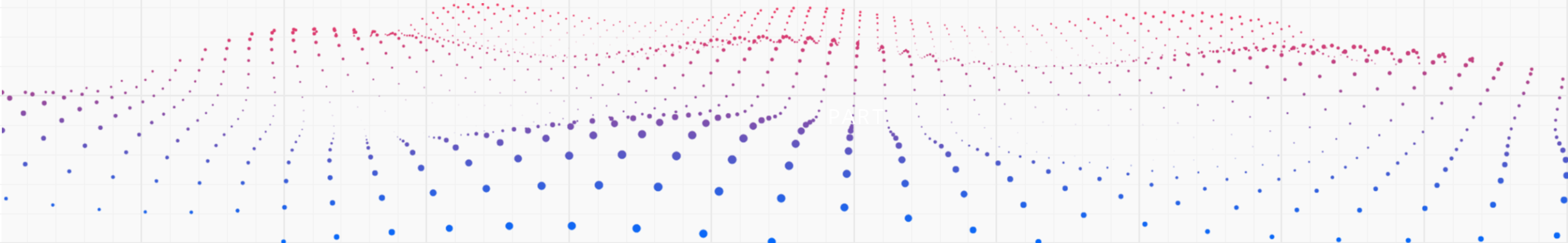
中国数据治理现状调研报告（2021）

CIO时代
研究·培训·资讯

2021年12月

CONTENT



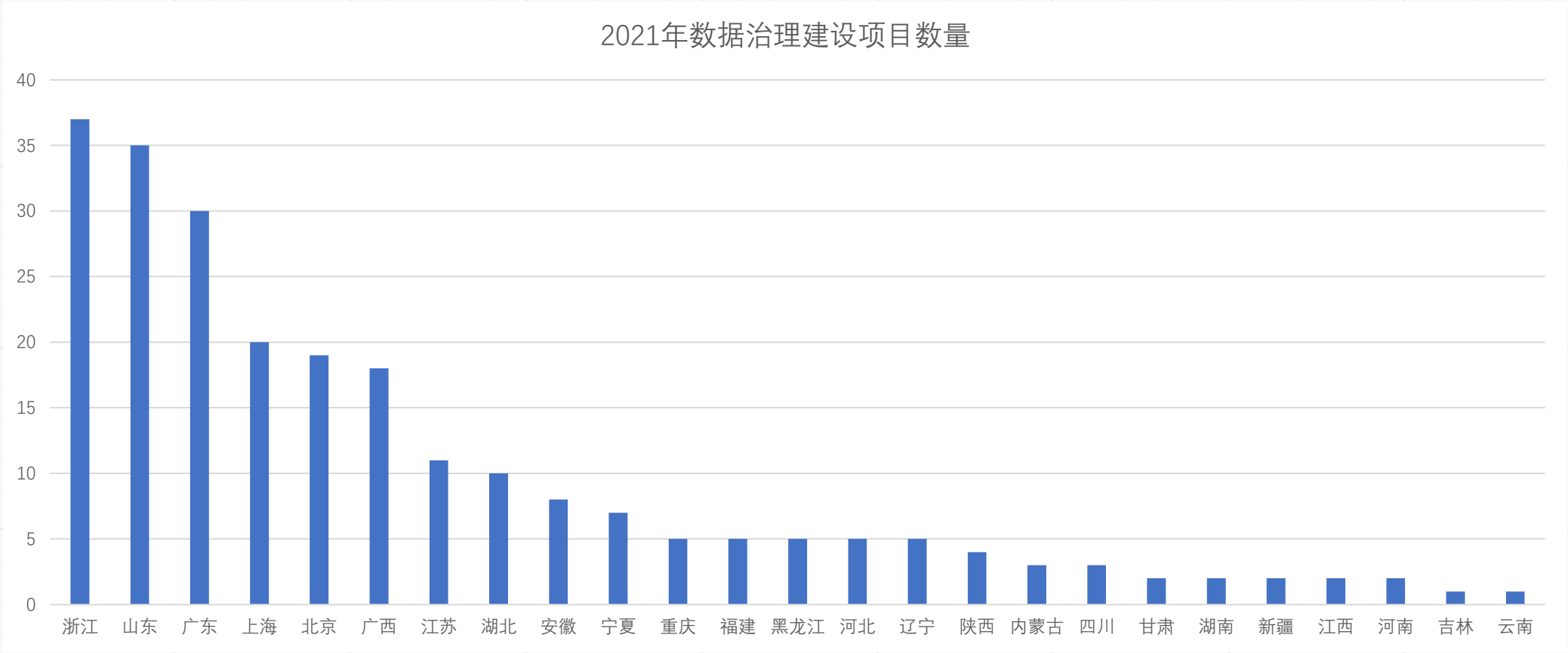
数据治理规范化、合法化势在必行

在大数据时代，数据资源已成为重要的战略资源和生产要素。保障数据安全、个人隐私，促进数据治理规范化、合法化，已经成为社会各界的共识。

《促进大数据发展行动纲要》	《中华人民共和国个人信息保护法》	《中华人民共和国数据安全法》	《工业互联网数据安全保护要求》
• 2015年8月，国务院印发，系统部署大数据发展工作。	• 于2021年11月正式施行，从多方面对个人信息保护进行了全面规定，建立起个人信息保护领域的基本制度体系。	• 2021年9月，作为首部关于数据安全的律法正式施行，这标志着我国在数据安全领域有法可依，同时对数据治理提出了更高要求。	• 2021年7月正式施行，推动了工业数据管理能力，提高数据安全保护要求。

数据治理市场增长迅速，但发展不均衡

全国公共资源交易数据显示，2021年成交的数据治理相关的建设项目已经超过240个，连续两年同比增长超过30%。数据治理项目主要分布在浙江、山东、广东、上海、北京等经济较发达区域。



技术进步极大增强了数据治理的可操作性

来源：《数据治理工具图谱研究报告》（2021版）

数据治理已经成为企业急需解决的难题

数据质量差	原始数据获取时不够规范，出现的数据缺失、重复、损坏，入库之后，数据加工不够规范
数据多源异构	数据来源多样并且数据之间结构可能不一致，成为数字化转型的重要难题
数据不一致	同样的统计口径产生不一样的结果；同样的对象同样的属性，却是不同的数据
数据溯源能力不足	不能及时、完整地追踪数据变化情况和数据质量问题

调研目的

为掌握国内数据治理的实际现状，CIO时代于 2021年 9月开展数据治理调研工作。通过调研各单位数据治理现状，以及数据治理工作的难点、痛点，结合相关专家和企业的访谈，力争全面客观的描绘国内数据治理在技术、管理、人才等方面的现状、发展趋势，深度分析国内数据治理面临的共性问题，并尝试提出相应的解决方法，希望借此推动国内数据治理工作的开展。

调研方法和样本说明

调研对象：企业中高级管理人员、信息技术人员、IT管理人员

调研时间：2021年9月

调研方式：文献研究、调查问卷、小组访谈

有效问卷：618份

文献研究

- 通过行业研究报告、公开新闻资料、公共资源交易数据等进行研究分析

调查问卷

- 在线问卷调查

CIO时代数据治理专委会(筹) 数据治理发展情况调研问卷

数字社会高速发展，对数据治理提出了新的更高的要求，也让人们看到数据治理是保证数字社会健康发展的重要基础。

为推进数据治理的有序发展，CIO时代将成立数据治理专委会。现进行数据治理发展情况调查，旨在为数据治理专委会未来的工作提供依据。

参与本次调研的人员将有机会成为数据治理专委会理事或成员，也将邀请在本次调查中有突出贡献的专家，共同编撰《数据治理发展研究报告》。

填写问卷过程中若有疑问，请咨询：刘胜文(院长助理)13810528344　鲁四海 13141231801

1、基本信息

单位名称

所在部门

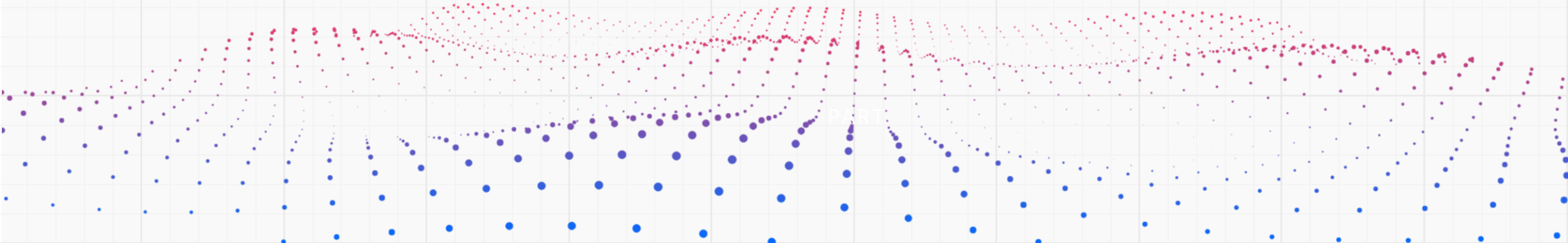
姓名

电话

小组访谈

- 分别在北京、上海、深圳举办三次数据治理座谈会

CONTENT



主要结论

1

数据治理已具备较好的技术基础，近90%的机构已有或正在建设大数据平台。但同时也存在，重建设，轻管理，少应用的情况。

2

开源是最广泛被采用的技术体系。其中Hadoop、Spark选用率最高；ClickHouse作为新晋技术，在互联网行业已得到广泛应用，传统行业应用正在崛起。

3

数据治理的驱动力来源主要是业务，数据治理内驱力需要提升。同时在数据治理能力建设方面，管理滞后、考评缺位，成为限制数据治理能力的最大瓶颈。

主要结论

4

数据治理对人工还存在较强依赖，自动化、实时性还有待提升。

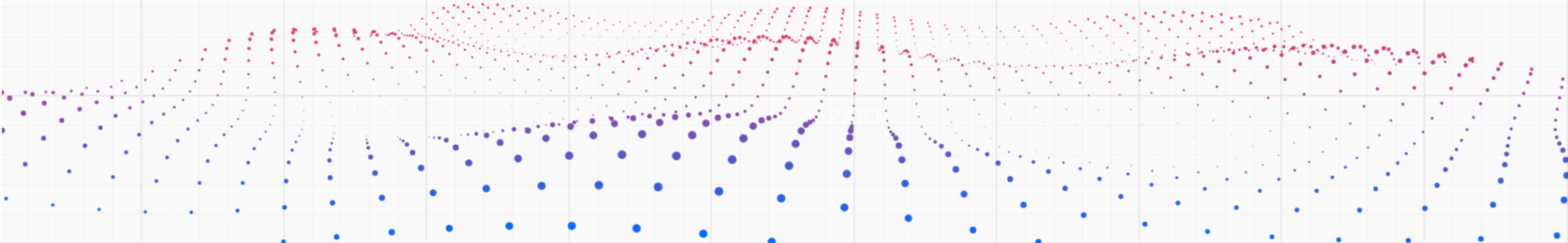
5

数据安全正由“保护数据”扩展到“保护数据成果”、“保护数据资产”，各行业根据自身的业务特点，均采用了与之匹配的多样化保护手段。

6

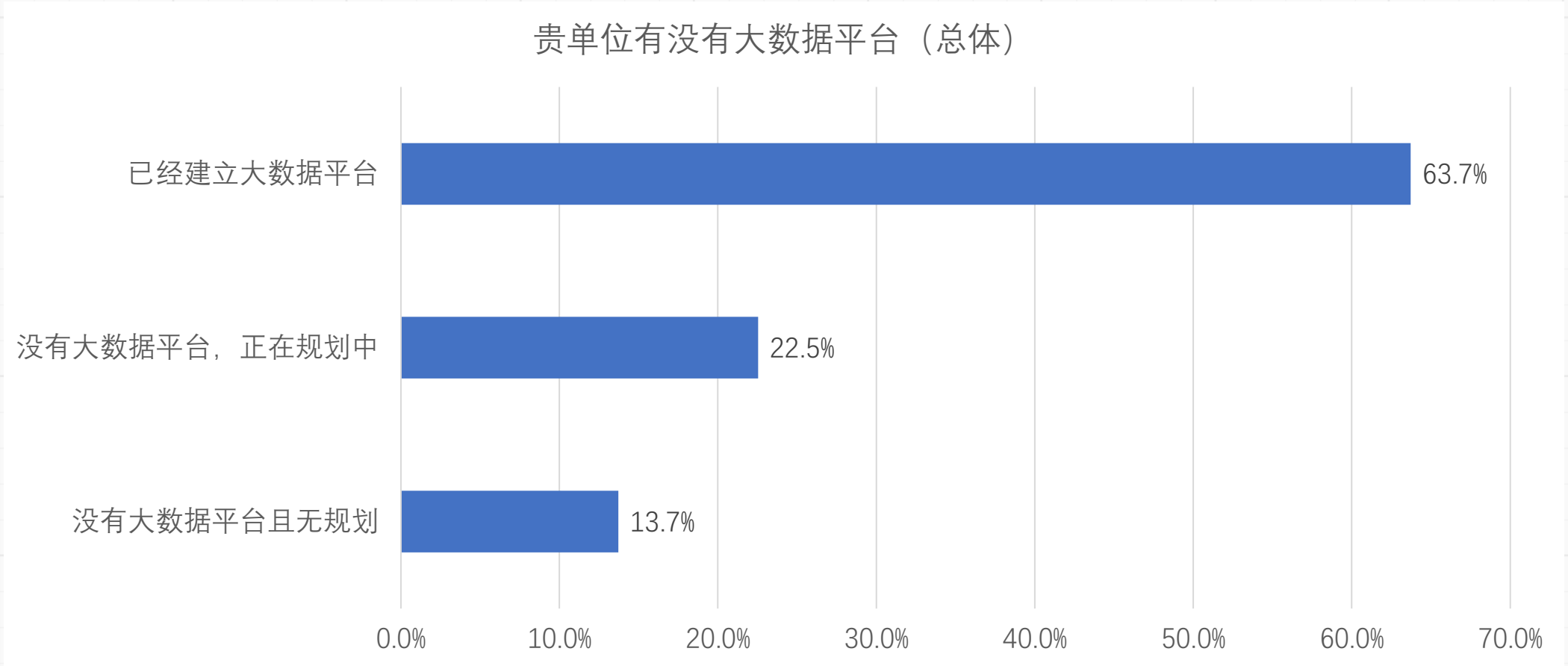
各行业均存在人才不足情况，高端人才、数据分析人才缺口较大，各行业均在通过购买服务的方式解决人力资源不足问题。

CONTENT



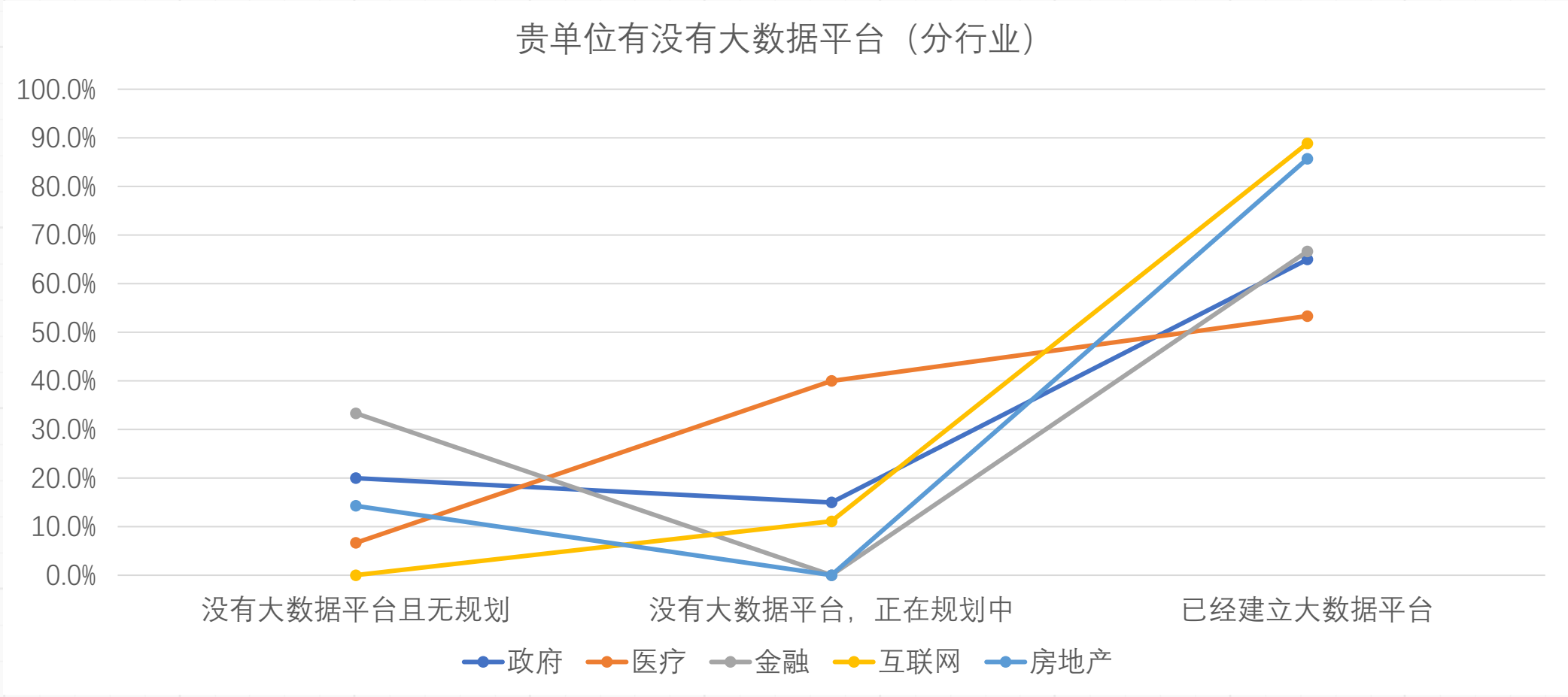
近九成机构已有或正在规划建设大数据平台

63.7%的机构已经建立大数据平台，22.5%的机构正在规划建设。大多数机构都已经意识到大数据平台对数据治理的重要性。



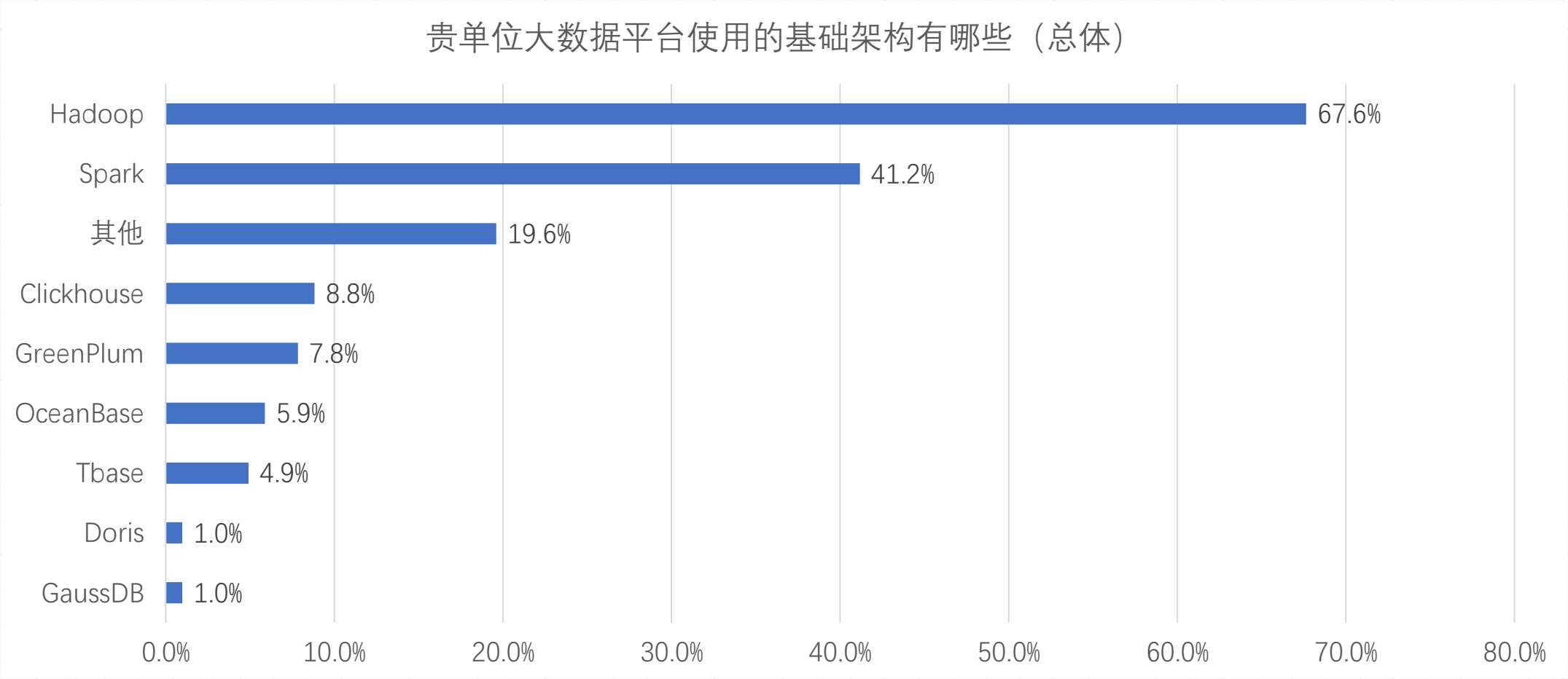
大数据平台建设互联网领先，政府和医疗正迎头赶上

接近90%的互联网企业都建立了大数据平台；未建设大数据平台的政府机构和医疗单位绝大多数已将大数据平台建设纳入未来发展规划中。



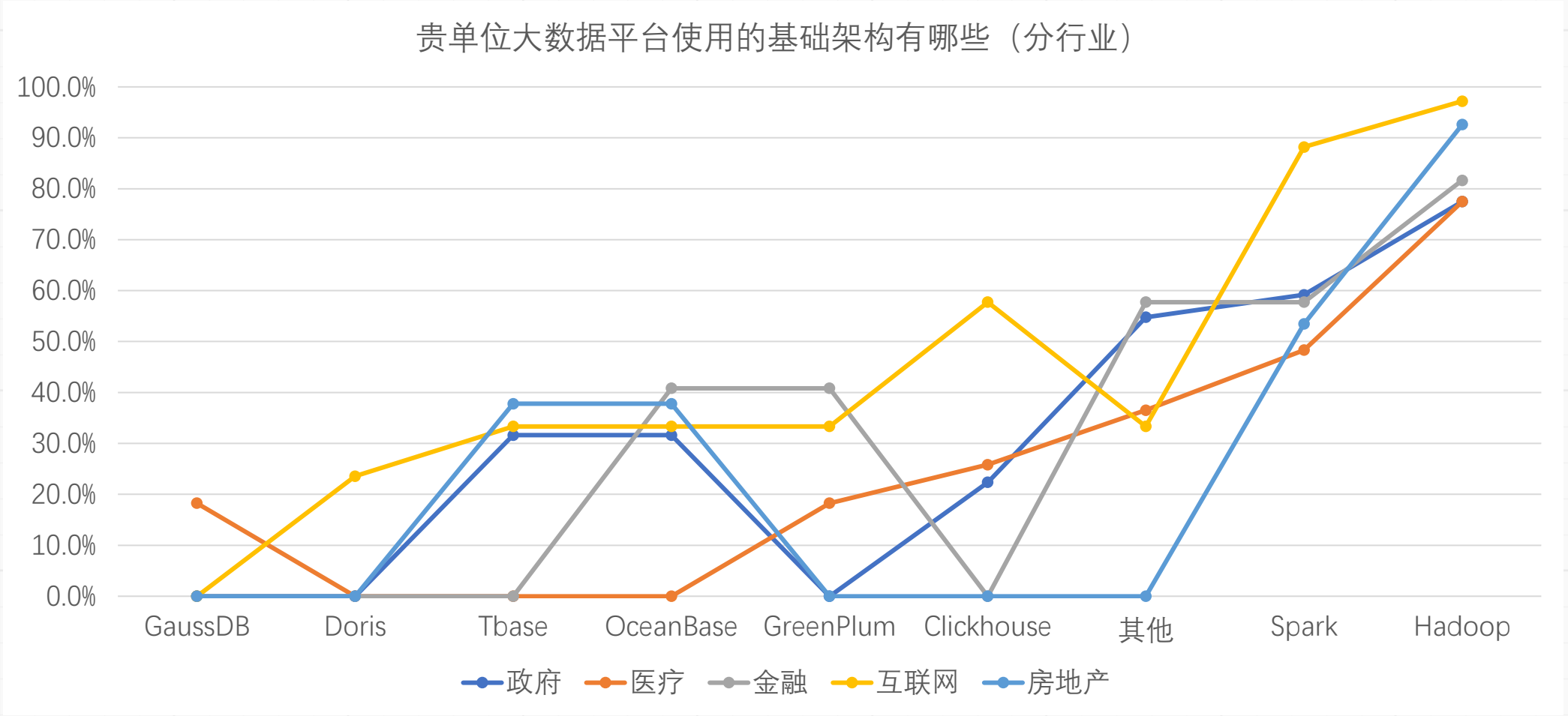
开源是最广泛被采用的技术架构，其中Hadoop应用最广泛

大数据平台技术架构，Hadoop、Spark处于绝对领先位置，比率分别为67.6%、41.2%； ClickHouse作为新晋的数据处理技术平台紧随其后。



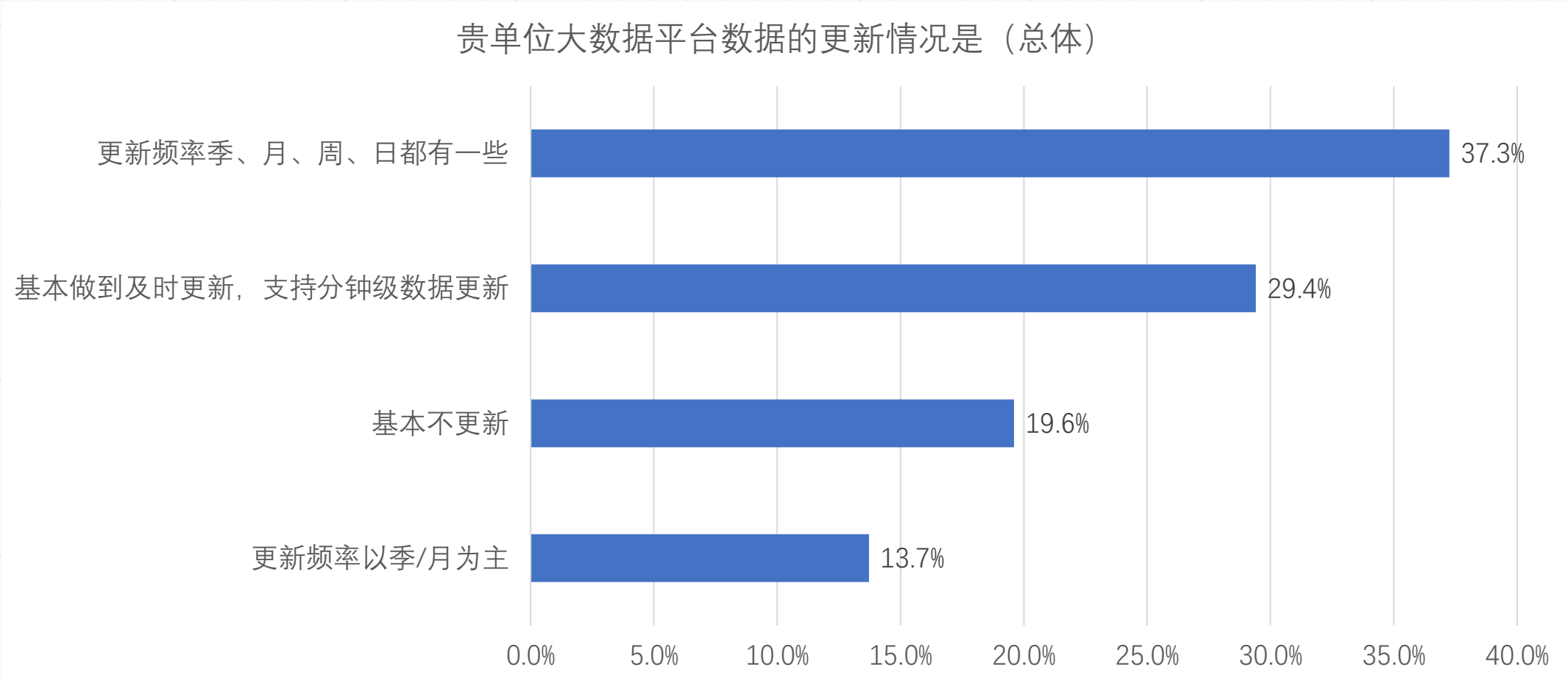
ClickHouse在互联网行业已得到广泛应用

Hadoop、Spark作为经典的大数据技术，在调研行业被广泛应用。但是，将近60%的互联网企业已经采用Clickhouse作为大数据的基础设施。



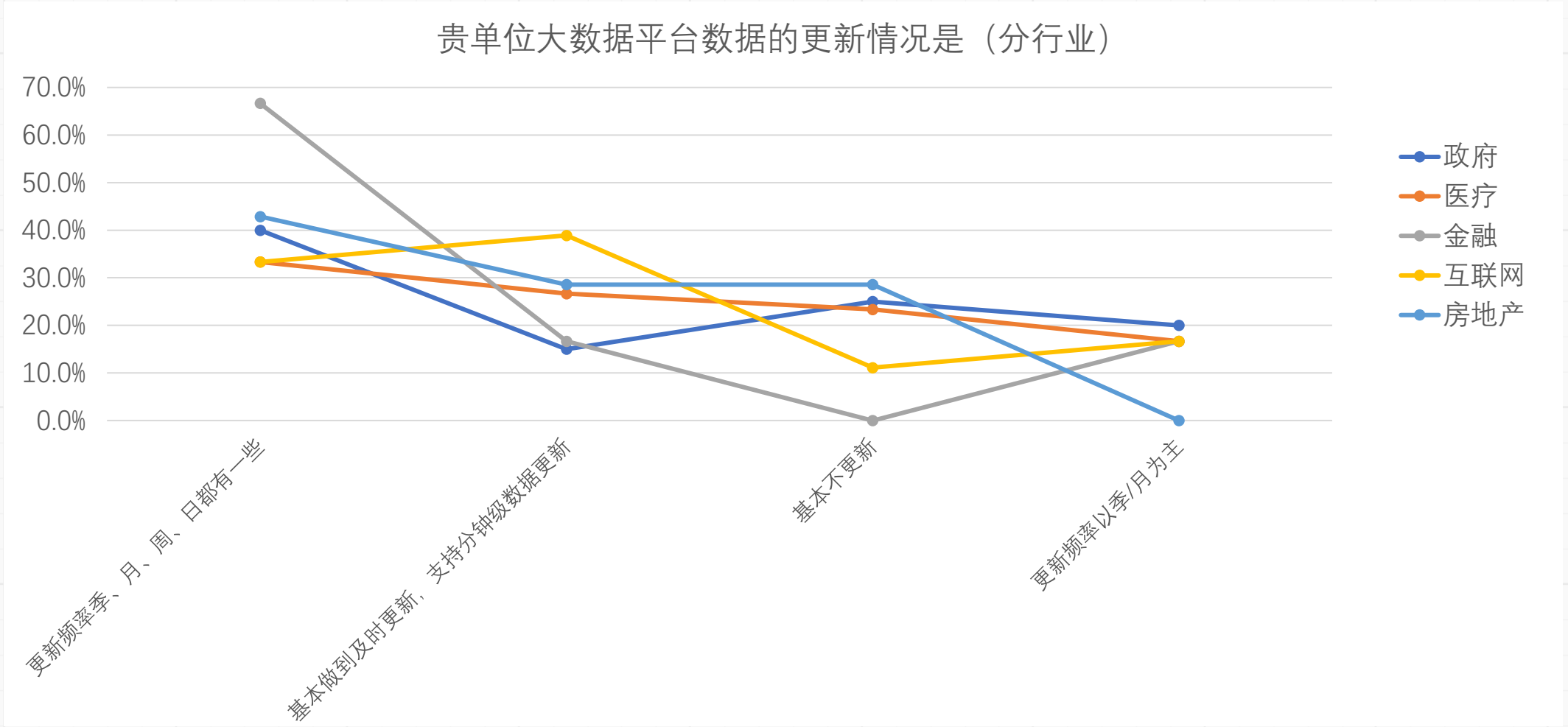
数据实时性有待提高

80.4%的企业大数据平台设立了数据更新机制，但能做到分钟级数据更新的不到30%，其余的以季、月、周、日定期更新为主。



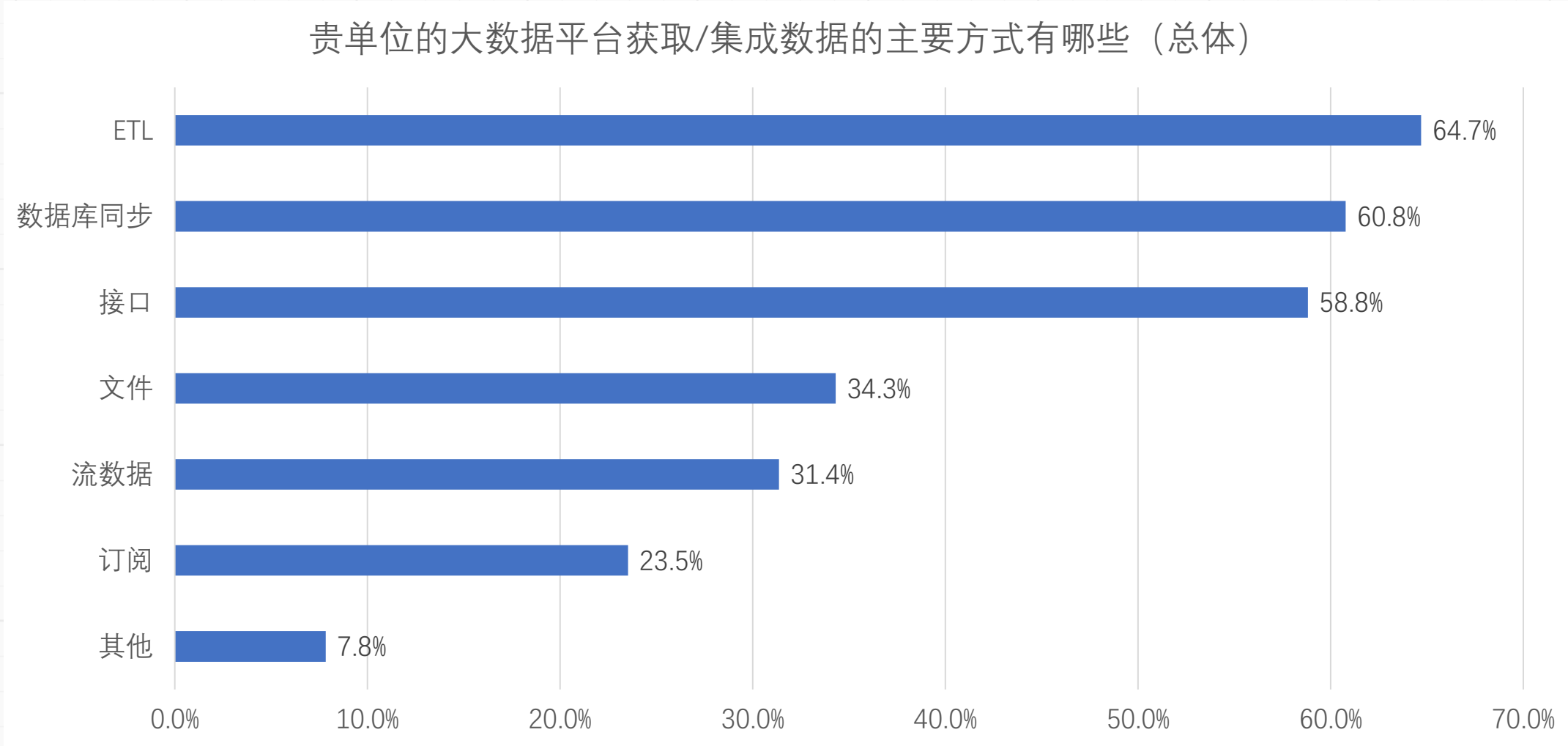
互联网数据更新实时性较好，其他行业有待提高

其中，互联网企业能做到即时更新、分钟级更新的比率为38.9%，房地产、医疗、金融、政府行业能做到及时更新，支持分钟级更新的比率分别为28.6%、28.5%、16.7%、15%。



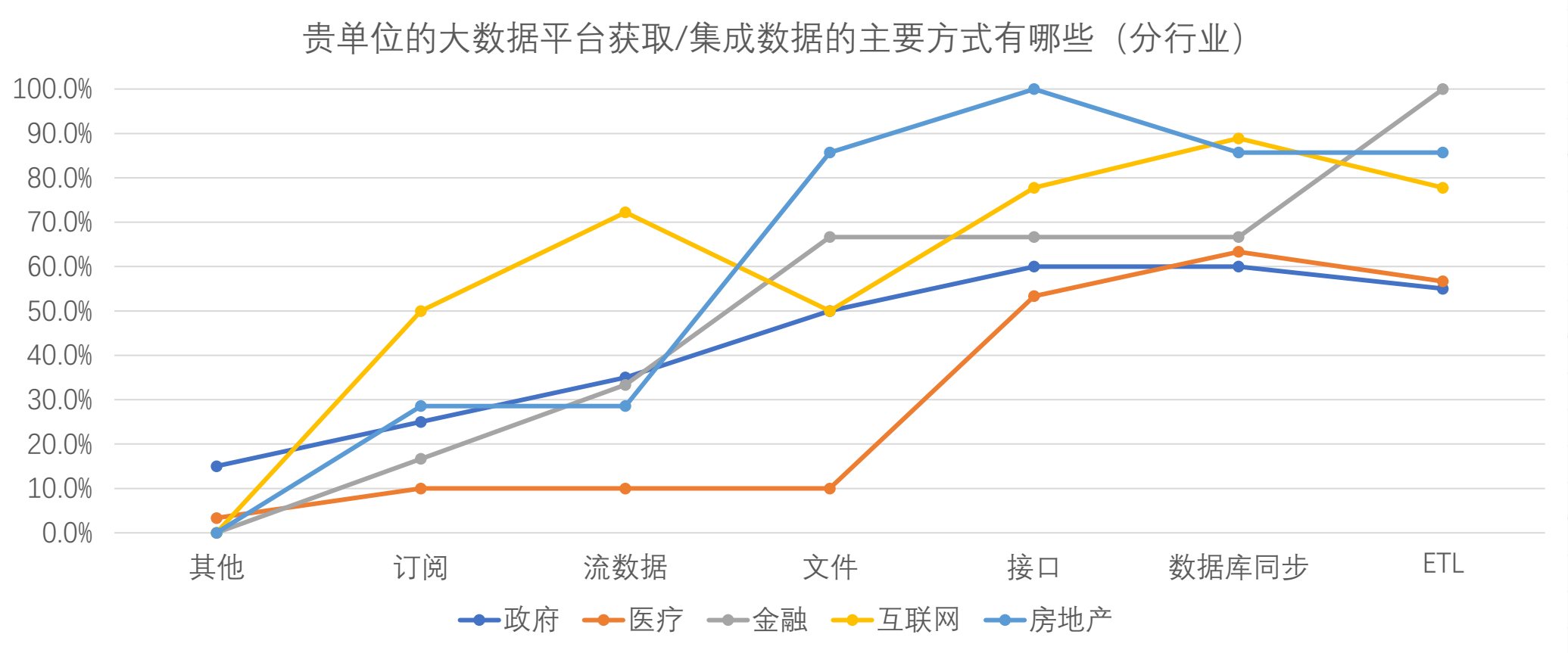
数据集成的主要方式是ETL、数据库同步和接口调用

数据集成方式第一梯队为ETL、数据库同步、接口，第二梯队为文件、流数据、订阅。



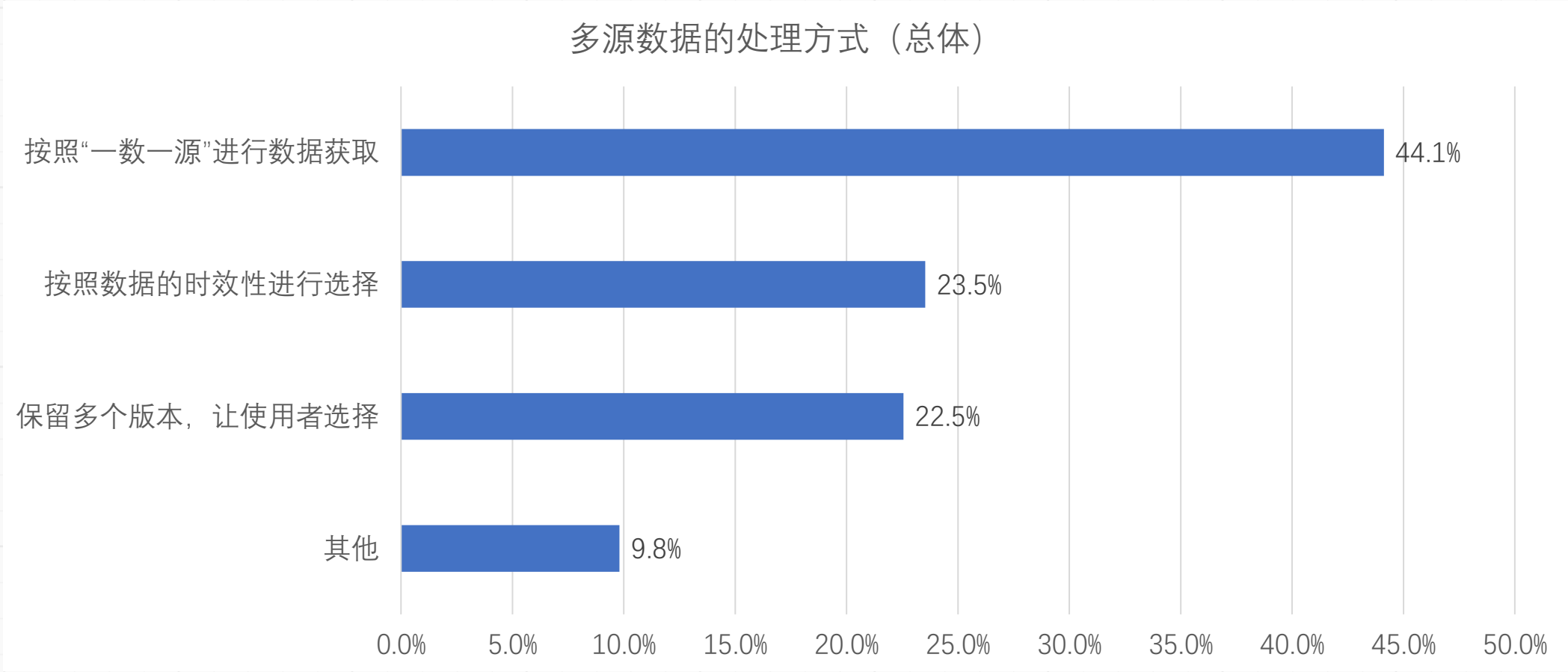
各行业数据集成方式呈多样性，同时体现出行业差异性

数据集成方式体现出不同行业业务特点，金融行业普遍采用ETL、数据库同步、文件和接口集成数据；房地产行业普遍采用接口、文件、数据库同步集成数据；互联网行业以数据库同步、ETL、接口、流数据集成数据。



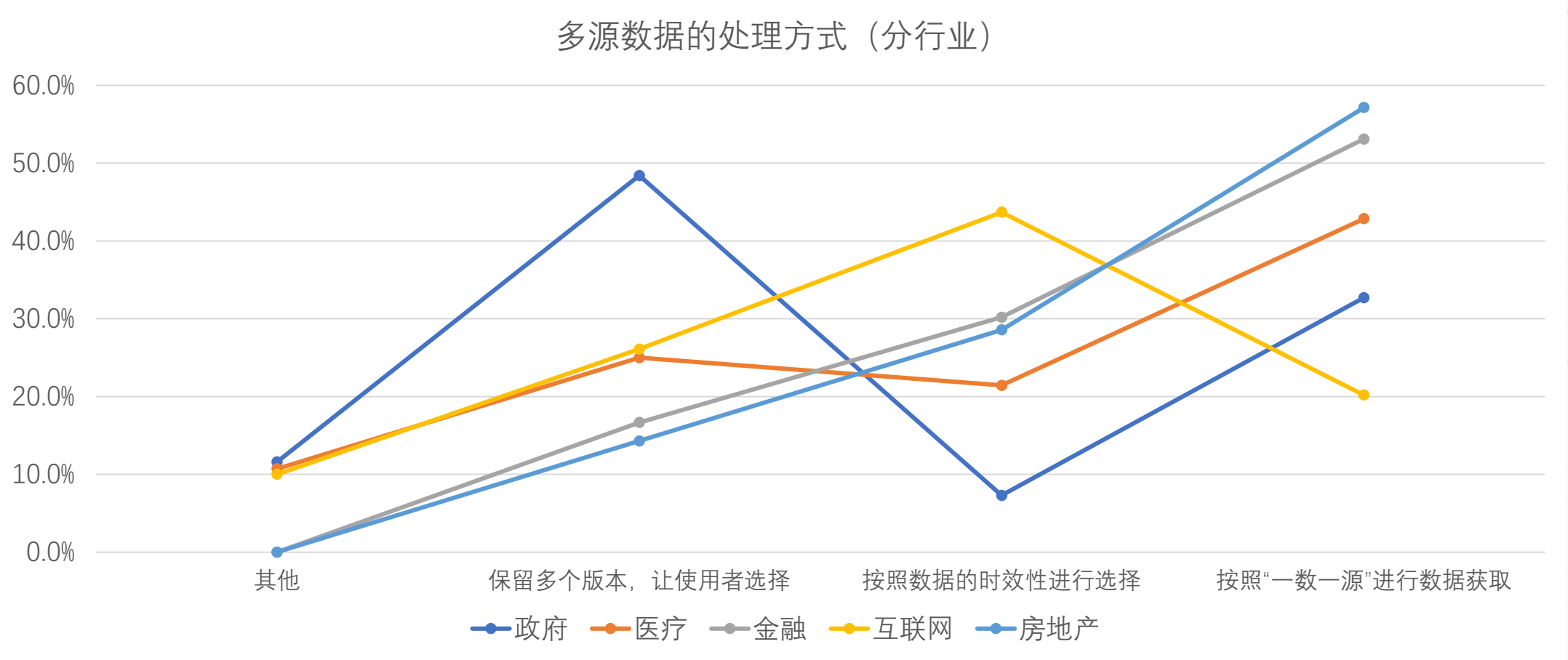
多源数据处理机制倾向“一数一源”

44.1%按照“一数一源”的方式获取数据； 23.5%按照数据的时效性进行选择；22.5%的企业会保留数据的多个版本，让使用者根据使用场景进行选择



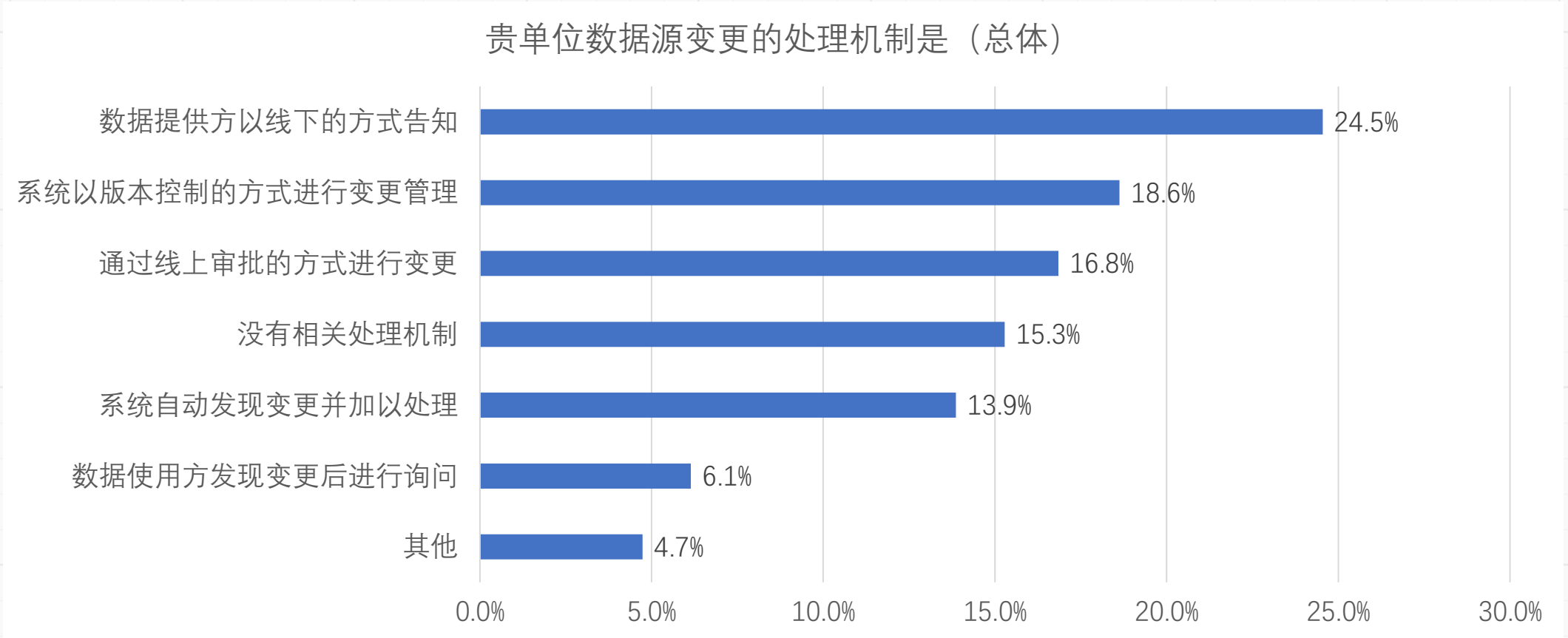
多源数据政府侧重由使用者选择，互联网侧重根据时效性选择

政府机构的数据获取的方式主要是保留数据的多个版本，让使用者选择，比率为48.4%；互联网企业比较倾向于根据数据的时效性进行选择，比率为43.7%。



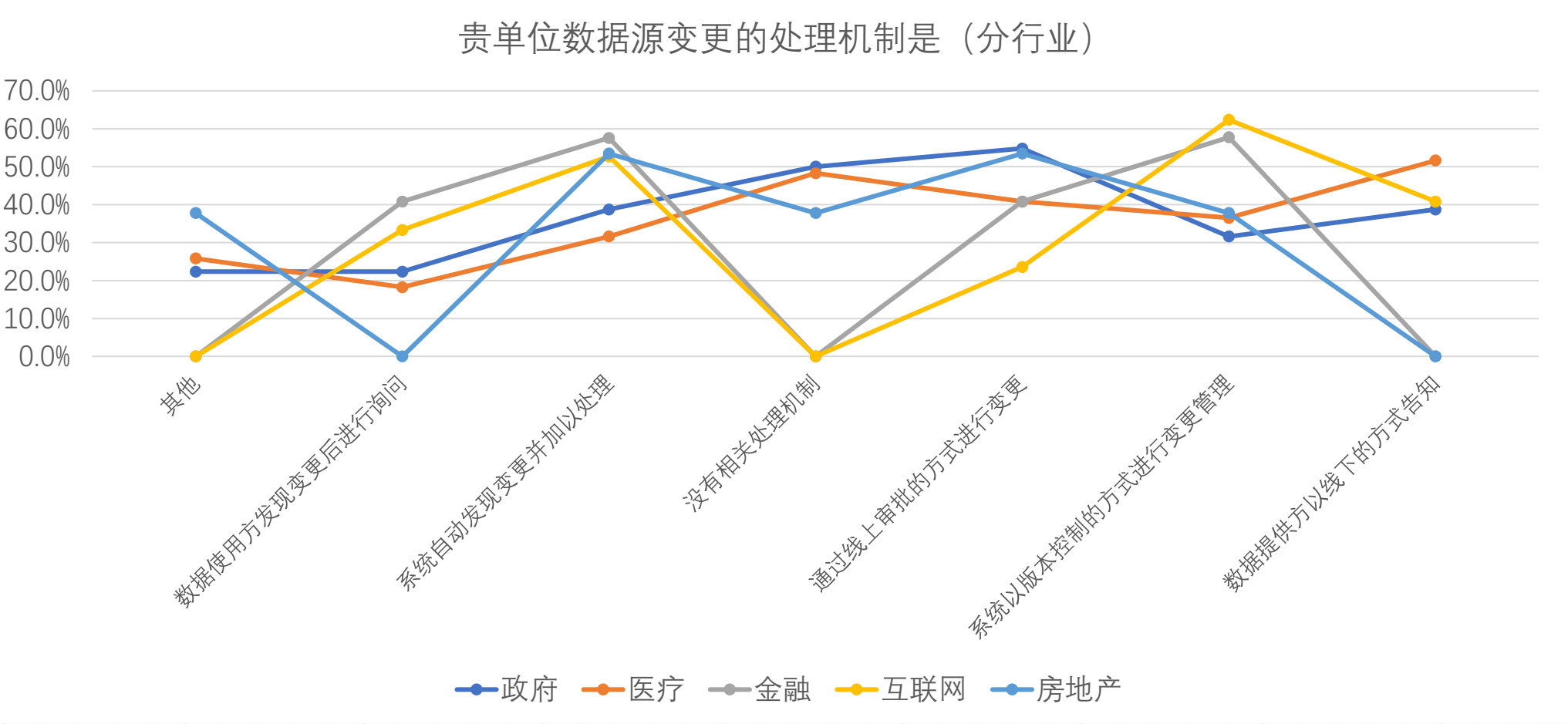
数据源变更近一半还需要人工处理，不能实现自动化

47.4%的机构数据源变更的处理机制是人工处理，32.5%的机构数据源变更由系统自动处理，同时还有15.3%的没有相关数据源变更处理机制。



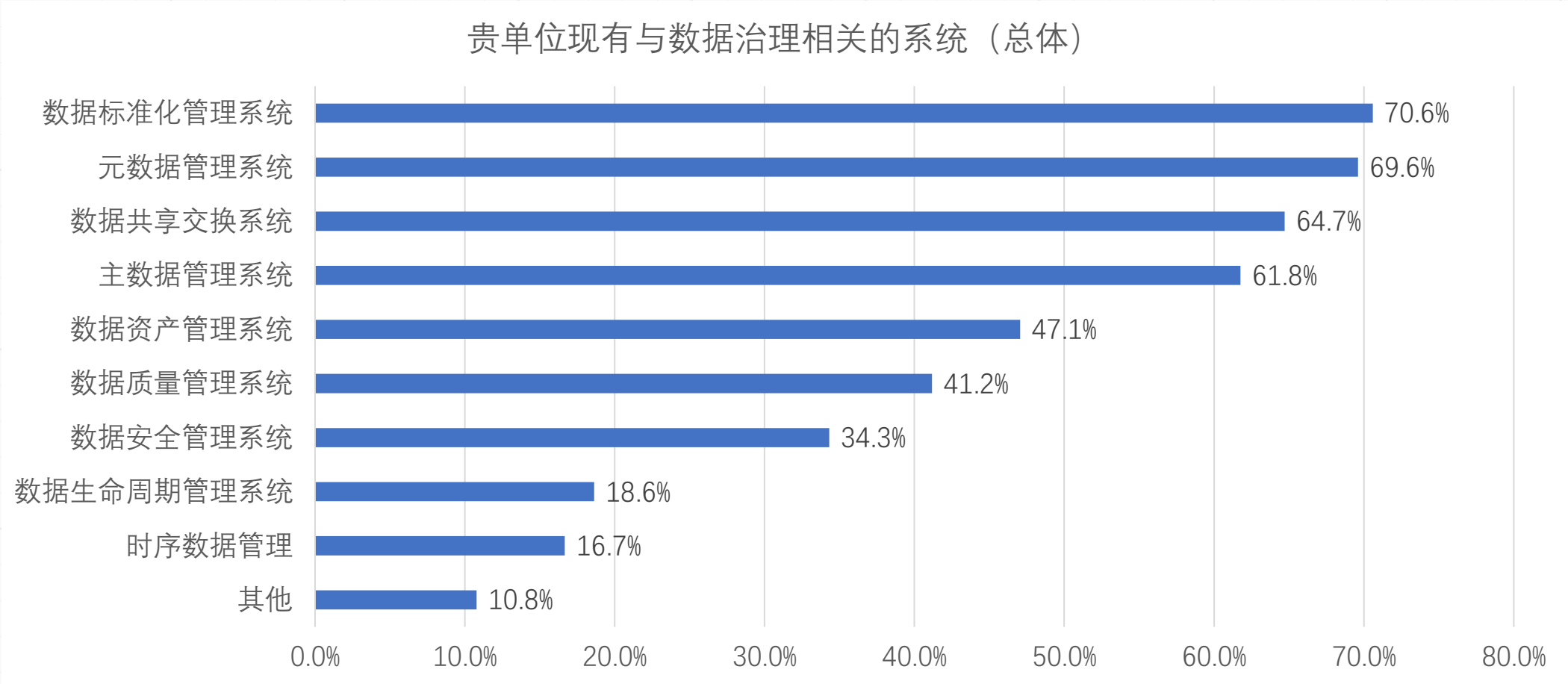
金融、互联网行业数据源变更处理自动化程度更高

其中，互联网、金融的数据源变更处理自动化比率分别达到62.4%、57.6%。值得注意的是，政府、医疗、房地产数据源变更没有设计相关处理机制的比率分别为48.3%、48.1%、37.8%



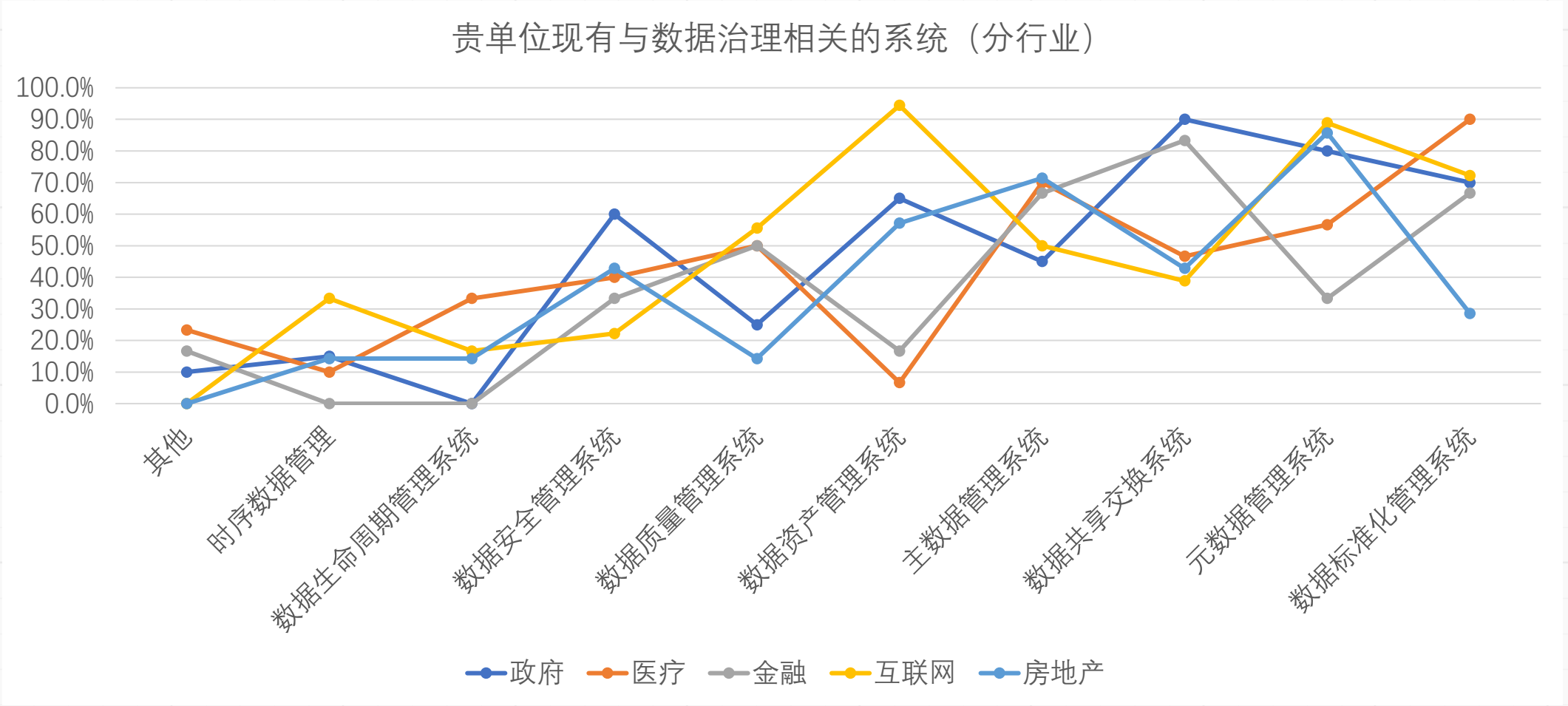
数据治理目前侧重系统建设，治理管理有待加强

60%以上的机构集中于数据标准化管理、元数据管理、数据共享交换、主数据管理等系统的建设，而在数据资产管理、数据质量、数据安全、数据生命周期、时序数据管理等数据治理能力建设不足50%。



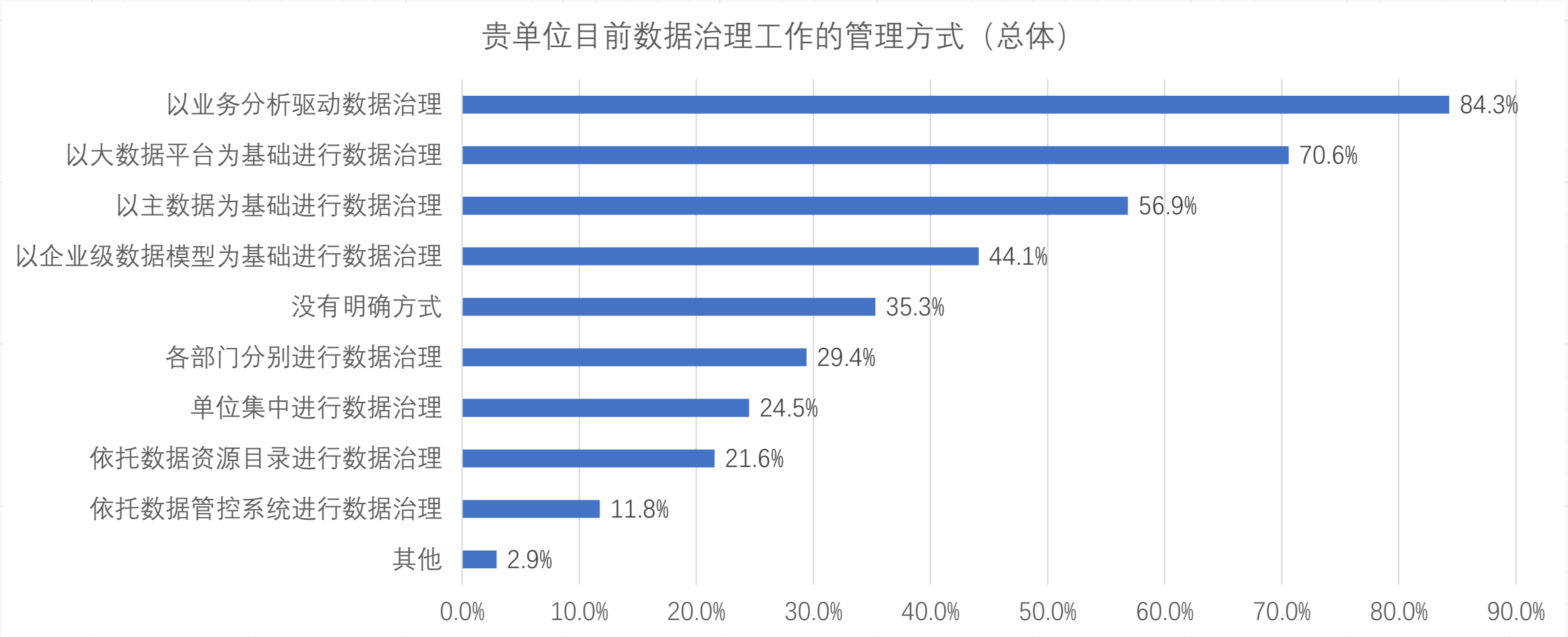
数据资产、数据共享交换、元数据系统建设呈两极化分布

政府和金融注重数据安全管理和数据共享交换；互联网注重数据资产管理和元数据管理；医疗注重数据标准化管理和主数据管理。



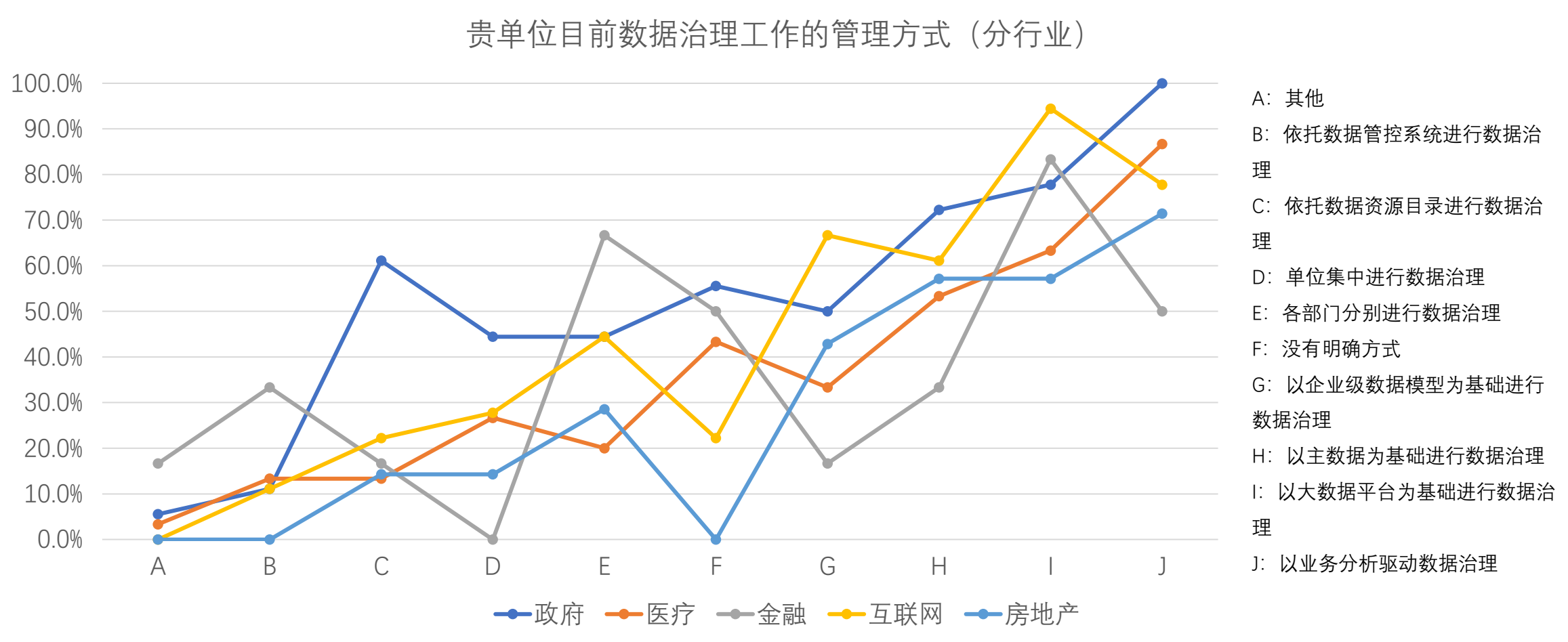
业务驱动是数据治理的核心动力

84.3%的机构以业务分析驱动数据治理工作；其次， 70.6%的机构以大数据平台进行数据治理工作；同时还有35.3%的机构没有明确数据治理工作管理方式，值得重视。



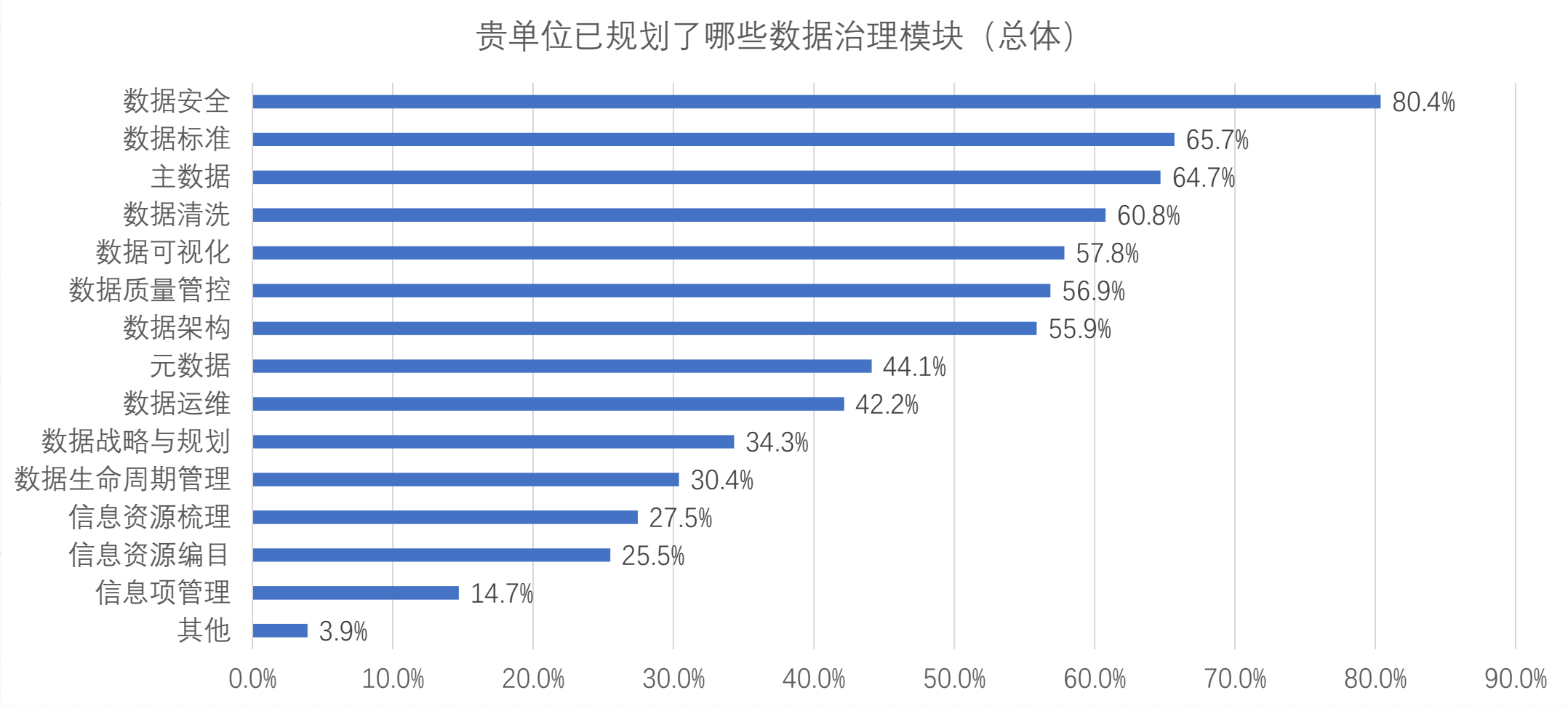
不同行业数据治理工作管理方式存在较大差异

金融行业既有集中式又有分散式的数据治理方式；政府更侧重依托数据资源目录开展数据治理工作；互联网行业94.4%以大数据平台为基础进行数据治理。



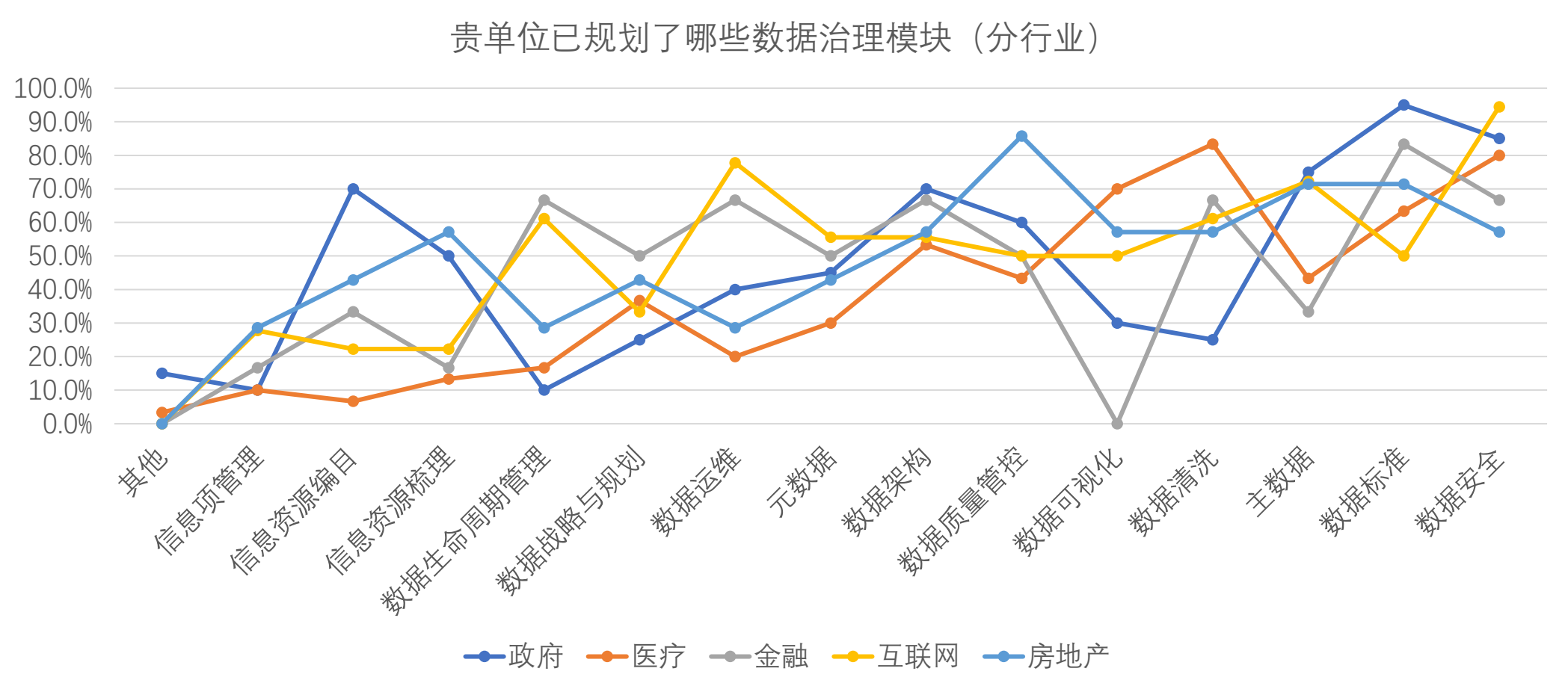
数据安全、数据质量是当前数据治理的建设重点

80.4%的机构都规划了数据安全相关能力建设，同时提升数据质量也是数据治理的重点，数据标准、数据清洗、数据质量管控等数据治理相关模块的比率分别为65.7%、60.8%、56.9%。



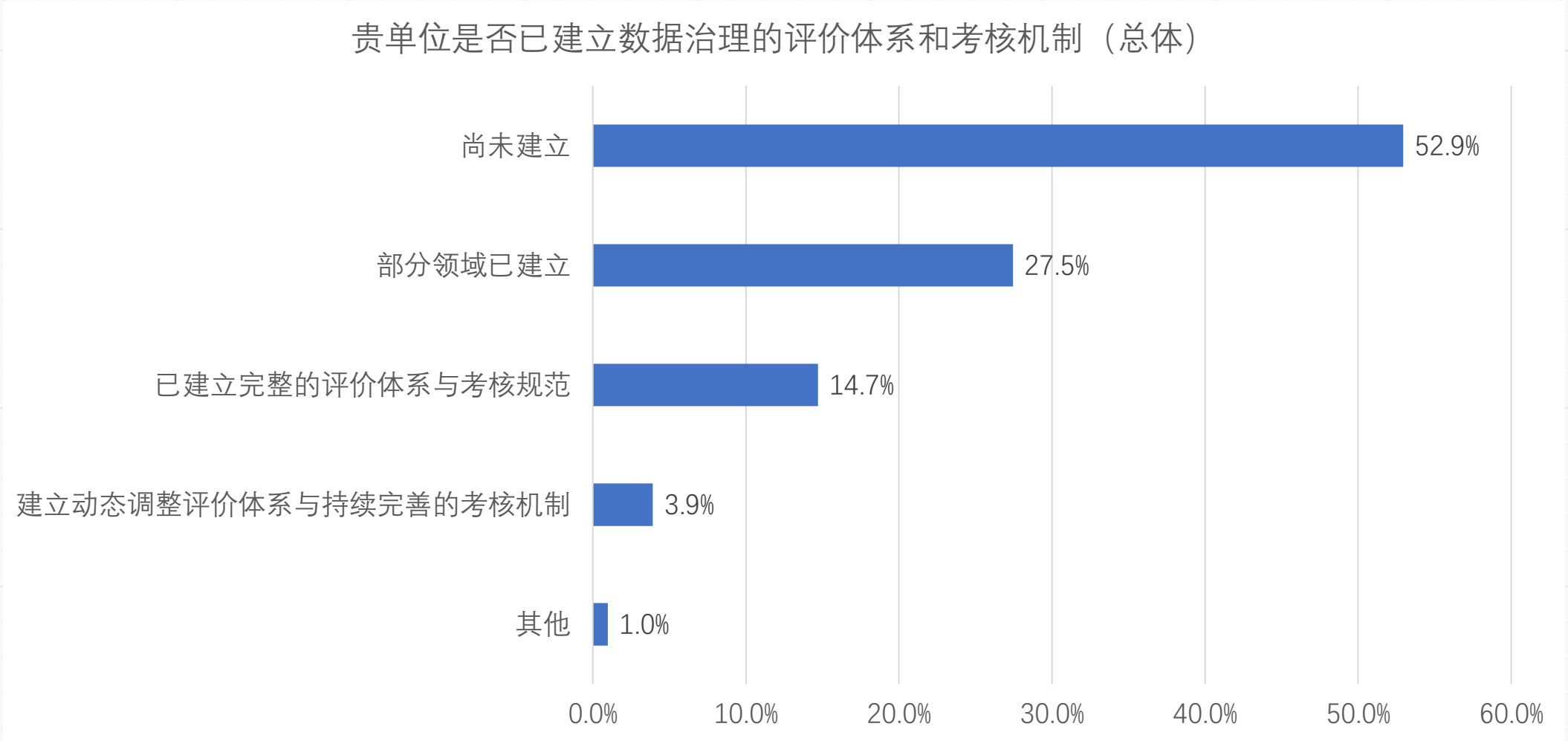
各行业对数据治理的规划重点差异性较大

政府侧重于数据标准和信息资源编目；医疗行业侧重于数据清洗；金融行业侧重数据标准和数据架构；互联网更侧重数据运维；房地产行业更侧重数据质量管控。



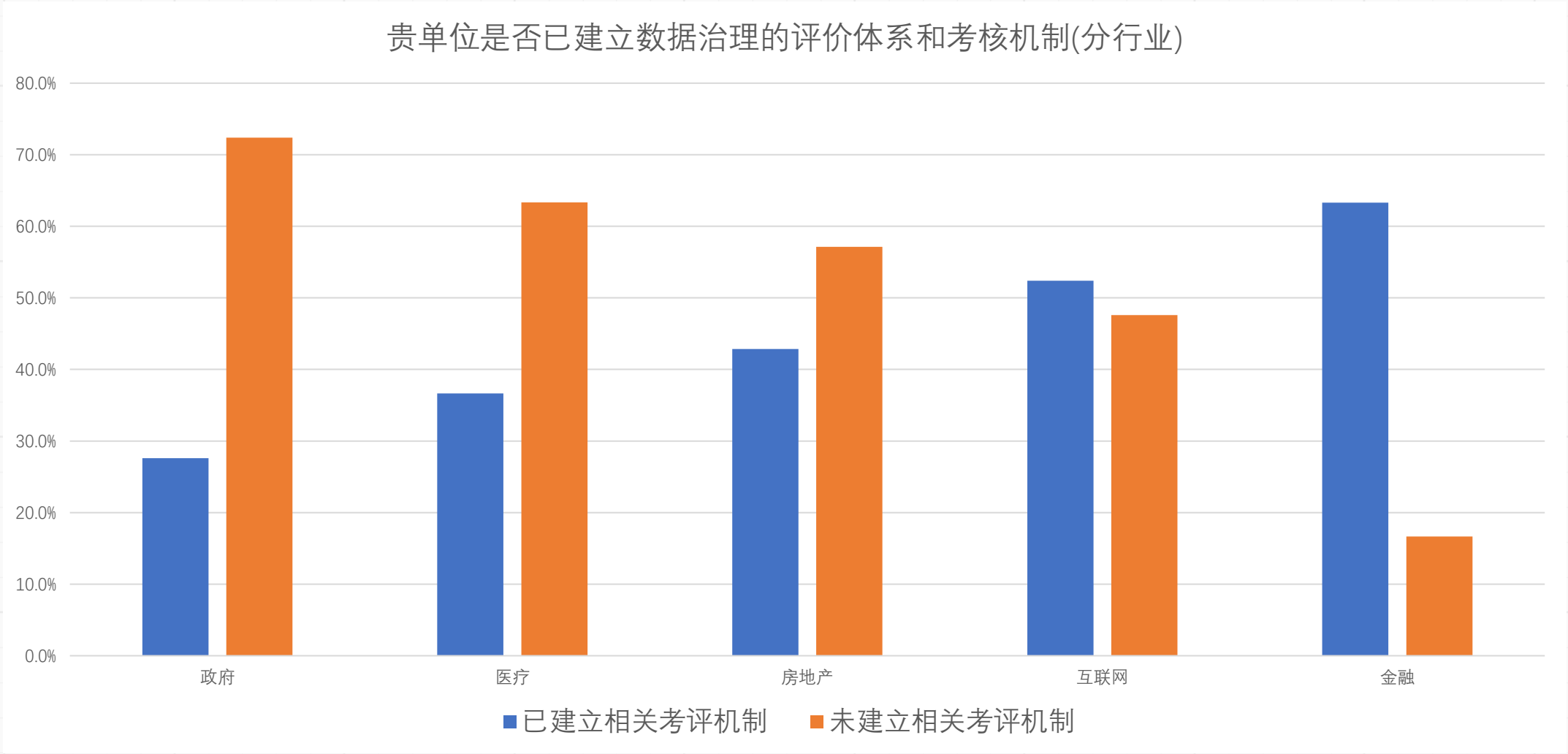
数据治理考评机制严重缺位

52.9%的机构没有建立数据治理相关的评价体系和考核机制，27.5%只在部分领域建立，数据治理评价体系和考核机制的建立任重道远。



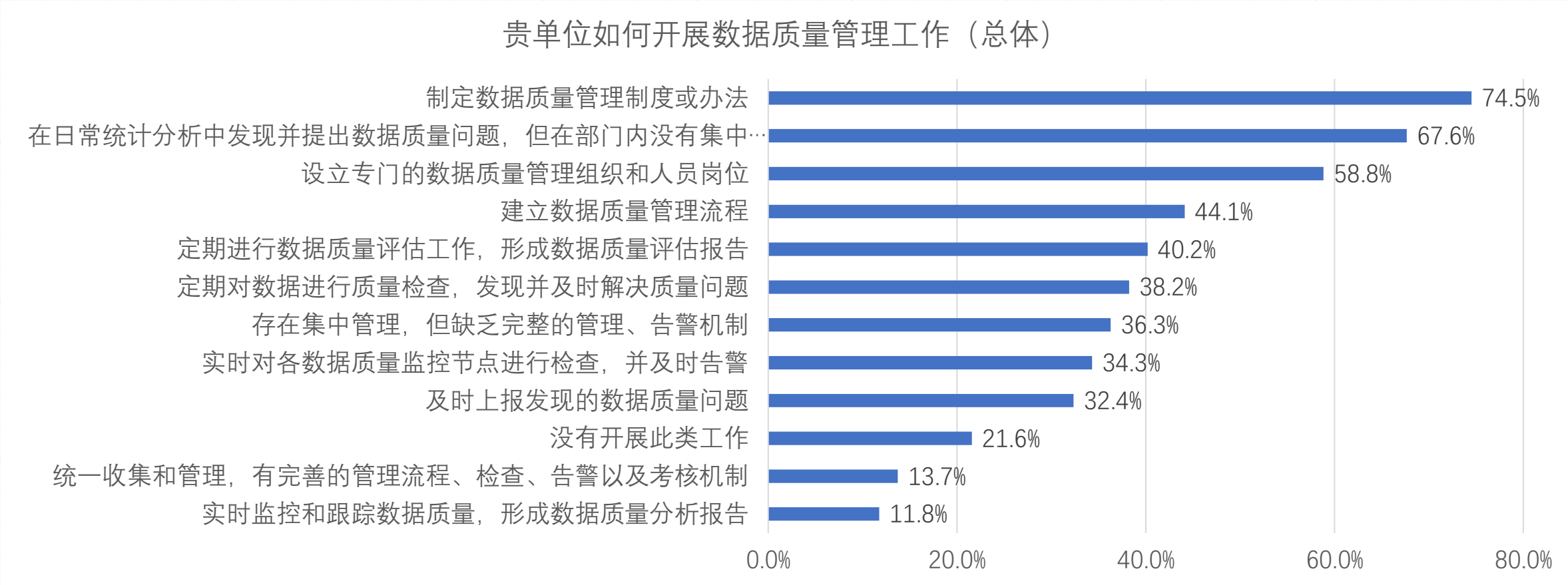
金融行业评价体系和考核机制建设表现相对较好

金融和互联网数据治理的评价体系和考核机制的普及率最高；政府、医疗行业的普及率相对较低。



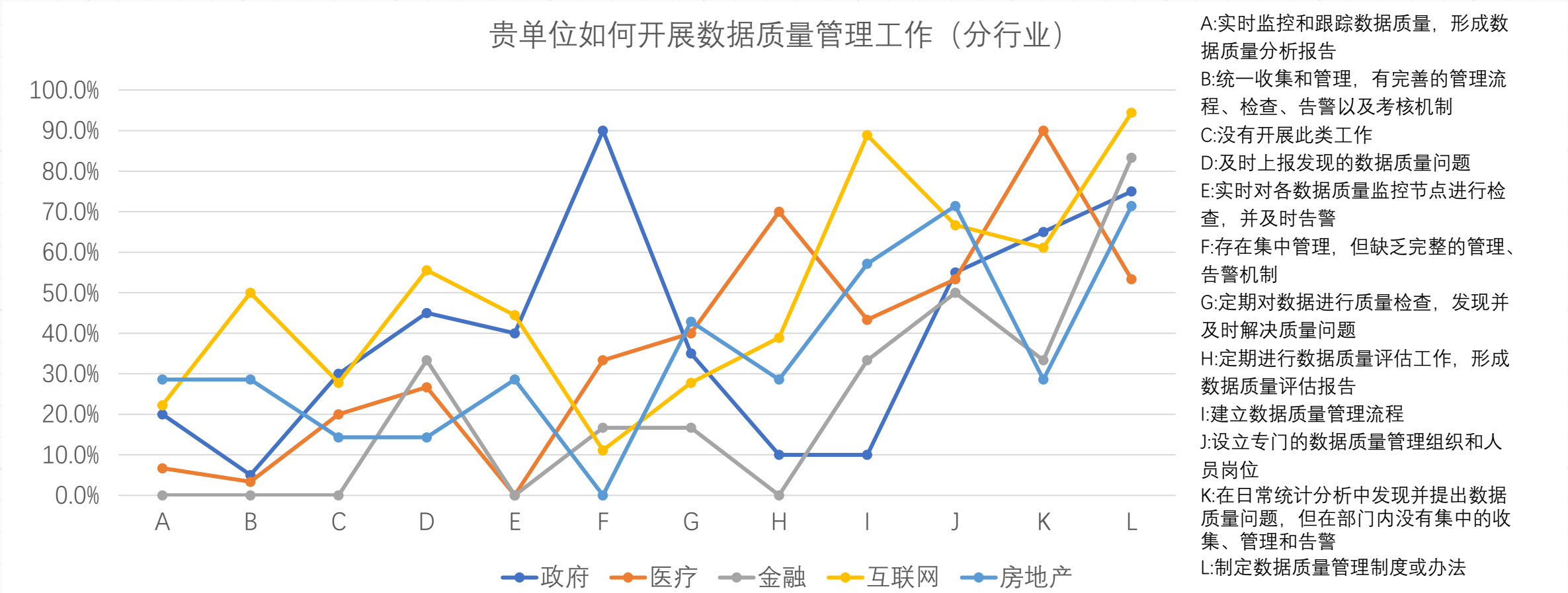
数据质量管理工作最大困境：有制度无执行

数据显示，74.5%的机构制定数据质量管理制度或办法，但同时在日常统计分析中发现并提出数据质量问题，但在部门内没有集中的收集、管理和告警的比率达到67.6%。58.8%的机构设置了专人专岗，但同时存在缺乏完整的管理、告警机制的比率也达到36.3%。



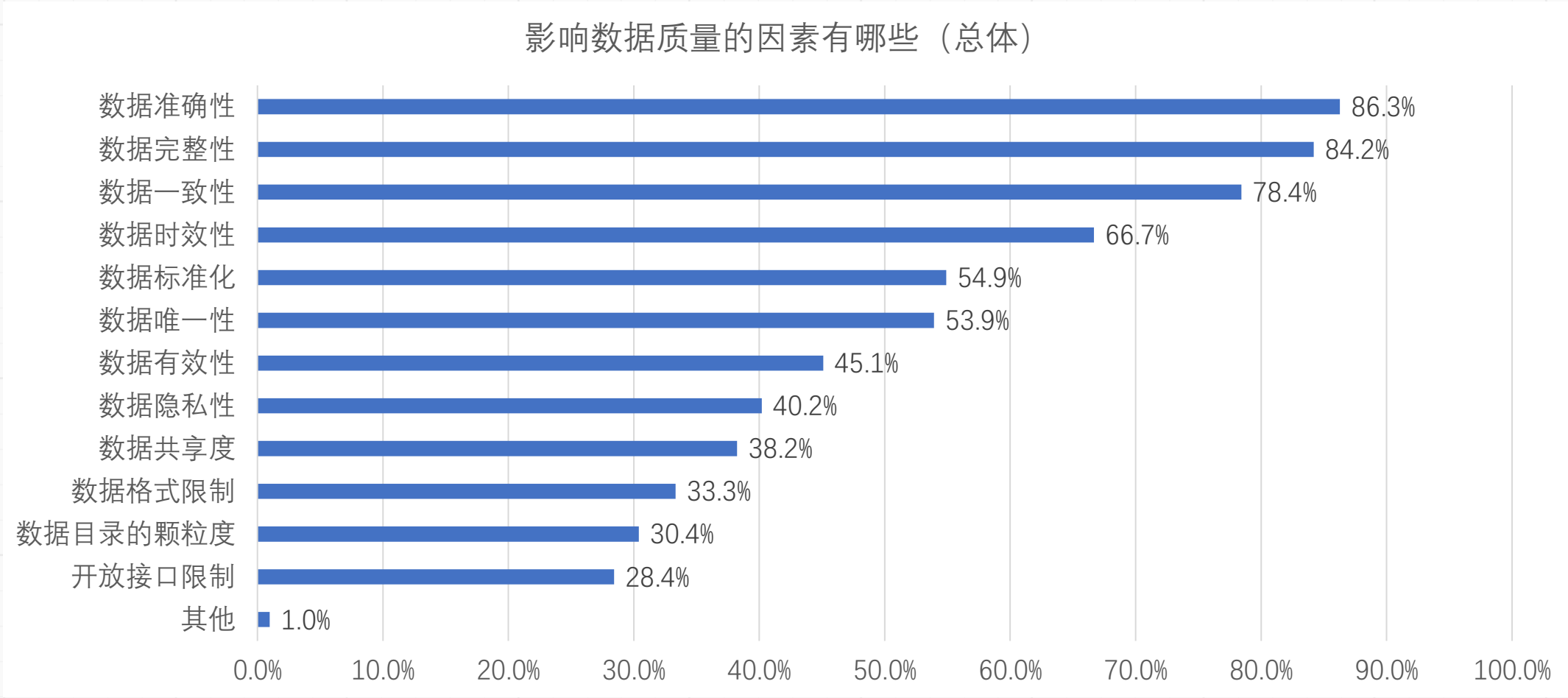
互联网行业数据治理管理工作制度最健全、执行最到位

政府、医疗普遍存在有制度但无落地的情况，政府在“存在集中管理，但缺乏完整的管理、告警机制”的比率分别达到90%、33.5%；医疗“在日常统计分析中发现并提出数据质量问题，但在部门内没有集中的收集、管理和告警”的比率达到90%。



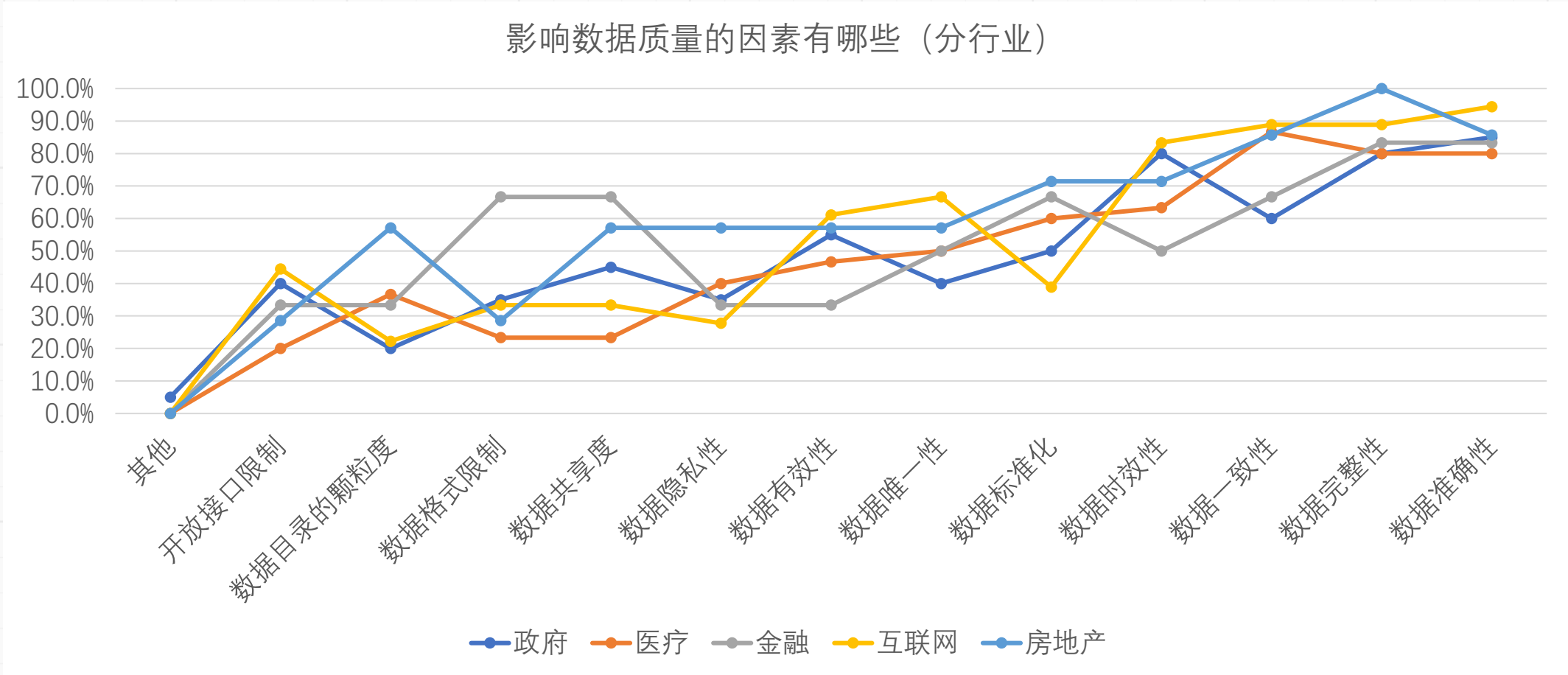
影响数据质量的因素复杂化多元化

数据准确性、数据完整性、数据一致性、数据时效性都是影响数据质量的突出原因，比率均在60%以上；其次数据标准化、数据唯一性也是影响数据质量的主要原因，比率均在50%以上。



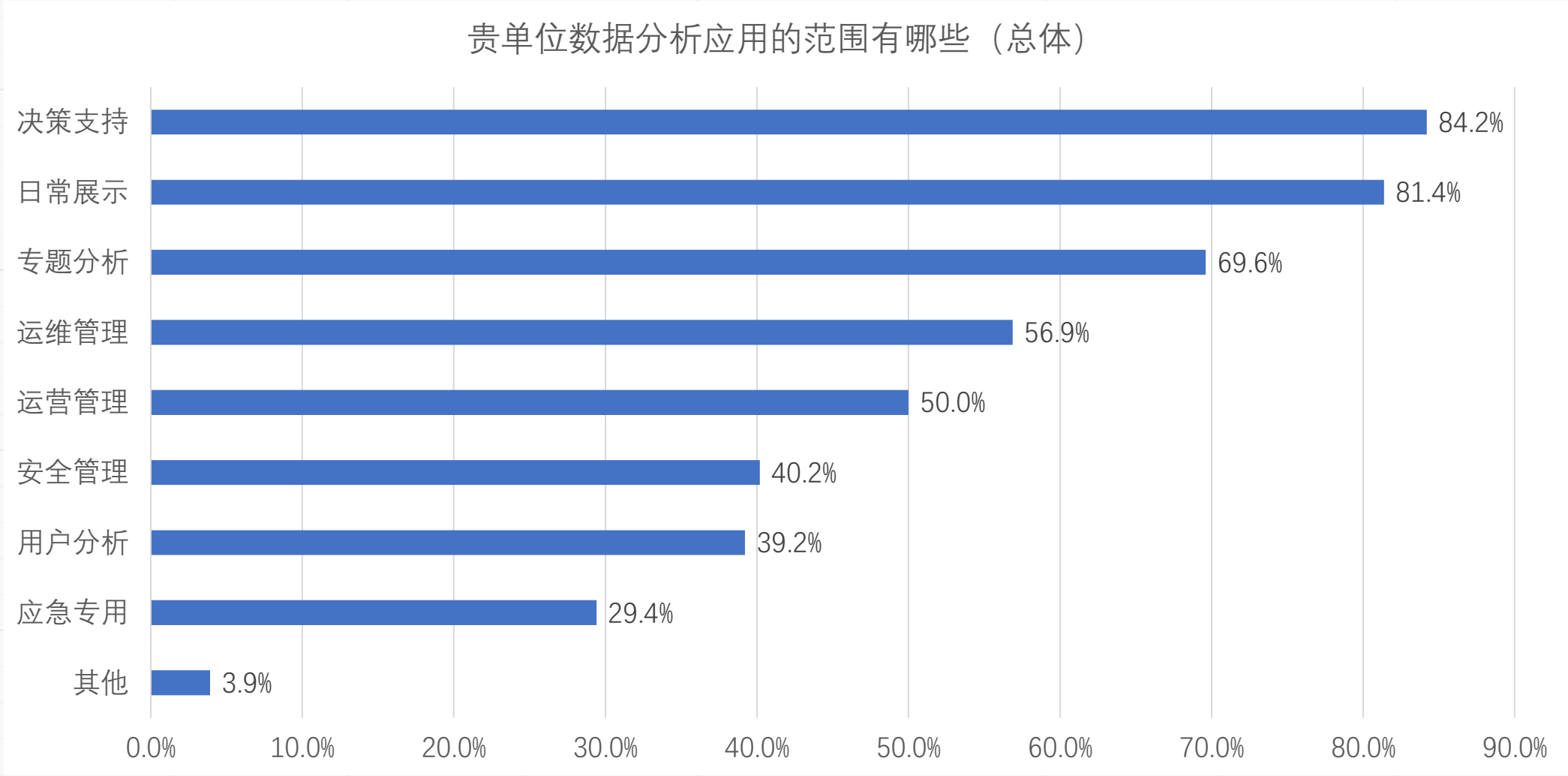
影响数据质量的原因具有共性，同时也有行业差异性

就政府行业而言，数据目录的颗粒度是一个重要因素；就金融行业而言，数据格式和数据共享度是影响数据质量的重要因素；就房地产行业而言，数据隐私性和数据完整性是影响数据质量的重要因素。



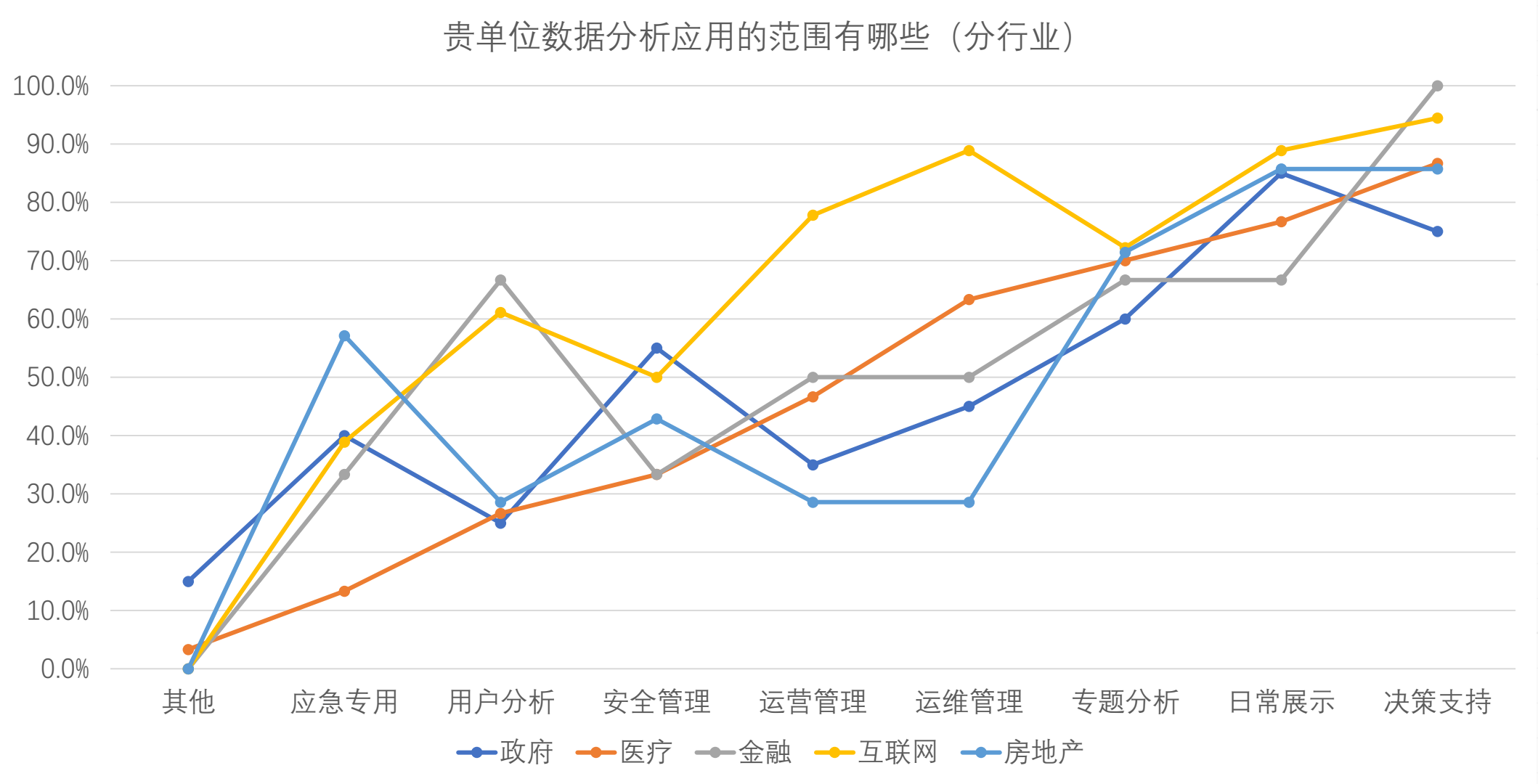
数据分析主要应用场景集中在决策支持和日常展示

各机构数据分析应用场景超过80%以上的需求都集中在决策支持、日常展示。



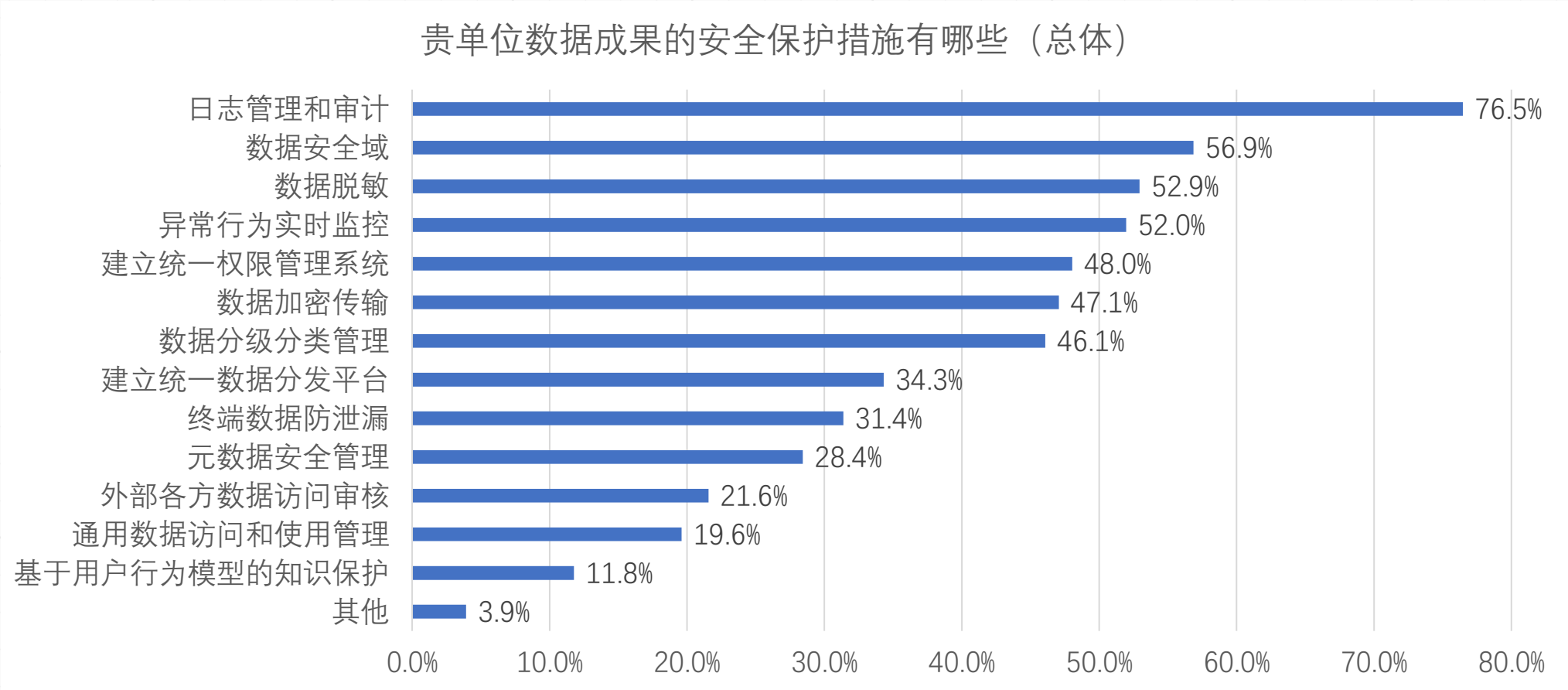
数据分析应用场景与行业属性紧密相关

互联网、金融行业主要用于决策支持和用户分析；政府在应急专用和日常展示方面比率较高。



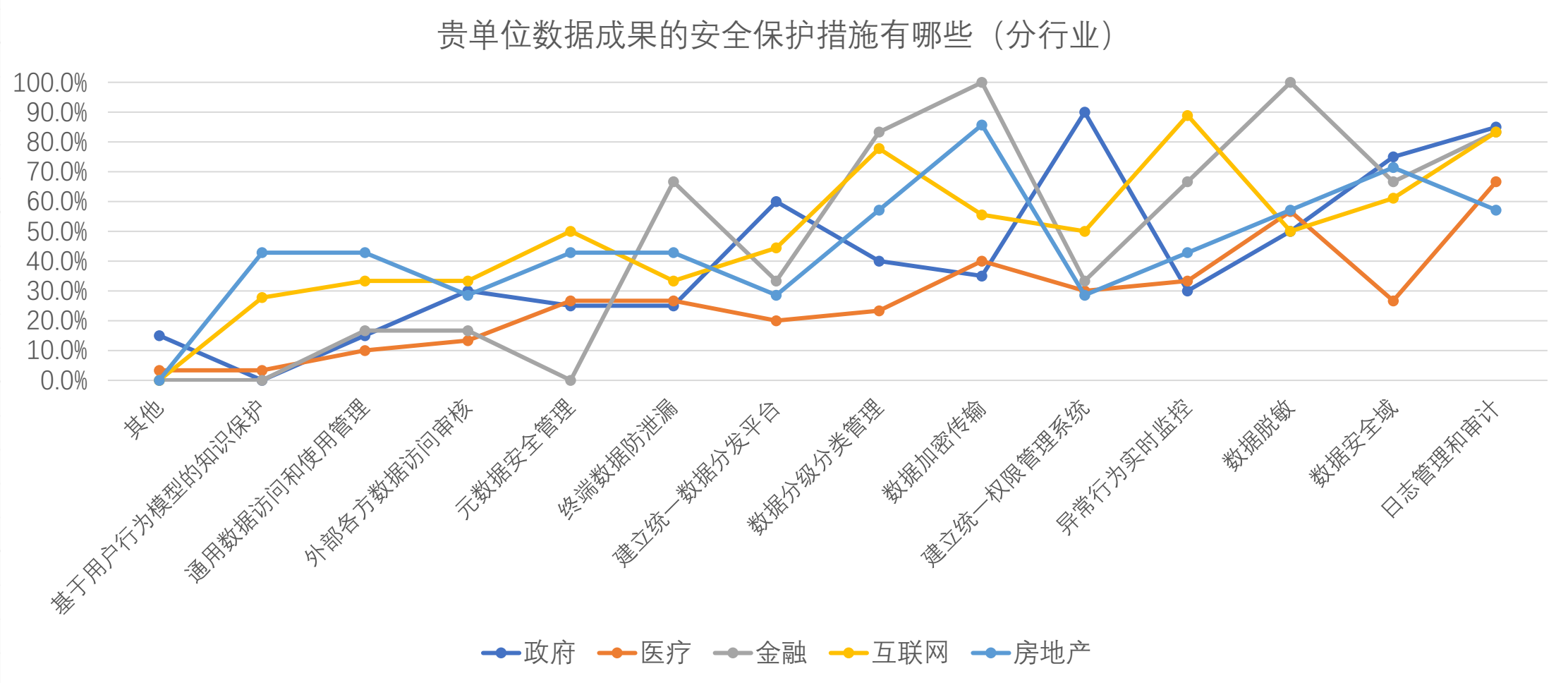
各行业均采用多种手段保护数据成果

76.5%的机构数据成果安全保护措施为日志管理和审计；56.9%的机构数据成果安全保护措施为建立数据安全域；52.9%的机构数据成果安全保护措施为数据脱敏；52%的机构数据成果安全保护措施为异常行为实时监控。



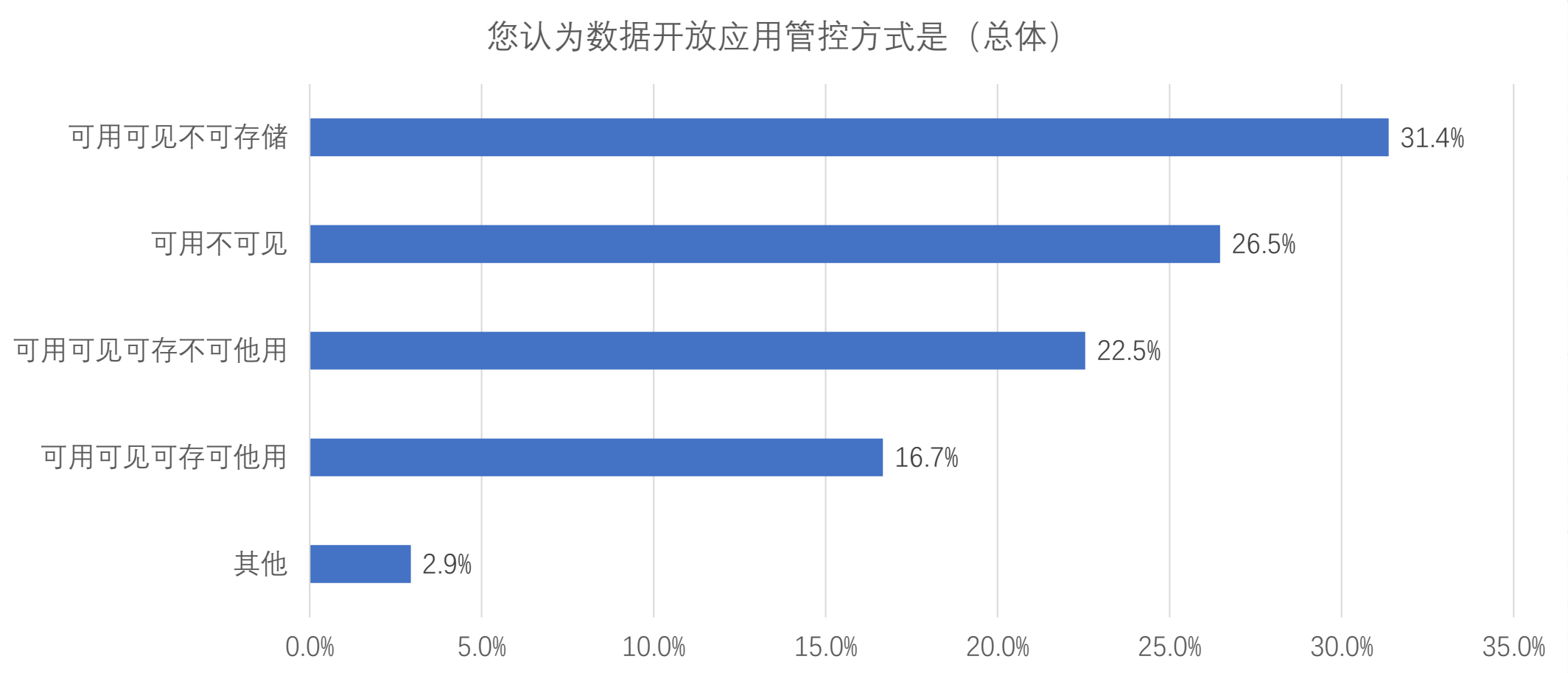
各行业数据成果安全保护措施侧重点有所差异

金融行业数据成果安全保护措施集中在数据脱敏和数据加密传输；互联网行业异常行为实时监控普及率明显高于其他行业；医疗行业的数据安全域普及率为26.7%。



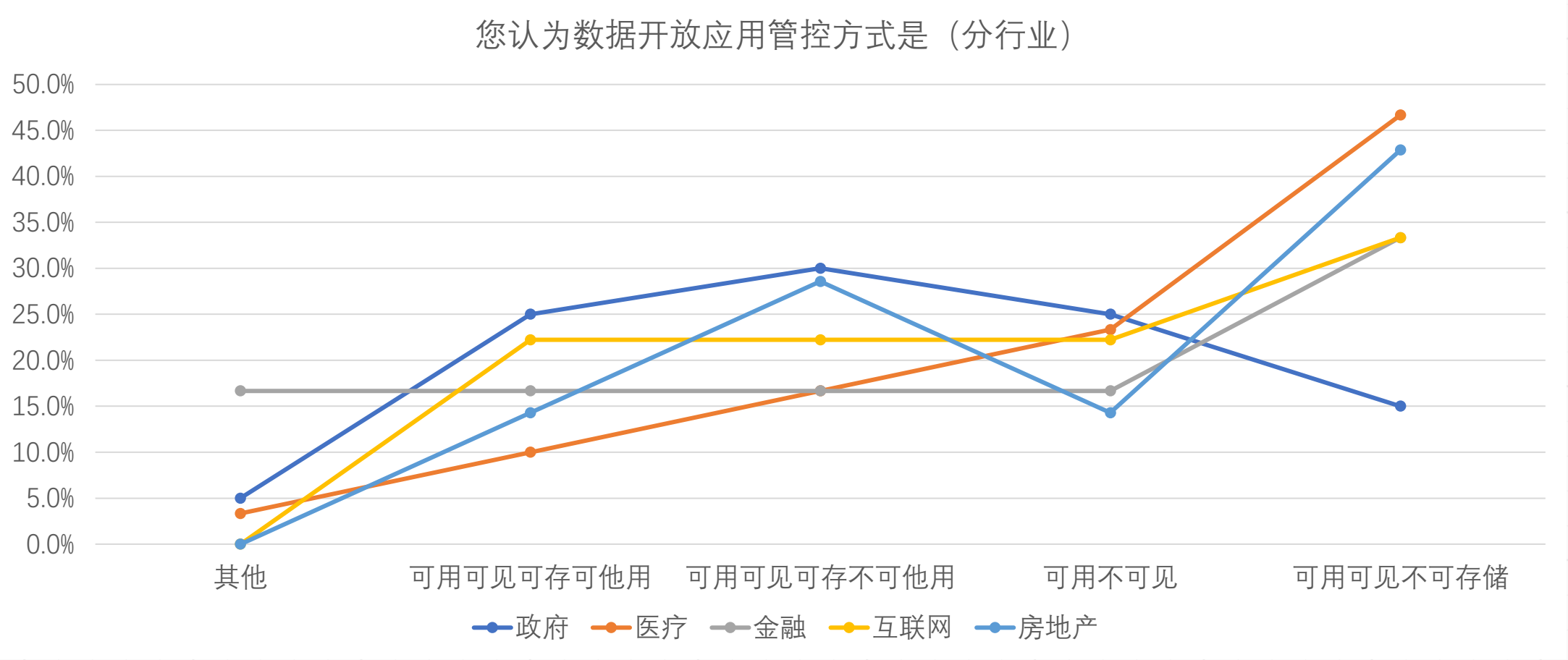
数据开放应用管控方式以可用可见为主

数据开放应用管控的方式，可用可见不可存储、可用可见可存不可他用和可用可见可存可他用的比率分别是31.4%、22.5%、16.7%；可用不可见的比率是26.5%。这将会对安全和隐私保护产生极大挑战。



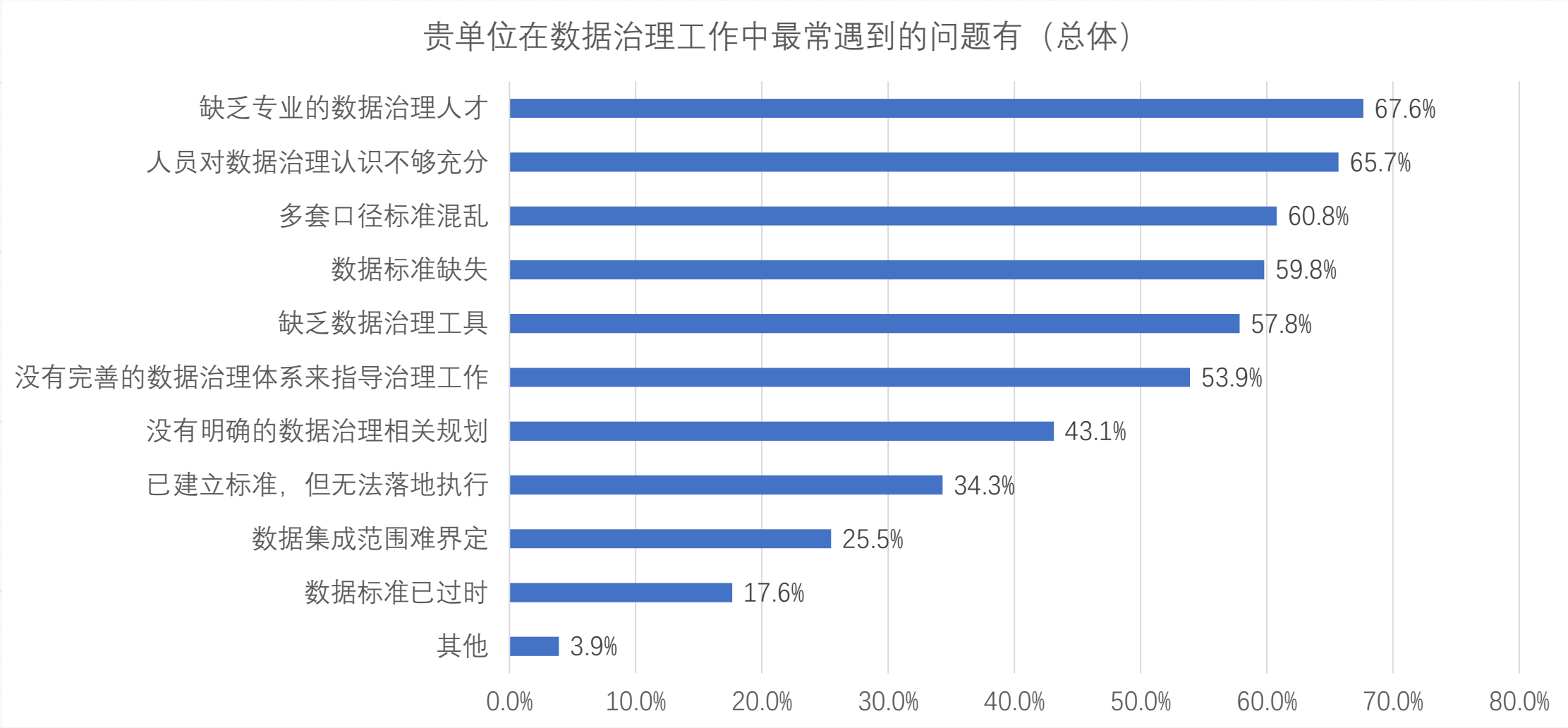
数据应用管控政府以面向公共服务为主，其他行业以自用为主

医疗、房地产、互联网、金融行业的主要管控方式是可用可见不可存储；政府行业的数据开放应用管控方式主要是可用可见可存不可他用。



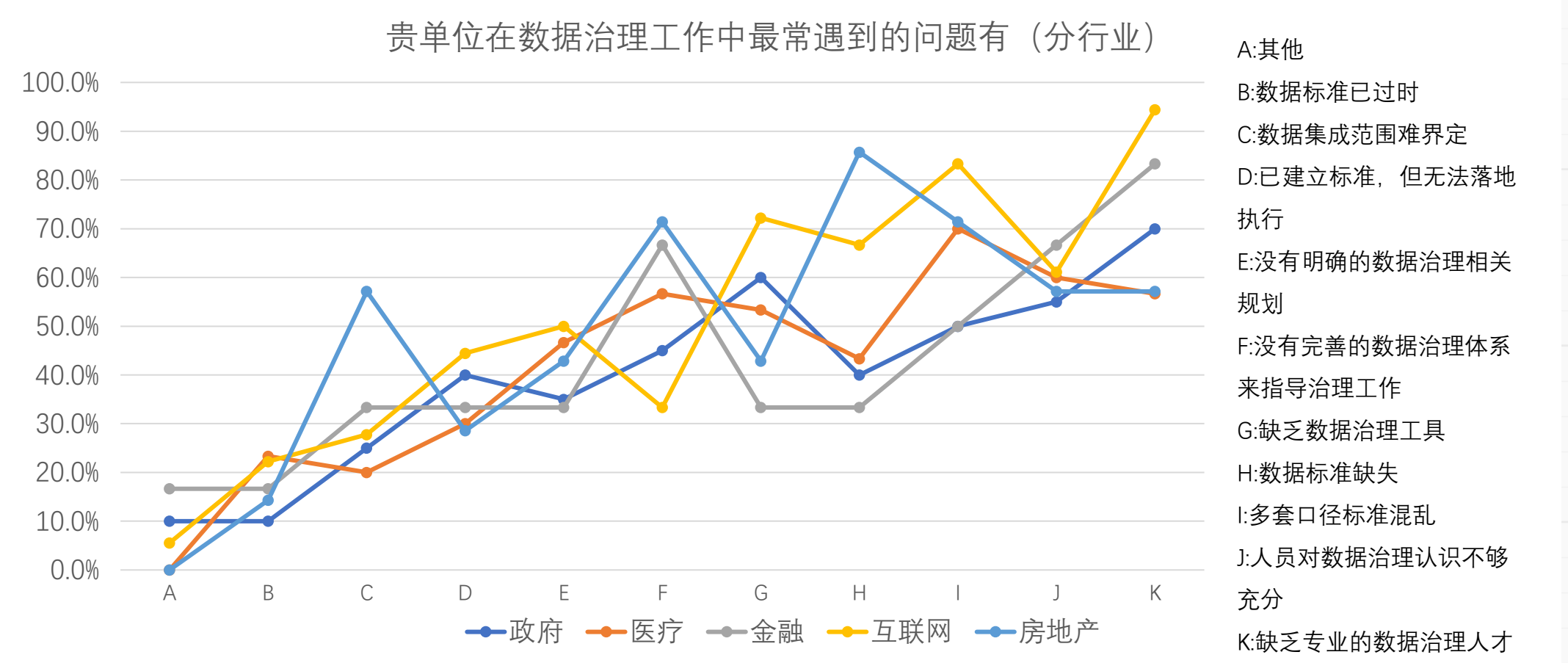
人才问题是数据治理面临的主要问题

人才问题主要体现在：缺乏专业的数据治理人才和人员对数据治理的认识不够充分。



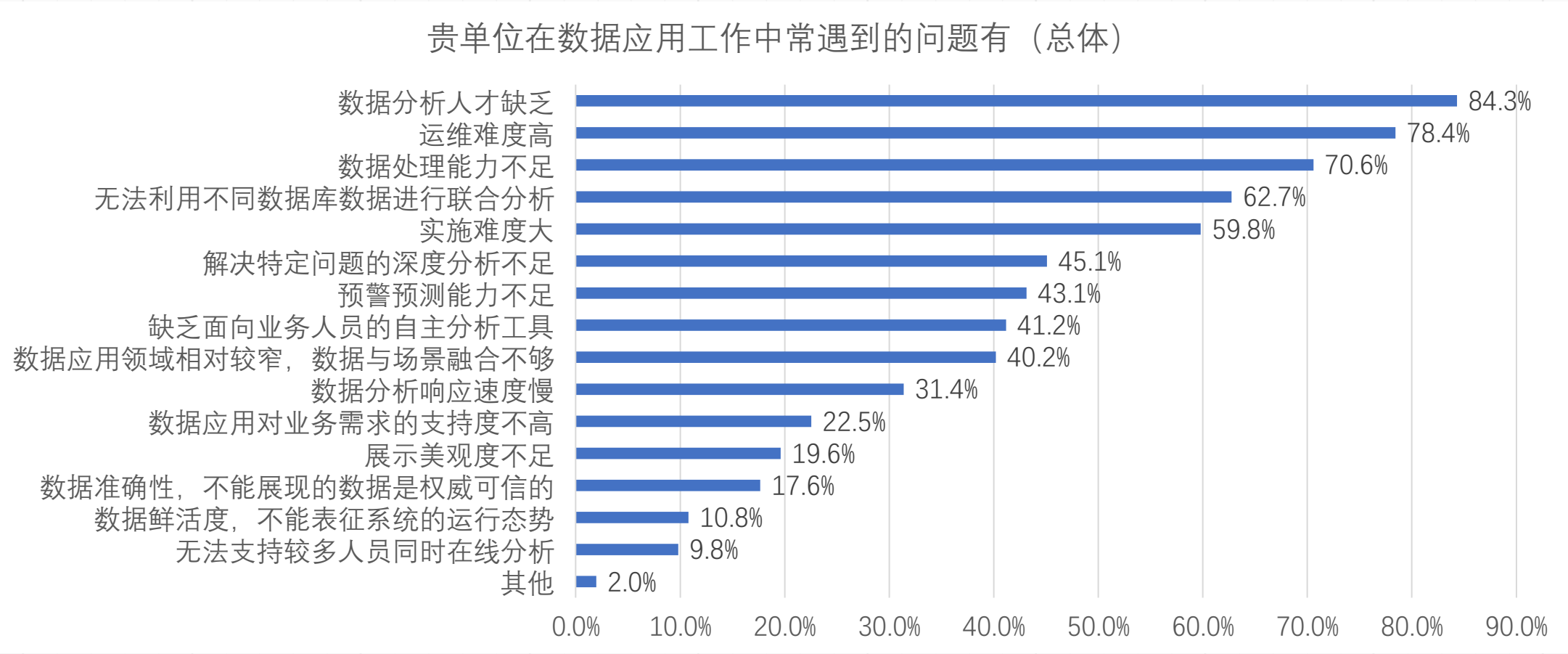
数据治理不同阶段面临的问题有显著区别

其中，互联网数据治理工作开展比较久，面临的问题主要是缺乏专业的数据治理人才、多套口径标准混乱、缺乏数据治理工具；房地产行业刚开展数据治理工作，面临的问题主要是数据标准缺失和没有完善的数据治理体系指导治理工作。



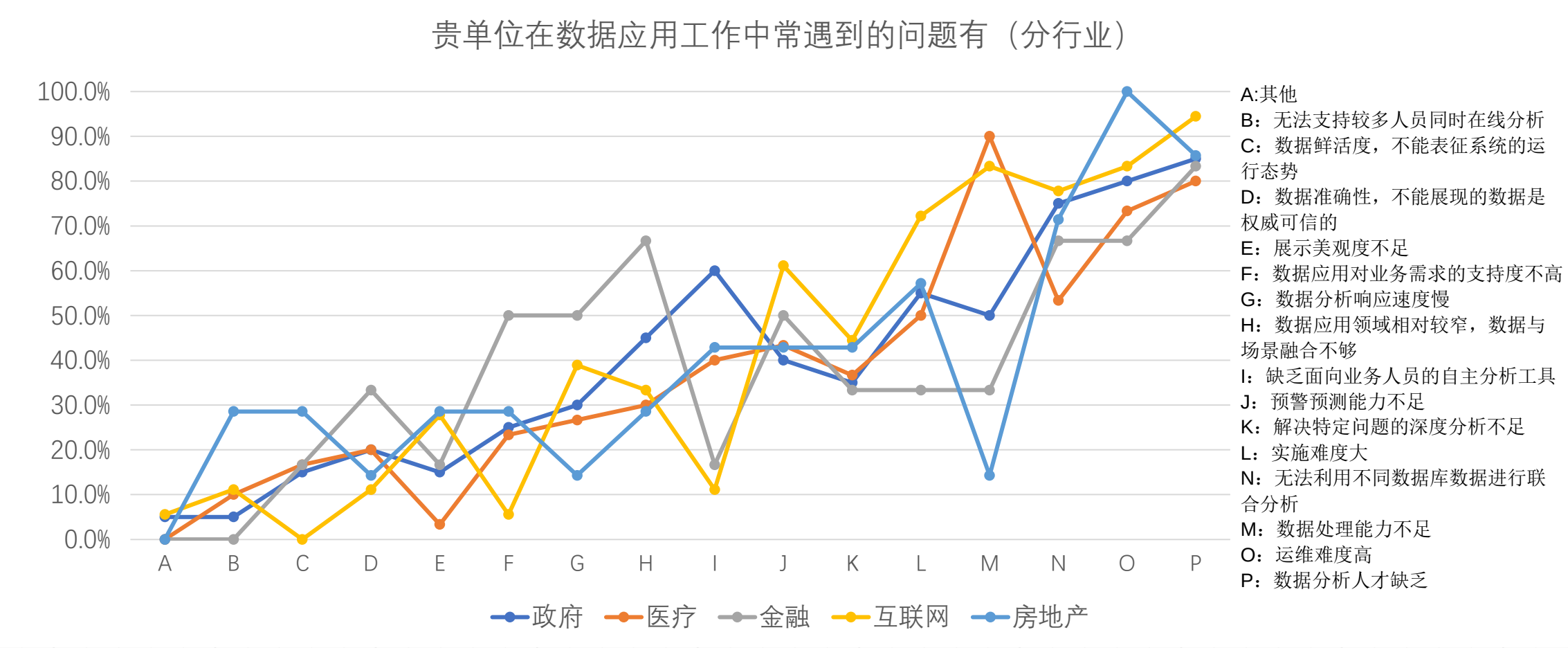
数据分析、处理能力不足导致数据应用难

数据应用工作最常遇到的问题中，84.3%是数据分析人才缺乏的问题；78.4%是数据应用运维难度高的问题；70.6%是数据处理能力不足的问题；62.7%是无法利用不同数据库进行联合分析的问题；59.8%是实施难度大的问题。



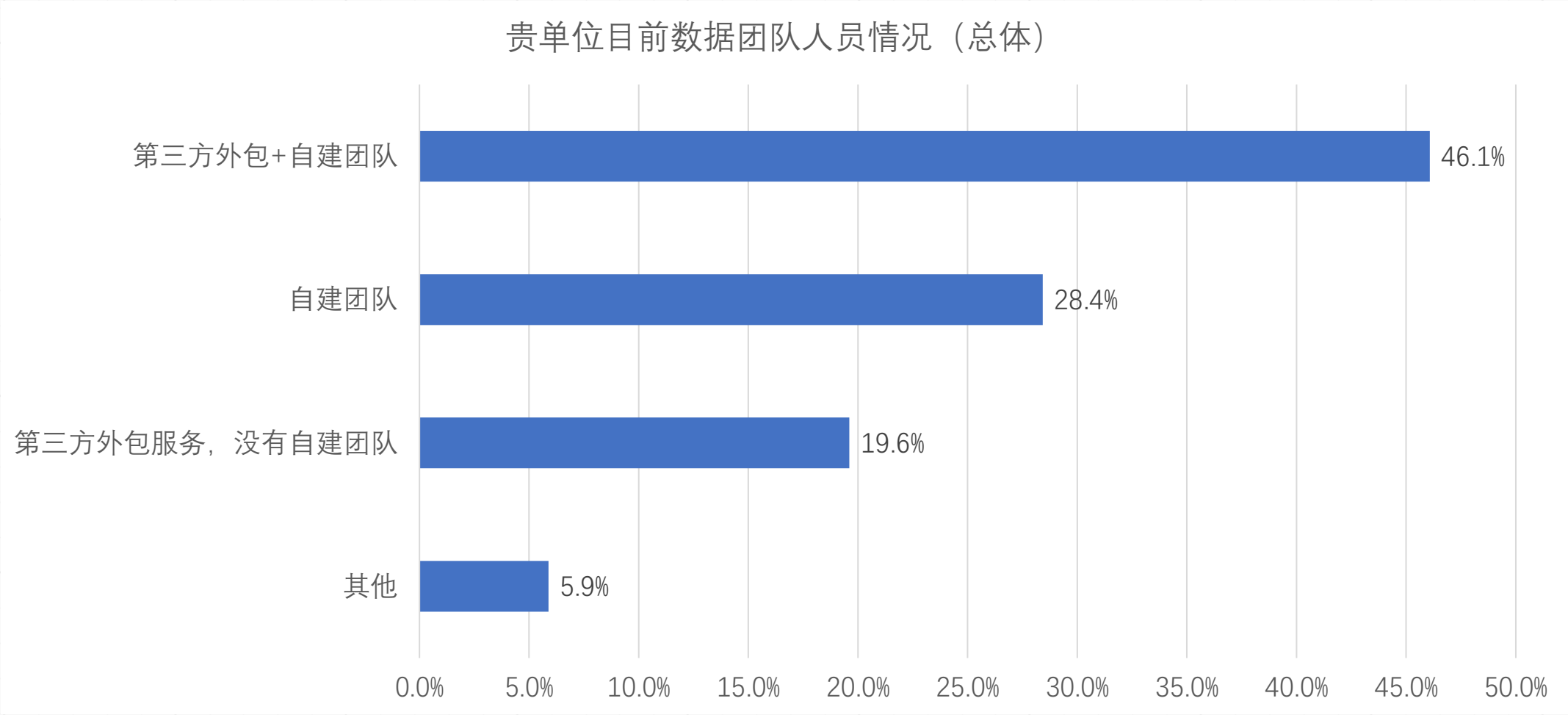
各行业业务不同，数据应用遇到的问题也不同

94.4%的互联网企业最常遇到的问题是数据分析人才缺乏；房地产最常遇到的问题是运维难度高；医疗最常遇到的问题是无法利用不同数据库数据进行联合分析；金融最常遇到的问题是数据应用领域相对较窄，数据与场景融合不够；政府最常遇到的问题是缺乏面向业务人员的自主分析工具。



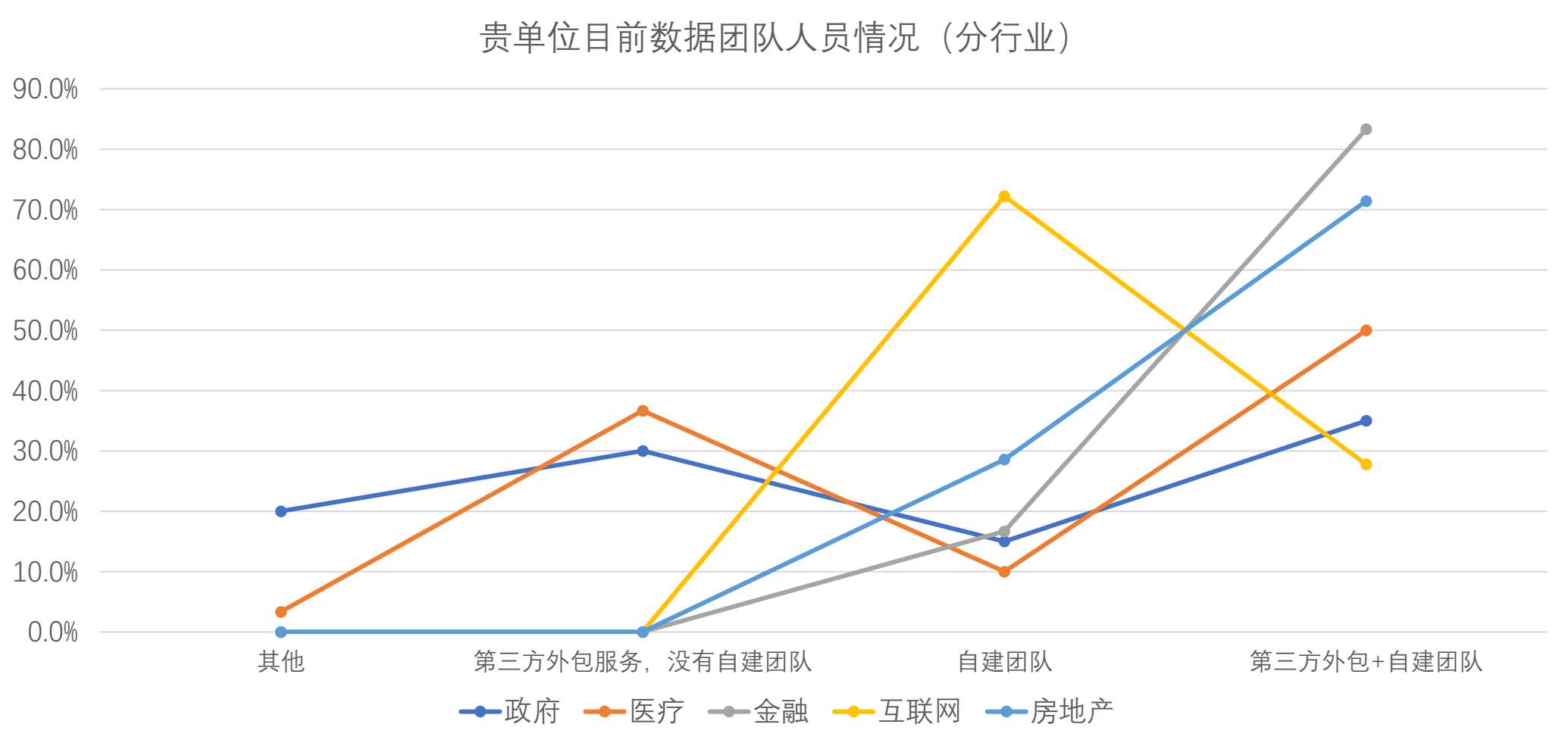
企业大数据团队构成的主要方式是外包+自建

企业的大数据团队由第三方外包+自建团队的比率为46.1%；完全为自建团队的比率为28.4%；完全为第三方外包服务，没有自建团队的比率为19.6%。



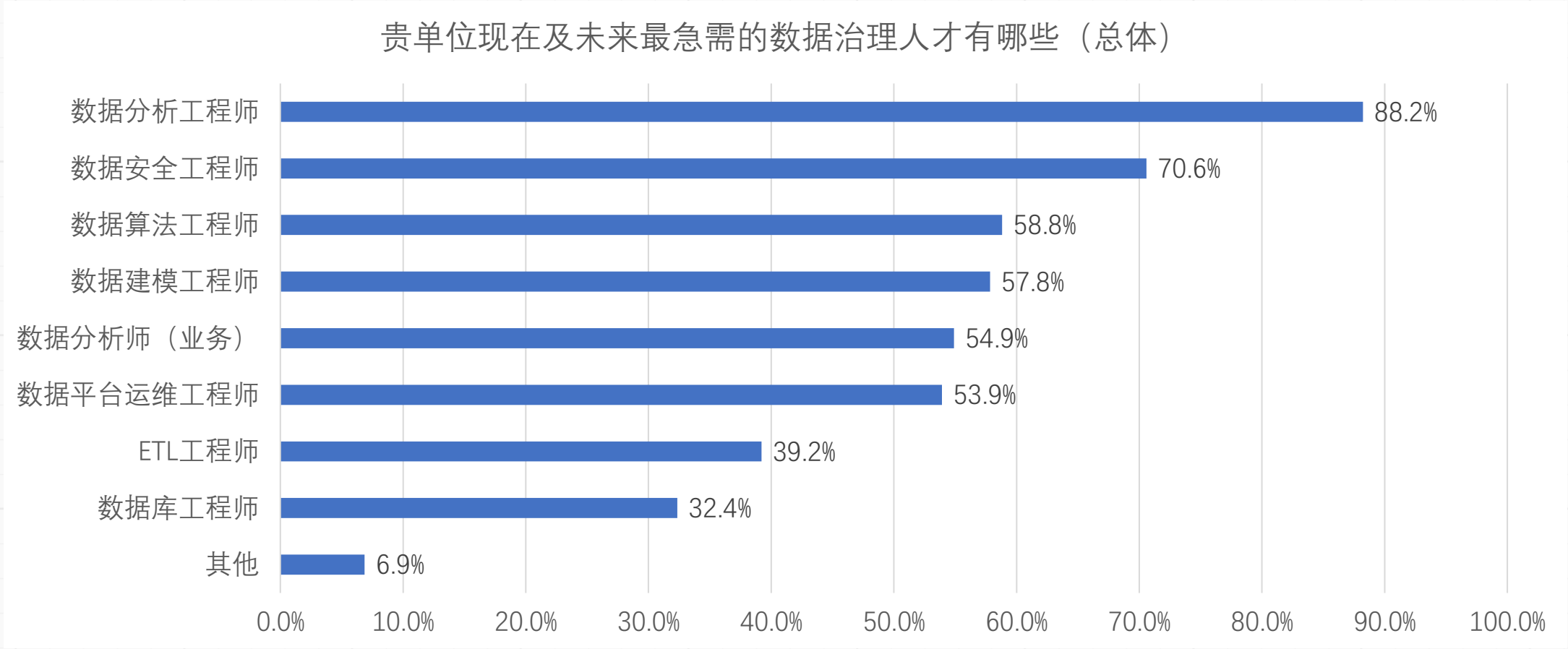
互联网更倾向于自建，其他行业更倾向于购买服务

72.2%的互联网企业大数据团队为自建团队，政府机构和医疗单位以及金融、房地产企业的数据团队主要为第三方外包+自建团队。



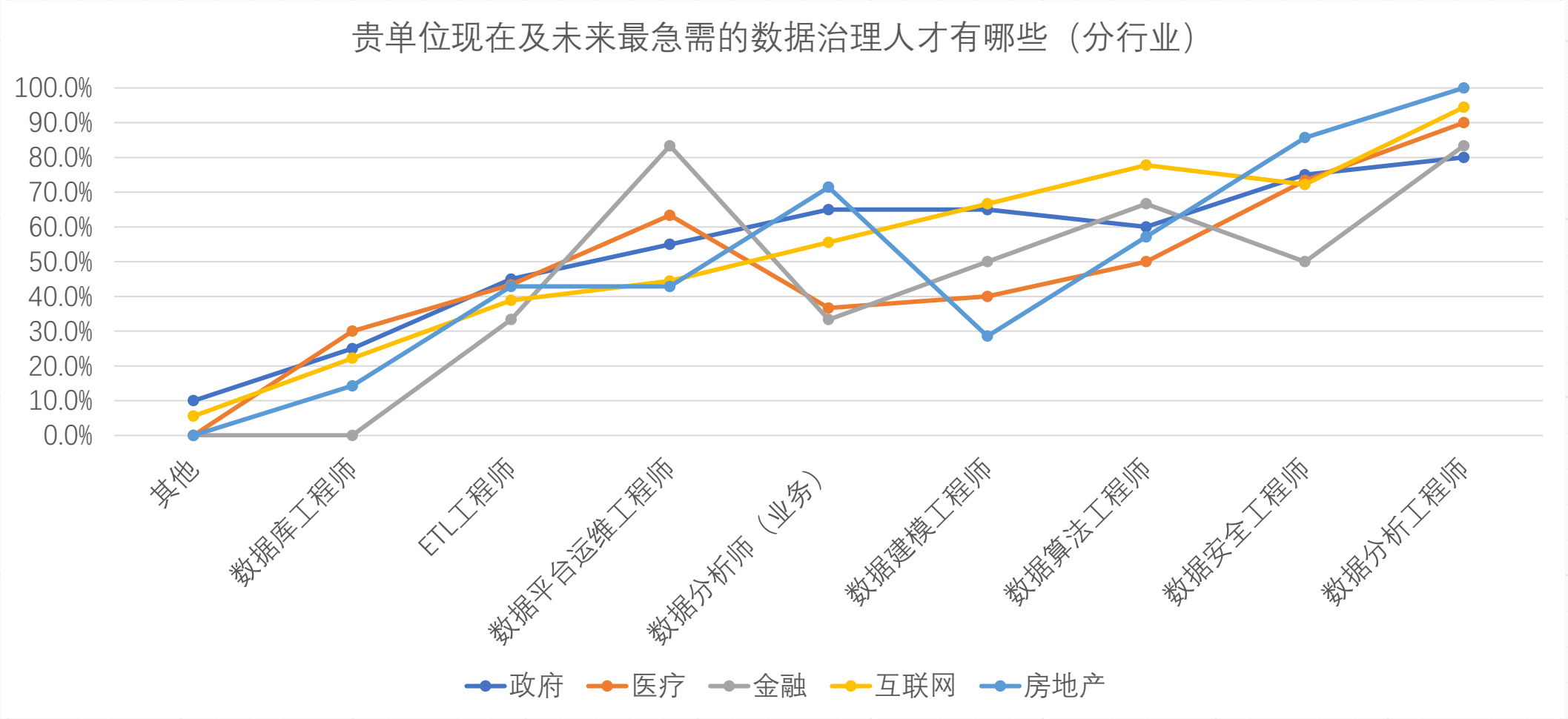
未来急需数据分析工程师和数据安全工程师

70%以上的企业缺乏数据分析工程师、数据安全工程师；同时，数据算法工程师、数据建模工程师、业务方向的数据分析工程师、数据平台运维工程师的比率分别为58.8%、57.8%、54.9%和53.9%。



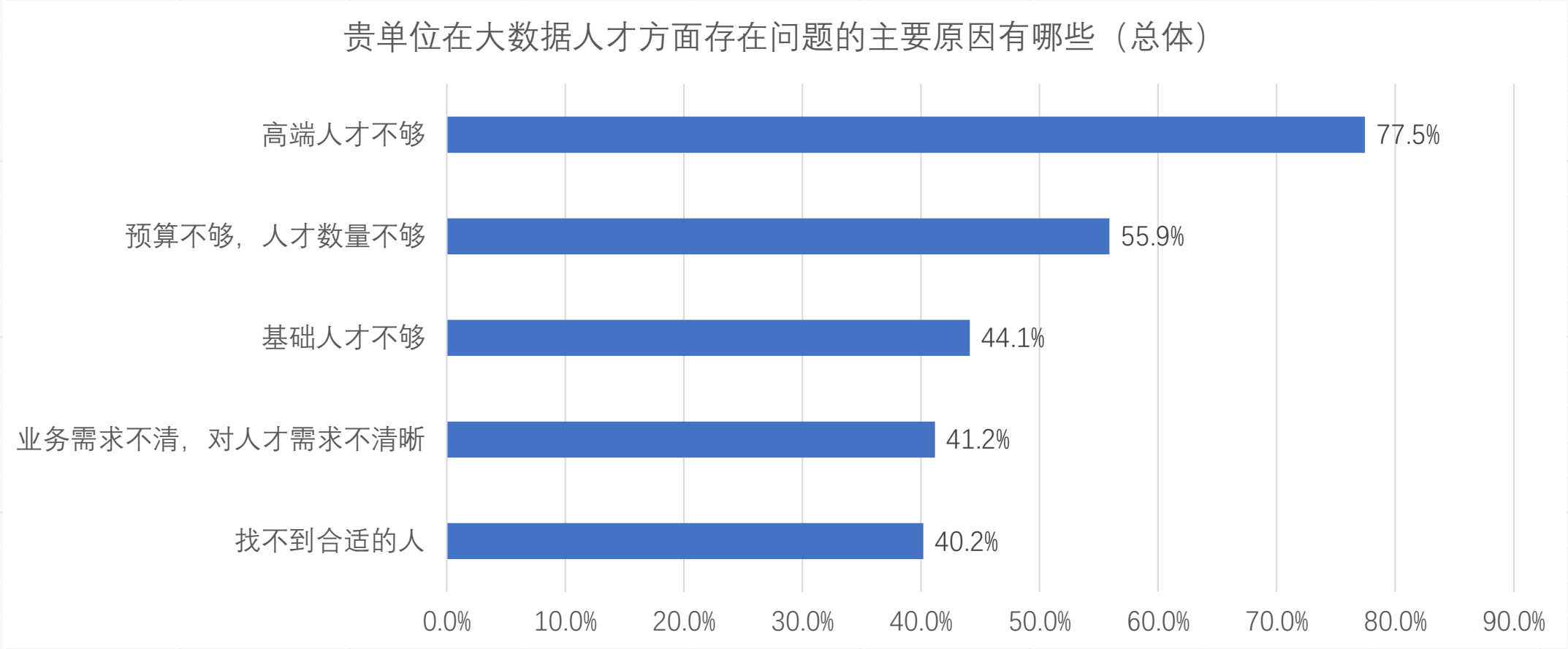
数据分析师不足在各行业都比较突出

房地产、互联网、医疗、政府行业未来最急需的数据治理人才是数据分析工程师；金融行业未来最急需的数据治理人才为数据平台运维工程师和数据分析工程师。



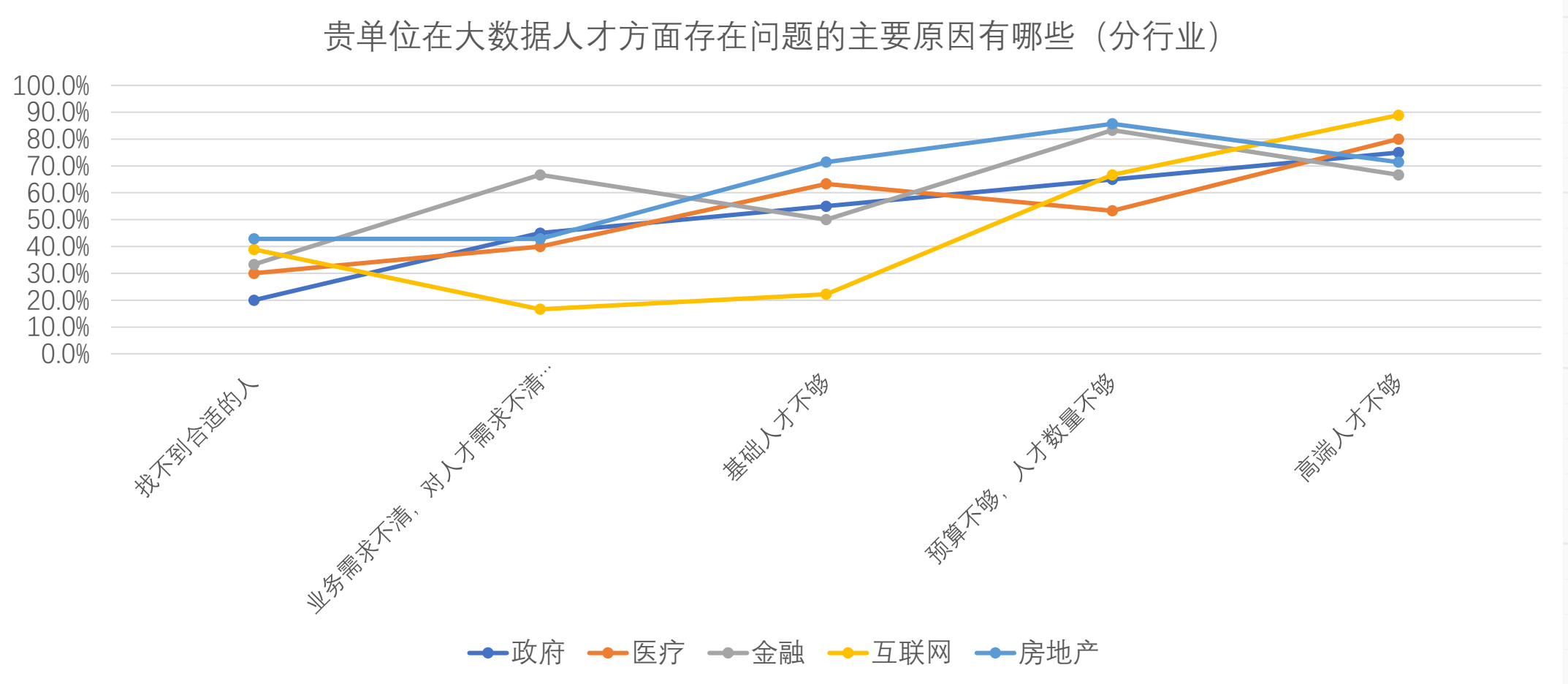
大数据高端人才的缺乏是各行业普遍面临的困境

77.5%的机构在大数据人才方面存在的主要问题是缺乏数据治理高端人才；55.9%的主要原因是预算不足、人才数量不够；44.1%的主要原因是基础人才不够；41.2%的主要原因是业务需求不清，对人才需求不清晰；40.2%是因为找不到合适的人。

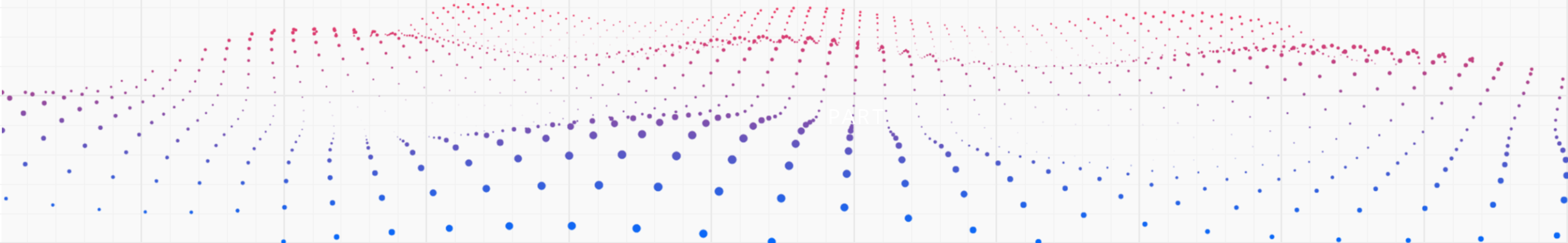


金融、房地产、互联网行业对人才需求的突出矛盾存在差异

金融行业面临的主要人才矛盾是业务需求不清，对人才需求不清晰；房地产行业面临的主要人才矛盾是预算不够，人才数量不够；互联网行业面临的主要人才矛盾是高端人才不够。



CONTENT



以建设“汇管用评”闭环，推进数据治理体系建设

1. **汇**——按照“需求牵引，问题导向；统筹规划，全面汇聚”的原则汇聚数据。
2. **管**——按照“进出有序，全程留痕，分级控制，成果管控”的原则对数据进行全面管控。
3. **用**——为数据用户提供“场景化，智能化，自主化”的数据服务。
4. **评**——以全生命周期的数据质量评估和数据应用服务评估，促进数据价值挖掘。

以数据应用为导向，以数据实时质量为抓手进行数据治理系统建设

1. 在业务中找数据应用场景，在场景中明确数据治理的目标和标准，才能让数据治理产生真正的价值。

2. 通过建立各环节的实时数据质量监测手段，实现可追溯的数据质量管理，才能实现数据问题精准、快速定位，才能保障数据的有用性、可用性、及时性。

以“可用不可见”让数据应用与数据安全从对立走向融合

1. 隐私计算技术已经得到快速发展，已经可以支撑“数据可用不可见”的应用方式，解决数据应用过程中的数据安全和隐私保护问题。

2. 通过在数据应用端增加用户行为模型，通过用户行为模型实现数据及成果的精准供给，降低数据泄露等安全风险。

数据治理人才队伍需要招聘和自己培养两手抓，以自己培养为主

数据治理人才的培养，需要放到数字化转型的大背景下考量。立足短期用人需求，着眼长期人才规划，建立面向中高层、面向业务岗位、业务IT部门等不同层级、不同职能的人才培训体系。从大数据素养、大数据管理、大数据实操技能等角度，根据课程要求和掌握程度要求采用线上、线下或者线上线下结合的方式开展人才的培训。

编写团队

专家顾问团

赖茂生　北京大学信息管理系教授 博士生导师

徐 斌　旭辉集团副总裁兼首席数字官 企业管理中心总经理

颜廷方　山东五征集团CIO&CDO

田宗梅　首都医科大学附属北京世纪坛医院信息中心主任 高级工程师

李郁鸿　郑大一附院信息处长、教授级高工 中国医院协会信息专业委员会常委

曾啟焯　广州功夫投资控股集团IT总监

王彦博　龙盈智达（北京）科技有限公司首席数据科学家

乐勇斌　TATA木门CIO

杨振良　北京影合众新媒体技术服务有限公司运维总监

赖安徽　大家信科基础设施负责人

鲁四海　万山数据创始人

编写团队

报告编写委员会

主编：

姚乐　CIO时代创始人兼研究院院长

副主编：

刘胜文　CIO时代研究院院长助理　点用实训联合创始人

王雪娜　CIO时代新基建创新研究院执行院长

编组成员：

李遗　李赛　刘祺　刘松　杨启洁　王菁　徐石磊　张少东

编写单位：CIO时代

技术支持：北京万山数据科技有限公司

报告编制说明

如对报告有任何意见和建议，请与我们联系。

联系人：刘胜文　杨启洁

联系方式：13810528344　18885401560