Radhika (00:05):

Welcome to Code Together, a discussion series exploring the possibilities of cross architecture development with those who live it. I'm your host, Radhika Sarin. This is a special episode. I'm excited to introduce our next host and moderator for our Code Together podcast, Tony Mongkolsmai. Tony is a software architect and technical evangelist at Intel. He joined Intel in 2003 and since then he has worked on several software developer tools. Most recently he led the software engineering team that built the Data Center platform which enabled Habana's scalable MLPerf solution. Tony, I'm so excited to welcome you. And thank you for joining us today.

Tony (00:57):

Thanks, Radhika, it's really great to join you guys as well. I'm really looking forward to talking about a lot of cool technology things as we go forward. With that, we'll start with our first one today. Today, we'll be talking about how AI continues to play an important role in applications and decision-making, and also the challenges that developers and data scientists continue to face as AI workloads expand. In today's podcast, we'll talk to our guests about how Intel and Red Hat's partnership is tackling these problems. We'll learn about how Intel's popular AI developer optimizations on Red Hat OpenShift Data Science are creating opportunities for developers. We're lucky enough today to have Audrey Reznik. Audrey joins us from Red Hat, and it's a pleasure to have you, Audrey.

Audrey (01:41):

It's great to be here. Hey, everybody.

Tony (01:43):

Audrey is a senior principal software engineer in Red Hat Cloud Services. She assists Red Hat OpenShift Data Science team, focusing on helping customers with managed services, AI/ML workloads, and next generation platforms. She's very passionate about data scientists and, in particular, the current opportunities with AI/ML at the edge and open source technologies. We also have today from Intel, Rachel Oberman, she is an AI software solutions engineer who helps customers optimize their workflows with data analytics and artificial intelligence optimizations. She holds a bachelor's degree in computer science and data science. And she's currently pursuing her master's in computer science with a focus on machine learning at Columbia University. It's wonderful to have you today, Rachel.

Rachel (02:27):

Thank you. It's great to be here.

Tony (02:29):

So, to start off, Audrey, can you talk a little bit more about this joint project between Intel and Red Hat and what problems you're trying to solve?

Audrey (02:36):

Yeah, for sure. Let me back up a bit and just talk about the Red Hat OpenShift Data Science platform. That's a platform that we put together to help data scientists, data engineers, developers, MLOps engineers, to basically build or develop, train and deploy models. That's all in the context of what you would do within a model life cycle. Now, when you have those various parts to an actual platform, and

you can imagine the platform will have a lot of underlying infrastructure, as Red Hat, we don't want to go ahead and develop a lot of the services on our own. We feel that there are fantastic solutions out there for such things as, say, optimizing workflows. That's where we have partnered with Intel, because Intel has a lot of services such as the Intel AI Kit, which, again, will help to develop and train models along with OpenVINO, which goes ahead and helps with the optimization.

(03:36):

This is really important when you're looking at such a platform because you are not only giving a solution from Red Hat, but you're making use of a large partner ecosystem. Again, Intel is a part of that ecosystem, which we're very happy to have on board as a partner. One thing that really is exciting is that they help us with some of the problems that customers have in that not only do they have to optimize their workflows, but sometimes the hardware is very expensive, when talking about GPUs.

(04:09):

So, if you have a customer that is trying to get some work done and they're stymied, because the cost of the GPU is there, that's where it's absolutely fantastic to bring Rachel and her team along where they can talk to that customer about optimizing workloads and using some of the great tools that they've developed. So, really addressing the infrastructure needs without more infrastructure, which again, that's really going to help the customers out in terms of cost. And I would just almost pull it over to Rachel and get her to chime in on some of the challenges that Intel sees when they're building a solution to just some of the problems that I've talked about.

Rachel (04:56):

Yeah. Thanks, Audrey. A lot of great things that you've talked about so far and just to touch on a few of those, especially in that last part, about utilizing resources to the most. There's a lot of things that Intel is seeing in terms of AI challenges and infrastructure challenges that we want to try and solve to the best of our ability. The first part is making sure for AI developers and data scientists that the packages that they're using, that they're all working together, they all mesh well together. There's no problem with that.

(05:27):

The second part is, of course, as you mentioned towards the end, resources, making sure that you're utilizing your resources to the best of your ability, having enough memory, having good infrastructure, whether that be just CPU or just GPU and taking advantage of those just to the best of your ability. So not just having a GPU, but how can we utilize it to its best potential, same with the CPU.

(05:50):

Similar to that, when you're working with these resources, again, you want to look at what's the best to your ability in terms of how can you accelerate it, is there's something more I can be doing? Can I parallelize this? Can I do distributed processing to better improve my workload? Besides that, with all these resources and software and so on, a lot of developers do not want to be restricted in terms of their infrastructure, as well as in terms of their software and cloud providers.

(06:20):

For example, if you're using a particular type of software that may limit in terms of what other kind of software that you can use with that, or if you're using a particular cloud provider or dataset, maybe certain tools and infrastructure is not going to be available for you that may be for other use cases. So, how can we make this more flexible to our developers and to just our technologists as a whole? How

can we make this easy and less frustrating? Which is the most common problem, frustrating in terms of having to overhaul your code, overhaul your infrastructure, overhaul your resources, something that no one wants to do when they have this great product so far, and they're just trying to scale or they're trying to deploy.

(07:02):

So, I want to pass it back to you, Audrey, and focusing on these last two challenges I mentioned about the ease of use to avoid that frustration and avoiding any sort of restriction. Can you talk about what Red Hat is currently doing to help solve these problems?

Audrey (07:16):

Yeah, for sure. I have to say that this is a problem with a lot of folks that are just starting out as well, too. So, I can take an example where I was in a conference a week ago for IEEE, where we had a lot of electronics engineers trying to figure out how do they go ahead and share their solutions? It really brought home the point that they don't want something complicated. They really want something easy. They want to be able to have a platform that is very easy to use, but most importantly, they don't want to be locked in as well, too. And I think that's where Red Hat OpenShift Data Science comes in, because we can, with that platform, help you either on prem or in the cloud. In the cloud, right now, we run on AWS.

(08:05):

But I would say a real cause of frustration is just not only getting that infrastructure in place, but again, how do you work together? How do you work together with your data engineer, your data scientists and your MLOps engineers in one platform? That's a huge thing is, how do you share your code? How do you show people that you're working with that you have a problem? Again, having a platform where it's very easy to maneuver around and also easy to use different solutions, because you don't want to be boxed in by a vendor, is really, really important.

(08:37):

Again, especially for some of our customers that are smaller in nature, I'm going to bring up that hardware issue again, it is really costly sometimes if you are a small organization to think about GPUs. That's where bringing in a joint solution, particularly with Intel, is very, very beneficial. Again, for Red Hat, we don't have to worry about that so much because we let you or Intel worry about that.

(09:03):

What we have, again, is this partnership where you bring in the Intel AI Toolkit with some of your packages that are optimized such as Modin, pandas. So that if we're going ahead and we're working on a particular piece of code using regular pandas, say, if you're loading some data or something, instead of taking up to a minute to load in something, you can very easily just use that package distribution from Intel and be able to load that item in under two to three seconds, of course, depending on the size of your file that you're working with.

(09:43):

When we look at partnerships, this is kind of the ideal partnership where we have a platform, we have users that are having some issues in terms of either building out some of their models or processing some of their data or getting their models to provide answers very quickly. And Intel just kind of seamlessly comes in very nicely and says, "If you're interested, customer, here's something that we can help you with in order to help you with your current workload and speed up your processing." I've

mentioned, this is very important, because if we can help them speed up the processing on their decisions, they're going to make their decisions faster. They'll probably have a lower TCO.

(10:27):

Really what we're also doing is, with the current framework that we have and that we have Intel in as a partner, we don't have to pull in IT so much. A lot of the data scientists, data engineers and MLOps engineers can really go and self-manage the memory that they're using, the CPU that they're using. And it almost takes that pain of trying to set up IT on a platform out of the picture.

Tony (10:57):

Yeah, I definitely agree with that. My last job, which was trying to maintain a Data Center targeting MLOps platform was super challenging. We were trying to maintain Kubernetes and pieces of the ML stack that data scientists were interested in, making sure that they get access to their Jupyter Notebooks. How do I store the data? All of these things were so complex and it's really great that you guys are able to provide a simple solution for people, where they don't have to worry about that. Because I'm sure as a data scientist or as a MLOps person, I don't really want to deal with infrastructure if I don't really have to.

Audrey (11:29):

Yeah. That's so very true. I'm just going to push it back to Rachel, because obviously I've talked about the Red Hat OpenShift Data Science platform in that we're really excited and really happy to partner with somebody like Intel that can bring a lot of solutions such as the Intel AI Analytics Toolkit and OpenVINO. Rachel, can you talk more about the AI Kit and some of its benefits of using it on this RHODS platform?

Rachel (11:56):

Yeah, definitely. Just [inaudible 00:11:59] what you mentioned, Intel is also very happy to collaborate with Red Hat on this solution. So, as far as the Intel AI Analytics Toolkit on the Red Hat OpenShift Data Science platform, the Intel AI Kit is really targeting to make AI development easier and faster for users on Intel hardware. So, instead of learning a brand new software stack or having to overhaul your code to fit a specific hardware or resource or something, you could go ahead and use your beloved framework such as TensorFlow, scikit-learn, pandas. And what we've done through the Intel AI Analytics Toolkit is that we basically packaged a bunch of optimization, very simple, minimal code change, drop-in acceleration into this AI Kit.

(12:47):

For example, we could take Intel extensions for scikit-learn or just in a few lines of code, you could get a significant 10 to 100 times, even beyond that, performance gain just by adding a few lines of code here. They could learn algorithms, which really helps to speed up processing on to the GPU or in your Intel hardware stack, which in turn, as, Audrey, you were talking about is really reducing that total cost. Again, those struggles of, "Oh, I need to speed up my hardware in order to reduce my cost. How can I do that without going insane?" Well, we are trying to provide that answer with the Intel AI Analytics Toolkit.

(13:26):

So, we have different optimizations for data pre-processing. We have different optimizations for training as well as inference and quantization. What we're trying to do through these accelerations is basically be able to get data scientists and AI developers back to the core of what they were working on, which is really trying to make decisions faster, get back to their analytics and beyond just these accelerations, we are also trying to introduce new capabilities.

(13:54):

For example, we are trying to increase the functionality of these loved Python AI packages. For example, we could take, again, another part of the AI Kit, which is Intel distribution of Modin through a single line code change, be able to distribute your pandas data frames from just a single core, which is just typical for pandas, to all available cores or just a subset of cores that you could specify. That way you're able to go ahead and accelerate and scale your pandas workloads much higher than typical and avoid having to go ahead and overhaul your data pre-processing part down the line to another software stack.

(14:31):

Again, being able to use all of these same packages while you're scaling, without having to deal with the struggle of having to change it down the line. This just makes it easier for developers overall to make it more efficient with their developer time. And that, in combination with the Red Hat OpenShift Data Science platform, being able to make the struggle for scaling much easier as well as making the infrastructure needs and just accessibility of all these packages much easier on the Red Hat OpenShift Data Science platform.

(15:04):

So, it's a great long-term solution that we're trying to build here with Intel, as well as working on being able to get these optimizations available to all Intel hardware, so not just CPU, but GPU and beyond. Additionally, to AI Kit, where we also have available on Red Hat OpenShift Data Science is the Intel OpenVINO toolkit, which beyond the deep learning inference optimizations and deep learning optimizations in AI Kit, we also have deep learning inference optimizations and deployment opportunities available through the OpenVINO toolkit. So, a lot of great things that we currently have available already on the Red Hat OpenShift Data Science platform, which is all just great things to talk about. Now that I've gone ahead and listed all of the great benefits that we could see just from AI Kit being available on RHODS, Audrey, can we take some time to talk about some potential scenarios that customers may benefit from by using the AI Kit on RHODS?

Audrey (16:02):

Yeah, for sure. The first thing that I'm going to mention is a lot of the customers that I've dealt with, they really are starting from scratch. They're updating some of their older technology stacks to newer ones. Again, that's really their intent is to make use of AI/ML and of course GPUs. Well, as soon as they decide to do this, the first thing that they're doing is they're kind of going bare metal to cloud. What do you do when you're in the cloud? What type of platform do you use? How can you get all your folks working together? This is where we position our customers with RHODS.

(16:35):

Then there's the next kind of issue that comes in is, great, I have this platform that I'm working on, but I'm finding that I'm not getting that uptake in terms of some of the processing. Now I'm having that challenge, do I really want to put a GPU online? And I think Rachel and I have both alluded to this,

there's that cost associated. How can we be more cost-effective? And that's something that everybody, especially today in the current economic environment, is looking at.

(17:04):

I would say that people really do need to take a look at some of those challenges for the GPU costs and looked at optimizers or look at what you can use with what Intel has done with the AI Kit, because it's one thing to develop your data science application, your model, but it's also pulling in that data. And how does data processing fit into it? How many transactions are you pulling in per day? Are you working on the edge with some edge sensors, where you're even pulling in more data than most people would do? Instead of a million rows per week, might be having a million rows per hour.

(17:42):

So, we really have these challenges of being able to really optimize the workflows that we have as data scientists and as well, too, when we're working with those workflows, we also have to keep in mind that some of the stuff may be very time sensitive, whether it's in the financial industry, where you need to go through a lot of your workflows very quickly, and you may have limited developers helping you out. Also, to the biomedical industry, where you're looking to go over some CT scans to get a prognosis for your patient. All of those things really can benefit from the use of various optimizers or just the way you can have various options that are created with your Intel AI Analytics Toolkit. I can't emphasize enough that people are very concerned about cost. Yet, they really want that bang for their buck. They want lower costs, but they want to have their solutions that are working very efficiently and giving answers very quickly.

(18:47):

Really, in my mind, the only way that you can do that is either buy a GPU or a couple GPUs, or invest in some of the fantastic software that Intel has to offer. So, you really can optimize your workflows for a lower cost rather than using multiple GPUs. Dare I say it, it's also very easy to use, as a data scientist, I've gone ahead and tested a lot of the workflows that I use with Intel products. I'm very pleasantly surprised, should that come out in this podcast, I'm very happy with the results. I mean, it's not that difficult to use. Again, as a data scientist, I just want to do my coding. I just want to create my models. I want to look at what stories my data is telling me. I don't want to get bogged down in any infrastructure. I don't want to get bogged down setting up anything. I just want my answers really quickly. From what I've worked with on the Intel side, it really shows that you can do that very easily, which is fantastic.

Tony (19:46):

Yeah, productivity is king, right? That's really what we always like to say for us, for our customers.

Rachel (19:51):

Yeah. This is a problem facing all different aspects of the industry, even beyond just the technology industry and the biomedical industry, it could affect financial industry, media. I've seen all sorts of use cases where this really becomes applicable.

Audrey (20:09):

I think we've talked a lot about these great use cases and some of the information that we've both kind of given out to our users today. So, I'm going to flip it over to you, Rachel, if the podcast listeners want to find out more about some of the things that we've talked about, where can they go to?

Rachel (20:27):

Yeah, sure. So, to learn more about the Intel AI Analytics Toolkit and the great work we're doing with Red Hat, I definitely encourage you to go ahead and try it out using the Intel AI Analytics Toolkit Operator, as well as look at a few of our blogs, which is one blog I have here, One-Line Code Changes to Boost Your pandas, scikit-learn and TensorFlow Performance, as well as going the Intel AI Analytics Toolkit homepage to learn more about that. We also have a blog that we recently have done with Red Hat regarding what we talked about today, as well as the Red Hat and Intel joint website.

Radhika (21:04):

Great. Audrey, do we have any resources for this great collaboration that developers and data scientists can leverage?

Audrey (21:10):

Yeah. So, one of the things that I want to talk about is, besides some joint white papers that we've created and some Business Wire articles that we've worked on, talking about how orgs can really go ahead and accelerate their data science workflows, Intel and RHODS actually have a public-facing sandbox where you can go ahead and try Red Hat OpenShift Data Science, and actually try Intel AI Kit, and Intel OpenVINO for free. I think that if you're interested in seeing what benefits can be realized from using these products, you go give it a test drive and see if you like it.

Radhika (21:46):

Great. We're almost at the end of this podcast, but definitely it's been a pleasure, what a great conversation for developers, data scientists and all those in the AI industry to definitely use. I wanted to thank, Audrey. Thank you so much for being here and for your time.

Audrey (22:06):

My pleasure. Thank you.

Radhika (22:07):

Thanks, Rachel. It's great to have you back here for our podcast.

Rachel (22:12):

Thank you.

Radhika (22:13):

And Tony, it's awesome, it's great to have you and cannot wait to hear back from you on other podcasts.

Tony (22:22):

Oh, thanks, Radhika, yeah, I'm really looking forward to having these conversations with all kinds of interesting people on interesting topics for developers.

Radhika (22:29):

All right. Great. Thank you to all of our listeners as well today for joining us and let's continue the conversation at oneAPI.com.