12/10/2025

II.3 Técnicas de limpieza de datos

Extracción de Conocimiento en Bases de Datos

Luis Eduardo Aguilar Sarabia

IDGS91N

Docente: Luis Enrique Mascote Cano

Introducción

El presente caso de estudio tiene como finalidad aplicar técnicas de procesamiento, limpieza y modelado de datos al conjunto “International Migration – March 2021”, el cual contiene información sobre el número de migrantes clasificados según país de ciudadanía, tipo de visa y país de última residencia permanente. El objetivo principal es transformar este conjunto en una base de datos estructurada, confiable y optimizada que pueda servir de soporte para análisis estadísticos o sistemas de inteligencia de negocios relacionados con los flujos migratorios internacionales.

Durante la primera etapa se desarrollaron procedimientos de limpieza y depuración de datos, necesarios para garantizar su calidad, homogeneidad y consistencia. Posteriormente, se definieron las tablas de hechos y dimensiones que permiten representar de manera lógica el fenómeno migratorio dentro de un esquema analítico. Finalmente, se construyó un modelo relacional normalizado hasta la Tercera Forma Normal (3FN), junto con su script SQL, que refleja las relaciones entre las entidades principales y mantiene la integridad referencial de los datos.

Proceso

Limpieza de datos

1) Cargar y explorar

```
Primeras filas:
  year_month month_of_release     passenger_type direction citizenship  \
0    2020-02          2021-03  Long-term migrant  Arrivals      non-NZ
1    2020-09          2021-03  Long-term migrant  Arrivals      non-NZ
2    2020-07          2021-03  Long-term migrant  Arrivals      non-NZ
3    2020-07          2021-03  Long-term migrant  Arrivals      non-NZ
4    2020-01          2021-03  Long-term migrant  Arrivals          NZ
5    2020-02          2021-03  Long-term migrant  Arrivals          NZ
6    2020-09          2021-03  Long-term migrant  Arrivals          NZ
7    2020-12          2021-03  Long-term migrant  Arrivals          NZ

                        visa country_of_residence  estimate  standard_error  \
0                   Resident              Andorra         1               0
1                   Resident              Andorra         1               0
2                    Visitor              Andorra         1               0
3  NZ and Australian citizens              Andorra         1               0
4  NZ and Australian citizens              Andorra         1               0
5  NZ and Australian citizens              Andorra         3               0
6  NZ and Australian citizens              Andorra         1               0
7  NZ and Australian citizens              Andorra         0               0

        status
0  Provisional
1  Provisional
2  Provisional
3  Provisional
4  Provisional
5  Provisional
6  Provisional
7  Provisional
```

2) Resumen de calidad (nulos / tipos / únicos)

column	dtype	null_count	null_pct	unique
year_month	object	0	0.0	243
month_of_release	object	0	0.0	7
passenger_type	object	0	0.0	1
direction	object	0	0.0	1
citizenship	object	0	0.0	3
visa	object	0	0.0	7
country_of_residence	object	0	0.0	246
estimate	int64	0	0.0	4045
standard_error	int64	0	0.0	173
status	object	0	0.0	2

3) Normalizar nombres de columnas a snake_case

```
['year_month', 'month_of_release', 'passenger_type', 'direction', 'citizenship', 'visa', 'country_of_residence', 'estimate', 'standard_error', 'status']
```

4) Limpiar texto: strip, quitar no imprimibles, y casing por heurística

```
Textos normalizados. Ejemplo:
  year_month month_of_release     passenger_type direction citizenship  \
0    2020-02          2021-03  Long-term migrant  Arrivals      Non-Nz
1    2020-09          2021-03  Long-term migrant  Arrivals      Non-Nz
2    2020-07          2021-03  Long-term migrant  Arrivals      Non-Nz
3    2020-07          2021-03  Long-term migrant  Arrivals      Non-Nz
4    2020-01          2021-03  Long-term migrant  Arrivals          Nz

                        visa country_of_residence       status
0                   RESIDENT              Andorra  Provisional
1                   RESIDENT              Andorra  Provisional
2                    VISITOR              Andorra  Provisional
3  NZ AND AUSTRALIAN CITIZENS              Andorra  Provisional
4  NZ AND AUSTRALIAN CITIZENS              Andorra  Provisional
```

5) Detectar y convertir columnas numéricas almacenadas como texto

```
def looks_numeric(series, threshold=0.9):
    s = series.dropna().astype(str)
    if len(s)==0: return False
    numeric_like = s.str.replace(r'[,\s]','', regex=True).str.match(r'^-?\d+(\.\d+)?$')
    return numeric_like.mean() >= threshold

for c in df.columns:
    if df[c].dtype == 'object' and looks_numeric(df[c]):
        df[c] = df[c].astype(str).str.replace(',','').str.replace(' ','')
        df[c] = pd.to_numeric(df[c], errors='coerce')
```

6) Parsear `year_month` y crear `time_key` (si existe)

```
Time_key creado. Ejemplo:   year_month year_month_parsed  time_key
0    2020-02        2020-02-01    202002
1    2020-09        2020-09-01    202009
2    2020-07        2020-07-01    202007
3    2020-07        2020-07-01    202007
4    2020-01        2020-01-01    202001
```

7) Eliminar duplicados exactos

```
Duplicados exactos encontrados: 0
```

8) Elegir columnas clave para agrupar

```
['year_month', 'month_of_release', 'passenger_type', 'direction', 'country_of_residence', 'status']
```

9) Agregar (sumar `estimate`) y recomputar `standard_error` correctamente

```
Agrupado. Filas antes: 401772 -> filas después: 45621
  year_month month_of_release     passenger_type direction  \
0    2001-01          2020-09  Long-term migrant  Arrivals
1    2001-01          2020-09  Long-term migrant  Arrivals
2    2001-01          2020-09  Long-term migrant  Arrivals
3    2001-01          2020-09  Long-term migrant  Arrivals
4    2001-01          2020-09  Long-term migrant  Arrivals
5    2001-01          2020-09  Long-term migrant  Arrivals

         country_of_residence status  estimate  standard_error
0                 Afghanistan  Final        24             0.0
1  Africa And The Middle East  Final      3012             0.0
2                  Antarctica  Final        12             0.0
3                   Argentina  Final        40             0.0
4                        Asia  Final     11860             0.0
5                   Australia  Final      5428             0.0
```

10) Detectar outliers en la medida (IQR) y marcar

```
...
Outliers IQR bounds: -226.0 398.0
Outliers encontrados: 7642
```

11) Eliminar filas con estimate negativo

```
Filas con estimate negativo: 0
```

12) Guardar CSV limpio

```
...
CSV limpio guardado en: staging_migration_clean_final_estimate.csv
>>>
```

Determinación de hechos y dimensiones

Tabla de Hechos (Fact Table)

- **FactMigration:** Esta tabla contendrá las métricas cuantitativas del proceso que se está midiendo: la migración.
    - **Propósito:** Almacenar los valores numéricos clave del conjunto de datos, como el número estimado de migrantes y su error estándar. Es el núcleo del modelo, desde donde se realizarán las principales consultas analíticas.
    - **Medidas:** estimate, standard_error.

Tablas de Dimensiones (Dimension Tables)

Estas tablas contendrán los atributos descriptivos que dan contexto a los datos de la tabla de hechos.

- **DimDate:**

    - **Propósito:** Proporcionar el contexto temporal. Permitirá analizar las tendencias de migración a lo largo del tiempo (por año, mes, etc.).
    - **Atributos:** DateKey, YearMonth, MonthOfRelease, Year, Month.
- **DimPassenger:**
    - **Propósito:** Describir el tipo de pasajero.
    - **Atributos:** PassengerKey, PassengerType.
- **DimGeography:**
    - **Propósito:** Almacenar la información geográfica, en este caso, el país de residencia.
    - **Atributos:** CountryKey, CountryOfResidence.
- **DimCitizenship:**
    - **Propósito:** Segmentar los datos por la ciudadanía del migrante.
    - **Atributos:** CitizenshipKey, Citizenship.
- **DimVisa:**
    - **Propósito:** Describir el tipo de visa utilizada por el migrante.
    - **Atributos:** VisaKey, Visa.
- **DimDirection:**
    - **Propósito:** Indicar la dirección del movimiento migratorio (llegadas o salidas).
    - **Atributos:** DirectionKey, Direction.
- **DimStatus:**
    - **Propósito:** Indicar si los datos son provisionales o finales.
    - **Atributos:** StatusKey, Status.

Normalización y almacenamiento

```
CREATE DATABASE MigrationData

USE MigrationData

-- Dimensión de Tiempo
CREATE TABLE DimDate (
    DateKey INT PRIMARY KEY,
    YearMonth VARCHAR(7) NOT NULL,
    MonthOfRelease VARCHAR(7) NOT NULL,
    Year INT,
    Month INT
);

-- Dimensión de Tipo de Pasajero
CREATE TABLE DimPassenger (
    PassengerKey INT PRIMARY KEY,
    PassengerType VARCHAR(50) UNIQUE NOT NULL
);

-- Dimensión Geográfica
CREATE TABLE DimGeography (
    CountryKey INT PRIMARY KEY,
    CountryOfResidence VARCHAR(100) UNIQUE NOT NULL
);

-- Dimensión de Ciudadanía
CREATE TABLE DimCitizenship (
    CitizenshipKey INT PRIMARY KEY,
    Citizenship VARCHAR(50) UNIQUE NOT NULL
);

-- Dimensión de Visa
CREATE TABLE DimVisa (
    VisaKey INT PRIMARY KEY,
    Visa VARCHAR(50) UNIQUE NOT NULL
);

-- Dimensión de Dirección
CREATE TABLE DimDirection (
    DirectionKey INT PRIMARY KEY,
    Direction VARCHAR(20) UNIQUE NOT NULL
);

-- Dimensión de Estado
CREATE TABLE DimStatus (
    StatusKey INT PRIMARY KEY,
    Status VARCHAR(20) UNIQUE NOT NULL
);

-- Tabla de Hechos de Migración
CREATE TABLE FactMigration (
    MigrationID INT PRIMARY KEY,
    DateKey INT,
    PassengerKey INT,
    CountryKey INT,
```

```
    CitizenshipKey INT,
    VisaKey INT,
    DirectionKey INT,
    StatusKey INT,
    Estimate INT,
    StandardError INT,
    FOREIGN KEY (DateKey) REFERENCES DimDate(DateKey),
    FOREIGN KEY (PassengerKey) REFERENCES DimPassenger(PassengerKey),
    FOREIGN KEY (CountryKey) REFERENCES DimGeography(CountryKey),
    FOREIGN KEY (CitizenshipKey) REFERENCES DimCitizenship(CitizenshipKey),
    FOREIGN KEY (VisaKey) REFERENCES DimVisa(VisaKey),
    FOREIGN KEY (DirectionKey) REFERENCES DimDirection(DirectionKey),
    FOREIGN KEY (StatusKey) REFERENCES DimStatus(StatusKey)
);
```

Conclusión

La aplicación de un proceso estructurado de limpieza, transformación y modelado de datos permitió convertir el conjunto "*International Migration – March 2021*" en una base de datos confiable, coherente y analíticamente útil. A través de la limpieza sistemática en Python se corrigieron errores, se homogenizaron formatos y se revisaron registros duplicados, asegurando la calidad del dataset. La determinación de hechos y dimensiones permitió identificar los elementos esenciales para el análisis de los flujos migratorios, mientras que la normalización hasta 3FN garantizó la integridad y eficiencia del almacenamiento de la información.

El resultado final es una base de datos estructurada y lista para ser utilizada en contextos de análisis estadístico o inteligencia de negocios, ofreciendo una representación clara de los movimientos migratorios según ciudadanía, tipo de visa y país de residencia. Este caso evidencia la importancia de la preparación y modelado adecuado de los datos como base fundamental para la generación de conocimiento, la planificación estratégica y la toma de decisiones basadas en información confiable.