UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

TECNOLOGÍAS DE LA INFORMACIÓN

Extracción de Conocimiento en Bases de Datos

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

IDGS91N

PRESENTA:

JORGE ALEJANDRO HERNANDEZ CONTRERAS

DOCENTE:

Enrique Mascote

Chihuahua, Chih.,12 Octubre de 2025

Introducción

El archivo que se trabajó tiene información sobre personas migrantes, como su país de ciudadanía, tipo de visa, país de residencia, tipo de pasajero y fechas. El propósito de este trabajo fue limpiar los datos, identificar tablas de hechos y dimensiones para un modelo de datos y crear un diseño relacional normalizado. Todo esto sirve para que la información se pueda analizar correctamente y se pueda usar en un data warehouse sin errores.

Limpieza de datos

Al revisar los datos, se encontró que no había valores nulos en las columnas principales. Aun así, se analizaron otros posibles problemas que afectan la calidad de la información. Estas son las acciones aplicadas:

Eliminación de duplicados

Se buscó si existían filas repetidas completamente. Para eso se usó quitar duplicados en Excel. Las filas idénticas fueron eliminadas para evitar que los conteos se dupliquen.

Normalización de texto

Se detectó que algunas columnas como citizenship, visa, country_of_residence y passenger_type podían tener inconsistencias como:

- Espacios de más
- Mayúsculas y minúsculas mezcladas
- Variaciones en nombres

Se aplicaron funciones como:

- =ESPACIOS()
- =NOMPROPIO()
- =MAYUSC()

Esto se hizo para que todo quedara estandarizado.

Formato de fecha

La columna year_month estaba como texto, así que se separó en:

- Año (year)
- Mes (month_num)

Revisión de datos numéricos

Las columnas estimate y standard_error se convirtieron a número en caso de que se detectaran como texto.

Valores únicos en status

Se revisó que los valores fueran consistentes, corrigiendo posibles variantes como minúsculas o errores.
Todo cambio realizado se registró en una hoja de auditoría con la siguiente información:

- Valor original
- Valor corregido
- Motivo

Hechos y dimensiones

Después de limpiar los datos, se identificaron las tablas necesarias para un modelo tipo data warehouse.

Tabla de hechos

Fact_Migration contiene los valores medibles:

- estimate
- standard_error
- Llaves de todas las dimensiones relacionadas

Tablas de dimensiones

Dim_Date

- Año, mes, nombre del mes, trimestre.

Dim_PassengerType

- Tipo de pasajero (ej. Long-term migrant).

Dim_Direction

- Arrivals o Departures.

Dim_Citizenship

- País de ciudadanía.

Dim_Visa

- Tipo de visa.

Dim_CountryResidence

- País de residencia anterior.

Dim_Status

- Estado del dato (Provisional o Final).

Modelo relacional

```sql
CREATE TABLE dim_date(
    date_key DATE PRIMARY KEY,
    year INTEGER NOT NULL,
    month INTEGER NOT NULL,
    month_name VARCHAR(20),
    quarter INTEGER
);

CREATE TABLE dim_passenger_type(
    passenger_type_key SERIAL PRIMARY KEY,
    passenger_type VARCHAR(100) UNIQUE NOT NULL
);

CREATE TABLE dim_direction(
    direction_key SERIAL PRIMARY KEY,
    direction VARCHAR(50) UNIQUE NOT NULL
);

CREATE TABLE dim_citizenship(
    citizenship_key SERIAL PRIMARY KEY,
    citizenship VARCHAR(150) UNIQUE NOT NULL,
    iso_code VARCHAR(10)
);

CREATE TABLE dim_visa(
    visa_key SERIAL PRIMARY KEY,
    visa_type VARCHAR(150) UNIQUE NOT NULL
);

CREATE TABLE dim_country_residence(
    country_residence_key SERIAL PRIMARY KEY,
    country_name VARCHAR(150) UNIQUE NOT NULL,
    iso_code VARCHAR(10),
    region VARCHAR(100)
);

CREATE TABLE dim_status(
    status_key SERIAL PRIMARY KEY,
    status VARCHAR(50) UNIQUE NOT NULL
);

--Tabla de hechos
CREATE TABLE fact_migration(
    fact_id SERIAL PRIMARY KEY,
    date_key DATE NOT NULL,
    passenger_type_key INTEGER NOT NULL,
    direction_key INTEGER NOT NULL,
    citizenship_key INTEGER NOT NULL,
    visa_key INTEGER NOT NULL,
    country_residence_key INTEGER NOT NULL,
    status_key INTEGER NOT NULL,
    estimate BIGINT NOT NULL,
    standard_error BIGINT,
    CONSTRAINT fk_date FOREIGN KEY(date_key) REFERENCES dim_date(date_key),
    CONSTRAINT fk_passenger FOREIGN KEY(passenger_type_key) REFERENCES dim_passenger_type(passenger_type_key),
    CONSTRAINT fk_direction FOREIGN KEY(direction_key) REFERENCES dim_direction(direction_key),
    CONSTRAINT fk_citizenship FOREIGN KEY(citizenship_key) REFERENCES dim_citizenship(citizenship_key),
    CONSTRAINT fk_visa FOREIGN KEY(visa_key) REFERENCES dim_visa(visa_key),
    CONSTRAINT fk_countryres FOREIGN KEY(country_residence_key) REFERENCES dim_country_residence
        (country_residence_key),
    CONSTRAINT fk_status FOREIGN KEY(status_key) REFERENCES dim_status(status_key)
```

Conclusiones

Este trabajo ayudó a entender cómo limpiar y organizar un conjunto grande de datos para que se pueda usar en un modelo relacional y después en un data warehouse. Aunque el archivo no tenía valores faltantes, sí era importante estandarizar el texto, revisar duplicados y convertir los formatos de fecha.

También se aprendió cómo separar la información en hechos y dimensiones, lo que facilita el análisis y evita datos repetidos. El modelo normalizado permite mantener integridad y trabajar de forma más profesional con la información.

Referencias

Microsoft Support. (2024). *Limpiar datos en Excel*. support.microsoft.com. https://support.microsoft.com/es-es/office/limpiar-datos-en-excel

Talend. (2024). *Introducción a la normalización de bases de datos*. talend.com. https://www.talend.com/es/resources/database-normalization/