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.