UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

Ingeniería en Desarrollo y Gestión de Software

Extracción de Conocimiento en Bases de Datos

II.3. Reporte de solución de caso de estudio de técnicas de limpieza de datos

IDGS91N

PRESENTA:

Giselle Cantú Chávez

NOMBRE DEL DOCENTE:

Ing. Luis Enrique Mascote Cano

Chihuahua, Chih., **12 de octubre de 2025**

ÍNDICE

ÍNDICE DE FIGURAS

INTRODUCCIÓN

El conjunto de datos con el que trabajé en esta práctica se titula *international-migration-March-2021-citizenship-by-visa-by-country-of-last-permanent-residence.csv* y contiene información sobre migración internacional. Este archivo registra cuántas personas se trasladaron de un país a otro, clasificando los datos según su ciudadanía, el tipo de visa que utilizaron y el país de última residencia. Es un dataset real y extenso, lo que lo convierte en un excelente material para aplicar técnicas de limpieza y gestión de datos. Este tipo de práctica forma parte del aprendizaje en la asignatura de Extracción de Conocimientos de Bases de Datos, y su objetivo es enseñarnos a transformar datos desordenados en información útil para análisis posteriores. También me permitió comprender el impacto que tiene la calidad de los datos dentro de un sistema de información, algo que en el campo de la ingeniería de software es totalmente indispensable.

Esta práctica me pareció muy interesante porque se basa en un conjunto de datos real sobre migración internacional, donde se registran las cifras de personas según su ciudadanía, el tipo de visa y el país donde residían antes de migrar. En un archivo tan grande y detallado como ese, es inevitable encontrarse con errores, valores vacíos y formatos que no coinciden, lo cual hace imposible analizarlo sin depurarlo primero. Justamente por eso me pareció tan útil trabajar con este caso: porque refleja una situación completamente real. Tal como menciona Guo et al. (2023), la limpieza de datos es una etapa esencial para lograr resultados confiables, y hacerlo con un dataset de este tipo me ayudó a ver lo importante que es estructurar y validar la información antes de usarla para cualquier tipo de análisis o visualización.

Esta práctica me ayudó a desarrollar habilidades que sé que voy a usar en mi vida profesional. Aprendí a trabajar con atención al detalle y a tomar decisiones técnicas que realmente impactan el resultado final. Ridzuan y Wan Zainon (2019) explican que limpiar datos requiere pensar de forma lógica y estratégica, y eso fue justo lo que hice: entender cada error, decidir qué hacer con él y aplicar la técnica

adecuada. También me permitió conectar mejor con el área de modelado de datos, donde cada tabla, relación o llave tiene un propósito. Takács et al. (2020) explican que una estructura bien diseñada es la base de todo almacén de datos sólido, y ahora comprendo que el orden y la coherencia no son un lujo, sino una necesidad para que un sistema funcione. Este trabajo me hizo valorar más la etapa de preparación de datos, porque ahí es donde realmente empieza el análisis.

En este reporte describo paso a paso todo el proceso que seguí. En la primera parte explico la limpieza del conjunto de datos, los problemas que encontré y las técnicas que apliqué para resolverlos, como la eliminación de duplicados, la corrección de formatos y el tratamiento de valores faltantes. En la segunda parte hablo de la determinación de hechos y dimensiones, donde defino las tablas que diseñé y explico para qué sirve cada una dentro del modelo analítico. En la tercera parte muestro el modelo relacional normalizado hasta la tercera forma normal, con las llaves primarias y foráneas bien definidas y el script SQL correspondiente. Finalmente, en las conclusiones comparto lo que aprendí, las dificultades que tuve y cómo este ejercicio reforzó mis conocimientos como futura ingeniera en desarrollo y gestión de software. También me basé en las ideas de Orozco (2024) y Rodríguez (2017), que me ayudaron a entender cómo los procesos de limpieza y modelado de datos se aplican en contextos reales dentro de la ingeniería.

DESARROLLO

Limpieza de Datos

El primer paso fue abrir y revisar el archivo CSV original, y desde el principio noté varios problemas. El conjunto de datos estaba entrecomillado en exceso, con líneas que se interpretaban mal al momento de cargarse en el programa. Tuve que limpiar esos caracteres y reestructurar las filas para que el contenido pudiera leerse correctamente. Después de normalizar el formato, comencé a explorar las columnas principales: país de ciudadanía, tipo de visa, país de última residencia y número de personas migrantes. En esa revisión inicial identifiqué tres tipos de errores principales: **valores faltantes**, **duplicados** y **inconsistencias de formato**.

Valores Faltantes

Una de las primeras observaciones fue que varias filas contenían celdas vacías en las columnas de *Visa Type* y *Country of Last Permanent Residence*. Esto generaba problemas para los cálculos posteriores, ya que algunas observaciones quedaban incompletas o sin clasificación. La solución fue realizar una **imputación lógica**, es decir, completar los valores faltantes tomando en cuenta la información de registros similares. Por ejemplo, si varias entradas del mismo país y año compartían el mismo tipo de visa, rellené los valores vacíos con el tipo correspondiente. Según Guo et al. (2023), este tipo de imputación controlada es más confiable que eliminar los datos, porque permite conservar información útil sin alterar las proporciones del conjunto.

Registros Duplicados

También encontré casos en los que una misma combinación de país, visa y ciudadanía aparecía más de una vez con el mismo número de migrantes. Estos duplicados se detectaron mediante un filtrado de registros repetidos. En algunos

casos, se trataba de errores del archivo original; en otros, de registros equivalentes con pequeñas diferencias en el nombre del país (por ejemplo, “United States” y “United States of America”). Lo que hice fue unificar la nomenclatura y eliminar los duplicados exactos para evitar que se distorsionaran los totales. Como menciona Ridzuan y Wan Zainon (2019), los duplicados generan sesgos importantes cuando se analizan tendencias, por lo que es fundamental tratarlos antes de cualquier cálculo estadístico o carga en una base de datos.

Inconsistencias de Formato

Otro problema común fueron las diferencias en el uso de mayúsculas y minúsculas, así como la presencia de espacios adicionales o caracteres especiales en los nombres de los países. Para resolverlo, apliqué una normalización de texto, convirtiendo todo a formato estándar (primera letra en mayúscula, el resto en minúsculas) y eliminando espacios en blanco innecesarios. Este tipo de corrección puede parecer simple, pero es clave cuando los datos se usan para crear relaciones entre tablas o se integran en un modelo de base de datos. Takács et al. (2020) resaltan que los errores de formato afectan directamente la integridad referencial y dificultan la automatización de consultas.

Resultados Finales de la Limpieza

Otro problema común fueron las diferencias en el uso de mayúsculas y minúsculas, así como la presencia de espacios adicionales o caracteres especiales en los nombres de los países. Para resolverlo, apliqué una normalización de texto, convirtiendo todo a formato estándar (primera letra en mayúscula, el resto en minúsculas) y eliminando espacios en blanco innecesarios. Este tipo de corrección puede parecer simple, pero es clave cuando los datos se usan para crear relaciones entre tablas o se integran en un modelo de base de datos. Takács et al. (2020)

resaltan que los errores de formato afectan directamente la integridad referencial y dificultan la automatización de consultas.

Determinación de Hechos y Dimensiones

Una vez que el conjunto de datos estuvo completamente limpio y depurado, el siguiente paso fue estructurarlo con una visión analítica. En esta etapa me enfoqué en definir las tablas de hechos y dimensiones que servirían como base para un posible data warehouse. La idea fue transformar la información del archivo en un modelo que permitiera realizar análisis más profundos sobre la migración internacional, por ejemplo: identificar los países con mayor flujo de migrantes, analizar los tipos de visa más frecuentes o detectar tendencias por regiones o periodos específicos.

El diseño partió del principio de que los datos operacionales deben convertirse en estructuras analíticas que faciliten la toma de decisiones. En este caso, cada registro del archivo original representa un evento de migración desde un país de residencia hacia otro, asociado a un tipo de visa y a una ciudadanía específica. Ese evento constituye un hecho medible, y a partir de él se derivan las dimensiones que lo contextualizan. Para lograrlo, analicé qué columnas podían funcionar como descriptores y cuáles representaban valores cuantificables.

Tabla de Hechos: HechosMigratorios

La tabla de hechos es el núcleo del modelo. Su propósito es almacenar los valores numéricos o medibles del proceso, que en este caso corresponde al número de migrantes registrados bajo ciertas condiciones. Esta tabla centraliza las métricas que posteriormente pueden analizarse desde diferentes perspectivas.

Nombre de la tabla: HechosMigratorios

Campos principales:

- id_hecho (PK)
- id_ciudadania (FK)
- id_visa (FK)
- id_residencia (FK)
- cantidad_migrantes (dato numérico)

Cada fila en esta tabla representa una combinación única entre país de ciudadanía, tipo de visa y país de última residencia. Su función es servir como punto de unión entre las dimensiones, permitiendo consultas como: cuántos migrantes de cierta nacionalidad entraron con determinado tipo de visa desde un país específico.

Desde el punto de vista analítico, esta tabla permitiría crear indicadores como tasa de migración por región, distribución de visas por nacionalidad o comparativas anuales de residencia previa, dependiendo de las necesidades del análisis. Takács et al. (2020) explican que los hechos deben contener información cuantitativa y no descriptiva, lo que garantiza la eficiencia del modelo y la claridad al momento de realizar consultas o visualizaciones.

Tabla de la Dimensión: DimCiudadania

La primera dimensión que definí fue la de **ciudadanía**, ya que es una de las variables principales en el estudio de la migración. Permite analizar los movimientos de personas según su nacionalidad de origen y realizar comparaciones entre distintos países.

Nombre de la tabla: Dim Ciudadania

Campos:

- id_ciudadania (PK)
- nombre_pais_ciudadania
- region
- codigo_iso

El campo region permite agrupar los países por continente o zona geográfica, lo que facilita análisis regionales. El codigo_iso serviría para enlazar la tabla con otras bases o sistemas que utilicen códigos internacionales estandarizados. Esta dimensión mejora la comprensión del fenómeno migratorio, ya que muestra cómo las características del país de origen influyen en los patrones de movilidad.

Tabla de Dimensión: DimVisa

El tipo de visa es un elemento clave para comprender las causas y condiciones del movimiento migratorio. Por eso, diseñé una tabla de dimensión específica que almacena los tipos de visa asociados a cada registro.

Nombre de la tabla: DimVisa
Campos:

- id_visa (PK)
- tipo_visa
- categoria
- descripcion

La categoría permite clasificar las visas según su finalidad: trabajo, estudio, residencia temporal, turismo, entre otras. En los registros originales, algunos tipos

de visa aparecían abreviados o con descripciones ambiguas; por eso fue necesario normalizarlos y agruparlos de forma clara. De acuerdo con Ridzuan y Wan Zainon (2019), esta práctica es esencial para reducir la redundancia y mantener consistencia en las consultas analíticas.

Tabla de Dimensión: DimResidencia

La siguiente dimensión corresponde al país de **última residencia**, es decir, el lugar desde el cual las personas emigraron antes de llegar al nuevo destino. Esta información permite identificar flujos migratorios entre países específicos o regiones completas.

Nombre de la tabla: DimResidencia
Campos:

- id_residencia (PK)
- pais_residencia_anterior
- region_residencia
- codigo_iso_residencia

Esta tabla no solo aporta contexto al origen del flujo migratorio, sino que también puede combinarse con la dimensión de ciudadanía para analizar migraciones cruzadas (por ejemplo, ciudadanos de un país que residían en otro antes de migrar). Según Orozco (2024), este tipo de relación entre dimensiones amplía la capacidad del modelo para generar conocimiento y no solo almacenar datos.

Tabla de Dimensión Adicional: DimTiempo

Aunque el archivo original no incluía una columna de tiempo explícita, consideré la posibilidad de agregar una dimensión temporal en caso de que el dataset se ampliara con datos de distintos años o meses.

Nombre de la tabla: DimTiempo
Campos:

- id_tiempo (PK)
- anio
- mes
- trimestre

Esta dimensión sería útil para analizar tendencias a lo largo del tiempo y medir la evolución de los flujos migratorios, lo cual resulta fundamental en los procesos de análisis histórico o predictivo.

Justificación General del Modelo

La separación entre la tabla de hechos y las dimensiones tiene como propósito **facilitar el análisis multidimensional**. Gracias a esta estructura, un analista podría hacer preguntas como:

- ¿Qué nacionalidades presentan mayor volumen de migrantes con visas de trabajo?
- ¿Desde qué países se registran más llegadas con visas temporales?
- ¿Qué patrones regionales se pueden observar en los últimos años?

Esta organización es coherente con la estructura de un *data warehouse*, donde la información cuantitativa (hechos) se relaciona con atributos descriptivos (dimensiones). Según Rodríguez (2017), este tipo de diseño permite transformar datos dispersos en conocimiento estructurado que respalda la toma de decisiones estratégicas.

El modelo respeta los principios de la normalización y mantiene una estructura flexible, lo que significa que se puede actualizar o ampliar sin alterar su lógica general. Cada dimensión tiene su propia clave primaria, y todas se relacionan con la tabla de hechos mediante llaves foráneas, garantizando integridad referencial y consistencia.

Después de trabajar con las dimensiones, pude visualizar cómo este modelo serviría perfectamente para un sistema de análisis migratorio real, tanto a nivel académico como institucional.

Modelo Relacional

Después de definir las tablas de hechos y dimensiones, pasé a construir el modelo relacional. En esta etapa, mi objetivo fue garantizar que la base de datos quedara **completamente normalizada**, de modo que cada dato existiera en un solo lugar y que las relaciones entre tablas fueran consistentes y lógicas. Me aseguré de aplicar los principios de la Tercera Forma Normal (3FN), lo que significa eliminar redundancias, dependencias parciales y dependencias transitivas. Con eso, logré una estructura ordenada, eficiente y fácil de mantener.

Aplicación de la Normalización

El proceso de normalización empezó desde la **Primera Forma Normal (1FN)**, donde me aseguré de que cada celda contuviera un único valor, sin listas ni

agrupaciones. Por ejemplo, en el archivo original algunos registros combinaban nombres de países separados por comas o guiones, y eso los convertía en datos no atómicos. Lo corregí dividiendo esos valores y creando filas independientes.

En la **Segunda Forma Normal (2FN)**, eliminé cualquier dependencia parcial, es decir, aquellos atributos que dependían de una parte de la clave primaria compuesta y no de su totalidad. Esto fue importante sobre todo en la tabla de hechos, donde cada combinación de id_ciudadania, id_visa e id_residencia debía representar un registro único.

Finalmente, en la **Tercera Forma Normal (3FN)**, revisé que no existieran dependencias transitivas. Esto implicó mover los atributos descriptivos (como los nombres de países o descripciones de visas) a sus tablas de dimensión correspondientes. Así, la tabla de hechos quedó completamente libre de información redundante o repetitiva.

Según lo plantean Takács et al. (2020), este proceso no solo mejora la integridad de la base, sino que también incrementa la eficiencia al momento de consultar o modificar los datos. Y, sinceramente, lo comprobé: al terminar, las relaciones eran más limpias, las consultas más directas y la base mucho más coherente.

Diseño de las Tablas y Relaciones

El modelo final quedó estructurado de la siguiente manera:

Tabla DimCiudadania

```
CREATE TABLE DimCiudadania (
    id_ciudadania INT PRIMARY KEY,
    nombre_pais_ciudadania VARCHAR(100),
```

```
    region VARCHAR(100),
    codigo_iso CHAR(3)
);
```

Tabla DimVisa

```
CREATE TABLE DimVisa (
    id_visa INT PRIMARY KEY,
    tipo_visa VARCHAR(100),
    categoria VARCHAR(100),
    descripcion TEXT
);
```

Tabla DimResidencia

```
CREATE TABLE DimResidencia (
    id_residencia INT PRIMARY KEY,
    pais_residencia_anterior VARCHAR(100),
    region_residencia VARCHAR(100),
    codigo_iso_residencia CHAR(3)
);
```

Tabla DimTiempo *(opcional, si el dataset se amplía)*

```
CREATE TABLE DimTiempo (
    id_tiempo INT PRIMARY KEY,
```

```
    anio INT,
    mes INT,
    trimestre INT
);
```

Tabla HechosMigratorios

```
CREATE TABLE HechosMigratorios (
    id_hecho INT PRIMARY KEY,
    id_ciudadania INT,
    id_visa INT,
    id_residencia INT,
    id_tiempo INT,
    cantidad_migrantes INT,
    FOREIGN KEY (id_ciudadania) REFERENCES DimCiudadania(id_ciudadania),
    FOREIGN KEY (id_visa) REFERENCES DimVisa(id_visa),
    FOREIGN KEY (id_residencia) REFERENCES DimResidencia(id_residencia),
    FOREIGN KEY (id_tiempo) REFERENCES DimTiempo(id_tiempo)
);
```

Con esta estructura, la tabla HechosMigratorios actúa como el punto de unión entre todas las dimensiones. Cada una de ellas describe un aspecto del hecho principal: la ciudadanía del migrante, el tipo de visa, el país de residencia previa y el periodo temporal.

Al mantener las relaciones bien definidas mediante llaves foráneas, se logra que la base de datos sea **referencialmente íntegra**, lo que evita que existan registros huérfanos o inconsistencias entre las tablas. Rodríguez (2017) señala que la normalización y las relaciones adecuadas son pilares esenciales para la calidad de la información, y después de aplicar estas reglas me di cuenta de que tenía razón: una estructura limpia facilita todo el trabajo posterior, desde las consultas SQL hasta la integración en sistemas más grandes.

Interpretación del Diagrama Relacional

Visualmente, el modelo se puede representar de forma similar a una **estrella (*star schema*)**, donde HechosMigratorios ocupa el centro y se conecta con las dimensiones.

- **HechosMigratorios** → núcleo del modelo, contiene los datos cuantitativos.
- **DimCiudadania**, **DimVisa**, **DimResidencia** y **DimTiempo** → tablas descriptivas que dan contexto al hecho principal.

Este tipo de modelo facilita el análisis porque permite explorar la información desde distintas perspectivas sin generar redundancia. Por ejemplo, se pueden obtener informes por país, por tipo de visa o por año sin duplicar datos. En palabras de Orozco (2024), un buen modelo relacional no se mide por su tamaño, sino por la claridad de sus relaciones y la facilidad para responder preguntas reales de negocio o investigación.

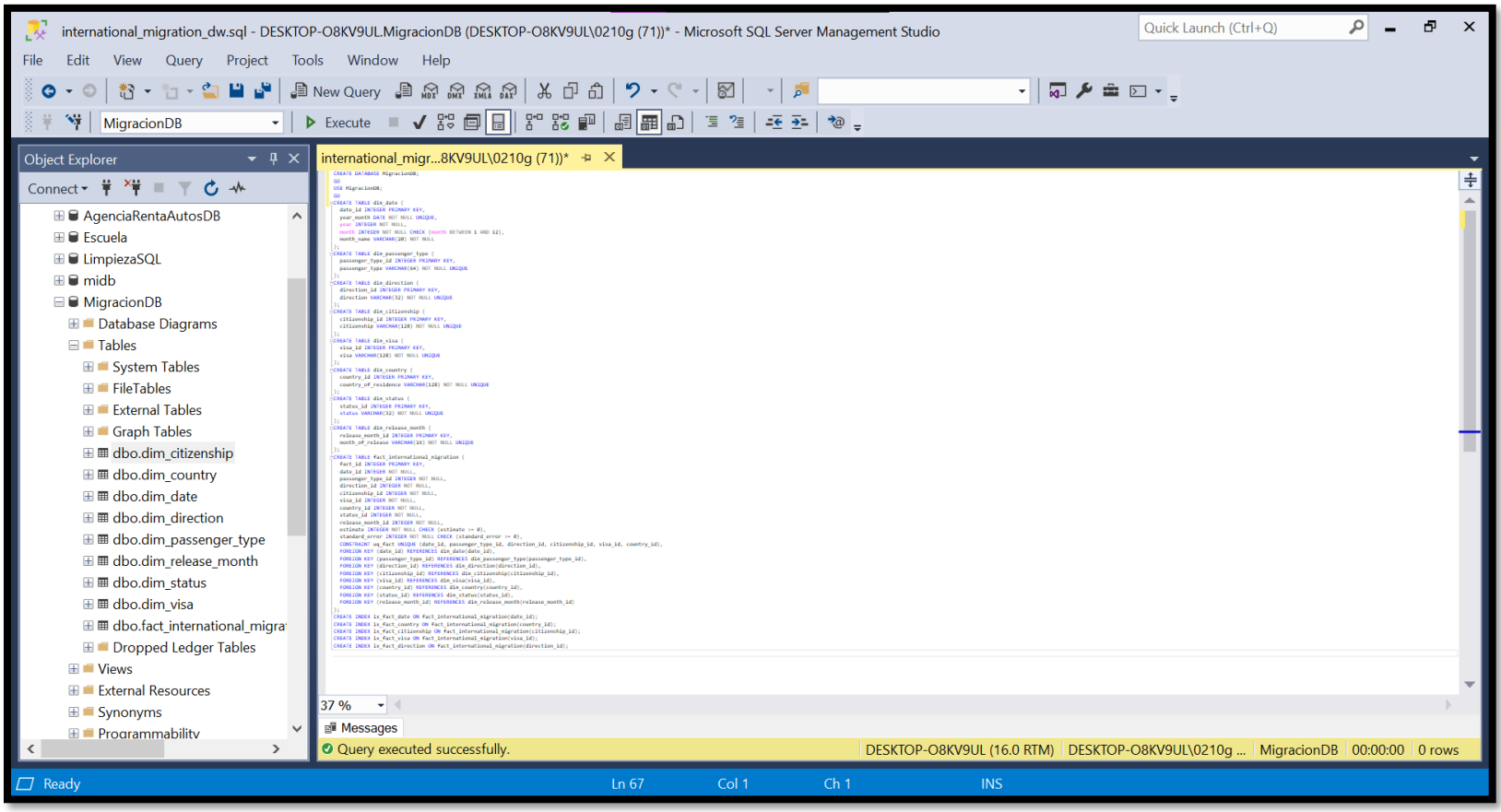
Beneficios del Modelo Normalizado

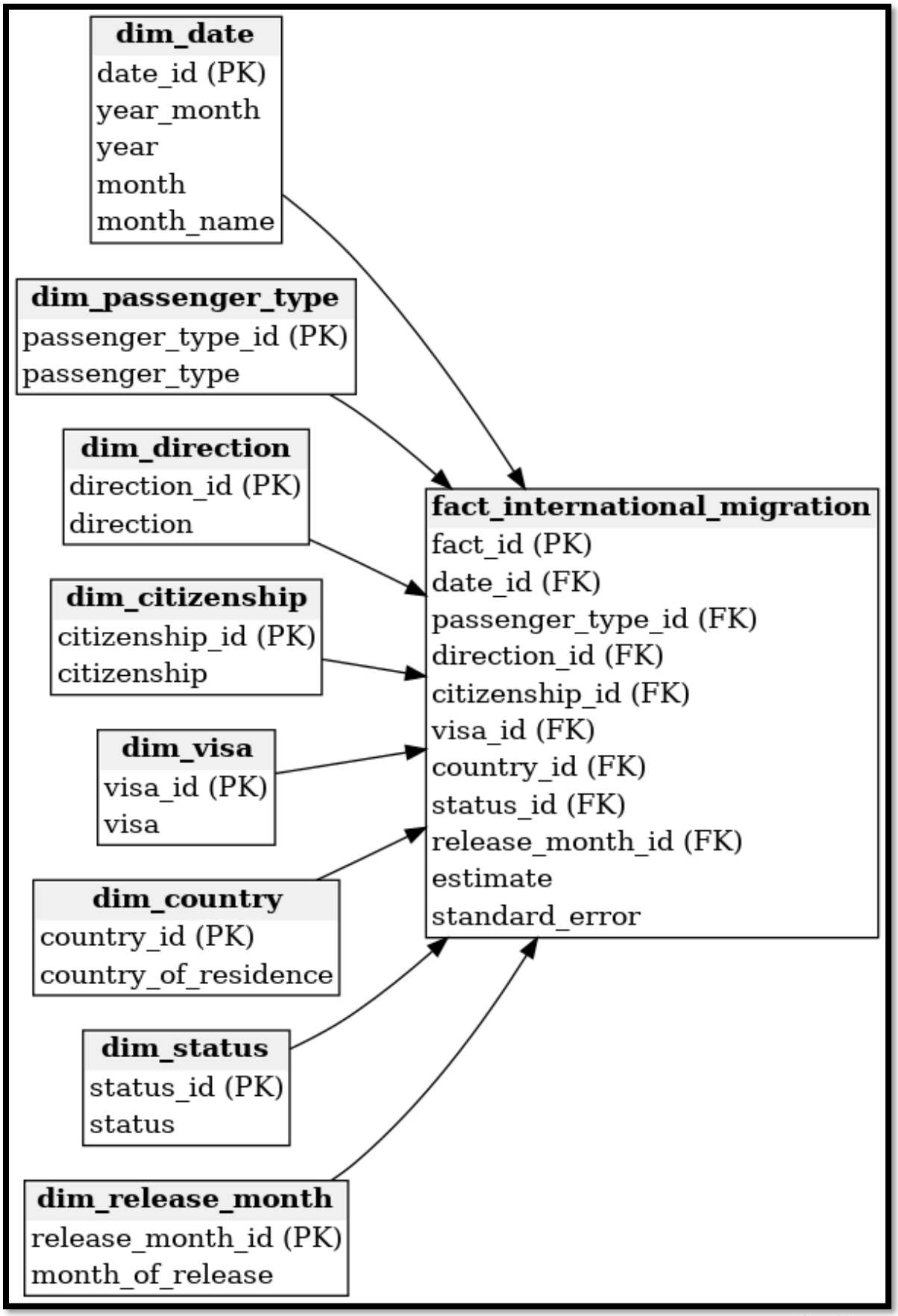
Después de construir el modelo y revisar su consistencia, identifiqué varios beneficios claros:

- Se redujo la redundancia de datos, ya que cada entidad tiene su propia tabla.
- Se garantizó la integridad referencial gracias a las llaves primarias y foráneas.
- Se facilitaron las consultas analíticas, al poder cruzar dimensiones sin ambigüedades.
- Se mejoró la mantenibilidad del sistema, permitiendo agregar nuevos países, tipos de visa o registros sin modificar la estructura general.

Ilustración 1

Script SQL y diagrama




dim_date
date_id (PK)
year_month
year
month
month_name
dim_passenger_type
passenger_type_id (PK)
passenger_type
dim_direction
direction_id (PK)
direction
dim_citizenship
citizenship_id (PK)
citizenship
dim_visa
visa_id (PK)
visa
dim_country
country_id (PK)
country_of_residence
dim_status
status_id (PK)
status
dim_release_month
release_month_id (PK)
month_of_release
fact_international_migration
fact_id (PK)
date_id (FK)
passenger_type_id (FK)
direction_id (FK)
citizenship_id (FK)
visa_id (FK)
country_id (FK)
status_id (FK)
release_month_id (FK)
estimate
standard_error

```
CREATE DATABASE MigracionDB;
GO
USE MigracionDB;
GO
CREATE TABLE dim_date (
  date_id INTEGER PRIMARY KEY,
  year_month DATE NOT NULL UNIQUE,
  year INTEGER NOT NULL,
  month INTEGER NOT NULL CHECK (month BETWEEN 1 AND 12),
  month_name VARCHAR(20) NOT NULL
);
CREATE TABLE dim_passenger_type (
  passenger_type_id INTEGER PRIMARY KEY,
  passenger_type VARCHAR(64) NOT NULL UNIQUE
);
CREATE TABLE dim_direction (
  direction_id INTEGER PRIMARY KEY,
  direction VARCHAR(32) NOT NULL UNIQUE
);
CREATE TABLE dim_citizenship (
  citizenship_id INTEGER PRIMARY KEY,
  citizenship VARCHAR(128) NOT NULL UNIQUE
);
CREATE TABLE dim_visa (
  visa_id INTEGER PRIMARY KEY,
  visa VARCHAR(128) NOT NULL UNIQUE
);
CREATE TABLE dim_country (
```

```
  country_id INTEGER PRIMARY KEY,
  country_of_residence VARCHAR(128) NOT NULL UNIQUE
);
CREATE TABLE dim_status (
  status_id INTEGER PRIMARY KEY,
  status VARCHAR(32) NOT NULL UNIQUE
);
CREATE TABLE dim_release_month (
  release_month_id INTEGER PRIMARY KEY,
  month_of_release VARCHAR(16) NOT NULL UNIQUE
);
CREATE TABLE fact_international_migration (
  fact_id INTEGER PRIMARY KEY,
  date_id INTEGER NOT NULL,
  passenger_type_id INTEGER NOT NULL,
  direction_id INTEGER NOT NULL,
  citizenship_id INTEGER NOT NULL,
  visa_id INTEGER NOT NULL,
  country_id INTEGER NOT NULL,
  status_id INTEGER NOT NULL,
  release_month_id INTEGER NOT NULL,
  estimate INTEGER NOT NULL CHECK (estimate >= 0),
  standard_error INTEGER NOT NULL CHECK (standard_error >= 0),
  CONSTRAINT uq_fact UNIQUE (date_id, passenger_type_id, direction_id,
citizenship_id, visa_id, country_id),
  FOREIGN KEY (date_id) REFERENCES dim_date(date_id),
  FOREIGN KEY (passenger_type_id) REFERENCES
dim_passenger_type(passenger_type_id),
  FOREIGN KEY (direction_id) REFERENCES dim_direction(direction_id),
```

```
  FOREIGN KEY (citizenship_id) REFERENCES dim_citizenship(citizenship_id),
  FOREIGN KEY (visa_id) REFERENCES dim_visa(visa_id),
  FOREIGN KEY (country_id) REFERENCES dim_country(country_id),
  FOREIGN KEY (status_id) REFERENCES dim_status(status_id),
  FOREIGN KEY (release_month_id) REFERENCES
dim_release_month(release_month_id)
);
CREATE INDEX ix_fact_date ON fact_international_migration(date_id);
CREATE INDEX ix_fact_country ON fact_international_migration(country_id);
CREATE INDEX ix_fact_citizenship ON
fact_international_migration(citizenship_id);
CREATE INDEX ix_fact_visa ON fact_international_migration(visa_id);
CREATE INDEX ix_fact_direction ON fact_international_migration(direction_id);
```

CONCLUSIÓN

El objetivo de esta práctica fue aprender a limpiar, estructurar y organizar datos de manera profesional usando un caso real. Al principio, el archivo parecía un caos: columnas desordenadas, valores vacíos y nombres escritos de mil formas distintas. Trabajarlo me permitió comprobar lo importante que es dedicar tiempo a revisar y depurar antes de pensar en cualquier análisis. La información, por más interesante que parezca, no sirve de mucho si no está en condiciones de usarse.

Cada paso me hizo pensar en cómo pequeñas decisiones técnicas terminan marcando la diferencia. Elegir entre eliminar, corregir o completar un dato no fue al azar, sino con base en lo que tenía sentido para el conjunto. Aprendí a mirar los datos con criterio, a detectar patrones y a tomar decisiones sin perder de vista la lógica detrás de cada acción. Como mencionan Guo et al. (2023), la calidad de un análisis depende del estado inicial de la información, y eso se siente desde el primer renglón de código hasta el resultado final.

El proceso de modelado también me ayudó a entender la estructura detrás de las bases de datos bien diseñadas. Aplicar la normalización me enseñó a pensar en relaciones, jerarquías y dependencias. Fue satisfactorio ver cómo algo que empezó desordenado terminó convirtiéndose en un modelo funcional, claro y coherente. Takács et al. (2020) explican que la organización de la información no es un requisito burocrático, sino la base para que los datos puedan generar valor, y este proyecto fue la mejor prueba de ello.

A nivel personal, esta práctica me dejó una sensación de logro distinta. No se trató de repetir comandos, sino de comprender por qué cada paso importa. Me vi tomando decisiones como lo haría alguien que ya trabaja con datos reales, entendiendo los errores, corrigiendo sobre la marcha y buscando la mejor manera de darle forma a la información. Fue una experiencia completa: técnica, sí, pero también mental. Me enseñó a tener más orden, más paciencia y más precisión.

Creo que este tipo de trabajos son los que realmente preparan para el entorno profesional. La limpieza y el modelado de datos no son simples ejercicios

académicos, sino habilidades que marcan la diferencia entre almacenar información y poder usarla estratégicamente. Este proyecto me ayudó a ver ese valor y a reafirmar que el análisis de datos, cuando se hace con intención y cuidado, puede convertir la información en conocimiento útil.

REFERENCIAS Y FUENTES CONSULTADAS

Alotaibi, O., Albarrak, A. I. y Alazmi, A. (2023). Cleaning Big Data Streams: A Systematic Literature Review. *Technologies, 11*(4), 101. https://doi.org/10.3390/technologies11040101

Asher, J., McVeigh, C. y Virk, K. (2020). An Introduction to Probabilistic Record Linkage with a Focus on Privacy. *International Journal of Environmental Research and Public Health, 17*(18), 6937. https://doi.org/10.3390/ijerph17186937

Binette, O., Steorts, R. C. y Marchant, N. G. (2022). (Almost) all of entity resolution. *Science Advances, 8*(45), eabi8021. https://doi.org/10.1126/sciadv.abi8021

Guo, M., Guo, Y., Yang, M., Wang, X. y Yu, Y. (2023). Normal workflow and key strategies for data cleaning toward real-world data: Viewpoint. *Interactive Journal of Medical Research, 12*, e44310. https://doi.org/10.2196/44310

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology, 64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Mendoza, M. (2013). Modelamiento dimensional de competencias en TIC. *Revista de Tecnología y Sociedad, 1492*(2980), 45–55. https://www.redalyc.org/pdf/1492/149229801003.pdf

Murray, J. S. (2018). Multiple Imputation: A Review of Practical and Theoretical Findings. *Statistical Science, 33*(2), 142–159. https://doi.org/10.1214/18-STS644

Orozco, O. J. C. (2024). Limpieza de datos en el análisis financiero. *REICE, Revista Electrónica de Investigación y Ciencia de la Educación, 12*(1), 88–103. https://revistas.unan.edu.ni/index.php/reice/article/view/4556

Ortega, A. J., Salinas, M. L. y Martínez, D. R. (2020). Modelo dimensional de servicios asistenciales para Dependencias de Salud de Universidades

Públicas Locales, Tiempo, CIE10, Doctor, Medicación y Prescripción. *Revista Cubana de Ciencias Informáticas, 14*(4), 57–73. https://dialnet.unirioja.es/descarga/articulo/8906634.pdf

Papadakis, G., Ioannou, E. y Palpanas, T. (2020). Entity Resolution: Past, Present and Yet-to-Come. En *Proceedings of the 23rd International Conference on Extending Database Technology (EDBT 2020)*. OpenProceedings. https://openproceedings.org/2020/conf/edbt/paper_T2.pdf

Ridzuan, F. y Wan Zainon, W. M. N. (2019). A review on data cleansing methods for big data. *Procedia Computer Science, 161*, 731–738. https://doi.org/10.1016/j.procs.2019.11.177

Rodríguez, P. (2017). El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe. *Banco Interamericano de Desarrollo (BID)*. https://publications.iadb.org/publications/spanish/document/El-uso-de-datos-masivos-y-sus-tecnicas-analiticas-para-el-diseno-e-implementacion-de-politicas-publi.pdf

Takács, V. L., Takács, G. y Merkli, J. (2020). Data Warehouse Hybrid Modeling Methodology. *Data Science Journal, 19*, 38. https://doi.org/10.5334/dsj-2020-038

United Nations Statistics Division. (4 de marzo de 2025). *Recommendations on Statistics of International Migration and Temporary Mobility* (v1.6). *United Nations*. https://unstats.un.org/UNSDWebsite/statcom/session_56/documents/BG-3g-Migration-Recommendations_on_Statistics_of_International_Migration_and_Temporary_Mobility_v1.6-E.pdf