Sumerian helps global investment bank to optimise FX trading by quantifying relationship between latency and order fill rate

The challenge

- Complex FX trade application within global investment bank, driving multi-billion dollar global revenue stream
- Seeking to gain improved visibility into its latency position - specifically if and how latency impacts fill rates, currency pairs and client streams

Value delivered

- Sumerian quantified relationship between latency and various performance criteria: fill rate, currency pairing, client streams
- For one region, Sumerian forecasted that a 20% reduction in latency could result in 15% more trade orders
- Holistic latency model identified 2 components to focus latency reduction efforts on
- Findings providing valuable intelligence to prove the value of latency reduction to the business and optimise FX trading

Client challenge

The rapid growth of high frequency and algorithmic trading in the competitive Foreign Exchange (FX) market continues to fuel investment in today's low-latency trading platforms. Driving down latency tails through the optimisation of internal application and messaging layers delivers the potential to stream quotes and execute orders without discernible delay – resulting in faster, more competitive trading. However, given the complex, variable nature of latency and the very high volumes of transactions involved in FX trading, it can be extremely difficult for teams to gain quantified visibility on where trading performance is sensitive to latency, and crucially, whether this is resulting in diminishing returns on the order fill rate.

In one such scenario, a global investment bank with a multi-billion dollar revenue stream was interested in exploring this specific challenge. The IT team had an established relationship with Sumerian having used the IQ service to gain improved understanding of its FX quote latencies and - after applying subsequent recommendations raised by the analysis - had achieved a considerable 75% reduction in quote latency.

With the continual high performance of its FX platform critical to its market position and growth, the IT team wanted to explore and widen its understanding of latency even further – namely, to understand if there was a measurable relationship between latency and the order fill rate, and whether certain currency pairs or client streams were more latency sensitive than others. If found, these metrics would provide highly beneficial intelligence back to the business into the value of latency reduction and how it has a direct impact on overall revenue.

Sumerian solution

Sumerian's analytics approach to gaining advanced visibility into latency starts by using a very low touch, highly secure data capture process. By utilising existing data captured from the bank's trading environment, the Sumerian IQ service captured, cleansed and integrated data from across the FX platform's application logs. The unique combination of Sumerian's powerful analytics platform and human expertise enabled a precise end-to-end model of the trading application to be built rapidly and accurately. By then calculating the latencies residing in each component step of the quote and executed trade flows, Sumerian could then analyse the model to determine the deeper level of fill rate intelligence the team required, answering questions such as:

- *How is latency impacting order fill rates across each trading region?*
- *Are certain currency pairs and venues more sensitive to latency?*
- *Are there any differences in fill rates across clients?*

Outcome and results

By providing direct, quantified answers to these questions, Sumerian uncovered a number of findings and recommendations that the team could take forward and act on quickly. For example, it was found that latency varied dramatically between RFQs and streamed quotes and, within each, for specific combinations of currency pair and region (see Fig. 1).

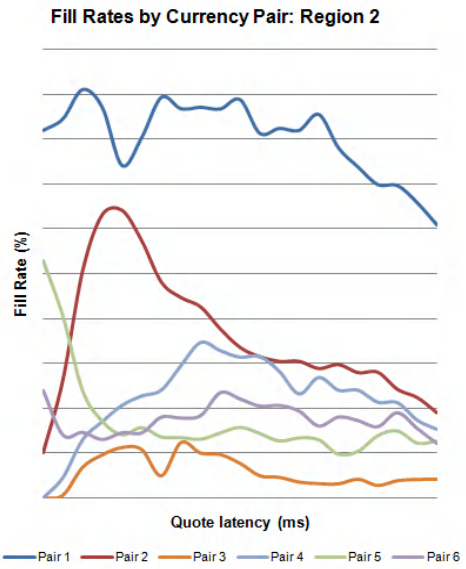

Fig. 1 – Sumerian visualisation showing comparison of currency pairs quote latency and fill rate %

For stream quotes, fill rates were driven by individual quote latency, not average performance across a given time interval. In examining this relationship, 2 regions showed very different profiles, both in terms of the average fill rate response by currency pair and in terms of the fill rate response to latency. One of these regions showed little or no change in fill rates for quote latencies. However, the other region (region 2) showed a noticeable reduction in fill rate as quote latency rose, driven largely by the performance of one particular currency pairing – specifically, when latency increased, the fill rate fell dramatically by ~70%.

By then creating a model of this relationship (see Fig. 2) Sumerian forecasted that if a 10% reduction in latency was made here it would generate an additional 7% trades being executed, with most of this additional volume arising from improvements in the lowest latency band. If latencies were improved even more, by 20% across the board, the bank could potentially see nearly 15% more trades booked for this particular currency pair.

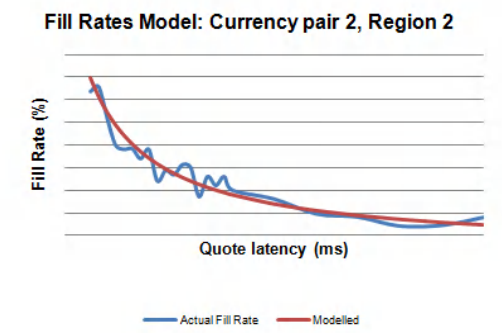

Fig. 2 – Sumerian model visualisation showing the expected fill rate at a given latency

To help the team to target these improvements and focus re-configuration and development in the right places, Sumerian examined the holistic distribution of latency across the FX platform. This step identified that 2 application components were contributing most to the increased latency, and that the greatest returns for improving the situation lay here.

To aid the business' understanding of the efforts being undertaken by the IT team in this space, Sumerian assisted in a presentation workshop, explaining the analysis methodology and its quantified findings. The IT team is now beginning a phase of work to optimise latencies in the areas where currency pairs and venues are more sensitive to latency. Once these adjustments have been made, the bank is planning for Sumerian to re-run the analysis and report on the gains that have been made. In addition, the bank is also seeking to further its use of the analysis to track and trace the causes of trade rejections, enabling an even greater advancement for its FX trading optimisation objectives.

Overall, by using Sumerian to provide a rigorous and progressive analytical approach to its latency reduction efforts, the IT team has the quantified level of visibility it needs to demonstrate measurable, tangible results back to the business – ultimately enhancing the bank's use of IT to increase its competitiveness and FX market position.

More information

For further information on Sumerian or to arrange a demonstration of our services, contact us on 0141 229 7580, e-mail us at **info@sumerian.com** or visit our Web site at **www.sumerian.com**