Data Warehouse Analysis and Design based on Research and Service Standards

Lasmedi Afuan, Nurul Hidayat, Dadang Iskandar, Arief Kelik Nugroho, Bangun Wijayanto, Ana Romadhona Yasifa
Informatics, Engeneering Faculty, Universitas Jenderal Soedirman, Purwokerto, Central Java, Indonesia

***Abstract*—Data are not easy to organize, especially if the data are big in quantity and stored manually and in a non-computerized way. Therefore, in the last few years, many organizations or companies used information systems to help their activities organize and manage the data. Universitas Jenderal Soedirman (UNSOED) is a state college that has existed for a long time and has many Study Programs and Faculties, including the Faculty of Engineering. Data organization in UNSOED is mainly performed through computerization. However, retrieving data needs to be improved because UNSOED has various information systems, and the data produced keeps increasing over time. The data have yet to be evaluated following the needs of Tri Dharma, with indicators of achievement as expressed in the Regulation of Minister of Education and Culture (PERMENDIKBUD) on the National Standard of Higher Education (SNDIKTI). Data warehouse technology can be applied to storing, collecting, and processing media within a specific time from various data sources. The data processing results in the data warehouse are later displayed using the tools Knowage which may help the executives of the Faculty of Engineering make a decision and monitor the businesses, mainly about research and service, by the society of academicians in the Faculty of Engineering regularly from time to time.**

***Keywords*—Data warehouse; knowage; SNDIKTI; UNSOED**

I. Introduction

In this digital era, any information can be easily acquired by anyone. In this era, data are essential and valuable, and any meaningful information can be retrieved for use from there, especially for organizations or companies that deal with data daily. Important information can be produced if the data in possession are processed appropriately. In addition, past data in possession can provide general and other important information that an organization or company may consider in performing their business.

The urgency of an organization or company's data storing keeps increasing, given that the longer an organization or company has been established or developing, the more data they produce. If the data produced are not well managed or stored, the organization or company will find it difficult to retrieve any critical information that it will take a long time to analyze the data, and applying a data management or storing system to the organization or company will take a long time since the data produced will be abundant and relations between data will be unclear. There will be possible duplicate data, requiring cross-checking for their validity.

UNSOED is one of the State Colleges in Indonesia established in 1963 and today has 12 Faculties, including the Faculty of Engineering consisting of five Study Programs: Electrical Engineering, Civil Engineering, Geological Engineering, Informatics, and Industrial Engineering [1]. The Faculty of Engineering took a long time to establish. This started in 2000 with only the Electrical Engineering and Civil Engineering study programs under the Undergraduate Program of Engineering, and finally, in 2014, the Faculty of Engineering was established [2].

In its implementation, Colleges, Faculties, and Study Programs must follow and comply with the existing standards and regulations issued by the Ministry of Education, Culture, Research, and Technology, including the Regulation of Minister of Education and Culture of the Republic of Indonesia Number 3 of 2020 on the National Standard of Higher Education (SNDIKTI), for Colleges to achieve the quality of learning, research and public service when they have exceeded the criteria defined in the National Standards of Higher Education [3].

Fulfilling the whole conditions or indicators in the SNDIKTI requires various data, especially related to Research and Public Service. In UNSOED, an institution manages the whole Research and Service activities performed by the society of academicians, the Research, and the Public Service Institution. In the data recording process, UNSOED has an information system, SINELITABMAS (Research and Public Service Information System) [4], in which the data are interconnected with other systems, including SIHURA (Human Resources Administration System) and SISTER (Integrated Resources Information System) that contain data of lecturers and educational workers.

The data collection, not only to satisfy SNDIKTI but also to satisfy other standardization such as accreditation or quality assurance, still needs to be done and takes a relatively long time for the UNSOED. The reason is that it has too many information systems. However, the data produced still needs to be evaluated following the needs of Tri Dharma Perguruan Tinggi (three pillars of higher education). Therefore, a working instrument is needed to process and store large-scale data from which a variety of important information can be displayed for the executives' use as an instrument and tool of evaluation in decision making.

By implementing a data warehouse and using Business Intelligent tools of Knowage, any data needed will be accessed more quickly, and they will be interconnected with each other. Besides, since the data warehouse is historical, it can be used in the long run. Previous data can be used for analysis and reporting to help executives in management decision-making.

II. Research Method

This research was conducted in phases by applying the Data Warehouse Life Cycle development method as follows [5]-[10]:

1) Literature study: The initial phase was conducted by learning some literatures such as textbooks and journals on data warehouse and, in addition, learning the content of Regulation of Minister of Education and Culture of the Republic of Indonesia Number 3 of 2020 on the National Standard of Higher Education for indicators to use.

2) Design: The initial process was interview with the executives of the Faculty of Engineering UNSOED to examine the kinds of data in possession and a business analysis on the process of and need for data was conducted, and from results of which, what data and where from the data were to be used were derived, were determined. Besides, database designing started from inter-table scheme, designing of dimension table and fact table along with attributes and sizes.

3) Prototype: This process was conducted using the results of designing and the need for data in the previous phase into data visualization or presentation dashboard designing. The prototype was made to adjust and describe user's needs. In case of shortcoming, developer can return to early phase and return to the prototype phase when added needs have been obtained. Second repetition in this phase can keep occurring until final prototype meets users all needs.

4) Deploy: Data presentation dashboard prototype agreed upon was then built into the environment or the tools of production used. In addition, integrated data processing was also conducted in this phase that completely processed data would be immediately stored into the data warehouse environment built.

5) Operate: In this phase, the data warehouse built was operated and tested. Test was conducted only on the developer part (Black Box Testing) and the operation in this phase could be observing suitability of information displayed, inspecting data processing performed, and other operation.

6) Enhance: In this final phase, some aspects lacking in the previous phases were modified and added covering physical components, operating process and data management, and logical scheme designing. One of the corrections included adding one dimension data with dummy data in order to produce information that was more suitable to the content of SNDIKTI.

III. Result And Discussion

A. Design

The design phase was conducted to examine the data needed for use. In this research, the needed data were collected through learning the document SNDIKTI No. 3 of 2020 and reconstructing the model or determining the format of data to be applied, and interview with the executives of the Faculty of Engineering UNSOED. The result of model reconstruction shown in Table I.

TABLE I. Model Reconstruction Based on SNDIKTI

Aspect in SNDIKTI	Description
Research Result	Research result is not confidential, does not disturb and/or does not harm public or national interest, must be disseminated through seminar, publication, patent, and/or other means that can be used to deliver the research results to the public.
Research Funding and Financing	College is obliged to provide fund for internal research and research management fund. Funding can also be derived from the government, cooperation with other domestic of foreign institutions, or fund from the public.

From the results of model reconstruction and interview, database was designed consisting of conceptual design, logical design, and physical design processes.

1) Conceptual design: In this phase, data were determined as needed, from process business identification and determining from which data were to acquire. The result of process business identification and some data mappings conducted can be seen on Table II.

TABLE II. Result of Process Business Identification

Process Business	Description	Function Involved
Research	Is business process related to research, from registration and submission of research proposal, selection, socialization, monitoring and evaluation, seminar of research results, and research publication.	LPPM
Public Service	Is business process related to service from registration and submission of service proposal, selection, socialization, monitoring and evaluation, seminar of public service results, and service publication.	LPPM

Research data could be acquired from some sources, such as data of lecturers from SISTER (Integrated Resources Information System) and data file analitik monev in which the data were contained in Excel file, data of study program could be generated from file of research data of PKM, research data could be acquired from file of research data of PKM and file of research data, data of types of publication and activity categories could be generated from file of publication data. The Mapping of Data Sources in Research Standards show in Fig. 1.

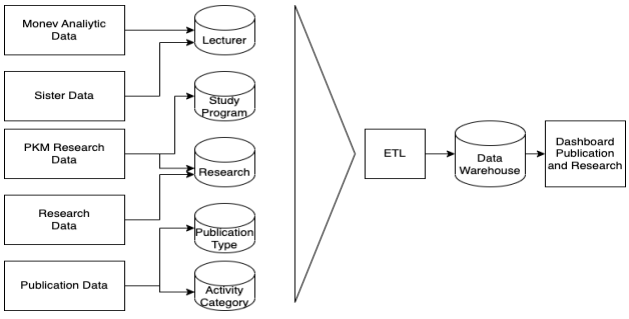

Fig. 1. Mapping of Data Sources in Research Standards.

After determining business process and data mapping, grains were determined to be fact table. Some grains that could later be taken as fact table shown in Table III.

TABLE III. GRAIN DETERMINATION

Grain	Description	Related Business Process
Research Transaction	Number of researches conducted along with publication and funding received	Research
Service Transaction	Number of public services conducted along with publication and funding received	Public service

The next phase was determining dimensions to be used pursuant to the predetermined grains. The description of dimension table to be made shown in Table IV.

TABLE IV. DIMENSION TABLE DETERMINATION

Dimension	Description	Grain
Lecturer	Data of lecturers were taken from data of SISTER in Excel file and from data of Excel file Analitik Monev	Research and Service Transactions
Researcher's Role	Information of researcher's role from Excel file	Research and Service Transactions
Time	Information of data of time including year, date, month, day, and semester generated from 2016 to present	Research and Service Transactions
Activity Category	Existing activity categories from data of research and service sources	Research and Service Transactions
Study Program	Data of study program	Research and Service Transactions
Type of SKIM	Data of the existing types of SKIM from data of research and service sources	Research and Service Transactions
Type of Publication	Data of the existing types of from data of research and service sources	Research and Service Transactions
Source of Dana	Information of source of fund of each research and service	Research and Service Transactions

After dimensions were determined, fact table was designed along with corresponding content of dimension. The fact table design produced can be seen on Table V.

2) Logical design: In this phase, dimensional modeling was made for dimension tables and facts made. The scheme used was star scheme which was a simple and easy-to-understand scheme. A dimensional modeling with star scheme designed to integrate data in data warehouse are shown in Fig. 2.

3) Physical design: In this phase, all dimension table and fact table designs were analyzed for metadata, covering information on the name of the table, detail of contents of the table such as a primary key or in data warehouse commonly known as a surrogate key, name of the field, type of field, size of the field, and source of data.

TABLE V. FACT TABLE DESIGNING

Fact	Description	Content
Fact_penelitianpublikasi	Facts of published research cover number of researches and funding obtained	Judul, biayadisetujui, dim_dosen_scd, dim_peranpeneliti, dim_waktu, dim_kategorikegiatan, dim_prodi, dim_skim, dim_jenispub, dim_sumberdana.
Fact_pengabdian	Facts of published service covers number of researches and amount of funding obtained	Judul, biayadisetujui, dim_dosen_scd, dim_peranpeneliti, dim_waktu, dim_kategorikegiatan, dim_prodi, dim_skim, dim_jenispub, dim_sumberdana.

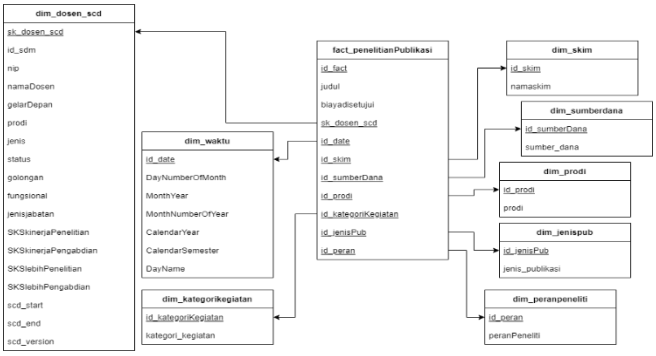

Fig. 2. Star Schema 'Fact_penelitianPublikasi'.

B. Prototype

In the prototype phase, the dashboard was designed to depict how data or information was visualized with various graphical forms as needed. This was designed using Google Spreadsheets as the tools of visualization and using some dummy data for graphical display to be easily understood by both developer and executive. Some designs of the dashboard regarding the research are shown in Fig. 3.

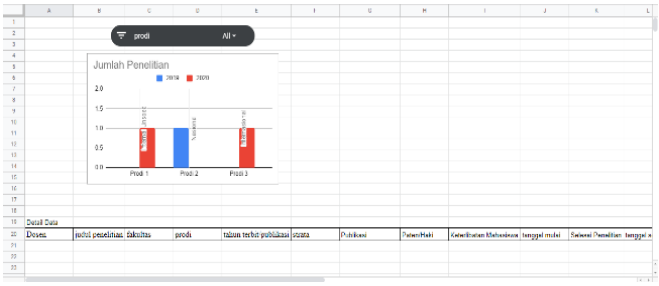

Fig. 3. Research Dashboard Design Aspects of Research Results.

Based on the model reconstruction, in the research result aspect, information of publication along with the respective number by study program in certain years could be displayed.

C. Deploy

The next phase was building data warehouse along with data visualization or presentation in the development environment used. In this research, software Talend Open Studio was used as the instrument in building ETL process

and software Knowage as the instrument in building data visualization or presentation.

ETL process started with storing data in source data that would be used in staging area and in this early phase data transformation processing was still not performed [11]. If the whole source data to be used have been stored in staging area, dimension table or fact table could be formed out of the data. Below is ETL process in building dimension table of lecturer:

1) Extract: Lecturer data stored were retrieved from data of staging area 'stg_sdm' and 'stg_dosen' in which 'stg_dosen' served as lookup while data 'stg_sdm' served as main data. Out of the two data staging, left outer join was performed on attribute of data with the same value for the data to complete each other (see Fig. 4).

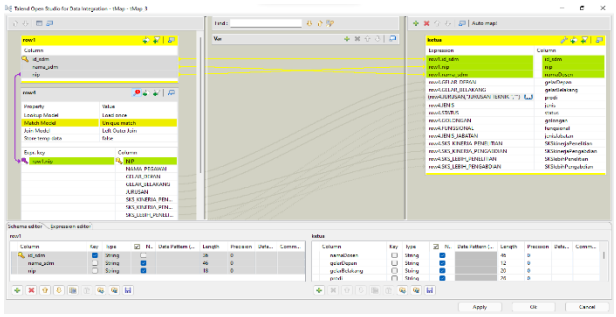

Fig. 4. Fields or Data Attributes used in Compiling Tables 'dim_dosen_scd'.

2) Transform: The attributes of data to be stored into lecturer dimension table were selected. The attribute 'JURUSAN' on 'stg_dosen' or on *row4* would be changed for more effective writing pursuant to agreed format. For example, regarding data change, data that were previously 'JURUSAN TEKNIK INFORMATIKA' would be changed and stored into 'INFORMATIKA'. *Loading*. The data were forwarded into target database or database 'dawer' such as the configuration shown in Fig. 5.

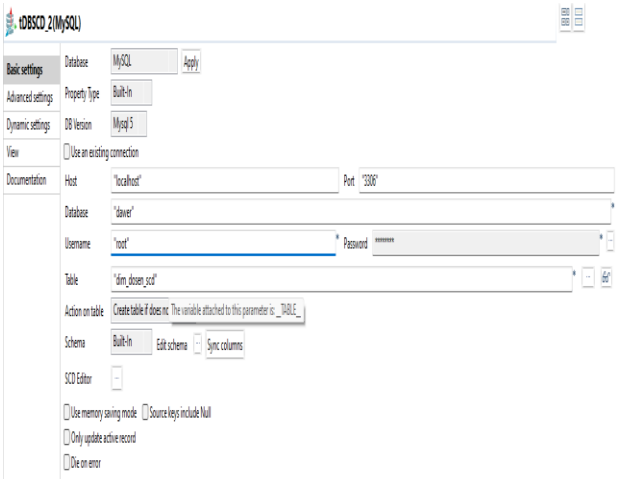

Fig. 5. Output Database Configuration for Lecturer Dimension Table.

Since dimension table applied SCD concept, further setting was needed on each attribute pursuant to the type of SCD used.

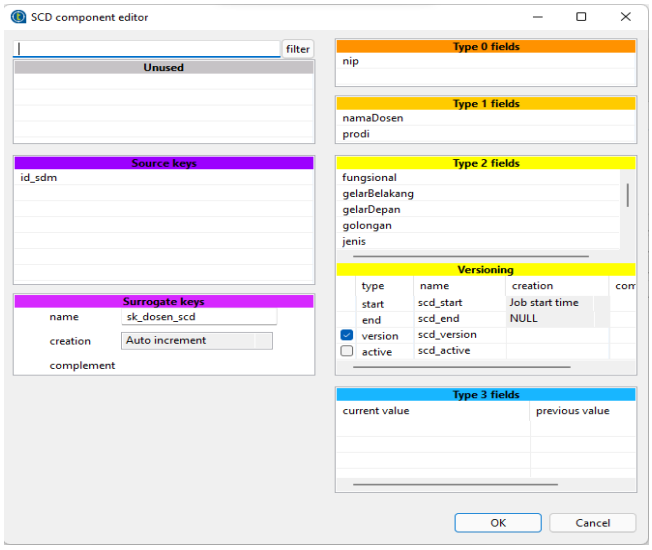

Fig. 6. SCD Component Settings.

Based on Fig. 6 above, 'id_sdm' was set to be the key attribute on the data source, attribute 'nip' was set to type 0 so that data could not change if the data were input, 'namaDosen' and 'prodi' were set to type 1 thus data change might occur by overwriting previous data. Other attributes were set to type 2 for changes to the other attributes to form a new record in the table. In addition, a key attribute or the surrogate key was needed for each record with setting to auto-increment, and this SCD concept would form additional attributes such as start and end dates record was a valid and higher version of the record equaled the latest, and that would be used in the subsequent data processing.

After dimension table was formed, ETL would be performed for fact table [12]-[13], one of which was service fact table are shown in Fig. 7.

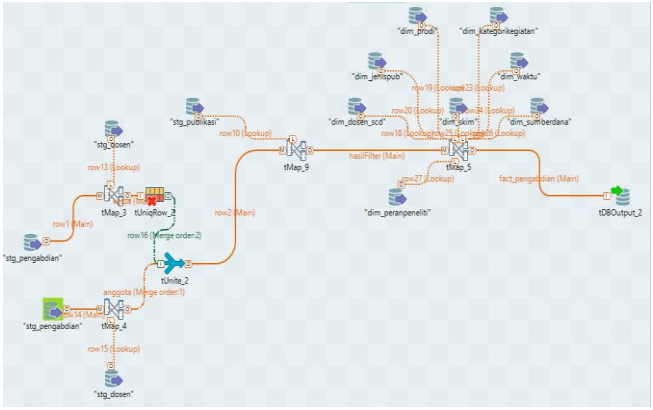

Fig. 7. ETL Process Forming Fact Table 'Fact_pengabdian'.

3) Extract: This phase started with retrieving and selecting data from the staging area of service data, the staging area of the lecturer, and the staging area of publication. Afterward, data from the staging area selected were connected with a dimension table designed in dimensional modeling as lookup for data in the staging area by applying Left Outer Join as a join model.

4) Tranform: 'Datajudul' and 'biayadisetujui' from the staging area and all data of the surrogate key in the dimension table were selected or retrieved for use in the fact table. In addition, in this phase surrogate key was made for the fact table, and the field biayadisetujui was conditioned in case of null data. They would be changed to 0.

5) Loading: Valid data would then be stored in the table 'fact_pengabdian' in dawer database.

The results of ETL processed stored in data warehouse were then connected with Knowage to present data that were processed into the form of dashboard, OLAP, and report. Below are some outputs of data presented made using research data. The result show in Fig. 8 - 11.

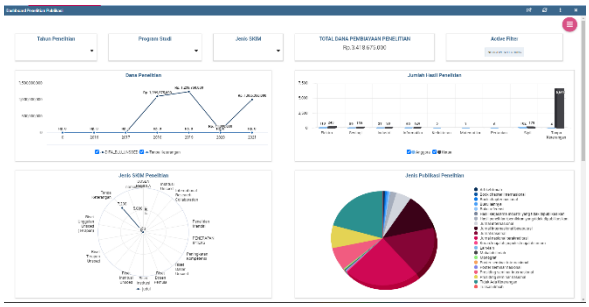

Fig. 8. Presentation of Data in Dashboard Form for Research Data Publications 1.

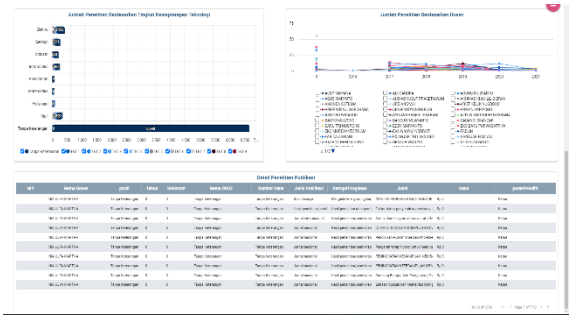

Fig. 9. Presentation of Data in Dashboard Form for Research Data Publications 2.

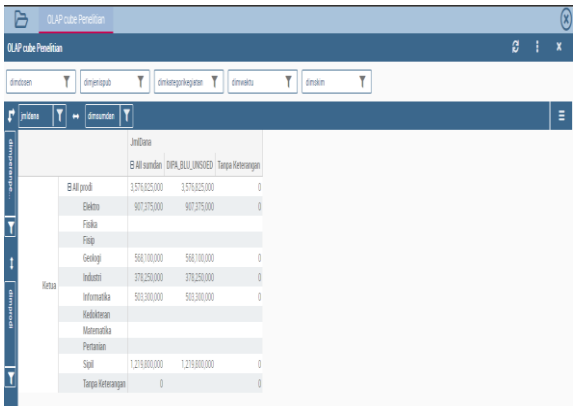

Fig. 10. Data Presentation in OLAP Form for Lecturer Research Financing Data.

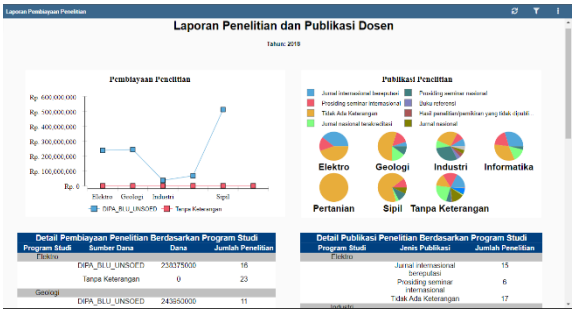

Fig. 11. Presentation of Data in the Form of Reports for Lecturer Research and Publication.

All forms of data presented above can be used by the executives to monitor the lecturer's performance of research and publication as needed.

D. Operate

After Data Warehouse and data presentation had been made, a test was performed using Black Box Testing by developer [14]-[15]. The results of black box testing shown in Table VI.

TABLE VI. RESULTS OF BLACK BOX TESTING

Test Scenario	Test Case	Output	Test Result
Data completeness	Input or load all data from data source into data warehouse	Number of data stored in data warehouse is equal to number of data in data source	Valid
Data transformation	Transform the data and compare data resulted from transformation with data in data source	Value of data resulted from transformation remains valid pursuant to data source	Valid
Data quality	Input false or invalid data into ETL processing	Data are not input into data warehouse and data value was changed to null, 0, or no remark if the data input were empty.	Valid
	Input the same data into ETL processing	Data do not increase and the value of data remains the same	Valid
	Specifically on dimension table of lecturer, data update was performed: 1. NIP update 2. Name update 3. Title update	Data are updated pursuant to the type of SCD applied 1. Data are not updated except data before update were null 2. Data are updated in the record or in the same key 3. Data are updated in the record or in the new key	Valid
Scalability and performance	Perform the whole ETL process	Not take too long time in processing	Valid
	Run dashboard, OLAP, and report	Aggregated data displayed in data presentation are valid	Invalid

Calculation was performed based on the test results in accordance with the number of valid scenarios. The calculation using Eq. (1).

$$evaluation\ result = \frac{Number\ of\ valid\ data}{Total\ test}\ x\ 100\% \quad (1)$$

The results of black box testing above shows that data warehouse made shows valid results of 86%, that it was necessary to further correct the testing part with invalidity.

E. Enhance

In this phase, ETL processing and data presentation were corrected pursuant to the test results on previous phase or based on other suggestion. The corrections and additions include:

1) Correct Aggregation Data Displayed in Data Presentation.

2) Add data of Research and Service Technology Readiness Level (TKT).

3) Present TKT data on the dashboard of research publication and service.

4) Correct data presentation in the dashboard regarding number of researches and services by lecturer.

IV. Conclusion

Based on the research, data warehouse designing can be performed using the Data Warehouse Life Cycle and through interviews to examine the needed data. The data will be subject to the ETL process for data integration to be presented into the dashboard, OLAP, and report to facilitate the executives in monitoring the performance of research and service by the lecturer in the Faculty of Engineering UNSOED.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. The research was supported and funding by the LPPM Universitas Jenderal Soedirman (Institutional Research Grant Schema with Contract Number T/984/UN23.18/PT.01.03/2022).

References

[1] Universitas Jenderal Soedirman, 2021. https://unsoed.ac.id/ (accessed Nov. 25, 2021).

[2] Tentang Fakultas – Fakultas Teknik Unsoed, 2021. http://ft.unsoed.ac.id/tentang-fakultas/ (accessed Nov. 25, 2021).

[3] Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 3 Tahun 2020 Tentang Standar Nasional Pendidikan Tinggi. 2020.

[4] Arti kata eksekutif - Kamus Besar Bahasa Indonesia (KBBI) Online. https://kbbi.web.id/eksekutif (accessed Nov. 29, 2021).

[5] W. H. Inmon, Building the Data Warehouse, 4th ed., vol. 13, no. 401. Canada: Wiley Publishing, Inc., 2005.

[6] R. Kimball and M. Ross, The Data Warehouse Toolkit, 3rd ed. Canada: John Wiley & Sons Inc., 2013.

[7] I. P. A. E. Pratama, Handbook Data Warehouse Teori dan Praktik Berbasiskan Open Source. Bandung: Informatika, 2018.

[8] T. M. Connolly and C. E. Begg, Database Systems : A Practical Approach to Design, Implementation, and Management, 4th ed. USA: Longman Inc, 2005.

[9] A. Supriyatna, "Sistem Analisis Data Mahasiswa Menggunakan Aplikasi Online Analytical Processing (OLAP) Data Warehouse," J. Pilar Nusa Mandiri, vol. 12, no. 1, pp. 62–71, Mar. 2016, Accessed: Mar. 18, 2022. [Online]. Available: http://ejournal.nusamandiri.ac.id/index.php/pilar/ article/view/260/230.

[10] M. Demarest, "Data Warehouse Prototyping: Reducing Risk, Securing Commitment and Improving Project Governance," Jan. 25, 2008. http://dssresources.com/papers/features/demarest08/demarest01252008. html (accessed Mar. 18, 2022).

[11] M. T. Rumondor and D. C. Irawati, "Designing an Online Analytical Processing (OLAP) for Project Feasibility Study in Siau Tagulandang Biaro District," Int. J. Innov. Sci. Res. Technol., vol. 4, no. 12, pp. 620–625, Dec. 2019, Accessed: Mar. 18, 2022. [Online]. Available: www.ijisrt.com620.

[12] Imelda, "BUSINESS INTELLIGENCE," Maj. Ilm. UNIKOM, vol. 11, no. 1, pp. 111–122, 2013, Accessed: Nov. 29, 2021. [Online]. Available: https://jurnal.unikom.ac.id/_s/data/jurnal/volume-11-1/09-miu-11-1-imelda.pdf/pdf/09-miu-11-1-imelda.pdf.

[13] R. Akbar, A. Soniawan, J. Adrian, R. Azim, and A. Zikri, "Implementasi Business Intelligence untuk Menganalisis Data Persalinan Anak di Klinik Ani Padang dengan Menggunakan Aplikasi Tableau Public | Zikri | Jurnal Online Informatika," J. Online Inform., vol. 2, no. 1, pp. 20–24, Jun. 2017, Accessed: Mar. 18, 2022. [Online]. Available: http://join.if.uinsgd.ac.id/index.php/join/article/view/v2i14/52.

[14] E. P. Putra and H. Juwitasary, "Evaluasi Datawareouse dan Tools untuk meningkatkan Efesiensi dan Efektifitas (Case Study PT XYZ.)."

[15] N. Leite, I. Pedrosa, and J. Bernardino, "Open Source Business Intelligence Platforms' Assessment using OSSpal Methodology," in Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE 2018), 2018, pp. 190–196, Accessed: Nov. 22, 2021. [Online]. Available: https://www.scitepress.org/Papers/ 2018/69101/69101.pdf.