



Historical Data Solutions: The Devil is in the Details

*Best Practices & Pitfalls
Whether You Build or Buy*

Presented by

Tom Myers

Vice President of Business Development

December 2, 2008

Sun London Customer Briefing Centre – London

What We'll Discuss Tonight

- Overview of sound historical data solution
- Sourcing your research data
- Corporate actions & ticker mapping
- Filtering data for bad ticks
- Survivorship bias
- The Build vs. Buy decision
- How to get more information

Sound Historical Data Solutions

- Reliable source of complete & accurate:
 - Price Data (quotes & trades)
 - Corporate Actions (e.g. dividends, splits, mergers, etc.)
 - Ticker/Identifier Mapping (e.g. symbol changes, exchange hops, etc.)
- Ability to identify (and correct) bad ticks
- Process for keeping data sets current

Sourcing Research Data

- Research data collected in real-time is fraught with pitfalls:
 - Dropped packets of data
 - Inability to go back to fill in gaps in data
 - Incomplete corporate actions and ticker mapping
 - Must collect enough data to begin research
 - Cost vs. benefit for new markets

Sourcing Research Data

- Obtain data from multiple sources & via multiple connections
- Obtain data after the close (not real-time)
- Sourcing research data directly from the exchange vs. from an aggregator – which is “best” depends on the market
- Even exchange data files contain omissions and errors

Sourcing Research Data

- Source for corporate actions and ticker mapping is **CRITICAL**
- Integrating price data, corporate actions and ticker mapping files is tedious
- Need to streamline and automate process for it to be scalable
- Make process flexible

Corporate Actions & Ticker Maps

- You need to know the:
 - Symbol
 - Company name
 - Exchange
 - Identifier (CUSIP, ISIN, SEDOL)
 - First & last date of data
 - Unique ID for the security
- Map symbols to Unique ID's

Corporate Actions & Ticker Maps

- ID# 8347 (Exxon Mobil) consists of symbol 'XON' from 1/4/1993 to 11/30/1999 and 'XOM' from 12/1/1999 to present.
- The ticker 'C' has been Chrysler and Citigroup – you must be able to differentiate between the two historically

Corporate Actions & Ticker Maps

- You need to track all stock splits, stock dividends, and cash dividends
- You need to track all exchange hops. For example...
 - JOYG moved from the NYSE to OTC_BB on 12/9/1999, and returned to NASDAQ on 7/31/2001

Filtering Data for Bad Ticks

- What exactly is a “bad tick”?
- How often do bad ticks occur?
- Why do bad ticks occur?
- *“So what if a bad tick gets into my system”,* said the young Padawan
- Unfiltered trade data is VERY suspect

Filtering Data for Bad Ticks

- Filtering data involves a tradeoff:
 - Filter too loosely and still have unusable data
 - Filter too much and remove the “reality” from the data
- Models developed with over-filtered data are doomed in real-time trading
- Treat historical vs. real-time the same

Filtering Data for Bad Ticks

- Data must be filtered at its finest granularity to ensure a usable time series
 - Daily bars may appear fine, but actually contain 30-minute bars with obvious outliers
 - Those filtered 30-minute bars may also like, fine but contain 10-minute bars with obvious outliers
 - So filter tick data to create the most usable time series

Filtering Data for Bad Ticks

- Goal of any data filter should be to produce a time series that:
 - Removes false outliers in your base unit of analysis (e.g. tick, 1-minute, 15-minute, etc.)
 - Supports back-testing without removing the statistical properties (“reality”) of the data
 - Does not introduce excessive latency
 - Adapts to different securities, tick frequency rates and price levels

Survivorship Bias

- What is Survivorship Bias?
 - Symbol universes are often based on broad indices where capitalization determines membership
 - Such indices reward (include) outperforming stocks, and punish (remove) underperformers
 - Many traders include the outperforming stocks and never see the impact of the underperformers

Survivorship Bias

- How Survivorship Bias gets worse...
 - Start trading with a “static” symbol universe
 - You omit the next stock added to the index
 - You include the underperforming company that was just crowded out of the index
 - Now you are NOT trading the outperforming company
 - You ARE trading the underperforming company
 - Both are in direct opposition to what you did in your research

Survivorship Bias

- What can you do about Survivorship Bias?
 - Maintain common practices in real-time and in simulation
 - Do not use a “static” symbol universe
 - Determine the criteria for inclusion in your research universe
 - Identify how often universe membership can change (i.e. be rebalanced) using your criteria

Survivorship Bias

- What can you do about Survivorship Bias?
 - Recreate “as of” universe membership lists on each date membership changes
 - Build a database that identifies the date range for which each security met membership criteria and is eligible to be traded in simulation
 - Obtain pricing and related data on all companies in the Comprehensive Research Universe
 - Dead companies are challenging

The Build vs. Buy Decision

- Depends on your:
 - Budget
 - Expertise with price data, corporate actions, ticker mapping & data filtering
 - Human resources to throw at the project
 - Time to market deadlines

How to Get More Information

*White Papers available at
www.TickData.com*

Tom Myers

+1.972.747.8815

tmyers@tickdata.com