Data-driven Modeling and Service based on Big Data Analytics and Perception Process

Wooyung Lee[1], Eunji Jang[2], Joon Lee[1],*

Graduate School of Convergence Science and Tech.,
Seoul National University[1],,*
Dept. of Computer Science and Eng.,
Seoul Women's University, Seoul, Korea[2]
*w.lee@xiilab.com[1], s111559@swu.ac.kr[2], joonlee8@snu.ac.kr[1],**

In this paper, we propose a data-driven service and model which trades the data products generated by the big data analytics and perception process. Existing global data brokers have traded the data products in categorized and listed. Data has been traditionally managed in closed form or limited to specific industries and eco-systems. However, we are faced on the fourth Industrial Revolution, where diverse data is based on utilization, data is becoming a core competence for change and the future. Finding the value in Big Data has been generalized and responded to the change. Also, there is a great demand for revitalization of data processing and utilization of data for industrial application. To support these requirements, the data must be traded in a way that can be easily utilized according to the data analytics and perception process. This is a key technology in the data broker industry, which will enhance data-driven industries.

1. Introduction

Previous industrial systems have been focused on the value of automation and this produced a result that generated data was confined to a closed system. However, in the 4th industrial revolution, which is emerging in recent years, data has become a material for creating value of artificial intelligence unlike the past [1]. Data is leading ICT to the change of

***Journal of Cognitive Science* 18-2: 201-214, 2017**
Date submitted: 05/25/17 Date reviewed: 06/17/17
Date confirmed for publication: 06/28/17

revolution and economic paradigm to the intelligent society. At the same it been an important task to have valuable data from various objects and people.

From the intelligent society, the perspectives on data determine how to make the data meaningful and create the value. As a result, many companies are focusing on deriving meaningful conclusions from their data in their industries and various analysis techniques and solutions for Big Data are emerging [2]. However, existing data broker industry provides bucket data in a simple category format, which requires further efforts to find data value for data users.

In this study, we propose a new data brokerage type that classifies the data brokers. The brokers are increasing in necessity but not yet entered the activation stage. Trading data based on the data analytics process will enable users who can not easily utilize data services to perform more competitive data analysis and become a driving force for constructing a data ecosystem [3].

The data broker model proposed in this paper is a service model that continuously distributes the profits after the transaction to the supplier that supplies the data. It is a data transaction service model that creates a virtuous circle of data transactions that can share the benefits of data acquisition time and cost savings to the data consumer.

2. Related Research

2.1 Background of Data Brokers

Data brokers collect and sell information for a variety of purposes including for fraud prevention, credit risk assessment, and marketing. One of the primary ways data brokers package and sell data is by putting consumers into categories or "buckets" that enable marketers – the customers of data brokers – to target potential and existing customers [4]. These data brokers now handling the industry has amassed trillions of digital consumer records, or 'big data', that are stockpiled, analyzed, and sold [5].

Data brokers can be divided into two groups according to the price setting method and the differentiated price method. The price setting criteria are cost base price, competitive base price, and demand base price. Differentiation can be made according to quality, time, user attribute, and

congestion by pricing method according to discrimination price [6]. The proposed model is based whether there is a data model and a data is defined from existing broker model.

2.2 Global Data Brokers

To select reference models for data trading or distribution, four representative global data broker service model were compared.

Acxiom was founded in the United States in the 1960s and began work on providing a list of electoral postal addresses [7]. Now, it is one of the companies with the largest personal information in the world, mainly selling and consulting private information data to multi-national global companies.

Kaggle is a representative platform for data sharing and is also famous for handling data for contests [8]. The analysts who solve the problem by solving the data set are not Kaggle 's own manpower but the teams who are teamed around the world. When the participating team solves the problem and achieves the goal, it receives various amount of money as the prize money. In addition to receiving license fees, the data provider gets a title called Data Expert. Therefore, Kaggle datasets are diverse. And analysts, analyzing models, and algorithms that solve data sets are also diverse.

Palantir was founded by Peter Thiel, a founder of Paypal, an online payment company, its main customers are mainly CIA (Central Intelligence Agency), NSA (National Security Agency), FBI (Federal Bureau of Investigation) [9]. Ministry of Defense and Marine Corps. And much more on private companies. Palantir is analyzing the data of the organization and performing various analytic activities such as terrorist detection, money flow tracking, missing and missing persons tracking.

Quandl in Canada distributes some free-of-charge data, such as time series data and social data, mainly on stock-based financial data, and offers data in various forms such as JSON, xml, CSV in data formats of API type, R, Python, Excel, Ruby.

2.3 Existing Data Brokers Models

The data brokers mentioned above are representative data brokers in the global market, each having its own unique characteristics: there is a certain types of data brokers (1) who has the defined data professionally but not in

the model, (2) who has undefined data set and not having data model and, and (3) who has no defined data but has its own data model.

Figure 1. shows the data brokers according to the above categories (1) to (3). The horizontal axis shows whether the broker has set the defined data set, and the vertical axis shows whether the data analytic model has its own. Acxiom and Quandl each have a data set that represents each company as personal information data and financial data. However, there is no specific data analysis model because it deals with the data itself. In the case of Kaggle, there is no data that specially handled the data set provided by the participant company in the format of the participant, and it is not the form that has the analysis model itself. On the other hand, in the case of Palantir, it is the case that it has the characteristic of receiving the data of the clients and analyzing the data by using the analysis model. Therefore, we can say that their data set is not clear but they have analysis model which is the basis for analyzing data.

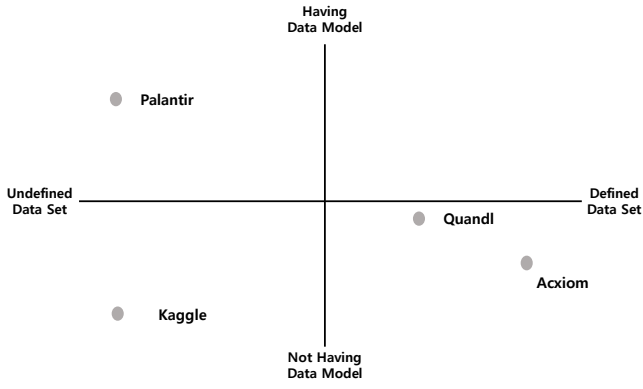

Figure 1. Data Brokers with model and data

2.4 A Proposed Data Broker

From the Figure 2, this study proposes a model that has a data broker service model based on data analytic process and can analyze the defined data.

Many existing data brokers are trading on the dataset itself without a data analysis model. This is primarily aimed at solving the difficulties of instantaneous use of data analysis. The proposed data model is named DataFarm. It should be noted that DataFarm is a data broker service

platform based on data analytic process.

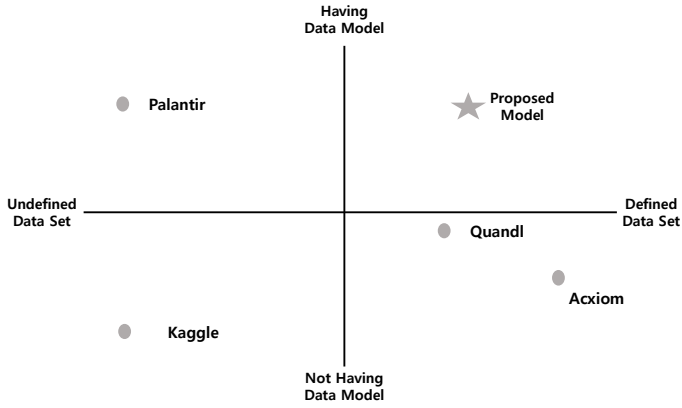

Figure 2. Proposed data broker model, DataFarm

3. Proposed Model

3.1 Data Model based on Analytic Process

When trading data, it is important to know the format of data. For example, the data broker must decide whether to provide only the source of data requested by the user, or to provide the analysis data together. To do so, a whole data process from data collection to analysis and processing is required.

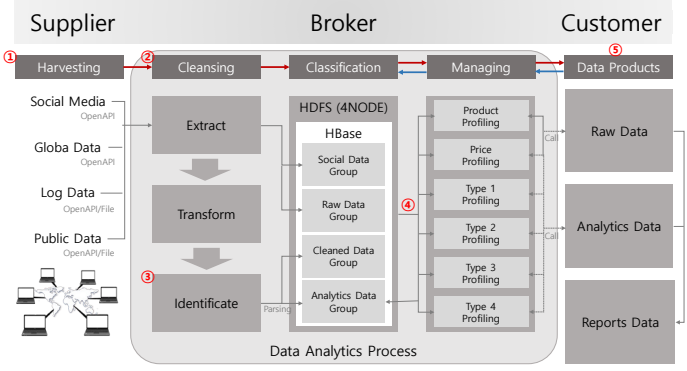

Figure 3. Data Broker Service based on Data Analytic Process

Figure 3. shows the overall process of data analysis from the data harvesting to managing.

① It is a step of harvesting data. In the case of social data, OpenAPI collects data in the form of OpenAPI or File in the case of public data.
② It is the step of data cleaning and loading in consideration of data harvesting type, data size, etc.
③ It is a step of parsing the loaded data and storing it in the database preprocessing or analysis data groups.
④ It is a managing step of classifying and profiling the data stored in the preprocessing data group.
⑤ It is the step of reporting the output data as source data, analysis data, and analysis report. Also, these are types of data products.

3.2 Data Harvesting

Data is collected through acquisition, purchase, and linkage [10]. In this study, we used government data, trend data from search engine, and social data. The collected data is searched as meta information, sourced as data for web development for providing data products, and an environment is created by API to link with the web.

In the case of FTP / SFTP, we used NoSQL according to the data type and size to send / receive files from the server via TCP/IP. When receiving data from a collection agency such as a private or public organization, we collect data in the form of OpenAPI and uses the private API through authentication according to the data characteristics. In addition, Hadoop-based Big Data Technology was used depending on the size and range of data.

3.3 Data Cleaning

The way of data is cleaned depends on the type and size of the data. For data preprocessing and storage, and for file systems, data that is not needed for analysis after data collection should be removed. After leaving only the data necessary for analysis, it is decided whether to store the data in the general file system or the distributed file system according to the size of the data, and then the classification operation is performed.

Because big data can not be processed in the general file system, it is stored in distributed file system. In the case of preprocessed data or structured data, it is stored in RDB of MySQL, Oracle, etc.

When storing unstructured data that does not require preprocessing, NoSQL such as HBase or MongoDB is used [11]. In addition, the appropriate architecture is applied considering data characteristics.

3.4 Data Classification and Mining

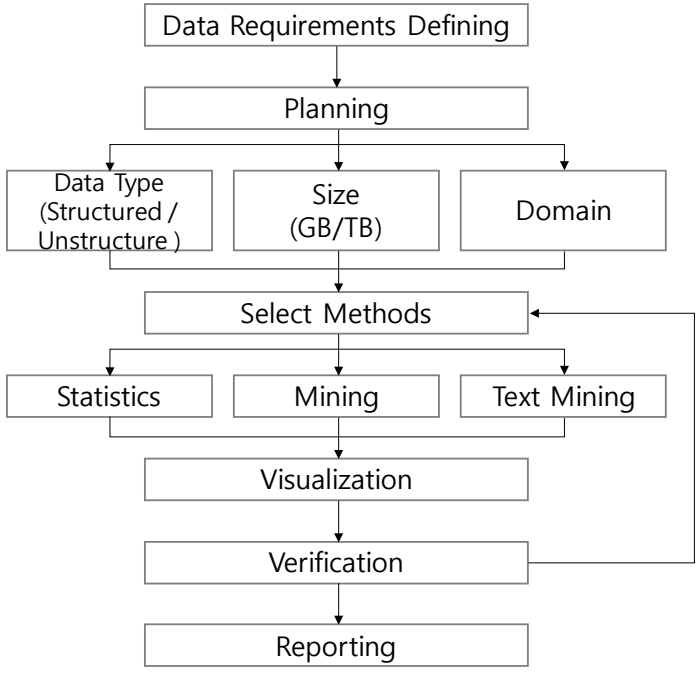

Figure 4. Mining Process according to analysis requirements

Figure 4. is shown a flow-chart showing the overall process of data mining. It is also shown that the process of analyzing the data according to the user's requirements.

First, the user receives the data analysis request, and when the requirements are listed, the data for the analysis that meets the needs is

selected. And the analysis plan is established by grasping the characteristics, size and category range of the data.

We selected statistical analysis, data mining, and text analysis to determine the appropriate data analysis method and analysis algorithm.

The results of the analysis were integrated and visualized to provide services to users.

4. Model Evaluation

As a result of this study, DataFarm is designed based on data analytics process and launched through website. DataFarm sought to secure an integrated channel of public and private data related to agriculture to provide farmers with the data they needed to conduct their business and to access data more conveniently.

With data product planning and design, DataFarm provided data related to cultivar, environment, and price related to agriculture in the form of dashboard analyzed by topic so as to construct data integration environment and to increase accessibility of data analysis. To access real data through category classification, we provide data service by API and table.

Figure 5. DataFarm Web Service

Figure 5. actually shows the appearance of the web service of DataFarm. DataFarm provides the convergence data such as source data, analysis data,

and data analysis report reflecting the real time market environment.

DataFarm provides users with the ability to select the necessary information as a subject language, and provided integrated environmental data analysis for production and sales activities through a diversified decision-making system based on the data base.

In addition to the basic data of existing services, it provides integrated data utilization environment such as environment, price, and index data, and provides detailed data such as revenue, flow population information, and production amount information that can be utilized by actual companies.

4.1 Systems Environments

Prior to the development of the data broker platform, hardware such as server and DB were provided to provide services [12]. Also, we constructed shared OSS (Operating Supporting Systems) / BSS (Business Supporting Systems) based on data broker service on the constructed server and provided UI / UX interface for web service.

Web services provided functions such as user management, payment service linkage, service policy management, and service-specific statistics. Hadoop-based high-performance computing technology environment was supported for data analysis [13].

In the case of large data, Hadoop-based high-performance computing architecture design technology for distributed processing has been developed to overcome the time-consuming work [14].

We used a high-performance processor that has a high degree of integration in processing operations and can process large numbers of operations quickly when large-scale data is distributed.

4.2 Types of Data Products

When providing data from the proposed data model, it is user 's experience to provide information for making by providing them on a dashboard basis, rather than providing simple information. The analytical data is based on price, shipping, environment.

Automatic data query generation and dashboard according to the selected topic are implemented and data query is generated and executed by user's request and analysis service is provided [15].

The data related to topics was provided by API or data tables by category and search. And the data reflecting the valuation of the data was provided. In addition, we provided data creation and transformation services that matched the request. Search-based data matchings are provided in data types and libraries such as CSV, JSON, Excel, Python, and R [16].

Data was processed using an internal data processing platform for data products. Based on the source data, the result data of the analysis, the analysis PaaS, the analysis report, and the fusion data were calculated. The data transaction broker platform provided a data catalog that minimizes the difference between the amount of data to be provided.

4.3 Data Broker Platform Considerations

The work of data collection among the broker platform is performed in the Mashup-Platform Layer. When designing the Mashup Platform Layer, it is considered whether the collected data is flowed into the real-time streaming, the API is collected at a certain time interval, data conversion process, etc [17].

In order to analyze data processing with different properties before going to the analysis step, it is necessary to convert the raw data for each data attribute into a database and convert between different types of data to sort and block according to specific criteria [18].

The data that has passed through the Mashup- Platform Layer is converted into a form that can be used in the Analysis Platform Layer. In the Analysis Platform Layer, the standard analysis function is provided according to the user's selected keyword.

When designing the Analysis Platform Layer, the security issues related to accessibility should be considered in the data, and the designated data is limited according to the data required [19].

In the dashboard, core analysis and monitoring functions, task-specific configuration functions, and graphic data standards were considered.

When analyzing topics, it is necessary to analyze them through keywords such as topics, keywords, and trends [20]. And the results are visualized through Open Platform and provided to users.

When designing a visualization platform, a standard interface for accessing various Data Bases, various access devices, and a graphical

representation covering the user environment were considered. When visualization results are extracted, they must be extracted in standard format, and abnormal patterns, singularities, and associations are detected and extracted together with analysis results.

5. Conclusions

As the interest in big data related technologies and data industries grows, the application area and scale are expanding every year, and the actual demand for data is increasing. As a result, the data utilization in the public sector continues to be led by the government, but at the same time, it is difficult to extend to the private sector. This is urgent to secure the public and private big data supply system.

While commercialization of solutions for data collection, analysis and processing in various public and private sectors is actively being carried out, the service platform of 'sharing' perspective for spreading data utilization is not able to use wide range of data. To do this, it is necessary to expand the service of the data broker platform.

In this study, a model that can create and trade data products based on data analysis process is designed as a data broker service named DataFarm. Through data broker in the agricultural sector, we wanted to deal with data products that meet the needs of converged data users and provide users with important information when making business decisions. This effort helps users who need the data and the datasets they handle also solve other problems. The data brokers, from (1) to (3) previously compared, did not have this model. Data analysis and visualization are required for data broker. And the proposed model is constructed based on the data broker environment with user's purpose of use.

Although data broker services are still at the level of data monitoring and gathering of data now, the need for data utilization increases as the society enters the future and data broker market will be increased every year.

In order to activate the data brokerage industry, it is necessary to solve the problem of retaining the technology that can process the data and the policy problem. To maintain and expand the data brokers several follow-up studies

are needed. We believe that the important issues required for further study are data pricing issues, data processing methods, and detection issues related to fraud that occur during data transactions

References

Shem Yin, Okyay Kaynak, "Big Data for Modern Industry: Challenges and Trends ", Proceedings of the IEEE, Vol 103 No.2, pp. 143-146, Feb. 2015

Avita Katal, Mohammad Wazid, R. H. Goudar, "Big data: Issues, challenges, tools and Good practices", Contemporary Computing (IC3), 2013 Sixth International Conference, Aug. 2013

Smitha Sundareswaran, Anna Squicciarini, Dan Lin, "A Brokerage-Based Approach for Cloud Service Selection", Cloud Computing (CLOUD), 2012 IEEE 5th International Conference, Jun. 2012

Kuempel Ashley, "The Invisible Middlemen: A Critique and Call for Reform of the Data Broker Industry", Northwestern Journal of International Law & Business, Vol 36 No. 1, 207-234, Win 2016

Roderick Leanne, "Discipline and Power in the Digital Age: The Case of the US Consumer Data Broker Industry", Critical Sociology, Vol 40 No. 5, pp 729-746, Sep. 2014

United States Senate, Office of Oversight and Investigations Majority Staff, "A Review of the Data Broker Industry: Collection, Use, and Sale of Consumer Data for Marketing Purposes," STAFF REPORT FOR CHAIRMAN ROCKEFELLER, pp 3-10, Dec. 18, 2013

Roderick, L., "Discipline and Power in the Digital Age: The Case of the US Consumer Data Broker Industry", Critical Sociology, 729-746, Jan. 2014

Korea Database Agency, "Public Data Pricing Policy", pp.11-21, Sep. 2016.

Kim,S.G., Lee,S.J., Kim, J.G., "A Study on the Development of Phased Big Data Distribution Model Based on Big Data Distribution Ecology", Journal of Digital Convergence),Vol.14 No.5, pp.95-98, 2016. 5.

Wayne, L.D., "The Data-Broker Threat: Proposing Federal Legislation to Protect Post-Expungement Privacy", The Journal of criminal law & criminology, Vol.102 No.1 253-282, 2012

Goli-Malekabadi. Z, Sargolzaei0Java, M, Akbari. MK, "A performance comparison of SQL and NoSQL databases", computer methods and programs in biomedicine, Vol 132, 75-82, Aug. 2016

Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, "Business Intellligence And Analytics: From Big Data to Big Impact", MIS Quarterly, Vol 36 No. 4, pp. 1165-1188, Dec. 2012

Jens Dittrich, Jorge-Arnulfo Quiane-Ruiz, "Efficient Big Data Processing in Hadoop MapReduce", Proceedings of the VLDB Endowment, Vol 5 No. 12, pp 2014-2015, Aug. 2012

Nordber Henrik, Bhatia Karan, Wang Kai, "BioPig: a Hadoop-based analytic toolkit for large-scale sequence data", Bioinformatics, Vol 29 No. 23, pp. 3014-2019, Dec. 2013

Scott Bateman, Jaime Teevan, Ryen W. White, "The search dashboard: how reflection and comparison impact search behavior", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp 1785-1794, May. 2012

Muenchen, Robert A. "The popularity of data analysis software." URL http://r4stats.com/popularity (2012).

Liu, Xuanzhe, et al. "Towards service composition based on mashup." Services, 2007 IEEE Congress on. IEEE, 2007.

Pitt, Janice S., and John Lawton. "Method and apparatus for conversion of database data into a different format on a field by field basis using a table of conversion procedures." U.S. Patent No. 5,493,671. 20 Feb. 1996.

Cuzzocrea, Alfredo. "Privacy and security of big data: current challenges and future research perspectives." Proceedings of the First International Workshop on Privacy and Secuirty of Big Data. ACM, pp45-47, 2014.

Boyd, Danah, and Kate Crawford. “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon.” Information, communication & society. Vol 15 No. 5, pp 662-679, May. 2012