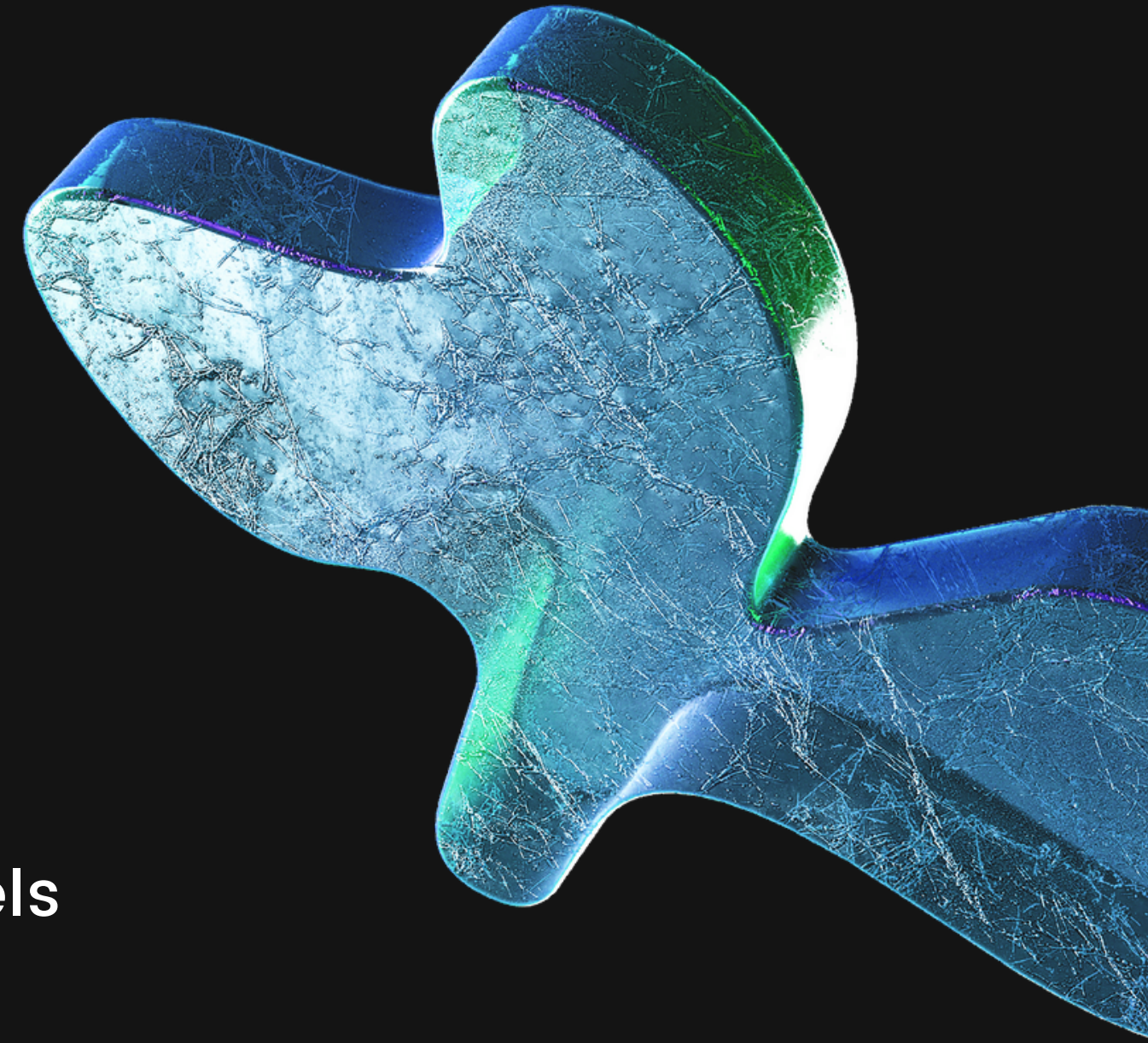


Seminario

Intune

Data pipeline optimization for deep Recommendation models

Ipinza, Vergara, Mariz, Hernández



CONTEXTO

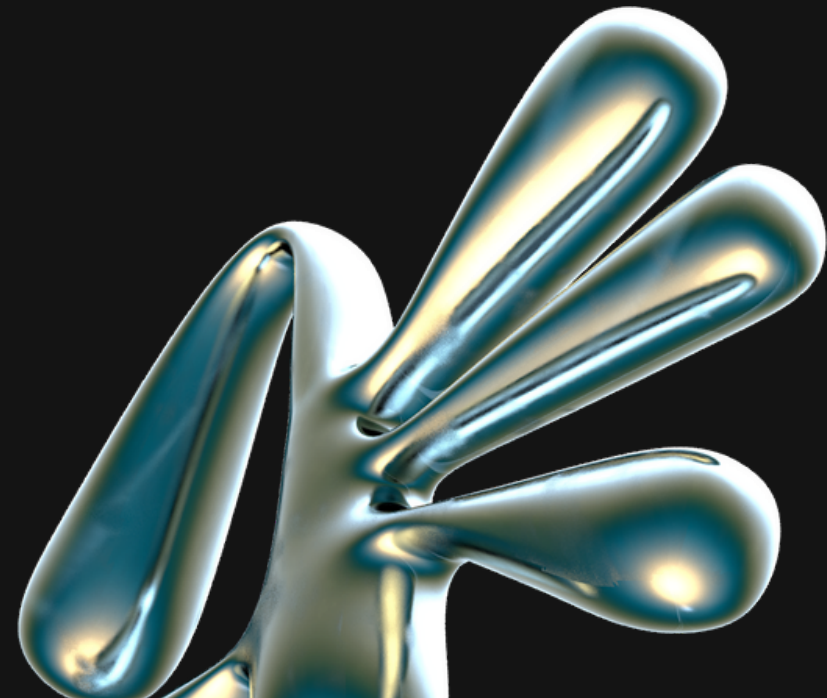
- Los modelos de aprendizaje profundo y su creciente popularidad en el mundo de los sistemas recomendadores

Introducción

- Problemas con el procesamiento de datos en sistemas de aprendizaje profundo.
- Destacar la importancia de mejorar el procesamiento en busca de mejorar el desempeño y reducir costos

Antes de InTune: AutoTune

- Baja eficiencia
- Altas tasas de error
- Poco apoyo a reescalamiento

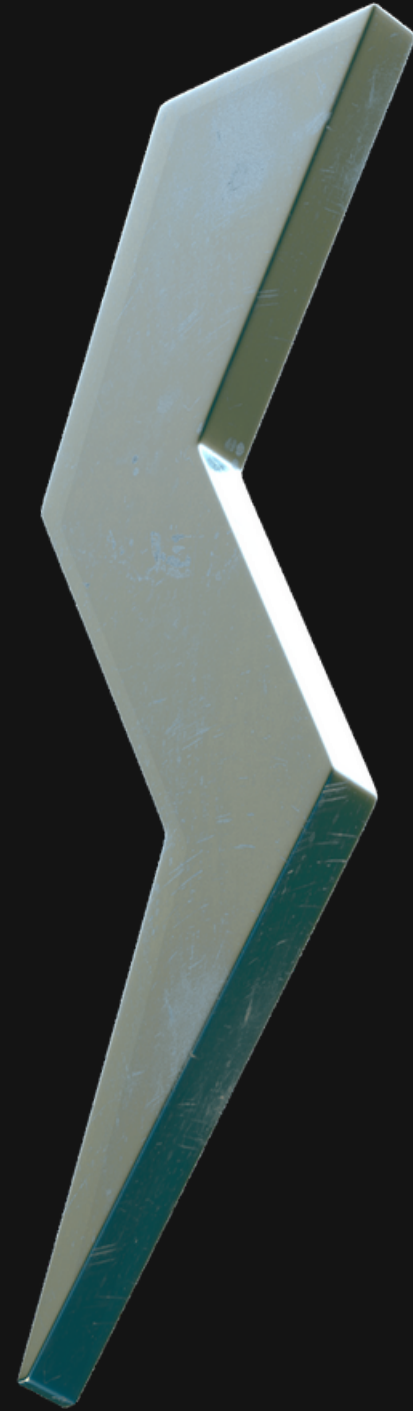


60% del tiempo
se dedica a la
ingesta de datos

Oportunidad: Se
crea InTune

Objetivos

Qué necesitas saber



¿Qué es Intune?

¿Por qué se creó Intune?

¿Qué se espera de Intune?

¿Cuáles son los beneficios de Intune?

¿Cómo fue el desarrollo de InTune?

Marco Teórico

Pipeline

AUTOMATIZACION

- Automatizar procesos
- Flujo de trabajo

MODULARIDAD

- Módulos independientes
- Reutilización y adaptación

SECUENCIALIDAD

- Ejecución módulos de forma secuencial
- Compatible con ejecución en paralelo

Deep learning recommender models

DRLM DATA PROCESSING

EMBEDDING TABLE

- Matriz de vectores embedding
- Reducción de dimensionalidad

ESCALABILIDAD

- Las embedding tables ocupan grandes cantidades de memoria

PROCESAMIENTO DE DATOS

- Disk Load
- Batch
- Shuffle
- UDF
- Prefetch

Reinforcement Learning

RECOMPENSA Y CASTIGO

- Maximizar recompensa
- Premio
- Castigo

EXPLOTACIÓN VS EXPLORACIÓN

- Explotación de las recompensas conocidas
- Exploración de nuevas posibles recompensas

Herramientas Existentes

AUTOTUNE

- Optimización de data Pipeline.
- Errores de memoria
- Mala escalabilidad

DATA PREPROCESSING SERVICE

- Servicio de Meta para optimización de ingreso de datos.
- Esta hecho a medida para la arquitectura de meta

NVT TABULAR

- Herramienta de Nvidia para cargar datos con GPU
- Problemas de Memoria de GPU
- Comparte recursos con entrenamiento

AUTOTUNE

La mejor herramienta actualmente

BAJO RENDIMIENTO

- Mal rendimiento en DLRM

ERRORES DE MEMORIA

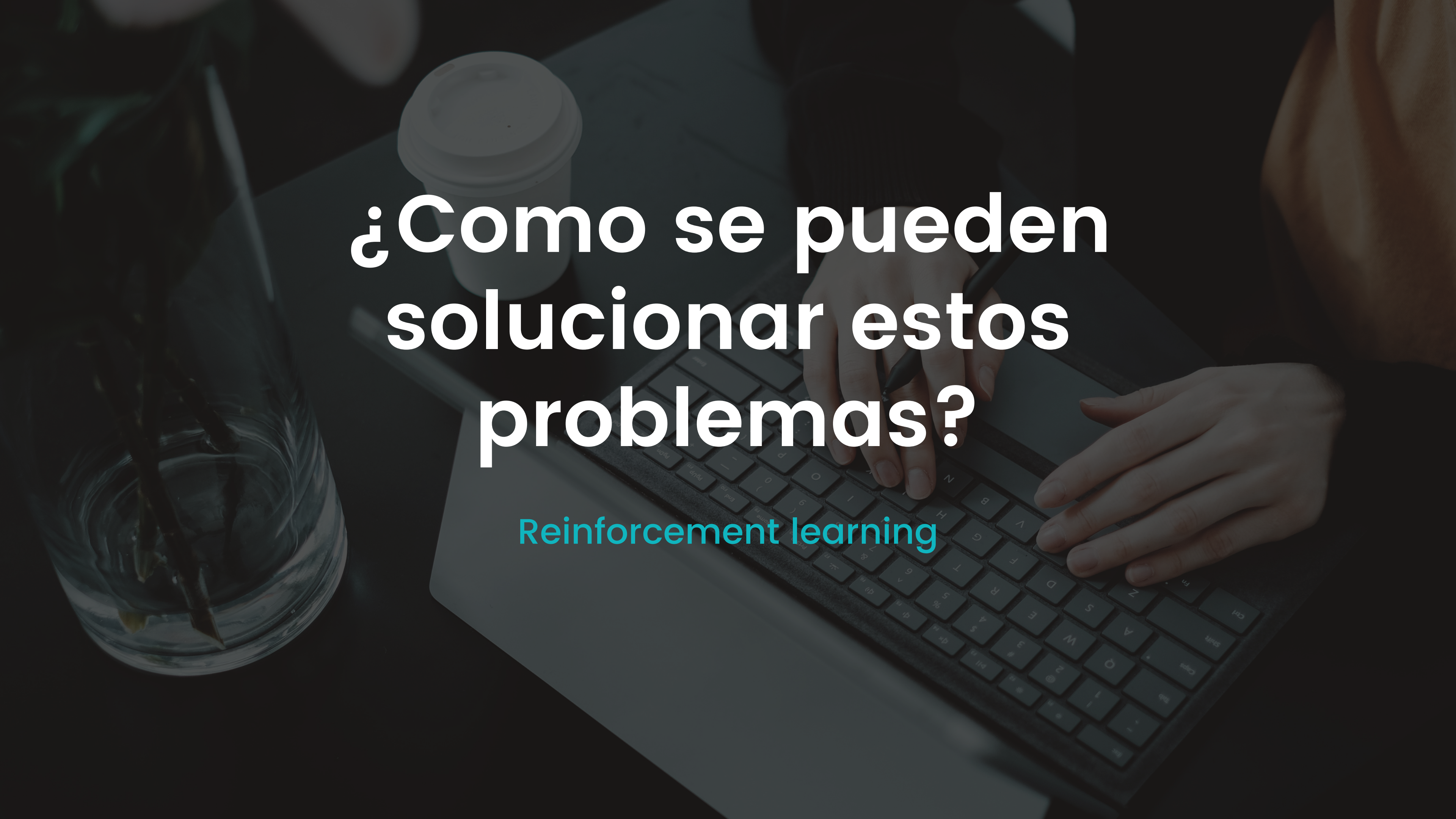
- Prefetch sobrepasa memoria de sistema

BAJO RENDIMIENTO FUNCIONES UDF

- UDF generan cuellos de botella en pipeline

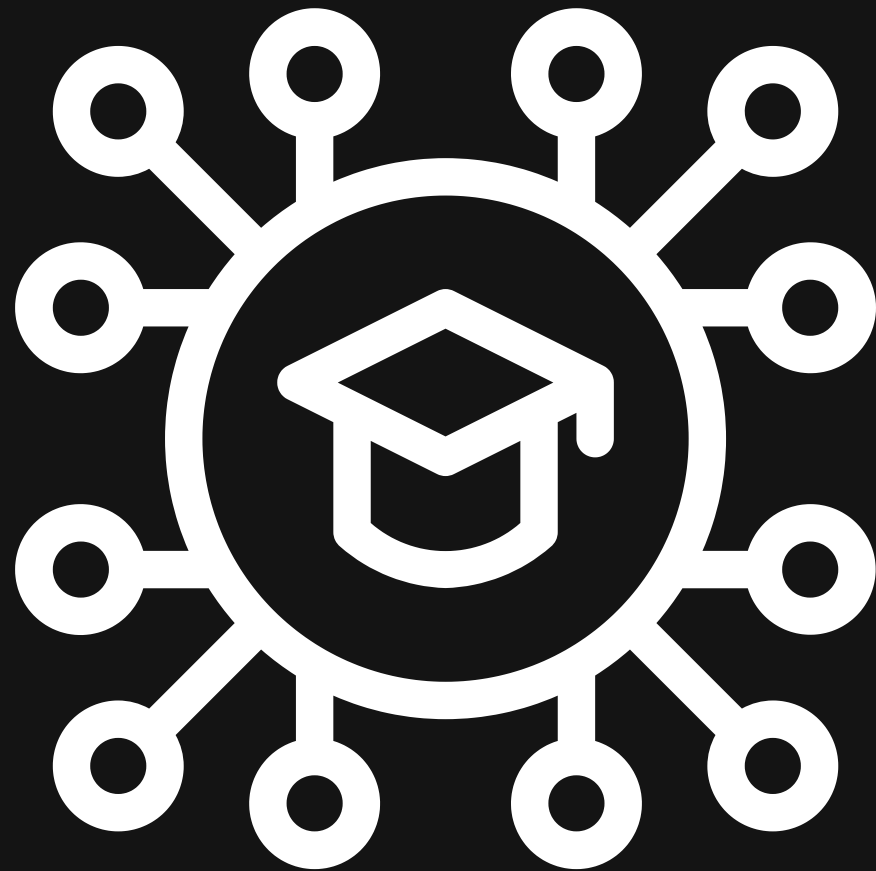
PROBLEMAS CON REESCALADO

- No responde a la modificación de recursos

A person's hands are shown typing on a laptop keyboard. A white coffee cup is on the desk to the left, and a glass of water with a straw is in the foreground. The background is dark and out of focus.

**¿Como se pueden
solucionar estos
problemas?**

Reinforcement learning



Reinforcement learning

- Basado en feedback
- Ajuste de recursos y parametros
- Aborda problemas de AUTOTUNE

Factores del Environment

MODIFICABLES POR AGENTE

- Pipeline Latency
- CPU libre
- Memoria libre

SIN CORRELACION

- Model Latency

ESTATICOS

- Ancho de banda DRAM
- Velocidad CPU

Modelo InTune

Características del modelo

PREMIO

Premio calculado en base al uso de memoria y pipeline latency

$$R = throughput \times (1 - memoryused / memorytotal)$$

3 CAPAS RELU

Los negocios pueden mover grandes cantidades de datos sin preocuparse por los problemas de la red que pueden afectar al negocio.

ACTION SPACE INCREMENTAL

Los negocios no tendrán que preocuparse por agregar más dispositivos conectados que muevan datos esenciales a su red.

Evaluación

Problemas a abordar

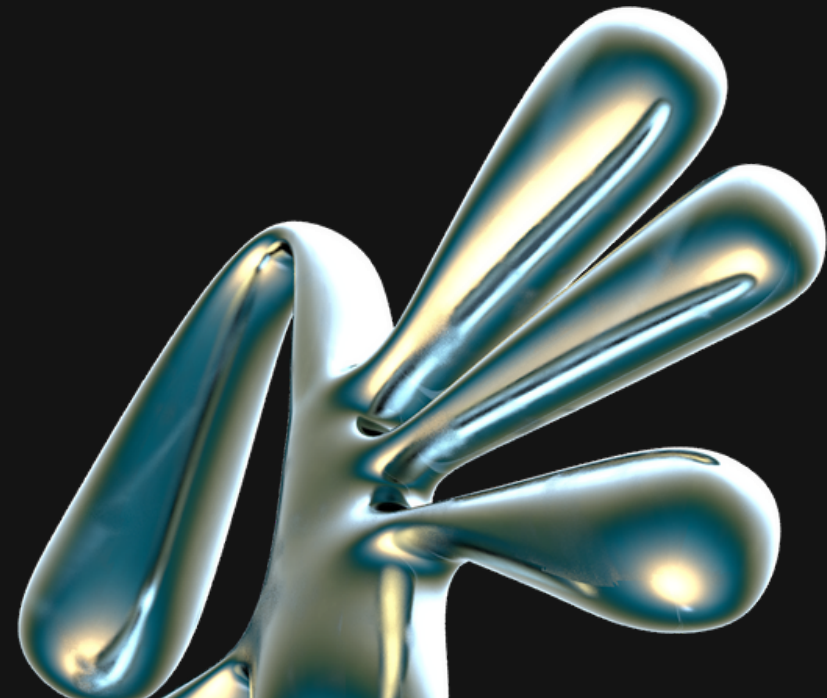
¿TIENE MENOS ERRORES DE
MEMORIA?

¿RESPONDE AL RE ESCALAMIENTO?

¿PERMITE MANEJAR MAYOR
VOLUMEN DE DATOS QUE
AUTOTUNE?

Caso Custom

- DRLM Predicción productos
- Docenas de columnas sparse, 5 continuas, 10 mil filas
- 5 Millones de parámetros
- GPUA100 40 GB Modelo
- 32 CPU Intel Xeon 3.0 Ghz Pipeline



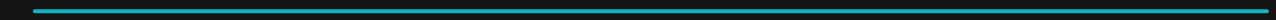
Caso Meta

- DLRM de Meta para predicción de clicks
- 26 columnas sparse, 13 continuas y 24 mil filas
- 25 Biliones de parámetros
- 2 GPUA100 40 GB Modelo
- 32 CPU Intel Xeon 3.0 Ghz Pipeline

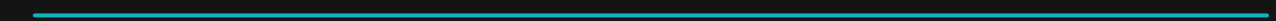
Baseline

Escalado Piramidal

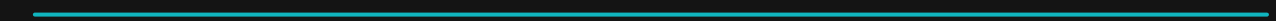
Unoptimized



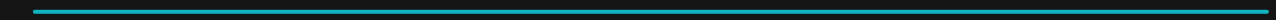
Autotune



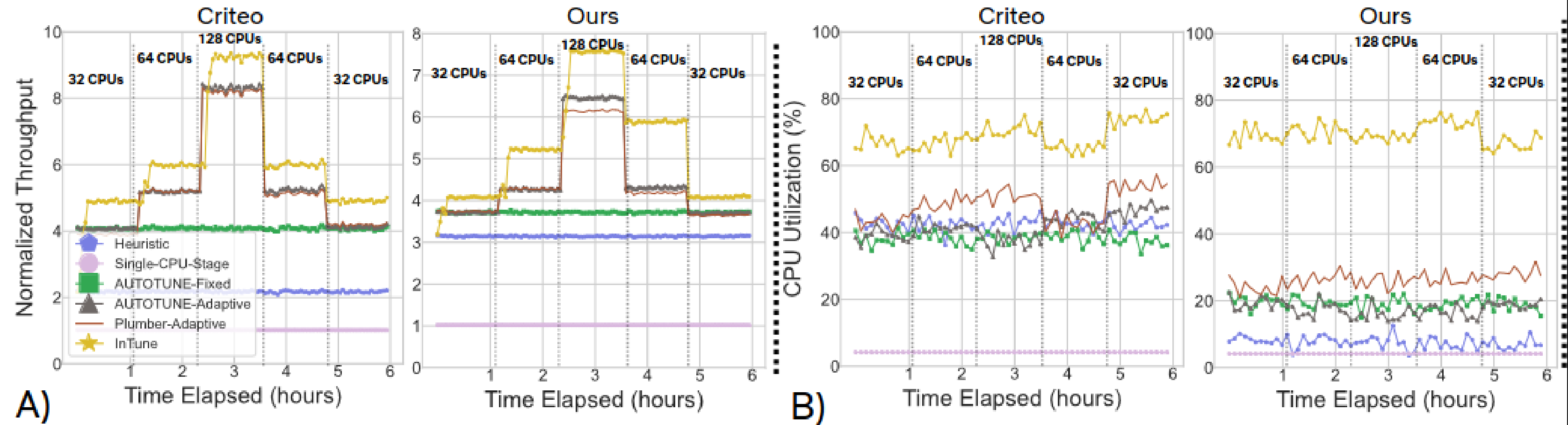
AUTOTUNE-Adaptative



Plumber-Adaptative



Heuristic



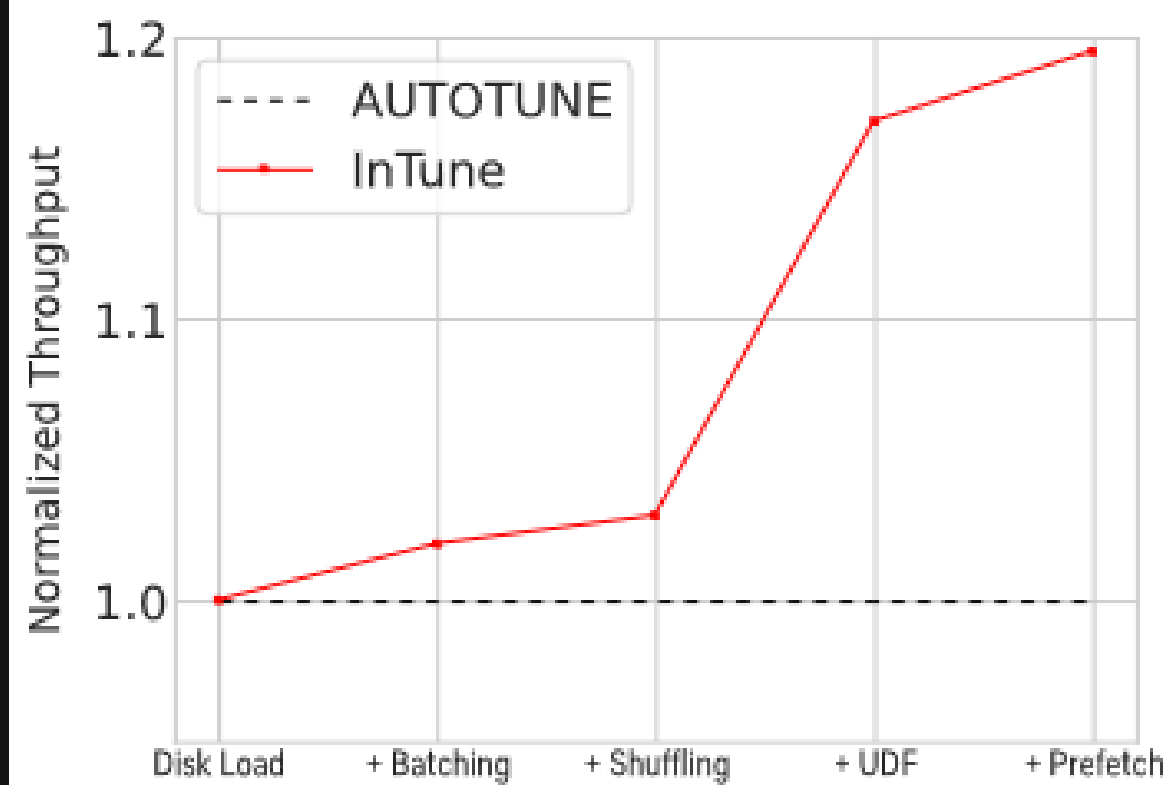
UNOPTIMIZED COMO BASE

RENDIMIENTO

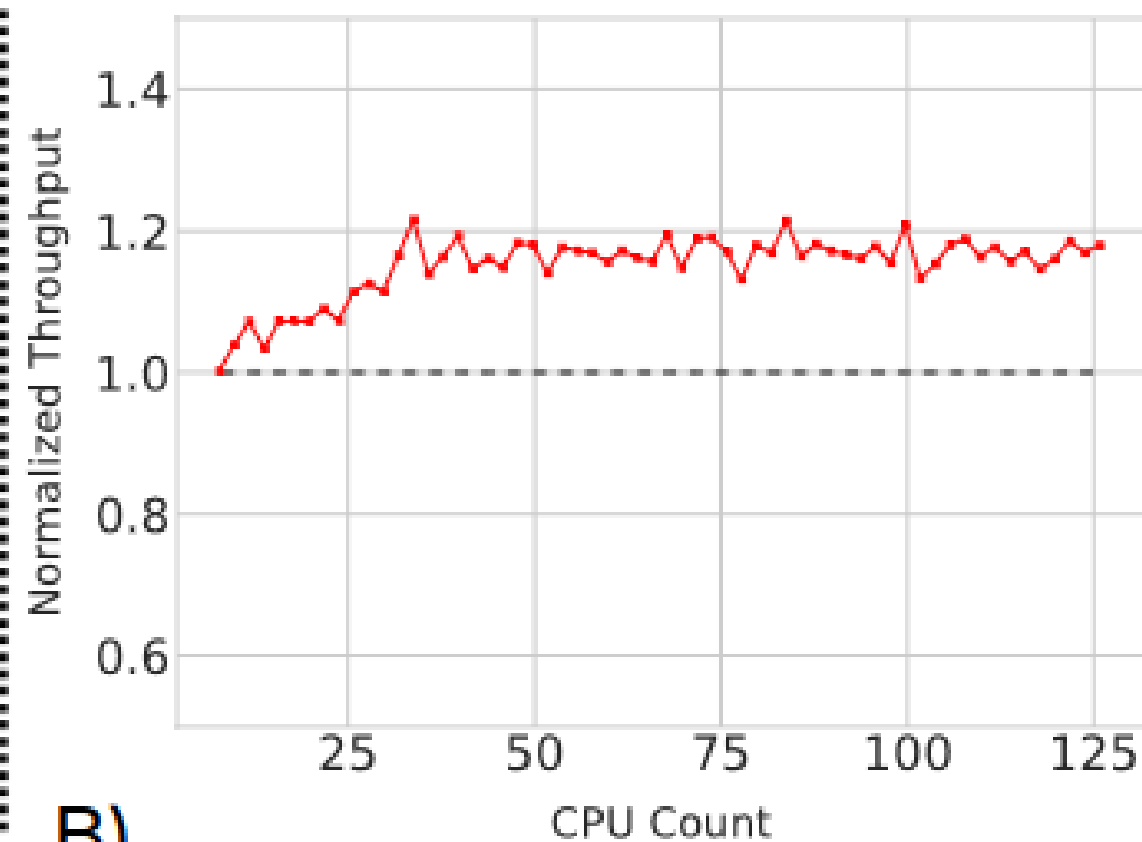
- 2.05X a 2.29X Por sobre AUTOTUNE
- 10%-20% Sobre Metodos con intervencion humana

ERRORES DE MEMORIA

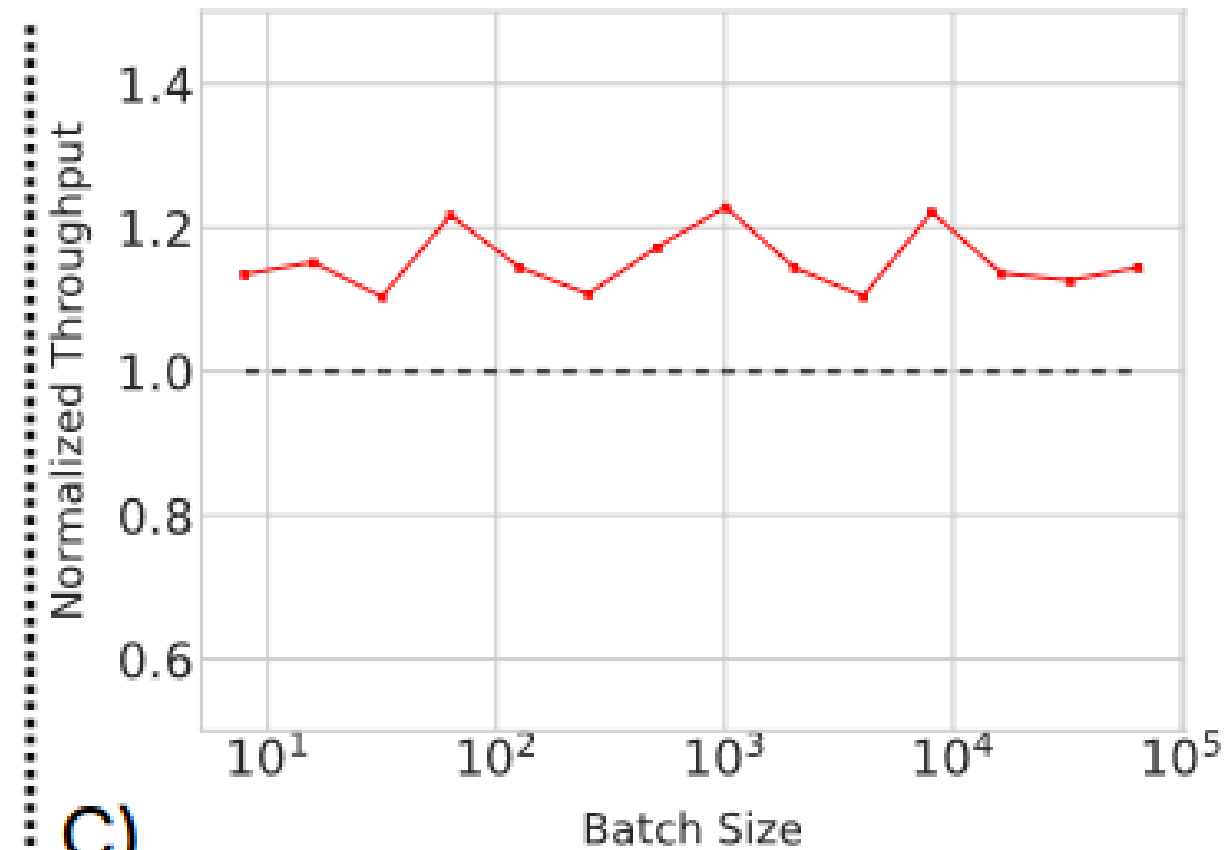
- Autotune 8% errores de memoria
- Intune 0% errores de memoria



A)



B)



C)

COMPLEJIDAD PIPELINE

- Mejora significativa en UDF

RE ESCALADO

- Mejora del 20%

BATCH SIZE

- Mejora significativa respecto AUTOTUNE

Conclusión

Problema DLRM

- Mal rendimiento pipeline DLRM

AUTOTUNE

- Herramienta aceptada optimizador.
- Tiene problemas en la práctica que reducen eficiencia.

InTune

Se propone un optimizador automatizado de canalización de datos

Rendimiento

InTune supera las líneas de base por un factor de 1.18 a 2.29 veces

Referencias

- Michael Kuchnik, Ana Klimovic, Jiri Simsa, Virginia Smith, and George Amvrosiadis. 2022. Plumber: Diagnosing and removing performance bottlenecks in machine learning data pipelines. *Proceedings of Machine Learning and Systems* 4 (2022), 33–51.
- Mark Zhao, Niket Agarwal, Aarti Basant, Buğra Gedik, Satadru Pan, Mustafa Ozdal, Rakesh Komuravelli, Jerry Pan, Tianshu Bao, Haowei Lu, Sundaram Narayanan, Jack Langman, Kevin Wilfong, Harsha Rastogi, Carole-Jean Wu, Christos Kozyrakis, and Parik Pol. 2022. Understanding data storage and ingestion for large-scale deep recommendation model training. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*. ACM. <https://doi.org/10.1145/3470496.3533044>