RWANDA POLYTECHNIC

HUYE COLLEGE

MODULE NAME: DATA MINING AND WAREHOUSING.

DEPARTMENT : ICT

OPTION : IT

LEVEL : 8 YEAR 4

CLASS : IT Btech

LIBRARY USAGE DATA WAREHOUSE SYSTEM

A Comprehensive Data Integration Solution for Modern Library Management

GROUP 1: MEMBERS

NAMES	REG NO	MARKS
HIMBAZWA Rebbeca	25RP21086	
NISINGIZWE Jean Pierre	25RP20888	
UMUGWANEZA Aimee Providence	25RP18890	

Date : 30th Jan 2026

1. Executive Summary

The Library Usage Data Warehouse System is a centralized analytical platform designed to integrate, clean, and transform library operational data into meaningful insights for decision-making. Modern libraries generate data from multiple sources such as book borrowing, digital downloads, and room bookings. Without integration, this data remains fragmented and underutilized.

This project implements a MySQL-based data warehouse using a star schema and a Python-driven ETL pipeline to consolidate and standardize data from CSV and Excel sources. The result is a high-quality, query-optimized warehouse that supports executive, departmental, and operational reporting.

Key Achievements:

- Designed and implemented a star schema data warehouse
- Built automated ETL processes using Python and Pandas
- Applied NULL handling and data quality validation
- Implemented role-based access and data masking/encryption
- Created SQL views for dashboard visualization

2. Introduction

This report documents the design and implementation of a Library Usage Data Warehouse System developed as part of the Data Mining and Warehousing module. The system supports analytical reporting and strategic planning for library operations.

2.1 Background

Libraries manage diverse services including physical lending, digital resource usage, and room reservations. Each service generates valuable data, but when systems operate independently, holistic analysis becomes difficult.

2.2 Problem Statement

The major challenges identified include:

- Disparate and inconsistent data sources
- Manual and slow report generation
- Limited historical and cross-functional analysis
- Poor data quality due to missing and inconsistent values

2.3 Solution Overview

The project proposes a centralized data warehouse with a star schema architecture to integrate and transform all library usage data into a unified analytical repository.

3. Project Overview

The system integrates data from:

Source System	Format	Description
Book Usage Logs	CSV	Downloads and reading activity
Student Records	CSV / Excel	Student IDs and departments
Room Bookings	Excel	Study room usage and time slots

Objectives:

- Create a single source of truth
- Enable fast analytical queries
- Improve data quality and consistency
- Automate ETL processes
- Support dashboards and BI tools

4. System Architecture

The system follows a three-tier architecture:

1. Source Layer – CSV and Excel operational data
2. ETL Layer – Python + Pandas for transformation and validation
3. Warehouse Layer – MySQL snowflake schema for analytics

Data Flow:

Sources → Staging → Cleaning → Transformation → Dimensions → Fact → Reporting

5. Data Warehouse Schema Design

A snowflake schema is used with:

5.1 Fact Table

fact_library_usage

Contains metrics such as:

- downloads
- duration_minutes
- booking_hours

With foreign keys to:

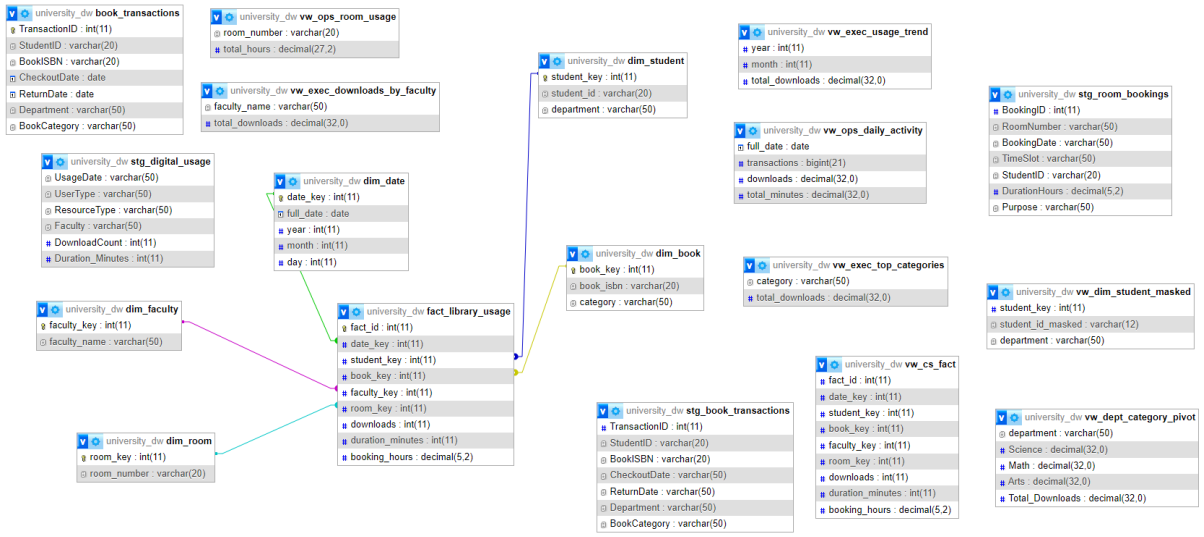
- date_key
- student_key
- book_key
- faculty_key
- room_key

5.2 Dimension Tables

- dim_date
- dim_student
- dim_book
- dim_faculty
- dim_room

6. Entity Relationship Diagram (ERD)

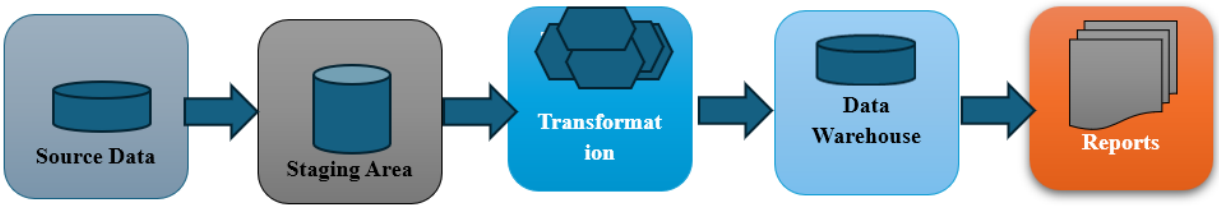
The ERD places fact_library_usage at the center with many-to-one relationships to each dimension, ensuring referential integrity and efficient querying.



7. ETL Process Flow

1. Extract data from CSV/Excel
2. Validate formats and types
3. Handle NULL values
4. Clean and standardize text fields
5. Load dimension tables
6. Generate surrogate keys
7. Load the fact table

ETL Process Flow Diagram



8. NULL Handling Strategy

- Preserve meaningful NULLs
- Replace mandatory NULLs with defaults
- Log unexpected NULLs for review

9. Data Quality Framework

Checks include:

- Completeness
- Consistency
- Accuracy
- Uniqueness
- Referential integrity

10. Dimensional Model Justification

The star schema was chosen because it:

- Improves query performance
- Is simple and business-friendly
- Supports OLAP operations
- Scales well for future expansion

11. Technology Stack

Component	Technology
Database	MySQL 8.0
ETL	Python 3 + Pandas
BI	Power BI / Excel
Docs	Word / PDF

12. Implementation Details

- MySQL used for schema creation and indexing
- Python scripts handle ETL
- SQL views created for dashboards
- User roles defined for admin, HOD, and analysts

13. Performance Considerations

- Indexed surrogate keys
- Optimized joins
- Pre-aggregated views

14. Security and Compliance

- Role-based access control
- Encrypted student_id
- Masked views for non-admin users

15. Testing and Validation

- Unit testing for ETL
- Referential integrity checks
- Query performance testing

16. Dashboards and Reporting

Three dashboards were designed:

- Executive Dashboard
- Department Dashboard
- Operational Dashboard

Each uses MySQL views as data sources in Power BI / Excel.

17. Future Enhancements

- Predictive analytics
- Machine Learning Models
- Real-time dashboards
- Mobile BI access

18. Conclusion

This project successfully demonstrates the design and implementation of a modern data warehouse for library analytics. By integrating multiple data sources into a star schema and automating ETL processes, the system provides reliable, scalable, and secure access to insights that support data-driven decision-making