

Manual técnico

[SS2] Práctica 1

Kevin Martinez

202004816 | USAC

INTRODUCCIÓN

Este Manual Técnico proporciona una guía concisa sobre el Proceso de Extracción, Transformación y Carga (ETL) de datos, así como una descripción detallada del modelo utilizado. Su objetivo es facilitar la comprensión del proceso de ETL y del modelo de datos entre los desarrolladores y equipos técnicos. A lo largo del manual, se presentarán los fundamentos del ETL, se describirá el modelo de datos, incluyendo las tablas de hechos y dimensiones, y se ofrecerán recomendaciones para garantizar el éxito del proyecto. Este manual se presenta como una herramienta esencial para todos los involucrados en el desarrollo e implementación del proceso de ETL.

PROCESO ETL

Extracción:

- La primera etapa del proceso de ETL es la extracción de datos desde la fuente de origen, que en este caso fue un archivo CSV.
- Los datos del archivo CSV se extraen y se cargan en la tabla Temporal del modelo de datos. Esto se logró ejecutando un script SQL para insertar los datos en la tabla Temporal de la base de datos SQL Server.

Transformación:

- Una vez que los datos se han cargado en la tabla Temporal, comienza la etapa de transformación. Durante esta etapa, los datos se limpian, se estructuran y se preparan para su inserción en las tablas de hechos y dimensiones.
- Se seleccionan y se transforman los datos relevantes de la tabla Temporal para insertarlos en las tablas Time, Location y Tsunami.
- Los datos se insertan en las tablas de hechos y dimensiones correspondientes utilizando consultas SQL que toman los datos de la tabla Temporal y los insertan en las tablas finales.

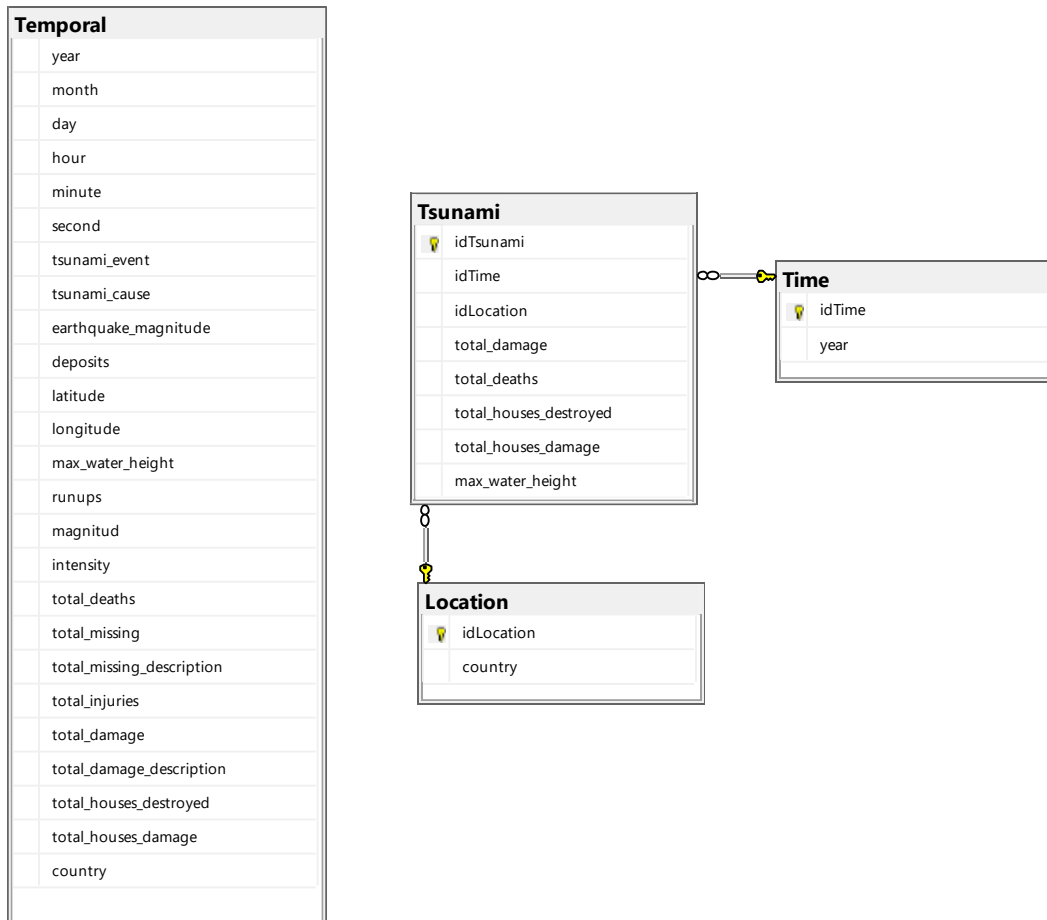
Carga:

- Una vez que los datos se han transformado adecuadamente, comienza la etapa de carga. Durante esta etapa, los datos transformados se cargan en las tablas de hechos y dimensiones del modelo de datos.
- Se realizan inserciones masivas de datos desde la tabla Temporal a las tablas Time, Location y Tsunami utilizando consultas SQL que seleccionan los datos transformados y los insertan en las tablas finales.
- Se asegura la integridad referencial de los datos mediante la verificación de las claves primarias y foráneas durante la inserción en las tablas de hechos y dimensiones.

En resumen, el proceso de ETL comienza con la extracción de datos desde la fuente de origen, seguido de la transformación de los datos en la tabla Temporal y, finalmente, la carga de los datos transformados en las tablas de hechos y dimensiones del modelo de datos. Este proceso garantiza que los datos se estructuren de manera eficiente y estén disponibles para su análisis y consulta posterior.

DESCRIPCIÓN DE MODELO

Modelo generado:



Temporal:

- Esta tabla actúa como un área de preparación para la carga masiva inicial de datos. Contiene columnas que representan las diferentes características de un evento de tsunami, como año, mes, día, hora, minuto, segundo, evento de tsunami, causa del tsunami, magnitud del terremoto asociado, depósitos, latitud, longitud, altura máxima del agua, etc.
- Es una tabla temporal que se utiliza durante el proceso de transformación de datos antes de dividir la información en las tablas finales.

Time:

- Esta tabla almacena información relacionada con la dimensión del tiempo. Contiene una columna idTime como clave primaria y una columna year que representa el año en que ocurrió el evento de tsunami.

- Se utiliza para realizar análisis temporales y segmentar los datos por año.

Location:

- Esta tabla almacena información relacionada con la dimensión de la ubicación. Contiene una columna idLocation como clave primaria y una columna country que representa el país donde ocurrió el evento de tsunami.
- Se utiliza para realizar análisis geoespaciales y segmentar los datos por país.

Tsunami:

- Esta tabla representa la tabla de hechos principal en el modelo dimensional. Contiene columnas que representan las métricas y medidas relacionadas con un evento de tsunami, como el total de daños, el total de muertes, el total de casas destruidas, la altura máxima del agua, etc.
- También contiene dos claves foráneas idTime e idLocation que hacen referencia a las tablas Time y Location, respectivamente. Esto permite relacionar cada evento de tsunami con su ubicación y tiempo correspondientes.

CONCLUSION

La práctica realizada se enfoca en implementar un proceso de ETL para analizar eventos de tsunamis. Utilizando Python y SQL Server, se diseñó un modelo de datos con tablas de hechos y dimensiones. Se extraen datos de archivos proporcionados, se transforman para ajustarlos al modelo y luego se cargan en las tablas correspondientes. La tabla "Tsunami" captura detalles específicos de cada evento, mientras que las dimensiones "Location" y "Time" brindan contextos geográficos y temporales. Este enfoque no solo destaca la importancia del ETL y el modelado de datos, sino que también ilustra su aplicación práctica en la gestión y análisis de eventos naturales, como los tsunamis, facilitando la toma de decisiones en áreas como la gestión de riesgos y la planificación de respuestas ante desastres.