

Maestría en Análisis y Visualización de Datos Masivos

---

# Ciencia de Datos Aplicada

Ciencia de Datos Aplicada

---

# Tema 1. La ciencia del dato y los datos masivos

# Índice

[Esquema](#)

[Ideas clave](#)

[1.1. Generalidades](#)

[1.2. La Cadena de Valor](#)

[1.3. ¿Qué son los datos masivos?](#)

[1.4. ¿Qué es la Ciencia de Datos?](#)

[1.5. La toma de decisiones basadas en datos](#)

[1.6. Problemas empresariales y soluciones de ciencia de datos](#)

[1.7. Referencias bibliográficas](#)

[A fondo](#)

[El ciclo de vida del científico](#)

[Los datos son el nuevo petróleo – Y eso es bueno](#)

[Ciencia de datos: ¿la profesión más sexy del siglo 21?](#)

[Test](#)



## 1.1. Generalidades

La ciencia de datos y los datos masivos son disciplinas que se encuentran en constante crecimiento y evolución en el ámbito empresarial. Su importancia radica en su capacidad para extraer conocimiento y generar valor a partir de grandes volúmenes de datos, permitiendo así tomar decisiones estratégicas fundamentadas en información precisa y relevante.

En este sentido, es fundamental comprender la cadena de valor que se establece desde la recopilación de los datos, pasando por su limpieza y procesamiento, hasta llegar a su análisis y visualización. Este proceso, conocido como "*pipeline de datos*", implica la aplicación de diversas técnicas y herramientas para asegurar la calidad y la integridad de la información.

Asimismo, es relevante conocer los conceptos clave de los datos masivos y la ciencia de datos para comprender su impacto en las empresas y su potencial para resolver problemas empresariales complejos. Los datos masivos, también conocidos como "*big data*", se refieren a conjuntos de información de gran tamaño, velocidad y variedad, que no pueden ser fácilmente gestionados con herramientas tradicionales. Por otro lado, la ciencia de datos se basa en la combinación de disciplinas como la estadística, la informática y el aprendizaje automático, con el objetivo de extraer conocimiento y patrones significativos de los datos para tomar decisiones informadas.

El análisis de datos a gran escala tiene una amplia gama de aplicaciones en el mundo empresarial. Por ejemplo, puede utilizarse para identificar oportunidades de mercado, predecir cambios en la demanda de productos o servicios, optimizar procesos y mejorar la eficiencia operativa. Además, la ciencia de datos y los datos masivos son fundamentales para los negocios en la era digital, donde la generación constante de información requiere de técnicas avanzadas de análisis para convertirla en ventaja competitiva.

A continuación, se definen los objetivos principales de este tema.

- ▶ Comprender la transformación de datos en valor agregado para las empresas.
- ▶ Definir los datos masivos y su relevancia en el entorno empresarial.
- ▶ Explicar el papel y los métodos de la ciencia de datos en la toma de decisiones.
- ▶ Describir cómo la ciencia de datos ayuda a resolver problemas empresariales complejos.

## 1.2. La Cadena de Valor

La cadena de valor en la ciencia de datos y los datos masivos comprende el proceso de transformación de los datos en valor agregado para las empresas. Esta cadena se inicia con la recopilación de los datos, seguida de su procesamiento y análisis para convertirlos en información útil. A partir de esa información, se genera conocimiento que permite la toma de decisiones basada en datos. Por último, la acción basada en esas decisiones permite crear valor para la organización. Cada etapa de esta cadena es fundamental y requiere de herramientas y técnicas específicas para maximizar el potencial de los datos en beneficio de la empresa.

La cadena de valor en la ciencia de datos y los datos masivos es un proceso fundamental que implica la transformación de los datos en valor agregado para las empresas. Este proceso comienza con la recopilación de los datos, seguida de su procesamiento y análisis para convertirlos en información útil y valiosa.

A partir de esa información, se genera conocimiento que permite la toma de decisiones basada en datos sólidos y confiables. Por último, la acción basada en esas decisiones permite crear valor tangible y cuantificable para la organización, brindando ventajas competitivas significativas en el mercado.

Cada etapa de esta cadena de valor tiene su importancia y requiere de herramientas y técnicas específicas para maximizar el potencial de los datos en beneficio de la empresa. En la fase de recopilación de datos, es crucial utilizar fuentes confiables y métodos eficientes para garantizar la calidad y veracidad de la información recopilada. En el procesamiento y análisis de datos, se requiere el uso de algoritmos y técnicas avanzadas para identificar patrones, tendencias y relaciones entre los datos, lo que permite obtener información valiosa y perspicaz.

A partir de esta información valiosa, se genera conocimiento que proporciona una base sólida para la toma de decisiones basada en datos. El conocimiento generado a través del análisis de datos permite a las empresas comprender mejor su entorno, anticipar tendencias y tomar decisiones informadas y estratégicas. Esta toma de decisiones basada en datos sólidos puede resultar en un mejor rendimiento empresarial y una ventaja competitiva significativa en el mercado.

Por último, la acción basada en estas decisiones permite a las empresas crear valor tangible y cuantificable para la organización. Esto se logra mediante la implementación efectiva de las decisiones tomadas, lo que puede implicar cambios en los procesos internos, desarrollo de nuevos productos o servicios, mejora de la experiencia del cliente y optimización de la eficiencia operativa. Estas acciones orientadas a crear valor brindan resultados tangibles y medibles, lo que contribuye al crecimiento y éxito a largo plazo de la empresa.

Es importante recordar que la cadena de valor en la ciencia de datos y los datos masivos es un proceso integral que comprende la recopilación, procesamiento, análisis y acción basada en los datos para crear valor agregado para las empresas.

Cada etapa de esta cadena tiene su importancia y requiere de herramientas y técnicas específicas para maximizar el potencial de los datos en beneficio de la organización. Al comprender y aprovechar adecuadamente esta cadena de valor, las empresas pueden obtener ventajas competitivas significativas en el mercado y lograr el éxito a largo plazo.

## 1.3. ¿Qué son los datos masivos?

Los datos masivos, también conocidos como big data, se refieren a conjuntos de datos de gran volumen, velocidad y variedad que superan ampliamente la capacidad de las herramientas tradicionales de procesamiento y análisis. Estos datos, provenientes de diversas fuentes como las redes sociales, transacciones en línea, sensores, y muchas más, han demostrado tener un potencial enorme para proporcionar información valiosa y estratégica que puede ser utilizada para tomar decisiones efectivas y obtener ventajas competitivas en el mercado.

La gestión y análisis de los datos masivos requieren de tecnologías avanzadas, tales como el aprendizaje automático y la inteligencia artificial, para poder extraer patrones, tendencias y conocimientos relevantes para las empresas. Estas tecnologías están en constante evolución y se han convertido en una parte integral de los procesos de toma de decisiones basados en datos. La utilización de algoritmos de aprendizaje automático permite a las empresas descubrir correlaciones, identificar tendencias y predecir comportamientos futuros, lo que a su vez les otorga una ventaja competitiva significativa.

Además, la cantidad de información generada en la actualidad es asombrosa. Diariamente se generan billones de gigabytes de datos, provenientes de millones de usuarios alrededor del mundo. Estos datos incluyen desde publicaciones en redes sociales hasta registros de compras en línea, desde información sobre el clima hasta datos biométricos recopilados por dispositivos wearables.

Todo este torrente de información, si se analiza correctamente, puede revelar patrones ocultos, relaciones complejas y tendencias emergentes, proporcionando una visión más completa y precisa de la realidad del mercado.

## Datos, información y conocimiento

Generalmente, los términos dato, información y conocimiento se utilizan de forma indistinta en ambientes no formales. Sin embargo, los ámbitos más formales (el profesional y el académico) requieren de una distinción precisa de estos conceptos. De esta forma, se evitan malinterpretaciones durante las distintas fases del análisis de datos.

Tal y como puede esperarse, son múltiples las aproximaciones que se utilizan para la distinción de estos términos. Seguiremos la proporcionada por Davenport y Prusak (1998).

**Un dato puede definirse como un hecho concreto y discreto acerca de un evento.**

La característica de ser discreto implica que, semánticamente, es la unidad mínima que puede comunicarse o almacenarse. Por sí solos, los datos no brindan detalles significantes del entorno del que fueron obtenidos.

Ejemplos de datos pueden ser:

- ▶ 2010.
- ▶ 443.
- ▶ DE.

**La información puede definirse como un mensaje formado por la composición de varios datos.**

A diferencia del dato, la información sí posee un significado para un receptor u observador. Por ejemplo, utilizando los datos anteriores se podría obtener la siguiente información:

- ▶ El año de fundación de la empresa ACME fue 2010.
- ▶ La altura del edificio Empire State es 443 metros.
- ▶ DE es el código ISO que identifica al idioma alemán.

**Los datos deben ser transformados para añadirles valor** y convertirlos en información. Estas transformaciones incluyen métodos como:

- ▶ Contextualización: conocer el propósito del dato obtenido.
- ▶ Categorización: conocer la unidad de medida y los componentes del dato.
- ▶ Cálculo: realizar una operación matemática sobre el dato.
- ▶ Corrección: eliminar errores del dato.
- ▶ Agregación: resumir o minimizar un dato de forma más concisa.

**El conocimiento implica una combinación de experiencias, información contextual y relevancia sobre cierta información.**

Así como la información se genera a partir de datos, el conocimiento surge de la agregación de información.

Ejemplos de métodos que generan esta transformación son:

- ▶ Comparación: relación entre información obtenida en distintas experiencias.
- ▶ Repercusión: implicación de la información en decisiones y acciones.

- ▶ Conexión: relación entre distintos tipos de información.
- ▶ Conversación: opinión de otras personas sobre la información.

La jerarquía del conocimiento suele representarse gráficamente por una pirámide en la que los datos suelen ser la base, mientras que el conocimiento se identifica con la cima, tal y como se representa en la Figura 1.



Figura 1. Representación piramidal de la jerarquía del conocimiento. Fuente: elaboración propia.

Distinguir estos conceptos básicos proporciona un nivel de abstracción útil para la separación de características en el proceso de análisis.

El hecho de que un dato sea inválido o erróneo debe distinguirse  
fácilmente de que la información que se obtiene de dicho conjunto de  
datos sea adecuada o no al problema que se intenta resolver.

Sirva como ejemplo la importancia de conocer si la malinterpretación de un análisis de datos se debe a un error en la fuente de datos, a un problema al combinar los datos en el proceso de análisis o a una confusión por parte del usuario final debido a experiencias en otros contextos.

## Tipos de datos

Los datos pueden ser representados de diferente forma en función de su naturaleza.

De forma básica podemos, identificar los siguientes tipos de datos:

- ▶ Numéricos: aquellos que pueden representarse mediante números.
- Discretos: solo toman un número finito de valores enteros (por ejemplo, el número de personas en un evento).
- Continuos: pueden tomar cualquier valor (número infinito de valores) dentro de un intervalo (por ejemplo, las magnitudes físicas como una temperatura y la longitud).
- Unidimensionales: aquellos que solo se pueden medir en una dirección (por ejemplo, clasificar personas en función de su altura).
- Multidimensionales: aquellos que se pueden medir en varias direcciones (por ejemplo, clasificar personas en función de su altura y su peso).
- Series temporales: sucesión de datos y su variación en función del momento temporal en el que han sido medidos.
- ▶ Texto: los datos también se pueden definir con caracteres alfabéticos formando cadenas de caracteres (por ejemplo, palabras que definen un término).
- ▶ Lógicos: son aquellos datos que pueden tomar dos valores: verdadero o falso (*true* o *false*). También se conocen como booleanos (por ejemplo, para definir el estado de una determinada acción, verdadero si se ha realizado o falso en caso contrario).

## 1.4. ¿Qué es la Ciencia de Datos?

La ciencia de datos es una disciplina que hace uso del método científico para extraer conocimiento de un conjunto de datos disponibles. Su principal diferencia con respecto a otras ciencias es que puede actuar de manera agnóstica al campo de aplicación de donde estos datos vengan.

Esto quiere decir que, por ejemplo, en química existen una serie de reglas, fórmulas, etc. a las cuales se les ha ido dando forma con el tiempo a medida que el ser humano va formulando teorías. En definitiva, la ciencia de datos trata de determinar qué acciones se pueden llevar a cabo con los datos para dar respuesta a preguntas, por ejemplo, ¿cuál es la proporción de hierro y carbono que necesito para formar acero?

La ciencia de datos es una disciplina aún más fundamental, no intenta encontrar fórmulas, sino que busca obtener información de los datos, sin entrar en métodos o fórmulas existentes, ayudando así a tener un mejor conocimiento de los datos de los que se dispone.

Jim Gray, reconocido científico de computadores y ganador del premio Turing, definía la ciencia de datos como el cuarto paradigma de la Ciencia. A los ya establecidos métodos empíricos, teóricos y computacionales, ahora habría que considerar el basado en los datos, debido a que está cambiando cómo se entiende todo lo que está relacionado con la ciencia. Para ello, la ciencia de datos hace un uso sinérgico de otras herramientas, como las matemáticas, la estadística, la informática, la ciencia de la información, la inteligencia artificial o el aprendizaje automático, entre otras.

Gracias a la ciencia de datos, las organizaciones tienen la capacidad de descubrir patrones ocultos, identificar tendencias significativas y obtener conocimientos profundos que pueden ser fundamentales para la toma de decisiones basadas en datos y la resolución de problemas empresariales. A través del análisis y la interpretación exhaustiva de los datos, es posible identificar oportunidades de crecimiento, optimizar las operaciones, mejorar los procesos, generar estrategias competitivas y maximizar el rendimiento en una amplia variedad de sectores e industrias.

En la era digital actual, la ciencia de datos juega un papel fundamental debido al cada vez más abrumador volumen y diversidad de datos generados diariamente. Ha pasado a ser un recurso imprescindible para aquellas organizaciones que buscan aprovechar al máximo los datos disponibles y utilizarlos de manera efectiva para alcanzar sus objetivos y mantenerse competitivos en un entorno empresarial en constante cambio.

## 1.5. La toma de decisiones basadas en datos

La toma de decisiones basada en datos es un enfoque holístico y riguroso que utiliza información objetiva y análisis estadístico profundo para respaldar decisiones estratégicas y tácticas en las empresas. Al aprovechar plenamente los datos y el conocimiento obtenido a través de la ciencia de datos, las organizaciones pueden tomar decisiones más informadas y fundamentadas, lo que a su vez les permite alcanzar sus objetivos organizativos de manera más eficiente y efectiva.

En este sentido, la toma de decisiones basada en datos implica el uso de una amplia gama de técnicas y herramientas de análisis de datos avanzadas para evaluar diferentes opciones y escenarios, identificar tendencias y patrones, analizar riesgos y oportunidades, y predecir resultados futuros. Esto se logra a través de la aplicación de algoritmos y modelos estadísticos complejos que permiten a las organizaciones extraer información valiosa de grandes conjuntos de datos, en tiempo real.

Además, también implica la adopción de medidas proactivas para garantizar la integridad y calidad de los datos utilizados en el proceso de toma de decisiones. Esto implica la implementación de prácticas de gestión de datos sólidas, como la limpieza y transformación de datos, la estandarización y normalización de variables, y la implementación de controles de calidad rigurosos.

Al utilizar este enfoque, las organizaciones pueden minimizar la incertidumbre y maximizar las probabilidades de éxito en la implementación de estrategias empresariales. La toma de decisiones basada en datos proporciona una base sólida y confiable para la formulación de estrategias, la asignación de recursos y la toma de decisiones operativas diarias. Además, ayuda a las empresas a identificar y capitalizar oportunidades emergentes, optimizar procesos internos y mejorar su ventaja competitiva en el mercado.

## 1.6. Problemas empresariales y soluciones de ciencia de datos

Los problemas empresariales pueden abordarse de manera más eficiente y efectiva utilizando soluciones basadas en la ciencia de datos. Esta disciplina permite analizar minuciosamente y comprender los desafíos que enfrenta una organización, identificar los factores clave involucrados y utilizar datos significativos para encontrar soluciones óptimas y efectivas en todos los aspectos empresariales.

Por ejemplo, la ciencia de datos puede ayudar de manera excepcional a una empresa a mejorar la eficiencia operativa, optimizar de forma precisa la cadena de suministro, detectar fraudes en tiempo real, personalizar y perfeccionar la experiencia del cliente, segmentar y alcanzar el mercado objetivo de manera precisa o predecir de manera acertada la demanda esperada de los productos o servicios ofrecidos.

Las soluciones de ciencia de datos se fundamentan en el análisis profundo y precisa de datos masivos, el desarrollo meticuloso de modelos predictivos y la aplicación detallada de técnicas avanzadas, que permiten convertir de forma eficaz y precisa los datos en conocimiento accionable para resolver de manera integral y precisa los problemas empresariales, evitando cualquier tipo de suposiciones o conjeturas innecesarias. La ciencia de datos ofrece una amplia gama de metodologías y herramientas, tales como el aprendizaje automático, la minería de datos, la visualización de datos, la inteligencia artificial y el análisis de sentimiento. Estas técnicas permiten a las empresas desbloquear el valor oculto en sus datos y aprovecharlos para tomar decisiones fundamentadas y estratégicas.

Además, la ciencia de datos no solo es aplicable a las grandes corporaciones, sino que también puede ser utilizada por pequeñas y medianas empresas. Al implementar soluciones basadas en datos, las empresas pueden optimizar sus operaciones, reducir costos, aumentar la eficiencia y mejorar la toma de decisiones en todos los niveles. Las ventajas de la ciencia de datos son múltiples y van más allá de solo resolver los problemas empresariales actuales, también permite a las empresas anticiparse a los desafíos del futuro y adaptarse rápidamente a los cambios del mercado.

## 1.7. Referencias bibliográficas

Bhageshpur, K. (15 de noviembre de 2019). Data Is The New Oil—And That's A Good Thing. *Forbes*.

<https://www.forbes.com/sites/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing/?sh=144132eb7304>

Borne, K. D., Jacoby, S., Carney, K., Connolly, A., Eastman, T., Raddick, J., Tyson, J. A. y Wallin, J. (2009). *The revolution in astronomy education: Data science for the masses*.

[https://www.researchgate.net/publication/45873844\\_The\\_Revolution\\_in\\_Astronomy\\_Education\\_Data\\_Science\\_for\\_the\\_Masses](https://www.researchgate.net/publication/45873844_The_Revolution_in_Astronomy_Education_Data_Science_for_the_Masses)

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, 69(1), 21-26.

Cukier, K. (2010). Data, data everywhere. *Economist*, 394(8671), 3-5.

Davenport, T. H. y Prusak, L. (1998). *Working Knowledge: How Organizations Manage what They Know*. Harvard Business Press.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4), 745-766.

Guillarranz, M. (14 de mayo de 2019). *El ciclo de vida de los datos: las 5 fases para llevar a éxito un proyecto de Big Data*. PiperLab. <https://piperlab.es/2019/05/14/el-ciclo-de-vida-de-los-datos-las-5-fases-para-llevar-a-exito-un-proyecto-de-big-data/>

Naur, P. (1974). Concise survey of computer methods. Petrocelli Books.

B12. Tech4Business. (10 de septiembre de 2019). *Qué es la ciencia de datos*. <https://agenciab12.com/noticia/que-es-ciencia-de-datos>

Routley, N. (7 de abril de 2018). *Data Visualization and Cholera: An Unexpected Connection*. Visual Capitalist. <https://www.visualcapitalist.com/data-visualization-cholera/>

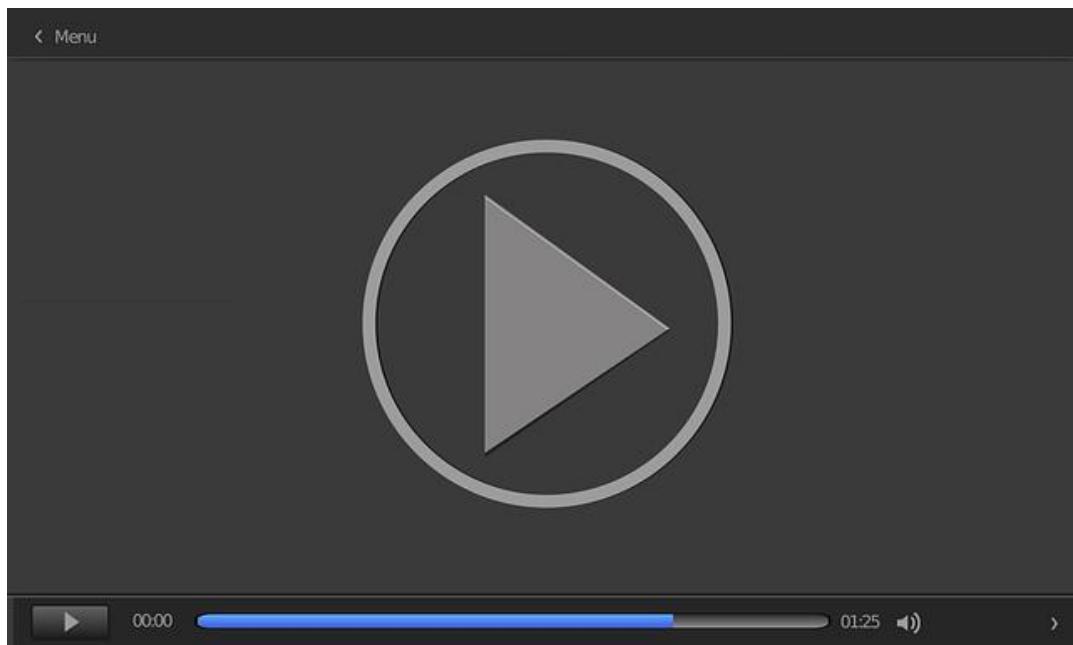
Simberloff, D. et al. (2005). Long-lived digital data collections: Enabling research and education in the 21st century. National Science Foundation.

Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*, 33(1), 1-67.

## El ciclo de vida del científico

Aprende IA con Ligdi González. (6 de octubre de 2020). *CICLO DE VIDA DEL CIENTÍFICO DE DATOS | #21 Aprende más sobre Inteligencia Artificial* [Archivo de vídeo]. YouTube. <https://www.youtube.com/watch?v=Z5qD8QPVpPY>

Lidgi González aborda en este vídeo el ciclo de vida del científico de datos desde la perspectiva de un proyecto. En él se enfatizan aspectos clave y habilidades necesarias para cada una de las fases del ciclo.



Accede al vídeo:

<https://www.youtube.com/embed/Z5qD8QPVpPY>

## Los datos son el nuevo petróleo – Y eso es bueno

Bhageshpur, K. (15 de noviembre de 2019). *Data Is The New Oil -- And That's A Good Thing.* Forbes.

<https://www.forbes.com/sites/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing/?sh=77e0dc6e7304>

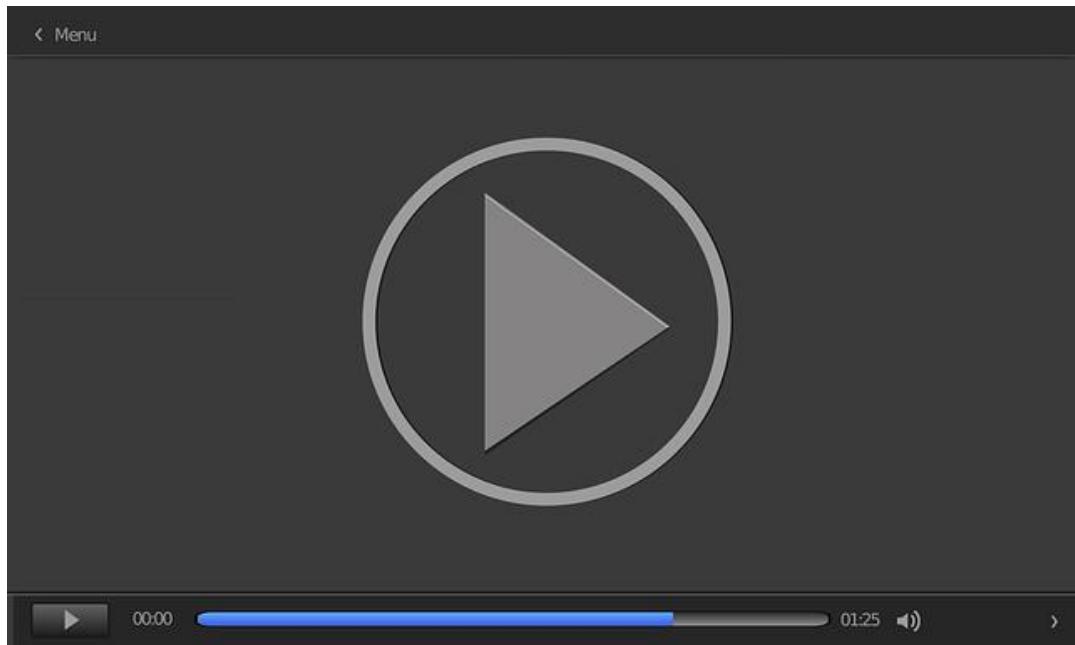
En este breve artículo publicado en Forbes se justifica por qué los datos se han convertido en el nuevo petróleo, en la nueva fuente de riqueza para cualquier institución. La información es importante y más en el siglo XXI.

## Ciencia de datos: ¿la profesión más sexy del siglo 21?

TEDx Talks. (17 de enero de 2020). *Ciencia de Datos: ¿La Profesión Más Sexy del Siglo 21? | Fredi Vivas | TEDxComodoroRivadavia* [Archivo de vídeo]. YouTube.  
<https://www.youtube.com/watch?v=AaoM5XhdnG0>

En esta TEDx Talk, Fredi Vivas (ingeniero en Sistemas de Información, especializado en *big data* y profesor en Disciplinas Industriales) aborda el atractivo de dedicarse profesionalmente a la ciencia de datos.

Esta disciplina se ha convertido en clave para que las organizaciones tomen mejores decisiones basándose en el análisis de la información, estimándose que, en el futuro a medio plazo, la ciencia de datos será el centro de todas las tendencias y una de las profesiones más demandadas. Fredi Vivas es un líder innovador con más de quince años como experto en tecnología para empresas multinacionales, universidades e instituciones públicas.



Accede al vídeo:

<https://www.youtube.com/embed/AaoM5XhdnG0>

- 1.** ¿Qué es la cadena de valor en la ciencia de datos?
  - A. Un software específico para análisis de datos.
  - B. Un proceso que transforma los datos en valor agregado para las empresas.
  - C. Una metodología de visualización de datos.
  - D. Un tipo de algoritmo de aprendizaje automático.
  
- 2.** ¿Qué caracteriza a los datos masivos?
  - A. Pequeños volúmenes de información.
  - B. Procesamiento manual de datos.
  - C. Grandes volúmenes de datos que requieren herramientas especiales.
  - D. Datos siempre estructurados.
  
- 3.** ¿Cuál es una aplicación de la ciencia de datos en las empresas?
  - A. Reducir la cantidad de datos a analizar.
  - B. Ignorar las tendencias del mercado.
  - C. Tomar decisiones basadas en conjeturas.
  - D. Optimizar procesos y mejorar la eficiencia operativa.
  
- 4.** ¿Qué se requiere en la fase de recopilación de datos?
  - A. Fuentes confiables y métodos eficientes.
  - B. Limitar la cantidad de datos recogidos.
  - C. Usar exclusivamente datos estructurados.
  - D. Evitar el uso de tecnología moderna.

5. ¿Qué permite el aprendizaje automático en el contexto de datos masivos?
  - A. Reducir la cantidad de datos necesarios para análisis.
  - B. Automatizar tareas complejas de análisis de datos.
  - C. Evitar el uso de cualquier tipo de datos.
  - D. Utilizar exclusivamente datos antiguos.
6. ¿Cuál es el primer paso en la cadena de valor de la ciencia de datos?
  - A. Visualización de datos.
  - B. Análisis de datos.
  - C. Recopilación de datos.
  - D. Toma de decisiones.
7. ¿Qué tecnologías son cruciales para manejar datos masivos?
  - A. Aprendizaje automático e inteligencia artificial.
  - B. Calculadoras básicas.
  - C. Herramientas de escritura manual.
  - D. Procesamiento de texto simple.
8. ¿Qué beneficios ofrece la ciencia de datos a las empresas?
  - A. Limita las fuentes de datos utilizadas.
  - B. Ignora las tendencias y patrones en los datos.
  - C. Ayuda a tomar decisiones basadas en datos y resolver problemas complejos.
  - D. Fomenta decisiones basadas únicamente en la intuición.

**9.** ¿Qué implica la acción basada en decisiones dentro de la cadena de valor de los datos?

- A. No tomar ninguna medida basada en los análisis.
- B. Implementar decisiones que crean valor para la organización.
- C. Desconocer los resultados del análisis.
- D. Revertir todas las decisiones previas.

**10.** ¿Qué se necesita para maximizar el potencial de los datos en beneficio de la empresa?

- A. Herramientas y técnicas específicas en cada etapa de la cadena de valor.
- B. Evitar la adopción de nuevas tecnologías.
- C. Reducir la cantidad de datos analizados.
- D. Concentrarse en datos irrelevantes.

Ciencia de Datos Aplicada

---

## Tema 2. Ciclo de vida de los datos masivos

# Índice

[Esquema](#)

[Ideas clave](#)

[2.1. Introducción y objetivos](#)

[2.2. Recolección](#)

[2.3. Análisis](#)

[2.4. Visualización](#)

[2.5. Interpretación](#)

[2.6. Referencias bibliográficas](#)

[A fondo](#)

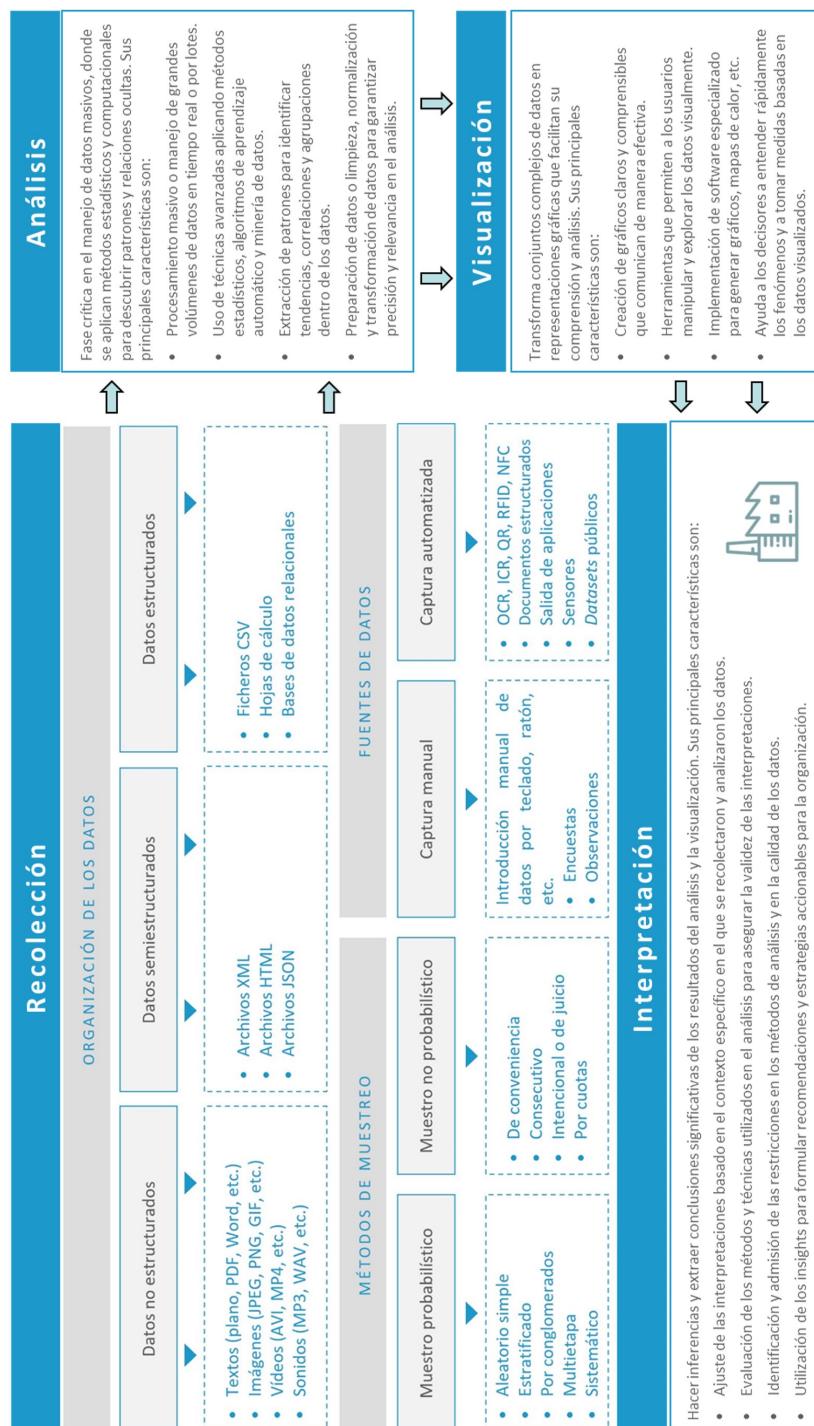
[Kaggle](#)

[OpenML](#)

[Google Dataset Search](#)

[DataPortals.org](#)

[Test](#)



## 2.1. Introducción y objetivos

Dentro de los sistemas de ciencia de datos, los datos, independientemente del contexto en el que se incluyan, han de ser sometidos a un proceso desde antes de su obtención hasta la explotación de estos, de forma que se facilite la toma de decisiones en función de la información que proporcionen.

A continuación, se presentan las principales etapas de su ciclo de vida. Como ejemplo, imaginemos que somos una empresa que vende ropa y cosméticos, para poder dar forma a cada una de las diferentes etapas (Guillarranz, 2019).

### Recolección

La primera fase en este proceso consiste en identificar las fuentes de información de las que se van a obtener los datos y la forma en que estas los proporcionarán. Es importante determinar de qué forma transmiten los datos (a través de una API, un fichero plano, directamente de un sensor, etc.), así como el volumen de datos que generan.

En la fase de captura o recolección de datos pueden implementarse tareas de limpieza de datos, tratamiento de datos ausentes, pequeñas transformaciones o comprobación de esquemas, entre otras, las cuales también pueden realizarse en la etapa de tratamiento.

Una vez que los datos son capturados y recibidos por el sistema diseñado, estos han de ser almacenados. En función del tipo de datos y de la información recibida debe definirse el mejor método de almacenamiento para que pueda disponerse de ellos en las siguientes etapas. Así, podrán diseñarse *data lakes*, *data warehouses* o *data marts* que estarán formados por sistemas de almacenamiento de ficheros, bases de datos relacionales, bases de datos NoSQL, etc.

Con los datos ya almacenados comienza el proceso de su preparación para su análisis. En esta etapa se realizan tareas de evaluación de la calidad (Quality Assurance) y preprocesamiento de los datos. La evaluación de la calidad analiza formatos, completitud, integridad, disponibilidad, etc.

Por otro lado, la tarea de preprocesamiento se encarga de realizar transformaciones, filtrados, eliminación de datos ruidosos u otras acciones que permitan disponer de los datos apropiados en un formato adecuado para su posterior análisis.

## Análisis

Con los datos ya preparados, la siguiente etapa se encarga de analizarlos. Es importante que se obtenga una descripción de estos, de forma que, por ejemplo, se especifique su cantidad, se entienda su distribución o se realicen análisis estadísticos que faciliten su comprensión (por ejemplo, a través de representaciones gráficas). De forma más concreta pueden realizarse tres tipos de análisis:

- ▶ Análisis descriptivo: describe lo que ha pasado con estadísticas, gráficos, tablas e informes.
- ▶ Análisis predictivo: realiza predicciones que van a ser útiles en el futuro siendo de especial importancia la precisión de la predicción.
- ▶ Análisis prescriptivo: ayuda a entender qué debe hacerse para obtener los resultados esperados en el futuro.

En esta etapa es habitual utilizar algoritmos de inteligencia artificial y, más concretamente, machine learning. Con ellos se facilita el aprendizaje a través de datos históricos, permitiendo extraer información valiosa de estos como asociaciones o patrones. Algunas de las técnicas que pueden emplearse son modelos de regresión, clasificación, segmentación y recomendación.

## Visualización

Finalmente, debe definirse la forma en la que se van a utilizar los resultados obtenidos de las etapas anteriores. Dependiendo del caso de uso concreto, los datos recogidos y la información generada pueden presentarse con informes periódicos o herramientas de visualización, o bien pueden integrarse en sistemas para la toma de decisiones de forma automática.

En el caso de presentar los datos de forma visual, la visualización es el proceso mediante el cual la información analizada se presenta de manera gráfica. Esta fase es esencial porque transforma grandes cantidades de datos complejos en representaciones visuales más accesibles y comprensibles, como gráficos, diagramas y mapas de calor. La visualización efectiva de datos ayuda a los stakeholders a comprender los resultados del análisis de manera intuitiva y rápida, facilitando así el proceso de toma de decisiones.

## Interpretación

Finalmente, la interpretación de los datos es el último paso, donde los resultados visualizados se examinan para tomar decisiones informadas. Esta etapa requiere una comprensión profunda del contexto en el que los datos fueron recolectados y analizados, así como de las limitaciones inherentes a los métodos estadísticos utilizados. La interpretación adecuada es crucial para aplicar los insights de manera efectiva y puede influir significativamente en las estrategias futuras de la organización.

Cada una de estas etapas es crucial y contribuye al ciclo de vida general de los datos masivos, asegurando que la información no solo sea recopilada y analizada, sino también correctamente interpretada y aplicada para cumplir con los objetivos

estratégicos. En las siguientes secciones, exploraremos cada una de estas etapas en detalle, profundizando en las técnicas, herramientas y consideraciones clave que marcan la diferencia en el manejo exitoso de datos masivos.

Veamos con detenimiento la descripción de las principales etapas de El ciclo de vida de los datos a través de un ejemplo práctico moderno.

En lo que a este tema se refiere, los contenidos en él expuestos buscan conseguir los siguientes objetivos:

- ▶ Describir con más énfasis la etapa de recolección en el ciclo de vida de los datos masivos, detallando las técnicas y desafíos asociados, ya que constituye la base fundamental para la calidad y eficacia del análisis posterior de los datos.
- ▶ Justificar la cobertura menos extensa de la etapa de análisis de datos dentro de este texto, considerando que existen asignaturas específicas dedicadas a profundizar en técnicas estadísticas avanzadas y modelos de aprendizaje automático aplicados a grandes volúmenes de datos.
- ▶ Explicar por qué se aborda de manera más concisa la visualización de datos, teniendo en cuenta que hay cursos especializados que se centran exclusivamente en la representación gráfica de información compleja, herramientas de visualización y metodologías de interpretación gráfica.

## 2.2. Recolección

Antes de describir los diferentes métodos de recolección, almacenamiento, tratamiento, análisis y visualización de datos, es conveniente mencionar los distintos tipos de organizaciones de datos con los que nos podemos encontrar. La forma en que se encuentren organizados los datos depende de su naturaleza y de cómo hayan sido capturados.

En este sentido, es importante mencionar que en un ciclo de vida de datos en particular algunas fuentes de datos suelen ser, a su vez, fuentes que ya fueron capturadas y almacenadas previamente, de modo que es importante conocer la estructuración de los datos antes incluso de ver con mayor profundidad los métodos de almacenamiento. Además, su organización influirá en el método que elijamos para almacenarlos.

Podemos clasificar la forma de organización de los datos en datos no estructurados, datos estructurados (o completamente estructurados) y datos semiestructurados (Sint, Schaffert, Stroka y Ferstl, 2009; Giudice, Musarella, Sofo y Ursino, 2019).

Es importante destacar que, en algunas ocasiones, la forma en la que organizamos los datos puede cambiar en función de en qué etapa del ciclo de los datos se encuentren. Por ejemplo, puede suceder que en algunos escenarios partamos de datos no estructurados y, tras el tratamiento de estos, pasemos a una organización de datos semiestructurados o completamente estructurados, siempre que sea necesario y posible.

Esto es análogo al tratamiento de datos con relación a la ETL (*extract, transform, load*) y el paso de *data lakes* a *data warehouses* y *data marts*. Aunque este proceso

puede verse también como parte de la captura de datos, antes es necesario que entendamos otros conceptos como las posibles fuentes de datos o los repositorios empleados para el almacenamiento de estos.

## Datos no estructurados

Los datos no estructurados son la forma «más cruda» de los datos sin una estructura identificable y pueden ser cualquier tipo de dato como textos, imágenes, sonidos o vídeos. Una forma de determinar si los datos son no estructurados es si no podemos almacenarlos en filas y columnas en una base de datos relacional (Sint et al., 2019).

Fue tal el golpecillo, que me desatinó y sacó de sentido, y el jarrazo tan grande, que los pedazos de él se me metieron por la cara, rompiéndomela por muchas partes, y me quebró los dientes, sin los cuales hasta hoy día me quedé.

Desde aquella hora quise mal al mal ciego, y, aunque me quería y regalaba y me curaba, bien vi que se había holgado del cruel castigo. Lavóme con vino las roturas que con los pedazos del jarro me había hecho, y, sonriéndose, decía:

—¿Qué te parece Lázaro? Lo que te enfermó te sana y da salud —y otros donaires que a mi gusto no lo eran.

Ya que estuve medio bueno de mi negra trepa y cardenales, considerando que, a pocos golpes tales, el cruel ciego ahorraría de mí, quise yo ahorrar de él; mas no lo hice tan presto, por hacerlo más a mi salvo y provecho. Y aunque yo quisiera asentar mi corazón y perdonarle el jarrazo, no daba lugar el maltratamiento que el mal ciego dende allí adelante me hacía, que sin causa ni razón me hería, dándome coscorrones y repelándome.

Y si alguno le decía por qué me trataba tan mal, luego contaba el cuento del jarro, diciendo:

—¿Pensaréis que este mi mozo es algún inocente? Pues oíd si el demonio ensayara otra tal hazaña.

Santiguándose los que lo oían, decían:

—¡Mirad quién pensara de un muchacho tan pequeño tal ruindad!

Y reian mucho el artificio y decíanle:

—¡Castigadlo, castigadlo, que de Dios lo habréis!

Y él, con aquello, nunca otra cosa hacía.

Y en esto yo siempre le llevaba por los peores caminos, y adrede, por hacerle mal y daño; si había piedras, por ellas; si lodo, por lo más alto; que, aunque yo no iba por lo más enjuto, holgábame a mí de quebrar un ojo por quebrar dos al que ninguno tenía. Con esto, siempre con el cabo alto del tiento me atentaba el colodrillo, el cual siempre traía lleno de tolondrones y pelado de sus manos. Y, aunque yo juraba no hacerlo con malicia, sino por no hallar mejor camino, no me aprovechaba ni me creía, mas tal era el sentido y el grandísimo entendimiento del traidor.

Y porque vea Vuestra Merced a cuánto se extendía el ingenio de este astuto ciego, contaré un caso de muchos que con él me acaecieron, en el cual me parece dio bien a entender su gran astucia. Cuando salimos de Salamanca, su motivo fue venir a tierra de Toledo, porque decía ser la gente más rica, aunque no muy limosnera. Arrimábase a este refrán: «Más da el duro que el desnudo». Y vinimos a este camino por los mejores lugares. Donde hallaba buena acogida y ganancia, deteniamonos; donde no, a tercero dia hacíamos San Juan.

Acaeció que, llegando a un lugar que llaman Almorox al tiempo que cogían las uvas, un vendimiador le dio un racimo de ellas en limosna. Y como suelen ir los cestos maltratados, y también porque la uva en aquel tiempo está muy madura, desgranábasele el racimo en la mano. Para echarlo en el fardel, tornábase mosto, y lo que a él se llegaba. Acordó de hacer un banquete, así por no poder llevarlo, como por contentarme, que aquel día me había dado muchos rodillazos y golpes. Sentámonos en un valladar y dijo:

Figura 1. Ejemplo de datos no estructurados. Fragmento de eBook en formato EPUB de la obra *El Lazarillo de Tormes*. Fuente: epublibre.

Antes de que las veamos, podemos pensar, por el momento, en las bases de datos relacionales como en tablas conformadas por filas y columnas y donde las tablas están relacionadas entre sí (clientes y pedidos, por ejemplo). En realidad, se suele afirmar que el 80 % de los datos con los que cuenta una organización son datos no estructurados (Das y Kumar, 2013).

Algunos ejemplos de datos no estructurados serían:

- ▶ Textos: archivos de texto plano (.txt), archivos de procesadores de texto como Word u OpenOffice, archivos PDF, archivos de e-book (EPUB, MOBI, etc.), correos electrónicos, presentaciones de PowerPoint, etc.
- ▶ Imágenes y animaciones: archivos JPEG, PNG, GIF, etc.
- ▶ Sonidos: archivos MP3, OGG, WAV, etc.
- ▶ Vídeos: archivos MPEG / MP4, AVI, MKV, OGG, etc.

Estos datos pueden estar almacenados en un repositorio de archivos similar a un directorio organizado en el sistema de ficheros del disco duro de un equipo informático. Extraer información de este tipo de datos implica un mayor esfuerzo que en datos más estructurados. Para ello, en primer lugar, es necesario extraer características estructuradas de los datos que describen o se abstraen de ellos.

En el caso de los textos, se pueden aplicar búsquedas en el texto completo (*full-text search queries*), cuya ventaja radica en que no necesitan conocer ninguna estructura (esquema) previa en el texto. Para obtener más información de los textos, se pueden aplicar técnicas de procesamiento del lenguaje natural (*natural language processing*) (Beysolow II, 2018) que permiten extraer características estructuradas de los datos o incluso *análisis de sentimiento* de los textos (si un comentario en Twitter es positivo o negativo, por ejemplo) (Lima, de Castro y Corchado, 2015).

En el caso de las imágenes, es posible emplear técnicas Deep Learning como las Redes Neuronales Convolucionales que nos permitan identificar si una imagen se corresponde con una determinada persona, animal u objeto (Krizhevsky, Sutskever y Hinton, 2017).

## Datos estructurados

Los datos estructurados o completamente estructurados siguen un esquema (*schema*) que requiere cierto esfuerzo elaborar. El caso más habitual de este tipo de esquema en los datos estructurados es el de tablas o esquema tabular. En los datos tabulares podemos estructurar los datos en filas y columnas, como si de una tabla se tratara. La cabecera (*header*) o primera fila nos indica el nombre (y tipo) de cada uno de los atributos. Una fila o un conjunto de atributos es conocida como un registro o una instancia de la especificación dada por el esquema de datos.

Los archivos CSV (*comma-separated values*) son un ejemplo claro de datos estructurados, siendo ampliamente utilizados como *datasets* en ciencia de datos.

Show_id	Type	Title	Director	Country	Release_year	Duration
s1	TV show	3 %		Brazil	2020	4 seasons
s2	Movie	7:19	Jorge Michel Grau	Mexico	2016	93 min
s3	Movie	23:59	Gilbert Chan	Singapore	2011	78 min
s4	Movie	9	Shane Acker	United States	2009	80 min
s5	Movie	21	Robert Luketic	United States	2008	123 min
s6	TV show	46	Serdar Akar	Turkey	2016	1 season

Tabla 1. Ejemplo de datos estructurados. Tabla representando datos parciales de un dataset en formato CSV. «TV Shows and Movies listed on Netflix» de Flixable. Fuente:

<https://www.kaggle.com/shivamb/netflix-shows>

Entre los ejemplos de datos estructurados, podemos incluir:

- ▶ Ficheros CSV.
- ▶ Hojas de cálculo (Excel o similares).
- ▶ Bases de datos relacionales (SQL u OLAP).

En realidad, **el ejemplo más extendido en la práctica de datos estructurados es el de aquellos almacenados en una base de datos relacional**. En ella, el esquema ha de ser definido antes de almacenar datos. Dicho esquema define el tipo de los datos, la estructura de los datos y las relaciones entre dichos datos, es decir, entre registros de diferentes tablas en la misma base de datos.

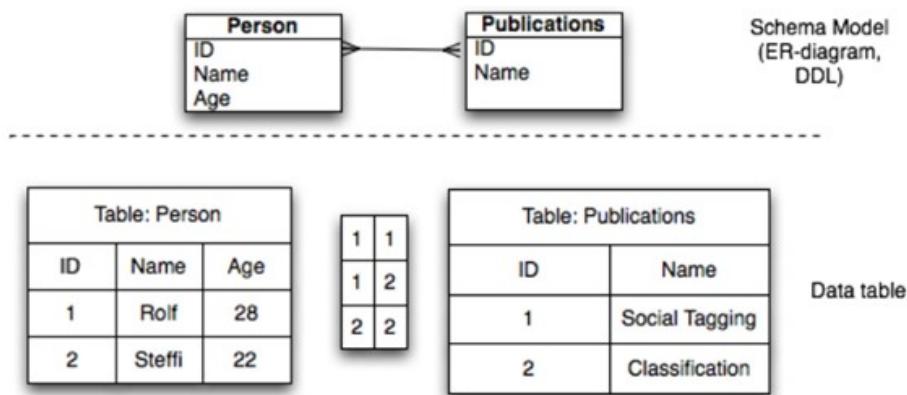


Figura 2. Ejemplo de datos estructurados: esquema y datos en una base de datos relacional. Fuente: Sint et al., 2019.

Los datos estructurados presentan como ventajas la eficiencia a nivel de rendimiento a la hora de navegar y trabajar con dichos datos (por ejemplo, cuando le pedimos a una base de datos relacional que realice operaciones sobre los datos mediante secuencias SQL).

Sin embargo, como desventajas, estos datos son menos flexibles y escalables, ya que una vez que especificamos un esquema es difícil cambiarlo (por ejemplo, añadir un atributo «Temporada» a una tabla con títulos de Netflix implica añadir una columna con un nuevo valor a todas las filas de la tabla, aunque no todos los registros empleen dicho atributo, como, por ejemplo, todos los títulos que sean películas y no series).

## Datos semiestructurados

En los datos semiestructurados el esquema no está separado de los datos, sino que estos son autodescriptivos. Esto no significa que el esquema como tal por separado no exista, sino que, en realidad, este es opcional (Sint et al., 2019). Además, un esquema también puede definirse en función de instancias ya existentes (*a posteriori*).

Los datos semiestructurados son estructuras más flexibles que los datos estructurados y no tienen por qué ser tabulares. Por ejemplo, un atributo puede presentar un tipo de datos variable, siendo una cadena de texto para algunas instancias, un valor numérico para otras o incluso directamente no existir en absoluto en algunas instancias.

Imaginemos una lista de productos en una tienda *online* como Amazon, quizás casi todos los productos tengan dimensiones (alto, ancho y largo) y peso, pero Amazon vende también licencias *software* que no son bienes tangibles. En el caso de los contenidos en Netflix, esto permite que haya registros con el atributo «Temporada» (las series) y otros no (las películas).

Así, un ejemplo típico de datos semiestructurados son las secuencias o archivos JSON (*javascript object notation*), como el ejemplo que podemos observar a continuación:

```
{"empleados": [ { "firstName":"Pedro", "lastName":"García" },  
    { "firstName":"María", "lastName":"Alonso" },  
    { "firstName":"Rosa", "lastName":"Pérez" }  
]}
```

De este modo, como ejemplos de datos semiestructurados tendríamos:

- ▶ Archivos XML (*extended markup language*), utilizado para intercambiar información en una amplia variedad de contextos, incluyendo API accesibles mediante SOAP (*simple object access protocol*).
- ▶ Archivos HTML (*hypertext markup language*), utilizando para mostrar contenidos con semántica en las páginas web.
- ▶ Archivos JSON (*Javascript object notation*), muy extendidos para intercambiar datos en las REST API, especialmente.
- ▶ Archivos YAML (*YAML ain't markup language* u, originalmente, *yet another markup language*), muy empleado en la actualidad en archivos de configuración.
- ▶ Tripletas RDF (*resource description framework*), empleadas en web semántica para codificar expresiones sujeto-predicado-objeto.

Aunque los datos semiestructurados suelen almacenarse en forma de archivos como, por ejemplo, ficheros JSON o XML, algunos tipos de datos semiestructurados pueden almacenarse en bases de datos orientadas a documentos (como MongoDB, por ejemplo, que trabaja con datos BSON, es decir, *Binary JSON*). Esas bases de datos permiten consultar los datos semiestructurados mediante una API de acceso adecuada.

## Mecanismos de muestreo

En este apartado vamos a describir, en primer lugar, los principales mecanismos de muestreo a la hora de recoger los datos desde diferentes fuentes, así como las categorías de las fuentes de datos más utilizadas en la actualidad. Finalizaremos la sección describiendo las Interfaces de Programación de Aplicaciones (API), por ser un mecanismo de interoperabilidad con fuentes de datos, repositorios de almacenamiento e incluso capas de tratamiento de datos ampliamente utilizadas.

Aunque el valor de los datos es innegable y el coste de almacenamiento de información en la nube en la actualidad es realmente bajo, no siempre es posible (ni necesario) almacenar absolutamente todos los datos o mediciones sobre un fenómeno (fuente de datos) en particular.

De hecho, almacenar datos que no aportan información no es útil y, a la postre, complica el manejo de estos por parte de los científicos de datos, que deberán efectuar tareas de limpieza y descartar datos sin valor, y redundarán en mayores costes de procesamiento. Los costes de la computación en la nube (uso de CPU y RAM en el tiempo) son comparativamente más elevados que el almacenamiento en la misma (Hassan, Nasir, Khairudin y Adon, 2017). Asimismo, el propio proceso de recogida de datos redundantes que no aportan información implica costes de recursos humanos, computacionales y energéticos innecesarios.

## Muestro de señales y el teorema de Shannon-Nyquist

Un ejemplo evidente es la recogida de datos desde **redes de sensores o dispositivos IoT**, ampliamente utilizados en la medición de datos acerca del entorno, de los propios usuarios o de su contexto. Estos elementos tienen una numerosa variedad de aplicaciones como las ciudades inteligentes (*smart cities*), la monitorización de variables biomédicas en aplicaciones sociosanitarias (*healthcare*) o en el ámbito de la Industria 4.0, por citar solo algunas de ellas (Alonso, Sittón-Candanedo, García, Prieto y Rodríguez-González, 2020).

Imaginemos una estación meteorológica en una ciudad midiendo la temperatura del aire, la humedad del aire, los niveles de polución, etc. y enviando por red de datos 4G los valores medidos a una base de datos en la nube.

En primer lugar, aunque la temperatura sea una variable continua, no podemos medir, enviar y almacenar infinitos valores en la base de datos, dado que los sistemas informáticos actuales procesan y almacenan información de forma discreta.

Por otra parte, si el sensor de temperatura solo proporciona una resolución de 0.1 °C, probablemente medir la temperatura cada segundo no va a aportar ninguna información adicional, ya que existirán muchas medidas seguidas repetidas sin variación alguna. Medir y enviarlas todas implica un consumo energético innecesario (uso de la red 4G), especialmente en dispositivos alimentados con baterías y paneles solares, además de crear un *dataset* con numerosos datos sin valor adicional.

A este respecto, para caracterizar toda la información de una magnitud (una señal) que varía en el tiempo (una serie temporal, al fin y al cabo) es suficiente con tomar muestras de esta a una frecuencia determinada.

Según el **teorema de Shannon-Nyquist**, si una función  $x(t)$  no contiene frecuencias superiores a  $B$  hertzios, podemos caracterizar completamente esta si tomamos muestras cada  $1/(2B)$  segundos o, lo que es lo mismo, con una frecuencia igual a  $2B$  (Vaidyanathan, 2001).  $2B$  es llamada también la tasa de Nyquist, mientras que la máxima  $B$  permitida por un equipo de muestreo (un conversor analógico-digital) es la frecuencia de Nyquist.

Esto es extrapolable a datos que varían en el espacio (en lugar de en el tiempo) o datos multidimensionales. Un ejemplo claro es la digitalización de una fotografía (datos que varían en el espacio en dos dimensiones).

Un ejemplo que ilustra esto es el siguiente. Si el oído humano solo alcanza a percibir sonidos por debajo de 20 000 Hz (en términos generales), cuando estamos digitalizando música (la voz humana tiene frecuencias mucho más bajas, por debajo de 4000 Hz, en cualquier caso), eliminamos mediante filtros de sonido aquellas frecuencias por encima de los 20 000 Hz y pasamos a muestrear. Dado que la

frecuencia máxima tras dicho filtrado es de 20 000 Hz, tomaremos muestras cada 0,000025 s, es decir, cada 25  $\mu$ s, o, lo que es lo mismo, con una frecuencia de muestreo de 40 000 Hz, y podremos almacenar muestras digitales de una canción con la mínima cantidad de datos necesaria que permite reconstruir posteriormente la señal sin ninguna pérdida.

De este modo, en este ejemplo, si **sobremuestreamos** la señal tomando muestras con una frecuencia mayor a 40 000 Hz (la tasa de Nyquist para dicho ejemplo), estaremos desperdiciando recursos sin aportar información adicional.

Sin embargo, si **submuestreamos** tomando muestras con una frecuencia inferior a los 40 000 Hz no podremos reconstruir la señal sin error, dado que existen infinitas señales posibles que pueden corresponderse con las muestras tomadas. Este efecto es lo que se conoce como *aliasing* o solapamiento y se muestra de forma gráfica en la figura 3.

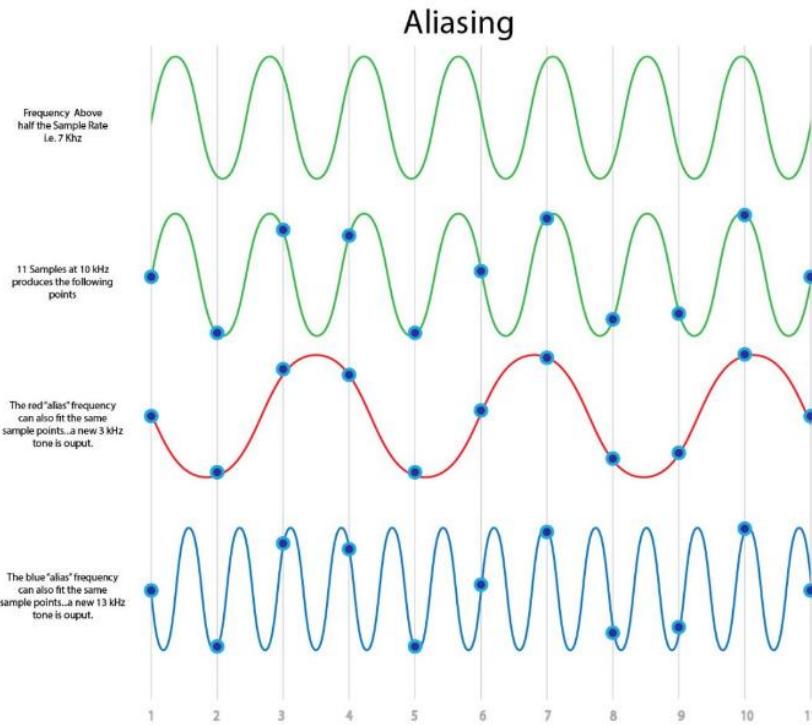


Figura 3. *Aliasing* o solapamiento resultado de muestrear por debajo de la tasa de Nyquist. Fuente:

[https://www.realhd-audio.com/wp-content/uploads/2014/05/140528\\_aliasing\\_illustration.jpg](https://www.realhd-audio.com/wp-content/uploads/2014/05/140528_aliasing_illustration.jpg)

Como se puede observar en la figura, al no haber muestreado la señal original con una frecuencia de al menos el doble de la frecuencia del tono (en audio, un sonido simple con una única frecuencia, es decir, una única componente sinusoidal, es conocida como un tono), existen otros tonos con frecuencia diferente cuyo muestreo a la frecuencia de muestreo empleada darían como resultado las mismas muestras, resultando confusa la reconstrucción de la señal.

## Muestreo de poblaciones: métodos probabilísticos, métodos no probabilísticos y sesgo

Otro ejemplo evidente son los barómetros de opinión, encuestas de intención de voto o encuestas de población activa. El coste de preguntar a absolutamente todos los

habitantes de un país sobre su situación laboral o sobre su intención de voto en unos comicios cercanos o lejanos en el tiempo haría prohibitivo este tipo de encuestas. Para ello, se llevan a cabo **muestreos escogiendo solo un subconjunto de la población** a la cual realizar la encuesta. Esto **reduce los costes sensiblemente, pero introduce un error de muestreo** que es necesario tener en cuenta.

Existen, a este respecto, diferentes métodos de muestro posibles (Garvin, 1987; Biscobing, 2018). Estos pueden dividirse en **métodos probabilísticos y métodos no probabilísticos**.

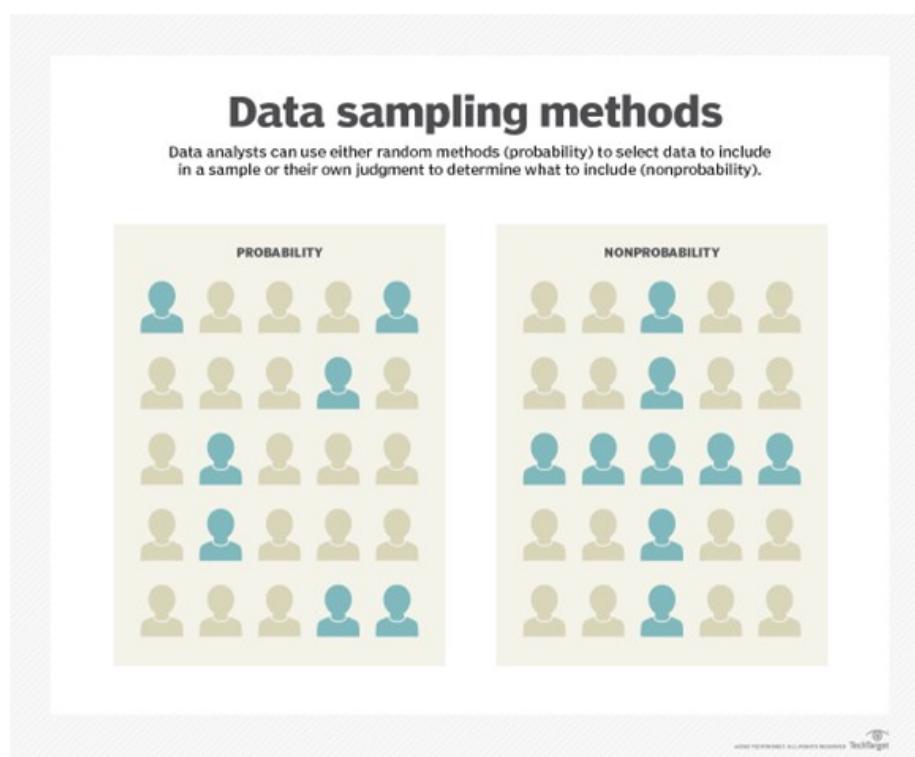


Figura 4. Métodos de muestro de datos probabilísticos y no probabilísticos. Fuente: Biscobing, 2018.

El muestreo puede basarse en la probabilidad, un enfoque que utiliza números aleatorios que corresponden a puntos del conjunto de datos para garantizar que no haya correlación entre los puntos elegidos para la muestra. Entre los métodos probabilísticos podemos destacar:

- ▶ **Muestreo aleatorio simple** (*simple random sampling*): se utiliza un proceso informático para seleccionar aleatoriamente a los sujetos de toda la población.
- ▶ **Muestreo estratificado** (*stratified sampling*): se crean subconjuntos de los conjuntos de datos o de la población en función de un factor común y se recogen muestras al azar de cada subgrupo (por ejemplo, estratificar la población por sexo, edad, provincia, ingresos, etc.). Así nos aseguramos de que no tenemos, por ejemplo, un 70 % de muestras de hombres, cuando estos no representan tal porcentaje en la población total. En caso contrario, introduciríamos un sesgo en el muestreo.
- ▶ **Muestreo por conglomerados** (*cluster sampling*): el conjunto de datos más amplio se divide en subconjuntos (*clusters* o clústeres) en función de un factor definido, y luego se analiza un muestreo aleatorio de los clústeres. Por ejemplo, crear subconjuntos por región o por provincia o crear subconjuntos por tipos de consumidores que se comportan de forma similar. Es importante que el comportamiento dentro de cada clúster sea relativamente homogéneo, mientras que el comportamiento entre individuos de diferentes clústeres sea relativamente heterogéneo.
- ▶ **Muestreo multietapa** (*multistep sampling*): se trata de una forma más complicada del muestreo por conglomerados. Este método también implica la división de la población más grande en una serie de clústeres. A continuación, los clústeres de la segunda etapa se dividen en función de un factor secundario y esos clústeres se muestrean y analizan. Este escalonamiento podría continuar a medida que se identifican, agrupan y analizan múltiples subconjuntos.
- ▶ **Muestreo sistemático** (*systematic sampling*): en este método, se crea una muestra estableciendo un intervalo en el que se extraen los datos de la población mayor. Por ejemplo, seleccionando cada 10 filas en una hoja de cálculo o un fichero CSV de 2000 elementos para crear un tamaño de muestra de 200 filas para analizar.

El **muestreo también puede basarse en la no probabilidad**, un enfoque en el que se determina y extrae una muestra de datos **basada en el juicio del científico de**

**datos** (el investigador, analista de datos o científico de datos). Como la inclusión la determina el científico de datos (introduciendo su sesgo particular), puede ser más difícil extraer si la muestra representa con exactitud a la población más amplia que cuando se utiliza el muestreo probabilístico. Entre los métodos probabilísticos podemos destacar (Biscobing, 2018):

- ▶ **Muestreo de conveniencia** (*convenience sampling*): los datos se recogen de un grupo fácilmente accesible y disponible.
- ▶ **Muestreo consecutivo** (*consecutive sampling*): los datos se recogen de cada sujeto que cumple los criterios hasta alcanzar el tamaño de muestra predeterminado.
- ▶ **Muestreo intencional o de juicio** (*purposive o intentional sampling*): el investigador selecciona los datos a muestrear basándose en criterios predefinidos.
- ▶ **Muestreo por cuotas** (*quota sampling*): el investigador garantiza una representación equitativa dentro de la muestra para todos los subgrupos del conjunto de datos o población.

Debemos tener en cuenta que el método de muestreo es fundamental para no incurrir en sesgos que impidan generalizar los resultados obtenidos de una investigación o resultado de aplicar procedimientos de ciencia de datos a un *dataset* creado. **Si una muestra no es representativa de la población, entonces está sesgada.**

El sesgo puede ser accidental (por ejemplo, seleccionar al azar una mayoría de individuos de un pequeño subgrupo) o intencionado, es decir, el sesgo se introduce para sesgar las opiniones o promover un punto de vista particular.

## Fuentes de datos

A la hora de recopilar datos, los científicos de datos pueden recurrir a diferentes fuentes. En este sentido, hablaremos de fuentes primarias y fuentes secundarias (Garvin, 1987):

- ▶ Fuentes primarias: son las que proporcionan datos originales, como experimentos, redes de sensores propias o encuestas realizadas por los propios investigadores o científicos de datos.
- ▶ Fuentes secundarias: son las que proporcionan datos recogidos de otros, como *datasets* previamente existentes, artículos de revistas, informes de periódicos, consultoras o encuestas realizadas por otros.

Al realizar un estudio sobre datos **es importante referenciar todas las fuentes de datos utilizando un formato adecuado**, especialmente cuando son fuentes secundarias. Esto permite, así, reproducir la investigación realizada, tanto si se trata de una investigación en el ámbito académico como en el entorno empresarial.

Una vez vistos los diferentes métodos de muestreo, los métodos de captura de datos propiamente dichos pueden clasificarse en función de las características del elemento (la fuente de datos) que genera el conjunto de datos. A continuación, exponemos algunas de las categorías más utilizadas en la actualidad (Taylor, 2020).

## Captura manual de datos

Es la categoría más tradicional y también una de las más habituales en el contexto de la investigación en los ámbitos social y natural, especialmente en las investigaciones de campo. Entre los métodos que encajan en esta categoría se incluyen el uso de encuestas y las mediciones a través de observaciones.

Aunque es la única categoría que no depende directamente de las tecnologías de la información, su uso requiere que la información sea ulteriormente digitalizada, ya sea en el momento de la captura o durante el procesamiento posterior.

Así, en el proceso de captura de datos manual, los datos son introducidos manualmente por un operador que utiliza dispositivos de entrada como el teclado, las pantallas táctiles, el ratón, etc., para introducir los datos en forma de cifras o texto en un *software* concreto como Excel o cualquier otro programa de procesamiento de datos o de textos.

Este método de recopilación de datos requiere mucha mano de obra y tiempo, por lo que las empresas consideran eficiente migrar a métodos automatizados de captura de datos. Sin embargo, el método manual no está totalmente extinguido y todavía encuentra aplicación en muchos procesos empresariales. De hecho, existen operaciones de introducción de datos que necesitan ser realizadas por humanos, como el etiquetado de datos en *datasets*, para lo cual en ocasiones se hace uso de *marketplaces* para la externalización de estas tareas, como Amazon Mechanical Turk.

Los métodos de captura manual de datos incluyen el uso de elementos como el teclado, el ratón, tabletas gráficas digitalizadoras o pantallas táctiles, entre otros posibles.

## Captura automatizada de datos

La captura de datos automatizada implica el uso de tecnología computarizada para capturar datos. Este método puede presentar un alto coste inicial debido a la inversión inicial requerida como, por ejemplo, la compra de tecnología, pero, a medida que avanza el proyecto, se encuentra que reduce significativamente los costes operativos debido a la baja necesidad de mano de obra (Taylor, 2020).

Además, dado que la mayoría de los datos existen hoy en día en formato electrónico, el coste de utilizar dicha tecnología automatizada también se ha reducido. Por lo tanto, ha existido una proliferación de técnicas y tecnología de métodos automatizados de captura de datos, cada uno adecuado para un tipo de datos o fuente de datos en particular.

La captura de datos automatizada incluye el uso de diferentes tecnologías como OCR (*optical character recognition*), utilizado para reconocer texto en documentos impresos, ICR (*intelligent character recognition*), para reconocer texto manuscrito, OMR (*optical mark recognition*), para reconocer respuestas en formularios y encuestas, códigos de barras, códigos bidimensionales (códigos QR), tarjetas inteligentes basadas en tecnologías inalámbricas RFID (*radio-frequency identification*) o, más concretamente en la actualidad, NFC (*near-field communications*), especialmente interesantes en sistemas para la trazabilidad o la logística, o incluso tecnologías de reconocimiento de voz.



Figura 5. Ejemplos de tags RFID pasivos. Fuente: Ministerio de Educación, 2019.

El resto de las fuentes de datos que vemos a continuación forman parte de lo que se conoce como captura automatizada de datos en mayor o menor medida, es decir, con menor o mayor intervención humana.

## ► Procesamiento de documentos estructurados

Consiste en la extracción directa de los datos disponibles en documentos cuya finalidad original no es ser consultados como fuente de datos y, por tanto, no fueron ya preparados para ello.

Uno de los métodos más comunes en esta categoría es el procesamiento de las páginas HTML de un sitio web, conocido como *web scraping*, en el que se emplean *bots* para extraer información de una web existente a partir del contenido HTML aprovechando la semántica de dicho lenguaje (la estructura de los contenidos web haciendo uso de etiquetas de cabecera, títulos de nivel 1, títulos de nivel 2, etc.).

Otro ejemplo es el análisis de los *logs* o registros, archivos que contienen una bitácora o lista secuencial de eventos ocurridos en un sistema y que fueron creados con el propósito de registrarlos para que un usuario humano, habitualmente un administrador de sistemas pueda averiguar qué ha sucedido históricamente con un proceso *software* (qué usuarios han accedido a una base de datos o a un portal web), pero no específicamente para que otras aplicaciones puedan acceder a ellos.

```
216.239.46.60 - - [04/Jan/2003:14:56:50 +0200] "GET  
/~lpis/curriculum/C+Unix/Ergastiria/Week-7/filetype.c.txt HTTP/1.0"  
304 -  
216.239.46.100 - - [04/Jan/2003:14:57:33 +0200] "GET  
/~oswinds/top.html HTTP/1.0" 200 869  
64.68.82.70 - - [04/Jan/2003:14:58:25 +0200] "GET /~lpis/systems/r-  
device/r_device_examples.html HTTP/1.0" 200 16792  
216.239.46.133 - - [04/Jan/2003:14:58:27 +0200] "GET  
/~lpis/publications/crc-chapter1.html HTTP/1.0" 304 -  
209.237.238.161 - - [04/Jan/2003:14:59:11 +0200] "GET /robots.txt  
HTTP/1.0" 404 276  
209.237.238.161 - - [04/Jan/2003:14:59:12 +0200] "GET  
/teachers/pitas1.html HTTP/1.0" 404 286  
216.239.46.43 - - [04/Jan/2003:14:59:45 +0200] "GET  
/~oswinds/publications.html HTTP/1.0" 200 48966
```

Figura 6. Ejemplo de un fichero de *log* de un servidor web. Fuente: Vakali, Pokorný y Dalamagas, 2004.

#### ► Salida de aplicaciones

Si bien un *log* o registro visto anteriormente pudiera entenderse también como la salida de una aplicación, no encaja en esta categoría por un motivo: las fuentes de datos englobadas dentro la categoría de salida de aplicaciones sí están pensadas para que puedan ser ingeridas por otros procesos *software*.

Así, dentro de esta categoría, podríamos tener ficheros con valores separados por comas (CSV), archivos XML, ficheros binarios para el manejo de datos (AVRO), formatos de fichero propios o específicos de una aplicación concreta (ficheros con información geoespacial en el caso de los GIS o sistemas de información geográfica) o, muy frecuentemente, almacenes de datos propiamente dichos como **bases de datos relacionales o SQL, bases de datos NoSQL o almacenes de objetos**.

#### ► Datos obtenidos a través de sensores

En este conjunto se pueden mencionar ejemplos como sensores meteorológicos (termómetros, higrómetros, pluviómetros, anemómetros, etc.), sensores ambientales (ruido, luz, control de afluencia midiendo la presencia de teléfono móviles a partir de señales *bluetooth* y *Wi-Fi*), sensores biomédicos corporales (termómetros, ritmo cardíaco, conductividad de la piel, cámaras termográficas para detectar posible fiebre

para control de epidemias, etc.) o sensores de dispositivos móviles (acelerómetros, giróscopos o magnetómetros). Actualmente existe un gran interés en el uso de sensores para la captura de datos en el contexto personal a través de *wearables* como pulseras de actividad o *smartwatches*, este movimiento es conocido como *quantified self*.

Actualmente las redes de sensores se encuentran ampliamente extendidas en nuestra vida cotidiana y ya forman parte indispensable de lo que se conoce como el **Internet de las cosas** o *Internet of things* con aplicaciones en múltiples ámbitos como la domótica (temperatura, presencia, videovigilancia), la agricultura inteligente, las ciudades inteligentes o la industria 4.0, donde se habla ya de Internet de las cosas industrial (**IIoT**) (Alonso et al., 2020).

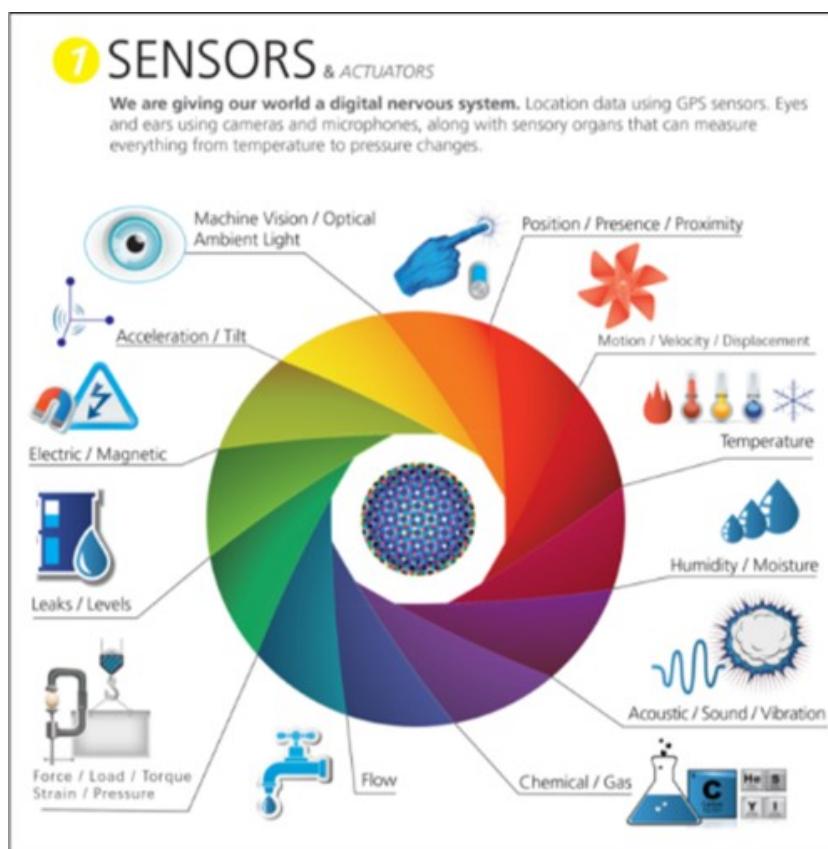


Figura 7. Tipos de sensores por tipo de magnitudes medidas. Fuente: Girard, 2020

Tanto es así, que en las principales plataformas *cloud* para el tratamiento de datos en la nube, como Amazon Web Services, Azure, Google Cloud o IBM Watson, se encuentran módulos específicos para recoger, procesar y almacenar datos IoT. Asimismo, las principales herramientas de visualización de datos, como Power BI, Tableau o Google Data Studio, se encuentran cada vez más especializadas para la visualización de datos provenientes de este tipo de elementos.

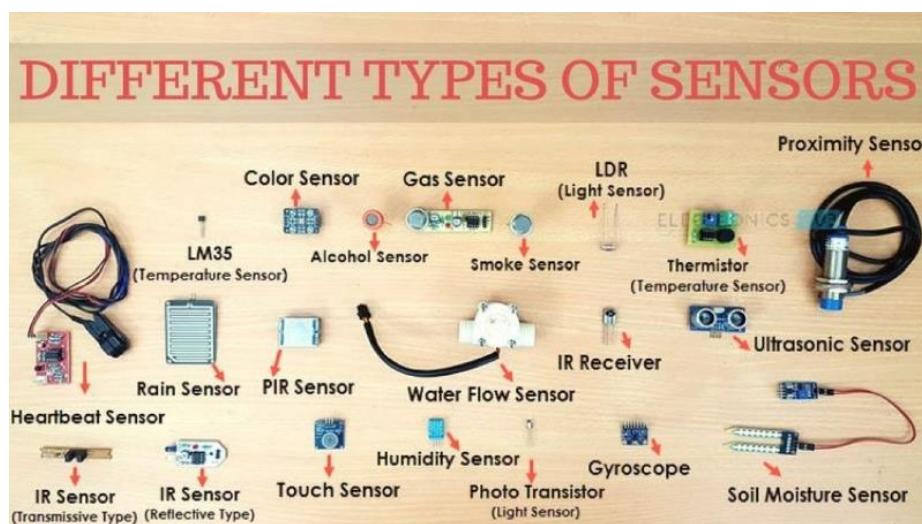


Figura 8. Diferentes tipos de sensores. Fuente: DEWEsoft, 2020.

## ► Acceso a datos públicos, fuentes de datos abiertas y CKAN

En este caso, estamos hablando de fuentes de datos secundarias o bien de fuentes de datos primarias que ha confeccionado el científico de datos mediante algún método de captura de los anteriormente mencionados y en algún momento pasan al ciclo de vida del dato dentro del estudio llevado a cabo.

Estamos hablando de la descarga de *datasets* o conjuntos de datos (generalmente como CSV, XML, Avro, etc., u otro formato de datos apropiado) o a través de interfaces de programación de aplicaciones **A P I** (*application programming interfaces*).

Existen muchas entidades públicas, como gobiernos centrales y locales, así como organizaciones, centros de investigación e incluso empresas privadas o investigadores particulares que publican catálogos de datos para el aprovechamiento a través de analítica de datos y el desarrollo de aplicaciones. La publicación de estos datos de carácter público o privado suele realizarse en un portal dedicado.

Ejemplos notables de portales donde podemos buscar y acceder a *datasets* abiertos para realizar nuestros propios estudios o pruebas de técnicas de ciencia de datos son:

- ▶ [Kaggle.](#)
- ▶ [OpenML.](#)
- ▶ [Google Dataset Search.](#)

A este respecto, existe un **movimiento de datos abiertos** (desde el Open Data Institute) cuyo principal objetivo es que los datos sean publicados bajo una licencia abierta para que cualquier usuario (otra organización o un particular) pueda acceder a los mismos y reutilizarlos con fines públicos o privados (en función del tipo de la licencia de los datos).

Por ejemplo, en el caso de datos gubernamentales, podemos acceder a portales como:

- ▶ En España: <http://datos.gob.es>.
- ▶ En Reino Unido: <http://data.gov.uk>.
- ▶ En Estados Unidos: <http://data.gov>.

Este concepto está estrechamente relacionado con los paradigmas *linked open data (LOD)* y *big and open linked data (BOLD)*, que permiten ingerir datos que no se encontraban previamente vinculados. Mediante este paradigma, se utiliza la web

para reducir las barreras que impiden vincular datos que actualmente están vinculados por otros métodos. Esto permite exponer, compartir y enlazar datos, información y conocimientos en la web semántica utilizando un identificador uniforme de recursos o *uniform resource identifier* (URI) y las especificaciones del marco de descripción de recursos o *resource description framework* (RDF) del Consorcio W3C.

The screenshot shows the homepage of the Junta de Castilla y León open data portal. At the top, there's a search bar with the placeholder "Buscar Datasets" and a red "BUSCAR" button. To the right of the search bar is a "Ayuda" link. Below the search bar, the page title "BÚSQUEDA DE DATOS" is displayed. Underneath, the title "CONJUNTOS DE DATOS DESTACADOS" is shown. A list of six datasets is presented, each with a title in red and three download icons (CSV, XLS, PDF) to its right. The datasets are:

- Situación epidemiológica coronavirus (COVID-19) en Castilla y León por provincias
- Situación epidemiológica coronavirus (COVID-19) en Castilla y León por hospitales
- Situación enfermos por coronavirus detectados en atención primaria
- Situación enfermos por coronavirus detectados en atención primaria por tramos de edad y sexo
- Tasa de enfermos por zonas básicas de salud
- Situación de solicitudes de ERTEs coronavirus (Marzo-Septiembre)

Figura 9. Portal de datos abiertos de la Junta de Castilla y León (España). Fuente: Junta de Castilla y León, s. f.

A este respecto, **CKAN** (*comprehensive knowledge archive network*) es una herramienta web de código abierto ampliamente extendida, especialmente en el ámbito de las entidades públicas, que permite almacenar y publicar datos abiertos para que sean empleados por otros usuarios (otras administraciones, empresas o particulares), que son accedidos a través de un conjunto de API estandarizadas (Herrera-Cubides, Gaona-García y Orjuela, 2017).

Por otro lado, algunas empresas proporcionan acceso público a los datos, uno de cuyos fines es formar parte de una plataforma de desarrollo, como es el caso de los servicios de redes sociales (Facebook o Twitter), que proporcionan a los desarrolladores API para obtener información sobre un usuario concreto o acerca de

las interacciones de dichos usuarios en la red social (*posts*, *likes*, menciones a un tema o *hashtag*, etc.).

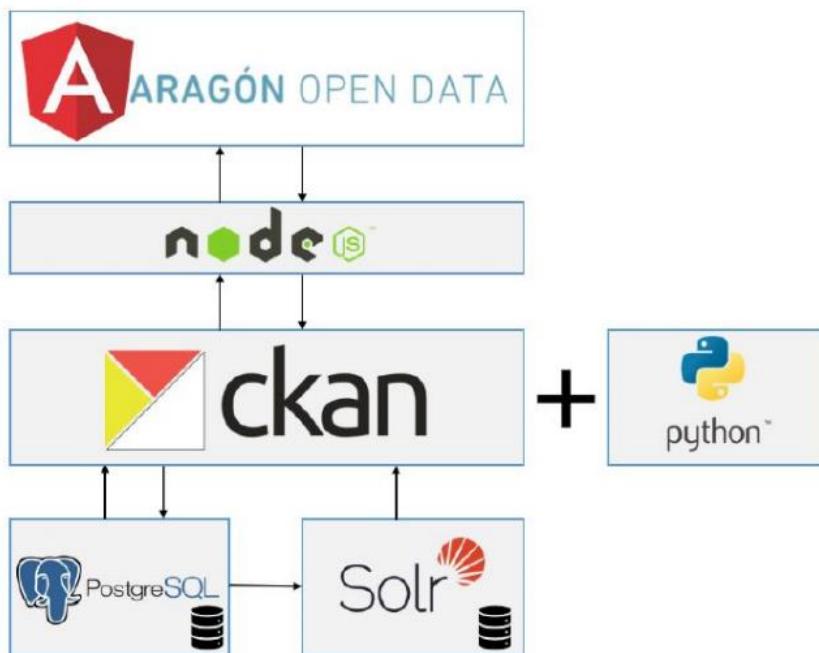


Figura 10. Aragón Open Data mediante tecnología CKAN. Fuente: datos.gob.es, 2020.

## Interfaces de programación de aplicaciones (API)

Dedicamos esta sección a describir con detalle las interfaces de programación de aplicaciones o *application programming interfaces* (API), pues **son empleadas ampliamente en los métodos de captura automatizada de información, así como para que otros procesos software** (en la misma o en distintas máquinas) **interoperen entre sí**. Si bien no es el único método posible de interoperabilidad, en la actualidad es uno de los más empleados.

Las API permiten a una aplicación o *software* determinado (el *software* consumidor) acceder a un conjunto de funcionalidades ofrecidas o proporcionadas por otro *software* (una biblioteca o repositorio que la ofrece, actuando como el *software*

servidor). Esto permite que estas funcionalidades puedan ser reutilizadas sin tener que ser diseñadas, desarrolladas e implementadas de nuevo. Esto permite reutilizar el código, evitar la repetición de funcionalidades ya implementadas y reducir el tiempo y los costes de desarrollo.

Imaginemos un ejemplo clásico. Prácticamente todas las aplicaciones con interfaz gráfica de usuario en un sistema operativo de escritorio como Windows, Mac OS o Linux utilizan elementos como botones, barras de menú o barras de desplazamiento (*scroll*). No tendría sentido que cada desarrollador que creara una nueva aplicación tuviera que diseñar y desarrollar (escribir el código en C++, Python o Java) cada componente desde cero (el aspecto del botón, su comportamiento al pulsarlo, etc.). Si así fuera, el desarrollo de las aplicaciones llevaría una enorme cantidad de tiempo y supondría enormes costes.

En su lugar, se utilizan **librerías o bibliotecas de componentes**, que ya contienen todos los componentes más utilizados, por ejemplo, los botones o los *sockets* (entidades utilizadas para manejar una comunicación entre dos procesos en la misma o diferentes máquinas) a través de Internet. En el caso de Windows, por ejemplo, existe la WinAPI para la interacción entre las aplicaciones y el sistema operativo y el framework .NET, que hace uso de ella y permite que las distintas aplicaciones se comuniquen entre sí. Lo mismo ocurre con otros entornos como Android, iOS, etc.

**Una API es un conjunto de subrutinas, funciones y procedimientos** (metodologías frías en la programación orientada a objetos) **que se utilizan en una biblioteca, para que puedan ser utilizados por otros programadores**. El *software* que utilizan las API está vinculado a cómo se implementan estas funciones internamente en la biblioteca (las API funcionan como una capa de abstracción). Se trata de una mancha y de la introducción de todos los parámetros en cada uno de los campos. Se trata de la información sobre el estado de las cosas, el tiempo y el lugar

en el que se encuentra, cuando se trata de la misma, y cuando se trata del mismo código.

A *l software* que utiliza la API no le importa cómo se implementan estas funcionalidades internamente en la biblioteca, solo lo utiliza y establece algunos parámetros de forma sencilla. Elige el texto del botón, el color y lo que ocurre cuando se pulsa, pero no escribe todo el código.

Hay API que permiten generar funciones de *software* en el mismo ordenador, pero también hay API que permiten desarrollar funciones y establecer comunicación con los programadores, que se utilizan en ordenadores de otros países (a través de una red, por ejemplo, una red local o una red interna). **Este tipo de API son las que realmente nos interesan cuando se trata de ingerir o capturar datos de fuentes externas.**

Existen diferentes formas de hacer que el *software* de diferentes máquinas se comunique entre sí a través de las API. Una de ellas es a través de las arquitecturas orientadas a servicio (SOA) dentro de las arquitecturas cliente-servidor, donde un *software* cliente utiliza los servicios proporcionados por un *software* servidor.

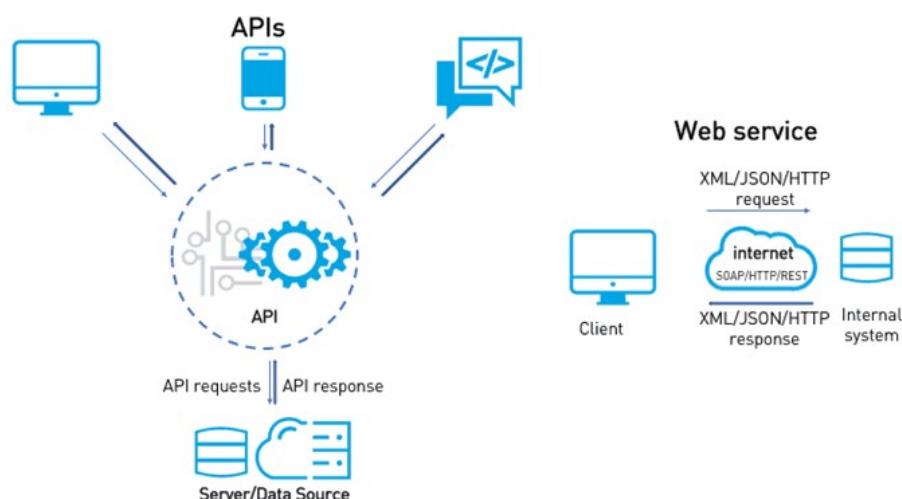


Figura 11. Diferencias entre API y Web Services. Fuente: BeltranC, 2019.

Dentro de las arquitecturas orientadas a los servicios web o *web services*, por un lado, están las API basadas en **SOAP** (*simple object access protocol*), donde los datos se intercambian en forma de XML y donde los servicios se definen mediante WSDL (*web services description language*).

Por otro lado, dentro de los *web services* existen las API REST (*representational state transfer*) o **REST API**, donde los servicios invocados no mantienen ningún estado, es decir, cada invocación de un servicio contiene toda la información que el servidor necesita conocer para realizar las acciones deseadas y devolver el resultado. Las REST API también se centran en los recursos (datos) más que en las acciones o procedimientos (son **no RPC** – *no remote procedure call*, a diferencia de SOAP).

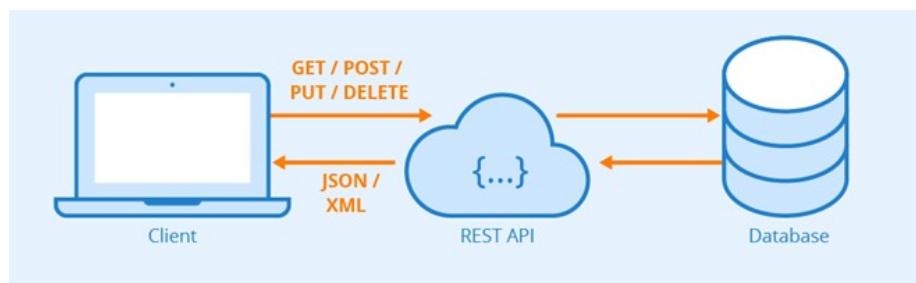


Figura 12. Uso de una REST API por un *software* cliente consumidor de funcionalidades (o datos). Fuente: Seobility, s. f.

Los datos intercambiados al utilizar una REST API pueden ser representados mediante XML o JSON, aunque este último es el formato más utilizado en la práctica por su sencillez. El protocolo utilizado para transportar los datos es **HTTP** (*hyper-text transfer protocol*), el mismo que se utiliza por un navegador web cuando solicita una página a un servidor web remoto. Sin embargo, cualquier *software* puede solicitar información a un servidor que ofrezca una web API de cara a compartir datos.

Asimismo, se utilizan los cuatro *verbos básicos* de HTTP para realizar las operaciones **CRUD** (*create read update delete*):

- ▶ **GET:** utilizado para obtener (leer) un conjunto de datos. Mediante diferentes parámetros se puede filtrar la información a recibir (se usa para ello una *query* tras un signo de interrogación, tal y como lo vemos cuando, por ejemplo, realizamos búsquedas en Google Maps – probar y ver cómo cambia la URL en nuestro navegador – equivalente al **WHERE** en SQL).
- ▶ **POST:** comando para agregar un nuevo recurso en el servidor remoto.
- ▶ **PUT:** permite la modificación de un recurso en el servidor remoto.
- ▶ **DELETE:** como su nombre indica, permite eliminar un recurso o conjunto de recursos en el servidor remoto.

Las REST API fueron empleadas por primera vez por compañías como eBay y Amazon, y son muy utilizadas para recopilar (o añadir, en algunos casos) información de lugares como Google Maps, Google Search, Twitter, Instagram, LinkedIn, etc. También es posible ser empleada para recoger datos IoT de lugares como *the things network*, por poner un ejemplo.

Un ejemplo muy simplificado de llamada para obtener datos de Instagram sería:

HTTP Method: GET

URL: <https://api.instagram.com/v1/users/>

Parameters: user={best\_friends\_user\_id}

En respuesta a esta u otras peticiones, el cliente recibe *respuestas* HTTP del proveedor con un estado de OK (normalmente códigos HTTP 2xx, como 200) o ERROR (normalmente códigos HTTP 4xx o 5xx). Por otra parte, se obtiene información en el cuerpo (*body*) de la respuesta en formato JSON, generalmente. Por ejemplo, la lista de usuarios amigos de un usuario determinado, en el caso de Instagram.

Es posible inspeccionar esta URL (al ser un GET) en el navegador, **pero es habitual que el software que utiliza el servicio necesite autenticación** (a no ser que se traten de datos de solo lectura o abiertos, e incluso en dicho caso puede ser necesaria) mediante algún método apropiado, como **OAuth** (creado por Twitter) o, lo que es común actualmente, **OAuth 2.0** (utilizado por la mayoría de las empresas, como Google, Microsoft, Facebook, Twitter, GitHub, etc.).

En este tipo de métodos de autenticación, la entidad consumidora utiliza en sus llamadas un *token* (similar a una clave muy larga, con unos permisos y duración de dichos permisos determinados) que le ha sido concedido previamente por la entidad que le autoriza a acceder a los datos.

Por otro lado, los métodos como POST y PUT requieren un cuerpo con los datos a enviar, normalmente en formato JSON, lo que no se puede hacer en el navegador sin un complemento o aplicación especial de Chrome. Se pueden especificar parámetros adicionales en la cabecera de la petición HTTP.

Para todos estos fines, incluida la autenticación, es posible probar la API utilizando *software* como curl (una herramienta de línea de comandos para Linux) o Postman (que se ofrece como una aplicación de escritorio o un *plugin/app* para el navegador Chrome).

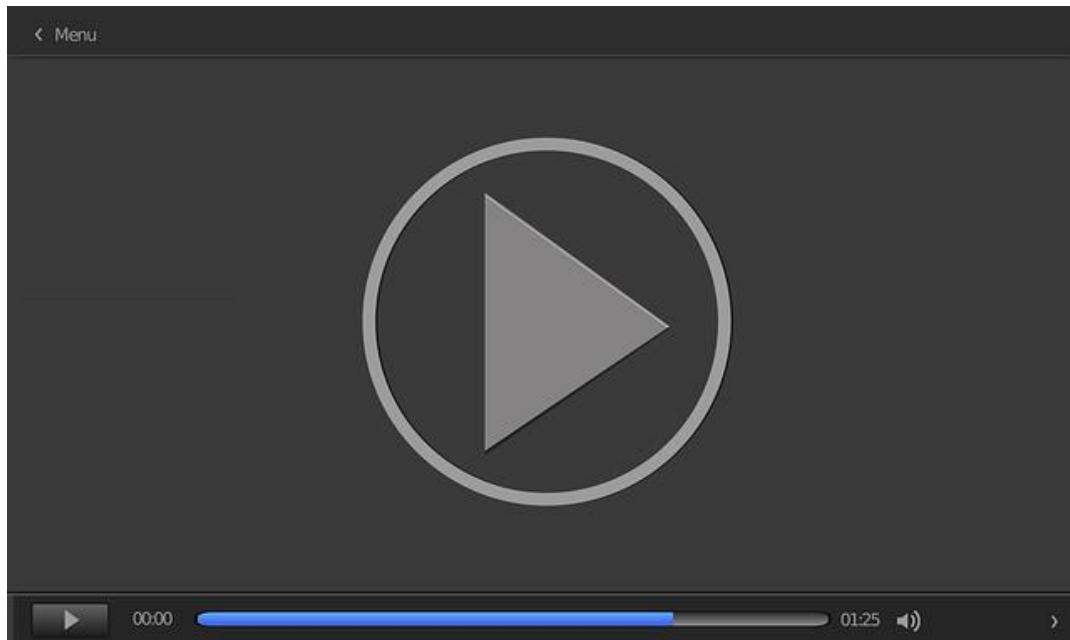
# Ideas clave

The screenshot shows the Postman application interface. On the left, there's a sidebar with sections like 'Collections', 'APIs', 'Environments', 'Mock Servers', 'Monitors', and 'History'. The main area shows a 'Customer APIs' collection with a 'Postman Echo' item. Under 'Postman Echo', there are several request methods: 'GET Request', 'POST Request', 'PUT Request', 'PATCH Request', and 'DELETE Request'. The 'GET Request' is selected, showing a URL: 'https://postman-echo.com/get?id=1&type=vip'. The 'Params' tab is active, displaying 'id' with value '1' and 'type' with value 'vip'. Below the params, there are 'Query Params', 'Body', 'Cookies', 'Headers', and 'Test Results' tabs. The 'Body' tab shows a JSON response with the key-value pairs from the query parameters. The status bar at the bottom indicates 'Status: 200 OK Time: 451 ms Size: 824 B'.

Figura 13. Ejemplo de uso de la aplicación Postman. Fuente: elaboración propia a partir de

<https://learning.postman.com>

En el siguiente vídeo, titulado *Exploración de un dataset público*, se explorarán datos en Kaggle, obteniendo métricas básicas sobre ellos.



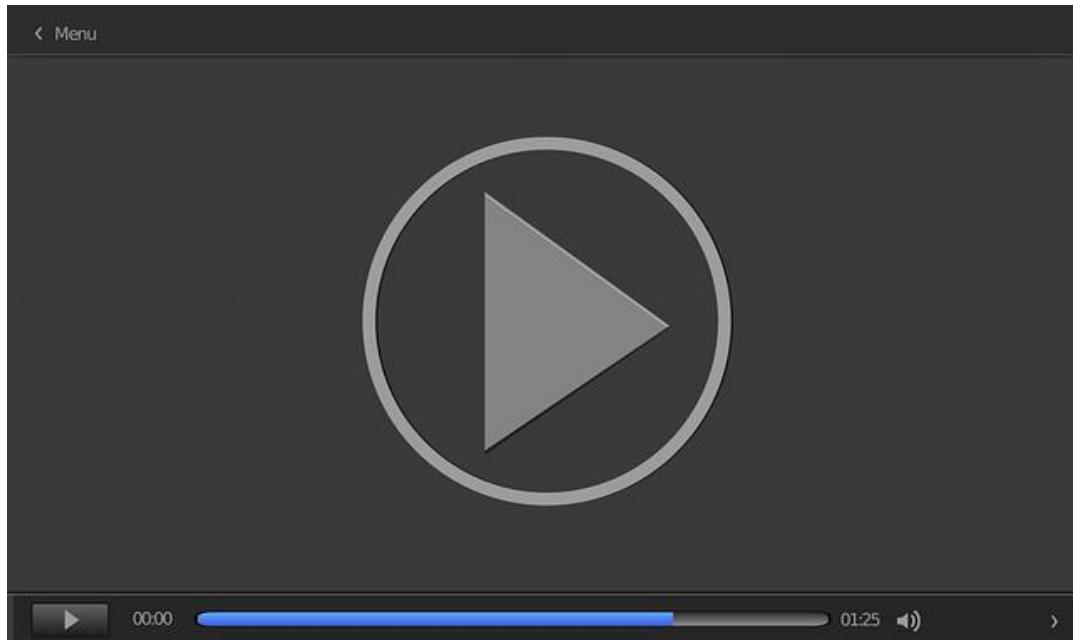
---

Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=4cc8b839-f6ee-424b-a885-b16800f12e12>

---

Por otro lado, en el vídeo *REST API y uso de Postman*, veremos el uso de la herramienta Postman capturando datos meteorológicos mediante una API pública y comprenderemos el funcionamiento de HTTP y las REST API para ingerir datos.



---

Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=ccaaec70-fdd9-4e49-b52b-b16800f12d3d>

---

## 2.3. Análisis

El análisis de datos e información está ampliamente soportado por la estadística. La estadística es uno de los cimientos del análisis de datos, de ahí que en este tema se haga una breve introducción a los conceptos más básicos de esta. Este tema afronta la primera necesidad que plantea el análisis de los datos: su organización y presentación para comprender la información que contienen. Además, se introducen los conceptos de probabilidad condicional y variables aleatorias. Una primera referencia que cubre una gran parte de los conceptos expuestos la tenemos en Ríus (1998).

Una primera aproximación de la estadística, simplificando mucho su definición y ámbito, podría entenderse como una colección de datos cualquiera. Referencias comunes a ella son las estadísticas de ventas de coches, de paro, etc. Sin embargo, esta aproximación informal hace referencia a estudios concretos, no facilitando una visión de la estadística como ciencia que estudia los datos.

De forma más exhaustiva, podría definirse la estadística como la ciencia que maneja los datos a través de un proceso que va desde el diseño del estudio, recogida de los datos, análisis, para finalmente organizar, resumir y mostrar la información contenida en ellos para sacar conclusiones. De forma más resumida, podemos adoptar la definición de Moore (2006):

**La estadística es la ciencia que permite aprender de los datos.**

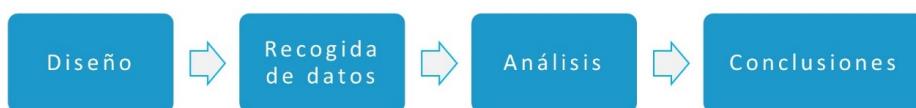


Figura 1. Fases de un estudio estadístico completo. Fuente: elaboración propia.

La ilustra las fases de un estudio estadístico completo: el diseño del estudio estadístico, la recogida de datos, su análisis y, finalmente, la extracción de conclusiones en función de los resultados que se han obtenido del análisis. **Todas las fases de un estudio estadístico son igualmente importantes**, pero es importante destacar que los datos deben recogerse siguiendo unos criterios estadísticos mínimos, siendo la etapa de recogida de datos muy importante y delicada. Existen autores que incluyen una fase extra al inicio de las aquí indicadas: **la identificación del problema de estudio.**

¿Para qué sirve entonces la estadística? Siguiendo nuevamente la definición de Moore, el objetivo de la estadística es «ganar en comprensión de un fenómeno a partir de los datos que se manejan sobre este» (Moore, 2006). Dependiendo del uso que se haga de esta, la estadística podrá ser de dos clases:

- ▶ Estadística descriptiva: se limita a describir una población basándose en la información recogida de su muestra.
- ▶ Estadística inferencial: extrae conclusiones sobre la población de estudio.

En otras asignaturas del curso se brinda información detallada sobre esta etapa del ciclo de vida de los datos masivos.

## 2.4. Visualización

Una de las herramientas más potentes para el análisis de datos es, sin duda, su visualización. En este sentido, dentro de las capas de consumo de datos en una arquitectura para el tratamiento de datos, la capa de presentación de datos, información y conocimiento juega un papel fundamental.

La representación de grandes (y no tan grandes) cantidades de información de forma gráfica facilita en gran medida su comprensión y asimilación. Sin embargo, el proceso de creación de infografías, gráficas u otras visualizaciones no es trivial y requiere un amplio conocimiento del ámbito para poder realizarlo de forma eficiente y eficaz.

La primera pregunta que nos interesa considerar sería «¿por qué visualizar los datos?». **Los humanos están dotados de un potente sistema visual que es capaz de establecer patrones a gran velocidad.** Esta capacidad de percepción inherente a nuestra condición humana nos permite alcanzar altos niveles de conocimiento.

El rápido e imparable aumento del volumen de información en las últimas décadas ha hecho imprescindible acercarse al conocimiento de diversos modos y con diversas técnicas para representarla visualmente y mostrarla al público, de modo que este sea capaz de extraer múltiples conclusiones.

**Recopilar y analizar los datos para finalmente diseñar una visualización  
es el punto final a todo un proceso que tiene su origen en la búsqueda  
del mensaje subyacente a un conjunto de datos.**

La ciencia de datos, especialmente cuando se manejan grandes datos, requiere visualizaciones potentes que ayuden a entender los datos mostrados, a explicar

mejor un concepto o, simplemente, a sintetizar el contenido de un volumen elevado de datos. Algunos ejemplos prácticos los encontramos en la elaboración de mapas, la representación de valores recogidos por estaciones meteorológicas o dispositivos *wearables* en gráficas e infografías en diferentes portales web y aplicaciones móviles o en los cuadros de mando en forma de *dashboards* que nos proponen los nuevos sistemas de *business intelligence*.

## Infografía y visualización de datos

A menudo se habla de los términos infografía y visualización de datos como campos diferentes e independientes el uno del otro. Si bien es cierto que las primeras representaciones infográficas cuentan con varios siglos de antigüedad, las visualizaciones de datos tampoco son un invento reciente. Además, en ocasiones una infografía y una visualización de datos pueden diluirse y dar lugar a una sola estructura con una finalidad común.

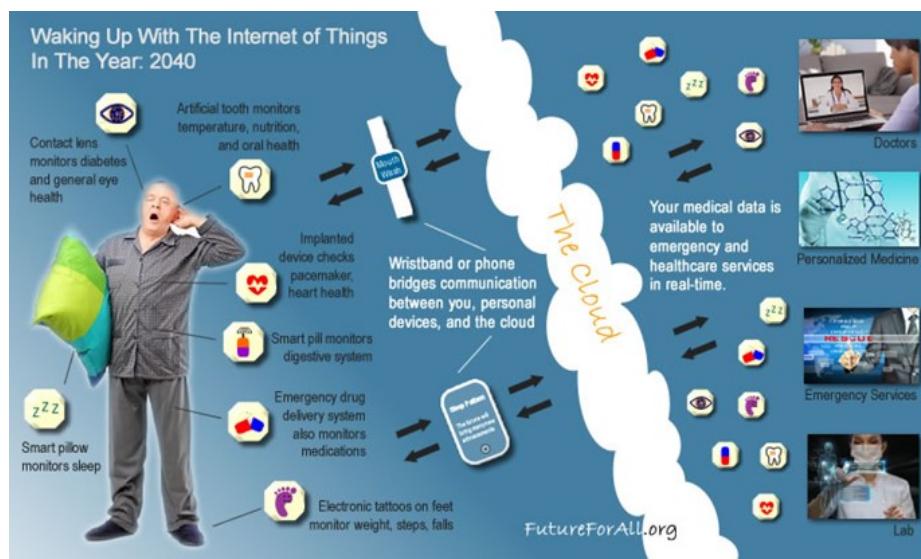


Figura 1. Infografía sobre las aplicaciones IoT en el ámbito del healthcare desarrollada por FutureForAll.

Fuente: <https://www.futureforall.org/images/iot-medical-750.jpg>

- ▶ Una infografía es una forma de representar información a través de la combinación de gráficos (ilustraciones, diagramas, mapas y otros recursos) y texto.
- ▶ Una visualización de datos es un término tradicionalmente más cercano a la comunidad científica con el que referirse a la creación de representaciones visuales (realizadas informáticamente y, normalmente, dotadas de interactividad) de datos abstractos con el fin de que el público pueda analizarlos y ampliar su conocimiento.

Utilizaremos los términos infografía y visualización indistintamente por su naturaleza común y los muchos elementos estructurales que comparten.

Ambos términos responden a un mismo objetivo común: **ayudar a comprender una información e invitar al análisis y la reflexión sobre el mismo**. De este modo, el receptor de la información puede descubrir patrones y relaciones hasta ahora desconocidas, es decir, permite inferir conocimiento.

## Importancia de la infografía y la visualización de datos

¿Por qué visualizar los datos? La infografía y la visualización permiten que el público pueda percibir rápidamente las **relaciones entre los datos**. Uno de los grandes beneficios de la visualización es la cantidad de información pura que puede ser interpretada en décimas de segundo. Esto es posible gracias a que utiliza el mismo lenguaje que nuestro propio sistema cognitivo.

Un ejemplo práctico de aplicación IoT industrial de la importancia de la infografía y la visualización de datos podemos encontrarlo en el análisis de consumos energéticos de las distintas sedes de una industria multinacional.

En primer lugar, se dispone de los datos en bruto, los cuales forman tablas enormes y difíciles de entender.

La representación de estos valores en gráficas de barras nos ayudaría a entender mejor la información, pudiendo apreciar comparaciones y evoluciones, entre otros aspectos.

Por otro lado, podría desarrollarse una infografía consistente en la representación de todas las sedes en un mapa mediante círculos de diferentes áreas en función de su consumo energético.

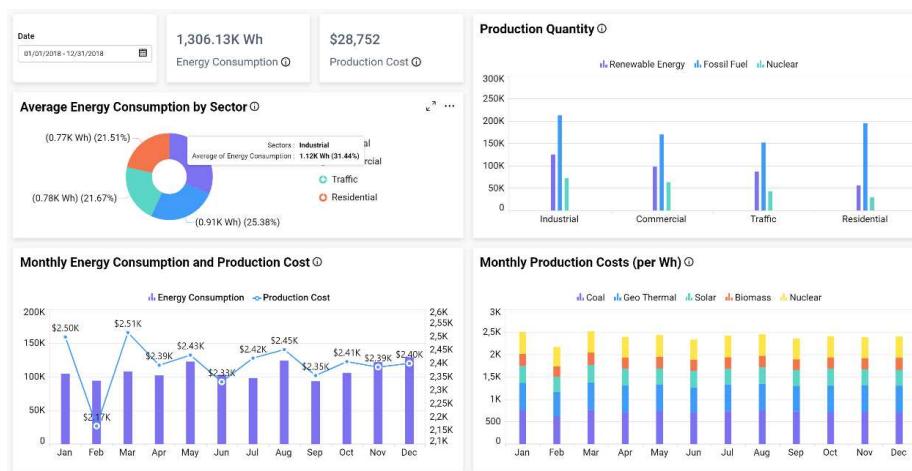


Figura 2. Ejemplo de dashboard para la monitorización de la producción y consumo energético (Bold BI).

Fuente: <https://samples.boldbi.com/solutions/energy/energy-monitoring-dashboard>

En definitiva, infografía o visualización juegan un papel fundamental en el momento actual debido a que **la cantidad de información disponible se multiplica sin cesar y necesita ser filtrada y organizada**. Así, el público podrá informarse desde un enfoque más analítico y dar respuesta a una pregunta o formular cuestiones nuevas.

En otras asignaturas del curso se brinda información detallada sobre esta etapa del ciclo de vida de los datos masivos.

## 2.5. Interpretación

La interpretación de datos es una fase crucial en el ciclo de vida de los datos masivos, donde se evalúan los resultados obtenidos a través de diversas técnicas de análisis y visualización para generar conocimientos aplicables y decisiones estratégicas. Esta etapa no solo se enfoca en entender lo que los datos indican, sino también en contextualizar estos hallazgos dentro de los límites operativos y estratégicos de la organización.

### Características Principales de la Interpretación de Datos

La interpretación de datos se caracteriza por ser un proceso profundamente analítico y crítico, que requiere una combinación de habilidades técnicas y sectoriales. Algunas de las características principales incluyen:

- ▶ **Contextualización:** Los datos nunca existen en el vacío; siempre están influenciados por el entorno en el que se recogen. Entender este contexto es fundamental para interpretar correctamente los datos. Esto incluye tener en cuenta las condiciones económicas, sociales, políticas y tecnológicas que puedan afectar los resultados.
- ▶ **Crítica de métodos:** Parte de la interpretación implica evaluar críticamente los métodos de análisis utilizados. Esto incluye considerar la adecuación del modelo estadístico, la precisión de los algoritmos de machine learning y la integridad de los datos utilizados.
- ▶ **Consideración de limitaciones:** Todo proceso de análisis tiene limitaciones, ya sean relacionadas con la calidad del dato, la metodología de análisis o las inferencias estadísticas. Reconocer estas limitaciones es crucial para la interpretación correcta de los resultados.

## Contexto de los Datos y del Análisis

El contexto en el que se recolectan y analizan los datos puede tener un impacto significativo en la interpretación de estos. Por ejemplo, los datos recolectados durante un evento anómalo, como una crisis económica o un cambio tecnológico disruptivo, deben ser interpretados de manera diferente a los datos recolectados en períodos de estabilidad. Además, el contexto del análisis, incluyendo el objetivo del estudio y las hipótesis planteadas, debe ser considerado para alinear los resultados con las expectativas y necesidades organizacionales.

## Limitaciones en la Interpretación de Datos

Las limitaciones en la interpretación de datos pueden surgir de varias fuentes:

- ▶ **Calidad de los datos:** Datos incompletos, inexactos o sesgados pueden llevar a interpretaciones erróneas.
- ▶ **Metodología de análisis:** La elección de técnicas de análisis inapropiadas para el tipo de datos o la pregunta de investigación puede distorsionar los resultados.
- ▶ **Sobreinterpretación:** Existe el riesgo de hacer inferencias más allá de lo que los datos pueden legítimamente sostener, especialmente cuando se utilizan métodos de análisis complejos como el machine learning.

## Resumen de Técnicas para la Interpretación de Datos

La interpretación efectiva de los datos requiere el uso de técnicas que puedan facilitar la comprensión y evaluación de los resultados. Algunas de estas técnicas incluyen:

- ▶ **Análisis de sensibilidad:** Evalúa cómo diferentes inputs afectan los outputs del análisis, ayudando a identificar variables críticas.
- ▶ **Comparación con benchmarks o estándares de la industria:** Permite contextualizar los resultados dentro de un marco más amplio, comparando los hallazgos con otros similares o con estándares reconocidos.
- ▶ **Técnicas de visualización avanzadas:** Herramientas como mapas de calor, gráficos de árbol y diagramas de red pueden ayudar a interpretar patrones complejos y relaciones entre variables.
- ▶ **Discusión crítica y revisión por pares:** Involucrar a expertos en el campo para revisar los análisis y resultados puede proporcionar nuevas perspectivas y validar la interpretación.

La interpretación de datos es una fase integral y compleja del ciclo de vida de los datos masivos que requiere no solo un profundo entendimiento técnico, sino también una fuerte capacidad para contextualizar y criticar los resultados. A través de una interpretación cuidadosa y considerada, las organizaciones pueden tomar decisiones más informadas y estratégicamente sólidas.

## 2.6. Referencias bibliográficas

Ríus, F. (1998). *Bioestadística: Métodos y aplicaciones*. Universidad de Málaga. Publicaciones.

Guillarranz, M. (14 de mayo de 2019). *El ciclo de vida de los datos: las 5 fases para llevar a éxito un proyecto de Big Data*. PiperLab. <https://piperlab.es/2019/05/14/el-ciclo-de-vida-de-los-datos-las-5-fases-para-llevar-a-exito-un-proyecto-de-big-data/>

Alonso, R. S., Sittón-Candanedo, I., García, Ó., Prieto, J., y Rodríguez-González, S. (2020). An intelligent Edge-IoT platform for monitoring livestock and crops in a dairy farming scenario. *Ad Hoc Networks*, 98.

BeltranC. (3 de mayo de 2019). Diferencia entre API y Servicio web. *BeltranC*. <https://medium.com/beltranc/diferencia-entre-api-y-servicio-web-5f204af3aedb>

Beysolow II, T. (2018). What Is Natural Language Processing? En Autor (ed.), *Applied Natural Language Processing with Python* (pp. 1-12). Apress.

Biscobing, J. (2018, septiembre). *What is data sampling*. Techtarget. <https://searchbusinessanalytics.techtarget.com/definition/data-sampling>

Brajkovic, H., Jaksic, D. y Posicic, P. (2020). Data warehouse and data quality—an overview. En V. Strahonja, W. Steingartner y V. Kirinić (eds.), *Central European Conference on Information and Intelligent Systems CECIIS 2020*. University of Zagreb.

Das, T. K. y Kumar, P. M. (2013). Big data analytics: A framework for unstructured data analysis. *International Journal of Engineering Science & Technology*, 5(1), 153.

Datos.gob.es. (25 de noviembre de 2020). *Cómo implementar CKAN: caso real del portal Aragón Open Data.* <https://datos.gob.es/es/documentacion/como-implementar-ckan-caso-real-del-portal-aragon-open-data>

DEWEISoft. (9 de marzo de 2020). *¿Qué es un sensor y qué hacer?* <https://deweissoft.com/es/daq/que-es-un-sensor>

Garvin, J. (1987). *Sampling methods.* Jongarvin.com. [http://jongarvin.com/up/MPM1D/slides/sampling\\_methods\\_handout.pdf](http://jongarvin.com/up/MPM1D/slides/sampling_methods_handout.pdf)

Girard, M. (8 de septiembre de 2020). *Standards for Cybersecure IoT Devices: A Way Forward.* G20 Insights. [https://www.g20-insights.org/policy\\_briefs/standards-for-cybersecure-iot-devices-a-way-forward/](https://www.g20-insights.org/policy_briefs/standards-for-cybersecure-iot-devices-a-way-forward/)

Giudice, P. L., Musarella, L., Sofo, G. y Ursino, D. (2019). An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake. *Information Sciences*, 478, 606-626.

Hassan, H., Nasir, M. H. M., Khairudin, N. y Adon, I. (2017). Factors influencing cloud computing adoption in small medium enterprises. *Journal of Information and Communication Technology*, 16(1), 21-41.

Herrera-Cubides, J. F., Gaona-García, P. A. y Orjuela, K. G. (2017). A view of the web of data. case study: use of services CKAN. *Ingeniería*, 22(1), 46-64.

Junta de Castilla y León. (s. f.). *Datos abiertos de Castilla y León.* <https://datosabiertos.jcyl.es/web/es/datos-abiertos-castilla-leon.html>

Krizhevsky, A., Sutskever, I. y Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84-90.

Lima, A. C. E., de Castro, L. N. y Corchado, J. M. (2015). A polarity analysis framework for Twitter messages. *Applied Mathematics and Computation*, 270, 756-767.

Ministerio de Educación. (14 de enero de 2019). Cómo elegir RFID pasivo. *Wristbands*. <https://www.wristbandhy.com/es/how-to-choose-passive-rfid-tags/>

Seobility. (s. f.). *Rest API*. [https://www.seobility.net/en/wiki/REST\\_API](https://www.seobility.net/en/wiki/REST_API)

Sint, R., Schaffert, S., Stroka, S. y Ferstl, R. (2009). Combining unstructured, fully structured and semi-structured information in semantic wikis. *CEUR Workshop Proceedings*, 464, 73-87.

Taylor, J. (20 de agosto de 2020). *10 Effective Ways to Data Capture*. Invensis. <https://www.invensis.net/blog/10-effective-ways-to-data-capture/>

Vaidyanathan, P. P. (2001). Generalizations of the sampling theorem: Seven decades after Nyquist. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 48(9), 1094-1109.

Moore, D. S. (2006). *Introduction to the practice of statistics* (5.<sup>a</sup> ed.). Freeman and Company.

## Kaggle

Kaggle. (s. f.). Datasets. <https://www.kaggle.com/datasets>

En esta página web se puede encontrar, entre otros, un listado de conjuntos de datos abiertos y con la calidad requerida como para utilizarlos con fines de investigación. Kaggle incluye más de 30 000 *datasets* para trabajar con ellos, además de incluir otros recursos *machine learning* y permitir compartir nuestros propios *datasets* con la comunidad.

## OpenML

OpenML. (s. f.). <https://www.openml.org/>

En esta página web se puede encontrar, entre otros, un listado de conjuntos de datos abiertos y con la calidad requerida como para utilizarlos con fines de investigación. El portal OpenML, entre otros recursos, incluye más de 20 000 *datasets* abiertos para trabajar con ellos, además de permitir la publicación de *datasets* propios.

## Google Dataset Search

Google. (s. f.). Dataset Search. <https://datasetsearch.research.google.com/>

Google Dataset Search es un motor de búsqueda a través de los metadatos de millones de conjuntos de datos en miles de repositorios en toda la web. En enero de 2020 dejó de ser un servicio en versión beta y actualmente cuenta con más de veinticinco millones de *datasets* indexados.

## DataPortals.org

Data Portals. (s. f.). 590 Data Portals listed. <http://dataportals.org/>

Sitio web que contiene un listado de catálogos de datos abiertos. Los catálogos están organizados por nivel (local, estatal, nacional, etc.) y por grupos.

1. ¿Cómo clasificarías un fichero JSON en función de su organización?

  - A. Datos estructurados.
  - B. Datos semiestructurados.
  - C. Datos completamente estructurados.
  - D. Ninguna de ellas.
  
2. ¿Cómo clasificarías una base de datos relacional en función de su organización?

  - A. Datos estructurados.
  - B. Datos semiestructurados.
  - C. Datos completamente estructurados.
  - D. Ninguna de ellas.
  
3. ¿Cuáles de los siguientes serían ejemplos de datos no estructurados?

  - A. Imágenes, vídeos y sonidos.
  - B. Bases de datos SQL y OLAP.
  - C. Archivos JSON y XML.
  - D. Archivos CSV.
  
4. Si estoy creando un instrumento para capturar secuencias de voz humana para su posterior procesamiento, ¿a qué frecuencia mínima debería muestrear el audio para no tener *aliasing*?

  - A. 10 Hz.
  - B. 4 KHz.
  - C. 8000 Hz.
  - D. 16 KHz.

5. ¿En qué método de muestreo probabilístico se crean subconjuntos de los conjuntos de datos o de la población en función de un factor común y se recogen muestras al azar de cada subgrupo?
- A. Muestro multietapa.
  - B. Muestro sistemático.
  - C. Muestreo por conglomerados.
  - D. Muestro estratificado.
6. En el método de muestreo no probabilístico intencional:
- A. Los datos se recogen de un grupo fácilmente accesible y disponible.
  - B. Los datos se recogen de cada sujeto que cumple los criterios hasta alcanzar el tamaño de muestra predeterminado.
  - C. El investigador selecciona los datos a muestrear basándose en criterios predefinidos.
  - D. El investigador garantiza una representación equitativa dentro de la muestra para todos los subgrupos del conjunto de datos o población.
7. Los datos recogidos por el propio científico de datos mediante encuestas se considerarían:
- A. Fuentes de datos primarias.
  - B. Fuente de datos secundarias.
  - C. Fuentes de datos terciarias.
  - D. Ninguna de las demás respuestas es correcta.

**8.** El método de captura de datos de forma automatizada mediante el procesamiento de páginas HTML de un sitio web se conoce como:

- A. Web semántica.
- B. *Web service*.
- C. *Web scraping*.
- D. Ninguna de las demás respuestas es correcta.

**9.** ¿En qué categoría de captura o fuente de datos encaja la lectura de información del pulso cardíaco por una pulsera de actividad?

- A. Captura manual.
- B. Sensores.
- C. Captura automatizada.
- D. B y C son correctas.

**10.** La infografía y la visualización de datos tienen como objetivo principal:

- A. Presentar la información de una manera muy atractiva.
- B. Informar y ampliar el conocimiento.
- C. Mostrar una información diferente al lector.
- D. Buscar y organizar datos.

Ciencia de Datos Aplicada

---

# Tema 3. Arquitecturas Típicas en Proyectos de Datos Masivos

# Índice

[Esquema](#)

[Ideas clave](#)

[3.1. Introducción y objetivos](#)

[3.2. Fuentes heterogéneas](#)

[3.3. Extracción, Transformación y Carga](#)

[3.4. Almacenamiento](#)

[3.5. Tratamiento de los Datos](#)

[3.6. Visualización](#)

[3.7. Referencias bibliográficas](#)

[A fondo](#)

[Google Dataset Search](#)

[What is Object Storage?](#)

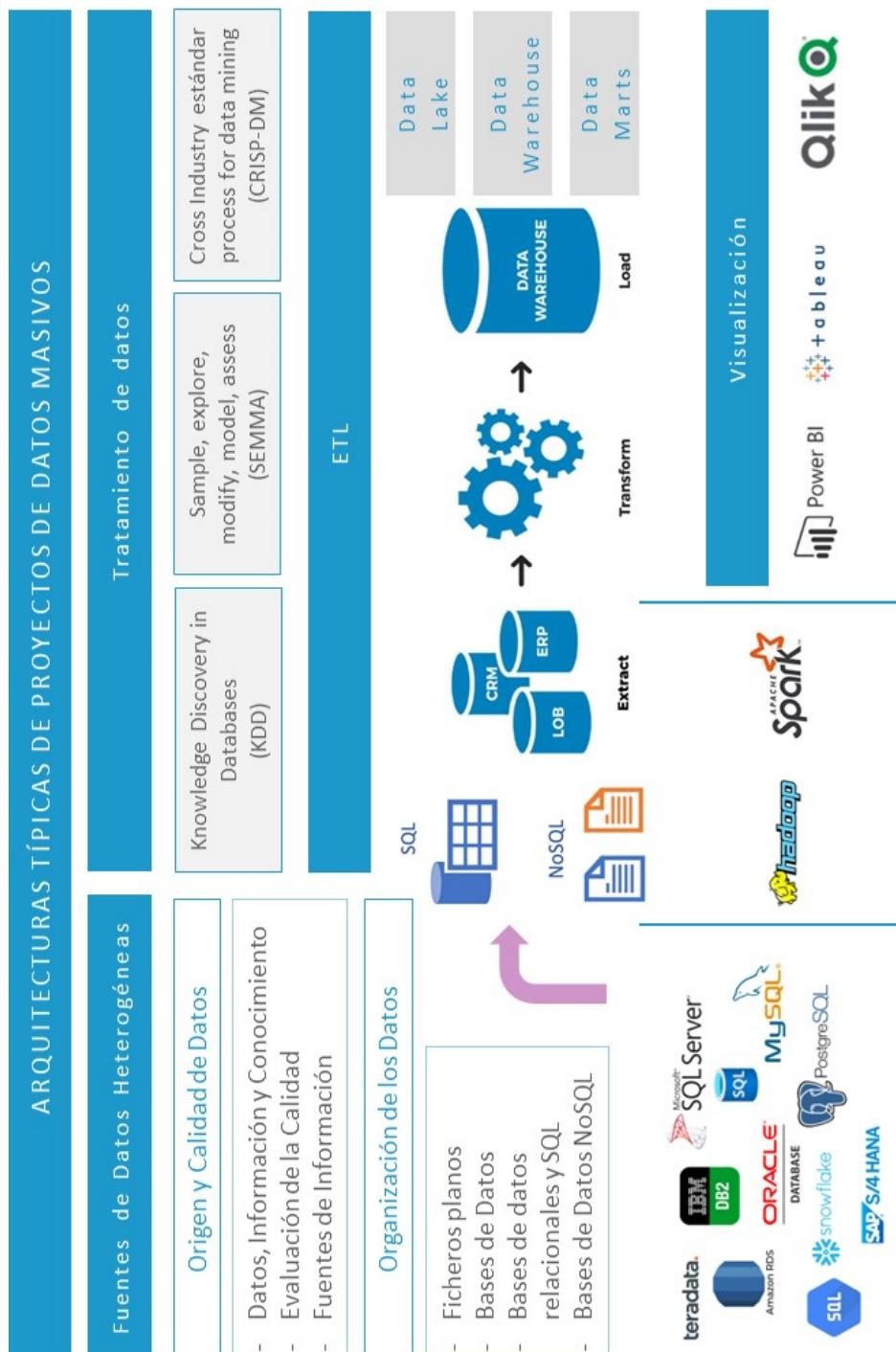
[SQL vs NoSQL or MySQL vs MongoDB](#)

[Preparar los datos y convertirlos en información](#)

[Procesos ETL. Enrique Onieva](#)

[Herramientas de procesado y visualización de datos](#)

[Test](#)



## 3.1. Introducción y objetivos

En el campo del Big Data y la Ciencia de Datos, la correcta estructuración de arquitecturas es fundamental para administrar eficientemente grandes volúmenes de datos, garantizando así la eficacia y escalabilidad de los proyectos analíticos. Estas arquitecturas se componen de diversos componentes interrelacionados que facilitan desde la recopilación hasta el análisis avanzado de datos. La claridad en la organización de estos elementos permite a las organizaciones maximizar el valor de sus datos, transformando la información cruda en insights operativos y estratégicos. En este tema, el estudio detallado de cada componente y su función dentro del sistema general de datos es esencial para cualquier profesional del ámbito. Este tema de introducción tiene como objetivo proporcionar una comprensión clara y estructurada de las arquitecturas de datos masivos en proyectos de Big Data y Ciencia de Datos.

Los objetivos de este tema de introducción son:

- ▶ **Comprender el Rol de las Arquitecturas en Big Data y Ciencia de Datos :** es importante comprender cómo las arquitecturas de Big Data facilitan la gestión y el análisis de grandes conjuntos de datos. Este entendimiento incluye conocer las características de las arquitecturas como su capacidad para manejar la velocidad, volumen y variedad de datos, y cómo estos factores influyen en la selección y diseño de la arquitectura apropiada para diferentes proyectos.
- ▶ **Identificar los Componentes Clave de las Arquitecturas de Datos Masivos :** este objetivo se enfoca en introducir a los estudiantes a los diversos componentes de una arquitectura típica de Big Data, desde las fuentes heterogéneas de datos hasta los sistemas de almacenamiento y procesamiento de datos. Los estudiantes aprenderán la importancia de cada componente y cómo interactúan entre sí para crear un sistema eficiente y escalable.

## ► Reconocer la Importancia de las Estrategias de Integración y Análisis de

**Datos:** finalmente, es crucial reconocer la importancia de las técnicas de integración de datos, como los procesos ETL (*extract, transform, load*), y de análisis para convertir los datos en información útil. Comprender estas estrategias permitirá a los estudiantes apreciar cómo la arquitectura de datos apoya la toma de decisiones basada en evidencias y la generación de conocimiento dentro de las organizaciones.

Al cubrir estos objetivos, el tema pretende dotar a los estudiantes de una base que le permita entender el papel de las arquitecturas de datos masivos y la importante de su implementación en contextos reales de Big Data y Ciencia de Datos.

## 3.2. Fuentes heterogéneas

El primer paso en cualquier arquitectura de Big Data es la integración de datos provenientes de fuentes heterogéneas. Esto incluye datos estructurados, semiestructurados y no estructurados, originados desde diversas plataformas como redes sociales, sensores, registros de transacciones, y más. La capacidad de amalgamar y manejar esta diversidad es crucial para un análisis de datos comprensivo.

En el contexto de las arquitecturas de Big Data, las fuentes heterogéneas representan uno de los componentes fundamentales y más desafiantes del sistema. Las fuentes heterogéneas se refieren a la variedad de formatos y tipos de datos que un sistema de Big Data puede necesitar procesar y analizar. Estos datos pueden provenir de múltiples orígenes, cada uno con sus propias características y desafíos específicos. La capacidad de integrar y procesar eficazmente datos de fuentes heterogéneas es crucial para el éxito de cualquier proyecto de Big Data, ya que permite a las organizaciones obtener una visión completa y detallada de su entorno operativo y de mercado.

### Principales Fuentes de Datos Heterogéneos

#### Datos Estructurados

- ▶ Descripción: Los datos estructurados están organizados en formatos predefinidos, típicamente en tablas con filas y columnas, como bases de datos SQL, hojas de cálculo y otros sistemas de gestión de bases de datos.
- ▶ Ejemplos de Uso: Bancos y entidades financieras utilizan datos estructurados para almacenar información transaccional y de clientes, lo cual es crucial para operaciones, análisis de crédito, y cumplimiento regulatorio.

## Datos Semiestructurados

- ▶ Descripción: Los datos semiestructurados no tienen una estructura rígida como los datos estructurados, pero contienen etiquetas o marcas que permiten agrupar y jerarquizar la información, facilitando su análisis.
- ▶ Ejemplos de Uso: Documentos XML y JSON utilizados en aplicaciones web y móviles, que son esenciales para el intercambio de datos en aplicaciones de comercio electrónico y plataformas de redes sociales.

## Datos No Estructurados

- ▶ Descripción: Los datos no estructurados son formas de información que no se ajustan a modelos de datos convencionales y no se organizan en forma de filas y columnas. Incluyen texto, imágenes, audio y video.
- ▶ Ejemplos de Uso: Las empresas de medios y entretenimiento analizan contenido de video y audio para recomendaciones personalizadas y análisis de sentimientos. Las redes sociales analizan publicaciones y comentarios para obtener insights sobre las preferencias y comportamientos de los usuarios.

## Datos en Tiempo Real

- ▶ Descripción: Los datos en tiempo real se generan continuamente y necesitan ser procesados rápidamente después de su creación para maximizar su valor y utilidad.
- ▶ Ejemplos de Uso: Sensores en la industria de la manufactura para monitorizar el rendimiento de las máquinas y predecir fallos antes de que ocurran, o datos de localización GPS en servicios de logística para optimizar rutas de entrega en tiempo real.

## Datos de Redes Sociales

- ▶ Descripción: Datos generados por usuarios en plataformas de redes sociales, que incluyen texto, enlaces, imágenes, y videos, junto con metadatos como ubicación, tiempo e interacciones entre usuarios.
- ▶ Ejemplos de Uso: Empresas de marketing y publicidad utilizan estos datos para analizar tendencias, conducta del consumidor y efectividad de campañas publicitarias.

El manejo eficiente de estas fuentes heterogéneas requiere tecnologías y estrategias especializadas, como sistemas de gestión de datos que puedan adaptarse a la diversidad de formatos y la escala de los datos. Los sistemas de ETL, plataformas de Big Data como Hadoop y herramientas de análisis en tiempo real son esenciales para transformar estos datos en información útil y accionable.

Al profundizar en estos aspectos, las organizaciones pueden aprovechar plenamente el potencial de sus datos para impulsar la innovación y mantener una ventaja competitiva en el mercado.

### 3.3. Extracción, Transformación y Carga

El ETL (*Extraction, transform and load*: extracción, transformación y carga), mostrado de forma simplificada en la Figura 1, es un proceso consistente en la siguiente secuencia típica:

- ▶ **Extraer** los datos de diferentes fuentes, generalmente heterogéneas (bases de datos relacionales, bases de datos NoSQL, archivos, REST API, etc.).
- ▶ **Transformar** dichos datos (realizando cálculos, limpiando datos, completando elementos ausentes, eliminando duplicados, fundiendo información, etc.).
- ▶ **Cargar** los datos en un sistema de almacenamiento de datos final, conocido como *data warehouse system*.

Existen diferentes ejemplos de **herramientas ETL** como Amazon Redshift, Marklogic o las soluciones de Oracle.

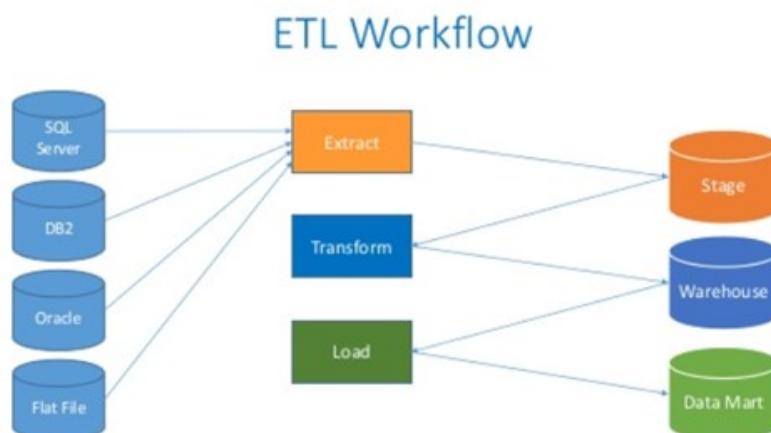


Figura 1. Esquema del proceso ETL básico. Fuente:

[https://fr.wikiversity.org/wiki/Int%C3%A9gration\\_de\\_donn%C3%A9es/Les\\_principales\\_approches\\_d%27int%C3%A9gration\\_de\\_donn%C3%A9es#/media/Fichier:KrisangelChap2-ETL.png](https://fr.wikiversity.org/wiki/Int%C3%A9gration_de_donn%C3%A9es/Les_principales_approches_d%27int%C3%A9gration_de_donn%C3%A9es#/media/Fichier:KrisangelChap2-ETL.png)

## Extracción

En el proceso de extracción los datos necesarios se analizan y se extraen de una o más fuentes diferentes, como sistemas de bases de datos y aplicaciones. Es posible que el tamaño de los datos extraídos sea siempre distinto, puede ser desde cientos de KB hasta GB, dependiendo totalmente del sistema de la fuente y de las situaciones de negocio.

El objetivo principal de este proceso es extraer todos los datos deseados del sistema fuente con los menores recursos posibles. El proceso de extracción debe construirse de manera que no afecte negativamente al sistema fuente en cuanto a rendimiento, tiempo de respuesta o cualquier tipo de bloqueo. Existen dos métodos principales de extracción:

### Métodos de extracción lógica

La extracción lógica se puede categorizar en tres métodos:

- ▶ **Notificación de actualización.** Si el sistema fuente notifica que un registro ha sido modificado y explica los cambios, esta es la forma más simple de obtener los datos.
- ▶ **Extracción incremental.** Puede ser que algunos sistemas no sean capaces de proporcionar una notificación de que se ha producido un cambio, pero pueden averiguar qué registros se han actualizado y proporcionar un extracto de esos registros.

- ▶ **Extracción completa.** Puede ser que algunos sistemas no sean capaces de averiguar qué datos han sido actualizados en absoluto, por lo que una extracción completa es la única manera de obtener los datos del sistema. La extracción completa depende de mantener una copia de la última extracción en el mismo formato para poder averiguar las modificaciones. La extracción completa también maneja las eliminaciones.

## Métodos de extracción física

Por su parte, existen dos tipos de extracción física:

- ▶ **Extracción en línea (*online*).** Los datos se recuperan directamente del propio sistema fuente. El proceso de extracción puede conectarse directamente al sistema de origen para acceder a las tablas de origen o a un sistema central que almacena los datos de una manera predefinida (por ejemplo, registros de instantáneas o tablas de cambios). El sistema central no difiere físicamente del sistema fuente.
- ▶ **Extracción fuera de línea (*offline*).** Los datos no se extraen directamente del sistema fuente, sino que se organizan explícitamente fuera del sistema fuente original. Los datos ya tienen una estructura existente (como, por ejemplo, registros de rehacer, registros de archivo) o fueron creados por una rutina de extracción. Cuando se utilizan extracciones incrementales o completas, la recurrencia de la extracción es extremadamente importante. Especialmente para las extracciones completas. Los volúmenes de datos pueden ser de decenas de GB.

## Transformación

En este proceso los datos se transforman a un formato aplicable que puede ser fácilmente almacenado en un sistema de almacén de datos. El proceso de transformación se asocia con la aplicación de cálculos, operaciones DML (*Data Manipulation Language*, como puede ser SQL), uniones, restricción, clave primaria y claves externas en los datos. Por ejemplo, si se desea el promedio de la anualidad total, se aplicará la fórmula de promedio en la transformación y se cargarán los datos. En el caso de algunos datos no es necesario realizar ninguna transformación que se pueda trasladar directamente al almacén de datos. Dichos datos se conocen como *datos de traslado directo o de paso*.

El proceso de transformación de datos también implica la **corrección** de estos, la **limpieza**, la **eliminación** de los incorrectos y duplicados, la **formación** de datos incompletos y la **corrección de los errores** de estos, su **integridad** y el **formateo** de los que son incompatibles antes de cargarlos en el sistema del *data warehouse*.

## Carga

El proceso final del ETL es la carga de datos en la estructura multidimensional objetivo. En este proceso, los datos extraídos y transformados se almacenan en las estructuras dimensionales a las que acceden los usuarios finales y los sistemas de aplicación.

El proceso de carga incluye tanto la carga de tablas de dimensiones como la carga de tablas de hechos. Es importante asegurarse de que la carga se realice correctamente y con la menor cantidad de recursos posible. El destino principal es una base de datos del proceso de carga. Para que todo el proceso sea eficiente, es beneficioso deshabilitar cualquier restricción e índice antes de la carga y habilitarlos de nuevo solo después de que el proceso se complete. Para garantizar la coherencia, la integridad referencial debe mantenerse mediante el proceso de carga.

Se pueden considerar los siguientes tipos de carga:

- ▶ Carga inicial (*initial load*). Es decir, poblando todas las tablas del *data warehouse*.
- ▶ Carga incremental (*incremental load*), aplicando periódicamente cambios continuos según sea necesario.
- ▶ Refresco completo (full refresh), borrando el contenido de una o más tablas y recargando con datos frescos.

## 3.4. Almacenamiento

El almacenamiento en el contexto del proceso de Extracción, Transformación y Carga (ETL) es un componente esencial que desempeña un papel crucial en la arquitectura de Big Data. Tras la extracción de datos de diversas fuentes heterogéneas y su transformación para garantizar su calidad y formato adecuado, el almacenamiento adecuado de estos datos es fundamental para facilitar el acceso eficiente y el análisis posterior.

El almacenamiento implica la selección y uso de sistemas de almacenamiento que pueden mantener grandes volúmenes de datos de forma segura y eficiente, permitiendo un acceso rápido y confiable. Este almacenamiento debe ser capaz de soportar las operaciones de carga de datos transformados y posibilitar su recuperación para análisis futuros. Es un eslabón crítico que afecta tanto el rendimiento de los procesos de ETL como la eficiencia de las operaciones de consulta y análisis que siguen.

Para comprender mejor este proceso, conviene identificar en primer lugar los términos *bases de datos*, *data lake* y *data warehouse*, así como algunos conceptos derivados de los mismos. Ambos términos se refieren a tipos de repositorios de datos, pero con diferentes características (Vassiliadis y Simitsis, 2002).

### Tipos de Almacenamientos

#### Bases de Datos Relacionales (RDBMS)

Los sistemas de bases de datos relacionales son ampliamente utilizados para almacenar datos estructurados. Son ideales para datos que necesitan ser almacenados en formatos normalizados con relaciones estrictas.

La integridad de los datos y la capacidad de realizar consultas complejas con SQL son puntos fuertes. Además, los RDBMS son ampliamente soportados y entendidos en la industria.

## **Data lake (lago de datos)**

Almacenan datos en bruto de todo tipo y tamaño (no estructurados, semiestructurados y estructurados), generalmente sin límite de capacidad en la práctica. Este tipo de datos se encuentran disponibles tan pronto como son creados y permiten llevar a cabo análisis de datos sobre los mismos. Siguen una estructura de procesamiento *schema on read*, es decir, la estructura de los objetos de la base de datos no está definida previa a la carga y no existe, por lo tanto, un proceso de validación contra estructura alguna.

La estructura de la base de datos se define en la lectura, con la flexibilidad de poder cambiar la estructura en función de los datos que se desean obtener en cada tipo de lectura (este tipo de procesamiento es el más habitual en las arquitecturas big data).

## **Data swamps (pantanos de datos)**

Cuando las entidades almacenan de forma desorganizada datos en los *data lakes*, estos corren el riesgo de convertirse en *data swamps*. Esto sucede cuando se almacenan continuamente datos sin metadatos que ayuden a describir el contenido de los datos almacenados, si se almacenan continuamente datos irrelevantes, si no existe una gobernanza de datos claramente definida (cómo se tratan y quién los gestiona en la organización), si no existen procesos de automatización en la gestión o si no se consigue llevar a cabo una estrategia de limpieza en el tiempo.

## **Data warehouse (almacén de datos)**

Son repositorios de datos altamente estructurados y previamente procesados y que varían de forma más lenta en el tiempo en comparación con un *data lake*. En este caso, el diseño del esquema de datos se establece previamente al comienzo del almacenamiento de datos. Siguen así una estructura de procesamiento *schema on write*, es decir, se necesita que la estructura del objeto de base de datos esté definida de forma previa a la carga y los datos estén validados contra dicha estructura (esta es la forma habitual de funcionar en las BBDD relacionales tradicionales).

## **Data mart (mercado de datos)**

Como los almacenes de datos son muy grandes y llevan un tiempo y esfuerzo en crearse, se pueden crear entretanto *data marts*, son más pequeños que los anteriores y están destinados a conservar los datos de una parte de la organización (es decir, un departamento de la empresa o entidad), mientras que el almacén de datos guarda los de toda la organización.

Estos mercados de datos pueden ser construidos por separado. Otra alternativa es partir de un *data warehouse* existente y que una parte de dicho almacén de datos se destine a una funcionalidad o departamento específicos. En ese caso, puede extraerse esta parte concreta para crear un *data mart* (y así para cada departamento).

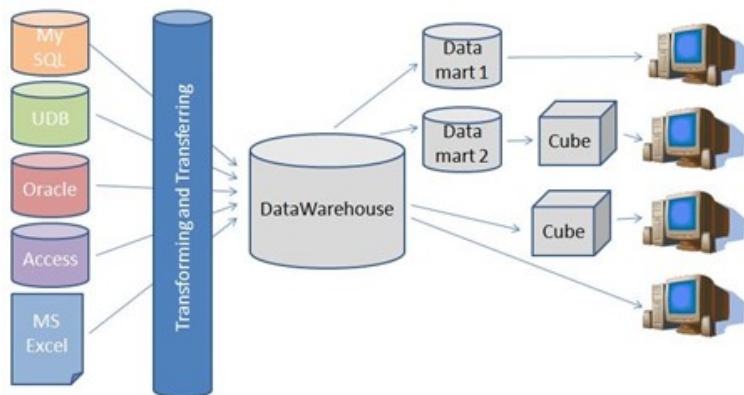


Figura 2. *Data warehouses, data marts* y cubos o hipercubos. Fuente:

<https://upload.wikimedia.org/wikipedia/commons/3/33/Etl2.JPG>

Las diferencias entre los *data lakes* y *data warehouses* se resumen en la tabla 1, según el trabajo de Khine y Wang (2018).

COMPARACIÓN	DATA LAKE	DATA WAREHOUSE
Datos	Datos estructurados Datos semiestructurados Datos no estructurados Datos en crudo Datos sin procesar	Datos estructurados Datos procesados
Procesamiento	<i>Schema on read</i>	<i>Schema on write</i>
Almacenamiento	Almacenamiento de bajo coste	Costoso y confiable
Agilidad	Configuración flexible y ágil	Configuración restringida y poco ágil
Seguridad	En maduración	Madura
Usuarios	Científicos de datos	Profesionales de los negocios

Tabla 1. Comparación entre *data lakes* y *data warehouses*. Fuente: Khine y Wang (2018).

Habiendo aclarado estos conceptos, se describen a continuación los procesos de extracción, transformación y carga de acuerdo con la descripción de Vyas y Vaishnav (2017).

## 3.5. Tratamiento de los Datos

**Los datos deben ser transformados para añadirles valor y convertirlos en información.** Recordemos que estas transformaciones incluyen métodos como:

- ▶ Contextualización: conocer el propósito del dato obtenido.
- ▶ Categorización: conocer la unidad de medida y los componentes del dato.
- ▶ Cálculo: realizar una operación matemática sobre el dato.
- ▶ Corrección: eliminar errores del dato.
- ▶ Agregación: resumir o minimizar un dato de forma más concisa.

Recordemos, por otra parte, que:

**El conocimiento implica una combinación de experiencias, información contextual y relevancia sobre cierta información.**

Así como la información se genera a partir de datos, **el conocimiento surge de la agregación de información.** Ejemplos de métodos que generan esta transformación son:

- ▶ Comparación: relación entre información obtenida en distintas experiencias.
- ▶ Repercusión: implicación de la información en decisiones y acciones.
- ▶ Conexión: relación entre distintos tipos de información.
- ▶ Conversación: opinión de otras personas sobre la información.

Así, antes de poder extraer conocimiento a partir de la información, esta ha de encontrarse debidamente formateada y correlacionada a partir de los datos capturados y almacenados (ingeridos, al fin y al cabo).

Esto requiere una **serie de tareas de tratamiento de datos que incluyen la preparación de ellos**, para lo cual es necesario llevar a cabo un proceso de recogida de datos (a partir de diferentes fuentes de datos ya ingeridas, es decir, hablamos ya de datos capturados y almacenados), el descubrimiento de datos y elaboración de perfiles, la limpieza de los datos (etapa fundamental), la estructuración de los datos en un formato unificado, la transformación y enriquecimiento de los mismos y, finalmente dentro de la preparación, la validación y publicación de los datos.

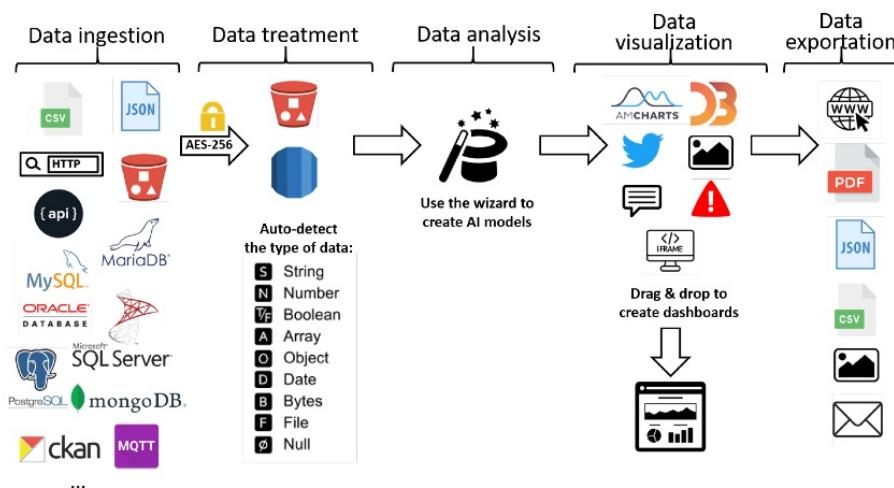


Figura 2. Ejemplo de plataforma Deepint para el tratamiento de datos en sus diferentes etapas. Fuente: Corchado et al., 2021.

Así, es habitual que existan plataformas que ya proporcionen a los usuarios funcionalidades orientadas a la gestión de la información contenida en las fuentes de datos (ver figura 2 y figura 3). En este tipo de plataformas, el proceso de preparación de los datos se realiza en parte de forma automatizada o asistida para el usuario, precisamente porque el tratamiento y, en particular, la preparación de los datos es una tarea altamente onerosa para el científico de datos (ver figura 3).

En herramientas como esta (Corchado et al., 2021), la plataforma puede detectar automáticamente el tipo de datos y el formato (para datos decimales o fechas), por lo que el usuario no tiene que dedicar tiempo a especificarlo. Sin embargo, en el caso de determinados gráficos o modelos puede ser importante o necesaria la asistencia del usuario de cara a especificar el tipo de datos de forma que la herramienta permita al usuario especificarlo manualmente o cambiar el tipo que se ha detectado automáticamente.

También permiten al usuario generar características a partir de campos existentes utilizando expresiones definidas por el usuario. En este punto del flujo, **es posible la creación de fuentes de datos derivadas a partir de fuentes de datos existentes**. Más concretamente, es habitual que se puedan realizar diferentes tipos de operaciones sobre las distintas fuentes de datos, como filtros sobre registros o parámetros de una fuente de datos o la fusión de dos fuentes de datos con los mismos parámetros, entre otras.

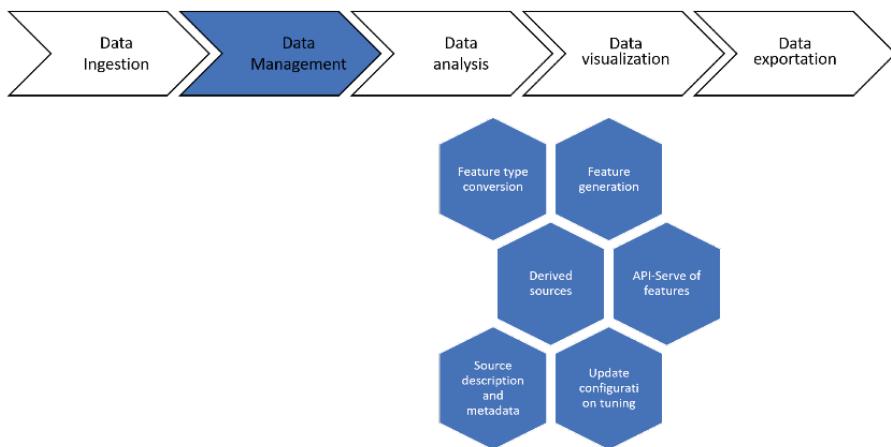


Figura 3. Etapa de tratamiento de datos en la plataforma de ejemplo Deepint . Fuente: Corchado et al., 2021.

## La minería de datos y el descubrimiento de conocimientos en bases de datos

Con el fin de entender mejor por qué son necesarias las diferentes operaciones de tratamiento y preparación de los datos, en el marco del conocimiento de esta materia es importante conocer las etapas empleadas para descubrir conocimiento en bases de datos (u otro tipo de repositorio o conjunto de datos).

**Esto es lo que se conoce como minería de datos.** El término «descubrimiento de conocimiento en bases de datos», del inglés *knowledge discovery in databases* (KDD), es frecuentemente utilizado como sinónimo de la minería de datos (Fayyad, Piatetsky-Shapiro y Smyth, 1996; Fayyad y Stolorz, 1997).

Sin embargo, **este término se refiere al procedimiento completo necesario para extraer conocimiento potencialmente útil y previamente desconocido** a partir de los datos en una base de datos. El KDD es un proceso iterativo que incluye etapas previas a la fase de minería de datos propiamente dicha, para la extracción y preparación de los datos, así como etapas posteriores para el análisis de los resultados y toma de decisiones.

El término «descubrimiento de conocimientos en las bases de datos», o KDD para abreviar, se acuñó en 1989 para referirse al amplio proceso de búsqueda de conocimientos en los datos y para destacar la aplicación de alto nivel de determinados métodos de minería de datos (Fayyad et al, 1996).

Fayyad considera la minería de datos (MD) como una de las fases del proceso de KDD y considera que la fase de minería de datos se refiere, principalmente, a los medios por los que se extraen y enumeran los patrones a partir de los datos (Azevedo y Santos, 2008).

En esta sección se trata del proceso global de KDD, así como las particularizaciones SEMMA y CRISP-DM. SEMMA fue desarrollado por el Instituto SAS. CRISP-DM fue desarrollado gracias a los esfuerzos de un consorcio compuesto inicialmente por DaimlerChrysler, SPSS y NCR (Azevedo y Santos, 2008).

A pesar de que SEMMA y CRISP-DM suelen denominarse metodologías, vamos a referirnos a ellas como procesos, en el sentido de que consisten en un curso de acción concreto destinado a lograr un resultado.

## KDD

El proceso de KDD (*knowledge discovery in databases*), tal y como se presenta por Fayyad et al. (1996), es el proceso de utilizar métodos de DM para extraer lo que se considera conocimiento según la especificación de medidas y umbrales, utilizando una base de datos junto con cualquier preprocesamiento, submuestreo y transformación de la base de datos que se requiera.

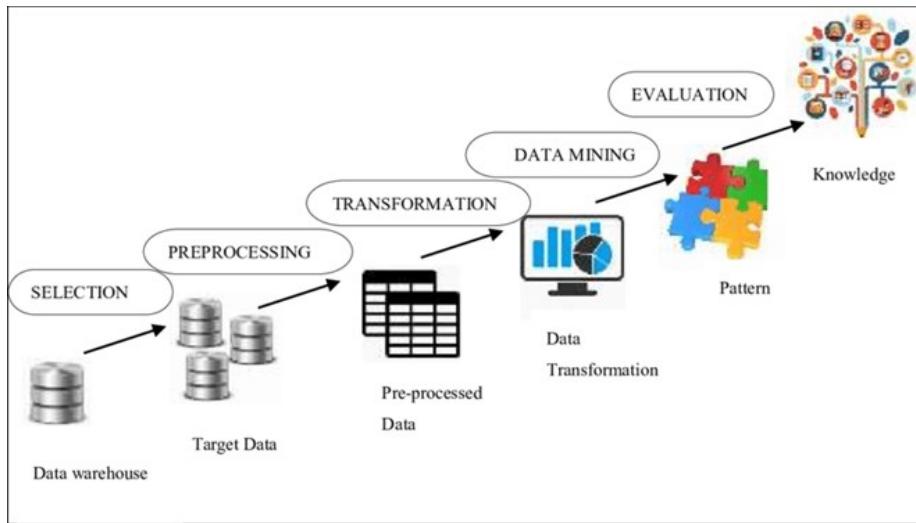


Figura 4. Etapas del proceso de descubrimiento de conocimiento (KDD). Fuente: Sabri, Man, Bakar y Rose, 2019.

Se consideran cinco etapas:

- ▶ Selección. Esta etapa consiste en crear un conjunto de datos objetivo, o centrarse en un subconjunto de variables o muestras de datos, sobre el que se va a realizar el descubrimiento.
- ▶ Preprocesamiento. Esta etapa consiste en la limpieza y preprocesamiento de los datos objetivo para obtener datos consistentes.
- ▶ Transformación. Esta etapa consiste en la transformación de los datos utilizando métodos de reducción de la dimensionalidad o de transformación.
- ▶ Minería de datos. Esta etapa consiste en la búsqueda de patrones de interés en una forma de representación particular, dependiendo del objetivo de la minería de datos (normalmente, la predicción).
- ▶ Interpretación/evaluación. Esta etapa consiste en la interpretación y evaluación de los patrones extraídos.

El proceso de KDD es interactivo e iterativo, e implica numerosos pasos con muchas decisiones tomadas por el usuario. Además, el proceso de KDD debe estar precedido por el desarrollo de una comprensión del dominio de la aplicación, el conocimiento previo relevante y los objetivos del usuario final. También debe ser continuado por la consolidación del conocimiento mediante la incorporación de este conocimiento al sistema (Fayyad et al., 1996).

## SEMMA

El proceso SEMMA fue desarrollado por el Instituto SAS. El acrónimo SEMMA significa ***sample, explore, modify, model, assess***, y se refiere al proceso de realización de un proyecto de minería de datos. El Instituto SAS considera un ciclo con cinco etapas para el proceso (Santos y Azevedo, 2008):

- ▶ Muestra. Esta etapa consiste en el muestreo de los datos mediante la extracción de una porción de un gran conjunto de datos lo suficientemente grande como para contener la información significativa, pero lo suficientemente pequeño para manipular rápidamente. Esta etapa se señala como opcional.
- ▶ Explorar. Esta etapa consiste en la exploración de los datos mediante la búsqueda de tendencias y anomalías imprevistas con el fin de obtener comprensión e ideas.
- ▶ Modificar. Esta etapa consiste en la modificación de los datos mediante la creación, selección y transformación de las variables para enfocar el proceso de selección del modelo.
- ▶ Modelar. Esta etapa consiste en modelar los datos permitiendo que el *software* busque automáticamente una combinación de datos que prediga de forma fiable un resultado deseado.

- ▶ Evaluación. Esta etapa consiste en evaluar los datos, valorando la utilidad y la fiabilidad de los resultados del proceso de minería de datos y estimando su rendimiento.

Aunque el proceso SEMMA es independiente de la herramienta de MD elegida, está vinculado al *software* SAS Enterprise Miner y pretende guiar al usuario en la implementación de aplicaciones de MD. SEMMA ofrece un proceso fácil de entender, que permite un desarrollo y mantenimiento organizado y adecuado de los proyectos de gestión de proyectos. Así, confiere una estructura para su concepción, creación y evolución, ayudando a presentar soluciones a los problemas de negocio, así como a encontrar los objetivos de negocio de la DM.

## CRISP-DM

El proceso estándar interindustrial para la minería de datos (CRISP-DM) es un marco para traducir los problemas empresariales en tareas de minería de datos y llevar a cabo proyectos de minería de datos independientes del área de aplicación y de la tecnología utilizada (Huber, Wiemer, Schneider y Ihlenfeldt, 2018). Es una implementación orientada a la industria ampliamente adoptada del proceso genérico de descubrimiento de conocimiento (KD).

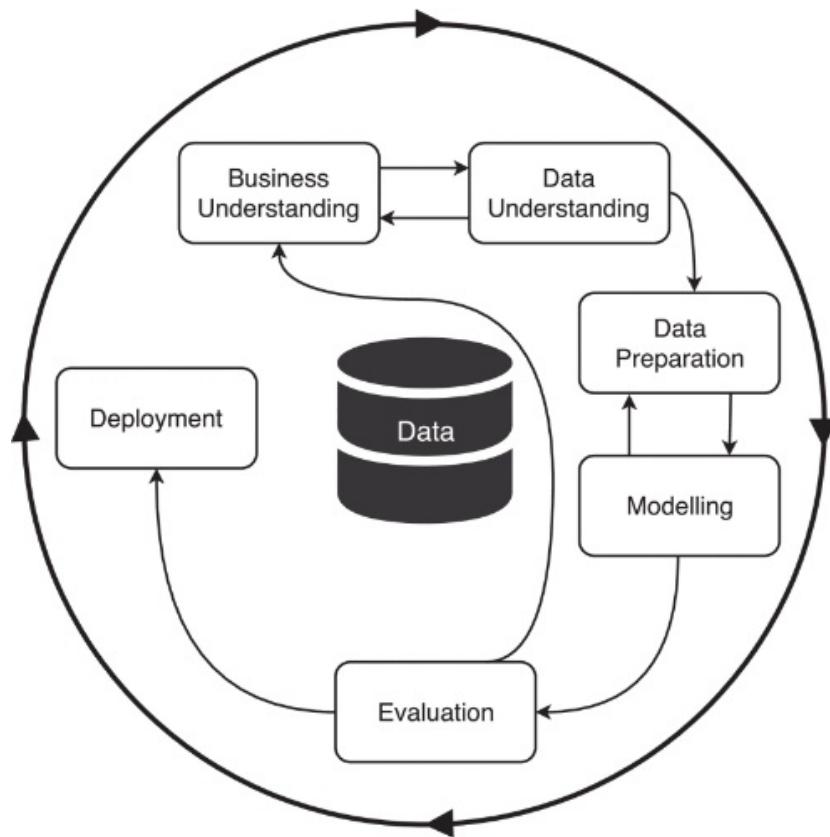


Figura 5. Etapas del proceso CRISP-DM. Martínez-Plumed et al., 2019.

La figura 5 muestra las seis fases del modelo de proceso CRISP-DM y sus interacciones.

- ▶ **Comprensión del negocio.** Esta fase inicial se centra en la comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial, para luego convertir este conocimiento en una definición del problema de minería de datos y un plan preliminar diseñado para lograr los objetivos.
- ▶ **Comprensión de los datos.** La fase de comprensión de los datos comienza con una recopilación inicial de datos y continúa con actividades para familiarizarse con los datos, para identificar los problemas de calidad de los datos, para descubrir los primeros conocimientos de los datos o para detectar subconjuntos interesantes para formar hipótesis de información oculta.

- ▶ **Preparación de los datos.** La fase de preparación de los datos abarca todas las actividades para construir el conjunto de datos final a partir de los datos brutos iniciales.
- ▶ **Modelización.** En esta fase se seleccionan y aplican diversas técnicas de modelización y se calibran sus parámetros hasta alcanzar los valores óptimos.
- ▶ **Evaluación.** En esta fase se evalúa más a fondo el modelo (o los modelos) obtenidos y se revisan los pasos ejecutados para construir el modelo con el fin de estar seguros de que alcanza adecuadamente los objetivos empresariales.
- ▶ **Despliegue.** La creación del modelo no suele ser el final del proyecto. Aunque el propósito del modelo sea aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse de forma que el cliente pueda utilizarlo.

La secuencia de las seis etapas no es rígida. CRISP-DM es extremadamente completo y documentado. Todas sus etapas están debidamente organizadas, estructuradas y definidas, permitiendo que un proyecto pueda ser fácilmente comprendido o revisado (Santos y Azevedo, 2008). Aunque el proceso CRISP-DM es independiente de la herramienta de DM elegida, está vinculado al *software* SPSS Clementine.

## La preparación de los datos

Dentro del tratamiento de datos, la preparación de datos es una actividad de vital importancia que **convierte datos dispares, sin formato y desordenados en una vista limpia y coherente** (IBM, 2021). **El proceso incluye la búsqueda, la limpieza, la transformación, la organización y la recogida de datos.** La preparación de datos es clave, pero requiere mucho tiempo. De hecho, los equipos de datos pasan hasta el 80 % de su tiempo convirtiendo datos en bruto en una salida de alta calidad y preparada para el análisis (IBM, 2021).

El trabajo de preparación de datos lo realizan los equipos de tecnología de información (TI), BI (*business intelligence*) y gestión de datos cuando integran conjuntos de datos para cargarlos en un almacén de datos, una base de datos NoSQL o un *data lake*, o cuando se desarrollan nuevas aplicaciones de análisis. Además, los científicos de datos, otros analistas de datos y los usuarios empresariales pueden utilizar **herramientas de preparación de datos de autoservicio** para recopilar y preparar los datos por sí mismos (Burns y Pratt, 2021).

La preparación de datos suele denominarse informalmente como *data prep*. También se conoce como *data wrangling*, aunque algunos profesionales utilizan este término en un sentido más estricto para referirse a la limpieza, estructuración y transformación de los datos como parte del proceso general de preparación de datos, distinguiéndolo de la etapa de preprocesamiento de datos.

Uno de los principales objetivos de la preparación de datos es garantizar que los datos brutos que se preparan para el procesamiento y el análisis de datos sean precisos y coherentes, de modo que los resultados de las aplicaciones de BI y análisis sean válidos.

Los datos suelen crearse con valores que faltan, inexactitudes u otros errores. Además, los conjuntos de datos separados suelen tener formatos diferentes que deben conciliarse. Corregir los errores de los datos, verificar su calidad y unir los conjuntos de datos constituye una gran parte del proceso de preparación de datos.

La preparación de los datos **también implica la búsqueda de datos relevantes** para incluirlos en las aplicaciones analíticas con el fin de garantizar que ofrezcan la información que buscan los analistas o los usuarios de la empresa. Los datos también pueden enriquecerse y optimizarse para hacerlos más informativos y útiles, por ejemplo, mezclando conjuntos de datos internos y externos, creando nuevos campos de datos, eliminando valores atípicos y abordando conjuntos de datos desequilibrados que podrían sesgar los resultados de los análisis.

Además, los equipos de BI y de gestión de datos pueden utilizar el proceso de preparación de datos para conservar los conjuntos de datos que analizarán los usuarios de la empresa. Esto ayuda a agilizar y guiar las aplicaciones de BI de autoservicio para los analistas empresariales, los ejecutivos y los trabajadores.

## Ventajas de la preparación de los datos

Los científicos de datos a menudo se quejan de que pasan la mayor parte de su tiempo localizando y limpiando datos en lugar de analizarlos. Una gran ventaja de instituir un proceso eficaz de preparación de datos es que ellos y otros usuarios finales pueden dedicar menos tiempo a encontrar y estructurar los datos y, así, centrarse más en la minería y el análisis de datos, las actividades relacionadas con el BI que aportan valor al negocio. Por ejemplo, la preparación de los datos puede realizarse con mayor rapidez y los datos preparados pueden ser suministrados automáticamente a los usuarios para realizar análisis repetitivos.

Un programa de preparación de datos bien gestionado también ayuda a una organización a realizar lo siguiente (Burns y Pratt, 2021):

- ▶ Garantizar que los datos utilizados para el BI, el aprendizaje automático, el análisis predictivo y otras aplicaciones analíticas tengan niveles de calidad suficientes para producir resultados fiables.
- ▶ Evitar la duplicación de esfuerzos en la preparación de datos que puedan utilizarse en múltiples aplicaciones.
- ▶ Preparar los datos para el análisis de forma rentable y eficiente.
- ▶ Identificar y solucionar problemas de datos que de otro modo no se detectarían.
- ▶ Tomar decisiones empresariales más informadas porque los ejecutivos tienen acceso a mejores datos, y obtener más valor empresarial y un mayor retorno de la inversión (ROI — *return on investment*) de sus iniciativas de BI y análisis.

Una preparación eficaz de los datos puede ser especialmente beneficiosa en entornos de *big data* con lagos de datos, a menudo construidos en torno a clústeres Hadoop, que almacenan grandes cantidades de datos estructurados, semiestructurados y no estructurados, a menudo en bruto.

En muchas aplicaciones de *big data*, la preparación de los datos es, en gran medida, una tarea automatizada: los algoritmos de aprendizaje automático pueden acelerar las tareas examinando los campos de datos y rellenando automáticamente los valores en blanco, corrigiendo los errores o renombrando los campos para garantizar la coherencia cuando se unen los conjuntos de datos.

## 3.6. Visualización

El aumento considerable del volumen de datos disponibles en la última década ha propiciado la proliferación de herramientas estándar para la visualización de estos. Mostrar los datos adecuadamente y hacerlos comprensibles a los consumidores de la información es una labor actualmente al alcance de cualquier usuario que requiera proporcionar dichas visualizaciones de forma eficiente y eficaz, sin necesidad de ser un desarrollador, ingeniero de datos o analista de datos experto. A lo largo de esta sección listaremos algunas de las herramientas más utilizadas para la visualización de datos provenientes de diferentes fuentes.

### Herramientas para la visualización de datos que se integran en las arquitecturas modernas

#### **Datawrapper**

Datawrapper un programa de código abierto para la creación de visualizaciones de manera gratuita. La herramienta permite el acceso a opciones adicionales y características mejoradas a través de diversos planes de pago y también cargar los datos y generar gráficos de manera muy sencilla. Basta con crear un perfil con una dirección de correo electrónico o a través de una cuenta de Twitter.

## Timeline JS

Timeline JS es una herramienta de código abierto que permite construir líneas de tiempo interactivas a partir de una hoja de cálculo. Podemos enriquecerlas incluyendo material multimedia procedente de fuentes como Twitter, Flickr, Google Maps, YouTube, Vimeo, etc. Una de las ventajas que ofrecen las cronologías creadas con Timeline JS es que son *responsive*, con lo que adaptan su diseño al dispositivo desde el cual se van a visualizar: escritorio, tableta o móvil.

## RAWGraphs

RAWGraphs es una herramienta de código abierto centrada en la creación de visualizaciones con D3.js de manera sencilla sin necesidad de tener conocimientos de programación. Podemos descargar las visualizaciones en formato vectorial (SVG) o como imágenes, así como obtener el código HTML para incrustarlas en nuestro sitio web.

## CartoDB

CartoDB es una aplicación en la nube que permite almacenar y visualizar datos en la web a través de la realización de mapas interactivos. Se ofrece bajo la modalidad de negocio *freemium* (el usuario parte de unos servicios básicos gratuitos y tiene acceso a opciones más avanzadas mediante una serie de planes de pago).

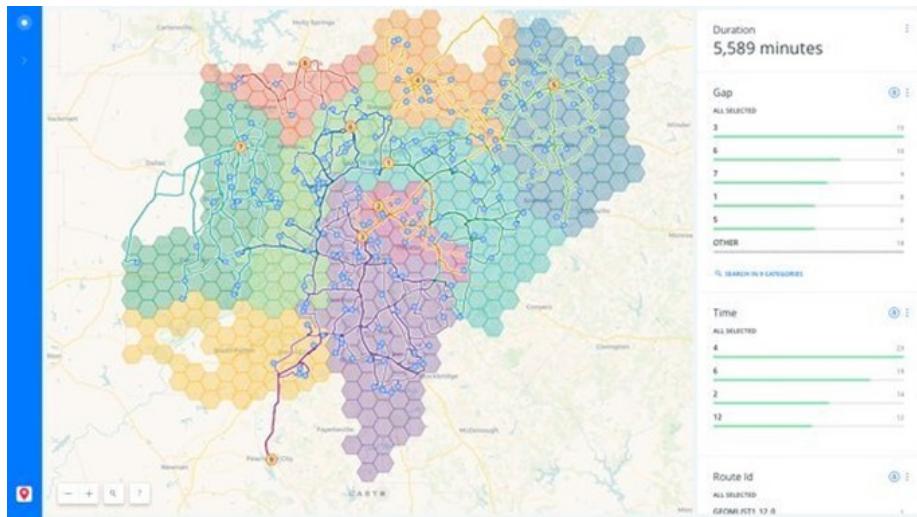


Figura 6. Ejemplo de mapa realizado con CartoDB en aplicaciones de logística. Fuente: Broderick y Debatte, s. f.

## Soluciones de presentación de datos e inteligencia empresarial

En la mayoría de las ocasiones, contaremos ya con capas de ingestión (redes sociales, encuestas, datos IoT, fuentes externas abiertas, bases de datos de la empresa, sistemas existentes en la empresa como ERP, CRM, etc.), almacenamiento (*data warehouses*, bases de datos NoSQL, etc.) y procesamiento de datos (herramientas *big data* como Apache Spark, etc.), siendo nuestra intención presentar informes y cuadros de mandos acerca de la información obtenida por la fusión de datos y el conocimiento inferido a partir de modelos *machine learning* sin necesidad de desarrollar capas de aplicación personalizadas.

Para ello, existe una gran variedad de soluciones de **inteligencia empresarial o business intelligence** orientadas a la creación de *dashboards* de forma sencilla eligiendo las fuentes de datos a representar, así como los tipos de gráficos más apropiados, disponiendo los mismos en *dashboards* personalizados.

Asimismo, los principales proveedores Cloud como AWS, Microsoft Azure y Google Cloud Platform facilitan la integración de sus fuentes de datos y capas de procesamiento de datos para la presentación de estas con este tipo de herramientas y viceversa, es decir, dichas herramientas se han desarrollado pensando en la presentación de datos almacenados en repositorios en la nube ofrecidos por los distintos proveedores *cloud*.

Es muy importante mencionar que los proveedores *cloud* no restringen las herramientas en función de que proporcionen otras herramientas competitivas en parte similares. Es decir, como ejemplos, es perfectamente posible utilizar Power BI para mostrar datos existentes en repositorios BigQuery en Google, así como utilizar Google Data Studio para mostrar información existente en AWS Redshift.

## Tableau

Tableau Software es una empresa que proporciona soluciones de *software* de presentación de datos e inteligencia empresarial basada en investigaciones originadas en la Universidad de Standford. Ofrece diferentes versiones de su *software* con distintos niveles de funcionalidades y precios. Permite conectarse con diferentes fuentes de datos como Amazon Redshift, Google BigQuery o bases de datos locales, entre otras, para crear cuadros de mando muy atractivos a la hora de presentar información.

**Tableau Public** es una avanzada e intuitiva aplicación gratuita de creación y publicación de gráficos interactivos, mapas, tablas, etc. Permite crear tableros que incluyan varias representaciones gráficas de los datos dentro de la misma visualización.

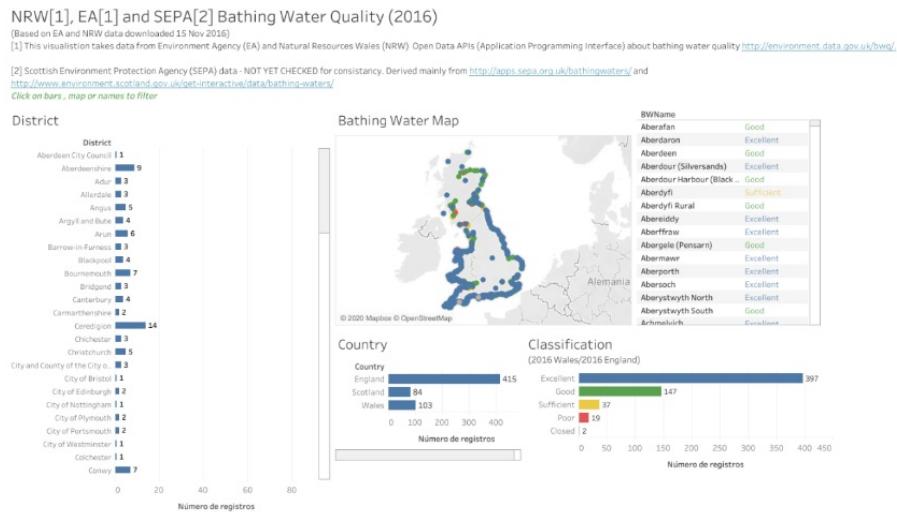


Figura 7. Ejemplo de *dashboard* en Tableau Public mostrando la calidad del agua de baño en Gran Bretaña. Fuente: Pshab, 2016.

Sin embargo, Tableau ofrece versiones *premium* no gratuitas con una mayor capacidad de almacenamiento de nuestros datos y más opciones para conectar fuentes de datos. **Tableau Desktop** es la versión de escritorio que permite crear cuadros de mando sin más que arrastrar diferentes fuentes de datos para su análisis. **Tableau Online** permite realizar análisis de inteligencia empresarial en la nube conectándose al mismo a través de un navegador web, mientras que **Tableau Server**, aunque también se encuentra en la nube, permite a los miembros de las organizaciones empresariales colaborar entre sí de forma más segura, ya que sigue una política diferente a la hora de usar recursos en el *cloud*.

## Google Data Studio

Google Data Studio es una herramienta de Google para la creación de *dashboards* completamente gratuita y que permite conectar numerosas fuentes de datos con una amplia variedad de gráficos. Una de las principales ventajas de Google Data Studio, además de su gratuidad, es que cuenta con un *marketplace* para utilizar conectores

desarrollados por terceros. Esto incluye (julio de 2021) más de ochocientas fuentes de datos sobre más de cuatrocientos treinta conectores posibles, incluyendo los propios de Google (BigQuery, Cloud Spanner, Google Analytics, Youtube Analytics, etc.), así como conectores de terceros (algunos con coste) que permiten conectarse a otras fuentes de datos y bases de datos en otros *cloud* (AWS, Azure, etc.).



Figura 8. Ejemplo de *dashboard* creado con Google Data Studio. Fuente:

[https://datastudio.google.com/u/0/reporting/1rBC8woDruwE-f\\_gsheY7kwlx0BPtgzEn/page/c2P1](https://datastudio.google.com/u/0/reporting/1rBC8woDruwE-f_gsheY7kwlx0BPtgzEn/page/c2P1)

## Power BI

Al igual que otras herramientas de inteligencia empresarial, Power BI (de *business intelligence*) es una solución de pago desarrollada por Microsoft cuyo objetivo inicial era proporcionar visualizaciones de inteligencia empresarial orientadas a *marketing*, ventas o datos financieros. Power BI permite extraer datos de fuentes como Excel, Google Analytics, Salesforce o redes sociales (LinkedIn, Twitter, Facebook, etc.). Una de las ventajas más importantes de Power BI es que permite a los usuarios interactuar mediante reconocimiento por comandos de voz usando técnicas de procesamiento de lenguaje natural, lo cual la hace muy atractiva por parte de los usuarios para obtener información y conocimiento adquiridos a partir de los datos.

Al igual que en el caso de Tableau, existen diferentes versiones de la herramienta en función de las necesidades de los usuarios: Power BI Desktop, Power BI Pro, Power BI Premium, Power BI Mobile (versiones de aplicaciones nativas para Windows Mobile, Android, así como iOS) y Power BI Embedded (para ser integrado en aplicaciones propias desarrolladas).

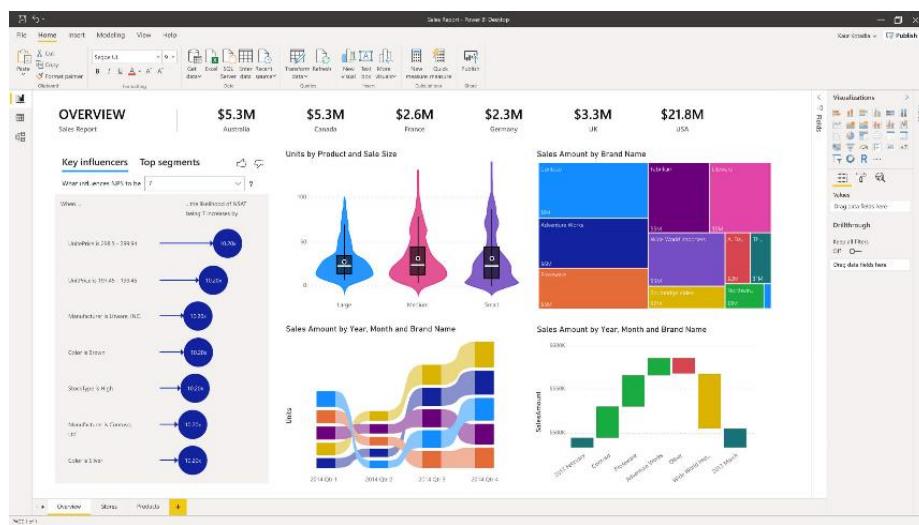


Figura 9. Cuadros de mando creados por Power BI Desktop. Fuente: <https://powerbi.microsoft.com/es-es/>

Con el paso del tiempo, e influenciada por la plataforma Azure y, especialmente, los componentes de Azure IoT, Power BI permite ingerir datos provenientes de fuentes IoT, lo cual la hace muy interesante para visualizaciones de datos de mediciones de sensores IoT, ya sea IoT industrial, *smart meters* o IoT de consumo.

## Grafana

Grafana es una herramienta *open-source* y gratuita que nació inicialmente para la visualización de estadísticas de uso de la CPU en ordenadores y servidores. Existen versiones para Linux, Windows, Mac, Docker (sistema de contenedores que permite ir un paso más allá en la virtualización de nuestro equipo, sin necesidad de instalar

una copia completa del sistema operativo en cada contenedor en funcionamiento y fácilmente orquestables en la nube mediante tecnología Kubernetes de Google) e incluso ARM.

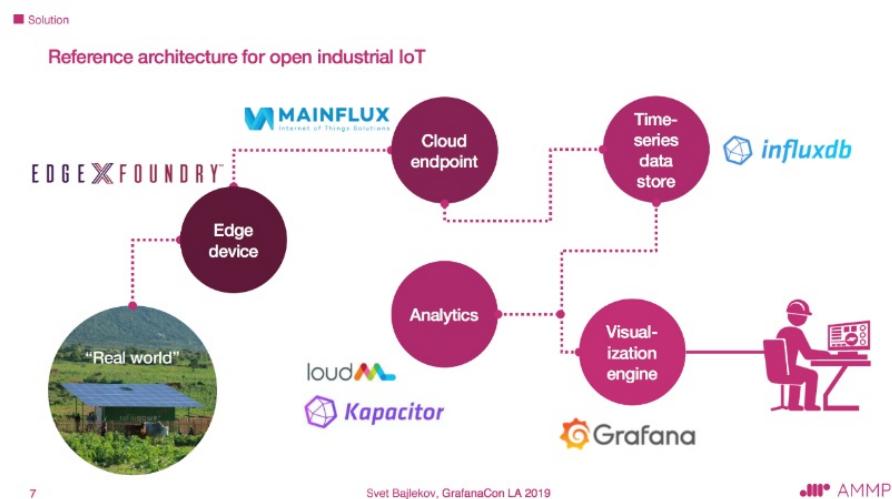


Figura 10. Ejemplo de arquitectura de ingestión, almacenamiento, procesamiento y visualización de datos IoT industriales en la cual se involucra a Grafana. Propuesta por la empresa AMMP Technologies. Fuente: Dam, 2019.



Figura 11. Ejemplo de uso de Grafana para mostrar datos IoT, creado por Dan Cech para el IoT Workshop GrafanaCon 2019. Fuente: Cech, 2019.

En la actualidad, gracias a su *marketplace de plugins*, Grafana es una herramienta que permite ingerir y mostrar datos de una gran variedad de fuentes como MySQL, InfluxDB (series temporales), Prometheus (eventos en tiempo real), Google BigQuery, AWS IoT SiteWise (servicio gestionado de AWS para tratar con datos de equipos industriales), etc., así como incorporar nuevos elementos de visualización a los *dashboards* que creamos.



Figura 12. *Dashboard* en Grafana leyendo datos de Prometheus. Fuente:

<https://prometheus.io/docs/visualization/grafana/>

## Looker

Looker Data Sciences fue fundada en 2012. En 2019 fue adquirida por Google y ahora Looker forma parte del ecosistema de componentes de Google Cloud Platform. Looker es una herramienta de pago muy completa de inteligencia empresarial que

permite ingerir datos de múltiples fuentes, aplicar modelos de aprendizaje automático y visualizar dicha información de forma profesional.

Uno de los aspectos más interesantes de Looker es su sistema de bloques *Looker Blocks* con licencia MIT que permite reutilizar piezas de código de terceros a partir de un amplio directorio estructurado en diferentes clases: Analytics, Source (*data sources*), Data Blocks, Data Tools, Viz Blocks (visualizaciones) y Embedded Blocks (para incluirlos en nuestros portales web, por ejemplo, así como entornos de realidad virtual, entre otras posibilidades).



Figura 13. Ejemplo de uso de Looker con Stitch (para ingerir fuentes de datos en un *data warehouse*).

Fuente: <https://looker.com/blog/announcing-looker-blocks-and-stitch>

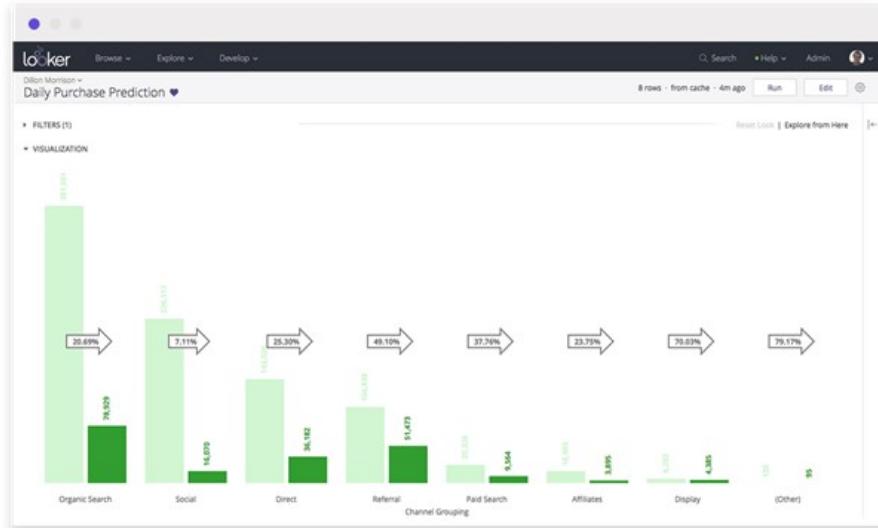


Figura 14. Uso de *bigrquery machine learning* (BQML) sobre Looker. Fuente:  
<https://looker.com/platform/blocks/source/bigquery-machine-learning-by-google>

## Qlik

Qlik nació en Suecia en 1993 (originalmente como Qliktech) y actualmente tiene su sede en Estados Unidos.

Qlik ofrece soluciones de pago orientadas a la inteligencia empresarial y visualización de datos con una gran variedad de posibles componentes y siendo destacable módulos específicos para datos IoT, especialmente interesante en la Industria 4.0.

Mientras que sus productos Qlik View están pensados para la visualización y el análisis de datos basado en su motor asociativo, Qlik Sense se presenta como la evolución de los productos de Qlik con un motor mejorado según información promocional de la propia empresa.

Ofrecen diferentes esquemas de precios basados en diferentes niveles de funcionalidad, tanto como un SaaS en la nube gestionada directamente por Qlink como *on-premise*.

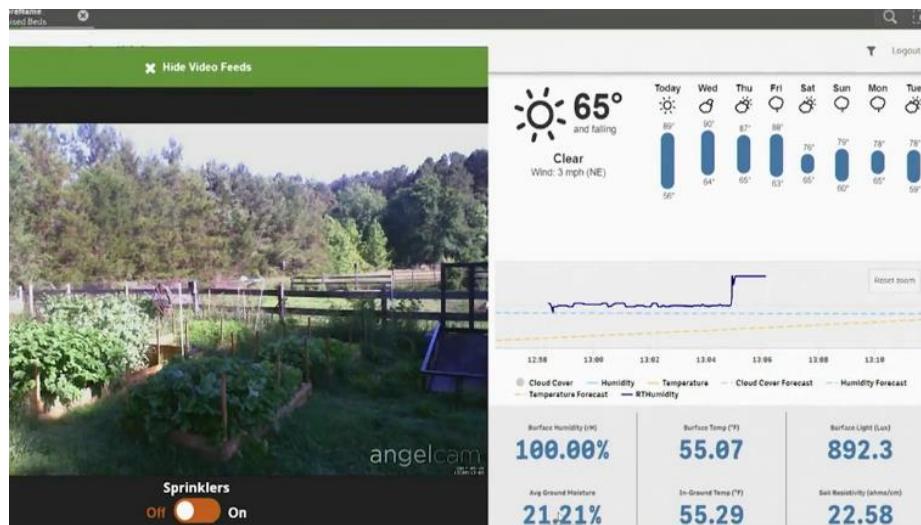


Figura 15. Ejemplo de visualización de datos IoT usando Qlik. Fuente: <https://blogqlik.com/an-iot-journey-through-connections-2017>

## Adverity

Adverity es una solución propietaria de pago que permite ingerir fuentes de datos mediante más de cuatrocientos conectores diferentes (Amazon Redshift, MongoDB, Twitter, Slack, BigQuery, Google Analytics, Shopify, etc.), procesarlos, mostrarlos mediante *dashboards* propios y aplicar modelos inteligentes para la creación de campañas.

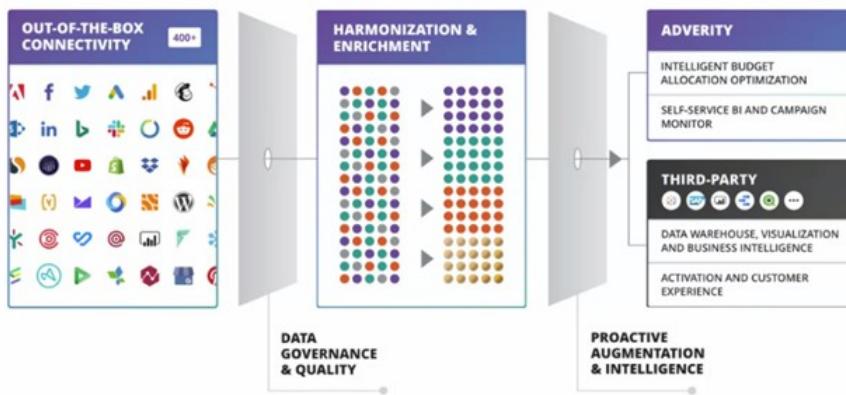


Figura 16. Plataforma Adverity para la ingestión, procesamiento y presentación de datos. Fuente:

<https://www.adverity.com/analytics-platform/>



Figura 17. Ejemplo de *dashboard* creado con Adverity. Fuente: <https://www.adverity.com/analytics-platform/data-visualization/#insightsdashboards>

Asimismo, Adverity permite volcar los datos finalmente a *data warehouses* o colas de procesamiento en la nube (AWS, Azure, Google Cloud, Hadoop, MySQL) u otras herramientas de visualización de terceros (Tableau, Google Data Studio, etc.).

## Funnel

Funnel es otra alternativa a Adverity y sigue una filosofía similar. Una plataforma de pago que permite ingerir datos de más de quinientas fuentes de entrada diferentes y conectarlas a su salida a múltiples destinos, incluyendo una API propia (para su conexión con herramientas desarrolladas por terceros), AWS S3 / Redshift, Google BigQuery / Cloud Storage, Azure, etc.

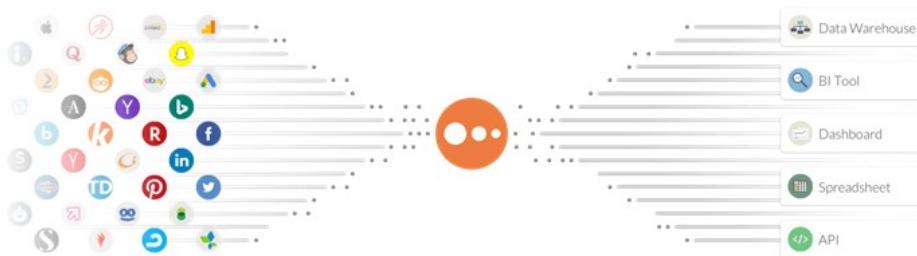


Figura 18. Imagen promocional de Funnel y su concepto como plataforma de tratamiento de datos.

Fuente: <https://funnel.io/>

## Lenguajes de programación para la presentación de datos personalizadas en el ámbito de la ciencia de datos

Finalmente, y como ya se ha comentado, **existe la posibilidad de desarrollar una capa de presentación de datos personalizada dentro de nuestra capa de aplicación**. Tal y como se describió anteriormente, cuando se trata de interfaces web, usaremos HTML, CSS y JavaScript, además de *frameworks reactivos* como Angular, React.js o Vue.js en combinación con librerías para la construcción de gráficos como Google Charts o D3.js, entre otras posibilidades.

En el caso de que queramos personalizar nuestras presentaciones de datos haciendo uso de lenguajes de programación desde un **punto de vista más enfocado a la ciencia de datos o la aplicación de técnicas de inteligencia artificial** los dos lenguajes más empleados son Python y R.

## El lenguaje Python y las librerías Matplotlib y Seaborn

Python es un lenguaje de programación creado por Guido van Rossum y lanzado en 1991 (y sí, el nombre viene del célebre grupo humorístico británico Monty Python). Es el lenguaje más empleado en el ámbito de la inteligencia artificial y el *machine learning* (Bansal, 2019), además de ser utilizado también para el desarrollo de aplicaciones web del lado del servidor, el desarrollo de *software* en general, aplicaciones matemáticas o *scripting* en la gestión de sistemas operativos.

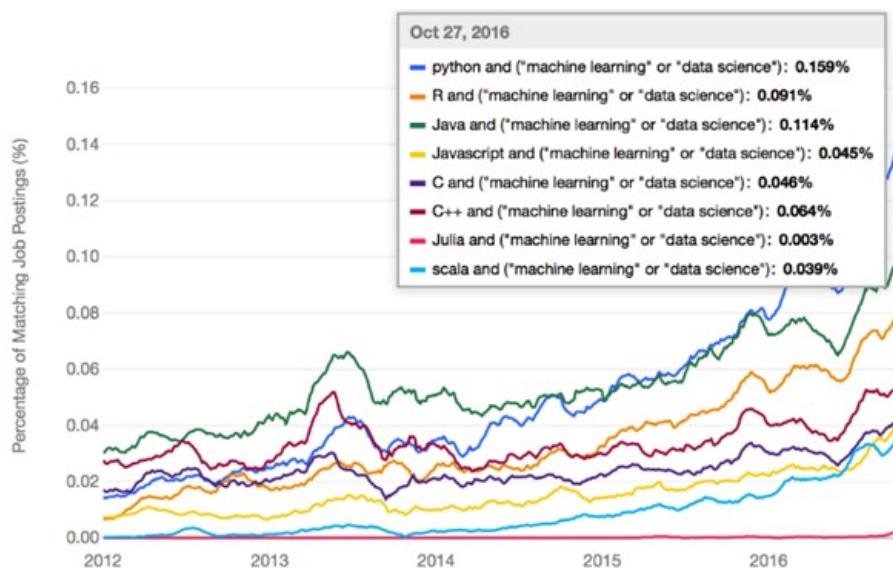


Figura 19. Ofertas de empleo en *machine learning* y ciencia de datos en el portal Indeed según lenguaje de programación. Fuente: Puget, 2016.

Una de las ventajas de Python es que cuenta con numerosas librerías desarrolladas por la comunidad que pueden ser utilizadas en función de las necesidades de nuestros proyectos. En el caso de la aplicación de técnicas de inteligencia artificial, ya hemos comentado librerías como Scikit-learn (para trabajar con aprendizaje automático clásico) y TensorFlow y Keras (para trabajar con redes neuronales y *deep learning*). Para implementar nuestros propios gráficos para la presentación de datos,

las librerías más extendidas son Matplotlib y Seaborn, que en realidad trabaja sobre la primera.

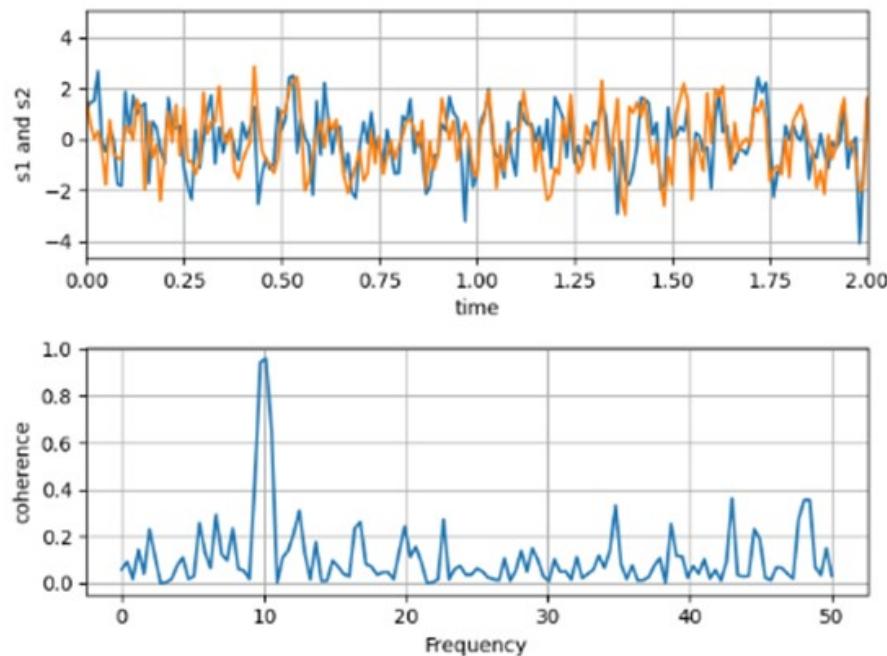


Figura 20. Ejemplo de diagramas de líneas mostrando series temporales con Matplotlib. Fuente:

[https://matplotlib.org/gallery/lines\\_bars\\_and\\_markers/cohere.html#sphx-glr-gallery-lines-bars-and-markers-cohere-py](https://matplotlib.org/gallery/lines_bars_and_markers/cohere.html#sphx-glr-gallery-lines-bars-and-markers-cohere-py)

Matplotlib es una biblioteca de Python muy popular para la visualización de datos. Es particularmente útil cuando un programador quiere visualizar los patrones de los datos. Es una librería de ploteo en 2D usada para crear gráficos y diagramas en 2D. Un módulo llamado Pyplot facilita a los programadores el trazado, ya que proporciona características para controlar los estilos de línea, las propiedades de las fuentes, los ejes de formato, etc.

Matplotlib proporciona varios tipos de gráficos y diagramas para la visualización de datos, es decir, histogramas, tablas de error, chats de barras, etc. Por su parte, Seaborn está construida sobre Matplotlib para hacer más sencilla la tarea de dotar a los gráficos de una mayor calidad y detalle si es necesario un acabado más atractivo.

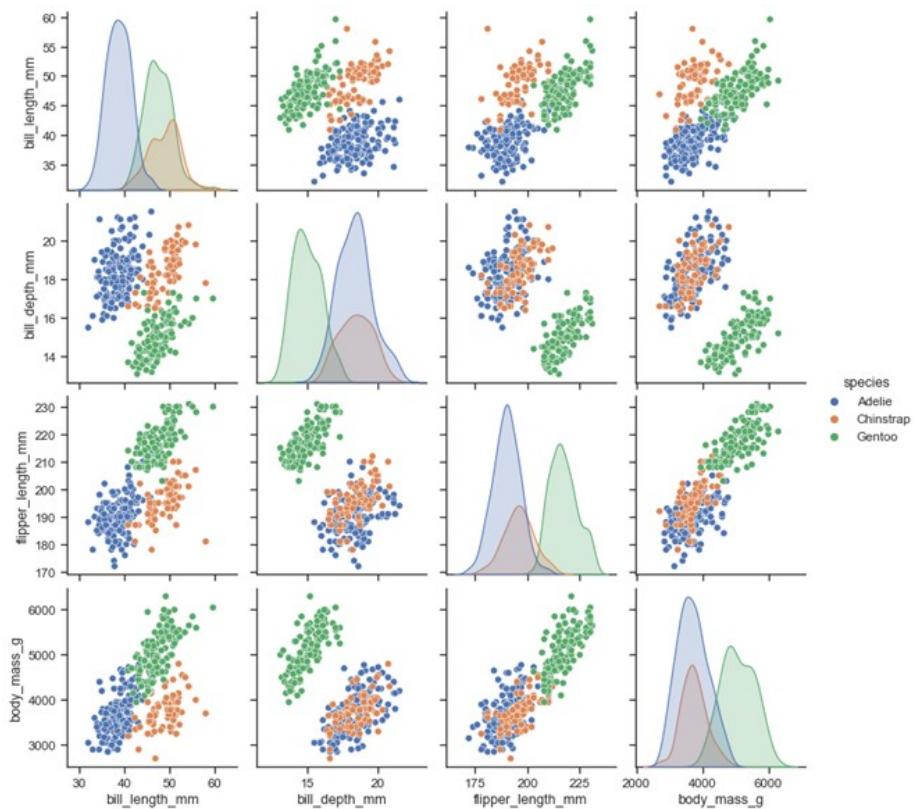


Figura 21. Ejemplo de matriz de diagramas de dispersión utilizando Seaborn. Fuente:

[https://seaborn.pydata.org/examples/scatterplot\\_matrix.html](https://seaborn.pydata.org/examples/scatterplot_matrix.html)

## El lenguaje de programación R

R es un dialecto del lenguaje S, desarrollado por John Chambers en Bell Labs, iniciando su desarrollo en 1976 mediante bibliotecas Fortran como un entorno interno para análisis estadístico. R es un lenguaje popular de modelado estadístico que es utilizado por los científicos de estadísticas y datos. Proporciona apoyo a un paquete estadístico diverso que se utiliza más ampliamente para el análisis y la modelización de datos. Ross Ihaka y Robert Gentleman desarrollaron juntos R en 1995 en la Universidad de Auckland. Para varios roles de análisis de datos y computación estadística, R es una elección popular.

Según Bansal (2019), R se encuentra también en tercera posición como lenguaje más empleado para la aplicación de técnicas de *machine learning* tras Python y Java. Según dichas fuentes, R es un lenguaje basado en gráficos que se utiliza para la computación estadística, el análisis y las visualizaciones en el aprendizaje de las máquinas. Para aquellos que quieren explorar datos estadísticos a través de gráficos, es la plataforma perfecta. También es usado para una variedad de propósitos por científicos de datos en Facebook, Google y muchas otras grandes compañías.

Si bien su curva de aprendizaje es más lenta que en otros lenguajes como Python y no está concebido como un lenguaje de propósito general que permita implementar *back-end* o trabajar con dispositivos IoT de consumo (Raspberry Pi, etc.), esta herramienta es de gran interés para el análisis estadístico, **siendo una de las más potentes en este campo** y una de las más extendidas en el ámbito académico y de la investigación.

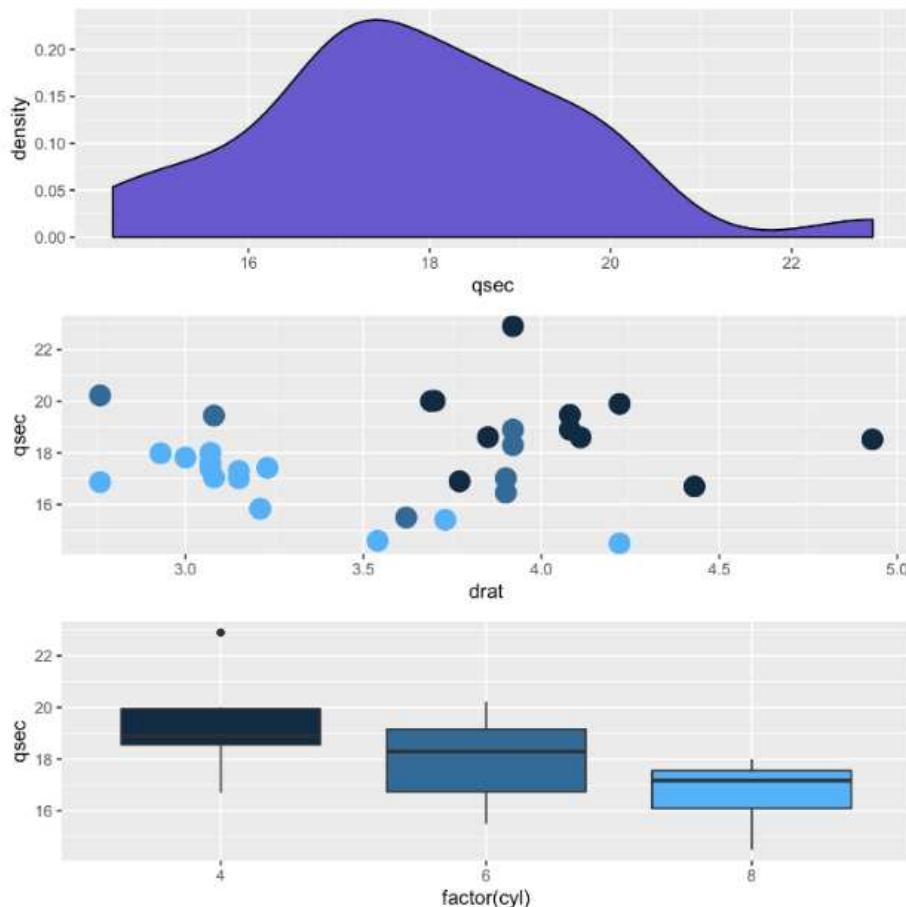
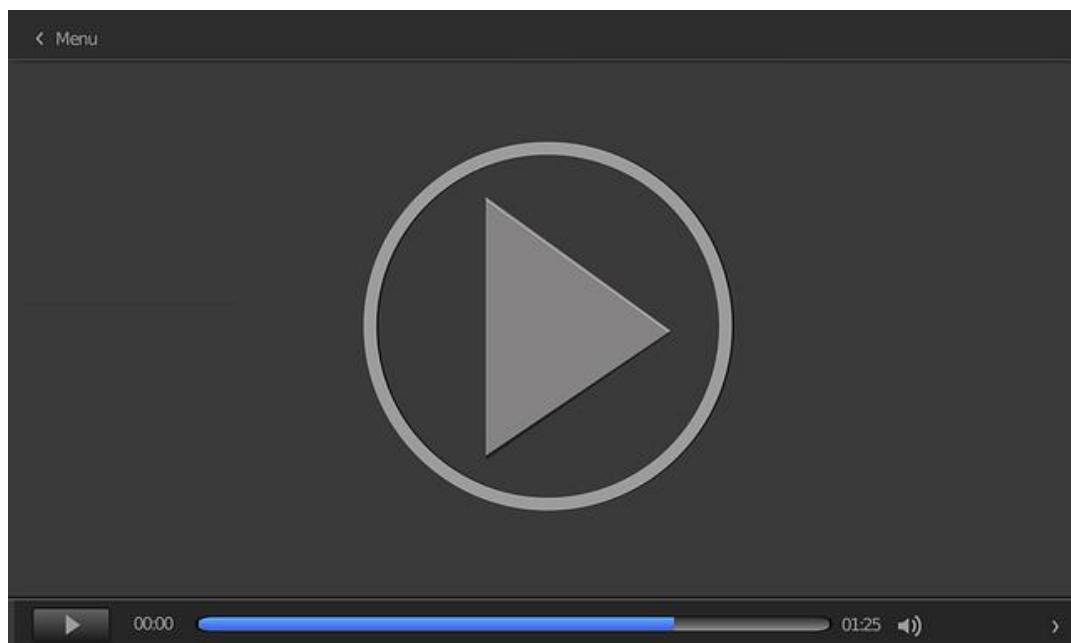


Figura 22. Ejemplo de múltiples gráficos en una única figura utilizando R y ggplot2. Fuente: <https://www.r-graph-gallery.com/261-multiple-graphs-on-same-page.html>

Existen más de diez mil paquetes en el repositorio de la distribución de la biblioteca CRAN de R. Estos paquetes están hechos a medida para una variedad de aplicaciones estadísticas. Mientras que R puede ser un lenguaje estadístico de núcleo duro, proporciona un soporte extensible para varios campos, que van desde la salud a la astronomía y la genómica.

R es popular por su amplio soporte de visualización. **En el caso de la presentación de datos, uno de los paquetes más extendidos es ggplot2.** Este paquete proporciona una amplia gama de capacidades gráficas que hacen que los datos sean interactivos para los usuarios. Con la ayuda de ggplot2, los usuarios pueden aprovechar las extensiones para aumentar la usabilidad y la experiencia personal.

Veamos ahora un tutorial introductorio a la *Visualización de datos con Google Data Studio*, configuración de la cuenta, conexión con una fuente de datos básica y creación de un *dashboard* de ejemplo.



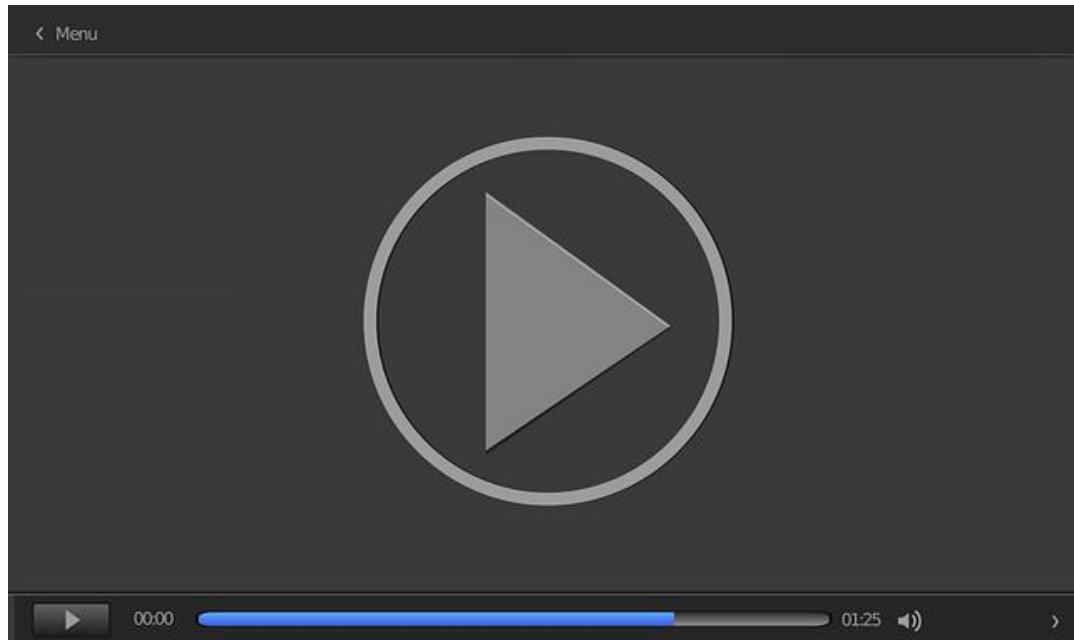
---

Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=f0f3b425-e370-402c-9ae1-b16f00d18643>

---

A continuación, podemos ver otro tutorial de *Visualización de datos con Tableau*, configuración de la cuenta, conexión con una base de datos SQL y creación de un *dashboard* de ejemplo.



---

Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=56f312d7-c1fa-47cb-9245-b16f00d185f5>

---

## 3.7. Referencias bibliográficas

Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002). Conceptual modeling for ETL processes. \*Data & Knowledge Engineering\*, 46(3), 265-300.

Lima, A. C. E., de Castro, L. N. y Corchado, J. M. (2015). A polarity analysis framework for Twitter messages. *Applied Mathematics and Computation*, 270, 756-767.

Azevedo, A. y Santos, M. F. (2008). *KDD, SEMMA and CRISP-DM: a parallel overview*. IADIS.

Burns, E. y Pratt, M. K. (2021). *Data preparation*. TechTarget.  
<https://searchbusinessanalytics.techtarget.com/definition/data-preparation>

Corchado, J. M., Chamoso, P., Hernández, G., San Román, A., Rivas, A., González-Briones, A., Pinto-Santos, F., Goyenechea, E., García-Retuerta, D., Alonso-Miguel, M., Bellido, B., Valdeolmillos, D., Sánchez-Verdejo, M., Plaza-Martínez, P., López-Pérez, M., Manzano-García, S., Alonso, R. S., Casado-Vara, R. ... y Omatu, S. (2021). Deepint. net: A rapid deployment platform for smart territories. *Sensors*, 21(1), 236.

Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11), 27-34.

Fayyad, U. y Stolorz, P. (1997). Data mining and KDD: Promise and challenges. *Future generation computer systems*, 13(2-3), 99-115.

Hazelcast. (2021). *Event Stream Processing*. <https://hazelcast.com/glossary/event-stream-processing/>

Huber, S., Wiemer, H., Schneider, D. y Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications—a holistic extension to the CRISP-DM model. *Procedia Cirp*, 79, 403-408.

IBM. (18 de julio de 2019). *ESB (bus de servicio empresarial)*.  
<https://www.ibm.com/es-es/cloud/learn/esb>

IBM. (2021). *Preparación de datos*. <https://www.ibm.com/es-es/analytics/data-preparation>

Khine, P. P. y Wang, Z. S. (2018). Data lake: a new ideology in big data era. En *ITM web of conferences* (Vol. 17). EDP Sciences.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J. H., Kull, M., Lachiche, N., Ramírez-Quintana, M. J. y Flach, P. A. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*. <https://research-information.bris.ac.uk/en/publications/crisp-dm-twenty-years-later-from-data-mining-processes-to-data-sc>

Rodríguez, S. (18 de noviembre de 2016). *Qué es un ESB*. Sergio Rodríguez Calvo. Medium. <https://serrodcal.medium.com/qu%C3%A9-es-un-esb-256f95b08ec5>

Sabri, I. A. A., Man, M., Bakar, W. A. W. A. y Rose, A. N. M. (2019). Web data extraction approach for deep web using WEIDJ. *Procedia Computer Science*, 163, 417-426.

Talend. (2021). *What is data preparation?* <https://www.talend.com/resources/what-is-data-preparation/>

Vyas, S. y Vaishnav, P. (2017). A comparative study of various ETL process and their testing techniques in data warehouse. *Journal of Statistics and Management Systems*, 20(4), 753-763.

## Google Dataset Search

Google. (S. f.). *Data Search*. <https://datasetsearch.research.google.com/>

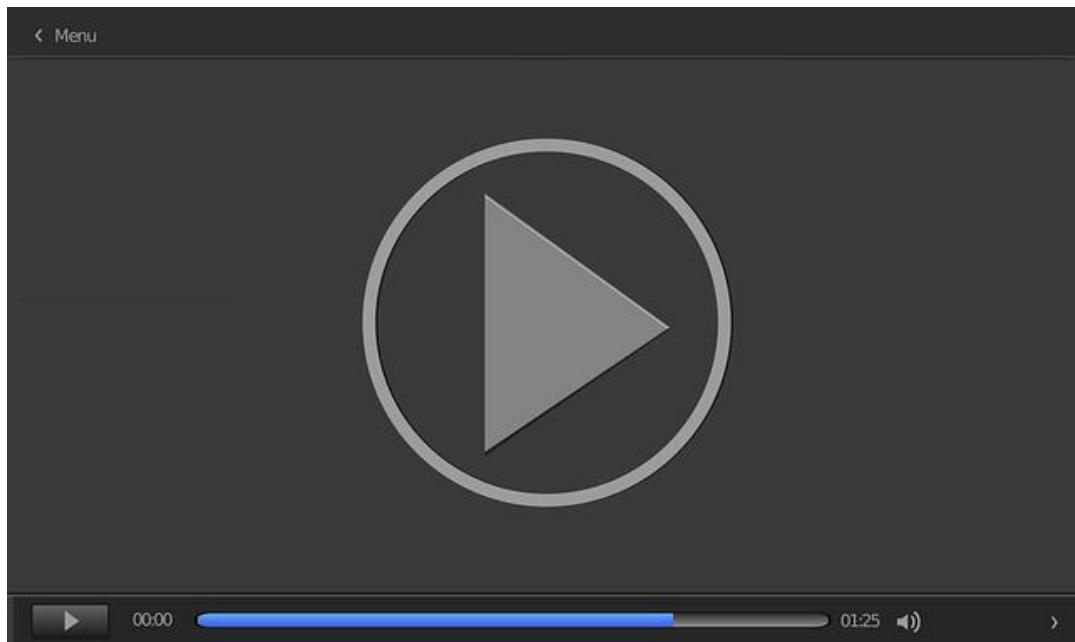
Google Dataset Search es un motor de búsqueda a través de los metadatos de millones de conjuntos de datos en miles de repositorios en toda la web. En enero de 2020 dejó de ser un servicio en versión beta y actualmente cuenta con más de 25 millones de *datasets* indexados.

## What is Object Storage?

IBM Technology. (25 de abril de 2019). *What is Object Storage?* [Vídeo]. YouTube.

<https://youtu.be/FLp88DzvtUk>

Vídeo introductorio de IBM Cloud explicando qué es el almacenamiento de objetos y casos posibles de aplicación.



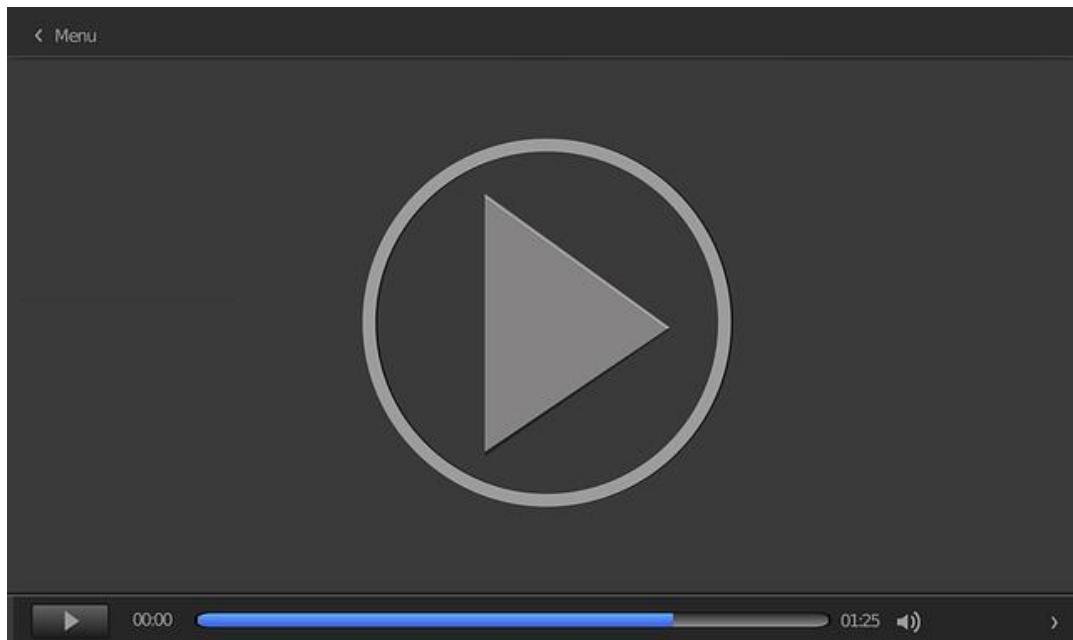
Accede al vídeo:

<https://www.youtube.com/embed/FLp88DzvtUk>

## SQL vs NoSQL or MySQL vs MongoDB

Academind. (25 de julio de 2018). *SQL vs NoSQL or MySQL vs MongoDB* [Vídeo]. YouTube. [https://youtu.be/ZS\\_kXvOeQ5Y](https://youtu.be/ZS_kXvOeQ5Y)

Este vídeo hace una comparación entre las bases de datos NoSQL y SQL, además de mostrar algunos ejemplos de bases de datos NoSQL.



Accede al vídeo:

[https://www.youtube.com/embed/ZS\\_kXvOeQ5Y](https://www.youtube.com/embed/ZS_kXvOeQ5Y)

## Preparar los datos y convertirlos en información

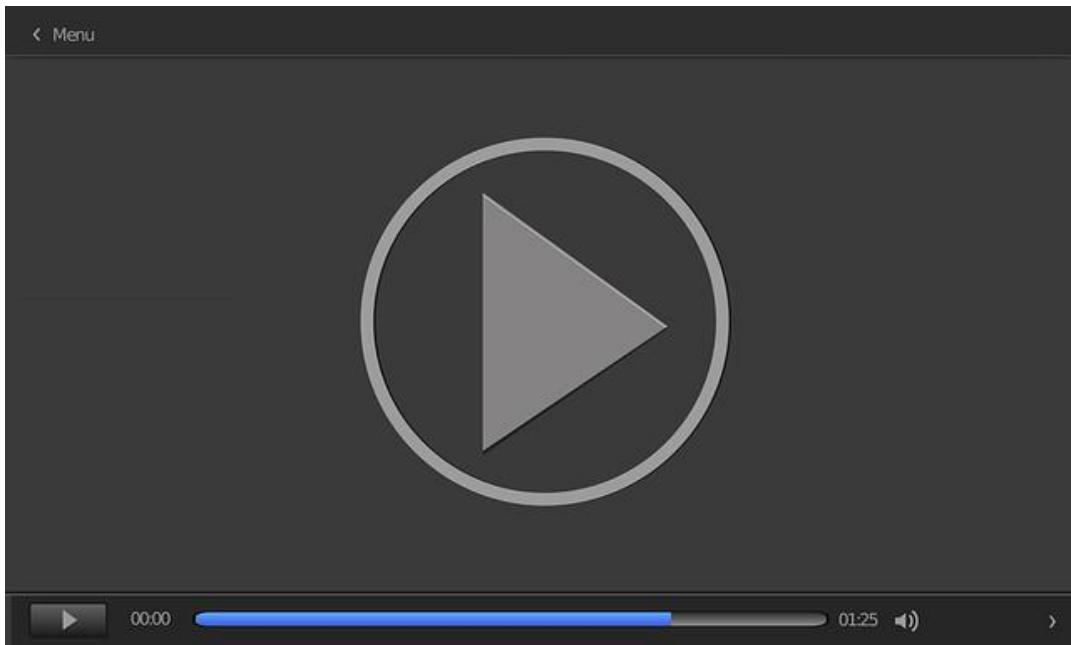
El arte de medir. (19 de enero de 2021). *Preparar los datos y convertirlos en información*. <https://elartedemedir.com/blog/preparar-los-datos-y-convertirlos-en-informacion/>

El arte de medir es una empresa privada que proporciona servicios de análisis y ciencia de datos, entre otros. Como otras empresas del sector, su blog tiene interesantes entradas sobre diferentes procesos y etapas de la ciencia de datos. En este caso incluimos una entrada en la que tratan la preparación de datos.

## Procesos ETL. Enrique Onieva

Universidad Deusto / Deustuko Unibertsitatea. (17 de marzo de 2016). *Procesos ETL. Enrique Onieva* [Archivo de vídeo]. YouTube. <https://www.youtube.com/watch?v=u3Le4lFePnQ>

Vídeo de introducción a los procesos de extracción, transformación y carga (ETL) impartido por el Dr. Enrique Onieva de la Universidad de Deusto. Un buen punto de partida para contar con una explicación gráfica acerca de sus diferentes etapas.



Accede al vídeo:

<https://www.youtube.com/embed/u3Le4lFePnQ>

## Herramientas de procesado y visualización de datos

Gobierno de España. (s. f.). Datos.gob.es.

<https://datos.gob.es/es/documentacion/herramientas-de-procesado-y-visualizacion-de-datos>

Datos.gob (el portal de datos abiertos del Gobierno de España) presenta una interesante recopilación de las herramientas de procesado y visualización de datos que podemos encontrar gratuitamente en la actualidad. Puedes descargar el documento en diferentes formatos.

1. ¿Qué nombre recibe un conjunto de datos persistente utilizado por un sistema de *software*?

  - A. Archivo.
  - B. Base de datos.
  - C. Registro.
  - D. Las respuestas A y B son correctas.
  
2. ¿Qué tipo de datos puede almacenar un *data warehouse*?

  - A. Datos estructurados.
  - B. Datos no procesados.
  - C. Las respuestas A y B son correctas.
  - D. Ficheros planos.
  
3. Al proceso de utilizar métodos de minería de datos para extraer lo que se considera conocimiento según la especificación de medidas y umbrales, utilizando una base de datos junto con procesos de transformación de los datos se lo conoce como:

  - A. CRISP-DM.
  - B. ETL.
  - C. KDD.
  - D. *Machine learning*.

- 4.** Entre las ventajas de la preparación de los datos nos encontramos con las siguientes (marca todas las correctas):
- A. Preparar los datos para el análisis de forma rentable y eficiente.
  - B. Garantizar que los datos utilizados para el BI tengan niveles de calidad suficientes.
  - C. Crear duplicados de los datos para que puedan utilizarse en múltiples aplicaciones de forma segura.
  - D. Todas son correctas.
- 5.** La limpieza de datos corrige problemas como:
- A. Datos duplicados.
  - B. Datos redundantes.
  - C. Datos no estructurados.
  - D. Datos incoherentes.
- 6.** ¿Cuáles de los siguientes repositorios de datos almacenan datos no estructurados, semiestructurados y estructurados?
- A. *Data warehouses*.
  - B. A y C son correctas.
  - C. *Data lakes*.
  - D. *Data marts*.
- 7.** ¿Cuáles de los siguientes repositorios de datos siguen una estructura de procesamiento *schema on write*? Marca todas las correctas:
- A. *Data swamps*.
  - B. *Data warehouses*.
  - C. *Data lakes*.
  - D. *Data marts*.

**8.** ¿Cuál de las siguientes afirmaciones describe mejor el propósito del proceso ETL?

- A. Procesar eventos en tiempo real para análisis instantáneo.
- B. Integrar datos de múltiples fuentes en un formato homogéneo.
- C. Enviar y recibir mensajes entre diferentes aplicaciones.
- D. Visualizar datos para reportes y paneles de control.

**9.** ¿Cuál es una fase crítica del proceso ETL donde se aplican reglas para corregir o eliminar datos incorrectos o incompletos?

- A. Extracción.
- B. Transformación.
- C. Carga.
- D. Almacenamiento.

**10.** ¿Qué aspecto del proceso ETL se enfoca principalmente en mejorar el rendimiento de las consultas y la escalabilidad del sistema de almacenamiento de datos?

- A. Optimización de la extracción.
- B. Paralelización de la transformación.
- C. Incremento en la frecuencia de carga.
- D. Diseño del esquema de datos.

Ciencia de Datos Aplicada

---

# Tema 4. El Perfil del Científico de Datos

# Índice

[Esquema](#)

[Ideas clave](#)

[4.1. Introducción y objetivos](#)

[4.2. Ciencias de la computación](#)

[4.3. Matemáticas](#)

[4.4. Comunicación](#)

[4.5. Negocios](#)

[4.6. Referencias bibliográficas](#)

[A fondo](#)

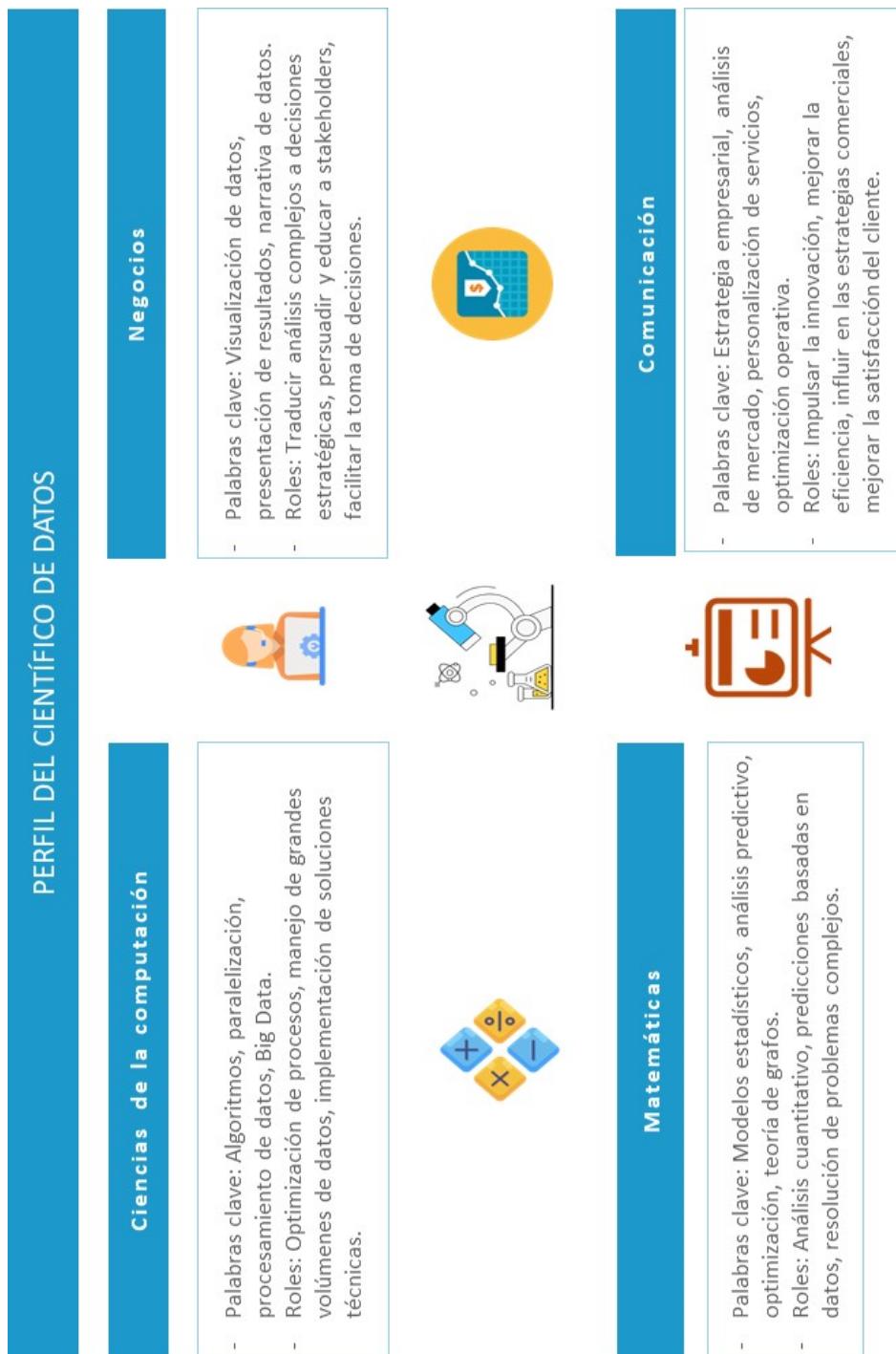
[Sobre la figura del CDO](#)

[Club CDO España](#)

[Club CDO Spain: La evolución del Chief Data Officer](#)

[How and where to find great data science jobs](#)

[Test](#)



## 4.1. Introducción y objetivos

En el mundo actual, dominado por datos y tecnología, el papel del científico de datos se ha vuelto indispensable en las organizaciones que buscan capitalizar la vasta cantidad de información disponible. Este perfil profesional combina habilidades en ciencias de la computación, matemáticas y estadística, comunicación y conocimientos de negocios para extraer patrones significativos, predecir tendencias futuras y proporcionar recomendaciones basadas en datos que impulsan las decisiones estratégicas.

La importancia de los científicos de datos radica en su capacidad para no solo manejar grandes volúmenes de datos sino también en transformar estos datos en métricas accionables que pueden traducirse en ventajas competitivas para las empresas.

Los científicos de datos actúan como puentes entre los datos técnicos y las decisiones de negocio, empleando su experiencia técnica para solucionar problemas complejos y comunicando sus hallazgos de manera efectiva a los stakeholders de la empresa. En una era donde los datos se generan a una velocidad y volumen sin precedentes, estos profesionales son clave para navegar por el ruido informativo y descubrir la información valiosa que subyace.

Los objetivos de este tema se centran en:

- ▶ Comprender la Convergencia de Habilidades en Ciencia de Datos: analizar cómo la combinación de ciencias de la computación, matemáticas, habilidades de comunicación y conocimientos de negocio forman la base del trabajo del científico de datos, y por qué cada área es crítica para el éxito en este campo.
- ▶ Identificar las Aplicaciones Prácticas de la Ciencia de Datos en Negocios: explorar casos de estudio y ejemplos reales donde los científicos de datos han transformado datos en insights valiosos que han generado un impacto significativo en las decisiones de negocio.

## 4.2. Ciencias de la computación

En el campo de la ciencia de datos, las ciencias de la computación desempeñan un papel esencial debido a que proporcionan la base técnica que permite el análisis y manejo eficaz de grandes volúmenes de datos. La ciencia de la computación ofrece las herramientas y técnicas necesarias para crear sistemas capaces de procesar, almacenar y analizar datos a una escala sin precedentes.

### Importancia de la Ciencia de la Computación en la Ciencia de Datos

- ▶ Desarrollo de Algoritmos y Modelos de Aprendizaje Automático: los científicos de datos utilizan su conocimiento en ciencias de la computación para desarrollar algoritmos complejos que pueden aprender de los datos y hacer predicciones o clasificaciones. Estos modelos son fundamentales para transformar grandes conjuntos de datos en insights accionables que pueden influir en decisiones críticas de negocio.
- ▶ Gestión de Grandes Volúmenes de Datos: con habilidades en bases de datos, estructuras de datos y algoritmos, los científicos de datos están equipados para manejar y optimizar bases de datos y sistemas de almacenamiento de datos, asegurando que los datos se almacenen de manera eficiente y sean accesibles para el análisis.
- ▶ Optimización del Rendimiento de las Consultas: el conocimiento en ciencias de la computación también permite a los científicos de datos optimizar las consultas a bases de datos para mejorar el rendimiento y la velocidad del análisis de datos, lo cual es crucial en entornos empresariales donde el tiempo de respuesta es crítico.

- ▶ Desarrollo de Software y Herramientas de Análisis: la programación es una habilidad central en ciencias de la computación que los científicos de datos utilizan para escribir scripts y desarrollar software que automatiza la recopilación, el procesamiento y el análisis de datos. Herramientas como Python y sus bibliotecas de ciencia de datos, R, y plataformas de big data como Apache Hadoop y Spark, son ejemplos de cómo se aplican estas habilidades en el día a día.

## Retos y Oportunidades

Los científicos de datos enfrentan el reto de mantenerse actualizados con los rápidos avances en tecnología de la computación y metodologías de análisis de datos. Esto implica una continua educación y adaptación a nuevas herramientas y lenguajes de programación, así como a las mejores prácticas en el desarrollo de software y la seguridad de los datos.

Al mismo tiempo, esta intersección entre ciencia de datos y ciencias de la computación ofrece oportunidades significativas. Los avances en el procesamiento y análisis de datos permiten a los científicos de datos desempeñar un papel clave en la innovación y en la creación de nuevas formas de valor a partir de los datos, lo que es crucial en la economía basada en datos de hoy.

El dominio de las ciencias de la computación es fundamental para el éxito de cualquier científico de datos, proporcionando no solo las habilidades técnicas necesarias para manejar y analizar datos a gran escala, sino también la capacidad para innovar y adaptarse a un campo que está en constante evolución.

## 4.3. Matemáticas

Las matemáticas son un pilar fundamental en el trabajo del científico de datos, proporcionando el marco teórico y las herramientas necesarias para realizar análisis precisos y efectivos de los datos. Desde estadísticas hasta modelado predictivo y optimización, las matemáticas permiten a los científicos de datos cuantificar incertidumbres, modelar complejidades y tomar decisiones basadas en evidencias sólidas.

### Aplicaciones Específicas de las Matemáticas en Ciencia de Datos

- ▶ Estadística y Probabilidades: la estadística es crucial para entender la distribución, la variabilidad y las tendencias de los datos. Por ejemplo, los científicos de datos utilizan pruebas de hipótesis para validar asunciones o inferencias, y modelos de regresión para predecir relaciones entre variables. Un ejemplo concreto es el uso de la regresión logística en la predicción de la probabilidad de que un cliente realice una compra basada en su historial de navegación web y patrones de compra.
- ▶ Análisis de Series Temporales: en sectores como las finanzas o la meteorología, el análisis de series temporales permite modelar y predecir comportamientos futuros basados en datos históricos. Un científico de datos puede utilizar modelos ARIMA (AutoRegressive Integrated Moving Average) para predecir precios de acciones o patrones climáticos, aprovechando la capacidad matemática para entender las dependencias temporales y estacionales en los datos.

- ▶ Algebra Lineal: el álgebra lineal es fundamental en muchas técnicas de machine learning, especialmente en métodos que involucran grandes volúmenes de datos. Por ejemplo, los métodos de factorización de matrices, como la descomposición en valores singulares (SVD), son esenciales para algoritmos de recomendación en sistemas como Netflix o Spotify, donde se busca predecir las preferencias del usuario reduciendo la dimensionalidad de los datos de interacción usuario-item.
- ▶ Optimización y Métodos Numéricos: los científicos de datos a menudo enfrentan problemas de optimización, como maximizar la eficiencia de una campaña de marketing o minimizar los costos en una cadena de suministro. Utilizan métodos numéricos para resolver estos problemas, aplicando técnicas de programación lineal y no lineal para encontrar soluciones óptimas que a menudo requieren el manejo de múltiples variables y restricciones.

## Retos y Sinergias

Los retos para los científicos de datos en el campo de las matemáticas incluyen la necesidad de comprender profundamente teorías complejas y aplicarlas correctamente a problemas prácticos. Esta aplicación no siempre es directa, ya que los modelos matemáticos deben ser ajustados y validados con datos reales, lo que a menudo requiere un enfoque iterativo y experimental.

La sinergia entre la ciencia de datos y las matemáticas surge de la capacidad de utilizar modelos matemáticos para estructurar y resolver problemas de datos. Esta colaboración se manifiesta en la capacidad de transformar teorías matemáticas en herramientas prácticas que pueden predecir, optimizar y analizar fenómenos en el mundo real, haciendo que los datos "hablen" y revelen insights que de otro modo permanecerían ocultos.

A continuación, se describen algunos casos de uso comunes donde las matemáticas son fundamentales:

- ▶ Predicción de Demanda: utiliza modelos estadísticos y de series temporales para prever la demanda futura de productos o servicios, permitiendo a las empresas ajustar la producción, el inventario y la planificación de la logística.
- ▶ Sistemas de Recomendación: emplea técnicas de álgebra lineal y algoritmos de aprendizaje automático para recomendar productos, películas o música a los usuarios basándose en sus intereses y comportamientos pasados.
- ▶ Detección de Fraude: aplica algoritmos de clasificación y patrones estadísticos para identificar transacciones o comportamientos anómalos que puedan indicar fraude en sectores como banca y seguros.
- ▶ Optimización de Rutas: usa algoritmos de optimización para determinar la ruta más eficiente en términos de costos y tiempo para la entrega de mercancías o la planificación de rutas de transporte público.
- ▶ Análisis de Sentimiento: implementa modelos matemáticos para analizar y clasificar opiniones de los usuarios en datos textuales, como reseñas o comentarios en redes sociales, determinando si son positivas, negativas o neutrales.
- ▶ Segmentación de Mercado: utiliza técnicas de clustering y análisis de componentes principales (PCA) para identificar segmentos de clientes basados en características similares, lo que ayuda a las empresas a dirigir sus estrategias de marketing de manera más efectiva.
- ▶ Evaluación de Riesgos: emplea modelos de regresión y simulaciones Monte Carlo para evaluar y cuantificar los riesgos financieros, como el crédito o el mercado, ayudando a las instituciones financieras en la toma de decisiones.

- ▶ Modelado de Propagación de Enfermedades: aplica modelos matemáticos de epidemiología para predecir la propagación de enfermedades y evaluar la efectividad de las intervenciones de salud pública.
- ▶ Análisis de Redes Sociales: usa teoría de grafos y algoritmos para analizar redes sociales, identificando patrones de conexión, influenciadores clave y comunidades dentro de las redes.
- ▶ Valoración de Activos: emplea modelos financieros y estadísticos para determinar el valor justo de diversos activos, incluyendo acciones, bonos y derivados.

Cada uno de estos casos de uso demuestra cómo las matemáticas no solo proporcionan la base para manejar grandes conjuntos de datos, sino que también aportan la estructura necesaria para modelar y resolver problemas complejos, haciendo que los científicos de datos puedan proporcionar soluciones efectivas y basadas en evidencia en diversos contextos.

Para el científico de datos las matemáticas no solo enriquecen su capacidad analítica, sino que también amplían el alcance y la precisión de las investigaciones en ciencia de datos, permitiendo abordajes más sofisticados y soluciones más eficaces a problemas complejos.

## 4.4. Comunicación

La comunicación es una habilidad crítica en la ciencia de datos, vital para el éxito de cualquier proyecto de análisis de datos. El científico de datos no solo necesita ser competente en técnicas estadísticas y de programación, sino también en la habilidad de comunicar hallazgos complejos de manera clara y persuasiva a un público diverso que puede incluir expertos técnicos, ejecutivos de negocio y otros stakeholders no técnicos.

### Comunicación de Resultados

Uno de los roles principales del científico de datos en el campo de la comunicación es traducir los resultados técnicos en indicadores que puedan ser comprendidos por todos los stakeholders del proyecto. Esto implica presentar los datos de manera que resalten las conclusiones clave sin perderse en detalles técnicos innecesarios.

Herramientas como visualizaciones de datos, dashboards interactivos y presentaciones claras son fundamentales. Por ejemplo, un científico de datos puede usar una visualización de gráfico de calor para demostrar áreas de alta actividad en un estudio de mercado, facilitando la comprensión rápida de datos complejos.

## Presentación de Avances de los Proyectos de Datos

Durante la gestión de proyectos de datos, comunicar los avances de manera efectiva asegura que todas las partes interesadas estén informadas sobre el progreso, los desafíos y los cambios en los objetivos del proyecto. Esto es crucial para mantener alineados a todos los miembros del equipo y para gestionar las expectativas de los stakeholders. Ejemplo de esto sería la actualización periódica a través de reuniones regulares donde se presentan métricas de progreso y se discuten las necesidades de ajustes en la estrategia o recursos del proyecto.

## Impacto de los Proyectos en la Sociedad

Los científicos de datos también tienen la responsabilidad de comunicar cómo los proyectos de datos impactan en la sociedad. Esto puede incluir la discusión de implicaciones éticas, privacidad de datos y potenciales beneficios o daños. Por ejemplo, en proyectos que involucran datos de salud pública, es vital comunicar cómo se manejan y protegen los datos para evitar preocupaciones sobre privacidad.

Además, es importante resaltar los beneficios, como mejoras en la atención sanitaria o en la eficiencia de servicios públicos, que pueden convencer y tranquilizar al público sobre la utilidad y seguridad de tales proyectos.

La comunicación en ciencia de datos es esencial no solo para la ejecución efectiva de los proyectos, sino también para asegurar que los resultados son entendidos, valorados y aplicados correctamente. Un científico de datos eficaz debe ser capaz de narrar una historia con los datos, destacando su relevancia y asegurando que las métricas derivadas conduzcan a decisiones informadas y mejoren las intervenciones y políticas basadas en evidencia.

## Aspectos relevantes de la comunicación según el ámbito

La comunicación en ciencia de datos varía considerablemente entre diferentes sectores, reflejando las necesidades específicas, expectativas y sensibilidades de cada ámbito. Vamos a explorar cómo se adapta la comunicación en los sectores sanitario, gubernamental, ético y de sostenibilidad:

### Sector Sanitario

En el ámbito sanitario, la comunicación debe manejar con cuidado la privacidad y la sensibilidad de la información personal de salud. Los científicos de datos que trabajan en este sector a menudo deben presentar sus hallazgos de manera que respeten la confidencialidad y el cumplimiento de normativas como HIPAA en EE. UU. o GDPR en Europa.

Por ejemplo, al comunicar resultados de investigaciones clínicas, es esencial utilizar un lenguaje que sea tanto médicaamente preciso como accesible para los no médicos, incluyendo a pacientes y administradores. Visualizaciones claras y comprensibles ayudan a demostrar cómo los análisis pueden mejorar los tratamientos o la eficiencia operativa sin comprometer la seguridad del paciente.

### Sector Gubernamental

La comunicación en el sector gubernamental debe ser transparente y diseñada para fomentar la confianza pública. Los proyectos de ciencia de datos pueden abarcar desde la optimización de los servicios públicos hasta la seguridad nacional. Aquí, es crucial que los científicos de datos comuniquen cómo los datos son recolectados, procesados y utilizados, asegurando que se mantienen dentro de los límites éticos y legales.

Al presentar datos a funcionarios o al público, es importante hacerlo de una manera que sea fácilmente interpretable, con énfasis en cómo los resultados del análisis benefician o impactan a la comunidad.

## Ética en Ciencia de Datos

La comunicación en temas de ética implica discutir cómo se manejan los datos y las implicaciones de los proyectos de ciencia de datos. Esto incluye temas como el sesgo en los algoritmos y la equidad en el análisis de datos. Los científicos de datos deben ser capaces de explicar y justificar las decisiones metodológicas y las salvaguardas que se implementan para prevenir resultados perjudiciales o injustos. Es esencial que la comunicación en este ámbito sea abierta y honesta, promoviendo una cultura de responsabilidad y reflexión continua sobre el impacto de la ciencia de datos.

## Sostenibilidad

En el campo de la sostenibilidad, la comunicación se centra en cómo los proyectos de datos pueden ayudar a resolver problemas ambientales o mejorar la eficiencia de recursos.

Los científicos de datos deben comunicar sus hallazgos de manera que resalten la importancia de la acción sostenible y cómo la tecnología puede ser una parte de la solución. Esto podría incluir la visualización de tendencias de consumo de energía, la optimización de rutas para reducir emisiones, o el análisis de patrones climáticos para predecir desastres naturales. La comunicación debe ser motivadora y dirigida a fomentar cambios en comportamientos y políticas.

Cada uno de estos sectores presenta desafíos únicos en términos de cómo los datos deben ser comunicados. Lo esencial es que la comunicación en ciencia de datos debe ser adaptable, precisa y sensible al contexto, asegurando que los datos no solo se entiendan y utilicen correctamente, sino que también se manejen de manera ética y responsable.

## 4.5. Negocios

En el sector empresarial, la ciencia de datos se ha convertido en un componente crítico para la toma de decisiones, la optimización de procesos, y la innovación de productos y servicios. Los científicos de datos en el ámbito de los negocios no solo proporcionan análisis y modelos basados en datos, sino que también deben comunicar sus hallazgos de manera que influyan en las estrategias y operaciones de la empresa. Veamos cómo se manifiesta esto en empresas relevantes y casos de uso reales.

### Ejemplos de Empresas y Casos de Uso

- ▶ Amazon - Optimización de la Cadena de Suministro: Amazon utiliza la ciencia de datos para optimizar su inmensa cadena de suministro, desde la gestión de inventarios hasta la logística de entrega. Mediante el uso de algoritmos predictivos, Amazon puede prever la demanda de productos, ajustar los niveles de stock en tiempo real y optimizar las rutas de entrega para reducir costos y tiempos de envío. Este enfoque basado en datos no solo mejora la eficiencia operativa, sino que también eleva la satisfacción del cliente.
- ▶ Netflix - Personalización de Contenidos: emplea técnicas avanzadas de análisis de datos para personalizar las recomendaciones de contenido a sus usuarios. Utilizando algoritmos de aprendizaje automático, Netflix analiza los patrones de visualización y preferencias de los usuarios para sugerir películas y series que probablemente sean de su interés. Esta personalización aumenta la retención de usuarios y también es central en la estrategia de marketing y contenido de Netflix.
- ▶ Starbucks - Análisis de Ubicación: utiliza modelos de ciencia de datos para determinar las ubicaciones óptimas de sus nuevas tiendas. Analizando datos demográficos, de tráfico y de la competencia, junto con el rendimiento de tiendas

existentes, Starbucks puede seleccionar ubicaciones que maximicen la visibilidad y el tráfico de clientes potenciales.

- ▶ Goldman Sachs - Análisis de Riesgos Financieros: en el sector financiero, empresas como Goldman Sachs aplican modelos predictivos para analizar el riesgo de crédito, de mercado y operacional. Utilizando técnicas de análisis cuantitativo, pueden predecir y mitigar riesgos financieros, lo que es crucial para la toma de decisiones de inversión y la gestión de carteras.

## Comunicación e Impacto en los Negocios

La comunicación en este ámbito implica traducir complejos análisis de datos en decisiones estratégicas y tácticas. Los científicos de datos deben presentar sus hallazgos de manera que sean accesibles para los ejecutivos y tomadores de decisiones, a menudo a través de visualizaciones de datos e informes concisos que destaque las implicaciones comerciales de los análisis. Además, deben asegurarse de que sus comunicaciones sigan siendo alineadas con los objetivos estratégicos del negocio, promoviendo una cultura basada en datos.

La ciencia de datos en el ámbito empresarial es fundamental para impulsar la innovación, optimizar operaciones y personalizar las interacciones con los clientes. Las empresas que invierten en capacidades de ciencia de datos no solo obtienen una ventaja competitiva significativa, sino que también están mejor equipadas para responder a las dinámicas cambiantes del mercado. Los científicos de datos, como puente entre los datos crudos y las decisiones estratégicas, son así actores clave en el éxito empresarial moderno.

## 4.6. Referencias bibliográficas

Davenport, T. H. y Patil, D. J. (2012). *Data scientist: the sexiest job of the 21st century*. Massachussets: Harvard Business Review.

Chatfield, A. T., Shlemon, V. N., Redublado, W. y Rahman, F. (2014). Data scientists as game changers in big data environments. *Proceedings of the 25th Australasian conference on information systems (ACIS)* (pp. 1-11). New Zealand: Auckland University of Technology.

Mikalef, P., Framnes, V., Danielsen F., Krogstie, J. y Olsen, D.H. (2017). Big data analytics capability: antecedents and business value. *Proceedings of the 21st Pacific Asia conference on information systems (PACIS)*, 137.

Mikalef, P., Pappas, I. O., Krogstie, J. et al. (2018). Big data analytics capabilities: a systematic literature review and research agenda. *Inf Syst E-Bus Manage*, 16, 547.

Mohanty, S., Jagadeesh, M. y Srivatsa, H. (2013). *Big data imperatives: enterprise «Big Data» warehouse, BI implementations and analytics*. Nueva York: Apress.

Prescott, E. M. (2014). Big data and competitive advantage at Nielsen. *Management Decision*, 52(3), 573–601.

## Sobre la figura del CDO

CDO Club. Página oficial. <https://cdoclub.com/>

Club. Página oficial del Club CDO (*Chief Digital Officer*), la comunidad internacional con más miembros inscritos y que proporciona información sobre eventos, puestos de trabajo relacionados, publicaciones y diferentes materiales de apoyo.

## Club CDO España

Chief Data Officer Club Spain. Página oficial. <http://clubcdo.com>

Página web oficial de la asociación de CDO de empresas españolas creada con el objetivo de ser una red de apoyo para estos profesionales.

## Club CDO Spain: La evolución del Chief Data Officer

The Valley. (1 de marzo de 2019). Club CDO Spain: La evolución del Chief Data Officer (Blog post). Recuperado de <https://thevalley.es/blog/chief-data-officer-evolucion/>

En este artículo se explica brevemente la figura del *chief data officer* y su evolución, además de realizar un pronóstico sobre esta figura para dentro de cinco años.

## How and where to find great data science jobs

---

Custer, C. (29 de marzo de 2019). How and where to find great data science jobs (Blog post). Recuperado de <https://www.dataquest.io/blog/career-guide-find-data-science-jobs/>

En este artículo se explica de forma pormenorizada cómo y dónde encontrar puestos de empleo como *data scientist*.

1. ¿Qué técnica utilizan las empresas como Amazon para optimizar su cadena de suministro?

  - A. Análisis de sentimiento.
  - B. Algoritmos predictivos.
  - C. Minería de texto.
  - D. Análisis de redes sociales.
  
2. ¿Qué método utiliza Netflix para personalizar las recomendaciones a sus usuarios?

  - A. Clustering.
  - B. Regresión lineal.
  - C. Aprendizaje automático.
  - D. Análisis factorial.
  
3. ¿Qué herramienta matemática es crucial en la evaluación de riesgos financieros en empresas como Goldman Sachs?

  - A. Cálculo diferencial.
  - B. Teoría de grafos.
  - C. Modelos predictivos.
  - D. Geometría analítica.
  
4. En el contexto de la ciencia de la computación, ¿qué técnica es fundamental para el procesamiento de grandes volúmenes de datos?

  - A. Programación funcional.
  - B. Algoritmos de ordenamiento.
  - C. Paralelización de procesos.
  - D. Uso de variables estáticas.

5. ¿Qué técnica matemática es ampliamente utilizada para segmentar audiencias en marketing digital?
  - A. Cálculo integral.
  - B. Análisis de cluster.
  - C. Álgebra lineal.
  - D. Probabilidad y estadística.
6. ¿Cómo contribuye la teoría de grafos en la ciencia de datos?
  - A. Optimización de algoritmos de búsqueda.
  - B. Mejora de interfaces gráficas.
  - C. Análisis de redes sociales.
  - D. Desarrollo de juegos.
7. ¿Qué metodología utiliza Starbucks para determinar las ubicaciones óptimas para sus nuevas tiendas?
  - A. Simulaciones Monte Carlo.
  - B. Análisis de ubicación.
  - C. Modelos de regresión.
  - D. Análisis de la competencia.
8. ¿Cuál es un ejemplo de aplicación de regresión lineal en ciencia de datos?
  - A. Predecir el precio futuro de las acciones.
  - B. Codificar datos para algoritmos de cifrado.
  - C. Crear gráficos interactivos.
  - D. Diseñar bases de datos.

**9.** ¿Qué representa la comunicación efectiva de los resultados en proyectos de ciencia de datos en negocios?

- A. Publicar papers académicos.
- B. Convencer a los stakeholders del valor de los hallazgos.
- C. Implementar directamente los cambios en la producción.
- D. Ninguna de las anteriores.

**10.** ¿Qué rol juegan las visualizaciones de datos en la comunicación científica de datos?

- A. Solo para presentaciones académicas.
- B. Para simplificar el código.
- C. Para hacer los hallazgos comprensibles y accesibles.
- D. Para aumentar la carga computacional.

Ciencia de Datos Aplicada

---

# Tema 5. Áreas de aplicación de la Ciencia de Datos

# Índice

## Esquema

### Ideas clave

- 5.1. Introducción y objetivos
- 5.2. Comercio
- 5.3. Industria
- 5.4. Salud
- 5.5. Seguridad y Ciberseguridad
- 5.6. Finanzas
- 5.7. Conducción Autónoma
- 5.8. Otros
- 5.9. Referencias bibliográficas

### A fondo

La aplicación de inteligencia artificial a la analítica de datos

La ética del Data

¡El futuro pasa por el big data!

## Test

ÁREAS DE APLICACIÓN DE LA CIENCIA DE DATOS	
COMERCIO	SALUD
FINANZAS	
<ul style="list-style-type: none"> <li>- Palabras clave: Optimización de precios, gestión de inventario, personalización, análisis de clientes.</li> <li>- Técnicas: Algoritmos predictivos, análisis de clúster, machine learning, sistemas de recomendación.</li> </ul>	<ul style="list-style-type: none"> <li>- Palabras clave: Predicción de enfermedades, análisis de registros médicos, epidemiología, personalización de tratamientos.</li> <li>- Técnicas: Modelos predictivos, machine learning, análisis de series temporales, simulación Monte Carlo.</li> </ul>
<ul style="list-style-type: none"> <li>- Palabras clave: Mantenimiento predictivo, optimización de producción, control de calidad, cadena de suministro.</li> <li>- Técnicas: Machine learning, visión por computadora, análisis predictivo, optimización de procesos.</li> </ul>	<ul style="list-style-type: none"> <li>- Palabras clave: Detección de amenazas, análisis de comportamiento, prevención de fraudes, respuesta a incidentes.</li> <li>- Técnicas: Análisis de comportamiento, machine learning, detección de anomalías, inteligencia artificial.</li> </ul>
	<ul style="list-style-type: none"> <li>- Palabras clave: Percepción sensorial, predicción de comportamiento, optimización de rutas, entrenamiento de sistemas.</li> <li>- Técnicas: Visión por computadora, redes neuronales profundas, análisis predictivo, simulaciones.</li> </ul>

## 5.1. Introducción y objetivos

La implantación en nuestro día a día de las nuevas tecnologías permite que dispongamos de una creciente oferta de servicios centrados en el usuario y automatizados. Esto lleva a que gran número de organizaciones públicas y privadas recojan, procesen, almacenen y hagan uso de grandes cantidades de datos que nos ayudan a resolver infinidad de tareas de un modo antes inimaginable. Los sistemas *big data* dan cobertura al procesamiento y la gestión de la información en todos los ámbitos de nuestra sociedad, como son el entretenimiento, el sanitario, la industria, la investigación, la sociología, la política o el financiero, entre muchos otros.

Esta amplitud en el uso de herramientas de ciencia de datos no está exenta de riesgos. A lo largo de los últimos años son comunes las noticias relacionadas con la filtración de datos o de un uso inapropiado de la información de la que disponen muchas empresas.

Entre estos riesgos se encuentra la gestión de la información personal, un aspecto en el que los Gobiernos están poniendo especial énfasis en regular, como en el caso del Reglamento General de Protección de Datos (RGPD) de la Unión Europea (Reglamento (UE) 2016/679) o el *California Consumer Privacy Act* (Departamento de Justicia del Estado de California, 2021).

Estas regulaciones son solo dos ejemplos en los que se pretende reducir el riesgo a la relación entre los sistemas *big data* y la sociedad, ya que un uso inoportuno de la información puede desembocar en resultados devastadores, suponiendo una amenaza real para la sociedad y el entorno que nos rodea (Yamin, 2019).

El objetivo principal de este tema es introducir algunos de los ámbitos de aplicación de la ciencia de datos y, por ende, de los sistemas *big data* dentro de nuestra sociedad. Una vez conocidos estos ámbitos, se destacarán las implicaciones más importantes que esta ciencia y su trabajo tienen sobre la sociedad. Finalmente, se introducirán algunas de las implicaciones éticas más relevantes que derivan de su intervención en la sociedad.

## 5.2. Comercio

La ciencia de datos ha revolucionado el sector comercial al permitir una toma de decisiones más informada y estratégica. Mediante el análisis de grandes volúmenes de datos, las empresas comerciales pueden optimizar operaciones, personalizar experiencias de clientes y mejorar la eficiencia en la cadena de suministro. La capacidad de extraer insights valiosos de los datos es ahora un diferenciador competitivo clave en el comercio.

Las técnicas avanzadas de ciencia de datos, como el machine learning, la inteligencia artificial, y la analítica predictiva, están al frente de esta transformación. Empresas líderes en el sector comercial utilizan estas tecnologías para abordar desafíos específicos y mejorar su rendimiento. A continuación, se presentan ejemplos específicos y concretos de la aplicación de la ciencia de datos en el comercio, resaltando los beneficios obtenidos.

### Optimización de Precios Dinámicos

- ▶ Ejemplo: Amazon utiliza modelos de precios dinámicos basados en machine learning para ajustar los precios de millones de productos en tiempo real, considerando factores como la demanda, el inventario, y la competencia.
- ▶ Beneficio: Esta estrategia permite maximizar los ingresos y mantener la competitividad de precios, adaptando los precios a las condiciones cambiantes del mercado.

## Gestión de Inventario y Cadena de Suministro

- ▶ Ejemplo: Walmart implementa algoritmos de forecasting para prever la demanda futura de productos a nivel de tienda. Este enfoque utiliza datos históricos de ventas, tendencias de mercado, y eventos especiales para anticipar las necesidades de stock.
- ▶ Beneficio: La mejora en la precisión de la previsión reduce los costos de sobre inventario y falta de stock, optimizando así la gestión de la cadena de suministro.

## Personalización de la Experiencia del Cliente

- ▶ Ejemplo: Starbucks usa análisis predictivo para ofrecer recomendaciones personalizadas a través de su aplicación móvil. El sistema analiza las compras anteriores y las preferencias de los usuarios para sugerir nuevos productos que podrían gustar al cliente.
- ▶ Beneficio: Incremento en la satisfacción del cliente y aumento de las ventas por la promoción de productos ajustados a los gustos individuales de los consumidores.

## Detección de Fraudes y Anomalías

- ▶ Ejemplo: PayPal emplea modelos de machine learning para detectar y prevenir fraudes en transacciones en tiempo real. Estos modelos analizan patrones de comportamiento anormal y comparan transacciones contra perfiles de fraude conocidos.
- ▶ Beneficio: Reducción de pérdidas financieras y aumento de la confianza del consumidor en la plataforma, asegurando transacciones más seguras.

La figura 1 muestra un cuadro resumen de cómo los sistemas *big data* y la ciencia de datos ayudan en diferentes áreas de múltiples sectores. Esto nos lleva a identificar la

creciente y urgente necesidad de «pensar en grande para crear el importante valor añadido que se espera de los sistemas *big data*, al tiempo que se establecen las políticas y los mecanismos necesarios para garantizar que la nueva era siga respetando los derechos de las personas, así como proponer y gestionar políticas y mecanismos necesarios para garantizar que la nueva era que se avecina siga respetando los derechos individuales» (Thonnet y Nicolas, 2017).

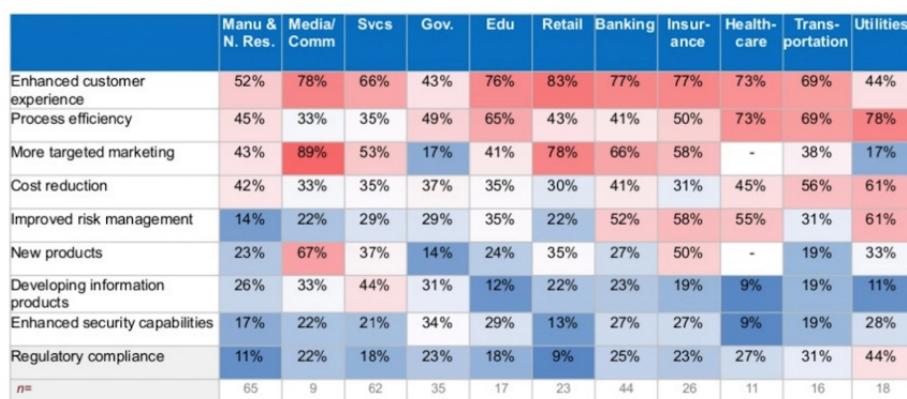


Figura 1. Mapa de calor de los problemas que ayudan a resolver los sistemas *big data* por sectores.

Fuente: Incibe-cert, 2016.

A lo largo de las siguientes subsecciones se desgranarán las implicaciones que los sistemas *big data* tienen en tres de los ámbitos más importantes de nuestra sociedad, como son el gubernamental, el económico, el financiero y bancario, así como el de la salud.

La aplicación de la ciencia de datos en el comercio no solo facilita operaciones más eficientes, sino que también mejora la interacción con el cliente, personalizando experiencias y anticipando sus necesidades.

El uso estratégico de los datos se ha convertido en un imperativo para las empresas que buscan mantenerse competitivas en un mercado cada vez más digitalizado y orientado al análisis de datos.

## 5.3. Industria

Uno de los pilares de la cuarta revolución industrial (Industria 4.0) ha sido la implementación de sistemas *big data* que han permitido hacer un uso más eficiente de los datos empresariales a todos los niveles, gestionar la ingente sensorización de entornos industriales, vehículos o robots, o mejorar la toma de decisiones a través de la implementación de soluciones *business intelligence*.

El sector industrial ha experimentado una transformación significativa con la incorporación de la ciencia de datos, que permite optimizar procesos, reducir costos operativos y mejorar la calidad del producto. Las técnicas avanzadas de analítica y machine learning son esenciales para aumentar la eficiencia y la sostenibilidad en la manufactura y otras actividades industriales.

### Estado del Arte en la Ciencia de Datos Industrial

La integración de IoT (Internet de las Cosas) con modelos de machine learning y sistemas de análisis predictivo está revolucionando la manera en que las industrias operan. A continuación, se presentan ejemplos específicos de cómo la ciencia de datos está siendo aplicada en el sector industrial, destacando los beneficios concretos.

### Mantenimiento Predictivo

- ▶ Ejemplo: General Electric utiliza sensores en sus máquinas y analítica predictiva para predecir fallos antes de que ocurran. Los datos recogidos de las máquinas se analizan utilizando modelos predictivos que estiman la probabilidad de fallos y sugieren mantenimientos preventivos.

- ▶ Beneficio: Esto reduce el tiempo de inactividad no planificado y los costos asociados con paradas de emergencia y reparaciones costosas, mejorando la eficiencia operativa.

## Optimización de la Producción

- ▶ Ejemplo: Siemens implementa sistemas avanzados de simulación y modelos de optimización para ajustar los procesos de producción en tiempo real. Estos modelos analizan variables como velocidad de producción, consumo de energía y calidad del producto para optimizar el rendimiento de las líneas de montaje.
- ▶ Beneficio: Mejora la eficiencia energética y la productividad, al tiempo que asegura la calidad del producto final.

## Control de Calidad Automatizado

- ▶ Ejemplo: Fabricantes de componentes electrónicos utilizan técnicas de visión por computadora y machine learning para inspeccionar visualmente las placas de circuitos en las líneas de montaje. Estos sistemas son capaces de detectar imperfecciones y componentes defectuosos con alta precisión.
- ▶ Beneficio: Reduce el número de productos defectuosos y mejora la confiabilidad del producto, lo que a su vez aumenta la satisfacción del cliente.

## Gestión de la Cadena de Suministro

- ▶ Ejemplo: Caterpillar utiliza modelos de ciencia de datos para gestionar su compleja cadena de suministro global. Los modelos predictivos y de simulación ayudan a anticipar retrasos en los suministros y ajustar los planes de producción en consecuencia.

- ▶ Beneficio: Asegura la continuidad de la producción y minimiza los costos adicionales debido a retrasos o escasez de insumos.

La ciencia de datos en el sector industrial mejora la eficiencia operativa y contribuye en una mayor sostenibilidad y reducción de costos. Las empresas industriales que adoptan estas tecnologías están mejor posicionadas para responder a las demandas cambiantes del mercado y mantener una ventaja competitiva en un entorno cada vez más tecnológico y automatizado.

## 5.4. Salud

El tercer ámbito que destacamos en esta sección es la salud. El *big data* dentro de la medicina suele referirse a los registros sanitarios, hospitalarios, de tratamiento, reclamaciones médicas, datos administrativos, datos de ensayos clínicos, aplicaciones para teléfono o *tablet*, datos recogidos por dispositivos médicos, sistemas de identificación de pacientes, datos de investigación, etcétera.

Se espera que la cantidad de datos sobre la salud a nivel global aumente drásticamente en los próximos años. Las primeras estimaciones hechas hacia 2013 sugieren que hubo alrededor de 153 *exabytes* de datos de salud generados en ese año, proyectando una generación de hasta 2314 *exabytes* de nuevos datos generados en 2020 (Steward, 2020). Parece obvio que el manejo de los datos que la digitalización del sistema sanitario maneja requiere de sistemas *big data* y de la aplicación de la ciencia de datos.

La aplicación de múltiples técnicas relacionadas con el *big data* permite que los servicios relacionados con los sistemas de salud mejoren sustancialmente. Por ejemplo, con la inteligencia artificial (IA) se pueden realizar simulaciones de diagnósticos o análisis de imágenes que ayuden a la toma de decisiones de los médicos.

En un contexto de consulta, los médicos pueden analizar pruebas pasadas, tendencias y datos actuales. Con esta información y el acceso a los datos relativos al estilo de vida, el historial y la genética de un paciente, se facilita una imagen holística que puede ayudar a los médicos a proporcionar la atención más adecuada.

Por otro lado, del mismo modo en que las empresas han adoptado políticas *customer-centric* para personalizar la atención a sus clientes, la asistencia sanitaria puede también proveer este tipo de servicio. Los sistemas sanitarios disponen de

suficientes datos personales sobre cada paciente como para ofrecer un tratamiento personalizado. Además, los pacientes podrían desempeñar un papel más activo en su salud y bienestar proveyendo información a través de dispositivos inteligentes, monitorizando a aquellas personas que presenten patologías de riesgo o demencias, o simplemente optando a una mejor interpretación de la información que ha sido adaptada a sus necesidades específicas.

La investigación médica también se ha visto beneficiada por los sistemas *big data*. Sirvan como ejemplo los trabajos llevados a cabo para la secuenciación del genoma humano. A medida que la tecnología *big data* se expande, la posibilidad de procesar cada vez cantidades mayores de datos hace posible descubrir nuevas partes de este. A medida que se secuencien más personas, los científicos y los médicos tendrán acceso a un conjunto de datos más amplio para conocer los genes que antes no podían entender, y así identificar la relación existente entre ciertos genomas y enfermedades.

El reto principal que presentan los sistemas de salud es cómo gestionar un abanico muy amplio de datos que, además, son extremadamente sensibles. La figura 4 muestra la complejidad de este ecosistema y trata de ordenar y estructurar qué tipo de información manejan estos sistemas.

De su análisis cabe destacar algunos aspectos interesantes que determinarán cómo gestionar la información. Por un lado, **la diversidad en lo que a la procedencia se refiere**: las fuentes de datos son múltiples y heterogéneas, lo que presenta grandes retos a la hora de trabajar con ellas, especialmente de forma conjunta. Por otro lado, puede observarse que existen también una **gran variedad en los consumidores de esta información**, desde médicos a pacientes, pasando por investigadores o empresas terceras que proveen servicios.

Finalmente, puede comprobarse que los retos no solo se plantean a nivel técnico, sino que el análisis y tratamiento de los datos también será complejo, sin olvidar que su gobierno y regulación han de ser especialmente importantes debido a la delicadeza de esta información personal.

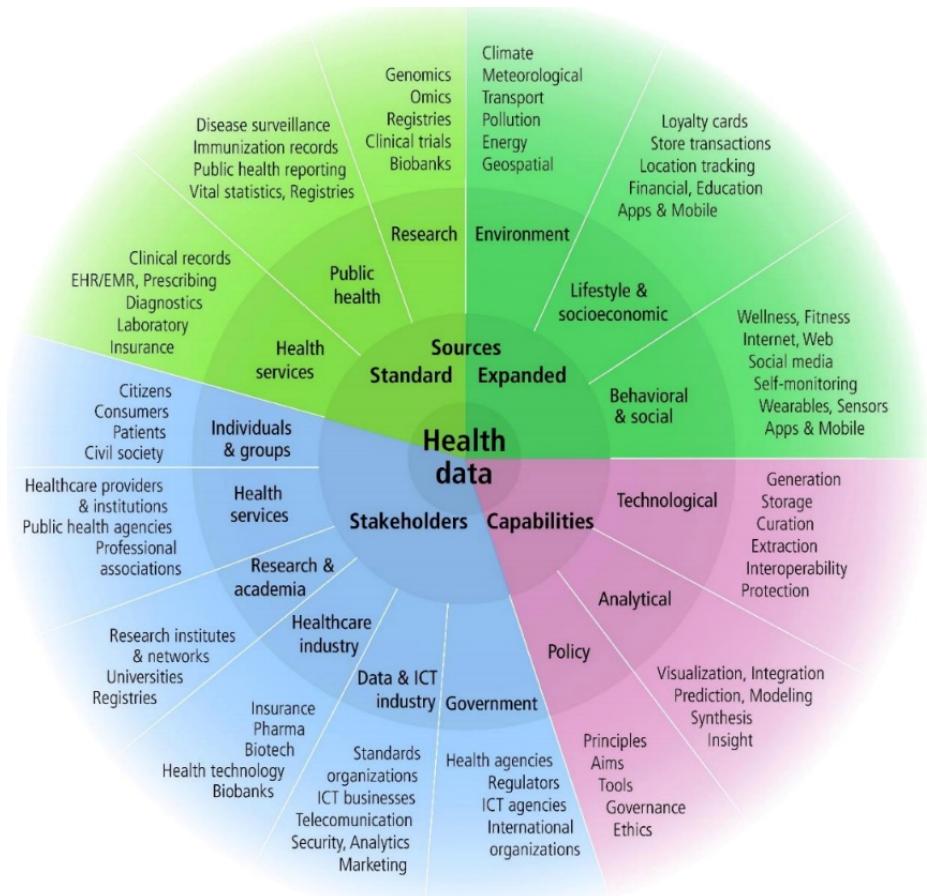
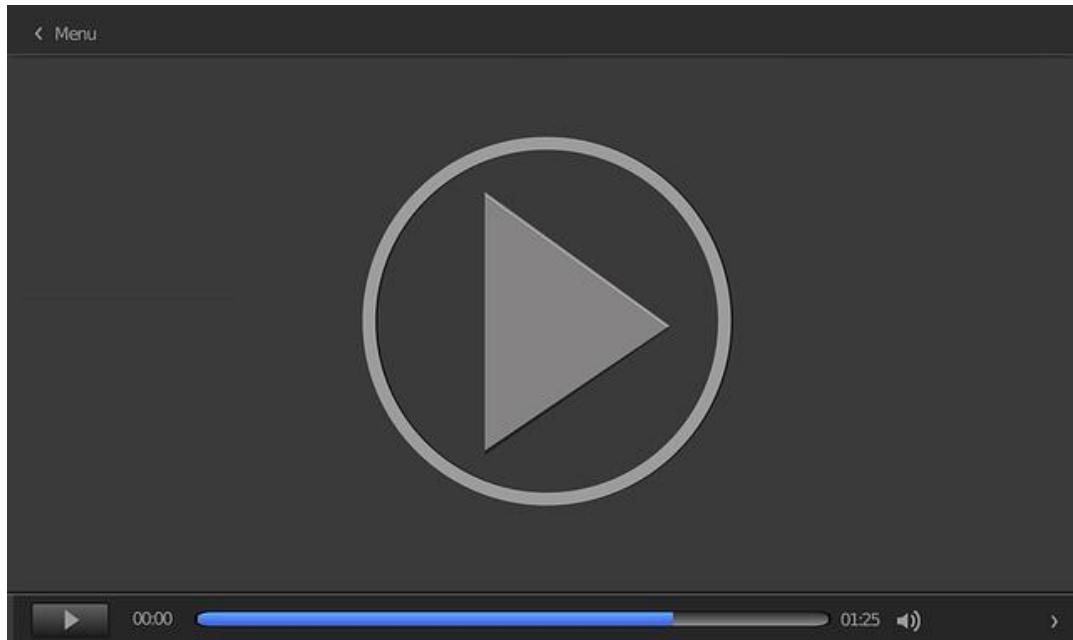


Figura 2. Ecosistema de la digitalización de los sistemas de salud y la implicación de los datos en ellos.

Fuente: Thonnet y Nicolas, 2017.

Observemos una introducción al Internet de las cosas, al Internet de las cosas industrial y su combinación con las *Aplicaciones de la ciencia de datos en la Industria 4.0.*



---

Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=a28f530f-8e39-472d-812a-b16f00e04cbc>

---

## 5.5. Seguridad y Ciberseguridad

En el ámbito de la seguridad y ciberseguridad, la ciencia de datos ha emergido como una herramienta crítica para detectar, prevenir y responder a amenazas. Utilizando algoritmos de machine learning y análisis de grandes volúmenes de datos, las organizaciones pueden identificar patrones anómalos y comportamientos sospechosos que podrían indicar intentos de intrusión o brechas de seguridad.

El uso de técnicas avanzadas como el aprendizaje profundo, el análisis de redes y la inteligencia artificial está transformando la forma en que las organizaciones protegen sus activos digitales. A continuación, se presentan ejemplos específicos de cómo la ciencia de datos está siendo aplicada en seguridad y ciberseguridad, destacando los beneficios directos.

### Detección de Amenazas en Tiempo Real

- ▶ Ejemplo: Cisco utiliza sistemas de detección de intrusiones alimentados por machine learning para monitorear el tráfico de red en busca de signos de actividad maliciosa. Estos sistemas aprenden de millones de eventos de seguridad para detectar anomalías que podrían pasar desapercibidas por métodos tradicionales.
- ▶ Beneficio: Permite una respuesta más rápida a las amenazas, reduciendo el tiempo de exposición a ataques y minimizando potenciales daños.

## Análisis de Comportamiento del Usuario

- ▶ Ejemplo: Exabeam aplica modelos de análisis de comportamiento para identificar acciones de usuarios que se desvían de patrones normales de trabajo. Al analizar datos de comportamiento de usuario, puede identificar señales tempranas de cuentas comprometidas o insiders maliciosos.
- ▶ Beneficio: Incrementa la seguridad interna y ayuda a prevenir el robo de datos desde el interior de la organización.

## Phishing y Prevención de Fraudes

- ▶ Ejemplo: Empresas como PayPal utilizan técnicas de aprendizaje automático para analizar el comportamiento de las transacciones y detectar intentos de fraude y phishing. Estos modelos examinan detalles de la transacción, el dispositivo usado, la ubicación del usuario y patrones históricos de transacciones.
- ▶ Beneficio: Mejora la capacidad de distinguir actividades legítimas de intentos de fraude, protegiendo tanto a los usuarios como a la infraestructura financiera de la empresa.

## Respuesta Automatizada a Incidentes

- ▶ Ejemplo: Symantec ha desarrollado sistemas que utilizan inteligencia artificial para automatizar la respuesta a incidentes de ciberseguridad. Estos sistemas pueden ejecutar acciones correctivas predefinidas basadas en el tipo y la gravedad del ataque detectado.
- ▶ Beneficio: Reduce la carga sobre los equipos de respuesta a incidentes y acelera la resolución de problemas, manteniendo la integridad del sistema de TI.

La integración de la ciencia de datos en la seguridad y ciberseguridad es fundamental para adelantarse a las amenazas en un panorama digital que evoluciona rápidamente. Las organizaciones que adoptan estas tecnologías mejoran su capacidad de defensa y optimizan la gestión de recursos de seguridad, lo que resulta en un entorno más seguro y resiliente.

## 5.6. Finanzas

Los sectores económico, financiero y bancario han experimentado cambios importantes y profundos gracias a la aplicación de los sistemas *big data*, los cuales eran difíciles de imaginar no hace mucho tiempo. Como ejemplo sencillo, **el *big data* permite a los sistemas financieros disponer de información en tiempo real de cualquier mercado en cualquier parte del mundo.**

El análisis de estos datos permite reducir riesgos, mientras que la aplicación de técnicas de *machine learning* permite predecir el valor del precio de una acción en el futuro. De igual modo, el análisis de datos ha ayudado a los bancos a entender mucho mejor el comportamiento de los clientes, basándose en los datos recibidos de los patrones de inversión, las tendencias de compra, la motivación para invertir y los antecedentes personales o financieros.

Bajo el paraguas de los sistemas *big data* y tecnologías como *blockchain* han surgido en este ámbito las *fintech*. Maestre (2020) define a estas como «empresas financieras tecnológicas que tratan de aportar nuevas ideas y que reformulan gracias a las nuevas tecnologías de la información, las aplicaciones móviles o el *big data*, la forma de entender y prestar los servicios financieros».

Este tipo de empresas cubre unas áreas o sectores innovadores que la banca tradicional tardaría más tiempo en implementarlos, como son las criptomonedas, los préstamos *peer-to-peer* (P2P), la microfinanciación o las herramientas de *trading*.

Para ilustrar mejor el impacto que el *big data* y la ciencia de datos tienen en el sector bancario, profundizaremos en ocho áreas de especial interés para el sector, tal y como ilustra la figura 3.

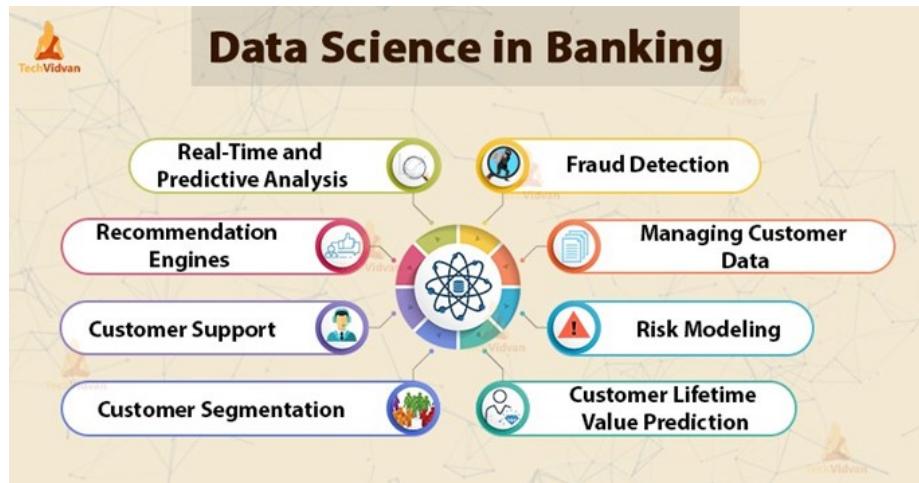


Figura 3. Uso de la ciencia de datos en el sector bancario. Fuente: TechVidvan, 2020.

- ▶ Gestión de datos en tiempo real y análisis predictivo. Las transacciones que se llevan a cabo en los sistemas bancarios son una gran fuente de información. Sobre ellas se pueden aplicar técnicas que permiten extraer información o hacerlas seguras. Pueden destacarse la **analítica en tiempo real**, que permite determinar a los bancos la fiabilidad de la transacción, detectando fraudes o ataques, y la **analítica predictiva**, mediante la que se ayuda a los bancos a identificar patrones de riesgo o consumo.
- ▶ Detección de fraude. La monitorización de transacciones y su posterior análisis deriva en una mejora en la detección de fraude. En este sentido, **el análisis de datos históricos y análisis forenses de ataques o fraudes previos permite que los bancos estén más preparados** para repeler situaciones en los que exista riesgo de fraude. Las técnicas de análisis de datos y *machine learning* son algunas de las más utilizadas para este propósito.

- ▶ Gestión de datos de clientes. La gestión de datos de clientes ha evolucionado enormemente gracias a los sistemas *big data*. Ahora son más y de mejor calidad los datos que se recogen de ellos. La ciencia de datos provee las herramientas y técnicas que permiten una mejor gestión y extracción de información de ellos que, a su vez, da luz verde para que los bancos puedan mejorar sus servicios, tal y como veremos a continuación.
- ▶ Atención al cliente. Los bancos han mejorado la atención que prestan a sus clientes gracias a la gestión más eficiente de los datos. La implementación de canales de comunicación alternativos a los tradicionales, como los chats, o el uso de asistentes virtuales, como los *chatbots*, buscan mejorar la relación con los clientes. Se facilitan, así, espacios virtuales en los que se mejora la accesibilidad a los gestores.
- ▶ Sistemas de recomendación. La disponibilidad de mucha más información sobre sus clientes ha permitido a los bancos aplicar técnicas de sistemas de recomendación.  
**El disponer de más parámetros que alimentan el perfil de los clientes permite identificar mejor qué servicio es el más conveniente en cada momento.** Este tipo de sistemas no solo ayuda a la mejora de los servicios, sino que tiene un claro enfoque comercial con el que los bancos pueden aumentar el número de productos que cada cliente contrata con ellos.
- ▶ Segmentación de clientes. Siguiendo con el uso de la información de los clientes, la segmentación de estos permite a los bancos generar perfiles en función de características comunes. Estos perfiles permiten definir de forma más precisa y personalizada los productos que ofrecen. Así, los bancos tratan de maximizar el éxito y el margen de los servicios que comercializan con sus clientes.

- ▶ Protección y mantenimiento de clientes. De los puntos anteriores se desprende que el modelo bancario está girando a un modelo centrado en el usuario (el cliente). Toda la información que se recoge sobre ellos será utilizada también para fidelizar y hacerlos sentir seguros y, en cierto modo, únicos. Este tipo de experiencias, enriquecidas gracias a la aplicación de sistemas *big data*, **permite que los clientes permanezcan más tiempo y contraten más productos.**

**Modelos de riesgo.** Toda la información de la que disponen los datos trabaja para uno de los objetivos principales: la minimización del riesgo en sus operaciones. Los vastos volúmenes de información que ahora manejan, ayudados por la aplicación de herramientas que permiten un procesado y análisis rápido, dan lugar a modelos de evaluación de riesgo cada vez más precisos. Además, este riesgo ya no solo se calcula sobre productos u operaciones a ofrecer, sino también en aquellas operaciones o transacciones que se están realizando en un momento dado, siendo monitorizadas en tiempo real y pudiendo detenerlas en caso de que exista riesgo de fraude.

## 5.7. Conducción Autónoma

La conducción automática o autónoma representa uno de los campos más innovadores donde la ciencia de datos tiene un impacto profundo y transformador. La integración de algoritmos avanzados machine learning y grandes volúmenes de datos de sensores está permitiendo desarrollar vehículos que pueden operar sin intervención humana, mejorando la seguridad y eficiencia del transporte.

Las técnicas de visión por computadora, procesamiento de señales, y redes neuronales profundas son fundamentales en el desarrollo de sistemas autónomos de conducción. Estos sistemas aprenden de vastas cantidades de datos recopilados durante miles de horas de conducción para tomar decisiones en tiempo real. A continuación, se presentan ejemplos específicos de cómo la ciencia de datos está siendo aplicada en la conducción automática, destacando los beneficios tangibles.

### Percepción y Procesamiento Sensorial

- ▶ Ejemplo: Tesla utiliza redes neuronales avanzadas para interpretar datos de cámaras, radares y ultrasonidos, permitiendo a sus vehículos "ver" y "entender" el entorno circundante. Esta tecnología es crucial para la detección de objetos, vehículos, peatones y señales de tráfico.
- ▶ Beneficio: Aumenta la seguridad al proporcionar una percepción precisa del entorno vehicular, reduciendo la probabilidad de accidentes causados por errores humanos.

### Predicción de Comportamiento de Otros Conductores y Peatones

- ▶ Ejemplo: Waymo, la empresa de vehículos autónomos de Alphabet, implementa modelos predictivos que anticipan las acciones de otros conductores y peatones.

Estos modelos utilizan históricos de datos para prever movimientos y ajustar la navegación del vehículo.

- ▶ Beneficio: Mejora la capacidad de respuesta del vehículo ante situaciones imprevistas en la carretera, como un peatón cruzando inesperadamente.

## Optimización de Rutas y Gestión del Tráfico

- ▶ Ejemplo: Uber ATG (Advanced Technologies Group) emplea algoritmos de optimización para calcular rutas eficientes basadas en condiciones del tráfico en tiempo real, eventos en la ciudad, y patrones meteorológicos.
- ▶ Beneficio: Minimiza el tiempo de viaje y maximiza la eficiencia del combustible, reduciendo a su vez la huella de carbono de los vehículos.

## Entrenamiento y Validación de Sistemas Autónomos

- ▶ Ejemplo: Nvidia ha desarrollado plataformas de simulación que permiten a los vehículos autónomos entrenarse en miles de escenarios virtuales antes de ser desplegados en las carreteras. Estas simulaciones proporcionan una gran cantidad de datos que ayudan a mejorar la toma de decisiones del vehículo.
- ▶ Beneficio: Asegura que los sistemas de conducción autónoma estén bien preparados para enfrentarse a todo tipo de situaciones en el mundo real, mejorando su seguridad y confiabilidad.

La ciencia de datos es esencial en el desarrollo y la implementación de tecnologías de conducción autónoma, ofreciendo soluciones que aumentan la seguridad, la eficiencia y la sostenibilidad del transporte. A medida que esta tecnología continúa evolucionando, su capacidad para manejar situaciones cada vez más complejas sin intervención humana promete revolucionar nuestra forma de viajar.

- ▶ **Sistemas de entretenimiento:** plataformas como Amazon Prime, Netflix o Spotify basan sus recomendaciones de contenido en sistemas *big data* para así proporcionarnos contenido acorde a nuestros gustos.
- ▶ **Información meteorológica:** las predicciones meteorológicas han ganado en precisión y granularidad gracias a la aplicación de técnicas de gestión y análisis de datos masivos.

La aplicación de técnicas de ciencia de datos busca el descubrimiento de nuevo conocimiento extraído de múltiples fuentes de información para, de esta forma, apoyar la toma de decisiones inteligentes, oportunas e informadas. Los sistemas *big data* han supuesto un *big bang* en este sentido, ya que la información de la que disponemos ha crecido exponencialmente y, además, ahora ya no solamente se analiza en entornos aislados, sino que fluye de forma transversal entre múltiples ámbitos.

La recogida y manipulación de datos se realiza tanto a nivel público como privado. Por un lado, las administraciones disponen de una cantidad ingente de información con la que pueden mejorar los servicios que prestan a los ciudadanos, además de aumentar el tipo de información que recogen gracias a su digitalización. Por otro lado, las empresas privadas también recogen vastas cantidades de información que utilizan para su propio beneficio o que ofrecen, bajo pago o gratuitamente, a otras para que empresas terceraas mejoren las prestaciones de sus soluciones. Un claro ejemplo son los gigantes tecnológicos conocidos como GAFA (Google, Amazon, Facebook y Apple), empresas que recogen, manejan y ofrecen información de todo tipo. Esta información fluye entre unos sistemas y otros mediante contratos entre entidades o bien de forma abierta, como es el caso de [OpenStreetMap](#) o [MarineExplore](#), favoreciendo así no solo la mejora de productos o servicios públicos y privados, sino también la investigación y el desarrollo tecnológico, industrial o social.

Echando un vistazo rápido a nuestro alrededor podemos encontrar múltiples situaciones en las que hacemos uso a diario de sistemas que confían en los sistemas *big data* para proporcionarnos un servicio más personalizado, preciso y de mejor calidad. Entre otros, podemos encontrar (Yamin, 2019):

- ▶ **Compras:** las tarjetas de fidelización permiten identificar más fácilmente nuestros hábitos de compra o preferencias, de forma que podamos recibir ofertas o descuentos.
- ▶ **Transporte:** el uso de la tarjeta de transporte permite registrar cuándo y dónde viajamos, de forma que las empresas de gestión pueden regular mejor los servicios.
- ▶ **Deporte:** los relojes inteligentes, las pulseras de actividad o el propio teléfono móvil recogen información de nuestra actividad diaria o cuando practicamos algún tipo de deporte, de forma que podemos estar monitorizados, seguir un plan de entrenamiento o recibir incentivos para llevar un estilo de vida más saludable.
- ▶ **Publicidad:** Las redes sociales (Facebook, Instagram, Twitter, TikTok), buscadores (Google, Bing) o *marketplaces* (Amazon, Alibaba) hacen uso de nuestra huella digital para personalizar la publicidad que recibimos al hacer uso de los servicios que nos ofrecen de forma gratuita.
- ▶ **Salud:** la información y la investigación en el ámbito de la salud se han visto beneficiadas gracias a los sistemas *big data*, tal y como pudo comprobarse con la información, casi en tiempo real, que recibimos sobre la pandemia provocada por la COVID-19.
- ▶ **Educación:** la implantación de la educación *online* o la mejora de los métodos tradicionales mediante el uso de herramientas innovadoras permiten aumentar el rendimiento de los estudiantes y proporcionar realimentación a los profesores que anteriormente era difícil de captar y digerir.

- ▶ **Seguros:** las aseguradoras utilizan estas tecnologías para ajustar las primas e incluso incentivar a sus usuarios a reducir el riesgo, como sucede en aquellas que recopilan la información proveniente de la actividad física de sus clientes gracias a la conexión con las aplicaciones específicas (y la autorización de estos).
- ▶ **Agricultura y ganadería:** el uso de sistemas *big data* en la agricultura y la ganadería ha permitido el desarrollo de sistemas predictivos y preventivos de gran ayuda para la toma de decisiones, facilitando, por ejemplo, la detección de infecciones en cultivos o la reducción de la consanguinidad en la cría de ganado en pureza.
- ▶ **Turismo:** el sector turístico es otro de los mayores usuarios de sistemas *big data*, tanto en la optimización de los sistemas de gestión de reservas como a nivel más práctico, incluyendo la implementación de guías en dispositivos móviles o realidad aumentada en museos, entre otros muchos.
- ▶ **Telecomunicaciones:** el sector de las telecomunicaciones es también un gran consumidor de *big data* debido a la cada vez mayor cantidad de información que deben gestionar sus redes.

## 5.8. Otros

Los Gobiernos y la Administración pública se han convertido en los últimos años en grandes usuarios de los sistemas *big data*. La diversidad de funciones que cubren las Administraciones públicas hace que el uso que se hace de los grandes conjuntos de datos sea igualmente diverso.

La primera utilidad que podemos encontrar es más política que enfocada al ciudadano: **el análisis de la intención de voto**. Existen un gran número de empresas dedicadas a estimar cómo votaríamos los ciudadanos en un momento determinado, de forma que la información que se recoge permita identificar y analizar patrones que nos dejen descubrir el resultado de unas elecciones.

Este tipo de análisis no solo pretende ser meramente exploratorio, sino que, en ocasiones, tratan de influir en el comportamiento de la audiencia: la información es poder y como tal se usa, por lo que debemos ser conscientes de las implicaciones éticas que tiene el uso de la información.

Por este y otros motivos, la aplicación de técnicas de ciencia de datos en los Gobiernos está sujeta al debate y la crítica, dada la importancia de los datos personales, lo cual puede entenderse como una limitación o freno en los Gobiernos en los que las libertades individuales y la privacidad del individuo están ampliamente regulados. Sin embargo, este no ha sido un motivo para desincentivar el desarrollo de políticas que fomenten el uso de datos para ayudar a resolver problemas que mejoren la vida de los ciudadanos.

Un primer ejemplo que podemos encontrar del uso del *big data* en organismos gubernamentales es la colaboración entre diferentes administraciones, entre diferentes Gobiernos de diferentes países e incluso entre el sector público y el privado. **La gestión de emergencias**, como en caso de terremotos u otras

catástrofes son un claro ejemplo de coordinación entre Administraciones y empresas: disponibilidad de recursos, ubicaciones, gestión de riesgos, etcétera. La calidad de los datos es en estas situaciones es primordial para priorizar la información y minimizar el uso incorrecto de unos recursos que, en estas ocasiones, siempre suelen ser limitados.

Por otro lado, y muy relacionada con el ejemplo anterior, **la seguridad es otro de los ámbitos en los que los Gobiernos se benefician de la aplicación de sistemas *big data*.** En múltiples países, como, por ejemplo, Estados Unidos, se hace uso de modelos de riesgos que predicen la delincuencia identificando edificios o zonas potencialmente peligrosas. Esto ayuda a que los cuerpos y fuerzas de seguridad del estado dispongan de herramientas útiles que les permitan una reacción más rápida ante situaciones que comprometen la seguridad ciudadana, utilizando planos, información en tiempo real o índices de riesgo, entre otros muchos recursos. De este modo, se puede reducir el número de robos, el uso indebido de armas de fuego y, en definitiva, la delincuencia en general.

Otra de las áreas en las que el *big data* ayuda a los Gobiernos es a través de la **digitalización de la administración.** El desarrollo de certificados digitales, como pueden ser el [DNI electrónico \(DNI-e\)](#) y el [certificado de la Seguridad Social](#) en España, permite a los ciudadanos disponer de una identidad digital oficial que les facilita realizar trámites de la administración y de índole privada. Gracias a esta identidad se ha simplificado la realización de trámites burocráticos o firma de documentos, pudiendo tramitarse *online* desde cualquier ubicación, reduciendo la presencialidad, tiempos de espera y otras incomodidades asociadas a estos servicios.

Podemos encontrar otro ejemplo de la colaboración entre administraciones en las tramitaciones expuestas anteriormente. En alguna ocasión hemos tenido que lidiar con la provisión de la misma información para diferentes entidades lo cual, a simple vista, es totalmente ineficiente. Los Gobiernos se están dando cuenta de que una

mejor experiencia de usuario (UX, *user experience*) facilita su relación con los ciudadanos, de forma que es clave el desarrollo y puesta en marcha de redes de intercambio de información seguras. La coordinación de información entre estamentos públicos permitirá que, con una sola actualización, todos aquellos con los que estemos relacionados puedan disponer de ella al instante.

Mejorar la relación de los usuarios con la Administración también implica mejorar la participación de los primeros en la segunda. Incentivar a participar a los ciudadanos en la toma de decisiones hace que estos adquieran un mayor compromiso y mejora la transparencia de los Gobiernos. Este nivel de participación se lleva a cabo a todos los niveles, no solo en aquellas Administraciones con más recursos. Sirvan como ejemplo la consulta pública sobre el plan estratégico de subvenciones 2021-2023 del [Ayuntamiento de San Pedro del Pinatar \(Murcia\)](#) o el proceso de participación ciudadana del Ayuntamiento de Godella para decidir las inversiones que se ejecutan (El periodic, 2018). Existen múltiples formas de recoger la información proporcionada por los ciudadanos, ya sea en forma de encuestas, propuestas, etcétera.

Finalmente, cabe destacar que **la transición a las ciudades inteligentes (*smart cities*) como otra de las áreas en las que los Gobiernos están haciendo hincapié**. El uso intensivo de sistemas de captura de información (como, por ejemplo, la implantación de sistemas de internet de las cosas y su gestión a través de sistemas *big data*) está permitiendo que los servicios que se proporcionan a los usuarios sean cada vez más precisos y eficientes.

La ciudad inteligente permite gestionar mejor la energía, el agua, la recogida de basuras, los flujos de movilidad, la comunicación o la vivienda, teniendo como objetivo último la mejora de la calidad de vida de los ciudadanos. La figura 4 muestra las doce claves que convierten a una ciudad en una ciudad inteligente identificadas por Iberdrola (s. f.).

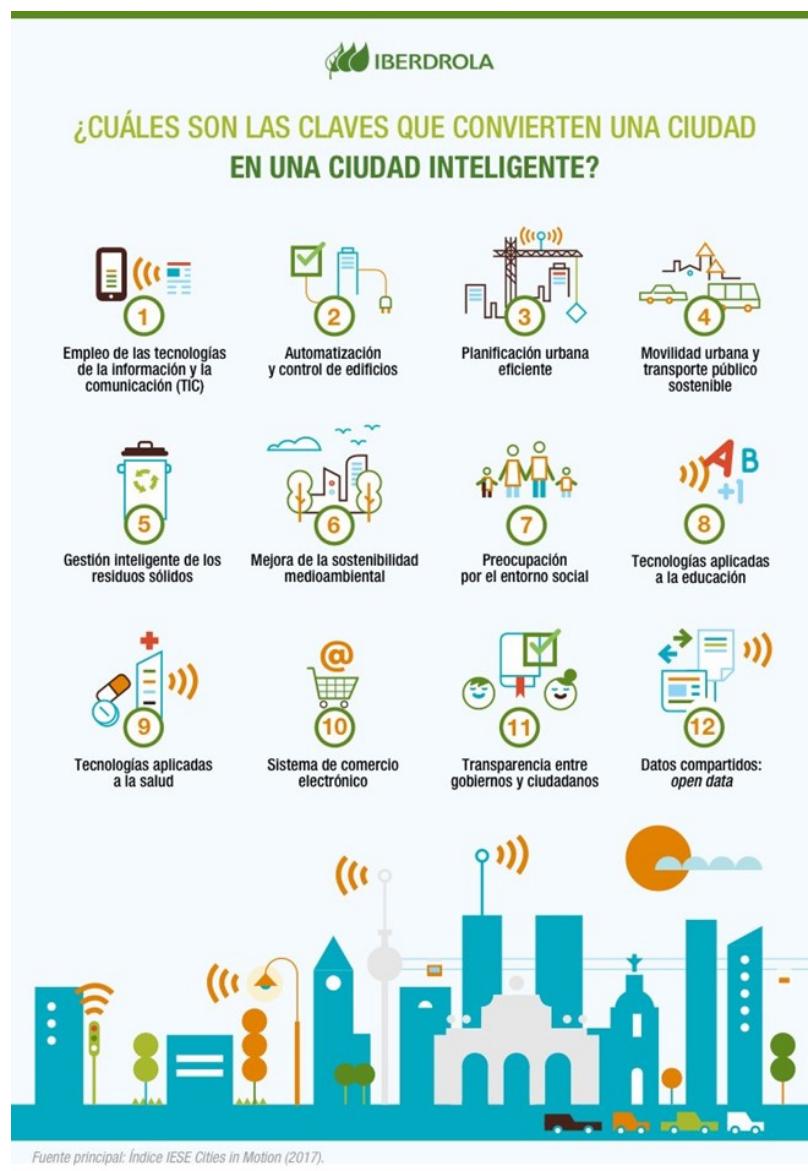


Figura 4. Claves que convierten una ciudad en una ciudad inteligente identificadas por Iberdrola. Fuente: Iberdrola, s. f.

De forma general, se observa que las claves que nos muestra la figura 4 están orientadas al beneficio social y al comportamiento responsable con el medio ambiente, espacios en los que el uso de los datos tiene gran relevancia y, obviamente, la tendrá en el futuro.

El uso de grandes conjuntos de datos en múltiples ámbitos conlleva implicaciones sociales que no ha de pasarse por alto, tal y como se desprende de las secciones anteriores. El uso de la tecnología en nuestro día a día parece que aporta grandes beneficios, pero deja una importante duda en el aire: ¿es su aplicación neutral? La digitalización y uso de la tecnología tiene importantes implicaciones que a menudo son imprevisibles o están mal entendidas dentro del conjunto de la sociedad. ¿Qué seguridad tienen nuestros datos? ¿Hasta qué punto son privadas nuestras conversaciones? ¿Manejarán los algoritmos nuestras vidas?

Dejando a un lado estas dudas, que corresponden más a aspectos éticos, las implicaciones sociales de los sistemas *big data* son claramente beneficiosas y desafiantes para todas las partes interesadas. Empresas y Gobiernos están inmersos en la implementación de sistemas de gestión de grandes volúmenes de datos que mejoren la calidad, planificación y dimensionamiento de sus servicios, al igual que existe una política de fomento del uso de datos abiertos que desbloqueen en cierta medida su uso para así incentivar la innovación continua de las empresas y, por ende, de los servicios de los que nos beneficiamos como parte de la sociedad.

Sin embargo, las innovaciones relacionadas con los datos deben aplicarse y desplegarse de forma responsable. Los derechos de los consumidores deben ser satisfechos mediante la aplicación de principios de privacidad y seguridad por diseño. Esto dará lugar a la creación de modelos sólidos de gobernanza de datos que pongan en cuestión cómo y por qué se recopila, almacena, procesa y hace uso de determinada información, teniendo en cuenta los efectos potenciales que estas acciones tengan en la sociedad en general.

Las implicaciones sociales del uso de la tecnología nos acompañan desde que los humanos nos organizamos como sociedad, pero es ahora cuando se hace más evidente, debido a que esta ha entrado de lleno en todos los ámbitos de nuestra vida. El uso de dispositivos que recogen información de nuestro día a día presenta

grandes retos que permitirán mejorarlo. El contrapunto a esta ventaja se encuentra en cómo se aplican las técnicas y los algoritmos que gestionarán nuestra información, un aspecto que aún parece opaco a la vista tanto de los usuarios comunes como de expertos en la materia.

Existen ciertos retos que deben abordarse para mejorar la relación de la sociedad con esta nueva forma de gestionar la información, entre otros:

- ▶ Que las empresas tengan una profunda conciencia de las implicaciones de su trabajo más allá de la aplicación final, habitualmente sesgada por la consecución de objetivos empresariales.
- ▶ Que los responsables políticos y legisladores comprendan las implicaciones de los desarrollos tecnológicos y cómo estos interactúan con las leyes y políticas existentes.

De este modo nos aproximamos a las implicaciones sociales de la tecnología desde dos puntos de vista en ocasiones enfrentados, pero, como sociedad, los intereses se cruzan cuando los modelos de recogida de información, su automatización y el uso de algoritmos puedan impactar en el valor social de los mismos, ya sea reduciéndolo o incrementándolo.

Desde hace décadas las empresas y Gobiernos han considerado las implicaciones sociales que la tecnología, más concretamente en nuestro caso la gestión de los datos, en diferentes aspectos como son el desarrollo sostenible, la ética y los valores humanos, así como el compromiso con la comunidad científica.

Las implicaciones sociales de la tecnología son un tema extenso y omnipresente sobre el que se está haciendo un gran esfuerzo un esfuerzo por fomentar un pensamiento crítico y un nivel pragmático que marcarán la diferencia en el futuro.

## Implicaciones éticas

En los sistemas *big data* y la aplicación de ciencia de datos, todos los datos e información, independientemente del contexto en el que se manejen, deben ser sometidos a una serie de procesos desde antes de su obtención hasta la explotación de estos.

El reto que se presenta aquí es comprender las implicaciones e impactos éticos y legales del uso de estos recursos, especialmente en contextos potencialmente sensibles. Por ejemplo, la gestión de datos relacionados con la salud de los individuos es especialmente delicada, pero también podemos plantearnos otra serie de situaciones en las que se manejen datos sin consentimiento de los usuarios para, por ejemplo, identificar situaciones de terrorismo, discriminación, violencia de género, etcétera.

Para garantizar que el uso de estas herramientas no se salga del control de la sociedad que las aplica debe tenerse en cuenta quién y cómo gestionará nuestros datos. Además, es clave considerar qué tipo de algoritmos se aplicarán sobre ellos y las implicaciones que tendrán sus resultados. Por último, y sin que este aspecto sea menos importante, es determinante fijar un marco legal que ha de imponer la sociedad para que los resultados generados sean positivos y justos.

Dos de los ámbitos que generan más controversia a nivel ético son la automatización y la inteligencia artificial. Existen múltiples ejemplos en los que el impacto de su utilización podría ser negativo a nivel social. Por ejemplo, las empresas podrían utilizar sistemas de recomendación que incluyesen sesgos culturales de forma que se beneficiase más a una parte de la sociedad que a otra, generando estereotipos y discriminaciones.

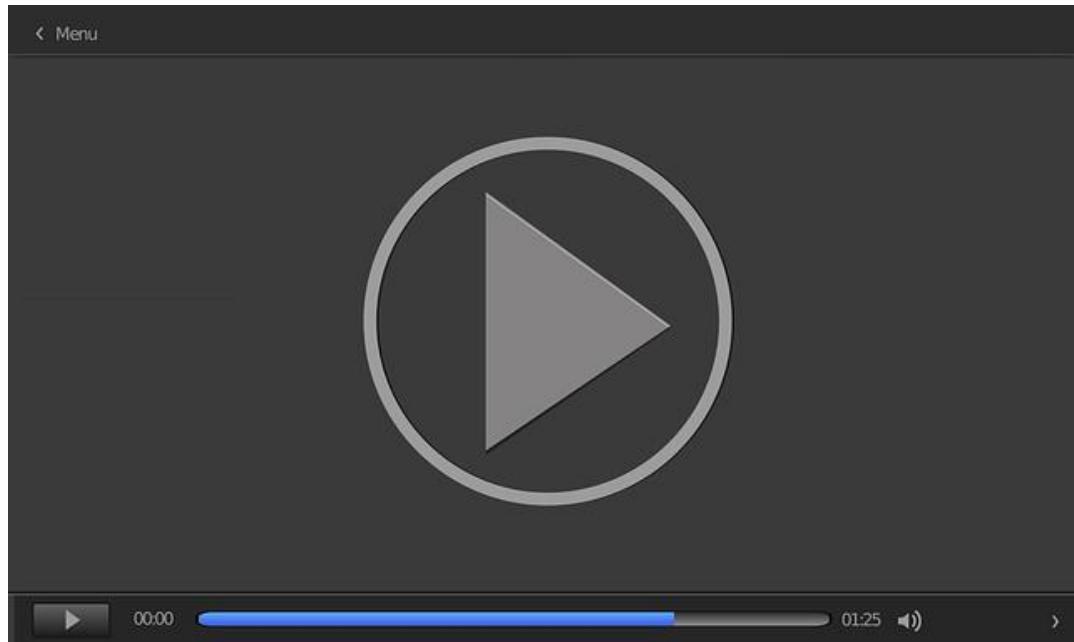
La evaluación, la revisión, la supervisión, la rendición de cuentas y los marcos legales parecen apropiados si, por ejemplo, el uso de los sistemas *big data* sirve para mejorar nuestra seguridad, elaborando, por ejemplo, perfiles de terroristas, pero esto

podría tener impactos nefastos en algunas comunidades si nuestro algoritmo falla. El problema no es el concepto de análisis de datos, sino cómo se desarrollan, utilizan, entienden y evalúan.

En todo este proceso, independientemente del ámbito de aplicación, debemos asegurarnos de que las soluciones implantadas se evalúan rigurosamente en función de métricas que comprueben no solo la precisión y la eficacia, sino también la disparidad de impacto y las cuestiones morales. El factor humano en este caso es determinante pues no solamente ayudará a implementar la tecnología, sino que debe ser capaz de apoyarse en ella para que su uso sea justo y se minimice su impacto negativo.

En los últimos años se han empezado a desarrollar políticas públicas que dan importancia a la ética dentro de la tecnología. En este sentido, debe tenerse en cuenta tanto la ética profesional como la ética en el diseño y aplicación de las nuevas tecnologías. No solo es importante explotar los datos, debemos ser capaces de hacerlo de forma objetiva y éticamente responsable.

Veamos otra introducción a la *Ética y ciencia de datos: inteligencia artificial explicable e inteligencia artificial ética*, donde descubriremos la relación entre todos estos conceptos.



---

Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=3b1bc959-17f9-4a6f-a6bf-b16f00e304f7>

---

## 5.9. Referencias bibliográficas

Departamento de Justicia del Estado de California. (2021). *Ley de Privacidad del Consumidor de California (CCPA)*. <https://oag.ca.gov/privacy/ccpa>

El periodic. (6 de marzo de 2018). El Ayuntamiento de Godella comienza un proceso de participación ciudadana para decidir las inversiones. *Elperiodic.com*. [https://www.elperiodic.com/godella/ayuntamiento-godella-comienza-proceso-participacion-ciudadana-para-decidir-inversiones\\_553308](https://www.elperiodic.com/godella/ayuntamiento-godella-comienza-proceso-participacion-ciudadana-para-decidir-inversiones_553308)

España. Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el se deroga la Directiva 95/46/CE (Reglamento general de protección de datos). *Diario Oficial de la Unión Europea*, 4 de mayo de 2016, L 119/1-88.

Incibe-cert. (9 de agosto de 2016). *Thinking in Big (Data) and industrial security*. <https://www.incibe-cert.es/en/blog/thinking-big-data-and-industrial-security>

Maestre, R. J. (2020). *Qué es el Fintech, definición, sectores y ejemplos de startups*. IEBS. <https://www.iebschool.com/blog/que-es-fintech-finanzas/>

Steward, C. (24 de septiembre de 2020). *Total amount of global healthcare data generated in 2013 and a projection for 2020*. Statista. <https://www.statista.com/statistics/1037970/global-healthcare-data-volume/>

TechVidvan. (3 de marzo de 2020). *Data Science in Banking—8 Remarkable Applications with Case Study*. <https://techvidvan.com/tutorials/data-science-in-banking/>

Thonnet, M. y Nicolas, L. (2017). *Information paper on supporting preparatory convergence meetings between the eHN and WHO*. JASEHN.

[https://ec.europa.eu/health/sites/default/files/ehealth/docs/ev\\_20170509\\_co15\\_en.pdf](https://ec.europa.eu/health/sites/default/files/ehealth/docs/ev_20170509_co15_en.pdf)

Yamin, M. (2019). Information technologies of 21st century and their impact on the society. *International Journal of Information Technology*, 11(4), 759-766.

## La aplicación de inteligencia artificial a la analítica de datos

Requejo, P. (6 de abril de 2016). Big data for Social Good o la aplicación de Inteligencia Artificial a la analítica de datos. *BlogThinkBig*.  
<https://blogthinkbig.com/big-data-for-social-good-o-la-aplicacion-de-inteligencia-artificial-a-la-analitica-de-datos>

Pablo Requejo presenta en este artículo su conversación con Nuria Oliver, entonces directora científica en Telefónica I+D, sobre los cambios que está experimentando la sociedad en relación con los sistemas *big data* y la inteligencia artificial.

## La ética del Data

Calle, C. (2018). *La ética del Data.* KPMG Tendencias.

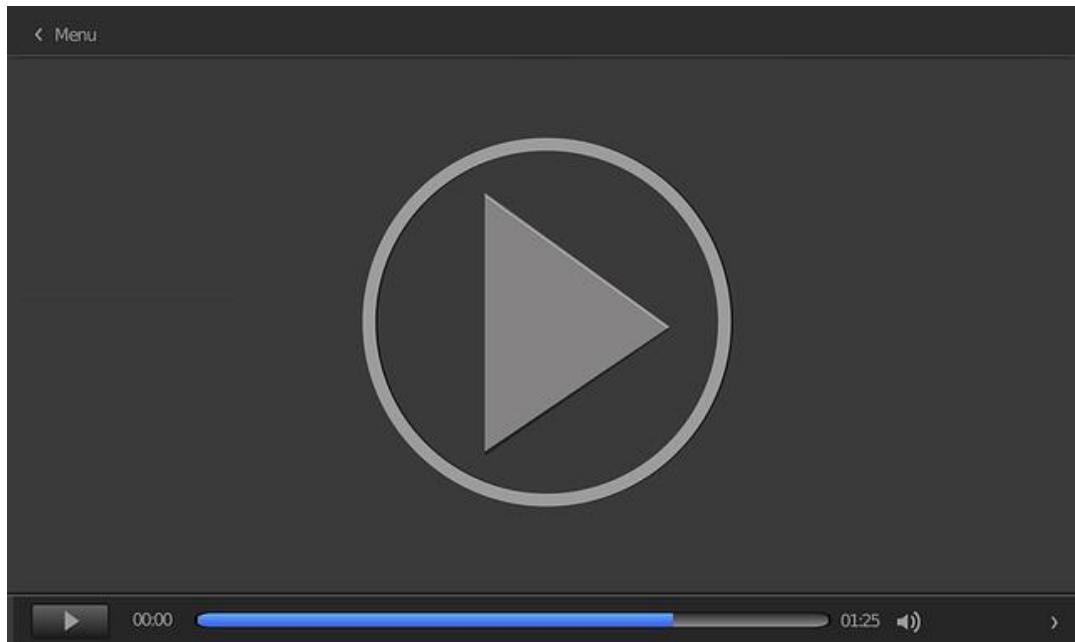
<https://www.tendencias.kpmg.es/2018/04/etica-big-data/>

Lee este artículo de Consuelo Calle (KPMG) en el que expone la importancia de instruir y utilizar de forma correcta la aplicación de la tecnología, la cual es, por naturaleza, imparcial.

## ¡El futuro pasa por el big data!

BlogThinkBig. (6 de abril de 2016). *¡El futuro pasa por el Big data! Hablamos con Nuria Oliver* [Archivo de vídeo]. YouTube. [https://www.youtube.com/watch?v=EqvKssq\\_yZ0&t=1s](https://www.youtube.com/watch?v=EqvKssq_yZ0&t=1s)

Además, no dejes de ver el vídeo de la interesante conversación en la que se basó el anterior artículo.



Accede al vídeo:

[https://www.youtube.com/embed/EqvKssq\\_yZ0](https://www.youtube.com/embed/EqvKssq_yZ0)

- 1.** ¿Qué técnica utilizan empresas como Amazon para ajustar los precios en tiempo real?

  - A. Regresión lineal.
  - B. Clustering.
  - C. Algoritmos predictivos.
  - D. Análisis factorial.
  
- 2.** ¿Cuál es una aplicación común de la ciencia de datos en la industria manufacturera?

  - A. Publicidad en redes sociales.
  - B. Optimización de la cadena de suministro.
  - C. Análisis de sentimientos.
  - D. Gestión de recursos humanos.
  
- 3.** ¿Qué tecnología es fundamental en los vehículos autónomos para 'ver' el entorno?

  - A. Análisis de texto.
  - B. Visión por computadora.
  - C. Bases de datos NoSQL.
  - D. Blockchain.
  
- 4.** ¿Qué método utilizan las instituciones financieras para detectar fraudes?

  - A. Análisis predictivo.
  - B. Minería de datos.
  - C. Modelos de regresión logística.
  - D. Todas las anteriores.

5. ¿Cómo ayuda la ciencia de datos en la atención médica?

- A. Optimización de precios.
- B. Predicción de enfermedades.
- C. Mejora de algoritmos de redes sociales.
- D. Desarrollo de juegos.

6. ¿Cuál es un beneficio clave del uso de la ciencia de datos en la seguridad cibernetica?

- A. Aumento de ventas.
- B. Detección de amenazas en tiempo real.
- C. Mejora en la gestión de inventarios.
- D. Optimización de rutas de transporte.

7. ¿Qué rol juega la ciencia de datos en la personalización de experiencias de usuario en plataformas como Netflix?

- A. Aprendizaje automático.
- B. Análisis de redes.
- C. Simulaciones Monte Carlo.
- D. Regresión lineal.

8. ¿En qué sector es crítico el uso de simulaciones para entrenar sistemas antes de su implementación real?

- A. Educación.
- B. Conducción autónoma.
- C. E-commerce.
- D. Hostelería.

**9.** ¿Qué aplicación de la ciencia de datos ayuda a Starbucks a decidir dónde abrir nuevas tiendas?

- A. Optimización de menús.
- B. Análisis de ubicación.
- C. Gestión de recursos humanos.
- D. Campañas de marketing digital.

**10.** ¿Qué técnica es usada en el sector industrial para predecir el mantenimiento de maquinaria?

- A. Mantenimiento predictivo.
- B. Gestión de la calidad del producto.
- C. Optimización de la experiencia del cliente.
- D. Análisis de sentimientos.

Ciencia de Datos Aplicada

---

# Tema 6. Estrategias en Almacenamiento Masivo

# Índice

[Esquema](#)

[Ideas clave](#)

[6.1. Introducción y objetivos](#)

[6.2. Data Mart](#)

[6.3. Data Warehouse](#)

[6.4. Data Lake](#)

[6.5. Referencias bibliográficas](#)

[A fondo](#)

[La importancia de los Data Marts en la inteligencia empresarial](#)

[Data Warehouse: la base de la inteligencia empresarial moderna](#)

[El papel de los Data Lakes en la era del Big Data](#)

[Test](#)

ESTRATEGIAS EN ALMACENAMIENTO MASIVO	
DATA MARTS	DATA WAREHOUSE
<b>Enfoque Específico</b> <ul style="list-style-type: none"> <li><u>Características:</u> Orientado a un área específica de la empresa (como ventas, marketing o finanzas).</li> <li><u>Ventajas:</u> Facilita el análisis detallado y relevante para departamentos específicos.</li> <li><u>Desventajas:</u> Ofrece una visión limitada ya que solo cubre una parte de los datos de la organización.</li> </ul>	<b>Centralización de Datos</b> <ul style="list-style-type: none"> <li><u>Características:</u> Repositorio centralizado que integra datos de múltiples fuentes.</li> <li><u>Ventajas:</u> Proporciona una visión global y consolidada de los datos de la organización.</li> <li><u>Desventajas:</u> Complejidad y coste de implementación y mantenimiento elevados.</li> </ul>
<b>Rápida Implementación</b> <ul style="list-style-type: none"> <li><u>Características:</u> Menor tamaño Y complejidad en comparación con un Data Warehouse.</li> <li><u>Ventajas:</u> Implementación más rápida y económica.</li> <li><u>Desventajas:</u> Puede resultar en duplicación de datos y problemas de integridad si no se gestiona adecuadamente..</li> </ul>	<b>Orientación al Análisis</b> <ul style="list-style-type: none"> <li><u>Características:</u> Diseñado para análisis y generación de informes complejos.</li> <li><u>Ventajas:</u> Facilita la toma de decisiones estratégicas basadas en datos históricos y actuales.</li> <li><u>Desventajas:</u> Menos adecuado para análisis en tiempo real debido a su estructura rígida.</li> </ul>
<b>Optimización de consultas</b> <ul style="list-style-type: none"> <li><u>Características:</u> Datos organizados y optimizados para consultas rápidas.</li> <li><u>Ventajas:</u> Mejor rendimiento en análisis específicos y tiempos de respuesta más rápidos.</li> <li><u>Desventajas:</u> Menos flexible para cambios en los requisitos de datos y análisis.</li> </ul>	<b>Datos Históricos</b> <ul style="list-style-type: none"> <li><u>Características:</u> Almacena grandes volúmenes de datos históricos para análisis a largo plazo.</li> <li><u>Ventajas:</u> Permite el análisis de tendencias y patrones a lo largo del tiempo.</li> <li><u>Desventajas:</u> Requiere procesos ETL complejos y puede tener problemas de rendimiento con datos muy grandes.</li> </ul>
	<b>Acceso Múltiple</b> <ul style="list-style-type: none"> <li><u>Características:</u> Permite el acceso simultáneo a los datos por parte de diferentes herramientas y lenguajes de análisis.</li> <li><u>Ventajas:</u> Facilita el trabajo de científicos de datos, analistas y desarrolladores.</li> <li><u>Desventajas:</u> Menos eficiente para consultas estructuradas y análisis de datos altamente estructurados.</li> </ul>

## 6.1. Introducción y objetivos

En la actualidad, el manejo eficiente de grandes volúmenes de datos es crucial para el éxito de cualquier organización. Las empresas generan y recopilan datos de múltiples fuentes, lo que crea la necesidad de estrategias robustas de almacenamiento masivo. Este capítulo ofrece una visión general de tres enfoques principales para el almacenamiento de datos: Data Mart, Data Warehouse y Data Lake, así como una discusión sobre las nuevas tendencias en este campo.

Al finalizar la lectura de este tema, los lectores deberán ser capaces de:

- ▶ Comprender las diferencias clave entre Data Mart, Data Warehouse y Data Lake y cómo cada uno se adapta a diferentes necesidades organizacionales.
- ▶ Reconocer las ventajas y limitaciones de cada estrategia de almacenamiento masivo, proporcionando una base sólida para elegir la mejor solución para diversos contextos empresariales.
- ▶ Identificar las nuevas tendencias en almacenamiento masivo y su impacto potencial en la gestión de datos a nivel organizacional.

## 6.2. Data Mart

Un Data Mart es una estructura de almacenamiento de datos optimizada para un departamento o una función específica dentro de una organización, como ventas, marketing o finanzas. Se trata de una versión más pequeña y especializada de un Data Warehouse, diseñada para satisfacer las necesidades de un grupo particular de usuarios con datos específicos y relevantes (Inmon, W. H., Strauss, D, 2008).

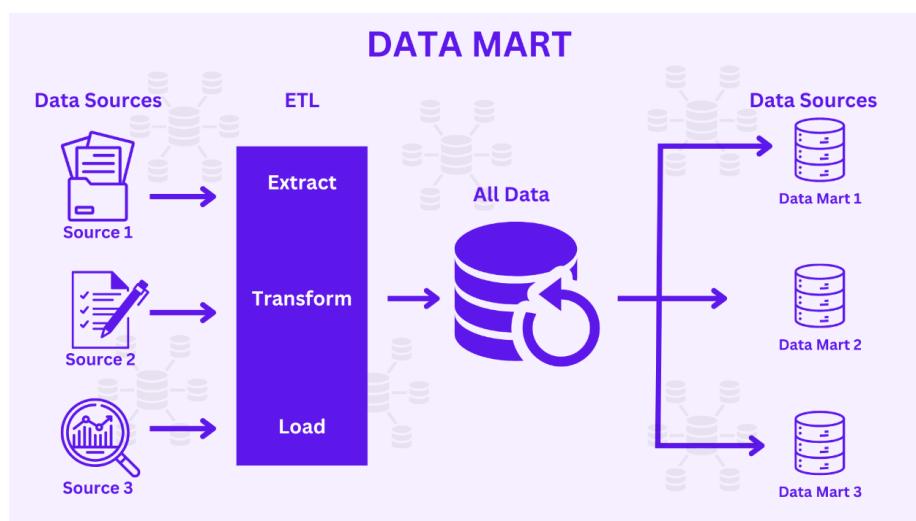


Figura 1. Enfoque general de construcción de un Data Marts. Fuente: Naveen Chandra

<https://blog.naveenchandra.co.in/all-articles/data-mart/>

### Principales características

- ▶ **Enfoque específico:** Un Data Mart está orientado a un área concreta de la empresa, proporcionando datos que son relevantes y útiles para ese departamento en particular.
- ▶ **Menor tamaño:** Comparado con un Data Warehouse, un Data Mart maneja un volumen de datos más pequeño, lo que facilita el acceso y la gestión de la información.

- ▶ **Rápida implementación:** Debido a su tamaño reducido y a su enfoque específico, los Data Marts pueden ser implementados más rápidamente que los Data Warehouses.
- ▶ **Facilidad de uso:** Los datos en un Data Mart están organizados y optimizados para su consulta y análisis, lo que permite a los usuarios realizar análisis detallados sin la necesidad de conocimientos técnicos avanzados.
- ▶ **Estructura simplificada:** Los Data Marts suelen tener una estructura más simple y menos compleja en comparación con los Data Warehouses, lo que facilita su gestión y mantenimiento.

## Desventajas

- ▶ Visión limitada: Al estar enfocados en un área específica, los Data Marts no proporcionan una visión global y consolidada de los datos de toda la organización.
- ▶ Duplicación de datos: En organizaciones con múltiples Data Marts, puede ocurrir duplicación de datos, lo que puede llevar a inconsistencias y problemas de integridad.
- ▶ Mantenimiento: Gestionar múltiples Data Marts puede ser más complicado y costoso en términos de mantenimiento y actualización, especialmente si no están integrados adecuadamente.
- ▶ Escalabilidad: Los Data Marts pueden no ser adecuados para empresas que necesitan escalar rápidamente el almacenamiento y el procesamiento de datos a nivel global.

## Diferenciación respecto a otros sistemas de almacenamiento

- ▶ **Data Warehouse:** Mientras que un Data Warehouse es una colección centralizada de datos integrados de múltiples áreas de la organización, un Data Mart se centra en una única área o función específica. Los Data Warehouses ofrecen una visión completa y global de los datos de la empresa, mientras que los Data Marts están diseñados para necesidades específicas y localizadas (Kimball, R, 2013).
- ▶ **Data Lake:** A diferencia de los Data Marts, los Data Lakes almacenan datos en su formato original y sin procesar. Los Data Lakes son más flexibles y pueden manejar datos estructurados, semi-estructurados y no estructurados, mientras que los Data Marts están optimizados para datos estructurados y análisis específicos.

## Casos de uso

- ▶ **Ventas:** Un Data Mart de ventas puede contener datos sobre transacciones, clientes, productos y tendencias de ventas. Esto permite a los gerentes de ventas analizar el rendimiento de ventas, identificar patrones de compra y tomar decisiones informadas para mejorar las estrategias de ventas.
- Mejora: Ayuda a identificar productos con mejor rendimiento, segmentar clientes según comportamientos de compra y optimizar campañas de ventas.
- ▶ **Marketing:** Un Data Mart de marketing puede incluir datos sobre campañas, canales de marketing, respuestas de clientes y métricas de rendimiento. Los equipos de marketing pueden utilizar esta información para analizar la efectividad de las campañas y ajustar las estrategias en tiempo real.
- Mejora: Permite realizar análisis de campañas más precisos, optimizar el presupuesto de marketing y mejorar la segmentación y personalización de las campañas.

- ▶ **Finanzas:** Un Data Mart financiero puede contener datos sobre ingresos, gastos, presupuestos y previsiones financieras. Los analistas financieros pueden utilizar estos datos para realizar análisis detallados, crear informes financieros y apoyar la toma de decisiones estratégicas.
- Mejora: Facilita el seguimiento y control de gastos, mejora la precisión en las previsiones financieras y apoya la toma de decisiones estratégicas basadas en datos precisos y actualizados.
- ▶ **Recursos Humanos:** Un Data Mart de recursos humanos puede almacenar datos sobre empleados, contrataciones, desempeño y retención. Los departamentos de RRHH pueden utilizar esta información para gestionar mejor su fuerza laboral y desarrollar estrategias de retención de talento.
- Mejora: Ayuda a identificar tendencias en el desempeño de los empleados, optimizar procesos de contratación y mejorar las estrategias de retención de talento.

Los Data Marts proporcionan una solución efectiva y eficiente para el análisis de datos específico de cada área de negocio, mejorando la toma de decisiones y optimizando los procesos operativos.

### 6.3. Data Warehouse

Un Data Warehouse (almacén de datos) es una base de datos centralizada que recopila, integra y almacena grandes volúmenes de datos procedentes de diversas fuentes heterogéneas. Está diseñado para el análisis y la generación de informes, proporcionando una visión coherente y consolidada de la información a lo largo del tiempo. Los datos en un Data Warehouse suelen ser históricos y están organizados para facilitar consultas complejas y análisis avanzados.

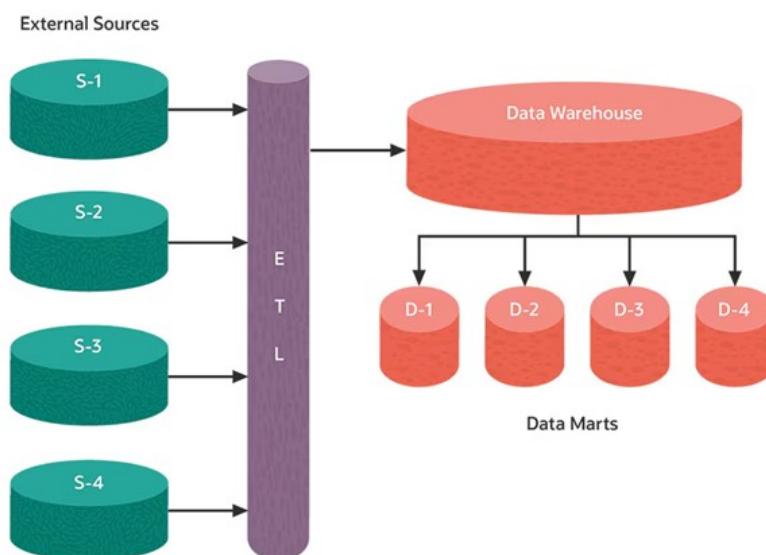


Figura 2. Enfoque general de Data Warehouse Vs. Data Marts. Fuente: Oracle

<https://www.netsuite.com/portal/resource/articles/data-warehouse/data-mart.shtml>

## Principales características

- ▶ Integración de datos: Un Data Warehouse unifica datos de múltiples fuentes, como bases de datos operacionales, sistemas transaccionales, archivos y otras fuentes externas, en un formato coherente y homogéneo.
- ▶ Orientado al análisis: A diferencia de los sistemas transaccionales que manejan operaciones diarias, los Data Warehouses están optimizados para consultas complejas, análisis y generación de informes.
- ▶ Datos históricos: Almacena datos históricos que permiten realizar análisis de tendencias y comparativas a lo largo del tiempo.
- ▶ Alta disponibilidad y rendimiento: Diseñado para soportar un gran volumen de consultas y análisis sin afectar el rendimiento de los sistemas operacionales.
- ▶ Modelado de datos: Utiliza esquemas de datos específicos como el esquema estrella o el esquema copo de nieve, que facilitan el acceso y la consulta de datos.

## Desventajas

- ▶ Coste de implementación: La creación y el mantenimiento de un Data Warehouse pueden ser costosos debido a la necesidad de hardware especializado, software y personal capacitado.
- ▶ Complejidad: La integración de datos de múltiples fuentes y el mantenimiento de la coherencia pueden ser complejos y requerir una gestión cuidadosa.
- ▶ Tiempo de carga: La extracción, transformación y carga (ETL) de datos en un Data Warehouse puede ser un proceso largo y complicado.
- ▶ Rigidz: Los Data Warehouses son menos flexibles para cambios rápidos en la estructura de datos o en los requisitos de análisis en comparación con sistemas más modernos como los Data Lakes.

## Diferenciación respecto a otros sistemas de almacenamiento

- ▶ Data Mart: Un Data Mart es una versión más pequeña y específica de un Data Warehouse, enfocada en una única área de negocio. Mientras que un Data Warehouse proporciona una visión global y consolidada de los datos de la organización, un Data Mart se centra en datos específicos para usuarios particulares.
- ▶ Data Lake: Un Data Lake almacena datos en su formato original y sin procesar, permitiendo una mayor flexibilidad para análisis avanzados y machine learning. En contraste, un Data Warehouse almacena datos estructurados y procesados, optimizados para consultas y análisis de negocios.

## Casos de uso

- ▶ Análisis de ventas y marketing: Un Data Warehouse puede integrar datos de transacciones, campañas de marketing, interacciones con clientes y datos externos de mercado. Esto permite a los equipos de ventas y marketing realizar análisis detallados de desempeño, segmentar clientes y optimizar estrategias de marketing.
- Mejora: Facilita el análisis de la efectividad de campañas, mejora la segmentación de clientes y permite el seguimiento de tendencias de ventas a lo largo del tiempo.
- ▶ Informes financieros: Un Data Warehouse puede consolidar datos financieros de diferentes departamentos y sistemas, proporcionando una visión única y coherente de la situación financiera de la empresa. Esto es esencial para la generación de informes financieros precisos y la toma de decisiones estratégicas.
- Mejora: Mejora la precisión de los informes financieros, facilita el cumplimiento normativo y apoya la planificación financiera a largo plazo.

- ▶ Gestión de la cadena de suministro: Un Data Warehouse puede integrar datos de proveedores, inventarios, producción y logística. Esto permite a las empresas optimizar la gestión de la cadena de suministro, reducir costos y mejorar la eficiencia operativa.
- Mejora: Ayuda a identificar ineficiencias, optimizar niveles de inventario y mejorar la planificación y ejecución de la cadena de suministro.
- ▶ Análisis de recursos humanos: Un Data Warehouse puede almacenar datos sobre contratación, desempeño, retención y desarrollo de empleados. Los departamentos de recursos humanos pueden utilizar esta información para mejorar la gestión del talento y desarrollar estrategias efectivas de recursos humanos.
- Mejora: Facilita el análisis de tendencias en la fuerza laboral, mejora la retención de empleados y optimiza los procesos de contratación y desarrollo profesional.
- ▶ Cumplimiento normativo y auditoría: Un Data Warehouse puede almacenar datos necesarios para cumplir con regulaciones y facilitar auditorías. Esto incluye datos históricos y transacciones detalladas que pueden ser necesarios para demostrar el cumplimiento.
- Mejora: Asegura el cumplimiento normativo, reduce el riesgo de sanciones y facilita el proceso de auditoría interna y externa.

Los Data Warehouses son esenciales para proporcionar a las organizaciones una visión integrada y consolidada de sus datos, permitiendo una toma de decisiones informada y estratégica basada en un análisis profundo y detallado de la información histórica y actual.

## 6.4. Data Lake

Un Data Lake es una arquitectura de almacenamiento de datos que permite guardar grandes volúmenes de datos en su formato original y sin procesar. A diferencia de los sistemas tradicionales de almacenamiento, que requieren una estructura predefinida y esquemas rígidos, un Data Lake ofrece una flexibilidad sin precedentes al permitir que los datos se almacenen tal como llegan, sin necesidad de estructurarlos previamente.

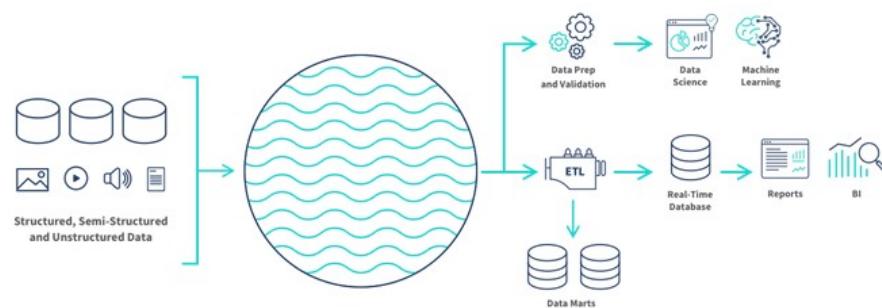


Figura 3. Arquitectura general de un Data Lake. Fuente: Qlik <https://www.qlik.com/us/data-lake>

Esto incluye datos estructurados, semi-estructurados y no estructurados, como archivos de texto, registros de transacciones, datos de sensores, imágenes, videos, archivos de audio, y datos de redes sociales (ver Figura 3).

### Principales características

- ▶ Almacenamiento de datos crudos: Los Data Lakes permiten almacenar datos en su forma original sin necesidad de procesarlos o estructurarlos previamente, lo que facilita la ingesta de datos desde múltiples fuentes.

- ▶ Escalabilidad: Los Data Lakes están diseñados para escalar horizontalmente, lo que significa que pueden manejar grandes volúmenes de datos en crecimiento continuo sin afectar el rendimiento.
- ▶ Flexibilidad: Ofrecen la capacidad de almacenar diferentes tipos de datos, desde estructurados hasta no estructurados, permitiendo un análisis más completo y diverso.
- ▶ Economía: El almacenamiento en Data Lakes es generalmente más económico en comparación con los sistemas de almacenamiento estructurados, ya que no requieren costosos esquemas de preprocesamiento.
- ▶ Acceso múltiple: Permiten el acceso simultáneo a los datos por parte de diferentes herramientas y lenguajes de análisis, facilitando el trabajo de científicos de datos, analistas y desarrolladores.
- ▶ Metadatos y catalogación: Los Data Lakes suelen incorporar herramientas de gestión de metadatos y catalogación que ayudan a los usuarios a encontrar y utilizar los datos de manera eficiente.

## Desventajas

- ▶ Calidad de datos: Al permitir la ingestión de datos sin procesar, los Data Lakes pueden enfrentarse a problemas de calidad de datos, como duplicación, datos incompletos o inconsistentes.
- ▶ Complejidad de gestión: Sin una gestión adecuada, un Data Lake puede convertirse en un "data swamp" (pantano de datos), donde los datos son difíciles de encontrar y utilizar.
- ▶ Rendimiento: La consulta de datos crudos en un Data Lake puede ser menos eficiente en comparación con los datos estructurados en un Data Warehouse.

- ▶ Seguridad y gobernanza: La flexibilidad de los Data Lakes puede dificultar la implementación de políticas de seguridad y gobernanza de datos, lo que puede llevar a riesgos de privacidad y cumplimiento normativo.

## Diferenciación respecto a otros sistemas de almacenamiento

- ▶ Data Mart: Mientras que los Data Marts están diseñados para áreas específicas de negocio con datos estructurados y optimizados para análisis rápidos, un Data Lake almacena todos los datos en su forma original, proporcionando una visión más amplia y menos estructurada de la información.
- ▶ Data Warehouse: A diferencia de un Data Warehouse, que organiza datos en estructuras definidas para consultas y análisis optimizados, un Data Lake permite almacenar datos tal como llegan, ofreciendo mayor flexibilidad para análisis avanzados y big data (NetSuite, 2021).

## Casos de uso

- ▶ Análisis de Big Data: Empresas que manejan grandes volúmenes de datos de diversas fuentes, como registros de clics, datos de sensores y redes sociales, utilizan Data Lakes para almacenar y analizar esta información de manera efectiva.
- Mejora: Permite realizar análisis complejos y correlacionar datos de múltiples fuentes, proporcionando insights profundos y detallados que mejoran la toma de decisiones.

- ▶ Machine Learning y AI: Los Data Lakes son ideales para proyectos de machine learning y AI, donde los algoritmos pueden necesitar acceso a grandes volúmenes de datos crudos para entrenamiento y validación.
- Mejora: Facilita el acceso a datos variados y en grandes volúmenes, mejorando la precisión y efectividad de los modelos de machine learning.
- ▶ Internet de las Cosas (IoT): Las organizaciones que trabajan con dispositivos IoT generan grandes cantidades de datos en tiempo real. Un Data Lake puede almacenar estos datos de manera eficiente para análisis en tiempo real y a largo plazo.
- Mejora: Permite el análisis en tiempo real y la detección de patrones en los datos de IoT, optimizando el funcionamiento de dispositivos y mejorando la toma de decisiones.
- ▶ Análisis de Sentimientos y Redes Sociales: Las empresas que monitorean las redes sociales para entender la percepción del cliente y analizar tendencias pueden beneficiarse de un Data Lake.
- Mejora: Facilita la ingesta y análisis de grandes volúmenes de datos de redes sociales, permitiendo una respuesta más rápida y efectiva a las tendencias del mercado.
- ▶ Integración de Datos Multisectoriales: Organizaciones que necesitan integrar datos de diversas áreas como ventas, marketing, finanzas y operaciones pueden usar un Data Lake para almacenar todos estos datos en un solo lugar.
- Mejora: Proporciona una visión integrada y holística de la organización, facilitando el análisis transversal y la toma de decisiones basada en datos integrados.

- ▶ Investigación y Desarrollo: Las instituciones de investigación que manejan datos de múltiples experimentos y estudios pueden utilizar Data Lakes para almacenar datos brutos y procesados para análisis posteriores.
- Mejora: Facilita el acceso a datos históricos y actuales para la investigación continua, optimizando la reutilización de datos y el descubrimiento de nuevos insights.

Los Data Lakes son una solución poderosa y flexible para el almacenamiento y análisis de grandes volúmenes de datos diversos. Al proporcionar una plataforma unificada para datos estructurados y no estructurados, permiten a las organizaciones aprovechar al máximo sus activos de datos para una amplia gama de aplicaciones y análisis avanzados.

## 6.5. Referencias bibliográficas

Inmon, W. H., Strauss, D., & Neushloss, G. (2008). DW 2.0: *The Architecture for the Next Generation of Data Warehousing*. Morgan Kaufmann.

Kimball, R., Ross, M., Thorntwaite, W., Mundy, J., & Becker, B. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley.

Qlik. (n.d.). What is a Data Mart? Definition, Benefits, Types. Retrieved from Qlik (Qlik).

Teradata. (n.d.). What is a Data Mart? Types and Best Practices. Retrieved from Teradata (Teradata).

NetSuite. (2021). Data Mart Defined: What It Is, Types & How to Implement. Retrieved from NetSuite (Netsuite).

## La importancia de los Data Marts en la inteligencia empresarial

Oracle. (n.d.). ¿Qué es un Data Mart?. Oracle.

<https://www.oracle.com/es/autonomous-database/what-is-data-mart/>

Este artículo de Oracle explica cómo los Data Marts permiten a las empresas segmentar y analizar datos específicos de diferentes áreas de negocio, mejorando así la toma de decisiones y la eficiencia operativa. Destaca cómo los Data Marts pueden proporcionar acceso rápido y enfocado a los datos necesarios para cada departamento, como ventas, marketing o finanzas.

## Data Warehouse: la base de la inteligencia empresarial moderna

IBM. (n.d.). Data Warehouse. IBM. <https://www.ibm.com/topics/data-warehouse>

En este artículo, IBM analiza el papel fundamental de los Data Warehouses en la inteligencia empresarial, destacando su capacidad para integrar datos de múltiples fuentes y proporcionar una visión unificada y coherente de la información. Se describen las principales características de los Data Warehouses, como la consolidación de datos históricos, la capacidad de realizar análisis complejos y la mejora en la toma de decisiones estratégicas. También se aborda cómo los Data Warehouses soportan las necesidades de análisis de grandes volúmenes de datos en las organizaciones modernas, facilitando informes detallados y la detección de tendencias.

## El papel de los Data Lakes en la era del Big Data

PowerData. (n.d.). Data Lake: Qué es, Ventajas y Desventajas. PowerData.  
<https://www.powerdata.es/data-lake>

PowerData explica cómo los Data Lakes están revolucionando la gestión de grandes volúmenes de datos al permitir el almacenamiento flexible y escalable de datos crudos. Se destaca la capacidad de los Data Lakes para manejar datos estructurados, semi-estructurados y no estructurados, lo que los hace ideales para el análisis avanzado y proyectos de Big Data. Además, se discuten las ventajas, como la flexibilidad y la capacidad de escalado, así como las desventajas, incluyendo los desafíos de la calidad de datos y la complejidad en la gestión. El artículo proporciona una visión integral de cómo los Data Lakes pueden impulsar el análisis avanzado y la inteligencia artificial en las organizaciones modernas.

- 1. ¿Qué es un Data Mart?**
  - A. Un sistema de almacenamiento en la nube.
  - B. Una versión simplificada y específica de un Data Warehouse.
  - C. Un tipo de base de datos relacional.
  - D. Un método de análisis de datos en tiempo real.
  
- 2. ¿Cuál es una característica principal de los Data Marts?**
  - A. Integración de datos de múltiples fuentes.
  - B. Almacenamiento de datos en su formato original.
  - C. Escalabilidad horizontal.
  - D. Enfoque específico en un área de negocio.
  
- 3. ¿Cuál de las siguientes es una desventaja de los Data Marts?**
  - A. Alta escalabilidad.
  - B. Flexibilidad en la gestión de datos.
  - C. Pueden llevar a duplicación de datos.
  - D. Integración de múltiples fuentes de datos.
  
- 4. ¿Qué tipo de esquema es comúnmente utilizado en los Data Marts?**
  - A. Esquema en red.
  - B. Esquema estrella.
  - C. Esquema jerárquico.
  - D. Esquema de malla.

- 5.** ¿Para qué tipo de proyectos son especialmente útiles los Data Marts?

  - A. Proyectos a corto plazo.
  - B. Proyectos a largo plazo.
  - C. Proyectos de infraestructura.
  - D. Proyectos de migración de datos.
  
- 6.** ¿Qué es un Data Warehouse?

  - A. Un repositorio centralizado de datos integrados de múltiples fuentes.
  - B. Un sistema de análisis en tiempo real.
  - C. Una base de datos distribuida.
  - D. Un sistema de almacenamiento temporal.
  
- 7.** ¿Cuál es una ventaja clave de un Data Warehouse?

  - A. Menor coste de implementación.
  - B. Flexibilidad en la estructura de datos.
  - C. Proporciona una visión global y consolidada de los datos de la organización.
  - D. Almacenamiento de datos en su formato original.
  
- 8.** ¿Cuál es una desventaja de los Data Warehouses?

  - A. Facilidad de implementación.
  - B. La implementación y mantenimiento pueden ser costosos.
  - C. Alta flexibilidad para cambios rápidos.
  - D. Baja escalabilidad.

**9.** ¿Qué es un Data Lake?

- A. Un sistema de almacenamiento temporal.
- B. Una base de datos relacional.
- C. Un almacenamiento flexible y escalable de datos crudos.
- D. Un sistema de procesamiento en tiempo real.

**10.** ¿Cuál es una característica principal de los Data Lakes?

- A. Almacenan datos altamente estructurados.
- B. Requieren preprocesamiento de datos antes de almacenarlos.
- C. Almacenan datos en su forma original sin procesar.
- D. Son exclusivamente para datos transaccionales.

Ciencia de Datos Aplicada

---

# Tema 7. Estrategias de Aplicación de la Ciencia de Datos y Datos Masivos

# Índice

## Esquema

### Ideas clave

- 7.1. Introducción y objetivos
- 7.2. Inteligencia de Negocio
- 7.3. Analítica de negocio
- 7.4. Minería de Datos
- 7.5. Aprendizaje Automático
- 7.6. Inteligencia Artificial
- 7.7. Referencias bibliográficas

### A fondo

La importancia de los Data Marts en la inteligencia empresarial

Componentes clave de la Inteligencia de Negocio

La relevancia de la analítica predictiva en el entorno empresarial

Ventajas y desventajas de la analítica de negocio

Fundamentos y aplicaciones del aprendizaje automático

Desafíos y consideraciones éticas en el aprendizaje automático

Innovaciones recientes en inteligencia artificial

Visión futura de la inteligencia artificial

## Test

ESTRATEGIAS DE APLICACIÓN DE LA CIENCIA DE DATOS Y DATOS MASIVOS	
INTELIGENCIA DE NEGOCIO	ANALÍTICA DE NEGOCIO
<ul style="list-style-type: none"> <li>• Fundamentos de la inteligencia de negocio: Transformación de datos en información útil para la toma de decisiones.</li> <li>• Componentes clave: Fuentes de datos, almacén de datos, herramientas ETL, análisis y visualización.</li> <li>• Herramientas populares: Tableau, Power BI, QlikView.</li> <li>• Fases de implementación: Definición de objetivos, recopilación de requisitos, diseño de arquitectura, desarrollo e integración, pruebas, despliegue y mantenimiento.</li> <li>• Aplicaciones prácticas: Optimización de inventarios, detección de fraudes, mejora de la calidad del cuidado en salud.</li> </ul>	<ul style="list-style-type: none"> <li>• Introducción a la analítica de negocio: Uso de datos y modelos para mejorar la toma de decisiones.</li> <li>• Herramientas de analítica: SAS, R, Python, IBM SPSS.</li> <li>• Métodos y técnicas: Análisis estadístico, modelos predictivos, análisis de series temporales, análisis de cohortes.</li> <li>• Aplicaciones prácticas: Optimización de procesos, mejora de la toma de decisiones, personalización del marketing, gestión del riesgo, Optimización de la Cadena de Suministro, mejora de la Experiencia del Cliente, optimización de precios, previsión de la demanda, detección de fraudes, análisis de sentimiento.</li> </ul>
MINERÍA DE DATOS	APRENDIZAJE AUTOMÁTICO
<ul style="list-style-type: none"> <li>• Fundamentos de la minería de datos: Descubrimiento de patrones en grandes volúmenes de datos.</li> <li>• Herramientas y software: Weka, RapidMiner, KNIME, Apache Mahout.</li> <li>• Proceso de minería de datos (CRISP-DM): Comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación, despliegue.</li> <li>• Aplicaciones y casos de estudio: Marketing y ventas, detección de fraudes, salud, manufactura.</li> </ul>	<ul style="list-style-type: none"> <li>• Introducción al aprendizaje automático: Algoritmos y modelos que permiten a las máquinas aprender de los datos.</li> <li>• Algoritmos y técnicas: Regresión lineal, árboles de decisión, SVM, redes neuronales, K-means.</li> <li>• Herramientas: Scikit-learn, TensorFlow, Keras, PyTorch.</li> <li>• Evaluación y validación de modelos: Validación cruzada, métricas de rendimiento, conjunto de prueba, regularización.</li> <li>• Implementación en la industria: Sector financiero, salud, marketing, automoción.</li> </ul>
<p><b>INTELIGENCIA ARTIFICIAL</b></p> <ul style="list-style-type: none"> <li>• Introducción a la inteligencia artificial: Simulación de procesos de inteligencia humana por máquinas.</li> <li>• Algoritmos y técnicas: Aprendizaje supervisado, no supervisado, por refuerzo, redes neuronales profundas.</li> <li>• Herramientas y entornos: TensorFlow, PyTorch, Keras, OpenAI Gym.</li> <li>• Aplicaciones actuales de la IA: Salud, finanzas, automoción, atención al cliente, marketing.</li> <li>• Desafíos y consideraciones éticas: Privacidad de los datos, transparencia, sesgo, regulación.</li> <li>• Visión futura de la IA: IA explicable (XAI), IA general, ética y regulación, integración en la vida cotidiana.</li> </ul>	

## 7.1. Introducción y objetivos

La ciencia de datos combina principios y técnicas de matemáticas, estadísticas, informática y conocimiento del dominio para analizar datos y obtener información útil.

El Big Data, por otro lado, se refiere al manejo y procesamiento de grandes volúmenes de datos que no pueden ser tratados con las herramientas tradicionales debido a su tamaño, velocidad y variedad.

Este apartado explora diversas estrategias de aplicación de la ciencia de datos y el Big Data, abarcando desde la inteligencia de negocio hasta la inteligencia artificial. Se examinan las herramientas, técnicas y enfoques que permiten a las organizaciones transformar datos en conocimientos accionables. Estas estrategias no solo mejoran la toma de decisiones y optimizan los procesos empresariales, sino que también impulsan la innovación y proporcionan una ventaja competitiva.

A lo largo de esta sección, se analizarán aplicaciones prácticas y casos de estudio que ilustran cómo diferentes industrias están utilizando estas estrategias para resolver problemas complejos y crear nuevas oportunidades. Desde la analítica de negocio que ayuda a predecir tendencias de mercado hasta el aprendizaje automático que automatiza tareas repetitivas, el objetivo es proporcionar una comprensión integral de cómo los datos pueden transformar organizaciones y sectores enteros.

Además, se abordarán temas emergentes y complementarios como el Internet de las Cosas (IoT), la computación en la nube y la visualización de datos, que juegan un papel crucial en el ecosistema de datos masivos. También se discutirán aspectos críticos como la ética y la seguridad de los datos, subrayando la importancia de gestionar los datos de manera responsable y segura.

En resumen, este apartado ofrece una guía exhaustiva sobre cómo aplicar la ciencia de datos y el Big Data para obtener ventajas competitivas y mejorar las operaciones empresariales, preparándonos para un futuro donde la capacidad de manejar y aprovechar los datos será fundamental para el éxito organizacional.

Al finalizar la lectura de este tema, los lectores deberán ser capaces de:

- ▶ Proveer una comprensión profunda de los principios fundamentales, conceptos y terminología asociada con la ciencia de datos y el Big Data, incluyendo las diferencias y relaciones entre ambas disciplinas.
- ▶ Analizar y discutir las principales estrategias y metodologías utilizadas en la ciencia de datos y Big Data, abarcando áreas como inteligencia de negocio, analítica de negocio, minería de datos, aprendizaje automático e inteligencia artificial, sin enfocarse en la implementación práctica.
- ▶ Examinar el impacto transformador de los datos masivos en diversas industrias y en la sociedad en general, así como las implicaciones éticas, sociales y económicas del uso extensivo de datos y tecnologías asociadas.

## 7.2. Inteligencia de Negocio

La inteligencia de negocio (BI) es una disciplina esencial que integra diversas metodologías, procesos, arquitecturas y tecnologías para transformar datos en bruto en información significativa y útil. BI proporciona a las organizaciones las herramientas necesarias para tomar decisiones informadas y estratégicas basadas en datos precisos y relevantes. Este campo ha evolucionado significativamente desde sus inicios, adaptándose a las crecientes necesidades de las empresas en un entorno cada vez más competitivo y complejo (Negash, 2004).

El propósito principal de la BI es ayudar a las empresas a optimizar su rendimiento mediante la recopilación, integración, análisis y presentación de datos empresariales. Esto incluye la utilización de herramientas de visualización, almacenamiento de datos y técnicas de minería de datos para descubrir patrones y tendencias ocultas. Al proporcionar una visión integral de la empresa, BI permite a los tomadores de decisiones identificar oportunidades, mejorar la eficiencia operativa y obtener una ventaja competitiva.



Figura 1. Herramientas empleadas por la inteligencia de negocios y la administración del conocimiento.

Fuente: Unidad de Apoyo para el Aprendizaje [https://repositorio-uapa.cuaied.unam.mx/repositorio/moodle/pluginfile.php/2789/mod\\_resource/content/1/UAPA-Inteligencia-Negocios-Administracion-Conocimiento-Organizaciones/index.html](https://repositorio-uapa.cuaied.unam.mx/repositorio/moodle/pluginfile.php/2789/mod_resource/content/1/UAPA-Inteligencia-Negocios-Administracion-Conocimiento-Organizaciones/index.html).

En la práctica, la implementación de BI implica varios componentes clave:

- ▶ Fuentes de datos: Estas incluyen bases de datos transaccionales, sistemas de gestión de relaciones con los clientes (CRM), sistemas de planificación de recursos empresariales (ERP) y datos externos de mercado.
- ▶ Almacén de datos: Un repositorio centralizado que almacena datos integrados de múltiples fuentes, diseñado para facilitar el análisis y la generación de informes.
- ▶ Herramientas ETL (Extract, Transform, Load): Procesos que extraen datos de diversas fuentes, los transforman en un formato adecuado y los cargan en el almacén de datos.
- ▶ Herramientas de análisis y minería de datos: Software que permite descubrir patrones, tendencias y relaciones en los datos.

- ▶ Dashboards y visualización de datos: Interfaces que permiten a los usuarios visualizar y explorar datos a través de gráficos interactivos y visualizaciones (Turban et al., 2011).
- ▶ La capacidad de BI para integrar y analizar datos de diversas fuentes hace posible la toma de decisiones más rápida y precisa. Esta capacidad es fundamental en un entorno empresarial donde las decisiones basadas en datos pueden significar la diferencia entre el éxito y el fracaso.

## Herramientas y tecnologías

Las herramientas y tecnologías de BI han evolucionado para atender diversas necesidades y niveles dentro de las organizaciones. Estas herramientas están diseñadas para facilitar la recopilación, el análisis y la presentación de datos de manera efectiva.

### Herramientas de visualización de datos:

- ▶ Tableau: Una de las herramientas de visualización más populares, conocida por su capacidad para crear visualizaciones interactivas y dashboards de forma intuitiva. Tableau permite a los usuarios conectar, visualizar y compartir datos rápidamente, y sus capacidades de análisis son robustas y fáciles de usar (Chen et al., 2012).
- ▶ Power BI: Una herramienta de Microsoft que se integra perfectamente con otros productos de la suite Office y Azure. Power BI es conocido por su capacidad para transformar datos en visualizaciones coherentes y compartir insights en toda la organización.
- ▶ QlikView: Ofrece análisis de datos en memoria, conocido por su rapidez y eficiencia. QlikView permite a los usuarios explorar grandes volúmenes de datos y descubrir patrones que podrían no ser evidentes con otras herramientas (Few, 2006).

## Sistemas de gestión de bases de datos:

- ▶ SQL Server: Proporciona capacidades robustas de gestión de datos y análisis. SQL Server es ideal para manejar grandes volúmenes de datos y realizar análisis complejos.
- ▶ Oracle: Ofrece soluciones avanzadas para la gestión de grandes volúmenes de datos y análisis complejos, siendo una de las plataformas más utilizadas en BI.
- ▶ Hadoop: Ideal para almacenar y procesar grandes volúmenes de datos no estructurados. Hadoop permite a las organizaciones manejar y analizar datos masivos de manera eficiente (Dean & Ghemawat, 2008).

## Herramientas de minería de datos:

- ▶ RapidMiner: Plataforma que permite realizar minería de datos y aprendizaje automático sin necesidad de conocimientos profundos de programación. RapidMiner facilita el proceso de descubrir patrones ocultos en los datos y generar modelos predictivos.
- ▶ KNIME: Software de código abierto que facilita el análisis de datos a través de flujos de trabajo visuales. KNIME es flexible y extensible, lo que permite a los usuarios integrar diferentes tipos de análisis y técnicas de minería de datos en sus proyectos (Fayyad et al., 1996).

## Implementación de BI

Implementar un sistema de BI efectivo requiere un enfoque estructurado y meticuloso. El ciclo de vida de un proyecto de BI incluye varias fases críticas:

- ▶ **Definición de objetivos y alcance:** El primer paso en la implementación de un proyecto de BI es definir claramente los objetivos que se desean alcanzar. Estos objetivos deben ser específicos, medibles, alcanzables, relevantes y con un marco

temporal definido. Además, es crucial determinar el alcance del proyecto, es decir, qué áreas de la organización se verán afectadas y qué datos se analizarán (Yeoh & Koronios, 2010).

- ▶ **Recopilación de requisitos:** En esta fase, se recopilan los requisitos de los usuarios y las especificaciones técnicas necesarias para el proyecto. Es fundamental involucrar a los stakeholders clave desde el principio para asegurarse de que el sistema de BI satisfaga sus necesidades y expectativas. Esto incluye entrevistas, encuestas y talleres para comprender las necesidades de información y los procesos empresariales actuales (Watson & Wixom, 2007).
- ▶ **Diseño de la arquitectura:** El diseño de la arquitectura del sistema de BI implica planificar la infraestructura tecnológica necesaria, que incluye el almacén de datos, las herramientas ETL, y las herramientas de análisis y visualización. Es esencial diseñar una arquitectura que sea escalable y flexible para adaptarse a futuros cambios y expansiones (Turban et al., 2011).
- ▶ **Desarrollo e integración:** En esta fase, se construyen y configuran las soluciones de BI. Esto incluye el desarrollo de procesos ETL para integrar datos de diversas fuentes, la configuración del almacén de datos y el desarrollo de dashboards y reportes personalizados. La integración de diferentes sistemas y fuentes de datos es una tarea compleja que requiere una planificación cuidadosa y pruebas exhaustivas (Yeoh & Koronios, 2010).
- ▶ **Pruebas y validación:** Una vez desarrollado el sistema de BI, es crucial realizar pruebas exhaustivas para asegurar que funciona correctamente y cumple con los requisitos definidos. Esto incluye pruebas funcionales, de rendimiento y de seguridad. La validación de los datos es esencial para asegurar su calidad y precisión (Wixom & Watson, 2001).
- ▶ **Despliegue y capacitación:** Tras las pruebas, el sistema de BI se despliega en el entorno de producción. Es fundamental capacitar a los usuarios finales en el uso de las nuevas herramientas y procesos. La capacitación debe ser continua para

asegurar que los usuarios puedan aprovechar al máximo las capacidades del sistema (Turban et al., 2011).

- ▶ **Mantenimiento y mejora continua:** Una vez que el sistema de BI está en funcionamiento, es necesario realizar un mantenimiento continuo para asegurar su rendimiento y relevancia. Esto incluye la actualización de los datos, la incorporación de nuevas fuentes de datos y la mejora de las funcionalidades existentes. La mejora continua es esencial para mantener el sistema alineado con las necesidades cambiantes del negocio (Yeoh & Koronios, 2010).

## Casos de uso

La inteligencia de negocio se aplica en diversos sectores para resolver problemas específicos y mejorar la eficiencia operativa.

Sector minorista:

- ▶ En el sector minorista, BI se utiliza para optimizar la gestión de inventarios y mejorar la experiencia del cliente. Al analizar patrones de compra y comportamiento del cliente, las empresas pueden predecir la demanda de productos y ajustar sus inventarios en consecuencia. Esto no solo reduce los costos de almacenamiento, sino que también asegura que los productos estén disponibles cuando los clientes los necesiten (Chen et al., 2012).

Sector financiero:

- ▶ En el sector financiero, BI juega un papel crucial en la detección de fraudes y la gestión de riesgos. Las herramientas de BI permiten analizar transacciones en tiempo real para identificar patrones sospechosos que podrían indicar fraude. Además, los modelos de análisis predictivo ayudan a las instituciones financieras a evaluar y gestionar riesgos, asegurando la estabilidad financiera y el cumplimiento de las regulaciones (Negash, 2004).

## Sector salud:

- ▶ En el sector salud, BI se utiliza para mejorar la calidad del cuidado del paciente y la eficiencia operativa. Al analizar datos clínicos y administrativos, los proveedores de salud pueden identificar áreas de mejora en la atención al paciente y optimizar los recursos. Esto incluye la gestión de costos, la optimización de los flujos de trabajo y la mejora de los resultados clínicos (Turban et al., 2011).

## Sector manufacturero:

- ▶ En la manufactura, BI ayuda a mejorar la eficiencia de la producción y la gestión de la cadena de suministro. Al analizar datos de producción y logística, las empresas pueden identificar cuellos de botella y optimizar los procesos de producción. Esto resulta en una reducción de costos y un aumento en la productividad (Watson & Wixom, 2007).

## 7.3. Analítica de negocio

La analítica de negocio (BA, por sus siglas en inglés) se refiere al uso de datos, análisis estadísticos y técnicas de modelado predictivo para tomar decisiones empresariales informadas y optimizar los resultados. La analítica de negocio se enfoca en la interpretación de datos históricos y actuales para prever tendencias futuras y comportamientos. Esta disciplina se ha convertido en un pilar fundamental para las organizaciones que buscan mejorar su competitividad y eficiencia operativa (Davenport & Harris, 2007).

Componentes clave de la analítica de negocio:

- ▶ **Análisis descriptivo:** Se centra en la interpretación de datos históricos para entender qué ha ocurrido en el pasado. Utiliza técnicas como informes, dashboards y visualizaciones para proporcionar una visión clara de los eventos históricos.
- ▶ **Análisis predictivo:** Utiliza modelos estadísticos y algoritmos de aprendizaje automático para predecir futuros eventos basados en datos históricos. Esta técnica es esencial para prever tendencias y tomar decisiones proactivas.
- ▶ **Análisis prescriptivo:** Va un paso más allá del análisis predictivo al no solo prever lo que puede suceder, sino también recomendar acciones específicas que pueden influir en esos futuros eventos. Utiliza técnicas como simulación y optimización para sugerir las mejores opciones de acción (Shmueli & Koppius, 2011).

La integración de estas técnicas permite a las organizaciones comprender mejor sus datos y tomar decisiones más informadas y efectivas.

## Herramientas de analítica

Las herramientas de analítica de negocio son variadas y están diseñadas para ayudar a las organizaciones a recolectar, procesar, analizar y visualizar datos de manera efectiva.

Herramientas populares de analítica de negocio:

- ▶ SAS (Statistical Analysis System): Un software líder en el campo de la analítica avanzada que ofrece una amplia gama de capacidades analíticas, desde la gestión de datos hasta la analítica predictiva y la minería de datos. SAS es conocido por su robustez y flexibilidad en el manejo de grandes volúmenes de datos (SAS Institute, 2012).
- ▶ R: Un lenguaje de programación y entorno de software para la estadística y el análisis gráfico. R es ampliamente utilizado en la analítica de negocio debido a su capacidad para manejar datos complejos y realizar análisis estadísticos avanzados (R Core Team, 2013).
- ▶ Python: Un lenguaje de programación versátil que se ha convertido en una herramienta esencial para la analítica de negocio debido a sus bibliotecas específicas para análisis de datos como Pandas, NumPy, y Scikit-learn. Python es apreciado por su facilidad de uso y su capacidad para integrarse con otros sistemas (Van Rossum & Drake, 2009).
- ▶ IBM SPSS (Statistical Package for the Social Sciences): Ofrece herramientas avanzadas para la analítica estadística y predictiva. SPSS es popular en la investigación de mercado, salud y educación debido a su facilidad de uso y potentes capacidades analíticas (IBM, 2011).

## Métodos y técnicas

La analítica de negocio utiliza una variedad de métodos y técnicas para analizar datos y obtener insights significativos.

Métodos y técnicas clave en analítica de negocio:

- ▶ Análisis estadístico: Incluye técnicas como regresión, análisis de varianza (ANOVA), y pruebas de hipótesis. Estas técnicas son fundamentales para identificar relaciones y patrones en los datos.
- ▶ Modelos predictivos: Utilizan algoritmos como árboles de decisión, redes neuronales y máquinas de soporte vectorial (SVM) para prever futuros eventos basados en datos históricos. Los modelos predictivos son esenciales para anticipar comportamientos del cliente, detectar fraudes y gestionar riesgos (Friedman, 1997).
- ▶ Análisis de series temporales: Involucra la evaluación de datos secuenciales a lo largo del tiempo para identificar tendencias, patrones estacionales y ciclos. Esta técnica es particularmente útil en la previsión de ventas y la planificación de la demanda (Box & Jenkins, 1976).
- ▶ Análisis de cohortes: Permite segmentar datos en grupos basados en características compartidas y analizar sus comportamientos a lo largo del tiempo. Esto es útil para entender la retención de clientes y el impacto de las campañas de marketing.

## Aplicaciones prácticas

La analítica de negocio se aplica en diversos contextos para resolver problemas específicos y mejorar la eficiencia operativa de las organizaciones.

Aplicaciones prácticas de la analítica de negocio:

- ▶ **Optimización de procesos:** La analítica de negocio permite identificar ineficiencias en los procesos operativos y sugerir mejoras. Por ejemplo, en la manufactura, el análisis de datos de producción puede revelar cuellos de botella y oportunidades para optimizar la cadena de suministro (Chopra & Meindl, 2013).
- ▶ **Mejora de la toma de decisiones:** Utilizando modelos predictivos, las organizaciones pueden anticipar tendencias del mercado y adaptar sus estrategias en consecuencia. Esto es crucial en sectores como el retail y las finanzas, donde las condiciones del mercado pueden cambiar rápidamente.
- ▶ **Personalización del marketing:** La analítica de datos permite segmentar a los clientes y personalizar las campañas de marketing. Al analizar el comportamiento de compra y las preferencias del cliente, las empresas pueden diseñar campañas más efectivas y dirigidas (Wedel & Kannan, 2016).
- ▶ **Gestión del riesgo:** En el sector financiero, la analítica predictiva se utiliza para evaluar y gestionar riesgos. Esto incluye la identificación de posibles fraudes y la evaluación de la solvencia crediticia de los clientes (Jorion, 2007).

## 7.4. Minería de Datos

La minería de datos, también conocida como data mining, es un proceso analítico diseñado para explorar grandes cantidades de datos en busca de patrones consistentes o relaciones sistemáticas entre variables. Utiliza técnicas avanzadas de análisis estadístico, aprendizaje automático y bases de datos para descubrir y validar estos patrones. La minería de datos se ha convertido en una herramienta crucial para las organizaciones que buscan transformar datos en conocimiento útil y aplicable (Han, Kamber & Pei, 2012). Los componentes clave de la minería de datos son los siguientes:

- ▶ **Recopilación de datos:** La primera fase de la minería de datos implica la recopilación de grandes volúmenes de datos de diversas fuentes. Estos datos pueden provenir de sistemas transaccionales, bases de datos relacionales, archivos de texto, imágenes, videos y más.
- ▶ **Preparación de datos:** Esta fase implica limpiar y transformar los datos para que sean adecuados para el análisis. Incluye la eliminación de duplicados, el manejo de datos faltantes y la normalización de los datos.
- ▶ **Selección de datos:** No todos los datos recopilados son relevantes para cada análisis. Por lo tanto, es crucial seleccionar un subconjunto de datos que sea representativo y relevante para el problema que se está tratando de resolver.
- ▶ **Modelado:** Implica la aplicación de algoritmos y técnicas de minería de datos para identificar patrones en los datos. Los modelos pueden ser descriptivos (identificando patrones existentes) o predictivos (prediciendo futuros eventos basados en datos históricos).
- ▶ **Evaluación:** Una vez que se ha construido un modelo, se evalúa su efectividad y precisión. Esto puede implicar la validación cruzada y el uso de conjuntos de datos de prueba para medir el rendimiento del modelo.

- ▶ **Despliegue:** Los modelos validados se implementan en el entorno de producción donde pueden ser utilizados para tomar decisiones empresariales informadas (Fayyad, Piatetsky-Shapiro & Smyth, 1996).

## Herramientas y software

Existen diversas herramientas y software de minería de datos que ayudan a las organizaciones a realizar análisis complejos de manera eficiente.

Herramientas populares de minería de datos:

- ▶ Weka: Un conjunto de herramientas de aprendizaje automático para minería de datos que incluye una colección de algoritmos y visualización de datos. Weka es conocido por su facilidad de uso y su capacidad para integrar múltiples técnicas de minería de datos en un solo entorno.
- ▶ RapidMiner: Una plataforma de análisis de datos que permite realizar procesos de minería de datos y aprendizaje automático. RapidMiner es popular por su interfaz intuitiva y su capacidad para manejar grandes volúmenes de datos.
- ▶ KNIME: (Konstanz Information Miner) es una plataforma de integración de datos, procesamiento y análisis de datos de código abierto. KNIME permite la creación de flujos de trabajo visuales y la integración de diversas técnicas de minería de datos.
- ▶ Apache Mahout: Una biblioteca de aprendizaje automático escalable para sistemas distribuidos. Mahout es utilizado principalmente para realizar minería de datos en grandes conjuntos de datos distribuidos a través de sistemas como Hadoop (Witten, Frank, Hall & Pal, 2016).

## Proceso de minería de datos

El proceso de minería de datos es un ciclo iterativo que sigue una serie de pasos sistemáticos para asegurar que los resultados obtenidos sean precisos y útiles. CRISP-DM, que son las siglas de Cross-Industry Standard Process for Data Mining, es un método probado para orientar sus trabajos de minería de datos.

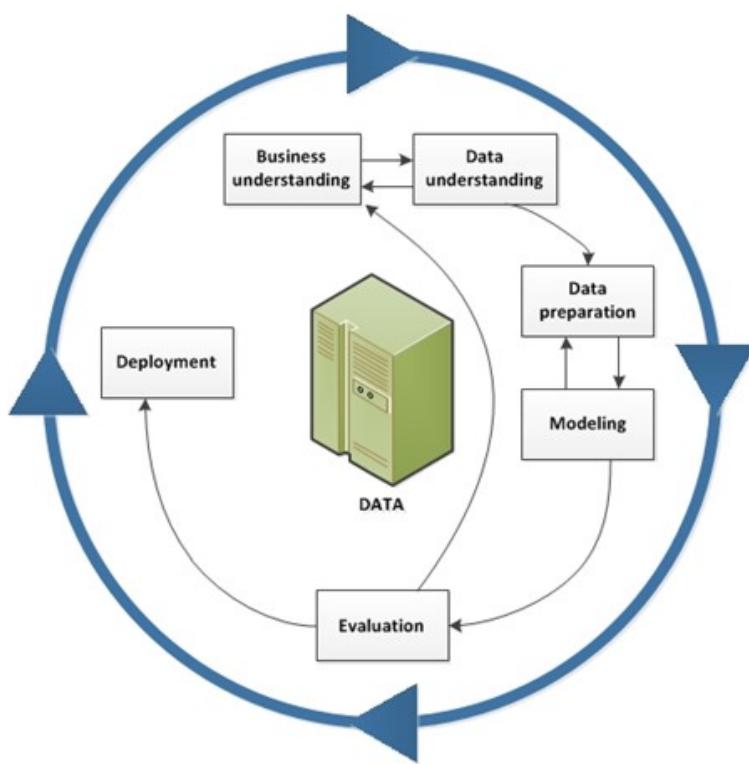


Figura 2. Ciclo de vida de minería de datos. Fuente: <https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview>.

Como **metodología**, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.

Como **modelo de proceso**, CRISP-DM ofrece un resumen del ciclo vital de minería de datos.

Fases del proceso de minería de datos (CRISP-DM):

- ▶ **Comprendión del negocio:** El primer paso es comprender los objetivos del negocio y los requisitos del proyecto desde una perspectiva empresarial. Esto ayuda a establecer un plan de proyecto y a definir los criterios de éxito.
- ▶ **Comprendión de los datos:** Implica la recopilación inicial de datos y la exploración de estos para identificar problemas de calidad de los datos, descubrir patrones preliminares y obtener una comprensión general de los datos disponibles.
- ▶ **Preparación de los datos:** Incluye todas las actividades necesarias para construir el conjunto de datos final a partir de los datos iniciales. Las tareas pueden incluir selección de atributos, limpieza de datos, creación de nuevas variables y transformación de datos.
- ▶ **Modelado:** Aplicación de técnicas de modelado y algoritmos a los datos preparados. Esta fase puede requerir la selección de técnicas específicas y la calibración de parámetros del modelo.
- ▶ **Evaluación:** Evaluación de los modelos construidos para asegurarse de que cumplen con los objetivos del negocio. Esto incluye validar los modelos utilizando datos de prueba y analizar los resultados para garantizar su relevancia.
- ▶ **Despliegue:** La fase final implica la implementación del modelo en un entorno de producción. Esto puede incluir la integración del modelo en sistemas de toma de decisiones y la monitorización de su rendimiento a lo largo del tiempo (Chapman et al., 2000).

El modelo de CRISP-DM es flexible y se pueden personalizar fácilmente. Por ejemplo, si su organización intenta detectar actividades de blanqueo de dinero, es probable que necesite realizar una criba de grandes cantidades de datos sin un objetivo de modelado específico. En lugar de realizar el modelado, su trabajo se centrará en explorar y visualizar datos para descubrir patrones sospechosos en datos

financieros. CRISP-DM permite crear un modelo de minería de datos que se adapte a sus necesidades concretas.

## Aplicaciones y casos de estudio

La minería de datos se aplica en diversas industrias para resolver una amplia gama de problemas empresariales y obtener insights valiosos.

Aplicaciones de minería de datos:

- ▶ **Marketing y ventas:** Las empresas utilizan la minería de datos para segmentar clientes, predecir comportamientos de compra y diseñar campañas de marketing dirigidas. Por ejemplo, los modelos predictivos pueden identificar qué clientes son más propensos a responder positivamente a una oferta de marketing específica.
- ▶ **Detección de fraudes:** En el sector financiero, la minería de datos se utiliza para detectar patrones inusuales en las transacciones que podrían indicar fraude. Los algoritmos de aprendizaje automático pueden analizar grandes volúmenes de datos transaccionales para identificar comportamientos anómalos.
- ▶ **Salud:** Los proveedores de servicios de salud utilizan la minería de datos para analizar registros médicos y descubrir patrones asociados con enfermedades y tratamientos. Esto puede mejorar la precisión del diagnóstico y personalizar los tratamientos para los pacientes.
- ▶ **Manufactura:** En la industria manufacturera, la minería de datos ayuda a optimizar los procesos de producción mediante el análisis de datos de sensores y maquinaria. Esto puede reducir el tiempo de inactividad y mejorar la calidad del producto (Han et al., 2012).

## 7.5. Aprendizaje Automático

El aprendizaje automático (ML, por sus siglas en inglés) es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender de los datos y mejorar su desempeño con la experiencia. En lugar de seguir instrucciones explícitas programadas, los sistemas de ML identifican patrones y relaciones en los datos, permitiendo hacer predicciones y tomar decisiones de manera autónoma (Mitchell, 1997).

Componentes clave del aprendizaje automático:

- ▶ **Datos:** El fundamento del aprendizaje automático son los datos. Los algoritmos de ML requieren grandes volúmenes de datos etiquetados para entrenarse y mejorar su precisión.
- ▶ **Algoritmos:** Los algoritmos son las matemáticas y las reglas que rigen el aprendizaje. Entre los más comunes se encuentran la regresión lineal, los árboles de decisión, las redes neuronales y las máquinas de soporte vectorial (SVM).
- ▶ **Modelos:** Un modelo es la salida del entrenamiento de un algoritmo con datos. Representa una abstracción matemática de los patrones encontrados en los datos.
- ▶ **Entrenamiento:** El proceso de ajustar los parámetros del modelo para minimizar el error y mejorar la precisión de las predicciones.
- ▶ **Validación y prueba:** Evaluar el rendimiento del modelo con un conjunto diferente de datos para asegurar que funcione bien y no esté sobreajustado a los datos de entrenamiento (overfitting).

## Algoritmos y técnicas

El aprendizaje automático abarca una variedad de algoritmos y técnicas, cada uno adecuado para diferentes tipos de problemas y conjuntos de datos.

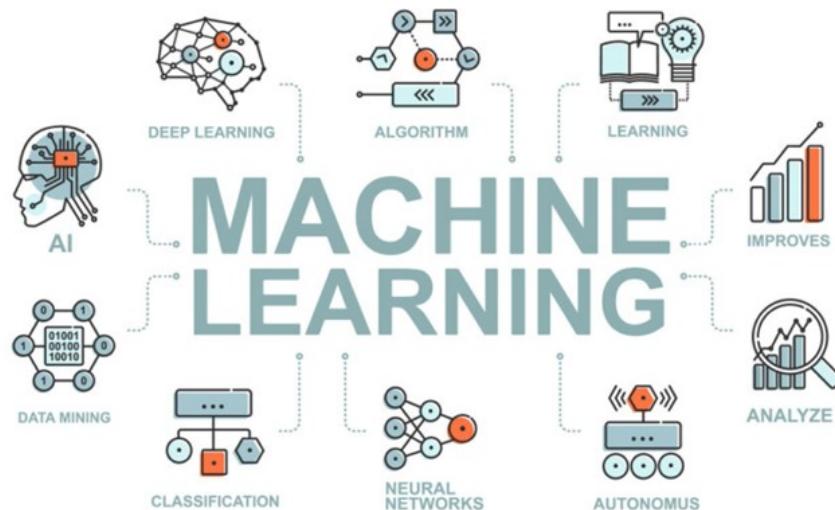


Figura3. Beneficios y desafíos del Machine Learning. Fuente: <https://sensoricx.com/conocimiento/machine-learning/>

## Algoritmos y técnicas clave en aprendizaje automático:

- ▶ **Regresión lineal:** Un método estadístico para modelar la relación entre una variable dependiente y una o más variables independientes. Es comúnmente utilizado para predicción y análisis de tendencias.
- ▶ **Árboles de decisión:** Algoritmos que dividen repetidamente los datos en subconjuntos basados en valores de características, formando una estructura similar a un árbol. Son útiles para tareas de clasificación y regresión.
- ▶ **Máquinas de soporte vectorial (SVM):** Algoritmos de clasificación que encuentran el hiperplano óptimo que separa las diferentes clases en el espacio de características. Son efectivos en problemas de alta dimensionalidad.

- ▶ **Redes neuronales:** Modelos inspirados en la estructura del cerebro humano, compuestos por capas de nodos (neuronas) que procesan información y aprenden patrones complejos. Las redes neuronales profundas (Deep Learning) han revolucionado campos como la visión por computadora y el procesamiento del lenguaje natural (Goodfellow, Bengio & Courville, 2016).
- ▶ **K-means:** Un algoritmo de clustering que agrupa datos en k clusters basados en características similares. Es útil para segmentación de clientes y análisis exploratorio de datos.

## Herramientas y entornos

Existen numerosas herramientas y entornos de desarrollo que facilitan la implementación y experimentación con algoritmos de aprendizaje automático.

Herramientas populares de aprendizaje automático:

- ▶ **Scikit-learn:** Una biblioteca de Python que proporciona herramientas simples y eficientes para el análisis de datos y minería de datos. Es ampliamente utilizada por su facilidad de uso y su extensa documentación (Pedregosa et al., 2011).
- ▶ **TensorFlow:** Una biblioteca de código abierto desarrollada por Google para la implementación de redes neuronales y otros algoritmos de aprendizaje profundo. Es conocida por su flexibilidad y escalabilidad.
- ▶ **Keras:** Una biblioteca de alto nivel que se ejecuta sobre TensorFlow y facilita la construcción y el entrenamiento de modelos de redes neuronales. Keras es popular por su simplicidad y modularidad.
- ▶ **PyTorch:** Una biblioteca de aprendizaje automático de código abierto desarrollada por Facebook. PyTorch es conocida por su capacidad para trabajar con gráficos computacionales dinámicos, lo que la hace muy adecuada para la investigación en aprendizaje profundo (Paszke et al., 2019).

## Evaluación y validación de modelos

La evaluación y validación de modelos es un paso crucial en el proceso de desarrollo de modelos de aprendizaje automático. Es esencial para garantizar que los modelos funcionen correctamente y se generalicen bien a datos no vistos.

Métodos de evaluación y validación:

- ▶ **Validación cruzada:** Un método que divide los datos en múltiples subconjuntos y entrena el modelo en algunos de estos subconjuntos mientras lo valida en los restantes. Este proceso se repite varias veces para asegurar una evaluación robusta.
- ▶ **Métricas de rendimiento:** Dependiendo del tipo de problema (clasificación o regresión), se utilizan diferentes métricas como precisión, recall, F1-score para clasificación y error cuadrático medio (MSE) para regresión.
- ▶ **Conjunto de prueba:** Después del entrenamiento, el modelo se evalúa utilizando un conjunto de datos que no se utilizó durante el entrenamiento ni la validación. Esto proporciona una medida imparcial del rendimiento del modelo.
- ▶ **Regularización:** Técnicas como L1 y L2 regularización se utilizan para prevenir el sobreajuste ajustando los parámetros del modelo para que no se ajusten demasiado a los datos de entrenamiento (Bishop, 2006).

## Implementación en la industria

El aprendizaje automático se ha implementado exitosamente en diversas industrias, transformando procesos y creando nuevas oportunidades.

Aplicaciones de aprendizaje automático en la industria:

- ▶ **Sector financiero:** Utilizado para la detección de fraudes, evaluación de riesgos y predicción de precios de acciones. Los modelos de ML analizan transacciones en tiempo real y detectan patrones sospechosos, ayudando a prevenir fraudes.
- ▶ **Salud:** En el sector salud, ML se aplica en la predicción de enfermedades, personalización de tratamientos y análisis de imágenes médicas. Los modelos de aprendizaje profundo pueden analizar imágenes de resonancias magnéticas para detectar anomalías que indican enfermedades (Obermeyer & Emanuel, 2016).
- ▶ **Marketing:** En marketing, ML se utiliza para la segmentación de clientes, personalización de campañas publicitarias y predicción de churn (deserción de clientes). Los modelos analizan el comportamiento del cliente y personalizan las ofertas para aumentar la retención y satisfacción del cliente.
- ▶ **Automoción:** Los vehículos autónomos utilizan algoritmos de ML para procesar datos de sensores y tomar decisiones en tiempo real. Esto incluye el reconocimiento de señales de tráfico, detección de obstáculos y navegación autónoma (Bojarski et al., 2016).

## 7.6. Inteligencia Artificial

La inteligencia artificial (IA) se refiere a la simulación de procesos de inteligencia humana por parte de máquinas, especialmente sistemas informáticos. Estos procesos incluyen el aprendizaje (la adquisición de información y reglas para el uso de la información), el razonamiento (usar reglas para llegar a conclusiones aproximadas o definitivas) y la autocorrección. Las aplicaciones de IA incluyen sistemas expertos, procesamiento del lenguaje natural (NLP), reconocimiento de voz y visión artificial (Russell & Norvig, 2020).

Componentes clave de la inteligencia artificial:

- ▶ **Algoritmos:** Los algoritmos son las secuencias de instrucciones que las máquinas siguen para resolver problemas y tomar decisiones. Estos pueden variar desde algoritmos simples de búsqueda y clasificación hasta complejas redes neuronales.
- ▶ **Datos:** La IA requiere grandes cantidades de datos para entrenar y mejorar sus modelos. Los datos pueden provenir de diversas fuentes como sensores, bases de datos, y plataformas en línea.
- ▶ **Modelos:** Un modelo de IA es una representación matemática de un sistema o proceso que se ha entrenado para realizar tareas específicas.
- ▶ **Entrenamiento y aprendizaje:** El proceso mediante el cual un modelo de IA mejora su rendimiento ajustando sus parámetros con base en los datos de entrenamiento.
- ▶ **Interfaz de usuario:** La manera en que los usuarios interactúan con sistemas de IA, que puede incluir interfaces de voz, visuales, y textuales.

## Algoritmos y técnicas de IA

La IA abarca una amplia gama de algoritmos y técnicas, cada uno adecuado para diferentes aplicaciones y problemas.

Algoritmos y técnicas clave en IA:

- ▶ **Aprendizaje supervisado:** Este enfoque utiliza datos etiquetados para entrenar modelos que pueden hacer predicciones o clasificaciones. Ejemplos incluyen la regresión lineal, máquinas de soporte vectorial (SVM) y redes neuronales.
- ▶ **Aprendizaje no supervisado:** Se utiliza para encontrar patrones o agrupaciones en datos no etiquetados. Ejemplos incluyen algoritmos de clustering como k-means y análisis de componentes principales (PCA).
- ▶ **Aprendizaje por refuerzo:** Un enfoque en el que un agente aprende a comportarse en un entorno realizando acciones y recibiendo recompensas o castigos. Este método es fundamental en el desarrollo de sistemas de IA para juegos y robótica (Sutton & Barto, 2018).
- ▶ **Redes neuronales profundas:** Estas son un tipo de red neuronal con múltiples capas entre la entrada y la salida, que son capaces de aprender representaciones complejas de datos. El aprendizaje profundo ha revolucionado áreas como la visión por computadora y el procesamiento del lenguaje natural (Goodfellow, Bengio & Courville, 2016).

## Herramientas y entornos

La implementación de IA se facilita mediante diversas herramientas y plataformas que permiten desarrollar y desplegar modelos de IA de manera eficiente.

Herramientas populares de IA:

- ▶ **TensorFlow:** Una biblioteca de código abierto desarrollada por Google que se utiliza ampliamente para el desarrollo de modelos de aprendizaje profundo.
- ▶ **PyTorch:** Desarrollada por Facebook, esta biblioteca es popular en la investigación y desarrollo de IA debido a su flexibilidad y facilidad de uso.
- ▶ **Keras:** Una biblioteca de alto nivel que se ejecuta sobre TensorFlow, que facilita la creación y el entrenamiento de redes neuronales.
- ▶ **OpenAI Gym:** Un toolkit para el desarrollo y la comparación de algoritmos de aprendizaje por refuerzo.

## Aplicaciones actuales de la IA

La IA se está utilizando en una amplia gama de aplicaciones en diversas industrias, transformando procesos y creando nuevas oportunidades.

Aplicaciones de IA en la industria:

- ▶ **Salud:** La IA se utiliza para el análisis de imágenes médicas, predicción de enfermedades, y personalización de tratamientos. Modelos de IA pueden analizar imágenes de resonancias magnéticas para detectar anomalías que indican enfermedades.
- ▶ **Finanzas:** Utilizada para la detección de fraudes, evaluación de riesgos y análisis de mercado. La IA puede analizar transacciones en tiempo real para detectar patrones sospechosos.
- ▶ **Automoción:** Los vehículos autónomos utilizan IA para procesar datos de sensores y tomar decisiones en tiempo real, incluyendo el reconocimiento de señales de tráfico y la detección de obstáculos.

- ▶ **Atención al cliente:** Chatbots y asistentes virtuales impulsados por IA están mejorando la eficiencia y la experiencia del cliente en diversas industrias.
- ▶ **Marketing:** La IA se utiliza para la segmentación de clientes, personalización de campañas publicitarias y análisis de sentimiento en redes sociales (Kaplan & Haenlein, 2019).

## Desafíos y consideraciones éticas

El desarrollo y la implementación de la IA presentan varios desafíos técnicos, sociales y éticos que deben ser considerados.

### Desafíos en IA

- ▶ **Privacidad de los datos:** El uso de grandes volúmenes de datos personales plantea preocupaciones sobre la privacidad y la seguridad de la información.
- ▶ **Transparencia:** La complejidad de los modelos de IA, especialmente las redes neuronales profundas, puede hacer difícil entender y explicar cómo toman decisiones (black box problem).
- ▶ **Sesgo:** Los modelos de IA pueden heredar y amplificar sesgos presentes en los datos de entrenamiento, lo que puede llevar a resultados injustos o discriminatorios.
- ▶ **Regulación:** La falta de regulaciones claras y estandarizadas en torno a la IA plantea riesgos en cuanto a la seguridad y la responsabilidad (Mittelstadt et al., 2016).

## Visión futura de la IA

La inteligencia artificial continúa evolucionando rápidamente, y su futuro promete aún más innovaciones y desafíos.

Tendencias futuras en IA:

- ▶ IA explicable (XAI): Desarrollar técnicas que hagan que los modelos de IA sean más transparentes y comprensibles para los humanos.
- ▶ IA general: Avances hacia la creación de IA con capacidades cognitivas generales similares a las humanas, capaces de realizar cualquier tarea intelectual que un humano puede hacer.
- ▶ Ética y regulación: Se espera un aumento en la creación de marcos éticos y regulaciones para guiar el desarrollo y uso responsable de la IA.
- ▶ Integración en la vida cotidiana: La IA será cada vez más omnipresente, integrándose en dispositivos del hogar, vehículos, y la infraestructura urbana para mejorar la calidad de vida y la eficiencia operativa (Bostrom, 2014).

## 7.7. Referencias bibliográficas

- Baars, H., & Kemper, H. (2008). Management support with structured and unstructured data—an integrated business intelligence framework. *Information Systems Management*, 25(2), 132-148.
- Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165-1188.
- Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107-113.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37-54.
- Few, S. (2006). *Information Dashboard Design: The Effective Visual Communication of Data*. O'Reilly Media, Inc.
- Negash, S. (2004). Business intelligence. *Communications of the Association for Information Systems*, 13(1), 177-195.
- Turban, E., Sharda, R., & Delen, D. (2011). *Decision support and business intelligence systems*. Pearson Education.
- Watson, H. J., & Wixom, B. H. (2007). The current state of business intelligence. *Computer*, 40(9), 96-99.
- Wixom, B. H., & Watson, H. J. (2001). An empirical investigation of the factors affecting data warehousing success. *MIS Quarterly*, 25(1), 17-41.
- Yeoh, W., & Koronios, A. (2010). Critical success factors for business intelligence systems. *Journal of Computer Information Systems*, 50(3), 23-32.

- Yeoh, W., & Popović, A. (2016). Extending the understanding of critical success factors for implementing business intelligence systems. *Journal of the Association for Information Science and Technology*, 67(1), 134-147.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. *The New England Journal of Medicine*, 375(13), 1216-1219.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32, 8026-8037.
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
- Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence. *Business Horizons*, 62(1), 15-25.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3(2), 2053951716679679.
- Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. Pearson.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

## La importancia de los Data Marts en la inteligencia empresarial

Oracle. (n.d.). ¿Qué es un Data Mart?. Oracle.

<https://www.oracle.com/es/autonomous-database/what-is-data-mart/>

Este artículo de Oracle explica cómo los Data Marts permiten a las empresas segmentar y analizar datos específicos de diferentes áreas de negocio, mejorando así la toma de decisiones y la eficiencia operativa. Destaca cómo los Data Marts pueden proporcionar acceso rápido y enfocado a los datos necesarios para cada departamento, como ventas, marketing o finanzas.

## Componentes clave de la Inteligencia de Negocio

Oracle. (n.d.). *Oracle Business Intelligence*. Oracle. Oracle.

<https://www.oracle.com/business-analytics/business-intelligence/>

En este recurso se describen los componentes esenciales de las soluciones de inteligencia de negocio de Oracle, como Oracle BI Answers, Oracle BI Interactive Dashboards, Oracle BI Publisher y Oracle Real-Time Decisions. Estos componentes permiten a las organizaciones analizar datos de manera interactiva, crear reportes y tomar decisiones en tiempo real basadas en datos.

### La relevancia de la analítica predictiva en el entorno empresarial

IBM. (n.d.). *Predictive Analytics*. IBM. <https://www.ibm.com/analytics/predictive-analytics>

Este recurso de IBM explora cómo la analítica predictiva utiliza datos históricos para predecir eventos futuros, permitiendo a las empresas anticiparse a las necesidades del mercado y mejorar su toma de decisiones. Se destaca la aplicación de modelos predictivos en sectores como el marketing, la gestión de riesgos y la cadena de suministro.

## Ventajas y desventajas de la analítica de negocio

SAS. (n.d.). *What is Business Analytics?* SAS.

[https://www.sas.com/en\\_us/insights/analytics/big-data-analytics.html](https://www.sas.com/en_us/insights/analytics/big-data-analytics.html)

Este artículo de SAS proporciona una visión general de los beneficios y limitaciones de la analítica de negocio. Entre las ventajas se incluyen la capacidad de tomar decisiones informadas y mejorar la eficiencia operativa, mientras que las desventajas pueden incluir la complejidad de implementación y la necesidad de competencias especializadas en análisis de datos.

## Fundamentos y aplicaciones del aprendizaje automático

Google Cloud. (n.d.). *What is Machine Learning?* Google Cloud.  
<https://cloud.google.com/learn/what-is-machine-learning>

Google Cloud explica los conceptos básicos del aprendizaje automático y cómo se aplica en diversas industrias, desde la detección de fraudes hasta la personalización de la experiencia del usuario. El artículo destaca la importancia de los algoritmos y modelos en la automatización de tareas y la mejora de la precisión en las predicciones.

## Desafíos y consideraciones éticas en el aprendizaje automático

Microsoft Azure. (n.d.). *Machine Learning and AI*. Microsoft Azure.

<https://azure.microsoft.com/en-us/products/machine-learning>

Microsoft Azure aborda los desafíos y cuestiones éticas asociadas con el aprendizaje automático, como la privacidad de los datos, el sesgo en los modelos y la transparencia en los procesos de decisión automatizados. Se enfatiza la necesidad de desarrollar prácticas responsables para garantizar la equidad y la confianza en las aplicaciones de aprendizaje automático.

## Innovaciones recientes en inteligencia artificial

NVIDIA. (n.d.). *AI Computing*. NVIDIA. <https://www.nvidia.com/es-es/ai-data-science/>

NVIDIA describe las últimas innovaciones en el campo de la inteligencia artificial, incluyendo avances en hardware y software que permiten entrenar modelos más complejos y precisos. Se destacan aplicaciones en áreas como la conducción autónoma, la atención médica y la robótica.

## Visión futura de la inteligencia artificial

MIT Technology Review. (n.d.). *The Future of AI*. MIT Technology Review.  
<https://cdn.technologyreview.com/artificial-intelligence/>

Este artículo de MIT Technology Review ofrece una perspectiva sobre el futuro de la inteligencia artificial, explorando cómo la IA podría transformar diversas industrias en los próximos años. Se discuten tendencias como la IA explicable, la integración de IA en la toma de decisiones estratégicas y el impacto de la IA en la fuerza laboral global.

1. ¿Cuál es uno de los beneficios principales de los Data Marts en la inteligencia de negocio?

  - A. Almacenan grandes volúmenes de datos no estructurados.
  - B. Facilitan la segmentación y análisis de datos específicos de diferentes áreas de negocio.
  - C. Automatizan procesos financieros.
  - D. Mejoran la velocidad de las transacciones.
  
2. ¿Qué herramienta de Oracle BI permite a los usuarios crear visualizaciones interactivas?

  - A. Oracle BI Publisher.
  - B. Oracle BI Interactive Dashboards.
  - C. Oracle BI Answers.
  - D. Oracle Real-Time Decisions.
  
3. ¿Qué técnica analítica es esencial para anticipar cambios en el mercado según IBM?

  - A. Análisis descriptivo.
  - B. Análisis predictivo.
  - C. Análisis de componentes principales.
  - D. Análisis de cluster.
  
4. ¿Cuál es una limitación importante de la analítica de negocio mencionada por SAS?

  - A. Mejora la eficiencia operativa.
  - B. Facilita la toma de decisiones informadas.
  - C. Requiere competencias especializadas en análisis de datos.
  - D. Simplifica la implementación de soluciones.

5. ¿Qué describe mejor el proceso de preparación de datos en la minería de datos?
  - A. Almacenamiento de datos en un data warehouse
  - B. Limpieza y transformación de datos para análisis.
  - C. Visualización de datos en dashboards interactivos.
  - D. Automatización de procesos transaccionales.
6. ¿Cuál es una aplicación común de la minería de datos en el sector financiero?
  - A. Optimización de inventarios.
  - B. Detección de fraudes.
  - C. Personalización de campañas de marketing.
  - D. Análisis de la cadena de suministro.
7. ¿Qué herramienta de aprendizaje automático es conocida por su capacidad para trabajar con gráficos computacionales dinámicos?
  - A. Scikit-learn.
  - B. TensorFlow.
  - C. PyTorch.
  - D. Keras.
8. ¿Qué método de evaluación asegura que un modelo de aprendizaje automático no se ajuste demasiado a los datos de entrenamiento?
  - A. Regularización.
  - B. Análisis de componentes principales.
  - C. Validación cruzada.
  - D. Clustering.

- 9.** ¿Cuál es una aplicación de la inteligencia artificial en la atención al cliente?
- A. Optimización de procesos de manufactura.
  - B. Detección de enfermedades.
  - C. Chatbots y asistentes virtuales.
  - D. Análisis de imágenes médicas.
- 10.** ¿Qué avance en inteligencia artificial se espera que haga los modelos más transparentes y comprensibles para los humanos?
- A. IA general.
  - B. IA explicable (XAI).
  - C. Aprendizaje no supervisado.
  - D. Clustering.

Ciencia de Datos Aplicada

---

# Tema 8. Aplicaciones en Minería de Datos

# Índice

[Esquema](#)

[Ideas clave](#)

[8.1. Introducción y objetivos](#)

[8.2. Conceptos de Minería de datos](#)

[8.3. Objetivos de la Minería de Datos](#)

[8.4. Procesos de descubrimiento de conocimiento](#)

[8.5. Ejemplos de aplicaciones de Minería de Datos](#)

[8.6. Referencias bibliográficas](#)

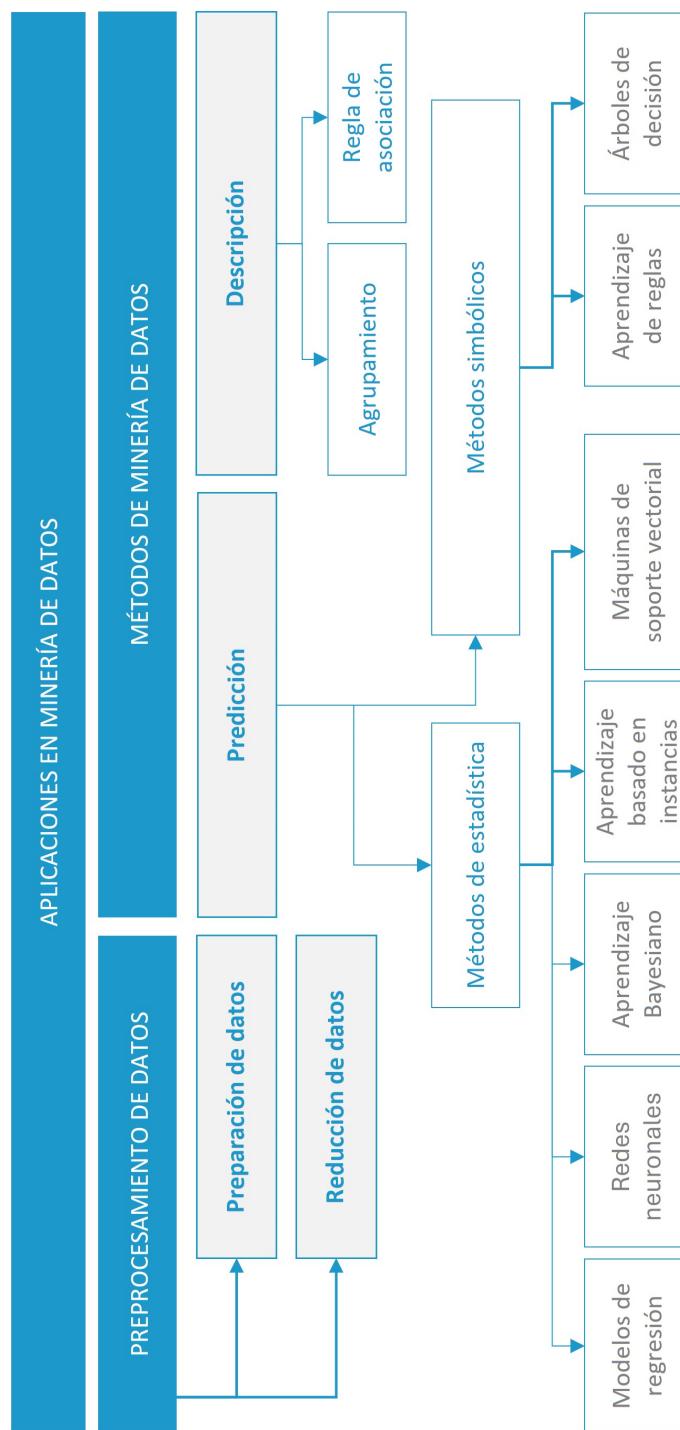
[A fondo](#)

[Avances históricos de la minería de datos](#)

[Herramientas y aspectos importantes de la minería de datos](#)

[Test](#)

# Esquema



## 8.1. Introducción y objetivos

Este tema tiene como propósito dar una visión más profunda del tratamiento de datos desde el contexto de la **minería de datos** (DM, por sus siglas en inglés *Data Mining*). En los primeros apartados se presenta una **introducción a DM**, donde se abordan los **conceptos básicos** y las **etapas necesarias** para su aplicación en la resolución de problemas. A continuación, se describen los **principales métodos usados** en DM divididos en dos grupos según su enfoque: **métodos de predicción, métodos estadísticos y simbólicos**. Más adelante, se presentan los **paradigmas de aprendizaje supervisado y no supervisado**, acompañados de una clara diferenciación entre ambos. Finalmente, se establecen dos **etapas primordiales** en el tratamiento de datos: **la preparación y la reducción de datos**.

Los **objetivos de aprendizaje** que se pretenden con este tema serán los siguientes:

- ▶ Conocer las características elementales de los principales métodos en minería de datos y su relación con los procedimientos de tratamiento de datos.
- ▶ Identificar los diferentes paradigmas de aprendizaje en minería de datos.
- ▶ Entender las diferentes etapas que se requieren en el tratamiento de datos.

## 8.2. Conceptos de Minería de datos

Durante los últimos años han sucedido cambios significativos en el proceso de la gestión de la información debido, principalmente, al **crecimiento exponencial** de los dispositivos que se interconectan a través de la **nube**. El **avance de la tecnología** ha permitido el desarrollo de sistemas que funcionan con un consumo mínimo, por lo que son capaces de capturar enormes cantidades de datos.

Dentro de los **tipos de datos** que más se generan se encuentran: los archivos de la web, las interacciones financieras, la interacciones entre usuarios y el Internet de las cosas (IoT) (Winn, 2020).

De esta manera, la DM surge con el propósito de describir los diferentes aspectos relacionados con la **exploración y descubrimiento de los datos**. Alrededor de este campo de estudio existe un amplio número de aplicaciones, formulaciones y representaciones de datos ajustados a problemas reales como detección de fraudes, previsión financiera y empresarial, diagnóstico médico, *marketing* dirigido y diseño de productos.

La **minería de datos** y el descubrimiento de **conocimiento en bases de datos** (KDD, por sus siglas en inglés *Knowledge Discovery in Databases*) se tratan con frecuencia como sinónimos. No obstante, la DM es solo una parte del proceso KDD, si bien es la porción central y más importante.

Un **aspecto clave** que caracteriza el proceso KDD es su desarrollo a través de etapas; para este caso se adopta un esquema híbrido que categoriza las **etapas en seis pasos** como se muestra en la Figura 1:

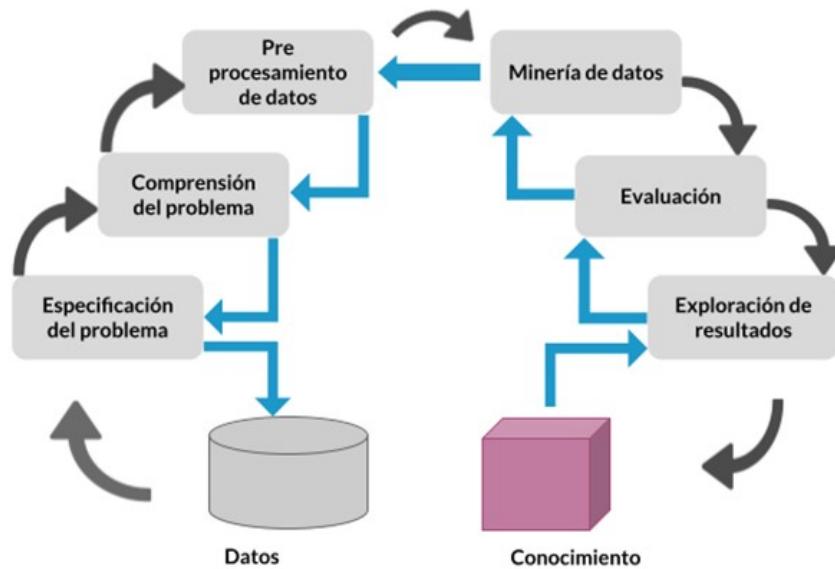


Figura 1. Proceso de KDD. Adaptado de: Preprocesamiento de datos en minería de datos (p.3), por García, S., Luengo, J. y Herrera, F. (2015). Fuente: <https://doi.org/10.1007/978-3-319-10247-4>

En la Figura 1 se muestra la minería de datos como un paso en el proceso de descubrimiento del conocimiento. El preprocesamiento está antes, dado que no existe un control de las estructuras de los datos y, generalmente, la información se encuentra en un formato que no es procesable. De allí la importancia de incorporar **etapas de preprocesamiento** donde los datos sean recopilados, limpiados y transformados en un **formato estandarizado** capaz de ser leído y procesado por un sistema de cómputo (Agarwal, 2014).

## 8.3. Objetivos de la Minería de Datos

La minería de datos surge de la intersección de varias áreas, como la estadística, la inteligencia artificial, el aprendizaje automático y las bases de datos. Su objetivo principal es extraer patrones, conocimientos y tendencias útiles a partir de grandes conjuntos de datos. Estos descubrimientos pueden ser utilizados para mejorar la toma de decisiones, optimizar procesos, identificar oportunidades y detectar problemas antes de que se conviertan en críticas. Los objetivos de la minería de datos están íntimamente relacionados con los métodos que se utilizan para alcanzarlos. Cada objetivo específico de la minería de datos requiere el uso de técnicas y algoritmos particulares que permitan extraer el conocimiento deseado de los datos.

Este apartado proporciona una breve revisión de **algunos métodos** seleccionados como los más frecuentes e importantes en DM. El propósito de esta revisión no es detallar ni explicar por completo cada una de las técnicas, sino destacar sus **principales características** y explicar resumidamente **cómo operan** manteniendo el enfoque en su relación con el preprocesamiento de datos.

En la Figura 2 se observan algunos métodos clasificados según el proceso de obtención de conocimiento: predicción y descripción.



Figura 2. Esquema general de métodos de DM. Adaptado de: Preprocesamiento de datos en minería de datos (p.4), por García, S., Luengo, J. y Herrera, F. (2015). Fuente: <https://doi.org/10.1007/978-3-319-10247-4>

## Métodos estadísticos

Los **métodos estadísticos** suelen caracterizarse por la representación del conocimiento a través de modelos matemáticos.

### Modelos de regresión

El **análisis de regresión lineal** se puede dividir en **regresión lineal simple y múltiple**. Ambos modelos se enfocan en la predicción de resultados continuos.

En general, para este tipo de análisis se cuenta con un número de variables explicativas y una variable de respuesta continua. El **objetivo** es encontrar una relación capaz de predecir el resultado deseado (Rong y Bao-wen, 2018).

Esta técnica se utiliza para **pronosticar, modelar series de tiempo y encontrar la relación del efecto causal entre las variables**. Dentro de los **beneficios** de utilizar el análisis de regresión se encuentran (Sunil, 2015):

- ▶ Identificación de relaciones significativas entre la variable dependiente y la variable independiente.
- ▶ Valoración de la fuerza del impacto de múltiples variables independientes sobre una variable dependiente.
- ▶ Comparación de los efectos de variables medidas en diferentes escalas.

## Redes neuronales artificiales (RNA)

Se definen como **sistemas de mapeos no lineales**, cuya estructura se inspira en el funcionamiento del sistema nervioso humano. Están compuestas por unidades denominadas neuronas y cuentan con ciertas características como entradas, salidas, conexiones con pesos y funciones de activación (Ponce Cruz, 2020).

Pueden realizarse **formulaciones complejas** dependiendo de sus conexiones, y dentro de los más utilizados se encuentran el **perceptrón multicapa** (MLP, del inglés *Multi Layer Perceptron*).

Las **capas de entrada** dependen de la información disponible, mientras que en las capas de salida el número de nodos es igual a la cantidad de clases. Las neuronas de una capa se conectan con las de la capa siguiente mediante **sinapsis**, determinados a través del proceso de entrenamiento (Villada et al., 2016).

**Otras redes** también son usadas en muchas aplicaciones: redes de función de base radial (RBFN, del inglés *Radial Basis Function Networks*), cuantificación de vectores de aprendizaje (LVQ) y mapas de autoorganización (SOM) (Kwon, 2011).

## Aprendizaje bayesiano

Estas son redes que emplean la **teoría de probabilidad bayesiana** para representar en un grafo dirigido acíclico las dependencias entre sus variables aleatorias.

El método bayesiano más aplicado es **Naive Bayes**, un algoritmo que funciona con atributos categóricos, dado que el cálculo de la probabilidad solo puede realizarse en dominios discretos. Además, el **supuesto de independencia** entre atributos hace que estos métodos sean sensibles a la redundancia.

También existen modelos complejos basados en estructuras de dependencia como las redes bayesianas (Barber, 2010).

Finalmente, el aprendizaje bayesiano asume que **los atributos son condicionalmente independientes**, dada la clase que obtiene mayor probabilidad. Para calcular estas probabilidades, se utiliza la **matriz de confusión** (Titterington et al., 1981).

## Aprendizaje basado en instancias

Este tipo de aprendizaje es usado **para predecir qué instancias de una base de datos se encuentran más cerca de una nueva instancia**.

El aprendizaje basado en instancias a menudo se denomina *Lazzy Learning*, ya que no hay una «transformación» ni entrenamiento de las instancias.

Los **métodos pueden diferir** en relación con: la métrica de distancia utilizada, el número de instancias usadas, los mecanismos de ponderación de votos y el uso de algoritmos eficientes para encontrar las instancias más cercanas (tales como KD-Tree, Ball-Tree y Brute-Force).

El **método de los k vecinos más cercanos** (k-NN, por sus siglas en inglés *k-Nearest Neighbour*) es una de las técnicas más aplicadas y útiles, principalmente, cuando los valores de los atributos son continuos. La idea es estimar la clase basado en la clasificación de las instancias más cercanas. Dado que el conjunto de entrenamiento es obtenido en tiempo de ejecución y cambia con cada nueva clasificación, este método es considerado no paramétrico.

Algunos de los principales **inconvenientes** del método son: altos requisitos de almacenamiento, baja eficiencia en la respuesta de predicción y la precisión puede verse afectada por el ruido (Cunningham y Delany, 2007).

## Máquinas de soporte vectorial

Las **máquinas de vectores de soporte** (SVM, del inglés *Support Vector Machine*) componen uno de los métodos de aprendizaje supervisado perteneciente a la familia de los clasificadores lineales. El propósito de estos algoritmos de aprendizaje es **transformar un conjunto de datos de una dimensión n hacia un espacio de dimensión superior aplicando una función kernel** (Sánchez Anzola, 2016). Además, es considerado como una extensión del **perceptrón**; la diferencia radica en que el algoritmo del perceptrón busca minimizar los errores de clasificación, mientras que en SVM el objetivo de optimización es maximizar el margen (Martínez Heras, 2020).

De manera explicativa, se puede decir que SVM traza una coordenada de cada dato en un espacio de n características. Por tanto, la **finalidad** es establecer un separador que divida los datos en categorías justo en la frontera del hiperplano (Mirjalili y Raschka, 2019).

## Métodos simbólicos

Los **métodos simbólicos** prefieren representar el conocimiento mediante **símbolos y conectivos**, produciendo modelos más interpretables para los humanos.

## Aprendizaje de reglas

Los **algoritmos de reglas** tienen como operación principal encontrar una regla que explique el conjunto de datos. Estas reglas son una de las maneras de formalizar la representación del conocimiento, puesto que su expresividad simbólica es considerada más comprensible y natural para los humanos que otros formalismos (Filiberto et al., 2011; Stefanowski y Wilk, 2009).

Hay muchas **formas de interpretar** las reglas obtenidas y utilizarlas en el mecanismo de inferencia. Desde el punto de vista del preprocesamiento de datos, los algoritmos de reglas requieren datos nominales o discretizados y disponen de un selector innato de atributos de los datos. Algunos ejemplos de estos algoritmos son el AQ, CN2, RIPPER, PART y FURIA.

## Árbol de decisiones

Los **árboles de decisión** consisten en trazar todos los caminos posibles, considerando la importancia de cada atributo y utilizando particiones recursivas para clasificar los datos.

El árbol de decisión proporciona una **herramienta predictiva** que ayuda a determinar qué atributos tienen mayor incidencia para dividir los datos en función de dichos atributos.

El árbol se compone de ramas y hojas; los **nodos intermedios** (las ramas) representan soluciones, mientras que los **nodos finales** (las hojas) dan la predicción que se requiere.

Los árboles de **decisión** pueden usarse para resolver problemas tanto de **clasificación como de regresión**. Algunos algoritmos, como el *Top Down Induction Tree*, utilizan como heurística principal buscar el mejor atributo y ubicarlo en la raíz del árbol para hacer uso de un estadístico llamado mayor ganancia de información (Santa Chávez et al., 2013).

## Clustering

El problema del agrupamiento de datos ha sido ampliamente estudiado en la DM debido a sus numerosas aplicaciones. En ausencia de información etiquetada específica, se puede considerar el agrupamiento de datos o Clustering. Aquí, el **objetivo** es agrupar de forma natural los datos, de manera que los elementos comparten algunas propiedades o similitudes.

Los algoritmos calculan una medida de distancia multivariante entre las observaciones que se encuentran estrechamente relacionadas (Jain, 2010).

En términos generales, se puede hablar de **cinco tipos de métodos de agrupación**: agrupamiento particional, agrupamiento jerárquico, agrupamiento difuso, agrupamiento basado en densidad y agrupamiento basado en modelos.

Uno de los algoritmos más populares de agrupamiento se denomina ***k-means***, el cual pertenece a la categoría de agrupamiento particional. Este algoritmo comienza con un número  $k$  de observaciones del conjunto de datos y los utiliza como centroides iniciales en cada iteración, intentando reducir el valor de la suma de los errores al cuadrado. Una métrica para indicar cuán bien los centroides representa a los miembros de su grupo es la suma de los errores al cuadrado.

## Reglas de asociación

En MD, a veces se utiliza un conjunto de datos para **descubrir patrones y relaciones** que existen entre los atributos, generalmente en forma de reglas conocidas como **reglas de asociación**. Hay muchas reglas de asociación posibles derivadas de cualquier conjunto de datos, lo cual permite establecer relaciones de coocurrencia y no causalidad entre variables cualitativas (Martínez, 2018). Dentro de las principales **ventajas** se encuentran:

- ▶ La posibilidad de asociaciones entre cualquiera de los atributos.
- ▶ La obtención de muchas conclusiones al tener muchas reglas.
- ▶ Comprobación y fiabilidad altas, dado el número de restricciones al que está sujeto.

## 8.4. Procesos de descubrimiento de conocimiento

El proceso de descubrimiento de conocimiento en bases de datos (KDD, por sus siglas en inglés) y los paradigmas de aprendizaje en la minería de datos están estrechamente relacionados. El KDD es un proceso estructurado que incluye varias etapas, y en cada una de estas etapas se aplican diferentes paradigmas de aprendizaje para extraer conocimiento valioso de los datos.

En este apartado se describirán brevemente tres **tipos de aprendizaje: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje reforzado** (ver Figura 3). Además, se establecerán algunas **pautas** que permiten diferenciar los tres tipos distintos de aprendizaje.



Figura 3. Tipos de aprendizaje en DM. Adaptado de: Python *machine learning*. Aprendizaje automático y aprendizaje profundo con Python, *scikit-learn* y TensorFlow, por Raschka, S., y Mirjalili, V. (2019). Fuente: <https://www.ecobook.com/libros/python-machine-learning-aprendizaje-automatico-y-aprendizaje-profundo-con-python-scikit-learn-y-tensorflow/9788426727206/>

### Aprendizaje supervisado

El **propósito del aprendizaje supervisado** es aprender una estructura llamada modelo a partir de un conjunto de datos etiquetados con el fin de realizar

predicciones sobre nuevos datos. Generalmente, un modelo describe y explica experiencias que están ocultas en los datos y que pueden ser utilizadas en la predicción del valor del atributo destino, si se conocen los valores de los atributos de entrada.

Típicamente, los **atributos** pueden ser nominales, categóricos o numéricos. Los dos problemas básicos y clásicos que pertenecen a la categoría de aprendizaje supervisado son **clasificación** y **regresión**.

La **clasificación** es una de las aplicaciones más comunes para la minería de datos y corresponde a una tarea que ocurre con frecuencia en la vida cotidiana, la de discriminar **instancias dentro de un conjunto de datos de diferentes tipos de clases**. Finalmente, una vez tenemos un modelo que se ajusta a las instancias de entrenamiento, podemos hacer **predicciones fiables** para nuevas instancias.

Para entender mejor estos conceptos, echa un vistazo al **ejemplo 1** que se muestra a continuación:

### Ejemplo 1. Filtro de correo no deseado

Considera el desarrollo de un correo no deseado que podemos entrenar con algún algoritmo de aprendizaje supervisado. Los datos de entrenamiento corresponden al cuerpo de correos electrónicos etiquetados con dos clases: correo no deseado, correo deseado. Esta situación corresponde a un ejemplo de clasificación binaria, donde el algoritmo aprende un conjunto de reglas y aprende a distinguir entre dos posibles clases.

Existe otro **problema** alrededor de la DM denominado «**regresión**» que, en ciertas ocasiones, puede llegar a presentar más dificultades que el problema de clasificación porque los recursos de cálculo necesarios y la complejidad del modelo son mayores.

A través del ejemplo 2, se muestra un **problema de regresión**:

## Ejemplo 2. Resultados de un examen

Supón que se quiere predecir los resultados del examen del curso de tratamiento de datos. Conocemos datos previos de simulaciones que han hecho los alumnos con sus respectivos resultados. Es así como se podría aprovechar esta información como datos de entrenamiento para generar un modelo que prediga los resultados del examen final de la asignatura.

Se podría afirmar que el proceso de regresión puede volverse más complejo, dado el número de atributos de un conjunto de datos, y no siempre es tan sencillo de resolver con una regresión lineal simple, como se plasmó en el ejercicio 2.

Existe **otro tipo de aprendizaje supervisado** que involucra series de tiempo y se ocupa de hacer predicciones en el tiempo. Las aplicaciones típicas de este aprendizaje incluyen: **el análisis de precios de acciones, tendencias del mercado y previsión de ventas**.

## Aprendizaje no supervisado

Anteriormente se vio cómo el **objetivo del aprendizaje supervisado** es utilizar las **etiquetas** de las clases proporcionadas por un experto para entrenar un modelo capaz de **predecir nuevos datos**. Por el contrario, en el **aprendizaje no supervisado** no existe tal supervisor y solo los **datos de entrada** están disponibles. Así, el objetivo ahora es **encontrar regularidades, irregularidades, relaciones, similitudes y asociaciones en la entrada** (ver ejemplo 3).

## Ejemplo 3. Segmentación de clientes

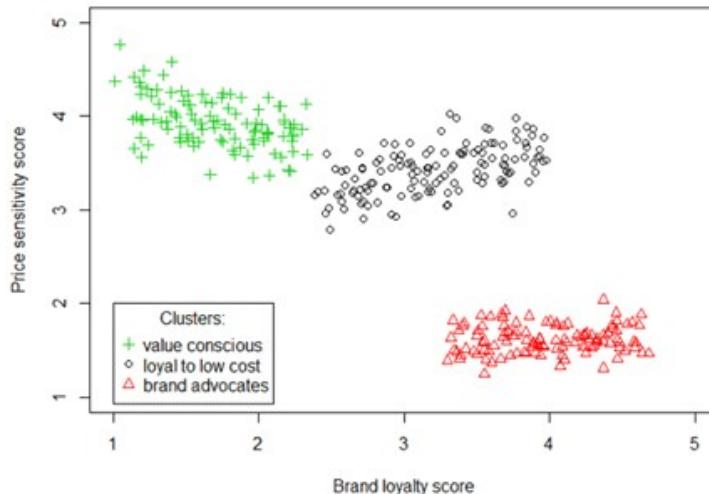


Figura 4. Segmentación por grupos.

Supón que tienes una tienda en línea y deseas segmentar los clientes por características similares para poder ofrecer productos y servicios de acuerdo con sus preferencias. Se han seleccionado dos variables: lealtad de la marca y sensibilidad al precio. Una vez ejecutado el algoritmo, se identifican tres grupos: preocupados por el precio (Verde), leales a los precios bajos (Negro) y defensores de la marca (Rojo). Por supuesto, este algoritmo no es capaz de obtener este tipo de inferencias, dado que solo arroja los grupos. Un experto en el tema puede llegar a dar ese tipo de explicaciones una vez analice cada grupo.

## Aprendizaje reforzado

En el aprendizaje reforzado el objetivo es **desarrollar un sistema que mejore su rendimiento con base en las señales del entorno**. Adicional al estado actual del

entorno, se requiere una **señal de recompensa**. A diferencia del aprendizaje supervisado, el *feedback* no es el valor de la etiqueta, sino una métrica obtenida por una función de recompensa. Mediante la **interacción con el entorno**, un sistema puede aprender una serie de acciones que maximicen esta recompensa, utilizando un enfoque experimental.

## Preprocesamiento de los datos

Una vez revisados algunos conceptos básicos de DM, se puede intuir que, si los datos de entrada son incorrectos, se obtendrán modelos igualmente incorrectos.

El **conjunto de datos** debe proporcionarse en la **cantidad, estructura y formato** que se adapte a la actividad a desarrollar.

Las bases de datos en el mundo real son muy diferentes a lo que pensamos. Se encuentran altamente influenciadas por **factores negativos** como ruido, atributos perdidos, datos inconsistentes o un tamaño grande de la información. Estas características, sin duda, afectarán a la calidad de los datos y, por supuesto, al rendimiento del modelo de DM.

A continuación, se proporciona un breve resumen de las **categorías en las que se divide el conjunto de técnicas de preprocesamiento de datos**.

## Preparación de los datos

En este paso de preprocesamiento, los datos se convierten o consolidan para que el resultado del **proceso de minería** pueda aplicarse y sea más eficiente (ver Figura 4).

Dentro de las **técnicas** se encuentran: el suavizado, la construcción de características, la agregación o resumen de datos, la normalización, la discretización y la generalización. La mayoría de ellas son tareas independientes.

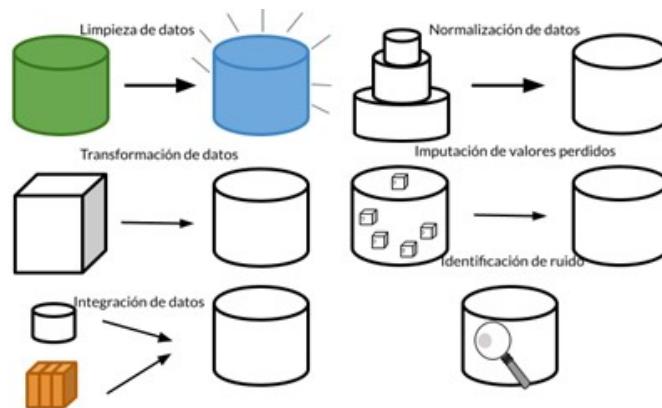


Figura 4. Etapas para la preparación de los datos. Adaptado de: *Preprocesamiento de datos en minería de datos* (p.12), por García, S., Luengo, J. y Herrera, F. (2015). Fuente: <https://doi.org/10.1007/978-3-319-10247-4>

Etapas de preparación de los datos		
1	Limpieza de datos	La limpieza de datos es un concepto general que comprende o se superpone a otras técnicas de preparación de datos muy conocidas, que incluyen el tratamiento de datos perdidos, datos con ruido, detección de discrepancias y datos sucios.
2	Transformación de datos	Dentro de las técnicas se encuentran: el suavizado, la construcción de características, la agregación o resumen de datos, la normalización, la discretización y la generalización.
3	Integración de datos	Esta etapa se encarga de la fusión de datos de múltiples almacenes de datos, cuyas operaciones típicas son la identificación, unificación de variables, el análisis de correlación de atributos, la duplicidad de los datos y la detección de conflictos.
4	Normalización de datos	Todos los atributos deben expresarse en las mismas unidades de medida y deben utilizar una escala o rango estándar. La normalización de los datos intenta dar a todos los atributos el mismo peso con el fin de mejorar el rendimiento de algunos métodos.
5	Imputación de valores perdidos	Es una forma de limpieza de datos, donde la finalidad es llenar las variables que contienen datos vacíos con una estimación razonable.
6	Identificación de ruido	Se conoce como suavizado en la transformación de datos, cuyo objetivo es detectar errores aleatorios o variaciones en una medida variable; proceso basado en la corrección que podría implicar algún tipo de operación.

Tabla 1. Etapa de preparación de los datos.

## Reducción de los datos

La **reducción de datos** comprende el conjunto de técnicas empleadas para obtener una representación reducida de los datos originales. Mientras que la preparación acomoda los datos para que sirvan como entrada en una tarea de minería, la

reducción de datos generalmente **mantiene la estructura esencial y la integridad de los datos originales**, pero disminuye la cantidad de datos.

Los **algoritmos** tienen una cierta complejidad temporal que depende de varios parámetros.

En minería de datos, uno de estos **parámetros** es directamente proporcional al tamaño de la base de datos de entrada. Si el tamaño excede el límite, la ejecución del algoritmo puede resultar prohibitiva. Entonces, la tarea de reducción de datos es tan crucial como la preparación de datos.

Respecto a otros factores, como la disminución de la complejidad y la mejora de la calidad de los modelos arrojados, el papel de la reducción de datos vuelve a ser determinante. Las **cuestiones básicas** que deben resolverse en la reducción de datos se presentan en la siguiente tabla:

Etapas para la reducción de los datos		
1	Selección de características	Logra la reducción del conjunto de datos eliminando características (o dimensiones) irrelevantes o redundantes. Su objetivo es encontrar un conjunto mínimo de atributos. Por ejemplo, lograr que la distribución de probabilidad resultante de los atributos de salida de datos (o clases) sea lo más cercana posible a la distribución original con todos los atributos. Facilita la comprensión del patrón extraído y aumenta la velocidad de la etapa de aprendizaje.
2	Selección de instancias	Consiste en elegir un subconjunto de la totalidad de datos disponibles para aplicar la minería como si se hubieran utilizado todos los datos. Constituye la familia de métodos que realizan de manera inteligente la elección del mejor subconjunto de ejemplos posible de los datos originales, usando algunas reglas o heurísticas. La selección aleatoria de ejemplos generalmente se conoce como muestreo y está presente en una gran cantidad de modelos de minería de datos para realizar validaciones internas y evitar el sobreajuste.
3	Discretización	Este procedimiento transforma datos cuantitativos en cualitativos, es decir, atributos numéricos en atributos discretos o nominales con un número finito de intervalos, obteniendo una partición no superpuesta de un dominio continuo. Así, se establece una asociación entre cada intervalo con un valor discreto numérico. Una vez realizada la discretización, los datos pueden tratarse como datos nominales durante cualquier proceso de minería de datos. La discretización es en realidad una técnica híbrida de preprocesamiento de datos que implica tanto la preparación de datos como las tareas de reducción de datos. La discretización puede verse como un método de reducción de datos, ya que mapea datos de un amplio espectro de valores numéricos a un subconjunto muy reducido de valores discretos.
4	Extracción de características/generación de instancias	Extiende tanto la selección de características como la selección de instancias al permitir la modificación de los valores internos que representan cada ejemplo o atributo. En la extracción de características, además de la eliminación de atributos, los subconjuntos de atributos pueden fusionarse o pueden contribuir a la creación de atributos sustitutos artificiales. En cuanto a la generación de instancias, el proceso es similar en cuanto a los ejemplos. Permite la creación o el ajuste de ejemplos sustitutos artificiales que podrían representar mejor los límites de decisión en el aprendizaje supervisado.

Tabla 2. Etapas de reducción de los datos.

## 8.5. Ejemplos de aplicaciones de Minería de Datos

La minería de datos tiene aplicaciones significativas en diversos sectores, proporcionando valor mediante la mejora de la toma de decisiones, la detección de patrones, la optimización de procesos y la mejora de la experiencia del cliente. A continuación, se presenta un ejemplo detallado de la aplicación de la minería de datos en el sector de la salud, ilustrando los pasos, riesgos y características de la planificación involucrada.

### Aplicación de Minería de Datos en la Salud: Detección de Fraude en Seguros Médicos

La detección de fraude en seguros médicos es una aplicación crítica de la minería de datos en el sector de la salud. El fraude en las reclamaciones de seguros médicos puede llevar a pérdidas significativas. Utilizando técnicas de minería de datos, las organizaciones de salud pueden identificar patrones sospechosos y prevenir reclamaciones fraudulentas.

#### Pasos del Proceso

##### Selección y Adquisición de Datos:

- ▶ Recopilación de datos de diversas fuentes, como registros electrónicos de salud (EHR), sistemas de gestión hospitalaria, y bases de datos de reclamaciones de seguros.
- ▶ Garantizar la precisión y representatividad de los datos seleccionados.

## Preprocesamiento y Transformación de Datos:

- ▶ Limpieza de datos para eliminar duplicados, corregir errores y manejar valores faltantes.
- ▶ Normalización y transformación de los datos para adecuarlos a los algoritmos de minería de datos.

## Minería de Datos:

- ▶ Aplicación de algoritmos de detección de anomalías y análisis de patrones para identificar comportamientos sospechosos.
- ▶ Técnicas como el análisis de asociación y la clasificación ayudan a descubrir relaciones inusuales entre las reclamaciones y detectar posibles fraudes.

## Evaluación e Interpretación de Resultados:

- ▶ Validación de los modelos utilizando conjuntos de datos de prueba y técnicas de validación cruzada.
- ▶ Interpretación de los patrones descubiertos y formulación de insights valiosos para la toma de decisiones.

## Implementación y Monitoreo:

- ▶ Implementación de los modelos en sistemas operativos para la detección en tiempo real.
- ▶ Monitoreo continuo y actualización de los modelos para adaptarse a nuevos datos y patrones emergentes.

## Riesgos

### Datos Incompletos o Sesgados:

- ▶ La calidad y representatividad de los datos son cruciales. Datos incompletos o sesgados pueden llevar a resultados incorrectos.

### Falsos Positivos/Negativos:

- ▶ Los modelos pueden generar falsos positivos (identificación incorrecta de fraude) o falsos negativos (no detectar fraude real), lo que puede afectar la eficiencia y credibilidad del sistema.

### Privacidad y Seguridad:

- ▶ El manejo de datos de salud sensibles requiere estrictas medidas de seguridad y cumplimiento de regulaciones de privacidad como HIPAA.

## Características de la Planificación

### Definición Clara de Objetivos:

- ▶ Establecer objetivos claros y medibles, como la reducción de reclamaciones fraudulentas en un porcentaje específico.

### Selección de Técnicas y Herramientas Adecuadas:

- ▶ Elegir técnicas de minería de datos y herramientas que mejor se adapten a los datos y objetivos específicos.

### Evaluación Continua:

- ▶ Implementar un ciclo iterativo de evaluación y mejora continua para adaptar los modelos a nuevos patrones de fraude y mejorar su precisión.

## Capacitación y Colaboración:

- ▶ Involucrar a expertos en datos, profesionales de la salud y personal de TI en el proceso para asegurar una implementación exitosa y una comprensión completa de los resultados.

## 8.6. Referencias bibliográficas

Agarwal, S. (2014). *Data mining: Data mining concepts and techniques*. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement*, ICMIRA 2013. <https://doi.org/10.1109/ICMIRA.2013.45>

Barber, D. (2010). *David Barber - Machine Learning and Bayesian Reasoning*.

Cunningham, P., y Delany, S. J. (2007). K -Nearest Neighbour Classifiers. *Multiple Classifier Systems*, May, pp. 1-17.

Filiberto, Y., Bello, R., y Caballero, Y. (2011). En la teoría de los conjuntos aproximados extendida algorithm to learn clasification rules based on the extended rou ... January.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), pp. 651-666. <https://doi.org/10.1016/j.patrec.2009.09.011>

Kwon, S. J. (2011). *Artificial neural networks*. In *Artificial Neural Networks* (Issue July 2006). <https://doi.org/10.15864/jmscm.1104>

Martínez, C. D. J. (2018). *Reglas de Asociación*. [https://rstudio-pubs-static.s3.amazonaws.com/367334\\_353f1bbf1b3543e180bb9210e711a73f.html](https://rstudio-pubs-static.s3.amazonaws.com/367334_353f1bbf1b3543e180bb9210e711a73f.html)

Martinez Heras, J. (2020). *K-Means con ejemplos en Python*. <https://www.iartificial.net/clustering-agrupamiento-kmeans-ejemplos-en-python/>

Ponce Cruz, P. (2020). *Inteligencia Artificial*. Alphaomega.

Rong, S., y Bao-wen, Z. (2018). The research of regression model in machine learning field. *MATEC Web of Conferences*, 176, 01033. <https://doi.org/10.1051/matecconf/201817601033>

Sánchez Anzola, N. (2016). Máquinas de soporte vectorial y redes neuronales artificiales en la predicción del movimiento USD/COP spot intradiario. *ODEON*, 9, 113. <https://doi.org/10.18601/17941113.n9.04>

Santa Chávez, J., Veloza Mora, J., y Arias Montoya, R. (2013). Aplicación del aprendizaje automático con árboles de decisión al estudio de las variables del modelo de indicadores de gestión de las universidades públicas. *Scientia Et Technica*, 18(4), pp. 725-731. <https://doi.org/10.22517/23447214.8841>

Stefanowski J. y Wilk S. (2009) Ampliación de clasificadores basados en reglas para mejorar el reconocimiento de clases desequilibradas. En: Ras ZW, Dardzinska A. (eds.) *Avances en la gestión de datos. Estudios en Inteligencia Computacional*, vol. 223. Springer. Heidelberg. [https://doi.org/10.1007/978-3-642-02190-9\\_7](https://doi.org/10.1007/978-3-642-02190-9_7)

Sunil, R. (2015). 7 Types of Regression Techniques you should know. *Analytics Vidhya*, 14, pp 1-25. <https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>

Titterington, D. M., Murray, G. D., Murray, L. S., et al. (1981). Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients. *Journal of the Royal Statistical Society. Series A (General)*, 144(2), 145. <https://doi.org/10.2307/2981918>

Villada, F., Muñoz, N., y García-Quintero, E. (2016). Redes Neuronales Artificiales aplicadas a la Predicción del Precio del Oro. *Información Tecnológica*, 27(5), pp. 143-150. <https://doi.org/10.4067/S0718-07642016000500016>

Winn, Z. (2020). *The factory of the future, batteries not included*. MIT News Office. <https://news.mit.edu/2020/everactive-sensors-0820>

## Avances históricos de la minería de datos

Riquelme, J. C., Ruiz, R. y Gilbert, K. (2006). Minería de datos: conceptos y tendencias. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 10(29), pp. 11-18. <https://www.redalyc.org/articulo.oa?id=92502902>

Este recurso abarca información acerca de aspectos relevantes e históricos sobre el papel de la minería de datos en el mundo tecnológico y digital, de importancia para la contextualización de la temática por parte del estudiante; además, en este artículo podrás profundizar sobre aspectos tratados en este tema.

## Herramientas y aspectos importantes de la minería de datos

Rodríguez Suárez, Y., y Díaz Amador, A. (2009). Herramientas de Minería de Datos. *Revista Cubana de Ciencias Informáticas*, 3(3-4), pp. 73-80.  
<https://www.redalyc.org/articulo.oa?id=378343637009>

Este recurso consta de un artículo científico donde se trata información importante acerca de las principales herramientas de minería de datos, su desarrollo, definiciones claves, aplicaciones, así como algoritmos y técnicas útiles para complementar el contenido estudiado en este tema.

- 1.** ¿Cuáles son los métodos usados en DM para resolución de problemas?
  - A. Métodos de clasificación, estadísticos, simbólicos y de grafos.
  - B. Métodos de predicción, estadísticos y simbólicos.
  - C. Aprendizaje supervisado y no supervisado.
  - D. Aprendizaje de máquina y aprendizaje profundo.
  
- 2.** Son aplicaciones de la minería de datos:
  - A. Detección de fraudes.
  - B. Previsión financiera y empresarial.
  - C. Diagnóstico médico.
  - D. Todas las respuestas anteriores son correctas.
  
- 3.** No es un método estadístico dentro de la minería de datos:
  - A. Modelos de regresión.
  - B. Redes neuronales.
  - C. Árbol de decisión.
  - D. Máquina de soporte vectorial.
  
- 4.** Es un algoritmo que utiliza el aprendizaje basado en instancias:
  - A.  $K$  vecinos más cercanos o KNN.
  - B. Árboles de decisión.
  - C. Máquina de vectores de soporte.
  - D. Random Forest.

5. Desde la obtención de los datos hasta generar conocimiento se deben seguir las siguientes etapas:
  - A. Todas son válidas.
  - B. Especificación y comprensión del problema.
  - C. Preprocesamiento de los datos.
  - D. Minería de datos, evaluación y exploración de resultados.
6. Es una afirmación falsa sobre las máquinas de soporte vectorial:
  - A. Pertenece a la familia de clasificadores lineales.
  - B. Transforma el conjunto de datos de una dimensión  $n$  a un espacio de dimensión superior.
  - C. Es considerado una extensión del perceptrón.
  - D. Su objetivo es minimizar los errores.
7. Es un método de agrupamiento:
  - A. Agrupamiento jerárquico.
  - B. Agrupamiento denso.
  - C. Agrupamiento central.
  - D. Agrupamiento no denso.
8. Es una ventaja de las reglas de asociación:
  - A. Posibilidad de asociación.
  - B. Obtención de muchas conclusiones al poseer muchas reglas.
  - C. Fiabilidad alta.
  - D. Todas las respuestas anteriores son correctas.

- 9.** Es un tipo de aprendizaje basado en recompensas, aprende de una serie de acciones y lleva a cabo un proceso de decisión.
- A. Aprendizaje supervisado.
  - B. Aprendizaje no supervisado.
  - C. Aprendizaje reforzado.
  - D. Todas las respuestas anteriores son correctas.
- 10.** Es característico de un problema de regresión:
- A. Es típico en aplicaciones que incluyen análisis de precios, tendencias y previsión.
  - B. Discriminación de instancias.
  - C. Discriminación entre una o varias clases categóricas.
  - D. Ninguna de las respuestas anteriores es correcta.

Ciencia de Datos Aplicada

---

# Tema 9. Aplicación en Aprendizaje Automático

# Índice

[Esquema](#)

[Ideas clave](#)

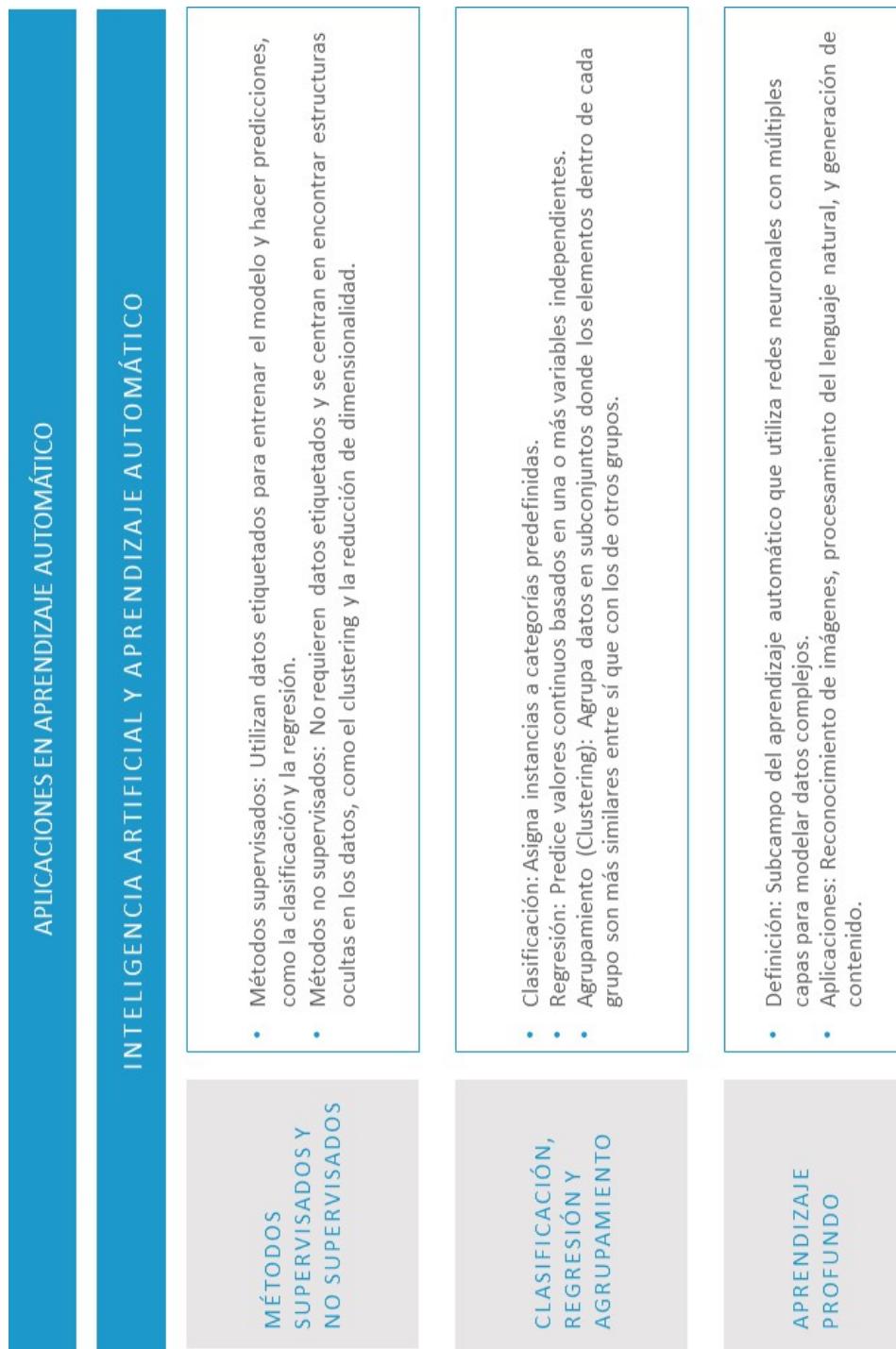
- [9.1. Introducción y objetivos](#)
- [9.2. Objetivos del Aprendizaje Automático](#)
- [9.3. Métodos supervisados y no supervisados](#)
- [9.4. Clasificación, regresión y agrupamiento](#)
- [9.5. Aprendizaje Profundo](#)
- [9.6. Ejemplos de aplicaciones de aprendizaje automático](#)
- [9.7. Referencias bibliográficas](#)

[A fondo](#)

[Introducción al aprendizaje automático](#)

[El aprendizaje automático dentro de la inteligencia artificial](#)

[Test](#)



## 9.1. Introducción y objetivos

Este tema introduce los conceptos generales del aprendizaje automático. Comenzamos describiendo de manera general algunas de sus **técnicas para extraer conocimiento**.

Este tema muestra, además, la importancia de estos campos y su gran diversidad de aplicaciones. Explicaremos conceptos fundamentales del **aprendizaje automático**: definición de aprendizaje, elementos que intervienen en una tarea de esta índole, tipos de aprendizaje, etc. Una vez abordados los conceptos básicos que envuelven esta rama de la inteligencia artificial, se desgrana la clasificación de los tipos de aprendizaje automático.

A continuación, se definen los objetivos principales de este tema.

- ▶ Introducir a los estudiantes al aprendizaje automático.
- ▶ Conocer su origen.
- ▶ Poder identificar las fases de extracción de conocimiento.
- ▶ Entender la clasificación general de aprendizaje automático.

## 9.2. Objetivos del Aprendizaje Automático

Este es la rama de la inteligencia artificial referida a la construcción de programas computacionales que automáticamente mejoran su rendimiento en una tarea determinada basándose en su experiencia. Habitualmente se nombra por su término en inglés, *machine learning* (ML).

El aprendizaje automático consiste en extraer conocimiento **a partir de los datos**. Se trata de un campo de investigación en el que convergen estadística, inteligencia artificial, ingeniería informática y lo que se conoce como análisis o aprendizaje predictivo.

Las aplicaciones de aprendizaje automático en los últimos años han pasado a formar parte de nuestro día a día. Desde sistemas de recomendación de películas o música, hasta diferentes tipos de comida, o bien reconocer en las fotografías de nuestros dispositivos móviles a nuestros amigos.

A lo largo de los años, la financiación de las iniciativas de IA ha pasado por una serie de ciclos activos e inactivos. El término «invierno» se utiliza para describir los períodos de inactividad, cuando el interés de los clientes por la IA disminuye. Tras el «invierno de la IA» de los años 80 y 90, el interés por la aplicación de técnicas no ha dejado de aumentar en varios campos de la ingeniería, como el análisis del habla y de las imágenes (Hinton et al., 2012), así como también de las comunicaciones (Ibnkahla, 2000).

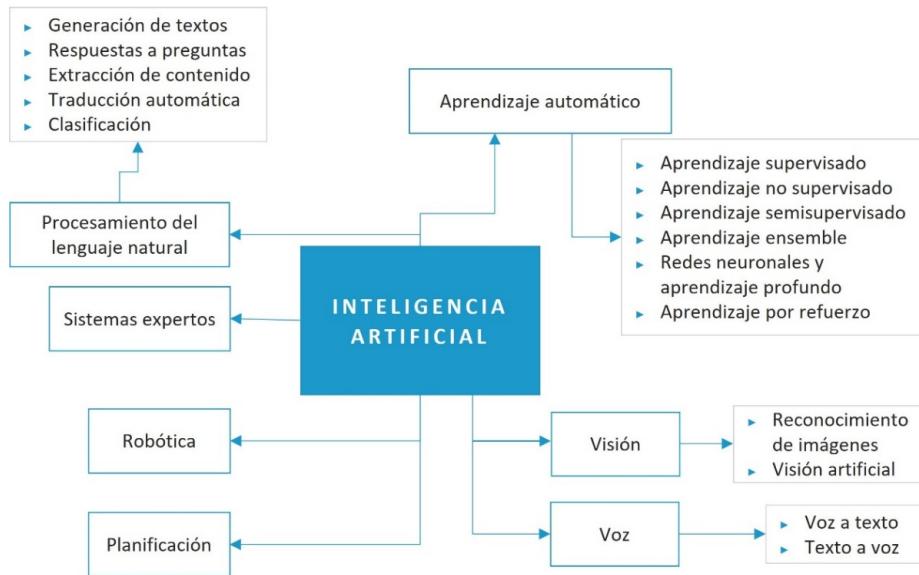


Figura 1. Taxonomía del aprendizaje automático dentro de la inteligencia artificial. Fuente: adaptado de Panesar, 2019.

A diferencia de los sistemas expertos basados en la lógica, que predominaban en los primeros trabajos sobre IA (Nilsson, 1986), la renovada confianza en los métodos basados en datos está motivada por los éxitos de las herramientas de **reconocimiento de patrones** basadas en el aprendizaje automático.

## 9.3. Métodos supervisados y no supervisados

Los tipos más famosos de este aprendizaje son aquellos que permiten **automatizar la toma de decisiones** a partir de generalizaciones desde ejemplos aprendidos. En este grupo entra lo que se conoce como **aprendizaje supervisado**, en el que un usuario provee al algoritmo con pares de entradas y salidas deseadas, y este se encarga de encontrar la forma de producir el resultado de salida esperado dados unos datos de entrada. En concreto, el algoritmo es capaz de crear unos datos de salida a partir de unos datos de entrada que no ha visto nunca y sin la intervención humana.

Uno de los casos más utilizados para la comprensión del aprendizaje supervisado sería el ejemplo de la bandeja de entrada de un correo, en el que se da como datos de entrada los diferentes *mails* con información correspondiente, así como los *mails* que han sido identificados como *spam*. Con la entrada de un nuevo correo, el algoritmo es capaz de identificar si es *spam* o no.

Como se ha visto en temas anteriores, la **minería de datos** utiliza técnicas de aprendizaje automático para, por ejemplo, detectar transiciones fraudulentas en tarjetas de crédito. El aprendizaje automático tiene múltiples aplicaciones en otro tipo de sistemas como en aquellos relacionados con la robótica, o en sistemas de reconocimiento del habla (Rogers,2017). En la Figura 1 podemos ver las diferentes ramas de la inteligencia artificial y el espacio que ocupa en dicha taxonomía el aprendizaje automático, así como sus diferentes subramas.

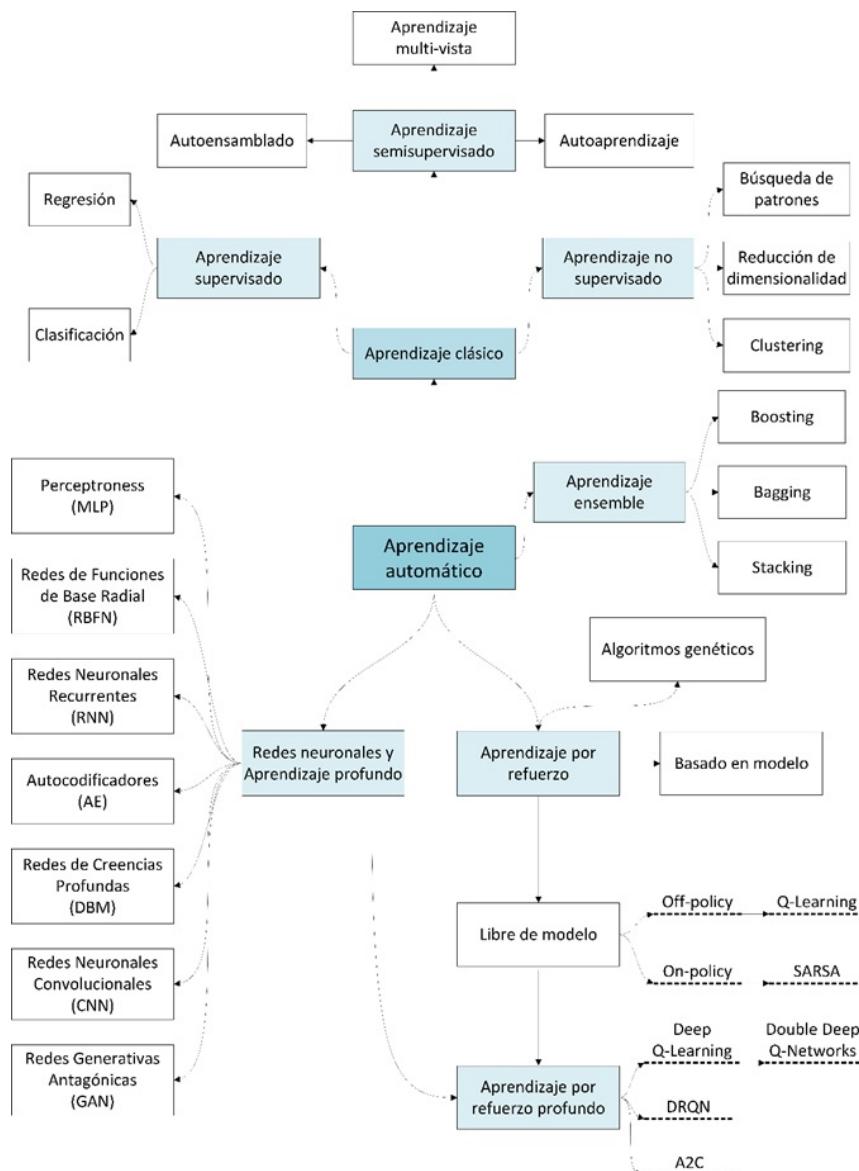


Figura 2. Taxonomía de las diferentes ramas del aprendizaje automático. Fuente: Alonso Rincón, 2020.

El aprendizaje de conceptos se plantea, a menudo, como una búsqueda en un espacio de posibles hipótesis (esto es, posibles soluciones al problema de aprendizaje) con el fin de encontrar la que mejor encaje con los datos de entrenamiento. En este **aprendizaje inductivo** se puede garantizar que la hipótesis encontrada es la que mejor encaja con los datos de entrenamiento, pero ¿encajará esa hipótesis también con nuevas instancias? Se asume que sí y se plantea, por

tanto, lo siguiente: cualquier hipótesis que encaje «suficientemente» bien con un conjunto «suficientemente» grande de ejemplos de entrenamiento, también encajará bien en instancias nuevas.

Como se muestra en la Figura 3, dentro del aprendizaje de conceptos podremos definir tres grandes grupos dentro del aprendizaje automático:

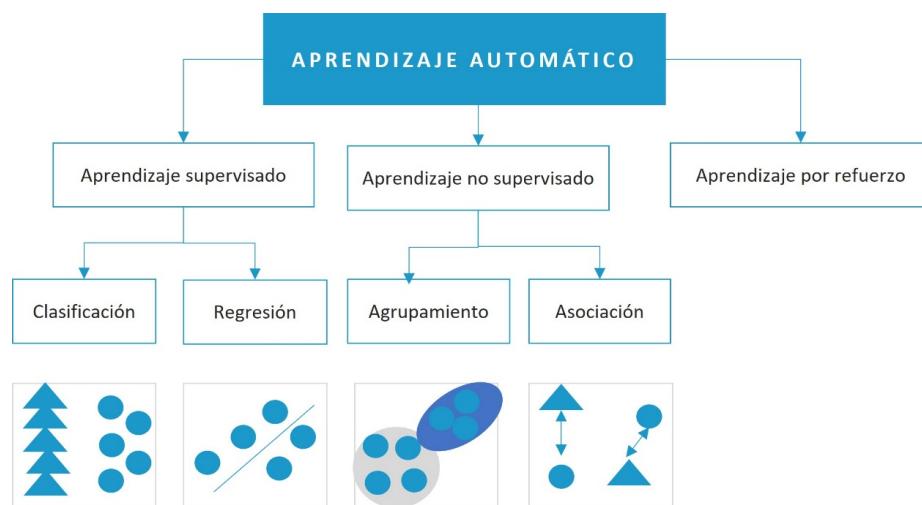


Figura 3. Taxonomía de las diferentes ramas del aprendizaje automático. Fuente: elaboración propia.

- ▶ **Aprendizaje supervisado:** pretende caracterizar o describir un concepto a partir de instancias de este. Para ello, se parte de un conjunto de **datos de entrenamiento etiquetados**, es decir, un humano u otro proceso previamente han especificado en los datos de entrenamiento a qué clase corresponde cada instancia con el fin de aprender a identificar a qué clasificación (o valor numérico, en el caso de la regresión) corresponde una instancia nueva no etiquetada.
- ▶ **Aprendizaje no supervisado:** pretende caracterizar un concepto desconocido a partir de instancias de este. Aquí no existen clases definidas y, por tanto, se trata de describir un nuevo concepto o clase. Es decir, para ello se parte de un conjunto de **datos de entrenamiento no etiquetados**. No se conocen las clases *a priori*. Este proceso, en ocasiones, precede al aprendizaje supervisado.

- ▶ **Aprendizaje por refuerzo:** este se basa en el entrenamiento de modelos de aprendizaje automático para tomar decisiones de forma secuencial. El método emplea prueba y error para encontrar una solución al problema, reforzando o debilitando las decisiones con recompensas o penalizaciones a las acciones que realiza, tratando de maximizar la recompensa final. Aquí pueden o no existir datos de entrenamiento previos. Un agente inteligente evalúa el estado del entorno en el que se encuentra y en cada estado decide llevar a cabo una acción dentro de las posibles que puede realizar, con el fin de maximizar las recompensas (o minimizar las penalizaciones) obtenidas por sus acciones.

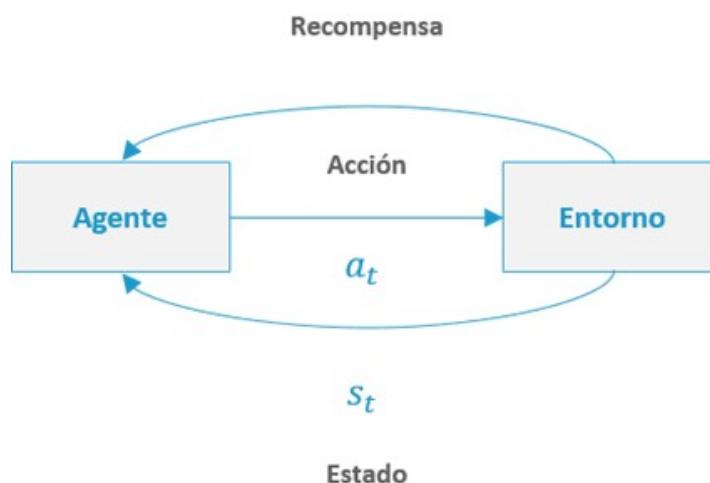


Figura 4. Interacción de un agente con su entorno en el aprendizaje por refuerzo. Fuente: adaptado de Alonso Rincón, 2020.

De forma habitual, un sistema de aprendizaje constará de las siguientes etapas:

- ▶ Selección del objetivo de aprendizaje.
- ▶ Selección del conjunto de datos de entrenamiento.
- ▶ Selección de una función objetivo y su representación.
- ▶ Selección del algoritmo de aprendizaje que aproximará la función objetivo.
- ▶ Evaluación y validación de los resultados.

## 9.4. Clasificación, regresión y agrupamiento

### Clasificación

El aprendizaje supervisado también se aplica a problemas de clasificación. La clasificación también es una de las herramientas importantes en la toma de decisiones. Básicamente, la clasificación es un problema de predicción de categorías. En un problema de predicción el objetivo es estimar una cantidad (de productos, económica, etc.).

En un problema de clasificación se intenta estimar a qué categoría pertenece la observación de interés. Por ejemplo, una empresa puede tener interés en estimar si un cliente tiene o no capacidad de pago, si puede o no realizar operaciones fraudulentas, si es o no un buen conductor. Se puede clasificar una empresa como solvente o no, potencial cliente o no, etc. En el contexto empresarial existen infinidad de ejemplos de la utilidad de un modelo de clasificación.

Un clasificado es básicamente una división del espacio de las variables independientes. A cada división se le asigna una clase o categoría. Es importante indicar que la clasificación binaria se refiere a casos donde solo hay dos categorías o clases. Los métodos de clasificación se pueden generalizar a un mayor número de categorías.

También podemos encontrar modelos de clasificación basados en:

- ▶ **Redes neuronales.** Se entrena una red neuronal para que su predicción sea la categoría o clase que corresponde a cada observación.
- ▶ **Boosting.** Consiste en combinar varios clasificadores sencillos para obtener uno más robusto. El aprendizaje es secuencial, cada nuevo clasificador sencillo se especializa en aprender las observaciones mal clasificadas por los anteriores.

- ▶ **Bagging.** A partir del conjunto original de observaciones de entrenamiento se generan  $m$  conjuntos nuevos de datos con las mismas propiedades. Con cada uno de estos conjuntos se entrena un clasificador y finalmente se combinan los  $m$  clasificadores para obtener uno robusto.

## Medidas de calidad de un modelo de clasificación

A la hora de decidir qué modelo de clasificación es el más adecuado para cierto proyecto es necesario disponer de métricas que midan de forma objetiva la precisión de los distintos modelos. Esta sección presenta varias de dichas medidas. Se comienza presentando qué es la matriz de confusión.

Una matriz de confusión es una herramienta para medir la calidad de un modelo de clasificación. Dado un conjunto de observaciones, las cuales tienen etiquetadas la categoría a la que pertenecen, la matriz de confusión se construye de la siguiente forma. Las columnas representan las categorías reales de las observaciones.

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Figura 5. Matriz de Confusión y fórmulas de Previsión y Recall. Fuente: Natassha Selvaraj

<https://www.kdnuggets.com/2022/11/confusion-matrix-precision-recall-explained.html>.

Las filas las categorías predichas por el modelo para cada observación. Por tanto, en la diagonal se tiene el número de observaciones con predicción acertada. El resto de las posiciones de la matriz indica cuántas observaciones de cada categoría se han predicho erróneamente y cuál ha sido la predicción realizada.

## Regresión

La regresión es un método de aprendizaje automático utilizado para modelar la relación entre una variable dependiente (objetivo) y una o más variables independientes (predictoras). La finalidad principal de la regresión es predecir un valor continuo basándose en las características observadas.

## Principales Características

- ▶ **Predictibilidad:** La regresión es útil para predecir valores numéricos continuos, como el precio de una casa, la temperatura o la demanda de un producto.
- ▶ **Linealidad:** La forma más sencilla de regresión es la regresión lineal, que asume una relación lineal entre las variables independientes y la variable dependiente.
- ▶ **Interpretabilidad:** Los modelos de regresión son a menudo fáciles de interpretar, especialmente los modelos lineales, que permiten entender la influencia de cada predictor en la variable dependiente.

## Técnicas de Regresión

- ▶ **Regresión Lineal Simple:** esta técnica modela la relación entre dos variables mediante una línea recta.
- ▶ **Regresión Lineal Múltiple:** extiende la regresión lineal simple para incluir múltiples variables independientes
- ▶ **Regresión Polinómica:** modela la relación entre la variable dependiente y las variables independientes como un polinomio de  $n$ -ésimo grado. Es útil cuando la relación no es lineal.
- ▶ **Regresión Logística:** aunque su nombre sugiere lo contrario, se utiliza para problemas de clasificación binaria. Modela la probabilidad de un resultado binario utilizando una función logística
- ▶ **Regresión Ridge y Lasso:** son técnicas de regularización que modifican la función de costo para incluir términos que penalizan grandes coeficientes, ayudando a prevenir el sobreajuste. Ridge utiliza la penalización L2 mientras que Lasso utiliza la penalización L1.

## Otras Consideraciones

- ▶ **Evaluación del Modelo:** Las métricas comunes para evaluar los modelos de regresión incluyen el error cuadrático medio (MSE), el coeficiente de determinación y la raíz del error cuadrático medio (RMSE).
- ▶ **Suposiciones:** Los modelos de regresión lineal asumen que existe una relación lineal entre las variables, que los errores son independientes y distribuidos normalmente con media cero y varianza constante (homocedasticidad).
- ▶ **Aplicaciones Prácticas:** Los modelos de regresión se utilizan ampliamente en economía para prever precios y tendencias, en biología para modelar crecimiento de poblaciones, en ingeniería para predecir la vida útil de componentes, y en muchos otros campos.

## Ventajas y Limitaciones

- ▶ **Ventajas:** Simplicidad, interpretabilidad, y eficiencia computacional.
- ▶ **Limitaciones:** Sensibilidad a los valores atípicos y a la multicolinealidad, y posibles problemas de sobreajuste con muchos predictores.

## Agrupamiento o clustering

El clustering es uno de los métodos de aprendizaje no supervisado más importantes y, como cualquier otro método de aprendizaje no supervisado, busca caracterizar conceptos desconocidos a partir de instancias de estos. En este tipo de problemas de aprendizaje no supervisado, la clase es desconocida y, precisamente, el descubrimiento de esta clase es el objetivo a través de la agrupación de instancias con base en un esquema de similitud.

Podemos definir entonces clustering como un método de aprendizaje no supervisado que permite agrupar objetos en clústeres o agrupamientos, cuyos miembros son similares entre sí en cierto modo. Por otro lado, definimos clúster como una colección de objetos similares entre sí y diferentes a los objetos que pertenecen a otros clústeres.

El aprendizaje no supervisado pretende caracterizar un concepto desconocido a partir de instancias de este. En este caso no existen clases definidas y, por tanto, se trata de describir un nuevo concepto o clase.

Las técnicas de agrupamiento o clustering son muy utilizadas en problemas de aprendizaje no supervisado. Mediante clustering, las instancias se agrupan de acuerdo con un esquema de similitud. En este tipo de aprendizaje no supervisado, los datos de entrenamiento no especifican qué se está intentando aprender (los agrupamientos), mientras que, en el aprendizaje supervisado, las clases que se están intentando describir sí están especificadas.

El clustering se puede utilizar para el análisis de datos como un primer paso en la construcción de un modelo de estos. El algoritmo de clustering típicamente proporciona visualizaciones (véase la figura 6) a partir de las cuales se pueden encontrar atributos similares en las instancias y tratar de extraer conclusiones. Se puede utilizar, además, en tareas de generalización, descubriendo instancias similares, que comparten propiedades, y pudiendo incorporar futuras instancias en los agrupamientos generados.

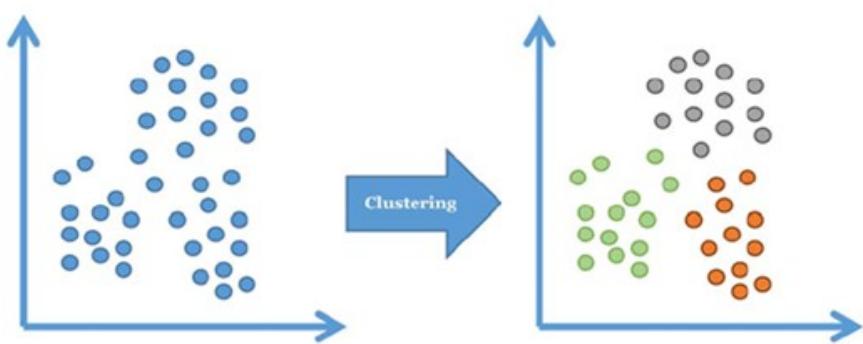


Figura 6. Ilustración básica del concepto de clustering.

Diferentes algoritmos de clustering pueden dar lugar a diferentes agrupamientos y no es siempre fácil determinar qué es un buen agrupamiento. La calidad de los agrupamientos creados dependerá de la aplicación final, por lo cual es el usuario el que determinará la calidad con base en si los agrupamientos creados son útiles y se ajustan a sus necesidades. Por lo tanto, habrá que seleccionar un algoritmo u otro según el objetivo del problema a tratar.

Por ejemplo, no es lo mismo tratar de detectar el espacio donde hay una mancha de fuel en el mar, espacio que puede tener una forma irregular, que tratar de obtener agrupamientos circulares y de similar tamaño con el fin de segmentar un grupo de clientes para ofrecer productos.

A continuación, se enumeran diferentes tipos de algoritmos de clustering en función del tipo de agrupamientos que producen:

- ▶ Agrupamientos exclusivos. Pueden generarse por métodos que partitionan los datos creando un número  $k$  determinado de clústeres. Cada uno de los clústeres tiene, al menos, un objeto y los objetos se agrupan de modo exclusivo, pudiendo pertenecer únicamente a un clúster.
- Generalmente, estos métodos de partición utilizan una medida de distancia para generar los clústeres. Sería el caso del algoritmo K-means.
- También existe otro tipo de algoritmos que típicamente generan clústeres exclusivos tales como los métodos basados en la densidad, que van incorporando nuevos objetos a un clúster si la densidad de objetos en «los alrededores» supera un umbral. Los algoritmos basados en densidad (en vez de en medidas típicas de distancia) son útiles para generar clústeres de formas irregulares (p. ej.: mean-sift o DBSCAN).
- ▶ Agrupamientos jerárquicos. Existen algoritmos jerárquicos que dan lugar a una estructura jerárquica de clústeres. En el primer nivel de la jerarquía se tiene un único clúster que, en una primera iteración, se divide en clústeres. Cada uno de estos se divide a su vez y se van generando nuevos clústeres en siguientes iteraciones. Esto es lo que se llama una aproximación divisoria (p. ej.: DIANA). También existen algoritmos que generan clústeres en el sentido inverso, llamados aglomerativos, generando primeramente los clústeres más pequeños y agrupándolos progresivamente para generar la estructura jerárquica (p. ej.: AGNES).
- ▶ Agrupamientos solapados. Los objetos se agrupan a través de conjuntos difusos y cada objeto puede pertenecer a uno o más clústeres con diferentes grados de pertenencia. Un algoritmo que produce agrupamientos solapados es el Fuzzy C-means.

- ▶ Agrupamientos probabilistas. Los clústeres se generan mediante un método probabilístico como es el algoritmo EM (Expectation- Maximization).

## Medida de distancia

Las medidas de distancia son utilizadas en un importante número de algoritmos de clustering, puesto que muchos algoritmos se basan en medidas de distancia entre objetos para, en función de su «cercanía», incluirlos en el mismo clúster o en clústeres separados.

Según la medida de distancia que se escoja, un algoritmo dará lugar a diferentes clústeres y, por tanto, seleccionar la medida de distancia apropiada es importante, aunque no siempre se puede saber a ciencia cierta cuál es la medida óptima. Es muy utilizada la medida de distancia euclídea, por ejemplo. Pero existen otros tipos de medidas muy utilizadas también típicamente por algoritmos de clustering no tan extendidas. A continuación, se describen algunas de estas medidas de distancia conocidas como linkage measures (medidas de conectividad):

Enlace sencillo (single-linkage). La similitud entre dos clústeres se calcula como la similitud de los dos puntos más cercanos pertenecientes a los diferentes clústeres. Así, se considera la distancia más pequeña existente entre uno de los puntos del primer clúster y otro del segundo (ver figura 7).

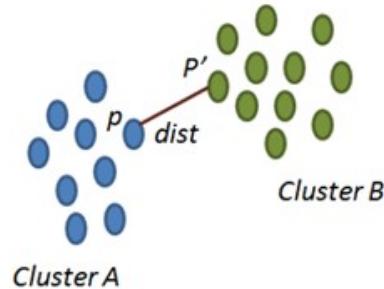


Figura 7. Enlace sencillo entre dos clústeres. Fuente: <https://pabloruizruiz10.com/resources/Curso-Machine-Learning-Esp/5---Aprendizaje-No-Supervisado/Intro-Clustering.html>

Enlace completo (complete-linkage). Caso opuesto al enlace sencillo, ya que se tiene en cuenta la distancia mayor existente entre cualquier punto de uno y otro clúster (ver figura 8).

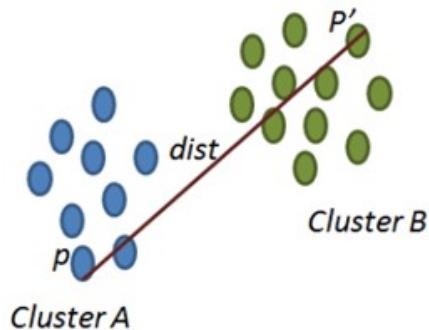


Figura 8. Enlace completo entre dos clústeres. Fuente: <https://pabloruizruiz10.com/resources/Curso-Machine-Learning-Esp/5---Aprendizaje-No-Supervisado/Intro-Clustering.html>

Enlace promedio (average-linkage). La distancia entre dos clústeres se calcula como la distancia media entre cualquier punto del primer clúster con cualquier punto del segundo (ver figura 9).

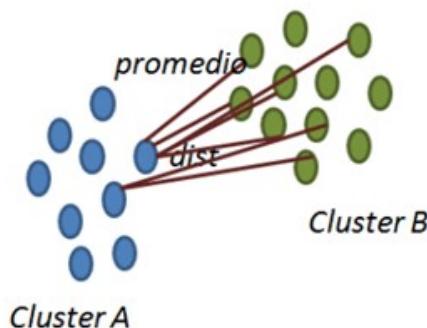


Figura 9. Enlace promedio entre dos clústeres. Fuente: <https://pabloruizruiz10.com/resources/Curso-Machine-Learning-Esp/5--Aprendizaje-No-Supervisado/Intro-Clustering.html>

Algunos ejemplos de aplicaciones prácticas del clustering en entornos de Industria 4.0 son:

- ▶ Encontrar objetos representativos de grupos homogéneos. Por ejemplo, en un sistema de ahorro energético, identificar usuarios con características similares para determinar diferencias de comportamientos y consumos. Analizando cada grupo, se pueden definir estrategias de ahorro energético atendiendo a criterios a priori desconocidos.
- ▶ En un sistema de gestión de clientes, las técnicas de clustering pueden agrupar los diversos clientes según características comunes y permitir aplicar estrategias empresariales adecuadas a los diversos grupos, posicionar productos, etc. (Abbasimehr y Shabani, 2019).
- ▶ Encontrar grupos y describirlos en función de sus propiedades. Por ejemplo, en biología, clasificación de plantas o, en medicina, clasificación de enfermedades raras o, en producción hortofrutícola, para la clasificación de los productos (Pacheco y López, 2019).
- ▶ Detección de casos anómalos. Por ejemplo, en un sistema de detección de errores en líneas de comunicación al detectar patrones de ruido que no correspondan con ningún caso descrito.

## 9.5. Aprendizaje Profundo

A pesar de resolver múltiples problemas eficazmente, las redes neuronales de retropropagación presentan una serie de limitaciones que deben ser resueltas. Estas limitaciones reducen principalmente la eficacia de este tipo de redes en problemas complejos en los que el número de nodos intermedios es elevado. El elevado número de nodos aumenta el número de conexiones y, por ende, el cálculo de los pesos asociados. De este modo, el entrenamiento de las redes se convierte en un proceso ineficiente y costoso, por lo que han de buscarse soluciones que nos permitan la creación de redes neuronales con decenas y cientos de nodos y capas.

Es entonces cuando aparece el aprendizaje profundo, más conocido por traducción al inglés deep learning (Graupe, 2019; Pouyanfar et al., 2018). El deep learning es un área específica del aprendizaje automático que se basa en la utilización de redes neuronales con un alto número de nodos y capas. Este nuevo concepto dentro del aprendizaje automático lida con procesos complejos que trabajan con volúmenes elevados de datos, así como con la interconexión necesaria entre los diferentes sistemas que forman parte de la solución completa.

Se puede definir entonces el aprendizaje profundo o deep learning como una clase de algoritmos de aprendizaje automático que utiliza múltiples capas para extraer progresivamente características de nivel superior de la entrada bruta. Esto incluye, por tanto, una cascada de capas conectadas entre sí, con unidades de procesamiento no lineal en cada una de ellas. De este modo, las primeras capas (inferiores) se corresponden con niveles de abstracción menor, mientras que las últimas (superiores) se basan en las anteriores para formar una representación jerárquica de conceptos.

Tal y como se vio en la sección anterior, existe una amplia y creciente variedad de redes neuronales artificiales.

Dentro de los sistemas deep learning pueden destacarse por su relevancia las siguientes (Liu, Wang, Liu, Zeng, Liu, y Alsaadi, 2017):

- ▶ Redes prealimentadas (Feedforward networks) y redes prealimentadas profundas (Deep Feedforward Networks).
- ▶ Redes neuronales recurrentes (RNN, Recurrent Neural Networks) y redes recurrentes profundas (Deep Recurrent Networks).
- ▶ Autoencoders (AE).
- ▶ Redes neuronales convolucionales (CNN, Convolutional Neural Networks).
- ▶ Redes Generativas Antagónicas (GAN, Generative Adversarial Networks).

Además de estas grandes tendencias, existen muchos otros tipos que deben ser mencionados. Las redes de creencias profundas o deep belief networks (DBN) (Hinton, 2009) son arquitecturas apiladas de mayormente RBMs (Restricted Boltzmann Machines o máquinas Boltzmann restringidas) o de autocodificadores variacionales (VAE).

Las redes neuronales y, por ende, las redes neuronales profundas pueden utilizarse como sustitutos de los métodos supervisados, no supervisados y semisupervisados, algunos vistos en secciones anteriores y algunos que se verán a continuación, obteniendo una precisión mucho mayor (obviamente a cambio de un mayor coste de computación tanto en la etapa de entrenamiento como en la etapa de testing o producción).

Además de para resolver cualquier problema que ya podíamos solventar mediante los métodos de machine learning clásico, también se utilizan, entre otras posibles aplicaciones, para la identificación de objetos en imágenes y vídeos (Shah, Bennamoun y Boussaid, 2016), el reconocimiento de voz (Tandel et al. 2020), la síntesis de voz (Arik et al., 2017), el análisis de sentimientos y reconocimiento de emociones del habla (Fayek, Lech, y Cavedon, 2017), el procesamiento de imágenes (Razzak, Naz y Zaib, 2018), la transferencia de estilos (por ejemplo, aplicación del estilo de pintura de Van Gogh a cualquier fotografía) (Luan, Paris, Shechtman y Bala, 2017), el procesamiento del lenguaje natural (natural language processing) (Costa-Jussà, Allauzen, Barrault, Cho y Schwenk, 2017), la traducción automática (Deng y Liu, 2018), etc.

## 9.6. Ejemplos de aplicaciones de aprendizaje automático

El aprendizaje automático está transformando diversas industrias con aplicaciones innovadoras que van más allá de los usos tradicionales. A continuación, se presentan algunos ejemplos actuales y vanguardistas de cómo se está utilizando esta tecnología:

### Medios Sociales y Personalización

- ▶ Funcionalidades en Redes Sociales: Plataformas como Facebook utilizan algoritmos de aprendizaje automático para analizar las actividades de los usuarios y ofrecer sugerencias personalizadas de amigos y páginas. Estas funcionalidades mejoran la experiencia del usuario al adaptar el contenido a sus preferencias individuales.

### Recomendación de Productos

- ▶ E-commerce: Sitios web como Amazon emplean algoritmos para rastrear el comportamiento de compra, historial de búsquedas y patrones de carrito de compras para recomendar productos que probablemente interesen a los usuarios. Esto no solo incrementa las ventas, sino que también mejora la satisfacción del cliente al proporcionar una experiencia de compra más personalizada.

## Reconocimiento de Imágenes

- ▶ Diagnóstico Médico: Algoritmos de redes neuronales convolucionales (CNN) se utilizan en la detección de cáncer de piel, proporcionando tasas de precisión muy altas en comparación con métodos manuales. Esta tecnología analiza miles de imágenes para identificar lesiones benignas y malignas.

## Análisis de Sentimientos

- ▶ Marketing y Atención al Cliente: Las herramientas de análisis de sentimientos ayudan a las empresas a entender las opiniones y emociones de los clientes en tiempo real, a partir de reseñas, correos electrónicos y publicaciones en redes sociales. Esto es crucial para mejorar la atención al cliente y adaptar las estrategias de marketing.

## Optimización del Viaje del Cliente

- ▶ Marketing Digital: Las técnicas de aprendizaje automático optimizan el viaje del cliente al analizar y predecir los puntos de interés en tiempo real, mejorando las estrategias de adquisición y retención de clientes mediante recomendaciones personalizadas.

## Automatización de Control de Acceso

- ▶ Seguridad en el Lugar de Trabajo: Las organizaciones implementan algoritmos para determinar los niveles de acceso de los empleados basados en sus perfiles de trabajo, mejorando la seguridad y eficiencia en la gestión de recursos humanos.

## Preservación de Vida Marina

- ▶ Conservación Ambiental: Algoritmos de aprendizaje automático se utilizan para modelar el comportamiento de especies marinas en peligro, ayudando a los científicos a regular y monitorear sus poblaciones de manera más efectiva.

## Predicción de Fallos Cardíacos

- ▶ Medicina Preventiva: Algoritmos analizan las notas de los médicos y patrones en el historial cardiovascular de los pacientes para identificar riesgos de fallos cardíacos, facilitando diagnósticos más rápidos y precisos.

## Traducción de Idiomas

- ▶ Comunicación Global: Herramientas de traducción de idiomas basadas en aprendizaje automático permiten traducciones precisas y contextuales entre múltiples idiomas, eliminando barreras lingüísticas y facilitando la comunicación global.

## Detección de Bots Maliciosos

- ▶ Ciberseguridad: Los sistemas de detección de bots en plataformas como Twitter utilizan técnicas de aprendizaje supervisado para identificar y clasificar bots buenos y malos, reduciendo la propagación de información falsa y amenazas cibernéticas.

## Ejemplos en un Escenario Común

Imagina un ecosistema de e-commerce avanzado que combina todas estas aplicaciones de aprendizaje automático. En este entorno, la plataforma no solo recomienda productos basados en las preferencias de compra del usuario, sino que también ofrece traducciones automáticas del contenido del sitio para usuarios internacionales. Mientras tanto, el sistema de seguridad identifica y bloquea bots maliciosos que intentan acceder a cuentas de usuario.

Las herramientas de análisis de sentimientos y chatbots mejoran la atención al cliente, ofreciendo respuestas personalizadas y en tiempo real. Además, la plataforma utiliza algoritmos para predecir y gestionar problemas de salud de sus empleados y usuarios, asegurando un ambiente seguro y saludable. Todo esto se integra en una experiencia de usuario fluida y altamente personalizada, que optimiza la satisfacción del cliente y la eficiencia operativa de la empresa.

Estas aplicaciones demuestran el impacto transformador del aprendizaje automático en la vida cotidiana y en diversos sectores, marcando una evolución significativa hacia sistemas más inteligentes y eficientes.

## 9.7. Referencias bibliográficas

Abbasimehr, H. y Shabani, M. (2019). *A new methodology for customer behavior analysis using time series clustering: A case study on a bank's customers*. Kybernetes. <https://doi.org/10.1108/K-09-2018-0506>

Alonso, R. S. (2020). *Deep Reinforcement Learning para la gestión de Redes Definidas por Software en arquitecturas Edge Computing para el Internet de las Cosas* [tesis, Universidad de Salamanca]. Repositorio documental Credos. <https://gredos.usal.es/handle/10366/145257>

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11), 27-34.

G. Hinton et al. (2012, noviembre). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, 6(29), 82-97.

Ibnkahla, M. (2000). Applications of neural networks to digital communications—a survey. *Signal processing*, 80(7), 1185-1215.

Nilsson, N. J. (1986). Probabilistic logic. *Artificial intelligence*, 28(1), 71-87.

Panesar, A. (2019). What Is Machine Learning? En *Machine Learning and AI for Healthcare* (75-118). Springer.

Rogers, S. y Girolami, M. (2017). *A first course in machine learning* (2.<sup>a</sup> ed.). CRC Press; Taylor & Francis Group; Chapman & Hall book.

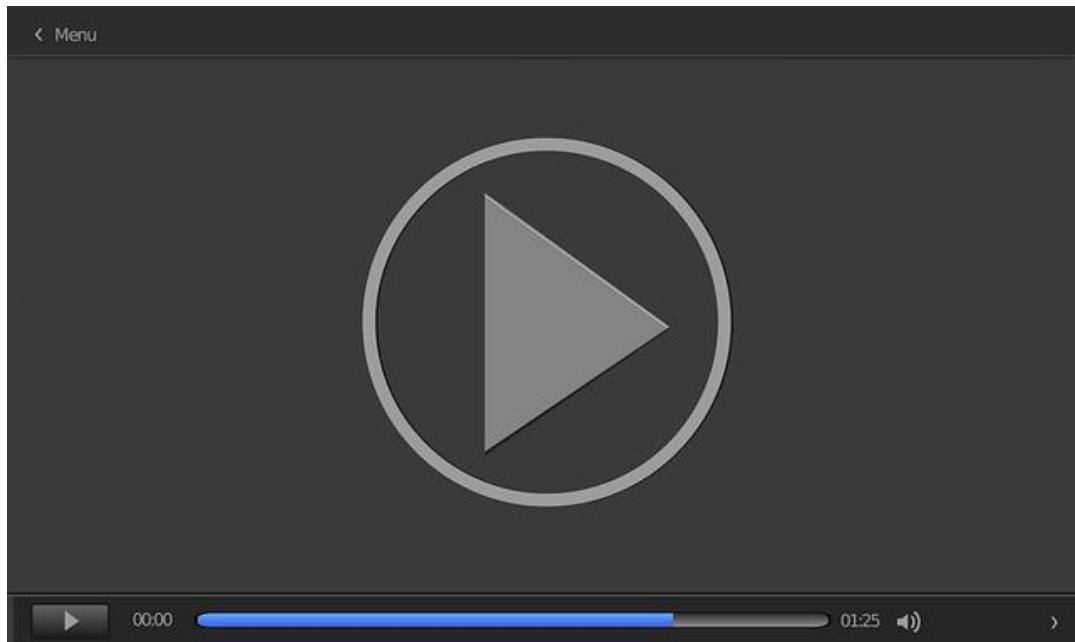
Deng, L. y Liu, Y. (2018). *Deep learning in natural language processing*. Springer.

- Costa-Jussà, M. R., Allauzen, A., Barrault, L., Cho, K. y Schwenk, H. (2017). Introduction to the special issue on deep learning approaches for machine translation. *Computer Speech & Language*, 46, 367-373.
- Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., Moor, B. D. y Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.
- Graupe, D. (2019). *Deep Learning Neural Networks: Design and Case Studies*. World Scientific.
- Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M. L., Chen, S. C., & Iyengar, S. S. (2019). A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Computing Surveys*, 51(5). <https://doi.org/10.1145/3234150>.
- Luan, F., Paris, S., Shechtman, E., & Bala, K. (2017). *Deep Photo Style Transfer*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4990-4998).
- Razzak, M. I., Naz, S., & Zaib, A. (2017). *Deep Learning for Medical Image Processing: Overview, Challenges and the Future*. In *Classification in BioApps* (pp. 323-350). Springer. doi:10.1007/978-3-319-65981-7\_12. Retrieved from [https://scihub.se/10.1007/978-3-319-65981-7\\_12](https://scihub.se/10.1007/978-3-319-65981-7_12).
- Shah, A. R., Bennamoun, M., & Boussaid, F. (2016). Automatic deep metalearning for 3D image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 27(6), 1229-1242. <https://doi.org/10.1109/TNNLS.2015.2441159>
- Liu, Z., Wang, Y., Liu, J., Zeng, N., Liu, X., & Alsaadi, F. E. (2017). A novel robust fractional-order fuzzy neural network sliding mode control for nonlinear systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(8), 2047-2059. <https://doi.org/10.1109/TSMC.2017.2679183>

## Introducción al aprendizaje automático

MIT OpenCourseWare. (2017, mayo 19). 11. *Introduction to Machine Learning* [Vídeo]. YouTube. <https://www.youtube.com/watch?v=h0e2HAPTGF4&t=1s>

El siguiente recurso se trata de una clase magistral del MIT, en la que el profesor Eric Grimson hace una introducción al pensamiento computacional y a la ciencia de los datos.



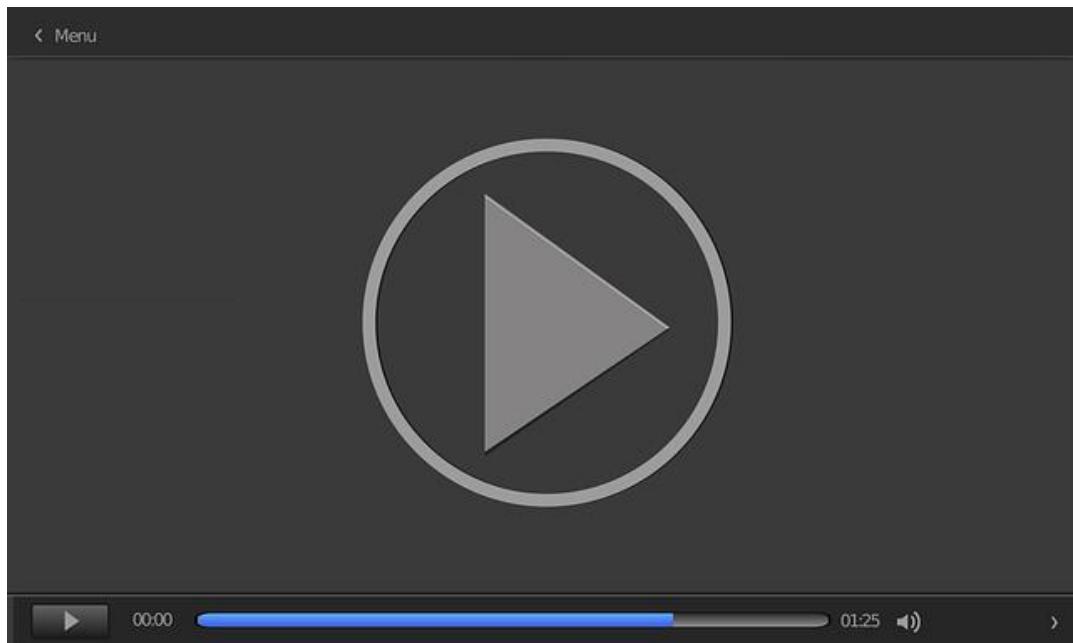
Accede al vídeo:

<https://www.youtube.com/embed/h0e2HAPTGF4>

## El aprendizaje automático dentro de la inteligencia artificial

MIT OpenCourseWare. (2020, junio 25). 1. *Artificial Intelligence and Machine Learning* [Vídeo]. YouTube. [https://www.youtube.com/watch?v=t4K6lney7Zw&list=RDCMUCEBb1b\\_L6zDS3xTUrIALZOw&index=2](https://www.youtube.com/watch?v=t4K6lney7Zw&list=RDCMUCEBb1b_L6zDS3xTUrIALZOw&index=2)

El siguiente recurso trata de una clase magistral del MIT, en la que se resuelven y se plantean los fundamentos de las matemáticas del *big data* y el aprendizaje automático.



Accede al vídeo:

<https://www.youtube.com/embed/t4K6lney7Zw>

- 1.** El aprendizaje automático consiste en extraer conocimiento a partir de los datos.

  - A. Verdadero.
  - B. Falso.
  
- 2.** Los tres grandes grupos que se pueden definir dentro del aprendizaje automático son:

  - A. Aprendizaje por refuerzo.
  - B. Aprendizaje supervisado y no supervisado.
  - C. Aprendizaje semisupervisado.
  - D. Las tres anteriores son correctas.
  
- 3.** Dentro del aprendizaje supervisado existen diferentes técnicas, entre ellas:

  - A. Clasificación y regresión.
  - B. Agrupamiento y asociación.
  - C. Aprendizaje por refuerzo.
  - D. Ninguna de las anteriores.
  
- 4.** El aprendizaje por refuerzo se basa en el entrenamiento de modelos de aprendizaje automático para tomar decisiones de forma secuencial.

  - A. Verdadero.
  - B. Falso.
  
- 5.** ¿Cuáles de los siguientes puntos pertenecen al aprendizaje por refuerzo?

  - A. Agente, acción, entorno.
  - B. Recompensa, intercambio, acción.
  - C. Estado, lugar, tiempo.
  - D. Ninguna de las anteriores.

- 6.** ¿Cuáles de los siguientes puntos pertenecen a las fases del proceso de extracción de conocimiento visto en clase?

  - A. Datos, datos objetivo.
  - B. Datos transformados, patrones, conocimiento.
  - C. Datos preprocesados.
  - D. Todas las anteriores.
- 7.** El proceso KDD es interactivo e iterativo.

  - A. Verdadero.
  - B. Falso.
- 8.** La elección de los algoritmos de minería de datos incluye:

  - A. Seleccionar el método o métodos que se utilizarán para buscar patrones en los datos.
  - B. Búsqueda de patrones de interés en una forma de representación particular.
  - C. La interpretación de los patrones descubiertos.
  - D. Ninguna de las anteriores.
- 9.** Un sistema de aprendizaje consta, en parte, de las siguientes etapas:

  - A. Selección del objetivo de aprendizaje.
  - B. Selección del conjunto de datos de entrenamiento.
  - C. Selección de una función objetivo.
  - D. Todas las anteriores.

**10.** El aprendizaje automático tiene múltiples aplicaciones en otro tipo de sistemas, como en aquellos relacionados con la robótica, o en sistemas de reconocimiento del habla.

- A. Verdadero.
- B. Falso.

Ciencia de Datos Aplicada

---

# Tema 10. Aplicación en Inteligencia Artificial

# Índice

[Esquema](#)

[Ideas clave](#)

[10.1. Introducción y objetivos](#)

[10.2. Objetivos de la Inteligencia Artificial aplicada](#)

[10.3. Datos masivos en la Inteligencia Artificial](#)

[10.4. La visión, el lenguaje natural y conocimiento](#)

[10.5. Ejemplos de aplicación de la Inteligencia Artificial](#)

[10.6. Referencias bibliográficas](#)

[A fondo](#)

[Redes neuronales y el aprendizaje profundo](#)

[Procesamiento del Lenguaje Natural con Redes Neuronales](#)

[Test](#)

## APLICACIONES EN INTELIGENCIA ARTIFICIAL

### Inteligencia Artificial

#### Conceptos

- Aprendizaje Automático
  - Supervisado
  - No Supervisado
  - Por Refuerzo
- Procesamiento del Lenguaje Natural (NLP)
- Visión por Computadora
- Robótica
- Sistemas Expertos

#### Objetivos

- Automatización de Tareas
  - Reducción de Tareas Repetitivas
  - Mejora de la Toma de Decisiones
  - Análisis Predictivo
  - Personalización de Experiencias
  - Recomendaciones Personalizadas
  - Innovación Tecnológica
  - Descubrimiento de Fármacos
  - Interacción Humano-Máquina
  - Asistentes de Voz
- Resolución de Problemas Complejos
  - Modelado Climático
  - Optimización de Recursos
  - Logística
- Visión, lenguaje natural y conocimiento

#### Datos masivos

- Importancia de Big Data
  - Calidad de Datos
  - Cantidad de Datos
- Aplicaciones
  - Salud
    - Diagnóstico y Medicina Personalizada
  - Finanzas
    - Detección de Fraudes y Análisis de Riesgos
  - Marketing
    - Segmentación de Clientes y Recomendaciones

## 10.1. Introducción y objetivos

En este tema, nos centraremos en los **conceptos fundamentales de la inteligencia artificial (IA)** desde una perspectiva aplicada, con el fin de comprender cómo estos principios se utilizan en la práctica para resolver problemas reales. Asumiendo que las definiciones básicas ya se han cubierto en apartados anteriores, nos enfocaremos en cómo los conceptos teóricos se traducen en aplicaciones concretas.

La IA ha transformado múltiples sectores, desde la medicina hasta la industria del entretenimiento, permitiendo la **automatización de procesos, la toma de decisiones basada en datos y la creación de sistemas inteligentes capaces de interactuar con los humanos de manera natural**. Para entender esta transformación, exploraremos las principales áreas de la IA, los modelos y algoritmos más utilizados, y los componentes esenciales que forman la base de cualquier sistema de IA. Además, discutiremos las implicaciones éticas y las tendencias emergentes que moldean el futuro de la inteligencia artificial.

### Aplicabilidad de los conceptos de IA

#### Aprendizaje automático (*Machine Learning*)

- ▶ Aplicaciones: diagnóstico médico, detección de fraudes, recomendaciones personalizadas en plataformas de Streaming.
- ▶ Modelos comunes: redes neuronales profundas para reconocimiento de imágenes, algoritmos de clustering para segmentación de clientes.

#### Procesamiento del lenguaje natural (NLP)

- ▶ Aplicaciones: asistentes virtuales (como Alexa y Siri), traducción automática, análisis de sentimientos en redes sociales.

- ▶ Técnicas usadas: modelos de lenguaje como BERT y GPT para comprender y generar texto.

## Visión por computadora

- ▶ Aplicaciones: sistemas de seguridad y vigilancia, vehículos autónomos, diagnóstico por imágenes en medicina.
- ▶ Herramientas: redes neuronales convolucionales (CNN) para el reconocimiento de patrones en imágenes.

## Robótica

- ▶ Aplicaciones: automatización en manufactura, robots de servicio, exploración espacial.
- ▶ Integración: uso de algoritmos de IA para navegación autónoma y manipulación de objetos.

## Sistemas expertos:

- ▶ Aplicaciones: diagnóstico médico, sistemas de soporte a decisiones en negocios, planificación logística.
- ▶ Componentes: bases de conocimiento y motores de inferencia que simulan el juicio humano.

## Componentes de un sistema de IA

### Datos

- ▶ Rol crucial: los datos son el combustible de los modelos de IA. La calidad y cantidad de datos disponibles determinan el éxito de las aplicaciones de IA.
- ▶ Preprocesamiento: técnicas para limpiar y preparar los datos antes de su uso en modelos de IA.

## Modelos y algoritmos

- ▶ Creación y entrenamiento: selección del algoritmo adecuado y entrenamiento con datos relevantes.
- ▶ Evaluación: uso de métricas para medir el rendimiento y ajuste de los modelos.

## Infraestructura

- ▶ *Hardware*: GPUs y TPUs que aceleran el procesamiento de grandes volúmenes de datos.
- ▶ *Software*: frameworks y bibliotecas como TensorFlow y PyTorch que facilitan el desarrollo de modelos de IA.

## Ética y futuro de la IA

### Implicaciones éticas

- ▶ Privacidad y sesgo: consideraciones cruciales para asegurar que las aplicaciones de IA sean justas y responsables.
- ▶ Transparencia: la necesidad de que los modelos de IA sean interpretables y auditables.

### Tendencias emergentes

- ▶ Innovaciones recientes: avances en inteligencia artificial explicable (XAI), modelos generativos y aplicaciones en salud y educación.
- ▶ Desafíos futuros: seguridad, sostenibilidad y el impacto socioeconómico de la IA.

En este tema se persiguen los siguientes objetivos:

- ▶ Comprender los conceptos teóricos de la IA.
- ▶ Ver cómo se aplican estos conceptos para crear soluciones innovadoras y efectivas en el mundo real.

## 10.2. Objetivos de la Inteligencia Artificial aplicada

La inteligencia artificial (IA) no solo se define por sus capacidades técnicas, sino también por los objetivos que persigue y las soluciones que ofrece en diversos campos. Este apartado explora los principales objetivos de la IA, mostrando cómo estos se traducen en aplicaciones prácticas que transforman industrias y mejoran la vida cotidiana. Entender estos objetivos nos permite apreciar la profundidad y el alcance de la IA, así como su potencial para resolver problemas complejos y optimizar recursos.

### Automatización de tareas repetitivas y monótonas

La **automatización** de tareas repetitivas y monótonas ha sido uno de los **principales motores del desarrollo de la IA**. Al delegar estas tareas a sistemas inteligentes, las empresas pueden mejorar la eficiencia operativa, reducir errores humanos y disminuir costos laborales. Esto permite a los trabajadores humanos enfocarse en tareas que requieren creatividad, juicio y habilidades interpersonales.

El objetivo es liberar a los seres humanos de tareas repetitivas y tediosas para que puedan centrarse en actividades más creativas y estratégicas.

#### Aplicaciones:

- ▶ Industria manufacturera: robots industriales que ensamblan productos en líneas de producción, como en la fabricación de automóviles.
- ▶ Servicios al cliente: chatbots y asistentes virtuales que responden consultas frecuentes, gestionan reservas y solucionan problemas básicos.
- ▶ Oficinas: sistemas de gestión documental que automatizan la clasificación y archivo de documentos, así como la generación de informes financieros y contables.

## Mejora de la toma de decisiones

La capacidad de la IA para analizar grandes conjuntos de datos y extraer patrones y tendencias permite a las organizaciones tomar decisiones más informadas y precisas. La toma de decisiones basada en datos reduce el riesgo de errores y mejora la efectividad de las estrategias empresariales y operativas.

Su objetivo es utilizar grandes volúmenes de datos (*big data*) y algoritmos avanzados para mejorar la precisión y la rapidez en la toma de decisiones.

Aplicaciones:

- ▶ Finanzas: algoritmos de IA que analizan datos de mercado y predicen movimientos financieros para la gestión de inversiones y la detección de fraudes.
- ▶ Salud: sistemas de apoyo a la decisión clínica que analizan historiales médicos y datos de pacientes para proporcionar diagnósticos más precisos y personalizados.
- ▶ *Marketing*: plataformas que segmentan clientes y personalizan campañas publicitarias basadas en el análisis del comportamiento de los usuarios y sus preferencias.

## Personalización de experiencias y servicios

La IA permite a las empresas ofrecer productos y servicios altamente personalizados, mejorando la experiencia del cliente y aumentando la satisfacción y fidelidad. Al analizar datos de comportamiento y preferencias, los sistemas de IA pueden anticipar las necesidades del usuario y adaptar sus ofertas en consecuencia.

Su objetivo es ofrecer experiencias y servicios personalizados que se adapten a las necesidades y preferencias individuales de cada usuario.

## Aplicaciones:

- ▶ *E-commerce*: recomendaciones de productos basadas en el historial de navegación y compras del usuario, como las que ofrece Amazon.
- ▶ Educación: plataformas de aprendizaje adaptativo que ajustan el contenido y el ritmo de enseñanza según el progreso y las necesidades del estudiante.
- ▶ Entretenimiento: servicios de *streaming* como Netflix que sugieren películas y series basadas en los gustos y hábitos de visualización del usuario.

## Innovación y desarrollo tecnológico

La IA impulsa la innovación al abrir nuevas posibilidades tecnológicas y **permitir la creación de soluciones** que antes eran inimaginables. Desde el descubrimiento de nuevos materiales hasta avances en biotecnología, la IA está en el centro de las revoluciones tecnológicas contemporáneas.

Su objetivo es fomentar la innovación y el desarrollo de nuevas tecnologías que pueden revolucionar diversos campos y mejorar la calidad de vida.

## Aplicaciones:

- ▶ Investigación científica: utilización de algoritmos de IA para modelar fenómenos complejos y acelerar el descubrimiento de nuevos fármacos y materiales.
- ▶ Automoción: desarrollo de vehículos autónomos que utilizan la IA para navegar y operar de manera segura, reduciendo el riesgo de accidentes y mejorando la eficiencia del tráfico.
- ▶ Energía: optimización de redes eléctricas inteligentes y gestión eficiente de recursos energéticos mediante el análisis de datos de consumo y producción.

## Mejora de la interacción humano-máquina

La mejora de la interacción humano-máquina es fundamental para aumentar la accesibilidad y usabilidad de las tecnologías avanzadas. Interfaces intuitivas y amigables permiten a los usuarios aprovechar al máximo las capacidades de la IA sin necesidad de conocimientos técnicos avanzados.

Su objetivo es crear interfaces más naturales y eficientes para la interacción entre humanos y máquinas.

Aplicaciones:

- ▶ Asistentes de voz: sistemas como Alexa y Google Assistant que mejoran continuamente en el reconocimiento de voz y la comprensión del lenguaje natural, permitiendo interacciones más fluidas y naturales.
- ▶ Realidad aumentada y virtual: interfaces inmersivas que se utilizan en educación, entretenimiento y formación profesional, ofreciendo experiencias interactivas y atractivas.
- ▶ Robótica colaborativa: robots diseñados para trabajar junto a humanos en entornos compartidos, mejorando la eficiencia y seguridad en sectores como la manufactura y la logística.

## Resolución de problemas complejos

La capacidad de la IA para analizar grandes volúmenes de datos y realizar cálculos avanzados permite abordar problemas complejos en diversas áreas. Esto incluye desde modelar sistemas climáticos hasta personalizar tratamientos médicos, ofreciendo soluciones innovadoras a desafíos críticos.

Su objetivo es abordar problemas complejos que son difíciles o imposibles de resolver mediante métodos tradicionales.

## Aplicaciones:

- ▶ Ciencias ambientales: modelado y predicción de fenómenos climáticos para entender mejor el cambio climático y desarrollar estrategias de mitigación.
- ▶ Biomedicina: análisis de datos genómicos y moleculares para identificar patrones y desarrollar tratamientos personalizados basados en las características individuales de cada paciente.
- ▶ Seguridad: detección y prevención de ciberataques mediante el análisis de patrones de comportamiento y la implementación de respuestas automáticas.

## Optimización de recursos

La IA permite una gestión más eficiente de los recursos mediante la optimización de procesos y la predicción de necesidades. Esto no solo reduce costes, sino que también mejora la sostenibilidad y el rendimiento en diversas industrias.

Su objetivo es maximizar la eficiencia en el uso de recursos para reducir costes y mejorar el rendimiento.

## Aplicaciones:

- ▶ Logística: optimización de rutas de transporte, gestión de inventarios y reducción de tiempos de entrega mediante el uso de algoritmos de optimización.
- ▶ Agricultura: implementación de agricultura de precisión utilizando drones y sensores para monitorear cultivos y optimizar el uso de agua, fertilizantes y pesticidas.
- ▶ Energía: gestión inteligente de redes eléctricas que permite una distribución eficiente de la energía, integración de fuentes renovables y reducción de pérdidas.

## Métodos de enseñanza:

- ▶ Estudio de casos: análisis de ejemplos reales de cómo la IA ha logrado estos objetivos en diversas industrias.
- ▶ Discusión y debate: reflexión sobre los beneficios y posibles desventajas de la implementación de la IA.
- ▶ Proyectos prácticos: desarrollo de pequeños proyectos que aborden uno o más de los objetivos mencionados utilizando herramientas de IA.

## Evaluación:

- ▶ Exámenes teóricos: para evaluar la comprensión de los objetivos y su importancia.
- ▶ Proyectos prácticos: implementación de soluciones de IA que demuestren cómo se pueden lograr estos objetivos en situaciones reales.
- ▶ Ensayos críticos: reflexiones sobre el impacto de la IA en la sociedad y su potencial futuro.
- ▶ Esta estructura detallada nos permitirá comprender no solo los objetivos fundamentales de la inteligencia artificial, sino también cómo estos objetivos se materializan en aplicaciones prácticas que benefician a diversos sectores y mejoran la vida cotidiana.

## 10.3. Datos masivos en la Inteligencia Artificial

En la era digital, la **inteligencia artificial** (IA) y los **datos masivos** (*big data*) se han convertido en **componentes esenciales para la innovación y la mejora de procesos** en diversas industrias. Los datos masivos son grandes volúmenes de información que pueden ser estructurados, semiestructurados o no estructurados y que, debido a su tamaño y complejidad, requieren tecnologías avanzadas para su procesamiento y análisis. Este apartado se centrará en cómo los datos masivos se aplican en la IA, destacando casos de uso actuales y discutiendo sus ventajas y desventajas.

La relación entre datos masivos y la IA es simbiótica: los datos masivos proporcionan la materia prima necesaria para entrenar modelos de IA, mientras que los algoritmos de IA ofrecen las herramientas necesarias para analizar y extraer valor de estos datos.

A continuación, se presentan algunos ejemplos destacados de cómo se utilizan los datos masivos en la IA en diferentes sectores:

### Salud:

- ▶ Predicción y diagnóstico: los sistemas de IA analizan vastas cantidades de datos médicos, desde historiales clínicos hasta imágenes de diagnóstico, para identificar patrones y predecir enfermedades. Por ejemplo, Google Health ha desarrollado modelos de IA capaces de detectar enfermedades oculares y cáncer de mama a partir de imágenes médicas con alta precisión.
- ▶ Medicina personalizada: la integración de datos genómicos y clínicos permite a los sistemas de IA desarrollar tratamientos personalizados para pacientes, optimizando así la eficacia de las terapias.

## **Finanzas:**

- ▶ Detección de fraudes: los algoritmos de IA analizan transacciones en tiempo real, identificando comportamientos anómalos que podrían indicar fraude. Esto es esencial para bancos y empresas de tarjetas de crédito como JP Morgan y Visa, que protegen así a sus clientes y sus activos.
- ▶ Análisis de riesgos: las instituciones financieras utilizan datos masivos para evaluar riesgos de crédito y de inversión, mejorando la toma de decisiones y minimizando pérdidas.

## **Marketing y publicidad:**

- ▶ Personalización: empresas como Amazon y Netflix analizan los datos de comportamiento de los usuarios para recomendar productos y contenidos personalizados. Esto no solo mejora la experiencia del cliente, sino que también incrementa las ventas y la fidelidad del cliente.
- ▶ Análisis de sentimientos: herramientas de IA analizan grandes volúmenes de datos de redes sociales para evaluar la percepción pública de marcas y productos, permitiendo a las empresas ajustar sus estrategias de marketing en tiempo real.

## **Transporte y logística:**

- ▶ Optimización de rutas: empresas como UPS utilizan datos masivos combinados con IA para optimizar las rutas de entrega, reduciendo costes y mejorando la eficiencia operativa. El sistema ORION de UPS, por ejemplo, ha ahorrado millones de galones de combustible al año.
- ▶ Mantenimiento predictivo: en la industria del transporte, los datos de sensores y el análisis predictivo permiten anticipar fallos mecánicos, reduciendo el tiempo de inactividad y los costes de reparación.

## Ventajas y desventajas

El uso de datos masivos en la IA ofrece numerosas ventajas, pero también presenta desafíos significativos.

Entre las primeras, podemos encontrar:

- ▶ Mejora de precisión y eficiencia: los datos masivos permiten entrenar modelos de IA con gran precisión, mejorando la toma de decisiones y optimizando procesos.
- ▶ Personalización: la capacidad de analizar datos detallados permite ofrecer productos y servicios personalizados, mejorando la satisfacción del cliente.
- ▶ Identificación de patrones y tendencias: los datos masivos permiten descubrir insights que no serían visibles en conjuntos de datos más pequeños, facilitando la innovación y la toma de decisiones estratégicas.

Entre sus desventajas:

- ▶ Privacidad y seguridad: el manejo de grandes volúmenes de datos plantea riesgos significativos para la privacidad y la seguridad de la información, requiriendo robustos mecanismos de protección y cumplimiento normativo.
- ▶ Costos y complejidad: el almacenamiento, procesamiento y análisis de datos masivos requieren inversiones significativas en infraestructura y tecnología, además de habilidades especializadas.
- ▶ Sesgo de datos: los conjuntos de datos masivos pueden contener sesgos que, si no se gestionan adecuadamente, pueden llevar a decisiones erróneas o injustas. Es crucial garantizar la calidad y representatividad de los datos utilizados.

La integración de datos masivos en la inteligencia artificial está **transformando industrias** y ofreciendo nuevas oportunidades para la innovación y la eficiencia. Sin embargo, también plantea desafíos que deben ser abordados para maximizar los beneficios y minimizar los riesgos. La clave está en **equilibrar las ventajas con una gestión adecuada de las desventajas**, aprovechando al máximo el potencial de los datos masivos en la IA para crear soluciones que mejoren la vida cotidiana y la operatividad empresarial.

## 10.4. La visión, el lenguaje natural y conocimiento

La inteligencia artificial (IA) ha logrado avances significativos en tres áreas clave: la visión por computadora, el procesamiento del lenguaje natural (NLP) y la representación del conocimiento. Estas tecnologías permiten a los sistemas de IA interpretar y comprender el mundo visual, comunicarse en lenguaje humano y organizar información de manera eficiente. En este apartado, exploraremos aplicaciones prácticas y actuales de estas tecnologías, destacando sus beneficios y los desafíos que enfrentan.

### Visión por computadora

La visión por computadora permite a los sistemas de IA interpretar y analizar información visual del entorno. Esta tecnología se basa en algoritmos avanzados y redes neuronales convolucionales (CNN) para realizar tareas como el reconocimiento de objetos y la navegación autónoma. Entre sus aplicaciones prácticas se encuentra la automatización industrial, donde los sistemas de inspección visual en fábricas detectan defectos en productos, mejorando la calidad y reduciendo desperdicios.

También se aplica en la agricultura de precisión, con drones equipados con cámaras que monitorean cultivos, identifican plagas y optimizan el uso de recursos agrícolas. En seguridad pública, las cámaras de vigilancia inteligentes pueden detectar comportamientos sospechosos y alertar a las autoridades en tiempo real.

Las principales ventajas de la visión por computadora incluyen su capacidad para analizar imágenes y videos con gran precisión y velocidad, superando las capacidades humanas en muchos casos. Esto permite la automatización de procesos visuales repetitivos, liberando a los trabajadores para realizar tareas más complejas. Sin embargo, también presenta desafíos, como las preocupaciones sobre la privacidad debido al uso extensivo de cámaras, y la necesidad de grandes volúmenes de datos etiquetados para entrenar los modelos con precisión.

## Procesamiento del lenguaje natural (NLP)

El procesamiento del lenguaje natural (NLP) permite a las máquinas comprender, interpretar y generar lenguaje humano, lo cual es crucial para mejorar la interacción entre humanos y máquinas. Entre las aplicaciones más comunes se encuentran los asistentes virtuales, como Google Assistant y Amazon Alexa, que responden a comandos de voz y realizan tareas como controlar dispositivos inteligentes y proporcionar información.

En el ámbito del servicio al cliente, los chatbots manejan consultas y soporte, mejorando la eficiencia y disponibilidad del servicio. Además, las herramientas de análisis de texto se utilizan para analizar grandes volúmenes de texto y extraer *insights* valiosos, como en el análisis de opiniones en redes sociales para estrategias de *marketing*.

El NLP facilita una comunicación más intuitiva y eficiente entre humanos y máquinas, permitiendo a las empresas analizar datos no estructurados y extraer información útil. Sin embargo, la complejidad del lenguaje humano, con sus sutilezas y contextos variados, sigue siendo un desafío. Además, los modelos de NLP pueden perpetuar sesgos presentes en los datos de entrenamiento, lo que requiere una gestión cuidadosa.

## Representación y gestión del conocimiento

La representación del conocimiento implica la organización y estructuración de información de manera que pueda ser utilizada eficientemente por los sistemas de IA. Esto incluye la creación de ontologías, bases de conocimiento y grafos de conocimiento que permiten a las máquinas comprender relaciones complejas entre datos. Los motores de búsqueda, como Google, utilizan grafos de conocimiento para mejorar la precisión de los resultados, ofreciendo información contextual y relevante.

En el ámbito médico, los sistemas expertos asisten en el diagnóstico y tratamiento de enfermedades mediante bases de conocimiento especializadas. Además, plataformas educativas como Coursera y Khan Academy utilizan estas representaciones para personalizar el aprendizaje y recomendar contenidos específicos a los estudiantes.

La principal ventaja de la representación del conocimiento es la eficiencia en la recuperación de información, lo que mejora la capacidad de los sistemas para encontrar y utilizar datos relevantes rápidamente. Esto es especialmente útil en la toma de decisiones complejas, donde la integración de vastos conjuntos de datos interrelacionados es crucial. Sin embargo, la creación y mantenimiento de estas bases de conocimiento pueden ser laboriosos y costosos. También, la gestión de la ambigüedad y la resolución de conflictos en la información representan desafíos continuos.

La integración de la visión por computadora, el procesamiento del lenguaje natural y la representación del conocimiento en la inteligencia artificial ha permitido avances significativos en diversas aplicaciones prácticas. Estas tecnologías presentan tanto ventajas como desafíos, pero su desarrollo continuo promete transformar aún más la interacción humano-máquina y la capacidad de los sistemas de IA para comprender y responder al mundo de manera más efectiva y eficiente.

## 10.5. Ejemplos de aplicación de la Inteligencia Artificial

### IA Generativa: un ejemplo actual de la aplicación de la IA

La inteligencia artificial generativa, una rama avanzada de la IA, ha emergido como una tecnología transformadora capaz de crear contenido nuevo a partir de datos existentes. Esta tecnología utiliza modelos como las redes generativas adversarias (GANs) y transformadores como GPT-3 y DALL-E para generar textos, imágenes, música y más. En este apartado, exploraremos algunas de las aplicaciones más destacadas de la IA generativa, destacando sus elementos principales y el impacto que están teniendo en diversos sectores.

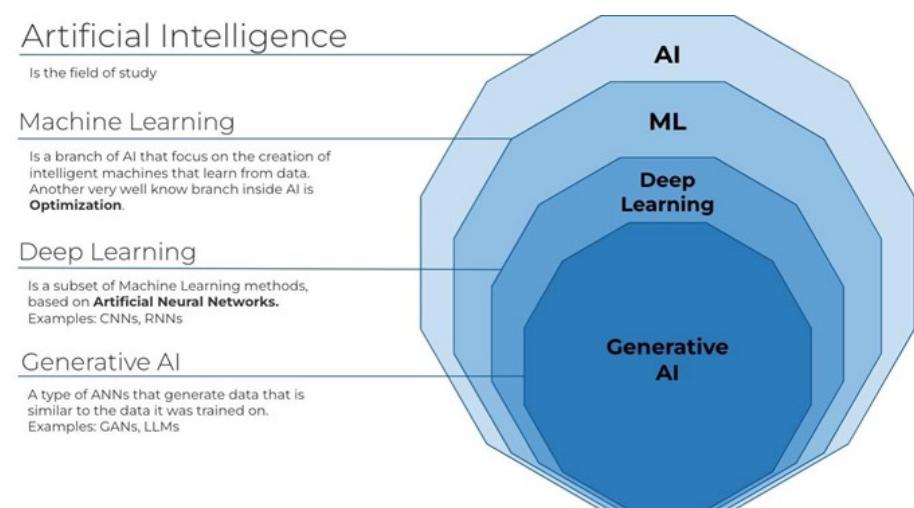


Figura 1. Esquema general de la IA (Fuente: tomado de <https://gcloud.devoteam.com/es/blog/creatividad-sin-lmites-como-la-ia-generativa-esta-transformando-el-mundo-de-la-innovacion/>)

### Diferentes aplicaciones

#### Generación de textos

La IA generativa ha revolucionado la creación de contenido textual. Modelos como GPT-3, desarrollado por OpenAI, pueden generar texto coherente y contextualmente

relevante a partir de unas pocas palabras de entrada. Estas capacidades se aplican en:

- ▶ Asistentes de redacción: herramientas como Grammarly y AI Dungeon utilizan modelos generativos para ayudar a los usuarios a escribir mejor, ofreciendo sugerencias de mejora y completando frases y párrafos.
- ▶ Contenido automatizado: medios de comunicación y blogs emplean IA generativa para producir artículos, resúmenes y noticias de manera rápida y eficiente, adaptando el estilo y tono según las necesidades específicas.

## Creación de imágenes

Modelos como DALL-E y GANs pueden generar imágenes realistas y creativas a partir de descripciones textuales. Esto tiene aplicaciones en:

- ▶ Diseño y arte: artistas y diseñadores utilizan IA generativa para crear nuevas obras de arte, explorar variaciones de diseño y generar contenido visual para proyectos multimedia.
- ▶ Publicidad y *marketing*: agencias de publicidad emplean estas tecnologías para crear imágenes personalizadas y visualmente atractivas para campañas publicitarias.

## Música y Audio

La IA generativa también se aplica en la creación de música y contenido de audio. Modelos como Jukedeck y OpenAI's MuseNet pueden componer música en varios estilos y géneros, ofreciendo nuevas herramientas para:

- ▶ Compositores y productores: músicos y productores utilizan IA generativa para inspirarse, experimentar con nuevos sonidos y generar pistas de acompañamiento.
- ▶ Entretenimiento y medios: plataformas de *streaming* y videojuegos integran música generada por IA para personalizar la experiencia del usuario y crear ambientes sonoros únicos.

## Modelado 3D y animación

La generación de contenido 3D ha sido potenciada por la IA generativa, permitiendo la creación de modelos y animaciones complejas con menor esfuerzo humano. Esto es particularmente útil en:

- ▶ Cine y videojuegos: estudios de animación y desarrolladores de videojuegos utilizan IA para crear personajes, escenarios y efectos visuales detallados, reduciendo el tiempo y los costos de producción.
- ▶ Arquitectura y diseño: arquitectos y diseñadores utilizan modelos generativos para visualizar y iterar rápidamente sobre diseños de edificios y espacios.

## Elementos principales de la IA generativa

La IA generativa se basa en varios elementos clave que permiten su funcionamiento y aplicación:

- ▶ Modelos de aprendizaje profundo: redes neuronales avanzadas, como las GANs y los transformadores, son fundamentales para la generación de contenido. Estos modelos aprenden de grandes volúmenes de datos y pueden generar nuevas muestras que son coherentes con los datos de entrenamiento.
- ▶ Entrenamiento y *fine-tuning*: los modelos generativos requieren entrenamiento en vastos conjuntos de datos y ajustes finos (*fine-tuning*) para especializarlos en tareas específicas, como la generación de texto o imágenes en un estilo particular.
- ▶ Interfaz de usuario y experiencia: la integración de IA generativa en aplicaciones prácticas requiere interfaces intuitivas que permitan a los usuarios interactuar fácilmente con la tecnología, proporcionando entradas y obteniendo resultados deseados.

## Impacto y futuro de la IA generativa

La IA generativa está transformando diversas industrias al automatizar la creación de contenido y ofrecer nuevas herramientas creativas. Sin embargo, también plantea desafíos, como la necesidad de gestionar el uso ético de la tecnología, evitar la generación de contenido malicioso o engañoso, y garantizar que las creaciones de IA respeten los derechos de propiedad intelectual.

En el futuro, es probable que veamos una mayor integración de la IA generativa en aplicaciones cotidianas, desde asistentes personales hasta herramientas profesionales en áreas creativas y técnicas. Con avances continuos en la tecnología, la capacidad de la IA para generar contenido nuevo y útil seguirá expandiéndose, ofreciendo posibilidades ilimitadas para la innovación y la creatividad.

En conclusión, la IA generativa ejemplifica el impacto actual de la inteligencia artificial en la creación de contenido. Sus aplicaciones prácticas están revolucionando sectores como el diseño, la música, la escritura y el entretenimiento, mostrando el potencial transformador de la IA en nuestra vida diaria y en la industria creativa.

## 10.6. Referencias bibliográficas

Boston Consulting Group (BCG). (2023). *How Generative AI is Transforming Healthcare*. BCG.

Brown, T. B., et al. (2020). *Language models are few-shot learners*. *arXiv preprint arXiv:2005.14165*. Recuperado de <https://arxiv.org/abs/2005.14165>

Chapel Hill, J. (2023). *How Epic is charging ahead to bring generative AI into healthcare*. Fierce Healthcare.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24-29. doi:10.1038/s41591-018-0316-z.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). *Generative adversarial nets*. En *Advances in neural information processing systems* (pp. 2672-2680). Recuperado de <https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf>

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245), 261-266. doi:10.1126/science.aaa8685

McKinsey & Company. (2023). *Generative AI in healthcare: Emerging use for care*. McKinsey & Company.

Reddy, S. (2024). Generative AI in healthcare: An implementation science informed translational path on application, integration, and governance. *Implementation Science*, 19(27). Recuperado de <https://implementationscience.biomedcentral.com/articles/10.1186/s13012-024-01357-9>

## Redes neuronales y el aprendizaje profundo

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Recuperado de <https://www.deeplearningbook.org/>

El libro «Deep Learning» de Ian Goodfellow, Yoshua Bengio y Aaron Courville es una referencia esencial para cualquier persona interesada en el campo del aprendizaje profundo. Este libro cubre desde los fundamentos de las redes neuronales hasta las técnicas avanzadas utilizadas en la actualidad. Se incluyen temas como la retropropagación, las redes convolucionales, y las redes recurrentes, así como aplicaciones prácticas en áreas como el reconocimiento de voz y la visión por computadora. El recurso es altamente recomendado para aquellos que buscan una comprensión profunda y detallada del aprendizaje profundo.

## Procesamiento del Lenguaje Natural con Redes Neuronales

Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3rd ed.). Pearson. Recuperado de <https://web.stanford.edu/~jurafsky/slp3/>

«Speech and Language Processing» es una obra fundamental escrita por Daniel Jurafsky y James H. Martin. Este libro ofrece una cobertura exhaustiva del procesamiento del lenguaje natural (NLP) con un enfoque particular en el uso de redes neuronales y técnicas de aprendizaje profundo. Los autores explican cómo los modelos de lenguaje, las redes de traducción automática y las herramientas de análisis de sentimientos se desarrollan y aplican en la práctica. Este recurso es invaluable para estudiantes y profesionales que desean profundizar en el uso de IA en el procesamiento del lenguaje.

1. ¿Cuál es una de las principales aplicaciones de la IA generativa en el campo de la salud?

  - A. Diagnóstico de enfermedades utilizando imágenes médicas.
  - B. Creación de algoritmos de búsqueda.
  - C. Desarrollo de videojuegos.
  - D. Automatización de procesos financieros.
  
2. ¿Cuál de las siguientes es una ventaja clave de la visión por computadora?

  - A. Requiere menos datos para entrenar modelos precisos.
  - B. Puede analizar imágenes y videos con gran precisión y velocidad.
  - C. Es menos costosa que otros métodos de IA.
  - D. No plantea preocupaciones de privacidad.
  
3. ¿Qué tecnología se utiliza comúnmente en el procesamiento del lenguaje natural para comprender el contexto de las palabras en una oración?

  - A. Redes convolucionales.
  - B. Algoritmos genéticos.
  - C. Redes neuronales recurrentes.
  - D. Redes de retropropagación.
  
4. ¿Cuál es uno de los desafíos principales de los modelos de IA generativa?

  - A. Requieren una infraestructura de hardware muy simple.
  - B. No pueden generar datos sintéticos.
  - C. Pueden perpetuar sesgos presentes en los datos de entrenamiento.
  - D. Son fáciles de implementar sin conocimientos técnicos.

5. ¿Cuál de las siguientes opciones describe mejor una red generativa adversaria (GAN)?
- A. Un algoritmo que clasifica datos en categorías predefinidas
  - B. Una red neuronal que traduce texto de un idioma a otro
  - C. Un modelo de IA que genera datos nuevos y realistas mediante la competencia entre dos redes neuronales
  - D. Un sistema que optimiza rutas de transporte
6. ¿Qué es el aprendizaje supervisado en el contexto del aprendizaje automático?
- A. Un método donde la máquina aprende a partir de datos no etiquetados.
  - B. Un enfoque que no requiere datos de entrenamiento.
  - C. Un método donde la máquina aprende a partir de datos etiquetados proporcionados por humanos.
  - D. Un proceso que implica solo la optimización de *hardware*.
7. ¿Cuál es una aplicación práctica de los sistemas expertos en la medicina?
- A. Automatización de la fabricación de automóviles.
  - B. Diagnóstico de enfermedades basado en el conocimiento médico acumulado.
  - C. Desarrollo de videojuegos.
  - D. Análisis de sentimientos en redes sociales.
8. ¿Qué herramienta se utiliza comúnmente para la representación del conocimiento en IA?
- A. Redes convolucionales.
  - B. Algoritmos genéticos.
  - C. Grafos de conocimiento.
  - D. Redes neuronales recurrentes.

- 9.** ¿Cuál de las siguientes es una desventaja de la IA generativa en el ámbito de la privacidad?
- A. Requiere menos datos para entrenar modelos precisos.
  - B. Puede generar contenido malicioso o engañoso.
  - C. Facilita la creación de modelos predictivos.
  - D. Mejora la personalización de servicios.
- 10.** ¿Qué ventaja ofrece la personalización a través del uso de IA en el *marketing*?
- A. Reducción de costos de producción.
  - B. Mejora en la precisión del diagnóstico médico.
  - C. Incremento en la satisfacción y fidelidad del cliente.
  - D. Optimización de rutas logísticas.

Ciencia de Datos Aplicada

---

# Tema 11. Aplicaciones en Industria 4.0

# Índice

[Esquema](#)

[Ideas clave](#)

[11.1. Introducción y objetivos](#)

[11.2. Conceptos de la Industria 4.0](#)

[11.3. Objetivos de la Industria 4.0](#)

[11.4. Datos masivos en la industria 4.0](#)

[11.5. Cadena de suministros, producción y distribución](#)

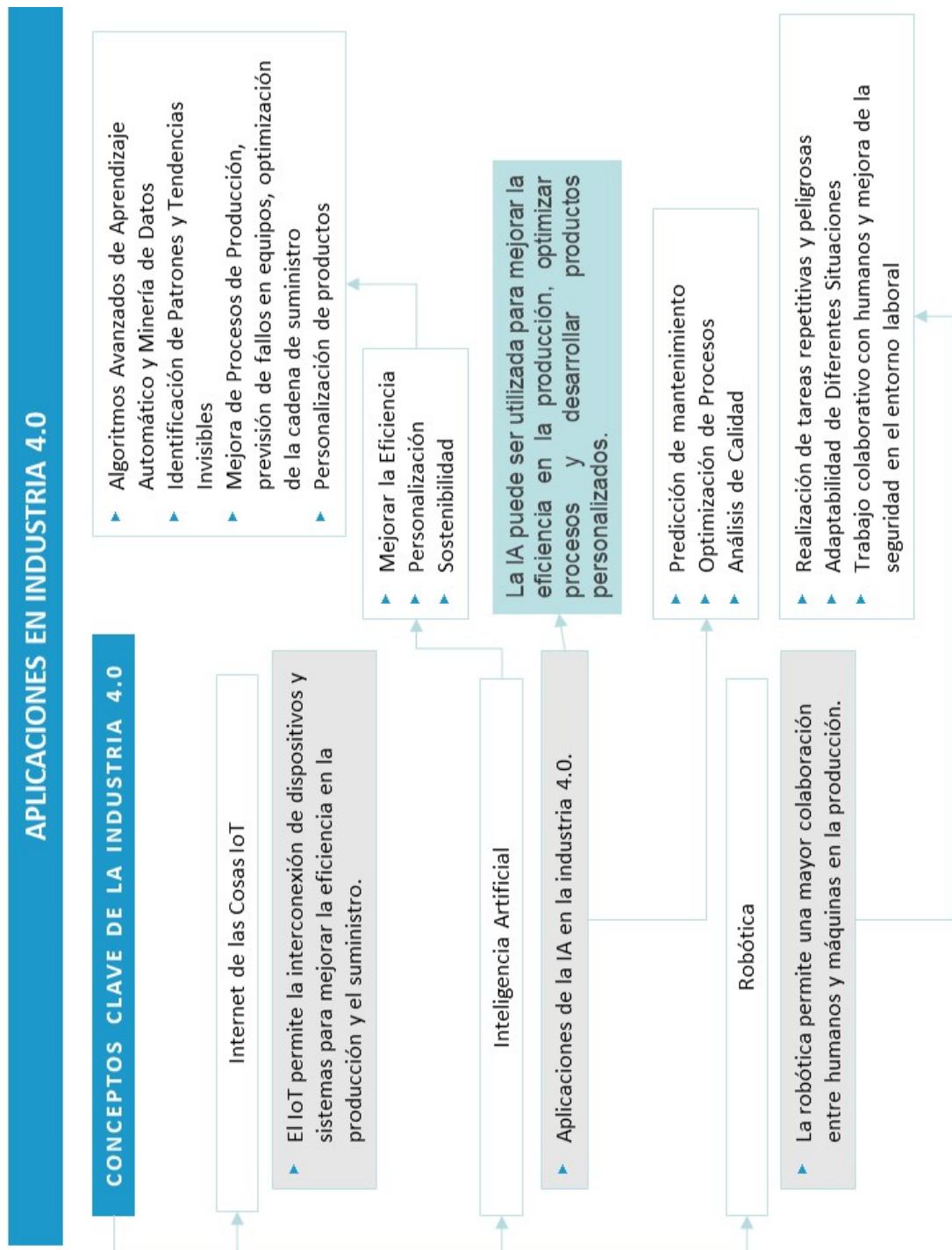
[11.6. Referencias bibliográficas](#)

[A fondo](#)

[La cuarta revolución industrial](#)

[Industry 4.0: Managing The Digital Transformation](#)

[Test](#)



## 11.1. Introducción y objetivos

La Industria 4.0 es una revolución industrial que combina tecnologías emergentes como Internet de las cosas (IoT), inteligencia artificial (IA), robótica y análisis de datos para mejorar la eficiencia y la productividad en la producción y el suministro. Esta transformación industrial busca crear un entorno más flexible, personalizado y sostenible, donde la automatización y la interconexión de dispositivos y sistemas permiten una mayor colaboración entre humanos y máquinas.

En la actualidad, la Industria 4.0 es un tema de gran interés y debate en el ámbito académico y empresarial. La adopción de tecnologías como la IA y el análisis de datos ha aumentado significativamente en los últimos años y se espera que siga creciendo en el futuro. Sin embargo, también existen retos significativos que deben ser abordados, como la seguridad de los datos, la privacidad de los usuarios y la necesidad de capacitación para los trabajadores.

La IA, en particular, es un tema de gran interés en la Industria 4.0 debido a su capacidad para analizar grandes cantidades de datos y tomar decisiones informadas. La IA puede ser utilizada para mejorar la eficiencia en la producción, optimizar procesos y reducir costos. Además, la IA puede ser utilizada para desarrollar productos y servicios personalizados, lo que puede ser un factor clave para la competencia en el mercado.

Sin embargo, la adopción de la IA también plantea desafíos significativos. Por ejemplo, la IA puede reemplazar a los trabajadores en ciertas tareas, lo que puede tener un impacto negativo en el empleo. Además, la IA puede ser utilizada de manera inapropiada, lo que puede tener consecuencias negativas para la sociedad.

Los objetivos de este tema son los siguientes:

- ▶ Comprender los conceptos clave de la Industria 4.0: entender los fundamentos de la Industria 4.0, incluyendo Internet de las cosas (IoT), inteligencia artificial (IA), robótica, y análisis de datos.
- ▶ Analizar los objetivos y beneficios de la Industria 4.0: identificar los objetivos y beneficios de la Industria 4.0, como la mejora de la eficiencia, la personalización y la sostenibilidad.
- ▶ Explorar las aplicaciones de la ciencia de datos en la Industria 4.0: verificar cómo la ciencia de datos se aplica en la Industria 4.0 para mejorar la predicción de mantenimiento, la optimización de procesos y el análisis de calidad.

Estos objetivos nos permitirán profundizar en los conceptos y aplicaciones de la Industria 4.0 y entender cómo esta revolución industrial puede transformar la producción y el suministro.

## 11.2. Conceptos de la Industria 4.0

La Industria 4.0, o cuarta revolución industrial, es un término que se refiere a la transformación completa de los procesos de producción mediante la integración de tecnologías digitales avanzadas.

Esta evolución implica un cambio fundamental en cómo las fábricas y empresas manufactureras operan (ver Figura 1), promoviendo una mayor interconexión y automatización a través de sistemas ciberfísicos, el Internet de las Cosas (IoT), la inteligencia artificial (IA), y el análisis de grandes volúmenes de datos (*big data*).

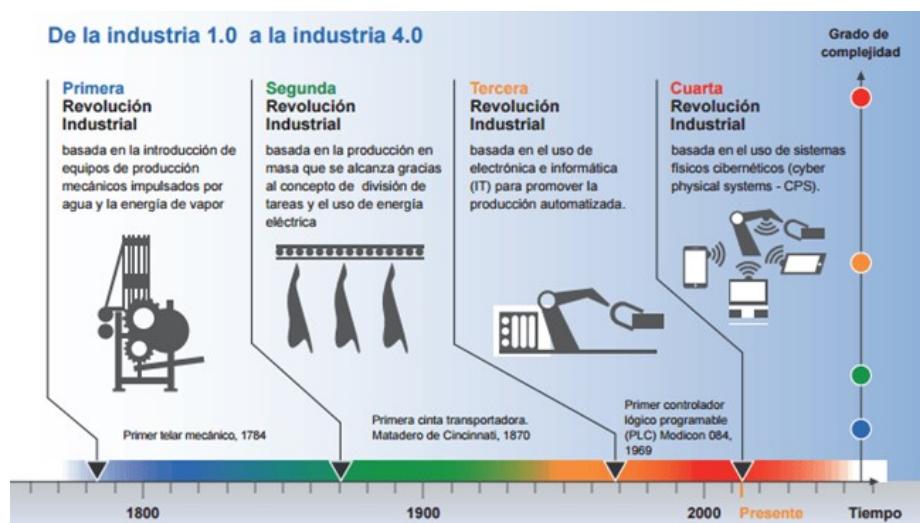


Figura 1. Evolución de la Industria 4.0 (Fuente: <https://papelesdeinteligencia.com/que-es-industria-4-0/>)

### Sistemas ciberfísicos (CPS)

Los sistemas ciberfísicos son la columna vertebral de la Industria 4.0. Estos sistemas integran el mundo físico con el digital, utilizando sensores, actuadores y redes de comunicación para recopilar y procesar datos en tiempo real. Los CPS permiten la monitorización y control de procesos físicos a través de algoritmos computacionales y modelos digitales.

Esto no solo mejora la eficiencia y precisión de los procesos industriales, sino que también facilita la creación de fábricas inteligentes donde las máquinas pueden comunicarse y coordinarse entre sí y con los humanos de manera autónoma (SAP) (papeles de inteligencia competitiva).

## Internet de las cosas industrial (IIoT)

El IIoT extiende el concepto del internet de las cosas al ámbito industrial. En este contexto, los objetos físicos como máquinas, herramientas y equipos están equipados con sensores que recopilan datos sobre su funcionamiento y estado. Estos datos son transmitidos a través de redes industriales a sistemas centralizados donde se analizan para optimizar la producción, realizar mantenimiento predictivo y mejorar la toma de decisiones.

La capacidad de monitorear y analizar datos en tiempo real permite a las empresas industriales reaccionar rápidamente a problemas y cambios en las condiciones operativas, mejorando la eficiencia y reduciendo el tiempo de inactividad (Grupo Novatech) (InnovaciónDigital360).

## *Big data*

### y análisis de datos

El análisis de grandes volúmenes de datos es otro pilar crucial de la Industria 4.0. La cantidad de datos generados por los sensores y dispositivos en una fábrica inteligente es inmensa. *Big data* se refiere a la capacidad de manejar y analizar estos datos para extraer información valiosa que puede utilizarse para mejorar los procesos de producción, prever fallos en los equipos, optimizar la cadena de suministro, y personalizar productos según las preferencias de los clientes.

Las herramientas de análisis de *big data* utilizan algoritmos avanzados de aprendizaje automático y minería de datos para identificar patrones y tendencias que de otra manera serían invisibles (Papeles de Inteligencia Competitiva).

## Robótica y automatización avanzada

La robótica avanzada y la automatización son elementos centrales en la Industria 4.0. Los robots modernos no solo realizan tareas repetitivas y peligrosas, sino que también son capaces de adaptarse a diferentes situaciones gracias a la inteligencia artificial. Los robots colaborativos, o cobots, pueden trabajar junto a los humanos, asistiendo en tareas complejas y mejorando la seguridad y la eficiencia en el entorno de trabajo. La automatización avanzada permite a las fábricas reducir costos, aumentar la producción y mejorar la calidad del producto final (SAP) (InnovaciónDigital360).

## Simulación y gemelos digitales

La simulación y los gemelos digitales son herramientas poderosas en la Industria 4.0. Un gemelo digital es una réplica virtual de un producto, proceso o sistema físico que se actualiza en tiempo real con datos del mundo real. Esto permite a las empresas realizar simulaciones y pruebas virtuales antes de implementar cambios en el entorno físico, reduciendo riesgos y costos. La simulación también se utiliza para optimizar procesos de producción y prever el comportamiento de sistemas complejos bajo diferentes condiciones (Grupo Novatech) (UNIR).

## Ciberseguridad

Con el aumento de la conectividad y la dependencia de los datos, la ciberseguridad se convierte en una preocupación primordial en la Industria 4.0. La protección de datos sensibles y la prevención de ataques ciberneticos son esenciales para mantener la integridad y continuidad de las operaciones industriales. Las empresas implementan tecnologías avanzadas como blockchain, aprendizaje automático y arquitecturas de confianza cero para proteger sus sistemas contra amenazas y garantizar la seguridad de la información (SAP) (Papeles de Inteligencia Competitiva).

## Integración vertical y horizontal

La integración vertical y horizontal se refiere a la conectividad y coordinación de todos los niveles de la producción, desde el piso de la fábrica hasta los sistemas de gestión empresarial, y a lo largo de toda la cadena de valor, incluyendo proveedores y clientes. La integración vertical permite una mayor eficiencia y control de los procesos internos, mientras que la integración horizontal facilita la colaboración y el intercambio de información entre diferentes actores de la cadena de suministro, mejorando la planificación y la respuesta a la demanda del mercado (UNIR) (Papeles de Inteligencia Competitiva).

## Fabricación aditiva

La fabricación aditiva, o impresión 3D, es una tecnología emergente que está revolucionando la producción industrial. Permite la creación de piezas y componentes a partir de modelos digitales, capa por capa, sin necesidad de moldes o herramientas especiales. Esto no solo reduce los costos y tiempos de producción, sino que también permite una mayor personalización y la fabricación de geometrías complejas que serían imposibles de producir con métodos tradicionales (Papeles de Inteligencia Competitiva).

La Industria 4.0 es una revolución que transforma radicalmente la manufactura mediante la integración de tecnologías digitales avanzadas. Los conceptos fundamentales incluyen sistemas ciberfísicos, IoT, big data, robótica avanzada, simulación, ciberseguridad, y la integración vertical y horizontal. Juntas, estas tecnologías permiten la creación de fábricas inteligentes que son más eficientes, flexibles y capaces de adaptarse rápidamente a las cambiantes demandas del mercado (SAP) (Grupo Novatech) (InnovaciónDigital360) (UNIR) (Papeles de Inteligencia Competitiva).

## 11.3. Objetivos de la Industria 4.0

La Industria 4.0 se centra en la transformación digital de la manufactura y otros sectores industriales mediante la integración de tecnologías avanzadas. Los principales objetivos de la Industria 4.0 son mejorar la eficiencia operativa, aumentar la flexibilidad, mejorar la calidad de los productos y servicios, y desarrollar nuevos modelos de negocio que aprovechen la conectividad y la automatización. A continuación, se detallan estos objetivos fundamentales.

La mejora de la eficiencia operativa es uno de los objetivos primordiales de la Industria 4.0. La implementación de tecnologías como los sistemas ciberfísicos (CPS), el Internet de las Cosas (IoT) y la inteligencia artificial (IA) permite la monitorización y control en tiempo real de los procesos productivos. Estos avances tecnológicos facilitan la optimización de los recursos, la reducción de desperdicios y la minimización de tiempos de inactividad mediante el mantenimiento predictivo. Las fábricas inteligentes, equipadas con sensores y conectadas a través de redes, pueden ajustar sus operaciones de manera autónoma para maximizar la productividad y la eficiencia.

La capacidad de adaptarse rápidamente a las cambiantes demandas del mercado es crucial en la Industria 4.0. Este objetivo se logra mediante la integración de sistemas flexibles y escalables que permiten la personalización masiva de productos. La digitalización y la interconexión de los procesos de producción permiten a las empresas cambiar rápidamente las líneas de producción y responder de manera ágil a las necesidades de los clientes. La fabricación aditiva, también conocida como impresión 3D, es un ejemplo de cómo la tecnología puede proporcionar la flexibilidad necesaria para producir piezas personalizadas y responder rápidamente a las demandas del mercado.

La Industria 4.0 busca mejorar la calidad de los productos y servicios mediante el uso de tecnologías avanzadas para la monitorización y control de los procesos de producción. Los sistemas de control de calidad automatizados, equipados con sensores y algoritmos de aprendizaje automático, pueden detectar defectos en tiempo real y corregir problemas antes de que afecten la producción. Además, los gemelos digitales permiten la simulación y optimización de procesos y productos antes de su implementación en el mundo real, reduciendo así los errores y mejorando la calidad general.

La conectividad y la automatización proporcionadas por la Industria 4.0 abren nuevas oportunidades para el desarrollo de modelos de negocio innovadores. Las empresas pueden ofrecer productos y servicios basados en datos, como el mantenimiento predictivo y los servicios personalizados, que mejoran la experiencia del cliente y crean nuevas fuentes de ingresos. La integración de la cadena de suministro digital permite una colaboración más estrecha entre proveedores, fabricantes y clientes, optimizando la logística y reduciendo costos. Además, la economía de la suscripción y los modelos de negocio basados en plataformas digitales están emergiendo como formas efectivas de aprovechar las capacidades de la Industria 4.0.

Los objetivos de la Industria 4.0 están orientados a transformar la manufactura y otros sectores industriales mediante la implementación de tecnologías avanzadas que mejoran la eficiencia operativa, aumentan la flexibilidad, mejoran la calidad de los productos y servicios, y desarrollan nuevos modelos de negocio. La adopción de estos objetivos permitirá a las empresas mantenerse competitivas en un mercado global en constante cambio y responder de manera efectiva a las demandas de los clientes y las oportunidades del mercado.

## 11.4. Datos masivos en la industria 4.0

En la industria 4.0, los datos masivos, conocidos como *big data*, desempeñan un papel crucial en la optimización de los procesos de producción y en la toma de decisiones estratégicas. La capacidad de recopilar, almacenar y analizar grandes volúmenes de datos provenientes de diversas fuentes permite a las empresas mejorar la eficiencia operativa, predecir fallos en los equipos, personalizar productos y optimizar la cadena de suministro.

A continuación, se presentan ejemplos concretos, aplicaciones prácticas del mundo real, plataformas conocidas del mercado, y las ventajas y desventajas del uso de los datos masivos en la industria 4.0.

### Ejemplos concretos y aplicaciones prácticas

#### Mantenimiento predictivo

- ▶ Ejemplo: General Electric (GE) utiliza sensores en sus turbinas y motores aéreos para recopilar datos en tiempo real. Estos datos son analizados para predecir posibles fallos y programar el mantenimiento antes de que ocurran problemas, lo que reduce significativamente el tiempo de inactividad y los costos de reparación.
- ▶ Aplicación práctica: la integración de sensores IoT en maquinaria industrial permite monitorizar continuamente el estado de los equipos y anticipar fallos, mejorando la eficiencia y prolongando la vida útil de los activos.

#### Optimización de la cadena de suministro

- ▶ Ejemplo: Walmart utiliza análisis de *big data* para optimizar su cadena de suministro. Analiza datos de ventas, inventarios y patrones de compra para predecir la demanda de productos y ajustar el stock en tiempo real.

- ▶ Aplicación práctica: las empresas pueden utilizar datos masivos para obtener una visibilidad completa de la cadena de suministro, optimizando la logística, mejorando la precisión en la previsión de la demanda y reduciendo los costos operativos.

## Personalización de productos

- ▶ Ejemplo: Nike emplea *big data* para personalizar las experiencias de compra de sus clientes. A través del análisis de datos de comportamiento de compra y preferencias de los clientes, Nike ofrece recomendaciones personalizadas y diseños de productos a medida.
- ▶ Aplicación práctica: las empresas pueden analizar datos de clientes para ofrecer productos personalizados, mejorando la satisfacción del cliente y aumentando la lealtad a la marca.

## Plataformas conocidas del mercado

### Apache Hadoop

Es una de las plataformas más utilizadas para el almacenamiento y procesamiento de grandes volúmenes de datos. Hadoop permite a las empresas manejar datos estructurados y no estructurados a gran escala de manera eficiente.

### Apache Spark

Conocido por su velocidad y capacidad de procesamiento en tiempo real, Spark es ideal para análisis rápidos y complejos de *big data*. Es ampliamente utilizado en aplicaciones de machine learning y análisis predictivo.

### Microsoft Azure:

Azure ofrece servicios de *big data* y análisis a través de su plataforma en la nube. Permite a las empresas almacenar, procesar y analizar grandes volúmenes de datos con herramientas como Azure HDInsight y Azure Synapse Analytics.

## Google BigQuery

Es un almacén de datos completamente administrado que permite realizar consultas SQL rápidas y en tiempo real. BigQuery es ideal para el análisis de grandes conjuntos de datos y se integra con otras herramientas de Google Cloud Platform.

### Ventajas del uso de datos masivos

#### Mejora de la toma de decisiones

El análisis de grandes volúmenes de datos proporciona información valiosa que ayuda a las empresas a tomar decisiones informadas y basadas en evidencia.

#### Aumento de la eficiencia operativa

La capacidad de analizar datos en tiempo real permite optimizar los procesos de producción, reducir desperdicios y mejorar la eficiencia general.

#### Personalización y satisfacción del cliente

El análisis de datos de clientes permite ofrecer productos y servicios personalizados, mejorando la experiencia del cliente y aumentando la lealtad a la marca.

#### Reducción de costos:

Las técnicas de mantenimiento predictivo y la optimización de la cadena de suministro pueden reducir significativamente los costos operativos y de mantenimiento.

### Desventajas del uso de datos masivos

#### Complejidad y costo de implementación

La implementación de soluciones de Big Data puede ser costosa y compleja, requiriendo inversiones significativas en infraestructura y capacitación del personal.

## Problemas de privacidad y seguridad

El manejo de grandes volúmenes de datos sensibles plantea riesgos de privacidad y seguridad y requiere medidas estrictas de ciberseguridad para proteger la información.

## Calidad de los datos

La precisión de los análisis de *big data* depende de la calidad de los datos. Datos incompletos o incorrectos pueden llevar a conclusiones erróneas y decisiones ineficaces.

## Gestión del cambio

La adopción de tecnologías de *big data* requiere cambios en la cultura organizacional y en los procesos empresariales, lo que puede ser un desafío para algunas organizaciones.

En conclusión, los datos masivos son fundamentales para la Industria 4.0, ofreciendo numerosas ventajas en términos de eficiencia operativa, personalización y toma de decisiones informadas. Sin embargo, también presentan desafíos significativos que las empresas deben abordar para aprovechar plenamente su potencial.

## 11.5. Cadena de suministros, producción y distribución

La Industria 4.0 no solo transforma la producción en las fábricas, sino que también impacta significativamente la cadena de suministro y los procesos de distribución. La adopción de tecnologías avanzadas mejora la visibilidad, eficiencia y flexibilidad en toda la cadena de valor, desde la adquisición de materias primas hasta la entrega del producto final al cliente. Este apartado analiza cómo las tecnologías de la Industria 4.0 están revolucionando estos aspectos clave.

La cadena de suministro en la Industria 4.0 se caracteriza por la integración y la visibilidad en tiempo real. Las tecnologías como el Internet de las Cosas (IoT), el Big Data y la inteligencia artificial permiten a las empresas monitorear y gestionar cada eslabón de la cadena de suministro de manera más eficiente.

### Visibilidad en tiempo real

Las empresas pueden rastrear el movimiento de materias primas y productos a lo largo de la cadena de suministro mediante sensores IoT y sistemas de seguimiento. Esta visibilidad permite a las empresas anticipar y responder rápidamente a interrupciones, optimizando así el flujo de materiales y productos.

### Optimización de inventarios

El análisis de *big data* ayuda a las empresas a prever la demanda y ajustar sus niveles de inventario en consecuencia. Esto reduce los costos asociados con el exceso de inventario y minimiza las interrupciones por falta de materiales.

### Producción inteligente

La producción en la Industria 4.0 se apoya en tecnologías como los sistemas ciberfísicos (CPS), la fabricación aditiva (impresión 3D) y la robótica avanzada. Estos sistemas permiten una producción más flexible y eficiente.

## Sistemas ciberfísicos

Los CPS integran procesos físicos y computacionales, permitiendo una monitorización y control precisos de las operaciones de producción. Esto resulta en una mejora en la calidad del producto y una reducción en los tiempos de ciclo.

## Fabricación aditiva

La impresión 3D permite la producción de componentes complejos y personalizados con rapidez y precisión. Esto no solo reduce los tiempos de producción, sino que también disminuye los costos al eliminar la necesidad de herramientas y moldes específicos.

## Automatización y robótica

La robótica avanzada y los robots colaborativos (cobots) trabajan junto a los humanos en el entorno de producción, mejorando la eficiencia y reduciendo los errores. Estos robots pueden adaptarse a diferentes tareas, ofreciendo una flexibilidad significativa en las líneas de producción.

## Distribución eficiente

La distribución en la Industria 4.0 se optimiza mediante la integración de tecnologías que permiten una gestión más efectiva de la logística y la entrega.

## Logística inteligente

Las plataformas digitales y el análisis de datos permiten optimizar las rutas de entrega y gestionar la flota de transporte de manera más eficiente. Esto no solo reduce los costos de transporte, sino que también mejora los tiempos de entrega y la satisfacción del cliente.

## Automatización de almacenes

Los almacenes inteligentes utilizan robots y sistemas automatizados para gestionar el almacenamiento y la preparación de pedidos. Estos sistemas aumentan la precisión y la velocidad de las operaciones de almacén, reduciendo los errores y mejorando la eficiencia general.

## Plataformas digitales de gestión

Herramientas como los sistemas de gestión de la cadena de suministro (SCM) y los sistemas de planificación de recursos empresariales (ERP) integran y automatizan los procesos de distribución, facilitando una coordinación efectiva entre diferentes actores de la cadena de suministro.

## Ventajas de la transformación digital en la cadena de suministro, producción y distribución

### Mejora en la eficiencia y reducción de costos

La automatización y el análisis de datos permiten optimizar los procesos, reducir los tiempos de inactividad y disminuir los costos operativos.

### Mayor flexibilidad y capacidad de respuesta

Las empresas pueden adaptarse rápidamente a cambios en la demanda del mercado, personalizando productos y ajustando sus operaciones en tiempo real.

### Mejora en la calidad del producto

La monitorización en tiempo real y el control preciso de los procesos de producción resultan en productos de mayor calidad y menores tasas de defectos.

## Satisfacción del cliente

Una cadena de suministro eficiente y una distribución optimizada mejoran los tiempos de entrega y la precisión de los pedidos, aumentando la satisfacción del cliente.

## Desafíos de la transformación digital

### Inversión inicial alta

La implementación de tecnologías avanzadas requiere una inversión significativa en infraestructura y capacitación del personal.

### Complejidad en la integración

Integrar nuevas tecnologías con los sistemas existentes puede ser complejo y requerir una reestructuración de los procesos empresariales.

### Seguridad y privacidad de los datos

El manejo de grandes volúmenes de datos sensibles plantea riesgos de seguridad y privacidad, requiriendo medidas estrictas de ciberseguridad.

### Resistencia al cambio

La adopción de nuevas tecnologías y procesos puede encontrar resistencia por parte del personal, requiriendo una gestión efectiva del cambio.

La Industria 4.0 transforma la cadena de suministro, la producción y la distribución mediante la integración de tecnologías avanzadas que mejoran la eficiencia, la flexibilidad y la calidad. Aunque presenta desafíos significativos, las ventajas superan con creces las dificultades, permitiendo a las empresas mantenerse competitivas en un mercado global en constante evolución.

## 11.6. Referencias bibliográficas

Gilchrist, A. (2016). *Industry 4.0: The Industrial Internet of Things*. Apress.

Grupo Novatech. (n.d.). *Conceptos básicos de la Industria 4.0*. Recuperado de <https://www.grupo-novatech.com/conceptos-basicos-de-la-industria-4-0>.

InnovaciónDigital360. (n.d.). *Industria 4.0: qué es, en qué consiste y ejemplos*. Recuperado de <https://www.innovaciondigital360.com/industria-4-0/industria-4-0-que-es-en-que-consiste-y-ejemplos/>

Kagermann, H., Wahlster, W., & Helbig, J. (Eds.). (2013). *Recommendations for Implementing the Strategic Initiative INDUSTRIE 4.0: Securing the Future of German Manufacturing Industry. Final report of the Industrie 4.0 Working Group*. Forschungsunion.

Lee, J., Bagheri, B., & Kao, H. A. (2015). A cyber-physical systems architecture for Industry 4.0-based manufacturing systems. *Manufacturing Letters*, 3, 18-23. <https://doi.org/10.1016/j.mfglet.2014.12.001>

SAP. (n.d.). *¿Qué es la industria 4.0? Definición, tecnologías, beneficios*. Recuperado de <https://www.sap.com/products/scm/industry-4-0/what-is-industry-4-0.html>

UNIR. (n.d.). *¿Qué es la Industria 4.0? Definición y objetivos*. Recuperado de <https://www.unir.net/empresa/revista/industria-alimentaria-40/>.

Xu, L. D., Xu, E. L., & Li, L. (2018). Industry 4.0: state of the art and future trends. *International Journal of Production Research*, 56(8), 2941-2962. <https://doi.org/10.1080/00207543.2018.1444806>.

## La cuarta revolución industrial

Schwab, K. (2017). *The Fourth Industrial Revolution*. Crown Business. Recuperado de <https://www.weforum.org/about/the-fourth-industrial-revolution-by-klaus-schwab>

El libro *The Fourth Industrial Revolution* de Klaus Schwab ofrece una visión integral de la transformación que la Industria 4.0 está trayendo a la sociedad. Schwab, fundador del Foro Económico Mundial, explora cómo las tecnologías emergentes están fusionando el mundo físico, digital y biológico. El texto abarca una amplia gama de temas, desde la inteligencia artificial y la robótica hasta la biotecnología y la computación cuántica, destacando el impacto de estos avances en la economía global, el mercado laboral y la vida cotidiana.

## Industry 4.0: Managing The Digital Transformation

---

Oswald, G., & Kleinemeier, M. (Eds.). (2016). *Industry 4.0: Managing The Digital Transformation*. Springer. Recuperado de <https://www.mintur.gob.es/Publicaciones/Publicacionesperiodicas/EconomiaIndustrial/RevistaEconomiaIndustrial/410/PRIMERA%20CR%C3%8DTICA.pdf>

---

*Industry 4.0: Managing The Digital Transformation*, editado por Gerhard Oswald y Michael Kleinemeier, es una colección de ensayos que exploran cómo las empresas pueden gestionar la transformación digital impulsada por la Industria 4.0. El libro ofrece una visión detallada de las tecnologías clave, los modelos de negocio y las estrategias de gestión que las empresas pueden emplear para aprovechar las oportunidades de la digitalización. Los autores destacan casos de estudio y prácticas recomendadas para guiar a los líderes empresariales en la adaptación a este nuevo paradigma industrial.

1. ¿Cuál de las siguientes tecnologías es un pilar fundamental de la Industria 4.0?

  - A. Blockchain.
  - B. Realidad Virtual.
  - C. Sistemas Ciberfísicos.
  - D. Redes Sociales.
  
2. ¿Qué objetivo principal tiene el uso de Big Data en la Industria 4.0?

  - A. Reducir el uso de tecnología.
  - B. Incrementar el volumen de producción.
  - C. Optimizar procesos de producción y mantenimiento.
  - D. Limitar la conectividad de dispositivos.
  
3. ¿Cuál es una aplicación práctica de la fabricación aditiva en la Industria 4.0?

  - A. Producción en masa de productos idénticos.
  - B. Creación de piezas personalizadas y prototipos.
  - C. Aumentar los costos de producción.
  - D. Reducción de la conectividad entre dispositivos.
  
4. ¿Qué plataforma es ampliamente utilizada para el almacenamiento y procesamiento de grandes volúmenes de datos en la Industria 4.0?

  - A. Apache Hadoop.
  - B. Microsoft Excel.
  - C. Facebook.
  - D. Instagram.

5. ¿Cuál es una ventaja clave del uso de robots colaborativos (cobots) en la producción industrial?
  - A. Incremento de errores en la producción.
  - B. Eliminación de la intervención humana.
  - C. Mejora de la eficiencia y reducción de errores.
  - D. Disminución de la calidad del producto.
  
6. ¿Cuál de las siguientes es una plataforma conocida para el análisis de *big data* en tiempo real?
  - A. Microsoft Word.
  - B. Apache Spark.
  - C. Adobe Photoshop.
  - D. Google Docs.
  
7. ¿Qué tecnología permite la creación de réplicas virtuales de productos y procesos en la Industria 4.0?
  - A. Realidad Virtual.
  - B. Gemelos Digitales.
  - C. Redes Sociales.
  - D. Computación Cuántica.
  
8. ¿Cuál es una desventaja significativa del uso de Big Data en la Industria 4.0?
  - A. Reducción de costos operativos.
  - B. Mejora en la toma de decisiones.
  - C. Complejidad y costo de implementación.
  - D. Aumento de la eficiencia operativa.

**9.** ¿Qué ventaja proporciona la integración de la cadena de suministro digital en la Industria 4.0?

- A. Aumento de la burocracia.
- B. Reducción de la visibilidad de los procesos.
- C. Mejora de la coordinación y optimización logística.
- D. Disminución de la capacidad de respuesta.

**10.** ¿Qué herramienta en la nube es conocida por permitir consultas SQL rápidas y en tiempo real para grandes conjuntos de datos?

- A. Microsoft Azure.
- B. Google BigQuery.
- C. Dropbox.
- D. iCloud.

Ciencia de Datos Aplicada

---

# Tema 12. Nuevas aplicaciones y tendencias

# Índice

[Esquema](#)

[Ideas clave](#)

[12.1. Introducción y objetivos](#)

[12.2. Internet de las cosas](#)

[12.3. La importancia de contar historias con tus datos o data storytelling](#)

[12.4. Tecnología blockchain](#)

[12.5. Referencias bibliográficas](#)

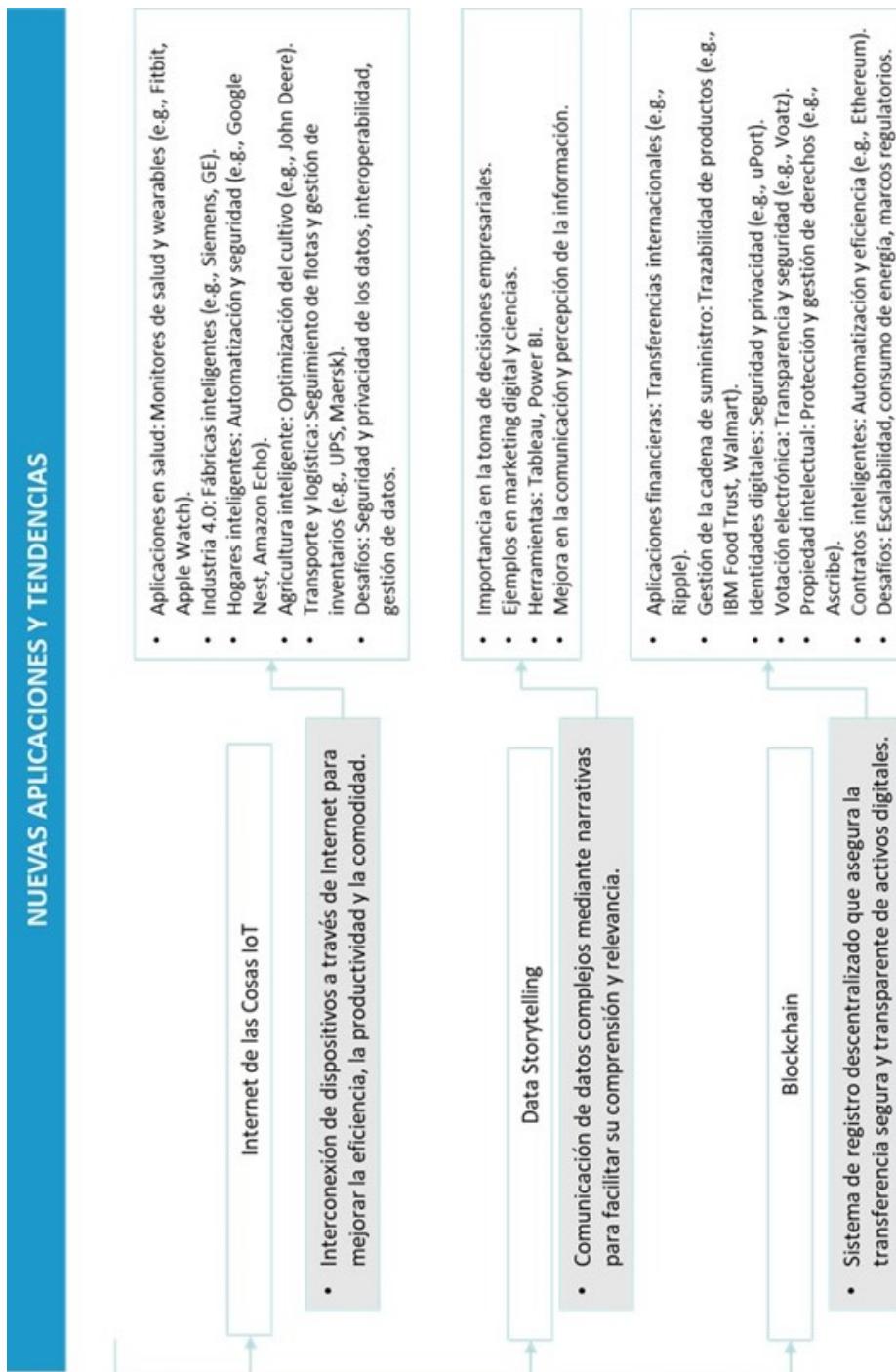
[A fondo](#)

[Internet de las Cosas \(IoT\)](#)

[A Survey on Enabling Technologies, Protocols, and Applications](#)

[The Business Blockchain: Promise, Practice, and the Application of the Next Internet Technology](#)

[Test](#)



## 12.1. Introducción y objetivos

En la era digital actual, la innovación tecnológica avanza a pasos agigantados, transformando la forma en que interactuamos con el mundo y con la información. Las nuevas aplicaciones y tendencias tecnológicas no solo redefinen industrias enteras, sino que también afectan profundamente nuestras vidas cotidianas. Entre estas tendencias emergentes, tres destacan por su impacto y potencial disruptivo: el Internet de las Cosas (IoT), el *storytelling* de datos y la tecnología *blockchain*. Estas tecnologías no solo representan avances técnicos, sino que también ofrecen nuevas oportunidades para la creación de valor, la mejora de la eficiencia y la transformación de modelos de negocio tradicionales.

Los objetivos de este tema son los siguientes:

- ▶ Comprender las aplicaciones y el impacto del Internet de las Cosas (IoT): analizar cómo la interconexión de dispositivos está cambiando sectores como la salud, la industria y el hogar, y evaluar los desafíos y oportunidades que presenta.
- ▶ Explorar la importancia del *storytelling* de datos: examinar cómo la capacidad de contar historias con datos puede transformar la toma de decisiones, la comunicación y la percepción de la información en diferentes contextos empresariales y sociales.
- ▶ Investigar la tecnología *blockchain* y sus aplicaciones: profundizar en el funcionamiento de la *blockchain*, sus aplicaciones más allá de las criptomonedas, y los posibles futuros desarrollos y retos regulatorios y de seguridad.

## 12.2. Internet de las cosas

El Internet de las cosas (IoT, por sus siglas en inglés) se refiere a la interconexión de dispositivos y objetos a través de Internet, permitiendo que se comuniquen entre sí y con los usuarios. Este fenómeno está impulsado por la miniaturización de los sensores, la mejora de las tecnologías de comunicación y la creciente capacidad de procesamiento y almacenamiento de datos.

La IoT tiene aplicaciones en una amplia variedad de sectores. En la salud, por ejemplo, dispositivos como monitores de actividad física y sensores de salud permiten un seguimiento constante y en tiempo real del estado de los pacientes, lo que puede mejorar significativamente la gestión de enfermedades crónicas y la atención médica preventiva.



Figura 1. Conceptos para entender el IoT (Fuente red.es)

<https://www.acelerapyme.es/recursos/infografia/5-conceptos-para-entender-el-iot>

En el ámbito industrial, la IoT facilita la implementación de la Industria 4.0, donde las fábricas inteligentes utilizan sensores y sistemas de control automatizados para optimizar los procesos de producción, reducir el tiempo de inactividad y mejorar la eficiencia energética. Un ejemplo real es el uso de sensores en la maquinaria de manufactura para predecir fallos antes de que ocurran, lo que reduce costos y aumenta la productividad.

En el hogar, la IoT se manifiesta a través de dispositivos inteligentes como termostatos, cámaras de seguridad y asistentes virtuales, que no solo mejoran la comodidad y seguridad, sino que también contribuyen a la eficiencia energética. A pesar de sus beneficios, la IoT también presenta desafíos significativos, como la seguridad y privacidad de los datos, la interoperabilidad entre dispositivos de diferentes fabricantes y la gestión de grandes volúmenes de datos.

## Ejemplos que ilustran el Internet de las Cosas (IoT)

### **Salud y bienestar: monitores de salud y *wearables***

Los dispositivos portátiles como Fitbit y Apple Watch han revolucionado la forma en que las personas monitorean su salud. Estos dispositivos pueden rastrear una variedad de métricas de salud, como la frecuencia cardíaca, los niveles de actividad física, la calidad del sueño y los pasos diarios. La recopilación y el análisis de estos datos permiten a los usuarios mantener un control constante sobre su salud y hacer ajustes en su estilo de vida para mejorar su bienestar general. Además, estos dispositivos pueden enviar alertas y recordatorios personalizados, así como proporcionar datos detallados a los médicos para una mejor atención y diagnóstico.

### **Industria 4.0: fábricas inteligentes**

Siemens y General Electric (GE) son pioneros en la implementación de fábricas inteligentes utilizando la tecnología IoT. En estas fábricas, los equipos y las máquinas están equipados con sensores que recopilan datos en tiempo real sobre su

rendimiento y estado. Estos datos se analizan para prever el mantenimiento necesario y evitar fallos antes de que ocurran. Por ejemplo, Siemens ha implementado soluciones IoT en sus plantas de producción que han permitido reducir los tiempos de inactividad y optimizar la eficiencia operativa mediante el análisis predictivo y el mantenimiento preventivo.

## **Hogar inteligente: automatización y seguridad**

Google Nest y Amazon Echo son ejemplos populares de dispositivos IoT que están transformando los hogares en espacios inteligentes y más eficientes. Google Nest ofrece termostatos inteligentes que aprenden las preferencias del usuario y ajustan automáticamente la temperatura para optimizar el confort y el ahorro de energía. Además, los dispositivos de seguridad, como cámaras y timbres inteligentes, proporcionan vigilancia constante y notificaciones en tiempo real sobre cualquier actividad inusual.

Amazon Echo, con su asistente virtual Alexa, permite a los usuarios controlar una amplia gama de dispositivos domésticos mediante comandos de voz. Esto incluye luces, electrodomésticos, sistemas de entretenimiento y más, lo que mejora la conveniencia y la interactividad en el hogar.

## **Agricultura inteligente: optimización del cultivo**

John Deere, un líder en maquinaria agrícola, ha integrado la tecnología IoT en sus equipos para mejorar la eficiencia y la productividad agrícola. Los tractores y otros equipos agrícolas de John Deere están equipados con sensores que recopilan datos sobre las condiciones del suelo, la humedad, el clima y el crecimiento de los cultivos. Estos datos se analizan para proporcionar a los agricultores información precisa y en tiempo real sobre cuándo sembrar, regar y cosechar, optimizando así el rendimiento de los cultivos y reduciendo el uso de recursos.

## Transporte y logística: seguimiento de flotas y gestión de inventarios

UPS y Maersk han adoptado la tecnología IoT para mejorar la eficiencia y la transparencia en sus operaciones de transporte y logística. UPS utiliza sensores y dispositivos IoT para rastrear sus vehículos y paquetes en tiempo real, optimizando las rutas de entrega y reduciendo los tiempos de tránsito. Además, los datos recopilados permiten a UPS prever problemas potenciales y tomar medidas preventivas para garantizar entregas puntuales.

Maersk, una de las principales empresas de transporte marítimo, ha implementado IoT en sus contenedores para monitorear la ubicación y el estado de las mercancías durante el tránsito. Los sensores dentro de los contenedores registran datos sobre temperatura, humedad y movimiento, lo que permite a Maersk asegurar la integridad de los productos y responder rápidamente a cualquier incidente durante el transporte.



Figura 2. Esquema general de una Plataforma IoT (Fuente: <https://innovacion-tecnologia.com/iot/plataformas-iot/>)

Estos ejemplos ilustran cómo el Internet de las Cosas está transformando diversos sectores al mejorar la eficiencia, la seguridad y la calidad de vida (ver Figura 2). Desde la salud y el bienestar hasta la industria, el hogar, la agricultura y la logística, la IoT ofrece soluciones innovadoras que están redefiniendo las expectativas y las capacidades en cada área. A medida que la tecnología continúa avanzando, es probable que veamos aún más aplicaciones revolucionarias de IoT en el futuro.

## 12.3. La importancia de contar historias con tus datos o data storytelling

El *storytelling* de datos es la práctica de comunicar información compleja y estadística mediante narrativas que faciliten su comprensión y relevancia. Esta técnica se basa en la idea de que los seres humanos son naturalmente atraídos por las historias, lo que hace que los datos presentados en forma narrativa sean más fáciles de entender y recordar. El *storytelling* de datos es especialmente valioso en el contexto empresarial, donde los líderes deben tomar decisiones basadas en datos y comunicar estos *insights* de manera efectiva a diferentes audiencias.



Figura 3. Elementos del Data Storytelling (Fuente: <https://datos.gob.es/>)

Por ejemplo, en una empresa de *marketing* digital, presentar el rendimiento de una campaña publicitaria mediante gráficos y estadísticas puede ser informativo, pero puede no ser suficiente para transmitir el impacto total de la campaña. Al integrar estos datos en una narrativa que explique cómo las diferentes métricas se relacionan entre sí y con los objetivos comerciales, se puede proporcionar un contexto más rico y una visión más clara para la toma de decisiones estratégicas.

Además, el *storytelling* de datos es crucial en el ámbito científico y de investigación, donde los resultados complejos pueden ser difíciles de comunicar a un público no especializado. Utilizar historias y visualizaciones efectivas permite a los investigadores destacar la importancia de sus hallazgos y su relevancia para la sociedad. Empresas como Tableau y Power BI han desarrollado herramientas que facilitan la creación de visualizaciones interactivas, ayudando a los usuarios a contar historias con sus datos de manera más efectiva.

## 12.4. Tecnología blockchain

La tecnología *blockchain*, o cadena de bloques, es un sistema de registro descentralizado que permite la transferencia segura y transparente de activos digitales. Aunque la *blockchain* es conocida principalmente por su uso en criptomonedas como Bitcoin, sus aplicaciones se extienden mucho más allá. Una de las características clave de la *blockchain* es su inmutabilidad, lo que significa que una vez que se registra una transacción, no puede ser alterada ni eliminada. Esto proporciona una capa adicional de seguridad y confianza en la gestión de registros.

En el sector financiero, la *blockchain* tiene el potencial de transformar las transacciones bancarias y de valores, haciendo que sean más rápidas, seguras y menos costosas. Por ejemplo, Ripple, una empresa de tecnología de pagos, utiliza la *blockchain* para facilitar transferencias internacionales de dinero en cuestión de segundos, comparado con los días que pueden tomar las transferencias tradicionales.

Otra área de aplicación significativa es la gestión de la cadena de suministro. Las empresas pueden utilizar la *blockchain* para rastrear el origen y el estado de los productos a lo largo de toda la cadena de suministro, desde la producción hasta la entrega final. Esto no solo mejora la transparencia y la eficiencia, sino que también ayuda a combatir el fraude y la falsificación. Un ejemplo de esto es el uso de la *blockchain* por parte de Walmart para rastrear la procedencia de los alimentos frescos, mejorando la seguridad alimentaria y reduciendo el tiempo de respuesta en caso de retiradas de productos.

Además, la *blockchain* está emergiendo como una herramienta poderosa en la gestión de identidades digitales, proporcionando a las personas un control más seguro y eficiente sobre su información personal. A pesar de sus numerosas ventajas, la adopción de la *blockchain* enfrenta desafíos, como la escalabilidad, el consumo de energía y la necesidad de desarrollar marcos regulatorios adecuados.

El Internet de las Cosas, el *storytelling* de datos y la tecnología *blockchain* son tres tendencias tecnológicas que están redefiniendo el panorama actual y futuro en múltiples sectores. Su comprensión y aplicación efectiva pueden proporcionar ventajas competitivas significativas y abrir nuevas oportunidades para la innovación y el crecimiento.

## Ejemplos que ilustran la tecnología

### *blockchain*

#### **Sector financiero: transferencias internacionales**

Es una plataforma de pagos que utiliza tecnología *blockchain* para facilitar transferencias internacionales de dinero de manera rápida, segura y económica. A diferencia de los sistemas bancarios tradicionales, que pueden tardar varios días en completar una transacción y cobrar tarifas altas, Ripple permite la transferencia de fondos en cuestión de segundos con tarifas mínimas. Esto es posible gracias a su protocolo de consenso, que valida las transacciones casi instantáneamente. Ripple ha sido adoptado por diversas instituciones financieras, como Santander y American Express, para mejorar la eficiencia de sus operaciones transfronterizas.

#### **Gestión de la cadena de suministro: trazabilidad de productos**

Es una plataforma de identidad digital que utiliza *blockchain* para proporcionar a los usuarios control sobre su información personal. A diferencia de los sistemas tradicionales, donde las identidades digitales están gestionadas por entidades centralizadas, uPort permite a los individuos crear y gestionar sus identidades de

manera descentralizada. Esto significa que los usuarios pueden compartir solo la información necesaria con terceros, manteniendo la privacidad y la seguridad de sus datos personales. Esta tecnología es particularmente útil en regiones con sistemas de identificación inefficientes o inexistentes, proporcionando una solución segura y accesible.

## **Votación electrónica: transparencia y seguridad**

Es una plataforma de votación electrónica que utiliza *blockchain* para asegurar la integridad y la transparencia del proceso electoral. La tecnología *blockchain* garantiza que cada voto sea registrado de manera segura y que no pueda ser alterado una vez emitido. Voatz ha sido utilizado en varias elecciones, incluidas las elecciones municipales de West Virginia en Estados Unidos, donde permitió a los votantes emitir sus votos de manera remota a través de una aplicación móvil segura. Esta solución es especialmente relevante en contextos donde la manipulación de votos y el fraude electoral son preocupaciones importantes.

## **Propiedad intelectual: protección y gestión de derechos**

Ascribe es una plataforma que utiliza *blockchain* para ayudar a los creadores a proteger y gestionar sus derechos de propiedad intelectual. Los artistas, escritores y músicos pueden registrar sus obras en la *blockchain*, creando un registro inmutable de propiedad. Esto no solo facilita la prueba de autoría y la protección contra el plagio, sino que también permite la gestión de licencias y la monetización de las obras. Los creadores pueden vender o licenciar sus trabajos directamente a los consumidores, eliminando intermediarios y garantizando un mayor control sobre sus creaciones.

## **Contratos inteligentes: automatización y eficiencia**

Es una plataforma de *blockchain* que permite la creación y ejecución de contratos inteligentes, que son acuerdos autoejecutables con los términos del contrato

directamente escritos en código. Estos contratos se ejecutan automáticamente cuando se cumplen las condiciones predefinidas, eliminando la necesidad de intermediarios y reduciendo el riesgo de errores o fraudes. Los contratos inteligentes de Ethereum tienen aplicaciones en una variedad de sectores, incluyendo el sector inmobiliario, donde pueden facilitar la compraventa de propiedades, y en la industria del seguro, donde pueden automatizar el procesamiento de reclamaciones.

Estos ejemplos destacan cómo la tecnología *blockchain* está siendo utilizada en diversas industrias para mejorar la seguridad, la transparencia y la eficiencia. Desde la simplificación de las transferencias internacionales y la mejora de la trazabilidad en la cadena de suministro, hasta la protección de identidades digitales y la automatización de contratos, la *blockchain* ofrece soluciones innovadoras que están transformando la manera en que operan las empresas y los gobiernos.

## 12.5. Referencias bibliográficas

Atzori, L., Iera, A., & Morabito, G. (2010). The Internet of Things: A survey. *Computer Networks*, 54(15), 2787-2805. <https://doi.org/10.1016/j.comnet.2010.05.010>

Almajali, S., & Danoy, G. (2020). Edge-Computing Architectures for Internet of Things Applications: A Survey. *Sensors*, 20(22), 6441. <https://doi.org/10.3390/s20226441>

Knaflic, C. N. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Wiley. ISBN: 978-1119002253

Duarte, N. (2019). *DataStory: Explain Data and Inspire Action Through Story*. Ideapress Publishing. ISBN: 978-1940858982

Nakamoto, S. (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System*. <https://bitcoin.org/bitcoin.pdf>

Mougayar, W. (2016). *The Business Blockchain: Promise, Practice, and the Application of the Next Internet Technology*. Wiley. ISBN: 978-1119300311

## Internet de las Cosas (IoT)

Greengard, S. (2015). *The Internet of Things*. MIT Press. ISBN: 978-0262527736

El libro *The Internet of Things* de Samuel Greengard ofrece una exploración profunda de cómo la IoT está transformando diversos aspectos de la vida cotidiana y la industria. Greengard analiza cómo los dispositivos conectados están cambiando la forma en que trabajamos, vivimos y nos comunicamos. El autor aborda las oportunidades y los desafíos que presenta la IoT, incluyendo temas de privacidad, seguridad y regulación. Además, el libro destaca ejemplos prácticos y casos de uso en diferentes sectores como la salud, la industria y el hogar.

## A Survey on Enabling Technologies, Protocols, and Applications

Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications. *IEEE Communications Surveys & Tutorials*, 17(4), 2347-2376.  
<https://doi.org/10.1109/COMST.2015.2444095>

Este estudio académico proporciona una visión integral de las tecnologías habilitadoras, los protocolos y las aplicaciones del Internet de las Cosas. Los autores presentan una revisión exhaustiva de los avances tecnológicos que permiten la IoT, incluyendo sensores, redes y plataformas de procesamiento de datos. También se analizan los desafíos técnicos y las soluciones propuestas en áreas como la interoperabilidad, la escalabilidad y la seguridad. El artículo destaca aplicaciones prácticas de la IoT en sectores como la salud, el transporte y las ciudades inteligentes.

## The Business Blockchain: Promise, Practice, and the Application of the Next Internet Technology

Mougayar, W. (2016). *The Business Blockchain: Promise, Practice, and the Application of the Next Internet Technology*. Wiley. ISBN: 978-1119300311

*The Business Blockchain* de William Mougayar es una guía exhaustiva sobre el impacto y las aplicaciones de la tecnología blockchain en los negocios. El autor explica los principios fundamentales de la blockchain y cómo esta tecnología está transformando sectores como las finanzas, la cadena de suministro y la gestión de identidades. Mougayar discute casos de uso reales y proporciona una visión clara de cómo las empresas pueden aprovechar la blockchain para mejorar la transparencia, la seguridad y la eficiencia operativa. El libro también aborda los desafíos regulatorios y de adopción que enfrenta la tecnología.

1. ¿Qué es el Internet de las Cosas (IoT)?

  - A. Una red de dispositivos interconectados que pueden comunicarse entre sí y con los usuarios.
  - B. Una nueva red social.
  - C. Un tipo de software de inteligencia artificial.
  - D. Una plataforma de comercio electrónico.
  
2. ¿Cuál es un ejemplo de aplicación de IoT en la salud?

  - A. Videojuegos.
  - B. Monitores de actividad física como Fitbit.
  - C. Plataformas de redes sociales.
  - D. Aplicaciones de *streaming* de música.
  
3. ¿Qué tecnología se combina frecuentemente con IoT para mejorar la eficiencia operativa en la industria?

  - A. Realidad virtual.
  - B. *Edge computing*.
  - C. Redes sociales.
  - D. Publicidad en línea.
  
4. ¿Cuál de las siguientes es una preocupación principal del IoT?

  - A. Falta de contenido de entretenimiento.
  - B. Seguridad y privacidad de los datos.
  - C. Costos de desarrollo de *software*.
  - D. Baja velocidad de Internet.

5. ¿Qué sector ha implementado IoT para la trazabilidad de productos?
  - A. Industria de los videojuegos.
  - B. Sector agrícola.
  - C. Servicios de *streaming*.
  - D. Fabricación de automóviles.
6. ¿Qué es la tecnología *blockchain*?
  - A. Una base de datos centralizada.
  - B. Un sistema de registro descentralizado.
  - C. Un tipo de red social.
  - D. Un motor de búsqueda.
7. ¿Cuál es un uso común de la tecnología *blockchain* fuera de las criptomonedas?
  - A. Juegos en línea.
  - B. Gestión de la cadena de suministro.
  - C. Mensajería instantánea.
  - D. Servicios de entrega de alimentos.
8. ¿Qué característica clave de la *blockchain* asegura que una vez que se registra una transacción, no puede ser alterada?
  - A. Escalabilidad.
  - B. Inmutabilidad.
  - C. Facilidad de uso.
  - D. Compatibilidad con todos los dispositivos.

- 9.** ¿Qué plataforma utiliza contratos inteligentes para automatizar acuerdos?
- A. Facebook.
  - B. Ethereum.
  - C. Netflix.
  - D. Google Docs.
- 10.** ¿Qué beneficio proporciona la *blockchain* en el sector financiero?
- A. Mayor costo de transacciones.
  - B. Procesamiento más lento.
  - C. Transacciones más rápidas y seguras.
  - D. Mayor consumo de energía.