ISSN: 2614-8404

Data Pipeline Architecture with Near Real-Time Streaming Multiple Source Indonesian Online News Data Lake

Angelina Pramana Thenata[1*]
[1]Program Studi Teknik Informatika, Fakultas Teknologi dan Desain, Universitas Bunda Mulia
email: [1]angelina.pramana31@gmail.com

Abstract – The rapid development of information has made online news increasingly needed. Online news attracts readers' attention by providing convenience and speed in presenting news from various fields. However, the large amount (volume) of online news that spreads in a short time (velocity) and the public's need to consume news in various references (variety) can affect people's lives. Therefore, the government as the regulator and news agencies need to monitor online news circulating. Based on these problems, the researcher proposes a data lake architectural design that is suitable for online news and can run in real-time. Data lakes can solve the main problems of Big Data (volume, velocity, variety). In proposing this data lake architecture, the researcher conducted a literature study and analyzed the flow of the data lake architecture according to online news. Furthermore, the researcher will use this architecture to combine and uniform the online news data structure from several online news channels and then stream it in real-time to fill the data lake. The results of using the data lake architecture for online news will be stored on MongoDB which functions as a database to store all data for both the short and long term. Finally, this data lake will be a means to accommodate, dive into, and analyze the circulating online news data.
Keywords – Data Lake, Online News, Real-Time

I. INTRODUCTION

Information technology in modern life has developed rapidly. This is marked by the presence of the internet which makes it easier for humans to communicate and get information without recognizing the limitations of space and time [1]. This development has had an impact on the existence of the media, including online news which provides a variety of information and news. The online news that is displayed can be in the form of events, opinion events, ideas, and so on. Online news attracts readers' attention by providing convenience and speed in presenting news from various fields, including economy, politics, technology, lifestyle, and others [2].

As information develops, online news is increasingly needed. Judging from Similar Web regarding traffic visits in June 2020, several Indonesian online news channels had total visits, namely kompas.com (120,500,000), detik.com (140,000,000), and tempo.co (16,800,000) [3]. The large amount (volume) of online news that is spread in a short time (velocity) and the public's need to consume news in various references (variety) can affect people's lives [4]. Therefore, the government as the regulator and news agencies needs to monitor online news circulating quickly and accurately. Also, there are mass media companies, especially in the field of online news, researchers or analysts need various news data generated by various news sources which will be used for analysis purposes which eventually become a valuable view.

Based on these problems, 3 problems arise, namely (1) the large volume of online news that is spread, (2) the high speed or velocity of news dissemination, (3) the diversity or variety of formats from various reference sources, and (4) the need to analyze data. The first three points of these problems are part of the 3 main problems of Big Data that can be solved using a Data Lake that can run in real-time [5], [6]. The data lake itself can be a means to accommodate, dive into, and analyze circulating online news data.

The term data lake comes from James Dixon as founder and chief technology officer (CTO) at Pentaho. He described the data lake as water in its natural state, or it can be said that the data is stored in a large amount of raw format [7]. The data flows from sources to fill the lake, and users have access to the lake to inspect, sample, or even dive into it.

Thus, the responsibility of the data lake is to be central, high endurance storage for any type of data that may wish to be stored over multiple periods. This study proposes a data lake architectural design that accommodates large amounts of online news data from several online news channels.

Researchers will conduct literature studies, and analyze the flow of data lake architecture following online news. Furthermore, the researcher will use this architecture to combine and uniform the online news data structure from several online news channels and then stream it in real-time to fill the data lake. Finally, this data lake can be used by users to dive into and analyze both to monitor online news circulating in Indonesia and for analysis purposes to obtain news patterns, trends, and so on so that it can help companies develop their business strategies.

II. LITERATURE REVIEW

Recently, various systems exhibit similar properties but have different architectures. Therefore, we review architectures that enhance the processing associated with collecting data from heterogeneous data sources. Sarathkumar Rangarajan researcher Huai Liu et al. (2018) present a new data lake architecture to reduce data ingestion time and improve the accuracy of health analytics. All available data is digestible without ETL processing but requires proper metadata management. Because the data

ISSN: 2614-8404

lake without proper metadata management will make it a data swamp [8]. Sophisticated metadata management combines work with rapidly changing data structures, as well as sub-second request response on highly structured data [9].

Furthermore, researchers Mohamed Saifeddine Hadj Saasi et al. (2019) propose a Cognitive Internet of Things and big data architecture that combines a data warehouse, data lake, and heterogeneous data collection tools both in batch and stream processing. With the method integrated into the tool through algorithms can make the device think like a human and recognize what is in the data collected. Data Lake (DL) utilities as a data store in highlighting solutions to many critical challenges such as 3V requirements in big data. Data can be digested as is. Therefore, users can add DL with their native format. This leads to ensuring architectural scalability. Thus, DL can handle an unlimited amount of data and has the power of handling complex analysis quickly and efficiently [10].

Besides, Hassan Alrehamy and Coral Walker's (2018) research presents a data integration approach by combining data from various sources and building a unified view for users. Compared to the monolithic view of the single data model emphasized by the ETL process, the data lake is a more dynamic environment that relaxes data retrieval constraints and hinders data modeling and integration requirements to the next stage in the data cycle, resulting in an almost infinite potential to ingest and store multiple types of data irrespective of their source and schemas that often change, which are often unknown beforehand [11].

Most of them propose an architecture that starts with collecting data and then absorbing the data which only loads the data into a database table that is not completely clean, but this can slow down the process of data analysis. As they also let users have limited reviews without analyzing solutions to technical constraints of each architecture. Knowing what techniques and tools are suitable for the data flow process is important [10]. Therefore, this study will build a data lake flow architectural design that can combine and homogenize data structures from several sources with a real-time approach starting from the data absorption process. The architectural model that was built aims to ensure that every data that is streamed quickly into the data lake will be in the form of raw data that is clean, structured, and ready to be processed for real-time data analysis needs.

III. DATA LAKE CONCEPT

Data lakes are a new concept that is emerging and is increasingly being used to build and manage next-generation systems that can address the challenges of today's big data era. The concept of a data lake resembles water in its natural state flowing from a source to fill a lake, and users have access to the lake to inspect, take samples or even dive into it [7].

The data lake has been introduced as an architecture that can support a broader analysis of various types of data from various sources. Data lakes can store data regardless of size, schema, format, and complexity. Furthermore, the data lake can process, convert, and secure data, as well as make data can be processed with speed and value according to user needs [12].

With this capability, the data lake becomes a place for flexible and task-oriented data structuring only where and for what is needed. Data lakes can also be used by knowledge management, log management, business intelligence, or researchers to obtain information and calculation analysis to minimize costs, risks, provide benefits, and so on.

Besides, data lakes have several components, such as (1) collect data, which collects data from various sources and types of data, both structured, semi-structured, and unstructured. (2) Ingest data, which functions as a place for receiving data, and users can access the data contained therein. (3) Data processing is the ability to process raw data both batch and real-time so that it can be analyzed. (4) Data analysis, which is a component that performs data analysis to obtain analysis results for both user needs and new perspectives on business.

This concept was accepted as a way of describing large data sets in which the data schema and requirements were not defined until the data was queried. Thus, the data lake has several capabilities, namely [13]:

- Retrieves and stores raw data at low cost
- Stores many types of data in the same repository.
- Perform transformations on new data processing.
- Conduct a single subject analysis based on a very specific use case.

IV. DATA LAKE ARCHITECTURE PROPOSAL FOR ONLINE NEWS

The researcher proposes a data lake architecture for online news which can be seen in Figure 1. This data lake will contain data in the raw online news format and support data transformation. The data lake architecture has four layers which will be explained as follows. In the first layer, there are data sources, namely online news data sources that come from various online news channel sources, such as kompas.com, detik.com, and tempo.co. These online news channels have different structures for their stories. Furthermore, the second layer is the Data Ingestion Layer which consists of Data Extraction and Preprocessing and Messaging Queue.

This Data Extraction section will extract data from various existing sources. Then the Preprocessing section will combine and uniform the data structure into one standard form (structured data). This is necessary because the data lake source, even within the same domain, can be very heterogeneous in how data is organized, labeled and described (e.g., naming conventions for JSON keys, XML tags, or CSV headers), exhibiting considerable variation even for data with almost the same attributes [14]. Meanwhile, this section is carried out at the beginning before going to MongoDB so that any data that is streamed quickly into the data lake is in the form of raw data that is clean, structured, and ready to be processed in real-time.

ISSN: 2614-8404

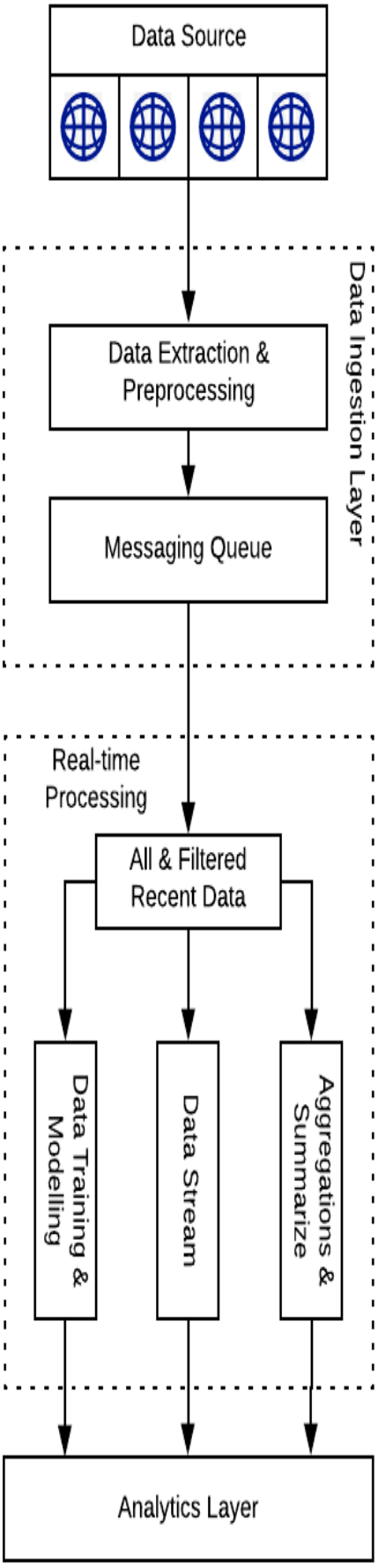

Figure. 1 Data Lake Architecture For Online News

After going through the data extraction and preprocessing, the Messaging Queue will be carried out, which is the part that regulates the distribution of data flow to the next process, namely to MongoDB. Researchers use MongoDB as a database to store data with short and historical data retention times. MongoDB tools can manage data from any structure without the burden of expensive data warehouse and without redesigning the database, no matter how often it changes [15].

Furthermore, the third layer is real-time processing. In the Real-time processing section, the stored data is separated into 2 types of data, namely short time retention data and historical data. Short time retention data is intended so that stored data can be processed quickly and have good relevance (for example, yesterday's economic news data does not affect the stock/bond market today) for real-time analysis needs. Then the aggregations and summary process can be done. This process is the process of collecting raw data and is presented in a summary format for further analysis. [16]. Also, aggregations and summaries are processed in real-time so that data movements can be easily monitored and analyzed, based on a certain time frame. Even though data aggregations and summaries are always moving/growing, overall, the data will not be too large. So that the benefits of using data aggregations and summaries in real-time can be felt maximally, without major risks/shortcomings.

Then in the part of the historical data that is stored, it is all data that has been extracted and streamed to the data lake. Data is stored in MongoDB permanently or forever without time retention. This data serves to view historical data, analyze data characteristics, and process data training and modeling that require complete data. Historical data is processed in real-time so that data has high update ability.

As for real-time processing, data stream processing can also be done. This process is responsible for consuming data in the message queue and processing data [17]. This data stream is directly streamed to the analysis section (for example, to input into the predictive analytics model to get real-time predictions) or display (such as the Twitter / Facebook / Instagram timeline which is always updated every time).

Furthermore, the fourth layer is the Analytics Layer which is the part where all the data that has been stored and processed in the Data Lake is used for analysis. Data lake supports analysis systems with storage and processing of all types of data [5]. The available data can be analyzed whatever is possible using the data. This is also the part where short-time and historical data retention times can work together to produce quality analyzes. For example, layer analysis will use training data and modeling results from historical data as a guideline to determine analysis/prediction results instantaneously based on input from data streams generated by short time retention data. The results of this analysis can be displayed in the form of reports, graphs, and others that can be used directly by the user.

V. RESULTS AND DISCUSSION

This section presents the experimental results of the use of the data lake architecture for online news proposed by the researcher. The collection of news data in this research

 ISSN: 2614-8404

is from January 1, 2020, to July 19, 2020. The news data consists of 55,599 data in the news category from kompas.com, 13,155 data with news categories from detik.com, and 10,431 data with national categories from tempo.co. Data sets from various sources will use field names defined in JSON format. The fields that have been defined are the "page" field containing the news URL page. The "news_id" field contains the initial 3 letters of the news channel such as KMP for kompas.com, DTK for detik.com, and TEM for tempo.co, followed by the date using the "ddmmyyyy" format, and followed by the 7 digits a serial number of the news on the day that.

Then the "news_title" field contains the online news title. Field "news_url" contains the URL address of the online news data. Field "news date" contains the format equivalent to date (dd / mm / yyyy) and time (hh: mm: ss), field "category" contains groups of online news types. The "subcategory" field contains more specialized categories of online news types. Next, the news_writer_code field contains the initials of the online news article author's name. The news_writer_name field contains the name of the online news writer. The news_editor_code field contains the initials of the online news editor. The news_editor_name field contains the name of the online news editor. The news_location field contains the location related to online news. The news_content field contains online news content, and the news_tag field contains online news labels. Meanwhile, some sample data from online news channels kompas.com, detik.com, and tempo.co are shown in Figures 2, 3, and 4.


```
{
"page": "https://news.kompas.com/search/",
"news_title": "Luhut Pastikan Sudah Koordinasi dengan Gubernur soal Larangan Mudik",
"news_url": "https://nasional.kompas.com/read/2020/04/21/21012111/luhut-pastikan-sudah-
"news_date": "21/04/2020, 21:01 WIB",
"news_category": "news",
"news_subcategory": "Nasional",
"news_writer_name": "Haryanti Puspa Sari",
"news_editor_name": "Kristian Erdianto",
"news_content": "JAKARTA, KOMPAS.com - Menteri Perhubungan ad interim Luhut Binsar Pand
"news_tag": ["Luhut Binsar Pandjaitan", "Larangan Mudik Berlaku 24 April 2020"]
}
```


Figure. 2 Sample data from the kompas.com channel


```
{
"page": "https://news.detik.com/berita/indeks?date=01%2F01%2F2020",
"news_title": "BPBD: Jabodetabek Diguyur Hujan Sedang-Lebat Hingga Dini Hari",
"news_url": "https://news.detik.com/berita/d-4843172/bpbd-jabodetabek-diguyur-
"news_date": "Rabu, 01 Jan 2020 23:36 WIB",
"news_category": "Berita",
"news_writer_code": "maa",
"news_writer_name": "Matius Alfons",
"news_editor_code": "aud",
"news_location": "Jakarta",
"news_content": "Jakarta - Badan Penanggulangan Bencana Daerah (BPBD) DKI Jaka
"news_tag": ["hujan", "banjir","bpbd"]
}
```


Figure. 3 Sample data from the detik.com channel


```
{
"page": "https://www.tempo.co/indeks/2020/01/01/nasional",
"news_title": "Malam Tahun Baru, Brimob Baku Tembak dengan KKB di Mimika",
"news_url": "https://nasional.tempo.co/read/1290089/malam-tahun-baru-brimob
"news_date": "Rabu, 1 Januari 2020 21:36 WIB",
"news_category": "nasional",
"news_subcategory": "hukum",
"news_writer_name": "Antara",
"news_editor_name": "Ninis Chairunnisa",
"news_content": "TEMPO.CO, Mimika - Kepala Kepolisian Resor Mimika Ajun Kom
"news_tag": ["Brimob", "KKB", "Malam Tahun Baru", "Mimika"]
}
```


Figure. 4 Sample data from the tempo.co channel

Sample data such as Figures 2, 3, and 4 are data that still require preprocessing. The process combines and uniform data structures. All data sources are made to have the same fields, otherwise they can be left blank (null). For example, the data source originating from the kompas.com and tempo.co channels does not have the news_writer_code and news_editor_code metadata fields so that it will generate 3 letters from the names of news authors and editors, then news_location is taken from news_content because there are no news location metadata. Also, on the detik.com channel, the news_editor_name field is left blank because there is no name of the online news editor. Furthermore, both the data from the kompas.com, tempo.co and detik.com channels in the news_date field are equated with the format (dd / mm / yyyy, hh: mm: ss), then the fields news_category, news_subcategory, news_writer_code, news_editor_code, news_location, news_tag is equivalent to lower case format.

After the data is processed and becomes clean and structured, it will be streamed to the data lake using MongoDB as the database. These data will be stored for both the short and long term. The results of several documents in the data lake in the form of JSON format for online news can be seen in Figures 5, 6, and 7.


```
{
"_id":{"$oid":"5f05afa92dcc90465655f8a9"},
"page":"https://news.kompas.com/search/",
"news_id":"KMP210420200000001",
"news_title":"Luhut Pastikan Sudah Koordinasi dengan Gubernur soal Larangan Mudik",
"news_url":"https://nasional.kompas.com/read/2020/04/21/21012111/luhut-pastikan-sud
"news_date":"21/04/2020, 21:01:00",
"news_category":"news",
"news_subcategory":"nasional",
"news_writer_code":"hps",
"news_writer_name":"Haryanti Puspa Sari",
"news_editor_code":"keo",
"news_editor_name":"Kristian Erdianto",
"news_location":"jakarta",
"news_content":"JAKARTA, KOMPAS.com - Menteri Perhubungan ad interim Luhut Binsar P
"news_tag":["luhut binsar pandjaitan","larangan mudik berlaku 24 april 2020"]
}
```


Figure. 5 Online news document data from the kompas.com channel

 ISSN: 2614-8404


```
{
"_id":{"$oid":"5f05afa92dcc90465655f8aa"},
"page":"https://news.detik.com/berita/indeks?date=01%2F01%2F2020",
"news_id":"DTK010120200000001",
"news_title":"BPBD: Jabodetabek Diguyur Hujan Sedang-Lebat Hingga Dini Hari",
"news_url":"https://news.detik.com/berita/d-4843172/bpbd-jabodetabek-diguyur-h
"news_date":"01/01/2020 23:36:00",
"news_category":"berita",
"news_subcategory":"berita",
"news_writer_code":"maa",
"news_writer_name":"Matius Alfons",
"news_editor_code":"aud",
"news_editor_name":"",
"news_location":"jakarta",
"news_content":"Jakarta - Badan Penanggulangan Bencana Daerah (BPBD) DKI Jakar
"news_tag":["hujan","banjir","bpbd"]
}
```


Figure. 6 Online news document data from the detik.com channel


```
{
"_id":{"$oid":"5f05afa92dcc90465655f8ab"},
"page":"https://www.tempo.co/indeks/2020/01/01/nasional",
"news_id":"TEM010120200000001",
"news_title":"Malam Tahun Baru, Brimob Baku Tembak dengan KKB di Mimika",
"news_url":"https://nasional.tempo.co/read/1290089/malam-tahun-baru-brimo
"news_date":"1/01/2020 21:36:00",
"news_category":"nasional",
"news_subcategory":"hukum",
"news_writer_code":"ant",
"news_writer_name":"Antara",
"news_editor_code":"age",
"news_editor_name":"Ninis Chairunnisa",
"news_location":"mimika",
"news_content":"TEMPO.CO, Mimika - Kepala Kepolisian Resor Mimika Ajun Ko
"news_tag":["brimob", "kkb", "malam tahun baru", "mimika"]
}
```


Figure. 7 Online news document data from the tempo.co channel

The data lake can later be used for data streams, aggregations, and summarizes, as well as training and modeling data, which can then be analyzed as needed. The results of the analysis are displayed in the form of reports, graphics, etc. which can be used directly by the user. These results can be used for government purposes in monitoring online news circulating in Indonesia. Besides, the results of the analysis can provide a business view of the analyzed news patterns so that it can assist the company in formulating its business strategy to minimize costs, risks, and provide benefits.

VI. CLOSING

This study conducted a literature study and analyzed the flow of data lake architecture following online news. Unlike previous studies, they propose an architecture that starts with collecting data. Then data absorption is carried out which only loads the data into a database table that is not completely clean, however, this can slow down the process during data analysis. Thus, the researcher built a data lake flow architectural design with a real-time approach starting from the data absorption process. The architectural model that was built aims to ensure that every data that is streamed quickly into the data lake will be in the form of raw data that is clean, structured, and ready to be processed for real-time data analysis needs. Furthermore, researchers used this architecture by using MongoDB as a database for both short-time and historical data retention. In the future, researchers intend to conduct aggregations, summaries, data streams, or analyze online news such as fake news classification. The results of the analysis can then be displayed in the form of reports, graphs, and others that can be used directly by the user. Finally, this data lake can be used by users to dive into and analyze both to monitor online news circulating in Indonesia and for analysis purposes to obtain news patterns, trends, and so on so that it can help companies develop their business strategies.

BIBLIOGRAPHY

[1] C. Juditha, "News Accuracy in Online Journalism (News of Alleged Corruption The Constitutional Court in Detiknews)," *J. Pekommas*, vol. 16, no. 3, pp. 145–154, 2013.

[2] Nurkinan, "Dampak Media Online Terhadap Perkembangan Media Konvensional," *J. Polit. Indones.*, vol. 2, no. 2, pp. 28–42, 2017.

[3] SimilarWeb Ltd, "SimilarWeb," *www.similarweb.com*, 2020. https://www.similarweb.com/top-websites/indonesia/category/news-and-media/ (accessed Jul. 01, 2020).

[4] D. S. Adhiarso, P. Utari, and Y. Slamet, "Pemberitaan Hoax di Media Online Ditinjau dari Konstruksi Berita dan Respons Netizen," *J. Ilmu Komun.*, vol. 15, no. 3, pp. 215–225, 2017.

[5] M. Chessell, D. Wolfson, and T. Vincent, "Architecting to Deliver Value From A Big Data and Hybrid Cloud Architecture," in *Software Architecture for Big Data and the Cloud*, 1st ed., I. Mistrik, R. Bahsoon, N. Ali, M. Heisel, and B. Maxim, Eds. Elsevier Inc., 2017, pp. 33–48.

[6] M. R. Llave, "Data lakes in business intelligence: Reporting from the trenches," *Procedia Comput. Sci.*, vol. 138, pp. 516–524, 2018, DOI: 10.1016/j.procs.2018.10.071.

[7] H. Fang, "Managing data lakes in big data era: What's a data lake and why has it became popular in the data management ecosystem," *IEEE Int. Conf. Cyber Technol. Autom. Control Intell. Syst.*, pp. 820–824, 2015, DOI: 10.1109/CYBER.2015.7288049.

[8] S. Rangarajan, H. Liu, H. Wang, and C. L. Wang, "Scalable Architecture for Personalized Healthcare Service Recommendation Using Big Data Lake," *Springer*, vol. 234, pp. 65–79, 2018, DOI: 10.1007/978-3-319-76587-7_5.

[9] F. Ravat and Y. Zhao, "Metadata Management for Data Lakes," *Springer*, vol. 1064, pp. 37–44, 2019, DOI: 10.1007/978-3-030-30278-8.

[10] M. S. Hadj Sassi, F. G. Jedidi, and L. C. Fourati, "A new architecture for cognitive internet of things and big data," *Procedia Comput. Sci.*, vol. 159, pp. 534–543, 2019, DOI: 10.1016/j.procs.2019.09.208.

[11] H. Alrehamy and C. Walker, "SemLinker: automating big data integration for casual users," *J. Big Data*, vol. 5, no. 1, 2018, DOI: 10.1186/s40537-018-0123-x.

[12] N. Miloslavskaya and A. Tolstoy, "Big Data, Fast

ISSN: 2614-8404

Data, and Data Lake Concepts," *Procedia Comput. Sci.*, vol. 88, pp. 300–305, 2016, DOI: 10.1016/j.procs.2016.07.439.

[13] R. Benaissa, F. Benhammadi, O. Boussaid, and A. Mokhtari, "Clustering Approach for Data Lake Based on Medoid's Ranking Strategy," *Springer*, vol. 50, pp. 250–260, 2019, DOI: 10.1007/978-3-319-98352-3.

[14] A. Abelló, "Big data design," *Dol. Proc. ACM Int. Work. Data Warehouse. Ol.*, vol. 23-Oct-201, pp. 35–38, 2015, DOI: 10.1145/2811222.2811235.

[15] H. Abbes and F. Gargouri, "Big Data Integration: A MongoDB Database and Modular Ontologies based Approach," *Procedia Comput. Sci.*, vol. 96, pp. 446–455, 2016, DOI: 10.1016/j.procs.2016.08.099.

[16] T. Wen, "Data Aggregation," *Encylopedia of Big Data*. Springer, Cham, 2020, DOI: https://doi.org/10.1007/978-3-319-32001-4.

[17] W. Jiang, L. G. Xu, H. B. Hu, and Y. Ma, "Improvement design for distributed real-time stream processing systems," *J. Electron. Sci. Technol.*, vol. 17, no. 1, pp. 3–12, 2019, DOI: 10.11989/JEST.1674-862X.80904011.