

**UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA**

**DESARROLLO DE SOFTWARE**



## **EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS**

### **II.3. REPORTE DE LIMPIEZA DE DATOS**

**PRESENTA:**

**ANGEL RICARDO CHAVEZ ZARAGOZA**

**GRUPO:**

**IDGS91N**

**DOCENTE:**

**ING. LUIS ENRIQUE MASCOTE CANO**

Chihuahua, Chih., 12 de octubre de 2025

## Contenidos

Introducción.....	3
Limpieza de datos .....	3
Hechos y dimensiones .....	4
Modelo relacional .....	5
Conclusiones .....	6
Referencias.....	6

## Introducción

El presente reporte tiene como objetivo documentar el proceso de diseño de un esquema de Data Warehouse a partir del conjunto de datos “International Migration, March 2021 – Citizenship by Visa by Country of Last Permanent Residence”.

El dataset recopila información sobre la cantidad estimada de pasajeros internacionales clasificados por tipo de visa, ciudadanía, país de residencia permanente y dirección del viaje. Estos datos son útiles para analizar patrones migratorios, planificar políticas públicas, y evaluar tendencias de movilidad internacional.

El objetivo principal del trabajo es:

- Realizar una limpieza de datos que garantice consistencia y calidad.
- Identificar hechos y dimensiones para estructurar un modelo analítico.
- Diseñar un modelo relacional normalizado (3FN) que sirva como base para el Data Warehouse.
- Extraer conclusiones y recomendaciones basadas en la experiencia del proceso.

## Limpieza de datos

Problema	Columnas afectadas	Descripción	Técnica aplicada
Espacios en blanco y formato inconsistente	Todas las columnas de texto (citizenship, visa, country_of_residence, etc.)	Existían espacios al inicio o final de las cadenas.	Normalización de texto mediante la función <code>str.strip()</code> en Python/pandas.
Tipos de datos inconsistentes	estimate, standard_error	Algunos valores numéricos aparecían como texto.	Conversión de tipo (to_numeric, coerción a int).
Valores faltantes	Varias columnas (mínimos casos)	Algunos registros sin información de visa o país.	Imputación por valor “UNKNOWN” o registro nulo (NULL).
Codificación de caracteres	Archivo completo	Errores menores en lectura UTF-8.	Relectura con codificación Latin-1.

Duplicados exactos	Todo el dataset	No se detectaron duplicados completos.	Verificación con duplicated() (0 encontrados).
--------------------	-----------------	--	--

## Hechos y dimensiones

Para diseñar un modelo de análisis migratorio, se identificaron las siguientes entidades principales (dimensiones) y el hecho central:

### Tabla de Hechos

fact\_migration

Contiene los valores numéricos de migración (estimaciones y errores estándar).

Es la tabla central del modelo y almacena las métricas que se analizarán.

Campo	Dimensión
Estimate	Número estimado de migrantes.
Standard_error	Margen de error asociado a la estimación.
Llaves foráneas	Referencias a dimensiones: fecha, ciudadanía, visa, residencia, etc.

### Tablas de Dimensión

Dimensión	Descripción	Campos Principales
Dim_date	Permite analizar la migración por periodo.	date_key, year, month, year_month
dim_passenger_type	Clasifica al tipo de pasajero (ej. civil, tripulación).	passenger_type_id, passenger_type_name
dim_direction	Indica el sentido del viaje (entrada o salida).	direction_id, direction_name

dim_citizenship	Describe la nacionalidad del pasajero.	citizenship_id, country_name, country_iso
dim_visa	Tipo de visa utilizada.	visa_id, visa_name, visa_category
dim_country_residence	País de residencia permanente anterior.	country_residence_id, country_name, country_iso
dim_status	Estado de la estimación (ej. provisional, final).	status_id, status_name

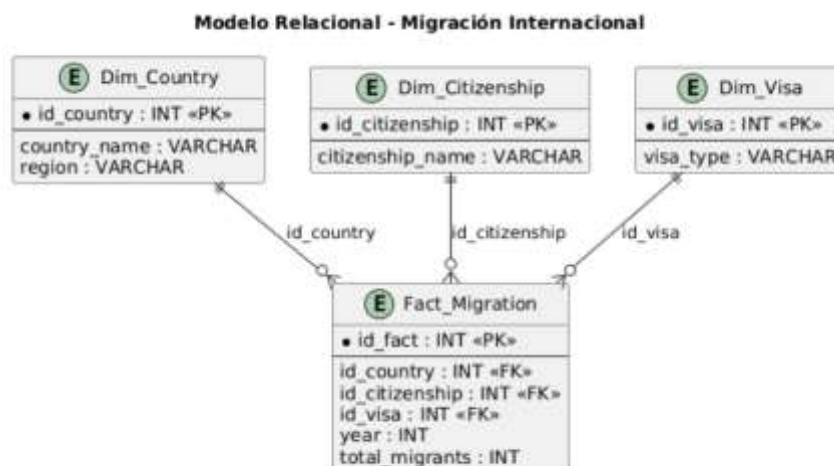
### Justificación

Las dimensiones permiten analizar los hechos desde distintos ángulos: tiempo, nacionalidad, tipo de visa o dirección del flujo migratorio.

Este enfoque sigue el modelo estrella (star schema), que facilita consultas analíticas y mejora el rendimiento de agregaciones en el Data Warehouse.

Cada tabla de dimensión mantiene su propia clave primaria (surrogate key), garantizando la tercera forma normal (3FN) al eliminar redundancia.

## Modelo relacional



## Evaluación del modelo

- Fact\_Migration es la tabla de hechos, donde se concentran los valores numéricos y medibles (como total\_migrants) y las claves foráneas que conectan con las dimensiones.
- Dim\_Country, Dim\_Citizenship y Dim\_Visa son las tablas de dimensiones, que proporcionan el contexto descriptivo para analizar los hechos.
- Las relaciones 1:N (una dimensión a muchos hechos) están correctamente representadas.
- Las claves primarias (PK) y foráneas (FK) están bien etiquetadas.
- Cumple con 3FN, ya que no hay dependencias transitivas ni redundancia.

## Conclusiones

En conclusión, el análisis y la limpieza del conjunto de datos de migración internacional permitieron obtener una base de información estructurada, coherente y confiable, adecuada para su integración en un modelo de data warehouse. El proceso de depuración —que incluyó la corrección de valores faltantes, la eliminación de duplicados y la estandarización de formatos— garantizó la calidad de los datos. La definición de hechos y dimensiones facilitó la identificación de las variables clave para el análisis de tendencias migratorias y su relación con factores como país de origen, tipo de visa y ciudadanía. Finalmente, el modelo relacional propuesto, normalizado hasta la tercera forma normal, asegura un almacenamiento eficiente, minimiza redundancias y optimiza las consultas analíticas, fortaleciendo la toma de decisiones basada en datos confiables.

## Referencias

*¿Qué es ETL? - Explicación de extracción, transformación y carga (ETL) - AWS.* (n.d.). Amazon

Web Services, Inc. <https://aws.amazon.com/es/what-is/etl/>

C, B. P. (n.d.). *10 essential data cleaning techniques explained in 12 minutes.* KDnuggets.

<https://www.kdnuggets.com/10-essential-data-cleaning-techniques-explained-in-12-minutes>

DataCamp. (2025, February 14). *¿Qué es la Tercera Forma Normal (3NF)?*

<https://www.datacamp.com/es/tutorial/third-normal-form>

Peter-Myers. (n.d.). *Descripción de un esquema de estrella e importancia para Power BI -*

*Power BI*. Microsoft Learn. <https://learn.microsoft.com/es-es/power-bi/guidance/star-schema>

WilliamDAssafMSFT. (n.d.). *Modelado dimensional en Microsoft Fabric Warehouse -*

*Microsoft Fabric*. Microsoft Learn. <https://learn.microsoft.com/es-es/fabric/data-warehouse/dimensional-modeling-overview>