Introducción al concepto de los datos y su análisis

Tema 1

Profesores:

Juan C. Trujillo
Alejandro Reina Reina
LUCENTIA Research Group

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos

Índice

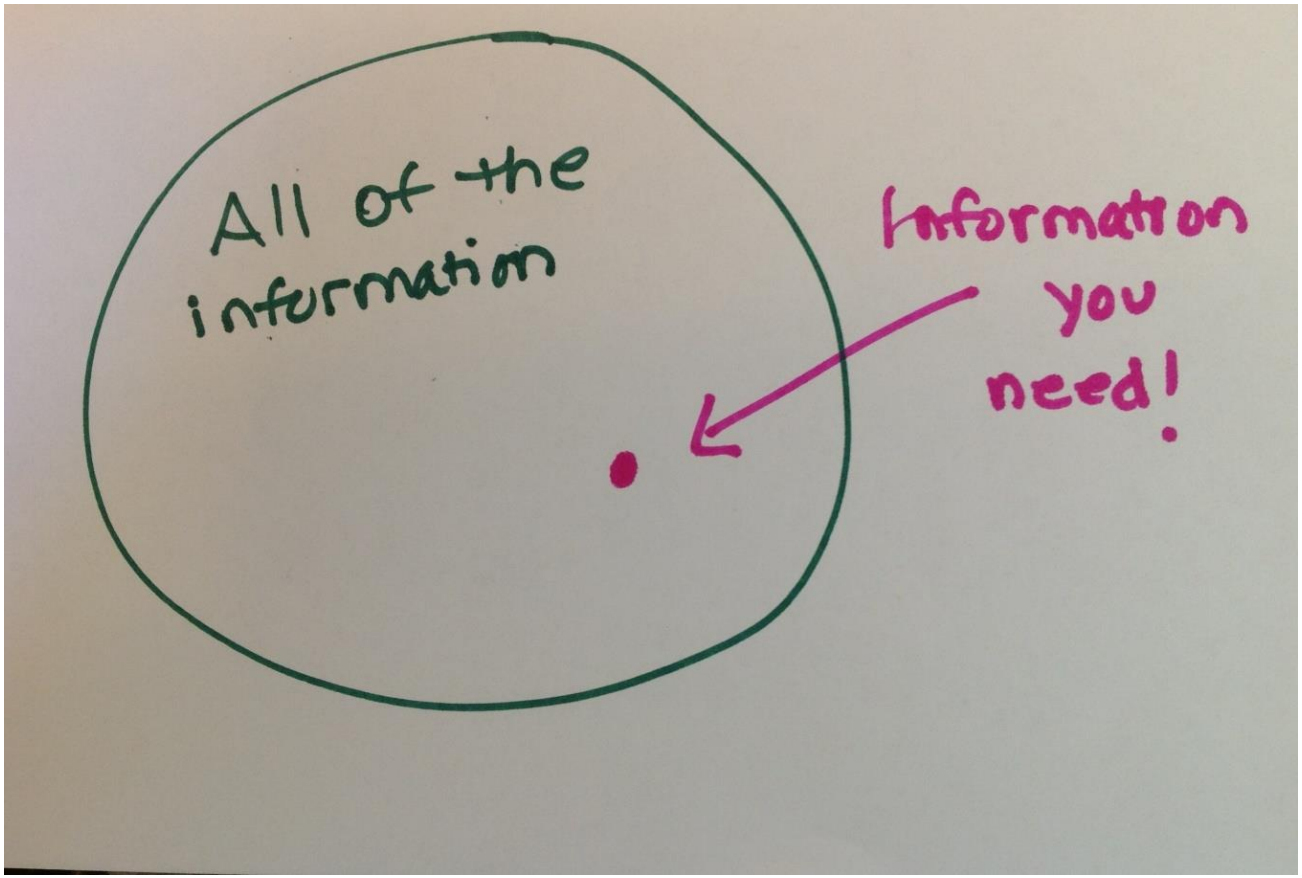
- Planificación de la asignatura en Moodle de la UA
  - Acceso a través de UACLOUD

Índice

- Introducción a Big Data
- Nuevas necesidades de gestión de datos
- El almacén de datos: primera aproximación
- Diferencias entre sistemas operacionales y analíticos
- Arquitecturas de DWs
- Esquema tradicional de un DW

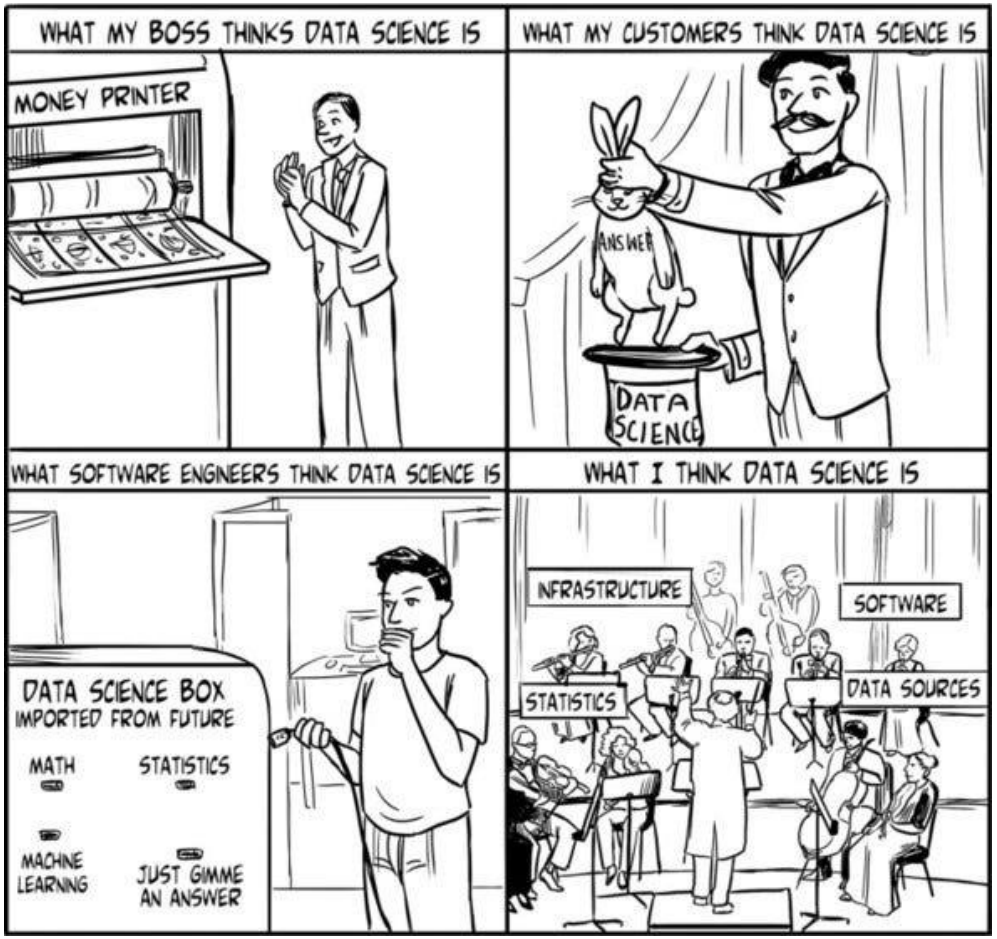
Introducción a Big Data

- Introducción a Big Data



Introducción a Big Data

- ¿Qué es la Ciencia de datos?



Introducción a Big Data

- Un proyecto de Big data no es un proyecto de una única persona
  - Servidores
  - Arquitectura
  - Programación
  - Diseño
  - Análisis
  - Dirección
    - DevOps, Backend, Frontend, Data scientist…

Introducción a Big Data

- ¿Quién realiza esta tarea?
  - Ingeniero de datos
    - Lidia con problemas de recolección, gestión, transformación y publicación de datos
    - Tiene conocimientos de: Modelado, SQL, NoSQL, ETL/ELT, Spark, MapReduce, Kafka, Flume,
    - Desarrollo y despliegue en la nube…

Introducción a Big Data

- ¿Quién realiza esta tarea?
  - Data scientist
    - …a data scientist is 1) a data analyst in California or 2) a statistician under 35 (Gartner blog post by analyst Svetlana Sicular).
    - Lidia con problemas de recolección, gestión, transformación y publicación de datos
    - Tiene conocimientos de: Modelado, SQL, NoSQL, ETL/ELT, Spark, MapReduce, Kafka, Flume, Estadística R, Matlab, SAS, SPSS, Minería de datos, Procesamiento de lenguaje natural, Machine Learning, Map/Reduce, Hadoop, Hive, Python, Desarrollo y despliegue en la nube…
  - La separación entre ambas "figuras" no siempre es clara a la hora de abordar un proyecto

Introducción a Big Data

- Introducción a Big Data
  - The notion of a Data Scientist is a little mad but then so is Big Data.

    Removing the buzzwords just leaves you with….Data.

Introducción a Big Data

Introducción a Big Data

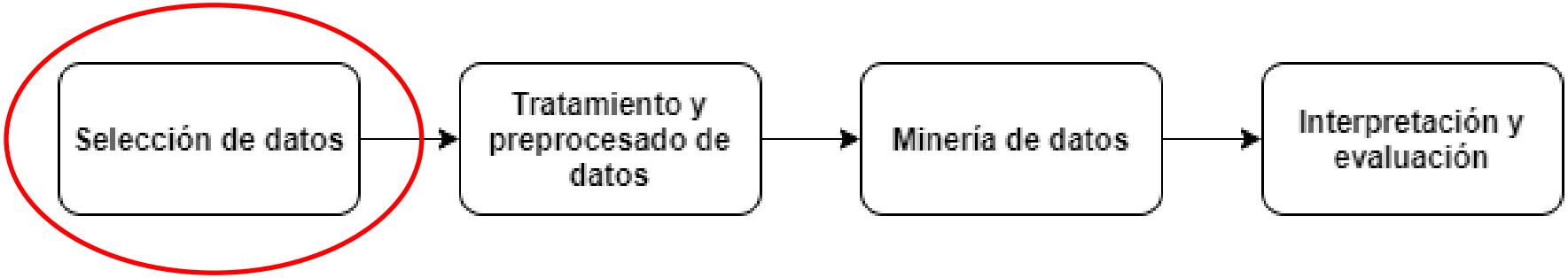
- Con el auge del "Big Data" y el crecimiento de técnicas de inteligencia artificial, ha hecho que el ML sea una practica común tanto en educación como en empresa.
- Un Proyecto de Machine Learning (ML) requiere no solo conocimientos en técnicas de procesamiento de datos y algoritmos, si no también conocimiento profundo del dominio.
- Fuerte dependencia y necesidad de comunicación entre desarrolladores y stakeholders donde además se deben de capturar numerosos requerimientos (funcionales y no funcionales)

Introducción a Big Data

- Big Data e Inteligencia artificial
  - Uno de los principales problemas para el éxito de los proyectos analiticos, es que los proyectos de data warehouses, visualizaciones de usuario y proyectos de ML han sido abordados de manera independiente.
    - Ignorando posibles requerimientos cruzados
    - Restricciones colectivas
    - Dependencias entre salidas de diferentes sistemas
    - Objetivos de negocio

Introducción a Big Data

- Ciclo de vida del dato



Introducción a Big Data

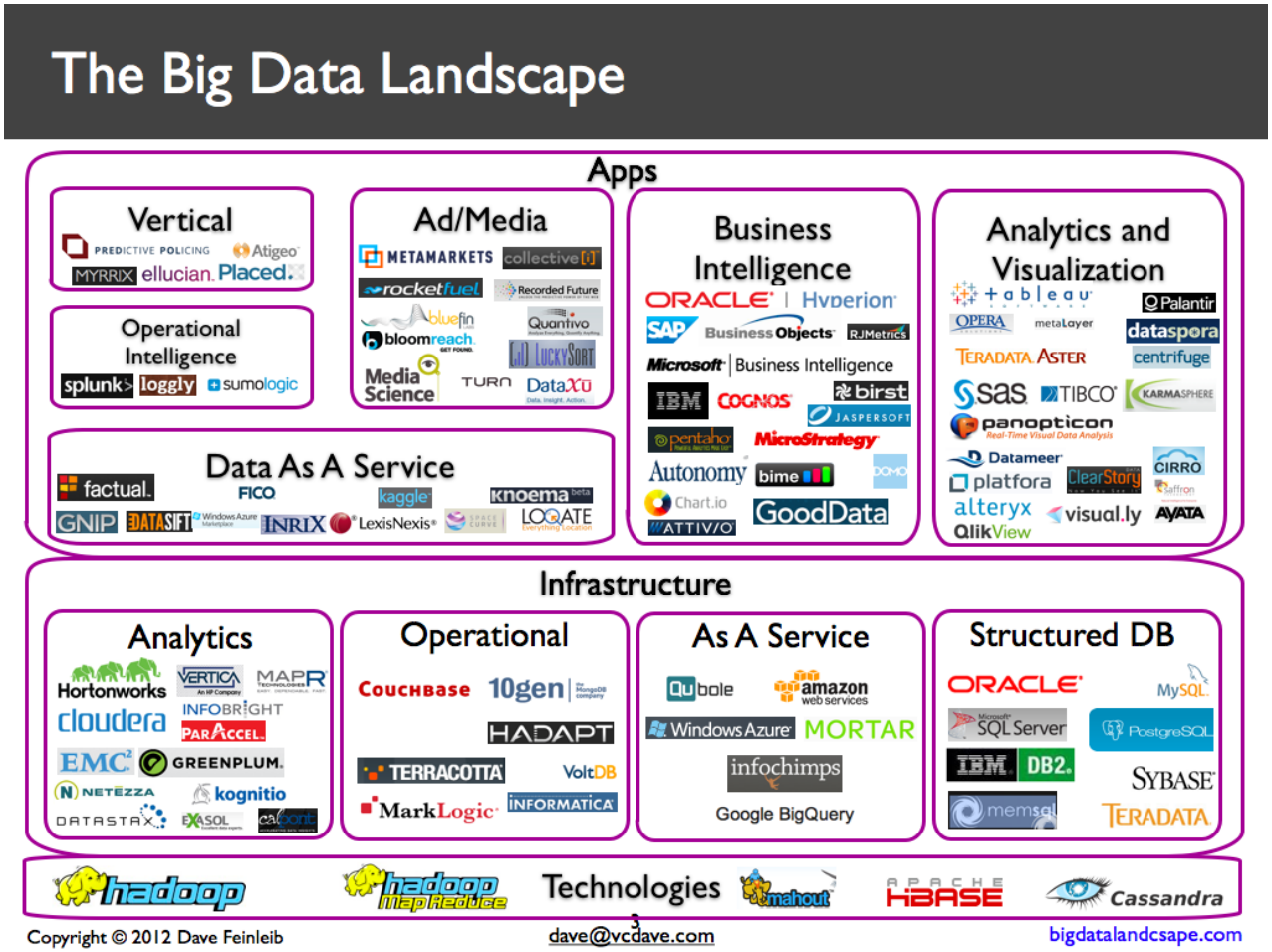
- Adquisición de datos
  - Tratamos de obtener tantos datos de valor como sea posible, esto incluye fuentes internas de la organización, como fuentes externas.
    - Bases de datos, ficheros Excel, informes, APIs…
  - El objetivo es obtener los datos necesarios para el análisis, automatizando en la medida de lo posible la extracción de los mismos (ETLS).
  - En este punto “no importa” tanto los datos si no las entidades y fuentes.

Introducción a Big Data

- BIG DATA para salvar el mundo
  - Siempre hemos tenido mucha información
  - Pero ahora gracias a nuevas herramientas y técnicas se pueden analizar e interpretar
  - También se pueden almacenar más cantidad de información
    - Genoma Humano
    - Datos de Enfermedades
    - LHC

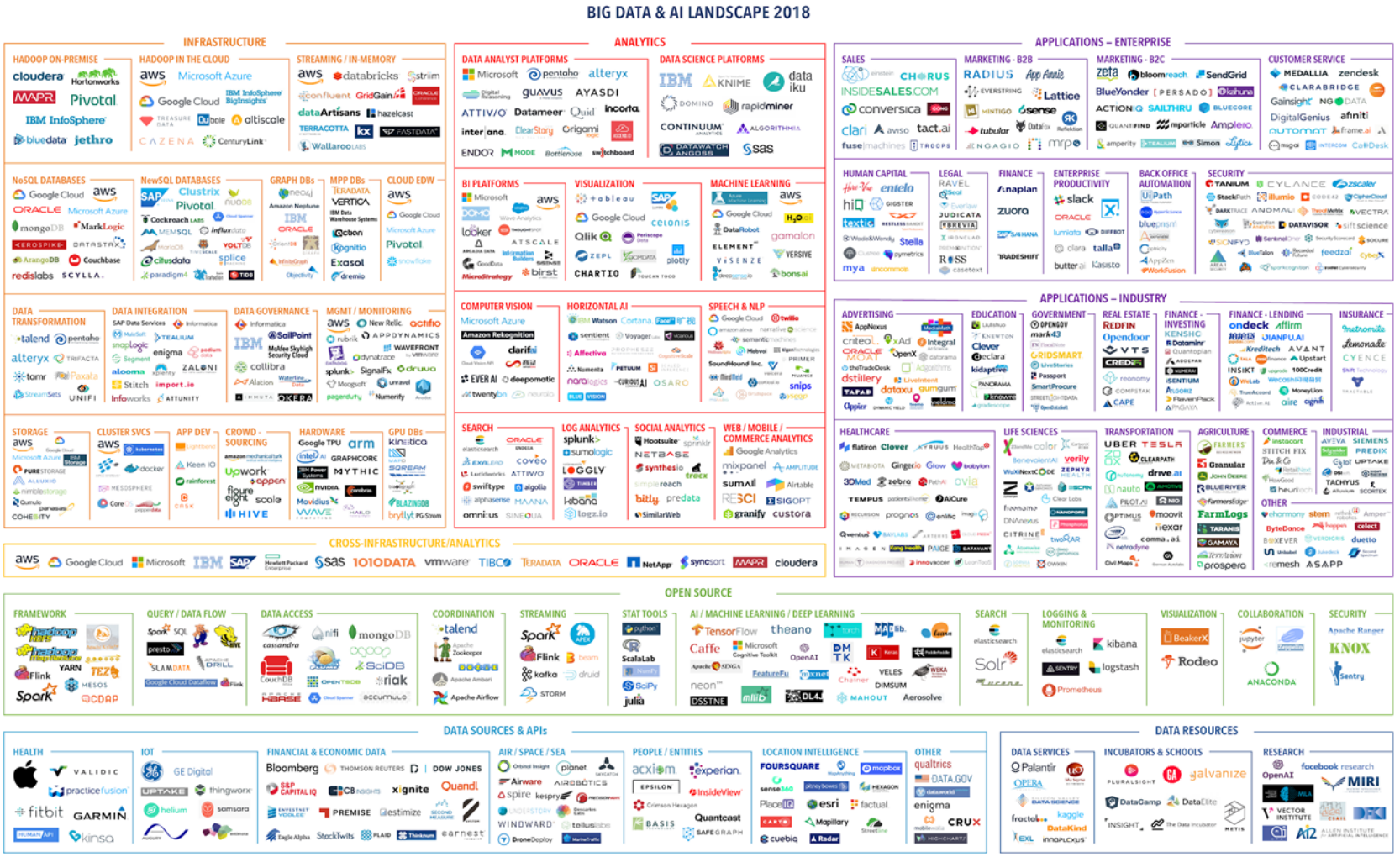
Introducción a Big Data

- Ya hay muchos jugadores…



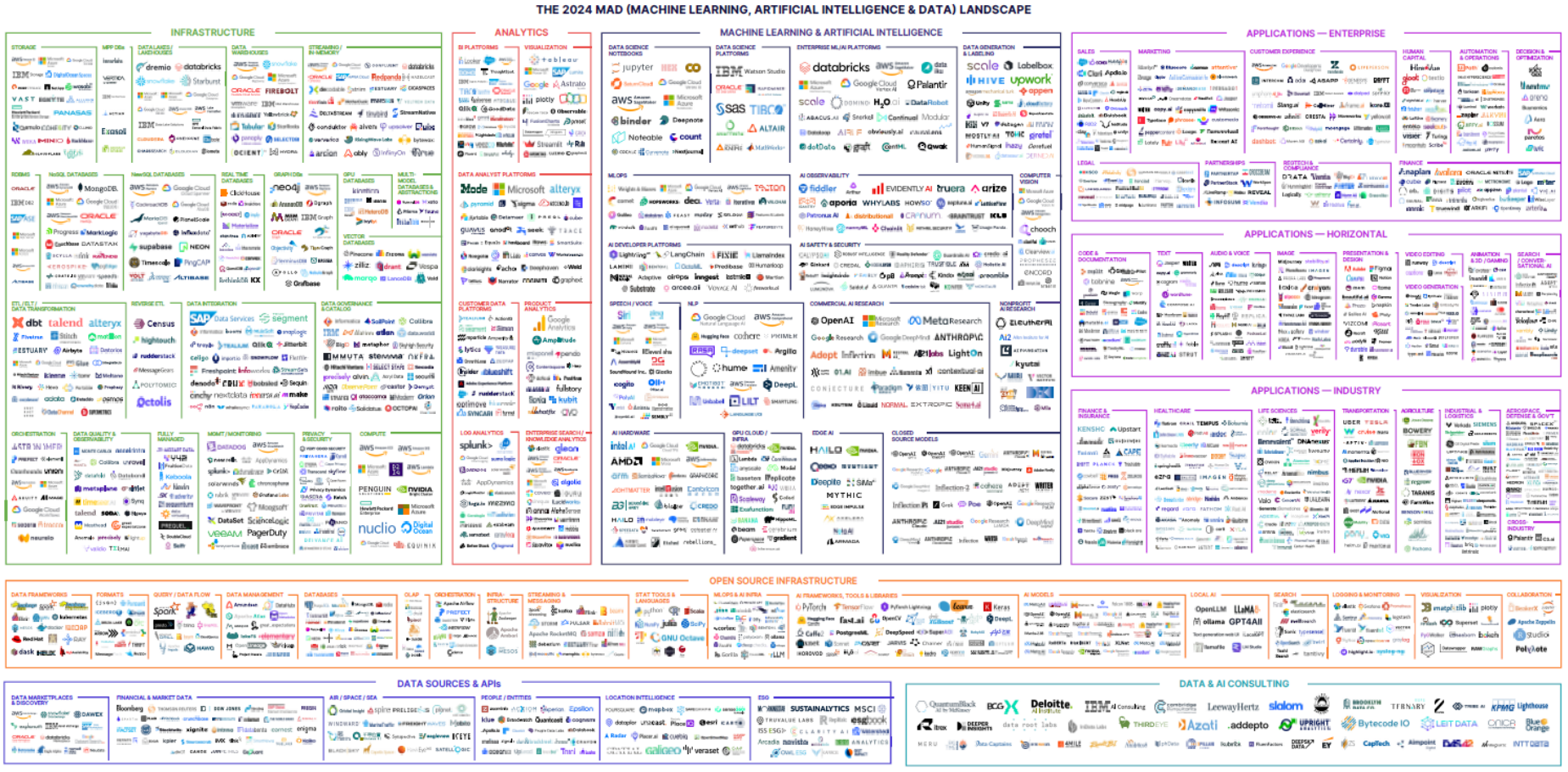
Introducción Big Data

2018



Introducción Big Data

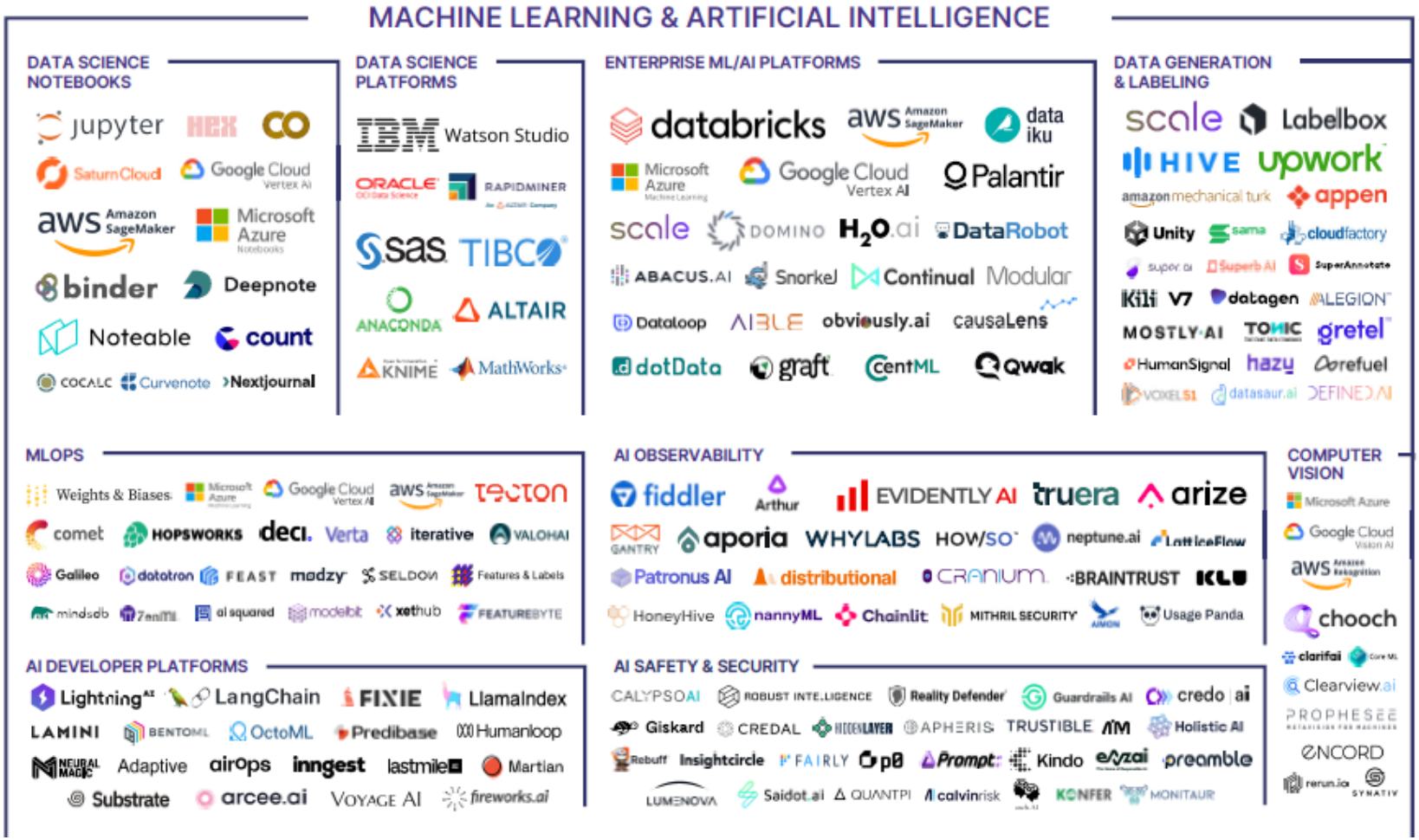
2024



APD. 2025

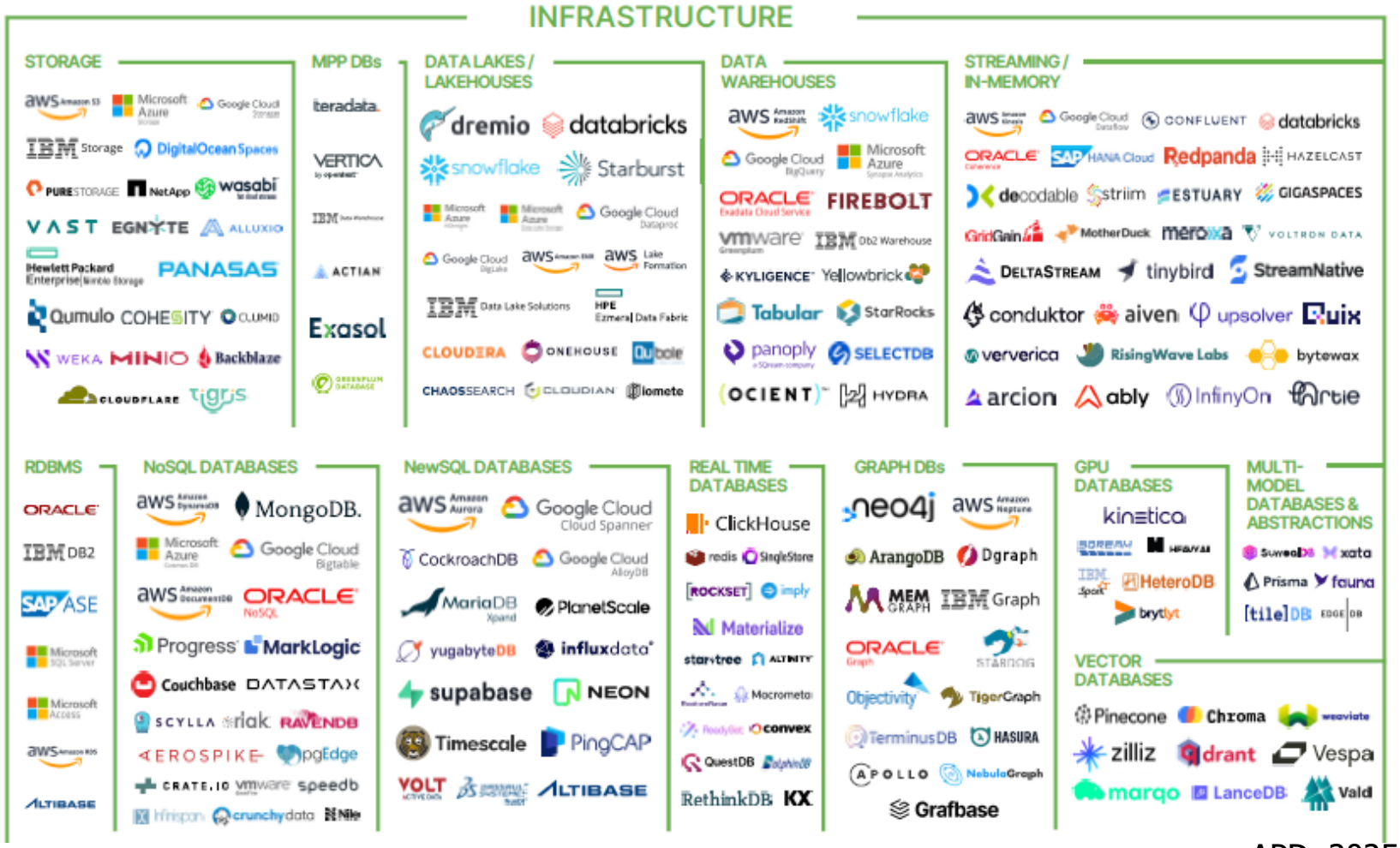
Introducción Big Data

2024



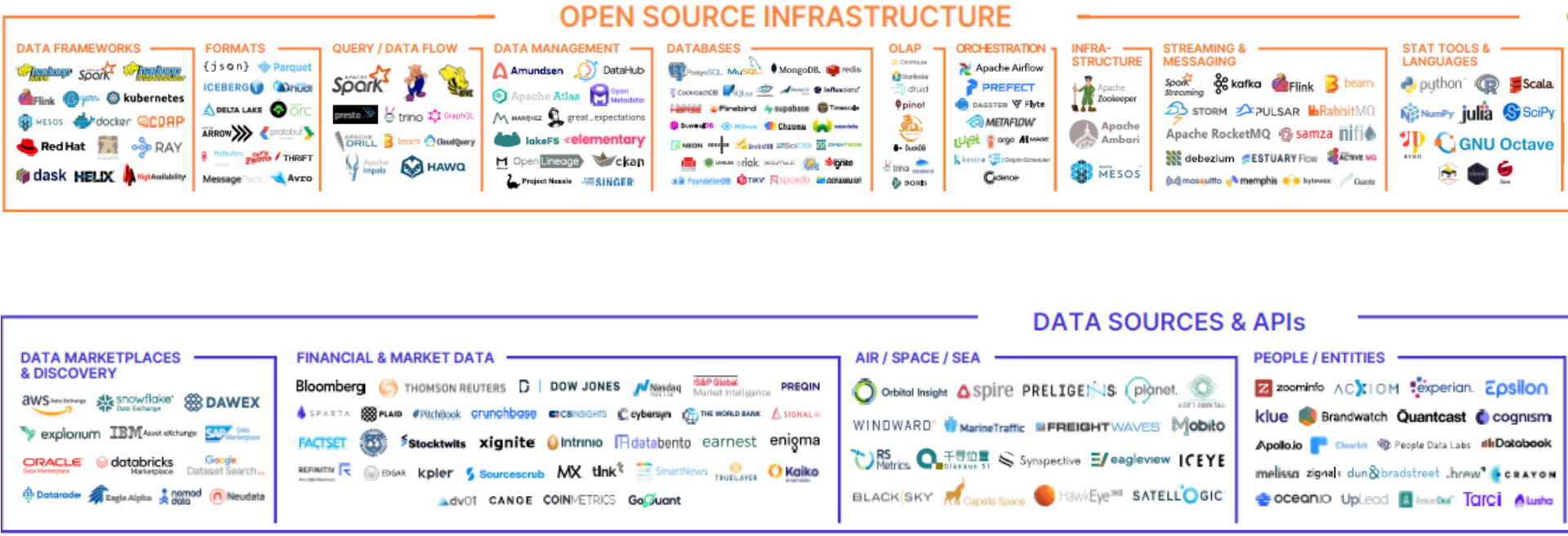
Introducción Big Data

2024



Introducción Big Data

2024



Introducción a Big Data

- Dentro de Big Data se engloba o tiene que ver:
  - Smart City
  - Sensores
  - Seguridad
  - Privacidad
  - Inteligencia artificial
    - …
  - Tendencias
  - Marketing
  - Psicología
    - …

Introducción a Big Data

- ¿Qué es Big data?

  "Forma de afrontar el procesamiento o análisis de grandes volúmenes de información que por su naturaleza desestructurada no pueden ser analizados, y en un tiempo aceptable, usando los procesos y herramientas tradicionales de BI" (IBM)

- Características (5v's)

Veracidad

Velocidad

Valor

Volumen

Variedad

Introducción a Big Data

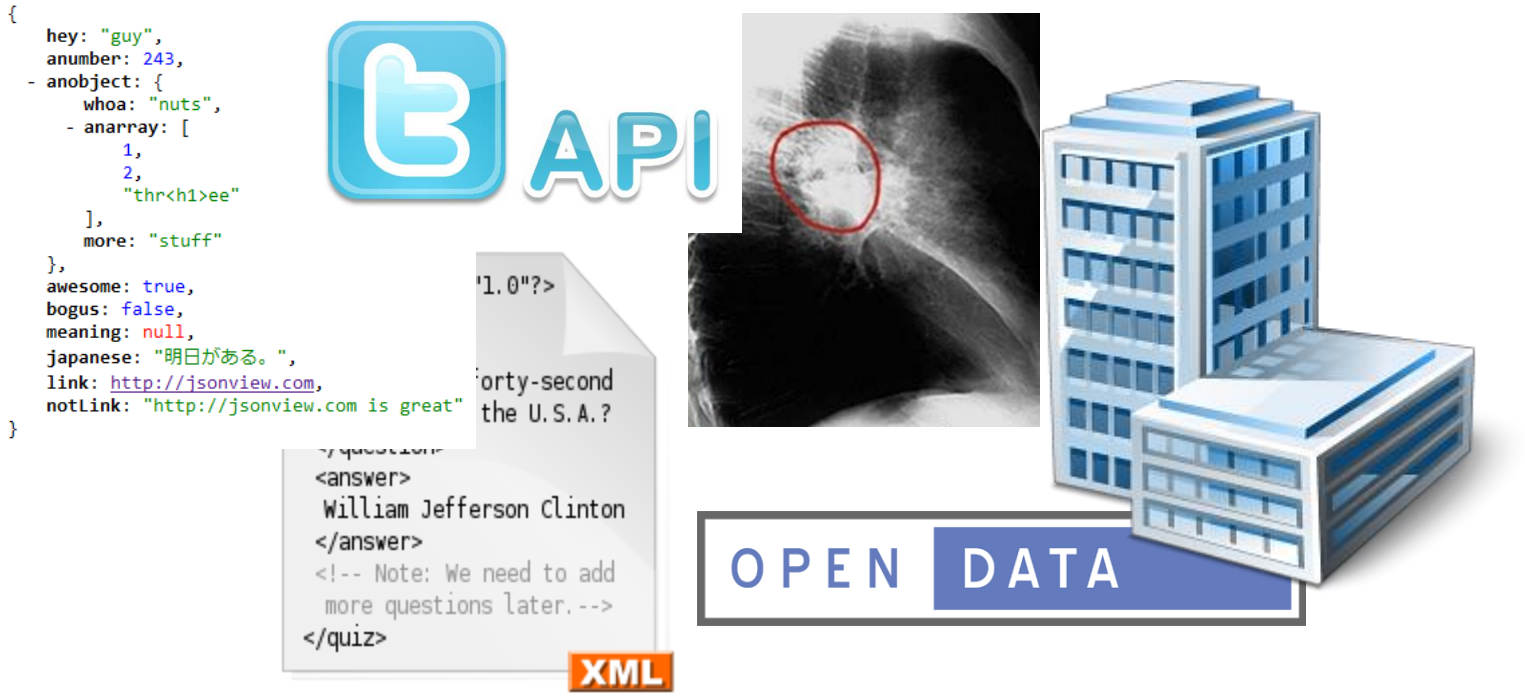
- Volumen: capacidad para procesar **grandes volúmenes de datos**

1 Zettabyte = 10^6 Pettabytes = 10^9 Terabytes = 10^{12} Gigabytes

APD. 2025

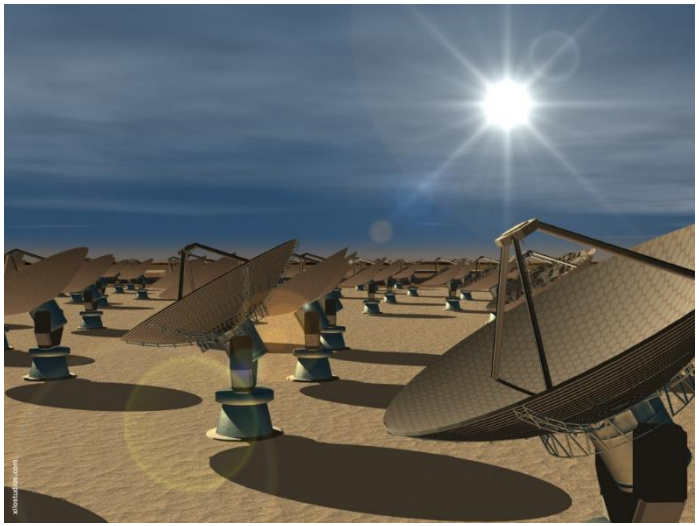
Introducción a Big Data

- Variedad: capacidad para soportar el aumento en la **heterogeneidad** de las **fuentes** a procesar.



Introducción a Big Data

- Velocidad: velocidad a la que fluye la información.

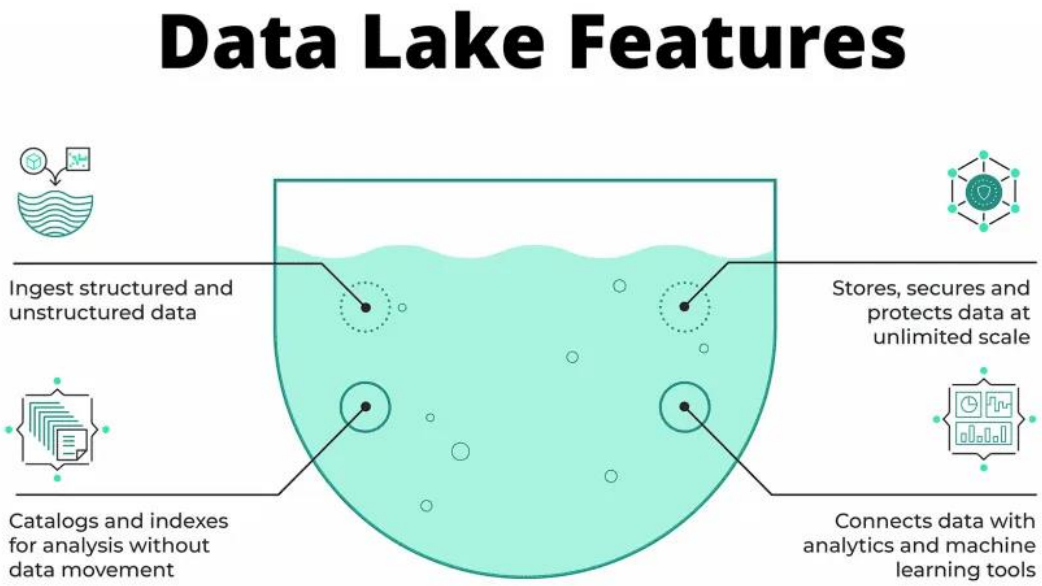


Telescopio SKA
10 petabytes / hora

X
350.000 por min
200 billones por año

Introducción a Big Data

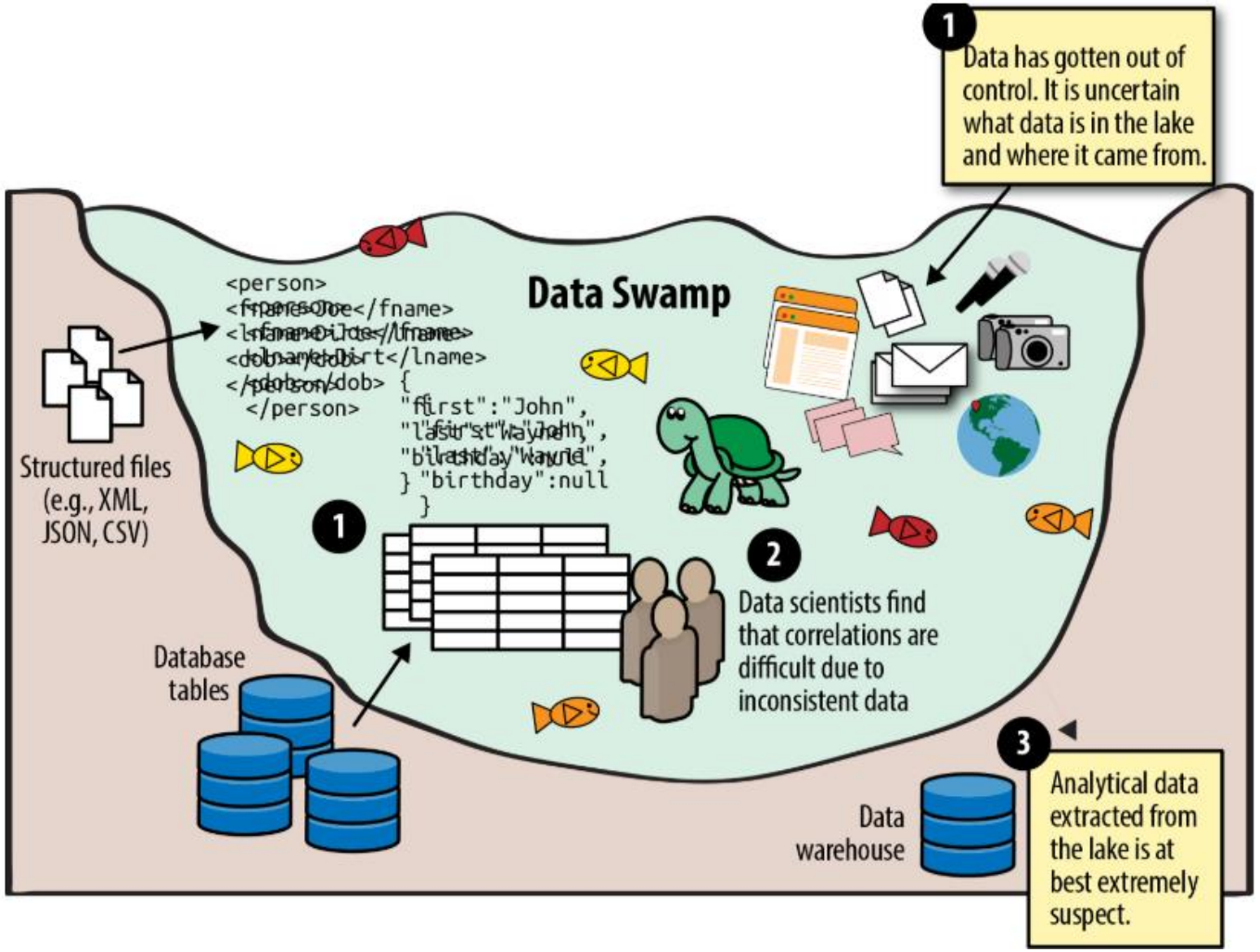
- ¿Dónde se guarda toda esa ingente cantidad de datos?
  - Data Lake: Almacén centralizado de datos en bruto (raw data)



Introducción a Big Data

- ¿Dónde se guarda toda esa ingente cantidad de datos?
  - Data Lake: Almacén centralizado de datos en bruto (raw data)
    - Múltiples formatos: estructurados, semiestructurados y no estructurados
    - Escalable y flexible, pensado para grandes volúmenes de datos
    - Ideal para análisis exploratorio, ciencia de datos e IA
    - Bajo coste de almacenamiento, pero ¡requiere buena gobernanza y gestión o se convertirá un data swamp!

Introducción a Big Data

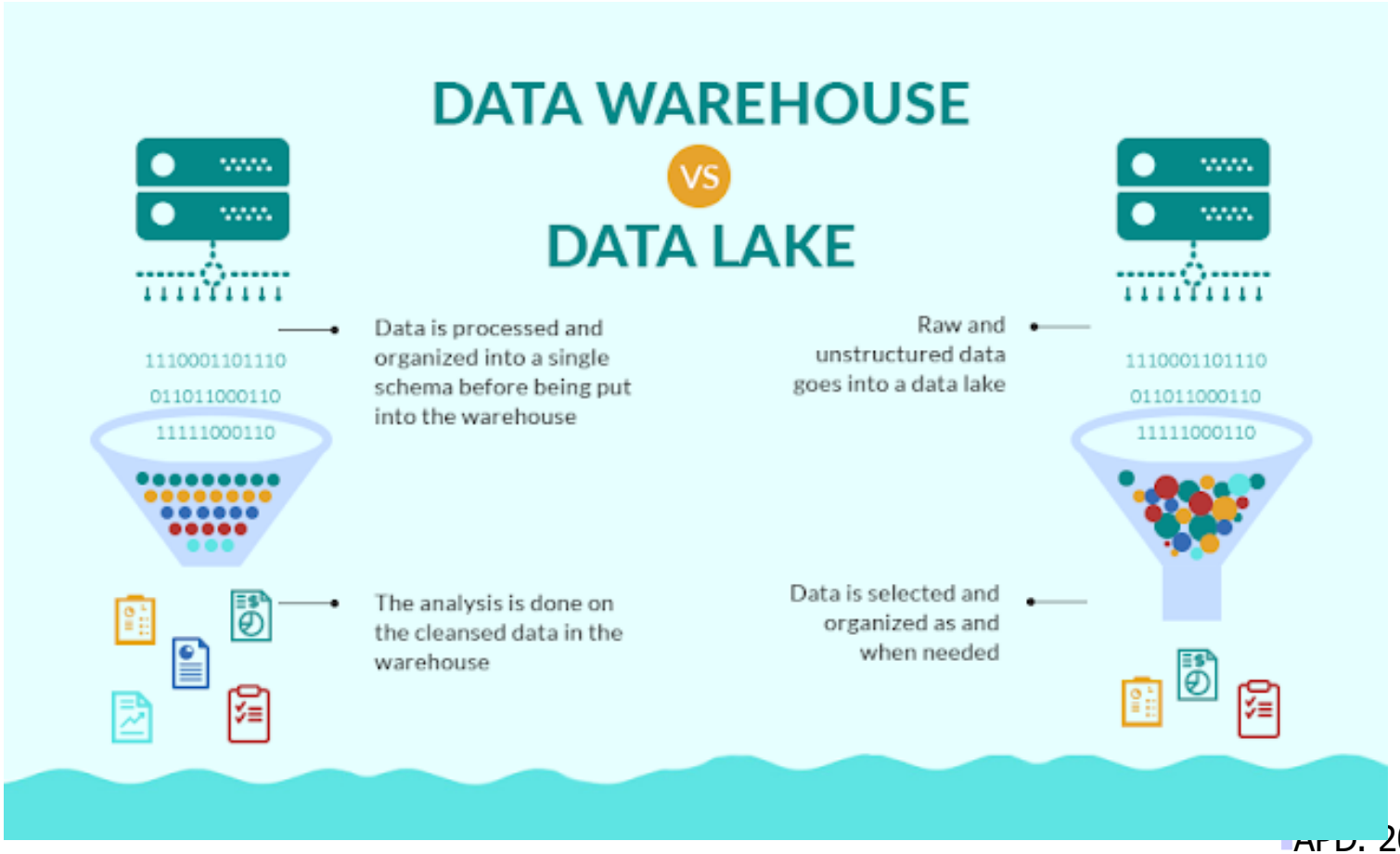


Introducción a Big Data

- ¿Dónde se guarda toda esa ingente cantidad de datos?
  - Data Swamp: ¡Es un Data Lake MAL GESTIONADO!
    - Datos desorganizados, falta de metadatos ni controles de calidad.
    - Difíciles de explorar (¿Qué son estos datos?), de mantener (de donde vienen estos datos?, y de extraer valor (¿Para qué nos sirve esto?)
    - Duplicidades, inconsistencias -> Perdida de confianza

Introducción a Big Data

- ¿Dónde se guarda toda esa ingente cantidad de datos?



Índice

- Introducción a Big Data
- Nuevas necesidades de gestión de datos
- El almacén de datos: primera aproximación
- Diferencias entre sistemas operacionales y analíticos
- Arquitecturas de DWs
- Esquema tradicional de un DW

Nuevas necesidades de gestión de datos

- Situación actual: Entornos económicos altamente competitivos
  - Típica pregunta a SGBDR
    - ¿Cuántos zapatos vendimos el último mes?
  - Empresas necesitan adoptar decisiones estratégicas
    - ¿Cuántos zapatos del 41 de color rojo se vendieron el último mes en la zona norte, este y sur; comparados con las ventas del mismo mes el año pasado ?
    - ¿Qué tipo de cliente me ha estado comprando el BMW 320i durante los últimos 10 años ?
      - ¿Directivo, profesor, trabajador escala básica,...?
  - Empresas quieren conocer que pasará
    - ¿Cuántos zapatos venderé…? ¿Quién comprará…?

Nuevas necesidades de gestión de datos

- Req. 1. Gran volumen de datos (varios años, clientes, productos, almacenes, etc.)
  - Históricos y procedentes de distintas fuentes
- Req. 2. Tienen que ser presentados en un entorno amigable y fácil de usar
  - Entender el tipo de preguntas "estratégicas"

Nuevas necesidades de gestión de datos

- ¿Son válidos los sistemas OLTP para tales decisiones?
  - Problemas
    - Datos históricos no disponibles en sistemas diarios OLTP (OnLine Transaction Processing)
    - Normalmente en distintas fuentes de datos
      - Proveedores, Clientes, componentes, productos defectuosos, etc.
    - Los directivos no saben manejar tales sistemas y,
    - Rendimiento, errores, etc….

Nuevas necesidades de gestión de datos

- El uso de OLTP requeriría:
  - Integrar datos → Consumir tiempo (req. 1)
  - ¿Podría un analista manejar herramientas transaccionales ? (req. 2)

- Vamos a hacer breve historia sobre las soluciones planteadas para manejar datos históricos

Nuevas necesidades de gestión de datos

- Datos procedentes de sistemas heredados
  - 1970's se hace uso de IBM mainframes
    - Cobol, CICS, IMS, DB2, etc
  - 1980's plataformas AS/400 y VAX/VMS
  - Hoy en día muchas aplicaciones de negocio "corren" sobre estos sistemas
    - Muchos años recogiendo datos y reglas de negocio → dificultad para llevarlos a otro sistema
    - Los datos se vuelcan en librerias donde accederán otras aplicaciones de negocio
  - Coste de aplicaciones de negocio es grande

Nuevas necesidades de gestión de datos

- Datos extraídos en el escritorio (1990's)
  - Se reduce la distancia entre usuario final y programador
    - PC con hojas de cálculo, herramientas de análisis, etc.
  - Herramientas de análisis que acceden a datos producidos por sistemas heredados
  - Problema: datos permanecen fragmentados y están orientados a necesidades específicas de usuarios final
    - Soluciones parciales
    - No todos usuarios tienen la misma destreza

Nuevas necesidades de gestión de datos

El Excel se ha usado en la Fórmula 1 hasta que se han dado cuenta que no es la mejor forma de controlar las 20.000 piezas del coche

James Vowles, nuevo jefe de Williams, encontró uno de los motivos por los que el histórico equipo de la Fórmula 1 estaba tan atrasado

2025

Nuevas necesidades de gestión de datos

- Sistemas de apoyo a la decisión (SAD) y Sistemas de Información de Ejecutivos (SIE):
  - SAD: Datos en detalle. Ejecutivos medios y bajos
  - EIS: Datos consolidados. Altos ejecutivos
    - Más orientados a vista multidimensional de los datos
  - Son similares y solapan funcionalidades
  - Son los precursores de los Almacenes de datos
  - Alto precio y de nuevo descoordinación sobre los datos necesarios para efectuar el análisis

Nuevas necesidades de gestión de datos

- Resumen características SAD y EIS:
  - Datos descritos en términos de negocio, en lugar de términos técnicos como tupla, fila o tabla.
    - Sistemas enfocados a usuarios no técnicos
  - Datos preprocesados siguiendo patrones de reglas de negocio
    - Beneficios por la venta de productos en distintos almacenes
  - Vistas consolidadas de datos
    - Aunque permiten ver datos en detalles, rara vez pueden acceder a todos los datos en detalle

Índice

- Introducción a Big Data
- Nuevas necesidades de gestión de datos
- El almacén de datos: primera aproximación
- Diferencias entre sistemas operacionales y analíticos
- Arquitecturas de DWs
- Esquema tradicional de un DW

El almacén de datos: primera aproximación

- Evolución de los sistemas de almacenes de datos
  - Desde los primeros sistemas de gestión de ficheros (a. 1970) hasta los Sistemas de Gestion de Bases de Datos actuales (SGBD)
    - SGBD eficientes
    - SGBD robustos
    - Múltiples herramientas de alto nivel que facilitan su manejo
      - Servidor
      - Cliente

El almacén de datos: primera aproximación

- Empresas en la actualidad
  - Aproximadamente el 90% de SGBD son relacionales
    - Integridad de los datos
    - Independencia de los datos, etc.
  - Orientados a los procesos diarios de la empresa
    - Sistemas de Procesamiento Transaccional en Línea (*On-Line Transactional Processing, OLTP*)
    - Compras de productos, ventas, pedidos, gestión de clientes, etc.
  - Datos históricos → almacenamientos externos (Data Centres, almacenamiento en la nube, discos externos,…)

El almacén de datos: primera aproximación

- El Almacén de datos (Data Warehouse, DW)
  - Sistemas que almacenan datos históricos para ser utilizados por los SAD
  - Son sistemas eminéntemente de consulta enfocados a extraer conocimiento de los datos históricos almacenados
  - El análisis de los datos → On-Line Analytical Processing (OLAP)
    - Utilizan el modelado multidimensional (cubos, hipercubos, etc)

El almacén de datos: primera aproximación

- Definición según W. Inmon (1992)

  - "Una colección de datos orientados por tema, variables en el tiempo y no volátiles que se emplea como apoyo a la toma de decisiones estratégicas"

El almacén de datos: primera aproximación

- Orientados por tema
  - El diseño enfocado a responder eficientementa a consultas estratégicas
    - Actividades de interés: compra, ventas, alquileres,…
    - Contexto de análisis: clientes, vendedores, productos, etc…
  - El modelado Multidimensional (primera aproximación)
    - Hechos → actividades de interés
    - Dimensiones → contexto de análisis

El almacén de datos: primera aproximación

- Integrados
  - Datos integrados de distintas fuentes de datos operacionales
- Variables en el tiempo
  - Datos relativos a un periodo de tiempo y se incrementan periódicamente
- No volátiles
  - Los datos almacenados no se modifican ni actualizan nunca, sólo se añaden nuevos datos

El almacén de datos: primera aproximación

- Objetivo técnicas de almacenamiento de datos (Data Warehousing)
  - Ofrecer información histórica para la toma de decisiones integrando la información procedente de distintas fuentes de datos operacionales

El almacén de datos: primera aproximación

- Las tecnologías de almacenamiento de datos integran eficientemente tecnología de bases de datos con análisis de datos
  - Bases de Datos: Un SGBD que soporte el almacén
  - Análisis de datos: herramientas que permitan llevar a cabo un análisis sencillo de los datos
    - Las más extendidas: herramientas OLAP
    - Análisis Multidimensional basado en el modelo multidimensional

El almacén de datos: primera aproximación

- Ejemplo: conocer los ventas de productos con respecto a los productos vendidos, los almacenes donde se vendieron y el tiempo

Ventas			Producto.Grupo = "Supermercado"			
			Comida		Bebida	
			Cong	Fresco	Refresco	Alcohol
Almacén. comunidad = "Comunidad Valenciana"	Alicante	Albatera	100	200	300	400
		Elche	500	600	700	800
	Valencia	Sagunto	900	1000	1100	1200
		Cullera	1300	1400	1500	1600

El almacén de datos: primera aproximación

- Ventajas para las empresas
  - Decisiones soportadas por datos fiables
  - Rentabilidad de inversiones
  - Aumentar la competitividad en los nuevos entornos hostiles
  - Entornos amigables → los directivos analizan por ellos mismos los datos
    - Por fin se entienden con los ordenadores → lo hemos conseguido !!!

El almacén de datos: primera aproximación

- Inconvenientes
  - Infravalorar los recursos necesarios para alimentar el DW a partir de datos operacionales
  - No planificar el alto esfuerzo para lograr un buen diseño
  - Nunca está acabado → incremento continuo de requerimientos ad-hoc
  - ¡¡¡¡ Cuidado !!!! A más cantidad de datos → Más seguridad

Índice

Sistemas operacionales VS Analíticos

	OLTP	OLAP
Usuario	Profesional de TI	Analista de Información
Función	Operaciones diarias	Apoyo a la decisión
Diseño de BD	Orientada a aplicación (Basado en EE-R)	Orientado al tema (esquema estrella, copos)
Datos	Actuales, Aislados	Históricos, Consolidados
Vistas	Detallados, Planos, Relac.	Agregados
Destino/utilización	Estructuradas, repetitivas	Ad-Hoc
Unidades de trabajo	Transacciones simples	Consultas complejas
Acceso	Lectura/escritura	Lectura principalmente
# Registros accedidos	Decenas	Millones
# Usuarios	"Miles"	"Centenares"
Tamaño de la BD	100 MB-GB	100 GB-TB
Medidas de rendimiento	Nº transacciones	Nº consultas, Respuesta

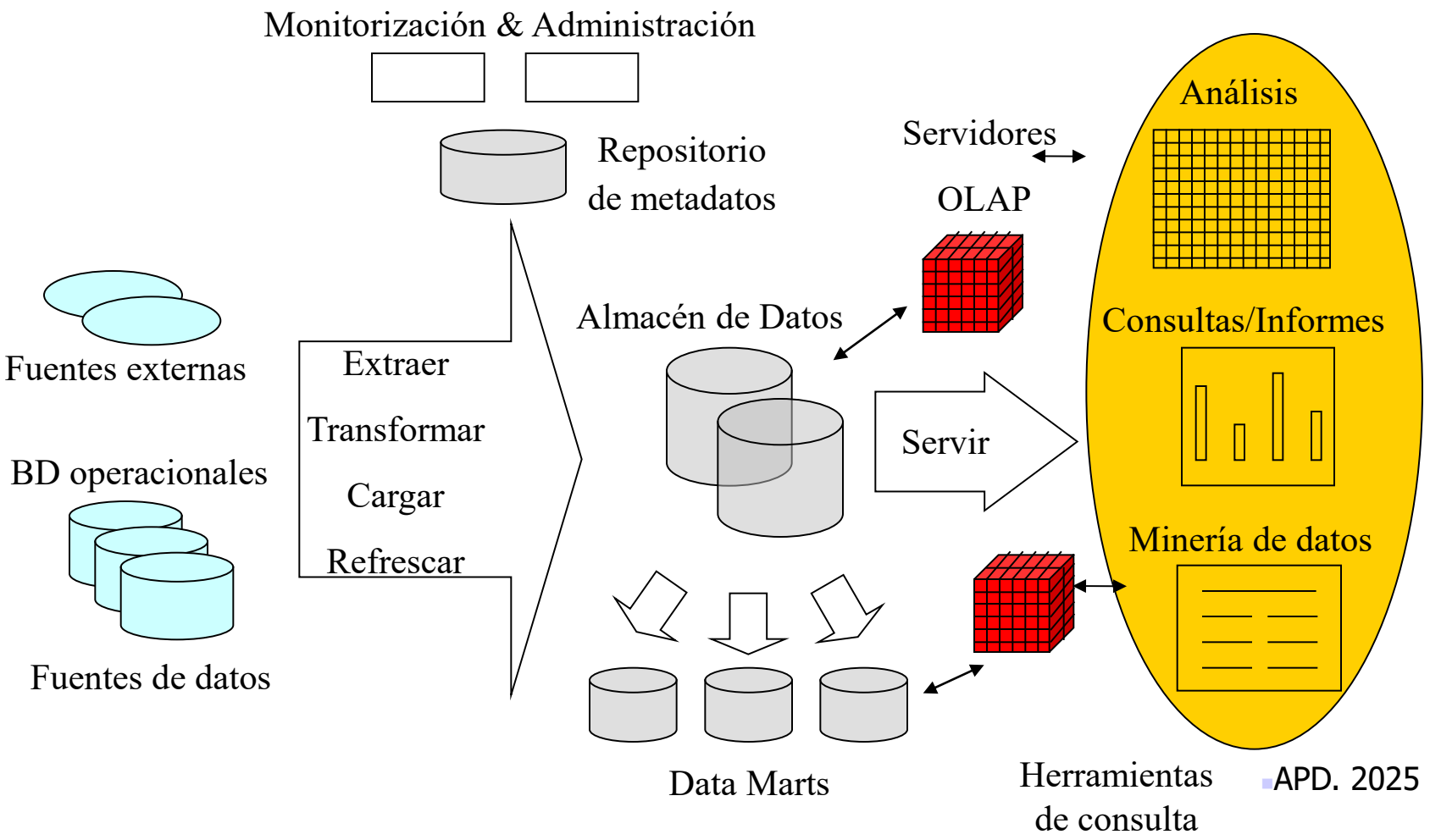
Indice

Arquitecturas de DataWarehouse

- Existen distintas variantes de la arquitectura de un DW según las necesidades del negocio
- Algunas de ellas:
  - Arquitectura tradicional: Diseñada para análisis de grandes cantidades de datos estructurados
  - En tiempo real: Cuando las decisiones requieren de datos tan pronto como se generan
  - Para "Big Data": Cuando la información a tratar incluye información no estructurada (redes sociales!)

Arquitecturas de DataWarehouse

- Esquema general:



Arquitecturas de DataWarehouse

- Arquitectura de tres capas
  - Servidor del repositorio o base de datos del almacén de datos
    - Casi siempre un SGBD Relacional
  - Servidores OLAP
    - Relational OLAP (ROLAP)
      - Extiende SGBD relacionales para permitir operaciones MD
    - Multidimensional OLAP (MOLAP)

      Directamente implementa el modelo MD en vectores

Arquitecturas de DataWarehouse

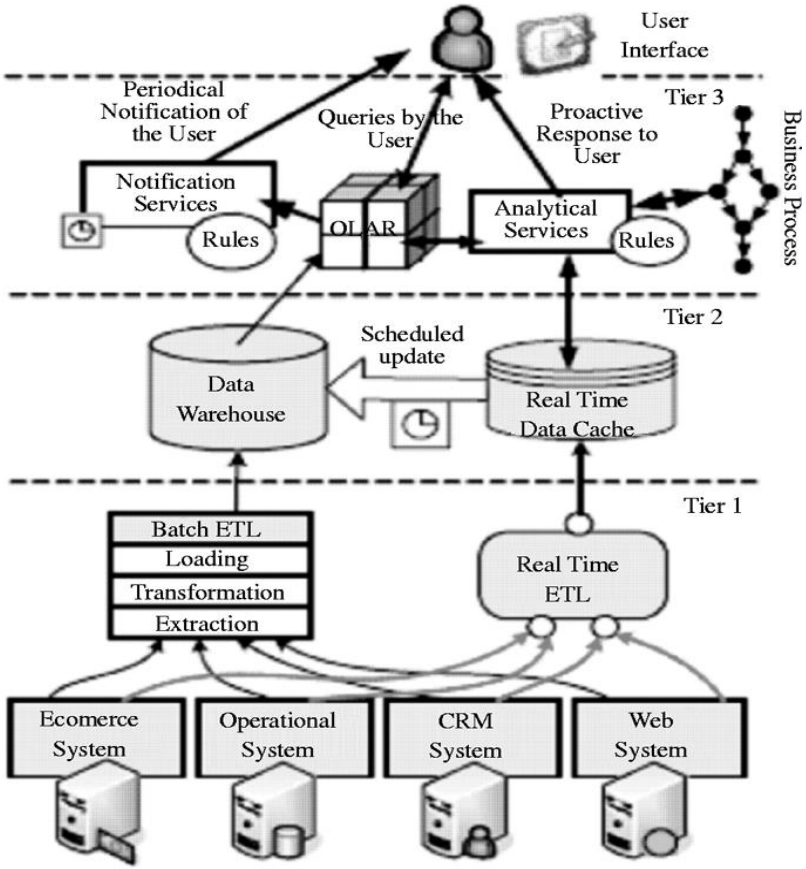
- Arquitectura de tres capas (II)
  - Clientes → Herramientas
    - Informes y consultas
    - OLAP (On-Line Analytical Processing)
    - Data Mining

Arquitecturas de DataWarehouse

- ¿Por qué Data Warehouse separado?
  - Rendimiento (Performance)
    - Consultas OLAP complejas → Ralentización del servidor
    - Métodos de implementación, accesos, etc. distintos
  - Funcionalidad
    - Datos no existentes → históricos
    - Datos consolidados (agregados, sumados, resumidos, etc.) de distintas fuentes
    - Calidad de datos
      - Diferentes fuentes → representaciones distintas, etc…

Arquitecturas de DataWarehouse

- Arquitectura tradicional: más en detalle en la siguiente sección
- En tiempo real:



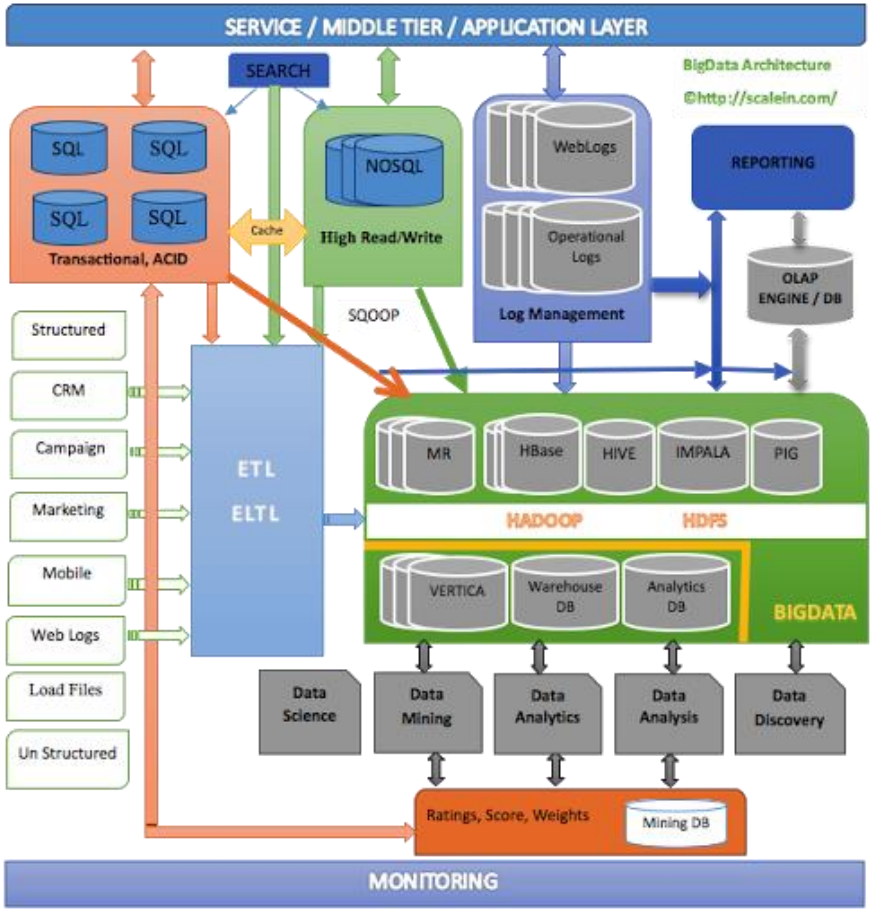
Source: Nguyen Manh *et al.* (2005)

Arquitecturas de DataWarehouse

Arquitectura BigData típica

Características:

- Almacenamiento de diferentes tipos de datos
  - Semi-estructurados (Marketing/ campañas/ móbil/ web logs)
  - Estructurados
  - Ficheros de log
- Carga de datos desde diferentes bases de datos (MySQL, Oracle, PostgreSQL, MongoDB, etc)
- Minería de datos
- Analíticas
- Almacenes de datos para reporting
- Análisis por lotes (Hadoop)
- Web caching
- Search



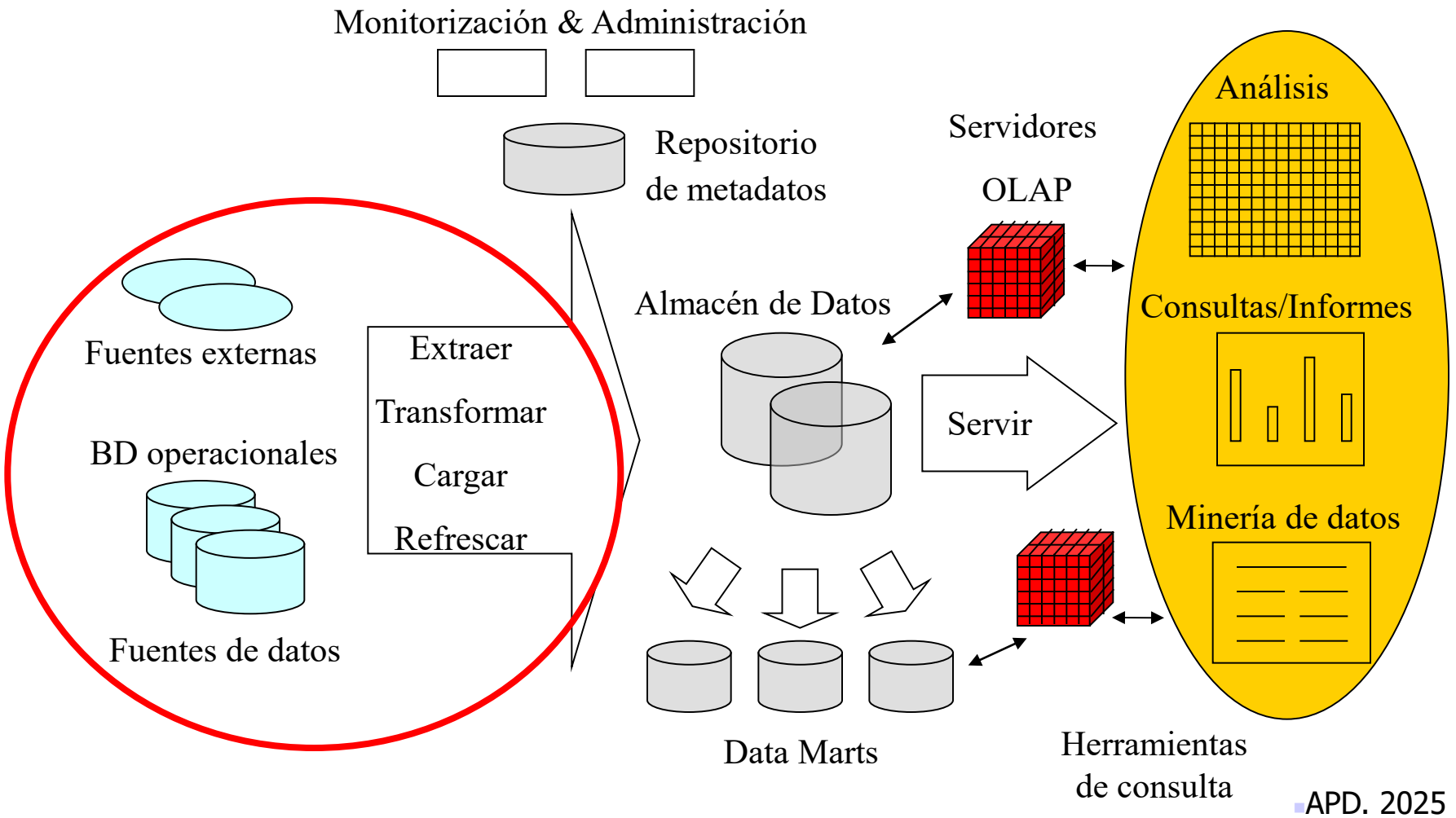
APD. 2025

Imagen via (http://scalein.com/)

Indice

Esquema de una arquitectura de DW

Fuentes de datos



Esquema de una arquitectura de DW

Fuentes de datos

- Fuentes de datos
  - Fuentes de datos operacionales de la empresa
  - Bases de datos externas (públicas o privadas)
  - Ficheros planos
  - Datos en formato tradicional:
    - documentos, facturas, albaranes, etc.
  - Internet → Cantidad ingente de datos

Esquema de una arquitectura de DW

Fuentes de datos

- Procesos para alimentar de datos el almacén (ETL)
  - Extracción (Extraction)
  - Limpieza (Cleaning) y Transformación (Transformation)
  - Carga (Loading)
  - Refresco

- Extracción
  - Procesos que recogen los datos necesarios del almacén

Esquema de una arquitectura de DW

Fuentes de datos

- Limpieza (Cleaning)
  - Fundamental que los datos del almacén sean correctos
    - Decisiones estratégicas
  - Muchas fuentes de datos → alta probabilidad de error y anomalías
    - Longitud de campos inconsistentes
    - Descripción inconsistente (¿¿¿¿¿ Qué es dirección ????)
    - Valores incoherentes ("U. Alicante" vs "UA")
    - Valores nulos

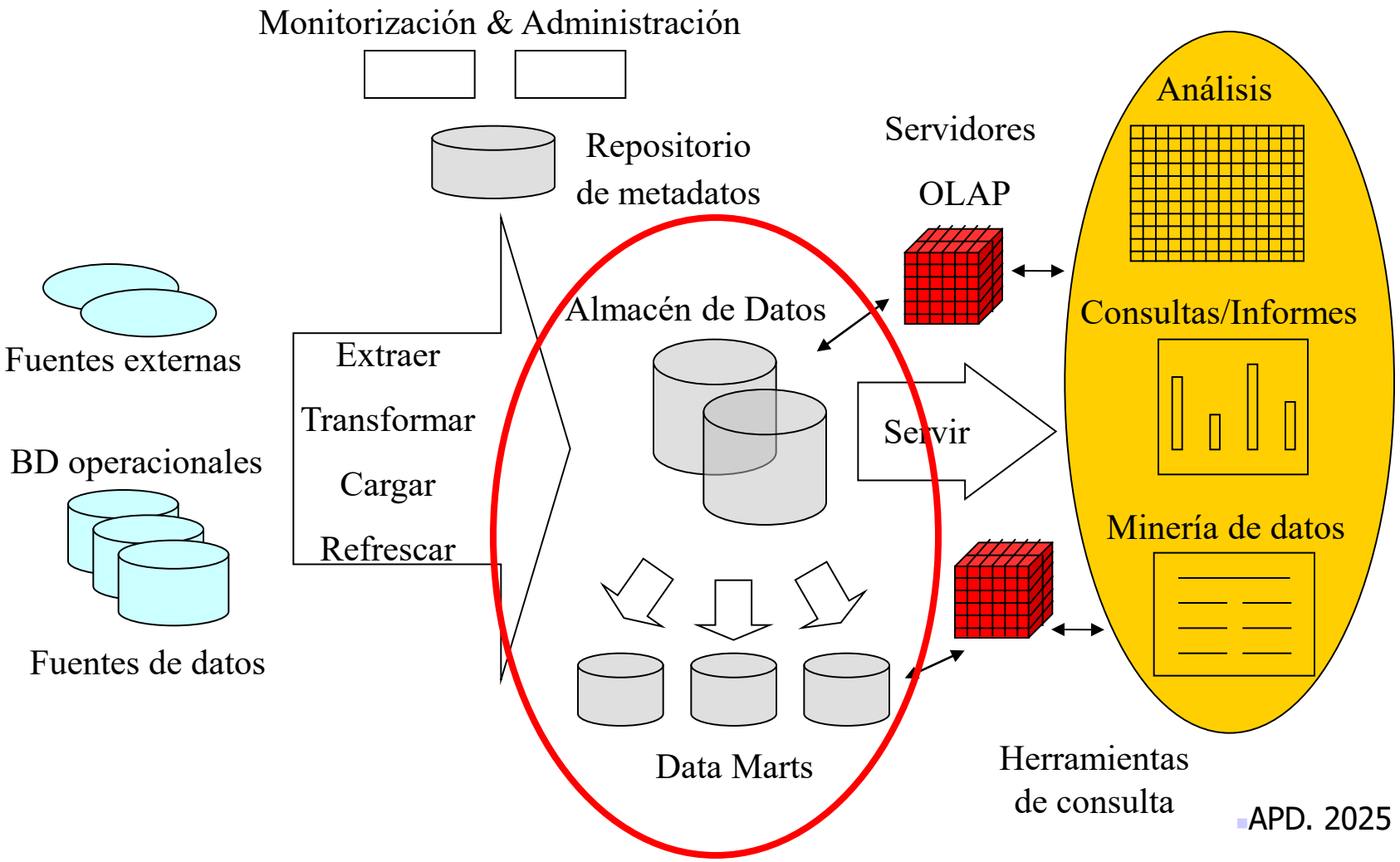
Esquema de una arquitectura de DW

Fuentes de datos

- Carga (Loading)
  - Una vez que los datos se extraen, limpian y transforman → CARGAR
  - Se puede necesitar más pre-proceso antes de carga
    - Comprobar reglas de integridad de nuevo
    - Calcular datos agregados
    - Construir Tablas derivadas y virtuales e intermedias
    - Construir índices
    - Calcular tiempo → Muchas veces por la noche
      - ¿¿ Y si multinacional ??

Esquema de una arquitectura de DW

El almacén o repositorio



Esquema de una arquitectura de DW

El almacén o repositorio

- Datos orientados por tema
  - Los sistemas OLTP están optimizados para las transacciones
    - NORMALIZAR
      - Muchas transacciones con pocos datos
        - Ej. datos de clientes en varias tablas
  - ¿Almacén de datos normalizado?: PROBLEMAS
    - ¿Directivo es capaz de tener visión de todas las tablas y relaciones?
    - Pocas transacciones que incluyen muchos datos
      - Operación MAS cara en BD: unión de tablas

Esquema de una arquitectura de DW

El almacén o repositorio

- Los datos están orientados por tema
  - En un solo lugar (digamos tabla) datos referentes a un concepto que es el objeto de estudio
    - Ej. Tabla para clientes
    - Ventas
    - Compras
    - Vehículos, etc.

Esquema de una arquitectura de DW

El almacén o repositorio

- Integrados
  - Están coherentemente agrupados a partir de datos de las fuentes de datos
  - También hay datos derivados
  - Para ello: procesos de limpieza y transformación
  - Hay errores difíciles de detectar: ¡¡¡¡¡ Cuidado !!!!!
    - Ej. Código producto válido → Exhaustivo análisis de datos

Esquema de una arquitectura de DW

El almacén o repositorio

- Integrados,.....
  - Problemas de incoherencia: resumir en 4 tipos
    - Descripción
      - J. A. Rodríguez vs. Jose A. Rodriguez
    - Codificación
      - Varón “V”, Hembra “H”; en otra BD Varón “H”, Hembra “M”
    - Unidades
      - Estatura: 1,70 mts; 170 cm
    - Formato
      - Número de teléfono como cadena de caracteres (965- 90 34 00) vs.como entero (965903400)

Esquema de una arquitectura de DW

El almacén o repositorio

- No volátiles
  - En sistemas OLTP se pueden modificar datos (ej. tuplas)
    - Unidades de pedido 200; si cliente modifica, se cambia.
  - En DW nunca se modifican (salvo excepciones), se añaden nuevos datos para el análisis
    - Un pedido con una fecha de 200 uds.
    - Un pedido con otra fecha de 150 uds.

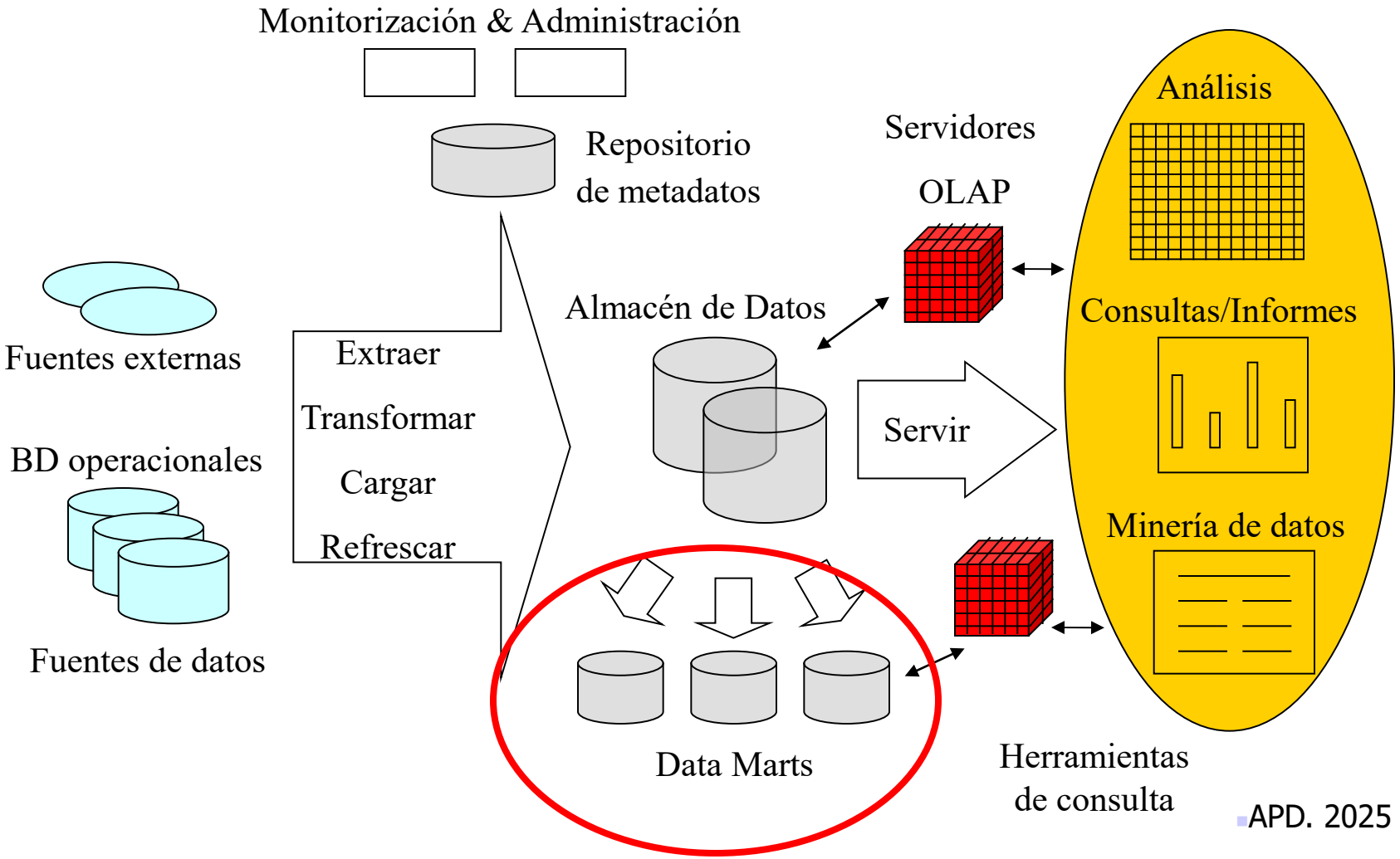
Esquema de una arquitectura de DW

El almacén o repositorio

- Variables en el tiempo
  - No volatibilidad → Dimensión básica: El TIEMPO
    - Datos analizados en función del tiempo
    - Ej. Anterior del pedido
      - ¿ Por qué un cliente ha variado la cantidad de su pedido en una semana ?
      - ¿ Por qué han cambiado los gustos de un cliente en varios años ?
      - Etc.

Esquema de una arquitectura de DW

El almacén o repositorio



Esquema de una arquitectura de DW

El almacén o repositorio

- Data Marts
  - Es como una vista del almacén de datos
  - Se definen para satisfacer las necesidades de un departamento o sección dentro de una empresa
  - Normalmente, en la práctica, suelen contener más cantidad de información agrupada que en detalle

Esquema de una arquitectura de DW

El almacén o repositorio

- Data Marts.....
  - Para su construcción se pueden seguir dos aproximaciones:
    - (I) Definir primero el almacén de datos y, a partir de él, definir los data marts, ó
    - (II) Definir primero los data marts departamentales y, posteriormente integrarlos en un almacén de datos global para la organización
  - Nota: Si la envergadura de la empresa es considerable y, la experiencia en construir DW poca, es aconsejable seguir la aproximación 2

Esquema de una arquitectura de DW

El almacén o repositorio

- Resumen: Data warehouse vs. Data Marts
  - Enterprise DW
    - Información sobre "temas" de toda organización
    - Requiere complejo modelado de negocio
    - Puede llevar AÑOS para construir e implementar
  - Data Mart
    - Departamental → sub-temas
      - Ej. Marketing data mart, Clientes, productos, ventas !!!!!
    - Más rápido agregar
    - OJO !!! Integración con DW puede ser compleja

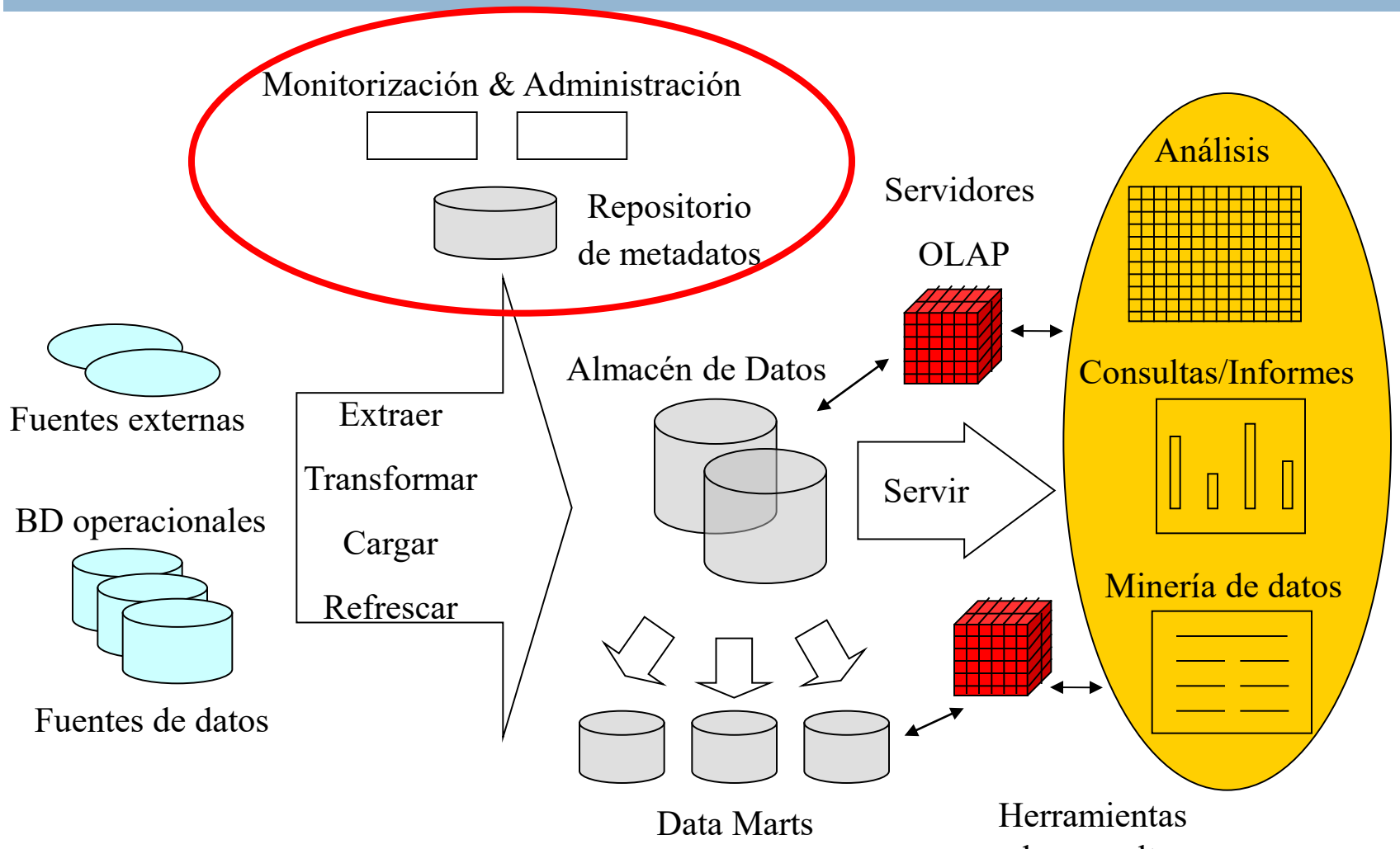
Esquema de una arquitectura de DW

El almacén o repositorio

- Virtual Data Warehouse
  - Vistas sobre bases de datos operacionales
    - Materializan algunas vistas agregadas → consultas eficientes
    - Mayor facilidad en la construcción
    - Requieren exceso de capacidad del servidor operacional
    - Muchas empresas dicen tener DW cuando el administrador ha creado un VIRTUAL DW

Esquema de una arquitectura de DW

Los metadatos



Esquema de una arquitectura de DW

Los metadatos

- Los metadatos
  - Son datos sobre datos
    - Qué dato se guarda (ej. clientes)
    - Dónde se guarda (tabla clientes)
    - Campos de la tabla
    - Con qué datos de las fuentes se corresponden
    - Niveles de agregación
    - Procesos de carga → ¿ Cuándo se actualizan ?
      - ¿ Cuándo fue la última actualización?
    - Patrón de dato válido (Ej. *Apellido1 Apellido2, Nombre*)

Esquema de una arquitectura de DW

Los metadatos

- Los metadatos …
  - Son datos sobre datos …
    - Reglas de transformación
    - ¿ Cuándo se incorporan al almacén de datos ?
    - Y muchos más…

Esquema de una arquitectura de DW

Los metadatos

- Tipos
  - Administrative metadata (Toda la información necesaria para el DW)
    - Fuentes de datos y contenidos
    - Esquema del data warehouse, vistas y datos agregados
    - Dimensiones de análisis con sus jerarquías
    - Consultas e informes predefinidos
    - Localización y contenido del los Data Marts
    - Diseño físico → particionamiento de datos

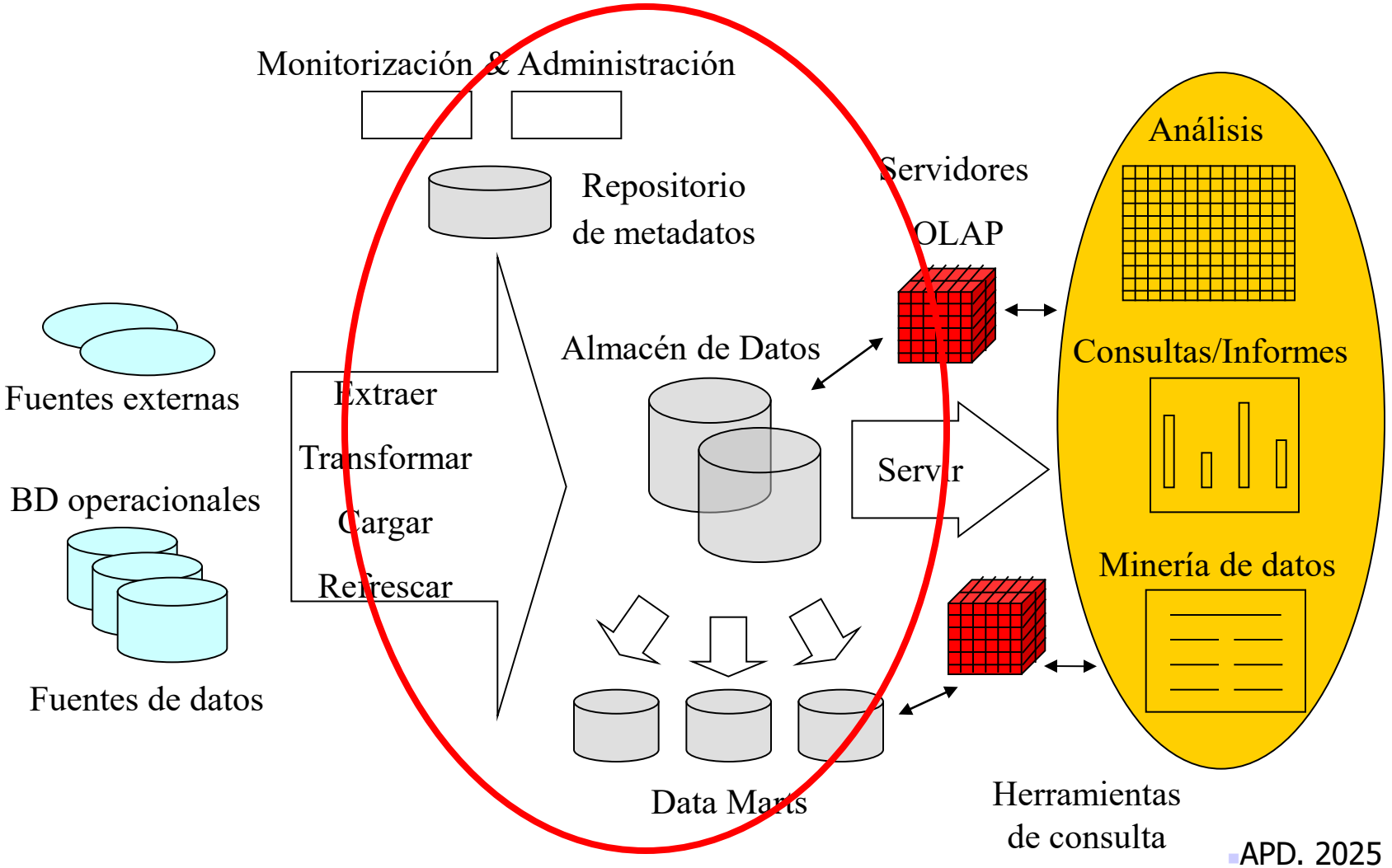
Esquema de una arquitectura de DW

Los metadatos

- Tipos…
  - Business metadata
    - Información y términos de negocio
    - Políticas de posesión de datos
    - Políticas de permiso de datos por usuarios (seguridad)
  - Operational metadata
    - Obtiene información recogida durante el proceso del “almacén de datos”
      - Datos migrados y secuencia de transformaciones aplicadas
      - Auditoría
      - Informes de error

Esquema de una arquitectura de DW

Servidor del almacén de datos



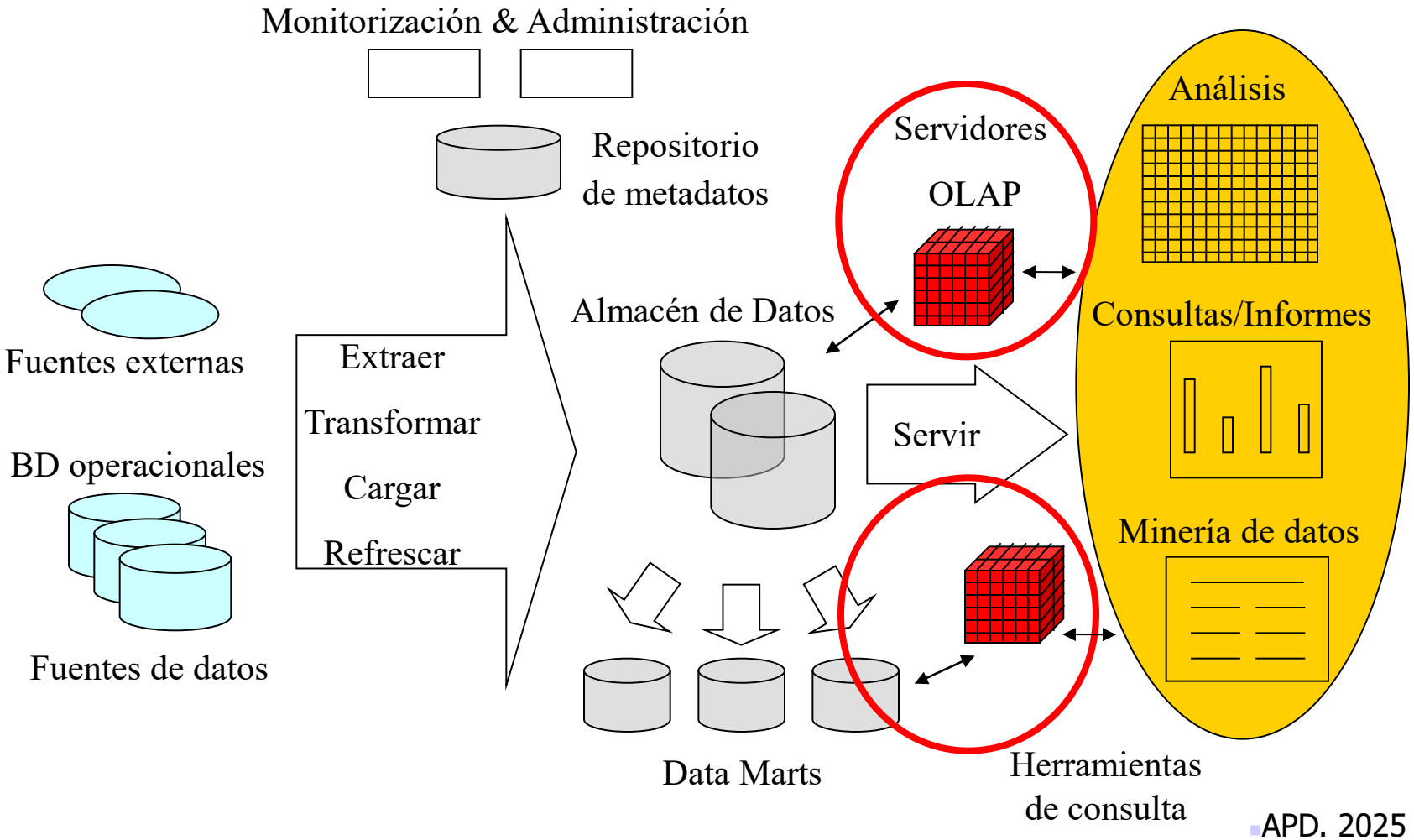
Esquema de una arquitectura de DW

Servidor del almacén de datos

- El servidor es un SGBD que se encarga de
  - Gestionar el repositorio propio del almacén de datos
  - Coordinar los procesos ETL que alimentan el DW
  - Procesan las consultan lanzadas sobre el almacén y devuelven los datos
  - Generalmente son servidores relacionales

Esquema de una arquitectura de DW

Servidor del almacén de datos



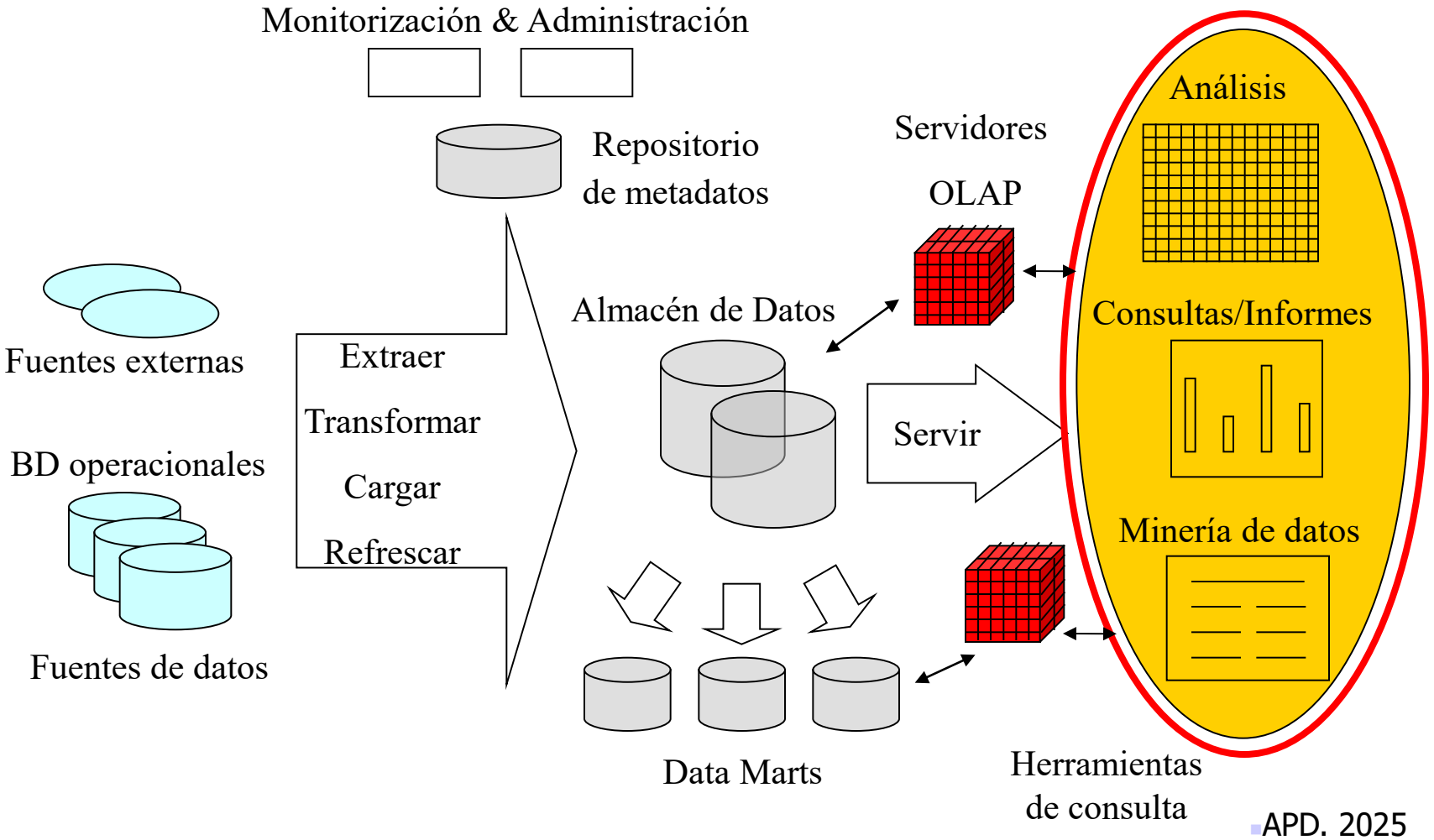
Esquema de una arquitectura de DW

Servidor de consultas

- El servidor de consultas
  - En la mayoría de las arquitecturas se utiliza un servidor distinto al del almacén de datos
    - Rendimiento y mantenimiento
  - La mayoría de herramientas funcionan con esta arquitectura
    - Ejplo. MicroStrategy
  - Dos tecnologías ampliamente utilizadas
    - ROLAP
    - MOLAP

Esquema de una arquitectura de DW

Herramientas de consultas



Esquema de una arquitectura de DW

Herramientas de consultas

- Generación de informes
  - Consultas ad-hoc e informes
    - Permiten acceso a los datos base
    - Informe se construye con “point-and-click”
  - Entornos consulta/informes de directivos
    - Muestra datos base en términos de negocio
    - Utiliza consultas predefinidas/almacenadas
    - Soporte limitado para consultas ad-hoc
    - Informes se pueden presentar como documentos

Esquema de una arquitectura de DW

Herramientas de consultas

- Data Mining (Minería de datos)
  - Descubre tendencias y patrones (minería interpretativa)
  - Crean modelos y hacen predicciones (minería predictiva)
    - Utilizan los datos existentes para detectar tendencias y crear modelos
    - Aplican los patrones y modelos a datos nuevos
    - Importante → Adaptación de modelos según nuevos datos

El almacén de datos

Bibliografía

- Juan Trujillo, Jose Norberto Mazón, Jesús Pardillo. Diseño y explotación de almacenes de datos: Conceptos Básicos de Modelado Multidimensional. Ed. Club Universitario. 2011.
  - Tema 1
- Inmon (2005). Building the Data Warehouse (4ª ed.)
  - Tema 1 (introducción)
- Kimball & Ross (2013). The Data Warehouse Toolkit (3ª ed.)
  - Tema 1 (introducción)
- Thomsen (2000). OLAP solutions: Building Multidimensional Information Systems
  - Tema 1 (introducción)
- Giovinnazo (2000). Object-Oriented Data Warehouse Design: Building a star schema
  - Tema 1 (introducción)
- Barrera, J. M., Reina-Reina, A., Lavalle, A., Maté, A., & Trujillo, J. (2024). An extension of iStar for Machine Learning requirements by following the PRISE methodology. *Computer Standards & Interfaces*, *88*, 103806.
- Lavalle, A., Reina-Reina, A., Maté, A., Trujillo, J. (2024). Social Modeling Using the i* Framework.
  - Capitulo: Data Analytics from a Social Perspective.

Introducción al concepto de los datos y su análisis

Tema 1

Profesores:

Juan C. Trujillo
Alejandro Reina Reina
LUCENTIA Research Group

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos