

SOLUTION BRIEF

Geo-distributed Analytics with NetApp StorageGRID and Alluxio

Simplify and accelerate
your geo-distributed analytics
platform at scale



The challenge of architecting a data platform across a heterogeneous environment

Today, many organizations run a multitude of data-driven applications and data platforms that span geographic regions and heterogeneous environments—public, private, hybrid, and multicloud. Further, the trend of separating compute resources from storage resources makes it easier to scale compute and storage independently, allowing organizations to keep up with trends in data analytics and AI. In response, more organizations are modernizing their data platforms to meet their needs.

However, data storage and access across mixed environments have always been challenging. How can you architect your data platform to access data across regions and data centers efficiently? How can you optimize cost, performance, security, and agility at the same time? And how do you scale the infrastructure nondisruptively? These questions have to be answered on your modernization journey.

Many organizations are adopting multicloud and hybrid infrastructure strategies. NetApp has partnered with Alluxio to help you implement these strategies faster and optimize your spending. With our joint solution, you'll gain a modern data infrastructure that's ready for enormous scale.

Alluxio data orchestration

Alluxio is an open-source data orchestration platform for analytics and machine learning (ML) applications. It's a software solution that sits between disaggregated compute and storage to connect data analytics applications to a mixture of data sources and bring data close to compute.

Key benefits

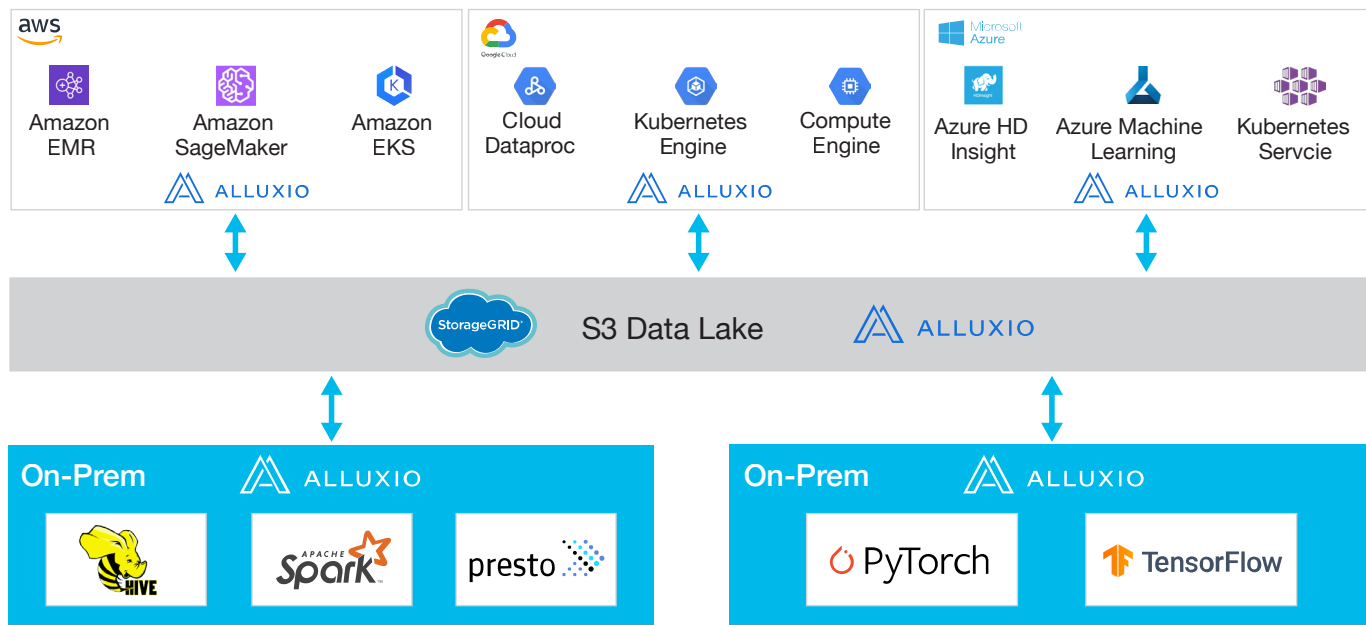
- Provide data-driven applications with unified access to storage across geographic regions, in hybrid and multicloud environments.
- Use an enterprise-grade, cloud-integrated, and scalable software-defined solution with security and disaster recovery.
- Boost performance, reduce storage cost, and avoid cloud vendor lock-in.
- Gain the agility you need to meet evolving requirements.

To deliver data to a wide variety of data-driven applications with agility and simplicity, Alluxio provides the following features:

- **Unified namespace to break down data silos.** Alluxio provides a unified view of data that's spread across regions for agility across applications.
- **Transparent data tiering.** Alluxio manages hot and cold data across expensive and cheaper storage tiers. Tiering is transparent to applications and uses a single data management interface.
- **Data caching across regions.** Alluxio provides distributed caching that enables higher read/write throughput with data locality.

StorageGRID object storage

NetApp® StorageGRID® is an enterprise-grade, on-premises object storage solution that supports the native AWS Simple Storage Service (S3) API. StorageGRID is software defined, which means that you can run it on different platforms—bare metal, VMware-based environments, or NetApp purpose-built appliances—and mix platforms in a grid. In a single namespace, StorageGRID can scale up to 16 data centers worldwide. StorageGRID offers massive S3 object storage and dynamic data management, enabling you to run next-generation workflows on the premises, alongside your public cloud. The solution's industry-leading data management policy engine helps you optimize levels of performance and durability and adhere to data locality requirements.



How Alluxio and NetApp work together

StorageGRID and Alluxio give you the agility to employ cloud innovation for analytics wherever your data resides—and to avoid performance issues, high costs, and complex migration processes. The combination of Alluxio and StorageGRID provides the flexibility to build hybrid cloud environments anywhere.

By using StorageGRID as your object store and Alluxio as your data orchestration platform, you'll keep up with evolving requirements, avoid data silos across geographic regions, and optimize storage costs while maximizing performance.

The following three use cases demonstrate the solution's benefits.

Use case 1: Migration from HDFS to object storage

To move away from legacy infrastructure, more companies are offloading their data analytics and AI/ML workloads on Hadoop Distributed File System (HDFS). Object storage is a popular choice due to its low cost, scalability, and ease of management, especially for workloads such as Spark and Presto.

With StorageGRID and Alluxio, you can seamlessly migrate data from HDFS to object storage. Alluxio mounts HDFS and object storage into a unified namespace to facilitate migration. By using a single

point of access, data can be migrated incrementally according to defined policy without interrupting the application side. As a result, you can now truly decouple compute and storage and scale them independently.

After your data is migrated to StorageGRID, you can create different tiers and retire the less frequently used data to a more cost-effective archive tier to reduce costs.

Use case 2: Expose data on premises to compute engines in the cloud

With StorageGRID and Alluxio, compute engines in the cloud can access on-premises data with high performance and a fraction of the cost. Because the solution brings the data locally to the analytics and AI applications, the performance is the same as having the data co-located in the cloud. Also, because Alluxio caches the data on demand, you no longer need to repeatedly fetch data directly from the cloud storage—which reduces many of the network egress costs.

From the storage side, the tiered storage capabilities of StorageGRID and Alluxio give you full control over where your data resides and keep your data reliable and highly available. StorageGRID supports assigning the right grade of data to the right tier. Alluxio's intelligent data tiering manages hot and cold data across expensive and cheaper storage tiers transparently to applications with a single data management interface.

Use case 3: Enable multicloud workloads

With Alluxio and StorageGRID, you gain the flexibility of using cloud services from any vendor—while minimizing data movement. For example, suppose you normally use AWS for most workloads but would like to run a specific service in Azure. By using Alluxio, you can run workloads in both AWS and Azure as needed, without moving data from one cloud to another. You have complete control over the placement of your data, which is especially useful when you need to switch vendors.

Alluxio serves as the data orchestration layer that connects any compute framework and any underlying storage. Together, Alluxio and StorageGRID let you take advantage of all the services that are available, in both public and private cloud, across different vendors. You can continue to adapt as your requirements change, without worrying about vendor lock-in.

[Learn more](#) about this joint solution.

About Alluxio

Proven at global web scale in production for modern data services, [Alluxio](#) is the developer of open source data orchestration software for the cloud. Alluxio moves data closer to data analytics and machine learning compute frameworks in any cloud across clusters, regions, clouds and countries, providing memory-speed data access to files and objects. Intelligent data tiering and data management deliver consistent high performance to customers in financial services, high tech, retail and telecommunications. Alluxio is in production use today at [eight out of the top ten](#) internet companies. Venture-backed by Andreessen Horowitz, Seven Seas Partners, Volcanics Ventures, and Hillhouse Capital. Alluxio was founded at UC Berkeley's AMPLab by the creators of the Tachyon open source project.

About NetApp

In a world full of generalists, NetApp is a specialist. We're focused on one thing, helping your business get the most out of your data. NetApp brings the enterprise-grade data services you rely on into the cloud, and the simple flexibility of cloud into the data center. Our industry-leading solutions work across diverse customer environments and the world's biggest public clouds.

As a cloud-led, data-centric software company, only NetApp can help build your unique data fabric, simplify and connect your cloud, and securely deliver the right data, services, and applications to the right people—anytime, anywhere. www.netapp.com

