Módulo 1. Introducción al big data

1. Ecosistemas de datos y big data

Cada día se generan más de **402 millones de *terabytes* de datos** en el mundo (Bartley, 2025). Cada búsqueda que hacemos en internet, cada paso que registran nuestros teléfonos, cada transacción con tarjeta, foto subida a redes sociales, historia visualizada, sensor que mide la temperatura o el tránsito, mensaje enviado o producto escaneado, se transforma en información que circula, se almacena y se analiza. Estas acciones cotidianas, muchas veces automáticas, forman parte de un entramado digital que crece a una velocidad sin precedentes y que modifica la manera en que las personas, las organizaciones y los gobiernos toman decisiones.

Ese conjunto de interacciones y registros digitales es lo que hoy se conoce como ***big data***: un conjunto de tecnologías, procesos y capacidades que permiten trabajar con volúmenes de información masivos y diversos. Su desarrollo ha abierto nuevas posibilidades para entender fenómenos complejos, anticipar comportamientos y generar soluciones innovadoras en campos tan diversos como la salud, el transporte, el comercio, la

educación o la planificación urbana. Comprender qué es y cómo funciona el *big data* resulta indispensable para interpretar los entornos actuales de producción y uso de la información.

En esta unidad abordaremos este fenómeno desde una mirada técnica y contextual. Se repasará la evolución del tratamiento de datos en la era digital, se presentarán las características propias del *big data* —como el volumen, la velocidad, la variedad, la veracidad y el valor—, y se explorarán los modelos de almacenamiento y procesamiento masivo, junto con herramientas ampliamente utilizadas como Hadoop, Spark y Kafka. A lo largo de este recorrido, se busca comprender cómo se articulan los componentes técnicos y conceptuales que intervienen en la gestión de datos a gran escala.

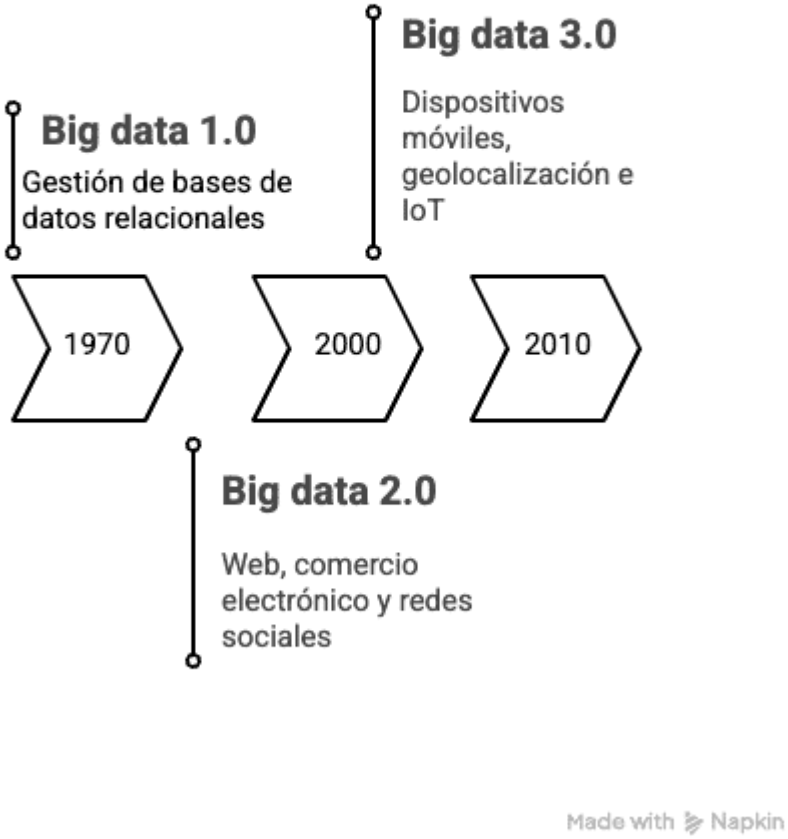
De los datos analógicos al entorno digital

El análisis de datos acompaña a las sociedades desde tiempos remotos. Civilizaciones como la egipcia o la romana registraban y sistematizaban información para organizar actividades y tomar decisiones. Con el paso del tiempo, esas prácticas fueron evolucionando, primero con registros físicos organizados en papel, luego con sistemas informatizados, hasta llegar a los entornos digitales actuales. En ese recorrido, el volumen, la velocidad y la diversidad de los datos cambiaron radicalmente. El crecimiento de internet, los dispositivos conectados y las

plataformas digitales hicieron que la producción de información alcanzara niveles sin precedentes, dando lugar a lo que hoy conocemos como *big data*.

El término *big data* comenzó a circular en la década de 1990, y se atribuye su popularización a John R. Mashey, entonces en la empresa Silicon Graphics. Su significado, sin embargo, fue moldeándose a medida que se transformaban los procesos de almacenamiento y análisis de información. Para comprender cómo se llegó a este escenario, es útil reconocer tres fases principales que marcaron la evolución del *big data* (Big Data Framework, 2025). Estas etapas permiten identificar cambios tecnológicos, nuevas necesidades analíticas y modificaciones en las formas de producir conocimiento a partir de los datos.

Figura 1. Fases del desarrollo del *big data*



Fuente: elaboración propia con base en Big Data Framework, 2025

La primera fase, que comenzó en la década de 1970, se basa en las tecnologías tradicionales de gestión de bases de datos. Durante esta etapa, el almacenamiento y el análisis se centraban en datos estructurados, organizados en sistemas relacionales (*RDBMS*). Estas herramientas ofrecían una estructura estable para realizar consultas, generar informes y construir análisis básicos. Técnicas como SQL, procesamiento analítico en línea y reportes

tabulares definieron el funcionamiento de esta etapa, que permitió consolidar prácticas de gestión informativa ampliamente utilizadas en sectores administrativos, financieros y operativos.

La segunda fase, iniciada a principios de los años 2000, estuvo marcada por la explosión de internet y el auge del comercio electrónico. Empresas como Amazon, Yahoo y eBay comenzaron a recolectar y analizar información generada por la navegación, las búsquedas y las compras online. Esta etapa introdujo el manejo de datos semiestructurados y no estructurados, como registros webs, ubicaciones IP, clics y comentarios en redes sociales. Esta expansión generó la necesidad de desarrollar nuevas tecnologías capaces de almacenar y procesar este tipo de información, lo que motivó la creación de sistemas distribuidos y enfoques analíticos más flexibles (Big Data Framework, 2025).

La tercera fase, que se inicia alrededor del año 2010, se vincula con la masificación de los dispositivos móviles y el desarrollo del Internet de las Cosas (*IoT*). A través de teléfonos inteligentes, sensores, dispositivos portátiles y electrodomésticos conectados, se genera información constante sobre comportamiento físico, movimientos, ubicaciones y hasta variables de salud. Estos datos permitieron ampliar las posibilidades de análisis a nuevos sectores, como la movilidad urbana, el transporte inteligente, la gestión energética y la salud digital. La información ya no solo se recolecta en entornos digitales tradicionales, sino que proviene

de la interacción cotidiana con objetos y espacios físicos conectados.

En este recorrido, el *big data* se transformó en un fenómeno técnico y social que excede el almacenamiento masivo. Cada una de sus fases aportó capacidades distintas y nuevos desafíos, desde las bases relacionales hasta los sistemas distribuidos e inteligentes. Esta evolución permite entender cómo se configuran hoy los ecosistemas informacionales en los que trabajamos, tomamos decisiones y desarrollamos soluciones.

Concepto y características del *big data*: volumen, velocidad, variedad, veracidad y valor

El concepto de *big data* surge como respuesta al crecimiento exponencial de la información en circulación. A diferencia de los sistemas de datos tradicionales, que operaban con estructuras bien definidas y volúmenes controlables, los entornos actuales exigen capacidades de procesamiento y análisis que puedan adaptarse a la magnitud, diversidad y dinamismo de los datos generados. Esta transformación tiene su origen en las fases evolutivas ya descritas: desde las bases relacionales de los años 70 hasta los entornos conectados y móviles actuales, se fue gestando un ecosistema donde los datos se convirtieron en materia prima para la toma de decisiones estratégicas.

Big data se refiere a conjuntos de datos masivos y complejos que no pueden ser tratados eficazmente con herramientas convencionales. Se trata de información que puede provenir de múltiples fuentes —como redes sociales, sensores, sistemas financieros o dispositivos móviles— y que, al ser recopilada, gestionada y analizada con enfoques adecuados, permite generar conocimientos valiosos y accionables (Badman y Kosinski, 2024). Este fenómeno se consolidó en paralelo con el avance de tecnologías digitales e impulsó la transformación de sectores como el comercio, la salud, la logística o la educación, al ofrecer nuevas capacidades para anticipar comportamientos, optimizar procesos o personalizar servicios.

A pesar de que existen similitudes en el propósito final —analizar para comprender y decidir—, los datos tradicionales y el *big data* presentan diferencias marcadas en cuanto a volumen, estructura y métodos analíticos. Mientras que los primeros operan sobre conjuntos reducidos y organizados de forma estructurada, el *big data* abarca grandes volúmenes de datos en múltiples formatos, que requieren soluciones avanzadas de análisis y procesamiento distribuido.

Tabla 1. Diferencias entre datos tradicionales y *big data*

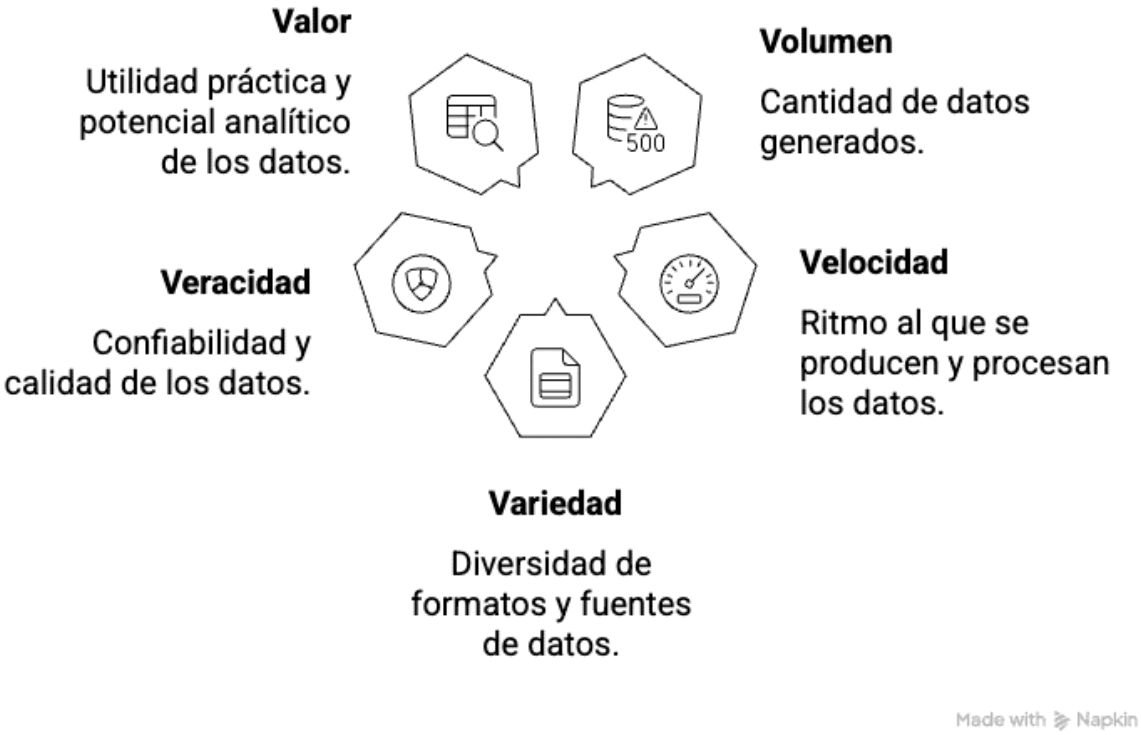
Dimensión	Datos tradicionales	*Big data*
Volumen	Bajo a medio	Muy alto (terabytes a petabytes)
Formato	Estructurado	Estructurado, semiestructurado y no estructurado
Velocidad	Carga y análisis periódicos	Generación y análisis en tiempo real
Infraestructura	Servidores centralizados	Procesamiento distribuido en la nube

Fuente: elaboración propia con base en Badman y Kosinski, 2024

Una forma común de caracterizar al *big data* es a través del enfoque de las 5V: **volumen, velocidad, variedad, veracidad y valor.** Estas dimensiones permiten describir qué lo hace diferente respecto de otros sistemas de información y por qué requiere enfoques específicos de gestión. Cada una aporta una perspectiva particular sobre los desafíos que presenta el manejo de grandes volúmenes de datos, y en conjunto configuran las

condiciones técnicas y organizacionales necesarias para operar eficazmente en estos entornos.

Figura 2. Las 5V del *big data*



Fuente: elaboración propia

Veamos, a continuación, que implica cada una de estas dimensiones:

Volumen

Uno de los rasgos más evidentes del ***big data*** es la magnitud de los datos que se generan constantemente. Estos pueden originarse en registros de transacciones, sensores, aplicaciones, plataformas digitales, entre otros. Este volumen creciente supera con facilidad las capacidades de los sistemas tradicionales de almacenamiento, por lo que requiere arquitecturas flexibles y escalables, muchas veces basadas en la nube. La gestión eficiente de estos volúmenes garantiza que la información no quede atrapada ni fragmentada, y que pueda ser analizada sin pérdida de contexto (Badman y Kosinski, 2024).

Velocidad

La información en entornos digitales circula con una rapidez que obliga a pensar en el procesamiento en tiempo real. Desde redes sociales hasta sistemas financieros, los datos se generan y requieren análisis inmediato. Esto habilita una toma de decisiones más ágil, pero también implica desafíos técnicos importantes. Las herramientas de procesamiento en flujo y los sistemas en memoria permiten capturar, procesar y actuar sobre los datos en movimiento, reduciendo los tiempos entre la recolección y la acción.

Variedad

En el ***big data***, los datos pueden presentarse en múltiples formatos. Pueden ser textos, imágenes, sonidos, videos, señales, formularios o registros abiertos. Esta diversidad exige herramientas que permitan reunir e interpretar datos muy distintos entre sí. Para lograrlo, es necesario

contar con sistemas flexibles, que puedan reconocer estos formatos y permitir que se conecten entre sí. Gestionar esta variedad implica —además de un almacenamiento adecuadamente— asegurar que esa información pueda organizarse, relacionarse y analizarse de forma útil.

Veracidad

La confianza en los datos es una condición para generar análisis significativos. Sin embargo, al provenir de múltiples fuentes, los datos pueden contener errores, ruido o ambigüedades. La veracidad apunta a la calidad de los datos: su precisión, consistencia y relevancia. Esto requiere procesos de validación, limpieza y verificación sistemática, que permitan filtrar información inexacta y evitar conclusiones erróneas.

Valor

Finalmente, el propósito del ***big data*** es generar conocimiento útil. La capacidad de convertir grandes volúmenes de datos en información significativa es lo que define su impacto. Este valor puede expresarse en múltiples formas: optimización de recursos, mejora de servicios, innovación de productos, prevención de riesgos, entre otros. La analítica avanzada, el ***machine learning*** y la inteligencia artificial permiten transformar conjuntos de datos en modelos predictivos o en indicadores accionables, integrando así la información a los procesos de toma de decisiones.

Estas características muestran que el trabajo con *big data* requiere infraestructuras pensadas para responder a nuevas demandas. Por eso, surgieron modelos de almacenamiento y procesamiento capaces de adaptarse a estas condiciones y escalar según las necesidades de los datos.

Modelos de almacenamiento y procesamiento masivo

Para poder trabajar con grandes volúmenes de datos, no alcanza con usar las herramientas tradicionales. A medida que los datos crecieron en cantidad, complejidad y velocidad, fue necesario desarrollar nuevas formas de almacenarlos y procesarlos. Estos modelos permiten organizar la información, hacerla accesible y, sobre todo, analizarla de manera eficiente según las necesidades de cada organización.

A continuación, se presentan los principales modelos de almacenamiento y procesamiento utilizados en contextos donde se trabaja con *big data*. Cada uno responde a distintos tipos de demandas y se adapta a situaciones concretas.

Tabla 2. Modelos de almacenamiento y procesamiento en entornos de *big data*

Modelo	¿Para qué sirve?
Servidores centralizados	Almacenar y procesar datos en una única computadora o servidor.
Sistemas distribuidos	Repartir los datos entre varios equipos que trabajan en conjunto.
Procesamiento por lotes	Analizar grandes cantidades de datos agrupados en bloques.
Procesamiento en tiempo real	Analizar los datos al mismo tiempo que se generan.
Almacenamiento en la nube	Guardar datos y acceder a ellos desde internet.
Arquitectura desacoplada	Separar el lugar donde se almacenan los datos del lugar donde se procesan.

Fuente: elaboración propia

Estos modelos se aplican de manera práctica en diferentes contextos. Por ejemplo, empresas de comercio electrónico o plataformas de *streaming* combinan almacenamiento distribuido y procesamiento en tiempo real para ofrecer servicios inmediatos

y fiables a millones de usuarios. La combinación de modelos permite que los datos se muevan y procesen sin interrupciones, optimizando la experiencia del usuario y la eficiencia operativa.

En otros casos, organizaciones con flujos de información que no requieren decisiones inmediatas aprovechan el procesamiento por lotes para analizar grandes volúmenes históricos y generar reportes periódicos. Por ejemplo, un banco puede revisar todas las transacciones de la semana para elaborar informes financieros o detectar patrones de comportamiento de sus clientes, optimizando sus decisiones sin necesidad de procesar cada operación en tiempo real. Para gestionar estos volúmenes de datos de manera eficiente y permitir que distintos equipos accedan a ellos según sus necesidades, se recurre a soluciones de almacenamiento más flexibles y escalables.

El almacenamiento en la nube y las arquitecturas desacopladas ofrecen, justamente, esa flexibilidad y escalabilidad. Permiten que distintas áreas de una empresa trabajen con los mismos datos, utilizando herramientas diferentes según sus necesidades, sin depender de un único sistema físico. Esto facilita la integración de análisis avanzados y modelos predictivos, potenciando la toma de decisiones basada en información actualizada y confiable.

Estos modelos de almacenamiento y procesamiento masivo establecen las bases para trabajar con grandes volúmenes de datos de manera eficiente y confiable. Sobre estas infraestructuras se apoyan las herramientas y tecnologías más utilizadas en entornos *big data*, como Hadoop, Spark y Kafka, que permiten almacenar, procesar y analizar información masiva en distintos formatos y tiempos. Comprender cómo funcionan los modelos antes de explorar estas tecnologías facilita interpretar su aplicación práctica y aprovechar al máximo sus capacidades en escenarios reales.

Herramientas y tecnologías utilizadas en entornos *big data*: Hadoop, Spark, Kafka

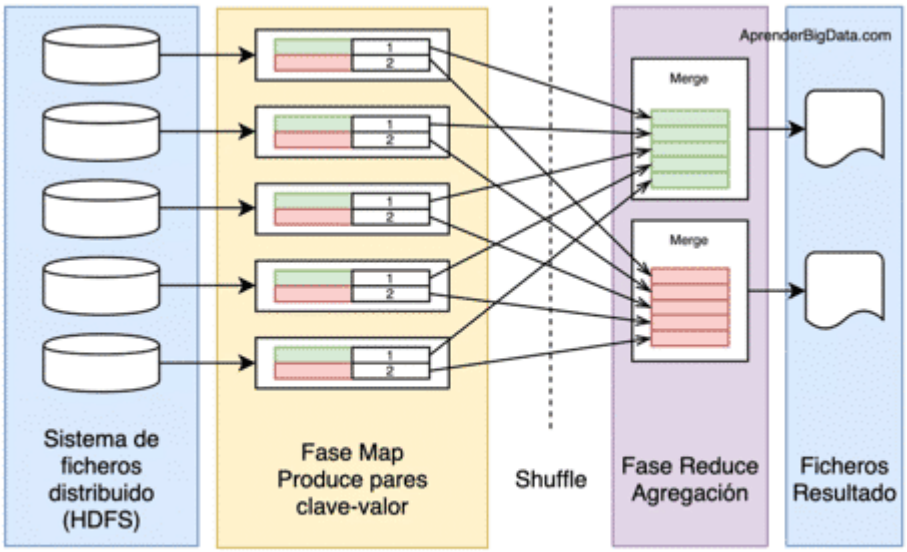
En los entornos *big data*, las herramientas tecnológicas permiten procesar, mover y almacenar cantidades de información que superan ampliamente las capacidades de los sistemas convencionales. Estas plataformas están diseñadas para operar en arquitecturas distribuidas, gestionar flujos de datos heterogéneos y ejecutar tareas analíticas en tiempos optimizados. Hadoop, Spark y Kafka son tres de las soluciones más extendidas en este campo, cada una con funcionalidades específicas que responden a diferentes necesidades del ecosistema de datos.

Hadoop es una tecnología que permite almacenar y procesar grandes volúmenes de información a través de una red de computadoras conectadas. Su funcionamiento se apoya en dos componentes principales: un sistema de archivos distribuido (HDFS) y un modelo de procesamiento denominado ***MapReduce***. El primero divide los datos en fragmentos que se distribuyen entre distintos nodos, garantizando tolerancia a fallos y eficiencia en el acceso; el segundo se encarga de ejecutar tareas en paralelo, organizando la información en pares de valores y claves, para luego agrupar y combinar esos resultados.

Esta lógica puede observarse en el siguiente esquema, que resume las etapas del proceso: primero, el sistema HDFS fragmenta los datos y los distribuye; luego, la fase ***Map*** organiza los datos en pares clave-valor; finalmente, la fase Reduce agrega esos resultados y los transforma en archivos útiles para su análisis posterior.

Entonces, supongamos que una empresa de telecomunicaciones necesita analizar millones de registros de llamadas que recibe diariamente para detectar fraudes. Estos datos se almacenan en el sistema HDFS, donde son divididos y distribuidos entre distintos servidores. A continuación, entra en juego la fase Map, que identifica pares clave-valor, por ejemplo, número de origen y duración de la llamada. Luego, la fase ***Reduce*** agrupa estos valores y calcula, por ejemplo, cuántas llamadas sospechosas provienen de un mismo número. Finalmente, los resultados consolidados se guardan como archivos accesibles para el equipo de seguridad, que puede actuar rápidamente ante comportamientos anómalos. Este flujo permite que tareas complejas se ejecuten de forma distribuida y eficiente, incluso sobre millones de registros diarios.

Figura 3. Funcionamiento de Hadoop: HDFS, Map y Reduce



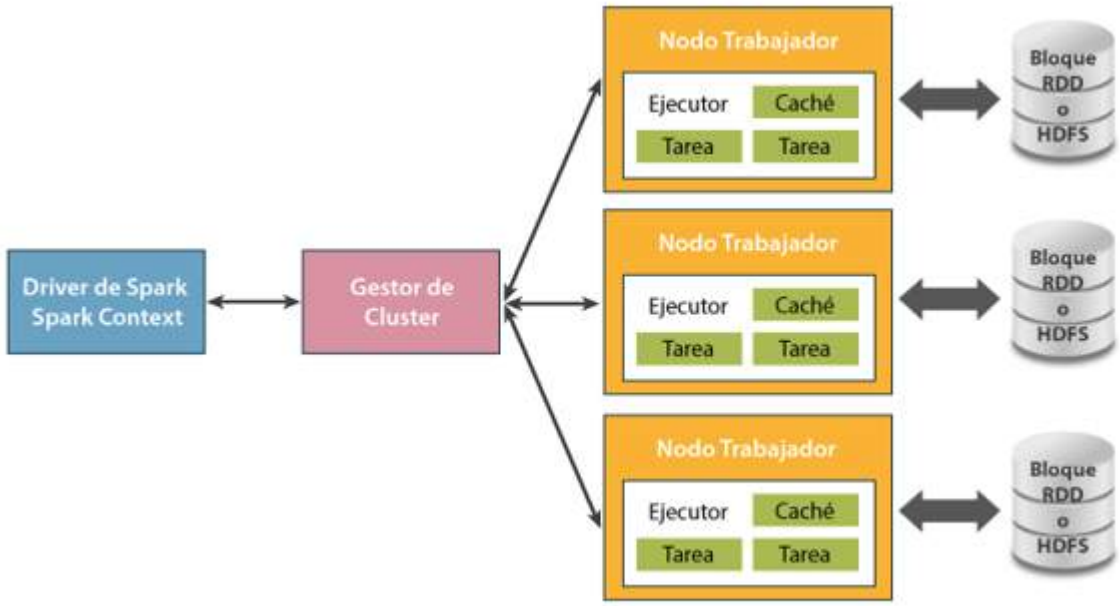
Fuente: Fernández, 2025, https://goo.su/g4OW5

Spark es una tecnología diseñada para el procesamiento de datos a gran velocidad sobre múltiples nodos. Su arquitectura distribuye las tareas entre distintos componentes: el ***driver***, que coordina la ejecución; el ***gestor de clúster***, que asigna los recursos; y los ***nodos trabajadores***, que ejecutan las tareas concretas. Cada nodo accede a bloques de datos almacenados en sistemas como HDFS o en memoria (***cache***), lo que permite acelerar el análisis de grandes volúmenes de información en tiempo real.

En la práctica, una empresa de transporte urbano podría utilizar Spark para optimizar sus rutas a partir de datos de geolocalización de sus unidades. A medida que cada colectivo envía su ubicación, esos datos se almacenarían de forma distribuida. El ***driver*** organizaría las tareas de análisis, el gestor asignaría los recursos disponibles y los nodos ejecutarían distintas funciones: algunos identificarían desvíos, otros cruzarían datos históricos de tráfico y otros predecirían posibles demoras. El procesamiento en

conjunto permitiría ofrecer rutas alternativas casi en tiempo real, mejorando la eficiencia operativa y la experiencia de los usuarios.

Figura 4. Funcionamiento de Spark



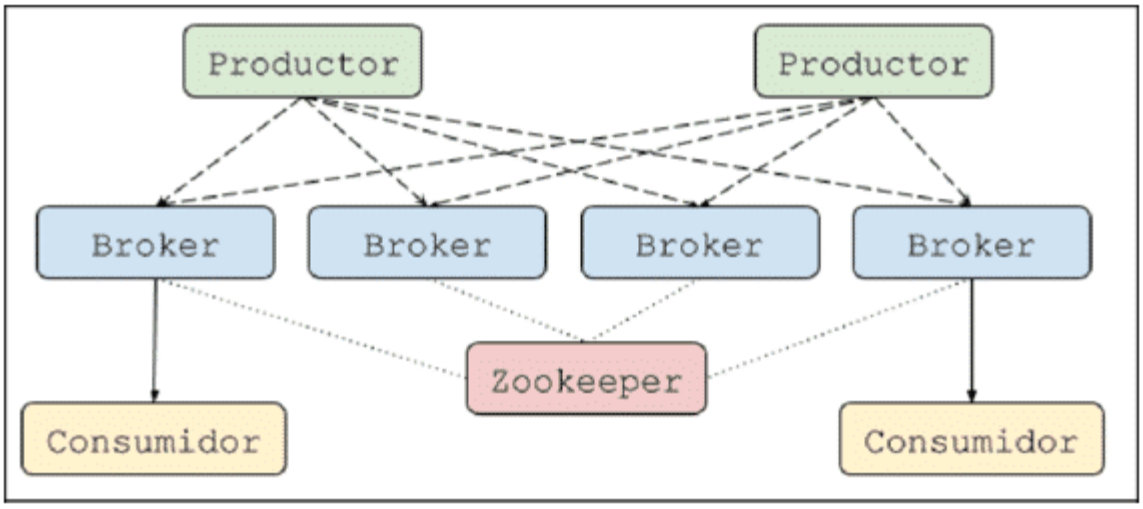
Fuente: [imagen sin título sobre esquema de funcionamiento distribuido de Apache Spark], (s.f.), https://goo.su/XK4SaG

Apache Kafka es una tecnología orientada a la transmisión distribuida de datos en tiempo real. Su arquitectura se basa en el intercambio de mensajes entre productores, ***brokers*** y consumidores, todo coordinado por un sistema auxiliar llamado Zookeeper. Esta estructura permite que grandes volúmenes de datos se transmitan, almacenen y procesen de forma escalable y tolerante a fallos.

Para entender su funcionamiento, pensemos en el siguiente ejemplo: una plataforma de comercio electrónico recibe millones de acciones por minuto, como clics en productos, búsquedas, compras o reseñas. Cada uno de estos eventos se origina en diferentes fuentes (los productores) y se transmite a través de ***brokers*** que organizan y mantienen los datos disponibles para los consumidores. Estos consumidores pueden ser motores de recomendación, sistemas de detección de fraude o tableros de control en tiempo real. Zookeeper, por su parte, se encarga de asegurar que la comunicación entre todos estos componentes se mantenga sincronizada y estable, incluso ante picos de actividad o fallos temporales.

Si bien estas herramientas ofrecen funciones ampliamente utilizadas en entornos ***big data***, la sugerencia es siempre investigar su funcionamiento y evaluar otras alternativas disponibles. Según el tipo de datos, el objetivo del análisis y la escala del proyecto, algunas tecnologías pueden resultar más adecuadas que otras para resolver necesidades específicas.

Figura 5. Arquitectura de Kafka



Fuente: Halabi et al., 2018, https://goo.su/plOU40

CONTINUAR

2. Estructura y funcionamiento de los ecosistemas de datos

En la unidad anterior nos centramos en comprender qué es el *big data*, cómo evolucionaron los entornos de datos y qué tecnologías permiten almacenar y procesar grandes volúmenes de información. A partir de esa base, aparecen nuevos interrogantes: **¿qué pasa con los datos una vez que se generan? ¿Quiénes intervienen en su gestión? ¿Cómo se articulan los distintos actores y sistemas para que esos datos se transformen en información útil?** Estas preguntas permiten ampliar el enfoque y comenzar a pensar los datos como parte de un **ecosistema** dinámico, donde intervienen aspectos técnicos, organizacionales y sociales.

En esta segunda unidad, analizaremos cómo se estructuran estos ecosistemas de datos, qué dimensiones los conforman, quiénes participan de su funcionamiento y cuáles son los principales desafíos para asegurar su interoperabilidad, calidad y gobernanza. También veremos cómo se comparten y circulan los

datos entre plataformas, servicios y sistemas abiertos en la vida cotidiana.

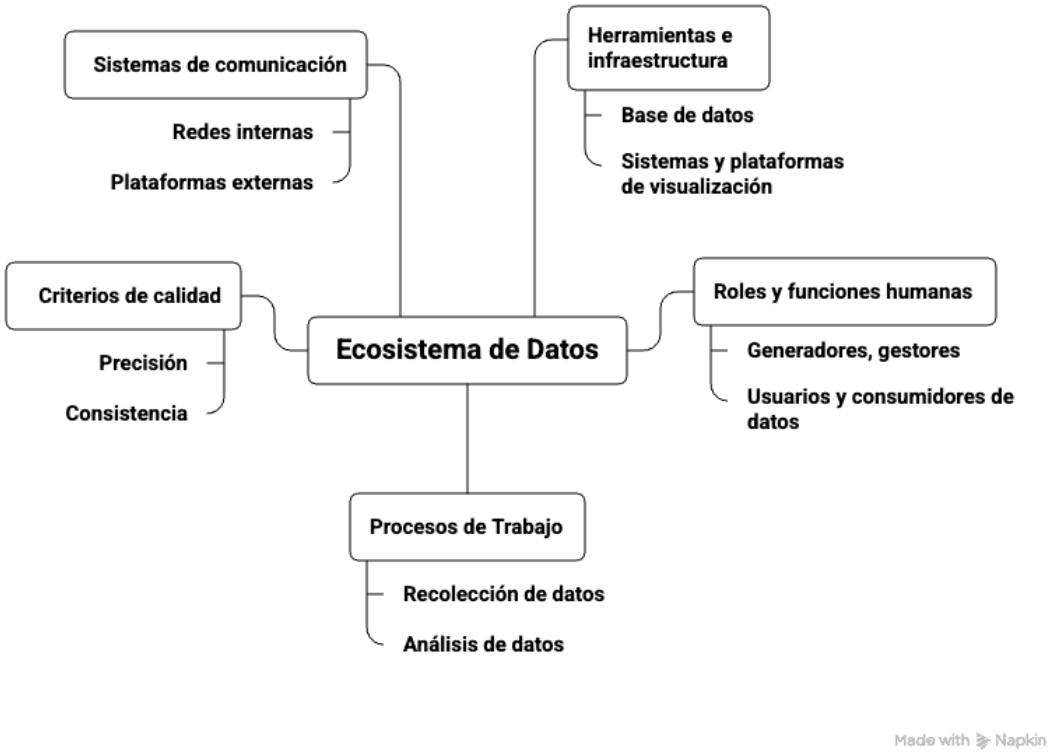
¿Qué es un ecosistema de datos?

Cuando hablamos de *big data*, solemos imaginar grandes volúmenes de información circulando entre servidores, plataformas y dispositivos. Pero detrás de ese movimiento existe una estructura compleja que permite que los datos sean útiles, accesibles y confiables. A ese entramado de componentes interdependientes se lo conoce como ecosistema de datos. En él convergen infraestructuras técnicas, personas que trabajan con los datos y procesos que guían su uso, gestión y aprovechamiento. Comprender cómo se organizan estos elementos permite entender cómo se transforma la información dispersa en recursos útiles para diferentes sectores.

Un ecosistema de datos no se define solamente por la tecnología disponible. También implica decisiones organizativas, roles definidos, criterios de calidad y prácticas compartidas. Por eso, al pensar en estos entornos, es necesario considerar tanto las herramientas y sistemas como los actores humanos y las dinámicas de trabajo que los articulan.

Una forma práctica de identificar los componentes que integran un ecosistema de datos es la siguiente:

Figura 6. Componentes de un ecosistema de datos



Fuente: elaboración propia

Las **herramientas e infraestructuras** incluyen desde bases de datos tradicionales hasta entornos donde se almacenan grandes volúmenes de información en distintos formatos. Estos sistemas permiten guardar, combinar y recuperar datos cuando se necesitan. Se usan, por ejemplo, para registrar transacciones de

clientes, consultar historiales médicos o hacer seguimientos de *stock*. También entran en esta categoría las plataformas que organizan los datos para su análisis o visualización.

En cuanto a los **recursos humanos**, se trata de los distintos perfiles que participan en el manejo de datos: quienes organizan los sistemas, quienes analizan la información y quienes la usan para tomar decisiones. En el próximo tema, nos detendremos en cada uno de estos actores y sus roles dentro del ecosistema.

Los procesos de trabajo refieren a las etapas que siguen los datos: desde su recolección, su integración con otros datos, el almacenamiento, el análisis y la posterior comunicación. Por ejemplo, una empresa puede recopilar información de compras, combinarla con datos de navegación web y usarla para ajustar su estrategia comercial. Para que estos datos sean útiles, deben seguir criterios de calidad: estar actualizados, ser confiables, y estar organizados de modo que se puedan interpretar. Esto es lo que se entiende por gestión de la información.

Finalmente, los sistemas de comunicación y distribución aseguran que los análisis lleguen a quienes los necesitan. Pueden tomar forma de reportes, visualizaciones en tableros o herramientas interactivas que permiten consultar la información según distintos criterios. Su propósito es facilitar el acceso a los datos y su uso en decisiones cotidianas.

Actores del ecosistema de datos

En un ecosistema de datos, diferentes actores participan activamente en la creación, gestión, análisis y uso de la información. Comprender quiénes son y cómo se articulan permite visualizar los recorridos que siguen los datos desde su origen hasta su aprovechamiento práctico. A continuación, se presentan los cuatro principales perfiles involucrados, desde una perspectiva funcional y organizacional.

Generadores de datos

Son quienes producen la información original que da inicio al ciclo de los datos. Pueden hacerlo a través de acciones cotidianas, decisiones operativas o incluso mediante dispositivos que capturan información de manera automatizada. Por ejemplo, una persona que consulta una app de salud y registra su presión arterial está generando datos personales que pueden ser útiles para el sistema sanitario en su conjunto. También lo hace una empresa logística al usar sensores para monitorear la temperatura en camiones de carga, alimentando el sistema con información en tiempo real.

Gestores de datos

Son los responsables de recolectar, organizar, almacenar y preparar los datos para su posterior análisis. Su tarea no se limita a aspectos técnicos, sino que también incluye decisiones sobre calidad, integración y disponibilidad de la información. En una plataforma de educación virtual, por ejemplo, el equipo técnico que centraliza las calificaciones, las interacciones y los registros de acceso actúa como gestor de datos, ya que transforma múltiples fuentes dispersas en un sistema unificado y accesible para otros actores.

Usuarios organizacionales

Utilizan los datos ya procesados para tomar decisiones, definir estrategias o evaluar resultados. Están presentes en áreas como ***marketing***, operaciones, planificación o innovación, y su acceso a la información puede estar mediado por herramientas de visualización o reportes periódicos. En una cooperativa agrícola, por ejemplo, quienes organizan la distribución de productos consultan reportes sobre rendimiento de cosechas y comportamiento de compra en distintas regiones para decidir volúmenes de envío y campañas promocionales.

Consumidores de datos

Reciben los productos finales del análisis de datos, ya sea a través de recomendaciones, información pública, visualizaciones o funcionalidades dentro de un servicio digital. No suelen intervenir directamente en la generación o el procesamiento, pero su experiencia se ve influida por el uso que se da a los datos. Un ciudadano que accede a un mapa de calidad del aire en su ciudad, actualizado en tiempo real, está actuando como

consumidor de datos, al utilizar el resultado final de múltiples procesos técnicos y organizativos.

Estas cuatro figuras conforman una cadena dinámica en la que cada actor cumple un rol específico dentro del ciclo de vida de los datos. Si bien su participación varía en complejidad y nivel de intervención, la articulación efectiva entre generadores, gestores, usuarios y consumidores permite que los datos se transformen en insumos útiles para decisiones, estrategias o servicios. Esta interacción organizada es la que sostiene la circulación informativa en un ecosistema funcional.

Gobernanza e interoperabilidad de los datos compartidos

La circulación de datos en entornos organizacionales requiere algo más que capacidad técnica: **implica asegurar que la información compartida entre actores y sistemas sea precisa, compatible y utilizada de forma confiable**. Conceptos como interoperabilidad y gobernanza permiten entender cómo se construyen condiciones organizativas, técnicas y normativas que garanticen un uso estratégico y ético de los datos. Aunque suelen abordarse de manera conjunta, cada uno representa un aspecto específico del ecosistema informativo, y su

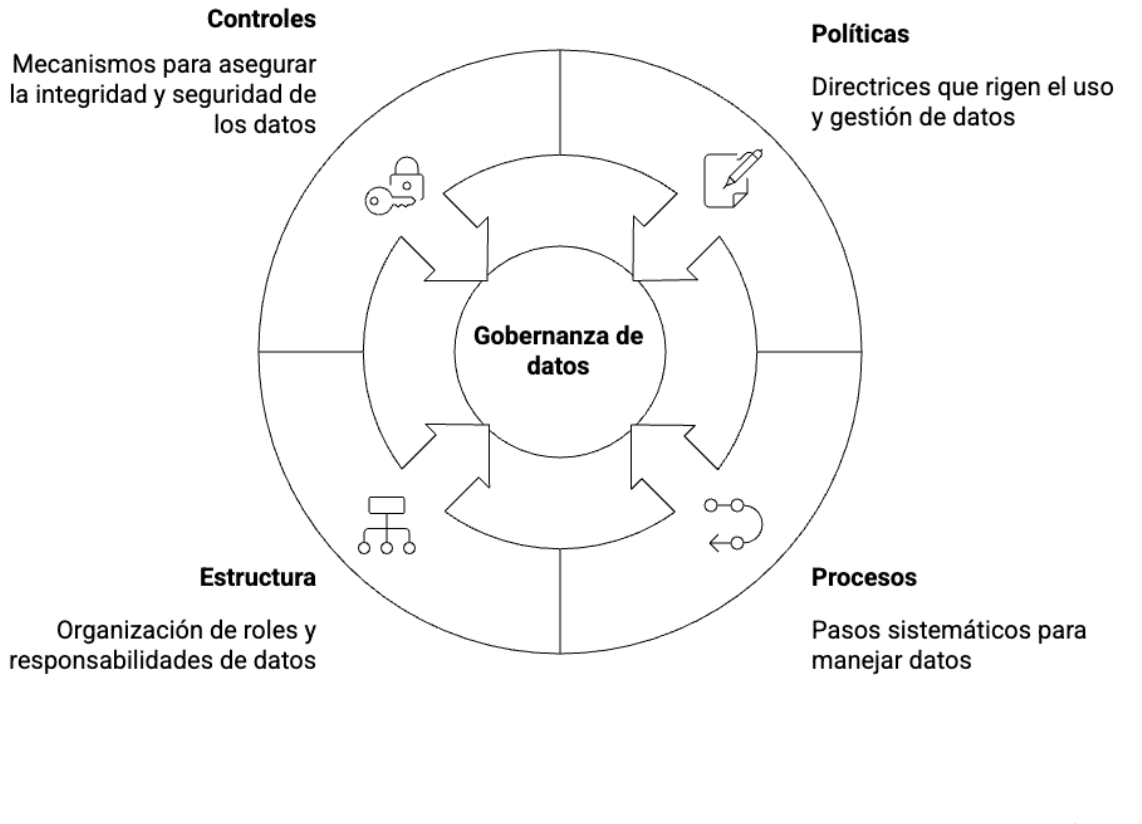
implementación efectiva incide directamente en la toma de decisiones y en la generación de valor a partir de la información.

Gobernanza de los datos

La gobernanza de datos es el conjunto de principios, políticas, estructuras y mecanismos que aseguran que los datos se gestionen de manera segura, consistente y alineada con los objetivos de la organización. Lejos de ser una actividad puntual, constituye una práctica transversal que involucra a múltiples áreas y niveles jerárquicos. A través de ella, se busca garantizar la integridad, calidad, trazabilidad y disponibilidad de los datos, facilitando su uso responsable y estratégico.

Una forma habitual de representar los componentes de un marco de gobernanza es el siguiente esquema:

Figura 7. Componentes de la gobernanza de datos



Fuente: elaboración propia con base en Fortinet, 2025

Este enfoque incluye cuatro elementos principales:

- Las **políticas** marcan los criterios sobre qué datos se recopilan, cómo se usan y quién accede a ellos. Por ejemplo, una empresa puede definir que los datos personales de sus clientes solo se utilicen con fines de atención o soporte.

- Los **procesos** organizan las tareas cotidianas: cómo se incorporan nuevos datos, cómo se corrigen errores o cómo se controla su calidad. Esto permite que distintas áreas trabajen con información coherente y actualizada sin duplicaciones o inconsistencias.

- La **estructura** establece quién se encarga de cada cosa. Desde equipos técnicos que validan los datos, hasta responsables que definen reglas de acceso o supervisan su cumplimiento. Tener funciones definidas reduce conflictos y acelera la toma de decisiones.

- Los **controles** ayudan a monitorear el uso de los datos: quién los consultó, cuándo se modificaron o si hubo accesos no autorizados. Son indispensables para auditar, proteger la información sensible y generar confianza en los sistemas internos.

Contar con estas prácticas —además de mejorar el uso interno de los datos— sienta las condiciones necesarias para conectar sistemas, compartir información entre áreas y garantizar que los datos mantengan su valor en cada instancia del proceso. De ahí que resulte clave complementar la gobernanza con el criterio de interoperabilidad, tema que abordaremos a continuación.

Interoperabilidad de los datos

La interoperabilidad es la capacidad de distintos sistemas, plataformas y organizaciones para intercambiar datos de manera fluida, comprensible y útil. Permite que la información circule entre áreas, instituciones o dispositivos sin perder consistencia ni requerir transformaciones manuales. Esto es especialmente valioso en contextos como la salud, el comercio, la administración pública o la logística, donde múltiples actores necesitan acceder y utilizar la misma información de forma coordinada.

Entre sus principales beneficios, se destacan la mejora en la eficiencia, la reducción de errores, el ahorro de costos y la posibilidad de tomar decisiones más informadas. La siguiente tabla resume los impactos más relevantes que produce su implementación:

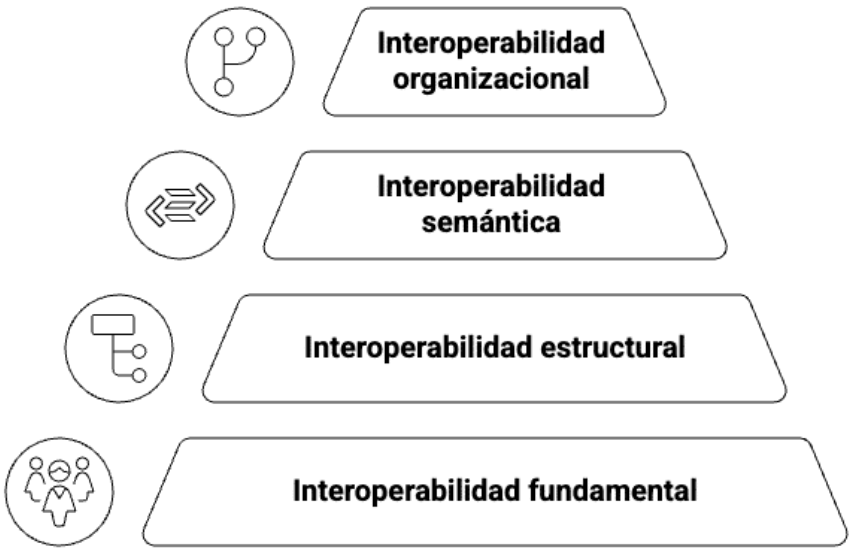
Tabla 3. Beneficios de la interoperabilidad de datos

Beneficio	Descripción breve
Acceso a los datos	Permite consultar múltiples fuentes sin reformateo previo
Mayor eficiencia	Reduce tareas manuales y acelera flujos de trabajo
Mejora en la colaboración	Facilita la coordinación entre equipos y organizaciones
Decisiones más informadas	Amplía la visibilidad sobre los datos disponibles
Escalabilidad	Permite ampliar operaciones sin duplicar sistemas de datos
Reducción de costos	Disminuye gastos de integración y procesamiento de información externa

Fuente: elaboración propia

Para lograr esta conectividad funcional, los sistemas pueden alcanzar diferentes niveles de interoperabilidad. Estos niveles determinan hasta qué punto los datos pueden compartirse y utilizarse entre organizaciones y plataformas.

Figura 8. Niveles de interoperabilidad de datos



Fuente: elaboración propia con base en Kosinski, 2024

En el primer nivel, conocido como **interoperabilidad fundamental**, los sistemas logran intercambiar datos, pero sin entender su contenido. Por ejemplo, un archivo CSV puede enviarse de una institución a otra, pero quien lo recibe debe abrirlo y procesarlo manualmente para entender su contenido.

El segundo nivel es la **interoperabilidad estructural**, donde los datos se organizan bajo formatos estándar. Esto permite que distintos sistemas los reconozcan automáticamente. Un ejemplo sería un sistema de reservas turísticas que exporta datos de pasajeros clasificados por nombre, documento y destino, para que una aerolínea los procese sin necesidad de edición.

En el tercer nivel, la **interoperabilidad semántica**, los sistemas no solo leen los datos, sino que también entienden su significado gracias a un lenguaje común. Por ejemplo, distintos hospitales pueden usar códigos estandarizados para enfermedades, de modo que, aunque un paciente sea tratado en diferentes centros, su historial médico sea interpretado correctamente en todos.

Finalmente, la **interoperabilidad organizacional** implica que las instituciones no solo compartan tecnología, sino también procesos, protocolos y acuerdos legales para trabajar de forma coordinada. Es el caso de agencias estatales que unifican criterios para cruzar datos de empleo, salud y educación a fin de ofrecer servicios sociales integrales sin duplicar esfuerzos.

Ahora bien, esta posibilidad de cooperación solo se vuelve efectiva si los datos pueden circular fluidamente entre plataformas y sistemas. Esto exige entornos abiertos, accesibles y técnicamente compatibles. En el siguiente apartado, nos detendremos en el concepto de intercambio de datos como

condición indispensable para lograr ecosistemas digitales que operen de manera integrada, permitan la reutilización de información y habiliten procesos de decisión basados en evidencias confiables y oportunas.

Intercambio de datos

El intercambio de datos es una práctica cada vez más común en organizaciones que buscan tomar decisiones basadas en información confiable, actualizada y compartida de forma segura. Consiste en permitir que distintas áreas internas o actores externos accedan a ciertos datos, sin necesidad de replicarlos o duplicarlos innecesariamente. Para que esto funcione, es necesario que haya coordinación, tecnología adecuada y reglas claras. El objetivo es que la información fluya sin fricciones, pero sin perder control sobre su calidad o su privacidad.

Una manera más concreta de pensarlo es imaginar una empresa que tiene múltiples departamentos: comercial, logística, atención al cliente, entre otros. Si cada uno almacena y gestiona los datos por su cuenta, se corre el riesgo de tener información desactualizada, incompleta o incluso contradictoria. En cambio, si todos acceden a una misma fuente de datos compartida, se puede trabajar de forma más alineada, con menos errores y con mayor agilidad para resolver problemas.

Este enfoque permite, por ejemplo, que el área de ventas conozca rápidamente el historial de compras de un cliente, que logística planifique entregas con base en datos reales de demanda, o que el área de atención pueda responder con precisión a una consulta. En lugar de esperar reportes o pedir autorizaciones para ver cierta información, los equipos pueden tomar decisiones más rápidas y coordinadas. El intercambio de datos, en este sentido, se convierte en una herramienta para mejorar la eficiencia operativa.

Pero los beneficios no terminan dentro de la organización. Muchas veces, también es necesario compartir información con otras empresas, instituciones públicas o socios estratégicos. Para hacerlo, es imprescindible garantizar que esos datos estén protegidos, que se sepa quién los usa y para qué, y que el intercambio se realice de forma controlada. Esto se vuelve especialmente relevante cuando se trabaja con datos sensibles, como información personal de clientes o registros financieros.

A nivel global, el intercambio de datos ha crecido de manera exponencial gracias al desarrollo de tecnologías que permiten hacerlo en tiempo real y desde cualquier lugar. Sin embargo, también ha aumentado la necesidad de contar con buenas prácticas: desde definir quién tiene acceso, hasta asegurarse de que los datos estén bien organizados, correctamente documentados y sean comprensibles para quienes los consultan.

De lo contrario, el riesgo es compartir información que no se puede usar o que puede ser malinterpretada.

El intercambio de datos también plantea desafíos importantes. Por un lado, compartir información requiere confiar en que otros actores respetarán las reglas establecidas. Por otro, hay que asegurarse de que los datos estén bien protegidos frente a posibles filtraciones o usos indebidos. Por eso, muchas organizaciones definen políticas claras de manejo de datos, invierten en sistemas seguros y revisan periódicamente quién accede a qué. No se trata solo de abrir el acceso, sino de hacerlo de manera responsable.

A medida que las organizaciones avanzan en sus procesos de transformación digital, el intercambio de datos adquiere un papel central en la coordinación de acciones, el desarrollo de estrategias y la mejora continua. Establecer condiciones claras de acceso, roles definidos y criterios comunes permite que la información circule con propósito, sea comprensible y útil para todos los sectores involucrados. Esta dinámica favorece la creación de entornos colaborativos, impulsa la innovación y fortalece las decisiones basadas en evidencia dentro y fuera de la organización.

CONTINUAR

Referencias

Bartley, K. (2025). *Big data statistics: How much data is there in the world?* Rivery. https://rivery.io/blog/big-data-statistics-how-much-data-is-there-in-the-world/

Big Data Framework. (2025). *A short history of Big Data*. https://www.bigdataframework.org/short-history-of-big-data/

Fernández, O. (2025). *¿Qué es Hadoop MapReduce? Introducción*. Aprender Big Data. https://aprenderbigdata.com/hadoop-mapreduce/

Fortinet. (2025). *¿Qué es la gobernanza de datos? Mejores prácticas y componentes*. https://www.fortinet.com/lat/resources/cyberglossary/data-governance

Halabi, A. R., Rueda-Toicen, A., & Ospina, M. (2018). *Registros médicos permisados y distribuidos a través de Hyperledger Fabric e*

InterPlanetary Filesystem. https://doi.org/10.13140/RG.2.2.30327.02720

[Imagen sin título sobre esquema de funcionamiento distribuido de Apache Spark], (s.f.).https://adictosaltrabajo.com/2015/11/16/introduccion-a-apache-spark-batch-y-streaming/

Kosinski, M. (2024). *Interoperabilidad: conecte datos, sistemas y organizaciones para impulsar la innovación*. IBM. https://www.ibm.com/mx-es/think/topics/interoperability

Mucci, T. (2024). *¿Qué es el intercambio de datos?* IBM. https://www.ibm.com/es-es/think/topics/data-sharing

CONTINUAR