EXTRACCIÓN DE CONOCIMIENTOS EN BASES DE DATOS

ING. LUIS ENRIQUE MASCOTE CANO.

Universidad Tecnológica de Chihuahua

REPORTE DE SOLUCIÓN DE CASO DE ESTUDIO DE TÉCNICAS DE LIMPIEZA DE DATOS

Lic. Ricardo Hernández Martínez

Fecha de Entrega: 12/OCTUBRE/2025

Índice

Introducción

Se analizó el dataset international-migration-March-2021-citizenship-by-visa-by-country-of-last-permanent-residence.csv (con datos mensuales de migración por ciudadanía, tipo de visa, país de última residencia).

- Realizar limpieza de datos (identificar y tratar faltantes, inconsistencias, duplicados).
- Definir tablas de hechos y dimensiones para un DW.
- Diseñar un modelo relacional normalizado (≥3FN) y generar script SQL/diagrama.

Resumen técnico rápido (valores clave detectados):

- Filas originales: 401,772
- Filas tras eliminar duplicados exactos: 401,772 (no se eliminaron filas porque no había duplicados idénticos)
- Filas marcadas como agregadas (p. ej. visa = "TOTAL" u otros): 125,515

Análisis

Formato inconstante:

- year_month y month_of_release usan formato YYYY-MM (bueno). Confirmar que siempre respetan ese formato; normalizar a tipo date (representar como primer día del mes: YYYY-MM-01).
- visa y passenger_type pueden contener valores como TOTAL o Unknown — hay que tratarlos como categorías específicas o excluírlos según análisis.
- Técnica aplicada: Marcar las columnas como is_aggregate = True para que el modelado las trate de forma explícita (suelen ser útiles para comprobaciones o agregaciones, pero hay que distinguirlas de las filas detalladas). Recomendación: mantenerlas en un dataset raw pero excluirlas por defecto del fact table si se desea sólo granularidad por visa/pais.
- Observación: No se encontraron duplicados exactos en todas las columnas (duplicados contados sobre todas las columnas del dataset = 0).
- Técnica aplicada: drop_duplicates(subset=all original columns) — no cambió filas porque no había duplicados exactos.

2) Tablas sugeridas (fact & dims) y propósito

Propuesta general: esquema estrella con hecho principal fact_migration_monthly y dimensiones de soporte (fecha, país, visa, passenger type, direction, status, release date si se quiere separar). Además, mantenemos un staging limpio.

Dimensiones (cada una en 3FN si necesario):

- dim_date
    - Propósito: estandarizar year_month (periodo observado). Permite análisis por año/mes/trimestre etc.
    - Campos clave: date_id (surrogate), year, month, month_start_date (DATE), year_month (VARCHAR or ym code).

- dim_release_date (opcional)
    - Propósito: fecha de publicación/difusión del dato (month_of_release). Puede usarse para versionado (provisional/final).
    - Campos: release_date_id, release_year, release_month, release_month_start.
- dim_passenger_type
    - Propósito: categoría del pasajero (Student, Work, TOTAL, etc.)
    - Campos: passenger_type_id, passenger_type_code (business key), description.
- dim_direction
    - Propósito: Inbound / Outbound u otras direcciones.
    - Campos: direction_id, direction_code.
- dim_country
    - Propósito: normalizar países (tanto citizenship como country_of_residence). Incluir ISO codes. Evita duplicidad de nombres.
    - Campos: country_id, country_name, iso_alpha2, iso_alpha3, notes. (Usar la misma tabla para citizenship y residence.)
- dim_visa
    - Propósito: tipo de visa (Work, Student, etc.). TOTAL puede tratarse como categoría.
    - Campos: visa_id, visa_code, description.
- dim_status
    - Propósito: Provisional, Final, etc.
    - Campos: status_id, status_code.

Hecho:

- fact_migration_monthly
    - Propósito: almacenar la medida por combinación de dimensiones en un mes.
    - Campos: fact_id (surrogate), date_id (FK -> dim_date), release_date_id (FK), passenger_type_id, direction_id, citizenship_country_id, residence_country_id, visa_id, status_id, estimate (BIGINT), standard_error (FLOAT), row_count (optional audit), load_timestamp, source_filename, source_row_hash (para auditoría/dedup).

Conclusiones

- El dataset está mayormente consistente: no hay nulos en columnas clave (year_month, estimate) y no hay duplicados exactos.
- Hay una proporción relevante de filas de agregado (visa = TOTAL u otros) que hay que tratar con cuidado (marcadas con is_aggregate).
- Es importante estandarizar las cadenas (mayúsculas/minúsculas y nombres de países) antes de cargar al DW para evitar dimension explosion (p. ej. "NZ" vs "New Zealand" vs "nz").

1. Normalización de países: mapear country_of_residence y citizenship a códigos ISO (ISO-3166) y usar éstos en la dimensión. Evitar nombres libres que generen duplicación.
2. Reglas para filas agregadas: decidir si quieres:
    - conservar los registros TOTAL en el DW (marcados) y excluirlos en análisis por defecto, o
    - eliminarlos del fact y mantenerlos sólo en un dataset de comprobación.
3. Procesamiento ETL sugerido:
    - Cargar raw → limpieza (trim, tipo, imputaciones si aparecen nulos en futuras actualizaciones) → poblar dimensiones (con INSERT IGNORE/MERGE) → poblar fact con FK resolvidas.
4. Verificación de month_of_release: si para auditoría quieres la fecha de publicación, mantener month_of_release como atributo en una tabla de metadatos o añadirla a dim_date como release_month_label.
5. Calidad y trazabilidad: conservar el dataset original (sin modificar) y almacenar logs del proceso ETL (registro de filas transformadas/eliminadas) para auditoría.

Referencias

Stats NZ. (2021, March). *International migration: March 2021 – Citizenship by visa by country of last permanent residence (CSV file)*. Statistics New Zealand. https://www.stats.govt.nz/

Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed.). Wiley.

Coronel, C., & Morris, S. (2019). *Database Systems: Design, Implementation, & Management* (13th ed.). Cengage Learning.