

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.

OpenRefine ejercicio_limpieza_de_datos_IDGS91N [Permalink](#) Open... Export Help

Facet / Filter Undo / Redo 0/0 **401,772 rows** Extensions Wikibase

Show as: **rows** records Show: 5 10 25 50 100 500 1000 rows « first « previous 1 - 25 next » last »

Using facets and filters 
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.
Not sure how to get started?
[Watch these screencasts](#)

	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	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 personaliazdas revisamos si cada fila tiene espacios en blanco o nulos.

401,772 rows Extensions Wik

Show as: rows records Show: 5 10 25 50 100 500 1000 rows « first < previous 1 - 1000 next >

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

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 Language General Refine Expression Language (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()` 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:

uniq_key2	uniq_key
Facet	Andorranon-NZ non-NZ Resident 2020-02
Text filter	Andorranon-NZ non-NZ Resident 2020-09
Edit cells	Andorranon-NZ non-NZ Visitor 2020-07
Edit column	Andorranon-NZ non-NZ NZ and Australian citizens 2020-07
Transpose	AndorraNZ NZ NZ and Australian citizens 2020-01
Sort...	Word facet
View	Duplicates facet
Reconcile	Numeric log facet

Y como resultado podemos observar que efectivamente tenemos valores duplicados

uniq_key2
2 choices Sort by: name count
false 144851
true 256921
Facet by choice counts

Ahora procederemos a eliminar las filas que son iguales

Facet / Filter Undo / Redo 3 / 3

Refresh Reset all Remove all

uniq_key2 change invert reset

2 choices Sort by: name count

false 144851 exclude

true 256921

Facet by choice counts

144,851 matching rows (401,772 total)

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

All year_month month_of_release passenger_type

Transform...

Edit all columns

Facet

Add blank rows

Edit rows

Edit columns

View

Star rows

Unstar rows

Flag rows

Unflag rows

Remove matching rows

Remove duplicate rows

Remove duplicate rows

Select columns used to identify duplicate rows

- ☒ year_month
- ☒ month_of_release
- ☒ passenger_type
- ☒ direction
- ☒ citizenship
- ☒ visa
- ☒ country_of_residence
- ☒ uniq_key2
- ☒ uniq_key
- ☒ estimate
- ☒ standard_error
- ☒ status

Select all Deselect all

OK Cancel

13 filas fueron afectadas

OpenRefine ejercicio_limpieza_de_datos_IDGS91N Permalink Remove 13 rows Undo

Facet / Filter Undo / Redo 5 / 5 401,757 rows

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 periodos y tendencias temporales.

4. Normalización y almacenamiento

Modelo relacional 3FN

Tabla	Campos principales	Llave primaria	Llaves foráneas
Fact_Migration_Flows	fact_id, country_id, citizenship_id, visa_id, time_id, migration_count	fact_id	FK: country_id, citizenship_id, visa_id, time_id
Dim_Country	country_id, country_name, continent, income_level	country_id	—
Dim_Citizenship	citizenship_id, citizenship_name, region_origin	citizenship_id	—
Dim_VisaType	visa_id, visa_category, description	visa_id	—
Dim_Time	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)

AlejandraPC\ejerc...DGS91N - Diagram_0

SQLQuery3.sql - AL...NDRAPC\karla (65))

```

graph TD
    DimCountry[Dim Country] --> FactMigrationFlows[Fact Migration Flows]
    DimCitizenship[Dim Citizenship] --> FactMigrationFlows
    DimVisaType[Dim VisaType] --> FactMigrationFlows
    DimTime[Dim Time] --> FactMigrationFlows

```

Dim Country

- country_id
- country_name
- continent
- income_level

Fact Migration Flows

- fact_id
- country_id
- citizenship_id
- visa_id
- time_id
- migration_count

Dim Citizenship

- citizenship_id
- citizenship_name
- region_origin

Dim VisaType

- visa_id
- visa_category
- description

Dim Time

- time_id
- year
- quarter
- month

9

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 periodos. 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*.