BIG DATA: CHALLENGES AND OPPORTUNITIES
Big Data: Challenges and Opportunities

EXECUTIVE SUMMARY

The volumes and complexity of market data required by financial institutions today are immense and growing rapidly. Ongoing market changes are accelerating the growth in demand for data, and forcing financial institutions to address the challenges of what has come to be known as “Big Data”.

This demand is fueled as firms develop and deploy new, more sophisticated cross-asset investment strategies.

At the same time regulatory changes are also forcing firms to source and report increasingly larger amounts of trade data, as well as to adopt higher-quality – and usually data-hungry – risk and pricing models. Investors are making similar demands of their asset managers.

The cost of acquiring and analyzing Big Data has ballooned, with financial institutions globally projected to spend some $25 billion in 2012 on financial data alone.* But the real challenge is deploying the right technology and analytical capability to produce actionable information out of massive data sets, and to do it in a timely manner; repeatedly, during the day. Those that solve this problem most effectively could have a notable competitive advantage over their more slow-footed rivals.

* Burton Taylor Financial Market/Data/Analysis Global Share Segment Sizing 2011
UNWIELDY AND EXPENSIVE, BIG DATA CAN NONETHELESS PROVIDE COMPETITIVE ADVANTAGES TO FIRMS THAT TAME IT

Not a day goes by when “Big Data” fails to appear on financial institution radar screens. The growth in market data volumes and demands for its use and reuse are remarkable, outpacing even the most bullish expectations. In a single generation, the volume of market transactions has increased by several orders of magnitude, and is projected to push global spend on financial data alone to around $25 billion in 2012, according to Burton Taylor.* Capital, staffing, network capacity and technology all struggle to keep up. The problem is accelerating and should demand more attention than it currently receives.

Take the average daily number of pricing ticks processed by Interactive Data: in October 2011, across all traded asset classes in North America, an average of 10.7 billion ticks was processed every trading day. (See Table 1.) That translates into about 19.3 terabytes of data per year. However, the figures in the table are averages; peaks can and do occur. For example, on August 8, 2011, over 26 billion ticks were experienced. From an infrastructure perspective, that suggests the need for capacity to handle at least three times the daily average, which will boost costs measurably. Factor in that many firms purchase data feeds from several providers and one can begin to see the size of the problem.

However, sheer volumes are only part of the issue. The data is becoming referentially more sophisticated too as firms link together a growing array of data sets, build historical databases that include changes a security has undergone, and confirm details of each item on a more granular level. (An example of the increasingly granular nature of the data required for modern investment strategy development, back testing and simulation is shown here.)

### Table: Average Daily Ticks Processed by Interactive Data, October 2011*

<table>
<thead>
<tr>
<th>Daily Average - North American Data Sets</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>US OPTIONS</td>
<td>9,030,826,468</td>
</tr>
<tr>
<td>FUTURES</td>
<td>336,351,522</td>
</tr>
<tr>
<td>STOCKS</td>
<td>304,437,791</td>
</tr>
<tr>
<td>OPTION FUTURES</td>
<td>289,280,789</td>
</tr>
<tr>
<td>SPREADS</td>
<td>1,14,452,521</td>
</tr>
<tr>
<td>FOREIGN EXCHANGE</td>
<td>86,214,834</td>
</tr>
<tr>
<td>BONDS</td>
<td>3,290,921</td>
</tr>
<tr>
<td>INDICES</td>
<td>144,771</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Daily Average - International Data Sets Ticks</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>STOCKS</td>
<td>321,750,544</td>
</tr>
<tr>
<td>OPTIONS</td>
<td>214,810,163</td>
</tr>
<tr>
<td>INDICES</td>
<td>23,197,391</td>
</tr>
</tbody>
</table>

| Grand Total | 10,724,737,715 |

*These figures are based on six-day averages from Interactive Data's global real-time systems.

Source: Interactive Data

* Burton Taylor Financial Market/Data/Analysis Global Share Segment Sizing 2011
This additional complexity may be a boon for those creating multifaceted trading strategies or feeding voracious risk and pricing models, but it greatly amplifies the Big Data problem.

Firms need to normalize and analyze this fire hose of increasingly granular information, separating the useful information from the noise. This is the part that costs the most. A data industry maxim is that firms will spend four to five times as much on processing and integrating data as they spend on the content itself. This cost continues to grow.

While data firms and other service providers offer services for firms that do not want to store or cull the data themselves, this does not entirely solve the problem of escalating costs. ___<<

THE DRIVERS OF BIG DATA

>>>__ The increase in the volume and complexity of market data is primarily driven by three overlapping factors: market evolution, regulatory requirements, and the increasing sophistication of pricing and risk models.

Market Factors
The growth of Big Data points to an important economic trend: access to additional markets at a lower cost to trade. While share volumes remain fairly stagnant, the number of transactions continues to expand rapidly, due in part to the fragmentation and disaggregation of markets.

Another important driver is computer-driven, or algorithmic, trading. The volume of trades initiated by firms using these strategies is vastly higher than those placed by other asset managers pursuing other types of strategies.

Investor demand for greater transparency from hedge funds and asset managers is also boosting Big Data. Pension fund managers and insurers are increasingly using risk attribution and return attribution analyses to inform their allocation decisions and to monitor asset concentrations. All this requires data of a far more complex nature than simply adding more prices.

Regulatory Factors
Under Section 727 of the Dodd-Frank Act OTC derivatives market participants will be required to report swap data including price and volume as soon as technologically practicable after execution of the swap.
Dodd-Frank forces hedge funds and many other investment partnerships to register as investment advisors, putting them under the supervision of the Securities and Exchange Commission. The SEC, and the Commodity Futures Trading Commission, have made plain their desire for these firms to disclose more quantitative information, despite the fact that the government currently lacks the resources to analyze it. That may change, however, as the new Office of Financial Research – the data-gathering arm of the Financial Stability Oversight Council – expands its activities.

Regulators’ increased interest in vetting firms’ risk-based capital calculations, stress test results, Value at Risk (VaR) computations and other metrics will also boost demand for data to feed these exercises, as will regulators’ growing scrutiny of the inputs to firms’ Level 3 asset valuations.

**Model Complexity**

The pricing models for Level 3 assets are not the only ones becoming more data-intensive. In the past decade, risk models have come to require greater quantities of more complex data. For example, stochastic Monte Carlo VaR calculations, which at one time took firms’ supercomputers all night to run, are now being cranked out much more frequently, thanks to increases in off-the-shelf computer power and a steadily growing library of high-performance software tools.

Other data-hungry models have come to the fore due to the ascendance of the credit derivatives market. The so-called Merton or Firm Value models, once the province of Moody’s KMV and a few others, are now widely available. Using them to price credit or determine appropriate hedge ratios often requires enormous amounts of equity data.

---

## BIG DATA CHALLENGES

The common denominator behind these drivers is the need to acquire, manage, analyze and report on large data sets as close to a real-time basis as possible. Big Data, therefore, calls for the ability to adopt a real-time analysis and reporting approach that is both highly empirical and easy to monitor for risk purposes.

Put into economic terms, this trend points to some fairly predictable consequences. Significant investment is required in these competencies to set the baseline capability to compete and thrive in a world of Big Data. The problem is twofold: Yesterday’s technology almost certainly has a limited shelf life, and increased R&D is going to become increasingly important to take advantage of Big Data opportunities.
In the capital markets, this means mastering the ability to manage and analyze content across the security master, corporate actions, real-time and time series pricing and customer data domains, while at the same time complying with new regulatory demands. A key approach to achieving this goal is to deploy the power of statistical and empirical methods across a much broader universe of data sets and business functions.

When looked at as a whole, this positions Big Data at a very interesting intersection between data content, technology, and analytical capability. It involves the pre-processing of data before a firm makes strategic decisions based on it. The pressure to keep pace and thrive in this environment is going to place a premium on having the right software and other infrastructure in place. For data suppliers this means being able to offer a far wider (as well as deeper) universe of content on demand and in a form that is easily consumed. The historical partnership between supply and application of data is going to become even more important and sophisticated as a result.

One impediment to complete optimization of data management is the fact that it is rare to find a financial institution that relies on only one market data provider. Multiple data sources are used to support procedures such as validating vendor data based on a user-defined set of tolerances. While this is viewed as prudent and may be required by internal risk management or compliance departments – for occasions when the data from one provider seems suspect, or the provider has technical difficulties – having two or more providers makes managing the data that much harder.

MAKING VIRTUES OF NECESSITY

So the challenges are pretty clear, but what about the opportunities? Those firms that make the necessary investments will be in a position to begin offering a whole new set of services and investment strategies to their clients. Better yet, those services will be based on state-of-the-art statistical and empirical analyses, providing a far higher degree of risk assessment as an integral part of the investment process.

To be sure, that is going to require participants to embrace their inner statisticians, but the result will be the opportunity to investigate, back test, and simulate new and innovative investment strategies on demand. This new ability to survey the capital markets on a cross-asset, global basis opens up the opportunity to more easily evaluate risks, and balance investment goals against those risks before committing funds. It also allows the firm to run simulations and perform empirical testing, and then to communicate the salient results to customers, without having to send out reams of data “noise”.
To illustrate some Big Data best practices, the following example use case shows how a firm can involve market data in the development, back testing and simulation of a new investment strategy.

BIG DATA AND INVESTMENT STRATEGY DEVELOPMENT

>> A highly visible trend within the industry is the move to empirical and quantitative analysis in the identification of investment opportunities and related execution strategies. The workflow that needs to be supported is:

1. Analysts and traders need to connect to a data source containing a complete universe of fit-for-purpose content required to identify securities of interest based on defined criteria.

2. Users need to query the content of the data source based on logical criteria, commonly referred to as “variables”.

Note that this is a sample list only and is not intended to be definitive. The key point to recognize is that the firm approaches the data source with an undetermined list of securities – they are seeking to resolve for a list of securities of interest (and maintain this list as new securities are added). By definition there is no portfolio identified in advance.

With repeated complex querying, the firm resolves for a list of securities of interest and where they are traded (including the terms under which they trade).

Data Source Criteria

- Industry classification associated with the issuer
- Related firms in the corporate hierarchy
- Asset classification
- Currency
- Payment rates
- Cash generation (e.g., dividend payment rates)
- Trading locations
- Features such as convertibility, callability etc.
- Associated derivatives
The firm joins this list of securities of interest with available pricing data in order to analyze pricing patterns over time. Key signals that tend to be looked at include:

- Volatility
- Cash rate of return over time
- Total return over time (including associated cash flows)
- Arbitrage possibilities
- Performance versus index (or indices)
- Performance versus sector

This is an indicative list only – the analyses to which these scenarios are subjected vary considerably between firms based on investment policies and other business criteria (including acceptable levels of risk). These pre-execution strategy analyses act to increase the firm’s data appetite and put additional strain on infrastructure. However, this analytical diversity is viewed as a competitive differentiator between firms.

The firm then selects various portfolios of instruments resulting from the querying activities and assessments and collates corresponding deep historical pricing sets on either an end-of-day or (preferably) intraday basis or both.

The portfolio is subjected to historical validation based on the pricing history over selected time periods. This is an arbitrary choice that varies between firms; however, the general rule of thumb required by statistical analysis is that the more data available, the better. This process is referred to as “back testing”.

Based on the results of the back testing (typically measured by net return over time versus a selected index or other group of securities deemed to be risk free), the analyst executes a specific buy/sell strategy. It should be noted that this strategy is dynamic. Since the selection of the portfolio is based on database criteria, the list of securities that can be part of the analysis can (and does) change from day to day. This introduces additional complexity for the investment firm, but makes it more likely that they execute on a unique strategy, as it would be almost impossible for a competitor firm to produce precisely the same list of securities and execution rules.

Where a portfolio and related execution strategy is deemed to meet the firm’s risk profile, it is moved into final simulation for an ad hoc period of time. In this environment the strategy is connected to real-time pricing streams and runs in parallel to actual trading to verify the investment hypothesis.

With successful simulation, the strategy is then placed into full production. Signals generated from the investment model are routed to either traders or automated execution systems. A cycle of innovation is thus formed.
The whole process sounds relatively simple, but in practice it requires a significant expertise to operate at large scale. It involves whole teams of IT professionals, data analysts, investment researchers and portfolio managers.

For this process to work, this is what's needed from the data perspective:

**Reference data content** – In order to scan the capital markets to identify suitable instruments (and their associated derivatives) based on logical criteria, a comprehensive, cross-asset collection of data content is required. Ongoing investigation and research has identified a universe of approximately 300 fields of interest.

**Pricing content** – Financial statistical analysis requires the use of extensive pricing time series and associated statistical coefficients. These time series must be able to be linked to the universe of reference data in order to perform a join between the instruments of interest resolved from the reference data universe with measures of their economic performance in the capital markets. Tick-based records are particularly sought after as they provide precision on an intraday basis. Alternatively, end-of-day pricing values are acceptable for many forms of analysis.

**Interrogation** – Functionality that allows searching and screening of the data universe is a fundamental requirement. The entry criteria to performing empirical analysis require that the universe of content is scanned to identify those instruments that meet the selection and logical criteria of the hypothesized investment strategy. The industry predominantly uses Structured Query Language (and Transact-SQL in particular) to perform this task in a high performance manner.

**Statistical applications and libraries** – With the selection of instruments of interest (and their derivatives), the associated pricing histories are submitted to rigorous statistical analysis. This requires that the content can be passed through a statistical application in order to generate coefficients as required by the underlying model (these will vary by firm and the logic applied in developing the investment model). Two approaches are typically followed:

- Data are extracted and passed as a batch to an external analysis package – known to be a slow and cumbersome process.
- Application libraries are called as part of the request for historical pricing data and return the resulting coefficients making up the analysis. This is standard Transact-SQL functionality and is commonly used across the industry.
Data Quality – Successful quantitative analysis demands consistent data content. Typically referred to as “data quality” two fundamental requirements are:

- Support for adjustment in response to corporate actions (e.g., dividend policy changes, M&A, recapitalizations and so on).
- Support for management of symbol changes over time.

There are additional requirements, such as accuracy, but these two elements are fundamental to analysts in order to base their assessments on consistent pricing vectors.

A HOPEFUL VIEW

>> Big Data is the natural evolution of an increasingly globalized business. As such, it unleashes a deep reservoir of creativity in order to keep pace with and take advantage of opportunities. That makes for exciting times across the capital markets as new functionality, information services, and applications will make their way into investors’ hands.
### About Interactive Data

Interactive Data Corporation is a trusted leader in financial information. Thousands of financial institutions and active traders, as well as hundreds of software and service providers, subscribe to our fixed income evaluations, reference data, real-time market data, trading infrastructure services, fixed income analytics, desktop solutions and web-based solutions. Interactive Data’s offerings support clients around the world with mission-critical functions, including portfolio valuation, regulatory compliance, risk management, electronic trading and wealth management. Interactive Data is headquartered in Bedford, Massachusetts and has over 2,500 employees in offices worldwide.

For more information, please visit www.interactivedata.com.

### North America

**Interactive Data**  
100 William Street  
New York, NY 10038  
USA

Tel: 212-771-6908  
Fax: 212-771-6987  
email: info@interactivedata.com

### Europe

**Interactive Data**  
Fitzroy House  
13-17 Epworth Street  
London EC2A 4DL  
UK

Tel: +44 (0)20 7825 7700  
Fax: +44 (0)20 7825 7701  
email: eu-info@interactivedata.com

### Asia Pacific

**Interactive Data (Australia) Pty Ltd**  
1155 Malvern Road  
Malvern  
Melbourne  
Victoria 3144  
Australia

Tel: +61 3 9249 2000  
Fax: +61 3 9249 2001  
email: info@interactivedata.com

### Limitations

This document is provided for informational purposes only. The information contained in this document is subject to change without notice and does not constitute any form of warranty, representation, or undertaking. Nothing herein should in any way be deemed to alter the legal rights and obligations contained in agreements between Interactive Data Real-Time Services and its affiliates (collectively, “Interactive Data”) and their clients relating to any of the products or services described herein.

Interactive Data makes no warranties whatsoever, either express or implied, as to merchantability, fitness for a particular purpose, or any other matter. Without limiting the foregoing, Interactive Data makes no representation or warranty that any data or information supplied to or by it are complete or free from errors, omissions, or defects.

Interactive Data™ and the Interactive Data logo are service marks of Interactive Data Corporation. Other products, services, or company names mentioned herein are the property of, and may be the service mark or trademark of, their respective owners.

©2012 Interactive Data Corporation, Inc.