开发；测试组：进行平台测试。

工程实施组：应用示范组在集成各个系统和平台的基础上开展研制成果在领域用户的示范应用。

## 4.4. 产学研结合加快工作进展的设想

本项目由复旦大学研究团队牵头，并与万达信息股份有限公司、上海超级计算中心、上海产业技术研究院、上海科技网络通信有限公司进行产学研合作，复旦大学研究团队负责该项目的总体技术和关键算法研究，万达信息股份有限公司主要负责系统软件平台的开发，关键技术攻关和部署，上海产业技术研究院主要负责数据质量测评和优化，上海超级计算中心主要负责众创平台建设，并负责联系示范用户单位推广。

在合作过程中多方将建立风险共担机制，明确各方的责任权利义务，让研发者最大程度地降低研发风险性。围绕该项目需求开展多种形式、多种层面的多元化调研，扩大产学研合作范围，坚持紧密合作，提高产学研合作深度。在产学研合作研发和攻关过程中，积极推行紧密型合作模式，使高校研发机构与企业和联盟结成共同体，使产、学、研三方在资金、技术、设施、人力以及管理等方面有序结合为一个整体，充分利用各方在技术、资金、联盟、产业等方面的优势，实现该项目的总体攻关目标，不断增强合作的紧密度，确保大数据试验场关键技术研发的高效和成功，并以关键技术创新和示范应用带动大数据试验场持续扩大影响力，并持续拓展项目组多方的产学研合作空间。

# 第5章　成果形式和考核指标

## 1. 主要技术指标及成果形式

本项目致力于构建面向交易的大数据试验场，突破相关技术瓶颈，主要的成果形式包括：

1) 研制支持数据交易的大数据试验场平台软件和关键工具集（软件系统，具体系统名称与指标见后），构建支持数据交易的大数据试验场；
2) 20 项核心发明专利申请；
3) 15 项软件著作权申请；
4) 50 篇国际/国内学术论文；
5) 形成技术规范草案 4 份；
6) 培养硕士/博士 30 名。

项目的具体技术指标如下：

1) 支持数据交易的大数据试验场平台软件和关键工具集包括
   a) 一套数据质量评估和修复系统（软件），支持质量模型的定义、自动和半自动测量、质量评估报告生成，实现数据质量度量通用模型和评估方法的评估规则管理；实现数据质量的修复算法，算法种类涵盖数据质量模型的各种特性，在明确数据质量规则的基础上，能够 100%修复数据不满足数据质量规则的质量问题；
   b) 一套探索性大数据分析与价值评估系统（软件）， 实现数据选取、数据融合、工具选取、数据价值评估、数据推荐等功能；
   c) 一套交易数据管理和访问控制平台（软件），实现面向多源异构交易数据的元数据管理、数据目录管理、数据访问与安全授权、数据资产使用管控、数据血缘图谱检索等功能；

d) 一套交易过程协同管理平台软件（软件），建立大数据工具集接口规范，并面向用户提供数据服务、工具服务、存储服务、计算服务等功能；

e) 一套大数据试验沙箱（软件）,支持应用适配的软硬件集群（Hadoop、Spark）自动配置、全局优化的自适应资源调度功能，支持多用户在大数据试验场 2000 节点的共享集群上创建隔离的试验专用空间；

f) 支持公平可信数据交易的保障工具集软件（软件）：支持数据可信公平交换、交易总账管理、结果验证、试验沙箱安全初始化、可信审计数据采集

g) 大数据众创空间平台（软件），实现异构多源数据多层次共享，支持基于大数据试验场的多用户沟通交流、创新创业团队示范。

2) 构建适合大数据试验的基础设施，物理存储容量不少于 5PB，能够对 1PB 的数据资源实行并行化处理的计算能力试验场内部的骨干网络至少达到 40Gbps 的数据交换能力，集群管理软件需要能够同时管理至少 2000 个节点、支持监控大数据试验场中部署的主要软件的工作状态，能够实现统一配置、自动报警等管理功能。

3) 可提供 1PB 以上的大数据集（含样本数据、仿真数据等），涉及科学、医疗、交通等领域，达到至少 100 种不同的数据集。

4) 形成 4 份技术规范，主题与数据数据质量分析、公平交易、数据访问控制、工具集接口相关。

**2. 项目实施中形成的实验室、研发中心、示范基地、中试线、生产线及其规模等**

通过本项目的实施，建成支持数据交易的大数据试验场，用于支撑大数据技术及应用的探索及试错，支持包括上海市大数据交易中心在内的数据交易场所的数据交易业务。为建立上海大数据试验功能性平台打下技术、系统和运维基础。

**3. 经济考核指标**

本项目拟建设一个公益性，开放性的支持数据交易的大数据试验场平台，不以营利为目的，但为促进试验场的良性发展，拟建立一套可持续运行的机制，对社会提供大数

据试验服务。力争通过 3 年努力，通过服务收入能覆盖日常的运行成本。

**4. 团队建设情况**

本项目拟培养硕士/博士 30 名;培养一批具有国家前瞻视野的数据科学青年科学家，建立由技术专家和行业专家构成的试验场专家队伍。

**5. 社会效益指标**

本项目拟建成一个公益性、开放性的功能性平台，支持上海市大数据应用及产业发展，为上海乃至全国需要进行大数据研究及试验的组织及个人提供相关基础设施及专业咨询、培训服务。试验场通过提供交易前、中、后的技术支持服务促进包括上海大数据交易中心在内的数据交易场所的建设。

# 第6章　预期效果和风险分析

## 6.1. 项目成果对社会发展所起的作用

随着数据的飞速增长及大数据技术的快速发展，深入挖掘数据价值、开展数据应用已经成为政府、学校、科研机构、企业强化核心技术、提升服务能力、增强竞争力的重要手段，但对大数据的利用面临着数据难以获取、分析工具难以集成利用、分析环境难以自主搭建等交易、分析门槛。因此，我们迫切需要一个汇聚多方数据，集成大数据工具集，提供数据利用环境的大数据应用场所。本项目将基于复旦大学、万达信息、上海超级计算中心、上海产业技术研究院、上海科技网络通信有限公司在大数据领域数据管理、技术研究、业务应用等方面的优势，构建汇聚多源异构数据，整合大数据工具集，提供数据交易、分析环境的大数据试验场。

支持数据交易的大数据试验场能够为重大创新产业领域的大数据建设提供一个统一的开发、共享的技术验证、技术支持平台，为不同领域的海量数据存储提供服务，研究有效的数据管理方法，试验与领域应用相关的数据分析方法，培养数据专家，减少重复建设、避免技术弯路。本项目成果的应用将促进我国大数据产业的发展，通过在大数据基础技术研究，

数据融合管理，大数据工具集整合利用，交易、分析试验等方面的应用，促进探索大数据应用与技术的价值链与产业链，打造开放包容的合作体系，集聚上海乃至全球的优势资源，服务上海全球科创中心的目标，形成可持续发展的机制，探索大数据试验场良性健康发展的道路。

## 6.2. 经济效益和产业化前景

本项目成果为上海市大数据试验场功能性平台的建设提供技术储备、设施构建及经验积累，具备巨大的经济效益及产业化前景。大数据试验场功能性平台将对接上海数据交易中心，面向医疗、科研、教育等行业，提供数据存储、处理、分析利用等服务，进而推广到工商、交通、金融等行业，推动大众创业、万众创新。随着各行业对数据利用的逐步重视，未来大数据服务市场将爆炸式增长。

## 6.3. 对环境影响程度及资源综合利用情况

本项目为软件开发类项目，主要进行电子信息的搜集、加工、整理和发布，属于无污染项目。本项目将依托各单位现有基础设施，不涉及新建土建项目，不产生废水、废气、废液、废渣。在实施运营过程中对周围环境基本不造成污染,基本没有有害气体、废渣、废水排出，所采用的设备也不产生电磁污染，基本不产生设备噪声源。

本项目为软件开发类项目，固定资产投资主要在软硬件设备的采购，不涉及对自然资源各组成要素进行多层次、多用途的开发利用过程等资源综合利用问题。

## 6.4. 风险分析

### 6.4.1. 技术风险

在支持数据交易的大数据试验场关键技术研发上，在技术、数据上有可能存在 3 方面的风险，分别来自技术架构设计、软件功能部署和数据资源积累，我们这里分别描述并给出了应对和化解风险的措施。

（1）设计的系统架构不能适合大数据试验和数据交易的变化。大数据试验场和数据交易属于新兴事物，国内目前尚无可参考借鉴的成熟经验，目前的大数据试验场系统架构以及实施方案主要参考了云计算平台等解决方案，针对数据交易需求的特点进行设计，因此存在系统架构并不能完全适应各种数据交易相关的大数据试验的风险。

应对此风险，项目组将召集各方面专家对支持数据交易的大数据试验场技术方案进行充分论证，架构设计、设备选型过程中充分考虑可扩展、多用途和灵活性，避免建成后的大数据试验场只能服务于特定几种数据交易相关的大数据试验的问题。

（2）部署的软件功能不健全限制了大数据试验。由于建设资金有限，当前项目重点安排了试验场的硬件和支持数据交易的试验系统建设，对于大数据试验场所需的支撑软件、业务软件，常规的、免费或开源的产品可以部署，部分需要定制开发的软件（如先进的机器学习、数据分析软件）缺少研发经费支持，这些欠缺的软件会限制大数据试验，限制大数据试验场提供服务。

应对此风险，一方面可以尽量寻找可替代的免费或开源软件，满足大数据试验；另一方面积极寻找资金支持，组织力量研发必须的系统，逐步添置商业软件，使大数据试验场能够开展更多的数据试验，提供更好的服务。

（3）积累的数据资源无法完全支撑大数据试验。该风险主要表现在数据量积累得不够多，种类积累得不够多，数据质量不够高等，导致无法开展大数据试验，或是使得试验结果产生难以察觉的偏差。

应对此风险，项目组将通过 3 个途径积累足够用于大数据试验的数据资源：一是充分调动复旦大学和合作单位的优势，汇聚一批已有的科研和商业数据资源，充实大数据试验场的数据储备；二是通过互联网获取公开的、对数据交易有价值的数据资源；三是吸引、挖掘社会资源，尤其是与复旦大学合作关系良好、愿意共享数据的单位，获取一部分数据资源。对于数据质量不高的问题，可以通过使用数据清洗工具、多来源核校、人工清洗校验等方式提高数据质量。

总体看来，复旦大学与万达信息股份有限公司、上海超级计算中心、上海产业技术研究院、上海科技网络通信有限公司在多年的项目研发中积累了丰富的技术攻关和项目开发经验。本项目拥有一批优秀的技术研发团队，能够充分保障项目建设和营运管理，很好的协调相关方，使项目的总体技术风险降到最低。

### 6.4.2. 市场风险

本项目成果将面向科研、医疗、教育等多行业进行推广应用，覆盖行业数据提供方、大数据工具厂商、大数据分析服务厂商、产学研机构等多家单位，项目成果的市场推广所带来的市场风险将主要体现在产品的成熟度、稳定性、性能、可扩展性、售后增值等方面。我们考虑有如下对策：

（1）行业用户对大数据试验场的认可程度将会成为项目研究成果在市场竞争中存在的风险。在大数据领域，目前国内市场上尚无大数据试验场，还没有形成成熟的市场影响力。用户对本项目成果的接纳认可程度还存在一定的不确定性，而本项目成果在后期的推广实施将借助上海大数据产业联盟、复旦大学、万达信息、上海超级计算中心、上海产业技术研究院、上海科技网络通信有限公司在大数据科研、项目实施等领域的行业资源和良好口碑，展开面向典型行业的推广应用，进而在科委领导单位的支持下，面向更多的行业、市场企业进行推广，同时派遣更多的大数据技术专家、行业专家深入用户中间，进行宣传培训，试点应用，提升市场用户对大数据管理技术的认可程度，降低市场竞争风险。

（2）同类产品竞争产生的市场风险。随着近年来大数据供需市场的发展，相信在不久的将来，将会形成以龙头企业为首的大数据试验场同类产品市场竞争压力。但是由于产业还处于市场萌芽阶段，各类产品在行业应用中还处于探索调整阶段，不存在较大的竞争压力。