

Almacenes de Datos

Elaboró:

Dra. Anilu Franco Arcega

Dr. Virgilio López Morales

Dr. Joel Suárez Cansino

Fecha de elaboración: Mayo/2023

Tema

Almacenes de Datos

Resumen

En este documento se presentan los conceptos teóricos necesarios para comprender la importancia del uso de almacenes de datos en instituciones que desean obtener conocimiento a partir de los datos que generan en sus operaciones diarias.

Palabras Clave: Almacén de datos, arquitectura del almacén de datos.

Subject

Data Warehouse

Abstract

This document presents the theoretical concepts necessary to understand the importance of the use of data warehouses in institutions that want to obtain knowledge from data that they generate in their daily operations.

Keywords: Data warehouse and architecture.

Objetivo de Aprendizaje

Comprender el uso de almacenes de datos, como una herramienta para la preparación y almacenamiento eficiente de la información que genera en su día a día alguna institución, ya sea por el manejo de personas, productos o servicios.

Almacenes de datos

- Un almacén de datos es una **gran colección de datos** que **recoge información** de múltiples sistemas y cuya actividad principal se centra en la **Toma de Decisiones** en vez de en su captura.
 - Los sistemas que proveen de información al almacén son en su mayoría sistemas de tipo OLTP (On-Line Transaction Processing) que utiliza la institución para recoger información del exterior.

- Una vez reunidos los datos de los sistemas fuentes, se guardan durante mucho tiempo, lo que permite el acceso a **datos históricos**. El tamaño del histórico lo determina la institución en dependencia de sus necesidades.
- Los almacenes de datos proporcionan al usuario una **interfaz consolidada** única para los datos, lo que hace más fácil escribir las consultas para la toma de decisiones.

- Esa interfaz consolidada, permite que todos los datos de entrada al almacén, no importando de que sistema de adquisición de datos provengan, sean **homogeneizados y unificados** en convenciones establecidas para el almacén.
 - Si datos que describen a una misma característica provienen de dos o más fuentes con diferente formato, el almacén establece cuál es el formato que prevalecerá.

Ventajas de usar un DW



- Alto retorno de inversión
 - El uso de un almacén permitirá mejorar los procesos de toma de decisiones en una institución, y por tanto habrá un crecimiento en el desarrollo de la misma.

Ventajas de usar un DW



- Ventajas competitivas derivadas de su utilización
 - La institución podrá ser competitiva e incluso mejorar los procesos que se llevan a cabo en otras instituciones similares. Por lo anterior, habrá una ganancia palpable con su uso.

Ventajas de usar un DW



- Mayor productividad de los responsables de la toma de decisiones
 - El encargado de la toma de decisiones dentro de la institución podrá actuar de una forma objetiva, eficiente y eficaz en la solución de problemáticas reales y constantes.

Desventajas de usar un DW



- Subestimación de los recursos necesarios para la carga de datos.
 - Debido a que se tratan de proyectos de larga duración, ya que se usan para la solución de problemas complejos, puede existir una alta demanda de recursos que sobrepase los existentes.

Desventajas de usar un DW



- Problemas ocultos de los sistemas de origen.
 - Un almacén de datos se nutre de información proveniente de diversas fuentes de información, por lo que puede haber problemas en ellos:
 - No cuentan con los datos requeridos
 - Hay una variación en los datos, por lo que requiere procesos de homogeneización
 - Complejidad de integración
 - Propiedad de los datos

Desventajas de usar un DW



- Incremento de la demanda por parte de los usuarios finales.
- Altos costos de mantenimiento
 - Debido a que se cuenta con grandes volúmenes de información.

Requerimientos de un DW

- Para poder diseñar un almacén de datos es necesario considerar **cuatro requerimientos** indispensables:

- Tema
- Integración
- Tiempo
- Volatilidad



- Organizado de acuerdo a temas

- La información se clasifica con base en los aspectos que son de interés para la empresa.
- Este requerimiento permite poder observar y analizar la información desde diferentes perspectivas, las cuales se usan para resolver diferentes problemáticas.

- Integración

- La integración de datos consiste en especificar la definición que conservarán los datos que entren al almacén, a través del proceso ETL (Extraction, Transformation, Load). Esta definición incluye:

- Convenciones de nombres
- Modificaciones consistentes
- Medidas uniformes de variables, etc.

- Dependiente del tiempo

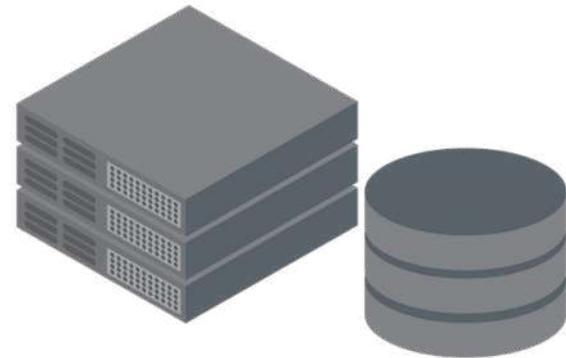
- La información representa los datos sobre un horizonte largo de tiempo – histórico de los datos.
- Cada estructura clave contiene (implícita o explícitamente) un elemento de tiempo (día, semana, mes, etc.).
- La información, una vez registrada correctamente, no puede ser actualizada (algo que ha sucedido, ya no puede modificarse).

- Volatilidad

- El almacén sólo permite cargar nuevos datos y acceder a los ya almacenados.
- No se permite borrar ni modificar los datos ya almacenados.

Arquitectura DW

- Para poder crear un almacén de datos, se considera una arquitectura que consta de **cuatro módulos** principales:
 - Datos de origen
 - Proceso ETL
 - Almacén
 - Herramientas de acceso



- Datos de origen

- Se refiere al origen de los datos que serán almacenados físicamente por el Almacén de Datos, estos pueden ser:

- Datos operacionales – Sistemas OLTP
 - Datos provenientes de Internet
 - Archivos
 - Datos provenientes de encuestas

- * Los más comunes son los primeros

- *Proceso ETL (extraction – transformation – load)*
 - **Extracción** de datos.
 - Selección sistemática de los datos operacionales que formarán parte del Almacén de Datos.
 - **Transformación** de datos.
 - Procesos para sumarizar y realizar cambios en los datos operacionales (procesos de unificación).
 - **Carga** de datos.
 - Inserción de datos en el Almacén.

- Almacén.

- Almacenamiento físico de los datos en la arquitectura del almacén.

- Herramientas de acceso.

- Herramientas externas que tienen acceso a los datos para poder procesarlos y analizarlos.

Diagrama de arquitectura



Utilización de un DW

Existen tres razones principales por las que se usa un DW:

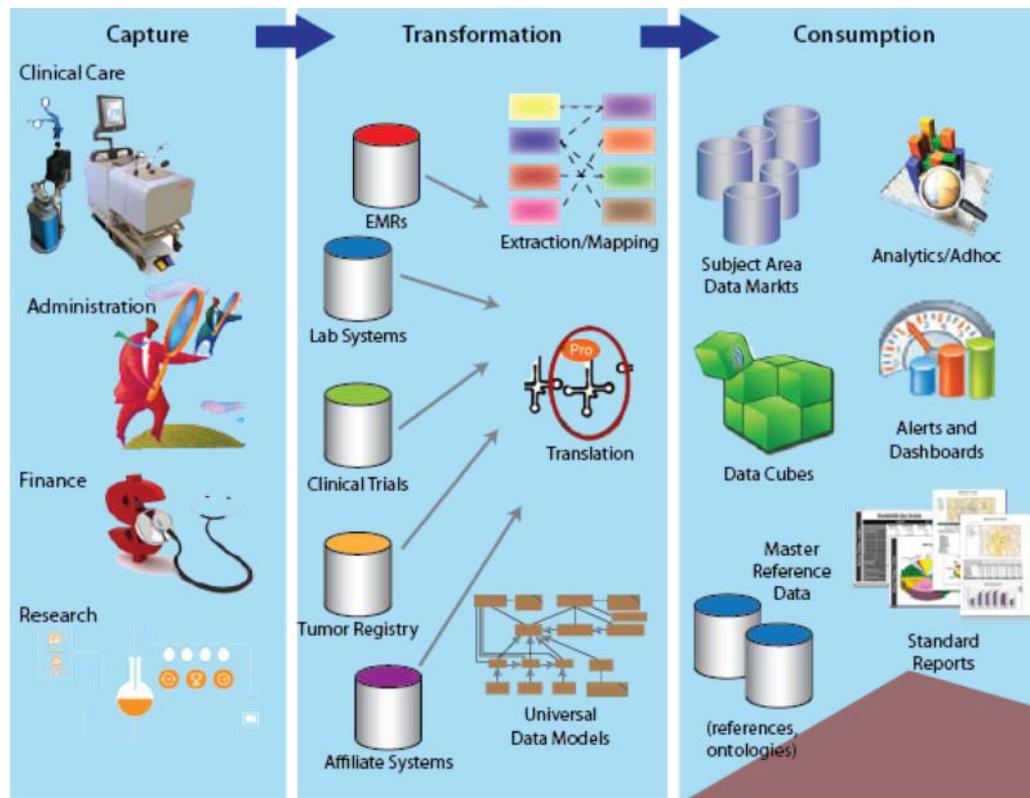
1. Integración de bases de datos heterogéneas (relacionales, documentales, geográficas, archivos, etc.)



Fuente: <https://www.elevondata.com/services/data-warehousing>

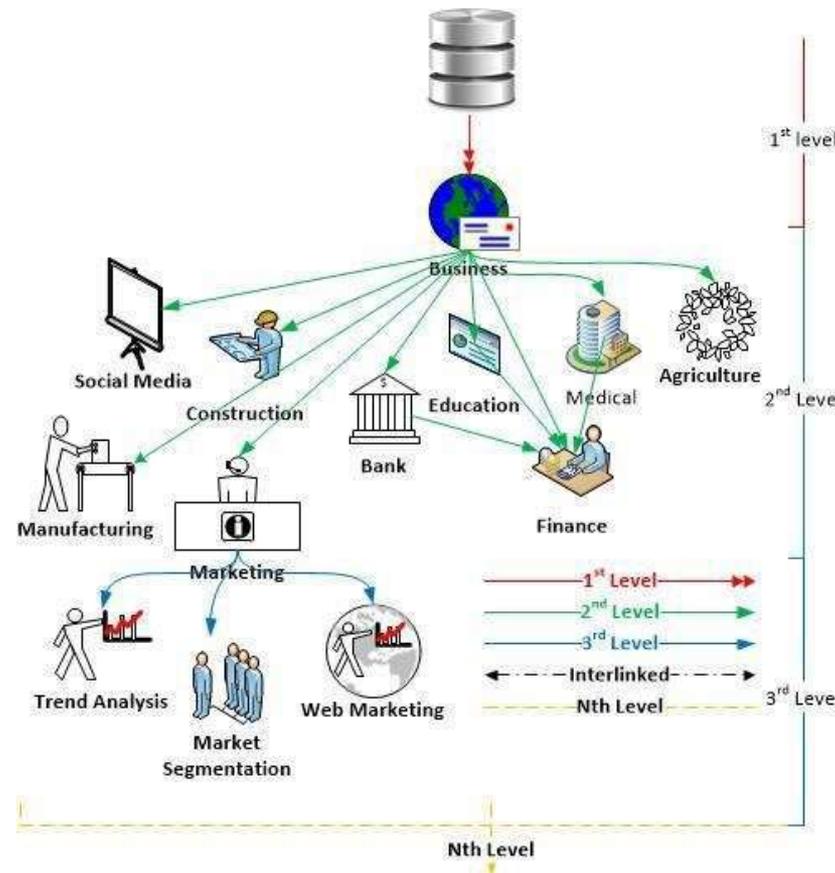
2. Análisis del problema en términos de dimensiones

- Con un mismo almacén se puede dar solución a diferentes problemáticas de la institución.



Fuente: <https://edgewater.tech.wordpress.com/2012/04/25/please-stop-telling-everyone-you-have-an-enterprise-data-warehouse-because-you-dont/>

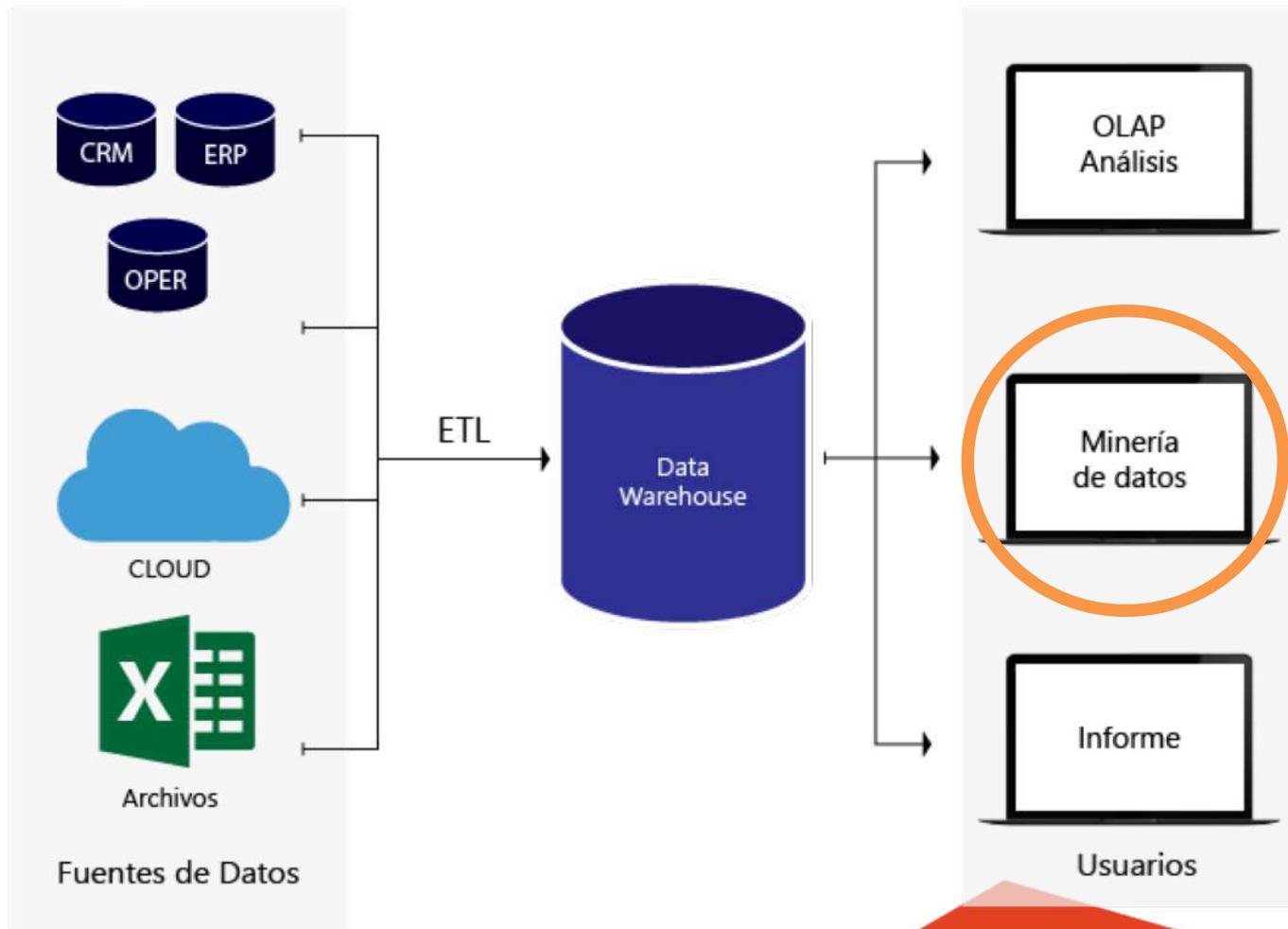
3. Ejecución de consultas complejas no predefinidas visualizando el resultado en forma gráfica y en diferentes niveles de agrupamiento y totalización de datos.



Fuente: https://www.researchgate.net/figure/Business-Application-of-Data-Warehouse_fig3_301770422

DW en DM

- Un desafío principal para las organizaciones es **identificar los datos más adecuados** para aplicar técnicas de Minería de Datos.
- Esta disciplina necesita una **fuentes de datos unificada, independiente, limpia, integrada y auto coherente**. Un almacén de datos está bien preparado para proporcionar los datos requeridos por la Minería de Datos.



Fuente: <https://gravitar.biz/datawarehouse/data-warehouse-tipos/>

Conclusiones

El uso de almacenes de datos por alguna institución permite que ésta conserve su información de una manera ordenada y limpia. Una principal ventaja es que ayuda a administrar los datos que provienen de diferentes fuentes de información con el fin de poder utilizarlos en forma posterior para mejorar la toma de decisiones en la institución.

Los almacenes son una fuente fundamental para el uso de técnicas de Minería de Datos.

Bibliografía

Chakrabarti, S., et al. (2009) Data Mining, know it all. Morgan Kaufmann Publishers.

Hernández Orallo, J., Ramírez Quintana, M.J., Ferri ramírez C. (2005). Introducción a la Minería de Datos. Pearson Education.

Khalaf Hamoud, A.A., Salah Hashim, A. and Akeel Awadh, W. (2018) Clinical Data Warehouse: A review. Iraqi Journal for Computers and Informatics, Vol. 44, No. 2, doi: 10.25195/2017/4424.

Ponniah. P. (2001) Data Warehousing Fundamentals. John Wiley & Sons.

Datos de contacto

Dra. Anilu Franco Arcega

afranco@uaeh.edu.mx

Dr. Virgilio López Morales

virgilio@uaeh.edu.mx

Dr. Joel Suárez Cansino

jsuarez@uaeh.edu.mx

Área Académica de Computación y Electrónica

Instituto de Ciencias Básicas e Ingeniería

Universidad Autónoma del Estado de Hidalgo

Teléfono: 7717172000 ext. 6734