Week 2 Task: Data Modeling and Integration

Internship: Virtual Power BI Data Insights Internship

Prepared by: Mohammad Saniya

Date: June 22, 2025

1. Objective

The objective of is to build a detailed Power BI data model by integrating multiple datasets, applying transformation procedures, and establishing structured relationships to support meaningful analysis and reporting. This task simulates real-world data modeling and focuses on clean, optimized, and interconnected data.

2. Identify Data Sources

- **Student Performance Dataset (Kaggle):** Includes student demographics, exam scores, parental education levels, and lunch types.
- **Student Attendance Dataset (Open Government Portal):** Records attendance on a per-day basis, linked to student IDs and academic calendar dates.

3. Data Cleaning and Preparation Methods

- Null & Missing Values: Removed rows with missing student_id, score, or attendance values.
- Standardization: Renamed columns using consistent format (e.g., student_id).
- Data Type Fixes: Converted date and numeric fields correctly.
- Categorical Mapping: Changed encoded values to readable terms (e.g., G → Female).
- Derived Columns: Created total_score, average_score, and attendance_percentage.

In Power BI's Power Query Editor, I applied various transformation procedures in addition to basic cleaning:

- **Split Columns**: Separated full name into first_name and last_name to normalize student information.
- **Replace Values**: Replaced null values in attendance status with "Absent" for consistency.

- **Created Derived Columns**:
    - total_score: calculated as the sum of scores across all subjects.
    - attendance_flag: flagged students as "Low" if attendance < 75%, else "Good".
- **Standardized Formats**: Converted gender and status values to uppercase using Text.Upper() for consistency.
- **Unpivoted Subject Data**: Transformed wide-format subject scores into long format (subject, score) to allow dynamic subject-level analysis.

Data types were explicitly defined during transformation to support consistent modeling and avoid calculation issues:

- student_id, subject, class, status, attendance_flag → **Text**
- score, total_score, attendance_percentage → **Decimal Number**
- date → **Date/Time**

These data types ensured that relationships, filters, and DAX calculations performed without error across the model.

4. Design Data Model

The data model follows a **star schema design**, with the students table acting as the central **dimension table.** It connects to two **fact tables —** performance and attendance — using **one-to-many relationships** via the common key student_id.

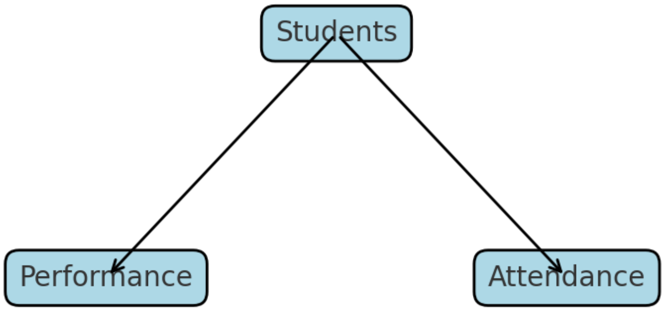
- Each student has multiple entries in the performance table representing their test scores across subjects.
- Similarly, each student has multiple records in the attendance table for each academic day.

These relationships are defined in **Power BI's Model View**, where:

- students[student_id] → performance[student_id] *(one to many)*
- students[student_id] → attendance[student_id] *(one to many)*

This structure allows for powerful aggregation and filtering based on student demographics (e.g., gender, parental education) while analyzing performance trends or attendance patterns. By following this relational structure, Power BI can apply slicers and filters correctly across tables, supporting interactive dashboards with accurate results.

The decision to use a **star schema** was based on Power BI's preference for simple, relational models that optimize performance and visual responsiveness. I placed Students at the center as the key dimension, with Performance and Attendance as fact tables linked via student_id. Other supporting dimension tables (Date, Subjects, Class) enable better filtering and time-based slicing of the data. This structure also supports accurate DAX calculations and user-friendly visual design.



5. Integration Techniques

- **Imported Data:** Used Get Data > Text/CSV.
- **Power Query Editor:** Cleaned and transformed raw data.
- **Merged Queries:** Joined tables on student_id.
- **Model View:** Defined one-to-many relationships.
- **DAX Measures:** Created average score and attendance %.

6. Issues & Resolutions

Issue	Resolution
Duplicate student IDs	Removed duplicates using composite keys in Power Query.
Mismatched date formats	Used 'Detect Data Type' to convert all date fields.
Missing attendance entries	Used 'Group By' and 'Fill Down' to impute averages.
Column name conflicts	Renamed overlapping columns before merging.

7. Summary

This report demonstrates practical application of Power BI's data modeling and integration tools in the context of an educational dataset. By cleaning raw data, defining relationships, and applying best practices in Power Query and DAX, I've constructed a scalable data model that can power insightful dashboards in the coming weeks.