Análisis de la pandemia a través de Data Werehouse

Andy Manuel Sequeira Garcia

Priscilla Aylin Brenes Ulloa

Joseph Antonio Rojas Leon

Luis Emerson Solano Artavia

Ingeniería en sistemas de Computación, Universidad Fidélitas

SC-602 Data Warehouse y Base de Datos Multidimensionales

Prof. Marvin Solano Campos

6 de diciembre del 2023

Índice

Contenido

Introducción

En el 2019 su sufrió una de las grandes epidemias de la historia humana. El COVID-19, que provocó grandes cambios a nivel mundial, en muchos ambientes de toda la vida. Debido a toda esta situación mundial se derivaron en datos recolectados masivamente de todo tipo, como, por ejemplo: Datos clínicos, demográficos, etc. Que se recolectaron a través de los años.

En este contexto, se plantea la necesidad de aprovechar las capacidades de los Data Warehouses y la técnica de minería de datos para la exploración, recopilación y análisis eficiente de esta vasta cantidad de información relacionada con el estudio del COVID-19. El objetivo es identificar patrones, tendencias y desafíos clave enfrentados por diferentes países y autoridades de salud durante la pandemia. Este análisis retrospectivo busca proporcionar información valiosa que permita la formulación de estrategias más efectivas para la gestión de futuras epidemias y situaciones similares.

Esta formulación del problema es más concisa y enfocada en el objetivo de tu estudio, que es utilizar Data Warehouses y técnicas de minería de datos para analizar los registros relacionados con el COVID-19 y extraer información relevante para la toma de decisiones y la planificación de futuras respuestas ante situaciones de crisis de salud pública.

Objetivo General

Analizar la información que está recopilada acerca del COVID-19 para mejorar la gestión de crisis de salud pública

Objetivos específicos

1. Limpiar los datos nulos, duplicados o con errores, para que haya la consistencia en el Data Werehouse
2. Identificar patrones y tendencias que se dieron en el transcurso de la epidemia
3. Crear Dashboards para visualizar e interactuar con los datos de manera efectiva para el análisis

Justificación y su importancia

Desde el inicio oficial de la pandemia en el año 2020 la recolección, manejo y análisis de datos han sido procesos que casi era todo el mundo se tuvo que realizar casi de manera inmediata. Durante en el inicio de la pandemia fue muy común leer y ver en las noticias las estadísticas como la taza de contagio, la mortalidad y la eficiencia de las vacunas, para poder sacar estas estadísticas se debió de hacer un análisis muy detallado, porque se demostró cuán importante es el manejo de este tipo de información y saber cómo emplearla.

Como se explicó anteriormente, el análisis de datos fue fundamentas en el transcurso de la pandemia, esto no solo para ver el avance a nivel mundial sino para poder tomar las acciones necesarias para poder disminuir los contagios. Por esto en este trabajo se empleará el uso de herramientas tecnologías para realizar un tipo de análisis similar al que se desarrolló durante la pandemia, con la diferencia que ya se tiene datos más concretos y una mejor compleción de la situación.

Tener los datos y saber cómo emplear las estragas necesarias para el futuro es de gran importancia, de esto la gran relevancia que tienen este tipo de trabajos donde se realiza un análisis de diferentes factores y se realizan acciones o como en este caso se visualizan los datos para un mejor entendimiento de la información.

Requerimientos funcionales

- El sistema mostrara los datos a través de diferentes dimensiones creadas por los estudiantes relacionado al COVID-19 y sus años más fuertes.
- El sistema permitirá a los usuarios ingresar los datos a través de las diferentes tablas creadas para llegar a visualizar los datos ingresados
- Se mostrará el nivel de población contagiada a través de los años que la pandemia estuvo en su máximo apogeo.
- A través del datawarehouse mostraremos el nivel de mortalidad que hubo durante la época de la pandemia.
- El sistema podrá mostrar el nivel de población por edad enfocándose en las personas mayores de 60 y 70 años.
- El sistema eliminara datos en nulos, lo cual no utilizara espacios en blanco.

Requerimientos no funcionales

- Uso de la aplicación SQL Server 2019 para mostrar los datos además del uso de Power BI Desktop esto para mostrar las diferentes gráficas
- El sistema debe ser capaz de leer todos los datos y además de eso debe de por lo menos crear más de tres dimensiones a través de los datos proporcionados del COVID-19.
- Toda la visualización de las tablas creadas por los usuarios debe de responder en menos de 10 segundos.
- El sistema debe de llevar un orden de sus datos, no puede mostrar datos acumulados, si esto ocurre los administradores deben de setear los datos.

Alcance

Este proyecto se llevará a cabo en un lapso de 4 meses, durante los cuales se planea llevar a cabo una serie de pasos mediante los cuales mostraremos una amplia cantidad de datos relacionados con el covid-19, a continuación de procederá a hablar acerca de ellos.

1. Obtención de Datos: En este punto buscaremos la mayor cantidad de información que podamos encontrar hacer del COVID en cada país, entre los datos más relevantes que buscamos serian, las fechas de ingreso, cantidad total de enfermos por país, las regiones, los nuevos casos que podrían aparecer, nuevas muertes y el total de muertes que encontremos hasta ese punto.
2. Después de haber obtenido toda esta información, el próximo paso a seguir seria transferir esta información a una base de datos desde la cual se pueda podamos tener una mejor distribución, un manejo de la información y mediante esto poder realizar mejor las vistas y todos los otros cometidos que surjan a través del proyecto.
3. Después de tener toda la información necesaria en la base de datos, se procederá a realizar algunos gráficos en los cuales podremos observar las épocas en las que hubo más casos del COVID como el promedio de muertes por cada mes a través del lapso durante el cual duro la pandemia.
4. Y por último con los gráficos creados a base de la información y las vistas que se crearon en el SQL server, esperamos que el público hacia el que va dirigido este proyecto tenga una conciencia mundial acerca de inmenso impacto que

tuvo en distintos países y algunos ejemplos de la alta tasa de mortalidad que tuvo durante el lapso en el que su apogeo era más grande.

El propósito de este proyecto poder brindar un panorama aún más amplio a nivel mundial acerca de cómo esta enfermedad trajo consigo múltiples consecuencias para la humanidad, la información a mostrar también incluirá el impacto que tuvo en aquellas personas las cuales su edad superaba los 60 y 70 años.

Este proyecto está centrado únicamente en el COVID-19, por lo cual no se tocaran temas aparte del mismo, como por ejemplo, el fallecimiento de cierta cantidad de personas las cuales fallecieron a causa de paros cardiacos u otro tipo de enfermedades, así también no se tocara el tema de aquellas personas las cuales se sospecha que fallecieron presuntamente a causa del COVID-19 nuestra investigación estará centrada únicamente en casos confirmados ya sea de muerte o de las personas las cuales pudieron haber sido hospitalizadas a causa de esta enfermedad, ni hablaremos de aquellas personas fumadoras o con algún tipo de enfermedad pulmonar.

Limitaciones

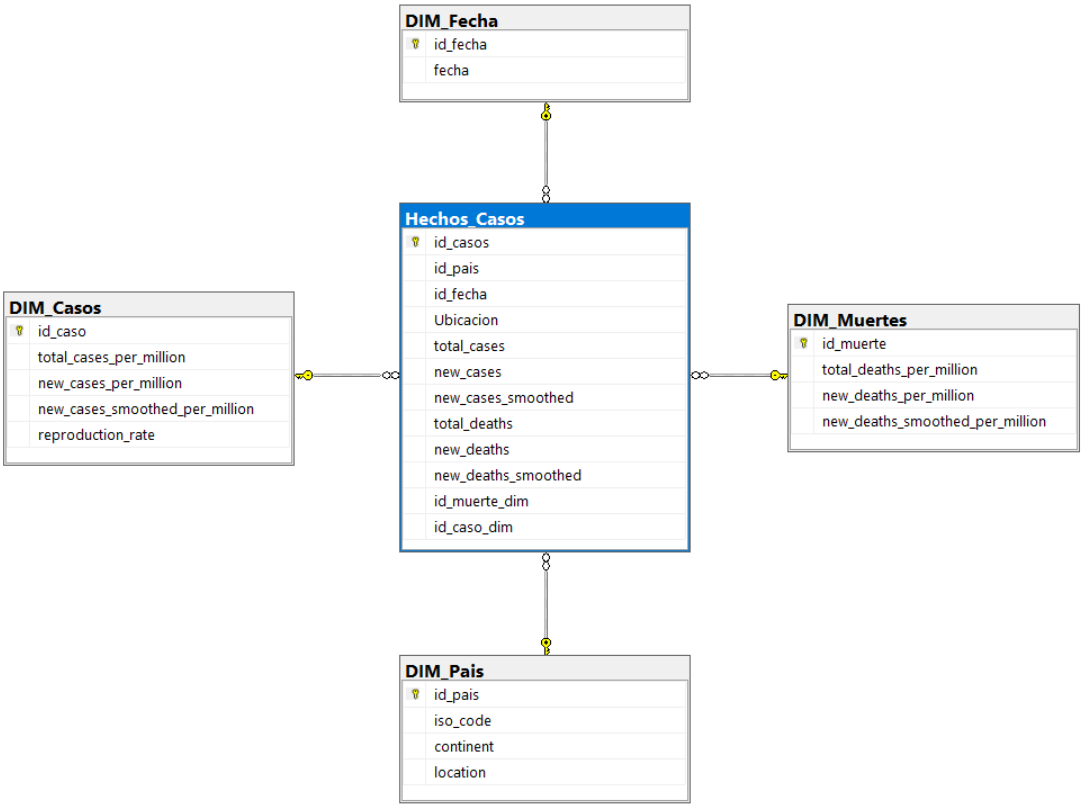
En este tipo de proyectos donde el trabajo de análisis es en base de una información encontrada o suministrada la principal limitación que se encuentra es el acceso a la información ya se está restringido a la información que se posee y el uso de información no oficial puede afectar negativamente el proyecto al no tener la certeza de la procedencia de los datos.

Otra limitación que se presenta en este proyecto es que no se cuenta con algún tipo de asesoría o contacto directo con el cual puedan realizar preguntas o asesorías sobre los datos y todo el proceso de limpieza y análisis será del equipo de trabajo de este proyecto.

Desarrollo del planteamiento

Estructura de la Base de datos

Se desarrolló una base de datos Data Mart con forma de estrella, con la tabla Hechos y 4 tablas DIM, las cuales van a almacenar los datos extraídos de nuestro DatawWrehouse, A continuación una foto de la estructura en Diagrama y el script de la generación de la Base de datos



Fuente (Creación propia diseñada y generado en el SQL Server, Management Studio)

Script:

```
Create Database DW_Proyecto

USE master

Use DW_Proyecto

CREATE TABLE DIM_Pais (

  id_pais INT IDENTITY(1,1) PRIMARY KEY,

  iso_code VARCHAR(10) NOT NULL,

  continent VARCHAR(50) NOT NULL,

  location VARCHAR(100) NOT NULL

);



CREATE TABLE DIM_Fecha (

  id_fecha int IDENTITY PRIMARY KEY,

  fecha varchar(30) not null

);
```

```
CREATE TABLE DIM_Casos (
  id_caso INT IDENTITY(1,1) PRIMARY KEY,
  total_cases_per_million INT NOT NULL,
  new_cases_per_million INT NOT NULL,
  new_cases_smoothed_per_million INT NOT NULL,
  reproduction_rate DECIMAL(10,2) NOT NULL
);

CREATE TABLE DIM_Muertes (
  id_muerte INT IDENTITY(1,1) PRIMARY KEY,
  total_deaths_per_million INT NOT NULL,
  new_deaths_per_million INT NOT NULL,
  new_deaths_smoothed_per_million INT NOT NULL
);

CREATE TABLE Hechos_Casos (
  id_casos INT IDENTITY PRIMARY KEY,
  id_pais INT FOREIGN KEY REFERENCES DIM_Pais(id_pais),
  id_fecha int FOREIGN KEY REFERENCES DIM_Fecha(id_fecha),
  Ubicacion VARCHAR(100),
```

```
  total_cases INT ,
  new_cases INT,
  new_cases_smoothed INT,
  total_deaths INT,
  new_deaths INT,
  new_deaths_smoothed INT,
  id_muerte_dim INT FOREIGN KEY REFERENCES DIM_Muertes(id_muerte),
  id_caso_dim INT FOREIGN KEY REFERENCES DIM_Casos(id_caso)
);

USE DW_Proyecto

select * from DIM_Muertes;

select * from DIM_Casos;

select * from DIM_Pais;

SELECT * FROM DIM_Fecha;
```

```
SELECT * FROM Hechos_Casos;



CREATE OR ALTER PROCEDURE sp_llenar_nulos

AS

BEGIN

  DECLARE @id_pais int = 1

  WHILE @id_pais <= (SELECT MAX(id_pais) FROM DIM_Pais)

  BEGIN

    UPDATE top (1) Hechos_Casos

    SET id_pais = @id_pais

    WHERE id_pais IS NULL



    SET @id_pais = @id_pais + 1

  END



  DECLARE @id_fecha int = 1

  WHILE @id_fecha <= (SELECT MAX(id_fecha) FROM DIM_Fecha)

  BEGIN

    UPDATE top (1) Hechos_Casos

    SET id_fecha = @id_fecha
```

```
  WHERE id_fecha IS NULL

  SET @id_fecha = @id_fecha + 1
END

DECLARE @id_muerte int = 1
WHILE @id_muerte <= (SELECT MAX(id_muerte) FROM DIM_Muertes)
BEGIN
  UPDATE top (1) Hechos_Casos
  SET id_muerte_dim = @id_muerte
  WHERE id_muerte_dim IS NULL

  SET @id_muerte = @id_muerte + 1
END

DECLARE @id_caso int = 1
WHILE @id_caso <= (SELECT MAX(id_caso) FROM DIM_Casos)
BEGIN
  UPDATE top (1) Hechos_Casos
  SET id_caso_dim = @id_caso
```

```
    WHERE id_caso_dim IS NULL


    SET @id_caso = @id_caso + 1
  END
END


EXEC sp_llenar_nulos



/* PROCEDIMIENTOS ALMACENADOS DE CONSULTAS */


--sp_casos_pais_tiempo: Obtiene los casos por país a través del tiempo. Útil para gráficos de tendencia.
CREATE PROCEDURE sp_casos_pais_tiempo
AS
SELECT p.iso_code, f.fecha, h.total_cases
FROM Hechos_Casos h
INNER JOIN DIM_Pais p ON h.id_pais = p.id_pais
INNER JOIN DIM_Fecha f ON h.id_fecha = f.id_fecha
```

```
EXEC sp_casos_pais_tiempo



--sp_tasa_mortalidad: Obtiene la tasa de mortalidad por país.

CREATE PROCEDURE sp_tasa_mortalidad

AS

SELECT p.iso_code,

      (h.total_deaths * 100.0) / h.total_cases AS mortality_rate

FROM Hechos_Casos h

INNER JOIN DIM_Pais p

  ON h.id_pais = p.id_pais



EXEC sp_tasa_mortalidad



--sp_comparacion_mortalidad Comparación de muerte entre paises

CREATE PROCEDURE sp_comparacion_mortalidad

AS

SELECT

  p.iso_code,
```

```
  MAX(h.total_deaths) AS total_muertes,
  MAX(h.total_cases) AS total_casos,
  (MAX(h.total_deaths)*100 / MAX(h.total_cases)) AS tasa_mortalidad
FROM Hechos_Casos h
INNER JOIN DIM_Pais p
  ON h.id_pais = p.id_pais
GROUP BY p.iso_code
ORDER BY tasa_mortalidad DESC

EXEC sp_comparacion_mortalidad

--Tendencia reproducción del virus:
CREATE PROCEDURE sp_tendencia_reproduccion
AS
SELECT AVG(c.reproduction_rate) AS tasa_reproduccion,
       DATEPART(month, f.fecha) AS mes
FROM Hechos_Casos h
INNER JOIN DIM_Casos c
  ON h.id_caso_dim = c.id_caso
```

```
INNER JOIN DIM_Fecha f
  ON h.id_fecha = f.id_fecha
WHERE f.fecha >= '2020-01-01'
GROUP BY DATEPART(month, f.fecha)
ORDER BY mes


EXEC sp_tendencia_reproduccion



--Tendencias Globales de Casos:
CREATE PROCEDURE sp_ObtenerTendenciasGlobalesCasos
AS
BEGIN
    SELECT
        F.fecha AS Fecha,
        SUM(HC.total_cases) AS TotalCasos,
        SUM(HC.new_cases) AS NuevosCasos,
        SUM(HC.total_deaths) AS TotalMuertes,
        SUM(HC.new_deaths) AS NuevasMuertes
    FROM
```

```
        Hechos_Casos HC
        INNER JOIN DIM_Fecha F ON HC.id_fecha = F.id_fecha
    GROUP BY
        F.fecha;
END;

EXEC sp_ObtenerTendenciasGlobalesCasos

--Estadísticas Generales de Guatemala, Belice y el Salvador:
CREATE PROCEDURE sp_ObtenerEstadisticasGeneralesPorPais
AS
BEGIN
    SELECT
        P.location AS Pais,
        SUM(HC.total_cases) AS TotalCasos,
        SUM(HC.total_deaths) AS TotalMuertes
    FROM
        Hechos_Casos HC
        INNER JOIN DIM_Pais P ON HC.id_pais = P.id_pais
    GROUP BY
```

```
    P.location;

END;


EXEC sp_ObtenerEstadisticasGeneralesPorPais
```

Diccionario completo del Datawerehouse

Tabla 1: DIM_Pais

Descripción: En esta tabla se tienen toda la información referente al pais.

- Columna 1: id_pais
  - Descripción: Este campo tiene la infromacion del id del pais.
  - Tipo de datos: int
  - Restricciones de la columna: no aplica
  - Longitud y precisión: 1,1
  - Valores permitidos: solo valores enteros
- Columna 2 : iso_code
  - Descripción: Este campo tiene el codigo identificador del país.
  - Tipo de datos: varchar
  - Restricciones de la columna: no aplica
  - Longitud y precisión: 10
  - Valores permitidos: solo valores varchar
- Columna 3 : continent
  - Descripción: Este campo tiene la informacion del continente del país.
  - Tipo de datos: varchar
  - Restricciones de la columna: no aplica

    - Longitud y precisión: 50
    - Valores permitidos: solo valores varchar
- Columna 4 : location
    - Descripción: Este campo tiene la informacion de la locación del país.
    - Tipo de datos: varchar
    - Restricciones de la columna: no aplica
    - Longitud y precisión: 100
    - Valores permitidos: solo valores varchar

Clave primaria: id_pais

Origen de datos: owid-covid-data.xlsx

Tabla 2: DIM_Fecha

Descripción: En esta tabla se tienen toda la información referente a la fecha.

- Columna 1: id_fecha
    - Descripción: Este campo tiene la infromacion del id de la fecha.
    - Tipo de datos: Date
    - Restricciones de la columna: no aplica
    - Valores permitidos: solo valores de tipo date
- Columna 2 : iso_code
    - Descripción: Este campo tiene la fecha.
    - Tipo de datos: Date
    - Restricciones de la columna: no aplica
    - Valores permitidos: solo valores de tipo date

Clave primaria: id_fecha

Origen de datos: owid-covid-data.xlsx

Tabla 3: DIM_Casos

Descripción: En esta tabla se tienen toda la información a los casos de Covid.

- Columna 1: id_caso
    - Descripción: Este campo tiene la infromacion del id de la caso.
    - Tipo de datos: int
    - Restricciones de la columna: no aplica
    - Longitud y precisión: 1,1
    - Valores permitidos: solo valores de tipo int
- Columna 2: total_cases_per_million
    - Descripción: Data el total de los casos por millon
    - Tipo de datos: Int
    - Restricciones de la columna: no aplica
    - Valores permitidos: solo valores de tipo int
- Columna 3: new_cases_per_million
    - Descripción: : Data los nuevos casos por millon
    - Tipo de datos: Int
    - Restricciones de la columna: no aplica
    - Valores permitidos: solo valores de tipo int
- Columna 4: new_cases_smoothed_per_million
    - Descripción: Data el total de los casos leves por millon

    - Tipo de datos: Int
    - Restricciones de la columna: no aplica
    - Valores permitidos: solo valores de tipo int
- Columna 5: reproduction_rate
    - Descripción: Este campo tiene la taza de produccion.
    - Tipo de datos: Decimal
    - Restricciones de la columna: no aplica
    - Longitud y precisión: 10,2
    - Valores permitidos: solo valores de tipo decimal

Clave primaria: id_fecha

Origen de datos: owid-covid-data.xlsx

Tabla 4: DIM_Muertes

Descripción: En esta tabla se tienen toda la información a los muertes asociadas al Covid.

- Columna 1: id_muerte
    - Descripción: Este campo tiene la infromacion del id de las muertes.
    - Tipo de datos: Int
    - Restricciones de la columna: no aplica
    - Longitud y precisión: 1,1
    - Valores permitidos: solo valores de tipo int
- Columna 2: total_deaths_per_million
    - Descripción: Data el total de muertes por millon

- Tipo de datos: Int
- Restricciones de la columna: no aplica
- Valores permitidos: solo valores de tipo int

- Columna 3: new_deaths_per_million
  - Descripción: : Data los nuevas muertes por millon
  - Tipo de datos: Int
  - Restricciones de la columna: no aplica
  - Valores permitidos: solo valores de tipo int
- Columna 4: new_deaths_smoothed_per_million
  - Descripción: Data el total de los muertes leves por millon
  - Tipo de datos: Int
  - Restricciones de la columna: no aplica
  - Valores permitidos: solo valores de tipo int

Clave primaria: id_muerte

Origen de datos: owid-covid-data.xlsx

Tabla 5: Hechos_Casos

Descripción: En esta tabla se tienen toda la información a los casos y hace refencia a las otras tablas.

- Columna 1: id_caso
  - Descripción: Este campo tiene la infromacion del id de las casos.
  - Tipo de datos: Int

- Restricciones de la columna: no aplica
- Valores permitidos: solo valores de tipo int

- Columna 2: id_pais
  - Descripción: Este campo tiene la infromacion del id del pais.
  - Tipo de datos: int
  - Clave foránea: DIM_Pais(id_pais)
  - Valores permitidos: solo valores enteros
- Columna 3: id_fecha
  - Descripción: Este campo tiene la infromacion del id de la fecha.
  - Tipo de datos: Date
  - Clave foránea: DIM_Fecha(id_fecha)
  - Valores permitidos: solo valores de tipo date
- Columna 4 : location
  - Descripción: Este campo tiene la informacion de la locación del país.
  - Tipo de datos: varchar
  - Restricciones de la columna: no aplica
  - Longitud y precisión: 100
  - Valores permitidos: solo valores varchar
- Columna 5: total_cases
  - Descripción: Data el total de los casos
  - Tipo de datos: Int
  - Restricciones de la columna: no aplica
  - Valores permitidos: solo valores de tipo int
- Columna 6: new_cases
  - Descripción: Data los nuevos casos

- Tipo de datos: Int
- Restricciones de la columna: no aplica
- Valores permitidos: solo valores de tipo int

- Columna 7: new_cases_smoothed
  - Descripción: Data los nuevos leves
  - Tipo de datos: Int
  - Restricciones de la columna: no aplica
  - Valores permitidos: solo valores de tipo int
- Columna 8: total_deaths
  - Descripción: Data el total de muertes
  - Tipo de datos: Int
  - Restricciones de la columna: no aplica
  - Valores permitidos: solo valores de tipo int
- Columna 9: new_deaths
  - Descripción: Data las nuevas muertes
  - Tipo de datos: Int
  - Restricciones de la columna: no aplica
  - Valores permitidos: solo valores de tipo int
- Columna 10: new_deaths_smoothed
  - Descripción: Data las nuevas muertes
  - Tipo de datos: Int
  - Restricciones de la columna: no aplica
  - Valores permitidos: solo valores de tipo int
- Columna 11: id_caso_dim
  - Descripción: Este campo tiene la infromacion del id de la caso.

- Tipo de datos: int
- Clave foránea: DIM_Casos(id_caso)
- Valores permitidos: solo valores de tipo int

- Columna 12: id_muerte_dim
  - Descripción: Este campo tiene la infromacion del id de las muertes.
  - Tipo de datos: Int
  - Restricciones de la columna: no aplica
  - Longitud y precisión: 1,1
  - Valores permitidos: solo valores de tipo int

Clave primaria: id_muerte

Origen de datos: owid-covid-data.xlsx

Desarrollo en Knime

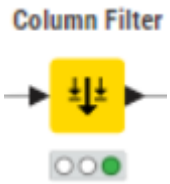
Se desarrollo un ETL en la herramienta Knime la cual consideramos una herramienta fácil y amigable para usarse, en esta sección se explicará lo hecho dentro de la herramienta.

En este caso ya que Knime no permite cierta cantidad de datos se opto por visualizar las estadísticas de los países centroamericanos de lo vivido en la pandemia de Covid 19 vivida de 2020 a finales de 2021 en ella se utilizaron los siguientes componentes:

Excel Reader

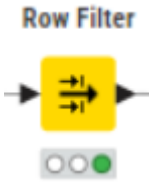
Este componente es utilizado para extraer datos directamente de un Excel, en ella se extrayeron los detalles de los países centroamericanos.

Column Filter



En este componente es utilizado para filtrar columnas por su nombre, esto sirve mucho a la hora de separar datos de una tabla.

Row Filter

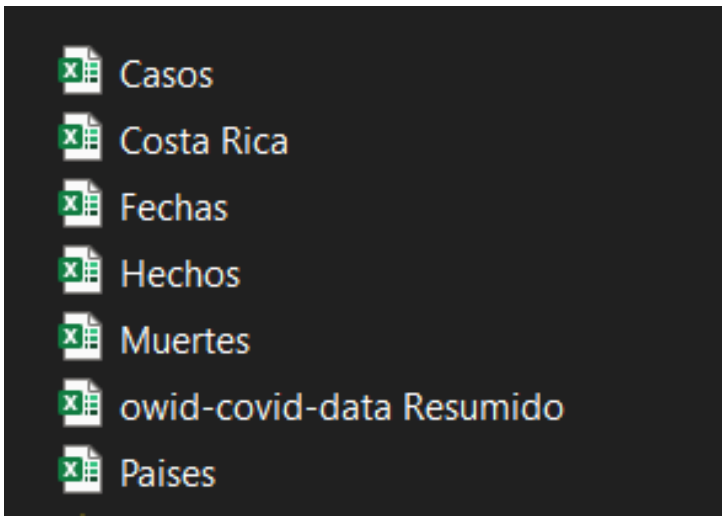


Este componente se utiliza para filtrar las filas ya sea por datos determinados o por números en específicos, es uno de los filtros más útiles a la hora de depurar datos.

Excel Writer

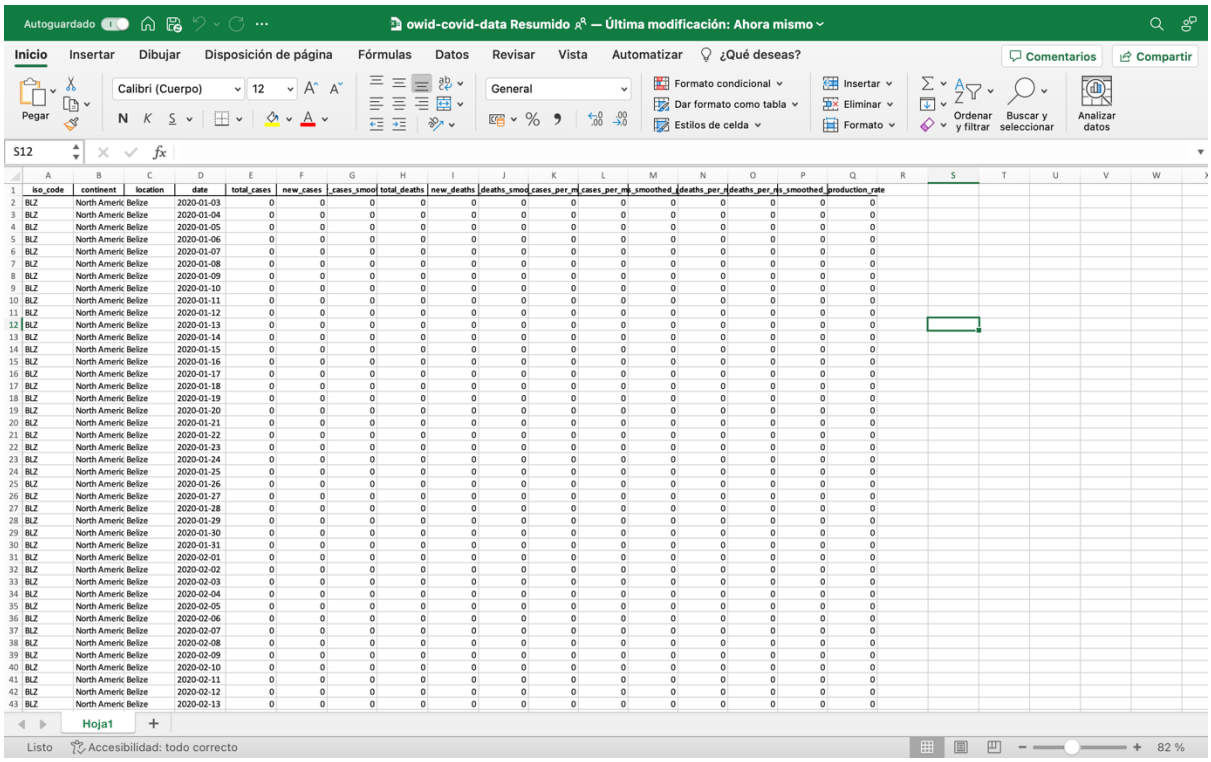
Por último el Excel Writer, este componente es utilizado para crear archivos Excel en base a todos los filtros de columnas y filas ya depurados.

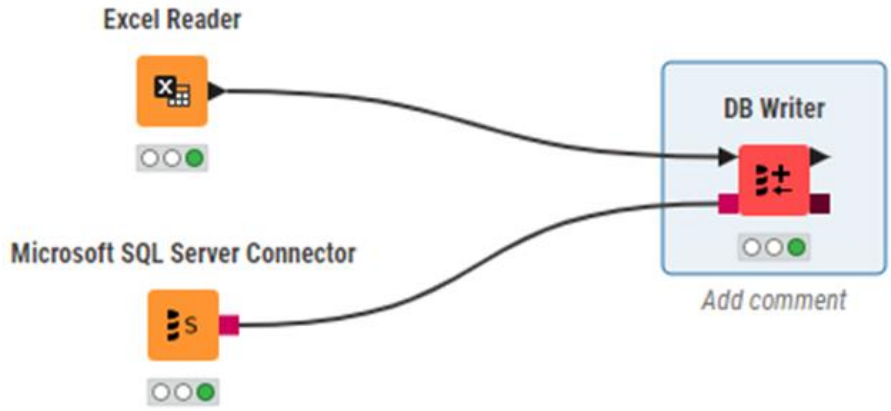
Con estos Excel Writer se crearon los siguientes Excels:



Desarrollo de un ETL

Se desarrollo un ETL con el fin de poner depurar los datos, en el cual se eliminaron todos los países que no fueran de Centroamérica para poder tener un mejor enfoque con los datos.





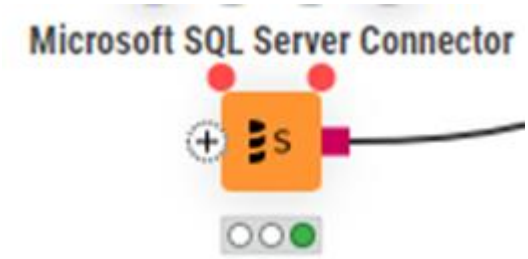
Mediante este circuito de componentes que nos brinda Knime logramos realizar la inserción de datos a una base de datos en sql, la cual llenamos de información acerca

del covid-19, en el ejemplo mostrado observamos el método usado para llenar la tabla Países.

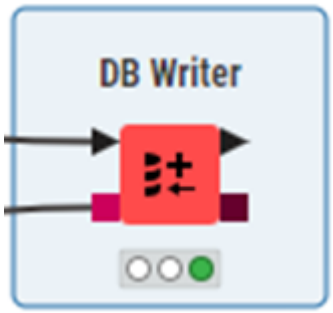
	id_pais	iso_code	continent	location
9528	9528	PAN	North America	Panama
9529	9529	PAN	North America	Panama
9530	9530	PAN	North America	Panama
9531	9531	PAN	North America	Panama
9532	9532	PAN	North America	Panama
9533	9533	PAN	North America	Panama
9534	9534	PAN	North America	Panama
9535	9535	PAN	North America	Panama
9536	9536	PAN	North America	Panama
9537	9537	PAN	North America	Panama
9538	9538	PAN	North America	Panama
9539	9539	PAN	North America	Panama
9540	9540	PAN	North America	Panama
9541	9541	PAN	North America	Panama
9542	9542	PAN	North America	Panama
9543	9543	PAN	North America	Panama
9544	9544	PAN	North America	Panama
9545	9545	PAN	North America	Panama
9546	9546	PAN	North America	Panama
9547	9547	PAN	North America	Panama
9548	9548	PAN	North America	Panama

Aquí podemos observar como la inserción de datos hacia la base de datos se realizo de manera correcta y se insertaron los 9548 datos que habían en este Excel.

Mediante este componente logramos obtener la información del Excel para luego mediante una conexión lograr la inserción de datos asía el sql.

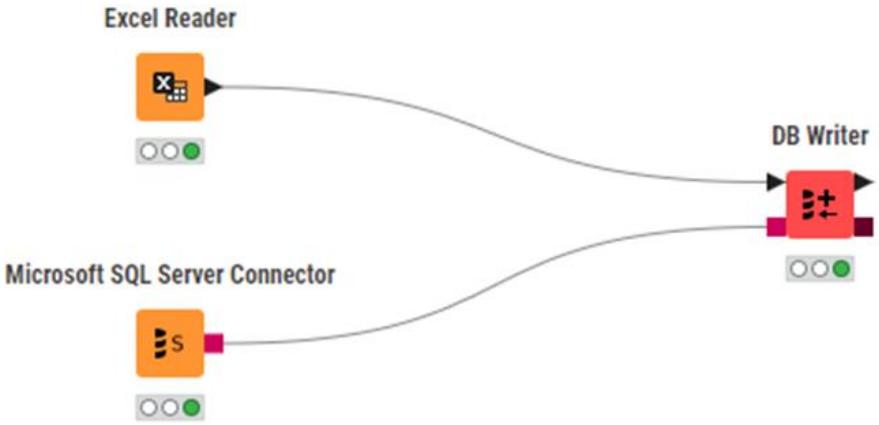


En este componente establecemos una conexión con el sql y definimos a que base de datos deseamos realizar la inserción de la información obtenida del Excel anterior.



Y finalmente mediante la ayuda de este componente definimos que datos del Excel deseamos transferir a la base de datos y así también definimos que tablas vamos a llenar con los datos ya extraídos del Excel.

En este ejemplo podemos observar como la taba DIM_Casos se encuentra vacía por el momento, pero al poner en funcionamiento el siguiente circuito.



Y al ponerlo en funcionamiento se llenará la base de datos de casos.

Results | Messages

	id_caso	total_cases_per_million	new_cases_per_million	new_cases_smoothed_per_million	reproduction_rate
9452	9452	237564	0	0	0.00
9453	9453	237564	0	0	0.00
9454	9454	237564	0	0	0.00
9455	9455	237564	0	0	0.00
9456	9456	237564	0	0	0.00
9457	9457	237564	0	0	0.00
9458	9458	237564	0	0	0.00
9459	9459	237564	0	0	0.00
9460	9460	237564	0	0	0.00
9461	9461	237564	0	0	0.00
9462	9462	237564	0	0	0.00
9463	9463	237564	0	0	0.00
9464	9464	237564	0	0	0.00
9465	9465	237564	0	0	0.00
9466	9466	237564	0	0	0.00

```
CREATE TABLE DIM_Fecha (
  id_fecha int IDENTITY PRIMARY KEY,
  fecha varchar(30) not null
);
```

En la tabla DIM_Fecha, se tuvo que realizar un cambio el cual permitiera realizar la inserción de datos de manera optima, el cambio realizado fue el cambiar el date del id_fecha por un int identity

Para realizar la última inserción en la tabla de hechos, se le tuvieron que realizar unos ajustes a la ultima tabla, el resultado final de dicha tabla es el siguiente

```
CREATE TABLE Hechos_Casos (
  id_caso INT PRIMARY KEY,
  id_pais INT NOT NULL FOREIGN KEY REFERENCES DIM_Pais(id_pais),
  id_fecha int NOT NULL FOREIGN KEY REFERENCES DIM_Fecha(id_fecha),
  Ubicacion VARCHAR(100) NOT NULL,
  total_cases INT NOT NULL,
  new_cases INT NOT NULL,
  new_cases_smoothed INT NOT NULL,
  total_deaths INT NOT NULL,
  new_deaths INT NOT NULL,
  new_deaths_smoothed INT NOT NULL,
  id_muerte_dim INT NOT NULL FOREIGN KEY REFERENCES DIM_Muertes(id_muerte)
);
```

Así quedan las tablas llenas de valores

100 %

Results | Messages

	id_muerte	total_deaths_per_million	new_deaths_per_million	new_deaths_smoothed_per_million
1	1	2	2	0
2	2	2	0	0
3	3	2	0	0
4	4	2	0	0
5	5	2	0	0
6	6	4	2	0

	id_caso	total_cases_per_million	new_cases_per_million	new_cases_smoothed_per_million	reproduction_rate
1	1	2	2	0	0.00
2	2	2	0	0	0.00
3	3	4	2	0	0.00
4	4	4	0	0	0.00
5	5	4	0	0	0.00

	id_pais	iso_code	continent	location
1	1	BLZ	North America	Belize
2	2	BLZ	North America	Belize
3	3	BLZ	North America	Belize
4	4	BLZ	North America	Belize

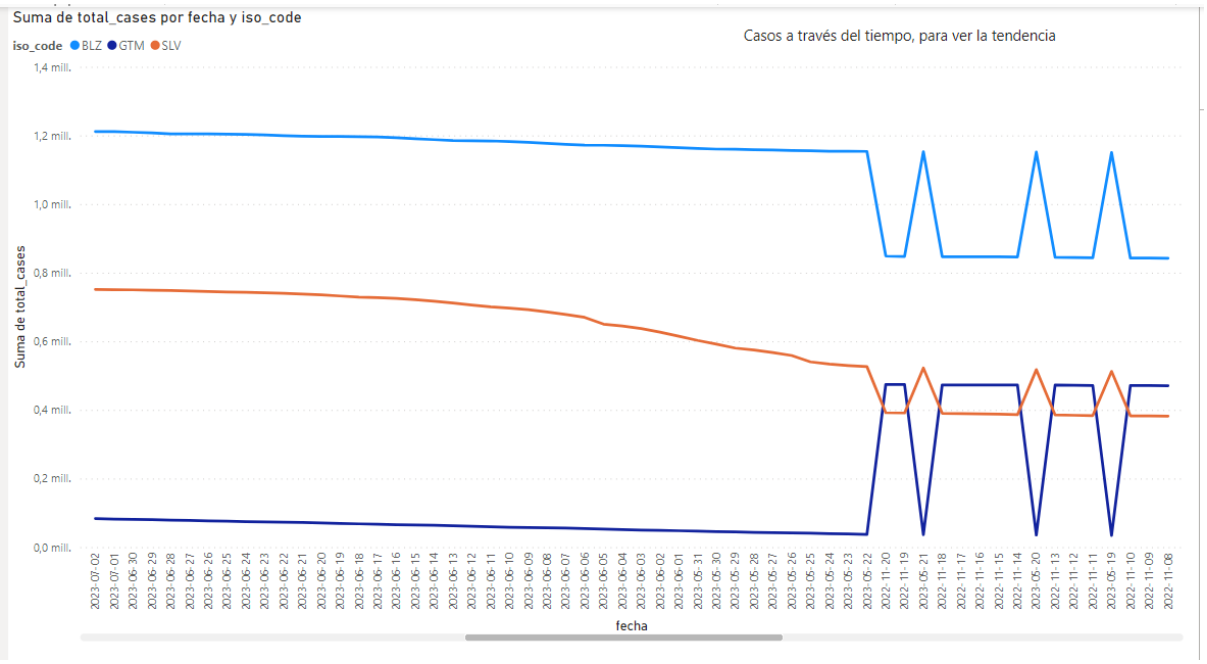
	id_fecha	fecha
1	1	2020-03-25
2	2	2020-03-26
3	3	2020-03-27
4	4	2020-03-28
5	5	2020-03-29

	id_casos	id_pais	id_fecha	Ubicacion	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	id_muerte_dim	id_caso_dim
1	1072256	1	1	Belize	7	2	0	1	1	0	1	1
2	1072257	2	2	Belize	388	32	34	3	1	0	2	2
3	1072258	3	3	Belize	475	23	42	4	1	0	3	3
4	1072259	4	4	Belize	668	20	30	6	1	0	4	4
5	1072260	5	5	Belize	713	27	34	10	4	0	5	5
6	1072261	6	6	Belize	760	30	29	11	1	0	6	6
7	1072262	7	7	Belize	818	58	30	12	1	1	7	7
8	1072263	8	8	Belize	964	94	42	13	1	1	8	8

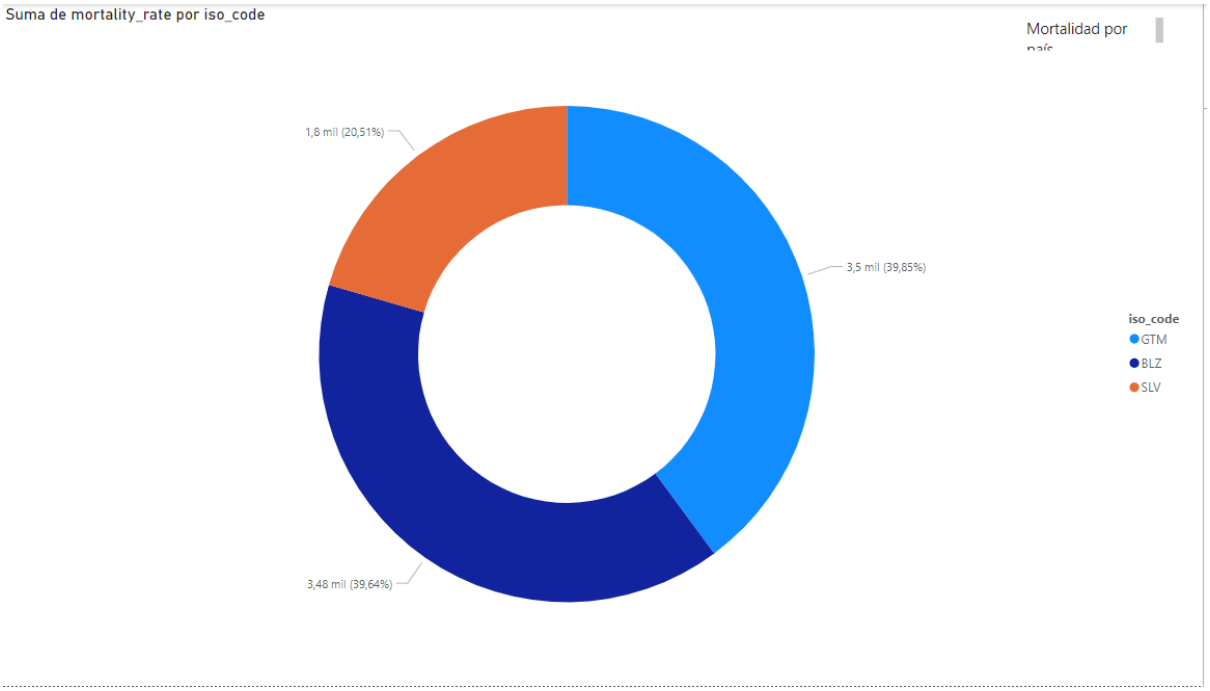
Dashboard Dinámico

Usando toda la información recopilada vamos a realizar un proceso de análisis de la información y lo vamos a mostrar y aplicar mediante la herramienta de Power BI que nos permite hacer DashBoard Dinámicos, para ejemplificar bien los datos que se trabajaron. Las consultas se manejaron mediante procedimientos almacenados en el SQL Server y se muestran a continuación de forma gráfica.

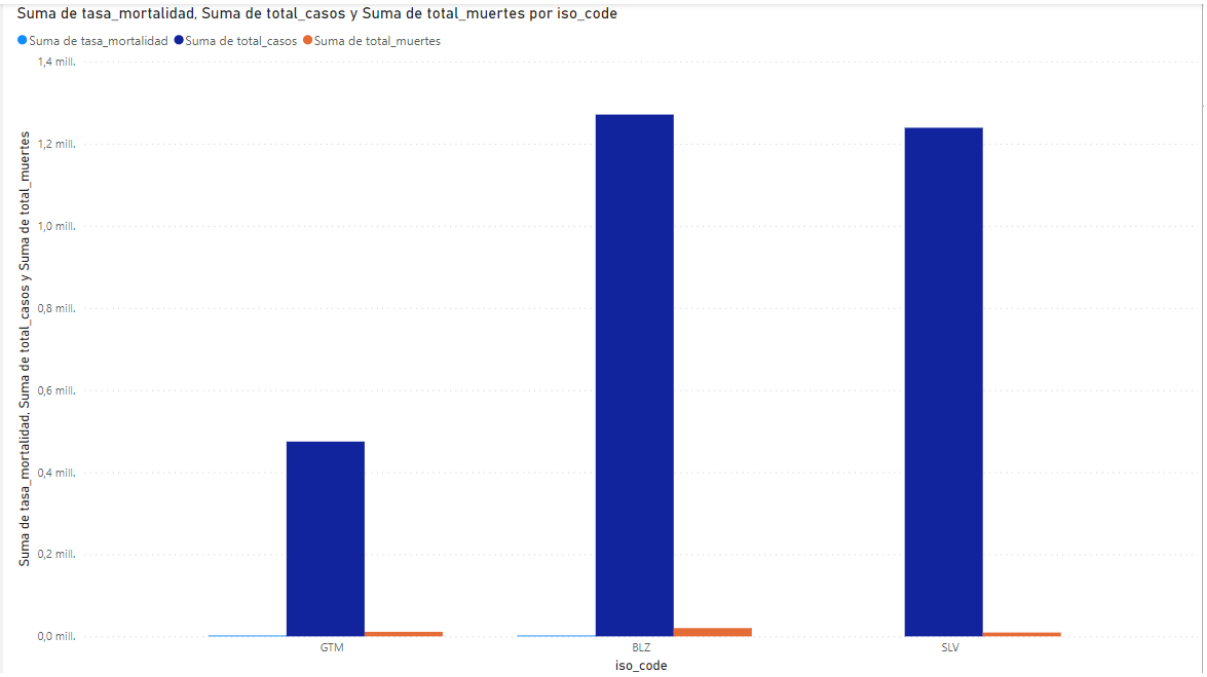
Tendencia de los casos a través del tiempo



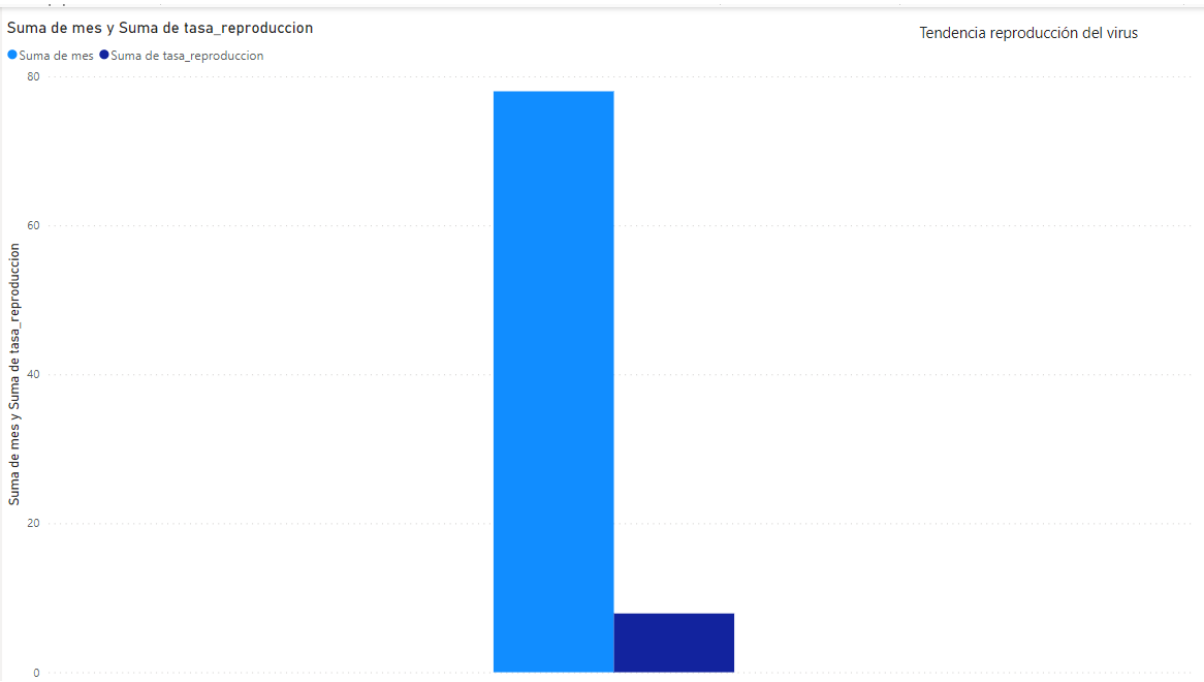
Mortalidad por país



Muerte por país



Tendencia a reproducción del virus en un mes

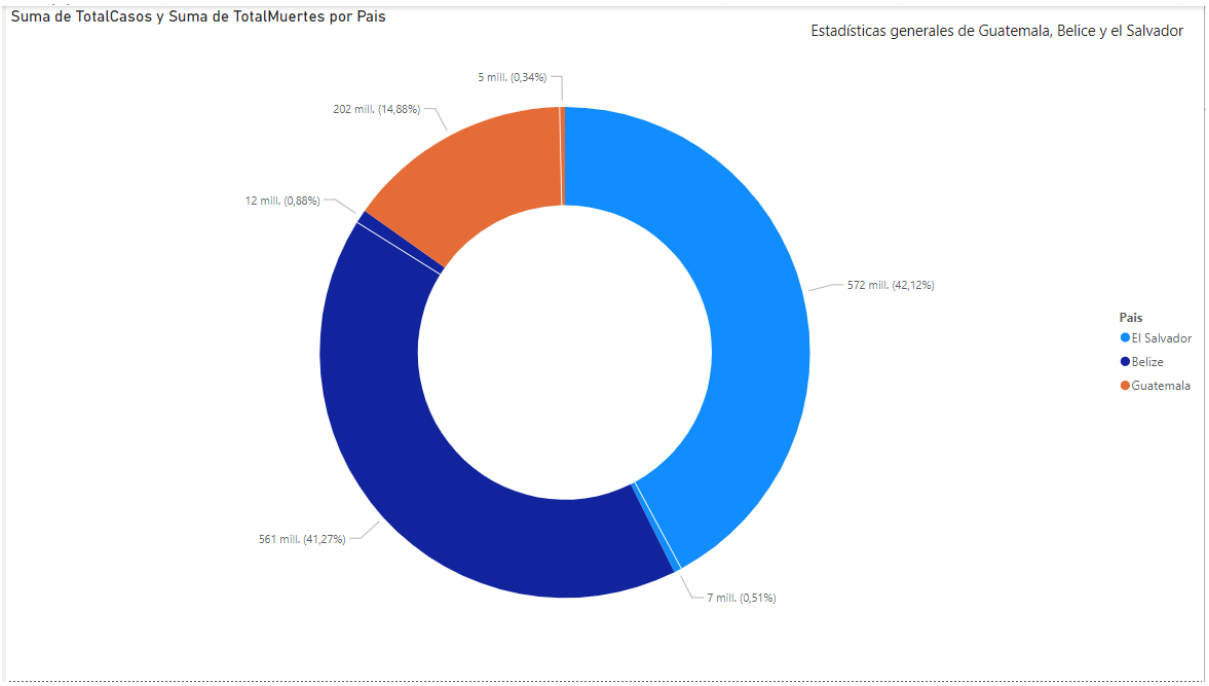


Tendencias globales por país

Fecha	Suma de NuevosCasos	Suma de TotalCasos	Suma de TotalMuertes
2020-03-12	1143	903529	8323
2020-03-13	826	904355	8325
2020-03-14	2080	906435	8329
2020-03-15	2055	909602	8355
2020-03-16	2185	911787	8360
2020-03-17	1899	913686	8367
2020-03-18	1659	915345	8376
2020-03-19	1198	916628	8378
2020-03-20	2486	1093099	10609
2020-03-21	2426	1095525	10626
2020-03-22	2877	1098496	10651
2020-03-23	2844	1101408	10673
2020-03-24	2362	1103770	10696
2020-03-25	1980	1105755	10720
2020-03-26	2288	1108392	10746
2020-03-27	2658	1111114	10770
2020-03-28	10771	1125329	10800
2020-03-29	7501	1132984	10840
2020-03-30	7065	1140066	10872
2020-03-31	7143	1147209	10900
2020-04-01	7599	1154945	10928
2020-04-02	7100	1162288	10962
2020-04-03	6018	1168415	10995
2020-04-04	4107	1172576	11020
2020-04-05	3415	1175991	11051
2020-04-06	2560	1178719	11074
2020-04-07	1944	1180796	11098
2020-04-08	1730	1182568	11118
2020-04-09	1493	1184113	11139
2020-04-10	1307	1185420	11155
2020-04-11	1341	1187430	11179
2020-04-12	2167	1190435	11202
2020-04-13	4029	1194498	11229
Total	**4172152**	**1335335378**	**23538393**

Tendencias globales

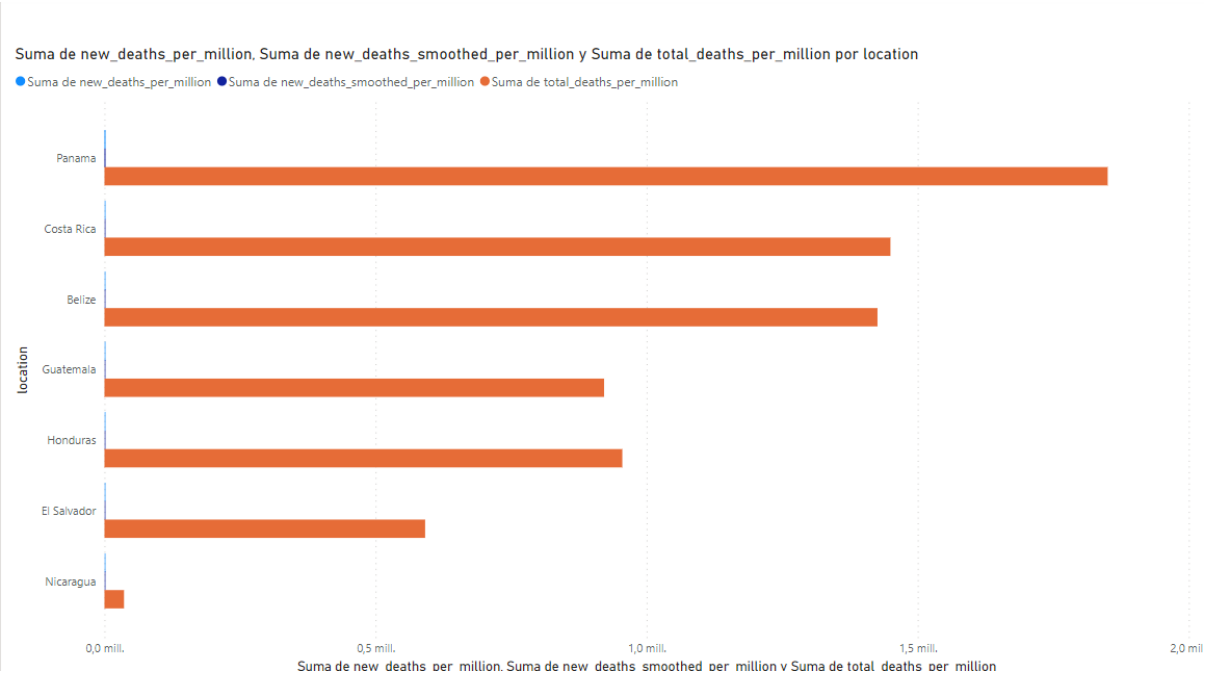
Estadísticas generales de Guatemala, El Salvador y Belize



Casos en Costa Rica

location	Suma de new_cases	Suma de total_cases	Suma de total_deaths
Costa Rica	1238884	828079125	7510469
Total	**1238884**	**828079125**	**7510469**

Suma de muertes por país

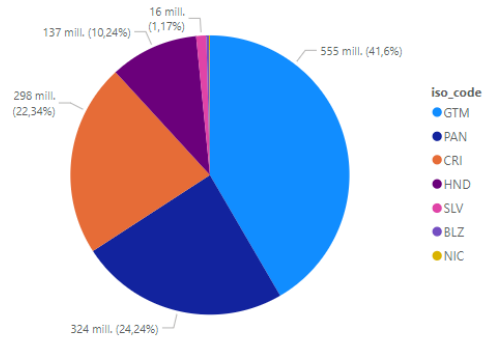


Tasa de reproducción del virus por país

Suma de reproduction_rate por Ubicacion y Ubicacion

Recuento total de casos por país y suma de casos nuevos

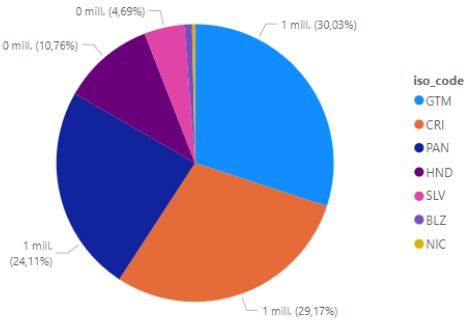
Suma de total_cases y Recuento de total_cases por iso_code



iso_code	Recuento de total_cases
BLZ	250
CRI	721
GTM	1014
HND	667
NIC	115
PAN	825
SLV	332
Total	**3924**

location	iso_code
Belize	BLZ
Costa Rica	CRI
El Salvador	SLV
Guatemala	GTM
Honduras	HND
Nicaragua	NIC
Panama	PAN

Suma de new_cases por iso_code



iso_code	Suma de new_cases
GTM	1253006
CRI	1216847
PAN	1006025
HND	449036
SLV	195654
BLZ	37006
NIC	14578
Total	**4172152**

Suma total de casos por país

Conclusiones:

1. El análisis retrospectivo de los datos de COVID-19 utilizando técnicas de data warehouse y minería de datos permitió identificar patrones y tendencias claves durante el apogeo de la pandemia, proporcionando información valiosa para la toma de decisiones.
2. Se logró consolidar y analizar eficientemente una gran cantidad de datos relacionados con el COVID-19, superando las limitaciones de desempeño de bases de datos transaccionales.
3. Los tableros de control dinámicos desarrollados facilitan la visualización y el análisis continuo de métricas críticas como casos acumulados, mortalidad relativa, estacionalidad del virus, entre otros hallazgos.

Recomendaciones:

1. Ampliar el data warehouse para incorporar otras fuentes de información como datos de vacunación, estudios clínicos, movilidad de la población, que permitan un análisis aún más profundo.
2. Desarrollar data marts específicos por temas como efectividad de vacunas, tratamientos médicos, impactos económicos, etc., para soportar análisis especializados.
3. Incorporar capacidades avanzadas de inteligencia artificial como machine learning y procesamiento de lenguaje natural para potenciar la toma de decisiones informada por los datos.

Referencias Bibliográficas

1. Big Data preprocessing – knime. (s.f.). KNIME Community Hub. https://hub.knime.com/knime/spaces/Examples/10_Big_Data/01_Big_Data_Connectors/01_Big_Data_Preprocessing_Example~P69kx252fdUB-aFR/most-recent
2. DB reference row filter – tommy. (s.f.). KNIME Community Hub. https://hub.knime.com/tommy/spaces/Public/DB%20Reference%20Row%20Filter~VZmRcQMf4qUUZFn-/most-recent
3. ExtracciÃ³n, transformaciÃ³n y carga de datos (ETL) - Azure Architecture Center. (s.f.). Microsoft Learn: Build skills that open doors in your career. https://learn.microsoft.com/es-es/azure/architecture/data-guide/relational-data/etl
4. In-Database processing on SQL server – knime. (s.f.). KNIME Community Hub. https://hub.knime.com/knime/spaces/Examples/40_Partners/01_Microsoft/01_SQL_Server_InDB_Processing(Azure)~44JVorS4AKoeiFHQ/current-state
5. Las diez formas principales de limpiar los datos - Soporte técnico de Microsoft. (s.f.). Microsoft Support. https://support.microsoft.com/es-es/office/las-diez-formas-principales-de-limpiar-los-datos-2844b620-677c-47a7-ac3e-c2e157d1db19
6. Open for Innovation | KNIME. (s.f.). Open for Innovation | KNIME. https://www.knime.com/
7. ¿Qué es ETL? - Explicación de extracción, transformación y carga (ETL) - AWS. (s.f.). Amazon Web Services, Inc. https://aws.amazon.com/es/what-is/etl/#:~:text=El%20proceso%20ETL%20funciona%20en,base%20de%20datos%20de%20destino.
8. Search | kaggle. (s.f.). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/search?q=covid
9. Tutorial: Connect to on-premises data in SQL server - power BI. (s.f.). Microsoft Learn: Build skills that open doors in your career. https://learn.microsoft.com/en-us/power-bi/connect-data/service-gateway-sql-tutorial