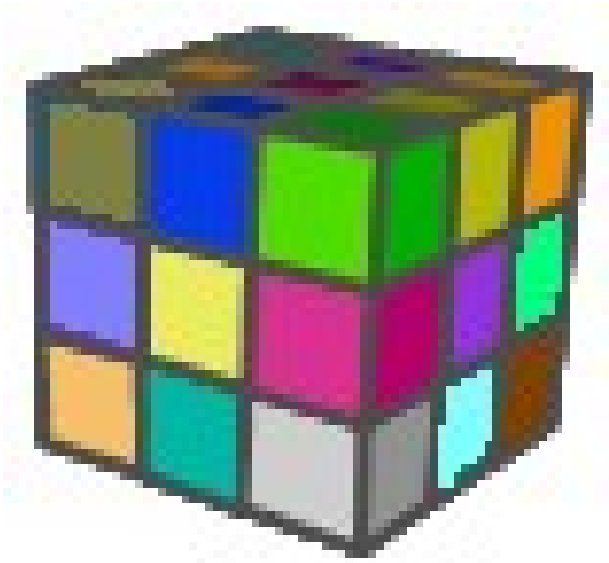
Módulo
Minería de Datos
Diplomado

Por
Elizabeth León Guzmán, Ph.D.
Profesora
Ingeniería de Sistemas
Grupo de Investigación MIDAS



Análsis Dimensional

OLAP

On-Line Analytical Processing

Estructura del Proceso de Acceso a Datos y Entrega de Información en BI

Data Quality

Fuentes de datos

Calidad de datos

ETLC

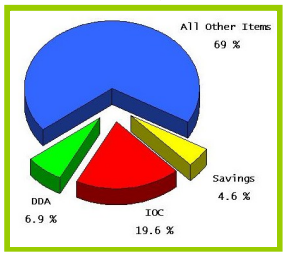
Producto
Cliente
Geografía
Empleado
Data Warehouse

OLAP

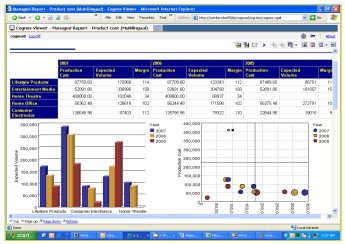
Diseño y Construcción Modelos OLAP

Modelos Molap / Rolap

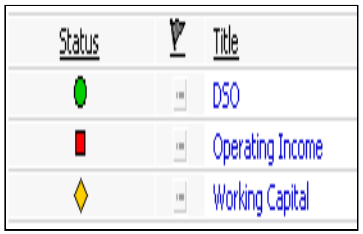
Metadata Integrada para Acceder Bases de Datos y Fuentes OLAP



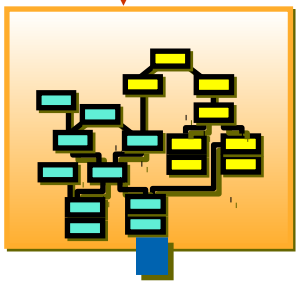
Análisis OLAP Agregaciones



Reportes y Consultas



Score Cards



Minería de datos

OLAP

- On-Line Analytical Processing
- Técnica de Análisis Multidimensional
- Diseñado para lograr un buen rendimiento en consultas ad-hoc
- Vista Multidimensional de los datos
- Mecanismo para almacenar un **cubo**
- Puente entre como los datos están almacenados en la bodega y en como son presentados al usuario

OLAP

Fácil de usar por los analistas del negocio

- Navegar en los Datos
- Velocidad de las consultas
- Esconde complejidad
- Riqueza analítica

OLAP permite mas fácilmente:

- Analizar datos
- Generar reportes
- Acceder los datos por navegadores de web
- Visualizar datos
- Importar datos

Creación de una matriz multidimensional

- Dos pasos clave en la conversión de los datos tabulares en una matriz multidimensional.
- En primer lugar, identificar qué atributos deben ser las **dimensiones** y que es atributo de ser el atributo de destino cuyos valores aparecen como entradas en la matriz multidimensional.
  - Los atributos utilizados como dimensiones deben tener valores discretos

Creación de una matriz multidimensional

- El valor objetivo es típicamente un recuento o valor continuo, por ejemplo, el costo de un elemento, precio de venta

- En segundo lugar, encontrar el valor de cada entrada de la matriz multidimensional mediante la suma de los valores (del atributo de destino) o recuento de todos los objetos que tienen los valores de los atributos correspondientes a esa entrada.

Cubo

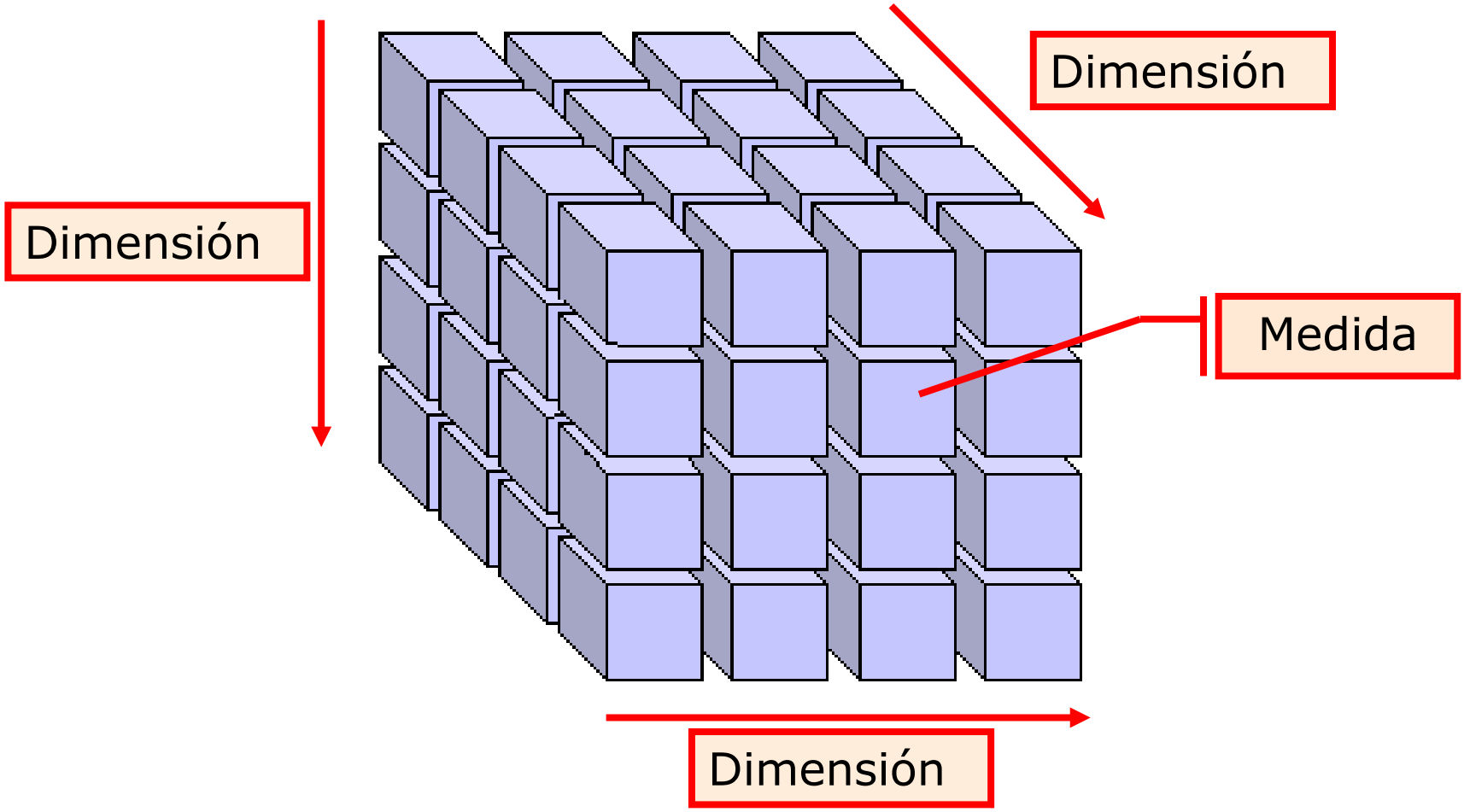
Contiene datos de primer interés para los usuarios

Es un subconjunto de los datos que están en la bodega. Contiene valores **agregados** a todos los **niveles** de las dimensiones

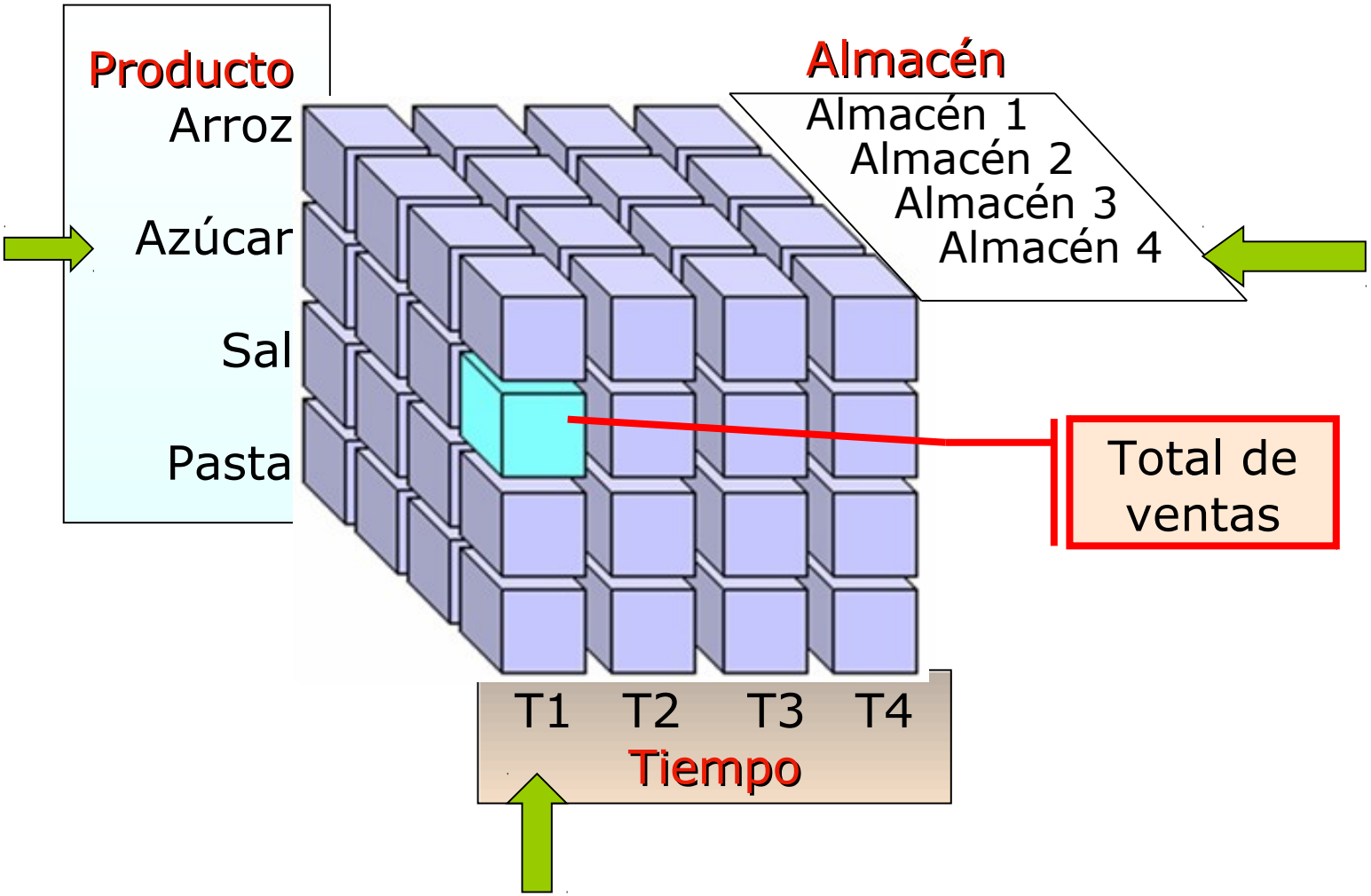
Usados para organizar los datos en dimensiones y medidas

Mejoran la velocidad de consulta

Cubo



Cubo de Ventas



¿Cual fue el total de ventas de azúcar en el almacén 4 durante el tiempo T1?

Cubo

El cubo puede responder preguntas que incluyan tres dimensiones y una medida

- Dimensión producto: contiene categorías del producto
- Dimensión almacén: contiene almacenes
- Dimensión tiempo: contiene periodos del año
- Medida ventas: cantidad numérica que puede ser sumarizada

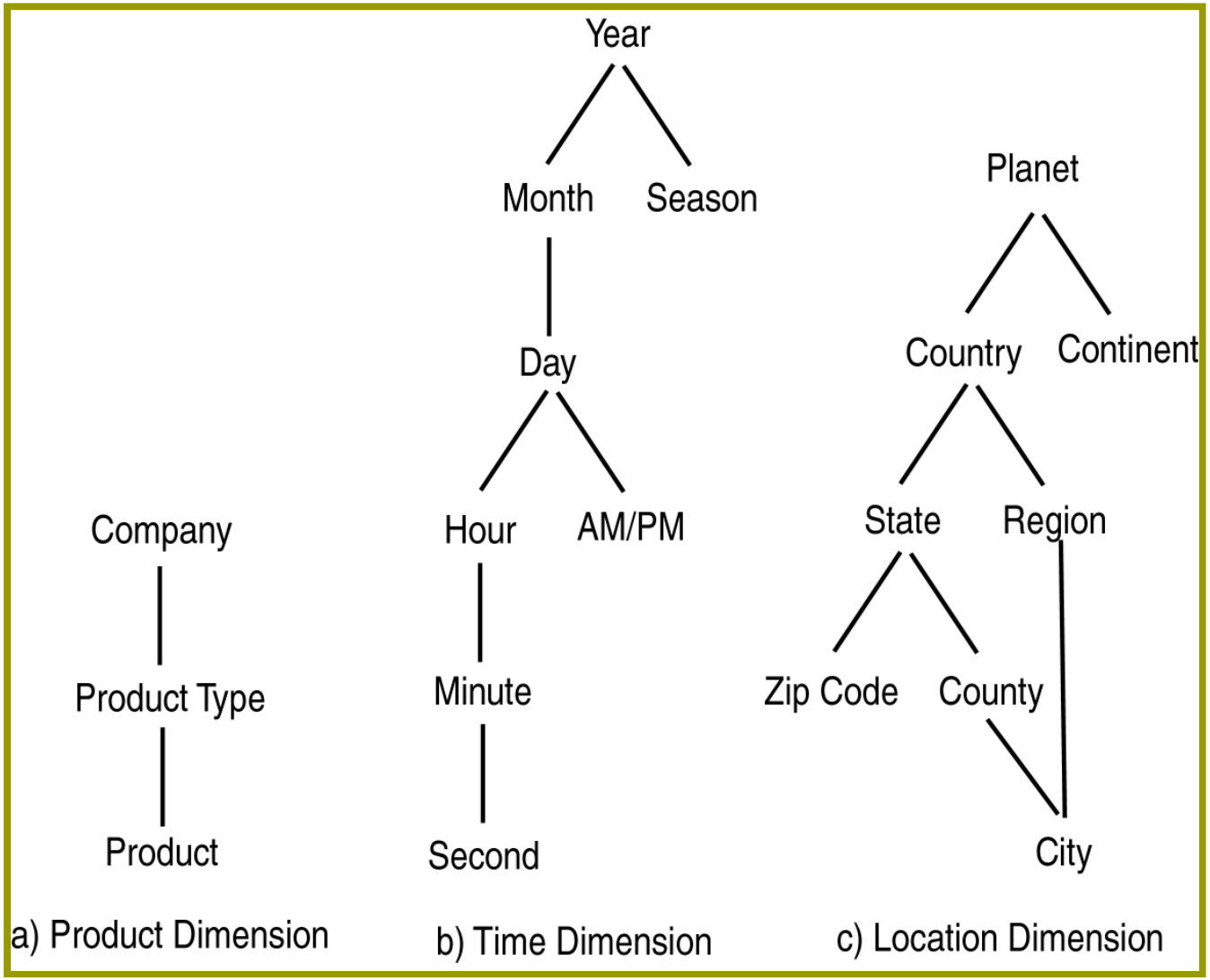
Cubo

Un cubo puede tener hasta 64 dimensiones

- Cada celda del cubo tiene un valor
- El valor de cada celda es la intersección de las dimensiones
- El dato en la celda es una agregación

Para obtener las ventas totales anuales por producto y localización: seleccionar el producto y la localización y **suma** por las cuatro celdas de tiempo

Jerarquía de agregación

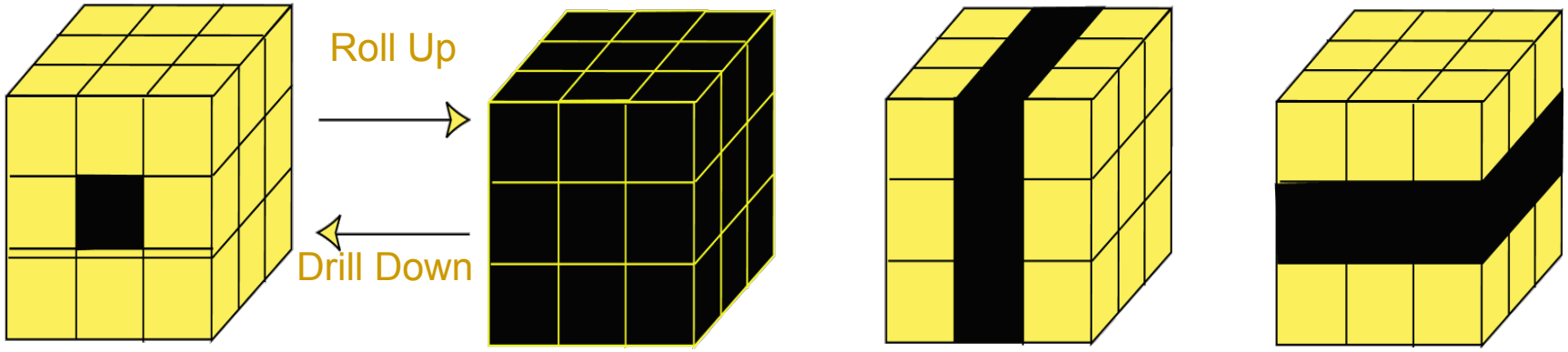


Tomado de [5]

Operaciones OLAP

Roll Up: dimensiones generales

Drill Down: dimensiones especificas

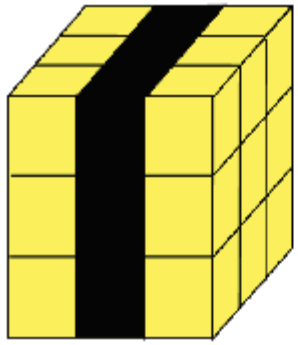


Single Cell Multiple Cells Slice Dice

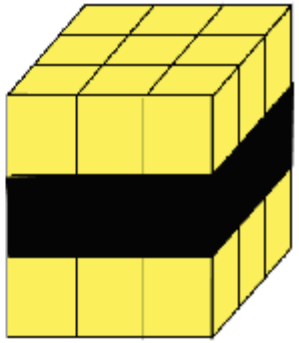
OLAP Operaciones

- "Slicing" es la selección de un grupo de celdas de la matriz multidimensional especificando un valor para una o más dimensiones.

- "Dicing" rotar el cubo para mirar otras dimensione



slice



dice

- Esto es equivalente a la definición de una submatriz de la matriz completa.

- En la práctica, ambas operaciones también puede ser acompañado por la agregación sobre algunas dimensiones.

OLAP Operaciones: Roll-up y drill down

- **Valores de los atributos a menudo tienen una estructura jerárquica.**

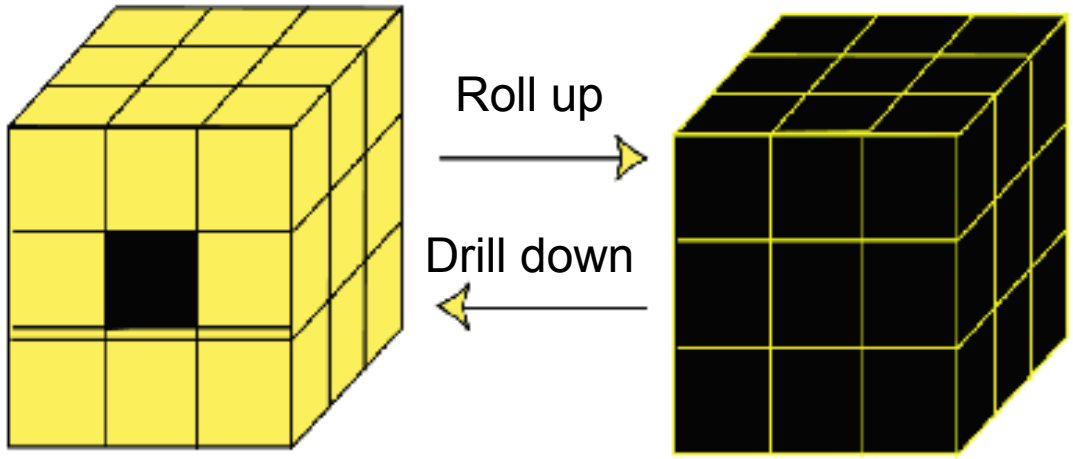
- Cada día se asocia a un año, mes y semana.
- Una localización se asocia con un continente, país, estado (provincia, etc), y la ciudad.
- Los productos se pueden dividir en varias categorías, tales como ropa, electrónica y muebles.

- **Tenga en cuenta que estas categorías suelen anidar y la forma de un árbol o red.**

- Un año contiene meses, que contiene días.
- Un país contiene un estado que contiene una ciudad.

OLAP Operaciones: Roll-up y drill down

- **Esta estructura jerárquica da lugar a la enrollable y perforador de fondo de operaciones.**



- Para los datos de ventas, podemos agregado (enrollar) las ventas a través de todas las fechas en un mes.

- Por el contrario, dado un punto de vista de los datos, donde se rompe la dimensión del tiempo en meses, podríamos dividir los totales de ventas mensuales (drill down) en los totales de ventas diarias.

- Asimismo, se puede profundizar o rodar sobre la ubicación o identificación de los atributos del producto.

Ejemplo: Conjunto de Datos Iris

- Se muestra cómo los atributos: longitud, ancho del pétalo, y el tipo de especie, se pueden convertir en una matriz multidimensional
  - En primer lugar, discretizar la anchura y la longitud del pétalo de tener valores categóricos: bajo, medio y alto.
  - Recibimos la siguiente tabla - tenga en cuenta el atributo de número

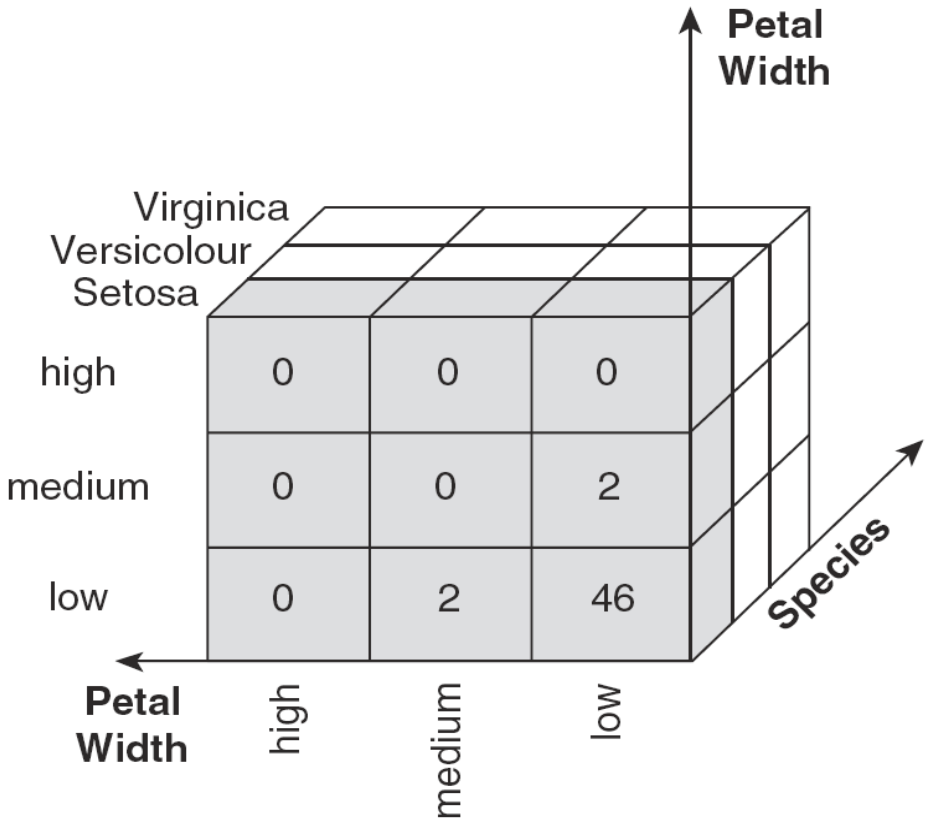
Petal Length	Petal Width	Species Type	Count
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44

Ejemplo: Iris de datos (continuación)

- Cada tupla única: ancho de pétalo, longitud de pétalo, y el tipo de las especies identifica un elemento de la matriz.

- Este elemento se le asigna el valor de recuento correspondiente.
- La figura ilustra el resultado.

- Todo aquello no especificado tuplas son 0.



Ejemplo: Iris de datos (continuación)

- Las rebanadas de la matriz multidimensional se muestran los siguientes tabulaciones cruzadas

- ¿Qué nos dicen estas tablas?

		Width		
		low	medium	high
Length	low	46	2	0
	medium	2	0	0
	high	0	0	0

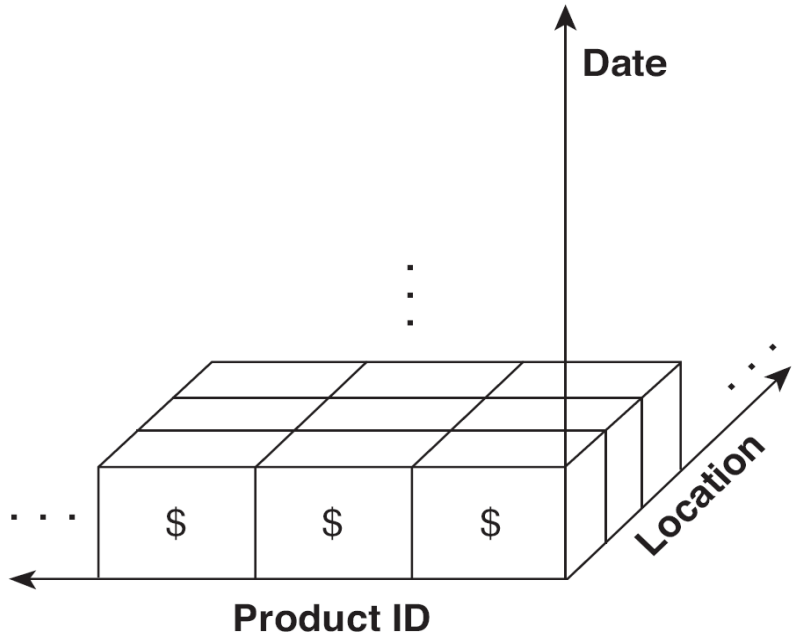
		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	43	3
	high	0	2	2

		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	0	3
	high	0	3	44

Ejemplo de datos de cubos

- Conjunto de datos que registra las ventas de productos en una serie de tiendas de la compañía en fechas diferentes.
- Estos datos pueden ser representados como una matriz de 3 dimensiones.

- 3 bidimensional agregados
- 3 unidimensionales agregados
- 1 cero-dimensional agregado (el total)



Ejemplo de datos de cubos (continuación)

- Agregaciones por dos dimensiones (valores en las casillas), una dimensión (totales de columnas y filas) y cero dimensiones (total general)

product ID \ date	Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	total
1	$1,001	$987	...	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	...	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	...	$631,221	$227,352,127

MOLAP y ROLAP

Las arquitecturas de los Bodegas han evolucionado en base al crecimiento, en funcionalidad y rendimiento, de los DBMS.

Es por ello, que la arquitectura reinante de los ’80 fue el modelo **MOLAP** (**Multidimensional OLAP**).

En los ’90, surge una nueva arquitectura de DW denominada **ROLAP** (**Relational OLAP**)

MOLAP

Capacidad de análisis

- Ofrece **vistas** de objetos **multidimensionales**
- Tiempo de **respuesta cero**, pues tiene todo precalculado.
- Si no se precalcula todo (en general todo el precálculo tiene volúmenes inaceptables) la capacidad de análisis se limita a aquellas porciones del cubo que fueron precalculadas.

MOLAP

- Existen variantes de los MOLAP donde en caso de no poder responder a una consulta, se precalcula el cubo que responde a esa pregunta y a la periferia, pudiendo llevar esta generación varias horas de construcción (generalmente batch).

MOLAP

Sistema de diseño propietario

- Generalmente el cubo se trata de una *"caja negra"* de datos encriptados que pueden residir de forma local o en un servidor MOLAP.
- Flexibilidad y escalabilidad **limitados**.
- Cambios en el modelo dimensional del negocio implican la **generación** de todos los cubos nuevamente.

MOLAP

Ambientes adecuados

- Modelos dimensionales **pequeños** y **estáticos**.
- Instalaciones dónde el tiempo de respuesta sea crítico.
- **Pocos** volúmenes de datos.
- Análisis de información a **nivel agregado**.

ROLAP

Capacidad de análisis

- Ofrece **vistas** de objetos **multidimensionales.**
- Tiempos de respuestas que rondan entre los **segundos** y los **minutos**.
- Existen **técnicas de tuning**, **caching**, **materialización de vistas**, **indexación** y **esquema de diseño** que mejoran la performance de respuesta de los ROLAP.

ROLAP

Sistema de diseño abierto

- El cliente interactúa **directamente** contra el RDBMS vía **SQL** en distintos motores.
- Provee **flexibilidad** y **escalabilidad**.
- Los cambios en el modelo dimensional del negocio son trasladados al DW e **inmediatamente** se encuentra disponible para las consultas pertinentes.
- La ventana de carga del data warehouse es menor pues no existe el tiempo de generación de los multi-cubos.

ROLAP

Ambientes adecuados

Modelos dimensionales **grandes** y **dinámicos**.

Grandes volúmenes de datos.

Necesidad de análisis a **nivel transaccional**.

OLAP

Demos y presentaciones:

- Pentaho:

  http://demo.pentaho.com/pentaho

Referencias

[1] Wiley - Mastering Data Warehouse Design - Relational And Dimensional Techniques – 2003.

[2] Wiley - Data Analysis -The Data Warehouse Toolkit - Second Edition.

[3] Wiley - Building The Data Warehouse - Third Edition

[4] Wiley - The Data Warehouse ETL Toolkit -2005.

[5] Wiley - The Data Warehouse Lifecycle Toolkit 1998

[6] MicroStrategy - Business Intelligence - 2006.

[7] Data Mining: Introductory and Advanced Topics by Margaret Dunham, Publisher: Prentice Hall – 2006

[8] Introduction to Data Mining. Tan, Steinbach, Kumar. 2006