EPISODE 669

[INTRODUCTION]

[0:00:00.3] JM: A stock trader has access to high volumes of information to help that stock trader make decisions about whether or not to buy an asset. A trader who's considering buying a share of Google stock can find charts, reports, and statistical tools to help with their decision. There are a variety of machine learning products to help a technical investor create models of how a stock price might change in the future.

Real estate investors on the other hand do not have access to the same data and tooling. Most people who invest in apartment buildings are using a combination of experience, news and basic reports. Real estate data is very different from stock data. Real estate assets are not fungible. Each one is arguably unique from all others, whereas on share of Google stock is the same as another share, but there are commonalities between real estate assets, just like collaborative filtering can be applied to find a new movie that is similar to the ones that you've already watched on Netflix.

Comparable analysis can be used to find an apartment building that's very similar to another apartment building which recently appreciated in asset value, and this comparable analysis, by finding traits that are similar in apartment building purchase opportunities, is the backbone of skyline.ai. Skyline.ai is a company that is building tools and machine learning models for real estate investors.

Or Hiltch is the CTO at skyline.ai and he joins the show to explain how to apply machine learning to real estate investing. He also describes the mostly serverless architecture of the company. This is one of the first companies that we've talked to that is so heavily built on managed services and functions as a service. So it was really fascinating to hear about their backend, because it's marketably different than a lot of the other companies that we've talked to, the fact that they are so heavily built on managed services and "serveless infrastructure".

[SPONSOR MESSAGE]

[00:02:08] JM: Leap.ai matches you with high quality career opportunities. You are more than just your skills, and a job description, and a resume. These things can't fully capture who you are. Leap.ai looks beyond these details to attempt to match you with just the right opportunities. You can see it for yourself at leap.ai/sedaily.

Searching for a job is frustrating, and Leap tries to reduce the job search from an endless amount of hours, days, weeks to as little as 30 seconds trying to get you matched to a job instantly by signing up based on your interests, your skills and your passions.

Leap works with top companies in the Bay Area, unicorns and baby unicorns. Just to name a few; Zoom, Uber, pony.ai, Cloudera, Malwarebytes and Evernote. With Leap, you are guaranteed a high response and interview rates. You can save time by getting direct referrals and guaranteed interviews through leap.ai. Get matched to jobs based on your interests, your skills and your passions instantly when you sign up.

Try it today by going to leap.ai/sedaily. You would also support Software Engineering Daily while looking for a job. Go to leap.ai/se daily. Thank you to leap.ai.

[INTERVIEW]

[00:03:46] JM: Or Hiltch, you are a cofounder and CTO at Skyline.ai. Welcome to Software Engineering Daily.

[00:03:53] OH: Thanks for having me on.

[00:03:54] JM: Yeah, it’s great to talk to you about this set of challenges that you’re tackling. Can you give a brief description of the problem you’re trying to solve at Skyline AI?

[00:04:04] OH: Certainly. Yeah. So the problem that we’re trying to solve is mostly around being able to make real estate investments more precisely in the same scientific manner that we used to from doing pretty much everything else these days, right? I could be selling stuff on Amazon or trying to get a loan or whatever, and by doing that these days almost all of these things would be highly powered by machine learning and other data-driven approaches.

The thing about real estate investments, that it's kind of like still remained traditional in the sense that it's very, very much intuition-based. Sometimes that intuition works. In fact, some of what we're trying to build is exactly that intuition, only like mathematically so to speak. But a lot of that intuition is sometimes based on some cognitive bias that turns out to be wrong. So our main mission is to basically try to advance this world of recent investments into the age of technology or to make it somehow more comparable with what we used to from the stock market, where robotic trading is currently responsible for about 70% of 80% of trading.

[00:05:18] JM: So real estate investing is a topic I'm somewhat familiar with because my mom has been in real estate for about 10 or 12 years in Austin over the course of which Austin, Texas has grown. It's expanded a lot, and it's one of those markets that there's asymmetric advantage that you can accumulate by experience essentially.

So over the past 10 or 12 years, she's had her ups and downs, but now she has so much knowledge and so many relationships that she can really get a good grasp on the information that's out there, but even so, I feel like this is the kind of set of tools that would be really useful to her, because if she had a dashboard in front of her that said, "Hey, here's a lot of heuristics or insights," and maybe you could drill down further into this, I feel like that would be really useful.

Now, obviously, the ideal would be that you would have so much information that it would be like you're sitting in front of a Bloomberg terminal for real estate, which is a Bloomberg terminal, for those who don't know is for traders, a trader sitting in New York or in Chicago has this operating system of information in front of them for how to trade stocks, or bonds, or currencies, and you don't need to be in Turkey to know about currency fluctuations of Turkey. All you need is the Bloomberg terminal in front of you.

So where do you think you're at in terms of progress between the sort of tool that gives you some insights that are useful, but you still need to be in the area. You still need to have some domain expertise, versus the Bloomberg terminal.

[00:06:58] OH: All right. So that's a great question. So there's no doubt that real estate experts; brokers and those guys, know their market inside and out. The good broker or the good experts,

like your mother, has seen dozens or even hundreds of different assets during the past 10 or 20 years, but eventually their field of view is still limited by that experience. The AI that we're building, it has like about 50 years of experience and has seen all assets in all times, and it also able to basically extrapolate the insides from all sorts of commonalties that aren't necessarily dependent on the geography.

So for example, you would imagine an asset in Seattle, which is maybe shares a few things with an asset in New York just because both of those assets are relatively close to an Amazon office and the tenants are roughly the same. The parts in the area are often the same. Yeah, definitely even the best of brokers could definitely use something like a tool that powers it to basically uncover all of those hidden correlations.

Regarding where we're actually at at the moment, so currently technology-wise our focus is on basically being able to predict things like market value of assets. So we're operating in the commercial real estate environment. So that would be mostly family assets and above, so not single family homes. In that market, it could be fairly difficult to assess the value of an asset. So our algorithms is basically able to generate really, really high-accuracy prediction for those market value and also about generating future rent predictions.

So these are the areas where currently I would say we're definitely already providing a lot of value. We like to call that the Ironman suit for real estate underwriting. So it doesn't act as to completely replace the analyst. It basically just arms them with super human abilities when it comes to getting to know the asset really, really deeply.

[00:08:55] JM: Yeah. So it kind of like the chess player who has a computer helping them to evaluate the set of moves, the Centaur Chess Model, the augmentative model. Why are you focused on commercial real estate instead of also including the single family occupancy homes as well right now?

[00:09:16] OH: Basically Skyline was born out of a personal lead that we had, myself and my cofounders, where we actually invested in ourselves in multifamily assets, and that asset class has actually a lot of advantages when it comes to things like defensibility, because in a

multifamily asset, it's basically a rental property, right? So whenever the economy is down, the occupancy levels are still fairly high.

So there are a lot of things that make this sort of asset really interesting for us, because we kind of like enjoy the fact that it's residential in a sense, because people actually live there. So a lot of features that are really interesting for residence, like schools, the crime rates, walk scores, things like that are relevant to that asset.

But on the other hand, it's an asset that is traded by huge companies, which have commonalities, like holding periods, like certain areas we like to invest, like certain things that happen to the assets before they try to dispose of it. So it's really an asset class that we figured was perfect for us to attack as the first go-to-market when it comes to developing accurate models.

[00:10:26] JM: Are the datasets more available in commercial real estate than residential single family occupancy?

[00:10:33] OH: So it's debatable, because there are datasets around mostly family. So we are actually paying quite a lot of money for data these days, but the truth is that none of the data sources – So there is no real one or even 10 datasets that you could use to get the complete picture when it comes to multifamily, just as there isn't one in single family. The truth is that we actually spend a lot of our work.

So in our R&D group, we basically have two teams. One team is the daily engineering team, and these are the guys who are responsible for getting the data and normalizing it and eventually building this one nice clean homogenous layer that the data science team could then later on connect to to generate all of the insights.

So it's really a game of being able to periodically download and stream hundreds of different of data sources with many dependencies between them in order to basically gain these understanding. It's not a picnic, but I would assume that the situation is not really different for single families as well.

[00:11:38] JM: What I've heard about building these kind of machine learning companies is often times you are gated by access to the data. Whether you're talking about building a machine learning model for predicting the severity of a stroke, or you're trying to build something around mapping. In all of these cases, it comes down to can you get the data?

Is there even somebody that you can buy the data from? Once you find that person who you can buy the data from, do you have the budget to afford it?

[00:12:10] OH: Yeah. So it's a really, really big barrier to entry to this market. That's definitely true. I think that a part of what we do to mitigate – So this problem exists obviously in real estate as well. In real estate, at least in commercial real estate, there are several so called commercial real estate analytics companies that you can buy data from, but that data is not always that accurate.

So one thing that we do in order to solve this problem, is we actually partner with funds, private equities and other real estate players in sort of relationship where we provide them with some insights about some of the deals and properties that they're looking at in exchange to getting some of their proprietary. One example is an agreement that we have with Greystone, which is one of the largest lenders for multifamily in the states where we actually enjoy this partnership in a way that definitely helps us get more quality data.

[00:13:03] JM: Before we get into the guts of how your company works, I'd like to step back and talk a little bit about the core hypothesis of your company, which is that you can use technical analysis for real-estate purchases. So there are some people who will, in financial trading, securities trading, stock trading, some people are technical analysts. Some people are fundamental analysts, like Warren Buffett will say, "You should look at the fundamentals of a specific business. You should look at each business atomically."

Rather than looking at how the market is fluctuating and the large collection of signals and trying to derive the wisdom of the crowds from these small technical signals and just accumulating technical wisdom. Then there's a giant amount of people who fall somewhere in between that gradient between fundamental and technical analysis.

So I'm sure if there're people who are involved in real estate who are listening to this, they might be thinking, "This is a losing game. You can't use these raw stats to figure out how to make a real estate decisions." It all depends on, "Do you know the market? Do you know the people? Do you know the way that the neighborhoods interact with each other? Do you have the macro perspective? Because the historical data is not going to help you." So I'd like to hopefully counter some of the upfront arguments against this kind of platform existing in the world.

[00:14:33] OH: Yeah. So I think that eventually what we're building, as I called it a few minutes ago, is kind of like the Ironman suit for real estate investments. As it is right now, it is not meant to completely replace humans in the process, because the process of a recent acquisition is actually quite a long one, which includes a lot of manual steps. For example, actually closing the deal, you cannot close a deal in real estate by pushing a button like you can do in the stock market. However, you can definitely use predictions about things like what the asset is actually worth, or what the rent is right now and what it is probably going to be in a year from now. That is on a sphere of data that you're partly exposed to right now when you're making investments.

Most analysts, when they would look into acquiring a multifamily deal or a multifamily asset, then would check a few things like, "The schools around the properties and some of the properties comps." Actually, the properties comps is a really good example of how you could actually use this sort of technology to exponentially improve your performance.

[00:15:41] JM: What did you say? Property discount?

[00:15:43] OH: So it's called property comps. So when you examine a specific investment or a property, one of the first things that you do is you construct a peer group, which is basically a set of properties which are comparable to your asset. So you want to know that you could buy this asset and then maybe make some value add investment. Then your thesis is that you could charge some premium following that value at investment.

One of the key things to do in this understanding is to look at comparable assets, so assets which are – For example, traditionally, roughly from the same vintage have the same number of units are in a close proximity to the asset and so forth. But usually when the analysts does that, they are fairly limited in their field of view, right? So they would pick maybe four or five, or I could

go crazy, like 10 assets, and then they would try to get their understanding of, “Okay, where this asset is standing versus these guys?”

Obviously, using machine learning, when you can construct the group, which is comprised out of thousands of similar assets and not just four or five, your analysis improve drastically, right? That’s one example. Trying to figure out what is the optimal renovation budget for such a property based on all sorts of parameters. It’s also something which is not trivial to do without having access to such an abundance of data.

Concerning the actual knowledge of the locality. So basically the types of data that we currently have, it could be like really, really local. So I mentioned the actual assigned schools for each property and its grades. The pattern of roads around the property, all sorts of things like that. But it’s also a macro-level data.

So for each transaction that we’re looking at, for example value prediction algorithm, the machine learning algorithms learns by looking at past transactions. For each such transactions, it has a really huge features vector, which is also comprised out of macro-level data. For example, what was the interest rate back then? What was the 10-year treasury constant maturity rate back then? What were the library rates back then? So it’s both looking at really high-resolution local properties, but also in real macro-level properties.

[00:17:55] JM: A couple of things there. The idea that you can use your product of product like it, use machine learning to, for example, predict how much a certain type of renovation is going to cost. That’s pretty useful if, for example, you’re considering buying a commercial property that has an HVAC system that’s malfunctioning in this particular way and you say, “I have no idea how to find out.” Is there going to be budget creep in this HVAC reinstallation? What if there’s also wiring issues in the building and I need to get those done too. Is that kind of overlap with the problems of my HVAC reinstallation?

I could imagine being able to predict the prices of those kinds of things would really help you in buying decisions and other kinds of decisions. But the core of what you said there is that comparable analysis has been a big windfall. It’s been an insight that has really been useful to you.

In some ways, I think this is nothing new. It's collaborative filtering. This is the same thing that gives you Netflix recommendations based off of people like you, or Amazon recommendations based on people like you or people who also liked this item. This is a well-studied area of machine learning, or computer science. We know it works if you have the right signals, if you have the right data. So what are the signals that you have found to be the highest signal in trying to construct a comparable analysis model?

[00:19:27] OH: So that's a great question. Obviously, some of the features are things that I cannot expose, because these are some of what makes our secret sauce, but I could definitely mention a couple of those.

[00:19:37] JM: Other ones are just latent, right?

[00:19:39] OH: Yeah, indeed. Yeah, a couple of examples is maybe the more trivial ones would be all sorts of trends in the property rent. So once we have access to the history, to like a time series of rent data, of rental performance data that obviously represents like similarity in some sense. Other ones would be things like – More advanced ones would be actually kind of like a digital signature of the properties environment. I'll explain what I mean by that.

So one of the nicest things that we're doing at Skyline is we're actually using machine learning to generate new features for more machine learning models. So we have our supervised learning algorithms, which is responsible for things like the value prediction. But in fact some of the features that's feed into this model are generated by unsupervised learning models.

So just as an example, for each and every asset that we're looking at, we're using Google Maps static API to basically take a snapshot of the properties in the immediate environment emphasizing things like roads or other points of interest and then using auto encoders powered by continents, we're able to learn the representation of that environment. So the ratio and distances between the asset and all sorts of points of interest, for example, things like roads, schools, parks, bodies of water and things like that.

So the use of auto encoders basically allows us to translate that image into a really small vector that eventually is just another feature in our value prediction. So in a sense it's historical rents, whatever compounds, the properties environment, a macro-level financial figures, like I mentioned the U.S. bonds, so the 10-year treasury constant maturity rates, which are actually an alternative investment to real estate and things like that.

[SPONSOR MESSAGE]

[00:21:34] JM: Citus Data can scale your PostgreS database horizontally. For many of you, your PostgreS database is the heart of your application. You chose PostgreS because you trust it. After all, PostgreS is battle tested, trustworthy database software, but are you spending more and more time dealing with scalability issues? Citus distributes your data and your queries across multiple nodes. Are your queries getting slow? Citus can parallelize your SQL queries across multiple nodes dramatically speeding them up and giving you much lower latency.

Are you worried about hitting the limits of single node PostgreS and not being able to grow your app or having to spend your time on database infrastructure instead of creating new features for you application? Available as open source as a database as a service and as enterprise software, Citus makes it simple to shard PostgreS. Go to citusdata.com/sedaily to learn more about how Citus transforms PostgreS into a distributed database. That's citusdata.com/sedaily, citusdata.com/sedaily.

Get back the time that you're spending on database operations. Companies like Algolia, Prosperworks and Cisco are all using Citus so they no longer have to worry about scaling their database. Try it yourself at citusdata.com/sedaily. That's citusdata.com/sedaily. Thank you to Citus Data for being a sponsor of Software Engineering Daily.

[INTERVIEW CONTINUED]

[00:23:19] JM: You mentioned an application there of auto encoders, and I don't think this is a topic that we can completely describe on a podcast. But could you give a few other applications for auto encoders? When are auto encoders generally useful for a machine learning application?

[00:23:36] OH: So auto encoders are classically used in a noise reduction and anomaly detection. So the basic idea is that using – An auto encoder is a mechanism. Its underlying architecture could be various types of neural networks. When it comes to using auto encoders over visual data, for example, something like a graph, or a map, usually you would train an auto encoder to be able to represent an image using fewer data points.

So the auto encoder would take an image and its job would be to restore the exact same image using less data. The result is that the places where the auto encoder isn't able to reconstruct the image as well as it can. So those areas would usually be classified as anomalies.

So you would see people using auto encoders in anomaly detections. For example, for monitoring server activity and things like that. Also, like I said, in noise reduction. It's fairly common to use them in a noise reduction scenarios when you get a noisy signal and you want to actually figure out what is the signal and what is the noise. Those would be two common uses for that.

[00:24:45] JM: Interesting. So it's kind of like compression, compression algorithms. The more commonality you have across your data, the more efficient your compression is going to be. With auto encoders, you can use that to say, "Well, if we're not able to compress it effectively or encode it effectively, that's in some sense an anomaly by definition."

But more generally speaking, in the non-anomaly detection example, you can use it to just reduce the dimensionality of an image, for example, or a set of images. So your machine learning model can process the data in a collection of images more effectively.

[00:25:23] OH: Yes, exactly. So if some of the listeners are familiar with PCA, principle component analysis, which is a method of reducing the dimensionality of a vector. So you could definitely think of an auto encoder as a non-linear PCA. Basically being able to apply the same principle, but not using like orthogonal transformation, which only works when the properties or the features are linearly co-related.

[00:25:48] JM: Let's get a feel for the timeline of this business. I was looking at your work history and you've done a couple of other companies in the past, and so this was nothing new to you. What was the process of figuring out what this product was going to be? Did you know it upfront? You said, "Okay. Let's grab a dataset and just start hacking on it and see if there's a there there." What was the founding sequence of steps?

[00:26:16] OH: Yeah. So it's a great question. I think it's funny that we actually had several other companies in the past prior to Skyline, but Skyline is probably the first company that we established that actually sprung out of a personal need. So prior to Skyline, we had a company around cyber security, which was acquired by AVG, the antivirus software company. Then we had a company around video optimization, a company called StreamRail, which was acquired by IronSource in '06.

Then what happened is that we actually started investing in real estate personally, and we kind of like figured out that it's not working too well, because we would have a hard time finding the best deals. Once we reached an operator, they would provide us with the OM, with the offering memorandum, depicting all of the deals characteristics, like the expected yield and things like that. We had no idea of figuring out if that data was correct or not and if whether or not those predictions about the yield are worth anything.

So we started out by just looking around. We said, "Okay. Maybe we should look at what the big boys are doing. The ones who managed the investment for some of the largest institutions in Israel and in the states, and we got a few introductions to some of the people who had the global real estate operations for some of the largest private equities in New York City, and we met with them and we saw that they are not doing a lot of data capabilities or not as good as they could be.

Just as an example. So my little brother, he's a dentist, but he have this hobby of building Irish tin whistles. He builds them from scratch, and it's a just a hobby, because in day-to-day work, he's a dentist.

[00:27:57] JM: Irish tin whistles?

[00:27:58] OH: Yes. Yes. It's a type of flute. It's a pretty nice [inaudible 00:28:01]. Anyways, and he likes to sell them on eBay and Amazon just for the fun. The type of tooling that he gets as a merchant on eBay and Amazon to optimize his sales is a million times more sophisticated and more data-driven than the type of data tools that some of the largest private equities in real estate today have.

So that was when we realized, "Okay. There is something here that we can probably try to improve," and that was how Skyline was born. The next steps were trying to figure out what can we get from around the web. Had gone through a really long process of understanding which of the data providers are more accurate than the others. We've actually built and trained – Today we are actually building machine learning models just to tell us which of the data providers is likely to be more accurate in the property level. We've gone through quite a big journey, and just in setting up the infrastructure for doing this thing.

[00:28:58] JM: This is one of those things where you are probably faced with the question "Why doesn't this exist? It seems like this should have existed. Therefore, maybe we shouldn't even try 0it. There's probably very rational reasons that it doesn't exist. Let's go move on to something else." Instead you said, "There is a reason why this doesn't exist in the market," whether it was the lack of overlap between somebody that saw this opportunity and somebody that had the machine learning insights to figure it out, or maybe the machine learning tools have gotten easy enough to work with that now you can do this, or maybe the data is more accessible. Do you have any understanding of why this problem had not been successfully tackled before?

[00:29:43] OH: Yeah. So I think it's a combination. I think I have like my own theory regarding the DNA of the types of players that work in this field. There is a growing understanding in all of the commercial real estate investment world that AI and machine learning are going to change the face of the field in the next coming years. Some of these organizations have been able to recruit the data science team, try to figure out how they can improve things.

But the truth is that the type of challenge that we're facing here is just as much as an engineering challenge as it is an AI challenge, right? So just being able to construct this data

layer that is fed constantly from dozens or hundreds of different data sources with dependencies between them and concurrency requirements and things like that.

Just this part, it's quite a big engineering effort, and you do need an engineering culture. I would even say like startup DNA to be able to manage that. Certainly to converge it with AI disciplines, right? Traditionally, people coming from the AI and data science world are more research-oriented and they're able to do magical things given the correct data.

But in this business, it's really about merging both engineering and data science, and this is something that I think the big financial institutions have really been struggling with. That's, I guess, my R&D manager point of view. But there are a lot of other reasons, right? It's like one of the largest traditional industries in the world today, right? A lot of things are still based on lack of transparency, so the bidding process. You can see what a property is sold for, but you cannot necessarily tell what were the bids prior to that. Yeah, I think these are some of the reasons for that not to have happened yet.

[00:31:43] JM: I love your data engineering hypothesis, where you take a big financial institution. They have the data advantage. They have lots of historical data. But it's spread throughout the company and the people at the top who are asking for that data, because they want to make high-impact investment decisions based on it. They have to ask a business analyst. The business analyst has to ask a data scientist. The data scientist has to ask a data engineer, and the data engineer has to go find the 50 data sources throughout the company and figure out how to get them together.

By the time they get all that and they hand it off to the data scientist, the data scientist hands it back to the data analyst, the data analyst hands it back to the VP of real estate investments. It's been a 8 months and the opportunity has passed.

[00:32:29] OH: Exactly. Yeah. So I think one of the things I like the most about Skyline is that it's really a weird bunch of people rocking on the office here. We've somehow been able to merge those three disciplines into one company that works like a single organism. So we have the engineering guys on one hand, the AI guys on the other hand and the real estate people and

they're all working together continuously to improve this. This is something that I think it's really hard to reconstruct certainly in a large financial organization.

[00:32:55] JM: Oh, yeah. Not to mention the fact that you can start with modern productivity tools. You can start with Slack and whatever project management tools you use and the org structure is small enough so that those tools haven't scaled and fractured yet. So you can communicate more effectively. So it makes a lot of sense.

I do want to get a sense of the data engineering pipeline you have. So as far as I understand so far, the input is lots of sources of data from different real estate data systems. I believe you have all of the commercial real – Or the vast majority of the commercial real estate in the United States across the union of all of those datasets. Is that right?

[00:33:37] OH: Yeah. Right now we're actually focused on multifamily assets. So we do cover pretty much all of them. The way that our data pipeline – So we have basically –

[00:33:45] JM: By the way, that's like apartments. You have like all the apartments.

[00:33:48] OH: Right.

[00:33:48] JM: Okay.

[00:33:49] OH: So it's apartment complexes, exactly. We basically have two types of pipelines running on a nightly basis. One is the data pipeline. So that's basically starting out with an in-house infrastructure that we developed in order to orchestrate the running of the ETLs. So ETLs, an application that extracts, transforms and loads data. We have a bunch of these, like 130. Each application, it's a standalone application written in Go in our case, which is responsible for downloading the data from somewhere. It could be from one of our data partners. It could be from using Google Maps API. It could by downloading PDF documents and then extracting a structured data from the handwritten signatures and things like that.

So it's really, really varied. So each ETL is a standalone Go application, which is then Dockerized. So we run inside a Docker container, and we have an orchestrator that basically

knows when and how to run each of these Docker containers. So the thing is once you have a lot of these, just as an example, I could run the ETL that fetches the weather data in parallel to running the ETL that fetches the weather data, or the crime data, right? Because they're both not related.

But I would not want to call the process that generates the new prediction before those two are complete. So if you look at the variety of those ETL applications, eventually you would end up with something like a graph, like a directed acyclic graph, where each node represents an ETL application and they have all sorts of dependencies between them, which enable you to run some of them concurrently or to know that there's just no use in running something because its dependency has failed.

So we have built an in-house infrastructure to handle that, which is basically running – It's responsible for executing the graph each night. One cool thing that we did in this process is basically leverage serverless capabilities in order to not be tied into something like a huge machine that would handle all of these load.

In our process, our orchestrator basically kicks off each ETL inside a Docker container, and the runtime for that Docker container is a platform called hypershell. Hypershell, it's a serverless container platform, kind of like AWS Fargate, which is kind of new. Basically in two words, that means that you could basically tell that platform, "Look, here is my Docker image. Just run it. I don't want to have to worry about the location or the type of machine or anything like that," and that has scaled pretty well for us, because since each and every ETL is – It has its own different needs, it definitely makes sense to run them completely separately on a serverless infrastructure where you don't have to care about things like the memory consumption or the network and bandwidth and stuff like that.

[00:36:39] JM: Is that Hyper.sh? Is it that company?

[00:36:42] OH: Yes, it is. Yes, indeed.

[00:36:44] JM: That's also Zeit, right? Or it's a company under Zeit, or is it different?

[00:36:48] OH: Yeah. Hypershell, it's a completely different company, but is fairly similar to Zeit.

[00:36:54] JM: Okay. All right Cool. Are they using AWS Lambda or Google Cloud Functions or something underneath?

[00:37:00] OH: No. They've actually built their own infrastructure. I think they do support being deployed on AWS, but it's mostly their own stuff.

[00:37:08] JM: I've talked to some people that use Airflow for orchestrating these kinds of jobs and the dependency graphs, because like you said, you have these series of ETL pipelines and some of them may depend on each other. So that happens before relationship can be important. You can just spin up all these jobs willy-nilly. I believe Airflow is a scheduler out of Airbnb that some people use for this application. Have you looked at that?

[00:37:33] OH: Yeah, we have. Yeah. So we've looked into Airflow and Luigi, which is not a Python-based DAD execution engine. Back when we started – So we started about almost a year ago. Airflow was really, really not mature at that stage. I think right now it's already under the Apache incubator. So it's probably a lot more production ready than it was almost a year ago when we started looking into that. But back then, it didn't quite work well for us.

I think another issue is that we really wanted our platform, our ETL platform to be dynamic in the sense that you would not have to write code in add nodes to the graph, and Airflow is imperative in that sense if adding new nodes to the graph means that you – At least back then it meant that you had to write some code to do that. We really wanted to do that through an API so we could play around with it or get things running through Slack or through our own user interface and things like that.

Yeah, but I did hear that it actually – It went a long way since back then. So maybe it's a good time for us to check it out again.

[SPONSOR MESSAGE]

[00:38:45] JM: Logi Analytics is an embedded business intelligence tool. It allows you to make dashboards and reports embedded in your application. Create, deploy and constantly improve your analytic applications that engage users and drive revenue.

You focus on building at the best applications for your users while Logi gets you there faster and keeps you competitive. Logi Analytics is used by over 1,800 teams, including Verizon, Cisco, GoDaddy and J.P. Morgan Chase. Check it out by going to logianalytics.com/datascience. That's logianalytics.com/datascience.

Logi can be used to maintain your brand while keeping a consistent familiar and branded user interface so that your users don't feel like they're out of place. It's an embedded analytics tool. You can extend your application with advanced APIs. You can create custom experiences for all your users, and you can deliver a platform that's tailored to meet specific customer needs and you could do all that with Logi Analytics. Logianalytics.com/datascience to find out more, and thank you to Logi Analytics.

[INTERVIEW CONTINUED]

[00:40:12] JM: These ETL jobs run. You get the data. You normalize it all and then you throw it into what? A database?

[00:40:20] OH: Into a data warehouse.

[00:40:22] JM: Data warehouse.

[00:40:22] OH: Yeah. So we're using Google BigQuery as our data warehouse, which has a lot of benefits for our use case. I think just to mention a couple of those. So Google BigQuery, it's kind of like a serverless data warehouse where you don't have to worry about scaling storage or compute or anything like that. Google takes care of that for you.

They used to compare it to AWS Red Shift, but now I guess it's more similar to AWS Athena. So Google BigQuery, it has a lot of advantages for us. First, it has a really, really strong API. So you could pretty much create a table from anything using a really simple command line interface. So

you could tell it, “Hey, go out and load these bunch of CSVs and use wild cards and everything.” Some of them could be Gzip, some of them could be plain out CSVs and they could be a bunch of different locations and it would handle that pretty well.

It also has pretty neat automatic schema detection mechanisms so you could treat it in a sense like a schema-less database. You don’t really have to worry about structure changing, and it also supports snapshot queries. So if you want to query tables to get the data that they held a few hours ago and things like that, it’s built in to the syntax. So it’s really easy to do.

Basically, at the end of the day, it allows us the peace of mind of having petabyte scale data all in the same data warehouse without having to worry about the dev ops part of babysitting it, indexing it and things like that.

[00:41:53] JM: It goes all into BigQuery, and then do you build your applications on BigQuery basically, or do ever put this data into Mongo, or Redis, or something?

[00:42:03] OH: Right. Yeah. So after it’s in BigQuery, it’s basically the data warehouse, where the main user of this database is our data science team. So what they would do is they would basically source and prepare queries from the BigQuery, then they would code their models. This could be all sorts of models. Some of them like deep learning models. Some of them are things like a gradient boosted trees. It’s quite a large variety of models.

Then once they have coded the model and experimented with it, usually using Jupyter Notebook, we would use Google Cloud ML engine to basically train, evaluate and tune the model. So Google Cloud ML engine, it’s a cloud platform for doing everything after the model is already coded. So taking care of, training it efficiently using GPUs or even TPUs, and if you’re using TensorFlow. Being able to deploy the model and also getting predictions from it so they could actually – The Google ML engine has the ability to serve an API for the predictions once the model is trained. Then it also provides you with the capability of monitoring the ongoing predictions so you can basically compare things like training times, accuracy levels, any weird stuff going on with loss functions after data has changed. Then finally, you could also manage your models and versions. So you could decide which of the models you want to use in production right now. You can A-B test different version of the model and things like that. That

would be the machine learning pipeline. So after the data is already in BigQuery. So that's like one user, one type of user for the platform.

The second type of user is the engineering team who's responsible for maintaining our web application. So we also have like a Google Maps style web application where you could basically search for an asset by owner, or by history, by address, by name or you name it. Then you would basically be able to get tons of insights about the property, like predictions about the future and things like that in a few seconds. That application works by querying an AWS Lambda API, and that AWS Lambda API fetches the data either from Redis if it is indeed cached there, or if there is a cached miss, then it goes out all the way to Google BigQuery.

[00:44:17] JM: Wow! Okay. So you've described most of the stack there. We can talk a little bit more about managed services hopefully. So you've mentioned AWS Lambda and Google BigQuery. So it sounds like you're pretty cloud – And hyper.hs. It sounds like you're pretty cloud-agnostic.

[00:44:33] OH: Yeah. One of our first priorities is being completely serverless in the sense that we don't have to spend times on worrying about up time and things like that. Yeah, it could be our data warehouse, which is BigQuery. It could be our API functions, which are served over Lambda, or the ETLs, which we currently run Hyper.

Yeah, we basically just try to pick the best solution for each part of the problem regardless of whether it is on Google, or Amazon, or whatever. This approach, it's proven itself to be quite efficient in our case.

[00:45:05] JM: I love it. I think that's really sophisticated. I've been talking to a lot of companies that are migrating to Kubernetes, but it sounds like you have built in a way where you don't even really need a Kubernetes cluster.

[00:45:17] OH: Yeah. I think that if you – Nowadays, I really think that as long as you could use serverless – So if, for example, your use case is supported on something like AWS Lambda, then there really isn't any reason not to do it, right? I guess that on some, there are definitely a

lot of use cases. For example, like the ETL example, where it's not a good fit for something like Lambda, because it's long-running and it has to manage state and things like that.

So for those applications, Kubernetes is definitely a good solution, especially if cost is involved. So our users, at the end of the day, the users of our platform are only – As I said, we're using the platform to make better real estate deals, which means that either it is us using the platform to actually track source and acquire the assets, or it is someone from our real estate partners.

So in that sense, the scale doesn't come from the user facing APIs. In this regard, cost is not a huge concern for us. It's not that we have to manage like a million request per second and then at which point certainly managing your own servers on something like Kubernetes makes more sense than using Lambda. Our scale cost is more in the data part, like training the models, running the queries.

So luckily, it's been possible for us to get the best of our world when it comes to serverless, and we don't have to carry the burden of managing something like a Kubernetes cluster.

[00:46:45] JM: If the bulk of your processing costs are coming from training machine learning models, does the Google hosted machine learning service, are they always going to be able to schedule and train your models in a way that's more — do they do that in a way that's more cost efficient than if you were trying to do it on your own Kubernetes cluster for example?

[00:47:05] OH: So you could implement what Google has built with ML engine on your own, but then it comes back to the tradeoff you always have between implementing your own thing. So I guess that on a certain scale, currently at Skyline AI, we are about 20 people and about 12 of that is engineering. So I guess that the larger the company becomes and the more cost-sensitive we will be in the sense.

If you look at Facebook, Facebook has their own CDN, right? So it never ends. Whenever you have a big knee, then you're big enough, you would eventually probably go out and build your own.

[00:47:42] JM: Well, who knows if Facebook would need a CDN if they had started in the age of serverless stuff? Hard to know.

[00:47:49] OH: Yeah, I think that it is. But I guess, again, it probably comes back to the tradeoff. We don't know how much effort and money you need to put in to build your own, versus just leverage an off-the-shelf solution.

[00:48:01] JM: Yeah. Have you been able to avoid downtime and heavy operational issues and heavy debugging issues through this serverless approach?

[00:48:10] OH: Yes. Yes. So we haven't actually had any downtime since we started the company since we went to production. That's largely due to the fact that it's quite difficult to bring down something like AWS Lambda. We are using a bunch of logging stuff and we have some automation around making sure that production shows what it should show. We definitely get sometimes bugs, where like the visual aspects of things and things like that. But the infrastructure side, we'd like to say we've been really fortunate in having quite a perfect uptime thus far.

[00:48:43] JM: So there's a lot of stuff we didn't cover, but I guess I do want to take a step back, because we're running out of time. This company architecture, the software architecture you have, probably is dramatically different than the previous companies you've had. How does the difference in going heavily on serverless, managed services, AWS Lambda, etc. How does it change the company building side of things and how does it translate into advantages when you compare the company building exercise today to your pervious companies?

[00:49:19] OH: Yeah. So Skyline AI is very, very different indeed. I think it's mostly different in the sense that, well, in a couple of ways. First of all, it's that we really have, I would say, a mutually respectful size of data science team versus an engineering team. So we actually have more data scientists than engineers at the company, which is also a first for us, and that actually creates a lot of interesting challenges in basically being able to maintain kind of like the Agile mentality, when you have to deal with a lot of people doing research. That's one I think really interesting challenge that we had the chance to cope with here, which is fairly different from the previous companies.

We also have the fact that there's really large business side that is really, really different from our world is tech guys, and merging that discipline into the company is also something that is quite refreshing and interesting for us. There are actually a lot of similarities, because eventually we did it in cyber and then we did it in video, but both VisioNize and StreamRail were around – At the end of the day, using data to optimize stuff.

While we haven't actually had the luxury of working with completely serverless infrastructure in the past, we have gotten a taste of them, and it was quite clear to us that whenever these things becomes a reality, then we will be the first to adopt this. So I think it definitely helped us scale the technology business, because we don't have to – These days, we could a data scientist fresh out of university, doesn't know nothing about dev ops, and they could pretty easily integrate their code into our workflow, which is kind of a complex architecture when you think about it. So it's the same – I will guess that if we would want to do that before the days of Cloud ML and BigQuery and Lambda, then we'll probably have to have ops guys and dev ops people. Yeah. It definitely helped us grow really, really quickly when it comes to launching the product to production.

[00:51:25] JM: Well, I think the hangover from proprietary, like the 90s proprietary databases, proprietary operating systems, people getting locked into those things. The hangover has been really severe. So people have been really resistant to jumping on managed services and going all-in on them, at least some people. That's just my sample from talking to people. But I think as productivity, I'm like, "If you're a random hacker in a bedroom, build on Heroku, build on Firebase. If you're starting a new company, build on as much managed services as you can, unless you're on a really low-margin kind of business and costs are really that important."

[00:52:07] OH: Yeah. Yeah. I totally agree. I think that whatever can be done like serverless. There's no really justification in doing it in any other way if it is available in a serverless way and if it's a good fit.

[00:52:20] JM: So the company is doing fantastic, and you've already got all these commercial properties in the United States. What are the challenges that you're encountering? What are the biggest challenges and what's in the future for Skyline AI?

[00:52:34] OH: So thus far we've been focused on executing deals on a deal-by-deal basis, which meant that whenever we source and do it in a deal that we like, we went out to our investor-base, which is mostly around like family offices, or high-net worth individuals globally.

[00:52:52] JM: Right. I'm sorry. I didn't even allow you to explain really the business model. So you've surfaced deals and then you bring them to people and then people have a chance to invest in the real estate opportunities that you discover?

[00:53:06] OH: Right. Basically, our model is kind of similar to what is sometimes referred to as private equity. On one end of the equation, you have capital, and that capital can come from institutions, like pension funds, or insurance companies, or it could come from family offices, which are companies that manage capital for some rich families, or it could come from what's referred to as high-net worth individuals. People with enough capital to be able to afford investments in an expensive asset class, like multifamily where the minimum check sizes could be several millions of dollars. So that's like one end of the equation.

Then on the other end of the equation, they are the operators. So these are the guys or the companies that are really doing the hands-on real estate part. So they would be the ones who are actually physically acquiring the property, taking care of, managing it, applying their innovation part and so forth.

In the middle, there are the private equities. So companies like Blackstone, for example. It's one of the most well-known private equities for real estate. The area where we operate is there, it's in the private equity slot. The difference is that we're doing everything according – Enhanced by technology. So it could be in the sourcing of the deals or figuring out a certain area that we want to focus on. It could be the underwriting of the deal. So once we already have a deal in the table, we would use our AI in order to make sense of it.

Eventually, the business model is roughly the same. We go out and we raise money whenever we think we can make a good investment. Thus far, we've done that on a deal-by-deal basis. Meaning that once we had a deal, that we thought, "Okay, we're going to go after this one. We would go out and raise the money."

But right now we are working on changing this a little bit into more like something like a fund structure, where we'd basically partner with a large operator in the states, and together we would be raising a large fund to allow us to deploy more efficiently and more quickly. So that's one of our main goals for the next couple of months. We're already doing pretty good there. We're already also working – From a technology perspective, we're working on basically expanding to additional asset classes. So we started off with multifamily, but we also have office logistics industrial and a lot of more asset classes coming up. Yeah. So a lot to do.

[00:55:27] JM: Very cool. Or, it's been really great talking to you. I appreciate you coming on the show and making time to tell me about Skyline AI.

[00:55:33] OH: Sure. Thanks for having me.

[END INTERVIEW]

[00:55:37] JM: Cloud computing can get expensive. If you're spending too much money on your cloud infrastructure, check out DoIt International. DoIt International helps startups optimize the cost of their workloads across Google Cloud and AWS so that the startups can spend more time building their new software and less time reducing their cost.

DoIt international helps clients optimize their costs, and if your cloud bill is over $10,000 per month, you can get a free cost optimization assessment by going to D-O-I-T-I-N-T-L.com/sedaily. That's a D-O-I-T-I-N-T-L.com/sedaily. This assessment will show you how you can save money on your cloud, and DoIt International is offering it to our listeners for free. They normally charge $5,000 for this assessment, but DoIt International is offering it free to listeners of the show with more than $10,000 in monthly spend. If you don't know whether or not you're spending $10,000, if your company is that big, there's a good chance you're spending $10,000. So maybe go ask somebody else in the finance department.

DoIt International is a company that's made up of experts in cloud engineering and optimization. They can help you run your infrastructure more efficiently by helping you use commitments, spot instances, rightsizing and unique purchasing techniques. This to me sounds extremely domain

specific. So it makes sense to me from that perspective to hire a team of people who can help you figure out how to implement these techniques.

DoIt International can help you write more efficient code. They can help you build more efficient infrastructure. They also have their own custom software that they've written, which is a complete cost optimization platform for Google cloud, and that's available at reoptimize.io as a free service if you want check out what DoIT International is capable of building.

DoIt International art experts in cloud cost optimization, and if you're spending more than $10,000, you can get a free assessment by going to D-O-I-T-I-N-T-L.com/sedaily and see how much money you can save on your cloud deployment.

[END]