Navigating the Hype
What’s the big deal about big data?
There is a lot of big talk about “big data” these days. The financial services industry is awash in stories about machine learning and artificial intelligence (AI) applications. Definitions vary widely and discussions are often vague, so it can be hard to determine what is real and what is “spin.” Without consistency in what these terms mean, it can be a challenge understanding how these advances in processing power have changed the way algo engines work, or enhanced the tools that deliver analytics and insights.

What follows is a series of FAQs about the subject of big data, and a glossary of terms that may be useful in deciphering the jargon. We also offer a few best practices related to the decision making surrounding a big data program.

**What are the criteria that define big data?**

**Is there a standard definition?**

The 3Vs: Volume, Velocity, and Variety, are often used to differentiate simple data from big data. Any big data project would factor in these criteria, but we would include a 4th V: Value. Value refers to the quality of the data, as well as the quality of its return on investment (ROI) relative to how it is being used. This is an element we believe all constituencies should actively evaluate, define, and build into their programs.

**THE “4Vs”**

**Volume**

Volume is a function of the depth and breadth of data and their sources. The term “alternative data” means non-traditional sources that can now be applied to quantitative analysis. We see many more sources, as well as larger quantities of data, and possibly also greater frequency and/or lower-latency real-time increments of data—all of which combine to dramatically increase the overall volume of data that has to be captured, stored, processed, and analyzed.

**Velocity**

With new sources of data such as social media, machine data, and mobile applications streaming into the ecosystem in real time, support for high velocity extends to not only how swiftly data is captured and collected, but all the way through the process until the application of that data drives a resulting business action. The time horizons between data capture and results have been massively compressed, especially for industries with business models that critically depend upon low-latency capabilities and capacity, such as financial services.

**Variety**

Data variety refers to the many sources and types of data being consumed, both structured and unstructured.

- **Structured data:** This includes the typical market or tick data and transaction reference data that traders or quants have contended with for years. These data sets have predetermined formats that are designed to fit into systems analyzed easily.

- **Unstructured data:** Examples of unstructured data include social, sentimental, and voice data. You can find drastic variations between these data points, and they will need to be constructed into a machine-readable format for analysis (becoming structured data). Trade emails, voice, and IM data are good examples of what is captured for compliance and risk analysis.

**Value**

The final V—Value—is by far the most important. It characterizes the potential ROI and strategic impact of big data to your day-to-day business activities or organization. As we think about the Value factor, the first order of business is to assess the quality of your data. If the content you are collecting is not trustworthy or clean, the entire process is corrupted. More isn’t necessarily more if you cannot be assured that the data being collected is going to add value.

By the same token, if the way in which you are applying the data is not well considered, i.e., if you are not “asking the right questions of that data,” then you will not extract benefit from the process. The age-old story of “garbage in, garbage out” certainly applies to the process of big data management, but we can add a new maxim, as well: bad question, bad result. It’s important to be mindful that big data isn’t a virtue unto itself. Its value lies in its effective application to a specific problem or model.
**NAVIGATING THE HYPE**

Why has the term “big data” become so pervasive across financial services?

So-called “tipping points” occur in all technological evolutions, the point at which advanced technology becomes available to a wide enough group of users at a time when it is truly needed. In financial services, market complexity, compressed client commission pools, and intensifying regulation have converged with a dramatic expansion of available technological capabilities, such as big data, machine learning, and AI. So, a lot of people are talking about these topics, and they are garnering a fair bit of media attention. Organizations of all types, across every industry, continually seek new ways to gain efficiencies or find a competitive edge. Financial institutions in particular understand the value of technology in better meeting the needs and demands of clients and regulators. And like other sectors, Wall Street is also looking to leverage data to enhance their clients’ experience, information security, fraud prevention, and regulatory compliance.

Why doesn’t everyone use big data analysis?

A tall order.

Big data digitizes the sheer volume of information that is being produced globally and synthesizes that information to deliver benefits, improve efficiencies, or advance an organization’s goals. That’s a rather tall order. To do this, an enterprise must first be able to develop strategies, operations, and the right resources to plan and manage the logistics of all this information.

Requires enterprise-wide change.

There are several imperatives:

- A cultural change needs to happen. Traditional financial firms are not set up to take advantage of the data that’s available today. They must adopt new mindsets and skill sets in order to realize the benefits that new technology can bring.
- You must have a modern data platform in place to support your big data strategy across the enterprise. Some financial institutions lack the systems and technologies to integrate siloed data and model data to produce insights that they can incorporate into their operations.
- Traders and analysts need to be comfortable and effective in applying new techniques. Workflow needs to change, along with their tools and strategies.

Committing to these imperatives is not easy. It requires a change in your business model that aligns the organizational structure, your processes, and technology to create a robust, secure, and scalable data management infrastructure. It also requires having uniquely talented people—not only data scientists, but IT and business people—who know how to pursue the 4Vs and ask the right questions. This mix of talent can be difficult to find, especially when so much of this technology is still new.


The upfront investment of time and resources can be a challenge to firms of all types. Depending upon the nature, size, and mission of a firm, committing to a major, long-term investment such as a big-data project can be hard to sell to executive management and boards, since quantifying the tangible benefits and understanding the timeline for reaping the ROI can require a leap of faith.

**Where does the data come from?**

Data can come from various sources:

- Internal data systems, including client relationship management platforms (CRMs), execution management and order management platforms (EMS/OMS), tick database, and risk systems.
- External data feeds for securities prices and news (exchange feeds, the SIP, etc.).
- Internal systems that may be generating machine data.
- Alternative data from third-party and public domain sources. An alternative data strategy could include non-financial data, such as location and foot traffic, energy supply and demand, weather, sensors, patent filings, satellite imagery, or telecommunications-related information. These data sets can provide sources of alpha and uncorrelated signals, but they are time-sensitive and can be costly.

**Big Data: Frequently Asked Questions**

**Why has the term “big data” become so pervasive across financial services?**

**Why doesn’t everyone use big data analysis?**

**Where does the data come from?**

**Big Data Feeds**
What are the challenges related to acquiring, storing, and managing big data?

Information needs to be optimally stored and protected, carefully organized with its usage applications in mind, and set up to be easily accessed and processed in order to provide the kind of value a firm will expect from big data.

To manage and process this data effectively and in compliance with regulations, financial institutions must be able to manage an array of business and technical challenges:

**Inefficient legacy systems**
Some firms face pre-existing technology issues. If they are already struggling with data integration, the ability to incorporate historical data in a real-time environment, or if they don’t have clean mechanisms to add context on the fly that improves trading strategies, they will struggle more with big data facilities.

**Information silos**
Silos of information pose another challenge: data access. If firms operate multiple systems that store data, each system has its own access controls. Silos need to be mitigated or broken down before any significant big data effort can be effectively implemented.

**Data governance**
Some financial firms are still trying to build a sustainable data governance practice to comply with regulatory obligations. Some organizations are using “point” solutions or existing legacy systems, which were not designed to handle the data surge.

Many different internal and external applications
Brokerage firms use numerous external and internal applications. Internally, they have systems to consolidate and analyze market data. Externally, they manage platforms to help customers manage and settle trades. Since many applications access and update common data sets, it’s critical that information used by these applications is consistent and up-to-date.

Specialized storage needs
Storing big data sets is very different from the simpler storage mechanisms of the past. An onsite server room won’t cut it. The most challenging of all data going forward, including data from internal sources, is unstructured data, which comes in different forms with various levels of accuracy, security, and quality. This exposes organizations to new risks and regulations.

Analytics have always been important to trading. How has big data analysis changed the way data is used in financial services?

Since Instinet launched electronic trading in 1969, technology has been playing an ever-expanding role in the financial sector. Big data is a significant factor in the most recent rapid evolution of electronification. It is pushing the industry to new heights and across functions such as idea generation, analytics, execution, risk management, regulatory compliance, marketing, and client relationship management.

Big data technology has enabled storage and analysis of data sets not possible before. Many firms are putting greater emphasis on new data management platforms that enable them to integrate and deliver data and analytics in real time, rather than using numerous separated analytics engines. Using the latest technologies, data infrastructure, and processing methods to harvest greater intelligence from increasingly higher volumes of data is becoming a competitive necessity.

The existence of big data makes the following more possible:

- **Real-time responsiveness.**
  Incorporating low-latency stimuli from alternative sources into existing strategies and the behavior of live orders.

- **Heuristic capabilities.**
  Going beyond static models.

- **Machine learning.**
  Combining advanced computational analysis and simple automation.

- **Artificial intelligence.**
  Handing over decision-making discretion to the platform.
How has the use of big data and streaming events data impacted the quality of trading and analytics?

Trading advantages are often a function of the precision of their algorithm’s undertaking, in microseconds, to receive, analyze, and act upon new data. Stream processing found its first uses in the financial sector as stock exchanges moved from floor-based trading to electronic trading. Today, as data increase in volume, variety, and velocity, the ability to stream event data quickly, and at scale, has made it easier to assess and manage risk and opportunity in real time.

Real-time stream processing has the ability to ingest multiple high-volume, high-speed streams of transactions from each major exchange, and enable automated high-frequency traders to apply sophisticated tools, such as scoring models, to the data stream and take action based on those scores.

Stream processing is not only being used for trading. It is now prevalent in risk management, regulatory compliance, client service, and fraud detection. These applications provide continual awareness of conditions to predict and act on opportunities and threats exactly in the moment.

**THE BENEFITS OF BIG DATA CAN BE SEEN ACROSS THE ENTIRE LIFECYCLE OF THE INVESTMENT**

**DEVELOP**

**Idea generation**

For fundamental investors, having large and diverse historical data sets available can help to greatly refine the idea generation process. The ability to incorporate real-time streaming data can also assist a portfolio manager or analyst to react more immediately to developments with regard to public securities. If the investor uses “quantimental” or quantitative strategies, the benefits of big data are even more direct. Quantitative analytics and quant models can be dramatically enhanced by the use of the more varied and deeper data sets big data provides.

**Pre-trade analytics**

New pre-trade analytics can be built on top of historical data sets applying the power of big data. Whether the investor is trading equities (portfolios or single stock), derivatives, rates, credit, or FX, there are significant advantages to the incorporation of advanced analytics powered by amplified structured and unstructured data.

**Pre-trade risk management and portfolio risk ratio analysis**

Since this is a highly quantitative process, it’s intuitive to understand why big data can enhance the process of managing risks. The big benefit relates to the ability to apply alternative data sets to a process that has traditionally been limited to financial data sources. Using alternative and unstructured data in the process can alert the investor to issues that may not yet have impacted in-market behavior, but likely will. Those extra incremental microsecond can often make a major difference to short-term alpha and transaction costs.

**Strategy selection**

The opportunity to integrate more and better data into the process of trading strategy selection is clear. Having at your fingertips a quick assessment of how similar orders under similar conditions have traded most efficiently in the past is enormously valuable, and can increase conviction regarding the best course of action and the parameters within which your order should be traded.

**EXECUTE**

**Order entry and routing decisions**

Traders tell their brokers that they want a more streamlined process for managing their algos and electronic order types. What that really means is that they want these order types to be smarter, as well as simpler to use. Enter big data. The effectiveness and efficiency of an order management process is entirely dependent upon the order (be it the parent or the tactical child order) to process information in real time and react accordingly, based on the trader’s goals. The better an order handles data, the more likely it is to be successful in achieving the objective.

**Execution monitoring and management**

There was a step change in electronic trading when the industry was able to incorporate real-time data into the order monitoring and management process. Big data will be yet another step change that will greatly enhance the effectiveness and efficiency of the process.

**Trade execution and reporting**

The point at which an order is executed and reported to the tape offers another set of opportunities to capture meaningful data, both immediately before and for some time period after the execution. The data generated around this point in the investment lifecycle—particularly related to volume and price—is already critically important to the analysis of best execution. When big data capabilities and processes come into play, there is the possibility for even more nuanced insights.

**Post-trade analytics**

Post-trade transaction cost analysis (TCA) can be powered by the same platform where all the execution and market data was captured. A scalable post-trade analytics platform would mean quicker turnaround of new analytics. It also democratizes data, providing researchers and algo developers access to details not easily available before.

**PROCESS**

**Clearing and settlement**

One of the gross costs associated with the investment process relates to the trade clearing and settlement process. What kind of efficiencies may we be able to extract by the application of greater analytics around where, when, and how orders clear and settle?

**Commission management and research procurement**

A good deal of effort goes into the analysis of the net costs of a trade. What was the execution price? What was the transaction cost? How does that cost basis impact the performance of the position? But as more data becomes available, we can better consider the overall gross costs of the investment process, including clearing, settlement, and the costs associated with managing the investment process and acquiring the research that drove the original idea.
Where else is big data being applied in trading?

**New strategy creation**
The process of quantitative research and development is in part focused on the formulation of models and strategies that enhance the order management process such as new order types, algos, and automation strategies. Big data has an immediate impact on a quantitative analyst’s ability to gather intelligence, look for patterns, and back-test ideas to see if they can achieve a client’s execution goals more efficiently and/or effectively.

**Strategy backtesting**
After developing a new trading strategy, traders need to know whether the strategy will work. Big data optimizes backtesting, enabling it to be done as quickly as possible by providing fast access to historical market data and market conditions.

**Mining more signal data**
To identify trading opportunities, traders track signals from a variety of data sources: price data, large volumes of news data, and regulatory filings, as well as web and social media data (e.g., blogs, Twitter, etc.). Combining structured data, such as price, with unstructured data, such as news headlines, has shown great potential to enhance the process of protecting or capturing incremental alpha. For example, some quant traders are analyzing news feeds to help predict movements of markets for everything from government bonds to commodities.

**Exposure and risk metrics**
For risk management, firms are increasingly processing large data sets to obtain an array of risk metrics in real time, usually multiple times during the day. Finding the places where hidden risks are aggregated could save a lot of slippage in the trading process, and assist an investor in better diversifying their positions.

What are some best practices for the use of big data in algorithmic trading?

There are some key decisions and strategic approaches to using big data that can help drive success in algorithmic trading.

**Pick one clear thing to focus on first.**
One best practice is to focus a big data project initially on one specific usage goal. Thoughtfully and simply articulating that first goal is critical. There’s no single starting point that’s best for everyone. Instead, examine your needs (immediate and long-term) and focus on one specific goal or project where you can attain significant ROI. Once you’ve undertaken the first project, it tends to lead to other next steps rather organically.

**Explore new data formats.**
At the outset of each project, begin with a data exploratory process by getting a handle on what data you already have, where it is, who owns and controls it, and how it is currently used. This process will provide a lot of insights. Consider leveraging new data formats. Access to new data formats, including semi-structured, unstructured, and nested data, greatly expands your opportunities to tap the benefits of social media, combine insights from diverse sources, and reduce development time through more efficient workflows.

**Integrate new tools and technologies.**
It is often necessary to source new hardware, software, or many different tools in order to achieve your goals. While these tools can be extremely powerful, setting each of them up, integrating them and keeping them working in harmony can be challenging. Open source projects are not packaged as complete solutions. It may be necessary to get big data expert training and support when needed.

**Continually test your algorithms and your assumptions.**
The laboratory method and continuous testing should become part of your new quantitative process. Cause and effect patterns and probabilistic analysis in an environment like the financial markets can be very hard to get right. It is also important to remember that introducing machine learning, AI, and big data applications that leverage new data sources and massive amounts of data may introduce latency, which might not be desirable depending upon at which point in the trading life cycle they are interposed.
What is the relationship between big data, machine learning, and artificial intelligence?

**Big data** is the core “fuel” that drives technologies like machine learning and AI. These technologies are dependent upon the advanced computational capabilities and characteristics (the 4Vs) of the underlying data.

**Machine learning** is a way of distilling patterns and/or achieving automation that is genuinely heuristic. Instead of writing millions of lines of code with complex rules to perform a task, you can develop technology that can look at a lot of data, recognize patterns, and learn from the data. “Learning” requires feeding huge amounts of data to the algorithms and allowing them to adjust and improve. While machine learning may be considered an evolution or extension of known statistical methods, it requires new data logistics and analytical skill in order to derive signals that are relevant to the investment process, and drive conclusions or actions in a way that delivers against the goals with precision and consistency.

**Artificial intelligence (AI)** is the attempt to build machines that can perform tasks that are characteristic of human intelligence. This includes understanding language, recognizing objects and sounds, learning, and problem solving. AI in financial services puts greater discretion over decision making into the technology versus the human operators. This means that AI offerings are replacing certain aspects of human labor or effort, empowering technology to perform these explicit tasks with a degree of pre-determined control.

Don’t try to replace all human involvement. Businesses need to be mindful of where to apply machine learning and other new data processing technologies. Machine learning is the first step in the integration of big data analytics with automation. This is where technology utilizes big data to learn and respond or adapt. Machine learning still allows for human insights and judgment to drive it forward—it allows you to use automation to recognize patterns, or remove bias, in a way that is faster than what a human could do.

Using artificial intelligence means you are asking the technology to make decisions on your behalf. This level of discretion is something that must be designed and weighed carefully. It’s analogous to the difference between using a GPS guidance tool versus a self-driving car.

At the end of the day, machine learning and other data processing technologies should not seek to replace the experience, judgment, and insight of the human trader, but rather they should amplify his/her capabilities and complement his/her intuition.

**AI AND THE BLACK BOX**

Many AI applications are so complex that even developers of the algorithms are unsure of why a program makes a certain decision. Efforts, driven by compliance and internal objectives, have been underway to better understand what is going into these black boxes, how they are analyzing data, and why decisions are made.

Transparency is needed in other industries, such as digital marketing, which heavily leverages algos and other advanced quantitative approaches to drive digital advertising and better product placement. For example, the EU’s GDPR privacy mandate requires data collection of any algorithmic or machine learning applications that use consumer data for audits.
Big Data: Frequently Asked Questions

What do I need to consider when evaluating a big data initiative?

To get the most value out of your data, these factors are important:

**Start with business needs.**

Big data programs need to be led by key decision makers because they will impact the business model. Develop a big data strategy based on business needs and challenges.

**Focus on IT alignment.**

You need a flexible, high-performance data management infrastructure to support a holistic data analytics strategy. This new data management solution will have the scalability to handle the 4Vs of big data, and support machine learning and streaming analytics. It’s important to eliminate silos.

**Identify new relevant data.**

Engage in a discovery process. Know where to find the data you need and how to use it. Taking the time to understand the sensitivity and classification of the data will prevent counterproductive processes.

**Build a data-centric culture.**

Putting data at the center of the process may be a paradigm shift in some cases. Financial firms will realize value by effectively managing and processing increasing volumes of new and existing data, and putting the right processes and skills in place to better understand their operations and the marketplace as a whole.

**Actively manage data quality.**

The need for high-quality data cuts across all aspects of the trade lifecycle. If the data is not complete, consistent, accurate, or timely, it will not serve the purpose you need or deliver the insights you are anticipating.

**Cultivate a cross-functional team and skills.**

Last but not least, cultivate the right team with the right skills. The right mix of data scientists, data engineers, analytics architects, IT, and business translators are all equally important to the success of big data initiatives, and need to work in concert with one another.

How should the cloud factor into this process?

Today, cloud computing and big data are increasingly working hand-in-hand. The use of cloud computing accelerates a firm’s ability to process and incorporate big data programs. In regulated industries such as financial services, compliance with a wide variety of rules and laws is often cited as a concern with regard to using the cloud. It is possible, however, for the cloud to provide the financial industry with the agility to innovate and provide value for customers, while still meeting compliance standards and maintaining control of the data. Across industries, we see workloads that are increasingly being moved to the cloud to take advantage of cost and operational efficiencies.

**Blue skies.**

When you integrate your data infrastructure with a cloud platform, you can analyze and process vast amounts of data quickly. This means more nimble “go-to-market” applications. For financial services, for example, firms will have more opportunities to launch new front-office applications quickly, and automate back-office processes and improve human productivity, while reducing errors.

Research has shown that firms that put a data strategy in place, and leverage the right big data and cloud technologies to effectively implement their strategy, are able to capitalize more quickly on the insights their data can unlock.

**Cloud control.**

Despite all of the benefits the cloud brings, migrating your analytics and workloads to the cloud is complex and can negatively impact business operations if not well planned and executed.

Some data and workloads may need to remain on-premises for security or compliance reasons. Maintaining security is a shared responsibility between your firm and your cloud vendor. Although many cloud providers understand the importance of security in the cloud and are compliant with PCI-DSS, FRTB, Consolidated Audit Trail (CAT), MiFID II, and AML, you must ensure during the RFP process that the cloud vendor is in compliance with every rule to which you must adhere.

Financial services firms can further ensure compliance on their own by using a data platform with security built in. That means having a platform that has encryption, access controls, global high availability, and other security features unified across all different kinds of applications and data to help maintain compliance and enhance security.
Big Data: Frequently Asked Questions

What questions should I ask my broker about their use of big data?

These subjects are being promoted more aggressively each day. Here are some questions you can ask that may help to “level set” the jargon and get a clearer picture of how your brokers are using big data-driven technologies:

• What are your machine learning or AI offerings enabling you to do differently?
• How will it affect my experience with your strategies/tools? What in your suite is/ is not touched by this new technology?
• How do these complex analytics work? Can you be transparent about the behavior?
• Where can I expect to see efficiencies?
• Are data quality standards integrated into your day-to-day operations?
• What new techniques are you using to manage risks?
• What about latency? What new processes are in place to enable fast access to all data?
• Describe how you integrate data from multiple sources, from legacy systems to streaming systems, without negatively impacting any existing strategies or platforms.
• How often do you retest your strategies and assumptions?
• Who on your team is managing these big data analytics processes?

Glossary of Terms

Artificial Intelligence (AI)
AI is the attempt to build machines that can perform tasks that are characteristic of human intelligence. This includes understanding language, recognizing objects and sounds, learning, and problem solving.

Byte Scale (Peta, Tera, etc.)
The term refers to information in the terabyte, petabyte, or exabyte range. Organizations working with this range of data or big data often use a distributed file system to rapidly transfer data.

Cloud Computing
Cloud computing refers to the delivery of hosted services over the Internet. Cloud computing enables organizations to leverage computer resources or applications without having to build and maintain computing infrastructure in-house.

Converged Data
Data convergence represents the unification of real-time event streaming, database tasks, and data storage all in one place. A converged data platform is designed to do this, with data governance controls built in.

Cluster
A cluster is a group of servers and other resources that act like a single system and enable parallel processing, high availability, and load balancing.

Data Agility
Data agility refers to how fast an organization can extract value from their mountains of data and how quickly they can translate that information into action.

Data Fabric
The data fabric is a new concept that has been emerging to provide a scalable and flexible solution that converges data management and processing capabilities in real time across data types and across locations to speed up application development.

Data Lake
A data lake is a storage repository that holds a vast amount of raw data in its native format, including structured, semi-structured, and unstructured data. The data structure is defined when the data is needed.

General Data Protection Regulation (GDPR)
GDPR is a regulation in the European Union (EU) law on data protection and privacy for all individuals within the EU, and addresses the export of personal data outside the EU. GDPR took effect on May 25, 2018.

Machine Learning (ML)
Machine learning is the science of teaching computers to learn and adapt using data, without being explicitly programmed.

NoSQL
NoSQL describes high-performance, non-relational databases. NoSQL databases typically do not enforce a schema. They use a variety of data models, including document, graph, key-value, and columnar. NoSQL databases are designed to scale out using distributed clusters of low-cost hardware to increase throughput without increasing latency.

General Data Protection Regulation (GDPR)
GDPR is a regulation in the European Union (EU) law on data protection and privacy for all individuals within the EU, and addresses the export of personal data outside the EU. GDPR took effect on May 25, 2018.

NoSQL
NoSQL describes high-performance, non-relational databases. NoSQL databases typically do not enforce a schema. They use a variety of data models, including document, graph, key-value, and columnar. NoSQL databases are designed to scale out using distributed clusters of low-cost hardware to increase throughput without increasing latency.

Structured Data
Structured data refers to any data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets. Structured data has the advantage of being easily entered, stored, queried, and analyzed.

Unstructured Data
Unstructured data is data that can’t be readily classified and fit into a rigid structure. Unstructured data includes videos, photos, audio, streaming events, web pages, emails, blog entries, word processing documents, etc.

Visualization
Data visualization in big data projects provides a way for organizations to quickly and easily get an overview of their data via dashboard. Most of today’s data visualization tools come with connectors to popular data sources, including common relational databases, Hadoop, and a variety of cloud storage solutions.