



INTRODUCCIÓN A BIG DATA

Big Data se caracteriza por las **5V**: **volumen**, la cantidad masiva de datos generados; **velocidad**, rapidez en la recopilación y análisis; **variedad**, tipos de datos estructurados y no estructurados; **veracidad**, calidad y confiabilidad de los datos; y **valor**, información útil para decisiones estratégicas. Herramientas como Hadoop y Spark permiten procesar y analizar grandes volúmenes de datos, diferenciándose por su enfoque: Hadoop es ideal para procesamiento por lotes, mientras que Spark destaca en análisis en tiempo real. Bases de datos NoSQL, como MongoDB y Cassandra, facilitan la escalabilidad y flexibilidad en entornos dinámicos.

Dominar Big Data impulsa la innovación y abre caminos hacia un futuro transformador.

INICIAR





INTELIGENCIA DE NEGOCIOS

INTRODUCCIÓN A BIG DATA

Big Data se caracteriza por las **5V**: **volumen**, la cantidad masiva de datos generados; **velocidad**, rapidez en la recopilación y análisis; **variedad**, tipos de datos estructurados y no estructurados; **veracidad**, calidad y confiabilidad de los datos; y **valor**, información útil para decisiones estratégicas. Herramientas como Hadoop y Spark permiten procesar y analizar grandes volúmenes de datos, diferenciándose por su enfoque: Hadoop es ideal para procesamiento por lotes, mientras que Spark destaca en análisis en tiempo real. Bases de datos NoSQL, como MongoDB y Cassandra, facilitan la escalabilidad y flexibilidad en entornos dinámicos.

Dominar Big Data impulsa la innovación y abre caminos hacia un futuro transformador.

INICIAR



TECNOLÓGICA DEL ORIENTE
INSTITUCIÓN DE EDUCACIÓN SUPERIOR

Todo el contenido de este curso es propiedad intelectual de la Corporación Tecnológica del Oriente y está protegido por derechos de autor. No puede ser reproducido, distribuido, modificado ni compartido sin su autorización por escrito.

UNIDAD 1. INTRODUCCIÓN A BIG DATA

INTRODUCCIÓN

El crecimiento exponencial de los datos en la era digital, ha dado lugar a la necesidad de herramientas y enfoques que permiten procesar, analizar y extraer valor de grandes volúmenes de información. *Big Data* surge como una solución clave, al abordar el manejo de datos complejos, provenientes de diversas fuentes y formatos.

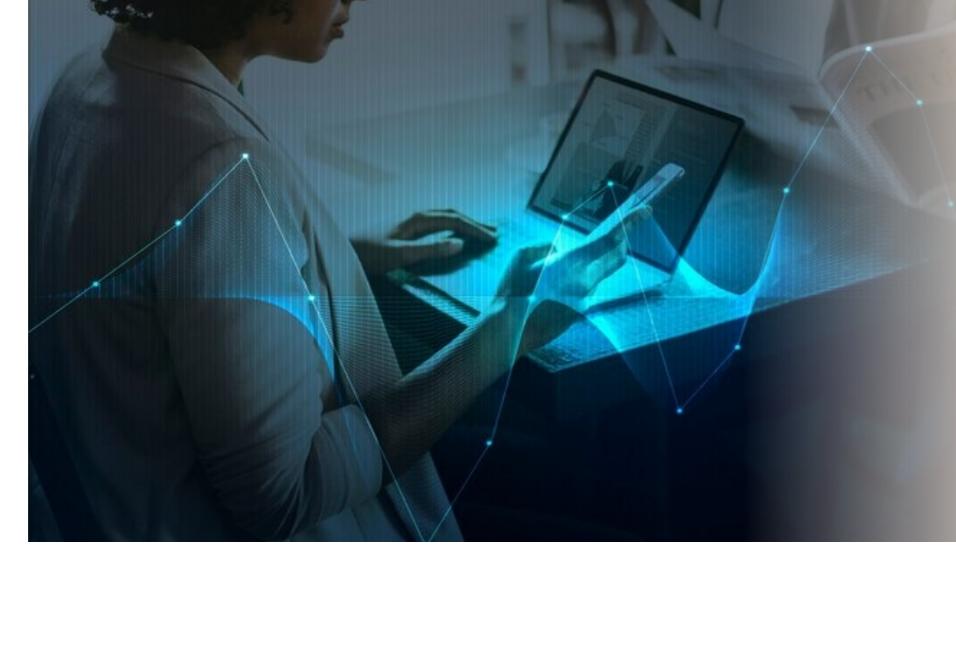
Sus características fundamentales, conocidas como las 5V: volumen, velocidad, variedad, veracidad y valor, son la base para comprender cómo los datos pueden transformar organizaciones y procesos en múltiples sectores. La combinación de estas propiedades, permite convertir información cruda en conocimiento estratégico.



La adopción de tecnologías como Hadoop y Spark, junto con bases de datos NoSQL, facilita el análisis eficiente de datos, permitiendo a las organizaciones optimizar decisiones y adaptarse a un entorno cada vez más dinámico y competitivo.

UNIDAD 1. INTRODUCCIÓN A BIG DATA

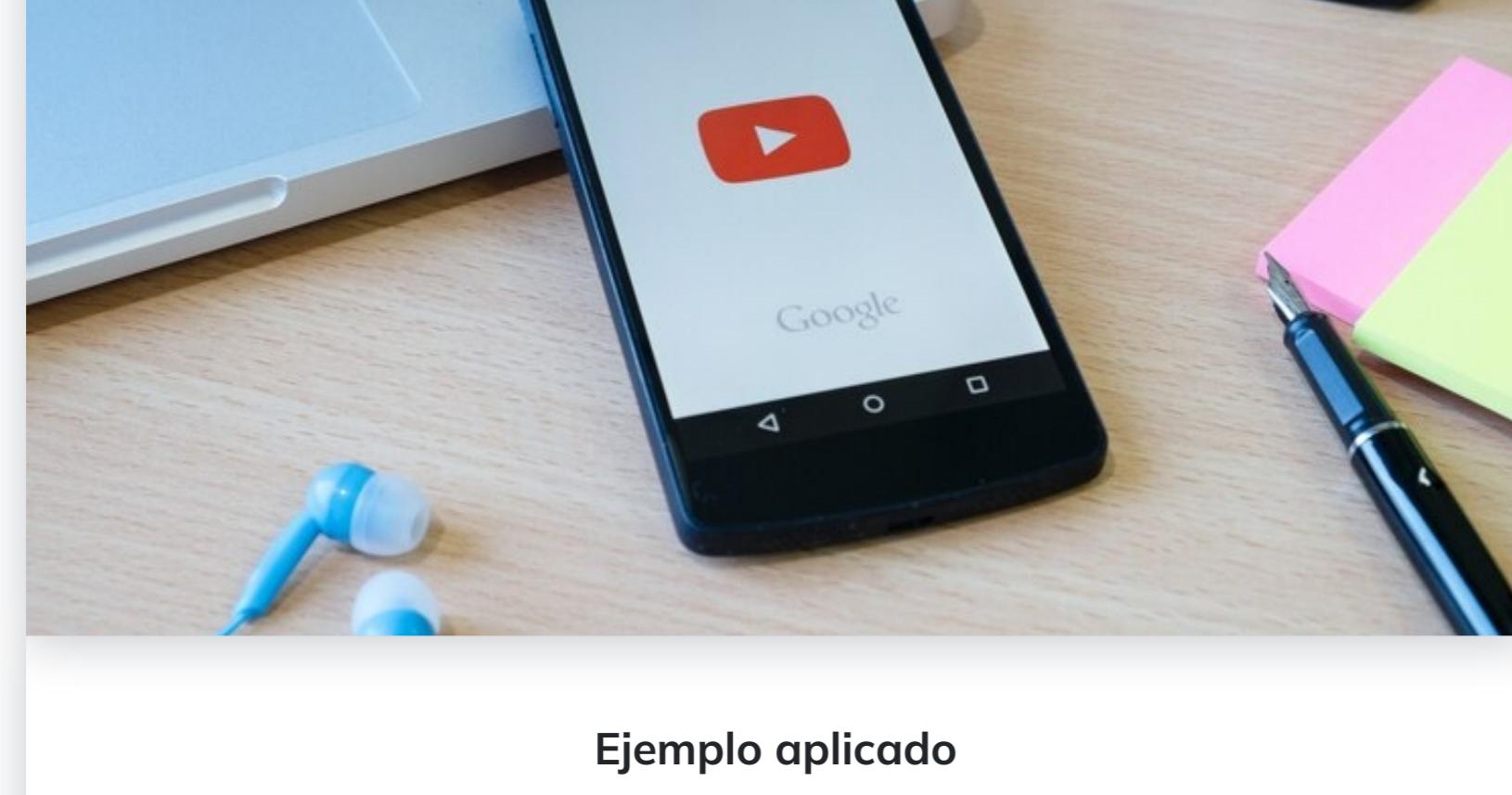
1. CONCEPTOS BÁSICOS DE BIG DATA



Big Data engloba el manejo de grandes volúmenes de datos que exceden la capacidad de procesamiento de las herramientas tradicionales. Sus cinco características esenciales, conocidas como las "5V", son fundamentales para comprender y gestionar el fenómeno. A continuación, se explican en detalle, con ejemplos prácticos y análisis crítico.

Volumen

Hace referencia a la enorme cantidad de datos generados y almacenados diariamente debido al crecimiento de las plataformas digitales, redes sociales y dispositivos IoT.



Ejemplo aplicado

Cada minuto, se suben más de 500 horas de video a plataformas como YouTube, lo que requiere sistemas robustos de almacenamiento y organización.

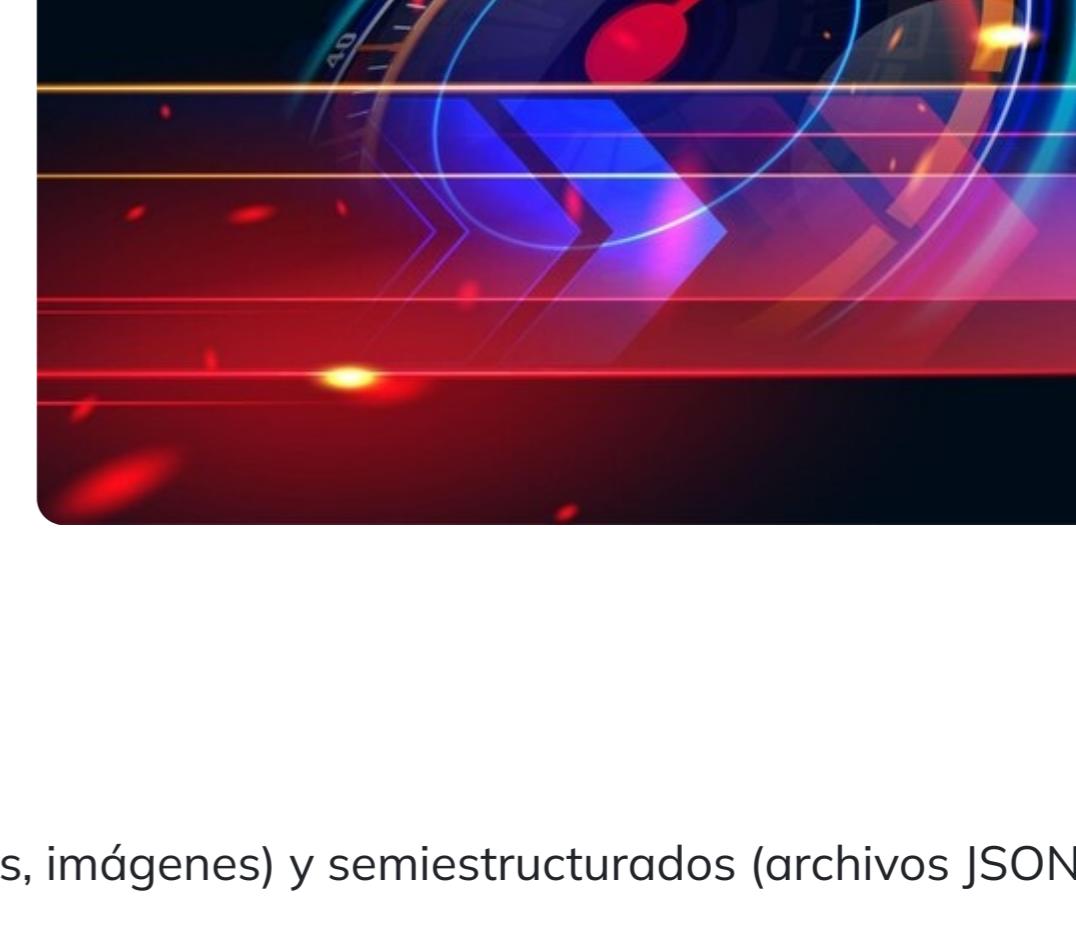
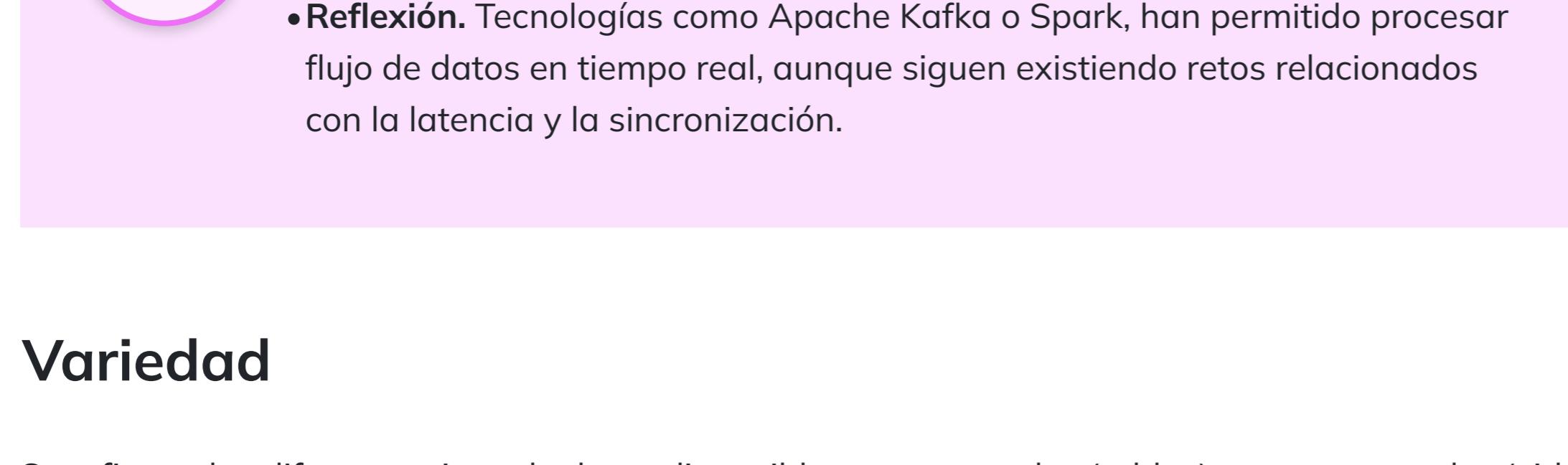


Análisis críticos

Las empresas enfrentan desafíos significativos para almacenar estos volúmenes de datos, de manera eficiente y rentable. Soluciones como el almacenamiento en la nube y los sistemas distribuidos, se han convertido en estándares para abordar este reto.

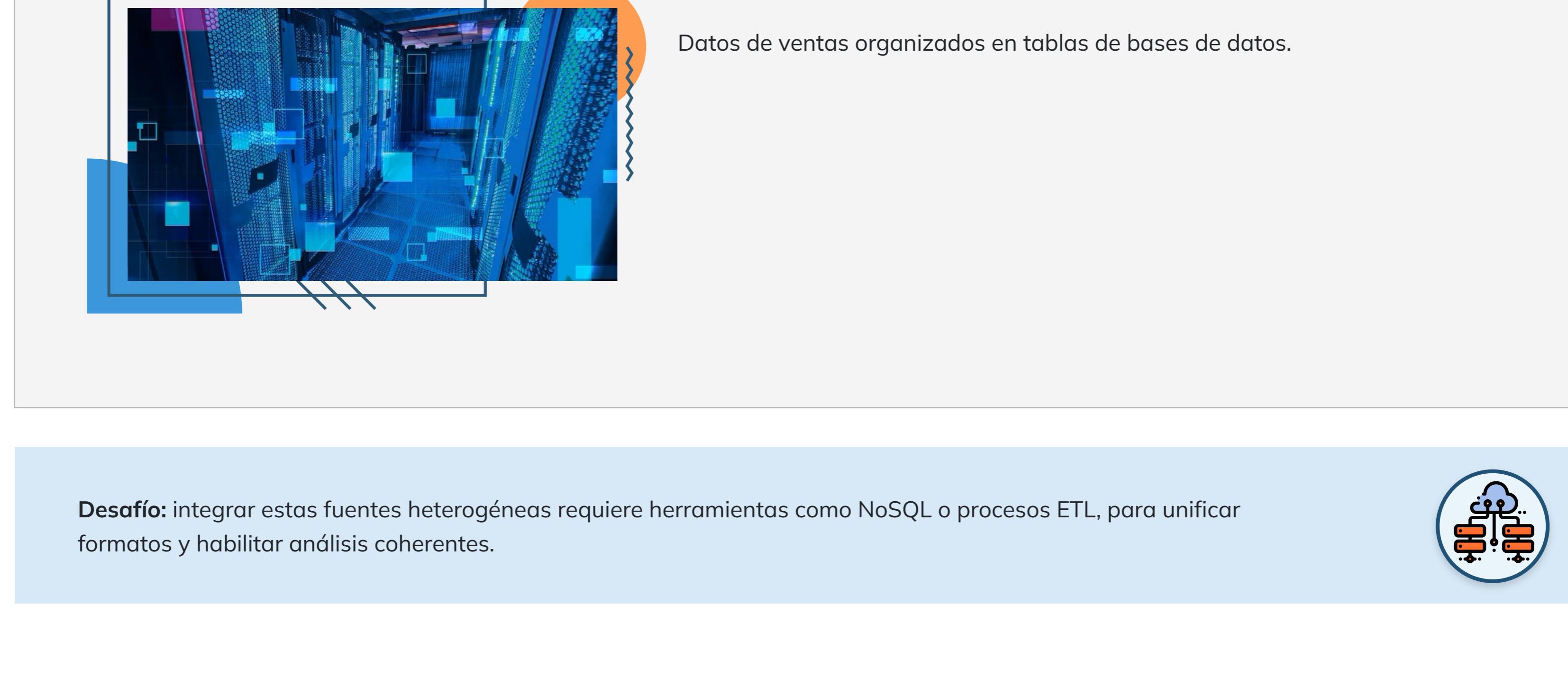
Velocidad

Indica la rapidez con la que los datos son generados, recolectados y procesados, siendo crucial para aplicaciones que demandan análisis en tiempo real.



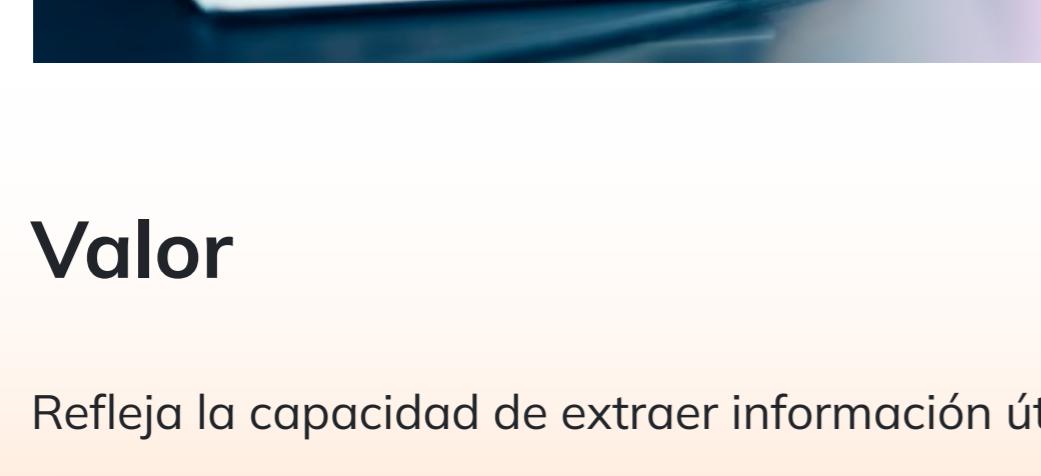
Variedad

Se refiere a los diferentes tipos de datos disponibles: estructurados (tablas), no estructurados (videos, imágenes) y semiestructurados (archivos JSON o XML).



Veracidad

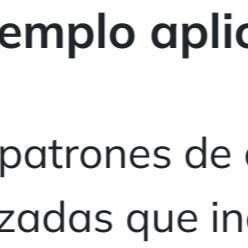
Se enfoca en garantizar la calidad, precisión y confiabilidad de los datos, dado que la información incorrecta puede llevar a decisiones equivocadas.



- Ejemplo aplicado.** En el sector salud, datos imprecisos en los historiales clínicos, pueden comprometer diagnósticos y tratamientos.
- Técnicas clave.** Algoritmos de validación y herramientas de limpieza de datos, se emplean para identificar errores y mejorar la confiabilidad de los análisis.

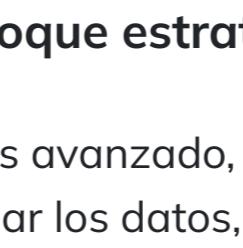
Valor

Refleja la capacidad de extraer información útil y relevante, de los datos recopilados, para tomar decisiones estratégicas.



Ejemplo aplicado

En marketing, el análisis de patrones de comportamiento, permite crear campañas personalizadas que incrementan la retención.

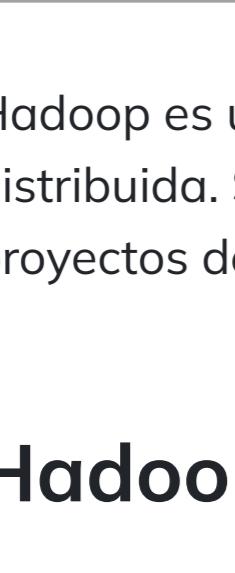


Enfoque estratégico

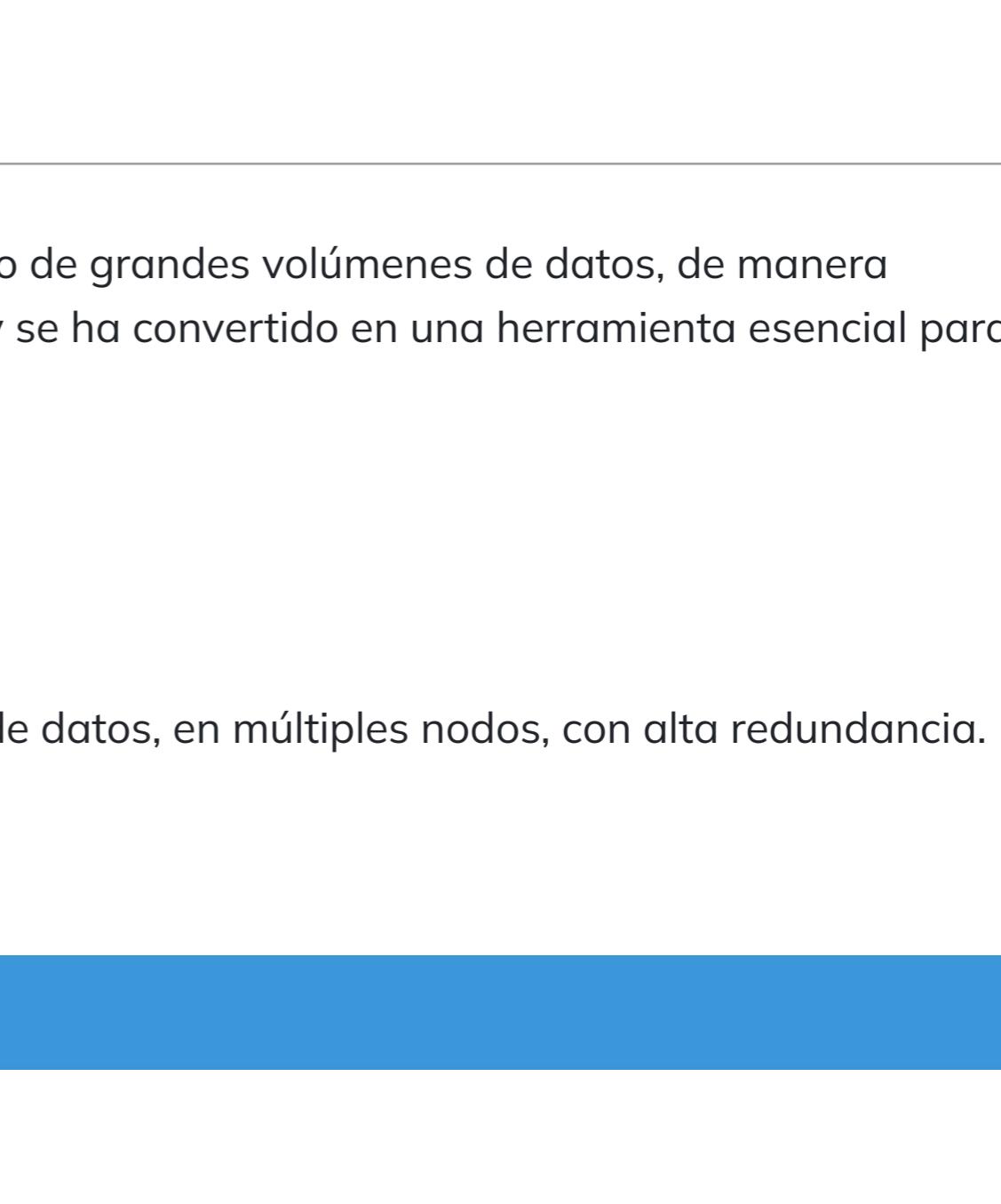
Las herramientas de análisis avanzado, como Tableau o Power Bi, Son esenciales para transformar los datos, en acciones de alto impacto.

El éxito de una estrategia de Big Data depende del equilibrio entre estas cinco características. Ignorar cualquiera de ellas puede comprometer la utilidad y confiabilidad de los datos. Por ejemplo, un sistema que maneja datos a alta velocidad, pero con baja veracidad, puede generar conclusiones inexactas. Del mismo modo, sin una clara orientación hacia el valor, los esfuerzos en Big Data pierden relevancia estratégica.

2. HERRAMIENTAS Y TECNOLOGÍAS PARA BIG DATA



El manejo eficiente de *Big Data* requiere no solo de una comprensión teórica, sino de herramientas tecnológicas avanzadas que permitan procesar, almacenar y analizar grandes volúmenes de datos, de manera escalable y eficiente. A continuación, se presentan dos de las tecnologías más destacadas: *Hadoop* y *Apache Spark*, explicando su arquitectura, aplicaciones y casos de uso, en diferentes industrias.



2.1 Introducción a *Hadoop*

Hadoop es una plataforma de código abierto, diseñada para el almacenamiento y procesamiento de grandes volúmenes de datos, de manera distribuida. Su arquitectura permite manejar tanto datos estructurados como no estructurados, y se ha convertido en una herramienta esencial para proyectos de *Big Data*.

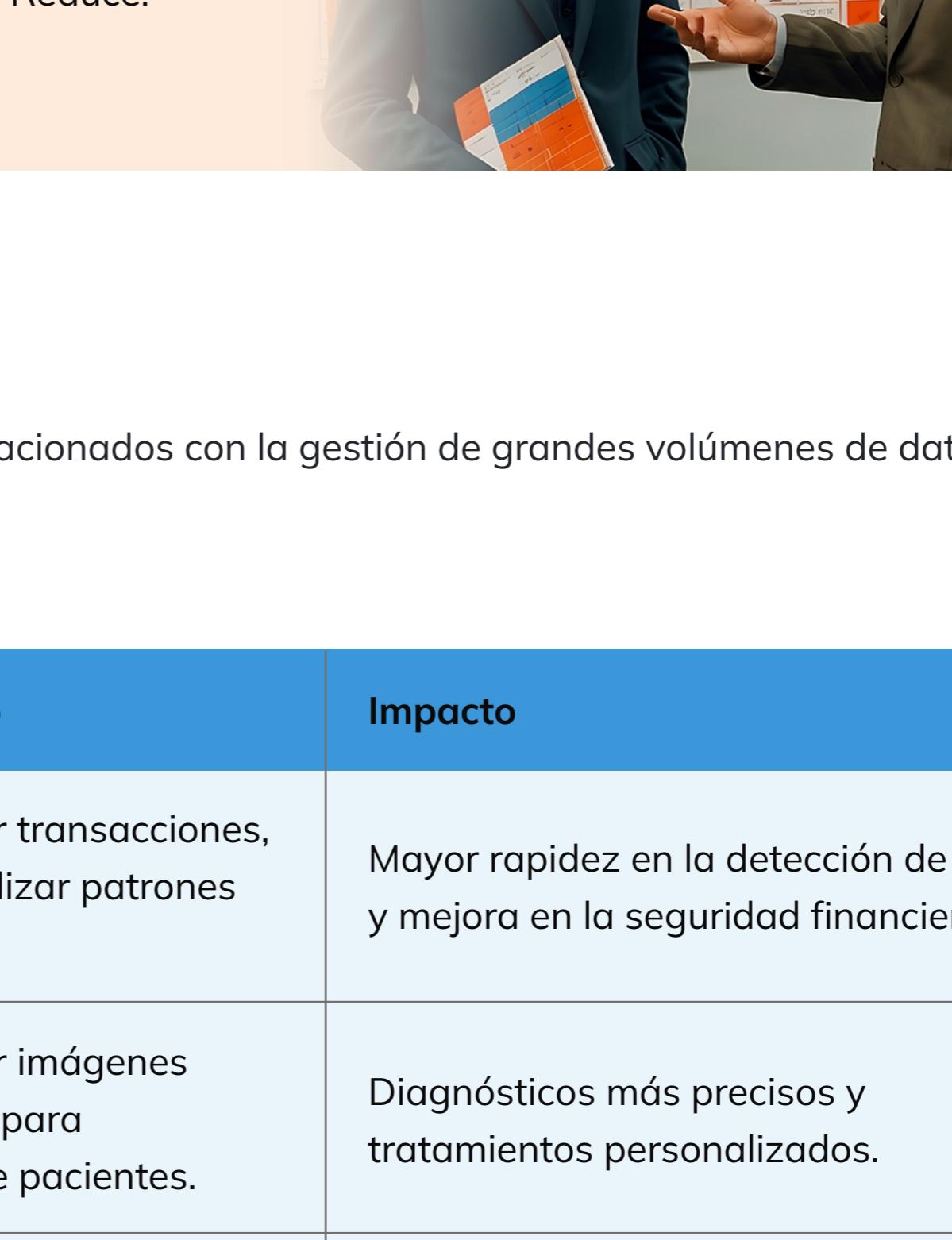
Hadoop Distributed File System (HDFS)

Es el sistema de archivos distribuido de *Hadoop*, diseñado para almacenar grandes volúmenes de datos, en múltiples nodos, con alta redundancia.

Las características principales son:

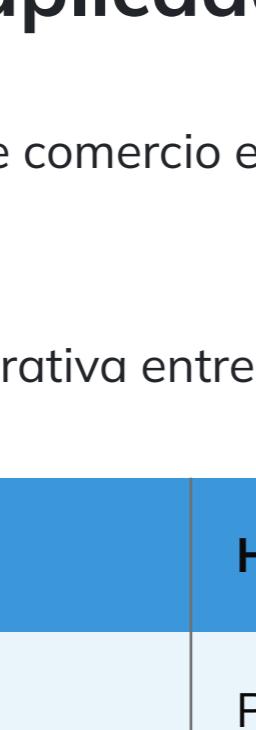
Escalabilidad horizontal

Capacidad de agregar nodos para aumentar el almacenamiento y la potencia de procesamiento.



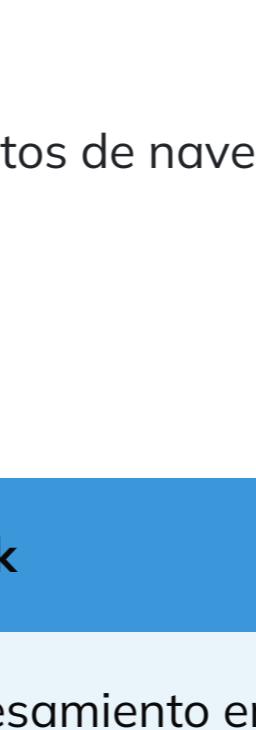
MapReduce

Es el modelo de procesamiento distribuido de *Hadoop*, que divide las tareas en dos etapas:



Map

Divide los datos en fragmentos procesables en paralelo.



Reduce

Combinan los resultados de las tareas Map para generar una salida final.

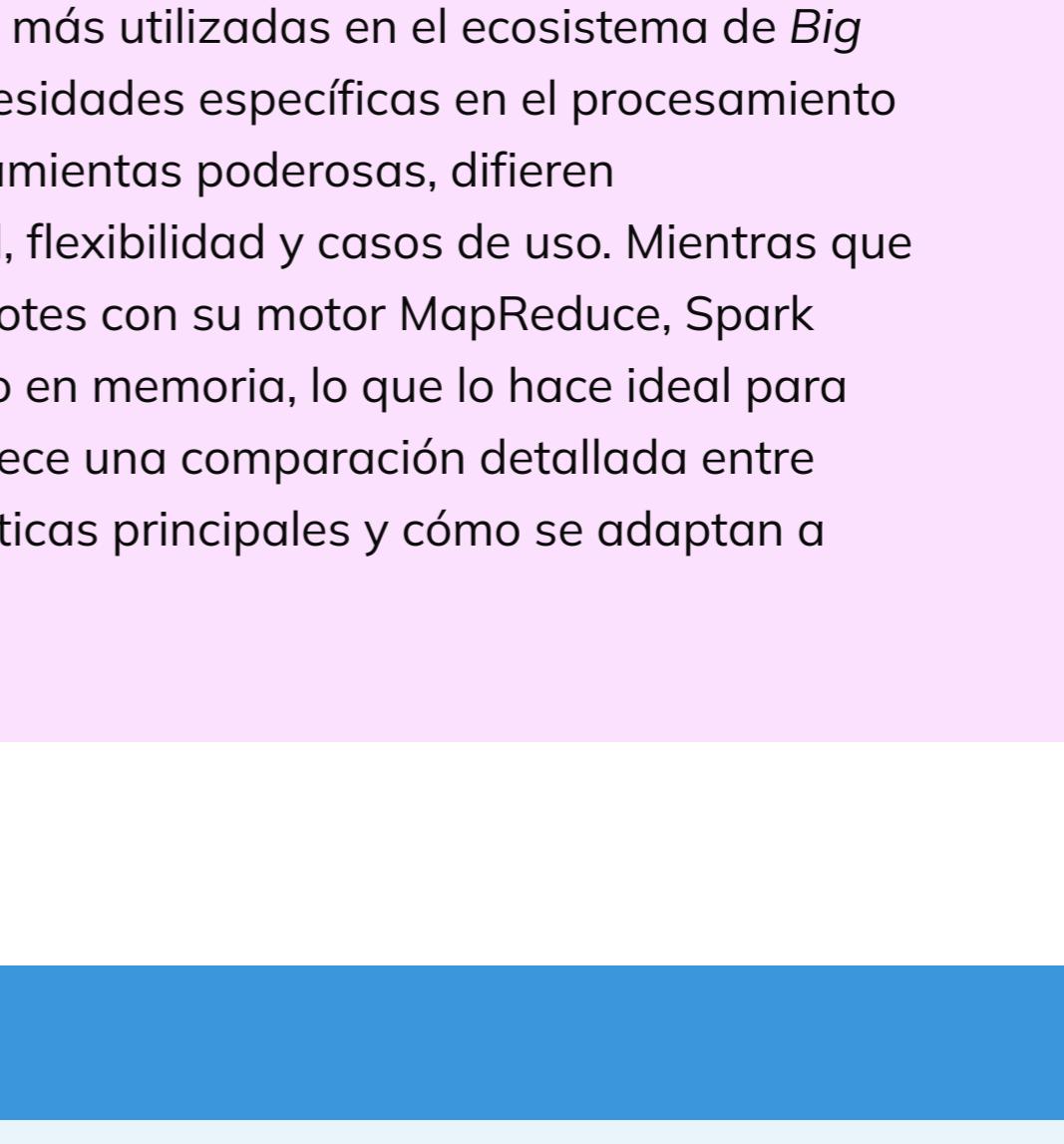


Ejemplo práctico

Una empresa de telecomunicaciones usa MapReduce, para analizar registros de llamadas y detectar patrones de uso que optimicen su infraestructura.

Se describen las partes claves del sistema y su función, dentro del procesamiento de grandes volúmenes de datos distribuidos.

- **HDFS**. Almacena datos distribuidos entre nodos, con tolerancia a fallos.
- **MapReduce**. Procesa datos en paralelo, dividiendo tareas en las fases Map y Reduce.



Casos de uso en industrias clave

Hadoop se ha utilizado en una amplia gama de sectores, para resolver problemas relacionados con la gestión de grandes volúmenes de datos.

Tabla 1. Aplicaciones de *Hadoop* por industria

Industria	Problema	Solución con <i>Hadoop</i>	Impacto
Banca y finanzas	Detección de fraudes.	HDFS para almacenar transacciones, MapReduce para analizar patrones anómicos.	Mayor rapidez en la detección de fraudes y mejora en la seguridad financiera.
Salud	Procesamiento de datos médicos.	HDFS para almacenar imágenes médicas, MapReduce para correlacionar datos de pacientes.	Diagnósticos más precisos y tratamientos personalizados.
Telecomunicaciones	Análisis de registros de llamadas.	HDFS para datos de torres, MapReduce para predecir picos de tráfico.	Optimización de la red y mejoría de la experiencia del cliente.

2.2 Spark como herramienta de procesamiento en tiempo real

Apache Spark, es una plataforma de análisis distribuido, diseñada para manejar grandes volúmenes de datos con rapidez y flexibilidad. A diferencia de *Hadoop*, *Spark* puede realizar procesamiento en tiempo real, lo que lo convierte en una herramienta clave en aplicaciones críticas, como aprendizaje automático y análisis predictivo.

Las ventajas principales de *Spark*:

1

Velocidad

Procesa datos en memoria, hasta 100 veces más rápido que MapReduce.

2

Flexibilidad

Admite lotes y flujos en tiempo real.

3

Competitividad

Analiza datos en tiempo real, con mayor velocidad y eficiencia.

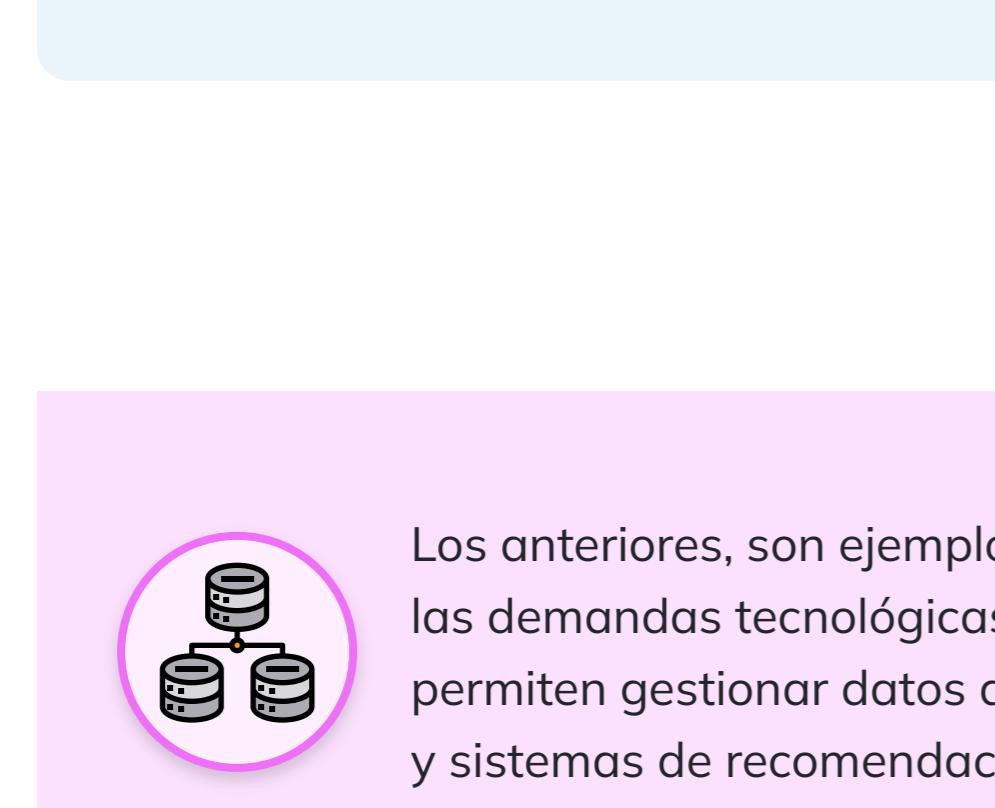


Tabla 2. Comparativa entre *Hadoop* y *Spark*

Aspecto	<i>Hadoop</i>	<i>Spark</i>
Velocidad	Procesamiento más lento con MapReduce.	Procesamiento en tiempo real, con Spark Streaming.
Tipo de datos	Enfocado en grandes lotes.	Admite lotes y flujos en tiempo real.
Casos de uso	Almacenamiento masivo y análisis por lotes.	Ánalisis predictivo, aprendizaje automático, etc.

Ambas tecnologías, *Hadoop* y *Spark*, son esenciales en el ecosistema de *Big Data*, y su aplicación combinada puede maximizar el potencial de los proyectos en diversos sectores. Estas herramientas permiten a las organizaciones transformar datos complejos en información valiosa, facilitando decisiones informadas y estrategias efectivas.

Comparativa entre *Hadoop* y *Spark*: velocidad y flexibilidad



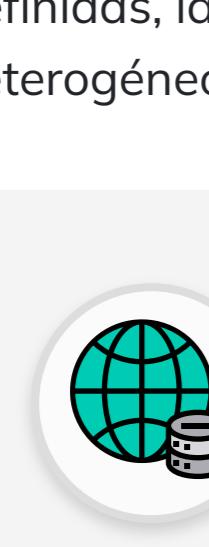
Hadoop y *Spark*, son dos de las tecnologías más utilizadas en el ecosistema de *Big Data*, cada una diseñada para abordar necesidades específicas en el procesamiento de datos masivos. Aunque ambas son herramientas poderosas, difieren significativamente en términos de velocidad, flexibilidad y casos de uso. Mientras que *Hadoop* se centra en el procesamiento por lotes con su motor MapReduce, *Spark* destaca por su capacidad de procesamiento en memoria, lo que lo hace ideal para tareas en tiempo real. La siguiente tabla ofrece una comparación detallada entre *Hadoop* y *Spark*, destacando sus características principales y cómo se adaptan a diferentes escenarios de análisis de datos.

Tabla 3. Comparativa entre *Hadoop* y *Spark*: velocidad y flexibilidad

CRITERIO	HADOOP	SPARK
Velocidad	Procesamiento por lotes: <i>Hadoop</i> se basa en MapReduce, lo que significa que los datos se procesan en bloques, con múltiples etapas de lectura y escritura en disco. Esto puede hacer que sea más lento para tareas que requieren respuestas rápidas.	Procesamiento en memoria: <i>Spark</i> realiza la mayoría de las operaciones directamente en memoria (RAM), eliminando la necesidad de escribir y leer constantemente en disco. Esto lo hace hasta 100 veces más rápido que <i>Hadoop</i> , en ciertas tareas.
Flexibilidad	Enfocado principalmente en el procesamiento por lotes y con un modelo limitado para manejar datos en tiempo real.	Compatible con múltiples lenguajes de programación (Scala, Python, Java, R). Soporta para múltiples cargas de trabajo: análisis de datos, aprendizaje automático (MLlib), procesamiento de grafos (GraphX) y consultas SQL (Spark SQL). Funciona tanto con datos por lotes como en tiempo real, lo que amplía significativamente su rango de aplicaciones.

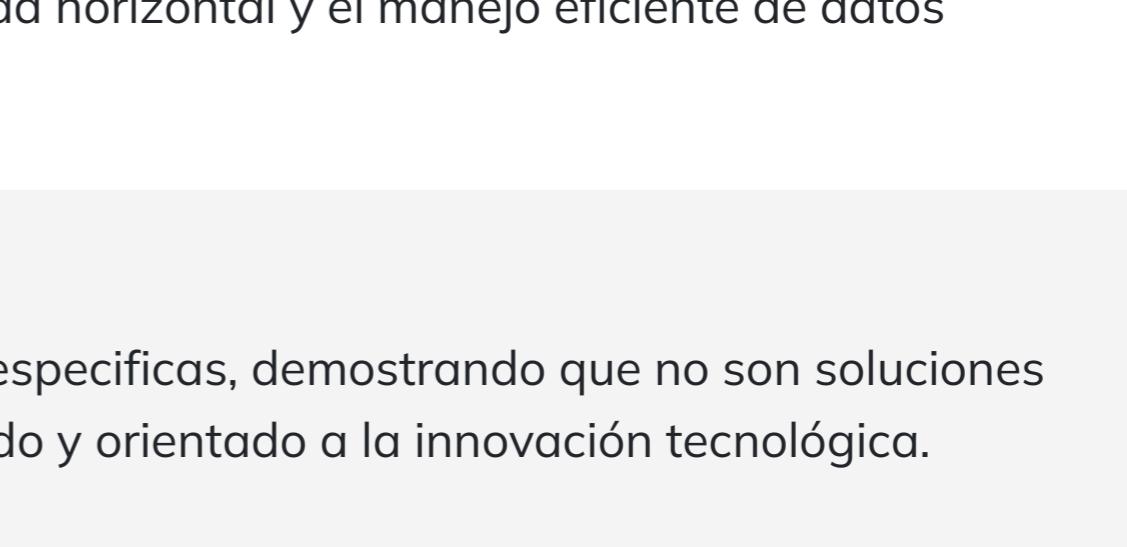
Mientras que *Hadoop* es ideal para el almacenamiento y el procesamiento de datos por lotes a gran escala, *Spark* sobresale en tareas que requieren rapidez, flexibilidad y análisis en tiempo real. Muchas organizaciones combinan ambas tecnologías, utilizando *Hadoop* para almacenamiento (HDFS) y *Spark* para procesamiento ágil.

La capacidad de *Spark* para procesar flujos de datos en tiempo real, lo ha convertido en una herramienta esencial para industrias que necesitan decisiones rápidas y basadas en datos. A continuación, se presentan ejemplos destacados:



• Características de las bases de datos NoSQL (escalabilidad horizontal, flexibilidad de esquemas).

• Ejemplos de bases de datos como MongoDB y Cassandra.



Sistemas de recomendación

Empresas como Netflix y Amazon utilizan *Spark* para analizar patrones de comportamiento en tiempo real, generando recomendaciones personalizadas de contenido o productos. *Spark* procesa datos como histórico de visualización, búsquedas recientes y preferencias de otros usuarios similares.

Impacto: incremento en la satisfacción del cliente y en la conversión de ventas o tiempo de permanencia en la plataforma.

Monitoreo de fraudes financieros

Optimización de redes de telecomunicaciones

Análisis de redes sociales y sentimiento del consumidor

Estas aplicaciones no solo mejoran la eficiencia operativa y la seguridad, sino que también potencian la capacidad de adaptación en entornos dinámicos, reafirmando el valor estratégico del análisis de datos, en tiempo real, en un mundo cada vez más conectado y competitivo.

2.3 Bases de datos NoSQL

Las bases de datos NoSQL (Not Only SQL), han surgido como una alternativa fundamental a las bases de datos relacionales tradicionales, para enfrentar los retos asociados con *Big Data*. Diseñadas para manejar grandes volúmenes de datos no estructurados y semiestructurados, estas bases de datos destacan por su escalabilidad y flexibilidad, lo que las convierte en la opción preferida para aplicaciones modernas, como análisis en tiempo real, IoT y sistemas distribuidos.

• Características de las bases de datos NoSQL (escalabilidad horizontal, flexibilidad de esquemas).

• Ejemplos de bases de datos como MongoDB y Cassandra.

Características principales de las bases de datos NoSQL

Seguidamente, se presenta una síntesis de las características principales de las bases de datos NoSQL, destacando sus diferencias fundamentales frente a las bases de datos relacionales tradicionales. Este enfoque permite comprender las fortalezas y limitaciones de las bases de datos NoSQL en diversos contextos, facilitando su elección, según las necesidades particulares de almacenamiento y procesamiento de información.

Escalabilidad horizontal

Una de las principales ventajas de NoSQL es la capacidad de escalar horizontalmente, es decir, distribuir datos y carga de trabajo en múltiples servidores (nodos), en lugar de depender de un solo servidor más grande (escalabilidad vertical).

Además, ayuda a la reducción de costos al utilizar servidores básicos en lugar de hardware especializado, y posee mejor tolerancia a fallos: si un nodo falla, los datos están replicados en otros nodos del clúster.

Estas características las posicionan como una solución ideal para los desafíos de la era del *Big Data* y las aplicaciones en tiempo real.

Ejemplos destacados de bases de datos NoSQL

Han surgido múltiples implementaciones destacadas que abordan necesidades específicas, como el manejo de datos estructurados, semiestructurados y no estructurados, la alta disponibilidad y el procesamiento en tiempo real. Veamos algunos ejemplos representativos de bases de datos NoSQL, explorando sus características principales y casos de uso.

Tipo: base de datos orientada a documentos.

Características claves: utiliza JSON (o BSON) para almacenar datos. Ideal para estructuras flexibles y dinámicas.

Caso de uso: un sitio de comercio electrónico que almacena información de productos con atributos variados (color, tamaño, precio) y permite modificaciones frecuentes.

Ventajas destacadas: consulta eficiente mediante índices e integración, con herramientas de análisis como Tableau o Power BI.

MongoDB

Cassandra

Redis

Neo4j

Los anteriores, son ejemplos destacados de bases de datos NoSQL demuestran su versatilidad y capacidad para adaptarse a las demandas tecnológicas actuales. Soluciones como MongoDB, Cassandra, Redis y Neo4j, ilustran cómo estas herramientas permiten gestionar datos de manera eficiente en aplicaciones como comercio electrónico, redes sociales, análisis en tiempo real y sistemas de recomendación. Su adopción continúa creciendo gracias a su capacidad para ofrecer rendimiento, escalabilidad y flexibilidad en escenarios cada vez más complejos y dinámicos.

Comparativa con bases de datos relacionales

La comparación entre bases de datos NoSQL y bases de datos relacionales, surge de la necesidad de entender cómo estas tecnologías responden a los crecientes desafíos del manejo de datos en la era digital. Mientras que las bases de datos relacionales, con su estructura rígida y lenguaje SQL, han sido el estándar durante décadas, las bases de datos NoSQL ofrecen un enfoque más flexible y escalable, para aplicaciones modernas. Analizar sus diferencias en aspectos como estructura, rendimiento, escalabilidad y casos de uso específicos, permite evaluar cuál es la solución más adecuada según los requerimientos del proyecto.

La elección entre ambas tecnologías, depende del contexto y las necesidades específicas, demostrando que no son soluciones excluyentes, sino complementarias en un ecosistema cada vez más diversificado y orientado a la innovación tecnológica.

Table 4. Comparativa con bases de datos relacionales

Aspecto	Relacionales	NoSQL
Estructura de datos	Estructura fija (tablas).	Estructura flexible (documentos, etc.).
Escalabilidad	Vertical (un servidor más grande).	Horizontal (múltiples nodos).
Flexibilidad	Limitada.	Alta
Casos de uso	Transacciones	

UNIDAD 1. INTRODUCCIÓN A BIG DATA

3. DESAFÍOS EN EL MANEJO DE GRANDES VOLÚMENES DE DATOS



Trabajar con *Big Data* implica enfrentar retos técnicos y estratégicos que van desde el almacenamiento eficiente, hasta el análisis de información en tiempo real. Este tema se centra en comprender estos desafíos y en explorar las soluciones tecnológicas disponibles para superarlos.

Escalabilidad y costos en el manejo de *Big Data*

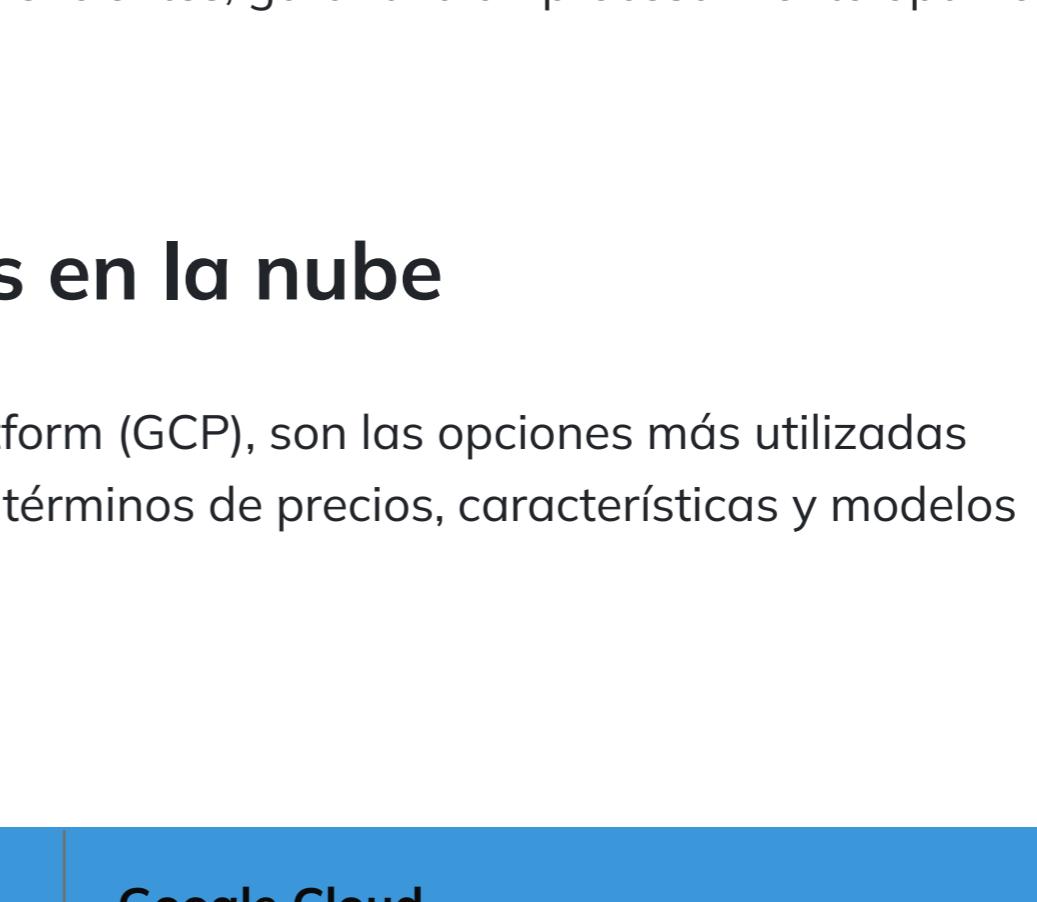
La capacidad de escalar de forma eficiente y el control de costos, son factores esenciales en el manejo de grandes volúmenes de datos. A medida que la generación de datos aumenta, las organizaciones deben adoptar estrategias sostenibles que combinen escalabilidad técnica y optimización de recursos financieros.

Escalabilidad horizontal	Optimización de costos	Estrategias de eficiencia
	<p>Escalabilidad horizontal</p> <p>Añadir más servidores o nodos, para distribuir la carga de almacenamiento y procesamiento.</p>	

Ejemplo aplicado

Una empresa de comercio electrónico enfrenta el desafío de almacenar y analizar millones de registros de ventas diarias:

- **Escalabilidad:** implementa hadoop, para distribuir los datos entre varios nodos.
- **Optimización:** utiliza servicios en la nube con políticas de pago por demanda, para evitar sobrecostos.
- **Procesamiento:** integra Apache Spark para realizar análisis en tiempo real, sobre patrones de compra y comportamiento del cliente.
- **Resultado:** la organización mejora la eficiencia de sus operaciones, reduce costos innecesarios y optimiza la toma de decisiones basada en datos precisos.



En conclusión, el manejo de grandes volúmenes de datos, requiere abordar los desafíos técnicos relacionados con la infraestructura, la velocidad, la escalabilidad y la integración de datos diversos. La adopción de tecnologías avanzadas y estrategias eficientes, garantiza un procesamiento óptimo y sostenible en el tiempo.

Comparativa de costos de almacenamiento en plataformas en la nube

Las plataformas en la nube como Amazon Web Services (AWS), Microsoft Azure y Google Cloud Platform (GCP), son las opciones más utilizadas para almacenar y procesar grandes volúmenes de datos. Sin embargo, estas plataformas difieren en términos de precios, características y modelos de uso.

Tabla 5. Comparativa de costos de almacenamiento en plataformas en la nube

Aspecto	AWS	Azure	Google Cloud
Modelo de precios	Pago por uso; descuentos por uso reservado y a largo plazo.	Pago por uso; descuentos por uso reservado y a largo plazo.	Pago por uso con descuentos para uso sostenido.
Costo por almacenamiento estándar	\$0.023/GB por mes para S3 Standard.	\$0.0184/GB por mes para Blob Storage.	\$0.020/GB por mes para Standard Storage.
Almacenamiento frío (Cold Storage)	\$0.004/GB por mes (Glacier).	\$0.001/GB por mes (Archive).	\$0.004/GB por mes (Coldline).
Ubicación de los centros de datos	Más de 25 regiones globales.	Más de 60 regiones globales.	Más de 35 regiones globales.
Integración de servicios	Integración robusta con servicios como Redshift, Athena y EMR.	Amplia compatibilidad con aplicaciones empresariales de Microsoft.	Herramientas avanzadas de análisis como BigQuery y Vertex AI.

En conclusión, se tiene que la AWS es ideal para proyectos que requieren una amplia gama de servicios y herramientas personalizadas, aunque puede ser más costoso en aplicaciones intensivas. Pero Azure ofrece una ventaja competitiva para empresas que ya utilizan herramientas de Microsoft, como Office 365 o Dynamics. Por su lado, GCP destaca por su simplicidad y su enfoque en análisis y aprendizaje automático, siendo una opción rentable para proyectos orientados al análisis de datos.

Estrategias para optimizar costos, mediante tecnologías eficientes

A pesar de las ventajas de las plataformas en la nube, los costos pueden aumentar rápidamente, si no se implementan estrategias adecuadas. A continuación, se presentan enfoques claves para optimizar los costos, asociados al manejo de datos masivos:

Uso de almacenamiento jerarquizado (Tiered Storage)	
<p>Dividir los datos en diferentes niveles de almacenamiento, según su frecuencia de uso:</p> <ul style="list-style-type: none"> • Hot Storage: datos de acceso frecuente (más costoso). • Cold Storage: datos de acceso ocasional. • Archive Storage: datos históricos o raramente utilizados. <p>Ejemplo práctico: una empresa de análisis financiero, puede almacenar datos recientes en almacenamiento caliente para análisis en tiempo real y mover registros históricos al almacenamiento archivado.</p>	

Ejemplos de soluciones aplicadas en Big Data

Se invita a conocer ejemplos de soluciones aplicadas en *Big Data*, donde herramientas como **Hadoop** y **Spark** han optimizado procesos y mejorado los servicios en sectores como **retail**, **banca**, **telecomunicaciones**, **salud** y **educación**. Los casos destacan logros como detección de fraudes en tiempo real, diagnósticos médicos más rápidos y estrategias educativas personalizadas, demostrando el impacto transformador del análisis de datos masivos.



Anexo. Ejemplos de soluciones aplicadas en Big Data





UNIDAD 1. INTRODUCCIÓN A BIG DATA

SÍNTESIS

Esta unidad brinda una base sólida sobre los principios y herramientas esenciales para gestionar grandes volúmenes de datos. Se analizan las "5V" de *Big Data* (volumen, velocidad, variedad, veracidad y valor), junto con tecnologías como **Hadoop**, **Spark** y bases de datos **NoSQL**, además de herramientas de visualización como **Power BI**, **Tableau** y **Qlik**.

