**Smarter RAG Starts with Metadata (Beginner's Guide)**

**Transcrito por** [**TurboScribe.ai**](https://turboscribe.ai/pt/?ref=docx_export_upsell)**.** [**Atualize para Ilimitado**](https://turboscribe.ai/pt/subscribed?ref=docx_export_upsell) **para remover esta mensagem.**

Hoje, eu vou falar sobre as diferentes maneiras de usar metadata para fazer seus agentes RAG muito mais inteligentes. Não só a metadata nos permite organizar e filtrar nossos dados, mas também nos enriquece, para que possamos ter muito mais contexto sobre as partes que estamos olhando. Se ainda não faz muito sentido para você, não se preocupe, vamos desmaiá-lo tudo.

Vou passar por alguns exemplos ao vivo e, no final, você definitivamente estará pensando em como você pode aplicar a metadata para seu próprio uso. Ok, então não quero perder tempo. Vamos começar com uma demo rápida, para que eu possa mostrar o que essa coisa de metadata parece, falando com o nosso agente RAG do transcripto do YouTube.

Então, eu vou enviar essa mensagem que diz qual é a diferença entre um databases relacionados e um databases vectores. Nosso agente vai procurar pelo Superbase, ele vai usar um re-ranker para escolher esses resultados, e agora nós vamos obter nossa resposta baseado em um transcripto do YouTube. E você pode ver, não só o nosso agente conseguiu procurar por uma resposta para nós, ele também nos deu o mesmo vídeo do YouTube que ele retirou esse dado, o mesmo tempo que ele retirou o dado, e então, nós também podemos ir aqui e clicar no link se nós queremos assistir o vídeo inteiro.

E a única razão pela qual ele nos deu esse contexto extra é porque nós enriquecemos os pedaços com essa informação na metadata. Então, esse é o caso de uso que nós vamos abordar nesse vídeo, mas metadata, essencialmente, só significa dados sobre dados. Então, o que você está carregando, você poderia enriquecê-lo com metadata, como o dia, ou o departamento, ou o autor, ou o nome do arquivo, ou o tamanho do arquivo, o que você quiser.

Antes de eu descer essa pipelina onde nós estamos recebendo o dado do transcripto, preparando-o, standardizando-o, e depois enriquecendo-o com metadata para vectorizar-o, vamos subir para uma pequena visualização de ExcalDraw para que eu possa explicar rapidamente por que a metadata é importante e como funciona com a busca por vetores. Se você está seguindo o meu canal por um tempo agora, você provavelmente já viu essa visualização exata de ExcalDraw, mas, de qualquer forma, vamos passar por isso rapidamente. E eu também vou manter esse exemplo relevante para o que nós estaremos passando através do end-to-end, só para manter tudo consistente.

Então, o que começamos com aqui é o nosso transcripto. E então, o que acontece é que nós temos que trancá-lo e passá-lo através de um modelo de embedimentos para que ele seja transformado em diferentes vetores. E cada um desses pontos, ou pontos de vetores, é basicamente associado com um pedaço do transcripto.

Então, o transcripto pode ter 20 pedaços, 30 pedaços, 50 pedaços, dependendo de quão longo o vídeo é. E então, o problema aqui é que, sem metadata, quando nós estamos olhando para esse pedaço específico, nós na verdade não sabemos qual vídeo do YouTube ele veio de. Então, se nós colocarmos três transcriptos de vídeo do YouTube em nossa database de vetores, transcripto A, transcripto B e transcripto C, com metadata, nós podemos dar a cada pedaço mais informação, como o título do vídeo inteiro que ele veio de, ou a URL do vídeo inteiro que ele veio de, ou o timestamp de esse pedaço específico no transcripto. E sem esse tipo de metadata, nós não teríamos ideia do tipo de insights que nós estaríamos recebendo.

E, honestamente, nós não teríamos tanta confiança quando nosso agente responde para nós, porque ele não pode recuperá-lo com uma fonte real. E então, uma coisa a manter em mente é que a metadata é apenas dados sobre dados. Então, a metadata não tem efeito no sentido real do pedaço.

Ela não tem efeito no lugar onde o pedaço fica colocado na database de vetores. Então, quando nós estamos procurando pela nossa database de vetores, o código fica embedido, ele fica colocado na database de vetores, e então o pedaço mais próximo, ou o pedaço mais próximo, como você for, é puxado de volta, apenas baseado no sentido real do transcripto. Mas então, uma vez que nós pegamos o transcripto de volta e nós percebemos que ele é relevante para o código, então nós vamos puxar a metadata e olhá-lo para mais contexto.

E então, é claro, você pode fazer coisas como filtrar metadata, onde você pode dizer, ok, eu tenho transcripto A, B e C no meu database de vetores, mas para esse código específico, eu só quero olhar para transcripto C, porque eu sei que esse é o vídeo que eu quero algumas informações de. Então, os três grandes benefícios de ter metadata para sua database de vetores, RAG, é que você pode ter mais contexto quando você está puxando coisas de volta, você pode manter sua data um pouco mais organizada e segmentada, e então, é claro, você pode filtrar usando metadata para receber apenas o que você quer de volta. Então, agora que isso está fora do caminho, vamos subir para o end-to-end e vamos dar uma olhada nessas diferentes pipelines e como a metadata está realmente funcionando.

Ok, então nós só vimos a funcionalidade de conversa, onde estamos falando com a Superbase. Agora, vamos olhar a pipeline de transcriptos de como nós estamos recebendo dados para o nosso supermercado de vetores de Superbase com metadata. Então, antes de a gente lançar essa pipeline, deixe-me, em breve, entrar na Superbase para mostrar para vocês o transcripto que está atualmente aqui.

Este é este vídeo do YouTube que vocês viram na demonstração, mas, como vocês podem ver, em cada um destes diferentes pedaços, nós temos o conteúdo, que é apenas texto, e então, no metadata de cada um, nós temos o timestampo de onde este pedaço veio, nós temos o título do vídeo, e então nós temos a URL do vídeo. Então, é exatamente o que acontece com cada vídeo do YouTube que nós processamos para o nosso metadata. Ok, então eu tenho isso preparado para, basicamente, ativar uma submissão de formas se nós queremos colocar um vídeo no nosso metadata.

Então, eu vou clicar em Execute Workflow, vai puxar essa forma, que nos promete a colocar um título de vídeo e uma URL. Então, eu vou colocar este vídeo do Anthropic, onde eles estão sentados no sofá e estão falando de dicas para construir agentes AIs efetivos. O que acontece primeiro é que nós vamos no Appify para esmagar o transcripto do YouTube.

Nós vamos juntar o transcripto completo do YouTube, nós vamos pegar os timestamps, unir tudo junto, e então isso é o que fica vetorizado. E então, também nos atualiza no nosso site do Google, então nós podemos entrar aqui rapidamente e ver quais vídeos existem atualmente no nosso metadata. E agora, vamos descer em cada um destes nodos para que vocês possam ver o que está acontecendo.

O primeiro é o nosso pedido HTTP para o Appify para esmagar o transcripto. É um pedido bastante simples, nós basicamente estamos só dando-lhe a URL do vídeo, que veio aqui, da submissão de formas. E então, o que nos dá é este tipo de item estranho, onde nós temos, tipo, centenas de objetos aqui, e cada objeto tem um tempo de começo, uma duração, e então, texto.

E então, é realmente separado e não exatamente do jeito que nós o queríamos. E, por acaso, se vocês acessarem este template e vocês querem se arrumar, tudo o que vocês têm de fazer é ir para o Appify e colocar sua própria chave API do Appify aqui. Quando vocês chegarem ao Appify, façam certeza de usar o código 30NATEHURK para 30% de desconto por três meses, e então vocês vão descer aqui para os settings, vão clicar em API e integrações, e então aqui é onde vocês vão copiar sua chave API pessoal, e aí você só tem que escrever, como eu disse, aqui.

Apenas façam certeza de manter essa chave privada porque é tipo o seu passaporte. Então, a partir daí, o que eu decidi fazer foi, eu queria limpar e pegar uma linha do contexto inteiro. Então, eu decidi usar um código para isso.

Agora, o que eu fiz aqui neste código, só para mostrar para vocês o meu processo de pensamento, eu venho aqui e eu percebo, ok, eu tenho este JSON vindo e eu quero usar código para limpar. O que eu vou fazer é, eu vou copiar um pouco deste JSON aqui, só para que eu possa entender o esquema, e então eu vou para o Cloud e eu digo, me ajudem a escrever um código no end-to-end que receberá este JSON vindo. Eu copio o esquema e então eu digo, eu preciso separar todo o texto como uma linha gigante porque estamos processando um transcript.

Então, eu quero que todos os textos flutuem juntos como uma linha em um item. Ele espalha o código, eu o copio, eu tento, e o que vocês sabem, nós recebemos um item, que é o transcripto inteiro em uma linha. Então, é exatamente o que nós queríamos.

Mas, não é exatamente a informação que eu queria vectorizar e cortar, porque, como vimos na visualização, nós não sabíamos os tempos de cada área que foi cortada. Então, o que eu decidi fazer foi outro código, e, desta vez, eu queria separá-lo, mas mantendo os tempos relativos a cada pedaço de transcripto. Então, eu fiz a mesma coisa aqui, eu copiei o esquema, eu fui para o Cloud e, basicamente, eu disse, o que eu quero fazer agora é juntar 20 objetos de dados por tempo, e eu quero que os tempos fiquem relevantes a cada pedaço.

Então, isso não pode fazer muito sentido, eu gostaria de vir aqui e mostrar o que eu quero dizer com isso. Então, o node de transcripto ejecta objetos diferentes chamados de dados, e cada objeto de dados tem um tempo de começo, uma duração, e, então, o texto nesse pequeno objeto. Então, o que eu estou, basicamente, dizendo ao Cloud fazer é que eu quero juntar 20 objetos de dados por tempo, e, então, quando eu junto esses objetos de dados, eu quero que você coloque e combina todos os fios de texto juntos, mas, para manter o tempo relevante, eu quero obter o primeiro tempo de começo desse primeiro objeto de dados, e, então, do último, o que eu quero fazer é obter o tempo de começo mais a duração, o que seria igual ao tempo final.

E, então, você pode ver neste lado esquerdo, nós obtemos 25 pedaços diferentes de transcripto, e cada pedaço tem um tempo de começo formado e um tempo final formado. Então, é por isso que nós estamos obtendo nossos bons pedaços para vectorizar que têm nossos verdadeiros timestamps com eles. Então, isso funcionou maravilhosamente, eu estou agora juntando tudo e, então, nós o colocamos em Superbase, e a única coisa que é especial acontecendo aqui é no default data loader.

É aqui que nós estamos dizendo qual data vectorizar, e também é aqui que nós estamos dando os três fios de metadata, título de vídeo, timestamp e URL de vídeo. Então, a primeira coisa é o dado para vectorizar. Eu não queria fazer todos os dados de input, nós queríamos loadar dados muito específicos, o que seria apenas o texto que realmente foi dividido com timestamps.

Nós não queríamos vectorizar o inteiro texto combinado. Então, para cada um desses pedaços, nós queríamos dar o metadata. Então, para o título de vídeo, eu só fui para a submissão de forma e coloquei o título de vídeo.

Para o URL de vídeo, eu fiz a mesma coisa daqui, coloquei o URL de vídeo. Mas então, para o timestamp, o que eu tinha que fazer foi ir para o nodo Merge, onde nós colocamos tudo juntos, e aí eu queria colocar o tempo de começo formado, e então eu coloquei um dash manualmente, e aí eu coloquei o tempo de fim formado. Então, agora nós estamos recebendo uma espécie de 40 segundos para cada um dos nossos pedaços.

E agora que nós temos tudo no nosso database de vetores e enriquecido, nós estamos escrevendo isso no nosso site do Google, para que nós possamos ver novamente aqui os vídeos que existem no nosso database de vetores. E, por acaso, se vocês querem ter uma melhor sensação e acessar esse workflow, vocês podem fazer isso grátis. Vocês só têm que se juntar à minha comunidade de escola grátis.

O link para isso estará na descrição. Vocês vão então procurar o título deste vídeo, ou vocês podem clicar em Recursos do YouTube, e então o post associado a este vídeo terá o workflow aqui para acessar. Também haverá um link aqui para este template Google Sheet se você quiser importar o mesmo exato, para que você não tenha que setar nada mais.

Ok, legal. Então, agora o que nós podemos fazer é conversar com nosso agente e ver se ele pode acessar esse novo vídeo. Então, ele respondeu.

O time de Antropox compartilha várias insígnias sobre construir agentes efetivamente. Este é o primeiro dicas deste vídeo, do timestamp 12 para 42 com o link do vídeo do YouTube. Você pode ver que temos mais insígnias aqui embaixo, e esta é de muito mais tarde neste vídeo, então de 1745 para 1816.

E então, é claro, isso nos dá mais alguns pontos, mas você pode ver que está pulando de um lado para o outro. Ok, então agora que vocês viram como isso funciona, antes de chegarmos para o Pipeline aqui embaixo para automaticamente eliminar insígnias do nosso Vector Database, eu queria falar sobre, agora que nós temos dois vídeos do YouTube no nosso Vector Database, o que se tivéssemos que fazer uma busca e dizer que eu só quero que você olhe por este único vídeo? Bem, este é o lugar onde nós usaríamos filtro de metadatas, porque nós podemos literalmente dizer busque pelo nosso Vector Database para este, mas apenas pule as regras em que o título do vídeo é igual a este. Então, o que nós podemos fazer para isso? Eu tenho um set-up diferente aqui, que é o nosso agente com uma submissão de forma para fazer o filtro de metadatas.

Então, nós falamos com este através de uma submissão de forma, eu vou abrir a forma aqui, e ele vai nos pedir um vídeo do YouTube para olhar através, e então nós podemos dar-lhe uma query. Então, eu estou dizendo que eu só quero olhar através do vídeo de Tips para Criar Agentes de Anthropic, e eu só quero três passagens claves. Então, eu vou enviar isso, nosso agente vai pegar isso, procurar pelo Vector Database apenas por partes, onde o título do vídeo é igual a o que nós colocamos ali, e então ele vai nos dar uma resposta.

E também, este não vai responder no chat, porque nós falamos com ele através de uma submissão de forma, então eu só tenho que clicar no agente para ler a resposta. Então, aqui vamos nós, aqui são três passagens claves do vídeo Tips para Criar Agentes de Anthropic. A primeira é, quando você está criando agentes, é importante desenhar o seu produto para que, enquanto os modelos de AI melhorem, o seu produto melhora também.

A segunda passagem é sobre desenvolvedores, e ela vem de 517 nesse vídeo. E então a última passagem vem do mesmo vídeo, e do marco de 12 segundos. Então, lembre-se, isso não é perfeito, e se você realmente queria três passagens claves ou uma grande suma, você provavelmente não iria querer usar um método de pesquisa de vectores, porque os pedaços ainda não têm um contexto holístico de todos os pedaços combinados.

Mas o ponto que eu estou tentando fazer aqui é que é assim que você pode controlar quais pedaços você está procurando, porque nesse nodo de Superbase Vector Store, nós temos um filtro de metade. E é aqui que nós estamos dizendo apenas pule os pedaços de volta se o título do vídeo é igual a esse input dinâmico, que, no nosso caso, é o que nós colocamos na submissão do formato. Tudo bem, e a parte final é que nós temos esse pequeno pipeline para deletar coisas se nós não queremos eles no nosso database de vectores.

Então, digamos que estamos cansados de conversar com esse vídeo do YouTube do Nate, e nós queremos saí-lo do nosso database de vectores. Tudo o que nós faríamos é chegarmos aqui e mudarmos o status para remover, e então este fluxo de trabalho irá ligar se é um fluxo de trabalho ativo, porque nós atualizamos um pedaço. Isso vai garantir que só vai processar pedaços onde o status é igual a remover.

Então, deixe-me ligar isso e mostrar para vocês. E então, nós colocamos o fluxo. Apenas caso você vá lá e marque 5 como remover, ele irá ir em frente e processar todos os 5. Então, você pode ver que foi muito rápido.

Já terminou. E se eu for para a página do Google, você pode ver que isto agora está marcado como removido. E se eu for para o nosso database de vectores e refazermos esta coisa, devemos ver que este primeiro pedaço aqui não terá o título do YouTube de tudo que eu aprendi sobre agentes de construção, aquele vídeo que acabamos de deletar.

Então, agora nosso database de vectores está atualizado. Então, só para mostrar para vocês como isso funciona, o botão sai quando um pedaço está atualizado, mas ainda vai puxar tudo naquela folha. Então, o que nós fazemos é que nós vamos para um filtro e nós apenas dizemos que se a coluna status é removido, então nós vamos mantê-la.

E é por isso que ele manteve este vídeo, mas ele não manteve os dicas para construir agentes de ciência. Então, nós só vamos realmente remover o que nós marcamos como removido. Então, nós temos o fluxo.

Nós só temos um item neste caso, então não importa. Mas então, nós estamos fazendo um set, apenas para ter certeza que nós setamos a URL do vídeo, porque é isso que nós vamos usar como filtro de metadeira para remover os vetores. Então, nós temos a URL aqui do vídeo do YouTube, e então nós vamos para um nodo de super base onde nós vamos remover uma coluna.

Nós estamos removendo ela da tabela que nós tínhamos setada e então, basicamente, nós estamos dizendo apenas remover colunas onde na metadeira, o campo chamado vídeo URL é igual a este, que é o URL de vídeo do YouTube dinâmico que nós acabamos de setar. Então, ele volta para nós e diz, ok, nós encontramos 32 vetores onde o campo do vídeo URL o campo de metadeira é igual a esta coluna. Então, nós vamos apenas deletá-los todos.

Ele vai em frente e o deleta e então nós atualizamos na nossa folha do Google que esta URL foi marcada como removida. Então, esse é o sistema completo que eu queria compartilhar com vocês hoje. Se vocês acessarem este template gratuito, vocês poderão brincar com tudo isso e testá-lo e ver, sabe, como você poderia começar a usar metadeira para o seu caso específico porque sempre vai ser diferente para diferentes tipos de dados.

Mas a chave realmente é esta pipelinha de RAG aqui, onde vocês vão procurar diferentes tipos de dados, mas para cada dado vocês devem saber exatamente o que vocês querem. Então, neste caso, eu sabia que sempre que eu recebo um transcript, eu quero a coisa inteira, eu quero os timestamps e então eu vou saber exatamente que campos vão ser metadeiras. E então, uma vez que você tem essa pipelinha decidida onde você está tomando dados, você está preparando, você está standardizando, tudo é previsível, então você pode realmente começar a enriquecer todo esse dado em um database para retirá-lo.

E se você está procurando uma divulgação profunda em algumas das coisas e você está procurando mais experiência de mão a mão, então eu definitivamente recomendo conferir a minha comunidade pagada, o link para isso está também na descrição. Temos uma ótima comunidade de membros que estão sempre compartilhando e construindo com a Anadan todos os dias e nós temos dois cursos inteiros, um é chamado Agents Zero, que é a Fundação para Automação de A.I. E então nós temos 10 horas para 10 segundos, onde você aprende como identificar, desenhar e construir automações de tempo salvo. Então, eu adoraria ver vocês nessa comunidade, mas isso vai fazer para este vídeo.

Espero que vocês tenham gostado ou aprendam algo novo. Se fizer, eu realmente apreciaria um like, me ajuda um monte. E, como sempre, eu aprecio vocês fazendo isso até o final do vídeo.

Vejo vocês no próximo. Obrigado, pessoal.

**Transcrito por** [**TurboScribe.ai**](https://turboscribe.ai/pt/?ref=docx_export_upsell)**.** [**Atualize para Ilimitado**](https://turboscribe.ai/pt/subscribed?ref=docx_export_upsell) **para remover esta mensagem.**