LUMEN-RA

Digital Allyship Analytics Warehouse

Data Science & Engineering Documentation

Team 39 · Women Techsters Fellowship 2025 · SDG 5 — Gender Equality

Field	Details
Project Name	Digital Allyship Analytics Warehouse
Team Name	Lumen-Ra — Team 39
SDG Alignment	SDG 5 — Gender Equality
Programme	Women Techsters Fellowship 2025
Documentation Author	Data Science & Engineering Team
Date	February 2026
Repository	github.com/WANGAR1/Lumen-Ra-Analytics-Dashboard

1. Project Overview

Lumen-Ra is a Digital Allyship Toolkit designed to equip men aged 16–44 with practical, trauma-informed skills to support women experiencing gender-based violence or health-related distress. The platform is built around structured learning modules, scenario-based simulations, and progress tracking.

The Data Science and Engineering team was responsible for designing and building the data infrastructure that supports analytical reporting and impact measurement for the platform. This document covers the full scope of work completed during the capstone project.

2. Problem Statement

Despite high awareness of gender-based violence among men aged 18–40, many lack the practical skills and structured guidance required to respond appropriately when women disclose experiences of violence. The platform addresses this knowledge–action gap.

From a data perspective, the challenge was:
- The platform was still under development during the capstone timeline
- No real user data was available to analyze
- The data science team needed to demonstrate analytical capability before the platform launched
- A full data warehouse architecture needed to be designed and implemented

3. Data Architecture

3.1 Medallion Architecture

The data warehouse follows the Medallion Architecture — an industry-standard approach that organizes data into three progressive layers, each adding more structure and business value than the last.

Layer	Name	Purpose
Bronze	Raw Data	Stores data exactly as received — CSV files loaded into SQL Server with no transformation
Silver	Cleaned Data	Removes duplicates, standardizes formats, validates score ranges, handles missing values
Gold	Analytics Layer	Business-ready star schema optimized for reporting — fact and dimension tables powering dashboards

3.2 Database Schema

The Bronze layer schema was designed using dbdiagram.io and contains five core tables that model the LMS platform data. All tables follow proper relational database design with primary keys and foreign key relationships.

- lms_users — 50 registered users (39 learners, 11 admins excluded from analysis)
- lms_courses — 5 learning courses across GBV awareness, communication, referral and recovery
- lms_modules — 15 modules across all courses, each with content_type (video or text)
- lms_lesson_activity — 150 activity events tracking user progress per module
- lms_quiz_attempts — 120 quiz attempts measuring knowledge acquisition

Key relationships: lms_courses → lms_modules (via course_id), lms_users → lms_lesson_activity (via user_id), lms_users → lms_quiz_attempts (via user_id), lms_modules → lms_lesson_activity (via module_id), lms_modules → lms_quiz_attempts (via module_id).

4. Synthetic Data Generation

4.1 Approach & Rationale

Because the Lumen-Ra platform was still in active development during the capstone timeline, no real user data was available. The team used Mockaroo — a professional synthetic data generation platform — to create realistic dummy datasets that mirror what real platform data would look like.

This approach allowed the data science team to build and validate the full analytics pipeline before the platform launched, ensuring that when real data becomes available it can be loaded directly into the existing warehouse with no structural changes required.

4.2 Data Generation Process

Tables were generated in a specific dependency order to ensure referential integrity across all foreign key relationships:

- Step 1 — lms_users generated first (50 rows) as the root anchor table
- Step 2 — lms_courses generated (5 rows) as the content reference
- Step 3 — lms_modules generated (15 rows) using course_id values from Step 2
- Step 4 — lms_lesson_activity generated (150 rows) using user_id and module_id from Steps 1 and 3
- Step 5 — lms_quiz_attempts generated (120 rows) using same ID references

User IDs from each parent table were copied into Mockaroo's Custom List field type for child tables, ensuring every foreign key value exists in the referenced parent table. ID prefixes (ACT, QZ) were applied post-generation using Excel formula: =`"ACT"&TEXT(A2,"000")`.

4.3 Admin Exclusion

Upon analysis, 11 of the 50 users were identified as admin accounts (role = 'admin'). All analytics and dashboards were produced using learner-only data (39 learners) to ensure accurate representation of platform learning outcomes. Admin activity accounts for 39 of the 150 activity records, which were excluded from all KPI calculations.

5. Tools & Technologies

Tool	Purpose	Usage in Project
Mockaroo	Synthetic data generation	Generated all 5 CSV datasets with realistic values
SQL Server	Data warehouse database	Hosts Bronze, Silver and Gold layer tables

DBeaver / SSMS	Database management	Used to run SQL scripts and manage the warehouse
dbdiagram.io	Schema design	Designed Bronze layer schema with visual ER diagram
Power BI Desktop	Dashboard analytics	Built initial versions of both analytics dashboards
HTML / CSS / JS	Interactive dashboards	Final production dashboards built as standalone web files
GitHub	Version control	Repository for all project files, CSVs, and dashboards
VS Code	Development environment	Used for all file editing, Git operations, and README creation
Excel	Data post-processing	Applied ID prefixes and verified data integrity

6. Analytics Dashboards

6.1 Dashboard 1 — Platform Analytics

The Platform Analytics Dashboard answers five core business questions about learning effectiveness and engagement on the Lumen-Ra platform.

Key Performance Indicators

- Total Learners: 39 (after admin exclusion)
- Total Activities: 111
- Completed Activities: 41 (37% completion rate)
- Average Quiz Score: 62.9 out of 100
- Average Progress: 51.8%

Business Questions Answered

- Which modules have the highest completion rates? — Emotional Support leads with 5 completions
- Which user segments show the highest engagement? — Management department with 37 activities
- How effective are different content formats? — Text modules score 65.96 vs video at 61.25
- Is knowledge improving over time? — Score trends visualized by date
- Which modules perform best overall? — Responding to Disclosure Calmly leads with avg score 82.0

6.2 Dashboard 2 — Retention & Impact

The Retention and Impact Dashboard answers four business questions about user behaviour, return rates, drop-off patterns, and what predicts learning success.

Key Performance Indicators

- Return Rate: 79% (31 of 39 learners returned for more than one session)
- Average Sessions Per User: 3.2 activities
- Drop-off Events: 18 activities where users started but never progressed
- Top Quiz Score: 99.2 (Guenna FitzAlan, M014 — Long-Term Allyship)
- Learner Funnel: 39 registered → 35 active → 31 returned → 27 completed at least one module

Business Questions Answered

- Are users returning after learning? — Yes, 79% return rate confirmed
- Which users create the most value? — Mariejeanne Lehr and Cornell Raikes lead with 7 activities each
- What behaviour predicts completion? — Learners with 4+ activities score 8.4 points higher on average
- Where do users drop off? — Self Care and Long-Term Allyship modules show highest drop-off

6.3 Dashboard Versions

Two versions of each dashboard were produced during the project:

- Power BI version (.pbix) — Built during initial development using Power BI Desktop, stored in /dashboard_1 and /dashboard_2 folders
- HTML version (.html) — Final production version built as standalone interactive web files, stored in /latest_dashboards folder, color matched to the Lumen-Ra prototype design

The HTML dashboards use the official Lumen-Ra color palette — cream background (#F5F0E8), dark teal (#1C3333), terracotta orange (#C4622D), and sage green (#7A9E87) — and are accessible in any web browser without any software installation.

7. Key Findings

7.1 Learning Effectiveness

- Text-based modules outperform video modules by 4.7 points on average (65.96 vs 61.25)
- Responding to Disclosure Calmly is the highest scoring module at 82.0 average

- Emotional Support Without Overstepping has the most completions (5) making it the most engaging module
- Asking the Right Questions has the lowest average score (45.9) suggesting content may need revision

7.2 Engagement & Retention

- 79% of learners returned for more than one session — strong early retention signal
- Management department is the most active segment with 37 activities
- Community department has low engagement (14 activities) suggesting outreach opportunity
- HR has the lowest engagement (4 activities) with only 1 learner
- Learners who engage 4 or more times score 8.4 points higher on average

7.3 Drop-off Analysis

- 18 activity events show status of 'started' with no progression — 16% drop-off rate
- Self Care module has the most drop-offs (3 started events) despite also having 4 completions
- Safe Referral has 3 started events suggesting the content may feel intimidating to users
- Body Language module has 0 completions across 2 activities — requires urgent content review

8. Repository Structure

All project files are version controlled on GitHub at:
github.com/WANGAR1/Lumen-Ra-Analytics-Dashboard

- /data — All five CSV files generated via Mockaroo
- /schema — Bronze layer ER diagram exported from dbdiagram.io
- /dashboard_1 — Power BI file for Platform Analytics Dashboard
- /dashboard_2 — Power BI file for Retention and Impact Dashboard
- /latest_dashboards — Final HTML versions of both dashboards
- /documentation — This documentation file

Each folder contains its own README.md file explaining the contents. The root README.md provides a full project overview.

9. Skills Demonstrated

- Data Architecture — Designed Medallion Architecture (Bronze, Silver, Gold) for a real analytics use case
- Database Schema Design — Created relational schema with primary and foreign keys using dbdiagram.io
- Synthetic Data Engineering — Generated realistic multi-table datasets with referential integrity using Mockaroo
- Data Cleaning — Identified and excluded admin accounts, validated data types and ranges
- ETL Pipeline Design — Planned Extract, Transform, Load pipeline from CSV to SQL Server to analytics layer
- Business Analytics — Translated business questions into measurable KPIs and data visualizations
- Dashboard Development — Built interactive dashboards in both Power BI and HTML/CSS/JavaScript
- Version Control — Organized and maintained project repository on GitHub using Git
- Data Storytelling — Communicated findings through visual dashboards matched to product design language

10. Limitations & Future Work

10.1 Current Limitations

- All data is synthetic — findings reflect simulated patterns, not real user behaviour
- No Silver or Gold layer SQL scripts were written during the capstone timeline due to time constraints
- Survey tables (pre and post assessment confidence data) were designed but not populated
- Certification tracking was planned but not implemented in the current schema

10.2 Recommended Next Steps

- Connect dashboards to live SQL Server database when platform launches
- Build Silver layer cleaning scripts in Python (Pandas) to automate data quality
- Implement Gold layer star schema with fact and dimension tables for optimized querying
- Add confidence assessment tracking using pre and post survey data
- Expand user demographics table to capture age range and region for segment analysis
- Set up automated KPI refresh schedule so dashboards update without manual intervention

Data Science & Engineering · February 2026