EPISODE 1330

[00:00:00] JM: Shinji, welcome to the show.

[00:00:02] SK: Thanks, Jeff. Good to be here.

[00:00:05] JM: So we're talking today about data engineering?

[00:00:08] SK: Data discovery.

[00:00:09] JM: Data discovery, which is a subfield of data engineering?

[00:00:14] SK: I would say that data discovery is applicable to anyone that works with data, or consume data.

[00:00:24] JM: Got it. So if I'm a data scientist, I need to do this. If I'm a data engineer, I need to do this. If I'm an application developer, I need to do this?

[00:00:31] SK: Yep, because you do touch data. You either produce data as a software engineer, or you manipulate data as a data engineer, or you build other datasets on top of that data, or models, or predictions as a data scientist. But even for like ops people, when somebody is trying to contact their customer and understand where their customer falls in between other customer cohorts, they would have to look at their data, customer data. So I consider that as data consumers, anyone that needs to access data. And they do need to have a good data discoverability in order for them to figure out which data to access, how to use that data, and what to do with it.

[00:01:21] JM: Data ingestion typically takes two paths. There's the OLTP path, which is the transactional stuff, like your account data, my name, my password, my email address, stuff like that, or the OLAP data, which is if you're using Uber and you're moving around in space, you've got geospatial data that's being streamed to a server, and probably being thrown into Kafka, and then thrown into some other places. Do you agree with that distinction?

[00:01:48] SK: Sure. Regarding data collection perspective, and I guess storage and processing perspective, yes.

[00:01:56] JM: How does the schema across those different data types get discovered? Generally. Not talking about select star or something. Just talking general.

[00:02:05] SK: Yeah. I mean, whether it's OLAP or OLTP, you do have information schema underneath, or some kind of meta store that has the structure of data. I guess, for events-based models, not as much. So most of the discoverability right now, I will say, is more focused on structured data. This is primarily because a lot of the data consumption happens on top of a data warehouse, because data warehouse is like the place where all the data eventually arrives that or copied into so that you can join multiple datasets to get like a new insights out of it, or run the analysis on top of it. So yeah, primarily, like Snowflake, BigQuery, Redshift, those types of data warehouses.

I would say, data discovery – I mean, I define data discovery as finding and understanding data. So I want to run analysis about my users. Then I would need to understand or know which table to query or which column I should look into and the understanding part of the data. So that's like finding which table to query, and then understanding the data is going to be about. So is this – When was this last updated? Where did the data come from? Was it only the mobile app users, or website users, or people that have like signed up on some marketing channel? And are there other people that have actually used this data to build other reports or dashboards? Those are like the types of things that is required to really get the understanding of the data, because that's what tells you whether this is a worthwhile data set that you can use to run your analysis.

[00:04:00] JM: As a company is developing, usually it's going to start with some OLTP thing, like a MySQL database, Postgres, or Mongo. You do have this account data, basic user data. So if I'm a new company, I start with my OLTP thing. It's got accounts. It's got users and stuff. And then over time I do the higher throughput data collection kind of stuff, OLAP stuff. In any case, I've got data that's being thrown in a lot of different places, and different kinds of databases. The age of one size fits all is over, right? It's lots of different databases, lots of different data systems, and a bunch of disconnected teams with different permission systems. And the idea of a data cataloging system that overlays across all these different data sets and

data sources is very enticing, but it's somewhat impractical/impossible it seems. How can you actually index and aggregate data across all these different database systems?

[00:04:58] SK: So that, I guess, concept itself is definitely hard, because there are so many different sources and so many different ways that the data has been treated. And that is how a lot of traditional data catalogs have approached the problem, because it's built for the enterprises that has, Oracle, SAP, MySQL, Postgres. And you do have multiple different connectors working in different ways of getting those data. But today, most of those data gets replicated and copied over to either a single data lake or a data warehouse. So the catalog –

[00:05:37] JM: Got it. So data warehouse is the single point of communion for all this data.

[00:05:44] SK: That's right. And when it comes to that single communion area, it may not be exactly the same as what you have in, let's say, your Kafka queue. It will be going through some ETL process before it gets loaded. Or it could be one-to-one replicated. Like if it's mostly like OLDP databases if you are just copying over a user data, then you can. But most of the time, the term modern data stack, that's really all about let's dump our data into one single place. Call it a data mart. So that anyone inside the company can look at the data, not just the one that they created, but also what other people have created. Like I'm in marketing. I can see all my data from Google Analytics, or Marketo, or MailChimp, let's say, but I want to marry that data with customer user engagement data inside of our product, which usually resides in MongoDB, or Postgres, different databases. So how do I join those data together so that I can see what the engagement of my mailing list look like based on whoever that has recently signed into the product? I mean, traditional way to do that is to either query the transactional database to get that events, and then marry it on top of a Google spreadsheet, or Excel. Or you can also copy that data over to a place where you can query both of them, or join them together. And that's what the data mart is for. And that's more of a conceptual thing. But usually, these data marts are built on top of data warehouses, like Snowflake, BigQuery, Redshift, or data lakes, like S3 buckets.

[00:07:53] JM: Well, why don't we just totally launch into something different, which is what you're doing? And then we'll jump around to all kinds of different things. We don't even have

time. We don't enough time. But we'll do as much as we can in the shortest amount of time possible. Let's just jump straight to what you're doing. What is Select Star?

[00:08:09] JM: Okay. I mean, it is a segue. I think it can be a segue. So if you say catalog, it's really just about collecting all different types of metadata in one place. What data discovery is anything that you've connected to regardless of number of data sources. Discovery is really around finding and understanding the data that you have in that data warehouse, or in that data lake. And it's most applicable to this notion of modern data stack and the new cloud data warehouses and data lakes, mainly because these data stacks have hundreds and thousands of tables, in general. So it's not about managing five different, or looking into, or trying to query against five different data sources. It's about querying one data source that has thousands of tables and fields. So how do I find what I'm looking for? How do I know which data exists inside the company?

So going into Select Star, what we are doing is an automated data discovery tool. And we automatically bring out the metadata. We will analyze each metadata based on how that has been used inside the system, used by other data consumers. So what I mean here is we will look at SQL queries, the history of how each user have used or accessed different tables or columns. And we will also look at what type of reports were created and how they were viewed. And with all of that information, we'll give you a summarized view, so that whenever you're looking at a table, it's not just the schema that you see, but we will also tell you where did the data come from. Who are the top users of this inside the company? And what are the dashboards reports that were built on top of this data set? And how are other people have actually used this data in the past? What kind of SQL queries have they run? What tables did they join this with? So it's really kind of like the whole context of that data that today you have to go ask other people around that we will provide that automatically to anyone that wants to learn about the data set.

[00:10:49] JM: Okay. So when do I start using Select Star? How big of a company do I have to have?

[00:10:54] SK: So usually, the sweet spot is when companies hit around 100 to 200 employees. This is just about a time where you cannot just go to like one person or one data team to find

out what this field is about, or whether you can create this table. Because as organizations grow, the data team evolves from being like a centralized team to become like supporting multiple divisions or business units, like finance, marketing, sales, so on and so forth. And then it morphed into each different business units having their own analyst, like sales ops, marketing ops, financial ops. And these different ops analysts will start creating their own data sets and their own reporting. And you'll get to a point where you can't really control it all on their one data team's control.

But this is more severe problem for larger companies. The smaller companies, though, so a couple smaller companies that we are working with right now, like 70-person, 100-person companies, they decided to use Select Star because head of data or the CTO of their company saw how messy it can get from their previous companies at a larger scale if they don't put a good – Like a data hygiene or a structure in place, or what I should call data governance in place. So they are trying to build that with Select Star. And building governance means the way that we help our customers, one is the discovery of the data. Or yeah, like I mentioned, like per data set, you get all these contexts, right? We actually provide that context in aggregate to the data platform team as well so that the data platform teams can find out which are the broken dashboards, which are the titles that nobody's querying anymore? And what are the datasets that everyone is querying so that they can either remodel the data, like build a new materialized views so that these data sets can load faster, versus like which data sets that needs to be documented first, for instance.

So for smaller companies that wants to build a good governance program from the get go from early on, I think, definitely, they can use Select Star. But usually, data teams are very busy. So they get to a point where they are starting to hire more data analysts because the business is growing fast. Each business divisions, like sales and operations marketing, they need more help from the data analysts. And hiring data analysts takes a long time. Ramping them up, training them takes a while. So getting Select Star usually is one of the ways to mitigate, providing a tool that even like less technical users can use to find datasets that they're looking for, as well as to train and ramp up new analysts quickly.

[00:14:19] JM: So I understand what you're doing at a high level. It's data governance, data access, data indexing, data set indexing, right?

[00:14:27] SK: Metadata indexing. We actually never look at your data. We also –

[00:14:31] JM: Right. Yeah. Yeah, obviously. So metadata indexing, meaning, if I work at Uber, for example, I've got so – There're so much data, and I need to find the telemetry data, right? I want to do a join across – Let's say I just want to find the average length of a ride, right? The average length of a ride across Uber. I'm going to need to first find the average length of any ride. Or I need to find the start and end time. Maybe it's the start and end time. I want to find the start time every ride on Uber. Maybe it's actually across separate geos. Maybe like if I have across all my data infrastructure, there's different geos, and each of those geos has a ride start time, right? Indonesia versus the United States has some – Maybe it even has a different schema. Maybe it has a different name for the start time in these different places. And I want to have the global average ride length. I need to find the start and end time, and it may have different names. Your system could potentially index the metadata across that entire system such that I could find the start time in both Indonesia and the United States.

[00:15:38] SK: Yep. So regarding search perspective, we will help you find that field that you're looking for. But from your example, if you were to, let's say, search for start time, there can be many different start time fields in many different tables, right? So am I getting –

[00:15:59] JM: Yeah. My fault. My fault. My fault. Many different start times in many different fields, meaning that in all the different geos, Indonesia, Malaysia, United States, Canada –

[00:16:09] SK: Or geo can be just like another column in that events data base, but there can be many different events database, right? So what we do is we look at how these datasets are being used against others. Meaning have a notion of popularity, and if there is a table that everyone is using for their reports, then we will tell you to use that field. You use that start time, not the other start time of events that so and so created for their test report. So even if you have thousands of matches when you are running a search, a lot of our users find what they're looking for very quickly.

[00:16:53] JM: Now, I really want to understand how do you index this stuff? You got metadata across all this different heterogenous data systems across Uber. Uber has thousands and

thousands and thousands of employees. I'm an engineer in one part of the company, and there's engineer another part of the company. If you come in and say, “Okay, you run Select Star.” You sell the Select Star product to some particular engineer, maybe it's the VP of engineering in some division of Uber’s engineering task force, how do you get them to allow you to hoover in the metadata for every single part of their infrastructure?

[00:17:29] SK: So first of all, we don't need to see every single part of their infrastructure because, again, it's only just really for their meta, or like their data mart, like data warehouse perspective. But even then, how do we get that access? I mean, we've passed their security, I guess.

[00:17:47] JM: Seriously speaking. Okay, like you pass the security rules. But can you actually have some – Do they just have an edict from on high. They say, “Hey, we're using Select Star for our data catalog system. We need to be able to data catalog every single thing in the infrastructure.” Or do you just have to wind your way through the infrastructure and win over teams one by one? You see what I’m asking?

[00:18:08] SK: Yeah. Usually, we partner with the data platform team that manages that data infrastructure part. And I'm talking about data warehouses, BI tools or –

[00:18:19] JM: And there usually is a data platform team that has visibility into everything.

[00:18:22] SK: That's right.

[00:18:23] JM: Got it.

[00:18:23] SK: And what we are asking for is create a service account for Select Star to use to query your meta store. And we only need the access to that meta store and query history. So they create a very special account that doesn't have access to the data values, but only the metadata. Or if when they integrate, or even we integrate with their BI tools, they create a read-only API access. Then we use the API to get the metadata for BI perspective. Because the other big part about this visibility, especially for the data team, is I change this table, then is there any business dashboard that's going to work? Understanding that impact can happen by us

connecting to their BI tool, Tableau, Looker, Mode, and tell them, “Hey, here are the reports and dashboards that is using or utilizing the data set in that table today.” So if you change that column, this dashboard is going to crash. And here are the top users that are going to get affected. And here are the people that have created those dashboards. So maybe you should let them know first.

[00:19:41] JM: Which is one thing I want to get very clear here, because I want to move past this at a personal level. To me, mentally, mentally, in my mental framing, it's hard to sell a tool across an entire organization. Like as big as Uber, it's very, very hard to push a tool that's on an entire organization or even go bigger, like Amazon. Amazon is so big. There's no way you could sell a tool that would actually aggregate all of the data across Amazon.

[00:20:06] SK: That’s very true. Yeah, so it will be division by division.

[00:20:10] JM: So you’d have to do division by division. So you could have a catalog for one division of Amazon, but you could never have a catalog across the entirety of Amazon.

[00:20:18] SK: I wouldn't say never, but it will take a long time.

[00:20:20] JM: Not never. But you’d basically need to build PageRank across the data catalog system to actually have an effective data catalog for Amazon.

[00:20:28] SK: That's right. The whole popularity system is kind of like PageRank, because we look at how people are actually utilizing the data and use that as a popularity score, which is the score to rank.

[00:20:38] JM: How do you get to that? Does that mean that you have some listener on all the different data sources?

[00:20:43] SK: So you keep going back to all the data sources. Again, we're only connected to like your main data warehouse.

[00:20:49] JM: Got it.

[00:20:50] SK: So main data warehouse. Why do we do that? Because of transactional data, it's mostly like write, ingestion, any production data that's really for the application itself. What we are wondering about is how is the data actually consumed inside the organization. And inside the organization, how is it being consumed? It's usually consumed by business stakeholders, or product managers, for them to make decisions about whether they want to launch this feature, or which customers to tap into, or what type of marketing that they should run against which customer, so on and so forth. A lot of business decisions get made. And that's where the data consumption happens.

And that happens in the medium of BI tools, or analysts running their SQL queries, mostly select queries. So how we get that information is we look at the query history. And we don't look at like all the history. We don't look at just all the logs. We are only looking at particular queries that will give us information about how the data is being used, or how the data is being created. So how is the data being used? We look at the select query. We look at which table, which column has been accessed from this query. And who ran this? And how often did this run? That type of information is what we gather in order to calculate our popularity or PageRank? So did that make sense regarding, yeah, how we get the

[00:22:24] JM: It does make sense. It does make sense. But it kind of makes me curious about the hooks for data warehouse, right? Because you basically – So the hooks into the data warehouse though, so is this how pretty good listener infrastructure for like Snowflake and BigQuery and all this stuff?

[00:22:41] SK: So a lot of modern data warehouses today have a view that has this history that you can query. Not everything does it. So like Postgres, we have to like listen to the raw log. And if it's not configured, then we don't have enough history to generate like a sound popularity. But it just means that we just have to continue listening to it after, yeah, we integrate with those databases. So yeah, it depends on the data store. Same what BI tools.

So for something like Mode, we cannot get like granular information directly through the API today. So usually we encourage our customers to integrate their “discovery database”, which

has more activity information inside Mode. So we query that database instead of using just the API. So things like that.

[00:23:39] JM: What's your take on – Okay, you and I talked about this a little bit in the past. So this is like kind of a hot button thing. And it's like not exactly on topic, but you build a streaming company. You sold it Akamai. You know what I'm asking. No, you don't know what I'm asking. Okay. Streaming was like this mirage. We all thought we were going to go for streaming, right? And then we all went to the data warehouse, right. We all thought we were going to have like – No. NO. Didn't we all think we're –

[00:24:06] SK: I don't know we think we all did, but –

[00:24:09] JM: Okay. Well, you clearly did.

[00:24:11] SK: I did. Yeah, I did.

[00:24:11] JM: You thought it was a bet on. I remember all the streaming systems, Apache Beam, Apache Storm, Apache Spark. What is it? Google Beam? Beam, Sparks, Storm, Flink, Heron –

[00:24:26] SK: Samza.

[00:24:27] JM: Samza, Concord.

[00:24:29] SK: Yes, Concord.

[00:24:30] JM: Which is yours.

[00:24:30] SK: That's right.

[00:24:31] JM: Very fast C++ streaming system.

[00:24:33] SK: Glad you remember.

[00:24:35] JM: C++ is way better than Java for this application, right?

[00:24:38] SK: Exactly. That's why we built it in C++.

[00:24:39] JM: Why do all these people build Java? It's crazy, right?

[00:24:44] SK: Ah, yeah, I mean it's also –

[00:24:45] JM: Although, Rust has got to be where it's at these days, right? If you wanted to build a streaming system, not that anybody is, you would do it in Rest, or Go maybe? Not C++.

[00:24:55] SK: C++ still works. But yeah, it's just the tool of choice.

[00:24:59] JM: But streaming – So there is this dream, right? Streaming, you should do streaming. Just move everything through pipes throughout your infrastructure. But instead we're doing the warehousing. We threw everything in the warehouse. Where are the pipes?

[00:25:11] SK: Pipe is still there.

[00:25:13] JM: But it's all Kafka.

[00:25:14] SK: Yes. And streaming is still there. I think – Well, back then, like 2012 or 2013, well, yeah, 2012, 2013 was when we were using Storm. Then I started Concord, which was 2014. Spark was still very young. Kafka didn't have KSQL, or any real processing mechanism. So, I mean, we have to build it. Today, there are a lot of options to process streams on top of Kafka. So does that mean that stream processing is not necessary? I don't believe that. I think there are certain use cases like a real time add serving, or like algorithmic trading, like things like that can still run on stream processing, and is the right thing to run on top of stream processing.

But why did we all go to data warehouse instead? I would say that's mainly because the primary consumption of data doesn't always happen, or it doesn't really always have to be real time.

There are certain things that's important to be real time. Like if you're Uber and you're looking at the traffic, and you need to support a driver or a user right away, and you do need a real time process data. And hence, yes, you need a streaming.

But if you are looking at which city to launch next month, or which your marketing campaign you'll want to run in the next week, you just have to look at your historical data. And what just happened in the last two minutes, five minutes, is not as helpful. Rather, you would rather have a data over the last year and two years' worth of data, because you want to look at long term trends. And a lot of data analysis, data science, or consumption of data regarding making business decisions usually happens by looking at the trends and long periods of time.

[00:27:33] JM: But we did think for a while that everything was going to be done with these streaming systems. Like if you wanted to calculate the coolest thing over time, I don't know, you have like a variable. That's the coolest thing over time. And then all the streams wash over that variable and just refresh it with the coolest thing?

[00:27:51] SK: That's true. At the same time, there're just too many dimensions to recalculate everything too.

[00:27:56] JM: Right. So we had to stuff everything in the warehouse.

[00:27:58] SK: Yeah. So you do have to index it in a columnar way. And I think the other big part is starting to look at the compute and storage separately, which I think has been another like a big breakthrough.

[00:28:14] JM: Everybody always says that. Snowflake is the breakthrough, because it separates compute and storage. And I'm always like, "What? I don't understand what you're talking about. What does that even mean?" I mean, I don't have a microprocessor and like a hard disk. Isn't that separation of compute and storage?

[00:28:30] SK: I think it comes down to the cost. So traditionally, the only data warehouse – Yeah, first cloud data warehouse used to be redshift. And then they charge for however much data you have. Enterprises have a ton of data. And most of companies are starting to have more

data. So it's like, “Well, I have all this data. And I do want to put them all in one place. So I have the option to run my analysis for the last three months, six months, a year and five years.” But in order for me to – If I were to run that in Redshift, or traditional data warehousing model, I have to pay for all that five-year worth of data just for that storage, even though I'm not doing anything with that data. Usually, like the most important data for me is the last month data, for instance, right? Whereas like Snowflake came in and said, “Hey, it's okay. You can put as much data as you want. And you can run that query only for the last month. And we will just charge for that computation regardless of the amount of data you already have.”

[00:29:46] JM: Gotcha.

[00:29:48] SK: So it's really on the cost side.

[00:29:51] JM: Can I switch topics again?

[00:29:53] SK: Yes. You're the host. So what am I going to say?

[00:29:56] JM: Why are there so few female entrepreneurs? Seriously?

[00:30:00] SK: Oh, you're like going a lot of different sideways.

[00:30:02] JM: Well, why not?

[00:30:04] SK: I don't know why you asked me, though.

[00:30:08] JM: Because you're an entrepreneur.

[00:30:09] SK: Yes, I'm an entrepreneur. Serial, serial entrepreneur. There's more than there should be. By the way, did you know the CEO of my company is a female?

[00:30:17] SK: Yeah, I met her. So I think there are a lot more female founders now, which is really awesome.

[00:30:23] JM: Yes, there. It's great.

[00:30:26] SK: Yeah, like CEO Front **[inaudible 00:30:28]**. Vanta, you interviewed, right? So yeah, it's really awesome to see these female entrepreneurs these days everyone like killing it.

[00:30:37] JM: Yes. But the sort of like societal or institutional pressures against it, are those fading away or are they still pretty predominant?

[00:30:49] SK: I think you need to be a little bit more specific.

[00:30:51] JM: Well, so there's clearly like some sort of societal pressures, or unless you're a believer, or evolutionary pressure, if you believe it, against – I mean, because proportionally, you would assume – So my mom is an entrepreneur. So you would assume that, "Okay, there's a 50-50 gender imbalance, or a 50-50 gender balance. Wouldn't you assume that there'd be 50-50 entrepreneur balance? Or maybe not?

[00:31:17] SK: Also, like are we talking about just tech? Or are we talking about like – There are so entrepreneurs outside of tech.

[00:31:26] JM: However you define entrepreneurship. My mom is a real estate professional.

[00:31:29] SK: Yeah. There are a lot of female real estate professionals. There are a lot of female like fashion business CEOs or people that run different freelancing agencies. So I think entrepreneurship – and I actually don't even know what the stats look like between female versus male in that regards. I totally understand that you don't see as many female tech entrepreneurs in the data space, or in the infrastructure space. That, I'm with you, and I think it's just, hopefully, over time more people will start companies. But everyone has their own – I don't know. Not everyone wants to always start a company, right?

[00:32:12] JM: Yeah, but I don't know about you. For me, working at a big company is much harder than operating my own company.

[00:32:20] SK: I agree. I feel the same way.

[00:32:23] JM: There's this narrative that it's actually harder to run a company. It's not harder. It's much easier, right? I mean, how much –

[00:32:30] SK: I don't know about much easier? It's just a different thing, I will say. So you just grok to more entrepreneurial working style. And I do too. And some people do, and some people don't.

[00:32:44] JM: Yeah, but it's almost objectively preferable to being enslaved by a giant corporation, right?

[00:32:52] SK: Okay.

[00:32:54] JM: Don't you agree?

[00:32:55] SK: I think it's a personal preference. I think some people do like having the security and resources.

[00:33:02] JM: That's what venture capital is for.

[00:33:05] SK: Not everyone can get venture capital, Jeff.

[00:33:08] JM: Sure they can these days. There's so much money.

[00:33:10] SK: That's what you say. No.

[00:33:11] JM: Who can't get venture capital?

[00:33:12] SK: A lot of people.

[00:33:13] JM: Who?

[00:33:13] SK: So many people still have trouble fundraising.

[00:33:15] JM: People with bad businesses and bad pitching skills.

[00:33:18] SK: That's what you say. Right. But, I mean, it could be either. It could be only one of them. It could be none of them. And they just don't have the network. They just don't have –

[00:33:27] JM: Don't have the network?

[00:33:30] SK: Right.

[00:33:31] JM: I mean, can't they just like email people or like go on LinkedIn and like message people?

[00:33:34] SK: Nobody responds to your cold emails if they don't know about you.

[00:33:37] JM: That's not true. Well, if you convince them.

[00:33:40] SK: Yeah. So you have to convince them, and that's a lot of work.

[00:33:43] JM: No it's not.

[00:33:46] SK: For you and I are, it's more – It could be more natural. For other people, this is something like that they have to learn that they didn't do before.

[00:33:53] JM: Before and everything.

[00:33:54] SK: That's right. And that's why it's hard because they don't have –

[00:33:57] JM: But it's not hard. It may be time consuming, but it's not hard.

[00:34:01] SK: Okay. So it's not hard. It's just maybe they get unlucky. I feel like –

[00:34:06] JM: Unlucky?

[00:34:07] SK: You have infinite shots on goal. You never –

[00:34:10] SK: I don’t think it’s always infinite, because –

[00:34:12] JM: Investors are a dime a dozen these days, right?

[00:34:15] SK: For you, I'm sure, but not for everyone. I am not kidding. I'm not kidding. I mean, I think there are a lot of founders out there that has – Well, I don't want to call it good businesses, but solid ideas or feasible ideas. And some people will get funded and some people just won't.

[00:34:37] JM: And it's a deterministic process. It's not a random process.

[00:34:40] SK: I think it's random.

[00:34:42] JM: I mean, it's random in the short term. Deterministic in the longer term.

[00:34:47] SK: Okay. I think it's really just like there is some luck involved in the –

[00:34:52] JM: In the process.

[00:34:53] SK: In the process of getting, yeah, funded. Meeting the right people and – I mean, because there's so many investors, right? You may never meet the right investors, and you only end up meeting hundreds of not applicable investors and just have to give up. Who knows?

[00:35:11] JM: Yeah, deciding to give up as a decision.

[00:35:13] SK: Sure.

[00:35:15] JM: No, I'm serious. Like it's a deterministic process. People make it out to be this luck-based non-determinism. It's just a straightforward – It's like going to McDonald's. It's like going to work at McDonald's.

[00:35:27] SK: What I'm saying is that for people that are getting funded, like think of it the other way around. People that got funded, do you think they got funded because they are actually the best of the best and it's only their married in the business? It all doesn't have to check the box. But there is some luck involved.

[00:35:51] JM: In the short term.

[00:35:52] SK: Sure, in the short term.

[00:35:54] JM: But if you continue working at it, and you have a sound business idea, you will get funded. It's that simple. Have you ever met somebody with a really good business idea that is unfunded for like two years and is also reasonably good and reasonably bent on improving? I've never seen that scenario.

[00:36:13] SK: Sure. I mean, if you put it that way, I'm with you. All I'm saying is when somebody like leaves their cushy job, because they are determined, ambitious, and they want to build, make this new company to life. And like, I don't know, because of some like unfortunate, not being lucky, they didn't get funded for, let's say, two years. I mean, it's hard for them.

[00:36:42] JM: Who is it hard for? Wait. It’s hard for who?

[00:36:45] SK: People that are out there. I mean, all I'm saying is like getting funding for first time founders and founders that isn't necessary like the type of good at fundraising is not that straightforward. It's all I'm saying.

[00:37:01] JM: I mean, how much material is there on this topic? There's like a bajillion blog posts, tons of books.

[00:37:05] SK: I agree. Yeah. I mean, I think 80% of the time, most people don't do their – Put in the work to actually.

[00:37:13] JM: 99% of the time.

[00:37:15] SK: They get work. I think for seed stage, yes, it is a much better environment than before, for sure. But at the same time, there're so many companies.

[00:37:25] JM: And?

[00:37:26] SK: So it's hard to differentiate, especially at the already stage.

[00:37:29] JM: Hard to differentiate for who? The investor?

[00:37:30] SK: Yeah.

[00:37:31] JM: So are we talking about investors now? Or are we talking about entrepreneurs?

[00:37:33] SK: For entrepreneurs. We just talked about funding. Funding is between the entrepreneur and the investor.

[00:37:40] JM: Okay. Are you expressing sympathy for the investors now? Or what?

[00:37:43] SK: No, no, no. It's the job of the entrepreneur to demonstrate the competitive advantage or differentiator of their product in the business, right? And that's harder to do for them, because there are so many other companies always to –

[00:37:58] JM: What? How many other good data catalog companies are there? Or data discovery companies?

[00:38:02] SK: Oh, not many.

[00:38:03] JM: Exactly. Exactly.

[00:38:04] SK: Just kidding. I mean –

[00:38:06] JM: Once you know your market, it's so easy to define the differentiation.

[00:38:10] SK: I agree. But some investors still get worried about competition.

[00:38:15] JM: If you're trying to tell me investors are incompetent, I won't argue with you.

[00:38:18] SK: Yeah, I mean, don't get me wrong. I get lot of nos.

[00:38:20] JM: Because they're a large class of people.

[00:38:22] SK: I heard a lot of nos on my fundraising, even though they all agree that our product looks the best and is the most advanced. Because they feel like, "Well, I mean, it still is pretty early." Or, "Oh, I think the competition is going to be really fierce."

[00:38:38] JM: Do you know what the real problem is with a lot of investors is like they're all envious of entrepreneurs. So they have to overcome their envy in order to actually make a bet on – Okay, we don't go that far. We won't go that far. Yes and no, fine.

[00:38:48] SK: Yes and no. That's a lot.

[00:38:51] JM: We won't go that far. But honestly, all I'm trying to say here is I think there's a demystification that needs to happen, because being an entrepreneur is actually easier than being an employee. I think for most people, every person I've seen who has gone from the transition of being an employee to being an entrepreneur has been happier, more relaxed. Okay, not everybody. Some people go through some stressful stuff. But usually, that's stressful stuff is just a speed bump on the road to a better life.

[00:39:20] SK: Sure. That's a big deal. That's not a sure. That's a big deal.

[00:39:24] SK: Yeah, but that's only if there are business is running, and sound, and can sustain.

[00:39:32] JM: Yeah, sure. I mean, yeah –

[00:39:34] SK: But that's the hard part. Either make it profitable –

[00:39:37] JM: We're all adults here, right? Like do you want to work in the corporate sweatshops? Or do you want to just go your own direction?

[00:39:47] SK: Actually, again, I think it all comes down to that formality as like your personal preference, right? Even if I'm like the amazing engineer, right? If I decided to, let's say, set myself to be like a dev shop or like a freelance agent, then like I still have to go out and get clients. Obviously, if I'm really good, then maybe I don't have to worry about the referrals. But I still have to kind of figure out how to manage those clients. And that's a different type of skill. Whereas if I'm in a company, then I just get paychecks and somebody's going to –

[00:40:19] JM: Oh, no. It's not you just get paychecks. It's you show up way too early, you stay way too late. You have to attend a bunch of boring meetings that stupid people force you to attend.

[00:40:28] SK: So that's what entrepreneurs would feel like. And for those people –

[00:40:34] JM: No. Those are facts. Those are not feelings. Those are facts.

[00:40:37] SK: I don't know. Some people would like just sitting in the office and sit in the – I don't know, random meetings too, right?

[00:40:46] JM: Another question.

[00:40:47] SK: I don't know what you want to get to.

[00:40:49] JM: What's going to happen to all those office buildings in downtown San Francisco? Those are vacant.

[00:40:52] SK: Oh, man. I have no idea.

[00:40:54] JM: I have two ideas.

[00:40:56] SK: Okay.

[00:40:58] JM: Can I pitch you? Do you play an investor? Are you an investor? Do you invest also?

[00:41:02] SK: I invest.

[00:41:03] JM: Oh, excellent. Okay, wonderful. Well, I'm looking for a pre-seed check for a few ideas.

[00:41:11] SK: Gosh! Well, you invest. So you think you would invest in these ideas?

[00:41:16] JM: I'm not going to say yes or no. I don't want to bias. I don't want to bias you.

[00:41:20] SK: Okay, what's your idea?

[00:41:21] JM: Are you ready? Okay. Do you want to hear about the business call – Actually, they both have the same name. I'm just not sure what the company does yet. Okay, the business is called Downtown. The idea is what do you do with all these office buildings in downtown? Downtown X, San Francisco, Atlanta, Austin, all these places that have downtown office space that is completely unused. It's never coming back. Nobody's going back to those offices. Nobody wants to work in the office. Hybrid is not a thing. Nobody wants go to the office. Do you agree with me there?

[00:41:51] SK: Yeah.

[00:41:52] JM: Okay. So what are you going to do with the offices? I see two options. Let's go with the most obvious one first. We live. We work. WeLive could have worked, right? Did you see the documentary, the WeWork documentary?

[00:42:04] SK: No.

[00:42:04] JM: Okay. Have you looked at WeWork? You know what WeWork is?

[00:42:06] SK: Yeah, yeah. Of course, of course.

[00:42:07] JM: So WeWork had their second business. WeLive, the idea is you live next to where you work.

[00:42:15] SK: I was at the very first WeWork in New York before.

[00:42:18] JM: Oh, really?

[00:42:18] SK: Yeah, 20 times.

[00:42:20] JM: That was magic.

[00:42:21] SK: In SoHo.

[00:42:21] JM: Did you see Adam? Whatever his name is?

[00:42:24] SK: Adam Newman?

[00:42:25] JM: Yeah.

[00:42:26] SK: Yeah, I don't know.

[00:42:27] JM: Did you hang out with him? Does he walk on water?

[00:42:30] SK: Oh my gosh!

[00:42:31] JM: No. But, seriously. Did you see WeLive? Do they have a WeLive there?

[00:42:36] SK: I've never been to any of the WeLives, but I know that the – I've looked into it, because I was looking for, I don't know, a place to stay in –

[00:42:44] JM: So you considered it. You evaluated it.

[00:42:46] SK: Yeah, but it didn't make sense.

[00:42:49] JM: It doesn't make sense. Why not?

[00:42:52] SK: So I don't exactly remember. I think it was just the price. It didn't make sense at the time, because I was looking for like a day or two stay, but the main purpose of WeLive is like communal living, right? You live for like a month or like you get like a whole lease and then you share like a lounge, and have like a community.

[00:43:15] JM: Which is awesome. I would do that.

[00:43:20] SK: Okay.

[00:43:21] JM: Who wouldn't do that? People with enough money, right?

[00:43:27] SK: So why would somebody want to do like downtown instead of like creating their own communal living? Because there are a lot of communal, I think, houses in San Francisco too.

[00:43:38] JM: What? No. If I just graduated college, I don't want to go create a commune. I want to go live somewhere. I want an apartment, right? But I want an apartment with cool people. Okay, you take those office buildings, just throw a bunch of –

[00:43:50] SK: Yeah. Actually, you should pitch OnDeck. OnDeck is doing something like that.

[00:43:53] JM: I'm pitching you.

[00:43:54] SK: Okay, great. Because I'm just telling you what I've seen so far.

[00:44:00] JM: I'm pitching you. You're the investor. You're playing investor.

[00:44:01] SK: Oh, okay. Yeah.

[00:44:02] JM: I'm just trying to get your like a thousand dollars or whatever to get this thing off the ground.

[00:44:06] SK: Okay. What else?

[00:44:07] JM: I can buy a lot of bunk beds for a thousand dollars. So you put up bunk beds.

[00:44:10] SK: No. It doesn't work.

[00:44:12] JM: It doesn't work? Why not? Put bunk beds in the office buildings, and then you got tables in the other room. That's all you need. Get fast WiFi, some free coffee.
[00:44:19] SK: Office buildings you can just have people sleeping over?

[00:44:23] JM: Well, you can technically do it today, but you can win over those land owners, right? Because there's nothing going on there. Nobody is in there. They're just losing money. Throw some bunk beds in there, rent them out for a thousand dollars a bed.

[00:44:34] SK: I really don't know enough of real estate to comment in regulations.

[00:44:37] JM: I mean, have you been a broke student before, or a broke graduate?

[00:44:40] SK: Oh yeah. Yeah, yeah, yeah.

[00:44:42] JM: Yeah. So would you pay a thousand dollars to live in an office building in downtown San Francisco? Hell yeah, right?

[00:44:47] SK: But there's nothing going on in downtown San Francisco either.

[00:44:51] JM: Except for downtown the coolest bunk bed area with cheap WiFi.

[00:44:57] SK: Okay. It's like going in circles.

[00:44:59] JM: There is some cost, right? Maybe it's $500. They're totally unused right now, right? Totally unused. There's nothing going on in those buildings, right? They're literally empty. Do you ever walk around downtown and you just go, "Hello? Hello? Hello?" There's nobody downtown in San Francisco.

[00:45:15] SK: Right. So why would anyone want to live in downtown? There's nobody. Nothing is going on.

[00:45:21] JM: You get the opportunity to be like the first colonist in the abandoned ghost town of San Francisco.

[00:45:26] SK: But you just said –

[00:45:27] JM: I would go there with my friends.

[00:45:30] SK: Okay. Okay. I don't buy into the idea. Next.

[00:45:34] JM: All right. Fine. Second one is a data center business. Fill up all these office buildings with data centers.

[00:45:40] SK: I just don't know if it would the cost margin will be there, because San Francisco real estate overall is so expensive.

[00:45:48] JM: What are you talking about? There's nobody in those buildings.

[00:45:50] SK: I understand, but still, compared to like, I don't know, New Jersey or –

[00:45:56] JM: There's nobody in those buildings. Salesforce Tower, there's nobody in that building.

[00:46:00] SK: Okay. Yeah, I'm sure Equinix and all these data center companies are starting to buy up these office space, for sure.

[00:46:07] JM: They could.

[00:46:07] SK: Yeah. I'm sure they are.

[00:46:09] JM: All I'm trying to say is what else are you going to do with this space? You can't knockdown the buildings. Micro-performance centers? Drone depots?

[00:46:17] SK: I don't know. I just never had the time to really think about it. I'll think about it a little bit.

[00:46:21] JM: Now you do. No. No. Okay. I can give you like 30 seconds.

[00:46:24] SK: Oh, no. I can't come up with this. Why is this Software Engineering Daily? There's nothing software engineering here.

[00:46:30] JM: You want to talk about some data catalogs?

[00:46:32] SK: I don't know.

[00:46:32] JM: I mean, you need a place to put the data catalog, right? You need the data centers.

[00:46:37] SK: It just goes on to the cloud. We don't take that much space.

[00:46:42] JM: That's true. Because you're just kind of – You're basically an index, right? You're just an index over all your metadata pretty much?

[00:46:48] SK: Yeah, index and the analysis on top, the queries. So, yeah.

[00:46:53] JM: The queries. Right. Okay. So if I query, you save the query so that people later on have a fast reference to previously saved queries.

[00:47:04] SK: I guess, yes. That's one thing we do.

[00:47:07] JM: That sounds useful.

[00:47:07] SK: Yeah. I mean, the useful part is for you to look at user table and look at the top five queries that everyone has used around the table. So even if I have never used the user table before, I can see that, "Oh, Jeff always runs this query in user table. I should run that too. And, oh, here are other things I can join with."

[00:47:33] JM: So there you have like basically –

[00:47:34] SK: I can join it with user table, with marketing table, with customer table.

[00:47:38] JM: Right. So this is critical. This is the social networkization of the enterprise, right? Because it's like one click to share your query.

[00:47:47] SK: It's not even share your query. It's already there.

[00:47:49] JM: It's already there.

[00:47:50] SK: Yeah. And I can visit your profile and see which queries you ran.

[00:47:55] JM: I'm not sure I want you to see that.

[00:47:56] SK: And which tables you work with the most.

[00:47:58] JM: I'm not comfortable with you seeing what I query.

[00:48:00] SK: Or all the reports that you have created.

[00:48:01] JM: Okay. This is getting creepy.

[00:48:03] SK: I know. Some customers have mentioned that. I mean, we plan to customize it. But most companies, this is already an open information.

[00:48:11] JM: I only want to share my queries with my close friends.

[00:48:15] SK: Okay.

[00:48:16] JM: Can you do that? Do you have stories yet?

[00:48:19] SK: What do you mean stories?

[00:48:21] JM: Like Instagram stories for my data queries?

[00:48:26] SK: No, not yet. Next year. So I guess regarding that, because it's like docs. You create your thing. And then it's technically owned by the company. And it is all about like – It's all about like company data. So who cares, right? So most people, most tech companies are okay with having this available and being able to index, by user, by table, by team. So I can go find out what finance team use all the time, marketing team uses all the time, or sales team uses time. So this is really helpful for new analysts to ramp up with the data that they work with.

But something that definitely is important for financial services customers, and a lot of enterprise customers is this notion around gating the access. Hence, I shouldn't be able to see your queries, because most of the time, this is mostly for investment researchers. I'm working on this model, like a trading model. And I should just come up with my own instead of like looking over your shoulder to see what you're doing. And so that's a new customization that we plan to add so that each user can discover the categories of the data set that they have access to, but then may not have the full access of how the data is being used by others.

At the same time, this data is still really relevant to and important for the data managers. If I'm in the bank and I'm buying a bunch of external data, for me to get to see how much of these

external data is being used by which division and by whom, I can calculate the ROI of these data sets that I spend a lot of dollars on, and decide whether it's worthwhile keep doing it or not.

[00:50:24] JM: And so all your competitors, they're not very good, right? Or wait, some of them might have been on the show recently, or I can't remember. Who are the best ones?

[00:50:33] SK: You have everbody here. I don't know. I mean, so here's the thing. Traditionally, companies like Alation and Collibra have been in this space for a long time. They have a lot of features.

[00:50:44] JM: They define the data catalog category, right?

[00:50:46] SK: Yeah. They have a lot of features. I mean, we don't have that many features.

[00:50:49] JM: By the way, you don't call yourself data catalog, right? You're a data discovery.

[00:50:51] SK: No, we don't. Yeah, we call it data discovery. Data catalog is part of the feature of the discovery.

[00:50:56] JM: Gotcha.

[00:50:58] SK: The new players, most of them are still building, I guess, in stealth. So it's hard for me to tell how much hand-in-hand we are. But from what we saw from the open source alternatives and what we hear from other customers, currently, it's a lot of work to feed in that metadata yourself, for you to build your own popularity engine, for you to customize a search engine. All that type of stuff is the part that will take a lot of engineering resources. And that's when some companies decide to choose a vendor solution like ours, or all-in in solution that's hosted and managed for them.

[00:51:42] JM: But seriously, though, so other main competitors, the modern competitors, you can't compare yourself like the old ancestral competitors. Compare yourself with the modern competitors. So there's the stuff out of LinkedIn, right?

[00:51:53] SK: DataHub?

[00:51:54] JM: DataHub.

[00:51:54] SK: Mm-hmm. It's an open source project.

[00:51:55] JM: And then the other one. There're two companies around? Oh, Amundsen, that's the other one.

[00:51:59] SK: Yeah. So that's the other data open source.

[00:52:02] JM: What's the company built around that one?

[00:52:04] SK: Stemma.

[00:52:04] JM: Stemma, right. I think I invest in that one. No, I didn't.

[00:52:09] SK: What?

[00:52:10] JM: There's some company at Lyft I invested in. It was either that one or the data engineering one. Okay, so you got Stemma. You got Dataflow, or DataHub.

[00:52:21] SK: And then Acryl Data. I think that's also from –

[00:52:23] JM: Agryl Data?

[00:52:23] SK: Acryl.

[00:52:24] JM: Acryl.

[00:52:24] SK: Yeah, Acryl.

[00:52:25] JM: What's the story behind that one?

[00:52:27] SK: That one is also from LinkedIn, DataHub.

[00:52:29] JM: Right. Okay. This is like the Rift, whatever. Are you open source?

[00:52:36] SK: No.

[00:52:37] JM: Why not?

[00:52:40] SK: I mean, open source is a very different approach.

[00:52:43] JM: Which is more work. It's more work, right? And not everyone wants to manage an entire community, right?

[00:52:50] SK: No. I think open source has its own merit. We just didn't come out of the open source foundation or commercial company that was already using the project inside a company. That's why. It's just because I built it from scratch. My team and I built it from scratch. We did look at DataHub, Amundsen. What's the other one? Atlas. Like all the open source data catalog that exist to see whether we should build Select Star on top of it. It was just simpler for us to build it from scratch. It's mainly also because there is a lot of like LinkedIn-specific, or Lyft-specific packages and libraries that we have to adapt in order to make that work. Some of them outdated or some of them was there just for the purpose – Only because it was used in the internal purposes of either LinkedIn of Lyft or any other companies. And, I mean, we had our own model of how we view metadata. So we just ended up building it ourselves.

[00:53:59] JM: What do you mean your own model for how you view metadata? Your own perspective on how –

[00:54:03] SK: Yeah, yeah, our perspective.

[00:54:04] JM: What's different between you and Stemma, and DataHub and whatever else?

[00:54:08] SK: Like, underneath? Or I think I don't even know how to explain this. What's the main difference regarding the architecture of the data catalogs that we have?

[00:54:18] JM: Yeah, yeah.

[00:54:20] SK: I mean, it really kind of comes down to – I feel like everything is different, the whole foundation. So for instance, Amazon is very much like Airflow-based. So you basically feed in everything into Airflow and it will generate the Airflow models, or jobs will generate the model. For DataHub, it's like you feed everything through Kafka queues. For us, like we just have direct connectors like either like Python library of Snowflake or ODBC driver that goes directly into the data sources. And then underneath, I guess it's also a slightly different like. I can't speak really deeply about like the other architectures, because it's also been a while since I looked into it. But the way that we see it is like we have what we call unified metadata model, where all databases will be treated with a certain structure of database, schema, table, columns and regardless of what it's called. And then we have separate meta data model for BI tools. So it could be dashboard, and charts, and queries. And dashboards are always a group of quarries, a group of charsts, and then charts can be group of quarries, like things like that. So it's more like logically just different. Whereas, I think my DataHub and Amundsen –

[00:55:51] JM: Stemma? Amundsen?

[00:55:52] SK: They all have different ways of treating these metadata. And I think Amundsen uses a lot of the ORM – I forgot the word the ORM that they use for collecting all the other stuff. And they don't do query parsing. That's another big part. For us, like on top of the metadata model, we have query parser. So query parser supports different dialects of the SQL queries, or Look ML, or your DBT model, things like that. And then that emits the popularity model and the data lineage model. Whereas for DataHub and Amundsen, there's no like built-in the query parser. So hence, what customer needs to do or the user needs to do is they have to build it or compute it somehow and then feed it in as part of the metadata through Airflow or Kafka. Yeah. So that's another difference. For us, it's really like all you have to do is you connect your warehouse and then wait 24 hours, and you get everything from Select Star. So we're just trying to make it really easy.

[00:57:07] JM: Interesting. I mean, at many companies, I guess what you're saying – I've never actually seen a Snowflake installation. But I guess it's pretty easy for you to just hook in – So let's say a company –

[00:57:23] SK: We even give our customers here are the five lines of SQL query that you run inside Snowflake.

[00:57:28] JM: So let's take a company like – I don't know if you work with Instacart. You don't have to mention whether you do or not. Let's say a company like Instacart. It's like a pretty big company. But I could imagine that Instacart has some simple agreement with Snowflake. And all the data is in Snowflake in a way that you could hook into very easily. Is that realistic? Or do you think there's like separate contracts for different Snowflake instances across Instacart?

[00:57:50] SK: Oh, yeah. No, no, no. It's just one –

[00:57:54] JM: One big contract.

[00:57:54] SK: Yeah, just our contract with Instacart, it's an Instacarts data, regardless of where it is, right? So we abide the rules of Instacart.

[00:58:03] JM: No, no, no. But all I'm saying is maybe I'm too obsessed with this integration problem. But I'm just assuming that if you go to Instacart and you're trying to integrate with Instacart, either there's one big Snowflake instance that has all the tables, or there're multiple Snowflake instances. Different data engineering teams have set up different Snowflake –

[00:58:21] SK: It could be both.

[00:58:22] JM: It could be both. It could be both.

[00:58:23] SK: Yeah. And we can integrate with the multiple instances too.

[00:58:27] JM: Got it. You can integrate multiple instances and basically have an index that that searches across both those instances. Can you do a join across those two instances?

[00:58:35] SK: No, we don't run queries.

[00:58:36] JM: You don't run queries.

[00:58:37] SK: Yeah. We just process queries.

[00:58:38] JM: Got it. Process queries.

[00:58:40] SK: Yeah. Well, we'll read the queries, but we will never execute queries again.

[00:58:46] JM: Do you become a layer over Snowflake? Or you layer over Snowflake? Like am I interfacing with Select Star and then like I'd never have to deal with Snowflake ever again?

[00:58:56] SK: No, no. We are like Google for your Snowflake.

[00:58:58] JM: Got it.

[00:59:00] SK: So it's like a side reference you use.

[00:59:03] JM: Do you have ads?

[00:59:06] SK: I should put up something right? You just search something. And then you go back to Snowflake, or you go back to your any SQL ID or your BI tool. Because before you query, you would need to know what the table is called, or what the column is called.

[00:59:24] JM: What's your infra? What kind of stuff you got?

[00:59:24] SK: Infra?

[00:59:25] JM: Infrastructure? What are you building with?

[00:59:27] SK: Okay, we're on AWS. Everything runs on EKS. That's how we manage the servers.

[00:59:33] JM: By the way, your company's more of like a design problem and like a developer preferences problem. There's not like any super hard algorithmic problems or anything. Is there? Or am I wrong? It's not like an offensive question. I'm just saying this is a developer experience product.

[00:59:50] SK: Our end users are mostly data analysts. It's not developers. Developers use it, but it's primarily for a data team data analyst as –

[01:00:00] JM: I'm using developers a very loose term. Technical user.

[01:00:03] SK: I see. Yes, technically the users are the main users. But there are also ops users that we're starting to see a lot more that are primarily – Like they don't write SQL, but they use Looker, Tableau, Mode all the time. And they want to learn about like what is activation rate? What's the definition of this? And they find that there are documentation inside Select Star.

[01:00:26] JM: Oh, interesting. So you have documentation and stuff in there. So it's like it's your friendly – It's like your AI assistant for data engineering.

[01:00:35] SK: Let's call it knowledge base.

[01:00:37] JM: Knowledge base? Wow, interesting. That's pretty cool. So there, I'm seeing the adjacencies. Or at least one adjacency. You're a knowledge base. You're a data engineering knowledge base. That's pretty cool. Can I like tag people and like –

[01:00:53] SK: Yeah, you can. Collaboration.

[01:00:55] JM: It's a social network for data engineers.

[01:00:58] SK: You can even definitely call it that. So they're any dashboard table or metric, there is a tab called discussion, where people can just ask questions, and they can mention

other people. They can mention tables or other data objects in there. And the owner of the table – Usually, there's an ownership we help customers to assign. They get notified. They get a Slack notification or email notification. When they respond, then they also – The person who's initially posted a message get notification.

The good thing about that is all of those discussions are searchable inside Select Star as well. So people don't ask repeated same question or similar types of questions, because whenever they are searching, all of those also come out in the search results.

[01:01:50] JM: So what's the hardest engineering problem you had to solve?

[01:01:53] SK: Oh, hardest engineering problem did you say?

[01:01:55] JM: Yeah. Yeah. Yeah.

[01:01:57] SK: I think query parsing in the beginning definitely took some time to get it right. Because –

[01:02:03] JM: I don’t understand. Why do you have to do query parsing? Doesn't Snowflake take care of it?

[01:02:07] SK: Now they do. But this is a very new feature. They didn't have that before.

[01:02:11] JM: When you say query parsing, what do you even mean?

[01:02:13] SK: We look at a SQL query and then output which tables, which columns, this query has touched.

[01:02:22] JM: Oh, parsing is not the right word for that.

[01:02:25] SK: Then what is it?

[01:02:26] JM: It's like a data touching.

[01:02:29] SK: Data touching?

[01:02:30] SK: We are parsing the text. It's just the text of select, whatever, blah-blah-blah, and then we pull out those. And then –

[01:02:41] JM: So it's which tables did this touch? Or even which rows did this touch? Okay, that is pretty complicated.

[01:02:47] SK: And think of applying that in many different SQL dialects.

[01:02:52] JM: By the way, how is that not knowing what data access? If you're knowing the rows that I touch and the tables that I touch, isn't that knowing what data I access?

[01:03:01] SK: Yeah.

[01:03:03] JM: But I guess it's not knowing the data itself?

[01:03:04] SK: Yeah. I don't know what the value because I never see the current value.

[01:03:08] JM: You're like Google Maps. You're Google Maps for my data. I'm just trying to put your company in its best light.

[01:03:14] SK: Great. Thanks.

[01:03:15] JM: I'm trying to sell you marketing stuff. We have ads on my show. Do you want to buy some ads?

[01:03:19] SK: Yeah. I think we need some ads, for sure.

[01:03:21] JM: Really?

[01:03:21] SK: Yeah. You want to call it Google Maps or data?

[01:03:25] JM: Google Maps and social network, and you're building the stories feature, right?

[01:03:30] SK: Google and Facebook for your data.

[01:03:31] JM: Yeah, I wouldn't say Facebook. You don't want to say that word.

[01:03:34] SK: Okay. Okay. Fine.

[01:03:36] JM: Well, you worked at Facebook for a little bit.

[01:03:37] SK: Yeah.

[01:03:37] JM: What do you think of that company?

[01:03:38] SK: It's amazing.

[01:03:40] JM: Amazingly scary?

[01:03:43] SK: Everything. Let's not go into too much details here.

[01:03:45] JM: We don't have to go into too much detail.

[01:03:47] SK: Yeah. But I mean, Facebook was amazing when I was working there, 2009.

[01:03:49] JM: When you were working there. It was amazing in 2009. Back when it wasn't surveilling everything you do and affecting your actions and stuff. My words, not yours.

[01:04:00] SK: Sure. Yeah. I mean, I don't know. Like when you say surveilling, you're talking about employees or you're talking –

[01:04:07] JM: Everybody.

[01:04:07] SK: Everybody. Oh, okay. Yeah.

[01:04:09] JM: Except Mark Zuckerberg. I mean, is it a little unfair that he knows everything that we do and we don't know anything about what he does? Like what goes on in building eight?

[01:04:18] SK: But it's not just Facebook that knows everything about you or what you do. It's every other company.

[01:04:23] JM: So what do we do about that? Shouldn't that be like the number one political issue, right?

[01:04:27] SK: I think it has been before?

[01:04:30] JM: Not really.

[01:04:32] SK: They were questioning Mark Zuckerberg, right?

[01:04:34] JM: Oh, do you remember the questions?

[01:04:36] SK: It was something around that, right?

[01:04:38] JM: Something around that. Do you remember any of the questions that those old people asked?

[01:04:43] SK: It was a bit stupid, but whatever.

[01:04:45] JM: More than a bit?

[01:04:46] SK: Yeah. It was like it was like, "Are spying me?"

[01:04:48] JM: No, no, no. That would be at least legitimately sophisticated. It was like, "How much do you charge for your subscription business?"

[01:04:55] SK: Really?

[01:04:56] JM: Yes. Really?

[01:04:57] SK: Okay. Well, anyway.

[01:04:58] JM: No. No. Seriously. Like how is this Not a gigantic political issue like the fact that these companies just have too much power? I mean, I don't like to say that because I love these companies and I love the utility they provide. But we have to admit to ourselves these are just too powerful. They're way too powerful. They're more powerful than the geopolitical entities in our world. They're simply more powerful, right? It's like what do you do when a corporation becomes bigger than a nation state? It's kind of an issue.

[01:05:27] SK: When you say bigger?

[01:05:29] JM: More powerful.

[01:05:33] SK: What do you mean by powerful?

[01:05:35] JM: I mean, AWS can shut down anything in the world, except things that are in China perhaps, right? That's kind of weird.

[01:05:46] SK: So you think we need to have surveillance on those companies that has too much power.

[01:05:51] JM: I don't know. I don't know. I don't have a solution to this yet. But it seems like a pretty big issue, right?

[01:06:02] SK: Yeah, when you put it that way.

[01:06:05] JM: What do you do when there's only one power grid and it's entirely proprietary. And it's run by a guy who seems pretty nice, if you're talking about Jeff Bezos or Andy Jassy. They both seem like pretty nice guys, thankfully. But it's kind of weird, right?

[01:06:20] SK: What is in their interest to abuse that power?

[01:06:25] JM: it doesn't have to be in their interest. It just has to be like in – Once these companies get so big, there's like VP of shutting things down, right? There could be a VP of shutting things down. And as long as the VP of shutting things down thinks that something should be shut down, it get shut down. And Andy Jassy doesn't even really have oversight or opinion on that, because he's too busy in meetings all day, right? He's like determining should we go with like AWS SageMaker for kids, or like AWS S3 for geopolitical scenarios or something? He's too busy. He doesn't have time to even have an opinion on whether or not you should shut down Parler. So parler just gets shut down. And he's like, “Whatever. That's fine. It doesn't matter. Who cares?” Which is kind of fine. Thank goodness, it's just Parler, but it's kind of creepy, right? Really, the bookseller gets to shut down a social network? That's kind of weird. Isn't that weird? At a fundamental level, isn't that weird?

[01:07:26] SK: Yeah, I mean, when you put it that way. But I don't know. I just haven't really thought about it that way.

[01:07:31] JM: I don't know either. But it's strange. And it doesn't seem like anybody's talking about it. It seems like a big deal. It seems like a fairly big deal.

[01:07:41] SK: I feel like, even traditionally, like carriers, AT&T or Verizon, they could have just like shut down their network. And everybody would go dark, or any electric company.

[01:07:52] JM: And that’s a problem as well.

[01:07:54] SK: But we’ve been living like that the whole time.

[01:07:56] JM: Right, thankfully, thankfully. But it's not in the public consciousness that this is an issue, right? You want to be aware of these things before they happen. And you want checks and balances and like battle plans.

[01:08:09] SK: So are you saying that, compared to the traditional companies, the new tech companies don't have enough checks and balances in place to control those –

[01:08:19] JM: I think we need more open discourse around how to put checks and balances in place so that these companies cannot do things like that. Or if they do, we can know how to route around that, right?

[01:08:38] SK: I think there is a power that government can exercise to –

[01:08:45] JM: Certainly not this government. I mean, a government theoretically, but probably not this government.

[01:08:51] SK: I mean, it depends on what we're talking about. Obviously, like if somebody likes to be shut down, actually shut down everything, then that's a whole different story. But I'm talking about like they have – The government can ask for certain things for any corporation.

[01:09:10] JM: Certain things. Yeah, sure, but the government doesn't even know what to ask for in this scenario. Like what are they going to say, “Hey, you can't shut down an EC2.” Or, “Hey, you can't get rate limit this company.” Or, “Hey, you can’t –”

[01:09:20] SK: Yeah. I mean, those things are just like business to business contract. That's not what government should concern about.

[01:09:26] JM: Are you sure about that?

[01:09:28] SK: Yeah. They will concern about it if the businesses like sue.

[01:09:33] JM: Okay. What if the government has a Palantir contract and AWS decides to shut down Palantir?

[01:09:39] SK: I don't think those government servers run on AWS public cloud.

[01:09:46] JM: But let's say they do.

[01:09:48] SK: I mean, maybe that's the whole reason why.

[01:09:51] JM: Huh?

[01:09:51] SK: Maybe that's the whole reason why governments don't run their –

[01:09:55] JM: Gov cloud.

[01:09:57] SK: Yeah, but that's not – I mean, that's just the public gov cloud? Governments have their own –

[01:10:03] JM: No, there's gov cloud. There's the United States gov cloud run by Amazon or Azure. I'm not sure which one. Maybe a combination of the two. Maybe it's Oracle. But there is a gov cloud run by a corporation, right? And let's say there's some kind of on-prem – Is too political for you? We can talk about something else. We can talk about quest bars, or data infrastructure, downtown. You don't want to take pitches any more. Different pitch. Fundraising, entrepreneurship.

[01:10:28] SK: I just don’t know where this is going.

[01:10:29] JM: It’s not going anywhere. It's curious. There's nobody talking about this as far as I can tell.

[01:10:36] SK: Because nothing is happening.

[01:10:39] JM: But it could.

[01:10:40] SK: But it could have happened even before, and it has never happened.

[01:10:47] JM: Well, but it did with Parler, or with the Daily Stormer. These are not websites that I'm a fan of, by the way. I have never even been there. I'm agnostic on them. But they got shut down. You know?

[01:11:02] SK: Okay.

[01:11:04] JM: I mean, what have you got shut down, right? What if Select Start –

[01:11:08] SK: Yeah, that would suck.

[01:11:09] JM: Well, I think that whole streaming versus data warehouse thing is interesting. Why Snowflake win I think is interesting. Why did Snowflake win? And don't say separation of storage and compute?

[01:11:18] SK: I don't know why Snowflake won.

[01:11:20] JM: I think I know.

[01:11:22] SK: Why?

[01:11:22] JM: They were ex-Oracle.

[01:11:25] SK: What part is their ex-Oracle? The founder founders or –

[01:11:26] JM: The founders, I believe, right? Ex-Oracle.

[01:11:29] SK: I don't know. I never –

[01:11:30] JM: I think they're xe-Oracle.

[01:11:31] SK: Okay. So what?

[01:11:32] JM: So Snowflake is the next Oracle.

[01:11:34] SK: Okay.

[01:11:36] JM: That's my thesis. Agree? Disagree. Can't say? Deep partnership can't say anything.

[01:11:45] SK: No. I mean, like, okay, so what if it's ex-Oracle?

[01:11:49] JM: What if Snowflake shut you down?

[01:11:51] SK: Oh, that would be sad.

[01:11:53] JM: That would be sad, right? So you can't actually say anything bad about Snowflake, because Snowflake could shut you down.

[01:12:00] SK: What do you mean? I mean, it's not correlated to one another.

[01:12:07] JM: Well, I just think it's interesting, right? The way that large entities combat each other these days is software-based. It's low-level, right? Anyway, we could talk about something more or less touchy feely.

[01:12:25] SK: I don't still get it.

[01:12:26] JM: Seriously – Okay, let's just talk more brass tacks. Why did Snowflake beat BigQuery, Redshift –

[01:12:34] SK: I want to say they just fully beat them.

[01:12:37] JM: I mean, they're by far the most –

[01:12:39] SK: They are dominating, yes.

[01:12:40] JM: They're by far the most dominant. How about this? Snowflake versus the Databricks ecosystem?

[01:12:46] SK: Okay.

[01:12:48] JM: Not a question? Not interesting? That's not interesting.

[01:12:49] SK: No. I mean, this is all very interesting. I feel like this is something that Snowflake is dominant right now. But there are a lot of other things that's also happening in the data ecosystem that it's going to be a long game.

[01:13:07] JM: Right.

[01:13:09] SK: What I mean here is there are customers moving to BigQuery from Snowflake.

[01:13:14] JM: Come on.

[01:13:15] SK: No kidding. I know a customer.

[01:13:17] JM: Why?

[01:13:20] SK: I think cost. This is the advantage that cloud providers have. If I'm already committing seven, eight figures to a cloud provider, then why not just – Then cloud providers will give me like a bunch of credits to use BigQuery, or Redshift, or whatever. So cost. And then also integrations. I mean, Google bought Looker, right.

[01:13:53] JM: That's a good one. That's true. Snowflake doesn't really have an integrated thing like Looker, right?

[01:13:58] SK: Well, they bought Numeracy, which is an IDE.

[01:14:00] JM: What?

[01:14:02] SK: Numeracy. It's a SQL client.

[01:14:04] JM: Like a data IDE. Like PopSQL?

[01:14:06] SK: Mm-hmm. Yeah. And you can also run like visualization, charts, things like that.

[01:14:12] JM: That's pretty cool.

[01:14:14] SK: Yeah, yeah.

[01:14:16] JM: But Snowflake can't ever really be a platform, because it's proprietary.

[01:14:20] SK: What do you mean? That doesn't make sense? It can't be a platform because it's proprietary?

[01:14:25] JM: Yeah. I mean, who's going to make their entire bet on a truly, truly, truly closed source system like Snowflake? It's sort of like are you going to build your entire business on Airtable? Like it's a little bit too Microsoft Windowsy, right?

[01:14:40] SK: Got it. I hear you.

[01:14:41] JM: As much as I love Airtable. So internally, in Software Engineering Daily, we're a spreadsheets company, right? That's all we are. We're just a bunch of spreadsheets of like ads and like shows that are scheduled and stuff. I would love to move entirely to Airtable, but I'm a little bit nervous about it, because ir table looks like Microsoft Windows to me.

[01:14:58] SK: Whereas Google Sheets is not? And Microsoft Excel is not?

[01:15:02] JM: Google doesn't have any reason to abuse their power in this category, right?

[01:15:07] SK: Why?

[01:15:09] JM: Because they're making money hand over fist in search engs.

[01:15:12] SK: They just have so many other businesses? Is that the reason why?

[01:15:15] JM: Yeah. Yeah, exactly.

[01:15:17] SK: Okay. Yeah, I think there's like pros and cons for – So we were talking about Snowflake and BigQuery, right. What I was going to say also earlier is that there are a lot of other data warehouse players also coming out, like Firebolt is a new one.

[01:15:33] JM: Firebolt is great.

[01:15:34] SK: Yeah. Yeah, Databricks has their new –

[01:15:37] JM: I trust Databricks. I trust Databricks. If I had to go with one data warehouse, I'd probably go with Databricks at this point.

[01:15:42] SK: Why? Why is that?

[01:15:44] JM: I like the CEO. I like the founders. I like the ethos of the company. I like the fact that Delta is open source.

[01:15:52] SK: Delta is open source?

[01:15:53] JM: Yeah.

[01:15:54] SK: I didn't know.

[01:15:54] JM: Yeah. I mean, what's the best open source data warehouse? Databricks.

[01:16:00] SK: Okay.

[01:16:01] JM: What else is there Clickhouse? Not really.

[01:16:04] SK: I don't see Clickhouse as a warehouse, but sure.

[01:16:07] JM: What is it though? A house?

[01:16:09] SK: The last I would remember would be like just HDFS. But that's a file system. And Hive.

[01:16:16] JM: What? HDFS is not a data warehouse. It's is a file system.

[01:16:18] SK: Yeah. I just said that's a file system. But that's basically how you would store all the data.

[01:16:25] JM: That's a data lake. Do you know the term data lake?

[01:16:28] SK: I mean, I think, to me, can be either.

[01:16:30] JM: It's the same thing, data lake house. Not falling for that one?

[01:16:37] SK: Yeah, no. So yeah, I think that there's like still a lot more to go. It's not just Snowflake game.

[01:16:43] JM: Alright. Here's another question. If you weren't building Select Star today, do you have another business idea in the data engineering space or another business idea? Actually, two business ideas. I'm sure you have two. One in the data engineering space. One outside of it

[01:16:57] SK: Oh, that's a hard question.

[01:16:59] JM: Is it?

[01:17:00] SK: Yeah, because I had to go with – Not Snowflake. Select Star. I had other ideas before, but Select Star makes the business. So you want to hear about other ideas?

[01:17:15] JM: Yeah.

[01:17:17] SK: Okay.

[01:17:17] JM: I need something to pivot to. Podcasts are dying. Do you have any podcast adjacent ideas? Please?

[01:17:27] SK: Yeah. We should do data engineering thing. But there is actually a data engineering podcast.

[01:17:32] JM: Yeah. Tobias.

[01:17:33] SK: Yeah.

[01:17:33] JM: Have you gone in his show?

[01:17:34] SK: I did.

[01:17:35] JM: Nice. God! He beat me to this? That's my fault.

[01:17:38] SK: It's your fault. I'm just kidding. Actually, Select Star came off from my original idea, which was part of my longer term roadmap from Concord, which is determining, basically writing your ETL jobs automatically based on the types of data that you are collecting. So when you're looking at an event, you know that you're collecting either a number, or some string, or timestamp. And based on that, we will just calculate. If it's timestamp, we'll just give you the aggregation by day, by month, or week, or month. And then run also aggregation for numbers, min, max, sum, average, things like that. So that you don't have to write full freaking ETL job yourself.

[01:18:30] JM: That's cool. Alright, I'll do it. I'll do it.

[01:18:32] SK: Or more like at least determining like where and how it's coming from, and kind of try to do it that way, instead of having to like wrangle everything manually. But I realized a lot of the issues around streaming was starting to get solved. And it wasn't as applicable in so many use cases. So I decided to focus on data discovery.

[01:18:55] JM: So we have to wrap up. But wait, can you give me – Jesus. Can you give me one other business idea in one minute?

[01:19:01] SK: Non-tech idea?

[01:19:03] JM: The non-data engineering business idea in one minute. And then I have to say goodbye to you in like five minutes.

[01:19:07] SK: Oh, that's so sad.

[01:19:08] JM: I know.

[01:19:09] SK: Okay. The other idea that I also was very passionate about in the past is I can't say that this is a good business. It just wasn't the right business for me. But I think there is a lot of problem around patient advocacy. This is like personally something I went through when I was an advocate for my mom going through health issues. And I realized that for people that like may not have this like knowledge, and most of the time, patients are already in so much pain and a lot of things that they go through. It's hard for them to clear, like truly advocate themselves. So they need to rely on others. And patient advocates that you can hire are either very expensive, or it's just very hard to find the right person. So I wanted to build something like Uber for patient advocates, where it's a service that can really take care of like your needs. But it requires a lot of operational work to make it work.

[01:20:05] JM: That is a great idea actually.

[01:20:07] SK: So that is – Yeah. Flavor that has a lot of.

[01:20:11] JM: Alright, until next time. Until next time. So you're going to fund that that, right? I'm doing that. You're going to fund me for $5,000. And we're done. All right. Great. Shinji Kim, Thank you. $5,000. I'll take the check.

[01:20:21] SK: Thanks. Yeah. Great. We'll find you.

[END]