**Ed Enuff:** 0:00

I can also feed it the facts at the time, rather than having it deal with its foggy memory. So if I can give it the set of facts, then I can eliminate the hallucinations as well. So those combination of things, it's basically if I can have the LLM either I can tell it your preference is the information that I'm supplying to you, and if you genuinely don't know, say you don't know. That combination ends up with very low hallucination output.

**Craig Smith:** 0:32

Hi, this episode is sponsored by Solonus, the global leader in process mining. Ai has landed and enterprises are adapting, giving customers slick experiences and the technology to deliver. The road feels long, but you're closer than you think. You see your business processes run through many systems, creating data at every step. Solonus reconstructs this data to generate process intelligence a common business language With process intelligence. Ai knows how your business flows across every department, every system and every process. With AI solutions powered by Solonus, enterprises get faster, more accurate insights, a new level of automation and a step change in productivity, performance and customer satisfaction. Process intelligence is the missing piece in the AI-enabled tech stack. Search Solonus C-E-L-O-N-I-S to find out more. Hi, I'm Craig Smith and this is I on AI. This week, I speak to Ed Enough, chief Product Officer at Datastacks, about vector databases and their role in AI applications. We explore how retrieving trusted data into large language models addresses the hallucination problem. Ed explains the growth and demand for vector databases and key factors like production costs and benchmarking relevancy. I hope you find the conversation as informative as I did.

**Ed Enuff:** 2:19

I have been over at Datastacks for the last close to four years now, a couple of months short of that. Prior to that, I was at Google for about four years and joined Google when it acquired Apigee, which was the API management company that I was part of for a number of years. I've worked in a variety of companies enterprise software, as well as things like blogging, software and a few consumer things as well. I've been in tech for a long time now, originally an RPI graduate, rensselaer Polytechnic. Institute way back in the day.

**Craig Smith:** 3:08

Yeah, yeah. What made the move to Datastacks? I'm curious. Vector databases in particular are all the rage right now because of the hallucination problem with LLMs, but what was Datastacks doing before the GPTs came along?

**Ed Enuff:** 3:33

Well, so we make the Cassandra database. It's a very popular open source database that is designed for scale out data, and Cassandra is the database that's used by Apple, by Netflix, by Federal Express, by any of a large number of companies that need to deal with scale out data that they're distributing to users on a global basis. So part of the attraction of coming to Datastacks was the fact that it was dealing with very large data sets and the goal was to make that data available for AI and machine learning use cases, and that was something that we were already well underway on. A number of our customers were well underway on. Obviously, some of the users that I mentioned are very well known for their use of predictive AI. For example, uber as well also uses Cassandra as one of their primary databases, and they've published a lot of great research papers and done a lot of talks about how a lot of the predictive AI that powers the Uber application is built on top of Cassandra. So our goal was always to bring that to the customers as part of the product. And then the middle of that, generative AI hit and obviously has really sent everything into Overdrive.

**Craig Smith:** 5:06

Yeah, Cassandra. What kind of a database is that?

**Ed Enuff:** 5:10

So it's a NoSQL database. It is similar to databases like Mongo and CouchBase or DynamoDB or some of these others, which means that it, although it does let you use a query language like SQL standard query language, which is what most relational databases use, but it's more designed for situations where you're dealing with large amounts of data and the data schema can change or can be very flexible at run time. So you find a lot of the. These types of databases became very popular for the large internet services. Google, Facebook, Amazon were really the pioneers of this type of database, and Cassandra itself spun out of a project at Facebook and the technology was open sourced by Facebook. So that's where you tend to see this type of technology is where people are dealing with large amounts of data that they're using in their interaction with their end users.

**Craig Smith:** 6:19

And the. I mean I've got a few questions. So the data stacks builds itself as a real time data company, a real time database service company, where I guess you build on prem as well. But is that exclusively for now that we're dealing with Gen AI for vector databases, how different are vector databases from Cassandra, and why this sudden shift that I'm seeing to vector databases from other forms of storage?

**Ed Enuff:** 7:03

So couple of different questions there. So first, you know we do talk about ourselves as a real time database. We do so in the context of AI as well, and the reason is because Cassandra is designed for very high throughputs of data. You can write to the database very quickly, but, more importantly, that data is immediately available for reading as well. And it turns out that becomes really important in Gen AI, and I'll talk about that in a minute. But along the way, what we saw, and basically starting last year and going into this year, is that a new application pattern emerged, and it was inspired by chat GPT.

**Craig Smith:** 7:49

And it's.

**Ed Enuff:** 7:50

I think we've all experienced chat GPT, so it's useful to start with that because it helps explain why vector databases are important when you show up a chat GPT. Part of what makes chat GPT interesting is not just the question and then you get a response, it's actually the chat part of it. So you ask a question, it gives you a response. You ask it for some elaboration, or perhaps you're having it help write you some code. You'll go and say, oh, that isn't exactly right, could you make it do this? And so it has a sense of history. And if you know about how LLMs large language models work, they're completely stateless. The model itself. When you ask the model a question, it doesn't change. That model is frozen. It can later be tuned or retrained. But the model itself, the data, comes in. It actually can be represented from a software development set, but it's a very simple black box. I have an input which is often called the context or the prompt, which is a form of context, and then it has a response. Where do things like the history come from? Where do those actually come from? A database. So ChatGPT has a database that sits next to GPT form and it knows the history. It captures the response and it's able to create a personalized conversation with you, and that's where the magic happened. Then OpenAI, the company behind ChatGPT, did a blog post in December of last year, because at the time they weren't fully in the ChatGPT business. They actually had launched us an example. What they were really trying to do was get people to use their models. They wanted you to use GPT-3, they wanted you to use GPT-4. So they said if you like this thing that we've built and you wanted to build your own, what you need is a database that sits well with the large language model. Oh yeah, part of what helps that database sit with the large language model is if it is able to store and query vectors. Is the vector is the numerical representation of essentially the concept. What it does is it takes that whole string of text and it reduces it down to a very well, I say reduce, but in some cases actually it can be much larger A very long, multi-dimensional number that represents what that word or phrase or sentence or paragraph reduces it down into a context or concept, concept, semantic concept, and that's how LLMs think. They don't really literally think, but they represent what we consider ideas or concepts. They represent it in a large multi-dimensional space, and it's the distance between things in that space that says how similar or different they are. So this is an important capability that lets a database work well with LLM. Now there are some other uses for vectors. Some databases had been supporting vectors for other reasons basically for doing better search results for a while, and some of those were among the first crop of databases that people started using with LLMs. All the other database vendors very quickly followed suit. By this time we're almost a year since a couple of months short of a year since opening. I wrote that blog post that kicked off this whole thing. At this point, most major databases have some vector facility because it becomes really important. There's lots of people who are building these types of applications in enterprises and startups and so on. They need a database that can work with these models. That's why you're seeing everybody talking about vector databases. Sorry, that was a little bit of a long response.

**Craig Smith:** 11:58

No, that's wonderful. That really clarifies it quite a bit For data stacks at this point. When did you introduce vector databases and is your business 60, 40 with traditional databases and now vector databases? Where is the future of the database business going from your point of view?

**Ed Enuff:** 12:28

So really good question. So we introduced the vector capability. Since we're built on top of open source, we were able to get the code out there very quickly early in the year. We had it live in our Cloud service in April of this year. We're at a point right now where 40 percent of new signups people coming to our Cloud service are doing so for vector usage. Actually let me amend that, actually it's actually more than half are using it for vector usage. About 40 percent actually have never used Cassandra before, so they're completely new users who came to the service because they were looking for vector databases. So this is pretty important for us. Now I suspect that some of the other database companies are seeing similar stuff. I'd love to say, oh, we're uniquely suited. I do think we've got some great important capabilities that others don't have, but I also do. I think that the bigger point is that there's a major catalyst this year with people trying to build these types of applications, and it's causing them to go and rethink what databases they're looking at and evaluate different databases, and so on. So long and short of it is. That's why, if you're following the database industry, the entire conversation is around vector databases, because new application types don't come out that often. The last time we had a new application type of this level was probably mobile. That catalyzed a whole bunch of stuff. Then prior to that was probably e-commerce and websites and the web itself. You start to get that hyperbole as to the, but I've been around for a while. I'm sure you've been following this stuff for a while. We all have seen this thing where there's a new application type and companies rush to respond to it and it really changes everything. And within the database space, again probably the web, and then, prior to that, probably client server being the biggest catalyst. So very important to us as long and short of it.

**Craig Smith:** 15:05

Yeah, can we back up just a little bit because I want to talk about some of the products, but I have questions about data that I've never satisfied myself with answering. I was close for a while to a company called Labelbox. That is a platform for labeling data and it fascinated me during the supervised learning, when that was the dominant technology in AI and in reading and understand. My background is not as a technologist. I was a foreign correspondent covering politics and conflicts and stuff like that, so the amount of data that's being produced is increasing, it seems, exponentially. I don't have the numbers in front of me and so when you talk about real time data, it's writing and reading data in real time, but presumably some of that data then gets stored for long term usage or retrieval that it's not being retrieved frequently, but you don't want to lose it. And where does all this data go eventually? Because it's not being erased and at the time that I was looking at this, there's kind of a hierarchy where it gets moved from one medium to another until it's finally on literally tape drives stored in a mineshaft somewhere. But do you understand that? Where does this data go and is it being accumulated to the point that someday there's going to be more data than any other artifact in the world.

**Ed Enuff:** 17:26

It probably already is the longest short of it is? It somewhat depends on the use case and it really ends up. There's two pieces to it. There's sort of cost and policy and depending on the systems and depending on the. So, first of all, you're correct in your basic statement, particularly when we talk about real time. You're capturing data, you are perhaps doing analysis over ranges of it, you're keeping it around and you are accumulating it, and a lot of companies do accumulate and that becomes how long do I keep it for? In some cases, depending on what you're doing, the cost of getting rid of it might be less than the cost of keeping it. The cost of storing data has dropped over the years. What ended up happening was a lot of companies, particularly in the internet, he's just kept on accumulating that data. Because, again, querying and deleting a range of data is actually, from a computation standpoint, for the process of doing that It'd be just as expensive as writing that data, so it could be much more actually, and then the cost of storing the data can be very low. So then you see a lot of companies just going to say, okay, we're just going to keep this data, just because it's not like in as much as a business decision. It's more just like okay, the cost of doing something will get around to it, and maybe they do it on certain basis. Then you have the policy questions. The policy questions come from either business or regulatory. Then you have the questions of in some cases, some policies require you to retain the data for a period of time. Other policies go and say you shouldn't retain it. Many people on corporate email systems and so on, things are auto-deleted after 30 or 60 days as a company policy, because if that's your standard policy, then the data isn't there, and so on. It may not be desirable to keep it for various reasons, so that becomes a whole decision-pointed in and of itself. Then there's security reasons as to why you don't want to keep data around for a long period of time as well. So, for example, companies that are very concerned with privacy actually make it an advertiser. They advertise that they don't retain any personal data whatsoever. So there's a whole set of things around that. So the long and short of it ends up being that you probably could look at a hundred different companies with their data retention strategies and find a hundred different answers, because a lot of it will have to do with the quantity that's being accumulated and the cost associated with it. The companies and the ones that we talk about, for example, like financial statements and things like that that only get updated and aggregate on a monthly basis. That's why you can get, for example, a bank statement. They increasingly made it. It becomes harder to do because they shift it to what's essentially called cold storage or less accessible forms of data, but they still keep it around for 20 years in some cases if you need to get something. Whereas the internet companies that are capturing and literally storing a row in the database every time you load a page view, because they're doing it for personalization and ad tracking purposes those they will collapse and compress and create summaries and delete the rest. The interesting thing is that those summaries end up getting created through ML. That actually ends up being an important usage of machine learning is to go and condense those 10,000 clicks into something that they can then just store and they can delete the records of the previous 10,000 clicks. It is a whole thing in and of itself. I don't think there's a one size fits all strategy on that. It ends up, as I said, ends up being with how much data you're dealing with and how quickly it's changing.

**Craig Smith:** 22:18

Yeah, with data stacks. Is there a drop-off then that a customer can set that it's reading and writing, but after a certain period of time it goes into some storage bin that's a little more difficult to access. How does that work?

**Ed Enuff:** 22:42

Well, you definitely can do that in a variety mechanisms too, but it actually gives me an opportunity to talk about one of the cool features, which is we do actually have the ability for at the individual row level in the database. It's pretty unique to Cassandra that you can make the data expire as a consequence. You have lots of users of Cassandra that take advantage of that. They're like okay, I only care about this data for a month or whatever. Then after that, I don't want to have to go through the problem of crawling through all my data and checking each row and saying, oh, is it older than this amount of time? Let me delete it or whatever. Rather than that, the database just automatically does it. It just drops it off of the database. Then you have this dataset that always contains your last 30 days or whatever the period of time that you want it to be. That's why that becomes important. Is that again, where you get into these situations is where you have these very large sets of data. It is one of the areas where we play and where Cassandra is uniquely suited is for the extremely large datasets.

**Craig Smith:** 23:58

Yeah, my familiarity with vector databases comes from addressing the hallucination problem with LLMs. A lot of companies are building a vector database. Even after they tune in LLM, they still have this hallucination problem. They build a vector database with trusted data. The LLM becomes simply a language interface. Is that how data stacks customers are using your vector database?

**Ed Enuff:** 24:34

Yeah, that is a description of a process called retrieval, augmented generation, or we often hear it called RAG. If you're at a conference, an AI conference or something, everybody will be like, oh, RAG, RAG, RAG, and you're like, what is this? It is this process of unless the name implies, retrieval, augmented generation goes and says that the generation is that's the thing would apply being augmented by retrieval from data sources that are fed into the LLM at the point of inference. In fact, what you can do is tell the LLM that it should only consider the knowledge that is supplied to it at the point of inference. In some cases, it's just to supplement it. But if you really want to eliminate hallucinations, you just go and say, look, use your reasoning powers, but don't use your knowledge. The reason is because one of the aspects of hallucinations there's a lot of cause of hallucinations, but one of them is that the LLMs have a foggy memory, if you will, just like humans, although perhaps a different mechanism, Although not that different a mechanism, but that would be an entirely different conversation. So, rather than it going and trying to remember exactly, you're like here's the set of relevant content, choose among them and choose among these things and give me an output. You do that for two reasons. As you said, it's trusted, but there's also an other important piece of that which might be that it's also sensitive. So, for example, you could have an LLM that is providing you with medical information, medical recommendation. You don't want the language model fine-tuned on a set of electronic medical records. The LLM is not good with private information. Generally speaking, anything that goes into the LLM it's going to leak out, either inadvertently or inadvertently. There's no way that you put access control on the knowledge that an LLM has. You could do something convoluted, which I see some people say is all going to have a second LLM spilled the beans and filter it out. But now you're getting to this Rube Goldberg architectures. It's much better to just supply the LLM with that sensitive information as it needs it. Remember, an LLM has no memory in and of itself. So I can say here's Ed asking a healthcare question and here's his electronic medical record. And the LLM looks at that, goes through and says well, ed, you probably need to exercise more because your weight has gone up over the last six months. But it doesn't remember that fact and then later be in a conversation with Craig and say hey, craig, you want to exercise more because you don't want to end up like Ed. I don't want the LLM to know anything about Ed when he's talking to Craig and I don't want it to know anything about Craig when he's talking to Ed. So that's another piece of it. Now the other aspect of it is, again, I can also feed it the facts at the time, rather than having it deal with its foggy memory. So if I can give it the set of facts, then I can eliminate the hallucinations as well. So those combination of things it's basically, if I can have the LLM either, I can tell it your preference is the information that I'm supplying to you. If you genuinely don't know, say you don't know that combination ends up with a very low hallucination output. A lot of these things are more art than science. Which is frustrating to many is the building these systems. You can make it more exact. Again, we've done this. We do this for we have an AI co-pilot that helps you use our products. What we've done is we supply content out of our documentation into it and we tell it you exclusively use that, and then it gives a result that doesn't have any hallucinations. Now we also do a little bit of fine tuning, a squall, but if you really want to eliminate the hallucinations, it's a good mechanism.

**Craig Smith:** 29:24

Right now I know you have a product called AstraDB. What is your primary product and what are the use cases that people are turning to it for?

**Ed Enuff:** 29:41

Sure. So we have a cloud product called AstroDB and we also have a self-managed software that people can run themselves called Datastax Enterprise. They're both built on top of the Cassandra open source database. Our fastest growing product is the cloud product that one can imagine. We do have many, many enterprises actually the majority of the Fortune 500 is using Datastax Enterprise, running it in their own data center. Both of those groups are very, very interested in the vector database capabilities. I talk to customers every day that are doing things with their sensitive data that they need to keep in their own data center. Most people who are new to our product and our company do come to the cloud service. They come to AstroDB, they do a sign up and they're just right within a couple of minutes. There they are able to create a database and connect it to their applications. The type of applications that people build are all over the place in terms of applications that are powering mobile apps, applications that are powering websites, applications that are being used for things that are controlling internet of things, devices. As I started out by saying at this point, for brand new people coming to the service, the number one catalyst is generative AI. Maybe that'll even out or slow down over time. But I do think that just I think it just has to do with the number of new applications people are building in general that a lot of that focuses AI related right now, yeah, does the suite of products include the vectorization of data?

**Craig Smith:** 31:59

because in order to put it in a vector database, there need to be vectors.

**Ed Enuff:** 32:06

That's a really good question. So what we do is we actually do have the capability within the products for being able to use open source models. The selection of the model that you use for the vectorization is actually a really important decision. Some people want to use open AI, some people want to use Google's models, some people want to go and use open source models. The different models all have different costs associated with it and, as a consequence, you see that this selection process, people will use a smaller model for vectorization because it's much cheaper. So we enable all of that. As I said, we do provide an open source model where that makes sense, but majority of people are using something like open AI or they're using one of the Google models. Those tend to be the preferences.

**Craig Smith:** 33:11

Yeah, but on your platform you have those available, or does someone have to Well?

**Ed Enuff:** 33:16

yeah, I mean, we let you integrate your account. You can go and put in your open AI account.

**Craig Smith:** 33:23

Yeah, and I've been writing recently about the cost of inference and rate limits on large language models, which are constraints for enterprises. Does using a vector database lower the cost of inference and does it reduce the number of tokens that you're putting into or pulling out of an LLM because you've got the data vectorized already someplace else?

**Ed Enuff:** 34:04

Yes, so it depends on what it is that you're trying to do. If you're doing a simple search use case, the vector database can be a very good substitute for being able to go and take a natural language query, meaning a sentence that somebody's going to ask and give you the result. But it'll be a search result, meaning that if you're asking a question about, for example, what is the statue somewhere in the center of my town, it's going to give me a standard answer. But what's going to do is it's going to understand the question, it's going to turn it into a vector and then it's going to give me a result. That'll give me the top result, or how many results I want, but generally it'll give me the top result. It tends to be the way people build these things out that matches that question. So that gives me a little bit of the AI experience, because I'm able to ask this human question, but then I'm getting a standard response. If I want that response to be customized, the vector database can't write for you. What the vector database can do is it can read for you, but writing that response, then I need the language model, and so this ends up being one of these things that people look at for knowledge bases, for example, where they're like I don't really need a GPT personalized response to every query, I just want to give that result. Somebody goes and says to me, comes in and asks a question how do I, whatever change channels on my TV set, I just want to get that stock answer that says you know, that gives me the page from the reference manual on the remote control, right, but I want to handle the person asking that question one of a hundred different ways, including in different languages. The vector, the vectorization, the vector query process handles that very well, and so so, yes, as a low cost option you see that you see that sort of thing quite a bit but but it really depends on, like I said, how much of, how much of are they trying to do, of having sort of a full conversational AI experience. If you need that, then you're going to invoke the model at least once through every interaction.

**Craig Smith:** 36:42

Yeah, um is. So I would imagine that this business is growing leaps and bounds. You were saying 40% of all new signups are for use with an LLM. What? What kind of growth are we talking about?

**Ed Enuff:** 37:03

Well. So keep in mind, as we talk about these things, usage growth is is is pretty dramatic, yeah, but, but the interesting thing is that, um, a lot of what people are doing right now is experimentation. So this is one of these interesting things, as we talk about vectors and vector database and business growth and obviously we're running a commercial business. So I have this type of conversation all the time with investors and such and who are looking across the database industry and they go and say how soon are we going to actually see one of these public database companies announcing their results and see huge growth? And I said, well, we're probably another six months away from that, because what we're seeing right now is a lot of experiments. But we are all what are. All. Every database company is what's called a consumption business. If people aren't actually using it live on their website and their business processes or whatever, they're not going to be consuming more. They're not going to be consuming more database software. They're not going to be consuming cloud services. They're not going to be consuming open AI. That will all come from this stuff going into production. So what I would suggest for people as you're trying to look and understand this stuff is sort of look at the world around you as you go into this holiday season. You're going to see we're about a month away from Black Friday, which is the Friday after Thanksgiving largest retail day of the year, and I genuinely don't know the answer to this. We're working with a number of those retailers and many of them are working at breakneck pace to try to get their stuff live. But, as I said, something that sort of everybody who's listening to this can do themselves is sort of pay attention to that. When you go to, when you start doing your shopping, is there an agent on the? Do you see a conversational agent on the website saying, hey, what can I help you with today that lets you do a chat conversation and is suggesting products to you by this time next year, very highly likely, you're going to see those on the majority of websites. Just the same way that when we saw the mobile transition right, remember that you know you saw oh, it took about 18 months for, but it was a steady progression Everybody was like, hey, try out our app, and you'd started to do it. And now we're at a point right now where, if you're like me, the majority of stuff that you buy from the store down the street. You're going and looking at it like Home Depot which, by the way, as a data stacks customer, runs everything in their e-commerce site, uses our Cassandra database. But if I'm going to go and get something from the Home Depot, I've got some DIY, I've got to fix something and I'm trying to figure out which, first of all, which store is closest to me and which one has that thing in inventory and what aisle and row that it's in. You know, all of that I do through the mobile app, right, and so we're going to. A lot of us think, and with a lot of evidence, that we're seeing a similar transition to AI, and so what that means is, over the next few months, you'll see a set of companies that do this and the rest will the majority of them, in fact, because we have what's called the crossing the chasm. You've got the early companies that do everything first and then the rest, but you'll see a few of them race out over the next couple of months with these types of AI services and the expectation is again over the next year you'll be. Every website you log into, every mobile app. When you open it up, you'll see various forms of this sort of personal ized AI experience. Some of it will be classic chat, some of it will be a little bit more. That's one of the things I often get people asking me like does it all have to be via chat? The answer is no. It actually can be done in a variety of different user experience modalities. But you start to see these things and you'll be like, oh, that's AI powered and I think that's going to be the. You know, we're going to go through this process where it's pretty novel right now, but then it'll just become a standard thing. Keep in mind, the majority of the AI hype that we have experienced this year has been experiments. You definitely you know, I suspect if you're like me majority of the interactions you do, majority of things you buy online and so on. Right now, you're not seeing Gen AI in the loop on them, and that will change. That'll change pretty dramatically and that's when you're going to see the business growth. We're already seeing the usage growth, as I told you right now in terms of signups. But what are people doing once they sign up? They're building their demos, and the demos don't use a lot. They neither consume a lot nor are they paying for anything, yet because they haven't launched these services.

**Craig Smith:** 42:18

Yeah, well, actually, that's why I've been interested in rate limits in particular, because, for listeners that don't know what rate limits are, the LLM companies limit the number of tokens you can use per hour, let's say, and that limits how you can scale. And so there's been a lot of experimentation, not a lot of enterprise I shouldn't say not a lot, but it's mostly experimentation. And the question is will this rate limit, which is directly related to the availability of GPUs, which we all know are all sold out, how do you increase? How many of these experiments will succeed in going into production? Do you have any sense of that? I know it's off topic from what we're talking about.

**Ed Enuff:** 43:19

No, it's actually. It's very on topic and I get that question quite a bit because, again these questions a lot of folks are trying to figure out what is the size of the vector database market. So I have variations of this conversation all the time with analysts, and both industry analysts like Gardner and Forster, and financial analysts who are trying to make recommendations of which database stocks to buy and all that. And so what we know is this we know that the cost of for a similar amount of data and a similar amount of retrieval and it doesn't matter which database you're using, because we benchmark them all there is a significant amount of compute used in doing a vector retrieval. It is almost 10 times the amount of compute that a regular database query would do, and so that's an increase in cost. Some of that compute is GPU-based compute, but there's a lot of work being done to shift that to minimize the amount of GPU compute that's necessary, and in fact our systems can do all of the. In fact they do all of the vector retrieval without GPU compute. So that's not true of all the vector databases, but majority of them are moving in that direction. We actually get a lot more requests in the opposite direction, which is, if we want this to be even faster, can we use GPU compute, even though it is, it's scarce, but so the compute is expensive and so doing GenAI has an expense. And what we then go and say is okay, as we look at that, everybody's doing these experiments, but probably some subset probably anywhere from one in three to one in 10, actually have the business use case that justifies it. And when they do, it's not going to be a 10X cost, it has to be a two to five X cost. So so, where we look at what we're doing, and again, when I say 10X cost, we, we, we benchmarked everybody and we averaged it across across what, every, every, you name it you can throw names at me all day long and we did, we, we worked it into our benchmarks. And, because we needed to know this, where we are a database data scale company, so so all of our customers are very focused on cost. So what I think you're going to see is that you know every vector database company is going to be talking about and and commenting and innovating around costs and performance. For example, we, we created a piece of technology called J vector that is built on top of Java, because that's the language we use. That dramatically cuts the cost, and I know that you'll see similar approaches or announcements from other companies because they have to. So we have to get costs down. But even with that, you also have to have the use cases that actually drive a performance business outcome Right. So going back to that's often why I use some of the retail examples, because they will. They are the the retail industry tends to be and the variations of it won't literally be people selling product, might be people selling travel, but they have the outcome measurements where they're able to go and say oh, we delivered this AI recommendation service and we converted, we got this much more business. The companies that don't have that are going to struggle because AI is so nice to have, no matter how cool and innovative it seems, it has a cost involved and we're seeing this already. You know people come and say to me oh, I use chat, gpt and it was slower or the answer wasn't as good. Did they change something? And my response is I don't really know what goes on behind the scenes there, but I say look, what they've said has been that they haven't changed their model at all. So what I think that they're doing is they're doing things to control costs, right, and, and some of the things you do to control cost, what people perceive as quality and relevance from the model is not always a property of the model. It can be a property of the vector database you're using. It can be a property of the of the type of compute you're using. As similar thing, you go and see that that. You know Microsoft recently with their some of their very popular co-pilot services, it. You know it's costing them more to provide it than then then what they charge users for, and so they'll also be. So all of these, these companies that you know, the big, the big tech companies that have been rushing a AI out. They can do that and they can operate at a loss. Most enterprises won't do that. They will. They may, they might operate on an individual request. They'll lose money, but what they'll have to be able to go and say is our overall sales increased by x percent by using this model and that increase of In, that increase of revenue that we got, has to be greater than the amount we spent. Right, very similar, very simple lemonade stand economics, but but that's the way enterprises, you know, need to operate or they're gonna get beaten up on on Wall Street, right? So so you know we're gonna see that, and this is what we call the production filter, and we talk about this a lot, which is Everybody's looking at gen AI and this is the year year of experiments, and it should be like people should be getting out there Learning what the stuff is good for. That's the only way you can figure it out, but the production filter will be a Much smaller set of those will be the ones that you see live on on the websites or or live in in whatever Format makes sense for the business. A lot of this stuff is stuff that won't actually be presented directly to the user. It might be something that your customer support agent that you're talking to they're using a gen AI based system that's giving them Stuff to tell you right, or or or how to process a claim, or something of that sort. So so, yeah, that that the cost will be one of the biggest gating factors. Yeah, probably second to cost will be hallucinations, but, but it'll be cost by a long shot.

**Craig Smith:** 49:47

Yeah, and and to be fair too, were we're only a year in the, the GPT 3, gpt 4 Era, and and a lot of the you know everyone's working on Reducing costs or increasing tokens. Yeah.

**Ed Enuff:** 50:10

Exactly have to look. I'm old enough to remember, you know, the web 1.0 days when, when you know, startups were racking and stacking Sun micro system servers and Cisco routers and all of that and and you know as much as people sort of complain about the, you know, irrational exuberance of the tech industry and and all of this stuff. The reality is that we've all seen these situations where and it's baked into the institutional memory of most these companies that that that you can only get, you know, so far ahead on your ski is in terms of of you know, of this stuff, and some of the big and the big tech companies can can afford to operate at a loss as they do these things. But even when they do it, they're doing it in very careful, measured ways To to make sure that that that it's something that they have a path to sustainability and and you know, and profitability around. And all of the enterprises that I interact with very few enterprises that are like that are they're going and saying, look, we don't care how much it costs, we just want to get this thing out there there. Everyone of them is going and saying you know, okay, what is this going to look like? A production and, frankly, actually this is one of the ways that we're able to actually gauge. As we're Talking to one of these companies, our way of gauging how close are they to put it into production Ends up being is the cost conversation happening if they're not talking about the cost at all? What we know, is, they're still just basically experimenting, which is okay, but but obviously it's a business You're always trying to figure out how far along is, or are these folks you know in terms of of of where they are as customers? And so one of our check offs, checkboxes as we're going through this is like okay, are they starting to ask the hard questions about what does production costs look like? And then and we want to get them there Because we know that they're not going to go live until they go through that process yeah, obviously, I think you know we as a company have very good answers for it for that stage of it, but we know that, that they have to be asking those questions, right, and so yeah, yeah, well, I guess I'm going to go through that one more time.

**Craig Smith:** 52:21

I think Well, I, you know, I've this conversation. I've been kind of following my, my own Interest, which is not necessarily the most sophisticated Train of of thought what, what have I not talked about? That that you'd like people to know? About data stacks?

**Ed Enuff:** 52:42

Well, I think you know I've dropped a few pieces into it. We we do feel you know Data stacks has been the company that has served most of the companies in, you know, in the fortune 500, on their digital transformation journeys, meaning when they went to mobile, they and they had the scale that they had to deal with. You know data stacks and the Cassandra database. Where was the database that was your production, reliable production at scale and At a cost basis. We've been the company that powers that and the brands have draw have name dropped in the car in the course of conversation. You know, whether using somebody like price line, whether using somebody like, every time you scan a package or or or, for that matter, every time you track your package on federal express, every time they scan the package on the 20 different or a hundred different points on its journey. All of that is is is going through data stacks as databases. Every time that you you know you use Netflix, that's using the Cassandra database, the open source Cassandra database. Apple is uses Cassandra as well. So so Cassandra and data stacks have played a really important role in making this data available at scale and We've done all of the work to make it not just the best for those types database use cases, but for vector database use cases. And so I think you've asked all the right questions, which are are what happens when you put the stuff into production? What does it end up costing? You know what? What are the challenges around dealing with these large amounts of data and preserving the relevancy around it? And, and as people are thinking about this stuff, you know I always recommend, you know, make sure you try everything. But the one thing that isn't happening enough in this conversation, in this a? I conversation, is the experiments are great, but but Asking those questions about what happens when I go into production and when you look at the vector database as you pointed out at the start of this, there's every database company seems to be talking about vector databases Make sure that you are thinking about what are, what is the cost, the reliability, the accuracy, you know the performance of this stuff into production. Otherwise, you're gonna have a really cool experiment that you know looks really good and you know maybe you presented it to your board of directors and everybody's like ooh, ah, but it'll just be. You know It'll be just the AI equivalent of the concept cars that you know that Detroit used to wheel out Right. It won't be actually something that you're actually able to get out there and have the impact on your business.

**Craig Smith:** 55:20

Yeah, just on on that last point. Are there benchmarks that people should be looking at If, when they're evaluating databases?

**Ed Enuff:** 55:31

There are there are standard Benchmarks for relevancy and recall, and you should also be asking you know you should be looking at, as you look at, the databases. We're publishing these now and I'm seeing other people are too, but it's not widespread yet. So so I would definitely ask you, know a database vendor, database provider, whether it's open source database you're looking at or whether it's one from a vendor. I would go and ask to see the benchmarks, and just Not just from a performance standpoint. Some of these benchmarks are also related to relevancy. There are standard tests that are done to evaluate the relevancy. Not all vector databases are created equal in terms of of the relevancy of the results that they get, and that is something that you will end up Discovering as you build this stuff out yourself. Most common thing that we see is somebody loads in a bunch of data and, by the way, relevancy also changes With the number of records in the database. So any of the vector databases when you're doing that little test where you load in a couple hundred or even a thousand you know date entries, then you go and see you're like, wow, this is really good. And then you go and load in a hundred thousand or more, and then you start to see this drop-off and at that point you're in real-bind, and so we, you know, we think that's going to be important, certainly, as, as we spend the next 18 months with people putting these projects into production, you're going to hear a lot more about this and and it already is a topic that you see on blog posts and Articles and people write, like you see these how to improve your route, your, the relevancy, of your vector results. You're going to see a lot of that. It's going to be the next hot button issue. Hi, this episode is sponsored by Salonis, the global leader in process mining.

**Craig Smith:** 57:26

AI has landed and enterprises are adapting, giving customers slick experiences and the technology to deliver. The road feels long, but you're closer than you think. You see your business processes run through many systems creating data at every step. Salonis reconstructs this data to generate process intelligence, a common business language with process intelligence. Ai knows how your business flows across every department, every system in every process. With AI solutions powered by Salonis, enterprises get faster, more accurate insights, a new level of automation and a step change in productivity, performance and customer satisfaction. Process intelligence is the missing piece in the AI enabled tech stack. Search Salonis C E L O N I S To find out more. That's it for this episode. I want to thank Ed for his time. If you want to read a transcript of today's conversation, you can find one on our website. I on AI. That's E Y, e hyphen on dot AI. In the meantime, remember the singularity may not be the same. Remember the singularity may not be near, but AI is already changing your world, so best pay attention.