Big Data Modern Stack for District Government

Mutiara Auliya[1(✉)], Abdul Aziz[2], and Wahyu Nurharjadmo[3]

[1] Administrative Management, Vocational School of Sebelas Maret University, Surakarta, Indonesia
mutiaraauliya@staff.uns.ac.id
[2] Informatics Engineering, Vocational School of Sebelas Maret University, Surakarta, Indonesia
[3] Department of Public Administration, Sebelas Maret University, Surakarta, Indonesia

Abstract. Data is the new oil for all circles, including the district government. The more information systems are used, the more data will be collected both in structured and unstructured data, which is often called Big Data. The obstacle faced by the district government in handling Big Data is that the data is stored in different stores. Therefore the analysis process becomes complicated. For this reason, it is time for the district government to implement the Big Data Modern Stack that refers to a collection of technologies that comprise a cloud-native data platform. Big Data Modern Stack will combine cloud technology with on-premises technology in the district government system to store all the data collected in its data warehouse. This research explores how big data modern stack could be implemented for the district data system. We propose ELT (Extract, Load, Transform) using Fivetran for data ingestion from various forms of stored data, then unite all data into a data warehouse to be displayed in a Business Intelligence Tool. The paper concludes that this modern big data stack approach can unite some of the data owned by the district government so that it can be used to support the decision-making process for related stakeholders.

Keywords: Big Data · Modern Stack Infrastructure · District Government

1 Introduction

Nowadays, data is the new oil for all circles. The more information systems that are used, the more data will be collected. The more developed an agency or organization, the more technology is used to facilitate work and decision-making. Agencies will generate vast amounts of structured and unstructured data referred to as Big Data. Data is growing very fast, complex, and quite massive. Big Data holds huge volumes of sets of data. There are various types, quite large and generated from several sources [1]. Big Data is something that cannot be processed using conventional database systems. The data is too big, complex, massive, and does not fit into the single database format. The main characteristics of Big Data include three things – commonly abbreviated as 3V – volume, velocity, and variety. Volume is related to the amount of data that must be managed at a super large size. Velocity relates to the speed of data processing which must keep pace with the rapid growth in data. At the same time, variety refers to the characteristics of

© The Author(s) 2023
D. Lukitasari et al. (Eds.): INCODEC 2021, ASSEHR 702, pp. 103–109, 2023.
https://doi.org/10.2991/978-2-494069-53-4_12

very diverse data sources, both from structured databases and unstructured data. Big Data is usually structured, unstructured and semistructured for the type of data [2].

Various information and documentation in big data are sourced from various channels, including social media, sensors, video surveillance, and smart grids. All of these data channels lead to big data technology. The abundance and variety of data can be utilized by all parties, given the variety of information and the complexity of the data according to the needs of each party. For example, the data collected in big data can be used to make a strategic policy for the government. This policy is matched with the accumulation of data or information obtained. The pile of data or information should be one of the things that can be processed and used to make decisions [3]. Big Data consists of several phases: Big Data generation, Big Data Acquisition, Big Data Storage, and Big Data Analysis. In the Data Analysis, a method is needed to extract, transform, load into some model to extract the value. Such as Descriptive Analytics, Predictive Analytics, and Prescriptive Analytics for decision-making [4].

In addition, to the different types of data types, one of the obstacles for government is how to collect all the different resources into one container, such as a data warehouse. Because currently, government information systems are usually not centralized [5]. The exchange of data in real-time is also influential. One method of ingestion data that has been widely used is Extract, Transform, Load. This method is widely applied in various sectors of trade, economy, and government. This Extract, Transform, Load process is a process of raw data then put together in the data warehouse. ETL aims to collect, filter, process, and combine relevant data from various sources/databases in a data warehouse. This ETL process is fundamental because it plays an essential role in the quality of the data in the data warehouse to be used for data analysis processes [6]. One of the weaknesses of ETL is that it requires human resources who understand the various ETL tools used, such as *Pentaho Data Integration, Talend Data Integration, Informatice Data Integration*, and *Microsoft SQL Server Integration Services* [7] [8]. The employees in the district government lack knowledge about the ETL process in the data warehouse.

This paper aims to design and implement the Big Data Modern Stack, a collection of technologies that comprise a cloud-native data platform. Big Data Modern Stack will combine cloud technology with on-premises technology in the district government system. The designed system accommodates ELT (Extract, Load, Transform) using Fivetran for data ingestion from various forms of stored data, then unites all data into a data warehouse to be displayed in a Google Data Studio for business analysis tools.

The main contributions of this paper are as follows: First, the designed Big Data Modern Stack using on-premises and cloud technology using structured, semistructured, and unstructured data. Second, The system provides solutions to the top-level stakeholders to know the big data technology without having to worry about the state of its human resources. Third, the proposed system could be implemented in other districts government need a decision support system for big data. Forth, since the research that discusses the Big Data Modern Stack is still few, especially for the government, this research is used as a reference for future research.

The paper is organized as follows. Section 1 has described the relationship between big data and district government. Section 2 describes some corresponding research of

big data in government and it explains the proposed method. The implementation and experimental results are presented in Sect. 4. Finally, in Sect. 5 we conclude the research.

2 Literature Review

The Modern Data Stack (MDS) is a suite of tools used for data integration. The goal is to analyze your business's data to uncover new areas of opportunity and improve efficiency proactively. Modern data stack is hosted in the cloud and requires trim technical configuration by the user [9]. The Modern Data Stack primary infrastructure architecture includes things like moving from on-premises to cloud or working together and changing from Extract, Transform, Load (ETL) to Extract, Load, Transform (ELT) processes [10].

Modern Data Stack consists of 4 parts: data sources, ingestion, data warehousing, data transformation, data visualization and analysis, data syncing, and automation. Modern Data Stacks usually do not use the Extract, Transform, Load (ETL) process but instead use Extract, Load, Transform (ELT). Due to the limited processing power in legacy data warehouses, data engineers wrote transformation jobs before loading data in, leading to the term ETL (Extract-Transform-Load). Now, with the advancement of high-performing cloud-based columnar data warehouses, data engineers can run petabyte-scale queries in minutes. A Modern Data Stack can provide and start loading data into the data warehouse in minutes (ELT, Extract-Load-Transform). Analysts no longer need to rely on engineers to transform the data [11].

Modern Data Stack that is currently developing combines several infrastructures that become one unit to form a modern big data architecture. For ingestion data, you can use Fivetran, Stitch, for data warehouse, Snowflake, Bigquery, Redshift, and the ELT process; you can use dbt. Furthermore, for business intelligence software, you can use Looker, Google Data Studio, Metabase, Tableau, PowerBI [7].

3 Proposed Method

The proposed method can be seen in Fig. 1. In the method, there are data sources.

All of the sources will be added to the data ingestion process. Both structured and unstructured data will all be included in the ingestion data. After the data is entered, it will be processed to the data warehouse. Then the latter will be analyzed in BI Tools. As is shown in Fig. 2, the architecture of Modern Data Stack was presented. All data will be processed in data sources, including the MySQL database, which is widely used in the district government, several documents such as excel, csv, or google spreadsheet.

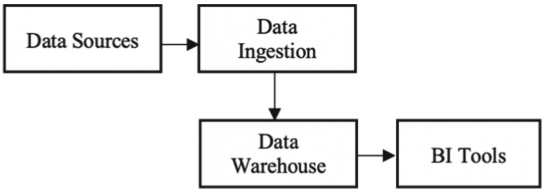

Fig. 1. The proposed method.

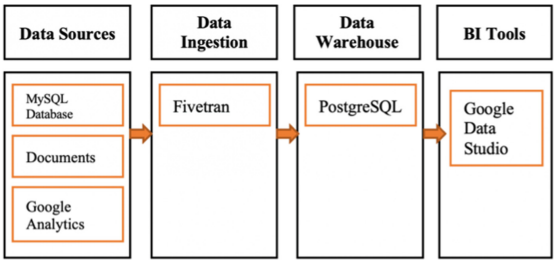

Fig. 2. The architecture of Modern Data Stack.

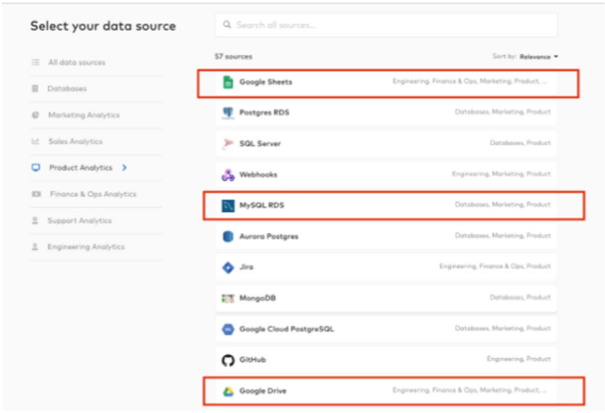

Fig. 3. Data Ingestion using Fivetran.

Then some third-party sources such as Google Analytics. All data will be connected to the data warehouse through the data ingestion process. For ingestion data, the author uses Fivetran [12].

Fivetran is used for how data gets moved and normalized from your data source to your data storage. In Fivetran, the ELT process can be carried out before entering the data warehouse. After all the data is collected, the data will be entered into BI Tools to make a dashboard for decision-making purposes.

4 Implementation and Results

This section will be discussed about the implementation of the proposed method with the analysis and system design. The structured data comes from MySQL databases, while unstructured data comes from the csv file and Google Spreadsheet. In Fivetran, we also connect Google Drive for easy entry of data. Figure 3 shows the data ingestion using Fivetran.

It seems to be a lot that can be connected to Fivetran in order to get into the data warehouse. First, please select Google Sheets, MySQL RDS and Google Drive and then sync with your account. Next, prepare your PostgreSQL data warehouse. Then sync

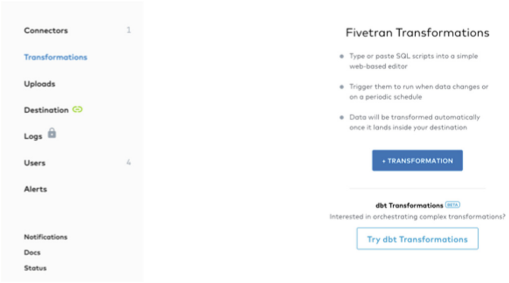

Fig. 4. Transformation Process in Fivetran.

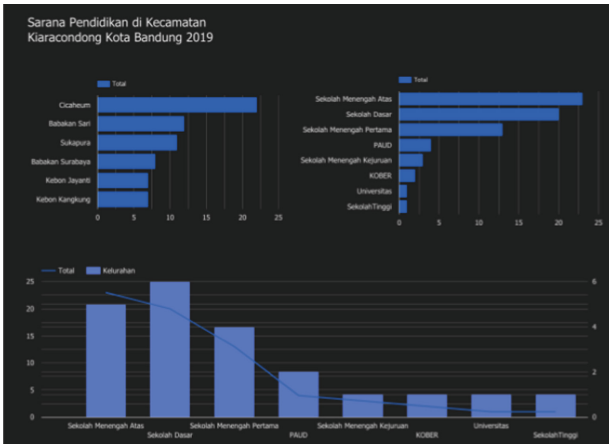

Fig. 5. Education Facility Dashboard in Kecamatan Kiaracondong, Kota Bandung 2019.

username, password, and port to Fivetran. So that Fivetran can directly enter all data into the data warehouse. Wait until the synchronization process is complete. Then, the Extract Load Transform process can be done on Fivetran (Fig. 4).

After the data is successfully entered into the data warehouse, then switch to BI Tools. For Google Data Studio, there is no need to install specific software. For BI Tools, we use Google Data Studio for displaying the data of each district government. As is shown in Fig. 5, we can make the dashboard for the district government stakeholders more interactive and attractive.

5 Conclusion

This paper has explored the modern data stack for district government. We use Big Data Modern Data Stack from the data source, data ingestion, data warehouse, and BI Tools. The Big Data Modern Stack approach will help the district government reduce difficulties associated with traditional database infrastructure and fix the problem related to the lack of employee knowledge in Big Data related. Moreover, Fivetran is a data ingestion tool

based on the cloud, and we should pay as we go. So, it depends on the amount of the data. However, we think it will not be a problem. Because the government usually has the funds for the development of its IT infrastructure. For future work, we can try another data warehouse approach using dbt for more complex ELT and using Snowflake.

References

1. J. Patel, 'An Effective and Scalable Data Modeling for Enterprise Big Data Platform,' in *2019 IEEE International Conference on Big Data (Big Data)*, Dec. 2019, pp. 2691–2697. DOI: https://doi.org/10.1109/BigData47090.2019.9005614.
2. C. K. Leung, Y. Chen, C. S. H. Hoi, S. Shang, and A. Cuzzocrea, 'Machine Learning and OLAP on Big COVID-19 Data', in *2020 IEEE International Conference on Big Data (Big Data)*, Dec. 2020, pp. 5118–5127. DOI: https://doi.org/10.1109/BigData50022.2020.9378407.
3. Al-Badi Ali, A. Tarhini, and A. I. Khan, 'Exploring Big Data Governance Frameworks', *Procedia Comput. Sci. Sci.*, vol. 141, pp. 271–277, 2018, DOI: https://doi.org/10.1016/j.procs.2018.10.181.
4. S. Malhotra, M. N. Doja, B. Alam, and M. Alam, 'Bigdata analysis and comparison of bigdata analytic approaches, in *2017 International Conference on Computing, Communication, and Automation (ICCCA)*, May 2017, pp. 309–314. DOI: https://doi.org/10.1109/CCAA.2017.8229821.
5. J. Ju, L. Liu, and Y. Feng, 'Citizen-centered big data analysis-driven governance intelligence framework for smart cities,' *Telecommun. Policy*, vol. 42, no. 10, pp. 881–896, Nov. 2018, doi: https://doi.org/10.1016/j.telpol.2018.01.003.
6. M. J. Denney, D. M. Long, M. G. Armistead, J. L. Anderson, and B. N. Conway, 'Validating the extract, transform, load process used to populate a large clinical research database,' *Int. J. Med. Inf.*, vol. 94, pp. 271–274, Oct. 2016, doi: https://doi.org/10.1016/j.ijmedinf.2016.07.009.
7. M. Hendayun, E. Yulianto, J. F. Rusdi, A. Setiawan, and B. Ilman, 'Extract transform load process in banking reporting system', *MethodsX*, vol. 8, p. 101260, 2021, doi: https://doi.org/10.1016/j.mex.2021.101260
8. M. Souibgui, F. Atigui, S. Zammali, S. Cherfi, and S. B. Yahia, 'Data quality in ETL process: A preliminary study, *Procedia Comput. Sci.*, vol. 159, pp. 676–687, 2019, DOI: https://doi.org/10.1016/j.procs.2019.09.223.
9. 'The Modern Data Stack: Past, Present, and Future,' *dbt blog*, Dec. 01, 2020. https://blog.getdbt.com/future-of-the-modern-data-stack/ (accessed Oct. 30, 2021).
10. M.-H. Su, T.-H. Yang, W.-H. Lin, and C.-H. Wu, 'Answer segmentation for question answering using latent Dirichlet allocation and delta Bayesian information criterion, in *2016 International Conference on Orange Technologies (ICOT)*, Dec. 2016, pp. 9–12. DOI: https://doi.org/10.1109/ICOT.2016.8278967.
11. 'The Modern Data Stack (updated for 2021)', *Metabase.* https://www.metabase.com/blog/The-Modern-Data-Stack/ (accessed Oct. 30, 2021).
12. 'Fivetran Inc'.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.