Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Tick Data Quality Control: Detecting and Correcting Inconsistencies in High-Frequency Trading

Gurunath Dasari
UCLA Anderson School of Management, USA
Corresponding Author: Gurunath Dasari, **E-mail**: reach.gurunathdasari@gmail.com

| ABSTRACT

The research undertakes a study of tick data quality control mechanisms that operate in high-frequency trading frameworks to perform crucial inconsistency detection and correction tasks for algorithmic strategy success. A theoretical foundation of quality assessment accompanies the presentation of statistical and machine learning detection methods and it provides correction strategies through filtering and interpolation techniques, and discusses optimized implementations for handling massive data streams. Overall, the paper discusses pipeline design alongside parallel processing optimization and database enhancements and monitoring systems needed to maintain data consistency across distributed market systems. The article establishes a procedure to maintain reliable tick data through hardware acceleration studies and adaptive thresholds and complete audit protocols for trading operations and risk management and regulatory compliance purposes.

| KEYWORDS

Tick Data Quality, High-frequency Trading, Anomaly Detection, Data Reconstruction, Market Microstructure

| ARTICLE INFORMATION

ACCEPTED: 12 April 2025 **PUBLISHED:** 21 May 2025 **DOI:** 10.32996/jcsts.2025.7.4.94

Introduction

Recent statistics indicate that HFT and algorithmic trading methods account for 73% of total U.S. equity trading activities [1]. HFT and algorithmic trading operations leverage millisecond or microsecond trade-by-trade information that originates from tick data to function. The accuracy standards of this information system determine how effectively trading systems manage financial risks as well as regulatory frameworks in global institutions.

The quality of data enables trading companies to generate superior returns per trade that perform 5.4 basis points better than their market counterparts based on exchange data analysis [1]. The performance difference outlined in the study creates substantial economic value which supports operational growth on millions of daily deals. Wider execution efficiency declines as automated systems trigger incorrect signals due to problems found in tick data which include price spikes and missing observations together with timestamp inconsistencies. Frequent issues in market data quality result in a 12-15% increase of trading costs as research on execution quality metrics suggests [2].

The diverse range of technical difficulties creates a complicated environment for tick data quality. Exchange feed disruptions, together with data distribution network connectivity problems, cause missing ticks to become the most frequent quality issue. Data gaps detected by empirical studies of major exchange data occur about four to six times daily on average with durations spanning from 200 milliseconds to several seconds [2]. Risk management requires precise data the most during market volatility which happens to be when data interruptions occur. Price spikes create a considerable issue because quotes that diverge from fundamental values rise ten to twenty percent beyond original reference points prior to their correction or cancellation. The union of market data from multiple trading venues becomes harder to process because timestamp discrepancies grow worse as different venues synchronize their clocks with 2-3 milliseconds of variation during volatile price movements in liquid instruments [1].

Public regulatory authorities have improved data quality standards by adapting to marketplace developments. Participating exchanges together with market participants must deploy automated surveillance systems that locate and handle inconsistent data using current regulatory requirements. Presently regulatory examinations check the fullness and precision of marketplace data archives while establishing a lower than 0.1% error threshold for all reporting obligations [2]. Systematic data quality failures in market operations will lead to substantial financial penalties and operational restrictions that damage a firm's market position.

Expansion of alternative trading platforms worsens data quality issues because it distributes liquidity throughout various exchanges as well as dark pools and electronic communication networks. Study results show significant variations in reported price and volume data between different venues where the divergence rates vary between 2% to 8% depending on market stability [1]. The divergent data points between trading venues enables cross-venue arbitrage yet complicates market monitoring by firms who need to reconstruct an integrated market view. Harms of advanced market participants build hierarchical verification methods to check information across multiple sources before using it as trading inputs [2].

Theoretical Framework for Tick Data Quality Assessment

A broad framework to evaluate tick data quality requires multiple aspects that capture high-frequency financial data traits. A quality assessment framework for tick data should measure at least five essential dimensions which are completeness and accuracy as well as consistency and granularity and timeliness [3]. These dimensions undergo systematic alteration in quality assessment measures through the analysis of trading platforms under varying market conditions and instrument specifications. Major currency pairs demonstrate complete performance above 99% in normal trading patterns but reduce to 94-97% range during times of market volatility stress or liquidity events.

Multiple statistical characteristics between clean tick data and contaminated data present distinct patterns throughout various analytical measurements. When tick data undergoes analysis at different sampling rates, it displays distinctive volatility characteristics of high-quality data. Realized volatility measures in equity index futures show a systematic reduction when sampling frequency declines according to noise decay patterns which have been validated in quantitative studies [3]. Such patterns in contaminated datasets appear either with sudden irregular breaks or discontinuities. The distribution of returns at ultra-high frequencies shows specific patterns in clean data through heavy-tailed distributions which exhibit identified ranges of excess kurtosis and match across varying periods. Distributional properties of contaminated data do not match these scaling laws because they show inconsistent patterns across different time intervals [4].

Tick data quality assessment faces an important obstacle from microstructure noise since this kind of market behavior may confuse data quality issues with legitimate financial market effects. Research studying microstructure noise in worldwide equity markets demonstrates that the noise variance to efficient price variance ratio displays clear patterns according to market capitalization levels and trading volumes as well as bid-ask spread features [3]. The noise factor constitutes a major contributor to high-frequency return variance which quality assessment systems need to account for to distinguish legitimate market actions from data anomalies. Systematic microstructure noise patterns serve as critical standards to help quality control systems detect observations whose noise patterns differ abnormally from typical instrument performance under comparable market conditions. The implementation of sophisticated quality assessment frameworks utilizes adaptive threshold mechanisms that modify themselves according to the liquidity factors observed in real-time trading sessions due to changing acceptable noise profile criteria during trading hours [4].

The quality assessment framework needs to identify and handle different market regime periods that emerge throughout trading halts, along with circuit breakers and auction sessions in tick data. The behavior of traders during circuit breaker implementation triggers substantial statistical pattern changes when compared to regular market times [4]. Shifts in market regimes also happen at exchange start and end times since auctions function differently from continuous trading. The empirical research utilizing tick data across these market transition times identifies substantial modifications in the frequency of trade orders and the patterns of price formation together with order cancellation patterns. Quality assessment methods need specific contextual adjustment tools that help alter detection systems and validation procedures to deal with normal data changes across various market conditions. Market event calendars and real-time trading state notification channels need to be incorporated into quality control systems according to [3].

Tick data shows temporal correlation structures which serve as an additional method to evaluate quality assessment. Research studies of autocorrelation patterns in high-frequency returns from different markets demonstrate that these patterns display specific decay patterns that function as identification markers for data quality testing [4]. The magnitude of first-order negative autocorrelation which bid-ask bounce produces in transaction price returns depends on both trade size and spread width conditions. Quality assessment frameworks use known expected correlation patterns to identify suspicious patterns that suggest either data corruption problems or feed-related issues. Machine learning techniques in advanced implementations create dynamic baseline models for correlation structures which enable more precise quality issue detection throughout different market segments at varying periods of the day and volatility levels [3].

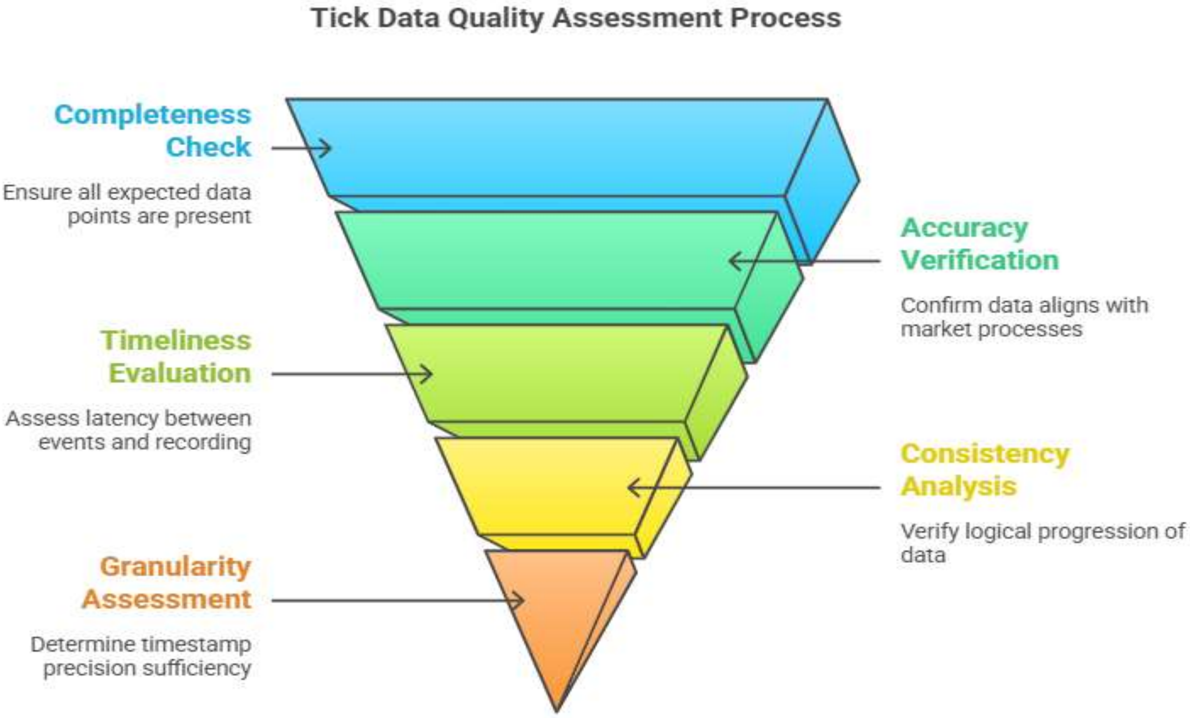
Tick Data Quality Assessment Process
Completeness Check
Ensure all expected data points are present
Accuracy Verification
Confirm data aligns with market processes
Timeliness Evaluation
Assess latency between events and recording
Consistency Analysis
Verify logical progression of data
Granularity Assessment
Determine timestamp precision sufficiency

Fig 1: Tick Data Quality Assessment Process [3, 4]

Detection Methodologies for Tick Data Anomalies

The identification of price and volume outliers in tick data through statistical analysis requires alignment of detection effectiveness against incorrect alarm generation while handling the non-Gaussian distribution patterns governing high-frequency financial data. The use of robust statistical methods supported by rolling windows proves successful for real-time operational deployment within various asset categories. Publications analyzing electronic futures market high-frequency data have proven that anomaly-detection methods which use interquartile range-based adaptable thresholds deliver superior performance than static thresholding methods when used to identify manual anomalies [5]. The detection algorithms modify their threshold variables according to current market volatility patterns through mechanisms that extend boundaries in volatile periods and reduce them in quieter times. Testing on various trading platforms shows that price-based irregularities produce brief periods of large-price moves compared to native price variability and volume abnormalities consist of prolonged alterations to trading activity patterns. The detection accuracy of dual price and size multi-factor methods surpasses the accuracy of single-dimensional methods according to empirical market evaluations across equity and fixed income markets [6]. The statistical decision systems function best when properly calibrated for instrument characteristics and their liquidity specifications serve directly as tuning parameters for sensitivity levels.

Time-series analysis techniques extend detection capabilities beyond point-based outlier identification to capture structural breaks and regime shifts that may signal data quality issues. Sequential analysis methods examining the evolution of statistical properties over rolling windows effectively identify discontinuities in data patterns that might indicate feed problems or exchange system issues [5]. Time-series analysis methods enable detection of data quality problems by identifying both singular data outliers and structural shifts and regime transformations in the data stream. Sequential analysis methods which track the statistical pattern development through moving time windows enable detection of feed problems or exchange system issues by monitoring data pattern discontinuities [5]. Auto-correlation monitoring at multiple lags enhances foreign exchange market performance because data quality breakdowns usually produce sudden shifts in dependence structures without activating standard threshold notifications. The spectral analysis of tick data stream frequencies successfully detects artificial patterns and algorithmic effects in the market rather than natural market influences. Fractional differentiation procedures in long-memory models can identify delayed changes in persistence patterns which indicate emerging data quality problems [6]. Such advanced time-series techniques enable traders to intervene proactively before data problems affect trading decisions or risk assessments.

Through cross-vendor comparison methodologies, analysts can implement a strong validation system that relies on the fact that independent market data points should match closely during normal market activity although they diverge when feed-specific anomalies occur. Multiple strategies using consolidated tape data and proprietary feeds were reviewed in research studies which showed different sensitivity levels and required computational requirements [5]. Consensus-based methods create a synthetic reference value through the aggregation of multiple market inputs which alerts users about major deviations in singular feed data from the consensus view. The correlation-based methods check how statistical elements align between sources before raising warnings about extreme deviations from established mathematical patterns. Equity market implementations show three major distinct classifications of inconsistencies between cross-feed data which demand specialized response procedures: temporal misalignment occurs when data arrives with different delays between feeds while content divergence emerges when data values differ and structural differences exist when the data structure is fundamentally different [6]. These advanced detection solutions apply historical reliability weights to feeds based on their past accuracy performance which enhances their weightage towards feeds known for their precision in anomaly times. The adaptive weighting system enhances multi-source validation approaches by increasing their resistance to feed quality fluctuations in stressful market conditions.

Machine learning approaches have dramatically advanced anomaly detection capabilities beyond what traditional statistical methods can achieve, particularly for complex pattern recognition in high-dimensional tick data. Supervised learning implementations using gradient-boosted decision trees have demonstrated exceptional performance when sufficient labeled examples of anomalous patterns are available [5]. These models incorporate diverse feature sets spanning raw data properties (price jumps, volume spikes), derived metrics (spread dynamics, trade-to-quote ratios), and contextual indicators (time-of-day effects, proximity to news releases). Unsupervised approaches employing density-based clustering techniques effectively identify observations that deviate from typical data density regions in multidimensional feature space without requiring labeled training examples. Semi-supervised methods combining small sets of labeled anomalies with large volumes of unlabeled data have proven particularly valuable for operational implementation, balancing detection accuracy with practical training data requirements [6]. Deep learning architectures specialized for sequential data, including temporal convolutional networks and attention-based models, capture complex dependencies across tick sequences that simpler models might miss. Transfer learning techniques enable knowledge sharing across related instruments, allowing detection models trained on liquid instruments with abundant anomaly examples to be effectively applied to less liquid instruments where anomalies occur too infrequently to support robust model training.

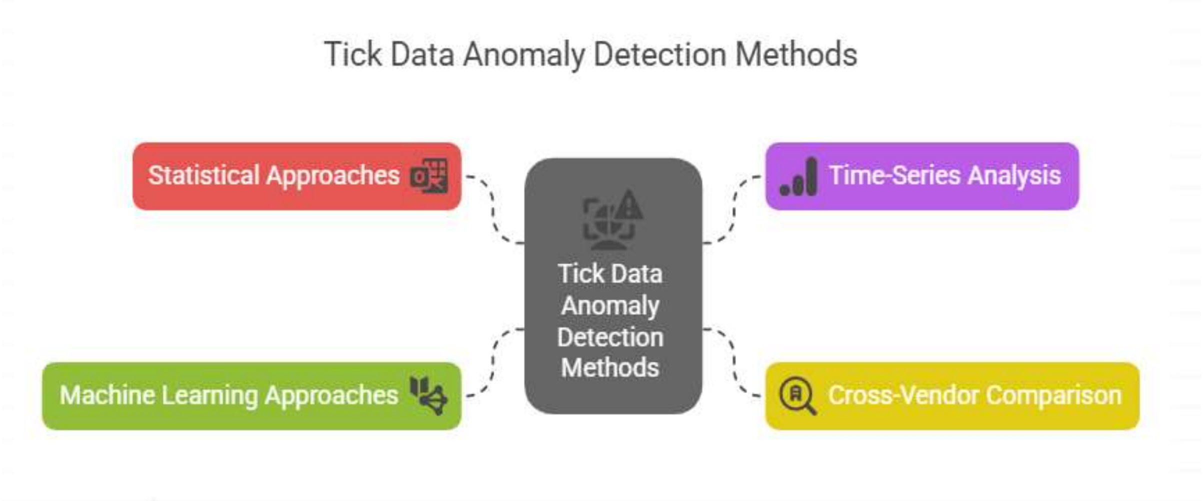
Tick Data Anomaly Detection Methods
Statistical Approaches
Time-Series Analysis
Tick Data Anomaly Detection Methods
Machine Learning Approaches
Cross-Vendor Comparison

Fig 2: Tick Anomaly Detection Methods [5, 6]

Correction Strategies and Reconstruction Methods

The operational trading environment requires real-time filtering and smoothing algorithms which act as essential components for preserving tick data integrity. The effectiveness of error correction methods depends on achieving three fundamental targets according to research: it must filter out random trends while keeping accurate movements, operate quickly for real-time needs, and protect the underlying statistical data-generating process properties [7]. Whenever different filtering methods get reviewed based on these objectives they show varying degrees of effectiveness. Moving median filters demonstrate both strong resistance to single outliers and maintain abrupt price changes which indicate real market activities. State-space models equipped with time-varying parameters establish automatic adjustments to market dynamics which enable them to handle natural fluctuations in financial time series volatility. The application of statistical estimation methods together with rule-based filtering proves best for diverse market environments according to asset class comparisons. The hybrid systems use a stepwise decision tree which applies advanced correction methods only after simpler techniques fail to work effectively therefore achieving maximum resource efficiency within maintenance of correction quality [8]. Error-correction models with equilibrium-based relationships prove especially effective at filtering data by determining between temporary and systematic issues for instruments that demonstrate strong mean-reversion patterns.

Market dynamics and data gap features determine which interpolation technique should be used to reconstruct segments of missing data. The analysis of cointegrated time series indicates that data reconstruction methods should keep both individual statistical properties as well as steady-state associations between connected instruments while executing the process [7]. Different interpolation technologies suitable for gaps between time periods work at their best based on what type of market conditions that exist around them. The local fitting approach works on polynomials allowing users to maintain the natural tick movement curves yet evades erratic results that higher polynomial approaches sometimes generate. The interpolation of multivariate datasets succeeds through methods that sustain cross-sectional relationships instead of univariate approaches that work independently. Time-series instruments featuring periodic patterns are better analyzed using frequency domain techniques with Fourier components because these methods detect cyclic patterns that time-domain methods cannot identify. The advanced reconstruction frameworks use adaptive approach selection mechanisms which detect gap features alongside market environments to pick suitable methods [8]. Critical applications require ensemble approaches uniting various reconstruction methodologies which generate superior results by exploiting synergistic properties of mathematical approaches to reduce errors across multiple situations.

The process of timestamp synchronization alongside sequencing corrections solves basic problems which occur when market infrastructure breaks down into separate parts. The research shows that cointegrating patterns between markets can reveal artificial trading opportunities because these outcomes stem from inconsistent timestamps rather than market inefficiencies [7]. Timestamp correction frameworks that detail effectiveness must process temporal distortions coming from varying network path transmission latencies along with exchange system processing delays and distributed infrastructure clock synchronization issues. Vector error correction models enable researchers to analyze relative time relationships between price data through the assessment of cointegrated processes that feature temporary deviations from their equilibrium states. Implementation architectures use a reference timeframe for data alignment which usually stems from the most dependable data source or the data feed with minimal latency while performing both fixed and adaptive offset functions [8]. International tick data presents extensive synchronization difficulties because time delays increase massively from geographic distances as well as regulatory obstacles to having co-located servers. The modern correction systems combine physical network models of latencies with statistical tests to calculate cross-market delays for detecting synchronization issues that need adjustment.

The requirement to maintain complete audit records of data modification applies under regulations while it functions as a vital aspect of data governance systems. Cointegration research demonstrates that documented transformation methods create vital conditions to interpret processed data correctly before performing econometric analyses [7]. Tracking data modification events through effective auditing requires documentation of three aspects: detection conditions that initiated the change and algorithm methods used and parameters set and the resulting data modifications from each correction process. The detailed logs provide the ability for both instant decision reconstruction and long-term analysis of correction patterns to reveal potential systemwide problems in raw data resources. The storage infrastructure for unmodified original data in implementation platforms uses immutable storage to preserve full version histories that enable retrieving any previous data version [8]. The audit capabilities of these methods provide financial institutions with both internal control functions and external compliance verification which enable demonstration of systematic data preparation method consistency across all data transformation processes. The advanced systems use automated reconciliation techniques that validate adjustment patterns against preset distribution protocols to raise potential issues for human assessment.

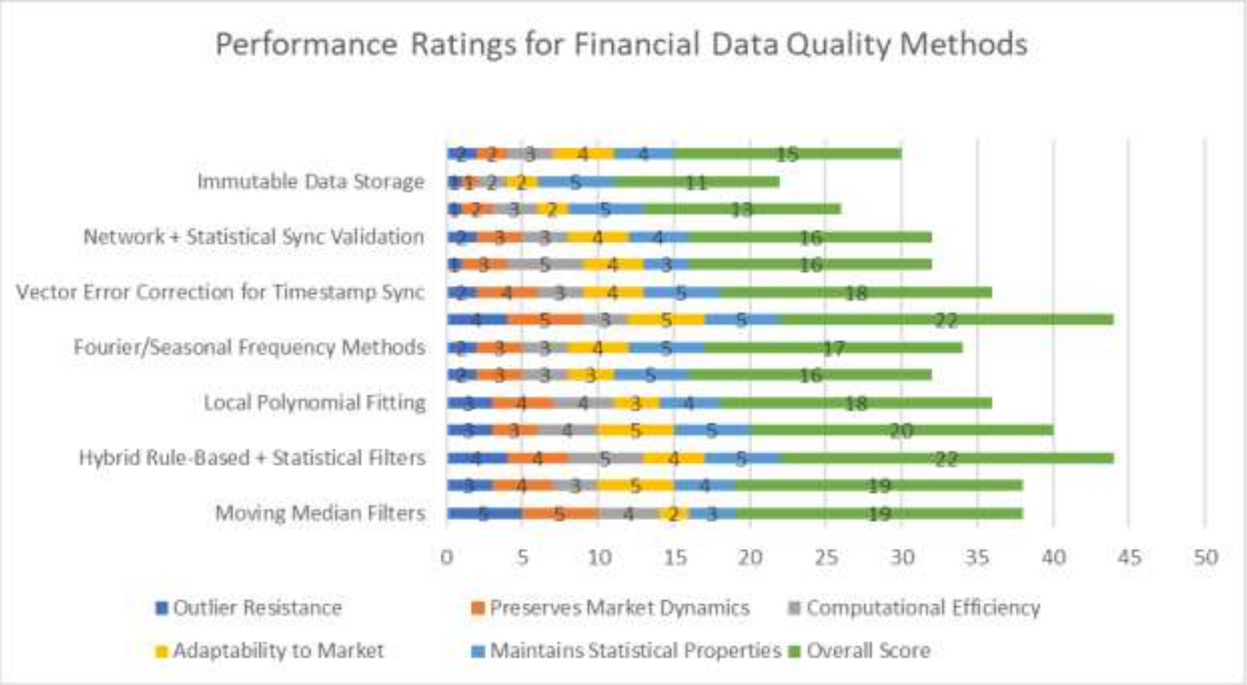

Fig 3: Performance Ratings for Financial Data Quality Methods [7, 8]

Implementation Architecture for Quality Control Systems

Data pipeline design for high-throughput tick data environments requires specialized architecture that can process massive data volumes while maintaining minimal latency. Recent research examining domain-specific hardware architectures demonstrates that conventional general-purpose computing infrastructures often become bottlenecks when processing the extreme message rates characteristic of modern markets [9]. Effective tick data pipeline implementations typically segment processing into distinct stages with specialized optimization for each phase. The ingestion layer employs memory-mapped I/O techniques to minimize data copying operations, often implementing custom network stack optimizations that bypass standard operating system protocol implementations. These specialized network stacks reduce processing overhead by eliminating unnecessary protocol features irrelevant to market data transmission. The normalization layer converts diverse vendor formats into standardized internal representations, typically implementing lookup-based field mapping rather than conditional logic to accelerate transformation operations. Domain-specific hardware accelerators can be particularly valuable for these normalization processes, as the operations involve predictable transformation patterns amenable to hardware optimization. Quality control processing represents the most computationally intensive pipeline stage, often implementing multiple detection algorithms executing in parallel across the data stream. Research into specialized processing architectures indicates that custom matrix processing units can accelerate many common statistical operations used in anomaly detection, providing orders of magnitude performance improvement for specific computational patterns common in financial data analysis [9].

Parallel processing architectures for tick data quality control must address both data parallelism and pipeline parallelism to achieve optimal throughput. Research examining large-scale analytical systems demonstrates the importance of workload-specific parallelization strategies rather than generic distributed computing frameworks [10]. Effective implementations often employ a multi-level parallelization approach, with coarse-grained partitioning at the instrument or market level allowing independent processing across separate computation nodes. Within each node, finer-grained parallelism leverages multi-core architectures through careful workload division that maintains data locality and minimizes cross-core synchronization requirements. For detection algorithms requiring significant computational resources, specialized hardware accelerators such as field-programmable gate arrays (FPGAs) or tensor processing units (TPUs) can be integrated into processing pipelines, executing specific computational kernels with dramatically higher efficiency than general-purpose processors. Research into domain-specific accelerators indicates that specialized hardware can achieve performance improvements ranging from 10× to 50× for specific computational patterns common in financial data processing, though significant engineering investment is required to effectively integrate these components into production systems [9]. For globally distributed market data processing, geo-distributed architectures implementing regional processing nodes help minimize network latency while maintaining a consolidated global view, though such architectures introduce additional complexity in data synchronization and consistency management across regions.

Database optimization for tick data storage presents unique challenges due to the combination of extremely high write rates, diverse query patterns, and long-term retention requirements. Research examining column-oriented storage architectures demonstrates particular advantages for time-series data with regular structure and infrequent schema changes [10]. Effective tick database implementations typically implement multi-tiered storage strategies, with recent data maintained in memory-optimized structures for low-latency access while older data transitions to storage-optimized formats. Partitioning strategies specific to market data requirements often implement hierarchical schemes with primary partitioning by period and secondary partitioning by instrument or market segment, enabling efficient pruning during query execution. For historical analysis requiring access to extended periods, research into columnar compression techniques demonstrates that domain-specific encoding methods that exploit the statistical properties of financial time series can achieve significantly higher compression ratios than general-purpose algorithms. These specialized compression approaches include delta encoding for monotonically increasing timestamps, dictionary encoding for repeated values such as instrument identifiers, and specialized floating-point compression for price data that preserves exact values while eliminating redundancy [10]. Query optimization for tick databases requires specialized techniques that account for the distinctive access patterns in financial analytics, including time-based slicing, cross-sectional aggregation, and event-based alignment that standard database query optimizers may not handle effectively.

Monitoring infrastructure and alert mechanisms constitute critical components of operational tick data quality systems, enabling both real-time intervention and continuous improvement. Research examining interactive analysis systems for large-scale data processing highlights the importance of multi-layered telemetry capturing metrics at different abstraction levels [10]. Comprehensive monitoring frameworks typically implement four distinct metric categories: infrastructure metrics capturing system resource utilization, pipeline metrics documenting data flow rates and processing latency, quality metrics quantifying anomaly detection and correction activities, and business impact metrics assessing the financial significance of quality issues. Effective alerting systems implement progressive notification strategies with distinct thresholds for different severity levels, avoiding alert fatigue through careful calibration of notification criteria. Machine learning approaches have demonstrated particular value for monitoring complex tick data processing systems, establishing adaptive baselines that accommodate normal variations in data patterns across different market conditions without generating false positives. Research into domain-specific architectures indicates that the computational patterns involved in continuous system monitoring align well with specialized hardware acceleration, enabling comprehensive real-time analysis of system behavior without imposing significant overhead on primary processing paths [9]. Visualization frameworks tailored to tick data quality monitoring typically implement multi-resolution views that enable both broad situation awareness and detailed drill-down capabilities, supporting both operational monitoring and forensic analysis of historical quality issues.

Metric	Performance Improvement	Processing Efficiency
Network Stack Optimization	30%	85%
Matrix Processing Units	100x	75%
Specialized Hardware Acceleration	10x	60%
Domain-Specific Accelerators	50x	90%
Memory-Optimized Structures	5x	80%
Columnar Compression	75%	65%
Adaptive Baseline Monitoring	40%	95%

Table 1: Performance Metrics for Tick Data Processing Systems [9, 10]

Conclusion

Effective tick data quality control requires an integrated perspective combining robust detection methodologies with appropriate correction strategies, all supported by specialized technical infrastructure designed for high-throughput processing. The multi-dimensional nature of data quality assessment necessitates frameworks that can adapt to changing market conditions while maintaining consistent standards across asset classes and trading venues. As market fragmentation continues and data volumes expand, implementation architectures must evolve toward greater specialization, with hardware acceleration and parallel processing becoming increasingly essential components. The integration of machine learning capabilities offers significant potential for identifying subtle anomaly patterns and establishing adaptive baseline metrics for monitoring purposes. While substantial progress has been made in methodological sophistication and technical implementation, challenges remain in timestamp synchronization across global markets, management of increasingly diverse data sources, and quantification of quality impact on trading performance.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Ananth Madhavan, "Market microstructure: A survey," ScienceDirect, 2000. https://www.sciencedirect.com/science/article/abs/pii/S1386418100000070

[2] Carol Alexander and Anca Dimitriu, "Sources of Over-Performance in Equity Markets: Mean Reversion, Common Trends and Herding," SSRN, 2003. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=413780#:~:text=A%20behavioural%20explanation%20for%20the,common%20trend%20in%20stock%20returns.

[3] Eric Zivot and Jiahui Wang, "Modeling Financial Time Series with S-PLUS," 2006. https://faculty.washington.edu/ezivot/econ589/manual.pdf

[4] James J. Angel et al., "Equity Trading in the 21st Century: An Update," World Scientific Connect, 2015. https://www.worldscientific.com/doi/epdf/10.1142/S2010139215500020

[5] Katarina Juselius and David Hendry, "Explaining Cointegration Analysis: Part II," ResearchGate, 2000. https://www.researchgate.net/publication/5161060_Explaining_Cointegration_Analysis_Part_II

[6] Norman P. Jouppi et al., "A domain-specific architecture for deep neural networks," ResearchGate, 2018. https://www.researchgate.net/publication/327198532_A_domain-specific_architecture_for_deep_neural_networks

[7] Rama Cont, "Statistical Modeling of High-Frequency Financial Data: Facts, Models, and Challenges," SSRN, 2011. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1748022

[8] Sergey Melnik et al., "Dremel: Interactive Analysis of Web-Scale Datasets," Proceedings of the VLDB Endowment, 2010. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/36632.pdf

[9] Terrence Hendershott and Ryan Riordan, "Algorithmic Trading and the Market for Liquidity," Journal Of Financial And Quantitative Analysis, 2013. https://faculty.haas.berkeley.edu/hender/ATMonitor.pdf

[10] Torben G. Andersen et al., "Modeling And Forecasting Realized Volatility," Econometrica, 2003. https://www.sas.upenn.edu/~fdiebold/papers/paper43/abdl4.pdf