The Distributed Data Show Podcast Episode #134

The (near) Future of Databases with Jonathan Ellis - TRANSCRIPT

Release Date: February 4, 2020

Jonathan Ellis: What I think we can do with machine learning is not only match the job that humans can do, but do better.

David Gilardi: Do better, yah.

Jonathan Ellis: Because when you have a problem space that large, even the best human administrator is only going to be looking at a dozen or so of those configurations and say, that's what I'm going to focus my effort on.

Speaker 3: From DataStax, this is The Distributed Data Show podcast.

David Gilardi: Well, hello everybody and welcome to another episode of The Distributed Data Show. I'm David Gilardi, here with Jonathan Ellis, founder of DataStax and knower of everything Cassandra and just everything. Honestly, I'm just going to say you know everything. So, hi there, Jonathan.

Jonathan Ellis: Hi David. Thanks for having me on this show.

David Gilardi: So you gave a pretty cool talk today at Data Day Texas where you gave your predictions in the database sphere, the datasphere, right? Five predictions. Five different classes of predictions of what you think is going to happen in the next five years, right? So I kind of want to talk about that and kind of see, you know.

Jonathan Ellis: Yeah. It's January, 2020, starting a new decade and I wanted to, like you said, talk about what's coming up next. I think the next 10 years, like that's too far out for me to have much of an idea, but for the next five years, I think I've got a pretty good idea of some of the things we're going to see.

David Gilardi: Okay. So the first section you had was machine learning, right?

Jonathan Ellis: Yeah.

David Gilardi: What's going on with machine learning? Where are we going with this?

Jonathan Ellis: I think where we're going is you're going to see databases become a lot more self-driving. And in particular around self-optimizing, like there's hundreds of tuneable configuration parameters, whether you're looking at

Cassandra, PostgreSQL, it's really complex systems. And not only, what I think we can do with machine learning is not only match the job that humans can do, but do better.

David Gilardi: Do better, yeah.

Jonathan Ellis: Because when you have a problem space that's that large, even the best human administrator is only going to be looking at a dozen or so of those configurations and say, that's what I'm going to focus my effort on. When you have an AI in control, it can be infinitely patient and really dial in the best possible performance for your specific workloads.

David Gilardi: And to that point, when I watched your presentation, I knew Cassandra already had a ton of configuration parameters, like way too many honestly. I had no idea that MySQL was at like 550, that's insane. So how can anyone ever manage all that?

Jonathan Ellis: It's ridiculous. I don't think you can.

David Gilardi: The prediction then being is that we're going to see more AI management of the systems, right? Are there any systems that exist today that are already starting or doing something like this?

Jonathan Ellis: In some of the areas, kind of the old guard relational databases are doing better than the NoSQL open source databases. So, Microsoft SQL Server has an index optimization wizard for instance. So they've been putting some effort into some of these areas already. In terms of the actual configuration optimization that we started out talking about, I think Andy Pavlo's research group at CMU is at the forefront of research.

David Gilardi: I remember seeing him do a talk on, was it last year?

Jonathan Ellis: We actually invited him to DataStax last year-

David Gilardi: Okay. So he was at Accelerate.

Jonathan Ellis: ... can talk about that.

David Gilardi: Okay. Okay. We've got now more AI automatic control of the configuration parameters and the tuning of your databases and such. That's in the machine learning area. The next section was cloud. What do you think is coming in the cloud arena?

Jonathan Ellis: I think what you're seeing is that, you see this today, so this isn't a prediction, this is a description of the world today. You've got two kinds of

cloud databases. You've got kind of the, I'm going to manage a legacy database, if you will, for you. We're going to automate that as much as possible. And this is what Amazon's done with RDS. That's kind of the state of the art there. They've done a really good job with that.

Jonathan Ellis: But because of the design of that legacy database, there's a limit to how much you can automate. It's designed around this single tenant philosophy. I'm going to be as full-featured as possible in the query language I'm giving you. It's going to be able to handle everything from ingesting time series to really sophisticated analytical queries. And so, what you end up with is there is some administration overhead where humans need to be involved as part of owning that RDS instance you're signing up for.

Jonathan Ellis: And so, the other category is a more native cloud database where they, to get that administration down to zero, we're giving up some of the query functionality and we're optimizing for something that's simpler and faster. And by doing that, I can give you a DynamoDB or a DataStax Apollo where it really is your administration and it's much more of a black box and you don't have to care about going in and spending a lot of time optimizing your index and your query planning because it is a simpler model.

David Gilardi: Right. So in a case like that, thinking about the Apollo standpoint or something, the way that we look at that is, from a development standpoint, for this fully managed databases, why do people do that? Well, they either don't have the IT, don't have the expertise, they don't want to spend the time administrating these systems and such. And from the developer experience, a lot of times, if I'm coding a new app, I don't want to have to worry about tuning all of this database and data layer and that kind of deal. So I can just spin this thing up and go and make the developer experience really easy.

David Gilardi: So are you saying that like separate from like Apollo, Dynamo, Cosmos, they all do that, right? They have a much simpler model that allows people to just go, databases handled and move on. So are we seeing or thinking that that really is where things are pushing and where they're going to go is more and more to these fully managed platforms.

Jonathan Ellis: Yeah, it really is like the platform as a service where give me an API, I'll call into the API and the service provider takes care of the plumbing.

David Gilardi: Okay. Okay. And then, the next session was open source. That one gets fun, especially with some of the things that have been going on in the open source world in the last year or two. Where do you see us going or where's the open source world going at least in your predictions?

Jonathan Ellis: I think that there's a very strong wind blowing that we saw in the 90s and the 2000s. Open source was a mega trend, that's the thought leaders term, right?

David Gilardi: Yeah.

Jonathan Ellis: It's a mega trend and a lot of people got a lot of funding to build infrastructure companies around open source products. And the reason driving that, it wasn't the original no free software foundation. I'm open sourcing it because it's the only ethical choice. I'm open sourcing it because I'm pragmatic and open source is a better way to get adoption. And what I see happening is, if your open source project is successful enough, then Amazon is going to deploy it and compete with.

David Gilardi: It's totally true.

Jonathan Ellis: They will. That's what they do. They look at what workloads people are running on the infrastructure they rent from Amazon, and if there's enough money to be made, they'll totally go and stand up an elastic search service, or a Kafka service or a Cassandra service.

Jonathan Ellis: And so, I think that coupled with what we just talked about, as people want infrastructure as a service, there is a new way to get adoption for your infrastructure startup and that is by providing a free level of service that it's not open source, but nobody cares because you're running it for them. So you've addressed the adoption pain point in a different way, but a way that's actually even more effective in 2020.

David Gilardi: Okay. So, it sounds like what I'm hearing is then we will see more adoption by more of the open source capabilities or whatever the actual projects are getting these managed services with a free tier that essentially lock in the adoption that way, but then don't have to in fact open source their work. Is that kind of what we're seeing there?

Jonathan Ellis: Exactly.

David Gilardi: Okay. Okay. So then let's move into hardware. You talked about some pretty cool stuff in the hardware section, like where do you see things going there? What will we see in the next five years?

Jonathan Ellis: I think that the cloud is going to drive the conversation or database hardware. If you can't get it from Amazon or Google or Microsoft, then you're not going to deploy it in your own data centers either because you want to have a coherent data layer across your hybrid deployments.

David Gilardi: Do you feel like there's going to be even more of a decrease on people going with their own on-prem installations and essentially pushing even more into the cloud?

Jonathan Ellis: I do think that's true, but that's not actually the point I was trying to make, which was that even if I am running my own hardware, I'm going to want to standardize on the kinds of stuff that I can get from the cloud vendors because that allows me to have a common denominator there.

David Gilardi: I see.

Jonathan Ellis: And so my point was, I don't think you're going to have either specialized compute in the form of FPGA and ASIC or specialized storage in the form of Intel Optane DIMMs for instance. I think those are both going to be pretty niche. I did put a caveat on this one, which is, this is the point on which the smart people that I know are most likely to disagree with me on. I still think it's true.

David Gilardi: Okay. All right, and then the last section, it's just one that's like near and dear to my heart, which is graph, you had some interesting things to say about graph.

Jonathan Ellis: My headline for graph was I think graph is a feature and not a database. And so, the reason is that there's a bunch of things that graph databases do well, they're good at tree retrieval, they're going at recursive relationships and pattern matching and pathfinding, reachability queries. But then there's a lot of stuff that they don't do well like user activity logging and sensor or IOT data, feature vectors for machine learning.

Jonathan Ellis: So, I don't think developers are well-served by having to do those pieces of the application in different systems sort of different databases. So I think what you're going to see is enterprising database vendors are going to pull graph features into a non-graph substrate and give you the best of both worlds. I think that's been a long time coming because the kind of the incumbent database vendors, they're all relational databases built on B-trees, which is a shocking, it's like the worst possible storage engine to implement graph on.

Jonathan Ellis: But now you have kind of that next generation of scale-out databases like Cassandra. Cassandra is built on log structured merge trees, which are actually really good at representing graph adjacency lists. So now you have vendors like DataStax being able to bring the graph features into the Cassandra layer and I think that's the future.

David Gilardi: So instead of then the folks out there that are digesting or ingesting this stuff, again, applications hooking up and having to now use multiple disparate systems to get different capabilities, have graph over here, have my data store, my database over here, what I think I'm hearing there is that we're going to see more of a marrying of those, where those graph, those basic graph capabilities are going to be provided through that one system along with whatever their data storage mechanism is. And that way, it'll simplify essentially their whole experience and kind of increase the capabilities they get from that one database.

Jonathan Ellis: Exactly.

David Gilardi: Okay. Very cool. We covered all the sections of your predictions. Not that it was in your presentation, but any Jonathan Ellis predictions you have coming for the next five years? I'm just totally putting you on the spot here.

Jonathan Ellis: I think I made enough friends with this set here. Let's check back in five years.

David Gilardi: Yeah, right. We'll come back in another five years and take a look. So, thank you Jonathan for giving us a recap of your talk and thank you everybody for listening to another episode of the Distributed Data Show.

Jonathan Ellis: Thanks David.

Speaker 3: Looking forward to seeing you at DataStax Accelerate, where we'll provide a whole set of community talks selected by community leaders for everyone in the open source Cassandra world. And don't forget to register in either San Diego or London by going to www.datastax.com/accelerate.