


# Explotación y administración de Base de datos

 Juan Carlos Otaegui  
jotaegui@unlam.edu.ar



# Data WareHouse

El almacén de datos



# Introducción

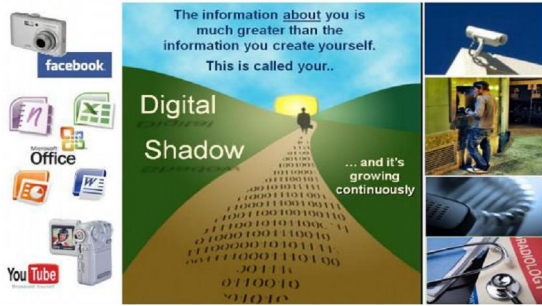
- Reseña histórica: volumen de información en la base de datos y problemas para su análisis

- ▣ Contexto actual

- ▣ Herramienta fundamental para la toma de decisiones

- ▣ Los sistemas transaccionales no tienen el mismo objetivo

# Introducción

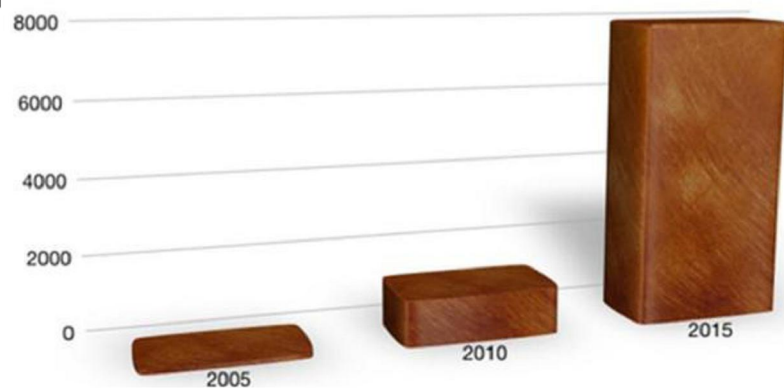


- La información digital en el mundo crece más del doble cada dos años.
- La cantidad de información que crean los particulares es mucho menor que la cantidad de información creada sobre ellos mismos en el universo digital.

• El 75% de la información en el universo digital la generan particulares.

• Las empresas son responsables del 80% de esta información en algún punto de su vida digital.

A Decade of Digital Universe Growth: Storage in Exabytes



# Definición

  Data warehouse es:

- «Una colección de datos no volátil, integrada y variante en el tiempo, orientada al usuario que tiene como objetivo ser una herramienta para la toma de decisiones» Bill Inmon

- «Una colección de datos derivada de las transacciones diseñada para la realización de consultas y análisis» Ralph Kimball

# Definición

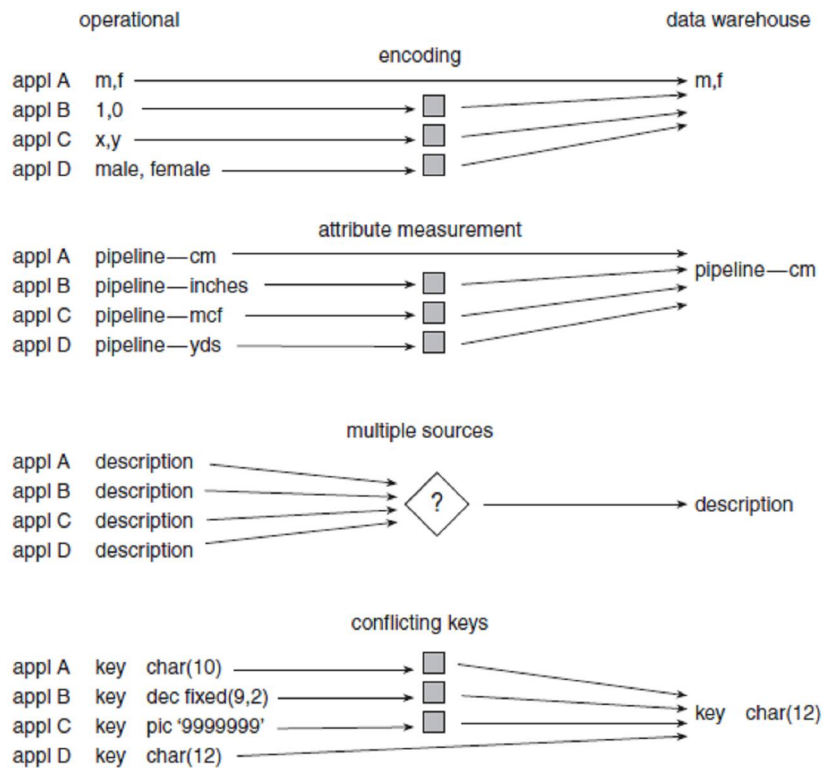
Orientada al usuario

- OLTP Software que utiliza la organización esta orientado a una tarea operacional.
- Ej.: POS, RTTx, CRM, Logística, RRHH
- El DW tiene como objetivo tener una visión global que permita ver a los objetos o entidades que son interesantes para el usuario.  
Ej. Clientes, Ventas, Stock, Costos.



# Definición

## Integration



# Definición

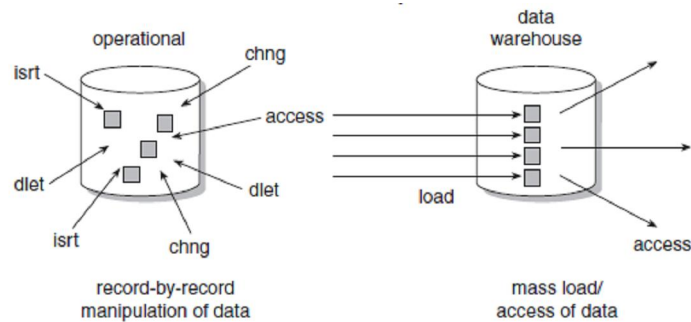
- No volátil y variable en el tiempo

- Es un repositorio de datos históricos.

Los datos perduran mas tiempo que en las bases operacionales.

- El concepto tiempo es fundamental para la utilidad de los análisis y como clave lógica física.

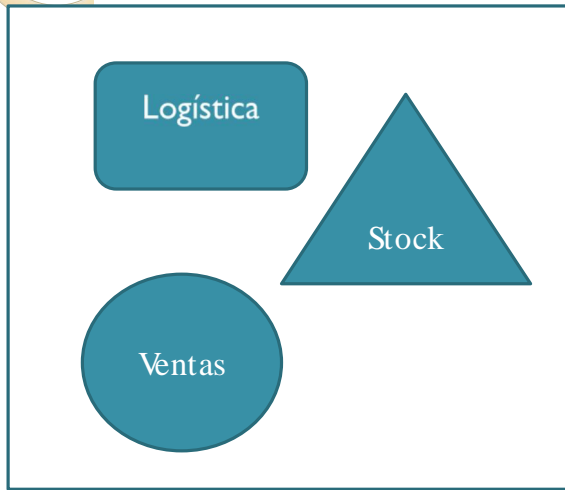
- Se guarda una nueva versión para cada cambio del sujeto de análisis en contraste con la actualización de los operacionales.



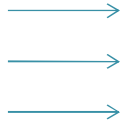


# Definición

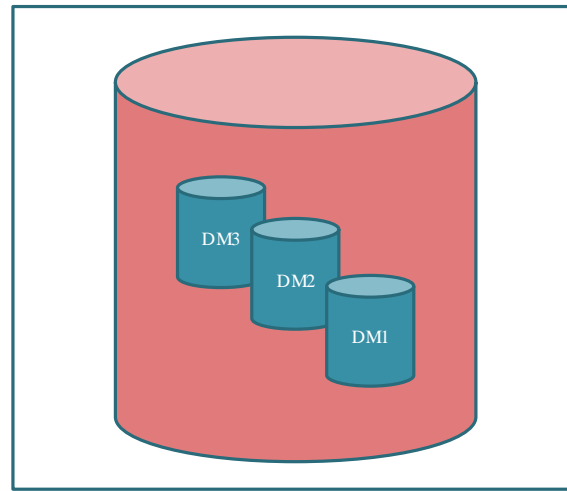
Operacionales



Cargas  
Masivas



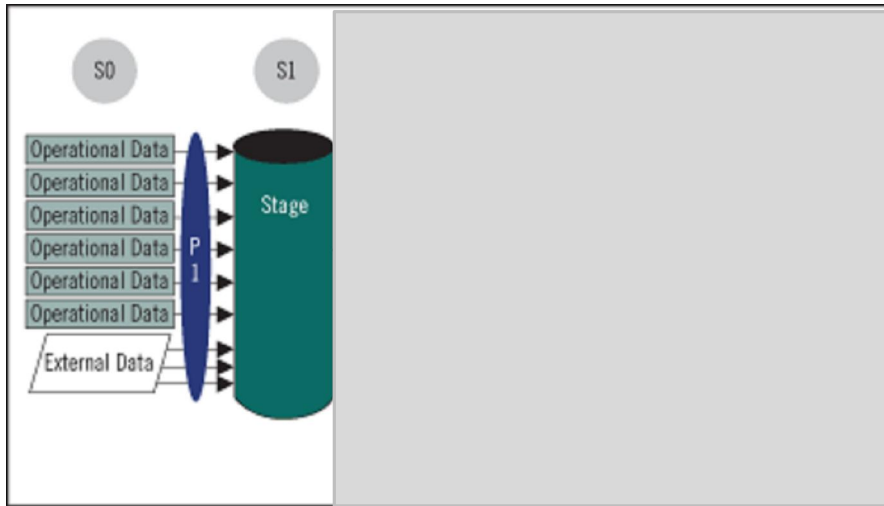
Data werehouse



SELECT

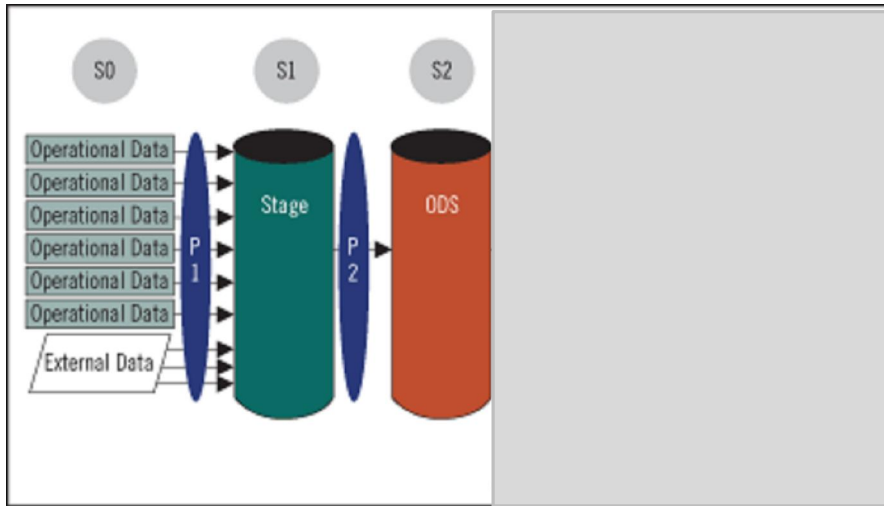
# Arquitectura

- S0/SI: Denominado Staging área es casi una copia de las fuentes de datos transaccionales.
- No se realizan transformaciones.
- No se aplica modelado lógico.
- En este paso es donde se ajusta la latencia de la información.



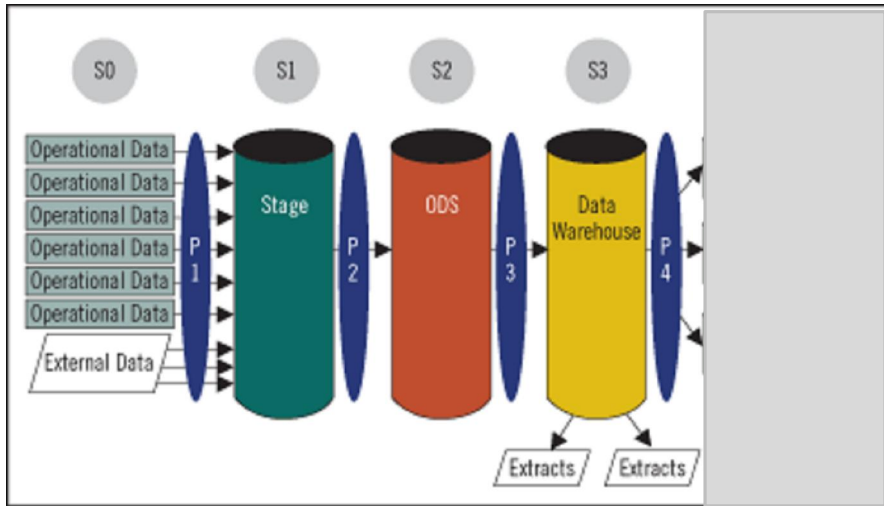
# Arquitectura

- S2 ODS: Operational Data Store
- Se utiliza para tareas de reporting de los sistemas operacionales.
- Si las bases operacionales pueden soportar reporting sin perjudicar performance este paso puede ser omitido.



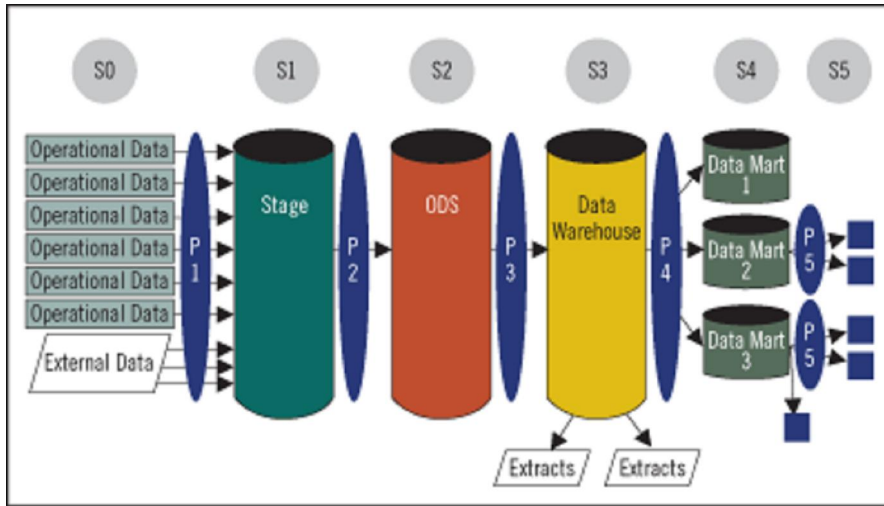
# Arquitectura

- S3 DW: Se aplica la normalización elegida en el modelado lógico y físico.
- La integración de los datos de distintas fuentes se realiza en este paso.
- Funciona como repositorio central de la empresa.



# Arquitectura


- S4/S5 Reporting y Explotación: Se crean vistas personalizadas por departamentos según requerimientos.
- Se realizan sumalizaciones y agregaciones dependiendo de la granularidad deseada.
- Se generan datasets, reportes, visualizaciones, análisis OLAP y técnicas de Data Mining.





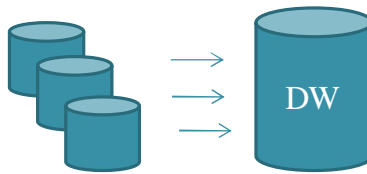
# Data Marts

- Es una vista integrada diseñada para un sector o departamento particular.
- El grado de granularidad, latencia e información disponible depende de los requerimientos del departamento.

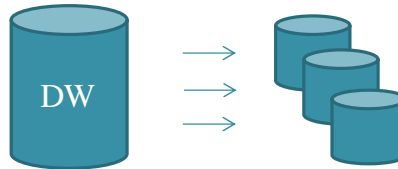
 Es un subconjunto de Data Warehouse empresarial.

# Arquitectura – DM y DW

Enfoque Ralph Kimball



Enfoque Bill Inmon



Enfoque Mixto





# Diseño de Data Warehouse

## Seleccionar el proceso de negocio

Se puede seleccionar o por la oferta basada en datos disponibles e identificando cuales de ellos son clave para la toma de decisiones.

- Se puede seleccionar según los requerimientos del usuario.

## Delinear la granularidad

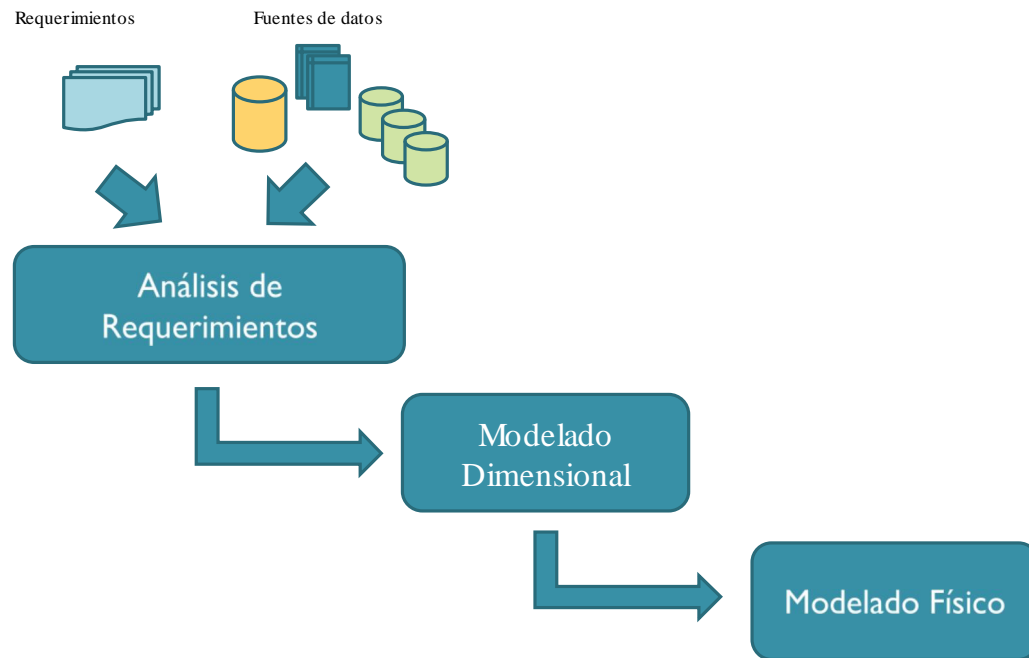
Que nivel de detalle necesitan los usuarios finales?

## Detallar las dimensiones

## Detallar los hechos



# Fases



# Modelado dimensional

- Es una técnica que presenta las dimensiones, hechos y sus relaciones de forma estándar e intuitiva.

- **Dimensión / Atributo**

  - En general de tipo texto

  - Son agrupables

  - Valores cualitativos de una transacción

  - Ejemplos: Día, Local, Municipio, Vendedor, Producto.

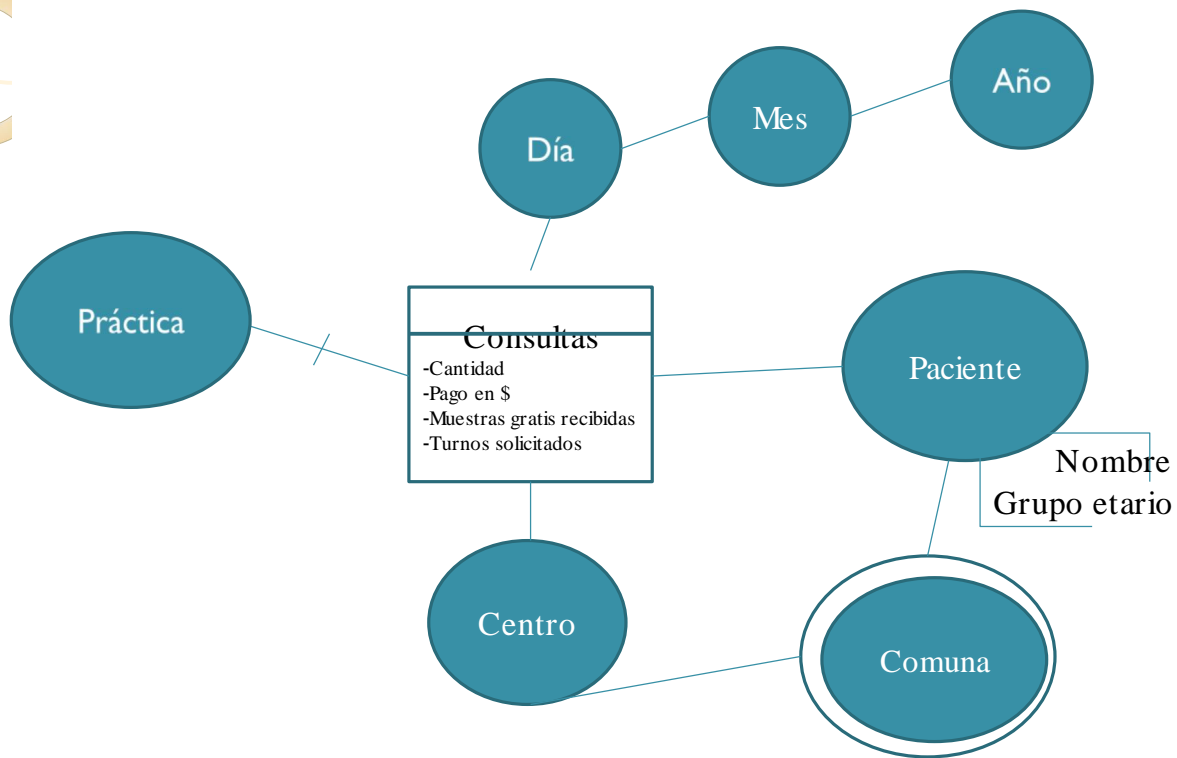
- **Hechos / métricas**

  - En general de tipo numérico

  - Son agregables (Sum, Avg, Count etc.)

  - Valores cuantitativos de una transacción


# Modelo dimensional






# Diseño Físico

- Debemos pasar del modelo conceptual al modelo físico.
- Armar un modelo físico es dar un soporte en tablas a todos los objetos identificados en el modelo multidimensional.

 Lo que se define en esta etapa son las tablas del data warehouse.

- El modelo físico se compone exclusivamente de tablas y columnas.

 Las tablas look up o de dimensión son las que almacenan los elementos de un atributo.

# Tipos de tablas - nomenclatura



# Diseño Físico - Normalización

- **Completamente normalizado:** id propio, descripción, id del padre.

- ▣ **Moderadamente normalizado:** idem anterior pero cada tabla tiene todas las referencias a los ancestros.

- **Completamente desnormalizado:** idem anterior más todas las descripciones.

# Diseño Físico

## Claves Subrogadas

Clave propia del DW

Para independizarse de cambios en sistemas fuentes.

- ¿Qué pasa si se utilizan códigos depurados?
- ¿Qué pasa si cambian las claves de los sistemas fuentes?

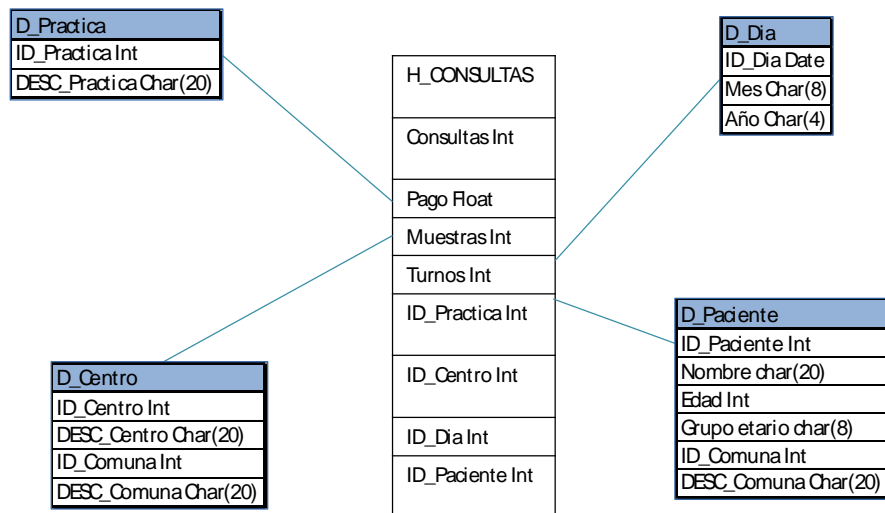
# Diseño Físico

## Tipos de hechos

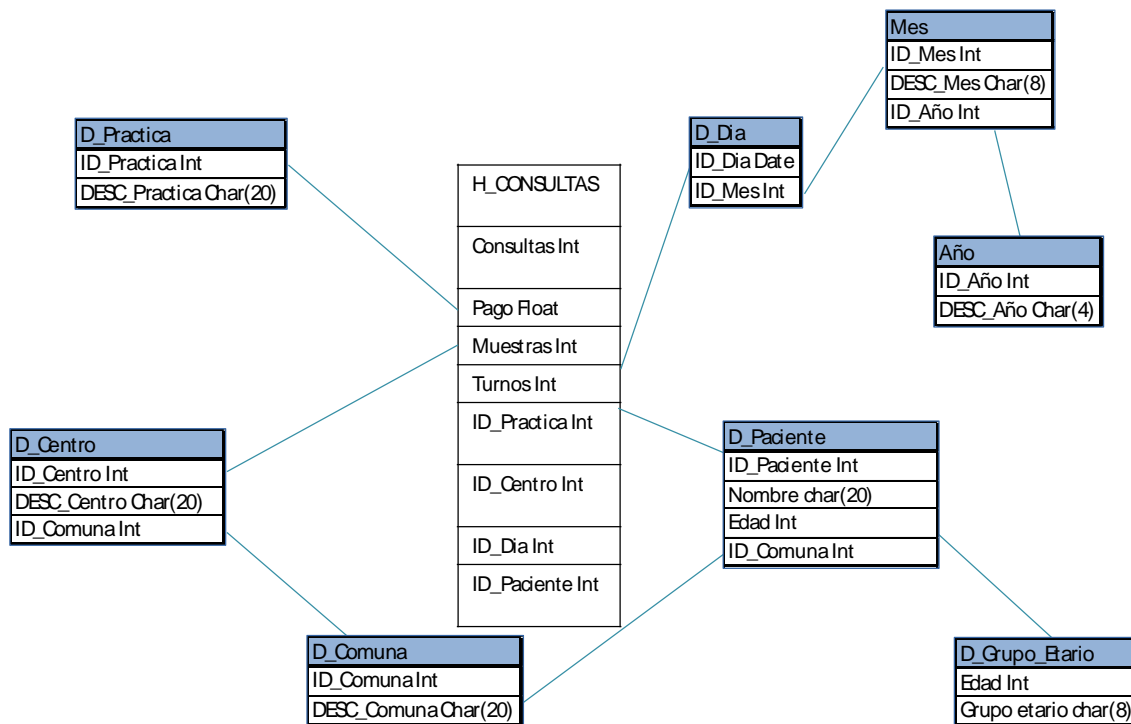
- Aditivos: Se aplican operadores distribuidos y se puede agregar los datos sin perder información (sum, max, min etc.)
- Semi Aditivos: Se pueden agregar en algunas dimensiones y en otras no. (Ej. ¿Cuántos productos teníamos en stock el 1/6?)
- No Aditivos: No es posible obtener la misma información. (Ej. Moda, Mediana)



# Diseño Físico – Esquema Estrella



# Diseño Físico – Esquema Snowflake





# Slowly Changing Dimensions

Las descripciones cambian cada tanto.

- Nuevo Registro

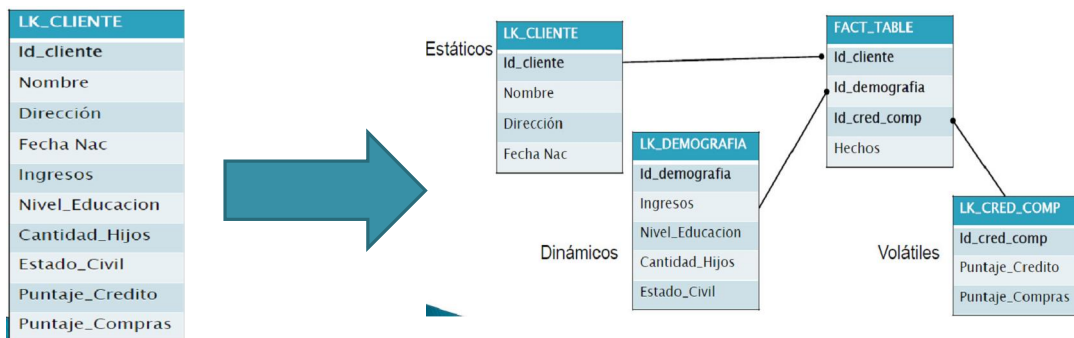
- Marca para auditar los cambios

- Versión

- En algunos casos solo se guarda el estado anterior

# Rapidly Changing Dimensions

- Dimensión muy grande que cambia constantemente.
  - Dejarla como originalmente se diseñó
  - Separar en datos estáticos y dinámicos
  - Separar también los volátiles
  - Llevar los datos volátiles a las fact tables.





# Procesos ETL

❏ Extraer Transformar y Cargar (o ELT si se utiliza Staging)

❏ Se definen Entradas

- Conexión a fuentes de datos

- Se definen Filtros

- Cleaning y Data Quality

- Se generan variables calculadas

❏ Se definen Salidas

❏ Durante todo el proceso es fundamental la MetaData



# Carga y actualización

Se caracterizan 2 tipos de ETL

## Carga Inicial


- Consiste en la generación inicial del DW
  - Se define que cantidad de datos históricos serán disponibles
  - Se debe informar a los usuarios que información no estará disponible
- Probablemente tenga un gran impacto en los sistemas fuentes por el alto volumen de datos.

# Carga y actualización

## Refresh

◦ Se define para cada requerimiento que entidades serán actualizadas por hora, día, semana, mes etc.

Como tratar los cambios en las dimensiones

-  Hay dimensiones que cambian lentamente: Ej. Estado Civil
  - Estrategias: Actualiza la forma, nueva versión, versión anterior y actual, fechas auditoria.
  - Hay dimensiones que cambian rápidamente: Ej. Puntaje de compras
  - Estrategias: Separar parte dinámica y estática, separo los volátiles, llevar la información a fact table.
  - Dimensión Junk: Atributos que no pertenecen a una dimensión.
  - Estrategias: Fact table, cada atributo como una dimensión aparte, sacarlos del diseño, Junk.



# Carga y actualización

## Problemas a tratar en proceso ETL

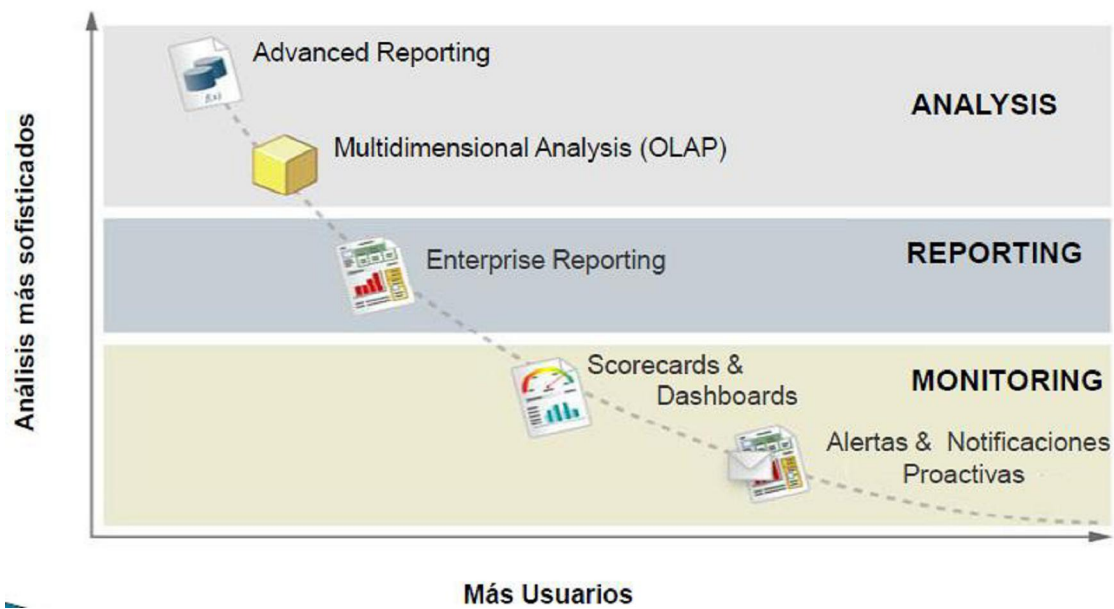
Data Quality

- Detección de los cambios en sistemas fuentes.
- Normalización

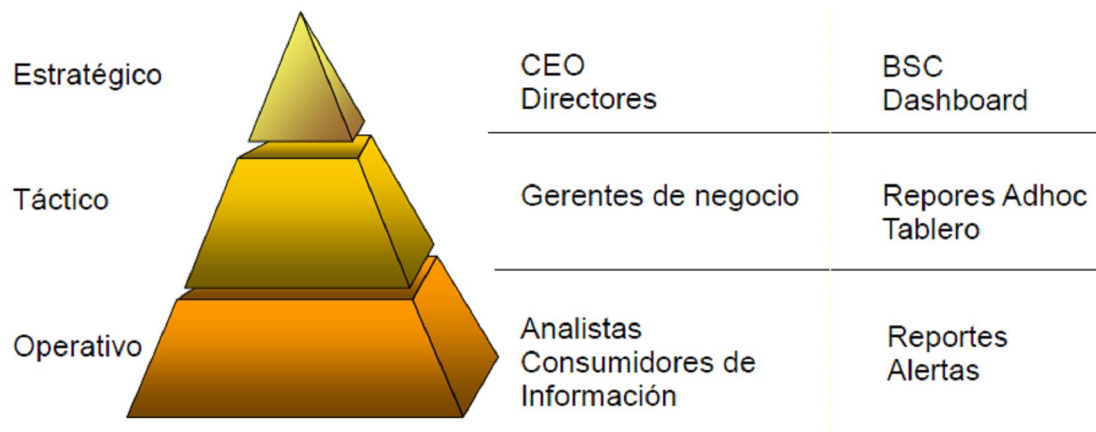
Procedimientos por contingencia



# Visualización de datos

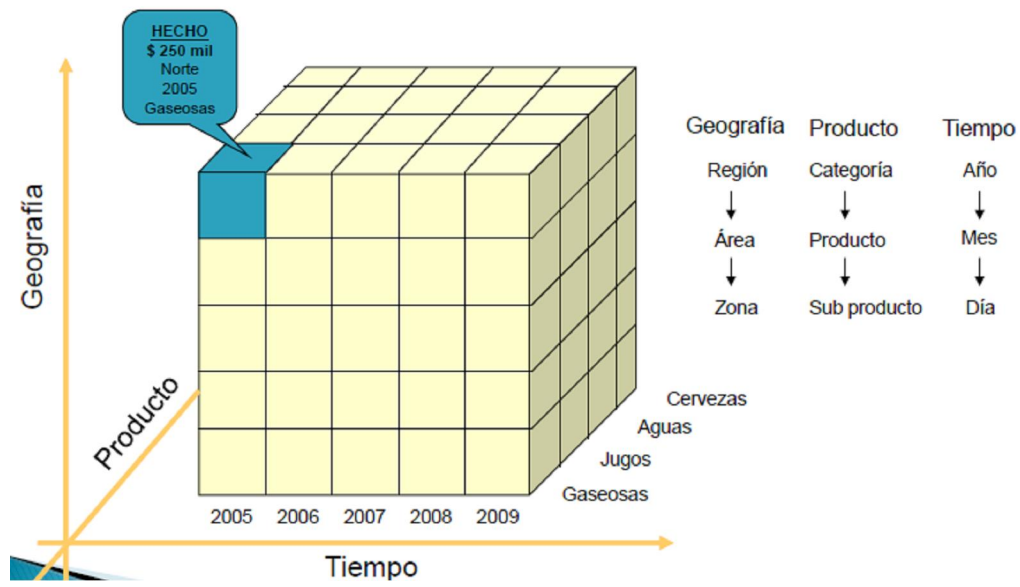


# Visualización de datos



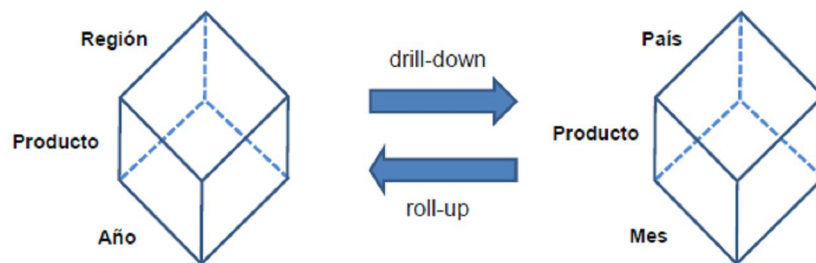
Ejemplos de dashboards on-line: <http://www2.microstrategy.com/software/business-intelligence/dashboards-and-scorecards/gallery/>

# Cubos



# Operaciones OLAP

- roll-up: Agregar datos a un nivel mayor en una jerarquía
- drill-down: Agregar datos a un nivel menor en una jerarquía



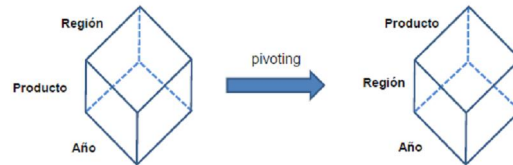
# Operaciones OLAP

- Pivoting: cuando se mueve una o mas

Sum of ventas		región	país		LATAM	
año	mes	Europa	Italia	Argentina	Brasil	
2008	Enero 08	\$ 4,800	\$ 9,100	\$ 1,800	\$ 7,500	
	Febrero 08	\$ 5,100	\$ 9,600	\$ 2,400	\$ 7,900	
	Marzo 08	\$ 5,250	\$ 9,700	\$ 3,000	\$ 9,200	



Sum of ventas		año	mes		
			2008		
región	país	Enero 08	Febrero 08	Marzo 08	
Europa	España	\$ 4,800	\$ 5,100	\$ 5,250	
	Italia	\$ 9,100	\$ 9,600	\$ 9,700	
LATAM	Argentina	\$ 1,800	\$ 2,400	\$ 3,000	
	Brasil	\$ 7,500	\$ 7,900	\$ 9,200	



 Slice & Dice: Reducir la dimensionalidad.



# Molap, Rolap y Holap

## MOLAP Multidimensional On-Line Analytical Processing

- Se almacenan los datos de forma multidimensional.

- El lenguaje de consulta es MDX

- Las agregaciones ocupan mucho espacio y pueden no ser utilizadas

- Pocos flexibles para refresh ya que generalmente se debe actualizar gran parte del cubo

- Se utiliza para datos que se consultan muy frecuentemente.



# Molap, Rolap y Holap

## ROLAP Relational On-Line Analytical Processing

- Los datos se guardan en una base de datos relacional.

- El lenguaje de consulta es SQL

- Provee gran escalabilidad

- Se requieren agregaciones pre calculadas para suplir inconvenientes de performance



# Molap, Rolap y Holap

## HOLAP Hybrid On-Line Analytical Processing

- Es una combinación de las anteriores
- Se guarda una parte de la información MOLAP y otra en ROLAP
- Se guarda la información en MOLAP mas critica en cuanto al tiempo de respuesta de las consultas
- Se guarda la información en ROLAP para optimizar los tiempos de carga y procesamiento/refresh del Cubo.





# Aplicaciones

Mostrar ejemplos de aplicaciones utilizadas para BI disponibles en el mercado.

ETL: Informatica, ODI Oracle, MSIS Microsoft, etc.

BI OLAP y Visualizacion: MicroStrategy, Cognos, Business Objects etc.