InContext Paper by Bloor

Author **Paul Bevan**

Publish date **December 2022**

Financial Trading Technology and RUMI™ from N5

> “In this paper, we will look at how a mix of huge volumes of market data that need to be analysed, continual and continuous development of trading algorithms, strict regulatory compliance, sophisticated risk management, real-time communications with trading partners and financial exchanges, and lightning-fast execution of trades have driven the development, often in-house, of sophisticated, expensive IT systems that bear little relation to the general purpose IT infrastructure seen in most other businesses.”

Introduction

It goes without saying that advances in Information Technology (IT) have had a profound, and growing, impact on business over the last decade. It is, perhaps, less obvious that some business sectors, having seen the competitive advantages offered by these advances, have exerted immense pressure on IT to deliver higher and higher levels of performance and availability. Financial Trading companies are not unique in seeking a business edge through higher IT performance. However, a unique mix of business requirements have combined to create technical challenges that, in turn, have created an IT arms race which potentially raises barriers to entry for new or smaller companies trying to compete and innovate.

In this paper, we will look at how a mix of huge volumes of market data that need to be analysed, continual and continuous development of trading algorithms, strict regulatory compliance, sophisticated risk management, real-time communications with trading partners and financial exchanges, and lightning-fast execution of trades have driven the development, often in-house, of sophisticated, expensive IT systems that bear little relation to the general purpose IT infrastructure seen in most other businesses.

In the context of these challenges, we will highlight a new IT platform, RUMI, and describe how, and how well, it meets the technical challenges of financial trading while allowing companies to focus on developing and deploying new applications without having to wrestle with the design and deployment of a complex IT infrastructure.

The Financial Trading Environment

IT, in the form of compute and data handling is a critical ingredient for a successful financial services company and sometimes its prime, or even only source, of competitive advantage. Strategies for investing and trading are data-intensive and very computational and may have a limited successful shelf life. As a result, these organisations develop, implement and perfect electronic trading strategies, and the underlying IT infrastructure continually.

IT is used for front office, which is where the trader sits, accesses information and places the trade, middle office when the trade is made with both parties sending electronic confirmation, and back office where the legal details and reporting is done. All these steps occur electronically.

The financial industry is heavily regulated and needs to provide many different reports to its regulators, clients and partners. In Capital markets this often requires heavy computation and a need to keep up-to-date information from both their own internal systems and from capital market exchanges.

Automated tools have also picked up a role in compliance monitoring, where they automatically update the legal and reporting institutions so that traders only focus on trading.

New platforms now permit live or near-live computation where ten years ago, computationally demanding tasks such as capital computations, credit or capital value-adjustments to derivatives books and variable annuity hedging programs involved overnight batch runs or weekend runs with complicated planning and inevitable frustration when errors triggered reruns of critical and long-awaited outcomes. Expansion into alternative data and use of machine learning also trigger more demand for scalable computation.

The Impact on IT of Algorithmic Trading

Arguably, the biggest impact on the IT infrastructure recently has been the adoption of automated algorithmic trading which has been supplementing and/or replacing electronic trading. This is most evident in equities, but is also growing as a percentage of all trades in other markets such as futures, foreign exchange and bonds

Algorithms have a short shelf life. Once they start executing trades, the competition reacts to counter the

> “…a unique mix of business requirements have combined to create technical challenges that, in turn, have created an IT arms race which potentially raises barriers to entry for new or smaller companies trying to compete and innovate.”

algorithm with new ones of their own. Since the algorithm cannot be static anymore and must adapt continuously to new input to remain relevant, this places significant stress on the hardware infrastructure. As such, the backend infrastructure must be able to accommodate live-data feeds and quick processing of large amount of data. Databases must be able to feed the compute engine in real, or near real-time to update the algorithm. This creates a loop of data movement and processing that must remain secure and reliable, i.e., experiencing near to zero down time.

The Algorithmic Development Platform needs to aggregate massive amount of data in parallel as quickly as possible. Therefore, databases that allow fast IOPS, ingest various data type and make it easy to query for languages such as Python, Spark, SQL or R are preferred.

The risk analysis or data-science models must be supported by powerful parallel compute power, especially in the back-testing phase. Models' size can commonly fit on one or two graphic cards. Therefore, GPUs and CPUs with large number of cores and high-memory bandwidth are preferred. Since application downtime translates into pure profit loss, high availability of the platform is a prerequisite.

On the other hand, for the trading platform, the algorithm is simply put in use. Latency is key in this case since it must connect to exchanges as fast as possible, ideally before market and competition reaction. Since the algorithm has been optimised, it is similar to the inference phase of an AI model. In this case, FPGAs or high-frequency over-clocked CPUs are commonly deployed for best performance.

Nevertheless, the need of a second environment for the trading part is not always necessary. Indeed, as some firms find their competitive advantage from the amount/variety of data and their ability to process and develop better algorithms, latency has become less of a determinant factor. As a result, a separate trading platform can be optional for some firms.

> **"Data is the fuel of a well-performing algorithm, in the past having more data was enough, but now variety – of data type – and speed (ingestion and computation) are essential."**

Data types

Data is the fuel of a well-performing algorithm. In the past having more data was enough, but now variety of data type and speed (ingestion and computation) are essential. Before moving on to designing a robust data infrastructure, one must understand what kind of data is fed into models.

Exchange data

Tick-data and time series coming from the exchanges, they are structured but must be updated as quickly as possible to update the model. The most used data is by far price data, which is highly structured, it is most commonly called tick-data due to its 'tick' format.

In terms of amount, tick data created from exchanges has only been increasing. In 2013, the NYSE averaged half a billion trades and quotes per day, in 2018, there were approximately 4 billion trades and quotes per day with peaks going over 8 billion trades per day. As a result, the financial services industry now receives multiple terabytes of streaming tick data per day which further stresses the data-base infrastructure.

External data

Comes in different forms, it can be historical data extracted from exchanges themselves or from third party vendors such as Bloomberg or Reuters. Alternatively, news data, such as FED (Federal Reserve Board)/Bank of England and even Twitter feeds, is very important to be fed to the model as these can affect price movements greatly. Nevertheless this data often comes as unstructured. Alternative data has made its entry as a differentiator, though due to its amount and diversity, leveraging its value still remains a challenge.

Internal data

The ability to correlate data is key to outperforming the competition and, as such, having data that no one else has is truly the differentiating factor in an algorithm. As a result, large institutions such as banks are looking at correlating sources like customer, credit, fraud and other types of data with trading and market behaviours. This enables pattern identification in market prices, especially with the advent of machine-learning techniques, analysing uncorrelated data can yield a winning trading strategy.

Storage and databases

As a result of the complexity and variety of data, databases must possess the ability to get the data from various sources simultaneously. For algorithmic trading, databases have traditionally needed the ability to rapidly ingest a large number of events into durable storage; process vast amounts of historical data for patterns and trends and deliver real-time analytics. Consequently, choosing the right database is often a choice between consistency, durability and performance

High Frequency Trading

Trading speed used to be the main differentiator for traders. Achieving the lowest latency meant placing trades before the market updated its information. This is commonly referred to as High-Frequency Trading. Accessing data ahead of market participants meant being able to see price changes and reacting accordingly before anyone else could do so. The rise of high-frequency trading has gained much interest and has consequently resulted in higher entry barriers. Indeed, reaching the lowest latency can only be attained by a few, not a majority.

For those few with the financial and technical resources, attaining the lowest level of latency can be achieved by reducing the distance between the exchange and trading platform through co-location, housing the trading platform in the same building as the exchange, usually an Equinix co-location data centre, and directly connecting the systems via cables. Additionally, specific hardware such as InfiniBand high performance, low latency network cabling, network interface cards (NICs) and data processing Units (DPUs) that bypass the server kernel, Graphics Processor Units (GPUs) that optimise many AI based high performance compute (HPC) tasks and field-programmable gate arrays (FPGA).

Where next?

As we have suggested, markets where the focus is almost exclusively on achieving the lowest possible latency, will be dominated by a few large players. They will have deep pockets and can afford to hire the best developers, IT architects and operations staff and work with vendors at the bleeding edge of technology to design and implement their own, very sophisticated, hardware and networks that bear little resemblance to general purpose IT systems.

New Opportunities. New Technologies.

It is becoming clear that time to market with new propositions, new strategies driven by the best data and analytics that don't need the absolute lowest latency offer smaller, newer and more nimble financial trading organisations the ability to compete and thrive in the market. Even the largest organisations can see that parts of their business model don't need such low latencies and can see opportunities for significant costs savings.

While smaller organisations don't want, and probably can't afford the time and cost, to build out and maintain their own infrastructure, the largest players are looking at new technologies and platforms that can remove the costs and challenges of hand-crafting trading systems. The use of the Cloud, long dismissed in financial trading markets, is now being investigated and trialled seriously. Also, the whole idea of developers being able to develop applications without having to worry about the idiosyncrasies of the target infrastructure, which is becoming more common with DevOps methodologies, containerisation and serverless computing models, is something that would be appealing to financial traders if a suitable underlying technology platform was available.

> “Accessing data ahead of market participants meant being able to see price changes and reacting accordingly before anyone else could do so.”

Solution: RUMI from N5 Technologies

N5 Technologies, Inc. and its founders originated from the financial services sector. Drawing from that experience, its technology is designed to power mission critical, real-time risk management solutions, ultra-low latency equity trading systems, eCommerce, and other similarly demanding applications. The company is based in San Francisco, California, and is privately funded.

> **Rumi is a software platform that enables financial service companies involved in market trading to embed rich, real-time analytical data processing directly into their transactional applications.**

What is it?

Rumi is a software platform that enables financial service companies involved in market trading to embed rich, real-time analytical data processing directly into their transactional applications. Specifically, it allows you to develop and run custom developed analytics on and across large volumes and varieties of stored raw and/or pre-analysed data, together with live streaming data, in real-time. The results can then be presented in a contextually relevant manner. Not only does it do this in real-time with very low latency, it also does so within predictable time limits. And finally, Rumi has been designed to support the sort of resiliency that financial markets require to support mission critical processes. In short, it addresses the need of financial service companies to derive business insights in real-time from massive volumes of historical and live data while integrating this capability directly into their customer serving, transactional applications.

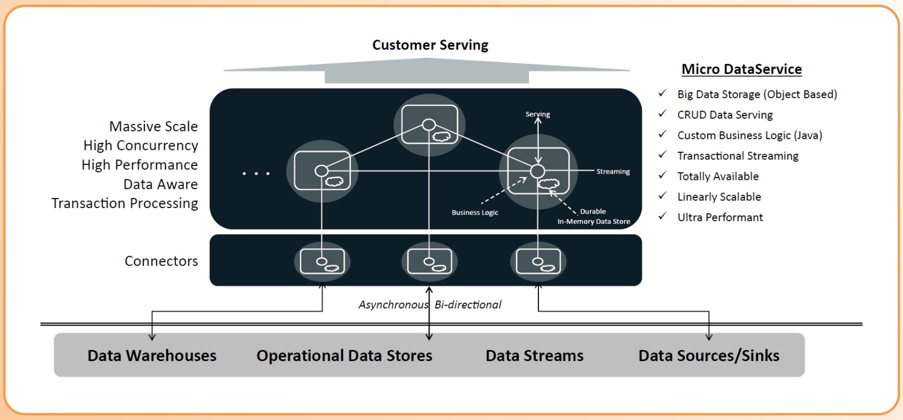

Figure 1 – Micro DataService architecture

Rumi can be purchased and deployed in public and private clouds, on-premises data centres and edge data centres, and can be configured on virtualized or bare-metal infrastructure. It currently supports Java applications, with other language support on the roadmap.

Rumi is also offered as a Managed Service. Known as RumiCloud, N5 has built a whole hardware and software platform that runs on bare metal servers within co-location facilities. Working with key hardware and application partners the managed service packages includes market data feeds, trade execution and a range of value-added services such as analytics, exposure, compliance and positions. Fundamentally, customers just need to focus on building applications without the need to architect the supporting infrastructure.

What does it do?

Rumi's architecture is based on a multi-node, massively scalable and resilient distributed processing system. Each node functions as a fault tolerant, highly available and elastically (and linearly) scalable Micro DataService (see ***Figure 1***). Each service houses data with co-located business logic and publishes integrated telemetry for monitoring and diagnostics. A service is independently capable of big data storage, fast data streaming, CRUD and/or analytical data serving. In other words, it provides comprehensive, native, stream processing capabilities within each and every node.

The stream processing and analytical data serving logic is implemented by the service's business logic. This logic is then activated by stream messages published by upstream nodes and service requests issued by the service's clients. The system is horizontally scaled by sharding Micro DataServices and by deploying multiple concurrently executing Micro DataServices interconnected via fire-and-forget message passing provided natively by Rumi, or over commodity messaging. The overall system of interconnected services is configured and managed as a single distributed deployment. Such a system of interconnected services is called the Micro DataService Fabric (MDF), as shown in ***Figure 2***. Note also that this kind of system can be implemented incrementally using microservices.

The MDF is the key innovation for Rumi, on which the application and data are deployed together for execution. It combines in-memory distributed data storage, data streaming, real-time business logic execution and analytical data serving with a microservices based application architecture. This optimises the platform to enable highly concurrent and scalable big and fast data processing with predictable (and ultra-low) latency performance. NoSQL-based data modelling is used to make it flexible and efficient to support hybrid application access patterns.

Rumi has been designed to process massive volumes and varieties of data in real-time and scale such processing as data volumes grow. Exactly-once processing is built-in. Furthermore, by co-residing application and data together for execution, Rumi provides high performing analytical data processing and serving. It also enables enterprises to rapidly develop and deploy applications that embed analytics with transactional processes. The MDF allows developers to focus on the application's business logic, as complex as needed, whether transactional or analytic or a combination of the two, while the fabric manages all non-functional application details. Moreover, by co-residing application and data together for execution, Rumi provides a single processing environment with a single code base. You can also integrate with AI/ML execution platforms if required.

Why should you care?

The cost and complexity of developing and deploying high-performance trading systems in financial markets can adversely impact margins and the agility required to bring new algorithms into production quickly, potentially damaging competitiveness. Having a managed platform that enables the application developers and data scientists to focus on delivering the most effective trading algorithms, without having to worry about the intricacies and performance of the underlying technology needed to manage the multiple data feeds or support AI/ML inferencing models, is now a credible alternative to in-house developed and maintained systems. Rumi enables you to support all of these from within a single development and deployment environment. Competitive offerings and in-house developments tend to require the use of multiple products, which potentially means significant extra cost and complexity and, in many cases, reduced business agility and competitiveness.

Further to this point, Rumi natively provides a highly available, fault-tolerant and massively scalable environment to support your hybrid applications, so it offers the sort of resiliency that enterprises will require. Moreover, it simplifies application design and development complexity, by depending on the platform to provide and manage enterprise level capabilities. As a result, the development and deployment of applications should be easier and more agile than would otherwise be the case.

Finally, Rumi's architecture has been designed to support high throughput performance combined with (ultra-) low latency. It is more than capable of processing large quantities of streaming data at very high speeds, in large part because it is able to distribute both data and compute responsibilities intelligently across its architecture. This is borne out by a just published benchmark of the Rumi platform that presents the results of latency benchmarks of a core equities trading flow using NVIDIA's Infiniband and Ethernet (RoCE) networking and SuperMicro's flagship 4 socket enterprise server cluster. I'll leave interested readers to view the details of the benchmark tests, but the conclusion is that the solution offered by N5 demonstrates incredibly low latencies and jitter even at extremely high throughputs. In conjunction with the other benefits mentioned, this should have a net positive effect on total cost of ownership.

> **"Rumi has been designed to process massive volumes and varieties of data in real-time and scale such processing as data volumes grow."**

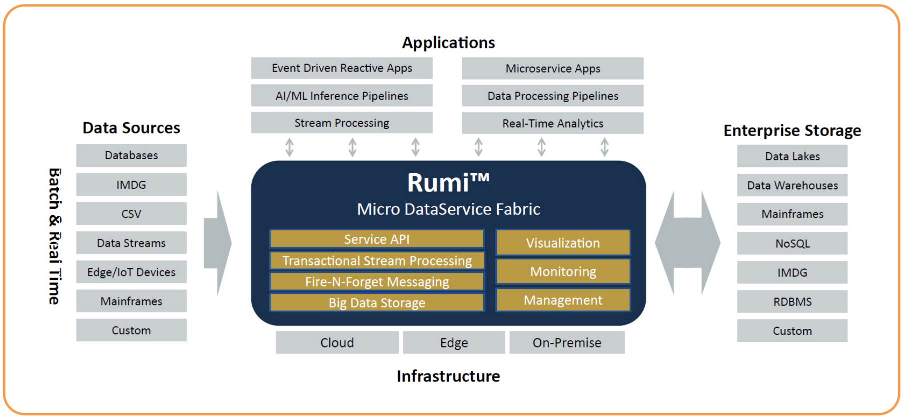

Figure 2 – Micro DataService Fabric architecture

The Bottom Line

Early customer trials and live deployments, with very large transaction volumes in, amongst others, equity trading and funds authorisation, running on-premises or as Amazon Web Services EC2 instances, have shown large scale improvements in latency and very significant cost savings over traditional development and deployment methodologies. It certainly appears to us that this will enable financial trading businesses, who may have feared they were being priced out of new markets by the cost and complexity of the necessary IT infrastructure, to easily and quickly author and deploy applications that exhibit sub-microsecond client-market trading latencies with zero loss across process, machine and network failures and that this will enable them to compete effectively from both a time to market, infrastructure performance and cost point of view.

> **“Early customer trials and live deployments, with very large transaction volumes … have shown large scale improvements in latency and very significant cost savings over traditional development and deployment methodologies.”**

FURTHER INFORMATION
Further information about this subject is available from
www.bloorresearch.com/company/n5-technologies/

About the author

PAUL BEVAN
Navigator, Research Director: IT Infrastructure

Paul has had a 40-year career in industry that started in logistics with a variety of operational management roles. For the last 33 years he has worked in the IT industry, mostly in sales and marketing, covering everything from mainframes to personal computers, development tools to specific industry applications, IT services and outsourcing. In the last few years he has been a keen commentator and analyst of the data centre and cloud world. Until recently he was also a non-executive director in an NHS Clinical Commissioning Group.

Paul has a deep knowledge and understanding about the IT services market and is particularly interested in the impact of Cloud, Software Defined infrastructure, OpenStack, the Open Compute Project and new data centre models on both business users and IT vendors. His mix of business and IT experience, allied to a passionate belief in customer focus and "grown-up" marketing, has given him a particular capability in understanding and articulating the business benefits of technology. This enables him to advise businesses on the impact and benefits of particular technologies and services, and to help IT vendors position and promote their offerings more effectively.

Bloor overview

Technology is enabling rapid business evolution. The opportunities are immense but if you do not adapt then you will not survive. So in the age of Mutable business Evolution is Essential to your success.

We'll show you the future and help you deliver it.

Bloor brings fresh technological thinking to help you navigate complex business situations, converting challenges into new opportunities for real growth, profitability and impact.

We provide actionable strategic insight through our innovative independent technology research, advisory and consulting services. We assist companies throughout their transformation journeys to stay relevant, bringing fresh thinking to complex business situations and turning challenges into new opportunities for real growth and profitability.

For over 25 years, Bloor has assisted companies to intelligently evolve: by embracing technology to adjust their strategies and achieve the best possible outcomes. At Bloor, we will help you challenge assumptions to consistently improve and succeed.

Copyright and disclaimer

Due to the nature of this material, numerous hardware and software products have been mentioned by name. In the majority, if not all, of the cases, these product names are claimed as trademarks by the companies that manufacture the products. It is not Bloor Research's intent to claim these names or trademarks as our own. Likewise, company logos, graphics or screen shots have been reproduced with the consent of the owner and are subject to that owner's copyright.

Whilst every care has been taken in the preparation of this document to ensure that the information is correct, the publishers cannot accept responsibility for any errors or omissions.

Bloor Research International Ltd
20–22 Wenlock Road
LONDON N1 7GU
United Kingdom

tel: **+44 (0)1494 291 992**
web: **www.Bloorresearch.com**
email: **info@Bloorresearch.com**