INTRODUCCION

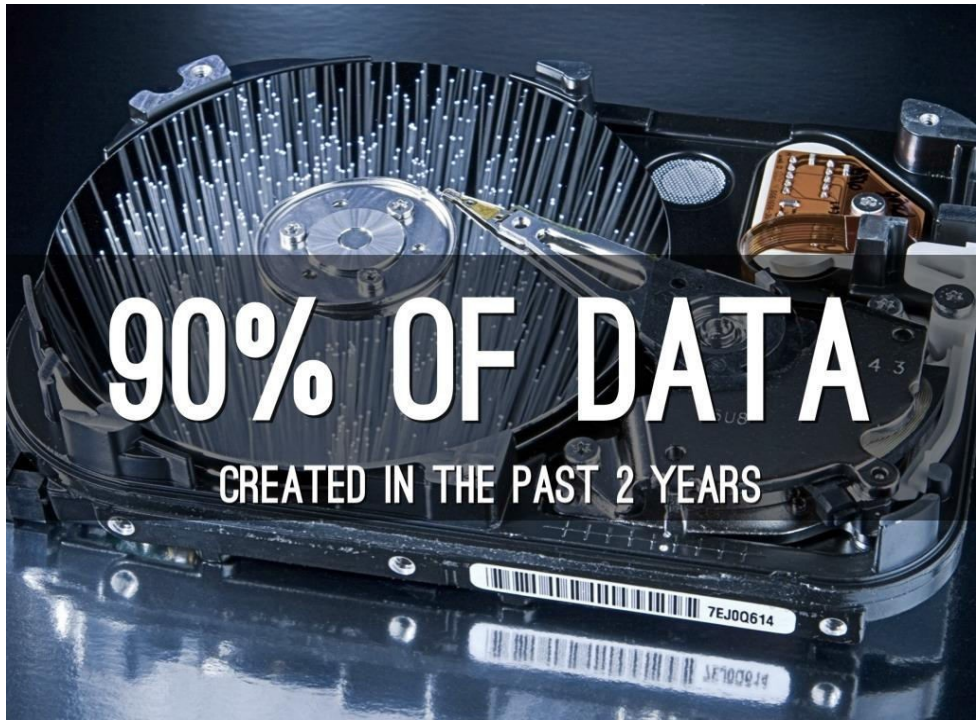
Oswaldo.solarte@correounivalle.edu.co

AGENDA

1. Introducción Big Data
2. Business Intelligence & Data warehouses
3. Data Mining

Explosión de los datos

EXPLOSIÓN DE LOS DATOS



Wall Mart Supermarket

Un millón de clientes cada hora

200 millones de transacciones por semana

2.5 petabytes de datos

EXPLOSIÓN DE LOS DATOS

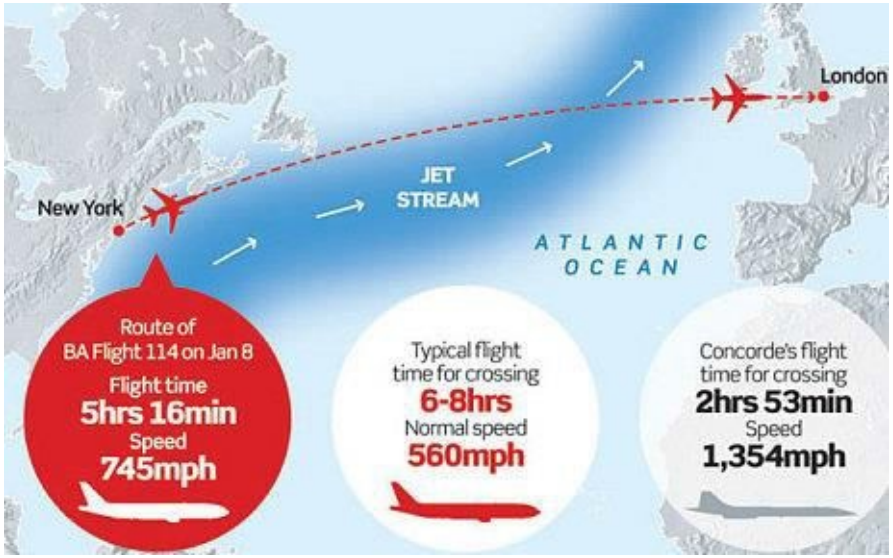
Quién genera información?

Qué es Big Data?

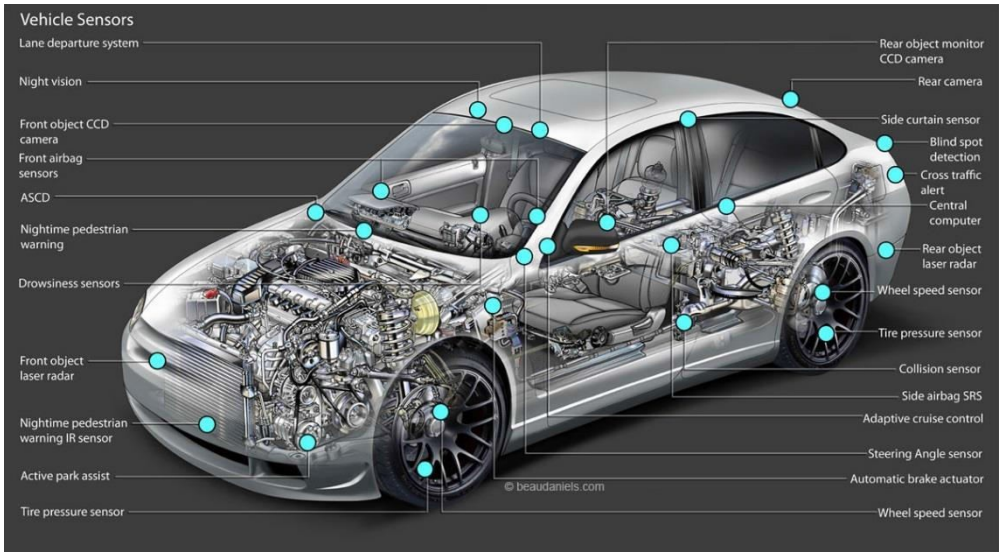
- Diferentes grados de **complejidad, ambigüedad** en los datos
- No pueden ser procesados utilizando **tecnologías tradicionales**

VOLUMEN

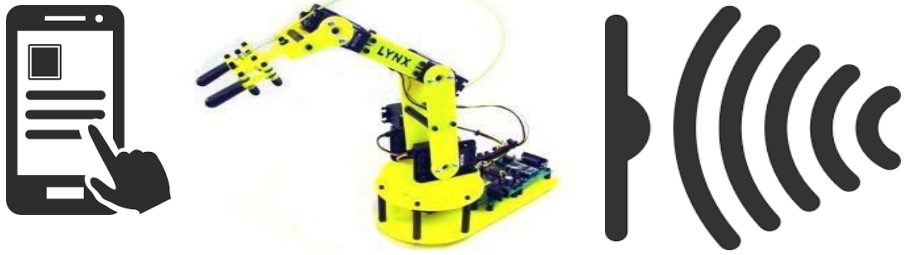
- Google procesa 20 PB de información por día
- 2,3 Trillones de GB se crean cada día
- Muchas empresas en USA tienen aprox. 100 TB de inf.

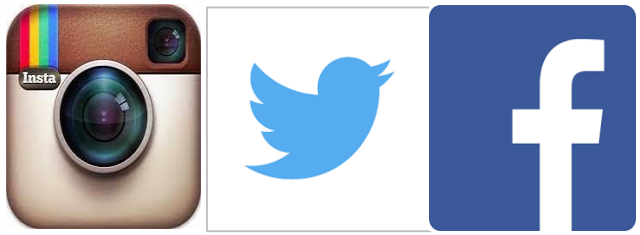


650 GB



VOLUMEN

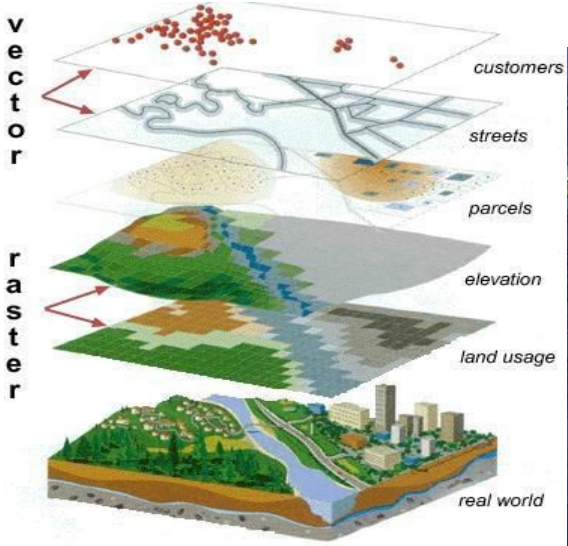




Machine data

Social media data

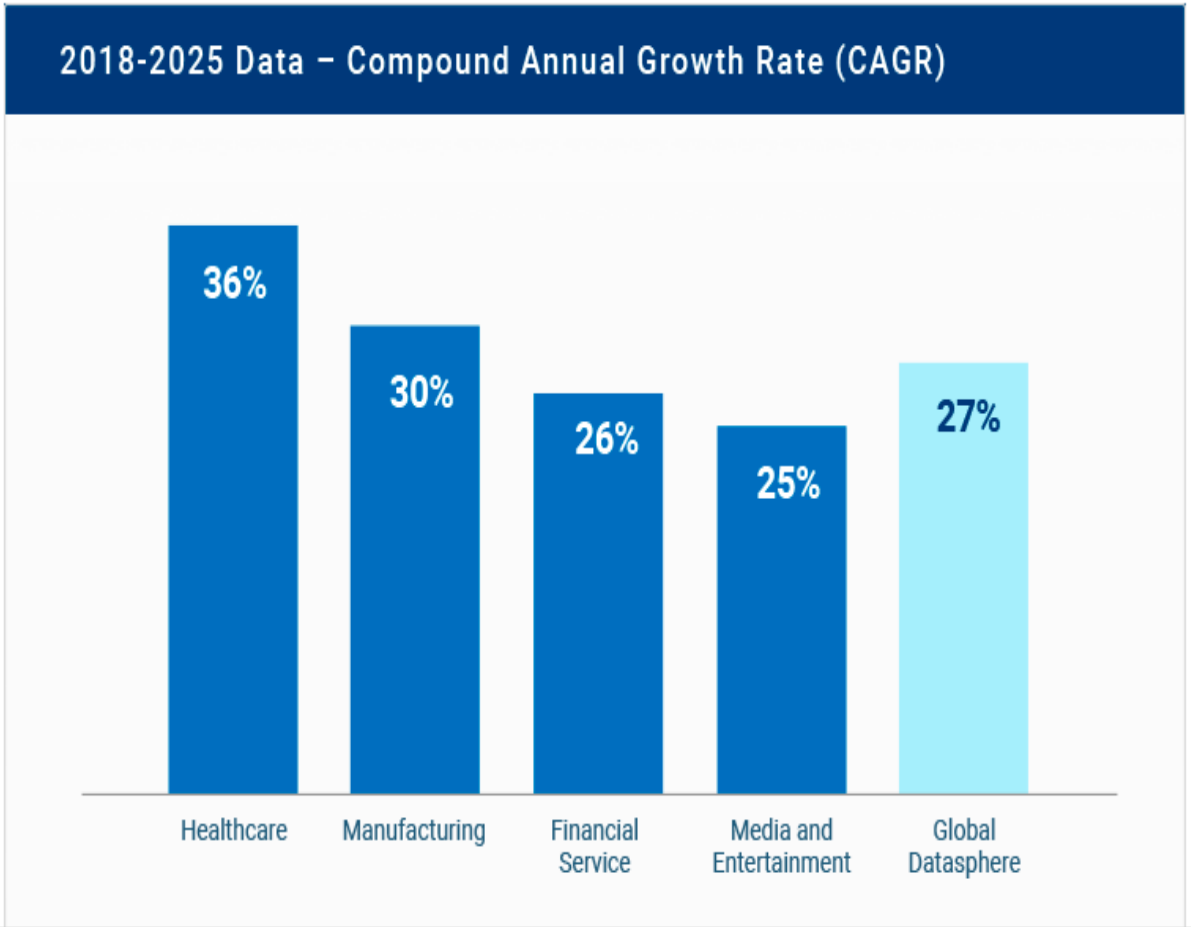
Geo spatial Data

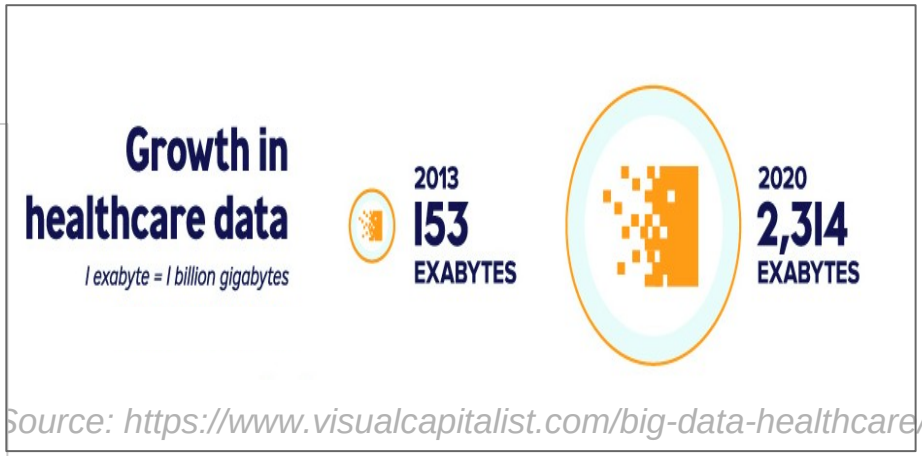


Email data

VOLUMEN

The health care process generates a huge amount of data. Hospitals produce **50 petabytes** of data **per year[1]**.





Source: https://www.visualcapitalist.com/big-data-healthcare/

VELOCIDAD

Cada minuto…..

- Google receives over 4 million search queries
- Facebook users share nearly 2.5 million pieces of content.
- Twitter users tweet nearly 300,000 times.
- Instagram users post nearly 220,000 new photos.

VELOCIDAD

Cada minuto…..

YouTube users upload 72 hours of new video content.

Apple users download nearly 50,000 apps.

Email users send over 200 million messages. Amazon generates over $80,000 in online sales.

Velocidad

Year	Global Internet Traffic
1992	100 GB per day
1997	100 GB per hour
2002	100 GBps
2007	2000 GBps
2014	16,144 GBps
2019	51,794GBps

Source: Cisco VNI, 2015

VARIEDAD

Unstructured Data

Internal Sources
- Email
- Call Center Transcripts
- Forums, Blogs

External Sources
- Social Networks
- Forums, Blogs
- Videos

Structured Data

Internal Sources
- CRM
- Point of Sale
- Service Tickets

External Sources
- Industry Research Data
- Financial Market Data

Text Mining & Human Analysis

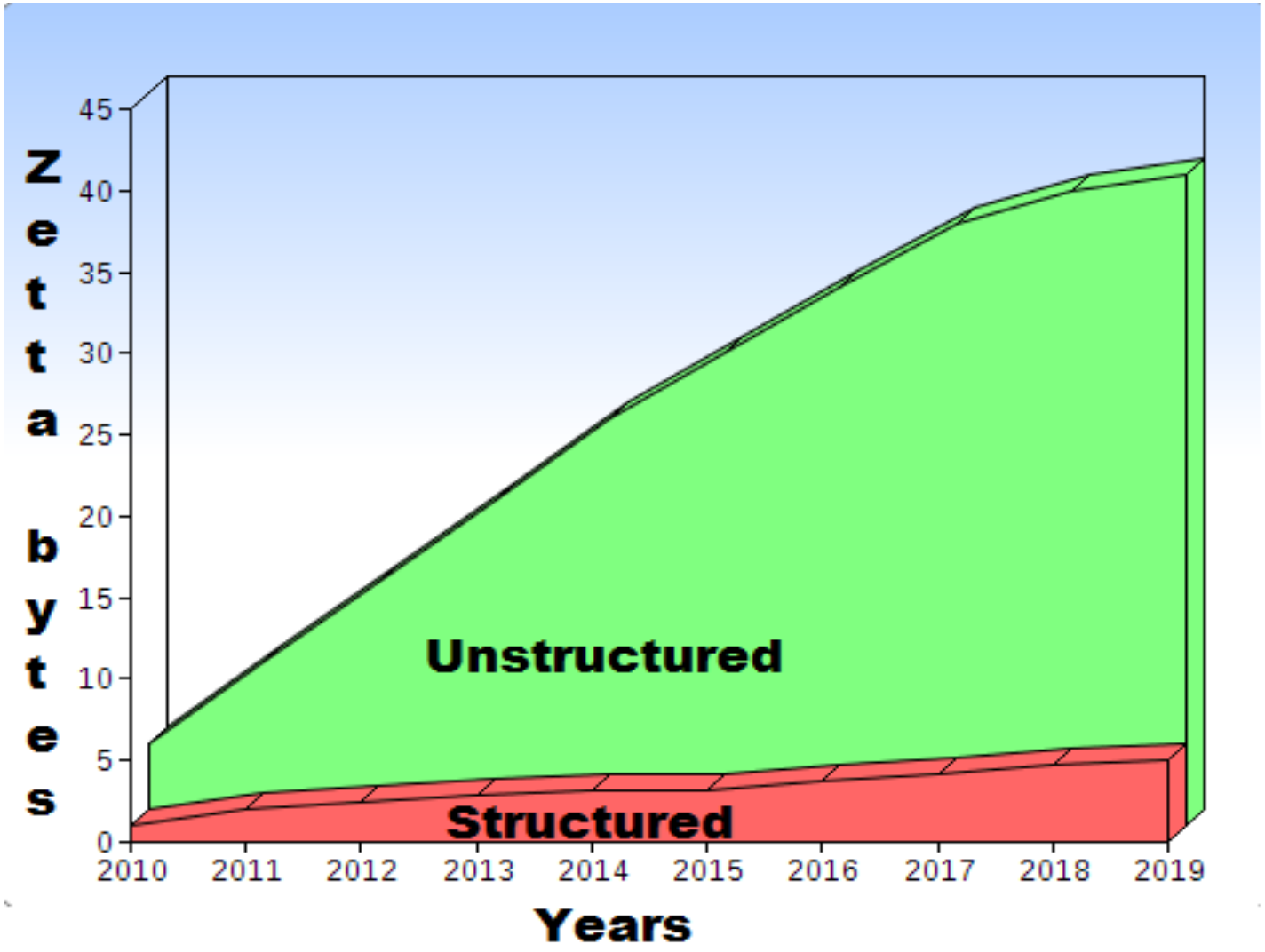
Dashboard & Alerts

Reports

Automatic Feeds to Other Systems

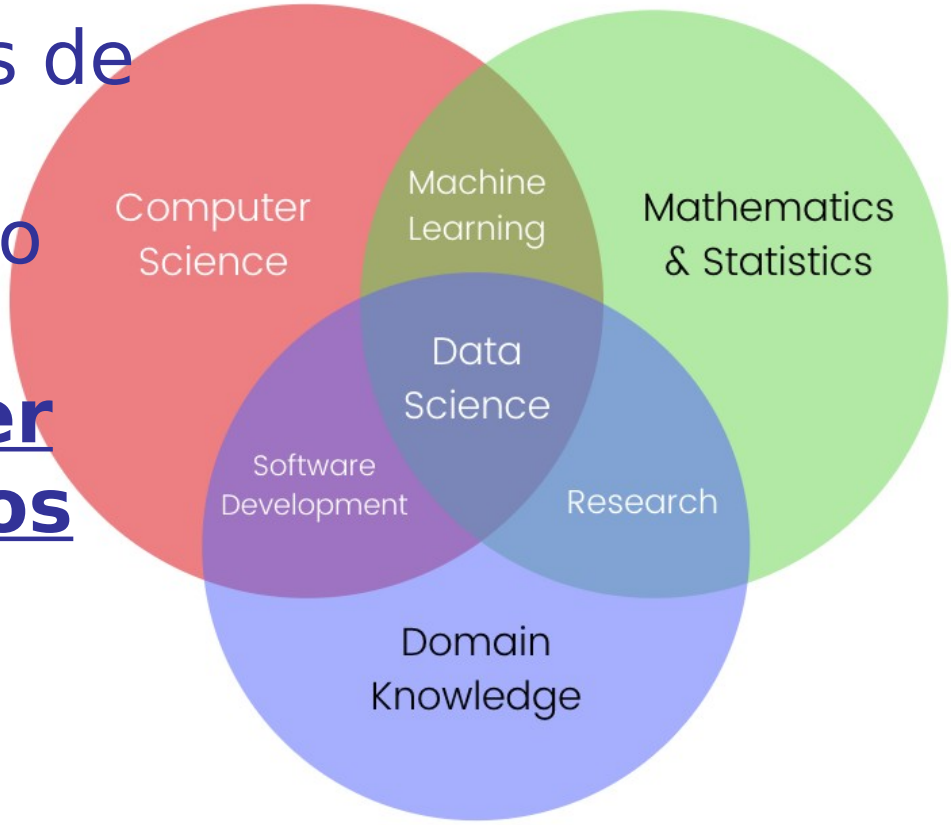

WHAT WE KNOW...
THE REST...
BIG DATA

NATURALEZA DE LOS DATOS



Data Science

Combina utiliza conceptos de estadística, matemática y programación, en conjunto con herramientas tecnológicas, para **extraer información de los datos** para tomar mejores decisiones.



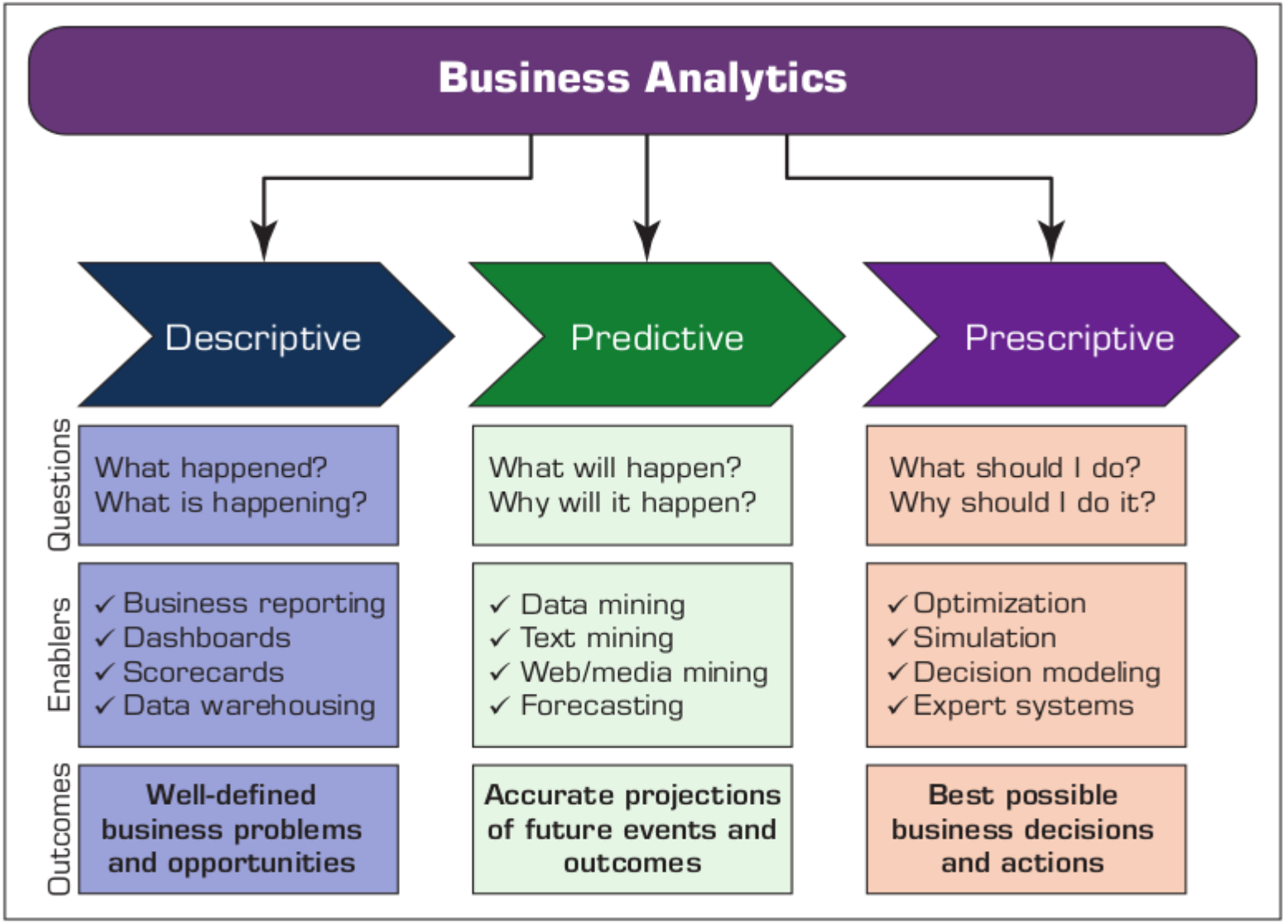
Data Vs Information

Table 2: Data and Information

Data	Information
Artificial signals intended to convey meaning	**Data with meaning for intended use**
Easily captured by machines	Requires human intervention to define meaning (relevance, purpose)
Easily manipulated	Requires consensus of meaning for action
Easily transferred	Can be replicated, but often hard to transfer accurately
In form suitable for quantification	Can be stored, but often difficult to recall economically
Easily stored	

http://hmi.ucsd.edu/pdf/HMI_2010_ConsumerReport_Jan_2011.pdf

Business Analytics

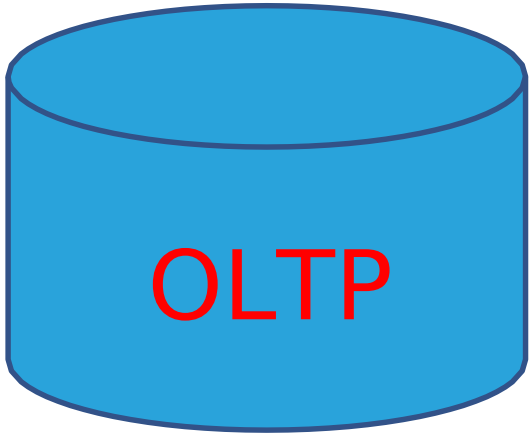


Source-Book:Analytics, DataSience, & AI. Ramesh et al, 2021

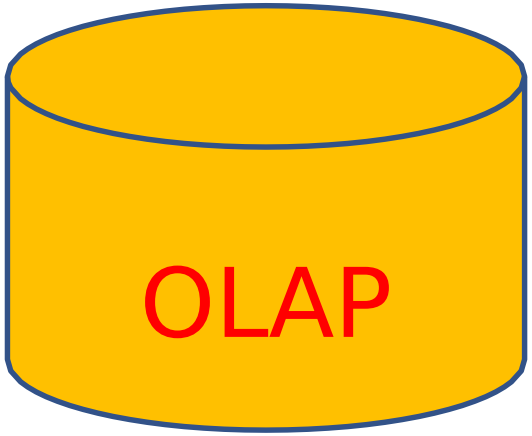
Business Intelligence (*BI*)

Pregúntame qué pasa

Pregúntame qué pasará



Vs

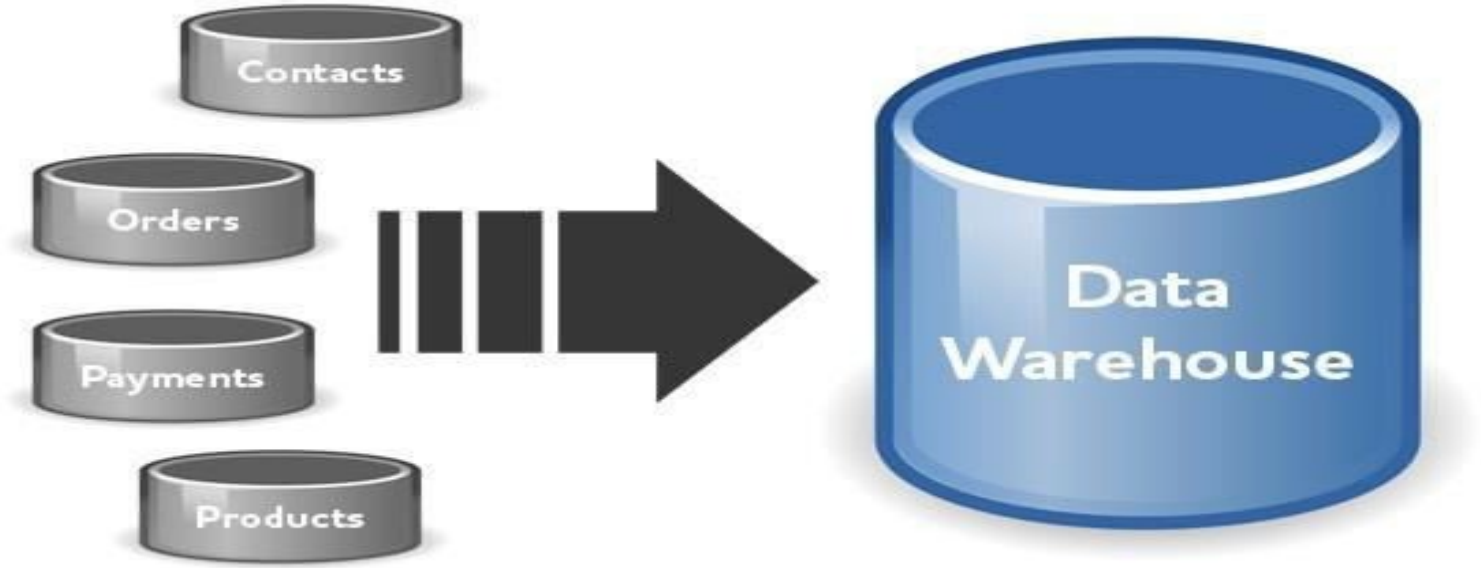


Online Transaction Processing

Online Analytical Processing

BUSINESS INTELLIGENCE (*BI*)

- Integración y tratamiento de los **datos** para **convertirlos** en **información** que permita apoyar a los tomadores de decisiones en la organización



Tomar decisiones a partir de los datos …


Business
Intelligence


BIG
DATA
ANALYTICS
STORAGE
SOFTWARE
TECHNOLOGIES
INTERNET
INFORMATION
COMPLEX
DATABASES
RESEARCH
MANAGEMENT
SIZE
NEEDED
PETABYTES
SAN
PARALLEL
SOCIAL
LARGE
EVERY
SETS
EXAMPLES
ABILITY
NETWORKS
SEARCH
RECORDS
INCLUDE
SYSTEMS
SHARED
TIME
DISK
HUNDREDS
BIOLOGICAL
MASSIVELY
GARTNER
USING
WORKING
USE
ONE
BUSINESS
MOVING
TOLERABLE
LOGS
PROCESSING
DISTRIBUTED
CAPTURE
MANAGE
GROW
STORE
CASE
LARGER
CURRENT
APPLIED
ELAPSED
ARCHIVES
SENSOR
TARGET
DIFFICULTY
CITATION
COST
CONTINUES
TERABYTES
PRACTITIONERS
CAPACITY
TENS
PERFORMANCE
ALSO
GENOMICS
TOOLS
SET
WITHIN
PROCESS
DEFINITION
MAY

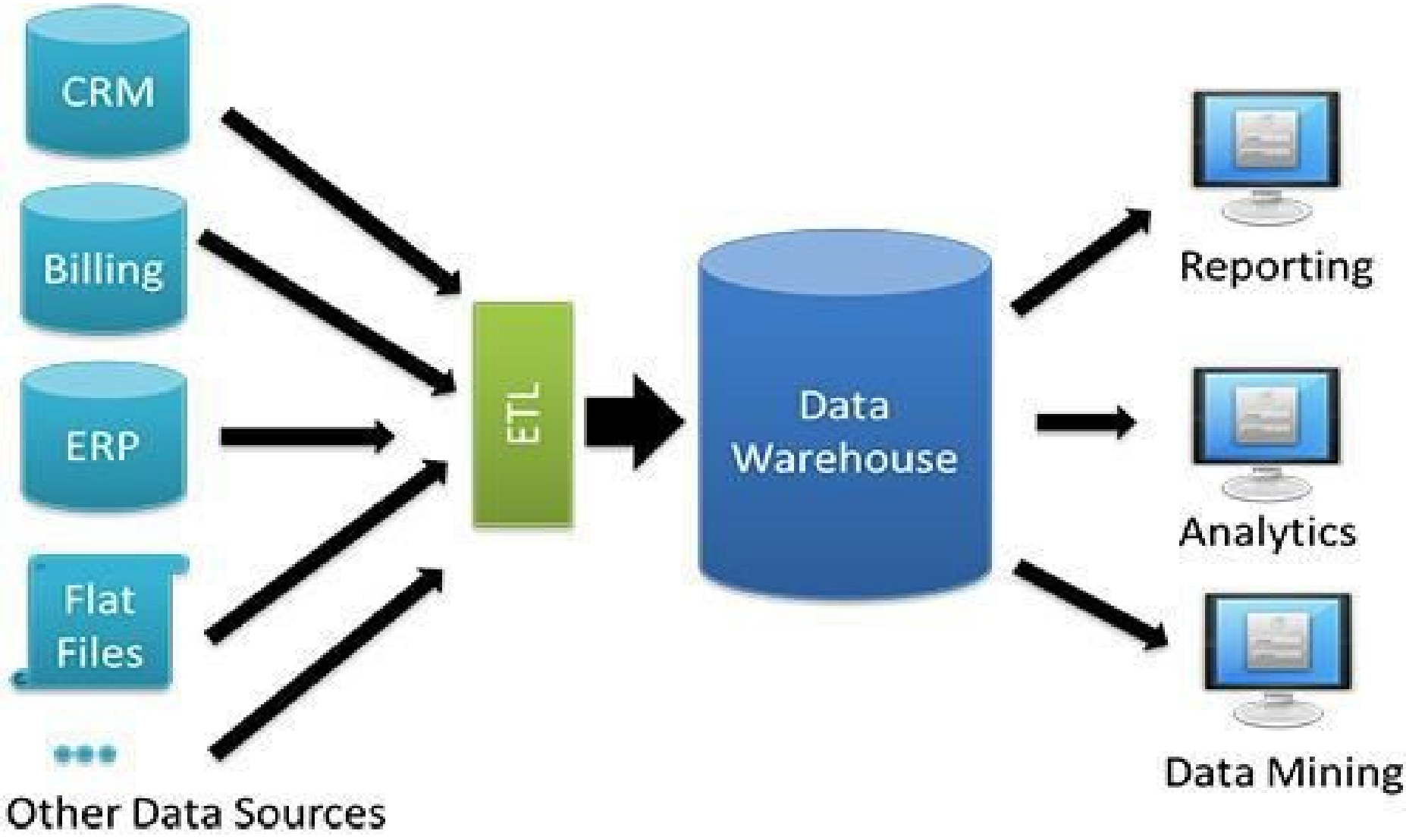
BUSINESS INTELLIGENCE

- **Ganar ventaja del negocio a partir de los datos**

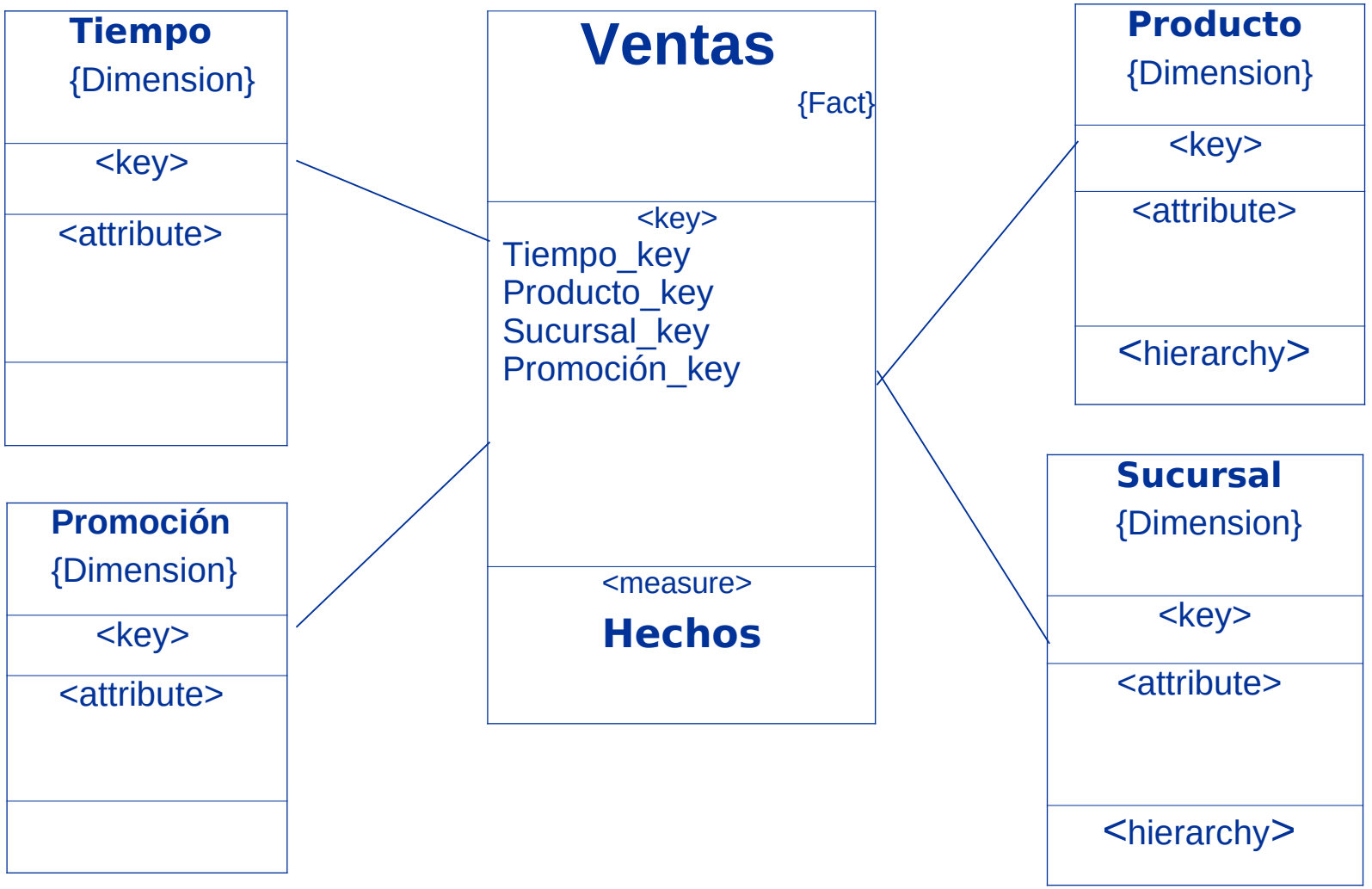
  - Cuáles son los mejores clientes?
  - Cuáles clientes se van a retirar?
  - Qué factores afectan las ventas?
  - Qué ventajas puede el negocio ofrecer a los clientes?
  - Dónde gano o pierdo?

BODEGA DE DATOS

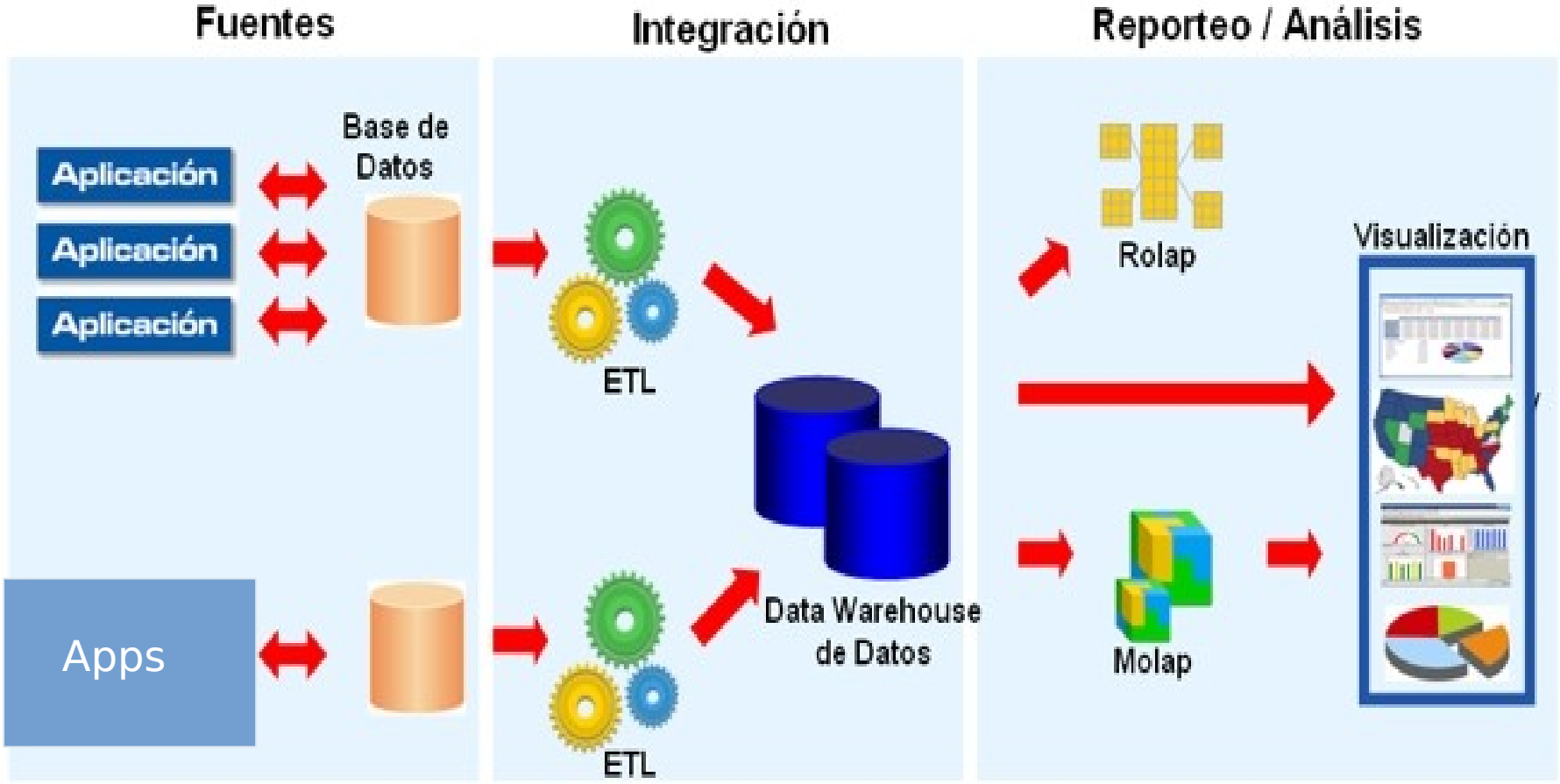
DATA WAREHOUSE



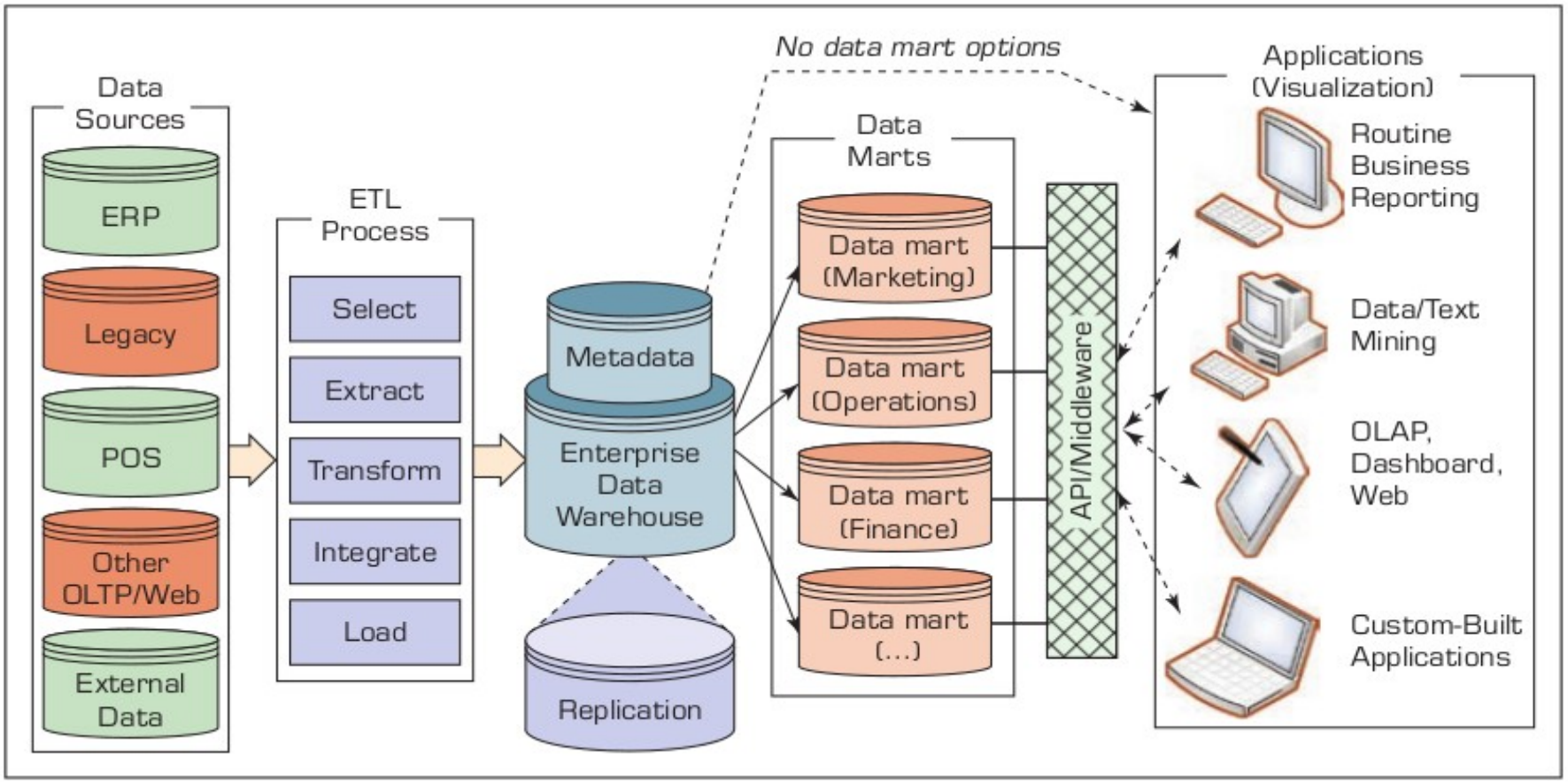
MODELO DIMENSIONAL



DATA WAREHOUSE

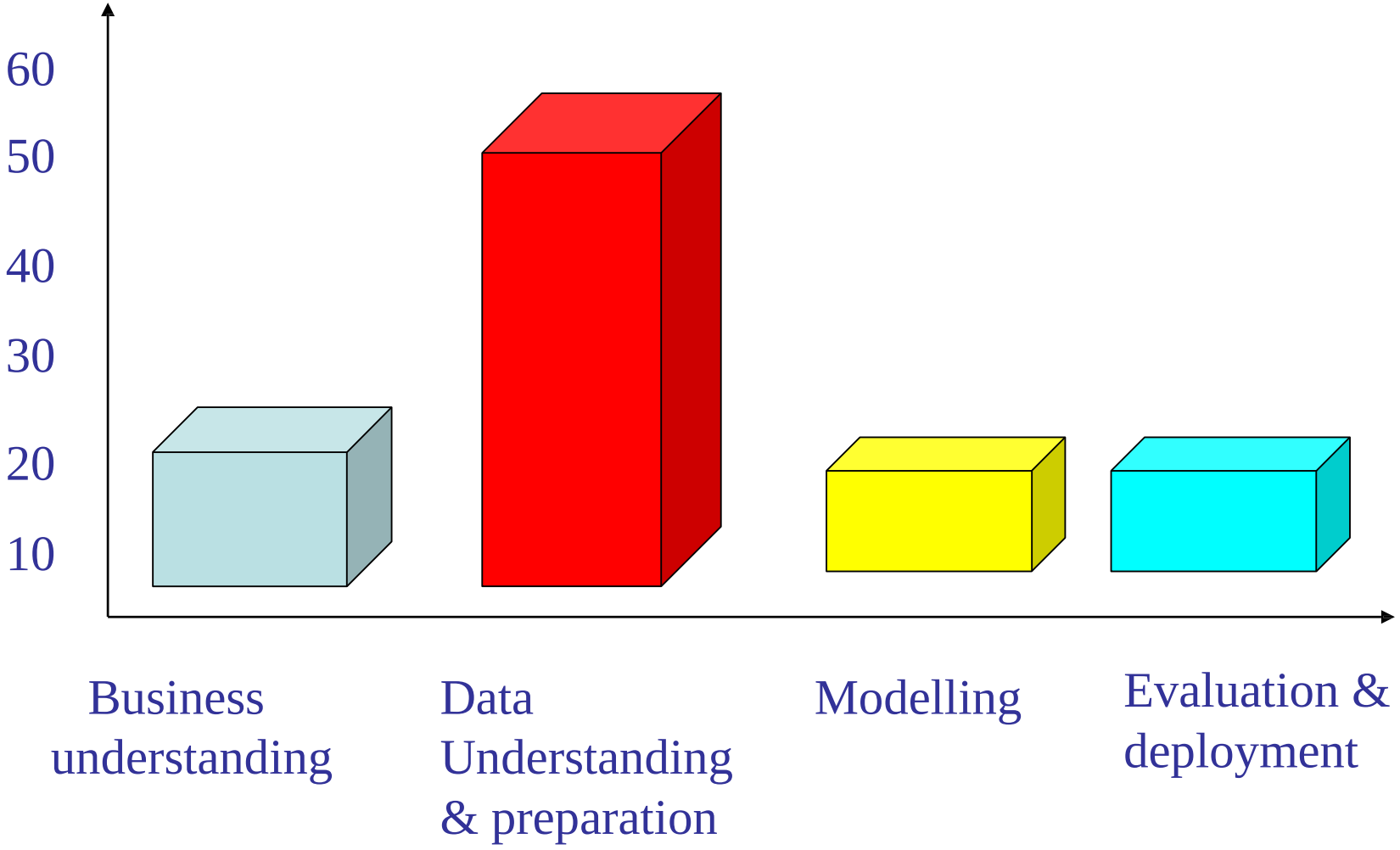


DATA WAREHOUSE

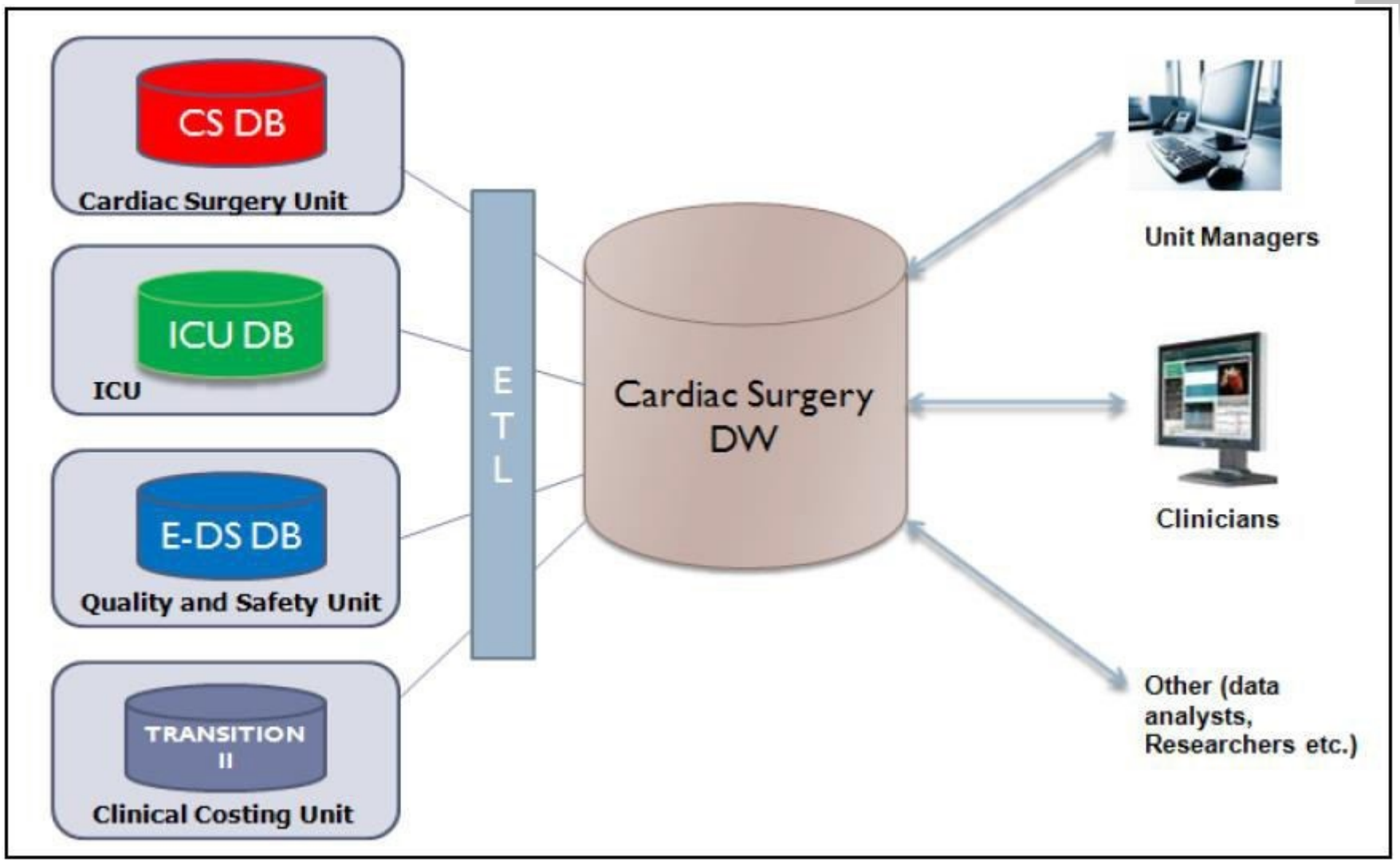


Source-Book:Analytics, DataSience, & AI. Ramesh et al, 2021

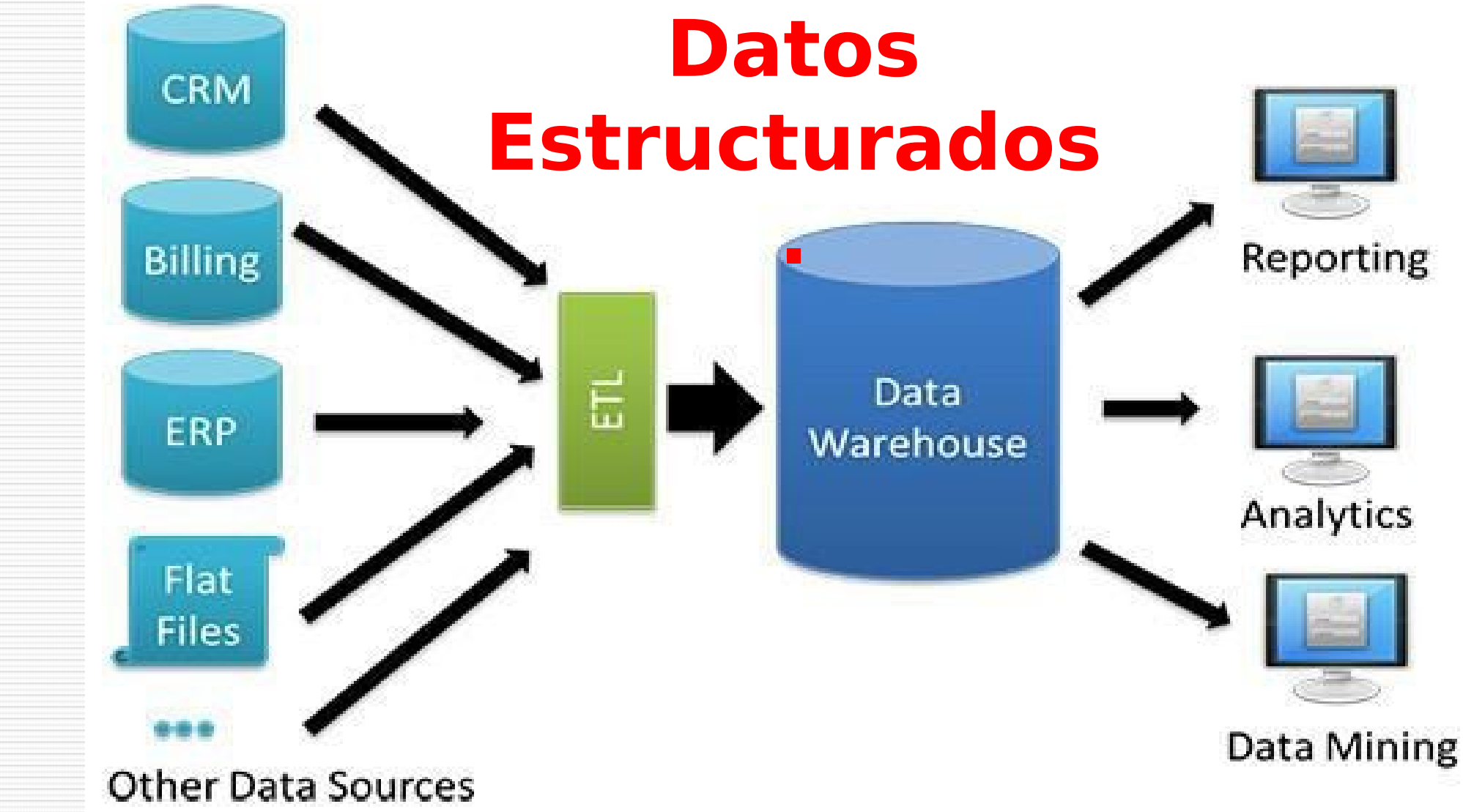
ESFUERZO ASOCIADO



CLINICAL DATA WAREHOUSE



DATA WAREHOUSE



NATURALEZA DE LOS DATOS

CLIENTE

NOMBRE	CEDULA	CUENTA	CIUDAD
CANO	7.245.310	C-101	CALI
PEREZ	1.352.851	C-121	PASTO
TORO	9.874.115	C-203	BOGOTA
LOPEZ	9.765.398	C-302	BUGA
SERNA	2.458.698	C-109	TADO
VEGA	4.111.119	C-230	LIMA
CANO	7.245.310	C-209	CALI
PEREZ	1.352.851	C-209	PASTO

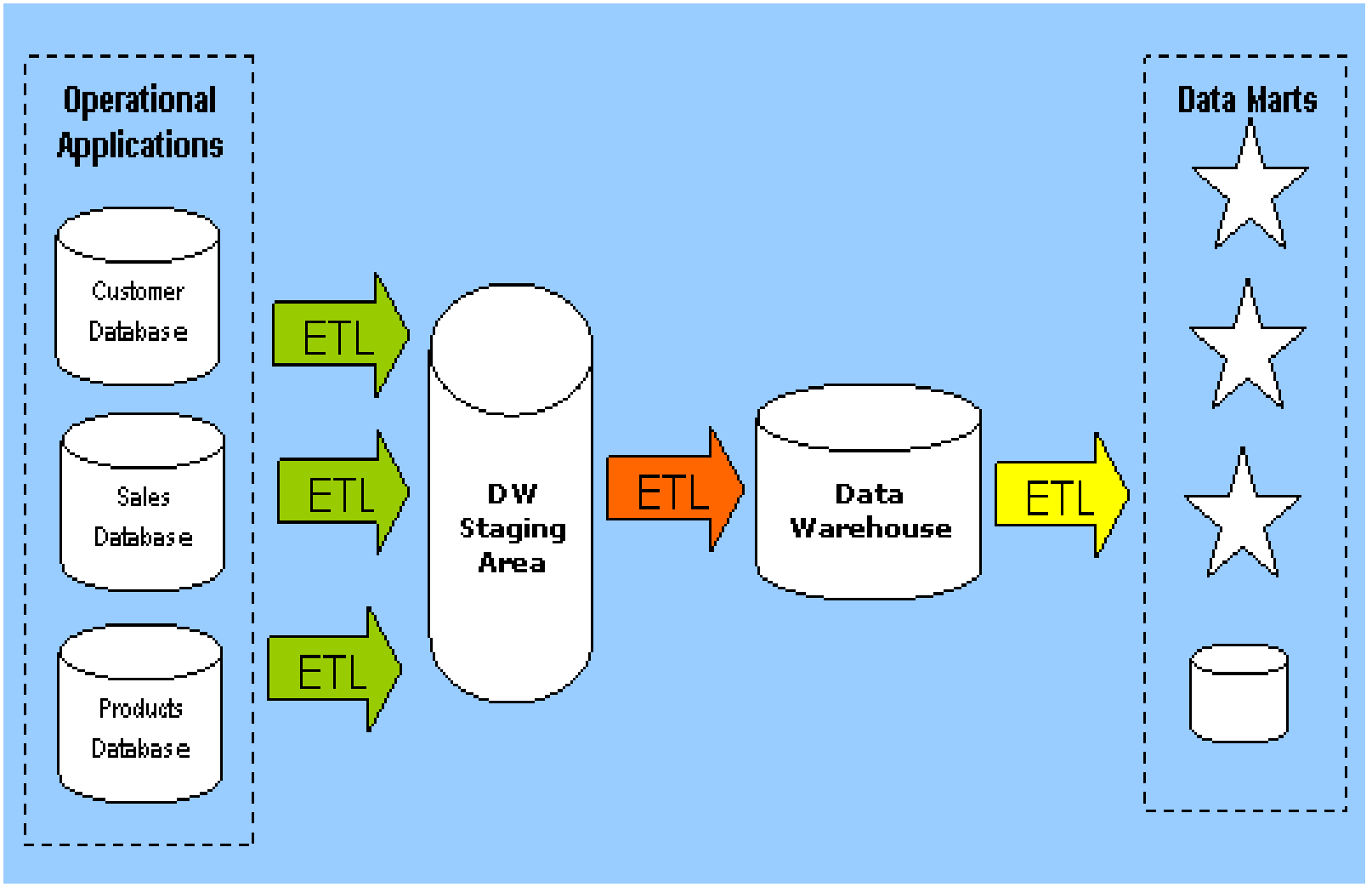
CUENTA

CUENTA	SALDO
C-101	50.000
C-121	120.000
C-203	70.000
C-302	98.000
C-209	42.000
C-109	108.500
C-230	59.000

Estructurados

No Estructurados

Componentes de un DWH



*ETL= **E**xtraction, **T**ransformation, **L**oad*

Componentes de un DWH

1) Sistemas operacionales fuentes (*Operational source systems*)

La fuente de los datos que se desean integrar y generalmente son sistemas legados que manejan la información con *tecnologías heterogéneas*

2) Area de preparación de datos (*Data Stagin Area*)

Herramientas **ETL (***Extraction, Transformation, Load***)**

Componentes de un DWH

3) Area de presentación de datos. (*Data presentation*)

Donde la información se organiza, almacena y se hace disponible para los usuarios. El área de presentación de datos frecuentemente se representa mediante un conjunto de *data marts* que están integrados

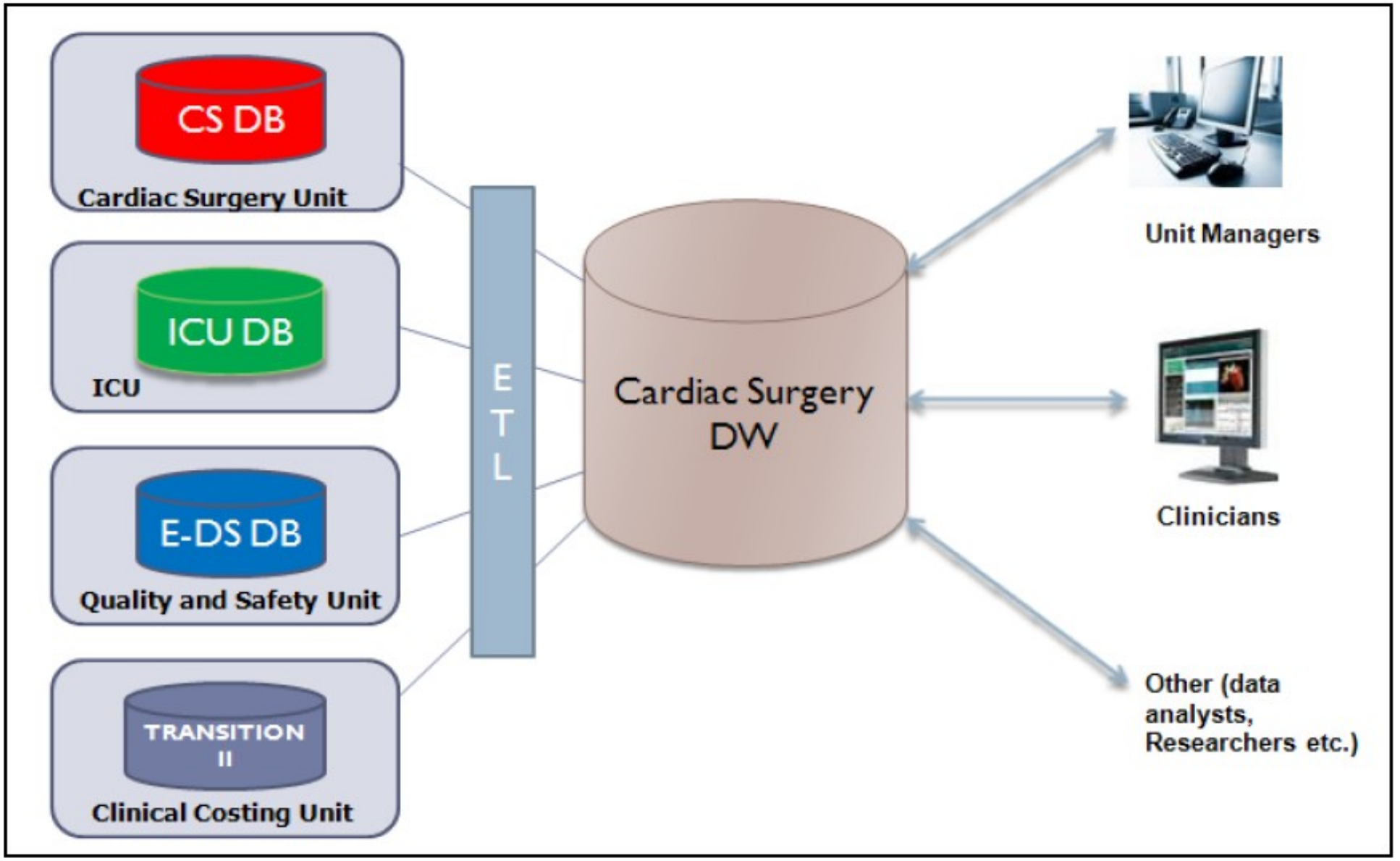
4) Herramientas de Acceso a los datos (Visualización)

Se ofrecen herramientas a los usuarios para hacer consultas, sacar reportes, imprimir informes.

Data Mart

- Técnicas de dwh aplicadas a una subconjunto de datos de la organización
- Igual conjunto de técnicas
- Pequeña escala

Ejemplo: Clinical DW



OLTP vs. OLAP

OLTP

Online Transation Processing

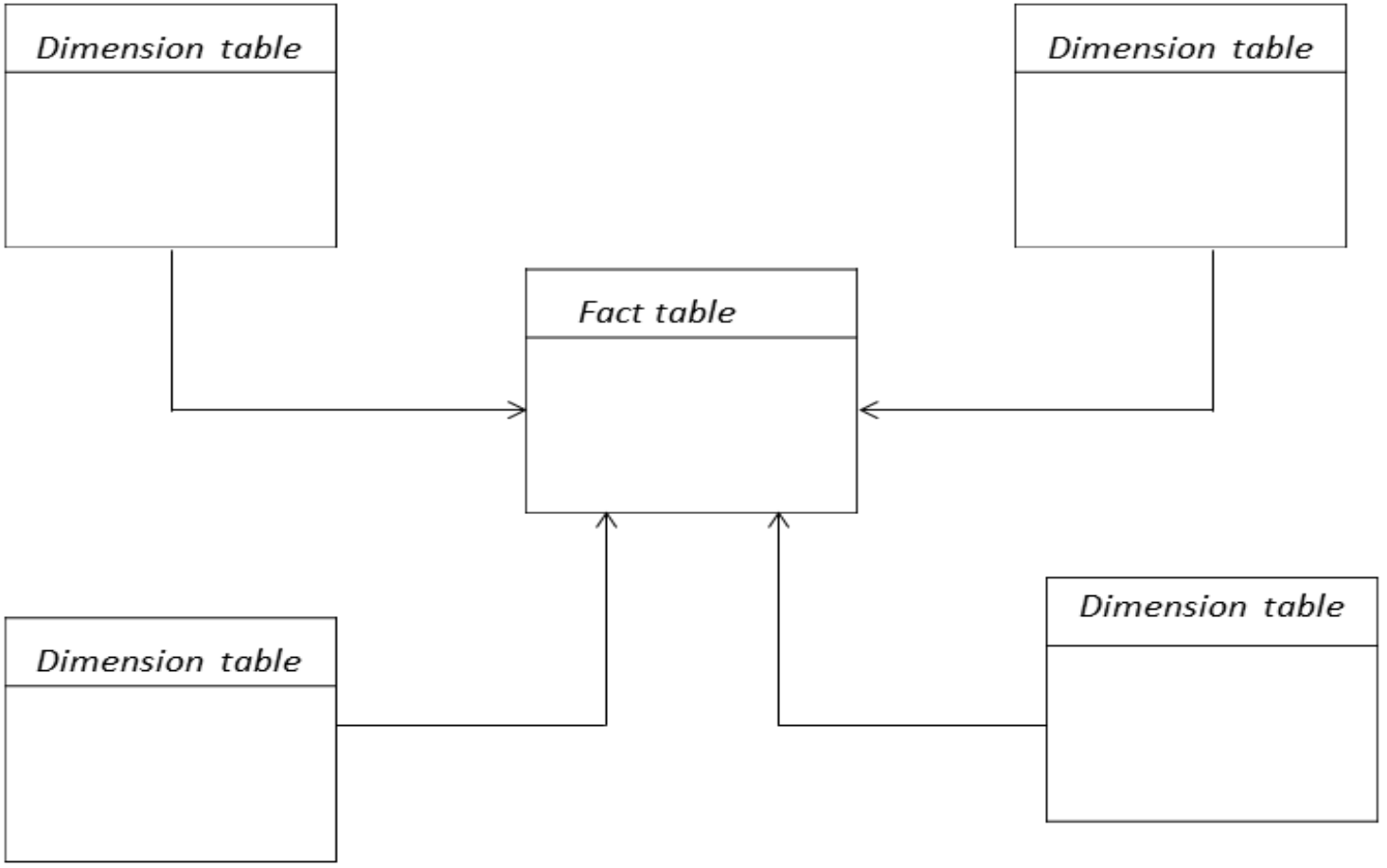
- Actualizaciones
- Muchas Transacciones pequeñas
- Mb-Tb de datos
- Datos crudos (Raw data)
- Datos vigentes

OLAP

Online Analytical Processing

- Lecturas,Consultas complejas y largas
- Gb-Tb de datos
- Resumidos y consolidados
- Toma de decisiones
- Información histórica

Esquema en estrella



Hechos y dimensiones

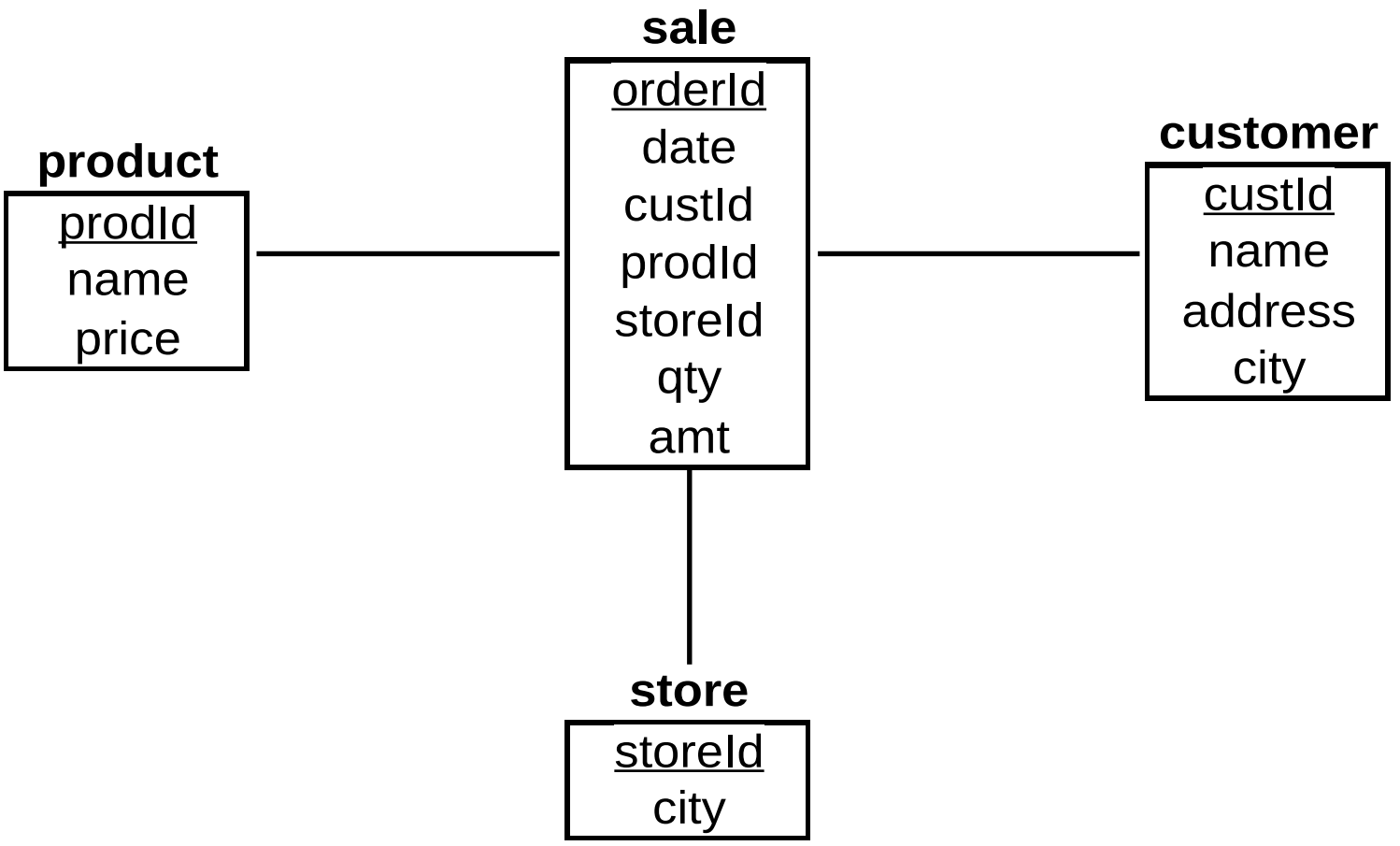
ETL

Extracción: Leer y entender los datos de diferentes fuentes. También se copia los datos que se usarán para ser manipulados posteriormente.

Transformación: Se resuelven conflictos entre los datos de dominio, se trata con elementos perdidos o faltantes o se mapea los datos a formatos estándares

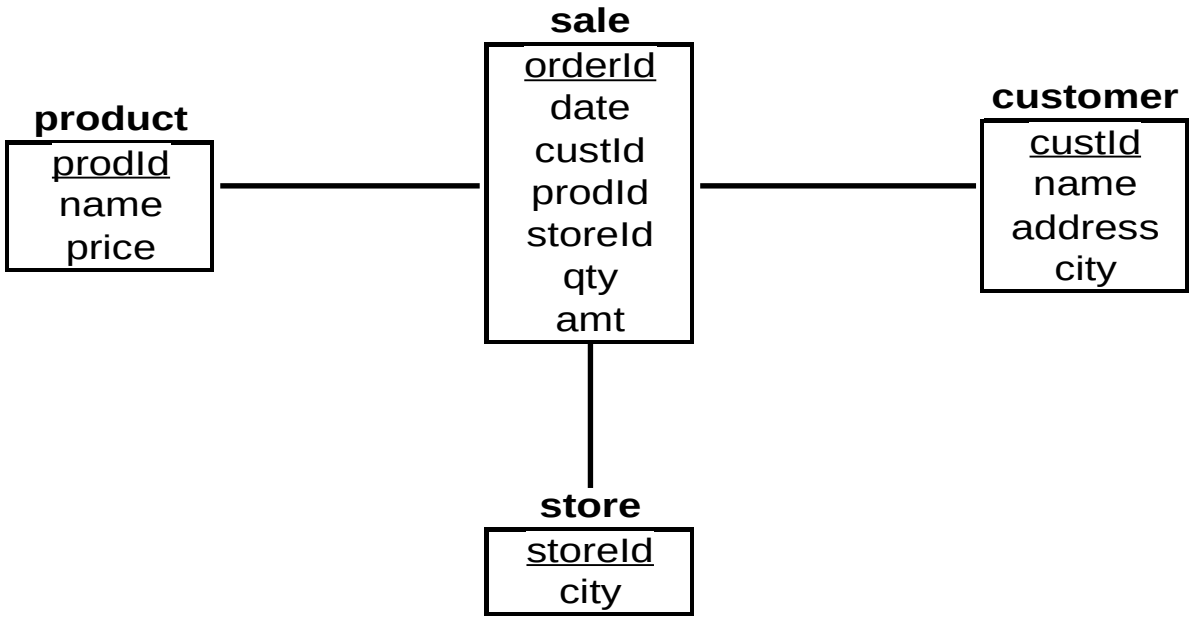
Carga: Se guardan los datos en el *data warehouse*, por lo general corren procesos que insertan datos en las dimensiones y las tablas de hechos.

Esquema en Estrella



Términos

- Tabla de Hechos
- Tablas de Dimensión
- Medidas

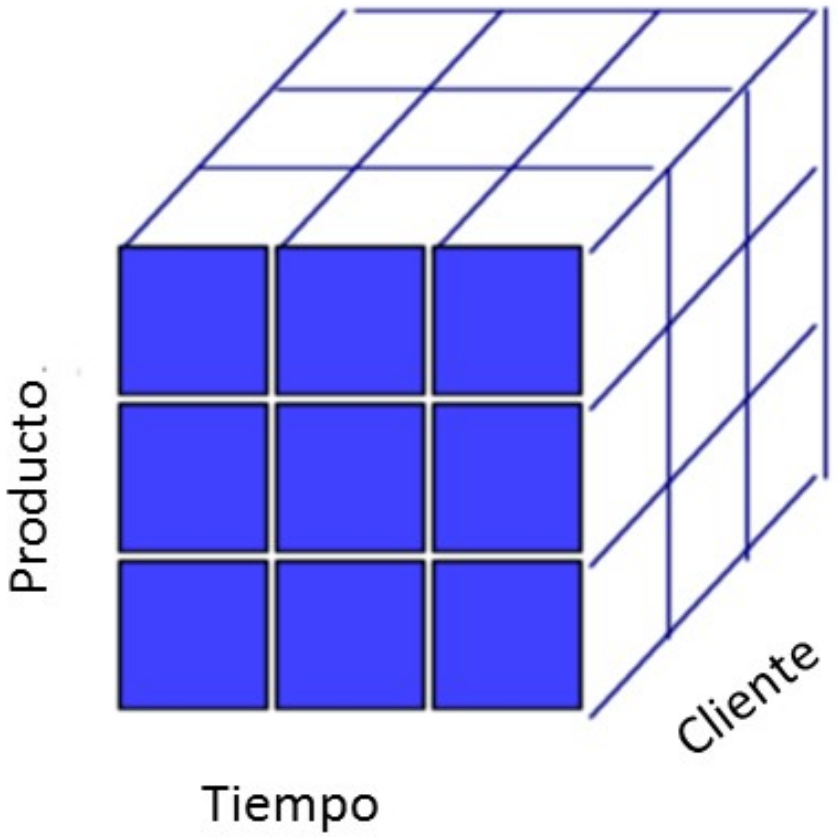


ROLAP vs. MOLAP

- ROLAP:
  Relational On-Line Analytical Processing

- MOLAP:
  Multi-Dimensional On-Line Analytical Processing

Cubo



DATA MINING

Data mining, also known as knowledge discovery in data (**KDD**), is the process of **uncovering patterns** and other valuable information from large data sets.

Evolution of data warehousing technology and the growth of big data, adoption of data mining techniques has rapidly accelerated over the last couple of decades, assisting companies by transforming their **raw data** into **useful knowledge**

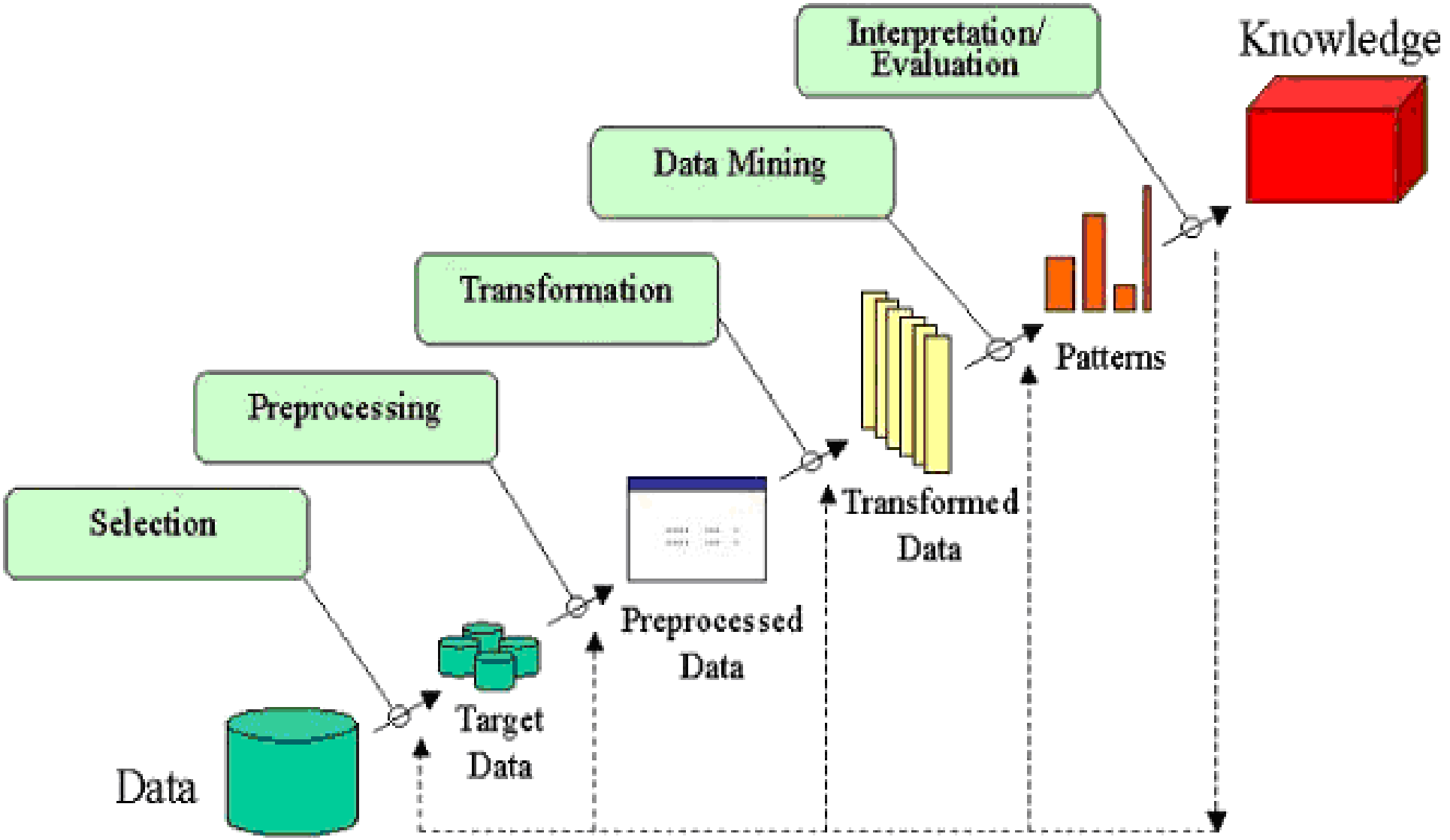
https://www.ibm.com/topics/data-mining

DATA MINING

- Análisis y la exploración de grandes volúmenes de datos para **descubrir patrones significantes** (automática o semi automáticamente)

- La meta: mejorar procesos de ventas, marketing y en general la relación con los clientes.

Data Mining es un proceso



Data Mining proporciona inteligencia

- Bases de datos proporcionan los datos
- Necesidad: explorar datos y encontrar patrones, reglas (entender qué está pasando)
- Predecir que pasará
- Se requieren: técnicas y herramientas para extraer el máximo beneficio de los datos.

DATA MINING

¿Cómo nos ayuda?

- ¿Quiénes son nuestros clientes fieles?
- Clientes que dejarán la compañía.
- ¿Dónde localizo la próxima sucursal?
- ¿Cuáles son mis productos más beneficiosos?
- ...

- Las respuestas están en los datos. Técnicas de *data mining* pueden ayudar a encontrarlas

¿Por qué ahora?

- Las técnicas existentes
- convergencia de una serie de factores:
  - Cantidad de datos
  - Datos integrados (data warehouse)
  - Más capacidad de cómputo de los computadores
  - Competencia feroz

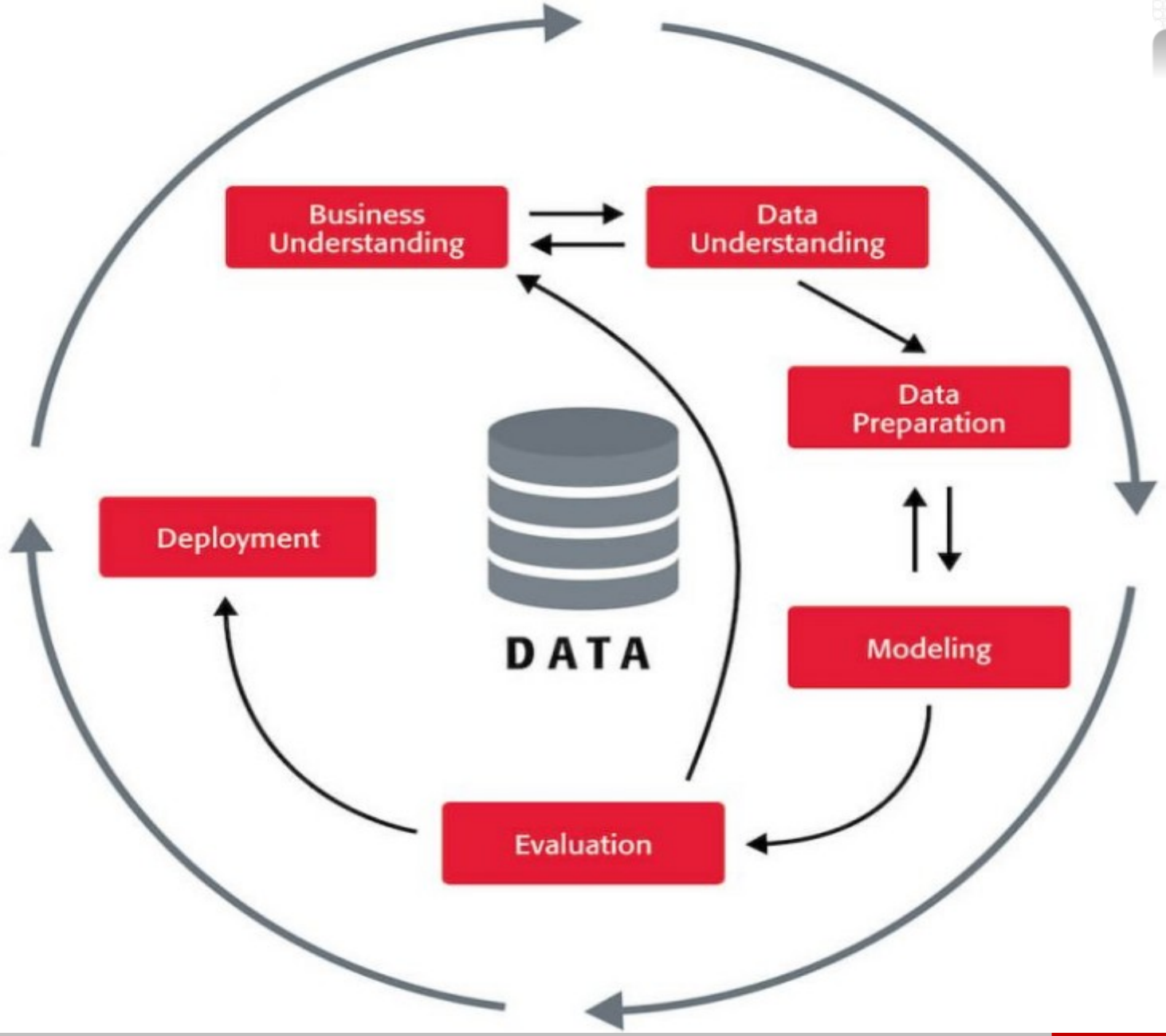
DATA MINING

Importante

- La promesa de Data Mining : encontrar los patrones
- Hallarlos no es suficiente
- Los patrones se tienen que entender y valorar
- El entendimiento de los patrones facilitan actuar
- se transforman en valor para la compañía.

Datos transformados en **Conocimiento** que permita actuar en forma rápida y eficiente

Metodologia CRISP-DM



Importante

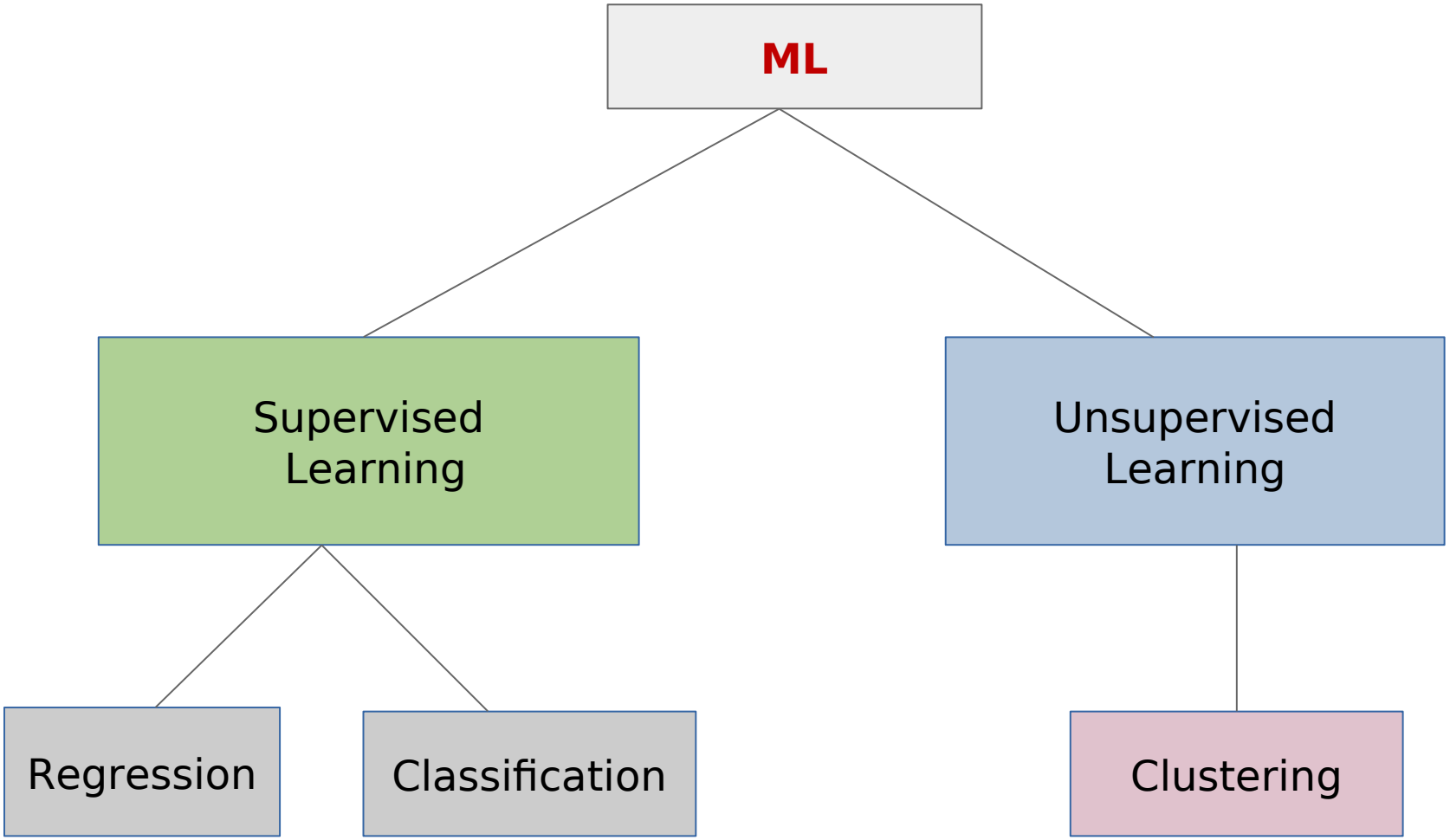
- La promesa de Data Mining : «Encontrar patrones en los datos»

Data Mining es un proceso

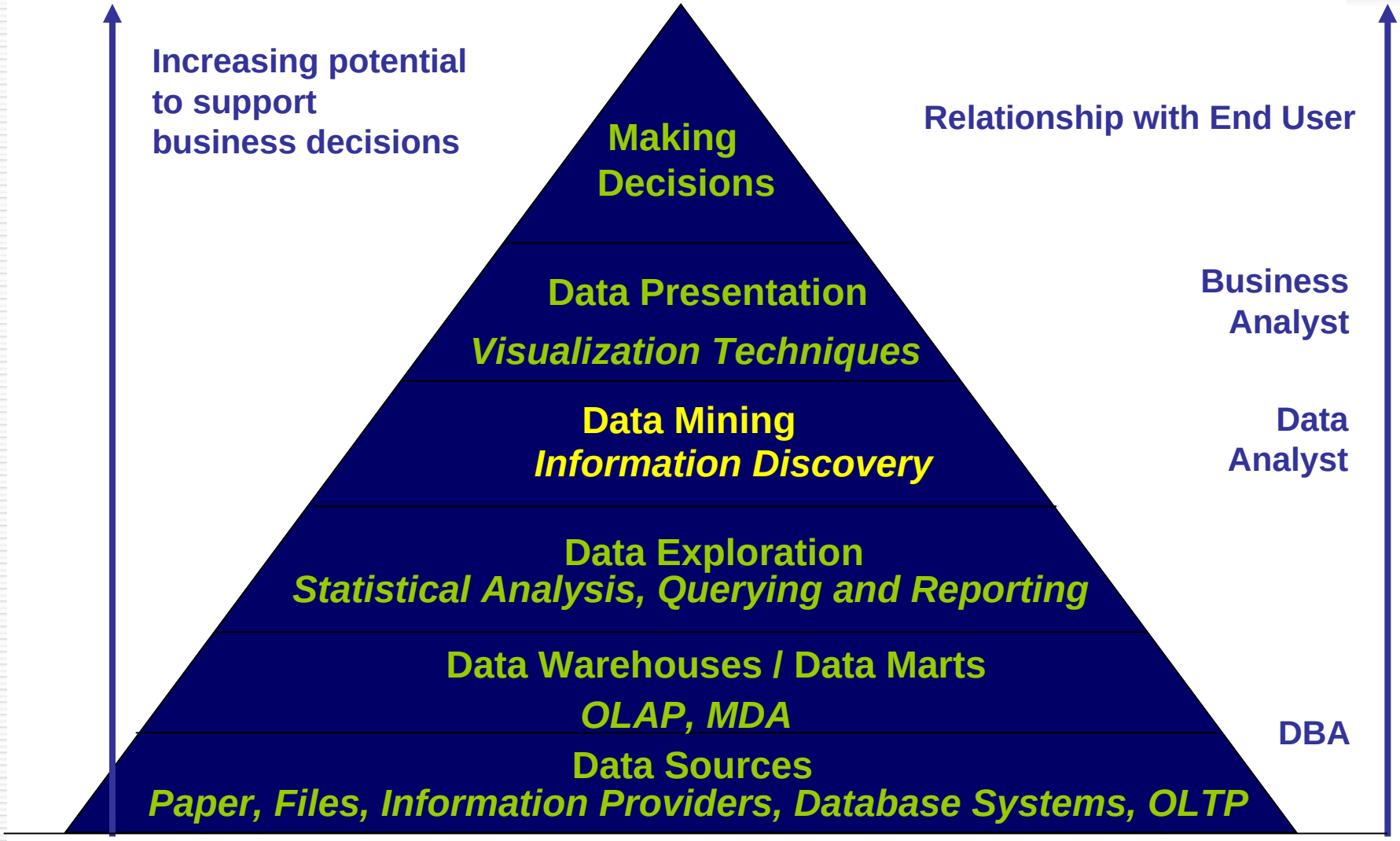
- orientado a las acciones que se derivan del hallazgo de los patrones
- algoritmos importantes, la solución al problema mas compleja que un conjunto de técnicas y algoritmos
- Las técnicas se aplican sobre datos apropiados en el momento apropiado
- Los datos operacionales requieren preprocesamiento

MACHINE LEARNING

Campo de la **Inteligencia Artificial** que usa algoritmos que tienen la capacidad de identificar (**Aprender**) patrones en datos masivos y elaborar **predicciones**.

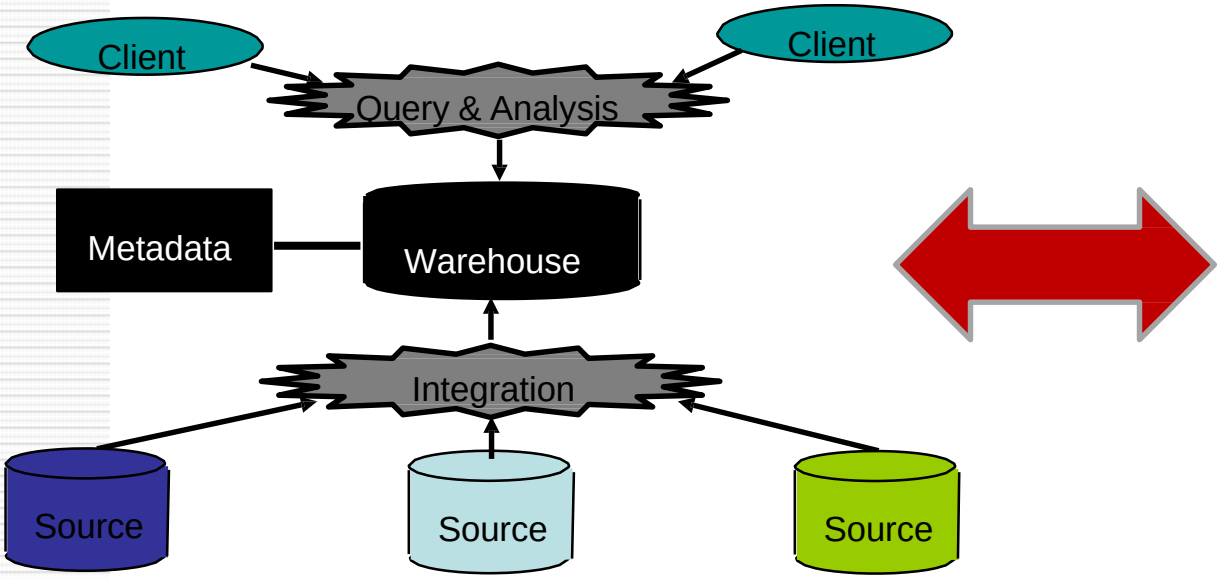


DATA MINING & BI

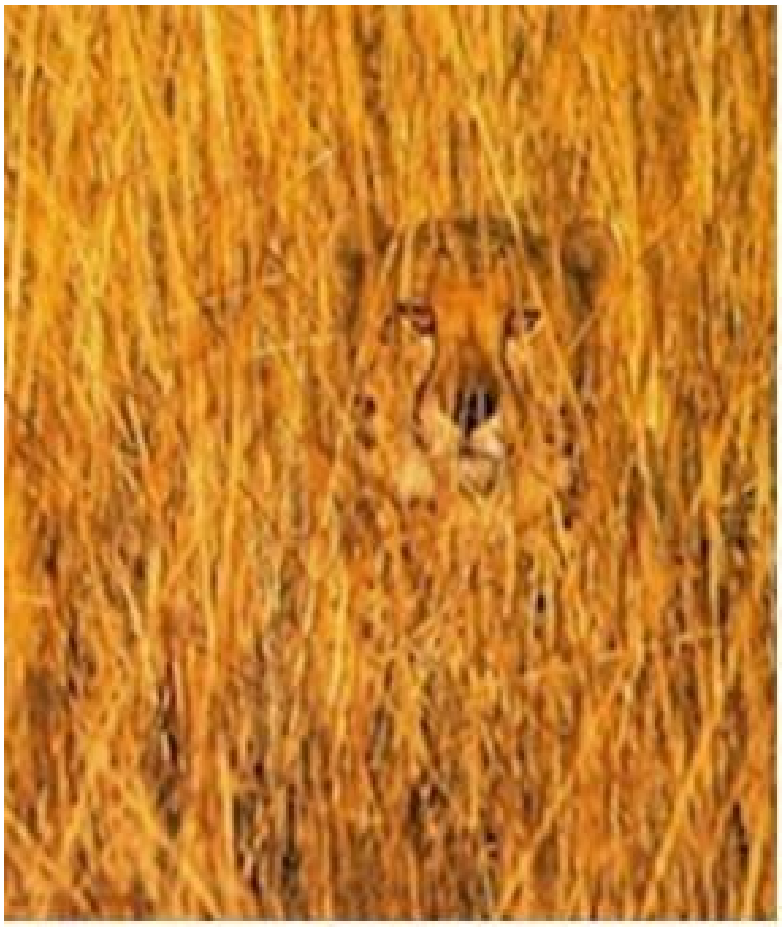


En resumen....

- Data warehousing



- Data mining




DATA
DATA EVERYWHERE

Científico de Datos

Gracias