Data Management Fundamentals
Módulo 2 – Big Data & Data Science

GOLD SPONSORS

Capgemini · NTT DATA · denodo · MERCANZA · World · Qlik · Collibra

SILVER SPONSORS

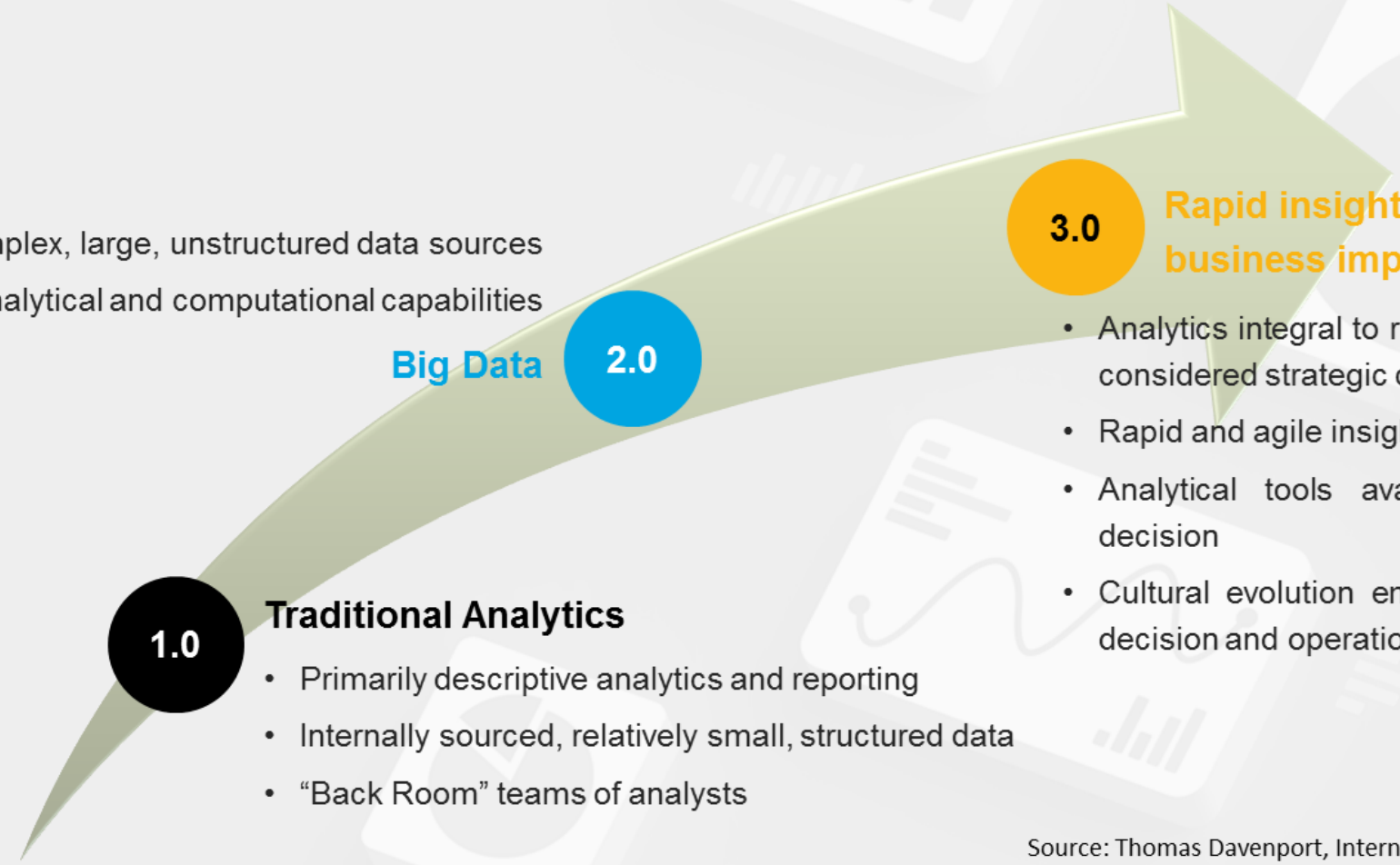
MINSAIT · rebAI · VARGROUP · Quest · bmc

Entidades académicas

UCLM Universidad de Castilla-La Mancha · MONDRAGON UNIBERTSITATEA

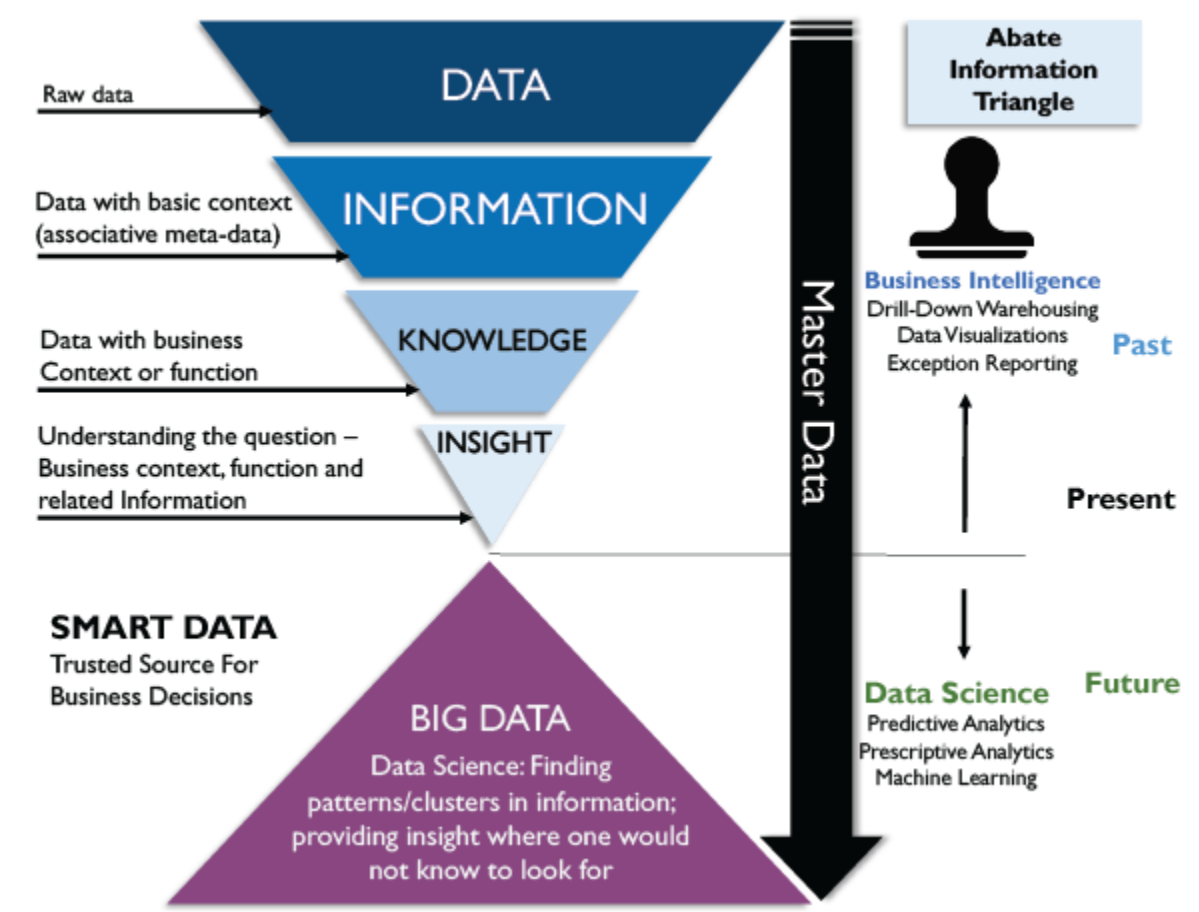
DAMA SPAIN

Evolución del análisis de los datos



Source: Thomas Davenport, International Institute for Analytics

¿Por qué necesitamos Big Data y Data Science?



Entre otras:

- Velocidad de proceso (Streaming)
- Complejidad de los datos (no estructurados)
- Escalabilidad
- Volumen de datos

¿Qué es Big Data?

No existe una definición consensuada, sino que se basa en las V's de Big Data.
Tratamiento y acumulación de datos de todo tipo (estructurados o no) de los que se desea extraer valor.

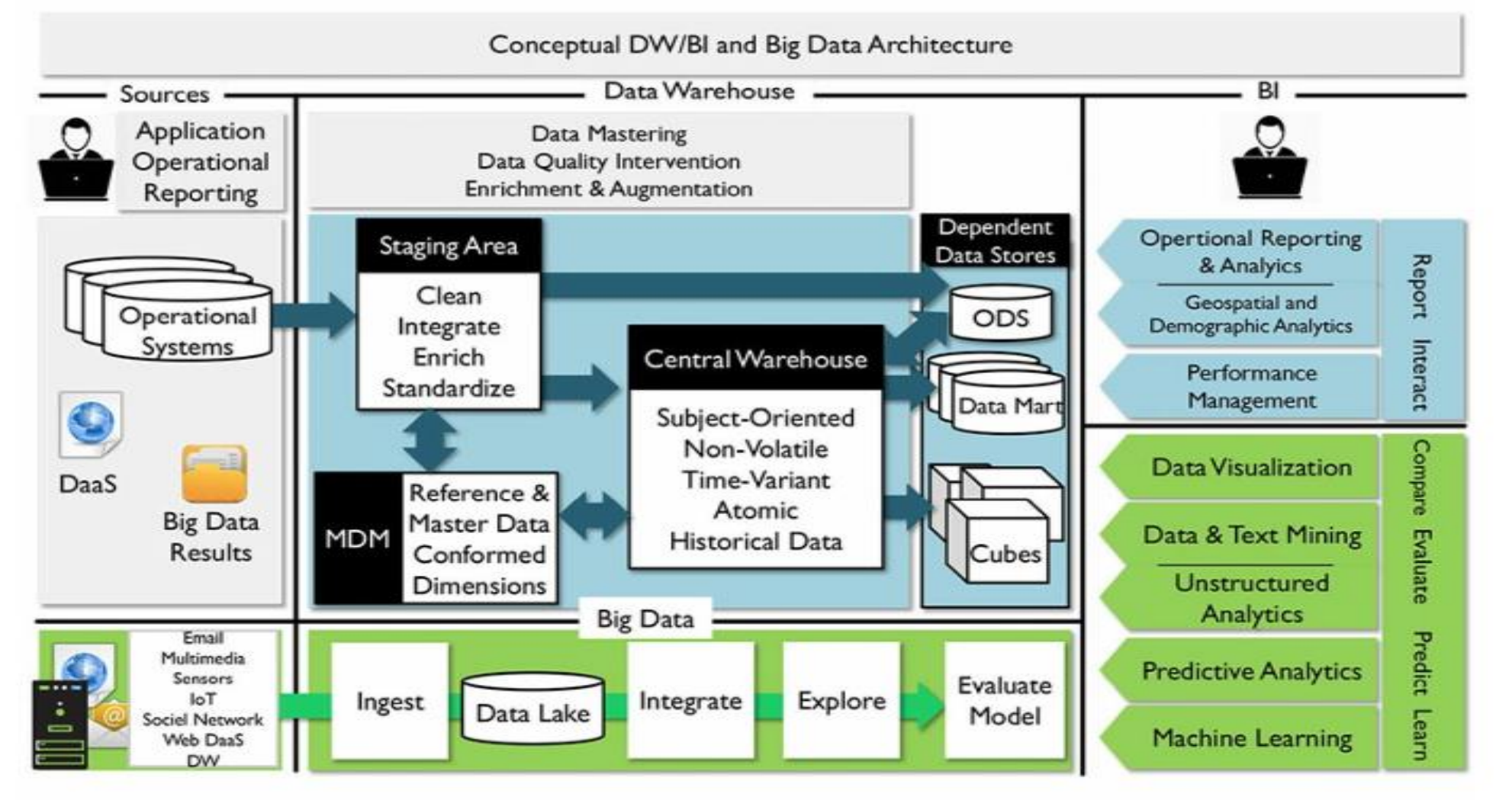
Las 3 V's (Doug Laney, 2001)

- **Volume**, cantidad de información
- **Velocity**, velocidad de captura
- **Variety (Variability)**, diversidad de formatos y formas de captura y entrega

Las nuevas V's

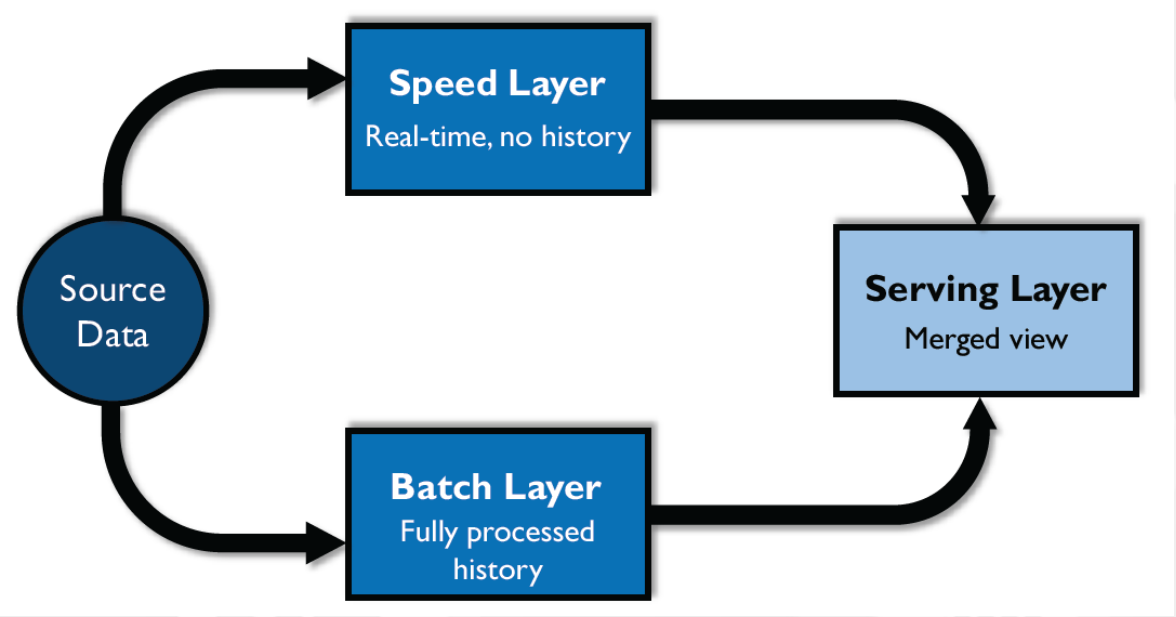
- **Viscosity**, dificultad de uso o integración de los datos
- **Volatility**, frecuencia de cambio de los datos, cuánto tiempo son útiles
- **Veracity**, cómo de fiables son los datos

Componentes de la arquitectura de Big Data



Data Lake

Entorno en el que una gran cantidad de datos de diferentes tipos y estructuras son ingestados, almacenados, evaluados y analizados.
Para procesar los datos se basa en proceso ELT en los que la transformación se ejecuta a demanda.



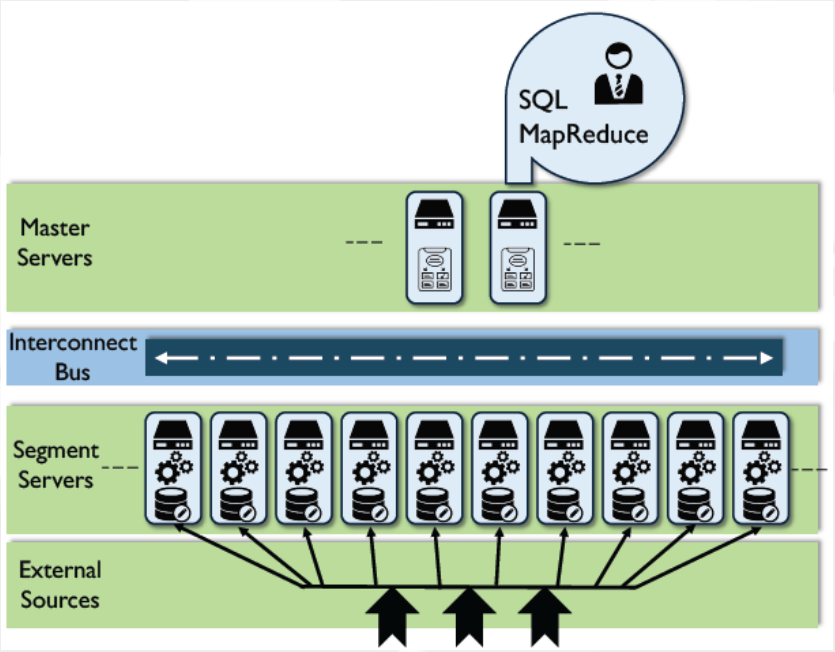
Arquitectura orientada a servicios, también se la conoce como arquitectura LAMBDA

Procesamiento masivamente paralelo (MPP)

- Los datos se distribuyen lógicamente en **múltiples servidores de procesamiento**
- **El procesamiento se distribuye** en el conjunto de servidores que trabajan con conjuntos más pequeños de datos
- **Escalabilidad lineal** a medida que se añaden nuevos servidores

Map Reduce, sistema de procesamiento de grandes conjuntos de datos. Implementado en Hadoop. Consta de 3 pasos principales que pueden ejecutarse tanto en secuencia como en paralelo:

- **Map**: Identificación y obtención de los datos
- **Shuffle**: Combinación de datos en función del patrón analítico
- **Reduce**: Eliminación de duplicados o ejecución de agregados



¿Qué es Data Science (ciencia de datos)?

La ciencia de datos **combina la minería de datos, el análisis estadístico y el aprendizaje automático** con la integración de datos y las capacidades de modelado de datos, para construir modelos de predicción que exploran los patrones de contenido de los datos.

El desarrollo de modelos predictivos a veces se denomina ciencia de datos porque el analista de datos, o científico de datos, utiliza el método científico para desarrollar y evaluar un modelo.

Dependencias de Data Science

El desarrollo de soluciones de Data Science implica la inclusión iterativa de fuentes de datos en modelos que generan insights. La ciencia de datos depende de:

- **Fuentes de datos ricas:** Datos con el potencial de mostrar patrones invisibles en el comportamiento de la organización o del cliente
- **Alineación y análisis de la información:** Técnicas para comprender el contenido de los datos y combinar conjuntos de datos para formular hipótesis y probar patrones significativos
- **Entrega de información:** Ejecutando modelos y algoritmos matemáticos contra los datos y produciendo visualizaciones y otros resultados para obtener una comprensión del comportamiento
- **Presentación de las conclusiones y de los datos:** Análisis y presentación de las conclusiones para que se puedan compartir los insights

El proceso de Data Science

DAMA

CRISP-DM

Ciclo de desarrollo de Data Science (DAMA) I

Se sigue el **método científico** de refinar el conocimiento realizando observaciones, formulando y comprobando hipótesis, revisando los resultados y formulando teorías generales que los expliquen.

1. **Definir la estrategia de Big Data y las necesidades de negocio:** Definir los requisitos que identifican los resultados deseados con beneficios tangibles medibles

2. **Elegir las fuentes de datos:** Identificar las carencias en la base de activos de datos actual y encontrar fuentes de datos para eliminarlas

3. **Adquirir e ingerir fuentes de datos:** Obtener conjuntos de datos e incorporarlos

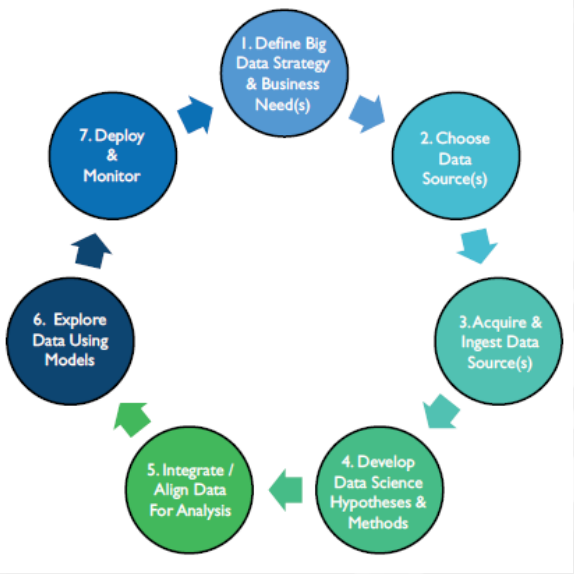
4. **Desarrollar hipótesis y métodos de ciencia de datos:** Explorar las fuentes de datos mediante perfilado, visualización, minería, etc.; perfeccionar los requisitos. Definir las entradas del algoritmo, hipótesis del algoritmo o los métodos de análisis

Ciclo de desarrollo de Data Science (DAMA) II

5. **Integrar y alinear los datos para el análisis:** La viabilidad del modelo depende de la calidad de los datos de la fuente. Aprovechar fuentes fiables y creíbles para reducir los trabajos de limpieza e integración

6. **Explorar los datos utilizando modelos:** Aplicar análisis estadísticos y algoritmos de machine learning contra los datos integrados. Validar, entrenar y, con el tiempo, hacer evolucionar el modelo. El entrenamiento implica ejecutar repetidamente el modelo contra los datos reales para verificar las suposiciones y hacer ajustes, como identificar los valores atípicos. Se pueden introducir nuevas hipótesis que requieran de nuevos conjuntos de datos, cambiar el modelo, las salidas o incluso los requerimientos

7. **Despliegue y monitorización:** Los modelos que producen información útil pueden desplegarse en producción manteniendo una vigilancia continua de su valor y eficacia



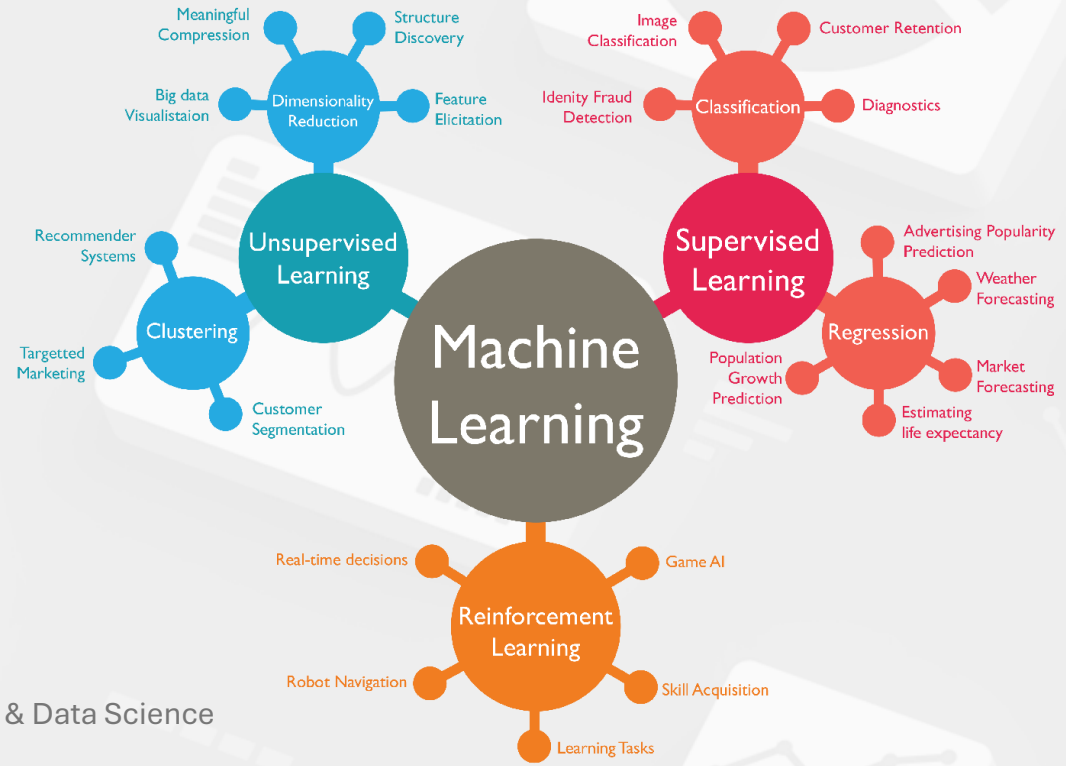
Tipos de algoritmo de Machine Learning

Tipos de algoritmos de **Machine learning:**

- **Aprendizaje supervisado**: se entrena con datos previos etiquetados
- **Aprendizaje no supervisado**: busca patrones no conocidos en los datos
- **Aprendizaje por refuerzo**: en cada iteración de entrenamiento, se recompensa al agente por éxitos y castiga por fracasos

Usos:

- Análisis de sentimiento
- Minado de datos
- Análisis predictivo
- Análisis prescriptivo
- Análisis de datos no estructurados
- Análisis operacional
- Visualización de datos
- Combinación de datos



Matriz de confusión

Ayuda a evaluar modelos supervisados.

En función de la aplicación del algoritmo se define la criticidad de cada tipo de error en la clasificación.

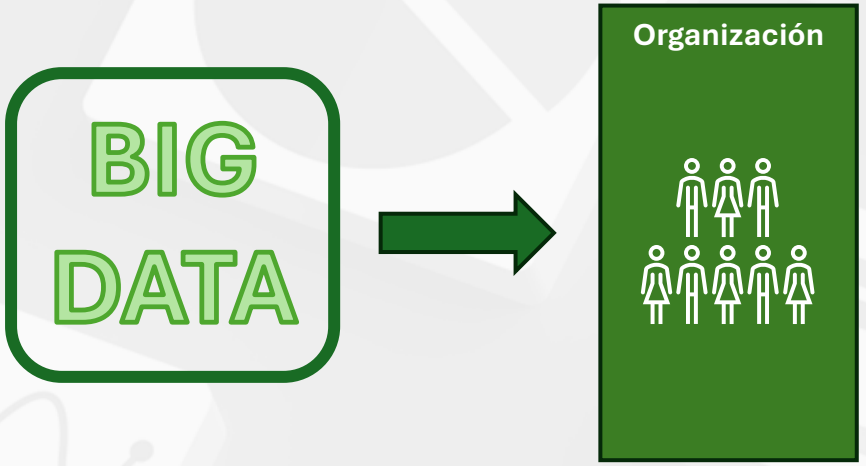
		Valor predicho	
		Positivo	**Negativo**
Valor real	**Positivo**	Positivo verdadero	Falso negativo
	Negativo	Falso positivo	Negativo verdadero

Guía de implementación

1. **Alineamiento con los objetivos estratégicos de la organización**, además de trabajar conjuntamente con iniciativas de datos corporativas tales como gestión de metadatos o calidad de los datos

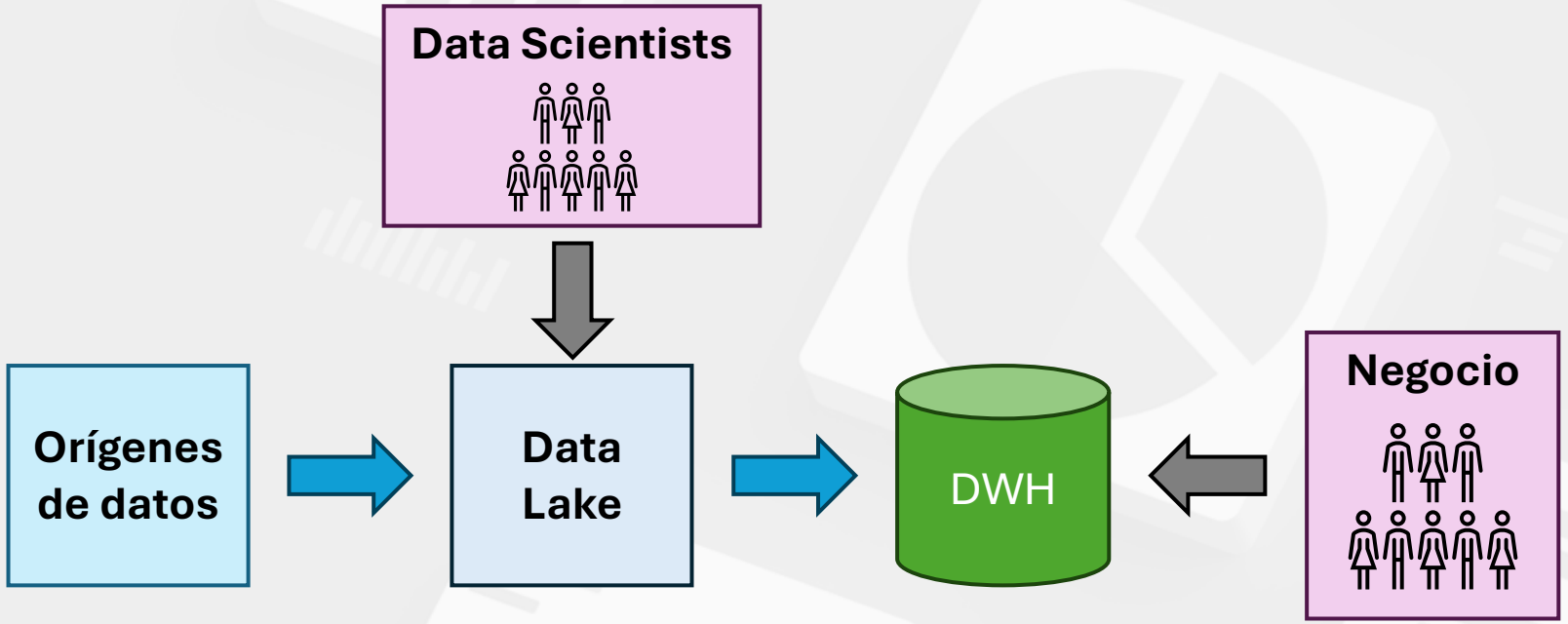
2. **Evaluación de preparación y riesgos**, comprobar la relevancia del Big Data para el negocio confirmando el alineamiento con sus necesidades y entendiendo el grado de madurez en el uso de los datos del mismo. Validar la viabilidad económica, así como la disponibilidad de presupuesto

3. **Cambio cultural y organizacional**, se debe lanzar un programa de culturización para que la organización pueda sacar todo el partido posible al Big Data. Puede ser buena idea crear un centro de excelencia. Van a ser necesarios nuevos perfiles en la organización tales como arquitectos o data scientist



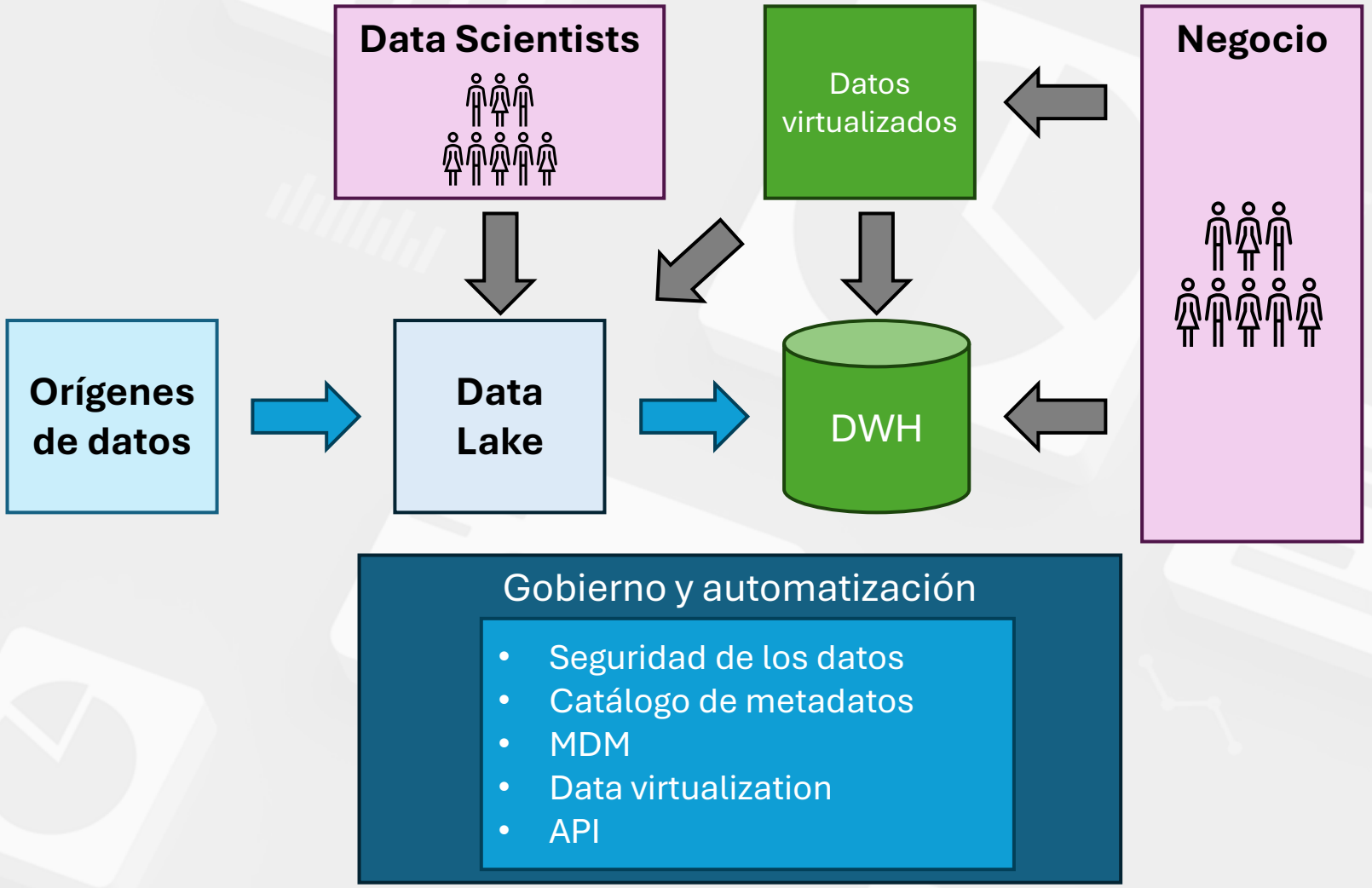
Modern Datawarehouse

- Se habilitan capacidades analíticas en el Data Lake
- El DWH se carga con datos del Data Lake
- Se enriquece el DWH con contenido procesado en tiempo casi real o proveniente de datos no estructurados
- El negocio sigue trabajando con el DWH



Data Fabric

- Es una evolución del Modern DWH
- Se añaden servicios de gobierno y automatización para toda la plataforma
- Se securiza el acceso a los datos
- Incorpora el uso del MDM
- Se crean servicios de automatización basados en metadatos
- Se ofrece acceso a los datos mediante APIs y virtualización

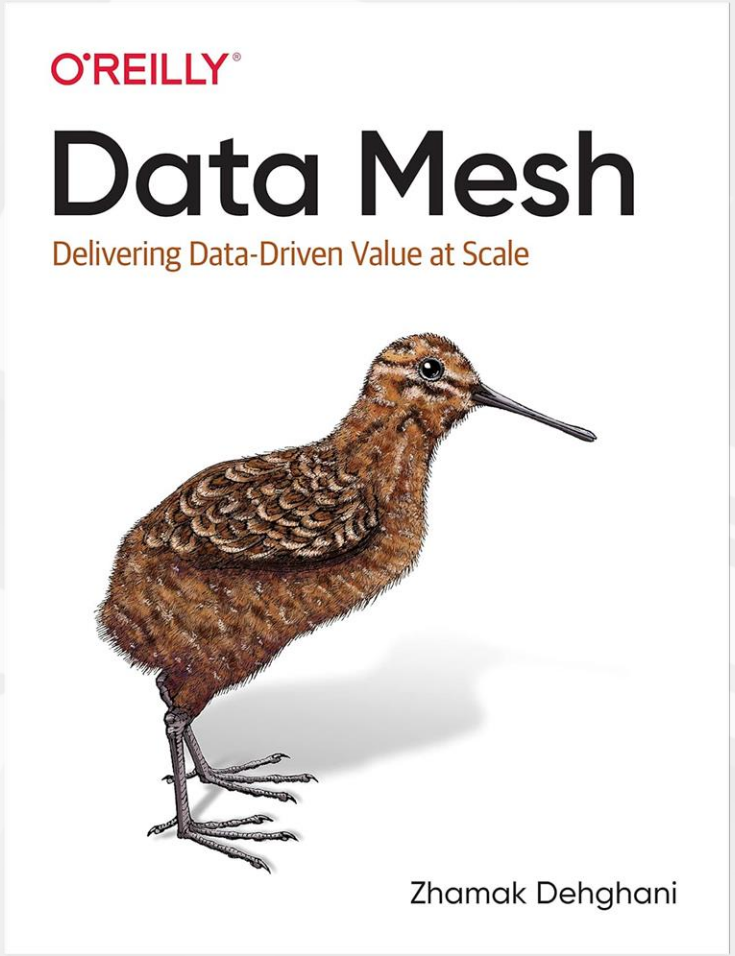


Data Mesh

- **Arquitectura descentralizada** basada en 4 principios:
  - **Equipos de dominio** independientes dueños de sus propios datos
  - Enfoque total hacia **productos de datos**
  - **Aprovisionamiento automatizado** de infraestructura
  - **Gobierno federado** con seguridad y reglas globales

- Requiere una **madurez y una cultura** específicas

- Propone el uso de **desarrolladores generalistas**, requiere de una capa de **arquitectura** muy potente

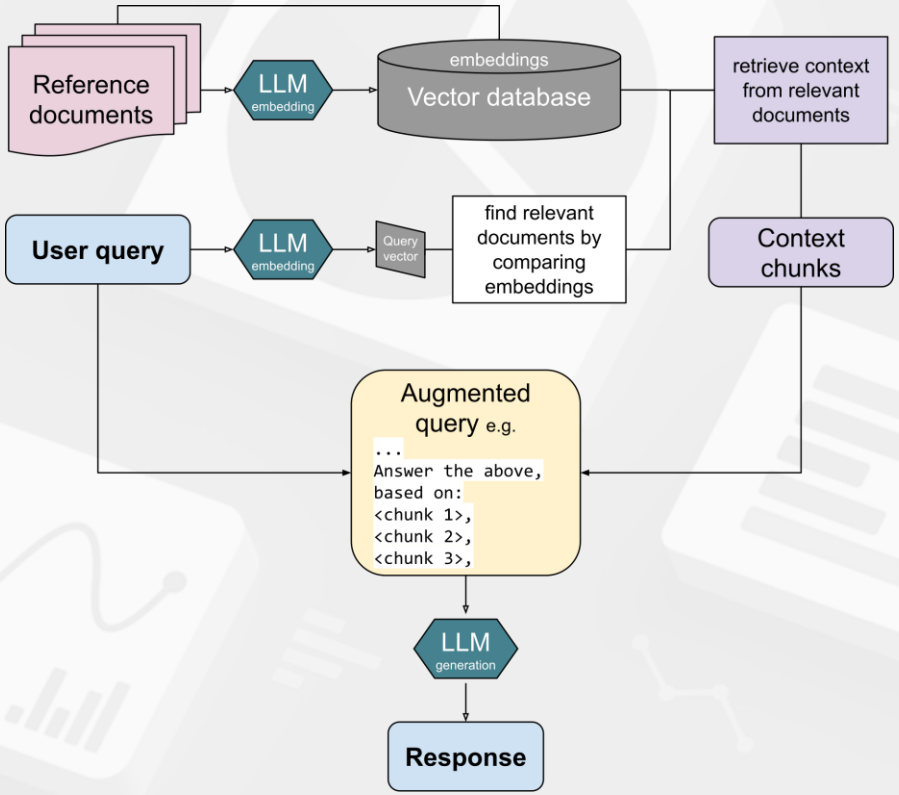


Modelos LLM y RAG

- Un LLM (Large Language Model) está **limitado por los datos** con los que ha sido entrenado

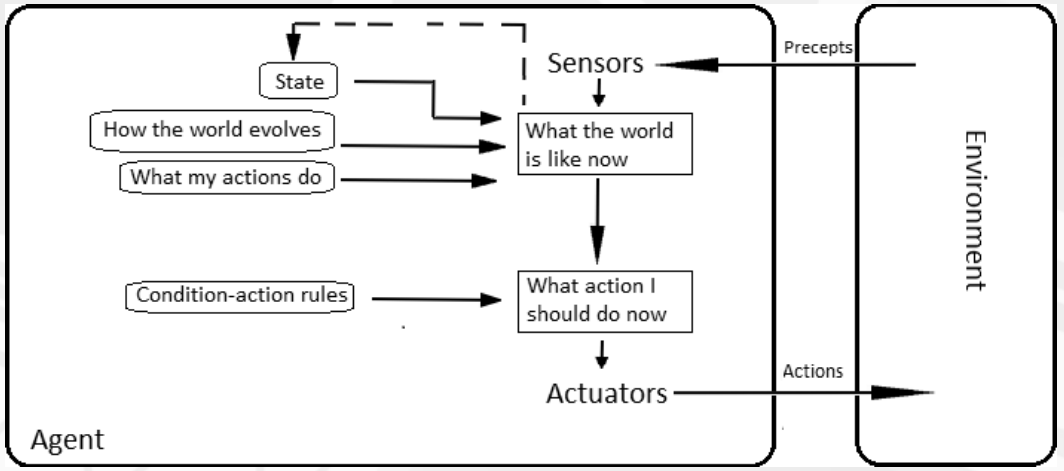
- Mediante un RAG (Retrieval-Augmented Generation) se puede **añadir contexto relevante** para la consulta

- El contexto puede estar indexado previamente o incluso localizarse mediante una búsqueda en Internet



Agentes IA

- Perciben el entorno y reaccionan a él
- Toman decisiones en base a un modelo
- Interactúan con el entorno tratando de llegar a su objetivo o tarea



¡Gracias!

GOLD SPONSORS

Capgemini
NTT DATA
denodo
MERCANZA
World
Qlik
Collibra

SILVER SPONSORS

MINSAIT
rebAI
VARGROUP
Quest
bmc

Entidades académicas

UCLM Universidad de Castilla-La Mancha
MONDRAGON UNIBERTSITATEA
DAMA SPAIN