Ingeniería de Datos

Data Warehouse

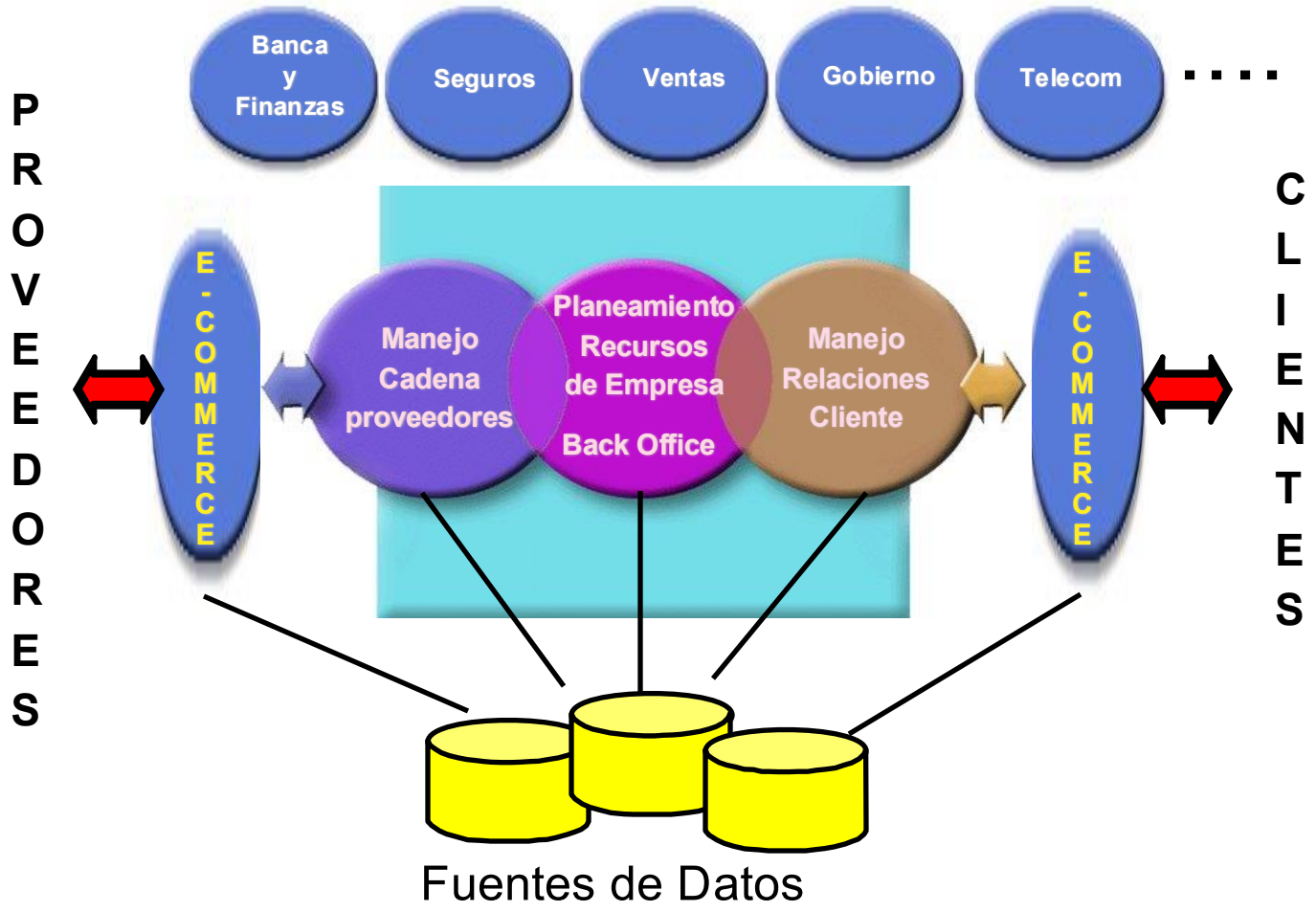
Rosario Guzmán, Milo Flores
Fernando Arimana, Pedro Wong

Junio 2006

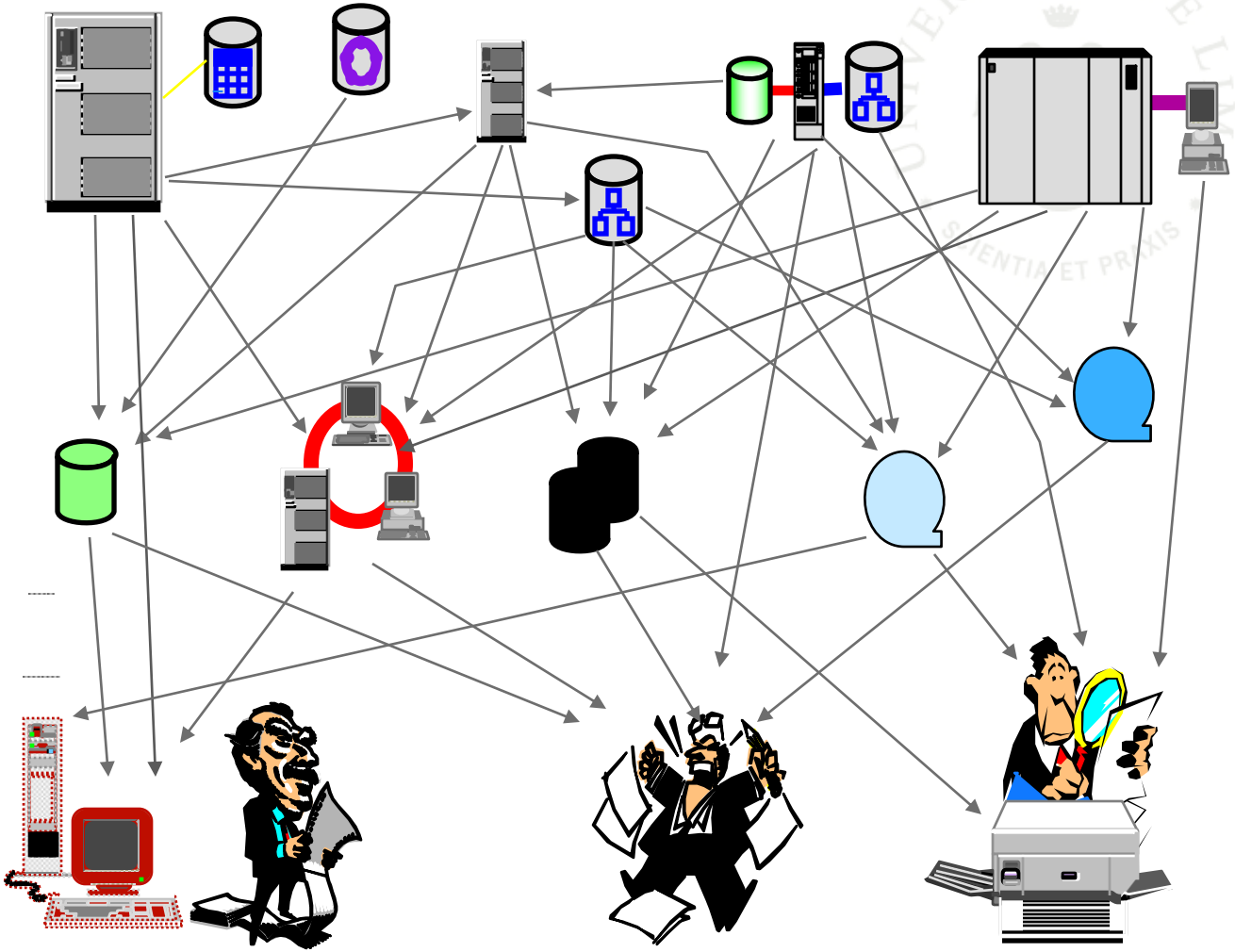
Agenda

- Introducción
- Concepto de Data Warehouse, Data Mart
- Arquitectura
- Tipos de implementación
- Extracción, transformación y carga de datos - ETL
- Metadatos
- Cubos de datos
- Explotación de datos
- Beneficios

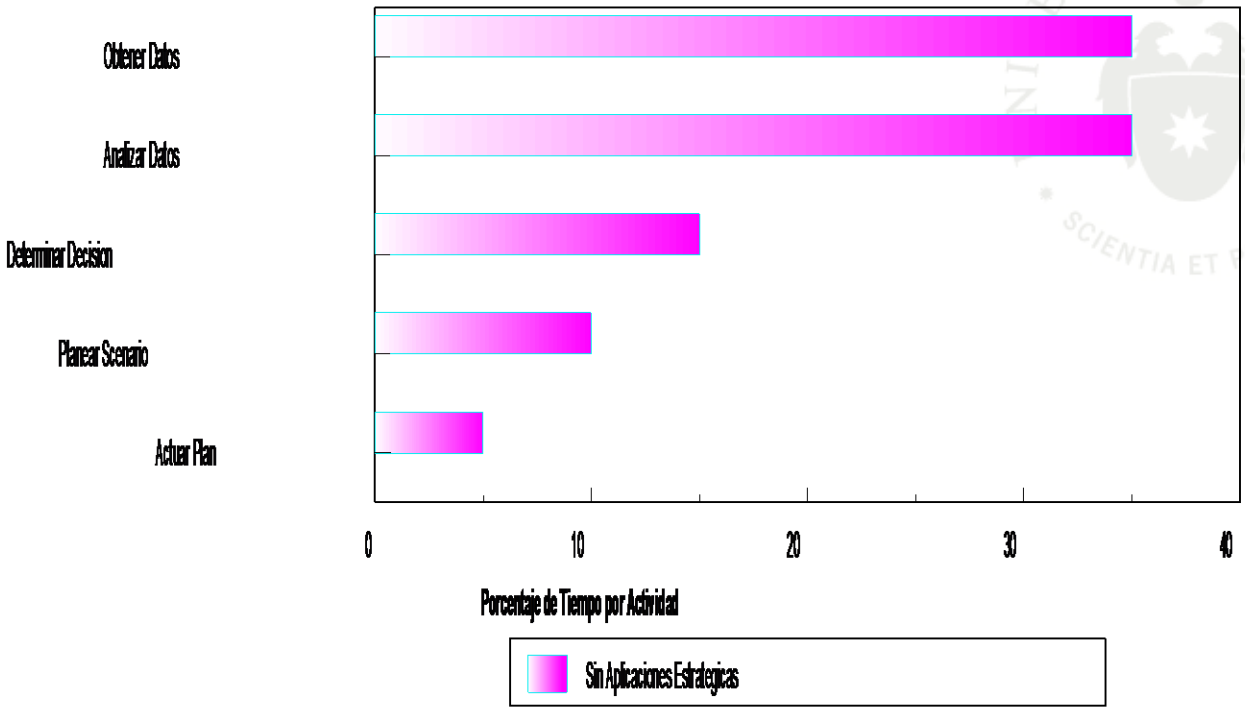
Infraestructura e-Business



Diversidad de Fuentes de Datos

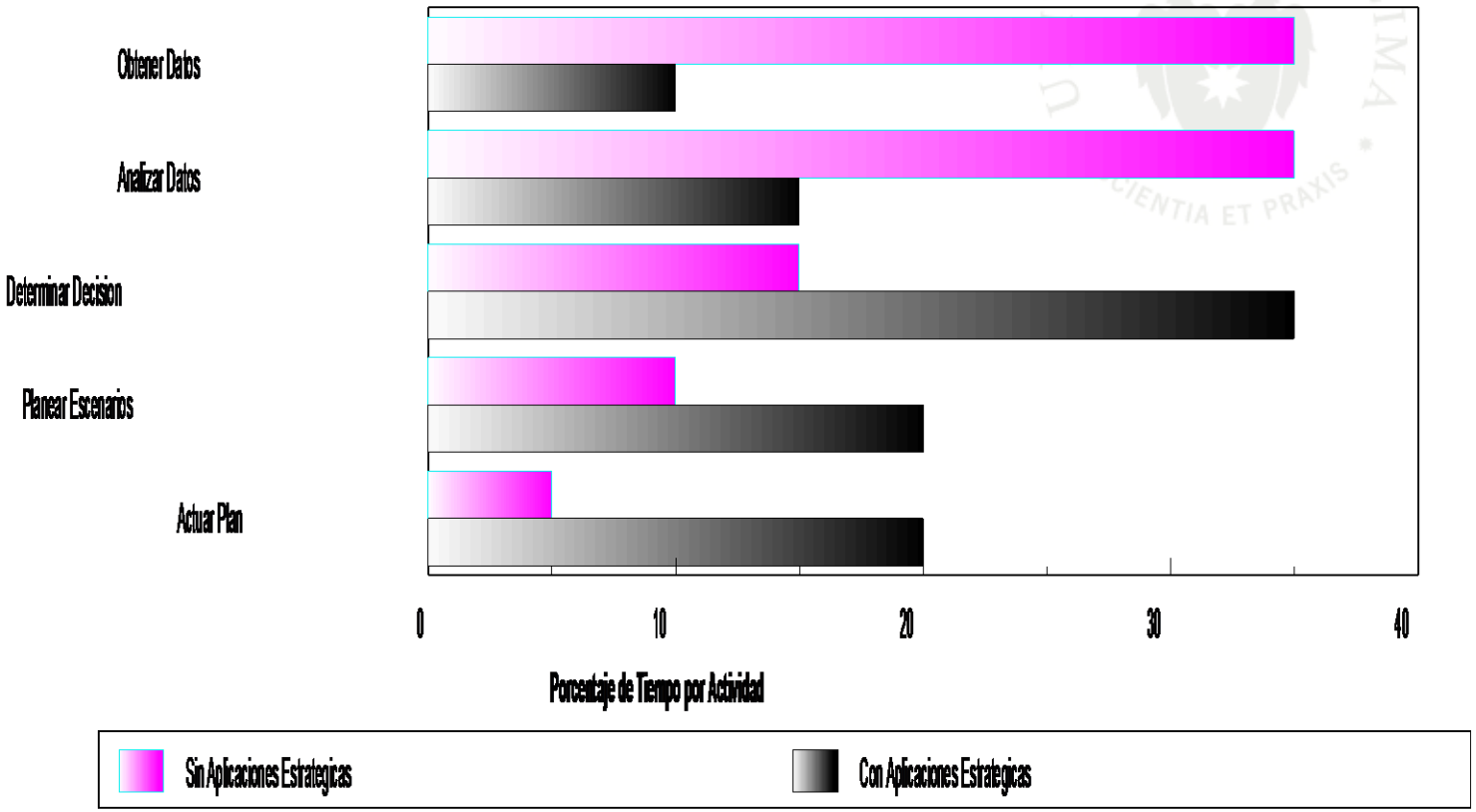


Productividad de los Usuarios de Negocios



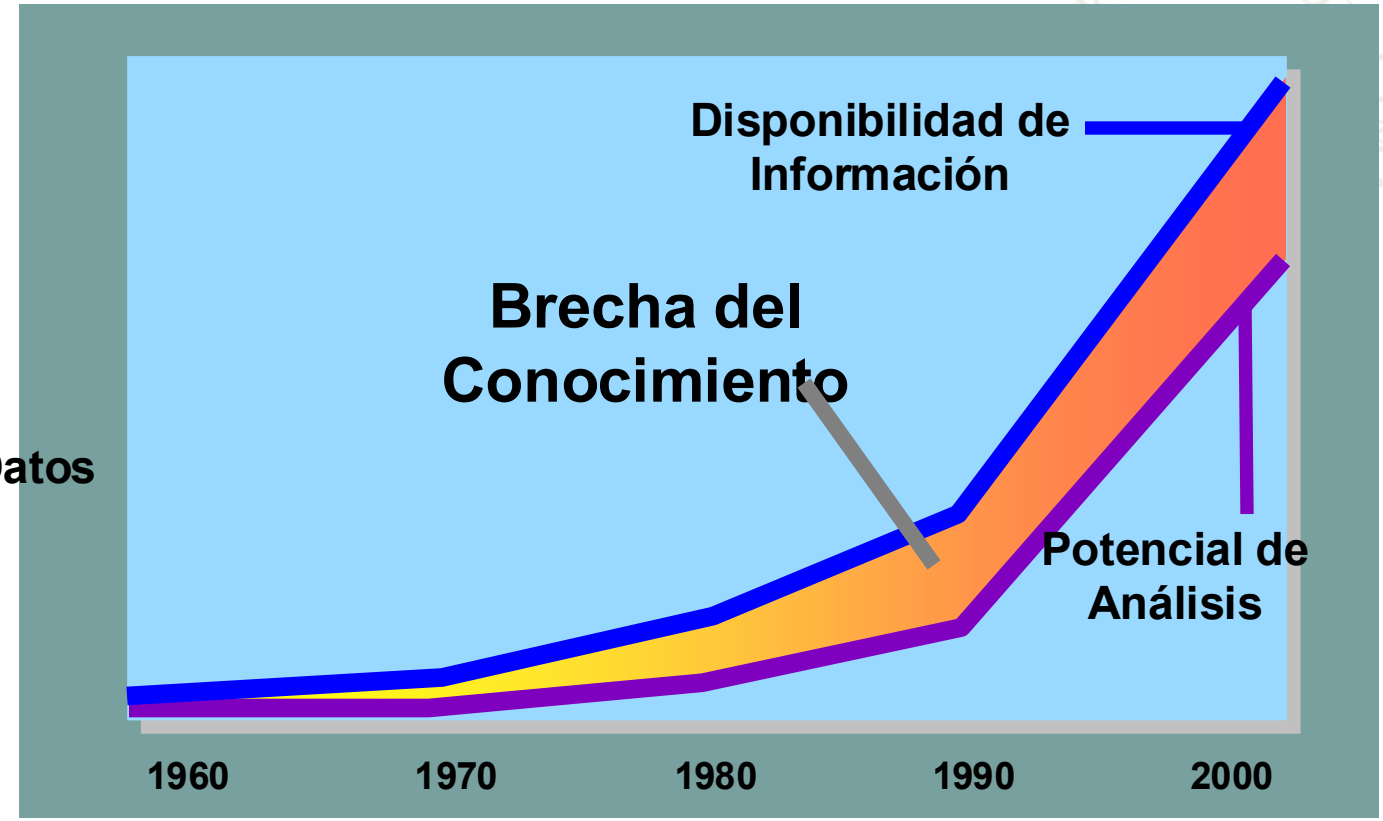
Source: Gartner Group

Productividad de los Usuarios de Negocios



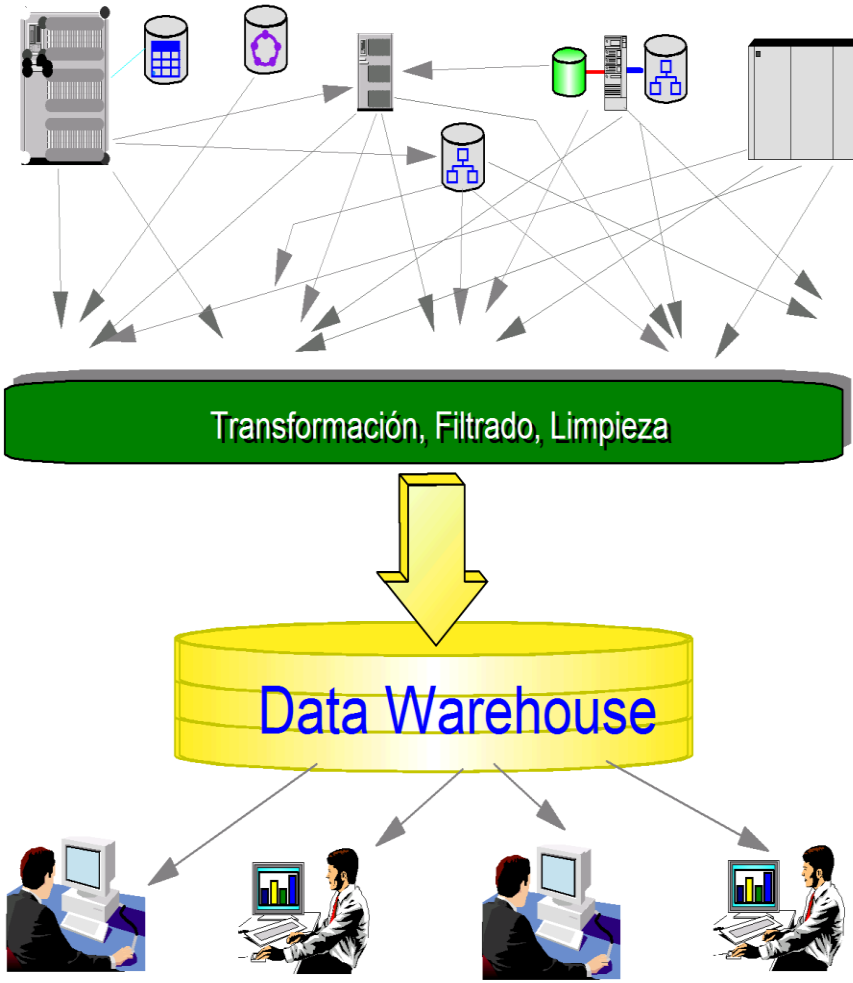
Source: Gartner Group

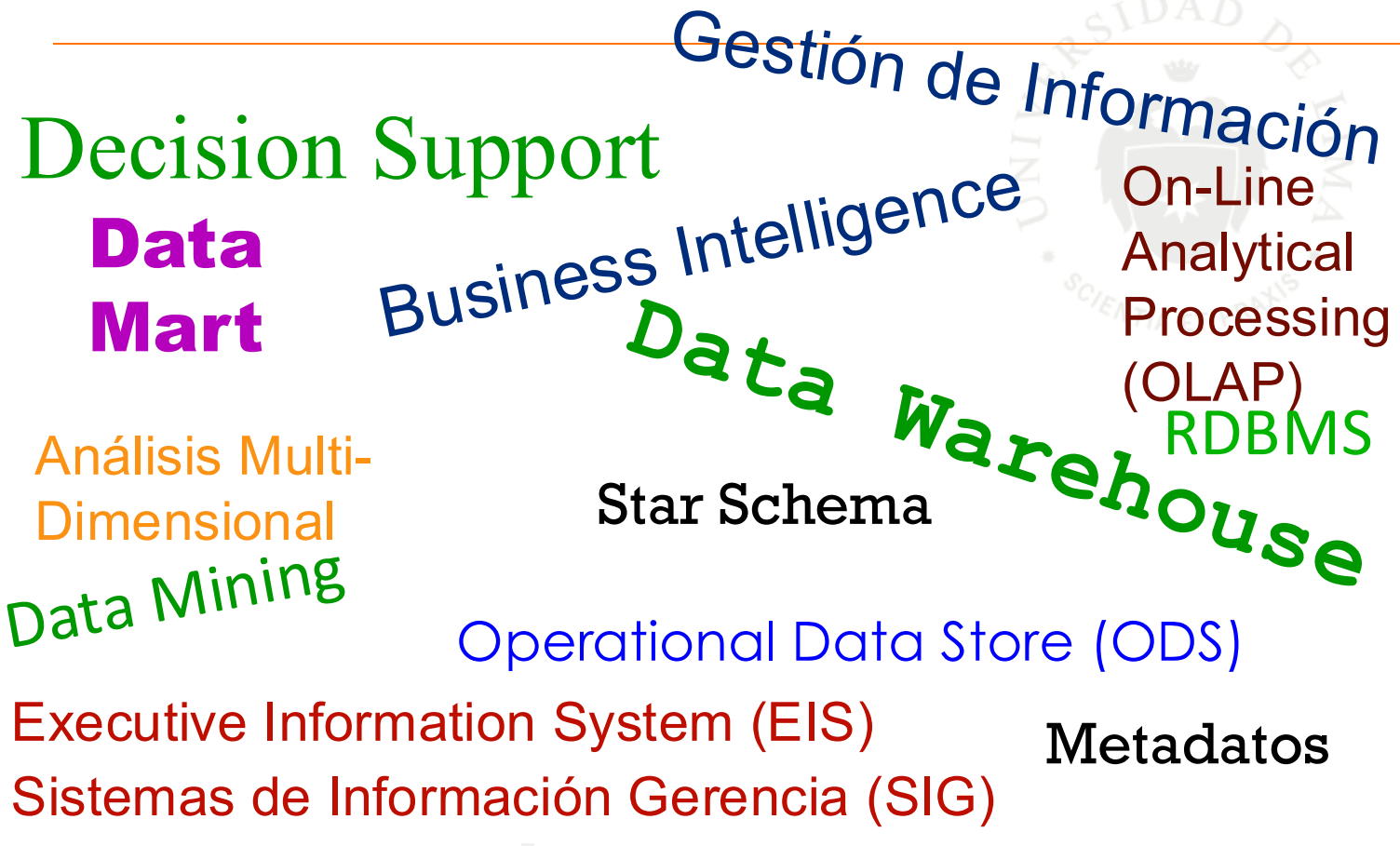
Brecha del Conocimiento Creciente



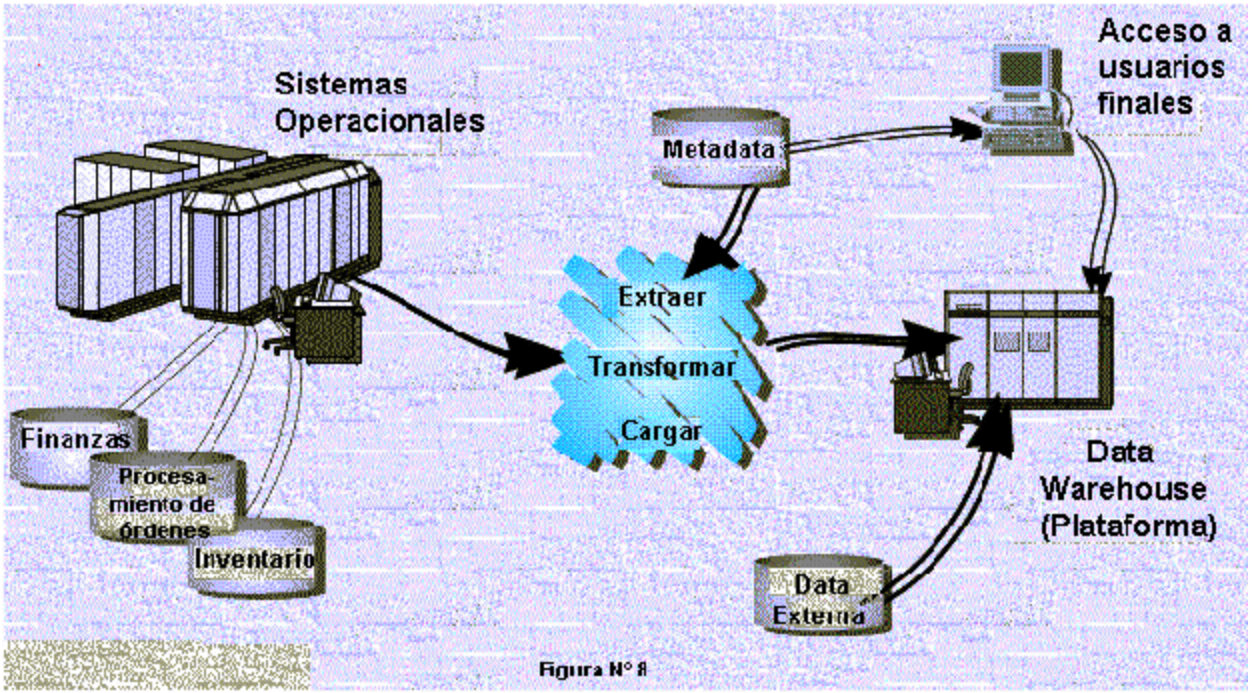
Source: Gartner Group

Data Warehousing




Gestión de Información
Decision Support
Data Mart
Business Intelligence
On-Line Analytical Processing (OLAP)
Data Warehouse
RDBMS
Análisis Multi-Dimensional
Star Schema
Data Mining
Operational Data Store (ODS)
Executive Information System (EIS)
Metadatos
Sistemas de Información Gerencia (SIG)

Data Warehousing



Data Warehouse

data warehouse *n.* Area de almacenamiento para Información como Soporte para Toma de Decisiones. **Almacena** Data recolectada de Sistemas Operacionales Diversos y Externos, **Integra** la Data en un Modelo de Negocios, **Permite el Análisis** de la Data y **Entrega Información** a las personas de Toma de Decisiones a lo largo de la Organización.

Data Mart

data mart *n.* Areas de Almacenamientos de **Grupos de Trabajo** o **Departamentales** que son **Pequeños** en tamaño y **Especializados** en Funciones. El Data Mart contiene Data Informacional que es ajustada a las Necesidades de Departamentos de Trabajo Específicos. También ofrece **Mejoras de Performance.**

En otras palabras .....

Presenta una vista de Unidad de Negocio de la Información

Usualmente Creada para Mejorar Performance de Acceso/Análisis

Según una definición formal de Bill Inmon:

Organizada por Temas: Información del mismo evento o tema relacionada.

Variante en el Tiempo: Lo cambios en la data son auditables,

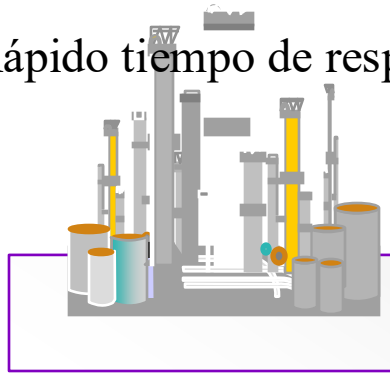
No Volatil: Información permanente,

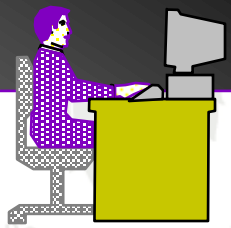
Integrada: Amplia y Consistente

Diferencias entre Sistemas Operacionales y Sistemas Analíticos

Sistemas Operacionales

- Ejecuta el negocio
- Empleado por
  - –Oficinistas/administradores
- Data Actualizada al segundo
- Altos Volúmenes de Transacciones simples
- Rápido tiempo de respuesta



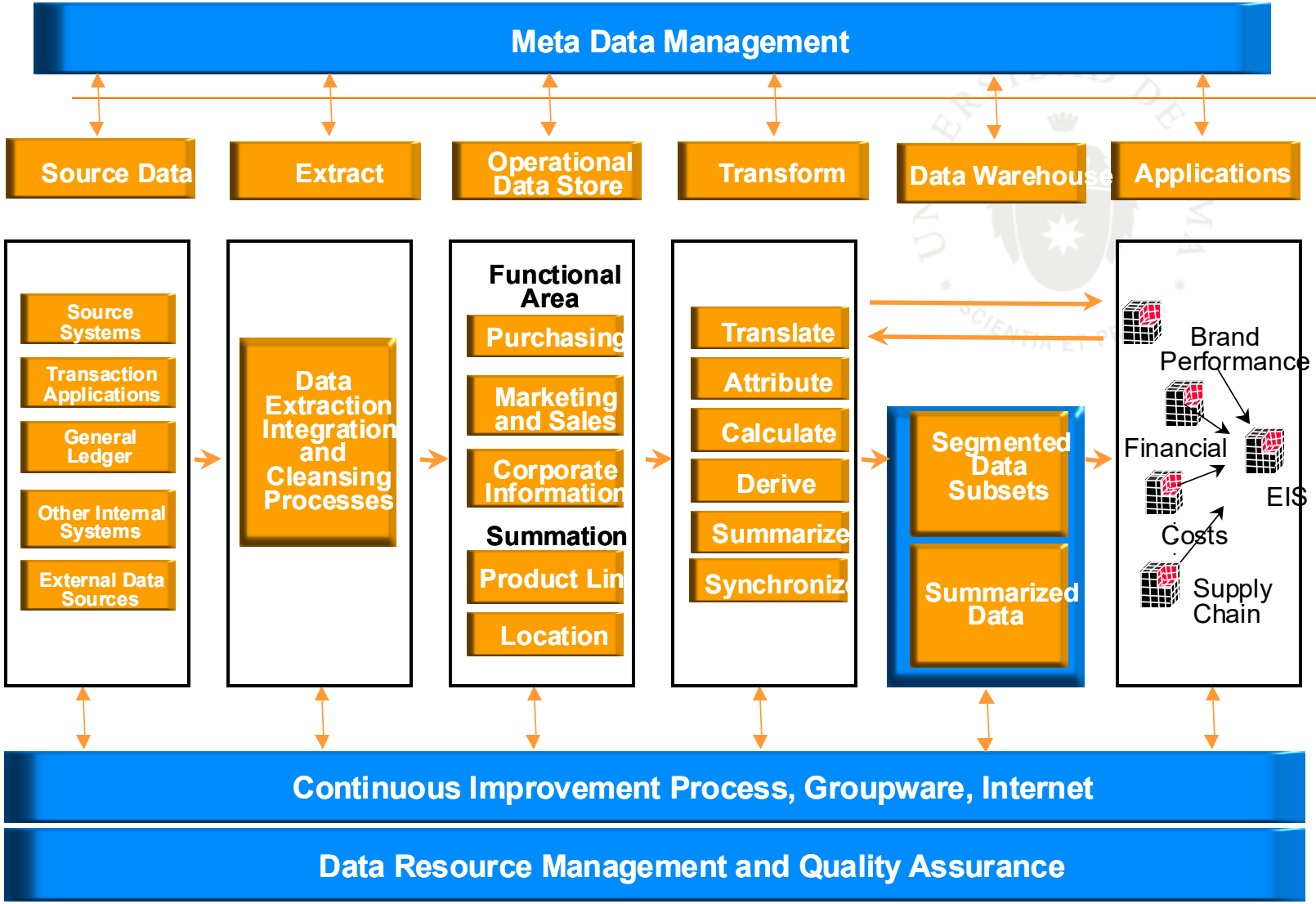


Sistemas Analíticos

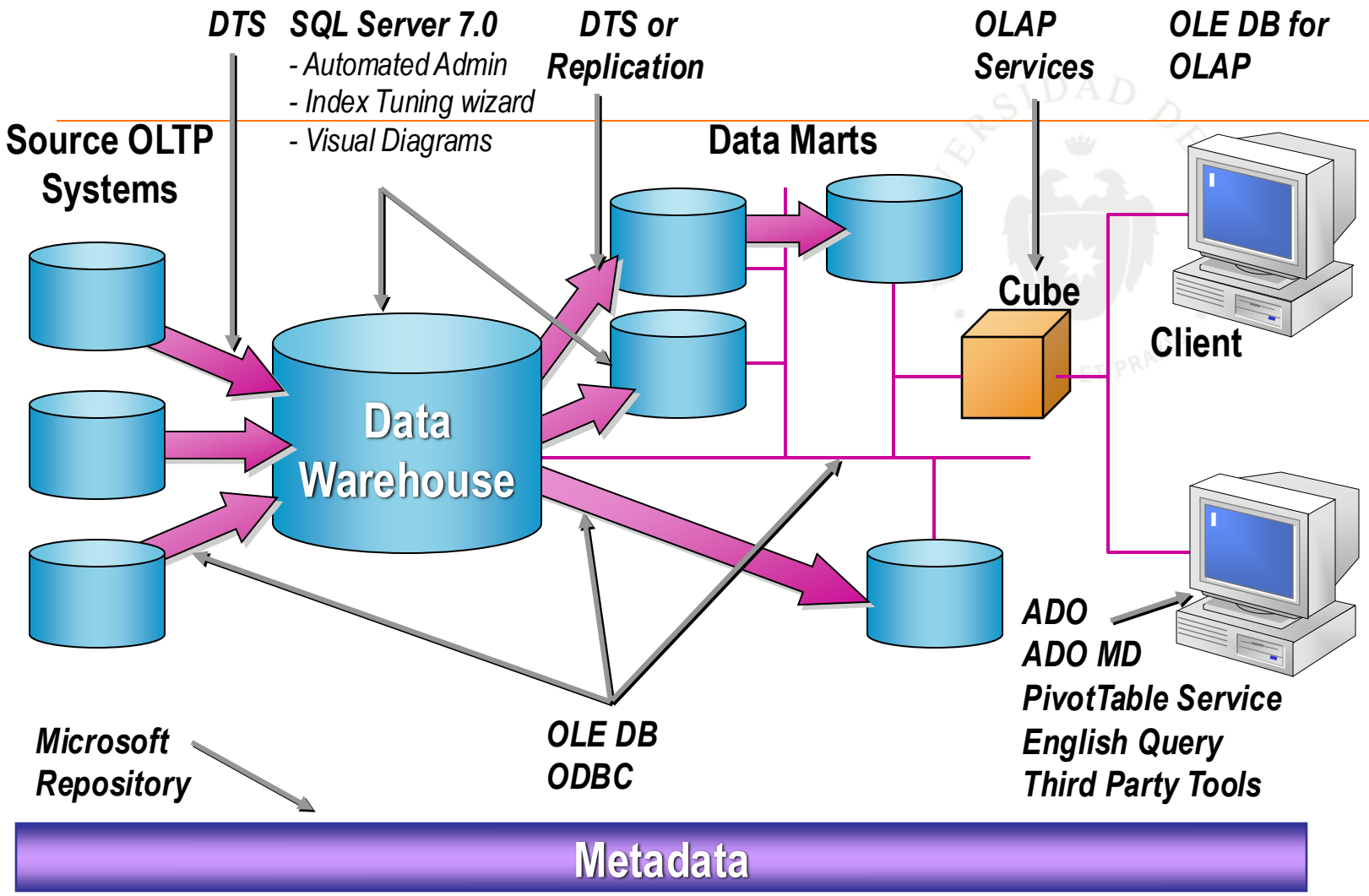
- Administra el negocio
- Empleado por
  - –Decision makers
  - –Empleados con experiencia
- Detallado y sumarizado
- Integrado
- Histórico

Arquitectura

Arquitectura Referencial

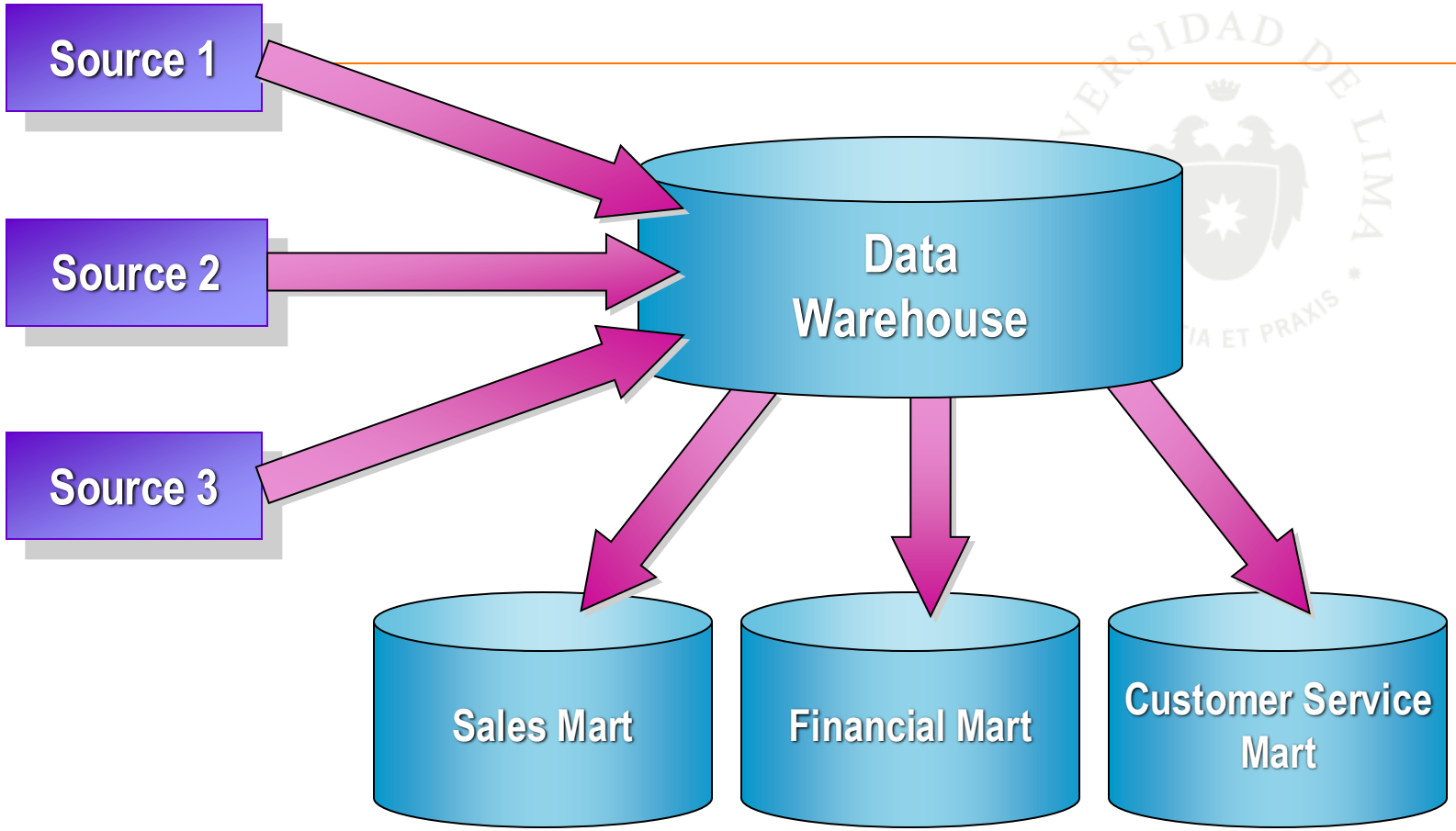


Ambiente de Trabajo MS

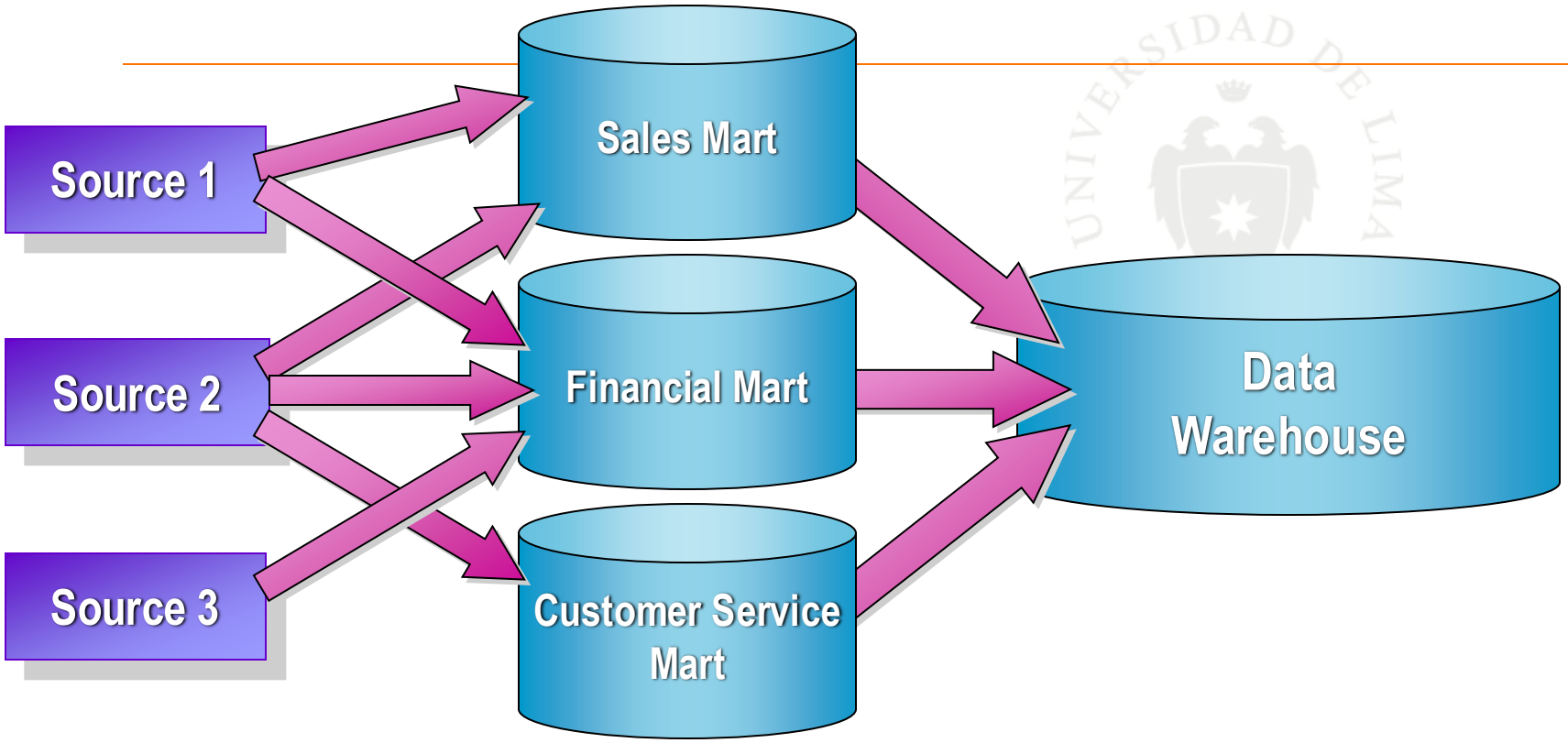


Tipos de Implementación

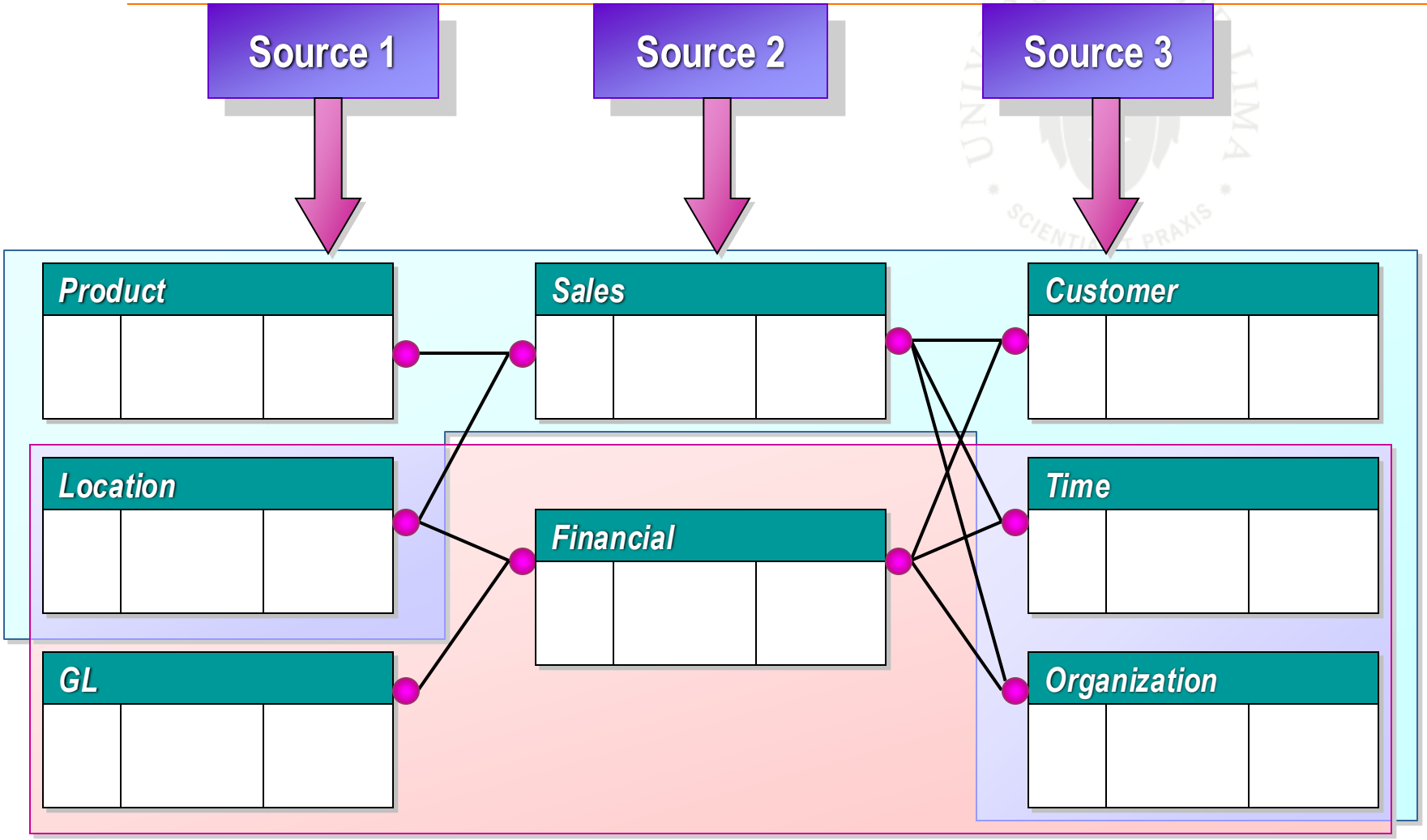
Implementación:
Data Warehouse -> Data Mart



Implementación:
Data Mart -> Data Warehouse



Integrando Data Marts



Extracción, Transformación y Carga de datos (ETL)

"80% del tiempo y recursos necesarios para crear y mantener un Data Warehouse son necesarios en la Extracción, Limpieza y Carga de la Data".

Source: Bill Immon

Extracción, Transformación y Carga de datos

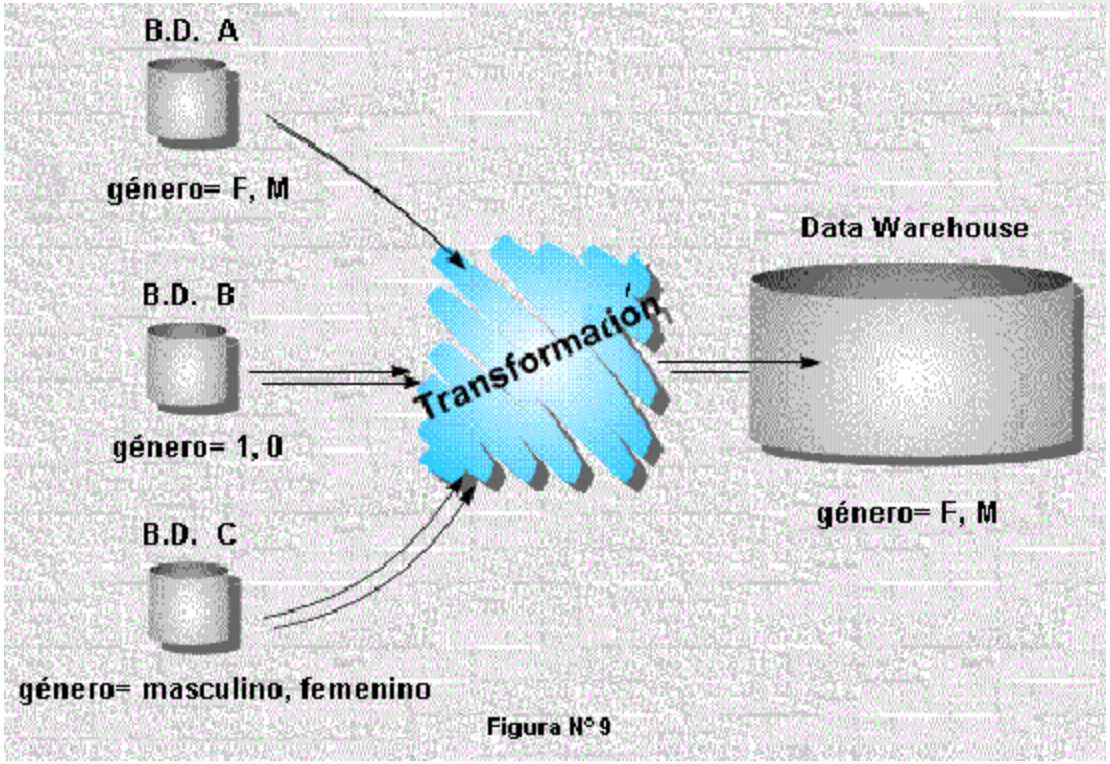
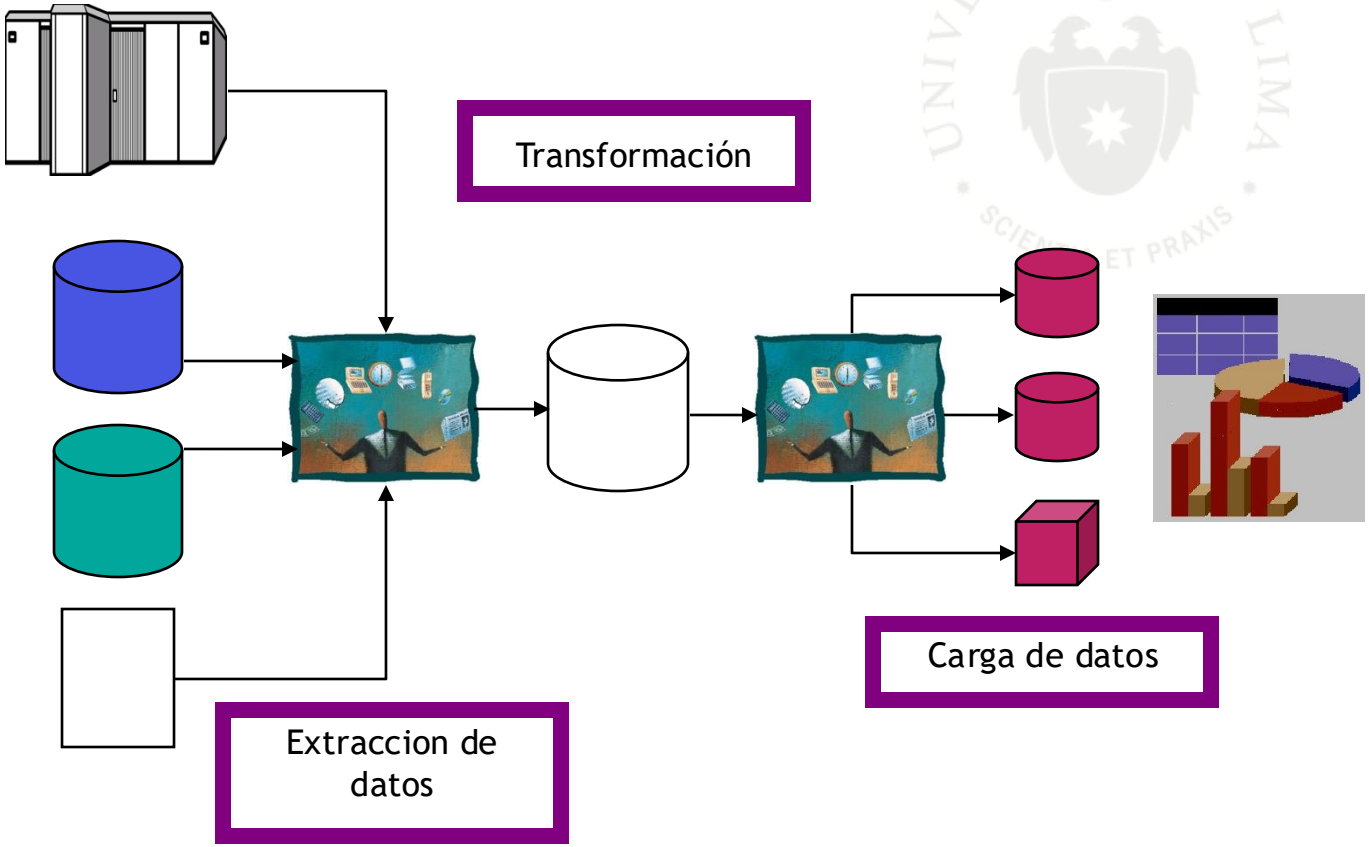
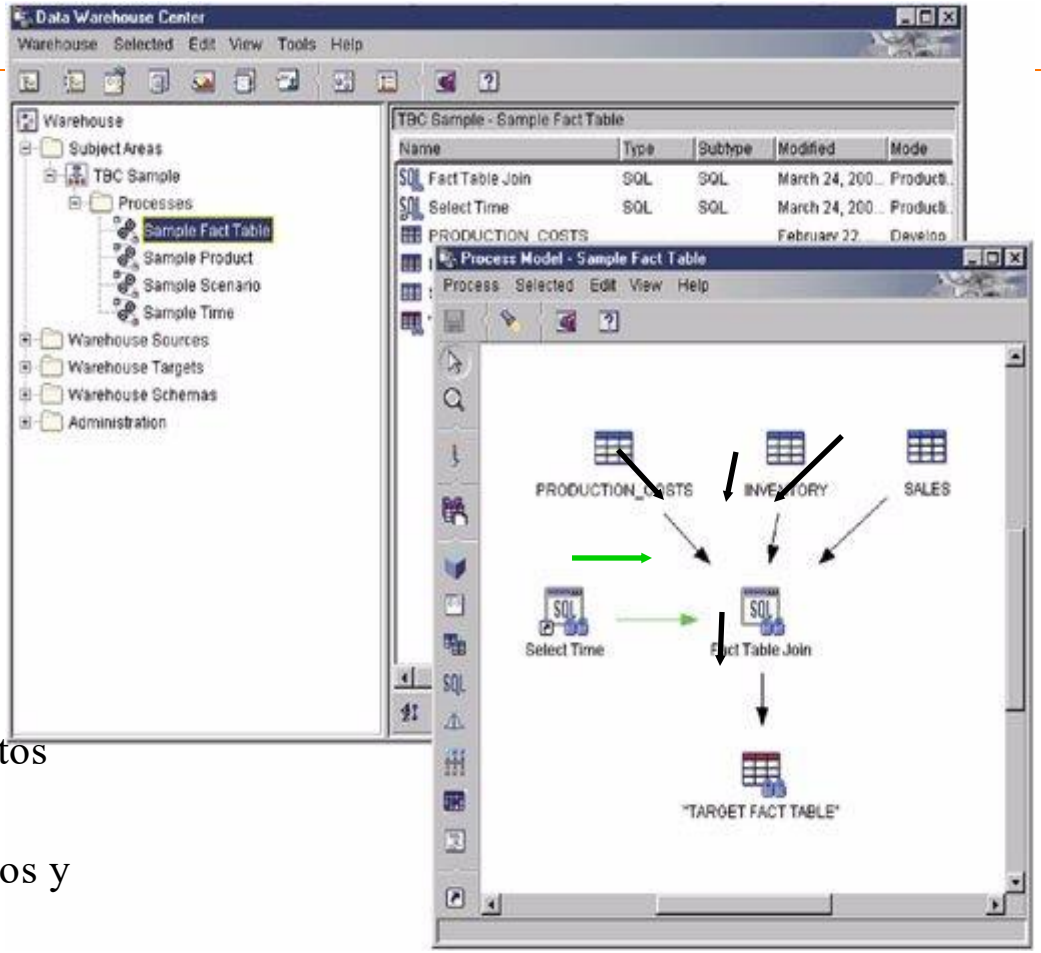


Figura N° 9

Herramienta ETL - Warehouse Manager



Warehouse Manager



Registra y accede a las fuentes de datos

- DB2, Oracle, Microsoft, Sybase, Informix, archivos planos y otros

Define extracciones y transformaciones

- Mas de 100 transformaciones incorporadas

Define carga y actualización del warehouse

- Full refresh, histórico, y movimento de datos

Modela, automatiza, y monitorea procesos

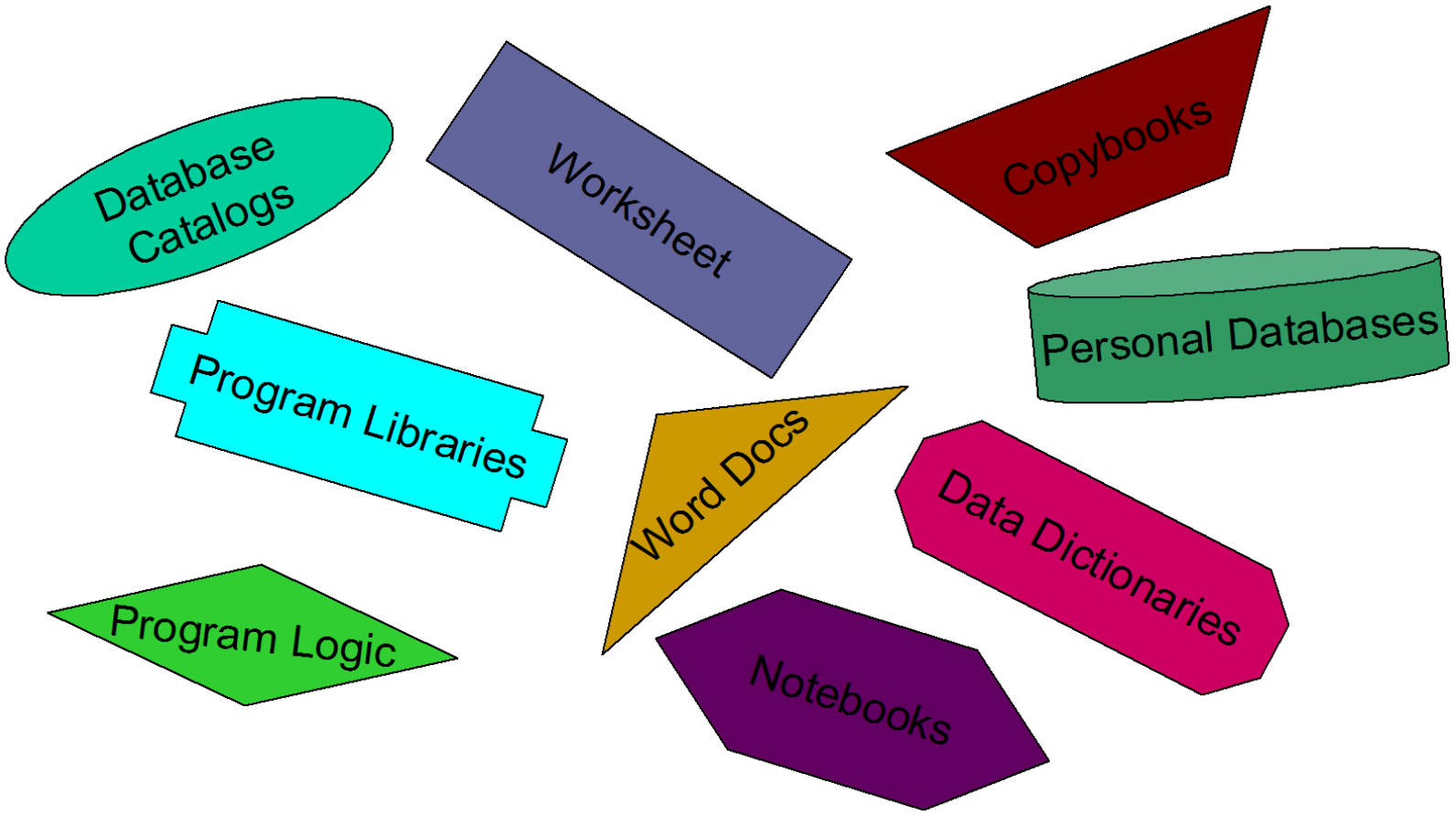
- Programa, triggers, dependencias, reintentos y notificaciones

Maneja e intercambia metadatos

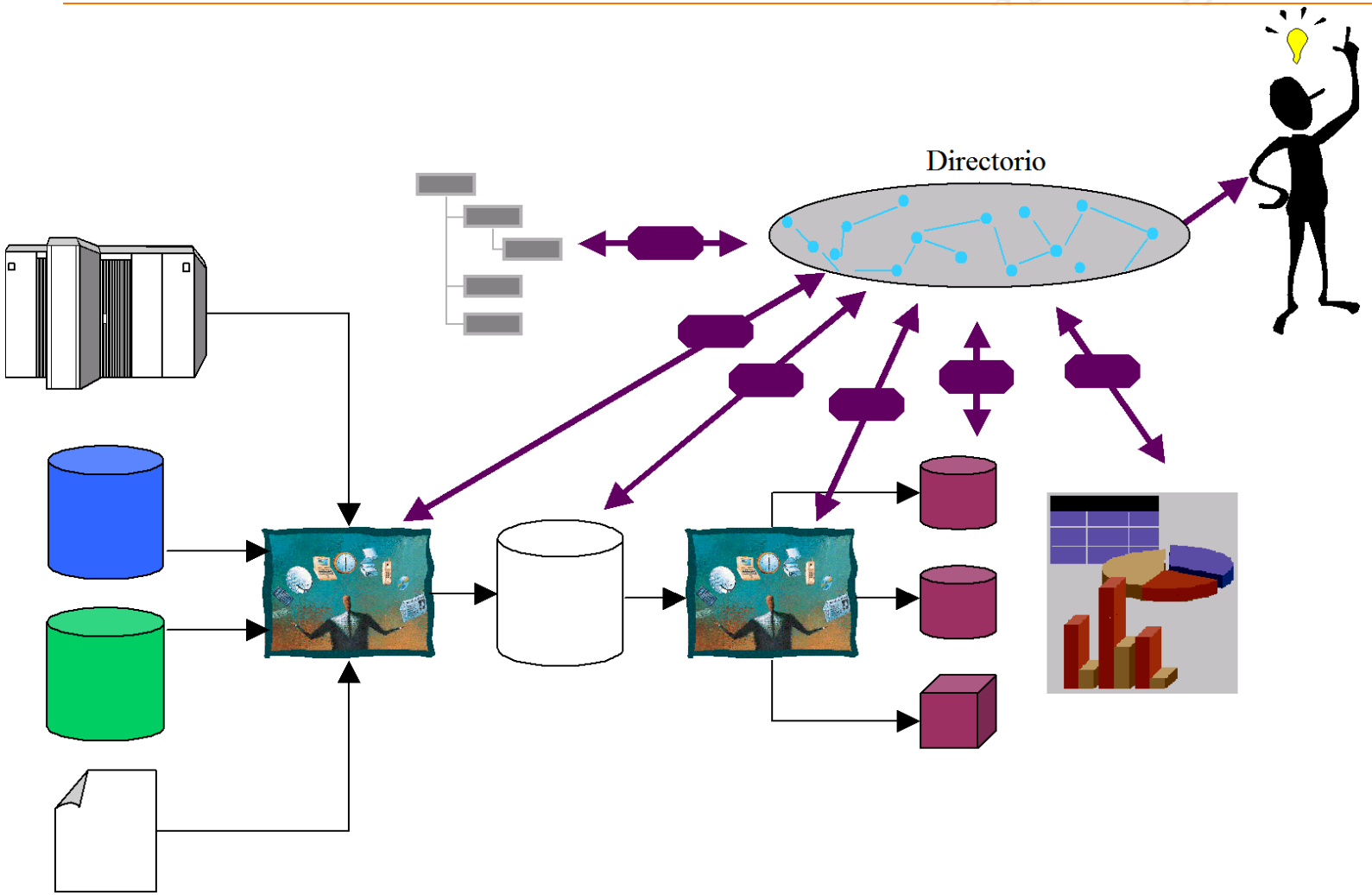
- Basados en estandars OMG CWMI

Metadatos

Metadatos - Diversidad de Componentes



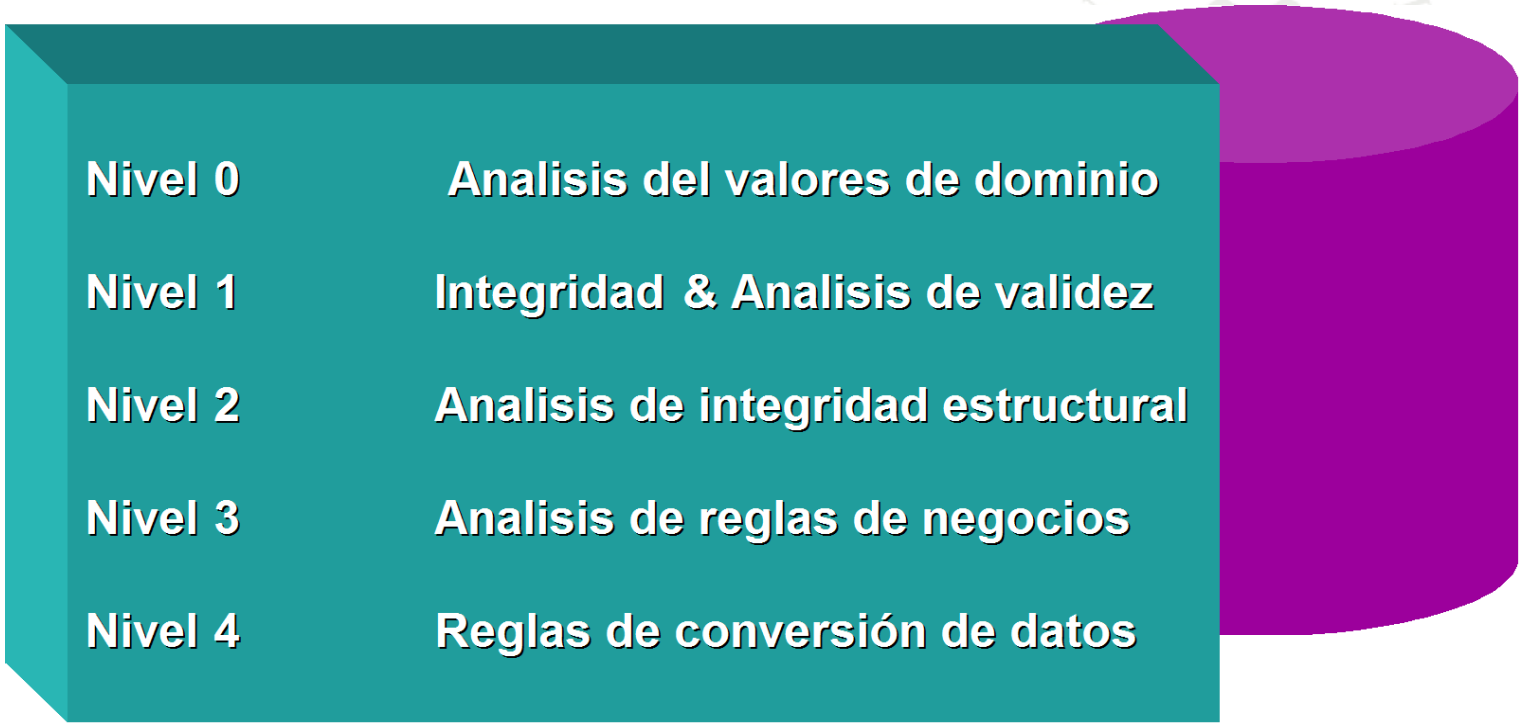
Metadatos - Arquitectura



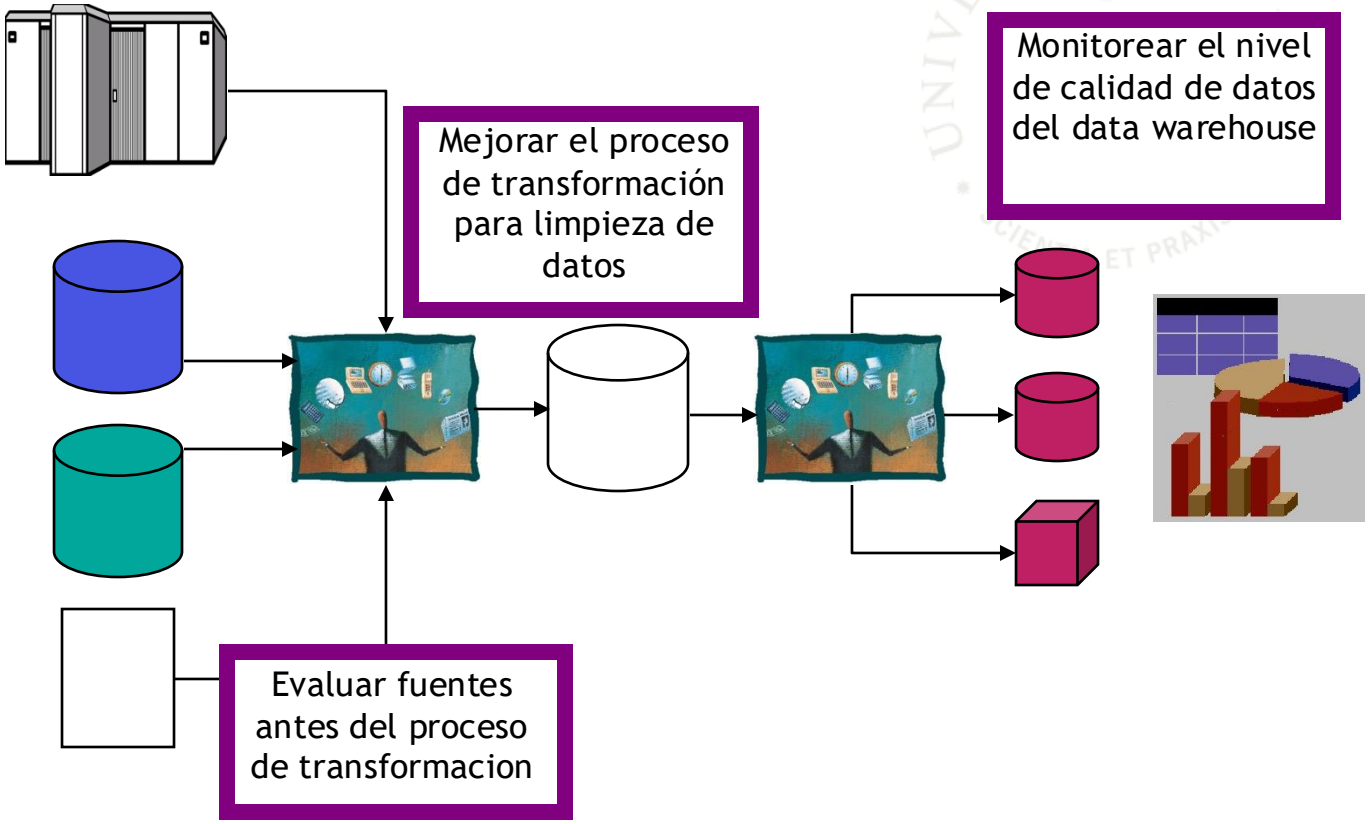
Metadatos - De dónde vienen los Datos?

Calidad de Datos

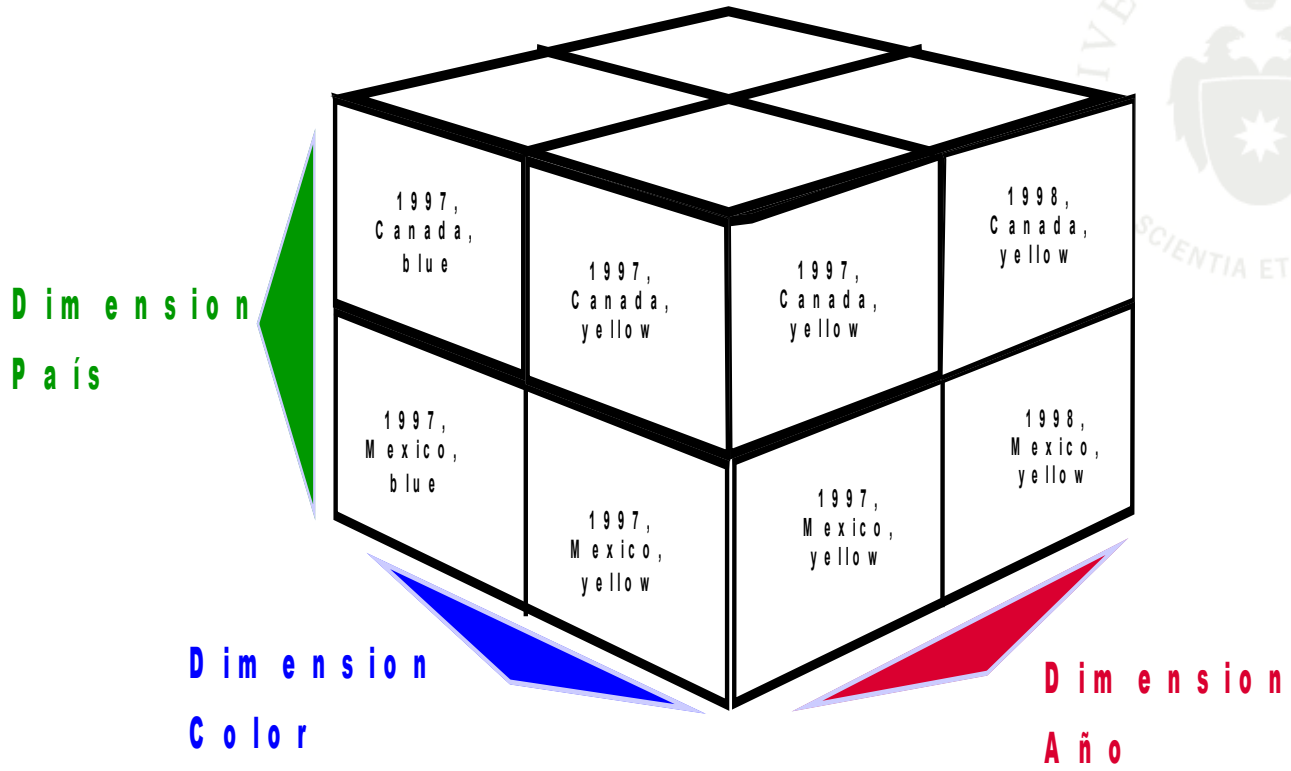
Calidad de Datos - Metodología



Calidad de Datos - Metodología



Cubos de Datos


Dimension
País
1997, Canada, blue
1997, Canada, yellow
1997, Canada, yellow
1998, Canada, yellow
1997, Mexico, blue
1997, Mexico, yellow
1997, Mexico, yellow
1998, Mexico, yellow
Dimension
Color
Dimension
Año

Cubos de Datos

- Los metadatos en data warehouse se representan a través de “cubo de datos” multidimensionales.
- Componentes:
  - Dimensiones
    - Ej: parte p, proveedor s, cliente c
  - Hechos o Fact (Medidas de interés)
    - Ej: ventas totales
  - Cubo de Datos 3-D
    - Ej: para cada celda (p,s,c),ventas totales de p que fueron compradas a s y fueron vendidas a c

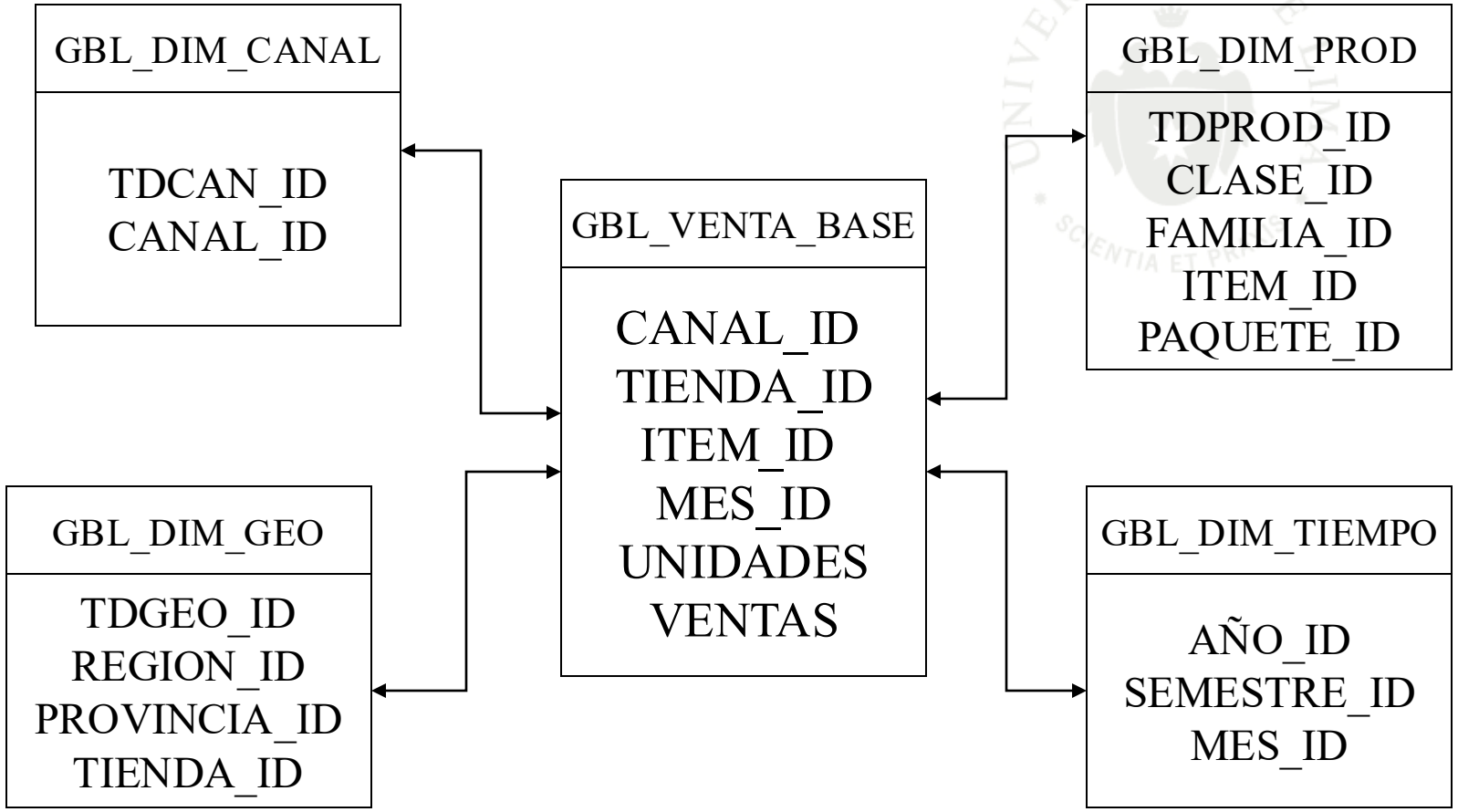
Definiciones

- Tablas Fact
  - Tambien conocidas como *Tablas de Hecho*, estas son las tablas centrales en un esquema estrella de un modelo DW. Los tablas fact representan el conocimiento del negocio y sus datos generalmente son numéricos y/o añadidos para ser analizados.
- Tablas Dimensión
  - También conocidas como *Tablas de la Referencia*, contienen los datos relativamente estáticos en el DW. Una dimensión es una estructura, integrada a menudo por unas o más jerarquías, que categoriza datos.
  - Jerarquías
    - Las jerarquías son las estructuras lógicas que utilizan niveles pedidos como los medios de ordenamiento de datos.

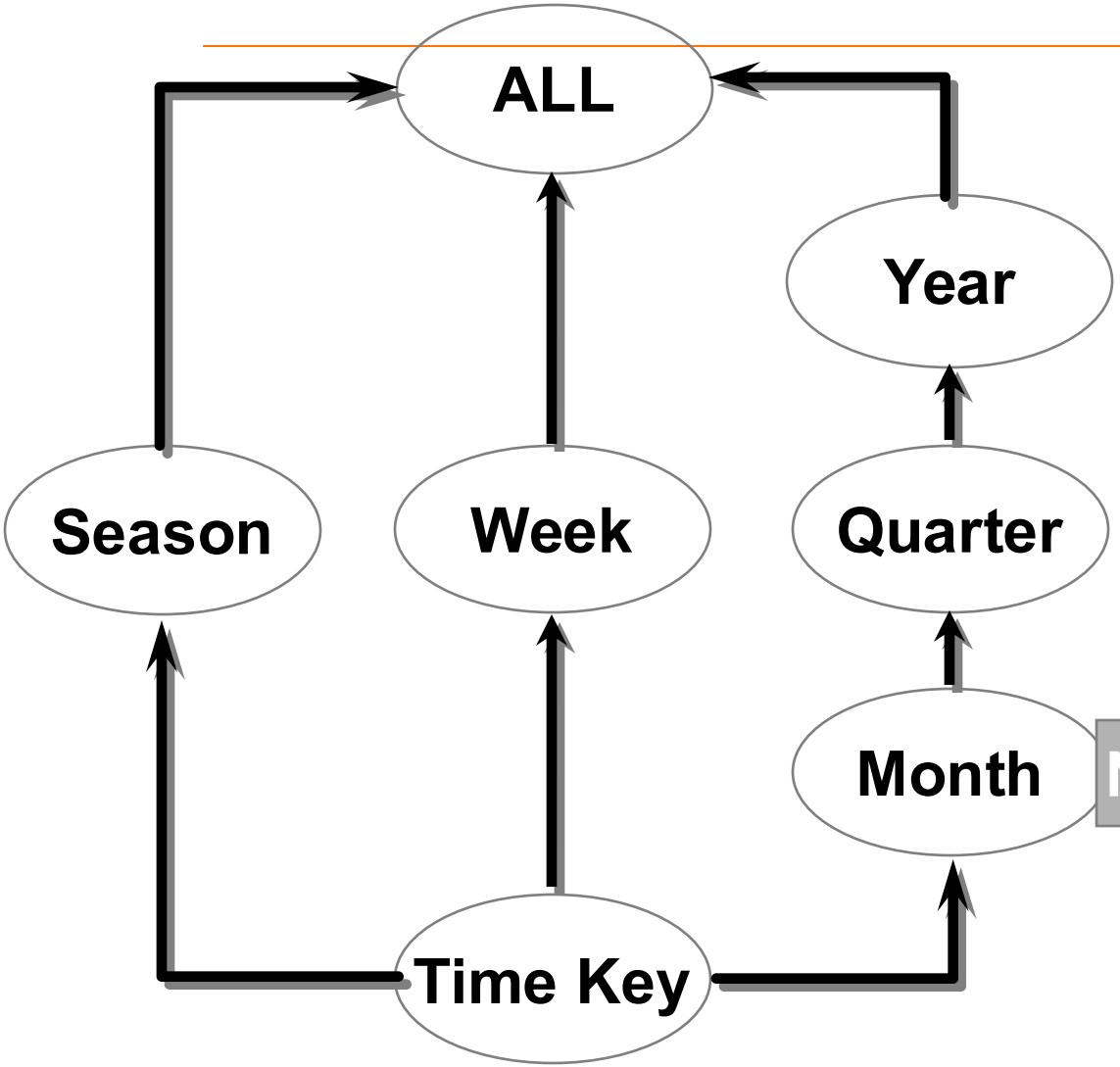
Índices

- Índices Bitmap
  - Ideales para búsquedas en gran cantidad datos y consultas ad hoc, reduce substancialmente del uso del espacio comparado con a otras técnicas de la indexación y su mantenimiento es muy eficiente durante consultas DML en paralelo y cargas de datos.
- Índices del B-tree
  - Un índice del b-tree ordena los datos como un árbol inverso. El nivel mas bajo del índice mantiene los valores reales de los datos y los punteros a las filas correspondientes. Este es utilizado mayormente para acceso de consultas típicas.

Esquema Básico



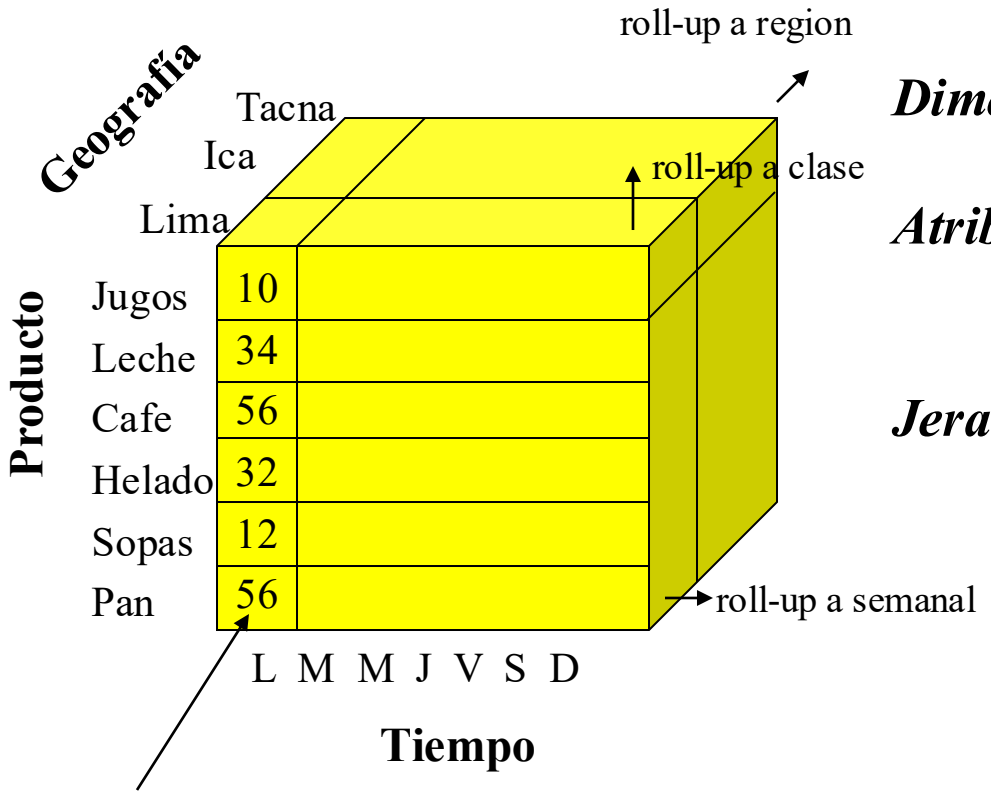
¿Qué es una Dimensión?



- Una dimensión describe las relaciones entre datos.
- Definido por la referenciación de tablas en la Base de Datos.
- Puede ser usada en una BD normalizada o desnormalizada.

Nombre del mes

Cubo de Datos



Dimensiones:

Tiempo, Producto, Geografía

Atributos:

Productos (upc, precio, …)

Geografia …

Jerarquias:

Producto → Clase → …

Dia → Semana → Semestre

Ciudad→ Region → Pais

56 unidades de la venta de pan en Lima el Lunes

Ejemplo de Dimensión

```
CREATE DIMENSION time_dim
    LEVEL time_key          IS time.time_key
    LEVEL month             IS time.month
    LEVEL quarter           IS time.quarter
    LEVEL year              IS time.year
    LEVEL week              IS time.week
    LEVEL fiscal_qtr        IS time.fiscal_qtr

    HIERARCHY calendar_rollup     (
              time_key    CHILD OF
              month       CHILD OF
              quarter     CHILD OF  year )
    HIERARCHY fiscal_rollup       (
              time_key    CHILD OF
              week    )

    ATTRIBUTE time_key DETERMINES
      (day_number_in_month, day_number_in_year)
    ATTRIBUTE  week     DETERMINES week_number_of_year
    ATTRIBUTE  month    DETERMINES full_month_name;
```

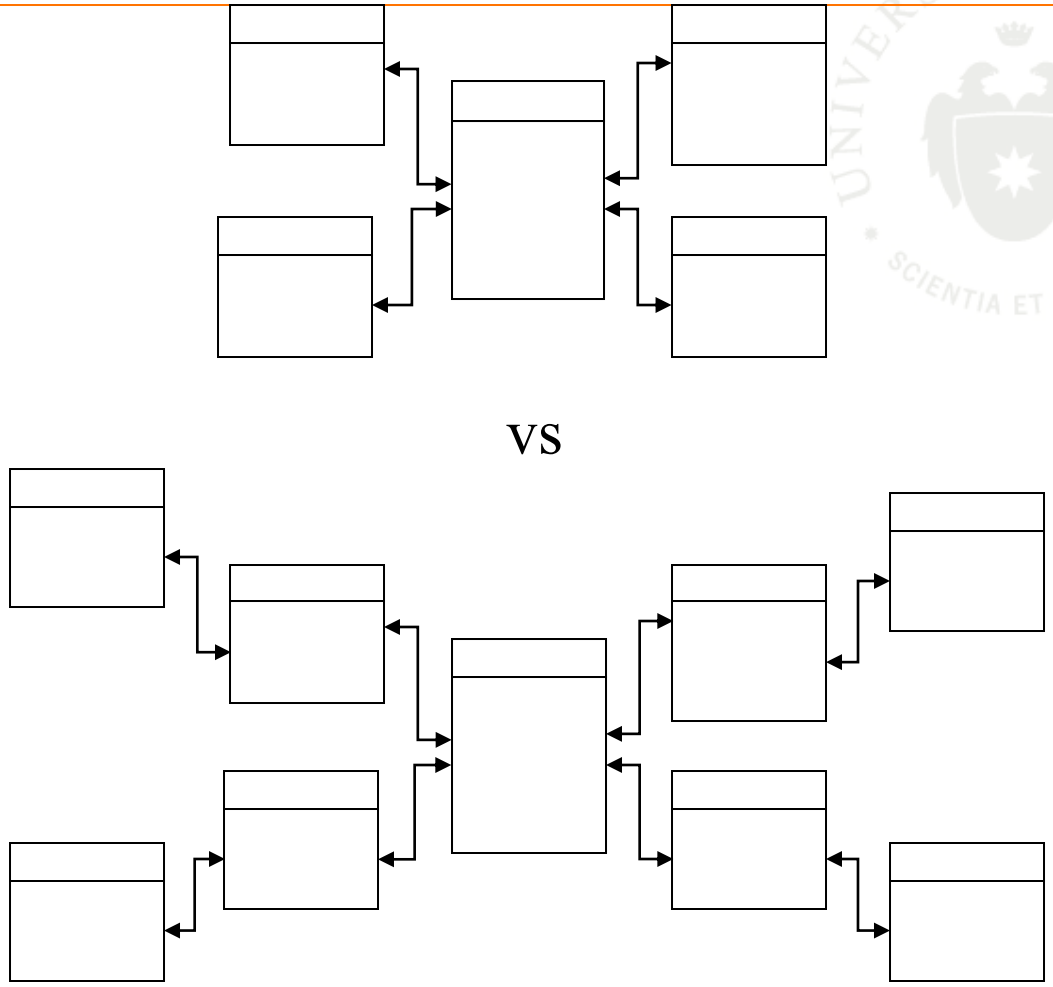
nombre

Define el nombre de cada nivel en la dimension y la columnas de la tabla

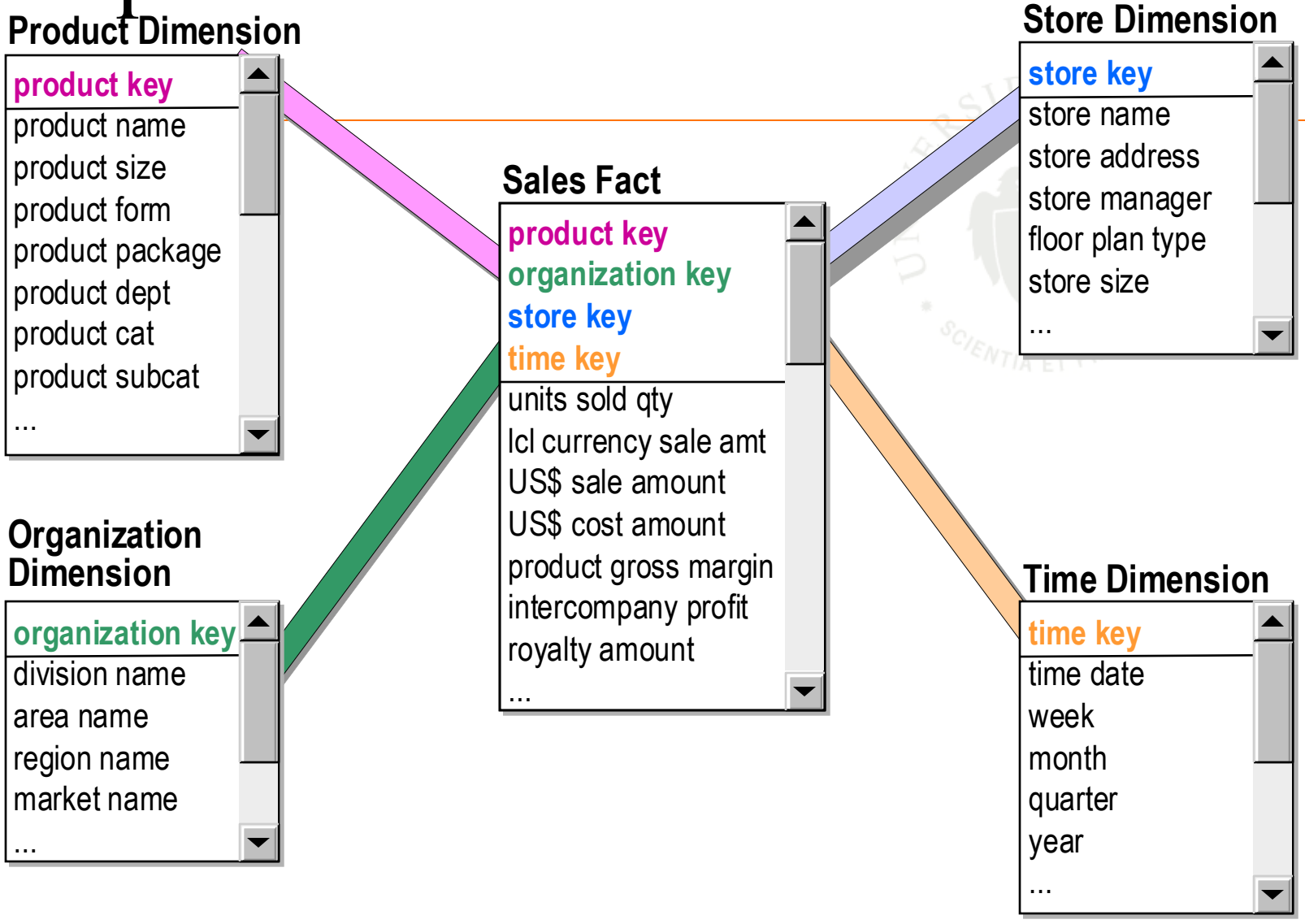
Describe la jerarquia usando niveles de nombres

atributos que son unicos por nivel

Desnormalización vs Normalización

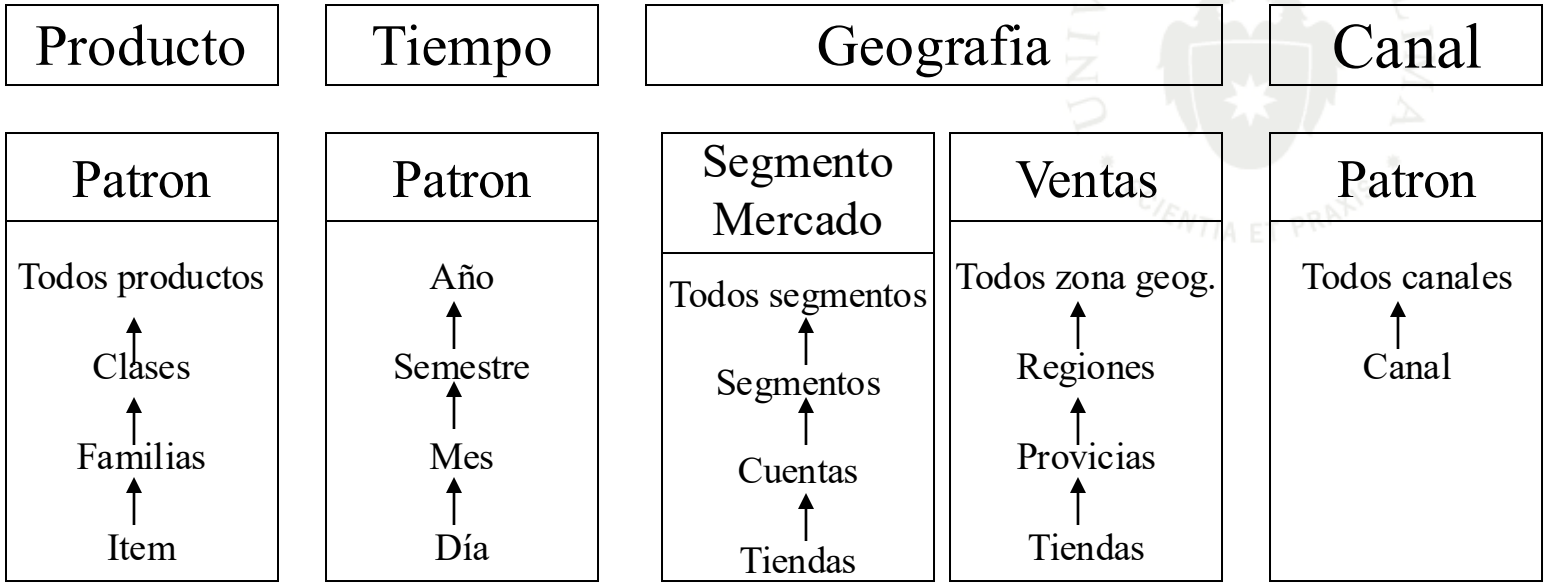


Esquema Estrella



Aplicación de un Modelo de Datos

Dimensiones, Niveles, Jerarquías y Atributos



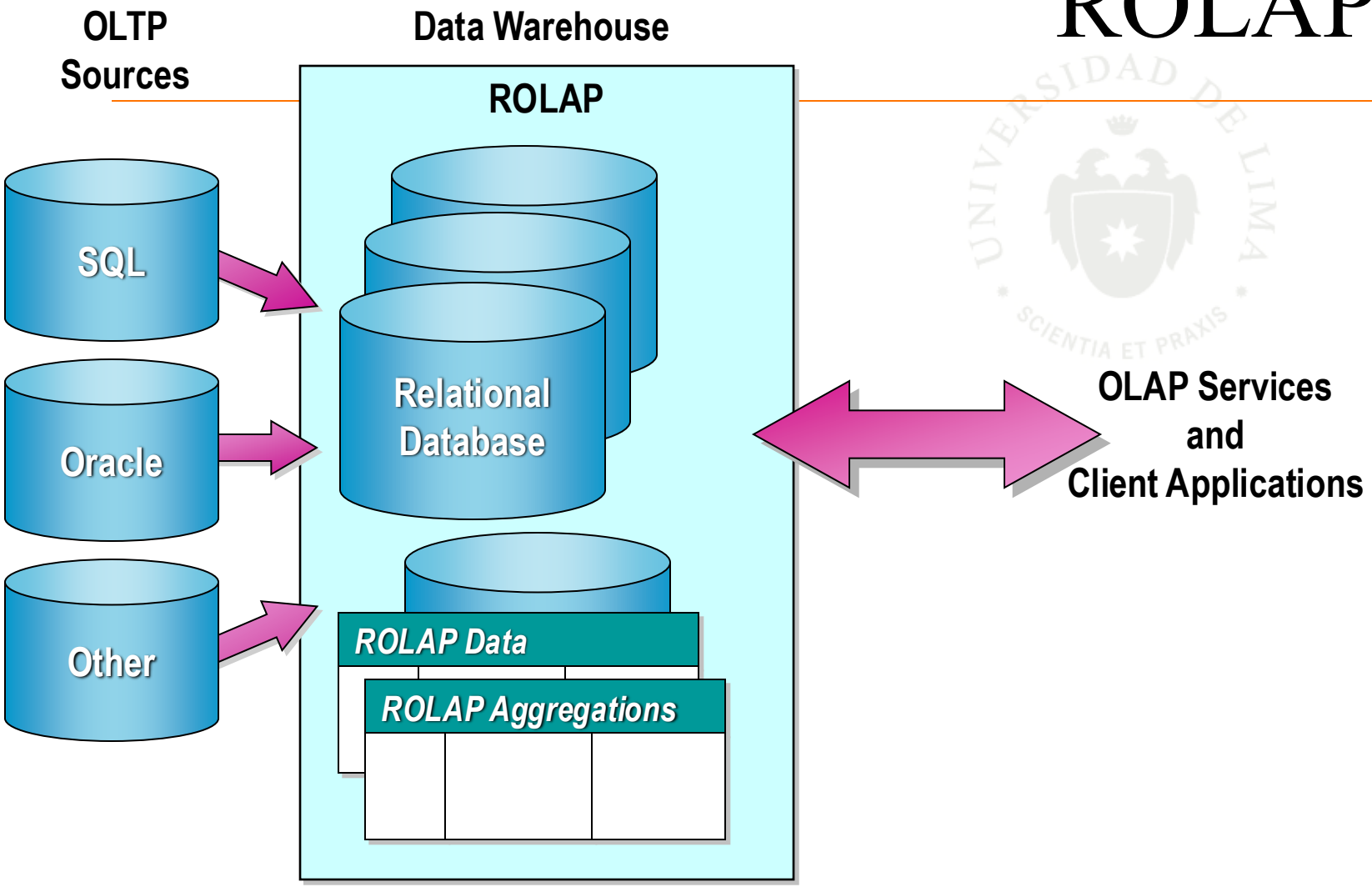
Variables

Dimensiones (producto, tiempo,geografía, canal)

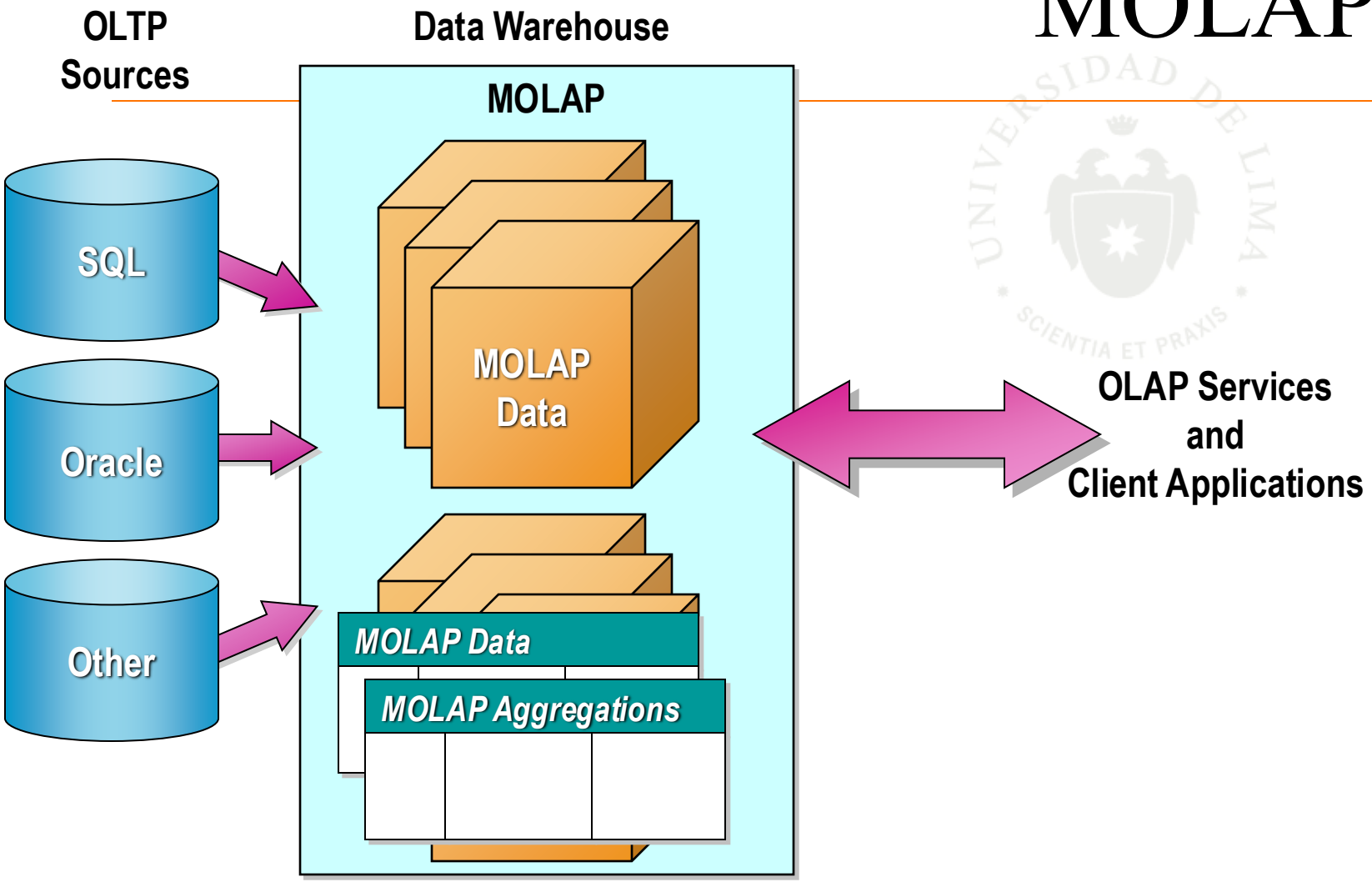
Unidades (patron{todos, algunos, uno}, segmentos{mercado, zonas})

Tipos de Cubos

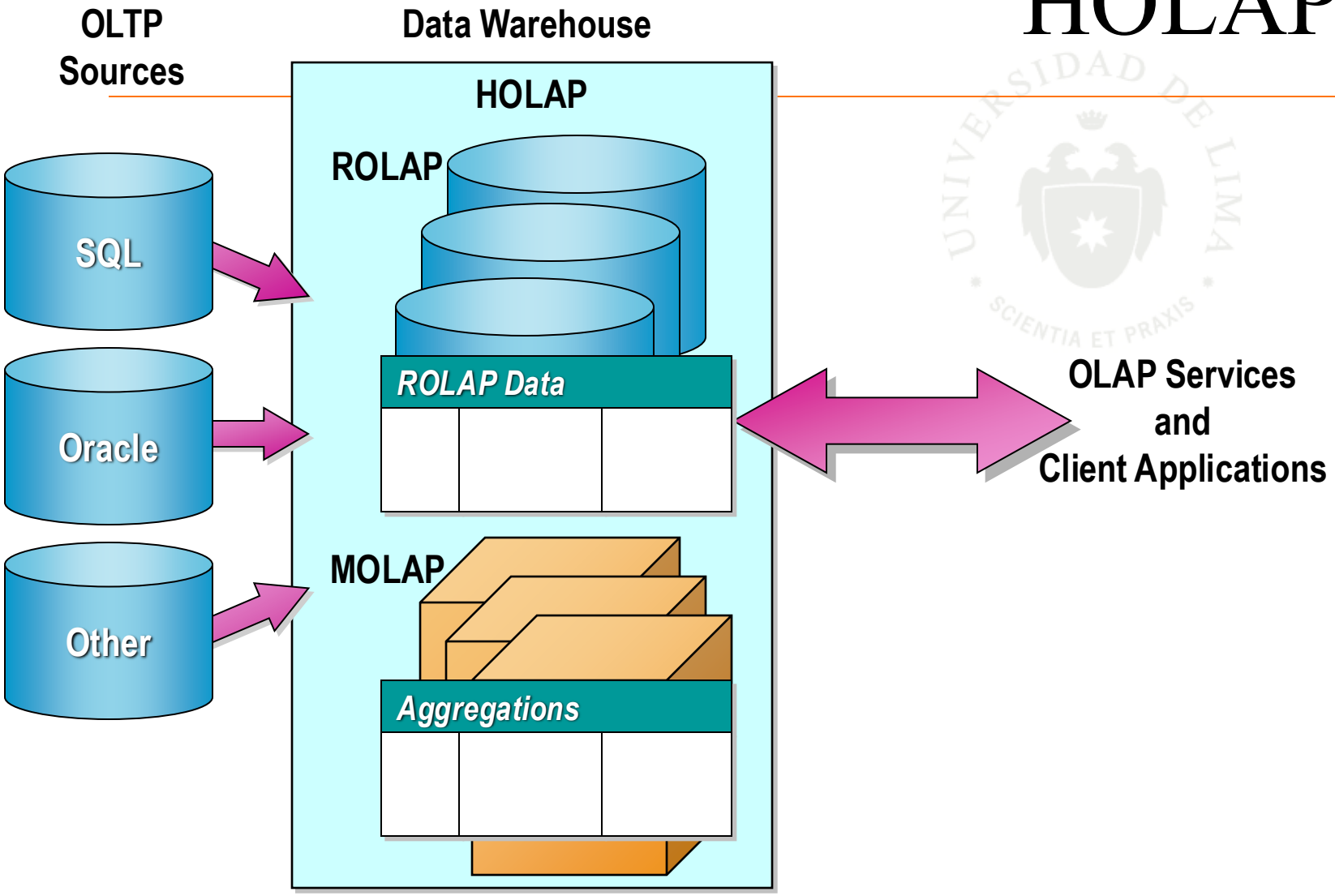
ROLAP


OLTP
Sources
Data Warehouse
ROLAP
SQL
Oracle
Other
Relational
Database
ROLAP Data
ROLAP Aggregations
OLAP Services
and
Client Applications

MOLAP


OLTP
Sources
Data Warehouse
MOLAP
SQL
Oracle
Other
MOLAP
Data
MOLAP Data
MOLAP Aggregations
OLAP Services
and
Client Applications

HOLAP


OLTP
Sources
Data Warehouse
HOLAP
ROLAP
SQL
Oracle
Other
ROLAP Data
MOLAP
Aggregations
OLAP Services
and
Client Applications

Explotación de Datos

Explotación de Datos

✓Query Ad hoc – Consultas Ad hoc
- Proceso Iterativo de Preguntas y Respuestas
- Herramienta que asiste a experto a encontrar información valiosa

✓OLAP - Análisis Multidimensional
- Preguntas Multidimensionales/Respuestas Multidimensionales
- Consultas Complejas - Herramienta que ayuda al usuario experto

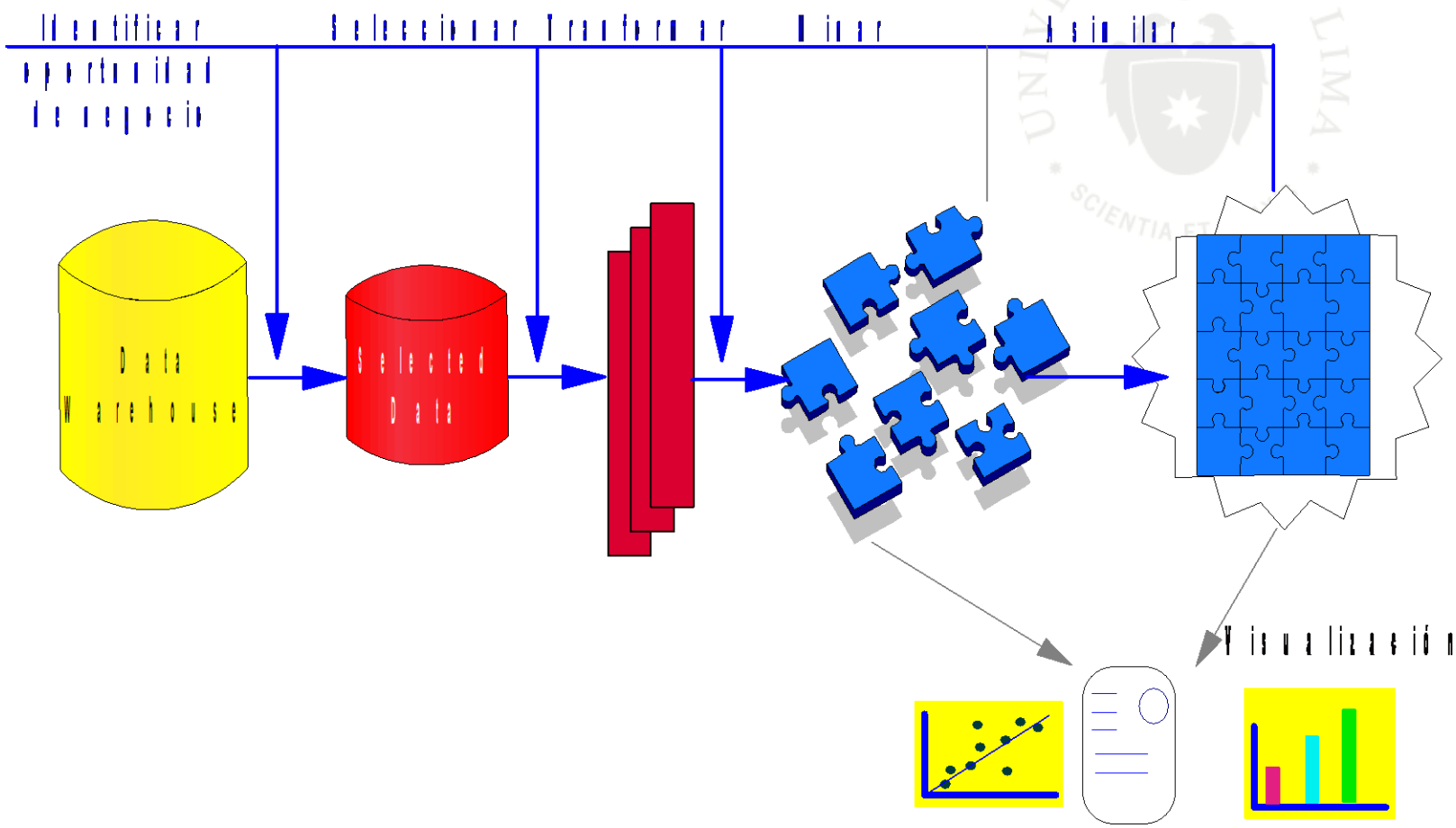
✓Data Mining - Minería de Datos
- Herramienta que automáticamente busca tendencias y patrones
- Herramienta que aprende y desarrolla experiencia
- Puede ser la base para análisis detallado

Data Mining

- Es el proceso que permite al usuario conocer la escencia y las relaciones entre sus datos.

- Data mining descubre patrones y tendencias en el contenido de la información almacenada en el warehouse a través de las diferentes técnicas que utiliza.

Proceso de Data Mining

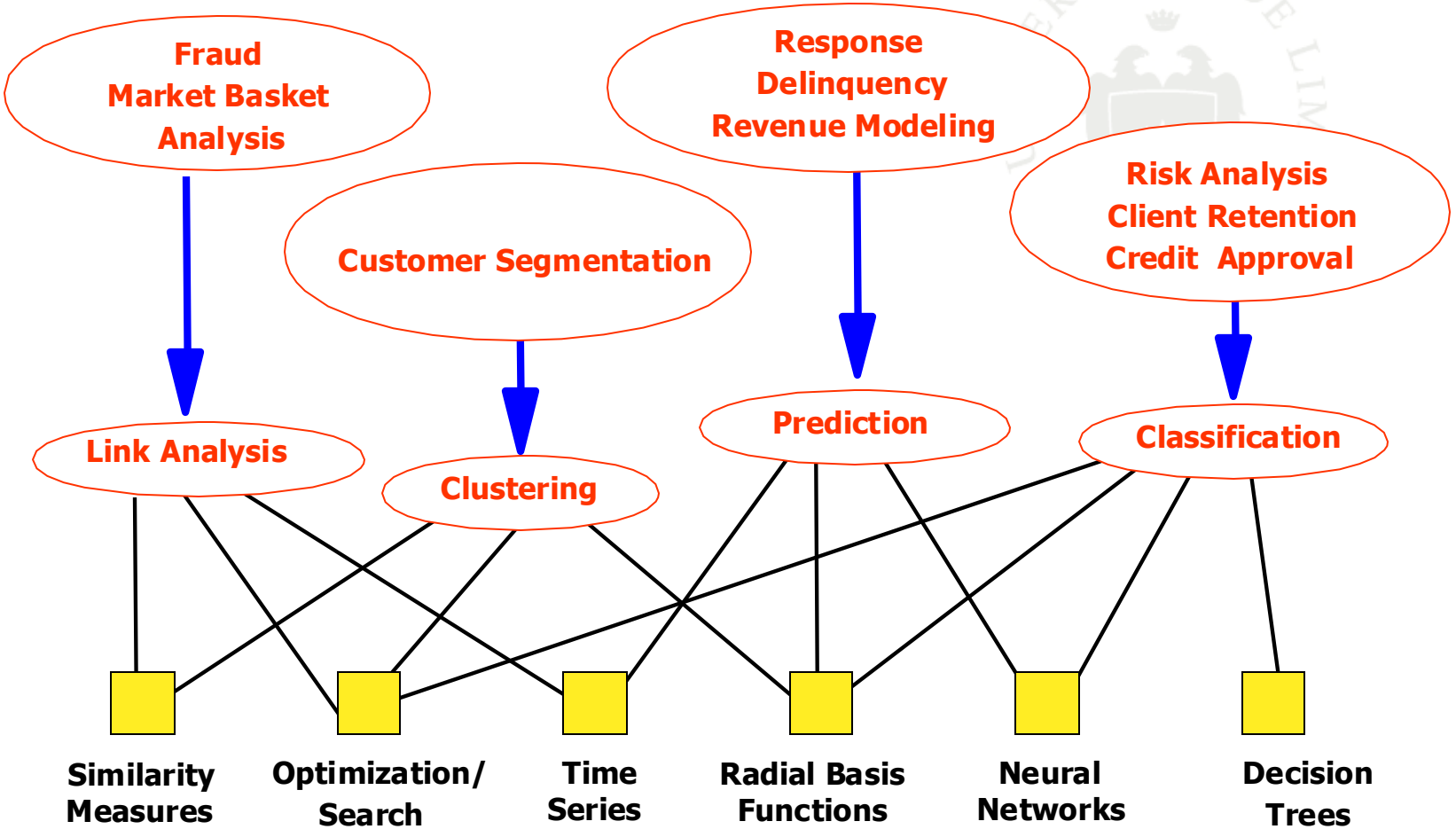


Tipos de Minería

- ✓ **Minería de Descubrimiento** - Encontrar patrones en Datos que pueden ser usados para guiar decisiones.
  - Minimiza dirección de la minería
  - Analiza todo el almacén de datos
  - Mineria de "fácil". Algunas veces la más valiosa.

- ✓ **Minería Predictiva** - Usar resultados conocidos para crear modelos que puedan predecir valores futuros.
  - Extensiva dirección del minador
  - Aprende del muestras para aplicar la población
  - Minería "Dura". Un experto usuario es requerido.

Técnicas Data Mining



Beneficios

ROI y Beneficios de la Solución

- ✓Las organizaciones que han implementado y utilizado aplicaciones analíticas han logrado retornos en un rango desde 17% a 2000% con una media de ROI de 112%.
- ✓Cerca de la mitad (49%) de estos han tenido un periodo de recuperación de menos de un año.
- ✓Los resultados en el área financiera utilizando sistemas analíticos tienen una media de ROI de 139%.
- ✓El costo de generación de reportes se reduciría en aprox. 60%.
- ✓El tiempo total de generación de reportes se reduce de 10 a 1

Source: IDC

Beneficios

- ✓ **Desde perspectiva IT**
  - Arquitectura flexible de data warehousing que me permita cambiar cuando lo necesite
  - Presupuestos limitados
  - Procesos que automáticamente actualicen los datos del data warehousing
  - Desgaste del area de IT por requerimientos de usuarios

- ✓ **Desde perpectiva de usuarios finales**
  - Tener una vista global de los datos de la organización, no importando de donde provenga la información
  - Tener una herramienta fácil de usar
  - Tener la capacidad de generar reportes de manera independiente
  - Acceso desde internet o intranet