Logotipo

Descripción generada automáticamenteLogotipo, nombre de la empresa

Descripción generada automáticamente

UNIVERSIDAD TECNOLÓGICA DEL VALLE DE TOLUCA

Dirección de Carrera de Tecnologías de la Información y Comunicación

Ingeniería en Desarrollo y Gestión de Software

Nombre la asignatura:

EXTRACCIÓN DEL CONOCIMIENTO DE BASE DE DATOS

Nombre de la profesora:

NELLY GUTIERREZ

“SEGUNDO ENTREGABLE”

CYNTHIA PAOLA DURO SANCHEZ 221811737

CAROLINA DÍAZ ROMERO 221811725

FRANCISCA CAMPOS QUIÑONES 221811678

BRENDA XIMENA DURO SÁNCHEZ 221712243

Grupo

IDGS – 91

Cuatrimestre

9o

Lugar

Santa María Atarasquillo, Lerma, México.

MAYO -AGOSTO 2022

**INDICE**

[**ESQUEMA DATA WARE HOUSE** 3](#_gjdgxs)

[**TIPOS Y FUENTES DE DATOS** 4](#_30j0zll)

[**TÉCNICAS DE LIMPIEZA DE DATOS** 5](#_1fob9te)

[**DATA CLEANING** 5](#_3znysh7)

# **ESQUEMA DATA WARE HOUSE**

Dentro del Esquema DWH de estrella se encuentra una tabla de hechos que contiene los datos para el análisis, rodeada de las tablas de dimensiones.

Las tablas de dimensiones tendrán siempre una clave primaria simple, mientras que en la tabla de hechos, la clave principal estará compuesta por las claves principales de las tablas dimensionales.

Diagrama

Descripción generada automáticamente

Este esquema nos ayuda a separar los datos del proceso de negocios en: hechos y dimensiones. Los hechos contienen datos medibles, cuantitativos, relacionados con la transacción del negocio, y las dimensiones son atributos que describen los datos indicados en los hechos

# **TIPOS Y FUENTES DE DATOS**

Tipos:

* Biométricos
* Maquina a Maquina
* Datos de transacciones
* Generados por los humanos
* Web y medios sociales

Web y medios sociales para el desarrollo de sistema de ventas:

Este tipo de fuente de datos nos beneficia ya que Son los que se originan en la red y configuran, según los expertos, el trozo más grande del pastel llamado Big Data y es una de las fuentes de datos más utilizadas en la actualidad. Hablamos de la información que se genera sobre clics en vínculos y elementos.

Pero también de toda aquella contenida en las búsquedas que realizamos por ejemplo en Google, las publicaciones en las Redes sociales (Twitter, Facebook, etc.) y el contenido web como páginas, enlaces o imágenes.

En este caso ya que Actualmente a través de Internet puede accederse a consultar fuentes de información imprescindibles en cualquier investigación.

Entre las primeras fuentes de información disponibles en Internet se encuentran los catálogos de las grandes bibliotecas, a estas se han unido importantes bibliografías, directorios e instituciones.

Además de los registros bibliográficos, en las bibliotecas virtuales se pueden consultar y leer textos completos. Para facilitar a nuestros usuarios sus consultas en Internet se han seleccionado diversos recursos que se han agrupado por tipología y materias.

Lo que nos ha ayudado como fuentes de información principal para el desarrollo de nuestro sistema de ventas.

# **TÉCNICAS DE LIMPIEZA DE DATOS**

La limpieza de datos ha hecho que la dependencia de la información de datos sea manejable al mantener la calidad de los datos y mantener la integridad como una prioridad principal para las empresas El proceso de limpieza de datos puede ser complejo si tiene diferentes conjuntos de datos provenientes de fuentes dispares. Tener una estrategia de limpieza de datos eficiente mantiene la integridad de los datos durante un proyecto de limpieza de datos.

Limpieza de datos, también conocida como depuración de datos o limpieza, es el primer paso en la preparación de datos. Implica identificar errores en un conjunto de datos y corregirlos para garantizar que solo se transfieran datos limpios y de alta calidad a los sistemas de destino lo que nos ayuda con el manejo más fácil y rápido de información.

## **DATA CLEANING**

**ETL**

Los procesos ETL son un término estándar que se utiliza para referirse al movimiento y transformación de datos. Se trata del proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y cargarlos en otra base de datos (denominada data mart o data warehouse) con el objeto de analizarlos. También pueden ser enviados a otro sistema operacional para apoyar un proceso de negocio.

Fases de un proceso ETL

* Las distintas fases o secuencias de un proceso ETL son las siguientes:
* Extracción de los datos desde uno o varios sistemas fuente.
* Transformación de dichos datos, es decir, posibilidad de reformatear y limpiar estos datos cuando sea necesario.
* Carga de dichos datos en otro lugar o base de datos, un data mart o un data warehouse, con el objeto de analizarlos o apoyar un proceso de negocio.

**La limpieza de datos como etapa separada de los procesos ETL**

Aunque podría entenderse como una acción integrada en la fase de transformación de datos, en la actualidad la tendencia es considerar la limpieza de datos como una fase separada del proceso ETL.

Esta visión corresponde a una concepción más moderna y práctica del proceso. Para ahorrar tiempo y ganar en efectividad es conveniente unificar criterios, por ejemplo, introduciendo “av” en vez de “avenida” en todos los registros de una base de datos de direcciones postales, ANTES de empezar el proceso ETL propiamente dicho.

Tan importante es tener la información consolidada como que todos los datos sean correctos y con una visión única para todos los usuarios. Solo así se pueden lograr unos circuitos de trabajo y análisis de dichos datos realmente óptimos y efectivos.

**Pasos que se realizan para la limpieza de datos:**

Paso 1: elimine las observaciones duplicadas o irrelevantes

Elimine las observaciones no deseadas de su conjunto de datos, incluidas las observaciones duplicadas o las observaciones irrelevantes. Las observaciones duplicadas ocurrirán con mayor frecuencia durante la recopilación de datos. Cuando combina conjuntos de datos de varios lugares, extrae datos o recibe datos de clientes o varios departamentos, existen oportunidades para crear datos duplicados. La reduplicación es una de las áreas más importantes a considerar en este proceso. Las observaciones irrelevantes son cuando notas observaciones que no encajan en el problema específico que estás tratando de analizar. Por ejemplo, si desea analizar datos sobre clientes de la generación del milenio, pero su conjunto de datos incluye generaciones anteriores, puede eliminar esas observaciones irrelevantes.

**Paso 2: corregir errores estructurales**

Los errores estructurales ocurren cuando mide o transfiere datos y observa convenciones de nomenclatura extrañas, errores tipográficos o mayúsculas incorrectas. Estas inconsistencias pueden causar categorías o clases mal etiquetadas. Por ejemplo, puede encontrar que aparecen "N/A" y "No aplicable", pero deben analizarse como la misma categoría.

**Paso 3: filtre los valores atípicos no deseados**

A menudo, habrá observaciones únicas que, de un vistazo, no parecen encajar en los datos que está analizando. Si tiene una razón legítima para eliminar un valor atípico, como una entrada de datos incorrecta, hacerlo ayudará al rendimiento de los datos con los que está trabajando. Sin embargo, a veces es la aparición de un valor atípico lo que probará una teoría en la que está trabajando. Recuerde: el hecho de que exista un valor atípico no significa que sea incorrecto. Este paso es necesario para determinar la validez de ese número. Si un valor atípico resulta ser irrelevante para el análisis o es un error, considere eliminarlo.

**Paso 4: Manejar los datos que faltan**

* No puede ignorar los datos faltantes porque muchos algoritmos no aceptarán valores faltantes. Hay un par de maneras de lidiar con los datos que faltan. Ninguno es óptimo, pero ambos pueden ser considerados.
* Como primera opción, puede eliminar las observaciones a las que les faltan valores, pero al hacerlo eliminará o perderá información, así que tenga esto en cuenta antes de eliminarlo.
* Como segunda opción, puede ingresar valores faltantes basados ​​en otras observaciones; nuevamente, existe la posibilidad de perder la integridad de los datos porque puede estar operando a partir de suposiciones y no de observaciones reales.
* Como tercera opción, puede modificar la forma en que se usan los datos para navegar de manera efectiva por los valores nulos.