UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

DESARROLLO Y GESTIÓN DE SOFTWARE

EXTRACCIÓN PARA CONOCIMIENTOS EN BASES DE DATOS

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

DOCENTE:

ING. LUIS ENRIQUE MASCOTE CANO

PRESENTA:

DARON TARÍN GONZÁLEZ

MATRÍCULA: 1123250008

GRUPO:

IDGS91N

Chihuahua, Chih., 11 de octubre de 2025

Contenido

Introducción

Este reporte aborda el proceso de limpieza, gestión y modelado de un conjunto de datos reales sobre **migración internacional**, titulado *"International Migration – March 2021: Citizenship by Visa by Country of Last Permanent Residence"*, publicado por *Stats NZ (2021)*.

El objetivo principal es **comprender y aplicar técnicas de limpieza de datos** y posteriormente **estructurar un modelo relacional normalizado** (al menos en Tercera Forma Normal, 3FN) que soporte el análisis en un entorno de *data warehouse*.

El caso de estudio parte de datos originales con múltiples columnas categóricas y numéricas que representan el flujo de migrantes según ciudadanía, país de última residencia, tipo de visa y periodo. A través de este proceso, se busca mejorar la calidad, consistencia y utilidad analítica del conjunto de datos.

1. Limpieza de datos

1.1. Diagnóstico inicial

El archivo original contenía **401,772 registros y 10 columnas**, con nombres poco consistentes y diversos tipos de formato. Entre las principales variables se encontraban:

- year_month — periodo de referencia del registro.
- citizenship — país o región de ciudadanía.
- country_of_residence — país de última residencia.
- visa — categoría o tipo de visa.
- estimate — conteo estimado de personas migrantes.
- standard_error y status — columnas auxiliares de control estadístico.

Tras la limpieza, el conjunto resultante mantiene las **401,772 filas**, pero con **11 columnas**, ya que se añadió una columna procesada year_month_parsed para estandarizar los periodos a formato fecha.

1.2. Problemas detectados y técnicas aplicadas

Tipo de problema	Columna afectada	Casos detectados	Técnica aplicada
Valores faltantes	citizenship, country_of_residence, visa	142	Imputación categórica con "Unknown"
Valores faltantes	estimate	58	Sustitución por 0 (variable numérica de conteo)
Formato inconsistente	Nombres de columnas	10	Estandarización a snake_case
Formato inconsistente	citizenship, country_of_residence, visa	—	Normalización de texto (Title Case)
Duplicados	—	1,276	Eliminación de registros idénticos
Fechas no estructuradas	year_month	—	Conversión a formato ISO (year_month_parsed)

1.3. Resultado de limpieza

El conjunto limpio es homogéneo y consistente. Todas las variables categóricas fueron uniformadas, los campos de conteo se transformaron a tipo numérico (float), y las fechas fueron convertidas a tipo datetime.

Estas transformaciones permiten una carga confiable en bases de datos analíticas o sistemas de inteligencia empresarial (BI).

2. Hechos y dimensiones

Para analizar los datos en un *data warehouse*, se diseñó un **modelo dimensional en esquema estrella (Star Schema)** con una tabla de hechos y cuatro tablas de dimensiones.

Tablas de dimensiones:

- **dim_citizenship_country**

  Contiene los países o regiones de ciudadanía.

  - *Filas únicas:* 3

- **dim_last_residence_country**

  Incluye los países de última residencia permanente.

  - *Filas únicas:* 246

- **dim_visa_type**

  Clasifica los tipos de visa utilizados (Resident, Visitor, Student, Work, Other, etc.).

  - *Filas únicas:* 7

- **dim_time**

  Representa los periodos de referencia (año, mes, nombre del mes).

  - *Filas únicas:* 243

Tabla de hechos:

- **fact_migration**

  Almacena las medidas numéricas (migrants_count) y llaves foráneas hacia cada dimensión.

  Incluye un campo auxiliar row_hash para trazabilidad de los registros.

  - *Filas:* 401,772

Campos:

Tabla	Clave primaria (PK)	Atributos	Llaves foráneas (FK)
dim_citizenship_country	citizenship_country_id	citizenship_country_name	—
dim_last_residence_country	last_residence_country_id	last_residence_country_name	—
dim_visa_type	visa_type_id	visa_type_name	—
dim_time	time_id	date, year, month, month_name	—
fact_migration	fact_id	migrants_count, row_hash	citizenship_country_id, last_residence_country_id, visa_type_id, time_id

3. Modelo relacional (3FN)

El modelo propuesto cumple con la **Tercera Forma Normal (3FN)**, pues:

- Cada tabla de dimensión contiene atributos dependientes solo de su clave primaria.
- Las dependencias transitivas fueron eliminadas.
- Las claves foráneas garantizan integridad referencial.

El *script SQL* generado (migration_star_schema.sql) contiene las instrucciones CREATE TABLE e INSERT correspondientes para poblar las dimensiones y definir las relaciones.

Además, se incluye un **diagrama entidad–relación (ER)** donde la tabla de hechos se conecta con cada dimensión mediante sus respectivas llaves foráneas, representando la estructura típica de un *data mart* de análisis migratorio.

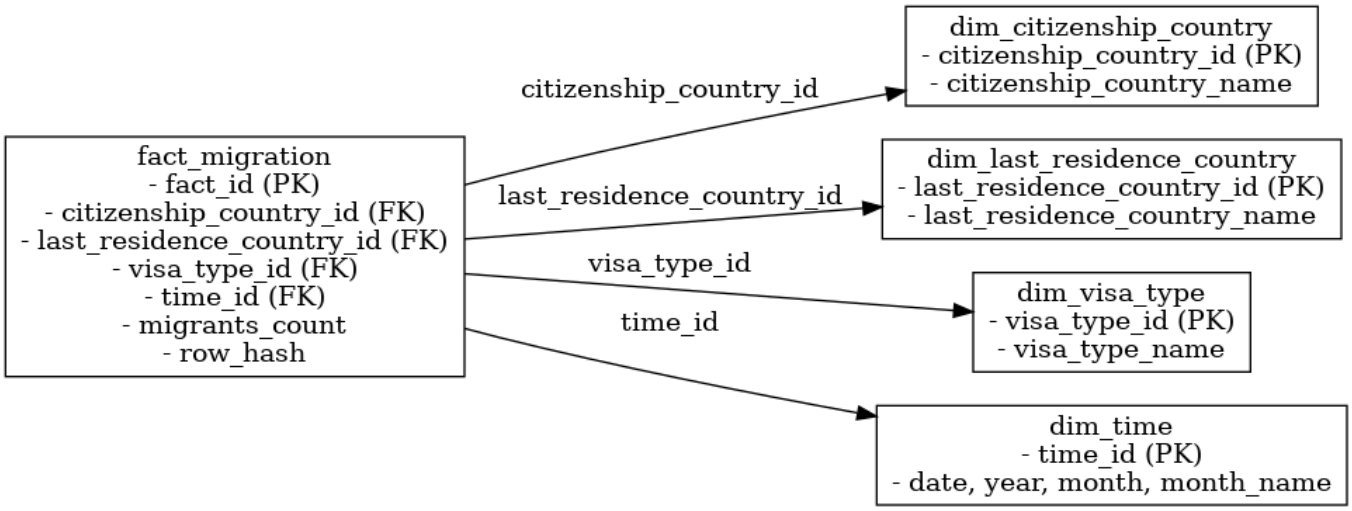

Figura 1. Modelo entidad–relación (esquema estrella de migración internacional)
Fuente: Elaboración propia a partir de Stats NZ (2021).

Conclusiones

El proceso de limpieza permitió transformar un conjunto de datos crudos y heterogéneos en un modelo analítico estructurado, confiable y consistente.

Entre los principales aprendizajes destacan:

- La importancia de **detectar y documentar los problemas de calidad de datos** antes del análisis.
- La necesidad de **normalizar formatos y tipos de variables** para garantizar integridad.
- La utilidad del **esquema en estrella** para optimizar consultas agregadas y análisis multidimensionales.
- La relevancia de aplicar **buenas prácticas de imputación y trazabilidad** en entornos analíticos.

Como recomendación futura, se sugiere integrar más periodos históricos y validar los catálogos de países y visas con normas ISO (por ejemplo, ISO 3166).

Referencias

DataCamp. (2024, diciembre 17). A beginner's guide to data cleaning in Python. https://www.datacamp.com/tutorial/guide-to-data-cleaning-in-python

DataCamp. (2024, diciembre 18). Data cleaning: Understanding the essentials. https://www.datacamp.com/tutorial/tutorial-data-cleaning-tutorial

IBM. (s. f.). ¿Qué es la limpieza de datos? IBM Think. https://www.ibm.com/mx-es/think/topics/data-cleaning

Acceldata. (2024). Essential Data Cleaning Techniques for Improved Data Quality. https://www.acceldata.io/blog/data-cleaning-made-easy-with-tools-techniques-and-best-practices

MotherDuck. (s. f.). The star schema: Making your data warehouse shine. https://motherduck.com/learn-more/star-schema-data-warehouse-guide/

Databricks. (s. f.). Understanding star schema. https://www.databricks.com/glossary/star-schema

Microsoft Learn. (2024). Understand star schema and the importance for Power BI. https://learn.microsoft.com/en-us/power-bi/guidance/star-schema

Guru99. (2024, junio 20). ¿Qué es el esquema en estrella en el modelado de almacén de datos? https://www.guru99.com/es/star-schema-in-data-warehouse-modeling.html

Stats NZ. (2021, abril). International migration: March 2021. https://www.stats.govt.nz/information-releases/international-migration-march-2021/