

## Predicting the Future in Trading 80TB at a Time

How a leading quantitative research company leverages massive data sets and large-scale analysis with cutting-edge machine learning for trading advantages



### Results achieved with Aerospike

- Previous “impossibly-long” data refreshes are now conducted in 12 minutes.
- Refreshed data is now 10-20% more accurate.
- Able to keep up to 20 years of data online and update it every five minutes.
- Now able to predict expenses around data storage growth needs.
- Readily scalable beyond 1PiB while maintaining performance.
- Conducting 95% of reads and writes in under 1 millisecond at a throughput greater than 1 million transactions per second.

### The Company

The use case presented in this profile is from a leader in algorithmic trading strategies that utilizes the latest in scientific data analysis methods for making new discoveries. The company's researchers derive trading strategies from deep analysis of large and often noisy historical and current market data, often enriching this data with 3<sup>rd</sup> party sources and calculating ‘implied’ pricing for less liquid or rarely trading securities.

The firm combines industry leading technology and resources to build a single, powerful platform for researching algorithmic trading strategies. They employ rigorous scientific methodology, including artificial intelligence, neural nets, NLP, deep learning, robust statistical analysis and pattern recognition to analyze the extensive and varied global financial data ecosystem, extracting deep insights from truly massive datasets. They then look to validate those ideas and to predict the movements in global financial markets with time series data. The firm conducts arbitrage, predicts individual company events, looks for investable trends and global macro strategies.

### Challenges to Scale and Growth

To develop the best version for any one of their time series models, the company was looking to track 200,000 products each with 200 unique attributes, trading on different global exchanges. In addition, there were concerns about the efficiency of their data compression leading to spiraling TCO costs. Prior to adopting Aerospike, they were using a SQL Server + caching layer but were unable to access as much data as they'd like in the time they needed, nor was it as current as they felt it could be. This created great inefficiencies and frustrations for their data science team and securities analysts.

### Goals

The firm was looking to optimize their data architecture, enabling them to refine their models for financial time series financial with the following attributes:

- Monitoring 200k products
- Monitor 200 fields for each of the 200k products
- Maintain up to 10 versions of each data point
- Conduct data point refreshes every 5 minutes
- Maintain 20 years of data online
- Have the scalability for 13.9PB of data
- Minimize data footprint
- Compute on data-in-place

Efficient storage and retrieval of the necessary data is critical to the firm's business, leading to their decision to find and deploy the highest performance, most scalable data architecture currently available.

## Why Aerospike

Compared to the other solutions being evaluated, the main drivers that made Aerospike so attractive to this firm was its ability to provide high performance at scale combined with a lower and predictable total cost of ownership in addition to the ability to incorporate mission critical security requirements.

Upfront during the presales process, the Aerospike engineering team was highly engaged during the POC. The Aerospike engineering team demonstrated deep understanding of the company's problems, giving the company confidence that investing in the Aerospike platform and product vision would be an asset to their future growth.

## Benefits with Aerospike

### Performance

During the initial evaluation with Aerospike, the firm was able to achieve 1 million transactions per second (TPS) with 95% of reads and writes completing in under 1 millisecond using 9 cloud-based VMs. It was stunning to the evaluation team; it was several times faster than what they were seeing before. Beyond being highly performant, Aerospike is predictably performant, delivering high write throughput at low latencies, enabling enterprises to build larger-scale applications more easily at a lower cost.

### Easily Scalable

Initial testing was conducted with 5TB of data, but during this initial evaluation phase linear scalability beyond 1PB while maintaining performance was demonstrated. Because the Aerospike Hybrid Memory Architecture™ database optimizes hardware efficiency, the firm anticipates being able to substantially reduce data footprint.

### Enterprise-Grade Security

Security is of paramount importance to this organization – only half in jest they describe themselves as 'being more security conscious than the NSA'. The InfoSec experts probed the various access points and links within an Aerospike architecture in order to ensure there were no viable 'attack

## Aerospike versus the Competition

This leading quantitative research firm was having a critical issue in obtaining reliable data in timely fashion. Part of the problem is that data comes in from multiple sources, albeit common ones in the financial services industry including Thomson Reuters, Bloomberg and FactSet. Information such as bond prices can be "a mess" and in conflict which needs to be resolved; algorithms can be further impacted by things such as corporate actions, (e.g. stock splits, special dividends or company divestitures).

***"Data refreshes that hadn't been possible due to length of time needed, are now possible to complete in 12 minutes on Aerospike..leading to data that is 10-20% more accurate."***

Vice President, Engineering  
Leading Quantitative Research Firm

vectors' – even at the most unlikely theoretical level – a scrutiny we were pleased to say we were able to satisfy, crucially without sacrificing performance.

Aerospike provides encryption at rest and on the network; offers role-based authorization, LDAP and Kerberos based authentication, integration with HashiCorp Vault as well as auditing.

### Data Compression

By carefully exploiting some regularity properties of the time series data, in conjunction with opportunities offered by Aerospike's complex data type API, data was compressible in some instances by as much as 90% - one data set moving from a footprint of 260Mb to 17.5Mb.

### Lowering Latency

In capital markets, low latency is critical for algorithmic trading in order to react to market events faster than the competition and thus increase profitability of trades. There are many technical factors which impact on the time it takes a trading system to detect an opportunity and to successfully exploit that opportunity. In this instance, the firm was seeking low latency middleware to store and access financial data in a way that let them keep their edge.

***"Everything you (Aerospike) said came true. Why are you (Aerospike) so much better than even the next best competitor??"***

Vice President, Engineering  
Leading Quantitative Research Firm

Thus, the company would scan the data to clean it – on partition at a time. It's data lakes, HDFS and HBase systems “were not cutting it”. The company's SQL Server + Memcached architecture had reached the point where they could no longer grow in a linear or predictable way. The ‘status-quo’ options included Mainframes (which charge on a MIPS basis) which were expensive, labor intensive to update regularly, and were not distributed geographically.

The other option of adding cache on top of RDBMs also was deemed too expensive, requiring lots of hardware, and was similarly complex to manage, maintain and keep running. In addition, it was susceptible to inconsistent performance (as caching systems are wont to do).

The team was open to making major changes to their data architecture and was examining many NoSQL options, both open source and proprietary. In the end, after examining eight others including Cassandra, Redis, and Redis Enterprise there was consensus that the Aerospike Hybrid-Memory Architecture™ could offer both superior performance and total cost of operations (TCO) savings over the other options.

Because Aerospike achieves its high performance with only indexes in memory while persisting data on SSD, it requires a much lower server footprint while also providing predictability. (This is the Aerospike Hybrid Memory Architecture™ advantage.) As a result: Aerospike always reads the data in the same, highly repeatable manner.

Aerospike also has node awareness features, where each node knows what data all the other nodes contain, regardless of data center geographic location.

With Aerospike, the firm is now able to keep up to 20 years of data online and update it every five minutes and have predictable expenses around data storage growth needs.

## Providing Optimized Data Platform for Machine Learning / Statistical Analysis

Capital markets firms routinely employ machine learning and artificial intelligence techniques to build algorithmic trading systems that learn from data without relying on rules-based systems. The firm had built a winning track record of financial performance by ingesting data from non-obvious attributes (tracking 200 elements per financial asset is notable in the industry). Their strategy to record and store as much as they could on each asset let them build models that can adapt to a changing environment, and volatility of markets is likely to continue to be a major factor that makes/breaks firms in this segment – giving it a competitive advantage.

## Conclusion

In the world of finance, few organizations move faster than quantitative research and Algorithmic traders where milliseconds matter and a machine learning model is only as good as the amount and accuracy of the data it is fed.

This leading quantitative research firm knew it was time to make a major change to accommodate scale and speed, looked at all of the best options today's data platforms had to offer, and chose to go with Aerospike for the benefits our unique technology provided them and our commitment to client success. Aerospike looks forward to having a long and mutually prosperous partnership together.

### About Aerospike

Aerospike is the global leader in next-generation, real-time NoSQL data solutions for any scale. Aerospike enterprises overcome seemingly impossible data bottlenecks to compete and win with a fraction of the infrastructure complexity and cost of legacy NoSQL databases. Aerospike's patented Hybrid Memory Architecture™ delivers an unbreakable competitive advantage by unlocking the full potential of modern hardware, delivering previously unimaginable value from vast amounts of data at the edge, to the core and in the cloud. Aerospike empowers customers to instantly fight fraud; dramatically increase shopping cart size; deploy global digital payment networks; and deliver instant, one-to-one personalization for millions of customers. Aerospike customers include Airtel, Banca d'Italia, Nielsen, PayPal, Snap, Verizon Media and Wayfair. The company is headquartered in Mountain View, Calif., with additional locations in London; Bengaluru, India; and Tel Aviv, Israel.