

Case Study

Data engineering for early-stage hedge funds:
how data science infrastructure as a service can
create a competitive advantage.

bmlitech.com

How a start-up hedge fund leveraged the BMLL data science platform to enable its Quantitative Research team to focus on research and development, reduce set up costs, and shorten time to market.

To develop effective investment and trading strategies, quantitative or systematic hedge funds must carry out extensive research and run complex computation and continuous model calibrations. They require robust infrastructure in place to do this. Building such a data engineering infrastructure in-house can be a significant investment for early-stage hedge funds, since it requires substantial financial resources to acquire the necessary hardware, software and associated technical support resources prior to generating any potential returns.

The quality of market data and connectivity to reliable data feeds also plays a key role in strategy calibration, plus consideration must be paid to the high cost of building a team of quant researchers to curate, cleanse and manage the various data feeds needed to optimise signal detection and deliver the alpha generation process.

Start-up quantitative and systematic hedge funds need to access critical data engineering infrastructure within a reasonable timeframe to shorten their time to market, but the cost of doing so must be achievable. To launch and become established, funds of this nature need to streamline the set up and ongoing management of data engineering processes and be able to continue these tasks going forward without risk of disruption.

Aleto (formerly Magma Capital Funds) is a start-up Hedge Fund that delivers cutting-edge product and technology offerings to empower RIAs. Gershie Vann, CEO and Founder speaks to BMLL about the complexities and challenges of setting up a new systematic hedge fund, and how Aleto was able to overcome these challenges.



Gershie Vann, CEO and Founder,
Aleto (formerly Magma Capital
Funds)



"Operating with a small team presents unique challenges as it requires the expertise of team members to function in their respective domains. In the case of Aleto, it is crucial for our Quantitative Research team to focus on research and development, which prompted us to avoid allocating their valuable time to data procurement. Instead, we entrusted our Operations team with this responsibility. However, since the Operations team lacked expertise in the data aspect of the business, this initiative posed significant challenges.

The data procurement process is also excessively expensive due to the numerous intermediaries involved, resulting in high costs to access the data. For a small, emerging manager such as Aleto, finding the funding for obtaining this data can be particularly difficult and time consuming. Bulk data purchases can amount to hundreds of thousands of dollars, and ongoing fees are required to regularly obtain live data.

The equipment and hardware necessary for receiving data, building, training, and testing models, as well as computing power can easily exceed a million dollars. This type of investment poses a significant hurdle for start-ups in the systematic hedge fund industry.

Moreover, historical data is often impractical to use. It takes weeks for the data vendor to obtain the data, followed by additional weeks to process it. Integrating this data into the internal infrastructure to make it usable by the team also takes weeks, further adding to the time constraints".

Gershie Vann, CEO and Founder,
Aleto

Reduce market data licensing fees

If the traditional aim of a systematic or quantitative hedge fund is to build a data science infrastructure in house, market data licensing fees and data security need to be taken into account. The costs associated with this can be extremely high, especially if the fund relies heavily on traditional market data directly from exchanges, other trading venues or incumbent vendors. For example, direct data licences may need to be put in place, usually having to cover both real-time and historical use cases from the outset, additional licence fees may be incurred when accessing the full depth venue order book, plus monthly vCPU cloud compute fees will increase the costs and preparation of setting up a fund for operation.

However, an alternative approach can solve the set-up problems for an early-stage hedge fund. By accessing a ready-to-use, quant-built data science platform, hedge funds can access the necessary research infrastructure to support their fund's specific operation and growth from the outset, without the high upfront costs and lengthy timelines normally associated with launching a quantitative or systematic fund.

Aleto sought a solution to address these difficulties. They wanted to find a single market historical data provider within their budget and who could deliver a solution quickly. Gershie Vann describes the process they undertook.

"One of our longstanding vendors introduced us to BMLL. We scheduled a meeting with BMLL to learn more about their offerings and capabilities. The information gathered from the meeting was shared with our Quant and Leadership teams to assess if BMLL could meet our needs. There was a two-week investigation period where our Quantitative Research Team worked with BMLL to understand the product before deciding. During this period, our team discovered that the data provided by BMLL contained more information than the processed data we were previously interested in, which was an added advantage. It was determined that partnering with BMLL would address most of our challenges and enable us to quickly evaluate their potential as a long-term partner.

During the implementation phase, we faced some challenges. Our team needed to familiarize themselves with new data formatting & how to operate in new environments, but BMLL provided excellent customer support to assist us in overcoming these challenges effectively. Our team needed to rewrite our API, and the implementation took approximately two months to build our infrastructure framework to process and migrate the data".

Gershie Vann, CEO and Founder,
Aleto

Data quality supported by highly reliable data engineering process

To carry out robust research, development and back testing of investment strategies, hedge funds require high quality historical data that captures full market activity. One of the challenges hedge funds encounter, that they must overcome, is poor quality historical market data from real-time data providers. This is because real-time data providers often offer historical data at zero cost to consumers of their real-time data service, and there is no incentive for them to carry out robust quality checks on free historical data.

Whilst this data may initially be considered great value, the commoditised historical data that is provided is typically nothing more than a collection of real-time data, printed as a side product or 'exhaust' of a real-time feed. As data quality is paramount in quantitative or systematic strategy development, even minor inaccuracies or gaps can negatively impact the performance of investment strategies. Unfortunately, this results in the need for a significant and ongoing quant resource allocation to cleanse and normalise this unformatted data to make it usable by hedge funds. Furthermore, a fund's ability to utilise full order book data can be severely restricted by the significant task in hand of cleansing and curating Level 2, market by price data.

As a hedge fund develops and grows, its need to run highly competitive investment and trading strategies will increase, and its need for continuous and effective research and backtesting grows accordingly. Its data engineering and processing requirements will expand exponentially. In-house infrastructure may struggle to keep pace with increasing data volumes, leading to more resource allocation to support a commoditized data solution, which in turn often leads to reduced availability of resources to focus on the core investment business, operating strategy, or alpha generation. This inevitably results in scaled back research capabilities or not being able to produce a defined production investment strategy before funds begin to run out.

Scalability and flexibility: superior computing power to speed up research and backtesting process

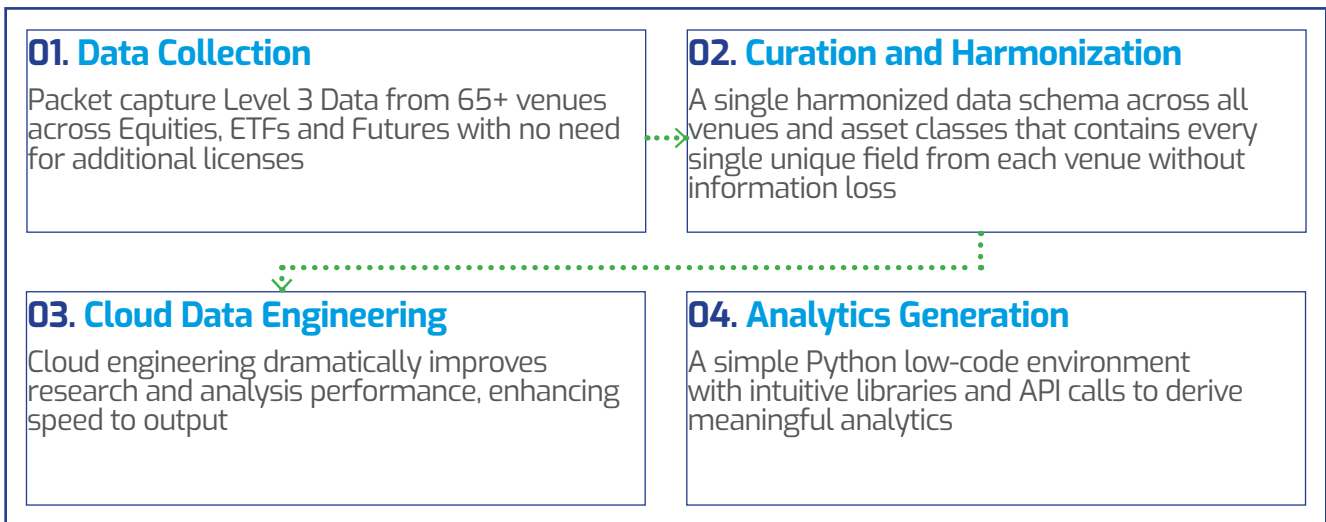
In a highly competitive landscape, time is critical for start-up hedge funds. Developing an in-house data engineering infrastructure requires extensive planning, resource allocation, implementation time, and last but not least, large scale funding. However, the time to market issues associated with building infrastructure in-house can be solved by using an out-of-the-box Data Science as a Service capability that provides quants with the highest quality derived data source available. By leveraging such a service, quants can gain easy access to full depth order book, including Level 3 data, greatly reducing the workload for in-house data engineering.

Dedicated historical data provider, BMLL, provides a cloud-based scalable Python research sandbox that delivers direct access to a harmonised data warehouse. It can be easily integrated into the hedge fund's quantitative research processes and production pipelines, and users gain the full breadth and scalability of Cloud computing power. By using a pre-built cloud-based platform from a specialist provider, early-stage hedge funds gain rapid access to engineered, high-quality, granular market tick data in a secure dedicated sandbox environment for carrying out research, signal detection and back testing, without the need to set up direct data licences themselves, nor incur additional licensing fees to access the full depth venue order book (Level 3) data.

The data engineering headache is removed

Specialist historical data and analytics provider BMLL completes the data engineering process so that quants' time can be focused on their business, and not on the time-consuming tasks associated with data licensing, collection, curation, and formatting. The BMLL platform (BMLL Data Lab) provides pre-wrangled data sets, tools, simulations, and analytics capabilities, enabling speedy access to a wide array of historical order book data science capabilities. BMLL undertakes approximately 200 million checks on the data within its central data repository each day to maintain data quality and accuracy, with robust collection, curation and harmonisation, cloud engineering, analytics generation and data management documentation methods to ensure that the data is complete and easy to reference. The BMLL Data Lab is a scalable and flexible infrastructure, built from the outset to address the needs of a systematic or quantitative business. For example, the BMLL data dictionary and data documentation library ensure users can quickly find and reference what they need.

Out of the box “DATA SCIENCE AS A SERVICE” from BMLL



BMLL also provides access to dedicated quant expertise that can be invaluable for funds seeking to optimise their quantitative strategies and navigate complex market structure dynamics. Hedge funds can easily access the most granular datasets available and manage complex analytics, ensuring consistent performance and responsiveness that scales with growth. They can rapidly research, develop, backtest and deploy their quantitative strategies, increase their data processing capabilities, and access new data from new markets without investing in additional resources or experiencing operational disruptions. Furthermore, BMLL acts as Vendor of Record for the content within the platform - users need no additional data licences to access the Level 1, 2 or 3 data within the BMLL Data Lab environment.

Aleto chose BMLL and the BMLL Data Lab (Data Science as a Service) product because of the speed at which its team could access business critical data.

“BMLL allowed us to bypass the complex and costly process of obtaining data from exchanges. As a result, we could focus our efforts on conducting research to determine usefulness and relevance of data to our specific needs. Having immediate access to the data eliminated any delays that would have otherwise hindered our research efforts. By removing these barriers, our team could dive straight into analyzing the data and extracting results”.

Gershie Vann, CEO and Founder, Aleto

The outcomes for Aleto

"By using BMLL, we can obtain data directly without paying the fees associated with exchanges and data processors.

Our Quantitative Research team has the flexibility to conduct research in a cloud environment, allowing for a more efficient and scalable research process. Because we can operate in the cloud, we do not need to download all of the data at once.

Aleto was quoted by a data processor that it would take at least 6 weeks of development to receive a portion of the data we needed. Additionally, we would need to purchase historical data from elsewhere, for over \$100,000.

Aleto was able to cut costs during the research period of product development by working in a cloud environment.

Previously, when Aleto purchased data from a data vendor, the data would be sent to us via our cloud compute provider. By removing this step, Aleto is now able to make substantial cost savings on cloud compute requirements every month."

Gershie Vann, CEO and Founder,
Aleto

Conclusion

For start-up quantitative hedge funds, leveraging the BMLL Data Lab as an Infrastructure-as-a-Service solution offers numerous advantages. It provides cost efficiency, accelerates time-to-market, ensures consistent data quality and completeness, offers scalability and flexibility, and provides access to expert support and collaboration. By outsourcing a fund's data engineering infrastructure to a trusted specialist provider like BMLL, start-up funds can focus on refining their investment strategies and driving growth whilst removing costs and complexity from their day-to-day business. By utilising the BMLL Data Lab, new hedge funds can level the playing field, enhance their competitiveness, and increase their opportunity for long-term success in the fast-paced world of quantitative finance.

Contact us

