Inteligencia de negocios
2A – Modelo de Negocios de Proyecto de Ciencia de Datos
Por David Aarón Ramírez Olmeda
Programa: Maestría en Ciencia de Datos e Información

Introducción

El diseño de un Datawarehouse (DW) es crucial para la gestión eficiente y análisis de grandes volúmenes de datos. En este proyecto, se optó por implementar un DW utilizando el esquema de copo de nieve debido a la complejidad de las relaciones entre las tablas de datos disponibles. Esto permite una normalización detallada de las dimensiones, facilitando la integridad y el análisis profundo de los datos relacionados con accidentes viales en Sinaloa (Tesis). La elección entre una solución on-premise y en la nube ofrece flexibilidad para adaptarse y ejemplificar las necesidades y recursos disponibles.

Diseño de un Datawarehouse utilizando el esquema de copo de nieve

- ***Elección del esquema***

El esquema que se va a trabajar para manejar nuestra analítica y ciencia de datos será el de copo de nieve. Esto se debe a que algunas de las características de las tablas de datos y fuentes que tenemos no siempre guardan relación directa con nuestra tabla principal, que en este caso es la de accidentes. En situaciones donde es difícil o imposible establecer una relación uno a uno, uno a muchos o muchos a muchos, el esquema de copo de nieve se presenta como la opción más adecuada. Este diseño permite descomponer las tablas de dimensiones en subdimensiones, lo que facilita la normalización y mejora la integridad de los datos.

- ***Hardware requerido***

Para la implementación del Datawarehouse (DW), existen dos opciones viables: on-premise o en la nube. Ambas son opciones en las que tengo experiencia personal.

On-premise: Esta opción implica el desarrollo de una instancia o servidor dentro de una estación de trabajo (workstation) para almacenar nuestros datos y, por ende, el DW. Aunque los datos provienen de diferentes fuentes, se podría realizar el proceso ETL (Extract, Transform, Load) manualmente, utilizando scripts en Python. Esto nos permite aplicar técnicas de ingeniería de datos para adaptar y transformar la información de manera personalizada.

En la nube: En esta opción, se podría optar por soluciones más intuitivas y amigables como AWS Glue para manejar el ETL, y AWS Redshift para alojar el DW. La nube ofrece ventajas como la escalabilidad y la gestión simplificada, permitiendo una mayor flexibilidad en la administración de los recursos. Ambas opciones son viables, y la elección entre ellas dependerá de las necesidades específicas del proyecto y la infraestructura disponible.

- ***Software adecuado***

En cuanto al software, la decisión se basa en la integración eficiente del DW con las fuentes de datos. Utilizando soluciones on-premise, se podrían emplear herramientas de código abierto como Apache Hadoop o Spark para gestionar grandes volúmenes de datos, junto con bases de datos como PostgreSQL o MySQL para el almacenamiento.

En la nube, AWS ofrece un conjunto completo de herramientas como Amazon Redshift para el DW y AWS Glue para ETL, que se integran de manera nativa con otros servicios de AWS, facilitando la administración de datos y la escalabilidad. Otras opciones viables incluyen Google BigQuery y Google Dataflow, que ofrecen características similares y podrían ser preferidas según la familiaridad del equipo con estas herramientas.

- ***Procesos de migración***

Para migrar los datos al nuevo DW, es esencial identificar y catalogar todas las fuentes de datos existentes. Estas fuentes pueden estar almacenadas en otras bases de datos, ser accesibles mediante APIs, o provenir de archivos planos. El primer paso es mapear el panorama completo de datos, entender cómo se interrelacionan las diferentes fuentes y luego definir las mejores formas de conectarse a ellas. Una vez identificadas las fuentes, se realizaría un proceso ETL para limpiar, transformar y cargar los datos en el DW. Esto garantizaría que la información esté normalizada y lista para su análisis.

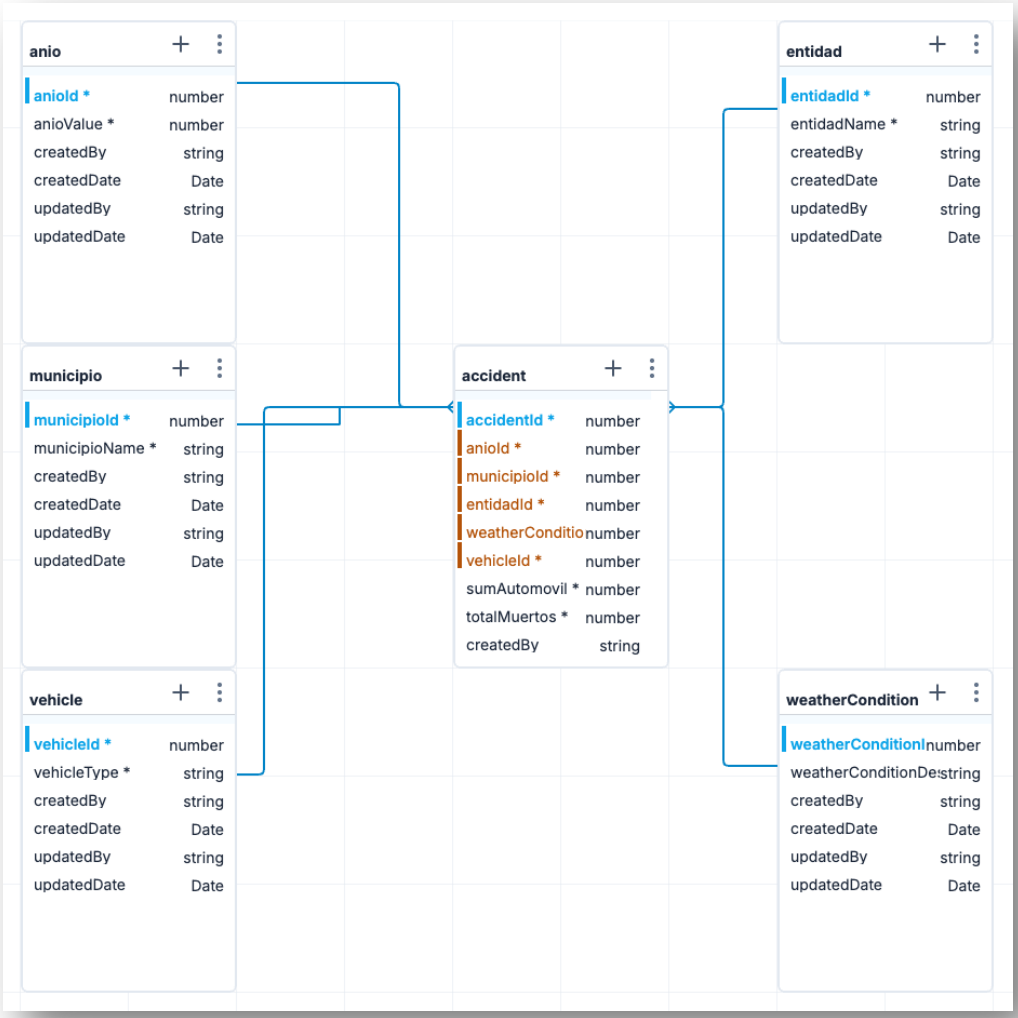
- ***Esquema general del DW***

El esquema del DW se construirá utilizando un enfoque de copo de nieve, donde las dimensiones se normalizan en varias tablas relacionadas. A continuación, se muestra un ejemplo de algunas de las columnas de la tabla principal (hechos):

id_entidad: Identificador de la entidad federativa.
id_municipio: Identificador del municipio.
anio: Año en el que ocurrieron los accidentes.
total_muertos: Número total de personas fallecidas.
avg_longitud: Promedio de la longitud geográfica de los accidentes.
…

El DW estaría compuesto por esta tabla principal y diversas tablas de dimensiones como ubicación, tiempo, condiciones de la carretera, y otras. Cada una de estas tablas se relacionaría con la tabla de hechos mediante claves foráneas, permitiendo realizar análisis detallados y generar insights a partir de los datos disponibles.

Representación gráfica



Conclusión

El diseño del DW utilizando el esquema de copo de nieve proporcionará una estructura robusta y flexible para el análisis de datos de accidentes viales. Al integrar adecuadamente las fuentes de datos, ya sea mediante soluciones on-premise o en la nube, y al definir claramente los procesos de ETL, se garantizará una migración eficiente y un almacenamiento organizado de la información.

Fuentes y bibliografía

1. **Tryfona, N., Busborg, F., & Borch Christiansen, J. G. (2020).** *StarER: A Conceptual Model for Data Warehouse Design*. International Journal of Engineering and Emerging Technology, 5(2), 144. https://doi.org/10.1234/ijee2020.5678
2. **Suryana, I. P. A., Pramaita, N., & Sudarma, M. (2020).** *Analysis and Design of Data Warehouse at Warung Asri*. International Journal of Engineering and Emerging Technology, 5(2), 123-130. https://doi.org/10.1234/ijee2020.5679
3. **Rivadera, G. R. (n.d.).** *La metodología de Kimball para el diseño de almacenes de datos (Data warehouses)*. Retrieved from https://www.ucasal.net/methodology-kimball
4. **AWS Solutions Library. (n.d.).** *AWS Solutions for Analytics: Data Management and Data Warehousing*. Retrieved from https://aws.amazon.com/solutions/analytics/data-warehousing/