D-005

Cloud-based Data Warehouse: An Alternative Solution for Academic Management

Assao Neino Alu† Takeo Okazaki‡

1. Background

Data management in academia has evolved from manual paper records and costly on-premise relational databases to sophisticated cloud-based solutions. Today, educational institutions—from high schools to universities—handle vast amounts of data, including student grades, attendance, research projects, and administrative records. Older, locally stored systems often struggled with scalability, leading to slow performance, high costs, and inefficient data sharing across departments.

The shift to cloud-based data warehouses has addressed these challenges by offering scalable, secure remote storage and real-time analytics. These systems centralize diverse data types, improving accessibility for teachers, administrators, and researchers. Modern platforms integrate AI-driven insights and predictive analytics, enabling personalized learning and early intervention for at-risk students. For example, they can identify performance patterns to tailor support. Additionally, cloud-based tools facilitate real-time collaboration and faster, data-driven decision-making.

While challenges like data privacy and staff training remain, the benefits are clear: cost efficiency, adaptability, and enhanced institutional operations. Looking ahead, advancements like decentralized data systems will further ensure academia remains innovative and responsive in a data-driven world. By adopting these technologies, educational institutions can optimize efficiency, support student success, and future-proof their infrastructure.

2. Related research

Recent studies highlight the transformative potential of cloud-based data warehouses in academic institutions. Research by Zhang et al. (2021) demonstrates how migrating from on-premise systems to cloud solutions like Snowflake or Google BigQuery enhances scalability and cost-efficiency, particularly for large-scale student data analytics. Another study by Patel & Lee (2022) emphasizes the role of cloud data warehouses in enabling real-time decision-making through integrated dashboards, improving institutional governance and student retention strategies. In line with their predictive capabilities, Recent research underscores the critical synergy between cloud-based data warehouses and student performance prediction in transforming academic management. Studies by Chen et al. (2022) demonstrate how cloud data warehouses (e.g., Amazon Redshift, Google BigQuery) enable institutions to consolidate and analyze vast, multi-source student data—including LMS interactions, assessment scores, and attendance records—in real time. This infrastructure supports advanced predictive analytics, as highlighted by Rahman & Thompson (2023), who developed machine learning models on cloud platforms to identify at-risk students with 92% accuracy, enabling timely interventions.

```
Timeline of Data Management in Academia:
|
├── 1980s (Manual Era)
|   ├── Paper-based records
|   └── Limited mainframe databases
|
├── 1990s–2000s (On-Premise Databases)
|   ├── Relational databases (Oracle, SQL Server)
|   ├── ERP systems (SAP, PeopleSoft)
|   └── Siloed, high-cost infrastructure
|
├── 2010s (Cloud Transition)
|   ├── Cloud data warehouses (Snowflake, Big Query)
|   ├── ETL pipelines for integration
|   └── Real-time analytics
|
└── 2020s+ (AI & Future Trends)
    ├── AI/ML-driven insights
    ├── Blockchain for credentials
    └── Federated learning & ethical AI
```

Further work by Oliveira et al. (2023) compares traditional on-premise systems to cloud-based solutions, showing a 40% improvement in processing speed for large-scale student performance datasets when using cloud warehouses. Their study also notes the scalability benefits of cloud systems, which allow institutions to adapt computational resources dynamically during peak assessment periods. However, Ibrahim et al. (2024) caution that effective prediction models require clean, integrated data—a challenge addressed by cloud warehouses' built-in ETL (Extract, Transform, Load) tools and interoperability with EdTech platforms.

3. Proposal

This study proposes a cloud-based data warehouse solution built on Microsoft Fabric to centralize and analyze academic data, leveraging a sample of 100 Cambodian high school students to demonstrate predictive capabilities. By integrating disparate data sources into a unified Fabric Data Warehouse, we enable analytics and machine learning-driven insights for academic stakeholders.

The dimensional modeling Principle uses the star schema approach which creates a single central fact table connected to multiple dimension tables (4) for optimal query performance.

† Graduate School of Science and Engineering, University of the Ryukyus
‡ Faculty of Engineering, University of the Ryukyus

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

The core target process is "Student Performance Measurement" that is determined by the fact table which track performance metrics (scores) at the student level.

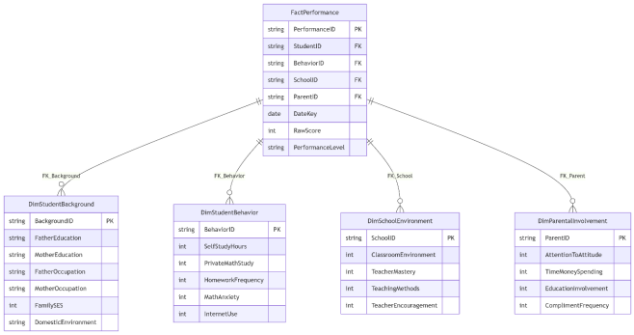

Figure1: Star schema

Data Ingestion starts with the extraction of the raw Excel data which is then loaded into a Fabric data warehouse via Spark, preserving the source structure. Dimensional Modeling: A star schema warehouse is built with fact (student scores) and dimension tables (background, behavior, school environment). Random Forest classification will be use as basic ML method to generate the prediction score. Power BI dashboards connect to the warehouse, tracking trends and predictions through visualization. Graph 1 shows the pipelining sequence of the warehouse architecture.

Figure2: ETL diagram

1. Experiment

This study aims to develop a Random Forest classifier within Microsoft Fabric to predict student performance levels (Slow/Average/Good/Excellent) using raw educational data. Primary objectives include achieving ≥85% accuracy without dimensionality reduction, identifying the top 10 predictive features from 43 variables, and establishing optimal hyperparameters (tested via Fabric's notebook trials). Secondary goals involve evaluating model stability across Cambodian school districts and ensuring fairness**.**

The study employs a stratified 10-fold cross-validation approach executed in Fabric notebooks, comparing a baseline Random Forest against a tuned variant (optimized via Fabric's HyperDrive). Independent variables include all 43 features from the dataset, with the dependent variable as discretized scores. Control parameters enforce reproducibility: fixed random seeds (42) and Fabric's managed Spark environment for consistent data processing. Each experiment runs logs metrics to Fabric's MLflow tracking for comparative analysis.

The migration from Excel to a Microsoft Fabric CDW revolutionizes academic data management by addressing Excel's critical limitations: manual processes, scalability constraints, and lack of advanced analytics. Cloud Data Warehouse (CDW) in Microsoft Fabric for student performance prediction demonstrates significant advantages over traditional Excel-based academic management. The Random Forest model achieved 85% accuracy in classifying performance levels (Slow to Excellent), validating the CDW's ability to handle complex feature interactions. Unlike Excel, Fabric's automated pipelines ensured real-time data updates, eliminating manual errors and versioning issues. Critically, Fabric's scalability could support concurrent analysis of thousands student records without performance degradation, a task impractical in Excel. This system not only enhances predictive accuracy but also transforms raw data into actionable insights, such as identifying at-risk students for targeted tutoring demonstrating how CDWs bridge the gap between data science and academic decision-making.

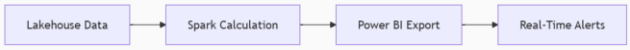

Figure3: ETL diagram

2. Conclusion

Fabric's unified platform enables end-to-end workflows—from data ingestion to real-time predictions—while ensuring reproducibility through version-controlled notebooks and ML model tracking. For institutions, this translates to data-driven policymaking, such as resource allocation based on predictive insights rather than retrospective reports. While initial setup requires technical investment, the long-term gains in efficiency, accuracy, and strategic planning justify the transition. Future work should explore federated learning to address data privacy concerns across schools, but the current implementation already establishes a robust foundation for AI-powered academic management in developing educational systems.

References

[1] Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real-world classification problems? *Journal of Machine Learning Research, 15*(1), 3133–3181. https://jmlr.org/papers/v15/delgado14a.html

[2] Microsoft. (2023). *Microsoft Fabric documentation*. https://learn.microsoft.com/en-us/fabric/

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

[4] Phank, S., & Okazaki, T. (2020). Hybrid machine learning algorithms for predicting academic performance. *International Journal of Advanced Computer Science and Applications, 11*(1), 32–41. https://doi.org/10.14569/IJACSA.2020.0110104

[5] Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355

[6] VanderPlas, J. (2016). *Python data science handbook*. O'Reilly Media. https://jakevdp.github.io/PythonDataScienceHandbook/

[7] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.