Big Data in Capital Markets

Michael Rauchman: michael.rauchman@gmail.com
Alex Nazaruk: anazaruk@comcast.net

We are grateful to the folks at Tradeworx for their help and for letting us use some of their unpublished materials.
Outline

- Introduction to modern financial markets
  - Basic concept and terminology
- Data collection and analysis challenges for the regulators
  - MIDAS
  - CAT
- Electronic trading and objectives of market participants
- Typical automated trading system
  - Trading model development process
- Implications for DBMS

©Middle Lake Partners LLC, Big Data Advisory and Investment
US Stock Market Structure Today

- Orders submitted via 2000+ Broker Dealers
- Regulated by SEC, SROs (FINRA and exchanges)
- Daily Volumes
  - $50-100 billion notional value traded
  - 5-10 billion shares traded
  - 2-6 billion orders submitted
  - Only ~1% of orders get executed

<table>
<thead>
<tr>
<th>Execution Venues</th>
<th>Number of Venues</th>
<th>Volume %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exchanges</td>
<td>14</td>
<td>66%</td>
</tr>
<tr>
<td>Dark Pools</td>
<td>Around 50</td>
<td>13%</td>
</tr>
<tr>
<td>Internalization</td>
<td>200+</td>
<td>21%</td>
</tr>
</tbody>
</table>
Inside an Exchange – Order Events

- **Incoming**
  - New Order (symbol, side, price, size, order type, etc.)
  - Order Cancel
  - Order CancelReplace / Modify

- **Outgoing**
  - Acknowledgements (New, Canceled, Modified)
  - Rejects
  - Executions
Market Data

- Rich data - anonymized stream of order events
  - Shows new orders, order modifications, cancellations and executions
  - Not all orders are there (marketable, hidden orders, etc.)
  - Most transparent paradigm
  - Increment based data – need full replay to reconstruct state of the market
- Order book snapshots
  - Aggregated by price level
  - Usually truncated to few best price levels
Race to the Bottom
High Frequency Trading (HFT)

- Fast systems
  - Exchange response time <100 µsec
  - Fastest trading systems ~15 µsec
  - CPU cache usage optimization
  - Exchanges can produce millions market data msg/sec
  - Dealing with data bursts - FPGAs, GPUs, etc.

- Dealing with geographic distribution
  - Multiple data centers and co-locations
  - Microwave NY to CHI and in NY metro
  - New transatlantic cable
Challenges for Regulators

- Health of the market structure
  - How HFT affects the markets?
  - What are the effects of dark pools on the markets?
  - What exactly is happening to a retail order?
- Effects of recent and potential rule changes
  - How would minimum time-in-force for orders affect liquidity?
  - How tick sizes affects visible/dark liquidity?
- Understanding unusual market events
  - Flash crash (5/6/2010)
  - Knight Capital Group breakdown (8/1/2012)
MIDAS - Market Information Data Analytics System

- Developed by Tradeworx – a 14 year old trading firm and financial technology provider
- System originally developed for use by trading firms (in-house and as a service)
- MIDAS went on-line at SEC in January 2013
- Currently 100+ daily users
- Examples of usage include:
  - Mini flash-crash analysis
  - Rule change impact assessment
  - Detection of abnormal patterns in message traffic
MIDAS - Data Capture

“Big Data”
- Stocks, options and futures
- 6 co-located datacenters
- 15 different protocols
- 1 terabyte data / day
- Millions of messages / second
- 1 µsec raw to normalized

Network
- Min 10 Gbps fiber network
- Microwave link between metro NY and CHI

©Middle Lake Partners LLC, Big Data Advisory and Investment
Data Capture Challenge – Accurate Timing

- Why relative timing is important?
  - Price protection rule
  - Need to see complete picture from any venue

- Why is it difficult?
  - Relative clock drift
  - Distance between data centers

- Solution
  - GPS time synch
MIDAS – Storage

- AWS VPC
- Forced Amazon into FedRAMP certification
- Normalized compressed text
  - Optimized for consecutive replay
  - Redundant storage with two file organization strategies:
    - Segregated by symbol, ordered by time
    - Whole universe segregated by time slice, ordered by time
MIDAS – Analysis Level 1

- **Graphical order book viewer**
  - Graphical representation of the book
  - TIVO functionality with microsecond resolution
  - combines books from any number of sources
  - allows “point-of-view” aggregation
MIDAS – Analysis Level 2

- **Interval-based data research platform**
  - Proprietary language for clustering and partitioning of data into fixed length bars
  - Standard statistical tool set optimized for time series analysis
  - Interfaces for Python, R, SAS, etc.
MIDAS – Analysis Level 3

- Tick data research platform
  - Replay framework for continuous tick-level data
  - C++ call-back based API
  - State-of-the-art multi-exchange simulator

“for the growing team of quant types now employed at the SEC, MIDAS is becoming the world’s greatest data sandbox.” – SEC Chairman Elisse Walter
Challenges for Regulators

- Health of the market structure
  - How HFT affects the markets?
  - What are the effects of dark pools on the markets?
  - What exactly is happening to a retail order?
- Effects of recent and potential rule changes
  - How would minimum time-in-force for orders affect liquidity?
  - How tick sizes affects visible/dark liquidity?
- Understanding unusual market events
  - Flash crash (5/6/2010)
  - Knight Capital Group breakdown (8/1/2012)
CAT – Consolidated Audit Trail System

- Mandated by Securities Exchange Act Rule 613, which was adopted by the SEC in July 2012.

- Bidding process for CAT is currently under way:
  - Process is run by SROs (national securities exchanges and national securities associations) - see [http://catnmsplan.com/](http://catnmsplan.com/)
  - RFP Responses due 6/30/2013
CAT – Scope of Data

CAT will consolidate comprehensive public and non-public data about the market into a single system.

Non-public information to be collected by CAT:

- Customer and account information
- Identities of the parties for every order and trade
- Activity on public markets not published in market data (marketable, hidden orders, etc.)
- Order events for customer-BD interaction
- Ability to trace the lifecycle of the order
Order Routing Scenario

Customer Sends Order to Firm A

Firm A creates #1234, Routes #4567 to Firm B

Order #4567

Firm B receives #4567 and creates #3456, Routes #8765 to Exch1

Order #8765

Exch1 Receives #8765, creates #7896

CAT

©Middle Lake Partners LLC, Big Data Advisory and Investment
CAT-Order-ID problem

- A series of unique order identifiers assigned by CAT Reporters are linked together by the CAT to create the lifecycle of an order and assigned a single CAT-Order-ID for the lifecycle.

- Simple problem, but very search intensive with severe time constraint
CAT – Data Collection and Validation

- Protocols will be specified by CAT implementer
- Will be collecting data from 2000+ participants
- The initial data checks required by the CAT include, but are not limited to:
  - Data format validation and syntax check
  - Identification of unregistered accounts
  - Identification of unregistered market participant identifiers
  - Identification of unlinked lifecycle events
- 4 hour validation window – needs to process 5mln records/sec!
CAT – Data Repository

- Rolling 5 year period on-line + 2 years of archived data

<table>
<thead>
<tr>
<th></th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated daily data size</td>
<td>5 TB</td>
<td>13 TB</td>
<td>16 TB</td>
<td>20 TB</td>
<td>24 TB</td>
</tr>
<tr>
<td>Estimated daily records</td>
<td>22bn</td>
<td>58bn</td>
<td>71bn</td>
<td>89bn</td>
<td>107bn</td>
</tr>
<tr>
<td>Accumulated total size of central repository</td>
<td>2 PB</td>
<td>6 PB</td>
<td>10 PB</td>
<td>15 PB</td>
<td>21 PB</td>
</tr>
</tbody>
</table>
CAT - Query and Extraction

• Online Query Tool
  • require a minimum set of criteria, including date/time range, symbol, Customer ID(s), CAT-Order-ID(s), etc.
  • must support approximately 3,000 registered users

• Bulk Data Extraction
  • bulk extraction and download of data, based on a specified date/time range, market, security, Customer ID and the size of the resulting data set

• Different levels of access depending on user role and function
# MIDAS vs CAT

<table>
<thead>
<tr>
<th></th>
<th>MIDAS</th>
<th>CAT</th>
</tr>
</thead>
</table>
| **Scope of data** | • US stocks, futures and options  
                    • Public market data only | • US stocks and options  
                    • Market data (public and non-public)  
                    • Customer data |
| **Users**         | • Trading firms  
                    • SEC                                      | • SEC  
                    • SROs |
| **Data Inputs**   | • Real-time capture (6 data centers)  
                    • Focus on accurate timings               | • Mostly batch submissions (2000+ submitters)  
                    • Focus on validation                     |
| **Repository**    | • Compressed text on a public cloud                                  | • TBD, likely an analytical database                             |
| **Usage paradigm**| • Analytical toolset                                                 | • Online query tools  
                    • Bulk data extraction                       |
Beyond CAT

“My long-term vision is a consolidated audit trail that spans products, markets and the globe” – SEC Chairman Elisse Walter
Further Topics

- Market participants and their trading objectives
- Types of financial data and its usage. How big is Big?
- How does typical automated trading system work?
- What is trading model?
- Trading model development dataflow
- Implications of Big Data In Electronic Trading for DBMS
# Market Participants

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>Orders/day</th>
<th>Hold Time</th>
<th>Price Sensitivity</th>
<th>Returns/transact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investors</td>
<td>Pension Funds, Retail Investors</td>
<td>1s - 100s</td>
<td>Months - Years</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Speculators</td>
<td>Hedge Funds, Day Traders</td>
<td>100s - 1000s</td>
<td>Seconds - Months</td>
<td>Low-High</td>
<td>Low-High</td>
</tr>
<tr>
<td>Market Makers</td>
<td>Prop Shops, NYSE DMM</td>
<td>1000s - MMs</td>
<td>Subseconds - Days</td>
<td>High</td>
<td>Very Low</td>
</tr>
</tbody>
</table>
Types Of Data

• Reference Data, a.k.a. Securities Master (ticker symbol, exchange, security description, corporate actions, etc.)

• Fundamental Data (corporate financials, analyst reports, filings, etc.)

• Market Data (orders, trades)

• News (earnings reports, economic news, etc.)

• Social Media (market sentiment, twits, Robinhood, etc.)
Who is interested in what?

Value of data for various types of market participants

- Investors
- Speculators
- Market Makers

Market Data Resolution:
- Fundamental
- Monthly
- Weekly
- Daily
- Hourly
- Minutes
- Seconds
- Tick-by-tick

©Middle Lake Partners LLC, Big Data Advisory and Investment
How Big is Market Data?

OPRA projections for 2014:
- 14.7 million MPS
- 24.6 billion messages/day
- Output rate for a single line 1 million MPS

ARCA data feed in Feb 2013 observed:
- 1 sec peak: 98.18 Mbps
- 10ms peak: 945.42 Mbps (a lot of microbursts)
Typical Automated Electronic Trading System

Market

Orders
Modifies
 Cancels
 Executions

Trading System

Securities Master

Trading Models
(pre-programmed trading instructions)

Telemetry Data

Price Ticks
What Is Trading Model?

- Trading model is a set of pre-programmed trading instructions that implement a particular trading strategy, e.g.:
  - Statistical arbitrage
  - Trend following
  - Mean reversion
  - Scalping
  - Market-making
  - etc.
Statistical Arbitrage as an example of trading strategy

- There are inherent relationships between prices of some securities (e.g., SP500 ETF vs. its component stocks; equity option and its underlying stock)
- There are statistical relationships between stocks in the same sector (e.g., PEP vs. KO, airline carriers)
- Securities prices/returns are often highly correlated
- …but markets are not perfectly efficient: there are temporary anomalies in relative pricing of correlated securities
Statistical Arbitrage as an example of trading strategy (continued)

- Typical Stat Arb strategy considers a basket(s) of hundreds (sometimes thousands) of correlated securities and is seeking to capitalize on short-term temporary relative-pricing anomalies.
  - employs complex statistical methods;
  - involves a lot of high-resolution market data;
  - is computationally extensive;
So, how do they develop trading models?
Trading Model Development Dataflow

- Exchange
  - Raw Format
  - Normalized Format
  - Quotes Data

- Storage (files in DFS and MPP DBs)
  - Securities Master
  - APIs, DSLs
  - Automated model backtesting and optimization

- Interactive: R, Python etc
  - Data Vendor

- Model development: human-driven interactive ad-hoc exploratory analysis and pattern discovery

©Middle Lake Partners LLC, Big Data Advisory and Investment
Trading Model Development Dataflow

©Middle Lake Partners LLC, Big Data Advisory and Investment
Model Development Process Design Requirements Summary

Enabling **humans** to generate new ideas (exploratory analysis)
- Human-centric
- Ad-hoc
- Interactive

Enabling **machines** to validate and optimize ideas
- Machine-centric
- Not interactive
- Highly automated tick data “replay”
- Many iterations with various combinations of model parameters
Idea development process must be efficient!

- Minimizing time-to-value requires:
  - variety and volume of available data
  - efficient queries (easy to write, quick response)
  - interactive ad-hoc analysis (R, SciPy, NumPy, visualization)
  - idea has to be easily expressible and sharable with other people; machines need to understand it too (DSL)
  - large chunk of quants’ time spent in data preparation, not in analysis; decouple data processing and analytics
Trading Model Development Dataflow

Model development: human-driven interactive ad-hoc exploratory analysis and pattern discovery

Interactive: R, Python etc

Storage (files in DFS and MPP DBs)

Securities Master

Data Vendor

APIs, DSLs

Automated model backtesting and optimization

Quotes Data

©Middle Lake Partners LLC, Big Data Advisory and Investment
Core Data System Zoomed-In:
It’s a very complex (and costly!) operational environment
So, what does it all mean for DBMS?
Financial Data Characteristics

- **Tick Data**
  - well-structured (when transformed into normalized format)
  - good fit for relational DB (for ad-hoc interactive analysis)
  - good fit for scientific formats, like HDF5 files (for automated back-testing)

- **Securities Master**
  - multi-structured
  - good fit for document-oriented DB

- **Fundamental Data**
  - multi- or loosely-structured
  - may benefit from a semantic data model/catalog
  - semantic DB
More Tick Data Characteristics

- normalized format -> well-structured
- time-series -> deterministic records order -> seq disk I/O (mostly) with efficient read-ahead
- can be easily segmented by date/exchange/ticker
- WHERE predicates are mostly predictable (date, exchange, ticker)
- key record attributes (date, exchange, ticker) are low cardinality -> good RLE compression within a column
- stock price changes in increments -> good DELTA-based compression within a column
Challenges

- Tick Data seems a good fit for column-store or array DB, but...
  - DBs are slow for retrieving large subsets
  - Model development requires additional context -> a common language to access and blend data from relational, document-oriented or semantic DBs
  - DBs need better build-in time-series analytics -> specialized SQL language extensions
  - DBs need to integrate statistical tools supporting parallelism in complex (computationally extensive) analysis
Core Data System Zoomed-In: It’s a very complex (and costly!) operational environment
Challenges: Market Data Platform

- Common Data API
- Data: Relational DB, Document-Oriented DB, Semantics DB, ArrayDB + Stats Tools
- ETL raw files into specialized DBs
- In-storage Parallel Processing Framework
- General Purpose DFS: Raw Files
- Hardware Cluster

- Provides access to raw files and heterogeneous DBs
- MPP DBs serve as specialized "materialized views" of raw file data
- Time-series friendly (TS file as a minimum processing unit?)
- Provide file-level metadata and semantics for data navigation and quality control
- A shared cluster hosting both raw files and DBs

©Middle Lake Partners LLC, Big Data Advisory and Investment
Challenges: Bringing Data Together

The hard way:

```sql
SELECT Date, Exchange, Ticker, Price, Dividend
FROM tbl_ticks AS Ticks
JOIN

  QUERY JSON:

    SELECT db.dividends.find(
      { ticker: "MSFT" },
      {date:1, ticker:1, dividend:1}
    ) AS Date, Ticker, Dividend

  ) AS Dividends ON Ticks.Date = Dividends.Date AND Ticks.Ticker = Dividend.Ticker
WHERE Ticks.Date = '6/25/13' AND Ticks.Exchange = 'NASDAQ' AND
Ticks.Ticker = 'MSFT'

```

The easy way:

```javascript
analytics.correlate("SP500", '6/25/13', "1 second").top(10)
```
Questions?

- Presented by Middle Lake Partners, LLC
- Boutique Big Data Technologies Advisory and Investment Firm
  - Enterprise Big Data Advisory Services:
    - Business Case Analysis
    - Strategy and roadmap
    - Architecture and design
    - Technology due diligence and selection
    - Implementation
  - Big Data Technology Investments
    - Alex Nazaruk  anazaruk@comcast.net
    - Michael Rauchman  michael.rauchman@gmail.com

©Middle Lake Partners LLC, Big Data Advisory and Investment