ALLUXIO | TWO SIGMA

CASE STUDY

Leverage the power of cloud bursting with Spark for on-premise Hadoop

Two Sigma, a leading hedge fund with more than $50 billion under management, turned to Alluxio for help with bursting Spark workloads in a public cloud to enable hybrid workloads for on-premise HDFS. With Alluxio, Two Sigma sees better performance, increased flexibility and dramatically lower costs with the number of model runs per day increased by 4x and the cost of compute reduced by 95%.

CUSTOMER

Two Sigma

INDUSTRY

Financial Services

USE CASE

Cloud bursting with Spark for on-premise Hadoop

APPLICATION STACK

Spark + Alluxio + HDFS

BENEFIT HIGHLIGHTS

- 4x faster model processing time
- 10x faster data load time
- 95% reduction in cost of compute in the cloud
- No changes to applications or existing infrastructure

The Challenge

Quantitative hedge funds rely on financial models to manage their business and drive investment strategy. The ongoing business challenge is to develop more powerful models so they can make intelligent investment decisions in a shorter period of time and at the lowest possible cost. The development and testing of investment models relies on Machine Learning techniques applied to vast amounts of data – the more data, the better the model. Data is collected from over 10,000 public and proprietary sources and totals over 35 Petabytes of ever-growing storage. The speed at which this data is processed is critical, as faster model runs enable multiple iterations and improved decision making.

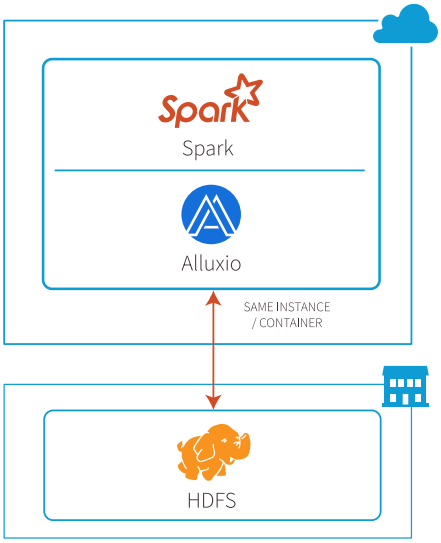
Initially, model runs were performed on-premise with a typical run taking about one hour on 1,000 data processing nodes. Apache Spark is used for the compute framework and data is stored using the Hadoop Distributed File System (HDFS). The workload profile is variable, with periodic load bursts significantly higher than average. This required significant overprovisioning of the infrastructure to ensure processing time would not slow down and be constrained by peak loads.

To address this, the company explored moving model processing to the cloud where the scalability and elasticity of the infrastructure is well suited to the workload profile. This presented several challenges:

- For security reasons data could not be stored in the cloud, requiring model data to be transferred from the on-premise data center prior to each run.
- Due to the size of the data and the physical transfer requirement, model run time in the cloud increased to approximately three hours and an average of only two model iterations per day.
- More expensive reserve compute instances were used to avoid interruptions in the cloud that would cause further delays.
- Any change to the model parameters would require a restart of the data loading process.

The Solution

The company turned to Alluxio to solve these problems with a 40 node Alluxio cluster deployed on reserve instances in the cloud. Data is loaded into Alluxio once, and subsequent data requests by the application are served from memory.



The faster data access enabled spot instances, rather than reserve instances, to be used for the 1,000 compute nodes. The Alluxio cluster provides temporary, non-persistent, storage of the data in memory so when the Alluxio instances are brought down the data is effectively removed. Additionally, the data in Alluxio in encrypted (by the client), so even if the cluster is compromised, the data is still secure.

The Results

With Alluxio deployed, Machine Learning run time was reduced by 75% the number of model iterations per day increased from two to eight. As the data sets grow in size Alluxio will be able to scale linearly to deliver the same performance. With the dramatic reduction in data access time enabling the use of spot instances, the company achieved a 95% reduction in cost of compute. Alluxio integrated seamlessly with the existing infrastructure, presenting the same API to the application and requiring no changes to applications or storage. All security requirements were met with data encrypted in Alluxio and no persistent storage in the cloud.

Looking Forward

With the Alluxio architecture, the business goal of faster model development with more iterations at the lowest possible cost was achieved. Following the initial success, the Alluxio deployment has been expanded to a second cloud provider and an internal data center. The company is now better positioned to take advantage of the ever-growing data sets that lead to more accurate models. Additionally, the company is pursuing increased adoption of the cloud, leading to an overall increase in the efficiency of the infrastructure. More variable workloads are being deployed in the cloud allowing internal data centers to handle the more predictable ones without wasteful overprovisioning. The result is a costeffective hybrid cloud infrastructure that ensures security of the data and significantly increases the performance of business-critical applications.

Stay Connected

Twitter: @Alluxio
LinkedIn: linkedin.com/company/alluxio-inc-/
Meetup: meetup.com/Alluxio/
Slack: alluxio.org/slack

About Alluxio

Alluxio, formerly Tachyon, is open source data orchestration for big data and machine learning in the cloud. By allowing applications to access data stored in disparate storage systems at memory speed, Alluxio enables enterprises to manage data efficiently, accelerate business analytics, and ease the adoption of hybrid cloud. Venture-backed by Andreessen Horowitz, Alluxio, Inc. was founded by the creators and top contributors to the Alluxio open source project. For more information, contact info@alluxio.com.

1825 S. Grant Street | Ste 600
San Mateo, California 94402

www.alluxio.com