**Documentation for Capstone Project on Data Warehouse Design and ETL Implementation Using Informatica**

**US CHRONIC DISEASE INDICATOR DATASET**

**~ KRUTHIK BALLARI ~**

**PROBLEM STATEMENT**

This is a scalable data warehouse design project aiming at the solution to the problem of complex and voluminous dataset analysis and the derivation of actionable insights from those datasets. Due to an increased demand in businesses for efficient mechanisms in data storage and processing, this capstone project is about building a reliable ETL pipeline using Informatica on transforming raw data into structured information housed within a well-designed data warehouse. The solution will be identifying relevant dimensions and facts, modeling the data warehouse schema, and implementing ETL workflows that ensure data accuracy, consistency, and scalability. Finally, SQL-based queries will be used to extract meaningful insights to support business decision-making processes.

**DATASET DESCRIPTION**

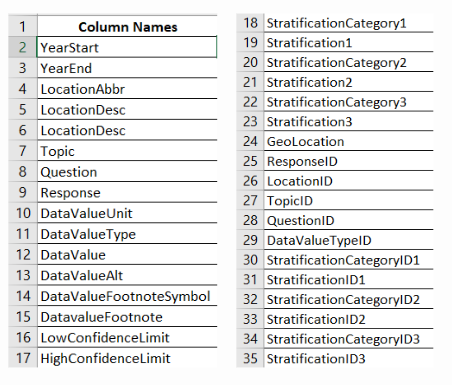
The US Chronic Disease Indicator dataset is a comprehensive data source that provides information on the prevalence and incidence of chronic diseases within the United States. This dataset typically includes key health indicators related to conditions such as diabetes, cardiovascular diseases, obesity, cancer, and respiratory diseases. The data is collected and reported by health agencies and public health departments to monitor trends, support policy-making, and improve public health interventions.

**Total Columns: 34**

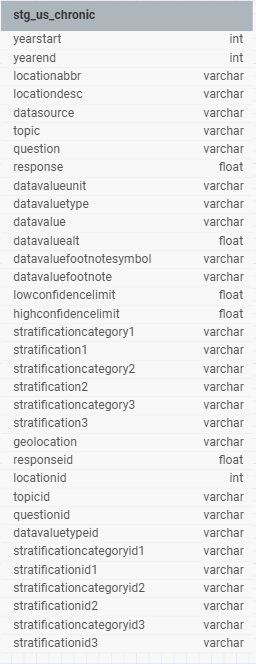
**Total Data Rows: 11,85,676**

**Source Data Type: CSV (Comma Delimited)**

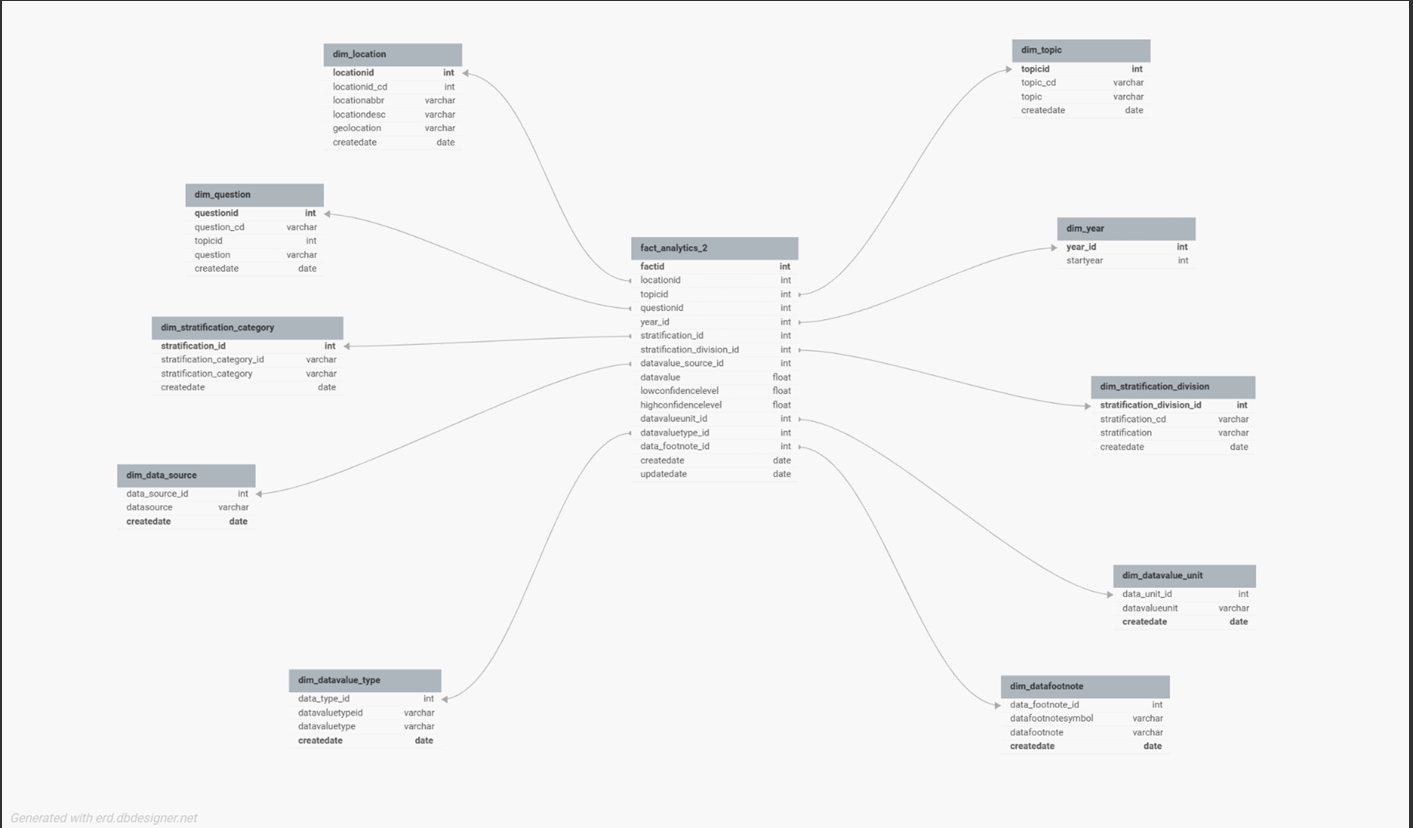
**Dataset Name: US Chronic Disease Indicator Dataset**

****

**Staging Table Design:**

****

**DIMENSION TABLE AND FACT TABLE DB DESIGN**

****

Total Number of Dimension tablesare 10 and the Fact tables is 1

**Steps for Informatica Project on US Chronic Disease Indicator Dataset**

**1. Data Understanding and Modeling**

1. **Dataset Analysis**
   * Analyzed the dataset to identify dimensions (e.g., Location, Time, Disease) and facts (e.g., Counts, Prevalence Rate, Age-Adjusted Rate).
   * Examined data attributes for primary keys, foreign keys, and granularity levels.
2. **Schema Design**
   * **Designed a staging table in MS SQL to hold raw data after extraction.**
   * **Developed a data model comprising:**
     + Dimension Tables: Location, Time, Disease Type, Demographics.
     + Fact Table: Chronic Disease Metrics (with keys to dimensions and measurable facts).

**2. ETL Implementation**

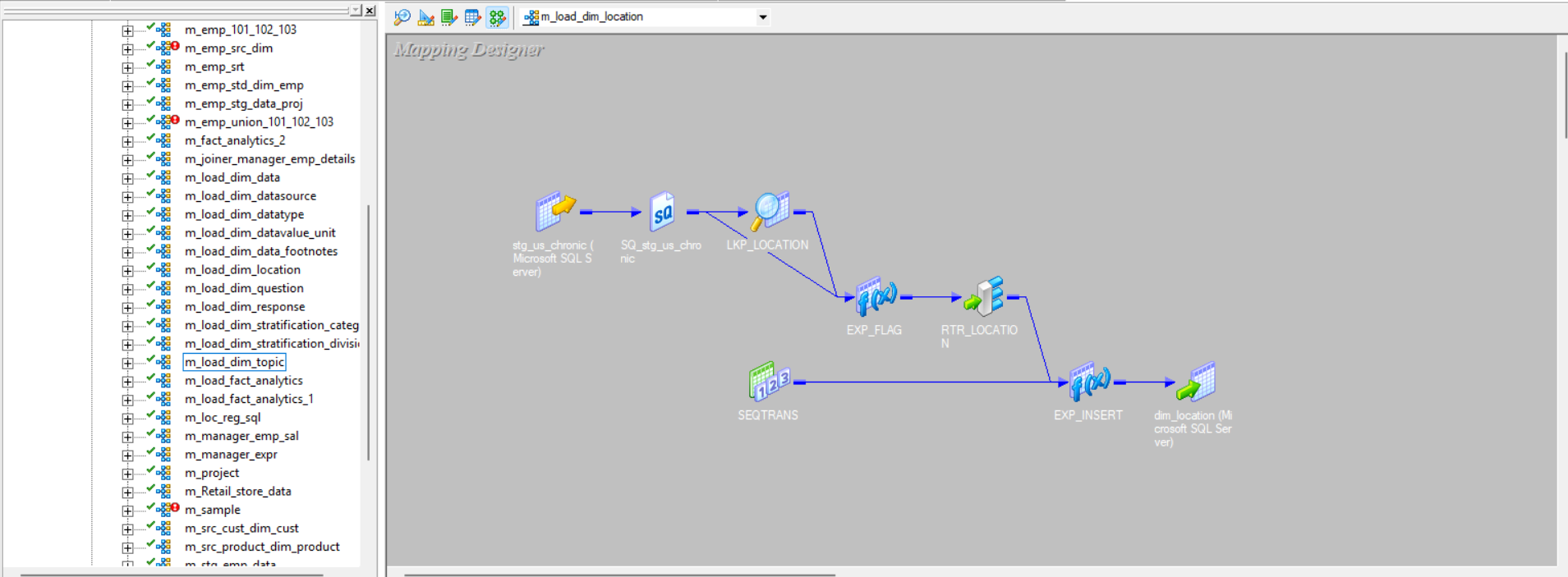
1. **Staging Table Population**
   * Loaded the CSV dataset into the staging table using Informatica mappings.
   * Validated data integrity and handled missing values during extraction and loading.
2. **Mapping for Dimension Tables**
   * **Created mappings for dimension tables using:**
     + Insert Only: For static reference data like locations.
     + SCD Type 1: For attributes where changes overwrite old data, e.g., updated disease descriptions.
     + SCD Type 2: For attributes where historical changes need tracking, e.g., demographic attributes with time validity.
3. **Fact Table Loading**
   * Defined and implemented mappings to load aggregated metrics into the fact table.
   * Ensured correct linkage with dimension tables using surrogate keys.
4. **Workflow Implementation**
   * **Designed workflows to automate the ETL process for all mappings:**
     + Used Event Wait Task to synchronize data loading steps.
     + Applied Command Task for external commands like generating success logs.
     + Parameterized database connections and table names for flexibility.

**3. Data Analysis**

1. **SQL Query Development**
   * **Wrote SQL queries on the dimension and fact tables to extract actionable insights.**
     + Example 1: Identify the top locations with the highest prevalence of a specific chronic disease.
     + Example 2: Analyze trends in age-adjusted rates over time for specific demographics.
2. **Data Validation**
   * Validated data accuracy by cross-checking derived metrics against the staging table and input dataset.

**DIMENSION TABLE MAPPINGS:**

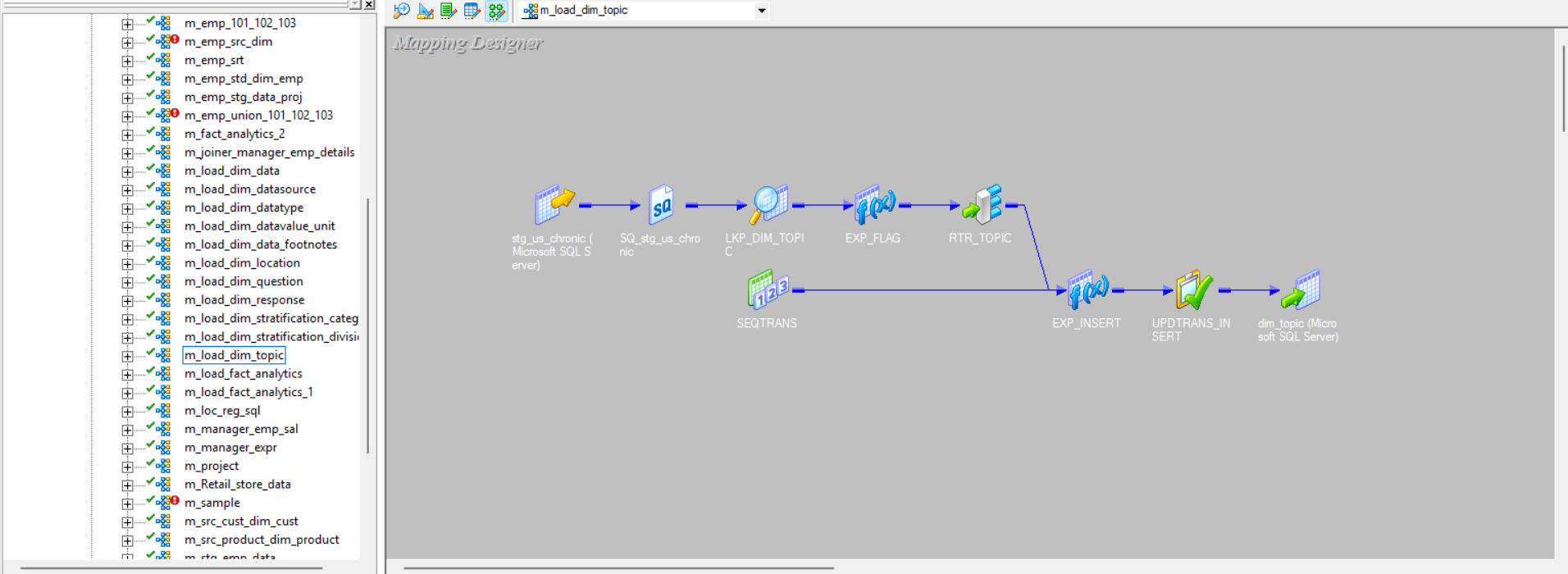
**DIM\_LOCATION:**

****

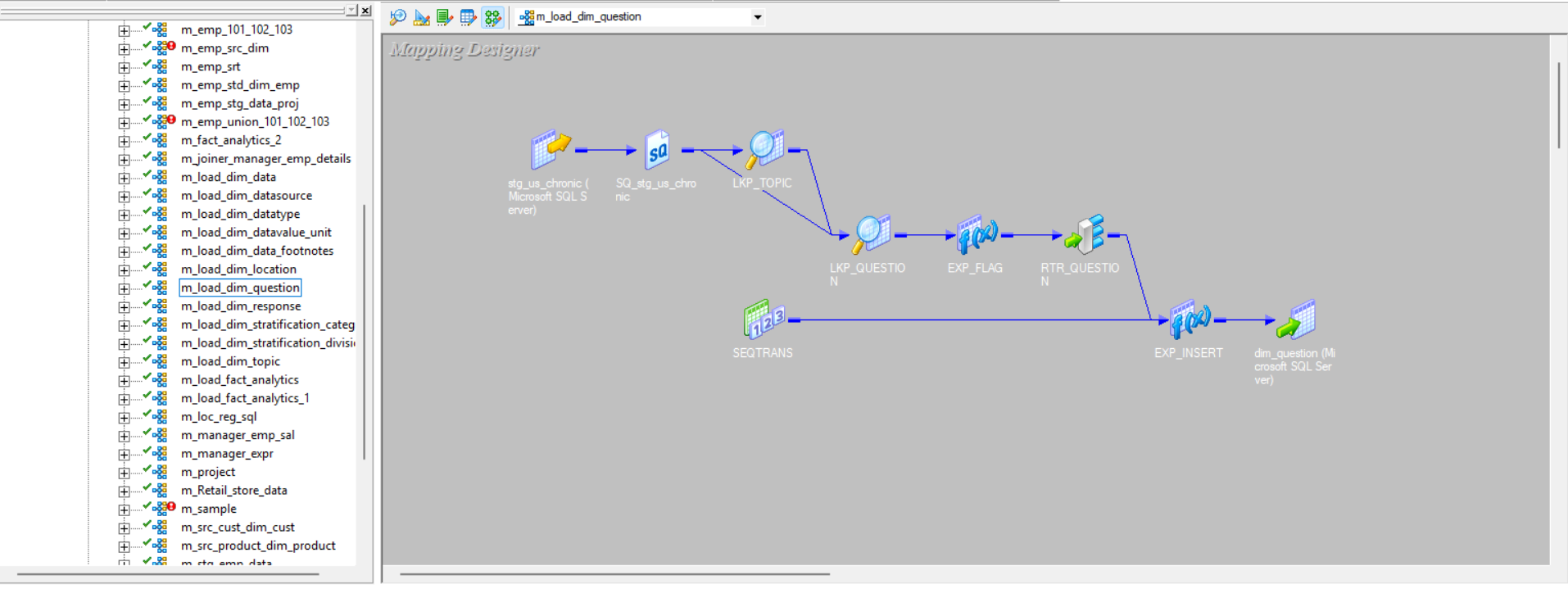
**DIM\_YEAR**

****

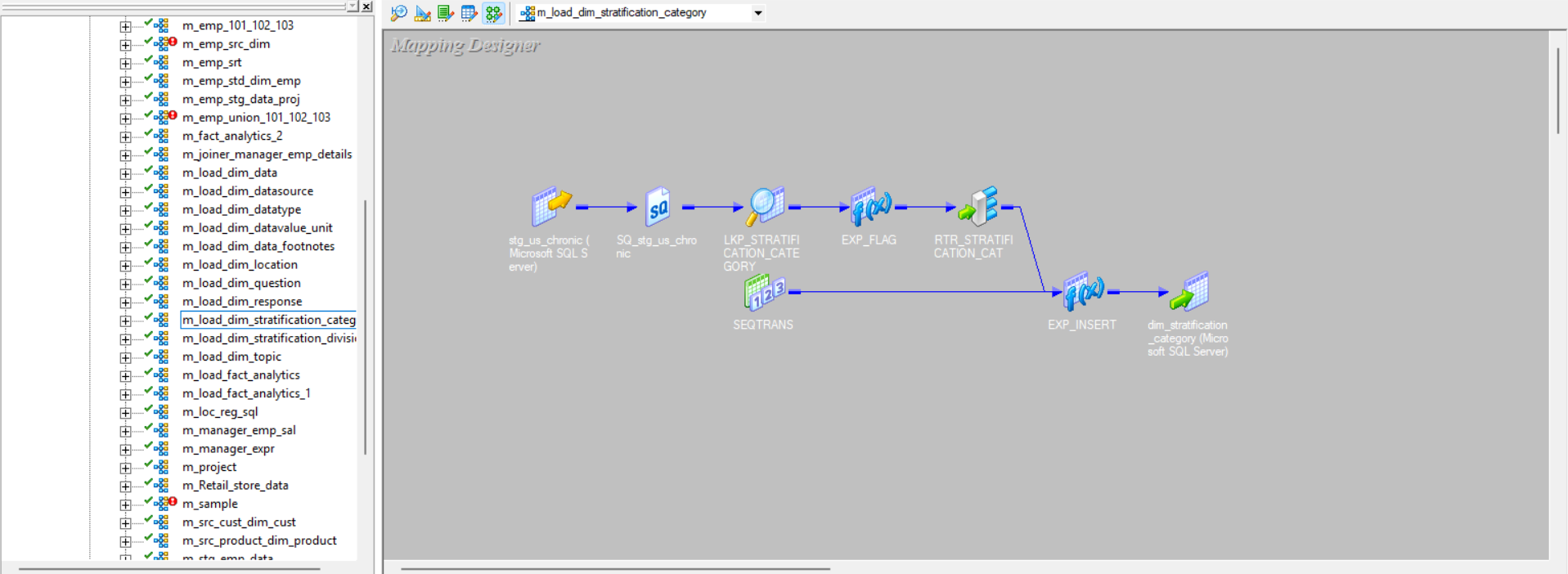
**DIM\_TOPIC**

****

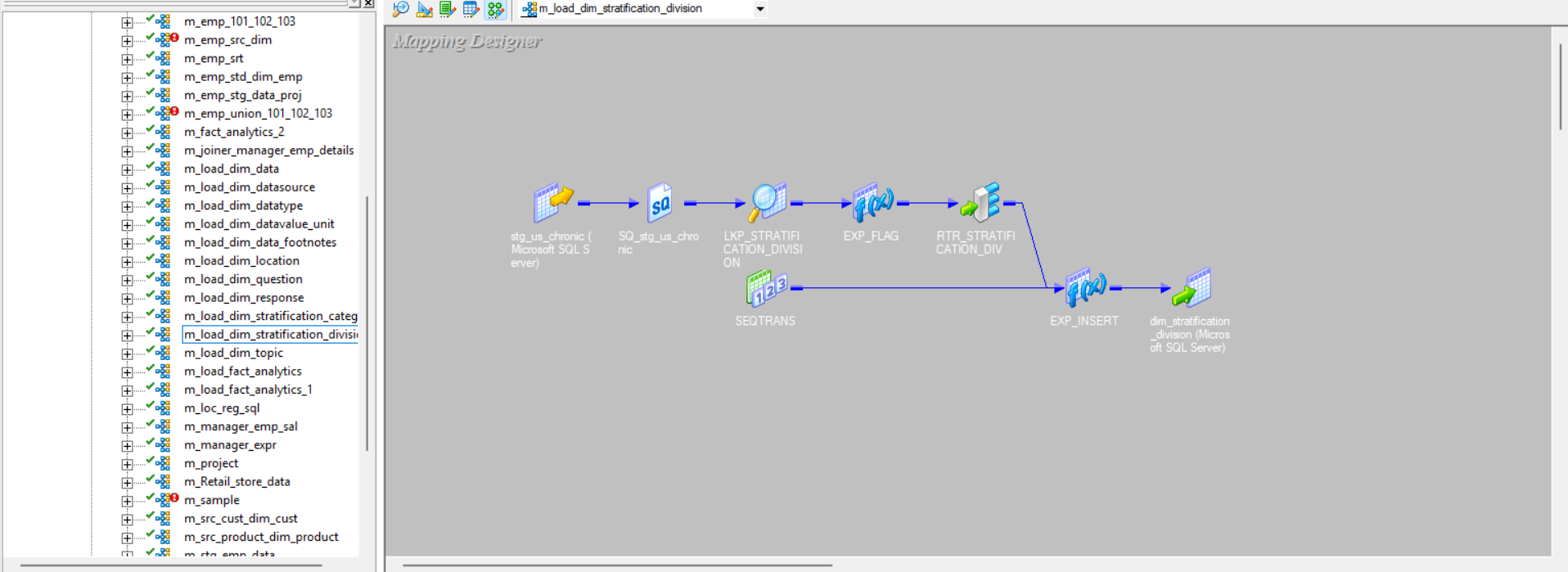
**DIM\_QUESTION**

****

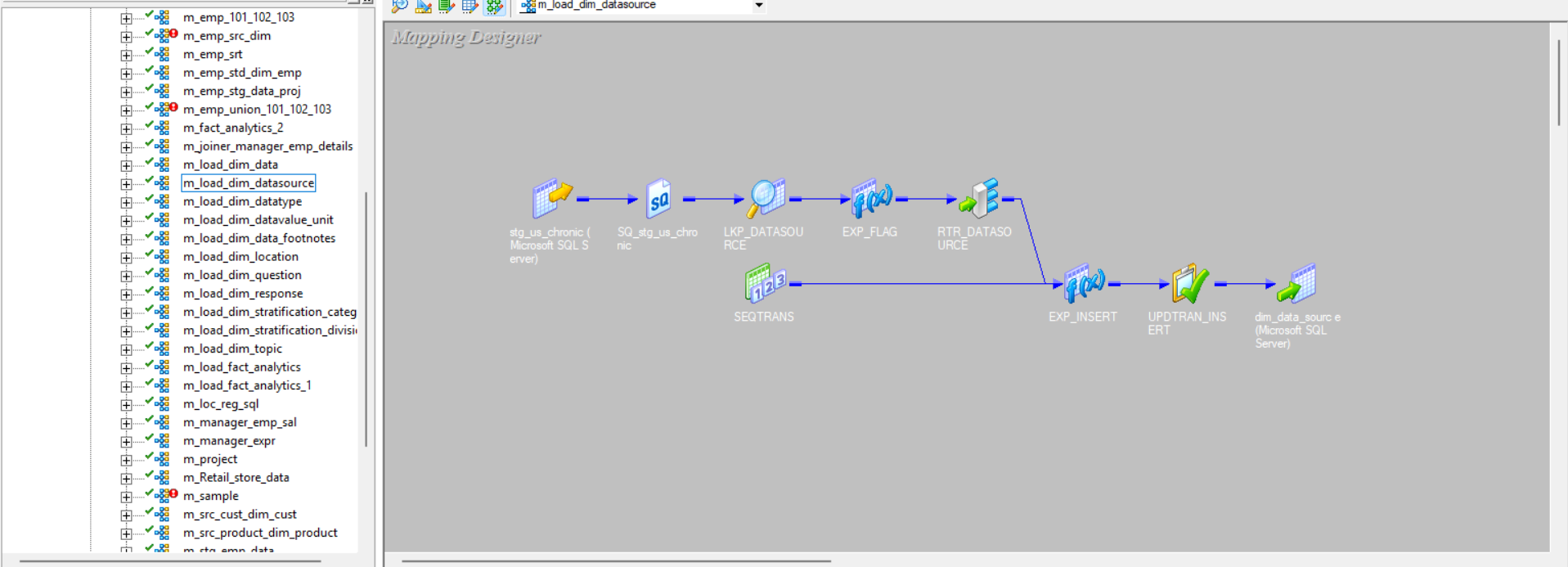
**DIM\_STATIFICATION\_CATEGORY**

****

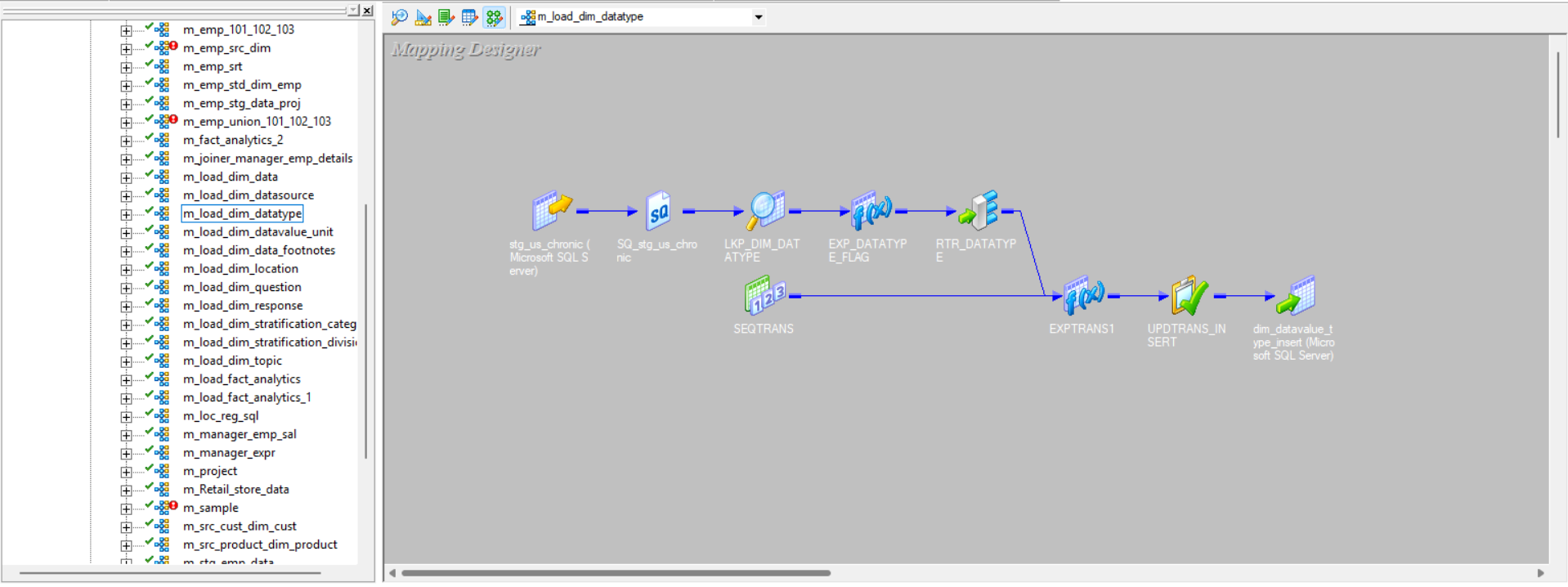
**DIM\_STRATIFICATION\_DIVISION**

****

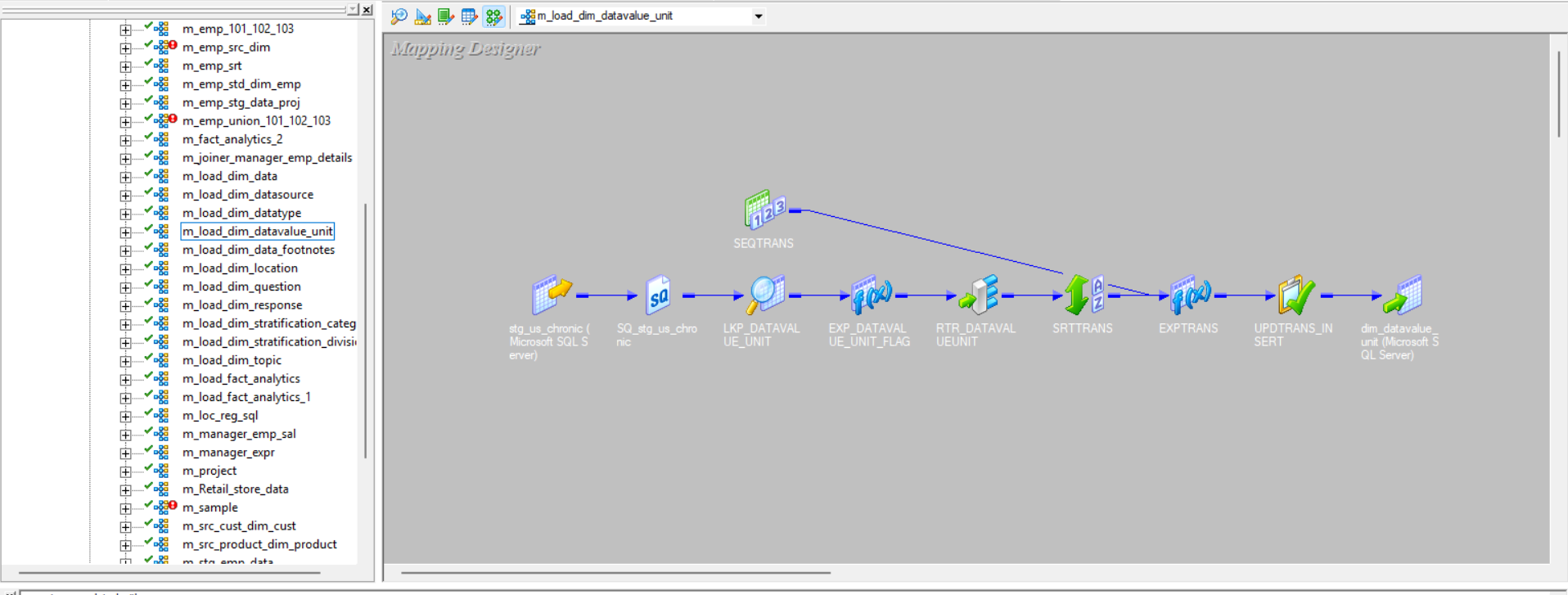
**DIM\_DATSOURCE**

****

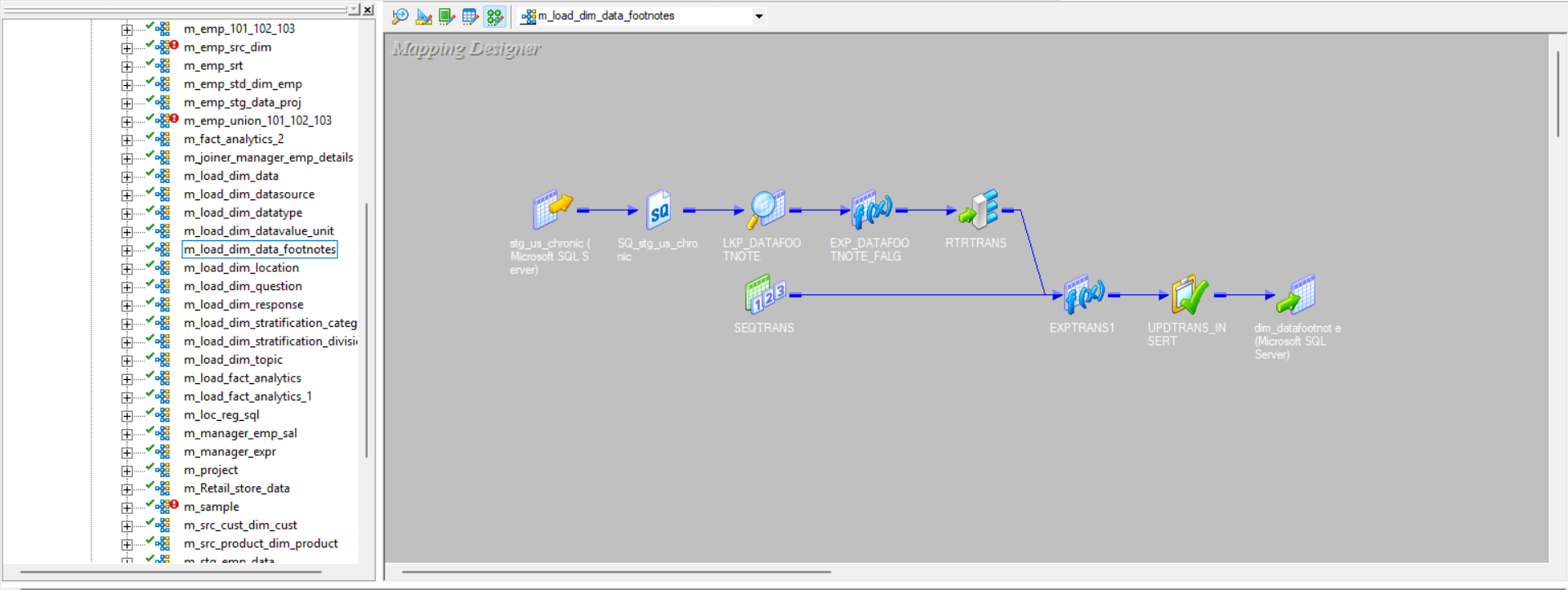
**DIM\_DATATYPE**

****

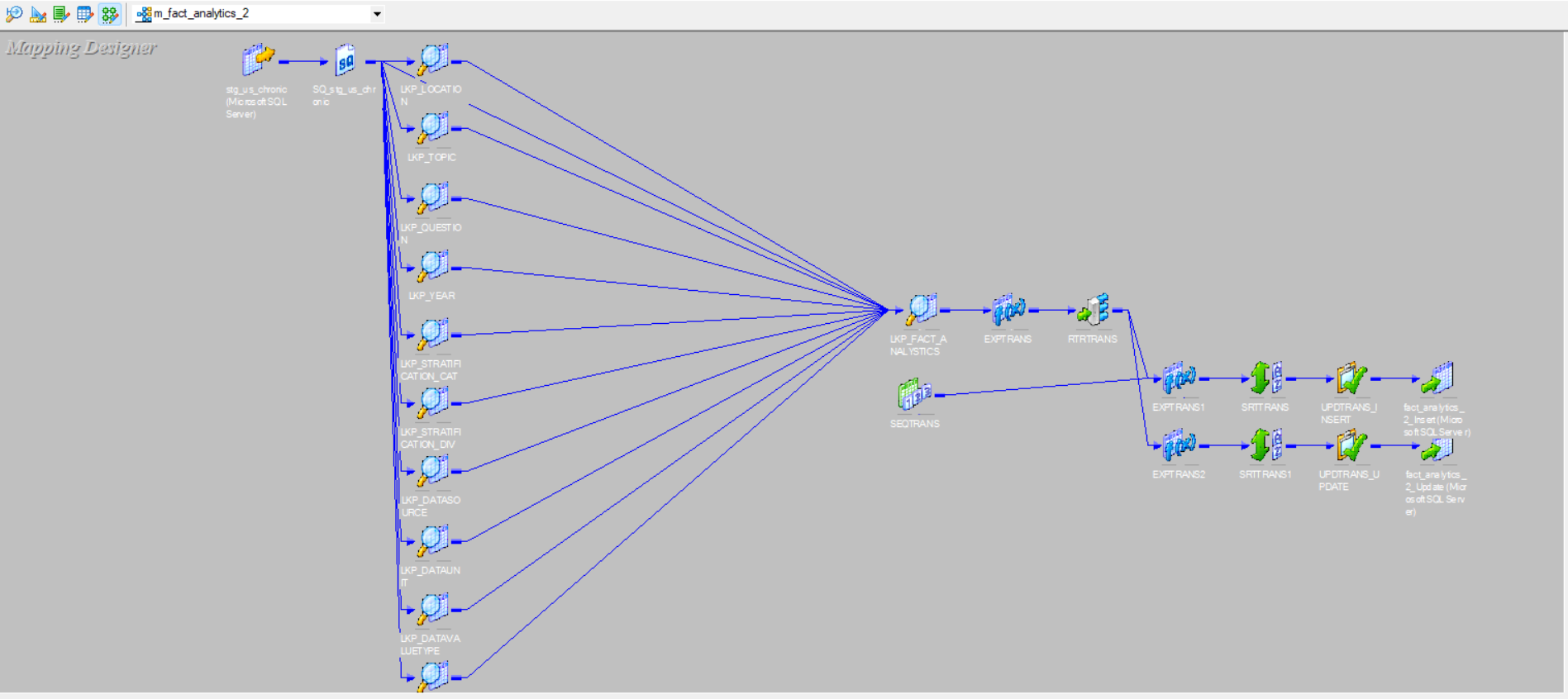
**DIM\_DATAVALUE\_UNIT**

****

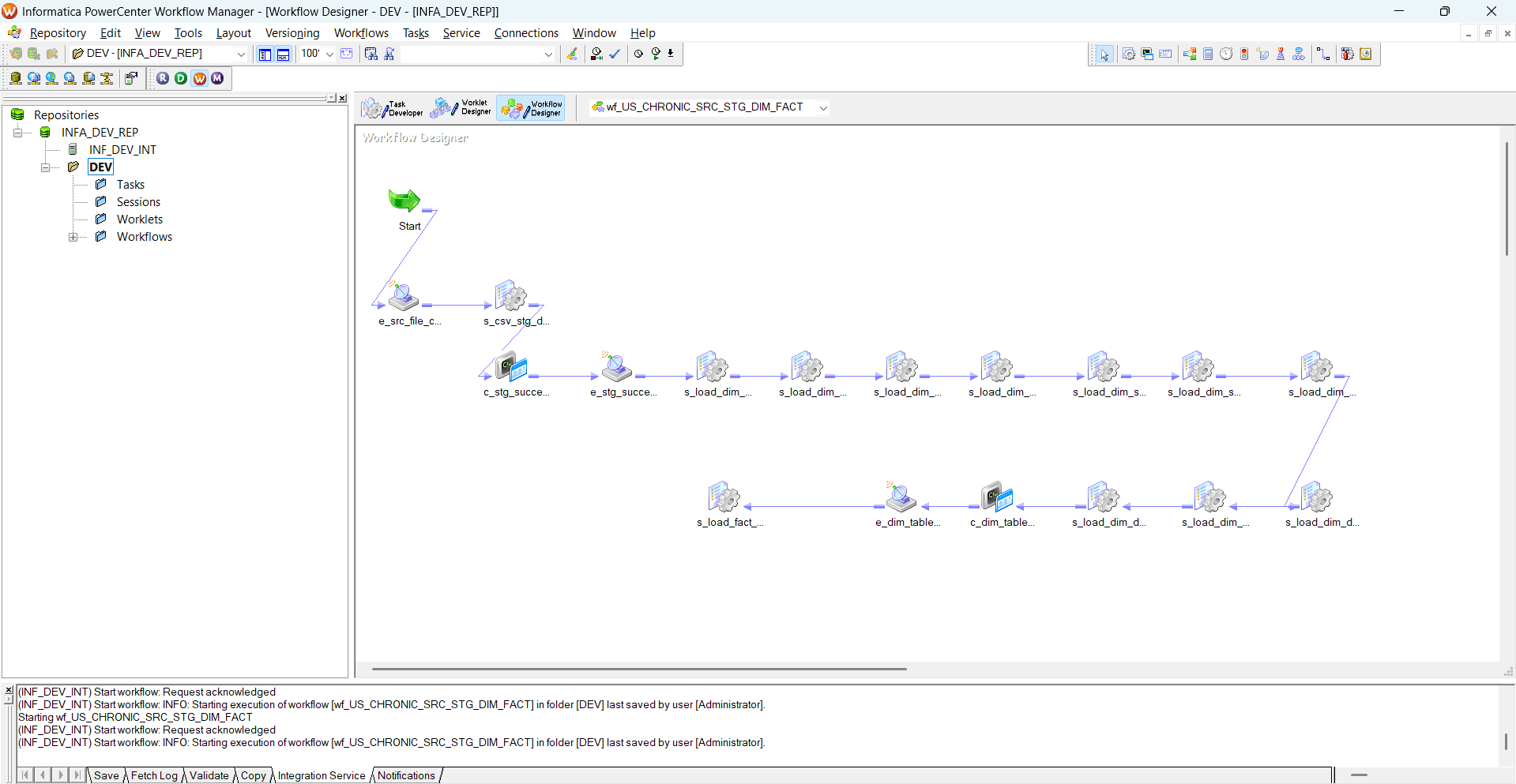
**DIM\_DATA\_FOOTNOTES**

****

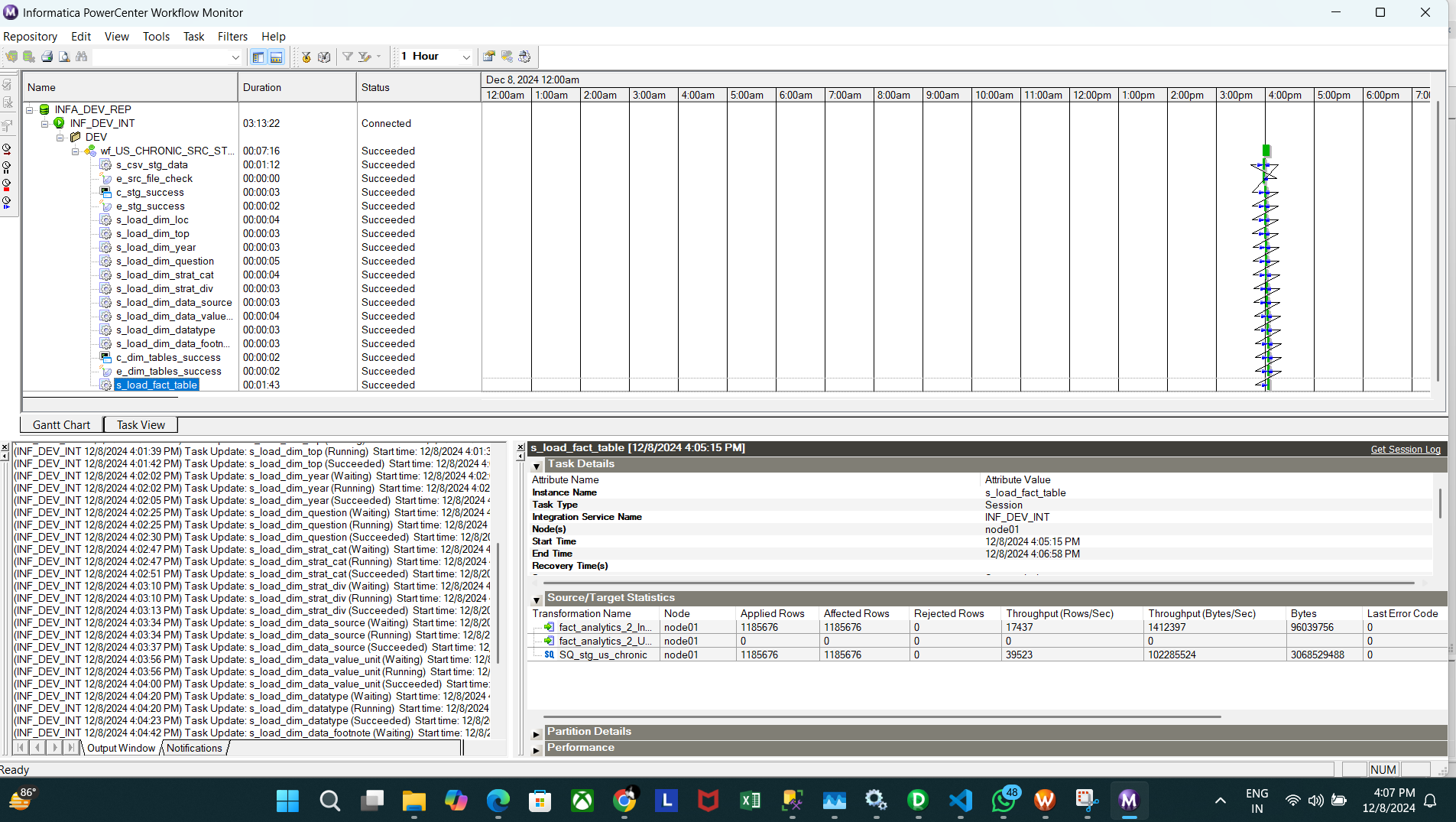
**FACT TABLE**

****

**FINAL WORKFLOW MANAGER DIAGRAM**

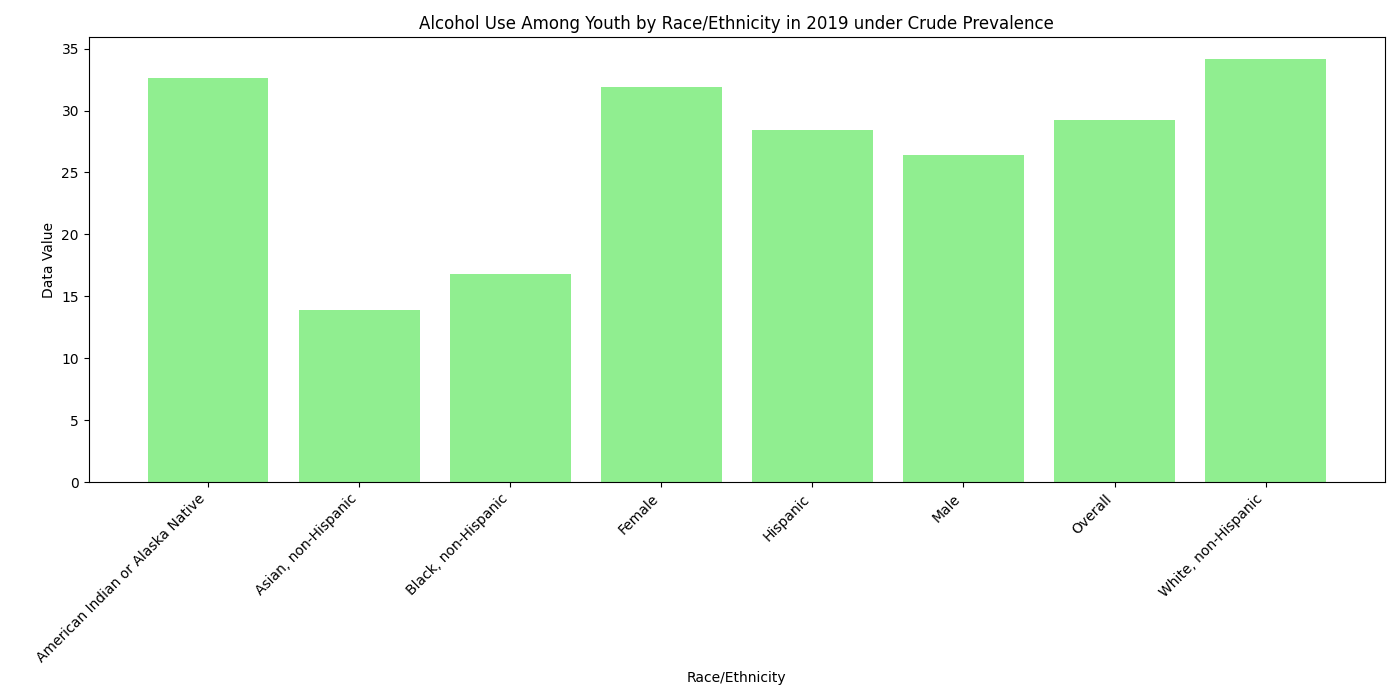
****

**FINAL WORKFLOW MONITOR OUTPUT**

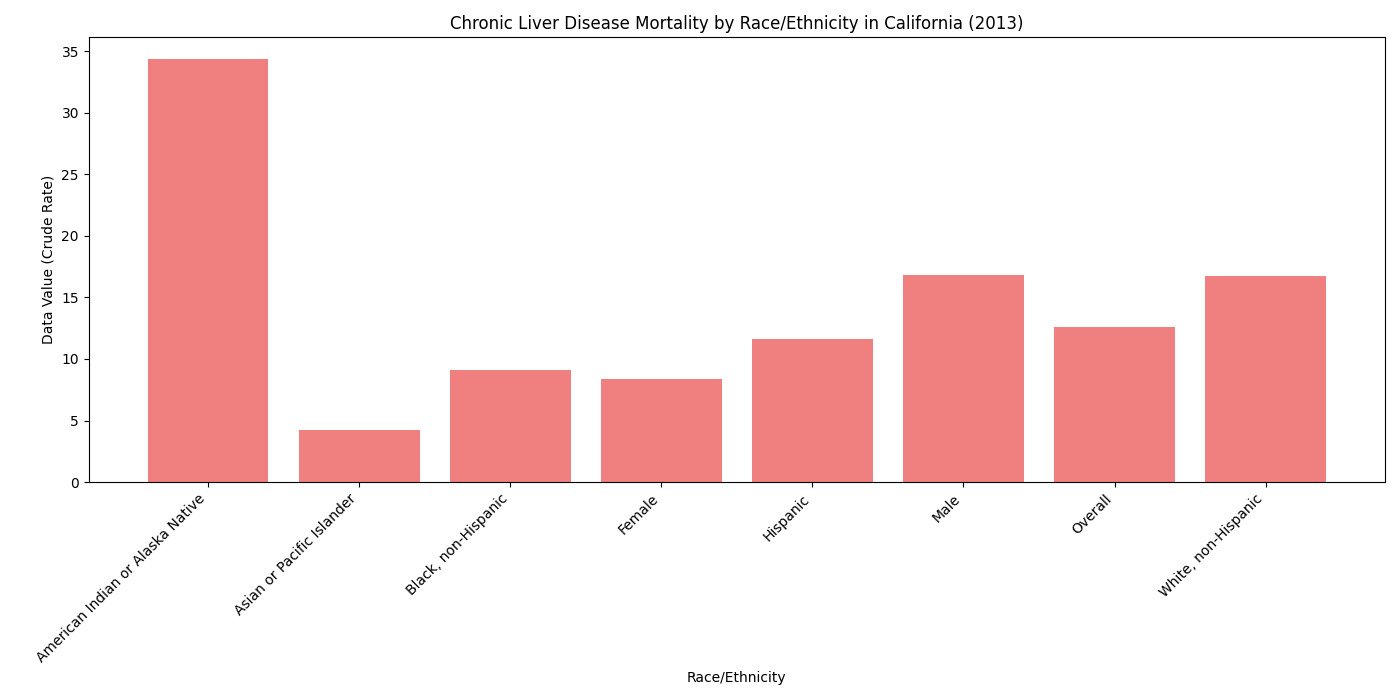


**VISUALIZATION**

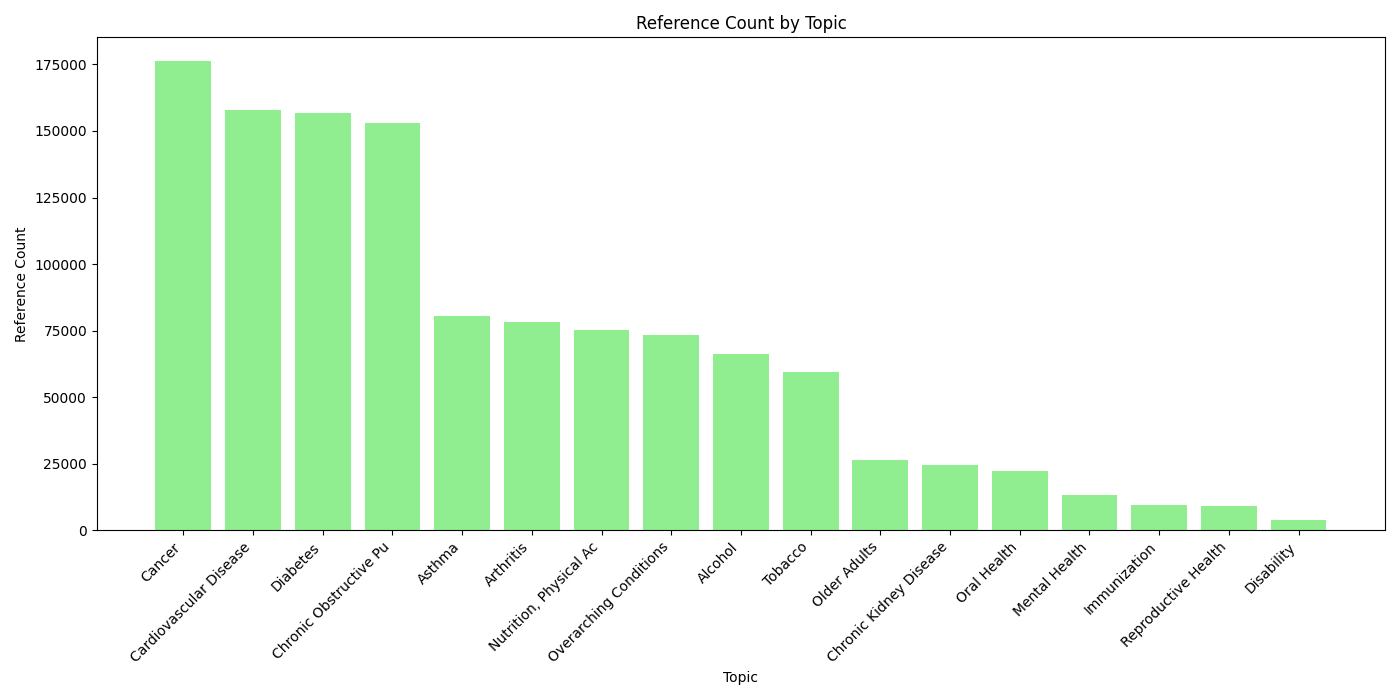
**Alcohol Use Among Youth by Race Ethnicity in 2019 under Crude Prevalence**



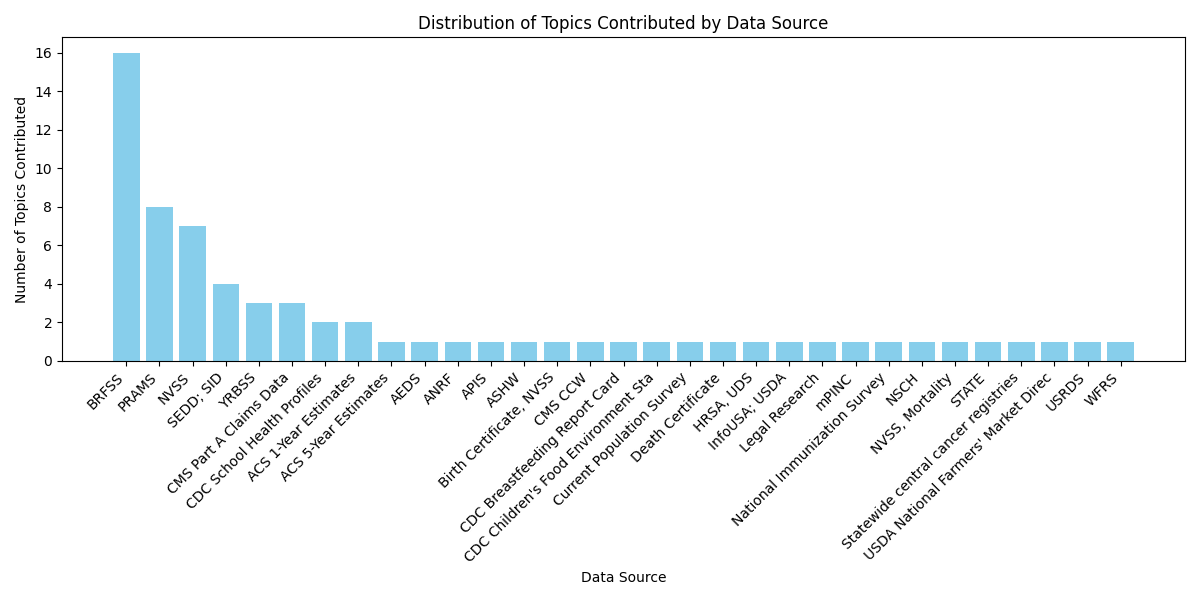
**Chronic liver disease mortality in California in 2013**



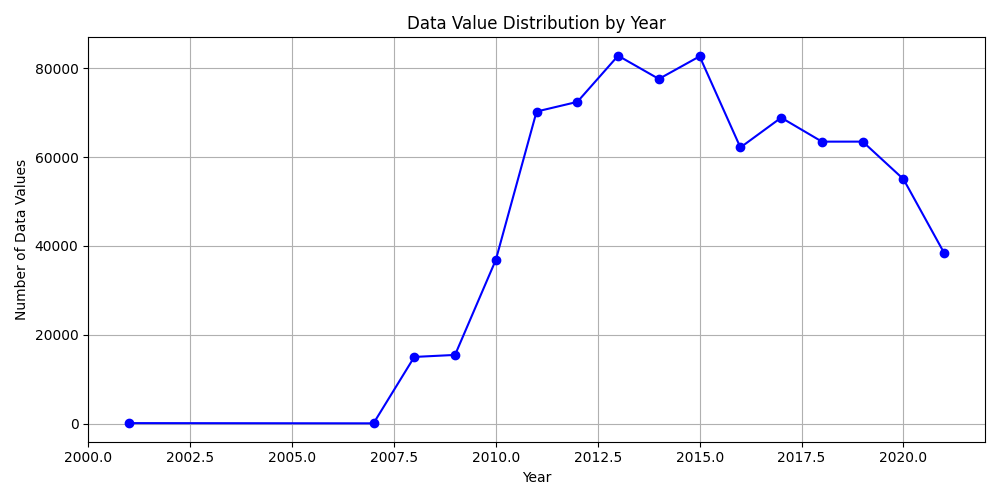
**Reference Count**



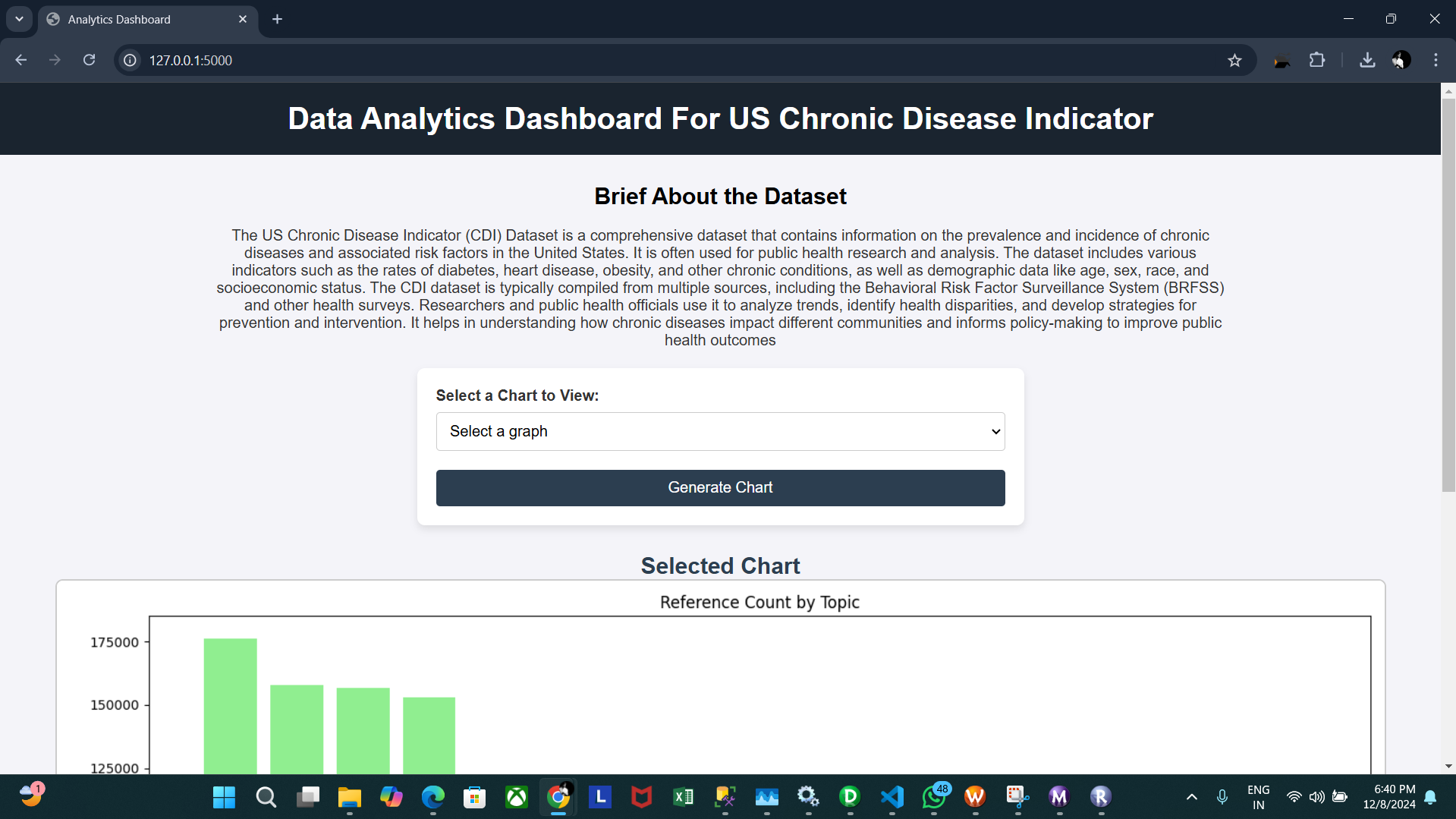
**Topics Distribution**



**Year Distribution**



**WEBSITE INTERFACE DEVELOPED FOR DATA CHARTS VISUALIZATION**

****

**CONCLUSION**

The successful completion of the project on Data Warehouse Design and ETL Implementation Using Informatica demonstrates the practical application of data warehousing concepts and ETL processes in addressing real-world data challenges. By transforming the raw dataset of US Chronic Disease Indicators into a structured data warehouse, actionable insights were derived to support decision-making and analysis.

The project achieved its objectives through:

1. Comprehensive Data Modeling: Designing dimension and fact tables with a clear understanding of the dataset's structure and requirements.
2. Efficient ETL Implementation: Leveraging Informatica to perform seamless data extraction, transformation, and loading, while adhering to best practices like parameterization and workflow dependencies.
3. Accurate Data Analysis: Writing and executing SQL queries to validate the data and extract meaningful insights, highlighting the significance of the data warehouse in real-time analytics.