Jillian Forde (00:13):

This is the AWS Podcast. I'm your host for today, Jillian Forde. And this is for all the Kubernetes lovers or the people who want to use Kubernetes, and particularly if you're interested in doing it for data analytics or machine learning, this is the episode for you. I've got two people here who know a lot about this topic and there's a really cool project that we're going to be talking about. But first, let's do some intros with these fabulous guests that we have. Vara, why don't you start?

Vara Bonthu (00:45):

Yeah, thanks Jillian. Hi everyone. My name is Vara Bonthu. I'm a principal solutions architect, mainly focused on open source and I'm also a technical lead for Data on EKS, which we are going to talk about today in this podcast. And yeah, I'm specialized in data analytics for 10 plus years and would love to have a chat with you guys today.

Jillian Forde (01:08):

So exciting. And Alex.

Alex Lines (01:10):

Thanks. Hey, everyone. Yeah, my name is Alex Lines. I am a senior container specialist at AWS. I'm Vara's partner in crime of sorts. I lead business development for the Data on EKS initiative. Been at Amazon for a little over seven years, been at AWS for about three and a half years. And I also am recently an ambassador of the Data on Kubernetes community.

Jillian Forde (01:30):

Wow, that is a super cool title. Both of you have super cool titles. Before we get into what the project is, I know we've got a lot of listeners who are new to AWS. And so for those who are interested in running data and machine learning workloads on EKS, can we walk through what are the options for them?

Alex Lines (01:49):

There's AWS managed services, right? There's SageMaker and there's EMR, there's Glue, there's all these managed services, but there are a number of customers that have decided that they are going to use Kubernetes. That could be because of a value for the open source community. It could be because they have some kind of hybrid requirement. We run in the cloud and on premises. So for a number of reasons, there are customers who have decided that Kubernetes is their option. And we at AWS, we can help them. So Amazon EKS is our managed Kubernetes offering and that is going to help with building and managing Kubernetes applications because there's a managed control plane.

(02:27):

So we handle control plane scaling. We also manage integrations with other services to make it easier to get up and running. And then specifically, the project that Vara and I work on, the Data on EKS project, helps customers go and build these key workloads for data for ML, the things that we've seen be most popular with our customers. It helps them go and build them in their environments.

Jillian Forde (02:52):

Wow, so cool. So I love that little teaser that you just gave us about the Data on EKS project. So tell us a little bit more about this project, what it is.

Vara Bonthu (03:02):

Yeah, sure. So Data on EKS Project, it's to start with everyone. It's a fully open source project. So it's mainly for customers who want to run data ML workloads on Kubernetes. They could use Data on EKS project from the GitHub and as a website with the documentation. It comes with a set of blueprints, and the blueprints are written in Terraform and CDK. And these blueprints will help you to build Spark cluster or a Flink based cluster or any other ML based clusters on EKS. And it comes with all the best practices with the Blueprints, which means how to scale some of your data workloads on Kubernetes.

(03:46):

And those best practices are kind of built in addition to the... You have a bunch of examples, which is part of this Data on EKS project. And these examples will help you to run the actual jobs in Spark, Flink or even ML jobs on Kubernetes by leveraging various instance types. It comes with a lot of resources and benchmarks, which helps you to actually kickstart your journey when you are moving to Kubernetes or Amazon EKS.

Jillian Forde (04:17):

That sounds super useful and I'm sure probably people are wondering really, why is it that we actually created this Data on EKS Project? So maybe Vara, can you go into that?

Vara Bonthu (04:28):

I'll tell my experience working with the customers and probably that might help. When I started to work with a few customers and I got a call from the customer saying, "Hey, we are thinking about migrating our Spark workloads, our Flink workloads onto Kubernetes and I know we have a bunch of open source resources and we try to actually deploy this, but we hit with a few issues and could you help us?" And that's the kind of a request that I received. And when I talked to these customers and I realized that... So everybody starts from the journey from the internet, how to deploy [inaudible 00:05:07]. But the key issues are how to run at scale and what kind of configurations you need to use to work with other AWS services mainly for the integrations and so on.

(05:18):

So, personally helped some of these customers to scale their workloads on Amazon EKS. Most of these questions are pretty common across all the customers. That's when I realized that, okay, so is there anything that we can build for these customers so that everyone can benefit from? People might be in different journeys and somebody who might be starting from the scratch or somebody who's middle of the journey, but the idea from this project is they can actually benefit from wherever they are. Since we started, we have 100+ customers using it and thinking about and in various ways.

Jillian Forde (05:56):

Wow. This sounds really helpful and I've got a question that I know is probably on a lot of people's minds. There's someone here who's probably listening, they're like, "Wait, but I don't use Terraform, so can I actually use the Data on EKS project?

Vara Bonthu (06:09):

That's a good question. A lot of people will say, when we opensourced this project, Data on EKS, and the Blueprints were created in Terraform, mainly because Data on EKS is like, are the project evolved from EKS Blueprints? I don't know if some of the audience knows about EKS Blueprints. We did this because we noticed a lot of customers who are using Terraform so that we can quickly help them, but we also

have blueprints in CDK as well. So coming back to your question, yes, you can benefit irrespective of the infrastructure as a code tool that you're using today, maybe Pulumi or maybe some other crossplane or other tools, you can use it. But Data on EKS will still give you how we implemented this solution that can be scaled on Amazon EKS. And it also gives you an examples of running Spark or a Flink or PyTorch or maybe LLMs with the Gen AI on Kubernetes.

(07:03):

So as a DevOps, and most of you know, the infrastructure as a code template can be easily switched to another template. I can use Terraform to write a Pulumi and use CDK to write Terraform. Answer is yes, you can. It is still helpful, because I hear from a lot of the customers who have already built their cluster and they come to us saying, "We need some help around the best practice of running Flink on EKS." So we have a cluster we don't want to deploy it, and we share these best practices and how they can actually modify their existing cluster setup by leveraging new tools like Karpenter as an autoscaler, which is a new thing for... Customers might be using an existing management grouped with a cluster autoscaler, but they can start thinking about Karpenter because we have some examples. It shows how to use Karpenter with various data and ML frameworks.

Jillian Forde (07:55):

Really cool. So I've got another question, and this one's for you Alex, because I know there's probably someone here who's listening and maybe they haven't used Kubernetes for their data or machine learning workloads and they're probably thinking, "I used it before for backend applications, but for data and machine learning, is that something that people use it for?" But you've seen that there are organizations that are actually using Kubernetes for these workloads. So why is it that these organizations aren't actually using Kubernetes for their machine learning and data workloads?

Alex Lines (08:29):

Yeah, absolutely. It's kind of two parallel stories going on here. When a customer starts to use Kubernetes and they start to adopt it across the organization, they're building tools like CI/CD that support it. They're building in their governance policies, they're building observability, they're maturing their Kubernetes platform, a lot of the time is how we see it. They're building it and the first things that they onboard there are stateless web applications because when... Or stateless applications, rather, like web applications.

(08:58):

And when Kubernetes was released back in 2015, that's what it was designed for, stateless microservices. And as the project has developed, more and more organizations and people in the community have wanted to use it for stateful applications. And we saw feature releases, like custom resource definitions, and that's what allows for operators for these different frameworks. So now you can start to use Kubernetes to manage Spark applications. We also saw GPU support come in 2019. That's a key thing that's going to unlock those ML workloads.

(09:32):

So as organizations have been out building their own practices for managing Kubernetes, the Kubernetes project has also been advancing and opening up the possibilities to run these stateful applications, which is something that is becoming more and more popular and ever-growing really. And quick shout out to the Data on Kubernetes community. So that was created in 2021, recognizing this groundswell of, hey, people want to use Kubernetes to manage stateful applications. How do we start to advance the techniques that are out there? It's becoming more and more popular.

Jillian Forde (10:05):

So interesting. And one example to that is the popularity is the SageMaker team seeing that popularity. And now that's the support for... Well, not now. It's been some time now, but support for Kubernetes operators and really being able to support MLOps with SageMaker as well. So yeah, I definitely see that it's really something that's more widely adopted now.

Vara Bonthu (10:29):

You're spot on there. Yeah.

Jillian Forde (10:32):

Vara, we got another question for you. From your experience of working with customers really closely on using the Data on EKS project. Have you seen that, does an organization have to adopt Kubernetes entirely in the company? Because I know there's probably some people here who are working at these larger companies, and I know you've worked very closely with a lot of similar situation. Does the entire organization has to adopt Kubernetes in order to use it?

Vara Bonthu (11:02):

That's a good question, right? I think many customers asked about this. I would say Kubernetes is very powerful. It helps you to deploy any type of workload. And today, now we are talking about data workloads, and nobody ever thought of it. When Kubernetes started, we only... They talked about running, they thought about running microservice workload like in applications, and now we are running data and ML workloads. But it's mainly around what customers want to do and how familiar with the Kubernetes. If they have a few applications in data which they wanted to achieve, like data processing on a nightly batches. And we have a number of managed services which are available like, I would say Glue, EMR, and a few other managed services which they can leverage without having to build the complex Kubernetes architecture.

(11:54):

But if the organization's already adopted Kubernetes, I know there are many customers moving on to Kubernetes, and it's a perfect place by leveraging Data on EKS, you can move the workloads to Kubernetes, even though you are full 100% on Amazon EKS, it's still having that balance of deciding on which workload will run on Kubernetes, which is not. I have simple applications.

(12:19):

I can even run it on Lambda. I wouldn't need to run it on Kubernetes. So just thinking about those things and a few applications. I can run it in a Glue, which is a lot cheaper and easier rather than managing Kubernetes. So really, we can help you decide that. But yeah, think of that perspective.

(12:36):

Another key point, and I think one customer I talked about, they have multiple teams within the organization. There are a few teams migrated to Amazon EKS, and other teams are thinking, "Hey, we want to migrate to Kubernetes as well, but we have centralized platform team who facilitate this, but we need help on that." So those are the best use cases. So there is a team already gone to existing team running these workloads on Kubernetes with a centralized platform. You can leverage that and start thinking about migrating yours because Kubernetes provides this multi-tenancy, further optimize the cost to use the same EKS cluster to accommodate multiple teams securely, which normally not the case when you use a typical Hadoop clusters.

(13:22):

These teams have to create that dedicated Hadoop cluster for every single team. But with the Kubernetes, with Amazon EKS, they can use one cluster and all the teams can work on that. So it's mainly balanced based on the workloads, and you don’t have to move everything Kubernetes.

Jillian Forde (13:39):

I love that answer and it really applies to any person here, regardless of what stage of business they're working at. I hear this concern a lot, even with startups where they're just getting started and they're afraid of making the wrong infrastructure choice and it's going to be so difficult. And then once they make this choice that they can't use Kubernetes, "Oh, we can't use Lambda, we can't use ECS or whatever it is." I love what you just said there. That's really what the beauty of AWS is, is that you can mix and match all these different types of workloads and different compute options based off of your use case.

Vara Bonthu (14:14):

Correct.

Jillian Forde (14:15):

So, another concern that some people have is they say, "Okay, this sounds great." And they've been probably planning a migration now for Kubernetes for a while, and maybe they've got some concerns. So I'd love, Vara, to hear from your perspective, because I know from all the customers that you've worked with, really, what advice do you have on some of those challenges that you see them encountering in these migrations?

Vara Bonthu (14:39):

Absolutely. I think most of the requests that I'm attending, or within Data on EKS team is that attending with the customers are challenges related to scaling. Kubernetes is highly scalable.. So the reason they adopt Kubernetes is it's highly scalable. You can optimize costs and so on. But how do we achieve that? The customers say, "We tried to run our workloads, but we couldn't get it working. It's not scaling. We are hitting some network issues," and so on. So the common challenges that we have seen is scalability. And the scalability is due to the incorrect network configuration, which is IP exhaustion. Within Amazon EKS, you have, when you deploy a workload and every pod takes one IP address and you have limited IPs within that VPC.

(15:25):

With the Data on EKS, we try to solve those problems by leveraging secondary CIDR ranges to attach to your VPC and using custom networking. These customers build these EKS clusters and running the workloads and then middle of the production, but then they reach out to the AWS saying, "Hey, we ran out of IPs. How do we extend the EKS cluster so that we can drop more workloads?" We wanted to actually avoid those problems and help these customers to build from the scratch. When they're thinking about migrating, "Hey, we are going to resolve all those problems when you look at our blueprints. We're going to help you solve that."

(16:05):

Other issues related to storage, like many customers asked about, "I know we can use data and ML workloads and how do we use FSx for Lustre? How do we use EFS? How do we use EBS? When to use these storage volumes with ML workloads?" We try to answer a lot of these questions, like say, take

an example of Spark. We have a lot of customers running Spark and Amazon EKS today. And they say, Spark does a shuffle data, which means it has to store the data on a... It requires a low latency shuffling.

(16:40):

So the question is mainly, "Should we use EBS volumes or should we use SSD volumes that comes with EC2 instances?" We've answered those things as a part of these blueprints again. Coming back to the ML stuff, talk about some of the latest LLM models, like BERT large model, Google's LLM. We shown that with one of the blueprint. How do you actually leverage FSx for Lustre as a high throughput file system to do the distributed training with their FSx? So that configuration, probably for some users might take some time to figure out how do we actually make these parts to connect to these FSx for Lustre and leverage this?

(17:21):

These are the common challenges that we noticed, and we built all of those challenges into solutions within Data on EKS today.

Jillian Forde (17:32):

Wow. So cool. So we've got templates for analytics, we've got templates for AI, including GenAI. This sounds amazing. So we've got time for just one more last question and then really where people can go to actually start using this. So I'd love for both of you to share one piece of advice based off of your experience of working with customers on the Data on EKS project. So Alex, I'd love to start with you.

Alex Lines (17:54):

Yeah, I think that when you're getting started, it's really important to... Well, I'd take a step back. If you're already doing this or if you're looking at doing a new project, I really think that our Blueprints are a great starting place. Whether you can take that sample code and use it or whether it's something where you're using that as your textbook guide to follow and it's like... Vara gave you the example earlier of you may use Pulumi instead of Terraform. Look at what we've done. We've built our Blueprints based on working with customers and helping them get to production and scale and build at a pretty massive scale.

(18:30):

So I would really encourage whether you can use the sample code or whether it's something that you're following as your guidebook or roadmap. There's a lot of good information in there and I would encourage anyone to go and find those templates and let us know if there's things that you don't see there. We have an open source GitHub roadmap and also GitHub issues where you can let us know, "Hey, I would like to see Blueprints for this, and I don't see anything yet."

Vara Bonthu (18:54):

I think one piece of advice to the customers are... Like I mentioned, we have two types of customers. And customers might be thinking, "We are already running Spark and Flink and other workloads on EKS. How does the Data on EKS is going to help us anyway?" And it's take a look at what we already have and then see whether it's meeting... Are you doing the same way what we have built here? You can always learn something from this and take weights out of it. And if there is anything that you want to know, even though you're running at a scale, raise an issue, like Alex said, we're going to take a look at it and speak to you. And happy to help you and solution architects within the team.

Alex Lines (19:32):

Yeah.

Vara Bonthu (19:33):

But if you are a newcomer, wanted to think about migrating to Amazon EKS, your workloads, definitely start with the Data on EKS, reach out to a solution architect. We will come and help you. We can help you customize how you can build that. In the past, we helped the customer to save months of the time figuring out these things. So, we are hoping that we can help you to complete your POC in weeks with this Data on EKS project.

Jillian Forde (20:01):

Super cool. I know a lot of people are really excited to start using it. Data on EKS project. Where can they find it?

Alex Lines (20:08):

There's two main places. So we have a public facing website where we have all these blueprints and that has the reference architectures and tutorials and a lot of that more descriptive information. So the website is the main place. And all of the patterns are on an open source GitHub repository, which we link back to. So those are the two main places where you can follow the Data on EKS project and see the things that we're releasing.

Jillian Forde (20:32):

Super cool. And we'll have a link to those as well. Alex, Vara, this was super exciting. I know a lot of people I'm sure are really excited. I'm excited for them. So thank you so much for being here on the AWS Podcast.

Vara Bonthu (20:44):

Thanks Jillian. Thanks Alex.

Alex Lines (20:46):

Thanks Jillian.