**DATA WAREHOUSE**

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.

1. ¿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.

Estructuras de un Data Warehouse

La arquitectura de un data warehouse puede ser dividida en tres estructuras simplificadas: básica, básica con un área de ensayo y básica con área de ensayo y data marts.

Con una estructura básica, sistemas operativos y archivos planos proporcionan datos en bruto que se almacenan junto con metadatos. Los usuarios finales pueden acceder a ellos para su análisis, generación de informes y minería.

Al añadir un área de ensayo que se puede colocar entre las fuentes de datos y el almacén, ésta proporciona un lugar donde los datos se pueden limpiar antes de entrar en el almacén. Es posible personalizar la arquitectura del almacén para diferentes grupos dentro de la organización.

Se puede hacer agregando data marts, que son sistemas diseñados para una línea de negocio en particular. Se pueden tener data marts separados para ventas, inventario y compras, por ejemplo, y los usuarios finales pueden acceder a datos de uno o de todos los data marts del departamento.

2. Pasado y presente del 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 forma naturales y fácil.

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 de la 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 de la data lake para ir a un data mart o a la 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 de la 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.

El data warehouse continúa evolucionando

Hay todavía un gran potencial en la evolución de la data warehouse y su uso sigue adelante.

El procesamiento analítico clásico de datos basado en transacciones se realiza en el data warehouse como siempre se ha hecho. Nada ha cambiado ahí. Pero ahora puede hacerse analítica sobre datos contextualizados, y esa forma de análisis es nueva. La mayoría de las organizaciones, hasta ahora no habían podido basar su toma de decisiones en datos textuales no estructurados. Y ahora hay una nueva forma de análisis posible en un Data Warehouse: la posibilidad de la mezcla de análisis. El análisis combinado se realiza mediante una combinación de datos transaccionales estructurados y datos contextuales no estructurados.

Existen muchas otras formas de análisis que también son posibles. Estas formas incluyen análisis predictivo y prescriptivo, así como tecnologías de machine learning y otras que están cambiando la forma en que los datos son administrados y analizados. El data warehouse ha sido durante mucho tiempo un elemento básico de la arquitectura de datos empresarial y sigue teniendo un gran futuro en el nuevo mundo de Big Data y la analítica avanzada.

3. Data Warehouse en la nube

Los data warehouses están atravesando actualmente dos transformaciones muy importantes que tienen el potencial de impulsar niveles significativos de innovación empresarial:

La primera área de transformación es el impulso para aumentar la agilidad general. La gran mayoría de los departamentos de TI están experimentando un rápido aumento de la demanda de datos. Los directivos quieren tener acceso a más y más datos históricos, mientras que, al mismo tiempo, los científicos de datos y los analistas de negocios están explorando formas de introducir nuevos flujos de datos en el almacén para enriquecer el análisis existente, así como impulsar nuevas áreas de análisis. Esta rápida expansión de los volúmenes y fuentes de datos significa que los equipos de TI necesitan invertir más tiempo y esfuerzo asegurando que el rendimiento de las consultas permanezca constante y necesitan proporcionar cada vez más entornos para equipos individuales para validar el valor comercial de los nuevos conjuntos de datos.

La segunda área de transformación gira en torno a la necesidad de mejorar el control de costes. Existe una creciente necesidad de hacer más con cada vez menos recursos, al mismo tiempo que se garantiza que todos los datos sensibles y estratégicos estén completamente asegurados, a lo largo de todo el ciclo de vida, de la manera más rentable.

La nube está demostrando ser un facilitador dominante. Permite a las organizaciones enfrentarse de forma activa a los desafíos que presentan estas dos transformaciones clave.

¿Por qué migrar a la nube?

Las 3 tecnologías que ya tienen y seguirán teniendo un gran impacto en la transformación de negocios son: la nube, datos y análisis, e Internet de las Cosas. Estas tres tecnologías son parte clave del ecosistema de la data warehouse actual. Por lo tanto, es posible sacar la conclusión de que el data warehouse en la nube tiene un gran potencial para generar un impacto significativo en los negocios.

La importancia de la nube para el data warehouse está directamente vinculado a tres factores clave:

Mejora de la agilidad

Muchos data warehouses están embarcados en una fase de actualización. Con mucho trabajo puesto ahora en torno al big data, las empresas buscan aprovechar los nuevos flujos de datos y nuevos tipos de análisis más ricos para apoyar e impulsar nuevas áreas, tales como:analítica de clientes de 360º, anaĺisis predictivo, detección de fraude, análisis de IoT y el establecimiento de los datos como centro de beneficio. Muchos de estos proyectos requieren el aprovisionamiento de nuevos entornos de hardware y el despliegue de software. Es más rápido, más fácil y más eficiente iniciar estos nuevos proyectos centrados en los datos utilizando los servicios en la nube.

Mayor control de costes

Muchos equipos de TI están buscando maneras de consolidar los data marts existentes, cada uno ejecutándose en hardware dedicado o incluso en hardware propietario, en un único entorno integrado. Los servicios en la nube de calidad ofrecen la oportunidad perfecta para iniciar este tipo de proyectos. El cumplimiento legal no puede ser visto como un extra opcional cuando se planifica un traslado a la nube. Los activos de datos deben estar protegidos a lo largo de todo el ciclo de vida. Los servicios en la nube deben facilitar todo esto, siendo más rentables ya que todas las características de seguridad se pueden habilitar de forma predeterminada y mejorada, y actualizada de forma transparente.

Co-Localización para una carga más rápida

La mayoría de los data warehouses generan datos directamente desde aplicaciones clave, como entrada de pedidos, ventas, finanzas y fabricación. Por lo tanto, tiene mucho sentido ubicar conjuntamente el data warehouse junto con los sistemas fuente que ya se estén ejecutándose en la nube. La ubicación conjunta ofrece una carga de datos más rápida, lo que significa que los usuarios obtienen un acceso más oportuno a sus datos.

Principales ventajas de mover el data warehouse a la nube

Existen tres ventajas principales para mover un data warehouse a la nube, y estas están directamente vinculados a los tres controladores clave enumerados anteriormente:

Más fácil consolidación y racionalización

Monetización más rápida de los datos en la nube

La nube ofrece mejor protección

**Características:**

**Integración de Datos:** Un Data Warehouse recopila y almacena datos provenientes de múltiples fuentes (bases de datos, archivos, sistemas externos) en un formato unificado. Esto facilita la comparación y el análisis de datos de diferentes orígenes.

**Orientado a Sujetos:** Organiza los datos en torno a áreas específicas de interés, como ventas, finanzas, o clientes, en lugar de estar centrado en las aplicaciones de negocio.

**Histórico:** A diferencia de las bases de datos transaccionales, que suelen almacenar datos actuales, un Data Warehouse guarda un historial de datos, permitiendo el análisis de tendencias a lo largo del tiempo.

**No Volátil:** Una vez que los datos se ingresan en un Data Warehouse, no se modifican ni se eliminan. Esto asegura la consistencia y la integridad de la información almacenada para análisis futuros.

**Optimizado para Consulta:** Está diseñado para realizar consultas y análisis complejos de manera eficiente, soportando herramientas de Business Intelligence (BI) que permiten la generación de reportes y análisis avanzados.

**Gran Volumen de Datos:** Está preparado para manejar y procesar grandes cantidades de datos, lo que es esencial para empresas que manejan mucha información histórica o de diversas fuentes.

**Diseño en Estrella o Copo de Nieve:** Los datos en un Data Warehouse se organizan comúnmente en esquemas en estrella o copo de nieve, donde una tabla de hechos está rodeada por varias tablas de dimensiones que describen los atributos de los datos.