Transcript

00:00:00 Interviewer

Yeah, so I'm recording this session after getting your permission, OK?

00:00:06 Interviewer

Thank you. So can you please introduce yourself and describe your job role in this company?

00:00:13 Interviewee

Yes. So I'm " " and chief data scientist.

00:00:20 Interviewee

My role involves.

00:00:24 Interviewee

A lot of things, for example.

00:00:26 Interviewee

I have to do.

00:00:27 Interviewee

Business development I have to design algorithms.

00:00:32 Interviewee

I have to come up with the mathematical models.

00:00:38 Interviewee

Decision of architectures and.

00:00:43 Interviewee

Choosing the cloud components and services.

00:00:48 Interviewee

Sometimes I have to do hands on implementation of models, but rarely.

00:00:54 Interviewee

Then my role also entails taking into hiring new people, talking to new clients.

00:01:02 Interviewee

Coming up with proposed ML models.

00:01:05 Interviewee

To improve their process.

00:01:07 Interviewee

In the company.

00:01:08 Interviewee

So a lot of.

00:01:09 Interviewee

Stuff is going.

00:01:12 Interviewee

It's like a leadership.

00:01:16 Interviewee

Strategy and at the same time.

00:01:20 Interviewee

And on algorithm design.

00:01:22 Interviewee

Architecture, design and model maintenance all at one place, but.

00:01:29 Interviewee

I have a good.

00:01:29 Interviewee

Team to support.

00:01:33 Interviewee

Most of the functions they.

00:01:35 Interviewee

You know, handle.

00:01:42 Interviewee

And I am therefore.

00:01:45 Interviewee

Just being responsible for the eventual projects.

00:01:51 Interviewee

That would be my role.

00:01:55 Interviewee

In this company, with my current company.

00:01:56 Interviewee

I've been working with four years.

00:02:04 Interviewee

My master thesis was around unsupervised learning image processing.

00:02:10 Interviewee

So that was the thesis.

00:02:19 Interviewee

About my current experience.

00:02:25 Interviewee

In this position I was able.

00:02:27 Interviewee

To count the forecasting model.

00:02:31 Interviewee

That I developed with my team and that forecasting model was later converted into a product.

00:02:38 Interviewee

Served on AWS.

00:02:44 Interviewee

We use AWS cloud services.

00:02:46 Interviewee

To automate the model and mostly.

00:02:48 Interviewee

Of course it is a web service.

00:02:50 Interviewee

That is consumed like a matter of service by applications like Microsoft Dynamics.

00:03:00 Interviewee

Most of the cloud providers also suggest that an ML model should be a microservice.

00:03:09 Interviewee

They allow for.

00:03:11 Interviewee

Making a highly scalable web service so that.

00:03:15 Interviewee

If there are multiple applications sending their data for prediction.

00:03:19 Interviewee

The service can scale and it can respond to many requests in a span of 1 seconds.

00:03:27 Interviewee

That's how that would, lasting went. We were also able to make a pricing engine for used cars that was again deployed on.

00:03:38 Interviewee

Microsoft Azure is using again web service.

00:03:44 Interviewee

Model as a web service which is microservices based architecture and the consumption layer was.

00:03:52 Interviewee

Web applications. Web applications used to collect data send.

00:03:55 Interviewee

It to the microservice price prediction the price.

00:03:58 Interviewee

Prediction will come.

00:04:00 Interviewee

From that microservice to the web browser again and Web browser will.

00:04:05 Interviewee

The results so that is 2 projects.

00:04:10 Interviewee

We also have.

00:04:13 Interviewee

Projects where ML needs to be embedded into the app, not.

00:04:21 Interviewee

Like in a distributed architecture, you need to have either a microservice based architecture or a client service architecture. But if you don't want to make distributed.

00:04:30 Interviewee

You can also just embed.

00:04:32 Interviewee

It and do the main application so.

00:04:36 Interviewee

In my previous company, your question is.

00:04:38 Interviewee

That how is your old experience? So we used to.

00:04:42 Interviewee

Apply ML and AI on.

00:04:46 Interviewee

Embedded boards.

00:04:49 Interviewee

And in embedded.

00:04:52 Interviewee

You don't want to, you.

00:04:53 Interviewee

Know call a third party service.

00:04:57 Interviewee

So ML is not distributed over there, So what you do is you embed your.

00:05:03 Interviewee

Pipelines grid as a function on top of the embedded engine.

00:05:10 Interviewee

All the ML AI is executed on the same.

00:05:15 Interviewee

Single board computer Dr unit is battery operated and.

00:05:21 Interviewee

Of course.

00:05:22 Interviewee

Editing it and changing the model parameter is a little trickier than.

00:05:27 Interviewee

If you have a.

00:05:29 Interviewee

Cloud based microservice based ML model but.

00:05:33 Interviewee

It does the job.

00:05:35 Interviewee

It comes under edge computing, so in edge computing.

00:05:39 Interviewee

Sometimes you prefer to process data on the edge rather than sending it to the cloud, so in those cases then microservice on this.

00:05:47 Interviewee

Architecture is not suitable but some.

00:05:50 Interviewee

Edge devices are such that they capture the data or image and then they send it to the cloud.

00:05:55 Interviewee

Just like you know.

00:05:57 Interviewee

Google, Lance, Google. Let's capture the image and then send it to the cloud and.

00:06:03 Interviewee

Size device.

00:06:04 Interviewee

Mobile phone.

00:06:07 Interviewee

We send it to the cloud so you know.

00:06:09 Interviewee

In Edge devices.

00:06:12 Interviewee

Which is.

00:06:13

You know.

00:06:14 Interviewee

Small hardware machines like mobile phone.

00:06:17 Interviewee

Single card computers.

00:06:19 Interviewee

You have a chance to not use.

00:06:23 Interviewee

Sometimes it is better to.

00:06:24 Interviewee

Not use distributed architecture and.

00:06:28 Interviewee

You just need to convert ML model into.

00:06:30 Interviewee

A function and then call it inside your application.

00:06:36 Interviewee

Then question #6 is.

00:06:41 Interviewee

Is my company.

00:06:42 Interviewee

Service based or product base we have.

00:06:43 Interviewee

A couple of products.

00:06:46 Interviewee

But mostly our business is service based.

00:06:48 Interviewee

And even the products that we have.

00:06:52 Interviewee

Aim to sell our services.

00:06:54 Interviewee

On top of that.

00:06:55 Interviewee

Product so.

00:06:57 Interviewee

We are mostly service based.

00:07:00 Interviewee

Hmm. Uh.

00:07:03 Interviewee

Software development model that we practise agile we.

00:07:07 Interviewee

Do some really.

00:07:10 Interviewee

Like in AI, we are not able to do waterfall in data science so.

00:07:15 Interviewee

It has to.

00:07:15 Interviewee

Be agile and there are so many iterations and.

00:07:18 Interviewee

Mostly some experiments.

00:07:23 Interviewee

Evolve in a Kanban fashion where we plan something and.

00:07:26 Interviewee

After a week, we decided that the rest of the.

00:07:28 Interviewee

Backlog was not.

00:07:30 Interviewee

Right. And we have to drop.

00:07:32 Interviewee

All the tickets.

00:07:35 Interviewee

So agile sometimes Trump sometimes Kanban that.

00:07:38 Interviewee

Is the model.

00:07:44 Interviewee

Yeah, the the interesting project that I.

00:07:46 Interviewee

Have worked on was.

00:07:50 Interviewee

The forecasting model for.

00:07:53 Interviewee

Predicting the sales of.

00:07:58 Interviewee

Product categories and.

00:08:02 Interviewee

Google Trends data using social mentions that are to predict.

00:08:08 Interviewee

Changing scenario.

00:08:09 Interviewee

Of pandemic, and you're also fortunate enough.

00:08:13 Interviewee

To use the.

00:08:15 Interviewee

Retail mobility data set, which is released by Apple.

00:08:19 Interviewee

That data shows how many people are going to read us first current times.

00:08:24 Interviewee

And that has a very good signal about how the sales is going up or down based on people's visit to the retail locations.

00:08:37 Interviewee

That would that would be the interesting project.

00:08:42 Interviewee

#9 is.

00:08:44 Interviewee

How many software architecture design techniques have you worked with? So I was telling you that non distributed single application.

00:08:57 Interviewee

And an embedded device and then distributed.

00:09:00 Interviewee

I have worked with microservices based.

00:09:02 Interviewee

Architecture we do client service client.

00:09:05 Interviewee

Client server based architecture.

00:09:10 Interviewee

Only for demos but.

00:09:13 Interviewee

We don't do it for actual client production, so that would be client microservice. Single application not distributed architecture and.

00:09:29 Interviewee

The third is science that will be.

00:09:33 Interviewee

Only for small demos.

00:09:36 Interviewee

Production is always microservice.

00:09:38 Interviewee

If it is on edge then it is not distributed. Everything on the edge.

00:09:46 Interviewee

For demos we use client server.

00:09:49 Interviewee

You know, make it simple because we're just showing it to the client once or twice and.

00:09:53 Interviewee

Then we have to.

00:10:01 Interviewee

In software architecture.

00:10:02 Interviewee

Design you found being used in most companies through.

00:10:05 Interviewee

Your experience? Yes, of course it is microservices based architecture.

00:10:11 Interviewee

For multiple business departments and multiple business entities.

00:10:16 Interviewee

You want to make.

00:10:17 Interviewee

Multiple models and each model has to be a microservice so that your application can choose which AI to consult for each type of.

00:10:26 Interviewee

Prediction microservices suits very well. Another benefit of microservices is.

00:10:33 Interviewee

That modifying your model and then.

00:10:37 Interviewee

Releasing the new model is very smooth in.

00:10:41 Interviewee

Microservice based solution.

00:10:43 Interviewee

You know.

00:10:46 Interviewee

The transitioning of like retiring 1 model.

00:10:49 Interviewee

And replacing it with another.

00:10:51 Interviewee

Is very easy in microservices because microservice has a very low adhesion and.

00:10:58 Interviewee

The app will not.

00:10:59 Interviewee

Crash if the service is not available.

00:11:03 Interviewee

But if you have a single non distributed model and you want to change it, the app has to be uploaded app that has to be done.

00:11:13 Interviewee

You can't run an app.

00:11:15 Interviewee

If the model is down because it's a very strong and has a.

00:11:18 Interviewee

Component of your application.

00:11:23 Interviewee

That is the beauty of microservices. Easy upgrades, easy retraining.

00:11:28 Interviewee

A lot of choice.

00:11:30 Interviewee

Of models.

00:11:32 Interviewee

Will be called in different.

00:11:33 Interviewee

Sections of your application if your application.

00:11:37 Interviewee

Needs detailed prediction for long term and a good prediction for short term. You can make two models, one for approximate prediction and one for detailed prediction.

00:11:49 Interviewee

And both models are doing the same job one.

00:11:51 Interviewee

Takes detail into account. Another takes.

00:11:54 Interviewee

Speed inside out and.

00:11:58 Interviewee

Microservice will support you in that.

00:12:01 Interviewee

You just need to choose different services, no?

00:12:11 Interviewee

I would say micro service.

00:12:12 Interviewee

Is most used in most companies.

00:12:15 Interviewer

OK.

00:12:19 Interviewee

I've already answered I think #11 also yeah.

00:12:26 Interviewee

Do you have any recommendations, software architecture, design, techniques of machine learning systems so?

00:12:35 Interviewee

I would say that of course it should be based mostly on microservices and.

00:12:42 Interviewee

The software components other than the ML.

00:12:47 Interviewee

Should be such that we can.

00:12:50 Interviewee

See the expandability of models also.

00:12:54 Interviewee

So for example, if.

00:12:58 Interviewee

Like in Microsoft Azure they have this.

00:13:02 Interviewee

Assumption that our model service will only create.

00:13:07 Interviewee

Input data.

00:13:10 Interviewee

And then it will only return one value.

00:13:14 Interviewee

So what I'm saying is the recommendation is even Microsoft was not.

00:13:18 Interviewee

Able to.

00:13:21 Interviewee

Make it right.

00:13:24 Interviewee

And why do I say that there are models that send prediction and explain ability information also so.

00:13:31

If you have.

00:13:31 Interviewee

Heard of explainable?

00:13:32 Interviewee

AI in explainable AI, you don't need just a regression prediction or a classification prediction.

00:13:38 Interviewee

You also want to.

00:13:39 Interviewee

Have answers from the model that why the model thinks.

00:13:43 Interviewee

A certain regression value is.

00:13:47 Interviewee

So you know, in a decision tree you need to give.

00:13:50 Interviewee

All the.

00:13:53 Interviewee

All the variable values.

00:13:54 Interviewee

Are all the.

00:13:54 Interviewee

Tree nodes that resulted in that prediction, right?

00:13:59 Interviewee

In decision tree you need to know.

00:14:01 Interviewee

The path of prediction.

00:14:02 Interviewee

You need to know how the.

00:14:03 Interviewee

Decision Tree calculated the value.

00:14:05 Interviewee

So in Microsoft Azure web service.

00:14:09 Interviewee

Microsoft Azure makes ML web services right Azure Azure ML Services which allows you to make web services out of.

00:14:16 Interviewee

Your models, but they don't have the capability to.

00:14:20 Interviewee

Send you the recommendation, send you the explain ability, and they only want to send.

00:14:24 Interviewee

You one value.

00:14:27 Interviewee

And that value is prediction of a tree, right? Regression prediction of a tree. If you're doing regret.

00:14:33 Interviewee

So let's say I'm doing.

00:14:34 Interviewee

My customer lifetime value prediction.

00:14:38 Interviewee

And customer land value prediction is a regression value and it will send me regression value. It will not send me the information how it calculated that and how?

00:14:47 Interviewee

Model thinks why?

00:14:48 Interviewee

The customer lifetime value is either less or more.

00:14:53 Interviewee

If I want to use that model.

00:14:54 Interviewee

For a band.

00:14:56 Interviewee

State Bank will ask.

00:14:58 Interviewee

Me a question.

00:14:58 Interviewee

Of why your model is not telling us the reason this model is useless. We can't we.

00:15:03 Interviewee

Can't approve it?

00:15:05 Interviewee

So where is the fault? The fault is Microsoft did not me provide me the architecture of the web service which can return to me many things other than the prediction.

00:15:15 Interviewee

The recommendation is.

00:15:16

For instance.

00:15:19 Interviewee

Have a micro service platform.

00:15:22 Interviewee

That is designed for ML.

00:15:26 Interviewee

And which does not.

00:15:29 Interviewee

It does not stop you from sending.

00:15:32 Interviewee

Other than the prediction and.

00:15:34 Interviewee

Information other than the prediction information.

00:15:38 Interviewee

Can send you.

00:15:40 Interviewee

A lot of things.

00:15:41 Interviewee

For example, model can send you probability or confidence of a certain prediction.

00:15:47 Interviewee

It can also send you.

00:15:48 Interviewee

The overall ability information of a certain prediction. It can send you.

00:15:56 Interviewee

Your service should be able to send you the variable importance.

00:16:00 Interviewee

Of different features.

00:16:03 Interviewee

So if I was to design the microservice based model productionization tools at Azure, I would design the microservice.

00:16:12 Interviewee

In such a way?

00:16:13 Interviewee

That it doesn't only return one.

00:16:15 Interviewee

But it has to.

00:16:17 Interviewee

Allow me to.

00:16:18 Interviewee

Return multiple just like it accepts.

00:16:19 Interviewee

Multiple values in a.

00:16:20 Interviewee

Jason, it should return to me.

00:16:23 Interviewee

A lot of values in a JSON that I want to send back to my.

00:16:29 Interviewee

Application from the microservice.

00:16:47 Interviewee

What would be the best, best practise that is helpful in applying software affected designing?

00:16:54

So the practise.

00:16:55 Interviewee

Of designing the architecture.

00:16:59 Interviewee

It starts from the input output flow of the process business first approach. So look at the business.

00:17:06 Interviewee

What are the deadlines?

00:17:08 Interviewee

For a prediction, is it real time? Is it near real time? Is it mode business can tell you or you can you can study the business and you will know that even even if I start predicting at night and I have predicted the value till the morning, the business will like it or the business will be able to.

00:17:26 Interviewee

Once you know.

00:17:27 Interviewee

The model flow.

00:17:28 Interviewee

And the feasibility, then you can design your.

00:17:30 Interviewee

Model for example, you want to use a cheap.

00:17:35 Interviewee

Cloud service where you can send a batch.

00:17:37 Interviewee

Of JSON 10,000 jasons.

00:17:40 Interviewee

Are like the input features right and then.

00:17:44 Interviewee

Present those 10.

00:17:45 Interviewee

1000 basins to the model and it can.

00:17:49 Interviewee

It can process those.

00:17:53 Interviewee

You know iterative fashion one at a time. Take each one each each, Jason.

00:17:58 Interviewee

Takes 5 seconds so.

00:18:00 Interviewee

In one minute you will be able to.

00:18:03 Interviewee

Process 12 cases were 12 inputs.

00:18:06 Interviewee

And then in 60.

00:18:08 Interviewee

In 60.

00:18:11 Interviewee

Second, you will be able to process 700.

00:18:18 Interviewee

Inputs in one minutes, 700 in one minutes and then.

00:18:29 Interviewee

7000.

00:18:32 Interviewee

It's like.

00:18:35 Interviewee

Around 40,000.

00:18:38 Interviewee

OK.

00:18:40 Interviewee

Prediction in one hour.

00:18:42 Interviewee

So if you have.

00:18:44 Interviewee

100,000 predictions it will take you 2 hours.

00:18:49 Interviewee

100,000 predictions 2 hours.

00:18:54 Interviewee

So it's a pretty much.

00:18:55 Interviewee

A bad job that you can run and.

00:18:57 Interviewee

At the night and.

00:19:01 Interviewee

Resume in the morning.

00:19:02 Interviewee

But if the business says that, hey, I have.

00:19:04 Interviewee

A customer that wants to add to cart.

00:19:06 Interviewee

A product and you need.

00:19:07 Interviewee

To do the prediction of.

00:19:09 Interviewee

Shipping date or prediction of.

00:19:12 Interviewee

Best price or best discount?

00:19:15 Interviewee

Then it's very real time. They need the answer and like.

00:19:19 Interviewee

2 seconds. They don't want to wait on.

00:19:21 Interviewee

The browser for the.

00:19:22 Interviewee

Price, right?

00:19:23 Interviewee

2 seconds even 2 seconds is more. It has to.

00:19:26 Interviewee

Be a second.

00:19:27 Interviewer

And you can't.

00:19:29 Interviewee

You can't think.

00:19:29 Interviewee

Of a software.

00:19:31 Interviewee

You can't think of a service that will run only once in a loop.

00:19:35 Interviewee

At night you have to have a real time highly available microservice which is highly parallelized which can respond to at least 200 requests within one second.

00:19:46 Interviewee

You can make it like 10,000 requests 1.

00:19:48 Interviewee

Second, also by using.

00:19:50 Interviewee

Cubanet is based scaling of the microservice.

00:19:56 Interviewee

So I would say the best practise to approach a architecture is to see what are.

00:20:00 Interviewee

Your deadlines and how real time.

00:20:02 Interviewee

Your answer is required and how many requests your your model will face.

00:20:08 Interviewee

In any given minute or second.

00:20:11 Interviewee

And then design accordingly state accordingly and use Kubernetes if.

00:20:16 Interviewee

You if you know that.

00:20:18 Interviewee

The model will receive requests from multiple, for example a customer.

00:20:24 Interviewee

Add to cart and generate model.

00:20:28 Interviewee

All right.

00:20:30 Interviewee

And if you.

00:20:30 Interviewee

Have 10,000 customers.

00:20:33 Interviewee

Interacting with your browser in one hour.

00:20:37 Interviewee

Then your model has to be scalable.

00:20:40 Interviewee

Sometimes even more than 10,000 customers interact with your browser.

00:20:45 Interviewee

Ecom browser.

00:20:47 Interviewee

And they generate.

00:20:47 Interviewee

A lot of modelling, a lot of production balls, so you have to have your service on top.

00:20:52 Interviewee

Of people so that.

00:20:56 Interviewee

Highly available web service can be designed.

00:21:00 Interviewee

Single web service without.

00:21:02 Interviewee

Parallelization or without.

00:21:05 Interviewee

Multithreading and not answer.

00:21:08 Interviewee

Multiple customers in the timeline so.

00:21:10 Interviewee

You know parallelization.

00:21:15 Interviewee

Using the parallel hiding scalable web service.

00:21:19 Interviewee

Is the key and some most of the cloud providers provide that scalability in their web service. For ML, you just need to be mindful of it and design accordingly. Whether you're running in.

00:21:31 Interviewee

A batch mode.

00:21:33 Interviewee

So you can have simple.

00:21:35 Interviewee

Single VM and make service and then.

00:21:37 Interviewee

Iteratively take the.

00:21:39 Interviewee

Input file and processes all time or you need a very real time service that will receive request in an hour from different places of the world different customers.

00:21:55 Interviewee

For best practise.

00:21:58 Interviewee

Designing software detection OK.

00:22:00 Interviewer

Yeah. Thanks.

00:22:03 Interviewee

What are the most common software architecture design challenges?

00:22:06 Interviewee

And machine learning systems the design challenges.

00:22:13 Interviewee

Or for example, not being able to scale the model and some requests just.

00:22:20 Interviewee

Lost in the air if your service.

00:22:23 Interviewee

Is busy answering.

00:22:24 Interviewee

Or predicting 1.

00:22:26 Interviewee

Answer and somebody else also called the service and it's not scalable. So your request will just vanish.

00:22:33 Interviewee

So for that you need to use queues so that every API call goes into the queue and then your service can.

00:22:39 Interviewee

Fetch those requests.

00:22:40 Interviewee

From the queue.

00:22:45 Interviewee

Models vanishing.

00:22:46 Interviewee

In the air as a problem.

00:22:52 Interviewee

Model requests when?

00:22:55 Interviewee

Used can be used.

00:23:00 Interviewee

Has drift and as the model drift you have to have.

00:23:05 Interviewee

Right architectural.

00:23:10 Interviewee

To detect the model drift and then retire the model and.

00:23:16 Interviewee

Before that, you should have already trained another model for training. You already need to have architectural components for training.

00:23:23 Interviewee

That training should be triggered by the drift, so there will be an automated.

00:23:30 Interviewee

Detector of drift which will detect that the drift is coming and you need to start retraining.

00:23:36 Interviewee

And when the retraining is complete, we need to.

00:23:38 Interviewee

Slowly and gradually retire the drifted model.

00:23:44 Interviewee

Canary testing or some other method slowly and gradually.

00:23:50 Interviewee

Introduced the new model that is fresh.

00:23:54 Interviewee

And replace the previous one.

00:23:57 Interviewee

That is.

00:23:58 Interviewee

A challenge that needs to be solved, the drift challenge and it makes the model more complex. It makes the training pipeline and the prediction.

00:24:09 Interviewee

Web service design more complex because you are.

00:24:12 Interviewee

Making it it's like a.

00:24:14 Interviewee

Continuous integration of modelling right continuous.

00:24:18 Interviewee

Redesigning of modelling and that makes the modelling complex but at the same time it can be elegantly designed.

00:24:27 Interviewee

Through using modern best practises and modern cloud tools and services that are provided by big cloud providers.

00:24:27

To use.

00:24:37 Interviewee

Yes. So the architectural decisions that need.

00:24:39 Interviewee

To be taken again.

00:24:40 Interviewee

I would. I would go back to the input output flow of the model.

00:24:47 Interviewee

Where each data source is going to come.

00:24:49 Interviewee

Are these data sources?

00:24:55 Interviewee

Very raw and require a lot of compute, so use data breaks or some other.

00:24:59 Interviewee

Distributed compute engine into clients that are converted into speech features or.

00:25:05 Interviewee

Or do we have the data which is pretty much ready and you just need to pick?

00:25:09 Interviewee

It in Python up.

00:25:10 Interviewee

And, you know, converted into features and feed it to the model. So what kind of data we are, what kind of data sources are we having and how much transformation is required? And based on that we choose the ETL tool that can be highly parallel spark.

00:25:26 Interviewee

In a database cluster or it can be?

00:25:31 Interviewee

Simple Python pandas query from a database or.

00:25:37 Interviewee

Coming from a file that is somewhere in the BLOB storage.

00:25:42 Interviewee

Then after the input decision is taken.

00:25:48 Interviewee

Then you have to think.

00:25:49 Interviewee

About input, output flow and real time.

00:25:52 Interviewee

Nature of the architecture and if you need real time.

00:25:56 Interviewee

Of course you have to have highly scalable matching services.

00:25:59 Interviewee

You don't need real time. You can have a service that you can call fashion all night.

00:26:05 Interviewee

So that you know real time versus non.

00:26:07 Interviewee

Real time decision is a key decision.

00:26:10 Interviewee

And then how would you retire models and retrain models and when will you retrain it? It will be triggered or you will have a set cycle of three months to retire a model and.

00:26:19 Interviewee

You train new one.

00:26:20 Interviewee

Would you train the model every time you want a prediction like in most forecasting scenarios?

00:26:26 Interviewee

We retrain the model every time we.

00:26:28 Interviewee

Need a prediction so that needs to be.

00:26:29 Interviewee

And then.

00:26:32 Interviewee

For example, in a forecasting system that we make.

00:26:35 Interviewee

We enable a retraining every time.

00:26:38

OK.

00:26:40 Interviewee

When we need a prediction now you have to keep your retraining.

00:26:44 Interviewee

Service up and then your you know.

00:26:48 Interviewee

Your retraining service and your.

00:26:53 Interviewee

Infinite Server is the same service.

00:26:57 Interviewee

And the compute happens one after another, but in a non forecasting scenario where you don't need retraining a lot of times.

00:27:06 Interviewee

Then your training pipeline has a separate branch and your.

00:27:10 Interviewee

Prediction or incidence pipeline is a separate branch and both of them are independent of each other.

00:27:19 Interviewee

That needs to be decided whether you're training and retraining your training.

00:27:23 Interviewee

Incidence is a separate.

00:27:26 Interviewee

Pipeline the answer.

00:27:28 Interviewee

Is same box in which you retrain and then predict.

00:27:35

And in the end.

00:27:37 Interviewee

How would you evaluate drift? That is also our decision.

00:27:41 Interviewee

Because soon your model will not generate and will stop generating business value.

00:27:48 Interviewee

How would you?

00:27:49 Interviewee

Make sure that you generate the.

00:27:50 Interviewee

Biggest value and then one decision of optimal.

00:27:55 Interviewee

Cloud usage.

00:27:58 Interviewee

Machine learning production systems.

00:27:59 Interviewee

Are very new and.

00:28:01 Interviewee

What is the architectural style which saves the most cost in cloud?

00:28:06 Interviewee

Is very important and.

00:28:09 Interviewee

You have.

00:28:10

To have a.

00:28:10 Interviewee

Good idea of time complexity, space complexity and then how the models and services scale so that you can decide what kind of a compute is optimal for pricing and considerations while keeping the performance.

00:28:25 Interviewee

An acceptable boundary.

00:28:27 Interviewee

I think these some of the decisions that.

00:28:29 Interviewee

I just mentioned.

00:28:31 Interviewee

Need needed to be considered before you go on.

00:28:34 Interviewee

Your drawing board.

00:28:36 Interviewee

To make that.

00:28:43 Interviewer

Yeah, I think so. You're helpful are very beneficial. I also learned a lot from you, especially so much knowledge and so much experience and so much discussion about the different scenarios.

00:28:55 Interviewer

And you try to combine the different scenarios and different scenarios by two.

00:29:02 Interviewer

Which practises we have to do?

00:29:08 Interviewer

Yeah. So it was really helpful for me.

00:29:15 Interviewee

Perfect.

00:29:21 Interviewee

Is there anything else that you want to discuss?

00:29:25 Interviewer

No, this question this answers I want you to have it, so I later on I can compile it soon enough sooner or later and then.

00:29:33 Interviewee

OK.

00:29:39 Interviewee

And would you be able to send?

00:29:40 Interviewee

Me the video.

00:29:42 Interviewee

Just for my records so that I can.

00:29:44 Interviewee

You know somebody wants to learn about.

00:29:50 Interviewee

You know productionized models, you know, we.

00:29:54 Interviewee

Instead of me designing my friend.

00:29:56 Interviewee

And just use this one.

00:29:58 Interviewer

Yeah, sure. I will upload on the Google Drive or.

00:30:02 Interviewer

And then I will share the link with you.

00:30:03 Interviewee

OK dear.

00:30:04 Interviewer

OK, I will do that.

00:30:06 Interviewee

That'd be great. That'd be great.

00:30:08 Interviewer

Yeah, I saw your session on LinkedIn like you yesterday.

00:30:13 Interviewer

Yeah, you're quite into the machine.

00:30:15

Right.

00:30:16 Interviewee

Right, right.

00:30:20 Interviewee

Yeah, you know.

00:30:21 Interviewee

That's what we do.

00:30:23 Interviewer

OK, so I will just stop the recording for now.