EPISODE 935

[INTRODUCTION]

[00:00:00] JM: In 2009, the Berkeley AMPLab was a center of innovation. Three projects from AMPLab have turned into successful open source projects and companies; Spark, Mesos and Alluxio. Haoyuan Li was the creator of Alluxio and he returns to the show to discuss his journey taking Alluxio from a research project to a company that has customers, including Alibaba, Baidu, Wells Fargo and Samsung.

Alluxio is a distributed, fault tolerant, tiered storage system. Alluxio allows application developers to think in terms of the latency that they require from their infrastructure rather than the details of different storage systems.

Haoyuan discusses the process of integrating with gigantic companies, like cloud providers, telecoms and huge ecommerce companies. Alluxio is also hosting the upcoming conference; The Data Orchestration Summit, November 7th at the Computer History Museum, and this is in Mountain View, California. If you're building a software project, post it on Find Collabs. Find Collabs is the company I'm working on. It's a place to find collaborators for your software projects. We integrate with GitHub and make it easy for you to collaborate with others on your open source projects and find people to work with who have shared interests so that you can actually build software with other people rather than building your software by yourself.

Find Collabs is not only for open source software, it's also a great place to collaborate with other people on low code or no code projects, or find a side project if you're a product manager or somebody who doesn't like to write code. Check it out at findcollabs.com.

[SPONSOR MESSAGE]

[00:01:58] JM: As a programmer, you think an object. With MongoDB, so does your database. MongoDB is the most popular document-based database built for modern application developers and the cloud area. Millions of developers use MongoDB to power the world's most innovative products and services, from crypto currency, to online gaming, IoT and more. Try

Mongo DB today with Atlas, the global cloud database service that runs on AWS, Azure and Google Cloud. Configure, deploy and connect to your database in just a few minutes. Check it out at mongodb.com/atlas. That's mongodb.com/atlas.

Thank you to MongoDB for being a sponsor of Software Engineering Daily.

[INTERVIEW]

[00:02:54] JM: HY, welcome back to Software Engineering Daily.

[00:02:57] HL: Thank you. Pleasure to be here again.

[00:02:58] JM: It's great to have you back. You started Alluxio about four years ago, and I think that's around the time I interviewed you. So that was based on your research around distributed memory.

[00:03:10] HL: Distributed virtual file system. That's my PhD thesis title.

[00:03:15] JM: Right.

[00:03:15] HL: Yeah.

[00:03:16] JM: What problems does a distributed virtual file system solve for a developer?

[00:03:22] HL: Essentially, there are different layers in the whole stack. So we are positioned to be a platform system for the data related stack. What does that mean? It means all the advanced analytics or the machine learning AI type data-driven applications, they need to interact with data. All these type of workloads in the end of the day would serve these type of workloads and developers. In this stack, you have platform engineers. You have application engineers, different type of engineers.

Essentially, our system is being deployed by platform engineers to serve the applications which is being developed by the data-driven application engineers. So we enable different type of

things for different type of engineers. For the application engineers, essentially, we can see all their data in their organization through our system under a unified global namespace and we make the data interaction very fast for all these different type of applications as well. That's our benefits, the bandwidth it provides for the application engineers.

For the platform engineers, essentially, they can use us to build their central data platform to provide very easy data access and data interaction for all the other engineers inside the organization. That's what we provide to these different type of engineers.

[00:04:56] JM: It's quite beautiful, and we've done a bunch of shows recently about this idea of a data platform. We talked to – One of my favorite shows that we did not too long ago was with Uber, about Uber's data platform, and it's got all these ETL going on and it's really beautifully crafted. I think one that help me understand how Alluxio works is the API, the APIs for writing data, which allows a developer to specify how they're writing it.

If you write to Alluxio, you can write directly to the memory storage and you can also write to the underlying storage or you can just write to the underlying storage or you can just write to the memory. You can do those two operations synchronously. You can do those two operations asynchronously. It gives the developer a ton of flexibility for where they're doing their write and potentially to multiple places. Why is that useful?

[00:05:54] HL: Actually, before answering that question, we can take a step back. We can talk about why do we do this in the end of the day? What's the vision and some of the use cases? We can use the use cases to talk about why people use this today and in the production environments.

Fundamentally, it's about a data revolution and we are in the data revolution era, right? Once you're at the early stage 20 years in and we have different type of data applications, like machine learning, like analytics, etc., and there'd be more of these type of data application being created. That's how the industry is moving forward.

With this, you have so much data, lots of data being created, generated, collected and stored. Then this triggered the industry to become very excited to create all type of storage systems to

store the data. On-premise storage, you have file system, object stores and now you have cloud storage, all the public clouds, like all the major vendors globally, like several in the U.S., several in APAC, maybe some in Europe as well. I'm not that familiar with that. At the same time, from the user perspective – That's the ecosystem. From the user perspective particularly, from the big enterprise perspectives, all the data siloed, all different type of storage deployments. How to access those data? Leverage those data become very challenging.

Many developers, many modelers, researchers, when they want to do some experiments, some new data they want to access, which they didn't have the access to in the past, they prompt to wait for either four weeks or several months to gather data. That's unbelievable. So then just to experiment the data, right? That's a challenge of these data siloes.

A way these data siloes traditionally industry tried to solve the issues, say, by creating a new storage. Then typically a store will be this new storage is so awesome. So you put all your data into this new storage, then you solve the data silo issue. You remove other siloes. That has always been the pitch over the past 20, 30 years. We fundamentally, based on the industry trend, we believe that will not solve the issue. Actually, there's cycle in our industry, storage industry. So every 5 to 10 years, another way of a generation of store systems. Therefore, they will not remove siloes. They'll continue to add more data siloes.

We at Alluxio, we actually take a fundamentally different approach from the past. It's not creating another storage, we're not doing that anymore. We essentially create a new layer, an abstraction layer we call the data orchestration layer on top of all the storage deployments.

Then that layer, the system, we call data orchestration system, essentially abstract the data. Virtualize the data from all different type of storage deployments and provide the data access efficiently to all the data-driven applications so that we made it happen that – We made it possible that any data-driven applications, those applications only need to face this data orchestration system which we implement as the open software called Alluxio only to interact with this system to interact read and write all the data in different type of data siloes, storage siloes. Maybe S3 from Amazon, maybe in HDFS, maybe [inaudible 00:09:21] from all different type of storage vendors. Essentially, that's what we do, what we achieved from the whole ecosystem perspective.

Then what are some values from the developer perspective? We enter the ecosystem into the middle between all the data-driven applications and all the storage systems, all the storage siloes. From the individual developer perspective, there are so many different type of use cases.

For example, a very popular use case today is a hybrid cloud use case. Many companies, many organizations, there are exploring the cloud initiative. They want to try the cloud. For them to try the cloud, say, any data-driven applications, like a machine learning, like Tensorflow, PyTorch, etc., or the analytics, like Presto, like Spark, like [inaudible 00:10:17], they need to access the data in the end of the day.

It's a big challenge for this enterprise to enable this journey, because they still have the data on-premise. So then before a solution like this data orchestration system, what they have to do is that they have to make a decision, say, "Oh! This organization is going to move a lot of data into the cloud and then start to explore the offerings in the cloud," which is very time consuming, and some organization takes two, three years to make any decision like this.

So then with the data orchestration system like us, essentially this enables a very, very simple, like hybrid cloud architecture, which means they run the applications in the cloud. We call it zero data copy cloud bursting solution. Essentially, application on top in the cloud, bursting to cloud, run data orchestration system like Alluxio alongside the data applications. Then this data orchestration system will manage the data movement between cloud and on-premise securely and also performantly.

Based on the policy, can have certain features like caching feature to cache the hot data in a cloud close to the compute so that these hybrid architecture works, this application in the cloud, still have the performance without having to migrate all the data into the cloud. So that effectively reduce the barrier for all these companies to try the cloud. That's very, very popular use case.

Then from the developer –

[00:12:04] JM: Wait. Just to make sure I understand correctly. Basically, the idea is you go to a company that's having trouble moving into the cloud or having trouble trying out the cloud and you say to them, "Look, we're going to give you an in-memory data layer in the cloud that allows you to have fast access to the data that you want to cache in-memory, and this will be a great way to onboard into the cloud.

[00:12:28] HL: 90%. It is correctly, but maybe I can add a little bit more.

[00:12:32] JM: Sure, yeah.

[00:12:33] HL: Directionally, absolutely. But instead of just only being memory, we actually have a feature called tiered storage feature so that we can manage all the storage tiers like Alluxio deployed in any computing environment. That means memory, SSD and HDD. We essentially manage all those resources in the cloud compute cluster and manage the data back and forth, moving back and forth. That's why we call the orchestration, data orchestration. That's the meaning of it, right? The system can intelligently decide when to move the data where. That's what we do.

Depends on the workload. Some users will absolutely want to get the best performance, then those users want to provision more memory resource to be managed by the system. Some users, they're happy with good performance already. So maybe just the memory plus SSD, or even HDD, to be managed by this data orchestration system. That's what we do. But you are absolutely right. This very, very popular use case, this hybrid cloud, to make the cloud experiment journey very, very simple. That's a very popular use case for these many companies. That's a use case.

[00:13:52] JM: Let's give this a little bit of contrast. There are common systems for caching data. Redis, for example. Why would I use Alluxio? I've got Redis. I can use Redis as my object cache. Why would I need anything else?

[00:14:08] HL: You exactly said it. Redis is a caching system. It's a very good caching system. It's a key value store caching and it's very popular for many database type of workloads. But we are more targeting two things. Number one, we're not a caching system. We have a caching

feature. Essentially, what the difference is that from the upper layer perspective, when they this particular open source system, they just see a file system [inaudible 00:14:37] other APIs. For example, like objects or APIs as well, but they see a file system with a file system namespace. There are many things behind-the-scene. You can mount HDFS, S3, GCP storage, [inaudible 00:14:53] and IBM Cleversafe, if anything, Ceph cluster into Alluxio as a folder.

It's very, very similar and analogous to like, say, when we are using our PC, and in the PC you'll have your local file system. But behind-the-scenes, that is essentially SSD, HDD or NFS and could be from Samsung, Intel, like Western Digital, Microsoft and that type, etc. But as a PC user, you never knew and you never asked.

That's the beauty. You just see a very simple namespace when you are interacting with your data in the PC environment. You're interacting with files and folders and all your PC application developers, when they are interacting with a data, they're not interacting directly with SSD, right? They are interacting with their file system happened to be maybe back upped by the SSD or backed up NFS. They don't care.

[00:15:58] JM: What people may not know, listeners have varying degrees of experiences, that a file system is an interface, and like that I understand it. A file system is basically an interface, and if you can build an in-memory system that serves that file system interface, it's going to be very fast. If you build it on top of tape drive, it will still be the same interface. It's going to be much slower.

[00:16:25] HL: Actually, from Alluxio perspective, we have this policy and we offshore the data. From the backend perspective we call it under storage for Alluxio. It can be any type of storage database, HDD-based, SSD-based, it doesn't matter. So we will, based on the policy, move the data around to make sure the applications will have the right performance with IO that they need. That's the beauty of it.

Essentially, go back to the early question you had, what's the value to those developers. To the application developers, like Presto developer, like Spark or Tensorflow developer, they never need to care about where to store the data and how they store the data anymore and how to access the data. They just don't need to care about it anymore. They just talk to a system like

this, Alluxio, a data orchestration system. Then their system, like platform administrators or operators will mount like the right under storage system into the folder so that they can access. That's in a nutshell make their life much easier.

Then on top of this, because of caching, as well as other functionalities this system has, it's much faster as well. In some cases, for example, a hybrid cloud architecture use cases, we've seen 10 times end-to-end performance improvement and we've seen some of the users. They have improved the modelers or the application users improve their working efficiency by four times. Previously a work, they need a year to get it done. Now you need three minutes to get it done, which is fabulous.

[SPONSOR MESSAGE]

[00:18:16] JM: I never liked searching for a job. It's painful. Engineers don't want to make a sacrifice of their time to do phone screens, and whiteboard problems and take-home projects. Everyone knows that software hiring is not perfect, but what's the alternative? Triplebyte is the alternative.

Triplebyte is a platform for finding a great software job faster. Triplebyte works with more than 400 tech companies including Dropbox, Adobe, Coursera and Cruise. If you've been hearing about Triplebyte for a while, you will be happy to know that Triplebyte just launched a brand-new machine learning track and they'll now be helping machine learning engineers find jobs in the same way that they've already helped generalist, and frontend, and mobile engineers. It's amazing seeing Triplebyte expand into these specific verticals, because they're so efficient about matching high-quality engineers to great jobs.

Go to triplebyte.com/sedaily to find out more about how Triplebyte works. You can take a quiz to get started, and if you end up taking a job with Triplebyte, you get an additional $1,000 signing bonus because you'll use the link triplebyte.com/sedaily.

If you make it through that quiz, you get interviewed by Ttriplebyte and you get to go straight to multiple onsite interviews. Its economies of scale for software engineering interviews is pretty sweet to see that centralized in a place that gives you those economies of scale. I'm a fan of

Triplebyte. I hope it gets bigger and bigger and creates more and more economies of scale in the miserable hiring process of getting a job as a software engineer. Make that painful process a little bit better with triplebyte.com/sedaily.

Thank you Triplebyte for being a sponsor.

[INTERVIEW CONTINUED]

[00:20:29] JM: I can think of all kinds of applications where this would be potentially useful, like backing Elasticsearch, or having my data in memory to train machine learning models, or having my data in memory because it's a machine learning model that I want fast access to. What are the most common applications that you see people using the in-memory part of Alluxio layered system, Alluxio system? What are the most common applications?

[00:21:02] HL: Funny that you mentioned our previous name, Tachyon.

[00:21:05] JM: Sorry, man. I can't get it out of my head.

[00:21:07] HL: No worries.

[00:21:08] JM: I like that name.

[00:21:09] HL: Yeah, I like that name as well. We rebranded because we got a legal letter. There's a company that –

[00:21:14] JM: That was what? Oh God!

[00:21:15] HL: They owned a trademark for Tachyon, so we cannot use it anymore. That's the reason. Yeah.

[00:21:18] JM: Oh! I thought it was just like Alluxio sounds cooler or something.

[00:21:22] HL: Both are cool.

[00:21:23] JM: Both are cool.

[00:21:23] HL: Alluxio means all user experience IO. Yeah, it also means all luxury IO. We want to be the luxury product, because we don't want to be a random product. We want them to be very useful, tasteful.

[00:21:39] JM: Beautiful. Nothing I love more than tasteful luxury UIs.

[00:21:44] HL: Yeah. Going back to the original question, you're absolutely right. All the case workloads use case you mentioned, we have seen them all, and there actually are [inaudible 00:21:56] materials for all the cases you mentioned, like all those out there [inaudible 00:22:01], like production deployments, etc.

Go back to what are the most popular workloads use cases today. From a stack perspective, our top three popular analytics and the machine learning frameworks on top of us today, they are like Presto for interactive SQL query, OLAP. They are like Apache Spark, and they are like Tensorflow for machine learning type of workloads. Those three type of frameworks, they are the most popular today on top of us in terms of percentage. Then that's a stack view.

From the sector view, we have many – Internet sector, like ecommerce, like financial service and like telecommunication. There are so many uses in those four sectors. All different topic is use cases. For example, like some ATM, like a refill routing, or like that's machine learning model thing. For example, like training the model for algorithm training. For example, like learning training a model for product recommendation, for ecommerce, all that all those use cases we've seen a lot is from the four sectors perspective. That's what the most popular today.

From the, let's say, framework level stack view, industry view, sector view as well as the last one I want to mention, is an architecture view. A very popular architecture today are like hybrid cloud, like I mentioned. You have the application, burst application in a cloud running with Alluxio. Then you have data on-premise. We have single cloud, which will speed up the performance so that, for example, we have many users running like Spark, Presto, in Amazon,

in Google, and top of Alluxio, on top of those clouds, like object storage. That's the second type of architecture, very popular.

The other type is an on-premise one. So we call it satellite-like architecture. You have big storage systems somewhere and you build several small satellite class compute clusters using Alluxio and maybe Presto, Tensorflow or Spark on top of us. Those are some very, very popular, very common use cases, simple use cases.

Of course, there are like more advanced use cases as well, and we have some users already running us as a really central data platform system for their whole organization. For example, there is a big bank. They are running us for all their data applications. So there are data applications running on top of us. All different type of frameworks in different data centers. Then they have like storage systems being plugged or mounted into us, like HDFS, like IBM Cleversafe, like Amazon S3, all these being plugged into us. They already achieved that [inaudible 00:25:13] realize this vision data orchestration. That's very exciting.

[00:25:17] JM: Data orchestration for a bank. Banks are not new companies. They are old companies. How do you integrate with a bank? That sounds long and somewhat difficult.

[00:25:30] HL: That's difficult. We have these type of users, customers actually, and has been with us for two years or more than two years. The first production deployment has been more than a year. Then after that –

[00:25:44] JM: It took a year to finish.

[00:25:46] HL: They have been in the staging cluster. They want to make sure absolutely nothing goes wrong. Two parallel systems running all the time for either 6 months or 9 months and then make sure this system is stable.

[00:26:03] JM: They switched it on.

[00:26:05] HL: It's not down. So then they started to put more and more workloads on to this and also develop more and more applications on to this as well as plug more storage into this. I

talked with these users, customers all the time. When I see this type of progress, I'm super impressed and very exciting as well.

[00:26:24] JM: It's scary. It's scary to do something like that.

[00:26:26] HL: In fact, so many users, like we have maybe 60, 70 public uses cases from 60, 70 huge companies.

[00:26:37] JM: Wow!

[00:26:38] HL: Maybe 50 of them are fortune 500 and the rest is not that big yet. So these public cases or website, is it a power by page? That's very exciting. Also towards the end of the year, November 7th, we're also hosting the first Data Orchestration Summit at Computer History Museum in the Bay Area. Lots of use cases will be presented there as well. I look forward to that. Very exciting.

[00:27:08] JM: Yeah. I saw Maxine Beauchemin speaking at Data Orchestration Summit.

[00:27:12] HL: Yeah. He and I are friends, and he's an awesome guys. He invented several very popular open source software.

[00:27:19] JM: I know. Unbelievable.

[00:27:20] HL: Unbelievable. He will be at an open source creator panel at the summit. It will be very interesting panel. So hopefully people will find it useful. The goal of the panel is to – We invited several different type of open source creators with different type of backgrounds. All different stories, how they started their open source projects, and the goal of the panel is to share and discuss all those experience so that we hope anyone, any engineers who are thinking of starting their own open source projects, they probably can learn something from those experience, whatever these people have done wrong or have done right will have to share those things in that panel.

[00:28:02] JM: You have gotten some of the biggest Chinese companies to integrate with Alluxio, which is amazing. One thing I've heard about the Chinese infrastructure companies is that because they're a little bit younger, they have fewer layers of this legacy storage. So it's sometimes easier to integrate with them, and the integrations can move a little bit faster. Is that your experience too?

[00:28:28] HL: That's part of that, but the other part is that – For those companies, they have more data. The reason is that because they have more data – I mean, we all know the reason. Because they have more data, their scale is bigger. When their scale is bigger, they got more value from a system like Alluxio. We have many deployments today, like single Alluxio deployment, more than a thousand servers, single. That's very exciting for us, and we see our technology being battle tested in those type of production environment. They have a lot of data that you can see value more and give you example.

We have telco users, customers all over the world, like China Unicom is using our technology. They're one of the three telcos there. They have 320 million subscribers.

[00:29:30] JM: That's a lot. Isn't that as many people in the United States?

[00:29:34] HL: Probably. They have 320 million users, subscribers in their network, and use our system as one of the core systems to process those data.

[00:29:46] JM: Wow! To like phone calls?

[00:29:49] HL: Billing, like learning, all those type of things.

[00:29:51] JM: Billing, machine learning. Wow!

[00:29:54] HL: Yeah, that scale is really huge. That also impose more technical challenges as well. Our system are being built, being designed from day one as a scaled our architecture and we're just so happy to see people are using us at that level of scale.

[00:30:13] JM: It also sounds like you're talking about – You mentioned this a couple of times, kind of this idea that you decouple the platform engineer from the application engineer.

[00:30:23] HL: Yes. Essentially, because they are data siloes in the end of the day. Traditionally, people try to solve the data silo issue by creating another storage, removing this silo. It never worked. It never worked. We have history to show this. Then we're creating this new layer, insert it into the middle between applications and storage, and storage deployments called data orchestration layer, data orchestration system. What we really enable is that to logically separate the application layer and the storage layer.

When this separation happens, so this essentially make two sides job easier. They can worry less about how to interact with more systems on either side, and hopefully that will help the ecosystem innovate faster. I'll give you an example. Similarly, in history, what has been done like this?

For internet, for example. For the internet, you have seven layers portal. Seven layers protocols. In the middle there's an IP layer as the narrowest. As long as whatever beneath an IP layer can talk to the IP layer based on the protocol, as long as upper layer can talk to the IP layer, so they can communicate. It doesn't matter anymore.

Lower layer doesn't need to worry about what's the innovation or changes in the upper layer. As long as in the middle, this narrowest continues to work. That's the beauty of it. That's more from the whole ecosystem stack view, which we really encourage platform distributed systems engineers to join us, either help us to contribute to our distributed system, this Alluxio open source data orchestration system or help us to build the ecosystem by integrating more and more like upper layer, like data-driven applications with us. Make integration better and more efficient, or like integrate us with more storage systems beneath us.

That's one that'll be very exciting to see the community to grow and more people to join the community to together realize the goal. If we can realize it, it'd be a very cool and interesting ecosystem future.

[SPONSOR MESSAGE]

[00:33:03] JM: When you listen to Spotify, or read the New York Times, or order lunch on Grubhub, you get a pretty fantastic online experience, but that's not an easy thing to pull off, because behind-the- scenes, these businesses have to handle millions of visitors. They have to update their inventory or the latest news in an instant and ward off the many scary security threats of the Internet. So how do they do it? They use Fastly.

Fastly is an edge cloud platform that powers today's best brands so that their websites and apps are faster, safer and way more scalable. Whether you need to stream live events, handle Black Friday, or simply provide a safe, reliable experience, Fastly can help. Take it for a spin tried for free by visiting fastly.com/sedaily.

Everybody needs a cloud platform to help you scale your company. Everybody needs a CDN. Check it out by visiting fastly.com/sedaily.

[INTERVIEW CONTINUED]

[00:34:17] JM: How do you think this will impact consumer file systems? Because I use Dropbox pretty aggressively, and Dropbox has this pretty cool thing where they can kind of give you basically a virtual file system. It gives you a window into everything you have in the cloud, but in order to sync it, you have to sync it. You have to download it from the internet. I've got big WAV files I deal with. We're recording a WAV file right now. I'm going to keep most of these in the cloud, and then when I download it, it's going to take me a while. Is that going to get significantly accelerated anytime soon?

[00:34:54] HL: That's a great question, and at the moment we're not focusing any workloads like that. We're mainly focusing like, say, if the developer, the Spark developer or the large scale data analytics developer, Spark, Presto, Hive, or machine learning, very large scale data-driven workloads developers. We're targeting these type of workloads today.

From the storage integration perspective, and we're integrating large scale storage systems as well. All these in a company type of scale and people typically categorize us into the infrastructure software. Not like Dropbox, a consumer use that. We don't have a direct impact to

these type of a consumer like interaction with applications today. Even though our users, they build up the systems, analytics, machine learning to better serve the other consumers. That's their job.

At the same time, we're being asked by the community that people will say, "Can you mount, besides mounting like HDFS, S3, like NFS, etc., into Alluxio. Can you also mount Google Drive into Alluxio?" They want to do learning. So using data inside Google Drive. People don't care. In the end of the day, from the application developer perspective, they shouldn't care about where the data is stores and how the data is stores.

[00:36:24] JM: Wait. Who's asking you to mount Google? Google Drive is not –

[00:36:28] HL: We have machine learning product and we already have done using mounting all different public cloud storage as well as typical HDFS, NFS, etc. Their users use their platform to do machine learning. Their users may have data from other places as well and they just want to plug those data in. That's fundamentally why they want a system like this.

[00:36:56] JM: I see. So they want to weigh for Google Drive interesting to be synced in to memory.

[00:37:04] HL: First of all, they don't care about the performance, number one. They care about accessibility. Because they are users –

[00:37:12] JM: This is exactly the use case I'm asking you for with Dropbox.

[00:37:15] HL: Yeah. Essentially, they have a machine learning application. Their user's data could be stores in NFS. It could be stores in S3. It could be stored in Google Drive. The machine learning application, or machine learning application developer doesn't want to care about it. They just want to access the data and then process it.

[00:37:39] JM: So when they're iterating through all the files in the file system, in the Alluxio point of view virtual file system, if those files can be in memory, it's going to be much faster to

access and process them. But if you got to go into the underlying storage layer, like I guess you would have – I guess in the Google Drive case, you have to go across the network.

[00:38:03] HL: Yeah.

[00:38:04] JM: And you should have to make individual network calls and then grab it from however Google Drive is storing it and pull those into memory before you're going to learn on them.

[00:38:11] HL: That's the integration part that our system need to do with Google Drive. Similarly, we can do other type of integration with other type of [inaudible 00:38:19].

[00:38:20] JM: That's just an API integration.

[00:38:21] HL: Exactly. I mean, API or driver, however you call it. Yes.

[00:38:26] JM: I assume the same would almost be true – I mean, people probably ask for the same thing with like Google Analytics, Stripe, QuickBooks. Wouldn't all these APIs, they're like, "Can you easily pull this into memory for –"Is that not your responsibility? Would that be something else? I mean, because the Google Drive, the interface between Google Drive and the Alluxio file system, that's basically networking API integrations that you need to write. Is the same true for something like if you want to get their data from Stripe and pull that into memory and always keep that in memory? I guess that's a different question.

[00:39:06] HL: Again, when people talk more about a Google Drive type of integration, there are less concerns, a bit of concern about the performance. But the number one issue they want to solve is accessibility. Essentially, they already build a machine learning or whatever data-driven application. That data-driven application will use one type of interface to interact with the data. A lot of machine learning type of applications, they use file system interface. Then it's not how they can talk to Google Drive.

Then if there's no system like us, like this data orchestration system like Alluxio, essentially, what can they do? One way they can do is somehow write another program to migrate the data

from Google Drive to some place. Some other storage they can to, their API can talk to. That's essentially created another data silo and another data copy. This migrating process is error prone and also is very time consuming process as well. That's one way.

The other way is that they need to modify their application to talk to another interface to interact with Google Drive. That's also time consuming. You need to change your application. Test your application. Somehow make your application production ready again. That's a long cycle. Depends on a company and depends on the person. That's never an easy thing to do. Because of this, a way the system like this essentially dramatically improve the productivity of these application developers they just don't need to carry anymore for how their application access data from different places and where the data is stores.

[00:41:02] JM: I think the reason I was thinking about that is I did a show recently with the company Fivetran, the data connector company. I talked to them recently. I was just thinking about ETL, but you're not doing ETL. You're an interface.

[00:41:17] HL: We're a system, and a system provide all the applications fast data access to data store anywhere. We don't care. They don't need to care anymore. There are people building, build like ETL type of applications on top of us. That's where the relationship is. Essentially, I've seen many companies running ETL workloads we have written on top of, like running those type of workloads on top of us. Yeah.

[00:41:49] JM: That would just be transformed from one system to another. If it's all built on – If all of those ETL jobs are just between two different systems that are on Alluxio, they're probably going to be a lot faster than one system that's on Alluxio and one that's like has a lot of disk usage.

[00:42:09] HL: Faster as well as more future prone, like easier to write. For example, from that application perspective, if they're interacting with Alluxio, maybe behind-the-scene is one is S3, one is HDFS behind-the-scene. Two different folders. From the ETL job perspective, it's just talking to Alluxio. It doesn't know what's behind-the-scene. It doesn't need to know.

Then the same application today run between HDFS and S3. Tomorrow can run between any other storage as well, for example, like Cleversafe and GCP. So that's the beauty of it as well.

[00:42:46] JM: Cool. Tell me about your plans for the future and what engineering problems you're working on right now. How do you want the company to look in five years?

[00:42:56] HL: Five years is a long-time.

[00:42:57] JM: Okay. Two years. One year.

[00:42:59] HL: I mean, say five years, I would hope we achieve the goal many, many companies like in the world have used this data orchestration system or make it the data platform in their organization, like the banking example I talked about earlier. Hopefully many of these companies have achieved that. That's five years.

But from a two-year perspective, we already have many, many use cases from these four sectors and will continuously to innovate the system. We're making the system more performance. Many, many people working on the performance side, or making the system much easier to use to help the adaption to reach more and more people in the ecosystem. We are developing features to orchestrate the data more. For example, how to effectively store the data in different storage systems to make the cold data in a cheaper storage and hot data in a more expensive storage.

Also, this is a new layer and there are so many different type of features. Our users are asking like the example of plugging, like mounting like Google Drive into us as well to connect us more with even bigger ecosystem. That's all the things we're trying to do and just more users, more impact and more people in our community and get value from the system from the community as well as contribute into the community. It'd be more fun.

[00:44:33] JM: Can you imagine any cloud service that you would offer? Because today, it's kind of – You get these large customers, you figure out how to do the integration. It can be a long cycle time, but then you have a great customer for a very long period of time. Can you imagine a self-serve version of this?

[00:44:52] HL: Yeah, definitely. There are several things here, several things. One is that there are certain cloud-based product offerings already embedded us and embedded our software technology. Some is visible, some is not.

[00:45:08] JM: They're offering your [inaudible 00:45:10]. I'm sorry!

[00:45:12] HL: No worries. They offer a product and in a product there's the component, like important component is Alluxio. We are working closely with those vendors in the world. That's exciting. That's one thing. The other thing is that if you go to this website today, Alluxio.io, go to the Try Alluxio page and you can see we have the tutorial to deploy battle tested version of Alluxio in the cloud. But it's not the hosted service, but it's the easy deployment for the cloud workloads. The reason is that we have seen many, many – Our users already running us in the cloud. We essentially DevOp this to make their life much easier, to more potential users life much easier to adapt this technology software in the cloud. That's the second way. The third way of the – Will be a manage of service from us, and we don't have that today.

[00:46:11] JM: To wrap up, what's the hardest part about managing a company?

[00:46:16] HL: I'll say I think the hard part will be at different time, there are different challenge and different issues. That's the hard part. It's never the same. The first 90-day is a different challenge. The second 90-day is another different challenge. The third 90-day is another different challenge. It's always different. Also, what I've learned is that it's fun to see the challenges. If you work on it, it will be solved. That's the challenge and at least the higher phase challenge, these type of challenges.

[00:46:51] JM: HY, thanks for coming back on the show.

[00:46:53] HL: Thank you, Jeff. I appreciate that.

[END OF INTERVIEW]

[00:47:03] JM: If you want to extract value from your data, it can be difficult especially for nontechnical, non-analyst users. As software builders, you have this unique opportunity to unlock the value of your data to users through your product or your service.

Jaspersoft offers embeddable reports, dashboards and data visualizations that developers love. Give your users intuitive access to data in the ideal place for them to take action within your application. To check out a sample application with embedded analytics, go to softwareengineeringdaily.com/jaspersoft. You can find out how easy it is to embed reporting and analytics into your application. Jaspersoft is great for admin dashboards or for helping your customers make data-driven decisions within your product, because it's not just your company that wants analytics. It's also your customers.

In an upcoming episode of Software Engineering Daily, we will talk to TIBCO about visualizing data inside apps based on modern frontend libraries like React, Angular, and VueJS. In the meantime, check out Jaspersoft for yourself at softwareengineering.com/jaspersoft.

Thanks to TIBCO for being a sponsor of Software Engineering Daily.

[END]