1. **DATOS GENERALES**

|  |  |  |
| --- | --- | --- |
| **RESPONSABLES EQUIPO DE PRODUCCIÓN** |  | Registro revisión |
| **Lider de producción** | Olga Constanza Bermúdez Jaimes | |
| **Nombre Asignatura:** | Big Data y Analítica de Datos | |
| **Experto Disciplinar** | Iván Guillermo Ortega Diez | Noviembre 2024 |
| **Asesor Tecno pedagógico** | Elizabeth Bermúdez Díez | Diciembre 2024 |
| **Diseñador Instruccional** | Paola Alexandra Moya Peralta | Diciembre 2024 |
| **Validador de contenidos** | María Sus |  |
| **Diseñador gráfico** |  |  |
| **Maquetador web** |  |  |

|  |  |
| --- | --- |
| **U1** | **Unidad 1.** Introducción a *Big Data* |
| **RESUMEN BANNER** | *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.* |

* + - 1. **INTRODUCCIÓN UNIDAD 1**

**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.

|  |  |
| --- | --- |
| <https://www.freepik.es/fotos-premium/big-data-sobre-fondo-digital-abstracto-almacen-datos-futurista-big-data-concepto-flujos-informacion-tecnologias-futuro-3d-render_29798942.htm#fromView=search&page=1&position=7&uuid=a6447873-6c03-4cf0-8663-fc4f4285d92e> | 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.

* + - 1. **SÍNTESIS UNIDAD 1**

|  |
| --- |
| 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**. |

**3. Desarrollo contenido Unidad 1**

**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.

|  |  |
| --- | --- |
| Fondo del concepto de visualización del diagrama Ilustración plana del fondo del concepto de vector de visualización del diagrama para diseño web  <https://www.freepik.es/vector-premium/fondo-concepto-visualizacion-diagrama-ilustracion-plana-fondo-concepto-vector-visualizacion-diagrama-diseno-web_65564027.htm#fromView=search&page=1&position=48&uuid=d3f35e3c-b9b4-44ba-844b-3084e7b18f8f> | **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 crítico**  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.

|  |  |
| --- | --- |
| <https://www.freepik.es/vector-gratis/fondo-movimiento-velocidad-velocimetro-rapido-fondo-velocidad-carreras_31288258.htm#fromView=search&page=1&position=9&uuid=84bb02e4-519f-4b02-a792-1c48181fb9c4> | * **Ejemplo aplicado.** En el comercio electrónico, Amazon procesa millones de transacciones por segundo, ajustando recomendaciones y precios casi instantáneamente. * **Reflexión.** Tecnologías como Apache Kafka o Spark, han permitido procesar flujo de datos en tiempo real, aunque siguen existiendo retos relacionados con la latencia y la sincronización. |

**Variedad**

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

| **PESTAÑAS** | | |
| --- | --- | --- |
| Estructurados | Datos de ventas organizados en tablas de bases de datos. | <https://www.freepik.es/vector-gratis/ilustracion-concepto-metodo-cascada_28901003.htm#fromView=search&page=1&position=12&uuid=bc534762-ff69-46ab-8729-007df53d912a> |
| No estructurados | Publicaciones en redes sociales, videos o documentos de texto. | <https://www.freepik.es/vector-gratis/telefono-movil-tema-purpura-rodeado-aplicaciones-coloridas_5453989.htm#fromView=search&page=1&position=1&uuid=bd821812-9221-4c7f-8766-a798d64e2963> |
| Semiestructurados | Registros de clientes en formularios web almacenados en JSON. | <https://www.freepik.es/vector-gratis/plantilla-formulario-contacto-diseno-plano_74880355.htm#fromView=search&page=1&position=2&uuid=a2df8f2e-4c16-496f-a7ea-f10b706152e2> |

**Desafío**: integrar estas fuentes heterogéneas requiere herramientas como NoSQL o procesos ETL, para unificar formatos y habilitar análisis coherentes.

**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 claves.** Algoritmos de validación y herramientas de limpieza de datos, se emplean para identificar errores y mejorar la confiabilidad de los análisis. | <https://www.freepik.es/vector-gratis/ilustracion-concepto-cinco-altos-trabajo-equipo_32318748.htm#fromView=search&page=1&position=8&uuid=5087c874-a646-47d2-820e-978ae8c26b6e> |

**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 del cliente.

**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. | <https://www.freepik.es/vector-gratis/ilustracion-concepto-redes-sociales_2807762.htm#fromView=search&page=1&position=1&uuid=bad93510-e881-40e3-a46c-7aa389aa4b0a> |

**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:

| **SLIDE** | | |
| --- | --- | --- |
| Escalabilidad horizontal | Capacidad de agregar nodos para aumentar el almacenamiento y la potencia de procesamiento. | <https://www.freepik.es/vector-gratis/ilustracion-siguiente-paso-degradado_37443123.htm#fromView=search&page=1&position=7&uuid=a68a3483-2b68-4e6b-94c3-ffa4a00ec2bd> |
| Tolerancia a fallos | Replicación automática de datos en diferentes nodos, para garantizar su disponibilidad. | <https://www.freepik.es/vector-premium/mujer-negocios-ayuda-rescatar-su-colega-escotilla_135040015.htm#fromView=search&page=1&position=10&uuid=42a20e34-a8b4-4aff-9856-156a38753743> |
| Ejemplo práctico | Una tienda en línea almacena historiales de compras y registros de usuarios en HDFS, distribuyendo los datos para garantizar acceso rápido y seguro. | <https://www.freepik.es/vector-gratis/concepto-guion-grafico-dibujos_10491221.htm#fromView=search&page=1&position=6&uuid=6114de7f-bdbe-449f-b4fb-4ecbbf5429ad> |

**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**

Combina 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.

|  |  |
| --- | --- |
| <https://www.freepik.es/vector-gratis/ilustracion-concepto-sala-servidores_82466230.htm#fromView=search&page=1&position=2&uuid=e3e9472f-cbe1-48c0-98a8-e142b4af6c69> | * **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 claves**

*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 Hadoop** | **Impacto** |
| **Banca y finanzas** | Detección de fraudes. | HDFS para almacenar transacciones, MapReduce para analizar patrones anómalos. | 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 mejora 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:

Velocidad

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

Flexibilidad

Compatible con múltiples lenguajes de programación (Python, Java, Scala).

Capacidades avanzadas

Incluye bibliotecas para aprendizaje automático (MLlib), procesamiento de gráficos (GraphX) y transmisión en tiempo real (Spark Streaming).

**Ejemplo aplicado**

Una empresa de comercio electrónico, usa Spark para procesar datos de navegación en tiempo real, ofreciendo recomendaciones personalizadas al instante.

**Tabla 2**

*Comparativa entre Hadoop y Spark*

|  |  |  |
| --- | --- | --- |
| **Aspecto** | **Hadoop** | **Spark** |
| **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. | Análisis 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 diversas industrias. 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: Hadoop 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: Spark 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 Hadoop, en ciertas tareas.  Capacidad para el procesamiento en tiempo real a través de **Spark Streaming**, una extensión que permite analizar flujos de datos continuos, casi en tiempo real. |
| 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).  Soporte 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:

| **ACORDEÓN** | | |
| --- | --- | --- |
| 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 historial 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. | <https://www.freepik.es/vector-gratis/eleccion-concepto-trabajador_9469552.htm#fromView=search&page=1&position=9&uuid=62d3e2dd-cbd0-4c5e-af93-285b71e0fe48> |
| Monitoreo de fraudes financieros | En instituciones bancarias, Spark analiza transacciones financieras en tiempo real, detectando patrones anómalos que podrían indicar fraudes. Por ejemplo, combinando datos de ubicación del usuario con patrones de gasto típicos, el sistema puede bloquear automáticamente actividades sospechosas.  **Impacto:** reducción de pérdidas por fraude y mayor seguridad para los clientes. | <https://www.freepik.es/vector-gratis/robar-concepto-datos_7940986.htm#fromView=search&page=1&position=1&uuid=9071e2e8-bb67-4256-af83-d79fe3e62a4a> |
| Optimización de redes de telecomunicaciones | Las empresas de telecomunicaciones utilizan Spark Streaming, para monitorizar en tiempo real, el tráfico en sus redes, identificando congestiones o interrupciones. Con estos datos, ajustan automáticamente la capacidad de la red o redirigen el tráfico.  **Impacto:** mejora de la calidad del servicio y reducción de tiempos de inactividad. | <https://www.freepik.es/vector-gratis/ilustracion-concepto-redes-sociales_3139755.htm#fromView=search&page=1&position=2&uuid=1941704c-ea63-4920-972e-cb3dae700da8> |
| Análisis de redes sociales y sentimiento del consumidor | Spark permite analizar millones de publicaciones en redes sociales, para detectar tendencias emergentes o cambios en la percepción de marca. Por ejemplo, en un lanzamiento de producto, las empresas pueden monitorear en tiempo real, las reacciones de los usuarios y ajustar estrategias de *marketing*.  **Impacto:** mejora en la agilidad de las campañas y en la comprensión del consumidor. | <https://www.freepik.es/vector-gratis/marketing-redes-sociales-sobre-diseno-telefonos_6600487.htm#fromView=search&page=1&position=4&uuid=a6004a52-5f21-4fbd-941f-503a711e5c64> |

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.

|  |  |
| --- | --- |
| <https://www.freepik.es/vector-gratis/ilustracion-alojamiento-sitio-web-degradado_22650925.htm#fromView=search&page=1&position=12&uuid=d0b66f3e-8f22-4b34-8f29-2511c01bd902> | * 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.

| **SLIDE** | | |
| --- | --- | --- |
| 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*. | <https://www.freepik.es/vector-gratis/logro-exito-aspiraciones-profesionales-promocion-laboral-crecimiento-personal-trabajador-motivado-empresario-volando-cohete-motivacion-determinacion-ilustracion-concepto_11668460.htm#fromView=search&page=1&position=6&uuid=c40f4859-c45c-4ad0-bdae-4da95589cc23> |
| Flexibilidad de esquemas | A diferencia de las bases de datos relacionales, NoSQL no requiere un esquema fijo. Esto significa que los datos pueden almacenarse en diversos formatos, como documentos JSON, columnas, grafos o claves-valor.  La adaptación a cambios dinámicos en la estructura de datos, es una de sus ventajas. Además es ideal para datos no estructurados, como imágenes, videos, *logs* de eventos y datos de sensores. | <https://www.freepik.es/vector-gratis/infografia-espina-pescado-diseno-plano_8514868.htm#fromView=search&page=1&position=9&uuid=76711eb8-ad74-4ea7-ba01-c21148dd5bf1> |
| Alto rendimiento | NoSQL está diseñado para ofrecer tiempos de respuesta rápidos, incluso con grandes volúmenes de datos y altas tasas de lectura/escritura.  Un ejemplo de ello, son las aplicaciones de redes sociales que deben manejar millones de interacciones, de usuarios, por segundo. | <https://www.freepik.es/vector-gratis/graduacion-universitaria-logro-educacion-superior-grado-academico-estudiante-exitoso-saltando-sosteniendo-birrete-desarrollo-personal_12085875.htm#fromView=search&page=1&position=4&uuid=21d3a2c7-e271-41bb-be5c-d74f5271cc38> |
| Distribución geográfica | Muchas bases de datos NoSQL están diseñadas para funcionar en sistemas distribuidos globalmente, asegurando acceso rápido y seguro a los datos, desde múltiples ubicaciones.  Por ejemplo, las bases de datos de servicios de *streaming,* donde los datos deben estar disponibles para usuarios, en diferentes regiones, en tiempo real. | <https://www.freepik.es/vector-gratis/concepto-moderno-comercio-internacional-diseno-plano_2533524.htm#fromView=search&page=1&position=0&uuid=ea799a32-c504-442d-9056-e7c01c442c17> |

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.

|  |  |  |
| --- | --- | --- |
| **INFOGRAFIA INTERACTIVA**  infografía api de gradiente  [**https://www.freepik.es/vector-gratis/infografia-api-gradiente\_25225772.htm#fromView=search&page=1&position=2&uuid=8703cf67-de3e-4007-a5dc-cbea503aa36b**](https://www.freepik.es/vector-gratis/infografia-api-gradiente_25225772.htm#fromView=search&page=1&position=2&uuid=8703cf67-de3e-4007-a5dc-cbea503aa36b) | | |
| MongoDB | **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. | <https://www.freepik.es/vector-gratis/conexion-red-cuatro-computadoras-portatiles-conexion-internet-almacenamiento-datos-nube-sala-servidores_3629653.htm#fromView=search&page=1&position=1&uuid=de7c9ff4-dd8a-42a8-9b86-b38ce33d4891> |
| Cassandra | **Tipo:** base de datos orientada a columnas.  **Características claves:** diseñada para manejar grandes volúmenes de datos, distribuidos en múltiples nodos con alta disponibilidad.  **Caso de uso:** empresas de telecomunicaciones que gestionan registros de llamadas (CDR) y analizan patrones de uso para optimizar redes.  **Ventajas destacadas:** escalabilidad sin interrupciones y rendimiento óptimo, con millones de transacciones por segundo. | <https://www.freepik.es/vector-gratis/ilustracion-colorida-mujer-programadora-haciendo-su-trabajo_5483074.htm#fromView=search&page=1&position=2&uuid=5c12e132-3dd3-409a-b642-875db50b49ca> |
| Redis | **Tipo:** almacén de datos en memoria, basado en clave-valor.  **Características claves:** velocidad ultrarrápida gracias al almacenamiento en memoria RAM y soporte de estructuras avanzadas como listas y conjuntos.  **Caso de uso:** sistemas de mensajería instantánea donde se requiere acceso casi inmediato a datos.  **Ventajas destacadas:** baja latencia, ideal para videojuegos en línea y análisis en tiempo real. | <https://www.freepik.es/vector-gratis/ilustracion-sql-tecnologia-abstracta_21743435.htm#fromView=search&page=1&position=1&uuid=76f2251b-7110-40a7-abc0-ab580c4dfb93> |
| Neo4j | **Tipo:** base de datos orientada a grafos.  **Características claves:** almacena datos en forma de nodos y relaciones, facilitando la representación de redes complejas.  **Caso de uso:** análisis de redes sociales para detectar comunidades, influenciadores y patrones de conexión.  **Ventajas destacadas:** consultas rápidas en datos interrelacionados. | <https://www.freepik.es/vector-gratis/ilustracion-concepto-arbol-redes-sociales_397052369.htm#fromView=search&page=1&position=2&uuid=8e13577d-7e4e-425b-a258-72c07db12cd6> |

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.

**Tabla 4**

*Comparativa con bases de datos relacionales*

|  |  |  |
| --- | --- | --- |
| Aspecto | Relacionales | NoSQL |
| Estructura de datos | Esquema fijo (tablas). | Esquema flexible (documentos, etc.). |
| Escalabilidad | Vertical (un servidor más grande). | Horizontal (múltiples nodos). |
| Flexibilidad | Limitada. | Alta. |
| Casos de uso | Transacciones bancarias. | *Big Data*, IoT, análisis en tiempo real. |

Aunque las bases de datos relacionales siguen siendo la opción preferida para sistemas que requieren consistencia estricta y estructuras bien definidas, las bases de datos NoSQL destacan en escenarios donde la flexibilidad, la escalabilidad horizontal y el manejo eficiente de datos heterogéneos son esenciales.

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.

**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.

| **PESTAÑAS** | | |
| --- | --- | --- |
| **Escalabilidad horizontal** | Añadir más servidores o nodos, para distribuir la carga de almacenamiento y procesamiento. | <https://www.freepik.es/foto-gratis/exitoso-hombre-negocios-caucasico-inteligente-caminando-acciones-comercio-escaleras-telefono-inteligente-comunica-felicidad-entusiasta-fondo-oficina-moderna_25118505.htm#fromView=search&page=1&position=9&uuid=f9ab60d8-364d-46fe-8c16-73b75bcf8a2a> |
| **Optimización de costos** | Utilizar plataformas de almacenamiento en la nube, como AWS, Azure o Google Cloud, que ofrecen modelos flexibles de pago por uso. | <https://www.freepik.es/foto-gratis/mujer-alto-angulo-trabajando-como-economista_36300788.htm#fromView=search&page=1&position=4&uuid=0b25da4b-1a61-4126-a410-e50fdd3fd54f> |
| **Estrategias de eficiencia** | Aplicar técnicas de compresión de datos y procesamiento en tiempo real, para reducir el consumo de recursos. | <https://www.freepik.es/foto-gratis/primer-plano-manos-sosteniendo-abierto-novela-taza-cafe_2862205.htm#fromView=search&page=1&position=4&uuid=da44df42-a6df-4fb5-9d32-56a2d770c536> |

**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. | <https://www.freepik.es/fotos-premium/foto-recortada-compras-linea-mujer-utilizando-tecnologia-dispositivo_3965844.htm#fromView=search&page=1&position=9&uuid=b65df767-2882-4a9b-ae2a-20d55593828f> |

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 compromisos 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:

| **SLIDE** | |
| --- | --- |
| **Uso de almacenamiento jerarquizado *(Tiered Storage)*** | Dividir los datos en diferentes niveles de almacenamiento, según su frecuencia de uso:   * + ***Hot Storage*:** datos de acceso frecuente (más costoso).   + ***Cold Storage:*** datos de acceso ocasional.   + ***Archive Storage*:** datos históricos o raramente utilizados.   **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. |
| **Compresión de datos** | Reducir el tamaño de los datos almacenados, mediante técnicas de compresión como *gzip, zstd o snappy.*  **Impacto:** menor consumo de almacenamiento y reducción de costos asociados. |
| **Eliminación de datos redundantes** | Implementar procesos de limpieza para identificar y eliminar datos duplicados o innecesarios.  **Beneficio:** optimización del espacio de almacenamiento sin afectar la integridad de los datos. |
| **Automatización de escalabilidad** | Configurar sistemas en la nube, para escalar automáticamente, según la demanda.  **Ejemplo:** servicios como AWS Auto Scaling o Azure Scale Sets, ajustan los recursos, según el tráfico o la carga de trabajo, evitando costos innecesarios en períodos de baja actividad. |
| **Adopción de tecnologías de código abierto** | Utilizar herramientas de código abierto como Hadoop y Spark en lugar de soluciones comerciales más costosas. Estas tecnologías ofrecen escalabilidad y flexibilidad, sin costos elevados de licencias.  **Consideración:** aunque iniciales bajos, requieren experiencia técnica para configuraciones complejas. |
| **Optimización del procesamiento** | Diseñar flujos de trabajo para minimizar la repetición de tareas intensivas en recursos.  **Ejemplo:** en lugar de procesar datos completos repetidamente, usar muestras o datos previamente preprocesados para tareas exploratorias. |
| **Monitoreo y auditoría continua** | Implementar herramientas de monitoreo como AWS Cost Explorer, Azure Cost Management o Google Cloud Cost Management, para identificar patrones de uso y oportunidades de ahorro.  **Práctica recomendada:** configurar alertas para evitar sobrecostos inesperados. |

La escalabilidad y los costos son dos caras de la misma moneda en *Big Data*. Aprovechar plataformas en la nube como AWS, Azure y Google Cloud, es esencial para manejar grandes volúmenes de datos, pero hacerlo de manera eficiente requiere estrategias claras para optimizar los costos. Desde el uso de almacenamiento jerarquizado, hasta la adopción de tecnologías de código abierto y monitoreo constante, las organizaciones pueden maximizar el retorno de inversión en sus proyectos de *Big Data* mientras garantizan la sostenibilidad a largo plazo.

|  |  |
| --- | --- |
| PDF Viewer(Reader) & PDF Creat - Aplicaciones en Google Play | **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. |

1. **GLOSARIO DE LA UNIDAD**

|  |  |
| --- | --- |
| **Relacione  las definiciones de los términos claves, requeridas para comprender adecuadamente los contenidos de esta unidad. Presentarlo en Orden Alfabético. Máximo 20 palabras.** | |
| **PALABRA, TÉRMINO O ABREVIATURA** | **SIGNIFICADO** |
| **Analítica predictiva** | Técnica que utiliza datos históricos y modelos estadísticos, para predecir eventos futuros. |
| **Analítica prescriptiva** | Proceso que sugiere acciones específicas basadas en los análisis predictivos y descriptivos. |
| **Analítica descriptiva** | Método que analiza datos históricos para comprender tendencias y patrones. |
| ***Big Data*** | Conjunto de datos grandes, complejos y variados, que requieren nuevas tecnologías y herramientas para su procesamiento. |
| **Hadoop** | Plataforma de *software* de código abierto, diseñada para el almacenamiento y procesamiento distribuido de grandes volúmenes de datos. |
| **HDFS *(Hadoop Distributed File System)*** | Sistema de archivos distribuidos que almacena datos en múltiples servidores, asegurando escalabilidad y tolerancia a fallos. |
| ***MapReduce*** | Modelo de programación utilizado por Hadoop para procesar grandes conjuntos de datos de manera distribuida. |
| **MongoDB** | Base de datos NoSQL orientada a documentos que utiliza JSON para el almacenamiento de datos. |
| **NoSQL** | Tipo de base de datos diseñada para manejar grandes volúmenes de datos no estructurados y semiestructurados. |
| ***Power BI*** | Herramienta de Microsoft para visualización y análisis de datos, útil en la toma de decisiones empresariales. |
| ***Spark*** | Plataforma de procesamiento de datos, en tiempo real, que opera en memoria para acelerar el análisis. |
| **SQL *(Structured Query Language*)** | Lenguaje utilizado para gestionar y consultar bases de datos relacionales. |
| **Tableau** | Herramienta de visualización que permite transformar datos en gráficos interactivos y *dashboards.* |
| **Variedad** | Característica de *Big Data* que se refiere a la diversidad de formatos de datos, incluyendo estructurados y no estructurados. |
| **Velocidad** | Propiedad de *Big Data* que mide la rapidez con la que se generan y procesan los datos. |
| **Veracidad** | Calidad y confiabilidad de los datos, esencial para tomar decisiones informadas. |
| **Valor** | Capacidad de los datos para proporcionar información útil y relevante. |
| **Visualización de datos** | Representación gráfica de datos para facilitar su análisis y comprensión. |
| **Volumen** | Cantidad masiva de datos que caracteriza a *Big Data*. |
| **Qlik** | Plataforma de visualización y análisis de datos utilizada para crear informes interactivos. |

1. **MATERIAL DE APOYO (SIEMPRE SE DEBE INCLUIR AL MENOS DOS RECURSOS BIBILOGRÀFICOS DE APOYO, DOS BASES DE DATOS Y DOS RECURSOS AUDIOVISUALES. SE DEBE INCORPORAR COMO MÍNIMO 2 RECURSOS DE E LIBRO)**

|  |  |  |
| --- | --- | --- |
| **RECURSOS BIBLIOGRÁFICOS DE REVISIÓN BÁSICA**  Incluye libros impresos, digitales, electrónicos, ebook, artículos de revistas impresas, electrónicas, informes, seminarios, congresos, ponencias, presentaciones o diapositivas online, etc. Desde el 2019 en adelante excepto casos excepcionales. Bases de datos (licenciadas por la institución), bases de datos libres, videos bajo licenciamiento creative commons | | **BASES DE DATOS**  Si las referencias bibliográficas son de las bases de datos institucionales mencione únicamente el nombre de la base de datos donde se encuentra el recurso. |
| T1 | Castillo Romero, J. A. (2019). *Big Data*. IFCT 128P. IC Editorial.  <https://elibro.net/es/lc/tecnologicadeloriente/titulos/124254> | [https://elibro.net](https://elibro.net/es/lc/tecnologicadeloriente/titulos/59093) |
| T2 | López Garzón, W. F. & Cárdenas López, J. C. (2019). Tecnología internet of things (IoT) y el Big Data. Mare Ingenii. *Ingenierías,* *1*(1), pp. 74-79. Fundación Universitaria San Mateo.  <https://elibro.net/es/lc/tecnologicadeloriente/titulos/127847> | [https://elibro.net](https://elibro.net/es/lc/tecnologicadeloriente/titulos/117744) |
| T3 | Nin, J. (2016). *Big Data: análisis de datos en entornos masivos*. Editorial UOC.  <https://elibro.net/es/ereader/tecnologicadeloriente/117744> | [https://elibro.net](https://elibro.net/es/lc/tecnologicadeloriente/titulos/59093) |
| **RECURSOS BIBLIOGRÁFICOS DE REVISIÓN COMPLEMENTARIA**  Incluye libros impresos, digitales, electrónicos, ebook, artículos de revistas impresas, electrónicas, informes, seminarios, congresos, ponencias, presentaciones o diapositivas online, etc. Desde el 2019 en adelante excepto casos excepcionales. Bases de datos (licenciadas por la institución), bases de datos libres, videos bajo licenciamiento creative commons | | **BASES DE DATOS**  Si las referencias bibliográficas son de las bases de datos institucionales mencione únicamente el nombre de la base de datos donde se encuentra el recurso. |
| T1 | Holmes, D. E. (2018). *Big Data: una breve introducción*. Antoni Bosch Editor.  <https://elibro.net/es/lc/tecnologicadeloriente/titulos/122682> | [https://elibro.net/](https://elibro.net/es/lc/tecnologicadeloriente/titulos/122682) |
| T2 | López Murphy, J. J. & Zarza, G. (2017). *La ingeniería del Big Data: cómo trabajar con datos.* Editorial UOC.  <https://elibro.net/es/lc/tecnologicadeloriente/titulos/59093> | [https://elibro.net/](https://elibro.net/es/lc/tecnologicadeloriente/titulos/122682) |
| T3 | Williamson, B. (2018). *Big Data en educación*. Ediciones Morata, S.L.  <https://elibro.net/es/lc/tecnologicadeloriente/titulos/119511> | [https://elibro.net/](https://elibro.net/es/lc/tecnologicadeloriente/titulos/122682) |
| **RECURSOS AUDIOVISUALES**  Incluye videos, podcast, audiolibros, grabaciones sonoras o musicales, que se encuentran alojados en internet. | | |
| T1 | LT DATA CHANNEL. (2024). *¿QUÉ ES EL BIG DATA Y PARA QUÉ SIRVE?*[video]. YouTube.  <https://www.youtube.com/watch?v=T31CpSQD-hc> | |
| T2 | UdeCataluña. (2020). *El valor del Big Data y la analítica avanzada en la era digital* [video]. YouTube  <https://www.youtube.com/watch?v=Rz_7C-mdk8c> | |
| T3 | UNRN. (2023). *El desafío del Gobierno de Datos en las Grandes Empresas* [video]. YouTube  <https://www.youtube.com/live/1qP5jQyUdIw?si=ni7eEEwdVWzwXZLS> | |
| **PÁGINAS WEB DE REVISIÓN BÁSICA**  Incluye blogs, hipertextos, animaciones, etc. | | |
| 1 | HubSpot. (2024). *¿Qué es Big Data?* HubSpot. <https://blog.hubspot.es/website/que-es-big-data> | |
| 2 | AICAD Business School. (s.f.). *Big Data: Qué es, para qué sirve y cómo se utiliza*. <https://www.aicad.es/big-data> | |
| 3 | Intelequia. (s.f.). *Big Data: el motor de la transformación empresarial en la era digital*. Intelequia. <https://intelequia.com/es/blog/post/big-data-el-motor-de-la-transformaci%C3%B3n-empresarial-en-la-era-digital>. | |

|  |  |  |
| --- | --- | --- |
| **LABORATORIOS, SIMULADORES, SOFTWARE, APP´S Y HERRAMIENTAS DE IA DE APOYO** | | **LINK DE ACCESO** |
| 1 | Comience su prueba gratuita de Tableau | <https://www.tableau.com/es-es/products/trial> |
| 2 | Obtener las herramientas y aplicaciones Power BI que se necesitan. | <https://www.microsoft.com/es-es/power-platform/products/power-bi/downloads> |
| 3 | Es un *software* de código abierto, que permite almacenar y procesar grandes cantidades de datos en clústeres de computadoras | https://hadoop.apache.org/ |