Tony (00:04):

Welcome to Code Together, a podcast for developers by developers, where we discuss technology and trends in industry. I'm your host, Tony Mongkolsmai.

(00:16):

Today, we're lucky enough to have two people who are interested in Red Hat OpenShift Data Science, which is really a platform that people can leverage in order to make their AI workloads and their AI use cases easier to use. We're joined by Steven Huels, who has more than 20 years of experience with data management and analytic platforms. He's a senior director for Red Hat's cloud services business, focusing on Red Hat's artificial intelligence strategy, ecosystem partners, and customer implementations. He's one of the founders of the Open Data Hub project, a reference architecture for building a data science platform on Red Hat OpenShift. Welcome, Steve.

Steven (00:52):

Thanks, Tony. Glad to be here.

Tony (00:54):

Also joining us today is Raghu Moorthy, who is a global director in Intel's sales and marketing group and focuses on driving get-to-market, go-to-market, and sales engagements for Intel's technologies across Red Hat's hybrid cloud portfolio. Prior to his current role, Raghu worked in the data center group at Intel with end customers piloting and deploying early intel technologies. Welcome to the podcast, Raghu.

Raghu (01:16):

Yeah, glad to be here.

Tony (01:18):

So as I mentioned, today we're going to talk a little bit more about Red Hat OpenShift Data Science, which we talked about in the past. Previously, we kind of talked about how Intel's AI analytics toolkit integrated into the OpenShift Data Science platform to make things a little easier for people to use Intel technologies within that. Today, we're going to talk a little bit more about the platform in general, and to start us off, Steven, can you tell us a little bit about why OpenShift Data Science is really a good platform choice for people?

Steven (01:45):

It's a good question, Tony. We get this one a lot. When it comes to different data science platforms, there's a lot of them in the market. Where OpenShift Data Science really excels is first in its hybrid deployment capabilities. So there are a lot of the environments out there that are dedicated to specific platforms or specific environments, whereas OpenShift Data Science is meant to run across all the cloud providers. It can also be deployed on-premise in customers' data centers, and we also have flavors for self-managed environments as well as fully managed environments. So we give a lot of flexibility in the form factor and consumption model that customers get access to OpenShift Data Science.

(02:25):

Another strength we have is in our ability to extend the core environment. A lot of the platforms out there today have a lot of capabilities built into them, but their ability to reach out and integrate with

others in the ecosystem, especially where there's a lot of innovation happening around AI/ML, is limited. OpenShift Data Science has this as one of the foundational elements of the architecture, and this is what you've seen in how we've integrated Intel's OpenVINO and API Toolkit into the environment. We can do this with other partners. So as there's advancements in various disciplines around data science, we're able to quickly and easily integrate those partners and give customers access to that.

([03:04](#)):

And finally, and this is really a Red Hat theme across the board, is how close we are to open source. We have seen a tremendous shift in customer interest toward being closer to the open source technologies behind these data science libraries, frameworks, and tools versus having a lot of proprietary features layered on top of it. And with OpenShift Data Science, you're as close to open source as you're going to get, right? This is what Red Hat's been doing for decades now, which means we do deliver more of the open source experience, but also that we can lifecycle the platform more quickly and get you access to that innovation in a much more rapid fashion.

Tony ([03:42](#)):

So if somebody's looking, actually then, to spin up their data science platform or their pipeline, you guys are providing basically the building blocks is what you're saying? So they can say, "Okay, I need something to do data ingestion, I need something to help me with feature extraction," you guys are actually setting all those pieces up so people don't have to do it on their own systems.

Steven ([04:02](#)):

We set up a lot of the foundational elements that you would have in one of those environments. And so let's kind of go through the model lifecycle development. From a data extraction and operationalizing data standpoint, Red Hat has a set of capabilities, of which Data Science can integrate with the Red Hat capabilities around integration with streaming data, let's say, or data caching with DataGrid. But then we also have partnerships with companies, like Starburst, who can do this distributed data aggregation to pull that data into the environment. And so, for customers who choose to use the Red Hat capabilities, we have those. For those that want more specialized capabilities or already have investments with other partners, we're able to take advantage of those.

([04:45](#)):

When we look at the core Data Science experimentation platform, that's where OpenShift Data Science has a lot of capabilities today, right? We're able to deliver the Jupyter notebooks that everyone uses for those data science experiments, we give you access to a set of predefined notebooks that we've created and curated for you with things like TensorFlow and PyTorch, and we give you the ability to bring your own notebooks. So a lot of companies have their own flavors of notebooks, their own packages they've developed over time. We're able to integrate those and make those widely available to the data science communities so they can collaborate and build for it on experiments.

([05:20](#)):

When we look then at how we put models into production, so we are going to be launching a model serving component with OpenShift Data Science, so you'll be able to quickly take an output from an experiment, deploy that model at scale, monitor it, and life cycle it. We're going to provide all of the pipelines for the management of that model at scale. You have to monitor these things and life cycle them. And then when it comes to acceleration, we have partnerships with companies like Intel. So for customers who want to take advantage of any of the existing infrastructure they have, they can take

advantage of CPU acceleration through OpenVINO and oneAPI, as well as access to other accelerators out there in their market, like GPUs and FPGAs. So our goal is to deliver on certain components very deeply of that data science workflow, partner in others, and give full of access to all the accelerators needed for data science.

Tony ([06:13](#)):

Okay. And so let's bring Raghu into the conversation a little bit. Raghu, Steve's talked a little bit about how the platform is built and the underlying goals that they have. How is Intel contributing to this? Specifically, we think of Intel kind of as a hardware vendor, and as Steve mentioned, their CPUs. Intel's obviously looking at GPUs with our Flex GPUs now and looking ahead to GPUs that we might have next year. Where does Intel fit into this and why are we bringing value to the table?

Raghu ([06:39](#)):

Yeah, absolutely. Thanks, Tony. Great question. So let me just kind of set some context here. If you look at it from an Intel perspective, we have a very rich portfolio of both hardware and a layer of software that really enables or brings our hardware to life. So if you look at our hardware platform, we have CPUs with built-in AI acceleration in them, and we also have, like you mentioned, specifically accelerators in Habana, which is an accelerator for Deep Learning, as well as our GPUs as well, our discreet GPUs that's now coming in and are coming online here with the Flex series now and then going forward next year into our training base GPUs as well.

([07:17](#)):

So for that entire portfolio of our Intel hardware, how do we bring it all together in a commercial setting? That's really where we've been collaborating very closely with Red Hat. So we have a number of software elements that really bring that to life, and if you look it from a layer perspective, at the very base layer, we have something called oneAPI. And within that, we have a number of highly performant libraries that really extract and also abstract, so that regardless of the hardware you're running it on, you're able to essentially use the same framework and get the benefit of the entire span of Intel hardware that is available to the data scientists or the developers.

([08:00](#)):

If I look at it from an end-to-end pipeline type of a perspective, we have optimized tool sets where you're able to do that whole from data extraction to data cleansing to machine learning and Deep Learning with our Python libraries, all the optimized Python libraries, and then in the machine learning category, like scikit-learn and XGBoost, as well as in the Deep Learning side with TensorFlow and PyTorch and all this. All of this, the beauty of all this, is you're essentially using the same open-source tool set that you already used to with extremely, literally one line of code change where now you're able to leverage all of this in a very seamless fashion.

([08:39](#)):

And then taking those models that you build and optimizing it for deployment is the other piece. So we have tool sets like OpenVINO, which let you take that model, optimize it, and then deploy that at any endpoints. And then bringing it all together, we're able to have MLOps toolsets like cnvrg, that pulls all of these things together and makes it usable in the context of Red Hat. So our collaboration here, what we've really done is taken all this I will call piece part or different components of Intel AI portfolio and really glued it together and made it available commercially on the Red Hat OpenShift and Red Hat OpenShift Data Science. So from an end customer perspective or end data scientist or a developer

perspective, they have one platform, which basically brings the goodness of Intel and the goodness of Red Hat OpenShift and Red Hat OpenShift Data Science that Steven talked about all in one piece together, so that it's as seamless an experience that they can get.

Tony ([09:37](#)):

Yeah. It's interesting, because you guys both kind of talked about different parts of the stack, actually. And Raghu was talking a little bit about things that data scientists would care about at the higher level when he is talking about the scikit-learn and other machine-learning libraries. And Steve, you also talked about that a little bit, Jupyter notebooks bringing things to where the customer and the data science person wants to work. But at the end of the day, this is actually a platform.

([09:59](#)):

So the goal of this, it sounds like to me, is to make it easy to get the data scientist person up and running, but really your target audience is the person who's having to deploy this infrastructure. And in the past, I had to do this for one of my teams when I was at Habana, and that was a very challenging thing to do. And I talked about it in the past in other podcasts, but to me, that seems like the big value-add here is being able to get somebody up and running and actually solving real problems as quickly as possible versus trying to beat your infrastructure into submission, so to speak.

Steven ([10:34](#)):

You're spot on, Tony, and in fact, that is largely the value delivered through OpenShift Data Science. A lot of the components we have in there are components that customers and users could deploy manually and stitch them together today. But through our unified approach to deploying these, you end up with a functional system in a couple of minutes, optimized for your platform.

([10:55](#)):

The other part you touched on was a lot of data science focuses on the experiments and the models themselves. That's the flashy part, and ultimately, that's the part that customers are looking to get value out of. Part of the other challenge is operationalizing those models. A couple of years back, having a platform for development was a challenge, and that was where a lot of customers were focused. Now that things like OpenShift Data Science are out there, they're able to have their data scientists out there experimenting and building models, and they've realized and moved on to the next order problem, which is how do I operationalize those models?

([11:31](#)):

And this again, when you talk about a platform through our partnership with cnvrg as well as the components we have built into OpenShift Data Science itself, we're able to address taking models and putting them out there as endpoints or rolling them into applications to help realize value out of those models. And from there then, the next order problem that customers run into is how do I do it at scale? Because once you've been able to successfully roll out one, you have to learn to monitor and life cycle it, but then you quickly find yourself having dozens if not thousands of models out there in the wild attached to various applications, a lot of them in combination with one another. And that's really where we're starting to focus a lot of our efforts, to be able to give transparency and explainability to the models that customers are putting into production, understanding how to life cycle them, and then giving all of the tools to be able to automate that and have reproducibility in the overall experience.

Tony ([12:25](#)):

One of the challenges there, I guess, is you guys, if you compare what you guys are offering versus what other cloud vendors are offering, it sounds to me like you guys have a slightly more flexible model into what goes into the pipeline versus potentially what the other cloud vendors have. And to be honest, I'll say I'm not necessarily familiar with building a production life cycle flow in the cloud, but to me, it sounds like you guys are pretty flexible about that. Is that something that's a differentiator, or is that something that everybody does, what maybe you guys have a unique spin on?

Steven ([12:54](#)):

I think everyone's trying to solve the problem of reproducibility in data science. I think where Red Hat as a whole excels is that from the beginning of OpenShift, one of the core premises there is that we were helping put applications into production for developers, which meant we had all of the plumbing and tools and integrations for a full end-to-end DevOps life cycle. So we had code management, artifact management, life cycles, triggers, auditability, security, all those components already built into the core fundamental platform OpenShift itself.

([13:30](#)):

So when we started to look at how do we operationalize data science models, we really leaned heavily on that experience and started applying basically that GitOps model to data science, treating models as code artifacts. So as developers are working with Jupyter notebooks and they commit changes, you have auditability in what changes were made, why they were made. You can commit the models as artifacts so you can see what impact those changes had on the outputs of the model, the decisions it was making. Once things were rolled into production, we already had enterprise-grade monitoring built into our platform that we were able to tie into to be able to give insights into things like performance, and then we could extend it to have a certain degree of AI explainability to detect model drift and bias and things that would cause you to want to retrigger those models.

([14:20](#)):

And so I think you're right, it is an advantage we've had, and it's one I think that we've leaned into now for several years. We also have the ability to deploy across all the platforms. So one of Red Hat's sort of core underpinnings has always been code-at-once, run-it-anywhere from an OpenShift standpoint. And so we do get the reproducibility, and we kind of get carried on the back of OpenShift with that one as well, as customers want to migrate different workloads from data center to cloud or even to edge deployments.

Tony ([14:48](#)):

Yeah, that's a good point. I guess I think of it as OpenShift Data Science platform, but really what you're saying is you have a legacy of having really consistent, good customer experiences on top of OpenShift, and now you're just kind of extending that to enable data science pipelines. So you kind of have this nice, "I've been working on this for a long time, everything is stable, and we've solved all of these problems of building this kind of hybrid cloud," but now you're just kind of adding another layer on top of it.

Raghu ([15:18](#)):

And maybe I can add to that, back to that flexibility comment you made, Tony. So I look at this as a few elements of flexibility. There is the flexibility with the number of ecosystem partners that are now available on the OpenShift platform and RHODS specifically. So what that lets you do is see absolutely

there is enough good tools already in place, but that's not there. There's always any number of, endless number of new tools and options available. So this is an open platform where all those tool chains can now land here, and then we can slowly expand the realm of what's possible and what's to do about it. That's one element of flexibility that you get.

([15:52](#)):

The other part of this is also one thing to remember, OpenShift is a very broad platform. You have commercial standard enterprise workloads that are only learning on it that are generating data, that are generating all the kind of existing stuff that's in place. Now AI is now essentially an add-on to that. You're essentially taking data from that or feeding recommendations back into those platforms. Now, you already have that one platform. Now, you're essentially adding this AI piece and making it extremely flexible and seamless from an integration perspective.

([16:27](#)):

The third aspect I think Steven just touched on it, from a hardware perspective, it's so important that when you touch the whole hybrid cloud model, you're able to do this anywhere essentially. I can train on cloud, deploy on edge, or creating a data center for any number of reasons, for data brevity reasons or for compliance reasons, we have any number of reasons why you may have to do that, right? So you're able to do that but still have that same consistent experience in moving back and forth. And that's why it's so easy from an Intel perspective. We have a lot of different elements that I could go into that I alluded to earlier. We were able to bring all of those pieces and kind of put it together on this one platform and making it commercially usable from a data scientist and a developer perspective.

Tony ([17:09](#)):

Yeah, that makes sense. And it's interesting, too, because a lot of the Intel components can be challenging to understand when to deploy them and where to hook them up, and that's true about the entire open source ecosystem around AI. And so,it's really nice that Intel is able to partner with Red Hat to provide something a little more stable there while still allowing that flexibility for people to pull in kind of their favorite go-to components. So when you guys deploy this on OpenShift, I assume you guys are deploying it on top of Kubernetes, some type of storage. Let's get into the details a little bit for people who might be interested in understanding what they're going to get if they deploy this. Can you talk about the infrastructure a little bit there, Steven?

Steven ([17:50](#)):

Sure. So you nailed it. Actually, you can give the presentation just as well here. The underlying orchestration platform is Kubernetes, which we deploy on top of OpenShift, so Red Hat's enterprise Kubernetes deployment. From a storage perspective, really, the world's your oyster, right? When we deploy into cloud hyperscaler environments, it's pretty common we go ahead and hook up something to S3, right? S3 is becoming an increasingly popular data platform for model exchange and for unstructured data, so that's a pretty common one. Obviously, customers have been compiling and storing data for years and years and years now, so there's always a need to tie into some storage layer behind the scenes, which we have the capability of doing.

([18:31](#)):

When we get to the hardware layer, it's whatever accelerator it is they want access to. So obviously, we've got our partnerships with Intel, and so if they want the CPU acceleration, we have the toolkits there. If there's GPUs, we support those. We recently announced support for Intel's Habana as well,

which is pretty exciting for Deep Learning use cases. Customers really get to pick and choose there on what they want us to integrate with.

(18:53):

As we move up the stack, the core components that can get laid down with OpenShift Data Science are the Jupyter Notebook environment as well as a set of notebook controllers to allow for collaboration. We have a standard set of notebook images that have optimized and secured things, like versions of the frameworks for TensorFlow, PyTorch, scikit-learn, all the standard ones you would expect for data science. And then coming up here in the not too distant future, possibly by the time this publishes, we'll have our model serving components, which work on distributed model serving. And then shortly after that, we're going to be adding our data science and ML pipeline components, which will build on top of Red Hat's core OpenShift pipelines. From there, customers have the ability to add any of our integrated partners, so this adds where they can deploy the OpenVINO or AI Toolkit components. And then in addition to that, Red Hat has a very extensive partner ecosystem. Not all of them are integrated into OpenShift Data Science, but we do have the ability to still integrate with those tools. Customers aren't limited in what they can deploy.

Tony (19:58):

Cool. Yeah, that sounds like a pretty comprehensive solution. I just love harping on the flexibility of things like this, because really, it's such a greenfield even now. When it comes to AI platforms, there's really no one answer. So having that flexibility, I guess that's my favorite word of the day, will allow you to really build what you need. Raghu, is there a particular Intel component here that you are really excited for? I know Steve actually mentioned Habana, which has a place in my heart having worked for them, but is there a particular component that you're excited about, one little piece that you think is really going to make a difference to data science users?

Raghu (20:31):

Actually, I would say the AI Toolkit and OpenVINO, definitely, because the beauty there is most customers already have an environment that's based on Xeons. So we are essentially able to, without any additional hardware investment, you're able to get an additional order of magnitude better performance for almost no cost. So extending that existing infrastructure and getting that extra boost is phenomenal. And I think you already brought up Habana. Now, customers that do want that specialized piece of hardware, Habana is absolutely a fantastic choice for Deep Learning workloads. And then I touched on this as well, kind of the last piece I'm really excited about this going forward, we already are working on the Flex series side industry GPU that should be tech-reviewing here later this year, and then really bringing this to a broader GPU adoption as well within the next year. So we have a fantastic roadmap map also ahead of us in terms of putting it all together.

(21:28):

We all know when you have a cloud platform, tech there's going to be dependent on what that cloud platform really offers from a hardware perspective. Now with OpenShift Data Science, since it's deployable anywhere, all these great things that even if there's a delay bringing this into your favorite cloud platform, you can still go experience this. You can actually go set up a cluster that's outside of that and start working on these things right away. So it opens up another whole set of possibilities that customers and data scientists can really jump into.

Steven (21:59):

And I want to echo what Raghu said there. It's really an untapped potential when we talk about CPU acceleration. I think by default, everyone sort of thinks of data science and I need a fancy accelerator or a specialized environment, and there's really so much you can do with your existing infrastructure, especially in times when supply chains are all at risk, and it's hard to get access to some of these other pieces of hardware. This is immediate value that you can start to realize. And even if it's just in the early experimentation phases of your life cycle, it's a great opportunity to try things out very easily, take advantage of hardware you have, ahead of trying to invest in something new. So it's one we really are very excited about and talk to customers a lot about.

Raghu (22:43):

Yep. And all at no additional cost or minimal additional cost to get that benefit.

Tony (22:49):

So if I was going to go and try this and actually kind of trial this, obviously, I'd want to look at it before I wanted to invest a lot into it to make sure it meets my needs, I'm looking at the Red Hat website, and actually see they have a sandbox option. So that's something, Steve, that people can kind of try it before they buy it, or what's that?

Steven (23:08):

That's exactly what it is. So our developer sandbox is meant as a playpen for data scientists and other developers to come and try out Red Hat technology. Obviously, we don't give you unlimited resources, we're going to constrain that a little bit, but it's good enough to build a simple object-detection model or fraud-detection model. You can get a sense of the tools. You can also get a sense of the integration with the Intel components and see how that works. From there, if you wanted to spin up another environment, we have the ability to do so. You can either contact us at Red Hat or there's self-service as well through various marketplaces. So yeah, it's absolutely meant to kick the tires, and if you use it and have feedback, I love hearing that stuff as well. That's how we improve things overall.

Raghu (23:52):

And then to add to that, Intel has a tool set to go get access to, playpen is a good moniker for that, and often learn all of this. So we also made that easier as well. So we actually launched last month a joint Intel-Red Hat developer program, and there's actually a joint website where you're essentially, it kind of lists all the assets, training assets, 101s, the 201s, the 301 type curriculum that you can come and use that as your learning tool, and then go and jump into the sandbox to go try it out. So it's a great opportunity for developers to come learn and then try it, all in that same, very easy to use context.

Tony (24:27):

Cool. And Steven, you said it was kind of a hybrid AI platform. What did you mean by that? Is that something where I can kind of run it on my machine or in my cluster, maybe in a data center and then scale out to the cloud using the OpenShift spillover model, or is that something else that you were talking about around the platform inner workings? What did you mean by hybrid AI platform?

Steven (24:49):

So we cover it in a couple of different use cases. I'll talk maybe a little bit about some of the cases that I get involved in a lot. Especially right now with a lot of the market volatility, capital markets ends up being a very interesting use case, where historically, they would get an evening's worth of data, they'd have 24 hours to train their model, redeploy that model, and then the next day people operate off of the outputs of that model. With all of the volatility now in the markets, they're being asked to retrain that model multiple times a day and sometimes multiple times an hour. And the existing infrastructure that they've typically run these jobs on simply can't scale to that kind of demand.

([25:28](#)):

And so what we have the ability to do is to be able to extend your perhaps on-prem data center to add additional capacity, whether it's across your data center, so you can add more nodes and do the dynamic scaling, or you can go into one of the clouds. So if it ends up being cheaper to be able to extend your overall Kubernetes cluster into one of the cloud environments, you can run it there. And so that's one example of hybrid training.

([25:52](#)):

Another area that we get into is when it comes to model deployment, it's very common that there is a separate production cluster where models are served than the development cluster where the experiments are happening. And so the way we've set up OpenShift Data Science is that for model serving, you have the ability to distribute models to any platform. It can be an OpenShift platform, it can be any Kubernetes platform, it can be an edge ARM-based platform. And at the end of the day, those models will phone home back to your core OpenShift cluster for monitoring and lifecycle purposes. So we're able to support multiple platforms and still be able to serve your model-serving needs.

Tony ([26:34](#)):

Yeah, that's cool. Yeah, that makes a lot of sense that the spillover model really seems to be something that is really great, especially like you said that you never know nowadays when you need to scale up and scale out. So it's good to have that flexibility.

Raghu ([26:46](#)):

And then maybe just to add one more point to that. While you're doing that, it's the same experience. Whether you're doing it on-prem in your data center or you're doing it in the cloud, you're essentially getting. So just from an ease of use and deployability and just having an instant process, it makes it extremely easy to train your data scientist on one platform from the whole end-to-end cycle perspective, right?

Tony ([27:09](#)):

Yeah, that sounds like a really great platform. I'm actually going to go try the sandbox right after this and see how it works.

Raghu ([27:16](#)):

And the beauty there is that within a matter of sub-five minutes, I should say even lower than five minutes, you can get access to the Intel tools, OpenVINO, AI Toolkit, in three or four minutes of signing up, and you can get access to and start trying it off right away. It can't get any easier than that.

Tony ([27:31](#)):

Yeah, no, that's great. Are there any resources that you guys want to point people to? We've talked a lot about Red Hat OpenShift Data Science, but Steve, are there places that people should specifically look to find out more information there?

Steven ([27:42](#)):

So you nailed the developer sandbox, so thank you. You put that one there right away. The other one we'd like to point folks to is our joint developer website, which talks about some of the joint solutions between Red Hat and Intel and can get you started with various quick starts and tutorials if you're new to it that you can take and run in your own cluster if you have one as well.

Tony ([28:03](#)):

Cool. Raghu, anything on your side that we might want to point people to from Intel?

Raghu ([28:07](#)):

Yeah, just reiterating that developer website's a great starting point and has everything in one place, both to learn and try. It's a fantastic starting point.

Tony ([28:17](#)):

Okay, well I think that's going to be the end of our time today. I'd like to thank Steve and Raghu for spending some time talking to us about data science platforms, and I'd like to thank you, the listener, for tuning in. We'll talk again next time about more technology topics. Thank you.