LSEG EXPERT TALK

What does it take to achieve high quality data?

Introduction

The need for very high quality exchange data, either raw or lightly normalised, has never been greater than it is today. Today, LSEG Tick History – PCAP data is captured at exchange data centres with nanosecond precision, and in a lossless manner to ensure the highest quality. And, of course, data quality is not just a point-in-time objective – it is something that needs to be delivered every day and constantly enhanced through technology changes and innovation.

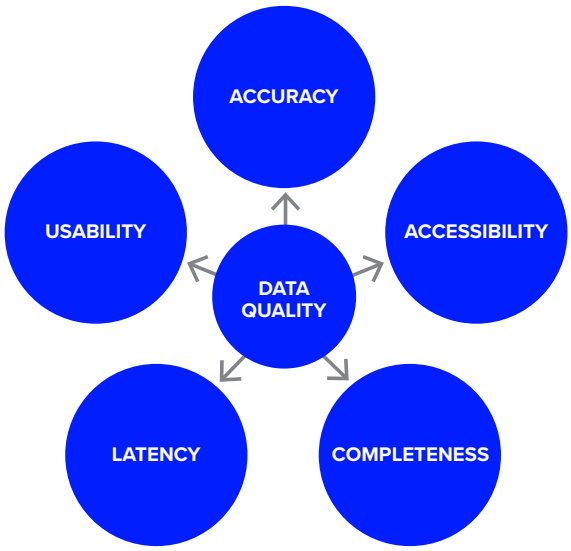
With the acquisition of MayStreet in May 2022 by LSEG there is significant potential to leverage LSEG's scale and continue to grow Tick History – PCAP data across several dimensions.

This Expert Talk explores key elements around data quality for PCAP tick data. It also looks ahead to how the data quality journey will continue to evolve for Tick History – PCAP within LSEG.

LSEG DATA & ANALYTICS

What is data quality?

Fundamentally, good quality data is data that is fit for purpose – the data needs to support the outcomes it is being used to achieve. So, to determine data quality, the condition of needs to be assessed based on factors such as accuracy, accessibility, completeness, latency and usability in the context of the particular use cases it will be applied to. So, the market data provided by Tick History – PCAP needs to be of extremely robust quality. of approaches.



What standard of data quality is needed for demanding tick data use cases?

To better understand the importance of data quality in the context of PCAP data, it is helpful to explore the concept in more depth by looking at the components of data quality. These are:

Accuracy – To be accurate, the market data needs to be valid. That is, the data must contain the values disseminated by the exchange and it needs to be of the highest fidelity.

For Tick History – PCAP, this means collecting the exchange's native format into a packet capture (PCAP) file with nanosecond-precision hardware timestamps – based on exactly what the execution venue published. Also, to achieve accuracy, the data collection must be on the direct connections to the exchanges, i.e., cross-connects on the fastest networks. Many of our data feeds are captured at sites that have Layer 1 switches to assign a timestamp directly at ingestion from the exchange cross-connect.

Additionally, the capture of feeds must happen at both primary and secondary data centres to ensure lossless data collection. Timestamps are GPS-synchronised with redundant clock sources. These features ensure that Tick History – PCAP data is as accurate as possible.

Accessibility – Quality data is accessible when and where you want it.

For Tick History – PCAP, the data – raw or lightly normalised – is available in a timely manner (intra-day or T+1) at the location of the firm's choosing: on-premises, in the cloud or hybrid cloud.

With Tick History – PCAP's shared storage offering firms can avoid the hassle and cost of storing a copy of the data and simply access PCAP data straight from our cloud storage buckets. By leaving data management and storage to LSEG, clients have immediate access to the highest quality PCAP or normalised data in our cloud storage buckets, combined with ease of use, easy integration and significant data storage savings.

Completeness – This means that data sets must contain all of the data elements provided by the exchange and that all data required for a use case is available. With Tick History – PCAP, completeness both within a feed and across feeds is required. Completeness within a feed requires redundant capture across multiple sites, including capturing A/B lines to ensure a clean and complete data stream. Completeness across feeds requires that the full range of Level 1, 2 and 3 data is captured. This includes depth-of-book feeds across all major markets, such as equities, equity options, derivatives, futures, commodities, and more.

Latency – By capturing the data in colocation with the exchange, we capture market data as the data comes out of the exchange, with minimum network latency & high-performance network infrastructure (switches & routers). Tick History – PCAP uses Real-Time – Ultra Direct, LSEG's ultra-low latency software library, to capture market data and store in PCAP format

Usability – This is the ability of financial services firms to use data in the format that they prefer. Tick History – PCAP is offered in PCAP, CSV, and Parquet formats that support front office and data science applications.

However, usability is about more than data. Having the right support is also important. Tick History – PCAP is supported by a team with deep expertise in this data. The importance of data quality for demanding use cases cannot be underestimated. Tick History – PCAP delivers data captured directly from exchanges with high levels of accuracy, availability, completeness, and usability, as well as low levels of latency, so that firms can trust the data for the wide variety of use cases they deploy it for.

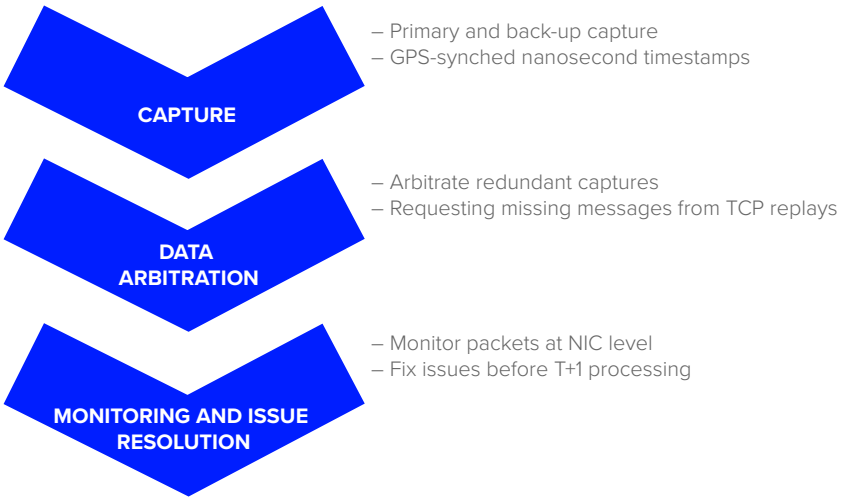
How does the data capture process support data quality?

LSEG captures and processes PCAP data in a range of different ways to ensure that the data is truly lossless. Tick History – PCAP captures raw packets from execution venues and other data sources (e.g., OPRA) and stores them in PCAP format, maintaining fidelity with what the exchange publishes, since an exact copy of the data distributed in its native form has been captured. Features of our capture process include:

- **Redundant & perpetual capture** – Data is captured from multiple lines, at multiple sites. Tick History – PCAP co-locates within execution venue data centres globally to capture data with the lowest possible network latency. In addition to capturing at primary sites, Tick History – PCAP also captures at one or more secondary sites for most venues to ensure completeness and resiliency. Redundant capture is an essential component of Tick History – PCAP's approach to data quality management. Approximately 500 unique data feeds are ingested in the capture systems. Every feed has at least one backup. The most-backed up feeds – the U.S. Equity SIP feeds – each have more than 30 unique captures associated with them. As part of the process, Tick History – PCAP captures both Line A and Line B of each feed for arbitration. The same is done for C, D, and E if they are offered. Tick History – PCAP always captures data at the highest-speed shape (if applicable) to ensure the quality of timestamps remains unparalleled. Data is captured 24/7/365. Even if a feed isn't expected to be publishing, Tick History – PCAP is still listening, just in case.
- **Nanosecond, GPS-synchronized timestamp granularity** – By capturing nanosecond timestamps, analysts are able to drill down to the most granular level and reconstruct the market with precision. When the sequence of orders and trades matters, nanosecond timestamps and exchange sequence numbers provide the clarity that is needed. The Tick History – PCAP team has successfully conducted analysis for clients who wanted options quotes with the underlying equity quotes in the same snapshot. Using nanoseconds, the Tick History – PCAP team deterministically matched individual options quotes to underlying quotes down to the single quote level.
- **Proprietary Capture Technology** – Tick History – PCAP's packet capture process is powered by Real-Time – Ultra Direct, a suite of tools and feed handlers developed in close collaboration with banks, quantitative asset managers, and other financial services industry clients. Our capture technology is tested in production with the most demanding clients.
- **Data & Engineering Expertise** – Beyond the technology, the Tick History – PCAP team has engineers responsible for the capture technology and processes that are thoughtful, highly skilled and extremely detail-oriented. The team monitors numerous metrics during live capture, including server health, network health, time accuracy and software performance. Even the smallest blip generates an actioned alert, and packet drops are treated like a bug, ensuring that root causes are addressed.

Beyond data capture, what are the additional steps in the ingestion process required to achieve data quality?

Simply capturing the data is not enough to achieve data quality, as illustrated by the end-to-end process flow.



The Data Arbitration Process

During the T+1 processing, Tick History – PCAP automatically uses primary and backup captures at the primary data centre only where available and when there would be no impact on processing time. Arbitrated data is examined for drops the next morning – if any are found, this is investigated to determine whether including backup sites would be beneficial.

Many drops actually occur at the exchange level, which TCP replays can help remedy.

Over the coming months, we plan to include TCP replays with all of our captures. If we find a drop, we will automatically capture the replays and deliver those with the feed. Longer term, we plan to automatically perform the drop examination processes before arbitration, allowing us to determine the best mix of sites to include in that night's arbitration.

Monitoring & Issue Resolution Process

Throughout the trading day, Tick History – PCAP monitors the networks for any packet drops at a NIC level (for example, if there has been an incident in a data centre). If any issues are detected, the Tick History – PCAP team modifies that night's processing to instead pull from non-degraded captures. Malformed or incomplete messages are detected at the network level to maintain the integrity of the data.

Issues not detected during the day – for example, a minor exchange issue that is unnoticed during market hours – will be caught T+1. The Tick History – PCAP team checks for drops daily during the early-morning check, then fill any drops from backup captures. The team also reviews sequence numbers daily to determine if there are any drops. If needed, they will request retransmission from the execution venue to fill in any gaps.

Longer term, our strategic development roadmap calls for us to automatically use the best captures, removing the need for manual intervention. We also plan to check for timestamp quality the same way we do for drops.

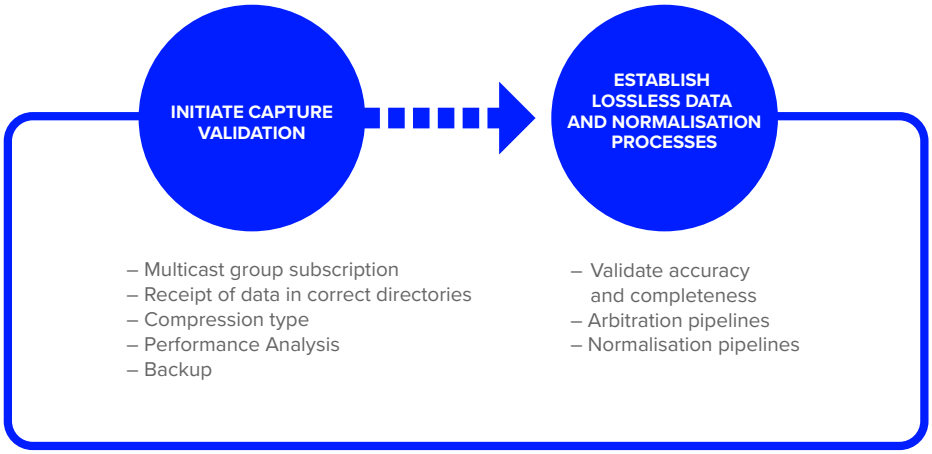
By monitoring Precision Time Protocol (PTP) live throughout the day and then cross-checking the time with our files on a T+1 basis, we ensure that all captured packets are time-stamped with nanosecond precision and synchronized to a GPS source.

How are new feeds onboarded to promote data quality?

High quality data begins with a robust onboarding process for new feeds in the primary, co-located data centre. Each aspect of data quality, including establishing hardware timestamping, is put in place as part of the onboarding process illustrated below. This process is repeated for each capture of the data feed.

Over the next year, Tick History – PCAP will be capturing new content via PCAP, leveraging the scale and infrastructure that LSEG can deliver. Announcements about this new data will be made as it becomes available.

New feed onboarding



How is data quality maintained during the normalisation process?

In addition to offering raw PCAP data, Tick History – PCAP also provides data in a lightly normalised format that allows users to work across multiple feeds or within a feed using a standard data dictionary. Normalised messages can be produced in CSV or Parquet formats.

In the process of normalising data, the Tick History – PCAP team monitors the quality of the incoming execution venue data as well as the entire transformation process. Any issues with normalisation are rectified with code updates so the normalisation process can be rerun against the raw PCAP data and any issues resolved.

What else does the acquisition of MayStreet by LSEG mean for the data quality of LSEG Tick History - PCAP

LSEG continuously invests in the technology that supports Tick History - PCAP, and innovates to enhance the feed handler – all with a focus on data quality. Ongoing improvements include:

– Switching to independent time sources – Doing this at every node at each site adds redundancy, reduces latency and improves time accuracy (e.g., Server 1: time source A, Server 2: time source B). This is currently in progress.
– Undertaking implementation improvements – This includes significant improvements in backhaul timing at all sites worldwide. Most deployments are now fully automated, and nearly one-third of upgrades are too. More automation is on the way.
– Upgrading our infrastructure – This will enable us to capture data as near to an exchange's matching engine as possible using capture switches that are 100 nanoseconds from the wire. We now have this in place for US equities, and should have this in place for CME data soon.

Conclusion

Today for most financial services firms, the quality of the historical data that is being used for ultra-low latency use cases needs to be very high. To determine the quality of data, it's important to define what quality means in the particular use case context. Then, it's necessary to examine the teams, processes and technology that support the data. Tick History – PCAP delivers high quality data across all of these assessment criteria, and it is used by dozens of firms around the globe. Moreover, as part of LSEG, significant investment is being made to expand coverage and enhance infrastructure, and will continue to be made in the future.
Our coverage list can be found here.

In short, LSEG Tick History – PCAP is a very robust data source that is continuing to evolve alongside the use cases that firms are focused on.

To request information about Tick History - PCAP, please visit our webpage

Related solutions

Real-Time – Ultra Direct

Real-Time – Direct