EPISODE 1264

[INTRODUCTION]

[00:00:00] JM: Apache Superset is an open source, fast, lightweight and modern data exploration and visualization platform. It can connect to any SQL-based data source through SQLAlchemy at petabyte scale. Its architecture is highly scalable, and it ships with a wide variety of visualizations. The company, Preset, provides a powerful, easy to use data exploration and visualization platform powered by Apache Superset. Preset enables team members with some to no programming experience to build interactive visualizations and dashboards with a no-code visualization builder and SQL editor. It works directly on top of popular cloud data warehouses and leading SQL engines. Preset delivers all the data visualization power of Apache Superset through their complete, easy to consume, enterprise-ready platform.

In this episode, we talk with Srini Kadamati, senior data scientist and developer advocate at Preset. He previously worked at Dataquest, and at Radius Data Intelligence before that. He's also a committer to Superset. We talked about data visualization, the power of big data, and Preset.

A few announcements before we get started. One, if you like Clubhouse, subscribe to the Club for Software Daily on Clubhouse. It's just Software Daily. And we'll be doing some interesting Clubhouse sessions within the next few weeks. And two, if you're looking for a job, we are hiring a variety of roles. We're looking for a social media manager. We're looking for a graphic designer. And we're looking for writers. If you are interested in contributing content to Software Engineering Daily, or even if you're a podcaster, and you're curious about how to get involved, we are looking for people with interesting backgrounds who can contribute to Software Engineering Daily. Again, mostly we're looking for social media help and design help. But if you're a writer or a podcaster, we'd also love to hear from you. You can send me an email with your resume, jeff@softwareengineeringdaily.com. That's jeff@softwareengineeringdaily.com.

[INTERVIEW]

[00:02:07] JM: Srini, welcome back to the show.

[00:02:09] SK: Hey, Jeff. Great to be here.

[00:02:11] JM: The last time we spoke on the show, I think we were talking about data science. You have a background in data science. You've done a lot of work in data science. And now you're working more on the frontend, the business intelligence layer. Tell me about the modern state of the business intelligence layer and how that relates to the data science or business analyst layer?

[00:02:39] SK: Yeah, that's a great question. I think it's definitely going through some interesting changes. So I generally think of BI as kind of focused on end users. So people who are trying to get value out of data, one of the ways to get value out of data is through charts, dashboards, especially if you're busy. And so that's kind of how I would differentiate the BI layer from more traditional data science tools where people may be doing kind of complex things in Python or Spark. They're doing really complicated data processing, machine learning, building machine learning models, and that the consumer of those types of products and interfaces are other code, backend systems, that type of thing.

So I think BI ultimately is kind of the frontend and user realization of data science. I think that's kind of the – Yeah, that's really an interesting part. And the reason why I transitioned into it is – So the last time I came on the show to talk about Dataquest and I was talking about teaching people data science tools. And that was a good learning experience for me. Ultimately, I decided that I also wanted to make the data tools themselves better and more accessible. And I think BI is kind of one of the most interesting areas right now in terms of accessibility for data, allowing really everyone in an organization to play with data.

[00:04:00] JM: Give me some of the example applications of business intelligence that are common to an organization today.

[00:04:10] SK: Yeah, so right now, it seems like it's growing like crazy every time kind of I work with an organization. So really, I think all jobs now are becoming data jobs. So you're seeing sales teams where they're trying to understand and optimize their own processes. And throughput, trying to understand how to close deals better. They're using data. They're using BI.

So definitely, I think sales is kind of a very common area now that we're seeing. You won't really find a modern sales team anymore, that isn't kind of aggressively looking at dashboards. And using that – Sales managers themselves, using it to try to understand where things can be better optimized. It's definitely a dominated marketing. So where do you invest your marketing dollars? How do you kind of try to make sense of things that are hard to measure? It's really hard to measure things more in brand advertising. But there are other things that are easier to measure. And, really, BI is used heavily by marketing teams there, whether it's paid advertising, whether it's more organic things.

At Preset, we have our own internal marketing dashboards. We're a pretty small marketing outfit, I would say. But even us, like we're able to use that data to understand where to invest more of our time and energy into, whether it's doing more events and webinars or writing more blog posts. So I think that's been cool to see too, that BI and data in general is spreading to more teams and organization.

Definitely, product. So, like product, I think the most classic example here is the product funnel. Right from when user discovers the product to getting to some form of user activation. And I know when I was in my last company, we use data heavily to try to understand and iterate on our funnel, as a way to kind of improve conversion and better understand our value proposition. So those are three examples. I do know that even nowadays, executives from really all teams, all business units, they're looking at dashboards. People are using dashboards in meetings. They're using it for OKR reviews. So it's really kind of just become this great indicator of progress towards the eventual outcome that each team is trying to hit.

[00:06:20] JM: Preset is a fairly new company. And the – One of the driving factors of the genesis of the company is open source, business intelligence. That said, open source in and of itself is not enough to make a company successful. I guess part of my thesis on why Preset is successful is really the revolution of React on the frontend and just better frontend practices in general. But I'd love to know why you think there's room for a newer BI company when we've had the tableau and whatever, micro strategy, or whatever the other legacy BI tools are out there.

[00:07:06] SK: Yeah, absolutely. It's a great, great question. I think ultimately, it comes down to really, if you go back to the original story for why Max Beauchemin started Superset. He was using Tableau. He was at Airbnb. He was at a really mature large scale data platform team. And they wanted to play around with Druid, and Presto, and Trino. They wanted to kind of have their BI tool work with their modern data stack. And at that time, Tableau just didn't support those new data tools. And so they were kind of just stuck. And they had to kind of – Data platform speed was in many ways bottlenecked by their BI speed. Because eventually, this data was going to feed into charts and dashboards that the proverbial teams that I mentioned earlier were going to consume. So that's kind of why Max started working on Superset originally at a hackathon, was he really wanted to focus on let's use open source components for the entire application.

Again, he'd worked at Airbnb, Facebook, Ubisoft. He saw kind of web scale technologies and how they could really cheaply and efficiently deliver value to millions, billions of users. And he just saw that compared to the BI space where people were still trying to sell you like cost per core and clusters. And like Tableau is based on Windows. You have to run Windows servers. And it's just a very different model than where the rest of kind of the web stack and just the modern data stack in general has moved. And so I think you're absolutely right, the React and TypeScript, which are both kind of key choices that were made early on, and those paid off really well and continue to pay off.

But even on the backend, we're using Python in the Superset community. We even embraced type hinting, which is like a new Python feature to help with scalable Python code, especially complex code bases. And we bet heavily in the community on SQLAlchemy. In fact, just this week, I gave a talk on how we use SQLAlchemy to support hundreds of databases. And so by focusing on these open standards and frameworks that aren't really going away. So even if there's a new database company started every two weeks it feels like. And shortly after their funding announcement or whatever, you'll see that they'll quickly announce a SQLAlchemy driver, which means that it's a pretty short path to supporting all the features in Superset.

So I think open source ultimately gives us and gives end users that want to use cutting edge features just a lot more speed. And we're even seeing some really interesting use cases where people are taking Superset, which kind of started out as a more traditional BI tool in terms of the user interface. And they're using it in all kinds of interesting and new ways that we didn't even –

The original contributors didn't think about. Like embedded has become a big use case for Superset that's being pushed by the community where you're taking charts that people are creating end users are creating and then embedding those charts and dashboards in other applications.

And so there're just all these kind of great remixability and extensibility benefits that you get in open source and the benefits of network effects. Again, by betting on these open standards, you know that as your data stack evolves and changes, your BI layer will also kind of just keep up, that you're not waiting for a vendor to hopefully support and update their functionality one day.

[00:10:35] JM: You touched on some of the pieces of the stack there. Can you just dive a little bit deeper into the macro view of the preset stack?

[00:10:46] SK: Yeah. So just to kind of clarify here. So Preset is basically offering a cloud-hosted version of Superset. But, really, we're kind of offering as a service, the same thing that's in core Superset that's really available to everyone. And what's interesting about that is that's kind of another advantage of open source, even on the business side, is we're able to kind of – When we speak to people, they've heard of Superset. They may even have teams in their organization that are using Superset. So it's kind of really interesting for us from like a business and a cloud-hosted standpoint.

In terms of the architecture, we're running Python in the backend and it's based heavily around a framework called Flask-AppBuilder, which is kind of a way to quickly bootstrap a lot of the backend routes that you'll see in common web applications. On the frontend, we have React and TypeScript. So even there, we're trying to innovate and pick the latest and greatest. So kind of the community just announced a few months ago that they were really betting on Apache ECharts. So another kind of vis framework, because Superset was originally based on NVD3, which is a layer on top of the D3 that everyone has come to know that gave Superset a little bit more like callbacks and stuff like that that NVD3 offered that the project needed. But now we're able to kind of move to Apache ECharts and get much higher performance. And if you go to their website, which I encourage everyone here to do, they have a beautiful kind of large gallery of amazing charts. And those can all be brought into Superset. And so that's kind of an important bet that the community is making. And it'll be really interesting to see where that goes.

We got a Helm chart in the repo. So we've spoken to many organizations, really large ones in fact that are running Superset at scale. And it just kind of works as another part of their Kubernetes stack that they're already deploying. Sometimes we spoke at organizations that are still using bare metal, all the way to private cloud. Being able to deploy superset since it's kind of always been built in mind for being cloud native. And there's a Docker image that's kind of constantly updated. Those are the main components. As I mentioned, on the backend, we're using SQLAlchemy, to connect into the different databases and even data-like engines out there. So things like Dremio, and Trino, and Presto that you don't classically think of as databases. So those are kind of the main components.

[00:13:17] JM: You might not be the best person to ask about this, but I'm going to ask anyway. The evolution of the data infrastructure world. So you have various large platform companies that have been built in the last five years. You have Confluent built around the Kafka ecosystem, now with KSQL, and all the other tooling they built around that. You have the Spark ecosystem, telling a story of a unified data lake and data warehouse with the Delta Lakehouse concept. You have Snowflake, just telling the story of throw everything in the Snowflake data warehouse and we'll be your unified data lake data warehouse tiered storage system. I realized you're more on the BI layer, but do you have a perspective for the preferred data infrastructure strategy of the average company or the most cutting edge companies these days and how that's been changing?

[00:14:16] SK: Yeah, that's a good question. I mean, like, I've spoken to teams and organizations that are really running all of those architectures. So it really kind of depends on the goals here. So obviously, kind of Deltalake is very focused kind of more on Spark and Databricks. We have Kafka with real-time data. And so there's kind of – And then you still have Snowflake and these kind of modern data warehouses are still kind of very popular. And so I would say, in these cutting edge startups where you see kind of really crazy growth, there's been this emerging – It's been kind of dubbed the modern data stack. And people kind of argue over what technically isn't it. Does it have to be all open source?, Do you put Fivetran in there? Like Snowflake is not technically open source, right?

I think, ultimately, what people – There're kind of a few common components. So one is definitely kind of the focus on being able to understand how your data flows throughout the entire system. And so whether you're using open source or not, having kind of a lot of like data security and privacy controls over your data. It's kind of been a big emerging theme, especially with kind of ongoing data privacy regulations that you're seeing prop-up everywhere. Two, you're seeing some tools just absolutely come in and set themself as a standard. So you have tools like DBT, which is trying to kind of become the universal standard for building models and running those models inside of your database or data warehouse.

And then for data warehouses, you're seeing everything. Like I'm speaking to teams that are using Snowflake. There's still people on Redshift, BigQuery. BigQuery and Snowflake are just kind of stwo of the most popular ones that I've been seeing, although it is a little bit of a biased sample. And now you're even seeing kind of other parts of the stack emerge and become kind of sexy again. So like data observability has become a big theme. So speaking to the earlier theme of like really knowing how your data is moving around and really knowing for a fact that it did move the way that you thought your code worked. And even you have reverse ETLs. Kind of four or five startups in this reverse ETL space where you're processing data in your warehouse and sending it back to Mixpanel, Hubspot, MailChimp, kind of your third-party SaaS tools that want to consume data that you're processing.

So yeah, and then as you move up to larger organizations, they're kind of traditionally – They embrace some of the kind of data lake concepts. Sometimes the use case and the data format is also really important. So if you're working with non-relational data, like geospatial genomic data, like this modern data stack for startups that I just described, it's just generally not as good of a fit. It's just not really optimized for that. So there're startups there. They're doing interesting work, like TileDB, for example. And then more established players, of course, like Databricks and Kafka both in those spaces.

And yeah, and I think like the third part, again, to like kind of go back to our earlier discussion, there's still kind of this interesting gap between – I think the terms are starting to shake up a little bit more, but you're seeing kind of more separation between kind of data science and analytics. Analytics is kind of usually a little bit more focused on end users. There’s kind of excitement around empowering everyone in our organization to write SQL, write DBT models, kind of self-

serve to some degree. And then you're seeing more specialization in data science, where you have a lot of still kind of PhD types doing really complex modeling. And they love these kind of more traditional tools, like for large scale data processing, whether it's Spark or Kafka, or kind of putting data in data lakes. And so it really kind of depends on the goals and what you're optimizing for. But these are kind of general trends that I've seen.

[00:18:12] JM: Do these trends that you've described, do they have any relationship to the BI layer, or like maybe the fact that all this infrastructure has been built out in the middleware, in the lower end of the stack, makes it easier for data availability on the frontend? Like, whatever, 10, 15 years ago, we just had Hadoop to get this kind of layer, this kind of data to the frontend. Now, we just have all this infrastructure. And it's so much easier to get data, basically, because of all these things like reverse ETL, and DBT, and so on just making things smoother.

[00:18:47] SK: Yeah, I think having more options, has a lot of benefit. So I think you're absolutely right. Like early on, if you wanted to manage large datasets, and like you were just stuck in that Hadoop world. And I think we had Spark come out later. And it's just general things are kind of moving because of just how low cost both compute and storage have become. Just the demands of these kind of older school, highly expensive frameworks are kind of slowly disappearing I would say, and you can kind of move a lot more quickly. And dare I say, agile.

And so I think what's great there is that there's a little bit less risk when you're making investments. So the stack, when you're a small team, as a startup of just – You may just have one full time data person even. The stack you can do there will be very different than as you kind of scale. And it's just much easier now to move between the different options. But because I think of open source and because of standards, and just in general, there's just been – people are kind of wary of walk-in in general. And so there's just been a lot more options that have spurred.

What that means for BI though, the ultimate thing is people keep trying to replace SQL. They want some replacement for SQL. But like, ultimately, that seems to be the thing that just it won't go away. If you're building a new database company, and it's amazing, it has all these performance guarantees and cool buzzwords, you're almost surely still having a sequel interface, because there's this great kind of network effect and ecosystem that's been built.

SQLAlchemy, as I mentioned earlier, is kind of a key part of this. Especially for us, it's Superset, the superset community. So I think what's exciting is that even as you have all of these lower costs, still high-performant, large data solutions, and kind of almost an unbundling that's been going on there where maybe in the past you would do everything in Spark or Hadoop. Now you have an entire tool just for ETL, for example, or ELT as it's called nowadays.

And so I think that what that really means though is, ultimately, you can use the tools that you really like for the heavy data lifting. And then as long as it's kind of going into a place that can interface to the SQL language for the BI layer and all the players in the space that we're in, it's great because it means that more organizations are embracing data and they're able to get on board and remix components in the backend. And then for the BI players, still speak in SQL. And so it's still kind of a great thing for us there.

[00:21:31] JM: So coming back to Preset itself, or Superset. Obviously, there's the opportunity to just build yet another BI tool and make it open source. That's definitely an opportunity. Are there opportunities for reimagining BI or building something completely new? In what ways has Preset just done things that are completely different or differentiated from the closed source competitors?

[00:22:02] SK: Yeah, absolutely. I would say that there're a few things that are – I don't think that are crazy, but I think they're kind of hints at the future. So one, one thing that Superset always had is an amazing API for people to work with. And this means you can version charts and dashboards and kind of manage and talk to the BI layer with code, which is like really powerful. I think there's only I think one closed source BI vendor that has that, and it's very limited in what you can do.

And yeah, so because it's open source, because it's kind of built like kind of by this great grassroots open source community, there's been a big push to open up the API even more or documented better, and stuff like that. And I think that's a really powerful thing, because that just means that it's going to lower the barrier even more. It'll allow more kind of dynamic chart creation with code that can still end up kind of powering the BI layer. Because usually, when you think of BI, you think, “Oh, it's end users. It’s humans creating charts and dashboards.” But then now, if you think about a world where you have this BI frontend mixing charts created by

machines and by humans, that's really unique. And I think that's something that Superset and potentially other – There are other open source BI players as well. If they have API's, then I think we're all well-positioned to take advantage of that.

Two, I think another big thing is because, again, of the open source core of Superset, we're able to kind of experiment and really extend features that were done by existing BI vendors, but in very limited ways. So as I mentioned earlier, embedded analytics, embedded BI. I think the new buzzword is actually headless BI is kind of been the term that's been coming up. So using code or an end user interface to create charts and dashboards, but then publishing that to the edges inside of your application. You can imagine a slack.com/analytics, or the analytics page of Stripe not being written by a kind of complicated web stack, but instead just written by data analysts in a BI tool like Superset and just pushed to the analytics part of the product, which is kind of really powerful. And I think it's bringing BI to even more places than originally would have thought of.

I think the last thing is not something, I will say, that's currently easy to do in a Superset project, but it's one that I'm personally very excited about, is being able to actually version control charts and dashboards. And that's definitely going to be kind of, again, enabled by the API and the open nature of Superset. And I think this is kind of a DBT nailed in many ways. Because at my last company, we had this problem. We used to write SQL queries in our BI tool and then literally put them in Git. Push them up to GitHub and try to review and collaborate there. And it was, as you can imagine, just a total disaster. And we just eventually stopped doing it.

And so I think if we're able to bring that idea into the BI layer, that's really, I think, compelling. If you really think about it, like the modern data stack, in many ways, it's really bringing over all these ideas from software engineering and DevOps, and web scale that have been battle tested and have it kind of proven to be the right way to do things and bring them over to the data stack. DBT is kind of get for your data transform models. And data observability is like observability for your cloud infrastructure, except for your data pipelines. And so when you kind of frame it that way, the BI layer has just always been 10, 20 years behind, closed source, proprietary, Windows-based, these like giant two gigabyte applications you open on your computer. And I think I'm excited to see how we can bring some of the ideas from software engineering, like versioning charts, like versioning dashboards, into the BI layer, and encourage those good hygiene and good habits into the BI layer itself.

[00:26:19] JM: As you're going to market, do you see people throwing out their old BI tools and replacing them with Superset? Or do they just add superset to their stack and just complement their existing dashboards?

[00:26:34] SK: We've seen a mix of both, but I'll say it's heavily biased towards the former. I think, essentially, there're two types of organizations generally. There're organizations where, let's say you have 1000 people, and you want to enable maybe 20, or 50 of those people to have the ability to create charts and create dashboards, and then the rest of their organization will consume those. And I think for that, like the existing BI vendors, like Tableau and Power BI, do have somewhat favorable pricing and licensing there were a viewer tier is much cheaper than the creator tier, significantly cheaper. But I think if you want to actually spread BI and data science thinking and turn every team into a data team, just kind of our informal motto and Preset, that you can't have these tearing of users, and that's something we deeply believe in. We want to kind of offer an alternative that actually is much closer to the viewer tier in terms of pricing, but actually gets you the creator functionality. So there's no tearing. Everyone has the same permissions, unless your admin kind of restricts your permission. But we really believe in the idea that everyone should be creating charts, creating dashboards and asking and answering their own questions.

And so when we have companies come approach us on the preset side, it's almost always that they have a mandate to spread this type of data literacy and data skills in the entire company, but the licensing of the existing BI tools is just very cost prohibitive. Paying $80 to $120 per user per month for a thousand people is a lot of money. And so I think that's really where our opportunity is at Preset.

[00:28:27] JM: What have you learned about how companies that are really successful with their business intelligence teams with their reporting teams? What have you learned about how they handle data and how they handle reporting? Are there certain processes or practices or ways of organizing their dashboards or organizing their data that make them more successful?

[00:28:55] SK: Yeah, so I've seen a few different things. I've seen organizations that are very sensitive about people making the wrong conclusions from data that they lock everything down.

You can only view things, view charts, view dashboards. They're not very dynamic. And I've seen also the very opposite, where they put guardrails in place, and they sincerely believe there's actual buy-in from the leadership level, that, “Hey, like this is an important thing for our organization. This needs to be in our hiring process. This needs to be in the tool access level as well,” which is we actually want everyone in a company to have some level of data literacy.

And I think, again, as I mentioned, for Preset, we're especially interested in the latter type of organization. And I think if you're going to do that well, it's definitely non-trivial. I think the main thing you have to do is you still need a data team. You don't throw those wonderful individuals away. But I think you really empower them to actually build the processes and the standards where they're able to create the sources of truth, whether that's the database tables themselves, whether that's views into the data, or things like that. And I think you help them help the rest of the organization. Where, basically, you get this amazing comparative advantage where, the salespeople, they know kind of more about the data in some ways and how to analyze it and how to like exclude things and clean it up. And so you want to empower them, but you also want to make sure that there're guardrails in place that they're not misinterpreting the data or using the wrong tables or the wrong data sources to reach the conclusions or to kind of power the charts and dashboards that they're hoping.

So I think that's kind of a feature that I'm personally very excited about. Having a data team that is kind of creating products in many ways, internal data products, which could be sources of truth, in terms of metadata and specific tables. It could be metrics. So that everyone is understanding what revenue means at a very consistent way. But then the rest of the organization is able to consume those data products and have the freedom and the creativity to use them and to kind of help basically improve their own problem, solve their own problems, and reach their own goals. So that's kind of a future that I'm really excited about. And that's something I think we're going to kind of advocate more as well on the Preset side.

[00:31:28] JM: Are there any particular challenges that the company has encountered in going to market either on the engineering side or the sales and kind of business development side?

[00:31:42] SK: Yeah. So the interesting caveat here is that – So Preset cloud is still in beta. And so we're still in closed beta. You can go to the site, preset.io and sign up if people want to check

it out. But we're kind of quickly marching towards GA, which we're hoping to do in the next two or three months. Essentially, that's kind of our soft goal there. And so yeah, I mean, the challenges we've had is that it's still – Like BI is a very competitive space, right? So even Tableau, which was acquired, I think, 20 billion by Salesforce, you would think that, "Oh, well, at that valuation and whatnot, like that they would be 30% or 60% of the market," but they're not. I think they're only 10% or 12% of the BI market when they were acquired. So it's not a classic winner take all market.

And what that also means is that there's tons of competition, there's tons of players, and people expect a lot out of BI tools. They expect it to do everything that all the existing vendors do at a lower price. And so I think for a company that's three years old, and the Superset project is a little bit more than four years old, that's definitely a big challenge. For us, I think we're going to have to shore up. And that's why I think differentiation is really important for all startups, but especially for us as we try to compete and differentiate against the other BI players that are really trying to offer every single bell and whistle at kind of the same price.

[00:33:20] JM: So your role at the company is developer advocate. Is that right?

[00:33:26] SK: Yep, that's right.

[00:33:26] JM: There's been a real rise in developer advocacy in the last several years. Tell me about how you see the role of the developer advocate and how that plays into what you do at Preset?

[00:33:41] SK: Yeah, developer advocacy is it's pretty interesting, and it's also kind of linked to this rise of the community strategy and community managers as well. I think it's super exciting, because it's basically turning teaching into kind of an amazing grassroots motion for growth. And so I think, especially, developer relations and developer advocacy for open source is especially interesting, because you kind of – In many ways, it feels like not a real job sometimes. You don't kind of have a traditional sprint and kind of hardcore deadlines and traditional KPIs. And I think developer relations succeeds especially well in smaller companies when the goal is more on end user empowerment, growth, things like documentation, doing advocacy, and kind of really evangelizing interesting use cases for the tool as well.

So I think the power of developer relations is really in education and community. So how do you bring people together who are looking to kind of accomplish similar set of goals? And how do you just help them and just expect nothing in return? Because the second you start kind of asking for things or you want to view it as a funnel exercise where you want to start looking at qualified leads or stuff like that, then it's kind of become a sales and marketing motion again. So I think developer relations is exciting because it's kind of a way of just selfless helping and educating and community building. And I think, if you look at a lot of the popular tools and software, like that's kind of just grown like super crazy quick. I think a lot of it is because of community, whether it's like things like Figma, or Notion, even chat applications like Discord, they have very powerful communities behind them and people – Like there's some developer relations of people I've met who actually have tattooed that project onto their body. Like that's how much they believe in this kind of – In that community and the ideas of that community.

So I think it's very powerful when done well. And yeah, I'm excited to see like more of it. Like I would love to go and participate in the community for all the tools that I love and use every day that kind of build great connections. You learn a lot about the tools and you get to use it in ways that you didn't think about. And very indirectly, some small percentage of those people could end up being your customers. And I don't think that's really the goal. Again, my goal, my MO at Preset is to grow the Superset project. And so I just evangelize it, advocate for it, and get people excited about open source Superset and not really concern myself too much with, "Will they convert to Preset cloud customers?"

[00:36:47] JM: So Superset is one of many companies in this lineage of an open source project that basically steals the thunder of some closed source, several closed source projects, closed source infrastructure projects. I wonder if you have any perspective on what domains this works for? Like another reason one is Airbyte, which is kind of the open source Fivetran, which has been gaining a lot of steam. But I wonder if there are domains for which this does not work? Like one company I saw recently was like an open source Intercom, basically, and Intercom, the customer messaging tool. And was thinking like, "That's kind of interesting. But does it make sense for there to be an open source Intercom? Does that actually matter? Do people actually care that much?" So do you have a perspective on for what domains the open source version of X actually makes sense or people actually care about that?

[00:37:50] SK: Yeah. I mean, even for Superset, right? Like we're an end user tool, it's a BI tool, and we have to have designers involved in the project and product managers if we really want to build a world-class BI tool. And it's definitely not easy, right? Even like the traditional collaboration mechanisms of GitHub, and leaving, like opening GitHub issues. It's just very software engineering-centric. And so I think that definitely – I think we do suffer from some challenges there.

But I think, in general, like I think it comes down to like why open source. So there're open source versions of every popular closed source SaaS product. There're open source Notion projects. There's open, there's tons of open source Slack projects. I remember, like 10 of them launched within a few months of Slack becoming super popular. And I think it's kind of – Somewhat, they've been a little bit hit or miss. But I think the ones that have succeeded have succeeded in use cases that the existing closed source vendors cannot do. And I think the most obvious one here is just companies with really strict infrastructure and data privacy requirements. So if you need a certain compliance – Or I spoke to one nonprofit a few months ago that they run everything on bare metal. They were excited about Superset, but they were like never going to use Preset cloud, for example, because they're not even in the cloud. They're still on bare metal. So in that case, they may not want to use Slack. Like that type of organization may want to use Zulip chat or Mattermost, or some other alternative.

So I think the most obvious reason for open source is just you just have your own constraint that is not well supported by the cloud and the kind of requirements there, or there's something about extensibility and customization that you want to be able to do that you can't. So as it relates to Superset, with traditional BI tools, if you want to add a new chart type, you're usually just kind of out of luck. Some tools will let you kind of write some JavaScript code if you know you happen to be data analysts that knows JavaScript, and you can do that way. But most tools are just kind of out of luck. And I think, yeah, if you're able to – But in Superset, you're able to bring your own chart types. You can extend Superset to work with. You could have built a database inside your company, which I know some companies do. And you can add your own kind of support for that. So I think that's kind of an important thing as well, is if there's kind of something about the extensibility, you want to be able to do, you want it to kind of more tightly integrate with the rest of your software stack. I think that's another reason to do open source.

But I think those are kind of the two biggest reasons. I think outside of that, like open source, it is a challenge. And I think even some of these, even like Mattermost and Zulip, these popular Slack alternatives, they offer a cloud-hosted version. And so then that kind of – I'm sure that's probably one of their biggest revenue drivers again. And so at that point, maybe people just prefer the interface or something different about it, but then there, again, much closer to Slack. So I think that's kind of a question for these open source projects to think about, is like there are people who care about open source for the sake of it just because it feels safer or feels more under control. There're people who actually want more extensibility. But I think you have to think about the open source version of a closed source product as its own product. So it has to compete in some ways. And if it can't do that, then I think that project will likely – Well, I hope, I don't think it will succeed as a business, but it might be okay as a kind of volunteer-led open source project.

[00:41:44] JM: As we begin to wrap up, I'd like to hear your perspective on what's in the near future for the company. So I assume Preset cloud is a big focus of the organization. What comes next? Are there are there other like BI features that need to be built or integrations that need to be built? What's in the near term like, I don't know, five years scope of the company that needs to be built out?

[00:42:16] SK: Yeah, so our biggest push right now and focus for the entire company is Preset cloud GA. So as I mentioned earlier, we're currently in closed beta. So getting to general availability on having – Kind of understanding what the technical constraints and guarantees we need there are in terms of reliability, and then also support, scaling up support, scaling up sales. So on the Preset cloud side, for the next few years, Preset cloud is going to be a great Superset experience. We'll have single sign-on, RBAC, all the kind of enterprise features you expect. But it will mostly be the same Superset that people are used to working with. I think on the Superset side, majority of the code that we still ship at Preset goes to the open source project. In fact, I think it's still 80% or 90%. And that's kind of still our crown jewel, is the Superset project.

And over there, there are going to be a lot of interesting developments. Like, as I mentioned, there's the big push to move to ECharts, such as supporting a larger variety of visualizations. And, again, because we're open source, we can experiment with like having a marketplace of visualizations or just even kind of this kind of visualization, a plug-in pack where people can

publish and maybe even build entire businesses off of types of specific plugins. So there's a lot on the Superset side there.

Yeah, but actually to circle back to Preset cloud, I mean, there's a lot of interesting things that I would like to see personally as kind of a data scientist and data geek. So I'm a big notebooks person. So I think there's definitely a chance that we kind of explore adding notebooks support into the preset cloud platform. I think that would really cater to every single persona, because right now, in Superset, we have a no-code chart builder. There's SQL Lab for people who are more savvy with SQL. But there isn't a way for Python or geeks like myself, or Julia geeks, who want to get their hands dirty and do more arbitrary data processing, analysis and visualization. So I think adding notebooks will really kind of complete the platform. And I think we'll make it super compelling.

I have some pet projects of my own that I've been pushing for as well. We're really excited about trying out the idea of instant dashboards. So you can imagine having just pointing to a data source, like Mixpanel, or Salesforce, and then just telling us we're databases or we can offer a temporary cloud-hosted database for you. And then for people who kind of aren't integrated into the data platform and the data platform team in your company, they have API keys for these third-party tools. It'd be really cool to just have the dashboards, have some default set of dashboards just magically appear and not really have to worry too much about how the data was moved around, and how it was transformed, and all that. So those are kind of like two examples of what we call like interesting bets that we'd like to do and we've kind of started to think a little bit more about.

[00:45:28] JM: Cool, well, anything else you want to add about Preset, or BI, or data infrastructure? Any interesting thoughts that we haven't explored?

[00:45:38] HG: The biggest thing I encourage everyone who's listening to think about is really kind of read up on this modern data stack that's been emerging. It's super exciting, and I think it's going to be a really big wave in software. And I think it's – Yeah, we're still just seeing the foundational parts of this. There's been kind of a lot more open source proliferate every part of the data stack. And, yeah, just like let's bring over more principles from software engineering into the data world. And it's really unfortunate, because I'm a data person first and kind of a

software person second. So I'm super excited about this this wave of technologies. And I think every month I'm seeing cool new open source projects and startups being started and funded to modernize the data stack. So I encourage everyone who's excited about these ideas to kind of join that movement. Obviously, if you're interested in Superset, I encourage you to join the community there as well. You might get a hello from me as I kind of hang out in that community Slack a lot. And yeah, just kind of be a part of this movement that's been happening around the modern data stack.

[00:46:47] JM: Okay, great. Well, Srini, thanks for coming back on the show. It's been a real pleasure.

[00:46:51] SK: Thanks for having me, Jeff. This was fun.

[END]