

Vamsi Talks Tech

Vamsi Chemitiganti's weekly musings on applying Big Data, Cloud, & Middleware technology to solving industry challenges. Published every Friday or Sunday (if I'm very busy). All opinions are entirely my own. I write this blog so my readers don't have to spend money on expensive consultants.

Category: Trade Surveillance

A POV on European Banking Regulation.. MAR, MiFiD II et al

Today's European financial markets hardly resemble the ones from 15 years ago. The high speed of electronic trading, explosion in trading volumes, the diverse range of instruments classes & a proliferation of trading venues pose massive challenges. With all this complexity, market abuse patterns have also become egregious. Banks are now shelling out millions of euros in fines for market abuse violations. In response to this complex world, European regulators thus have been hard at work. They have created rules for surveillance of exchanges with a view to detecting suspicious patterns of trade behavior & increase market transparency. In this blogpost, we will discuss the state of the regulatory raft as well as propose a Big Data led reengineering of techniques of data storage, records keeping & forensic analysis to help Banks meet the same.

A Short History of Market Surveillance Regulation in the European Union..



*The lobby of the European Securities and Markets Authority's (ESMA) headquarters in Paris, France.
Photographer: Balint Porneczi/Bloomberg*

As we have seen in previous posts, firms in typically most riskiest part of Banking – Capital Markets – deal in complex financial products in a dynamic industry. Over the last few years, Capital Markets have been undergoing a rapid transformation – at a higher rate perhaps than Retail Banking or Corporate Banking. This is being fueled by technology advances that produce ever lower latencies of trading, an array of financial products, differing (and newer) market participants, heavy quant based trading strategies and multiple venues (exchanges, dark pools etc) that compete for flow based on new products & services.

The Capital Markets value chain in Europe encompasses firms on the buy side (e.g wealth managers), the sell side (e.g broker dealers) & firms that provide custodial services as well as technology providers who provide platforms for post trade analytics support. The crucial link to all of these is the execution venues themselves as well as the clearing houses. With increased globalization driving the capital markets and an increasing number of issuers, one finds an ever increasing amount of complexity across a range of financial instruments assets (stocks, bonds, derivatives, commodities etc).

In this process, over the last few years the ESMA (European Securities and Markets Authority) has slowly begin to harmonize various pieces of legislation that were originally intend to protect the investor. We will focus on two major

regulations that market participants in the EU now need to conform with. These are the MiFID II (Markets in Financial Instruments Directive) and the MAR (Market Abuse Regulation). While both these regulations have different effective dates, together they supplant the 2003 passage of the original MAD (Market Abuse Directive). The global nature of capital markets ensured that the MAD was outdated to the needs to today's financial system. A case in point is the manipulation of the LIBOR (London Interbank Offered Rate) benchmark & the FX Spot Trading scandal in the UK- both of which clearly illustrated the limitations of dated regulation passed a decade ago. The latter is concerned with the FX (Foreign Exchange) market which is largest yet most liquid financial markets in the world. The turnover approaches around \$5.3 trillion as of 2014 with the bulk of it concentrated in London. In 2014, the FCA (Financial Control Authority) fined several leading banks 1.1 billion GBP for market manipulation. All of that being said, let us quickly examine the two major areas of regulation before we study the downstream business & technology ramifications.

Though we will focus on MiFiD II and MAR in this post, the business challenges and technology architecture are broadly applicable across areas such as Dodd Frank CAT in the US & FX Remediation in the UK etc.

MiFiD, MiFiD II and MAR..

MiFiD (Markets in Financial Instruments Directive) originally started as the investment services directive in the UK in the early 90s. As EU law # (2004/39/EC), it has been applicable across the European Union since November 2007. MiFiD is a cornerstone of the EU's regulation of financial markets seeking to improve the competitiveness of EU financial markets by creating a single market for investment services and activities and to ensure a high degree of harmonised protection for investors in financial instruments. MiFiD sets out basic rules of market participant conduct across the EU financial markets. It is intended to cover market type issues – best execution, equity & bond market supervision. It also incorporates statutes for Investor Protection.

The financial crisis of 2008 (<http://www.vamsitalkstech.com/?p=2758>) led to a demand by G20 leaders to create more safer and resilient financial markets. This was for multiple reasons – ranging from overall confidence in the integrity of the markets to exposures of households & pension funds to these markets to ensuring the availability of capital for businesses to grow. Regulators across the globe thus began to address these changes to create safer capital markets. After extensive work, it has been concluded from a political standpoint and has evolved into two separate areas – MiFiD II & MiFiR. MiFiD II expands on the original MiFID & goes live in 2018 [1], has rules built in that deal with breaching thresholds, disorderly trading and other potential abuse[2].

The FX market is one of the largest and most liquid markets in the world with a daily average turnover of \$5.3 trillion, 40% of which takes place in London. The spot FX market is a wholesale financial market and spot FX benchmarks (also known as “fixes”) are used to establish the relative value of two currencies. Fixes are used by a wide range of financial and non-financial companies, for example to help value assets or manage currency risk.

MiFiD II transparency requirements cover a whole range of organizations in a very similar way including –

1. A range of trading venues including Regulated Markets (RM), Multilateral trading facilities (MTF) & Organized trading facilities (OTF)
2. Investment firms (any entity providing investment services) and the Systematic internalizers (clarified as any firm designated as a market maker or a bank that has an ability net out counterparty positions due to it’s order flow)
3. Ultimately, MiFiD II affects the complete range of actors in the EU financial markets. This covers a range of asset managers, custodial services, wealth managers etc irrespective of where they are based (EU or no-EU)

The most significant ‘*Transparency*’ portion of MiFiD II expands the regime that was initially created for equity instruments in the original directive. It adds reporting requirements for both bonds and derivatives. Similar to the reporting requirements under Dodd Frank, this includes both trade reporting – public

reporting of trades in realtime, and transaction reporting, – regulatory reporting no later than T+1.

Beginning early January 2018[1] when MiFID II goes into effect – both EU firms & regulators will be required to monitor a whole range of transactions as well as store more trade data across the lifecycle. Firms are also required to file Suspicious Transaction Reports (STR) as and when they detect suspicious trading patterns that may connote forms of market abuse.

The goal of the Market Abuse Regulation (MAR) is to ensure that regulatory rules stay in lockstep with the tremendous technological progress around trading platforms especially High Frequency Trading (HFT). The Market Abuse Directive (MAD) complements the MAR by ensuring that all EU member states adopt a common taxonomy of definitions for a range of market abuse. The MAR

Meanwhile, MAR defines inside information & trading with concrete examples of rogue behavior including collusion, ping orders, abusive squeeze, cross-product manipulation, floor/ceiling price pattern, ping orders, phishing, improper matched orders, concealing ownership, wash trades, trash and cash, quote stuffing, excessive bid/offer spread, and ‘pump and dump’ etc.

The MAR went live on July 2016. It’s goal is to ensure that rules keep pace with market developments, such as new trading platforms, as well as new technologies, such as high frequency trading (HFT) and Algorithmic trading. The MAR also requires identification requirements on the trader or algorithm that is responsible for an investment decision.

MiFID II clearly requires that firms have in place systems and controls that monitor such behaviors and are able to prevent disorderly markets.

The overarching intent of both MiFID II & MAR is to maintain investor faith in the markets by ensuring market integrity, transparency and by catching abuse as it happens. Accordingly, the ESMA has asked for sweeping changes across how transactions on a range of financial instruments – equities, OTC traded derivatives

etc – are handled. These changes have ramifications for Banks, Exchanges & Broker Dealers from a record keeping, trade reconstruction & market abuse monitoring, detection & prevention standpoint.

Furthermore, MiFID II enhances requirements for transaction reporting by including venues such as High Frequency Trading , Direct electronic access (DEA) providers & General clearing members (GCM). The reporting granularity has also been extended to identifying the trader and the client across the order lifecycle for a given transaction.

Thus, beginning early 3rd January 2018 when MiFiD II goes into effect , both firms and regulators will be required to capture & report on detailed order lifecycle for trades.

Key Business & Technology Requirements for MiFid II and MAR Platforms..

While these regulations have broad ramifications across a variety of key functions including compliance, compensation policies, surveillance etc- one of the biggest obstacles is technology which we will examine as well as provide some guidance around.

Some of the key business requirements that can be distilled from these regulatory mandates include the below:

- **Store heterogeneous data** – Both MiFiD II and MAR mandate the need to perform trade monitoring & analysis on not just real time data but also historical data spanning a few years. Among others this will include data feeds from a range of business systems – trade data, valuation & position data, reference data, rates, market data, client data, front, middle & back office, data, voice, chat & other internal communications etc. To sum up, the ability to store a range of cross asset (almost all kinds of instruments), cross format (structured & unstructured including voice), cross venue (exchange, OTC etc) trading data with a higher degree of granularity – is key.

- **Data Auditing** – Such stored data needs to be fully auditable for 5 years. This implies not just being able to store it but also putting in place capabilities in place to ensure strict governance & audit trail capabilities.
- Manage a huge volume increase in data storage requirements (5+ years) due to extensive **Record keeping requirements**
- **Perform Realtime Surveillance & Monitoring of data** – Once data is collected, normalized & segmented, it will need to support realtime monitoring of data (around 5 seconds) to ensure that every trade can be tracked through it's lifecycle. Detecting patterns that could perform surveillance for market abuse and monitor for best execution are key.
- **Business Rules** – Core logic that deals with identifying some of the above trade patterns are created using business rules. Business Rules have been covered in various areas in the blog but they primarily work based on an IF..THEN..ELSE construct.
- **Machine Learning & Predictive Analytics** – A variety of supervised and unsupervised learning approaches can be used to perform extensive **Behavioral modeling & Segmentation** to discover transactions behavior with a view to identifying behavioral patterns of traders & any outlier behaviors that connote potential regulatory violations.
- **A Single View of an Institutional Client**– From the firm's standpoint, it would be very useful to have a single view capability for clients that shows all of their positions across multiple desks, risk position, KYC score etc.

The Design Ramifications of MiFiD II and MAR..

The below post captures the design of a market surveillance system to a good degree of detail. I had originally proposed it in the context of Dodd Frank CAT (Consolidated Audit Trail) Reporting in the US but we will extend these core ideas to MiFiD II and MAR as well. The link is reproduced below for review.

Design & Architecture of a Next Gen Market Surveillance System..(2/2)

This article is the final installment in a two part series that covers one of the most critical issues facing the financial industry – Investor & Market Integrity Protection via Global Market Surveillance. While the first (and previous) post discussed the global scope of the problem across multiple global jurisdictions – this post will discuss a candidate Big Data & ... Continue reading



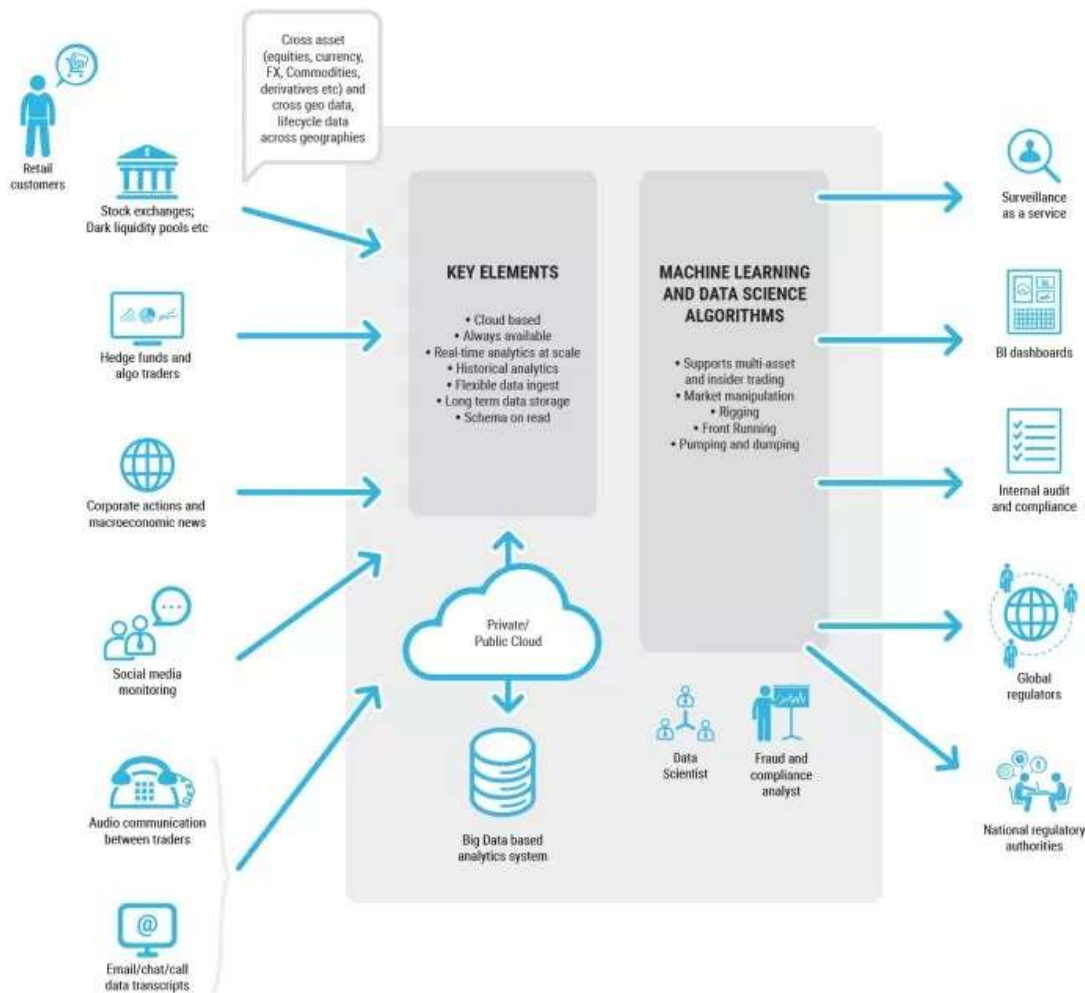
Vamsi Talks Tech

2

Architecture of a Market Surveillance System..

The ability perform deep & multi level analysis of trade activity implies the capability of not only storing heterogeneous data for years in one place as well as the ability to perform forensic analytics (Rules & Machine Learning) in place at very low latency. Querying functionality ranging from interactive (SQL like) needs to be supported as well as an ability to perform deep forensics on the data via Data Science. Further, the ability to perform quick & effective investigation of suspicious trader behavior also requires compliance teams to access and visualize patterns of trade, drill into behavior to identify potential compliance violations. A Big Data platform is ideal for these complete range of requirements.

MARKET SURVEILLANCE SYSTEM



Design and Architecture of a Market Surveillance System for MiFiD II and MAR

The most important technical features for such a system are –

1. Support end to end monitoring across a variety of financial instruments across multiple venues of trading. Support a wide variety of analytics that enable the discovery of interrelationships between customers, traders & trades as the next major advance in surveillance technology. HDFS is the ideal storage repository of this data.
2. Provide a platform that can ingest from tens of millions to billions of market events (spanning a range of financial instruments – Equities, Bonds, Forex, Commodities and Derivatives etc) on a daily basis from thousands of

institutional market participants. Data can be ingested using a range of tools – Sqoop, Kafka, Flume, API etc

3. The ability to add new business rules (via either a business rules engine and/or a model based system that supports machine learning) is a key requirement. As we can see from the above, market manipulation is an activity that seems to constantly push the boundaries in new and unforeseen ways. This can be met using open source languages like Python and R. Multifaceted projects such as Apache Spark allow users to perform exploratory data analysis (EDA), data science based analysis using language bindings with Python & R etc for a range of investigate usecases.
4. Provide advanced visualization techniques thus helping Compliance and Surveillance officers manage the information overload.
5. The ability to perform deep cross-market analysis i.e. to be able to look at financial instruments & securities trading on multiple geographies and exchanges
6. The ability to create views and correlate data that are both wide and deep. A wide view is one that helps look at related securities across multiple venues; a deep view will look for a range of illegal behaviors that threaten market integrity such as market manipulation, insider trading, watch/restricted list trading and unusual pricing.
7. The ability to provide in-memory caches of data for rapid pre-trade & post tradecompliance checks.
8. Ability to create prebuilt analytical models and algorithms that pertain to trading strategy (pre- trade models –. e.g. best execution and analysis). The most popular way to link R and Hadoop is to use HDFS as the long-term store for all data, and use MapReduce jobs (potentially submitted from Hive or Pig) to encode, enrich, and sample data sets from HDFS into R.
9. Provide Data Scientists and Quants with development interfaces using tools like SAS and R.
10. The results of the processing and queries need to be exported in various data formats, a simple CSV/txt format or more optimized binary formats, JSON formats, or even into custom formats. The results will be in the form of standard relational DB data types (e.g. String, Date, Numeric, Boolean).

11. Based on back testing and simulation, analysts should be able to tweak the model and also allow subscribers (typically compliance personnel) of the platform to customize their execution models.
12. A wide range of Analytical tools need to be integrated that allow the best dashboards and visualizations. This can be supported by platforms like Tableau, Qlikview and SAS.
13. An intelligent surveillance system needs to store trade data, reference data, order data, and market data, as well as all of the relevant communication from a range of disparate systems, both internally and externally, and then match these things appropriately. The matching engine can be created using languages supported in Hadoop – Java, Scale, Python & R etc.
14. Provide for multiple layers of detection capabilities starting with a) configuring business rules (that describe a trading pattern) as well as b) dynamic capabilities based on machine learning models (typically thought of as being more predictive). Such a system can also parallelize execution at scale to be able to meet demanding latency requirements for a market surveillance platform.

References..

[1] http://europa.eu/rapid/press-release_IP-16-265_en.htm

[2] http://europa.eu/rapid/press-release_MEMO-13-774_en.htm

November 3, 2016 / Opinion, Trade Surveillance / Leave a comment

Capital Markets Pivots to Big Data in 2016

Previous posts in this blog have discussed how Capital markets firms must create new business models and offer superior client relationships based on their vast data assets.

Firms that can infuse a data driven culture in both existing & new areas of operation will enjoy superior returns and raise the bar for the rest of the industry in 2016 & beyond.

Capital Markets are the face of the financial industry to the general public and generate a large percent of the GDP for the world economy. Despite all the negative press they have garnered since the financial crisis of 2008, capital markets perform an important social function in that they contribute heavily to economic growth and are the primary vehicle for household savings. Firms in this space allow corporations to raise capital using the underwriting process. However, it is not just corporations that benefit from such money raising activity – municipal, local and national governments do the same as well. Just that the overall mechanism differs – while business enterprises issue both equity and bonds, governments typically issue bonds. According to the Boston Consulting Group (BCG), the industry will grow to annual revenues of \$661 billion in 2016 from \$593 billion in 2015 – a healthy 12% increase. On the buy side, the asset base (AuM – Assets under Management) is expected to reach around \$100 trillion by 2020 up from \$74 trillion in 2014.[1]

Within large banks, the Capital Markets group and the Investment Banking Group perform very different functions. Capital Markets (CM) is the face of the bank to the street from a trading perspective. The CM group engineers custom derivative trades that hedge exposure for their clients (typically Hedge Funds, Mutual Funds, Corporations, Governments and high net worth individuals and Trusts) as well as for their own treasury group. They may also do proprietary trading on the banks behalf for a profit – although it is this type of trading that Volcker Rule is seeking to eliminate.

If a Bank uses dark liquidity pools (DLP) they funnel their Brokerage trades through the CM group to avoid the fees associated with executing an exchange trade on the street. Such activities can also be used to hide exchange based trading activity from the Street. In the past, Banks used to make their substantial revenues by profiting from their proprietary trading or by collecting fees for executing trades on behalf of their treasury group or other clients.

Banking and within it, capital markets continues to generate insane amounts of data. These producers range from news providers to electronic trading participants to stock exchanges which are increasingly looking to monetize data. And it is not just the banks, regulatory authorities like the FINRA in the US are processing peak volumes of 40–75 billion market events a day <http://www.vamsitalkstech.com/?p=1157> [2]. In addition to data volumes, Capital Markets has always possessed a variety challenge as well. They have tons of structured data around traditional banking data, market data, reference data & other economic data. You can then factor in semi-structured data around corporate filings, news, retailer data & other gauges of economic activity. An additional challenge now is the creation of data from social media, multimedia etc – firms are presented with significant technology challenges and business opportunities.

Within larger financial supermarkets, the capital markets group typically leads the way in being forward looking in terms of adopting cutting edge technology and high tech spends. Most of the compute intensive problems are generated out of either this group or the enterprise risk group. These groups own the exchange facing order management systems, the trade booking systems, the pricing libraries for the products the bank trades as well as the tactical systems that are used to manage their market and credit risks, customer profitability, compliance and collateral systems. They typically hold about one quarter of a Banks total IT budget. Capital Markets thus has the largest number of use cases for risk and compliance.

Players across value chain on the buy side, the sell side, the intermediaries (stock exchanges & the custodians) & technology firms such as market data providers are all increasingly looking at leveraging these new data sets that can help unlock the value of data for business purposes beyond operational efficiency.

So what are the different categories of applications that are clearly leveraging Big Data in production deployments.

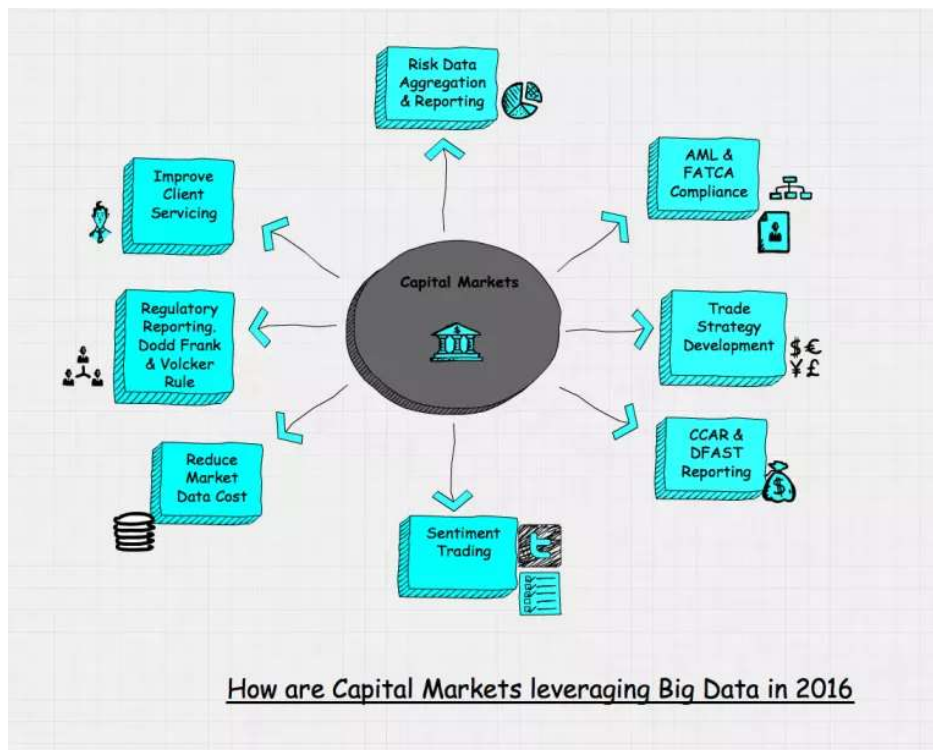


Illustration – How are Capital Markets leveraging Big Data In 2016

I have catalogued the major ones below based on my work with the majors in the spectrum over the last year.

1. **Client Profitability Analysis or Customer 360 view:** With the passing of the Volcker Rule, the large firms are now moving over to a model based on flow based trading rather than relying on prop trading. Thus it is critical for capital market firms to better understand their clients (be they institutional or otherwise) from a 360-degree perspective so they can be marketed to as a single entity across different channels—a key to optimizing profits with cross selling in an increasingly competitive landscape. The 360 view encompasses defensive areas like Risk & Compliance but also the ability to get a single view of profitability by customer across all of their trading desks, the Investment Bank and Commercial Lending.
2. **Regulatory Reporting – Dodd Frank/Volcker Rule Reporting:** Banks have begun to leverage data lakes to capture every trade intraday and end of day across it's lifecycle. They are then validating that no proprietary trading is occurring on on the banks behalf.

3. **CCAR & DFast Reporting**: Big Data can substantially improve the quality of raw data collected across multiple silos. This improves the understanding of a Bank's stress test numbers.
4. **Timely and accurate risk management**: Running Historical, stat VaR (Value at Risk) or both to run the business and to compare with the enterprise risk VaR numbers.
5. **Timely and accurate liquidity management**: Look at the tiered collateral and their liquidity profiles on an intraday basis to manage the unit's liquidity. They also need to look at credit and market stress scenarios and be able to look at the liquidity impact of those scenarios.
6. **Timely and accurate intraday Credit Risk Management**: Understanding when & if deal breaches a tenor bucketed limit before they book it. For FX trading this means that you have about 9 milliseconds to determine if you can do the trade. This is a great place to use in memory technology like Spark/Storm and a Hadoop based platform. These usecases are key in increasing the capital that can be invested in the business. To do this they need to convince upper management that they are managing their risks very tightly.
7. **Timely and accurate intraday Market Risk Management**: Leveraging Big Data to market risk computations ensures that Banks have a real time idea of any market limit breaches for any of the tenor bucketed market limits.
8. **Reducing Market Data costs**: Market Data providers like Bloomberg, Thomson Reuters and other smaller agencies typically charge a fee each time data is accessed. With a large firm, both the front office and Risk access this data on an ad-hoc fairly uncontrolled basis. A popular way to save on cost is to negotiate the rights to access the data once and read it many times. The key is that you need a place to put it & that is the Data Lake.
9. **Trade Strategy Development & Backtesting**: Big Data is being leveraged to constantly backtest trading strategies and algorithms on large volumes of historical and real time data. The ability to scale up computations as well as to incorporate real time streams is key to
10. **Sentiment Based Trading**: Today, large scale trading groups and desks within them have begun monitoring economic, political news and social media data to identify arbitrage opportunities. For instance, looking for correlations between news in the middle east and using that to gauge the price of crude oil in the

futures space. Another example is using weather patterns to gauge demand for electricity in specific regional & local markets with a view to commodities trading. The realtime nature of these sources is information gold. Big Data provides the ability to bring all these sources into one central location and use the gleaned intelligence to drive various downstream activities in trading & private banking.

11. **Market & Trade Surveillance**: Surveillance is an umbrella term that usually refers to a wide array of trading practices that serve to distort securities prices thus enabling market manipulators to illicitly profit at the expense of other participants, by creating information asymmetry. Market surveillance is generally out by Exchanges and Self Regulating Organizations (SRO) in the US – all of which have dedicated surveillance departments set up for this purpose. However, capital markets players on the buy and sell side also need to conduct extensive trade surveillance to report up internally. Pursuant to this goal, the exchanges & the SRO's monitor transaction data including orders and executed trades & perform deep analysis to look for any kind of abuse and fraud.
12. **Buy Side** (e.g. Wealth Management) – A huge list of usecases I have catalogued here – <https://dzone.com/articles/the-state-of-global-wealth-management-part-2-big-d>
13. **AML Compliance** – Covered in various blogs and webinars.
<http://www.vamsitalkstech.com/?s=AML>
<https://www.boozallen.com/insights/2016/04/webinar-anti-money-laundering>
—

The Final Word –

A few tactical recommendations to industry CIOs:

- Firstly, capital markets players should look to create centralized trade repositories for Operations, Traders and Risk Management. This would allow consolidation of systems and a reduction in costs by providing a single platform to replace operations systems, compliance systems and desk centric risk systems. This would eliminate numerous redundant data & application silos,

simplify operations, reduce redundant quant work, improve and understanding of risk.

- Secondly, it is important to put in place a model to create sources of funding for discretionary projects that can leverage Big Data.
- Third, Capital Markets groups typically have to fund their portion of AML, Dodd Frank, Volcker Rule, Trade Compliance, Enterprise Market Risk and Traded Credit Risk projects. These are all mandatory spends. After this they typically get to tackle discretionary business projects. Eg- fund their liquidity risk, trade booking and tactical risk initiatives. These defensive efforts always get the short end of the stick and are not to be neglected while planning out new initiatives.
- Finally, an area in which a lot of current players are lacking is the ability to associate clients using a Lightweight Entity Identifier (LEI). Using a Big Data platform to assign logical and physical entity ID's to every human and business the bank interacts can have salubrious benefits. Big Data can ensure that firms can do this without having to redo all of their customer on-boarding systems. This is key to achieving customer 360 views, AML and FATCA compliance as well as accurate credit risk reporting.

It is no longer enough for CIOs in this space to think of tactical Big Data projects, they must be thinking around creating platforms and ecosystems around those platforms to be able to do a variety of pathbreaking activities that generate a much higher rate of return.

References –

[1] “The State of Capital Markets in 2016” – BCG Perspectives

June 17, 2016 / Capital Markets, Trade Surveillance / Analytics, capital markets, Machine Learning, Money Laundering, Risk, Single View / Leave a comment

Design & Architecture of a Next Gen Market Surveillance System..(2/2)

This article is the final installment in a two part series that covers one of the most critical issues facing the financial industry – Investor & Market Integrity Protection via Global Market Surveillance. While the first (and previous) post discussed the global scope of the problem across multiple global jurisdictions – this post will discuss a candidate Big Data & Cloud Computing Architecture that can help market participants (especially the front line regulators – the Stock Exchanges themselves) & SROs (Self Regulatory Authorities) implement these capabilities in their applications & platforms.

Business Background –

The first article in this two part series laid out the five business trends that are causing a need to rethink existing Global & Cross Asset Surveillance based systems.

To recap them below –

1. The rise of trade lifecycle automation across the Capital Markets value chain and the increasing use of technology across the lifecycle contributes to an environment where speeds and feeds are contributing to a huge number of securities changing hands (in huge quantities) in milliseconds across 25+ global venues of trading; automation leads to increase in trading volumes which adds substantially to the increased risk of fraud
2. The presence of multiple avenues of trading (ATF – alternative trading facilities and MTF – multilateral trading facilities) creates opportunities for information and price arbitrage that were never a huge problem before in terms of multiple markets and multiple products across multiple geographies with different regulatory requirements. This has been covered in a previous post in this blog at –
<http://www.vamsitalkstech.com/?p=412>
3. As a natural consequence of all of the above – (the globalization of trading where market participants are spread across multiple geographies) it makes it all the more difficult to provide a consolidated audit trail (CAT) to view all activity under a single source of truth ;as well as traceability of orders across

those venues; this is extremely key as fraud is becoming increasingly sophisticated e.g the rise of insider trading rings

4. Existing application (e.g ticker plants, surveillance systems, DevOps) architectures are becoming brittle and underperforming as data and transaction volumes continue to go up & data storage requirements keep rising every year. This leads to massive gaps in compliance data. Another significant gap is found while performing a range of post trade analytics – many of which are beyond the simple business rules being leveraged right now and now increasingly need to move into the machine learning & predictive domain. Surveillance now needs to include non traditional sources of data e.g trader email/chat/link analysis etc that can point to under the radar rogue trading activity before that causes the financial system huge losses. E.g. the London Whale, the LIBOR fixing scandal etc
5. Again as a consequence of increased automation, backtesting of data has become a challenge – as well as being able to replay data across historical intervals. This is key in mining for patterns of suspicious activity like bursty spikes in trading as well as certain patterns that could indicate illegal insider selling

The key issue becomes – how do antiquated surveillance systems move into the era of Cloud & Big Data enabled innovation as a way of overcoming these business challenges?

Technology Requirements –

An intelligent surveillance system needs to store trade data, reference data, order data, and market data, as well as all of the relevant communications from all the disparate systems, both internally and externally, and then match these things appropriately. The system needs to account for multiple levels of detection capabilities starting with a) configuring business rules (that describe a fraud pattern) as well as b) dynamic capabilities based on machine learning models (typically thought of as being more predictive). Such a system also needs to parallelize execution at scale to be able to meet demanding latency requirements for a market surveillance platform.

The most important technical essentials for such a system are –

1. Support end to end monitoring across a variety of financial instruments across multiple venues of trading. Support a wide variety of analytics that enable the discovery of interrelationships between customers, traders & trades as the next major advance in surveillance technology.
2. Provide a platform that can ingest from tens of millions to billions of market events (spanning a range of financial instruments – Equities, Bonds, Forex, Commodities and Derivatives etc) on a daily basis from thousands of institutional market participants
3. The ability to add new business rules (via either a business rules engine and/or a model based system that supports machine learning) is a key requirement. As we can see from the first post, market manipulation is an activity that seems to constantly push the boundaries in new and unforeseen ways
4. Provide advanced visualization techniques thus helping Compliance and Surveillance officers manage the information overload.
5. The ability to perform deep cross-market analysis i.e. to be able to look at financial instruments & securities trading on multiple geographies and exchanges e.g.
6. The ability to create views and correlate data that are both wide and deep. A wide view will look at related securities across multiple venues; a deep view will look for a range of illegal behaviors that threaten market integrity such as market manipulation, insider trading, watch/restricted list trading and unusual pricing.
7. The ability to provide in-memory caches of data for rapid pre-trade compliance checks.
8. Ability to create prebuilt analytical models and algorithms that pertain to trading strategy (pre-trade models – e.g. best execution and analysis). The most popular way to link R and Hadoop is to use HDFS as the long-term store for all data, and use MapReduce jobs (potentially submitted from Hive or Pig) to encode, enrich, and sample data sets from HDFS into R.
9. Provide Data Scientists and Quants with development interfaces using tools like SAS and R.

10. The results of the processing and queries need to be exported in various data formats, a simple CSV/txt format or more optimized binary formats, JSON formats, or even into custom formats. The results will be in the form of standard relational DB data types (e.g. String, Date, Numeric, Boolean).
11. Based on back testing and simulation, analysts should be able to tweak the model and also allow subscribers (typically compliance personnel) of the platform to customize their execution models.
12. A wide range of Analytical tools need to be integrated that allow the best dashboards and visualizations.

Application & Data Architecture –

The dramatic technology advances in Big Data & Cloud Computing enable the realization of the above requirements. Big Data is dramatically changing that approach with advanced analytic solutions that are powerful and fast enough to detect fraud in real time but also build models based on historical data (and deep learning) to proactively identify risks.

To enumerate the various advantages of using Big Data –

- a) **Real time insights** – Generate insights at a latency of a few milliseconds
- b) A **Single View** of Customer/Trade/Transaction
- c) Loosely coupled yet **Cloud Ready** Architecture
- d) Highly **Scalable yet Cost effective**

The technology reasons why Hadoop is emerging as the best choice for fraud detection: From a component perspective Hadoop supports multiple ways of running **models and algorithms** that are used to find patterns of fraud and anomalies in the data to predict customer behavior. Examples include Bayesian filters, Clustering, Regression Analysis, Neural Networks etc. Data Scientists & Business Analysts have a choice of MapReduce, Spark (via Java,Python,R), Storm etc and SAS to name a few – to create these models. Fraud model development, testing and deployment on fresh & historical data become very straightforward to implement on Hadoop. The last few releases of enterprise Hadoop distributions

(e.g. Hortonworks Data Platform) have seen huge advances from a Governance, Security and Monitoring perspective.

A shared data repository called a Data Lake is created, that can capture every order creation, modification, cancelation and ultimate execution across all exchanges. This lake provides more visibility into all data related to intra-day trading activities. The trading risk group accesses this shared data lake to processes more position, execution and balance data. This analysis can be performed on fresh data from the current workday or on historical data, and it is available for at least five years—much longer than before. Moreover, Hadoop enables ingest of data from recent acquisitions despite disparate data definitions and infrastructures. All the data that pertains to trade decisions and trade lifecycle needs to be made resident in a general enterprise storage pool that is run on the HDFS (Hadoop Distributed Filesystem) or similar Cloud based filesystem. This repository is augmented by incremental feeds with intra-day trading activity data that will be streamed in using technologies like Sqoop, Kafka and Storm.

The above business requirements can be accomplished leveraging the many different technology paradigms in the Hadoop Data Platform. These include technologies such as enterprise grade message broker – Kafka, in-memory data processing via Spark & Storm etc.

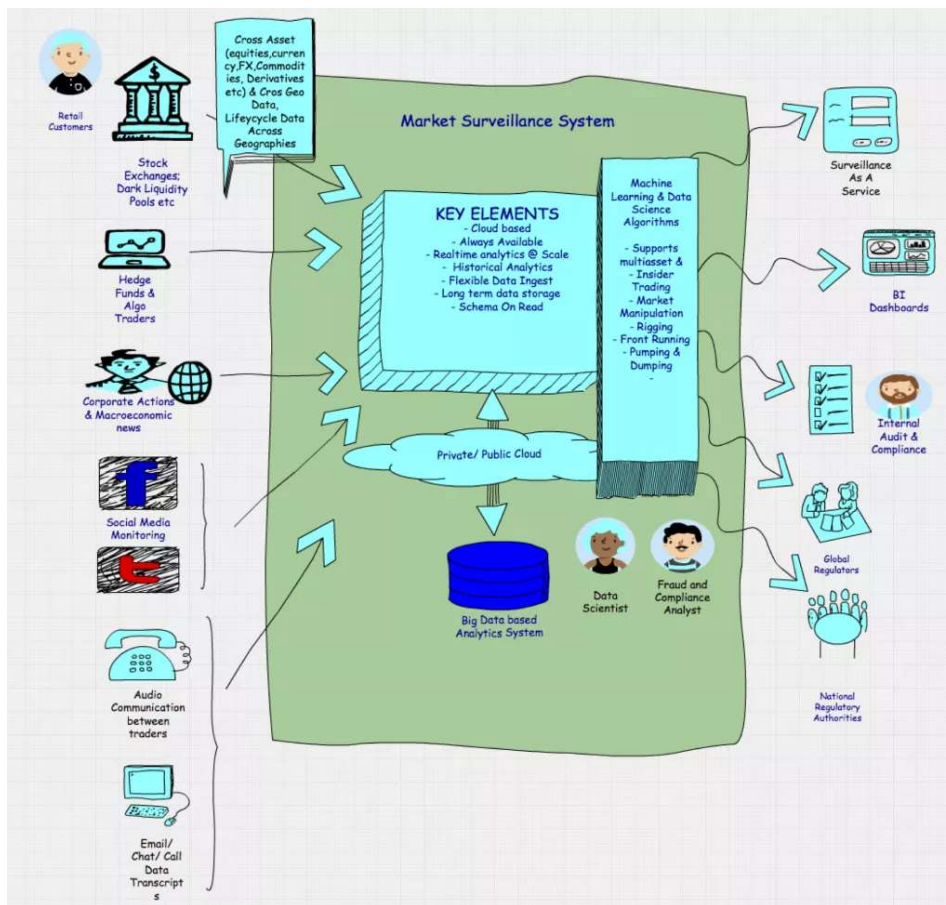


Illustration : Candidate Architecture for a Market Surveillance Platform

The overall logical flow in the system –

- Information sources are depicted at the left. These encompass a variety of institutional, system and human actors potentially sending thousands of real time messages per second or sending over batch feeds.
- A highly scalable messaging system to help bring these feeds into the architecture as well as normalize them and send them in for further processing. Apache Kafka is chosen for this tier. Realtime data is published by Payment Processing systems over Kafka queues. Each of the transactions has 100s of attributes that can be analyzed in real time to detect patterns of usage. We leverage Kafka integration with Apache Storm to read one value at a time and perform some kind of storage like persist the data into a HBase cluster. In a modern data architecture built on Apache Hadoop, Kafka (a fast, scalable and

durable message broker) works in combination with Storm, HBase (and Spark) for real-time analysis and rendering of streaming data.

- Trade data is thus streamed into the platform (on a T+1 basis), which thus ingests, collects, transforms and analyzes core information in real time. The analysis can be both simple and complex event processing & based on pre-existing rules that can be defined in a rules engine, which is invoked with Storm. A Complex Event Processing (CEP) tier can process these feeds at scale to understand relationships among them; where the relationships among these events are defined by business owners in a non technical or by developers in a technical language. Apache Storm integrates with Kafka to process incoming data. Storm architecture is covered briefly in the below section.
- HBase provides near real-time, random read and write access to tables (or 'maps') storing billions of rows and millions of columns. In this case once we store this rapidly and continuously growing dataset from the information producers, we are able to do perform super fast lookup for analytics irrespective of the data size.
- Data that has analytic relevance and needs to be kept for offline or batch processing can be handled using the storage platform based on Hadoop Distributed Filesystem (HDFS) or Amazon S3. The idea to deploy Hadoop oriented workloads (MapReduce, or, Machine Learning) to understand trading patterns as they occur over a period of time. Historical data can be fed into Machine Learning models created above and commingled with streaming data as discussed in step 1.
- Horizontal scale-out (read Cloud based IaaS) is preferred as a deployment approach as this helps the architecture scale linearly as the loads placed on the system increase over time. This approach enables the Market Surveillance engine to distribute the load dynamically across a cluster of cloud based servers based on trade data volumes.
- To take an incremental approach to building the system, once all data resides in a general enterprise storage pool and makes the data accessible to many analytical workloads including Trade Surveillance, Risk, Compliance, etc. A shared data repository across multiple lines of business provides more visibility into all intra-day trading activities. Data can be also fed into downstream systems in a seamless manner using technologies like SQOOP, Kafka and Storm.

The results of the processing and queries can be exported in various data formats, a simple CSV/txt format or more optimized binary formats, json formats, or you can plug in custom SERDE for custom formats. Additionally, with HIVE or HBASE, data within HDFS can be queried via standard SQL using JDBC or ODBC. The results will be in the form of standard relational DB data types (e.g. String, Date, Numeric, Boolean). Finally, REST APIs in HDP natively support both JSON and XML output by default.

- Operational data across a bunch of asset classes, risk types and geographies is thus available to risk analysts during the entire trading window when markets are still open, enabling them to reduce risk of that day's trading activities. The specific advantages to this approach are two-fold: Existing architectures typically are only able to hold a limited set of asset classes within a given system. This means that the data is only assembled for risk processing at the end of the day. In addition, historical data is often not available in sufficient detail. HDP accelerates a firm's speed-to-analytics and also extends its data retention timeline
- Apache Atlas is used to provide governance capabilities in the platform that use both prescriptive and forensic models, which are enriched by a given businesses data taxonomy and metadata. This allows for tagging of trade data between the different businesses data views, which is a key requirement for good data governance and reporting. Atlas also provides audit trail management as data is processed in a pipeline in the lake
- Another important capability that Hadoop can provide is the establishment and adoption of a lightweight entity ID service – which aids dramatically in the holistic viewing & audit tracking of trades. The service will consist of entity assignment for both institutional and individual traders. The goal here is to get each target institution to propagate the Entity ID back into their trade booking and execution systems, then transaction data will flow into the lake with this ID attached providing a way to do Customer & Trade 360.
- Output data elements can be written out to HDFS, and managed by HBase. From here, reports and visualizations can easily be constructed. One can optionally layer in search and/or workflow engines to present the right data to the right business user at the right time.

The Final Word [1] –

We have discussed FINRA as an example of a forward looking organization that has been quite vocal about their usage of Big Data. So how successful has this approach been for them?

The benefits Finra has seen from big data and cloud technologies prompted the independent regulator to use those technologies as the basis for its proposal to build the Consolidated Audit Trail, the massive database project intended to enable the SEC to monitor markets in a high-frequency world. Over the summer, the number of bids to build the CAT was narrowed down to six in a second round of cuts. (The first round of cuts brought the number to 10 from more than 30.) The proposal that Finra has submitted together with the Depository Trust and Clearing Corporation (DTCC) is still in contention. Most of the bids to build and run the CAT for five years are in the range of \$250 million, and Finra's use of AWS and Hadoop makes its proposal the most cost-effective, Randich says.

References –

[1] <http://www.fiercefinanceit.com/story/finra-leverages-cloud-and-hadoop-its-consolidated-audit-trail-proposal/2014-10-16>

November 18, 2015 / Cloud, Trade Surveillance / Trade Surveillance / 2 Comments

Vamsi Talks Tech / Proudly powered by WordPress