1era Ed.
2025

PUERTO MADERO
EDITORIAL

BIG DATA EN LA 4TA REVOLUCIÓN INDUSTRIAL

PROCESAMIENTO CON INTELIGENCIA ARTIFICIAL, ARQUITECTURAS ESCALABLES Y PRIVACIDAD EN ANÁLISIS AVANZADOS

DIEGO ALEJANDRO GARCÍA SARAGURO
MARÍA GABRIELA ARIAS GARNICA
HERNÁN PATRICIO MOYANO AYALA

puertomaderoeditorial.com.ar
La Plata - Argentina

BIG DATA EN LA 4ta REVOLUCIÓN INDUSTRIAL

Procesamiento con Inteligencia Artificial, Arquitecturas Escalables y Privacidad en Análisis Avanzados

AUTORES: Diego Alejandro García Saraguro, María Gabriela Arias Garnica
Hernán Patricio Moyano Ayala

ISBN: 978-631-6557-56-8

ISBN: 978-631-6557-56-8

ISBN: 978-631-6557-56-8

BIG DATA EN LA 4ta REVOLUCIÓN INDUSTRIAL

Procesamiento con Inteligencia Artificial, Arquitecturas Escalables y Privacidad en Análisis Avanzados

Autores
Diego Alejandro García Saraguro
María Gabriela Arias Garnica
Hernán Patricio Moyano Ayala

ISBN: 978-631-6557-56-8

García Saraguro, Diego Alejandro
Big data en la 4ta revolución industrial : procesamiento con inteligencia artificial, arquitecturas escalables y privacidad en análisis avanzados / Diego Alejandro García Saraguro ; María Gabriela Arias Garnica ; Hernán Patricio Moyano Ayala. - 1a ed. - La Plata : Puerto Madero Editorial Académica, 2025.
Libro digital, PDF/A

Archivo Digital: descarga y online
ISBN 978-631-6557-56-8

1. Revolución Industrial. 2. Inteligencia Artificial. 3. Análisis de Datos. I. Arias Garnica, María Gabriela II. Moyano Ayala, Hernán Patricio III. Título
CDD 006.3

ISBN: 978-631-6557-56-8

Primera Edición, Junio 2025

BIG DATA EN LA 4ta REVOLUCIÓN INDUSTRIAL Procesamiento con Inteligencia Artificial, Arquitecturas Escalables y Privacidad en Análisis Avanzados
ISBN: 978-631-6557-56-8

Editado por:
Sello editorial: ©Puerto Madero Editorial Académica
Nº de Alta: 933832

Editorial: © Puerto Madero Editorial Académica
CUIL: 20630333971
Calle 45 N491 entre 4 y 5
Dirección de Publicaciones Científicas Puerto Madero Editorial Académica
La Plata, Buenos Aires, Argentina
Teléfono: +54 9 221 314 5902
+54 9 221 531 5142
Código Postal: AR1900

Este libro se sometió a arbitraje bajo el sistema de doble ciego (peer review)

Corrección y diseño:
Puerto Madero Editorial Académica
Diseñador Gráfico: José Luis Santillán Lima

Diseño, Montaje y Producción Editorial:
Puerto Madero Editorial Académica
Diseñador Gráfico: Santillán Lima, José Luis

Director del equipo editorial: Santillán Lima, Juan Carlos

Editor: Santillán Lima, Juan Carlos

Hecho en Argentina
Made in Argentina

ISBN: 978-631-6557-56-8

ISBN: 978-631-6557-56-8

AUTORES:

Diego Alejandro García Saraguro
Escuela Superior Politécnica del Chimborazo. Ecuador
diego.garcia@espoch.edu.ec
https://orcid.org/0009-0008-7792-3969

María Gabriela Arias Garnica
Escuela Superior Politécnica del Chimborazo. Ecuador
mariag.arias@espoch.edu.ec
https://orcid.org/0009-0002-2535-9776

Hernán Patricio Moyano Ayala
Escuela Superior Politécnica del Chimborazo. Ecuador
patricio.moyano@espoch.edu.ec
https://orcid.org/0009-0008-4856-3925

ISBN: 978-631-6557-56-8

ISBN: 978-631-6557-56-8

ÍNDICE

ISBN: 978-631-6557-56-8

ISBN: 978-631-6557-56-8

ISBN: 978-631-6557-56-8

ISBN: 978-631-6557-56-8

RESUMEN

El libro *Big Data en la 4ta Revolución Industrial* ofrece una visión integral del papel que juegan los datos masivos en el contexto de la transformación digital de la industria. A través de cinco capítulos, se presentan los conceptos, tecnologías, aplicaciones y desafíos éticos relacionados con el Big Data, proporcionando al lector una base sólida para comprender su impacto en la era de la Industria 4.0.

En el **Capítulo 1**, se introducen los fundamentos del Big Data y su relación directa con la Cuarta Revolución Industrial. Se analiza cómo el crecimiento exponencial de los datos, impulsado por tecnologías emergentes, exige nuevas formas de organización, análisis y visualización, utilizando herramientas estadísticas y gráficas. El **Capítulo 2** se enfoca en los procesos de captura de datos y en las arquitecturas tecnológicas necesarias para los entornos industriales actuales. Se abordan sistemas de adquisición, sensores, bases de datos modernas como NoSQL, y tecnologías como *Hadoop* y *Spark*, subrayando la importancia de contar con datos de calidad y en tiempo real. En el **Capítulo 3**, se exploran los sistemas de procesamiento de datos masivos mediante inteligencia artificial. Se describen algoritmos y herramientas que permiten transformar grandes volúmenes de datos en conocimiento útil, destacando técnicas como árboles de decisión, aprendizaje no supervisado y redes neuronales. El **Capítulo 4** trata sobre la visualización de datos y su uso estratégico en la inteligencia empresarial. Se abordan métodos para representar datos de forma clara y útil para la toma de decisiones, incluyendo el uso de *dashboards*, KPIs y herramientas como el Cuadro de Mando Integral. Finalmente, el **Capítulo 5** se centra en los desafíos éticos y legales asociados al Big Data. Se discute la legislación sobre protección de datos personales, la necesidad de anonimización y los principios de transparencia, equidad y responsabilidad en el tratamiento de la información. Este libro está diseñado para estudiantes, profesionales y tomadores de decisiones interesados en comprender cómo el Big Data se convierte en un pilar fundamental en la evolución de la industria hacia modelos más inteligentes, eficientes y sostenibles.

Palabras Claves: Big Data, Industry 4.0, Digital Transformation, Data Analysis, Artificial Intelligence, Data Ethics.

ISBN: 978-631-6557-56-8

ABSTRACT

The book *Big Data in the 4th Industrial Revolution* provides a comprehensive overview of the role that massive data plays in the context of digital transformation in industry. Through five chapters, it presents the concepts, technologies, applications, and ethical challenges related to Big Data, offering readers a solid foundation to understand its impact in the Industry 4.0 era.

Chapter 1 introduces the fundamentals of Big Data and its direct connection to the Fourth Industrial Revolution. It examines how the exponential growth of data, driven by emerging technologies, demands new ways of organizing, analyzing, and visualizing information using statistical and graphical tools. **Chapter 2** focuses on data acquisition processes and the technological architectures required in today's industrial environments. It explores acquisition systems, sensors, modern databases such as NoSQL, and technologies like Hadoop and Spark, emphasizing the importance of real-time, high-quality data. **Chapter 3** delves into massive data processing systems powered by artificial intelligence. It describes algorithms and tools that transform large volumes of data into useful knowledge, highlighting techniques such as decision trees, unsupervised learning, and neural networks. **Chapter 4** covers data visualization and its strategic use in business intelligence. It presents methods for representing data in a clear and actionable way to support decision-making, including dashboards, KPIs, and tools like the Balanced Scorecard. Finally, **Chapter 5** addresses the ethical and legal challenges associated with Big Data. It discusses legislation on personal data protection, the need for anonymization, and principles such as transparency, fairness, and accountability in data management. This book is intended for students, professionals, and decision-makers interested in understanding how Big Data is becoming a fundamental pillar in the industry's evolution toward smarter, more efficient, and sustainable models.

Keywords: Big Data, Industry 4.0, Digital Transformation, Data Analysis, Artificial Intelligence, Data Ethics.

ISBN: 978-631-6557-56-8

INTRODUCCIÓN

La rápida evolución tecnológica, el desarrollo de las comunicaciones, la incorporación habitual de teléfonos inteligentes en nuestras rutinas, la creciente popularidad de los aparatos conectados o el auge de los sistemas IoT (Internet de las Cosas, por sus siglas en inglés *Internet of Things*) ya no representan novedades, sino que se han integrado completamente en nuestro quehacer diario. Esta realidad demuestra cómo nos encontramos sumergidos en un proceso de transformación digital y avance tecnológico cuya velocidad no tiene comparación en la historia. Esta situación ha provocado la creación de cantidades masivas de datos que superan la capacidad de gestión de los métodos convencionales, lo que ha impulsado el surgimiento del *big data* como una solución imprescindible más que como una simple opción. Naturalmente, las organizaciones, y especialmente aquellas vinculadas a la Industria 4.0, se encuentran completamente inmersas en este torbellino de innovación digital, constantes actualizaciones, desarrollo continuo y, en esencia, una completa reinvención de sus procesos.

El procesamiento de información representa un componente fundamental dentro de los entornos de big data, particularmente en el ámbito de la Industria 4.0. La evaluación de procesos productivos, los modelos predictivos y la mejora en la eficiencia energética, junto con numerosas aplicaciones adicionales en el sector industrial, encuentran sus fundamentos en métodos estadísticos y sus diversas herramientas analíticas derivadas. La estadística constituye uno de los pilares básicos para el tratamiento de información, por lo que este contenido ofrece una aproximación inicial a sus principios elementales y su conexión con el big data. Se revisan los fundamentos teóricos y metodológicos estadísticos más relevantes, abordando la estructuración y representación de conjuntos de datos para extraer significado de los mismos. Además, se detallan los modelos de representación visual más frecuentes, junto con recomendaciones para seleccionar el formato adecuado según la naturaleza de los datos y los propósitos específicos del estudio.

Antes de que la informática y los métodos modernos de adquisición de datos se integraran en la vida cotidiana, la recopilación de información solía ser un proceso complejo y costoso, que dependía en gran medida de la intervención humana para convertir documentos físicos en formato digital. Sin embargo, el constante desarrollo tecnológico ha transformado este panorama, facilitando la aparición de métodos de captura de datos cada vez más variados, automatizados y precisos. A medida que el costo de almacenamiento por unidad ha disminuido, los sistemas actuales son capaces de almacenar volúmenes masivos de datos provenientes de múltiples y diversas fuentes. Esta diversidad exige establecer una base robusta que garantice una captura eficiente y flexible, adecuada para los distintos usos que se pretende dar a la información. Esto implica comprender la diferencia entre dato, información y conocimiento; evaluar la calidad de los datos; y analizar casos reales de captura y almacenamiento.

ISBN: 978-631-6557-56-8

En la actualidad, los datos generados a nivel mundial provienen de redes sociales, mercados bursátiles, sensores IoT industriales, entre otras fuentes. Estos deben almacenarse en repositorios que permitan su análisis mediante algoritmos de inteligencia artificial como *machine learning* o *deep learning*, aplicados en escenarios como el mantenimiento predictivo. Para abordar estas necesidades, se han desarrollado tecnologías como las bases de datos NoSQL (por ejemplo, MongoDB y Cassandra) y sistemas de procesamiento masivo de datos como Hadoop y Apache Spark. Estas tecnologías permiten soluciones escalables en la nube, esenciales para manejar grandes volúmenes de información. En contextos críticos donde es necesario procesar datos en tiempo real, surge el enfoque de *smart data*, donde la información se analiza y procesa en el instante mismo de su captura.

Ante estas crecientes necesidades de procesamiento y análisis de datos masivos, especialmente en el contexto de la Industria 4.0, han surgido sistemas y herramientas fundamentales. Se destacan tecnologías clave como *Hadoop*, HDFS, *MapReduce y Apache Spark*, que han revolucionado el almacenamiento, procesamiento y análisis de información en entornos distribuidos. Estas tecnologías permiten superar los retos del crecimiento exponencial de datos provenientes de múltiples fuentes digitales, mediante enfoques paralelos y distribuidos. Su aplicación en la Industria 4.0 ha sido crucial para optimizar procesos industriales, tomar mejores decisiones y aplicar métodos de inteligencia artificial como árboles de decisión, *clustering*, redes neuronales y sistemas de recomendación, demostrando la sinergia entre big data e IA.

En este entorno, el análisis y la visualización de datos se han convertido en herramientas fundamentales para la toma de decisiones estratégicas. La representación gráfica de grandes volúmenes de información permite una comprensión más rápida y efectiva, facilitando la asimilación de conceptos complejos y el descubrimiento de patrones relevantes. La creación de infografías, gráficas y otras visualizaciones requiere conocimientos sólidos sobre los datos y las herramientas disponibles. Se abordan los conceptos de visualización de datos e inteligencia empresarial (BI), desde la idea inicial hasta su presentación final, con una clasificación de los distintos tipos de gráficos y visualizaciones, así como las herramientas (código abierto y comerciales) utilizadas para su desarrollo. Asimismo, se presenta una visión más amplia del concepto de BI, incorporando modelos como el Cuadro de Mando Integral (CMI) y su papel en la toma de decisiones organizacionales.

Por último, el notable aumento de dispositivos, plataformas y herramientas que recopilan información personal ha impulsado la implementación de regulaciones y normativas para proteger la privacidad y el uso adecuado de los datos. Se analizan los efectos de leyes como la Ley de Protección de Datos, la *California Consumer Privacy Act* y las normativas en países

ISBN: 978-631-6557-56-8

latinoamericanos. Se destaca la importancia de la anonimización como mecanismo para garantizar el tratamiento legal y seguro de la información personal. A su vez, se discuten los retos que enfrentan el big data y la Industria 4.0 en materia de privacidad, así como las tendencias y soluciones emergentes que buscan armonizar la innovación tecnológica con los derechos de los individuos.

ISBN: 978-631-6557-56-8

CAPÍTULO I

FUNDAMENTOS DEL BIG DATA Y LA INDUSTRIA 4.0

1.1. Introducción y Objetivo del Capítulo

La rápida evolución tecnológica, el desarrollo de las comunicaciones, la incorporación habitual de teléfonos inteligentes en nuestras rutinas, la creciente popularidad de los aparatos conectados o el auge de los sistemas IoT (Internet de las Cosas, por sus siglas en inglés Internet of Things) ya no representan novedades, sino que se han integrado completamente en nuestro quehacer diario. Esta realidad demuestra cómo nos encontramos sumergidos en un proceso de transformación digital y avance tecnológico cuya velocidad no tiene comparación en la historia.

Esta situación ha provocado la creación de cantidades masivas de datos que superan la capacidad de gestión de los métodos convencionales, lo que ha impulsado el surgimiento del big data como una solución imprescindible más que como una simple opción. Naturalmente, las organizaciones, y especialmente aquellas vinculadas a la industria 4.0, se encuentran completamente inmersas en este torbellino de innovación digital, constantes actualizaciones, desarrollo continuo y, en esencia, una completa reinvención de sus procesos.

El procesamiento de información representa un componente fundamental dentro de los entornos de big data, particularmente en el ámbito de la industria 4.0. La evaluación de procesos productivos, los modelos predictivos y la mejora en la eficiencia energética, junto con numerosas aplicaciones adicionales en el sector industrial, encuentran sus fundamentos en métodos estadísticos y sus diversas herramientas analíticas derivadas.

La estadística constituye uno de los pilares básicos para el tratamiento de información, por lo que este capítulo ofrece una aproximación inicial a sus principios elementales y su conexión con el big data. La exposición comienza revisando los fundamentos teóricos y metodológicos estadísticos más relevantes, abordando paralelamente el primer reto que surge al examinar conjuntos de datos: su adecuada estructuración y representación para extraer significado de los mismos. Por otra parte, se considera que los gráficos representan probablemente el método más claro y accesible para interpretar información. Debido a esto, el presente capítulo detalla los modelos de representación visual más frecuentes, junto con recomendaciones para seleccionar el formato adecuado según la naturaleza de los datos disponibles y los propósitos específicos del estudio.

El contenido presenta diversos fundamentos estadísticos, incorporando metodologías específicas aplicables al estudio de variables ya sea de manera aislada o interrelacionada junto con

ISBN: 978-631-6557-56-8

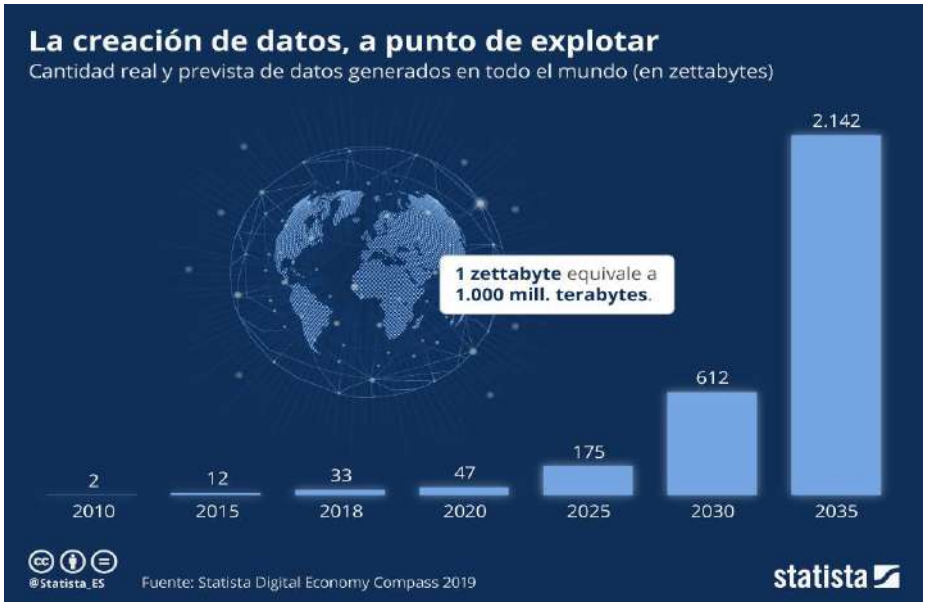
procedimientos para valorar categorizaciones. Su propósito radica en familiarizar al lector con estos elementos, facilitando su posterior ampliación mediante las fuentes referenciadas y los materiales complementarios incluidos en el apartado final. Para ello éste capítulo tiene por objetivos:

- Comprender el concepto de big data, reconociendo sus desafíos, posibilidades y las herramientas tecnológicas asociadas.
- Analizar cómo los avances tecnológicos y el crecimiento exponencial de datos pueden optimizar los procesos empresariales, especialmente en el ámbito de la industria 4.0.
- Reconocer el papel central de los sistemas big data en la transformación digital.
- Examinar y comprender los principales obstáculos que plantea el big data en los contextos de industria 4.0.

1.2. Fundamentos de los sistemas Big Data

La aceleración tecnológica de la última década ha traído consigo un crecimiento descomunal en la cantidad de información que gestionamos diariamente. Según datos de [1] esta expansión se manifiesta claramente en las proyecciones de la Figura 1: mientras que en 2025 se prevé una generación de 175 zettabytes de datos, esta cifra podría multiplicarse hasta alcanzar los 612 zettabytes para 2030 y superar los 2142 zettabytes en 2035, mostrando una curva de crecimiento acelerado.

Figura 1:Cantidad real y prevista de datos generados en todo el mundo (en zettabytes).



Fuente: [2]

Cabe destacar que los datos producidos presentan un carácter acumulativo, ya que la cantidad de información que se elimina resulta insignificante comparada con los nuevos registros generados,

ISBN: 978-631-6557-56-8

situación favorecida por la reducción progresiva en los costes de almacenamiento. Este escenario propicia la consolidación del big data como modelo para abordar los desafíos que plantea la gestión de estas cantidades masivas de información.

La mera existencia de estos volúmenes de datos, junto con las dificultades inherentes a su procesamiento, genera tanto obstáculos como posibilidades que el big data busca solventar y aprovechar. En esencia, la finalidad primordial de cualquier solución basada en big data radica en utilizar la información de manera óptima y productiva, permitiendo así una toma de decisiones más acertada. Esta circunstancia explica su relevancia en los sistemas de apoyo a decisiones basados en datos (*DSS - Data-driven Decision Support Systems*) [3].

Actualmente, la recopilación y procesamiento de información proveniente de diversas fuentes resulta compleja y requiere importantes recursos. Un caso típico ocurre al intentar integrar dispositivos de distintos fabricantes, como un sensor de temperatura de una marca con un sistema de iluminación inteligente de otra, lo que obliga a trabajar con plataformas separadas para obtener y administrar sus datos. Estas fuentes de información para sistemas big data abarcan desde sensores y wearables hasta equipos industriales, cámaras, bases de datos, fuentes públicas, observaciones directas, dispositivos energéticos e incluso seres vivos. La variedad no solo reside en el origen de los datos (personas, dispositivos o software), sino también en su formato de entrega. Según su estructura, según [4], podemos categorizar los datos en tres tipos principales:

1.2.1. Datos Estructurados

Estos datos siguen un esquema estructurado predeterminado, lo que facilita su procesamiento y análisis. Se organizan en formatos tabulares con relaciones claras entre filas y columnas, siendo los archivos de hojas de cálculo (como Excel) y las bases de datos relacionales (SQL) los ejemplos más representativos de este tipo.

1.2.2. Datos no Estructurados

Se refieren a información que carece de una organización predeterminada o de un esquema definido. Este tipo de contenido puede incluir textos, pero también valores numéricos, fechas y otros elementos factuales. Entre los ejemplos más frecuentes se encuentran archivos multimedia (audio y video) así como bases de datos no relacionales.

El manejo de esta clase de información ha experimentado avances significativos recientemente, impulsado por el desarrollo de tecnologías especializadas como MongoDB para la gestión documental. Este progreso adquiere especial importancia en el ámbito del big data, considerando que la mayoría de los datos empresariales pertenecen a esta categoría (documentos, material multimedia, etc.). La posibilidad de obtener información valiosa de estos contenidos

ISBN: 978-631-6557-56-8

desorganizados constituye uno de los factores clave que explican la expansión acelerada de las soluciones big data.

1.2.3. Datos Semiestructurados

Estos datos presentan una organización parcial, donde aunque no siguen el esquema rígido de las bases de datos tradicionales, incorporan marcas identificativas que permiten diferenciar componentes y establecer relaciones jerárquicas entre ellos (poseen cierta autodescripción). Formatos como CSV, JSON y XML son ejemplos característicos de esta categoría, cuyo análisis resulta más accesible que el de los datos no estructurados, gracias a las numerosas herramientas disponibles para su interpretación.

Cabe destacar que, en el caso particular del formato CSV, aunque aparentemente tabular, su diseño básico no permite representar por sí solo relaciones complejas o estructuras jerárquicas. Para lograr esto, se requiere utilizar múltiples archivos CSV interconectados mediante claves de referencia, característica que fundamenta su clasificación como semiestructurado.

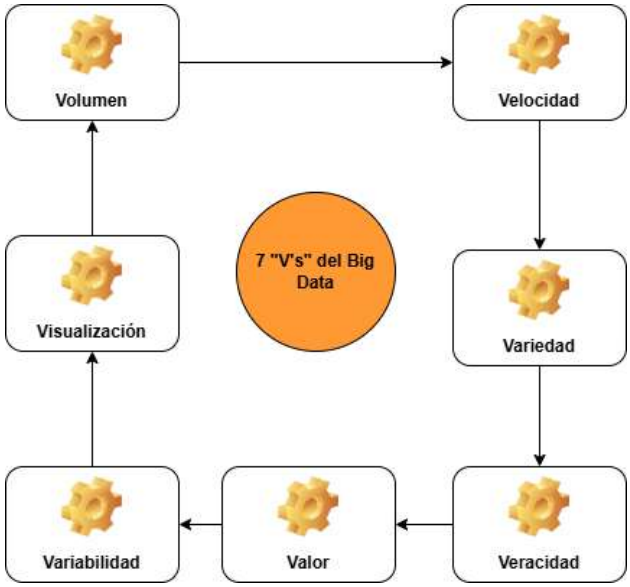
1.2.4. Metainformación

Los metadatos, aunque no constituyen una estructura de datos propiamente dicha, representan un componente fundamental para el análisis de información. Podemos definirlos como información descriptiva sobre otros datos, que aporta contexto y detalles adicionales sobre un conjunto específico. Por ejemplo, en un archivo de imágenes, los metadatos podrían incluir la fecha y ubicación donde fueron capturadas, transformando estos atributos en datos estructurados dentro de un contexto más amplio.

Sin embargo, el big data trasciende aspectos como el volumen, formato o diversidad de datos. Los sistemas de big data se definen fundamentalmente por los desafíos que buscan resolver, comúnmente conocidos como las "V" del big data. Este marco conceptual comenzó con tres dimensiones básicas, se expandió a cinco, y en algunas publicaciones especializadas incluso se extiende hasta abarcar diecisiete características distintas [5]. A continuación, se muestra la versión de las siete «V» del big data, tal y como muestra la Figura 2:

ISBN: 978-631-6557-56-8

Figura 2: Las 7 V's del Big Data



Fuente: Autores

A continuación, se detalla a que hace referencia cada una de las características:

Volumen.- Este aspecto representa el más evidente, aludiendo a los volúmenes masivos de información que deben procesar y almacenar las plataformas de big data.

Velocidad.- Este aspecto alude a la rapidez con que los sistemas deben capturar, analizar y poner a disposición la información. Un caso paradigmático son las plataformas de seguimiento instantáneo, donde la velocidad de procesamiento resulta determinante para su funcionamiento efectivo.

Variedad.- Esta característica aborda la diversidad en los orígenes y estructuras de los datos. Las soluciones de big data deben gestionar esta variedad de manera fluida y efectiva. Un caso ilustrativo sería un sistema agrícola inteligente que combine: registros estructurados de aplicaciones fitosanitarias (desde bases de datos tradicionales), mediciones ambientales en formatos semiestructurados (como JSON o XML) y contenido no estructurado como fotografías del cultivo obtenidas mediante drones.

Veracidad.- Este aspecto hace referencia a la confiabilidad y exactitud de la información recolectada. Contar con orígenes confiables -como fuentes institucionales o datos oficiales- que proporcionen información precisa, ayuda a evitar los errores y distorsiones que generan los conjuntos de datos inconsistentes o poco fiables. La calidad de los datos se convierte así en un factor determinante para garantizar análisis y conclusiones válidas.

ISBN: 978-631-6557-56-8

Variabilidad.- Este atributo (distinto de la variedad) alude a la naturaleza dinámica en la interpretación de los datos, más que a su formato u origen. Ilustrando este concepto, consideremos el caso de una bodega que produce vino: aunque mantenga invariables las proporciones de uva, las condiciones de crianza y los procesos año tras año, cada cosecha desarrolla características organolépticas únicas que difieren entre añadas. Esta fluctuación en los resultados finales, pese a mantener constantes los parámetros de producción, ejemplifica perfectamente el desafío que supone la variabilidad en el análisis de datos.

Visualización.- La representación gráfica se ha transformado en un componente fundamental de las plataformas de big data modernas. Emplear elementos visuales como diagramas, paneles interactivos o representaciones infográficas para presentar información derivada de conjuntos complejos resulta significativamente más eficaz que mostrar los datos sin procesar. Esta aproximación permite una comprensión más intuitiva e inmediata de patrones y relaciones complejas.

Valor.- La finalidad primordial de cualquier solución de big data radica en generar conocimiento valioso a partir del procesamiento de información. No basta con acumular grandes cantidades de datos, sino que es necesario transformarlos mediante procesos estructurados que permitan su interpretación clara y apoyen la toma de decisiones estratégicas.

Para lograr este propósito, se emplean métodos avanzados de inteligencia computacional, particularmente el aprendizaje automatizado, que agilizan el examen de información y facilitan el manejo de conjuntos masivos que exceden la capacidad de los métodos convencionales. Estas técnicas encuentran aplicación práctica en diversos ámbitos: desde redes neuronales para interpretación visual, hasta algoritmos de sugerencia personalizada en comercio electrónico o métodos de agrupamiento para segmentación de mercados, demostrando su versatilidad y eficacia en escenarios reales.

1.2.5. Ámbitos

En vista de lo anterior, se puede afirmar que los campos de aplicación del big data son prácticamente ilimitados, entre los cuales destacan:

Retail.- El crecimiento del comercio electrónico ha convertido al big data en el principal apoyo de las empresas minoristas. Mediante sus herramientas, estas organizaciones pueden implementar enfoques orientados al cliente (*customer centric*), ajustar sus estructuras de precios y mejorar sus procesos productivos.

ISBN: 978-631-6557-56-8

Logística y transporte.- Este sector figura entre los primeros en aprovechar las ventajas de los sistemas big data. La monitorización de flotas, el cálculo de rutas eficientes, la anticipación de demandas para control de inventarios y el seguimiento en tiempo real del estado de los vehículos representan algunas de las áreas más favorecidas por estas soluciones.

Banca.- A pesar de ser uno de los sectores más tradicionales y reticentes a los cambios (debido, como cabe esperar, a las grandes medidas de seguridad), la banca también ha sucumbido a la implantación y utilización de sistemas big data. Por ejemplo, la aparición de las *Fintech* basa gran parte de su valor añadido en la resolución de problemas de nicho mediante la aplicación de big data. Por otra parte, los propios bancos son capaces de mejorar sus previsiones de rentabilidad y riesgos mediante un uso intensivo de datos y algoritmos.

Marketing.- La revolución digital y, sobre todo, la aparición de las redes sociales y los motores de búsqueda en Internet han cambiado el mundo del marketing de tal forma que no puede entenderse sin que sea asociado al big data. La gran cantidad de información que generamos facilita que recibamos un marketing más personalizado que, como cabe esperar, intenta maximizar el número de conversiones.

Sector Salud.- Al igual que en otros ámbitos, los sistemas de salud han experimentado notables avances gracias al big data. Más allá de las mejoras en diagnóstico médico mediante el análisis de grandes volúmenes de información y el desarrollo de algoritmos avanzados, la digitalización del sector sanitario permite ahora acceder a historiales médicos desde cualquier centro de salud o beneficiarse de sistemas como la prescripción electrónica.

Ámbito Educativo.- El sector educativo tampoco ha quedado al margen de esta transformación digital. Algunas instituciones ya ejemplifican cómo la aplicación de principios del big data puede optimizar los servicios estudiantiles, facilitar el aprendizaje a distancia, mejorar los sistemas de clasificación del alumnado y permitir una mayor personalización de los programas académicos.

Gobierno electrónico.- La adopción de plataformas digitales en la administración pública ha necesitado resolver múltiples desafíos asociados al big data. Un ejemplo claro es la creación de registros documentales digitales, que requieren sistemas seguros de almacenamiento y gestión de información donde las tecnologías de big data resultan fundamentales.

Sector asegurador.- Las compañías de seguros han experimentado una notable evolución impulsada por el big data. La capacidad de procesar mayores volúmenes de información permite cálculos de riesgo más precisos. Además, muchas aseguradoras están incorporando datos de

ISBN: 978-631-6557-56-8

dispositivos portátiles como relojes inteligentes para analizar hábitos de los clientes y perfeccionar sus modelos de evaluación.

De la breve descripción de algunos de los ámbitos que se benefician del big data, pueden identificarse, entre muchas otras, las siguientes oportunidades:

- ✓ Mejora en la monitorización de procesos.
- ✓ Optimización de campañas de marketing.
- ✓ Mejora de las ventas.
- ✓ Mejora de la satisfacción y la experiencia del cliente.
- ✓ Mejora de la gestión logística y de almacén.
- ✓ Identificación de productos, tendencias y ventajas competitivas.
- ✓ Ahorro de costes.
- ✓ Mejora del proceso de toma de decisiones (más rápida y con más criterios).
- ✓ Aumento de la capacidad de pronosticar la demanda con mayor precisión.
- ✓ Resolución de problemas de redes de distribución más complejos.
- ✓ Mejora de la eficiencia de la planificación.
- ✓ Mejora de la planificación de los recursos humanos.
- ✓ Colaboración en la cadena de suministro.
- ✓ Monitorización de vehículos y maquinaria.

Es evidente que muchas de las aplicaciones y oportunidades mencionadas anteriormente resultan directamente relevantes para el contexto de la Industria 4.0. La siguiente sección abordará específicamente esta relación, examinando los principales desafíos que plantea la implementación de sistemas big data en entornos de producción industrial inteligente.

1.3. Desafíos y ventajas del Big Data en la Industria 4.0

Como se mencionó previamente, la industria 4.0 ha adoptado activamente las soluciones de big data, aunque cabe señalar que se trata de conceptos distintos pero sinérgicos. En el ámbito industrial, el big data resulta importante para diversas aplicaciones clave, particularmente en la manufactura inteligente, donde el análisis de información proveniente de sensores permite anticipar necesidades de mantenimiento y reparación en equipos productivos. Esta implementación permite a los fabricantes incrementar su eficiencia operativa, monitorear procesos en tiempo real, perfeccionar el mantenimiento preventivo y automatizar la administración de la producción [6].

La producción de datos en entornos industriales 4.0 no constituye un fenómeno reciente. No obstante, antes de la aparición de las soluciones big data, gran parte de esta información

ISBN: 978-631-6557-56-8

permanecía inexplorada en repositorios o silos de datos, al carecer de tecnologías adecuadas para su procesamiento y análisis.

Los sistemas big data vienen a resolver esta limitación práctica: transformar datos existentes en conocimiento útil. Más allá de aplicaciones puntuales, la industria 4.0 persigue implementar el big data como base para una estrategia global de inteligencia organizacional, integrando la captura, el procesamiento y el intercambio de información en todas las áreas de negocio [7].

Para alcanzar estos fines, la industria 4.0 está aprovechando las capacidades que brindan tanto el big data como otras tecnologías emergentes, entre ellas el IoT y la robótica, todas ellas estrechamente interrelacionadas. El propósito fundamental consiste en perfeccionar y automatizar los procesos productivos, abarcando toda la cadena de suministro. En cuanto a la sensorización, la meta última de la industria 4.0 es que los sensores incorporados en maquinaria, infraestructuras, objetos, personal, componentes y productos en fabricación transmitan información en tiempo real a los sistemas de información implementados.

La industria 4.0 también aprovechará los algoritmos de inteligencia artificial, especialmente el aprendizaje automático, para procesar y extraer conocimiento de los datos recolectados, permitiendo así la adaptación automática de los procesos según las necesidades detectadas. A modo ilustrativo, el big data sustenta los entornos de industria 4.0 en aplicaciones como:

Detección de factores ocultos.- Es frecuente que emerjan variables no visibles tanto en los procesos de fabricación como en otros ámbitos. El empleo de técnicas big data posibilita su localización, ayudando a evitar los estrangulamientos que provocan en la producción.

Optimización en tiempo real.- El análisis instantáneo de datos resulta fundamental para la industria 4.0, permitiendo perfeccionar la cadena logística, ajustar estrategias de precios, anticipar averías, innovar en productos y concebir instalaciones industriales avanzadas.

Plataformas de autogestión analítica.- La implementación de soluciones de autoservicio para el procesamiento de datos permite integrar y examinar los volúmenes de información generados en las plantas productivas. Un caso paradigmático es el de Intel, cuyos equipos en instalaciones inteligentes transmiten información a estos sistemas, que interpretan los datos al instante, identifican tendencias, descubren anomalías y generan representaciones gráficas para la dirección operativa.

Mantenimiento preventivo avanzado.- Representa un caso paradigmático del uso estratégico de datos para la gestión operativa. El análisis de información permite anticipar intervenciones y

ISBN: 978-631-6557-56-8

determinar las acciones prioritarias para prevenir paradas no planificadas o fallos en los equipos. En los entornos de Industria 4.0, el procesamiento de datos es equivalente a la implementación de sistemas de mantenimiento predictivo.

Automatización inteligente de procesos productivos.- Consiste en minimizar la intervención humana directa en las operaciones de fabricación. Este enfoque integra el análisis de datos históricos de producción con información en tiempo real del proceso específico, implementando ajustes automáticos mediante sistemas robóticos y actuadores conectados a plataformas de control. El software central interpreta los patrones identificados mediante análisis big data, generando instrucciones precisas que modifican físicamente la configuración de equipos y maquinaria sin requerir acción humana [8].

Optimización de la cadena de suministro.- Los beneficios alcanzan no solo las operaciones internas, sino que se amplifican mediante la incorporación de información de socios externos, incluyendo distribuidores, suministradores y otros colaboradores estratégicos.

En los próximos capítulos se analiza en detalle los principales elementos que conforman los sistemas de big data, incluyendo: los métodos de captura de información, las infraestructuras *cloud* que posibilitan su gestión, las aplicaciones de inteligencia artificial más relevantes, las plataformas de visualización más utilizadas, junto con otros aspectos vinculados a la toma de decisiones empresariales y al marco regulatorio aplicable.

1.4. Fundamentos teóricos

La estadística constituye una disciplina fundamental para el análisis de información. En una primera aproximación simplificada, podría concebirse simplemente como un conjunto de datos diversos. Ejemplos cotidianos incluyen cifras de ventas de automóviles o tasas de desempleo. No obstante, esta visión superficial alude únicamente a estudios específicos, sin capturar la esencia de la estadística como ciencia dedicada al estudio sistemático de los datos.

En términos más rigurosos, la estadística puede definirse como la disciplina científica que gestiona la información mediante un proceso integral que abarca: el diseño de estudios, la recolección sistemática de datos, su procesamiento analítico, y finalmente la organización, síntesis y presentación de los resultados para obtener conocimiento significativo. Como señala: *"La estadística es la ciencia que nos permite extraer conocimiento de los datos"* [9].

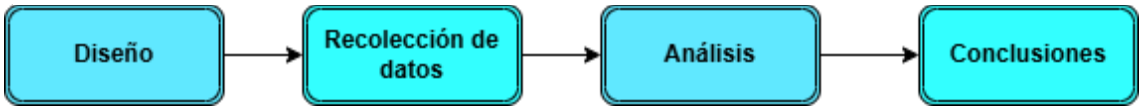
Esta formulación engloba tanto el aspecto metodológico (diseño y recolección) como el analítico (procesamiento e interpretación), destacando su papel transformador: convertir datos brutos en

ISBN: 978-631-6557-56-8

información accionable. La versión abreviada captura la esencia de su utilidad práctica en investigación y toma de decisiones.

La Figura 3 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.

Figura 3: Fases de un estudio estadístico completo.



Fuente: Autores

Cada etapa de un análisis estadístico reviste igual importancia, aunque cabe resaltar que la recopilación de información requiere seguir protocolos estadísticos rigurosos, constituyendo una fase particularmente crítica. Algunos expertos incorporan una fase preliminar adicional: la delimitación del problema de investigación.

¿Cuál es entonces la finalidad de la estadística? Retomando a [9], su propósito radica en *"profundizar en la comprensión de un fenómeno mediante el análisis de los datos disponibles"*. Según su aplicación, la estadística se clasifica en:

- **Estadística descriptiva:** Caracteriza una población mediante el examen de datos muestrales.
- **Estadística inferencial:** Deriva conclusiones generalizables para toda la población.

1.4.1. Dato, Población, muestra y muestreo

Dato: Representa el concepto fundacional en estadística. Desde esta perspectiva, los datos trascienden lo meramente numérico, incorporando siempre un contexto que los transforma en información significativa sobre algún aspecto concreto. Este referente contextual se denomina individuo, y al agruparse según criterios específicos, conforman un población estadística.

Población.- Corresponde al conjunto de individuos que serán objeto de estudio para obtener conclusiones relevantes. La estadística se enfoca específicamente en fenómenos colectivos que presentan componentes de variabilidad, a diferencia de los sistemas deterministas que estudian las ciencias exactas.

Muestreo.- Consiste en la selección sistemática de individuos para conformar una muestra. Este procedimiento asegura la calidad básica de los datos, fundamentales para análisis posteriores. La

ISBN: 978-631-6557-56-8

muestra debe reflejar la diversidad de la población original, garantizando así su representatividad estadística.

Error de muestreo.- Al emplear un subconjunto de la población (la muestra) para representar al conjunto completo, surge inevitablemente este tipo de error. Es una limitación intrínseca del proceso de muestreo, derivada de la naturaleza inferencial del análisis, por lo que su reducción constituye un objetivo fundamental.

Inferencia estadística.- Corresponde al procedimiento de generalizar las características observadas en la muestra a toda la población. Este proceso representa una rama especializada de la estadística, distinguiéndose así las dos grandes áreas: la estadística descriptiva (análisis de la muestra) y la inferencial (generalización a la población).

1.4.2. Tipos de Variables

Variables categóricas.- Sus valores (modalidades) no tienen una representación numérica inherente. Se subdividen en:

- Nominales: Categorías cualitativas sin orden intrínseco (ej: colores, estados civiles, tipos de material)
- Ordinales: Categorías con jerarquía o secuencia (ej: niveles de prioridad, tallas de ropa, escalas de satisfacción)

Variables cuantitativas.- Admiten valores numéricos sobre los que pueden realizarse operaciones matemáticas. Se clasifican en:

- Discretas: Adquieren valores enteros y contables (ej: unidades vendidas, defectos por lote, número de empleados)
- Continuas: Pueden tomar cualquier valor dentro de un intervalo (ej: temperaturas, pesos, tiempos de producción)

También podemos clasificar las variables según su enfoque metodológico:

Variables dependientes.- Son aquellas cuyos valores están determinados por otras variables, según una relación hipotética que establecemos en nuestros modelos estadísticos. Por ejemplo, en un análisis de regresión lineal, la variable dependiente es el resultado que intentamos predecir o explicar en función de otras variables.

Variables independientes.- Son los factores que, según nuestro modelo, influyen o explican la variable dependiente. Mantienen sus valores independientemente de otras variables en el estudio. Por ejemplo: "Nota final en un examen" (dependiente) podría estar influenciada por "Horas de estudio" (independiente).

ISBN: 978-631-6557-56-8

Intermediarias u omitidas.- Estas variables, no consideradas en el estudio o modelo, actúan como factores ocultos que pueden influir en la variable dependiente sin ser detectadas. Su identificación es crucial para evitar establecer correlaciones engañosas o suposiciones de causalidad incorrectas. Por ejemplo, al analizar el rendimiento académico de los estudiantes, el nivel formativo de los padres -a menudo no incluido en el modelo- podría ser un factor determinante que distorsione los resultados aparentes.

Variables dicotómicas.- Se trata de un tipo especial de variable estadística que solo puede adoptar dos valores posibles. Resultan particularmente útiles para representar situaciones de presencia/ausencia (1/0), como por ejemplo la detección de movimiento por un sensor o el éxito/fracaso en una prueba. Este grupo incluye las denominadas variables binarias, ampliamente utilizadas en modelos estadísticos y análisis de datos.

1.4.3. Diseño de experimentos

Existen dos tipos fundamentales de investigaciones estadísticas:

Estudios observacionales.- Consisten en la recopilación de datos mediante la observación directa, sin realizar ninguna intervención o modificación sobre los sujetos estudiados. Un ejemplo característico son las encuestas donde simplemente se registran respuestas sin alterar las condiciones de los participantes.

Estudios experimentales.- Implican la aplicación deliberada de tratamientos o condiciones específicas a las unidades de estudio (denominadas ahora unidades experimentales), para posteriormente medir sus efectos. Este enfoque es particularmente común en bioestadística, donde se comparan sistemáticamente los resultados de diferentes intervenciones médicas.

1.4.4. Contraste de Hipótesis

El contraste de hipótesis constituye un método estadístico riguroso para evaluar la plausibilidad de afirmaciones acerca de una población, basándose en evidencia muestral. Esta técnica resulta particularmente valiosa por su enfoque práctico, permitiendo tomar decisiones fundamentadas sobre cuestiones específicas.

El proceso implica establecer dos proposiciones mutuamente excluyentes:

Hipótesis nula (H_0): Representa la afirmación inicial que se somete a prueba, típicamente expresando "ningún efecto" o "ninguna diferencia"

Hipótesis alternativa (H_1): Corresponde a la negación de H_0 y se acepta cuando los datos proporcionan evidencia suficiente para rechazar la hipótesis nula

ISBN: 978-631-6557-56-8

El concepto de **significancia estadística** juega un papel central en las pruebas de hipótesis y está directamente vinculado con los niveles de significación utilizados en los intervalos de confianza.

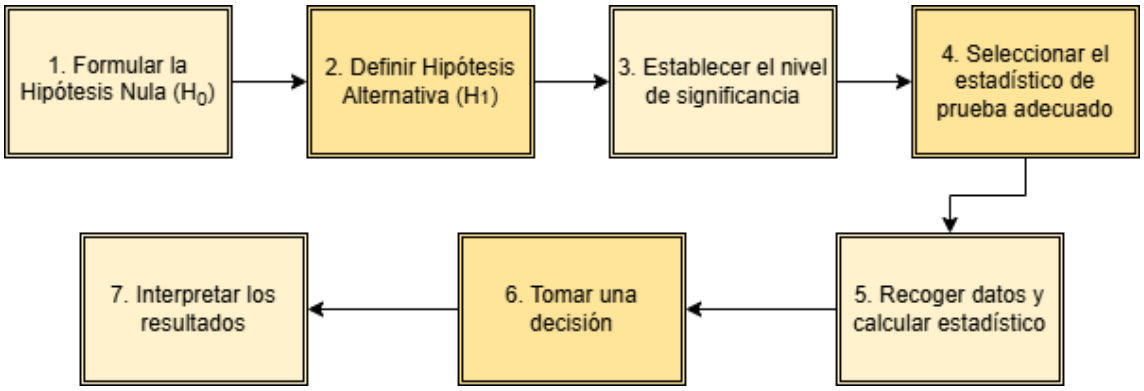
Se alcanza **significancia estadística** cuando el valor del estadístico de prueba supera un umbral crítico, lo que lleva al rechazo de la hipótesis nula (H_0). Este umbral está determinado por el nivel de significación α que representa la probabilidad máxima aceptable de cometer un error tipo I (rechazar (H_0) cuando en realidad es verdadera).

El **estadístico de contraste** es cualquier medida calculada a partir de los datos muestrales que permite evaluar la validez de (H_0) frente a (H_1). Su comparación con los valores críticos asociados a α determina si los resultados son estadísticamente significativos y, por tanto, si se rechaza o no la hipótesis nula.

Pasos para realizar un contraste de hipótesis

El procedimiento para contrastar hipótesis sigue una estructura sistemática que garantiza rigurosidad en las conclusiones. A continuación, se detallan los pasos en la Figura 4:

Figura 4: Pasos para realizar un contraste de hipótesis



Fuente: Autores

Explicación del proceso:

i.) **Formular la hipótesis nula** (H_1). (Expresar de manera simbólica la afirmación que se desea poner a prueba).

ii.) **Definir la hipótesis alternativa (H_1).** (Debe ser complementaria a (H_0) , de modo que no exista solapamiento entre ambas). Si (H_0) es falsa, (H_1) debe ser verdadera, y viceversa.

iii.) **Establecer el nivel de significancia (α).** Fijar la probabilidad máxima aceptable de cometer un error Tipo I (rechazar (H_0) cuando es verdadera).

ISBN: 978-631-6557-56-8

Valores típicos: α=0.05 α=0.05 (5%) o α=0.01 α=0.01 (1%).

iv.) **Seleccionar el estadístico de prueba adecuado.** Conocer su distribución muestral bajo (H_0) (asumiendo que (H_0) es cierta).

v.) **Recoger datos y calcular el estadístico**

- **Método tradicional**. Determinar las regiones de aceptación y rechazo según α.
- **Cálculo del estadístico** Calcular la probabilidad (p) de obtener un valor igual o más extremo que el estadístico observado, si (H_0) fuera cierta.

vi.) **Tomar una decisión**

Si se usa el método de regiones:

- **Aceptar** (H_0) si el estadístico cae en la zona de aceptación.
- **Rechazar** (H_0) si cae en la zona de rechazo.

Si se usa el p-valor:

- **Rechazar** (H_0) si $p \leq \alpha$ fijado
- **No rechazar** (H_0) si $p > \alpha$ fijado.

vii.) **Interpretar los resultados.** Si se rechaza (H_0) , se concluye que hay evidencia estadística a favor de (H_1) .Si no se rechaza (H_0) , no se prueba que sea cierta, solo que no hay suficiente evidencia en contra.

1.4.5. Medidas de Precisión de la clasificación

En el ámbito de la Industria 4.0 es frecuente el empleo de métodos de clasificación, muchos de los cuales se fundamentan en técnicas de machine learning (ML) o inteligencia artificial (IA) - siendo el aprendizaje automático una rama dentro de la IA-. Por ello, resulta importante poder evaluar el rendimiento de estos sistemas. En este contexto, esta sección introduce diversas métricas que permiten valorar la exactitud de un clasificador, con el fin de determinar su eficacia y comportamiento.

La **matriz de confusión** es una tabla que permite resumir y analizar el rendimiento de un modelo de clasificación, especialmente en aprendizaje supervisado. En ella se contabilizan las predicciones acertadas y las que no coinciden con la clase real. Específicamente, si existen n clases, la matriz tendrá dimensiones $n \times n$, y cada celda $p(i,j)$ indica cuántas veces las instancias de la clase i fueron asignadas por el modelo a la clase j.

La **exactitud** (*accuracy*) mide qué tan cerca se encuentra el resultado obtenido del valor real. Se calcula como la proporción de instancias que el modelo clasifica correctamente, dividiendo las predicciones correctas por el total de predicciones realizadas. Sin embargo, esta métrica no es adecuada cuando los datos están desbalanceados, es decir, cuando las clases no tienen una

ISBN: 978-631-6557-56-8

cantidad similar de instancias. En estos casos, es preferible usar métricas como precisión, *recall* o F1, que ofrecen una evaluación más precisa del desempeño del modelo. Por otro lado, el concepto de sesgo (*bias*) se relaciona con la exactitud y se define como la diferencia entre el promedio de las predicciones del modelo y el valor real que se desea estimar. El sesgo suele originarse por suposiciones incorrectas, lo que puede provocar que el modelo ignore información importante, generando un desajuste o *underfitting*.

La prevalencia (*prevalence*) representa la probabilidad de que una instancia sea positiva dentro del total de la muestra analizada.

$$prevalence = \frac{TP + FN}{TP + TN + FP + FN}$$

La **precisión**, también llamada valor predictivo positivo, expresa la proporción de instancias clasificadas correctamente como positivas en relación con el total de instancias que realmente son positivas.

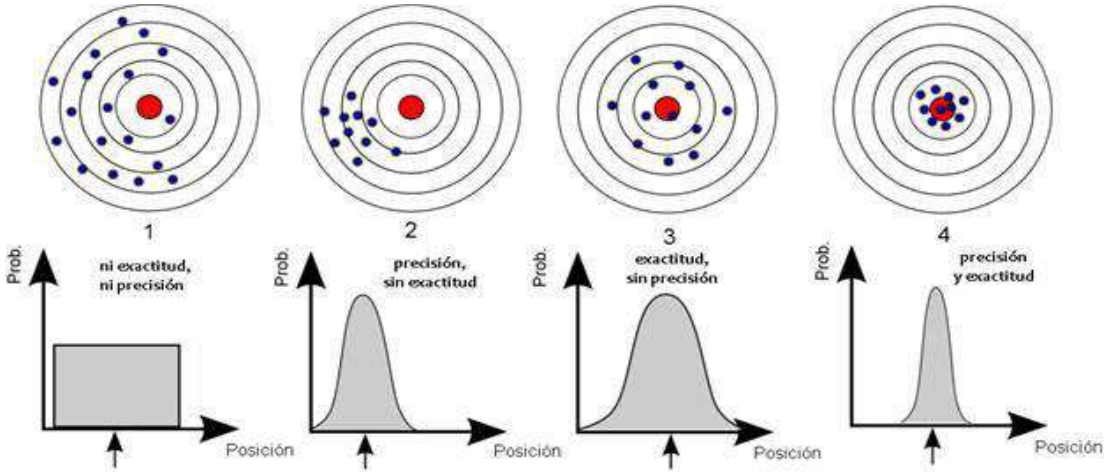
$$precision = \frac{TP}{TP + FP} = \frac{TP}{instancias\ positivas\ reales}$$

Está vinculado a la varianza del modelo, que refleja qué tan dispersas son sus predicciones. Cuando la varianza es elevada, es posible que el modelo se haya adaptado en exceso a los datos de entrenamiento, lo que suele provocar un bajo rendimiento al enfrentarse a datos de prueba o datos nuevos, fenómeno conocido como sobre-entrenamiento u *overfitting*.

Es importante distinguir entre la exactitud y la precisión de un sistema. La Figura 5 muestra claramente esta diferencia: un modelo se considera exacto cuando sus clasificaciones se acercan al valor objetivo, mientras que es preciso cuando las clasificaciones de instancias similares se concentran alrededor de un mismo valor.

ISBN: 978-631-6557-56-8

Figura 5: Repartición entre las instancias en función de su exactitud y precisión en la clasificación.



Fuente: [10]

La **sensibilidad** *(sensitivity o recall*) mide la capacidad del modelo para distinguir correctamente las instancias positivas de las negativas, centrándose en los verdaderos positivos. También se le denomina tasa de aciertos o tasa de verdaderos positivos (*true positive rate* – TPR – TP *rate*) y representa la proporción de instancias realmente positivas que fueron clasificadas de manera correcta por el modelo.

$$recall = TP\ rate = \frac{TP}{TP + FN} = \frac{TP}{resultados\ del\ clasificador}$$

La **especificidad** (*specificity*) mide la capacidad del modelo para diferenciar correctamente los casos negativos de los positivos, enfocándose en los verdaderos negativos. También se conoce como tasa de verdaderos negativos (*true negative rate* – TNR – TN *rate*) y representa la proporción de instancias realmente negativas que fueron clasificadas correctamente. Esta métrica es complementaria a la sensibilidad o *recall*.

La **tasa de falsos positivos** (*false positive rate* – FPR – FP *rate*) representa la proporción de instancias negativas que el modelo clasificó incorrectamente como positivas.

$$FP\ rate = 1 - specificity = \frac{FP}{TN + FP}$$

La **tasa de falsos negativos** (*false negative rate* – FNR – FN *rate*) es la proporción de instancias positivas que el modelo clasificó incorrectamente como negativas.

$$FP\ rate = \frac{FN}{TP + FN}$$

ISBN: 978-631-6557-56-8

La **puntuación F1** es una métrica que combina la precisión y la exhaustividad. Su valor máximo es 1, lo que indica una precisión y recuperación perfectas, y el mínimo es 0, lo que sugiere que el modelo confunde completamente las clases. Un valor de 0.5 indica que el modelo no puede distinguir adecuadamente entre las clases. En general, la puntuación F1 evalúa tanto la precisión como la eficacia del modelo.

$$F1 = 2\,x\,\frac{TP}{2TP + FP + FN} = 2\,x\,\frac{precision\,x\,recall}{precision + recall}$$

La **curva ROC** (*receiver operating characteristic curve*) es un gráfico que evalúa el rendimiento de un modelo de clasificación al representar la tasa de falsos positivos (FP *rate*) en el eje X y la tasa de verdaderos positivos (TP *rate*) en el eje Y, en función de diferentes umbrales de clasificación. Dado que calcular esta curva puede ser ineficiente debido a la necesidad de evaluar el modelo en múltiples umbrales (generalmente mediante regresión logística), se utiliza el algoritmo AUC (*area under the curve*). El AUC indica que, a mayor valor, mejor será la capacidad del clasificador para clasificar correctamente las instancias.

El índice **Kappa** se utiliza para evaluar los resultados de un clasificador comparándolos con la precisión que se esperaría de un clasificador aleatorio. En algunos programas, como Weka, esta comparación se realiza en función de cómo se han etiquetado las muestras, utilizando esta información como la fuente de verdad. De esta forma, un clasificador con un 90 % de precisión y una precisión esperada del 50 % tendría mayor mérito que otro con la misma precisión, pero con una precisión esperada del 80 %. Este valor puede calcularse utilizando los datos de la matriz de confusión, tomando en cuenta tanto la precisión observada (resultados del clasificador) como la precisión esperada (instancias positivas reales).

$$Kappa = \frac{Precisión\ observada - precisión\ esperda}{1 - precisión\ esperada}$$

1.5. Métodos de exploración univariable

Las técnicas de análisis univariante son una herramienta inicial clave para entender los datos, ya que son sencillas de aplicar y ofrecen una visión general de las variables que se desean analizar.

1.5.1. Distribución de frecuencias

La distribución de frecuencias aborda la etapa de organización y representación de los datos. La medida más sencilla que se puede realizar con los datos es contar cuántas veces se repite un valor o una categoría de una variable. Estas cantidades de repeticiones se conocen como frecuencias.

ISBN: 978-631-6557-56-8

Las frecuencias pueden clasificarse en: Absolutas, Relativas, Absolutas acumuladas, Relativas acumuladas.

1.5.2. Medida que resumen la información

La estadística proporciona diversas formas de resumir la información de una distribución en números. Estos valores son medidas que condensan las características de una muestra o población. En esta sección se presentarán las más relevantes.

Medidas de tendencia central.- Las primeras medidas que se estudiarán son aquellas relacionadas con el concepto de centro de la distribución de los datos. Son valores que se encuentran en el medio o la mitad de un conjunto o distribución de datos. Estas medidas buscan identificar valores que representen de manera general a todos los datos.

Medidas de dispersión.- Las medidas de dispersión nos muestran cuánto se desvían los datos, lo cual es importante para entender cómo se distribuye el conjunto de datos. La medida de dispersión más simple es el rango, que se calcula como la diferencia entre el valor mínimo y el valor máximo de las observaciones.

Medidas de posición y forma.- Las medidas de posición son útiles para determinar si un valor se encuentra alejado o cercano a la media, lo que nos permite evaluar qué tan extremo es en comparación con la mayoría de los datos del conjunto.

1.5.3. Datos atípicos y análisis exploratorio de datos

Los valores atípicos o extremos (*outliers*) son aquellos que se alejan significativamente de la mayoría de los datos. En un diagrama de caja y bigotes, los valores atípicos se marcan con puntos y los valores extremos con estrellas. Para diferenciar entre ambos, se utiliza el límite de 3 veces el rango intercuartil (RIC). Así, los datos que están entre 1.5 y 3 veces el RIC se consideran simples atípicos, mientras que los que superan las 3 veces el RIC se marcan como valores extremos. Aunque esta distinción es utilizada en programas como SPSS, en general se agrupan todos como datos atípicos o extremos, según el término que se prefiera, y es la clasificación comúnmente empleada en los gráficos de cajas y bigotes.

1.5.4. Distribución Normal

La distribución de probabilidad teórica más conocida es, sin lugar a dudas, la distribución normal. En el pasado, se pensaba que todas las variables aleatorias eran continuas, aunque hoy se sabe que esto no es cierto. Sin embargo, muchas de las variables aleatorias presentes en la naturaleza

ISBN: 978-631-6557-56-8

siguen una distribución normal [11]. A pesar de ello, la función matemática que describe esta distribución es compleja y no resulta práctica para su manejo.

1.6. Análisis estadístico bivariable

Es común desear estudiar la relación estadística entre dos variables. Existen diversas técnicas que se adaptan al tipo de información que se desea analizar. En esta sección se presentan algunas de las más importantes, tanto para el análisis de variables nominales, ordinales o métricas, como para la combinación de alguna de estas.

1.6.1. Tablas de frecuencia

Para resumir la información en tablas de frecuencias bidimensionales, se utiliza un formato en el que se coloca cada variable en un eje, y las frecuencias correspondientes a cada combinación de pares de categorías de ambas variables se sitúan en el interior de la tabla.

1.6.2. Covarianza

Al igual que en el caso de las variables unidimensionales, para las variables bidimensionales también existen estadísticos, pero en este caso no miden las propiedades individuales, sino las conjuntas de ambas variables en el conjunto de datos bivariados. El estadístico más relevante es la covarianza, que indica cómo varían juntas las dos variables. Técnicamente, se define como la media aritmética de los productos de las desviaciones de cada variable respecto a su media.

1.6.3. Correlación

Una de las relaciones más comunes entre dos variables es la lineal, probablemente debido a su simplicidad. Podemos observar los gráficos de dispersión para identificar un posible patrón lineal. Pero ¿cómo determinar si la magnitud de este patrón es significativa? Para ello, disponemos de un estadístico que mide esa magnitud. El coeficiente de correlación de Pearson es una medida que evalúa la fuerza de la relación lineal entre dos variables cuantitativas.

1.6.4. Regresión

Al igual que queríamos conocer la intensidad de la relación lineal, también nos interesa determinar la forma explícita de esa relación, es decir, la ecuación lineal que mejor describe la asociación observada. Este análisis se aborda en lo que se conoce como regresión lineal.

Conocer la ecuación de regresión nos permitirá hacer predicciones de una de las variables basándonos en los valores de la otra. Sin embargo, en este caso, la ecuación de regresión lineal

ISBN: 978-631-6557-56-8

está restringida al supuesto de que existe una variable que explica la relación, llamada variable explicativa, y otra que es explicada, conocida como la variable respuesta.

El método matemático que nos permite calcular las ecuaciones de regresión y sus componentes es el método de los mínimos cuadrados (MMC). El MMC es fundamental en la historia de la estadística, ya que su descubrimiento impulsó de manera significativa su evolución. Matemáticos clave como Legendre y Gauss fueron fundamentales en el desarrollo de este método [12].

1.6.5. Análisis de datos categóricos

Las **tablas de contingencia** tienen como objetivo resumir datos categóricos. Al analizarlas, lo más común es determinar si existe alguna relación entre la variable ubicada en las filas y la que se encuentra en las columnas de la tabla, incluyendo la distribución de frecuencias conjunta. Para este propósito, se utiliza la prueba de χ^2, que está diseñada para muestras discretas, aunque también es aplicable a muestras continuas que puedan ser agrupadas. En este caso, se demostrará una relación entre las variables si se puede rechazar la hipótesis nula basándose en el valor proporcionado por el estadístico.

La **prueba de Fisher** se emplea para estudiar la asociación entre dos variables cualitativas y determinar si las proporciones de una variable varían según el valor que tome otra variable. Generalmente se utiliza para comparar dos variables categóricas con dos niveles cada una (por ejemplo, si la formación de los trabajadores incrementa su productividad). La hipótesis nula establece que las variables son independientes [13].

1.6.6. Comparación entre grupos de muestras

La prueba **t de *Student*** es una distribución de probabilidad empleada cuando el tamaño de la muestra es pequeño (usualmente menor de 30). Se utiliza al estimar la media de una población con distribución normal, de la cual se desconoce la desviación estándar, y permite determinar si la diferencia entre las medias de dos poblaciones es estadísticamente significativa.

El **test Z** se utiliza cuando el número de muestras es suficientemente grande (generalmente mayor de 30). Esta prueba se basa en la premisa de que, a medida que el tamaño de la muestra aumenta, la distribución de los datos tiende a aproximarse a una distribución normal. El test Z funciona de manera similar al test t de *Student*, pero su principal diferencia es que puede detectar diferencias más pequeñas en los datos, reduciendo así el riesgo de cometer errores de tipo II.

El **test ANOVA** es una técnica que se utiliza para verificar si las medias de dos o más grupos son significativamente diferentes entre sí. Se aplica cuando la variable independiente tiene más de dos

ISBN: 978-631-6557-56-8

niveles, es decir, cuando se desea comparar tres o más grupos. Además, ANOVA también sirve para examinar el impacto de los factores en la variabilidad de una variable.

La hipótesis nula en ANOVA establece que las medias de los grupos comparados son iguales, es decir, que las muestras provienen de la misma población con las mismas medias y varianzas. En contraste, la hipótesis alternativa plantea que al menos una media de los grupos es diferente.

El test de *Mann-Whitney*, también conocido como **test U**, test de suma de rangos de *Wilcoxon* o test de *Wilcoxon-Mann-Whitney*, se emplea como una alternativa a la prueba t de *Student* cuando los datos no siguen una distribución normal.

Es una prueba no paramétrica que se utiliza para evaluar si existe una diferencia significativa entre dos poblaciones con muestras independientes, cuyas variables pueden ser ordinales o continuas. La prueba verifica si las muestras provienen de poblaciones con distribuciones similares. Es importante que las poblaciones tengan el mismo tipo de distribución y que los datos sean ordenables para que el test sea válido.

El **test de la mediana** es una variante del test $\chi 2$, ya que se basa en este. Su objetivo es utilizar la mediana como una medida de comparación para evaluar si dos muestras provienen o no de la misma población. Este test es no paramétrico y requiere que las variables sean ordinales. Resulta útil cuando se sabe que las poblaciones pueden contener valores extremos o estar sesgadas.

El **test de los rangos con signo**, también conocido como test de rango de signos de *Wilcoxon*, se utiliza en situaciones donde los datos solo pueden ser medidos en una escala ordinal, como ocurre con los sistemas de puntuación, en lugar de una escala de intervalo, que es necesaria para el test U.

El **test de Friedman** es la versión no paramétrica de ANOVA para datos dependientes. Representa una extensión de la prueba de *Mann-Whitney-Wilcoxon* (WMW) para más de dos grupos, comparando sus medianas. Esta prueba es adecuada cuando los datos deben estar ordenados para que tengan un sentido, es decir, cuando poseen un orden natural [14].

1.7. Técnicas complementarias de estudio

Como parte del resumen de técnicas de análisis de datos que se presenta en este tema, y antes de abordar la descripción de algunos de los gráficos más relevantes, esta sección reúne de manera concisa otras técnicas que el estudiante puede tener en cuenta según el tipo de problema que deba resolver.

ISBN: 978-631-6557-56-8

1.7.1. Multivariante

Las técnicas multivariantes permiten analizar simultáneamente las relaciones entre más de dos variables. La elección de la prueba a utilizar dependerá de los datos disponibles y del objetivo del estudio. Según [15], estas técnicas se pueden clasificar siguiendo tres criterios generales. Variables categóricas o con escala métrica, métodos de dependencia e interdependencia y número de variables.

1.7.2. Problemas de Clasificación

Un reto habitual en el análisis de datos es la clasificación de los mismos. Para abordar esta tarea existen diversas técnicas, siendo especialmente destacadas aquellas que se apoyan en la inteligencia artificial (las cuales se profundizarán en su tema respectivo). Dentro de estas técnicas se incluyen el análisis discriminante, la reducción de dimensionalidad, el análisis de componentes principales (PCA), los árboles de clasificación o regresión (CART), *Random Forest*, las reglas de asociación y el *clustering*.

1.7.3. Redes Neuronales

Las redes neuronales son una de las técnicas más sofisticadas y, al mismo tiempo, de las más utilizadas hoy en día para el análisis de datos. Estas buscan imitar el funcionamiento del cerebro humano con el fin de construir modelos de aprendizaje que permitan tomar decisiones con un alto grado de precisión. Dentro de esta rama de la inteligencia artificial existen diversos modelos que se ajustan de manera específica a distintos tipos de problemas, como el reconocimiento de texto, voz o imágenes, entre otros.

1.7.4. Test A/B

Esta es una de las técnicas de análisis más empleadas en el ámbito del marketing digital, especialmente en lo relacionado con la experiencia del usuario (UX), ya que permite identificar qué alternativa resulta más efectiva para los usuarios. Suele aplicarse en la validación de hipótesis durante el lanzamiento de nuevos productos, modificaciones en la interfaz o en el mensaje de una campaña, entre otros casos. Se puede considerar como una técnica bivariable, ya que consiste en presentar dos opciones distintas a diferentes grupos de usuarios y, a partir de la respuesta de cada uno, determinar cuál es la alternativa más adecuada.

ISBN: 978-631-6557-56-8

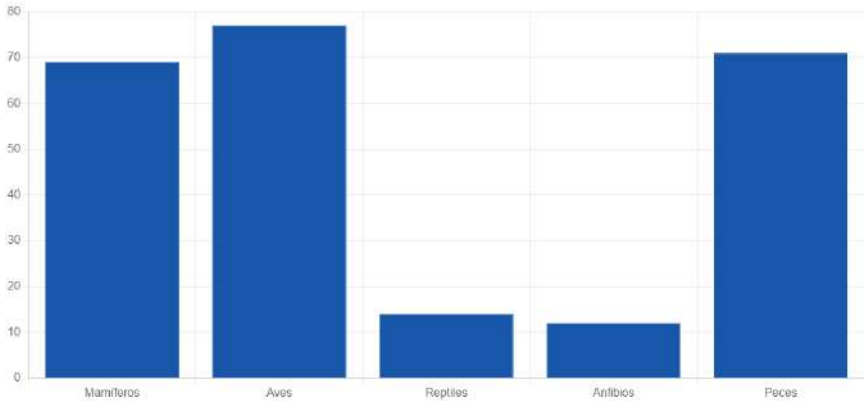
1.8. Visualización gráfica de resultados

Como bien dice el refrán, a veces una imagen comunica más que mil palabras, y en el caso de la estadística podríamos adaptarlo diciendo que «es preferible un buen gráfico antes que mil tablas de frecuencias». Esto se debe a que, en general, nuestra mente capta y comprende con mayor rapidez la información cuando se presenta de forma visual que cuando se encuentra codificada en formatos más complejos o analíticos. Para elegir el tipo de gráfico más adecuado al representar un conjunto de datos, es importante considerar el tipo de variable que se desea mostrar. A continuación, se describirán distintos gráficos que resultan más o menos apropiados según la naturaleza de la variable que se pretenda ilustrar.

1.8.1. Diagrama de barras

Cuando queremos representar variables de tipo categórico (aunque esta clasificación no es estrictamente formal, resulta práctica en su uso), ya sean cualitativas nominales u ordinales o cuantitativas discretas, considerando en este último caso cada valor como una categoría, recurrimos a los diagramas de barras. En otras palabras, todas las variables pueden representarse mediante diagramas de barras, salvo aquellas que son continuas.

Figura 6: Diagrama de barras para indicar las especies en peligro de extinción en EE.UU.



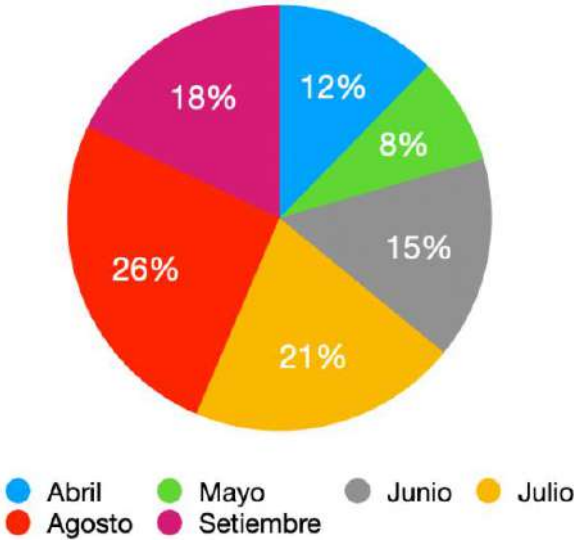
Fuente: [16]

ISBN: 978-631-6557-56-8

1.8.2. Gráficos de sectores

Este tipo de gráfico es el más adecuado para representar variables cualitativas de forma visual. También es conocido como gráfico circular, gráfico de porciones, gráfico de tarta o pie chart en inglés. Es una representación bastante común cuyo requisito principal es que los porcentajes de las diferentes categorías sumen siempre el 100 %. Cada sector del círculo es proporcional al porcentaje que representa dentro del conjunto total. Es recomendable utilizarlo cuando la cantidad de categorías no es demasiado elevada, ya que si las diferencias entre ellas son mínimas, podría ser más conveniente optar por un diagrama de barras, a continuación, en la Figura 7 se muestra un ejemplo de este tipo de gráficas.

Figura 7: Gráfico de sectores representando porcentaje por mes



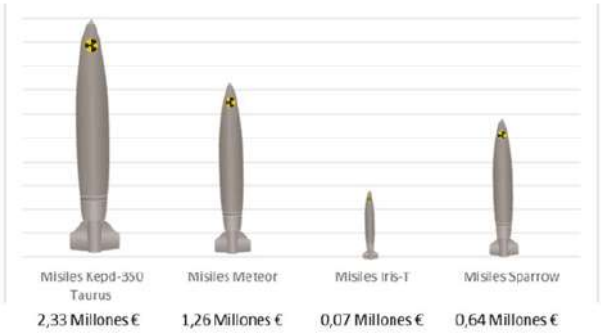
Fuente: [17]

1.8.3. Pictograma

Los pictogramas son gráficos que utilizan dibujos para reforzar visualmente las diferencias, ya que los elementos gráficos aportan un valor simbólico que capta la atención con facilidad. La Figura 8 muestra un ejemplo de este tipo de representación. Un error frecuente al trabajar con pictogramas es asignar el valor de cada categoría directamente al tamaño del dibujo, cuando en realidad lo correcto es que el área de cada figura sea proporcional a la magnitud que representa.

ISBN: 978-631-6557-56-8

Figura 8: Pictograma representativo de gasto militar.



Adaptado de: [18].

1.8.4. Histograma

Cuando se desea representar variables cuantitativas continuas, es común utilizar el histograma, un tipo de gráfico que, aunque se asemeja al diagrama de barras, refleja la continuidad de los datos mediante barras adyacentes sin separación entre ellas. Este gráfico se emplea, especialmente, cuando la información ha sido agrupada en intervalos, que es la forma habitual de trabajar con variables de este tipo.

Figura 9: Ejemplo de Histograma

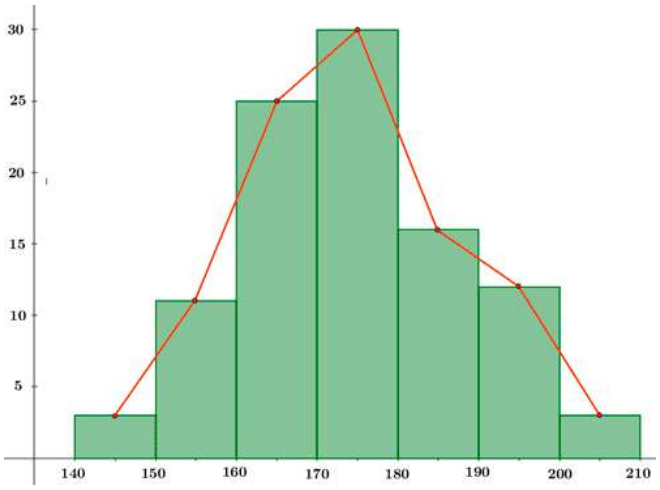
Adaptado de:[23]

1.8.5. Polígono de Frecuencias

El polígono de frecuencias es un gráfico que se utiliza con menor frecuencia que el histograma. Este se construye uniendo los puntos medios de las barras del histograma, como se observa en su superposición con dicho gráfico. Resulta especialmente útil cuando se pretende visualizar tendencias, ya que presenta la información mediante una línea continua. En la Figura 10 se muestra un ejemplo de polígono de frecuencias.

ISBN: 978-631-6557-56-8

Figura 10: Polígono de Frecuencias, distribución de tallas en alumnos

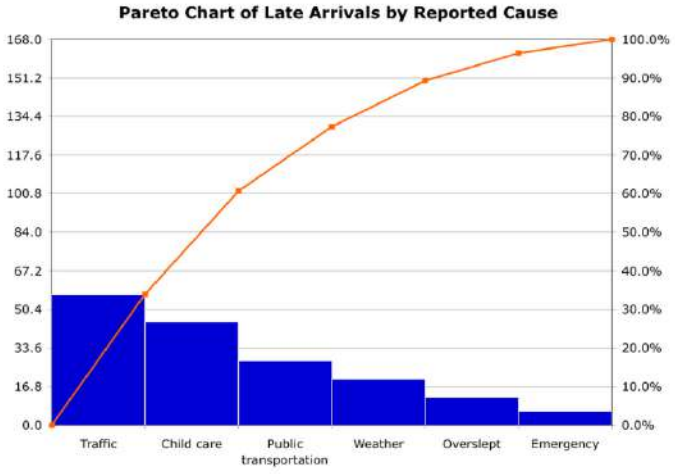


Fuente: [20]

1.8.6. Gráfico de Pareto

El gráfico de Pareto es una representación que combina un diagrama de barras, que muestra los valores individuales en orden descendente, junto a una línea que indica el total acumulado. El eje vertical izquierdo señala la frecuencia de ocurrencia, mientras que el derecho muestra el porcentaje acumulado de esas ocurrencias. Este tipo de gráfico permite identificar, por ejemplo, qué problemas conviene resolver para reducir en mayor medida su frecuencia, como las causas de piezas defectuosas, los productos con más devoluciones o los motivos de retraso al trabajo, como se muestra a continuación en la Figura 11.

Figura 11: Ejemplo de gráfico de Pareto



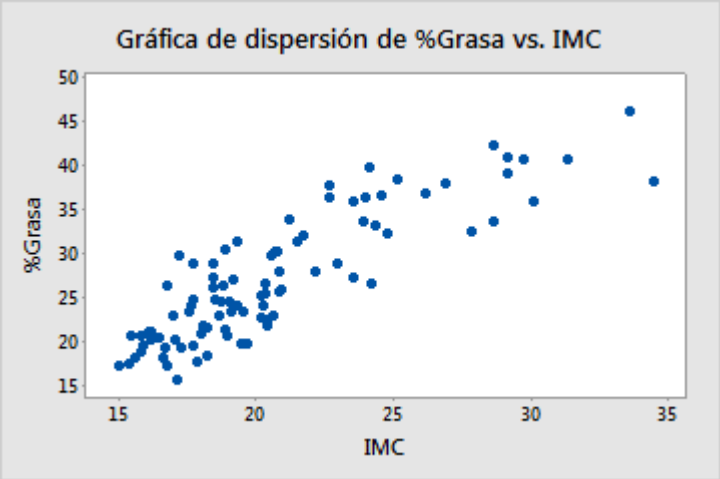
Adaptado de: [21]

ISBN: 978-631-6557-56-8

1.8.7. Dispersión

Este tipo de gráficos resulta especialmente útil para mostrar los valores de un individuo en dos variables continuas, Además, cuando se incluye una variable cualitativa, esta puede representarse diferenciando los puntos mediante colores o símbolos, tal como se muestra en la Figura 12.

Figura 12: Ejemplo de diagrama de dispersión

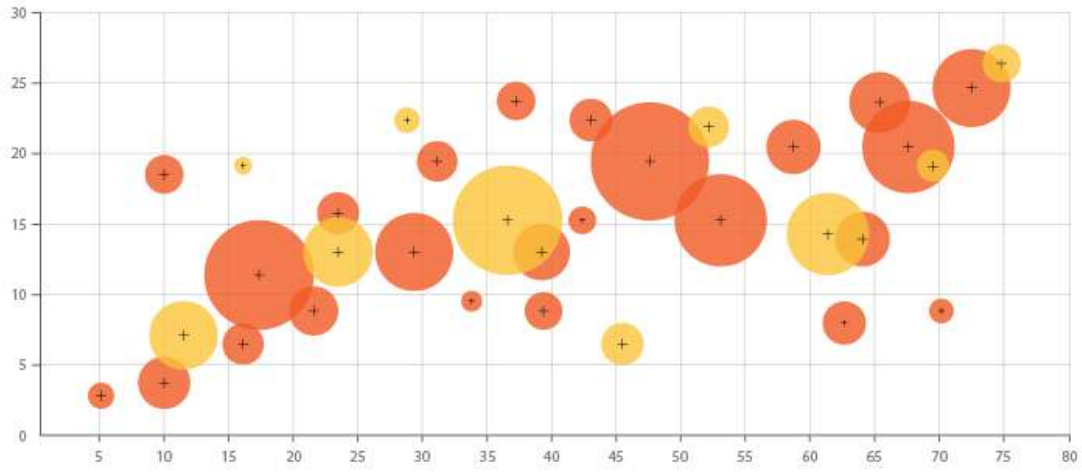


Adaptado de: [23]

1.8.8. Gráfico de burbujas

Un gráfico de burbujas se asemeja al diagrama de dispersión, ya que permite visualizar la distribución o relación entre variables; sin embargo, añade una tercera dimensión que se representa mediante el tamaño de cada burbuja.

Figura 13: Ejemplo de gráfico de Burbujas



Fuente: [23]

ISBN: 978-631-6557-56-8

1.8.9. Gráfico de Línea (Serie Temporal)

Las series temporales, conocidas en inglés como *time plot*, son una de las formas de representación gráfica más utilizadas. Es frecuente encontrarlas en medios económicos para ilustrar, por ejemplo, la evolución de índices bursátiles. En este tipo de gráfico, los valores se conectan mediante una línea que refleja su comportamiento a lo largo del tiempo. A veces, el área bajo la curva se rellena con color o textura, dando lugar a lo que se conoce como gráfico de área, a continuación en la Figura 14 se muestra un ejemplo.

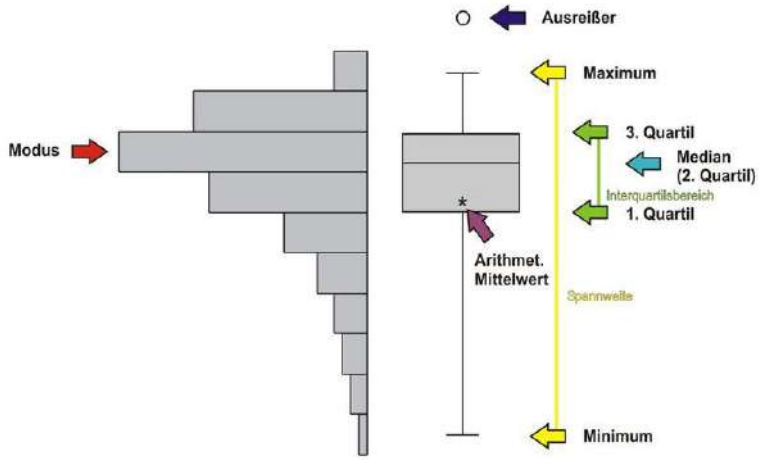
Figura 14: Ejemplo de serie temporal

Fuente: [23]

1.8.10. Diagrama de cajas y Bigotes

Los diagramas de cajas y bigotes son útiles para visualizar aspectos como la distribución de los datos, su simetría y la presencia o localización de valores atípicos. Este tipo de gráfico se construye a partir de los cuartiles, junto con los valores máximos y mínimos. La caja representa el rango intercuartílico (RIC), es decir, el conjunto de datos comprendido entre el primer cuartil ($Q1$) y el tercer cuartil ($Q3$). Los bigotes, por su parte, se extienden hasta 1.5 veces el RIC desde los cuartiles, alcanzando las posiciones $Q3 + 1.5 \cdot RIC$ y $Q1 - 1.5 \cdot RIC$.

Figura 15: Ejemplo de diagrama de cajas y bigotes



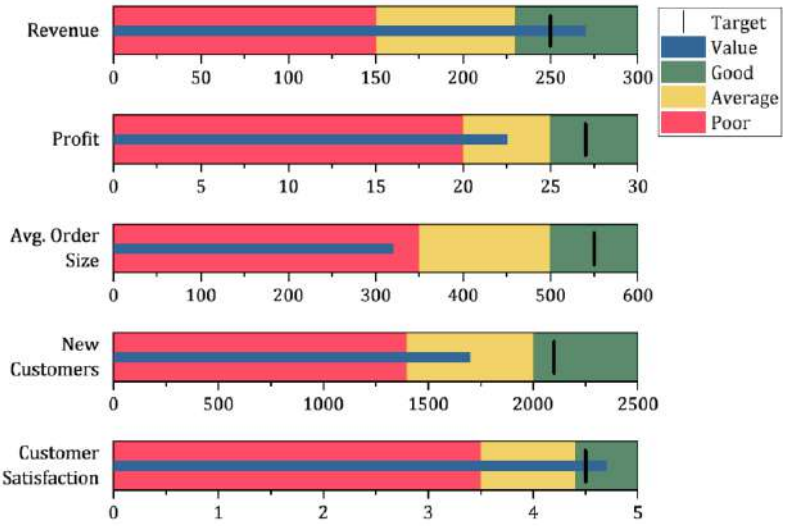
Fuente: [24]

ISBN: 978-631-6557-56-8

1.8.11. Gráficos de bala

Un gráfico de bala *(bullet graph)* es una representación visual compacta que muestra el progreso hacia un objetivo combinando una barra que indica el valor actual, una línea que marca la meta u objetivo y áreas sombreadas que representan rangos cualitativos como insuficiente, aceptable y óptimo; este tipo de gráfico, diseñado como una alternativa más informativa y eficiente a los medidores tradicionales, permite visualizar de forma clara el rendimiento en relación con un objetivo y es muy utilizado en *dashboards* o paneles de control por su capacidad de mostrar mucha información en poco espacio.

Figura 16: Ejemplo de gráfico de bala



Fuente: [23]

1.8.12. Mapa coroplético

Los mapas coropléticos son un tipo de representación cartográfica en la que se asignan colores o grados de saturación a distintas regiones geográficas según el valor de una variable, permitiendo visualizar de forma intuitiva patrones espaciales, comparativas y concentraciones de datos; este tipo de mapas puede considerarse un caso específico de los mapas de calor, ya que ambos utilizan variaciones cromáticas para representar intensidad, aunque en los coropléticos la segmentación se realiza por zonas geográficas definidas como países, provincias o distritos, un ejemplo de este tipo de gráficos se muestra en la Figura 17.

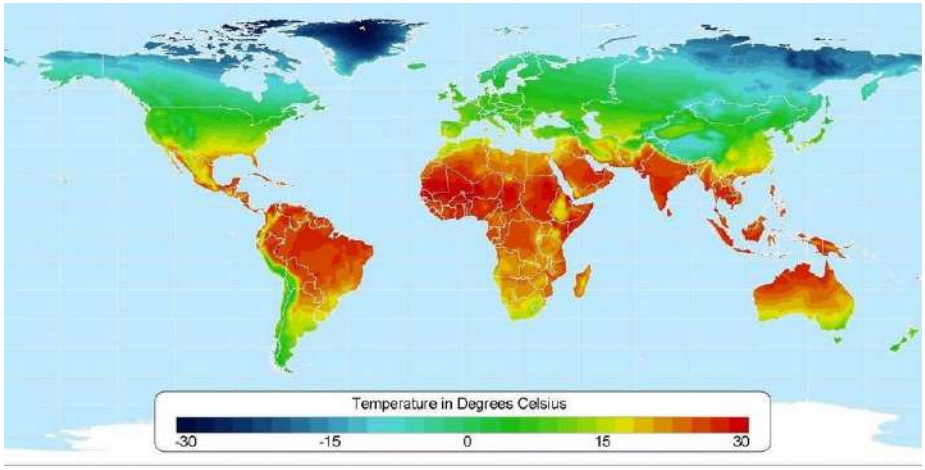
1.8.13. Mapa de calor

Un mapa de calor es una representación gráfica que muestra la relación entre dos elementos, donde la información sobre la intensidad o calificación de esa relación se traduce visualmente mediante una escala de colores o niveles de saturación, facilitando así la identificación rápida de

ISBN: 978-631-6557-56-8

patrones, concentraciones y contrastes dentro de los datos analizados, a continuación se muestra un ejemplo de un mapa de calor en la Figura 17.

Figura 17: Ejemplo de mapa de calor

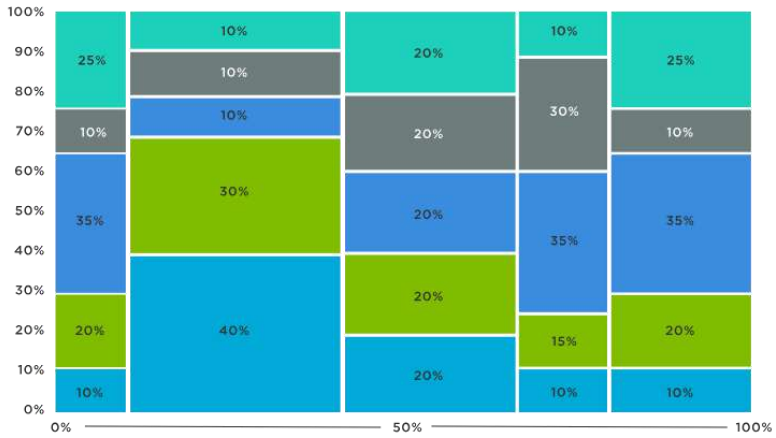


Fuente: [25]

1.8.14. Gráficos Mekko

El gráfico Mekko, también conocido como gráfico de Marimekko, permite comparar valores, medir la composición de cada uno y mostrar cómo se distribuyen los datos en cada segmento; es similar a un gráfico de barras apiladas, pero con la diferencia de que el eje X no representa una secuencia temporal sino que captura una dimensión adicional, siendo especialmente útil cuando se desea segmentar visualmente una población o dividir un conjunto de datos en proporciones significativas para su análisis.

Figura 18: Ejemplo de grafico Mekko



Fuente: [26]

1.8.15. Selección del grafico ideal

Uno de los problemas habituales al enfrentarse a la representación gráfica de un conjunto de datos es no saber por dónde empezar, ya que existen múltiples formas de abordar esta tarea; por ello, la

ISBN: 978-631-6557-56-8

elección del gráfico más adecuado dependerá fundamentalmente del tipo de variable que se desea representar y del objetivo del análisis, siendo la Tabla 1 una guía práctica que facilita esta elección al ayudar a seleccionar el gráfico que mejor se adapte a las características de los datos disponibles.

Tabla 1: Tipos de variables y opciones gráficas recomendadas para cada una de ellas.

Tipo de Variable		**Opciones Gráficas**	
Cualitativa	Normal	Diagrama de barras, sectores, pictogramas	Diagrama de barras acumuladas
	Ordinal		
Cuantitativa	Discreta	Histogramas, Dispersión, (dos continuas), serie temporal	Polígono de frecuencias
	Continua		

Fuente: Autores

Dependiendo del propósito del análisis, ciertos gráficos serán más adecuados que otros. Para comparar valores, dimensiones o categorías, se pueden utilizar barras, Mekko, sectores, línea, dispersión o gráfico de bala. Si el objetivo es seguir tendencias a lo largo del tiempo, los gráficos de línea y histograma son los más apropiados. Para resaltar un valor unitario, el *scorecard* es la mejor opción. Cuando se busca ilustrar cómo está formado un conjunto de datos, los gráficos de sectores, barras apiladas o *Mekko* son los más útiles. Si el propósito es entender la distribución de los datos, incluyendo la identificación de valores atípicos o rangos, se recomienda el uso de dispersión, *Mekko*, burbujas, línea o barras. Para analizar tendencias, los gráficos de línea, histograma y barras son los más indicados. En caso de querer visualizar la relación entre variables o conjuntos de datos, los gráficos de dispersión, burbujas y línea son ideales. Para representar datos en un mapa, se debe optar por un mapa coroplético. Finalmente, si se necesita comprobar cómo de bien los datos se comportan frente a un objetivo, el gráfico de bala es la opción más adecuada.

1.9. Procesamiento de datos en entornos industriales 4.0

La estadística clásica, definida tradicionalmente como la ciencia que recolecta y analiza los datos, enfrenta importantes retos en el contexto actual de la Industria 4.0. Con el crecimiento exponencial de la capacidad de computación y la llegada del Big Data, la cantidad de información disponible para el análisis ha superado los límites tradicionales, creando una situación en la que la magnitud de los datos es tan grande que es difícil procesarlos, analizarlos y, lo más importante, aprender de ellos. A pesar de la gran disponibilidad de datos, uno de los mayores problemas es la incapacidad de extraer conocimiento útil de ellos. La estadística clásica no estaba preparada para este entorno masivo de datos, lo que requiere una adaptación de sus métodos y prácticas. Para poder ser útil en este contexto, la estadística debe evolucionar e incorporar nuevas técnicas de

ISBN: 978-631-6557-56-8

análisis que permitan gestionar eficientemente grandes volúmenes de datos, identificar patrones relevantes y extraer conclusiones significativas. Este desafío implica no solo el desarrollo de nuevas metodologías, sino también la integración de herramientas avanzadas como el análisis predictivo, el aprendizaje automático y la inteligencia artificial, que se convierten en aliados clave para transformar datos masivos en información útil dentro de la Industria 4.0.

1.9.1. Retos

La **excesiva cantidad de información** y datos en el contexto actual de Big Data presenta uno de los mayores desafíos para la estadística clásica. Los métodos estadísticos tradicionales no fueron diseñados para manejar grandes volúmenes de datos, lo que genera problemas significativos al intentar aplicar estos métodos a conjuntos masivos de información. Esto se debe a que el tiempo y los recursos necesarios para realizar los cálculos requeridos pueden resultar inviables, especialmente cuando se manejan millones o incluso miles de millones de registros. Ante esta situación, es importante desarrollar soluciones eficientes que permitan superar estas limitaciones. Por un lado, es necesario optimizar los métodos estadísticos clásicos para que sean más rápidos y escalables. Por otro lado, también se deben desarrollar nuevos enfoques y algoritmos estadísticos que puedan manejar eficientemente grandes cantidades de datos, permitiendo realizar análisis complejos sin comprometer la rapidez y la precisión. La implementación de tecnologías avanzadas, como el procesamiento paralelo, la computación en la nube y algoritmos diseñados específicamente para Big Data, es crucial para afrontar este reto y aprovechar todo el potencial de los datos disponibles.

La **complejidad** de los datos en entornos de Industria 4.0 representa otro desafío significativo para la estadística moderna. A pesar de la enorme cantidad de información disponible, los datos no solo son grandes en volumen, sino también muy complejos y difíciles de interpretar. Esta complejidad se debe principalmente a la diversidad de fuentes y tipos de datos, que provienen de una variedad de dispositivos como sensores, robots, máquinas, y usuarios en plataformas digitales. Estos datos pueden estar en formatos muy variados, como texto, imágenes, registros de actividad, datos en tiempo real, entre otros, lo que dificulta su procesamiento y análisis.

Este tipo de datos, frecuentemente conocidos como "huella digital", son generados automáticamente por los usuarios y sistemas a través de la interacción con dispositivos y aplicaciones, creando una rica, pero desordenada, base de información. La integración y el análisis de estos datos, que a menudo se encuentran dispersos y sin una estructura clara, requieren técnicas avanzadas de procesamiento y análisis, como el aprendizaje automático y la inteligencia artificial. Estos métodos permiten extraer patrones y relaciones significativas de datos aparentemente caóticos, pero aún existen retos en la gestión eficiente de la calidad de los datos y la interpretación

ISBN: 978-631-6557-56-8

correcta de las métricas y relaciones que realmente aportan valor para las decisiones en la Industria 4.0. Además, la alta velocidad de los datos en tiempo real y su constante actualización obliga a que los métodos estadísticos sean ágiles y capaces de adaptarse rápidamente a la dinámica del entorno industrial.

La necesidad de **infraestructuras potentes** de análisis es un desafío en el manejo de grandes volúmenes de datos, especialmente en entornos como la Industria 4.0. Debido a la masiva cantidad de información generada por sensores, máquinas, robots y usuarios, es imperativo contar con entornos de computación que sean capaces de procesar estos datos de manera eficiente y en tiempos adecuados. Los métodos tradicionales de procesamiento de datos no son suficientes para manejar esta magnitud de información, lo que requiere soluciones de alto rendimiento.

Afortunadamente, la aparición de tecnologías como los clústeres de computación y la computación en la nube ha revolucionado la capacidad de análisis de datos. Estos clústeres permiten agrupar múltiples procesadores para trabajar de manera simultánea en el procesamiento de datos, reduciendo significativamente el tiempo necesario para realizar cálculos complejos. Además, la computación en la nube ofrece una opción accesible y escalable, permitiendo alquilar capacidad de procesamiento sin necesidad de invertir en infraestructura propia. De esta manera, las organizaciones pueden obtener el poder de cómputo necesario para realizar análisis avanzados y tomar decisiones informadas basadas en grandes volúmenes de datos, de forma más eficiente y rentable que nunca.

Las **políticas de privacidad** son fundamentales cuando se manejan los datos derivados de la "huella digital" de los usuarios en Internet. Estos datos, que incluyen información sobre comportamientos en línea, interacciones con páginas web, y otras acciones digitales, son valiosos para análisis estadísticos y de comportamiento, pero su uso está estrictamente regulado. Dado que se trata de información personal o sensible, es necesario obtener el consentimiento explícito de los usuarios para recolectarla y utilizarla en estudios o análisis. Además, las empresas que recopilan estos datos deben garantizar su manejo adecuado de acuerdo con las normativas de privacidad vigentes, como el Reglamento General de Protección de Datos (GDPR) en Europa, que regula el acceso, almacenamiento y uso de esta información.

En la estadística tradicional, el estudio se diseña primero y, posteriormente, se recolectan los datos necesarios, generalmente a través de encuestas o métodos de recolección específicos. El proceso sigue un enfoque estructurado donde el problema y el modelo de datos se definen antes de la obtención de la información. En contraste, en los entornos de big data dentro de la industria 4.0, los datos ya están disponibles y se debe ajustar el análisis al tipo de información existente. Esto

ISBN: 978-631-6557-56-8

puede requerir la realización de análisis adaptados a esos datos o la aplicación de procesos de transformación para extraer información relevante.

1.9.2. Casos de uso

Como se puede deducir, la implementación de técnicas estadísticas en los entornos de la industria 4.0 juega un papel importante para el éxito de estos sistemas. La estadística proporciona una base sólida para desarrollar herramientas que permitan describir, diagnosticar y descubrir patrones y tendencias, basándose en datos históricos y en tiempo real. Estas herramientas son fundamentales en la creación de dispositivos médicos o de fabricación, mejorando procesos al anticipar situaciones anómalas, optimizando la calidad de la producción y reduciendo costos. Además, facilitan la realización de predicciones precisas sobre eventos futuros. Por ejemplo, en los sistemas autónomos de almacén, estas predicciones ayudan a optimizar rutas o determinar cuándo es necesario reponer materiales. También se aplican en sistemas de carretillas autónomas, como los de *Linde Intralogistic Solutions*, que abordan los desafíos de la industria 4.0. Asimismo, los análisis estadísticos son importantes en sistemas de localización en tiempo real, ya que permiten procesar eficientemente el gran volumen de señales que estos sistemas deben gestionar, asegurando que los algoritmos funcionen correctamente.

En el sector agrícola, la aplicación de la estadística dentro de los sistemas de la industria 4.0 está cobrando una gran relevancia. Se están implementando sistemas de recolección de datos mediante sensores que, al combinarse con información meteorológica, permiten prever situaciones de riesgo en los cultivos. Un ejemplo de esto es la plataforma de Horta, que se especializa en la detección de hongos y otros riesgos para diversos tipos de cultivos. Además, las técnicas estadísticas son fundamentales en plataformas de ahorro energético, donde se utilizan para clasificar a nuevos usuarios y estimar su consumo en función, por ejemplo, del tipo de vivienda. Gracias a la estadística, también es posible realizar predicciones de consumo con alta confiabilidad. Un ejemplo de este tipo de plataformas es *SimbleSense*, que se enfoca en la optimización del uso de energía.

De manera más específica, a continuación, se presentan tres ejemplos en los que la estadística juega un papel fundamental en el análisis de componentes dentro de la industria 4.0:

La reducción del volumen de datos es una de las aplicaciones principales del análisis de componentes principales. En entornos de big data dentro de la industria 4.0, donde se maneja una cantidad masiva de información, estas herramientas son básicas para reducir los datos de manera que sean manejables. Un ejemplo de esto son aplicaciones como *Doeet*, que emplean sistemas de *business intelligence.*

ISBN: 978-631-6557-56-8

Otra aplicación relevante es la reducción del ruido en imágenes. Cuando las imágenes obtenidas están afectadas por factores como condiciones atmosféricas o interferencias, el análisis de componentes principales puede ayudar a eliminar el ruido al identificar y eliminar las últimas componentes del análisis. Esto mejora la claridad de las imágenes y es útil en sistemas de reconocimiento automático, como los clasificadores de reconocimiento facial utilizados en aeropuertos, especialmente en España, desarrollados por *Indra*.

Finalmente, el análisis de componentes principales también se usa para la detección de cambios en los datos a lo largo del tiempo. Al comparar datos de una misma población en diferentes momentos, se pueden identificar las variables que han permanecido constantes y aquellas que han experimentado cambios significativos. Este tipo de análisis es utilizado en sistemas de demanda-respuesta en la red eléctrica, aplicados por las empresas de servicios públicos, para optimizar la producción y el consumo de energía. En los siguientes códigos QR o enlaces se muestra información adicional relacionada al capítulo.

Introducción al Big Data	Conceptos básicos en análisis de datos
https://www.youtube.com/watch?v=zez2Tv-bcXY	https://www.youtube.com/watch?v=Hq0Ldt5S8Og

ISBN: 978-631-6557-56-8

CAPÍTULO II

CAPTURA Y ARQUITECTURAS DE DATOS PARA ENTORNOS INDUSTRIALES

2.1. Introducción y Objetivo del Capítulo

Antes de que la informática y los métodos modernos de adquisición de datos se integraran en la vida cotidiana, la recopilación de información solía ser un proceso complejo y costoso, que dependía en gran medida de la intervención humana para convertir documentos físicos en formato digital. Sin embargo, el constante desarrollo tecnológico ha transformado este panorama, facilitando la aparición de métodos de captura de datos cada vez más variados, automatizados y precisos. A medida que el costo de almacenamiento por unidad ha disminuido, los sistemas actuales son capaces de almacenar volúmenes masivos de datos provenientes de múltiples y diversas fuentes.

Esta diversidad de fuentes de datos exige establecer una base robusta que garantice una captura eficiente y flexible, adecuada para los distintos usos que se pretende dar a la información. Dicha base se estructura en torno a tres elementos clave: la clara definición y diferenciación entre dato, información y conocimiento; la evaluación de la calidad de los datos mediante dimensiones y criterios específicos; y la adquisición de experiencia práctica a través del análisis de casos reales donde se ha requerido capturar y almacenar datos de manera efectiva.

En la actualidad, la cantidad de datos generados a nivel mundial es inmensa y proviene de una variedad de fuentes, como redes sociales con opiniones de usuarios, mercados bursátiles en tiempo real o sistemas industriales basados en sensores IoT (Internet Industrial de las Cosas, IIoT). Estos datos necesitan almacenarse en grandes repositorios que permitan su posterior análisis y tratamiento mediante algoritmos de inteligencia artificial, como *machine learning* o *deep learning*, aplicados en escenarios como el mantenimiento predictivo propio de la Industria 4.0.

Para abordar estas necesidades, han surgido tecnologías especializadas como las bases de datos NoSQL por ejemplo, MongoDB y Cassandra y sistemas de procesamiento masivo de datos como *Hadoop* y *Apache Spark*. Estas herramientas permiten gestionar grandes volúmenes de datos distribuidos en la nube, ampliando su capacidad mediante la incorporación dinámica de nodos, lo que posibilita la implementación de soluciones big data escalables. Esto resulta fundamental para manejar información proveniente de fuentes como redes sociales, dispositivos IIoT, o datos enlazados abiertos y masivos, entre muchas otras [27].

ISBN: 978-631-6557-56-8

No obstante, existen escenarios en los que es fundamental que los datos sean procesados, analizados y sus predicciones visualizadas de manera inmediata, sin la posibilidad de esperar a que se apliquen técnicas de minería de datos por lotes o métodos como Map-Reduce, que suelen emplearse en entornos de big data. Esto es especialmente relevante en aplicaciones críticas, como algunas en el ámbito médico [28], o en sistemas que trabajan con grandes volúmenes de datos para identificar patrones de consumo y generar recomendaciones personalizadas en tiempo real. En este contexto, surge una transición del enfoque big data hacia el smart data, donde los datos son analizados y procesados instantáneamente en el mismo momento en que son capturados [29].

Así pues, el presente capítulo trata esta cuestión siguiendo los **objetivos** a continuación:

- Comprender la organización habitual de las arquitecturas big data, así como las distintas capas que las componen.
- Examinar las diversas fuentes de datos que pueden alimentar una arquitectura big data, especialmente en contextos industriales.
- Entender la función que desempeñan las capas de mensajería y almacenamiento, además de reconocer ejemplos concretos de cómo se implementan estas capas en arquitecturas reales.
- Familiarizarse con los principales elementos que conforman la capa de análisis dentro de una arquitectura big data.
- Reconocer las distintas capas que consumen datos en una arquitectura big data, incluyendo aquellas encargadas de presentar la información a los usuarios finales en entornos de industria 4.0.
- Comprender de qué manera las arquitecturas en la nube ofrecen soluciones eficientes para la implementación de técnicas de procesamiento big data, y cómo el edge computing actúa como complemento de dichas soluciones.
- Identificar las plataformas comerciales cloud más relevantes de uso general, y explorar cómo aprovechar sus capacidades en big data, machine learning, IoT y *edge computing*.
- Analizar casos prácticos donde se han implementado arquitecturas big data en aplicaciones reales dentro de la industria 4.0.

2.2 Origen y calidad de los datos

El valor de los datos en cualquier proceso de análisis o toma de decisiones depende directamente de su origen y de la calidad con la que han sido recolectados, almacenados y procesados. Comprender de dónde provienen los datos y garantizar que cumplan con criterios de integridad, precisión y fiabilidad es importante para obtener resultados significativos y confiables.

ISBN: 978-631-6557-56-8

2.1.1. Datos, Información y conocimiento

En contextos informales es común que los términos dato, información y conocimiento se utilicen como sinónimos. No obstante, en entornos profesionales y académicos es importante diferenciar estos conceptos con precisión, ya que una correcta distinción permite evitar confusiones durante las distintas etapas del análisis de datos. Un **dato** se entiende como un hecho puntual y específico relacionado con un evento. La **información** se define como el resultado de organizar y combinar varios datos, aportando así un significado. Finalmente, el **conocimiento** surge de la integración de experiencias, información contextual y criterios de relevancia, lo que permite interpretar y aplicar esa información de manera efectiva, a continuación en la Figura 19 se muestra la jerarquía que existe los diferentes niveles.

Figura 19: Pirámide de la jerarquía del conocimiento

Fuente: Autores

2.1.2. Evaluación de calidad

Las métricas o dimensiones que se utilizan para evaluar la calidad de un conjunto de datos pueden clasificarse según los distintos actores que interactúan con ellos.

Por un lado, los **diseñadores y administradores de bases de datos** emplean métricas orientadas al diseño o al esquema de los datos, más que a los datos en sí. Entre estas se destacan la **completitud**, que se refiere a la cobertura total de los tipos de datos requeridos, y el **minimalismo**, que implica la reducción de redundancias en la estructura del almacén de datos.

Por otro lado, los **desarrolladores de software** consideran métricas vinculadas al desarrollo de productos informáticos. Aunque estas métricas no siempre están directamente relacionadas con los datos, influyen de manera significativa en la forma en que se almacenan, acceden y manipulan.

ISBN: 978-631-6557-56-8

Finalmente, el **usuario final** quien interpreta y utiliza los datos para formular conclusiones y tomar decisiones valora métricas como la **disponibilidad de los datos** y el grado de **interpretabilidad**, es decir, la facilidad con la que la información puede ser comprendida una vez que ha sido presentada.

A partir de estas perspectivas [30], presentan un conjunto de **dimensiones** para evaluar las propiedades de un conjunto de datos:

Completitud o cobertura.- Esta métrica evalúa qué porcentaje de datos disponibles representa fielmente a la totalidad de la población que se desea describir. Por ejemplo, si se dispone de información sobre 90 de los 100 tratamientos médicos efectuados en un hospital, la cobertura alcanzaría un 90 %. Esta medida también puede aplicarse a características específicas dentro del conjunto de datos; por ejemplo, si el 5 % de los registros no incluye la fecha de finalización de los tratamientos, esa omisión afectaría negativamente la cobertura en relación a esa variable particular.

Credibilidad.- La credibilidad hace referencia al grado de confianza que se otorga a la fuente emisora del conjunto de datos. Esta confianza puede reflejarse en la lógica o coherencia de los valores contenidos. Por ejemplo, en un conjunto de datos sobre postres, se considera confiable si incluye elementos como tarta, helado o fruta, mientras que la aparición de valores poco congruentes —como lubina al horno— indicaría una baja credibilidad.

Precisión.- La precisión mide la proporción de datos correctos respecto al total registrado, expresándose generalmente en forma de porcentaje. Cuanto mayor sea esta proporción, mayor será la confianza en que los datos reflejan con exactitud la realidad que representan.

Consistencia.- Esta métrica determina el grado de coherencia interna de los datos. Por ejemplo, en un conjunto de datos geográficos, si a una misma entidad se le asocia simultáneamente una ciudad y un país que no guardan relación, se evidencia un problema de consistencia que puede afectar la fiabilidad global de la información.

Interpretabilidad.- La interpretabilidad mide qué tan comprensible es la información para un usuario final. Factores como la claridad en la documentación, la nomenclatura de los campos y el formato utilizado para presentar los datos influyen directamente en esta métrica, facilitando o dificultando la correcta comprensión de los datos.

ISBN: 978-631-6557-56-8

2.1.3. Fuentes de información

Los métodos empleados para la recolección de información pueden organizarse según las propiedades del elemento que origina los datos. En la actualidad, existen cinco categorías principales que se utilizan con mayor frecuencia para clasificar estos métodos:

Captura manual de datos.- Esta es la categoría más clásica y aún una de las más comunes, especialmente en investigaciones dentro de las ciencias sociales y naturales. Incluye métodos como encuestas y mediciones basadas en la observación directa. Aunque en esta categoría no existe una dependencia inmediata de las tecnologías de la información, los datos obtenidos requieren ser digitalizados en algún punto, ya sea durante su recolección o en una etapa posterior de procesamiento.

Procesamiento de documentos estructurados.- Este método se basa en extraer información directamente de documentos cuyo propósito original no era servir como fuente de datos. Un ejemplo representativo es el web *scraping*, técnica que permite capturar datos desde páginas web en formato HTML. Otro caso habitual es el análisis de logs, archivos que registran cronológicamente eventos en sistemas y que, aunque están diseñados para funciones de auditoría o seguimiento, pueden ser aprovechados como fuente de datos.

Salida de aplicaciones.- Se refiere a uno de los métodos más sencillos de adquisición de datos, que implica extraer información de sistemas de almacenamiento convencionales como bases de datos relacionales o archivos en formatos estructurados, por ejemplo, archivos CSV *(Comma-Separated Values).*

Captura de datos mediante sensores.- Este grupo comprende datos generados por sensores físicos de todo tipo. Entre los ejemplos se encuentran sensores meteorológicos (como pluviómetros y anemómetros), ambientales (ruido, luz), biomédicos (medición del ritmo cardíaco, conductividad dérmica) y sensores integrados en dispositivos móviles (acelerómetros, giroscopios, magnetómetros, entre otros). En la actualidad, destaca el creciente uso de sensores en dispositivos personales como pulseras inteligentes o relojes inteligentes, enmarcado dentro del movimiento conocido como *Quantified Self*, que promueve el autoanálisis de parámetros físicos mediante tecnología wearable.

Acceso a datos públicos.- Consiste en la obtención de datos abiertos, ya sea mediante su descarga directa en formatos estructurados como CSV o mediante la consulta de interfaces de programación de aplicaciones (API). Hoy en día, es habitual que organismos públicos, gobiernos locales y centrales, así como empresas privadas, pongan a disposición catálogos de datos que pueden ser reutilizados en procesos de análisis y desarrollo de nuevas aplicaciones.

ISBN: 978-631-6557-56-8

2.3. Organización de los datos

Los datos se presentan y se estructuran de diversas maneras, lo que implica que cada tipo de dato requiere soluciones de almacenamiento específicas y, en consecuencia, debe ser tratado de manera distinta. Como se abordará más adelante, la forma en que se organizan los datos puede variar dependiendo de la fase del proceso de gestión de la información en la que nos encontremos.

Datos no estructurados.- Son los datos más crudos y representan alrededor del 80 % de todos los datos. Pueden adoptar diversas formas, como textos (archivos de texto plano, PDFs), imágenes (ficheros JPEG, PNG), sonidos (archivos WAV, OGG o MP3) o vídeos (archivos MOV, MP4). Estos datos generalmente se almacenan en repositorios de archivos, organizados de manera similar a un directorio en el sistema de ficheros de un ordenador. Extraer valor de los datos no estructurados es un desafío, ya que no siguen una estructura común definida. Para aprovecharlos, es necesario extraer características estructuradas que los describan o abstraigan. Por ejemplo, en el caso de un texto, se podría necesitar identificar los temas tratados o realizar un análisis de sentimiento mediante técnicas de procesamiento del lenguaje natural (NLP). Para una imagen, podrían emplearse técnicas de *deep learning*, como redes neuronales convolucionales, para identificar características como paisajes o personas en la imagen.

Datos estructurados.- Son datos que están claramente definidos y se organizan de forma tabular, en filas y columnas. Se conoce qué columnas existen y qué tipo de datos contienen. Un ejemplo típico son los archivos CSV. Estos datos se almacenan comúnmente en bases de datos y se pueden consultar utilizando SQL, lo que facilita la creación de conjuntos de datos para diversas aplicaciones de ciencia de datos.

Datos semiestructurados.- Se encuentran entre los datos no estructurados y los estructurados. Aunque tienen una estructura definida, esta no es tan rígida como en los datos estructurados. El esquema de los datos es más flexible, permitiendo que los datos no sean necesariamente tabulares o que algunas partes de los mismos puedan estar incompletas o tomar diferentes tipos. Los datos semiestructurados suelen almacenarse en formatos como JSON o XML. Además, pueden ser gestionados por bases de datos orientadas a documentos (como MongoDB, que utiliza BSON, un formato binario de JSON), las cuales permiten realizar consultas a través de una API adecuada.

2.3.1. Ficheros Planos

Ficheros planos

Los ficheros planos son una forma sencilla y común de almacenar datos que se utilizan frecuentemente para el intercambio de información entre sistemas. Una de sus ventajas es que se puede ver y editar su contenido con una herramienta de edición de texto, lo que facilita su

ISBN: 978-631-6557-56-8

manipulación. Sin embargo, estos ficheros tienden a ser más verbosos que los ficheros binarios, lo que implica que su tamaño será mayor y las operaciones de procesamiento serán más costosas en términos de tiempo y recursos. Los formatos más comunes de ficheros planos son CSV, JSON y XML, cada uno con sus características particulares.

CSV *(Comma Separated Values).-* El formato CSV, que significa "Valores Separados por Coma", está documentado en la RFC 4180 y se utiliza para almacenar datos en forma tabular. Sus características incluyen:

- Cada registro está delimitado por un cambio de línea (combinación de los caracteres CR y LF, es decir, retorno de carro y salto de línea).
- Los valores dentro de un registro se separan por comas.
- El número de valores debe ser constante en todos los registros.
- Los valores pueden ir entre comillas dobles, especialmente si contienen comas, saltos de línea o comillas dobles.

JSON *(JavaScript Object Notation).-* JSON es un formato basado en el lenguaje de programación JavaScript y se utiliza principalmente para almacenar y transportar datos. Se organiza en dos estructuras principales:

- Un **objeto** o registro, que es un conjunto de pares nombre/valor.
- Un **array** o lista, que es una secuencia ordenada de valores. JSON es ampliamente utilizado debido a su simplicidad y facilidad de uso, especialmente en aplicaciones web.

XML *(eXtensible Markup Language).-* El formato XML es el más verboso de los tres y utiliza marcas o etiquetas para definir la estructura de los datos. Algunas de sus características son:

- El documento comienza con la línea <?xml version="1.0"?>.
- Un documento XML debe tener un solo **elemento raíz**.
- Los elementos se definen mediante etiquetas que se abren con <etiqueta> y se cierran con </etiqueta>.
- Los elementos pueden tener **atributos**, que se escriben como pares nombre/valor dentro de la etiqueta, separados por un signo de igual (=), y el valor debe estar entre comillas.
- El contenido de un elemento puede ser texto, otros elementos o una combinación de ambos.

ISBN: 978-631-6557-56-8

2.3.2. Bases de datos

Una base de datos es un conjunto organizado y persistente de datos, utilizados por sistemas de software para almacenar, gestionar y recuperar información. Los sistemas de base de datos están compuestos por cuatro elementos básicos: **datos, hardware, software y usuarios**. Los datos pueden ser integrados, cuando se almacenan de forma unificada y son accedidos normalmente por un solo usuario, o compartidos, cuando se gestionan con permisos diferenciados para varios usuarios.

El **hardware** de un sistema de base de datos incluye los dispositivos de almacenamiento (discos duros, unidades SSD), los procesadores que permiten ejecutar las operaciones sobre los datos, y la memoria principal, que facilita el acceso eficiente a la información. A nivel de software, el componente clave es el Sistema Gestor de Bases de Datos (DBMS), que actúa como intermediario entre los datos físicos y las aplicaciones o usuarios, gestionando la integridad, seguridad y accesibilidad de la información.

Existen tres tipos principales de usuarios en un sistema de base de datos: los **programadores**, que desarrollan aplicaciones para interactuar con la base; los **usuarios finales**, que utilizan dichas aplicaciones para consultar o modificar datos; y el **administrador** de base de datos (DBA), responsable de mantener la estructura, disponibilidad y rendimiento del sistema.

En el contexto de bases de datos, una entidad representa cualquier objeto o concepto que puede almacenarse, como por ejemplo un producto o una bodega. Las relaciones o vínculos permiten asociar entidades entre sí, facilitando la representación de dependencias o conexiones, como el caso de una relación entre productos y bodegas, que define en qué lugar se almacena cada producto.

Finalmente, los datos se organizan de manera jerárquica en tres niveles. El **campo** es la unidad más pequeña y define un tipo de dato, como texto, número o fecha. Un conjunto de campos relacionados forma un **registro**, que describe a una entidad específica. A su vez, un grupo de registros del mismo tipo constituye un archivo o **tabla**, que permite almacenar grandes volúmenes de datos de forma estructurada y eficiente.

2.3.3. Bases de datos relacionales y SQL

En un sistema de base de datos relacional, los archivos de datos se representan mediante tablas. En estas tablas, cada columna define un campo específico (como nombre, precio o fecha), mientras que cada fila corresponde a un registro individual que contiene valores para esos campos. Además, cuando un usuario ejecuta una operación sobre una tabla —por ejemplo, una consulta o

ISBN: 978-631-6557-56-8

filtro—, el resultado que se obtiene también se presenta en forma de tabla, manteniendo así la coherencia en la estructura de los datos durante todo el proceso.

2.3.4. Bases de datos NoSQL

El crecimiento acelerado de datos en diversas industrias ha impulsado el concepto de Big Data, lo que ha generado la necesidad de sistemas de bases de datos capaces de manejar grandes volúmenes de información. Inicialmente, esta demanda se cubría con bases de datos relacionales distribuidas y optimizadas, pero en muchos casos estas soluciones resultaron insuficientes, dando paso al surgimiento de nuevas alternativas: las bases de datos NoSQL.

Las bases de datos NoSQL (Not Only SQL) surgieron alrededor del 2009, destacándose por no depender de tablas tradicionales ni utilizar necesariamente el lenguaje SQL. Suelen usar modelos como documentos, grafos o claves-valor, y priorizan la flexibilidad y escalabilidad frente a la rigidez de las bases relacionales.

Una de sus principales ventajas es la facilidad para escalar horizontalmente, es decir, agregar nuevos servidores al sistema para repartir la carga y mejorar el rendimiento. Esto contrasta con el escalado vertical, propio de sistemas tradicionales, donde solo es posible mejorar el rendimiento añadiendo más recursos a un mismo servidor, lo cual es limitado y en ocasiones requiere

Las bases de datos NoSQL se dividen en varias categorías, cada una diseñada para necesidades específicas:

a. **Bases de datos clave-valor simples.-** Almacenan datos como pares clave-valor, sin estructura fija. Son ideales para sistemas de caché. Ejemplos: ***Memcached*** (muy usado en aplicaciones web) y **Redis** (rápido y en memoria, también admite almacenamiento persistente).

b. **Bases de datos clave-valor avanzadas.-** Mejoran la categoría anterior, permitiendo operaciones más complejas y estructuras de datos más flexibles. Ejemplos: Apache Cassandra, Dynamo, Voldemort y Riak.
Algunas, como Cassandra o HBase, también son conocidas como bases de datos columnar, ya que su modelo combina tablas y claves-valor.

c. **Bases de datos orientadas a documentos.-** Permiten almacenar documentos estructurados como JSON o BSON, lo que facilita la gestión de datos complejos y flexibles. Ejemplos: MongoDB, CouchDB, Couchbase, DocumentDB y MarkLogic.

ISBN: 978-631-6557-56-8

d. **Bases de datos orientadas a grafos.-** Usan nodos y aristas para representar datos y sus relaciones, siendo ideales para aplicaciones donde las conexiones entre entidades son relevantes. Ejemplos: Neo4J, AllegroGraph, InfiniteGraph y Stardog.

e. **Bases de datos multimodelo.-** Soportan varios modelos de datos (documentos, grafos, clave-valor, etc.) en un solo sistema. Ejemplos: OrientDB, ArangoDB, CosmosDB y FoundationDB.

f. **Bases de datos de series temporales (TSDB).-** Optimizadas para almacenar datos que se generan a lo largo del tiempo, como registros de sensores o métricas de rendimiento, muy útiles en entornos IoT e Industria 4.0. Ejemplos: InfluxDB, TimescaleDB, Prometheus y Apache Druid.

2.3.5. NoSQL vs SQL

Las bases de datos tradicionales, conocidas como **relacionales**, se destacan por usar un lenguaje estándar (SQL) y por ofrecer una gestión eficiente de la información gracias a su base matemática sólida y principios como **ACID** (Atomicidad, Consistencia, Aislamiento y Durabilidad). Estas bases de datos han sido ampliamente adoptadas en las organizaciones por su estructura definida, metodologías de diseño y variedad de herramientas.

Por otro lado, las **bases de datos NoSQL** surgen como una alternativa para manejar grandes volúmenes de datos y alta demanda de consultas. A diferencia de las relacionales, NoSQL no sigue un esquema fijo, no siempre utiliza SQL ni garantiza plenamente las propiedades ACID. Además, se diseñan para escalar horizontalmente y aprovechan mucho la memoria principal para mejorar el rendimiento.

Una de sus mayores ventajas es la **flexibilidad**: permiten almacenar registros con estructuras variables, algo ideal en contextos como tiendas en línea con productos muy diversos. Esto evita la rigidez de las tablas tradicionales, donde muchas columnas quedarían vacías.

Sin embargo, esta flexibilidad también representa un desafío, ya que puede facilitar errores en el desarrollo y manejo de datos si no se controlan bien las estructuras. A pesar de ello, NoSQL permite, si se desea, imponer modelos fijos para garantizar más consistencia sin perder rendimiento.

Ventajas de las bases de datos NoSQL:

- Ofrecen **escalabilidad horizontal** de forma sencilla, adaptándose a las crecientes necesidades de las empresas.

ISBN: 978-631-6557-56-8

- Evitan los **cuellos de botella** en el acceso y procesamiento de datos.
- Permiten manejar **grandes volúmenes de información** de forma eficiente.
- Es posible utilizar **diferentes bases de datos NoSQL para distintos proyectos**, según las necesidades específicas.
- Su **implementación es económica**, ya que no requieren equipos costosos.

Desventajas de las bases de datos NoSQL:

- **Soporte y reputación limitados**: Aunque esta situación ha mejorado con el tiempo gracias al respaldo de grandes empresas como Amazon, Microsoft y Google, en sus inicios las bases de datos NoSQL no contaban con el mismo prestigio ni soporte técnico que las soluciones tradicionales de Oracle, IBM o Microsoft.
- **Madurez tecnológica**: Durante sus primeros años, muchas soluciones NoSQL no estaban suficientemente probadas para entornos empresariales exigentes, aunque esto ha mejorado notablemente.
- **Dificultades con herramientas de inteligencia de negocios (BI)**: NoSQL no siempre se integra bien con herramientas de análisis como las bases relacionales, aunque hoy en día plataformas como Power BI, Google Data Studio y Toad ya ofrecen compatibilidad.
- **Falta de personal capacitado**: Dado que NoSQL es una tecnología más reciente (popularizada desde 2009), inicialmente existía poca oferta de profesionales con experiencia, aunque cada vez hay más especialistas.
- **Compatibilidad limitada**: A diferencia de las bases de datos relacionales, NoSQL carece de estándares comunes. Cada sistema tiene su propia API y formas de consulta, lo que dificulta la migración entre plataformas.

2.3.6. Interfaces de programación de aplicaciones (API)

Las interfaces de programación de aplicaciones, conocidas como API (*application programming interfaces*), permiten que un software o aplicación utilice las funciones que ofrece otro sistema, como una librería o biblioteca, sin necesidad de crearlas desde cero. Gracias a esto, es posible reutilizar código ya existente, evitar la duplicación de funcionalidades y optimizar tanto el tiempo como los costos de desarrollo.

De este modo, una API se define como un conjunto de subrutinas, funciones y procedimientos —denominados métodos en programación orientada a objetos— que una librería o biblioteca pone a disposición para que otros programas puedan utilizarlos.

ISBN: 978-631-6557-56-8

Existen diversas maneras de permitir la comunicación entre programas que se ejecutan en diferentes dispositivos mediante APIs. Una de las más comunes es a través de arquitecturas orientadas a servicios (SOA, Service-Oriented Architectures), dentro del modelo cliente-servidor, donde una aplicación cliente solicita y utiliza servicios que son ofrecidos por una aplicación servidor.

Dentro de este enfoque, se encuentran las API basadas en SOAP (*Simple Object Access Protocol*), donde la información se intercambia en formato XML y los servicios son descritos usando WSDL (*Web Services Description Language*).

Otra alternativa son las REST API (*Representational State Transfer*), las cuales funcionan sin mantener información del estado de las solicitudes, es decir, cada petición debe incluir todos los datos necesarios para que el servidor procese y responda. Además, las REST API se centran en los recursos o datos, en lugar de enfocarse en acciones o procedimientos específicos, a diferencia de las API basadas en SOAP que siguen un enfoque tipo RPC (*Remote Procedure Call*).

2.4. Proceso ETL

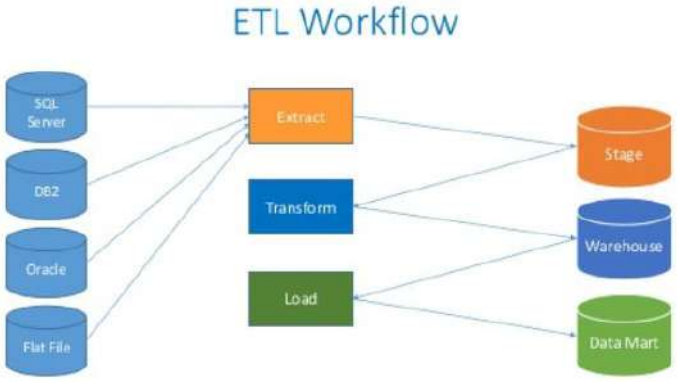
El proceso ETL (*Extract, Transform and Load* — extracción, transformación y carga), representado de manera simplificada en la Figura 20, sigue una secuencia habitual que comprende:

- La extracción de datos provenientes de diversas fuentes, que suelen ser heterogéneas, como bases de datos relacionales, NoSQL, archivos o APIs REST.
- La transformación de esos datos, lo que implica realizar cálculos, limpiar y depurar la información, completar valores faltantes, eliminar duplicados o combinar datos provenientes de diferentes orígenes.
- Finalmente, la carga de los datos transformados en un sistema de almacenamiento definitivo, conocido como data *warehouse*.

Entre las herramientas que implementan procesos ETL se pueden mencionar *Amazon Redshift*, Marklogic y diversas soluciones ofrecidas por Oracle.

ISBN: 978-631-6557-56-8

Figura 20: Esquema del proceso ETL básico.



Fuente: [31]

2.4.1. *Data lake* y *data warehouse*

***Data Lake* (Lago de Datos).-** Un data *lake* es un sistema de almacenamiento que permite guardar grandes volúmenes de datos en su formato original, sin importar si son estructurados, semiestructurados o no estructurados. No suele tener límites prácticos de capacidad y permite analizar los datos en cuanto están disponibles. Utiliza un enfoque *schema on read*, lo que significa que la estructura de los datos se define en el momento en que se consultan, lo que brinda gran flexibilidad, especialmente en entornos de big data.

Data Swamp (Pantano de Datos).- Cuando los datos en un data *lake* se almacenan de manera desordenada, sin metadatos, sin control de calidad ni una adecuada gobernanza, el sistema puede volverse un "pantano de datos". Esto dificulta su uso y puede hacer que la información almacenada pierda valor y utilidad.

***Data Warehouse* (Almacén de Datos).-** Un data *warehouse* es un repositorio donde se almacenan datos ya organizados y procesados, con un cambio más lento a lo largo del tiempo. Emplea el modelo **schema on write**, es decir, la estructura de datos debe definirse y validarse antes de que los datos sean cargados, lo que es característico de las bases de datos relacionales tradicionales.

***Data Mart* (Mercado de Datos).-** Un *data mart* es una versión reducida y enfocada de un data warehouse, destinada a almacenar información específica para un área o departamento de la organización. Puede construirse desde cero o derivarse de una porción de un data *warehouse* ya existente, permitiendo acceso rápido a datos específicos mientras se construyen o complementan almacenes de datos más grandes.

Las diferencias entre los *data lakes* y *data warehouses* se resumen en la Tabla 3.

ISBN: 978-631-6557-56-8

Tabla 2: Comparación entre data *lakes* y data *warehouses*.

COMPARACIÓN	DATA LAKE	DATA WAREHOUSE
Datos	Datos estructurados	Datos estructurados Datos procesados
	Datos semiestructurados	
	Datos no estructurados	
	Datos en crudo	
	Datos sin procesar	
Procesamiento	*Schema on read*	*Schema on write*
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

Adaptado de: [32]

2.4.2. Extracción

La extracción es la primera etapa del proceso ETL, en la que se recopilan datos desde una o varias fuentes, como bases de datos o aplicaciones. El volumen de datos extraído puede variar ampliamente, desde algunos kilobytes hasta varios gigabytes, dependiendo de la fuente y las necesidades del negocio.

El propósito de esta fase es obtener todos los datos requeridos sin sobrecargar ni afectar el rendimiento del sistema de origen. Para ello, la extracción debe planificarse cuidadosamente y utilizar el menor número de recursos posible.

Tipos de Métodos de Extracción

a) **Extracción Lógica**

- **Notificación de actualización:** El sistema fuente informa automáticamente cuándo y qué datos han cambiado, facilitando su captura.
- **Extracción incremental:** Si el sistema no notifica cambios, puede identificar registros modificados y extraer solo esos datos.
- **Extracción completa:** Cuando no es posible detectar qué datos cambiaron, se extraen todos los datos, comparándolos con una copia previa para identificar diferencias, incluyendo eliminaciones.

b) **Extracción Física**

- **En línea (Online):** Los datos se obtienen directamente del sistema de origen, accediendo a sus tablas o estructuras intermedias como registros o *snapshots*.

ISBN: 978-631-6557-56-8

- **Fuera de línea (Offline):** Los datos son preparados fuera del sistema original, ya sea desde registros específicos o a través de rutinas de extracción.

En especial, cuando se usan extracciones completas o incrementales, es importante definir bien la frecuencia de extracción, ya que los volúmenes pueden ser bastante grandes.

2.4.3. Transformación

En esta etapa, los datos extraídos se convierten en un formato adecuado para ser almacenados en un data *warehouse*. Durante el proceso, se aplican diversas operaciones como cálculos, manipulación de datos (por ejemplo, mediante SQL), combinaciones, restricciones, y la asignación de claves primarias y foráneas.

Por ejemplo, si se requiere obtener el promedio de un valor, como la anualidad total, el cálculo se realiza en esta fase antes de cargar los datos. En algunos casos, los datos no necesitan ajustes y pueden transferirse directamente al almacén; estos se conocen como datos de paso o traslado directo.

Además, la transformación también abarca la depuración, que incluye corregir errores, eliminar duplicados, completar datos faltantes y asegurar que el formato sea compatible e íntegro antes de ser almacenado.

2.4.4. Carga

En la fase final del proceso ETL, los datos transformados se almacenan en la estructura de datos multidimensional que utilizan los usuarios y las aplicaciones. Este proceso abarca tanto las tablas de dimensiones como las de hechos, y es importante realizarlo de manera eficiente y con el menor uso de recursos posible.

Para optimizar la carga, es recomendable desactivar restricciones e índices antes de la carga y volver a activarlos al finalizar. También se debe asegurar la integridad referencial durante todo el proceso.

Existen tres tipos de carga:

- **Carga inicial**: Consiste en llenar todas las tablas del data *warehouse*.
- **Carga incremental**: Realiza actualizaciones periódicas según los cambios necesarios.
- **Refresco completo**: Borra y recarga los datos en una o varias tablas.

ISBN: 978-631-6557-56-8

2.5. Casos de estudio

i. **Registro de Descargas de Documentos.-** En un sitio web donde los usuarios descargan documentos, sin una aplicación intermediaria para registrar estas acciones, se busca generar un informe sobre qué documentos son más descargados y quiénes son los usuarios más activos. La solución propuesta es procesar los logs del servidor web para capturar esta información.

ii. ***Web Scraping* de Cursos *Online*.-** Para obtener información sobre cursos masivos en línea (MOOC) de la página Class Central, donde no existe una API, se utiliza la herramienta de web scraping "*Scraper*" para capturar los datos estructurados de las tablas del sitio web y exportarlos a Google Docs.

iii. **Acceso a Transacciones Bancarias mediante API.-** Se necesita visualizar patrones de compra de personas utilizando tarjetas bancarias en una ciudad, con datos agregados por código postal, tipo de establecimiento y otros. Los datos se obtienen a través de un API público, y se han desarrollado scripts para capturarlos tras crear una cuenta en el portal de acceso.

iv. **Productos en Formato CSV.-** Se plantea el problema de definir el formato de un archivo CSV para almacenar un inventario de productos, con campos como identificador, nombre y cantidad, asegurándose de incluir los nombres de los campos para evitar confusión.

v. **Información Geolocalizada en Formato JSON.-** Se captura la actividad geolocalizada de una persona, que incluye su ubicación en términos de latitud y longitud, el identificador de usuario y el momento de la captura, en formato de fecha y hora. Esta información se representa en formato JSON.

vi. **Información de Clientes en una Base de Datos Relacional**.- Se almacena información sobre clientes en una base de datos relacional, que incluye el identificador de cliente, nombre, apellidos y fecha de suscripción. Los datos se organizan en formato de tabla.

2.6. Capas de una arquitectura Big Data

Las capas lógicas son una forma de organizar los componentes en una arquitectura, ya sea de software, hardware o red, dividiendo las funciones en secciones específicas. Estas capas no requieren que sus componentes se ejecuten en máquinas o procesos separados, lo que permite su distribución o centralización. En el contexto de una arquitectura big data, comúnmente se consideran cuatro capas lógicas, según [33].

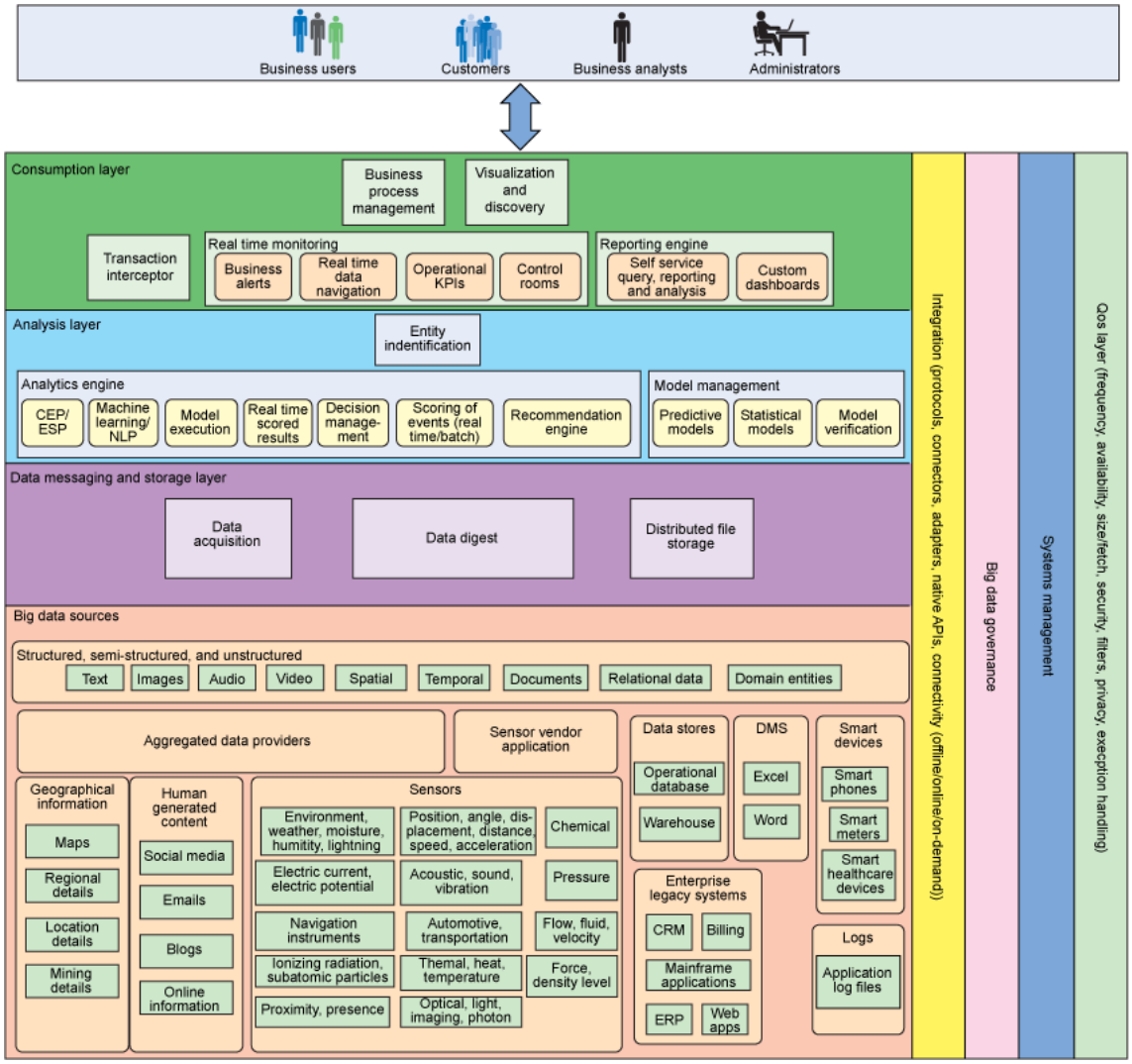
- **Fuentes de datos o fuentes big data**: Donde se generan o recogen los datos.
- **Capas de mensajería y almacenamiento de datos**: Encargadas de almacenar y transmitir los datos.
- **Capa de análisis**: Donde se procesan y analizan los datos.

ISBN: 978-631-6557-56-8

- **Capa de consumo o capas consumidoras**: Que permiten el acceso y uso de los datos procesados por los usuarios o sistemas.

En la Figura 21 se muestra la arquitectura Big Data.

Figura 21: Las cuatro capas principales en una arquitectura Big Data.



Fuente: [33]

2.6.1. Fuentes Big Data

Las fuentes de datos o fuentes big data comprenden toda la información que será utilizada para análisis, proveniente de diversos canales que varían según la solución implementada. Elegir las fuentes adecuadas es uno de los primeros pasos en un proyecto big data.

Una práctica recomendada es que los científicos de datos definan cuáles fuentes son necesarias para obtener resultados útiles para la empresa. Estas fuentes pueden diferir en:

- **Formato**: estructurado, semiestructurado o no estructurado.

ISBN: 978-631-6557-56-8

- **Velocidad y volumen**: varían según la fuente y el flujo de datos.
- **Punto de recolección**: los datos pueden obtenerse directamente, de terceros, en tiempo real o en procesos por lotes.
- **Ubicación**: las fuentes pueden ser internas o externas, y el acceso limitado a algunas de ellas puede restringir el análisis.

2.6.2. Capas de mensajería y almacenamiento

Esta capa, que a veces se divide en adquisición y almacenamiento, se encarga de recoger datos desde diversas fuentes y, cuando es necesario, transformarlos a un formato adecuado para su posterior análisis. Por ejemplo, una imagen capturada por una cámara infrarroja en una fábrica podría necesitar ser convertida antes de almacenarse en un sistema como HDFS de Apache Hadoop o en una base de datos relacional como MySQL, MariaDB o SQL Server. Además, es fundamental que durante este proceso se respeten las normativas de protección de datos vigentes en cada país, garantizando el almacenamiento seguro y adecuado según el tipo de información.

2.6.3. Capa de análisis

La capa de análisis se encarga de procesar los datos que han sido gestionados por la capa de mensajería y almacenamiento, aunque en algunos casos también puede acceder directamente a las fuentes de datos, especialmente si se trata de fuentes externas como *Big and Open Linked Data*, donde no siempre es necesario almacenar toda la información. En esta capa se aplican técnicas como *MapReduce, Apache Spark*, algoritmos de inteligencia artificial y aprendizaje automático. Su diseño exige una planificación cuidadosa, definiendo cómo se realizarán los análisis, cómo se extraerá información útil, cómo se identificarán las entidades relevantes, qué fuentes de datos se utilizarán y qué herramientas y algoritmos serán necesarios para obtener los resultados esperados.

2.6.4. Capa de consumo

Es la encargada de recibir y utilizar los resultados generados por la capa de análisis. Sus consumidores pueden ser aplicaciones de visualización, usuarios, procesos de negocio o servicios. En algunos casos, esta capa también es conocida como la capa de aplicación, y uno de sus retos es mostrar de manera clara y útil los resultados obtenidos del análisis. Además, es recomendable observar cómo otras empresas o competidores presentan sus resultados para optimizar esta etapa. Es importante recordar que tanto la seguridad como la privacidad son aspectos elementales en cualquier entorno industrial y en toda arquitectura de datos.

ISBN: 978-631-6557-56-8

2.6.5. Capas verticales

Capas verticales

En las arquitecturas Big Data, además de las capas lógicas, suelen existir capas transversales llamadas “capas verticales” que influyen en todos los componentes, encargándose de aspectos importantes como la seguridad, privacidad y calidad de los datos. Entre las más comunes se encuentran:

- **Integración de la información.-** Esta capa permite conectar diversas fuentes de datos —que pueden ser de diferentes orígenes y formatos— y unificarlas para su almacenamiento y posterior análisis en sistemas como HDFS, NoSQL o MongoDB. Se encarga no solo de ingresar datos al sistema (ingestión), sino también de integrar datos de distintas fuentes y preparar su almacenamiento o consulta mediante servicios y APIs.

- **Gobernanza de datos.-** Se centra en establecer normas y buenas prácticas para manejar los datos de forma eficiente y segura durante todo su ciclo de vida: desde que ingresan, se procesan, se almacenan, se analizan y, finalmente, se eliminan o archivan. Una gobernanza adecuada es clave para gestionar el volumen y la diversidad de datos.

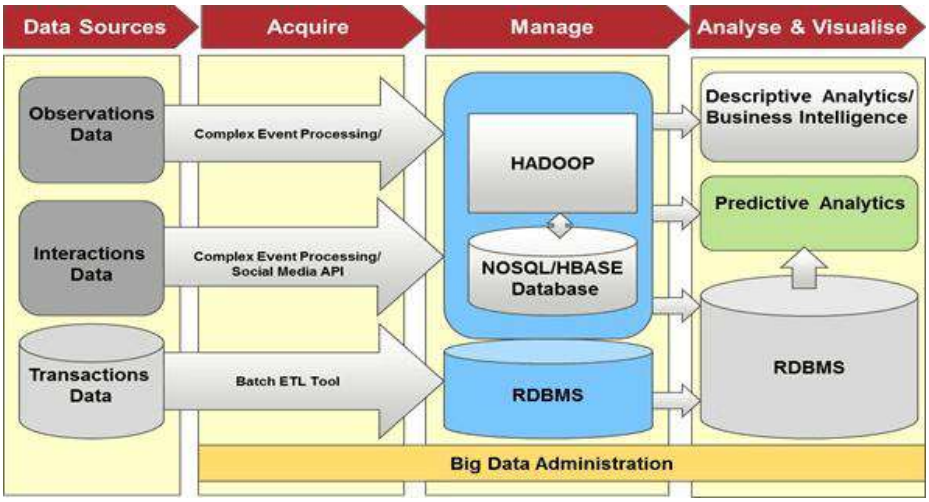
- **Capa de calidad de servicio.-** Define las políticas de calidad de datos, seguridad y privacidad, así como criterios sobre frecuencia de actualización, tamaño de los datos adquiridos y aplicación de filtros que depuran la información antes del análisis.

- **Gestión de sistemas.-** Supervisa el funcionamiento de la infraestructura Big Data, incluyendo servidores, redes, almacenamiento y sistemas virtualizados. Además, se encarga de la detección de fallos, gestión de alertas y del cumplimiento de acuerdos de servicio.

En algunas arquitecturas, estas capas pueden variar en número o funciones, pero siempre siguen una lógica de tareas similar. Es común ver distintas agrupaciones o divisiones según las necesidades de cada sistema, como sucede en modelos propuestos por otros autores, por ejemplo [34], donde las capas de adquisición, almacenamiento, análisis y visualización pueden combinarse o separarse según el enfoque como se muestra en la Figura 22.

ISBN: 978-631-6557-56-8

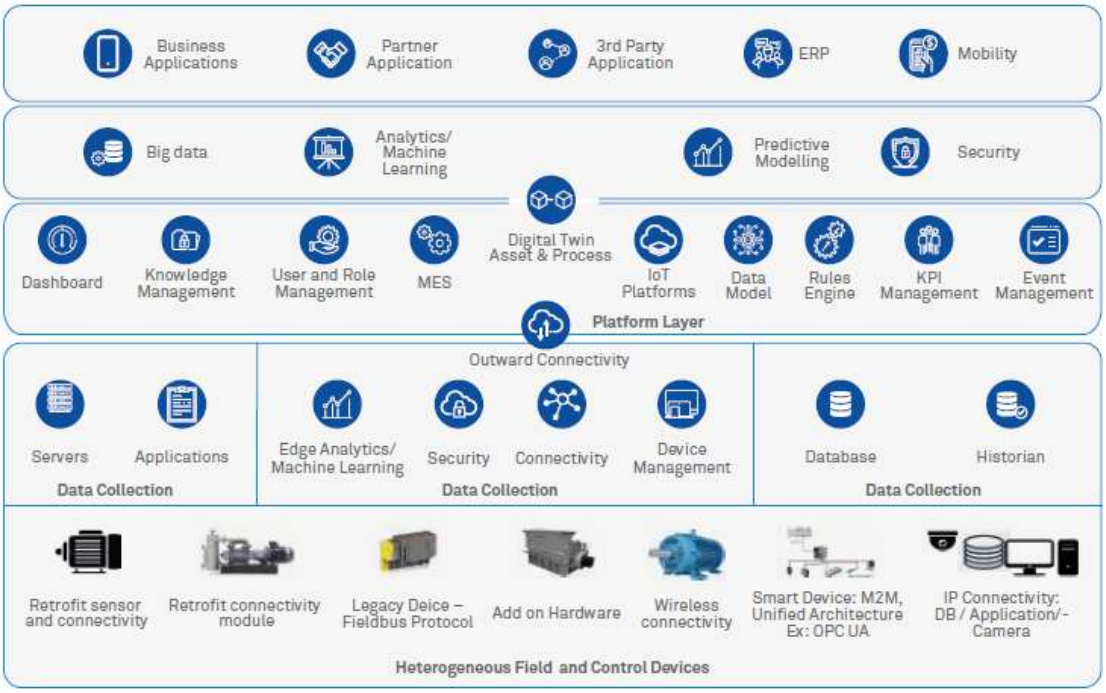
Figura 22: Variante de arquitectura big data de cuatro capas



Fuente: [34]

Por último, la Figura 23 muestra una arquitectura con diversas capas lógicas, propuesta por Nath (2020), diseñada para el procesamiento de datos en entornos IIoT dentro de la Industria 4.0. Aunque no es una arquitectura específica de Big Data, sigue un flujo de tareas muy similar, manteniendo el mismo orden lógico en las fases de tratamiento de datos.

Figura 23: Ejemplo de arquitectura IoT industrial



Fuente: [35]

2.7. Fuentes de datos

Esta capa incluye todas las fuentes de datos necesarias para proporcionar la información requerida a la hora de resolver la problemática que se haya planteado en el entorno industrial en particular. Como ejemplo, si estamos implantando una solución en agroindustria (producción de galletas, por ejemplo), las fuentes de datos podrán incluir estaciones IoT en los cultivos de cereales, pero

ISBN: 978-631-6557-56-8

también servicios externos de información meteorológica, fuentes externas que nos permitan acceder a las cotizaciones de los productos agrícolas en el mercado o incluso datos provenientes de redes sociales para analizar el sentimiento de los consumidores acerca de nuestros productos.

Sistemas existentes de la empresa o industria.- Esta categoría abarca las aplicaciones y plataformas que gestionan tanto las operaciones empresariales como las industriales, y que proporcionan datos importantes para el análisis y la toma de decisiones. El acceso a estos sistemas se realiza generalmente mediante APIs (por ejemplo, REST) o, en menor medida, directamente desde sus bases de datos. Entre estos sistemas se incluyen:

- CRM (gestión de relaciones con clientes).
- Sistemas de facturación.
- ERP (planificación de recursos empresariales).
- MES (ejecución de la fabricación).
- PLM (gestión del ciclo de vida del producto).
- Aplicaciones web y otros sistemas que amplían los datos disponibles en la empresa, muchas veces mediante protocolos propios.

Sistemas de gestión de datos (DMS).- Estos sistemas almacenan documentos legales, de procesos y políticas, los cuales pueden transformarse en datos estructurados útiles para análisis. Ejemplos incluyen hojas de cálculo y documentos de texto.

Data Warehouses.- Almacenan grandes volúmenes de datos estructurados (operativos y transaccionales) que pueden consumirse directamente o adaptarse para análisis, y que pueden integrarse en sistemas distribuidos.

Dispositivos inteligentes.- Equipos como smartphones o tablets permiten la captura, procesamiento y transmisión de datos en tiempo real o en lotes, por ejemplo en logística o control de inventarios.

Dispositivos IIoT.- En entornos industriales, sensores y dispositivos conectados permiten recolectar datos como temperatura, humedad, presión, vibraciones, consumo de energía o lecturas de etiquetas, lo que facilita la automatización, trazabilidad y control en las fábricas.

Tecnología Operacional (OT).- Incluye software y protocolos de comunicación propios de la industria, como OPC/OPC-UA para interactuar con sistemas SCADA y PLC, o PPMP para evaluar el rendimiento de producción.

ISBN: 978-631-6557-56-8

Proveedores de datos agregados.- Empresas especializadas suministran datos procesados, filtrados y en formatos específicos, como información meteorológica para predicciones en energía renovable.

Otras fuentes de datos.- Existen múltiples fuentes adicionales, tanto automatizadas como humanas, incluyendo:

- Información geográfica y de localización.
- Datos generados por usuarios en redes sociales, blogs y correos electrónicos, relevantes para análisis de percepción y sentimiento del mercado.

2.8. Capas de mensajería y almacenamiento

Dada la variedad en las características de los datos que se reciben —diferencias en frecuencia, formato, tamaño y canal de comunicación—, las capas encargadas de la mensajería y el almacenamiento deben contar con la flexibilidad necesaria para gestionar esta diversidad de manera eficiente (Mysore, Khupat y Jain, 2013). Dentro de estas capas se distinguen los siguientes componentes:

- **Adquisición de datos.-** Este módulo se encarga de recopilar información desde múltiples fuentes y enviarla al sistema de procesamiento o almacenarla en ubicaciones específicas. Además, debe tomar decisiones inteligentes sobre si los datos deben almacenarse, transferirse o enviarse directamente al análisis, en función de su naturaleza y prioridad.
- **Digestor de datos.-** Su función es transformar los datos al formato requerido para su posterior análisis. Esta transformación puede ser desde una conversión simple hasta la aplicación de algoritmos estadísticos complejos, dependiendo de las necesidades del motor analítico. Un reto significativo es la estandarización de datos no estructurados, como imágenes, audios o vídeos.
- **Almacenamiento distribuido de datos.-** Este componente se ocupa de guardar la información proveniente de las diferentes fuentes en estructuras adaptadas a sus características. Las opciones incluyen sistemas de archivos distribuidos (DFS), almacenamiento en la nube, bases de datos estructuradas, o bases NoSQL, dependiendo de los requisitos del proyecto.

2.9. Capa de análisis

La capa de análisis, también conocida como capa de procesamiento o capa de procesamiento y análisis, es el nivel encargado de transformar los datos en conocimiento útil para la toma de decisiones empresariales. En algunos casos, el procesamiento y el análisis se gestionan en capas

ISBN: 978-631-6557-56-8

diferenciadas, como se observó en la figura 3. Los componentes principales de esta capa son los siguientes:

- **Análisis (Identificación de entidades).-** Este componente detecta y construye entidades contextuales relevantes a partir de los datos. Se trata de un proceso exigente que demanda técnicas de alto rendimiento. Además, requiere de la colaboración con la síntesis de datos, la cual adapta la información al formato necesario para que los motores analíticos puedan operar con eficiencia.
- **Motor de análisis.-** Es el encargado de ejecutar el procesamiento de datos aplicando modelos, algoritmos y flujos de trabajo definidos. Este motor aprovecha componentes adicionales, como la identificación de entidades y la gestión de modelos, y es capaz de trabajar en entornos que permiten el procesamiento paralelo para optimizar su rendimiento.
- **Gestión de modelos.-** Administra, verifica y entrena modelos estadísticos o de aprendizaje automático. Este componente garantiza que los modelos se mantengan actualizados y precisos, permitiendo que puedan integrarse tanto en la identificación de entidades como en los procesos analíticos posteriores.

2.10. Capas Consumidoras de datos

La **capa de consumo** (también conocida como capa de salida) es la que hace accesible el conocimiento empresarial derivado de las aplicaciones de análisis. Este conocimiento es consumido tanto por usuarios internos de la organización como por entidades externas, tales como clientes, proveedores y socios. Los resultados del análisis son cruciales para diferentes sectores, permitiendo que las empresas personalicen sus ofertas para clientes, optimicen procesos industriales o incluso detecten fraudes en tiempo real.

Por ejemplo, en el sector de retail, una empresa puede utilizar la información sobre las preferencias y la ubicación de un cliente para ofrecerle productos personalizados mientras se encuentra cerca de la tienda. En el ámbito industrial, el análisis puede ser consumido por directivos y responsables de planta para identificar procesos susceptibles de mejora. En el sector bancario y comercial, los resultados del análisis permiten detectar transacciones fraudulentas, incluso durante su ejecución, para tomar medidas correctivas de inmediato.

A nivel de manufactura, el conocimiento obtenido también puede facilitar la predicción de fallos en equipos industriales, como en el caso de un aerogenerador, y permitir la ejecución de mantenimientos predictivos. Además, las interacciones con sistemas de gestión empresarial

ISBN: 978-631-6557-56-8

(como los CRM o MES) permiten automatizar diversos procesos, como la creación de pedidos o el bloqueo de tarjetas de crédito ante fraude.

Uno de los componentes clave de esta capa es el **motor de recomendación**, que proporciona sugerencias personalizadas en tiempo real a clientes, basándose en su historial y preferencias. En el sector industrial, estos motores también pueden ayudar a recomendar los mejores proveedores según los resultados de la producción.

La **capa de consumo** también permite a los usuarios internos acceder, comprender y explorar datos federados dentro y fuera de la organización. La creación de informes y tableros de mando facilita la toma de decisiones informadas, permitiendo, por ejemplo, la supervisión de indicadores clave de rendimiento (KPIs) y la recepción de alertas comerciales en tiempo real.

Entre los principales componentes de esta capa se encuentran:

- **Interceptor de transacciones**: Este componente gestiona transacciones de gran volumen en tiempo real y las convierte a un formato adecuado para su análisis. Además, se integra con diversas fuentes de datos, como sensores y dispositivos inteligentes.
- **Procesos de gestión empresarial**: Facilitan la automatización de funciones mediante el consumo de datos provenientes de la capa de análisis, mejorando el valor empresarial.
- **Monitorización en tiempo real**: Permite generar alertas a partir de los datos de la capa de análisis, brindando información sobre la eficacia operativa y el estado del sistema en tiempo real.
- **Motor de información**: Facilita la creación de informes y consultas ad hoc basadas en los resultados del análisis.
- **Visualización y descubrimiento**: Permite navegar a través de diversas fuentes de datos, integrando información estructurada, semiestructurada y no estructurada, y presentando vistas personalizadas a los usuarios en función de sus preferencias y permisos.

2.11. *Cloud Computing*

En este contexto, aún no hemos abordado específicamente el lugar donde se realiza el procesamiento y almacenamiento de las grandes cantidades de datos generadas. Muchas empresas, por diversas razones, no cuentan con servidores *on-premise* (en sus propias instalaciones) que puedan ofrecer la capacidad de procesamiento y almacenamiento que requieren las soluciones de **big data**. Adquirir estos equipos implica una elevada inversión inicial, costos de mantenimiento y la necesidad de amortizarlos a medida que se vuelven obsoletos con el tiempo. Aunque se puede mitigar este problema mediante el escalado vertical, cuando es necesario

ISBN: 978-631-6557-56-8

recurrir al escalado horizontal (aumentar la capacidad añadiendo nodos adicionales), las inversiones aumentan significativamente, incluso si los nuevos nodos son de bajo coste.

Además, algunas empresas o fábricas pueden necesitar una capacidad de cómputo y almacenamiento mucho mayor solo en ciertos picos de demanda, mientras que en otros períodos no requieren tal nivel de servicio. En este sentido, **la computación en la nube** (o *cloud computing*) ofrece una solución ideal, ya que permite acceder a recursos de computación y almacenamiento bajo demanda, sin necesidad de hacer inversiones iniciales. Los servicios se cobran según el uso real de los recursos, lo que permite a las industrias pagar solo por la capacidad que necesitan en cada momento.

2.11.1. Surgimiento de la computación en la nube

Gracias a la computación en la nube, las empresas pueden acceder a recursos computacionales y de almacenamiento de manera elástica, es decir, ajustando dinámicamente la capacidad según sus necesidades, sin preocuparse por la obsolescencia de los equipos o los costos de mantenimiento. Esto resulta fundamental para implementar soluciones de big data, que requieren grandes cantidades de procesamiento y almacenamiento distribuidos.

Históricamente, la computación en la nube fue conceptualizada por *R. K. Chellappa* en 1997, quien predijo que el modelo computacional del futuro estaría más enfocado en los intereses económicos que en las limitaciones tecnológicas. Aunque esta visión pudo haber parecido utópica en su momento, los avances tecnológicos han demostrado que, efectivamente, los intereses macroeconómicos de grandes empresas tecnológicas han impulsado el modelo de ***cloud computing***. Este paradigma está estrechamente vinculado a los intereses empresariales, lo que ha facilitado su rápido crecimiento.

El modelo comercial en la computación en la nube se basa en un esquema de pago por uso, similar al de servicios tradicionales como electricidad o agua. Esto permite a los usuarios negociar un **Acuerdo de Nivel de Servicio (SLA)** con el proveedor de servicios en la nube para acceder a los recursos necesarios en cada momento. De esta manera, los usuarios solo pagan por los recursos que realmente utilizan, lo que optimiza los costos. Un concepto fundamental en *cloud computing* es la elasticidad, que se refiere a la capacidad de ajustar dinámicamente la cantidad de recursos disponibles según la demanda. Este enfoque se basa en el modelo *just-in-time*, que asegura que los recursos solo se proporcionen en la cantidad necesaria para mantener un nivel constante de calidad. Aunque puede haber costos adicionales asociados a la flexibilidad de contar con capacidad a demanda, estos son generalmente mucho menores que los costos de mantener continuamente esa capacidad máxima.

ISBN: 978-631-6557-56-8

2.11.2. La arquitectura de referencia *cloud* del NIST

Para comprender la arquitectura cloud de manera general, sin entrar en los términos comerciales específicos de cada proveedor, se puede utilizar la **arquitectura de referencia cloud** propuesta por el **Instituto Nacional de Estándares y Tecnología** (NIST) de Estados Unidos. Según el NIST, los servicios *cloud* deben tener las siguientes características elementales:

- **Servicios a la carta**: Los servicios deben ser proporcionados automáticamente, sin intervención humana, y ajustados a las demandas del usuario.
- **Disponibilidad de servicios a través de Internet**: Los servicios deben ser accesibles a través de Internet, y los proveedores deben emplear este medio para ofrecer sus recursos.
- **Disponibilidad de recursos**: Los proveedores deben garantizar la disponibilidad de servicios, incluso con variaciones en la demanda, mediante la asignación dinámica de recursos, tanto físicos como virtuales.
- **Elasticidad**: Los recursos deben suministrarse de manera flexible y, en algunos casos, automática según la demanda, lo que es crucial para el ***cloud computing***.

Tipos de Servicios y Modelos de Despliegue

El NIST clasifica los servicios en tres categorías principales y los modelos de despliegue en cuatro tipos.

Tipos de servicios (capacidades):

- **SaaS** *(Software as a Service):* En este modelo, el proveedor ofrece aplicaciones directamente al consumidor. Estas aplicaciones se ejecutan en la infraestructura de la nube. Un ejemplo es Google Docs, Gmail, o Office 365. Aunque el modelo es ventajoso por su ubicuidad y el uso de clientes ligeros, también presenta desventajas, como la falta de control del consumidor sobre la infraestructura.
- **PaaS** *(Platform as a Service):* El proveedor ofrece herramientas que permiten a los usuarios crear sus propias aplicaciones. Estas herramientas incluyen bibliotecas, entornos de programación y otras utilidades. Un ejemplo es Google App Engine, que permite desarrollar aplicaciones en la nube utilizando lenguajes como Python, Java, C++, o Node.js.
- **IaaS** *(Infrastructure as a Service):* En este modelo, el consumidor accede a recursos de hardware como CPU, RAM y almacenamiento. Puede ser a través de máquinas físicas dedicadas (**bare metal**) o virtuales, donde los recursos físicos se comparten entre varios

ISBN: 978-631-6557-56-8

clientes. Un ejemplo es el acceso a máquinas virtuales a través de plataformas como **Amazon EC2**.

Modelos de despliegue:

- **Nube privada** *(Private Cloud):* En este modelo, la infraestructura *cloud* es utilizada por una única organización, aunque puede incluir diferentes consumidores dentro de la misma entidad.
- **Nube pública** *(Public Cloud):* Infraestructuras de uso público, accesibles para cualquier usuario. Los recursos son gestionados y mantenidos por el proveedor de la nube.
- ***Virtual Private Cloud (VPC)****:* Un modelo híbrido que combina la infraestructura privada sobre servicios públicos. Permite que una nube privada se construya sobre la infraestructura de una nube pública.
- ***Community Cloud****:* Utilizada por un grupo específico de consumidores u organizaciones que comparten un interés común o necesidades similares. La infraestructura es accesible solo para este grupo.
- **Nube híbrida** *(Hybrid Cloud):* Permite combinar cualquiera de los modelos anteriores, lo que facilita la interoperabilidad entre plataformas. Sin embargo, esto puede generar complejidades, aunque soluciones *como Google Anthos* o *Red Hat Openshift* han facilitado esta integración.

La arquitectura de referencia *cloud* del NIST describe cómo los servicios y modelos se combinan, y se enfocan en los roles clave que desempeñan los diferentes paradigmas dentro del *cloud computing.*

2.11.3. El consumo energético en el *cloud computing*

En cuanto a los costos energéticos, se ha demostrado que la computación en la nube genera un ahorro de energía en comparación con las soluciones tradicionales implementadas de manera local. Según Greenpeace, se calcula que el sector de las TIC representa alrededor del 7% del consumo mundial de electricidad [36]. Con un aumento proyectado en el tráfico global de Internet para 2020, se espera que la huella energética de Internet crezca aún más, impulsada por el mayor consumo individual de datos y por la expansión de la era digital, que alcanzará a una mayor parte de la población mundial, pasando de 3,000 millones a más de 4,000 millones de personas en todo el mundo. Otras fuentes, como la consultora Gartner, ofrecen conclusiones similares, pero con un enfoque específico en el sector TIC industrial. Según Gartner, la huella de carbono de los servidores privados es considerable, representando aproximadamente el 2% del total mundial, debido a los recursos inmobiliarios necesarios para alojarlos y almacenarlos [37].

ISBN: 978-631-6557-56-8

2.12. *Edge Computing*

En el contexto de la industria 4.0 y el Internet de las Cosas (IoT), los proveedores de servicios en la nube cobran a los usuarios según los recursos utilizados, como computación y almacenamiento. El IoT conecta millones de dispositivos que recogen datos y los envían a la nube para su procesamiento y análisis. A esta información se añaden los datos generados por robots, sistemas SCADA, ERP y otros componentes industriales de la IT y OT.

El *edge computing* emerge como una solución para reducir los costos de la nube, al procesar y filtrar los datos en el borde de la red, antes de enviarlos a la nube. Esto también mejora la respuesta ante eventos, ya que la toma de decisiones se traslada al borde, cerca de las fuentes de datos. Además, el IoT enfrenta retos como la heterogeneidad de los dispositivos, lo que complica la gestión de datos y la comunicación. Las arquitecturas de *edge computing* abordan estos desafíos al permitir la computación más cerca de los dispositivos, reduciendo la necesidad de recursos en la nube. A medida que más dispositivos IoT se conectan a través de diversas tecnologías, el procesamiento y almacenamiento de datos en la nube puede resultar costoso. El *edge computing* ayuda a minimizar el tráfico hacia la nube y los costos asociados. Por ejemplo, dispositivos como la Raspberry Pi o el Edge TPU de Google, junto con librerías como *TensorFlow Lite*, permiten ejecutar algoritmos de aprendizaje automático en el borde, mejorando la eficiencia y reduciendo la latencia. La evolución de arquitecturas como M2M (máquina a máquina), M2C (máquina a la nube) y M2G (máquina a puerta de enlace) muestra cómo el *edge computing* puede aliviar los problemas de comunicación y procesamiento, moviendo la carga computacional hacia nodos de borde. Esto no solo reduce la latencia, sino que también mejora la eficiencia energética de los dispositivos IoT.

Por otra parte, la arquitectura de IoT, que involucra capas IoT, Edge y Cloud, se compone de diferentes funcionalidades a nivel de Big Data, que se distribuyen en cada una de las tres capas, como se describe a continuación:

Capa IoT (IoT *Layer o Front-End*)

La capa IoT está compuesta por dispositivos IoT como sensores, actuadores y dispositivos inteligentes que recopilan o modifican la información del entorno. A nivel de aplicación, se encarga de las fuentes de datos y servicios de mensajería, y sus dispositivos son distribuidos, con capacidad de computación limitada y pequeño almacenamiento. En relación con Big Data, su principal función es actuar como la fuente de datos, siendo el punto de entrada de la información que luego será procesada y analizada en las capas posteriores.

Capa Edge (*Edge Layer o Near-End*)

ISBN: 978-631-6557-56-8

La capa Edge está compuesta por nodos en el borde (*edge nodes*) que actúan como pasarelas entre la capa IoT y la nube, permitiendo almacenar datos durante interrupciones de comunicación y reduciendo el consumo de energía de los dispositivos IoT. Además, realiza parte del procesamiento de datos, aliviando la carga computacional de la nube y proporcionando respuestas rápidas a los usuarios, y es capaz de ejecutar técnicas de aprendizaje automático en el borde de la red. A nivel de aplicación, se encarga del análisis de datos en tiempo real, la toma de decisiones y el almacenamiento temporal de datos. En relación con Big Data, la capa Edge se dedica al preprocesamiento y filtrado de datos, ofreciendo tiempos de respuesta más rápidos en comparación con la nube, ya que el procesamiento ocurre localmente.

Capa *Cloud (Cloud Layer o Far-End)*

La capa Cloud está compuesta por servicios desplegados en la nube, como máquinas virtuales, servicios distribuidos y plataformas como servicio (PaaS), entre otros. En términos de almacenamiento, puede incluir data *lakes*, data *warehouses*, bases de datos como servicio (DBaaS), entre otros. A nivel de aplicación, se encarga de tareas como la elaboración de informes, el análisis de datos a largo plazo y el almacenamiento de datos a largo plazo. En relación con Big Data, esta capa se dedica al procesamiento y almacenamiento de grandes volúmenes de datos, aunque ofrece mayores tiempos de respuesta en comparación con la capa Edge, debido a la distancia y la naturaleza centralizada de la nube.

En la Tabla 3 se muestra las funciones y capacidades de cada capa en una arquitectura *cloud-edge*-IoT y su relación con el big data.

Tabla 3: Funciones y capacidades de cada capa en una arquitectura *cloud-edge*-IoT y su relación con el big data

Capa	Funcionalidad	Big Data
IoT	Recopilación de datos, dispositivos inteligentes, mensajes en tiempo real	Fuentes de datos
Edge	Análisis y procesamiento en tiempo real, preprocesamiento, almacenamiento temporal	Preprocesamiento y filtrado
Cloud	Almacenamiento y análisis a largo plazo, informes y aplicaciones distribuidas	Procesamiento y almacenamiento

Fuente: Autores

2.13. Plataformas *Cloud* y *Edge*

Tras revisar los conceptos fundamentales de las arquitecturas Cloud y Edge, sus ventajas y su relación con el procesamiento Big Data, es importante explorar algunas de las soluciones comerciales más populares en el mercado.

ISBN: 978-631-6557-56-8

2.13.1. Máquinas virtuales, contenedores, Docker, *Kubernetes* y FaaS

Las máquinas virtuales (VM) permiten ejecutar diferentes sistemas operativos en un mismo servidor físico, pero consumen muchos recursos debido a la necesidad de replicar el sistema operativo en cada VM. Esto se resuelve con los **contenedores**, que son más ligeros y comparten el núcleo del sistema operativo anfitrión, lo que reduce significativamente el consumo de recursos y el tiempo de inicio. *Docker* es una tecnología de contenedores de código abierto que facilita el despliegue de estos contenedores en sistemas Windows o Linux. Para gestionar y escalar contenedores, se utilizan herramientas como *Kubernetes* o *Docker Swarm*. Además, existen soluciones de computación en la nube como *Function as a Service* (FaaS), que permiten ejecutar código bajo demanda, sin necesidad de servidores dedicados, lo que es ideal para escenarios de IoT y procesamiento de datos, donde se invoca el código al recibir nuevos datos de un sensor o un evento específico.

2.13.2. Amazon Web Services

La principal fortaleza de *Amazon* se encuentra en su dominio del mercado de infraestructura de nube pública, destacándose en el "*Cuadrante Mágico"* , donde AWS ha liderado el mercado de IaaS durante más de 10 años. Su popularidad se debe en gran parte a la amplitud de sus operaciones, con una amplia gama de servicios y una red global de centros de datos, AWS es el proveedor más maduro y preparado para empresas, con la capacidad de gestionar una gran cantidad de usuarios y recursos. Sin embargo, su debilidad radica en la complejidad de su estructura de costes, lo que dificulta su gestión eficaz, especialmente en grandes volúmenes de trabajo. A pesar de esto, las ventajas de Amazon superan sus desventajas, y organizaciones de todos los tamaños siguen utilizando AWS para diversas cargas de trabajo. En el ámbito de IoT y *edge computing*, Amazon ha lanzado AWS *Greengrass* para procesamiento en el borde, y ofrece servicios industriales como AWS IoT *SiteWise*, que permiten la recolección de datos industriales a través de *gateways* en las instalaciones del cliente [38].

2.13.3. Microsoft Azure

Microsoft, aunque llegó tarde al mercado de la nube, logró posicionarse en la segunda posición gracias a su estrategia de reutilizar software *on-premise* como Windows Server, Office, SQL Server, SharePoint y otros, adaptándolos a la nube. Esto le dio una ventaja, especialmente con las empresas que ya utilizan software de Microsoft, ya que la integración con Azure resulta más conveniente para ellos, generando lealtad entre sus clientes actuales y ofreciendo descuentos significativos en contratos de servicio. Sin embargo, una de las desventajas de Azure es la experiencia del servicio, que no está tan refinada como se esperaba de una compañía con el historial de Microsoft, con problemas reportados en soporte técnico, documentación, formación y ecosistema de socios. Se espera que Microsoft corrija estos problemas para mantenerse

ISBN: 978-631-6557-56-8

competitivo frente al avance de *Google Cloud Platform*. En cuanto a sus fortalezas, *Azure* se destaca en el ámbito industrial y de big data, con herramientas como Power BI para la visualización de datos y la integración con IoT. Además, Azure cuenta con servicios como IoT Hub e IoT Edge para la recolección y procesamiento de datos IoT, y sus dispositivos como *Hololens* ofrecen soluciones innovadoras para la industria 4.0.

2.13.4. *Google Cloud Platform* (GCP)

Se destaca por su fuerte oferta en contenedores, especialmente debido al desarrollo de *Kubernetes*, un estándar para la orquestación de contenedores que ahora también utilizan AWS y *Azure*. GCP es particularmente fuerte en áreas de alta computación, big data, análisis y aprendizaje automático, y cuenta con una infraestructura robusta, incluyendo una red propia de centros de datos y fibra óptica transoceánica, lo que le da una ventaja en términos de tiempos de respuesta rápidos. Sin embargo, Google ocupa el tercer lugar en cuota de mercado, ya que históricamente no ha tenido una relación establecida con clientes empresariales, habiéndose enfocado más en el usuario final. A pesar de esto, GCP está expandiendo rápidamente su red de centros de datos y ofertas. Según [38], las empresas suelen elegir GCP como proveedor secundario, aunque está ganando terreno como una alternativa estratégica frente a AWS, especialmente entre empresas centradas en código abierto o *DevOps*. En cuanto a big data, GCP ofrece herramientas como *BigQuery y BigQuery ML*, que permiten crear y ejecutar modelos de aprendizaje automático usando un lenguaje similar a SQL. En el ámbito de IoT y *Edge*, GCP ofrece *Cloud* IoT *Core* para gestionar dispositivos IoT y *Cloud* IoT *Edge* para ejecutar modelos de *machine learning* en el borde de la red, utilizando *TensorFlow Lite*.

2.14. Ejemplos de arquitecturas Big Data en el contexto de la Industria 4.0

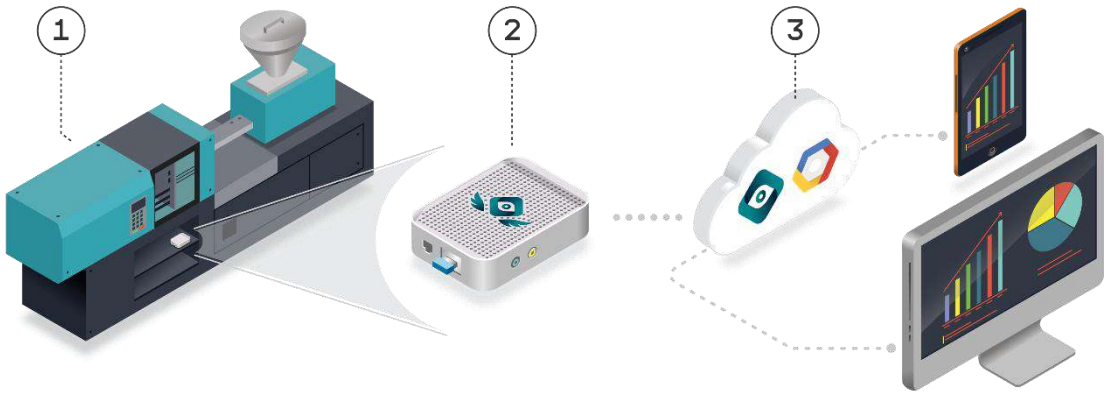
Para finalizar, se presentan algunos casos de uso industriales reales que aplican arquitecturas de *big data, cloud computing y edge computing* a través de plataformas comerciales. Un ejemplo destacado es *WirelessCar*, una empresa fundada por *Volvo, Telia y Ericsson* en 1999, que usa *Amazon Web Services* (AWS) para ofrecer servicios de control de flotas y vehículos conectados en ciudades inteligentes. A finales de 2020, contaba con más de 4 millones de vehículos conectados, gestionando picos de más de 300,000 mensajes por minuto. Este caso refleja la posición de AWS como líder en el mercado, utilizando tecnologías como las máquinas virtuales EC2 y el almacenamiento S3, que son fundamentales para el procesamiento de big data en la nube.

Otro ejemplo en el contexto de la Industria 4.0 es *Oden Technologies*, que ofrece software de gestión IoT industrial, como monitoreo de dispositivos, mantenimiento predictivo y análisis de rendimiento en manufactura. Esta empresa utiliza *Google Cloud Platform* (GCP) para

ISBN: 978-631-6557-56-8

proporcionar estos servicios a clientes industriales, este ejemplo se muestra a continuación en la Figura 24.

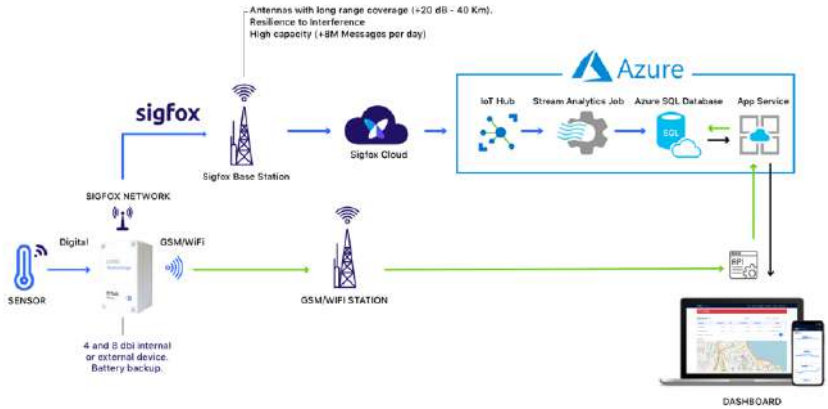
Figura 24: Soluciones IIoT sobre Google Cloud de Oden Technologies



Fuente: [39]

Finalmente, el caso de *Verse Technology* y la panificadora Bimbo ilustra cómo Microsoft Azure se utiliza para conectar y gestionar más de 1000 dispositivos IoT en la industria alimentaria. La solución incluye Azure IoT Hub para transmitir datos desde diversos dispositivos en el borde de la red y *Azure SQL Database* para almacenar información sobre variables como rpm, temperatura y consumo de gas. En el futuro, se planea expandir esta red a más de 10,000 dispositivos, el esquema de esta estructura se describe a continuación en la Figura 25.

Figura 25: Verse *Technology* para la panificadora Bimbo utilizando Azure



Fuente: [40]

ISBN: 978-631-6557-56-8

En los siguientes códigos QR o enlaces se muestra información adicional relacionada al capítulo.

Comparativa en BD SQL vs NoSQL	Caso de estudio *Even Better*
	
https://www.youtube.com/watch?v=ZS_kXvOeQ5Y	https://aws.amazon.com/es/solutions/case-studies/wireless-car/

ISBN: 978-631-6557-56-8

ISBN: 978-631-6557-56-8

CAPÍTULO III

INGENIERÍA Y PROCESAMIENTO DE DATOS CON IA

3.1. Introducción y Objetivo del Capítulo

En el presente capítulo se abordan los principales sistemas y herramientas que han surgido en respuesta a las crecientes necesidades de procesamiento y análisis de datos masivos, especialmente en el contexto de la Industria 4.0. A lo largo del contenido se presentarán tecnologías clave como *Hadoop*, HDFS, *MapReduce* y *Apache Spark*, las cuales han revolucionado la manera en que se almacenan, procesan y analizan grandes volúmenes de información en entornos distribuidos.

El recorrido inicia con la introducción de *Hadoop* y su ecosistema, surgido como solución a los desafíos que planteó el crecimiento exponencial de datos en la era digital, impulsado por la interacción en redes sociales, foros, comercio electrónico y servicios financieros online. Posteriormente, se profundiza en HDFS, el sistema de archivos distribuido que permite almacenar eficientemente grandes cantidades de datos, explicando sus componentes y funcionamiento básico.

Más adelante, se explora *MapReduce*, un paradigma de programación paralela que facilita el procesamiento distribuido de datos a gran escala, describiendo tanto su arquitectura como su flujo de ejecución. Complementando estas herramientas, se presenta *Apache Spark*, una plataforma moderna que amplía las capacidades de *Hadoop* mediante un enfoque más flexible y eficiente para el análisis de datos, destacando conceptos importantes como los RDD y las transformaciones que permiten operar sobre conjuntos de datos distribuidos.

Finalmente, se analiza la relevancia de estas tecnologías en aplicaciones reales dentro de la Industria 4.0, donde la inteligencia artificial y, en particular, el aprendizaje automático ha permitido optimizar procesos industriales y mejorar la toma de decisiones. Se presentan también algunos de los métodos más utilizados en este campo, como árboles de decisión, *clustering*, redes neuronales y sistemas de recomendación, demostrando cómo la sinergia entre big data e inteligencia artificial impulsa la innovación en sectores productivos.

En base a este antecedente el capítulo tiene por objetivos los siguientes:

- Identificar las razones que hacen necesaria la adopción de tecnologías de big data en el entorno industrial actual impulsado por la Industria 4.0.

ISBN: 978-631-6557-56-8

- Familiarizarse con los principales frameworks utilizados para el procesamiento de grandes volúmenes de datos, destacando componentes esenciales del ecosistema Apache Hadoop, como HDFS y Apache Spark.
- Comprender los principios fundamentales del paradigma de programación MapReduce y su aplicación en entornos de procesamiento distribuido.
- Analizar casos prácticos donde las tecnologías de big data han sido implementadas con éxito en escenarios reales de la Industria 4.0.

3.2. Necesidad de las tecnologías Big Data

En la era digital actual, la generación de datos crece de manera exponencial, superando ampliamente las cifras registradas en estudios de años anteriores. Ya en 2012, IBM advertía sobre el volumen masivo de datos que inundaba la red, destacando cifras como millones de correos electrónicos enviados cada día, terabytes subidos a redes sociales y exabytes generados en apenas días. Desde entonces, este crecimiento se ha acelerado impulsado por la expansión de las redes sociales, dispositivos IoT, sistemas de localización y la evolución constante de las tecnologías de la información y comunicación (TIC). El origen de estos datos puede rastrearse a tres tipos de interacciones: entre humanos a través de plataformas digitales; entre humanos y máquinas, como en la navegación web o sistemas de control remoto; y entre máquinas (M2M), donde sistemas como sensores y dispositivos IoT intercambian información de forma automática.

Más allá del volumen actual, el verdadero desafío radica en que esta tendencia seguirá creciendo a medida que más personas y dispositivos se conecten a la red, lo que exige nuevas soluciones para el almacenamiento y procesamiento eficiente de datos. Los sistemas tradicionales han quedado obsoletos ante esta explosión de información, por lo que el desarrollo de herramientas especializadas, tanto comerciales como de código abierto, se ha convertido en una necesidad fundamental para gestionar de manera óptima los datos en entornos modernos.

3.3. Hadoop

Hadoop es una plataforma de software libre, bajo licencia *Apache*, diseñada para facilitar la gestión y procesamiento de grandes volúmenes de datos. Su arquitectura se basa en dos componentes principales:

- **HDFS (*Hadoop Distributed File System*)**, que permite almacenar datos de forma distribuida en diferentes nodos de un *clúster*, asegurando redundancia y acceso eficiente, sin que el usuario necesite preocuparse por la gestión interna de esta distribución.

ISBN: 978-631-6557-56-8

- ***MapReduce***, un motor que divide y reparte las tareas de procesamiento entre los nodos de manera automática, optimizando el trabajo en paralelo.

Además, el ecosistema *Hadoop* incluye múltiples proyectos complementarios que amplían sus funcionalidades, simplifican tareas y mejoran la eficiencia. Con el paso del tiempo, herramientas como ***Apache Spark*** han cobrado gran relevancia, ya que ofrecen procesamiento de datos por lotes y en tiempo real, integrándose perfectamente con *Hadoop* y ampliando sus capacidades.

Hadoop ha sido adoptado por grandes empresas como Yahoo! y Facebook, que lo utilizan para procesar enormes cantidades de datos, demostrando su eficacia y escalabilidad. Por otra parte, además de las versiones estándar disponibles en la web de *Apache*, existen distribuciones comerciales que ofrecen paquetes de *Hadoop* ya probados y optimizados, con soporte adicional y, en algunos casos, entornos de prueba preconfigurados en máquinas virtuales llamados ***sandbox***, ideales para aprender y experimentar. Entre las distribuciones más conocidas se encuentran: ***Hortonworks*** (actualmente parte de Cloudera), ***Cloudera*** y ***MapR*** (actualmente HPE *Ezmeral Data Fabric*). Estas soluciones garantizan compatibilidad entre componentes y soporte especializado, facilitando el uso de *Hadoop* en entornos empresariales.

3.3.1. Despliegue de Hadoop

Para instalar Hadoop, es importante seguir una serie de pasos básicos generalizados:

- **Definir la arquitectura física**- Este primer paso consiste en planificar cuántos nodos formarán el clúster, qué funciones asumirá cada uno y cómo se organizarán físicamente. Aunque esta etapa requiere una visión clara del uso previsto, *Hadoop* permite escalar fácilmente la infraestructura añadiendo nuevos nodos según las necesidades futuras.

- **Elegir la distribución**- Las distribuciones de Hadoop ofrecen combinaciones de componentes ya probadas, junto con funciones adicionales y soporte técnico, lo que facilita una implementación estable.

En el contexto educativo, se recomienda usar una **máquina virtual** que simule un clúster de un solo nodo, ya que es poco probable contar con servidores físicos. Además, se sugiere optar por ***Hortonworks***, ya que su asistente de instalación, basado en *Apache Ambari*, simplifica mucho el proceso de configuración. A continuación se detalla estos pasos de manera más específica.

a) Creación de la máquina virtual, el primer paso es configurar una máquina virtual que alojará el entorno Hadoop. Para ello, se puede utilizar una herramienta de

ISBN: 978-631-6557-56-8

virtualización como ***Oracle VM VirtualBox***, compatible con los sistemas operativos más utilizados.

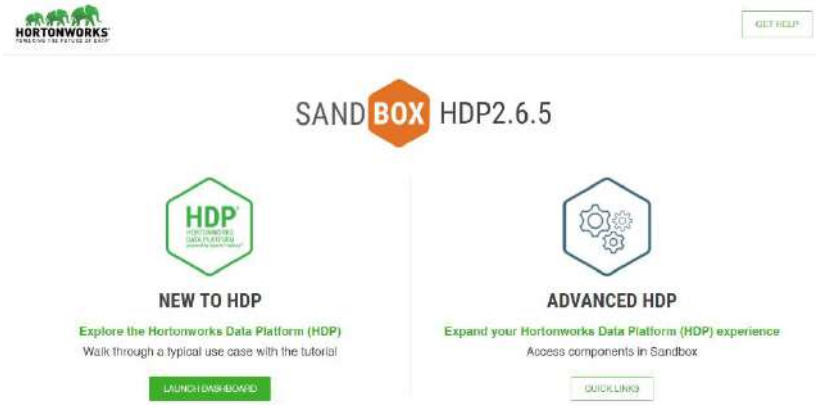
b) Una vez instalado VirtualBox, necesitamos descargar la última versión del *sandbox* de *Hortonworks Data Platform* (HDP), ahora a través del sitio web oficial de Cloudera. Se sugiere descargar el **sandbox** compatible con **VirtualBox**, el cual incluye una versión simplificada del ecosistema *Hadoop* lista para funcionar en un solo nodo. El archivo descargado, con extensión **.**OVA, permite importar la máquina virtual directamente en VirtualBox mediante la opción de "importar servicio virtualizado".

c) Una vez importado el servicio virtualizado ya se dispone en VirtualBox de una máquina virtual que contiene el *sandbox* de Hortonworks listo para ser ejecutado.

d) Una vez que está la máquina virtual importada, seleccionarla e iniciar su ejecución en el botón **Iniciar**.

e) Abrir un navegador web en la maquina máquina *host* y apuntamos a la dirección http://127.0.0.1:8888/ (localhost) para acceder a la interfaz de inicio como se observa en la Figura 26.

Figura 26: Interfaz del sandbox.



Fuente: [41]

f) Desde la página de inicio acceder a la interfaz de *Ambari*, disponible en la dirección http://127.0.0.1:8080/

g) Una vez iniciada la sesión correctamente con el usuario y la contraseña (que se proporcionan en el mismo enlace: **raj_ops/raj_ops**), se observa el cuadro de mando de *Ambari,* como se observa en la Figura 27.

ISBN: 978-631-6557-56-8

Figura 27: Cuadro de mando de Ambari

Fuente: [41]

h) Verificación**,** para confirmar que *Hadoop* se ha desplegado correctamente, es necesario comprobar que la instalación finalizó sin errores y que, en la interfaz de *Ambari*, todos los servicios aparecen activos, identificados por sus iconos en color verde en la pantalla principal.

3.4. HDFS

HDFS *(Hadoop Distributed File System)* es el sistema de archivos distribuido que utiliza *Hadoop*. Su principal fortaleza es la capacidad de almacenar grandes volúmenes de datos repartidos en múltiples equipos dentro de un *clúster*, algo importante cuando se manejan tamaños que superan los límites de una sola máquina, como terabytes o petabytes. HDFS está inspirado en el ***Google File System (GFS)***, presentado en 2003, que fue diseñado para almacenar grandes datos y permitir su procesamiento eficiente. Sin embargo, HDFS no es ideal para accesos aleatorios o actualizaciones frecuentes; su rendimiento es óptimo cuando se procesan archivos de gran tamaño de manera secuencial [42]. Una de las ventajas clave de ser un sistema distribuido es su escalabilidad: basta con añadir más nodos para aumentar la capacidad. Sin embargo, esta arquitectura también implica riesgos, como la posibilidad de fallos en los nodos. Para mitigar esto, **HDFS implementa redundancia**, almacenando copias de los datos en varios equipos, lo que garantiza la disponibilidad y tolerancia a fallos incluso con hardware económico.

3.4.1. Funcionamiento de HDFS

HDFS es el sistema que permite a *Hadoop* almacenar grandes volúmenes de datos de manera eficiente y tolerante a fallos, dividiendo los archivos en bloques de 128 MB. Estos bloques se distribuyen entre diferentes equipos del *clúster*, lo que permite:

- Superar las limitaciones de almacenamiento de una sola máquina.
- Acceso paralelo a los datos, acelerando las lecturas, algo clave en *frameworks* como *MapReduce.*

ISBN: 978-631-6557-56-8

Para evitar pérdidas de información, cada bloque se replica, normalmente tres veces, en nodos distintos.

Arquitectura: *Namenode y Datanode* **(**HDFS se basa en una arquitectura maestro-esclavo)

- ***Datanode*:** almacena físicamente los bloques de datos y atiende peticiones de lectura o escritura.
- ***Namenode*:** mantiene los metadatos, como la ubicación de los bloques y la estructura de directorios. También detecta fallos y se encarga de reubicar bloques cuando un nodo falla, gracias a un sistema de *heartbeat*.

El *Namenode* es el componente central y, por tanto, un posible punto único de fallo (SPOF). Para evitar esto, existen dos soluciones:

- **Federación:** distribuye la gestión de metadatos entre varios *Namenodes*, facilitando la escalabilidad.
- **Alta disponibilidad:** permite configurar un segundo *Namenode* en espera, listo para asumir el control si la principal falla.

Procesos de Lectura y Escritura

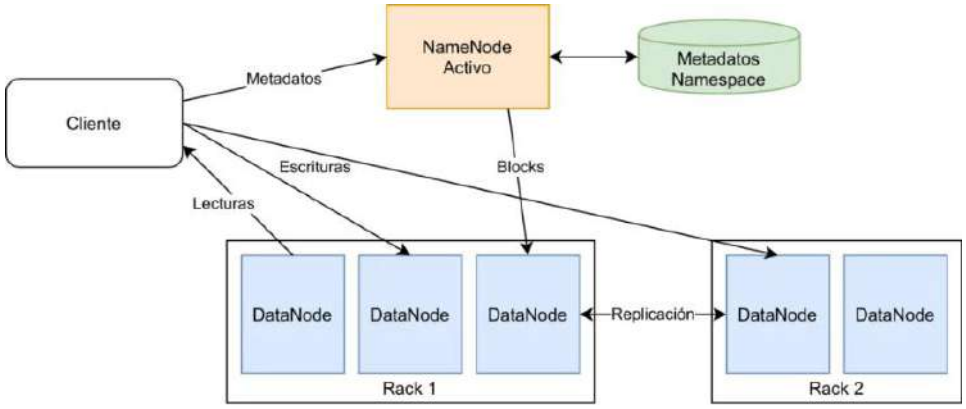
- **Lectura.-** El cliente pide al *Namenode* la ubicación de un archivo; este devuelve la lista de bloques y sus ubicaciones. Luego, el cliente descarga los bloques directamente de los *Datanodes* y reconstruye el archivo.
- **Escritura.-** El cliente solicita permiso al *Namenode* para guardar un archivo. Una vez autorizado, divide el archivo en bloques y los envía a los *Datanodes* seleccionados, asegurándose de que cada bloque se replica correctamente.

Monitorización con Ambari

Hortonworks distribuye *Hadoop* con ***Ambari***, una herramienta que simplifica tanto la instalación como la gestión del clúster, esta acción permite monitorizar el estado de HDFS mediante paneles visuales, facilita la configuración de parámetros como el tamaño de bloque o el número de réplicas y ofrece una interfaz web para explorar la estructura de archivos, revisar logs y descargar contenido. En la Figura 28 se describe el funcionamiento de HDFS en la arquitectura Hadoop.

ISBN: 978-631-6557-56-8

Figura 28: Funcionamiento de HDFS



Adaptado de: [43]

3.5. MapReduce

Es un modelo de programación diseñado por Google para procesar grandes volúmenes de datos de manera eficiente y paralela, especialmente cuando esos datos están distribuidos entre varios equipos, como sucede en sistemas que usan GFS o HDFS visto anteriormente.

El funcionamiento se basa en dos funciones principales:

1. **map()** Toma como entrada un par clave-valor y genera un conjunto de nuevos pares clave-valor intermedios. Luego, todos estos resultados se agrupan por clave, de manera que a cada clave le corresponde una lista de valores asociados.
2. **reduce()** Recibe las claves agrupadas con sus listas de valores y realiza una operación de consolidación o resumen, devolviendo un resultado final para cada clave.

Un ejemplo clásico es el conteo de palabras en documentos. En este caso:

La función **map()** procesa cada línea del texto y puede: Contar las repeticiones de cada palabra en la línea y devolver una tupla por palabra, o devolver una tupla con valor "1" cada vez que encuentra una palabra. Después, el sistema agrupa todas las tuplas por palabra, finalmente, la función **reduce()** suma todos los valores asociados a cada palabra, obteniendo su número total de apariciones.

El proceso es distribuido, cada nodo del *clúster* puede ejecutar varias instancias de **map()** y **reduce().** Generalmente, *Hadoop* procura que las tareas map se ejecuten en el mismo nodo donde están almacenados los bloques de datos, para reducir el tráfico de red. Además, puede usarse una función intermedia llamada ***combiner*** que actúa como un *mini-reduce* en cada nodo local, consolidando datos antes de enviarlos a los *reducers* finales, lo que mejora la eficiencia al reducir la cantidad de información que circula por la red.

ISBN: 978-631-6557-56-8

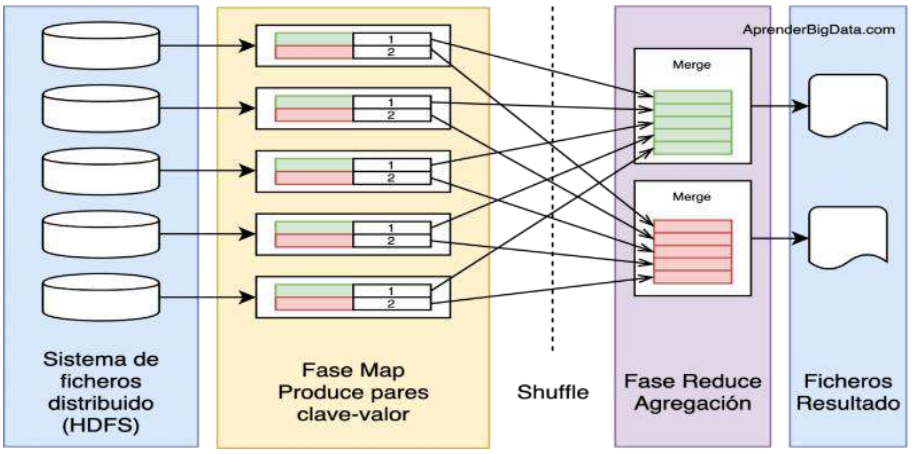
3.5.1. Funcionamiento de MapReduce

En sus primeras versiones, *Hadoop* consideraba cada trabajo *MapReduce* como una unidad de procesamiento, la cual incluía uno o varios archivos de entrada junto con la aplicación encargada de procesarlos. Dentro de este esquema, un nodo del *clúster* asume el rol de ***JobTracker***, cuya función principal es coordinar y supervisar que el trabajo se ejecute correctamente. El *JobTracker* se encarga de dividir el trabajo en múltiples tareas o *tasks*, que pueden ser de tipo *mapper* o *reducer*, dependiendo de si deben ejecutar la función map() o reduce() definida en la aplicación.

Una vez distribuidas, estas tareas son asignadas a diferentes nodos denominados ***TaskTrackers***, que se encargan de ejecutarlas y de informar continuamente al *JobTracker* sobre su estado de progreso. El *JobTracker* también se ocupa de particionar los archivos de entrada, generalmente alineando estas divisiones con los bloques de datos almacenados en HDFS, y asigna un *mapper* para procesar cada partición. En la medida de lo posible, Hadoop intenta que cada *mapper* se ejecute en el mismo nodo que contiene los datos que debe procesar, aplicando así el principio de **localidad de datos**, lo que permite minimizar el tráfico de red. Si alguna tarea falla o un *TaskTracker* deja de responder, el *JobTracker* reasigna la tarea a otro nodo disponible.

Una vez finalizada la etapa de *map*, los resultados se agrupan y se envían a los *reducers*, garantizando que todas las entradas con la misma clave sean procesadas por el mismo *reducer*. En esta fase, es común que haya un intercambio considerable de datos a través de la red, dado que ya no siempre es posible mantener la localidad. Para mitigar este efecto, es posible emplear ***combiners***, que actúan como *reducers* locales, reduciendo el volumen de datos que deben transferirse antes de pasar al proceso final de reducción. Finalmente, herramientas como **Ambari** permiten visualizar en detalle el estado y la evolución de los trabajos *MapReduce* durante su ejecución. A continuación, en la Figura 29 se describe el funcionamiento de MapReduce.

Figura 29: Funcionamiento de MapReduce



Fuente: [44]

ISBN: 978-631-6557-56-8

3.6. Apache Spark

En las secciones anteriores se ha expuesto el ecosistema *Hadoop*, destacando a HDFS como un sistema de almacenamiento distribuido y escalable para grandes volúmenes de datos, y a *MapReduce* como un modelo de procesamiento paralelo que también garantiza escalabilidad.

Ambas tecnologías, ampliamente adoptadas en entornos de producción, ofrecen bases sólidas, aunque es posible optimizar su rendimiento. Por ejemplo, *MapReduce* depende en gran medida de operaciones de lectura y escritura en disco, lo que, si bien es eficiente a gran escala, puede mejorarse mediante el uso de memoria RAM para ciertas transformaciones, reduciendo así el tiempo de procesamiento. Además, la comunicación entre *JobTrackers* y *TaskTrackers* implica una sobrecarga considerable.

En este contexto, ***Apache Spark*** ha ganado relevancia como motor de procesamiento de datos, no como reemplazo de HDFS, sino como una alternativa más flexible y eficiente al modelo *MapReduce*, ya que permite realizar operaciones similares, pero aprovechando el procesamiento en memoria y sin depender del motor *Yarn/MapReduce* de *Hadoop*.

3.6.1. Abstracción: Punto de clave de Spark (RDD)

Cuando hablamos de *Hadoop*, destacamos que una de sus mayores fortalezas es la capacidad de ofrecer abstracciones tanto para usuarios como para desarrolladores. Por ejemplo, en HDFS, aunque los datos se almacenan de forma distribuida, no es necesario preocuparse por su ubicación o por la gestión de fallos, ya que el sistema maneja automáticamente estos aspectos de manera transparente.

De forma similar, en *MapReduce*, el programador solo debe centrarse en definir las funciones de transformación y agregación, mientras que *Hadoop* se encarga de distribuir y ejecutar el procesamiento de manera paralela, incluso recuperando tareas en caso de fallos.

Por su parte, ***Apache Spark*** también mantiene este enfoque de abstracción, y lo simplifica aún más a través de los **RDD (*Resilient Distributed Dataset*)**. Un RDD es, en esencia, una colección distribuida de datos que se comporta de manera similar a una lista, facilitando la manipulación y el procesamiento sin necesidad de preocuparse por la distribución o la tolerancia a fallos.

3.6.2. Un ecosistema completo

Hasta el momento se ha identificado dos grandes ventajas de *Spark*: su velocidad superior frente a *Hadoop MapReduce* y la simplicidad que ofrece para desarrollar transformaciones y acciones sobre los datos. No obstante, *Spark* también destaca por una ventaja adicional: incluye, de manera nativa, diversos módulos que permiten abordar diferentes tipos de análisis sin necesidad de

ISBN: 978-631-6557-56-8

instalar componentes adicionales. Estos módulos cubren desde el procesamiento por lotes hasta el aprendizaje automático.A continuación, se describen brevemente sus principales componentes:

i. ***Spark Core*** es el módulo central de Spark. Proporciona la abstracción fundamental llamada RDD, sobre la que se pueden aplicar diversas operaciones para procesar datos en lotes.
ii. ***Spark Streaming*** facilita el procesamiento de datos en tiempo real mediante un enfoque conocido como *micro-batching*, que divide el flujo de datos en pequeños lotes para procesarlos de forma continua y eficiente, ofreciendo la sensación de inmediatez.
iii. ***Spark* MLlib** es la librería de Spark para machine learning, que permite construir modelos de clasificación, regresión y sistemas de recomendación, entre otros.
iv. ***Spark* SQL** habilita el análisis de datos estructurados a través de consultas SQL, simplificando la manipulación de este tipo de información.
v. ***Spark* GraphX** es el módulo diseñado para ejecutar operaciones sobre grafos de datos distribuidos, facilitando el análisis y modelado de redes complejas.

3.7. Casos de uso en la Industria 4.0

El procesamiento eficiente de grandes volúmenes de datos es uno de los principales desafíos en los entornos de la Industria 4.0. Existen múltiples aplicaciones prácticas de las herramientas y técnicas mencionadas, así como de otras soluciones disponibles en el mercado. La relevancia de esta área dentro del sector industrial puede apreciarse, por ejemplo, en el video "*Open Source Data Management for Industry 4.0*" presentado en [45].

Uno de los casos de uso destacados es el sistema de análisis energético ***SimbleSense***, desarrollado por la empresa *Simble*. Esta plataforma integra *Apache Spark*, que permite gestionar y analizar eficientemente grandes volúmenes de datos generados por sistemas de monitoreo que recolectan información cada poco segundos. *Spark* facilita tareas complejas, como segmentar el consumo energético según la fuente (maquinaria, oficinas, refrigeración, etc.), optimizando las búsquedas, reduciendo tiempos de respuesta y aumentando la precisión de los resultados. Además, su flexibilidad posibilita comparaciones entre diferentes clientes y sectores, aportando valor tanto al usuario como a los modelos de *machine learning* utilizados.

El sector logístico es otro gran beneficiado de estas tecnologías. La adopción de big data y técnicas de procesamiento masivo ha permitido a las empresas optimizar el seguimiento de mercancías, rutas y tiempos de entrega mediante tecnologías como códigos de barras, RFID o NFC. Plataformas basadas en *Hadoop*, por ejemplo, ayudan a los gestores a analizar patrones de transporte y a tomar decisiones que reducen costes, optimizan entregas y mejoran el servicio al cliente.

ISBN: 978-631-6557-56-8

En el campo del ***Business Intelligence***, herramientas como *Hadoop* y *Spark* son ampliamente utilizadas por empresas líderes, como HP, para mejorar tanto sus procesos internos como sus servicios al cliente. Estas soluciones permiten identificar tendencias de consumo, agilizar búsquedas, analizar datos provenientes de redes sociales y recopilar información sobre la experiencia de los usuarios, facilitando así una retroalimentación más rápida y eficaz. Diversos estudios respaldan los beneficios de estas tecnologías en la toma de decisiones estratégicas, siendo un claro [46], donde destaca su impacto en la creación de valor para el cliente.

Por último, el uso de sistemas basados en big data también se extiende al sector agrícola, especialmente en entornos **IoT**. Un ejemplo es la empresa Hort@, que emplea estas tecnologías para detectar enfermedades en cultivos a gran escala, recolectando datos de diversas regiones, correlacionándolos y filtrándolos para optimizar sus algoritmos y mejorar las recomendaciones a los agricultores.

En el ámbito de la fabricación en la nube, ***MapReduce*** también se aplica al diagnóstico automático de fallos en procesos productivos. Según [47], este enfoque permite reconocer patrones en sistemas con desequilibrio de datos, reduciendo costes de pruebas y mejorando la calidad final del producto, al apoyar la clasificación correcta tras aplicar modelos de *machine learning*.

3.8. Inteligencia artificial y aprendizaje automático

El propósito fundamental de la inteligencia artificial (IA), desde la perspectiva científica, es comprender los principios que sustentan el comportamiento inteligente en sistemas artificiales. Para ello, se analizan tanto agentes naturales como artificiales, se plantean hipótesis sobre cómo desarrollar sistemas capaces de realizar tareas que demandan inteligencia, y se prueban experimentalmente dichas hipótesis a través del diseño y construcción de sistemas inteligentes [48]. En el presente tema, la IA no se aborda desde un enfoque teórico o científico, sino desde una perspectiva práctica aplicada a la ingeniería, con especial énfasis en su implementación en entornos de industria 4.0. En este contexto, es importante clarificar tres conceptos básicos:

> **Inteligencia artificial:** área de la informática enfocada en la creación de agentes computacionales que, mediante estímulos externos y conocimientos previamente adquiridos ya sea a través del aprendizaje automático o de expertos humanos—, generan acciones que maximizan su rendimiento.
>
> **Minería de datos:** proceso que emplea técnicas de IA para analizar grandes volúmenes de datos y detectar patrones útiles que permitan obtener beneficios concretos.

ISBN: 978-631-6557-56-8

> **Aprendizaje automático (Machine Learning):** rama de la IA dedicada al desarrollo de programas que, a partir de la experiencia, optimizan su desempeño en tareas específicas [49].

El aprendizaje automático es un componente clave en la minería de datos, utilizado, por ejemplo, en la detección de fraudes en transacciones bancarias, y también se aplica en ámbitos como la robótica o el reconocimiento de voz. La IA, en general, tiene aplicaciones amplias en campos como la robótica, el entretenimiento, el marketing, la medicina o la predicción del clima. Entre los problemas que puede abordar la IA se encuentran: diagnóstico de fallos y propuesta de soluciones; selección de alternativas óptimas; predicción de comportamientos futuros; clasificación y agrupamiento de objetos según sus características; optimización de soluciones, y control de sistemas en tiempo real.

La industria, en especial la denominada industria 4.0, se beneficia ampliamente de la IA mediante soluciones de fabricación inteligente, análisis avanzado de datos empresariales y robótica, entre otros [50]. En el ámbito empresarial, las técnicas de IA permiten una mejor gestión de recursos y apoyan la toma de decisiones, como en el caso de la evaluación de la solvencia financiera [51]. Asimismo, los sistemas de recomendación utilizados por plataformas de comercio electrónico y proveedores de contenido digital, como Amazon, Netflix o Spotify, se basan en IA para anticipar las preferencias de los usuarios. En el ámbito de redes y sistemas, la IA también se aplica para detectar y corregir fallos, tanto en redes informáticas como en los propios equipos. Además, sistemas inteligentes son empleados en la planificación de rutas óptimas, ya sea para el transporte de mercancías o para la transmisión eficiente de datos, en función de criterios como el costo o el tiempo. En cuanto al ahorro energético, la IA se utiliza para optimizar el consumo, tanto en edificios como en hogares, mediante sistemas de localización y sensorización inteligente [52]. También permite a las empresas eléctricas predecir la demanda y planificar sus operaciones, incluyendo la integración de energías renovables y la programación de mantenimientos.

Por otro lado, la IA se ha utilizado en meteorología, en la gestión de catástrofes y en el sector agrícola, facilitando la planificación de cultivos, el control de plagas y la gestión eficiente de recursos. Un ejemplo concreto de esta aplicación es el uso de soluciones IoT en agricultura y ganadería para monitorizar y optimizar la producción. En conclusión, los casos mencionados demuestran cómo la inteligencia artificial ha logrado integrarse con éxito en una gran diversidad de sectores, aportando soluciones que optimizan procesos y potencian la toma de decisiones.

3.8.1. Tipos de aprendizaje

El aprendizaje de conceptos en inteligencia artificial suele abordarse como una búsqueda dentro de un espacio de hipótesis posibles, con el propósito de identificar aquella que mejor se ajuste a

ISBN: 978-631-6557-56-8

los datos de entrenamiento. Este enfoque inductivo garantiza que la hipótesis seleccionada es adecuada para los datos conocidos, aunque se asume que también funcionará con nuevos casos, siempre que el conjunto de entrenamiento sea suficientemente representativo. En este contexto, el aprendizaje de conceptos se clasifica en tres categorías principales:

a. **Aprendizaje supervisado.-** Consiste en aprender a identificar o predecir clases o valores numéricos a partir de datos etiquetados, es decir, datos previamente clasificados por un experto. El objetivo es que el modelo pueda asignar correctamente una clase a datos no etiquetados en el futuro.
b. **Aprendizaje no supervisado.-** Se utiliza para descubrir patrones o estructuras ocultas en conjuntos de datos que no tienen etiquetas. Aquí, el modelo intenta definir nuevas categorías o relaciones sin conocimiento previo de las clases.
c. **Aprendizaje por refuerzo.-** En este tipo de aprendizaje, un agente interactúa con su entorno tomando decisiones secuenciales. Mediante un proceso de prueba y error, el agente ajusta sus acciones en función de recompensas o penalizaciones, buscando maximizar su rendimiento final. A diferencia de los métodos anteriores, puede no requerir datos de entrenamiento previos.

El desarrollo de un modelo de aprendizaje generalmente sigue estas etapas: definir el objetivo, seleccionar los datos de entrenamiento, establecer una función objetivo y su representación, elegir un algoritmo de aprendizaje adecuado y finalmente evaluar y validar los resultados. En las secciones siguientes se describirán, a nivel general, algunas de las técnicas de inteligencia artificial más utilizadas, acompañadas de ejemplos prácticos que ilustran sus aplicaciones.

3.9. Árboles de decisión y reglas

Dos de las técnicas más utilizadas para la representación del conocimiento son los **árboles de decisión** y las **reglas de asociación**. Ambas se emplean en la resolución de problemas de **aprendizaje supervisado**. En esta sección se presentarán los conceptos fundamentales para su comprensión y se ofrecerá una breve descripción de los algoritmos más representativos que implementan estas técnicas. El aprendizaje supervisado tiene como objetivo principal identificar o definir un concepto a partir de ejemplos específicos. En el caso de la **clasificación**, este proceso suele dividirse en dos fases:

a. **Construcción del modelo.-** Consiste en generar un modelo descriptivo que represente el concepto, utilizando un conjunto de datos que incluye tanto instancias que pertenecen a la clase como otras que no.

ISBN: 978-631-6557-56-8

b. **Clasificación de nuevas instancias.-** Una vez definido el modelo, se utiliza para determinar si una nueva instancia puede ser clasificada como perteneciente a la clase aprendida.

A continuación, se enlistan algunos de los árboles de decisión y reglas

a) Árboles de decisión.- Es una de las técnicas más populares dentro del aprendizaje inductivo, ya que se considera un método robusto frente a la presencia de datos ruidosos. Generalmente, tanto las entradas como las salidas de la función objetivo son valores discretos, aunque en ciertos casos las entradas pueden tomar valores continuos. Esta técnica representa la función objetivo en forma de árbol, donde cada ruta desde la raíz hasta una hoja puede interpretarse como una secuencia de condiciones, lo que facilita su conversión en un conjunto de reglas claras. Cuando se necesita clasificar una instancia cuya clase aún es desconocida, los valores de sus atributos se utilizan para recorrer las distintas ramas del árbol, desde el nodo raíz hasta llegar a una hoja, donde se asigna la clase correspondiente. El uso de árboles de decisión se considera adecuado cuando se cumplen los siguientes criterios según [53]. Los **árboles de decisión** han demostrado ser herramientas eficaces en problemas reales, especialmente en tareas de clasificación, donde se requiere asignar ejemplos a categorías definidas. En el contexto de la **Industria 4.0**, se utilizan en áreas como diagnóstico de fallos, decisiones en sistemas autónomos o segmentación de clientes para optimizar estrategias, como el ahorro energético. El proceso de aprendizaje mediante árboles de decisión consiste en explorar todas las posibles estructuras de árbol hasta encontrar aquella que mejor se ajuste a los datos de entrenamiento ya etiquetados. Para cada clase, se busca que una combinación específica de atributos, representada en alguna de las ramas, permita clasificar correctamente las instancias. Un aspecto importante en la construcción de estos árboles es la **selección de atributos**, que define cómo se dividen las ramas en cada nivel. La elección depende del tipo de dato: Si el atributo es **discreto**, se crean ramas para cada uno de sus valores posibles, si es **numérico**, se establece un umbral que divide el conjunto, evaluando si el valor es mayor o menor que ese punto.
b) Algoritmo básico de aprendizaje de árboles de decisión (ID3).- El algoritmo **ID3** genera árboles de decisión siguiendo un enfoque descendente, es decir, desde la raíz hacia las hojas. Para decidir qué atributo usar en cada división, emplea un criterio basado en la **teoría de la información**. Según esta heurística, se selecciona el atributo que proporciona la mayor cantidad de información para mejorar la clasificación.
c) Sobreajuste y poda: algoritmo C4.5.- Uno de los inconvenientes más comunes al usar árboles de decisión es el **sobreajuste**. Este problema ocurre cuando el modelo se ajusta

ISBN: 978-631-6557-56-8

demasiado a los datos de entrenamiento, perdiendo capacidad de generalización frente a nuevas instancias, incluso si existen hipótesis alternativas que representarían mejor el conjunto completo de datos. Para reducir el sobreajuste, suelen aplicarse dos enfoques principales:

Prepoda.- consiste en limitar el crecimiento del árbol durante su construcción, evitando generar ramas innecesarias. Esto reduce el tiempo de procesamiento y mejora la simplicidad del modelo.

Pospoda.- se realiza una vez generado el árbol completo, eliminando ramas que no aportan mejoras significativas en la clasificación. En algunos casos, se combinan atributos antes de decidir qué partes podar, ya que algunas combinaciones pueden ser más informativas que atributos individuales. Un ejemplo de algoritmo que aplica pospoda es **C4.5**, una extensión del ID3. C4.5 trabaja tanto con atributos discretos como continuos, gestionando valores ausentes y aplicando la poda para mejorar la precisión final. Además, permite transformar el árbol en un conjunto de reglas para facilitar una poda más precisa: Construir el árbol a partir de los datos de entrenamiento. Convertir cada camino del árbol en una regla (condiciones unidas por AND). Simplificar cada regla eliminando condiciones innecesarias, siempre que la precisión aumente. Ordenar las reglas según su precisión y aplicarlas en ese orden al clasificar nuevas instancias.

Finalmente, para determinar el tamaño óptimo del árbol y su capacidad de generalización, se recomienda utilizar **validación cruzada**. Esta técnica consiste en dividir los datos en dos conjuntos: uno para entrenar el modelo y otro para validarlo. Cuando el volumen de datos es limitado, se aplica la variante ***K-fold cross-validation***, que distribuye los datos en *k* particiones, utilizando cada partición como conjunto de prueba en diferentes iteraciones, mientras el resto se usa para entrenar.

d) *Ensemble learning* y *random forest.*- El aprendizaje ensemble, o aprendizaje integrado, es una rama del *machine learning* que busca mejorar la precisión de los modelos combinando varios algoritmos simples, de modo que trabajen en conjunto y corrijan mutuamente sus errores. Aunque las redes neuronales son populares, los métodos ensemble como *random forest, boosting y bagging* son muy utilizados en aplicaciones reales por su alto rendimiento.

Este enfoque funciona especialmente bien cuando los algoritmos individuales son inestables, es decir, cuando pequeñas variaciones en los datos generan resultados diferentes. Por ello, modelos como árboles de decisión o regresión suelen ser ideales, mientras que algoritmos más estables, como *Naïve Bayes o K-Nearest Neighbors*, se usan con menos frecuencia en este contexto.

Los principales métodos ensemble son:

ISBN: 978-631-6557-56-8

i. ***Stacking (apilamiento).-*** Consiste en entrenar diferentes modelos con los mismos datos y combinar sus predicciones mediante votación o regresión, buscando una salida más precisa. Un ejemplo práctico es el método FWLS, usado en sistemas de recomendación.

ii. ***Bagging (Bootstrap Aggregating):*** Aquí se entrenan varios modelos (por lo general, del mismo tipo, como árboles de decisión) en subconjuntos aleatorios de datos. Luego, sus predicciones se combinan para obtener un resultado final. El algoritmo *Random Forest* es un claro ejemplo de este enfoque.

iii. ***Boosting:*** A diferencia de *stacking y bagging*, en *boosting* los modelos se entrenan de forma secuencial, ajustando el siguiente modelo en función de los errores del anterior. Aunque requiere más tiempo de cómputo, suele lograr una gran precisión en tareas de clasificación. Ejemplos destacados son *AdaBoost, CatBoost, LightGBM y XGBoost*, utilizados incluso en sistemas de búsqueda de Google y redes sociales.

- Reglas de clasificación y reglas de asociación.- Las **reglas** son una forma de representar conocimiento en inteligencia artificial. Cada regla se compone de dos partes: **Antecedente**: las condiciones que deben cumplirse y c**onsecuente**: la acción o conclusión que se sigue si las condiciones son verdaderas. Su estructura básica es: **SI** (condiciones) **ENTONCES** (conclusión).

  Las reglas pueden combinar múltiples condiciones usando operadores lógicos como **AND** y **OR**, aunque es recomendable no mezclarlos en la misma regla para mantener claridad. Estas reglas son una alternativa a los árboles de decisión, ya que ambos métodos se pueden transformar uno en otro: cada camino de un árbol equivale a una regla, y cada hoja del árbol corresponde a la conclusión de la regla. Existen dos grandes tipos de reglas: **Reglas de clasificación,** asignan una clase a una instancia y **Reglas de asociación,** identifican combinaciones frecuentes de atributos y valores en los datos.

  En el caso de las reglas de asociación, se utilizan medidas para evaluar su calidad. **Confianza,** probabilidad de que la conclusión sea cierta si las condiciones se cumplen y **Soporte**, frecuencia con la que ocurre una regla en el conjunto de datos. Una regla debe tener tanto confianza como soportes altos para considerarse relevante. El proceso para generar estas reglas suele dividirse en dos fases: i) Seleccionar reglas que superen un mínimo de soporte. ii) Filtrar las que también superen un umbral de confianza.

  En cuanto a los **algoritmos de aprendizaje de reglas**, uno de los más conocidos es ***PRISM***, que emplea un enfoque de recubrimiento secuencial. Este método: i) Aprende una regla que

ISBN: 978-631-6557-56-8

cubra algunos ejemplos de una clase. ii) Elimina esos ejemplos del conjunto. iii) Repite el proceso hasta cubrir todos los casos.

PRISM selecciona las reglas con base en su **precisión (confianza)**, garantizando que cada nueva regla sea lo más fiable posible.

e) Algoritmos de aprendizaje de reglas de asociación.- Existen diferentes algoritmos diseñados para extraer reglas de asociación, siendo **Apriori** uno de los más reconocidos [54]. Este método busca identificar conjuntos de elementos, llamados ***ítem-sets***, que superen un umbral mínimo de cobertura de manera eficiente. Un **ítem** representa un par atributo-valor, mientras que un ***ítem-set*** es un conjunto de estos pares. La **cobertura** se refiere a la cantidad de datos que cumplen con todas las condiciones del *ítem-set*, y esta medida es clave para determinar la validez de las reglas que se generen.

   El algoritmo **Apriori** se ejecuta en dos etapas: i) Identifica todos los ítem-sets que cumplen con el umbral mínimo de cobertura. ii) A partir de esos ítem-sets, genera reglas de asociación.

   Es importante señalar que, en conjuntos de datos muy grandes, la eficiencia del algoritmo puede verse afectada, ya que la cantidad de combinaciones crece significativamente en función del umbral de cobertura establecido.

3.10. Redes neuronales artificiales

Las **redes neuronales artificiales** son una de las principales técnicas del aprendizaje automático, ya que permiten que un sistema mejore su desempeño en una tarea a través de la experiencia. Inspiradas en la forma en que aprende el cerebro humano, estas redes pueden abordar problemas complejos e incluso superar las capacidades humanas en tareas específicas, como el reconocimiento de patrones. Son especialmente útiles cuando se trabaja con datos que poseen muchos atributos y cuando la salida puede tomar valores diversos, ya sea reales, discretos o una combinación de ambos. Aunque el proceso de **entrenamiento** suele ser largo, una vez aprendida la función, su aplicación a nuevos casos es rápida. Además, estas redes son resistentes al ruido, lo que las hace eficaces incluso cuando los datos de entrenamiento contienen errores.

Las **neuronas artificiales** imitan a las biológicas: reciben múltiples entradas (como las dendritas) y producen una única salida (como el axón). Las conexiones entre neuronas tienen **pesos** que determinan la importancia de cada entrada, y el aprendizaje consiste en ajustar esos pesos. Una red neuronal básica está formada por una capa de entrada, una o varias capas intermedias (ocultas) y una capa de salida. El diseño de la arquitectura de la red número de capas, número de neuronas y tipo de funciones de activación es un paso fundamental en su implementación.

ISBN: 978-631-6557-56-8

Las **funciones de activación** son las encargadas de calcular la salida de cada neurona. Existen distintas opciones, como: la función escalón, la función lineal, la función sigmoide, tangente hiperbólica, ReLU, Leaky ReLU y Softmax, cada una con características específicas según la tarea a realizar.

3.10.1. Tipos de redes neuronales

Existen diversos tipos de redes neuronales, siendo algunas de las más relevantes las siguientes:

i. **Redes neuronales multicapa.-** Estas redes, también conocidas como redes de alimentación hacia adelante (*feedforward*), permiten que la información fluya desde la entrada hasta la salida a través de las diferentes capas, sin retrocesos. Su estructura incluye al menos una capa oculta.

ii. **Redes con retropropagación**.- Su funcionamiento se basa en el ajuste de pesos mediante el método de **retropropagación del error** (*backpropagation*). En estas redes, cada neurona de una capa está conectada con todas las neuronas de las capas adyacentes, tanto anteriores como posteriores.

Estas redes tienen aplicaciones en una gran variedad de campos, entre ellos:

- Reconocimiento de escritura y de voz.
- Conducción de vehículos autónomos.
- Sistemas de guiado automático.
- Toma de decisiones en sistemas robóticos.
- Diseño de ecualizadores y sistemas de cancelación de eco adaptativos.
- Sistemas de localización, tanto en espacios interiores —para identificar la posición de un dispositivo móvil— como en servicios de mapas en línea, optimizando la predicción de ubicaciones y el uso eficiente de cachés.
- Gestión de la energía mediante sistemas de demanda-respuesta.

3.11. *Deep Learning*

Aunque las redes neuronales de retropropagación han demostrado ser eficaces en la resolución de diversos problemas, presentan limitaciones cuando se enfrentan a tareas complejas que requieren muchas capas y nodos intermedios. A medida que aumenta el número de nodos, también crecen las conexiones y los cálculos de pesos, lo que vuelve el entrenamiento lento y costoso. Para superar estas limitaciones surge el **aprendizaje profundo** o **deep learning**, una rama del aprendizaje automático que utiliza redes neuronales con muchas capas y nodos. Estas redes

ISBN: 978-631-6557-56-8

permiten abordar problemas complejos que implican grandes volúmenes de datos y múltiples niveles de abstracción. En ellas, las capas inferiores detectan características simples, mientras que las superiores combinan esas características para representar conceptos más complejos.

Existen diversas arquitecturas dentro del *deep learning*, entre las que destacan:

- Redes prealimentadas (*Feedforward*) y sus variantes profundas.
- Redes neuronales recurrentes (RNN) y sus versiones profundas.
- Autoencoders (AE).
- Redes convolucionales (CNN).
- Redes generativas antagónicas (GAN).
- *Deep Belief Networks* (DBN) y *Variational Autoencoders* (VAE).

Estas redes permiten resolver problemas tanto de aprendizaje supervisado como no supervisado, ofreciendo mayor precisión que los métodos tradicionales, aunque con un mayor costo computacional. Su aplicación se extiende a áreas como reconocimiento de imágenes y videos, síntesis y reconocimiento de voz, análisis de sentimientos, procesamiento de lenguaje natural, traducción automática, transferencia de estilos en imágenes, entre muchas otras.

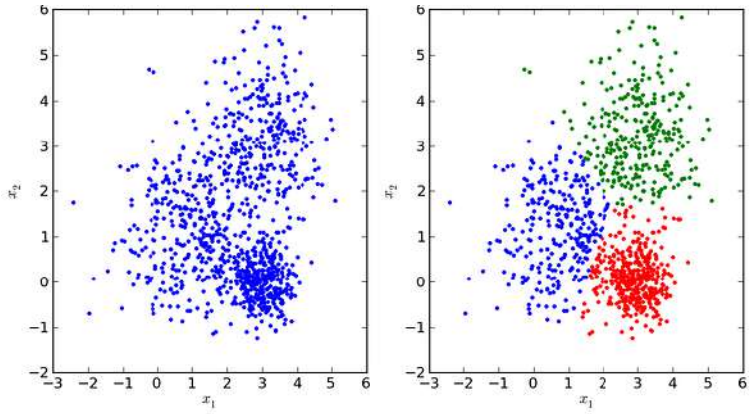
3.12. *Clustering*

Es un método de aprendizaje no supervisado que agrupa objetos en *clústeres* basados en su similitud. Su objetivo es descubrir clases desconocidas a partir de instancias, sin que estas clases estén previamente definidas. Un *clúster* es un conjunto de objetos similares entre sí y diferentes de los objetos de otros *clústeres*. En el aprendizaje no supervisado, no se especifican clases, y el objetivo es identificar nuevos conceptos a través de la agrupación de instancias según un esquema de similitud. A diferencia del aprendizaje supervisado, donde las clases están definidas, el *clustering* agrupa las instancias sin una guía preestablecida.

Las técnicas de *clustering* son útiles para el análisis de datos, ya que permiten descubrir patrones y propiedades comunes entre instancias. Además, estos métodos pueden ser utilizados para generalizar y encontrar instancias similares, y luego incorporar nuevas instancias en los agrupamientos existentes. Sin embargo, la calidad del *clustering* depende de la aplicación y los objetivos específicos del problema, por lo que el usuario debe evaluar qué agrupamiento es más útil según sus necesidades. Los diferentes algoritmos de *clustering* pueden generar agrupamientos distintos, por lo que la elección del algoritmo dependerá del objetivo de cada caso. En la Figura 30 se muestra un esquema básico de *clustering*.

ISBN: 978-631-6557-56-8

Figura 30: Ilustración básica del concepto de clustering.



Fuente: [55]

3.13. Sistemas de recomendación

La abundante información disponible en Internet puede resultar abrumadora para los usuarios, que deben tomar decisiones entre numerosas opciones. Por ello, las técnicas que permiten clasificar, ordenar y filtrar la información son importantes. Estas técnicas se basan en modelar contenidos y preferencias de usuarios, creando estereotipos y filtrando información según estos.

En [56] se describe los componentes clave de los sistemas de recomendación. El primer paso es definir cómo representar el perfil del usuario, considerando datos como el historial de compras o navegación para deducir preferencias y comportamientos. Luego, se aplica un tipo de filtrado para hacer recomendaciones:

i. **Filtrado demográfico**.- Utiliza descripciones de personas para relacionar los ítems con ciertos tipos de usuarios.
ii. **Filtrado colaborativo**.- Se basa en la retroalimentación explícita o implícita de los usuarios, como el historial de compras.
iii. **Filtrado basado en contenidos**.- Utiliza la metainformación de los ítems para determinar la relación entre un usuario y esos contenidos.

Es necesario elegir una técnica de filtrado adecuada y combinarla con una técnica de emparejamiento, como entre estereotipos y contenidos, perfiles de usuario y contenido, o entre perfiles de usuarios. Además, se pueden emplear técnicas de inteligencia artificial, como redes neuronales o árboles de decisión, y *clustering* para generar perfiles de usuario y asociarlos a grupos similares.

A medida que los usuarios interactúan, sus intereses cambian, por lo que es importante actualizar sus perfiles mediante retroalimentación explícita o implícita (por ejemplo, historial de navegación). Aunque los sistemas de filtrado demográfico pueden ser menos personalizados, el

ISBN: 978-631-6557-56-8

uso de redes sociales y datos colaborativos mejora la adaptabilidad y precisión de las recomendaciones.

3.14. Búsqueda

Existen problemas de inteligencia artificial que se pueden resolver mediante búsqueda, donde se define un espacio de estados y operadores que facilitan las transiciones entre estados, partiendo de un estado inicial y buscando uno objetivo. En estos problemas, desde un estado inicial, se trata de encontrar el camino hacia el estado objetivo. En cada estado, es posible aplicar una serie de operaciones para llegar a otro estado, que puede ser el final o no. Estas técnicas son ampliamente utilizadas en la programación de robots y autómatas en el contexto de la industria 4.0. Para resolver problemas mediante búsqueda, es necesario aplicar una estrategia de control que permita encontrar un camino desde el estado inicial hasta el objetivo. Esto implica analizar posibles secuencias de acciones y los estados que generan, seleccionando la secuencia que sea más adecuada según un criterio específico.

3.15. Sistemas expertos

Un sistema experto es un software diseñado para actuar de manera similar a un experto humano al abordar problemas dentro de un área de conocimiento específica. Estos sistemas serán capaces de:

i. Procesar conocimientos expresados en forma de reglas y razonar para resolver problemas en un campo determinado.
ii. Proporcionar conocimiento de manera comprensible, utilizando lenguaje natural.
iii. Explicar el proceso mediante el cual llega a sus conclusiones, rastreando las reglas empleadas en el razonamiento.
iv. Generar nuevo conocimiento al añadir nuevas reglas o modificar las existentes.

Al igual que un experto humano, un sistema experto podría cometer errores si los datos no son completos o ser capaz de razonar en situaciones de incertidumbre cuando los datos son incompletos. La figura 18 ilustra la arquitectura de un sistema experto y el perfil de las personas involucradas en su desarrollo.

3.16. Inteligencia artificial en industria 4.0

La inteligencia artificial (IA) se ha consolidado como una herramienta fundamental en múltiples sectores, incluyendo la robótica, los videojuegos, el marketing, la medicina y la predicción del clima. En el contexto de la Industria 4.0, la IA permite abordar diversos tipos de problemas mediante soluciones inteligentes. Algunos de sus usos más representativos son:

ISBN: 978-631-6557-56-8

i. **Diagnóstico**: identificar fallos en dispositivos o máquinas, analizando su comportamiento y recomendando soluciones.
ii. **Selección**: elegir el proveedor más adecuado basándose en experiencias previas y en la relación calidad-precio.
iii. **Predicción**: anticipar fallos en sistemas o equipos.
iv. **Clasificación**: ordenar imágenes o identificar estados de equipos mediante datos visuales, como imágenes termográficas.
v. **Agrupamiento**: segmentar clientes según patrones de comportamiento, facilitando la definición de mercados.
vi. **Optimización**: aplicar algoritmos, como los neuroevolutivos, para mejorar procesos industriales, por ejemplo, el paletizado en almacenes.
vii. **Control**: emplear algoritmos de aprendizaje por refuerzo para el manejo autónomo de robots en entornos logísticos.

Además, en un mundo cada vez más orientado a los datos, las empresas que logran extraer valor de grandes volúmenes de información obtienen ventajas competitivas notables. La **minería de datos**, impulsada por técnicas de IA, se ha convertido en un área clave, especialmente en sectores que buscan automatizar la toma de decisiones y optimizar procesos. Firmas como **Accenture** han destacado la importancia estratégica de la inteligencia artificial en el futuro de los negocios, subrayando su papel en la transformación digital de la industria.

Existen numerosos ejemplos que demuestran la efectividad de los sistemas inteligentes en diferentes áreas de la Industria 4.0. Un uso muy común es en **diagnóstico y mantenimiento predictivo**, donde estos sistemas permiten identificar fallos, prevenir problemas y proponer soluciones. Un caso representativo es el de ***Linde***, que con sus plataformas *Optima* y *Connect* ofrece gestión de flotas eficiente, control de uso, seguridad y reducción de daños mediante análisis de datos en tiempo real. En **logística y transporte**, la inteligencia artificial se utiliza para calcular rutas óptimas, minimizando tiempos o costes, como es el caso de ***Google Maps***, que ajusta sus recomendaciones en función del tráfico y otros factores.

Dentro del ámbito **energético**, la IA permite optimizar el consumo, predecir la demanda y diseñar estrategias que ayuden tanto a las empresas como a los usuarios a reducir gastos y mejorar la eficiencia. El grupo de investigación **GECAD**, del Instituto Politécnico de Oporto, es un referente en este campo. En **meteorología y gestión de desastres**, así como en **agricultura**, las soluciones basadas en IA contribuyen en la predicción, planificación de cultivos, control de plagas y gestión eficiente de la producción. Un ejemplo es la empresa **Horta**, que ha desarrollado sistemas que ayudan a los agricultores a mantener sus certificaciones ecológicas, y el **Instituto Nacional de**

ISBN: 978-631-6557-56-8

Tecnología Agropecuaria de Argentina, que ha implementado robots capaces de realizar tareas agrícolas como medición y fertilización dentro de invernaderos.

Por otro lado, la **robótica** ha sido una de las áreas con mayor desarrollo gracias a la inteligencia artificial. Desde robots de soldadura y ensamblaje en la industria automotriz hasta robots colaborativos como ***YuMi*** de ABB y ***NEXTAGE*** de *Rollmatic*, estas tecnologías permiten automatizar tareas complejas y mejorar la productividad en fábricas. Finalmente, estudios como los de ***Accenture*** demuestran el crecimiento acelerado de la automatización, destacando la instalación de más de 22,000 robots en entornos industriales, lo que subraya el papel clave de la IA en la transformación de la industria [57].

En los siguientes códigos QR o enlaces se muestra información adicional relacionada al capítulo.

Introducción a MapReduce	Conciencia Artificial y Test de Turing
https://www.youtube.com/watch?v=ht3dNvdNDzI	https://www.youtube.com/watch?v=Y90a6Xwwh3w

ISBN: 978-631-6557-56-8

ISBN: 978-631-6557-56-8

CAPÍTULO IV

VISUALIZACIÓN E INTELIGENCIA EMPRESARIAL (BI)

4.1. Introducción y Objetivo del Capítulo

En la era de la Industria 4.0, el análisis y la visualización de datos se han convertido en herramientas fundamentales para la toma de decisiones estratégicas. La representación gráfica de grandes volúmenes de información permite una comprensión más rápida y efectiva, facilitando la asimilación de conceptos complejos y el descubrimiento de patrones relevantes. No obstante, la creación de infografías, gráficas y otras visualizaciones requiere un conocimiento sólido de los datos y de las herramientas disponibles para garantizar resultados eficientes y significativos.

Este capítulo aborda de forma integral los conceptos de visualización de datos e inteligencia empresarial (*business intelligence* o BI), presentando tanto las bases teóricas como las aplicaciones prácticas. Se explorarán los pasos necesarios para desarrollar una visualización, desde la concepción de la idea hasta su presentación final al público, así como los distintos tipos de gráficos y su clasificación entre figurativos y no figurativos. Asimismo, se explicarán las diferencias entre visualizaciones estáticas, dinámicas e interactivas, destacando las principales herramientas de código abierto y comerciales utilizadas en su elaboración.

Paralelamente, se ofrecerá una visión ampliada del término inteligencia empresarial, más allá de su asociación con softwares específicos, para incluir aspectos clave del análisis empresarial y su rol en el soporte de decisiones en diversas áreas organizacionales. Se introducirán herramientas como el Cuadro de Mando Integral (CMI) como modelo de gestión estratégica y se ilustrará cómo los datos, junto con un enfoque adecuado de BI, pueden potenciar la eficiencia y competitividad de una empresa. Por lo tante en este capítulo se establen alcanzar los siguientes objetivos:

- **Analizar el proceso de tratamiento y visualización de datos**, comprendiendo sus diferentes fases, los tipos de gráficos más utilizados y las características de las visualizaciones estáticas, dinámicas e interactivas.
- **Conocer las tecnologías, lenguajes y herramientas empleadas en la visualización de datos**, tanto en entornos web como en ciencia de datos e inteligencia empresarial, especialmente en contextos industriales.
- **Explorar casos de uso reales de visualización de datos en la Industria 4.0**, destacando su impacto en la eficiencia operativa y la toma de decisiones.
- **Ampliar el concepto de inteligencia empresarial**, vinculándolo con procesos, metodologías y herramientas que permiten transformar datos en conocimiento útil para la gestión organizacional.

ISBN: 978-631-6557-56-8

- **Entender la relación entre datos, estrategia empresarial y toma de decisiones**, mediante el estudio del cuadro de mando integral y otros modelos de gestión estratégica apoyados por sistemas de BI.

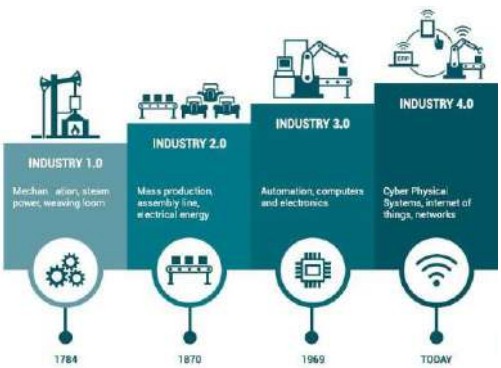
4.2. Introducción a la visualización de datos

Visualizar datos es de gran importancia porque el ser humano posee una gran capacidad para detectar patrones visuales de forma rápida, lo que facilita la comprensión y generación de conocimiento. Ante el crecimiento exponencial de la información, se vuelve elemental aprender a representarla gráficamente para facilitar su interpretación. La visualización es el resultado de un proceso que comienza con la búsqueda del mensaje oculto en los datos. En el contexto de la Industria 4.0 y el big data, estas representaciones permiten explicar conceptos, sintetizar grandes volúmenes de información y apoyar la toma de decisiones, como ocurre en mapas, gráficas energéticas, infografías o *dashboards* de sistemas de inteligencia empresarial.

4.2.1. Infografía y visualización de datos

Los términos infografía y visualización de datos suelen tratarse como conceptos distintos, aunque en la práctica pueden combinarse en una misma representación con un propósito compartido. Las infografías, con una larga historia, utilizan una mezcla de gráficos como diagramas, ilustraciones o mapas y texto para comunicar información. En cambio, la visualización de datos, más común en ámbitos científicos, se refiere a representaciones gráficas generadas por medios digitales, muchas veces interactivas, que permiten explorar datos abstractos y obtener una mejor comprensión de los mismos. A continuación, se muestran en las Figuras 31 y 32 ejemplos de infografías y visualización de datos.

Figura 32: Ejemplo de infografía



Adaptado de: [58]

Figura 31: Ejemplo de visualización de datos

Fuente: [59]

ISBN: 978-631-6557-56-8

4.2.2. Importancia de la infografía y la visualización de datos

Visualizar los datos permite al público identificar con rapidez las relaciones entre ellos. Uno de los principales aportes de este recurso es la capacidad de transmitir gran cantidad de información en apenas unos instantes, ya que se basa en una forma de comunicación que nuestro sistema cognitivo comprende de manera natural. Un ejemplo de su aplicación se encuentra en el análisis del consumo energético de las distintas sedes de una empresa internacional. Los datos en bruto, organizados en extensas tablas, pueden resultar difíciles de interpretar. Sin embargo, al representarlos en gráficos de barras, es más sencillo observar comparaciones y tendencias. También puede diseñarse una infografía que muestre un mapa con las ubicaciones de las sedes, utilizando círculos de tamaño proporcional al consumo energético de cada una.

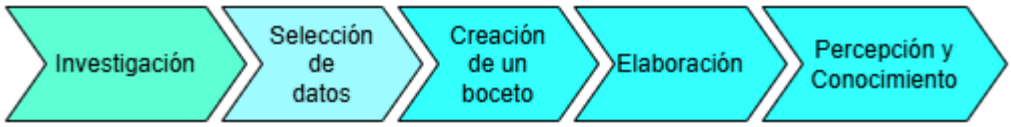
4.2.3. Estadios de la visualización

El proceso de crear una visualización de datos suele seguir cinco etapas fundamentales, aunque estas pueden adaptarse o recorrerse de forma no lineal dependiendo del proyecto:

i. **Investigación**: Comienza con la definición del tema y la recopilación de información confiable. Se deben identificar fuentes de datos precisas, muchas veces disponibles en línea, y contrastarlas para asegurar su veracidad.
ii. **Selección de datos**: Una vez reunida la información, se elige aquella más relevante y se organiza de forma que sea fácil de manipular y representar gráficamente.
iii. **Creación de un boceto**: Se desarrolla un esquema preliminar que actúa como una guía visual inicial. Este diseño básico, con una estructura simple, permite visualizar cómo se presentará la información antes de producir la versión final.
iv. **Elaboración**: En esta fase se transforman los datos seleccionados en una representación visual coherente y clara. Se eligen las herramientas adecuadas para desarrollar la visualización final que será presentada al público.
v. **Percepción y comprensión**: Una vez publicada, la visualización es interpretada por los usuarios. A través de la percepción visual, el cerebro comienza a analizar la información, facilitando su entendimiento y asimilación.

A continuación, se representa en la Figura 33, el diagrama de procesos que reúne todos estos pasos.

Figura 33: Estadios de la visualización



Fuente: Autores

ISBN: 978-631-6557-56-8

4.3. Trabajar con datos

Hoy en día es común oír que "la información es poder", sin embargo, esta idea no es del todo precisa. Según [60], sería más adecuado decir que "*el conocimiento es poder, la información no*". Esto se debe a que, aunque vivimos rodeados de enormes volúmenes de información gracias a Internet y a los servicios de almacenamiento en la nube como *Dropbox*, *Google Drive* o *iCloud*, el verdadero desafío no está en almacenar datos, sino en analizarlos y transformarlos en conocimiento útil. Los servicios en la nube funcionan sobre plataformas más complejas como *Amazon S3*, *Google Cloud Storage* o *Azure Blob*, y disponen de bases de datos tanto relacionales (como AWS RDS o *Google Cloud SQL*) como NoSQL (como *DynamoDB* o *Bigtable*), además de almacenes de datos masivos como *BigQuery.* Esta infraestructura permite acceder a grandes cantidades de datos, pero también hace evidente que lo más difícil no es recopilar esa información, sino interpretarla correctamente para comunicarla de forma clara, ya sea con representaciones estáticas o interactivas. Para que una visualización sea comprensible y útil en la toma de decisiones, es necesario analizar bien los datos de origen. Estos pueden variar desde hechos ya procesados hasta registros en bruto que requieren una exploración completa. Un paso clave antes de crear cualquier gráfico es identificar los tipos de variables presentes, ya que de eso depende el tipo de representación más adecuada. Las variables se clasifican así:

Categóricas.- Representan categorías. Por ejemplo: *Nominales*: No tienen un orden específico (ej.: tipo de libro: aventura, drama, ciencia ficción) y *Ordinales*: Sí tienen un orden establecido (ej.: nivel educativo: secundaria, licenciatura, posgrado).

Cuantitativas.- Representan cantidades. Por ejemplo: *Continuas*: Pueden tener decimales (ej.: estatura, peso)y *Discretas*: Solo toman valores enteros (ej.: número de hijos, cantidad de productos).

4.3.1. Recolección de datos

La forma en que se recopilan los datos influye directamente en la calidad del conocimiento que se puede extraer, ya que un mal proceso puede introducir sesgos. Esta recolección puede ir desde métodos específicos como encuestas diseñadas cuidadosamente para empleados, hasta sistemas avanzados que recogen información de sensores o dispositivos IoT en fábricas, destinados a ser analizados con herramientas de big data. Además de los datos generados internamente, es posible utilizar fuentes externas, que pueden ser de dos tipos:

i. **Públicas**.- Gracias a iniciativas de datos abiertos (como *open data*, *linked open data* o *big and open linked data*), gobiernos, instituciones y empresas publican datos accesibles

ISBN: 978-631-6557-56-8

para cualquier persona, fomentando la transparencia y la colaboración en la resolución de problemas comunes.

ii. **Privadas**.- En otros casos, el acceso a ciertos conjuntos de datos está restringido, ya sea por pertenecer a una red específica o porque se comercializan. Estos datos requieren acuerdos de licencia y el pago correspondiente para su uso.

4.3.2. Preparación y limpieza de datos

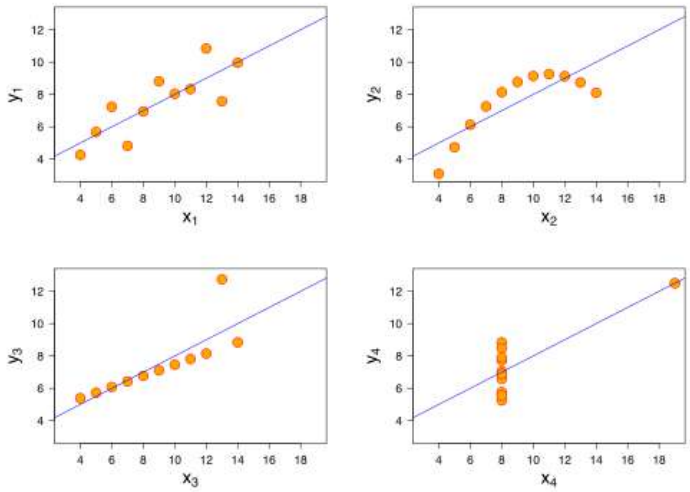
Antes de visualizar los datos, es fundamental realizar tareas previas de preparación, sobre todo si provienen de fuentes externas. Aunque muchas de estas actividades pueden automatizarse con herramientas digitales, casi siempre es necesario un trabajo manual que suele resultar tedioso. Esta fase incluye varias acciones complementarias: *i)* **Extraer los campos relevantes**, como separar la calle, ciudad y provincia en una dirección. *ii)* **Detectar información incompleta o vacía.** *iii)* **Unificar formatos y convertir unidades**, por ejemplo, estandarizar medidas expresadas en centímetros, pulgadas, metros o kilómetros. Una vez preparados, los datos deben limpiarse lo que corresponde a: *i)* **Detectar valores incorrectos o incoherentes**, como un número en un campo destinado a nombres. *ii)* **Eliminar registros duplicados.** *iii)* **Revisar si los valores están dentro de rangos plausibles**, por ejemplo, una temperatura de 224°C no sería válida en condiciones normales. *iv)* **Corregir errores ortográficos y de tipeo** y *v)***Validar formatos con patrones y expresiones regulares**, como en direcciones de correo electrónico.

4.3.3. Transformación de datos

La transformación de datos es una técnica común que permite descubrir patrones ocultos que no se perciben en su forma original. A menudo se usa para adaptar los datos a los requisitos de ciertas pruebas estadísticas, como la normalidad de la distribución (forma de campana), que muchas pruebas asumen. Estas transformaciones implican sustituir una variable por una función matemática, como el logaritmo o la raíz cuadrada, para facilitar su análisis. También es frecuente resumir los datos usando medidas de tendencia central, como la media o la mediana, aunque estas no siempre reflejan la verdadera forma de la distribución. Un ejemplo ilustrativo es el cuarteto de *Anscombe* mostrado en la Figura 34, donde diferentes conjuntos de datos comparten la misma media y recta de regresión, pero sus distribuciones son totalmente distintas. Esto demuestra que un solo valor promedio no siempre representa adecuadamente los datos.

ISBN: 978-631-6557-56-8

Figura 34: Cuarteto de Anscombe



Fuente: [61]

4.3.4. Visualización de datos

Para una visualización efectiva de datos, lo primero es definir con claridad el mensaje que se quiere comunicar. Es útil preguntarse: ¿Qué sé?, ¿qué significa?, ¿por qué es relevante? Si no se puede expresar en una frase breve, conviene repensarlo. Luego, hay que conocer a la audiencia para decidir cuántos y cuáles datos mostrar. Mostrar más no siempre es mejor; solo se debe incluir lo que aporte valor al mensaje. Usar una narrativa que guíe a la audiencia hacia el mensaje clave puede ser muy útil. Una buena prueba de esto es comprobar si el mensaje sigue siendo claro al cambiar el orden o eliminar elementos. Finalmente, se deben considerar aspectos visuales como el tipo de gráfica, el uso del color, la tipografía y el etiquetado. Visualizar datos es una combinación de ciencia y arte, y existen múltiples recursos para perfeccionar esta habilidad.

4.4. Definición y tipología de gráficos

Un gráfico es una forma visual de representar información. Esta representación puede ser figurativa o abstracta. Los gráficos abstractos (no figurativos) presentan los datos a través de sistemas visuales estandarizados y ampliamente comprendidos, aunque no guarden semejanza directa con lo que representan. A pesar de su carácter simbólico, permiten identificar de manera clara patrones, tendencias y relaciones dentro de un conjunto de datos, facilitando su interpretación de forma rápida y accesible.

4.4.1. Gráficos no figurativos

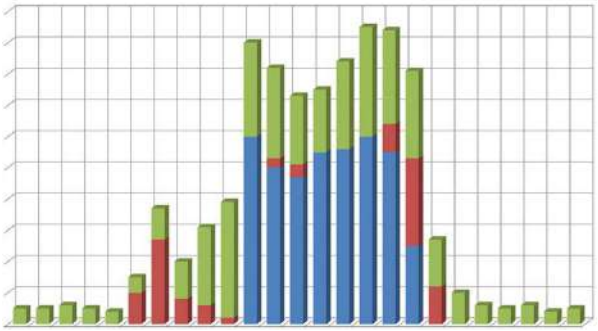
Los gráficos estadísticos son el ejemplo más representativo de gráficos no figurativos, ya que permiten visualizar grandes volúmenes de datos de forma clara y comprensible para el lector. Su diseño facilita la percepción inmediata de patrones, comparaciones y tendencias dentro de la información presentada. Entre los principales tipos se encuentran los gráficos o diagramas de

ISBN: 978-631-6557-56-8

barras, histogramas, polígonos de frecuencias, gráficos de líneas, diagramas de sectores (tartas) y los diagramas de cajas y bigotes; también se incluyen los pictogramas, aunque estos últimos se consideran gráficos figurativos por utilizar imágenes representativas. A continuación, se muestran este tipo de gráficos.

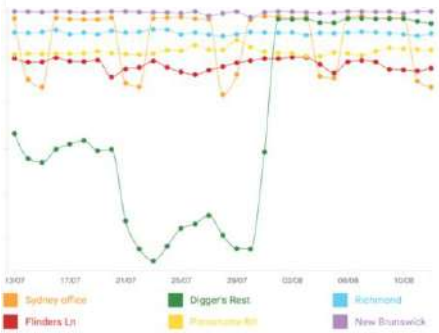
Los **gráficos de barras**, un tipo común de gráfico estadístico, representan datos mediante rectángulos paralelos cuya longitud es proporcional al valor que se quiere mostrar. Son útiles para comparar diferentes valores, analizar la evolución de un dato a lo largo del tiempo, visualizar la frecuencia de aparición de valores o rangos (como en los histogramas), y establecer clasificaciones ordenadas. Además, si se conectan los extremos superiores de las barras, se obtiene un polígono de frecuencias. También permiten observar desviaciones entre conjuntos de datos, desglosar valores totales en partes, examinar relaciones entre variables, y, mediante barras apiladas, mostrar la composición interna de cada valor. Cuando se emplean imágenes en lugar de barras, se denominan pictogramas. Un ejemplo de este tipo de gráficas se muestra en la Figura 35

Figura 36: Ejemplo de gráfico de barras



Fuente:[62]

Figura 35: Ejemplo de gráfico de líneas



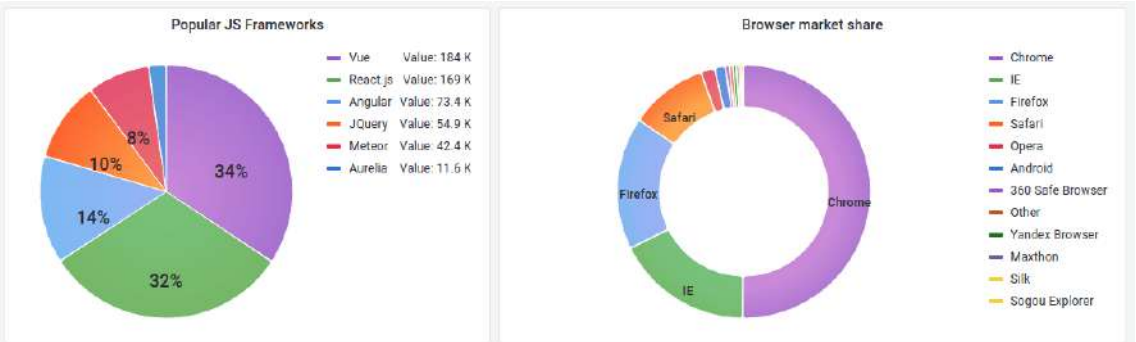
Fuente: [62]

Los **gráficos de línea** representan datos mediante puntos que se conectan con líneas, donde cada punto refleja un valor específico. Son especialmente útiles para mostrar la evolución de valores cuantitativos a lo largo del tiempo, comparar desviaciones entre dos conjuntos de datos y descomponer un valor agregado en sus componentes. También se emplean frecuentemente para mostrar múltiples tendencias o relaciones simultáneas en un mismo periodo temporal. Un ejemplo de este tipo de gráfico se observa en la Figura 36.

Los **gráficos de tarta** son representaciones circulares en las que cada sector muestra una parte proporcional de un total que equivale al 100 %. Su principal función es ilustrar las subdivisiones que componen un valor global. No obstante, cuando el número de particiones supera las cuatro o cinco, puede resultar más efectivo utilizar un gráfico de barras, ya que facilita una lectura más clara y comparaciones más precisas entre los valores. En la Figura 37 se observa un ejemplo de ese tipo de gráfico.

ISBN: 978-631-6557-56-8

Figura 37: Ejemplo de grafico de Tarta

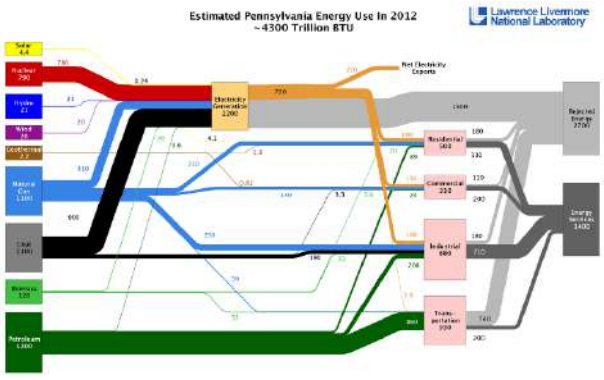


Fuente: [17]

4.4.2. Visualización de conjuntos de datos ligados temporales y espaciales

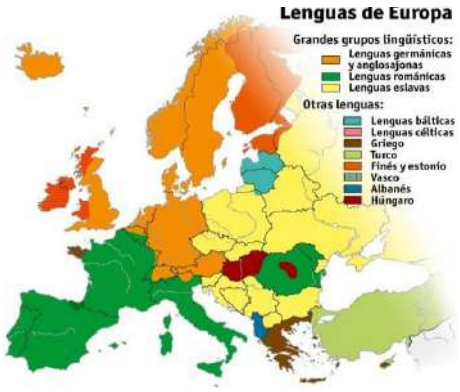
En algunos casos, los datos pueden presentar relaciones temporales o espaciales que es útil representar visualmente mediante líneas de tiempo o mapas. Las líneas de tiempo son una forma de mostrar datos a lo largo de un eje cronológico, permitiendo identificar la conexión entre los datos y un momento específico, además de revelar tendencias o patrones en el tiempo. Pueden presentarse tanto de manera horizontal como vertical. Por otro lado, los mapas de flujo ilustran el movimiento y la conexión entre puntos, utilizando líneas cuyo grosor refleja el valor representado, siendo Charles Minard el pionero en su sistematización, ejemplo en la Figura 38. Los mapas coropléticos, o de coropletas, muestran datos estadísticos dentro de áreas delimitadas, codificando estas regiones con diferentes colores, tonos, formas o texturas, y siempre deben incluir una leyenda para explicar la codificación empleada, ejemplo en la Figura 39.

Figura 39: Ejemplo de Mapa de Flujo



Fuente: [62]

Figura 38:Ejemplo de Mapa de coroplético



Fuente: [62]

4.4.3. Gráficos figurativos

Los gráficos figurativos presentan la información utilizando elementos que se asemejan visualmente a lo que representan, como mapas, fotografías o ilustraciones. Estos gráficos permiten al público identificar fácilmente lo que se muestra, ya que se relaciona directamente con el objeto o proceso que se desea explicar. Son efectivos para hacer comprensibles fenómenos complejos de forma sencilla, y suelen ser visualmente atractivos, captando rápidamente la

ISBN: 978-631-6557-56-8

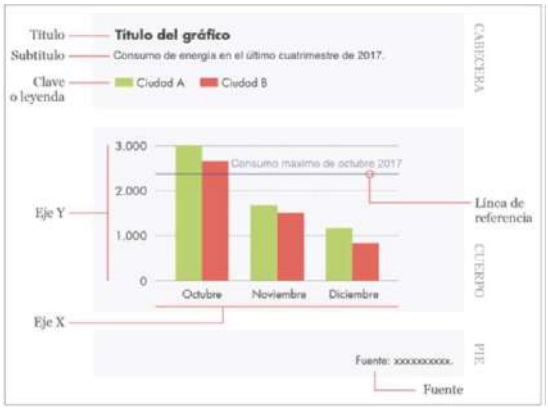
atención del espectador. Un ejemplo común de estos gráficos son los mapas. Otra variante son los pictogramas o pictografías, que son similares a los gráficos de barras, pero en lugar de barras, se utilizan imágenes representativas para ilustrar la magnitud de los datos.

4.4.4. Anatomía de un gráfico

Todos los gráficos tienen componentes clave que facilitan su comprensión. En los gráficos estadísticos, estos incluyen:

i. **Título**.- Debe ser directo y captar la atención del público objetivo, introduciendo claramente el tema.
ii. **Subtítulo**.- Proporciona detalles adicionales que no están en el título, como el período de los datos, y debe incluir las unidades de medida (por ejemplo, porcentaje, euros, personas).
iii. **Clave o leyenda**.- Explica cómo se ha codificado la información en el gráfico.
iv. **Eje Y (ordenadas)**.- Representa el eje cuantitativo, mostrando la frecuencia de los valores.
v. **Eje X (abscisas)**.- Representa el eje categórico, mostrando los valores de la variable.
vi. **Línea de referencia**.- Ayuda a contextualizar los datos mediante comparaciones.
vii. **Fuente**.- Siempre debe incluirse en el pie del gráfico, con un enlace en los gráficos
viii. interactivos para acceder a la fuente original de los datos.

Figura 40: Anatomía de un gráfico estadístico



Fuente: [62]

4.5. Visualización estática

Son representaciones gráficas que no cambian ni requieren interacción del usuario. Se utilizan tanto en medios impresos como digitales y su formato final es una imagen. Aunque no son interactivas, a veces comunican mejor la información que los gráficos dinámicos, como ocurre con los diagramas de montaje, que permiten seguir los pasos visualmente sin necesidad de interacción. Además, este tipo de visualizaciones fomenta un aprendizaje más profundo al requerir

ISBN: 978-631-6557-56-8

una interpretación mental del proceso. Estas visualizaciones son ideales para publicaciones impresas y también se integran fácilmente en entornos digitales. Una vez publicadas, no se modifican. Los formatos de salida dependen del medio:

i. **Soporte impreso**: se usan formatos de alta calidad como EPS, PDF o TIFF, generados con herramientas como *Adobe Illustrator*.
ii. **Soporte digital**: se prefieren formatos más ligeros como JPG, PNG o SVG. Además de *Illustrator*, se pueden usar herramientas como Inkscape, *Infogram, Creately* o *Easelly*. También es posible exportar capturas desde herramientas dinámicas para obtener gráficos estáticos.

4.6. Visualización dinámica

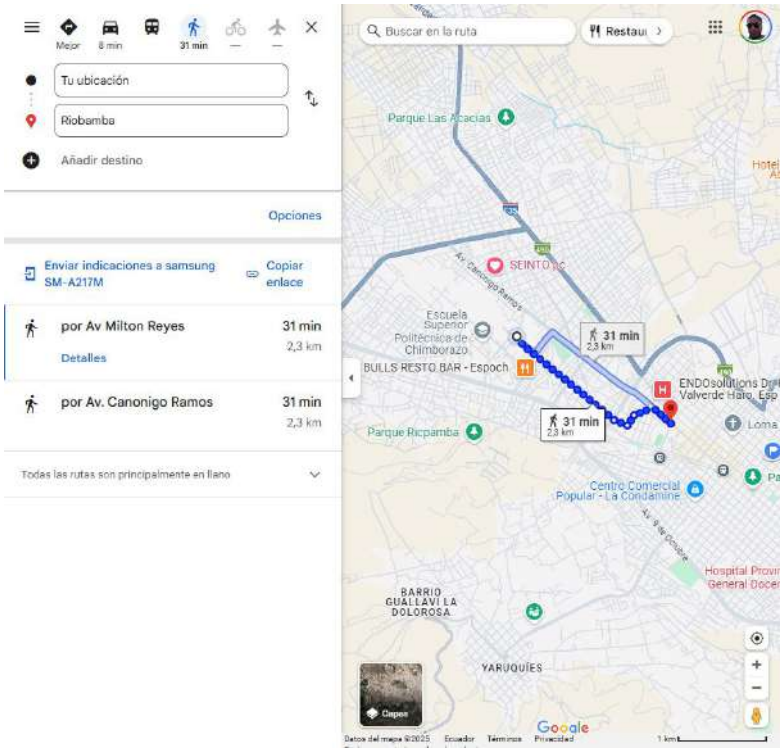
Las visualizaciones dinámicas muestran información que cambia con el tiempo y permiten la interacción del usuario, lo que facilita explorar y profundizar en los datos (como en mapas con actualizaciones en tiempo real). A diferencia de las visualizaciones estáticas, las dinámicas no están limitadas por el espacio físico y pueden contener múltiples pantallas o enlaces para ampliar información. Un ejemplo común es Google Maps, donde el usuario puede buscar lugares y obtener mapas detallados con opciones interactivas. Para crear visualizaciones dinámicas, existen herramientas como *Adobe Animate, Gapminder* o *Google Public Data Explorer*.

Los desarrolladores con conocimientos en programación pueden usar HTML, CSS y JavaScript para construir interfaces interactivas, integradas en soluciones tecnológicas. *Frameworks* como *Angular, React* o *Vue.js* facilitan esta tarea mediante arquitecturas como MVC o MVVM, permitiendo separar los datos de su presentación. Además, librerías como *Bootstrap* permiten crear interfaces adaptables a distintos dispositivos, y herramientas como *Google Charts*, D3.js, *Echarts, Highcharts* o *AmCharts* permiten generar gráficos visuales dinámicos. Estas herramientas ofrecen personalización en títulos, colores, tamaños, etc., y algunas, como D3.js, permiten trabajar con datos directamente en el navegador usando estándares web.

Por otro lado, si no se desea programar, plataformas como *Tableau, Grafana* o Google Data Studio permiten crear visualizaciones dinámicas fácilmente a partir de bases de datos, sin necesidad de conocimientos técnicos avanzados. En la Figura 41 se muestra un ejemplo de gráfico dinámico.

ISBN: 978-631-6557-56-8

Figura 41: Visualización dinámica de la ciudad de Riobamba



Fuente: [63].

4.7. Herramientas de visualización

El notable crecimiento en la cantidad de datos disponibles en los últimos años ha impulsado la aparición de numerosas herramientas estándar para su visualización. Actualmente, presentar los datos de manera clara y comprensible es una tarea accesible para cualquier usuario que necesite generar visualizaciones de forma eficiente, sin requerir conocimientos especializados como los de un desarrollador, ingeniero o analista de datos. En esta sección, se enumerarán algunas de las herramientas más populares en el ámbito profesional para representar datos provenientes de diversas fuentes, incluyendo algunas ya mencionadas anteriormente.

4.7.1. Herramientas para la visualización de datos

Existen diversas herramientas de código abierto y freemium que permiten crear visualizaciones de datos de forma sencilla, sin necesidad de conocimientos avanzados.

i. ***Datawrapper***.- Plataforma gratuita que facilita la creación de gráficos. Ofrece funciones adicionales mediante planes de pago. Se puede acceder con un correo electrónico o cuenta de Twitter.

ii. ***Timeline* JS**.- Permite generar líneas de tiempo interactivas usando hojas de cálculo, integrando contenidos multimedia de plataformas como YouTube o Google Maps. Las visualizaciones se adaptan automáticamente al dispositivo.

ISBN: 978-631-6557-56-8

iii. **RAWGraphs**.- Basada en D3.js, esta herramienta permite crear gráficos complejos sin necesidad de programar. Las visualizaciones se pueden descargar en SVG o como imágenes, o integrarse en sitios web mediante HTML.

iv. **CartoDB**.- Aplicación en la nube para crear mapas interactivos. Ofrece un modelo freemium con funciones básicas gratuitas y opciones avanzadas mediante suscripciones.

4.7.2. Soluciones de presentación de datos e inteligencia empresarial

Actualmente, muchas organizaciones ya cuentan con infraestructuras de ingestión, almacenamiento y procesamiento de datos a través de tecnologías como IIoT, bases de datos empresariales y herramientas Big Data. En este contexto, las soluciones de *business intelligence* (BI) permiten crear visualizaciones y cuadros de mando sin necesidad de desarrollar aplicaciones personalizadas, integrando múltiples fuentes de datos fácilmente.

Los proveedores de nube como **AWS**, **Google Cloud** y **Microsoft Azure** facilitan la interoperabilidad entre herramientas de visualización y sus servicios. Por ejemplo, es posible usar Power BI con datos en Google BigQuery o Google Data Studio con datos en AWS Redshift. A continuación, se muestras algunas herramientas destacadas de visualización de datos:

i. **Tableau**: Plataforma potente para análisis de datos y *dashboards*. Conecta con múltiples fuentes (*BigQuery*, *Redshift*, bases locales). Ofrece versiones como *Tableau Public* (gratuita), *Tableau Desktop*, Online y Server (de pago, con distintas funcionalidades y entornos).

ii. **Google Data Studio**: Gratuita y orientada a usuarios generales, permite crear *dashboards* conectando más de 800 fuentes mediante conectores, tanto propios de Google como de terceros.

iii. **Power BI**: Solución de Microsoft con diferentes versiones (Desktop, Pro, Premium, Mobile, Embedded). Admite comandos por voz y análisis de datos desde múltiples fuentes, incluyendo IoT.

iv. **Grafana**: Herramienta open-source inicialmente diseñada para monitorear sistemas, ahora permite dashboards avanzados conectando fuentes como MySQL, Prometheus, AWS IoT y Google BigQuery. Altamente extensible mediante plugins.

v. ***Looker***: Parte de Google Cloud, permite visualización avanzada y análisis de datos desde múltiples fuentes. Destaca por su sistema reutilizable "*Looker Blocks*" con licencia MIT.

vi. **Qlik**: Con enfoque en BI e IoT, ofrece productos como QlikView (motor asociativo) y Qlik Sense (evolución con más funciones), disponibles tanto en la nube como en instalaciones locales.

ISBN: 978-631-6557-56-8

vii. **Adverity**: Plataforma de pago para análisis de marketing y negocios. Soporta más de 400 conectores y puede enviar datos procesados a otras herramientas o almacenes.

viii. **Funnel**: Similar a Adverity, permite integrar más de 500 fuentes de datos y distribuir la información a plataformas de análisis, almacenamiento o visualización externas, incluyendo APIs personalizadas.

Estas herramientas facilitan la transformación de datos en información visual y útil, adaptándose a diferentes necesidades, presupuestos y niveles técnicos.

4.7.3. Lenguajes de programación para la presentación de datos personalizada

Es posible desarrollar una capa personalizada para mostrar datos dentro de una aplicación. En interfaces web, esto se logra con HTML, CSS, *JavaScript* y *frameworks* como Angular, React.js o Vue.js, junto con librerías de gráficos como Google Charts o D3.js. Cuando se busca una presentación de datos más avanzada, orientada a la ciencia de datos o inteligencia artificial, los lenguajes más utilizados son Python y R.

Python, creado por Guido van Rossum en 1991, es ampliamente usado en inteligencia artificial, desarrollo web, *scripting* y análisis de datos. Gracias a su comunidad, ofrece múltiples librerías especializadas. Para visualización, destacan *matplotlib*, una herramienta versátil para gráficos 2D, y *seaborn*, que se basa en *matplotlib* y facilita la creación de gráficos más estilizados y detallados.

R, por su parte, es un lenguaje derivado de S, desarrollado originalmente en *Bell Labs* y formalizado en 1995 por *Ross Ihaka* y *Robert Gentleman*. Es muy popular en estadística y análisis de datos, con más de 10,000 paquetes en su repositorio CRAN. En visualización, sobresale *ggplot2*, un paquete que permite crear gráficos interactivos y de alta calidad, adaptables a diversos campos como la salud, astronomía o genómica.

4.8. Visualización en la industria 4.0

En el contexto de la Industria 4.0 y el big data, la visualización de datos es clave para facilitar la comprensión, interpretación y toma de decisiones a partir de grandes volúmenes de información. Representar los datos de forma clara acelera el análisis y mejora la eficiencia.

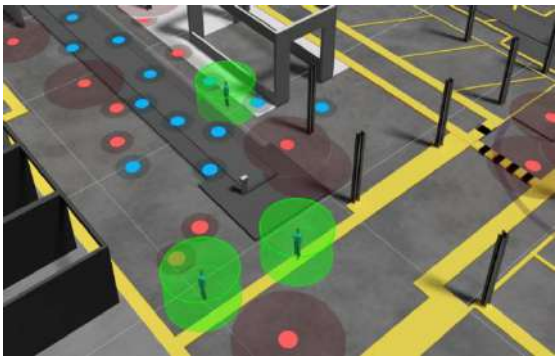
Existen múltiples aplicaciones prácticas, como el uso de mapas, mapas de calor o realidad virtual. Por ejemplo, la empresa Ubisense emplea estas técnicas en sistemas de localización en interiores con tecnología UWB, útiles para rastrear personas o equipos dentro de fábricas. Durante la pandemia de COVID-19, estas herramientas ayudaron a garantizar el distanciamiento entre trabajadores. Más allá de ese contexto, la localización de personas, vehículos o mercancías permite optimizar entregas, gestionar mejor los recursos y mejorar la eficiencia energética. En

ISBN: 978-631-6557-56-8

Business Intelligence, la visualización es imporante para crear cuadros de mando, donde el tipo de gráfico elegido depende del tipo de información a representar. Las gráficas temporales también son útiles, como muestra el sistema *SimbleSense*, que analiza el consumo energético diario y permite identificar picos y optimizar el uso de energía. Además, las infografías ayudan a simplificar procesos complejos y mostrar de forma clara flujos de datos o fases de un proceso, como lo hace *Siemens* en sus representaciones visuales.

Finalmente, los *dashboards* interactivos, como los que permite crear *Tableau*, integran distintos tipos de gráficos y datos en una sola vista, facilitando una visión global. En resumen, la visualización de datos es una herramienta en constante evolución que desempeña un papel importante en el análisis y uso eficiente de la información en la Industria 4.0. En la Figura 42 se muestra una solución de localización en interiores para asegurar la distancia de seguridad entre trabajadores en fábricas para evitar contagios por Covid-19 y la Figura 43 se muestra la información de ocupación estadística (mapas de calor) en la solución para asegurar la distancia de seguridad entre trabajadores en fábricas para evitar contagios por Covid-19, ambos casos son desarrollados por la empresa *Ubisence Contact Tracing*.

Figura 42: Solución de localización en interiores



Fuente: [64]

Figura 43: Muestra de información de ocupación estadística

Fuente: [64]

4.9. Definición de Inteligencia empresarial

La inteligencia empresarial (*Business Intelligence* o BI) abarca un conjunto de procesos, metodologías y tecnologías diseñadas para analizar datos y generar información útil para la toma de decisiones en una empresa. Estos datos pueden venir de diversas fuentes internas y externas, y se procesan mediante tareas como limpieza, integración y modelado antes de ser analizados.

Entre las herramientas más utilizadas están los almacenes de datos (*data warehouses*) y los *data marts*, donde se guarda información global o segmentada. También se emplean datos de terceros y fuentes como sensores o texto, en formatos estructurados o no estructurados. Una vez organizados los datos, se establecen metas (OKR) y se definen indicadores (KPI) para medir avances. Los OKR permiten establecer objetivos claros y medibles, mientras que los KPI ayudan

ISBN: 978-631-6557-56-8

a monitorear el desempeño de procesos específicos. Aunque a veces se confunden, los OKR se centran en metas generales, y los KPI en métricas asociadas a acciones concretas. Además, la BI incorpora técnicas avanzadas como minería de datos, análisis predictivo y uso de big data para anticipar escenarios y mejorar la planificación empresarial. En conjunto, estos elementos permiten tomar mejores decisiones y optimizar diversos aspectos del funcionamiento de una organización.

4.10. Importancia de la inteligencia empresarial

La inteligencia empresarial busca mejorar el funcionamiento de una organización a partir del análisis de datos. Al aplicar correctamente sus herramientas, las empresas pueden transformar la información en conocimientos útiles sobre sus procesos y estrategias, lo cual les permite tomar decisiones más acertadas, elevar la productividad e incrementar los ingresos. Sin BI, la toma de decisiones suele apoyarse en la experiencia o la intuición, lo que puede generar errores. En cambio, con estos sistemas es posible realizar un seguimiento continuo del rendimiento, identificar problemas y reaccionar ante oportunidades. Por ejemplo, se puede mejorar el enfoque de las áreas de marketing y ventas a partir del análisis de datos de clientes, o detectar fallos en la cadena de suministro antes de que causen perjuicios económicos. También permite evaluar la eficiencia del personal y controlar los costos laborales. Entre los principales beneficios del uso de BI se encuentran:

i. Mejorar la toma de decisiones.
ii. Optimizar procesos internos.
iii. Incrementar la eficiencia y la productividad.
iv. Detectar y resolver problemas.
v. Identificar nuevas tendencias de mercado.
vi. Formular mejores estrategias.
vii. Aumentar ventas e ingresos.
viii. Superar a la competencia.

4.11. Herramientas

Las herramientas de inteligencia empresarial (BI) incluyen una variedad de software, siendo los programas de visualización de datos los más comunes y populares. Plataformas como Qlik, Tableau, Data Studio y PowerBI dominan el mercado de herramientas BI de autoservicio. Aunque desarrolladores de herramientas tradicionales como Salesforce y SAP han seguido con sus propios planes, ahora incorporan funciones de autoservicio en sus soluciones, como la visualización de datos y consultas ad hoc. Las plataformas BI modernas incluyen características como:

ISBN: 978-631-6557-56-8

i. Software de visualización de datos para crear gráficos e infografías.
ii. Herramientas para diseñar informes y cuadros de mando que muestren métricas clave como OKR y KPI.
iii. Funciones de narración de datos que combinan visualizaciones y texto en presentaciones.
iv. Monitorización de rendimiento y control de seguridad.

Grandes proveedores de soluciones tecnológicas como IBM, Microsoft, Oracle, y SAP, junto con empresas emergentes como Alteryx y Sisense, ofrecen software de BI. También, empresas como Google han reforzado su presencia en este mercado al adquirir *Looker*. Las soluciones BI ahora son esenciales para integrar y visualizar datos de diversas fuentes, incluidas plataformas de big data como Hadoop, Spark, y bases de datos NoSQL, lo que permite obtener una visión unificada y facilita la participación de equipos multidisciplinarios en la mejora del rendimiento y la calidad de los datos.

4.12. Dirección estratégica

Tras abordar los pilares de la inteligencia empresarial, es esencial considerar aspectos estratégicos que permitan definir de manera efectiva los OKR y KPI, estableciendo una base sólida para su uso en la toma de decisiones. El concepto de estrategia proviene de la antigua Grecia, donde "strategos" refería al arte del general en la guerra.

En el contexto actual, caracterizado por un entorno cambiante y globalizado, las empresas deben adoptar una actitud estratégica similar a la de un general liderando su ejército, no solo para superar a la competencia, sino para guiar a la empresa hacia sus metas y misión. Aunque existen diversas definiciones de estrategia, [65] describe como un enfoque integral sobre cómo la empresa competirá, sus objetivos y las políticas necesarias para alcanzarlos.

La dirección estratégica se organiza en tres fases principales:

1) El **análisis estratégico externo** implica evaluar factores que afectan a la empresa desde fuera de su control, como cambios y tendencias en su entorno. Este análisis se desglosa en varias partes:

i. Análisis del entorno general, examina el contexto dinámico y competitivo donde la empresa opera, lo cual condiciona la formulación de estrategias.
ii. Análisis PESTEL, evalúa los factores políticos, económicos, sociales, tecnológicos, ambientales y legales que impactan a la empresa, fuera de su control directo.
iii. Análisis del entorno específico o sector industrial, estudia los factores externos relacionados con el sector donde opera la empresa.

ISBN: 978-631-6557-56-8

iv. Análisis de la estructura de la industria, evalúa cómo la estructura del sector afecta la rentabilidad y competitividad de la empresa.
v. Análisis de competidores y grupos estratégicos, consiste en estudiar a las empresas competidoras, predecir su comportamiento y utilizar esa información para formular estrategias.

2) El **análisis estratégico interno** se enfoca en identificar los recursos, capacidades y conocimientos dentro de la empresa, los cuales son fundamentales para evaluar sus fortalezas y debilidades. Este análisis abarca:

i. Análisis de la identidad: Incluye la edad, ciclo de vida, tamaño, ámbito geográfico y tipo de propiedad de la empresa.
ii. Análisis de recursos y capacidades: Evalúa tanto los recursos tangibles como intangibles y las capacidades funcionales y culturales.
iii. Análisis funcional: Examina las áreas clave de la empresa, como el área comercial, financiera, tecnológica, entre otras, para identificar las variables principales.
iv. Cadena de valor: Identifica las actividades que agregan valor a la empresa y son cruciales para su competitividad.

El perfil estratégico y el análisis DAFO ayudan a visualizar la situación de la empresa al comparar sus fortalezas, debilidades, oportunidades y amenazas. El análisis DAFO proporciona un resumen de los diagnósticos internos y externos de la empresa. La definición de objetivos y formulación estratégica requiere claridad sobre la misión, visión y los objetivos estratégicos de la empresa, asegurando que sean medibles, alcanzables y alineados con la visión a largo plazo. Los objetivos estratégicos deben ser específicos, medibles, realistas y desafiantes.

i. La estrategia competitiva se divide en tres tipos principales:
ii. Liderazgo en costos: Buscar ser el líder en costos del sector.
iii. Diferenciación: Ofrecer una diferencia apreciada por los consumidores.
iv. Enfoque o especialización: Centrarse en un nicho de mercado.

Existen también estrategias de desarrollo (expansión, penetración de mercado, desarrollo de productos) y estrategias de diversificación (tanto homogénea como heterogénea). Al evaluar opciones estratégicas, es importante considerar si la estrategia es conveniente, aceptable y factible, asegurando que sea adecuada para la empresa y los recursos disponibles.

3) La **implantación y control estratégico** son procesos clave para cumplir los objetivos. La implantación asigna tareas y establece un plan de acción, mientras que el control estratégico se

ISBN: 978-631-6557-56-8

enfoca en asegurar que los objetivos establecidos se cumplan, evaluando la eficacia y ajustando según sea necesario.

4.13. Cuadro de mando integral

Tras las etapas de análisis y planificación estratégica, el control estratégico se convierte en el proceso mediante el cual se asegura que la estrategia se ejecute correctamente para alcanzar los objetivos establecidos. Durante la era industrial, que se extendió hasta mediados de los años setenta, las empresas operaban en un entorno estable, con baja presión competitiva, ciclos de vida largos para los productos y una gran dependencia de la mano de obra y los costos de fabricación. En este contexto, la estrategia empresarial estaba centrada principalmente en el control de costos, conocido hoy como control de gestión. Sin embargo, a partir de mediados de los años noventa, el entorno cambió con la llegada de la era de la información. Los mercados se volvieron más dinámicos, competitivos y con ciclos de vida de los productos más cortos, mientras que la tecnología comenzó a jugar un papel fundamental. Las empresas tuvieron que adaptar o ampliar su estrategia para centrarse no solo en la reducción de costos, sino también en la satisfacción del cliente, la puntualidad en las entregas, la calidad del producto e innovación, lo que requería un enfoque más integral del control estratégico.

El control de gestión tradicional, centrado exclusivamente en el control de costos, resultó insuficiente para abordar esta nueva realidad. Fue necesario integrar nuevas variables que iban más allá del ámbito financiero. El Cuadro de Mando Integral (CMI), desarrollado por Kaplan y Norton en los años noventa, surgió como una respuesta a esta necesidad. Este modelo de gestión empresarial traduce la estrategia en objetivos interrelacionados que se miden mediante indicadores y están vinculados a planes de acción para alinear el comportamiento de los miembros de la organización con los objetivos estratégicos. El CMI se considera una herramienta estratégica porque permite comprender de manera precisa los objetivos estratégicos, tanto financieros como no financieros, y establece métodos para alcanzarlos. Además, facilita la comunicación de la estrategia y ayuda a gestionar la misma, enfocando y alineando equipos directivos, unidades de negocio, recursos y procesos con la estrategia global de la organización.

Las cuatro perspectivas del CMI permiten equilibrar los objetivos a corto y largo plazo, y la relación entre los resultados deseados y los factores que influyen en esos resultados. La **perspectiva financiera**, por ejemplo, está orientada a la generación de rendimientos a largo plazo para los inversores. Para ello, las empresas deben establecer objetivos financieros que aborden áreas como crecimiento, rentabilidad, costos y riesgo financiero, alineados con los análisis estratégicos previos. Esto incluye indicadores como rentabilidad económica, reducción de costos, aumento de ventas e ingresos, y mejora de la estructura financiera y de productividad. Estos

ISBN: 978-631-6557-56-8

objetivos financieros deben conectarse con metas en las demás perspectivas del CMI, como la del cliente, los procesos internos, y el aprendizaje y crecimiento, lo que permite tener una visión integral de la estrategia empresarial.

Por otro lado, la **perspectiva del cliente** destaca su papel central en la viabilidad de la empresa. La satisfacción del cliente es clave para generar ingresos, y se basa en comprender sus preferencias y necesidades, ofrecer productos y servicios que las satisfagan mejor que la competencia, y lograr su fidelización. Los indicadores en esta perspectiva incluyen la satisfacción del cliente, la calidad del producto, la cuota de mercado, y la retención de clientes. Esta perspectiva subraya la importancia de identificar a los clientes, comprender lo que valoran y evaluar el grado en que se cumplen sus expectativas.

En cuanto a la **perspectiva de procesos internos**, esta se enfoca en identificar los procesos críticos que deben ejecutarse con excelencia para lograr los objetivos económicos y satisfacer a los clientes. Aquí se busca la calidad del proceso, es decir, ofrecer productos de calidad al menor costo posible. El análisis se puede hacer mediante la cadena de valor de Porter, que ayuda a identificar tres procesos clave: la innovación, la eficiencia en las operaciones y el servicio postventa. En cuanto a los indicadores, se incluyen la calidad del producto, los costos de fallos, tiempos de entrega, y el número de reclamaciones, entre otros.

Finalmente, la **perspectiva de aprendizaje y crecimiento** aborda la necesidad de crear una infraestructura que permita a la empresa estar en un proceso continuo de aprendizaje y mejora. Esto implica invertir en la cualificación del personal, en proporcionar información estratégica a los empleados y en crear un clima organizacional que fomente la motivación y el compromiso para alcanzar los objetivos a largo plazo. Los indicadores en esta perspectiva incluyen la satisfacción y motivación de los empleados, su formación, la productividad y el ambiente laboral, así como la innovación a través de inversiones en investigación y desarrollo. A pesar de su importancia, los indicadores de esta perspectiva son a menudo los menos desarrollados en las empresas actuales, lo que indica una desconexión entre los objetivos estratégicos y el enfoque en el aprendizaje y crecimiento.

4.14. Inteligencia empresarial como soporte a la industria 4.0

A lo largo de este capítulo, se ha resaltado la relevancia de la inteligencia empresarial al establecer objetivos y utilizar indicadores en el ámbito corporativo. La Industria 4.0 está profundamente vinculada con la gestión y el análisis de datos, y en particular con los procesos de *business intelligence.* Los avances recientes en las tecnologías de la información y las comunicaciones (TIC) han tenido un impacto significativo en la gestión empresarial dentro de la industria, beneficiando enormemente estos procesos.

ISBN: 978-631-6557-56-8

La recopilación y el análisis de datos han sido fundamentales en el desarrollo del *business intelligence* desde sus inicios. Debido al gran volumen de datos generados por las tecnologías de la Industria 4.0, los procesos tradicionales de *business intelligence* deben adaptarse a estos nuevos retos. Este beneficio mutuo ha impulsado tanto a las soluciones comerciales como a las investigaciones a integrar la Industria 4.0 y el BI para mejorar los sistemas productivos, la gestión empresarial, la rentabilidad y la satisfacción del cliente. Específicamente, dentro de la Industria 4.0, la inteligencia empresarial respalda diversas áreas de la empresa, como la producción, las operaciones de marketing y ventas, la cadena de suministro, la investigación y el desarrollo, y las sinergias y la economía circular.

Por ejemplo, en la producción, se realiza la monitorización de maquinaria, el análisis de datos de producción y la predicción de demanda. En el ámbito de marketing y ventas, se emplean herramientas para prever la demanda y las ventas, analizar campañas de marketing y mejorar la relación con los clientes. En la cadena de suministro, se facilita la trazabilidad de productos y materias primas, así como el análisis de la satisfacción de los clientes y la determinación de stocks. En investigación y desarrollo, se lleva a cabo la monitorización de la inteligencia competitiva y se generan bases de conocimiento para facilitar el análisis. Además, en términos de sinergias y economía circular, se identifican oportunidades de especialización, se fomentan las economías de cercanía y se promueve la realimentación de los clientes.

Estos ejemplos muestran los diversos nichos de investigación y trabajo en el ámbito de la Industria 4.0 y el BI. Como bien se resume en el trabajo de [66], el uso de sistemas de big data en la Industria 4.0 facilita el diseño, seguimiento y mejora de las estrategias empresariales. Un claro ejemplo de esto es el sistema de gestión de carretillas industriales de **Linde MHI,** que mejora tanto la experiencia del cliente como los procesos internos mediante la automatización de alertas de mantenimiento. Otro ejemplo es la plataforma **IoT Cloud Service de Oracle**, que utiliza el big data y la inteligencia artificial para optimizar la producción, mejorar la capacidad de respuesta ante el mercado y reducir costos, todo basado en tecnologías como la realidad aumentada y el análisis automatizado de datos.

La mejora de los procesos internos en la Industria 4.0 también se ha beneficiado de tecnologías innovadoras, como las impresoras 3D. Empresas como ***Eschmann Textures*** han reducido los costos y tiempos de producción al imprimir sus diseños, lo que les permite validar rápidamente sus productos sin necesidad de realizar costosos ensayos previos. En sectores como la agricultura y ganadería, los sistemas de predicción de enfermedades en cultivos y la detección de partos en ganado vacuno, como los sistemas ***Hort@ y Moocall***, han mejorado la productividad y reducido los riesgos. Finalmente, el mantenimiento predictivo es otro ejemplo clave de cómo el big data y

ISBN: 978-631-6557-56-8

la BI apoyan la Industria 4.0. A través del análisis de datos, se pueden detectar fallos y defectos en la maquinaria antes de que ocurran, lo que mejora la gestión de recursos y reduce los costos, alargando la vida útil de la maquinaria y optimizando los períodos de amortización.

En los siguientes códigos QR o enlaces se muestra información adicional relacionada al capítulo.

La belleza de la visualización	Cuadro de mando Integral
https://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization#t-314777	https://www.youtube.com/watch?v=3pzTlYo0Ppw

ISBN: 978-631-6557-56-8

ISBN: 978-631-6557-56-8

CAPÍTULO V

PRIVACIDAD Y DESAFÍOS ÉTICOS EN LA ERA DEL BIG DATA

5.1. Introducción y Objetivo del Capítulo

El notable aumento de dispositivos, plataformas y herramientas que recopilan información personal ha llevado a las autoridades de diversas naciones y regiones a implementar regulaciones y medidas punitivas para controlar rigurosamente su manejo. Para profundizar en este tema, revisa las ideas clave que se presentan más adelante; en ellas examinaremos de manera concisa los efectos de la nueva Ley de Protección de Datos en las empresas, así como las normativas de privacidad en otras jurisdicciones, como la *California Consumer Privacy Act* en Estados Unidos o las leyes vigentes en países y estados de Latinoamérica. También se abordará la relevancia de la anonimización de datos como herramienta clave para garantizar el tratamiento seguro y legal de la información personal. Además, se explorarán los desafíos que representan el *big data* y la Industria 4.0 en materia de privacidad y protección de datos, así como las soluciones y tendencias emergentes para enfrentarlos. Al finalizar este capítulo se habrán alcanzado los siguientes objetivos:

- Analizar los fundamentos clave de la protección de datos en la Unión Europea, destacando sus normativas más relevantes.
- Examinar los principios elementales establecidos por la *California Consumer Privacy Act* (CCPA) en materia de privacidad de datos en Estados Unidos.
- Explorar los marcos regulatorios más importantes implementados en Latinoamérica (LATAM) para la protección de datos personales.
- Comprender las técnicas de anonimización, su aplicación y su relevancia en el manejo seguro de información personal.
- Evaluar el impacto del tratamiento de datos en el contexto de la Industria 4.0, reconociendo su importancia no solo en la fase operativa, sino también desde el diseño inicial de los sistemas.

5.2. Definiciones previas

Para un análisis adecuado de las normativas, primero precisamos clarificar términos elemetales que contextualicen su propósito, enfoque y aplicación práctica.

ISBN: 978-631-6557-56-8

5.2.1. Reglamentos de protección de datos

La normativa sobre protección de datos busca garantizar que las personas tengan el máximo control sobre la recolección y uso que las empresas hacen de su información personal. Aunque estas regulaciones surgieron en los años ochenta, su relevancia ha crecido significativamente en la era digital, impulsada por la explosión de fuentes de datos y el auge de Internet.

5.2.2. Datos personales

Los datos personales son cualquier información que permita identificar a una persona, ya sea de forma directa (como nombre, teléfono o dirección) o indirecta (a través de identificadores digitales o combinación de datos). Entre ellos se incluyen: documentos oficiales (DNI, NIF, pasaporte), contactos (correo electrónico, domicilio), rastros digitales (cookies, historial web, interacciones), ubicación precisa, perfiles en redes sociales y datos pseudoanonimizados. No se consideran personales los datos anonimizados o agregados.

5.2.3. Información de identificación personal

El término PII (Información de Identificación Personal) se usa principalmente en Estados Unidos. Según su definición oficial, corresponde a "datos que permiten distinguir o rastrear la identidad de un individuo, como su nombre, número de seguridad social o registros biométricos" [67], lo que podría interpretarse como limitado a información offline. En términos generales, el concepto de información personal es más amplio que el PII, ya que abarca también identificadores en línea como direcciones IP, cookies, dispositivos y datos de ubicación geográfica.

5.2.4. Datos sensibles

Ciertos datos requieren especial consideración por su naturaleza delicada, planteando dilemas éticos sobre su uso. En esta clasificación se encuentran información sobre salud, orientación sexual, finanzas personales, convicciones religiosas y afiliaciones políticas.

5.2.5. Datos de geolocalización precisa

Los datos PLD (del inglés *Precise Location Data*) son aquellos que indican la ubicación geográfica exacta de un dispositivo. Estos datos pueden obtenerse a través de diferentes tecnologías, siempre que permitan determinar con precisión razonable la localización física real de una persona o dispositivo. Entre los ejemplos se incluyen las coordenadas de latitud y longitud proporcionadas por sistemas de navegación por satélite (GNSS) como GPS, GLONASS o Galileo, así como la posición calculada mediante triangulación de señales de radio, como las usadas en redes de telefonía móvil.

ISBN: 978-631-6557-56-8

5.2.6. Gobierno de los datos

El gobierno de datos (o *Data Governance* en inglés) consiste en un conjunto de buenas prácticas, metodologías, procesos, tecnologías y comportamientos orientados a garantizar una gestión segura, eficaz y eficiente de los datos. Esto abarca aspectos como la seguridad y privacidad, el cumplimiento normativo (*compliance*), la integridad y usabilidad de los datos, la gestión de socios tecnológicos y de datos, los flujos de información, las taxonomías empleadas, así como las funciones y responsabilidades de los encargados de su administración.

5.2.7. Privacidad como premisa de diseño

El enfoque de Privacidad desde el Diseño (*Privacy by Design*) establece que la protección de datos debe integrarse como un componente fundamental en todas las etapas de creación de sistemas, servicios, productos o procesos. Esto implica considerar los aspectos de privacidad como elementos fundamentales, no solo durante la fase inicial de diseño, sino a lo largo de todo el ciclo de vida de la solución implementada.

5.3. Reglamento general de protección de datos (Contexto Europeo)

Como se ha mencionado previamente, la evolución tecnológica y social de los últimos años, junto con la necesidad de adaptar el marco normativo a los nuevos desafíos, condujo a la creación del Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 (en adelante, RGPD). Esta normativa, que protege los derechos de las personas en el tratamiento de datos personales y su libre circulación, entró en vigor el 25 de mayo de 2018.

Antes de profundizar en aspectos específicos sobre protección de datos, es fundamental comprender que las distintas legislaciones en esta materia buscan salvaguardar las libertades públicas y derechos fundamentales de los individuos, particularmente su honor e intimidad personal y familiar, en todo lo relacionado con el procesamiento de datos personales. Los principios fundamentales del RGPD que se aborda en este capítulo se muestran en la siguiente lista.

1) Datos personales.
2) Tratamiento.
3) Limitación del tratamiento.
4) Pseudoanonimización.
5) Fichero.
6) Responsable del tratamiento.
7) Encargado del tratamiento.
8) Destinatario.
9) Tercero.
10) Consentimiento del interesado.
11) Violación de la seguridad de los datos personales.
12) Autoridad de control.
13) Tratamiento transfronterizo.
14) Objeción pertinente y motivada.

ISBN: 978-631-6557-56-8

15) Organización internacional.

A continuación, se describe cada una de estas características.

El RGPD define los **datos personales** como cualquier información relativa a una persona física identificada o identificable (el interesado), siendo especialmente relevante el carácter interpretativo del término "identificable". Este concepto abarca información que permite la identificación indirecta, como direcciones IP, números telefónicos fijos, ADN, huellas dactilares u otros datos biométricos. El Considerando 26 del RGPD, alineado con la LOPD, especifica que para determinar si una persona es identificable deben considerarse todos los medios razonablemente disponibles -tanto técnicos como económicos- que permitan su identificación directa o indirecta. Esta definición amplia resulta crucial para las tecnologías de la Industria 4.0, donde identificadores únicos en dispositivos o aplicaciones pueden facilitar la identificación personal, encajando perfectamente en este concepto de "identificabilidad".

Se entiende por **tratamiento** de datos personales cualquier operación o conjunto de operaciones realizadas sobre datos personales o conjuntos de datos personales, tanto mediante procedimientos automatizados como manuales. Esto incluye, entre otros: la recogida, registro, organización, estructuración, conservación, adaptación o modificación, extracción, consulta, utilización, comunicación por transmisión, difusión o cualquier otra forma de habilitación de acceso, así como el cotejo o interconexión, limitación, supresión o destrucción de los mismos, conforme a lo establecido en el artículo 4.2 del Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016.

El marcado de los datos de carácter personal, según el artículo 4.3 del RGPD, es un mecanismo destinado a **limitar el tratamiento** futuro de esos datos. Sustituye el concepto de bloqueo y establece criterios específicos para el tratamiento de los datos. Este marcado puede implicar, por ejemplo, un plazo máximo durante el cual los datos pueden ser utilizados. Después de este periodo, el tratamiento de los datos debe limitarse o cesar, garantizando que no se utilicen de manera inapropiada o para fines no autorizados.

Pseudoanonimización, el tratamiento de datos personales consiste en transformarlos de manera que no puedan ser vinculados a un individuo específico sin recurrir a información adicional. Dicha información debe mantenerse separada y protegida mediante medidas técnicas y organizativas adecuadas, que aseguren que los datos personales no puedan asociarse a una persona física identificada o identificable, tal como establece el artículo 4.5 del RGPD.

ISBN: 978-631-6557-56-8

Fichero, es un conjunto organizado de datos personales que pueden ser consultados siguiendo criterios específicos, sin importar si se encuentran centralizados, distribuidos o repartidos funcional o geográficamente, según lo establece el artículo 4.6 del RGPD.

Responsable del tratamiento o responsable, la persona física o jurídica, autoridad, servicio u otro organismo que, solo o junto con otros, determine los fines y medios del tratamiento (art 4.7 RGPD).

Encargado del tratamiento o encargado, la persona física o jurídica, autoridad, servicio u otro organismo que trate datos personales por cuenta del responsable del tratamiento (art 4.8 RGPD).

Destinario, se entiende como destinatario a cualquier persona física o jurídica, autoridad, entidad u organismo que reciba datos personales, independientemente de que sea o no un tercero. Sin embargo, no se consideran destinatarios a las autoridades que accedan a esos datos en el contexto de una investigación específica, siempre que dicho acceso se ajuste a la legislación de la Unión o de los Estados miembros; en estos casos, el tratamiento de la información deberá cumplir con las normas de protección de datos correspondientes a su finalidad, conforme indica el artículo 4.9 del RGPD.

El término **tercero** hace referencia a cualquier persona física o jurídica, autoridad, entidad u organismo que no sea el propio interesado, ni el responsable o encargado del tratamiento, ni tampoco alguien autorizado para tratar los datos personales bajo la supervisión directa del responsable o encargado, según lo establecido en el artículo 4.10 del RGPD.

Consentimiento del interesado, se entiende como toda expresión de voluntad que sea libre, específica, informada y clara, mediante la cual la persona interesada acepta, ya sea de forma verbal o mediante una acción afirmativa evidente, que sus datos personales sean tratados (art. 4.11 RGPD). En el Reglamento, este consentimiento adquiere un carácter más estricto, por lo que no se admite la presunción de consentimiento.

Violación de la seguridad de los datos personales, se refiere a cualquier incidente que comprometa la seguridad y provoque la destrucción, pérdida o modificación, ya sea accidental o intencionada, de datos personales que hayan sido transmitidos, almacenados o tratados, así como el acceso o divulgación no autorizada de dicha información (art. 4.12 RGPD). En algunos contextos, es habitual que este tipo de incidentes se denominen como «brechas de seguridad».

Autoridad de control, es el organismo independiente designado por cada Estado miembro conforme a lo establecido en el artículo 51 del RGPD (art. 4.21). De manera general, se trata de

ISBN: 978-631-6557-56-8

las agencias responsables de velar por el cumplimiento de la normativa de protección de datos en cada país. En el caso de España, esta función recae en la Agencia Española de Protección de Datos.

Tratamiento transfronterizo, se refiere a dos situaciones específicas:

- Cuando el tratamiento de datos personales se lleva a cabo en el marco de las actividades de establecimientos ubicados en más de un Estado miembro, ya sea por parte del responsable o del encargado del tratamiento dentro de la Unión Europea.

- Cuando el tratamiento es realizado desde un único establecimiento, pero sus efectos impactan de manera significativa o probablemente afectarán a interesados de varios Estados miembros.

Objeción pertinente y motivada, se refiere a la oposición a una propuesta de decisión relacionada con la existencia de una infracción del RGPD o con la conformidad de las acciones previstas por el responsable o encargado del tratamiento con dicho reglamento. Esta objeción debe estar claramente justificada, mostrando los riesgos significativos que el proyecto de decisión podría implicar para los derechos y libertades fundamentales de los interesados, y en su caso, para la libre circulación de los datos personales dentro de la Unión Europea (art. 4.24 RGPD).

Organización internacional, se refiere a una entidad internacional y sus organismos subordinados de derecho público internacional, o cualquier otro organismo establecido mediante un acuerdo entre dos o más países, o en virtud de dicho acuerdo (art. 4.26 RGPD).

El Reglamento General de Protección de Datos (RGPD) introduce cambios significativos respecto a la anterior Ley de Protección de Datos. Principalmente, otorga a los individuos un mayor control sobre sus datos personales [68]. Este aumento de poder se debe en gran medida al auge de las redes sociales y la gran cantidad de información que se intercambia a través de ellas. La digitalización ha incrementado la exposición de datos personales, lo que ha llevado a la Unión Europea a reforzar la protección de la privacidad individual. La aplicación de este reglamento es muy amplia, ya que cualquier entidad, ya sea persona o empresa, que maneje datos personales dentro de la Unión Europea debe cumplir con sus disposiciones.

La afectación del Reglamento General de Protección de Datos (RGPD) es considerable, ya que toda entidad, ya sea un individuo o una empresa, dentro de la Unión Europea que maneje

ISBN: 978-631-6557-56-8

datos personales debe ajustarse a sus disposiciones. Según [69], las nuevas obligaciones que implica este reglamento son las siguientes:

Las nuevas obligaciones del Reglamento General de Protección de Datos (RGPD) incluyen:

- **Rendición de cuentas.-** Cualquier brecha de seguridad que afecte a los datos debe ser comunicada a la Agencia Española de Protección de Datos (AEPD) en un plazo máximo de 72 horas. Si la brecha involucra datos sensibles, como los relacionados con la salud, también debe informarse a los usuarios afectados.
- **Responsabilidad proactiva.-** Las entidades deben tomar medidas para prevenir incidencias que puedan comprometer la integridad de los datos. Además, deben registrar las actividades preventivas si la empresa tiene más de 250 empleados.
- **Delegado de Protección de Datos .-** Las entidades que trabajen con datos sensibles deben contar con un responsable de seguridad de datos, encargado de garantizar el cumplimiento del reglamento.
- **Derecho al olvido.-** Los usuarios tienen el derecho de solicitar la eliminación de sus datos, y la entidad debe cumplir con esta solicitud de manera obligatoria.
- **Derecho a la portabilidad.-** Los usuarios pueden solicitar que sus datos sean entregados en un formato que permita su transferencia a otros sistemas digitales.
- **Obtención del consentimiento.-** El consentimiento de los usuarios debe ser libre, informado, específico e inequívoco.
- **Tratamiento de datos por parte de terceros.-** Las subcontrataciones o servicios contratados a otras entidades también deben cumplir con el RGPD, y estas empresas deberán emitir un certificado que acredite su cumplimiento.

Trabajar con diversas normativas en un mundo globalizado puede hacer que sea complicado cumplir con todas ellas. Los entornos de la industria 4.0 se ven considerablemente impactados por estos cambios, especialmente cuando las soluciones implementadas son utilizadas por usuarios en varios países. Un ejemplo de esto es un sistema de localización que ubica a trabajadores en diferentes partes del mundo, o cuando se crea un sistema de robótica en un país, pero sus datos se gestionan en servidores situados en otro. Dado que las exigencias fuera de la UE, y particularmente en países de LATAM y EE. UU., varían, es necesario buscar convergencias en la aplicación de los principios de privacidad.

5.4. Privacidad en EE. UU.: California *Consumer Privacy Act*

Muchos consideran que en EE. UU. se encuentran los antecedentes de la privacidad y la protección de datos personales, y que la normativa europea se inspiró en gran medida en estos principios en sus orígenes. En 1798, la Cuarta Enmienda de la Constitución de EE. UU. reconoció

ISBN: 978-631-6557-56-8

el derecho de los ciudadanos a la inviolabilidad de su domicilio. Sin embargo, no sería hasta el siglo XIX cuando el concepto moderno de privacidad empezaría a desarrollarse en la sociedad estadounidense, en respuesta al creciente debate social sobre la protección de la vida privada por parte de la ley. A finales de ese siglo, los juristas norteamericanos Samuel Warren y Louis Brandeis jugarían un papel clave en la consolidación del derecho a la vida privada, al publicar un artículo en la revista de la Universidad de Harvard titulado *The Right to Privacy*. En él, introdujeron por primera vez el concepto de privacidad como una acción civil, abogando por la creación de un nuevo derecho civil que protegiera el espacio personal frente a su divulgación no autorizada al público.

Sin embargo, tuvieron que pasar varios años hasta que el derecho a la privacidad se consolidara, especialmente en Estados Unidos, donde se integra en cuatro ámbitos: la esfera privada, la apropiación del nombre, la distorsión de la imagen y la difusión pública de hechos privados. Solo después de un conjunto de normas y a través de la jurisprudencia y la legislación, se fue consolidando este derecho, en resumen se muestran las normas que consolidan el derecho de la privacidad en EE.UU en la Tabla 3.

Tabla 4: Normas que consolidan el derecho de privacidad en EE. UU

PrivacyAct	Ley de Protección de la Intimidad de 1974
Freedmon of information Act	Ley de Libertad de la Información, FOIA
FairCreditReportingAct	Ley de Equidad Financiera de 1978
PrivacyAct	Ley de Protección de la Intimidad de 1974

Fuente: Autores

5.4.1. Información personal que se recopila

Los consumidores tendrán el derecho de conocer, mediante una política o aviso de privacidad claro y general, qué datos personales ha recolectado la empresa sobre ellos, cuál ha sido su origen y con qué finalidad se están utilizando dichos datos.

5.4.2. Conocer destino de la información

Se entiende que las empresas que comparten o facilitan a terceros los datos personales de los consumidores, ya sea mediante su venta o cualquier otro método que les genere beneficios, están obligadas a informar de ello a los usuarios. Además, los consumidores tienen el derecho a excluirse de esta práctica a través de un enlace visible, identificado como «No vender mi información personal», que debe estar disponible en la página principal de la empresa, tal como exige la ley. Cabe señalar que el concepto de venta no se limita exclusivamente a la transacción comercial de datos, sino que también abarca cualquier forma de intercambio de información personal con terceros.

ISBN: 978-631-6557-56-8

5.4.3. Acceso a la información personal que ha sido recopilada

Los consumidores tienen el derecho de pedir a las empresas detalles específicos sobre sus datos personales, incluyendo las fuentes de las cuales se obtuvo dicha información, los datos concretos que han sido recopilados y con qué terceros se han compartido. La legislación establece que las empresas deben ofrecer canales claros y accesibles como teléfono, formularios o correo electrónico para que los consumidores puedan realizar estas solicitudes. Una vez recibida la petición, la empresa tiene un plazo máximo de 45 días para entregar la información solicitada, sin que esto genere ningún costo para el solicitante.

5.4.4. Eliminación de la información personal

Los consumidores tienen la posibilidad de solicitar que la empresa elimine los datos personales que haya recopilado sobre ellos. Sin embargo, aquellos datos cuya conservación sea obligatoria por disposición legal no estarán sujetos a esta solicitud y deberán mantenerse.

5.4.5. Antidiscriminatoria por ejercer sus derechos bajo la ley

La CCPA garantiza a los consumidores el derecho a acceder a los mismos productos, servicios y precios por parte de una empresa, aun cuando decidan ejercer sus derechos de privacidad contemplados en la ley. En este sentido, las empresas no pueden aplicar ningún tipo de trato discriminatorio hacia los consumidores por hacer uso de dichos derechos, lo que implica que no pueden negarles bienes o servicios, cobrar tarifas distintas ni ofrecer una calidad inferior en los productos o servicios.

5.5. Privacidad en LATAM

La normativa sobre protección de datos personales varía considerablemente a nivel internacional, mostrando diferencias significativas tanto en su alcance como en los derechos que otorgan a las personas. Al comparar las leyes de distintos países, es posible observar dos enfoques legislativos claramente diferenciados, marcados en gran medida por la tradición jurídica de cada lugar. Por un lado, existen Estados que han adoptado marcos legales amplios y coherentes, con leyes generales conocidas como leyes ómnibus que, en algunos casos, se ven complementadas por regulaciones específicas según el sector. Por otro lado, hay países que disponen de un entramado de normas dispersas, con legislaciones sectoriales o territoriales que, en conjunto, resultan complejas de identificar y aplicar en su totalidad. Además, se observa una gran diversidad en cuanto a los derechos que reconocen estas leyes y las obligaciones que imponen, así como en el nivel de protección que garantizan a los ciudadanos frente al tratamiento de sus datos personales.

Un ejemplo evidente de estas diferencias se observa al comparar la normativa de la Unión Europea con la de Estados Unidos. Para ampliar esta visión y mostrar aún más matices sobre cómo distintos

ISBN: 978-631-6557-56-8

países enfrentan el desafío de regular el tratamiento de datos personales, en esta sección se detallan las principales características de las leyes vigentes en varios países de América Latina.

5.5.1. Protección de datos en Argentina

El artículo 43 de la Constitución Federal reconoce a los ciudadanos el derecho a recurrir a la vía judicial para acceder a la información personal que figure sobre ellos en bases de datos, tanto públicas como privadas, y solicitar su corrección, actualización o eliminación si dicha información es incorrecta. Por su parte, la Ley de Protección de Datos Personales N.º 25.326 (LPDP), promulgada en octubre de 2000, establece un marco legal más amplio y detallado, tomando como referencia la legislación española en esta materia. Gracias a esta alineación normativa, el 30 de junio de 2003, la Comisión Europea declaró que Argentina garantiza un nivel «adecuado» de protección de datos personales conforme a la Directiva de Protección de Datos (95/46/CE). Además, el país cuenta con una autoridad central, la Dirección Nacional de Protección de Datos Personales (DNPDP), que supervisa el cumplimiento de la ley, establece la obligación de notificar los ficheros y exige, en ciertos casos, la designación de un responsable de seguridad para el tratamiento de datos.

5.5.2. Protección de datos en Brasil

Desde septiembre de 2020, Brasil cuenta con la Ley General de Protección de Datos (LGPD – 13.709/18), la cual entró en vigor con el objetivo de reforzar la protección y gestión transparente de los datos personales. Inspirada en el Reglamento General de Protección de Datos (RGPD) de la Unión Europea, esta normativa establece una serie de obligaciones para las empresas, que deben implementar y mantener procedimientos que garanticen la seguridad y correcto tratamiento de la información. Entre los elementos clave que contempla la ley se encuentran:

- La diligencia en el manejo de datos.
- La realización de auditorías.
- La implementación de modelos de gobernanza.
- El desarrollo de planes de comunicación y gestión de incidencias.
- La obtención del consentimiento y, cuando corresponda, el tratamiento mediante anonimización.
- La designación de un delegado de protección de datos.
- Adopción de medidas de seguridad para la protección de la información.

ISBN: 978-631-6557-56-8

5.5.3. Protección de datos en Chile

La protección de datos personales en Chile se regula a través de diversas leyes específicas y normas complementarias que establecen principios y obligaciones sobre el manejo de la información personal. Entre las principales disposiciones destacan:

- La Constitución de la República de Chile, en su artículo 19 nº 4, garantiza el respeto y la protección de la vida privada, el honor personal y familiar. Además, faculta a cualquier persona afectada por una acción u omisión arbitraria o ilegal a interponer una acción de amparo constitucional para proteger estos derechos.
- La Ley 19.628 Sobre la Protección de la Vida Privada (LPVP) regula el tratamiento de datos personales en bases públicas y privadas, siendo modificada por última vez el 17 de febrero de 2012.
- La Ley 20.285, sobre el acceso a la información pública, establece como principio la transparencia en la Función Pública y garantiza el derecho de acceso a la información de los órganos estatales, detallando procedimientos y excepciones.
- La Ley 20.575 consagra el "principio del destino" en el tratamiento de datos personales, añadiendo disposiciones específicas sobre el uso de datos económicos y de deudas.
- La Ley General de Bancos, en su artículo 154, consagra el secreto bancario, limitando el acceso a la información de depósitos únicamente al titular o su representante, salvo excepciones legalmente previstas.
- La Ley 19.223 tipifica conductas delictivas vinculadas al uso indebido de información contenida en bases de datos electrónicas, estableciendo sanciones para quienes accedan o utilicen dicha información de manera ilícita.

Además, en Chile:

- Las bases de datos del sector público deben registrarse obligatoriamente, mientras que no existe actualmente un requerimiento legal para notificar incidentes de seguridad relacionados con datos personales.
- Todo el personal que participe en el tratamiento de datos personales tiene la obligación legal de confidencialidad respecto a aquella información no pública, incluso después de terminada la relación laboral o contractual.
- La persona responsable de una base de datos tiene la obligación de aplicar medidas de seguridad con la debida diligencia para proteger los datos personales contenidos en ella.

ISBN: 978-631-6557-56-8

5.5.4. Protección de datos en Colombia

La protección de datos personales en Colombia se basa en un marco normativo extenso que abarca diversas leyes, decretos y disposiciones que buscan garantizar la privacidad y los derechos de los ciudadanos en cuanto al tratamiento de su información personal. A continuación se presentan las principales leyes y normas relacionadas con la protección de datos personales en Colombia:

a) **Constitución de Colombia**: El artículo 15 establece el derecho fundamental a la privacidad, el buen nombre, la reputación y la protección de los datos personales. Este derecho se encuentra respaldado por una serie de leyes y decretos adicionales que profundizan y regulan su alcance.

b) **Ley 1266 de 2008**: Regula la recolección, uso y transferencia de datos personales relacionados con el crédito, servicios financieros y bancarios. Fue revisada por la Corte Constitucional de Colombia en la Decisión C 1011-1008, centrando su atención en el manejo de la información crediticia.

c) **Ley 1581 de 2012 (Ley de Protección de Datos Personales)**: Esta ley establece el marco normativo principal para la protección de los datos personales en Colombia. Su objetivo es hacer efectivo el derecho constitucional de acceder, actualizar y rectificar la información recopilada sobre los ciudadanos en las bases de datos públicas o privadas. La ley se aplica a todos los datos personales almacenados en bases de datos o archivos, tanto en Colombia como en personas fuera del país que estén bajo la jurisdicción colombiana.

    - Esta ley establece que el titular de los datos debe dar su consentimiento previo, expreso e informado para cualquier actividad relacionada con la recopilación, uso o transferencia de sus datos personales, a menos que esté exento por la ley, como en el caso de la información de crédito bajo la Ley 1266.

d) **Decreto 1377 de 2013**: Regula la implementación práctica de la Ley 1581 y aborda aspectos clave como:

    - La autorización para el tratamiento de datos personales.
    - El tratamiento de datos sensibles.
    - Las políticas de tratamiento de datos personales.
    - Las transferencias internacionales de datos.
    - Las responsabilidades de los responsables del tratamiento de los datos.
    - El ejercicio de los derechos de los titulares de los datos, como la rectificación y eliminación de datos.

ISBN: 978-631-6557-56-8

e) **Autoridades de Protección de Datos**:

- **Superintendencia de Industria y Comercio (SIC)**: Es la autoridad principal encargada de supervisar el cumplimiento de la normativa de protección de datos. Tiene la facultad de exigir pruebas de cumplimiento a los responsables del tratamiento de datos.
- **Superintendencia Financiera de Colombia (SFS)**: Tiene competencias en la regulación de los tratamientos de datos personales en el ámbito financiero.

f) **Registro Nacional de Bases de Datos**: Es obligatorio que los responsables del tratamiento de datos notifiquen los tratamientos al Registro Nacional de Bases de Datos de la SIC.

g) **Medidas de Seguridad**: La normativa exige la aplicación de medidas de seguridad físicas, técnicas y organizativas para proteger los datos personales. Aunque no se exige un delegado de protección de datos, sí se deben tomar las medidas necesarias para garantizar la confidencialidad y seguridad de la información.

h) **Notificación de Violaciones de Seguridad**: La Ley 1581, en su artículo 17-N, establece la obligación de notificar al Departamento de Asuntos Políticos ciertos riesgos de seguridad o violaciones de las políticas de seguridad en el tratamiento de datos personales.

5.5.5. Protección de datos en Ecuador

En 2019, el Gobierno de Ecuador presentó un borrador de la Ley Orgánica de Protección de Datos Personales, inspirado en el RGPD europeo. Entre los aspectos clave del proyecto se encuentra la responsabilidad activa y comprobada de las empresas en la gestión de los datos personales. Los principios fundamentales que guían la ley incluyen, entre otros, la transparencia, el consentimiento, la confidencialidad, la legitimidad de la finalidad, la conservación y la seguridad.

Además, el proyecto propone la creación de un responsable de gestión de datos personales, quien deberá implementar medidas derivadas de un análisis de riesgos y su impacto. Los usuarios también verán asegurados sus derechos a acceder, rectificar, actualizar, eliminar, borrar, oponerse, anular o portar sus datos personales, entre otros.

ISBN: 978-631-6557-56-8

5.5.6. Protección de datos en México

La Ley Federal de Protección de Datos Personales en Posesión de los Particulares fue promulgada el 5 de julio de 2010 y entró en vigor al día siguiente. Junto a esta ley, se han emitido diversas normativas complementarias, tales como:

- El **Reglamento** de la Ley, publicado el 21 de diciembre de 2011, con vigencia desde el 22 de diciembre de 2011.
- Las **Directrices de Confidencialidad** de enero de 2013, en vigor desde abril de 2013.
- Los **Parámetros** para el Reglamento en relación con datos personales, del 29 de mayo de 2014, con vigencia a partir del 30 de mayo de 2014.

El Reglamento es aplicable a todos los tratamientos de datos personales cuando:

- El tratamiento se realiza dentro de México.
- El procesamiento se hace en nombre de un responsable de tratamiento ubicado en México, independientemente de la ubicación del tratamiento.
- La legislación mexicana es aplicable debido a la adhesión de México a convenios internacionales o la ejecución de contratos.
- El responsable del tratamiento no está en México, pero utiliza medios en territorio mexicano, salvo cuando se utilicen solo para tránsito de datos.

La normativa se enfoca en las entidades privadas y no afecta a las entidades gubernamentales ni a las sociedades de información crediticia, que están reguladas por una ley específica. Las principales autoridades encargadas de la regulación en México son el **Instituto Federal de Acceso a la Información y Protección de Datos (IFAI)** y el **Ministerio de Economía** para el sector correspondiente.

Aunque no se exige la notificación de los tratamientos de datos en todos los casos, algunos Estados federales tienen requisitos específicos. En comparación con otras regulaciones del continente, la ley mexicana sí requiere que exista un **responsable o departamento de Datos Personales** dentro de las organizaciones, tal como se establece en el artículo 30 de la Ley.

5.5.7. Protección de datos en Perú

Perú regula la protección de datos personales a través de la **Ley 29733 de Protección de Datos Personales** de 2011, la cual es de cumplimiento obligatorio para las entidades del Estado, así como para empresas, otros tipos de personas jurídicas y personas físicas. Su objetivo principal es

ISBN: 978-631-6557-56-8

garantizar, conforme a la Constitución, el derecho fundamental a la protección de los datos personales. Entre las medidas que deben cumplir las empresas se incluyen:

- Registrar los Bancos de Datos Personales (BDP) que posean.
- Obtener el consentimiento informado de los titulares de los datos personales.
- Aplicar medidas de seguridad eficaces para proteger los datos.
- Establecer procedimientos para atender los derechos de los usuarios, tales como la rectificación, la oposición, el olvido o la eliminación de datos personales.
- Comunicar el uso transfronterizo de los datos cuando sea necesario.

5.6. Disociación y anonimización

En esta sección se aborda la anonimización de datos como respuesta a las necesidades que surgen en el tratamiento de grandes volúmenes de información, especialmente en los ámbitos científico, estadístico y de marketing, y las implicaciones que ello tiene en la protección de datos personales.

Una de las herramientas clave que discutiremos es la **disociación de datos**, un proceso que permite obtener conjuntos de datos completamente anonimizados. La disociación implica separar los datos personales de aquellos que permiten identificar a los individuos, asegurando que no sea posible reidentificar a las personas a partir de los datos y, por lo tanto, estos ya no estén sujetos a la normativa de protección de datos.

Se observará también algunas **técnicas de anonimización**, así como los **riesgos asociados** a la reidentificación en datos que han sido parcialmente anonimizados. La disociación es una herramienta eficaz para realizar análisis de datos sin las restricciones impuestas por las leyes de protección de datos personales, siempre y cuando se gestionen adecuadamente los riesgos de reidentificación.

Es importante destacar que, si el riesgo de reidentificación se materializa, los datos volverán a estar sujetos a las obligaciones derivadas de la normativa de protección de datos.

En consecuencia, este conjunto de datos podría tratarse sin las restricciones impuestas por la normativa sobre datos personales, ya que estaríamos tratando datos disociados. Sin embargo, es importante destacar que los procesos de anonimización pueden eliminar información crucial para ciertos análisis. Además, un conjunto de datos que, en principio, se considera anónimo, al combinarse con otros conjuntos de datos, podría permitir la identificación de uno o más individuos.

ISBN: 978-631-6557-56-8

Por otro lado, existe un riesgo inherente a la anonimización debido a que, con el avance de la tecnología informática y la disponibilidad masiva de información, cada vez es más complicado generar conjuntos de datos auténticamente disociados a partir de datos personales. Aunque algunos autores minimizan este riesgo y consideran que es bajo, este factor debe ser cuidadosamente evaluado al aplicar cualquier técnica de anonimización.

En este sentido, es importante tener en cuenta no solo los posibles usos de los datos anonimizados, sino también el impacto potencial sobre las personas afectadas. Además, la probabilidad de que se materialice el riesgo de reidentificación debe ser valorada adecuadamente para determinar la viabilidad de la técnica de anonimización elegida.

5.6.1. Anonimización parcial

No siempre será posible lograr una disociación o anonimización completa debido a la naturaleza del procesamiento de datos. En casos donde se necesite volver a identificar a los sujetos de los datos o se necesiten detalles más granulares de los datos, lo que podría facilitar la identificación indirecta, se puede recurrir a un proceso de anonimización parcial. En estos casos, el conjunto de datos no podrá considerarse completamente anonimizado, pero se habrá dificultado la identificación de los individuos.

Para lograrlo, se pueden emplear diversas técnicas, como la pseudoanonimización, *keycoding*, *keyed hashing*, eliminación de identificadores y valores atípicos, sustitución de identificadores únicos, introducción de "ruido" y otras. La elección de la técnica más adecuada dependerá de garantizar que la identificación sea lo más difícil posible.

Dado que no se contará con información completamente anonimizada, es necesario implementar medidas de seguridad adicionales. En el contexto de España, entre estas medidas estarán aquellas relacionadas con el nivel de seguridad aplicable a los datos personales en el archivo original. Algunas de estas medidas incluyen:

- Adopción de medidas de seguridad adicionales específicas, como el cifrado.
- En el caso de la pseudoanonimización, asegurarse de que las claves que permiten vincular la información con los datos identificables hayan sido codificadas o encriptadas y almacenadas por separado.
- Incluir a un tercero de confianza cuando varias organizaciones deseen anonimizar los datos personales para un proyecto colaborativo.
- Restringir el acceso a los datos personales aplicando el principio de necesidad de saber, equilibrando los beneficios de una mayor difusión con los riesgos de divulgación

ISBN: 978-631-6557-56-8

inadvertida a personas no autorizadas. Esto podría implicar permitir acceso solo de lectura en entornos controlados o asegurar el acceso únicamente en ambientes seguros y comunidades cerradas.

5.6.2. Técnicas de anonimización

Existen diversas técnicas de anonimización, y en los últimos años, impulsadas por los riesgos asociados con la reidentificación, la comunidad científica ha mejorado o introducido nuevas metodologías. En general, el proceso de anonimización de un conjunto de datos requiere la combinación de varias de estas técnicas para mitigar las debilidades inherentes a cada una frente al riesgo de reidentificación. La selección de las técnicas más adecuadas dependerá de factores como la sensibilidad de la información, la necesidad de mantener una alta correspondencia con los datos originales para los fines previstos y el riesgo de reidentificación comprometida.

Las técnicas de anonimización pueden clasificarse según diferentes criterios, tales como su naturaleza, el efecto que tienen sobre los datos y su aplicación. A continuación se presentan algunas de estas clasificaciones:

Clasificación según su naturaleza:

- **Aleatorización**.- Modifica la veracidad de los datos, reduciendo el vínculo entre los datos y las personas, disminuyendo la probabilidad de inferencia entre los datos.
- **Generalización**.- Consiste en generalizar o diluir un valor de un atributo o columna. Por ejemplo, reducir un código postal a sus primeros dos dígitos. Esto reduce la significación de un individuo a partir de un valor de atributo único.

Clasificación según el efecto sobre los datos:

a) **Técnicas de reducción de atributos**: Estas técnicas eliminan los valores en las tablas de datos que facilitan la reidentificación:
    - Supresión de identificadores directos.
    - Agregación: Reducción del nivel de detalle de la información.
    - Muestreo: Selección de datos a partir de una amplia muestra.

b) **Técnicas de modificación de los datos**: Estas alteran los datos para reducir la posibilidad de reidentificación:
    - Generalización.
    - Adición de "ruido".
    - Asignación al azar (aleatorización de los valores).

ISBN: 978-631-6557-56-8

- Permutación o intercambio de datos.
- Privacidad diferencial.
- Supresión de datos.
- Pseudoanonimización.

c) **Métodos basados en restricción**: Introducen restricciones en el conjunto de datos para eliminar atributos que favorecen la reidentificación:

- Supresión de celdas.
- Cambio del esquema de clasificación.

Clasificación según su aplicación:

- Técnicas para reducir los riesgos de identificación en microdatos.
- Técnicas para reducir los riesgos de identificación en macrodatos.

La combinación de estas técnicas busca garantizar un adecuado equilibrio entre la utilidad de los datos y la protección de la privacidad.

5.6.3. Principios a la hora de construir un *data warehouse*

La construcción y gestión de data *warehouses* (almacenes de datos) es crucial en la Industria 4.0, especialmente cuando se trata de gestionar grandes volúmenes de datos sensibles. A medida que las empresas se enfocan en mejorar sus procesos de análisis de datos, también es importante tomar medidas para reducir la exposición a eventos de revelación de datos y proteger la privacidad de los individuos. Para lograr esto, podemos seguir un conjunto de principios que faciliten la construcción de un data *warehouse* más seguro y eficiente. Estos principios incluyen:

a) **Separación funcional.-** Este principio se refiere a la restricción de acceso a los datos según las funciones desarrolladas dentro de la organización. Es decir, solo aquellos usuarios que están autorizados y necesitan acceder a ciertos datos para cumplir con sus responsabilidades deben tener acceso a ellos. Esto limita la posibilidad de acceso no autorizado a información adicional y reduce la probabilidad de filtraciones de datos.

b) **Agregación de datos.-** La agregación de datos es una técnica fundamental en la construcción de un *data warehouse*. Al agregar los datos (por ejemplo, sumando valores, promediando o agrupando categorías), se reduce la granularidad de los mismos y se dificulta la identificación de un individuo. Esto hace más difícil que se realice una **reidentificación**, aunque no debe usarse de manera excesiva, ya que puede afectar la utilidad de los datos. Siempre es importante considerar

ISBN: 978-631-6557-56-8

que los **registros no agregados** suelen ser más fáciles de identificar que aquellos que han sido agregados.

c) **Reidentificación de los registros.-** Un paso básico para proteger los datos personales es eliminar los **identificadores directos e indirectos** siempre que sea posible. Los identificadores directos incluyen información como nombres, números de identificación o direcciones, mientras que los indirectos pueden incluir datos como fechas de nacimiento o códigos postales, que pueden llevar a la identificación de un individuo si se combinan con otros datos.

Para reforzar esta protección, **sustituir los identificadores** por **pseudónimos** puede ofrecer una capa adicional de seguridad, creando una barrera entre los datos y los individuos que los representan. Sin embargo, es crucial no tener expectativas excesivas sobre el nivel de anonimización tras la eliminación o sustitución de estos identificadores. Los datos aún podrían ser susceptibles a la **reidentificación**, especialmente si un atacante tiene acceso a un conjunto de atributos que puedan ser combinados con fuentes externas de información.

Consideraciones clave sobre la reidentificación:

- **Combinación de atributos**: A veces, incluso si los identificadores directos e indirectos se eliminan, los **atributos** restantes pueden ser lo suficientemente específicos para permitir la reidentificación de un individuo, especialmente si esos atributos se combinan con otros datos externos.
- **Nivel de anonimización**: No se debe considerar que los datos son completamente anónimos solo porque se hayan eliminado ciertos identificadores. Las técnicas de anonimización deben aplicarse de manera holística, considerando el **riesgo residual** de

5.7. Protección de datos personales en industria 4.0

El **Big Data** se basa en la creciente capacidad tecnológica para almacenar y procesar grandes volúmenes de datos, así como en la habilidad de analizar esta información para extraer su máximo valor. La capacidad de hacer análisis avanzados sobre estos grandes conjuntos de datos abre un amplio potencial para la innovación. Algunos lo comparan con el "petróleo del siglo XXI" debido a su enorme potencial para transformar industrias y mejorar la toma de decisiones.

Gracias al avance tecnológico y la facilidad de recolectar datos, Big Data permite descubrir correlaciones que anteriormente no eran evidentes. Al aplicar técnicas de análisis predictivo y modelos de *machine learning*, se pueden identificar patrones, tendencias y comportamientos que resultan en avances significativos en diversas áreas, como la salud, el marketing, y la economía.

ISBN: 978-631-6557-56-8

Esta capacidad de obtener valor de grandes volúmenes de datos representa un cambio importante en cómo se generan y utilizan los conocimientos.

Además de los avances técnicos, Big Data también trae consigo un cambio filosófico. Mientras que tradicionalmente se ha buscado comprender los fenómenos a través de relaciones de causa y efecto, ahora se prioriza la búsqueda de correlaciones. Este nuevo enfoque ofrece una forma diferente de entender el mundo, permitiendo identificar asociaciones en los datos que no necesariamente dependen de causas directas. Esto cambia la forma en que se abordan los problemas en disciplinas como el marketing, la salud y la optimización de recursos.

A pesar de las oportunidades que ofrece, Big Data también plantea desafíos, especialmente en cuanto a la **protección de la privacidad** y la **gestión ética** de la información. No obstante, con las medidas adecuadas, es posible minimizar los riesgos y maximizar las oportunidades que el análisis de grandes volúmenes de datos ofrece, transformando así la forma en que tomamos decisiones y resolvemos problemas complejos.

Hoy en día, existe un notable desconocimiento acerca del volumen real de datos almacenados por organizaciones, gobiernos y empresas, así como de los usos y finalidades que se le pueden dar a esta información. Sin embargo, hay un consenso generalizado de que este volumen aumentará de manera exponencial. La Industria 4.0 está viviendo este crecimiento del volumen de información, ya que la implementación de nuevos modelos de predicción, sistemas de robótica, realidad aumentada y otras tecnologías requiere una gran cantidad de datos. Tal como se ha mencionado en temas anteriores, muchas de las soluciones planteadas se enfocan en tratar datos personales, los cuales en muchos casos pueden ser anonimizados con mayor o menor facilidad.

En el ámbito de la Industria 4.0, varios ejemplos ilustran cómo se manejan grandes volúmenes de datos personales. Uno de ellos es ***Simble***, que ofrece soluciones de análisis para el ahorro energético y maneja datos personales de sus clientes, incluyendo tarifas asociadas al consumo energético. Otro caso es el ***SmartSpace*** de ***Ubisense***, un sistema de localización que gestiona datos de posicionamiento de usuarios tanto en interiores como en exteriores. Finalmente, el sistema de ***gestión de carretillas de Linde*** también trata datos tanto de personas como de vehículos, como parte de su solución de gestión de flotas.

La recopilación de datos de clientes o empleados siempre ha requerido la protección adecuada, pero en el caso de la Industria 4.0, la situación es aún más compleja. Las soluciones IoT, por ejemplo, no solo recopilarán datos de personas o empleados, sino que también lo harán de equipos o de procesos específicos. Un ejemplo de ello es el sistema ***Horta***, que facilita la protección de cultivos mediante estaciones meteorológicas. Los datos recopilados por estas estaciones, que

ISBN: 978-631-6557-56-8

podrían incluir información personal, deben ser tratados como si fueran datos personales. De manera similar, los datos relacionados con la proporción de recursos utilizados en procesos industriales también requieren un tratamiento especial para garantizar su protección y privacidad.

En la actualidad, existe un amplio debate sobre los riesgos que la acumulación masiva de datos representa para la protección de la privacidad y el derecho a la intimidad de las personas. Esta preocupación se acentúa en el contexto de la Industria 4.0, que no solo maneja datos personales proporcionados por individuos, sino que también interactúa con una variedad de tecnologías y sistemas (como el IoT, la robótica, y la realidad aumentada) que generan grandes volúmenes de datos frecuentemente asociados a personas o entidades. Aunque la tecnología utilizada en este campo no es cuestionada en términos de sus avances y beneficios, especialmente en áreas como la ciencia, sí se plantean interrogantes sobre el uso que se le puede dar a estos datos y las posibles repercusiones para los ciudadanos. Así, la presencia de gobiernos y empresas que poseen enormes cantidades de datos sobre los individuos genera temores no solo en relación con la privacidad, sino también con la libertad individual.

La amenaza principal se basa en la trazabilidad de nuestras actividades en línea, donde se recogen continuamente datos sobre nuestras interacciones. Esta recopilación de información puede reflejar aspectos muy privados de nuestra vida, ya que las personas tienden a tener una falsa percepción de anonimato en Internet. En este contexto, se hace posible vincular el "yo virtual" con el "yo real", lo cual, con la incorporación de tecnologías avanzadas de tratamiento de datos y fuentes de información adicionales, se convierte en una amenaza real y tangible.

Más allá de este debate "apocalíptico", la verdadera discusión debe centrarse en cómo encontrar un equilibrio que garantice tanto la **privacidad** y la **protección de datos** de los ciudadanos, como la satisfacción de las **necesidades** y los **intereses** de las empresas. Este balance es importante para asegurar que el desarrollo de tecnologías en la Industria 4.0 no se convierta en una amenaza para los derechos fundamentales de las personas.

En los siguientes códigos QR o enlaces se muestra información adicional relacionada al capítulo.

Derecho TIC	LSSI

ISBN: 978-631-6557-56-8

https://www.derechotics.com

https://lssi.digital.gob.es

ISBN: 978-631-6557-56-8

REFERENCIAS

[1] «Internet of Things: From sensing to doing», Deloitte Insights. Accedido: 18 de abril de 2025. [En línea]. Disponible en: https://www2.deloitte.com/content/www/us/en/insights/focus/tech-trends/2016/internet-of-things-iot-applications-sensing-to-doing.html

[2] «17734.jpeg (960×684)». Accedido: 18 de abril de 2025. [En línea]. Disponible en: https://cdn.statcdn.com/Infographic/images/normal/17734.jpeg

[3] T. Poleto, V. D. H. de Carvalho, y A. P. C. S. Costa, «The Roles of Big Data in the Decision-Support Process: An Empirical Investigation», en *Decision Support Systems V – Big Data Analytics for Decision Making*, B. Delibašić, J. E. Hernández, J. Papathanasiou, F. Dargam, P. Zaraté, R. Ribeiro, S. Liu, y I. Linden, Eds., Cham: Springer International Publishing, 2015, pp. 10-21. doi: 10.1007/978-3-319-18533-0_2.

[4] «Comprender los datos estructurados, semiestructurados y no estructurados», Astera. Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://www.astera.com/es/type/blog/structured-semi-structured-and-unstructured-data/

[5] «IRJET-V4I957.pdf». Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://www.irjet.net/archives/V4/i9/IRJET-V4I957.pdf

[6] F.-È. Bordeleau, E. Mosconi, y L. A. Santa-Eulalia, *Business Intelligence in Industry 4.0: State of the art and research opportunities*. 2018. Accedido: 19 de abril de 2025. [En línea]. Disponible en: http://hdl.handle.net/10125/50383

[7] A. Woodie, «Big Data Challenges of Industry 4.0», BigDATAwire. Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://www.bigdatawire.com/2019/04/25/big-data-challenges-of-industry-4-0/

[8] M. Sachdev, «The Role of Big Data Analytics in Industry 4.0». Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://blog.rgbsi.com/big-data-analytics-in-industry-4.0

[9] D. S. Moore, G. P. McCabe, y B. A. Craig, *Introduction to the practice of statistics*, Ninth edition. New York: W.H. Freeman, Macmillan Learning, 2017.

[10] «Precisión no es lo mismo que exactitud (Gráfico)». Accedido: 19 de abril de 2025. [En línea]. Disponible en: http://www.fogonazos.es/2014/05/precision-no-es-lo-mismo-que-exactitud.html

[11] J. M. Fernández y C. M. Abril, «ESTADÍSTICA BÁSICA PARA CIENCIAS DE LA SALUD».

[12] S. M. Stigler, «Gauss and the Invention of Least Squares», *Ann. Stat.*, vol. 9, n.º 3, pp. 465-474, may 1981, doi: 10.1214/aos/1176345451.

[13] «Prueba de los rangos con signo de Wilcoxon». Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://cienciadedatos.net/documentos/18_prueba_de_los_rangos_con_signo_de_wilcoxon

[14] «Test de Friedman». Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://cienciadedatos.net/documentos/21_friedman_test

[15] Á. E. Talaya y A. M. Collado, *Investigación de Mercados*. ESIC Editorial, 2014.

[16] A. Mendez, «Ejemplos de Gráficas de Barras - Ejercicios y Problemas Resueltos», Plan de Mejora. Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://www.plandemejora.com/ejemplos-grafica-barras/

[17] P. Rodó, «Diagrama de sectores», Economipedia. Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://economipedia.com/definiciones/diagrama-de-sectores.html

ISBN: 978-631-6557-56-8

[18] «Infografía GDAMS 2020: "Invirtamos en salud el gasto militar"», Delas. Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://centredelas.org/publicacions/infografia-gdams-2020-invirtamos-en-salud-el-gasto-militar/?lang=es

[19] exceltotal, «Cómo hacer un histograma en Excel», Excel Total. Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://exceltotal.com/como-hacer-un-histograma-en-excel/

[20] «Polígonos de frecuencia (artículo)», Khan Academy. Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://es.khanacademy.org/math/2-secundaria-pe/xf4e5558599a475b6:probabilidad-y-estadistica-2sec/xf4e5558599a475b6:preguntas-estadisticas-representacion-de-datos-a-traves-de-histogramas-y-poligonos-de-frecuencia/a/81817-artculo-polgonos-de-frecuencia

[21] M. at E. Wikipedia, *English: I created this file myself using Microsoft Excel and Microsoft PowerPoint. The data is purely hypothetical.* 2006. Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://commons.wikimedia.org/w/index.php?curid=6199854

[22] «Gráfico de Burbujas». Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://datavizcatalogue.com/ES/metodos/grafico_de_burbujas.html

[23] «Graph templates for all types of graphs - Origin scientific graphing». Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://www.originlab.com/index.aspx?go=products/origin/graphing

[24] OnkelDagobert, *Deutsch: Balkendiagramm und Boxplot und was man aus ihnen herauslesen kann.* 2009. Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://commons.wikimedia.org/wiki/File:Statistik_boxplot_balken.jpg

[25] G. Rivero, «Mapamundi de temperaturas medias | Gustavo Rivero». Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://gustavorivero.com/mapamundi-de-temperaturas-medias/

[26] «¿Qué es un gráfico Marimekko?», Jaspersoft. Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://www.jaspersoft.com/es/articles/what-is-a-marimekko-chart

[27] M. E. Pérez-Pons, A. González-Briones, y J. M. Corchado, «Towards financial valuation in data-driven companies», *Orient. J. Comput. Sci. Technol.*, vol. 12, n.º 2, pp. 28-33, jun. 2019.

[28] J. Wang, M. Qiu, y B. Guo, «Enabling real-time information service on telehealth system over cloud-based big data platform», *J. Syst. Archit.*, vol. 72, pp. 69-79, ene. 2017, doi: 10.1016/j.sysarc.2016.05.003.

[29] J. Luengo, D. García-Gil, S. Ramírez-Gallego, S. García, y F. Herrera, «Final Thoughts: From Big Data to Smart Data», en *Big Data Preprocessing: Enabling Smart Data*, J. Luengo, D. García-Gil, S. Ramírez-Gallego, S. García, y F. Herrera, Eds., Cham: Springer International Publishing, 2020, pp. 183-186. doi: 10.1007/978-3-030-39105-8_10.

[30] M. Jarke, M. A. Jeusfeld, C. Quix, y P. Vassiliadis, «Architecture and quality in data warehouses: An extended repository approach», *Inf. Syst.*, vol. 24, n.º 3, pp. 229-253, may 1999, doi: 10.1016/S0306-4379(99)00017-4.

[31] «Intégration de données/Les principales approches d'intégration de données — Wikiversité». Accedido: 20 de abril de 2025. [En línea]. Disponible en: https://fr.wikiversity.org/wiki/Int%C3%A9gration_de_donn%C3%A9es/Les_principales_approches_d%27int%C3%A9gration_de_donn%C3%A9es

[32] P. P. Khine y Z. S. Wang, «Data lake: a new ideology in big data era», *ITM Web Conf.*, vol. 17, p. 03025, 2018, doi: 10.1051/itmconf/20181703025.

ISBN: 978-631-6557-56-8

[33] Divakar Mysore, Shrikant Khupat, Shweta Jain, y Apache Cassandra, Dynamo, Voldemort y Riak, «Understanding the architectural layers of a big data solution», IBM Developer. Accedido: 20 de abril de 2025. [En línea]. Disponible en: https://developer.ibm.com/articles/bd-archpatterns3/

[34] T. R. Academy, «Big Data Layers – Data Source, Ingestion, Manage and Analyze Layer - RCV Academy». Accedido: 20 de abril de 2025. [En línea]. Disponible en: https://www.rcvacademy.com/big-data-layers/

[35] «IoT in the Manufacturing Industry Enabling Industry 4.0 - Wipro». Accedido: 20 de abril de 2025. [En línea]. Disponible en: https://www.wipro.com/engineering/iot-in-the-manufacturing-industry-enabling-industry-4-0/

[36] «Clicking Clean», Greenpeace International. Accedido: 20 de abril de 2025. [En línea]. Disponible en: https://www.greenpeace.org/international/publication/6826/clicking-clean-2017/

[37] «Gartner Says Global IT Spending to Grow 3.7% in 2020», Gartner. Accedido: 20 de abril de 2025. [En línea]. Disponible en: https://www.gartner.com/en/newsroom/press-releases/2019-10-23-gartner-says-global-it-spending-to-grow-3point7-percent-in-2020

[38] C. Harvey, «AWS vs Azure vs Google Cloud: Top Cloud Provider Comparison», Datamation. Accedido: 20 de abril de 2025. [En línea]. Disponible en: https://www.datamation.com/cloud/aws-vs-azure-vs-google-cloud/

[39] «Case Study: How Google Cloud Made Our Industry 4.0 Platform Even Better - Oden Technologies». Accedido: 20 de abril de 2025. [En línea]. Disponible en: https://oden.io/case-study-how-google-cloud-made-our-industry-4-0-software-better/

[40] «Grupo Bimbo transforma el análisis de datos de su área comercial, gracias a las soluciones de Microsoft – Centro de noticias». Accedido: 20 de abril de 2025. [En línea]. Disponible en: https://news.microsoft.com/es-es/2022/11/02/grupo-bimbo-transforma-el-analisis-de-datos-de-su-area-comercial-gracias-a-las-soluciones-de-microsoft/

[41] «Lab 1 - Installing HDP Sandbox», HackMD. Accedido: 21 de abril de 2025. [En línea]. Disponible en: https://hackmd.io/@firasj/BkSQJQ8eh

[42] «Google Research Publication: The Google File System». Accedido: 21 de abril de 2025. [En línea]. Disponible en: https://research.google.com/archive/gfs.html

[43] O. Fernandez, «¿Qué es HDFS? Introducción 2025», Aprender BIG DATA. Accedido: 21 de abril de 2025. [En línea]. Disponible en: https://aprenderbigdata.com/hdfs/

[44] O. Fernandez, «¿Qué es Hadoop MapReduce? Introducción», Aprender BIG DATA. Accedido: 21 de abril de 2025. [En línea]. Disponible en: https://aprenderbigdata.com/hadoop-mapreduce/

[45] DataWorks Summit, *Open Source Data Management for Industry 4.0*, (18 de abril de 2019). Accedido: 21 de abril de 2025. [En línea Video]. Disponible en: https://www.youtube.com/watch?v=hVxF5j2w4ZA

[46] Á. Rayón, «La aplicación del Big Data y Business Intelligence en la creación de valor para el cliente», Deusto Data. Accedido: 21 de abril de 2025. [En línea]. Disponible en: https://blogs.deusto.es/bigdata/la-aplicacion-del-big-data-y-business-intelligence-en-la-creacion-de-valor-para-el-cliente/

[47] A. Kumar, Shankar ,Ravi, Choudhary ,Alok, y L. S. and Thakur, «A big data MapReduce framework for fault diagnosis in cloud-based manufacturing», *Int. J. Prod. Res.*, vol. 54, n.º 23, pp. 7060-7073, dic. 2016, doi: 10.1080/00207543.2016.1153166.

[48] S. Pouyanfar *et al.*, «A Survey on Deep Learning: Algorithms, Techniques, and Applications», *ACM Comput Surv*, vol. 51, n.º 5, p. 92:1-92:36, sep. 2018, doi: 10.1145/3234150.

[49] S. Rogers y M. Girolami, *A first course in Machine Learning*, vol. 1. CRC Press, 2012.

[50] D. A. N. Venkatesh, «Industry 4.0: Reimagining the Future of Workplace (Five Business Case Applications of Artificial Intelligence, Machine Learning, Robots, Virtual Reality in Five Different Industries)», 19 de diciembre de 2018, *Social Science Research Network, Rochester, NY*: 3303732. Accedido: 21 de abril de 2025. [En línea]. Disponible en: https://papers.ssrn.com/abstract=3303732

[51] S. Afaq Ali Shah, M. Bennamoun, y F. Boussaid, «Iterative deep learning for image set based face and object recognition», *Neurocomputing*, vol. 174, pp. 866-874, ene. 2016, doi: 10.1016/j.neucom.2015.10.004.

[52] «Energy Efficiency in Public Buildings through Context-Aware Social Computing». Accedido: 21 de abril de 2025. [En línea]. Disponible en: https://www.mdpi.com/1424-8220/17/4/826

[53] T. M. Mitchell, «Artificial Neural Networks».

[54] «Mining association rules between sets of items in large databases | ACM SIGMOD Record». Accedido: 21 de abril de 2025. [En línea]. Disponible en: https://dl.acm.org/doi/10.1145/170036.170072

[55] G. Y. Alcantara, «¡Tenemos mucho en común!: Introducción al Cluster Analysis con R», Medium. Accedido: 21 de abril de 2025. [En línea]. Disponible en: https://medium.com/@gloria.yantas.alcantara/tenemos-mucho-en-com%C3%BAn-introducci%C3%B3n-al-cluster-analysis-con-r-c8fdedab7206

[56] M. Montaner, B. López, y J. L. de la Rosa, «A Taxonomy of Recommender Agents on the Internet», *Artif. Intell. Rev.*, vol. 19, n.º 4, pp. 285-330, jun. 2003, doi: 10.1023/A:1022850703159.

[57] «Accenture: "Hemos implantado 22.000 robots sin perder un solo empleo"», Expansión.com. Accedido: 21 de abril de 2025. [En línea]. Disponible en: https://www.expansion.com/economia-digital/innovacion/2017/10/06/59d64e7f46163fd5118b467a.html

[58] «La Industria 4.0 y la educación - Comunidad Virtual Externadista». Accedido: 24 de abril de 2025. [En línea]. Disponible en: https://micomunidadvirtual.uexternado.edu.co/la-industria-4-0-y-la-educacion/

[59] «Compromiso con la Industria 4.0 - BigdataBP». Accedido: 24 de abril de 2025. [En línea]. Disponible en: https://www.bigdatabp.com/compromiso-con-la-industria-4-0/

[60] «Visualización de la información: De los datos al conocimiento - Ignasi Alcalde Perea - Google Libros». Accedido: 24 de abril de 2025. [En línea]. Disponible en: https://books.google.com.ec/books?id=BpSnDAAAQBAJ&printsec=frontcover&source=gbs_atb#v=onepage&q&f=false

[61] «El cuarteto de Anscombe», Carlos J. Gil Bellosta. Accedido: 24 de abril de 2025. [En línea]. Disponible en: https://www.datanalytics.com/2013/08/30/el-cuarteto-de-anscombe/

[62] L. Sánchez, «Fundamentos del Big Data en la Industria 4.0». UNIR, 2021.

[63] M. Google, «Ubicación de la ciudad de Riobamba, Ecuador», 24 de abril de 2025. [En línea]. Disponible en: https://www.google.com/maps/dir/-1.6561032,-78.6747574/Riobamba/@-1.6594547,-78.6871011,14z/data=!4m10!4m9!1m1!4e1!1m5!1m1!1s0x91d3a8255b072981:0xcb8509cd0a3fdf99!2m2!1d-78.6588786!2d-1.6650227!3e2?entry=ttu&g_ep=EgoyMDI1MDQyMS4wIKXMDSoASAFQAw%3D%3D

[64] S. Manifold, «Ubisense Contact Tracing», Ubisense. Accedido: 24 de abril de 2025. [En línea]. Disponible en: https://ubisense.com/ubisense-contact-tracing/

ISBN: 978-631-6557-56-8

[65] ESTRATEGIA COMPETITIVA Técnicas para el análisis de los sectores y de la competencia». Accedido: 24 de abril de 2025. [En línea]. Disponible en: https://www.academia.edu/24621661/ESTRATEGIA_COMPETITIVA_T%C3%A9cnicas_para_el_an%C3%A1lisis_de_los_sectores_y_de_la_competencia

[66] F.-È. Bordeleau, E. Mosconi, y L. A. Santa-Eulalia, *Business Intelligence in Industry 4.0: State of the art and research opportunities*. 2018. Accedido: 24 de abril de 2025. [En línea]. Disponible en: http://hdl.handle.net/10125/50383

[67] «Privacy Program | U.S. Department of Commerce». Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://www.commerce.gov/opog/privacy-program

[68] «Ley Orgánica de Protección de Datos - LOPDGDD 3/2018 | Grupo Atico34». Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://protecciondatos-lopd.com/empresas/nueva-ley-proteccion-datos-2018/

[69] M. P. Ph.D, «Cómo afecta la nueva Ley de Protección de Datos 2018 • Montse Peñarroya», Montse Peñarroya. Accedido: 19 de abril de 2025. [En línea]. Disponible en: https://www.montsepenarroya.com/ley-de-proteccion-de-datos/

ISBN: 978-631-6557-56-8

ISBN: 978-631-6557-56-8

SEMBLANZA DE AUTORES

Diego Alejandro García Saraguro

• Ingeniero en Electrónica, Telecomunicaciones y Redes de la Escuela Superior Politécnica de Chimborazo. • Máster Universitario en Ingeniería del Software y Sistemas Informáticos por Universidad Internacional de la Rioja. • Máster en Estadística Aplicada por la Universidad Politécnica Estatal del Carchi.	• Técnico digitalizador del GADM – Riobamba, por 3 años • Técnico Docente Universitario por 2 años de la ESPOCH. • Profesor ocasional de la Escuela Superior Politécnica de Chimborazo • Autor de artículos científicos de alto impacto.
diego.garcia@espoch.edu.ec	
https://orcid.org/0009-0008-7792-3969	
Profesor en la Escuela Superior Politécnica de Chimborazo	

ISBN: 978-631-6557-56-8

María Gabriela Arias Garnica

Estudios	**Experiencia**
• Ingeniera Agrónoma por la Escuela Superior Politécnica de Chimborazo. • Magister en Cadenas Productivas Agroindustriales por la Universidad Nacional de Chimborazo. • Máster en Estadística Aplicada por la Universidad Politécnica del Carchi.	• Técnico Docente Universitario por 2 años de la ESPOCH. • Gerente Administrativa de la Empresa DISARPE CIA.LTDA. • Gerente propietario de la Empresa de Insumos Agrícolas INSUAGRO. • Autora de artículos científicos de alto impacto.
mariag.arias@espoch.edu.ec	
https://orcid.org/0009-0002-2535-9776	
Técnico Docente en Escuela Superior Politécnica de Chimborazo	

ISBN: 978-631-6557-56-8

Hernán Patricio Moyano Ayala

Estudios	**Experiencia**
• Máster en Ingeniería Civil: Estructuras y Construcciones - Universidad da Beira Interior. • Especialista en Gestión de Servicios Empresariales enfocados en la Construcción – ILAC, Toronto. • Ingeniero Civil – Pontificia Universidad Católica del Ecuador.	• Profesor de Topografía e Ingeniería de Caminos de la Facultad de Recursos Naturales ESPOCH. • Técnico Docente impartiendo principalmente clases de Álgebra Superior, Geometría y Trigonometría y Dibujo Técnico en el Centro de Admisión de Nivelación y en la Facultad de Informática y Electrónica ESPOCH. • Técnico Docente encargado de los laboratorios de Resistencia de Materiales, Térmica, y Mecánica de Fluidos en la Facultad de Mecánica ESPOCH.
patricio.moyano@espoch.edu.ec	
ORDID: 0009-0008-4856-3925	
Profesor en la Escuela Superior Politécnica del Ecuador	

PUERTO MADERO
EDITORIAL

ISBN 978-631-6557-56-8