Project Assignment

🚀 The Business Challenge: Unifying Our Data

Scenario: You've been hired as a Data Architect by a new company, "OmniCorp," that has recently acquired two smaller businesses: a music store (Chinook) and a food and beverage distributor (Northwind). Your first major task is to integrate the data from these two legacy systems into a single, unified data warehouse to enable cross-business analytics and support strategic decision-making. Your solution needs to be robust, scalable, and easy for business analysts to use.

Attachments:

- 2025 dsi310: Introduction Data Engineering
- 2025 dsi310: Data Lake Modeling
- Data Lake Design Template
- dsi310_northwind_chinook_eda_v1.0.ipynb

Lecture Note:

- Chapter 1: Intro to Data Engineering & Data Lake
- Chapter 2: SQL Data Sources & Ingestion
- Chapter 3: API & Log Data Ingestion
- Chapter 4: Data Modeling & OLTP vs. OLAP
- Chapter 5: Dimensional Modeling - Dimensions
- Chapter 6: Dimensional Modeling - Fact Tables
- Chapter 7: ETL - Landing Zone Data Management

---

1. Background Context

The provided schemas show two distinct business domains:

- **Chinook:** A digital media store. Its tables cover artists, albums, tracks, playlists, and sales (invoices).
- **Northwind:** A food and beverage company. Its tables cover products, suppliers, customers, employees, and orders.

The key to a unified data model is to find common entities or processes. Both databases contain information about sales transactions. Specifically, both have:

- **Customers:** Customer (Chinook) and Customers (Northwind).
- **Employees:** Employee (Chinook) and Employees (Northwind).
- **Products/Tracks:** Track (Chinook) and Products (Northwind).
- Transactions: Invoice and InvoiceLine (Chinook) and Orders and "Order Details" (Northwind).
- **Time:** InvoiceDate (Chinook) and OrderDate (Northwind).

These commonalities will form the basis of our unified star schema.

2. Kimball Star Schema Design

We'll design a single **fact table** to represent business transactions and several **dimension tables** that provide context for these transactions.

- **Fact Table:** We will create a FactSales table. This table will contain transactional data, such as sales quantity and dollar amount, and foreign keys linking to our dimension tables. This design allows us to query sales data from both Chinook and Northwind seamlessly.
- **Dimension Tables:** We will create conformed dimensions to unify the data. For example, DimCustomer will contain data from both Customer (Chinook) and Customers (Northwind).

Please use the data lake modeling template in the attached document!!!

2.1 Dimension Tables

Dimension tables are the "who, what, where, when, and how" of the business. They provide descriptive attributes for the data in the fact table.

2.1.1 DimCustomer

This dimension will store information about the customers. It will consolidate data from both the Customer and Customers tables.

- **Attributes:** CustomerID, CustomerName (combining first and last names), CompanyName, City, State, Country, PostalCode, Phone, Email.
- **Purpose:** To slice and dice sales data by customer, company, and location.

2.1.2 DimEmployee

This dimension will store information about the employees who handled the transactions. It will be built from the Employee and Employees tables.

- **Attributes:** EmployeeID, EmployeeName (combining first and last names), Title, City, Country, ReportsTo (for hierarchy).
- **Purpose:** To analyze sales performance by individual employee or by their reporting structure.

2.1.3 DimProduct

This dimension will contain details about the items sold. It will unify data from Track (Chinook) and Products (Northwind).

- **Attributes:** ProductID, ProductName (or Track Name), CategoryName (e.g., Rock, Beverages), GenreName, Composer, UnitPrice.
- **Purpose:** To analyze sales by product, category, genre, or composer.

2.1.4 DimTime

This dimension is crucial for time-series analysis. It will contain granular time attributes.

- **Attributes:** DateKey (YYYYMMDD), FullDate, DayOfMonth, DayOfWeek, Month, Quarter, Year.
- **Purpose:** To analyze sales trends over time, answering questions like "how much did we earn in January?"

2.1.5 DimSourceSystem

This is a simple but powerful dimension to distinguish the origin of the data.

- **Attributes:** SourceSystemID, SourceSystemName (e.g., 'Chinook', 'Northwind').
- **Purpose:** To allow filtering data by the original database, which helps with data quality checks and specialized analysis.

2.2 Fact Table

The fact table contains the quantitative metrics of a business process.

2.2.1 FactSales

This table represents each line item of a sales transaction.

- **Measures:** SalesQuantity (the number of items sold), SalesAmount (the total price for that line item, calculated as UnitPrice * Quantity).
- **Foreign Keys:** DateKey (from DimTime), CustomerID (from DimCustomer), EmployeeID (from DimEmployee), ProductID (from DimProduct), SourceSystemID (from DimSourceSystem).
- **Purpose:** To store the core metrics of sales, enabling calculations like total revenue, average sales, etc., and linking these metrics to the relevant dimensions for context.

A diagram of the proposed schema would look like this: . The fact table FactSales is in the center, surrounded by the dimension tables DimCustomer, DimEmployee, DimProduct, DimTime, and DimSourceSystem.

2.3 Example of a Fact Table Row

A single row in the FactSales table expresses a specific business event. The row links the event's measures (what happened) to the entities involved (who, what, where, when).
An example sentence describing a row in the FactSales table is:
On **January 15th, 2024**, customer **Bob Johnson** purchased **10 units** of **Chai Tea** for a total of **$180** from the **Northwind** business, with the transaction handled by **employee Nancy Davolio**.

3. Assignment Tasks

Task 1: Data Understanding & Source-to-Target Mapping (15%)

1. **Exploratory Data Analysis (EDA):** Use the provided schema and a Colab notebook to perform an EDA on both the Chinook and Northwind databases. Focus on identifying common business entities and processes. Create visualizations to show data volume, key relationships, and data types.
2. **Source-to-Target Mapping:** Based on your EDA, create a mapping document (you can use a spreadsheet, like the provided template) that outlines how columns from the source tables (Chinook.Invoice, Northwind.Orders, etc.) will map to the new unified tables in your data warehouse.

Task 2: Dimensional Modeling & Schema Design (40%)

1. **Design a Kimball Star Schema:** Propose a unified star schema to support common business intelligence questions like:
    - **"What is our total revenue across all business units (music vs. food)?"**
    - **"Who are the top 10 customers based on total spend?"**
    - **"Which products (tracks/items) are our top sellers?"**
    - **"How does sales performance vary by employee across both businesses?"**
2. **Define Your Tables:**
    - **Fact Table:** Define the FactSales table. Explain what a "fact" is in this context. List the specific measures you will include and the foreign keys that will link to your dimensions.
    - **Dimension Tables:** Define the key dimension tables (DimCustomer, DimProduct, DimEmployee, DimDate, etc.). For each, explain its purpose and list the descriptive attributes it will contain. Justify how you will handle different naming conventions and data types from the two source databases.

Task 3: Data Expression & Business Value (25%)

1. **Fact Table Row Expression:** Write an English sentence that a business analyst could understand, describing what a single row in your FactSales table represents. The sentence should clearly incorporate elements from at least three different dimensions to show the power of the star schema. For example: "On **[Date]**, **[Employee Name]** sold **[Quantity]** units of **[Product Name]** to **[Customer Name]** for a total of **[Sales Amount]**."
2. **Report Mock-up:** Design a simple dashboard or report mock-up (e.g., a hand-drawn sketch or a simple diagram) that answers one of the key business questions from Task 2.

Label which tables in your schema would be used to generate each part of the report. This demonstrates your understanding of how the data model supports business intelligence.

Task 4: Critical Thinking & Data Engineering Challenges (20%)

1. **Data Ingestion & Integration:** Discuss the key data engineering challenges you would face when ingesting data from these two disparate systems. Think about data types, primary key conflicts, and data quality issues.
2. **Schema Evolution:** How would your schema handle future changes, like OmniCorp acquiring a third business? What makes the star schema flexible for growth?

Data Lake vs. Data Warehouse: Briefly explain why this approach leverages a data lake (storing raw data) but ultimately serves a data warehouse's purpose (structured for analysis).

4. Grading Rubric and Test Cases

Here is a grading rubric and a set of test cases for the redesigned data science assignment.

Part 1: Grading Rubric

Task 1: Data Understanding & Source-to-Target Mapping (15 points)

- **EDA (10 points):** Comprehensive and insightful analysis of both schemas, correctly identifying common entities and data types. Visualizations are clear and effectively communicate findings.
- **Mapping (5 points):** The source-to-target mapping is logical, accurate, and clearly defines how source columns will be transformed or combined for the new data warehouse tables.

Task 2: Dimensional Modeling & Schema Design (40 points)

- **Star Schema Design (10 points):** The proposed schema is a true Kimball Star Schema with a central fact table and appropriate dimensions. It effectively unifies data from both sources.
- **Fact Table Definition (15 points):** The definition of the fact table is correct, measures are well-defined, and the foreign keys accurately link to the dimensions.
- **Dimension Table Definitions (15 points):** Each dimension table is correctly defined with appropriate attributes. The justification for handling data inconsistencies is logical and sound.
- Please use the data lake modeling template in the attached document!!!

Task 3: Data Expression & Business Value (25 points)

- **Fact Row Expression (15 points):** The example sentence is a clear, concise, and business-friendly expression of a fact table row. It correctly incorporates at least three dimensions.
- **Report Mock-up (10 points):** The mock-up is a clear visual representation of a business report. The annotations correctly link report elements to the proposed schema tables, demonstrating an understanding of how the model supports BI.

Task 4: Critical Thinking & Data Engineering Challenges (20 points)

- **Data Ingestion & Integration (7 points):** The response correctly identifies key challenges like data type conflicts, primary key management, and data quality. The discussion is thoughtful and goes beyond surface-level issues.
- **Schema Evolution (7 points):** The explanation of how the schema handles future growth is logical and highlights the flexibility of a star schema.
- **Data Lake vs. Data Warehouse (6 points):** The distinction between a data lake and a data warehouse is accurately and clearly explained, showing an understanding of their respective roles in a modern data architecture.

Part 2: Test Cases

Here are specific test cases to evaluate the quality and correctness of the student's solution.

Schema Design and Logic

1. **Customer Unification:** Can the student correctly identify and unify Chinook.Customer and Northwind.Customers into a single DimCustomer? Does their design handle the different CustomerID formats (integer vs. text)?
2. **Product Unification:** Does the DimProduct table correctly combine the disparate concepts of Track (Chinook) and Products (Northwind)? Does the student include relevant attributes from both, such as Genre from Chinook and Category from Northwind?
3. **Fact Table Granularity:** Is the FactSales table designed at the correct granularity (e.g., one row per line item per order/invoice)? This is a critical check for correctness. A common mistake is to model it at the invoice/order level, which prevents per-product analysis.

Querying and Analysis

1. **Top N Analysis:** Using their proposed schema, can a BI tool answer the question: "Who are the top 5 customers from the music business (Chinook)?" This requires using the DimSourceSystem and DimCustomer tables and the FactSales table.
2. **Cross-Business Comparison:** Does the schema design allow a simple query to compare total revenue between the two business units? The ideal query would aggregate SalesAmount and group by the SourceSystemName from the DimSourceSystem dimension.
3. **Employee Performance:** Can you slice sales data by employee title? For example, "What is the total sales amount for all 'Sales Support Agent' employees from Chinook?"

Critical Thinking and Justification

1. **Primary Key Strategy:** When combining CustomerIDs from Chinook (integers) and Northwind (text), what strategy did the student propose to create a single, unique primary key in the DimCustomer table? A correct solution would involve prefixing or a hashing function to prevent collisions.
2. **Handling Missing Data:** The Northwind Customers table has a Region column, while Chinook's Customer table does not. How did the student plan to handle this in the DimCustomer table? The correct approach would be to use NULL or a placeholder like "N/A."
3. **Star Schema Justification:** Does the student's explanation correctly justify why a star schema is a better fit for this analytics problem than a normalized schema (3NF)? The answer should highlight benefits like query performance and ease of use for business analysts.