

# Strategic HR Analytics & Cleaning and Modeling Stage

From Raw Data to a Robust Star Schema Architecture

**Project Context: Digital Egypt Pioneers Initiative (DEPI) - Graduation Project**

**Team Members:**

- **Osama El-Sarsawy** (Team Lead)
- Mohamed El-Bastawiy
- Alaa Osama
- Mariam El-Badry
- Taha Khalifa



# Project Roadmap

01

## Project Scope & Vision

Why are we doing this?

02

## The "Before" State

Assessing Data Chaos & Quality Issues.

03

## Phase I: Data Cleaning & Preparation (ETL)

Transformation Logic & Quality Assurance.

04

## Phase II: Advanced Data Modeling

Architecture & The "Reference Table" Strategy.

05

## Feature Engineering

Creating Intelligence via DAX.

06

## Technical Challenges & Solutions

Overcoming Complexity.

# Transforming HR Data into Strategic Assets

- **Objective** To move beyond static reporting and build a dynamic BI Ecosystem capable of diagnosing Attrition, predicting Burnout, and evaluating Talent Performance. This allows for proactive decision-making rather than reactive responses.
- **The Dataset Scope**
  - **Demographics:** 1,470 Employees (Age, Gender, Education levels, Job Role, etc.).
  - **Transactions:** 6,000+ Historical Performance Reviews, including ratings and feedback over time.
  - **Lookups:** 4 Dimension Tables (Education Field, Satisfaction Level, Performance Ratings, Experience Levels) to provide context.
- **Technical Goal** Achieve a "Single Source of Truth" through a rigorous Star Schema model. This ensures data consistency, reduces redundancy, and simplifies query design for analysts.



# Initial Data Assessment & Challenges

## The "Before" State (Data Audit)

### Fragmentation

Data existed in 5 disconnected CSV files (Employee.csv, PerformanceRating.csv, SalaryHistory.csv, etc.), making it impossible to get a holistic view of HR metrics without manual effort.

### Ambiguity

Critical metrics like Education and Satisfaction were coded as Integers (1-5) without clear text descriptions. This led to misinterpretation and a lack of actionable insights.

### Lack of History

No unified Timeline to track Month-over-Month (MoM) trends or historical changes, hindering longitudinal analysis of employee performance or satisfaction.

### Granularity Mismatch

Employee data (Headcount, Demographics) was at a different level of detail than Performance data (Transactional records), complicating direct analysis without careful aggregation.

Untitled - Power Query Editor

File Home Transform Add Column View Tools Help

New Recent Enter Data Data source settings Manage Parameters Export query results Refresh Preview Advanced Editor Properties Choose Columns Remove Columns Keep Rows Remove Rows Sort Split Column Group By Data Type: Text Use First Row as Headers Merge Queries Append Queries Combine Files

Queries

= Csv.Document(File.Contents("C:\Users\DELL\OneDrive\Desktop\دروازه طبقات HR\Employee.csv"),[Delimiter=",", Columns=23, Encoding=65001,

	EmployeeID	FirstName	LastName	Gender	Age	BusinessTravel	Department	Distance
1	3012-1A41	Leonelle	Simco	Female	30	Some Travel	Sales	27
2	CBCB-9C9D	Leonerd	Aland	Male	38	Some Travel	Sales	23
3	95D7-1CE9	Ahmed	Sykes	Male	43	Some Travel	Human Resources	29
4	47AO-559B	Ermentrude	Berrie	Non-Binary	39	Some Travel	Technology	12
5	42CC-040A	Stace	Savege	Female	29	Some Travel	Human Resources	29
6	C219-6C2E	Clerkclaude	Hinkins	Male	34	Some Travel	Sales	30
7	D906-8674	Uta	Melmar	Female	42	No Travel	Technology	45
8	3C7D-86ED	Joyan	Brason	Female	40	Some Travel	Sales	3
9	3D71-8DC2	Alix	Blazejewski	Male	38	Some Travel	Sales	20
10	5476-CA0D	Kayley	Snoad	Female	31	Frequent Traveller	Technology	4
11	73CF-4956	Hannis	Waslin	Female	32	Some Travel	Technology	42
12	277A-A6FA	Annabela	Pablos	Female	35	Some Travel	Technology	8
13	8BAB-B4A6	Torey	Abram	Male	38	Some Travel	Sales	35
14	111D-E5EF	Edna	Alison	Non-Binary	37	Some Travel	Technology	3
15	97F4-0B14	Vernen	Pownier	Male	33	Some Travel	Technology	4
16	5C03-1009	Willetta	Lurriman	Female	42	Some Travel	Technology	21
17	BD1B-53A3	Wendall	Dryden	Male	43	Some Travel	Sales	27
18	DFA9-990E	Cale	Holston	Male	43	No Travel	Sales	34
19	ED73-F078	Ernaline	Napolione	Female	45	Frequent Traveller	Technology	19
20	C6EC-FEB5	Charlena	Severwright	Female	38	Some Travel	Sales	1
21								

23 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 3:54 PM

27°C مدنی

Search

3:58 PM 11/29/2025

Untitled - Power Query Editor

File Home Transform Add Column View Tools Help

New Recent Enter Data Data source settings Manage Parameters Export query results Refresh Preview Advanced Editor Properties Choose Columns Remove Columns Keep Rows Remove Rows Sort Split Column Group By Data Type: Text Use First Row as Headers Merge Queries Append Queries Combine Files

Queries

= Csv.Document(File.Contents("C:\Users\DELL\OneDrive\Desktop\دروازه طبقات HR\PerformanceRating.csv"),[Delimiter=",", Columns=11, Encoding=65001,

	JobSatisfaction	RelationshipSatisfaction	TrainingOpportunitiesWithinYear	TrainingOpportunitiesTaken	WorkLifeBalance	OverallRating
1	5	1	0	4	4	4
2	4	1	3	4	4	3
3	4	1	2	3	5	4
4	5	2	0	2	3	2
5	3	1	0	4	4	3
6	2	1	0	4	4	3
7	3	2	0	4	4	4
8	4	2	1	5	4	3
9	5	1	1	3	3	2
10	5	1	1	4	5	4
11	4	2	3	4	5	4
12	3	2	0	4	3	3
13	4	2	2	4	5	5
14	5	3	2	3	5	5
15	3	2	2	2	4	3
16	5	4	1	0	5	5
17	4	3	2	3	4	4
18	4	2	1	4	5	5
19	4	1	1	2	3	3
20	4	3	2	3	4	3
21	4	2	0	4	4	3

11 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 3:55 PM

27°C مدنی

Search

3:59 PM 11/29/2025

# ETL Strategy & Data Quality Assurance

## Phase I - Data Cleaning & Preparation (ETL)

### → Strict Type Casting

Enforced specific data types (Dates vs. Text vs. Int64) across all columns to prevent calculation errors and ensure data integrity within the model.

### → Data Profiling

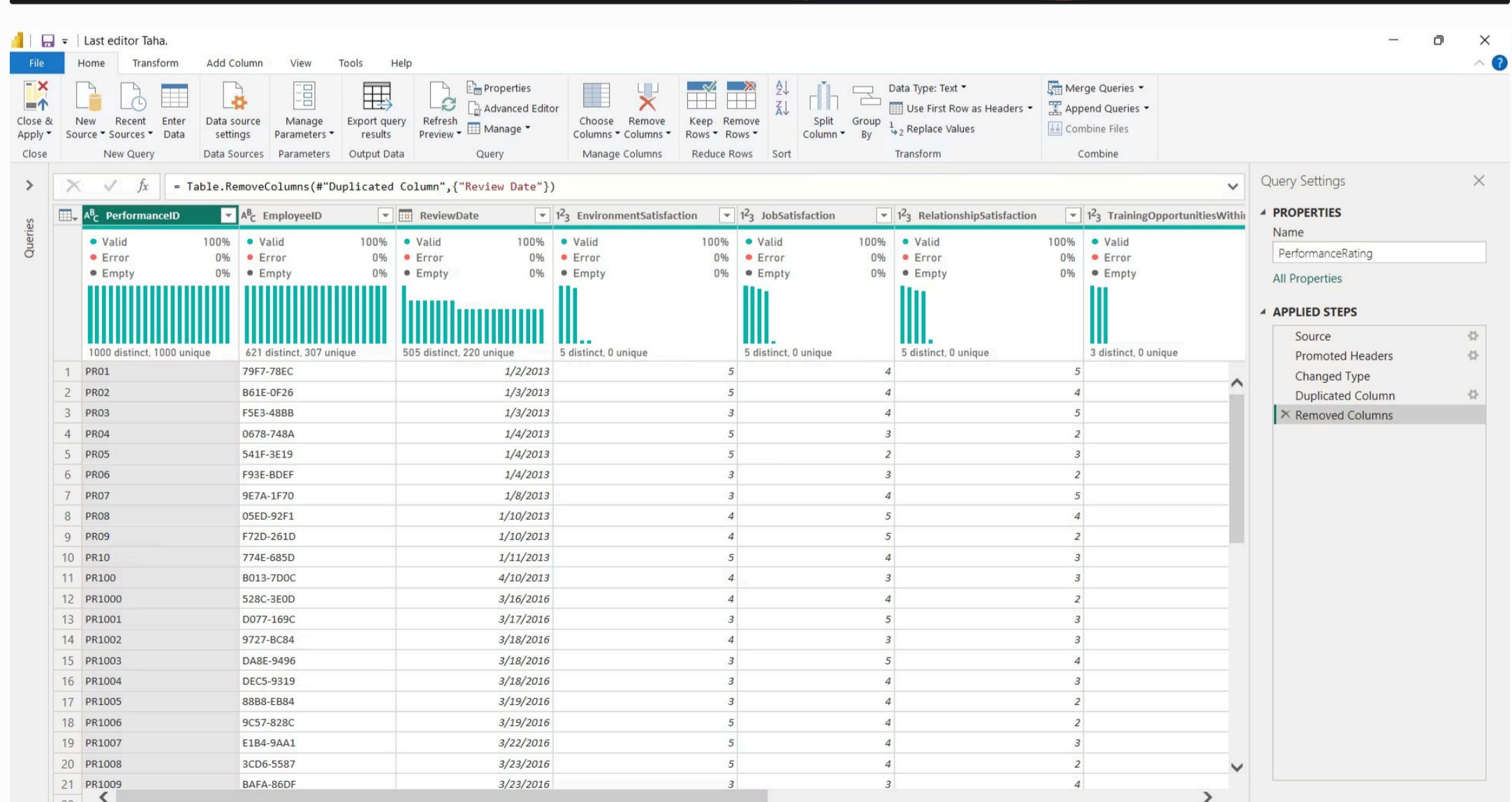
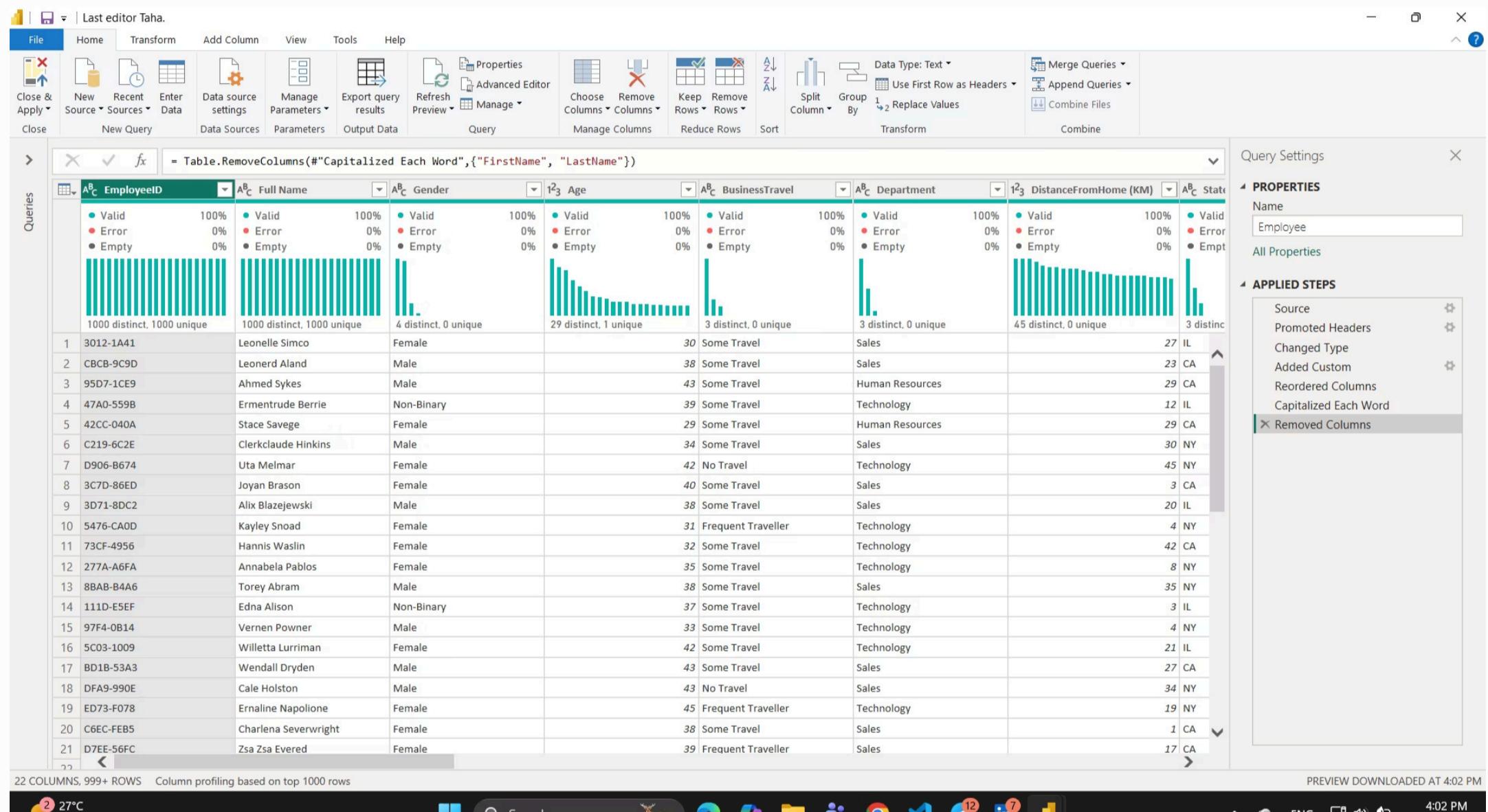
Achieved 100% Column Quality (0 Errors, 0 Empty values) across Primary Keys and critical attributes, ensuring a reliable foundation for analysis.

### → Standardization

Renamed inconsistent columns (e.g., EmpID to EmployeeID) to enable seamless auto-relationship detection and improve readability for future users.

### → Result

A clean, error-free dataset that serves as a robust and reliable source for advanced analytics and reporting, reducing the time spent on data preparation.



# The Architecture: Building the Star Schema

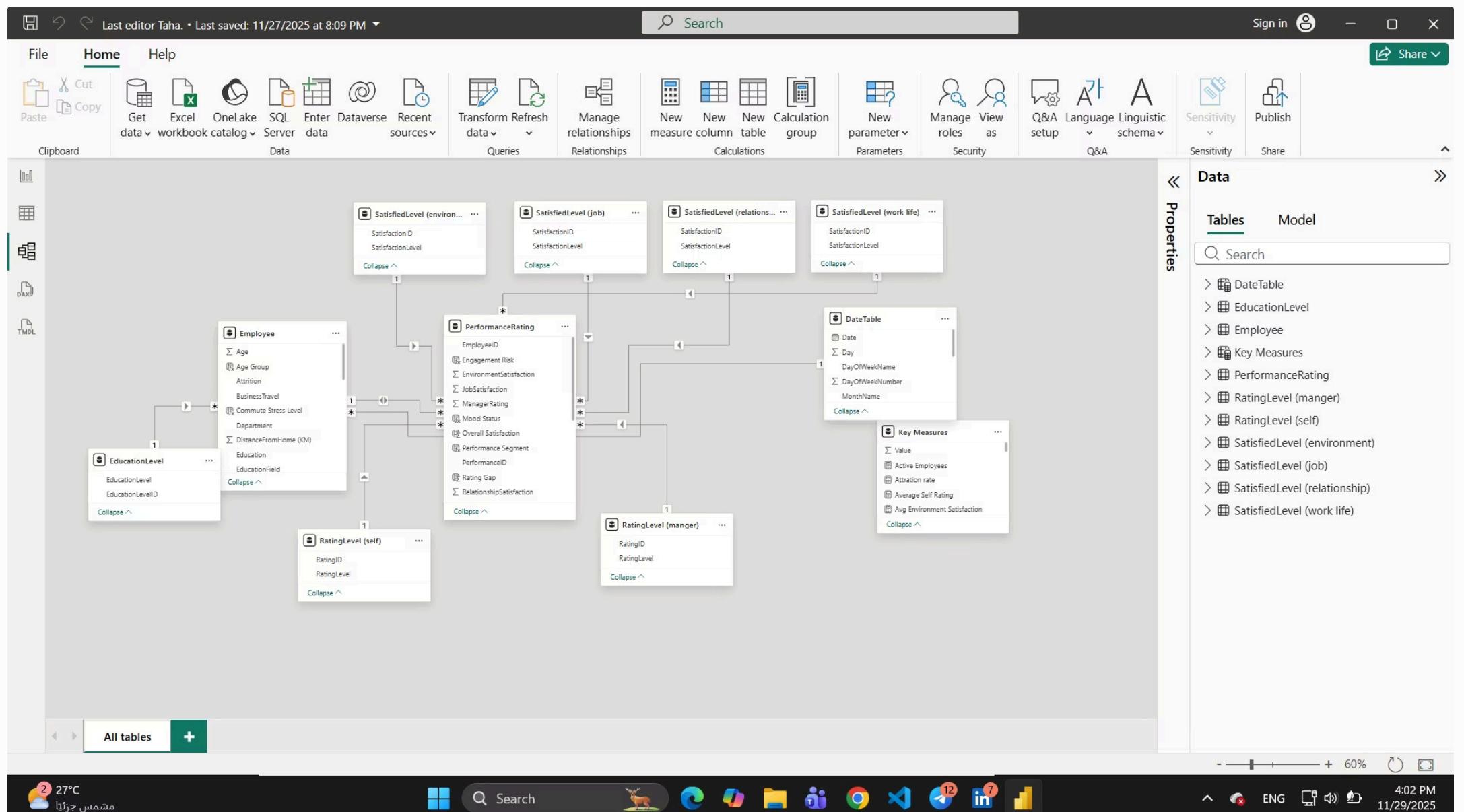
## Phase II - Advanced Data Modeling

### Model Overview

- Fact Table:** PerformanceRating (The central transactional table containing performance metrics and review dates).
- Dimension Tables:** Employee (SCD Type 1 for slowly changing attributes like job title), DateTable (for temporal analysis), Education, and Rating (for descriptive attributes).

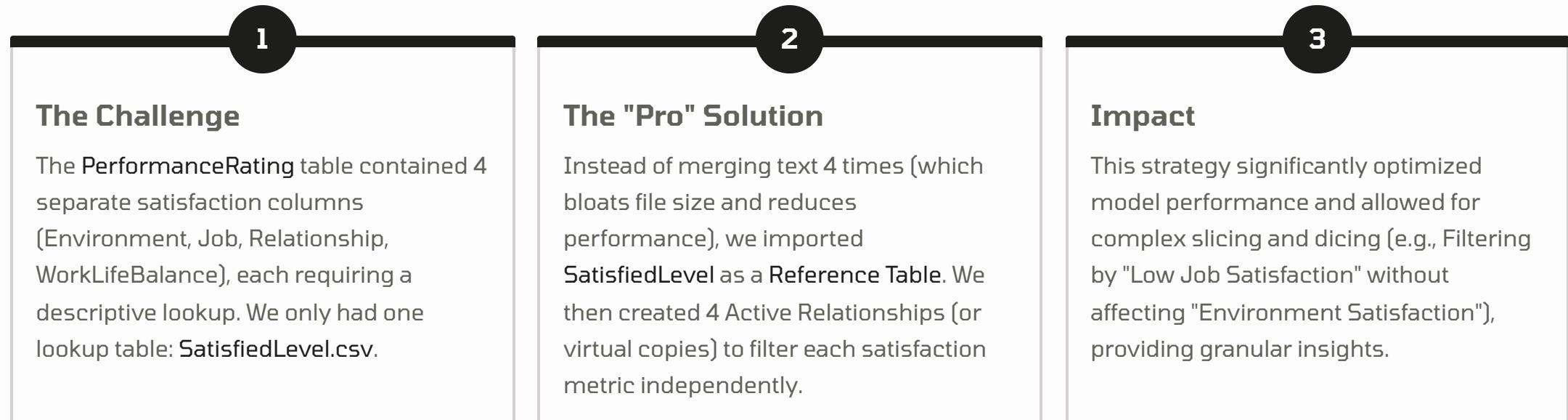
### The Logic

- Established One-to-Many (1:\*) relationships between dimension and fact tables to ensure data integrity and accurate aggregation.
- Enforced Single Direction filtering from dimension to fact tables to prevent ambiguity and circular dependencies, ensuring predictable data behavior.



# Solving the "Role-Playing Dimension" Challenge

## Technical Highlight - The "Reference Table" Strategy



The screenshot shows the Power BI Desktop interface with the following details:

- Home ribbon:** File, Home (selected), Help.
- Data ribbon:** Clipboard, Get data (with options for Excel, OneLake, SQL, Enter data, Data source catalog, Recent sources), Transform data (with Refresh, Queries, Manage relationships), New measure column, New table, New calculation group, New parameter, Manage roles, View as, Q&A setup, Language schema, Sensitivity, Publish, Share.
- Left sidebar:** Tables (Employee, EducationLevel, SatisfiedLevel, PerformanceRating, RatingLevel), DAX, TMDBL.
- Central workspace:** Five tables are shown in a grid:
  - Employee:** Age, Attrition, BusinessTravel, Department, DistanceFromHome (KM), Education, EducationField, EmployeeID, Ethnicity.
  - EducationLevel:** EducationLevel, EducationLevelID.
  - SatisfiedLevel:** SatisfactionID, SatisfactionLevel.
  - PerformanceRating:** EmployeeID, EnvironmentSatisfaction, JobSatisfaction, ManagerRating, PerformanceID, RelationshipSatisfaction, ReviewDate, SelfRating, TrainingOpportunitiesTaken.
  - RatingLevel:** RatingID, RatingLevel.
- Properties pane:** Shows the **Tables** tab with a list of tables: EducationLevel, Employee, PerformanceRating, RatingLevel, SatisfiedLevel.
- Bottom status bar:** Shows system icons (weather, battery, network), search bar, and navigation buttons.

# The Date Table & Temporal Analysis

## Time Intelligence Infrastructure

### The Problem

Raw data lacked a continuous and comprehensive timeline, making it challenging to perform time-based analyses such as trends, period-over-period comparisons, or year-to-date calculations.

### The Solution

We engineered a dynamic Date Table using DAX, populating it with a full range of dates and associated attributes like year, quarter, month, and day of week.

### Relationships

- **Active Link:** To PerformanceRating[ReviewDate] (for analyzing performance trends over time).
- **Inactive Link:** To Employee[HireDate] (activated via USERELATIONSHIP for recruitment analysis and tenure calculations).

### Enabled Analysis

This infrastructure enabled robust Month-over-Month (MoM) Growth calculations, Year-to-Date (YTD) analyses, and detailed Seasonality checks, providing deeper insights into HR dynamics.

### Data

Search

✓ DateTable

>  Date

$\Sigma$  Day

DayOfWeekName

$\Sigma$  DayOfWeekNumber

MonthName

$\Sigma$  MonthNumber

Quarter

$\Sigma$  Year

# Creating Intelligence: Advanced Calculated Columns

## Feature Engineering (Smart Columns)

Concept: We didn't just use the data; we enriched it.

### Highlight 2: The "Exit Reason Classifier"

**Concept:** A logic-based column that intelligently classifies leavers based on their performance and satisfaction data.

- **Burnout:** High Performance / Low Satisfaction
- **Salary Churn:** High Performance / High Satisfaction (suggesting external offers)
- **Involuntary:** Low Performance / Any Satisfaction (indicating performance-based exits)

### Highlight 1: The "Real Promotion" Logic

**Problem:** New hires often had "0 Years Since Promotion", mimicking recently promoted staff, leading to misinterpretation.

**Solution:** We developed a complex DAX query: Real Promotion  
Year = IF(YearsSincePromotion < YearsAtCompany, Year - YearsSincePromotion, BLANK()). This accurately identifies actual promotions versus new employment.

These engineered features provide a richer, more nuanced understanding of HR dynamics, moving beyond raw data to actionable intelligence.

```
1 Professional Turnover Rate (%) =
2 VAR TotalLeavers =
3     CALCULATE(
4         [Total Leavers],
5         ALLSELECTED('DateTable')
6     )
7 VAR CumulativeEmployees =
8     CALCULATE(
9         SUMX(
10            'Employee',
11            1
12        ),
13        FILTER(
14            ALL('Employee'),
15            'Employee'[HireDate] <= MAX('DateTable'[Date])
16        )
17    )
18 VAR TurnoverRate = DIVIDE(TotalLeavers, CumulativeEmployees, 0)
19 RETURN
20 FORMAT(TurnoverRate, "0%")
```

Data

Search

- C F MoM Stagnating
- C F MoM Training utilization
- Female % of Workforce
- High Performers %
- Hires vs. Leavers (This Period)
- Lifetime Avg Performance
- Lifetime Avg Satisfaction
- Male % of Workforce
- MAX
- Median Salary
- MIN
- MoM % Employee
- MoM % leavers
- Mood Status
- New Hire Attrition Rate
- Overall Satisfaction
- Overall Satisfaction (With Emoji)
- Overall Satisfaction MoM%
- Professional Turnover Rate (%)
- Promoted Employees Count
- Promoted Employees Count Mo...
- Role Stagnation Rate
- Stagnating

Key Measures

- Active Employees
- Attration rate
- Average Self Rating
- Avg Environment Satisfaction
- Avg Job Satisfaction
- Avg Job Satisfaction MoM%
- Avg Job Satisfaction MoM% 2
- Avg Manager Rating
- Avg Manager Rating MoM%
- AVG promoation years
- Avg Rating Gap
- Avg Relationship Satisfaction
- Avg Salary
- Avg Salary (High Performers)
- Avg Work-Life Balance
- Avg Years in Role
- C F MoM Job satisfacion
- C F MoM Leavers
- C F MoM manger ratin
- C F MoM overall satsfaction
- C F MoM Promoted
- CF MoM Satisfied

Data

Search

- Promoted Employees Count
- Promoted Employees Count Mo...
- Role Stagnation Rate
- Stagnating
- Stagnating MoM%
- Target Work Life BAlance
- Top Department
- Total Employees
- Total Leavers
- Total Promotions (Count)
- Total Salary
- Total Salary PY
- Training Utilization Rate
- Training Utilization Rate MoM%
- Σ Value
- YOY Salary %
- PerformanceRating
- RatingLevel (manger)
- RatingLevel (self)
- SatisfiedLevel (environment)
- SatisfiedLevel (job)
- SatisfiedLevel (relationship)
- SatisfiedLevel (work life)

# Overcoming Data Complexity

## Technical Challenges & Solutions

1

### Challenge 1: Granularity Mismatch

**Issue:** Trying to analyze "Employee Attributes" (one row per person) against "Historical Reviews" (multiple rows per person) directly could lead to incorrect aggregations.

**Solution:** Utilized DAX Context Transition to calculate "Lifetime Averages" for every employee, ensuring accurate metrics across different granularities.

2

### Challenge 2: Many-to-Many Traps

**Issue:** Connecting a single EducationLevel dimension to multiple fact tables (e.g., performance and hiring data) could create ambiguous or incorrect relationships.

**Solution:** Implemented Bridge Tables and distinct Reference Tables to keep the schema clean, maintain data integrity, and ensure filtering works as intended without circular dependencies.

## Final Deliverable: An Enterprise-Ready Model

### Status

The Data Model is fully optimized, validated, and meticulously documented, ready for deployment and immediate use by HR analysts and leadership.

### Scalability

The architecture supports adding future data sources (e.g., Attendance, Payroll, Training Records) without breaking existing logic, ensuring long-term utility.

### Next Step

This robust model is now the foundational layer for our 6 Strategic Dashboards covering Workforce Dynamics, Performance Management, Employee Well-being, and more.