

A review of Data Infrastructure for Education: A Proposal for Improved Decision Making in XYZ University

Lê Thắng Thục

School of Interdisciplinary Sciences and Arts, Vietnam National University, Hanoi, VIETNAM

e-mail : lethangthuc3321@vnu.edu.vn

Publisher's Note: JPPM stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Corresponding Author: Lê Thắng Thục

Abstract

Data is becoming a valuable resource for organisations in a variety of industries today, including education. Educational institutions are constantly gathering massive volumes of data from many sources. XYZ University (XYZU), one of the higher educational institutions, has issues with its current data infrastructure, which slows down its decision-making. The existing system relies on reporting and analytics that are derived directly from operational applications, which leads to data silos and discrepancies. To address these issues, this paper proposes a data architecture that combines data from several source applications to facilitate integrated reporting. The paper explores the background and problems in terms of data storage, management, and use of enterprise data, followed by a problem statement, a discussion of data integration, and a proposed technical architecture.

Keywords—Data Silos, Data Architecture, Integrated Reporting, Enterprise Data, Data Integration

1 Introduction

XYZ University (XYZU) is a higher education institution that was established in 1975. The university offers a wide variety of undergraduate and graduate programs across numerous disciplines. The university has a number of operational applications that manage student information, course information, faculty information and other information. The university is dedicated to fostering social responsibility, moral principles, and academic success among its faculty and student body. The university is also committed to promoting an innovative, creative, and entrepreneurial culture so that its graduates may compete on a global scale. As the university continues to grow, the need for a robust data infrastructure to support integrated reporting becomes increasingly important and necessary.

With a growing student population and increasing demand for data-driven decision-making, the university needs to improve its data management system to generate accurate and timely reports [1]. The university generates massive amounts of data from student records, assessments, and learning management systems. Nevertheless, the majority of this data is siloed across several systems and applications, making it challenging to extract useful insights. A data infrastructure that can store, manage, and integrate data from several sources and offer valuable insights for decision-making is proposed for XYZU.

2 Business Issues

XYZU is facing several business issues related to data management and use. The current system of reporting and analytics is inefficient and inconsistent, resulting in data silos and inaccuracies. This leads to difficulties in making data-driven decisions, which can impact the overall performance of the university [1]. Additionally, the lack of integration between operational applications limits the university's ability to gain insights from data and make informed decisions.

3 Background of Problem

Nowadays, educational institutions are operating in an increasingly complex and competitive environment [2]. They have access to a far larger collection of more relevant data than ever before. The challenge of storing and managing large amounts of data faced by educational institutions is compounded by the requirement to incorporate data from numerous sources in order to provide a holistic view of student progress and institutional performance. The use of data to inform decisions related to everything from resource allocation to course design is demanded. The current reporting and analytics systems in the education industry rely heavily on individual operational applications, making it difficult to analyse and interpret their data [3]. The storage, management, and use of enterprise data presents significant challenges for educational institutions. They are turning to technologies such as databases, data warehouses, data lakes or others to address their challenges.

A collective repository that captures elements of the worldly realities is established through technology like database. The Online Transaction Processing (OLTP) database was developed for current business processes, with the goal of retaining current transactions and allowing speedy processing and access to new transactions [4]. Databases that perform OLTP transactions enable the allocation and retention of a shared source of current time and exact data. However, it is difficult for educational institutions to run the analytical queries due to the joining of several tables and requires a database administrator with sufficient expertise or a developer who is familiar with the programme to submit queries that result in meaningful analysis.

A data warehouse, on the other hand, is a system of information that stores all historical data and communicative data acquired from both solitary and multiple sources [5]. Data is extracted, transformed, and loaded (ETL) into educational programmes for processing. The warehouse was developed for big historical data and enables the swift and intricate queries of the overall data generally utilised by Online Analytical Processing server (OLAP). Data warehouses employ OLAP to run a limited number of complex queries on vast aggregated past data sets. Nevertheless, it is challenging for educational institutions to make decisions and concessions on which data to keep and which data to delete [6]. A significant amount of effort and time is invested upfront determining on data storage and transformation. Less time is spent on executing data discovery, identifying patterns, or developing better hypothesis for commercial value addition.

On the contrary, data lake provides a scalable and flexible way to store raw data from a variety of sources. It is an upgraded version of the conventional data warehouse concept that underlies learning analytics applications in terms of data source type, processing type, and structure. Data that has been acquired from all of the universities is safely stored in the data lake. This includes school IT systems, learning management systems (LMS), research study findings, student surveys, and many others. The data lake enables education institutions to track how students and graduates learn throughout their lives. Similarly, the implementation of data lakes can be challenging due to data governance [7]. Because of the variety of data sources and stakeholders engaged in data management, data governance becomes more complicated. The utilisation of different data sources, inconsistent data formats, and incomplete or erroneous data can all lead to data quality challenges in a data lake setting.

In summary, education institutions face significant challenges in terms of data storage, management, and use of enterprise data. Database, data warehouse and data lake are technologies that can help education institutions to address their challenges and make more informed decisions. While these technologies offer potential benefits, there are also challenges associated with their implementation. Educational institutions must carefully evaluate their needs and capabilities before embarking on a data management and analytics strategy.

4 Problem Statement

XYZU now has disaggregated administrative information comprising of many years of spreadsheets, extracts and small databases managed by personnel to make reports. The current information systems are utilized to perform everyday transactions and store data including students' information, courses information, faculty information and others. The management relies on manual processes to extract data from different systems and transform it into a usable format, which is time-consuming and prone to errors. One of the problems XYZU is facing with its current relevant system is the lack of clean and integrated data to generate reports every semester. XYZU has many different operational systems and applications that collect data related to student information, academic programs, faculty information and others. They have data silos, where data is stored in different formats and locations and unable to interact with each other. This increases operational complexity and inefficiency since data becomes virtually imprisoned in numerous data applications. The inability to integrate data from multiple sources into a single reporting system result in a lack of data quality and consistency, which makes it difficult to access and analyze data holistically. The usage of diverse systems might make identifying connections and trends across various sections of the university difficult.

Furthermore, XYZU faces the challenge of inefficient data processing due to the lack of a centralized system capable of integrating and managing data from diverse sources. The existing systems are largely utilised for operational needs and are not intended for analytics or reporting. In some cases, data may be duplicated or have errors due to the employment of different operational applications, resulting in inaccurate insights and analysis. Staff members may have to wait for days or weeks to obtain insights because of the delayed data processing. Delays like these can result in lost opportunities or poor decision-making. Thus, an enterprise data warehouse system should be implemented for UOJ, allowing them to generate accurate reports on time every semester and identify trends, patterns and insights to support informed and effective decision making for their business.

5 Data Storage and Integration Perspective

5.1 The Need for Data Integration

The XYZU needs a data warehouse that integrates data from various sources for business analytics. This provides management with a one-way access mechanism to acquire and analyse information for decision making. Data Integration must be pursued as a strategic approach in order to assist the business with advanced analytics processes or to build multi-dimensional views of stakeholders. It provides them with real-time data in a simple format, allowing them to be proactive in exploring opportunities and identifying possible bottlenecks before they occur.

Data integration can help to address the problems faced by the university by offering a centralised platform that can aggregate data from many sources, cleanse and standardise the data, and provide a single source of truth for reporting. Once the data is integrated, it can be analysed to identify patterns, trends, and anomalies that can lead to improvements in academic programs and student outcomes. It can aid streamline data processing by consolidating data from different applications and providing a unified view of the data. This helps to reduce data processing time and improve the efficiency of data management processes. Without adequate data integration, data can be erroneous, partial, or inconsistent, resulting in ineffective decision-making.

5.2 Common Problems Faced by IT Professionals

When attempting to implement data integration, IT professionals commonly face several problems related to the quality of data [8]. Data consistency is a typical issue in which data is stored and displayed differently across numerous systems. This might lead to inconsistencies, duplication, and inaccuracies when merging the data. IT professionals may need to apply advanced data transformation techniques to make sure that data is appropriately mapped and transformed during the integration process to maintain data quality. This can be especially difficult when dealing with enormous data quantities or intricate data structures. Another problem is the lack of standardization across different data sources. Different data sources may use different formats, structures, or data definitions, making it difficult to integrate data. IT professionals may need to develop data mapping strategies to standardize the data and ensure that the data can be integrated correctly. They also need to automatically detect errors and keep an eye out for anomalies in data pipelines to prevent bigger problems from developing.

Moreover, data security is another problem faced by IT professionals. The integration process involves sharing data across multiple systems, which can increase the risk of data breaches, illegal access, and other security concerns. To safeguard data from these risks, IT professionals may need to deploy encryption, access controls, and other security measures. They may also be required to meet certain regulatory obligations concerning data protection and security. IT professionals must ensure that sensitive data is not exposed or compromised during the integration process and that data is transferred securely across different platforms.

5.3 Advantages and disadvantages of Using a Data Integration

Implementing a data integration solution can provide several advantages for XYZU, including improved decision-making and increased efficiency [9]. By combining data from different sources into a single, unified view, the administrators can access a more comprehensive and accurate understanding of the institution's operations and performance. This, in turn, can help them make better-informed decisions about resource allocation, student services, and academic programs. Moreover, data integration can automate data entry and processing tasks, reducing the time and effort required for manual data management. This increased efficiency can lead to cost savings and better use of staff time, allowing the university to focus on core educational and research missions. Overall, data integration can provide a valuable tool for the university to improve their operations and achieve their goals.

While data integration can provide significant benefits to the university, it also has several disadvantages. One of the significant disadvantages is the complexity of the integration process, which can involve combining data from multiple sources with different formats, structures, and protocols. This can require significant effort from IT

professionals and may lead to delays and errors in the integration process. Furthermore, data integration can be expensive, especially when dealing with large volumes of data or complex integration scenarios, which may require specialized tools and technologies. XYZU may need to invest in hardware, software, and staffing to support the integration process, which can add significant costs to their budgets. Therefore, XYZU needs to carefully weigh the potential benefits of data integration against the complexity and cost involved and determine if it is a viable option for their needs.

6 Technical Architecture and Conclusion

The University of XYZ should consider implementing a data warehouse to allow integrated reporting that incorporates data from many operational applications. A data warehouse is a centralised repository that contains data from different sources and is intended to aid in reporting and analytics. The technical architecture of the data warehouse consists of four main layers which includes data sources, data staging, data storage and data presentation.

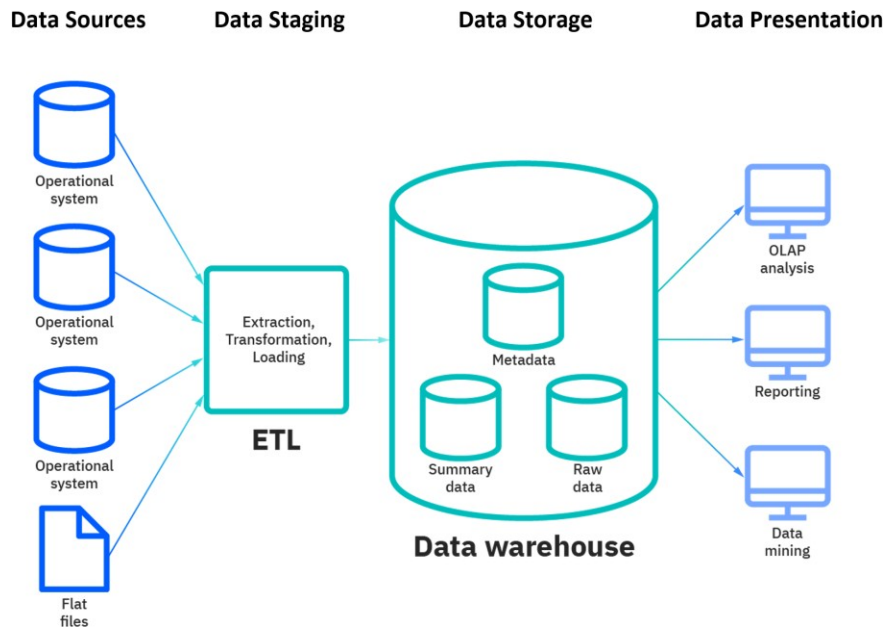


Figure 1 Data Warehouse Architecture

The proposed technical architecture can be seen in Figure 1. The data sources layer is the first layer and includes all the sources from where the data is being extracted. Data sources refer to the various systems or applications used by XYZU to collect and store data, including student information systems, financial systems, and other databases within the university that contains student data, financial data, courses data and others. This layer acts as the architecture's basis by providing the data required for analysis and reporting.

The data staging layer is the second layer which involves the ETL process that uses ETL tool such as Pentaho Data Integration (PDI) to extract data from various sources, transform and consolidate data into a consistent format from the data source layer, and load it into the data warehouse. It also provides an opportunity to validate the data, ensuring that it meets the quality standards set by the university. This involves comparing the extracted data against predefined business rules, data standards, or data profiling results. Any data that does not meet the quality standards is flagged for correction or rejection. This layer is where data is standardised, normalised, and verified before it is loaded into the data warehouse. The staging layer also provides the ability to reprocess data if necessary, ensuring that the data warehouse remains up-to-date and accurate.

The data storage layer is the heart of the data warehouse architecture, and it is where data is organized and stored in a way that makes it easy to retrieve and analyse. This layer consists of a data warehouse, which is a centralized repository that stores data from different sources in a consistent format. The data warehouse is designed to support multidimensional data analysis, allowing users to explore data across various dimensions, such as time, geography, or academic program. This is achieved through the use of a star or snowflake schema, which organizes data into fact tables and dimension tables. The fact tables contain the measures, or numerical data, while the dimension tables

LÊ THẮNG THỰC, A REVIEW OF DATA INFRASTRUCTURE FOR EDUCATION: A PROPOSAL FOR IMPROVED DECISION MAKING IN XYZ UNIVERSITY contain the attributes that provide context to the measures. The data storage layer is also responsible for managing data security, data backup, and recovery processes.

The final layer of the architecture is the data presentation layer, which is responsible for presenting data to staff in a user-friendly format to gain insights into university operations, student performance, financial performance, and more. This layer includes a variety of tools and technologies for reporting, dashboards, and analytics. The data presentation layer provides various reporting capabilities, such as tabular reports, graphs, charts, and scorecards. These reports can be accessed via web portals, mobile devices, or other interfaces. In conclusion, the proposed four-layered data warehouse architecture provides a comprehensive solution for the XUZ University data integration needs. It provides a single source of truth for the entire university. With all data stored in one location, the university can easily analyse and report on all aspects of the institution, from student enrolment to financial performance [10]. Moreover, the architecture is scalable and can handle large volumes of data, ensuring that the university can keep up with growing data demands.

Finally, the architecture is flexible and can be easily adapted to meet changing business requirements and new technologies. Overall, the proposed data warehouse architecture provides the university with a powerful tool for decision-making and strategic planning.

BIBLIOGRAPHY

- [1]. H. Ergun, G. Ghlonti, and I. Rodonaia, “Advancing University management decision - making through data warehousing and OLAP - driven decision support systems,” *International Journal of Science and Research (IJSR)*, vol. 13, no. 12, pp. 850–856, Dec. 2024. doi:10.21275/sr241209155502
- [2]. Countering violent extremism online: Understanding adversity and adaptation in an increasingly complex digital environment, 2023. doi:10.7249/rra2773-1
- [3]. “Interpret data,” *Assessing Impact: Evaluating Professional Learning*, pp. 159–173, 2018. doi:10.4135/9781506395999.n9
- [4]. J. Noverlita and H. Surbakti, “Streamlining stock price analysis: Hadoop ecosystem for Machine Learning Models and big data analytics,” *International Journal of Information Technology and Computer Science*, vol. 15, no. 5, pp. 25–34, Oct. 2023. doi:10.5815/ijites.2023.05.03
- [5]. B. wong, *Navigating the data architecture landscape: A comparative analysis of data warehouse, Data Lake, Data Lakehouse, and Data Mesh*, Oct. 2023. doi:10.20944/preprints202309.2113.v1
- [6]. O. Marzouk, J. Salminen, P. Zhang, and B. J. Jansen, “Which message? which channel? which customer? exploring response rates in multi-channel marketing using short-form advertising,” *Data and Information Management*, vol. 6, no. 1, p. 100008, Apr. 2022. doi:10.1016/j.dim.2022.100008
- [7]. R. Mahanti, “Data Governance Components and framework,” *Data Governance Success*, pp. 127–166, 2021. doi:10.1007/978-981-16-5086-4_5
- [8]. I. G. Azamovna, “Ecological edication of youth through the Edicational System,” *ACADEMICIA: An International Multidisciplinary Research Journal*, vol. 10, no. 12, pp. 1357–1360, 2020. doi:10.5958/2249-7137.2020.01845.5
- [9]. E. B. Mandinach, “Data-driven decision making in Education,” *Data-Driven Decision Making in Education*, May 2022. doi:10.4324/9781138609877-ree3-1
- [10]. A. Warokka, “Data Mining, financial performance and financial decisions on business: Evidence from Digital Marketing,” *International Journal of Psychosocial Rehabilitation*, vol. 24, no. 1, pp. 1160–1168, Jan. 2020. doi:10.37200/ijpr/v24i1/pr200217