CapTech

Keep Your Data Lake from Becoming a Data Swamp

CAPTECH TRENDS PODCAST | EPISODE 25

Vinnie

Hello, and welcome back to the CapTech Trends podcast. Back by popular demand we have Calli. One of the best parts of my job is when I get to work with people who are really passionate about what they do both for a living with inside projects and people who are very intelligent and who have knowledge and domain expertise that I don't have. That I can learn from and Calli is a great example of all three of those things. As I said, we had a lot of requests for her to come back on the show. So welcome, Calli.

Calli

Thanks. Great to be back, Vinnie.

Vinnie

Yeah, so we've had a couple podcasts recently speaking about data, advanced analytics, machine learning, AI, getting on the more of the consumption side of data, and we've said several times over and over again, how important good data is. We talked about the five V's so velocity, volume, veracity, variety, value, there's other V's. Right, but those are some of the V's. So, what I wanted to talk to you about today is getting a little bit closer to pragmatic approaches to improving the quality of your data so that you can do the advanced analytics, machine learning and AI and such. So, starting there, I know you've been on some projects recently, with some technologies that you're passionate about, why don't you start with the metadata driven ingestion?

Calli

Yeah, so something that I love about metadata driven ingestion is that it's very easy for users to get their data in. So basically, what we do with metadata driven ingestion is instead of a user saying, hey, I have this new file, I need it into the data lake or the data warehouse, or what have you, they're able to securely and with governance approval, and everything like that, be able to push that data up into the lake and be able to use it for their analytics. So, it's a really fantastic way to be able to get that volume,

or the velocity of the data and variety.

Vinnie

Yeah, so let's talk about that a bit. I like to get specific on so people can know like, actually what it is not just generally conceptually what it is. So, what we're talking about is power users, I guess business users?

Calli

Usually your power users, your analytics groups that really already have their shadow IT, whether you realize it or not, that are doing things that they have to do to be able to get those reports out. A lot of times, these are reports that are going to C- suites and stuff like that and they're just pulling data from the internet and pushing it from their local machine and using that to do these reports to push them out to you.

Vinnie

Right. So, I'm thinking of people who would normally have spreadsheets, some unstructured data, maybe some structured data that they're using in some operational systems and having a trusted method by which that can be ingested into a data lake. Do I understand that correctly?

Calli

Yeah, absolutely. I think the first time I built this, that I worked with a client on something like this, we had Zillow data, for instance, that we were pulling in we also had this was a large utility provider, we had smart meter data. That was JSON based. So, it didn't fit in the data warehouse, but we were able to put it into a data lake and as it changed, they were able to keep updating it themselves. They didn't have to get IT involved, but they could still do it in a self-governing kind of way.

Vinnie

So, what were some of the pains of the past then, people were doing this anyway and they didn't have this metadata driven approach. So, what did that lead to? Was it poor data, a data swamp as opposed to a data lake? Did it then require post processing to repair it? Like what were some of the why did this come about?

Calli

Yeah, absolutely. The data swamp is a big one. Another thing that really happened was being able to audit a lot of this stuff, they couldn't do it. So, for reports that are just general guidelines and stuff like that, that's great. I think of our financial clients, when you're making financial decisions, and approving or rejecting people based off data that's sitting on my machine that at any time could crash and burn, and then I can't get it back. We need to be able to audit those types of things.

Vinnie

Right. So, it's audit, or rules based on the upfront. Okay. That kind of takes the shadow out of shadow IT?

Calli

It does.

Vinnie

Instead of just asking people to stop doing what they're doing, you're allowing them to do it in a more. What's the word, governed? I hate to use that word, but more responsible.

Calli

Yeah, absolutely.

Vinnie

Yeah. What does that look like? So, for me as a more of a programmer background, less of a data background, I'm thinking, well, isn't this just a list of metadata tags and a rules-based engine?

Calli

You're not that far off, a lot of times I see it, where they'll take a tool like Collibra, which is a governance tool that has its own back end with all the metadata built in and marry that with an ETL tool. Depending on what you've got, there's a ton of them, a lot of them are homegrown as well, but be able to automatically read that metadata that's coming in and apply your rules. So, check to make sure of the basics, an integer is an integer, a string is a string, there's no funky non-ASCII characters in there, etc. If there's not supposed to be that sort of thing.

Vinnie

It reminds me back in the day of EDI Electronic Data Interchange when all these companies would send datasets back and forth and there were always rules-based engines to check for format, not just whether it's an integer or whatever but also range value range checking. That then turned into more abstracted rules-based engines. So, I guess my question there is not just showing how old I am. Is this a visual interface? Are people going in and looking at metadata tags and applying edit checks in a visual way or is this all scripted?

Calli

I've seen it both ways. The visual UI is obviously much better for your power users that aren't scripters or maybe you want to control the scripting and that sort of thing. We actually have an open-source tool that works with a dupe here at CapTech called Alfred.

Vinnie

Didn’t you create Alfred?

Calli

I did create Alfred. Self-plug. (laughs)

Vinne

Shameless plug. Yeah (laughs)

Calli

Works great with HDFS. I'm actually working on it this year to get it to work with AWS and be able to work in a data lake that way, but the UI to it allows the end users to put in exactly what they need, and the reasonableness checks and everything like that.

Vinnie

Is there a website people can go to, to pull that down?

Calli

It's on GitHub. It's on the CapTech GitHub.

Vinnie

Okay, under Alfred? Want to tell us why it's called Alfred?

Calli

I am a huge Batman nerd. Ben Harden actually came up with oh, it's like a data butler. That was the tagline and so for those of you who aren't Batman nerds, Alfred is Batman's butler.

Vinnie

Gotcha. Makes sense. Does this introduce other problems? So, what I'm thinking of is gosh, great, now

we are getting power users, and maybe some business users who aren't quite as powerful, or as knowledgeable and this sort of the freedom to import as much data as they want. So, do we end up getting more types of it? Does the variety and the volume start to exceed a good intent?

Calli

It does, you could, you could very easily end up with duplicates, you get somebody that's putting the same stuff in there. I think that's really where it comes down to at the end of the day data governance is people over process, you've really got to make sure that the people are keeping track of these things and looking into it and you can absolutely build in a ton of checks into these sorts of things where tasks can go back and forth between people to make sure that what's there works.

Vinnie

That sounds like workflow.

Calli

It does sound like workflow.

Vinnie

So, is that part of these tools or is that still largely a human process?

Calli

I usually see it largely as a human process but that doesn't mean we can't build it in, we can put workflow processes in to work with that.

Vinnie

Where does the metadata come from? Do I have to go in there and create a canonical model of my entire organization and then apply it to these data sources? Is it derived from the data sources

themselves? Is it stored separately, so there could be synchronization problems between what we have we think as metadata and what's really in the data sources?

Calli

The more sophisticated tools I've seen, is doing an extraction on the data themselves. So, like, you'd upload a CSV or a JSON file, something like that and it would be able to interpret that based off of that, and then as you upload the data, it's checking against that metadata. So, if the data changes, it can keep up with it and be able to report back and kick back an error and those sorts of things

Vinnie

Gotcha. So, there's a feedback loop. Where my head goes right away is, is it inferring what the metadata is based on the values in the fields? So, you better be checking a lot of rows, because you could have data that appears to be one type of data, right?

Calli

Yeah, absolutely. I mean, if you're uploading a sample file, for example, you'd want to make sure that that sample file has everything that's available, which is going to be difficult sometimes and there are times that there are errors in those sorts of things.

Vinnie

Gotcha. Makes sense. Are there tools for this? Are there vendors for this? I mean, Alfred, obviously, is something that you created, because I imagine you saw a gap in the market and needed a tool for it. Are there vendors that lead in the space or is it still largely custom?

Calli

It's largely custom. The other one I've seen is Kylo, which I believe is open sourced by IBM, I should double check that, but I'm almost positive. It's IBM, which also works on an HDFS cluster, but it's very

similar, has a beautiful front end has the data governance built into it and has ETL built into it for CSV, JSON, various other separated value type things, anything like that.

Vinnie

Gotcha. From a business standpoint, why do I care? I'm going to get the technical stuff, I'm excited that people's job easier from a business perspective, is it just getting I would imagine, and correct me, it's creating that foundation so that I can do the advanced analytics, machine learning and AI much more quickly and much more reliably? It's kind of quality of data.

Calli

Absolutely Yeah. So, from a business perspective, you think you really don't want bad quality data, you don't want data that you don't know where it came from. That's a huge thing that I see in a lot of analytics groups that they're using data that they don't know where it came, maybe one person knows where it came from, but generally, it's not really used well, so it comes back to that auditing and everything like that.

Vinnie

Right. Yeah. It reminds me of having owned a pool in the past. It's far harder to clean the pool. Once it's a mess, right versus just controlling everything up front. So, I get that. Do you see more people getting into this advanced analytics space, then realizing this as a problem and addressing it or do you see more people getting in the space knowing it's going to be important and doing it ahead of time?

Calli

The former, I definitely see a lot of people not realizing what advanced analytics entails, not realizing what the users are going to need. The scariest part of advanced analytics is that your users aren't governed by your same software processes, but they are building software. So, a lot of times your advanced analytics, your data scientists, those folks are on the business side of the house but are building software. So, they don't have a lot of the same governance processes and that sort of thing. So

how do you keep them from making the mistakes that we have tools in place on the engineering side of things without slowing them down?

Vinnie

It reminds me like of the maturity of application development, we got to the point where DevOps became obvious and mature, you know, CI, CD, automated builds, automated testing, all that kind of stuff and data seemed to trail that bit in terms of the maturity side. And maybe that's true because it was less productionized and that's that also was changing, right? So, is there is there a fun word like DevOps that applies to this style? Is it an extension of ML Ops?

Calli

It is an extension of ML Ops, we use DataOps, just to kind of envelop ML Ops and the DevOps side of things, and then apply to data.

Vinnie

Gotcha. So, let's switch gears a little bit, because there's other tools involved like Python, and R, and I hear people talking about using those to participate in these types of processes, too. Are those the two right tools to talk about? Are there others? And if they are, where do you see them being used appropriately and where do you see them kind of being used in a stretch where maybe they shouldn't be?

Calli

So, Python and R are the two most common languages being used on data in the advanced analytics space. I see Python, being more used in the enterprise space than I do R. I see R used more in academia. That's not to say that R isn't used in the enterprise space, I will refrain from venting too much, but R is definitely not a language that is a first-class language, as far as the cloud and a lot of support, it definitely comes as an afterthought. It's very difficult to productionize. A lot of the tools are written from an academic standpoint, which is a lot of times not sufficient for the enterprise.

Vinnie

Okay, so then why aren't there just more Python libraries to simulate or emulate those functions?

Calli

There are. A lot of times R is just faster and better at it.

Vinnie

So, it comes down to how it was built?

Calli

Right, exactly.

Vinnie

Right. help me understand, again, a very specific level, I know why I would want to use Python. I've dabbled in it. What's the problem I'm trying to solve that makes me say, oh, okay, gosh, I need to go use R now?

Calli

A lot of advanced statistical analysis.

Vinnie

Things I wouldn't be able to do anyway. (laughs)

Calli

(laughs) Exactly. A lot of the stuff that I mean, I can't do it either. That's what we have data scientists for.

That's why they get paid the big bucks.

Vinnie

(laughs) Gotcha. So, I don't have to worry about that in my professional career. Gotcha. Yeah. So, the last thing I can think of that I wanted to ask you is how do people get started? Is it a reaction to realizing that their lake is becoming a swamp, or they do it proactively in either case? What are the first couple things you did? What do you tell a team in terms of people process technology? What cloud vendor do you go to? What do you download first? What problems do you undertake first? Do you try to go broad and shallow or narrow and deep? Give me some rubber meets the road. First 30 days kind of thing.

Calli

First thing you do is figure out what you already have. I think that's the biggest thing that a lot of companies, maybe they know what's in their data warehouse, maybe they know what's in their operational data stores but there are pockets of data everywhere. There are data silos that they probably aren't even aware of.

Vinnie

So, this will be a never-ending task?

Calli

It's probably a never-ending task, finding the things that are important. You can't qualify everything. Like you said it's a never-ending task.

Vinnie

New data comes in all the time.

Calli

Right, finding the things that are in the important processes. So maybe even taking a step back and identifying what are those important processes? What are your analysts reporting on, what are they using and what gets sent to your CEO on a weekly basis? I work with a large healthcare provider and right now they have a COVID thing that runs monthly up to their CIO, and some of the data they're using is stored on someone's laptop and the first thing we said is, that needs to change. We can't have that just sitting on someone's laptop that could crash at any time.

Vinnie

Right. So, identify the data. I imagine there's a whole sort of information archaeology going around to getting all the different business rules that need to be applied for the edit checks and the ingestion and the range values. I mean, is that more of a discussion with the business or is that a discussion more with the DBAs? Or both?

Calli

I think it's both I think your DBAs are going to give you your technical pieces. They're going to be able to give you, that's definitely a number that shouldn't exceed 18 characters. Your business users are going to be able to give you your reasonableness checks. If you have score values coming out, maybe your score will never be bigger than one, type of thing.

Vinnie

Sort of random question. I've done some API work and modeling similar to what you're talking about where an organization grew by acquisition. So, what they considered a product number was different than what another person that they acquired, considered it. The rules were different. Some were alphanumeric, and some were not, some of the links were different. So, there wasn't a single product number in the organization. There were many, and they hadn't normalized that yet. So, I'm thinking about the metadata implications of that, are there ways to normalize that and create a single model? Are you dealing with those exceptions, kind of the same way, I had to with the API and just read a lot of encapsulation software around it?

Calli

A lot of times, it is a lot of encapsulations. It's a lot of trying to figure that out, you can normalize it, you can create a lookup table that would have your old product IDs and have your new product IDs type of thing. You can also just, at some point, honestly still need the lookup table, I was going to say, eventually just change them over. But you're always going to have that historical data somewhere.

Vinnie

Hopefully, over time, that'll age out, but it's going to be a decade.

Calli

Right, in a data warehouse, you'll probably have product number and old product number sitting side by side in the data warehouse.

Vinnie

Gotcha. So, two more things to talk about tools and people. So first 30 days, what are you downloading? Of course, Alfred, but what else? What else are you downloading, or looking at from a software or platform perspective?

Calli

I'm looking at things like Collibra and other data governance tools. I think that's absolutely where you start, you want to get the governance in place more so than you want to get the ETL in place, because you probably already have the ETL built in some way, shape or form. And then being able to marry that to your existing ETL processes in a way that's abstracted, which is always the most difficult part of this. How do you build an ETL process that can take a file and load it successfully?

Vinnie

Right. So last question on the people side, then, are there additional roles you have to hire for this or is it within the skill set of existing data scientists or data engineers, or even programmers like we're who has the skill set that this is an easy lift for?

Calli

To build this, you're going to need data engineers, and programmers. I think you can use your existing teams, a lot of times, that's sufficient. The big thing is data governance, I could talk for probably a half hour on just data governance, as boring as it is, but having people that are not just your data consumers, but your data producers, who are responsible. Your data stewards that are responsible for that data, the rules and everything that goes with it. A lot of times, that is somebody's job on top of their regular job but when you're really getting serious about keeping your data lake from becoming a data swamp, or getting this new data into a data warehouse, or what have you, it really needs to be someone's full time job.

Vinnie

Kind of an aside question and I'm going to get back to that last thing on the people on the full-time job. We talked about Python being a very common language here, if someone's a heavy JAVA shop, or heavy .NET shop, are there equally good tools in those realms, or are we saying, no, in addition to knowing either .NET or JAVA, it's going to be good to know Python for these purposes and learn it.

Calli

In your analytic space, Python is the language. .NET and JAVA just doesn't have the same analytical tooling and that sort of things built in. They're just not designed for that.

Vinnie

Gotcha. So, the question that people that I wanted to get to, is there, I'm thinking that yes, we talked about what it takes to understand and build this process. Is there a person on the business liaison side that is making sure the organization knows what data they have, the value of that data, analytics, what

sort of a person is that? A single point that you can go to and say, is our data clean? Do we have enough of it? What are we getting out of it? Is that person involved in this process?

Calli

Yeah, absolutely. That's going to be somebody in your data governance side of things. Hopefully, you have a data governance shop, when you've got probably one person that's usually a data steward. One of the more clever names I've seen for them, somebody called them a data angel.

Vinnie

Well that leaves space for data demons then too. (laughs)

Calli

(laughs) Those are your data consumers.

Vinnie

Well, good. That answers my questions, a great topic. Again, I wanted to get closer to the rubber hitting the road because we can talk about advanced analytics all day long. If we if we don't talk about data engineering, then it just becomes kind of a futurist discussion and not really a pragmatic discussion. So great. Thanks again for joining. I'm sure we'll have you back on soon and look forward to that.

Calli

Sounds great. Thanks, Vinnie.

podcast may be made without CapTech' s prior written permission. CapTech makes no warranty, guarantee, or representation as to the accuracy or sufficiency of the information featured in this podcast. The information opinions and recommendations presented in it are for general information only. And any reliance on the information provided in it is done at your own risk. CapTech. makes no warranty that this podcast or the server that makes it available is free of viruses, worms, or other elements or codes that manifest contaminating or destructive properties. CapTech expressly disclaims any and all liability or responsibility for any direct, indirect, incidental, or any other damages arising out of any use of, or reference to, reliance on, or inability to use this podcast or the information presented in it.