Lección 2.2 Análisis OLAP y Data Warehouses:

Los almacenes de datos, también llamados Data Warehouses, surgieron en la década de los 90 debido al aumento en el volumen de datos que las empresas necesitaban procesar para sus tareas de Business Intelligence. Este aumento dió lugar a una clara diferenciación entre las bases de datos de los sistemas transaccionales, denominados OLTP, y las analíticas, también llamadas OLAP.

1. El análisis OLAP:

- El análisis OLAP se basa en el análisis de grandes cantidades de datos de forma sencilla
    - Por ejemplo, cada vez que queremos consultar información acerca de las ventas del último mes estamos extrayendo los datos de todas las transacciones que han tenido lugar en los últimos 30 días.
    - El tiempo de respuesta es un factor crucial, necesitamos la información para tomar una decisión, por ejemplo si es necesario o no elaborar una campaña de marketing. La decisión necesitamos tomarla ahora, no dentro de otros 30 días.
    - Y no sólo eso, sino que después de obtener la respuesta inicial, lo más habitual es que queramos realizar otras preguntas que requieran volver a extraer datos. Por ejemplo, quienes son los clientes que han realizado un mayor número de compras durante estos 30 días.
- Las bases de datos OLTP no son adecuadas para tareas analíticas
    - Están optimizadas para la escritura de datos, no para la lectura. Al fin y al cabo, la empresa necesita que registrar todas las ventas que están ocurriendo simultáneamente durante el día
    - Las tareas de análisis requieren datos que se encuentran distribuidos por numerosas tablas en los sistemas OLTP, y la unión de los registros junto con el volumen de datos requerido pueden bloquear el sistema
    - Son difíciles de consultar debido a que la información se encuentra distribuida a través de múltiples bases de datos y de tablas, lo que requiere de la creación de consultas a la base de datos que pueden llegar a ser muy extensas, del orden de múltiples páginas para unas sola consulta

Por ello, la aproximación más habitual es que el análisis OLAP se sustente en una base de datos separada, que almacena la información de forma integrada y que se encuentra diseñada para facilitar el análisis. Esta base de datos es el almacén de datos.

***2. El almacén de datos*:**

- El almacén de datos es una base de datos integrada y diseñada para simplificar consultas y procesarlas de forma más eficiente
    - La estructura conceptual se diseña desde el punto de vista analítico siguiendo el modelado multidimensional para simplificar las consultas
    - El modelo lógico de tablas se organiza con el objetivo de optimizar las lecturas en lugar de las escrituras
    - La información almacenada es el resultado de la integración de las distintas fuentes de datos disponibles en la empresa y de las necesidades de análisis de toda la empresa
    - El almacén puede estar dividido en secciones, llamadas Data Marts, que dan respuesta a las necesidades de información de un departamento de la empresa en concreto
    - Dentro de cada Data Mart podemos la información se encuentra estructurada en cubos. Cuando estos cubos se cargan en memoria para su análisis, se denominan cubos OLAP y permiten al usuario navegar a través de la información, entrando en detalle en los datos o agregando la información.

Para comprender el funcionamiento del almacén de datos veamos más en detalle cómo se organiza su estructura.

3. El modelado multidimensional del almacén

- La información del almacén de datos se organiza de acuerdo al modelo multidimensional, propuesto inicialmente por Kimball. Existen dos conceptos básicos a diferenciar en la información que utilizamos para el análisis:
    - Por un lado, los hechos. Un hecho representa la ocurrencia de un evento relacionado con alguno de los procesos de la empresa. Por ejemplo, una compra por parte de un cliente en un supermercado, un encuentro de un partido de fútbol entre dos equipos en un campo y con una determinada fecha, ó la suscripción a un curso online por parte de una persona en algún lugar del mundo. Normalmente, los hechos van acompañados de medidas que permiten evaluar el rendimiento del proceso de la empresa. Por ejemplo el total de la compra, o el resultado del encuentro, aunque no siempre es así, como en el caso de la

suscripción a un curso. En ciertos casos simplemente es relevante tener constancia de que un evento ha ocurrido y de quiénes han sido los participantes de ese evento.
    - Por otro lado, las dimensiones. Una dimensión representa la información de contexto de un hecho, agrupada en la forma de una entidad identificable. Por ejemplo, a la hora de realizar una compra tenemos una serie de productos comprados, un cliente que realiza la compra, un cajero que atiende al cliente, etc. Esta información es muy útil a la hora de realizar análisis ya que permite segmentar los datos. Por ejemplo, analizar todas las compras de un cliente, o las ventas que realiza un supermercado en concreto. Como podemos ver, un hecho puede tener muchas dimensiones asociadas, y se pueden compartir dimensiones entre varios hechos.
- Además de estos conceptos básicos, de cara al usuario resulta importante explicar otro concepto más: las jerarquías de dimensión.
    - Una jerarquía representa las distintas formas posibles de agrupar las entidades de una dimensión.
    - Por ejemplo, piensa en los yogures de tu supermercado. Tienes yogures de distintos sabores, yogur natural, yogur de fresa, yogur de plátano/banana. Cada uno de estos yogures es un elemento distinto, pero todos pertenecen al mismo tipo de producto, todos son yogures. Si quisieras ver que sabor es el preferido por tus clientes consultarías los productos individuales, mientras que si quisieras saber si los yogures se están vendiendo bien consultarías el tipo de producto.
    - En realidad lo que estás haciendo es organizar en tu mente una jerarquía de la dimensión de producto, con un nivel de mayor detalle, que serán los productos individuales y un nivel agregado, que será el tipo del producto.

Como podemos ver, estos conceptos tan básicos permiten realizar análisis de forma muy potente siendo muy concisos. Combinar la información de varias dimensiones permite por ejemplo analizar de forma rápida qué cual es el rango de edad de tus clientes que te está comprando un tipo de producto, como los yogures. Después, podemos comprobar si algún producto individual de ese grupo se está vendiendo mal. Quizá el sabor a limón no gusta demasiado en una zona del país.

Pero podemos incluso hacer este análisis más sencillo y más rápido mediante las operaciones de navegación OLAP.

4. Las operaciones de navegación OLAP

- Las operaciones de navegación OLAP tratan de simplificar el análisis de datos, permitiendo pasar de un conjunto de datos a otro sin tener que especificar una consulta detrás de otra. Las operaciones de navegación vienen implementadas de serie en los servidores OLAP de los principales vendedores de BI y son las siguientes:
    - Roll-up: La operación de roll-up agrega una dimensión de los datos, pasando del nivel actual al nivel siguiente más agregado. Por ejemplo, si estamos analizando los datos de productos individuales podemos rápidamente hacer roll-up en la dimensión producto para ver los productos por tipo.
    - Drill-down: La operación contraria a roll-up, entra en detalle en una dimensión si hay algún nivel más detallado que el actual. Al contrario que antes, pasaríamos de analizar los productos por tipo a analizar los productos individuales.
    - Pivot: Hacer pivoting consiste en modificar la vista que tenemos de los datos, normalmente cambiando las dimensiones y medidas de posición. Por ejemplo, es posible que queramos organizar los territorios en las columnas y los productos en las filas, ya que tenemos muchos productos y nos simplifica el análisis de los datos.
    - Slice&Dice: La última operación de navegación OLAP consiste en filtrar los datos para obtener un subconjunto de los datos que estamos analizando. Por ejemplo si estamos analizando los productos pero sólo queremos ver qué ocurre con nuestros yogures, todo lo demás es ruido para nosotros, por lo que filtramos el cubo para que muestre sólo los datos referentes al tipo yogures.

Con esto hemos terminado de ver cómo se diseña y cómo funciona el núcleo del sistema de BI donde se ejecutan gran parte de las operaciones de análisis, pero esto no es todo.