Journal of Big Data

RESEARCH Open Access

Exploring the form of big data products and the supporting systems

Yazhen Ye[1,2*], Yao Zhang[1] and Yangyong Zhu[1,2]

*Correspondence:
yeyazhen@fudan.edu.cn
[1] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China
Full list of author information is available at the end of the article

Abstract

There have been studies and practices about forms and supporting systems of single-type data products. However, little literature is dedicated to these factors of big data products, which are huge in volume and contain a variety of different types of data. For example, we do not know the form of data composed of a relation table and a digital video file at the same time. The lack of certain forms of big data products makes it impossible to design its base unit for measuring. It also causes difficulties in pricing and valuation of big data products, and further causes issues of circulation and supporting services. In this paper, we analyze the challenges of productizing big data. By referring to information media such as books, a product form and supporting systems of big data products based on Data Box are proposed. Different from previous studies, we focus on the system-level design and systematically study the pipeline including containerization, circulation, pricing, protection and etc. The proposed Data Box-based big data products and supporting systems can have a positive influence on the research, exploitation, exchange, and circulation of big data products.

Keywords: Big data products, Data circulation, Data Box

Introduction

There have been many studies and applications about the form of single-type data products like digital songs, videos, images, eBooks, webnovels, etc. These data products all have basic characteristics like unified formats and volumes, complete contents, copyright identifiers, and unique IDs. Supporting systems for such data products usually contain the following basic elements: a data platform, a user interface, pricing rules, copyright protection management mechanisms, as well as data infrastructure. There are three prominent properties of these kinds of data products: (1) a data product only contains data of the same type; (2) the volume of a data product is small; (3) they can be transferred through the internet easily.

However, big data products show contrary properties: (1) a big data product is usually a combination of data of multiple types; (2) the volume of a big data product is relatively large; and (3) as a natural consequence, it is hard to transfer big data products purely through the internet. Therefore, the form and supporting systems of big data products need to be specifically studied and designed, since they are different from those of single-type data products.

Springer Open

Inspired by the form and properties of traditional media like books, magazines, and newspapers, in this paper, we design a novel form of big data products based on Data Box [1], which meets the necessities of data circulation, including unified format, certain scale, and valid pricing unit. In addition, we also design a supporting system of big data products, which consists of a big data product platform, a user interface, pricing mechanisms and strategies, as well as other critical contents. The issues of data publishing are also discussed. As the proposed pricing unit of big data products, Data Box is expected to have a positive impact on the trading, circulation, exploitation of big data products, and push related studies forward.

The rest of the paper is organized as follows. In section "The challenges of big data products", we present the design challenges of big data products and supporting systems. In section "The inspiration from books", we introduce the inspiration from books and how it can help us design big data products. In section "The form of big data products", we introduce the form of big data product, i.e., Data Box. In section "The supporting system of big data products", a supporting system of big data products based on Data Box is designed and implemented, and the data publishing is also discussed. At last, we conclude this work in section "Conclusion".

Related work

There have been many studies on productizing big data. For example, different types of data containers have been proposed [1–3]. Some researchers have studied data pricing from aspects of contents of data [4], demand of market [5], data quality [6], data measure [7], etc. Other related topics like data circulation [8] and openness of data [9] have also attracted much attention. However, the above studies only focus on some single aspect, and overlook the system-level designs of big data products. In this paper we discuss the form of big data products and propose the supporting systems that integrate the whole pipeline including containerization, circulation, pricing, protection and etc.

The challenges of big data products

The characteristics of single-type data products are single type, small scale of data in unit product and convenient transfer through internet. Therefore, data products of some typical industries operating and circulating in the market are mainly single-type ones. Most of these data products have their representative commercial platforms or service agencies. According to the characteristics of data products, the circulation process generally requires a two-step authorization [10], i.e., content authorization or platform authorization in the first step, and consumption authorization or terminal authorization in the second step. And then end-users purchase data products or consume data services through the network. There have been some studies of forms and supporting systems of data products such as digital songs, videos, images, eBooks, webnovels, etc. Some researchers have proposed the form of single-type data products and the architecture of their supporting platforms [10].

In contrast to single-type data products discussed above, big data products are usually made up of multiple types of data and are huge in volume [11, 12]. The analysis of a big data product needs to be conducted by treating it as an organic whole, so that we can reap valuable knowledge from it for decision-making. "Big data products" in this paper refers to legal datasets that have independent and complete meaning and contain multiple types of data with a certain scale.

The properties of big data products

According to the definition, the properties of a big data product are that it contains various types of data and is huge in scale [13]:

- **The variety of data types** A big data product may consist of multiple types of data [11, 14]. For example, a big data product of electronic medical records contains medical data of multiple formats, e.g., texts, fMRI data, CT images, as well as records of doctors and prescriptions, which is very complex [15].
- **Large scale of unit product** A big data product usually contains Terabytes (TB) of data or even more [16]. A big data product of electronic medical records must contain tens of thousands of individual records to satisfy the need for big data research on a particular disease [17]. In other words, the scale of big data products is much larger than that of single-type data products.

The design challenges

The above-mentioned properties of big data products cause the great challenges of designing the form of big data products and supporting systems.

- **Difficulties in designing the form of big data products** It is difficult to design big data products with a unified format and scale [11]. Variety of data types makes it hard to define the form of big data products in aspects of unified scale, unified format, and content integrity. In the case of unified scale, since big data products are made up of multiple types and formats of data, the scale of said products is decided by all these contents combined. Therefore, it is difficult to set up a base unit with fixed volume. In the case of unified format, since big data products usually contain at least two different types of data encoded in different standards, it is difficult to design a unified data format. In the case of content integrity, since big data products are a combination of various types of data, it is difficult to decide whether a piece of content is the smallest integrity and at the same time cannot be divided again.
- **Difficulty in internet transfer** It is hard to transfer Terabytes level data through the internet, which may take at least hours, sometimes weeks or months. For this reason, big data products cannot be downloaded or streamed online without effort [18]. Conventional networks cannot satisfy the need of transferring big data products, and dedicated technology needs to be developed to fulfill this task [19].
- **Difficulties in designing the supporting system** The challenges above cause difficulties in designing the big data products supporting system, including data platform, user interfaces, pricing rules, copyright management protection mechanisms, infrastructures, and so on [20].

  - *Data platform* A dedicated institutional platform that can handle a significant scale of big data products is necessary.

- *User interfaces* Common user interfaces are not capable of handling big data products. Therefore, a development and operation environment must be developed to ensure that big data products are readable, visible, and programmable.
- *Pricing rules* Since there is no proper pricing unit for big data products, it is difficult to conclude a universal pricing method.
- *Copyright management protection mechanisms* Due to the huge scale, various types of data, and multiple data sources of big data products, designing copyright management and protection mechanisms faces greater difficulties. There is a need for designing unified copyright and access identifiers.

The inspiration from books

Obviously, it is critical to make a breakthrough in the form of big data products in order to address the above challenges. The key question is how to design a form of big data products which involves a relatively unified format and scale, or in other words, how to design a big data product for a specific type of big data applications that meets the requirements of unified format and scale.

Books share some similarities with big data products from the aspect of containing multiple forms of data, such as words, symbols, images, photographs, charts, and so on. It can be said that the form of books solves the problems of unified format, unified scale and, content integrity when dealing with multiple types of data. In the following, we discuss the form of the book as information media and see if we can get some inspiration from books.

Definitions of books

The UNESCO[1] defines books as follows [21]: "A book is a non-periodical printed publication of at least 49 pages, exclusive of the cover pages, published in the country and made available to the public." ISO[2] defines books as "non-serial printed document in codex form". SAPPRFT[34] defines books as "regular books, maps, celebration artworks, images, picture books, calendars contain words and images, as well as other media that are recognized by the administration." In order to better identify book product around the globe, an International Standard Book Number (ISBN)[5] code is assigned to every single non-journal book publication [22], which in return promotes book commerce worldwide [23].

Page count as the core of books form

According to the definitions, we discovered that books bind many types of data together on a paper basis. As a product that carries data, the basic elements of books are as follows:

[1] http://portal.unesco.org/en/ev.php-URL_ID=13068&URL_DO=DO_TOPIC&URL_SECTION=201.html.

[2] https://www.iso.org/obp/ui/#iso:std:iso:2789:ed-5:v1:en.

[3] http://www.gov.cn/gongbao/content/2009/content_1399850.htm.

[4] http://lawinfochina.com/display.aspx?id=e0159be048a21ad0bdfb&lib=law.

[5] https://www.isbn-international.org/content/isbn-users-manual.

- Certain scale: A book usually contains more than 49 pages bound together except for front and back covers.
- Content integrity: The content of a book (including one of a series of books) is an organic whole, with the smallest integrity that cannot be subdivided, and a further subdivided part cannot be called a book.
- Pricing unit: The pricing unit of books is *copies*.
- Copyright ID: Published books have their own unique ID codes, e.g., ISBNs.

Pricing methods regardless of content value

Many countries, such as France, Germany, and Italy, choose fixed book price as their pricing policy [24]. Other countries, such as the United Kingdom, Finland, and Iceland, choose free prices, where publishing houses and book retailers are able to freely decide the price of books by considering the number of pages and copies [24, 25]. In China, publishing houses decide the price of a book either by the cost of printed pages, the overall cost or the expected profit margin [26].

In general, the pricing methods of books do not take content value into consideration. This can be demonstrated in the following two aspects.

- Pricing unit: Books are priced and sold in copies, which refers to content integrity, i.e., the smallest unit able to circulate and accepted by readers independently. In this way, "copies" becomes a practical unit for multiple-type data.
- Pricing mechanism: When deciding the price of a book, production cost, which is often indicated by the number of pages, is the most important factor rather than the value of content.

Effective binding of multiple data types through books

As paper-based products, books bind multiple types of data, such as words, charts, and images, together. Compared to a single image or a page of words, a book is much larger (at least 49 times larger) in volume. This can be regarded as an effective way of binding large-scale data into a dataset with a unified format. Details are as follows:

- **Various data types** Books are a kind of information medium product that binds various data types together. A page in a book may consist of different data types such as words, symbols, images, charts, photographs, and so on.
- **Relatively large in scale** Books are larger in scale than a piece of printed paper since a book must reach a minimum of 49 pages. A book may contain hundreds, even thousands of pages. In addition, books that share a common topic can be arranged and published in the form of book series.

Discussion and inspiration

By analyzing the form and the properties of books, we can conclude that a big data product, as a viable product, needs a certain scale, content integrity, a pricing unit, and a copyright ID, among which the scale is the most crucial factor. The minimum scale must be defined for big data products just like books, even though the minimum scale (49 pages) might not have been decided based on specific reasons. Therefore, for big data products, it might be a reasonable choice to define big data products based on the volume of data or the number of bytes.

The way of book pricing is developed naturally, and there is no connection between the value of content and the book prices, which might be a result of the difficulty in evaluating content value. The only practical method is to set a price based on the number of printed pages. Similarly, it is difficult to measure the value of the content of a big data product. In fact, the value of a big data product varies from person to person, and there is no widely agreed scheme of data value. By referring to the pricing mechanisms of books, it is possible to consider pricing big data products based on the volume of data rather than the content of data.

Books bind multiple types of data together by printing them on paper. When it comes to big data products, it is crucial to design a new data format to bind existing various types of data.

The form of big data products

The form of single-type data products and the supporting system is suitable for data products that are small-scale or can be streamed. Books, as a form of recording and disseminating information, have existed for hundreds of years. The information recorded through the form of books is various and involves a wide range of fields. It's very similar to big data products in some aspects, and so the design of books can be used for reference when designing the form of big data products.

The survey on the form of big data products

We conducted a survey on the form of big data products by delivering questionnaires to 110 data scientists who attended "The 3rd Data Scientist Conference"[6] at Deqing, Zhejiang Province, China. The questions in the questionnaire are listed in Appendix. The main conclusions are as follows:

- **The single-type data products are inappropriate to be considered as the big data products** 69.1% of the respondents did not agree that "a digital music file, a digital image file, a digital movie file or a webnovel file is a big data product". On the other hand, there was a contradiction about whether "1000 digital music files, 1000 digital image files, 1000 digital movie files or 1000 webnovel files are a big data product", where 55.5% of the respondents disagreed and 42.7% agreed.
- **Multi-type datasets with an acceptable scale can be recognized as big data products** 41.1% of the respondents agreed that "a digital music file, a digital image file, a digital movie file and a webnovel file combined can be considered as a big data product", while 55.1% disagreed. On the other hand, 68.2% of the respondents considered "1000 digital music files, 1000 digital image files, 1000 digital movie files and 1000 webnovel files combined" as a big data product.

[6] http://dsc.xintongconference.com/Page.

- **The scale of a big data product should at least reach Terabyte** On the issue of the minimum scale of a big data product, 14.7% of the respondents chose gigabyte level, while 55.6% chose terabyte level and 26.6% chose petabyte. It can be said that choosing terabytes as minimum scale of a big data product is acceptable.
- **Intuitively, a dataset with multiple data types and large scale is a big data product** About a dataset consists of “a high-resolution digital movie (10 GB), 10 music CDs (7 GB), 1000 10-million-pixel digital images (3 GB), 1000 500-thousand-word digital books (about 1 GB) and a data table with a billion records (10 GB)”, 52.3% of the respondents agreed that it exceeds the minimum scale of a big data product, while 28.4% disagreed.

Elements of big data products

The key reason why single-type data products such as digital songs, videos, images, eBooks, webnovels, etc., can be traded well in the market is that they have their own pricing units. However, it is difficult to determine the pricing unit for big data products with multiple types and large data sets, and there is no suitable dimension at present. Books and big data products are very similar in terms of forms, values, attributes, etc., thus the pricing dimension of books can be used as a reference.

By referring to the form of books and single-type data products, we propose the basic elements of big data products as follows.

- Certain scale: According to the results of the survey, the scale of a big data product should reach the Terabyte level. But at the same time, a dataset of roughly 30 GB of data including “a high-resolution digital movie (10 GB), 10 music CDs (7 GB), 1000 10-million-pixel digital images (3 GB), 1000 500-thousand-word digital books and a data table with a billion records” can be recognized as a big data product intuitively. Although 49 pages as the criteria of a book is not numerically large, “a book” can still be a sensible unit. For the same reason, 30 GB[7] may be considered as the base limit of the scale of a big data product, where it is already possible to include a large amount of different single type data products.
- Content integrity: The content of a book (including one of a series of books) is an organic whole, which cannot be subdivided and a further subdivided part cannot be called a book. What is the content integrity of a big data product? Firstly, single-type data products included in a big data product must be complete. (A data product, by definition, must be complete.) Secondly, other contents must have meanings that can be interpreted independently. Lastly, all the contents in a big data product combined together must have meanings that can be interpreted independently.
- Pricing unit: The pricing unit of books is copies. It is possible to design a structure called Data Box using the 30 GB volume scale, and utilize Box as the pricing unit for big data products.

[7] Whether 30 GB is a suitable criterion remains to be proven in future practice. This paper emphasizes the necessity of setting up a volume scale for future measurement and valuation of big data products.

Data Box: a form of big data products

The proposed Data Box is a kind of container for big data [1, 9]. A Data Box is an organized data storage model with an automated program unit and built-in computing capability. After a Data Box is loaded with data by data owners, the data within can only be accessed through automated program APIs in a controlled manner [27]. The data in a Data Box is visible, perceivable, and programmable from outside, while controllable, trackable, and recoverable from inside. In other words, data within Data Box can be circulated as a product, while the rights of data owners are protected.

For the time being, the concept of Data Box does not take the scale of the container into consideration. Under the circumstance where the basic scale of a Data Box is defined, it can be utilized as the pricing unit of big data products satisfying the need of circulation, valuation, and exchange. Only through the standardization of the Data Box may data be qualified for registration, publication and entering the market. The registration of data supports the claim of data ownership and rights, therefore an institution of data registration is a necessity for the purpose of defending these claims. A data product, whether it is produced by individuals or corporations, must be registered so that the rights are legally protected and the product would be able to trade in the data market. Copyright protection for books is mainly guaranteed by legislation. Similarly, piracy concerns of registered big data products should be a legal issue and should not be confronted with technical requirements. The registered and published data products would be able to circulate publicly on the data market. Data products that are registered but not published would also be able to trade, while the data owner retains the decision of whether to publish the data.

In the future, data would be stored in different standard formats according to the field they belong to in the same way as MP3 being a standard format of music data products. For example, a listed company can produce a Data Box with its annual reports, introductions of its main business, quotations of its shares and personal introductions of its senior managers. The Data Box, which at the same time is a big data product, would be registered at the data registration institution, and thus its ownership as well as other rights would be confirmed, and it can circulate in the market. The owner of the product may decide whether to publish the financial big data product. To publish this product, it is necessary to register the product at a specific administration beforehand.

Based on the discussions above, we propose our design for big data products: a big data product is a dataset packed into a Data Box with a minimum volume of 30GB, that contains multiple types of data, whose contents can be interpreted independently.

The supporting system of big data products

Based on big data products with Data Box as the pricing unit, the supporting system of big data products is discussed in this section.

Composition of single-type data products supporting system

First, we review the main phases that make up the single-type data products supporting system.

1. The platform owns the copyright of data products or obtains the copyright authorization of data products sources through various legal means.
2. The platform then standardizes the data product sources, format them into data products, and store these data products into a data product library.
3. The platform designs pricing mechanisms and strategies for data products in the library.
4. The platform builds and puts forward platform websites or online stores, and presents users the information of data products in the form of user interfaces through infrastructures such as the internet.
5. Users may purchase authorization and use data products through applications of user interfaces such as dedicated readers or universal client devices.

Composition of big data products supporting system

Since big data products are large in scale, it is hardly possible to present data within the product to users through the internet. For this reason, supporting systems for single-type data products obviously cannot support big data products.

Supporting system design The design of supporting systems for big data products must address the problem of circulation through the internet.

Circulation through Internet Since big data products are difficult to be transferred through the internet, a possible solution to this problem is to allow developers to upload applications to the supporting platform. In this way, data contained in big data products can be used directly on the platform, increasing security to prevent data leaks at the same time.

Now we describe details of the system design.

Data station design The Data Station for Self-governing Openness of Data [9] satisfies needs for accepting user-uploaded data mining software, therefore is capable to technically support big data supporting platform. Based on this system, we design authorization and pricing mechanisms as follows.

Authorization mechanism A two-step authorization process [10] is adopted. At the first step, the data source authorizes the operating group of the platform; at the second step, the platform authorizes the use rights of big data products to the end-users.

Pricing mechanism Data Box is chosen as the pricing unit. Big data products are priced based on the scale of their Data Boxes.

Elements of big data product supporting system

To include a broader definition of big data products, their supporting system should technically include a big data product platform, user interfaces, pricing mechanisms and strategies, and so on.

- **Big data product platform** The platform refers to institutions and systems that provide maintenance as well as other related services to big data products. After a platform is authorized by the data source, it processes the raw data into data products, i.e., goods, that are fit for circulation, which would be available to authorized end

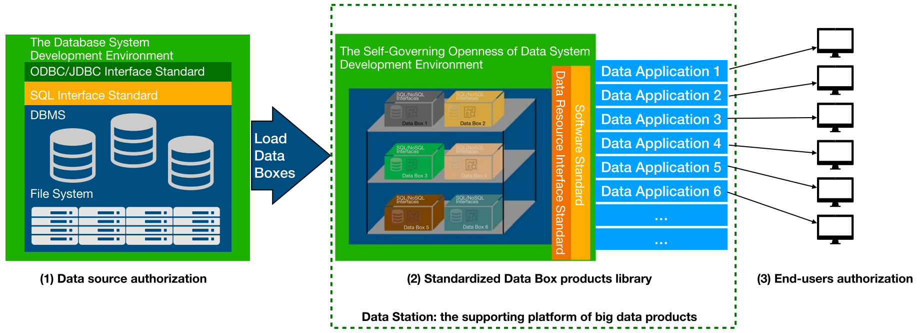
The Database System Development Environment
ODBC/JDBC Interface Standard
SQL Interface Standard
DBMS
File System
Load Data Boxes
The Self-Governing Openness of Data System Development Environment
SQL/NoSQL Interfaces
Data Box 1
Data Box 2
Data Box 3
Data Box 4
Data Box 5
Data Box 6
Data Resource Interface Standard
Software Standard
Data Application 1
Data Application 2
Data Application 3
Data Application 4
Data Application 5
Data Application 6
...
...
(1) Data source authorization
(2) Standardized Data Box products library
(3) End-users authorization
Data Station: the supporting platform of big data products

Fig. 1 The supporting system of big data products based on Data Box. Firstly, the platform gets authorization from data sources and load data into Data Boxes. Then the data station stores standardized Data Boxes and serves data applications with proper standards, which are finally provided to end-users through authorization

users. Such a platform is the core of big data supporting system, and therefore is a necessity.

- **User interfaces** In contrast to existing user interfaces for single-type data products such as players for digital songs or videos, the user interfaces for big data products aim to perform data mining and to extract value from data. Such user interfaces may not enable people to understand data directly, but instead provide access to data processing software. In this sense, user interfaces for big data products are a software development environment, which allows users to develop and upload data applications such that big data products can be utilized. User interfaces are the port linking users and big data products, therefore is a necessity.
- **Pricing mechanisms and strategies** The price of data usage is decided by the judgment of the data owner towards data markets. Pricing mechanisms and strategies directly affect the practicability of trading and circulation of big data products. From a technical perspective, a pricing unit for big data products is needed. One such pricing unit is Data Box, and there can also be other forms of pricing unit. Based on pricing units, pricing mechanisms and strategies can be formulated. As a platform to promote product circulation, pricing mechanisms and strategies are a necessity.

The supporting system of big data products based on Data Box is illustrated in Fig. 1. Firstly, following the two-step authorization process [10], the platform gets authorization from data sources. Then data are loaded into Data Boxes, where each box can be implemented with containerization techniques like Docker[8] [1] with certain data protection mechanisms [9]. Data Boxes are also equipped with standardized user interfaces including SQL, NoSQL, etc., with proper published ports. Then the data station stores standardized Data Boxes and provide services. End-users can access Data Boxes through the second phase of the authorization process [10], and upload their own applications. Each application can also be a Docker image with data mining tools installed. The application Docker containers can easily communicate with Data Boxes through networking. Since user applications are uploaded to the platform, it

[8] https://www.docker.com/.

is recommended to build the data station over clusters of servers where distributed computing environment like HDFS [28], Hive [29], Zookeeper [30] are deployed.

Legal environment for big data products

The main problem of data circulation in the legal context is the ownership of data. We may find that the circulation system of single-type data products resembles systems of books. Therefore, the confirmation of the ownership of most single-type data products is through the process of publication. For the same reason, it is natural to distribute data products through publication processes.

Regardless of other factors, all data products may claim their rights through publication, which requires the existence of data ownership registration institutions, copyright management and protection mechanism, data publication institutions, data publication standards and mechanisms. All these depend on the effort of the government and market entities to establish a market environment suitable for data circulation.

- **Data ownership registration institution** This is the organization that confirms intellectual property rights of data products and is the policymaker of data product standards and circulation mechanisms. Like ISBNs for books, an equivalent "International Standard Data Product Number"s may be introduced. Only registered big data products may confirm their ownership rights and enter the data market. The institution solves a series of difficulties, such as the difficulty in confirming the rights of productizing data resources due to the problem of data ownership, and the inability to trade due to authorization issues.
- **Copyright management and protection mechanism** Copyright of registered data products would be protected through the making of laws, regulations, policies, and the development of technical measures.
- **Data publication institutions** These are the publication management entities of big data products that own the copyright of products. Examples of these institutions are data journals, data centers, and publishing houses of big data products.
- **Data publication standards and mechanisms** We need to establish relevant standards to facilitate the circulation of big data products in a way that data rights are protected.

Data would be freely distributed after being published [31–33]. Big data products should also be freely distributed after the publication process, which would be beneficial for the promotion and circulation of said products. In order to provide better protection for the intellectual property rights of big data products, a dedicated publication standard and mechanism need to be established for big data products, which should be administrated by proper data publication institutions. However, in contrast to single-type data products, some kinds of big data products are not suitable to be published. Since publication leads to public availability, all institutions and individual can perform data mining and exploitation upon the big data products, which would have a negative effect on the rarity of the content of the said products and the premium of them, making it difficult to maximize the profit of content providers.

Conclusion

Data is the key element in the digital economy. Building a product system for big data is an urgent need if a digital economy is to be developed. The variety in types of big data products leads to the conflict in understandings about the form of these products. Vital problems of the ownership, accounting procedure, transaction and circulation pricing of big data products remain to be effectively resolved.

In this paper, the challenges of big data products are analyzed. By referring to information media such as books, a pricing unit of big data product, Data Box, is proposed. Unlike previous work, we systematically study the pipeline, including containerization, circulation, pricing, protection, etc., and design a pricing form and a supporting system of big data products based on Data Box. We use Data Box as an effective way of binding multiple types of data, which are then treated as the pricing unit. We tackle the difficulty in internet transfer of big data by allowing users to upload applications to the data station. We discuss the legal environment of big data products and suggest the necessary institutions and standards to be developed. The proposed Data Box-based solution can have a positive influence on the research, exploitation, exchange, and circulation of big data products.

The future directions may involve the following aspects:

- **Research the standardization of Data Box and its pricing mechanisms** Data Box is the minimum pricing and circulation unit of big data products on data markets, the standardization of which is the prerequisite and cornerstone for the mass production and standardized circulation of big data products. The minimum unit of big data products (which is Data Box) makes it more accurate and clear about the object of pricing research. Further study should focus on pricing mechanisms of big data products on the basis of Data Box.
- **Research accounting methods of big data products** Further study should include research about the categorization of big data products, whose minimum unit would be Data Box, and matching ways of processing in order to allow certain types of said products to be listed in fiscal statements.
- **Research the making of related laws and regulations** It is necessary to look into the making of laws and regulations concerning big data products, so that the legality of data products, which have been registered at data ownership registration institutions, would be protected. Moreover, the ownership of said products could be confirmed and the exchange regulations about big data products on data markets could be established.

Appendix

The questionnaire contains the following questions:

1. Should a piece of music, a photograph, a movie, or a web novel be considered as a big data product? (Yes/No/No idea)

2. Should 1000 pieces of music, 1000 photographs, 1000 movies or 1000 web novels be considered as a big data product? (Yes/No/No idea)
3. Should the bundle of a piece of music, a photograph, a movie, and a web novel be considered as a big data product? (Yes/No/No idea)
4. Should the bundle of 1000 pieces of music, 1000 photographs, 1000 movies and 1000 web novels be considered as a big data product? (Yes/No/No idea)
5. A big data product must contain data in multiple formats. (Yes/No/No idea)
6. The data volume of a big data product should reach (MB/GB/TB/PB or above) level.
7. A big data product must contain data in multiple formats and its volume should reach (MB/GB/TB/PB or above) level.
8. It is recognized internationally that a book should at least contain 49 pages of printed paper. In this sense, The data volume of a big data product should reach (MB/GB/TB/PB or above) level.
9. Imagine a bundle of data with a volume of about 30 GB, which contains a high-definition movie (10 GB), 10 music CDs (7 GB), 1000 photos each containing 1 million pixels (3 GB), 1000 books each containing 500 thousand words (1 GB), and a relation table containing 1 billion records (10 GB). Can all these be considered as the minimum scale of a big data product? (Yes/No/No idea)
10. Other questions on backgrounds of respondents.

Abbreviations

API: Application programming interface; eBook: Electronic books; CT: Computed tomography; DBMS: Database management system; ODBC: Open database connectivity; fMRI: Functional magnetic resonance imaging; GB: Gigabyte; HDFS: Hadoop distributed file system; ID: Identifier; ISBN: International Standard Book Number; ISO: International Organization for Standardization; JDBC: Java database connectivity; PB: Petabyte; RPM: Retail/resale price maintenance; SAPPRFT: The State Administration of Press, Publication, Radio, Film and Television of China; SQL: Structured query language; TB: Terabytes; UNESCO: The United Nations Educational, Scientific and Cultural Organization.

Acknowledgements

Not applicable.

Author contributions

Both the authors read and approved the final manuscript.

Funding

This work is funded in part by the Shanghai Science and Technology Development Fund No. 19DZ1200802, 19511121204, and the National Natural Science Foundation of China Projects No. U1936213.

Availability of data and materials

The raw results of the questionnaire are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China. [2]Shanghai Institute for Advanced Communication and Data Science, Shanghai, China.

Received: 22 November 2021 Accepted: 6 April 2022
Published online: 28 April 2022

References

1. Xiong Y, Zhu Y. Data box: a novel data model for self-governing openness of data. Big Data Res. 2018;4:2018015.
2. Fowler D, Barratt J, Walsh P. Frictionless data: making research data quality visible. Int J Dig Curation. 2017. https://doi.org/10.2218/ijdc.v12i2.577.
3. Qin Y, Wang Z, Wang H, Gong Q, Zhou N. Robust information encryption diffractive-imaging-based scheme with special phase retrieval algorithm for a customized data container. Opt Lasers Eng. 2018;105:118–24.
4. Jain S, Kannan P. Pricing of information products on online servers: issues, models, and analysis. Manage Sci. 2002;48(9):1123–42.
5. Oh H, Park S, Lee GM, Heo H, Choi JK. Personal data trading scheme for data brokers in IOT data marketplaces. IEEE Access. 2019;7:40120–32.
6. Yu H, Zhang M. Data pricing strategy based on data quality. Comput Ind Eng. 2017;112:1–10.
7. Ye Y, Zhang Y, Liu G, Zhu Y. A measure based pricing framework for data products. In: Web intelligence. IOS Press. 2021. p. 1–12.
8. Piattoeva N, Centeno VG, Suominen O, Rinne R. Governance by data circulation? the production, availability, and use of national large-scale assessment data. Politics of quality in education: a comparative study of Brazil, China, and Russia. 2018. p. 115–36.
9. Zhu Y, Xiong Y, Liao Z, Ye Y. Self-governing openness of data. Big Data Res. 2018;4:2018013.
10. Ye Y, Liu G, Zhu Y. The two-step authorization pattern of data product circulation. 2020.
11. Ardagna CA, Ceravolo P, Damiani E. Big data analytics as-a-service: issues and challenges. In: 2016 IEEE international conference on Big Data (big Data), IEEE. 2016. p. 3638–44.
12. Labrinidis A, Jagadish HV. Challenges and opportunities with big data. Proc VLDB Endow. 2012;5(12):2032–3.
13. Chen J, Chen Y, Du X, Li C, Lu J, Zhao S, Zhou X. Big data challenge: a data management perspective. Front Comput Sci. 2013;7(2):157–64.
14. Fang W, Wen XZ, Zheng Y, Zhou M. A survey of big data security and privacy preserving. IETE Tech Rev. 2017;34(5):544–60.
15. Kulynych J, Greely HT. Clinical genomics, big data, and electronic medical records: reconciling patient rights with research when privacy and science collide. J Law Biosci. 2017;4(1):94–132.
16. Li J, Tao F, Cheng Y, Zhao L. Big data in product lifecycle management. Int J Adv Manuf Technol. 2015;81(1):667–84.
17. Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. Clin Epidemiol. 2017;9:245.
18. Malik AW, Mahmood I, Ahmed N, Anwar Z, et al. Big data in motion: a vehicle-assisted urban computing framework for smart cities. IEEE Access. 2019;7:55951–65.
19. Marincic I, Foster I. Energy-efficient data transfer: bits vs. atoms. In: 2016 24th international conference on software, telecommunications and computer networks (SoftCOM), IEEE. 2016. p. 1–6.
20. Chiang M, Zhang T. Fog and iot: an overview of research opportunities. IEEE Internet Things J. 2016;3(6):854–64.
21. Whitney G. The UNESCO book production statistics. Book Res Q. 1989;5(4):12–29.
22. Agency II. ISBN user's manual. Internat. ISBN Agency. 2005.
23. Weissberg A. The identification of digital book content. Publ Res Q. 2008;24(4):255–60.
24. Løyland K, Ringstad V. Fixed or free book prices: is a hybrid system superior? Int J Cult Policy. 2012;18(2):238–54.
25. Stockmann D. Free or fixed prices on books-patterns of book pricing in Europe. Javnost-The Public. 2004;11(4):49–63.
26. Ragazzo C, de Costa e Silva Lima J. Fixed book price regimes: beyond the rift between social and economic regulation. Eur JL Reform. 2017;19:167.
27. Huang L, Li Y, Wang X, Zhao Y. Challenge of encryption technology for self-governing openness of data. Big Data Res. 2018;4:2018018.
28. Shvachko K, Kuang H, Radia S, Chansler R. The hadoop distributed file system. In: 2010 IEEE 26th symposium on mass storage systems and technologies (MSST), IEEE. 2010. p. 1–10.
29. Thusoo A, Sarma JS, Jain N, Shao Z, Chakka P, Zhang N, Antony S, Liu H, Murthy R. Hive-a petabyte scale data warehouse using hadoop. In: 2010 IEEE 26th international conference on data engineering (ICDE 2010), IEEE. 2010. p. 996–1005.
30. Junqueira F, Reed B. ZooKeeper: distributed process coordination. Newton: O'Reilly Media, Inc.; 2013.
31. Cinkosky MJ, Fickett JW, Gilna P, Burks C. Electronic data publishing and Genbank. Science. 1991;252(5010):1273–7.
32. Klump J, Bertelmann R, Brase J, Diepenbroek M, Grobe H, Höck H, Lautenschlager M, Schindler U, Sens I, Wächter J. Data publication in the open access initiative. Data Sci J. 2006;5:79–83.
33. Lawrence B, Jones C, Matthews B, Pepler S, Callaghan S. Citation and peer review of data: moving towards formal data publication. Int J Dig Curation. 2011;6(2):4–37.

Publisher's Note