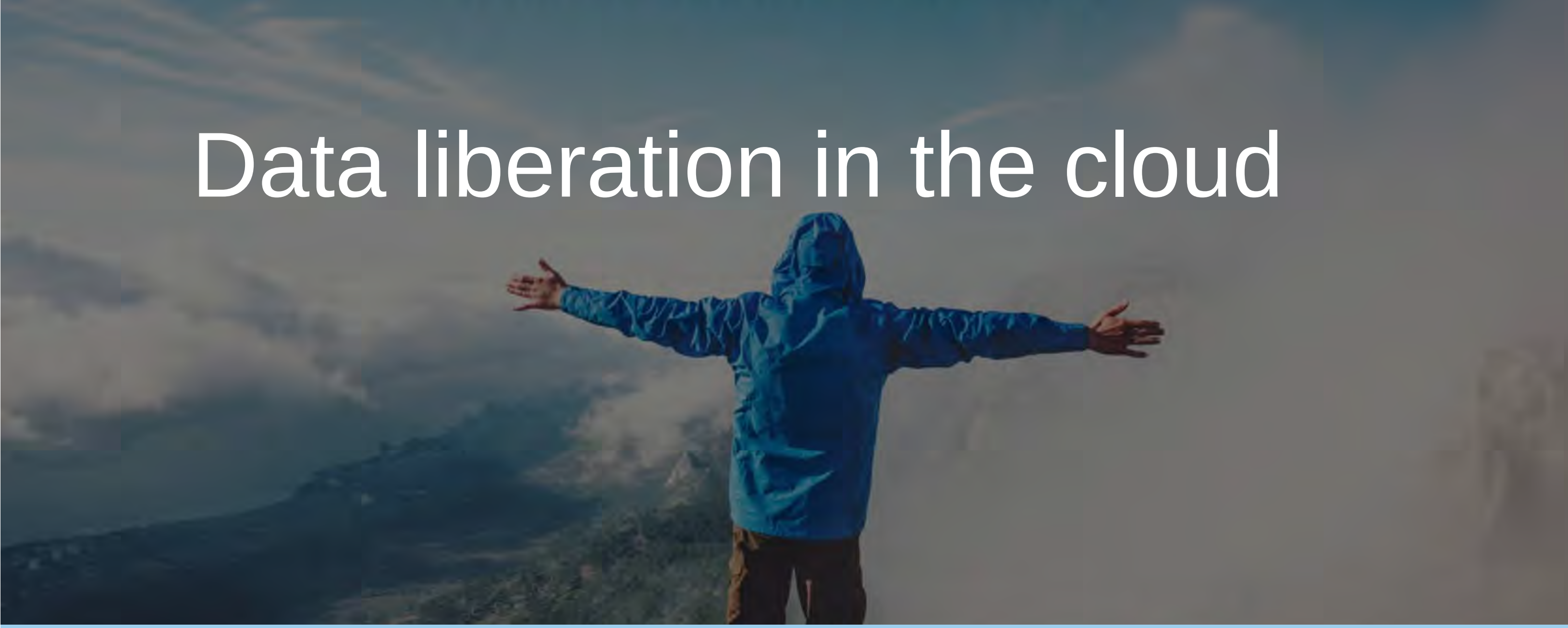
Data liberation in the cloud



SNIA CSTI | CLOUD STORAGE TECHNOLOGIES

Moving data from one system to another can be challenging. Ensuring the data arrives intact and usable is paramount.

Eric Lakin, Storage Engineering & Data Center Engineering Manager at the University of Michigan can envision a time when the university has five to 10 petabytes of data stored in a cloud service. “If we were to decide down the road for some business reason that we wanted to stop working with that cloud service and begin working with a different cloud service,” he says, “there really isn't a mechanism in place for moving that much data from one cloud provider to another without breaking the connections that that data may have with on-premises equipment.”

More specifically, Lakin’s team moves data that hasn’t been utilized in 180 days out to Amazon S3 and S3 Gov. The on-premises storage system, tracks the location of the data. “It uses what we used to call an HSM model, a hierarchical storage management model, to move that data out, so a stub file remains behind,” says Lakin.

Even if they were able to magically move a hypothetical five petabytes of data over to Azure or GCP, their storage system wouldn’t know what was happening. “When we were investigating solutions, we were not aware of tools in place for moving large quantities of data between cloud providers while maintaining whatever symbolic links we have back to our own organizations,” he says.

Even if that were somehow taken care of, using cloud storage natively with some kind of gateway, the same problem would exist, says Lakin. “Once you've made a significant commitment to putting a large amount of data in any one cloud provider, at some point you're limited in your ability to just seamlessly transition it over to a different provider,” he says.

There would be a large capital investment, as well, to bring all the data back to on-premises storage, as well or quite honestly, to bring back large quantities back on premises. If your entire storage strategy were to change, you would have to procure a pretty large amount of on-premises storage. It would be a large capital investment to bring all that data back. “What it's really about is having the flexibility and the option to change your mind down the road with regard to who you're doing business with,” says Lakin.

Mike Jochimsen, Director of Alliances at Kaminario, agrees. “Data liberation is the ability to easily and seamlessly move data between different cloud environments without restrictions due to data formats or on-ramp and off-ramp limitations to and from clouds.”

Jochimsen believes data liberation is both a theoretical and an explicit practice that should be defined and followed. The benefit for the end user is a world in which they have access to their data no matter where it is, and storage consumers can manage the actual location of the data to manage their own budgetary, business needs. “So the end customer benefits because it's transparent to them, they don't care where the data resides at any given point in time, all they care about is the data is accessible to them,” he says.

The hybrid, multi-cloud platform is in use by most companies today, explains IBM’s Program Director and Global Offering Manager for Hybrid Cloud Storage, Michelle Tidwell. “Most companies are using an average of five private and public clouds today in their enterprises and that's just going to increase,” she says.

Clients are looking to have the kind of flexibility that data liberation could provide to help transform their businesses to take advantage of the multi-cloud environments they’re already using. “It's about moving the data and having mobility around the data,” she says. “It's also about being able to manage that data seamlessly between on-premises and cloud environments such that resources can be monitored and adjusted depending on cost and performance. That’s something we’re really focused on. Clients also have made significant investments to operationalize their environments on-premises, and being able to reuse that with public cloud infrastructures could be a big savings in OpEx”

Orchestration, too, plays a critical role in allowing cloud-native applications like Kubernetes to access this data no matter where it’s hosted. “You define the composition of an environment through orchestration layers, the automation utilities then build the environment,” adds Jochimsen. “You need to have that extensibility out to these cloud environments to seamlessly migrate apps and data across cloud layers.”

Data liberation in today’s market is mostly being talked about as how to extend current data center storage setup to take advantage of public cloud infrastructure, says Tidwell. “In that context, the cloud is used more as remote storage and archive, sometimes targeting things like S3,” she says. “However with the latest software defined storage and hybrid cloud data management capabilities, it can also be used to setup a disaster recovery capability on the public cloud or just have a more seamless workflow migration back and forth.”

Depending on the vendor, the type of solution needed, and the end use case, the techniques involved can be anything from on-premises storage-level to network-level data replication.

“I think the way it's being done today between these various proprietary clouds is through API layers,” says Jochimsen. “And so each of the proprietary cloud vendors has its own API for accessing via applications that may span different layers,” he says. “So while you can build it today through APIs, there's no standardization at that API layer.”

There’s definitely room for some standardization here, as each vendor currently must build their own integrations via programming interfaces.

There are other concerns that need to be addressed, as well. “If you consider Europe and GDPR, there are concerns there,” says Tidwell. “(Users) would probably like the idea of flexibility but yet sometimes they're restricted about moving things across different countries.”

Another technical challenge to data liberation is networking. “That can be a very challenging problem, so we need technologies and standards around software-defined networking to help ease the physicality of setting up network connections,” says Tidwell.

SNIA, a non-profit global organization dedicated to developing standards and education programs to advance storage and information technology, has already begun the conversations needed to explore and surmount these challenges, with the development of CDMI, the Cloud Data Management Interface standard. CDMI is an ISO/IEC standard that allows disparate clouds to talk a common language and eases the data movement and migration between different clouds.

Jochimsen believes that SNIA will be instrumental in helping all the players that need to be involved get together and figure out the standards that need to be in place to make data liberation an actual process, and one of them will be CDMI. “Without standards it falls back to the end consumer to build their own or multiple vendors to build very expensive solutions,” he says. “SNIA has brought us all together and provided the forum for standardizing all of those touch points that allow us to do it more quickly, more cost efficiently and save the end consumer a lot of money hopefully.”

University of Michigan’s Lakin feels similarly. “I think SNIA's role has helped us identify data liberation as something that would definitely benefit from a solution, whether it's a third party provider or some other organization that can solve this problem,” he says. “And with CDMI, the technology issues can be more easily addressed. We're trying to adopt cloud services at scale and uncover problems that maybe one organization has either experienced or has identified as a future risk, and then everybody else becomes aware of it as well.”

The conversation then helps with decision-making and strategic planning, but that’s not all SNIA can do in fostering this sort of collaboration. “At some point,” Lakin continues, “we can also use our cooperative efforts to put pressure back on not only our storage vendors, but also cloud vendors to say, ‘Look, this is a business need and we need to have this kind of flexibility.’ So hopefully there's strength in numbers, and we can insist on adoption of CDMI to help us with our data liberation programs.”

About the SNIA Cloud Storage Technologies Initiative

The SNIA Cloud Storage Technologies Initiative (CSTI) is committed to the adoption, growth and standardization of storage in cloud infrastructures, including its data services, orchestration and management, and the promotion of portability of data in multi-cloud environments. To learn more about the CSTI’s activities and how you can join, visit snia.org/cloud.