The Distributed Data Show Podcast Episode #138:

Apache Cassandra™ Open Source Strategy with Sam Ramji - TRANSCRIPT

Release Date: March 3, 2020

Sam Ramji: We made a few substantial strategic errors a few years ago when we started moving away from the community. You never want to be in a position where you're competing with your community. The community by definition is where all of your opportunities will come from.

Speaker 2: From DataStax, this is the Distributed Data Show podcast.

Jeff Carpenter: On this week's episode of the Distributed Data Show, I have Sam Ramji with me. This is Jeff Carpenter and Sam Ramji and a welcome to the show, Sam.

Sam Ramji: Thank you very much, Jeff. I appreciate it.

Jeff Carpenter: You've recently joined up with DataStax and some previous stops, you were at a Cloud Foundry where I know a lot of people know you from, you were at Google, Autodesk and that's just the past few years and you don't have to give me the whole resume here, but maybe give me a little bit of highlights of the past few years and how you ended up here in the Cassandra community.

Sam Ramji: Yeah, well my family will tell you I'm really not good at anything except software. I started programming when I was nine years old. That was back in 1980. And 94 I finished studying neuroscience and artificial intelligence and entered into the second AI winter. Nobody was hiring anybody with AI. I became a software engineer working on educational software. One thing led to another, I ended up being the director of engineering for Ofoto, which some of you might know as the 2000/2001 version of Instagram. 10 years earlier and worth about $950 million less than Instagram.

Jeff Carpenter: Timing is everything, man.

Sam Ramji: It is everything. I got a chance to work in some pretty cool problems, distributed systems. Before that and then with BEA I worked on a bunch of distributed systems and integration. Then I worked at Microsoft and I was literally the open source and Linux guy at Microsoft. I ended up with a worldwide team and responsible for open source technology and marketing strategy. Got to work with Bill Gates, with Ray Ozzie, with other people to make the world safe from Microsoft in open source and then to get Microsoft turned around so it could participate well with open source.

And now kind of looking back just a few years, some of the things that my team was trying to get done in the late two thousands now look like a good idea. You might notice that Microsoft launched SQL server for Linux. Something we were talking about back in 2008.

Jeff Carpenter: Oh wow.

Sam Ramji: Became a reality in 2010 so yeah, good times. And I got to work at Apigee after that, which is where kind of the story with DataStax really begins, building out an API infrastructure. We acquired a company called Usergrid, which was led by a gentleman named Ed Anuff, which was an app server for Cassandra. We thought, wow, scale out information, scale out data. That's what APIs want. They're really hungry. They want an extremely high performance backend. Usergrid in front of Cassandra was the perfect thing.

Jeff Carpenter: Oh, got it. And so Ed's now here at DataStax as well.

Sam Ramji: That's right.

Jeff Carpenter: See how some of this community is forming here. Okay, keep going.

Sam Ramji: Yeah, that's exactly it. And then you take an opportunity that comes up in 2019 and Chet's talking with DataStax and I'm talking with DataStax and Ed's talking with DataStax and the three of us are talking with each other. We're like, man, there's something awesome to be able to do here with scale out data. That's kind of what brought me in. Just an opportunity to work with great people on a fantastic technology and figure out what the next decade looks like.

Jeff Carpenter: Awesome. Was there anything, what was it about Cassandra that was particularly attractive to you?

Sam Ramji: It solves an embarrassingly difficult problem. Distributed databases are really, really hard. First of all, the database does all the hard work for applications. For application developers to write simple code, the database has to handle all of these very difficult edge cases. Second, distributed systems and distributed data is even harder. And it takes a long, long time for those systems to harden. You kind of have to get it wrong and less wrong and less wrong and then find this other thing that fails over and oh my gosh, you got to really take all of the failure modes out of the software and that takes a decade. And so no surprise, here we are, 12 years after Cassandra was invented, 10 years since it was promoted to a top level Apache project and it's pretty darn good now. But that's how long it takes to solve those problems.

Jeff Carpenter: Right. And I've heard that 10 year mark cited in the context of just for a database engine, which is really, if you think about it, only half of Cassandra. There's also a whole distributed system built on top of it. Yeah, I could see where you actually have two levels of technology there that needs to kind of come together and mature over that 10 year period. And I would agree with you that they have.

Sam Ramji: Totally, right.

Jeff Carpenter: Cool.

Sam Ramji: It's so interesting to think about plugability there too. Storage engines are a point of view and not every point of view is right for all situations. You always end up in computer science with problems that are really, really thorny. And then you can solve maybe 80% of them, but for another 20% you might have three or four different solutions. You look at what's happened recently with AWS, they've launched managed Cassandra services, they've taken the Cassandra front end and they've taken the Instagram fork of Cassandra, which was designed to modify the storage engine so you could replace it with RocksDB. And they've taken that fork, they forked the fork, and now it talks to the private Amazon native infrastructure, which is kind of the same thing the Dynamo and other services talk to. This idea of having it to fork to get a storage engine is kind of interesting. Way more interesting to say, "What if we had pluggable storage engines as part of the Cassandra architecture? What would that mean to the future of distributed data?" And that gets me very excited.

Jeff Carpenter: Super cool. Yeah. I think that there are a number of cool opportunities that we're going to start see emerging based on the advent of having a Cassandra enhancement process that the community has recently ratified. That was pretty cool. We were at, Patrick and I were hanging out at Apache Con and Scott from Apple kind of put forward a proposal like, hey, this is really something that we should have a process for in the Cassandra community. And there was instantaneous agreement and adoption of that. And not too many conversations on the email list later, it was a thing.

Sam Ramji: That's great. And Apache's such a great community. I got a chance to start working with them back in 2006 when I was working at Microsoft. And one of the things that we realized was a lot of people wanted to run WAMP, Windows, Apache, My SQL and PHP.

Jeff Carpenter: I have not heard of the WAMP stack.

Sam Ramji: No, there was, there was a WAMP stack, it turned out. We started looking into it on a technical level and we're like, none of this is tuned and nobody's ever construed on anything. That gave me the first opportunity to start working with folks from the Apache technology community and I was just blown away by it. The licensing model, the way that the PMCs work, the way that the overall leadership rotates constantly, got opportunities to work with some super interesting people over time like Jim Cegielski in particular, Brian Behlendorf and just a great stalwart of the internet. We wouldn't have the internet with it without the Apache software foundation and we wouldn't have a lot of the commercial, private public partnerships that we've gotten as a result of the Apache software license 2.0. Is in fact the last meeting I ever had with Bill Gates, this is June 28, 2008 was one where we established green lighting of a particular set of open source licenses and the Apache software license 2.0 was one of the key ones that we approved. It's pretty cool.

Jeff Carpenter: Cool. Yeah. Since we're on the topic of open source and different communities, I actually, I wanted to ask you, because I know that you've had some involvement in leadership in Cloud Native Computing Foundation and so I know that there's been a number of different open source communities that you've been involved in over the years, but I'm interested in particular about CNCF, how would you compare and contrast or what are some things that we can learn from? You can take this question where whichever way you want, but what can we learn from Cloud Native Computing Foundation and its approach? Because it's kind of a novel approach, over these past couple years.

Sam Ramji: Yeah. It's super interesting. I had gotten a bunch of experience from the Cloud Foundry Foundation, where I was the founding CEO and we were trying to figure out how do we build a collaborative stance across all of these competing commercial organizations? We had HP and EMC and Pivotal and SAP and IBM, just a whole bunch of different companies trying to kind of come at each other and win in the market. But the point that I made was, we want to do this in a way that harms the user, keep the technology stable while competing. We took a very particular tack on that where we kind of brought the technical and the business operations together.

Sam Ramji: With CNCF, which was started by Craig McLuckie and a few others, the point of view was quite different, which said, let's let the technical team be supreme in their own domain and let's have no overlaps between that and the business team. The board of directors for the CNCF basically doesn't get to say anything about the technology. That's all managed.

Jeff Carpenter: That's interesting.

Sam Ramji: It's pretty cool. It's all managed by the TOC because what they wanted to do was to say, code speaks louder than words and if you're contributing, that's how you should do your governance. How do you do governance by contribution? And I think that's ended up being pretty clean. One of the big challenges that we had early on that I was Google's board member for the CNCF while I was there, actually joined Google to run the Kubernetes business in late, late 2016 was how do we avoid picking winners?

Sam Ramji: And I think this is one of those tricky things where somebody's going to really, really dislike you because you didn't pick a winner to make it easy for them to say, "If you want to do fu use bar." On the other hand, we wanted to be clear, we did not think that we were smarter than the market. What we wanted to do really was specify interfaces. How do we standardize on a service interface? How do we standardize on storage or a networking interface? And then how do we let the market play out and have people have different points of view about how to do those things? And I really think believing that the market is smarter is a good approach for any organization to take, particularly one that appears to be in the role of being a standard setting organization like the CNCF so easily maybe.

Jeff Carpenter: And that's I agree with that and that is the right philosophy. It seems like it's hard in practice because company X is always wanting to slip that little something into the API because they need just that one little hook that'll enable them to get that killer differentiating feature out there.

Sam Ramji: So well said.

Jeff Carpenter: And then company Y doesn't want that in there for the same converse reasons. And this is hard stuff.

Sam Ramji: You're totally right. Which is why I think, we're so far ahead of where we were in the early two thousands. When I was at BEA and we were part of the Java community process.

Jeff Carpenter: Oh my goodness, application servers.

Sam Ramji: Coming with all these other, oh my god.

Jeff Carpenter: UTEE, APIs.

Sam Ramji: Right. And so you effectively have this slow moving war between these different big organizations because BEA wants these things in the

standard and IBM once these others, and sometimes you feel like, well those folks are ahead of us here on the implementation so we're going to send somebody really smart to the committee and that person's only job is going to be to slow the committee down with really good technical complaints. Oh my God, like that's not good for the user.

Jeff Carpenter: There's a whole politics, bad politics that can be associated here.

Sam Ramji: It's a mess. On the one hand, how do you have a better standards organization? On the other hand, how do you embrace open source? And I think it's really letting open source be de facto standard. Letting code speak louder than words and letting the market choose. People are not going to be fooled into picking up something that doesn't work properly or that is one off or isn't reproducible by somebody else that doesn't have the right intellectual property guarantees that says, "Hey, just copy this and do whatever you want." I think that creates a lot more room for innovation, a lot more speed of movement and a lot more safety for users.

Jeff Carpenter: Excellent. Well, so I think people in all communities, CNCF, Apache, whatever community that you're in would agree with those things. We've seen some changes in terms of the DataStax posture, let's say. And not to be overly dramatic or anything, but people that are observers have probably noticed some things that we've started doing a little bit different lately. Just a couple of examples would be releasing the new versions of the drivers, single version of the driver that has additional features that we were previously saving only for working with DataStax Enterprise, releasing our a Kafka connector for Cassandra and making sure that that works with not only DSE but also open source. Really last week re released DS Bench. More of these tools that we're releasing are just as compatible with open source Cassandra as they are with DSE. And I don't know, maybe you want to speak to a little bit of what they're thinking there? What's behind all this?

Sam Ramji: Yeah. There's a silly T-shirt that you'll see on university campuses every so often, which will show you the speed of light and they'll say it's not just a good idea, it's the law. I think open source is kind of proved in that same way. What is the sound of a million developers collaborating? That sound is open source. If you want to create something really useful, if you want to create a really big market, it doesn't matter if you're coming at it with a commercial standpoint or if you're looking at it from a moving the industry forward through practice standpoint, more things in open source is better. I think we made a few substantial strategic errors a few years ago when we started moving away from the community. You never want to be in a position where you're competing with your community. The community by definition is where all of your opportunity, most of the good ideas will

come from and so you want to be in a stance where you're constantly nurturing the community and vice versa.

Sam Ramji: That means taking a really generous stand with competitors as well. And there's a great quote that Ed Anuff uses actually, which is, "It's important to avoid the narcissism of small differences." Said another way is, whatever it is that makes us look like we're fighting, the fact that we're able to fight about it means that we're a heck of a lot more similar to each other than we are to anybody else in the world. Let's figure out, how do we do this in a way that's collaborative because everybody could benefit. Maybe Sill has got a different point of view on how you should do the storage engine. Awesome. Maybe Instaclustr has got a different point of view on how you should, stand up, scale out infrastructure against Kubernetes. Awesome.

Sam Ramji: How do we do this in a way where we can all over a multi-month, multiyear process, move everybody forward? And I think that's by recognizing that the community's at the core, we've got to work effectively with the open source project, with the constellation of projects around open source and sort of bit by bit have many opinions start to congeal so that we can say, yep. And together we solve that problem in 2018, we solve the other problem in 2020 and we solve that problem in 2022 and now standing here in 2024, we're figuring this stuff out. It's about looking at this over the arc of history that lets us feel a little bit less combative and more collaborative.

Jeff Carpenter: Very nice. Very well put it. I'm going to make a topic change here in a second.

Jeff Carpenter: Okay. I'm just going to hit the pause button right here. I had such a great conversation with Sam that we blew past our regular episode time and just kept rolling, so we decided to make this into two episodes. The episode you just heard, we talked a lot about open source community and culture. Next week we're going to get Sam's thoughts on why open source projects should have multiple speed lanes to promote faster innovation. We're also going to talk about the future of databases and the potential of artificial intelligence and machine learning to improve database operations and why this is a moral imperative. We're also going to hear why the 2020s are going to be the decade of scale out data. See you next week.

Speaker 2: Thanks for listening to the Distributed Data Show. Please subscribe with your favorite podcast app, give us a rating and don't forget to share with your colleagues.