signal65

Compute-Ready Data with Dell AI Infrastructure

Data Preparation for Enterprise AI

Raw Data RAG Challenges

The Problem

Traditional document pipelines preserve format but not meaning

Documents treated as undifferentiated text—human-readable but computationally opaque

Fixed-length or "blind" chunking breaks contextual boundaries and internal logic

Why Raw Documents Fail

- Inconsistent Formats
- Missing Context
- Access Complexity
- Poor Chunking Size Determination

Hidden Costs

- Poor AI Responses
- Hallucinations
- High Latency
- Security Exposures
- Poor Chain of Custody Management

Probabilistic Risk vs.Deterministic Reliability

Traditional Methods

User asks question.

LLM "rolls the dice".

New answer generated.

New Methods

User asks question.

System Retrieves "Golden Record" answer from Semantic Twin (deterministic).

"Intelligence" Location: *In the Metadata (at ingestion) — not in the LLM (at query time)*

The Problem (Traditional RAG)

Every user query forces the LLM to re-read and re-interpret raw text in real-time.

Result:

The same question can yield different answers (inconsistency) and higher energy costs.

The Solution (TopicLake Semantic Twin)

One-Time Reasoning

Analytical reasoning, sentiment analysis, and logic extraction are performed once during ingestion.

Synthetic Q&A

The system generates and validates answers to likely questions before a user ever asks them.

Governed Retrieval

When a user queries, the system returns this pre-computed, governed answer rather than generating a new one from scratch.

Transforming Documents into Semantic Twins

Decomposition

Split by topic, not word count.

Instead of fixed-length "blind" chunking (e.g., splitting every 500 words), the document is broken into its composite pieces by topic boundaries. These compute-ready topics enable LLMs or agents to query specific topics directly, without having to read surrounding noise.

Enrichment

AI-generated summaries, sentiment, and Q&A pairs.

The ingestion engine uses LLMs to generate synthetic intelligence for each topic:

- **Summaries & Descriptions:** High-level overviews
- **Topic-Specific Keywords:** Arrays of keywords
- **Synthetic Q&A:** Curated questions and validated answers
- **Sentiment & Intent:** Topics tagged with emotional tone

Entity ID

Identifying People, Places, and Events.

The ingestion process extracts and identifies people, places, and events, creating a structured index of entities. Instead of searching for the word "London," the system treats "London" as a place entity, enabling more complex queries (e.g., "show me all events in London with negative sentiment").

Governance

Inherited security clearance and citation lineage.

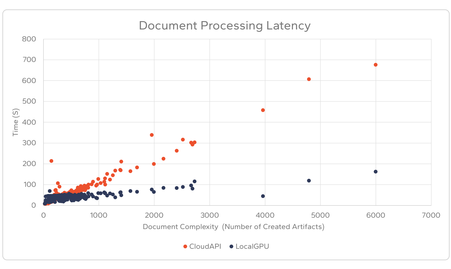
Topics inherit the unique ID, citation lineage, and security designation of parent documents. All enriched topics, summaries, and Q&A pairs from a document designated internal or secret carry the same designation. These inherited permissions dramatically simplify protection against data leakage.

Purpose-Built for Performance

- Dell PowerSwitch Z9864F-ON
- SONiC 4.4.0

Two Server Cluster, each with:

- 2 Dell PowerEdge XE7745
- Dual AMD EPYC 9555
- 2.3 TB DDR5 Memory
- 8 NVIDIA RTX Pro 6000 Blackwell Server
- 8 Broadcom BCM57608 400GbE RoCEv2 Network Controller
- Ubuntu 24.04 LTS
- CUDA 13.0 / Driver 580.95.05
- vLLM v0.10.2



Performance Advantages

- **Low and Predictable Ingestion Latency:** Eliminates network round-trips and queueing delays
- **Parallelized Reasoning:** Processes multiple semantic tasks concurrently
- **Deterministic Scaling:** Dedicated environment scales predictably under load
- **Low-Jitter Data Movement:** 400 GbE RoCEv2 fabric ensures consistent throughput

Why Shift to On-Prem?

Data Sovereignty

Complete control over document processing, model selection, and access governance.

Zero-Trust alignment

Higher Token Throughput

On-premises compute-ready document generation outperforms public cloud alternatives.

Predictable, low-latency processing

Energy Efficiency

Lower power consumption per token and reusable source data across applications.

Sustainable AI infrastructure

See how Dell can help create secure compute-ready data for your organization.

- Purpose-built infrastructure delivers superior token generation rates
- Lower energy consumption per document processed
- Accelerated ROI while reducing operational risk
- Complete chain of custody from source documents through AI consumption

Compute-ready data will define enterprise AI readiness. The competitive advantage belongs to organizations that recognize this fundamental truth: AI systems are only as valuable as the data they consume.

TESTING BY:

signal65

IN PARTNERSHIP WITH:

DELL Technologies

GADGET SOFTWARE