EPISODE 375

[INTRODUCTION]

[0:00:00.6] JM: Every software company backs up critical data sources. Backing up databases is a common procedure whether a company is in the cloud or on-prem. Backing up virtual machine instances is less common. Rubrik is a company that is known for building backup infrastructure for enterprises their main product is an appliance that sits on-prem at an enterprise and stores snapshots of virtual machines running within the enterprise. If a virtual machine dies, Rubrik can quickly restore the VM snapshot. The appliance also backs up to the cloud.

Kenny To is a founding engineer at Rubrik and he joins the show to discuss backups and how Rubrik is engineered. Enterprises that start backing up to the cloud through Rubrik start a path towards potentially more cloud services. For enterprises that have not been able to move to the cloud yet, this can be an appealing opportunity. This is a great episode. We discussed product development, distributed systems, engineering product development, how things are going at Rubrik, which is a superfast growing company. Really a lot of different areas we touched.

Software Engineering Daily is looking for sponsors for Q3. If your company has a product, or a service, or if you're hiring, Software Engineering Daily reaches 23,000 engineers listening daily. Send me an email, jeff@softwareengineeringdaily.com. I would love to hear from you.

[SPONSOR MESSAGE]

[0:01:45.8] JM: To build the kind of things developers want to build today, they need better tools, super tools, like a database that will grow as your business grows and is easy to manage. That’s why Amazon Web Services built Amazon Aurora; a relational database engine that's compatible with MySQL or PostgreSQL and provides up to five times the performance of standard MySQL on the same hardware.

Amazon Aurora from AWS can scale up to millions of transactions per minute, automatically grow your storage up to 64 TB if need be and replicate six copies of your data to three different

availability zones. Amazon Aurora tolerates failures and even automatically fixes them and continually backs up your data to Amazon S3, and Amazon RDS fully manages it all so you don't have to.

If you're already using Amazon RDS for MySQL, you can migrate to Amazon Aurora with just a few clicks. What you're getting here is up to five times better performance than MySQL with the security, availability and reliability of the commercial database all at a 10th of the cost, no upfront charges, no commitments and you only pay for what you use.

Check out Aurora.AWS and start imagining what you can build with Amazon Aurora from AWS. That's Aurora.AWS.

[INTERVIEW]

[0:03:25.4] JM: Kenny To is a founding engineer at Rubrik. Kenny, welcome to Software Engineering Daily.

[0:03:29.6] KT: Hi. Thanks for having me.

[0:03:31.0] JM: A core focus of your company; Rubrik, is a backup. What is a backup?

[0:03:37.5] KT: A backup to a business is basically insurance. It's insurance that the data that you rely on to operate your business will be there when you needed it. It won't be corrupt. It won't be lost even in the events of natural disasters or human errors that may have caused deletions or kind of corruptions.

[0:03:55.5] JM: Okay. What are the conditions under which a company would have to use a backup?

[0:04:01.4] KT: The most likely scenario I'd say these days is actually human error. Hardware failures are actually rare enough. Hard disc failures do happen but most systems are built with redundancy now. I'd say the most common would be somebody made a mistake, so they

deleted a file they shouldn't have or they changed something they shouldn't have and they need a point in time copy to restore from.

[0:04:20.5] JM: Rubrik was founded in 2014. What was the standard operating procedure for companies that wanted to have backups in place in 2014?

[0:04:31.3] KT: This was arguably a pretty icky story and that's what most of our customers — Well, most of our future customers still face today. There were a lot of moving parts for one thing. They'd have to buy, let's say — They'd have to look through a storage vendor that get the storage for actually storing the backups. They'd look for some d duplication engine to ensure they utilize the storage efficiently. They'd look for a backup application itself for scheduling the backup jobs, taking the backups and making the copies, managing a catalog. That catalog itself would typically be stored on another database and they also need servers to operate the backup software itself.

There are four or five moving parts here. They'll very frequently come from many different vendors, different vendors for each of them to set this up if you are a customer before Rubrik. This is after actually buying from each of the four different vendors to set this up. Then you'd have, typically, a month or two of actually either hiring your own consultant or working with the post-sales team at each of these different vendors to do the installations, do the configurations, connect all these parts together. That's just to get setup.

Then over the lifetime of actually managing and using these backups, it would typically be pretty complicated. You'd have to do — You as the customer would have to do a lot of work in connecting to the different things that you want to backup, staying up-to-date and setting up new jobs each time something else happened or a new application came in, as well as handling the kinds of exceptions of, "I can't take a backup this week because — Or this particular night because there's a critical application or event going on. I need to reschedule." Just managing all these was also, basically, a full-time job.

With Rubrik coming in — Sorry. Go ahead.

[0:06:15.4] JM: Yeah. We'll get to Rubrik in a second. I'm just curious just a little bit more about these legacy solutions, kind of like the conditions under which Rubrik was created. Did the legacy solutions lack any functionality, or was the problem just really bad user experience, all these coordination with different companies, like bad UIs? Did they actually get the job done?

[0:06:40.5] KT: Yes and no. If you did go through all that work and if you did invest everything you needed to get setup, they would usually get the job done with one exception. What a lot of customers actually didn't do even if they theoretically did everything else right is they didn't test restore. It's kind of well-known — Well, I hope it's a well-known fact in the backup industry that backup is actually easy and restore is hard.

A lot of these solutions, even after you did everything right, a couple of years later if you did have a disaster and you needed to restore, you'd actually find that, yes, the backup — Theoretically, each of the different pieces did their jobs and you have backups, but for one reason or another, restoring from them is actually either impossible or actually very very difficult.

I'd say the main functionality that actually is missing, even if you take away all the complexity and all the management and dealing with multiple vendors, the main thing that actually was still missing from a lot of the legacy solutions is easy restores.

[0:07:40.0] JM: Under those conditions, how did the first version of the Rubrik product come to pass?

[0:07:49.1] KT: Could you clarify what you mean by that actually?

[0:07:51.0] JM: Sorry. I should've just said that more simply. What was the first version of Rubrik when you guys decided you were going to build a better backup software product?

[0:07:59.8] KT: Okay. Got it. We took on these two aspects, these two primary aspects, making a very simple and integrated package together which at the time we called converged data management. Converged as in all the pieces come together. Then as far as the restore functionality, our first and very distinct feature of what we brought to the market was our instance recover feature.

I'm not sure how familiar you are with ecosystems. The first type of application that we targeted backing up is VMware virtual machines, and that's because they're by far the most common — Every enterprise today has some VMware going on somewhere. It's where a lot of their main applications run. It was the most important workload to target, and so the first version of Rubrik was effectively very simple VMware backup and recovery with this unique feature of instant recovery.

Instant recovery is our — And today is still a very important feature for us, is that with a traditional restore, even if everything is going well, the user would have to go into the backup storage and trigger some functionality to copy all the data out. This would be several terabytes of data out from what we call secondary storage back into the primary storage where it's going to be used. This would take hours and hours and those hours are actually basically business downtime.

What Rubrik has done is, as one of our first features, was actually make this an instant option. Our converged storage, since it comes with a little bit of compute and just enough to get you started, can actually restore an entire VM in spite of it being several terabytes, can restore it within a minute. It's not what we consider primary storage. It's not where you'd want to keep it running long-term, but it actually at least — The term is RTOs. It lowers the recovery time objective, which means that the business is at least up and running even if in kind of a — What's the term for it? A sort of degraded mode. The business would be up in a degraded mode, but at least they would be there, whereas with traditional restore, they'd just be completely down for the next few hours.

[0:10:08.6] JM: A lot of the shows that I've done on Software Engineering Daily are with two categories of company. One of the categories is some startup that's either really early and doing something new and weird or they're in the scaling position, and then I've also done a bunch of shows with companies like Amazon, and Google, and Facebook and so on. These companies kind of have the infrastructure of the future. It leaves out this large gap of what we call enterprises, and so there is this whole category of software stack that I haven't really explored, and that is the kind of software stack that companies like Oracle, and VMware, and Nutanix delivery to. I think Rubrik plays in that same space.

You mentioned this prototypical use case where the customer has their infrastructure. The developers are working in VMware instances. Shed a little more light on what their infrastructure looks like. Where are those virtual machines running, and if you want to back up a virtual machine, where are you inserting yourself and taking snapshots so that you can have something to back up to?

[0:11:27.9] KT: Yeah, I'd be happy to describe that. Actually, I'm glad you brought it up because I agree. It's actually a very big gap from my point of view in the understanding of kind of the tech world today. It does seem like here in Silicon Valley we tend to focus on kind of the new age infrastructure that most of our Silicon Valley startups use and which Amazon and Google and Facebook push very well, but there's this whole other world of every other company that's either a tech startup or a big tech company. Basically, exactly what you call enterprises.

They tend to be moving in the direction of the AWS and the Azure Clouds, but they're still — Because they're businesses, they're not focused entirely on software. They do still have a lot of what we might consider legacy. Actually, what's legacy to us is actually their bread and butter of their operations.

What the stack looks like, to answer your question, these days, ever since VMware showed up, I guess a little over a decade ago, VMware — I guess I call it the VMware on Microsoft Stack. Most businesses now will have a majority of their applications virtualized. They'll be running VMware as their hypervisor, so that's what's on the bare metal. VMware has a few components, named vCenter to manage all of these hosts.

Then the VMs that they're actually running are going to be kind of traditional Microsoft VMs. Before VMware that came around, these would have been physical Microsoft hosts. Now they're virtual. These traditional kind of Microsoft — This Microsoft ecosystem, I guess the centerpiece would be active director which stores all of their user accounts, stores their contact information, email addresses. Then right on top of that, they'll have Exchange, Exchange instead of Gmail or anything else.

Exchange would have their actual email content, their calendars, things like that, and that's kind of just to get the people side of the business up and running. Then there's always, always going to be some — Because that's a small amount of data. There's always a much larger footprint which really depends more on what the business is doing.

One of the more common pieces maybe Microsoft SharePoint, which basically is kind of an enterprise's equivalent having a Wiki system and Dropbox all converged into one. That's kind of the most general purpose application I'd say. Really, it definitely depends on kind of what industry they're in. They may have a lot of kind of enterprisey CRM or ERP types of applications or they may have some — If they're in manufacturing or healthcare or something, it will be something very specific to managing the logistics of manufacturing or healthcare.

I imagine that government does a lot that's pretty unique. I'm not too familiar. I know that there are other sectors like legal, which will be focused entirely on kind of the documents that they manage, the sheer number of documents that they have just for each of the clients that they have.

[0:14:22.9] JM: Where do you insert yourself? Because you've got all these people that are on some stack where they've got a virtual machine and the virtual machine is running all kinds of applications. Where does the backup strategy fit in?

[0:14:37.8] KT: The way Rubrik works — Our primary offering is a converged appliance, although we're also now extending into cloud offerings.

[0:14:46.3] JM: Converged appliance. What does that mean?

[0:14:48.2] KT: All those four or five vendors that I've described earlier, it's everything built into one unit more or less, one scale-out unit.

[0:14:59.5] JM: There's a hardware box.

[0:15:01.3] KT: There's a hardware box, but it's not just hardware with kind of general purpose compute. It's also capacity optimized storage as well as flash for performance optimization, plus

that d duplication and kind of content management software that I've mentioned that utilizes the storage efficiently. Plus, at the very top of it all is the actual backup and recovery application which drives backup jobs, configures the policies that the user wants and then reach the data from the environment into our storage. Then also offers the feature through our REST APIs to recover this data into either their primary environment or running off our live storage as well.

[SPONSOR MESSAGE]

[0:15:50.7] JM: VividCortex is the best way to improve your database performance, efficiency, and uptime. It's a cloud-hosted monitoring platform that eliminates your most critical visibility gap, providing insights at 1-second granularity into production database workload and query performance. It measures the execution and resource consumption of every statement and transaction, so you can proactively fix future database issues before they impact customers.

To learn more, visitvividcortex.com/sedaily and find out why companies like Github, DigitalOcean, and Yelp all use VividCortex to see deeper into their database performance. Learn more atvividcortex.com/sedaily, and get started today with VividCortex.

[INTERVIEW CONTINUED]

[0:16:53.3] JM: When you come in and install this for the enterprises, what is exactly is happening? What's the process for getting them on-boarded to a situation where they can install a backup under any sort of circumstances where otherwise they would have lost data or whatnot?

[0:17:14.8] KT: Sure. Our process, unlike the legacy process, will probably take about a day if everything is optimistic. Most of that day is actually spent on just the actual hardware side of racking and stacking and getting the right people into the rooms and such.

Once our appliance or multiple appliances are actually setup hardware-wise, connected to a network, then we'll have the user, the customer, go through what we call cluster bootstrap. I'm not a huge fan of that term. It's a little technical sounding. I guess it's sort of an initialization process where they'll set up the IP addresses that they want for our box. They'll configure some

basic things, like how do I talk to my active directory environment? I guess to some extent, I forget what our actual walk through looks like now.

Skipping a couple of details, the main part they'll get to is then they decide to connect which kind of applications they want to be backing up. For the case of VMware, which is still the most common, they'll just enter their vCenter credentials, so this is just a username or password that their organization has granted for their backup to use. They'll enter those into Rubrik, then Rubrik will go and scan their vCenter, and this is actually a very exhaustive scan that shows us a lot. Just by scanning the vCenter, we can pretty much see all the VMs that are of interest. From that point, we let them just drag and drop VMs into policies that they define.

These policies will be according to business requirements, like — We come with some default built-in which we call gold, silver, and bronze just as a high level of what kind of SLA they expect for their backups. Generally, a higher SLA will have more frequent backups and longer retention. They'll choose the policy, then by dragging and dropping VMs into them, this will take them about 15 minutes. Once that's done, there's actually nothing else they have to do.

From that point on, Rubrik will proactively figure out which jobs are necessary and figure out the timings, then each of these jobs will connect to the vCenter using those same credentials to basically use VMware's APIs to get access to the underlying data of these VMs and then also use VMware APIs to read that data in. Once we've read the data in and the data is on our system, that's where Rubrik just takes over and there's no more interactions with VMware, or in general, with whichever application we're backing up.

[0:19:38.8] JM: As I understand, after you get that data in a state where it can be backed up, you can also index the backup contents into something that can be searched, and I guess that's a useful value ads. Explain why that's useful.

[0:19:56.6] KT: Yeah, I'm glad you brought that up. I actually had forgotten about that, because we've been focusing entirely on kind of the backup. We've never really considered Rubrik to be a backup company, and that's why we say we're data management. Kind of our foot in the door for a customer and something that just brings them value from day one is offering backups, something that's a core business need but something that they often don't realize as very useful

is to actually get value out of their backups. It's not just data that you have to pay for and you buy these Rubrik boxes or alternatively incumbent boxes. Just in case something goes wrong a year later, just so that you can — It feels like a high cost just for insurance, if you get what I'm saying.

We offer these value ads on top of backups too. Search is definitely one of them and really the main one that differentiates us. Because of the fact that we started with this live mount functionality, that actually made it very easy for us to build indexing and search on top of this compared to other vendors because it would have taken hours to do the restore. It's not really practical to constantly index these things. Since we offer instant access to the data, one of our very first kind of non-backup features was to offer file search and recovery within the backups that we take.

Even if somebody is not a backup admin, the experience of just being able to search through kind of old archives of their VMs or their applications and find kind of this particular file — We all know the value of search. You don't know exactly what you're searching for, but you enter a search term then you see a lot of results. This being instantaneous is kind of essential to search being useful because you'll often search one term, realize that you don't get the results you want, then quickly search for another term. You go through a few cycles of this and that's how you actually identify what you need. Then ones you've identified what you need then you can actually go through a full recovery or even an individual file recovery if that's what you're looking for.

[0:22:02.1] JM: Sure. Now, we're kind of glossing over a lot of the software architecture. I'd like to understand in a little bit more detail, what does it mean to take a snapshot of a VM and how aggressively do you have to do that. How are those VM backups being snapshotted, and do you version them? Just give me some overview for the process of taking these snapshots, storing them, perhaps garbage collecting old VM snapshots, the whole picture.

[0:22:37.5] KT: We think about what snapshots are. A snapshot is pretty much a full point in time reference of a given VM or application, whatever you want to think of, or even just a set of files at that point in time. Ultimately, what backups are from a user point of view is a collection of these point in time snapshots. If I, as a business, have my exchange environment backed up,

then what I see in the Rubrik UI is that over the last two years, I have at least one snapshot per month of this exchange data.

Now, if this exchange data was originally — Let's say if it's 5 terabytes of user emails and other data that comes in, but if you have one full snapshot per month over the last two years, then that's 24 copies of this 5 terabyte data. That's actually a huge amount. You as a customer wouldn't actually have a primary system which as five terabytes of raw hard disc capacity. Likewise, you're not going to pay 20 times that just to have backups either. In order for this to be economical at all, it's really table stakes. All kind of backup solutions have to have some form of d duplication going on and some form of, as you point out, managing of the actual snapshots involved to actually expire and consolidate older data that's not needed anymore.

A lot of what — I forgot what your original question is, but a lot of what — The act of taking a snapshot is actually comparatively simple. That's telling the application or VMware or whatever the workload is, telling the VMware to bring the data into a consistent state so that when we read it it's actually valuable data rather than just kind of intermediate data that didn't actually mean anything to anybody.

The act of taking a snapshot is simple. That basically freezes the data into a consistent state, then we read it into Rubrik, but where we do a lot of work is the actual data management because if we were to store this 5 terabyte logical copy naively, we would very very quickly run out of storage.

A large amount of what we do on the backend, which is entirely within the Rubrik box is — What we call managing the data is, first, what d duplication is that even when you have multiple logical copies of the same data, usually not all of it has changed since the last time you looked at it. At the very first level, what everybody certainly does is track only the few things that have changed. In the exchange example, only the new emails are actually what's changed. You don't actually have 5 terabytes of new emails every day. You may have a couple of gigabytes of new emails each day. That's the first level. That's honestly where the biggest savings come from.

Then even when you do this, you still run into issues where over time — If you're taking backups every day, over time this till does add up because if you're growing at 5 gigabytes a

day, you add that up over a couple of years. That will still come out to utilizing all your storage. This is where expiration and consolidation come in.

There's a lot of logic that — Kind of computation that you do with these deltas in order to identify the fact that, "Okay, although I've gotten 5 gigabytes of changes each day, every day for the last month, if I only want to retain one copy according to the policy, then I'm going to figure out that, "Okay, here are the two points in times that I do actually care about and all the changes that went on in between, if they're kind of overlapping changes, then I don't actually need to keep the full record of changes that have gone on in between. I'm only going to keep the two endpoints."

I'm obviously glossing over a lot of the details for how this computation works.

[0:26:27.0] JM: That's okay.

[0:26:27.8] KT: But that is the idea that what this effective data management at the local level is — Basically, what it comes down to at the end of the day is finding a way to efficiently store and ultimately to restore the data that logically is very very huge because it's many many many point in time copies, but in some way — That's what we focus on a lot, is to reduce this data so that the physical footprint actually doesn't have to be so big.

[0:26:54.1] JM: Those backups that are sitting in out in an appliance on-premise with the rest of the enterprise, then do you copy those backups again to the cloud so that you have an off-premise backup?

[0:27:11.2] KT: Yeah, that's exactly right. Something we haven't talked about yet and something that customers are also very interested in from Rubrik in addition to just being backup and recovery is to be kind of a path to leveraging the cloud, because every business these days does want to leverage the economies of cloud. You see the prices that S3 offers. Similar things like Glazier and Google Storage, these are very attractive to them, and there's also a very obvious use case for having this off-premise storage as you point out, which is basically disaster recovery, that in case a hurricane, or more likely total power outage and power surge occurs

which wipes out your entire datacenter, you want to still be sure that you have something out there or something that's offsite to cover all your bases.

The other big side of Rubrik's data management is not just kind of optimizing how we store things locally, but actually pushing things out to the cloud so that customers can start putting their footprint into the cloud as well. It's actually a pretty similar process. We have two features related to this, they're called replication and archival. Replication is designed for kind of, in some sense, live replication. Keeping, basically, the full copy readily available either in the cloud or in other on-prem datacenter, and archival is more focused on trying to push only the older bits, the ones that are less likely to need to be restored, but which for peace of mind you kind of want a copy anyway. Kind of pushing these older bits off to cheaper storage.

These days, actually, our replication and archival features actually kind of blur together as well as our cloud versus off-site, on-prem targets. I'd actually generalize that to say that, now, Rubrik data management, in addition to the local optimizations is also really heavily involved in just, in general, pushing the current snapshots out to some remote location, whether that's be cloud. It can be one of multiple clouds, or a customer's other datacenter. Then also doing the same management now that the data lives in multiple sites, because all these problems that I've described about doing the d duplication, computing deltas, consolidating and expiring the deltas that are in the middle, now this applies equally on the remote end, because even though S3 is cheap, you still don't necessarily want to — You as the customer, don't really want your S3 bill to just climb forever. Rubrik helps to manage that as well, and it's a very similar problem but with different challenges, because now it's all done remotely instead of on the local box.

[0:29:48.9] JM: Yeah, that's a pretty interesting problem. Does the enterprise — The enterprise, now that we've described more of this product, is the enterprise is getting their VMs snapshotted and saved to this on-premise appliance, and then the appliance is sending the backups to a cloud service, like S3, and storing the VM snapshot backups on that cloud storage solution.

There's a lot of interesting things that we can approach from here now that we've talked about these different product offerings. I'm curious, does the enterprise, typical enterprise that you work with, do they care what cloud service their stuff is getting thrown on to? Do they want control over that cloud service, or do they just want some opaque — Are they okay with just

being, “Yeah, we’re interfacing with Rubrik. Yeah, Rubrik, just go find whatever is the cheapest S3 bucket equivalent or storage bucket. All we care about is some opaque backup solution. How involved do they want to be with what you’re doing under the covers with some cloud provider?

[0:31:02.7] KT: Good question. I think what I’ve seen is it actually varies. Some customers will have stronger requirements and stronger pickiness about the cloud they choose and how much they control it and others will have less. Right now, I think as we’re kind of pushing more to early adopters, the ones who are ready to get into the cloud and the ones that are more opinionated, I think we would lean a little towards being picky, but I think we haven’t hear any customer who just don’t care, who just want whatever the cheapest solution is.

To expand on that, the ones who are a little more picky are usually those who already have a footprint in the cloud for one thing, and so they just want to reuse that account. They don’t want to deal with multiple accounts, or they can have kind of local regulatory requirements about which clouds they’re allowed to us. For example — We are an international business by now, and so some countries have regulations that you must use a cloud with a local presence, for example.

We see a fair amount of being picky these days, but we have definitely heard a couple who just think, “Well, I just want to figure out what the lowest cost option is.” Sometimes that can even actually be in on-prem object storage. There are vendors these days that do sell on-prem objective stores. There are also larger enterprise clients who’ll even build their own on-prem objective stores because they started their business before objective stores became commodity. They’ll have kind of similar S3 compatible on-prem stores that they use in-house that they want to use.

[0:32:27.4] JM: What’s funny is I did a show recently about Microsoft’s IoT stuff and they now have an offering, or maybe they’ve had it for a while, where it’s on-prem. It’s basically Azure on-prem, and a lot of companies actually want this. If not for just compliance reasons, like for having an on-prem cloud, they want it for latency reasons. The discussion with the IoT people is you might want your machine learning models hosted on-prem because the latency is going to be so much less if you’re querying that machine learning model for something that’s really

mission-critical, like some kind of like hazard control system, or a centrifuge control system, something like that.

[0:33:17.9] KT: Yeah. Definitely, the quality of the network connection between kind of the application and the data that's being stored does matter a lot in some certain use cases. I can definitely imagine IoT being pretty latency sensitive. Maybe also throughput sensitive as well. This actually is one of the challenges that we hear for a lot of businesses that are trying to move into the cloud and something that even we internally at Rubrik kind of face as we shift some of our internal workloads between the cloud and on-prem is that this network in between these two locations becomes a very quick bottleneck — Sorry. Very quickly becomes a bottleneck and it actually is pretty interesting and unique challenge in itself, and this is something that we're starting to tackle as well. I don't think I can reveal too much, but we recognize the fact that this, basically, datacenter sprawl is starting to become a problem as we enable customers to leverage all these other kind of storage technologies. We think there's work to be done in that area as well.

[SPONSOR MESSAGE]

[0:34:27.9] JM: At Software Engineering Daily, we need to keep our metrics reliable. If a botnet started listening to all of our episodes and we had nothing to stop it, our statistics would be corrupted. We would have no way to know whether a listen came from a bot, or from a real user. That's why we use Encapsula to stop attackers and improve performance.

When a listener makes a request to play an episode of Software Engineering Daily, Encapsula checks that request before it reaches our servers and filters bot traffic preventing it from ever reaching us. Botnets and DDoS are not just a threat to podcasts. They can impact your application too. Encapsula can protect your API servers and your microservices from responding to unwanted requests.

To try Encapsula for yourself, got to encapsula.com/sedaily and get a month of Encapsula for free. Encapsula's API gives you control over the security and performance of your application. Whether you have a complex microservices architecture, or a WordPress site, like Software Engineering Daily.

Encapsula has a global network of over 30 data centers that optimize routing and cacher content. The same network of data centers that is filtering your content for attackers is operating as a CDN and speeding up your application.

To try Encapsula today, go to encapsula.com/sedaily and check it out. Thanks again Encapsula.

[INTERVIEW CONTINUED]

[0:36:13.1] JM: You don't have to go into specifics, but can you characterize — Another interesting topic we've explored a little bit on this show is the idea that as we get things like self-driving cars everywhere and drones everywhere, one way to think about these drones and these self-driving cars is they're roaming datacenters. I use those as examples to characterize that there is — We're going from a place where there's cloud and on-prem, to a place where there's cloud on-prem and then everything in between. There's self-driving cars that are basically datacenters on wheels. There's drones that are datacenters in the sky, and then there's just your smartphone which is a really powerful computer that has tons of unutilized CPU cycles you could be doing stuff with, maybe storage is getting better on the phone and maybe people aren't utilizing their storage enough. You could maybe have some kind of peer-to-peer storage solution.

Anyway, the point I make is just that the gradient between cloud and on-prem is thickening. As you described it, datacenter sprawl is only going to get worse. Could you shed some light on the conversations you're having with these companies about, I guess, the problems that they're having with that datacenter sprawl or how they're solving those problems?

[0:37:36.1] KT: The physical challenge I described, I'm not sure, to be honest, what is being done in these areas, and even within Rubrik I think we're still kind of spinning our wheels on that. I can say that there's definitely also kind of a more logical and management problem, which is that if you're a business now, especially if you're managing IT for a given business and now you're dealing with, basically, exactly this sprawl that even just from the point of view Rubrik, you have your server data which traditionally was stored in exactly one location, which is now going to be stored actually in one primary location plus one secondary, but still on-prem

location, plus one or more clouds. That's just for your server data. Then you'll also have your user data, which not too long ago, was almost entirely on a combination of desktops and file shares. Now, it's on desktops, laptops, smartphones, tablets, any other new devices that come in.

There's a pretty clear and big management problem with this, which is that, you, as an IT admin responsible for knowing where data lives and knowing who's responsible for which piece of data and knowing what policies you want to apply to the data and how you enforce these, whether it's from a security point of view or backup's point of view, this is getting pretty, pretty complicated.

Rubrik's answer to this aspect of the problem is something you alluded to earlier that some customers don't' really care about what's going on behind the scenes, whether it's cloud or whether it's some other on-prem site or whether it's local data. What they do see and what they do appreciate is Rubrik being that single pane of glass for all these data together and being a one-stop-shop for whatever their data needs actually are.

The more they put into Rubrik — Today, we're still primarily focused on server data, but for a given server that has that's data managed by Rubrik, they can be kind of assured that the one place they need to worry about going to is the Rubrik UI or the Rubrik API, then they can view whatever the report they need to from there and they'll have kind of a holistic view of that logical piece of data no matter how many different places it lives in.

[0:39:42.6] JM: Strategically, that opens up Rubrik to — You guys could either look at the different cloud providers and always find the lowest cost. If Rubrik's customer only cares about the opaque backup and application hosting service — If Rubrik is essentially a platform as a service to these companies, you could just find the lowest cloud provider or you could build your own cloud. Are there any strategies that you guys have settled on at this point or are you mostly just kind of focused on — I don't know. Yeah, talk more about that.

[0:40:20.9] KT: What's where I probably can't reveal too much, but I can say what's public. What we have technically today is I think we've announced our 3.2 and soon, later in this year, will be 4.0. As of today, what we'll have is mostly tactical, which is we give customers the option

to choose whichever cloud they want. For the past year, basically, the last 12 months or so, in the last few releases, we've been focusing on just expanding our service area so that customers can choose just exactly whoever functionality that they need from us. The point is, we've been focused entirely on actually having the functionality. Whereas a year ago, we were actually a much narrower focused product.

After we build the functionality, what we're also starting to spin up our RND on is actually figuring out exactly what strategy makes sense for — In addition to offering the functionality, making this very very simple, very opaque, very easy for Rubrik to be the one-stop-shop which can include, as you kind of allude, can include these kind of more creative options where if a customer doesn't necessarily care which cloud they pick, do we in fact pick one for them or do we in fact operate one for them as well?

[0:41:34.5] JM: Rather than talking about big, ambitious, futuristic product development things, let's talk about lower level concerns. What kinds of — The CAP theorem, this distributed systems tradeoff scheme. It's basically shorthand for — There's all kinds of tradeoffs that you need to make between consistency, availability, and partition tolerance. Basically, if you want to be consistent, then that often means you're going to be less available. If you're taking snapshots really aggressively — Sorry. If you want to be extremely up-to-date every time a user asks you for a snapshot, that might mean that you're not always available to restore a snapshot. I guess you could talk more, generally, what are the kinds of tradeoffs the distributed system's tradeoffs you make in terms of how aggressively you snapshot and what kinds of data loss you're willing to actually tolerate. Yeah, I guess just explore that discussion.

[0:42:41.3] KT: Sure. I'd love to. We should probably offer some backend actually. The reason why the CAP theorem applies to Rubik is that in addition to being so simple, something that I haven't talked about that which is the fact that we're a scale-out platform. Scale-out means we are a horizontal architecture. This is one of the ways that we're actually just very forward-facing compared to some of the more legacy incumbent solutions, which they're built on architectures in the '90s and in the early 2000s. They have their scalability limits. For customers, the main impact of this is that it's very difficult for them to either plan for the capacity that they need or to upgrade the capacity later because they basically have to do these forklift upgrades where they rip and replace everything and buy a bigger box next year when they have more needs.

Rubrik is scale-out in the sense that from a customer's point of view, they can start with a bare minimum today, which would be three or four nodes, but in the future if their needs grow, they'll just add another three or four nodes or whatever number makes sense for them. They don't actually have to preplan — Basically, upfront plan for exactly how much they're going to grow in each year to be cost-effective.

Because Rubrik is scale-out, this doesn't mean we're a distributed system, and that's where the CAP theorem applies. This is a pretty famous theorem, but often kind of misquoted and misunderstood. As you say, the main tradeoff it's describing is basically choosing between in certain scenarios of something going down. Usually, you'd be thinking of either a node going down or a network connection going down. In those situations, the generalization of it is that you're going to be forced to choose between consistency and availability.

If you think about it, as far why the CAP theorem makes sense, it's kind of intuitive that you have this tradeoff because if something did go down, then you have to lose something, and so you're either losing just access to that data or you're losing some kind of guarantee about how current or how consistent that data is.

As far as mostly what we've chosen with Rubrik and what I think is arguably kind of the trend these days for those who stay up-to-date on kind of the latest database, research topics, I think a lot of companies these days are choosing consistency, and that's what Rubrik has done as well. Kind of the trend towards consistency is the realization, I think. Broader than Rubrik is the realization that availability without consistency just often isn't very useful and often actually — You don't actually care if you get your data if you have no promise at all that it is the correct copy of the data or that it reflects all the most recent updates, things like that.

Yeah, for both Rubrik and for the broader industry, we're all starting to realize that consistency — Or availability without consistency isn't particularly valuable, and also moreover from kind of an internal facing software engineering point of view. We're all starting to realize that when you have availability and weak consistency and then you force your application developers to write these workarounds for those not so optimal consistencies, then you actually end up having very hairy, very messy hacky code that's really difficult to maintain actually. It always has weird bugs

that get introduced or weird interactions that go on. Then that's actually worse. Then you actually can almost guarantee that the data is inconsistent at some level, whether it's at the API level or later once you do some processing.

I'd say for the most part we've chosen consistency, but I'd say that also certainly it brings a lot of the challenges, because I think a lot of the problems that we do face, especially things that come up in the field, since we sell our product into customer datacenters, we don't get the luxury of Google or Facebook where we're operating our own datacenter and we maintain everything to the highest standards. Outages or network partitions are very rare.

We actually sell into a customer environment, and so we don't have control over that. A lot of bad things do happen in these environments. We'll have silly things like a power went out on this particular rack or this particular node, or the network really did go out because — We call it layer one problem where somebody just actually pulled the cable or the cable got damaged, or also certainly plenty of software problems with either the things that we talk to or the components within our own nodes.

A big problem that we do face since we — I guess we're improving on these things, but a big problems we've often faced is that since we've chosen consistency, we often do end up sacrificing availability in a very visible way. We've had quite a few field cases of a customer experiencing their UI for Rubrik is completely inaccessible, or that their backup job start but then something fails and then they can't write data into the database.

We face these kind of issues. I think more often than the likes of Google or Facebook because of the fact that more things can go wrong in these environments.

[0:47:43.8] JM: Now, eventually, you're going to have to compete, or maybe you're already competing with the backup solutions that are offered by Amazon and Google cloud products, maybe Microsoft cloud products. I know that today they are doing stuff with the Snowball, like the Amazon Snowball, where they send you this thing that you pipe all your data into and you can have a big backup of it. I don't know much about these products. I think the other service where Amazon brings a big truck with a bunch of servers on it and then they pipe all your data out of your datacenter into their truck and then they drive off and put it in a datacenter or

something like that. Can you talk a bit about these other — What the actual cloud providers themselves are offering in terms of this sort of backup area. I know that's not like exactly the same thing, because you're sort of offering this way that gets you gradually into the cloud by way of offering a backup service. It's a very creative way of getting a tow-hold into these enterprises and eventually offering them cloud services. Maybe you could talk about what the cloud providers are actually offering.

[0:49:03.6] KT: Yeah, that's actually very interesting. It might be surprising from an outsider point of view right now, but we don't see Amazon or the other cloud providers offering backup or backup-like services. We don't actually see them as competitors. Certainly, not today, and even looking into the future where we assume more services are moved into the cloud. Even then, they don't feel like a competitor.

Right now, actually, in the case of Snowball, for example, we're treating them actually as a partner. I don't remember if we chose to release this. I hope I'm not revealing too much, but we did actually have an internal project. It started as a hackathon project, to actually integrate with the Snowball. We entered the Amazon trial to get a Snowball in-house, started playing with it, started understanding its capabilities and actually built a tool to actually integrate the Rubrik S3 archival feature with the Snowball. I think it's really interesting. I think, as you say, it's kind of a funny site seeing Amazon build these kinds of features, build the actual Snowball, offer the services of having a big truck come out to you and kind of have a temporary mobile datacenter that you can archive to and then ship it off to Amazon.

I think if I were to say why it hasn't felt like a competitor so far, I'd say the main thing is I guess that Amazon's purpose isn't necessarily for backup, and I'd say backup isn't just the storage that's involved. Amazon really wants to get you on to their cloud, but they're not as opinionated about what you do there, whether you call that as backup or just storing a DR copy. For that matter, the main thing that I think they're trying to get customers on to is to do their primary workload on Amazon anyway.

Even if the storage is on Amazon — because you can actually look at that compared with our own archival feature today. We've had S3 archival since almost year one, and so we've never considered that to be in any way a competition for what we sell even though we sell on-prem

storage as well. We recognize that the on-prem side of Rubrik is far more expensive basically than S3 and that's why customers will buy Rubrik and then buy our feature to archive to S3. I think the realization is that there are a lot of different things going on with both of the backup world and the cloud world and that certainly Amazon has a part in actually moving the data and posting the data and being kind of the sheer economies of scale host of that data, but there are still a lot of services that so far doesn't look like Amazon is that interested in and the other cloud providers aren't interested, but that we as somebody kind of focused on the backup and the management aspects of these things still offer a pretty large value ad.

[0:51:54.7] JM: It's so funny, I have a lot of trouble understanding who are competitors and who are not competitors and who are coopetition. I was at Microsoft Build Conference recently and I was just walking around the Expo hall and there was one demo — I can't remember which Microsoft Cloud Service was demoing me. Oh! I think it was — Was it HDInsight, or something? He pulled up on the screen, he's like, "Okay, here we're going into the Amazon Azure Cloud dashboard," and he showed me the marketplace and he's like, "Here's how you install the Cloudera Hadoop distribution on Azure."

I was like, "Oh!" because I was assuming Cloudera is just an outright competitor with Azure, but actually there's kind of — I don't know how well-documented this is, or maybe a lot of people know this and I'm just naïve. I guess it's just like this entire marketplace of things that are being layered on to cloud services and kind of the early stages, but already you see — I saw a partnership yesterday where it was like Box is offering — Box in concert with Azure, and Azure is offering Box in concert with Azure.

You kind of see these partnerships between the cloud providers, which are — They're certainly doing infrastructure as a service and they're also doing some stuff layered on top of it, but they're also showing that they're completely happy to offer competing platform as a service and software as a service products on top of their infrastructure. They're perfectly willing to partner with stuff. I don't know. I don't have a question here, but it's such an interesting area to watch develop.

[0:53:48.8] KT: Yeah, I agree. It is actually very interesting and it feels a little unique to the software industry to be honest. I agree. I'm still kind of — I've been thinking about this for a

good while even before today as well. I don't really have a full answer to it myself, but I think you did kind of latch on to one of the core reasons, I think, which is the fact that software and services and everything related to these technologies has so many layers and so many pieces, so many use cases, that there actually is room for multiple companies which offer kind of very overlapping services to still both — Or all of them just to get the same customers.

I think layers in particular is where it gets interesting, because for the examples you described or for what we just talked about with Rubrik and Amazon, I think the recognition that they're multiple layers and the fact that different customers see different values out of different vendors of those different layers gets interesting. Basically, what it comes down to is, for example, are you going to run Hadoop on Azure or are you going to — Hadoop on top of Azure versus Microsoft's competitor to Hadoop on top of on-prem.

It's like there's just so many options. There's so many options and so many permutations that if you're one of these vendors. If you're Microsoft, or Amazon, or Rubrik, you offer solutions for all the multiple layers of the stack. Even though on the surface it sounds like you don't like a competitor — You don't like your customers buying some competitor product for this other layer of the stack. You also kind of realize that at the same time by being open to that, you are also getting more business because you get to compete in this other area of the stack which for whatever reason your competitor on some other layer isn't as strong.

I agree. The very interesting dynamic you end up where like almost by default by now, everybody is pretty much coopetition with everybody in some way. I think the main reason actually, to be honest, since I've been following the industry for a long of time too, one thing that's changed relatively compared to just, let's say, 20 years ago when I was just young and I was just kind of starting to learn about computers. In those days, if you remember, the attitude was actually kind of different. Especially from Microsoft, which was just so dominant back then, Microsoft's attitude was kind of Microsoft everything for the whole stack or nothing.

It was very clear, I guess, objective from Microsoft to want to control the entire stack. Whereas, obviously, things have changed now. I think there are more player. Now, because there are more players, everybody involved had started to realize that, "Yeah, we have some components that compete with each other, but at the same time our customers seem to like that diversity.

They seem to want something at the platform level — They have their preferences at the platform level and they have another preference at the application level." It's actually just a better kind of global picture if we are willing interoperate with each other's products and let customers have the choice.

It's kind of weird when you think about it also — I just realized this, consumer versus enterprise. Usually, the wisdom today is that consumers actually don't want too much choice. That's confusing. That creates a headache for researching which products they want. They want to pick from among one to three choices. The interesting thing is for enterprise, it does seem to be different that businesses would rather have choices between the 3x by 4x different permutations of different combinations you can put together and have that flexibility to actually exactly what suits their business.

[0:57:25.9] JM: Okay. I've got one last question. It's late June 2017 and we're seeing some crazy stuff around ransomware and you're also starting to see a lot of vendors talk about, "Okay, our solution is good against ransomware." Do you have any thoughts on how big of a problem ransomware is going to be? Is this just going to be a short-term thing while companies kind of figure out how to defend against it or they move away from Windows or whatever the solution is going to be? Is this going to be like an ongoing problem? Because I can see backups as being a pretty good defense because, "Okay, you get hit with ransomware, backup to before you were hit with ransomware."

[0:58:12.7] KT: Yeah, two answers to that question, I guess. The first one, the direct answer is I'm still keeping my eye on that situation as far as globally how big of a problem ransomware is. It's certainly trending right now, but it's hard for me to make out exactly what's going to happen, because on the one hand I see kind of the big impact that it's been having, a couple of attacks. We had one actually — Or not we, but the world, kind of had one just yesterday, I think. I see the big impact that's going on and I see all the fuzz that it's creating.

[0:58:40.3] JM: It's hard to know if it's overhyped.

[0:58:42.5] KT: Right. There are some signs that it might be overhyped and it may not be in my interest to say this for Rubrik, but it does feel a little bit like it might even be starting to fizzle,

because even for the incident yesterday, I'm kind of reading articles saying that, "Well, a lot of them just decided not to pay out," or even for the larger attack a couple of months ago, a lot of the ones who were trying to pay just couldn't even find a way to pay. It does seem like there's a component of hype, but that's why it makes it hard to say, in the next five years, how big of a problem this will be.

I'm afraid I don't have a great answer to that, because I myself am not really sure either. What I can say though, the second part of that is how it does relates to Rubrik. Intuitively, you do think backup should be the strong defense against ransomware, but one of the things that's actually surprisingly — This just kind of evolved organically for us, which has been also a differentiator for us in this Ransonware area, is that in order for those backups to be effective, you actually have to be guaranteeing that they're immutable. Immutable meaning nothing outside of Rubrik can possibly tamper with them.

This is huge, because actually it turns out there — I don't remember if these are public, but there were some cases of customers using legacy backup vendors who thought they were protected because they had a backup, but then it turned out that not only did the ransomware corrupted their primary copy but also their backup copy because of the fact that they weren't immutable. They actually didn't get that protection.

We at Rubrik have been kind of pointing this out, that we've had this kind of built in from day one because of our architecture, and that's been kind of a nice piece of buzz for us as well.

[1:00:25.0] JM: All right, Kenny. It's been great talking to you. I enjoyed this conversation a lot, and I look forward to seeing what develops with Rubrik.

[1:00:31.3] KT: Yeah, thank you for having me. It's been fun.

[END OF INTERVIEW]

[1:00:36.4] JM: Spring is a season of growth and change. Have you been thinking you'd be happier at a new job? If you're dreaming about a new job and have been waiting for the right time to make a move, go to hire.com/sedaily today.

Hired makes finding work enjoyable. Hired uses an algorithmic job-matching tool in combination with a talent advocate who will walk you through the process of finding a better job. Maybe you want more flexible hours, or more money, or remote work. Maybe you work at Zillow, or Squarespace, or Postmates, or some of the other top technology companies that are desperately looking for engineers on Hired. You and your skills are in high demand. You listen to a software engineering podcast in your spare time, so you're clearly passionate about technology.

Check out hired.com/sedaily to get a special offer for Software Engineering Daily listeners. A $600 signing bonus from Hired when you find that great job that gives you the respect and the salary that you deserve as a talented engineer. I love Hired because it puts you in charge.

Go to hired.com/sedaily, and thanks to Hired for being a continued long-running sponsor of Software Engineering Daily.

[END]