Achieving Hybrid and Multi-Cloud Architecture With Application Portability

Fortune 50 Technology Company

What's Inside

A Fortune 50 technology company has successfully implemented Alluxio to achieve a hybrid-cloud strategy, become multi-cloud ready, cut costs, and boost agility.

CASE STUDY

1 / Overview

This Fortune 50 company serves over 1 billion users worldwide and ingests and retains petabytes of data in both on-premises and cloud managed by the data platform team. On the data platform, multiple data science teams in different domains perform large-scale analytics and machine learning jobs by using applications built on open-source compute frameworks such as Spark and Trino.

Modernizing the data platform with agility and onboarding more teams is critical to this tech giant. A future-proof data infrastructure will accelerate the time-to-insights to enhance customer experience, boost operational efficiency, and more.

This case study details why the company chose Alluxio, the architecture of hybrid and multi-cloud, and how Alluxio helps the company standardize the data stack and access data anywhere across all environments.

2 / Highlights

About the Customer

- Industry: Technology
- Topology: Hybrid and multi-cloud
  - Application: both on-premises and in AWS
  - Data: both on-prem and in AWS
- Data stack
  - Compute: Spark, Trino, Hive
  - Storage: AWS S3, on-prem HDFS & object stores

Key Benefits with Alluxio

- Business Value
  - Gain agility for hybrid and multi-cloud
  - Reduce S3 egress costs
  - Shorten time-to-insights
- Technical Value
  - Access data anywhere with zero application reprogramming
  - Standardize on a common data stack
  - Achieve a future-proof architecture

3 / Background: The Journey to the Cloud

The data platform is key to the company’s ability to deliver innovative solutions for customers across the globe; it integrates data from multiple operational applications and supports analytics and machine learning applications.

In the past, both the compute and storage resources of the company’s data platform reside in many private data centers, serving multiple different teams.

As public cloud services bring significant benefits, the company is moving from all-on-premise to hybrid-cloud infrastructure, which consists of compute on-premises and data both on-premises and in AWS.

The data platform team has been growing and populating data in AWS S3 and stopped the data growth in HDFS while still utilizing on-premises Hadoop resources.

4 / Challenge: Unable to Bridge On-premises and Public Cloud

Prior to implementing Alluxio, the company could not bridge on-premises and cloud data access. On-premises Spark and Trino applications could not directly run on S3 data since the applications were only capable of running with HDFS APIs.

The data platform team had to use a copy method to make data available to the on-premises applications. When the applications needed to access data in the S3 bucket, the data platform team replicated the S3 data to HDFS and made the corresponding data available until Trino and Spark applications could perform analytics and machine learning jobs.

The company encountered the following challenges.

Huge S3 Egress Costs

Data scientists constantly need to retrieve data in S3 buckets, resulting in high egress costs. This significantly increased the long-term TCO of the data platform.

Poor End-User Experience and Long Time to Insight

Manually copying data means that data is not immediately available. This can lead to poor user experiences and complaints since the data is delayed by hours or days. The unavailability of data significantly slowed the time to insight.

Cloud Journey is Hindered by Lack of Application Portability

When data resides in both HDFS and S3, applications must be reprogrammed for data access unless data is copied over. This hinders the adoption of hybrid and multi-cloud, because applications are not portable between on-premises and the cloud.

5 / Motivation: Agility Brought by Hybrid and Multi-Cloud

In seeking a long-term solution to the data platform, the tech giant wants the agility to deploy their applications and compute capacity to any environment based on cost and operational overhead.

From a strategic viewpoint, hybrid and multi-cloud bring together private data centers and more than one public cloud provider to host the data platform in a scalable and agile way, and prevent vendor lock-in at the same time.

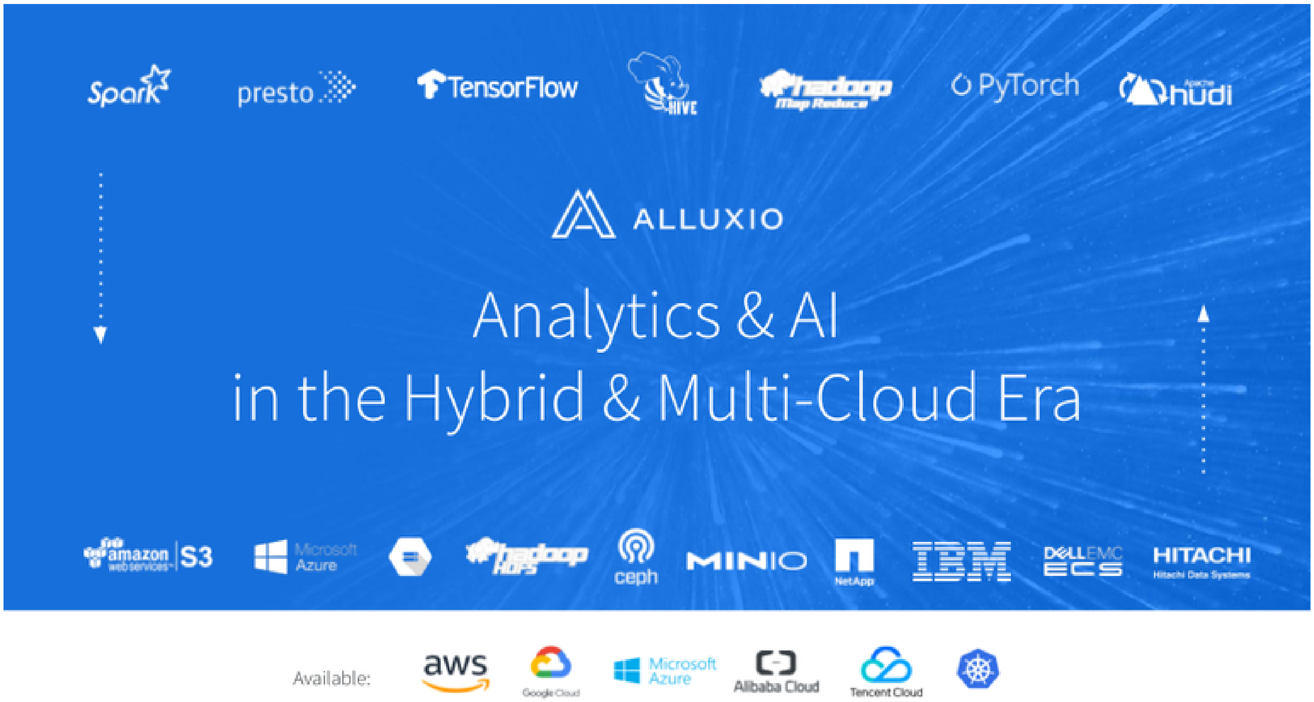
While planning and adopting hybrid and multi-cloud, designing a data architecture with application portability is the key. During migration, the data platform should continue serving data science teams with minimum impact on their applications.

As the company was looking for a solution to achieve hybrid and multi-cloud architecture, it turned to Alluxio.

6 / Solution: Achieving Hybrid and Multi-Cloud with Application Portability

Alluxio is the New Data Layer for Large-scale Analytics and Machine Learning

Alluxio is a new data layer between storage and compute engines for a variety of data-driven applications, such as large-scale analytics and machine learning.

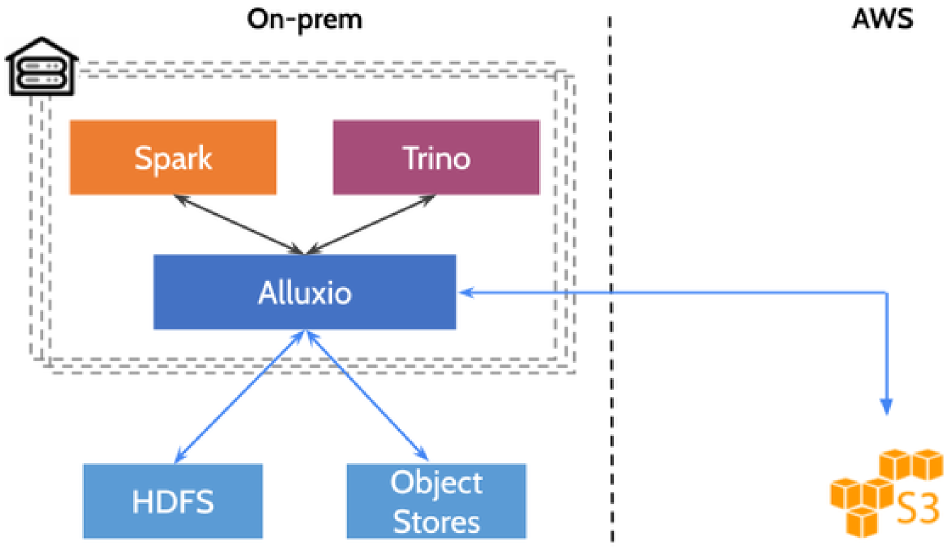


This new data layer provides complete virtualization across all data sources to serve data to applications that do not need to care about the location of data. The solution is applicable across environments, whether in the cloud or on-premises, bare metal or containerized.

7 / Why Alluxio

- Accelerate hybrid and multi-cloud
- Shorten time-to-insights
- Reduce S3 egress cost
- Access data anywhere without app change
- Standardize on a common data stack
- Achieve a future-proof architecture

Architecture: Standalone Alluxio Cluster On-Premises

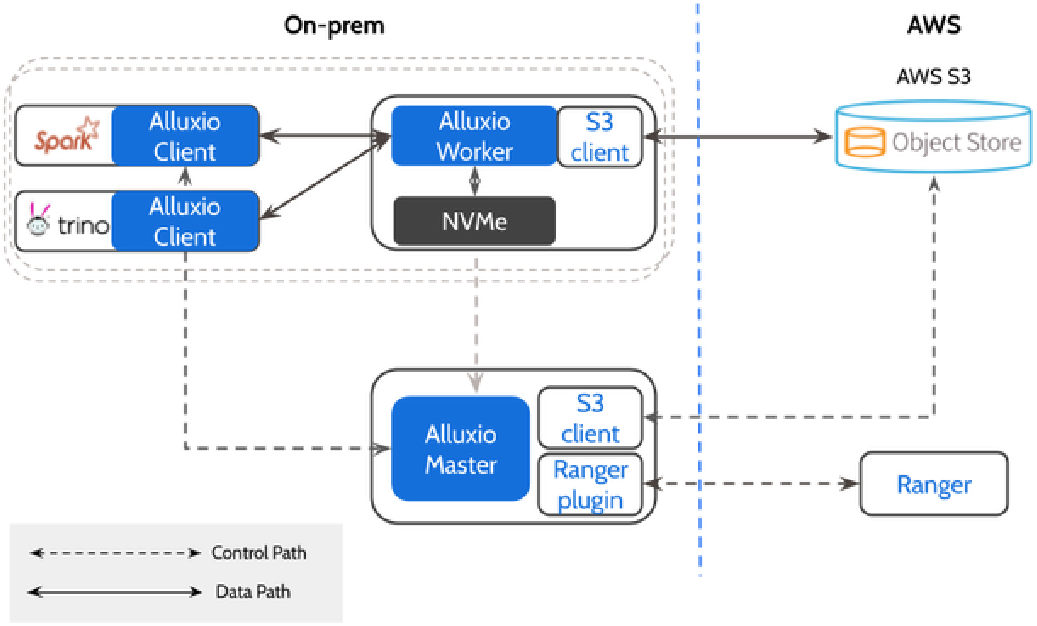


The company started by deploying Alluxio in the following architecture:

- Deployed a standalone Alluxio cluster.
- Alluxio Master is on-premises, which stores metadata for the filesystem namespace.
- Alluxio Workers are on-premises, which store the actual data. Alluxio Workers cache the data, and any subsequent accesses by the compute engines are served from local storage on the worker nodes.
- Mount HDFS and S3. All storage systems, including HDFS and S3, are mounted into Alluxio's namespace for access via Alluxio.
- Ranger resides in AWS for enterprise-grade data security.

7 / Why Alluxio

Product Features Supporting This Architecture



The following features support the above architecture:

- **Alluxio core features**
  - **Caching:** Alluxio workers leverage on-premises NVMe for cache. The first time data is accessed from the mounted storage system, Alluxio Workers cache the data on the worker nodes leveraging NVMe. Within one data center in a specific region, Alluxio serves as a regional cache with multiple Spark instances sharing caching.
  - **Server-Side API Translation:** Alluxio manages communication between applications and file or object storage. It transparently converts from a standard client-side interface to any storage interface. Thus, applications using HDFS API do not need to be reprogrammed while accessing S3 storage.
  - **Unified namespace:** Alluxio serves as a single point of access to multiple independent storage systems regardless of physical location. Both HDFS and S3 are mounted to a common Alluxio namespace, enabling unified access and a standard interface for applications.
  - **Metadata synchronization:** Metadata for the mounting location, such as file size and location, is loaded from remote storage into the Alluxio Master and synchronized seamlessly. Alluxio asynchronously fetches the metadata from both HDFS and S3 to make data consistent.

7 / Why Alluxio

- **Enterprise-grade security (Enterprise Edition Only)**
    - **Ranger plug-in:** Alluxio integrates with Ranger using a Ranger Plugin to support data authorization in AWS. Filesystem permissions and access controls are enforced on data accessed in the cloud along with prevailing user and group authentication.
    - **Others:** TLS (client to server, Alluxio master to worker) and AWS S3 AssumeRole (temporary access tokens that worker requests for master) are also implemented.
- **Catalog migration (Enterprise Edition Only)**
    - **Transparent URI:** Compute frameworks like Trino can connect to an existing Hive catalog on-premises without the need to re-define tables or maintain a replicated instance of Hive's metadata.

8 / Results: Significant Business and Technical Benefits Achieved

Business Values

Accelerate Hybrid and Multi-Cloud

With Alluxio, the company is achieving the hybrid-cloud strategy and moving closer to true multi-cloud. Having Alluxio as a standardized data abstraction layer, the applications have uniform and easy access to all the data. Alluxio ensures cloud migration agility with results on day one. The data platform can now evolve to hybrid and multi-cloud at a pace adaptable to business demands.

Reduce S3 Egress Cost

By implementing Alluxio, there is no need to manually copy data anymore. By providing unified data access, Alluxio presents the data to the applications no matter where the data resides. By caching data, Alluxio helps the company avoid repeatedly fetching data directly from the cloud storage, thus significantly reducing S3 egress cost.

Shorten Time-to-insight

Data is now immediately available to users without the need to wait for the long process of manual copy and validation. Data scientists and other end-users enjoy data availability and consistency, leading to shorter time-to-insights.

Technical Benefits

Access Data Anywhere with Zero Application Reprogramming

Because Alluxio translates data access requests from applications into underlying storage interfaces, Spark and Trino applications continue to use HDFS API with no S3 reprogramming. The company can now scale the data platform and onboard more applications with access to geo-distributed data either on-premises or in the cloud without the need for complex system configuration and management.

Standardize on A Common Data Stack Across Heterogeneous Environments

With the new architecture, Alluxio has enabled a standardized data stack with unified access to data. With this standardization, the application has gained portability, the platform is multi-cloud ready and applications can move seamlessly between on-premises and cloud and multiple cloud vendors.

Achieve a Future-proof Architecture for Compute and Storage Innovations

With Alluxio as the new data layer, the data platform has decoupled data management from elastic compute resources. The architecture is environment agonistic and ready to accommodate the future state for compute and storage technology.

9 / Looking Ahead: The Next-Generation Data Platform

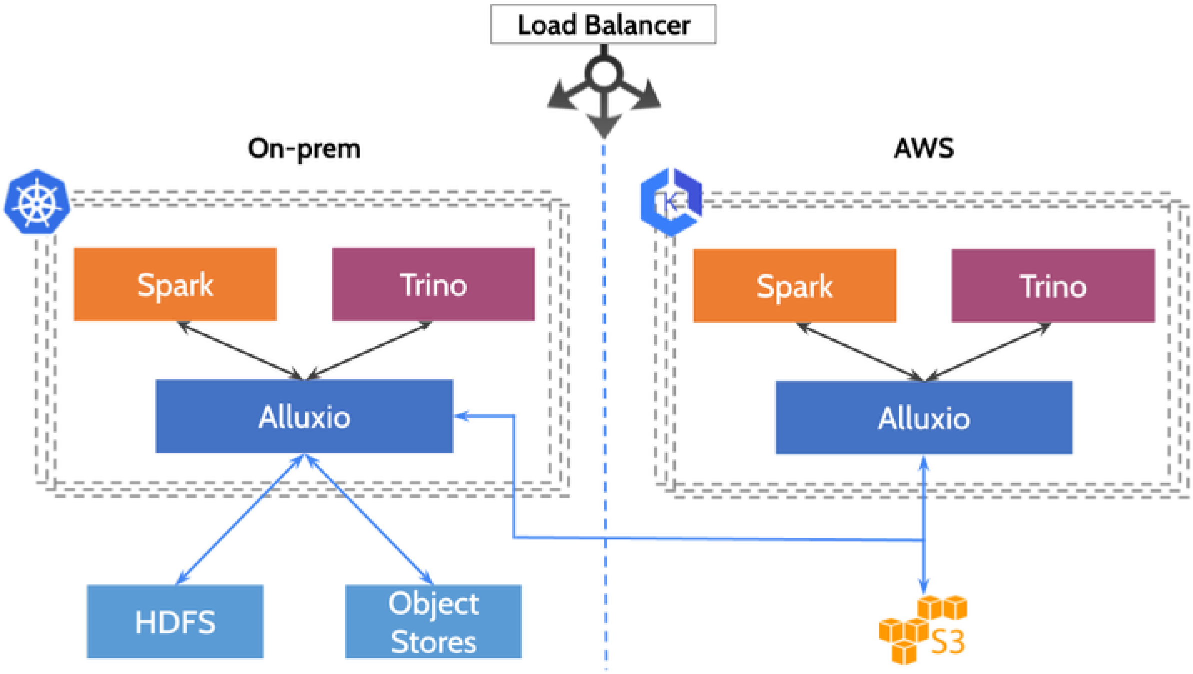
The customer is still early in its modernizing journey but looking forward to the possibilities that lie ahead.

The current Alluxio deployment has allowed the company's data platform team to focus on creating a solid foundation for future development. In the future, Alluxio will be supporting the longer-term goals of the data platform.

When looking ahead, the preferred final state is a balance of reduced operational complexity and cost optimization. Ultimately, the company wants the flexibility to spin up compute on-premises or in the cloud with access to data anywhere.

The next generation of the data platform should meet the three requirements:

- K8s everywhere with a load balancer to route traffic to containers, either on-premises or in the cloud.
- A single S3 API for everything, from Alluxio to HDFS, GCS, and S3. Deploy Alluxio clusters both on-premises and in multiple clouds by different CSPs.
- A single way of enforcing security: Ranger with Alluxio for authorization and accessing.



9 / Looking Ahead: The Next-Generation Data Platform

As the abstraction layer between compute and storage, Alluxio is expected to help the company achieve the above three goals and more teams can be onboarded to the data platform. By containerizing and unifying data access, the data platform will gain the flexibility to distribute compute resources anywhere based on availability and pricing.

10 / Summary

Overall, with Alluxio, this Fortune 50 technology company has achieved a hybrid-cloud strategy and become multi-cloud ready. By having Alluxio as the new data layer, the company is eliminating the hassle of copying data, achieving agility, reducing TCO, and accelerating the time to insights. Alluxio will also pave the path for the company to build the next generation platform in the future.

For more information about Alluxio, go to: https://www.alluxio.io/.