Accelerating Data Science

How Options Technology and Pure Storage Drive Faster Data Analysis

options

"Our vision is for the Options Data Platform to be the ultimate marketplace for capital markets data, enabling our clients to easily access and analyze data sets using our very high-performance data environment. Pure Storage® and FlashBlade are a big part of that vision."

JAMES LAMING, VP, GLOBAL HEAD OF INFRASTRUCTURE AT OPTIONS TECHNOLOGY

Overview

This interview is based on a real-life best practices story from Options Technology, a leading provider of IT infrastructure to global Capital Markets firms, supporting their operations and ecosystems.

To capture this story, Pure Storage® interviewed James Laming, VP, Global Head of Infrastructure at Options Technology with the goal of getting a sense of the overall company and what its new Data Platform offers clients. Questions are followed by answers from Laming.

Please tell us a bit about Options Technology and your clients.

Options provides high-performance managed trading infrastructure and cloud-enabled managed services to over 550 Capital Markets firms globally. We provide an agile, scalable platform in an Investment Bank grade Cybersecurity wrapper.

We are traditionally a Platform-as-a-Service (PaaS) vendor. Now we also have a data management platform, which provides both content and a full environment for quantitative analysis.

What was the business driver for Options to provide a data management platform?

Our clients work with huge data sets, routinely performing massive quantitative analyses. We saw that the current platforms available to them were enormously inefficient and costly. So, we built our Data Platform, which is conceptually an app store for data and analytics.

The Data Platform allows our clients to select and access data sets and easily query against them without incurring the time and cost of identifying, licensing, cleansing, and hosting the data sets. We also offer an environment complete with analytical tools and cost controls to make data engineering as cost efficient as possible. Clients can gain access to Data Platform and begin working with data in just a few minutes.

Fast access and analysis depend heavily on storage, so we use FlashBlade® from Pure Storage to make our platform highly performant.

That's great to hear. What about FlashBlade made you choose it for your storage infrastructure?

With FlashBlade, our clients access data through a folder structure, which is familiar and easy to use for them. Behind the scenes, FlashBlade manages the data as objects, which enables much greater efficiency and flexibility in storage, access, and delivery—key to our platform success. This duality makes for quite an accessible repository.

With FlashBlade managing the storage, we effectively have a content delivery network (CDN) platform. We leverage data gravity, bringing tools and applications closer to the data to make workloads highly efficient. Clients simply add Data Platform to their existing Options relationship, then data streams to them as needed.

Data is always accessed remotely, so clients are not pulling or copying the data anywhere. As they access the data, it is cached inside our HPC environment. It effectively streams because FlashBlade is incredibly performant.

Can you elaborate on the quantitative environment you mentioned earlier?

We help clients better control and dramatically cut their data analysis costs with Quantify—an app in the Data Platform. Quantify spins up a full grid environment for a quant in about two minutes. Accessible to all workloads, FlashBlade provides rapid access to data through its fast object store, speeding time to answer and avoiding CPU idle time.

Within the Data Platform environment, we have standard integrated development environments and common data analysis tools, including Jupyter Labs, Visual Studio Code, and more. We maintain all standard data science tool sets and all Python libraries in line with the tool sets. We support an open tool chain, including our own or the client's tool chain, and we can standardize them for teams, users, or companies. The tool sets follow users wherever they go.

Quantify also gives clients the ability to do fine grain cost control on systems, teams, projects, and more. When clients submit jobs, Quantify directs them to the lowest-cost or most performant environment based on spare capacity. It also provides a budgeting control overlay to help clients optimize their resource utilization. It is a true multi-cloud compute solution.

What infrastructure is required to make an effective quant environment?

Automation is key to data platforms that can host multiple services. Quantify uses an API to deploy needed infrastructure pieces of the HPC workloads together with S3 bucket and access management. FlashBlade fast object storage enables many of the central efficiencies of the Data Platform. Our app deployment is quite agile, and our storage with FlashBlade supports this agility by mounting to wherever users go within the environment. Available data can be securely accessed within seconds and new data is available as soon as it is stored.

We provide shared compute resources with great flexibility in cost and reservations (on demand, spot, reserve, reserve on demand, and spot on demand). We can even sequence on demand (pay by second), then spot (run when excess capacity is available), and then burst outwards (short duration extra high throughput) as well. We're sitting in about 18 AWS availability zones currently, which delivers massive redundancy. In addition to this flexible shared system, we offer dedicated compute, dedicated GPU compute, and FPGA compute.

We're in 52 data centers globally and directly connected to pretty much all capital markets. We also have a fantastic backbone network for access and delivery of data.

How do you maintain availability when there can be many concurrent jobs accessing many data sets by many clients?

We built our HPC infrastructure to be highly available and highly performant specifically for Capital Markets, from FlashBlade storage to networks and computing systems. We never stop jobs in favor of higher priority jobs. We build in resiliency, so APIs and processes do not hang. FlashBlade fast object ensures that jobs never fail because of timeout, storage bottlenecks, or unnecessary data transfers. Other general-purpose cloud computing platforms are not optimized for specific object workloads common in financial markets, which can lead to jobs that don't complete and must be rerun, incurring significant cost and wasting time.

We also have dedicated client networks and dedicated fabrics the keep the quantitative analysis separate from other activities so that client networks are not overwhelmed. Consistency and overall availability are extremely high.

What can you tell us about the data sets in the Data Store and available for analysis?

We have about 4.9 petabytes of historical tick data across 179 data sets. There are also enriched data sets, alternate data sets, and more. Data sets are available in cleansed and universal format, which is a great advantage to our clients as they don't have to fit their application to the data. Clients can choose to work with raw data as needed. FlashBlade does not force clients to choose a single format or discard data after transformation. Data is not siloed and is accessible quickly from one source. For active data sets, we support delayed release.

Data pipeline capture is as close as possible to the market data. The Data Platform takes care of data ingest, cleanse, and enrichment, and more data services as clients need. FlashBlade plays an essential role in all steps from data capture to query ready, providing fast object storage and access that keeps up with computing resources, shortening the overall processing time and minimizing network load.

We host all data sets within the Data Store, which is our agnostic data marketplace within the Data Platform. Through the Data Store, we connect clients and data vendors.

You mentioned data cleansing. Can you tell us more about that and how FlashBlade fits in?

All data sets have gaps, duplications, and other problems that need to be fixed before meaningful analysis. This process of cleansing data sets is expensive, taking up an estimated 40% to 50% of data scientist time—easily equivalent to $100,000 per quant. Data Store data sets are pre-cleansed, freeing data scientist time for higher value work.

Vendors who offer data sets through the Data Store have the option to use our data cleansing services. Clients have access to this service as well, relieving them of the cleansing task for their own data sets. They simply drop files to Options or we consume the data flow directly. We can provide API access, with the cleansed data written to whatever endpoint the client chooses.

Data cleansing is a compute and storage intensive process. With its all-flash and fast object design, FlashBlade ensures that storage is not the bottleneck. We can cleanse data very efficiently and make clean data sets available at a fraction of the cost clients would otherwise incur.

How does data enrichment improve data analysis and efficiency?

Data enrichment adds metadata tags that help categorize and characterize the data. This metadata is the basis for the folder structure view, essential for navigating through the massive data set across different clients and workloads. It also enables highly efficient queries. Both save significant time and provide additional angles for data interrogation.

As metadata is used to construct these folder views, lookups for certain data are now possible in a fraction of the time it takes on traditional filesystems. The usage is just as easy as using an API to query the metadata and load only the needed data for this tag. Queries may not need to go to storage to bring back an answer, because the answer is already in the metadata. When they do,

the lookup can pull specific bits of object out very quickly using key value pairs within the FlashBlade object store. This enables deep analysis without moving a vast quantity of data, most of which may be irrelevant.

How is Options better equipped to handle the challenges of data set storage specifically?

We offer an easy solution. Clients simply dip into data on FlashBlade as needed, do their analysis, and get their answer. There is no need for clients to store the data set or make backups, saving hardware costs and system administrator time. And there is no better, faster, cheaper way to get to answers.

If our clients were to take on the task of managing storage of these massive data sets, they would have an enormous cost outlay and administrative headache. The data sets are very large. For compliance, the clients would likely have to create backups, driving exponential storage growth. On top of that, analytical projects typically generate many iterations that need to be stored. While these may be purged at some point, storage must be sized for peak demand.

Beyond all of this, storage plays a crucial role in overall speed and efficiency. Our Data Platform relies on FlashBlade to feed CPUs as fast as they can take data, which minimizes wait time. FlashBlade minimizes network transfers by pinpointing needed data and makes it easy to recover from data loss incidents.

You mentioned minimizing CPU i/o wait time. How does this help your clients?

Yes, this is a great example of how we can help clients get the most from infrastructure investments. CPU wait time costs money. FlashBlade, delivers data to CPUs fast enough to eliminate the 15 to 20% CPU i/o wait time—idle time—typically incurred during data analysis, allowing clients to realize more value from their compute resources. In fact, CPUs sometimes now become the gating factor.

By optimizing at the storage layer, we make sure clients have a highly performant platform for data analysis. The power of our platform, supported by FlashBlade fast object storage, gives clients a significant advantage in performance when doing metadata queries.

What feedback have you received from clients? Are they gaining real benefits in dollars to answer?

We've had some fantastic client feedback. One client reported a more than 50% increase in speed on the Data Platform running on FlashBlade when passing three years of back data, dropping job run time from more than 80 minutes to 37 minutes. The client is no longer storage limited.

We've also done performance benchmarks against MinIO S3 and other object-based cloud computing systems. We're always roughly 40% better, normally split between performance and price. But effectively we found that there's quite a sizable difference in turnaround time. And that's without considering data cleansing, which they do not do.

We do field other systematic client workloads, but they use block and NFS on traditional servers and we sometimes see 40% i/o wait times.

From a cost of data science perspective, we believe there is an even bigger difference in dollars to answer. We combine the best of high-performance S3, FlashBlade fast object storage, and flexible grid computing to enable an environment best suited for the data science paradigm of 'fail often, fail fast.'

Thank you for sharing your story. What do you see coming up next?

We are working with our data vendor partners to ensure we can offer interesting trial data sets so clients can get a taste for the power and value. We also will be adding data sets with built-in Jupyter Lite so clients can do basic queries within their browser.

Above all, we want to deliver a very high-performance data environment that helps our clients be more effective. We want to connect our clients with new, exciting data sets and vendors. Ultimately, our vision is to become the go to reference data store for the financial vertical. Pure's FlashBlade is a big part of that vision because of its unmatched speed and simplicity at scale.

Learn more about Options Technology's journey and how Pure Storage FlashBlade can solve your unstructured data challenges.

PS2343-01 10/2022