BACKTESTING IN THE CLOUD

Avere Systems, in collaboration with Google Cloud, analyse using cloud computing with existing storage to boost analytic throughput

Scott Jeschonek is a leading cloud technology expert at Avere, navigating across customers, partners, executive management, sales teams, support teams and engineering teams. He works closely with GCP and other cloud technology providers to shape the firm's product vision and help influence go-to-market plans. He regularly participates in speaking engagements, webinars, guest blog posts and industry white papers.

Alex Vaysburd is an expert in hedge fund quantitative technology. He joined Google in 2016, having previously worked for over 15 years as a quant and portfolio manager at some of the largest hedge funds in the US.

In recent years we have witnessed a significantly increased usage of cloud computing among hedge funds, who have been quick to take advantage of this powerful technology. The cloud offers unprecedented capabilities in both scale and affordability. The level of adoption has varied — some hedge funds start by testing private cloud systems to get a feel for the technology, while others have gone "all in" and moved their systems to the cloud, removing the need to manage physical servers directly.

Many in the industry have shied away from taking either step. The possibility of operating without large datacentres is attractive, but committing to the cloud continues to present challenges for a variety of reasons, ranging from concerns around business continuity to the acknowledgement that many physical datacentre assets, which have not completed their lifecycle, must remain available for critical operations. Yet with cloud computing technology growing ever more powerful and cost-effective[1], many chief operating and technology officers at hedge funds are taking another look.

The cloud can provide an almost instant ability to improve a fund's analytic throughput, but concerns remain about how to access the data required for simulation, about how best to secure that data. Hedge funds need to have ultimate control and flexibility of how data will be stored and accessed. In this white paper, published in collaboration with our partners at Google Cloud Platform (GCP), we will explore how backtesting workloads can be moved to the cloud while keeping some or all of the actual data on-premises.

The hedge fund industry was built on innovation and technology, and will continue to employ whatever technologies are necessary to remain competitive. In a 2016 survey, 94% of respondents said technology will have at least some impact on competition in the next five years (KPMG/Aima/MFA Global Hedge Fund Survey). We hope you find this article interesting and we would be glad to hear from you with any questions.

COMMON CONSIDERATIONS

Let's start by looking at some of the common considerations hedge funds have as they explore the need for using cloud technology. Many of the questions centre on proprietary data and how it is used.

When it comes to the hedge fund industry, protecting their proprietary data is extremely important. It is the data owned by a hedge fund, used in simulations, that adds unique value, and so is always vigorously protected. For an industry built on "alpha", little is more important than the subsequent performance of the workloads that drive a firm's strategy.

The first question a hedge fund must address is its ability and willingness to use data in a cloud environment.

> "THE CLOUD CAN PROVIDE AN ALMOST INSTANT ABILITY TO IMPROVE A FUND'S ANALYTIC THROUGHPUT"

Not only must the data be protected, but, given the nature of the industry, the capability to scale performance is a necessity. For example, having a sufficiently low-latency network link is important if the data is to remain on-premises, but connect to the cloud.

Markets move quickly, making speed a crucial issue for hedge funds as they seek to run larger numbers of concurrent simulations. Improving time-to-insight is an ongoing quest. The need for speed means most hedge funds gather data in their own unique storage environment, in close proximity to their quants and also close to their on-premises HPC clusters. Having it all "local" helps hedge funds to optimise the performance of their analytic workflows – their quantitative analysis – at the fastest possible level.

One of the things we at Avere have heard from the hedge funds we work with is that they love the ability to run their analytic workloads both on-premises and on cloud solutions like Google Cloud Platform (GCP). With Google's preemptible VMs, a fund can orchestrate a larger cluster in a shorter period of time, resulting in a more rapid execution of simulations. However, as the number of instances participating in the simulation increases, demand for data residing on-premises can saturate a shared WAN interconnect. Having a clear perspective on the prevailing constraints and tuning workloads accordingly are necessary to achieve optimal analytic throughput.

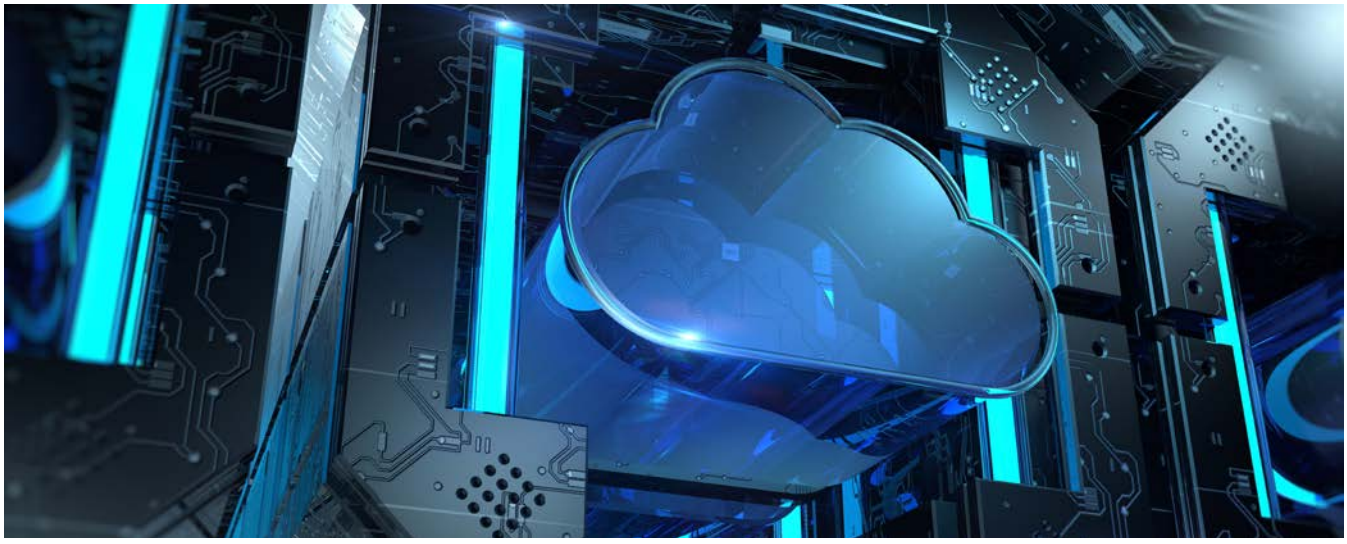
Further, funds must be able to decide where a simulation might be run, based on any number of variables (priority, time-to-run, budget and more), and so ensuring data is available when it is required becomes a major concern, especially when orchestrating between on-premises and cloud infrastructures.

The solution is to present a data access layer. By doing this, quantitative firms with large backtesting demands and multiple quant teams who want to do more alpha testing and quantitative analysis can turn to the cloud. The process has a simple name: cloud bursting.

CLOUD BURSTING FOR BACKTESTING WORKLOADS

Cloud bursting describes the ability to leverage cloud-based assets to offset on-premises limitations. The following scenario demonstrates what it means in practice.

A hedge fund client can programmatically spin up a 20,000-core cluster (using preemptible VMs) in Google Compute Engine (GCE) and run a simulation against an existing data set, whatever the data set is. It uses Avere



vFXT technology as a data access layer in GCE, while the data itself may be resident on a Dell EMC Isilon or NetApp system sitting in their datacentre. As each VM requests data, the data access layer first retrieves, then caches the requested data while presenting it to the VMs. Most simulations will likely read the same data many times, and caching will reduce overall simulation latency. That is the fundamental premise of backtesting in the cloud – using the data access layer minimises any latency the transmission of requested data may incur on the network link to GCE.

Cloud-based backtesting in a cloud bursting model allows the expanse of analytic throughput by simply allowing far more concurrent simulations without incurring more capital expenditure or clogging local compute queues.

HERE ARE SOME THOUGHTS FROM GOOGLE CLOUD PLATFORM'S ALEX VAYSBURD ON WHAT THE TECHNOLOGY CAN ACHIEVE

In some cases, simulations and backtests are executed as cron jobs that run at certain times or continuously throughout the day. The type and size of a cron job depends on the strategy being backtested.

One feature often required in design of those simulations is a division into parallel sub-computations. A single historical simulation of a trading strategy would run sequentially, one time period at a time, without necessarily an opportunity for parallelisation. The scalability of the cloud can now be used to allow firms to run simulations concurrently to simultaneously test multiple models and different combinations of settings for each model. Risk calculations and derivatives valuations make use of MCMC (Markov Chain Monte Carlo simulations), which can be parallelised as well. This parallelisation can now be achieved easily with cloud bursting, whereby hundreds or thousands of CPU cores are procured upon demand and released once simulations complete, with CPU usage charges applied based on the actual time duration of the simulations only.

THE COST AND INFRASTRUCTURE REQUIRED

The industry has been signalling for some time that it is ready for the cloud, that the cloud can be trusted, particularly with certain jobs. The innovative nature of the technology and very low cost barriers to entry of cloud-bursting are compelling.

From an executive perspective, a continuous tension between the desire to optimise the cost structure when considering either cloud or on-premises technology persists. These leaders are reasonably sceptical about the cost benefit. Yet in reality, cloud bursting can be an affordable and flexible solution for the following reasons:

- The cloud-bursting process can be turned on and off as needed. Adding physical assets cannot.
- Analytic throughput can be increased exponentially by spinning up cloud-based cores without the typical costs associated with owned infrastructure.
- The technology is easy to manage using industry-standard tools such as Ansible or python scripting.
- Firms make no commitment to permanently move data. Use of expensive block storage is minimised.
- Funds can retain existing data management procedures, whereas policies may need to be revised if the data leaves the local environment.
- Funds can control costs by leveraging Google Cloud Platform Preemptible Virtual Machines (short-lived compute instances suitable for batch jobs and fault-tolerant workloads that can be 80% less expensive than regular instances).
- One significant advantage to using cloud bursting with a data access layer is that the data brought to the cloud is ephemeral. Once the simulation is done running, you have the option of resetting the vFXT environment and wiping the cache.

The main logistical requirements for a cloud-bursting architecture involve networking and security. For data remaining in your existing on-premises network-attached storage (NAS), a WAN interconnect must be provisioned as a network link between your datacentre and GCP. Google offers a variety of network connectivity methods, but generally you should ensure at least a 1 GBPS link between your datacentre and GCP. The NAS environment must be accessible from GCE, and so any firewall rules must be configured to ensure network-level access between the two environments. Since the Avere virtual cluster (Avere vFXTs) will reside in GCE and will be the primary access method from GCE, you can lock down most, if not all other addresses on that link.[2]

Once the NAS environment is logically available to GCE nodes, you can then decide on your computational approach. Some groups want to replicate their on-premises compute cluster virtually, and in these cases you would proceed to set up your grid environment in the cloud. You can leverage tools such as Terraform or GCPs own Deployment Manager to set up the entire GCE environment as your compute jobs require. Avere provides a sim-

ple set-up script to enable the data access layer, assuming you've provided the above network connectivity this typically takes minutes to complete.

The environment set-up is mostly finished at this point; the data access layer allows your compute nodes to access the data as they would in your local environment – so no reworking of jobs is required. You are ready to cloud burst your backtests.

We are seeing a slow but steady uptake of this model. The rate of adoption is particularly high among start-ups not burdened by existing infrastructure. Cloud bursting removes some organisational and technical obstacles quickly, and allows for cloud resource on-boarding at the firm's own pace.

ALEX VAYSBURD ON GOOGLE CLOUD TECHNOLOGY

When a hedge fund operates a proprietary datacentre, efficient procurement and allocation of computing resources is a challenge. Quants/researchers may require hundreds or thousands of CPUs when running simulations, yet when simulations complete, much of the datacentre's CPU capacity will remain idle. With cloud bursting, a hedge fund may allocate CPU capacity on demand, run simulations, and release the virtual machines once simulations complete. This is a cost-effective way to manage variable CPU requirements, which (as an added benefit) relieves the hedge fund from the need to manage proprietary datacentres.

With cloud technologies such as TensorFlow (an open source machine learning framework developed by Google), parallelisation is implemented under the hood. When a TensorFlow model is trained in distributed mode using Google Cloud ML Engine, a cluster of multiple virtual machines is automatically configured and deployed for the duration of model training and released once the computation completes. The customers are only charged for GCE resources consumed during the model's training.

GCE's capability combined with local SSDs (flash memory storage devices), plus Google's network infrastructure, gives clients the performance they often wouldn't be able to achieve on-premises. When running distributed data processing applications on Google Cloud, the data usually travels between regions without leaving Google's high-performance network, which enables GCP to attain both efficiency and industry-leading level of security.

OTHER CONSIDERATIONS

Google Cloud has done a great job with offering flexibility in how resources are consumed and billed. For example, you can pay for VM usage by the minute. You can also do long-term contracts if your work flows will run 24 hours a day, 365 days a year, but when you start it is predominantly pay as you go.

Yet it is often a misconception that the main reason to use public cloud is that it is always cheaper than using your own datacentre resources. People often overlook the key advantage to the cloud, one that can be huge to financial firms; that is, a firm may run as many simulations as they like, meaning the time to success (or failure) can be greatly reduced.

So, if a hedge fund leverages Google Cloud, combining a flexible data access strategy with Avere, it can amplify the amount of analytic throughput offered by the current physical environment. In essence, what a quantitative hedge fund could previously do only linearly, can now be done exponentially, which can give them a significant competitive advantage in the marketplace.

ALEX VAYSBURD ON INDUSTRY UPTAKE:

CTOs at hedge funds will typically tell you they don't want to be in the business of running datacentres. In the next five years, as the existing hardware gradually becomes due for replacement, the question will arise as to whether to continue investing in new hardware or shift the data storage and computation load onto the cloud. The transition to the cloud won't happen overnight and will be a gradual process that can be expected to complete for many firms over the next five years.

CONCLUSION

An inflection point is happening. Just two years ago, many hedge funds were not ready to embrace the cloud. Moving backtesting workloads to the cloud, as we have outlined in this paper, is an example of how that can succeed in practice. Using Avere with Google Cloud, hedge funds can start today. In cloud-bursting for compute, data is always transient, optimising control of a firm's proprietary data.

If the customer becomes comfortable with using Google Cloud Storage as a permanent spot for data, Avere provides the technology to transfer the data from on-premises to cloud services offered by Google and others (while maintaining continuity with the simulation workflow software).

The cloud offers great advantages of affordability and scale, giving hedge funds the opportunity to enhance their operational layer and ultimately their trading capability. We look forward to hearing from any industry members of other individuals with questions about this exciting technology. ■

[1] See "Google Cloud Platform Live - Blending IaaS and PaaS, Moore's Law for the cloud" at https://cloudplatform.googleblog.com/2014/03/google-cloud-platform-live-blending-iaas-and-paas-moores-law-for-the-cloud.html

[2] Google Cloud's Direct Interconnect enables both on-premises and Google Cloud Platform resources to share a single RFC1918 network address space.

AVERE

Launched almost a decade ago by technology experts specialising in file systems, the firm has set out on a mission to reinvent storage by using fast, flash-based storage in the most efficient manner possible. The firm is located on the North Shore of Pittsburgh, Pennsylvania and works with many of the world's top brands in technology and finance, together with several other sectors.

GOOGLE

Google is a technology company based in California. Google Cloud Platform (GCP) provides customers with multiple options for computing and hosting. Users can choose to work with a managed application platform, leverage container technologies to gain lots of flexibility, or build their own cloud-based infrastructure.