Data Warehouse VS Data Lake

Ericka Esther Martinez Yufra, Brian Sebastian Anco Copaja
September 3, 2021

/beginResumen Si bien los data lake y los data warehouse se utilizan para almacenar grandes cantidades de datos, ambos términos poseen diferencias marcadas que son utilizadas para satisfacer necesidades específicas, misma razón por la que requieren un enfoque distinto para alcanzar una optimización óptima.

De manera breve, se puede entender a un data lake como un gran conjunto de datos que aún no cuentan con una finalidad específica, mientras que un data warehouse contiene datos ya estructurados y debidamente filtrados con un fin determinado. Así, las principales diferencias entre ambos son la estructura de datos, los métodos de procesamiento, el campo de aplicación y la finalidad con la que serán utilizados los datos.

Abstract

Although data lakes and data warehouses are used to store large amounts of data, both terms have marked differences that are used to satisfy specific needs, the same reason why they require a different approach to achieve optimal optimization.

Briefly, a data lake can be understood as a large set of data that does not yet have a specific purpose, while a data warehouse contains data that is already structured and properly filtered for a specific purpose. Thus, the main differences between the two are the data structure, the processing methods, the field of application and the purpose for which the data will be used.

1 Introducción

Junto al crecimiento de la tecnología, surgen a la par nuevos desafíos que deben enfrentar las empresas para mantenerse a la altura de las crecientes necesidades de sus clientes. Uno de ellos, es el enorme crecimiento de datos. A ello se suma que una gran parte de estos datos son información no estructurada, y otro importante porcentaje de los datos que maneja una empresa, son almacenados y gestionados fuera de sus centros de datos. Por ello, la recopilación de datos es una parte fundamental al buscar un mejor posicionamiento de las empresas frente a sus competidores, a través de la toma de decisiones basada en datos.

Para poder trabajar con estas grandes cantidades de datos, tenemos los data warehouse y los data lake, los cuales son paradigmas diferentes para almacenar y tratar datos, expandiendo el almacenamiento de datos a no solo su recolección y protección, sino logrando también aprender de ellos.

2 Desarrollo

2.1 Data Warehouse

Un data warehouse (almacén de datos) es una base de datos diseñada para almacenar, filtrar, extraer y analizar grandes colecciones de datos (de proveedores, clientes, marketing, administración, recursos humanos, bancos, etc.). La particularidad de estos sistemas es que están específicamente desarrollados para trabajar con big data, permitiendo visualizar y analizar de manera cruzada la información de forma simultánea, sin tener que mezclar y consolidar resultados procedentes de distintas fuentes de datos.

Está diseñado para separar los procesos de análisis y consulta de big data (más enfocados en la lectura de datos) de los procesos transaccionales (centrados en la escritura). Este planteamiento permite por lo tanto a una empresa multiplicar su poder de análisis sin impactar en sus sistemas transaccionales y las necesidades de la gestión del día a día.

Son una herramienta muy recomendable cuando se quiere garantizar que usuarios inexpertos en el manejo de sistemas y bases de datos puedan poner en riesgo la información de una empresa. Dada la arquitectura a tres niveles empleada en estas soluciones, los usuarios finales de los DWH pueden hacer consultas sobre sus almacenes de datos sin tocar o afectar en modo alguno la operación del sistema.

En síntesis, la arquitectura de un data warehouse se base en tres niveles:

- Nivel inferior: Es el servidor, donde se cargan y almacenan los datos.
- Nivel intermedio: Contiene el motor de análisis que se utiliza para acceder a los datos.
- Nivel superior: Es el cliente front-end que presenta los resultados de los análisis mediante herramientas de visualización de datos.

2.2 Data Lake

Un data lake es un gran repositorio de información que nos permite almacenar datos estructurados y no estructurados a cualquier escala. Como su nombre lo dice, tal y como los lagos y ríos reales, los datos de un data lake provienen de diferentes fuentes que llenan el lago, permitiendo a los usuarios sumergirse en ellos para poder examinarlos o tomar muestras. Almacenar de esta manera la información, nos permite conservar datos en un formato más flexible para su uso en el futuro.

Los data lakes son configurados generalmente en un clúster de hardware con un consumo económico y escalable, de esta manera podemos guardar datos sin preocuparnos por la capacidad de almacenamiento. Estos clúster pueden ser locales o en la nube.

Ya contenidos los datos, estos deben estar disponibles para todos sus usuarios. Los datos son introducidos sin mayor complejidad, y se gestionan a través de etiquetas de metadatos que permiten localizar y conectar la información cuando los usuarios la necesitan. Un data lake contiene tres características principales:

- Recibir todo: Contienen todos los datos, incluyendo fuentes procesadas y no procesadas, durante un largo periodo de tiempo.
- Fácil de usar: Permite a los usuarios refinar, explorar y enriquecer los datos como sean requeridos.
- Acceso flexible: Permite múltiplos formas de acceso a datos en una infraestructura compartida.

2.3 Comparación

Parámetros	**Data Lake**	**Data Warehouse**
Almacenamiento	Todos los datos se guardan independientemente de la fuente y su estructura, se mantienen en su forma original y solo se transforma cuando está listo para ser utilizado.	Consiste en datos que se extraen de sistemas transaccionales o datos que consisten en métricas cuantitativas con sus atributos. Los datos se limpian y transforman.
Tipo de datos	Datos estructurados y no estructurados de diversas fuentes de datos de la empresa	Datos históricos que se han estructurado para adaptarse a un esquema de base de datos relacional
Captura de datos	Captura todo tipo de datos y estructuras, semiestructurados y no estructurados en su forma original de los sistemas de origen.	Captura información estructurada y la organiza en esquemas definidos para propósitos de almacenamiento de datos.

Cronograma de datos	Los lagos de datos pueden retener todos los datos. Esto incluye no solo los datos que están en uso, sino también los datos que podrían usar en el futuro. Además, los datos se guardan para siempre, para retroceder en el tiempo y hacer un análisis.	En el proceso de desarrollo del almacén de datos, se dedica un tiempo considerable a analizar diversas fuentes de datos.
Usuarios	Ideal para los usuarios que se entregan a un análisis profundo. Dichos usuarios incluyen científicos de datos que necesitan herramientas analíticas avanzadas con capacidades como el modelado predictivo y el análisis estadístico.	Usuarios operativos debido a que está bien estructurado, es fácil de usar y comprender.
Costos de almacenamiento	Relativamente económico	Costoso y requiere más tiempo
Procesamiento de datos	Uso de Data Lakes del proceso ELT (Extract Load Transform)	Utiliza un proceso ETL (Extract Transform Load) tradicional.
Tipo de procesamiento	Permiten a los usuarios acceder a los datos antes de que hayan sido transformados, limpiados y estructurados. Por lo tanto, permite llegar a sus resultados más rapidamente en comparación con el almacén de datos tradicional.	Ofrecen información sobre preguntas predefinidas para tipos de datos predefinidos. Por lo tanto, cualquier cambio en el almacén de datos requerirá más tiempo.
Posición del esquema	Se define después de almacenar los datos. Esto ofrece una gran agilidad y facilidad de captura de datos, pero requiere trabajo al final del proceso.	Se define antes de almacenar los datos. Requiere trabajo al principio del proceso, pero ofrece rendimiento, seguridad e intregación.
Tarea	Contienen todos los datos y tipos de datos, permite a los usuarios acceder a los datos antes del proceso de transformación, limpieza y estructuración.	Proporcionan información sobre preguntas predefinidas para tipos de datos predefinidos.

Desventaja proncipal	Los datos se conservan en bruto. Solo se transforman cuando están listos para ser utilizados.	Incapacidad, el problema que se plantea al intentar realizar cambios en ellos.
Beneficio clave	Integran diferentes tipos de datos para plantear preguntas totalmente nuevas, ya que estos usuarios no suelen utilizar los almacenes de datos porque pueden necesitar ir más allá de sus capacidades.	La mayoría de los usuarios de una organización son operativos. A este tipo de usuarios solo les interesan los informes y las métricas clave de rendimiento.

3 Conclusiones

Los lagos de datos son más útiles cuando hay más datos para almacenar sin un tipo particular de estructuras definidas. Y no es necesario analizarlo todo de inmediato. La característica importante del lago de datos es su flexibilidad sobre el almacenamiento de datos. No habrá ninguna parte de procesamiento ETL en el lago de datos, que está destinado a una lógica empresarial específica.

Por otro lado, en los almacenes de datos tradicionales, procesa y transforma los datos para análisis avanzados y consultas en un ecosistema de base de datos altamente estructurado. Le brinda una funcionalidad más precisa y específica en las soluciones de BI y Reporting. Aunque la arquitectura y la capacidad de ambos tienen algunas similitudes, nunca supusieron un reemplazo directo el uno del otro. actúan como una tecnología coexistente que sirve a diferentes casos de uso con cierta superposición. Las soluciones de lago de datos generalmente se consideran complementarias a los almacenes de datos.

Hoy en día, la mayoría de las organizaciones mantienen un lago de datos para respaldar los almacenes de datos. Sin embargo, a medida que aumenta el volumen de datos, los almacenes de datos en la nube y los lagos de datos se están convirtiendo en la solución preferida sobre los almacenes de datos tradicionales. Las tecnologías modernas en la nube brindan soluciones como escalado, seguridad de datos, monitoreo, confiabilidad y mantenimiento a bajo costo.

4 Recomendaciones

- Si se va a decidir entre un lago de datos o un almacén de datos, revisar estas categorías y ver cuál se adapta mejor al caso de uso.

- No olvidar que a veces necesita una combinación de ambas soluciones de almacenamiento. Esto es especialmente cierto cuando se crean canalizaciones de datos.

References

- Fang, H. (2015, June). Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER) (pp. 820-824). IEEE.

- Inmon, W. H. (1996). The data warehouse and data mining. Communications of the ACM, 39(11), 49-51.

- John, T., Misra, P. (2017). Data lake for enterprises. Packt Publishing Ltd.

- Khine, P. P., Wang, Z. S. (2018). Data lake: a new ideology in big data era. In ITM web of conferences (Vol. 17, p. 03025). EDP Sciences.

- Linstedt, D., Olschimke, M. (2015). Building a scalable data warehouse with data vault 2.0. Morgan Kaufmann.

- Miloslavskaya, N., Tolstoy, A. (2016). Big data, fast data and data lake concepts.

Procedia Computer Science, 88, 300-305.

- Ramos, S. (2016). Data Warehouse, data marts y modelos dimensionales. Un pilar fundamental para la toma de decisiones. Albatera: SolidQ.

- Salinas, S. O., Lemus, A. C. (2017). Data warehouse and big data integration. Int. Journal of Comp. Sci. and Inf. Tech, 9(2), 1-17.