

How BlackForge measures the market

Updated August 2026 ·

How is BlackForge's data collected and measured?

BlackForge maintains a live order book for every pair on 9 spot venues from each venue's own WebSocket stream, computes order-book and trade-flow measurements every five minutes, and stores every window forever. Each row carries a quality bitmask that says which measurements to trust. Nothing is aggregated across venues and nothing is a trade call.

Nine venues, measured the same way

BlackForge collects order books and trades for every pair on nine spot exchanges: Binance, OKX, Bybit, Coinbase, Bitget, Kraken, Gate.io, MEXC and KuCoin. That is roughly ~11,800 pairs, and the exact figure drifts by a few pairs a day as venues list and delist.

The transport is WebSocket push, not REST polling. For each pair we maintain the live order book from the venue's diff stream, and we run the trade tape on a separate connection so a book resync never interrupts the trade tally. REST is used only to seed and reconcile the book: venues are compared against a deep snapshot (up to the venue's maximum, 500 levels for most, 5,000 on the venues that serve it) and aligned by sequence number, not wall clock. The result is a book we maintain, not a snapshot we poll.

How a five-minute window closes

The native window is five minutes. A row's timestamp is the UTC wall-clock five-minute boundary that ends the window, derived arithmetically, so any writer produces the identical key and duplicate rows collapse to the latest version in storage. A pair being re-seeded still always emits its row: an absent row means the collector was down, never that the market was quiet.

The nine collectors do not flush at the same instant. Each computes its columns at a deterministic per-venue offset inside the window, so nine flush spikes never land on the database together. The window

label is unaffected; only the moment the book is read shifts.

When each column is measured

This is the part most vendors leave undocumented, and it matters for cross-venue work. Two different clocks are in play, and BlackForge records both honestly rather than pretending everything happened at the window label.

The **trade columns** are bucketed by the exchange's own event time. A trade that printed at 12:04:59 lands in the 12:00 to 12:05 window wherever it arrives, so taker volumes and counts are directly comparable across venues. The **order-book columns** are different: the book is read at the venue's deterministic flush offset, and each reading carries its own observed-at timestamp. Measured on production, that offset ranges from about ten seconds after the boundary on the fastest venue to about 150 seconds on the slowest, so two venues' depth columns for the same window label can be observed up to roughly 150 seconds apart.

NOTE

The practical consequence: compare trade flow across venues freely, but treat two venues' book snapshots for the same label as near-simultaneous, not exact. When a window's events had to be bucketed by arrival rather than event time, the row says so with the `EVENTS_REATTRIBUTED` flag below.

The row-quality bitmask

The answer to “bought data turned out to be garbage” is that every row tells you, in the data itself, which of its measurements to trust. Each row carries a quality field: a bitmask, the names of the bits that are set, and the column families each set bit contaminates. If a bit is set, the columns in its families are placeholders for that window, not measurements, and you should drop or grey them out. No competitor ships this.

The 16 quality bits, their meaning, and what each contaminates

Bit	Value	Flag	What it means	Contaminates
0	1	BOOK_DESYNCED	The order book was not in sync with the venue for this window.	Book churn, depth bands, order ladders
1	2	BOOK_ONE_SIDED	Exactly one side of the book had no quotable level.	Book churn, depth bands, order ladders
2	4	BOOK_CROSSED	Best bid was at or above best ask with both sides present: the book was crossed at flush.	Book churn, depth bands, order ladders
3	8	BOOK_DEEP_STALE	The deep levels of the book were older than the shallow verified region.	Depth bands, order ladders
4	16	SYMBOL_DELISTED	The symbol was delisted or halted at the venue.	Book churn, depth bands, order ladders
5	32	NOT_MEASURED	A column in the contaminated families is a placeholder, not a measurement.	Book churn, depth bands, order ladders
6	64	PRICE_FROM_MID	The candle was derived from the book mid because no trade occurred.	Price, enrichment
7	128	PRICE_FROM_LAST_TRADE	The candle was carried forward from the last known trade.	Price
8	256	PRICE_UNAVAILABLE	No price could be established for this window at all.	Price
9	512	QUOTE_RATE_SUSPECT	The quote-to-USD conversion rate was stale or implausible.	Quality, enrichment

Bit	Value	Flag	What it means	Contaminates
10	1024	IDENTITY_CONFLICT	The symbol resolved to more than one candidate coin.	Enrichment
11	2048	DEPTH_BEYOND_COVERAGE	Depth was requested beyond the range our maintained book actually covers.	Depth bands, order ladders
12	4096	EVENTS_LOST	Book events were dropped in this window.	Book churn
13	8192	EVENTS_REATTRIBUTED	Book events were attributed to this window by arrival rather than event time.	Book churn
14	16384	BOOK_EMPTY	Both sides of the book had no quotable level: the book was empty.	Book churn, depth bands, order ladders
15	32768	QUALITY_UNKNOWN	No quality information exists for this row (the storage default; the collector never emits it).	All families

Transcribed from the collector's quality-flags definition. The families map to the groups on the metric reference.

On a response, the quality object looks like this. Here the row is clean apart from one flag, and it names exactly which families to distrust:

The quality object on a /v1/latest row

```
"quality": {
  "raw": 2048,
  "all": 2048,
  "flags": ["DEPTH_BEYOND_COVERAGE"],
  "contaminates": ["bookWalls", "orderLadders"],
  "observedAt": 1786581011523
}
```

The repairs we have run, with measured results

Data quality is not a claim; it is a programme with measured before-and-after numbers. Two repairs carry a clean acceptance test against an independent reference:

Historical repairs with a measured acceptance test

Repair	Rows	Measured before	Measured after
Surgical wide-book price repair	37,527	46.91% median error	1.78% median error
GBP quote-rate repair (Kraken XBTGBP vs XBTUSD, 3,664 windows)	95,262	-11.93%	+0.005%

The price repair is a 26-times improvement at the median, measured over 27,069 comparable rows. The GBP figure is the executed Kraken acceptance test.

Alongside those, a set of labelling and backfill passes made the archive self-describing rather than changing values:

- A base-asset backfill filled every historical row: 45,296,738 rows, zero left unfilled on all nine venues.
- A coin-identity pass labelled 84,831 rows and repaired 12,518 where a symbol had resolved to more than one candidate coin.
- 18,893,904 rows where the price had been carried from the book mid or the last trade were flagged with the PRICE_FROM_MID bit, with zero false positives, so a derived price can never be mistaken for a traded one.
- 38,362 empty-book windows were labelled rather than left looking like real, zero-depth books.

Kept forever

The earliest retained window is 14 July 2026, and retention is infinite: nothing has aged out. A plan's history window is an entitlement over that archive, not a limit on how long the data is kept. This is the one thing an exchange API and a self-recorder cannot give you at any price, because their history starts the day you start recording. More on that in free versus paid, recording it yourself, and whether order book liquidity is real.

Isn't this just the exchange's free feed, repackaged?

No. BlackForge maintains a live order book per pair from each venue's own WebSocket diff stream, computes measurements from it every five minutes, and stores the result forever.

The free feed gives you the raw stream from now on; it does not give you the maintained book, the computed columns, or any history before you started recording.

How do I know a value is real and not a placeholder?

Every row carries a quality bitmask. If a bit is set, the column families it contaminates are placeholders for that window, not measurements. The bitmask, its named flags and the families each flag contaminates are returned on every response, so you can drop or grey out contaminated windows rather than trusting them blindly.

Are all nine venues measured at the same instant?

The trade columns are: they are bucketed by exchange event time. The order-book columns are read at a deterministic per-venue offset inside the window, so two venues' book columns for the same window label can be observed up to about 150 seconds apart. Each book reading carries its own observed-at timestamp so the offset is never hidden.

How far back does the history go?

The earliest retained window is 14 July 2026, and retention is infinite: nothing is aged out. Each plan can read back a fixed distance (the history window is an entitlement), but the archive itself is kept forever.

[API documentation](#) · [Every column, by family](#) · [Coverage by venue](#) · [This page as a PDF](#)