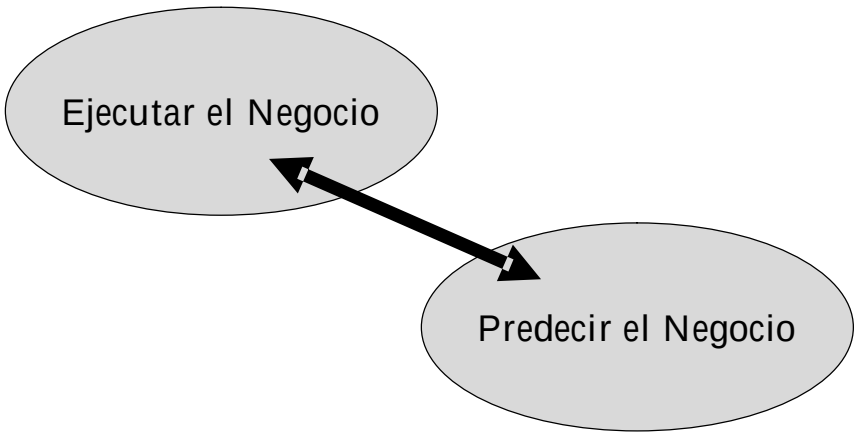
Introducción al Datawarehousing

Necesidades de Negocio



Caso

- Un supermercado tiene 20 sucursales.
- Cada sucursal tiene en promedio 50 cajas
- Cada caja atiende un promedio de 1 cliente cada 5 minutos durante 10 horas.
- Un cliente compra 30 artículos promedios.
- Cada sucursal tiene su propia base de datos donde se registran las operaciones.

- ¿ Cuántos registros de ítems vendidos se generan en un día ?
- Rta: 20 sucursales x 50 cajas x (10 horas x 60 min) / 5 x 30 = 3.600.000 registros
- ¿ Cuántos registros de ítems vendidos se generan en un año ?
- Rta: 3.600.000 x 360 = 1.296.000.000
- ¿ Y en los últimos 10 años ?
- Rta: 1.296.000.000 x 20 = 12.960.000.000

Un directivo quiere saber cuál hacer una comparación de la evolución de la venta de distinas marcas de cervezas durante los últimos 3 años.

Un directivo quiere predecir cuál será la evolución de la venta de cervezas en el futuro observando el comportamiento pasado.

¿ Como se soporta la información que permita responder estas preguntas ?

Solución

- Parece razonable almacenar la información histórica en un sistema separado y específicamente orientado a responder estas preguntas.

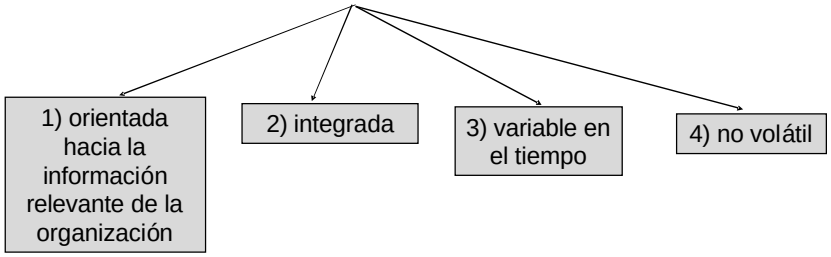
DATAWAREHOUSING

Definición de un Datawarehouse

Data Warehouse

definición

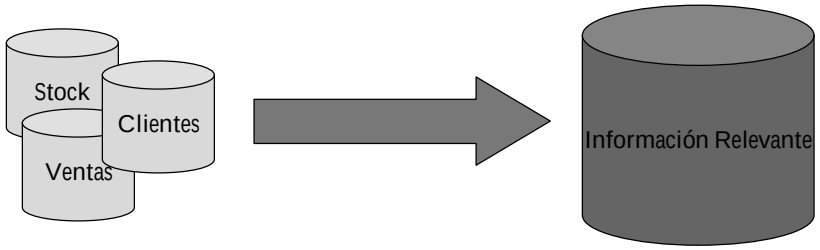
colección de datos diseñada para dar apoyo a los procesos de toma de decisiones



características

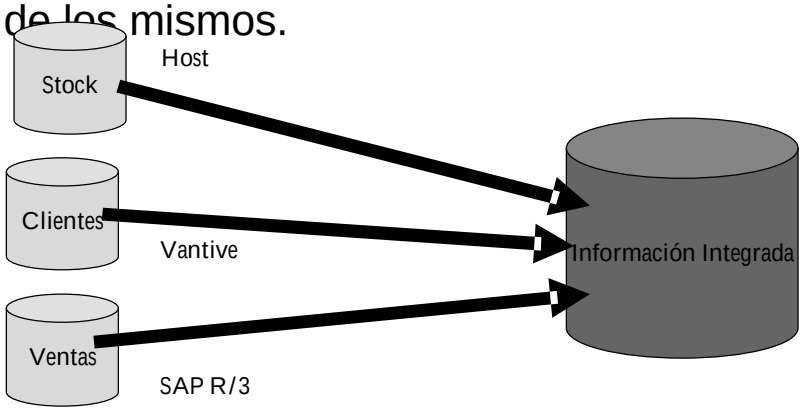
1) Orientada hacia la información relevante de la organización

- El diseño está orientado a la consulta eficiente de la información.
- No soporta el negocio, sino la toma de decisiones.



2) Integra información de distintos sistemas operacionales

- Provee una vista de los datos de la organización, más allá del soporte físico de los mismos.

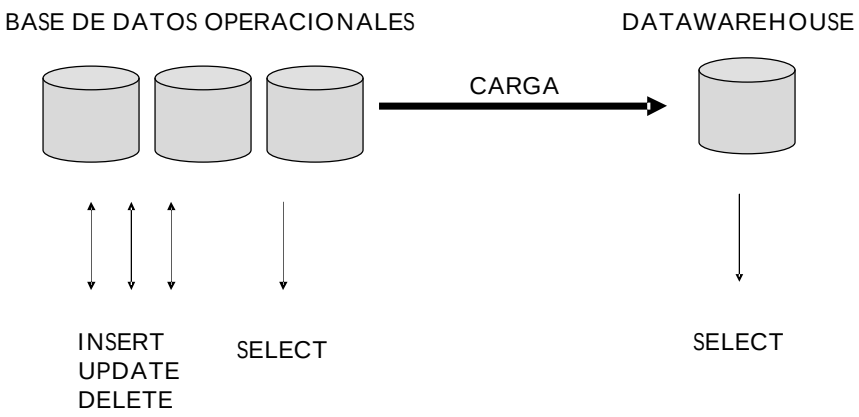


3) Variable en el tiempo

- Los datos son actualizados e incrementados periódicamente.
- Los datos son almacenados como fotos (snapshots) en el tiempo.

Mes	Ventas
200401	Ventas de Enero
200402	Ventas de Febrero
200403	Ventas de Marzo

4) No volátil



Los datos son incrementados y almacenan información histórica.

OLTP vs OLAP

- OLTP: on-line transaction processing

Son las aplicaciones que le dan soporte a la operatoria de la compañía.

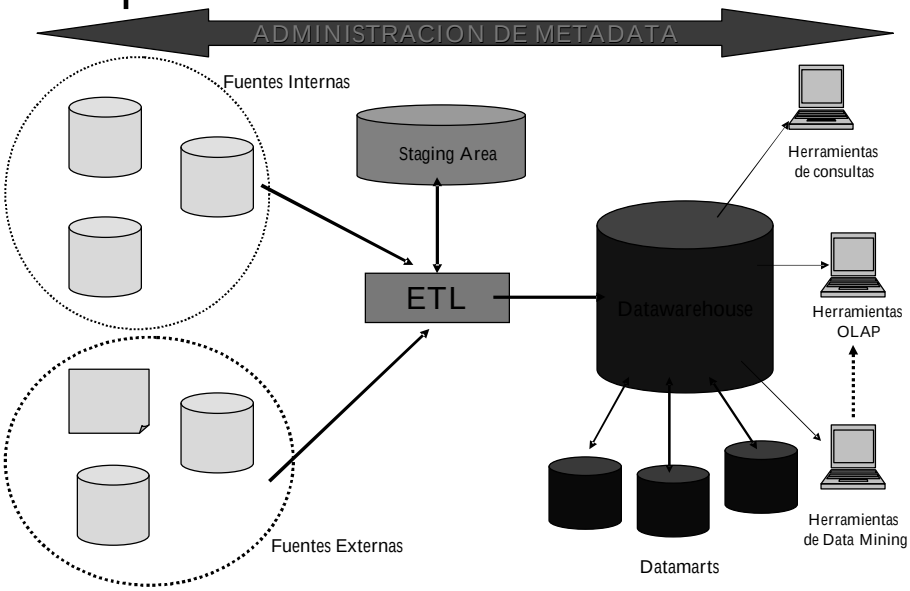
- OLAP: on-line analytical processing

Son las aplicaciones que analizan el negocio y de apoyo a la toma de decisiones (Decision Support Systems – DSS -)

OLTP vs OLAP

Característica	OLTP	OLAP
Dedicado a	Procesar transacciones	Análisis de datos
Tiempo de los datos	Actuales	Históricos
Tamaño de las DB	Medianas (100Mb – 1Gb)	Grandes (100 Gb – 1 Tb)
Granularidad de los datos	Datos de Detalle	Datos de detalle y agregados
Volatilidad de los datos	Estáticos	Dinámicos
Decisiones que soportan	Operativas	Estratégicas
Número de usuarios	Alto (operativos)	Bajo (directivos)
Tiempo de respuesta	Bajo	Variable (segundos a horas)
Orientado a	Procesos de la organiz.	Información relevante
Tipo de procesos	Repetitivos y previsibles	Distintos y no previsibles

Arquitectura de un Datawarehouse



Arquitectura de un Datawarehouse

Organización Externa de los Datos

Las herramientas de explotación de los Data Warehouses han adoptado un **modelo multidimensional de datos**.

Se ofrece al usuario una visión multidimensional de los datos que son objeto de análisis.

Arquitectura de un Datawarehouse

EJEMPLO

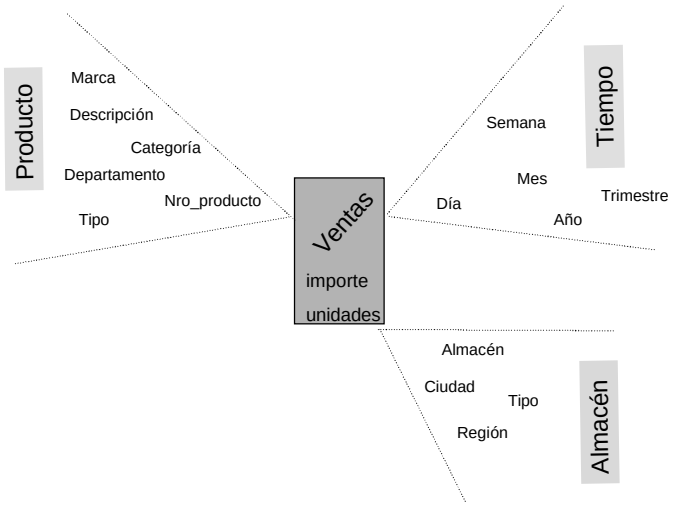
Organización: Cadena de supermercados.

Actividad objeto de análisis: ventas de productos.

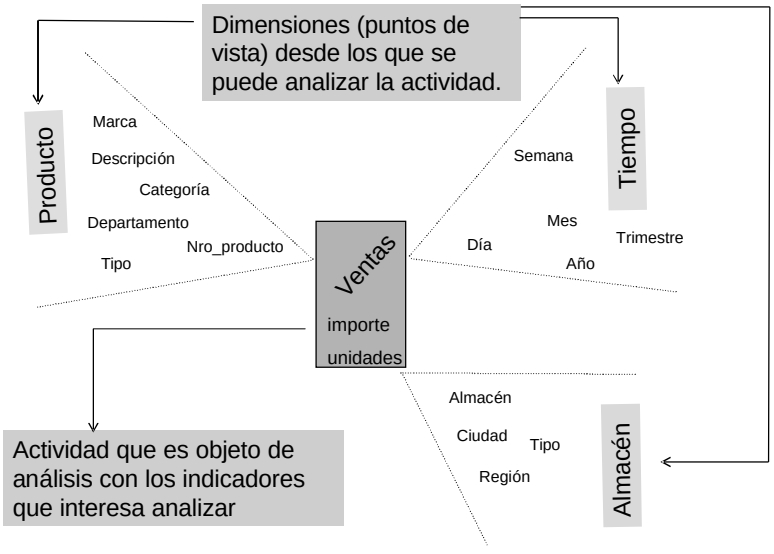
Información registrada sobre una venta: "del producto "Tauritón 33cl" se han vendido en el almacén "Almacén nro.1" el día 17/7/2003, 5 unidades por un importe de 103,19 euros."

Para hacer el análisis no interesa la venta individual (ticket) realizada a un cliente sino las ventas diarias de productos en los distintos almacenes de la cadena.

Arquitectura de un Data Warehouse



Arquitectura de un Data Warehouse



Arquitectura de un Data Warehouse

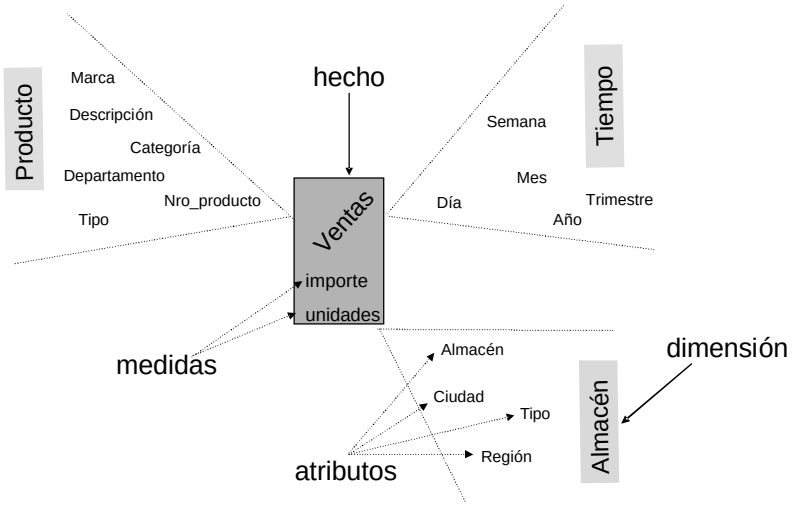
- Modelo multidimensional:

Ø en un esquema multidimensional se representa una actividad que es objeto de análisis (hecho) y las dimensiones que caracterizan la actividad (dimensiones).

Ø la información relevante sobre el hecho (actividad) se representa por un conjunto de indicadores (medidas o atributos de hecho).

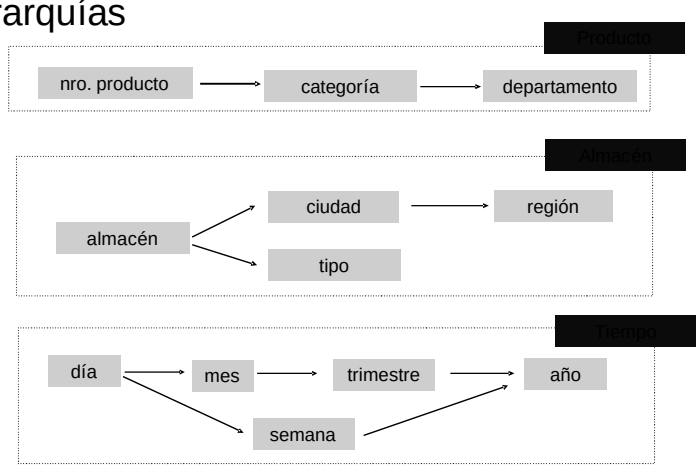
Ø la información descriptiva de cada dimensión se representa por un conjunto de atributos (atributos de dimensión).

Arquitectura de un Data Warehouse



Arquitectura de un Data Warehouse

Entre los atributos de una dimensión se definen jerarquías



Modelado de un Data Warehouse

- Las metas de un Sistema de Soporte de Decisión se logran por medio del Modelado Multidimensional.

- Su principal herramienta es el denominado Esquema Estrella.

Modelado de un Data Warehouse

Esquema Estrella:

- **Diseño de BD con los mejores tiempos de respuesta.**
- **Diseño fácilmente modificable.**
- **Paralelismo entre el diseño de la BD y cómo los usuarios visualizan y usan los datos.**

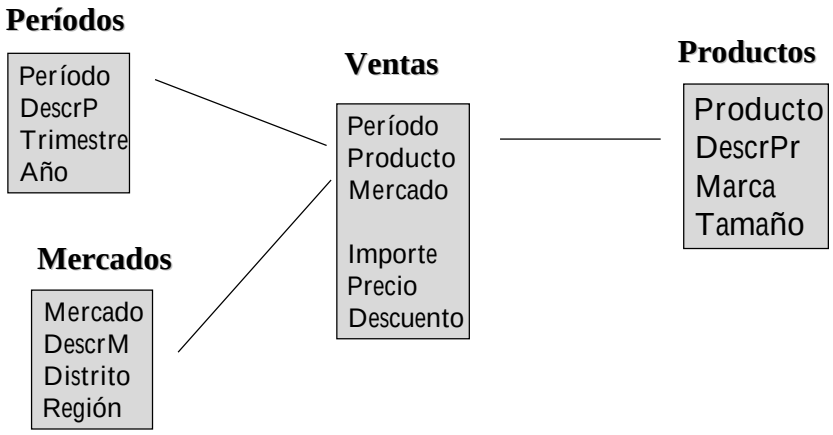
Modelado de un Data Warehouse

¿Por qué no el Modelo E/R?

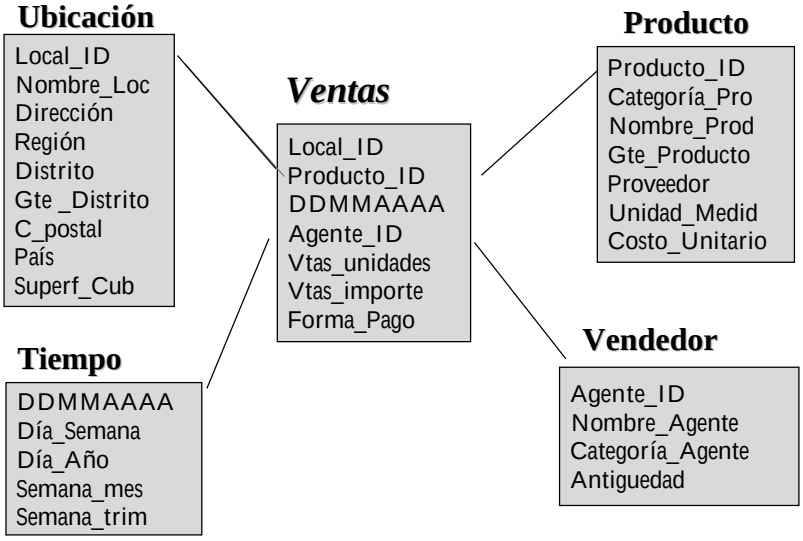
- **Pensado para obtener una BD altamente normalizada, adecuada para Sistemas con muchas transacciones que acceden a un número pequeño de registros.**
- **Adecuado para OLTP, donde los datos son fuertemente estructurados.**
- **En DSS la normalización afecta la eficiencia de las consultas(dificulta el drill-down y el roll-up).**

Modelado de un Data Warehouse

Esquema Estrella Simple: *Ventas*



Modelado de un Data Warehouse



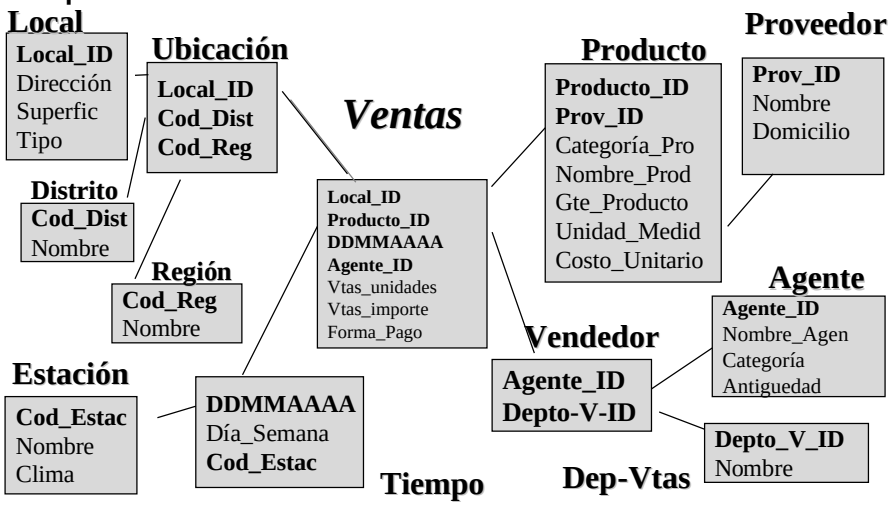
Modelado de un Data Warehouse

Normalización de Dimensiones

- Esquema "snowflake".
- Mayor complejidad en la estructura.
- Mejor performance?
- Mejor uso del espacio.
- Util en tablas de Dimensiones de muchas tuplas.

Modelado de un Data Warehouse

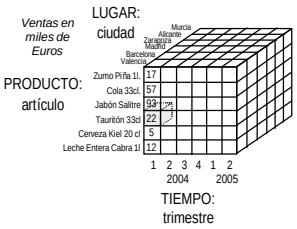
Esquema Snow Flaked: *Ventas*

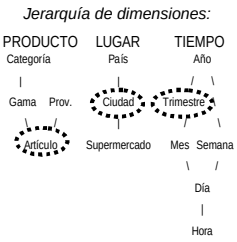


Arquitectura de un Data Warehouse

Se pueden obtener hechos a diferentes niveles de agregación:
obtención de medidas sobre los hechos parametrizadas por atributos de las dimensiones y restringidas por condiciones impuestas sobre las dimensiones

HECHO: *"El primer trimestre de 2004 la empresa vendió en Valencia por un importe de 22.000 euros del producto tauritón 33 cl."*





Un nivel de agregación para un conjunto de dimensiones se denomina cubo.

Arquitectura de un Data Warehouse

El Data Warehouse puede estar formado por varios datamarts y, opcionalmente, por tablas adicionales.

Data Mart ⇨ subconjunto de un Data Warehouse, generalmente en forma de estrella o copo de nieve.

- se definen para satisfacer las necesidades de un departamento o sección de la organización.
- contiene menos información de detalle y más información agregada.

Herramientas OLAP

- Presentan al usuario una visión multidimensional de los datos
- Permite que el usuario pueda formular consultas sobre la Base de Datos seleccionando atributos y abstrayéndose de la estructura interna.
- La herramienta genera la consulta (SQL) a partir de la selección del usuario.

Herramientas OLAP

- El usuario selecciona métricas sobre los hechos.
- La consulta es parametrizada por los atributos de las dimensiones.
- Y filtrada por condiciones de filtro impuestas sobre las dimensiones.

Herramientas OLAP

- Cantidad total de unidades de las ventas de la categoría lácteos por trimestre y marca, durante este año, en las sucursales de la provincia de Córdoba.

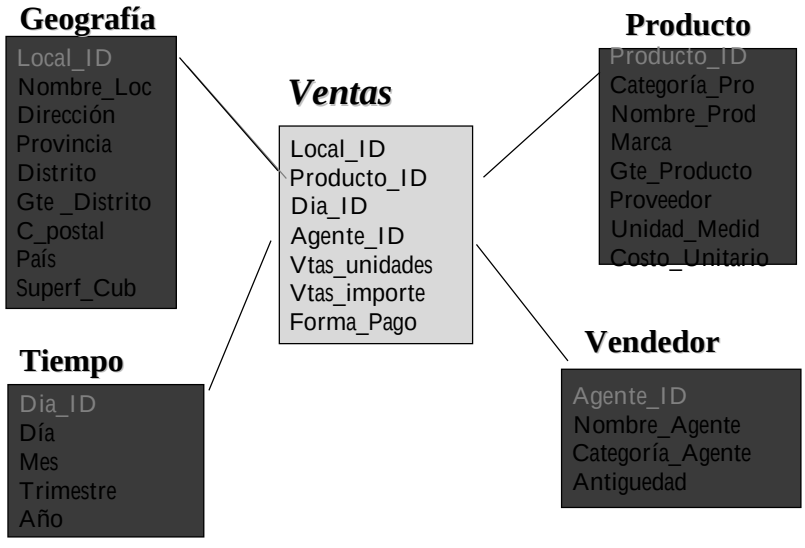
Hechos: Ventas

Métricas: Unidades vendidas

Parámetros de la consulta: por trimestre y marca

Restricciones: este año, sucursales de la provincia de córdoba, de la categoría lácteos.

Herramientas OLAP



Herramientas OLAP

- El usuario selecciona los componentes del query en la herramienta (no tiene que saber SQL)
- La herramienta transforma el reporte solicitado por el usuario en los queries necesarios para responder la consulta.

```
select trimestre, nombre_prod, sum(vtas_unidades) from ventas v,
    tiempo t, producto p, geografía g
where v.local_id = g.local_id
and     v.producto_id = p.producto_id
and     v.dia_id = t.dia_id
and     g.provincia = 'Cordoba'
and     p.categoria = 'Lacteos'
and     t.año = 2004
group by t.trimestre, p.marca
```

Herramientas OLAP

Marca	Trimestre	Unidades
La serenísima	T1	1254
La serenísima	T2	1258
La serenísima	T3	1287
La serenísima	T4	1298
Sancor	T1	327
Sancor	T2	385
Sancor	T3	390
Sancor	T4	402

Presentación tabular de los datos seleccionados

Herramientas OLAP

	T1	T2	T3	T4
La Serenísima	1254	1258	1287	1298
Sancor	327	385	390	402

Presentación matricial (multidimensional) de los datos seleccionados

Herramientas OLAP

- Lo interesante de las herramientas OLAP no es poder hacer consultas tradicionales: selección, proyección, agrupamientos.
- Lo interesante son los operadores adicionales de refinamiento

ü DRILL
ü ROLL
ü SLICE & DICE
ü PIVOT

Herramientas OLAP (Drill)

Drill: permite introducir un nuevo elemento de agrupación en el análisis, disgregando los datos existentes.

A la consulta anterior se agrega el atributo ciudad:

Cantidad total de unidades de las ventas de la categoría lácteos por trimestre, marca y ciudad, durante este año, en las sucursales de la ciudad de Córdoba.

Hechos: Ventas
Métricas: Unidades vendidas
Parámetros de la consulta: por trimestre, marca y ciudad.
Restricciones: este año, sucursales de la provincia de córdoba, de la categoría lácteos.

EL USUARIO NO NECESITA DISEÑAR ESTE NUEVO REPORTE

Herramientas OLAP (Drill)

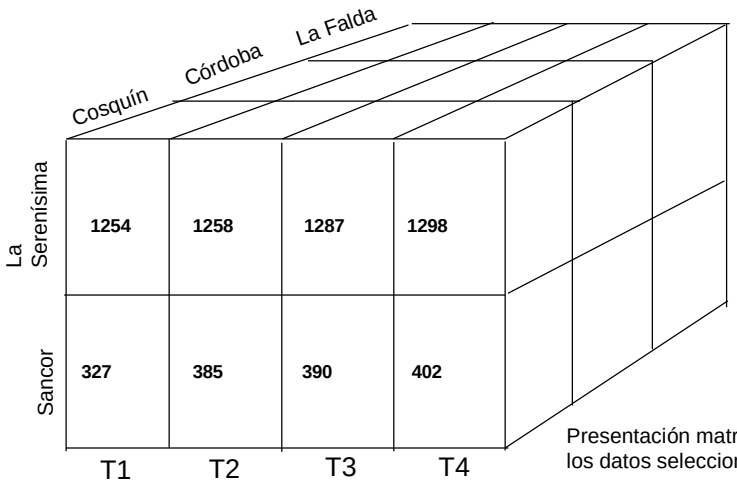
Marca	Trim.	Unidades
La serenísima	T1	1254
La serenísima	T2	1258
La serenísima	T3	1287
La serenísima	T4	1298
Sancor	T1	327
Sancor	T2	385
Sancor	T3	390
Sancor	T4	402

Marca	Trim	Ciudad	Unid.
La serenísima	T1	Cosquín	581
La serenísima	T1	Cord.	275
La serenísima	T1	La Falda	399

DRILL ACCROSS: cada grupo (marca, trimestre) de la consulta original se disgrega en n nuevos grupos (marca, trimestre, ciudad), siendo n la cantidad de ciudades.

Se asumen 3 ciudades: Cosquín, Córdoba y La Falda.

Herramientas OLAP (Drill)



Presentación matricial de los datos seleccionados.

Herramientas OLAP (Roll)

Roll: permite eliminar un elemento de agrupación en el análisis, agregando los datos existentes.

A la consulta anterior se le elimina el atributo trimestre:

Cantidad total de unidades de las ventas de la categoría lácteos por marca y ciudad, durante este año, en las sucursales de la ciudad de Córdoba.

Hechos: Ventas
Métricas: Unidades vendidas
Parámetros de la consulta: marca y ciudad.
Restricciones: este año, sucursales de la provincia de córdoba, de la categoría lácteos.

EL USUARIO NO NECESITA DISEÑAR ESTE NUEVO REPORTE

Herramientas OLAP (Roll)

Marca	Trim	Ciudad	Unid
La serenísima	T1	Cosquín	581
La serenísima	T1	Cord.	275
La serenísima	T1	La Falda	399
La serenísima	T2	Cosquín	576
La serenísima	T2	Cord.	321
La serenísima	T2	La Falda	402

Marca	Ciudad	Unid
La serenísima	Cosquín	1157
La serenísima	Cordoba	596
La serenísima	La Falda	801

ROLL ACCROSS: cada grupo (marca, trimestre, ciudad) de la consulta original se agrega en n nuevos grupos (marca, ciudad), eliminándose la disgregación en trimestre.

Se asumen 3 ciudades: Cosquín, Córdoba y La Falda.

Herramientas OLAP (Roll)

Las operaciones de agregación (ROLL) y disgregación (DRILL) se pueden hacer sobre:

- atributos de una dimensión sobre los que se ha definido una jerarquía: DRILL-DOWN, ROLL-UP
  - departamento – categoría - producto (Producto)
  - año - trimestre – mes - día (Tiempo)
- sobre dimensiones independientes: DRILL-ACROSS, ROLL-ACROSS
  - Producto – Ciudad -Tiempo

Herramientas OLAP(Roll)

Marca	Trim.	Unidades
La serenísima	T1	1254
La serenísima	T2	1258
La serenísima	T3	1287
La serenísima	T4	1298
Sancor	T1	327
Sancor	T2	385
Sancor	T3	390
Sancor	T4	402

DRILL DOWN

ROLL UP

Marca	Trim	Mes	Unid.
La serenísima	T1	Ene	465
La serenísima	T1	Feb	405
La serenísima	T1	Mar	384

DRILL DOWN: cada grupo (marca, trimestre) de la consulta original se disgrega en nuevos grupos (marca, trimestre, mes), donde mes es un elemento de disgregación de trimestre dentro de la jerarquía tiempo.

ROLL UP: cada grupo (marca, trimestre, mes) de la consulta original se agrega en nuevos grupos (marca, trimestre), donde trimestre es un elemento de agregación de mes dentro de la jerarquía tiempo.

Herramientas OLAP (Pivot)

Pivot: permite reorientar a las dimensiones del informe.

UNIDADES VENDIDAS

	Productos	Cosquín	Cordob
T1	La Serenísima	581	385
	Sancor	328	275
T2	La Serenísima	592	368
	Sancor	376	392

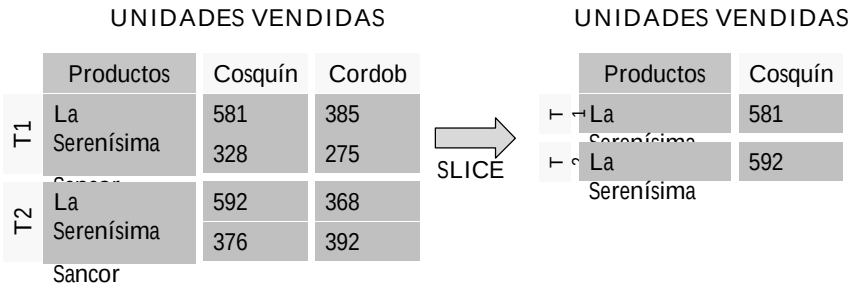
PIVOT

UNIDADES VENDIDAS

	Productos	T1	T2
Cosquí	La Serenísima	581	592
	Sancor	328	376
Córdoba	La Serenísima	385	368
	Sancor	275	392

Herramientas OLAP (Slice)

Slice & Dice: selecciona y proyecta datos del informe.



ROLAP y MOLAP

- El Datawarehouse y las herramientas que lo explotan pueden basarse en cuanto a su estructura de soporte físico de la información, en estas categorías:
- Sistemas ROLAP (OLAP Relacional): se implementan sobre bases de datos con tecnología relacional, las cuales disponen de algunas facilidades para optimizar las consultas.
- Sistemas MOLAP (OLAP Multidimensional): se implementan sobre estructuras de almacenamiento específicas orientadas a consultas y técnicas de compactación de datos (bases de datos multidimensionales).
- Sistemas HOLAP (OLAP Híbridos): utilizan una combinación de las tecnologías ROLAP y MOLAP.

ROLAP

- El datawarehouse se construye sobre un RDBMS relacional.
- Los fabricantes de RDBMS ofrecen extensiones y herramientas para poder utilizar el RDBMS como administrador de un datawarehouse:
  - Ø Indices BITMAPS
  - Ø Indices de JOIN
  - Ø Técnicas de Particionamiento de Datos
  - Ø Optimizadores de Consultas
  - Ø Extensiones a los operadores SQL.

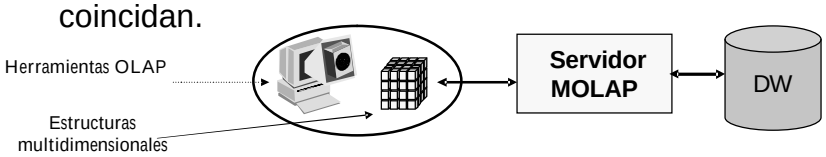
MOLAP

- Son sistemas de propósito específico, que ofrecen:

Ø Estructuras de datos específicas (arrays).

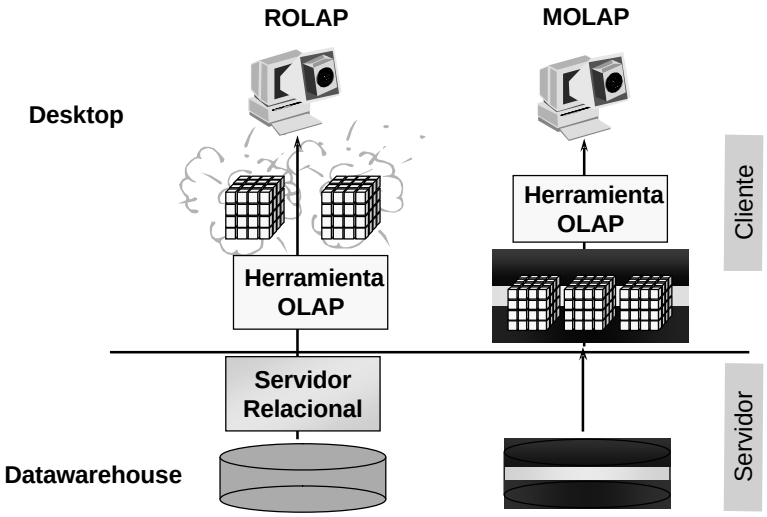
Ø Técnicas de Compactación de Datos

- El objetivo de estos sistemas es almacenar físicamente los datos en estructuras multidimensionales de forma tal que la representación externa y la interna de los datos coincidan.



Servidor MOLAP: construye y almacena datos en estructuras multidimensionales
Herramienta OLAP: presenta estas estructuras multidimensionales

ROLAP y MOLAP



ROLAP y MOLAP

- ROLAP
  - Aprovechan la tecnología relacional.
  - pueden utilizarse sistemas relacionales genéricos (más baratos o incluso gratuitos).
  - tiene mayor escalabilidad
  - por lo general tiene un mejor esquema de Drilling
- MOLAP
  - generalmente más eficientes que los ROLAP en volúmenes chicos y medios de información.
  - por lo general requiere mayor espacio de almacenamiento.
  - menor flexibilidad y escalabilidad que en ROLAP
  - el coste de los cambios en la visión de los datos.
  - la construcción de las estructuras multidimensionales.
  - tiene un mayor costo de administración ya que los cubos deben ser construídos.