



# Lenovo Analyzes Petabytes of Smartphone Data from Multiple Locations and Eliminates ETL with Alluxio

*Lenovo is the world's largest personal computer vendor and one of the world's largest smartphone vendors. The company has invested extensively in global information technology infrastructure, including multiple data centers worldwide collecting petabytes of smartphone data. Analyzing data located in multiple data centers worldwide is critical for Lenovo to understand and improve the usability and reliability of their products. With Alluxio, Lenovo unified data from multiple data centers and eliminated the ETL process while lowering storage cost due to multiple data copies.*

## Customer:

Lenovo

## Industry:

Mobile Communication

## Use Case:

Analytics

Remote Data Access

## Application Stack:

Spark SQL + Hive + Alluxio + HDFS

## Benefit Highlights

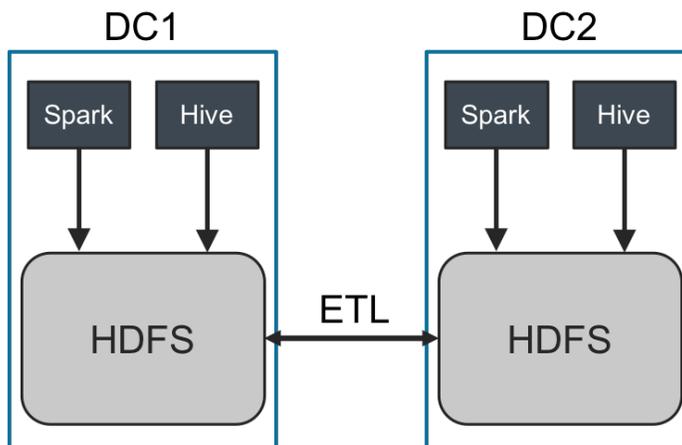
- Eliminated time consuming and error prone ETL process
- Lowered storage costs associated with multiple data copies
- Unified data from multiple data centers
- Regulatory compliance in multiple jurisdictions
- No changes to applications or existing infrastructure

## The Challenge

Smartphone data is processed on the Lenovo Enterprise Analytic Platform. Previously this required a time-consuming and error-prone ETL process to transfer the data from multiple locations to a single data center for analysis. Lenovo uses big data technologies like HDFS to store the data and Hive metastore to store the metadata associated with the structured data. Analytics is performed using Hive and Spark SQL to gain insight into user behavior, popular applications, log analysis and more. The volume of data and number of geographic locations presented multiple challenges:

- High storage cost due to duplication of data
- Bandwidth and performance limitations transferring data from multiple locations
- Regulations preventing the transfer of certain data and excluding it from analysis

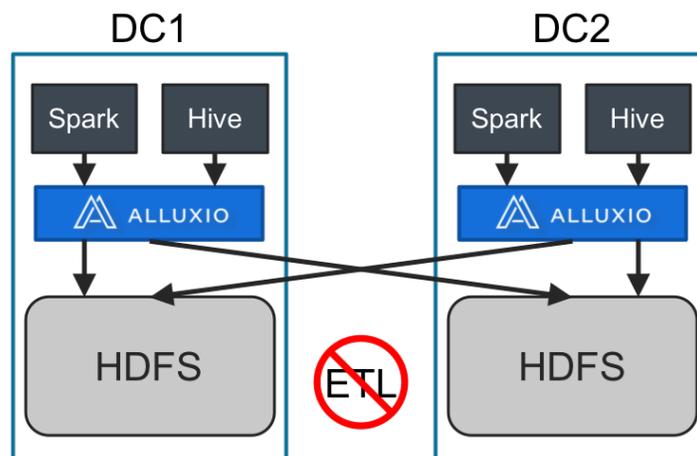
The following diagram shows the initial Lenovo infrastructure



## The Solution

Lenovo addressed the technical challenges by using Alluxio as the data management layer for all smartphone data collected worldwide. The HDFS data stores were connected to Alluxio, providing seamless access for multiple applications through the global namespace. No changes to the application were required. With this architecture, Lenovo performs advanced analytics involving cross data center data synchronization, joins and unions. Alluxio also temporarily stores the data in memory accelerating performance. Alluxio fits within existing security frameworks and enforces the policies in place, ensuring regulatory and compliance requirements from different countries and jurisdictions are met.

The following diagram shows the Lenovo infrastructure with Alluxio



## The Results

With this new architecture, Lenovo now has the infrastructure that allows them to analyze their worldwide data without the need for error prone, time consuming and costly ETL or the need for data duplication. Alluxio maintains the latest copy of the data in memory, or fetches it from HDFS for new requests, so data freshness is assured.

With Alluxio, the Enterprise Analytics Platform now stores data locally in memory from remote HDFS locations and provides transparent access for analytics applications. Alluxio presents the same API to the applications that they were already using. This allowed Lenovo to achieve the benefits without disrupting the existing stack or changing applications.

## Looking Forward

With Alluxio integrated in the data processing stack, Lenovo is now able to access and transform massive amounts of mobile data into valuable insights. This meets the business objectives of improved product quality and customer satisfaction at the lowest possible cost for their analytics platform.