ALLUXIO + Starburst = Better Together

Open Source Data Orchestration for Interaction Analytics in the Cloud

Presto With Alluxio

Presto with Alluxio brings together two open source technologies to give you better performance and multi-cloud capabilities for interactive analytic workloads. Presto's open source distributed SQL query engine coupled with Alluxio enables true separation of storage and compute for data locality and provides memory speed response time and aggregate data from any file or object store.

Interactive Analytic Workload Challenges

Network latency to query remote data is very high, making interactive Presto on remote datasets unattainable

Data copies made across environments must be kept in sync, which means more data storage and transfer costs
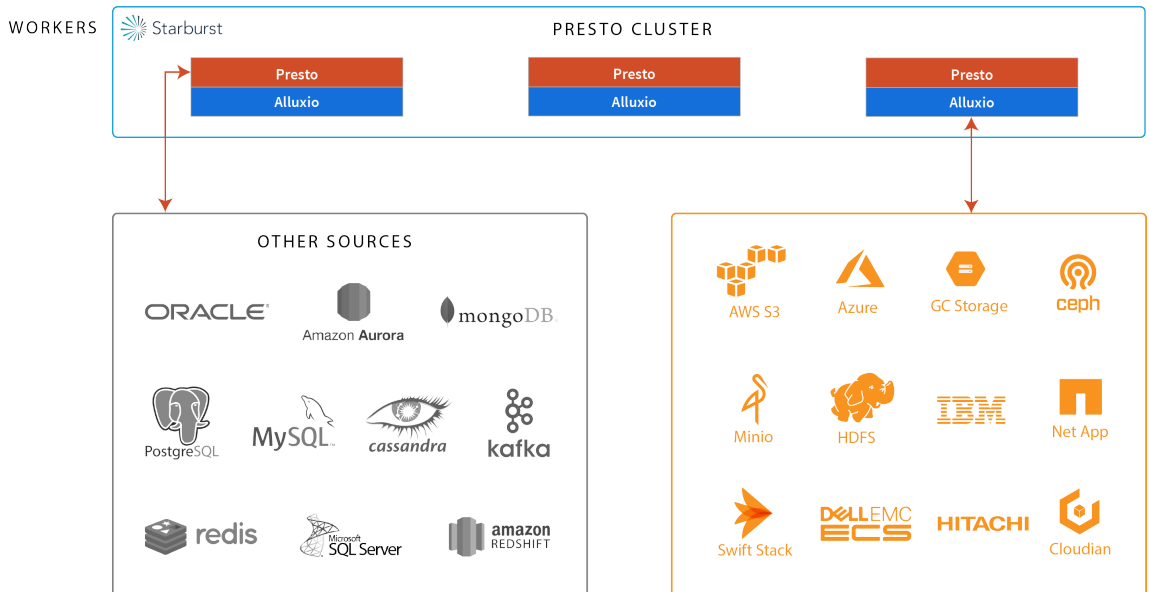
Slow, inconsistent query performance can occur even on frequently queried data due to slow metadata operations & remote data

AWS S3, Google Cloud Storage, and other object stores are not designed for analytical workloads

Presto with Alluxio, Better Together

Presto with Alluxio is a truly separated compute and storage stack, enabling interactive big data analytics on any file or object store.

✓ Alluxio provides a **multi-tiered data caching system for Presto**, enabling consistent high performance with jobs that run up to 10x faster

✓ Alluxio makes the **important data local to Presto**, so there are no copies to manage (and lower costs)

✓ Alluxio connects to a variety of storage systems and clouds so **Presto can query data stored anywhere**
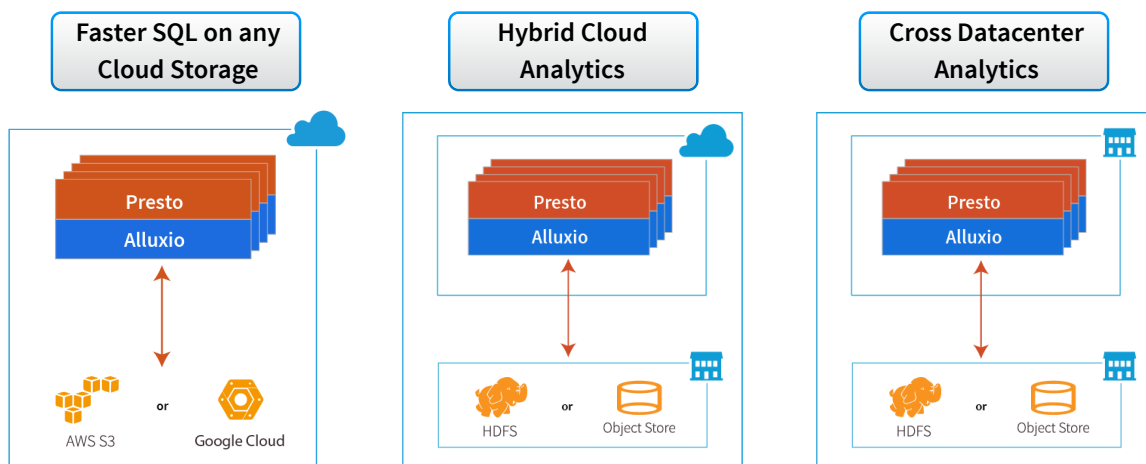
Presto with Alluxio Use Cases

Faster SQL on any Cloud Storage	Hybrid Cloud Analytics	Cross Datacenter Analytics

Reading cloud data into Presto and enabling data sharing is automated and transparent with Alluxio. Alluxio can be deployed colocated with Presto and be backed by a mounted remote storage.

Integrate on-prem data stores like HDFS with Alluxio and Presto and get high performance in your hybrid cloud environment, with all of your I/O offloaded. Burst Presto into the cloud on-demand, when you need it.

Access data anywhere it's located - across regions, sites, or data centers, in HDFS or object stores - for high performance analytics anywhere.

Case Studies

Leading online retailer JD.com built an ad-hoc SQL query engine to support 400,000 jobs (15+ PB) daily, on a system with more than 15,000 cluster nodes and a total capacity of 210 PB. Two challenges they faced were around Presto workers reading remotely from HDFS datanodes and a large query variance. With Alluxio and Presto together, JD.com has seen a 10x performance improvement, including enhanced syncing for better consistency between Alluxio/Presto and HDFS.

Online gaming company Netease, the operator of popular titles like "World of Warcraft" and "Hearthstone", needed a data platform to handle 30TB of raw data collected daily. That raw data is processed in ODS tables by ETL jobs which makes it an even larger amount of data. To support high performance ad hoc queries, they turned to Presto and Alluxio to speed up response time of queries for their massive datasets.

More Resources

- Learn more: Presto with Alluxio to Query Data Anywhere and Faster Compute
- Review Documentation: Running Presto with Alluxio - v1.8 (stable)
- Read the Blog: Top 5 Performance Tuning Tips for Running Presto on Alluxio
- Check out the Webinar: Interactive Big Data Analytics with the Presto + Alluxio stack