Milestone 1 Project Report: Data Engineering & Star Schema Modelling

Project Name: Hotel Revenue Analysis (2018–2020)

Internship Submission: Milestone 1

Author: Krishnarjun Shaji

1. Project Objective

The goal of Milestone 1 was to create a clean, high-performance data model for hotel revenue analysis. This involved taking raw, fragmented data from three different years and transforming it into a professional Star Schema. This architecture ensures that future reports (Milestone 2) will be accurate, fast, and easy to filter by hotel, date, room, or customer type.

2. ETL (Extract, Transform, Load) Process

Using the Power Query Editor, I performed the following data engineering steps to prepare the dataset:

- Data Consolidation: Appended the 2018, 2019, and 2020 historical datasets into a single master fact table named Raw_fact.
- Data Cleaning: Conducted a quality check on the children column. All identified errors and null values were replaced with 0 to ensure mathematical consistency.
- Unique Indexing: Created a booking_id column using an Index starting at 1 to serve as the Primary Key for every transaction in the fact table.

3. Advanced Feature Engineering

Custom logic was applied to translate raw categories into usable financial metrics and relational keys:

A. Financial Costing & Discount Mapping

New columns were created using conditional logic based on the organization's price list:

- Meal Costing: * BB: 12.99 | HB: 17.99 | FB: 21.99 | SC: 35.00 | Undefined: 0
- Segment Discounts: * Complementary: 1.0 (100%)
    - Online/Offline TA: 0.3 (30%)
    - Aviation: 0.2 (20%)
    - Corporate: 0.15 (15%)
    - Direct/Groups: 0.1 (10%)

B. Relational Key Development

- Hotel ID Mapping: Standardized the hotel categories where Resort Hotel is mapped to H01 and City Hotel to H02.
- Room ID Mapping: Converted the room type letters (A–P) into numeric IDs (1–10).
- Numeric Date Key: Created a specialized date_key using the formula:

$$Date.Year \times 10000 + Date.Month \times 100 + Date.Day$$

4. Dimensional Modeling (The Star Schema)

To optimize the data for business intelligence, I moved the data from a single "flat" table into a normalized Star Schema.

Dimension Table	Content	Relationship Key
Dim_Hotel	Hotel types and standardized IDs	hotel_id
Dim_Date	Unique calendar dates and numeric keys	date_key
Dim_Room	Room type descriptions and numeric IDs	room_id
Dim_Customer	Guest demographics and repeat status	customer_id

The Customer ID Bridge: A unique customer_id was created in the Customer dimension. This was then merged back into the Raw_fact table using a multi-column join (Adults + Children + Babies + Country) to ensure every transaction is linked to a unique guest profile.

5. Model Relationships & Integrity

In the Model View, the following logic was verified:

- Cardinality: All relationships are set to One-to-Many ($1:*$) from the Dimension tables to the Fact table.
- Filter Flow: The cross-filter direction is set to Single, ensuring that filters selected in a dimension table correctly propagate to the facts.
- Integrity: Verified that no duplicate IDs exist in the Dimension tables.

6. Conclusion

The foundation for the Hotel Revenue Dashboard is now complete. With a fully functional Star Schema, the model is prepared for Milestone 2, where I will implement DAX measures to calculate Total Revenue, Net Profit, and Occupancy Rates.