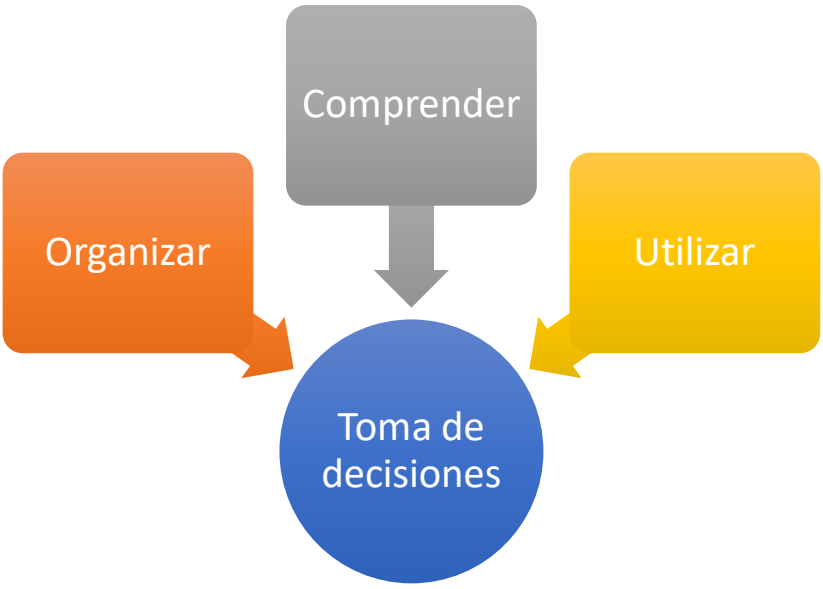
Lección 4: Data Warehouse (DW)

Un data Warehouse consiste básicamente en una bodega electrónica donde una empresa guarda una gran cantidad de información compuesta por los datos que se reúnen a través de las diferentes fuentes existentes, y que busca dejarlos a disposición del usuario para fines analíticos.

El concepto como tal surgió en 1988 a raíz del trabajo presentado por los investigadores de IBM, Barry Devlin y Paul Murphy, no obstante, fue William H. Inmon quien lo acuñó y se dio a conocer como el padre de Data Warehousing, este definió el DW como una compilación de datos en un tema específico, integrado y variable que apoya el proceso de la toma de decisiones.

Normalmente, el DW se encuentra en un servidor empresarial o en la nube, los datos allí alojados proceden de diversas aplicaciones de procesamiento de transacciones online (OLTP) y demás fuentes que son extraídas minuciosamente para su posterior uso. Este sistema permite a los empresarios organizar, comprender y utilizar los datos en pro de la toma de decisiones estratégicas.



El sistema de data warehouse se puede dividir en tres estructuras:

- ✓ Básica: que se compone de sistemas operativos y archivos planos que generan datos en bruto que son almacenados con metadatos (datos que describen otros datos) para que los usuarios tengan acceso a los mismos y tengan la posibilidad de generar análisis, informes e incluso minería.

- ✓ Básica con un área de ensayo: consiste en añadir el área de ensayo entre las fuentes de datos y el almacén como tal, así de logra obtener un espacio en el que los datos se puedan limpiar antes de que se integren al almacén.

- ✓ Básica con un área de ensayo y data marts: en cuyo caso se diseña un sistema para una línea de negocio específica, como ventas, inventario o compras.

Anteriormente, los data warehouse se formaban a partir de datos repetitivos y estructurados que se depuraban antes de su entrada en el DW, y que no podían ser mezclados con fines analíticos con datos textuales no estructurados, pero recientemente y gracias a su evolución, el data warehouse permite adjuntar datos no estructurados e información contextual de forma sencilla.

Los datos no repetitivos basados en textos generados con palabra escrita, hablada o leída como los comentarios en una encuesta, correos electrónicos o conversaciones tienen un tratamiento distinto a los datos repetitivos como el flujo de clics o las mediciones. Este tratamiento permite extraer el sentido de la información gracias a el análisis del contexto.

En el universo del Data Warehouse, se ha filtrado otro concepto denominado Data Lake, que lejano a ser un reemplazo para el data warehouse, es concebido como un respaldo o complemento. Es así entonces como se presentan algunas diferencias clave entre ambos conceptos.

- ✓ Datos: Data warehouse simplemente almacena datos que fueron previamente estructurados, mientras que un data lake almacena todo tipo de datos, ya sean estructurados, semiestructurados o no estructurados.

- ✓ Procesamiento: el enfoque del data warehouse en cuanto al procesamiento de datos discrepa del data lake, este primero realiza la estructura de los datos previo al ingresos de los mismos, mientras que data lake lo realiza después, cuando el usuario deba utilizar los datos.

- ✓ Almacenamiento: data warehouse presenta costos elevados en comparación con data lake, gracias a que este último posee mayoritariamente softwares de código abierto.

- ✓ Agilidad: un data lake no posee la estructuración en el almacén de datos que data warehouse si posee, y que permite configurar y reconfigurar los modelos y aplicaciones de forma ágil y sencilla.

- ✓ Seguridad: la madurez del data warehouse permite que su capacidad para asegurar y proteger los datos sea mayor en comparación con un data lake que es relativamente nuevo.

Teniendo en cuenta el proceso de transformación continuo en los data warehouse, se observa como estos importantes avances tienen la capacidad de impulsar distintas áreas en la innovación empresarial. Una de estas áreas de transformación esta relacionada con la agilidad general. Puesto que muchos de los departamentos atraviesan un rápido incremento en la demanda de datos, se hace necesaria la inversión de tiempo y esfuerzo para asegurar que el rendimiento de las solicitudes sea constante.

Por otro lado, otra área de transformación se relaciona con la necesidad de mejorar el control de los costes debido a que cada vez se hace mas recurrente que las empresas quieran hacer más, pero destinando menos recursos.

Dicho lo anterior, surge una posibilidad que permite que la empresa afronte ambas áreas de forma acertada. La nube, junto con los datos y el análisis y el internet de las cosas, se presenta como una solución para hacer frente a la transformación de las empresas.

La nube se vincula con el data warehouse a través de tres aspectos fundamentales:

Mayor agilidad	• Aprovechamiento de los nuevos flujos de datos • Tipos de análisis más eficientes • Facilidad para iniciar proyectos
Mayor control de costes	• Consolidación de datos existentes • Protección de datos más rentable
Co-localización	• Ubicación del data warehouse con las fuentes • Carga de datos más rápida

Después de observar las áreas de transformación y su posible mejoramiento con el uso de los servicios en la nube, se pueden resumir las ventajas de la utilización del sistema en tres puntos:

1. Fácil consolidación y racionalización de datos
2. Rápida monetización de los datos en la nube
3. Mejor protección de datos

El concepto de Data Warehouse se encuentra estrechamente relacionado con los conceptos de Big Data y de Business Intelligence. Los tres tipos de tecnologías comparten la forma en que el usuario maneja los datos, teniendo en cuenta los volúmenes y los formatos allí presentes. A fin de cuentas, el objetivo de todas estas es brindar al usuario una ventaja competitiva.

Toda empresa debe realizar un auto análisis para definir si un almacén de datos como data warehouse es la solución a algunas de las problemáticas que se pueden estar experimentando. Antes que nada, la empresa debe plantear el escenario, la complejidad que generan las operaciones asociadas al almacén y cual es su volumen de datos.

Si el volumen de datos se vuelve de alguna forma abrumador para el usuario, se podría plantear la idea se acceder a un data warehouse. Cuando las hojas de calculo se convierten en una herramienta insatisfactoria y el proceso de almacenamiento se vuelve cada vez mas tedioso se dificulta ampliamente el gobierno de datos.

Si una empresa afronta el hecho de que las hojas de calculo estén siendo utilizadas por casi todos los departamentos de la empresa, además existan varios propietarios, tienen la necesidad de trabajar sobre bases de datos mixtas con todo tipo de datos, generan informes manualmente y el volumen de la información crece de manera progresiva, es hora entonces de pensar en un almacenamiento de datos centralizado, con esto se lograría agilizar el reporting, reducir los tiempos de espera y obtener una única versión final.

Otro concepto importante es el de sistema de **planificación de recursos (ERP).** Dado que cada departamento de la empresa cuenta con un sistema propio, es necesario que se integre cada uno de ellos en un solo lugar, es así como los ERP sirven de intermediario a través de distintos módulos para que cada departamento realice su labor propia, sin dejar de estar toda la información en un mismo lugar.

Los datos que se encuentran en el ERP son fundamentalmente de tipo operacional o transaccional, por lo que no tienen la capacidad de ofrecer un análisis de tendencias o visiones del mercado, su utilidad resulta ser netamente operativa.

Referencias:

Alam Khan, F., Ahmad, A., Imran, M., Alharbi, M., Ur-rehman, M., & Jan, B. (2017). Efficient data access and performance improvement model for virtual data warehouse. Sustainable cities and society, 35, 232-240. Recuperado el 2 de agosto de 2018, de https://www.sciencedirect.com/science/article/pii/S2210670717302706

Microtech. (s.f.). Microtech. Recuperado el 2 de agosto de 2018, de https://www.microtech.es/blog/qu%C3%A9-diferencias-hay-entre-software-erp-y-software-bi-business-intelligence

PowerData. (2016). PowerDataq. Recuperado el 2 de agosto de 2018, de https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/que-es-un-data-warehouse-y-como-saber-cuando-lo-necesitas-implementar

PowerData. (s.f.). PowerData. Recuperado el 2 de agosto de 2018, de https://www.powerdata.es/data-warehouse