数据流动交易模式分类与规模估算初探

蔡跃洲[*1]　刘悦欣[2]

[1]中国社会科学院数量经济与技术经济研究所

[2]中国社会科学院大学商学院

摘　要：社会各界高度关注数据要素及其流动交易，但过于聚焦个人行为数据和数据交易所模式。为更为全面地反映数据流动交易状况，在对不同视角数据要素分类归纳的基础上，按照数据涉及行为主体，从个人数据、企业数据、公共数据三个方面对数据流动交易模式进行系统梳理，并综合不同来源的资料，对全球主要经济体当下数据资源和数据市场交易规模及其趋势特征进行了估算分析。研究结果表明：（1）丰富的现实场景决定了数据分类的多样性，而不同类型数据的涉及主体、权属划分、信息密度等特征存在较大差异，由此带来流动交易模式及收益分配的复杂性。（2）全球数据流动交易呈现产品化、服务化和平台化趋势特征。（3）全球主要经济体都存在数据生成规模与数据存储规模失衡，中国的情况尤为突出。（4）全球数据交易市场处于初期快速发展阶段，成交规模数量级大致为千亿美元，中国的数据要素市场发育更不充分，规模仅为全球的10%左右。社会各界应充分意识到数据要素分类和数据流动交易的多样性、复杂性以及相关制度体系构建的艰巨性、长期性，并顺应这些特点着力健全数据分级分类、加强存力算力基础设施建设、培育数据交易开发专业机构、完善数据治理体系。

关键词：数据要素；数据分类；流动交易模式；生成存储规模；交易成交规模

JEL 分类号：D83, O25

一、引言

随着移动互联网、3G/4G/5G通信、云计算、人工智能等新一代信息技术的大规模商业化应用，数据的生产、收集、传输、处理、分析发生了全方位革命性变化，产生并积累了大量数据资源。丰富的数据资源为更多数据分析应用场景提供了基础，催生出以平台经济（共享经济）为代表的各种新业态、新模式，成为重构经济社会生产生活组织模式的关键要素。2019年，党的十九届四中全会通过的《中共中央关于坚持和完善中国特色社会主义制度、推进国家治理体系和治理能力现代化若干重大问题的决定》提出“健全劳动、资本、土地、知识、技术、管理、数据等生产要素由市场评价贡献、按贡献决定报酬的机制”，明确将数据列为第七大生产要素。这表明，数字经济时代数据在微观生产运

* 联系人：蔡跃洲，邮箱：caiyuezhou@cass.org.cn。

基金项目：国家自然科学基金面上项目“新一代信息技术影响增长动力及产业结构的理论与经验研究”（71873144）；国家自然科学基金重大项目“宏观大数据建模和预测研究”（71991475）。

营、宏观经济增长及发展中所发挥的作用得到广泛认可。

数字经济时代的比特数据具有非竞争性、非排他性、低成本复制、网络外部性、即时性等技术-经济特征，能够提高微观运行效率，并在宏观层面实现价值创造的倍增效应（蔡跃洲、马文君，2021）。而发挥数据要素效率提升和价值倍增作用的一个重要前提在于实现其安全、有序、充分流动。为此，宏观调控部门针对数据流动交易和要素市场建设，先后出台多个文件予以指导。2021年12月，国务院办公厅发布《要素市场化配置综合改革试点总体方案》强调，完善公共数据开放共享机制、建立健全数据流通交易规则、拓展规范化数据开发利用场景、加强数据安全保护，探索建立数据要素流通规则，并提出“原始数据不出域、数据可用不可见”的交易范式和数据使用“可控可计量”原则。2022年6月22日，中央全面深化改革委员会通过《关于构建数据基础制度更好发挥数据要素作用的意见》，明确指出要加快构建数据基础制度体系，推进数据产权、流通交易、收益分配和安全治理，以促进数据流通使用，赋能实体经济。2016年以来先后通过的《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》则着重从数据安全角度对数据流动活动进行规范。

自2015年贵阳大数据交易所成立以来，地方政府纷纷着手推动数据交易所（中心）建设，作为规范数据流动交易的具体举措。目前，包括北京、上海、深圳、广州、福建、郑州等地都陆续成立了数据交易所（中心）。政府推动加上媒体报道和渲染，使得当下社会公众乃至学术界的注意力更多集中在个人行为数据和数据交易所模式上，相关的学术探讨和制度建设也随之倾斜。

然而，数字经济实践中所产生的数据类型众多，其流动交易模式也存在很大差别。基于消费互联网产生的个人行为数据仅仅是全社会海量数据资源的一部分，而数据交易所撮合交易也只是数据交易的一种具体实现方式，更多交易可以由供需双方通过点对点以数据产品/服务方式来达成。建立健全数据相关基础制度体系，切实发挥数据作为关键生产要素在效率提升、价值倍增中的支撑作用，有必要在数据内涵辨析和分类基础上，全面梳理数据流动交易模式、掌握数据资源规模结构等情况，归纳其特性趋势。为此，本文后续各部分将从数据相关基础概念辨析出发，系统梳理并分析不同类型数据流动交易模式，充分展示当下不同类型数据实际交易模式全貌，总结提炼特征及趋势，以期纠正认识上偏颇乃至误区；从存量数据资源和数据流动交易量两方面对数据生成、存储规模和特定数据交易方式下的成交金额进行展示及进一步估算，为考察数据资源和数据要素市场结构性特征提供更具体的量化支撑；从政策制定角度给出相应建议，以期为完善数据要素市场体制机制、促进数据流动交易共享提供基础性的参考。

二、数据相关概念辨析与数据分类

丰裕的数据资源既是数字技术大规模商业化应用的产物，也是支撑数字经济发展的新关键要

素。数字经济丰富的实践场景使得数据呈现出多种不同分类方式。有必要在辨析“数据”内涵和外延基础上，对数据分类进行系统性梳理，作为分析数据流动交易模式、开展数据规模估算的基础。

（一）比特数据、数据资源与数据要素

从物理和技术角度来看，数字时代的“数据”被广泛地界定为以二进制进行编码、作为信息载体的字符串，即比特数据（蔡跃洲、马文君，2021；Farboodi and Veldkamp, 2021）。比特数据的产生是基于数字技术对经济社会现实运行状况的观察和记录，因此，某种意义上可以看作是经济活动的副产品（Veldkamp and Chung, 2019）。OECD（2021）则提出“数据是通过访问和观察现象而产生，以数字格式记录、组织、存储、处理或进行访问的信息内容”，将数据看作是一种特殊的信息表现形式，即以二进制比特形式呈现的信息。加拿大统计局（Statistics Canada, 2019）出于测算的需要，将“数据”定义为已经被转化为数字/数码形式的观察结果，可以存储、传输或处理并从中获取知识；该定义旨在将“数据”的范围限定为某个时点对特定事物的观察，并且作为数字化记录能够被存储、检索、分析和调查，从而将数字化音乐、影视资料等排除在外。事实上，“数据”历来都被看作是一种信息（或事实），而数字时代“数据”与信息的关联更为密切，很多语境下等同于“信息”（蔡跃洲、马文君，2021）。

当然，以二进制字符串形式存在的比特数据与有效信息之间还存在差别，需要借助数据分析技术才能提炼出有效信息。经初次采集、以比特形式存储的海量原始数据，在未经处理、分析、提炼其内在有效信息之前无法直接应用于生产消费场景；因此，原始数据还不是能直接参与价值创造的生产要素/数据要素，而只是具备价值创造潜力或者说潜在价值的“数据资源”（Varian, 2018; UNCTAD, 2021）。原始数据经过清洗、聚合、处理、分析所形成数据集、数据库、信息报告、数据服务等不同形式的数据产品/数据服务，则可以根据现实需要应用于经济社会运行的不同场景，如市场营销、风险控制或是人员搜索（FTC，2014）。应用于不同场景的数据产品/数据服务，发挥了生产要素的作用，直接参与生产经营中的价值创造，是数据要素的具体呈现形式。

数据资源所具有的非竞争性、部分排他性、低成本复制等技术-经济特征，使得数据在理论上可以被大规模重复利用，从而缓解其他有形资本稀缺性带来的增长限制，实现宏观倍增效应（蔡跃洲、马文君，2021）。然而，企业生产经营实践中，需要投入大量的人力、物力对原始数据清洗、处理、分析，才能从中提取有效信息并将其切实转化成数据要素；这种转化能力本身对于企业来说也是非常稀缺的（OECD, 2013a）。当然，由于原始数据是有效信息的源头，而且有的原始数据也能直接提供有效信息，因此，包括原始数据、加工数据、数据产品、数据服务在内的各种数据资源，都可以算作是广义的数据要素。

（二）不同视角的数据分类与数据类型

数字技术的广泛渗透性和经济社会活动的复杂性决定了数据作为经济社会活动的数字化记

录必然呈现出多样性特征。为了促进数据资源/数据要素的流动交易，更好地发挥其对经济社会运行活动的支撑作用，有必要从不同视角对数据分类进行梳理。

数据作为经济社会活动的数字化记录，可以从记录对象的特征/类型入手梳理数据的分类。由于经济社会活动的复杂性，围绕被记录对象分类的维度可以有很多种。比较常见的方式是根据数据生成和应用场景所属领域进行分类，例如，比照既有的国民经济行业分类，将数据划分为通信大数据、金融大数据、医疗大数据、农业大数据、交通大数据、电力大数据等行业数据。也可以从经济社会活动角度，按照数据记录对应的每个具体环节随其进行分类，例如，2020年工业和信息化部办公厅引发的《工业数据分类分级指南（试行）》，将工业企业生产经营各环节进行细分，各环节对应的数据记录则划分为研发数据、生产数据、运维数据、管理数据和外部数据等。

更为常见的是根据数据记录对象涉及的行为主体进行分类，具体可分为“个人行为数据”“企业生产运营数据”和“政府公共部门数据”三大类。个人行为数据主要是各种互联网平台上，用户进行浏览、搜索、互动、交易等活动，被平台实时记录而形成的数据，如淘宝、京东的购物记录，微信、微博等社交媒体的互动交流内容等。企业生产运营数据是企业在生产经营过程中，对各环节、各流程状况进行记录、监测而形成的数据，如制造企业通过传感器实时监测反馈智能化生产线运行状况而收集沉淀的数据。政府公共部门数据则是各级政府部门、具有管理公共事务职能的组织和电力、公交、燃气、热力、给排水等公共服务运营单位，在提供公共基础设施、公共服务等过程中收集、产生沉淀形成的所有数据资源，如税务、海关等部门数据、企业经营相关资质和信用信息、自然资源数据、交通路况信息、电力供需调度数据、市政道路管网分布及运行状态数据等等。UNCTAD（2021）类似分类则将数据区分为消费者数据、商业数据、政府及公开数据。与行为主体分类密切关联的是数据收集、维护的资金来源以及数据权属等问题。按照数据创建、维护和持有所需资金来源，可将其分为私人部门数据和公共部门数据；而根据所有权和使用权等法律权属，则可划分为公共数据和专有数据，其中专有数据特指所有权明确，受知识产权或其他类似法律效力保护的数据（Swedish National Board of Trade, 2015; Nguyen and Paczos, 2020）。

除了记录对象外，还可以着眼于数据自身特征进行分类。从数据所提供信息含量特征来看，可以将数据划分为原始数据、加工数据、数据产品/数据服务以及元数据。而从数据流动范围来看，则可以分为境内流动数据和跨境流动数据。需要特别强调的是上述数据分类方式并不互斥，在不同分类标准下，同一数据（集合）可以同时归属于多个类型。

三、数据流动交易模式

充分流动是数据发挥其关键要素作用，提高运行效率、实现价值倍增效应的重要前提。数据流动的方式与数据类型密切相关。涉及的行为主体和权益归属不同，所包含的有效信息（及信息强度）

的差异，都会影响数据流动的方式。在数字经济运行实践中，数据流动交易模式呈现出复杂且多样的特征，而数据交易所仅仅是一种高社会关注度的模式。高效规范数据流动交易活动需要全面系统把握其流动模式、特征及趋势。为此，我们将依据上述行为主体分类，对个人数据、企业数据和公共数据三类数据的流动交易模式分别进行梳理和分析。需要特别指出的是，数据流动包括数据交易，后者是非公共部门数据流动所采取的方式，而公共数据通常是通过开放共享方式实现流动。

（一）个人行为数据交易模式

个人行为数据主要是指与个人消费行为相关活动的数据记录，即消费行为数据。对个人消费行为的记录在互联网出现之前已经存在，但大多散落于不同商家手中。往往需要第三方数据中介多方收集才可能形成相对完整的个人行为记录，并由此诞生了数据产业和“数据经纪人”等概念雏形。20世纪90年代中后期，个人电脑的大范围推广和互联网热潮的兴起，催生出以电子商务为代表的消费互联网模式，个人行为数据开始被（平台）系统性地收集，个人行为数据及其流动交易也开始进入学界研究的视野（Laudon,1996）。经过20多年的持续高速发展，众多细分领域的消费互联网平台，在彻底改变了人们生活消费模式的同时，也为个人数据的收集和生成创造了良好条件，并成为个人行为数据最主要的收集者。

从行为主体/数据当事人的角度来看，互联网平台对个人行为数据的收集和生成大致可以分为三种情形：（1）个人主动或接受条款进行分享的那些涉及其自身及第三方的数据，如产生的社交网络画像、在线购物刷卡记录等；（2）在记录用户活动过程中，无需用户授权即可合法观察和捕捉的数据，如网页浏览数据、移动电话使用时的定位数据等；（3）基于个人数据分析推断得到的衍生数据[①]，个别情况下个人数据也可以从几条看似“匿名”的数据中推断出来。而从渠道和来源来看，消费互联网平台收集的个人行为数据又可以分为两大类：一类是指在线平台基于自己的（数字）产品和服务而直接采集的一手数据；另一类是用户在平台之外的活动并由其他第三方记录收集的数据（OECD，2013b）。数据当事人与数据收集者的分离使得个人行为数据的交易更为复杂，至少涉及三方主体和两重交易。

对个人行为数据的记录和收集可以看作是第一重交易。交易双方分别是行为主体（平台消费者用户）和数据收集者（互联网平台），数据收集者作为买方免费或者以提供特定应用服务（免费使用App）作为对价获得收集记录用户个人（消费）行为的权利，并由此沉淀积累相应的原始数据资源。由于消费行为记录包含了消费者个人的隐私信息，这一重交易的本质也可以看作是消费者用户以其个人隐私为代价换取互联网平台的免费服务；这也使得有关隐私定价的研究通常都与消费者个人行为数据联系在一起（FTC, 2014; Acquisti et al., 2016）。另外，第一重（数据）交易是消费者

[①] 例如，可以根据与个人财务历史相关的许多因素计算信用分数。

用户与互联网平台之间“数据”和“数字化服务”的互换，属于“复合交易(composite transaction)”[②](Malgieri and Custers, 2018)。

数据收集者/互联网平台与其他第三方之间的数据交易是第二重交易。在此过程中，数据收集者通常既扮演买方角色从第三方获得额外的数据来丰富既有的个人行为数据，以实现产品创新、优化供应链结构和提高营销效率，据此提升平台对外的服务质量；也可以扮演卖方角色，向第三方提供原始数据或数据产品/数据服务，如营销服务、信用评价等，通过更广泛的数据流动和应用充分挖掘数据的潜在价值。而第三方既可以是其他数据平台，也可以是数据集成商，还可能是其他具有特定数据需求的微观经营主体或公共部门。

在第二重交易中，根据被交易数据的不同形态又分为直接交易和间接交易(田杰棠、刘露瑶，2020；Bergemann and Bonatti, 2019；Veldkamp and Chung, 2019)。在直接交易模式下，售卖方(供给方)仅对原始数据进行初步加工，而不是深入挖掘潜在信息，产品主要以数据集的形式对外出售。而在间接交易模式下，售卖方对原始数据进行清洗、整理、分析挖掘后，以定制数据产品或数据服务的形式对外销售。间接交易模式基本不涉及原始数据交换。此外，隐私计算(包括联邦学习、安全多方计算等技术)和区块链技术使得多方数据联合分析能够在数据不泄露的前提下进行，在技术上实现了数据可用不可见，其在多场景下的应用能提升对个人隐私信息的保护，因此成为间接交易模式的有益补充。

(二)企业数据交易共享模式

具有一定数字化基础的企业，在日常生产经营过程中能够产生并累积大量监测记录数据，主要包括基于传感器、物联网捕捉记录的机器设备运行状况数据和基于企业内部IT业务系统生成的生产、销售、物流等信息数据。同个人行为数据相比，企业数据的权属关系相对简单。大多数情况下，企业既是数据行为主体也是数据收集者，因此数据权属一般不存在争议。

从理论上讲，企业数据的流动和重复使用，不仅有利于增强产业链上下游的协同性，带来企业效益的提升，并在更大范围内提高社会整体福利。但实践中企业之间的数据交易共享并不一定会自发形成。事实上，企业出售或共享数据的意愿还要受到企业之间竞争行为的影响。只有当数据重复利用场景和数据来源企业的原始使用场景相互独立或形成互补时，企业才有意愿共享数据。具体来说，企业作为数据收集者和所有者出售或共享其数据资源/数据产品的动机大致可分为三种：(1)直接出售数据产品或数据服务以获得营业收入；(2)提高同关联企业之间的协同，优化产业链供应链，开发产品、改善服务、创新商业模式；(3)实现更高效的供需匹配。另外，从数据需求角度来看，购买方获取企业数据的主要目的包括开发新产品或改进既有产品、提高企业生产效率、改善客户关系、

[②] 法律概念，指单一合同中含有两个法律关系。

优化企业内部管理结构、精确定位市场目标等(European Commission et al., 2018)。

数字经济实践中,现有的企业数据交易共享模式主要有以下几类:(1)数据企业主导的直接交易模式,即数据所属企业直接向数据购买方出售相关数据产品或提供数据服务,并从中获得一定收益;交易实施的具体方式通常是由数据所属企业向购买方提供数据接口并授权访问。(2)基于数据平台的中介交易模式,即数据平台将数据供给方和数据需求方聚集在一起,由其充当可信第三方中介机构对接数据供需双方,促成交易并从每笔达成的数据交易中获得佣金收入。从交易实践来看,充当第三方中介的机构,既可以是诸如亚马逊AWS这样的综合性云服务平台(AWS Data Exchange数据交易平台),也可以是Dawex、大数据交易所等专业数据交易平台,还可以是邓白氏(Dun & Bradstreet)这样的专业数据集成商。(3)基于产业互联网的数据共享模式,即接入产业互联网平台的上下游企业,在相对安全的环境中自愿接入平台,并在一定范围内交换共享相关数据,以促进新产品开发或运营效率提升。接入企业通常会向产业互联网平台(运营方)免费开放共享数据,同时也从平台获得由其提供的服务或积累的其他数据。例如,空客公司于2017年建立了一家航空领域的产业互联网平台“Skywise(智慧天空)”,对用户(航空公司)提供的数据进行集成并匿名化后,免费向各航空公司提供基于其汇集数据形成的全球机队基准报告,并为航空公司和制造商提供数据接口获取上述集成的匿名化航空数据,以助力航空公司和制造商提高运营效率、改进飞机的软硬件设备。需要指出的是,尽管基于产业互联网的数据共享以业务创新和效率提升为主要目标因而更多采取免费方式,但它本质上还是一种复合交易;数据企业向产业互联网平台免费提供数据,作为获得平台免费服务的对价。另外,平台基于集成数据提供更高级数据产品和数据服务时,往往采用收费方式。

(三)公共数据流动模式

公共数据在时间、空间和样本覆盖范围方面具有显著的完备性,而其对政府公共部门和公用事业运行状况的记录,提供了个人、企业等微观主体相关活动的外部环境信息,与私人部门数据形成互补,成为数据挖掘和大数据分析不可或缺的数据资源,社会各界对于获取公共数据有着强烈的需求。实践中,公共数据流动主要采用数据开放的方式,通常由掌握公共数据资源的政府部门或公用事业机构,在充分评估数据安全等因素前提下,有选择地向公众开放数据(蔡跃洲、马文君,2021)。

从数据流动共享范围来看,公共数据流动又可以细分为公共部门之间的数据交换和面向社会各界的数据开放。一方面,构建涵盖不同政府部门和公共事业机构的统一数据平台,将分散于各部门、各机构的零散数据通过跨部门数据交换予以集成,能够形成更为统一完备的信息集,进而为政府部门迅速发现经济社会运行中存在的问题、及时采取应对措施政策、开展动态监测和事后评估提

供全面系统的一手信息。另一方面，公共部门积累的海量社会运行状况数据，涉及实时道路交通、水文气象、生态环境、公共安全等与生产经营活动密切相关的数据信息。如果能在安全有序前提下实现其中部分数据的共享和流动，将极大提升私人部门运营效率、促进私人部门开展创新活动。

开放政府数据历来是各国政府加快数字化转型的重要内容之一。早在2009年，美国一些地方政府便率先建立统一公共数据开放平台，英国、法国、加拿大、新加坡等国家紧随其后。2015年10月，“十三五”规划建议中明确提出实施“国家大数据战略”后，中国也自上而下着手推进公共数据平台建设和数据开放相关工作。在开放级别上，根据涉及内容的敏感性和安全性可以将公共数据划分为三类：无条件共享数据、受限共享数据和不共享数据。无条件开放的公共数据，通常可以从统一的公共数据开放通道（如公共信息门户平台）中获取，而受限开放公共数据的获取则需要向数据主管部门提出申请，经审核并签订协议和安全承诺。

公共数据对社会各界的开放原则上讲是免费的。不过，考虑到数据的收集、存储、加工整理，需要由数据科学专业技术人员完成，涉及大量人力和财力投入，仅靠政府部门或公共事业机构自身难以完成。为此，在国内外实践中，公共部门往往会引入商业机构参与公共数据的加工整理，涉及的部分数据服务也会采用收费方式。例如，英国将公共部门信息的持有者（Public Sector Information Holders, PSIHs）划分为三类：（1）免费提供未经信息精炼的PSI机构；（2）使用PSI数据改进或支持内部活动的机构；（3）具有商业激励的PSI创收机构。其中，第（3）类机构在对第三方提供基于原始公共数据形成的数据产品或数据服务时往往会采取收费模式。

目前，中国各地公共部门也在积极探索公共数据授权运营方式，以充分挖掘公共数据的潜在价值。如《上海市数据条例》提出，促进公共数据社会化开发利用。《浙江省公共数据条例》提出，县级以上人民政府可以授权符合相关规定的法人组织或非法人组织运营公共数据，授权运营单位应当依托公共数据平台对授权运营的公共数据进行加工，对加工形成的数据产品和服务，可以向用户提供并获取合理收益。上述授权运营在性质上类似于公共服务的特许经营，其收费标准应以弥补公共服务提供必要成本为限（常江和张震，2022）。

（四）数据流动交易模式特点

从前述个人数据、企业数据和公共数据三类不同行为主体数据的流动交易模式梳理可以看出，各类模式都是基于数据流动交易现实需要而自发形成的，总体呈现以下特点。

一是流动交易模式多样化，涉及主体关系复杂化。现实中数据需求和应用场景千差万别，而不同类型数据的记录对象、涉及主体和权属关系也各不相同；这些因素相互交织，使得数据流动交易的实现方式和供需双方及第三方的对价支付、收益分配呈现出纷繁复杂的局面。另外，数据流动交易不可避免会衍生出个人隐私和信息安全等问题，如何确定交易共享的范围、程度，需要在效率和

安全之间不断权衡。

二是交易(流动)对象/标的呈产品化、服务化趋势。从流动交易客体来看,尽管以原始数据为标的的直接销售模式和以数据产品/数据服务为标的的间接销售形式都占据一定比例,但后者有望成为未来数据交易市场的主要形式。毕竟基于买方需求提供经过加工的数据产品/数据服务,不会涉及原始数据复制转移,从而规避了由此带来的权属纠纷,同时也有利于保护个人隐私安全和国家信息安全。

三是平台或中介在数据流动交易中越来越发挥主导作用。尽管也有数据供需双方点对点直接交易的情形,但越来越多的数据流动特别是数据交易都是基于平台多边市场,由数据中介主导完成的。数据中介既可以是平台企业,也可以是数据集成商,或者数据经纪人:(1)一种情况下,数据经纪人通常只是撮合交易,买卖双方交易成功后从成交额中提取一定比例佣金;这种模式可替代性较高,数据中介/经纪人的作用有限,现实中交易双方甚至可以避开数据经纪人达成协议。(2)数据集成商不仅对外提供数据(产品/服务),自身也是数据的需求方。例如在消费金融领域,通常是银行和金融机构收集消费者数据(如个人账户信用历史等),这些公司将数据提供给征信机构的同时,也向征信机构购买关于自身客户和潜在新客户的额外信息(Bergemann and Bonatti, 2019)。数据集成商由于自身业务的特性汇集了大量数据,在数据质量、内容或范围或是核心算法技术等方面具有一定的竞争优势,一般很难被取代。(3)作为数据中介的平台企业,不仅仅提供撮合交易平台,同时在自身平台上提供数据交易所需的其他增值服务,例如,AWS数据交易平台和上海大数据交易平台,都可以提供数据交易的一站式服务,并充分保障数据安全。

四、数据资源及其流动交易规模估算

数据要素分类和流动交易模式的复杂多样性决定了数据要素涵盖范围可以有多种不同口径。客观认识数据资源和数据要素市场运行的特征规律需要科学准确估算数据要素规模,这就必须界定清楚估算对象所对应的口径标准。根据前述数据要素内涵分类和流动交易模式的梳理,估算数据要素规模,既可以围绕数据要素(资源)的规模进行,也可以针对数据流动交易的规模进行。前者从生产和存储角度反映数据要素规模,而后者则可以从流量角度间接反映数据要素使用的情况。由于数据类型和流动交易模式都较为繁杂,而现行国民经济统计体系也无法提供直接对应的数据,为此本部分将从生产存储和流动交易两个方面,综合多渠道数据信息,尝试对数据要素不同口径的规模进行估算,据此对数据要素规模呈现的结构性特征进行分析归纳。

(一)数据资源规模估算

伴随新一代信息技术的大规模商业化应用,数据生产(收集)成本大幅降低,边际成本接

近于零，数据资源则呈爆炸式增长。根据IDC等机构的估算，2000年，全球生成数据总量约为1000~2000PB；到2010年生成数据总量已达2ZB，增加1千倍以上。与此相对应，每年存储下来的数据也同样持续高速攀升，但其增速与生成数据增速之间却存在较大差距。根据IDC与希捷共同发布的研究报告《数字化世界——从边缘到核心》，2020年全球数据存储介质的出货容量大约为2ZB，而同年全球新产生的数据总量达到64ZB。这意味着，2020年全球实际被存储下来的数据大约仅占当年新生成数据总量的3.4%（见表1）。

虽然我们并不能确切统计每年新增的存储数据规模，但可以通过全球存储介质出货容量予以间接佐证IDC的上述判断。当前，数据主要是存储于机械硬盘和固态硬盘两类介质。近年来，在数据产量爆发式增长趋势下，固态硬盘在存取速度等技术层面展现出显著优势，其市场规模和存储容量也稳步攀升。然而，从技术升级迭代到实现最终替换本身需要较长时间，而且机械硬盘在价格和容量方面还具备绝对优势；因此，机械硬盘年出货容量占据全球数据存储出货容量比重仍保持在60%以上（Reinsel et al., 2017）。全球机械硬盘的生产供给主要集中于希捷、西部数据和东芝三家厂商。根据调研机构Trendfocus统计，2020年上述三家厂商机械硬盘出货容量分别达到479.99EB、422.23EB和116.1EB。据此，利用企业财务报告数据和前述60%比例结构，可以大致估算全球新增存储介质容量，并间接反映数据资源每年新增存量规模。相关估算数据与IDC报告数据从趋势到规模都较为接近，二者可以相互形成印证。

需要特别指出的是，每年新增存储介质容量与当年新存储数据之间并不会完全重合。一方面，当年新增存储介质的利用率不可能达到100%；另一方面，每年存量数据中会有一定比例被删除，由

表1：全球数据规模及存储出货容量规模

（单位：ZB）

年份	全球新产生数据总量	全球数据存储介质出货容量	占比	机械硬盘出货容量	全球存储介质出货容量估算
2010	2	0.5	25.3%		
2011	5	0.5	10.9%		
2012	6.5	0.6	9.6%		
2013	9	0.7	7.8%		
2014	12.5	0.8	6.3%		
2015	15.5	0.8	5.5%	0.5	0.7
2016	18	1.0	5.6%	0.5	0.8
2017	26	1.2	4.5%	0.6	0.9
2018	33	1.5	4.6%	0.7	1.1
2019	41	1.8	4.3%	0.8	1.2
2020	64.2	2.2	3.4%	1.0	1.6
2021	79	2.6	3.2%		

资料来源：（1）列2数据参考Statista，2010~2025 年全球创建、使用和存储的数据量；（2）列3数据参考IDC、希捷，《数字化世界——从边缘到核心》，2019；（3）列5数据参考希捷、西部数据公司财务年度报告；（4）列6数据由作者估算得出。

此腾出的存储空间也会被用于新增数据的存储。因此，每年生成数据被存储的比例与表2中第四列比例相比会有一定出入。考虑到存储介质购置和维护都是有成本的，存储介质购置规模（出货量）加上少量既有数据删除所形成的新增存储空间，相对于每年的实际新增存储数据不会有过多冗余。假定冗余比例、删除比例相对稳定，数据存储介质出货容量规模基本可以反映出新增存储数据规模。

2022年，华为公司联合罗兰贝格发布的《数据存力，高质量发展的数字基石》报告，提出了“数据存力”概念，并将“区域内数据存储设备容量与当年生成数据总量之比”作为衡量数据存力充足性的指标。报告对全球各国数据存力充足性进行测算，结果显示：美国数据存力充足性指数高达19.4%，排名全球第一；而中国数据存力充足性指数为8.9%，远低于美国，位列第11位。中美两国同为数据生产大国，每年新生成数据量稳居世界前2位，数据存力充足性的差距主要源于中国以往在存储方面投资的不足。近年来，各方也在加大存储领域投资力度以弥补前期投资不足；2017~2019年中国存储领域投资年均增长率高达45.4%（见表2）。

另外，根据《数字中国发展报告（2021年）》，2017~2021年，中国每年新生成数据从2.3ZB增长至6.6ZB，平均年增长率30.2%。[③]结合华为报告给出的数据存力充足性指数，2020年中国存量的数据存储设备容量大约为0.59ZB；再考虑一定的冗余量，当年累计存储下来的数据资源应该在0.5ZB左右。而中国信息通信研究院（2022）最新公布的估算数据显示，2021年中国存储总量容量达到800EB，此外，《国家数据资源调查报告（2020）》指出，2019年底中国数据总存量为332EB，[④]这两项

表2：2020年数据存储充足程度前十国家及中国情况

（单位：%）

存力充足性排名	国家	数据存力充足性	存储投资年均增长率（2017~2019年）
1	美国	19.4	3
2	新加坡	18.8	10.7
3	德国	18.4	20.4
4	瑞典	17.9	21.1
5	英国	15.5	18.5
6	加拿大	14.7	−0.4
7	南非	13.5	22.3
8	日本	11.8	17.8
9	法国	10	28.1
10	捷克	9.8	18.8
11	中国	8.9	45.4

资料来源：华为、罗兰贝格，《数据存力，高质量发展的数字基石》，2022。

注：数据存力充足性指标，通过数据存储设备容量与当年区域产生的数据总量比值来表示。

[③] 数据由国家互联网信息办公室2022年公布。

[④] 数据由中国信息通信研究院、中国网络空间研究院2021年公布。

数据在量级上正好验证了上述推算结果。如果新增存储数据占比以40%计算，[⑤]那么2020年新增数据存储量大约在0.2ZB左右，在5.1ZB新生成数据规模中的比重大约为3.9%。该比重与表1中基于数据存储介质出货容量计算出的全球比重大致相当，形成相互印证。

（二）数据流动交易规模估算

从前述第二部分可知，数据交易流动模式繁杂多样，不同模式对应着不同的统计口径，很难就各种模式对应流动交易数据规模一一进行估算。然而，除了公共数据共享外，大多数的数据流动都是通过交易方式实现的，其中很大一部分交易会以货币形式支付对价。每年数据交易的成交金额也不失为反映数据流动交易规模的一种方式。

现有的行业统计资料将大数据市场定义为广义的数据产业市场，涵盖范围包括了大数据相关硬件、大数据分析软件和大数据专业服务。其中，大数据专业服务主要是数据交易服务，其涵盖范围与本文的数据交易市场更为接近。从全球不同机构给出的统计数据来看，虽然统计覆盖的范围可能存在差异，数据市场规模的口径也有所不同，但基本都围绕与数据相关的交易和服务开展规模估算。

Statista和Wikibon两个机构针对全球数据市场的收入规模进行了估算和预测（见表3）。两个机

表3：2011~2027年全球大数据市场收入规模估算及预测

（单位：十亿美元、%）

年份	Statista 2020	Wikibon		
		总规模	专业服务规模	专业服务占比
2011	7.6			
2012	12.25	–	–	–
2013	19.6	–	–	–
2014	18.3	18.3	7.6	41.53
2015	22.6	22.6	9.1	40.27
2016	28	27.3	11.1	40.66
2017	35	33.5	13.4	40.00
2018*	42	40.8	15.8	38.73
2019*	49	49	18.2	37.14
2020*	56	57.3	20.3	35.43
2021*	64	65.2	22.0	33.74
2022*	70	72.4	23.3	32.18
2023*	77	78.7	24.3	30.88
2024*	84	84	25.1	29.88
2025*	90	88.5	25.8	29.15
2026*	96	92.2	26.3	28.52
2027*	103	–	–	–

资料来源：（1）Statista, Forecast Revenue big data market worldwide 2011-2027；（2）Wikibon, 2016-2026 Worldwide Big Data Market Forecast.

[⑤] 根据IDC的估计，全球存量数据中，大约有90%是在过去三年中创造的，以每年40%左右递增。

构对于全球大数据市场规模的估算基本一致。其中，Wikibon还专门估算并预测了“专业服务规模”。根据两个机构的界定，“大数据市场”是涵盖大数据相关硬件、软件和大数据专业服务的广义大数据产业市场；而大数据专业服务主要就是数据交易服务，其业务规模大致对应于数据交易服务收入总额。根据预测结果，2021年全球数据交易服务收入大约为220亿美元，结合数据交易服务收入占成交额的大致比重，则可以进一步估算数据要素的成交规模。如果假定数据交易服务收入占成交额的比重为20%~25%，那么对应的数据要素成交规模大致为880亿~1100亿美元。

2022年，IDC和The Lisbon Council主导的欧盟数据市场检测工具报告针对全球主要经济体的数据市场交易规模进行了估算和预测。该报告将“数据市场”定义为“将数据作为产品或服务进行交换的市场”（IDC and The Lisbon Council, 2022），其涵盖范围包括与大数据相关的信息和IT服务、研究、商业活动及大数据分析服务，与Statistica和Wikibon界定的“数据市场”存在较大差异。IDC和Lisbon Council报告对数据市场规模的估算参考了数据相关企业在本地区总收入以及从其他地区进口数据产品/服务的价值，因此，在一定程度上反映了数据要素（产品/服务）的交易规模。

报告公布的2019~2021各年估算结果见表4。其中，2021年美国的数据市场交易规模为2399.6亿欧元，在前四大市场中的占比约64%，远高于第2名欧盟27国的636.3亿欧元，约为中国的8倍；而中国的数据市场交易规模仅为316.5亿欧元，低于日本的399.7亿欧元，仅排名第四；美国、欧盟、日本、中国的数据市场交易总规模大约为3752亿欧元，由此推断全球数据市场交易总规模在4000亿欧元左右。

需要指出的是，上述数据市场交易规模涵盖范围，既包括大数据相关的软硬件服务，也包括诸如直接出售数据等大数据相关的商业活动。IDC和The Lisbon Council报告还首次披露了欧盟数据相关企业通过直接出售数据（Data Monetization）所得。2020年，这部分收入达到116.1亿欧元，占欧盟整体数据市场规模的19%。如果比照该比例推算，2021年全球数据相关软硬件服务、研发以及大数据分析活动等业务规模则应该在3000亿欧元左右；而数据企业直接出售数据所得，或者说全球数据要素成交规模应该在800亿欧元左右，中国的数据要素成交规模则在60亿欧元左右。

表4：2019~2021年全球前四大数据市场价值

（单位：十亿欧元、%）

国家	2019	2020		2021*	
	规模	规模	增长率	规模	增长率
美国	184.87	213.46	15.46	239.96	12.41
欧盟27国	58.43	60.64	3.78	63.63	4.93
日本	32.93	36.65	11.30	39.97	9.06
中国	24.23	27.47	13.40	31.65	15.22

资料来源：IDC and Lisbon Council, 2022。

注：*为预测值。

从增长速度来看，2019~2021年，中国、美国、日本均保持了高速成长的态势，年均名义增长率都在10个百分点以上，分别为14.3%、13.9%和10.2%；相比之下，欧盟的增速仅为4.35%，明显低于其他三大经济体。高成长速度从侧面反映出当前全球数据交易市场仍处于初期快速成长阶段。

国际数据服务商OnAudience对全球27个国家在线市场营销领域（主要为数字广告）的数据交易规模进行了追踪和测算。表5展示了2017~2021年全球及排名前三在线营销数据市场交易规模及其在全球占比。从中可以看出，2021年，全球在线营销数据交易规模达到523亿美元，其中，美国市场规模为306亿美元，位居全球第一，占比高达58.51%；中国市场虽排名全球第三，但市场规模仅为73亿美元，全球占比13.96%，远低于美国。

需要特别指出的是，基于现有可获取资料对数据交易规模/成交金额的估算，不仅由于资料来源、覆盖口径的不同而存在较大差距，更无法完全反映出数据流动交易的整体规模。一方面，大部分公共数据的流动共享以及基于平台的双重交易，本身并不采取交易或货币支付的方式，但也是数据流动交易的重要组成部分；另一方面，现有的统计资料以及可公开查询的企业财务数据，远不能涵盖所有以货币方式支付的数据交易，例如，众多企业微观主体间点对点的数据交易活动一般很难从公开渠道找到对应信息。因此，当下有关数据交易市场的统计，涵盖范围主要是那些依托公开交易平台达成的数据交易活动；而且成交金额往往还需要通过平台提供交易服务所获得收入进行间接推算。

（三）数据要素规模结构性特征分析

本部分着眼于生产存储和流动交易两个方面，基于不同渠道的公开资源，从数据资源体量和交易成交货币价值两个角度对数据要素规模进行了估算。尽管由于资料来源、范围口径等因素使得估算结果存在较大差异，但在此基础上仍可以提炼出当前数据要素规模方面的一些结构性特征。

首先，近年来全球数据生成规模和数据存储规模都在持续快速增长，但每年新增数据存储规模远小于数据生成规模，仅有4%左右的新生成数据能够被存储下来。另外，从存储能力来看，各主

表5：2017~2021年全球及排名前三在线营销数据市场交易规模及全球占比

（单位：十亿美元、%）

	2017年		2018年		2019年		2020年		2021年	
国家/地区	价值	占比	价值	占比	价值	占比	价值	占比	价值	占比
全球	18.9	100	26.5	100	34.6	100	41.4	100	52.3	100
美国	12.3	65.08	16.6	62.64	21.2	61.27	24.7	59.66	30.6	58.51
欧洲	2.8	14.81	4.1	15.47	5.3	15.32	6.3	15.22	7.6	14.53
中国	1.7	8.99	2.8	10.57	4.1	11.85	5.4	13.04	7.3	13.96

资料来源：OnAudience.com, “Global Data Market Size (2017-2021)”。

要经济体的数据存力充足性普遍偏低，排名最高的美国也不足20%，中国更是不足10%。

第二，中国数据生成规模与数据存储规模间的失衡更为明显。得益于超大规模市场和丰富应用场景，中国数据生产能力逐年大幅提升，数据生成规模与美国一起居世界前两位，但数据存储率仅为3%左右，明显低于世界平均水平。从存力充足性指标来看，中国在全球排名仅为11，与排名第1的美国存在显著差距。

第三，数据要素交易市场尚处于初期快速发展阶段，成交规模与数字经济规模相比显得微不足道。受到涵盖范围、统计口径以及资料获取等诸多因素的影响，全球数据流动交易规模估算目前尚无公认的准确测算结果，但不同测算结果在数量级上基本能相互印证。综合不同机构公布的测算结果推断，全球公开交易数据的成交规模数量级大致为千亿美元，与全球数字经济增加值规模相差两个数量级。巨大差距背后既有统计漏算的原因，更源于数据要素市场发展尚处于初期成长阶段。

第四，相比美国、日本等国家，中国的数据要素市场发育更不充分。一方面，数据市场交易规模、在线营销领域数据交易在全球占比均为10%左右，低于欧盟、日本，在主要经济体中排名第4。另一方面，相比美国、欧盟和日本，中国数据市场成交规模保持着更高的增长率，在四大经济体中位居第一。当然，中国数据要素成交规模在四大经济体中的占比可能存在较大幅度的低估，毕竟丰富应用场景必然带来充足的数据要素交易需求。

五、总结性评论及建议

社会各界高度关注数据要素及其流动交易，但存在过于聚焦个人行为数据和数据交易所模式的误区。为此，本文前述各部分以数据相关基本概念辨析为出发点，区分比特数据、数据资源和数据要素的内涵，并对数据分类的不同视角进行归纳。综合既有研究文献和相关实践，针对不同类型数据系统梳理其流动交易模式及特点，并综合不同来源的资料，对全球主要经济体当下数据资源规模和数据交易规模进行了估算和分析。旨在通过我们的分类、梳理和估算，尽可能客观全面地呈现数字要素流动交易的基本状况及趋势特征。具体来说有如下几点判断：

第一，丰富的现实场景决定了数据分类的多样性，而不同类型数据的涉及主体、权属划分、信息密度等特征存在较大差异，由此带来数据流动交易模式及收益分配的复杂性。促进数据要素安全、有序、充分流动，亟需顺应多样性和复杂性特点，从分级分类等基础性工作入手，结合数据流动交易实践需要，逐步构建完善数据分类和流动交易相关制度。

第二，数据流动交易呈现出产品化、服务化和平台化趋势特征。原始数据转移的流动交易模式依然存在，但对于数据需求方来说，原始数据并不能直接满足其特定应用场景的需要。数据供给方为需求方提供定制化的数据产品或数据服务，既省去了需求方自身加工数据的繁琐，又实现了数据增值，更为重要的是还能降低原始数据转移带来的隐私和信息泄露风险。至于数据交易平台化则更

是体现了数字经济运行实践最重要的特征，提高了数据流动交易的效率。

第三，现行统计核算体系是工业经济时代的产物，没有也不可能针对数据资源/数据要素提前预设专门的统计分类，不同机构给出的数据要素规模估算，受到资料来源、范围口径以及估算方法等因素的影响，存在较大差异。加强数据要素规模估算方面的探索，既是健全完善数据要素流动交易制度的一项基础性工作，也是数字经济研究特别是数字经济测算领域的前沿课题。

第四，全球新生成数据中仅有约4%能够被存储，主要经济体都存在数据生成规模与数据存储规模失衡，美国情况相对较好，而中国的数据生产与存储结构性失衡尤为突出。中美两国所拥有的数据资源和数据生产能力远远领先于其他经济体，但中国的存力充足性指标仅为全球第11，相比美国还有很大差距，数据存储能力提升空间巨大。

第五，全球数据交易市场整体还处于初期快速发展阶段，数据公开交易的成交规模数量级大致为千亿美元，与数字经济增加值规模相差达两个数量级。在主要经济体中，中国的数据要素市场发育更不充分，成交规模占比不足10%，不仅远小于美国，也低于欧盟和日本。

要促进数据要素充分流动，切实发挥数据作为新关键要素对经济增长和社会发展的支撑作用，社会各界都应提高对数字经济和数据要素的认知水平。在认识到数据要素重要性的基础上，更要充分意识到数据要素分类和数据流动交易的多样性、复杂性以及相关制度体系构建完善的艰巨性、长期性，并顺应这些特点从基础制度建设入手，对数据流动交易活动进行引导和规范。

第一，围绕数据要素本身，完善分类分级、权属界定等制度建设，为促进数据充分流动交易提供基础性的制度保障。由行业主管部门牵头，结合数据生产和数据应用的具体实践，比照工业和信息化部办公厅引发的《工业数据分类分级指南（试行）》，针对重点领域、重要场景陆续出台相应的分类分级标准（或指南），不断健全数据分级分类制度体系。完善数据要素权属划分相关的制度建设，在数据分类分级、确权授权使用基础上，细化、明确不同类型数据持有权、使用权、经营权等相关权属边界，为数据共享流动各环节参与者、利益相关者的正当权益提供法律和制度保障。

第二，加强存储中心、算力中心等数字基础设施建设，提高存力、算力充足性水平，保障数据资源的收集、存储。以“东数西算”等国家级战略为基础，对算力中心和存储中心建设规划进行细化；除了现有的8个国家级枢纽节点外，还可以考虑在其他具备气候、电力等相对优势的西部地区布局建设更多存储中心。

第三，培育一批数据要素交易和开发利用的专业化机构，充分发挥企业的市场主体作用及技术优势。数据要素的市场化配置及价值挖掘，对技术能力、专业积累等要求较高，应充分调动数据企业的积极性。一方面鼓励企业，特别是互联网巨头企业更多参与数据交易环节，通过提供数据产品和服务，更好盘活数据资源，推动数据要素市场规模增长。另一方面，借助企业在数据资源、数字技术、应用场景等方面的专业积累，为政府部门和各类市场主体提供专业化的产品和服务，提升数据

要素市场的交易和服务质量。

第四，建设并完善数据治理体系，统筹好数据流动中的安全与发展。充分保护数据信息安全、推动数据相关行业领域健康发展，需要参与各方加强对关键领域数据的保护，事前、事中、事后做好风险预警、安全评估和事后追溯；市场监管部门、行业主管部门和司法部门发挥数据治理主导作用，确保司法和执法全面到位，在鼓励和引导数据新兴商业模式创新发展的同时，防止出现数据垄断、价格歧视等问题。此外，健全数据跨境流动相关的安全审查制度和双边多边合作机制，为实现数据在更大范围内的流动交易保驾护航。

第五，支持并鼓励数字相关核心关键技术创新和突破。数据实现安全、有序、充分流动不仅需要制度和相关产业政策的保障和支持，更需要核心技术提供坚实的物质基础。一方面，守住数据安全底线，实现数据可靠溯源、安全存储和传输以及数据可用不可见，在数据流动交易中维护国家数据安全、防止个人隐私和商业机密泄露，需要安全多方计算、区块链等较为成熟可靠的技术提供保障。另一方面，数据资源的开发利用需要充足的数据存储能力和算力，存储介质方面固态硬盘占据主导的趋势已经形成，而其核心的闪存颗粒目前被国外寡头垄断。唯有切实解决好此类核心关键技术卡脖子问题，才能更好地发挥数据要素支撑经济发展作用。

参考文献

[1] 蔡跃洲、马文君，2021，《数据要素对高质量发展影响与数据流动制约》，《数量经济技术经济研究》第3期。

[2] 常江、张震，2022，《论公共数据授权运营的特点、性质及法律规制》，《法治研究》第2期。

[3] 田杰棠、刘露瑶，2020，《交易模式、权利界定与数据要素市场培育》，《改革》第7期。

[4] 中国信息通信研究院，2022，《中国存力白皮书（2022年）》。

[5] Acquisti, A., C. Taylor, and L. Wagman. 2016. “The Economics of Privacy.” *Journal of Economic Literature*, 54(2): 442–92.

[6] Bergemann, D. and A. Bonatti. 2019.“Markets for Information: An Introduction.” *Annual Review of Economics*, 11:85–107.

[7] CAICT and Chinese Academy of Cyberspace Studies (CACS). 2021. *National Data Resources Survey Report (2020)*.

[8] European Commission, Directorate-General for Communications Networks, Content and Technology, Scaria, E., Berghmans, A., Pont, M., et al. 2018. *Study on Data Sharing between Companies in Europe : Final Report*, Publications Office.

[9] Farboodi, M., and Veldkamp L. 2021. “A Growth Model of the Data Economy.” *NBER Working Paper*, No.28427.

[10] Federal Trade Commission. 2014. “Data Brokers: A Call for Transparency and Accountability.”

[11] IDC and The Lisbon Council. 2022. *European Data Market Study 2021-2023 (D2.1 First Report on Facts and Figures)*. Luxembourg: Publications Office of the European Union.

[12] Laudon, K. C. 1996. “Markets and Privacy.” *Communications of the ACM*, 39(9): 92–104.

[13] Malgieri, G., and Custers B. 2018. “Pricing Privacy - The Right to Know the Value of Your Personal Data.” *Computer Law & Security Review*, 34(2): 289-303.

[14] Mitchell, John, Daniel Ker and Molly Lesher. 2021. “Measuring the Economic Value of Data.” OECD Going Digital Toolkit Note, No.20.

[15] Nguyen, D. and M. Paczos. 2020. “Measuring the Economic Value of Data and Cross-border Data Flows: A Business Perspective.” *OECD Digital Economy Papers*, No. 297.

[16] OECD. 2013a. “Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value.” *OECD Digital Economy Papers*, No. 220.

[17] OECD. 2013b. “Introduction to Data and Analytics (Module 1): Taxonomy, Data Governance Issues and Implications for Further Work.” DSTI/ICCP (2013)13.

[18] OECD. 2021. *Issues Paper: Recording Observable Phenomena and Data in the National Accounts*. Paris: OECD Publishing.

[19] Reinsel, D., J. Gantz, and J. Rydning. 2017. *Data Age 2025*. IDC.

[20] Statistics Canada. 2019. “The Value of Data in Canada: Experimental Estimates.” *Latest Developments in the Canada Economic Accounts* (Working Paper Series), No.9.

[21] Swedish National Board of Trade. 2015. *No Transfer, No Production: The Importance of Cross–border Data Transfers for Companies Based in Sweden*. Stockholm: Swedish National Board of Trade.

[22] The State Internet Information Office. 2022. *Digital China Development Report (2021)*.

[23] United Nations Conference on Trade and Development(UNCTAD). 2021. *Digital Economy Report 2021 Cross–border Data Flows and Development: For Whom the Data Flow*. Geneva: United Nations Publications.

[24] Varian, H. 2018. “Artificial Intelligence, Economics, and Industrial Organization.” *NBER Working Paper*, No. 24839.

[25] Veldkamp, L., and C. Chung. 2019. “Data and the Aggregate Economy.” *Journal of Economic Literature*, forthcoming.

Data Flow & Transaction Mode Classification and An Explorative Estimation on Data Storage & Transaction Volume

Cai Yuezhou[*1] and Liu Yuexin[2]

[1] *Institute of Quantitative & Technological Economics (IQTE), Chinese Academy of Social Sciences (CASS), Beijing, China*

[2] *University of Chinese Academy of Social Sciences (CASS)*

Abstract: *The public has shown great interest in the data factor and data transactions, but the current attention is overly focused on personal behavioral data and transactions happening at Data Exchanges. To deliver a complete picture of data flow and transaction, this paper presents a systematic overview of the flow and transaction of personal, corporate and public data on the basis of data factor classification from various perspectives. By utilizing various sources of information, this paper estimates the volume of data generation & storage and the volume & trend of data market transactions for major economies in the world with the following findings: (i) Data classification is diverse due to a broad variety of applying scenarios, and data transaction and profit distribution are complex due to heterogenous entities, ownerships, information density and other attributes of different data types. (ii) Global data transaction has presented with the characteristics of productization, servitization and platform-based mode. (iii) For major economies, there is a commonly observed disequilibrium between data generation scale and storage scale, which is particularly striking for China. (iv) The global data market is in a nascent stage of rapid development with a transaction volume of about 100 billion US dollars, and China's data market is even more underdeveloped and only accounts for some 10% of the world total. All sectors of the society should be fully aware of the diversity and complexity of data factor classification and data transactions, as well as the arduous and long-term nature of developing and improving relevant institutional systems. Adapting to such features, efforts should be made to improve data classification, enhance computing infrastructure development, foster professional data transaction and development institutions, and perfect the data governance system.*

Keywords: *Data factor, data classification, data transaction mode, data generation & storage volume, data transaction volume*

JEL Classification Code: D83, O25

DOI: 10.19602/j.chinaeconomist.2022.11.04

1. Introduction

Around 2008, the New Generation of Information and Communication Technologies led by mobile internet, cloud computing and artificial intelligence (AI) have brought a revolutionary change

[*] CONTACT: Cai Yuezhou, email: caiyuezhou@cass.org.cn.

Acknowledgement: This paper is a staged achievement of the General Project of the National Science Fund of China (NSFC) "Theoretical and Empirical Studies on How New Generation of Information Technology Affects Growth Impetus and Industrial Structure (Grant No. 71873144); NSFC Major Project "Macroeconomic Big Data Modelling and Prediction" (Grant No. 71991475).

in data generation, collection, transmission, processing and analysis, and generated an abundance of data resources underpinning data analytics and applications. These developments have given rise to myriad new business modes such as the platform economy and sharing economy, spearheading a new round of socio-economic restructuring. The *Decisions of the CPC Central Committee on Adhering to and Improving the Socialist System with Chinese Characteristics and Advancing the Modernization of the State Governance System and Capacity* adopted at the Fourth Plenary Session of the 19th CPC Central Committee in 2019 called for "perfecting the mechanism in which the contributions of labor, capital, land, knowledge, technology, management, data and other factors of production are evaluated in a market-based manner, and return on those factors of production is determined according to their contributions." This policy statement explicitly identifies data as the seventh factor of production. In the digital economy era, it is generally recognized that data is pivotal to business operations and macroeconomic growth.

In the digital economy era, data in bits is characterized by the technical-economic attributes of non-competitiveness, non-exclusiveness, low-cost reproduction, network externalities, and instantaneity. Those attributes are conducive to economic efficiency at the firm level and help realize the multiplier effect of value creation more broadly (Cai and Ma, 2021). A key premise for those effects to be brought into play is the secure, orderly and sufficient flow of data. To this end, China's macroeconomic authorities have enacted a succession of policy documents on data transactions and factor market development. In December 2021, the State Council General Office released the *Overall Program for Comprehensive Pilot Reforms of Market-Based Factor Allocation*, which called for improving mechanisms to open and share public data, developing rules on data flow and transaction, broadening standard use cases for data development and utilization, enhancing data security protection, and creating rules on the circulation of data factor. It has also spelled out the principle that "raw data should stay within their origin and be available yet invisible."

On June 22, 2022, the Central Commission for Comprehensively Deepening Reform adopted the *Opinions on Creating Data Fundamental Institutional Systems to Better Leverage the Role of the Data Factor*, which called for speeding up the creation of fundamental institutions and advancing data property rights, data transactions, profit distribution and security governance to ensure data circulation and empower the real economy. The *Cybersecurity Law*, the *Data Security Law* and the *Personal Information Protection Law* enacted since 2016 have standardized data flow and transactions from the perspective of data security.

Since the Guiyang Big Data Exchange was launched in 2015, China's local governments - including Beijing, Shanghai, Shenzhen, Guangzhou, Fujian and Zhengzhou - have rushed to establish local data exchange centers to standardize data flow and transactions. Under government initiatives and media coverage, the public and academia have focused their attention on the data flow and transaction mode via exchange centers and personal behavioral data involving rather complex ownerships, and relevant academic discussions and institutional construction have also leaned towards those issues.

The digital economy, however, has generated a diverse range of data types, and the flow and transaction modes of different data types vary considerably. Personal behavioral data generated from the consumer internet is only one part of an ocean of society-wide data resources, and matchmaking at data exchange centers is one way to realize data transactions. More transactions can be conducted directly between suppliers and users of data products/services. The government should develop fundamental institutional systems for data as a critical factor of production to support efficiency improvement and value addition. It is necessary to take stock of data flow and transaction modes and identify the structure and tendency of data resources on the basis of data identification and classification.

Hence, the following sections of this paper will start by analyzing basic concepts of data, systematically review and analyze different data transaction types, present a panoramic view of actual

data transactions with identifying their characteristics and trends, so as to eliminate misunderstandings. In terms of the existing scale of data resources and data transaction volume, this paper presents and further estimates the scale of data generation and storage, as well as the transaction volume of each data type, which provide more specific quantitative support to examine the structural characteristics of the data market. Finally, this paper puts forth policy recommendations as a basic reference for improving institutional mechanisms for the data market and promoting data transaction and sharing.

2. Concept and Classification of Data

An abundance of data resources is both a result of the commercial application of digital technology on a large scale and a new critical factor undergirding the digital economy. Methods of data classification are varied due to the diverse use cases of the digital economy. It is necessary, therefore, to systematically review data classification on the basis of analyzing the connotations and denotations of data before investigating data transaction modes and estimating data volumes.

2.1 Bit Data, Data Resources and the Data Factor

From physical and technical perspectives, "data" in the digital era broadly refers to binary-coded character strings that serve as a vehicle of information, i.e. "data in bit" or "digital data" (Cai and Ma, 2021; Farboodi and Veldkamp, 2021), which is generated based on the observations and records of socio-economic reality. To some extent, therefore, digital data can be seen as a byproduct of economic activity (Veldkamp and Chung, 2019). According to the OECD (2021), "Data is information content that is produced by accessing and observing phenomena, and recorded, organized, stored, processed or accessed in the digital format," i.e. data is seen as a unique form of information expressed in binary bit.

For the purpose of estimation, Statistics Canada defines "data" as the result of observation that has been converted into numeric/digital form and can be stored, transmitted or processed to generate knowledge. This definition delimits the scope of data to the observation of a specific matter at a certain time point, which is digitally recorded for storage, search, analysis and investigation. Obviously, such information as digital music and films is excluded (Statistics Canada, 2019). As a matter of fact, data has always been regarded as information or fact. In the digital era, data is more closely related to information and equal to information in many contexts (Cai and Ma, 2021).

Digital data, which exists in the form of binary character strings, require data analytics to extract valid information. Massive raw data initially collected and stored in bit form cannot be directly applied in producer and consumer use cases without processing, analysis and extraction of valid information. Therefore, raw data is not a factor of production that may directly participate in value creation and is instead merely a "data resource" (Varian, 2018; UNCTAD, 2021) with the potentials to create value. After cleansing, agglomeration, treatment and analysis, raw data is processed into data sets, databases, information reports, and data services, among other forms of data products and services, which can be meaningfully applied in various socio-economic use cases such as marketing, risk control and person search (FTC, 2014). As a factor of production, data products and services directly contribute to business value creation.

Data resources are characterized by such technical-economic attributes as non-competitiveness, partial exclusiveness, and lowcost of reproduction, which allows data to be theoretically reusable on a large scale to mitigate growth constraints from the scarcity of other tangible capital and achieve the multiplier effect (Cai and Ma, 2021). Yet businesses have to invest numerous human and material resources to cleanse, process and analyze raw data to extract valid information and turn such information into the data factor. The ability for such conversion is scarce (OECD, 2013). Given that raw data is the source from which valid information can be derived, all sorts of data resources ranging from raw data to processed data, data products and data services may all count as the data factor in the broad sense.

2.2 Data Classification and Types of Data from Various Perspectives

There is a great deal of diversity of data as digital records of complex socio-economic activities that extensively employ digital technology. To promote data transactions and the role of data in supporting socio-economic activity, it is necessary to classify data from different perspectives.

As digital records of complex socio-economic activities, data can be classified on various dimensions according to the characteristics or types of recorded matters. A common approach is to classify data according to the domains of data creation and practical scenarios. For instance, data can be classified into various sectoral data according to the classification of national economic sectors, including big data related to communication, finance, health, agriculture, transportation, and electric power. Data may also be classified according to each specific process of data recordation from the perspective of socio-economic activity. For instance, the *Guidance for the Classification of Industrial Data (for Trial Implementation)* released by the Ministry of Industry and Information Technology (MIIT) in 2020 has classified various business operation processes of industrial enterprises and divided the data records of each process into R&D data, production data, operation and maintenance data, management data and external data.

A more common approach is to classify data into "personal behavioral data," "corporate data" and "government and public sector data" according to the actors subject to data recordation. Personal behavioral data refers to data of user activities such as browsing, search, interaction and transaction recorded by internet platforms on a real-time basis. Such data includes, among others, shopping records at e-commerce platforms like Taobao and JD.com, and chats & messages at social media platforms like WeChat and Weibo.

Corporate data is generated from the recoding and monitoring of various business processes, and includes data collected by manufacturers via sensors to monitor and report the operational status of intelligent production lines.

Government and public sector data refer to all data resources created by and collected from government agencies at various levels, public administration institutions, and public utilities such as electric power, public transit, fuel gas, heat supply and water supply and drainage.

Those data resources are related to the provision of public infrastructure and services, and include tax and customs data, corporate qualification and credibility information, natural resources data, traffic information, electric power scheduling data, as well as municipal road and pipeline distribution and operational status.

Similar classification by UNCTAD (2021) divides data into consumer data, commercial data, and government and public data. Issues of public concern include funding for data collection and maintenance, and data ownership, among others. According to the sources of funds for data creation, maintenance and possession, data can be divided into private-sector data and public-sector data. According to legal rights such as the right of ownership and the right of use, data can be divided into public data and proprietary data, and the latter refers to data with explicit ownership and protected by intellectual property rights or similar laws (Swedish National Board of Trade, 2015; Nguyen and Paczos, 2020).

Aside from the recorded matters, data may also be classified into raw data, processed data, data products/services and metadata according to the attributes of information content. Judging by the scope of data flow, data includes domestic and cross-border flows of data. Notably, the above methods of data classification are not exclusive of each other. Under different classification criteria, the same data (set) can be classified into various types of data simultaneously.

3. Data Flow and Transaction Modes

The sufficient flow of data is a key premise for data to serve as a critical factor of production,

contribute to efficiency, and help realize the value multiplier effect. Data flow is closely related to the type of data and influenced by differences in data subjects, ownership, the content of valid information, and information intensity. Data exchange centers are only one of the myriad modes of data flow and transaction in the digital economy but have received the most public attention. Only with insights about the modes, characteristics, and trends of data flows will policymakers be able to craft policies to promote efficient and standard data transactions. Hence, we will classify data according to the above-mentioned data subjects for an analysis of the flow and transaction modes of personal, corporate and public data. Notably, data flow includes data transactions, which refers to the mode of data flow in the private sector, and public data normally flows in open and sharing modes.

3.1 Personal Behavioral Data Transactions Modes

Personal behavioral data refers to data records related to personal consumption behaviors, i.e. consumer behavior data. Consumer behavior records predate the advent of the internet, but are scattered among various merchants. Complete personal behavior records are collected and compiled often by third-party data intermediaries, which gave rise to the embryonic concepts of the data industry and "data agents." In the mid-and late 1990s, consumer internet platforms led by e-commerce gained ground with the increasing penetration of personal computers and internet. It was not until then that personal behavioral data and its transactions entered the horizon of academic research and became systematically collected by internet platforms (Laudon, 1996). After two decades of rapid development, consumer internet platforms in various niche sectors have not only transformed people's ways of consumption but created favorable conditions for the generation and collection of personal data, becoming the primary collectors of personal behavioral data.

From the perspective of data subjects, internet platforms generate and collect personal behavioral data of the following three types: (i) Data shared by an individual related to him/herself or a third party at his/her initiative or under the terms and conditions of the platform, e.g. social network profiles and online shopping records; (ii) data that can be lawfully observed and captured by recording user activity without user authorization, e.g. webpage browsing data and mobile phone location data; (iii) derivative data obtained based on personal data analysis.[1]

In some cases, personal data may also be derived from a few entries of seemingly anonymous data. In terms of data source, personal behavioral data collected by consumer internet platforms can be divided into two categories: First, first-hand data collected by online platforms based on their digital products and services; second, data of user activity outside the platforms collected by a third-party (OECD, 2013). The separation between data subjects and data collectors has added to the complexity of personal behavioral data transactions, which are twofold and involve at least three parties.

Recordation and collection of personal behavioral data can be regarded as the first fold of transactions. The parties of transactions are data subjects (platform users) and data collectors (internet platforms). By providing free access to or provision of specific application services (free use of apps), data collectors obtain the right to collect and record data of users' personal (consumption) behaviors and thus accumulate raw data resources. Since consumption records include consumers' private information, such transactions can be seen as consumers' exchange of their private information for free services from internet platforms. As such, research on privacy pricing is often correlated with consumers' behavioral data (FTC, 2014; Acquisti et al., 2016). In addition, the first data transaction is a swap between data of consumers and digital services from internet platforms, which can be classified as a "composite transaction"[2] (Malgieri and Custers, 2018).

[1] For instance, credit score can be calculated according to many factors related to personal financial history.

[2] Composite Transaction means a contemporaneous sale, refinancing or other disposition of all three properties (exclusive of the Distribution Center).

Data transactions between a data collector/internet platform and a thirdparty are the second fold of transactions, in which the data collector buys additional data from a thirdparty to enrich the existing behavioral data to improve platform service quality by realizing product innovation, optimizing supply chain structure and increasing marketing efficiency. The data collector may also provide raw data or data products/services - such as marketing services and credit evaluation - to a thirdparty to bring about the potential value of data through more extensive data flow and application. The third party could be another data platform, a data integrator, or a business or public entity with specific data demand.

The second fold of transactions is further divided into direct transaction and indirect transaction (Tian and Liu, 2020; Bergemann and Bonatti, 2019; Veldkamp and Chung, 2019). Under the direct transaction mode, the seller (supplier) only performs preliminary data processing without exploring potential information, and sells products mainly in the form of data sets. Under the indirect transaction mode, the seller sells customized data products or services after cleansing, arranging, analyzing, and mining raw data. The indirect mode of transaction does not involve the exchange of raw data. In addition, privacy computing - such as federated learning and secure multi-party computation (SMPC) - and blockchain technology allow multi-party joint data analysis to be carried out free from data divulgence, achieving data availability and invisibility. Their applications in various use cases may enhance the protection of privacy and supplement the indirect mode of transaction.

3.2 Corporate Data Transaction and Sharing Modes

Enterprises with a knack for digital applications may generate and accumulate a wealth of monitoring and recording data in their daily business operations. Such data includes, for instance, equipment operation status data captured by sensors and the internet of things (IoT), as well as manufacturing, sales and logistics information generated by corporate internal IT systems. Compared with behavioral data, the ownership of corporate data is relatively simple. In most cases, enterprises are both data subjects and data collectors and free from ownership controversy.

Theoretically, the flow and reuse of corporate data not only help increase the synergy of upstream and downstream industrial chains and corporate profitability, but raise overall social welfare on a broader scale. In practice, however, data exchange and sharing between enterprises may not occur spontaneously. The intent of firms to sell or share data is also subject to firm competition. Only when the scenario of data reuse and the original use of source firms are independent of each other or complementary to each other will firms become motivated to share data. Specifically, firms as the collectors and owners of data have the following three motivations to sell or share their data resources/products: (i) To sell data products or services directly to generate business revenue; (ii) to increase synergy with affiliates and thus optimize supply chains, develop products, improve services, and innovate business modes; (iii) achieve a more efficient supply-demand match. From a data demand perspective, buyers acquire firm data mainly to develop new products or improve existing ones, boost productivity, build customer relations, optimize corporate internal management structure, and identify precise market goals (European Commission et al., 2018).

In terms of the digital economy practice, corporate data transactions and sharing fall into the following categories: (i) Direct transactions led by data firms, i.e. data owners sell data products or provide data services to data buyers directly to make a profit; such transactions is performed when a data owner provides a buyer with a data interface and authorized access.

(ii) Under the intermediary transaction mode, data platforms bring data suppliers in touch with data users and serve as trustworthy thirdparty intermediaries that match data supply with demand and facilitate transactions to earn commissions. In terms of transaction practice, thirdparty intermediaries may include integrated cloud service platforms like the Amazon Web Services (AWS) Data Exchange platform, specialized data transaction platforms such as Dawex and big data exchange centers, or professional data integrators such as Dun & Bradstreet.

(iii) Under the data sharing mode via industrial internet, upstream and downstream enterprises access industrial internet platforms in a secure environment and share a certain scope of data to promote new product development or efficiency improvement. Normally, connected enterprises will share data to industrial internet platforms (operators) for free and access services or other data from those platforms in exchange. Airbus, for instance, launched "Skywise," an industrial internet platform, in 2017, which provides airlines with global fleet benchmark reports for free based on anonymized data collected from users (airlines). Skywise also offers airlines and aircraft manufacturers data interfaces to obtain such integrated anonymized aviation data for the latter to raise operational efficiency and improve aircraft software and hardware equipment. Notably, although data sharing over industrial internet primarily aims to spur business innovation and efficiency improvement and is usually free of charge, it is still a type of composite transactions in nature; data firms feed data into industrial internet platforms free of charge in consideration of access to free platform services. In addition, platforms usually charge fees when offering more advanced data products and services based on integrated data.

3.3 Public Data Flow Modes

Public data has a complete coverage of temporal and spatial dimensions and samples. Public data records of government and public sectors and utilities provide information about the external environment for individuals, firms and other entities. As a supplement to private sector data, public data serves as an indispensable data resource for data mining and big data analysis, hence the strong demand for accessing public data. In practice, the flow of public data is realized via open access. Normally, the government sector or public institutions as owners of public data selectively open data to the public on the basis of sufficient evaluation of data security and other factors (Cai and Ma, 2021).

Judging by the scope of data flow, public data flow can be further divided into data exchanges between public sectors and publicly available data. On one hand, the government establishes integrated data platforms covering various government agencies and public institutions to bring together sporadic data from various departments and institutions and form a more complete information set in order to identify problems in socio-economic operations, make targeted responses, carry out dynamic monitoring, and perform an *ex-post* evaluation to provide comprehensive and systematic first-hand information. On the other hand, massive public sector data about social operations involves transportation, hydrological, environmental and public security information closely related to business activity. Certain public sector data - once made publicly available in a secure and orderly manner - will greatly spur private sector efficiency and innovation.

Open access to government data has been a key priority of various governments in speeding up the digital transformation. As early as in 2009, some local governments in the US took the initiative to establish open public data platforms, followed by countries such as the UK, France, Canada and Singapore. After the *Suggestions on the 13th Five-year Plan* released in October 2015 called for implementing a "national big data strategy," the Chinese government also set out to develop public data platforms and promote open access to data. Depending on the sensitivity and security of content, public data can be divided into the three categories of unconditionally shared data, conditionally shared data, and restricted data. Unconditionally shared data can be accessed from public avenues such as public web portals, and access to restricted public data is subject to review and approval by the data authority and requires signing an agreement and security commitment.

In principle, access to public data should be free for all. Yet it takes massive human and financial resources to collect, store and process data, which cannot be accomplished by the government or public institutions alone. In many countries, therefore, the public sector often involves a business entity in processing public data and charges a fee for certain data services. For instance, the United Kingdom classifies public sector information holders (PSIHs) into the following three categories: (i) PSIHs that provide unrefined information; (ii) institutions that use PSI data to improve or support internal activities;

and (iii) commercially motivated institutions that profit from PSI. Type (iii) institutions usually charge a fee from third parties when providing them with data products or services based on raw public data.

Chinese public authorities have also been proactively exploring the modes of authorized access to public data. For instance, Shanghai's data regulations have called for "encouraging the development and utilization of public data," and the *Regulations of Zhejiang Province on Public Data* stipulates that the government may authorize eligible incorporated or unincorporated entities to operate public data, and the authorized operator may process public data based on public data platforms and provide users with data products and services for a profit. The above-mentioned operation is akin to the authorized operation of public services, the pricing of which is limited by the necessary cost of rendering such public services (Chang and Zhang, 2022).

3.4 Characteristics of Data Flow and Transaction Modes

As can be seen from the above discussions, the supply of personal, corporate and public data among the three types of actors is spontaneous and driven by the needs of data flow and transaction with the following characteristics:

First, the supply of data is diversified and involves myriad stakeholders. Data demand and use cases are varied, and so are data subjects, stakeholders and ownerships. These intertwined factors have led to complexities in the realization, payment and profit distribution of data transactions among data suppliers, users and third parties. Considering the privacy and information security issues related to data transaction, it is necessary to balance efficiency with security in determining the scope and level of transaction sharing.

Second, data tends to be offered as a product or service. Although the direct sales mode of raw data and indirect sales of data products and services both account for a certain proportion of data transactions, the latter is poised to dominate data transactions going forward. The supply of processed data products/services does not entail the reproduction and transfer of original data and is free from ownership dispute, making it conducive to privacy protection and national information security.

Third, data transactions is increasingly led by platforms or intermediaries. Despite some instances of point-to-point data exchanges directly between data suppliers and users, there is a tendency for data flow and data transactions in particular to occur over multilateral data markets led by data intermediaries, which can be platform enterprises, data integrators or data agents: (i) In some circumstances, data agents facilitate deals and charge commissions from sellers and buyers. The role of data intermediaries or agents is limited and replaceable. In realty, both sides of transaction may reach an agreement without a data agent. (ii) Data integrators are not only providers of data products or services, but data users as well. In consumer finance, for instance, it is banks and financial institutions that collect consumer data (such as credit history) and send such data to credit reference institutions in exchange of additional information about their existing and potential customers (Bergemann and Bonatti, 2019). Given their massive pool of data, data integrators boast competitive data quality, content and scope, as well as core algorithms, which make them hard to be replaced. (iii) As data intermediaries, platform enterprises not only facilitate transactions, but provide other value-added data transaction services. For instance, the AWS Data Exchange and the Shanghai Big Data Exchange provide one-stop services for data transactions with sufficient data security assurances.

4. Estimation of Data Storage and Transaction Volume

We may follow different standards in determining the scope of the data factor given the complexity and diversity of data classification and transaction. It takes a scientific and accurate estimation of the data factor's scale in order to characterize data resources and data market operations. This requires a clear definition of the scope of what is to be estimated. Based on the above classification of the data factor and

identification of data transaction modes, we may estimate both the volume of the data factor (resources) and the volume of data flow and transaction. While the former reflects the generation and storage scale of data, the latter provides an indirect clue about the use of the data factor in terms of data flow. Given the complex data types and transaction modes and the lack of corresponding data in the national accounting system, this section will try to estimate the data factor's volume by different standards in terms of the generation and storage of data and the flow and transaction of data based on various sources of data and identify the structural characteristics of the data factor's volume.

4.1 Data Generation & Storage Volume

New-generation IT applications have sharply reduced the cost of data creation and collection with a marginal cost close to zero and led to an explosive growth of data resources. According to the International Data Corporation (IDC) and some other institutions, global data generation totaled some 1,000 to 2,000 petabytes (PB) in 2000, which increased over a thousand folds to reach 2 zettabytes (ZB) by 2010. Rapid growth in data generation was accompanied by - but still far outpaced - a jump in annual storage capacity. According to *The Digitization of the World from Edge to Core*, a research report jointly published by the IDC and Seagate, global data capacity shipments reached some 2ZB in 2020 - a mere 3.4% of the 64ZB of data generated globally in the same year (see Table 1).

While the annual growth of data storage volume is hard to measure, global data storage capacity - including hard-disk drives (HDDs) and solid-state drives (SSDs) - may still lend credence to the IDC's above-mentioned assessment. Amid the explosive growth of data generation over recent years, SSDs have made up a steadily growing market share and storage capacity thanks to their faster read and write speeds. Yet it takes time for new technology to supplant the existing stock of old technology. Besides, HDDs still boast absolute advantages in terms of price and storage capacity. As a result, HDD shipments still account for over 60% of global data storage capacity shipment (Reinsel et al., 2017).

Seagate, Western Digital and Toshiba represent a lion's share of global HDD supply. According to

Table 1: Global Data Volume and Storage Capacity Shipment (in ZB)

Year	Global total data creation	Global data storage capacity shipment	Share	Global shipment of HDD	Estimated global data storage capacity shipment
2010	2	0.5	25.3%		
2011	5	0.5	10.9%		
2012	6.5	0.6	9.6%		
2013	9	0.7	7.8%		
2014	12.5	0.8	6.3%		
2015	15.5	0.8	5.5%	0.5	0.7
2016	18	1.0	5.6%	0.5	0.8
2017	26	1.2	4.5%	0.6	0.9
2018	33	1.5	4.6%	0.7	1.1
2019	41	1.8	4.3%	0.8	1.2
2020	64.2	2.2	3.4%	1.0	1.6
2021	79	2.6	3.2%		

Source: (i) Data in Column 2 is referenced from Statista: Global Creation, Use and Storage of Data, 2010-2025; (ii) Data in Column 3 is referenced from the IDC and Seagate: *The Digitization of the World from Edge to Core*, 2019; (iii) Data in Column 5 is referenced from the annual financial reports of Seagate and Western Digital; (iv) Data in Column 6 is referenced from Column 5 and IDC: *Data Age 2025*, 2017.

Trendfocus, a market intelligence provider, the above-mentioned three HDD suppliers shipped 479.99 exabytes (EB), 422.23EB and 116.1EB in HDD capacity in 2020, respectively. Based on corporate financial statement data and the above-mentioned 60% ratio, we may arrive at a rough estimate of the global increase of storage capacity that indirectly reflects the annual growth of data stock. Our estimate tallies with the IDC's data in terms of both trend and volume.

Notably, the annual increase of storage capacity may not coincide with the increase of data storage in the same year. For one thing, the utilization of new storage capacity cannot reach 100%. For another, a certain proportion of data stock is deleted each year to make room for new data. As such, the proportion of data created each year that ends up in storage will be somewhat different from the proportion listed in Column 4. Given the cost to purchase and maintain storage capacity, the total shipment of storage capacity plus additional storage space made available by removing a small amount of existing data should not have too much redundancy than the actual increase of newly stored data each year. Assuming the redundancy and removal ratios to be stable, the shipment of data storage capacity should roughly reflect the size of additional data storage.

The report entitled *Data Storage Power Is the Digital Cornerstone of High-Quality Economic and Social Development* jointly released by Huawei and Roland Berger in 2022 puts forth the concept of data storage power and measures the adequacy of data storage capacity by the ratio of data storage capacity to the total data generation in a certain region and year. The report measures the computing power of various countries and ranks the US first in the world with an adequacy of data storage capacity as much as 19.4% of world total while China ranks 11th place with a much smaller data storage capacity of 8.9%. China and the US are both major generators of data and rank the top two in the world in terms of annual data generation volumes. Yet China lags far behind the US in terms of data storage capacity mainly due to its insufficient data storage investment in the past, which has increased over recent years to make up for the shortfall. In 2017-2019, China's data storage investment grew by an annual average of 45.4% (see Table 2).

According to the *Digital China Development Report (2021)*, China's annual data generation

Table 2: Top Ten Countries and China in Terms of Data Storage Adequacy in 2020 (in %)

Ranking of data storage adequacy	Country	Data storage adequacy	Growth rate of data storage investment (2017-2019)
1	US	19.4	3
2	Singapore	18.8	10.7
3	Germany	18.4	20.4
4	Sweden	17.9	21.1
5	UK	15.5	18.5
6	Canada	14.7	-0.4
7	South Africa	13.5	22.3
8	Japan	11.8	17.8
9	France	10	28.1
10	Czech Republic	9.8	18.8
11	China	8.9	45.4

Source: Huawei and Roland Berger: *Data Storage Power Is the Digital Cornerstone of High-Quality Economic and Social Development*, 2022.
Note: The adequacy of data power capacity is the capacity of data storage devices in the current year as a share of the total amount of data generated by the region.

grew from 2.3ZB to 6.6ZB, up 30.2% on an annual average basis.[3] Based on the adequacy of data storage capacity from Huawei's report, China's data storage capacity reached some 0.59ZB in 2020, and with a certain redundancy taken into account, the total storage of data resources in the same year should be about 0.5ZB. According to the latest estimate by the China Academy of Information and Communications Technology (CAICT), China's total storage capacity reached 800EB by the end of 2021. According to the *National Data Resources Survey Report (2020)*, China's aggregate data storage reached 332EB by the end of 2019.[4]

Those two figures have verified the above result of estimation. If the increase of data storage accounts for 40%[5], China's data storage should have increased by about 0.2ZB in 2020, or 3.9% of the 5.1ZB of newly created data in the same year. This proportion is roughly consistent with the global share estimated according to the storage capacity shipment as shown in Table 1.

4.2 Data Transaction Volume

As can be learned from the previous section, the flow and transaction of data occur in various modes, each corresponding to a different statistical scope, which makes it hard to separately estimate the flow and transaction of data under each mode. Except for open access to public data, however, most data flow is realized via transaction, a significant portion of which is paid in monetary consideration. The annual volume of data transaction may also reflect the scale of data flow and transaction.

In the existing sectoral statistics, big data market is defined as the data industry market in the broad sense, which encompasses big data hardware, big data analysis software and big data professional services. Among them, big data professional services are primarily data transaction services similar in scope to the data transaction market mentioned in this paper. Despite differences in statistical scope and data markets, most statistics from various global institutions are estimated based on data-related transactions and services.

Statista and Wikibon estimated and forecasted the global data market revenues in 2020 (see Table 3). The two institutions have arrived at roughly consistent estimates of the global big data market. Additionally, Wikibon estimated and forecasted the value of professional services. According to their definitions, the "big data market" refers to the big data industry market including big data hardware, software and big data professional services. Big data professional services are primarily data transaction services, which should be consistent with the amount of data transaction service revenues. Global data transaction service revenues in 2021 are projected to reach some 22 billion US dollars. Data transaction service revenues as a share of the big data market may offer a clue about the transaction volume of the data factor. Assuming the share of data transaction service revenues to be 20% to 25% of the total transaction volume, it can be learned that the total transaction volume of the data factor is in the range between 88 billion and 110 billion US dollars.

In 2022, the European Data Market Monitoring Tool report led by the IDC and the Lisbon Council provided an estimate and forecast of data market transaction volumes of major economies. The report defines "data market" as a "market in which data is exchanged as a product or service" (IDC and Lisbon Council, 2022), whose scope covers information and IT services, research, commercial activities and big data analysis services related to big data. This scope is rather different from Statistica and Wikibon's definition of "data market." In estimating the size of data market, the IDC and the Lisbon Council's report has referenced the total incomes of relevant data firms in their respective regions, as well as the value of data products and services imported from elsewhere, which to some extent reflects the transaction volume of the data factor, including data products and services.

[3] The State Internet Information Office. *Digital China Development Report (2021)*.
[4] CAICT and Chinese Academy of Cyberspace Studies (CACS). *National Data Resources Survey Report (2020)*.
[5] According to the IDC's estimate, some 90% of global data stock was created over the past three years with an annual growth rate of about 40%.

Table 3: Estimation and Forecast of Global Big Data Market Revenues in 2011-2027 (in billion USD, %)

Year	Statista 2020	Wikibon		
2011	7.6	Aggregate amount	Big data professional services	Share of big data professional services
2012	12.25	-	-	-
2013	19.6	-	-	-
2014	18.3	18.3	7.6	41.53
2015	22.6	22.6	9.1	40.27
2016	28	27.3	11.1	40.66
2017	35	33.5	13.4	40.00
2018*	42	40.8	15.8	38.73
2019*	49	49	18.2	37.14
2020*	56	57.3	20.3	35.43
2021*	64	65.2	22.0	33.74
2022*	70	72.4	23.3	32.18
2023*	77	78.7	24.3	30.88
2024*	84	84	25.1	29.88
2025*	90	88.5	25.8	29.15
2026*	96	92.2	26.3	28.52
2027*	103	-	-	-

Source: (i) Statista: Forecast Revenue big data market worldwide 2011-2027; (ii) Wikibon: 2016-2026 Worldwide Big Data Market Forecast.

Estimated results for various years from 2019 to 2021 released by the report are shown in Table 4. In 2021, US data transaction volume stood at 239.958 billion euros, which accounts for some 64% of the top four markets - far above the data transaction volume of 63.627 billion euros for the 27 EU member states combined, the second largest market, and eight times that of China. In the same year, China's data market transaction volume was only 31.651 billion euros, which was below Japan's 39.97 billion euros and came fourth. The aggregate data transaction volume of the US, the EU, Japan and China amounted to 375.2 billion euros, based on which it can be estimated that the aggregate volume of global data market transactions should be around 400 billion euros.

Notably, the scope of the above data market transaction volume includes both big data-related software and hardware services and commercial activities such as data monetization. The IDC and the Lisbon Council's report also disclosed the incomes of data firms in the EU from data monetization, which stood at 11.61 billion euros in 2020, accounting for 90% of the EU's overall data market. Following this ratio, global data-related software and hardware services, R&D and big data analysis, among others, should be some 300 billion euros, and the income of data firms from data monetization, or the volume of global data transactions, should reach some 80 billion euros, including 6 billion euros from data transactions in China.

During 2019-2021, China, the US and Japan have all maintained rapid growth momentum in their data markets with annual nominal growth rates averaging 14.3%, 13.9% and 10.2%, respectively. In contrast, the EU's growth rate stood at a mere 4.35%, which is significantly below those of the other three major economies. The rapid growth rates suggest that the global data market remains in the early stage of rapid growth.

Table 4: Value of the Global Top Four Data Markets in 2019-2021 (in billion euros, %)

Country	2019	2020		2021*	
	Market value	Market value	Growth rate	Market value	Growth rate
US	184.87	213.46	15.46	239.96	12.41
27 EU member states combined	58.43	60.64	3.78	63.63	4.93
Japan	32.93	36.65	11.30	39.97	9.06
China	24.23	27.47	13.40	31.65	15.22

Source: IDC and Lisbon Council, 2022.
Note: * is the forecast value.

OnAudience, an international data service provider, traced and estimated the volume of online marketing data transactions (mainly digital advertising) in 27 countries around the world. Table 5 show the transaction volumes of the global and top three online marketing data markets in 2017-2021, which reveals that global online marketing data transactions reached 52.3 billion US dollars in 2021. Among them, the US data market was worth 30.6 billion US dollars, ranking first in the world, or 58.51% of the global data market. China's data market ranked third, but was worth a mere 7.3 billion US dollars, or 13.96% of the global data market, far eclipsed by those of the US.

Notably, the estimates of data transaction volume based on available information are highly inconsistent due to differences in information sources and scope of coverage and cannot fully reflect the overall volume of data flow and transaction. For one thing, most public data contains an important part of data flow and transaction despite the non-transactional nature of their flow, sharing and platform-based exchange. For another, existing statistics and publicly accessible corporate financial data are far from enough to cover monetized data transactions. In most cases, it is hard to find public information about point-to-point data transactions between firms. Statistics about the data market, therefore, are chiefly related to data transactions over public platforms, the value of which needs to be indirectly estimated based on incomes from platform-based transaction services.

4.3 Structural Characteristics of the Data Factor

In this section, we estimate the data factor's size in terms of the volume of data resources and the monetary value of data transactions based on open information from various sources under the dual perspectives of the generation and storage of data and the flow and transaction of data. Despite great differences in the estimated results due to various sources and scope of information, we may still identify some structural characteristics regarding the size of the data factor.

First, data storage has increased at a much faster pace compared with data creation. As a result, only around 4% of data created ends up stored. Data storage capacity is low for most major economies. It is less than 20% for the US, which ranks first, and less than 10% for China.

Second, the imbalance between data creation and data storage is even more striking. Data creation in China has increased sharply over the years thanks to its ultra-large market and diverse use cases, ranking top two together with the US, but only 3% or so of such data ended up stored, which is significantly below world average. In terms of the adequacy of data storage power, China ranks only 11th in the world, which is far behind the US that ranks first.

Third, the data market remains in the early stage of rapid development with the volume of data transactions eclipsed by the heft of the digital economy. Due to various reasons such as the scope of coverage, statistical criteria and access to information, there is no universally recognized estimate of global data flow and transaction, but different estimates generally corroborate each other on the order of magnitude. Based on the estimates published by various institutions, the volume of global data

Table 5: Transaction Volumes and Shares of the Global and Top Three Online Marketing Data Markets, 2017-2021 (in billion USD, %)

	2017		2018		2019		2020		2021	
Country / region	Value	Share	Value	Share	Value	Share	Value	Share	Value	Share
Global	18.9	100	26.5	100	34.6	100	41.4	100	52.3	100
US	12.3	65.08	16.6	62.64	21.2	61.27	24.7	59.66	30.6	58.51
Europe	2.8	14.81	4.1	15.47	5.3	15.32	6.3	15.22	7.6	14.53
China	1.7	8.99	2.8	10.57	4.1	11.85	5.4	13.04	7.3	13.96

Source: OnAudience.com, "Global Data Market Size (2017-2021)".

transactions is on the order of 100 billion US dollars, which is two orders of magnitude below the value-added of the global digital economy. Such a discrepancy stems from not only statistical omissions, but more importantly, the nascency of the data market.

Fourth, China's data market is less developed compared with countries like the US and Japan. China accounts for around 10% of the global data market transactions and online marketing data transactions, ranking fourth also after the EU and Japan. Compared with the US, the EU, and Japan, China's data market transaction volume has maintained the highest growth rate. Indeed, China's share of the data transaction volume could be substantially underestimated given the robust demand for data transactions brought about by diverse use cases.

5. Concluding Comments and Policy Recommendations

The public has shown great interest in the data factor and data transaction, but attention is overly focused on behavioral data and data exchange centers. Based on an analysis of the basic concepts of data in the preceding sections, this paper distinguishes the connotations of data in bit/digital data, data resources and the data factor, and classifies data from various perspectives. Based on the literature review and practical experience, we have estimated the scale of data resources and data transaction volumes for major economies. The goal is to reveal the status and trends of the flow and transaction of the data factor. Specifically, the following conclusions can be made:

First, diverse use cases have caused data classifications to be diverse. Various types of data involve heterogenous entities, ownerships, information densities and other attributes, which have in turn led to complexity in the flow, transaction and profit distribution of data. Based on such diversity and complexity, the government should take steps to develop sound regulatory systems for data classification, flow, and transaction.

Second, data is exchanged and traded increasingly as a product or service over internet platforms. For data users, raw data cannot address the needs of certain use cases. By offering customized data products or services, data suppliers not only spare their clients the drudgery of making sense of piles of data, but create value addition and more importantly, eliminate the risk of privacy and information divulgence. Platform-based data transaction boosts efficiency, which is a key merit of the digital economy.

Third, the existing statistical accounting system is a product of the industrial economy era, which does not and cannot carve out a separate statistical category for data resources as a factor of production. Significant differences exist in the estimates of the size of the data factor due to the source of information, statistical scope and methods of estimation. Sizing up the data factor is both a fundamental task for improving the flow and transaction of the data factor and a frontier subject of research on the digital economy, especially the measurement of the digital economy.

Fourth, only around 4% of global data generation ends up in storage. Disequilibrium between data creation and storage exists in all major economies. While such disequilibrium is less severe in the US, the structural disequilibrium between data creation and storage is particularly striking in China. While China and the US far outstrip other economies in terms of their data resources and data creation capabilities, China's adequacy of data storage power ranks only 11th in the world, which is far behind the US. China has plenty of room to improve its data storage capacity.

Fifth, the global data transaction market remains in a nascent stage of rapid development with a public transaction volume reached the order of hundreds of billions of dollars, representing two orders of magnitude different from the value added of the digital economy. Among major economies, China's data factor market is even more underdeveloped, accounting for less than 10% of the world total, far smaller than not only the United States, but also lower than the European Union and Japan.

First, improve institutional development such as data classification and ownership identification to promote the flow and transaction of data. In accordance with the *Guidance on the Classification of Industrial Data (for Trial Implementation)* released by the Ministry of Industry and Information Technology (MIIT), industry authorities should enact data classification standards or guidelines for key sectors and use cases to improve the data classification system. Based on data classification and authorized access, the boundary of the right to possess, use and operate data should be demarcated to provide legal and institutional assurances for the legitimate rights and interests of participants and stakeholders in various processes of data exchange.

Second, strengthen digital infrastructure such as data storage and computing centers to increase the accuracy of data storage capacity and computing power and ensure the collection and storage of data resources. Based on national strategies such as "developing data centers and cloud computing centers in the western region to meet the needs of the eastern region," the government should refine development planning for cloud computing and data storage centers. In addition to the existing eight national nodes, consideration may be given to developing more data storage centers in the western region with a favorable climate, cheap electricity and other advantages.

Third, foster a group of professional institutions for the transaction, development and utilization of the data factor to give full play to the role and technological strength of enterprises as market entities. We should leverage the technological prowess of data firms in promoting the market-based allocation and value creation of the data factor. We should encourage firms - especially tech giants - to become more involved in data transaction, and contribute to the growth of the data market by making use of data resources. Companies with a knack for data resources, digital technologies and use cases should be utilized to provide the government and various market entities with professional products and services that improve the quality of data market transactions and services.

Fourth, improve data governance to balance security with development. Data security and healthy development of relevant sectors require all stakeholders to enhance the protection of critical data and perform risk early warning, security assessment and *ex-post* tracing. Market regulators and judicial authorities should supervise data governance, and prevent problems like data monopoly and price discrimination through appropriate judicial and law enforcement while encouraging the innovation and development of emerging business modes. In addition, the government should beef up security review and multilateral cooperation on the cross-border flow of data to secure the flow and transaction of data on a broader scale.

Fifth, catalyze digital technology innovations. Secure, orderly and sufficient flow of data cannot occur without institutional assurance, policy support, and more importantly, critical technologies. Concerning data security, priority should be given to data traceability, secure storage and data availability but invisibility. It takes secure multi-party computing, blockchain and other sophisticated technologies to protect national data security and prevent the divulgence of privacy and business secrets from the exchange of data. On the other hand, the development and utilization of

data resources require sufficient data storage capacity and computing power. With SSD as the dominant medium of storage, the problem is that its core component flash memory is monopolized by foreign oligarchs. Such choke-point technology much be obtained in order for the data factor to better support economic development.

References:

[1] Acquisti, A., Taylor C., and Wagman L. 2016. "The Economics of Privacy." *Journal of Economic Literature*, 54(2): 442-92.

[2] Bergemann, D., and Bonatti A. 2019. "Markets for Information: An Introduction." *Annual Review of Economics*, 11: 85-107.

[3] Cai, Yuezhou, and Wenjun Ma. 2021. "Data Factor's Effects on High-quality Development and Constraint of Data Flow." *Journal of Quantitative & Technical Economics*, No. 3.

[4] CAICT and Chinese Academy of Cyberspace Studies (CACS). 2021. *National Data Resources Survey Report (2020)*.

[5] CAICT. 2022. *White Paper on China's Data Storage Power (2022)*.

[6] Chang, Jiang, and Zhen Zhang. 2022. "Authorized Operation of Public Data: Characteristics, Nature and Regulation." *Research on Rule of Law*, No.2.

[7] European Commission, Directorate-General for Communications Networks, Content and Technology, Scaria, E., Berghmans, A., Pont, M., et al. 2018. *Study on Data Sharing between Companies in Europe: Final Report*, Publications Office.

[8] Farboodi, M., and Veldkamp L. 2021. "A Growth Model of the Data Economy." *NBER Working Paper*, No.28427.

[9] Federal Trade Commission. 2014. "Data Brokers: A Call for Transparency and Accountability."

[10] IDC and the Lisbon Council. 2022. *European Data Market Study 2021-2023 (D2.1 First Report on Facts and Figures)*. Luxembourg: Publications Office of the European Union.

[11] Laudon, K. C. 1996. "Markets and Privacy." *Communications of the ACM*, 39(9): 92–104.

[12] Malgieri, G., and Custers B. 2018. "Pricing Privacy - The Right to Know the Value of Your Data." *Computer Law & Security Review*, 34(2): 289-303.

[13] Mitchell, John, Daniel Ker and Molly Lesher. 2021. "Measuring the Economic Value of Data." OECD Going Digital Toolkit Note, No.20.

[14] Nguyen, D. and M. Paczos. 2020. "Measuring the Economic Value of Data and Cross-border Data Flows: A Business Perspective." *OECD Digital Economy Papers*, No. 297.

[15] OECD. 2013a. "Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value." OECD Digital Economy Papers, No. 220.

[16] OECD. 2013b. "Introduction to Data and Analytics (Module 1): Taxonomy, Data Governance Issues and Implications for Further Work." DSTI/ICCP (2013)13.

[17] OECD. 2021. *Issues Paper: Recording Observable Phenomena and Data in the National Accounts*. Paris: OECD Publishing.

[18] Reinsel, D., J. Gantz, and J. Rydning. 2019. *Data Age 2025*. IDC.

[19] Statistics Canada. 2019. "The Value of Data in Canada: Experimental Estimates." *Latest Developments in the Canada Economic Accounts* (Working Paper Series), No.9.

[20] Swedish National Board of Trade. 2015. *No Transfer, No Production: The Importance of Cross–border Data Transfers for Companies Based in Sweden*. Stockholm: Swedish National Board of Trade.

[21] The State Internet Information Office. 2022. *Digital China Development Report (2021)*.

[22] Tian, Jietang,and Luyao Liu. 2020. “Transaction Mode, Definition of Rights and the Fostering of a Data Market.” *Reform*, No.7.

[23] United Nations Conference on Trade and Development(UNCTAD). 2021. *Digital Economy Report 2021 Cross–border Data Flows and Development: For Whom the Data Flow*. Geneva: United Nations Publications.

[24] Varian, H. 2018. “Artificial Intelligence, Economics, and Industrial Organization.” *NBER Working Paper*, No. 24839.

[25] Veldkamp, L., and C. Chung. 2019. “Data and the Aggregate Economy.” *Journal of Economic Literature*, forthcoming.