RWANDA POLYTECHNIC

HUYE COLLEGE

MODULE NAME: DATA MINING AND WAREHOUSING.

DEPARTMENT: ICT

OPTION: IT

LEVEL: 8 YEAR 4

CLASS: IT Btech

Date: 06th Feb. 2026

LIBRARY USAGE DATA WAREHOUSE SYSTEM

GROUP 8: MEMBERS

NAMES	REG NO	MARKS
HIMBAZWA Rebecca	25RP21086	
NISINGIZWE Jean Pierre	25RP20888	
UMUGWANEZA Aimée Providence	25RP18890	

Contents

LIST OF TABLES iii
LIST OF FIGURES iv
1. EXECUTIVE SUMMARY 1
2. INTRODUCTION 1
2.1 BACKGROUND 1
2.2 Problem Statement 2
2.3 Solution Overview 2
3. PROJECT OVERVIEW 2
3.1 Data Sources 2
3.2 Project Objectives 3
4. SYSTEM ARCHITECTURE 3
4.1 Architecture Layers 3
4.2 Data Flow 4
5. DATA WAREHOUSE SCHEMA DESIGN 4
5.1 Dimension Tables 4
6. ENTITY RELATIONSHIP DIAGRAM 5
7. ETL PROCESS FLOW 5
7.1 ETL Process Flow Diagram 6
8. NULL HANDLING STRATEGY 6
9. DATA QUALITY FRAMEWORK 6
10. DIMENSIONAL MODEL JUSTIFICATION 7
11. ANALYTICS RESULTS AND BUSINESS INSIGHTS 8
PART 1: OLAP Operations 8
PART 2: Power BI Dashboards 10
Executive Dashboard 10
Department Dashboard 11
Operation Dashboard 12
12. TECHNOLOGY USED 12
13. IMPLEMENTATION DETAILS 13
14. SECURITY AND COMPLIANCE 13
15. TESTING AND VALIDATION 13
16. FUTURE ENHANCEMENTS 14
17. CONCLUSION 14

LIST OF TABLES

Table 1: Data Sources.....3
Table 2: Technology used.....12

LIST OF FIGURES

Figure 1: Architecture Layers....4
Figure 2: Entity Relationship Diagram....5
Figure 3:ETL Flow process....6
Figure 4:Star schema....7
Figure 5: Drill-down result....8
Figure 6: Roll-up result....9
Figure 7: Slicing result....9
Figure 8: Dicing result....10
Figure 9: Executive Dashboard....10
Figure 10: Department Dashboard....11
Figure 11:Operation Dashboard....12
Figure 12: Users....13
Figure 13: Student Masking....13

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

Libraries generate large volumes of data from multiple systems including book Usage Logs, student records, and room bookings. This project integrates these sources into a single data warehouse for analytical processing and visualization.

3.1 Data Sources

The data warehouse integrates information from three main library systems to provide a unified view of library usage. The first source is the Book Usage Logs (MySQL database), which records book checkouts, returns, student departments, and book categories. The second source is the Student records (Excel/CSV files), which tracks e-book and journal downloads, user types, resource types, and reading durations. The third source is the Study Room Booking System (CSV exports), which contains data on room reservations, booking dates, time slots, student identifiers, and booking purposes. These heterogeneous data sources were cleaned, standardized, and integrated through ETL processes before being loaded into the data warehouse for analysis and better visualization standardization.

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

Table 1: Data Sources

3.2 Project Objectives

The objectives of this project are to:

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

4. SYSTEM ARCHITECTURE

The system follows a **three-tier architecture** to ensure modularity, maintainability, and scalability.

4.1 Architecture Layers

- **Data Source Layer:** CSV and Excel from operational data
- **ETL Layer:** Python and Pandas for extraction, transformation, and validation
- **Data Warehouse Layer:** MySQL-based star schema optimized for analytics

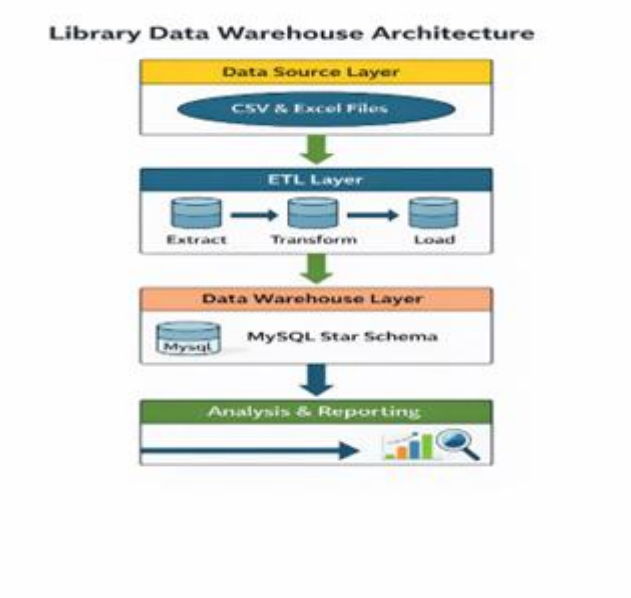


Figure 1: Architecture Layers

4.2 Data Flow

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

5. DATA WAREHOUSE SCHEMA DESIGN

A **star schema** design is used, consisting of one central fact table connected to multiple dimension tables.

Fact Table fact_library_usage Contains metrics such as: downloads, duration_minutes, booking_hours with foreign keys: date_key, student_key, book_key, faculty_key, and room_key.

5.1 Dimension Tables

- ✓ dim_date
- ✓ dim_student
- ✓ dim_faculty
- ✓ dim_room

6. ENTITY RELATIONSHIP DIAGRAM

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

.

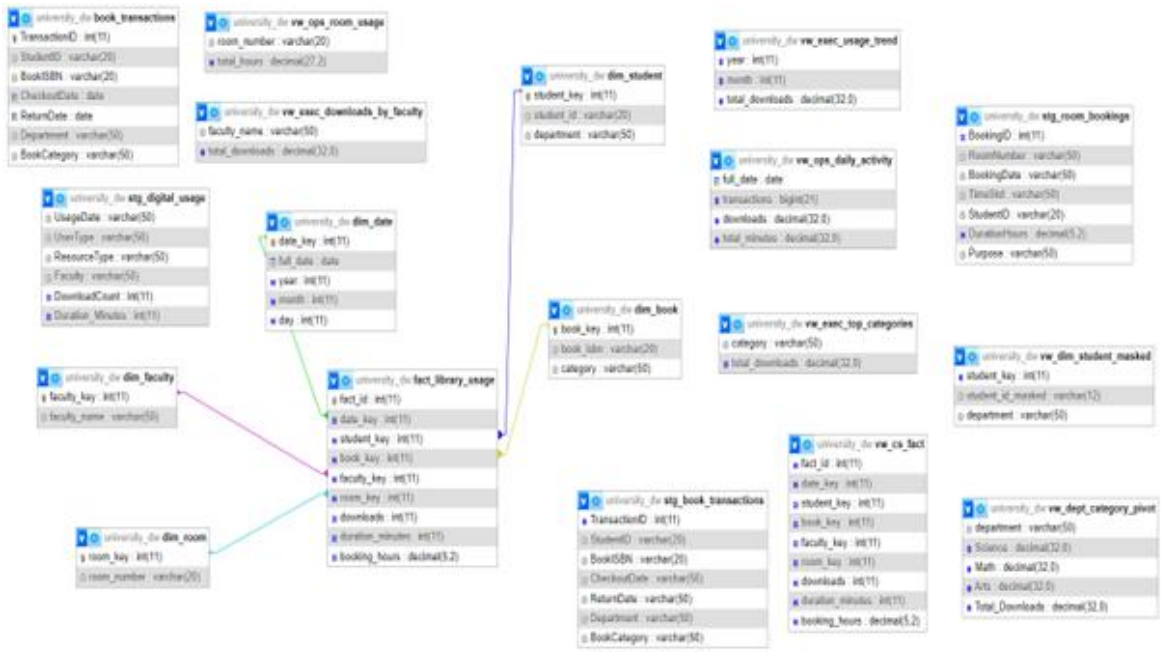


Figure 2: Entity Relationship Diagram

7. ETL PROCESS FLOW

- Extract data from CSV/Excel
- Validate formats and types
- Handle NULL values
- Clean and standardize text fields
- Load dimension tables
- Generate surrogate keys
- Load the fact table

7.1 ETL Process Flow Diagram

Figure 3:ETL Flow process

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:

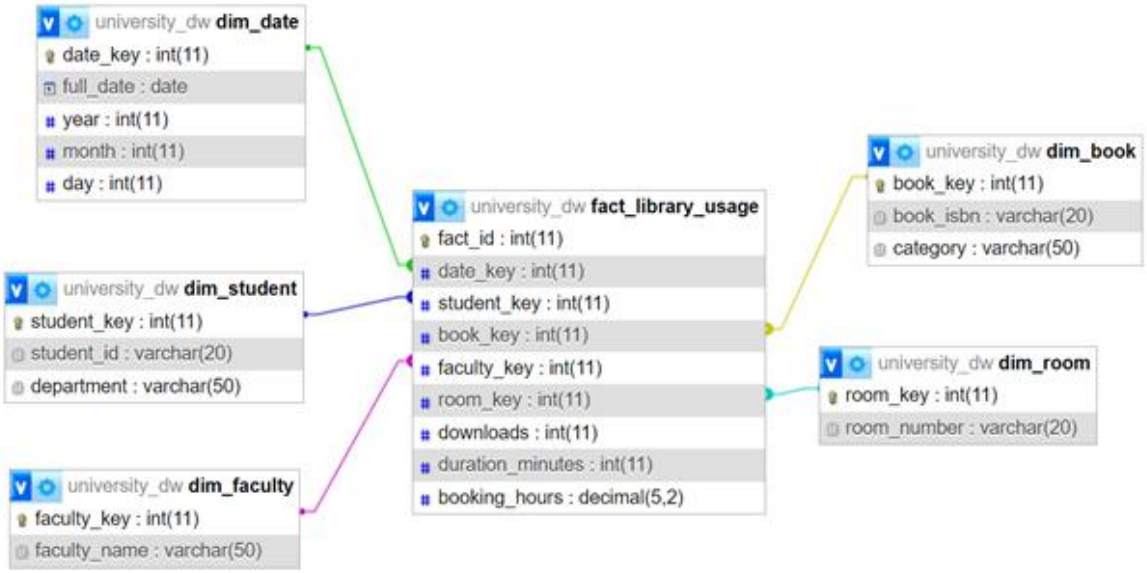


Figure 4:Star schema

The star schema was chosen because it:

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

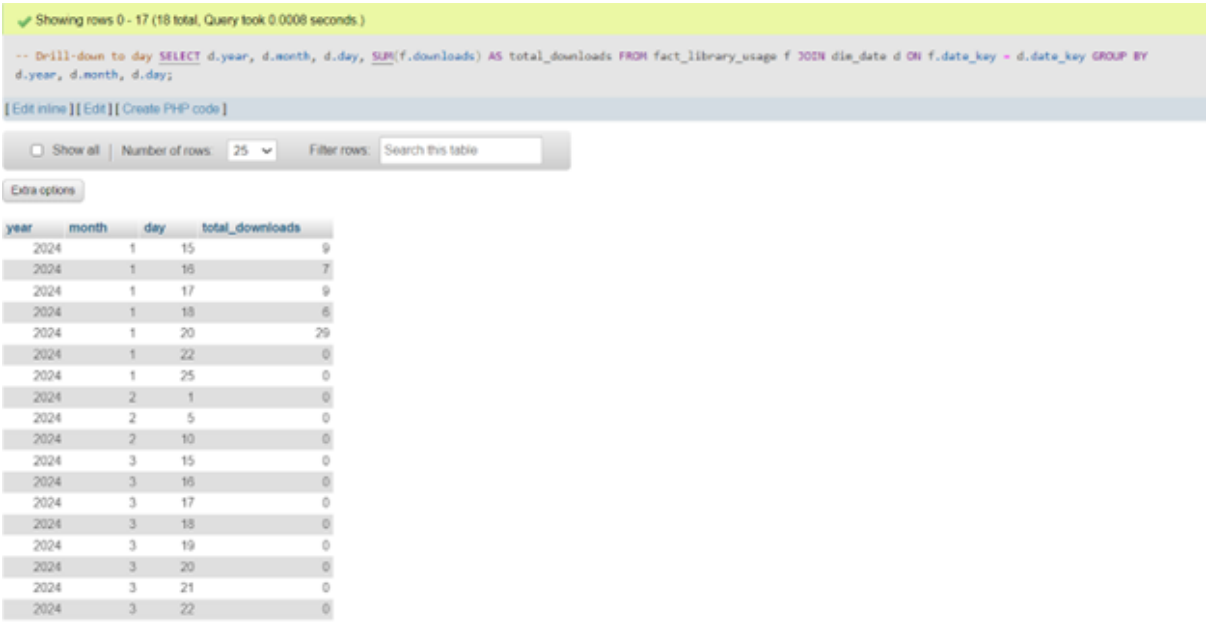
11. ANALYTICS RESULTS AND BUSINESS INSIGHTS

PART 1: OLAP Operations

The implemented warehouse enables standard OLAP operations for multidimensional analysis:

a) Drill-down (Year → Month → Day)

This figure shows the results of the drill down operation of total downloads starting from year, month and Day.



Showing rows 0 - 17 (18 total, Query took 0.0008 seconds.)

```
-- Drill-down to day SELECT d.year, d.month, d.day, SUM(f.downloads) AS total_downloads FROM fact_library_usage f JOIN dim_date d ON f.date_key = d.date_key GROUP BY
d.year, d.month, d.day;
```

[Edit inline] [Edit] [Create PHP code]

Show all | Number of rows: 25 | Filter rows: Search this table

Extra options

year	month	day	total_downloads
2024	1	15	9
2024	1	16	7
2024	1	17	9
2024	1	18	6
2024	1	20	29
2024	1	22	0
2024	1	25	0
2024	2	1	0
2024	2	5	0
2024	2	10	0
2024	3	15	0
2024	3	16	0
2024	3	17	0
2024	3	18	0
2024	3	19	0
2024	3	20	0
2024	3	21	0
2024	3	22	0

Figure 5: Drill-down result

These figures below represent a roll-up operation aggregating daily room bookings into monthly summaries.

Query selecting day, month, year and sum of booking duration hours from fact_library usage table joining the dimension date table.

b) Roll-up (Day → Month → Year)

Results showing the selected day, month, year and total booking hours.

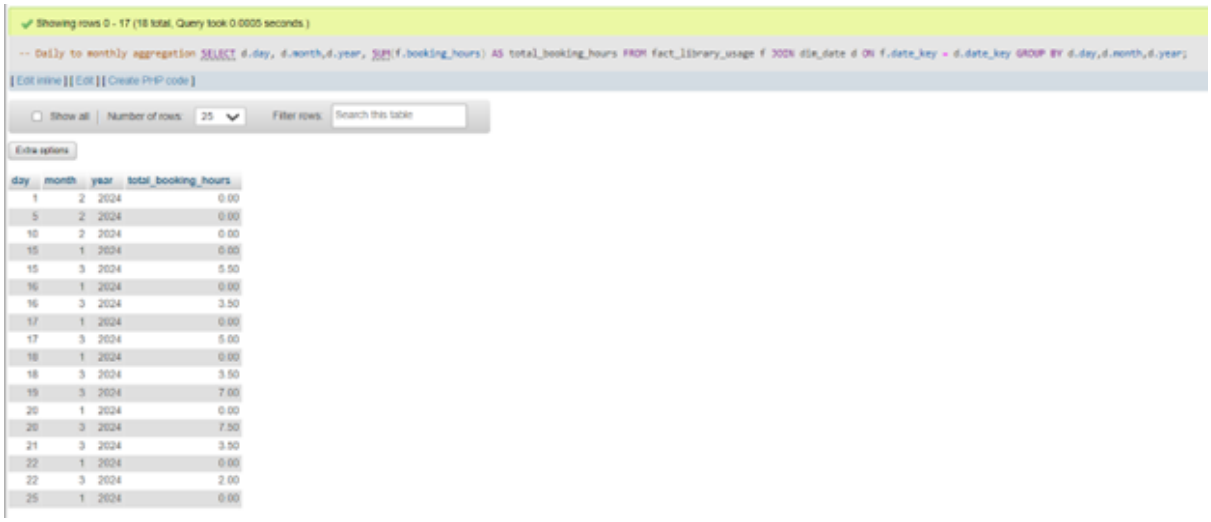


Figure 6: Roll-up result

c) Slicing (subset on one dimension)

Results displaying the total downloads and total booking hours for 2024.

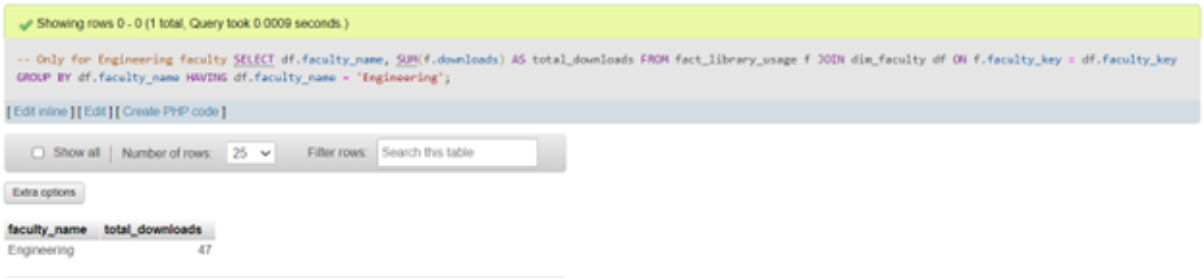


Figure 7: Slicing result

d) Dicing (subset on multiple dimensions)

Lastly, we have dicing operation showing the digital downloads of books in every department.

Result showing digital downloads in each department.

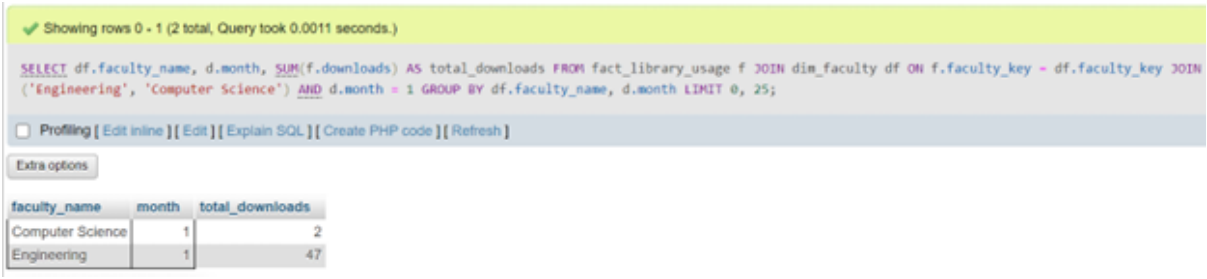


Figure 8: Dicing result

PART 2: Power BI Dashboards

Interactive dashboards were developed using Power BI to support different user roles. We have three dashboards which includes: Executive dashboard, Department dashboard and Operation dashboard.

Executive Dashboard

This dashboard provides high-level KPIs such as total download trends and booking duration hours, which is assigned specifically to the Library Director.

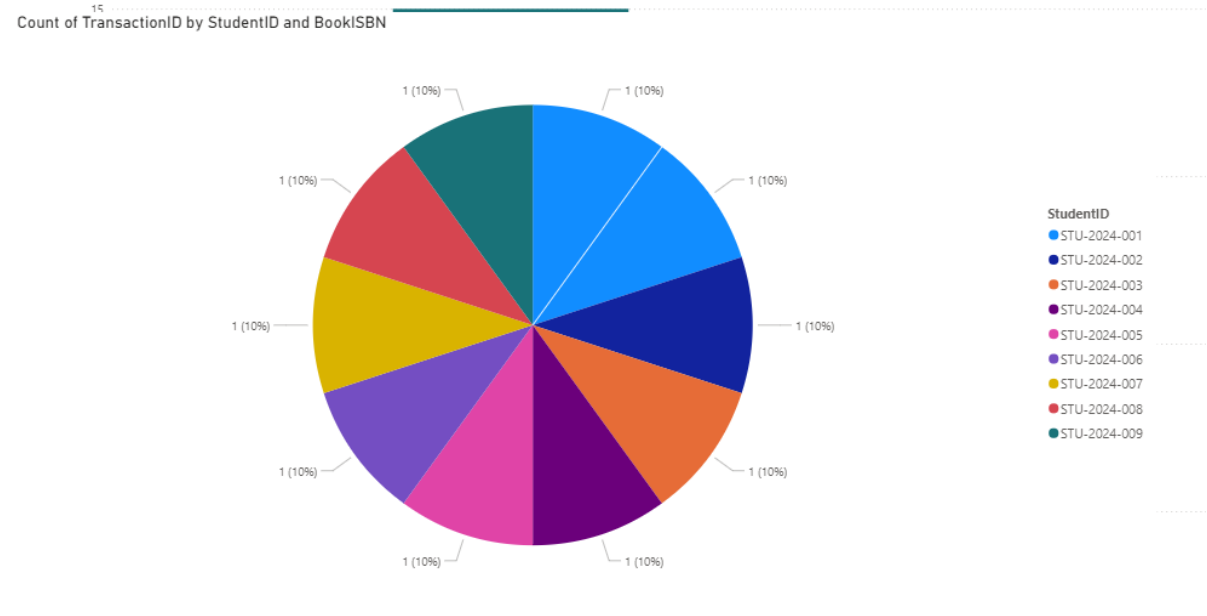


Figure 9: Executive Dashboard

Department Dashboard

The Department Dashboard enables department-level analysis of library usage. This dashboard is assigned to the Head of Department.

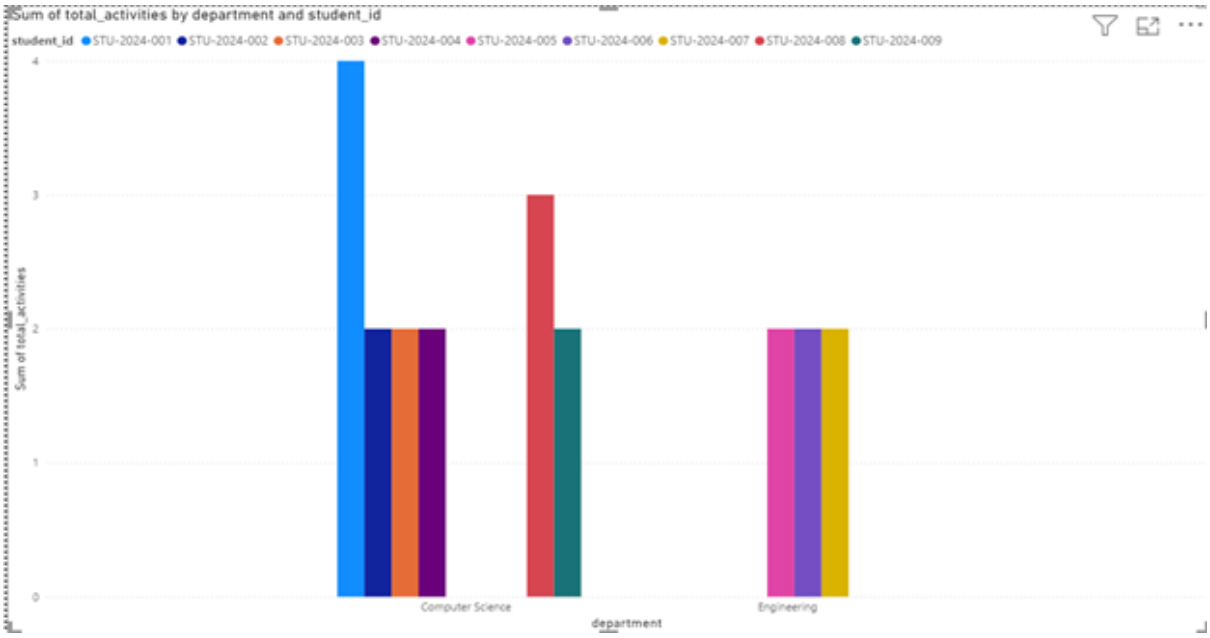


Figure 10: Department Dashboard

Operation Dashboard

The Operational Dashboard supports daily monitoring of room bookings and digital activity. This dashboard is assigned to the Analyst with read-only access.

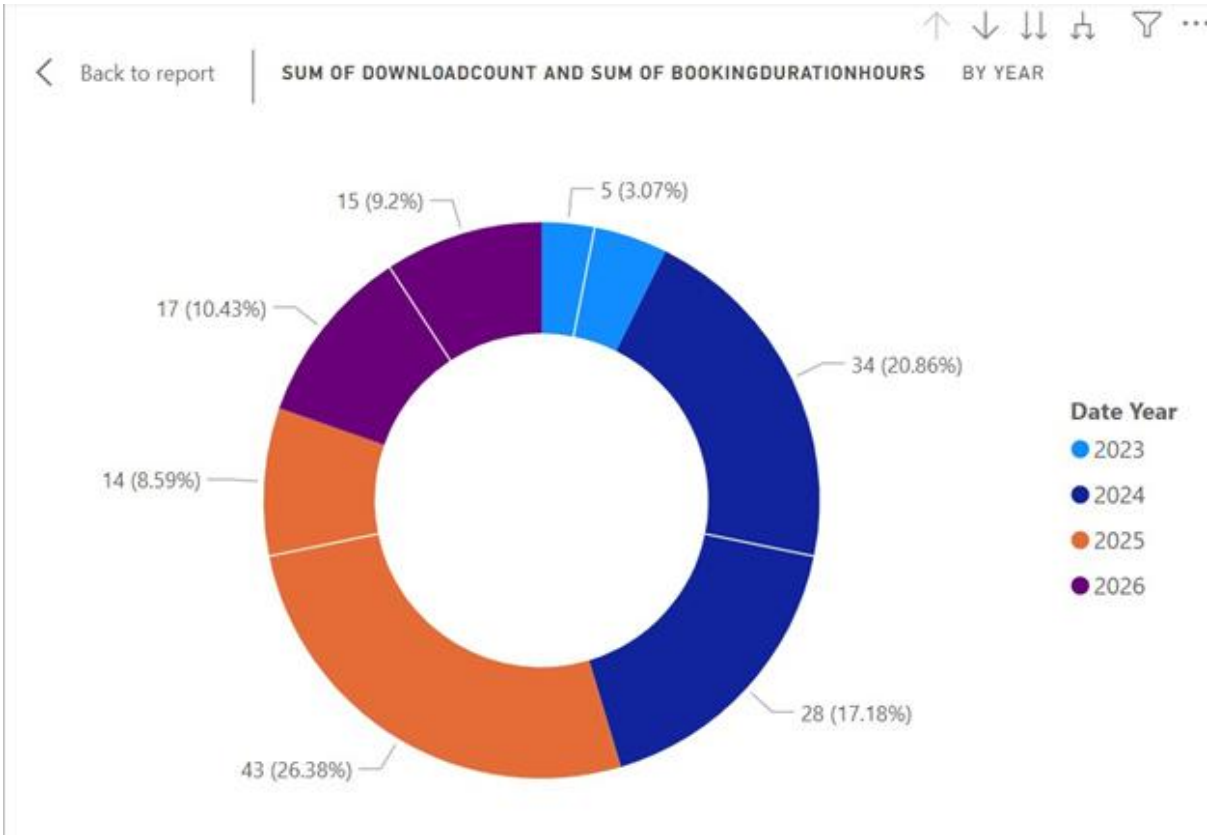


Figure 11:Operation Dashboard

12. TECHNOLOGY USED

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

Table 2: Technology used

13. IMPLEMENTATION DETAILS

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

User	Host
HOD	%
analyst1	%
dw_admin	%

Figure 12: Users

14. SECURITY AND COMPLIANCE

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

	student_key	student_id_masked	department
Edit Copy Delete	1	STU-XXXX-XXX	Computer Science
Edit Copy Delete	2	STU-XXXX-XXX	Computer Science
Edit Copy Delete	3	STU-XXXX-XXX	Computer Science
Edit Copy Delete	4	STU-XXXX-XXX	Computer Science
Edit Copy Delete	5	STU-XXXX-XXX	Engineering
Edit Copy Delete	6	STU-XXXX-XXX	Engineering
Edit Copy Delete	7	STU-XXXX-XXX	Engineering
Edit Copy Delete	8	STU-XXXX-XXX	Computer Science
Edit Copy Delete	9	STU-XXXX-XXX	Computer Science

Figure 13: Student Masking

15. TESTING AND VALIDATION

Testing includes unit, integration, performance, and user acceptance testing. Data accuracy and ETL reliability are validated through reconciliation and benchmarking.

16. FUTURE ENHANCEMENTS

- Predictive analytics: Forecast resource demand and usage patterns using historical data
- Machine Learning Models: Personalized book recommendations and automated categorization
- Real-time dashboards: Live data updates for immediate operational insights
- Mobile BI access: Mobile apps for on-the-go access to reports and analytics

17. 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.