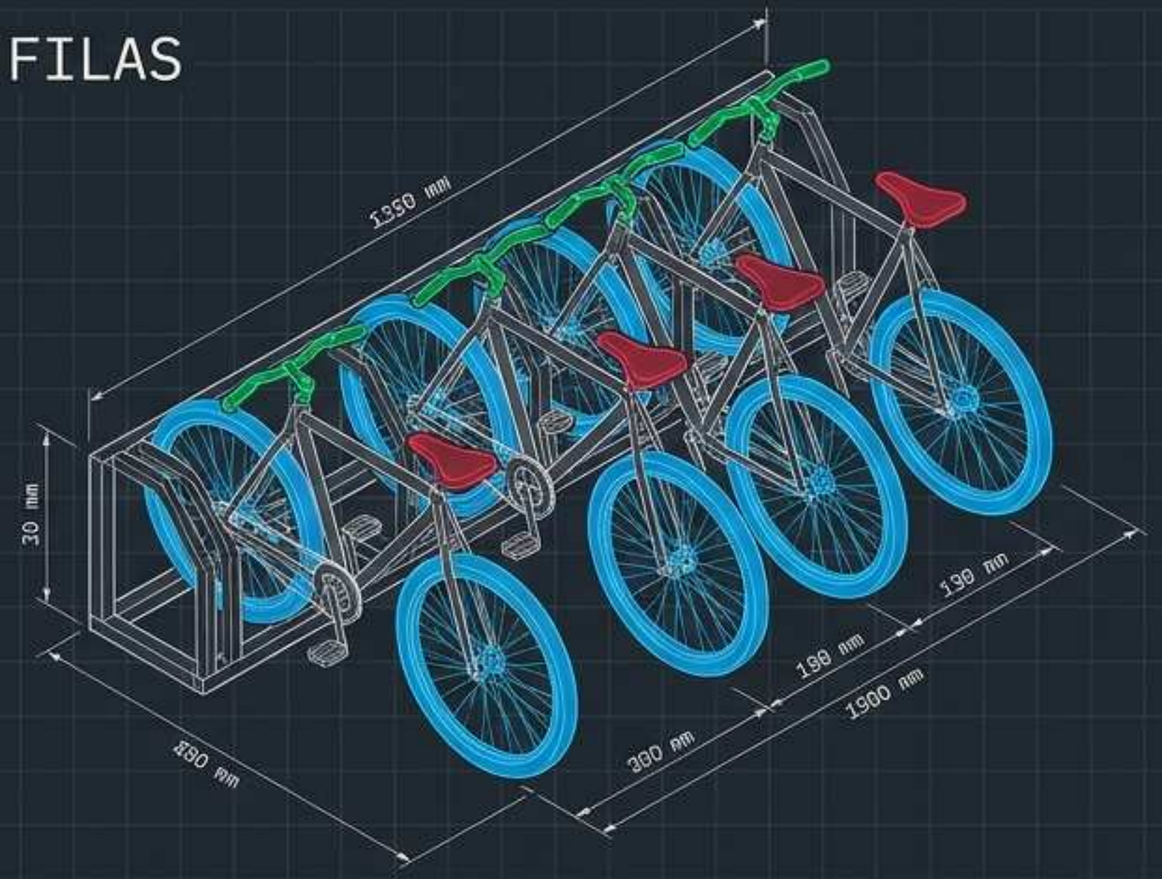
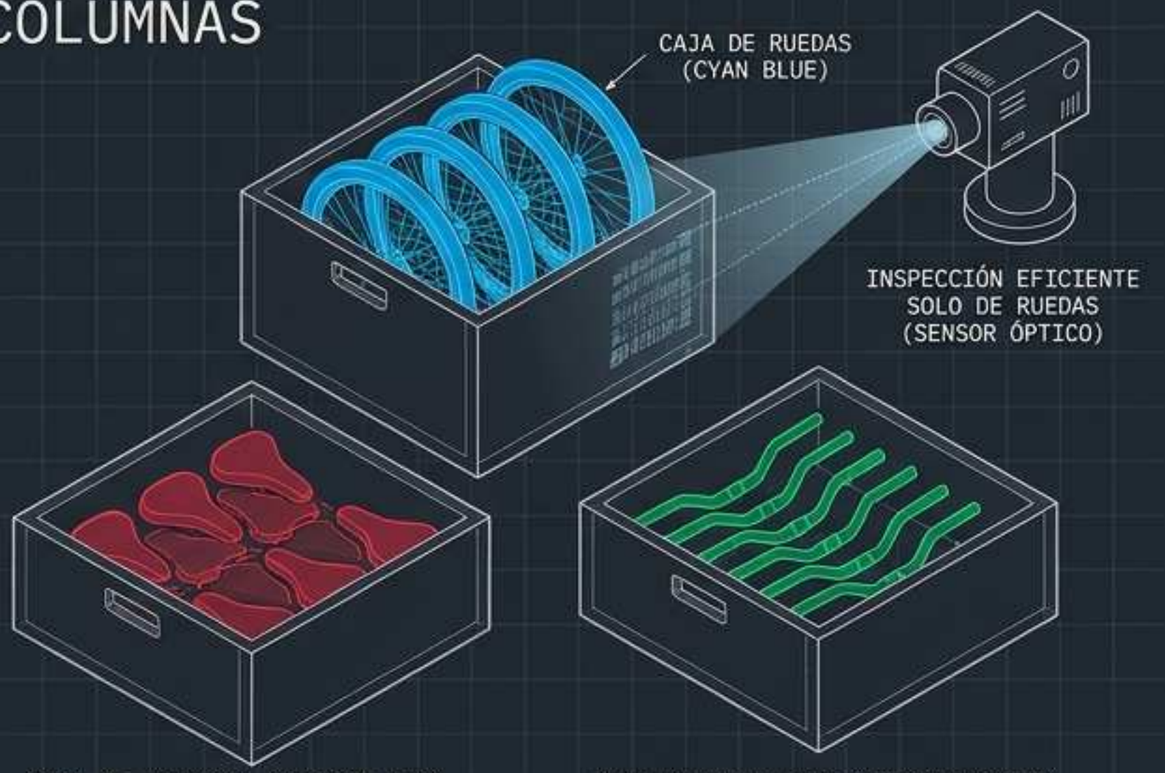
Bases de datos: Por qué una simple decisión física permite leer 12 veces menos memoria

FILAS



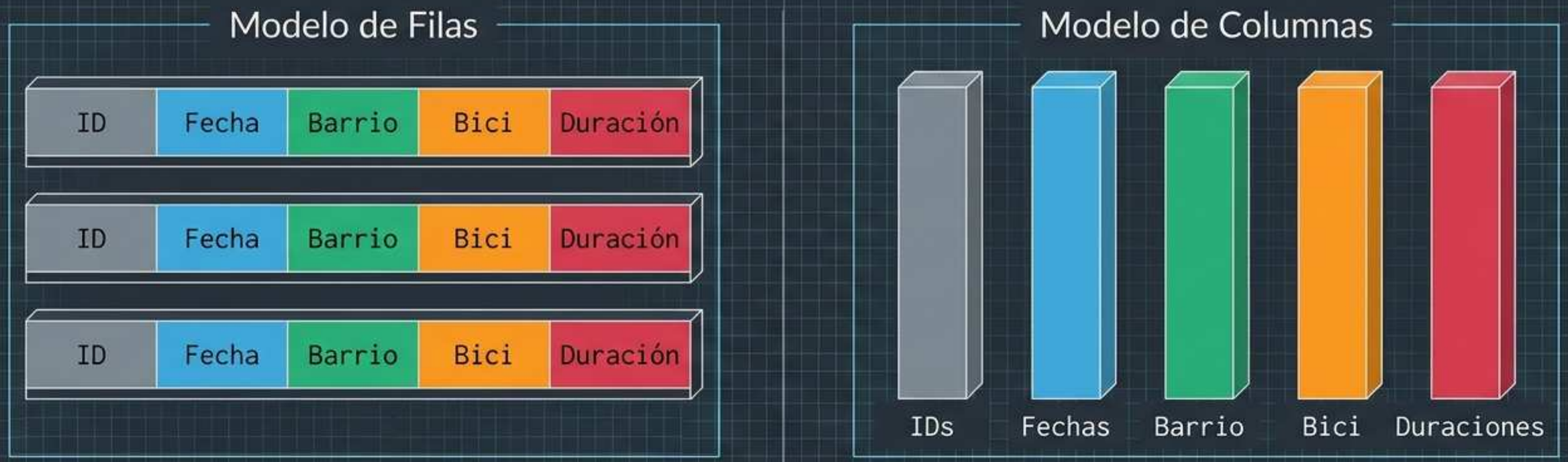
COLUMNAS



Pensamos en los datos como objetos completos, pero para analizar millones de registros rápidamente, el software moderno necesita desarmarlos.

Gemini Notebook

```
SELECT AVG(duration_min)
FROM rides;
```



El lenguaje SQL oculta el diseño físico. Ejecutamos la misma consulta, obtenemos las mismas respuestas, pero la forma en que los datos se asientan en el disco y en la RAM cambia drásticamente la mecánica del sistema.

Gemini Notebook

La especialización de las bases de datos modernas: OLTP frente a OLAP

Dimensión	OLTP (Procesamiento de Transacciones)	OLAP (Procesamiento Analítico)
Disposición de Datos	Almacén por filas (Row-store) 	Almacén por columnas (Column-store) 
Tecnologías Representativas	PostgreSQL, MySQL	DuckDB, ClickHouse
Patrón de Acceso Típico	Leer, escribir o modificar un único registro completo a la vez (ej. el inicio de sesión de un usuario o una reserva).	Leer unas pocas columnas a través de millones de registros para extraer totales, promedios y tendencias.

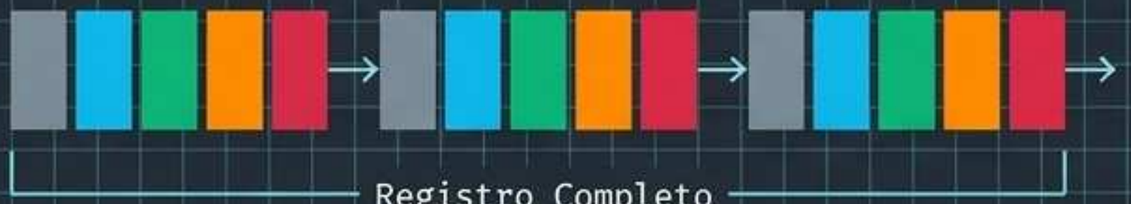
No existe una base de datos universal. La estructura física dicta el rendimiento, forzando a los ingenieros a elegir entre sistemas transaccionales y analíticos.

Gemini Notebook

La física de los datos en la memoria RAM

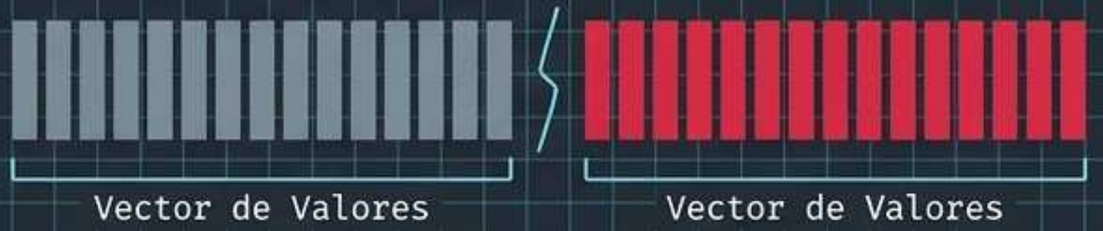
Array of Structs

```
struct Ride {
  ride_id: u64,
  started_at: Timestamp,
  start_area: String,
  bike_type: BikeType,
  duration_min: u32,
}
type Rows = Vec<Ride>;
```



Struct of Arrays

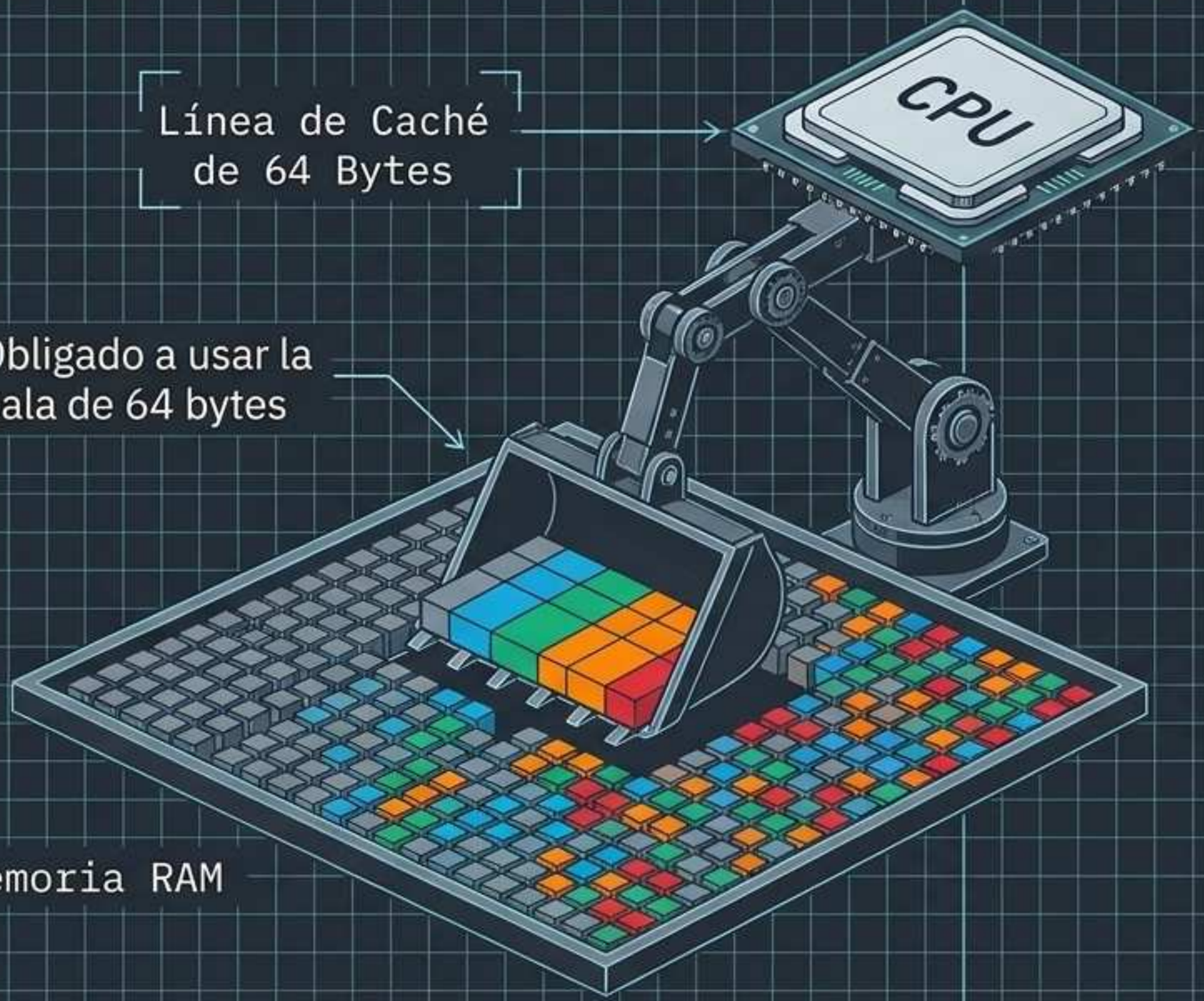
```
struct Columns {
  ride_id: Vec<u64>,
  started_at: Vec<Timestamp>,
  start_area: Vec<String>,
  bike_type: Vec<BikeType>,
  duration_min: Vec<u32>,
}
```



El modelo mental del programador dicta la arquitectura física. Pasar del registro completo al vector de valores transforma el pensamiento humano en pensamiento de máquina.

Gemini Notebook

El hardware no lee valores individuales, extrae bloques de tamaño fijo



La CPU tiene reglas físicas estrictas.

No puede importar la memoria byte a byte. Siempre importa bloques fijos llamados 'líneas de fijos llamados "líneas de caché", arrastrando todo lo que haya alrededor del dato solicitado.

Gemini Notebook

Por qué calcular un promedio requiere leer 12 veces menos datos

Modelo de Filas: Alto Desperdicio



Línea de Caché de 64 Bytes



Modelo de Columnas: Cero Desperdicio



Línea de Caché de 64 Bytes



En un modelo de filas, para leer una sola columna de 4 bytes en un registro de 48 bytes, el hardware se ve obligado a cargar la fila completa. Las columnas eliminan la contaminación de la caché, alcanzando una densidad de utilidad del 100%.

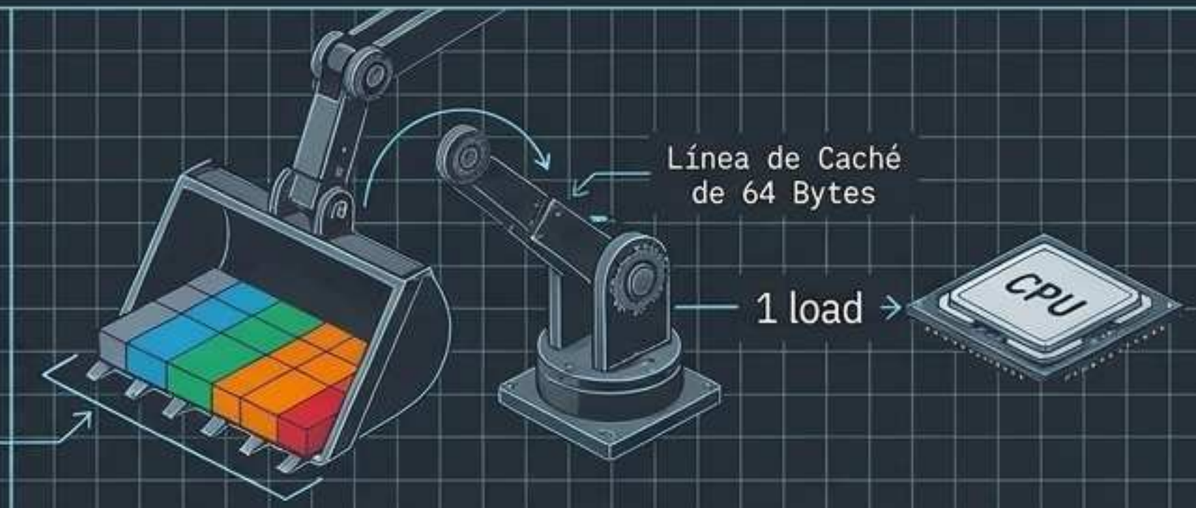
Gemini Notebook

El efecto inverso: Por qué buscar un solo individuo penaliza a las columnas

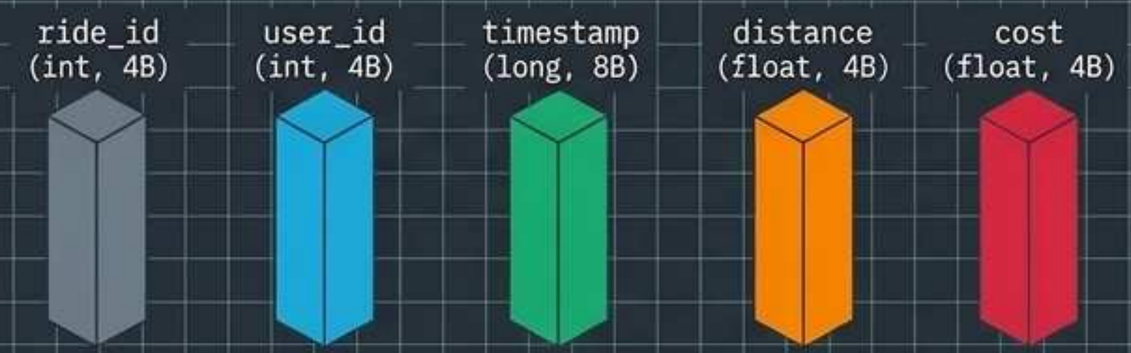
```
SELECT * FROM rides WHERE ride_id = 1004;
```

Fila: Operación Eficiente





Columna: Penalización por Reconstrucción



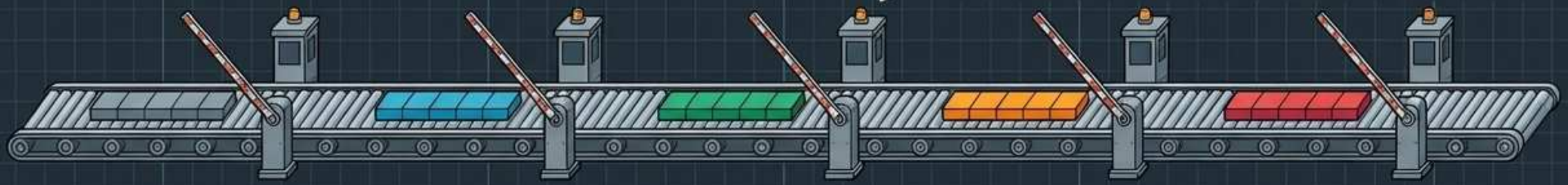


Desensamblar los datos hace que el análisis sea instantáneo,
pero reconstruir un individuo requiere múltiples saltos en memoria.

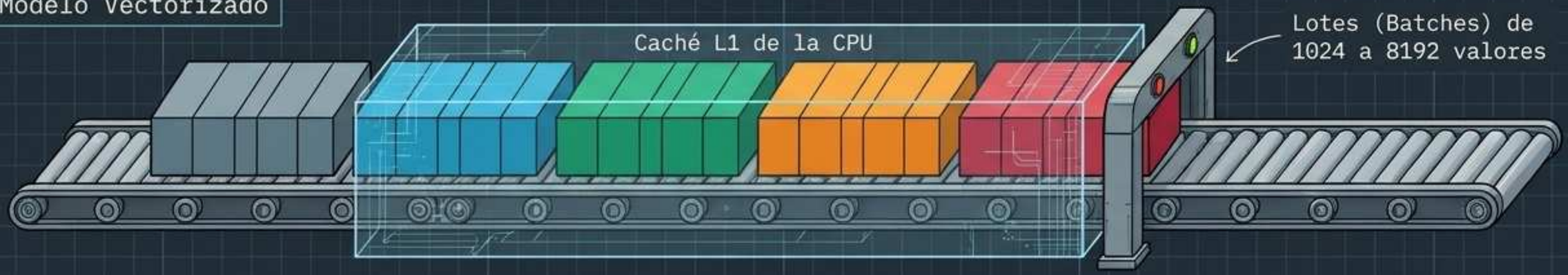
Gemini Notebook

Adiós al modelo Volcano: La revolución de la ejecución vectorizada

Modelo Tradicional (Volcano)



Modelo Vectorizado

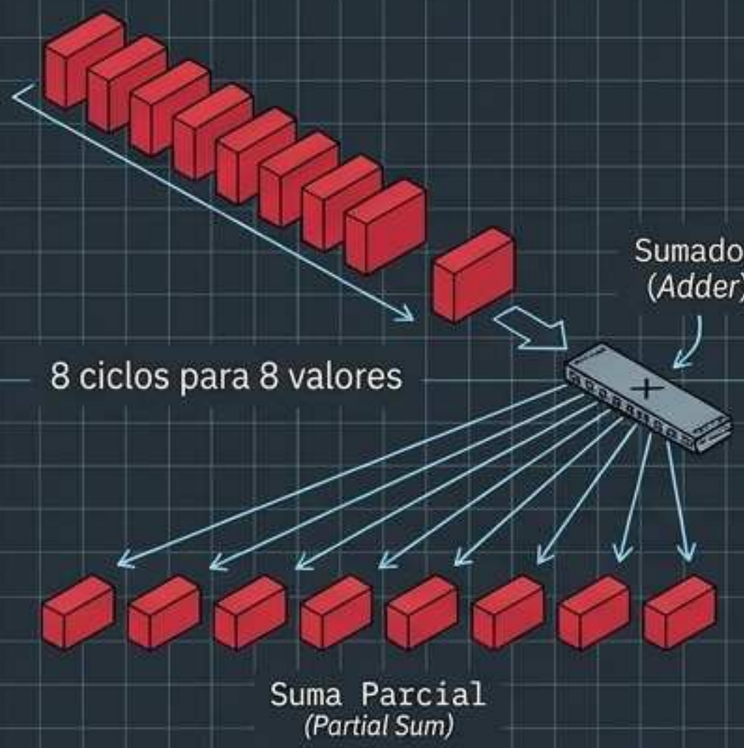


Los motores clásicos evalúan los datos fila por fila, pagando una penalización de sobrecarga en cada paso. Los motores columnares procesan 'lotes' de miles de valores diseñados matemáticamente para encajar en la caché rápida del procesador.

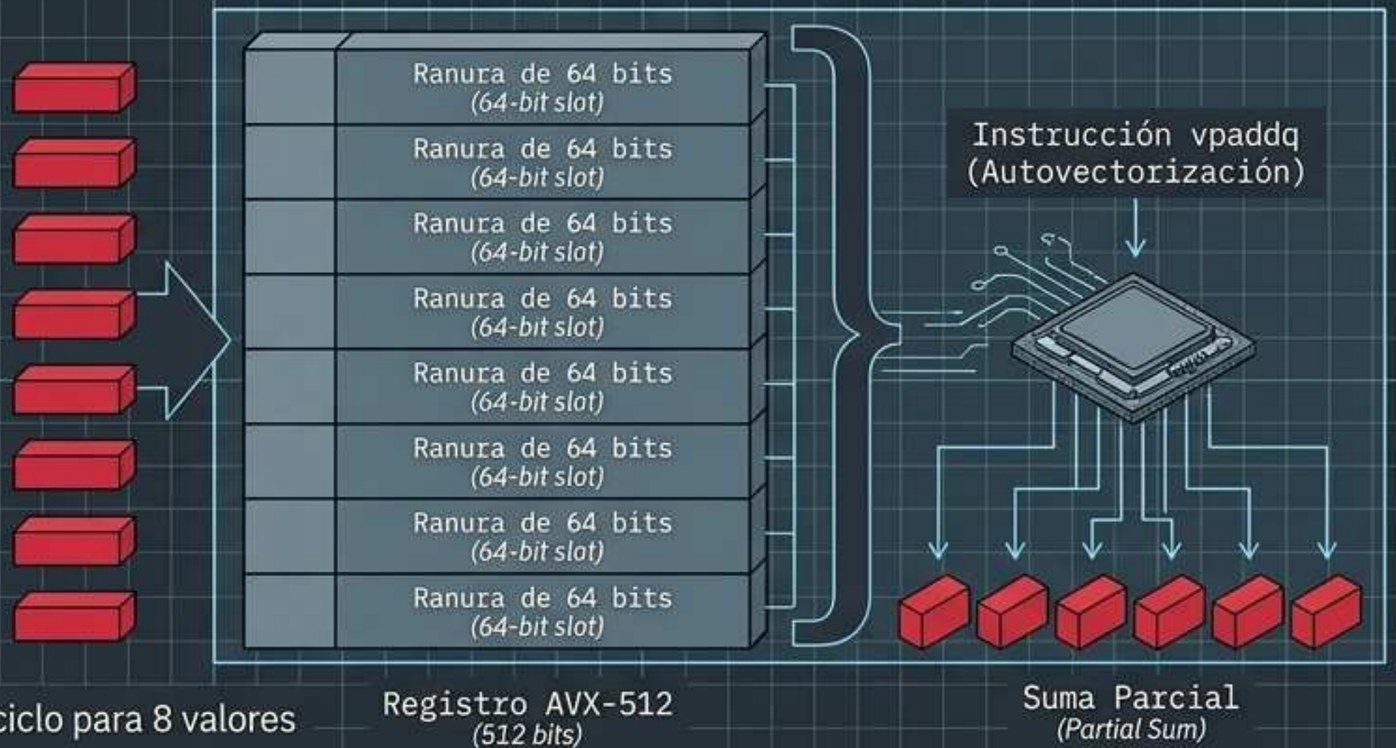
Gemini Notebook

SIMD: Sumar ocho valores en un solo ciclo de reloj

Procesamiento Secuencial



Paralelismo de Datos (AVX-512)

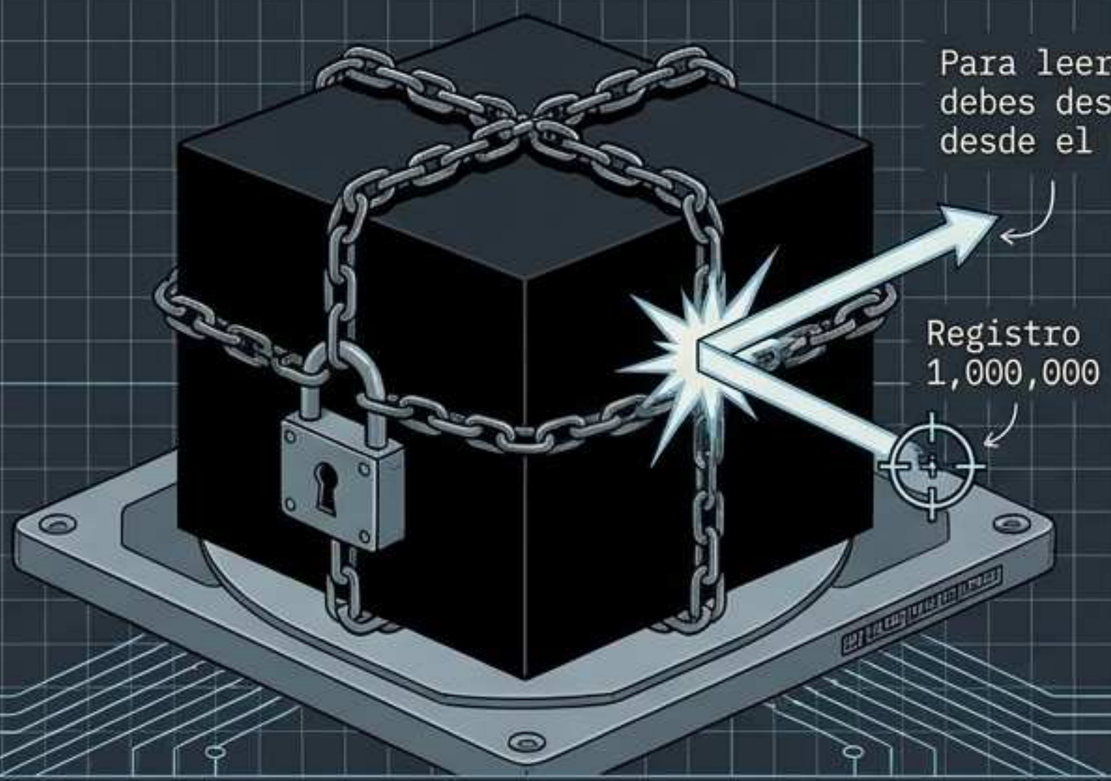


Cuando los valores del mismo tipo están contiguos, el compilador activa la Autovectorización. Las instrucciones SIMD de las CPUs modernas operan sobre múltiples puntos de datos simultáneamente, acelerando el bucle analítico sin intervención del programador.

Gemini Notebook

Almacenamiento: El cuello de botella del acceso aleatorio

Compresión de Propósito General (.ZIP)



Formato Columnar Estructurado

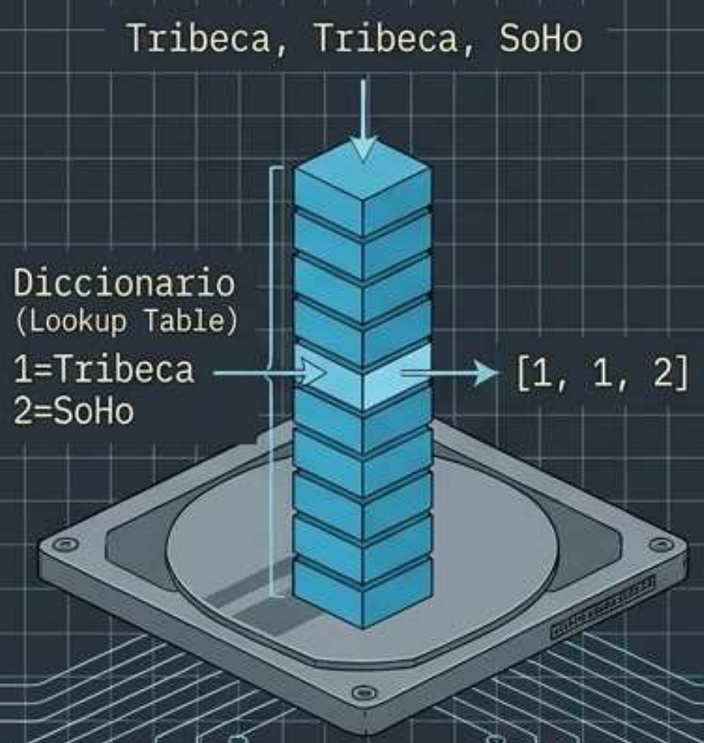


Leer desde el disco suele ser el límite de velocidad. La compresión estándar trata los datos como una caja negra, destruyendo la capacidad de saltar a un valor específico sin descomprimir todo el archivo.

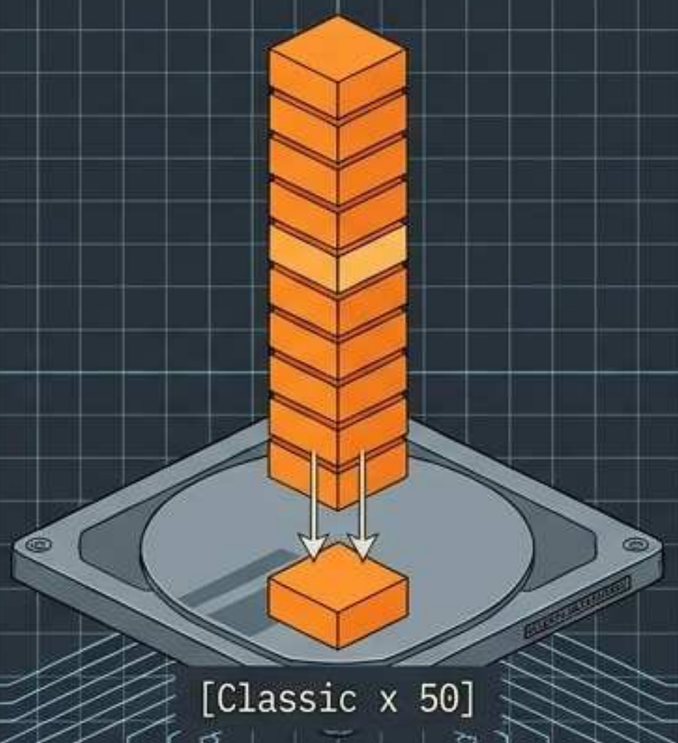
Gemini Notebook

Codificaciones ligeras: Comprimir sin destruir la estructura

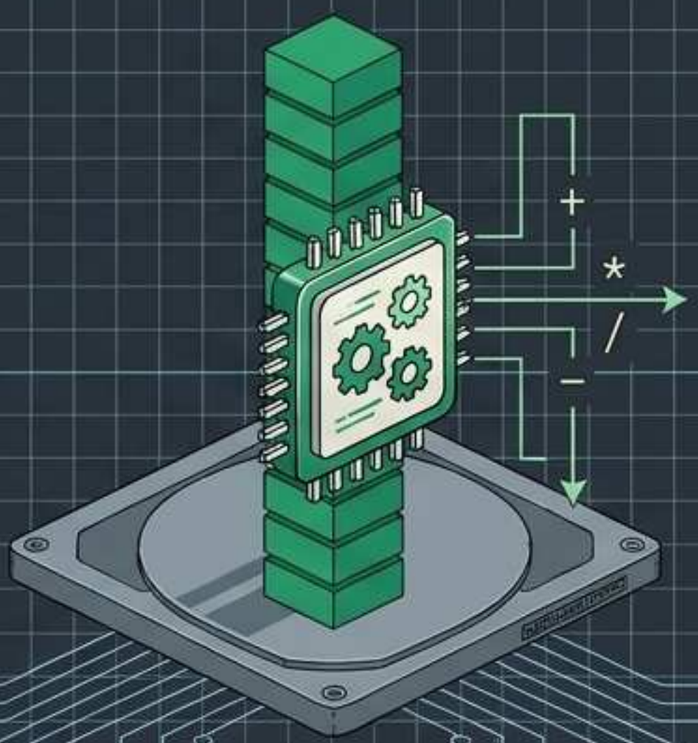
Diccionarios / Barrios



Run-Length Encoding (RLE) / Bici



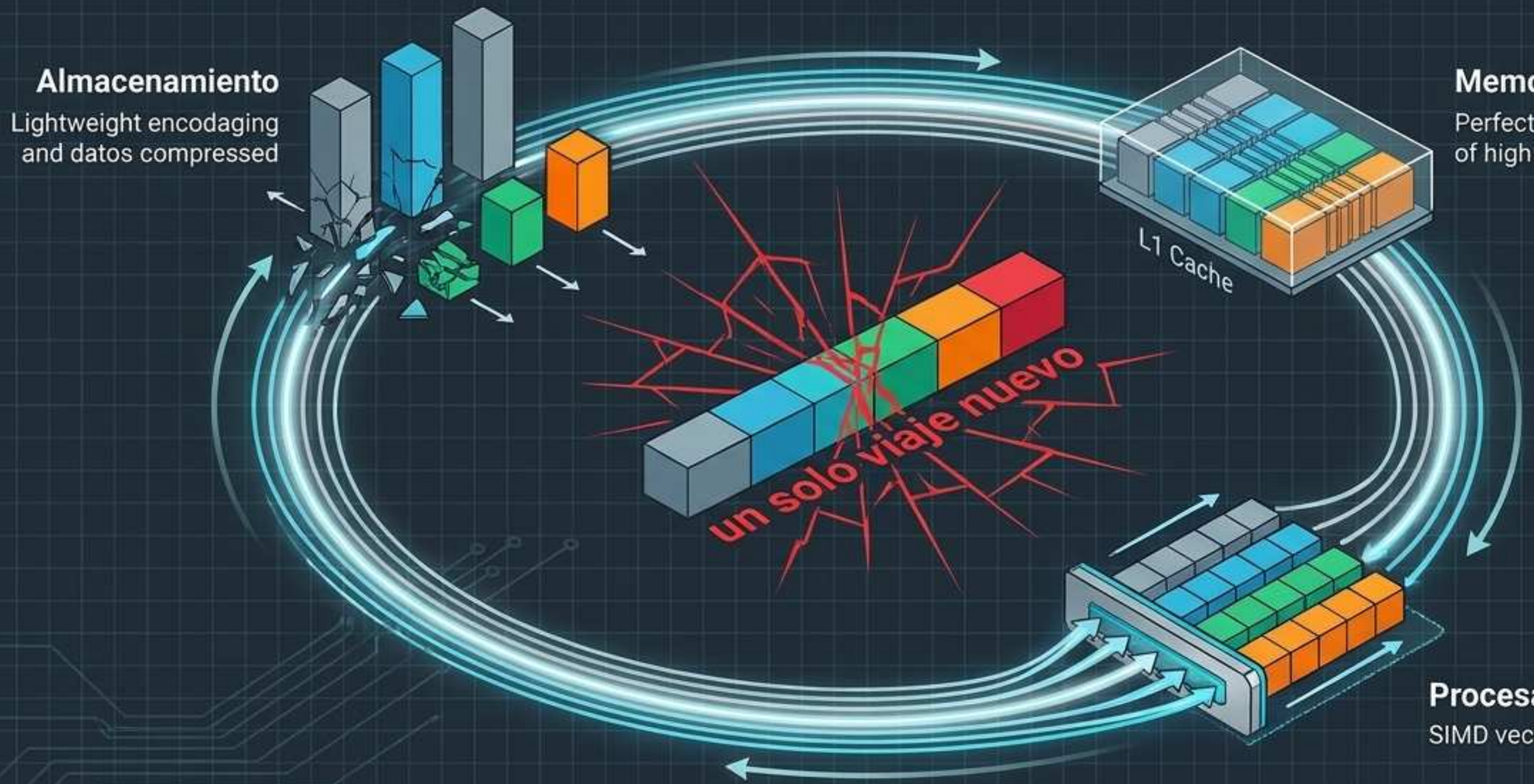
Materialización Tardía



Al agrupar datos similares, podemos usar compresión nativa. Permite saltar directo al dato necesario e incluso operar matemáticamente sin descomprimir.

Gemini Notebook

La anatomía de la especialización: Una simetría rota

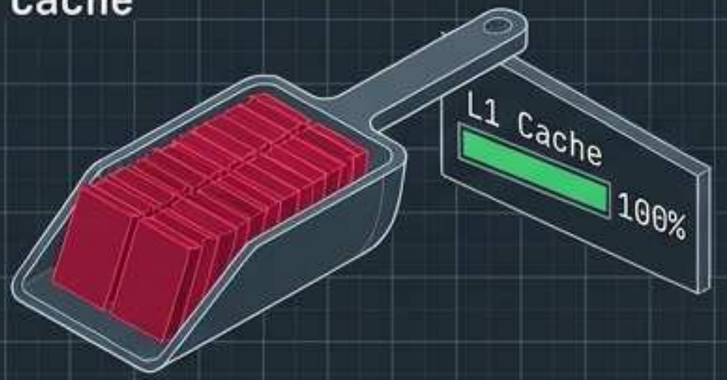


Lo que hace que un almacén de datos sea increiblemente rápido para el análisis (compresión predictiva, contigüidad en RAM, vectorización SIMD) es exactamente lo que lo hace ineficiente y torpe para insertar o modificar un solo registro.

Gemini Notebook

El hardware dicta las reglas: Pensar en columnas

Densidad de Caché (Memoria)



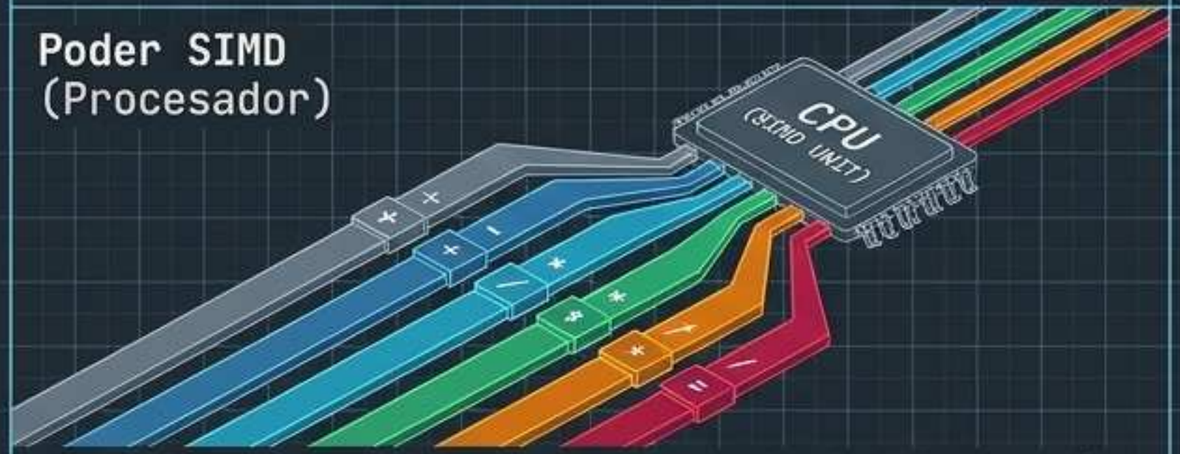
Cero desperdicio al cargar bloques de 64 bytes.

Modelo Vectorizado (Software)



Lotes que eliminan la sobrecarga de evaluación fila por fila.

Poder SIMD (Procesador)



Operaciones matemáticas en paralelo a nivel de instrucción.

Acceso Aleatorio (Disco)



Codificaciones ligeras que permiten materialización tardía.

Lógicamente, los humanos conceptualizan el mundo en filas.
Físicamente, el silicio moderno exige columnas para el análisis a gran escala.

Para profundizar en la ingeniería física de las bases de datos, los cursos del sistema Carnegie Mellon (CMU 15-445) documentan este estándar de la industria

Gemini Notebook