Datasent: A Technical Overview

Lossless Tokenisation, Trusted Transmission,
and Governed Reconstruction

Datasent Research `info@datasent.com` 2026

Abstract

Datasent is a lossless data encoding layer with two defining properties. First, it encodes any data — tabular, sensor, time-series, image, video, audio, text, embeddings, or graphs — as a compact structured token representation that admits exact reconstruction. Second, it implements a trusted setup protocol under which raw data never traverses organisational or network boundaries: only the unpredictable residual travels, and reconstruction is gated by explicit authorisation.

This paper provides a unified technical overview of the full system. We describe the lossless tokenisation primitive and its exact reconstruction guarantee, the trusted setup and residual-only transmission protocol, adaptive segmentation via joint MDL optimisation, multi-modal canonicalisation, governed reconstruction via custodian gating and threshold sharing, and the path to AI-native data workloads. We state the key mathematical results with proofs, characterise complexity, and compare Datasent to existing systems.

1. Introduction

Modern data infrastructure rests on a problematic assumption: that raw data must travel between systems to be used. Compression reduces the size of what travels but does not change the fundamental model — raw data still crosses network boundaries, accumulating cost, risk, and compliance exposure at every hop.

Datasent is designed around the opposite assumption. If sender and receiver agree on a deterministic representation upfront, most of the data can be regenerated locally. Only the unpredictable part — the residual — needs to move. Raw data stays where it originated.

This paper describes the full Datasent system. Section 2 covers the lossless tokenisation primitive. Section 3 describes the trusted setup and residual-only transmission protocol. Section 4 covers adaptive segmentation. Section 5 describes multi-modal support. Section 6 covers the custodian gating model. Section 7 characterises complexity. Section 8 compares to existing systems. Section 9 describes the AI readiness path.

2. The Lossless Tokenisation Primitive

2.1. Canonicalisation

All data enters Datasent through a deterministic canonicalisation step that produces one of three canonical types:

- `CanonicalMatrix`: integer matrix $Y \in \mathbb{Z}^{N \times D}$
- `CanonicalBlob`: byte sequence $\mathcal{B} \in \{0,1\}^*$
- `CanonicalBundle`: labelled collection of matrices and blobs (e.g. video: keyframes + motion + residual frames)

Canonicalisation enforces deterministic ordering, fixed-point quantisation for numeric inputs, and explicit metadata for scaling and layout. Two equivalent inputs — same records, different formatting or ordering — produce identical canonical representations. This determinism is a prerequisite for trusted reconstruction: the same canonical input must yield the same encoding decisions and the same decode behaviour across environments.

2.2. The Token Structure

For a canonical matrix $Y \in \mathbb{Z}^{N \times D}$ partitioned into segments $Y^{(1)}, \ldots, Y^{(S)}$, the encoding of segment

s is:

$$Y^{(s)} = \underbrace{\text{round}\big(B^{(s)}C^{(s)}\big)}_{P^{(s)}} + \underbrace{R^{(s)}}_{\text{exact residual}}$$

where $B^{(s)} \in \mathbb{R}^{L_s \times K_s}$ is a deterministic basis matrix, $C^{(s)} \in \mathbb{R}^{K_s \times D}$ is the coefficient matrix fitted on the segment, and $R^{(s)} = Y^{(s)} - P^{(s)}$ is the exact integer residual. Each segment becomes a **token**:

$$z_s = \big(\tau_s^\star,\ C^{(s)},\ R^{(s)},\ m_s\big)$$

where $\tau_s^\star$ identifies the model family and m_s is the metadata sufficient to regenerate $B^{(s)}$ deterministically. The full encoded dataset is $E(Y) = (z_1, \ldots, z_S)$.

2.3. Lossless Reconstruction

Theorem 1 (Exact reconstruction)**.** *If $B^{(s)}$ is deterministic and $R^{(s)}$ is stored exactly, then* $\text{round}(B^{(s)}C^{(s)}) + R^{(s)} = Y^{(s)}$ *for every segment s, and therefore $D(E(Y)) = Y$.*

Proof. By definition, $R^{(s)} = Y^{(s)} - P^{(s)}$. Therefore $P^{(s)} + R^{(s)} = P^{(s)} + (Y^{(s)} - P^{(s)}) = Y^{(s)}$. Concatenating segments gives $D(E(Y)) = Y$. □

> This is not a probabilistic or approximate result. Reconstruction is exact by construction, regardless of model quality. If the model captures nothing, $R^{(s)} = Y^{(s)}$ and the encoding is lossless at full raw size.

2.4. Compression Condition

Compression occurs when the tokenised representation is smaller than the raw data. Including the raw encoding as a baseline model τ_0 in $\mathcal{M}$ guarantees the encoding never increases description length: $L_{\text{tok}}(Y) \leq L_{\text{raw}}(Y)$ always holds.

Proposition 1 (Compression condition)**.** *Tokenisation achieves compression for segment s if and only if $H(R^{(s)}) < H(Y^{(s)})$, i.e. the residual entropy is lower than the raw entropy. This holds whenever the model is informative.*

3. Trusted Setup and Residual-Only Transmission

3.1. Definition

A **trusted setup** in Datasent means:

> All parties agree on deterministic representation parameters before any data-dependent computation occurs. This includes basis type and parameters, segmentation strategy, and quantisation and rounding rules.

Crucially, this is *not* a cryptographic ceremony. The setup parameters are public and require no secrecy. Their role is to ensure that both sender and receiver can independently regenerate the same basis $B^{(s)}$ from metadata m_s, without transmitting $B^{(s)}$ itself.

3.2. Roles and Parties

Three logical actors participate in the protocol:

1. **Data Holder (DH):** possesses raw data X and its canonical form Y. Performs local fitting and transmits residuals.
2. **Reconstruction Party (RP):** does not possess X. Receives residuals and reconstructs Y under authorisation.
3. **Custodian (C):** optional gatekeeper that holds shares of coefficient data and authorises reconstruction.

3.3. The Golden Path

All data in Datasent flows through the following invariant sequence:

Raw input → Canonicalise → Fit representation → Segment → Tokenise → Transmit residuals/shares → Authorised reconstruction

Raw data Y is never transmitted. What crosses the network is only:

- the residuals $R^{(s)}$, or
- secret shares of coefficients $C^{(s)}$ and/or $R^{(s)}$.

3.4. Bandwidth Analysis

Let $|Y|$ be the raw data size, $|C|$ the total coefficient size, and $|R|$ the total residual size. For structured data:

$$|C| \ll |Y|, \qquad |R| \ll |Y|.$$

The transmitted payload is $|R|$ or $|C| + |R|$, often orders of magnitude smaller than $|Y|$. Importantly, this saving is **architectural**: even a payload of size $|C|+|R| = |Y|$ would represent an improvement, because raw data stays local and only structured, governed artifacts cross boundaries.

4. Adaptive Segmentation

4.1. The Joint Optimisation Problem

Real datasets are structurally heterogeneous. Datasent jointly optimises segment boundaries and model family assignments to minimise total description length:

$$\min_{S,\,\{\tau_j\}} \sum_{j=1}^{S} \left[L(\tau_j) + L(C_j^{(\tau_j)}) + L(R_j^{(\tau_j)}) + L(m_j)\right] + \lambda S$$

where λ is a segment penalty preventing over-segmentation and each $L(\cdot)$ denotes description length in bits. This objective is derived from the Minimum Description Length (MDL) principle [1, 2].

4.2. Dynamic Programming Solution

The joint problem exhibits optimal substructure, enabling an exact polynomial-time solution under a maximum segment length constraint W:

$$\mathcal{V}(t) = \min_{1 \le s \le \min(t,W)} \left[\mathcal{V}(t-s) + \mathcal{C}(t-s,\, t) + \lambda\right], \quad \mathcal{V}(0) = 0$$

Theorem 2 (Optimal substructure)**.** *The globally optimal segmentation of $Y_{1:t}$ can be computed from the globally optimal segmentation of $Y_{1:s}$ for all $s < t$. The DP recurrence above solves the joint problem exactly in $O(NW|\mathcal{M}|K_{\max}^2 D)$ time with memory $O(WK_{\max}D|\mathcal{M}|)$.*

4.3. Supported Model Families

The admissible set $\mathcal{M}$ includes polynomial (Vandermonde) bases, Fourier and cosine bases, wavelet bases, data-driven SVD/PCA bases, autoregressive models, and a raw identity baseline. The MDL objective selects automatically — no configuration required.

5. Multi-Modal Canonicalisation

Every data modality is handled through a unified composition principle:

> **Composition principle:** every modality decomposes into integer matrices. Each matrix is tokenised independently using the lossless primitive. The full encoded representation is the structured composition of resulting token streams.

Modality	Canonical form
Tabular / time-series	$Y \in \mathbb{Z}^{N \times D}$
Sensor / scientific	$Y \in \mathbb{Z}^{T \times D}$
Audio (PCM)	$Y \in \mathbb{Z}^{T \times C}$
Images	Per-channel pixel matrix
Video	Keyframes + motion + residual
Text	Token ID / byte sequence
Embeddings	Fixed-point vector matrix
Graphs	Node features + edges + adjacency

Corollary 1 (Lossless composition)**.** *If each component token stream satisfies Theorem 1 independently, the composed representation is lossless for the full modality. The trusted setup and raw-data-local guarantee apply uniformly across all modalities.*

6. Governed Reconstruction

6.1. Threshold Sharing

In the 2-of-2 XOR sharing scheme, the coefficient blob c for a segment is split into shares:

$$c_1 = u, \qquad c_2 = c \oplus u,$$

where u is a uniformly random bitstring. Reconstruction requires both shares:

Theorem 3 (2-of-2 correctness)**.** *Given shares c_1, c_2, reconstruction by XOR recovers the original coefficient blob exactly: $c_1 \oplus c_2 = u \oplus (c \oplus u) = c$.*

Theorem 4 (2-of-2 secrecy)**.** *Given only c_1 (or only c_2), the distribution of c is uniform over bitstrings of that length. A single share reveals no information about the coefficients (information-theoretic secrecy).*

The k-of-n generalisation uses Shamir secret sharing [3] over a finite field: any k shares reconstruct; any $k-1$ shares reveal no information.

6.2. Custodian Gating

In a standard deployment, the Data Holder transmits residuals $R^{(s)}$ to the Reconstruction Party. The Custodian holds share c_2 of the coefficient data. Reconstruction requires the Custodian to release c_2, which it does only upon explicit authorisation. Every reconstruction event is logged.

This architecture reframes data sharing as a governed token exchange:

- Neither party possesses the full picture alone.
- Raw data never leaves the Data Holder.
- Reconstruction is auditable and revocable.

6.3. Integrity Verification

Because representation is deterministic, reconstructed data can be re-canonicalised and its hash compared to a stored digest:

$$h^{(s)} = H(P^{(s)}),$$

where H is a collision-resistant hash (e.g. SHA-256). Tampering with residuals or coefficients is detectable. This provides a verifiable chain of custody from raw collection to final output, without requiring zero-knowledge proofs.

> **Future work:** ZK proofs can be layered on top of this system to prove that residuals correspond to a valid fitting, that coefficients satisfy error bounds, and that reconstruction occurred under authorisation — without revealing the underlying data. The architecture is designed to support this extension.

7. Computational Complexity

Operation	Complexity
Encoding (total)	$O(N\lvert\mathcal{M}\rvert K_{\max}(K_{\max}+D))$
Decoding (total)	$O(NK_{\max}D)$
Segmentation (bounded W)	$O(NW\lvert\mathcal{M}\rvert K_{\max}^2 D)$
Streaming (per row)	$O(K_{\max}^2 D\lvert\mathcal{M}\rvert)$
Parallel (P processors)	$O(N\lvert\mathcal{M}\rvert K_{\max}^2/P)$

All operations scale linearly in N for fixed model complexity. Segments are fully independent — encoding, decoding, and residual computation parallelise without coordination overhead. In streaming mode, memory is bounded at $O(WK_{\max}D\lvert\mathcal{M}\rvert)$ regardless of N, making the system suitable for constrained edge hardware.

8. Comparison to Existing Systems

System	L	RL	GR	MM
Classical compression	✓	×	×	×
Lossy transform coding	×	×	×	×
Columnar storage	✓	×	×	×
Federated learning	×	✓	×	×
Secure data clean rooms	×	✓	✓	×
Datasent	✓	✓	✓	✓

L=Lossless, RL=Raw data stays local, GR=Governed reconstruction, M

vs. classical compression (Huffman [4], LZ [5]): these find statistical redundancy in raw byte streams and produce opaque compressed bitstreams. Raw data is still transmitted. Structured

information in the data is not exploited. Entropy coders can be applied on top of Datasent residuals as a complementary lower layer.

vs. columnar storage (Parquet, ORC): delta coding and RLE are special cases of degree-1 and degree-0 polynomial models. Datasent selects adaptively among these and more expressive families. Columnar formats do not implement trusted transmission or governed reconstruction.

vs. federated learning: FL frameworks keep raw training data local but are lossy (gradient aggregation discards information) and application-specific. Datasent provides lossless residual exchange across any data type.

vs. secure data clean rooms: clean rooms provide governed access to shared data but typically require raw data to enter a controlled environment. Datasent eliminates that requirement — raw data never leaves the Data Holder in any configuration.

9. AI Readiness

9.1. Tokens as ML Features

Each token $z_s = (\tau_s^\star, C^{(s)}, R^{(s)}, m_s)$ maps to a fixed-dimensional feature vector:

$$\phi(z_s) = \left[\mathrm{vec}(C^{(s)});\ \mathrm{vec}(R^{(s)});\ \tau_s^\star;\ L_s;\ K_s\right].$$

The coefficient matrix $C^{(s)}$ captures local trend, curvature, and spectral components — the features preprocessing pipelines typically compute from raw data. $R^{(s)}$ captures local deviations. $\tau_s^\star$ provides an unsupervised structural label per segment. These features are available at zero additional computational cost beyond encoding.

9.2. Sequence Length Reduction

The token sequence $E(Y) = (z_1, \ldots, z_S)$ has length $S = N/L_{\text{avg}}$, where L_{avg} is the average segment length. For a transformer processing time-series data with $N = 10^6$ and $L_{\text{avg}} = 10^3$, sequence length reduces from 10^6 to 10^3 — a three order-of-magnitude reduction in attention complexity from $O(N^2)$ to $O(S^2)$, with no loss of information.

9.3. Federated Training

The trusted setup architecture makes federated ML across organisational boundaries structurally simpler. Organisations exchange tokenised residuals rather than raw datasets. Raw training data never leaves each organisation's environment. The custodian model provides explicit authorisation and audit logging for every reconstruction.

9.4. Token-Native Compute

The operator compatibility framework established in the full mathematical white paper [6] provides a formal foundation for running ML inference directly on Datasent-encoded data — without a decode or preprocessing step. This direction is discussed in detail in Section 9 of that paper.

10. Conclusion

Datasent establishes a new foundation for data infrastructure built around two mathematical guarantees: exact lossless reconstruction from a structured token representation, and a trusted setup protocol under which raw data never traverses organisational or network boundaries.

The combination enables capabilities that no existing system achieves simultaneously: lossless compression, raw-data-local transmission, governed reconstruction with audit logging, and unified multi-modal support. The framework is extensible to cryptographic verification and designed to support AI-native compute workloads.

Full technical treatment. Complete proofs, complexity analyses, the token algebra, and ZKP integration details are in the full mathematical white paper [6]. The trusted setup protocol and threshold sharing correctness proofs are in the companion paper [7].

References

[1] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[2] M. H. Hansen and B. Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.

[3] A. Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

[4] D. A. Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952.

[5] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.

[6] Datasent Research. A mathematical framework for structured information encoding. Technical report, Datasent, 2026.

[7] Datasent Research. A trusted setup for bandwidth-minimal, lossless data tokenization. Technical report, Datasent, 2025.

Datasent Research `info@datasent.com`
`datasent.com`