NextGen Data Architects

A Comprehensive Data Engineering and Analytics Platform for Educational Institutions

Authors

[Your Name]
Co-Authors: [Names]
Supervisor: [Name]

Affiliations

Uganda Christian University
Faculty of Engineering, Design & Technology
Department of Computing

Introduction

Educational institutions generate vast amounts of data from academic, administrative, and financial operations. This research presents **NextGen Data Architects**, a unified data platform that integrates fragmented data sources, provides predictive analytics, and enables role-based data access for informed decision-making.

Objective

Design and implement a scalable data warehouse with ETL pipeline, role-based access control system, machine learning models for student performance prediction, and automated payment tracking with deadline compliance.

Methodology

Data Sources: Academic (DB1) and Administrative (DB2) databases

ETL Pipeline: Medallion Architecture (Bronze → Silver → Gold)

Data Warehouse: Star Schema with fact and dimension tables

ML Models:

- Random Forest (100 estimators)
- Gradient Boosting (100 estimators)
- Neural Network (MLPRegressor)
- Ensemble Model (average of all three)

Implementation: Python/Flask backend, React.js frontend, MySQL database

Related Literature

Star Schema design (Kimball & Ross, 2013), Medallion Architecture (Databricks, 2023), Educational Data Mining (Romero & Ventura, 2013), RBAC models (Sandhu et al., 1996), Ensemble methods (Breiman, 2001; Friedman, 2001).

Analysis

Data Integration: Successfully unified academic, financial, and administrative data sources into a single data warehouse.

Student Distribution by Department
Bar chart showing enrollment across departments

Caption: Distribution of students across all departments and programs

Payment Status Distribution
Pie chart: Paid, Pending, Overdue percentages

Caption: Payment compliance tracking with deadline analysis

Average Grades Over Time
Line chart showing performance trends by semester

Caption: Academic performance trends across semesters

Model Performance: Ensemble model achieved $R^2 > 0.85$, outperforming individual models with reduced variance.

Results/Findings

- 100% data integration success across all sources
- ML model accuracy: $R^2 > 0.85$ for performance prediction
- 9 user roles with effective RBAC implementation
- 95%+ accuracy in payment deadline compliance tracking
- Real-time dashboard updates with filter synchronization
- Scalable architecture supporting thousands of students

IMPORTANT!

The system enables early identification of at-risk students and automated financial compliance tracking, improving institutional decision-making capabilities.

Conclusion

Key Findings:

- Medallion Architecture ensures data quality and traceability
- Ensemble ML models achieve superior prediction accuracy
- RBAC system effectively enforces data access policies

Implications: Improved data-driven decision-making, early intervention for at-risk students, efficient financial management.

Future Work: Real-time streaming, deep learning models, mobile applications, multi-institution support.

Contact Information

Email: [your.email@ucu.ac.ug] | **Institution:** Uganda Christian University | **Department:** Computing