

Data Warehouse para una fundación de acompañamiento de duelo

Segundo Parcial

Bryan Alan Vargas Chávez
César Román Zúñiga

Visión General

01 Introducción

02 Objetivos del proyecto

03 Calendario de Desarrollo

04 Evaluar la calidad y consistencia de los datos. (limpieza)

05 Reglas de Proceso

06 Desarrollar el proceso de extracción, transformación y carga (ETL).

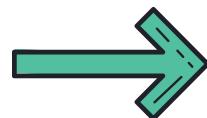
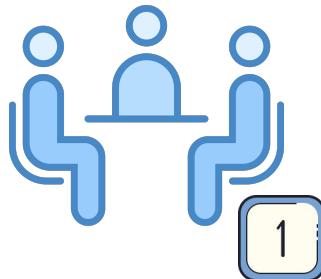
07 Diagrama ER: Constelación

08 Desarrollo del Data Warehouse

09 Realizar Pruebas Unitarias

10 Referencias

Introducción



Entrega de
Documentación de
Avances



Dudas sobre
datos específicos



Reglas de Negocio

Objetivos del proyecto



General

Analizar, diseñar y **desarrollar** un Data Warehouse



Específicos

- Planificar y definir los **requisitos** del Data Warehouse para su correcto uso.
- Analizar la información para el diseño y desarrollo del proceso de **extraer, transformar y cargar (ETL)** los datos al Data Warehouse.
- **Desarrollar el Data Warehouse** para la carga de los datos y posterior explotación.
- Conectar el Data Warehouse con una **herramienta de explotación** para visualizar los datos.

Calendario de Desarrollo

Diagrama de Gantt

Evaluar la calidad y consistencia de los datos (limpieza)



Se utilizó la herramienta de Ataccama DQ Analyzer para **crear reportes de perfilamiento** de los datos.

Se realizó la comparación de los datos antes y después de su limpieza.

Perfilamiento de datos: Resumen del estado de la información.
Permite evaluar **reglas de negocio** y observar **patrones de la información**

Evaluar la calidad y consistencia de los datos (limpieza)

Inputs and Roll Ups

- [BAJAS_20_21](#)
- [DOLIENTES_2020](#)
- [DOLIENTES_2021](#)

Basic	Frequency	Domains	Masks	Quantiles	Groups
Mask Analysis					
Value	Count	Percentage			
LLLLLL DD LL LLLL	20	5.45%			
LLLLLL DD LL LLLLLL	34	9.26%			
LLLLLL DD LL LLLLLL	19	5.18%			
LLLLLL DD LL LLLL	17	4.63%			
LLLLLL DD LL LLLL	16	4.36%			
LLLLLL D LL LLLL	14	3.81%			
LLLLLL DD LL LLLL	13	3.54%			
LLLLLL DD LL LLLL	13	3.54%			
LLLLLL DD LL LLLLLL	12	3.27%			
LLLLLL DD LL LLLLLL	11	3.00%			
LLLLLL D LL LLLLLL	10	2.72%			
LLLLLL DD LL LLLL	10	2.72%			
LLLLLL D LL LLLL	10	2.72%			
LLLLLL DD LL LLLL	10	2.72%			
LLLL L LL LLLL	9	2.45%			
LLLLLL DD LL LLLL	9	2.45%			
LLLL DD LL LLLL	8	2.18%			
LLLL DD LL LLLLLL	8	2.18%			
LLLL D LL LLLL	8	2.18%			
LLLLLL DD LL LLLLLL	8	2.18%			
LLLL DD LL LLLL	7	1.91%			
LLLLLL D LL LLLL	7	1.91%			
LLLL L LL LLLL	6	1.63%			
LLLL DD LL LLLL	6	1.63%			
LLLLLL D LL LLLLLL	6	1.63%			
LLLL D LL LLLL	5	1.36%			
LLLL DD LL LLL	5	1.36%			
LLLLLL DD LL LLLL	5	1.36%			
LLLL DD LL LLL	4	1.09%			
LLLLLL DD LL LLLLLL	4	1.09%			
LLLLLL D LL LLLLLL	3	0.82%			
LLLLLL DD LL LLL	3	0.82%			
LLLLLL D LL LLLL	3	0.82%			
LLLLLL D LL LLLLLL	3	0.82%			
LLLL DD LL LLLLLL	2	0.54%			
LLLL DD LL LLLL	2	0.54%			
LLLL D LL LLLL	2	0.54%			
LLLL DD LL LLL	2	0.54%			
LLLLDD LL LLLLLL	2	0.54%			
LLLLLL D LL LLLLLL	2	0.54%			
LLLLLL DD LL LLLL	2	0.54%			
LLLLLL D LL LLLL	2	0.54%			

Inputs and Roll Ups

- [Canalizados20_21](#)
- [Dolientes20_21](#)

Basic	Frequency	Domains	Masks	Quantiles	Groups			
Columns								
Expression	Type	Nulls	Not nulls	Distinct	Unique	Min	Max	Median
id_doliente	STRING	0	1,139	1,139	1,139	1	999	485
marca_temporal	STRING	0	1,139	373	67	1990-01-01	2021-12-30	2021-03-08
edad	STRING	0	1,139	67	1	-1	85	43
ciudad_pais	STRING	0	1,139	51	9	Aguascalientes	Zacatecas	Estado de MĂŠxico
preferencia_de_horario	STRING	0	1,139	4	0	AM	PM	PM
medio_de_enterarse	STRING	0	1,139	6	2	LĂnea Origen	Soy peregrino de Magdala	No Recopilado
quieres_recibir_info	STRING	0	1,139	1	0	FALSE	FALSE	FALSE
ser_querido_fecha_muerte	STRING	0	1,139	584	323	1931-01-29	2023-10-30	2020-11-15
ser_querido_tipo_relacion	STRING	0	1,139	5	0	Amistad	RelaciĂn Amorosa	Familia Directa
ser_querido_motivo_muerte	STRING	0	1,139	8	0	Accidentes	Problemas de salud mental	Enfermedad
ser_querido_edad_muerte	STRING	0	1,139	99	4	-1	99	5
genero	STRING	0	1,139	3	0	Hombre	No Recopilado	Mujer

Basic	Frequency	Domains	Masks	Quantiles	Groups
Mask Analysis					
Value	Count	Percentage			
DDDD-DD-DD	1139	100.00%			

Evaluar la calidad y consistencia de los datos (limpieza)

Inputs and Roll Ups

- [BAJAS_20_21](#)
- [DOLIENTES_2020](#)
- [DOLIENTES_2021](#)

Frequency Analysis		
Value	Count	Percentage
Enfermedad	2	0.54%
Coronavirus	97	26.43%
Accidente	22	5.99%
Suicidio	16	4.36%
Covid	10	2.72%
enfermedad	9	2.45%
COVID	6	1.63%
Homicidio	5	1.36%
Infarto	5	1.36%
coronavirus	4	1.09%
Paro cardiaco	4	1.09%
accidente	3	0.82%
Asesinato	3	0.82%
covid	3	0.82%
infarto	3	0.82%
Lo asesinaron	3	0.82%
Desaparecido	2	0.54%
leucemia	2	0.54%
A Julio le dio un infarto y Demetrio Neumonia	1	0.27%
a mi hijo lo mataron, Y mi hermano de Coronavirus	1	0.27%
Aborto	1	0.27%
accidente automovilistico	1	0.27%
Accidente automovilistico	1	0.27%
accidente, enfermedad	1	0.27%
Asesinato	1	0.27%
Ambos coronavirus	1	0.27%
Aún no nacían	1	0.27%
Broncoaspiración por sobredosis	1	0.27%
Caída, tal vez Infarto	1	0.27%
cáncer	1	0.27%
Cancer de páncreas	1	0.27%
Cancer, depression, y covid	1	0.27%
Gir cadera por caída	1	0.27%
Crugia de corazón	1	0.27%
Complicaciones de la diabetes y que no lo recibían en ningún hospital por el covid.	1	0.27%
Corazón	1	0.27%
Corona virus	1	0.27%
Corona vírus	1	0.27%
Covid 19	1	0.27%
COVID 19	1	0.27%
COVID-19	1	0.27%
De mi abuela y tía coronavirus y de mi esposo asesinato	1	0.27%
Dejo de respirar	1	0.27%

Inputs and Roll Ups

- [Canalizados20_21](#)
- [Dolientes20_21](#)

Dolientes20_21

Columns

Expression	Type	Nulls	Not nulls	Distinct	Unique	Min	Max	Median
<code>id_doliente</code>	STRING	0	1,139	1,139	1,139	1	999	485
<code>marca_temporal</code>	STRING	0	1,139	373	67	1990-01-01	2021-12-30	2021-03-08
<code>edad</code>	STRING	0	1,139	67	1	-1	85	43
<code>ciudad_pais</code>	STRING	0	1,139	51	9	Aguascalientes	Zacatecas	Estado de MÁSxico
<code>preferencia_de_horario</code>	STRING	0	1,139	4	0	AM	PM	PM
<code>medio_de_enterarse</code>	STRING	0	1,139	6	2	LÁnea Origen	Soy peregrino de Magdala	No Recopilado
<code>quieres_recibir_info</code>	STRING	0	1,139	1	0	FALSE	FALSE	FALSE
<code>ser_querido_fecha_muerte</code>	STRING	0	1,139	584	323	1931-01-29	2023-10-30	2020-11-15
<code>ser_querido_tipo_relacion</code>	STRING	0	1,139	5	0	Amistad	RelaciÁn Amorosa	Familia Directa
<code>ser_querido_motivo_muerte</code>	STRING	0	1,139	8	0	Accidentes	Problemas de salud mental	Enfermedad
<code>ser_querido_edad_muerte</code>	STRING	0	1,139	99	4	-1	99	5
<code>genero</code>	STRING	0	1,139	3	0	Hombre	No Recopilado	Mujer
<code>aporta</code>	STRING	0	1,139	2	0	FALSE	TRUE	FALSE

Basic Frequency Domains Masks Quantiles Groups

Frequency Analysis

Value	Count	Percentage
Enfermedad	442	38.81%
COVID-19	402	35.29%
Accidentes	86	7.55%
Otras causas	61	5.36%
Problemas de salud mental	56	4.92%
Homicidio	51	4.48%
Complicaciones del parto	34	2.99%
No Recopilado	7	0.61%

Evaluar la calidad y consistencia de los datos (limpieza)

Inputs and Roll Ups

- [BAJAS_20_21](#)
- [DOLIENTES_2020](#)
- [DOLIENTES_2021](#)

DOLIENTES_2020

Columns		
Expression		
BAJAS		
Q		
Por favor danos tu nombre solo primer nombre		
Qué edad tienes solo números		
SEXO		
Basic	Frequency	Domains Masks Quantiles Groups

Inputs and Roll Ups

- [Canalizados20_21](#)
- [Dolientes20_21](#)

Frequency Analysis

Value	Count	Percentage
267	64.34%	
mujer	112	26.99%
hombre	19	4.58%
MUJER	15	3.61%
Hombre	1	0.24%
Mujer	1	0.24%

Dolientes20_21

Columns

Expression	Type	Nulls	Not nulls	Distinct	Unique	Min	Max	Median
id_doliente	STRING	0	1,139	1,139	1,139	1	999	485
marca_temporal	STRING	0	1,139	373	67	1990-01-01	2021-12-30	2021-03-08
edad	STRING	0	1,139	67	1	-1	85	43
ciudad_pais	STRING	0	1,139	51	9	Aguascalientes	Zacatecas	Estado de MĂxico
preferencia_de_horario	STRING	0	1,139	4	0	AM	PM	PM
medio_de_enterarse	STRING	0	1,139	6	2	LĂnea Origen	Soy peregrino de Magdala	No Recopilado
quieres_recibir_info	STRING	0	1,139	1	0	FALSE	FALSE	FALSE
ser_querido_fecha_muerte	STRING	0	1,139	584	323	1931-01-29	2023-10-30	2020-11-15
ser_querido_tipo_relacion	STRING	0	1,139	5	0	Amistad	RelaciĂn Amorosa	Familia Directa
ser_querido_motivo_muerte	STRING	0	1,139	8	0	Accidentes	Problemas de salud mental	Enfermedad
ser_querido_edad_muerte	STRING	0	1,139	99	4	-1	99	5
genero	STRING	0	1,139	3	0	Hombre	No Recopilado	Mujer
aporta	STRING	0	1,139	2	0	FALSE	TRUE	FALSE

Inputs and Roll Ups

- [BAJAS_20_21](#)
- [DOLIENTES_2020](#)
- [DOLIENTES_2021](#)

DOLIENTES_2021

Columns		
Expression		
Field_0		
Q		
Por favor danos tu nombre solo primer nombre		
SEXO		
Basic	Frequency	Domains Masks Quantiles Groups

Frequency Analysis

Value	Count	Percentage
546	75.31%	
mujer	161	22.21%
hombre	18	2.48%

Frequency Analysis

Value	Count	Percentage
Mujer	958	84.11%
Hombre	140	12.29%
No Recopilado	41	3.60%

Reglas de Proceso



Origen	Código de Regla	Regla de Proceso
Reglas generales	RS	Formato necesario dado por el diagrama de Datos
Reglas generales	RS	Homologación de Datos
Columna fecha_canalizado	R1	Cuando no existe el dato, se considera la fecha de la última escucha y en caso de que no exista escucha se considera la marca temporal.
Columna aporta	R2	Caso 1 (Hojas de Cálculo): La columna aporte si tiene un aporte igual a 0 entonces es verdadero, de lo contrario es falso Caso 2 (Base de Datos operacional): Verificar en tabla de aporte si la cantidad aportada es mayor a 0 entonces es verdadero, de lo contrario es falso.
Tabla Escucha	R3	En caso de que un doliente no tenga escuchas, se considera en la tabla de canalizado como BAJA
Columnas Actualizadas	R4	No se almacena el histórico de cambios
Marca temporal	R5	Se revisa que la fecha sea menor a la primera escucha, en caso de que no. Se toma la primera escucha
Edad	R6	Las edades se redondean
Fecha de Fallecimiento	R7	Fechas donde solo se especifica el año, se toma el último día de ese año
Fecha de Fallecimiento	R8	Fechas donde no se especifica el año, se considera la fecha del año actual de la hoja de cálculo
Fecha de Fallecimiento	R9	Dolientes con más de una pérdida, se toma la más reciente
Fecha de Fallecimiento	R10	Fechas descritas (hace 3 años) se calcularán contra la marca temporal

Reglas de Proceso



Origen	Código de Regla	Regla de Proceso
Se_cumplio	R11	Para los registros que no se tenga si la escucha se cumplio: si el número total de escuchas es menor al número de fechas que se tienen, entonces la escucha no se cumplio
Sexo	R12	Si no aparece el género, se asigna manualmente
Aporto	R13	Si aporto algo es TRUE, de lo contrario FALSE
Escucha	R14	Si la escucha se realizó TRUE, si no FALSE. Originalmente esta con colores, por lo que requiere su que sea textual
Fecha de escucha	R15	Las fechas deben de estar en el formato ISO
ser_querido_tipo_relacion, ser_querido_motivo_muerte, ser_querido_edad_muerte	R16	Dolientes con más de un fallecido, se utiliza la información del primer valor

Desarrollar el proceso de extracción, transformación y carga (ETL).



Desarrollar el proceso de extracción, transformación y carga (ETL).

Campo	Transformación		Carga
	Descripción	Regla de Proceso	Destino
Sexo	Se asignó el género adecuado de acuerdo con los nombres ya existentes. Los nombres que no se pudieron definir se envían a la fundación para que se determine el género	R12	sexo
Fechas	Se categorizó la información que existía en esta columna en distintas dependiendo el contexto. (Status, Marca Temporal)	R0	Status, Marca Temporal
Edad	Eliminar caracteres innecesarios	R0	edad
¿Qué edad tiene tu ser querido?	Eliminar caracteres innecesarios	R0	ser_querido_edad_muerte
Fechas	Si es menor a la primera escucha, se utiliza como marca temporal, si no existe o es mayor se utiliza la primera escucha.	R5	Marca Temporal
Fecha Escuchas	La última Escucha se toma como Fecha de canalizado	R1	Fecha Canalizado
Marca Temporal	Formato de Fecha ISO	R0	Marca Temporal
Fecha Canalizado	Formato de Fecha ISO	R0	Fecha Canalizado
Cuando Falleció tu ser querido	Formato de Fecha ISO	R0	Cuando Falleció tu ser querido

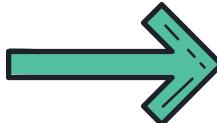
Desarrollar el proceso de extracción, transformación y carga (ETL).

Campo	Transformación		Carga
	Descripción	Regla de Proceso	Destino
Asignación 1 escucha	Formato de Fecha ISO	R0	Asignación 1 escucha
Asignación 2 escucha	Formato de Fecha ISO	R0	Asignación 2 escucha
Asignación 3 escucha	Formato de Fecha ISO	R0	Asignación 3 escucha
Marca temporal	Se revisa que la Fecha sea menor al primero escucha, en caso de que no. Se toma la primera escucha	R1	Marca Temporal
Edad	Las edades se redondean	R0	Edad
Cuando Falleció tu ser querido	Fechas donde solo se especifica el año, se toma el último día de ese año	R0	ser_querido_fecha_muerte
Cuando Falleció tu ser querido	Dolientes con más de una pérdida, se toma la más reciente	R9	ser_querido_fecha_muerte
Cuando Falleció tu ser querido	Fechas descritas (hace 3 años) se calcularán contra la marca temporal	R10	ser_querido_fecha_muerte
¿En qué ciudad y país vives?	Se homologó a nivel estado en los lugares de México y a nivel país en los lugares fuera de México	R01	ciudad_pais
Hora	Formato de Fecha ISO	R01	hora_termino

Limpieza con Inteligencia Artificial (OpenAI API)

=IF(ISDATE(AO7);AO7;GPT("Recopila la fecha en formato ISO";AO7))

26de febrero del 2020
02 de junio de 2020
2020-04-25
26 de Diciembre del 2019



2020-02-26
2020-06-02
2020-04-25
2019-12-26

REGEX

Buscar: `(\d+)$(\n\n^Input[^"]+$)?\n\n^Input:.+$\n^Output:.+?" ,)`

Reemplazar por: `$1$3`

2020-02-17
Input: 2020/03/25, 2020-04-01
Output: 2020-03-25, 2020-04-01
2020-05-17
Input: 2020/06/30, 2020-07-01
Output: 2020-06-30, 2020-07-01
2020-06-15
Input: 2020/07/20, 2020-08-10
Output: 2020-07-20, 2020-08-10



2020-02-17
2020-05-17
2020-06-15

Desarrollar el proceso de extracción, transformación y carga (ETL).

```
def dividir_escuchas(df, n, id_escucha):
    escucha_dividida = df[['id_doliente tabla escuchas', 'Acompañante',
f'Asignación {n} escucha', f'Hora Escucha {n}', f'Se cumpleo {n}']]
    escucha_dividida = escucha_dividida.dropna(subset=[ f'Asignación {n}'
escucha', f'Hora Escucha {n}'])

    # Agregar Ids de escuchas
    escucha_dividida = escucha_dividida.reset_index(drop=True)
    escucha_dividida.index += id_escucha
    escucha_dividida = escucha_dividida.reset_index()
    escucha_dividida = escucha_dividida.rename(columns={'index':
'id_escucha'})
    escucha_dividida['numero_escucha'] = n
    escucha_dividida = escucha_dividida[['id_escucha','id_doliente tabla
escuchas', 'Acompañante', f'Asignación {n} escucha', f'Hora Escucha
{n}', 'numero_escucha',f'Se cumpleo {n}']]
    print(escucha_dividida.shape[0] + id_escucha)
    return escucha_dividida
```

División de escuchas de formato Excel a formato deseado en la base de datos

```

def get_indexed_value(list, index):
    if len(list) > 1:
        if len(list) >= index + 1:
            return list[index]
        else:
            if index > 0:
                return list[-1]
    else:
        return list[0]

for i, row in df.iterrows():
    s_fecha_muerte = row['ser_querido_fecha_muerte'].split(', ')
    if len(s_fecha_muerte) == 1:
        s_fecha_muerte = row['ser_querido_fecha_muerte'].split(',')
    if len(s_fecha_muerte) > 1:
        # Encontrar el index del fallecimiento más próximo
        index_fecha = get_most_recent_date_index(s_fecha_muerte)
        print("index fecha = ", index_fecha)
        print(i, row['ser_querido_fecha_muerte'],
              row['ser_querido_tipo_relacion'], row['ser_querido_motivo_muerte'],
              row['ser_querido_edad_muerte'], sep="\t\t")
        # Tipo relacion
        s_tipo_relacion =
        get_indexed_value(row['ser_querido_tipo_relacion'].split(', '), index_fecha)
        # Motivo Muerte
        s_motivo_muerte =
        get_indexed_value(row['ser_querido_motivo_muerte'].replace(' / ', ' ', ' ').split(', ',
              index_fecha ))
        # Edad Muerte
        s_edad_muerte = get_indexed_value(row['ser_querido_edad_muerte'].replace(' / ', ' ', ' ').split(', ',
              index_fecha ))

```

Selección de datos en columnas con más de un valor

```

print(i, s_fecha_muerte[index_fecha], s_tipo_relacion, s_motivo_muerte, s_edad_muerte,
      end='\n\n', sep='\t\t\t')

df.loc[i,'ser_querido_fecha_muerte'] = s_fecha_muerte[index_fecha]
df.loc[i,'ser_querido_tipo_relacion'] = s_tipo_relacion
df.loc[i,'ser_querido_motivo_muerte'] = s_motivo_muerte
df.loc[i,'ser_querido_edad_muerte'] = s_edad_muerte

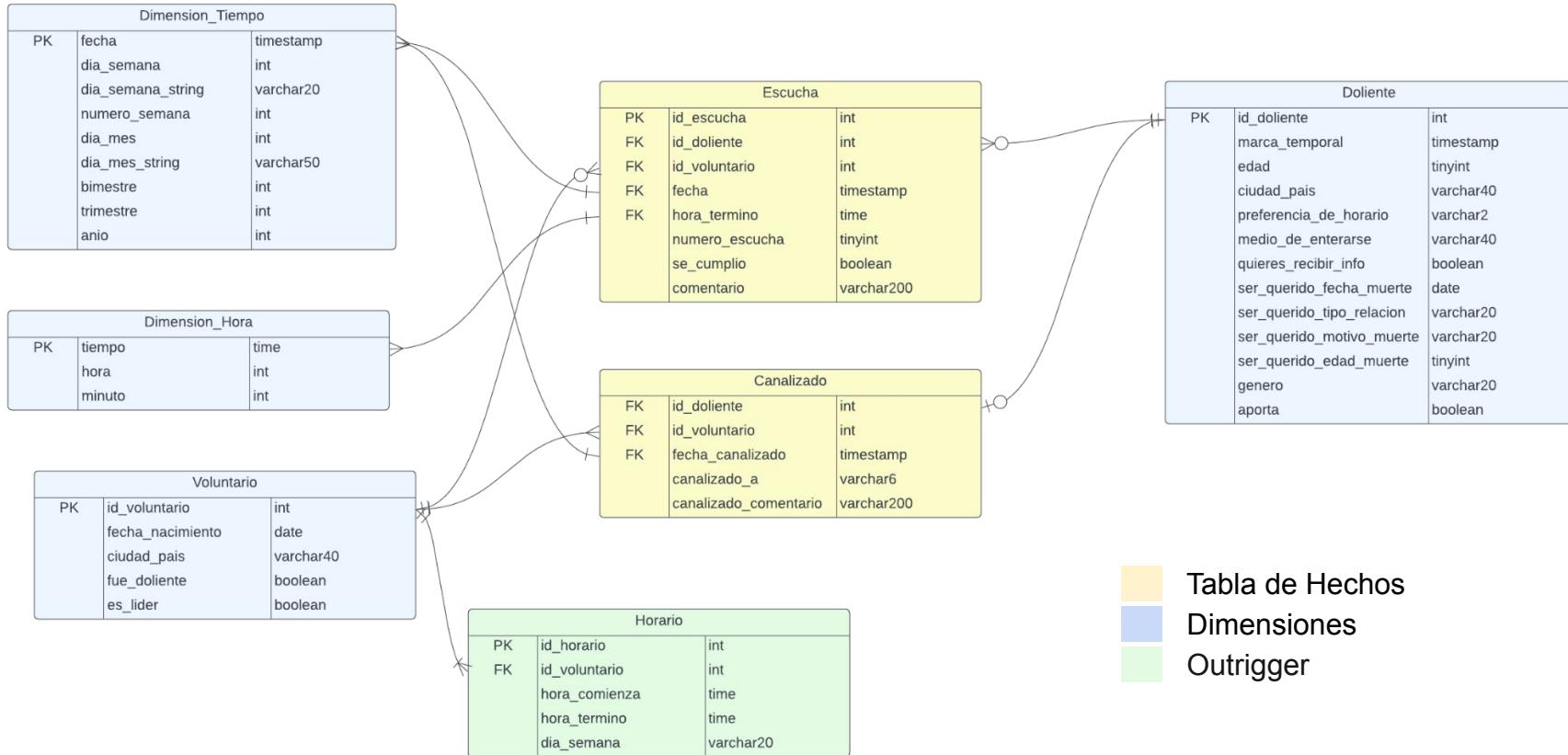
# Columnas con duplicados sin fecha duplicada
s_tipo_relacion = row['ser_querido_tipo_relacion'].split(', ')
df.loc[i,'ser_querido_tipo_relacion'] = s_tipo_relacion[0]

s_edad_muerte = row['ser_querido_edad_muerte'].replace(' ', ' ', ' ').split(', ')
df.loc[i,'ser_querido_edad_muerte'] = s_edad_muerte[0]

s_motivo_muerte = row['ser_querido_motivo_muerte'].replace(' / ', ' ', ' ').split(', ')
df.loc[i,'ser_querido_motivo_muerte'] = s_motivo_muerte[0]

# Limpieza general
df.loc[i,'ser_querido_tipo_relacion'] = df.loc[i,'ser_querido_tipo_relacion'].title()
df.loc[i,'genero'] = df.loc[i,'genero'].title()
df.loc[i,'aporta'] = df.loc[i,'aporta'].replace('VERDADERO', 'TRUE').replace('FALSO', 'FALSE')
df.loc[i,'quieres_recibir_info'] = df.loc[i,'quieres_recibir_info'].replace('No Recopilado',
      'FALSE')
```

Diagrama ER: Constelación



Desarrollo del Data Warehouse

- Tablas de Hechos

```
CREATE TABLE
    escucha (
        id_escucha int,
        id_doliente int,
        id_voluntario int,
        fecha timestamp,
        hora_termino time,
        numero_escucha tinyint,
        se_cumplio boolean,
        comentario varchar(200),
        FOREIGN KEY (id_voluntario) REFERENCES voluntario(id_voluntario),
        FOREIGN KEY (fecha) REFERENCES dimension_tiempo(fecha),
        PRIMARY KEY (id_escucha)
    );

CREATE TABLE
    canalizado (
        id_doliente int,
        id_voluntario int,
        fecha_canalizado timestamp,
        canalizado_a varchar(6),
        canalizado_comentario varchar(200),
        FOREIGN KEY (id_doliente) REFERENCES doliente(id_doliente),
        FOREIGN KEY (id_voluntario) REFERENCES voluntario(id_voluntario)
        FOREIGN KEY (fecha_canalizado) REFERENCES dimension_tiempo(fecha)
    );
```

- Tablas de Dimensiones

```
CREATE TABLE
    doliente (
        id_doliente int not null auto_increment,
        marca_temporal timestamp DEFAULT NULL,
        edad tinyint,
        ciudad_pais varchar(40),
        preferencia_de_horario varchar(2),
        medio_de_enterarse varchar(40),
        quieres_recibir_info boolean,
        ser_querido_fecha_muerte date,
        ser_querido_tipo_relacion varchar(20),
        ser_querido_motivo_muerte varchar(20),
        ser_querido_edad_muerte tinyint,
        genero varchar(20),
        aporta boolean,
        PRIMARY KEY (id_doliente)
    );

CREATE TABLE
    voluntario (
        id_voluntario int not null auto_increment,
        nombre varchar(100),
        fecha_nacimiento date,
        ciudad_pais varchar(40),
        fue_doliente boolean,
        es_lider boolean,
        PRIMARY KEY (id_voluntario)
    );
```

Desarrollo del Data Warehouse

```
-- ! Dolientes 2020-2021
```

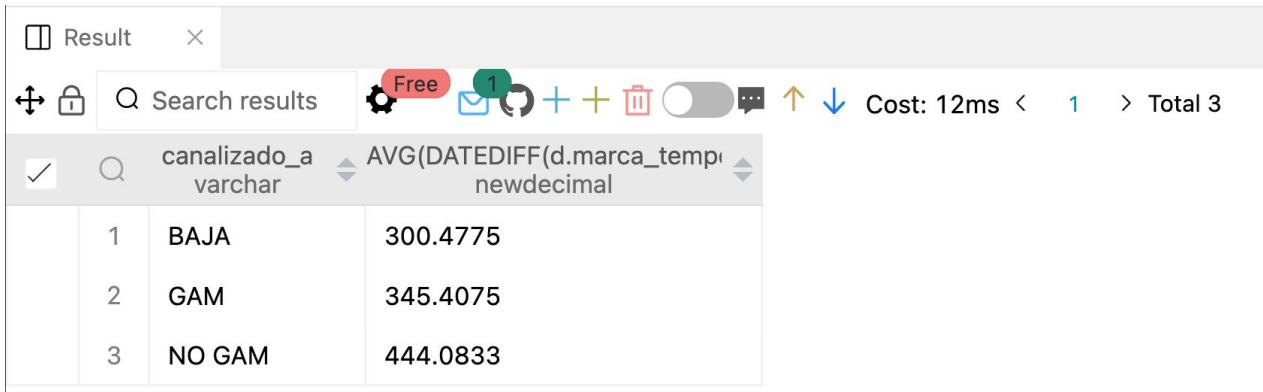
```
LOAD DATA
  LOCAL INFILE
  '/Users/bryanvargas/Documents/Anahuac/Practicum/5
  Documentación/DatosListos/Dolientes20-21.csv' INTO
  TABLE
    DWH_Fundacion_Acompana.doliente FIELDS
  TERMINATED BY ',' ENCLOSED BY "" LINES TERMINATED
  BY '\n' IGNORE 1 ROWS (
    id_doliente, marca_temporal, edad, ciudad_pais,
    preferencia_de_horario, medio_de_enterarse,
    quieres_recibir_info, ser_querido_fecha_muerte,
    ser_querido_tipo_relacion,
    ser_querido_motivo_muerte, ser_querido_edad_muerte,
    genero, aporta);
```

```
-- ! Dolientes 2022
```

```
LOAD DATA
  LOCAL INFILE
  '/Users/bryanvargas/Documents/Anahuac/Practicum/5
  Documentación/DatosListos/Dolientes2022.csv' INTO
  TABLE
    DWH_Fundacion_Acompana.doliente FIELDS
  TERMINATED BY ',' ENCLOSED BY "" LINES TERMINATED
  BY '\n' IGNORE 1 ROWS (
    id_doliente,marca_temporal,edad,ciudad_pais,prefere
    ncia_de_horario,medio_de_enterarse,quieres_recibir_
    info,ser_querido_fecha_muerte,ser_querido_tipo_rela
    cion,ser_querido_motivo_muerte,ser_querido_edad_mue
    rte,genero,aporta);
```

Realizar pruebas unitarias y de integración.

```
-- ! Fecha de registro vs Fecha de fallecimiento de familiar
SELECT c.canalizado_a, AVG(DATEDIFF(d.marca_temporal,d.ser_querido_fecha_muerte))
FROM doliente d INNER JOIN canalizado c ON c.id_doliente = d.id_doliente
WHERE d.ser_querido_fecha_muerte!= '1990-01-01' AND d.marca_temporal !='1990-01-01'
GROUP BY c.canalizado_a;
```



The screenshot shows a database query results window with the following details:

- Result**: The title bar of the results window.
- Search results**: A search bar with a magnifying glass icon.
- Free**: A red button with the word "Free".
- 1**: A green notification badge with the number 1.
- +**: A blue plus sign icon.
- : A red minus sign icon.
- Cost: 12ms**: The execution cost of the query.
- 1**: The current page number.
- Total 3**: The total number of pages.

	canalizado_a	AVG(DATEDIFF(d.marca_temporal,d.ser_querido_fecha_muerte))
1	BAJA	300.4775
2	GAM	345.4075
3	NO GAM	444.0833

Realizar pruebas unitarias y de integración.

```
-- ! Número de escuchas por rangos de edad
SELECT CASE
    WHEN d.edad BETWEEN 18 AND 25 THEN '18-25'
    WHEN d.edad BETWEEN 26 AND 35 THEN '26-35'
    WHEN d.edad BETWEEN 36 AND 45 THEN '36-45'
    WHEN d.edad BETWEEN 46 AND 55 THEN '46-55'
    WHEN d.edad BETWEEN 56 AND 65 THEN '55-65'
    ELSE 'Más de 65'
END AS rango_edad, COUNT(e.numero_escucha)
FROM doliente d INNER JOIN escucha e ON e.id_doliente = d.id_doliente
WHERE d.edad != 0
GROUP BY rango_edad
ORDER BY rango_edad;
```

Result

Free 1

Search results

	rango_edad	COUNT(e.numero_escucha)
1	18-25	173
2	26-35	587
3	36-45	762
4	46-55	654
5	55-65	420
6	Más de 65	234

Realizar pruebas unitarias y de integración.

```
-- ! Número de escuchas por preferencia de horario
SELECT
    preferencia_de_horario,
    COUNT(id_escucha)
FROM doliente d
    INNER JOIN escucha e ON e.id_doliente = d.id_doliente
WHERE
    se_cumplio=1
GROUP BY
    preferencia_de_horario;
```

Result

Free 1

Search results

	preferencia_de_horario	COUNT(id_escucha)
	varchar	bigint
1	PM	1913
2	AM	1429
3	NR	70
4	BO	3

Referencias

- [1] DQ Analyzer - DQ Analyzer. (2023). Ataccama.com. <https://support.ataccama.com/home/docs/dqa> (Accessed: Abril 19, 2023)
- [2] OpenAI API. (2023). Openai.com. <https://platform.openai.com/docs/guides/chat>
- [3] IO tools (text, CSV, HDF5, ...) — pandas 2.0.0 documentation. (2019). Pydata.org. https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html#excel-files
- [4] RegExr: Learn, Build, & Test RegEx. (2021). RegExr; RegExr. <https://regexr.com/> (Accessed: Abril 19, 2023).
- [5] User Guide — pandas 2.0.1 documentation. (2023). Pydata.org. https://pandas.pydata.org/docs/user_guide/index.html
- [6] MySQL :: MySQL 8.0 Reference Manual. (2023). Mysql.com. <https://dev.mysql.com/doc/refman/8.0/en/>
- [7] Salandra, G., Rubio, R. y Guakil, A. (2022) Practicum I Ingeniería - Protocolo “Aplicación web de acompañamiento de duelo.” Universidad Anáhuac México Norte.

Desarrollo del Data Warehouse

```
CREATE TABLE
dimension_tiempo (
    fecha timestamp,
    dia_semana int,
    dia_semana_string varchar(20),
    numero_semana int,
    dia_mes int,
    dia_mes_string varchar(50),
    bimestre int,
    trimestre int,
    anio int,
    PRIMARY KEY (fecha)
);
```

```
CREATE TABLE
dimension_hora (
    tiempo time,
    hora int,
    minuto int,
    PRIMARY KEY (tiempo)
);
```

```
CREATE TABLE
horario (
    id_horario int,
    id_voluntario int,
    hora_comienza time,
    hora_termino time,
    dia_semana varchar(20),
    FOREIGN KEY (id_voluntario) REFERENCES
voluntario(id_voluntario),
    PRIMARY KEY (id_horario)
);
```

Desarrollo del Data Warehouse

```
-- ! Canalizado 2020-2021
```

```
LOAD DATA
  LOCAL INFILE
  '/Users/bryanvargas/Documents/Anahuac/Practicum/5
  Documentación/DatosListos/Canalizados20-21.csv'
  INTO
  TABLE
    DWH_Fundacion_Acompana.canalizado FIELDS
  TERMINATED BY ',' ENCLOSED BY "" LINES TERMINATED
  BY '\n' IGNORE 1 ROWS (
    id_doliente,
    fecha_canalizado,
    canalizado_a,
    id_voluntario
  );
SELECT * FROM DWH_Fundacion_Acompana.canalizado;
```

```
-- ! Canalizado 2022

LOAD DATA
  LOCAL INFILE
  '/Users/bryanvargas/Documents/Anahuac/Practicum/5
  Documentación/DatosListos/Canalizados2022.csv' INTO
  TABLE
    DWH_Fundacion_Acompana.canalizado FIELDS
  TERMINATED BY ',' ENCLOSED BY "" LINES TERMINATED
  BY '\n' IGNORE 1 ROWS (
    id_doliente,
    fecha_canalizado,
    canalizado_a,
    id_voluntario
  );
SELECT * FROM DWH_Fundacion_Acompana.canalizado;
DELETE FROM
  DWH_Fundacion_Acompana.canalizado
WHERE id_doliente IS NOT NULL;
```

Desarrollo del Data Warehouse

```
-- ! Voluntarios

LOAD DATA
    LOCAL INFILE
    '/Users/bryanvargas/Documents/Anahuac/Practicum/5
    Documentación/DatosListos/Voluntarios.csv' INTO
    TABLE
        DWH_Fundacion_Acompana.voluntario FIELDS
    TERMINATED BY ',' ENCLOSED BY "" LINES TERMINATED
    BY '\n' IGNORE 1 ROWS (id_voluntario, nombre);
    SELECT * FROM DWH_Fundacion_Acompana.voluntario;
```

```
-- ! Escuchas

LOAD DATA
    LOCAL INFILE
    '/Users/bryanvargas/Documents/Anahuac/Practicum/5
    Documentación/DatosListos/Escuchas.csv' INTO
    TABLE
        DWH_Fundacion_Acompana.escucha FIELDS
    TERMINATED BY ',' ENCLOSED BY "" LINES TERMINATED
    BY '\n' IGNORE 1 ROWS (
        id_escucha,
        id_doliente,
        fecha,
        hora_termino,
        numero_escucha,
        se_cumplio,
        id_voluntario
    );
```

Desarrollo del Data Warehouse

```
-- ! Dimension_Tiempo

INSERT INTO dimension_tiempo (fecha, dia_semana,
dia_semana_string, numero_semana, dia_mes,
dia_mes_string, bimestre, trimestre, anio)
SELECT fecha, WEEKDAY(fecha) as dia_semana,
DAYNAME(fecha) as dia_semana_string, WEEK(fecha) as
numero_semana, MONTH(fecha) as dia_mes,
MONTHNAME(fecha) as dia_mes_string,
CEILING(MONTH(fecha) / 2) as bimestre,
QUARTER(fecha) as trimestre, YEAR(fecha) as anio
FROM (
    SELECT fecha FROM escucha
    UNION
    SELECT fecha_canalizado as fecha FROM
canalizado as fecha
) AS union_tables;

-- ! Dimension_Hora

INSERT INTO dimension_hora(tiempo, hora, minuto)
SELECT DISTINCT(hora_termino) as tiempo,
HOUR(hora_termino) as hora, MINUTE(hora_termino) as
minuto FROM escucha;
```