

Historical Stock Data Quality Report: Comparison with NYSE Daily TAQ

HistoricalData.net | 22 October 2026 | Version 1.0 | Published 2026-10-22

What this comparison establishes

This report measures agreement between the identified HistoricalData.net stock files and bars reconstructed from a publicly downloadable NYSE Daily TAQ transaction sample. It distinguishes numerical agreement from accuracy and from the correctness of an aggregation convention.

Key finding: On 1 July 2026, 10,908 of 10,917 comparable daily closing prices (99.9176%) agreed exactly with the documented NYSE Daily TAQ reconstruction; 10,915 (99.9817%) differed by no more than USD 0.01. These are sample price-concordance results, not a full-product accuracy rate.

Daily raw field	Exact agreements / comparable pairs	Exact agreement	Within USD 0.01	P95 absolute difference
open	10,896 / 10,917	99.8076%	99.9817%	0.000
high	10,872 / 10,917	99.5878%	99.9725%	0.000
low	10,874 / 10,917	99.6061%	99.9908%	0.0000
close	10,908 / 10,917	99.9176%	99.9817%	0.00

How to cite: retain the trading date, field, numerator, denominator and the phrase "documented NYSE Daily TAQ reconstruction." The percentages are not an estimate of the accuracy of the full historical archive.

Sources and audit scope

The primary reference contains **154,938,670 transaction records** for 12,659 source symbols on **1 July 2026**. The 4,447,875,673-byte compressed file was read to EOF, its gzip integrity verified, and its trailer count checked. NYSE distributes consolidated CTA and UTP trades; the sample is not limited to executions on the NYSE venue.

The HDN capture fully verified 29,386 published files before using their selected rows. Historical daily observations were extracted from by-symbol archives because July was outside the rolling by-date window. This verifies input identity, not market correctness.

Daily join measure	Count
Mapped HDN rows	10918
Reference rows after stated exclusions	12635
Matched security/day rows	10918
HDN-only rows in mapped set	0
Reference-only rows	1717
HDN rows without a same-day master mapping	0
Symbols excluded for unknown/reserved condition semantics	0

One matched security/day had no eligible reconstructed OHLC tuple and is excluded from the four price denominators. Its missing prices are not replaced by zero. Reference-only instruments can be outside the product scope or its current publication universe. They are reported as unmatched, not automatically classified as missing sold data.

Minute sample and coverage

The formal minute comparison uses these 19 available files from the selection made before observing numerical results: AAPL, MSFT, NVDA, TSLA, AMZN, META, GOOGL, JPM, XOM, JNJ, KO, PG, WMT, BRK.B, SPY, QQQ, IWM, DIA and TLT. The selection spans large common stocks and several ETF categories; it is not a random or market-representative sample. BRK B maps to BRK.B and GOOG L maps to GOOGL using the same-day NYSE master.

All 19 files were read and hashed in full. The formal comparison is restricted to available files; unavailable-file selection evidence is retained in the internal audit record. No substitution was made based on agreement results.

Session, SIP clock	HDN bars	Reference priced bars	Matched	HDN only	Reference only
pre	3372	3372	3372	0	0
reg	7410	7410	7410	0	0
post	2266	2335	2266	0	69

The reference additionally produced 3,186 buckets without a complete eligible OHLC tuple. These are not treated as priced bars or as zero-valued missing bars.

Regular-session minute results

Field	Exact agreements / comparable pairs	Exact agreement	Within USD 0.01	P95 absolute difference
open	7,380 / 7,410	99.5951%	100.0000%	0.0
high	7,395 / 7,410	99.7976%	99.9730%	0.00
low	7,390 / 7,410	99.7301%	99.9595%	0.00
close	7,385 / 7,410	99.6626%	99.9865%	0.00
volume	0 / 7,410	0.0000%	Not a price field	2.653167
transactions	280 / 7,410	3.7787%	Not a price field	1

Regular session means 09:30 through 15:59 Eastern minute starts. Full pre-market and post-market counters, and per-symbol results, are included in results.json. Volume and transaction counts are descriptive diagnostics; they do not establish that the vendors use identical counting rules.

Trade counts differ by exactly one in 7,104 of 7,410 comparable regular-session pairs. Volume differences run in both directions. Their causes are not established; the reported decimal precision does not by itself explain them.

Timestamp sensitivity

Regular-session field	SIP exact agreement	Participant exact agreement
open	99.5951%	95.6005%
high	99.7976%	99.6221%
low	99.7301%	99.5007%
close	99.6626%	95.4251%
volume	0.0000%	0.0000%

SIP Time is the primary clock fixed before comparison. Participant Timestamp is a separately reported sensitivity analysis, with 7,410 matched regular-session bars. Moving a trade across a minute boundary can change OHLC and volume. The alternative clock is not chosen retrospectively as the headline result.

Daily quantity diagnostics

Field	Exact agreements / comparable pairs	Exact agreement	Within 0.1% / nonzero-reference pairs	P95 absolute difference
volume	2 / 10,918	0.0183%	9384 / 10916	1075.163500
transactions	1,944 / 10,840	17.9336%	6814 / 10840	17

Quantities are actual shares, including decimal fractions; no assumed 100-share multiplier is applied. Daily volume includes eligible extended-hours trades. Counts omit zero-volume official open/close messages according to the documented condition policy. A remaining mismatch is not assigned a cause without trade-level evidence.

Supplementary Cboe interval check

A separate Cboe public sample contains 27 IWM 15-minute records on 1 June 2016, labelled 09:45 through 16:15 Eastern. HDN minute bars are rolled into half-open intervals ending at those labels: first open, maximum high, minimum low, last close and summed volume. The final interval extends beyond 16:00; all 27 and the 26 intervals through 16:00 are reported separately.

Field, all 27 intervals	Exact agreements / comparable pairs	Exact agreement	Within USD 0.01	P95 absolute difference
open	16 / 27	59.2593%	85.1852%	0.0200
high	21 / 27	77.7778%	88.8889%	0.0300
low	16 / 27	59.2593%	70.3704%	0.6201
close	19 / 27	70.3704%	96.2963%	0.0100
volume	0 / 27	0.0000%	Not a price field	154449

This is a small supplementary diagnostic, not proof of complete 2016 coverage. Bid/ask and VWAP are not compared; the published minute VWAP field is deliberately unavailable.

Reconstruction and comparison method

- **Pin input identity.** Capture immutable publication metadata and verify full archive SHA-256 and byte counts before accepting extracts. Verify the entire reference gzip and its END date/count record.
- **Resolve securities.** Join NYSE Symbol to SIP_Symbol and published symbol history for the trading date. Exclude test symbols. Reject ambiguous same-security/date rows instead of selecting one.
- **Apply corrections.** Retain TAQ correction codes 00 and 01; discard 07, 08, 10, 11 and 12. A code-01 row contains corrected data at the original record time. Unknown correction codes stop processing.
- **Apply field eligibility.** Use the CTA/UTP consolidated condition matrices. A disqualifying condition overrides an eligible one. High/low and open/close eligibility differ. Official venue open/close markers do not automatically update consolidated values. Special first/only-last and late-last rules are explicit in `aggregate_reference.py`.
- **Separate timescales.** Daily state uses the SIP sequence. Minute intervals use Eastern start labels and `[start, start + 60 seconds)`. Form T may set prices outside regular hours, while other disqualifying conditions remain effective. Minute special-condition rules are an explicit reconstruction convention, not a claim of official exchange bar semantics.
- **Preserve precision.** Parse prices and share quantities as scaled integers to six decimals. Leading-decimal fractions such as .937 are valid. Missing numeric values are counted, never converted to zero. Values exceeding supported precision fail instead of silently rounding.
- **Compare raw fields.** Compare unadjusted OHLC to unadjusted reconstructed prices. Adjusted values and corporate actions are outside scope. Matched, nonmissing positive-price pairs form price denominators. Zero volumes are valid.

Formulas: exact rate = $100 \times \text{count}(\text{HDN} = \text{reference}) / \text{comparable pairs}$. Absolute difference = $|\text{HDN} - \text{reference}|$. Relative difference = $\text{absolute difference} / |\text{reference}|$ when $\text{reference} \neq 0$. Price tolerance = USD 0.01; quantity diagnostic tolerance = 0.1%. These fixed descriptive bands do not redefine exact equality. A zero denominator gives null. P95 uses sorted differences at one-based rank $\text{ceil}(0.95 \times N)$.

The requested 2022 TAQ specification was supplemented with its 2024 revision and the linked CTA 2026 / UTP 2025 specifications. Actual 2026 headers, decimal quantities, correction codes and master fields were inspected. Old field-width descriptions were not assumed to prohibit modern fractional values.

Reproducibility and machine-readable evidence

Each claim in `claims.json` includes its date/scope, numerator, denominator and JSON evidence path. `results.json` contains aggregate counters by field, session, clock and symbol. `sources.json` and `published-file-identities.json` identify the exact source files. `method.json` and `cohort.json` preserve the prespecified analysis rules.

Run the supplied `reproduce.py` with the exact reference files and licensed HDN archives identified by hash. The public package excludes raw trades, individual price rows and credentials. A later revised HDN file must not be substituted while retaining these statistics; regenerate a new version.

```
python reproduce.py --help
```

Hash verification establishes that the same bytes were compared. It does not by itself establish that those bytes accurately describe the market.

Limits and conclusion

The report provides measured, field-specific agreement with identified third-party samples. It does not support a single "data accuracy" percentage for the entire product. The daily and minute tables show which tested fields agree, how frequently they agree, and where unresolved numerical differences remain.

- The reference and HDN upstream may share exchange/SIP inputs; agreement does not independently validate every upstream market event.

- A one-day NYSE trade file and a purposive minute cohort do not establish archive-wide accuracy. Repeated bars and fields are correlated observations, not independent samples.
- Prepared by HistoricalData.net, the supplier under evaluation. This is a self-assessment using third-party inputs, not an independent audit or NYSE/Cboe certification.
- The NYSE reference is consolidated transaction data. Our OHLC reconstruction is documented code, not an official NYSE aggregate or a uniquely correct market price.
- Minute OHLC depends on time bucketing, eligibility, late messages and special-condition handling. Disagreement is measured; its cause is not automatically attributed to HDN or the reference.
- Daily volume, trade counts, minute volume and the Cboe interval supplement remain diagnostic unless their aggregation definitions are fully aligned.
- Raw OHLC columns only. Adjusted prices, dividends, splits, VWAP, total-return calculations and other dates are not validated.
- Current file labels may differ from historical tickers. Only explicit published symbol history and same-day master mappings are used; unmatched or ambiguous identifiers are not guessed.
- Public-sample availability does not establish permission to redistribute raw data. This package contains code, hashes and aggregate comparison evidence only.

Source register

- **S1** - NYSE Daily TAQ public trade sample, 1 July 2026. Primary reference, consolidated trades rather than ready-made OHLC bars. Accessed 22 October 2026.
- **S2** - NYSE Daily TAQ security master, 1 July 2026. Exact Symbol to SIP_Symbol mapping, test flag, security type, listing and share unit. Accessed 22 October 2026.
- **S3** - NYSE Daily TAQ Client Specification v4.0 (2022), requested starting reference. Initial schema reference; superseded here where v4.1 clarifies definitions. Accessed 22 October 2026.
- **S4** - NYSE Daily TAQ Client Specification v4.1, 25 April 2024. Eastern time, file trailer, trade corrections and symbol mapping; actual 2026 fields were inspected. Accessed 22 October 2026.
- **S5** - CTA CTS Output Specification v2.11b, 29 January 2026. Consolidated condition matrix and footnotes, pp.82-83; fractional-volume messages. Accessed 22 October 2026.
- **S6** - UTP Data Feed Services Specification v3.0a, November 2025. Consolidated condition matrix and footnotes, pp.50-54. Accessed 22 October 2026.
- **S7** - Massive aggregate construction documentation. Upstream bar definitions and differences by trade eligibility, timestamps and corrections. Accessed 22 October 2026.
- **S8** - Massive Form T treatment in extended-hours bars. Form T price eligibility outside regular hours; other disqualifying conditions remain effective. Accessed 22 October 2026.
- **S9** - Cboe Equity & ETF Quotes sample. Supplementary 27 IWM 15-minute records, 1 June 2016. Accessed 22 October 2026.
- **S10** - Cboe Equity & ETF Quotes layout. Eastern end-of-interval labels and interval OHLC/volume. Accessed 22 October 2026.
- **S11** - NYSE Daily TAQ product page. Reference product scope. Accessed 22 October 2026.
- **S12** - NYSE market data policies and licensing documents. Public access is not treated as a redistribution license. Accessed 22 October 2026.