

**UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA  
DESARROLLO Y GESTIÓN DE SOFTWARE**



**REPORTE DE SOLUCIÓN DE CASO DE ESTUDIO DE  
TÉCNICAS DE LIMPIEZA DE DATOS  
EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS**

**PRESENTA:**

**KARLA ALEJANDRA DE LA CRUZ ZEA**

**DOCENTE:**

**ING. LUIS ENRIQUE MASCOTE CANO**

**12 de octubre de 2025**

## **Contenido**

1.	Introducción .....	2
2.	Limpieza de datos.....	2
	Preparacion de la llave compuesta .....	4
	Identificación de valores duplicados.....	5
3.	Determinación de hechos y dimensiones.....	7
	Definición .....	7
	Justificación .....	7
4.	Normalización y almacenamiento .....	8
5.	Conclusión .....	10
6.	Referencias .....	10

## 1. Introducción

La introducción debe ser breve y sintetizada pero bien explicada.

En esta practica mostrare como realizar una limpieza de datos con un programa que se llama Open Refine, es sencillo pero cumple bien con su función.

## 2. Limpieza de datos

Para comenzar a identificar los valores faltantes, las inconsistencias de formato y los duplicados utilizare el programa de Open refine, el cual permite que se puedan hacer filtrado de inconsistencias.

The screenshot shows the OpenRefine interface with a dataset titled "ejercicio\_limpieza\_de\_datos\_IDGS91N". The interface includes a sidebar for facets and filters, a header with search, sort, and filter options, and a main table view with various columns like year\_month, month\_of\_release, passenger\_type, direction, citizenship, visa, country\_of\_residence, estimate, standard\_error, and status.

year_month	month_of_release	passenger_type	direction	citizenship	visa	country_of_residence	estimate	standard_error	status	
1. 2020-02	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	Andorra	1	0	Provisional	
2. 2020-09	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	Andorra	1	0	Provisional	
3. 2020-07	2021-03	Long-term migrant	Arrivals	non-NZ	Visitor	Andorra	1	0	Provisional	
4. 2020-07	2021-03	Long-term migrant	Arrivals	non-NZ	NZ and Australian citizens	Andorra	1	0	Provisional	
5. 2020-01	2021-03	Long-term migrant	Arrivals	NZ	NZ and Australian citizens	Andorra	1	0	Provisional	
6. 2020-02	2021-03	Long-term migrant	Arrivals	NZ	NZ and Australian citizens	Andorra	3	0	Provisional	
7. 2020-09	2021-03	Long-term migrant	Arrivals	NZ	NZ and Australian citizens	Andorra	1	0	Provisional	
8. 2020-12	2021-03	Long-term migrant	Arrivals	NZ	NZ and Australian citizens	Andorra	0	0	Provisional	
9. 2021-01	2021-03	Long-term migrant	Arrivals	non-NZ	Other	United Arab Emirates	0	0	Provisional	
10. 2019-11	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	16	0	Final	
11. 2019-12	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	6	0	Provisional	
12. 2020-01	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	23	0	Provisional	
13. 2020-02	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	19	0	Provisional	
14. 2020-03	2021-03	Long-term migrant	Arrivals	edit	non-NZ	Resident	United Arab Emirates	4	0	Provisional
15. 2020-05	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	2	0	Provisional	
16. 2020-06	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	1	0	Provisional	
17. 2020-07	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	23	1	Provisional	
18. 2020-08	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	3	1	Provisional	
19. 2020-09	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	4	1	Provisional	
20. 2020-10	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	7	1	Provisional	
21. 2020-11	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	4	1	Provisional	
22. 2020-12	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	5	1	Provisional	
23. 2021-01	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	7	2	Provisional	
24. 2021-02	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	8	2	Provisional	
25. 2021-03	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	2	1	Provisional	

Comenzamos revisando si tenemos espacios en blanco. Utilizando las facetas personalizadas revisamos si cada fila tiene espacios en blanco o nulos.

401,772 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

Extensions Wiki

All		year_month	month_of_release	passenger_type	direction	citizenship	visa	country_of_residence	estimate	standard_error				
1.	2020-02	2021-03	Long-term migrant	Arrivals	Facet	Text facet			1	0	Pr	I		
2.	2020-09	2021-03	Long-term migrant	Arrivals	Text filter	Numeric facet			1	0	Pr	I		
3.	2020-07	2021-03	Long-term migrant	Arrivals	Edit cells	Timeline facet			1	0	Pr	I		
4.	2020-07	2021-03	Long-term migrant	Arrivals	Edit column	Scatterplot facet...			1	0	Pr	I		
5.	2020-01	2021-03	Long-term migrant	Arrivals	Transpose	Custom text facet...			1	0	Pr	I		
6.	2020-02	2021-03	Long-term migrant	Arrivals	Sort...	Custom numeric facet...			1	0	Pr	I		
7.	2020-09	2021-03	Long-term migrant	Arrivals	View	Customized facets	Word facet				Pr	I		
8.	2020-12	2021-03	Long-term migrant	Arrivals	Reconcile	Australian citizens	Andorra	Duplicates facet			Pr	I		
9.	2021-01	2021-03	Long-term migrant	Arrivals	NZ	NZ and Australian citizens	Andorra	Numeric log facet			Pr	I		
10.	2019-11	2021-03	Long-term migrant	Arrivals	non-NZ	Other	United Arab E	1-bounded numeric log facet			Pr	I		
11.	2019-12	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab E	Text length facet			Fit			
12.	2020-01	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab E	Log of text length facet			Pr	I		
13.	2020-02	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab E	Unicode char-code facet			Pr	I		
14.	2020-03	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab E	Facet by error			Pr	I		
15.	2020-05	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab E	Facet by null			Pr	I		
16.	2020-06	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	Facet by empty string	1	0	Pr	I		
17.	2020-07	2021-03	Long-term migrant	Arrivals	non-NZ	Resident	United Arab Emirates	Facet by blank (null or empty string)	23	1	Pr	I		

Se realizo el filtro en todas las filas y el resultado fue falso, es decir que no contamos con espacios en blanco o nulos. **false** → registros completos.

Refresh Reset all Remove all

**citizenship** change  
1 choice Sort by: name count  
false 401772  
Facet by choice counts

**country\_of\_residence** change  
1 choice Sort by: name count  
false 401772  
Facet by choice counts

**year\_month** change  
1 choice Sort by: name count  
false 401772  
Facet by choice counts

**month\_of\_release** change  
1 choice Sort by: name count  
false 401772  
Facet by choice counts

## Preparacion de la llave compuesta

Ahora para detectar duplicados vamos a crear una llave compuesta a la que vamos a llamar uniq\_key en donde vamos a ingresar las columnas a filtrar. Usando el lenguaje GREL:

```
cells["year_month"].value.trim() + " | " +
cells["month_of_release"].value.trim() + " | " +
cells["passenger_type"].value.trim() + " | " +
cells["direction"].value.trim() + " | " +
cells["citizenship"].value.trim() + " | " +
cells["visa"].value.trim() + " | " +
cells["country_of_residence"].value.trim() + " | " +
cells["estimate"].value.trim() + " | " +
cells["standard_error"].value.trim() + " | " +
cells["status"].value.trim()
```

**Add column based on column uniq\_key2**

New column name

On error  set to blank  store error  copy value from original column

Expression

```
cells["year_month"].value.trim() + " | " +
cells["month_of_release"].value.trim() + " | " +
cells["passenger_type"].value.trim() + " | " +
cells["direction"].value.trim() + " | " +
cells["status"].value.trim()
```

No syntax error.

Preview History Starred Help

row	value	cells["year_month"].value.trim ...
1.	2020-02   2021-03   Long-term migrant   Arrivals   non-NZ   Resident   Andorranon-NZ   1   0   Provisional	2020-02   2021-03   Long-term migrant   Arrivals   non-NZ   Resident   Andorranon-NZ   1   0   Provisional
2.	2020-09   2021-03   Long-term migrant   Arrivals   non-NZ   Resident   Andorranon-NZ   1   0   Provisional	2020-09   2021-03   Long-term migrant   Arrivals   non-NZ   Resident   Andorranon-NZ   1   0   Provisional
3.	2020-07   2021-03   Long-term migrant   Arrivals   non-NZ   Visitor   Andorranon-NZ   1   0   Provisional	2020-07   2021-03   Long-term migrant   Arrivals   non-NZ   Visitor   Andorranon-NZ   1   0   Provisional

OK Cancel

Resultado de la creación de la columna.

uniq_key2
2020-02   2021-03   Long-term migrant   Arrivals   non-NZ   Resident   Andorranon-NZ   1   0   Provisional
2020-09   2021-03   Long-term migrant   Arrivals   non-NZ   Resident   Andorranon-NZ   1   0   Provisional
2020-07   2021-03   Long-term migrant   Arrivals   non-NZ   Visitor   Andorranon-NZ   1   0   Provisional
2020-07   2021-03   Long-term migrant   Arrivals   non-NZ   NZ and Australian citizens   Andorranon-NZ   1   0   Provisional
2020-01   2021-03   Long-term migrant   Arrivals   NZ   NZ and Australian citizens   AndorraNZ   1   0   Provisional

## Identificación de valores duplicados

Ahora vamos a identificar aquellos valores que tengamos duplicados, utilizando el filtro de facetas en la columna que acabamos de crear el resultado es el siguiente:

The screenshot shows a data visualization interface with two main sections. On the left, a facet menu for the column 'unq\_key2' is open, listing various facet types like Text facet, Numeric facet, Timeline facet, etc. On the right, a facet list for the column 'unq\_key' is displayed, showing several unique values. A purple arrow points from the 'Duplicates facet' entry in the list to the 'true' entry in the facet menu, indicating that the 'true' value is a duplicate.

Y como resultado podemos observar que efectivamente tenemos valores duplicados

This screenshot shows a detailed view of the 'unq\_key2' facet. It displays a list of choices and their counts. The 'true' choice has a count of 256921, and the 'false' choice has a count of 144851. A purple arrow points to the 'true' entry, highlighting it.

Ahora procederemos a eliminar las filas que son iguales

The screenshot shows the OpenRefine interface with the following details:

- Facet / Filter:** Shows a facet for "uniq\_key2" with 2 choices: "false" (144851) and "true" (256921). A "Facet by choice counts" button is also present.
- Table View:** Displays 144,851 matching rows (401,772 total). The columns are "All", "year\_month", "month\_of\_release", and "passenger\_type". The data includes rows like "Transform...", "Edit all columns", "Facet", "Add blank rows", and various data points such as (2021-03, Long-term migrant, Ar).
- Action Bar:** Shows "Rows" and "records" buttons, and a "Show" dropdown with options 5, 10, 25, 50, 100, 500, 1000.
- Context Menu:** A context menu is open over a row, with the "Remove duplicate rows" option highlighted by a purple arrow.
- Dialog Box:** A modal dialog titled "Remove duplicate rows" is open, asking "Select columns used to identify duplicate rows". It lists numerous columns with checkboxes, all of which are checked:
  - year\_month
  - month\_of\_release
  - passenger\_type
  - direction
  - citizenship
  - visa
  - country\_of\_residence
  - uniq\_key2
  - uniq\_key
  - estimate
  - standard\_error
  - status

Buttons at the bottom include "Select all", "Deselect all", "OK", and "Cancel".

13 filas fueron afectadas

The screenshot shows the OpenRefine interface after the cleanup process:

- Header:** Shows the OpenRefine logo, the project name "ejercicio\_limpieza\_de\_datos\_IDGS91N", a "Permalink" link, and buttons for "Remove 13 rows" and "Undo".
- Facet / Filter:** Shows a facet for "uniq\_key2" with 2 choices: "false" (144851) and "true" (256921). A "Facet by choice counts" button is also present.
- Table View:** Displays 401,757 rows. The "Rows" and "records" buttons are visible, along with a "Show" dropdown.

### 3. Determinación de hechos y dimensiones

#### Definición

En un sistema de *data warehouse*, las tablas de hechos almacenan los eventos o medidas cuantitativas que se analizan (por ejemplo, cantidades o valores), mientras que las tablas de dimensiones contienen los atributos descriptivos que permiten contextualizar esos hechos (por ejemplo, país, tipo de visa, año o categoría de migrante).

#### Justificación

El conjunto de datos sobre **migración internacional** contiene información estructurada por país, ciudadanía, tipo de visa, año y número de migrantes. Para un *data warehouse* que analice movimientos migratorios, se pueden identificar las siguientes estructuras:

#### Tabla de hechos:

##### Hecho principal:

###### *Fact\_Migration\_Flows*

Contendrá los datos cuantitativos del número de personas migrantes por combinación de país, ciudadanía y tipo de visa.

##### Medidas principales:

`migration_count` → cantidad total de migrantes registrados.

`year` → periodo de registro (dimensión temporal).

#### Tablas de dimensiones:

##### **Dim\_Country**

Describe el país de residencia o destino.

Atributos: `country_id`, `country_name`, `continent`, `income_level`.

Propósito: permitir análisis geográficos y comparaciones regionales.

##### **Dim\_Citizenship**

Atributos: `citizenship_id`, `citizenship_name`, `region_origin`.

Propósito: identificar nacionalidades y analizar flujos por origen.

### **Dim\_VisaType**

Atributos: visa\_id, visa\_category, description.

Propósito: analizar migraciones según tipo de visa o permiso (trabajo, estudio, residencia, etc.).

### **Dim\_Time**

Atributos: time\_id, year, quarter, month.

Propósito: permitir análisis por períodos y tendencias temporales.

## **4. Normalización y almacenamiento**

Modelo relacional 3FN

Tabla	Campos principales	Llave primaria	Llaves foráneas
<b>Fact_Migration_Flows</b>	fact_id, country_id, citizenship_id, visa_id, time_id, migration_count	fact_id	FK: country_id, citizenship_id, visa_id, time_id
<b>Dim_Country</b>	country_id, country_name, continent, income_level	country_id	—
<b>Dim_Citizenship</b>	citizenship_id, citizenship_name, region_origin	citizenship_id	—
<b>Dim_VisaType</b>	visa_id, visa_category, description	visa_id	—
<b>Dim_Time</b>	time_id, year, quarter, month	time_id	—

Script SQL del modelo relacional

Object Explorer

Connect ▾

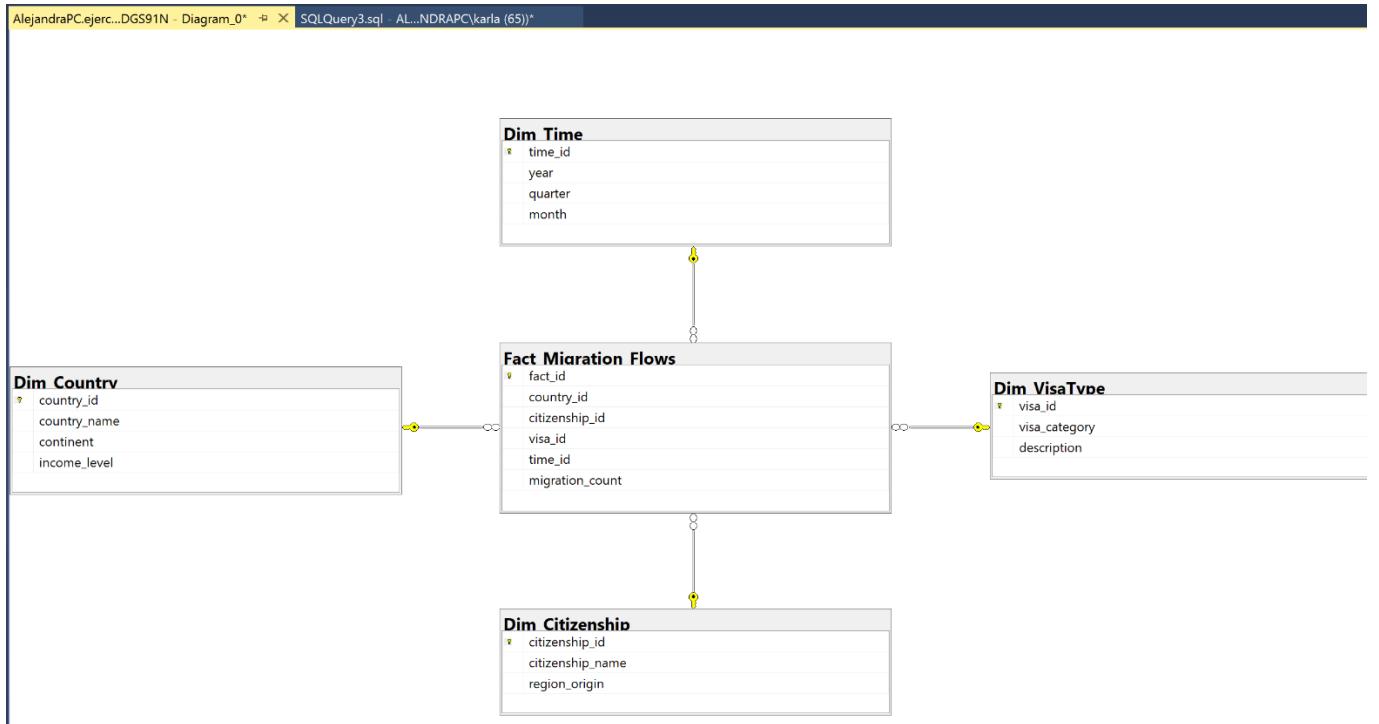
ALEJANDRAPC (SQL Server 16.0.11)

- Databases
- Security
- Server Objects
- Replication
- Always On High Availability
- Management
- Integration Services Catalogs
- SQL Server Agent
- XEvent Profiler

SQLQuery1.sql - AL...NDRAPC\karla (65)\*

```
-- Country
CREATE TABLE Dim_Country (
    country_id INT PRIMARY KEY,
    country_name VARCHAR(100) NOT NULL,
    continent VARCHAR(50),
    income_level VARCHAR(50)
);
-- Citizenship
CREATE TABLE Dim_Citizenship (
    citizenship_id INT PRIMARY KEY,
    citizenship_name VARCHAR(100) NOT NULL,
    region_origin VARCHAR(50)
);
-- Visa
CREATE TABLE Dim_VisaType (
    visa_id INT PRIMARY KEY,
    visa_category VARCHAR(50) NOT NULL,
    description VARCHAR(150)
);
-- Time
CREATE TABLE Dim_Time (
    time_id INT PRIMARY KEY,
    year INT NOT NULL,
    quarter VARCHAR(10),
    month VARCHAR(15)
);
-- Fact
CREATE TABLE Fact_Migration_Flows (
    fact_id INT PRIMARY KEY,
    country_id INT,
    citizenship_id INT,
    visa_id INT,
    time_id INT,
    migration_count INT,
    FOREIGN KEY (country_id) REFERENCES Dim_Country(country_id),
    FOREIGN KEY (citizenship_id) REFERENCES Dim_Citizenship(citizenship_id),
    FOREIGN KEY (visa_id) REFERENCES Dim_VisaType(visa_id),
    FOREIGN KEY (time_id) REFERENCES Dim_Time(time_id)
);
```

## Diagrama lógico (esquema estrella)



## 5. Conclusión

El trabajo con OpenRefine permitió limpiar y preparar correctamente los datos para poder analizar o diseñar un Data Warehouse. A través de la detección de valores faltantes, duplicados y errores de formato, se logró mejorar la calidad del conjunto de datos de migración internacional, haciéndolo más confiable y útil para su almacenamiento y análisis.

El diseño del modelo en forma de estrella ofrece una estructura clara que facilita consultas, reportes y comparaciones entre países, tipos de visa y períodos. Además, la normalización hasta la tercera forma normal evita redundancia, mantiene consistencia y permite crecer el sistema en el futuro sin perder orden.

## 6. Referencias

- OpenRefine.** (2024). *OpenRefine Documentation: Working with messy data.* Recuperado de <https://docs.openrefine.org>
- Van Hooland, S., Verborgh, R., & De Wilde, M. (2013, 5 agosto). **Cleaning Data with OpenRefine. Programming Historian.** <https://programminghistorian.org/en/lessons/cleaning-data-with-openrefine#facetting-and-clustering>
- OpenAI.** (2025). *ChatGPT definition guidance for data analysis and documentation.*