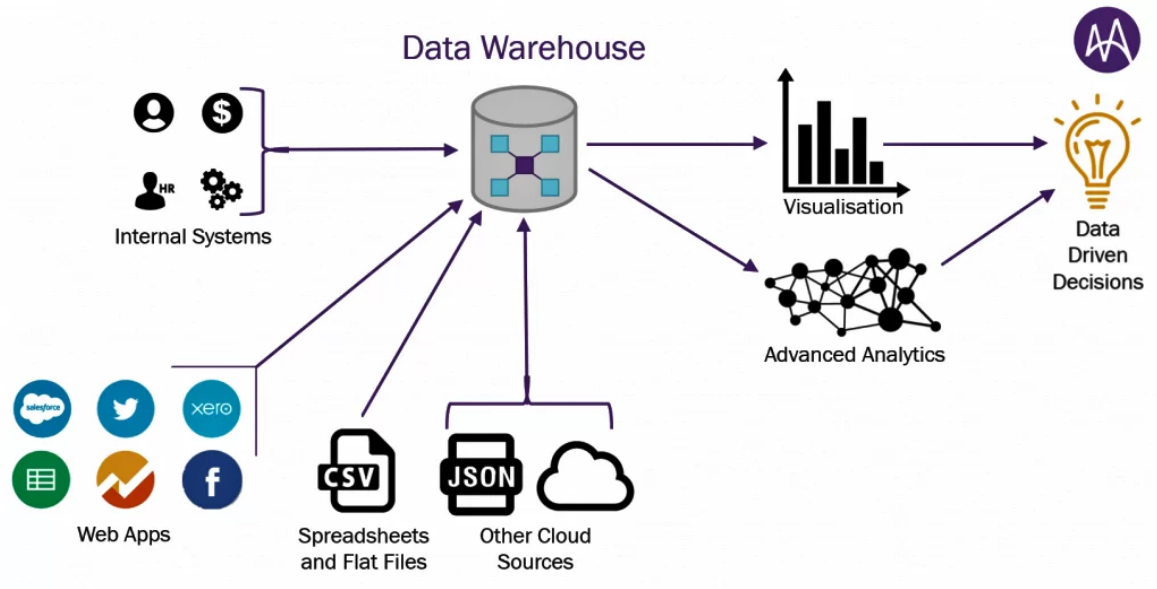
Data Warehousing



Un Data Warehouse es un almacén electrónico donde generalmente una empresa u organización mantiene una gran cantidad de información. Los datos de un data warehouse deben almacenarse de forma segura, fiable, fácil de recuperar y fácil de administrar.

El concepto de data warehouse se originó en 1988 con el trabajo de los investigadores de IBM, Barry Devlin y Paul Murphy aunque el término data warehouse fue acuñado por William H. Inmon, el cual es conocido como el padre de Data Warehousing. Inmon describió un data warehouse como una colección de datos orientada a un tema específico, integrado, variante en el tiempo y no volátil, que soporta el proceso de toma de decisiones.

¿Qué es un data Warehouse?

Un data warehouse es un repositorio unificado para todos los datos que recogen los diversos sistemas de una empresa. El repositorio puede ser físico o lógico y hace hincapié en la captura de datos de diversas fuentes sobre todo para fines analíticos y de acceso.

Normalmente, un data warehouse se aloja en un servidor corporativo o cada vez más, en la nube. Los datos de diferentes aplicaciones de procesamiento de transacciones Online (OLTP) y otras fuentes se extraen selectivamente para su uso por aplicaciones analíticas y de consultas por usuarios.

Data Warehouse es una arquitectura de almacenamiento de datos que permite a los ejecutivos de negocios organizar, comprender y utilizar sus datos para tomar decisiones estratégicas. Un data warehouse es una arquitectura conocida ya en muchas empresas modernas.

Pasado y presente de los Data Warehouse

Históricamente, los data warehouses se habían formado utilizando datos repetitivos estructurados que eran filtrados antes de entrar en el data warehouse. Sin embargo, en los últimos años, el data warehouse ha evolucionado debido a información contextual que ahora se puede adjuntar a los datos no estructurados y que también puede ser almacenada.

Aquellos primeros datos relacionales estructurados no podían ser mezclados y emparejados para temas analíticos con datos textuales no estructurados. Pero con el advenimiento de la contextualización, estos tipos de análisis ahora sí pueden hacerse de formas naturales y fáciles.

En el data warehouse, datos no repetitivos, como los comentarios en una encuesta, correos electrónicos y conversaciones, se tratan de forma diferente a las ocurrencias repetitivas de datos, como el flujo de clics, mediciones o el procesamiento máquina o analógico. Los datos no repetitivos son datos basados en textos que fueron generados por la palabra escrita o hablada, leída y reformateada y, lo que es más importante, ahora puede ser contextualizada. Con el fin de extraer cualquier sentido de los datos no repetitivos para su uso en el Data Warehouse, deben tener el contexto de los datos establecidos.

En muchos casos, el contexto de los datos no repetitivos es más importante que los datos en sí. En cualquier caso, los datos no repetitivos no pueden utilizarse para la toma de decisiones hasta que se haya establecido el contexto.

Data Lakes y Data Warehouses: ¿mutuamente exclusivos o partners perfectos?

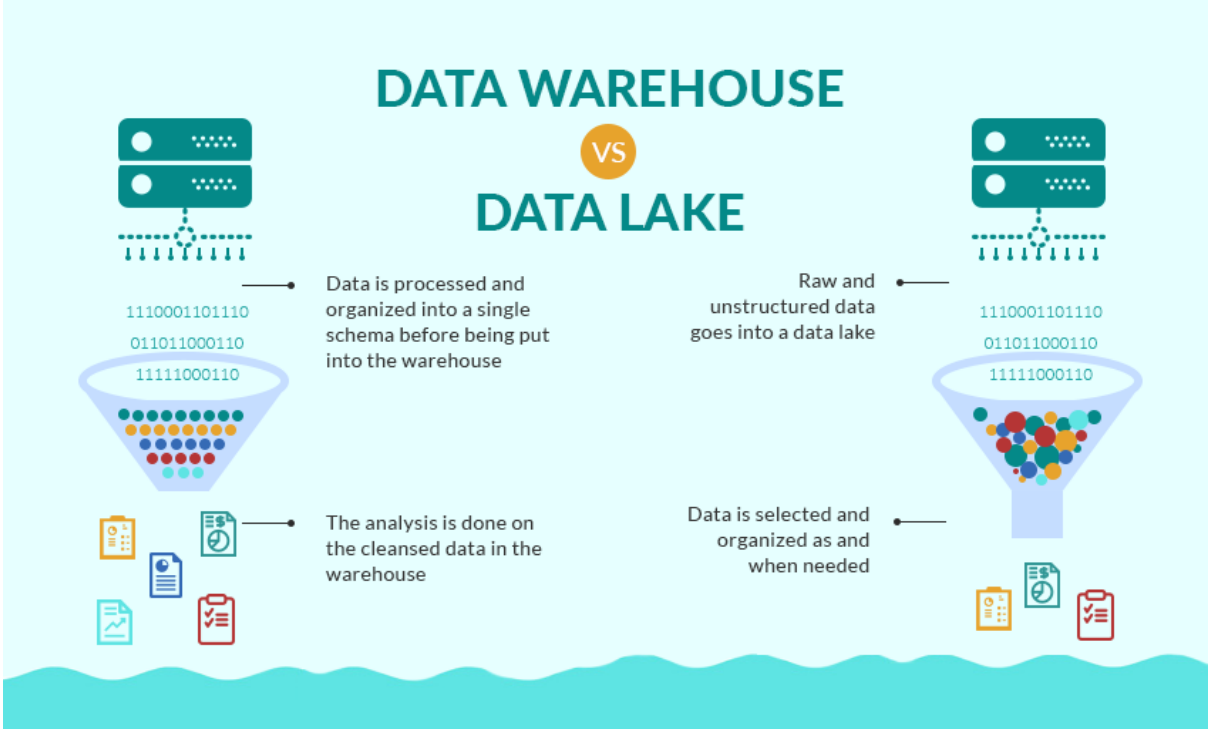
Los data lakes han surgido en el paisaje de Data Management en los últimos años, sin embargo, data lake no es necesariamente un reemplazo del data warehouse. En cambio, complementan los esfuerzos existentes y dan soporte al descubrimiento de nuevas preguntas. Una vez que se descubren esas preguntas se optimizan las respuestas. Y optimizar puede significar moverse fuera del data lake para ir a un data mart o al data warehouse

Estas son algunas diferencias clave entre data lake y data warehouse:

- **Datos**:
  Un data warehouse sólo almacena datos que han sido modelados o estructurados, mientras que un Data Lake no hace acepción de datos. Lo almacena todo, estructurado, semiestructurado y no estructurado.
- **Procesamiento**:
  Antes de que una empresa pueda cargar datos en un data warehouse, primero debe darles forma y estructura, es decir, los datos deben ser modelados. Eso se llama schema-on-write. Con un data lake, sólo se cargan los datos sin procesar, tal y como

están, y cuando esté listo para usar los datos, es cuando se le da forma y estructura. Eso se llama schema-on-read. Dos enfoques muy diferentes.

- **Almacenamiento**:
  Una de las principales características de las tecnologías de big data, como Hadoop, es que el coste de almacenamiento de datos es relativamente bajo en comparación con el de un data warehouse. Hay dos razones principales para esto: en primer lugar, Hadoop es software de código abierto, por lo que la concesión de licencias y el soporte de la comunidad es gratuito. Y segundo, Hadoop está diseñado para ser instalado en hardware de bajo coste.
- **Agilidad**:
  Un almacén de datos es un repositorio altamente estructurado, por definición. No es técnicamente difícil cambiar la estructura, pero puede tomar mucho tiempo dado todos los procesos de negocio que están vinculados a ella. Un data lake, por otro lado, carece de la estructura de un data warehouse, lo que da a los desarrolladores y a los científicos de datos la capacidad de configurar y reconfigurar fácilmente y en tiempo real sus modelos, consultas y aplicaciones.
- **Seguridad**:
  La tecnología del data warehouse existe desde hace décadas, mientras que la tecnología de big data (la base de un Data Lake) es relativamente nueva. Por lo tanto, la capacidad de asegurar datos en un data warehouse es mucho más madura que asegurar datos en un data lake. Cabe señalar, sin embargo, que se está realizando un importante esfuerzo en materia de seguridad en la actualidad en la industria de Big Data.



Data Warehouse vs Big Data vs BI: cuáles son las diferencias

Los tres conceptos están interconectados y la perspectiva es que, cada vez más, la mayoría de empresas utilicen el análisis generado por este tipo de tecnologías para tener una visión más analítica de su negocio y así poder tomar las mejores decisiones para crecer.

Se trata de tres conceptos completamente diferentes que tienen en común una nueva manera de lidiar con los datos, siempre teniendo en cuenta la existencia de una gran volumen de información en varios formatos que contribuyen, de forma estructurada o no estructurada, a la toma de decisiones estratégicas. El objetivo final de cualquiera de estas tecnologías es ofrecer una ventaja competitiva a las empresas, pero la forma en que se utiliza es la que marcará la diferencia.

- **Big Data:**
  Llamamos big data a un gran volumen de datos con una variedad, complejidad y velocidad de crecimiento enorme y que además tienen la característica de no ser estructurados. Eso significa que no son relacionales, estando además fuera del entorno corporativo. Es un tipo de tecnología que permite analizar los datos en tiempo real y puede provenir de diferentes fuentes y formas, tales como mensajería instantánea, redes sociales, registros de grabaciones, imágenes, mensajes de correo electrónico, etc.

  Por otro lado, un data warehouse almacena datos consolidados de diversas fuentes o sistemas de la empresa. Se trata de datos estructurados, que tiene como objetivo principal ser precisos y de alta calidad para de esta forma poder dar soporte a la toma de decisiones de la empresa. Se trata de conseguir tener todos los datos juntos para después poder dividirlos para hacer un análisis de determinados sectores o estrategias.

- **Data Warehouse:**
  Por otro lado, un data warehouse almacena datos consolidados de diversas fuentes o sistemas de la empresa. Se trata de datos estructurados, que tiene como objetivo principal ser precisos y de alta calidad para de esta forma poder dar soporte a la toma de decisiones de la empresa. Se trata de conseguir tener todos los datos juntos para después poder dividirlos para hacer un análisis de determinados sectores o estrategias.

- **BI:**
  Un Business Intelligence (BI) es una especie de “cuello de botella” de los datos recogidos del data warehouse, que llegan de forma exacta y útil para ayudar a la toma de decisiones. Business Intelligence transforma los datos en información útil para analizar no sólo los negocios, sino también las principales estrategias corporativas.

Los tres conceptos están interconectados y la perspectiva es que, cada vez más, la mayoría de empresas utilicen el análisis generado por este tipo de tecnologías para una visión más analítica de su negocio y así poder tomar las mejores decisiones para crecer.