Comparativa de metodologías de elaboración de DataWarehouses vs DataLakes

Allison Chino, Marco Garcia, Miguel

3 de septiembre de 2021

Resumen

Cada vez es más común que las empresas traten de buscar soluciones al almacenamiento de grandes volúmenes de datos recurriendo a la tecnología Big Data. En este contexto, para gestionar toda esta información los profesionales pueden optar por dos sistemas, Data Lake o Data Warehouse. A veces su elección genera dudas así que vamos a ver en detalle en qué consiste cada uno de ellos, así como sus principales diferencias para que cada empresa pueda tomar la mejor elección para sus proyectos.

Abstract

It is increasingly common for companies to try to look for solutions to the storage of large volumes of data by resorting to Big Data technology. In this context, to manage all this information, professionals can opt for two systems, Data Lake or Data Warehouse. Sometimes your choice generates doubts so let's see in detail what each of them consists of, as well as their main differences so that each company can take the best choice for their projects.

I. Introduccion

Internet y las nuevas tecnologías han provocado el acceso y el almacenamiento desmesurado de información de los clientes y potenciales.

Las empresas son cada vez más conscientes de la importancia que tienen esos datos para conocer mejor a los usuarios y así poder ofrecerles aquello que realmente piden, y no lo que nosotros pensamos que necesitan. Esto es lo que se llama, aplicar estrategias customer centric.

Para ello se necesita gestionar altos volúmenes de datos, tanto en tiempo real como organizados. Para ello, no hay nada mejor que un Data Warehouse o un Data Lake.

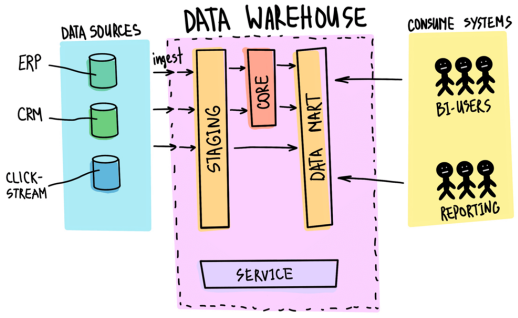
II. Desarrollo

II.1. ¿Qué es el Data Warehouse?

Data Warehouse es una combinación de tecnologías y componentes para el uso estratégico de datos. Recopila y gestiona datos de diversas fuentes para proporcionar información empresarial significativa. Es el almacenamiento electrónico de una gran cantidad de información diseñada para consultas y análisis en lugar de procesamiento de transacciones. Es un proceso de transformación de datos en información.

Las funciones importantes que se deben realizar son:

- Extracción de datos.
- Limpieza de datos.
- Transformación de datos.
- Carga y actualización de datos.



Ventajas de un Data Warehouse

- La data está lista para ser usada.
- Buena performance en el acceso a los datos.
- La mayoría de los usuarios de una empresa son operacionales, data warehouse es ideal para ellos.
- Es muy adecuado para generar reportes y métricas.

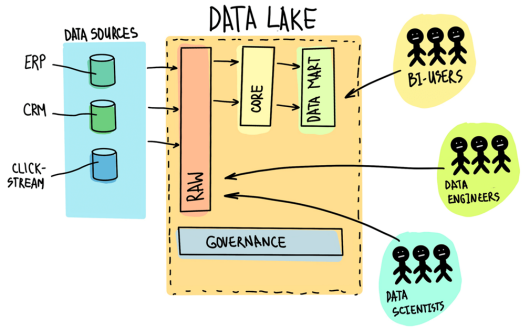
Desventajas de un Data Warehouse

- Mayores costos de almacenamiento, lo que implica pensar bien qué data es realmente necesaria.
- No es flexible a cambios.
- Demanda inversión de tiempo antes del almacenamiento para decidir esquemas, formatos y casos de uso.

II.2. ¿Qué es Data Lake?

Un Data Lake es un repositorio de almacenamiento que puede almacenar una gran cantidad de datos estructurados, semiestructurados y no estructurados. Es un lugar para almacenar todo tipo de datos en su formato nativo sin límites fijos en el tamaño de la cuenta o el archivo. Ofrece una gran cantidad de datos para un mayor rendimiento analítico e integración nativa.

Data Lake es como un gran contenedor que es muy similar a los lagos y ríos reales. Al igual que en un lago, entran múltiples afluentes. De manera similar, un lago de datos tiene datos estructurados, datos no estructurados, máquina a máquina, registros que fluyen en tiempo real.



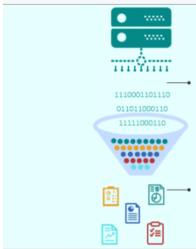
Ventajas de un Data Lake

- No hay necesidad de descartar datos.
- Puede nutrir a diversos usuarios de una empresa.
- Se adapta fácilmente a los cambios.
- Al poder integrarse tipos de datos muy distintos, se puede realizar todo tipo de análisis.
- Permite fácilmente agregar nueva data.

Desventajas de un Data Lake

- No está pensado para acceder a los datos de manera performante.
- Cada vez que se requieren datos, hay que transformarlos y curarlos para el uso que se les quiera dar.
- Demanda invertir en generar estándares de buenas prácticas a nivel organizacional.

II.3. Diferencia clave entre Data Lake y Data Warehouse

Data Lake	Data Warehouse
	
Se almacena toda la información **independientemente de la fuente y su estructura.**	Se almacena información, proveniente de sistemas transaccionales, **que pasa por un proceso de normalización y transformación.**
Funciona como un **"repositorio"** de datos estructurados, semi-estructurados y no-estructurados	Se **combinan tecnologías** para recopilar y gestionar datos procedentes de distintas fuentes (suelen almacenarse de manera estructurada)
Se define el esquema de la dará **después** de su almacenamiento, al momento de ser usada.	Se define el esquema **antes** del almacenamiento y **antes** de su uso.
Usa el proceso **extracción-carga-transformación.**	Usa el proceso **extracción-transformación -carga**.
Ideal para hacer **todo tipo de análisis en profundidad.**	Ideal para **usuarios operacionales.**

II.4. Comparación de Metodologías de Elaboración

Una de las diferencias que marcan la comparación entre Data Lake vs Data Warehouse es que, este último, por lo general, sigue una metodología como las definidas por Inmon y Kimball, a diferencia de Data Lakes.

Metodologías de Data Warehouse

Al diseñar una solución de Data Warehouse, las organizaciones se topan con distintas metodologías y enfoques a seguir, las cuales se deberán evaluar para seleccionar la que mejor se adapte a los requisitos del proyecto.

Las siguientes metodologías fueron diseñadas por Ralph Kimball, Bill Inmon y Dan Linsted correspondientemente.

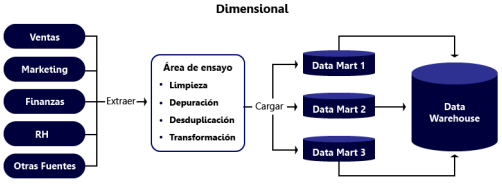
1. Dimensional (BUTTON UP / ASCENDENTE)

La metodología dimensional o metodología de Ralph Kimball, mantiene un diseño ascendente por lo que los Data Marts son los primeros en crearse y después se integran al Data Warehouse creando un almacenamiento más completo.

Busca que el almacenamiento de datos de los usuarios se ejecute de la forma más rápida posible. Según Kimball, un almacenamiento de datos es la copia de los datos transaccionales específicamente estructurados para consultas analíticas e informes con el fin de apoyar la toma de decisiones. Con su metodología al crear primero los Data Marts se proporcionan capacidades analíticas y de informes para procesos específicos de negocio y funcionales.

Principales Diferenciadores

- Para mantener el Data Warehouse solo se requiere de un equipo pequeño de desarrolladores y arquitectos de datos.
- Brinda buena funcionalidad para las métricas en cuanto al departamento y el seguimiento de KPI, al orientar los Data Marts a informes en cuanto a procesos de departamento o de negocios.



La metodología de Kimball propone crear una matriz de negocio que contenga los elementos comunes que son utilizados por los Data Marts, como conformed-shared dimension, measures, etc., teniendo esta información, el usuario puede desarrollar soluciones que apoyen el análisis a través de los procesos de negocio para la venta cruzada.

2. Relacional (TOP DOWN / DESCENDENTE)

La metodología relacional de Bill Inmon muestra un diseño descendente, donde se construye primero el Data Warehouse y por consiguiente los Data Marts. Ubicando el DW en el centro de la información corporativa lo que asegura

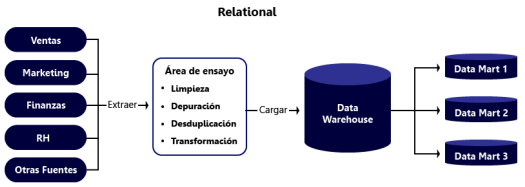
un marco lógico en los datos.

Crear una estructura de entidades procurando que no se repitan datos. Este modelo crea una única fuente de verdad para todo el negocio. La carga de datos se vuelve menos compleja debido a la estructura normalizada del modelo. Sin embargo, el uso de esta disposición para realizar consultas es complicado, ya que incluye gran cantidad de tablas y vínculos.

Este modelo propone la construcción de Data Marts por separado para cada departamento. Todos los datos que entran en el Data Warehouse están integrados. Para garantizar la integridad y la coherencia en toda la empresa, el Data Warehouse actúa como un único origen de datos para varios Data Marts.

Principales Ventajas

- El Data Warehouse proporciona una única versión de la verdad, al ser el único origen de datos para los Data Marts.
- Es para los usuarios comprender más fácilmente los procesos empresariales, ya que el modelo lógico representa entidades empresariales detalladas.
- Resulta más fácil y menos propenso al fracaso el proceso de ETL, puesto que en la actualización de los datos las anomalías se evitan al contar con una redundancia muy baja.



3. Data Vault

Es una metodología de seguimiento histórico orientado a los detalles y conjunto de tablas normalizadas vinculadas de forma única que admiten una o más áreas funcionales de negocios. Es un modelo de datos que está diseñado específicamente para cumplir las necesidades de uno o varios Data Warehouse empresariales.

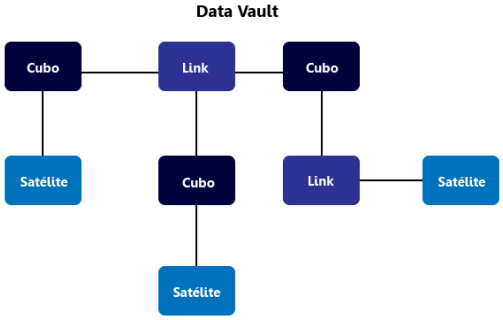
Principales Ventajas

- Diseñado especialmente para almacenar registros. Hace que el proceso de registro de datos sea más sencillo.
- Con esta metodología, es más fácil agregar un nuevo origen de datos sin modificar el ya existente.
- Automatiza fácilmente los procesos ETL.

Arquitectura de Data Vault

Data Vault contiene tres tablas básicas:

- Hub
- Links
- Satélites



III. Conclusiones

Tanto los Data Warehouses como los Data Lakes están destinados a convivir en las empresas que deseen basar sus decisiones en datos. Como se puede entender, ambos son complementarios, no sustitutivos, pudiendo ayudar a cualquier negocio a conocer mejor el mercado y el consumidor, de cara a poder realizar estrategias basadas en el conocimiento de estos, con comunicaciones cada vez más personalizadas, es decir, ser más customer centric.

COMPARACIÓN	DATA WAREHOUSE	DATA LAKE
Datos	Estructurado, datos procesados	Datos estructurados/ semiestructurados, no estructurados, raw data, datos sin procesar
Procesamiento	Esquema sobre escritura	Esquema en lectura
Almacenamiento	Costoso, confiable	Almacenamiento de bajo costo
Agilidad	Menos ágil, configuración no tan flexible	Ágil, configuración flexible
Seguridad	Madurado	Madurando
Usuarios	Negocio profesional	Científicos de datos (familiarizados con el dominio)

IV. Recomendaciones

El data warehouse responde a necesidades más maduras, cuando ya sabemos qué data es la más importante y qué tipo de trabajo vamos a hacer con ella.

Por el otro lado, los data lakes son alternativas más ágiles, baratas y flexibles: atributos muy deseables para empresas jóvenes, como puede ser tu startup.

Los data lakes pueden ser especialmente efectivos cuando no sabemos bien qué queremos hacer con toda nuestra data, pero sabemos que la información es potencialmente valiosa.

Referencias

[1] IBM(2021).¿Qué es un Data Warehouse?. Recuperado de
`https://cutt.ly/ZWQGWZF`

[2] AWS (2020). Data lake on AWS. Recuperado de
`https://cutt.ly/4nod8Lm`

[3] AWS (2019). Data Lake Foundation on the AWS Cloud. Recuperado de
`https://cutt.ly/lnod5w8`

[4] Data Kitchen (2020). Agile Data Lake and Warehouse. Recuperado de
`https://cutt.ly/Tnod6Bj`

[5] Google Cloud (2020). Cloud Storage as a data lake. Recuperado de
`https://cutt.ly/RnofwG5`

[6] IKimball, R. (2016). The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence Remastered Collection. Recuperado de
`https://cutt.ly/Jnofruq`