

QUERY ACCELERATION & DATA ACCESS AS A SERVICE

COMPANY	Shopee
INDUSTRY	E-commerce
LOCATION	Asia
EMPLOYEES	30k+
TOPOLOGY	Multiple on-prem data centers
COMPUTE	Trino, Python/C++/Go applications
STORAGE	HDFS

ALLUXIO BENEFITS

- Speed up query by ~55%
- Enhance developer productivity
- Simplify data access with decoupled architecture
- Enable a future-proof and flexible infrastructure

Shopee is a leading e-commerce platform in Asia. It is a platform tailored for the region, providing customers with an easy, secure and fast online shopping experience through strong payment and logistical support.

Shopee's data Infrastructure team is responsible for designing, building and maintaining the underlying systems that support data-centric operations across the company. The team's goal is to provide a robust data platform that supports a variety of big data applications for all teams within the company.

CHALLENGE: Slow Query Performance & Data Access Challenges When Onboarding New Applications

- **Trino queries were slow and unstable.** Shopee's Trino queries against HDFS suffer from performance issues, including slow and inconsistent query execution. This negatively impacts end-users, hindering their ability to effectively use BI and dashboarding applications to obtain results.
- **Applications in Python/C++/Go languages needed to be rewritten when accessing HDFS.** Current applications like Trino and Spark that access HDFS use the HDFS API. Onboarding new Python/C++/Go applications that prefer using S3 or POSIX APIs can be challenging with the HDFS interface, as it requires rewriting the applications. This delays the onboarding of new use cases onto the data infrastructure.

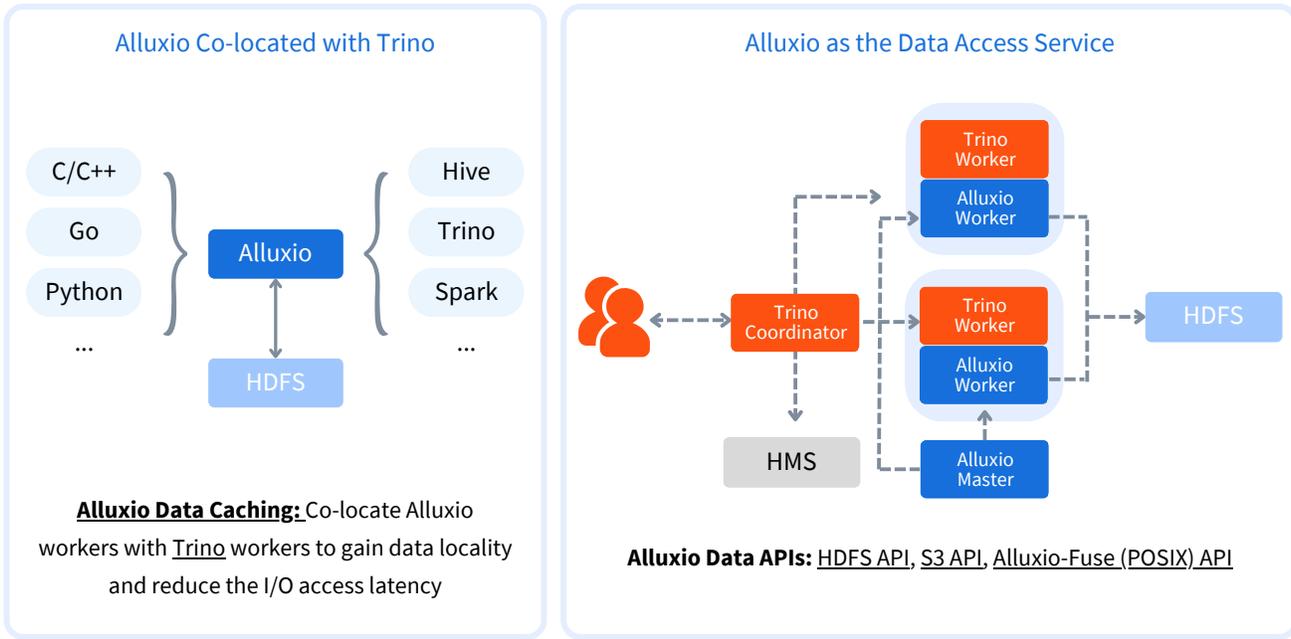
SOLUTION: Caching to Speed Up Queries and "Data Access as a Service"

- **Data caching for Trino.** Alluxio provides caching for Trino. Data stored in HDFS is cached in Alluxio upon first-time access, after which Trino will access the data from Alluxio instead of the HDFS cluster. This caching approach improves Trino query performance and prevents tail latency spikes.
- **Co-Located Trino and Alluxio architecture.** Shopee chooses to co-locate the Trino worker and Alluxio worker on the same machine. This architecture reduces query latency and saves network resource usage required to transfer data between Trino and Alluxio.

"Alluxio empowers our data infrastructure across multiple data centers, allowing us to provide 'Data Access as a Service' to application developers. Alluxio has helped us achieve faster query performance, increased developer satisfaction, and decreased management overhead."

– Li Luo, Director of Data Infrastructure, Shopee

- API translation for flexible data access.** Alluxio provides access to any type of storage via the HDFS API, Alluxio-Fuse (POSIX) API, and S3 API, eliminating the need for Python and C++ applications to rewrite their code. By translating data access requests to storage operations, Alluxio enables “Data Access as a Service” and improves the speed of onboarding new applications.



RESULTS: Performant, Scalable, and Cost-effective Data Infrastructure

- Speed up queries by ~55%.** Using Alluxio as the data cache, Shopee’s Trino queries have been accelerated by 55%. Moreover, applications don't need to be concerned with the performance and tuning of the underlying storage. This leads to faster data-driven insights and better experiences for internal and external customers.
- Enhance developer productivity.** Alluxio empowers Shopee to derive value from day one as data-driven use cases expand. Upon launching, applications can easily access data, freeing up developers to focus on building and writing code instead of adapting to different data access protocols.
- Simplify data access with decoupled architecture.** Alluxio enables a truly decoupled architecture of compute and storage, which allows for seamless integration with the current tech stack. Additionally, Alluxio unifies heterogeneous data access models, resulting in a simplified data platform.
- Enable a future-proof and flexible infrastructure.** Alluxio seamlessly integrates with all environments, data storage, and compute engines. With Alluxio, Shopee’s data platform team does not need to worry about whether technology decisions made today will impact the ability to support the business tomorrow.

ABOUT ALLUXIO

Alluxio, the developer of the open source data platform, makes it easy to manage your data and serve it from any storage to any compute engine in any environment – on premise, in the cloud, or across clouds. By removing complexities and toil from managing and accessing data infrastructure, Alluxio accelerates and future-proofs your data strategy, delivering performant, accessible, cost-effective, resilient, and secure data applications that power improved outcomes, at any scale. To learn more, contact info@alluxio.com or follow us on [LinkedIn](#), or [Twitter](#).