UNIVERSITAS UTPADAKA SWASTIKA
Jl. KS Tubun No.11 Tangerang 15112 Banten, Telp. (021) 5589161-62 Fax. (021) 5589163
JISI Jurnal Ilmiah Sistem Informasi
ISSN: 3046-711X

Big Data Trade Survey Pricing, Trading, and Data Protection

Lukas Umbu Zogara

Program Studi Teknologi Informasi, Utpadaka Swastika University, Tangerang, Indonesia, 15112
E-mail: lukasumbuzogara68@gmail.com

Accepted: March 21, 2024 | Published: April 30, 2024

ABSTRACT

In this era, Big Data is considered as the key to unlocking the growth of people's productivity in daily life. Internet usage, mobile applications and social networks, as well as the internet of things based on smart-grid and so on has an impact on the amount of data collected. The advancement of analytics data provided by machine learning and data mining technology supported by cloud computing, the resulting information will be more useful. Since the advent of Big Data, the dataset has become one type of "new money" in the digital world. Copyright protection mechanisms including digital encryption and watermarking need to be done to maintain data values. To maximize this, an effective container is needed to allow data owners and buyers to trade the data. The focus of the topic is on designing data trading platforms and schemes, supporting effective and efficient data trading, security, and maintaining the privacy of data owners.

KEYWORDS: Big Data Market, Data Privacy, Trading

1. Introduction

Technological advances are already integrated into our daily lives, such as mobile and social network applications, and Internet of Thing (IoT)-based smart world systems. Different types of sensors and smart devices generate huge data sets continuously from all aspects and domains [1]. Thus, the unprecedented, comprehensive, and complex data, namely Big Data, becomes more valuable. In addition, with advances in data analytics provided by machine learning, data mining, and computing capabilities supported by cloud infrastructure, the potential values of Big Data generated are becoming more useful. However, there are a number of significant challenges, including data collection, storage, analysis, sharing, approval, and others [2]. To maximize the usefulness of the collected data, one of the solutions that can be used is to design an effective and secure data trading market that involves data owners.

With increasing attention to the economic value of Big Data in improving efficiency and decision-making, customer experience, and more, several third-party big data commerce markets have been designed [3]. For example, Global Big Data Exchange (GBDEX) has 150 PB of official tradable data collected from thousands of companies and organizations. Nonetheless, due to the lack of viable protocols, the existing Big Data trading market is still in its infancy. To create an effective market for data trading, there are several challenges that need to be faced. The first issue is related to the price that needs to be determined precisely for the data to be traded as well as considerations regarding market structure in the design of an appropriate data pricing model. Through a fair price, the economic benefits of data owners and consumers can be ensured [4]. The second issue is related to the platform and the data trading scheme.

In this case, viable trading platforms and schemes should be designed to ensure profitability, fairness, honesty, and data privacy [5]. The third issue is related to data copyright protection, as digital products can be easily counterfeited or duplicated. In particular, if the purchased data is resold by the buyer, the value of the data from the original data owner as the seller will have a significant effect. Thus, data copyright protection schemes should be designed to guarantee the legal rights of their owners [6].

The total amount of data in the world explodes with an estimated 2.5 quintillion bytes of data generated every day [7]. Almost 90% of the data in the world was created in the last two years alone. Data sources are diverse, especially as IoT becomes increasingly involved in our daily lives, underpinning a wide range of IoT systems [8]. Such diverse data sources result in expansive volumes of data, while creating enormous commercial value potential. Data can have a variety of different and complementary formats, such as log data from various devices and applications, database files, XML files, and others. In addition, data can have unstructured data types (image, video and audio streams, etc.). Thus, Big Data is massive, sustainable, and comprehensive, and has high potential commercial value thanks to advances in data analytics techniques, such as machine learning and Data mining [9].

In the data trading market, data can be effectively changed/shared among individuals and organizations, the improvisation of utilizing data significantly in IoT applications also generates tremendous potential value in economics [10]. Especially in online data trading, data sellers are in the majority position. But, in most data trading mechanisms today, the privacy of the data seller is studied from the point of view of the data buyer, and the price of the data is determined by the data buyer [11].

To overcome problems regarding Big Data trading, in this study, ways that can be done such as reviewing existing research related to Big Data, and identifying Big Data cycles for data trading, including data collection, data analytics, data pricing, data trading, and data protection [12]. It then categorizes popular market structures, pricing strategy data, and pricing model data, and lists the advantages and limitations of each category. Investigate the process of data trading, and summarize the problems of data trading and find solutions to deal with those problems [13]. Then study the last part of the Big Data cycle which is about data protection. This study summarizes existing copyright protection schemes and describes the advantages and disadvantages of Stage 2: Customer Segmentation This stage is carried out using the K-Means algorithm to group customers in each segment. The results will be used to predict customer needs using the naïve bayes algorithm. The prediction aims to develop a business marketing strategy in the future [16]such methods, as well as outlining the challenges of Big Data copyright protection[14].

2. Methodology

There are several processes that need to be carried out in Big Data to determine customer segmentation as the main target in research, namely data accessing and computing, data privacy and data mining [15]. In this study, work focused on Big Data algorithms using RFM (Recently, Frequency, Monetary) analysis to separate customers based on their transaction activities.

Step 1: Identify Customer Patterns

The classification of customers is categorized based on the determination of weights for each marketing assessment factor. Next, the scores of each factor will be accumulated and sorted by highest score to lowest

Stage 2: Customer Segmentation

This stage is carried out using the K-Means algorithm to group customers in each segment. The results will be used to predict customer needs using the naïve bayes algorithm. The prediction aims to develop a business marketing strategy in the future [16]

Stage 3: K-Means Algorithm

This algorithm is used to group large amounts of customer data based on certain factors. There are five steps used in the K-Means algorithm, namely determining the value of a subset that is not empty and is randomly generated. Next, identify the cluster centroids on each partition and calculate the minimum Euclidean distance from each point to the cluster centroids. It then recalculated the mean distance of each cluster and assigned a new centroid cluster based on the results of the Euclidean distance calculation. The last stage is to compare the new centroid cluster with the initial centroid cluster until observations are completed [17].

Stage 4: Naïve Bayes

It is a classification method of Bayes' theorem. This method uses probability and statistical methods that predict future opportunities based on past experience. The main characteristic of this method is a very strong assumption that continues to be independent of each condition or event. Naïve Bayes works very effectively compared to other methods because it has a better level of accuracy. This algorithm is one of the classification methods used to predict the potential needs of customers from previous probability calculations and posterior probabilities. Thus, the company can make the right marketing strategy to meet the needs [8].

Furthermore, in the process of maintaining the privacy of transaction data, a method is needed that can guarantee this. The method used is homomorphic encryption. This method protects data privacy and confidentiality, while improving batch verification and data trading processes. In contrast to traditional encryption schemes, the signature component processes identity-based data in the ciphertext space [6].

This method is designed to perform complete encryption by utilizing an ideal vector network. In general, the design of an encryption scheme can evaluate its own debrance circuit. Due to the nature of the lattice ideal which is a linear combination with an integer of itself, addition and multiplication operations can be performed on its encryption [11].

In vector-based cryptography, encryption works through a multi-dimensional network and places it somewhere at the point of the vector circuit [4]. Each vector network has an internal structure so that an attacker will never know about the path or rate close to that point.

This method provides a fundamental solution on how to allow unknown people to perform the calculation process against encrypted personal data without needing to know the contents of the data. In other words, we still allow unauthorized parties to process encrypted data in order to provide effective and efficient calculation results [18].

3. Results and discussion

Big Data is now the most important resource in the era of data technology. To trade or share data resources, how to evaluate the commercial value for that data set is a fundamental issue. In addition, capturing and mining value from a data set can further increase the value of the data. To determine the commercial value of Big Data, we need to determine what the commercial value of the data set is. Nonetheless, there are some challenges involved with applying data mining to Big Data. The first challenge focuses on data access and computational procedures [19].

Due to distributed storage systems and ever-expanding data volumes, computing platforms must have the ability to handle distributed and large-scale data stores. Most data mining algorithms require loading all the necessary data into main memory, which is obviously a technical challenge in the case of Big Data, as moving data from distributed storage systems is expensive. The second challenge is the variety of Big Data applications. More specifically, applications exist in different domains, with different data privacy and data sharing schemes between data owners and consumers. The third challenge is designing effective machine learning and data mining algorithms. Learning and mining algorithms must handle the difficulties of large volumes, and distributed, complex and dynamic data characteristics [19].

With regard to data commerce, remind that data is a virtual or digital item and has its own characteristics. So, to trade data, it is necessary to build market data securely by paying attention to the privacy of the data owner. Before the customer decides to buy the data, many search processes are carried out. In handling this, a search system is needed to minimize costs incurred by consumers. The solution found is through optimization with a scheme to reduce the number of queries for the data trading process. This scheme is included in the execution engine to obtain certain information.

One of the most popular data trading mechanisms is through the data auction process. In general, auction is a scheme that aims to allocate and set the appropriate price between buyers and sellers. Auction theory has been well explored in several fields due to its ability to show great potential in data trading. In auction theory, there is an auction mechanism carried out by several aspects that are considered including bidders, auctioneers, sellers, valuations, and clearing prices. In the auction process, bidders are people who apply for a warning and aim to buy the commodity. Bidders are usually categorized as start-up companies that want to conduct application investigations using specific data sets. The auctioneer acts as someone who runs the auction process and determines the winner and until the transaction process takes place. The seller is the owner of the commodity from a large amount of data. Valuation in the auction process refers to buyers and sellers in each commodity union. Price clearing in the auction process is the process of bid prices from commodities or data to be utilized. The price is the price that has been agreed upon by both parties between the seller and the buyer.

With the digitization of traditional media growing every day, more and more data is stored in digital volumes so that the community is changing from practical to virtual. This causes the distribution of data easily can be spread quickly. So data protection is needed as a condition to secure data ownership [19].

The Digital Rights Management (DRM) has been established to prevent intentional digital content from being copied, shared, as well as stolen, acting more importantly as a guiding file in the development of digital copyright protection. All these DRM solutions require five key components [20] (i) Security. It focuses on content file encryption and hashing, watermarking, and signatures for digital content. (ii) Access control. It is responsible for identity and access management, and credential provision for users who need to access protected digital content. In addition, this component monitors the behavior of authorized users, and assigns different access rights to different users. (iii) Usage control. It monitors usage for each authorized user, and records usage as history. (iv) License management. It releases licenses (keys, XrML files, authentication codes) to authorized users, and controls and checks the lifetime (validity period) for a license. (v) Payment Management. This component works with the context of use. and calculate the fees that users must pay. This is the main goal of digital business.

Watermarking technology has developed widely and has been applied in video and audio. The watermarking algorithm is designed to embed watermarks into quantization index modulation to improve and avoid content corruption. It should be noted that images are also considered multimedia content that can be considered. Watermarking technology is the most common technology as an approach to protecting copyright.

4. Conclusion

Specifically, this study refers to research relevant to Big Data and identifies the Big Data life cycle for data trading. This includes data collection, data analytics, data pricing, data trading, and data protection. This research revisits the categories of market share and data security. For the data trading process, further investigation of data auction strategies as well as detailing various schemes regarding data trading, trading platforms, and related issues. Data protection is carried out as the final stage of the Big Data life cycle. The purpose of this study is to provide a clear and in-depth understanding of data trading while still paying attention to data protection as an effort in the development of Big Data.

References

[1] S. C. Wang, Y. Te Tsai, and Y. S. Ciou, "A hybrid *Big Data* analytical approach for analyzing customer patterns through an integrated supply chain network," *J. Ind. Inf. Integr.*, vol. 20, p. 100177, 2020, doi: 10.1016/j.jii.2020.100177.

[2] S. Ying, S. Sindakis, S. Aggarwal, C. Chen, and J. Su, "Managing *Big Data* in the retail industry of Singapore: Examining the impact on customer satisfaction and organizational performance," *Eur. Manag. J.*, no. xxxx, 2020, doi: 10.1016/j.emj.2020.04.001.

[3] P. Hajek and M. Z. Abedin, "A Profit Function-Maximizing Inventory Backorder Prediction System Using *Big Data* Analytics," *IEEE Access*, vol. 8, pp. 58982–58994, 2020, doi: 10.1109/ACCESS.2020.2983118.

[4] T. Kutuzova and M. Melnik, "Market basket analysis of heterogeneous data sources for recommendation system improvement," *Procedia Comput. Sci.*, vol. 136, pp. 246–254, 2018, doi: 10.1016/j.procs.2018.08.263.

[5] A. A. Khade, "Performing Customer Behavior Analysis using *Big Data* Analytics," *Procedia Comput. Sci.*, vol. 79, pp. 986–992, 2016, doi: 10.1016/j.procs.2016.03.125.

[6] X. Xu, Y. Shen, W. (Amanda) Chen, Y. Gong, and H. Wang, "Data-driven decision and analytics of collection and delivery point location problems for online retailers," *Omega (United Kingdom)*, vol. 100, no. xxxx, p. 102280, 2021, doi: 10.1016/j.omega.2020.102280.

[7] T. Hayashi and Y. Ohsawa, "The acceptability of tools for the data marketplace among firms using market research online communities," *Procedia Comput. Sci.*, vol. 176, pp. 1613–1620, 2020, doi: 10.1016/j.procs.2020.09.184.

[8] Q. Li, Y. Chen, J. Wang, Y. Chen, and H. Chen, "Web Media and Stock Markets : A Survey and Future Directions from a *Big Data* Perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 2, pp. 381–399, 2018, doi: 10.1109/TKDE.2017.2763144.

[9] S. Han, W. Reinartz, and B. Skiera, "Capturing Retailers' Brand and Customer Focus," *J. Retail.*, 2021, doi: 10.1016/j.jretai.2021.01.001.

[10] A. S. M. Systems, "environments."

[11] F. S. Parreiras, G. Groner, D. Schwabe, and F. D. F. Silva, "Towards a marketplace of open source software data," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2015-March, pp. 3651–3660, 2015, doi: 10.1109/HICSS.2015.439.

[12] F. Wang, M. Li, Y. Mei, and W. Li, "Time Series Data Mining: A Case Study with *Big Data* Analytics Approach," *IEEE Access*, vol. 8, pp. 14322–14328, 2020, doi: 10.1109/ACCESS.2020.2966553.

[13] A. Griva, C. Bardaki, K. Pramatari, and D. Papakiriakopoulos, "Retail business analytics: Customer visit segmentation using market basket data," *Expert Syst. Appl.*, vol. 100, pp. 1–16, 2018, doi: 10.1016/j.eswa.2018.01.029.

[14] J. S. K. Tan, A. K. Ang, L. Lu, S. W. Q. Gan, and M. G. Corral, "Quality Analytics in a *Big Data* supply chain: Commodity data analytics for quality engineering," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, pp. 3455–3463, 2017, doi: 10.1109/TENCON.2016.7848697.

[15] L. C. Boldt *et al.*, "Forecasting Nike's sales using Facebook data," *Proc. - 2016 IEEE Int. Conf. Big Data, Big Data 2016*, pp. 2447–2456, 2016, doi: 10.1109/BigData.2016.7840881.

[16] W. Gao, W. Yu, F. Liang, W. G. Hatcher, and C. Lu, "Privacy-Preserving Auction for *Big Data* Trading Using Homomorphic Encryption," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 2, pp. 776–791, 2020, doi: 10.1109/TNSE.2018.2846736.

[17] F. Dong, S. Yuan, H. Ou, and L. Liu, "New Cyber Threat Discovery from Darknet Marketplaces," *2018 IEEE Conf. Big Data Anal. ICBDA 2018*, no. December 2017, pp. 62–67, 2019, doi: 10.1109/ICBDAA.2018.8629658.

[18] Y. Gao, X. Chen, and X. Du, "A *Big Data* Provenance Model for Data Security Supervision Based on PROV-DM Model," *IEEE Access*, vol. 8, pp. 38742–38752, 2020, doi: 10.1109/ACCESS.2020.2975820.

[19] F. Liang, W. Yu, D. An, Q. Yang, X. Fu, and W. Zhao, "A Survey on *Big Data* Market: Pricing, Trading and Protection," *IEEE Access*, vol. 6, pp. 15132–15154, 2018, doi: 10.1109/ACCESS.2018.2806881.

[20] H. Oh, S. Park, G. M. Lee, J. K. Choi, and S. Noh, "Competitive Data Trading Model With Privacy Valuation for Multiple Stakeholders in IoT Data Markets," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3623–3639, 2020, doi: 10.1109/JIOT.2020.2973662.