FAST
FORWARD
XPERIENCE

web summit
LISBON

COBERTURA AO VIVO DA CURADORIA FFX

PALCO	DIA	LISBOA
AI SUMMIT	**10 NOV 2026**	**14:00**

SE A IA ENTENDE 100 MILHÕES DE FOTOS POR US$ 200, O QUE VOCÊ FAZ COM ISSO?

IA · DADOS · BUSCA · MODELOS ABERTOS

Alexey Milovidov

Cofundador e CTO da ClickHouse, começou o projeto em 2009, dentro da Yandex, para processar os dados do que a própria empresa descreve como o segundo maior sistema de web analytics do mundo.

EM 30 SEGUNDOS

Alexey Milovidov, da ClickHouse, transformou quase 100 milhões de fotos em vetores de significado com cerca de US$ 200 de GPU alugada.

Calculados uma vez, esses vetores servem para buscar e classificar sem pagar um modelo de linguagem a cada consulta.

NÚMEROS DO PALCO

US$ 200

GPU na AWS para vetorizar as fotos e os sites do teste

99,2 mi

Fotos da base YFCC100M usadas no teste da ClickHouse

4 bits

Por número do vetor, com busca perto de 100% de acerto

PARA LEVAR

1. **Classificar sentimento e tema não exige mandar cada texto a um modelo de linguagem: vetor pronto e um classificador simples resolvem.**
2. **Guardar vetores com pouca precisão é escolha de projeto: 32 ou 64 bits por número é desperdício.**
3. **Escolha o modelo pelo acerto no seu próprio português, não pela posição na classificação global.**

O ARTIGO

3 MIN DE LEITURA

Transformar quase 100 milhões de fotos em números que resumem o que cada uma mostra custa cerca de US$ 200 em placas de vídeo (GPU) alugadas na AWS, segundo Alexey Milovidov, cofundador e CTO da ClickHouse.

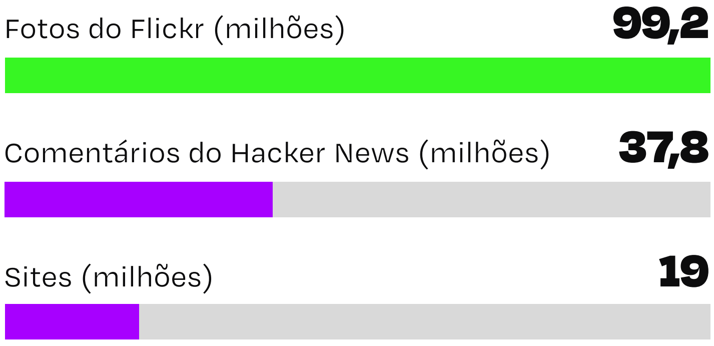
Com essa conta, o obstáculo deixa de ser o preço de fazer a IA entender imagem e texto em grande volume. Passa a ser o que a empresa faz com esses números depois de prontos e qual modelo escolhe para gerá-los.

Esses números, chamados de vetores (ou embeddings), vêm de modelos treinados para uma tarefa só: pôr perto o que tem sentido parecido. Milovidov os separou dos números tirados do meio de um modelo de linguagem, que, segundo ele, "vai meio que funcionar", mas não foi o que ele usou. Com os modelos certos, processou as três bases do slide: 99,2 milhões de fotos do Flickr, 37,8 milhões de comentários do fórum Hacker News e 19 milhões de sites.

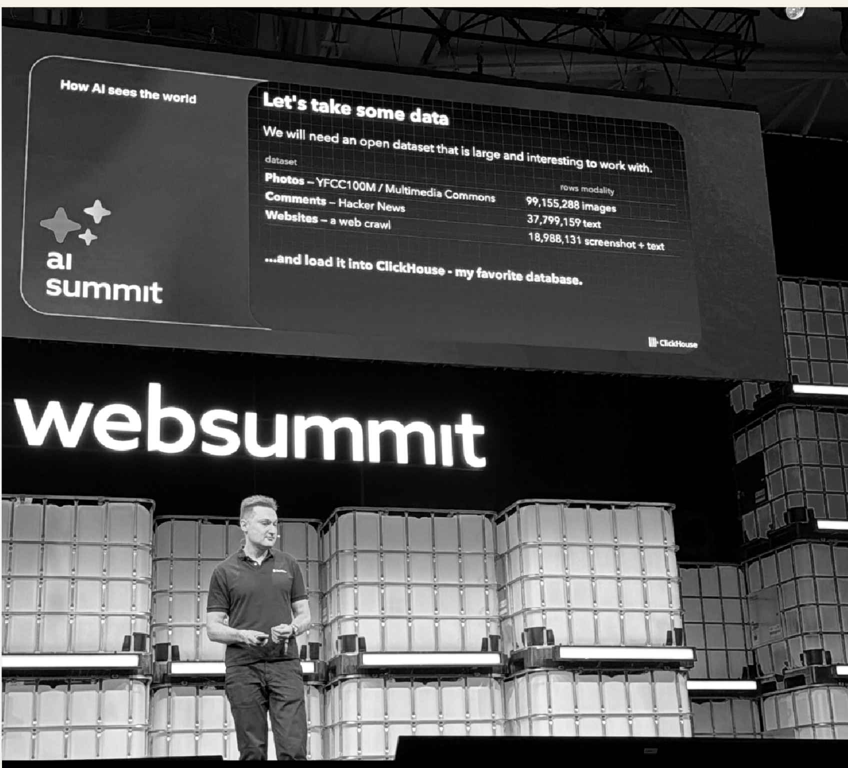
Com os vetores prontos, um modelo estatístico dos mais simples, uma regressão logística muito pequena, separou elogio de crítica nos comentários do Hacker News sobre a ClickHouse. Um comentário que parecia negativo era, lido de perto no palco, um elogio à empresa, e o modelo tinha acertado. Outro classificador do mesmo tipo pôs comentários sobre Elon Musk na categoria sociedade e política. **O vetor é calculado uma vez; cada tarefa nova pede só um classificador por cima.**

Calcular uma vez só compensa se guardar for barato, e Milovidov mostrou como. Para ele, guardar cada número com 32 ou 64 bits é desperdício, e um formato de 16 bits já é mais do que o necessário. A ClickHouse criou um tipo de dado que guarda cada bit do vetor em separado; a consulta pode ler 4, 2 ou só 1 bit, trocando precisão por velocidade. Com 4 bits, segundo ele, a busca devolve perto de 100% do que o vetor completo daria.

Três bases somam mais de 150 milhões de itens



Para a FFX, a conta se divide em duas: o custo de calcular o significado, pago uma vez por item, e o custo de perguntar, pago a cada consulta. **Quem manda cada avaliação de produto ou cada reclamação para um modelo de linguagem paga a segunda conta em tarefas que a primeira resolveria.** Com modelos abertos e GPU alugada, fazer em casa fica ao alcance de muitas empresas.



How AI sees the world

Let's take some data

We will need an open dataset that is large and interesting to work with.

dataset	rows	modality
Photos – YFCC100M / Multimedia Commons	99,155,288	images
Comments – Hacker News	37,799,159	text
Websites – a web crawl	18,988,131	screenshot + text

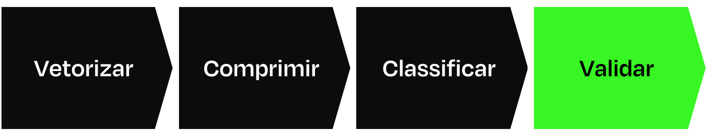
...and load it into ClickHouse - my favorite database.

Slide com as três bases do teste: fotos, comentários e sites. Foto: curadoria FFX no Web Summit 2026

> **O QUE OS EMBEDDINGS PODEM FAZER POR VOCÊ? É A PERGUNTA ERRADA. PERGUNTE-SE O QUE VOCÊ PODE FAZER PELOS EMBEDDINGS.**
>
> **Alexey Milovidov**, Co-founder & CTO da ClickHouse

A objeção mais forte vem do idioma: os textos do teste saíram do Hacker News, fórum em inglês. Um dado de fora do palco mostra o risco. No MTEB-BR, teste comparativo do Instituto Federal de São Paulo publicado em julho de 2026, um modelo que era o 3º colocado na classificação multilíngue caiu para 49º em tarefas originais em português. **Como gerar vetores ficou barato, testar vários modelos no próprio catálogo custa pouco perto do erro de escolher pela classificação global.**

Calcule uma vez, reaproveite sempre



Vetorizar	Comprimir	Classificar	Validar
Modelo de embeddings resume cada item uma única vez	4 bits por número preservam quase toda a busca	Uma regressão logística pequena para cada tarefa	Comparar modelos no próprio catálogo em português

A frase de Milovidov sobre a pergunta errada aponta para onde vai o trabalho. Com o cálculo barato, a IA já entrega a parte dela; falta a empresa montar a própria base, decidir quais tarefas vão usá-la e mantê-la atualizada. **Sai na frente quem trata esses vetores como matéria-prima de cada projeto novo, e não como experimento de um time só.**

LEITURA FFX

Para as empresas brasileiras, os vetores do catálogo e das conversas com clientes passam a ser um ativo, com dono e orçamento, como o cadastro de produtos.

A Fast Shop triplicou a taxa de conversão ao adotar a busca semântica da VTEX (Agência E-Plus, 2025); o Nubank seguiu o outro caminho e treinou um modelo próprio com as transações de mais de 100 milhões de clientes, que os times da empresa reaproveitam em tarefas diferentes (Nubank, 2025).

Varejistas, bancos e seguradoras escolhem entre contratar a busca pronta e montar a própria base, que custa mais no começo e menos a cada uso novo.

PROVOCAÇÃO FFX

SUA EMPRESA AINDA PERGUNTA O QUE A IA PODE FAZER POR ELA, OU JÁ SABE O QUE PRECISA FAZER PELOS PRÓPRIOS DADOS?

RETRATO DA EMPRESA

CLICKHOUSE

Banco de dados analítico de código aberto, com serviço em nuvem, criado por Alexey Milovidov em 2009 dentro da Yandex para consultas rápidas sobre volumes enormes de dados.

Avaliada em **US$ 15 bi** na Série D, ante US$ 6,35 bi em maio de 2025 (TechCrunch, jan/2026)

Série D de **US$ 400 mi**, liderada pela Dragoneer (ClickHouse, jan/2026)

Mais de **US$ 250 mi** de receita anualizada na nuvem, mais que o triplo de um ano antes (ClickHouse, mai/2026)

4.000 clientes, mais de mil novos desde janeiro (ClickHouse, mai/2026)

No Mercado Libre, 400 mi de registros de monitoramento por minuto em ClickHouse Cloud (ClickHouse, ago/2026)

FONTES

1. MTEB-BR: A Text Embedding Benchmark for Brazilian Portuguese (IFSP, 2026)
2. Agência E-Plus: IA na busca e recomendação, o futuro da VTEX (2025)
3. Nubank: Understanding Our Customers' Finances Through Foundation Models (2025)
4. Recall de codecs quantizados no Embeddings Explorer (ClickHouse, 2026)
5. TechCrunch: Snowflake, Databricks challenger ClickHouse hits $15B valuation
6. ClickHouse raises $400M Series D led by Dragoneer
7. ClickHouse tops $250M ARR and 4,000 customers
8. Mercado Libre observability on ClickHouse Cloud