UD1: INTRODUCCIÓN AL BIG DATA. DEL DATO A LA INFORMACIÓN

Introducción: La Revolución de los Datos y el Nacimiento del Big Data

El término "Big Data" ha trascendido el ámbito técnico para convertirse en un pilar de la estrategia empresarial moderna. Sin embargo, es fundamental delimitar su significado más allá de la simple noción de "muchos datos". Big Data se refiere a conjuntos de datos cuyo tamaño, velocidad de generación o complejidad son tales que las herramientas tradicionales de procesamiento y gestión de datos resultan insuficientes para su captura, almacenamiento, análisis y visualización.[2] Esta definición no solo establece un umbral tecnológico, sino que también implica la necesidad de nuevos paradigmas, arquitecturas y perfiles profesionales para extraer valor de este nuevo recurso.

Para comprender el propósito fundamental de trabajar con Big Data, es esencial visualizar el viaje que transforma los hechos brutos en decisiones estratégicas. Este proceso se conceptualiza a través de la Pirámide DIKW (Datos, Información, Conocimiento, Sabiduría), un modelo jerárquico que representa la cadena de valor de cualquier iniciativa de datos.[4]

- **Dato (Data):** Es el nivel más bajo, compuesto por hechos crudos, símbolos y observaciones sin procesar ni contextualizar. En su estado aislado, un dato tiene un significado limitado o nulo.[6] Por ejemplo, la secuencia de números
  12012012 es un dato sin contexto.[4] De manera similar, una lectura de un sensor que indica
  37.5 es solo un número.
- **Información (Information):** Los datos se convierten en información cuando se organizan, estructuran y se les añade contexto. La información responde a las preguntas fundamentales: "quién", "qué", "cuándo" y "dónde".[7] La secuencia 12012012 se transforma en información cuando se estructura como 12/01/2012, una fecha específica. La lectura 37.5 se convierte en información cuando se contextualiza como "la temperatura corporal del paciente A a las 08:00 es de 37.5 grados Celsius".
- **Conocimiento (Knowledge):** El conocimiento emerge de la interpretación y la conexión de la información, respondiendo a las preguntas de "cómo" y "por qué".[4] Se trata de identificar patrones, relaciones y principios a partir de la información disponible.[6] Siguiendo el ejemplo meteorológico, si la información es "la temperatura bajó, la humedad subió rápidamente y la presión atmosférica descendió en El Cairo a las 10:00 a.m.", el conocimiento es la comprensión del patrón: "la combinación de estos factores meteorológicos suele preceder a la lluvia".[7]

- **Sabiduría (Wisdom):** Es el pináculo de la pirámide y representa el conocimiento aplicado en acción para tomar decisiones óptimas. La sabiduría responde a la pregunta de "qué es lo mejor" y se enfoca en el futuro, utilizando el entendimiento adquirido para guiar las acciones.[4] Con el conocimiento de que es probable que llueva, la sabiduría se manifiesta en la decisión: "alertar a los servicios de emergencia de la ciudad para que preparen los sistemas de drenaje y advertir a la población sobre posibles inundaciones".

Este modelo no es meramente teórico; es el mapa de ruta que define el flujo de trabajo de un equipo de datos moderno y la interacción entre sus roles. El viaje de **Dato a Información** es el dominio principal del **Ingeniero de Datos**, quien construye la infraestructura para recolectar, limpiar y estructurar los datos brutos, dotándolos de contexto y haciéndolos accesibles.[8] La transición de
Información a Conocimiento es donde el **Analista de Datos** y el **Científico de Datos** entran en juego. El analista examina la información histórica para descubrir patrones y responder a preguntas de negocio, mientras que el científico de datos construye modelos predictivos para generar nuevo conocimiento sobre el futuro.[10] Finalmente, el salto a la
Sabiduría es la culminación del proceso, donde los líderes empresariales, asesorados por los equipos de datos, utilizan este conocimiento para tomar decisiones estratégicas que impulsan el negocio.[12] Así, la Pirámide DIKW no solo ilustra la transformación del dato, sino que también justifica la existencia y colaboración de los distintos roles en el ecosistema de Big Data.

Sección 1: Los Pilares Conceptuales - Las 5 Vs del Big Data

Para caracterizar los desafíos y las oportunidades del Big Data, la industria ha adoptado un modelo basado en cinco dimensiones conocidas como las "5 Vs". Estas no son características aisladas, sino que interactúan y se amplifican mutuamente, definiendo la complejidad inherente a cualquier proyecto de Big Data.

1.1. Volumen: La Escala del Desafío

El Volumen se refiere a la inmensa cantidad de datos que se generan, almacenan y procesan.[14] Hablamos de magnitudes que superan con creces las capacidades de los sistemas de bases de datos tradicionales, entrando en el terreno de los terabytes (TB), petabytes (PB) y exabytes (EB). Un petabyte equivale a un millón de gigabytes.[16]

- **Ejemplo del Mundo Real (Retail):** La cadena de supermercados Walmart es un caso paradigmático. Opera aproximadamente 10,500 tiendas en 24 países y gestiona más de un millón de transacciones de clientes cada hora. Este flujo constante de datos de

ventas, inventario y logística se traduce en la importación de más de 2.5 petabytes de datos por hora a su infraestructura, que constituye una de las nubes privadas más grandes del mundo.[16]

- **Ejemplo del Mundo Real (Redes Sociales):** Plataformas como Twitter, Instagram o Facebook generan un volumen masivo de datos cada segundo a través de publicaciones, comentarios, "me gusta" y subidas de contenido multimedia. Se estima que para 2025, la cantidad de información en línea alcanzará cifras difíciles de imaginar, impulsada en gran medida por estas interacciones.[15]

1.2. Velocidad: El Flujo Incesante de Datos en Tiempo Real

La Velocidad se refiere al ritmo al que los datos se generan, se mueven y deben ser procesados para ser útiles.[14] En muchos escenarios de negocio, el valor de los datos disminuye drásticamente con el tiempo, lo que exige capacidades de procesamiento en tiempo real o casi real.

- **Ejemplo del Mundo Real (Finanzas):** Los sistemas de detección de fraude con tarjetas de crédito son un ejemplo crítico de la necesidad de velocidad. Deben analizar flujos de millones de transacciones en tiempo real para identificar patrones anómalos y bloquear una compra fraudulenta en milisegundos. Un retraso de incluso unos pocos segundos puede suponer pérdidas financieras significativas.[17]
- **Ejemplo del Mundo Real (Salud):** En una unidad de cuidados intensivos, los monitores de constantes vitales (frecuencia cardíaca, saturación de oxígeno, presión arterial) generan un flujo continuo de datos. Estos datos deben ser procesados instantáneamente para generar alertas que puedan salvar la vida de un paciente.[17]

1.3. Variedad: El Ecosistema de Datos

La Variedad describe la diversidad de tipos y formatos de datos que las organizaciones deben gestionar.[14] El Big Data abarca mucho más que los datos numéricos tradicionales y se clasifica principalmente en tres categorías [20]:

- **Datos Estructurados:** Son datos altamente organizados que se ajustan a un modelo de datos predefinido, como las filas y columnas de una base de datos relacional o una hoja de cálculo. Son fáciles de almacenar, consultar y analizar. Ejemplo: un registro de ventas con campos como ID_Cliente, ID_Producto, Fecha y Precio.[15]
- **Datos No Estructurados:** No tienen un modelo de datos predefinido ni una organización interna. Componen la gran mayoría de los datos del mundo. Ejemplos incluyen texto libre (correos electrónicos, comentarios en redes sociales, documentos), imágenes, vídeos y archivos de audio.[15]

- **Datos Semiestructurados:** No se ajustan a la estructura formal de los modelos de datos relacionales, pero contienen etiquetas u otros marcadores para separar elementos semánticos y hacer cumplir jerarquías de registros y campos. Ejemplos comunes son los archivos JSON y XML.[15]
- **Ejemplo del Mundo Real (E-commerce):** Un gigante como Amazon consolida una enorme variedad de datos para personalizar la experiencia del usuario. Recopila historiales de compra (estructurados), reseñas de productos en texto (no estructurados), imágenes y vídeos de productos (no estructurados), búsquedas por voz a través de Alexa (audio no estructurado) y logs de clics en la web (semiestructurados).[17]

1.4. Veracidad: La Búsqueda de la Calidad y Confianza

La Veracidad se refiere a la calidad, precisión, fiabilidad y confianza de los datos.[17] Los datos del mundo real son inherentemente "sucios": pueden contener errores, sesgos, inconsistencias, valores ausentes y ruido. Garantizar la veracidad es uno de los mayores desafíos del Big Data, ya que las decisiones basadas en datos de mala calidad pueden ser peores que no tomar ninguna decisión.[16]

- **Ejemplo del Mundo Real (Análisis de Sentimiento):** Una empresa que intenta medir la percepción de su marca analizando comentarios en Twitter se enfrenta a un gran problema de veracidad. El sistema debe ser capaz de interpretar el sarcasmo, corregir errores tipográficos, filtrar el spam de bots y entender la jerga local. Un análisis que tome un comentario sarcástico como positivo puede llevar a conclusiones completamente erróneas sobre la opinión pública.[16]
- **Ejemplo del Mundo Real (Salud):** La veracidad en los historiales clínicos electrónicos es de vital importancia. Un error en la dosis de un medicamento, una alergia no registrada o un diagnóstico incorrecto pueden tener consecuencias fatales para la seguridad del paciente.[18]

1.5. Valor: El Objetivo Final

El Valor es, desde una perspectiva de negocio, la "V" más importante. Se refiere a la capacidad de transformar el vasto y complejo universo de datos en conocimientos accionables que generen un beneficio tangible, ya sea a través de la optimización de operaciones, la mejora de la experiencia del cliente, la creación de nuevos productos o la mitigación de riesgos.[14]

- **Ejemplo del Mundo Real (Entretenimiento):** Netflix es un maestro en la extracción de valor a partir de los datos. Analiza miles de millones de eventos diarios: qué títulos ven

los usuarios, cuándo pausan, qué búsquedas realizan, qué contenido abandonan, etc. Estos datos alimentan su sofisticado motor de recomendación, que es responsable de más del 80% del contenido consumido en la plataforma.[22] Se estima que este sistema de personalización ahorra a la compañía más de 1.000 millones de dólares al año en reducción de la tasa de cancelación de suscripciones (churn), ya que mantiene a los usuarios comprometidos y satisfechos.[23]

Las 5 Vs no son desafíos independientes, sino un sistema de complejidad interconectado. El verdadero reto del Big Data surge de su interacción. Por ejemplo, un alto **Volumen** y una alta **Velocidad** de datos con una gran **Variedad** hacen que garantizar la **Veracidad** sea exponencialmente más difícil. Es casi imposible limpiar y validar manualmente un flujo de datos de petabytes que llega en tiempo real. La capacidad de una organización para construir una arquitectura que gestione simultáneamente estas cuatro dimensiones es lo que, en última instancia, determina el **Valor** que puede extraer. Una estrategia de Big Data no consiste en resolver cada "V" por separado, sino en abordar su efecto combinado y compuesto.

Característica	Retail	Finanzas	Salud	Entretenimiento (Streaming)
Volumen	Millones de transacciones por hora y datos de inventario de miles de tiendas (ej. Walmart).[16]	Billones de registros de transacciones bursátiles y datos de clientes a nivel global.	Historiales clínicos electrónicos (EHR), imágenes médicas (Rayos X, RMN) y datos genómicos para millones de pacientes.	Petabytes de datos de visualización, interacciones de usuario y catálogo de contenidos para más de 200 millones de suscriptores (ej. Netflix).[23]
Velocidad	Actualizaciones de inventario en tiempo real y análisis de patrones de compra para ofertas dinámicas.	Detección de fraude en milisegundos analizando flujos de transacciones en vivo.[17]	Monitorización de constantes vitales en tiempo real desde dispositivos médicos en UCI.[17]	Procesamiento instantáneo de interacciones del usuario (play, pause, search) para actualizar recomendaciones en tiempo real.
Variedad	Datos de ventas (estructurados), feedback en redes sociales	Transacciones (estructuradas), correos electrónicos de	Registros médicos (estructurados), notas del doctor (texto), imágenes	Metadatos de contenido (estructurados), valoraciones de

	(texto), vídeos de seguridad (vídeo), logs de web (semiestructurados).	clientes (texto), llamadas de soporte grabadas (audio), documentos de análisis de mercado (PDF).	de RMN (imagen), datos de sensores wearables (series temporales).	usuarios (numéricos), búsquedas de texto (texto), imágenes de portada personalizadas (imagen).
Veracidad	Asegurar la consistencia del inventario entre el sistema online y las tiendas físicas. Filtrar reseñas de productos falsas.	Identificar y eliminar transacciones erróneas o duplicadas. Mitigar el sesgo en los modelos de calificación crediticia.	Garantizar la precisión de los datos del paciente para evitar errores de diagnóstico o medicación.[18]	Limpiar datos de visualización para distinguir entre visualizaciones intencionadas y accidentales. Eliminar el ruido de los datos de interacción.
Valor	Optimización de la cadena de suministro, personalización de ofertas y predicción de la demanda.	Detección de fraude, gestión de riesgos, trading algorítmico y marketing personalizado.	Medicina personalizada, predicción de brotes de enfermedades, optimización de la gestión hospitalaria.	Retención de clientes a través de recomendaciones personalizadas, optimización del gasto en contenido y mejora de la experiencia de usuario.[22]

Sección 2: El Recorrido del Dato - El Ciclo de Vida en la Ingeniería de Datos

Para gestionar la complejidad descrita por las 5 Vs, los datos siguen un viaje estructurado conocido como el ciclo de vida de los datos. Este ciclo describe las etapas por las que pasan los datos desde su creación hasta su eventual destrucción, proporcionando un marco para las operaciones de ingeniería de datos.

2.1. Fase 1: Generación y Captura

Esta es la fase de origen, donde los datos nacen. Se generan a partir de una multitud de fuentes, como interacciones de clientes en una web, transacciones en un punto de venta, sensores en dispositivos de IoT, publicaciones en redes sociales, logs de aplicaciones o la introducción manual de datos.[24] La calidad y la estructura de los datos en esta fase inicial tienen un impacto directo en la complejidad y el coste de todas las etapas posteriores.

- **Ejemplo:** Un vehículo autónomo genera continuamente terabytes de datos a partir de sus sensores (cámaras, LiDAR, GPS, radar) para percibir su entorno.

2.2. Fase 2: Ingesta y Almacenamiento

Una vez generados, los datos deben ser trasladados desde su fuente a un sistema de almacenamiento centralizado. Este proceso se conoce como ingesta. La ingesta puede ser de dos tipos principales: por lotes (batch), donde los datos se recopilan y procesan en grandes bloques a intervalos programados, o en tiempo real (streaming), donde los datos se procesan continuamente a medida que se generan.[26] El almacenamiento también es una decisión crítica, que implica elegir entre arquitecturas como Data Lakes o Data Warehouses. Una estrategia común es clasificar los datos por su "temperatura" según la frecuencia de acceso: datos "calientes" (acceso frecuente) se almacenan en sistemas rápidos, mientras que datos "fríos" (acceso infrecuente) se archivan en soluciones de bajo coste.[26]

- **Ejemplo:** Una plataforma de e-commerce utiliza Apache Kafka para la ingesta en tiempo real de los clics de los usuarios en su sitio web. Estos datos de eventos se almacenan en su formato crudo en un Data Lake basado en un servicio de almacenamiento en la nube como Amazon S3.

2.3. Fase 3: Procesamiento

Los datos crudos rara vez son útiles para el análisis directo. La fase de procesamiento se encarga de transformar estos datos en un formato limpio, consistente y utilizable. Este es el corazón de los procesos ETL (Extract, Transform, Load) o ELT (Extract, Load, Transform). Las tareas comunes incluyen la limpieza de datos (corregir errores, imputar valores ausentes), la transformación (convertir formatos de fecha, estandarizar unidades), la integración (combinar datos de múltiples fuentes) y el enriquecimiento (añadir información adicional, como datos demográficos a un registro de cliente).[24]

- **Ejemplo:** Un pipeline de datos construido con Apache Spark lee los datos crudos de ventas desde el Data Lake, elimina registros duplicados, estandariza las direcciones de los clientes, une los datos con información de productos de otra base de datos y calcula el margen de beneficio para cada transacción.

2.4. Fase 4: Análisis

Con los datos ya procesados y listos, comienza la fase de análisis. Aquí es donde se extraen los conocimientos y el valor. Los analistas y científicos de datos utilizan una variedad de técnicas, como modelado estadístico, algoritmos de machine learning, minería de datos e inteligencia de negocio, para identificar patrones, descubrir correlaciones, predecir tendencias y responder a preguntas de negocio complejas.[24]

- **Ejemplo:** Un científico de datos utiliza el conjunto de datos de ventas ya procesado para entrenar un modelo de aprendizaje automático que predice la probabilidad de que un cliente abandone la empresa en los próximos tres meses (análisis de churn).

2.5. Fase 5: Visualización y Consumo

Los resultados del análisis deben ser comunicados de una manera que sea comprensible y accionable para los responsables de la toma de decisiones. La visualización de datos juega un papel clave, transformando números y tablas complejas en gráficos, mapas y cuadros de mando interactivos a través de herramientas como Tableau o Power BI.[24] Además, los datos o los resultados del modelo pueden ser servidos a través de una API para que otras aplicaciones o sistemas puedan consumirlos.[10]

- **Ejemplo:** Un analista de negocio crea un dashboard en Tableau que muestra las predicciones de churn por región, segmento de cliente y producto, permitiendo al equipo de marketing diseñar campañas de retención específicas.

2.6. Fase 6: Gobernanza, Ética y Destrucción

Esta no es una fase final, sino una capa transversal que abarca todo el ciclo de vida. Incluye la **gobernanza de datos**, que establece políticas para garantizar la calidad, el linaje y la seguridad de los datos.[27] La
ética de datos aborda las consideraciones morales sobre cómo se recopilan y utilizan los datos para evitar sesgos y discriminación.[29] Finalmente, las políticas de retención y destrucción definen cuánto tiempo se deben conservar los datos y cómo eliminarlos de forma segura para cumplir con regulaciones como el GDPR.[25]

- **Ejemplo:** Una empresa define una política de gobernanza que establece que todos los datos personales de los clientes deben ser anonimizados después de 5 años y eliminados de forma segura de todos los sistemas después de 7 años para cumplir con las normativas de privacidad.

Es crucial entender que el ciclo de vida de los datos en la práctica moderna no es un proceso

lineal en cascada, sino un ciclo iterativo y continuo. Los hallazgos en la fase de 'Análisis' a menudo retroalimentan y provocan cambios en las fases de 'Procesamiento' o 'Ingesta' (por ejemplo, "necesitamos recopilar un nuevo punto de datos para mejorar la precisión de nuestro modelo"). Este enfoque, conocido como **DataOps**, aplica principios de DevOps al mundo de los datos, enfatizando la automatización, la monitorización y los bucles de retroalimentación continuos para acelerar la entrega de valor.[27]
Dentro de este ciclo, la elección entre **ETL** y **ELT** es una decisión arquitectónica fundamental. El enfoque tradicional **ETL**, común en los Data Warehouses, transforma los datos *antes* de cargarlos en el sistema de destino. Esto garantiza que los datos en el almacén sean de alta calidad y estén estandarizados, pero puede ser un proceso lento y rígido. El enfoque moderno **ELT**, habilitado por la flexibilidad y el bajo coste de los Data Lakes, carga los datos crudos primero y los transforma *después*, solo cuando es necesario para un análisis específico. Esto es mucho más rápido y flexible para la ingesta, pero traslada la responsabilidad de la limpieza y la validación al momento del análisis, lo que puede aumentar la complejidad para los usuarios finales.[31] La elección entre ambos modelos refleja la estrategia de datos de una organización: ¿se prioriza la consistencia y los informes estandarizados (ETL) o la flexibilidad y la exploración para la ciencia de datos (ELT)?

Sección 3: La Arquitectura Fundacional y su Evolución

La capacidad de gestionar el ciclo de vida del Big Data a escala requirió una revolución en la arquitectura de software. El punto de partida de esta revolución fue el ecosistema Hadoop, que sentó las bases para el procesamiento distribuido y que ha evolucionado hacia las arquitecturas modernas basadas en la nube que dominan la industria hoy en día.

3.1. El Ecosistema Hadoop Clásico: La Base de Todo

Apache Hadoop es un framework de código abierto que permitió, por primera vez, el almacenamiento y procesamiento de enormes volúmenes de datos utilizando clústeres de hardware de bajo coste (commodity hardware).[33] Su arquitectura clásica se basa en tres componentes principales:

- **HDFS (Hadoop Distributed File System):** Es la capa de almacenamiento. HDFS es un sistema de archivos distribuido diseñado para ser tolerante a fallos y proporcionar un alto rendimiento en el acceso a los datos. Su arquitectura maestro-esclavo consiste en un **NameNode** (maestro), que gestiona los metadatos del sistema de archivos (la estructura de directorios y la ubicación de los bloques de datos), y múltiples **DataNodes** (esclavos), que almacenan los bloques de datos reales. Para garantizar la tolerancia a fallos, HDFS divide los archivos grandes en bloques (por ejemplo, de 128

MB) y replica cada bloque varias veces (generalmente tres) en diferentes DataNodes del clúster.[35]

- **MapReduce:** Fue el paradigma de procesamiento original de Hadoop. Es un modelo de programación para procesar grandes conjuntos de datos en paralelo a través de un clúster. Un trabajo MapReduce se divide en dos fases principales: la fase **Map**, que toma los datos de entrada, los filtra y los transforma en pares clave-valor; y la fase **Reduce**, que recibe la salida de la fase Map, agrega los datos por clave y produce un resultado final resumido.[34] MapReduce fue diseñado para el procesamiento por lotes (batch) a gran escala.
- **YARN (Yet Another Resource Negotiator):** Es el gestor de recursos del clúster. En las primeras versiones de Hadoop, MapReduce era responsable tanto del procesamiento como de la gestión de recursos, lo que lo hacía inflexible. YARN se introdujo para desacoplar estas dos funciones, actuando como el "cerebro" o sistema operativo del clúster. Es responsable de asignar recursos del sistema (CPU, memoria) a las diversas aplicaciones que se ejecutan en el clúster y de programar las tareas a través de los nodos.[33]

3.2. La Transición a Arquitecturas Modernas: La Era de la Velocidad

A pesar de su carácter revolucionario, el modelo MapReduce presentaba limitaciones significativas que impulsaron la búsqueda de alternativas más eficientes:

- **Rendimiento Lento:** La principal desventaja de MapReduce es su dependencia intensiva del disco. Después de cada fase (Map y Reduce), los resultados intermedios se escriben en HDFS. Este constante I/O (entrada/salida) de disco genera una alta latencia y ralentiza significativamente el procesamiento, especialmente en comparación con las tecnologías en memoria.[38]
- **Limitado al Procesamiento por Lotes:** MapReduce fue diseñado exclusivamente para el procesamiento batch. No es adecuado para el procesamiento de datos en tiempo real o streaming, una necesidad cada vez más crítica para los negocios modernos.[40]
- **Ineficiente para Algoritmos Iterativos:** Los algoritmos de machine learning a menudo requieren múltiples pasadas (iteraciones) sobre el mismo conjunto de datos. En MapReduce, cada iteración implica leer los datos desde el disco, procesarlos y escribir los resultados de nuevo en el disco, lo que lo hace extremadamente ineficiente para estas cargas de trabajo.[38]
- **Complejidad de Programación:** Escribir un trabajo MapReduce en Java requiere una cantidad considerable de código repetitivo (boilerplate), lo que aumenta la complejidad del desarrollo y el tiempo de entrega.[41]

Para superar estas limitaciones, surgió **Apache Spark**. Spark es un motor de procesamiento distribuido de propósito general que se convirtió en el sucesor de facto de MapReduce.[33] Su

innovación clave es el
procesamiento en memoria (in-memory processing). Spark carga los datos en la memoria RAM del clúster y los mantiene allí a través de múltiples etapas de procesamiento, evitando el costoso I/O de disco. Esto le permite ser hasta 100 veces más rápido que MapReduce para ciertas cargas de trabajo, especialmente las iterativas como el machine learning.[38] Además, Spark proporciona un marco unificado que soporta procesamiento batch, streaming, consultas SQL (Spark SQL) y machine learning (MLlib), simplificando enormemente el stack tecnológico.[42]

3.3. Almacenamiento Moderno: Del Almacén al Lago y la Casa del Lago

Paralelamente a la evolución del procesamiento, las estrategias de almacenamiento de datos analíticos también han experimentado una transformación significativa, dando lugar a tres paradigmas principales:

- **Data Warehouse (Almacén de Datos):** Es un sistema de almacenamiento centralizado optimizado para el análisis de datos estructurados y limpios. Utiliza un enfoque de **esquema en la escritura (schema-on-write)**, lo que significa que los datos deben ser limpiados, transformados y conformados a un esquema predefinido (proceso ETL) *antes* de ser cargados. Esto garantiza una alta calidad y consistencia de los datos, haciéndolo ideal para la inteligencia de negocio (BI), la generación de informes y el análisis histórico por parte de los analistas de negocio.[31]
- **Data Lake (Lago de Datos):** Es un repositorio masivo y de bajo coste que almacena datos en su formato nativo y crudo, sin importar si son estructurados, semiestructurados o no estructurados. Utiliza un enfoque de **esquema en la lectura (schema-on-read)**, donde la estructura se aplica a los datos solo cuando se leen para un análisis específico. Esto proporciona una enorme flexibilidad y velocidad para la ingesta de datos. Los Data Lakes son la base para la ciencia de datos exploratoria, el machine learning y el análisis de Big Data, y son utilizados principalmente por científicos e ingenieros de datos.[32]
- **Data Lakehouse (Casa del Lago):** Es una arquitectura moderna e híbrida que busca combinar lo mejor de ambos mundos: la flexibilidad, la escalabilidad y el bajo coste del Data Lake con las capacidades de gestión de datos, rendimiento y transacciones ACID (Atomicidad, Consistencia, Aislamiento, Durabilidad) del Data Warehouse. Un Lakehouse se construye sobre un Data Lake (generalmente en almacenamiento de objetos en la nube) y añade una capa de metadatos y un motor de consultas de alto rendimiento. Su objetivo es servir como una única plataforma para soportar cargas de trabajo de BI, ciencia de datos y machine learning, eliminando la necesidad de mantener sistemas separados y duplicar datos.[43]

Esta evolución, tanto en el procesamiento (de MapReduce a Spark) como en el

almacenamiento (de Warehouse a Lakehouse), está impulsada por un objetivo de negocio fundamental: reducir el **"tiempo hasta el conocimiento" (time to insight)**. MapReduce y los Data Warehouses, con sus procesos lentos y rígidos, podían tardar días o semanas en convertir un dato crudo en un conocimiento accionable. Spark aceleró drásticamente el procesamiento, y los Data Lakes aceleraron la ingesta. El Data Lakehouse representa el último paso en este camino, buscando crear la ruta más corta y eficiente posible desde el dato crudo hasta el conocimiento fiable y accionable, respondiendo así a la presión competitiva que exige a las empresas tomar mejores decisiones, más rápido.

Característica	Data Warehouse	Data Lake	Data Lakehouse
Estructura de Datos	Principalmente estructurados y procesados.	Todos los tipos: estructurados, semiestructurados y no estructurados (datos crudos).	Todos los tipos, con capacidad para estructurar datos crudos.
Esquema	Schema-on-write (Esquema en la escritura): definido antes de la carga de datos.	Schema-on-read (Esquema en la lectura): aplicado durante el análisis.	Híbrido: soporta schema-on-read pero permite la aplicación y evolución de esquemas.
Modelo de Procesamiento	ETL (Extract, Transform, Load).	ELT (Extract, Load, Transform).	ELT, con capacidades de ETL y streaming.
Usuarios Principales	Analistas de negocio, ejecutivos.	Científicos de datos, ingenieros de datos.	Todos: analistas, científicos e ingenieros de datos.
Casos de Uso Típicos	Business Intelligence (BI), reporting corporativo, análisis histórico.	Ciencia de datos exploratoria, machine learning, procesamiento de datos no estructurados.	Plataforma unificada para BI, IA y machine learning.
Coste	Generalmente más alto debido al hardware especializado y al acoplamiento de cómputo y almacenamiento.	Más bajo, utiliza almacenamiento de objetos de bajo coste y separa cómputo de almacenamiento.	Bajo coste de almacenamiento (como un Data Lake) con rendimiento optimizado.
Soporte de Transacciones ACID	Sí, es una característica fundamental.	No, generalmente no soporta transacciones.	Sí, a través de formatos de tabla abiertos como Delta Lake, Apache Iceberg

			o Apache Hudi.

Sección 4: Los Protagonistas del Ecosistema de Datos

El ecosistema de Big Data no solo está compuesto por tecnologías y arquitecturas, sino también por los profesionales que las diseñan, las operan y extraen valor de ellas. Aunque las responsabilidades pueden solaparse, existen tres roles principales y bien definidos que colaboran para llevar los datos desde su estado crudo hasta la decisión de negocio.

4.1. El Arquitecto de los Datos: El Ingeniero de Datos (Data Engineer)

Los ingenieros de datos son los constructores y mantenedores de la infraestructura de datos. Su principal objetivo es desarrollar, probar y mantener arquitecturas robustas y escalables, como bases de datos, pipelines de datos y sistemas de procesamiento a gran escala. Son responsables de garantizar que los datos sean fiables, de alta calidad y estén disponibles de manera eficiente para su consumo por parte de otros roles.[8]

- **Responsabilidades:**
  - Diseñar y construir pipelines de datos para la ingesta (ETL/ELT).
  - Gestionar y optimizar bases de datos y sistemas de almacenamiento (Data Lakes, Warehouses).
  - Asegurar la calidad y la integridad de los datos a través de procesos de limpieza y validación.
  - Automatizar y monitorizar los flujos de trabajo de datos.
- **Habilidades y Herramientas Clave:**
  - Programación avanzada (Python, Java, Scala).
  - Dominio de SQL y modelado de datos.
  - Experiencia con herramientas de Big Data como Apache Spark, Apache Kafka y Apache Airflow.
  - Conocimiento profundo de plataformas en la nube (AWS, Azure, GCP) y sus servicios de datos.[8]

4.2. El Explorador y Traductor: El Analista de Datos (Data Analyst)

Los analistas de datos son los intérpretes que convierten los datos en información comprensible para el negocio. Se centran en el análisis de datos históricos para identificar tendencias, patrones y responder a preguntas de negocio específicas. Su gran fortaleza

reside en su capacidad para comunicar hallazgos complejos a audiencias no técnicas a través de informes, dashboards y visualizaciones, actuando como un puente entre los datos y la toma de decisiones.[10]

- **Responsabilidades:**
  - Recopilar, limpiar y analizar conjuntos de datos para extraer información relevante.
  - Crear visualizaciones de datos y cuadros de mando interactivos.
  - Generar informes periódicos y ad-hoc para diferentes departamentos.
  - Traducir los requisitos del negocio en preguntas analíticas.
- **Habilidades y Herramientas Clave:**
  - SQL avanzado para la consulta de datos.
  - Dominio de hojas de cálculo (Excel).
  - Experiencia con herramientas de BI como Tableau, Power BI o FineBI.
  - Conocimientos de estadística descriptiva.
  - Excelentes habilidades de comunicación y visualización de datos.[10]

4.3. El Visionario Predictivo: El Científico de Datos (Data Scientist)

Los científicos de datos son los visionarios que utilizan los datos para predecir el futuro. Aplican métodos estadísticos avanzados, algoritmos de machine learning y técnicas de modelado predictivo para descubrir conocimientos ocultos y construir sistemas que puedan tomar decisiones autónomas. A diferencia de los analistas que se centran en el "qué pasó", los científicos de datos se enfocan en "qué pasará" y "qué deberíamos hacer al respecto".[11]

- **Responsabilidades:**
  - Formular preguntas de negocio complejas que pueden ser respondidas con datos.
  - Realizar análisis exploratorio de datos para identificar patrones y oportunidades.
  - Construir, entrenar y validar modelos de machine learning.
  - Desplegar modelos en producción para crear "productos de datos" (ej. motores de recomendación, sistemas de detección de fraude).
- **Habilidades y Herramientas Clave:**
  - Sólidos fundamentos en estadística, matemáticas y probabilidad.
  - Programación en Python o R.
  - Experiencia con librerías de machine learning como Scikit-learn, TensorFlow o PyTorch.
  - Conocimiento de técnicas de modelado predictivo y prescriptivo.[46]

4.4. La Sinergia en Acción: Cómo Colaboran los Roles

La colaboración entre estos tres roles no es un proceso lineal, sino un ciclo de retroalimentación continuo y dinámico.[9] Un proyecto típico de Big Data ilustra esta sinergia:

- **Ejemplo del Mundo Real Detallado: Proyecto de Detección de Fraude en Tiempo Real**
    1. **Inicio del Proyecto:** El negocio plantea la necesidad de reducir las pérdidas por fraude en transacciones con tarjeta de crédito.
    2. **Ingeniero de Datos:** Construye un pipeline de datos robusto utilizando Apache Kafka para ingerir millones de eventos de transacción en tiempo real. Utiliza Apache Spark Streaming para procesar y enriquecer estos datos (por ejemplo, añadiendo el historial del cliente) y los almacena en un Data Lakehouse para análisis tanto en tiempo real como histórico.[48]
    3. **Analista de Datos:** Utiliza SQL para consultar los datos históricos de transacciones almacenados por el ingeniero. Identifica patrones en casos de fraude conocidos, como "las transacciones fraudulentas a menudo ocurren en nuevos dispositivos y superan el gasto medio del cliente". Crea un dashboard en Tableau para que el equipo de fraude pueda monitorizar estas tendencias y métricas clave.[47]
    4. **Científico de Datos:** Toma los patrones descubiertos por el analista como punto de partida para la ingeniería de características (feature engineering). Utiliza el gran conjunto de datos históricos, preparado y mantenido por el ingeniero, para entrenar un modelo de machine learning (por ejemplo, un Gradient Boosting) que calcula una puntuación de probabilidad de fraude para cada nueva transacción.[9]
    5. **Ciclo de Retroalimentación:** El ingeniero de datos colabora con el científico de datos para "produccionizar" el modelo, es decir, integrarlo en el pipeline de streaming de Spark. Ahora, cada transacción que llega a través de Kafka es evaluada por el modelo en tiempo real. Si la puntuación de fraude supera un umbral, se genera una alerta o se bloquea la transacción automáticamente. El analista de datos monitoriza el rendimiento del modelo en producción a través de su dashboard, y los resultados (falsos positivos, fraudes no detectados) se utilizan para reentrenar y mejorar el modelo, comenzando el ciclo de nuevo.[49]

Dimensión	Ingeniero de Datos	Analista de Datos	Científico de Datos
Objetivo Principal	Construir y mantener la infraestructura de datos para que sea fiable, escalable y eficiente.	Interpretar datos históricos para responder preguntas de negocio y comunicar insights.	Utilizar datos para predecir eventos futuros y construir modelos de decisión inteligentes.
Habilidades Clave	Programación (Python, Java, Scala), SQL, ETL/ELT, Modelado de	SQL, Estadística Descriptiva, Visualización de Datos,	Estadística, Machine Learning, Programación (Python,

	Datos, Arquitectura de Sistemas, Cloud Computing.	Comunicación, Conocimiento del Negocio.	R), Modelado Predictivo, Experimentación.
Herramientas Comunes	Apache Spark, Kafka, Airflow, dbt, Snowflake, Bases de Datos NoSQL, AWS, Azure, GCP.	Tableau, Power BI, Excel, SQL, Google Analytics, Python (Pandas, Matplotlib).	Python (Scikit-learn, TensorFlow, PyTorch), R, Jupyter Notebooks, Spark MLlib.
Pregunta que Responde	"¿Cómo podemos mover y almacenar estos datos de forma eficiente y fiable?"	"¿Qué ha pasado y por qué?"	"¿Qué pasará a continuación y qué podemos hacer al respecto?"

Sección 5: Mapa Conceptual del Stack Tecnológico Moderno

Para que los estudiantes puedan navegar por el complejo panorama de herramientas de Big Data, es útil presentar un mapa conceptual que organice las tecnologías clave según su función dentro del ciclo de vida de los datos. Este mapa sirve como una hoja de ruta para futuras unidades técnicas más profundas.

Una Visión General y Gráfica

El stack tecnológico moderno de Big Data se puede visualizar como una serie de capas funcionales, cada una con herramientas especializadas:

- **Ingesta de Datos (Streaming & Batch):**
  - **Apache Kafka:** Es el estándar de la industria para construir pipelines de datos en tiempo real. Funciona como un sistema de mensajería distribuido de publicación-suscripción, organizando los flujos de eventos en "topics" que pueden ser consumidos por múltiples aplicaciones de forma fiable y escalable.[50]
  - **Apache Flume:** Una herramienta diseñada específicamente para recolectar, agregar y mover grandes volúmenes de datos de logs desde diversas fuentes a un almacenamiento centralizado como HDFS.[36]
- **Almacenamiento:**
  - **Sistemas de Archivos Distribuidos y Almacenamiento de Objetos:** La base del almacenamiento de Big Data. Incluye el tradicional HDFS y, más comúnmente hoy en día, los servicios de almacenamiento de objetos en la nube como Amazon S3, Azure Blob Storage y Google Cloud Storage, que ofrecen escalabilidad casi

infinita y bajo coste.[34]
    - **Plataformas de Data Lakehouse:** Sobre el almacenamiento de objetos, formatos de tabla abiertos como **Delta Lake**, **Apache Iceberg** y **Apache Hudi** añaden capacidades transaccionales (ACID), gestión de esquemas y viajes en el tiempo a los Data Lakes, convirtiéndolos en Lakehouses.[43]
- **Procesamiento:**
    - **Apache Spark:** El motor de procesamiento de datos a gran escala por excelencia. Su capacidad para realizar procesamiento en memoria lo hace ideal para ETL, análisis interactivo, streaming y, especialmente, machine learning.[42] **PySpark** es su popular API para Python.[54]
    - **Apache Flink:** Otro potente motor de procesamiento de streams, conocido por su baja latencia y su manejo avanzado del tiempo de evento, lo que lo hace ideal para aplicaciones de streaming muy exigentes.[40]
- **Orquestación de Workflows:**
    - **Apache Airflow:** Es la herramienta líder para programar, orquestar y monitorizar flujos de trabajo de datos complejos. Los flujos de trabajo se definen como código (Python) en forma de Grafos Acíclicos Dirigidos (DAGs), lo que permite crear pipelines dinámicos, versionables y mantenibles.[56]
- **Bases de Datos Analíticas / Data Warehouses en la Nube:**
    - **Snowflake, Google BigQuery, Amazon Redshift:** Son plataformas de Data Warehouse nativas de la nube que separan el cómputo del almacenamiento, ofreciendo una escalabilidad y un rendimiento masivos para consultas SQL analíticas.[43]
- **Consumo y Visualización:**
    - **Herramientas de BI:** Tableau, Power BI, Looker, etc., que se conectan a los sistemas de almacenamiento para crear visualizaciones y dashboards.[24]
    - **APIs:** Para servir datos o predicciones de modelos a aplicaciones externas. Frameworks modernos como **FastAPI** en Python permiten construir APIs de alto rendimiento de manera rápida y sencilla, con validación de datos y documentación automática.[59]

El panorama tecnológico actual se define por dos tendencias clave: la **"Gran Desagregación" (Great Unbundling)** y el dominio de los **servicios gestionados en la nube**. Los primeros ecosistemas de Hadoop eran monolíticos, empaquetando almacenamiento, procesamiento y gestión en una única distribución. La nube rompió este modelo al ofrecer un almacenamiento de objetos (como S3) superior, más barato y más escalable que HDFS. Esto permitió "desagregar" o desacoplar el almacenamiento del cómputo.

Este desacoplamiento fomentó la aparición de empresas especializadas que ofrecen las mejores soluciones para cada capa del stack: Snowflake para el warehousing, Databricks para el procesamiento con Spark, Confluent para Kafka, etc. Si bien esta desagregación ofrece una flexibilidad sin precedentes para elegir la mejor herramienta para cada tarea,

también crea un enorme desafío de integración.
Por esta razón, la **orquestación** (con herramientas como Airflow) y la **interoperabilidad** (a través de formatos de datos abiertos como Parquet y formatos de tabla como Delta Lake o Iceberg) se han convertido en los componentes más críticos del stack moderno. Son el "pegamento" que une este ecosistema desagregado. Para un estudiante que ingresa a este campo, esto significa que ya no es suficiente aprender una única plataforma monolítica; es esencial comprender cómo integrar múltiples servicios especializados para construir una solución de datos completa y eficaz.

Obras citadas

1. BOE-A-2021-7686.pdf
2. Characteristics of Big Data: Types & the 5 V's Explained - Pickl.AI, fecha de acceso: septiembre 25, 2025, https://www.pickl.ai/blog/characteristics-of-big-data/
3. Big Data Defined: Examples and Benefits | Google Cloud, fecha de acceso: septiembre 25, 2025, https://cloud.google.com/learn/what-is-big-data
4. What Is the Data, Information, Knowledge, Wisdom (DIKW) Pyramid? - Ontotext, fecha de acceso: septiembre 25, 2025, https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/
5. From Knowledge to Wisdom: Looking beyond the Knowledge Hierarchy - MDPI, fecha de acceso: septiembre 25, 2025, https://www.mdpi.com/2673-9585/3/2/14
6. Transforming clinical data into wisdom - PMC - PubMed Central, fecha de acceso: septiembre 25, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC8018525/
7. DIKW Pyramid Explained: Turning Data into Insight & Wisdom - Symbio6, fecha de acceso: septiembre 25, 2025, https://symbio6.nl/en/blog/dikw-pyramid-model
8. Data Engineer vs Data Analyst: Key differences and opportunities - Synapxe, fecha de acceso: septiembre 25, 2025, https://www.synapxe.sg/blog/career-stories/data-engineer-vs-data-analyst
9. DataOps: Bridging the Gap Between Data Engineering and Data Science - IABAC, fecha de acceso: septiembre 25, 2025, https://iabac.org/blog/dataops-bridging-the-gap-between-data-engineering-and-data-science
10. Difference between Data Analyst, Data Engineer and Data Scientist? Which among these is more difficult to become and which is a more interesting role? : r/dataanalysis - Reddit, fecha de acceso: septiembre 25, 2025, https://www.reddit.com/r/dataanalysis/comments/1dvbwu3/difference_between_data_analyst_data_engineer_and/
11. In Layman's terms, what's the difference in role between Data Analyst, Data Engineer and Data Scientist [Serious Answers Only] : r/dataanalysis - Reddit, fecha de acceso: septiembre 25, 2025, https://www.reddit.com/r/dataanalysis/comments/130fxpc/in_laymans_terms_whats_the_difference_in_role/
12. The Data-Information-Knowledge-Wisdom Pyramid - DataCamp, fecha de

acceso: septiembre 25, 2025, https://www.datacamp.com/cheat-sheet/the-data-information-knowledge-wisdom-pyramid
13. What is the Data, Information, Knowledge, Wisdom (DIKW) Model? - Weje.io, fecha de acceso: septiembre 25, 2025, https://weje.io/blog/data-information-knowledge-wisdom
14. What are the 5 V's of Big Data? - Teradata, fecha de acceso: septiembre 25, 2025, https://www.teradata.com/glossary/what-are-the-5-v-s-of-big-data
15. Explain 5 Characteristics of Big Data – 5V of Big Data With Example E - Upskill Campus, fecha de acceso: septiembre 25, 2025, https://www.upskillcampus.com/blog/characteristics-of-big-data/
16. What are the 5 V's of Big Data? | Definition & Explanation - TechnologyAdvice, fecha de acceso: septiembre 25, 2025, https://technologyadvice.com/blog/information-technology/the-four-vs-of-big-data/
17. 5 Vs of Big Data: Meaning, Examples, Applications & Importance - Plutus Education, fecha de acceso: septiembre 25, 2025, https://plutuseducation.com/blog/5-vs-of-big-data/
18. The 5 Vs in Big data: what they are and how to apply them - SMOWL, fecha de acceso: septiembre 25, 2025, https://smowl.net/en/blog/big-data-5v/
19. Role of Cloud Computing in Big Data Analytics - GeeksforGeeks, fecha de acceso: septiembre 25, 2025, https://www.geeksforgeeks.org/blogs/role-of-cloud-computing-in-big-data-analytics/
20. (PDF) Big Data Characteristics, Challenges, Architectures, Analytics and Applications: A Review - ResearchGate, fecha de acceso: septiembre 25, 2025, https://www.researchgate.net/publication/317133629_Big_Data_Characteristics_Challenges_Architectures_Analytics_and_Applications_A_Review
21. 6V's of Big Data - GeeksforGeeks, fecha de acceso: septiembre 25, 2025, https://www.geeksforgeeks.org/data-science/5-vs-of-big-data/
22. Netflix Content Recommendation System – Product Analytics Case Study - HelloPM, fecha de acceso: septiembre 25, 2025, https://hellopm.co/netflix-content-recommendation-system-product-analytics-case-study/
23. Netflix Recommender System : Big Data Case Study | PPTX | Search ..., fecha de acceso: septiembre 25, 2025, https://www.slideshare.net/slideshow/netflix-recommender-system-big-data-case-study/248272316
24. Data Lifecycle: Definition and Best Practices - Acceldata, fecha de acceso: septiembre 25, 2025, https://www.acceldata.io/blog/data-lifecycle
25. 8 Data Life Cycle Phases Explained | Airbyte, fecha de acceso: septiembre 25, 2025, https://airbyte.com/data-engineering-resources/data-life-cycle
26. The Data Engineering lifecycle for beginners | by Chamuditha Kekulawala | Medium, fecha de acceso: septiembre 25, 2025, https://medium.com/@ckekula/the-data-engineering-lifecycle-for-beginners-057b12f2e2a8
27. Data Engineering Lifecycle - by Jayant Nehra - Medium, fecha de acceso:

septiembre 25, 2025, https://medium.com/towards-data-engineering/data-engineering-lifecycle-d1e7ee81632e
28. Ensure Good Data Ethics and Governance - Data to Policy Navigator, fecha de acceso: septiembre 25, 2025, https://www.datatopolicy.org/considerations/ensure-good-data-ethics-and-governance
29. What is the relationship between data ethics and data governance? - Milvus, fecha de acceso: septiembre 25, 2025, https://milvus.io/ai-quick-reference/what-is-the-relationship-between-data-ethics-and-data-governance
30. Data Governance and Ethics in the Age of Big Data - Scalar Solutions, fecha de acceso: septiembre 25, 2025, https://www.scalar-solutions.com/blogs/data-governance-and-ethics
31. Data Warehouse vs Data Lake vs Data Lakehouse: Comparison - Atlan, fecha de acceso: septiembre 25, 2025, https://atlan.com/data-warehouse-vs-data-lake-vs-data-lakehouse/
32. Data Lake vs Data Warehouse: 6 Key Differences - Qlik, fecha de acceso: septiembre 25, 2025, https://www.qlik.com/us/data-lake/data-lake-vs-data-warehouse
33. Hadoop Ecosystem - GeeksforGeeks, fecha de acceso: septiembre 25, 2025, https://www.geeksforgeeks.org/dbms/hadoop-ecosystem/
34. What is Hadoop? - Apache Hadoop Explained - AWS, fecha de acceso: septiembre 25, 2025, https://aws.amazon.com/what-is/hadoop/
35. Hadoop Architecture Explained-The What, How and Why - ProjectPro, fecha de acceso: septiembre 25, 2025, https://www.projectpro.io/article/hadoop-architecture-explained-what-it-is-and-why-it-matters/317
36. Hadoop Ecosystem | Hadoop Tools for Crunching Big Data | Edureka, fecha de acceso: septiembre 25, 2025, https://www.edureka.co/blog/hadoop-ecosystem
37. Apache Hadoop Architecture - HDFS, YARN & MapReduce - TechVidvan, fecha de acceso: septiembre 25, 2025, https://techvidvan.com/tutorials/hadoop-architecture/
38. Spark vs. MapReduce: What's the Difference? - Coursera, fecha de acceso: septiembre 25, 2025, https://www.coursera.org/articles/spark-vs-mapreduce
39. Hadoop vs Spark - Difference Between Apache Frameworks - AWS, fecha de acceso: septiembre 25, 2025, https://aws.amazon.com/compare/the-difference-between-hadoop-vs-spark/
40. 13 Big Limitations of Hadoop & Solution To Hadoop Drawbacks - DataFlair, fecha de acceso: septiembre 25, 2025, https://data-flair.training/blogs/13-limitations-of-hadoop/
41. Spark vs Hadoop MapReduce: 5 Key Differences | Integrate.io, fecha de acceso: septiembre 25, 2025, https://www.integrate.io/blog/apache-spark-vs-hadoop-mapreduce/
42. What is PySpark & Why Use It? - YouTube, fecha de acceso: septiembre 25, 2025, https://www.youtube.com/watch?v=VEzjGwOb6J0

43. Data Warehouses vs. Data Lakes vs. Data Lakehouses | IBM, fecha de acceso: septiembre 25, 2025, https://www.ibm.com/think/topics/data-warehouse-vs-data-lake-vs-data-lakehouse
44. Data Lake vs. Data Warehouse vs. Data Lakehouse: Understanding the Differences, fecha de acceso: septiembre 25, 2025, https://amplitude.com/blog/data-lake-vs-warehouse-vs-lakehouse
45. How do you decide between a database, data lake, data warehouse, or lakehouse? - Reddit, fecha de acceso: septiembre 25, 2025, https://www.reddit.com/r/dataengineering/comments/1mb3280/how_do_you_decide_between_a_database_data_lake/
46. Data Scientist vs Data Engineer | What's the Difference? | DataCamp, fecha de acceso: septiembre 25, 2025, https://www.datacamp.com/blog/data-scientist-vs-data-engineer
47. How Data Analysts Collaborate With Data Engineers And Scientists In Real Projects | ECA, fecha de acceso: septiembre 25, 2025, https://employabilityadvantage.com/how-data-analysts-collaborate-with-data-engineers-and-scientists-in-real-projects/
48. How do Data Engineers collaborate with Data Scientists and Analysts? - Lemon.io, fecha de acceso: septiembre 25, 2025, https://lemon.io/answers/data/how-do-data-engineers-collaborate-with-data-scientists-and-analysts/
49. Collaboration between data engineers, data analysts and data scientists | by Germain Tanguy | Dailymotion | Medium, fecha de acceso: septiembre 25, 2025, https://medium.com/dailymotion/collaboration-between-data-engineers-data-analysts-and-data-scientists-97c00ab1211f
50. What is Kafka? - Apache Kafka Explained - AWS - Updated 2025, fecha de acceso: septiembre 25, 2025, https://aws.amazon.com/what-is/apache-kafka/
51. Apache Kafka for Beginners: A Comprehensive Guide - DataCamp, fecha de acceso: septiembre 25, 2025, https://www.datacamp.com/tutorial/apache-kafka-for-beginners-a-comprehensive-guide
52. Introduction - Apache Kafka, fecha de acceso: septiembre 25, 2025, https://kafka.apache.org/intro
53. Kafka For Beginners. What is Kafka? | by Rinu Gour - Medium, fecha de acceso: septiembre 25, 2025, https://medium.com/@rinu.gour123/kafka-for-beginners-74ec101bc82d
54. www.coursera.org, fecha de acceso: septiembre 25, 2025, https://www.coursera.org/articles/what-is-pyspark#:~:text=PySpark%20is%20an%20open%2Dsource,data%20sets%20of%20all%20sizes.
55. What is Apache Kafka? - GeeksforGeeks, fecha de acceso: septiembre 25, 2025, https://www.geeksforgeeks.org/apache-kafka/apache-kafka/
56. www.qubole.com, fecha de acceso: septiembre 25, 2025, https://www.qubole.com/the-ultimate-guide-to-apache-

airflow#:~:text=Apache%20Airflow%20is%20an%20open,trigger%20tasks%2C%20and%20success%20status.
57. An introduction to Apache Airflow® | Astronomer Docs, fecha de acceso: septiembre 25, 2025, https://www.astronomer.io/docs/learn/intro-to-airflow
58. What is Apache Airflow? | Qubole, fecha de acceso: septiembre 25, 2025, https://www.qubole.com/the-ultimate-guide-to-apache-airflow
59. en.wikipedia.org, fecha de acceso: septiembre 25, 2025, https://en.wikipedia.org/wiki/FastAPI
60. Get Started With FastAPI - Real Python, fecha de acceso: septiembre 25, 2025, https://realpython.com/get-started-with-fastapi/
61. FastAPI, fecha de acceso: septiembre 25, 2025, https://fastapi.tiangolo.com/