Sistemas de Información para el Análisis de GVDatos

Instituto de Computación - Facultad de Ingeniería

Marzo 2024

Introducción

Temas

- Motivación
- Arquitecturas
- Modelos Multidimensionales
- Sistemas de Data Warehouse

Motivación

FUENTES DE DATOS

Agregar una línea de producto →

Cambiar el precio de un producto →

BD Productos

Cambiar cronograma de avisos →

Aumentar gasto en radio →

BD Publicidad

Aumentar límite de crédito →

Cambiar nivel de salario de cliente →

BD Clientes

ANÁLISIS PARA TOMA DE DECISIONES

Motivación

FUENTES DE DATOS

Agregar una línea de producto →

Cambiar el precio de un producto →

BD Productos

Cambiar cronograma de avisos →

Aumentar gasto en radio →

BD Publicidad

Aumentar límite de crédito →

Cambiar nivel de salario de cliente →

BD Clientes

ANÁLISIS PARA TOMA DE DECISIONES

Cuántos productos se vendieron el último mes de

Cuánto hemos gastado en nos

Cómo han evolucionado los precios de nuestros pro

Qué tipos de clientes han aumentado las compras en qué tipos de productos?

Motivación

FUENTES DE DATOS

BD Productos

BD Publicidad

BD Clientes

ANÁLISIS PARA TOMA DE DECISIONES

Cuántos productos se vendieron el último mes de

Cuánto hemos gastado en nos

Cómo han evolucionado los precios de nuestros pro

Qué tipos de clientes han aumentado las compras en

Qué nivel de aceptación tie de

Qué reputación tienen nuestros clientes?

Cómo nos posicionamos con respecto a los precios de la canasta familiar?

Factores Críticos

- **Tradicionalmente:**
  - **Accesibilidad**
  - **Tiempo de acceso y transformaciones**
  - **Integración de datos**
  - **Calidad de datos**

- **Al incluir Big Data al análisis:**

Volumen **Variedad**

Velocidad **Veracidad**

La información y las organizaciones

- Las organizaciones tienen necesidad de:
  - Conocimiento:
    - Materia prima para toma de decisiones.
    - Es lo que se desea construir.
  - Información:
    - Materia prima para conocer los fenómenos reales.
    - Un ítem de datos es información según el contexto.
  - Datos:
    - Materia prima de la información.
    - Generados por procesos que no necesariamente los explotan.

La información y las organizaciones

- Los datos existen, pero ...
  - No siempre se acceden fácilmente
  - No siempre se explotan

[1]

- ❖ Datos organizacionales
- ❖ Internet
- ❖ Social Media
- ❖ Comunicación
- ❖ Fotos
- ❖ Servicios
- ❖ Internet of things

  - Qué porcentaje de la información se usa? Muy pequeño…

[1] www.forbes.com

La información y las organizaciones

- Los sistemas de información tradicionales
  - Orientados a sistemas operacionales
  - Asociados a procesos productivos
  - Procesan grandes cantidades de transacciones
- Pueden resolver estas necesidades ?

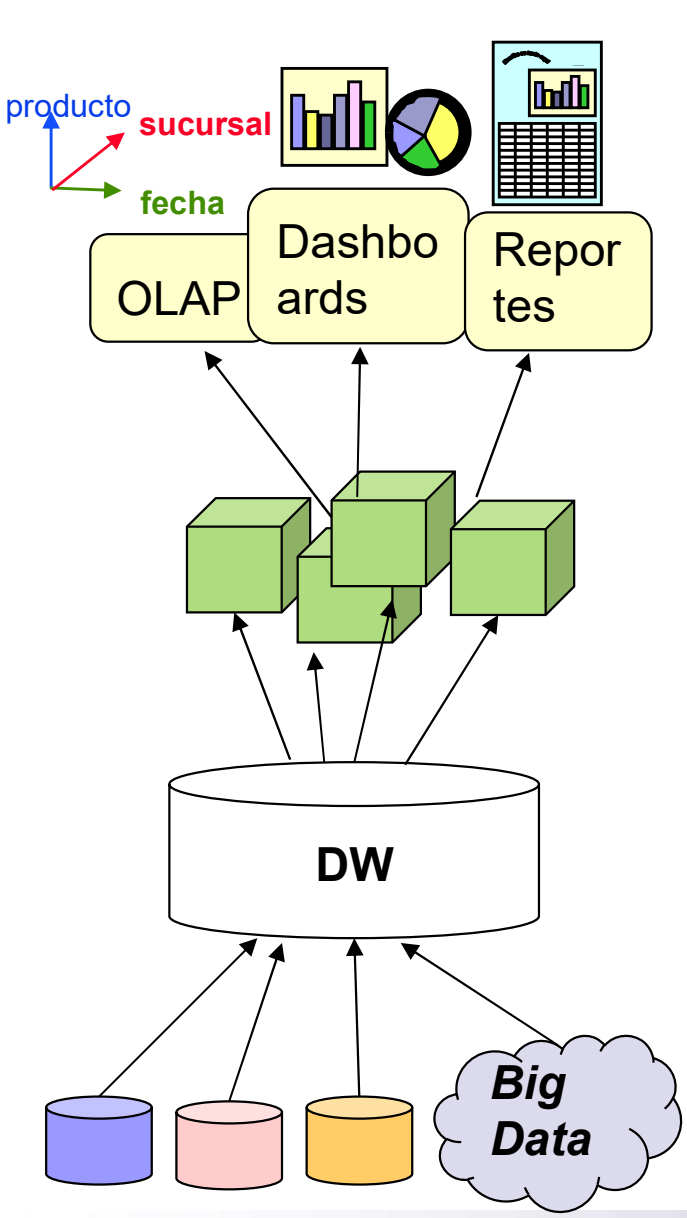
Sist. de Producción y de Decisión

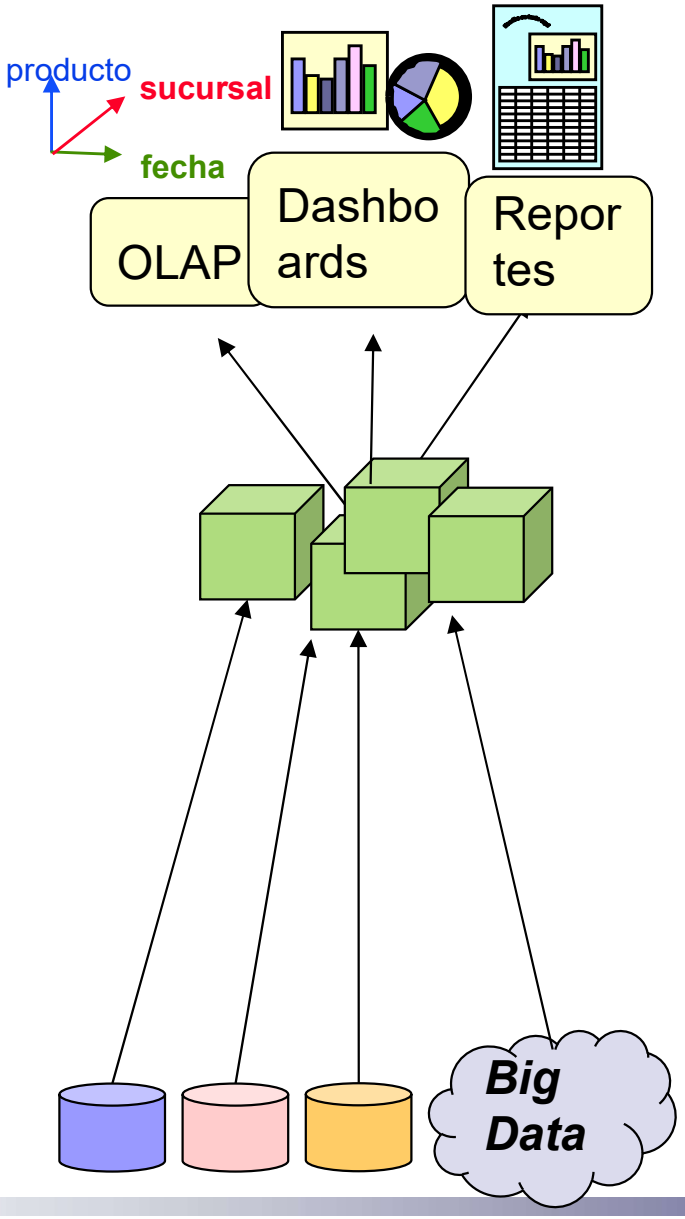
- Sistemas orientado a la Producción:
  - Prioridad:
    - tiempo de respuesta a transacciones read-write
  - Se manejan datos actuales muy detallados
  - Estables y de larga vida útil
- Sistema orientado a la Decisión:
  - Prioridad:
    - expresividad y eficiencia en consultas complejas
  - Datos actuales+históricos, resumidos
  - En constante evolución

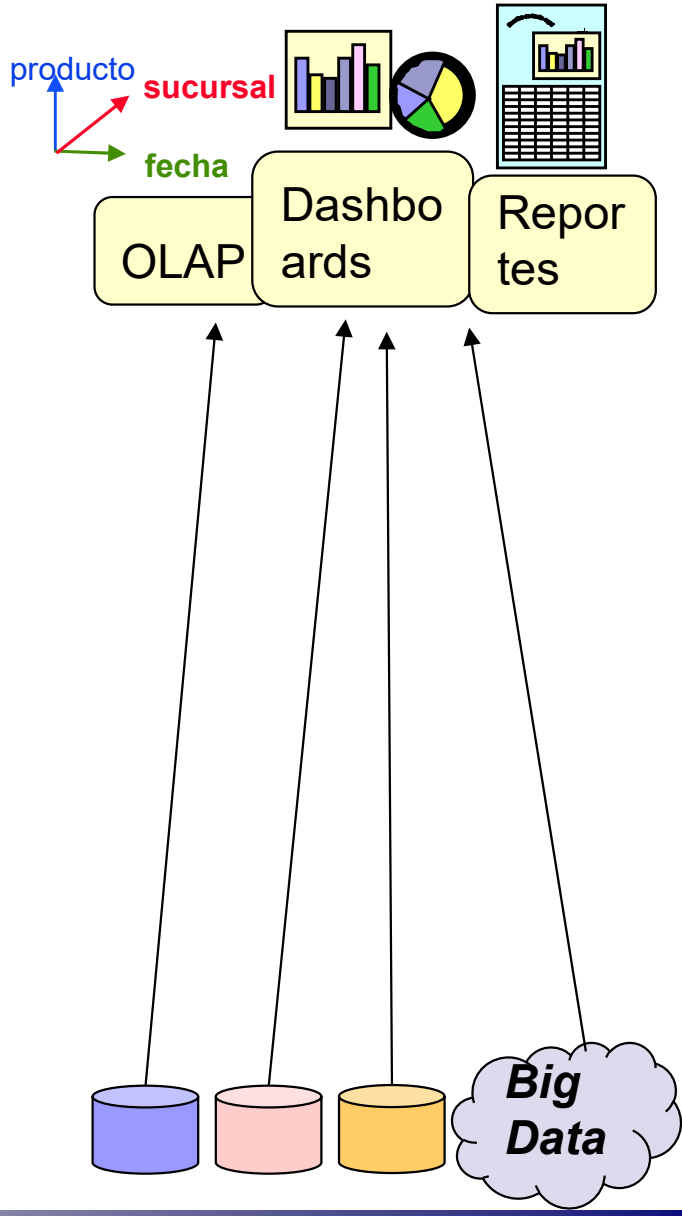
Sist. de Producción y de Decisión

- Conclusión
  - Se trata de sistemas con objetivos diferentes
  - Se construyen para ser eficientes en sus objetivos
- No es posible usar uno para las tareas del otro
- Sistemas de Decisión: de ahora en más le llamaremos **Sistemas de Información para el Análisis de GVD**

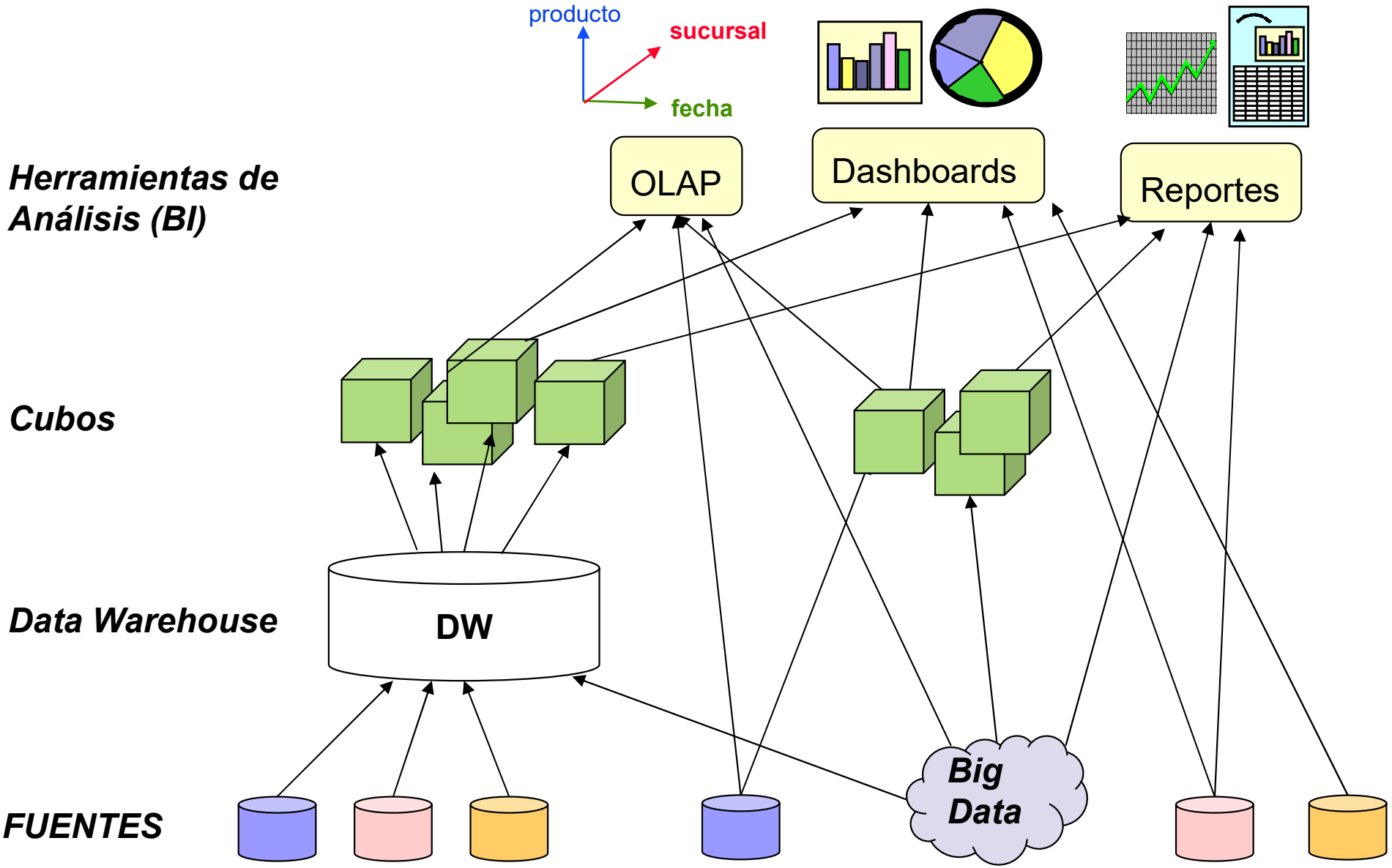
Análisis GVD - Arquitecturas







Análisis GVD - Arquitectura

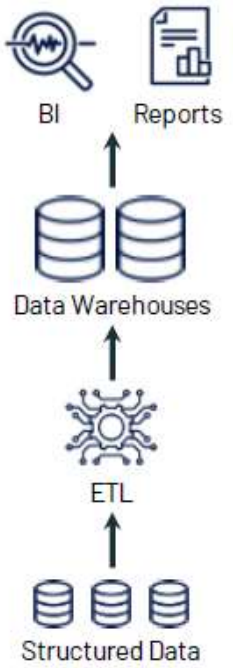


Arquitecturas para análisis GVD

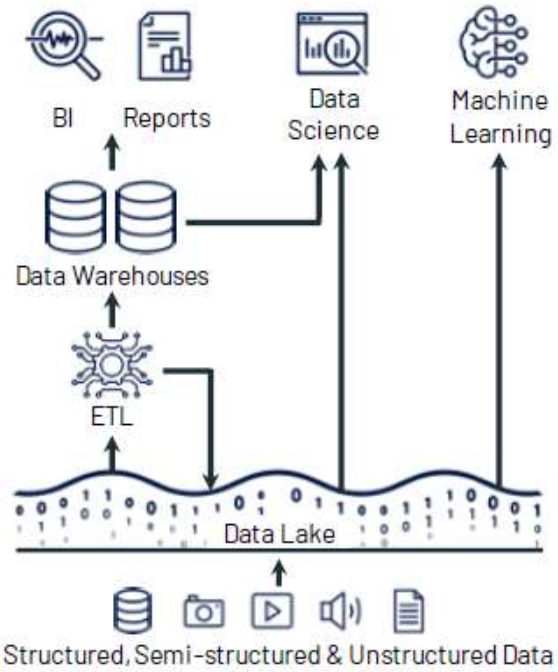
Lakehouse: A New Generation of Open Platforms that Unify DataWarehousing and Advanced Analytics

CIDR '21, Jan. 2021, Online

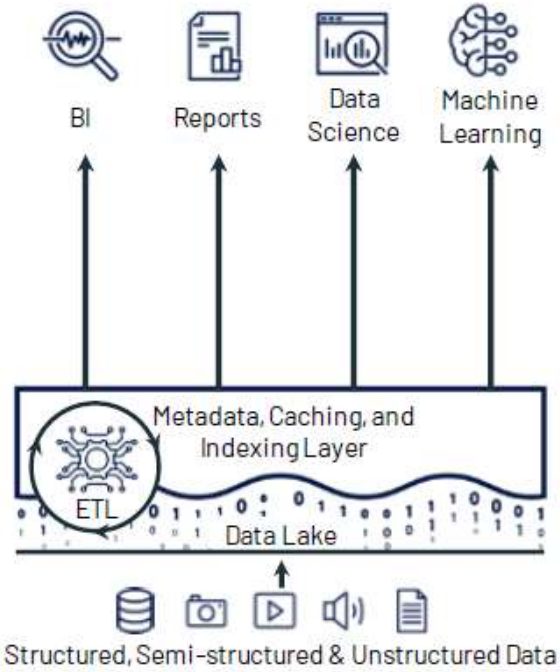
Michael Armbrust, Ali Ghodsi, Reynold Xin, and Matei Zaharia



(a) First-generation platforms.



(b) Current two-tier architectures.



(c) Lakehouse platforms.

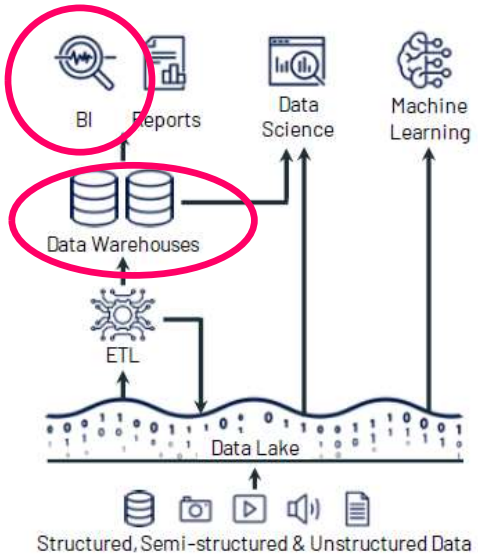
Definiciones

- Data Lake
  - Colección masiva de *datasets* que pueden
    - estar almacenados en distintos sistemas
    - tener **formatos variados**
    - no estar acompañados de metadatos
    - cambiar en forma autónoma
  - Objetivo
    - Obtener los datos rápidamente de las fuentes y darle a los usuarios la posibilidad de manejar la heterogeneidad y la escalabilidad

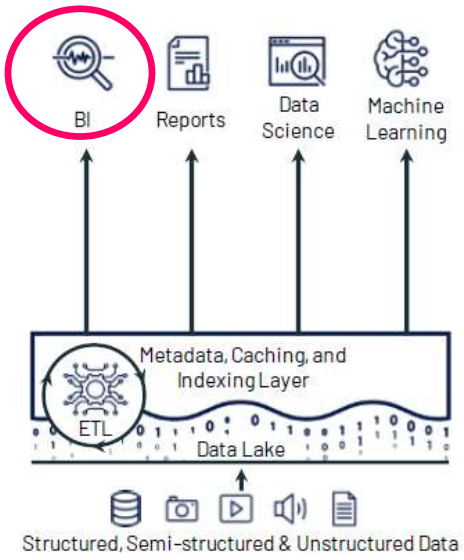
Definiciones

- Data Warehouse
  - Base de datos donde los datos, que provienen de diversas fuentes, fueron cuidadosamente transformados y corregidos, para conformar formatos, esquemas y semántica estándares de la organización.
- ETL – Extraction, transformation and loading
  - Proceso de transformación y carga de datos
- BI (Business Intelligence)
  - Análisis de datos orientada al negocio. En general, utilizando OLAP (On-line Analytical Processing), sistemas basados en el Modelo Multidimensional.

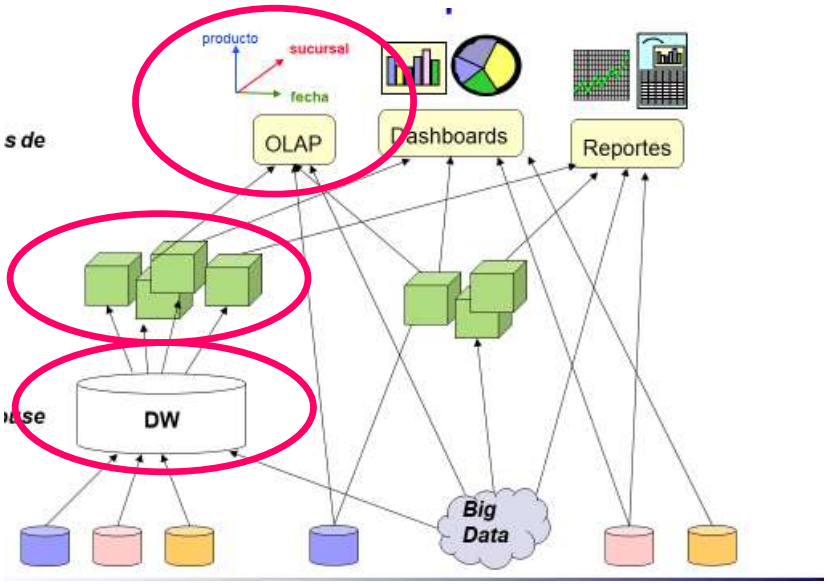
OLAP – Modelos multidimensionales



(b) Current two-tier architectures.



(c) Lakehouse platforms.

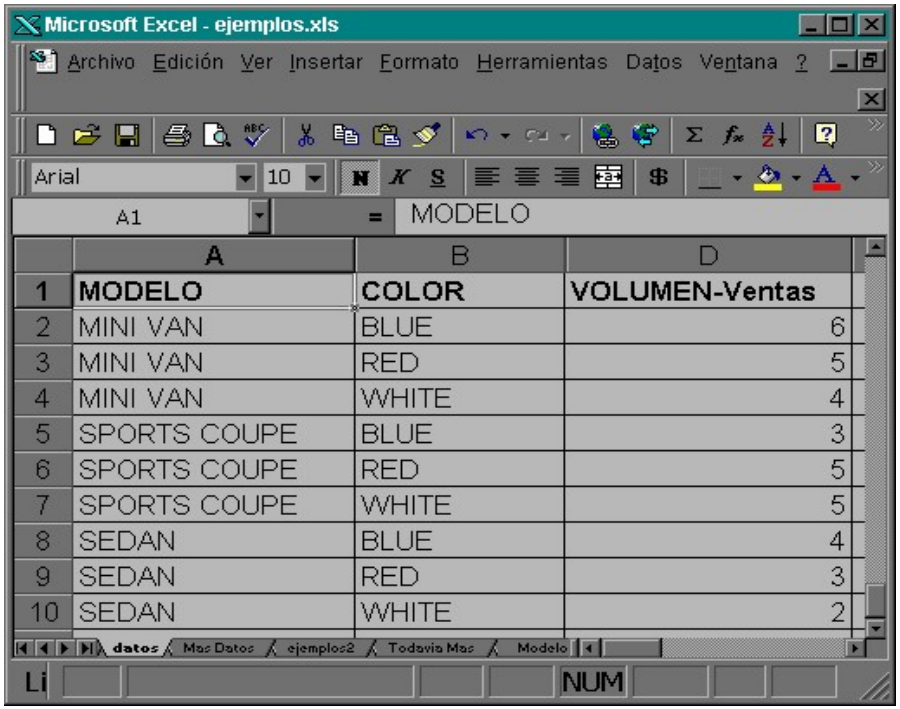


Modelos Multidimensionales

Motivación

Representación Tabular

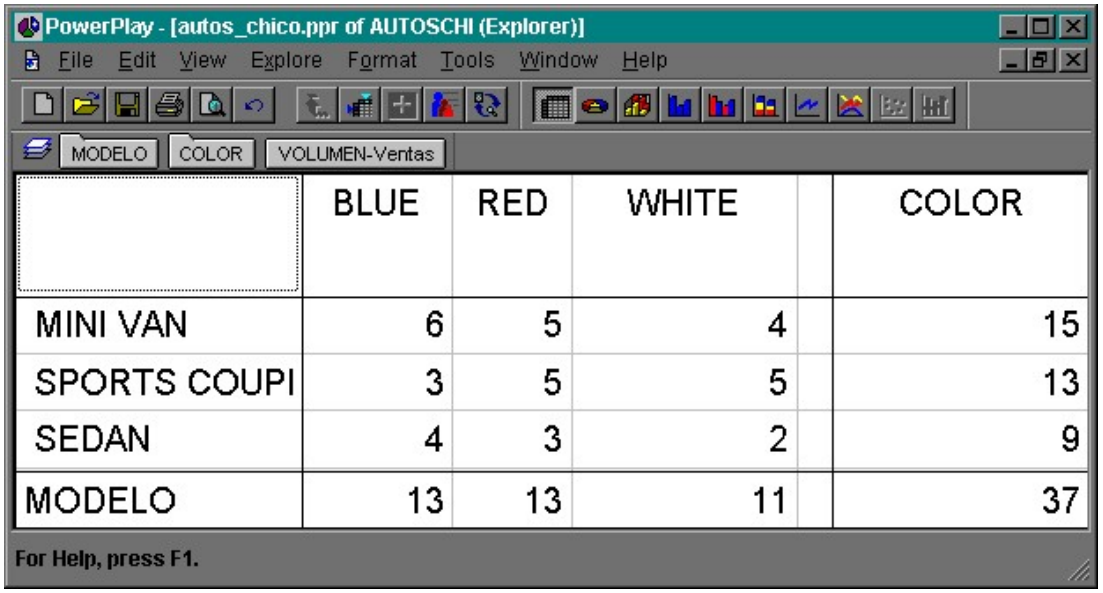
MODELO	COLOR	VOLUMEN-Ventas
MINI VAN	BLUE	6
MINI VAN	RED	5
MINI VAN	WHITE	4
SPORTS COUPE	BLUE	3
SPORTS COUPE	RED	5
SPORTS COUPE	WHITE	5
SEDAN	BLUE	4
SEDAN	RED	3
SEDAN	WHITE	2



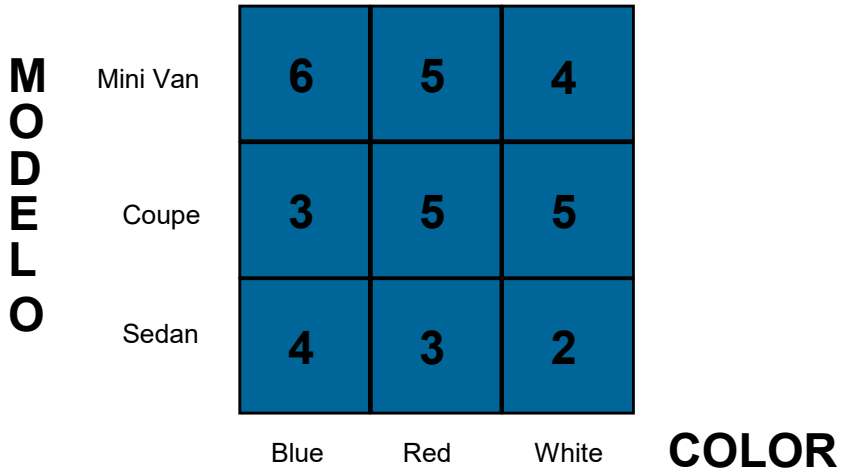
Ventas de autos en función de Modelo y Color.

Motivación

Representación Matricial

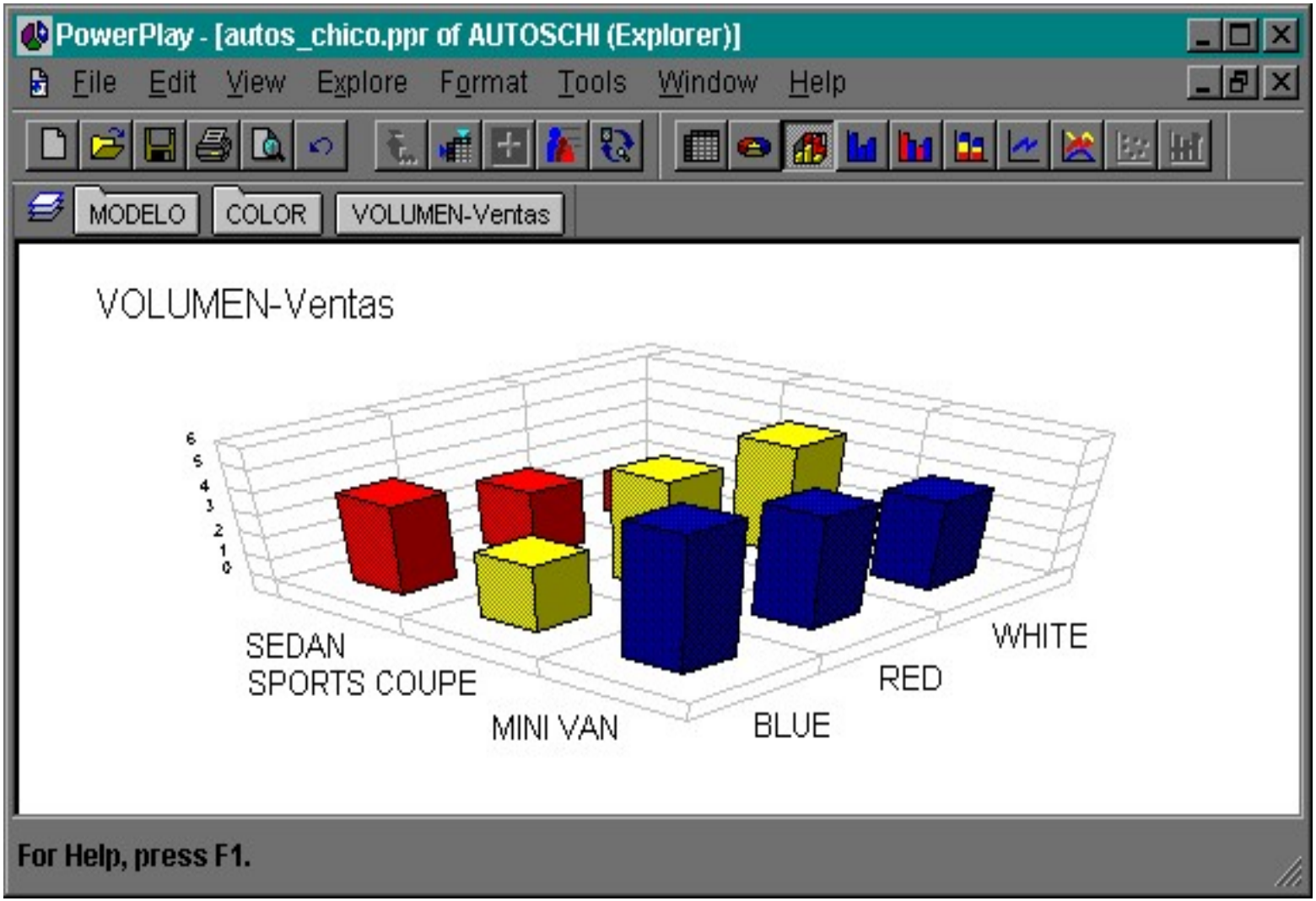


	BLUE	RED	WHITE	COLOR
MINI VAN	6	5	4	15
SPORTS COUPI	3	5	5	13
SEDAN	4	3	2	9
MODELO	13	13	11	37



Ventas de autos en función de Modelo y Color.

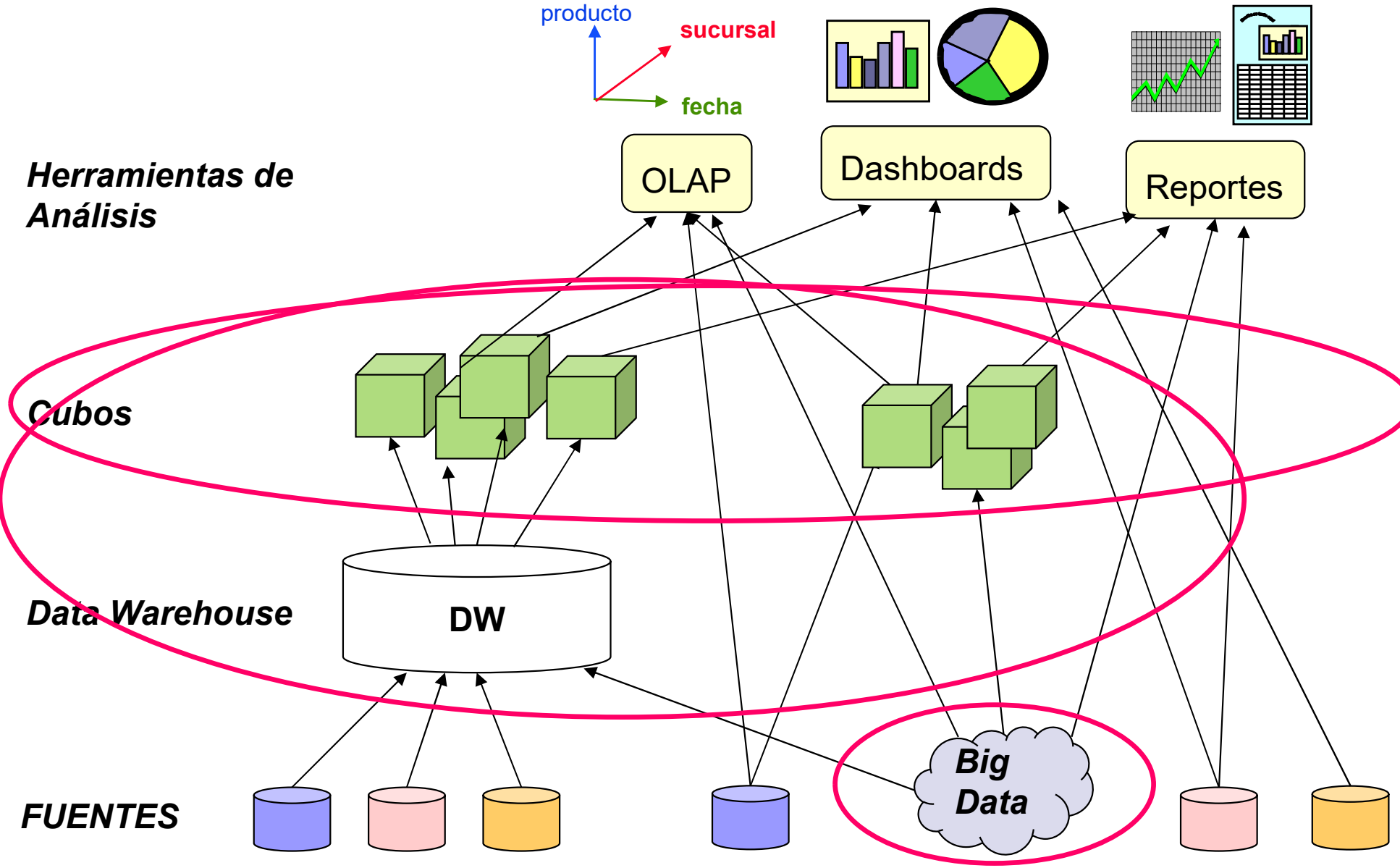
Motivación


PowerPlay - [autos_chico.ppr of AUTOSCHI (Explorer)]
File Edit View Explore Format Tools Window Help
MODELO COLOR VOLUMEN-Ventas
VOLUMEN-Ventas
6
5
4
3
2
1
0
SEDAN
SPORTS COUPE
MINI VAN
BLUE
RED
WHITE
For Help, press F1.

Modelos Multidimensionales

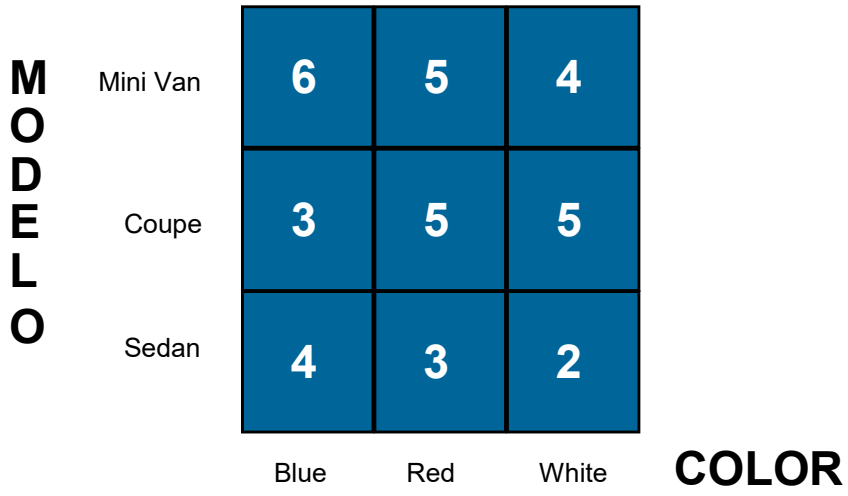
- Qué tienen en común estas dos últimas representaciones?
  - El usuario final (gerente) las entiende y maneja habitualmente.
- Objetivos de los MMD:
  - Representar los datos en forma cercana a la intuición del usuario.
  - Resolver problemas planteados en sistemas relacionales.

Sist. Inf. Análisis - Arquitectura



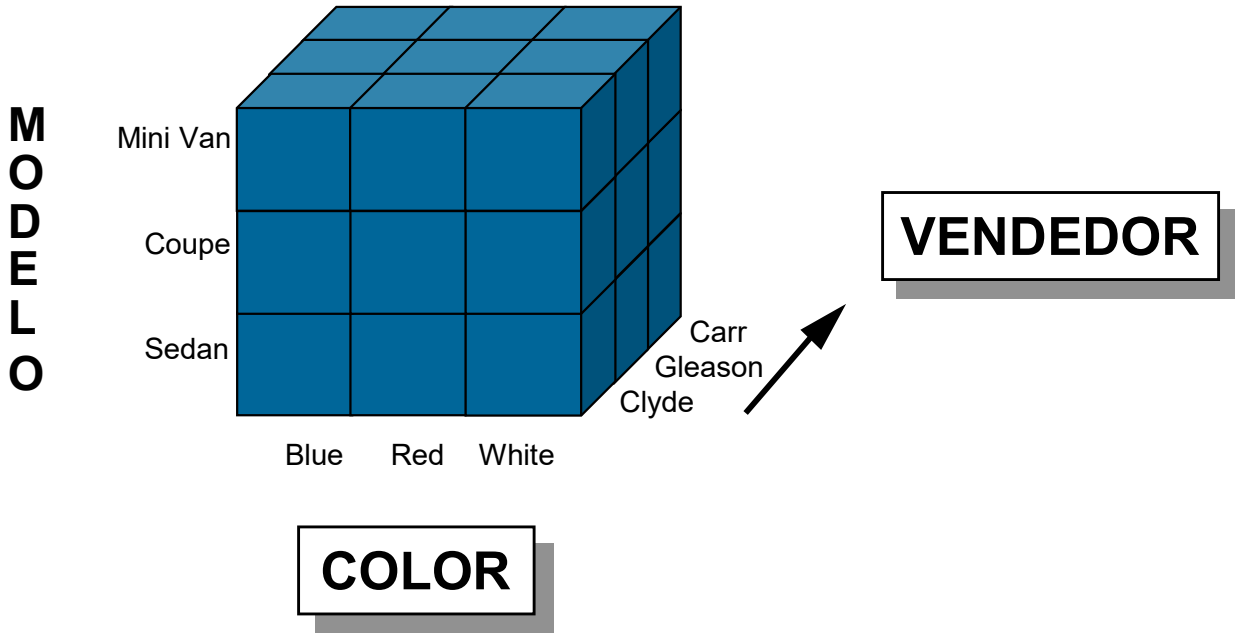
Características

- Se representan los datos como una matriz.
  - En los ejes están los criterios de análisis.
  - En los cruces están los valores a analizar.
  - A esta estructura se le llama ***Cubo*** o ***Hipercubo***.



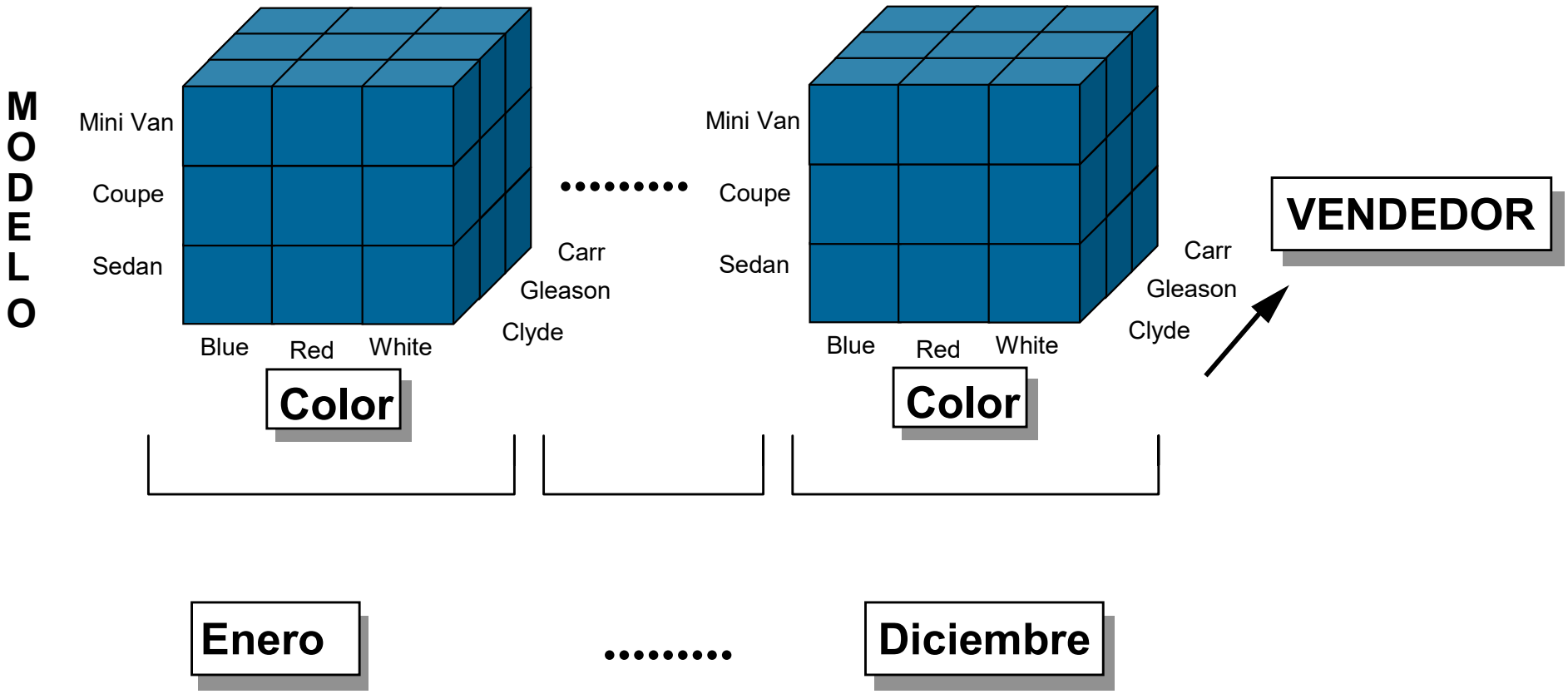
Características

- Agregando una 3a. dimensión:

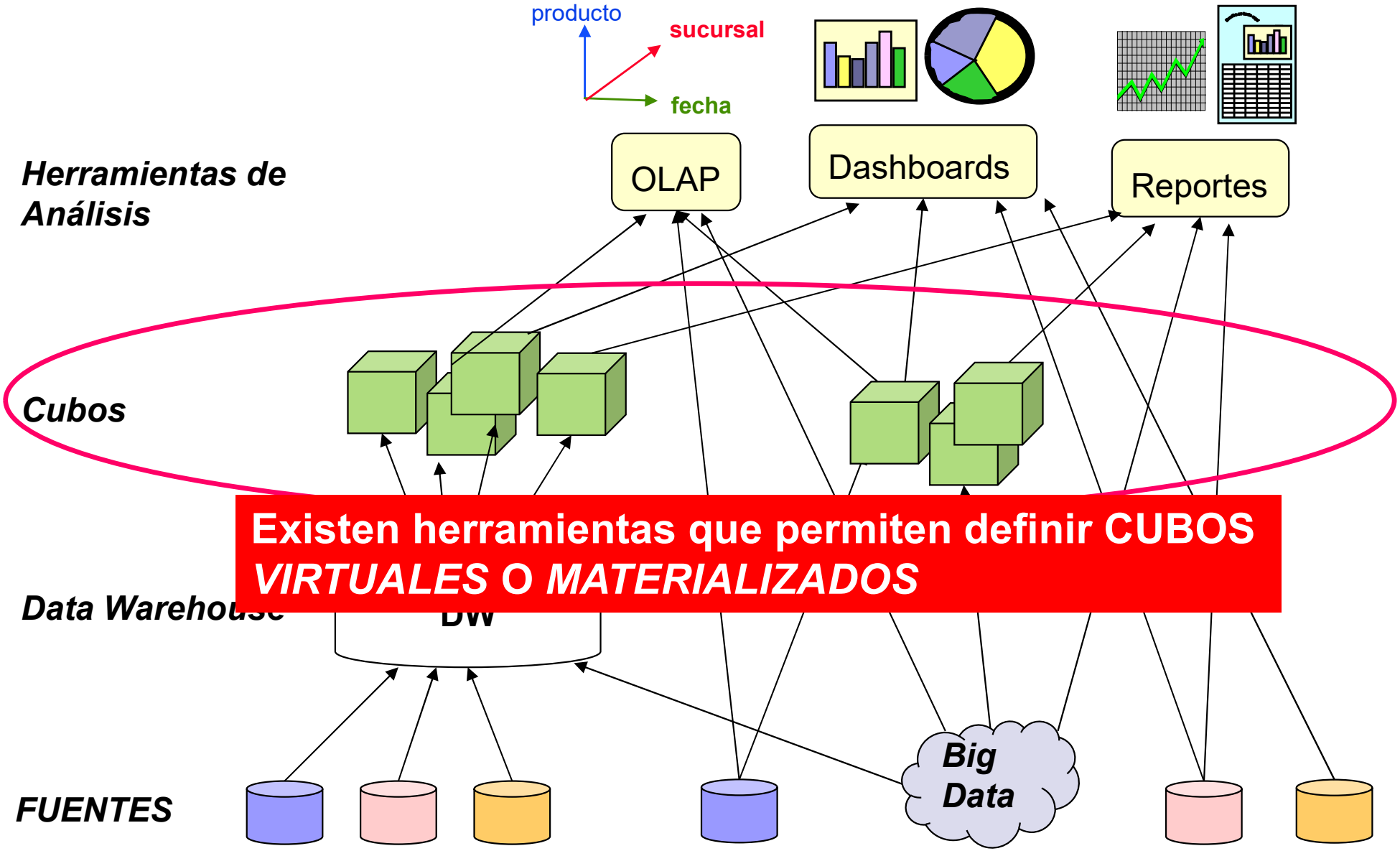


Características

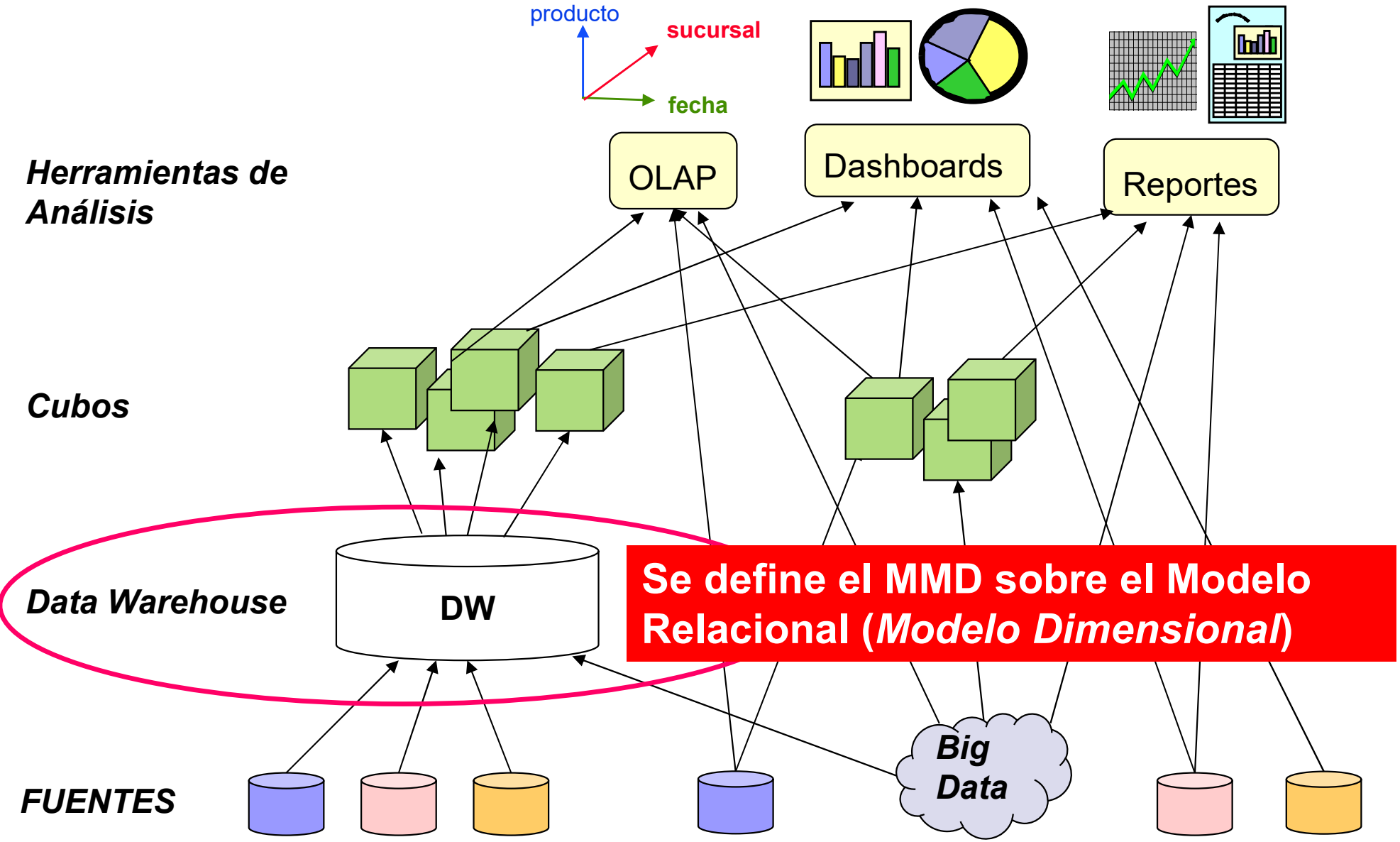
- Agregando una 4a. dimensión:



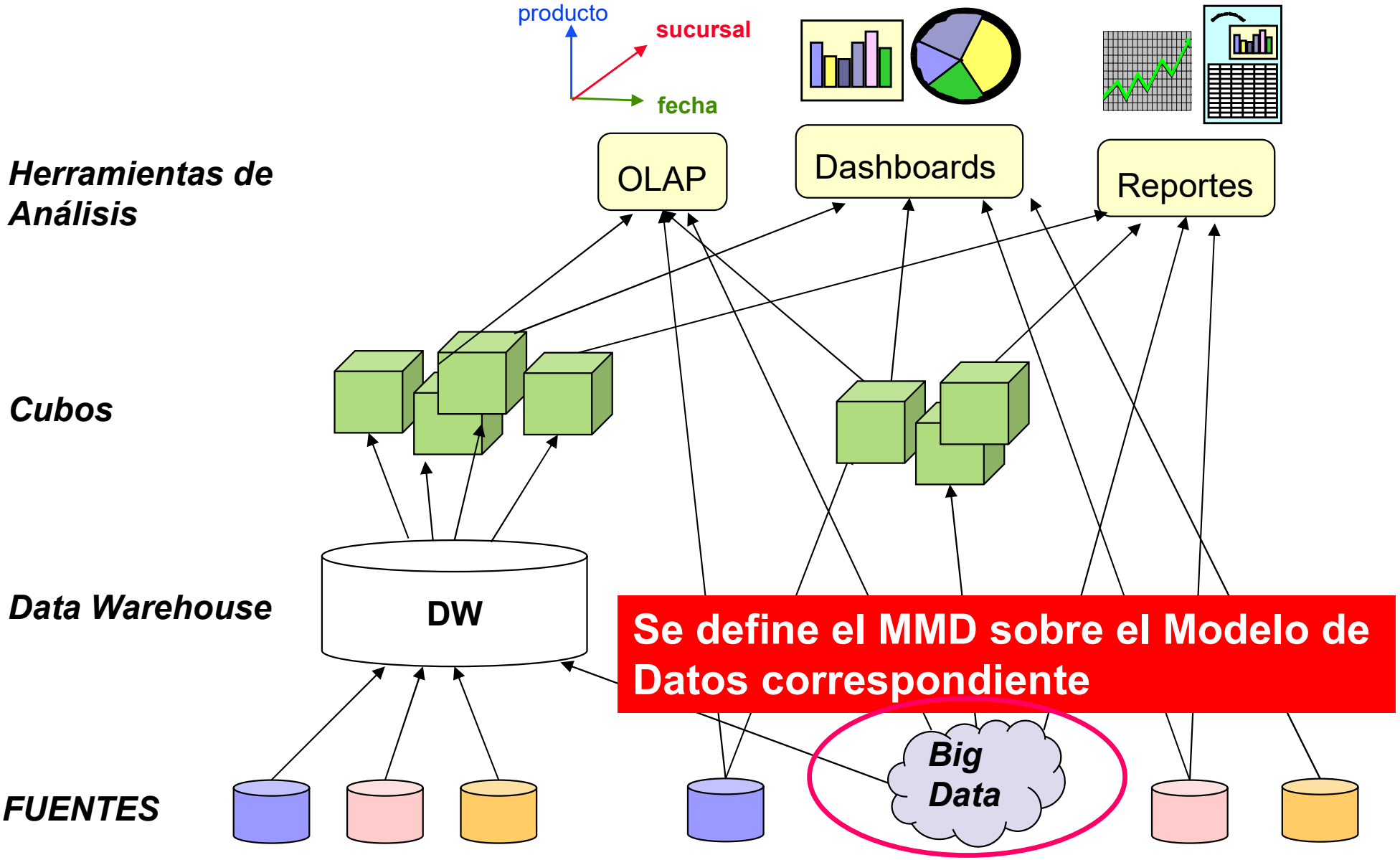
MMD en cada componente



MMD en cada componente

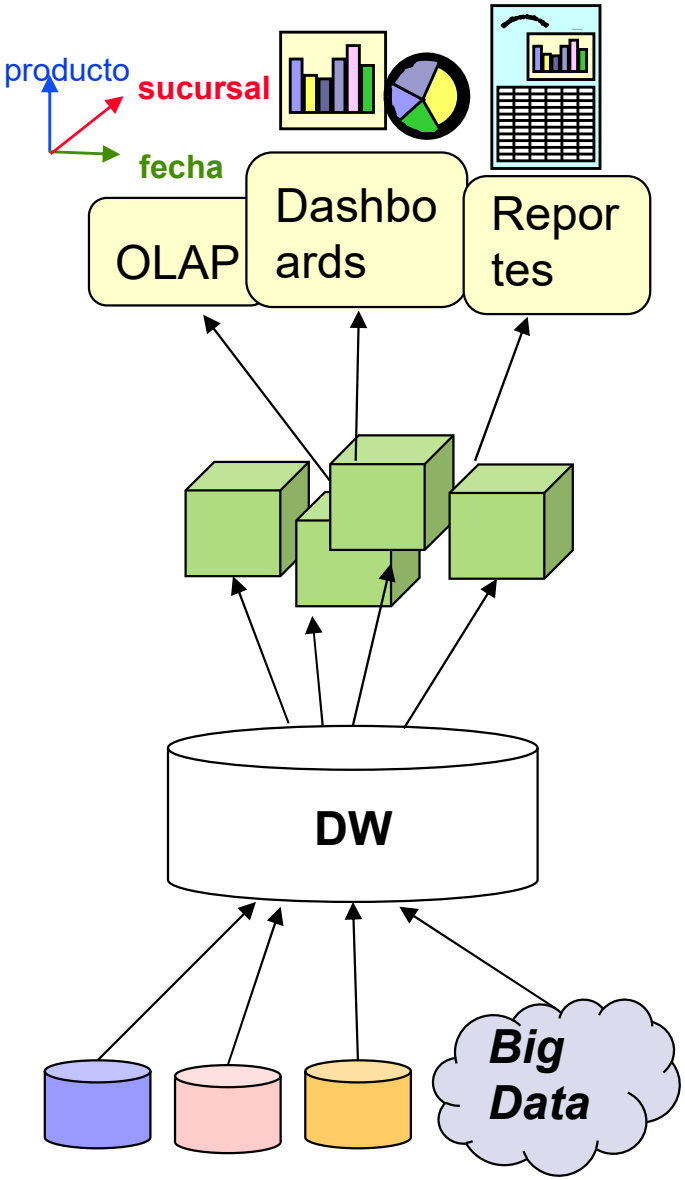


MMD en cada componente



Sistemas de Data Warehouse

Sistemas de Data Warehouse



Sistemas de Data Warehouse

- Definiciones:
  - Data Warehouse [Inmon 94]:
    - Es un conjunto de datos orientados a temas, integrados, no volátiles e históricos, organizados para soportar un proceso de toma de decisiones.
  - Sistema de Data Warehouse:
    - Es un sistema informático capaz de ofrecer información para toma de decisiones, y cuya pieza principal es un Data Warehouse.

Data Warehouse

- Datos Orientados a Temas:
  - En los DW, los datos se organizan en torno a los Temas principales de la organización
- Datos integrados:
  - Heterogeneidad de datos:
    - Diferentes áreas de la organización o externos a la organización
    - Diferentes tipos (bases de datos, datos semi-estructurados, geográficos, documentos, etc.)
  - Aspectos a resolver en la integración:
    - Unificación de conceptos
    - Construcción del dato integrado a partir de los fuentes

Data Warehouse

- Datos históricos:
  - Se deben manejar los datos con su referencia temporal
- Datos no volátiles:
  - Los datos deben ser lo suficientemente estables como para permitir análisis “largos” sin que cambien durante el mismo.
  - Esto se obtiene como consecuencia de:
    - La historización
    - La planificación de la carga

BD Operacional vs. DW

Aspect	Operational databases	Data warehouses
User type	Operators, office employees	Managers, high-ranking executives
Usage	Predictable, repetitive	Ad hoc, nonstructured
Data content	Current, detailed data	Historical, summarized data
Data organization	According to operational needs	According to the analysis problem
Data structures	Optimized for small transactions	Optimized for complex queries
Access frequency	High	From medium to low
Access type	Read, update, delete, insert	Read, append only
Number of records per access	Few	Many
Response time	Short	Can be long
Concurrency level	High	Low
Lock utilization	Necessary	Not necessary
Update frequency	High	None
Data redundancy	Low (normalized tables)	High (unnormalized tables)
Data modeling	ER model	Multidimensional model
Modeling and implementation	Entire system	Incremental

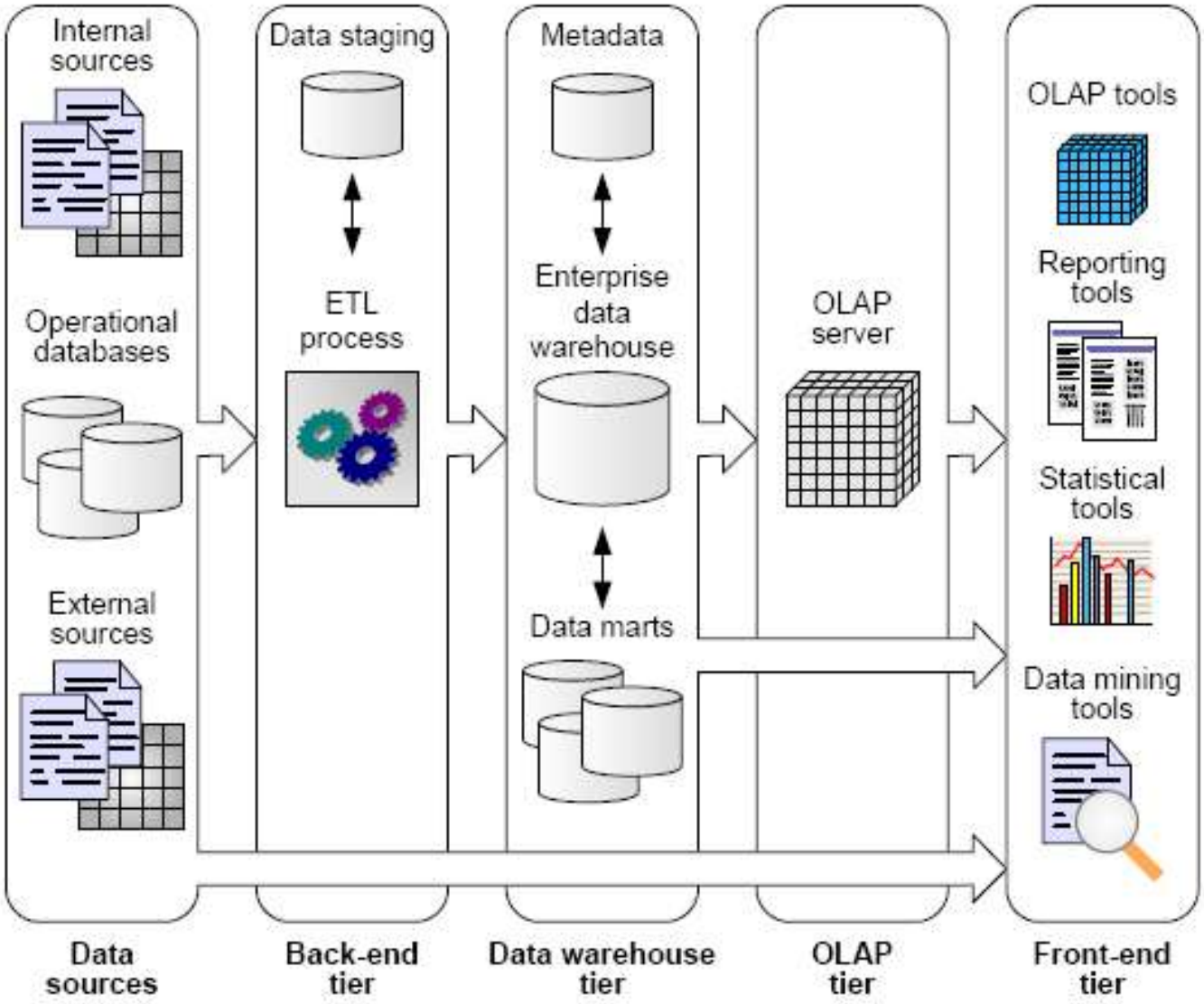
[Malinowski 2008]

Business Intelligence

- Conceptos y métodos para mejorar toma de decisiones de negocio utilizando sistemas de soporte basados en hechos. [Bouman 2009 - Pentaho]

- Conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la compañía) en información estructurada, para su explotación directa (reporting, análisis OLTP / OLAP, alertas...) o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio. [http://www.sinnexus.com/business_intelligence/]

DW clásico - Arquitectura



[Malinowski 2008]

DW clásico - Arquitectura

- Capa *back-end*
  - Extracción, transformacion y carga
- Capa *Data Warehouse*
  - DW Corporativo (Enterprise DW)
  - Data Marts
  - Metadata
- Capa *OLAP*
  - Datos multidimensionales provenientes del DW
  - Servidor OLAP: ROLAP, MOLAP o HOLAP
- Capa *front-end*
  - Herramientas clientes que permiten a los usuarios explotar los contenidos del DW

DW - Propiedades

- Relación adecuada con BD Fuentes:
  - Acceso a BDs heterogéneas y multiplataforma
  - Independiente de los Sistemas de Producción
    - Razones de performance: Un SDW "pesado" no debe acceder on-line a BD-Producción
      - Recargaria el Sistema de Producción
      - La performance de los dos se degradaría

Acceso a BD Fuentes heterogéneas

- BD Fuentes heterogéneas:
  - Diferentes modelos de datos:
    - Relacional.
    - Archivos legados (legacy)
    - Geográficos
    - Datos semi-estructurados (ej., JSON) y no estructurados (ej., documentos)
    - Fuentes externas de datos (ej.: datos de la Web, cotizaciones bolsa, indicadores socio-económicos)
  - Diferente semántica:
    - Por ej.
      - Claves diferentes para los mismos objetos
      - Igual nombre para objetos distintos

DW - Propiedades

- Permitir acceso efectivo a usuarios finales:
  - Soportar múltiples tipos de usuarios
  - Ofrecer Interfaces a usuario avanzadas

Diferentes niveles jerárquicos:

- Directivos
- Gerentes de área
- Mandos técnicos

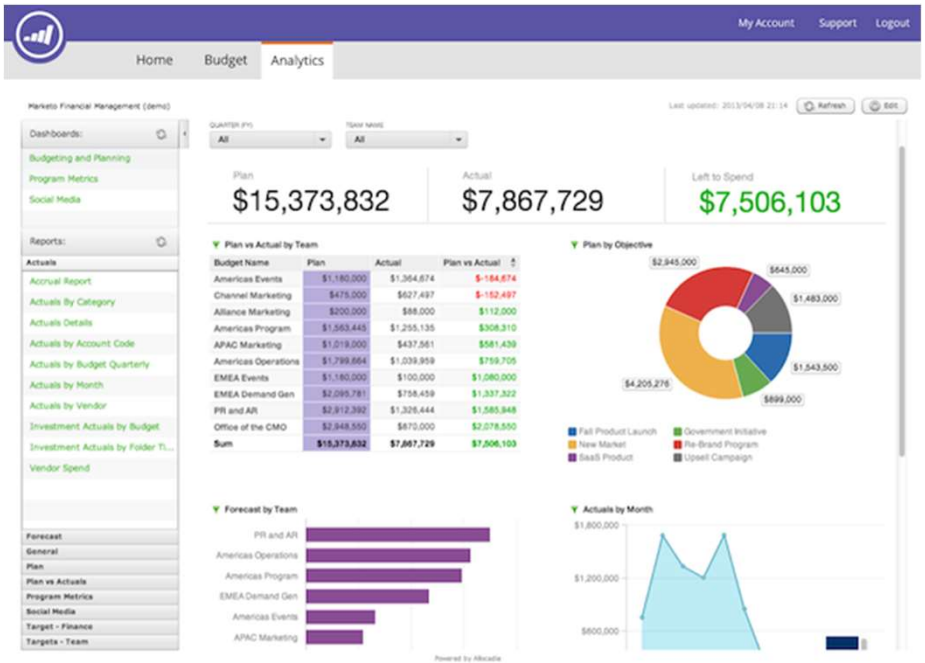
Diferentes funciones:

- Planificación
- Control
- Análisis

Interfaces de usuario avanzadas

- Interfaces de usuario especializadas.
- Por qué ?
  - Optimizar el tiempo del usuario
  - Principio:

    A cada tipo de usuario
    o aplicación se le ofrece
    la interfaz más adecuada.

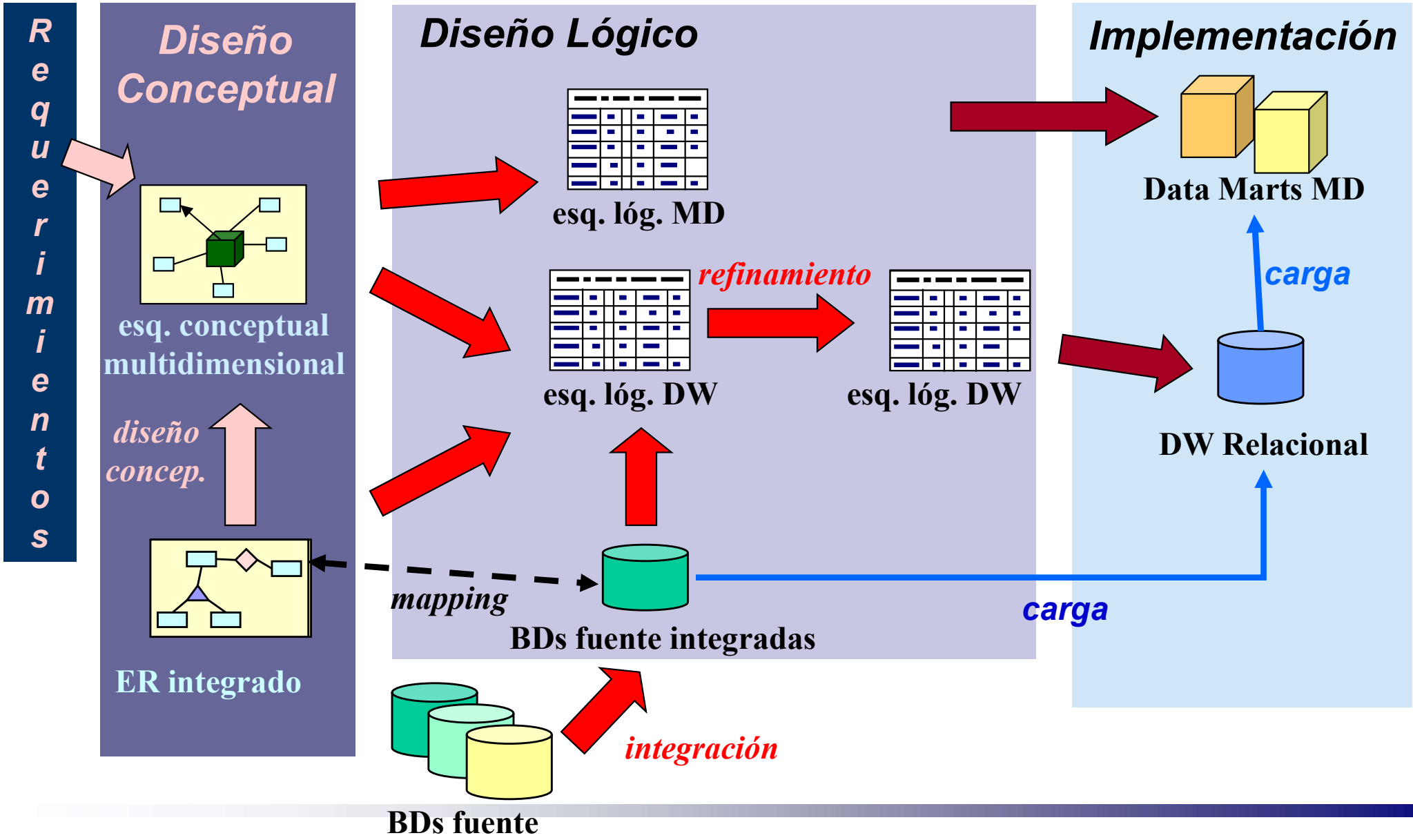


DW - Desarrollo

- Fases:
  - Especificación de Requerimientos
  - Diseño Conceptual
  - Diseño Lógico
  - Diseño Físico e Implementación

- Componentes a desarrollar:
  - Adquisición de datos
  - Almacenamiento del DW
  - Mecanismos de acceso por parte de usuarios

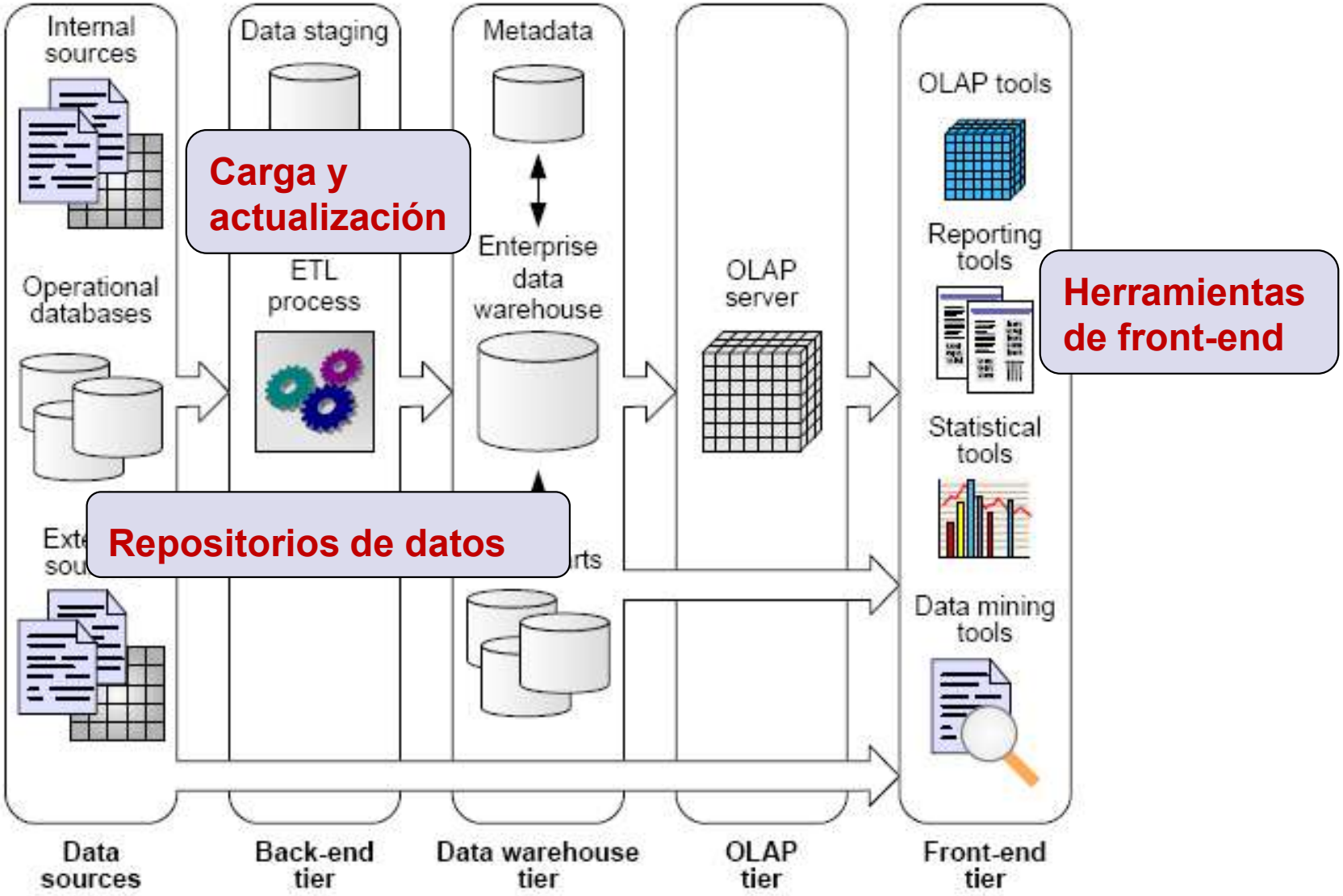
Proceso de Desarrollo



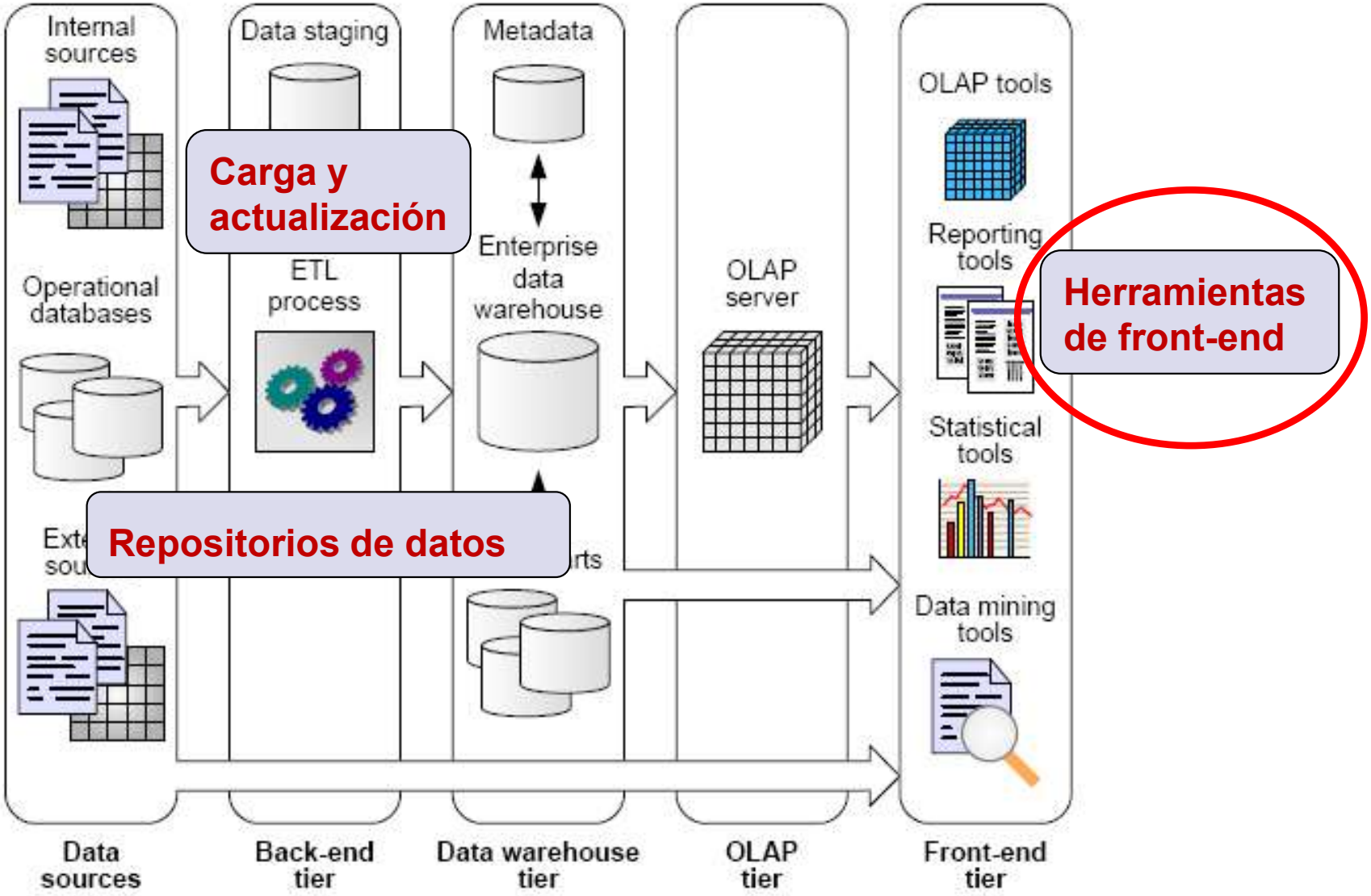
Requerimientos

- Enfoque orientado al usuario
  - Los usuarios determinan los hechos y dimensiones de análisis relevantes.
- Enfoque orientado al negocio
  - Las estructuras del DW se derivan del análisis de los requerimientos o procesos del negocio.
- Enfoque orientado a las fuentes
  - El DW se deriva de las fuentes de datos.
- Enfoque combinado
  - Combinacion de usuario/fuentes o negocio/fuentes

Entrando en detalle…



Entrando en detalle…



Herramientas Front-End

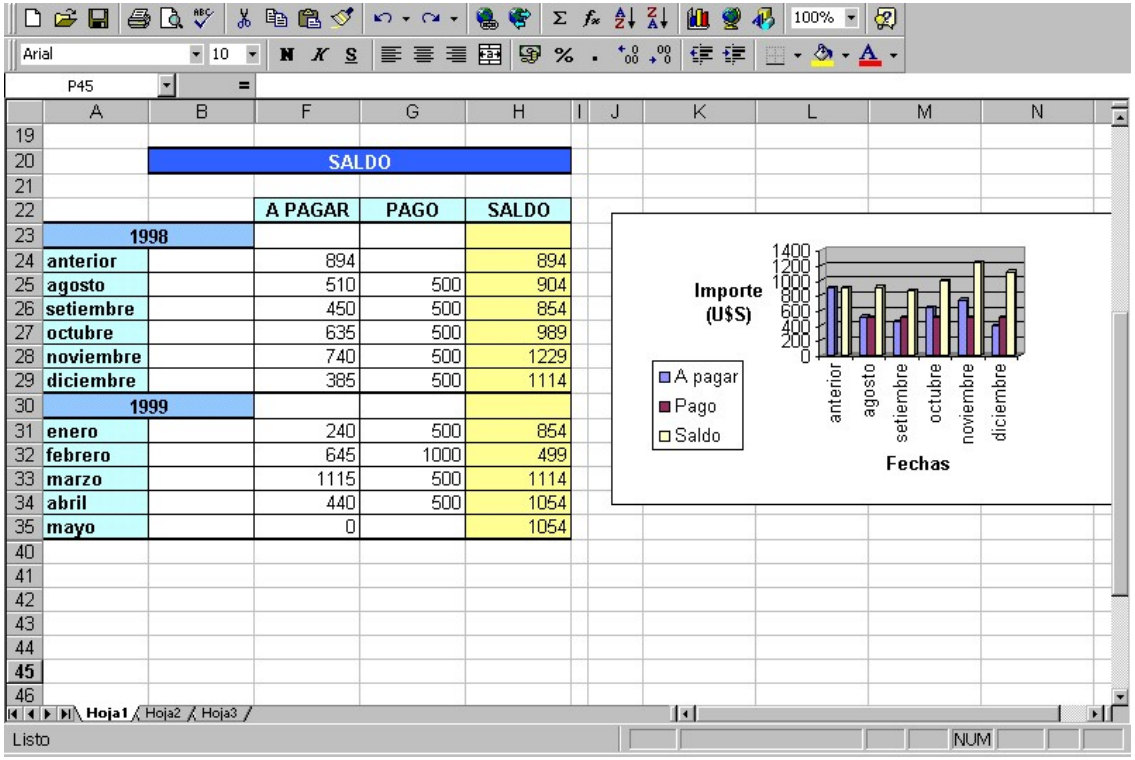
- Introducción:
  - Son herramientas usadas por el usuario para acceder a la información.
- Objetivos:
  - Ofrecer al usuario final mecanismos de acceso eficaces.
    - Mecanismos simples.
    - Mecanismos potentes.
  - Tener conexión eficaz al DW.

Herramientas Front-End

- Diferentes tipos:
  - Planillas Electrónicas
  - Herramientas de consulta y reportes
  - Herramientas OLAP
  - Herramientas de Dashboard
  - Herramientas de Data Mining

Planillas Electrónicas

- Estructuración de datos y operaciones cercanas a la visión del usuario.
- Conocidas por los usuarios en general.
- Proveen variedad de funciones y operaciones.

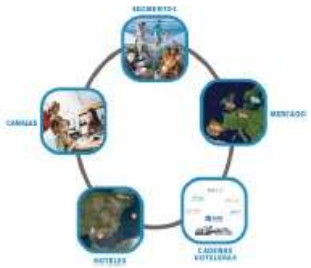


Herramientas Consultas y Reportes

- Construir fácilmente consultas/reportes complejos.
- Sirven para construir reportes no previstos.
- Ofrecen diferentes niveles de complejidad orientada a diferentes tipos de usuario:
  - Construcción de reporte complejo desde cero.
  - Construcción de reporte en base a templates.
  - Ejecución parametrizada de reportes.
  - Ejecución fija de reporte.

Herramientas Consultas y Reportes

VISITAS POR PAIS



DEPARTAMENTO: IA			
	CLIENTE: 1405165048347464		
		Importe	Visitas
		1.203,6	1
	CLIENTE: 3696111157712011		
		Importe	Visitas
		480,26	1
Total Dpto		1.683,86	2
DEPARTAMENTO: KY			
	CLIENTE: 2067662480768564		
		Importe	Visitas
		125,68	1
Total Dpto		125,68	1
DEPARTAMENTO: NC			

Herramientas OLAP

- Permiten consultar datos :
  - Interactivamente y en forma eficiente.
  - Usando mecanismos comprensibles para usuarios.
    - Una consulta corresponde a cruzar dimensiones y elegir la medida en el cruzamiento.
- Visualización gráfica

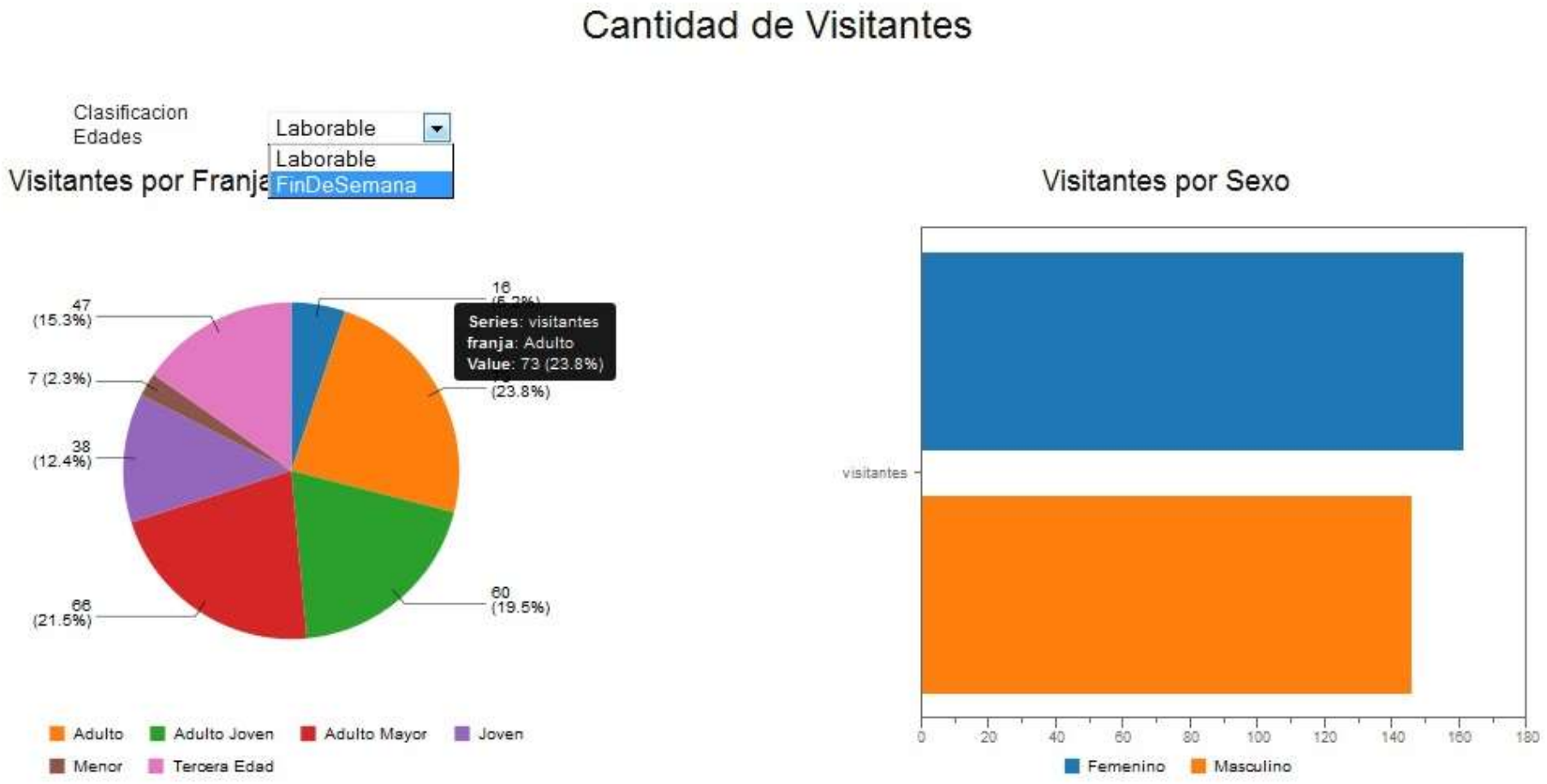
Herramientas OLAP

- Implementan Modelos Multidimensionales.
  - Los Modelos MD representan los datos como dimensiones en un cubo de “n” dimensiones.
- Diferentes alternativas tecnológicas:
  - ROLAP vs. MOLAP vs. HOLAP:
    - *ROLAPs*: actuan directamente sobre BD Relacional.
    - *MOLAPs*: trabajan sobre almacenamiento especializado.
    - *HOLAP:* intentan aplicar ambas estrategias.

Herramientas OLAP

- Cualidades de herramientas OLAP:
  - Acceso a BDs.
  - Valorización de datos
    - Capacidad de cálculos
    - Capacidad de operaciones de análisis
    - Variedad de presentación de resultados
  - Adaptación a diferentes tipos de usuario
  - Control de operaciones del usuario
    - Control de usuarios
    - Evitar cruzamientos de datos que no tengan sentido

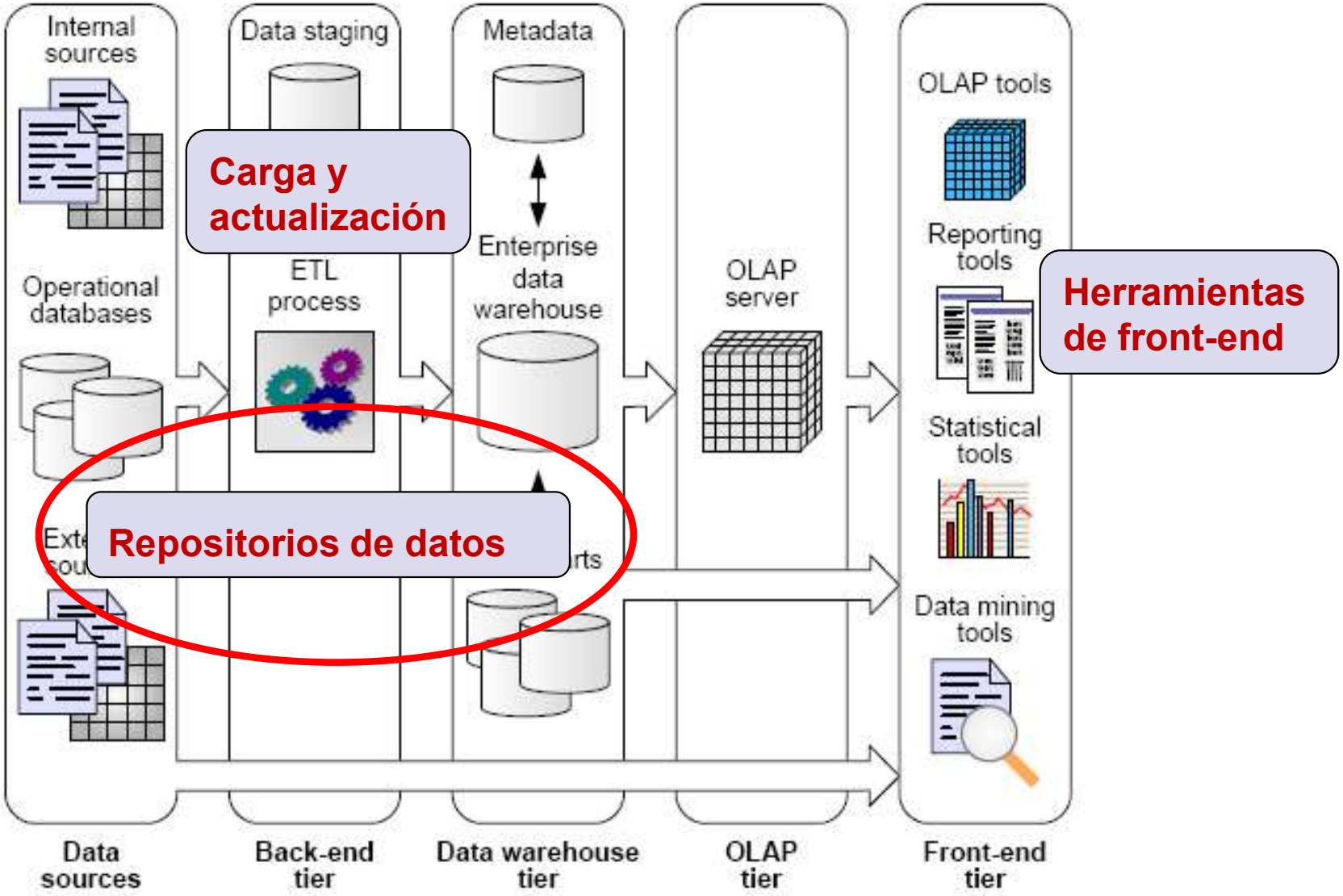
Dashboards


Cantidad de Visitantes
Clasificacion Edades
Laborable
Laborable
FinDeSemana
Visitantes por Franja
Series: visitantes
franja: Adulto
Value: 73 (23.8%)
16
47 (15.3%)
7 (2.3%)
38 (12.4%)
66 (21.5%)
60 (19.5%)
Adulto
Adulto Joven
Adulto Mayor
Joven
Menor
Tercera Edad
Visitantes por Sexo
visitantes
0
20
40
60
80
100
120
140
160
180
Femenino
Masculino

Data Mining

- Objetivos:
  - Explorar BDs buscando relaciones desconocidas entre los datos.
- Incluye un conjunto muy amplio y heterogéneo de técnicas y herramientas.
  - Pattern-matching, Clasificación, Deducción de reglas, y otros para determinar relaciones claves entre los datos
- La iniciativa es del algoritmo y no del usuario.

Entrando en detalle…



Repositorios de datos

- Data Warehouse Corporativo (CDW):
  - Brinda información gerencial a nivel de la organización
  - Alcance global
  - Información integrada, “limpia”, histórica
- Operational Data Store (ODS):
  - Contiene resultados intermedios de los procesos de carga y actualización
  - Información integrada
  - Puede ser temporal o perdurable

Repositorios de datos

- Data Marts
  - Brindan información gerencial a un área específica de negocios
    - Por ejemplo: Ventas, Marketing, Recursos Humanos
  - Basado en porciones de datos del CDW
  - Realizan la interacción directa con los usuarios
  - Fuertemente asociados a requerimientos concretos de usuarios/aplicaciones
    - Incluyen los datos más directamente relacionados
    - Aplican las tecnologías más adecuadas a c/caso
    - Administración y evolución relativamente autónoma
    - Información resumida

Clases de Datos

- Datos detallados:
  - En el ODS y en el DW
  - Diferencia con datos de BD-Fuentes:
    - Están en formato homogéneo
    - Pasaron un primer control de calidad
    - Pueden ya haber sido integrados
  - Generalmente se trata de volúmenes importantes de datos

Clases de Datos

- Datos agregados:
  - En el DW y Data Marts
  - Resultantes de aplicar funciones de totalización sobre los datos
    - Ej: total mensual de ventas por producto
  - Ventajas:
    - Información significativa para analizar
    - Permiten reducir volúmenes de datos
    - Su cálculo interactivo plantea problemas de performance

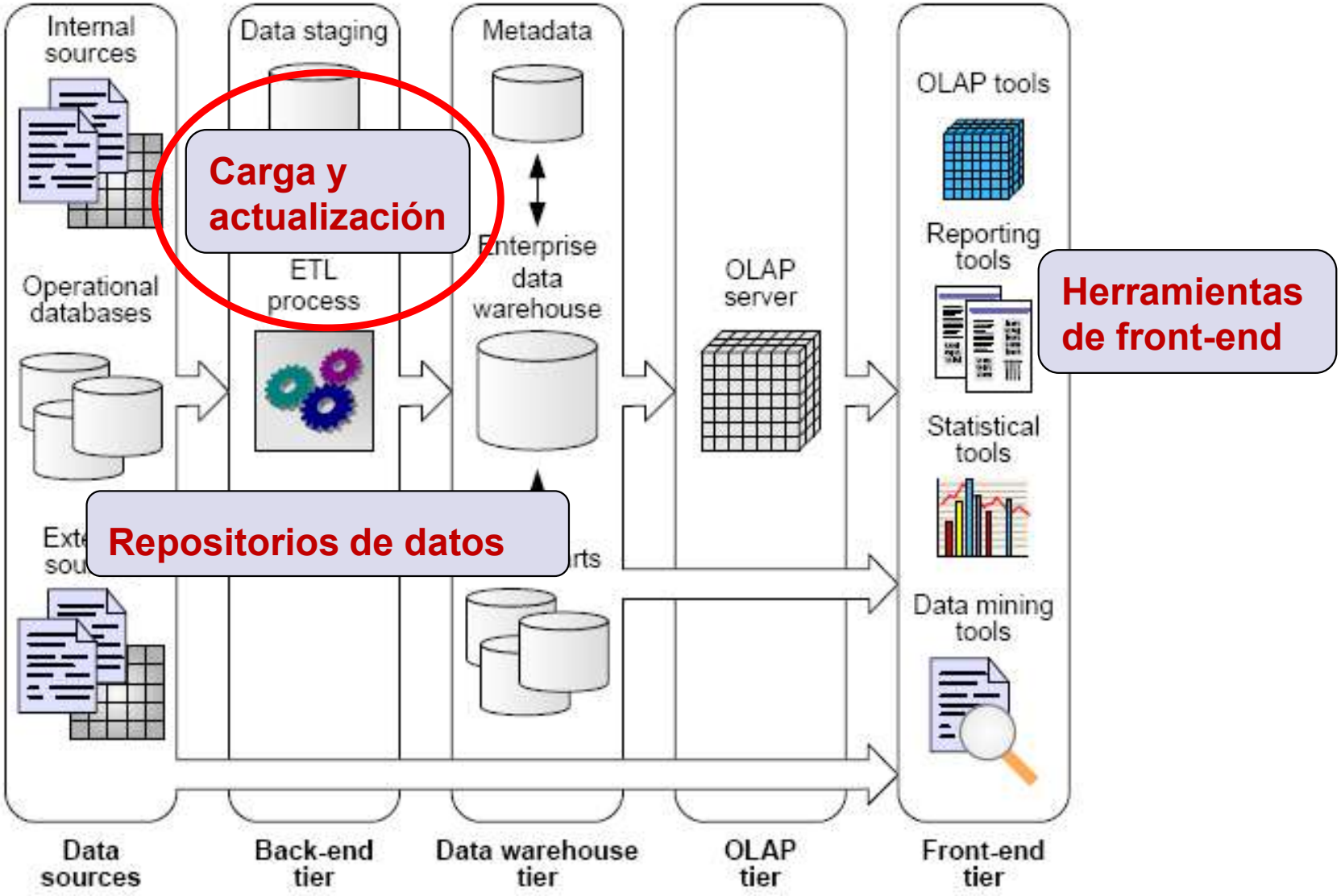
Clases de Datos

- Datos historizados:
  - En el DW y Data Marts
  - Datos (detallados o agregados) a los cuales se les agrega una marca de tiempo.
  - Generan volúmenes importantes de datos

Clases de Datos

- Metadatos:
  - Información sobre los datos del DW
    - Semántica de los datos y su localización en el DW
    - Localización de los datos en los sistemas de producción y reglas de transformación
    - Especificación de formulas de cálculo de agregados
    - Información sobre frecuencias de carga, mecanismo de historización, etc.
  - Pieza clave para:
    - El control de calidad de los datos
    - La explotación eficaz del DW
    - El mantenimiento del DW

Entrando en detalle…



Tipos de Operaciones/Transformaciones

- Extracción de datos.
  - Consiste en extraer los datos de la BD fuente y cargarlo en el ODS o DW
- Limpieza.
  - Filtrado de datos no admisibles en el DW
  - Modificación de formato o valores para que cumpla pautas definidas en el DW
- Integración.
  - Consiste en integrar datos provenientes de dos o más fuentes

Tipos de Operaciones/Transformaciones

- Cálculos y Agregaciones
  - Consiste en calcular indicadores a partir de datos base y generar agregaciones de los datos.
- Generación de datos históricos (historización).
  - Consiste en agregar marcas de tiempo a datos
- Generación de versiones.
  - Consiste en agregar atributos diferenciadores de diferentes versiones de un objeto base

Carga y actualización

- Problemas a resolver:
  - Diseño de la configuración:
    - Qué operaciones se realizan en qué momento
    - Cuáles son secuenciales y cuáles paralelizables
    - Qué dependencias hay entre ellas
  - Diseño e implementación de procesos:
    - Extracción, integración, etc.
  - Definición de frecuencia de operaciones
    - Grado de sincronización del DW con BDs fuente
    - Coordinar esta frecuencia con la lógica de las consolidaciones