eBook

In-Memory Computing for Financial Services eBook

Part 1: High-Frequency Trading, Fraud Prevention, and Real-Time Regulatory Compliance for Financial Services

A GridGain Systems In-Memory Computing eBook

August 2017

Table of Contents

Chapter One

Driving High-Frequency Trading with In-Memory Computing

With high-frequency, algorithmic, and quantitative trading becoming the norm for today’s financial services companies, everyone is looking for a technical edge. Companies are racing to beat each other on latency, performance, and analytical complexity. At the same time, they need to maintain transactional-level compliance and risk-management controls. As a result of these requirements, firms engaging in high-frequency trading face unprecedented technical challenges – and they are looking to in-memory computing for answers.

This paper looks at the current state of high-frequency trading – why it’s popular and what types of strategies and technologies are being used – and then explores how in-memory computing can meet the technological challenges and increase profits within this market segment.

What Is High-Frequency Trading?

High-frequency trading is a method of trading securities on the financial markets that involves high-speed, rules-based strategies, and multiple simultaneous trades – with all of the decisions driven by computerized, quantitative models. Basically, this method involves computer programs analyzing the situation in the market and making decisions on the right time to buy, sell, or perform other financial activities. The idea is to predict the market's movements and take actions that will cause you to benefit when those predictions come true – *if* they come true, and if your technology is fast enough to take advantage of the situation before other traders do.

High-frequency trading became very popular in the 2008 to 2009 timeframe, a period of market instability, market volatility, and financial crisis. High-frequency trading strategies work especially well in periods with a lot of price volatility. If prices move quickly and companies can be equally quick about getting market information and executing trades, they can make substantial profits – as Goldman Sachs did with their high-frequency trading in 2008-09.

While high-frequency trading encompasses a variety of distinct strategies, which we’ll discuss in the next section, all of these strategies involve the following basic steps:

1. Obtaining market information – at the fastest speed possible
2. Processing the information through prediction algorithms – at the fastest speed possible
3. Executing trades based on the information — again, at the fastest speed possible
4. Fine-tuning the prediction algorithms based on how they perform

Having the fastest speed possible for the first three steps is crucial – it’s what brings the most wins to traders and makes them more profitable than the rest of the market.

The fourth step, fine-tuning your algorithms based on transaction-cost analysis and back testing, is also extremely important. You need to analyze how well your algorithms performed, based on real-time data from the market, and ask these questions: Can the algorithms be improved? Should something have been done differently?

The first three steps are heavy transactional processes, requiring online transaction processing (OLTP), while the last one is primarily a heavily analytical process, requiring online analytical processing (OLAP). In high-frequency trading, both paradigms, OLTP and OLAP, need to work together for maximum profitability.

The next section takes a closer look at the types of strategies involved in high-speed trading to give a better idea of the type of analytics and transactional speeds that come into play.

Strategies Involved in High-Speed Trading

The general category of high-frequency trading includes a variety of individual strategies, including the following:

- Market making
- Market-taker
- Arbitrage
- Statistical arbitrage
- Pairs trading
- Momentum trading
- Pinging
- News and sentiment trading

Let’s take a closer look at each of these strategies.

Market making. Market-making strategists bet on both sides by placing limits on a sale order at slightly *above* the current market price, while at the same time placing limits on a buy order at slightly *below* the current market price.

This strategy is based on analyzing a security’s *bid-ask spread* where the bid is the price a buyer is willing to pay, and the ask is the price a seller is willing to accept. The wider the spread between bid and ask at a given moment, the more profit can be made. In the world of high-frequency trading, computers are used to quickly identify and act on those advantageous spreads.

If you’re an investor, the size of that spread is going to determine whether your bid or your ask is going to be the most successful. To determine what your spread should be, you need to analyze a lot of data. You need to analyze how other market makers are pricing the securities on the market in which you operate, as well as on other electronic markets. You also need to analyze the volatility of the instrument for which you are making markets, as well as how its price depends on other markets.

If your spreads are better than those of others, you're going to win most of the execution flow, which will bring most of the profit to you. But you need a great deal of analytical speed and transactional speed in order to perform all the necessary analysis and then act quickly in setting up your spread.

Market-taker. These days, in most of the liquid markets, traders either pay a fee to the market or take in a rebate from the market based on how they are affecting liquidity (whether they're taking it or adding it). By analyzing the rebates and setting up a strategy accordingly, a firm can find some ways to benefit. For example, for a given security, they might buy on a market that provides a rebate and sell on one that does not charge a fee, creating a difference in prices that benefits them. We refer to this approach as the market-taker strategy.

Arbitrage. Arbitrage basically involves finding inefficiencies within the fair markets and taking advantage of those inefficiencies. This strategy heavily exploits the idea of *latency*, which describes the time that passes between the electronic sending of a signal and the receipt of the signal. For example, when you send an order out to the exchange, some latency occurs before the exchange receives the order and then fulfills the order.

Being fast in responding to market inefficiencies and placing your order can sometimes bring huge profits. However, with more and more players watching the market for inefficiencies and finding them quickly, there have become fewer arbitrage opportunities. By the time you send your order, it may be too late. As a result, algorithms must now be smarter and more complex in order to extract profitability from market inefficiencies. A successful strategy requires speed and analytics working together.

Statistical arbitrage. This form of arbitrage is much more complex. It involves looking at short-term related securities and making determinations about what is happening in the short-term versus the long-term. Basically, it means trying to come up with a short-term strategy to extract value from knowing how different securities have traded historically versus how they are trading during this short spectrum of time.

This technique uses advanced mathematical strategies and requires both analytical speed and transactional speed, so a combination of OLAP and OLTP approaches is extremely useful. You need to observe the market quickly and make decisions quickly.

Pairs trading. This trading strategy, which is considered to be a type of statistical arbitrage, involves pairing a long-position stock with a short-position stock within the same economic sector or industry. For example, two airlines, such as Delta and United, would share the same cost structure and have similar expanses – oil, airport fees, and so on.

Suppose that, historically, the stock of these two airlines has traded more or less similarly, responding in much the same ways to factors such as oil prices, weather, and general economic prosperity. If you were to spot a difference between the normal stock trajectories of these two companies, you might project that this difference will be temporary – and act to take advantage of it. You could buy more stock in the airline that is currently performing less well, with the expectation that its value will move closer to that of the other airline. At the same time, you could sell the higher-performing airline's stock short, expecting it to decrease.

Sometimes pairs trading is also used to pair a derivative with an underlying asset, because derivatives are traded on one exchange and underlying assets are traded on a different exchange – for example, an options exchange versus an equities exchange or a futures exchange. You may see, historically, that certain derivatives and certain underlying assets typically move together: when the price of one goes up, the price of another goes down, according to a particular ratio. If you're seeing a discrepancy from that ratio in different markets, you can take advantage of that imbalance.

Momentum trading. This is a strategy used to stimulate the market by placing a number of trades on orders for a certain security in a particular direction. Usually, those orders are cancelled in less than a second, and this action triggers competing traders' computers to react by buying or selling. This reaction then creates an artificial price to change the market, and the momentum trader can act against this change.

In certain markets, momentum trading is considered a questionable practice. Critics say that momentum traders should pay for their activities because erroneous information is being sent to the market. However, this strategy has survived the criticism and continues to be used – most successfully in situations where the infrastructure supports speedy information access and fast transaction processing.

Pinging. This strategy is also quite controversial. It involves placing small orders, such as one lot (100 shares), in order to expose large hidden orders that exist on the exchange. Exchanges have traditionally hidden certain types of orders, such as discretionary orders and iceberg orders, which are large single orders divided into smaller lots in order to hide the full quantity. Pinging enables investors to find out if there are any hidden orders sitting on a particular security. By sending some small orders, they can see what happens – what they can detect in the way of liquidity – and how prices are affected.

This strategy requires heavy analytics because you need to make very quick decisions based on what happens after you place the small orders. You have to act very quickly to take advantage of what you've learned before other investors do so.

News and sentiment trading. This is a very recent phenomenon that has become quite popular on Wall Street nowadays. The strategy involves using historical knowledge to anticipate market reactions to news coming out about companies, sectors, countries, industries, or economic indicators. Investors employing this strategy will carefully monitor industries or companies they know to be susceptible to certain types of news or events.

There are feeds available from large market-data providers such as Bloomberg and Reuters that help investors interpret news from the market in ways that can digitally feed into algorithms. The algorithms analyze the news to determine decisions about what to buy or sell immediately based on the sentiment that this news is likely to cause in the market.

This strategy requires significant processing power and speed. You have to understand the sentiment of the news story, analyze how the prices of securities have changed historically based on that type of sentiment, and then make decisions and act upon them very quickly. You want to be the first to act on the news, before others act and change the pricing. Acting that quickly requires very heavy analytical capabilities, OLAP, as well as extremely fast transactional processing power, OLTP.

Technologies Used in High-Speed Trading

To implement the types of strategies we've been discussing for high-frequency trading, you have to have very strong, very fast infrastructure. Currently, the technologies that firms use to create this infrastructure include the following:

- **Dark fiber cabling**. To be very fast, you want to connect your decision engine directly to the backbone of the exchange with fiber-optic cable. That way, as soon as you detect the data, you can make a decision based on it and send your order very quickly to that exchange matching engine. The term *dark fiber* refers to a privately operated optical-fiber infrastructure that you can lease, rather than having to build your own.

- **Exchange co-location**. The closer you can get your trading computers to the exchange's servers, the better your high-frequency trading strategies will work. The ideal situation is to put your decision-making logic right on the same server or data center where your exchange is hosting its matching engine. That way, you will be closest to the liquidity, closest to the source of information, and closest to the processing engine of the orders that you will be sending out.

- **Hardware-based programming logic (FPGAs and GPUs)**. Field programmable gate arrays (FPGAs) and graphics processing units (GPUs) are technologies that involve pre-programming your decision logic onto hardware boards, so that decisions are made using hardware instead of software. While this technology is somewhat more expensive than using software, it can also provide responses that are hundreds of times faster. The faster performance comes from doing most of the processing on the hardware. High-frequency traders have been among the first sectors in the financial services industry to adopt this hardware-based processing technology.

- **Apache® Hadoop™ with MapReduce**. High-frequency trading often requires dealing with large amounts of data – for example, when you're doing historical analysis of data for momentum trading or transactional cost analysis to improve the performance of your algorithms. In such situations, you need a way of expediting access to your data – such as using the programming framework Hadoop. It facilitates rapid data transfer among nodes, using MapReduce data structures to divide applications into numerous small blocks that can be run on any node. With Hadoop, you can expedite both access to data and processing of historical data.

- **Complex Event Processing (CEP)**. This technology takes multiple streams of data coming in at the same time and processes them using advanced mathematical data structures and algorithms, trying to infer patterns or events. In high-frequency trading, CEP is used to analyze market-related data and figure out which trading actions make sense to take. Is it better right now to buy or to sell? If better to buy, how many shares? If better to sell, when and how should it be sold – and how much of it? With CEP, you have a lot of incoming data streams and outgoing decisions that feed the order-management system, the engine that's going to generate orders, and send them out directly to the exchange for execution.

- **Parallel processing clusters**. Traders are also using parallel processing clusters to expedite high-frequency trading, especially when dealing with heavy analytics. You need to expedite the processing and analysis of the data, and a single server usually won't perform adequately for these types of analytics. To handle these situations, companies and firms are using clusters made up of multiple nodes, so that whatever they're processing can be run on multiple nodes at the same time to expedite the results.

- **In-memory computing**. This is one of the newest and most successful technologies being used for high-frequency trading, as well as other Big Data applications. It involves keeping data in memory, instead of on disk, to provide massive improvements in performance and extreme scalability to handle massive sets of data.

Full-featured, in-memory computing platforms, such as the GridGain in-memory computing platform, combine several of the technologies discussed in this section – in-memory computing, data acceleration using Hadoop and MapReduce, complex event processing of multiple data streams, and parallel processing clusters that harness the power of large numbers of computers in a grid.

The next section takes a closer look at how in-memory computing has evolved into a technology that works extremely well for high-speed Big Data use cases such as high-frequency trading.

In-Memory Computing: Benefits and Use Cases

In-memory computing is an essential technology for use cases that involve analyzing deep levels of data, such as level one and level two market data, and use cases that need extremely fast response times. At GridGain, we've found that moving to 100-percent in-memory (RAM) processing typically provides response times that are 1,000 to 1,000,000 times faster than with a traditional disk-based approach.

In addition to putting data in memory for faster access, in-memory computing platforms such as GridGain's leverage clustered memory and the parallel-distributed-processing approach favored in today's high-frequency trading. Building all of that distributed processing and clustered processing into memory makes performance much faster, even compared to software-based parallel processing. Basically, in-memory computing platforms make Big Data become Fast Data.

A top use case for this technology is trading platforms with high-volume transactions, algorithmic trading, and ultra-low latency requirements. GridGain's clients include large banks, hedge funds, and financial technology firms such as Sberbank, Barclays, and Misys who use our technology to process and analyze large amounts of data for decision making. Whenever you are trading, you also have to have certain compliance checks done, and if you do them in-memory, you'll have a significant edge over other players in the market who are still using software-based checks for their compliance controls.

One of GridGain's most noteworthy financial services uses cases is Sberbank, the largest bank in Russia and the third largest in Europe. We helped them significantly improve performance after they analyzed several vendors in the in-memory computing space and found that GridGain provided the best performance results. We were able to generate 1 billion transactions per second using only 10 Dell blades with a combined memory of one terabyte. This system cost about $25,000, which is a significant reduction compared to the days when using in-memory technology for high-frequency trading and algorithmic trading cost millions of dollars.

In a January 2016 article in RBC, the CEO of Sberbank says that they liked GridGain's technology because of excellent performance and reliability, because it's cheaper than the technology that had been using, and because it provides machine-learning and analytical capabilities that they couldn't find from any of the other vendors they looked at.

Meeting the Challenges of High-Frequency Trading

As competition intensifies in the field of high-frequency trading, financial services firms need a new level of transactional speed and analytic power to beat their competition. In-memory computing can provide that speed and power. And GridGain's in-memory computing platform offers a scalable, comprehensive, and affordable solution – an elegant and efficient way to give traders the high-performance edge they need.

Chapter Two

Powering Financial Fraud Prevention with In-Memory Computing

Financial fraud is now a multi-billion-dollar business and growing rapidly, with Juniper Research predicting that online fraud alone will climb from $10.7 billion in 2015 to 25.6 billion in 2020. Failure to detect and prevent fraud can harm the reputations of financial firms and reduce confidence in the industry as a whole.

Protecting their customers from fraud and protecting themselves from fraud-related losses are high priorities for financial institutions. However, fraud prevention is not a simple task, and firms must tackle it simultaneously with other crucial tasks such as ensuring regulatory compliance. To accomplish these data-intensive tasks in a timely manner, financial firms need solutions that are flexible, scalable, reliable, and fast enough to analyze extremely large datasets in real-time.

Fortunately, today's in-memory technologies provide powerful tools for combatting fraud – tools that perform complex processing, modeling, and analysis of big data in real-time. This white paper will discuss what financial fraud is, how firms are addressing the problem, and why in-memory computing technologies such as the GridGain in-memory computing platform are perfectly suited to the task of detecting and stopping fraud wherever it occurs.

Where Financial Fraud Occurs

The term "financial fraud" encompasses a wide variety of illegal practices. Financial fraud typically shows up in the following arenas:

- **Checks** that are written by unauthorized institutions or officers
- **Credit cards** that are stolen and used illegally
- **Mortgages** that are illegally manipulated
- **Corporate financial statements** that are changed or illegally manipulated
- **Securities** that are traded using illegal techniques
- **Payments** that are requested fraudulently or rerouted to improper destinations
- **Identity theft** in which thieves steal financial information or impersonate others in order to make money
- **Forgery** of documents, signatures, banknotes, or works of art to produce financial gain
- **Computerized banking** and computer-based financial transactions employed improperly to produce financial gain
- **Tax evasion** in which corporations or people avoid paying taxes that they owe

The Financial Fraud Research Center at Stanford University estimates that Americans currently lose $50 billion dollars a year to fraudulent practices such as these. The magnitude of this loss provides strong incentives for financial firms to find more effective techniques for fraud prevention.

Evolving Techniques for Fraud Prevention

Traditional approaches to identifying fraud relied on manual verification and analysis. These approaches involved people auditing transactions directly, looking at who was providing passwords and other information. There were few apparent ways to automate this process. However, technology has progressed to the point where firms are now able to collect much more information and employ sophisticated techniques to automatically process and analyze data to detect fraud as it happens.

The techniques that banks and financial services firms currently use to detect fraud include the following:

- **Statistical and multi-channel analysis**: Calculating parameters (such as averages and performance metrics), linking together data from multiple sources (channels), and analyzing correlations between different data measurements to find patterns that help with fraud detection
- **Models and probability distributions**: Calculating models and probability distributions that predict how financial data will behave, so actual data can be compared against predictions and variations can be flagged as potential indications of fraud
- **User profiles**: Computing and maintaining user profiles associating personal data with specific users to help identify atypical behavior or attributes that may be fraudulent or indicate identify theft
- **Real-time algorithmic analysis:** Using algorithms to identify and validate user actions as they occur, in real-time
- **Data clustering and classification:** Analyzing known patterns and profiles and classifying them for use in algorithms and models – essentially creating a data repository for fraud detection
- **Artificial intelligence and machine learning:** Using machine learning techniques such as neural networks to refine automated fraud detection, reducing false positives (false alarms) and improving behavior-based predictions for current transactions and users

Performing these knowledge-intense activities in real-time on extremely large datasets requires high performance and highly scalable technologies, as the next section discusses.

Technologies Used for Fast Data Analysis

To apply the processing and analysis techniques needed for fraud detection in real-time or near real-time on a large scale, financial firms combine these techniques with technologies that can provide fast data analytics in a high intensity, transactional environment.

These technologies include the following:

Big Data. The first step in using financial data for fraud detection is to prepare data and make it available for analysis. Big data technologies provide ways to organize large datasets into multiple pools and connect them in real-time for immediate fraud detection and additional analysis.

Apache™ Hadoop® with MapReduce. Stopping fraudulent transactions in large datasets in real-time requires speed and efficiency. The average speed of executing a transaction may be only milliseconds, and within those single-digit milliseconds, the processing system must analyze the transaction, validate it, and check all available data pools without affecting the performance of processing the transaction. Hadoop with MapReduce is designed to help in situations exactly like these. It organizes hierarchical data to improve performance, allowing quick conclusions as to whether a transaction should be stopped.

Complex Event Processing (CEP) with data streaming. This technology, used in many financial institutions today, involves looking at multiple streams of incoming data and using artificial intelligence (AI) to identify meaningful events, such as potential fraud. It uses neural networks and other AI paradigms to decide how incoming data elements affect the behavior of the system as a whole as transactions are processed.

Near real-time systems. Trying to solve fraud issues after they occur is an expensive strategy and it poses risks to a company's reputation. A much better approach is to stop those transactions while they are happening. This approach requires extremely fast and efficient processing so financial firms are turning to near real-time systems.

Data partitioning and parallel processing clusters. When there are many transactions coming in at the same time, lining them up one by one to check them for fraud is not an option. To operate in real-time and maintain acceptable performance, the system must include multiple processors operating on the data simultaneously – that is, clusters of connected computers processing the data in parallel. It is also important to have the distribution available on the clusters to process those transactions regardless of where they occur, while maintaining data consistency. A system with data partitioning and parallel processing clusters is essential to meet these needs.

Scalable data architecture. We are operating in the world of constantly growing data. Large financial institutions are experiencing 20 to 30 percent data growth year over year, and they cannot risk running out of space. They must be able to add more storage while not losing performance, which means they need a scalable data architecture.

In-memory computing. Combatting fraud is an analytically intense process that uses performance-hungry models and it must be performed in the fastest possible way: using in-memory computing. Because in-memory computing involves keeping data in RAM for extremely fast access, with no disk-related slowdowns, it is faster than any other storage-based computing method.

In the next sections, we will discuss how in-memory solutions such as the GridGain in-memory computing platform have evolved to be fast, affordable, and comprehensive in their ability to combine all of the technologies listed above.

Financial Institutions Using In-Memory Computing

Financial institutions use GridGain for a variety of fraud detection use cases involving high-volume transaction processing and big-data analytics, such as checking for compliance with anti-money-laundering (AML) and "know your customer" (KYC) regulations, looking for market manipulation, or monitoring other regulated areas. They are using complex event processing for real-time or near real-time customer views and analysis of positions, so they require ultra-low latency in real-time or near real-time data processing and analytics.

Banks who have implemented the GridGain in-memory computing platform – including Barclays, Citi, Sberbank, and others – are seeing a measurable difference at the transactional level. They no longer need to export the data to another system for analysis and approval (or disapproval) of the transaction. That model too often involved performance degradation and post-transaction processing delays, with clients unable to complete transactions until all required steps were performed. In contrast, because the GridGain in-memory computing platform does most of the required computing in a distributed and in-memory fashion, it can process transactions with no noticeable slowdown to the clients. Because GridGain verifies the integrity of each transaction before allowing it to go through, the result is a safer environment for clients.

Customer Case Study: Sberbank. One of the most noteworthy GridGain Systems financial services customers is Sberbank, the largest bank in Russia and the third largest in Europe. The company had 130 million customers and had begun to struggle to validate transactions in real-time due to increasingly high volumes. Traffic was becoming too substantial, and traditional legacy systems were struggling to keep up with the transaction processing without slowing down transactions and creating user dissatisfaction. The need existed to be scalable and transact in real-time but this simply was not possible with legacy systems.

Sberbank analyzed more than ten potential solutions from vendors in the in-memory computing space and found that the GridGain in-memory computing platform provided the best performance results, allowing the bank to significantly improve performance. With GridGain, the company was able to generate one billion transactions per second in a test environment using only 10 Dell® blades with a combined memory of one terabyte. This system cost about $25,000, which is a significant reduction compared to the days when using in-memory technology cost millions of dollars.

The GridGain in-memory computing platform also provided several other important capabilities that Sberbank needed, including machine-learning and analytics, scalability, ease of deployment, hardware independence of cluster components, and a rigorous level of transactional consistency. Of particular importance was the ability to conduct integrity checking and rollback on financial transactions. Sberbank could not find that level of consistency with other in-memory computing solutions.

In a January 2016 article in RBC, Herman Gref, the CEO of Sberbank, said that the bank selected the GridGain Systems technology to build “a platform that will enable the bank to introduce new products within hours, not weeks.” He went on to state that the GridGain in-memory computing platform enables Sberbank to provide “unlimited performance and very high reliability” while being “much cheaper” than the technology used previously. Sberbank is using GridGain’s in-memory computing platform to implement capabilities such as “machine learning, flexible pricing, and artificial intelligence”, that could not be provided by the other vendors evaluated – a group that included Oracle®, IBM® and others.

Meeting the Challenges of Real-Time Fraud Prevention

As financial institutions and other companies are inundated with ever-increasing amounts of data to process and analyze for potential fraud, they are looking for high performance and highly scalable ways do so in real-time and near real-time in order to stop fraud before it affects their finances and reputations. Fortunately, in-memory computing solutions can now provide the level of performance and scale these companies need. The GridGain in-memory computing platform offers a scalable, comprehensive, and affordable solution – an elegant and efficient way to stop fraud in its tracks.

Chapter Three

Real-Time Financial Regulatory Compliance

Unprecedented and growing technical challenges face today's financial services organizations. Stringent regulations and client protection initiatives enacted in the wake of the 2008 financial meltdown pose tough requirements for validation of financial transactions. Banks and other financial institutions must monitor, collect, and analyze vast amounts of data from multiple, disparate sources in real-time. Coping with these challenges in an efficient way will require an extremely fast, scalable, and cost-effective data technology.

This paper provides an overview of the current economic and regulatory environment, focusing particularly on new and recent regulatory initiatives. It then discusses how the banking industry is addressing today's daunting challenges with new business strategies and innovative technologies.

Economic Outlook Uncertainty

The current economic outlook is full of uncertainty. There has been a modest amount of economic growth following the 2008 financial crisis, with low inflation and low interest rates. However, there has also been one of the slowest economic recoveries in recorded history.

Both developed markets and emerging markets are demonstrating stagnation and the forecast is pessimistic. It's not clear where economic growth will come from. Worldwide, there are many troubling geopolitical issues — such as Brexit — with financial implications that have yet to be resolved.

In addition to the effects of a weak economy, financial institutions are dealing with significant fallout from the 2008 crisis in the form of regulatory tightening and reduced profits. The ROE (return on equity) profitability measure is now below the normal 10 percent for the top ten global banks. "Easy money" is no longer available as banks allocate reserves for new expenses and requirements.

Post-2008 Banking: Tighter Regulations, Tighter Belts

In the post-2008 environment, the banking industry faces a host of new pressures, including:

- **Regulatory tightening** from governing bodies in the U.S., E.U., APAC and other regions to ensure that banks and overall economies are stable enough to prevent a repeat of the 2008 meltdown

- **Low stock prices** relative to historical values, pressing the banks to act more cautiously

- **Deleveraging pressures** pushing banks to reduce the debt and risky assets on their balance sheets, in order to demonstrate that they pose no risks to the economies in which they operate

- **Penalties and compliance costs** relating to issues going back to the 2009-to-2011 timeframe — and requiring banks to allocate funds for any additional such costs that may arise in the future

- **Litigation costs** for legal actions in the wake of the 2008 crisis, for which fines have exceeded $230 billion — more than the capitalization of some of the largest banks combined

Banks are responding to these pressures in multiple ways. They are cutting costs, restructuring, and optimizing certain business lines. They are selling non-core assets, exiting less profitable activities, and refocusing on more profitable ones. And, finally, they are embracing new technology that can help them operate efficiently in the new regulatory environment.

Before we look at the specifics of this new technology, let's look more closely at the regulatory situation that banks are dealing with today.

Major New Regulations Affecting Banks

Banks today must comply with significant new regulations. In the E.U., these regulations include Basel III and IV, MiFID II, the Net Stable Funding Ratio, and the Culture and Ethics Standard in Banking. In the U.S., they include CCAR, IHC, Enhanced Prudential Standards, Dodd-Frank Living Wills, Basel III, and Enhanced Consumer Protection.

Many of these regulations center around the riskiness of assets that the banks hold in their portfolios. They involve tracking a bank’s constantly changing assets, weighting them by risk level, and evaluating them against acceptable levels of exposure. Other regulations, such as the Culture and Ethics Standard in Banking in the E.U. and the Enhanced Consumer Protection regulations in the U.S., involve monitoring transactions and applying analytics to look for ethical violations.

Let’s look at these regulations in more detail.

Basel III and IV. Under the capital requirements directive in Basel III, E.U. banks must hold reserved capital greater than 8 percent of their risk-weighted assets (RWA). This requirement will be even higher in Basel IV.

Banks must track the values of their assets essentially in real-time to make sure they are meeting these requirements. While reports go to the regulators on a quarterly basis, monitoring must happen on a daily basis, so that banks can respond to changes in the market, changes in the profit and loss statement (P&L), changing prices of various assets within the portfolio, and changing conditions in various geographical regions.

Tracking and analyzing the assets of a large financial institution in real time involves a lot of data — and significant system expense in getting a system like this ready for compliance purposes.

CCAR (Comprehensive Capital Analysis and Review). CCAR is a regulatory framework introduced by the Federal Reserve in order to assess, regulate, and supervise large U.S. banks. As with Basel, it involves looking at all of a bank’s assets and how they are graded in terms of risk. Banks must run special stress tests based on various scenarios — for example, the interest rate growing 10 percent, or the currency rate changing significantly. Banks must analyze how their portfolios are affected by those changes.

The bottom line with CCAR is that the system must be agile enough and scalable enough to perform all of these stress-testing scenarios in more or less real time, so that banks can submit their reports to the Federal Reserve Bank in a timely manner.

IHC (Intermediate Holding Company). An important new requirement in 2016 is that every non-U.S. financial organization having U.S. legal entities with more than $50 billion of assets must have an operating IHC. Such organizations (which include the Deutsche banks, Credit Suisse, UBS and others) must create separate entities that are capitalized and operational and that participate in all of the capital analysis tests that the Federal Reserve Bank runs.

So, even though these are non-U.S. headquartered institutions, they must still participate in tests similar to those that larger U.S. institutions participate in. Many of these banks have been working hard and spending a lot of money to be in compliance with IHC.

MiFID II (Markets in Financial Instruments Directive). This huge E.U. initiative is a second version of standard MiFID, which covers execution requirements for customer orders. MiFID II extends the requirements to cover all instruments traded in Europe today, requiring banks to price markets and perform compliance checks both before and after executing orders. They must look at trends and prices, making sure that the price offered or bid is within a certain threshold of where that security is trading in the market.

In regulating previously unregulated trading facilities — all those alternative marketplaces where OTC securities are traded — MiFID II includes requirements for safety of algorithms and high-frequency trading activities. Orders in these markets will need to satisfy risk controls and compliance checks similar to those for traditional market orders. MiFID II also requires a similar type of governance and coverage of derivatives transactions, under supervision by ESMA, the European Security Market Association.

In addition, MiFID II includes stricter requirements for portfolio management and investment advice. There are new time-stamping requirements for all orders executed by portfolio managers. The goal of the requirements is to capture the following information for use in potential audits: when the intention to buy or sell securities was first demonstrated, when the first contact with the broker was established, when the order was placed, when the order was executed, and when the order was reported.

This data capture takes place in a more rigorous “best execution” scenario than before. Previously, with MiFID, a broker had to evaluate where the best price was and send the order to be executed on that exchange or in that marketplace. With MiFID II, brokers must not only look for the best prices but also capture the environment, as proof that orders are executed on the platform with best price.

The result is that much more data has to be captured and stored someplace — not just the order data, but also all of the information in the market at that particular time of execution. And banks need to be able to retrieve this data quickly. To meet all of MiFID II’s new investor protection and risk-control requirements and still maintain low-latency, microsecond-level results, banks need to invest in systems that can provide the fastest performance possible.

Net Stable Funding Ratio. This regulation, which comes from Basel committee and is obligatory in the USA, requires financial institutions to hold cash someplace to cover potential losses during the year. There is usually an audit by the regulators to determine the quota of cash that must be held, and banks must monitor their risk profiles to see whether the amount they need to hold is over or under that minimum.

Enhanced Prudential Standards. This initiative from the U.S. Federal Reserve requires large financial institutions to monitor risk management across the entire enterprise, not just within individual business units such as wealth management, brokerage, and retail. To meet this mandate, large amounts of data must flow from those individual units to a centralized unit, which will then use its own risk-management analytics to understand the risk-management situation for the enterprise as a whole.

Culture and Ethics Standard in Banking. This initiative from the Financial Stability Board in Basel requires banking institutions to implement analytics to look for unethical transactions such as bribery, money laundering, or conflicts of interest. Banks need a mechanism of capturing such transactions, monitoring them, reporting on them, acting on them, and creating a case around them if necessary. This is yet another situation requiring large amounts of data and real-time analytics.

Dodd-Frank Living Wills. Under the Dodd-Frank Act in the U.S., living wills are mandatory for banks with over $50 billion in assets. Each of these banks must create a trouble-resolution financial plan and file it with the government. If anything changes with respect to its financial situation or financial holdings, the bank must redefine the plan and notify the regulators. Banks are expected to monitor their situations actively and respond promptly to changes over a certain threshold.

Enhanced Consumer Protection Through Dodd-Frank and the Services Directive. Dodd-Frank in the U.S. and the Services Directive in the E.U. protects consumers by requiring banks to do more to validate suspicious transactions. Such validation procedures also protect banks from liability for customer wrongdoing — the fines can be very high in such situations.

Proactive Bank Controls to Avoid Litigation

To respond proactively to the new regulatory environment and avoid further litigation burdens, banks are implementing or strengthening the following types of measures:

- **Anti-money-laundering (AML) controls**. Banks are trying to monitor where money is going — tracking source and destination of transactions, whether amounts are in line with the previous transactions — and do so in real time, prior to approving the transaction. Successful monitoring requires real-time analytics so that the monitoring is not noticeable to consumers yet is still comprehensive enough to protect the bank from any potential suits from the government.

- **Know your customer (KYC) controls**. Banks are now monitoring a large data environment (reporting bureaus, social media, and so on) for the types of information about their customers that they must report and update. They record the information where it will be accessible by various units and systems if the customers perform certain banking activities.

- **Sanctions and cybercrime controls**. Many countries and companies around the world are under sanctions, and financial institutions face serious fines when they don't implement sanctions and cybercrime controls such as creating cross-reference tables that are constantly updated or purchasing specialized software.

- **Anti-fraud, anti-corruption, and anti-bribery controls**. Banks are evaluating every transaction against the customer's historical patterns, and they are making sure that any entities that are potentially on the bribery lists are flagged if there is a transaction going through them. This process involves maintaining the relevant information within their systems and implementing appropriate checks and validations before transactions can be completed.

- **Real-time trade compliance monitoring**. As noted earlier with respect to investment banking and wealth management, trades now require both pre-trade and post-trade compliance checks to be done in real time, rather than allowing post-trade checks to be done later. If certain actions occur — for example, if someone trades a lot of stock right before certain news comes out about that stock — this information must immediately be reported to compliance groups within the banks. At least one bank is also now monitoring voice communications among brokers and clients, checking for words that might indicate issues for concern. The overall result: much data recording and analysis to support active monitoring and actions to stop unethical or problematic transactions.

- **Supervisor accountability controls.** For each trade or transaction process there is a supervisor who must sign off on the process. Supervisors are monitoring the activities of their traders, their tellers, and their portfolio managers — often in real time. If something is not right, the monitoring systems need to alert the supervisors either prior to the compliance officers or simultaneously with compliance officers so that action can be taken right away.

- **Proactive monitoring of high-risk countries**. Client relationships in risky countries and countries with sanctions against them require extra transaction scrutiny. In today's banks, most of this scrutiny is happening electronically, from the capture and analysis of data to the issuing of alerts.

- **Internal ethics controls.** Ethics codes within financial institutions require accountability for conduct issues such as people accessing data they are not supposed to access. Here again, monitoring systems are important for quickly detecting and reporting problems.

Naturally, the need to implement all of these new and increased controls in an efficient and responsive manner is having technology implications.

Technology Implications and Trends

With today's environment, banks need to move away from their traditional modes of operation and make better use of technology. The technical needs of today's banking industry have sparked a number of trends, including the following:

- **Disruption from fintech firms.** Realizing that banks lack experience at addressing electronic issues, new technology companies are moving in on some profitable aspects of the banking business. Areas in which fintech firms have taken some business away from banks include electronic payments, personal finance management, lending, investments, and even core banking.

- **Bank investments in innovation.** As banks try to navigate evolving technologies and retain customers, they are making strategic investments in open source technologies and other areas. They're trying to partner with technology firms and create innovation labs to test new technologies. They know they need to find the right technology to help them deal with big data and perform real-time monitoring, analytics, and transaction processing.

- **Digitalization.** With most of the processes that were historically done at the end of the day or manually now being moved into real time, banks are moving to electronic, automated strategies to replace these traditional processes.

- **Cloud services for core activities.** Banks are investing in cloud services for numerous reasons: scalability, cost savings, better access to data, ease of moving and reusing data, and the need to remove silos and have all of the data in one place. At first, the focus was on private clouds, but banks are now experimenting with hybrid and public clouds as well.

- **Architecture simplification.** Banks are trying to simplify architecture and get rid of legacy systems to move toward faster, more agile technologies that improve time-to-market for new channels and products.

- **Blockchains and distributed ledgers.** Distributed-ledger technologies, such as the blockchains pioneered by Bitcoin, allow a ledger of digital records to be securely distributed across a network and quickly accessed by all computers running a specific protocol. Banks are using these technologies — in conjunction with faster, higher-performance systems for data exchange — to reduce the time needed to execute trades, clear trades, and change the registry on trades.

- **Big data and advanced analytics.** As data input has increased, along with the storage of extremely large datasets, banks have devoted an ever-higher proportion of their data initiatives (now about 72%) to advanced analytics: predictive analysis, data mining, big data/fast data, simulation, optimization, and location-based intelligence. Performing these advanced analytics quickly, particularly with disc-based systems, has become a major challenge. Speed is a challenge at the transactional level as well, as performing big-data related controls can cause customers to notice significantly slower transactions.

- **In-memory computing.** In-memory computing, including in-memory data grids, can address many of the problems financial organizations face in using disc-based systems for big data: performance issues — such as discs being too slow to trade, process events, or perform risk and compliance functions — and the need for scalability. Because in-memory computing allows data to be stored in RAM across a cluster of computers and processed in parallel, it operates hundreds of times faster than traditional computing and allows easy addition of more and more nodes to the cluster.

Chapter Four:

Banking on In-Memory Computing with GridGain and Apache® Ignite™

Addressing the challenges highlighted in this eBook requires speed, scalability, availability, security and flexibility. In short, they need distributed in-memory computing. In-memory computing eliminates the disk access bottleneck that slows down applications built on disk-based databases. An in-memory computing platform enables users to process transactions 1,000 times faster than disk-based solutions and enables scale out to terabytes of in-memory data by adding new nodes to the cluster.

In-memory technology has been around for decades. Until recently, however, the cost of RAM made in-memory computing practical only for the highest value applications. The cost of memory continues to fall, dropping an average of 30 percent per year, which makes in-memory computing platforms economical for a wider range of use cases each year. Gartner projects that the in-memory technology market will grow to $10 billion by the end of 2019, a 22 percent compound annual growth rate.

In-memory computing platforms include key features that are now essential for many financial applications. These features, available in Apache Ignite and the **GridGain in-memory computing platform**, include an in-memory data grids with strong distributed SQL capabilities, in-memory database capabilities, streaming analytics, and native integrations with a variety of other open source projects including Apache® Kafka™, Spark™, Cassandra™ and Hadoop™.

GridGain and Apache Ignite are deployed as an in-memory computing layer between the application and data layers. The products work with any RDBMS, NoSQL or Hadoop database. The **In-Memory Data Grid** with strong distributed SQL capabilities is a key-value store which can replicate and partition data caches across multiple nodes and deliver elastic on-demand scalability. Distributed in-memory ACID transactions are also supported. The Data Grid offers support for all popular RDBMSs, with read-through and write-through and support for write behind. Setup is flexible to address unique use cases.

The **In-Memory Compute Grid** enables distributed parallel processing of resource-intensive compute tasks. It offers adaptive load balancing, automatic fault tolerance, linear scalability, and custom scheduling. Built around a pluggable SPI design, it offers a direct API for Fork-Join and MapReduce processing.

The **Distributed SQL** is horizontally-scalable, fault-tolerant, and ANSI SQL-99 compliant with support for all SQL, DML and DDL commands such as SELECT, UPDATE, INSERT, MERGE, and DELETE queries or CREATE or DROP tables. It is a mature, in-memory solution to supplement or replace a disk-based RDBMS. Geospatial support is built into the product and all the communication to the SQL grid is done through ODBC and JDBC APIs without custom coding.

The optional **Persistent Store** feature in the memory-centric Apache Ignite architecture is a distributed disk store that transparently integrates with GridGain as an optional disk layer. It may be deployed on spinning disks, solid state drives (SSDs), Flash, 3D XPoint or other similar storage technologies. Persistent Store allows organizations to maximize their return on investment by establishing the optimal tradeoff between infrastructure costs and application performance by adjusting the amount of data that is kept in-memory.

The **In-Memory Service Grid** provides control over services deployed on each cluster node and guarantees continuous availability of all deployed services in case of node failures. It can automatically deploy services on node startup, deploy multiple instances of a service, and terminate any deployed service. It is a load-balanced and fault-tolerant way of running and managing services across the grid.

Stream analytics establish windows for processing and run either one-time or continuous queries against these windows. The event workflow is customizable and often used for real-time analytics. Data can be indexed as it is being streamed to make it possible to run extremely fast distributed SQL queries against the streaming data.

In-memory Hadoop acceleration provides easy-to-use extensions to disk-based HDFS and traditional MapReduce, delivering up to 10 times faster performance. GridGain and/or Ignite can be layered on top of an existing disk-based HDFS and used as a caching layer offering read-through and write-through while the GridGain Compute Grid can run in-memory MapReduce.

Conclusion

With the tight regulatory environment and cost pressures that banks are facing today, they need big data technologies that make their risk management, monitoring, and compliance processes much faster and more efficient. Large financial institutions accumulating massive amounts of data need to be able to perform analytics on that data in real time in a cost-conscious manner to ensure a good user experience. Many banks are finding in-memory computing platforms such as GridGain and Apache Ignite to be a key strategy for meeting these challenges.

Additional Resources

For more information about topics covered in this eBook, the following resources are available from the GridGain website:

Datasheet: **The GridGain In-Memory Computing Platform**

Case Study: **Misys Uses GridGain to Enable High Performance, Real-Time Data Processing**

Article: **How Russia's Oldest Bank Found Itself on the Edge of In-Memory Computing in CIO Magazine**

Webinar Recording: **In-Memory Computing to Achieve Real-Time Financial Compliance**

Webinar Recording: **Driving High-Frequency Trading and Compliance with In-Memory Computing**

Webinar Recording: Powering Fraud Prevention with In-Memory Computing

Contact GridGain Systems

To learn more about how GridGain can help your business, please email our sales team at sales@gridgain.com, call us at +1 (650) 241-2281 (US) or +44 (0) 7775 835 770 (Europe), or complete our contact form to have us contact you.

About GridGain Systems

GridGain Systems is revolutionizing real-time data access and processing by offering enterprise-grade in-memory computing solutions built on Apache® Ignite™. GridGain solutions are used by global enterprises in financial, software, eCommerce, retail, online business services, healthcare, telecom and other major sectors. GridGain solutions connect data stores (SQL, NoSQL, and Apache™ Hadoop®) with cloud-scale applications and enable massive data throughput and ultra-low latencies across a scalable, distributed cluster of commodity servers. GridGain is the most comprehensive, enterprise-grade in-memory computing platform for high volume ACID transactions, real-time analytics, and hybrid transactional/analytical processing. For more information, visit gridgain.com.

COPYRIGHT AND TRADEMARK INFORMATION