Conceptos Básicos de OLAP

Autor
Martín Vladimir Alonso Sierra Galvis

Gestión de Datos, Maestría en Ciencia de Datos
Pontificia Universidad Javeriana Cali

Versión 2.0
Santiago de Cali, marzo de 2023

Tabla de contenido

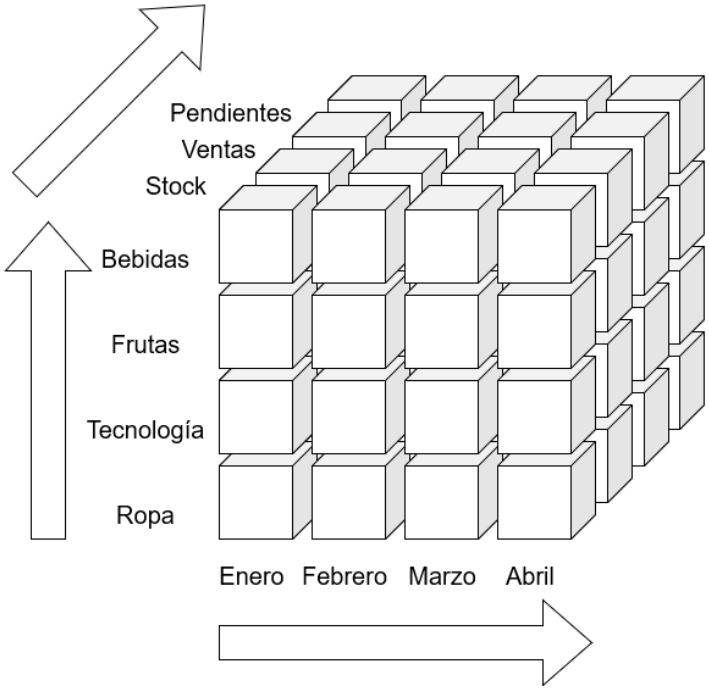
Estructuras para Análisis de Datos

Consideraciones

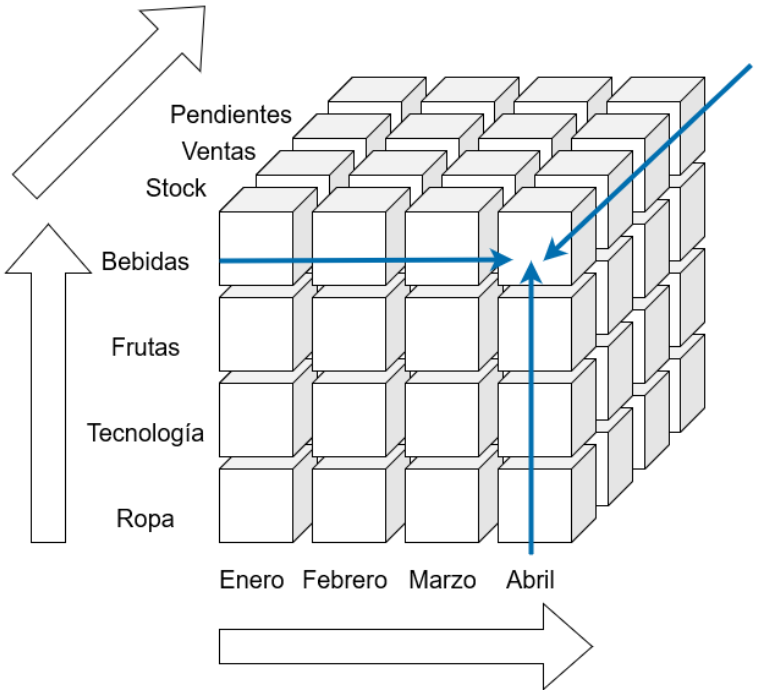
Antes de aprender cuál es el proceso que debemos seguir para implementar Bases de Datos OLAP, es importante que tengamos claros algunos conceptos básicos relacionados, como lo son los Modelos Multidimensionales, los Cubos OLAP, los Data Marts y los Data Warehouses.

Modelos Multidimensionales y Cubos

Cuando hablamos de una Estructura Multidimensional estamos haciendo referencia a determinada variante del Modelo Relacional que usa múltiples dimensiones para la organización de los datos. Si quisiéramos hacer una analogía de esta organización con algún elemento del mundo real que nos permitiera entenderla más fácil, podríamos decir que la figura que más se parece a una estructura de este tipo es el cubo. Cada cara del cubo es una dimensión y estas, a su vez, están conformadas por celdas, representadas por pequeños cubos. Cada una de estas celdas contienen datos agregados, relacionados con las diferentes dimensiones del cubo. Observemos la siguiente ilustración.



La imagen muestra los datos de un supermercado, divididos en las tres dimensiones que conforman el cubo. Cada una de estas dimensiones, identificadas con las flechas, representan distintos datos agregados: la dimensión cuya flecha apunta en sentido horizontal a la derecha contiene datos de fechas en meses; la dimensión cuya flecha apunta en sentido vertical hacia arriba contiene datos relacionados a las categorías de los productos; finalmente, la dimensión representada por flecha diagonal superior contiene datos de diferentes variables vinculadas a las ventas de los productos. Si un científico de datos o analista quisiera averiguar el número de **bebidas** que permanecieron en el **stock** del supermercado durante el mes de **abril**, entonces podría obtener los datos así.

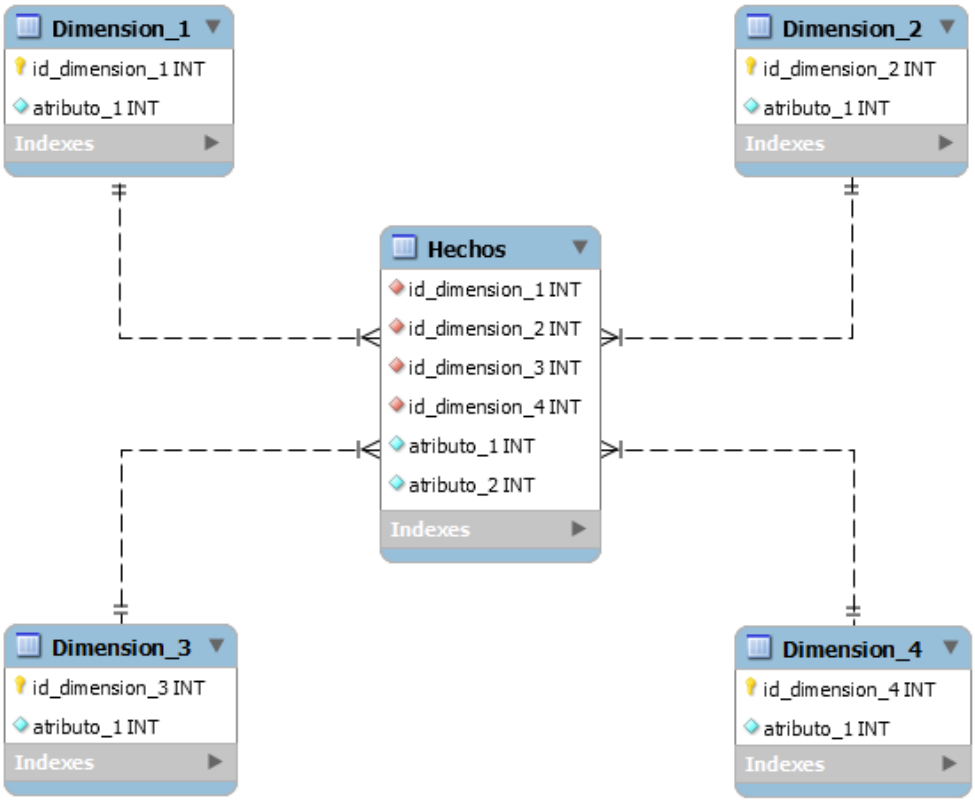


Uno de los beneficios de representar los datos de esta manera es su fácil visualización y rápida comprensión. Además, las consultas a realizar pueden llegar a ser más simples que las Estructuras Relacionales usadas, generalmente, en las Bases de Datos OLTP. Así mismo, una característica importante es que las Estructuras Multidimensionales facilitan la organización de los datos de manera jerárquica. Por ejemplo, en el cubo de la imagen, los datos se encuentran divididos por meses pero, si utilizaramos el concepto de **jerarquías**, podríamos organizarlos en trimestres, semestres e incluso en años si así lo quisiéramos. Por todo lo anterior, la Estructura Multidimensional de Cubo, llamada comúnmente como Cubo OLAP, es la más popular para trabajar con bases de datos analíticas que requiera respuestas rápidas a operaciones complejas de análisis de datos.

Ahora bien, al igual que pasaba con las Bases de Datos Relacionales, las Bases de Datos con Estructuras Multidimensionales se implementan a partir de modelos, denominados Modelos Multidimensionales. Los principales son el Modelo Estrella, el Modelo Copo de Nieve y el Modelo Constelación. Veamos en qué consiste cada uno.

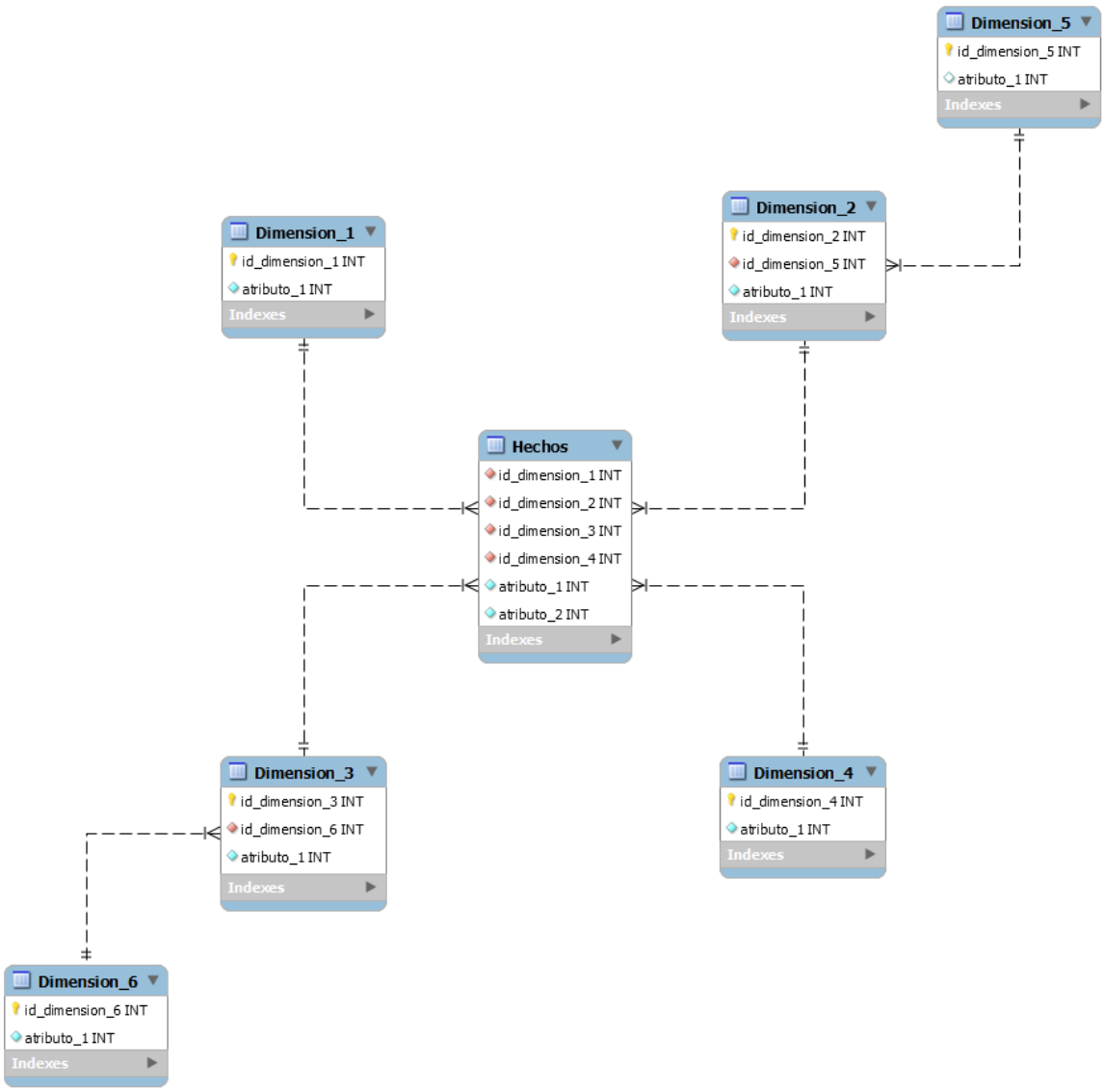
Modelo Estrella

El Modelo Estrella se basa principalmente en una tabla de hechos central que representa un proceso de negocio que se quiera analizar. Esta tabla de hechos está relacionada con otras tablas satélites, llamadas tablas de dimensiones, encargadas de representar puntos de vista que ayudan a analizar el proceso de negocio. Se le llama Modelo Estrella pues su diagrama tiene la forma de una estrella. Es el modelo más utilizado por su facilidad de comprensión y por poseer la optimización en tiempos de búsqueda y acceso a datos más alta. El diagrama general de un Modelo Estrella es el siguiente.



Modelo Copo de Nieve

El Modelo Copo de Nieve organiza los datos utilizando jerarquías de dimensiones. En este sentido, puede verse como una extensión del Modelo Estrella, donde se tiene una tabla de hechos central, relacionada con las tablas de dimensiones que, a su vez, pueden relacionarse con otras tablas de dimensiones. Aunque este modelo hace un mejor manejo del espacio de almacenamiento, pues trata de evitar la redundancia de datos, lo cierto es que el desempeño de las consultas puede ser menor debido a que se debe obtener los datos de diferentes dimensiones. El diagrama general de un Modelo Copo de Nieve es el siguiente.



Modelo de Constelación

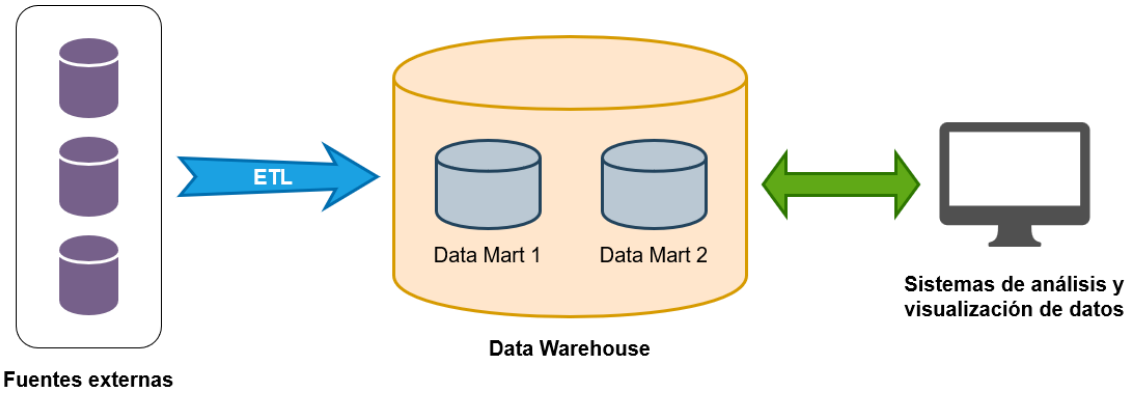
El Modelo de Constelación es, básicamente, una agrupación de varios Modelos de Estrella. Esto trae la ventaja de poder analizar más aspectos y procesos clave del negocio con un único Modelo Multidimensional. A pesar de esto, muy pocas herramientas de consulta y análisis de datos, por no decir ninguna, soportan este modelo. El diagrama general de un Modelo de Constelación es el siguiente.

Los Modelos Multidimensionales nos ayudarán a implementar las Estructuras Multidimensionales físicas de las Bases de Datos OLAP, conocidas como Data Marts. Veamos, a continuación, qué es un Data Mart y cuál es su diferencia con un Data Warehouse.

Data Marts y Data Warehouses

El Data Warehouse o Bodega de Datos es, en la actualidad, uno de los elementos más importantes para realizar Analítica de Datos y Business Intelligence. Es aquí donde se almacenan todos los datos que una organización obtiene de sus diferentes fuentes u orígenes de datos. Los datos se organizan en una estructura especial conformada por fragmentos independientes, denominados **Data Marts**. Estos Data Marts se encargan, a su vez, de almacenar los datos en una estructura física de un Modelo Multidimensional.

En otras palabras, un Data Warehouse puede contener Data Marts que organizan los datos siguiendo un Modelo Estrella. Gracias a esta forma de organizar los datos, se pueden realizar diversas operaciones complejas como generación de reportes, análisis de datos mediante Cubos OLAP, minería de datos, entre otras. La siguiente gráfica muestra un esquema general de un Data Warehouse.



Es importante que entendamos que un Data Warehouse no es un software específico ni una única base de datos gigante. En realidad el concepto hace referencia al conjunto de tecnologías que permiten a un profesional experto en datos, como un Científico de Datos, gestionar datos organizados de forma multidimensional, para que un ejecutivo, un administrador o un analista puedan tomar decisiones de manera más rápida y fácil. Así pues, un Data Warehouse es totalmente independiente de un Gestor de Bases de Datos. En realidad, un Data Warehouse lo que hace es alimentarse de los datos almacenados en diferentes fuentes de datos. A estas fuentes se les conoce como **fuentes externas** de datos y para alimentar la Bodega de Datos con ellas es necesario aplicar un Proceso ETL. El Proceso ETL es una herramienta que podremos utilizar para extraer datos de las fuentes externas de origen, transformar esos datos conforme a la estructura de la fuente destino y, finalmente, cargar los datos transformados en dicha fuente destino, que en este caso, serían los Data Marts del Data Warehouse. De ahí que las siglas ETL hagan referencia a las palabras **E**xtract, **T**ransform y **L**oad. Si quisiéramos una definición técnica de lo que es un Data Warehouse, podríamos decir que este es una **"colección de datos orientada a temas, integrada, no volátil y variable en el tiempo, que se usa principalmente en una organización para la toma de decisiones"**. De esta definición podemos enumerar las características de un Data Warehouse:

1. **Orientado a temas** relevantes para una organización pues está especialmente diseñado para consultar, de manera eficiente, datos que se relacionan con los procesos de negocio claves de la misma.

2. Sus datos son **integrados** pues se obtienen a partir de diferentes fuentes externas de datos que alimentan a la organización.
3. La naturaleza **no volátil** hace referencia a la no modificación de los datos almacenados. Estos solamente se incrementan.
4. Es **variable en el tiempo** pues al cambiar constantemente los datos de las fuentes externas, deben ser integrados con regularidad.

En contraste con las Bases de Datos OLTP, cuyas estructuras están enfocadas a la ejecución de operaciones transaccionales diarias como lecturas, registros, modificación y eliminación de datos, los Data Marts y, por ende, los Data Warehouse están enfocados a almacenar información histórica, resumida y consolidada, organizada de tal manera que permita un rápido acceso a una gran cantidad de datos con el objetivo de analizarlos y generar reportes y visualizaciones. Es por esta razón que, las Bases de Datos OLTP, al tener registros consistentes y poco redundantes, ocupan espacios del orden de los gigabytes, mientras que las Bases de Datos OLAP, al preferir la redundancia de datos, pueden llegar a ocupar terabytes de espacio. Finalmente, las Bases de Datos OLAP, como los Data Marts, están diseñadas y estructuradas de tal manera que permiten consultas muy complejas que de utilizarse en Bases de Datos OLTP presentarían problemas de rendimiento. Si alguna vez, como Científicos de Datos, vamos a implementar un Data Warehouse en un ambiente de producción, es recomendable que dicho Data Warehouse quede separado de las bases de datos transaccionales de la organización de forma que no se afecte el rendimiento del sistema de almacenamiento.

Referencias

1. Auribox Training. *¿Qué es un Data Warehouse | Business Intelligence*. Junio de 2017. [En línea] disponible en https://www.youtube.com/watch?v=jFsRdTcljeU. [Accedido en 2020].
2. Calle Sánchez, D. A. *Manual para el diseño e implementación de bases de datos OLAP y su aplicación en inteligencia de negocios*. 2009. Trabajo de grado para optar por el título de Ingeniero de Sistemas. Escuela de Ingeniería. Universidad EAFIT. Departamento de Ingeniería de Sistemas. Medellín, Colombia. [En línea]: disponible en https://core.ac.uk/download/pdf/47240196.pdf. [Accedido en 2020].
3. Chaudhuri, S y Dayal U. *An Overview of Data Warehousing and OLAP Technology*. Marzo de 1997. ACM Sigmod Record. [En línea]: disponible en https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/sigrecord.pdf. [Accedido en 2020].
4. CompuEL Cursos. *Cubo OLAP*. Mayo de 2017. [En línea]: disponible en https://www.youtube.com/watch?v=5NEuHT75G0Q. [Accedido en 2020].
5. Ibarra, M. de los A. *Procesamiento Analítico en Línea (OLAP)*. 2006. Trabajo de adscripción para la materia Diseño y Administración de Datos. Licenciatura en Sistemas de Información. Universidad Nacional del Nordeste. Facultad de Ciencias Exactas, Naturales y Agrimensura. Argentina. [En línea]: disponible en http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/OLAPMonog.pdf. [Accedido en 2020].