Procesamiento de Grandes Volumenes de Datos

Conferencia 1: La revolución de los datos

Deborah Famadas Rodríguez

Universidad de la Habana

September 1, 2025

Estructura del Curso

Componentes:

- 13 Conferencias teóricas
- 12 Clases prácticas / Seminarios
- 1 Proyecto final en equipo

Evaluación:

- Evaluación Parcial (Semana 7-8)
- Tareas y Participación
- Proyecto Final (Semana 16)

Metodología

Combinaremos una base conceptual sólida con una aplicación práctica intensiva para desarrollar habilidades.

Crecimiento de los datos

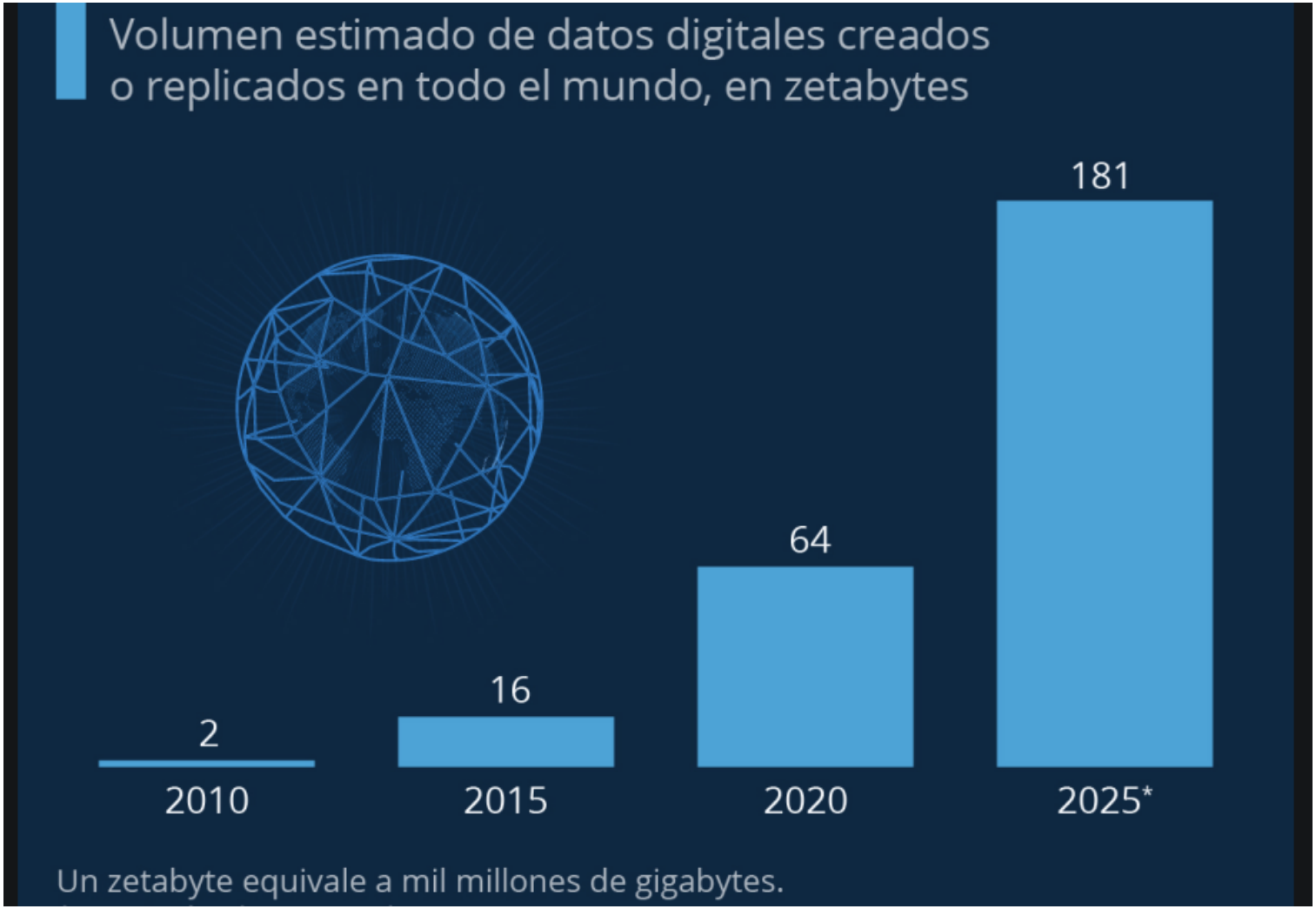


Figure: Crecimiento exponencial

El Desafío: Más de lo que se Ve a Simple Vista

- Estructurados
- Analizados
- Bases de datos, Hojas de cálculo

- No estructurados y semi-estructurado
- Crudos, sin procesar
- Texto, Imágenes, Logs, Redes Sociale

La Primera V: Volumen

La Escala del Problema

Hablamos de cantidades de datos que superan la capacidad de los sistemas tradicionales.

- Terabytes (TB), Petabytes (PB), Exabytes (EB).
- 1 Petabyte = 1,000 Terabytes.
- **Ejemplo:** Los datos generados por todos los vuelos comerciales del mundo en un día, o todas las transacciones de Amazon a nivel global.

La Segunda V: Velocidad

El Ritmo de Generación y Procesamiento

Los datos se generan a una velocidad vertiginosa y su valor a menudo depende de una respuesta inmediata.

- Procesamiento en tiempo real o casi real.
- **Ejemplo 1 (Finanzas):** Un sistema de detección de fraude debe analizar una transacción en milisegundos.
- **Ejemplo 2 (Media):** Un motor de recomendación de Netflix debe reaccionar a lo que acabas de ver.

La Tercera V: Variedad

La Complejidad de los Formatos

La mayoría de los datos del mundo no están ordenados.

Tipos de Datos: Estructurados

Definición:

Datos perfectamente organizados en tablas con filas y columnas definidas.

- Modelo de datos predefinido (esquema)
- Bases de datos relacionales (RDBMS)
- Fácil de consultar (SQL)

Esquema rígido y predefinido. Fáciles de gestionar y consultar.

Tipos de Datos: No Estructurados

Definición:

Datos sin un modelo predefinido. Constituyen ¿80 porciento de los datos mundiales.

- **Ejemplos:** Correos electrónicos, documentos PDF, imágenes, vídeos, audios, publicaciones en redes sociales. modelo de datos predefinido.
- Data Lakes / Bases de datos NoSQL.
- Difícil de analizar.
- Barato de ingerir.

No tienen estructura inherente, lo que los hace difíciles de procesar para los sistemas tradicionales.

Tipos de Datos: Semi-estructurados

Definición:

No residen en una base de datos relacional, pero contienen etiquetas o marcadores para separar elementos semánticos.

- **Ejemplos:** Archivos JSON, documentos XML.
- **Característica Clave:** Tienen una estructura jerárquica y flexible, un punto intermedio entre los otros dos tipos.

Las Otras V's: Veracidad y Valor

Veracidad

- ¿Son los datos fiables y precisos?
- La calidad de los datos es crucial.
- Un lago de datos lleno de "basura" lleva a decisiones erróneas.

Valor

- El objetivo final.
- Transformar datos en conocimiento, acciones y ventaja competitiva.
- Sin valor, los grandes volúmenes de datos son solo un coste de almacenamiento.

Análisis de los Sistemas RDBMS

Fundamentos del Modelo Relacional

El modelo relacional ha constituido la base de los sistemas de datos transaccionales, priorizando la consistencia y la integridad referencial.

- Diseñados para garantizar las propiedades **ACID** (Atomicidad, Consistencia, Aislamiento y Durabilidad), lo cual es crítico en sistemas transaccionales (OLTP).
- Su esquema rígido y la necesidad de normalización, si bien garantizan la integridad, introducen una falta de flexibilidad para datos semiestructurados o no estructurados.
- Presentan una escalabilidad vertical como principal opción, la cual resulta costosa e insuficiente para las cargas de trabajo masivas, que demandan una escalabilidad horizontal distribuida. [1]

Limitación 1: Escalabilidad Vertical

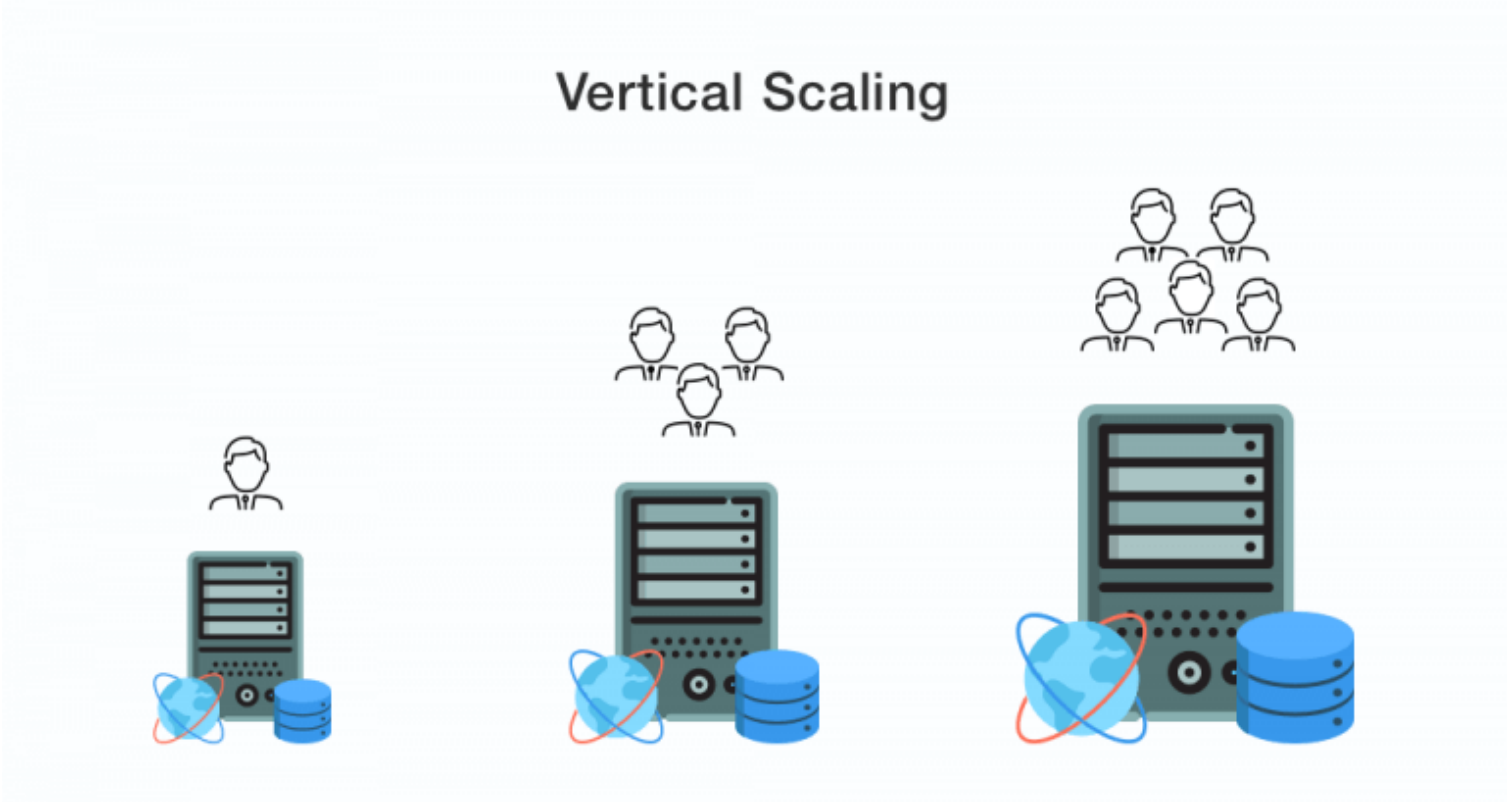


Figure: Scale-Up: Un único servidor, cada vez más grande y caro.

Es costoso, tiene límites físicos y no es tolerante a fallos. Si el servidor cae, todo el sistema cae.

La Solución: Escalabilidad Horizontal

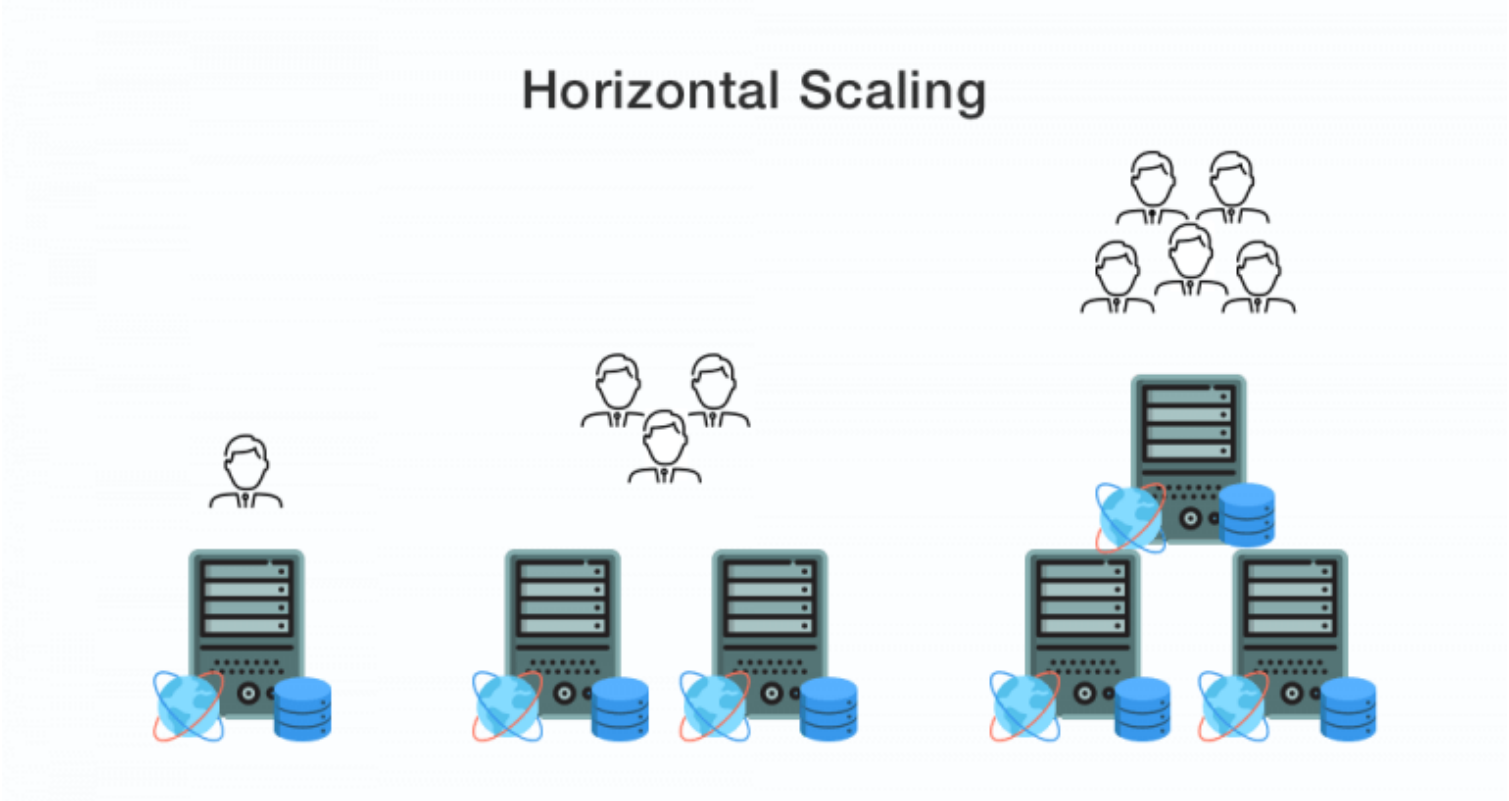


Figure: Scale-Out: Múltiples servidores estándar trabajando en clúster.

Es económico, casi infinitamente escalable y tolerante a fallos por diseño.
Si un nodo cae, el sistema sigue funcionando.

Limitación 2: El Esquema Rígido

Schema-on-Write (Esquema en la Escritura)

Los RDBMS te obligan a diseñar una estructura (esquema) perfecta y rígida **antes** de poder almacenar cualquier dato.

- Debes definir tablas, columnas y tipos de datos de antemano.
- Es inflexible. ¿Qué pasa si recibes un nuevo tipo de dato no previsto? No puedes almacenarlo sin un costoso rediseño.

La Solución: Flexibilidad

Schema-on-Read (Esquema en la Lectura)

Los sistemas como Hadoop popularizaron el enfoque opuesto: almacena los datos tal como vienen (brutos) y preocúpate de darles una estructura solo en el momento en que los necesites para un análisis.

- Máxima flexibilidad para almacenar cualquier tipo de dato.
- Permite el análisis exploratorio y el descubrimiento de patrones no anticipados.
- Es la filosofía detrás del concepto de **Data Lake**.

Limitación 3: Ineficiencia con Datos No Estructurados

- Los RDBMS no están diseñados para almacenar, y mucho menos para analizar, grandes volúmenes de texto, imágenes o logs.
- Sus motores de consulta están optimizados para búsquedas en índices sobre datos estructurados.
- Intentar usarlos resulta en un rendimiento pésimo y en la incapacidad de extraer valor real.

La Revolución de Hadoop

Un Nuevo Ecosistema

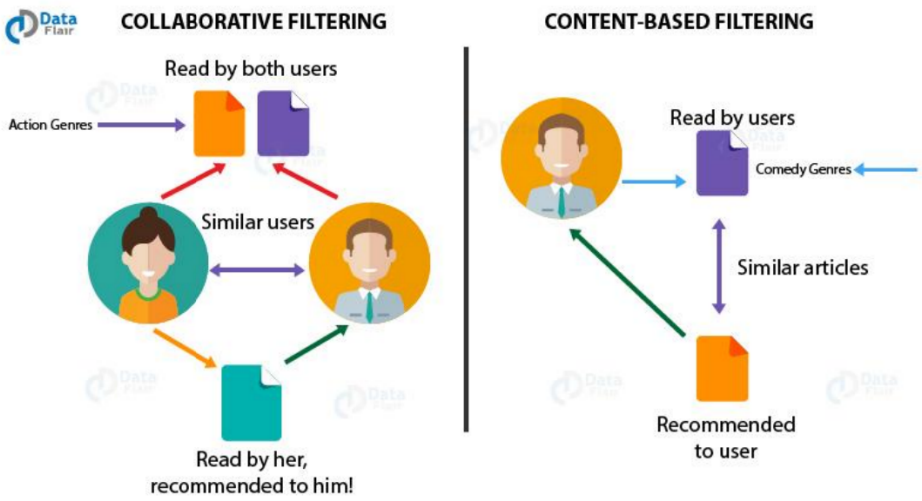
Apache Hadoop no es una herramienta, es un framework que cambió las reglas del juego.

- Basado en **hardware de bajo coste** (commodity hardware).
- Diseñado para la **escalabilidad horizontal**.
- **Tolerante a fallos** por diseño.
- Capaz de procesar **cualquier tipo de dato**.
- Introdujo el paradigma de procesamiento **MapReduce**. [4]

Uso de grandes volúmenes de datos

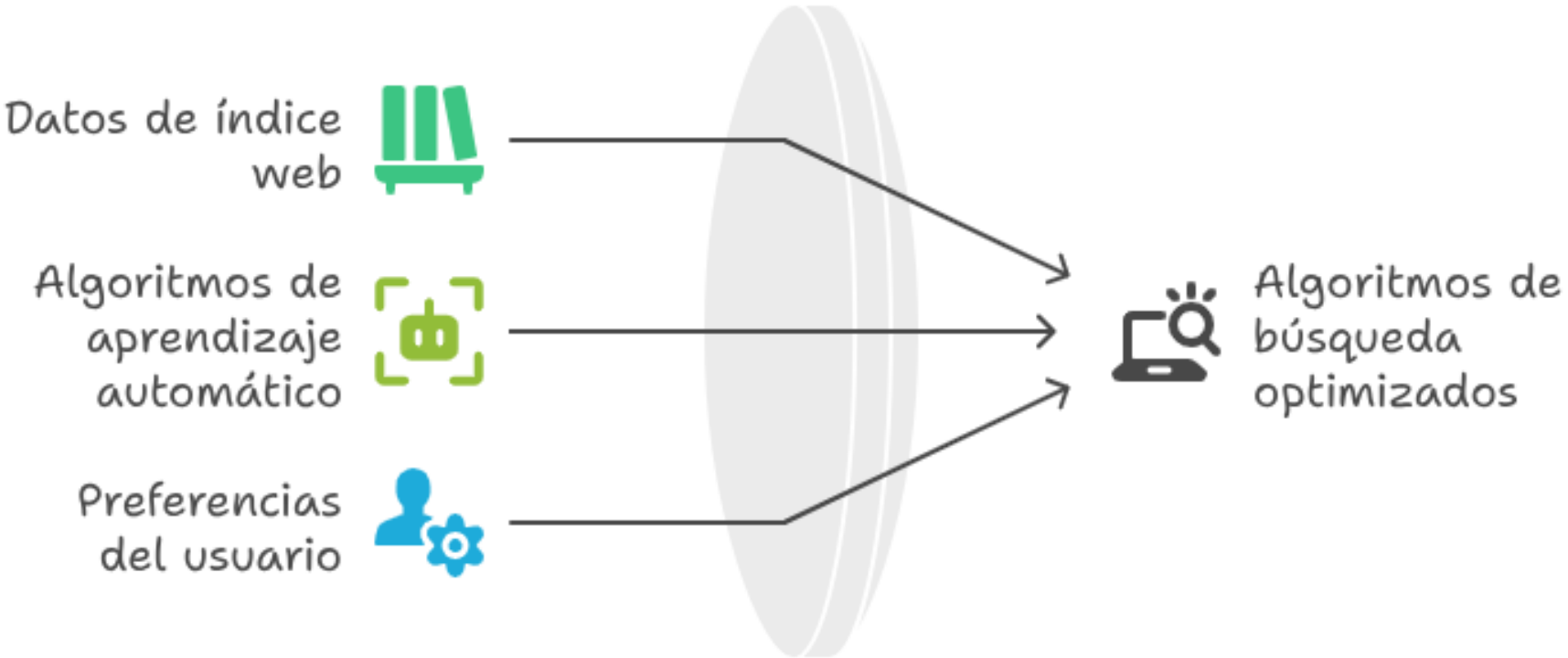
El análisis de grandes volúmenes de datos puede ayudar a las empresas a identificar nuevas oportunidades y los movimientos estratégicos adecuados que deben realizar.

Caso de Netflix



Netflix aplica modelos de análisis de datos para descubrir el comportamiento y los patrones de compra de sus clientes. A partir de esta información, recomienda películas y programas de televisión a sus clientes. Es decir, analiza la elección del cliente y sus preferencias y sugiere programas y películas en consecuencia.

Casos de Google



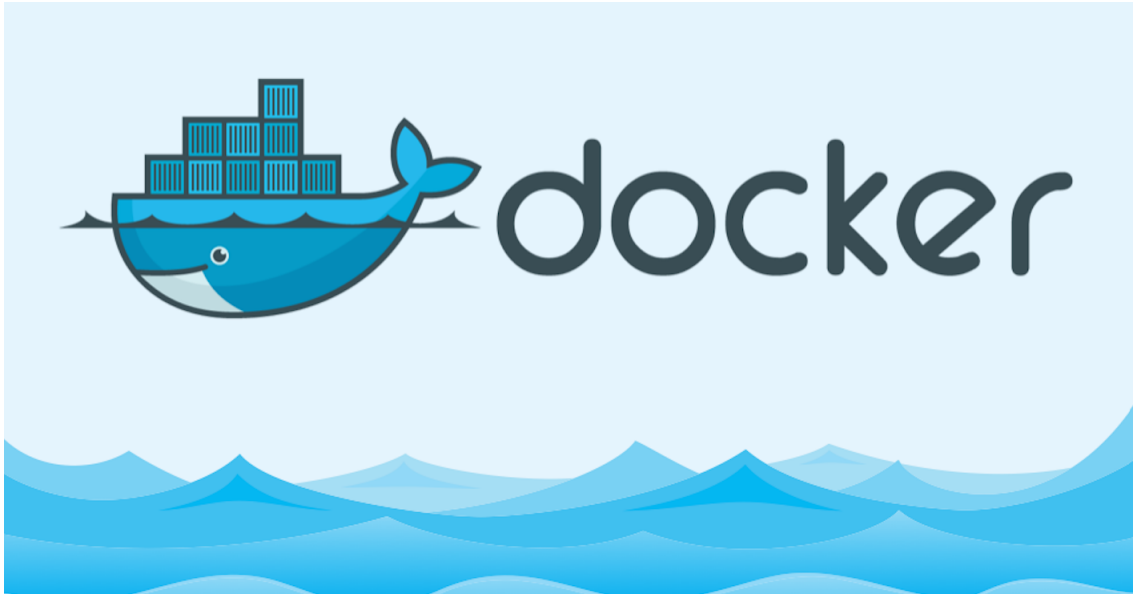
Nuestro Viaje por el Ecosistema

- **Semanas 1-4: Fundamentos.** Entenderemos el ecosistema Hadoop, su sistema de archivos (HDFS) y su modelo de procesamiento original (MapReduce).
- **Semanas 5-9: Arquitectura e Integración.** Diseñaremos Data Warehouses y Data Lakes, y moveremos datos con ETL y ELT.
- **Semanas 10-13: Procesamiento Avanzado.** Dominaremos Apache Spark, el estándar de la industria actual.
- **Semana 14: El Futuro.** Exploraremos la supercomputación (HPC).

La Herramienta Clave para la Práctica: Docker

¿Por qué Docker?

Para evitar problemas de configuración.



Docker empaqueta una aplicación con TODAS sus dependencias en un "contenedor" aislado.

Ventajas de Usar Docker en este Curso

- **Entorno Consistente:** Adiós al "en mi ordenador no funciona".
- **Instalación Rápida:** Ejecutamos un comando y tenemos un clúster funcional.
- **Limpieza y Aislamiento:** No instalamos software complejo directamente en nuestro sistema.
- **Habilidad Profesional:** Aprender a usar Docker es una habilidad muy demandada en la industria.

Próximos Pasos: La Clase Práctica

¡Manos a la Obra!

En nuestra primera clase práctica, nos dedicaremos a:

1. Instalar Docker Desktop en sus ordenadores.
2. Descargar y ejecutar nuestro contenedor de Hadoop.
3. Aprender a acceder a la terminal del contenedor.

Preguntas y Cierre

¿Preguntas? @dbytah @carlapvalera

Procesamiento de Grandes Volumenes de Datos

Conferencia 1: La revolución de los datos

Deborah Famadas Rodríguez

Universidad de la Habana

September 1, 2025