ALLUXIO | expedia group

CUSTOMER CASE STUDY

UNIFY DATA LAKES ACROSS MULTIPLE GEOGRAPHIC REGIONS IN THE CLOUD

Executive Summary: Expedia Group has implemented Alluxio to federate cross-region data lakes in AWS. Alluxio unifies geo-distributed data silos without replication, enabling consistent and high performance with ~50% reduced costs.

COMPANY	Expedia Group (NASDAQ: EXPE)
INDUSTRY	Technology, Information and Internet
LOCATION	Seattle, WA
EMPLOYEES	10k+
TOPOLOGY	Multi-region in AWS
COMPUTE	Spark, Trino, Hive, Databricks, JupyterHub
STORAGE	AWS S3 in Multiple Regions

ALLUXIO BENEFITS

- Unify cross-region data lake access without the need to replicate data repeatedly
- Reduce ~50% S3 egress cost per query
- Provide analytics & AI applications with consistent performance

Expedia Group is an American online travel shopping company with a portfolio of 20+ brands. The data platform team at Expedia builds petabyte-scale data platforms used by different brands. The analytics & AI platforms serve data scientists and ETL developers around the world.

To modernize the data platforms, Expedia has created a central data lake in AWS. However, each brand has historical silos of data lakes in other AWS regions. The data platform team has prohibited users from directly querying data from other regions because of cross-region data transfer costs. On a daily basis, the data platform team replicated petabytes of data to the central data lake.

CHALLENGE: Replicating Data Cross-region was Slow, Costly and Error-prone

- **Long time-to-insights.** It took hours or days for the replicated and validated data sets to be available, resulting in a poor user experience and long time-to-insights.
- **High S3 egress costs.** Because Hive does not support update/merge, the entire table had to be replicated, adding up to cross-region data transfer costs.
- **Error-prone.** Manual data synchronization and validation were error-prone and complex, increasing the data platform team's operational overhead.

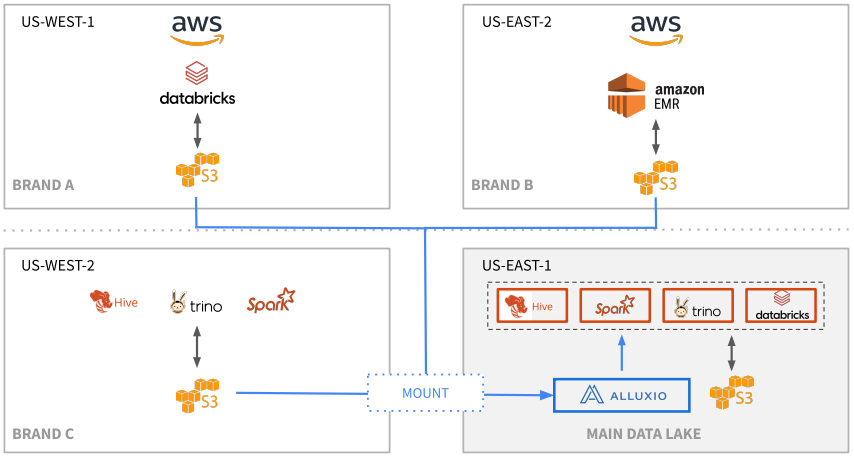
SOLUTION: Federate Data Lakes Without Replication and Serve Various Compute Engines

Expedia adopted Alluxio in the central data lake region, between S3 and compute engines.

“With the introduction of Alluxio, we are seeing better performance, increased manageability, and lowered costs. We plan to implement Alluxio as the default cross-region data access in all clusters in the main data lake.”

-JIAN LI, Senior Software Engineer at Expedia

- **Unify data lake silos without replication.** Alluxio unifies data lakes across multiple regions and provides single access to all compute engines, eliminating the need to replicate data from multiple storage silos to the main data lake.
- **Simplify data access for compute engines.** When end-users are querying cross-region data, Spark, Trino, Hive, Databricks and other compute engines only need to talk to Alluxio instead of fetching data from remote regions.
- **Caching to eliminate the network effect.** Alluxio caches the hot data and brings frequently accessed data to compute engines in the main region. These engines are also sharing data with Alluxio as a regional cache.



RESULTS: Cloud Data Platform with Better Performance and Lower Costs

- **Greatly enhanced query performance.** Data is now immediately available to users without the need to wait for the long process of manual replication and data validation. Performance is significantly increased because of Alluxio's caching. End-users enjoy consistently high performance, leading to a shorter time to value.
- **Significant cost savings.** Alluxio minimizes network egress costs by caching data, eliminating the need to fetch data from cross-region data lakes repeatedly. For frequently accessed tables, approximately 50% cost reduction is estimated.
- **Ease of management.** Data lakes in different regions are unified without manual replication. Alluxio also simplifies the process of setting up fine-grained access controls at the cluster level.

LOOKING AHEAD: Data Mesh Vision

When looking forward, the data platform's vision is to achieve data mesh. Expedia plans to adopt the domain-driven architecture in the main data lake. Having Alluxio as the unified cross-region access paves the way for data mesh, making it easy to share data regardless of which teams produce and consume it.

WHY ALLUXIO

- Unify data lake access without manual replication
- Provide analytics & AI with consistent performance
- Reduce S3 egress costs by ~50%
- Embrace decentralized data mesh paradigm

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 is in production use today at **eight out of the top ten internet companies**. Venture-backed by Andreessen Horowitz, Seven Seas Partners, Volcanic Ventures, and Hillhouse Capital. Alluxio was founded at UC Berkeley's AMPLab by the creators of the Tachyon open source project.

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