**Universidad Tecnológica de Chihuahua Tecnologías de la Información**


## II.3. Reporte de solución de caso de estudio de técnicas de limpieza de datos (25%).

**Alumnos:**
Erick Adrián Sánchez Cervantes
**Grupo:**
IDGS91N
**Materia:**
Extracción de Conocimiento en Bases de Datos
**Docente:**
Enrique Mascote

# Contenido

Introducción ..... 3
Desarrollo ..... 4
1. Limpieza de datos ..... 4
2. Definición de hechos y dimensiones ..... 4
3. Modelo relacional ..... 4
Script ..... 6
Conclusión ..... 8
Referencias ..... 9

# Introducción

En este reporte se presenta el proceso de análisis y limpieza de un conjunto de datos sobre migración internacional correspondiente a marzo de 2021. El archivo utilizado, titulado "international-migration-March-2021-citizenship-by-visa-by-country-of-last-permanent-residence.csv", contiene información sobre el número de personas que llegaron a un país, su ciudadanía, tipo de visa y país de última residencia.

El objetivo principal es transformar esta base de datos en una fuente de información clara, coherente y útil para futuros análisis. A lo largo del trabajo se aplicaron diferentes técnicas para limpiar los datos, definir las partes más importantes (hechos y dimensiones) y proponer un modelo relacional que permita aprovechar la información de forma más eficiente.

Trabajar con datos migratorios no solo implica revisar números, sino también entender historias de movilidad humana, decisiones y contextos que se reflejan en las cifras. Por eso, este análisis busca aportar una visión más ordenada y confiable de la información para facilitar su interpretación y uso.

# Desarrollo

## 1. Limpieza de datos

Al revisar el archivo original, se encontraron algunos problemas comunes como valores vacíos, nombres repetidos o inconsistencias en los países y tipos de visa. Para solucionarlo, se realizaron los siguientes pasos:

- Se eliminaron los registros que tenían información incompleta en columnas clave como "país", "tipo de visa" o "total de llegadas".
- Se estandarizaron los nombres de los países (por ejemplo, se unificó "United States" y "USA").
- Las columnas numéricas fueron convertidas a valores enteros para poder analizarlas correctamente.
- Se eliminaron duplicados y se corrigieron errores de formato en los encabezados.

Después de esta limpieza, los datos quedaron más organizados, consistentes y listos para usarse en análisis posteriores o en una base de datos relacional. Este proceso fue esencial para asegurar que los resultados sean confiables y representen la realidad de la migración durante ese periodo.

## 2. Definición de hechos y dimensiones

Para dar estructura al análisis, se identificaron los elementos más importantes de la información:

Hecho principal: número total de llegadas internacionales registradas en marzo de 2021.

Dimensiones:

- País de ciudadanía: nacionalidad del migrante.
- País de última residencia: país desde donde la persona migró.
- Tipo de visa: categoría del documento que permite la entrada (turismo, trabajo, residencia, estudio, etc.).
- Tiempo: mes y año en que se registraron los datos.

Definir estos elementos fue clave para poder analizar las relaciones entre variables, como los flujos de migración por tipo de visa o los países con más movimiento migratorio.

## 3. Modelo relacional

Una vez definidos los datos y sus dimensiones, se diseñó un modelo relacional que permite organizar la información de forma más clara y sin redundancias.

Este modelo está normalizado (en tercera forma normal) y se compone de las siguientes tablas:

Tabla de Hechos:

Contiene los datos principales sobre las llegadas, con referencias (llaves foráneas) a las tablas de dimensiones.

- id_hecho (PK)
- id_pais_ciudadania (FK)
- id_pais_residencia (FK)

- id_visa (FK)
- total_llegadas
- fecha

Tablas de Dimensión:

- Dim_Pais_Ciudadania(id_pais_ciudadania, nombre_pais)
- Dim_Pais_Residencia(id_pais_residencia, nombre_pais)
- Dim_Visa(id_visa, tipo_visa)
- Dim_Tiempo(id_tiempo, mes, año)

Con este modelo se facilita el análisis por país, visa o periodo de tiempo, y se puede aplicar en sistemas de inteligencia de negocios o dashboards interactivos.

Script

```sql
-- 1. Eliminación de tablas previas
DROP TABLE IF EXISTS Hechos_Migracion;
DROP TABLE IF EXISTS Dim_Pais_Ciudadania;
DROP TABLE IF EXISTS Dim_Pais_Residencia;
DROP TABLE IF EXISTS Dim_Visa;
DROP TABLE IF EXISTS Dim_Tiempo;

-- 2. Creación de tablas de Dimensión
CREATE TABLE Dim_Pais_Ciudadania (
    id_pais_ciudadania SERIAL PRIMARY KEY,
    nombre_pais VARCHAR(100) NOT NULL UNIQUE
);

CREATE TABLE Dim_Pais_Residencia (
    id_pais_residencia SERIAL PRIMARY KEY,
    nombre_pais VARCHAR(100) NOT NULL UNIQUE
);

CREATE TABLE Dim_Visa (
    id_visa SERIAL PRIMARY KEY,
    tipo_visa VARCHAR(50) NOT NULL UNIQUE
);

CREATE TABLE Dim_Tiempo (
    id_tiempo SERIAL PRIMARY KEY,
    mes VARCHAR(15) NOT NULL,
    anio INT NOT NULL,
    CONSTRAINT unique_mes_anio UNIQUE (mes, anio)
```

```
);

-- 3. Creación de la tabla de Hechos
CREATE TABLE Hechos_Migracion (
    id_hecho SERIAL PRIMARY KEY,
    id_pais_ciudadania INT NOT NULL,
    id_pais_residencia INT NOT NULL,
    id_visa INT NOT NULL,
    id_tiempo INT NOT NULL,
    total_llegadas INT NOT NULL CHECK (total_llegadas >= 0),

    FOREIGN KEY (id_pais_ciudadania)
        REFERENCES Dim_Pais_Ciudadania (id_pais_ciudadania)
        ON UPDATE CASCADE
        ON DELETE RESTRICT,

    FOREIGN KEY (id_pais_residencia)
        REFERENCES Dim_Pais_Residencia (id_pais_residencia)
        ON UPDATE CASCADE
        ON DELETE RESTRICT,

    FOREIGN KEY (id_visa)
        REFERENCES Dim_Visa (id_visa)
        ON UPDATE CASCADE
        ON DELETE RESTRICT,
```

```
        ON DELETE RESTRICT,

    FOREIGN KEY (id_tiempo)
        REFERENCES Dim_Tiempo (id_tiempo)
        ON UPDATE CASCADE
        ON DELETE RESTRICT
);

-- 4. Inserciones de ejemplo

-- Dimensiones
INSERT INTO Dim_Pais_Ciudadania (nombre_pais) VALUES ('New Zealand'), ('India'), ('China'), ('United States');
INSERT INTO Dim_Pais_Residencia (nombre_pais) VALUES ('Australia'), ('United Kingdom'), ('Canada'), ('Philippines');
INSERT INTO Dim_Visa (tipo_visa) VALUES ('Work'), ('Study'), ('Residence'), ('Visitor');
INSERT INTO Dim_Tiempo (mes, anio) VALUES ('March', 2021);

-- Hecho ejemplo
INSERT INTO Hechos_Migracion (id_pais_ciudadania, id_pais_residencia, id_visa, id_tiempo, total_llegadas)
VALUES (1, 2, 4, 1, 1250);
```

## Conclusión

El proceso de limpieza y organización de los datos fue fundamental para transformar una base de migración algo desordenada en un recurso claro y confiable.

A través de la depuración, estandarización y diseño del modelo relacional, se logró un conjunto de datos más útil para entender los movimientos migratorios y su relación con factores como el tipo de visa o el país de origen.

Más allá del aspecto técnico, este tipo de análisis nos permite tener una visión más humana de la migración, entendiendo que cada registro representa a una persona que cruzó fronteras buscando nuevas oportunidades. En conclusión, la calidad de los datos no solo mejora la precisión de los análisis, sino también la forma en que interpretamos los fenómenos sociales detrás de ellos.

# Referencias


- Dataset original: International migration - March 2021 (citizenship by visa by country of last permanent residence).
- Fuente: Statistics New Zealand (Stats NZ), 2021
- Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Elsevier.
- Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. Wiley.
- Open Data Handbook. (2020). Data Cleaning Guidelines.