**Matt Hicks:** 0:00

If you go back to the early 2000s, the debate was operating systems. They may sound boring, but like these are the things that run our hospitals, our airlines, our defense systems. How on earth could you have every line of code visible to the world? Because all the bad actors could just find their exploits in it and then risk health and travel and national security and all the things. Ai, with the right structure and creation to it, can have a lot more good than bad, but you always have to think through the bad and the malicious uses in technology.

**Craig Smith:** 0:38

Hi, I'm Craig Smith and this is my AI. In this episode, we're joined by Matt Hicks, ceo of Red Hat, to explore open sources' pivotal role in technology innovation, particularly focusing on large language models. Red Hat, known for its proficiency in infrastructure software, bridges the gap between model development and deployment. Matt, with his 18-year journey at Red Hat, shares his profound experience. Whether an open source enthusiast or curious about the future of generative AI, this episode is packed with riveting discussions and enlightening perspectives. Let me mention our sponsor, netsuite by Oracle. If you're a business owner, having a single source of truth is critical to running your operations. So remember these three numbers 36000, 25,1. 36,000 because that's the number of businesses that have upgraded to NetSuite by Oracle. Netsuite is the number one cloud financial system for streamlining accounting, financial management, inventory, hr and more. The number 25 because NetSuite turns 25 this year. That's 25 years of helping businesses do more with less, close their books in days, not weeks, and drive down costs. One because your business is one of a kind, so you get a customized solution for all of your KPIs in one efficient system. With one source of truth. You manage risk, get the most reliable forecasts and improve margins everything you need all in one place. I'm not the most organized person in the world, and there's real power to having all the information you need in one place to make better decisions. This is an unprecedented offer by NetSuite to make that possible. Right now, download NetSuite's popular KPI checklist, designed to give you consistently excellent performance. Download it for free at netsuitecom. They support us, so let's support them.

**Matt Hicks:** 3:08

Matt Hicks. I'm the CEO of Red Hat right now and I have been a Red Hat for almost 18 years now. So I actually joined the company when Red Hat was pretty much in startup mode. We just had Red Hat Enterprise Linux. I went to school for hardware, for computer engineering, but at the time I learned software really well. That introduced me to open source software, which drew me to Red Hat. I loved that model and I started in IT, actually in Red Hat, sort of applying software. I worked about every role you could imagine in IT and in the software groups. I moved up through the ranks to where now I get the privilege of getting to help try to lead the company. I enjoyed every minute of it for the last 18 years here.

**Craig Smith:** 4:00

Yeah, that's incredible. Just for listeners who aren't that familiar with Red Hat, can you take us back to Linus Torval? Was he involved in the creation of Red Hat, or was that spun out of the foundation? Anyway, how did Red Hat come to be and how does it relate to Linux?

**Matt Hicks:** 4:21

Red Hat at the time in the early 2000s, even a little before that, linus and a lot of the software around there. There's this belief and new model of creating open source software and that turned into a Linux distribution of how do you collect all of these things, the GNU technologies, the Linux kernel and put it together into something usable. I remember at the time there were lots of different distribution choices there, from VA Linux and Caviar, yellow Dog, all sorts of them. Red Hat was one of those distributions. The angle that Red Hat took was that they were going to serve enterprises with Red Hat Enterprise Linux. Red Hat actually started selling boxes of Linux and Best Buys and Circuit City at the time, but they made that pivot to provide predictable life cycles and long-term support for enterprises to be able to adopt Linux. That's really where Red Hat got their start. Was that Enterprise focus on it? Since then, the rest of the history, I think the model of open source and software development has really taken off. Linux, I would say, is the operating system that underpins a whole lot of modern enterprises these days. Red Hat's grown from 800 or so people when I started to over 20,000 worldwide. We do Linux products like OpenShift and containers, automations or the full gamut of software technology, all based on an open source model.

**Craig Smith:** 6:08

You're still primarily like a service company wrapped around an open source operating system. Is that right? You help take the Linux operating system and implement it and refine it or deepen it or add features or whatever it's built around the core open source system. Is that right? That's right.

**Matt Hicks:** 6:37

All the software we produce is under open source licenses, so it's freely available to the world. What we do is we focus on the Brayfix support. We obviously know the software, we write a lot of it, and then also things like security updates, when most enterprises can't stay on the latest versions of things. Open source moves so fast and so we make sure that security, bug fixes, those things are also applied to older versions so that even if you're running, in some cases, linux versions that are 10 years old, you can still be secure and on a stable platform. So we do the Brayfix support and we like to describe it as its innovation at a rate that enterprises can consume from open source. The open source innovation is tremendous and we give companies all of those choices on that life cycle of how they want to consume it.

**Craig Smith:** 7:35

Yeah, just generally on open source. I remember when Linux first came out or first sort of hit the public consciousness, it was considered, and you'll forgive me, kind of a poor cousin to commercial operating systems that were bundled with hardware largely and. But over time it's grown and I think at one point I had the impression that open source was sort of the future, because you have this hive mind, a fact of all these people improving, improving, and how could a commercial enterprise hiring expensive Engineers and developers keep up with that and and and to some extent that's born out. So can you talk first of all about the promise of open source and how it stacks up against proprietary software, and then we can talk about generative AI, which is, I said, is what I'm interested in, and how open source maybe breaks down in that arena because of the hardware constraints and cost of all those things.

**Matt Hicks:** 9:04

Yeah. So I'll give an example of open source from sort of my moment in the field and I actually, before red hat I was in consulting and as a consultant I was literally deploying unit systems right next to red hat enterprise links. It was actually red hat enterprise linux 2.1. It was the first version of red hat release and the difference to me that really brought the promise home to me as a consultant was If I hit an issue with the unit's deployment. They were at the time much more polished, they worked probably more consistently, but I was paralyzed with any issue. I was dependent on a vendor at that point to put in a support ticket. I was in the queue and you're a consultant. Your whole job is to make progress and get deployments finished on Linux with all of rel's rough edges at the time, and it was totally transparent to me. I could go, understand every line of code, I could debug anything myself, I could put in a support request to red hat, but I could actually work a lot of that issue myself. There was nothing hidden from me on it and I I fell in love with that model because I felt like this really lets me do my best work with it, and I think it wasn't just me, it was millions of people around the world fell in love with that model of just being able to not just be given something that they use like a black box, but to be able to understand it and learn from it and iterate on it and improve it. It really is. It's pretty captivating, and so I think that that's why you've seen that model applied to so many different forms of software at this point. But, yeah, that's sort of the promise, and that's why, you know, at Red Hat we often say, like passion runs deep. At Red Hat we have associates that have our logo tattooed on their body and we have changed our logo and they have both of them tattooed on them. So it's a different level of belief in sort of what we show up and do every day versus just producing software and that is, you know, we're trying to change the world for the better in the process.

**Craig Smith:** 11:24

Right, ok, so, and we can talk about all the different kinds of open source software that have come along since then and the successes and failures. I'm interested in that model now. You know Meta has open sourced the Lama models and you know I started talking when GPT two came out. I'm sure you're familiar with Connor Lehi, who is at Luther AI, who built the first open source large language model, and at the time that's that. It seemed kind of scary that this decentralized group of hackers could build a large language model to rival open AI or at least come to think of it, they didn't have the money to get as large as open AI, but they could build the model. Now you have met open sourcing models and I was just reading Yanlacoon's testimony at the tech hearing in Washington. I can't remember, frankly, which one it was, but he was talking about how open source will play a big role in the spread of large language models. But I've since had conversations with people that argue yeah, there will be small, commoditized LLMs that people can deploy for for specific use use cases, but the future of generative AI remains with with big companies and the proprietary proprietary model Because they're so costly to build and deploy and particularly, everyone knows about the training. So something like an operating system which I could put on my laptop. It's very different from a large language model where the cutting edge is going to be beyond the reach of the open source community because they can't marshal those kinds of resources. So, yeah, what are your thoughts on that whole debate? Whether Yanlacoon's right to open sources of the future for generative AI or these proprietary proponents, I think you can argue that the cost is beyond the reach of open source.

**Matt Hicks:** 14:25

Yeah, you know, I think I've certainly heard these arguments before. Take relational databases. I think it's a classic example of the days I was coming up in. Software is like no, no companies are ever going to be able to put in the R&D effort to build relational databases to complex the query optimizations that are saying won't be achievable within this model. And yet if you look today, I think there's always a use for both models in it. But I would argue that you take relational databases and there are different scaling patterns that have come up, not just vertical scaling but horizontal scaling. Much of that gap has been closed with it. When you look at large language models today, I do think there's always an element of truth. To build that first foundation model on it, it is an incredibly resource and time intensive process. We work pretty closely with a lot of different companies, but with IBM in this space, and it does take a sizable amount of resources to build those first models. What I think is pretty amazing that's happening right now, though, is the specialization of these models and the resources that it takes to refine them is a laptop operation. It's not a hundred or thousand GPU operation, and that areas like that is where we see the open source contribution and tuning and specialization and application exploding, because there is a lot of IP that goes into that original model, but I would say I'd probably see as much innovation that goes into the refinement of a general model and to specialize use cases, and that is pretty within reach of a really really wide base of developers at this point. So, and then I would say, as technology is going to change a lot in the next five to ten years, from a hardware aspect of what it takes to build those foundation models, I do think it might be today's statement that that's out of reach for a lot of people. On it, not the refinement aspect, but I don't know how many years that actually holds on it. But just putting more billions of parameters into the models hasn't necessarily shown to drive the same accuracy rate. So I do think we're going to learn a lot in the next couple of years of what drives the best outcomes from it. Is it in the refinement, is it in the core model, and when do you sort of hit diminishing returns in that? But it is certainly moving fast, I think, on both sides of the original model creation and then what we see people doing with it afterwards.

**Craig Smith:** 17:26

Yeah, that's interesting and that actually not contradicts but challenges one of the arguments that's been made. To me and I'm a journalist this stuff gets very confusing, you know, as whoever has the bigger office or the most prestigious PhD, you know, you tend to listen to them. So a foundation model, right, takes I don't know tens of millions, hundreds of millions to build, and if it's open source, you don't need to build and train that model. You download the weights but you still need infrastructure and a lot of power to put those weights on a model in the cloud or something. I mean these larger models are not going to fit on your laptop. Yeah, so I mean one of the arguments is that, yeah, open sourcing allows people to do that, to take a model and then fine tune it, but that's still the domain of companies that are extremely well resourced. It's because, to back up a little bit, one of the concerns, when you know a Luther AI, open sourced GPTJ, that first open source large language model, and now currently he is kind of switched sides and he's running an alignment startup and is very freaked out by the power of generative AI and how, if it falls in the wrong hands, and how open sourcing is a terrible idea because people can then take that model and fine tune it or refine it to do terrible things. So just talk a little bit from your understanding and I realize that you're in a different part of the business, but you know, downloading the weights of Lama 2 and how realistic is it that a small company or a group of people could take that model and deploy it and then Find tuna?

**Matt Hicks:** 20:13

Yeah, well, I'll go through an example. Um, I Know because I like playing around with this and we put a llama to the side for a second look at something like this, stable diffusion or For those might not be steeped in image generation and manipulation in this. It was not that long ago I mean we're talking probably a year where that lived in the land of big, well-resourced companies that could right Hugging face, which is an aggregate of all these refined models. It is stunning to me to see how many people, with their own custom data, have been able to specialize that image generation into domains that they know, whether it's Animation focused areas, or you see people that are, you know, specializing image generation on their animals or their pets so that they can create it with their dog and cat. There is a Tremendous amount of creativity that's being applied to these, where I'm sure that the creators that train stable diffusion these. They know art, but they're not gonna know Anime as well, as I'm someone who lives in that domain and yet we're seeing you know, creators that live in those domains that are able to Download these, you know, sizable models they're not at the same parameter counts that you're talking about but then be able to apply unique data to them, get a pretty impressive refined result and then share it back and, where hugging face comes in, as in, someone else can pick up that refined model and refine it again. And it's been interesting to me, you know it is not with code. I understand the code. It's a very linear progression. I can. I can get work, I can make an incremental change. I sort of know the outcome with it. With AI it's not quite as linear to it. Yeah, you, really the outcomes and refinements you might do might have really unexpected or Pretty amazing results on it. So it is a very incremental process to get there. But that aspect, I think it's amazing how accurate it's gotten on it and this is really for most people it's constrained by their laptop or their desktop on it. They're not going into Amazon and running a bunch of Nvidia GPUs to do that. But I think the same thing is we will get there on smaller, large language model Parameter counts for it where you get to a point where you can load on modern hardware and you can refine it, and One of the reasons I remind people that I think this will happen is In in our world and we're in the world of you know we connect this modern hardware to the Applications you run. So we're in that. It's called inference, like when your model is trained. How do we run it efficiently? One of the most Common places that we see these models that desire to run them is at what we call the edge. So you can think of the edges in your car or in your factory. Having a 20 billion parameter model that needs really Sophisticated hardware to just load into memory to run is not that practical in a modern-day factory with it. And so there is a boundary constraint, I think, that just Will be put in place for the cost of inference. How much does it cost to ask that question? And we're gonna want to ask a lot of questions about autonomous driving or lane correction or those things, but there's a cost constraint there. So I think we'll always see that pressure to get these really powerful models Small enough to be able to run on edge-based hardware, and the minute that happens it is usually within reach of Hardware in a laptop form, a desktop form that most developers can get. So I do think there's a self-correcting aspect. They'll always be the chat GPT is where you can have a big model. You can run it in the cloud with pretty impressive resources behind it. But I think that's just going to be one vein where we'll see AI applied Outside of the data center as much as inside of the data center. Yeah makes sense.

**Craig Smith:** 24:52

Yeah, and that brings up a couple of other questions. So the this Discussion I've had now with several people on open sourcing these large models, one sort of view is that it's going to fall into commoditized models Based on open source, as you say, hugging face, and then Very large foundation models, that that are are going to eventually develop agency and be able to use tools you know, connect to other tools and and that that that the the real value creation is is in that Higher form. That, yes, there will be commoditized, smaller models and there will be a lot of economic benefit to that, but the real value creation is going to be at the frontier, where these models that again are out of reach of open source developers Because of the resource constraints are the ones that will will be marching forward towards a GI or Whether you believe in that. I mean, does that sound like a reasonable scenario to you?

**Matt Hicks:** 26:22

And maybe I'm an incrementalist on this, but I think there's always gonna be the general purpose AI debate and risk. At what point can it learn by itself and do all the things, and is that good or bad? On it, I read a good analogy. That's sort of like you're traversing a mountain range and the only thing you can see in front of you is the next mountain for it. And so it's very hard to tell how far ahead that is, but I think most people would agree like we're not there right now. On it we have my experience in it. It's not great in novel work. It is very good at reproducing and analyzing things that have been created, not the best at creating something new that's not there. So we have some work to do there. But if you look at it like this for me, where will enterprises get the most value right now? Will it be in the race for a general AI knowledge or will the winner take all or will it be in the application of pretty impressive specialized models, like whether we call them commoditized or not. It's sort of like the application of Linux. You could certainly say an operating system at this point is a commodity. You have a lot of different choices. But an operating system that can run on ARM hardware that enables you at the edge, or run on the latest Intel gear to give you a better inference and capability or ties into the next generation GPUs for it, there is real value in that and so for us, I think most of the enterprise value, which is gonna change how we interact with things every day, is probably for a while gonna be applied at that specialized model where, openly, I think open source is gonna have much more influence and application. There we will have large resource efforts at government levels and large companies on general purpose AI and how does quantum affect that, and all of these things in parallel. But I think there's a tremendous amount of value unlocked in terms of how we can do some of this work that we're constrained on really, really well. We know how to do it and we have shown with large language models and refinement that the AI we have at our fingertips today is good, like impressively good at doing this work, but it can't be high level. I get it right 90% of the time. That specialization to get you to the 99, 99.9 accuracy it's needed for most businesses in an industry who are right. If you're in the medical profession, telecommunications, you need that precision and accuracy and that's where I think we're gonna see just a lot of. I mean the thing I love with open source. It's inspiring to me to see how much creativity is applied into software. In the early days it was the Linux from every corner on this planet and it's not always from the people that you would expect if you like the classically trained backgrounds and I feel like we're seeing the same thing. I'll give you a good example, a moment that I love. We have a big event each year called Red Hat Summit and I was talking to our director who coordinates all of the work at Summit. It's a large, it's director like a movie director type role on it and he was telling me how he has the video of all of us and we have to do animated work every once in a while and he was playing with a hugging face model and refining it with data of us presenting on stage to help drive like immediate animation view vector graphics of us as he needed them. And Dave in this case is not. He's not from engineering, he's from a creative background and the fact that he could apply that and was empowered to do it it's a pretty neat moment. I think that's where we're gonna see a ton of value in parallel to some really big companies and it's like the moon race type things to get to larger, more capable AI platforms. But I don't think that's the only place that value will come from there, yeah, okay, then to another part of the open source debate.

**Craig Smith:** 31:20

With regards to generative AI, there's a lot of concern that, just in the way that this fellow could take an open source model and train it with his data and do something interesting with it, a malicious group of actors could take an open source model and refine it to was it Mustafa Suleyman or somebody? Recently, I saw, as described, a test where they asked a model to create bio weapons or something, and overnight it created, like you know, some massive amount of super toxic molecules that then someone could synthesize. So, yeah, what do you think of that argument that open sourcing sort of takes the guardrails off and everything that people are worried about is suddenly possible?

**Matt Hicks:** 32:38

Yeah, I think this is tough being in technology. I think this has for as long as technology has been around. It's been that double edged sword with technology. If you go back to the early 2000s, the debate was you know operating systems. They may sound boring, but like these are the things that run our hospitals, our airlines, our defense systems. How on earth could you have every line of code visible to the world? Because all the bad actors could just find their exploits in it and then risk health and travel and national security and all the things on it. There's a very similar argument at the time. What played out, I think, was that visibility, the accessibility to understand, the creativity and the contribution to it closed those holes faster than proprietary models for it because bugs exist in software on it. But the open source models. There's the statement you know, with enough eyes, all bugs are shallow. I think Linus might have said that that really started to play out in that model. Now the thing that becomes uncomfortable with that is you know, I think as a nation, or a lot of nations, we get comfortable with controlling access to things Like that's your security basis as you control who sees what, who can get what. With open source it becomes the application of technology more than the control of it. I think there are always risks with these, but it is just different. It's a different use of that model where you're not gonna control who sees the code, because there can always be bad actors that see code that you don't want to. But I think we've shown that that model of everyone having access to it, being able to contribute, produces more secure code with it. In the case of artificial intelligence, the challenge, I think, is different, but it's similar if you look at it in another lens, which is how do you control these models with it? How do you make sure that it won't answer questions on your example of, you know, creating bio weapons? But it will help you optimize your business for it. Just as much creativity will have to go into that as avoiding time or hardware loops. You know, attack errors and operating systems on it, and I think we have to pick an approach. It can either be control access to it, where you only have a few people that can see how the model is created and we all hope they get it right, or there are a lot of people that control how those models are created, how controls are put in place, and you're tapping into the creativity of many, with the downside being, if everybody has access to it, the controlling the thing, the access to software is in your control point anymore. And you know, I don't think one system is right or wrong. We certainly have seen a better outcome using open source models with software on it. AI is still so early right now. It is slightly different in application, but any technology I think we've seen people do incredibly good things with it and I think you will always have bad actors that will try to use it for other purposes or find ways around it. Then we just have to find the model that amplifies the good while constraining the bad. I don't think it's, unfortunately, I don't think it's that new with AI? I think this has been going on for a long time.

**Craig Smith:** 36:42

No, that's right. The one thing and maybe I don't understand how open source works. But if you open source a model and someone downloads the weights and the code base and sets it up on their own servers, it's no longer what they do with. That code is no longer visible to anyone if they don't want it to be. Is that right?

**Matt Hicks:** 37:16

Yeah, I think it'd be the question of you downloading an operating system. You can always take out code, but it only affects you. I think with AI it's you download a model and weights but can refine your way out of those constraints. Can you use it to apply to other things? It's new. There's certainly a risk of that on it because you're using that to apply to other things. But at the same point, if you look at hackers in the world today, they have their own scripts and kits that aggregate all the known issues that run and apply those to it. Does AI speed that up? Absolutely, but it wouldn't be like it doesn't exist today. What I would argue is, as long as we can use AI to help defend against it, more so than you're speeding it up, you probably end up in a better spot. But it is with any technology that amplifies how fast innovation works and plays out. I think there's always that double-edged sword. I saw it in the early 2000s with the onset of things like Linux. I think we saw it in our space with the application explosion of Linux containers how portable that made software in all of these different environments. I think we're seeing that again with AI, luckily in the past, I think it's been on the more good than bad side. I'm an optimist, so I think AI, with the right structure and creation to it, can have a lot more good than bad. But you always have to think through the bad and the malicious uses in technology.

**Craig Smith:** 39:02

Yeah, the other thing is open source. The open source community, of course, is open. Researchers and academics are by nature collegial, but in a national security setting or geopolitical context it may matter which country has the strongest models or what they do with models by open sourcing these large models, very large models, and certainly not on the scale of GPT-4, but every government in the world can now build a pretty powerful LLM and then take it wherever they want it to go. There was a rumor that Ernie bought it with dues. They have a 70 billion parameter and I think, 130 billion parameter version, that was built, that they started with LLAMA, that they didn't have to build this from scratch. Is that a concern at all or yeah?

**Matt Hicks:** 40:34

I think at any point we rely on who can access technologies through some regulatory aspects to export control or those areas. On the other side, though, I think a lot of times with technology once the genie is out of the bottle so to speak, we're more responding to that existing. It is there at this point. However, we got to this point. There's a lot of power that is out there. Learning how we apply it well, learning how we make sure it evolves safely. From this point, I think, has to be the focus. We could put a lot of energy into how you put this genie back in the bottle, but it's not going to go away. Whether the LLAMA or LLMA2 or the hugging face models are right or wrong, I think being able to amplify good out of them is important on it, even if you're going to have bad actors that are going to try to use those for malicious means on it. I think that will happen at any boundary you want to come up with, whether it's country boundaries, state boundaries. Within that, I think you will always have malicious use. But there is a lot of power as well. Take like a network attack on it. Ai models trained to defend against network attacks are going to be able to respond faster than humans on it. My hope is, as long as we are sharp on the application side, you can end up in a more secure state To your point. I think it is always a risk with technology, it's always been a risk with material technology, it's always been a risk with computing technology that you have to get that balance right. But all that said, if you go back to the open versus the proprietary method of doing that, there is the argument of how important is that pace of innovation and sort of where do you want to apply? Because I do think we've seen in an open source model with software, if you say we can get the regulations and the constraints right and we can be on the balance of good more than the balance of chaos, and technology innovation can have a pretty profound impact, I think, for the better, and I think we've seen that the open source models can bring a lot of that innovation whether it's from academia or it's from just incredibly creative people around the world to contribute to something common. So it may not play out in AI because of geopolitical risks or those things, but I think the core potential is there if we can get those constraints right.

**Craig Smith:** 43:47

Yeah well, I'm feeling a little guilty because I've made you talk about all the stuff that I wanted to talk about. So tell me where you're going with Red Hat. What's happening with Red Hat today?

**Matt Hicks:** 44:01

Yeah, so for us, I think we've spent a lot of time. As we would say, how do we get developer code from a laptop into production faster? On it, we put terms like DevOps on it and we have a product line called OpenShift. That's really been our passion. It's distributed computing. How do we take all of these computing resources and then let developers use them? Because we think that creativity on the good side is there and, especially for enterprises, we think there's a lot of value with it. We've seen this. It was funny for me. At the Summit this year we had about 10 announcements. We've been working in the AI space for a while. I assumed it was yeah, we have OpenShift AI, which is tuned to running AI models. As we planned the Summit, I was expecting that it would be priority 10 on our list. We worked for, I would say, almost eight years at this point in the open source communities of evolving Post-Chat GPT. That was the only thing that we talked about in the press with customers, anything. But it is a really neat space to say these models that you're going to build, whether they're answering, they're a chatbot for support or they're making a breaking decision for a car and are very specialized, you're going to be running them side by side, the applications that you write. And there are two things I've sort of learned and suffered. One is things that are magic, that you can't incrementally change and debug and understand what change, don't go far in production. They make for really impressive demos. They don't run on it, and so we've been really passionate about the plumbing aspect of how do you not just train a model once that does cool parlor tricks for you, but as your data changes, how do you continually refine it to make it more accurate? How do you take a team of data scientists that can create a model but then be able to deploy it next to an operations team that knows how to run Linux, and how can you do that over and over and over again like you would modern applications, and so for us, that's where we spent a ton of time with OpenShift AI, the same platform of how can you move an app and a model side by side. How can you get data scientists teams working next to development and operations teams and unlock that incremental benefit? Because, again, while we talked about some of the big broad issues, I think we could dive into copyright and trademark implications. There are a lot of big topics that will be tough to understand what's right or wrong. But the thing we know in the short term is that there is that incremental value to unlock, not just in these massive models but in specialized areas, and so we've put a lot of time into making that within reach of mere mortals in the enterprise, because there are a lot of model options out there. How can they get them, refine them, deploy them to production and see if they can do a task more accurately than they were able to a few months ago? So it's a fun space. It is. Saying it has exploded is a bit of an understatement these days. But that's technology for you. You don't always get to guess what the next trend or change is going to be there.

**Craig Smith:** 47:46

Yeah, and has yours? Has Red Hat's role or road map changed at all since the IBM acquisition?

**Matt Hicks:** 47:58

Not really at all, to be honest, when IBM and it's funny, I'm a very long boomerang back to IBM. I started my career at IBM and consulting and about 18 years ago I left and joined Red Hat and then I came back at the acquisition time and at the time Jenny and Arvin were really clear we'll run you all independently on this. And the reason for that was pretty clear: Red Hat was a platform company. We want to make sure we underpin all the applications on top of all the hardware we can, which means we have to run on IBM's cloud and hardware, but we also have to run on Amazon's and Azure and Edge hardware and be ubiquitous with it if you want to be an effective platform, and so our mantra, really our goal, is being able to deliver that infrastructure software for what we call the open hybrid cloud. It's open source, it runs anywhere and it gives you cloud-like efficiencies as a developer whether you're running on-prem or in a public cloud with it. So that theme and priority hasn't changed with IBM. We've worked really closely with IBM in some exciting areas. They actually trained a model for Ansible Playbook generation, which you can think of as a chat. Gpt can certainly generate Ansible for you, but there are a couple nuances, like large language models, that tend to do it differently pretty regularly. That's how language works. If you asked me to explain something twice, I would try to do it differently to make it stick With coding. That's not always what you want. You sort of want to ask me 100 different ways and I give you the same most efficient line of code. So IBM is in the model space and they've worked really closely with us in some areas, like being able to bring a lot more people up to speed with Ansible knowledge in that. So we work incredibly closely with them. But our goal is still sort of lay down the open hybrid cloud everywhere, from Edge to public cloud to data center, so that the power of developers, they can learn that platform and apply to cloud native apps and to AI models, whatever the next generation of applications becomes.

**Craig Smith:** 50:28

Yeah, yeah. Well, that's fascinating. Describe how Linux and Red Hat works with an enterprise. Yeah, I mean, where are they deploying Linux on their servers to run their operations and you're advising on how to do that? Or are they just?

**Matt Hicks:** 51:04

Yeah. So I would say we can think of it in a couple different layers of our platform line. So Linux, to your point, a company is going to go make a purchase decision of a lot of servers and they want an operating system that will boot on it, but then also a vendor that's going to keep it current as they add or optimize that hardware, and that's sort of the rel layer. You got a thousand machines, you can boot them all, and if anything goes off, the company's call is where the single throat to choke or work with Intel, dell, hp, whoever, to make sure we can get that server back to an optimized running state, openshift tends to come in. When then the customer says I want all thousand of these boxes to act like one unit for it, which we call distributed computing, and that pulls in not Linux now, but technologies like Kubernetes, and it's the same thing. These technologies. They change so fast. In three years about 98% of Kubernetes is rewritten, and so we employ a lot of engineers to keep up with that. But we give customers updates to let them keep staying current without having to be as steeped in the technology. And then, if you look at a layer like OpenShift, it would do all those same things. I'm really interested in the GPU optimization for training or potentially in the inference side for it. And so again, it's distributed computing stacked on Linux and Linux's ability to talk to GPU, but now you're farming out AI training jobs on it, and so that really is where we play on both providing that innovation at a pace that enterprises can consume, where it just works for them, and these are thousands of integration points on it where the software is available to them. But it's a tremendous amount of work to make it work consistently across hardware variants, and we play in there, and then if it breaks, we're knowledgeable about it, we fix it and then if you fall off the latest, we'll continue to support you on life cycle, so that you've trained your model and maybe you don't want to do a big update. You want to keep OpenShift AI a little behind. We have longer life cycles so you can stay secure without consuming all the new features of the next drop. Does that make sense? That's what we do, and so if you're a bank, we want you to focus on banking, or our airlines on bookings, and then we focus on that plumbing layer to let developers really focus on their business domain.

**Craig Smith:** 53:52

Yeah, and what do I mean? Certainly in the early days that was a cost decision. You know, rather than going with a preparatory OS, Is it still primarily a cost consideration?

**Matt Hicks:** 54:12

I would say, and it's funny for me the shift from rel to OpenShift probably moved from cost to innovation, where all of a sudden it became. I mean, you look at the kernel, it's tens of millions of lines of code just in the kernel. Then you add a distributed computing layer on top. It's a lot to learn. It's the same story I had as a consultant. It's all available to you. You can go and really be an educated user. But having command of that is extremely expensive. And so if you want to move fast and use these technologies to move fast, I've seen that customers shift now towards using Red Hat to innovate faster because we need some stable foundation layer and it's not to your point. You know in the early days it was Unix to Linux takeout cost on it. So you know it's been fun to see that shift. It's a different game where you're pushed to go faster and faster and faster and anyone can step off that train and they want to be stable. But I think that's a more productive space to play in the innovation side there.

**Craig Smith:** 55:20

I also want to encourage you to visit net suite.com slash I on AI for an unprecedented offer by the number one cloud financial system you can download a custom KPI checklist designed to give you consistently excellent performance. It's absolutely free at net suite.com slash I on AI. That's net suite.com slash EYE O N A I all run together. So go to net suite.com slash I on AI to get your own KPI checklist. I hope you'll support them because they're supporting me. And remember the singularity may not be near, but AI is about to change your world, so pay attention. That's it for this episode. I want to thank Matt for his time, as always. 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 O N dot AI. And remember, the singularity may not be near, but AI is already changing our world, so pay attention.