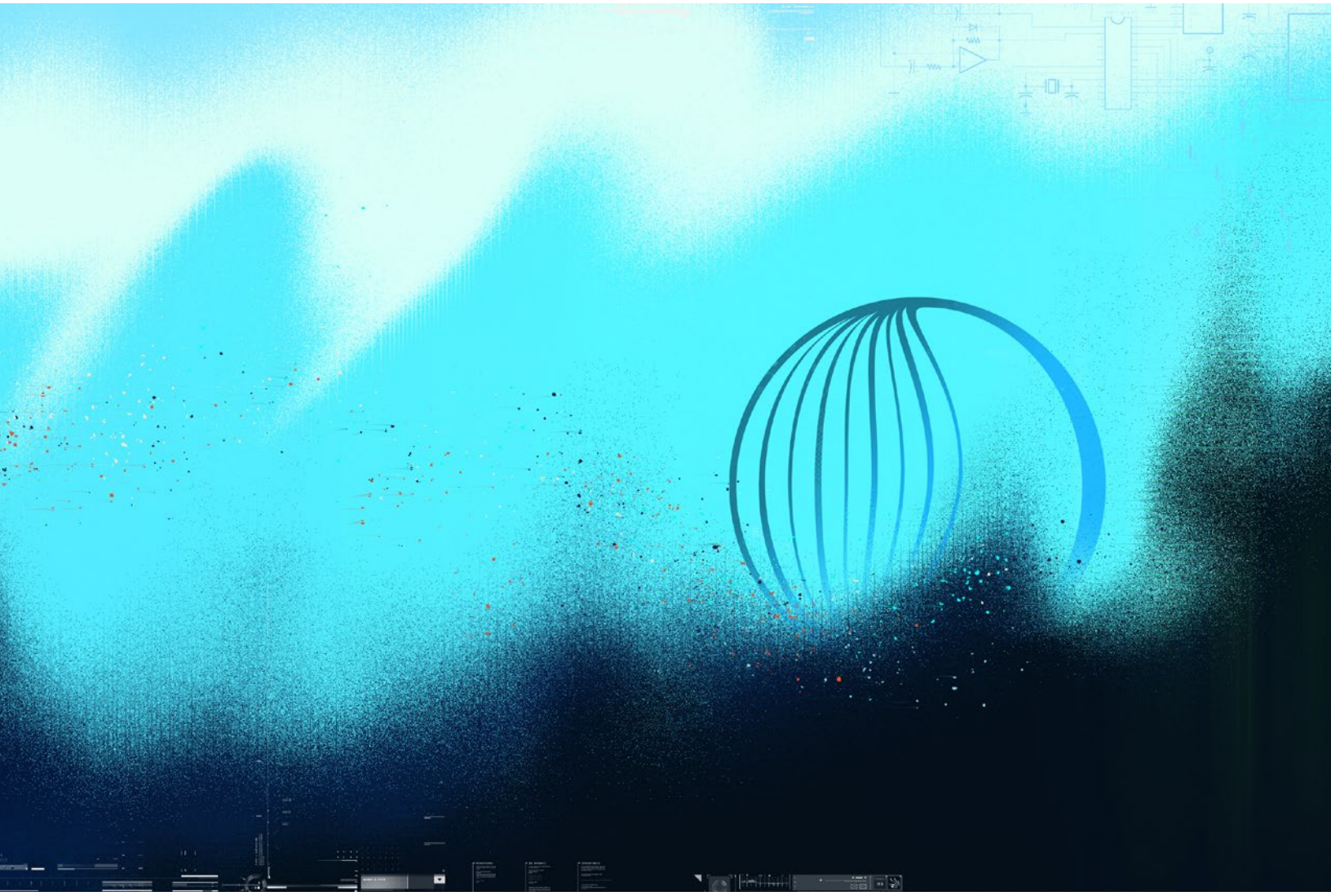
signal65



Data Preparation for Enterprise AI

Compute-Ready Data with TopicLake Insights and Dell AI Infrastructure

AUTHORS

Matthew Goldensohn
Performance | Signal65

Brian Martin
AI Data Center Performance | Signal65

JANUARY 2026

IN PARTNERSHIP WITH

DELL Technologies

GADGET SOFTWARE

Executive Summary

As enterprises accelerate AI adoption, the quality, structure, and accessibility of their data have become the primary determinant of success. While organizations invest heavily in large language models and inference infrastructure, many overlook a critical prerequisite: transforming unstructured enterprise data into compute-ready formats that AI systems can consume efficiently and reliably. Without this foundational step, even the most sophisticated AI deployments struggle with inconsistent outputs, security vulnerabilities, and unpredictable costs.

On-premises data preparation addresses these challenges by bringing the data transformation pipeline, from raw document ingestion to structured, AI-ready outputs, under direct organizational control. This approach ensures data sovereignty, enables consistent governance, and delivers measurable performance advantages over cloud-dependent alternatives. By processing documents locally using purpose-built infrastructure, enterprises can generate compute-ready data at higher throughput and lower energy consumption while maintaining complete visibility into provenance and access patterns.

Key Highlights

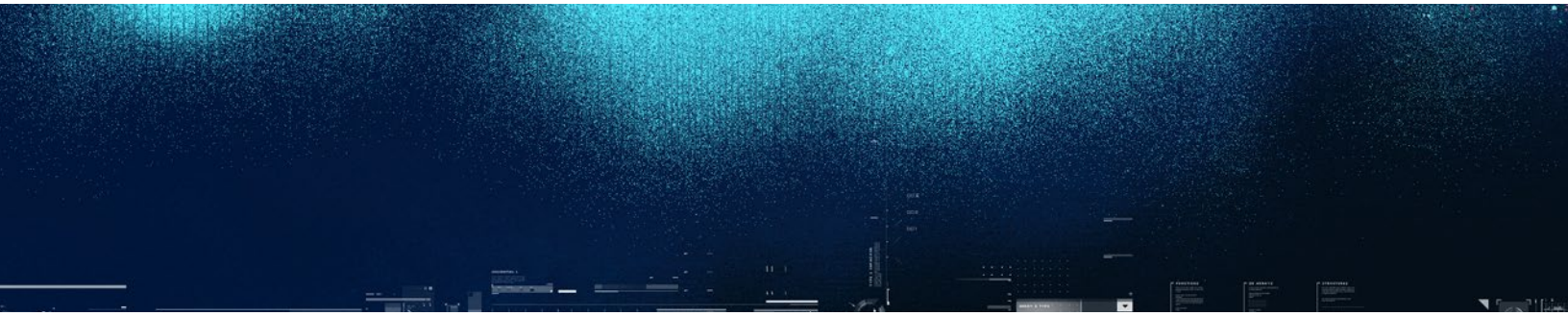
Data Sovereignty
Complete control over document processing, model selection, and access governance

Higher Token Throughput
On-premises compute-ready document generation on Dell platform outperforms public cloud alternatives

Energy Efficiency
Lower power consumption per token and reusable source data across applications

TopicLake Insights represents a new paradigm for enterprise data management: topic-oriented data lakes combining generative AI capabilities with structured, governed storage to normalize unstructured content into a unified, queryable format. Built on Dell PowerEdge servers with Broadcom high-performance NICs and Dell PowerSwitch switches, this architecture transforms documents, reports, and media into enriched digital artifacts, semantic twins that serve AI applications, business intelligence platforms, and operational systems simultaneously.

This paper examines the architectural patterns, performance characteristics, and business benefits of on-premises compute-ready document generation, demonstrating how organizations can establish data foundations that scale with evolving AI requirements while preserving security, governance, and operational efficiency.



A Data Preparation Imperative

The Gap Between AI Capability and Enterprise Data Readiness

Any organization seeking to use proprietary documents to generate accurate and compliant AI-driven responses must first augment that data to ensure controlled access, company-specific context, and possibly even regulatory compliance.

Why Raw or Unprocessed Documents Fail to Provide Accurate and Citable AI Responses

Poor generative AI responses based on raw data or documents are due to inconsistent formats, missing context, access complexity and chunking sizes determination when RAG pipelines are built.

The hidden costs of working with unprocessed documents or data are poor AI responses, hallucinations, latency, security exposures and poor chain of custody management.

Compute-Ready Documents for Enterprise AI

Generative AI responses based on raw or unprocessed documents often suffer due to inconsistent formats, fragmented context, and arbitrary chunk boundaries introduced during retrieval-augmented generation (RAG) pipeline construction. These limitations force models to reconstruct meaning at query time, increasing the likelihood of incomplete answers, broken citations, and inconsistent interpretations of the source material.

The hidden costs of relying on unprocessed data extend beyond response quality. Organizations experience higher latency, increased hallucination rates, expanded security exposure, and weakened chain-of-custody controls as AI systems repeatedly reprocess raw content. These challenges are compounded by the probabilistic nature of generative AI itself: even when prompted with the same question, models naturally produce different responses each time. Without a structured, governed data foundation, this inherent variability makes it difficult to deliver reliable, auditable, and repeatable AI outputs suitable for enterprise and regulatory use.

The Anatomy of a Compute-Ready Document

A compute-ready document comprises four distinct layers constructed through agentic ingestion.

1. Semantic Decomposition (the composite pieces)

Instead of fixed-length or "blind" chunking (for example, splitting every 500 words), the document is broken into its composite pieces by topic boundaries. This enables LLMs or agents to query specific topics directly, without having to read surrounding noise to determine where one idea ends and another begins.

2. Synthetic Enrichment

The ingestion engine uses one or more appropriate LLMs to generate synthetic intelligence for each topic:

- **Summaries & Descriptions:** High-level overviews for quick scanning
- **Topic Specific Keywords:** An array of keywords derived from the topic, allowing for topic relations to be identified

- **Synthetic Q&A:** A curated set of likely questions and validated answers, enabling question-to-question matching, significantly more accurate than question-to-text matching
- **Sentiment & Intent:** Topics are tagged with emotional tone (e.g., frustrated customer, formal policy, technical note).

3. Entity Identification

The ingestion process extracts and classifies identities such as people, places, and events from incoming documents, creating a structured index of entities for advanced querying. 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").

4. Governance and Security

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

Feature	Traditional RAG	Graph RAG	Agentic Compute-Ready
Ingestion Complexity	Low (Fast/Cheap)	High (Extracts Triplets)	High (Agentic Processing)
Unit of Storage	Raw Text Chunks	Knowledge Nodes/ Edges	Enriched Semantic Topics
"Intelligence" Location	In the LLM (at query time)	In the Graph Structure	In the Metadata (at ingestion)
Search Method	Semantic Vector Search	Traversal and Vector	Hybrid (Keywords and Metadata and Vector)
Auditability	Poor (Hard to cite source)	Moderate (Nodes link to source)	Excellent (Mandatory Citations and immutable IDs)

Table 1: *Document Pipeline Comparisons*

Concepts and Architecture for Compute-Ready Documents

Transforming unstructured enterprise data into a form that AI systems can consume effectively requires a fundamental shift beyond traditional document management approaches. Conventional systems are designed to store and retrieve content for human use, not to support computational reasoning at scale. Establishing a foundation of compute-ready data, therefore demands an architectural framework that treats documents as structured, semantically rich assets rather than passive files.

Enterprises manage a broad spectrum of document types, ranging from financial filings and technical specifications to multimedia reports and scanned records. Traditional document pipelines struggle in AI-driven environments because they preserve format but not meaning. Documents are often treated as undifferentiated text or images, making them human-readable but computationally opaque. Common practices such as fixed-length or "blind" chunking further degrade document integrity by breaking contextual boundaries and internal logic. The result is operational friction: AI systems produce hallucinations, experience higher latency, and introduce security and governance risks. To become compute-ready, documents must be transformed into structured, semantically dense, and governed data objects that preserve meaning, context, and provenance.

The TopicLake Insights architecture introduces a new paradigm for addressing this challenge through a topic-oriented data lake that normalizes unstructured content into a unified, queryable format. Within this model, source documents are transformed into enriched digital artifacts, or semantic twins, that retain the original content while embedding extensive AI-generated metadata. The repository employs Write-Once-Read-Many (WORM)–style storage patterns to support auditability, long-term retention, and cost efficiency. By exposing this enriched data through high-performance APIs, TopicLake functions as a shared utility layer, simultaneously supporting AI applications, business intelligence platforms, and operational systems without duplicating ingestion or processing effort.

To ensure documents are fully understood before storage, the ingestion pipeline applies a multi-layered processing approach that combines structured parsing with vision-enhanced analysis. For documents with predictable layouts, high-performance structured extraction techniques preserve hierarchy and semantic relationships through precise XML parsing. These workflows are optimized for Dell PowerEdge infrastructure, leveraging the high memory capacity and compute density of the XE7745 platform to sustain throughput at scale. When documents include complex visual elements such as charts, tables, or handwritten annotations, the architecture leverages advanced vision models to interpret visual data and convert it into machine-readable artifacts. These artifacts are further enriched by large language model–based agents that generate descriptive metadata, summaries, and synthetic question-and-answer pairs to support accurate retrieval.

Together, this architectural framework ensures that the analytical reasoning required to understand a document's context, structure, and intent is performed once during ingestion rather than repeatedly at query time. By pre-computing semantic understanding and governance controls, the system delivers faster, more accurate, and more secure AI responses while establishing a durable foundation for enterprise-scale AI workloads.

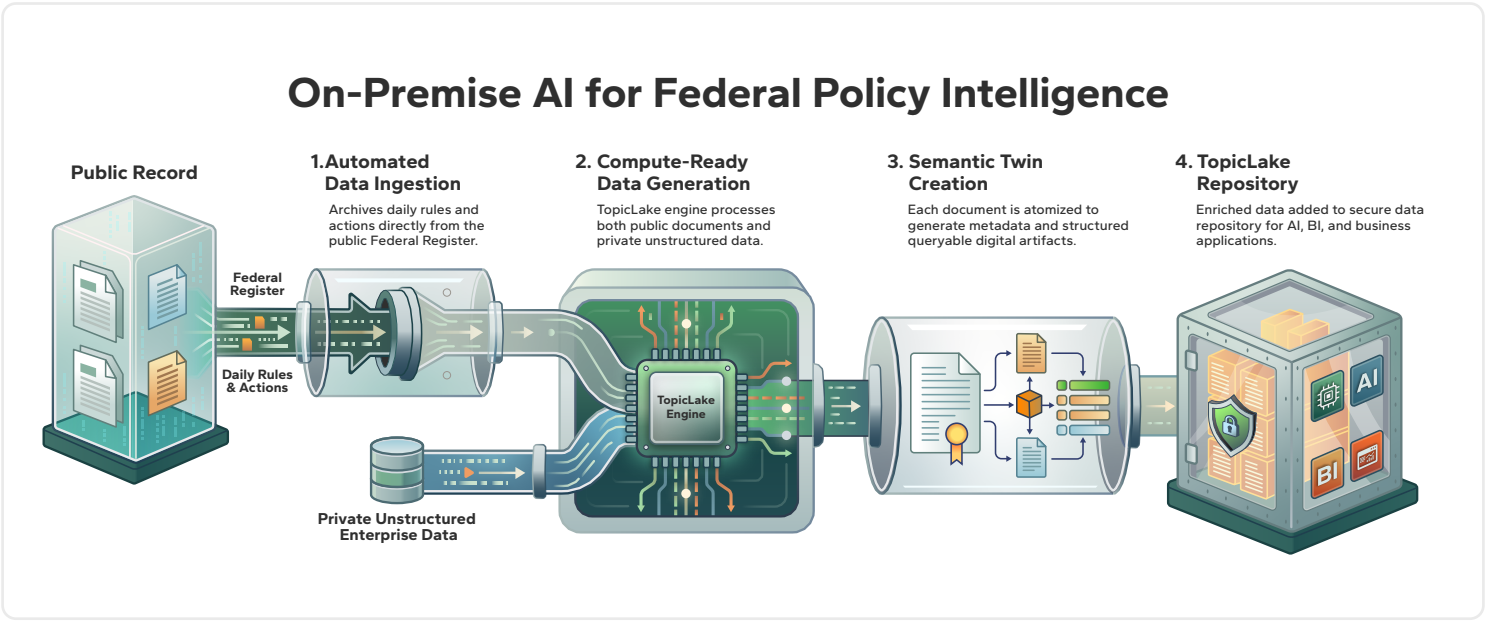


Figure 1: *On-Premise AI for Federal Policy Intelligence*

The Case for On-Premises Data Preparation

As enterprises accelerate their AI adoption, the location where data is transformed becomes a strategic decision. While cloud solutions are common, moving the data transformation pipeline on-premises—from raw document ingestion to AI-ready outputs—provides organizations with direct control over their most sensitive assets

This shift addresses the critical gap between raw AI capability and true enterprise readiness. Below is a detailed breakdown of the advantages in sovereignty, performance, and efficiency of a localized data preparation strategy.

Data Sovereignty and Security

The primary driver of on-premises data preparation is protecting proprietary intellectual property. Processing sensitive enterprise content in shared cloud environments introduces increased risk related to data exposure, third-party access, and jurisdictional ambiguity. Keeping data transformation pipelines within organizational boundaries ensures enterprises retain direct control over how their most valuable information is processed, stored, and accessed, reducing both security risk and regulatory uncertainty.

Local data preparation also enables organizations to meet stringent regulatory and compliance requirements across industries such as financial services, healthcare, and government. By ensuring sensitive data remains within controlled environments, enterprises can more easily satisfy obligations under frameworks such as GDPR, HIPAA, and sector-specific regulations. At the same time, on-premises pipelines allow organizations to

maintain full governance over model selection, versioning, and ingestion workflows, ensuring alignment with internal standards, risk policies, and audit requirements.

Finally, on-premises infrastructure establishes a documented, end-to-end chain of custody from the moment a source document is ingested through its consumption by downstream AI applications. This architecture supports seamless integration with existing enterprise security frameworks and access-aware retrieval systems, enforcing permissions directly at the data layer. As a result, organizations can prevent permission leakage, maintain auditable provenance, and deliver AI-driven insights without compromising security or trust.

Performance and Throughput

On-premises data preparation on purpose-built infrastructure delivers measurable performance advantages over cloud-dependent alternatives. By processing documents locally, organizations eliminate the latency variability, bandwidth constraints, and transfer overhead inherent in cloud-based pipelines. This results in more predictable performance, faster time-to-insight, and greater operational control, particularly for large, document-heavy workloads.

Purpose-built platforms such as the Dell PowerEdge XE7745 provide the computational density needed to achieve significantly higher token-generation rates during document transformation. High-memory, multi-GPU configurations enable massive parallelism during ingestion, allowing AI models to perform summarization, entity extraction, and synthetic enrichment at scale. This throughput advantage is especially important during the analytical reasoning phase of ingestion, where performance directly impacts pipeline efficiency and cost.

Equally critical is deterministic network performance. By leveraging Broadcom 400 GbE network controllers and a RoCEv2 fabric, on-premises systems deliver low-jitter, high-bandwidth data movement across large document corpora. Combined with pipelines optimized for batch processing, this architecture enables enterprises to normalize diverse content types into a unified, compute-ready format at higher sustained speeds without the unpredictability of shared cloud resources.

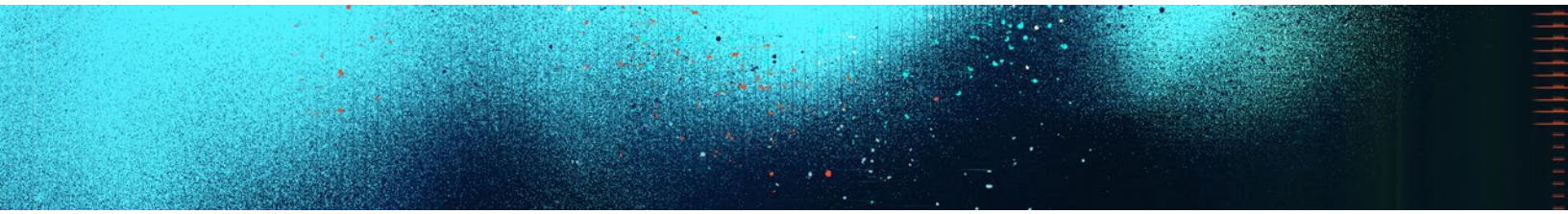
Energy Efficiency and Total Cost of Ownership (TCO)

Beyond security and speed, on-premises infrastructure provides a more sustainable and cost-predictable model for large-scale AI deployments.

Benefit	On-Premises Advantage
Energy Consumption	Lower power consumption per token compared to cloud-equivalent workloads.
Operational Costs	Avoids the escalating and often unpredictable costs associated with ongoing cloud processing and data egress.
Sustainability	Purpose-built infrastructure maximizes efficiency, supporting long-term corporate sustainability objectives.
ROI Acceleration	High-performance local processing leads to a faster return on investment by reducing the time-to-insight for document-heavy workflows.

Table 2: *On-Premise Infrastructure*

Ultimately, the choice of data preparation architecture determines long-term success more than any single AI model. By establishing a foundation on Dell PowerEdge servers and Broadcom networking, organizations transform their unstructured data into a strategic utility while maintaining the highest standards of security and efficiency.



Evaluation

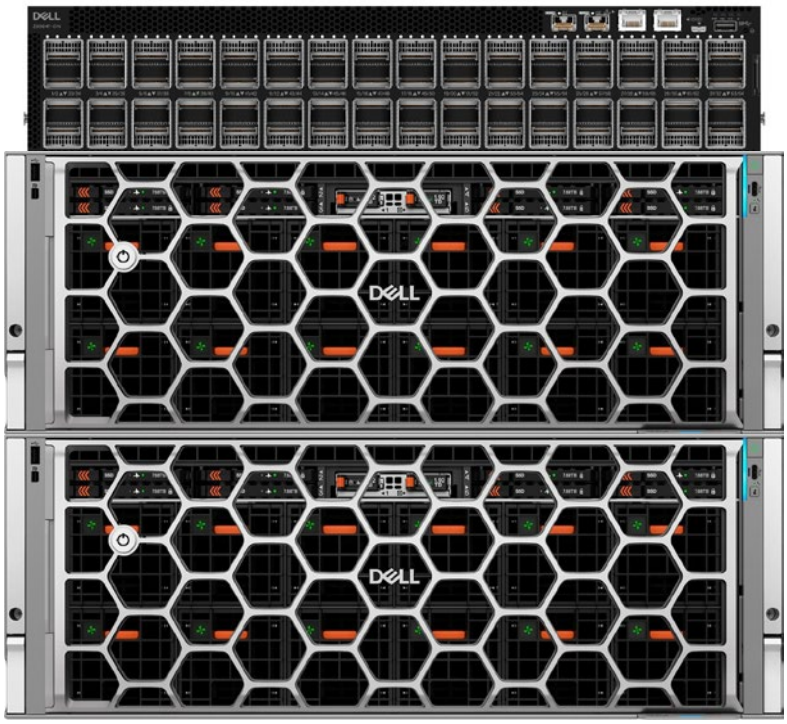
Use Case: Processing the U.S. Federal Register

The U.S. Federal Register represents one of the largest and most complex sources of regulatory information, comprising daily notices, proposed rules, final rules, and executive actions published across multiple agencies. These documents are lengthy, structured, and rich in legal and procedural context, making them difficult to use effectively with traditional retrieval-augmented generation (RAG) pipelines. When treated as raw text or blindly chunked, Federal Register content often loses critical context, citation integrity, and cross-reference relationships, limiting the accuracy and defensibility of AI-generated regulatory analysis.

Using the TopicLake architecture, Federal Register documents can be ingested and transformed into compute-ready documents that preserve regulatory structure while embedding semantic intelligence. Each notice or rule is decomposed along logical and topical boundaries, such as agency authority, statutory references, compliance requirements, and effective dates, and enriched with AI-generated summaries, entity extraction, and synthetic question-and-answer pairs. Security classifications, citation lineage, and provenance metadata are inherited at the topic level, enabling access-aware retrieval and auditable analysis. Once normalized, this compute-ready regulatory corpus can be queried by AI agents, compliance teams, and business intelligence systems to support impact analysis, rule comparison, policy monitoring, and regulatory change management—delivering faster, more reliable insights without sacrificing governance or traceability.

Testing Configuration

The Dell PowerEdge XE7745 delivers enterprise-grade AI infrastructure, providing a high-memory, multi-GPU platform that makes Agentic AI practical at scale. Equipped with dual AMD EPYC processors, up to 3 TB of DDR5 memory, support for eight NVIDIA L40S, H100, H200, or RTX Pro 6000 GPUs, and eight Broadcom BCM57608 400 GbE network controllers, each node delivers exceptional bandwidth and parallelism for large-context inference and high-volume retrieval. Clustered with Dell PowerSwitch Z9864F-ON switches in a RoCEv2 fabric, these solutions achieve deterministic, low-jitter throughput across workloads. This balance of memory capacity, compute density, and network efficiency enables businesses to deploy Agentic AI systems for the most demanding reasoning workloads.



- 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

Results Analysis

Testing conducted on the Dell PowerEdge XE7745 infrastructure demonstrates that on-premises data preparation provides a superior foundation for enterprise AI compared to generalized cloud services. By utilizing a high-memory, multi-GPU platform designed for agentic ingestion, organizations can achieve deterministic performance while optimizing operational costs.

Throughput and Latency

Document ingestion latency is a critical determinant of end-to-end pipeline efficiency, particularly as document complexity increases. As shown in Figure X, the on-premises XE7745-based pipeline consistently delivers lower and more predictable processing latency than cloud-based API workflows across a wide range of document complexities.

For cloud-dependent pipelines, latency increases sharply as documents grow in size and structural complexity. This behavior reflects a combination of network transit overhead, contention for shared infrastructure, and serialized execution during the analytical “thinking” phase of ingestion. In contrast, the on-premises pipeline exhibits a flatter latency curve, maintaining bounded response times even for highly complex documents.

Several architectural factors contribute to this latency advantage:

- **Low and Predictable Ingestion Latency:** By locating high-density GPU compute with the data source, the on-premises pipeline eliminates network round-trips and queueing delays inherent in cloud APIs. This results in consistently lower per-document processing times, particularly for complex documents that require extensive summarization, entity extraction, and synthetic Q&A generation..
- **Parallelized Reasoning at Ingestion Time:** The Dell PowerEdge XE7745 configuration enables massive parallelism during the reasoning-intensive stages of ingestion. Rather than serializing analysis across documents, the system processes multiple semantic tasks concurrently, preventing latency from scaling linearly with document complexity.
- **Deterministic Scaling Under Load:** Unlike cloud platforms, where latency variance increases under shared demand, the dedicated on-premises environment scales predictably. vLLM v0.10.2 efficiently manages large context windows and GPU memory, allowing high-complexity documents to be processed without disproportionate latency spikes.
- **Low-Jitter Data Movement:** The Broadcom BCM57608 400 GbE controllers and Z9864F-ON switches provide a low-latency, low-jitter RoCEv2 fabric, ensuring that multimodal data flows between processing stages without introducing additional delay.

Together, these characteristics explain the divergence observed in *Figure 2*, while cloud-based pipelines exhibit rapidly increasing, highly variable latency as document complexity increases, the on-premises architecture maintains stable processing times. This latency stability directly translates into higher sustained throughput, faster time-to-insight, and more predictable operational behavior for enterprise-scale document ingestion workloads.

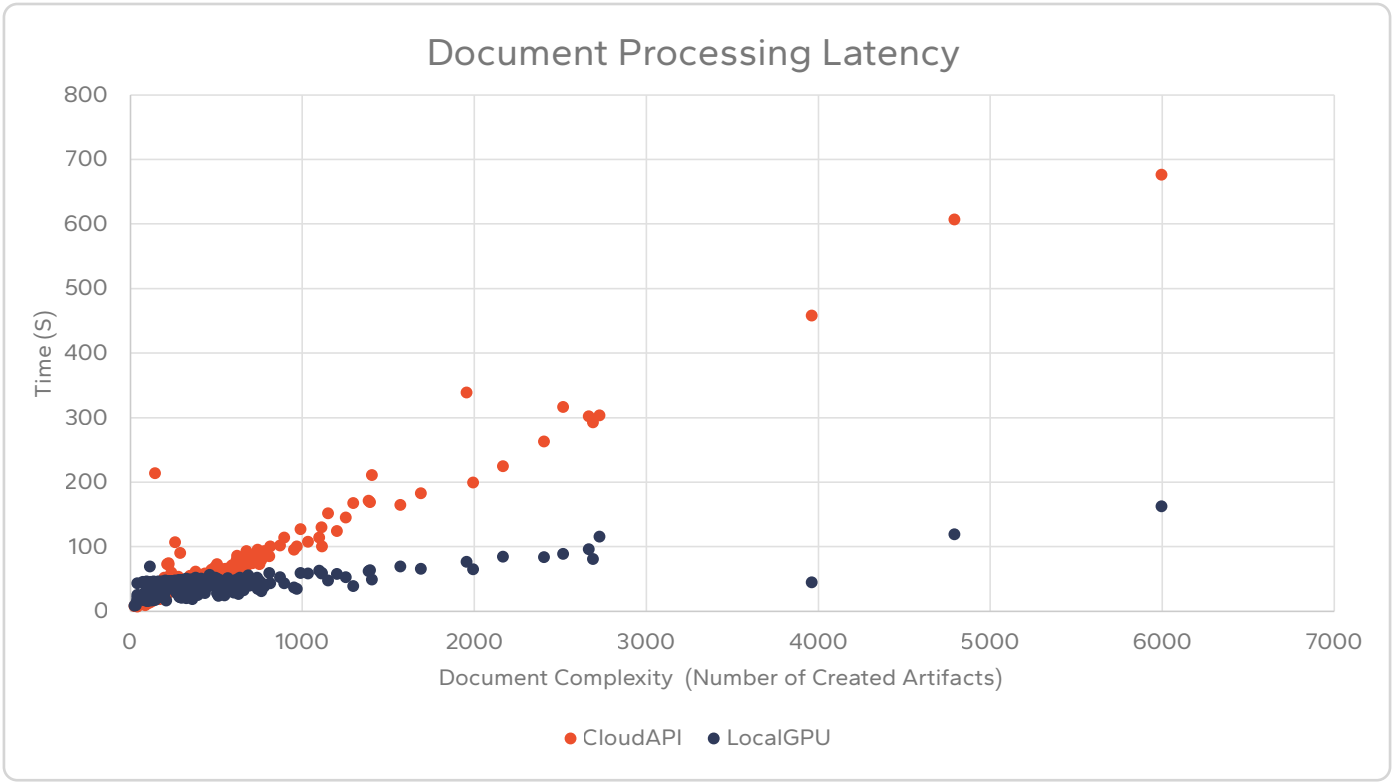


Figure 2: *Document Processing Latency*

Energy Efficiency

On-premises data preparation yields measurable sustainability benefits by reducing the power required per document processed.

- **Lower Energy per Token:** Purpose-built infrastructure is optimized for specific AI workloads, consuming less energy per generated token than cloud-equivalent workloads, which often run on generalized, less efficient hardware.
- **Power Optimization:** The Dell PowerEdge XE7745 is designed for high performance per watt, ensuring large-scale deployments align with enterprise sustainability objectives.
- **Reusable Source Data:** By transforming unstructured content into locally enriched semantic twins, organizations avoid the energy-intensive process of repeatedly re-ingesting or re-reading raw data across different applications.

Total Cost of Ownership (TCO)

While on-premises deployment requires an initial infrastructure investment, it offers a more favorable ROI timeline for steady-state enterprise AI operations.

- **Capital vs. Operational Expense:** By bringing the data pipeline in-house, enterprises eliminate the escalating costs of third-party API calls and unpredictable cloud egress fees.
- **WORM-like Data Lake Efficiency:** The TopicLake architecture uses a Write-Once-Read-Many (WORM) storage pattern, providing cost-efficient, auditable storage for long-term data retention.
- **Cost Avoidance:** Local processing mitigates the hidden costs of cloud processing, such as latency-induced performance degradation and the high financial risks associated with security exposures or regulatory non-compliance.

Business Benefits

The transition to on-premises data preparation using Dell PowerEdge and Broadcom infrastructure creates a strategic foundation that extends beyond technical performance, delivering tangible value across the enterprise. By shifting to a topic-oriented data lake, organizations move from reactive data management to a proactive utility model.

Enterprise Data Sovereignty

Maintaining the TopicLake repository on controlled Dell and Broadcom infrastructure enables organizations to enforce data sovereignty in accordance with regulatory, compliance, and zero-trust security frameworks. By keeping data preparation and enrichment within defined organizational boundaries, enterprises minimize

exposure to third-party environments while maintaining clear jurisdictional control over sensitive information. This approach supports compliance with data protection and residency requirements across regulated industries, including financial services, healthcare, and government.

On-premises deployment provides full policy control over document processing pipelines, including model selection, versioning, and access governance. Local execution of language models enables sensitive or classified content to be processed without sharing external data, aligning with zero-trust principles that assume no implicit trust in external systems. All access to data and models can be governed through explicit policy enforcement, continuous verification, and auditable controls, ensuring alignment with internal risk management and regulatory obligations.

Security is enforced natively at the data layer through Compute-Ready Documents, where every topic and enriched artifact inherits the security classification, access controls, and compliance attributes of its source document. This inheritance model enables access-aware retrieval that enforces least-privilege access by design, preventing unauthorized disclosure and eliminating permission leakage as data moves between storage, retrieval, and AI consumption layers. The result is a verifiable chain of custody and a defensible security posture that supports regulatory audits, incident response, and long-term compliance assurance.

Operational Improvements

Standardizing unstructured data into document semantic twins simplifies downstream operations and significantly accelerates time-to-insight. By transforming source documents into compute-ready assets at ingestion, the architecture eliminates the need to repeatedly retrofit data pipelines to accommodate new applications or evolving requirements. This approach reduces operational complexity and allows organizations to scale AI initiatives without continuously reengineering their data foundations.

By performing summarization, sentiment analysis, and entity extraction during ingestion, the system establishes a consistent, high-quality data layer that can be reused across AI workloads. Compute-ready documents are exposed through high-performance APIs, enabling seamless integration with existing business intelligence and operational systems. As a result, enterprises can deploy new analytics and AI-driven capabilities more quickly, with greater consistency and lower integration overhead.

Industry Applications

The versatility of the TopicLake architecture allows it to address the unique challenges of data-heavy industries.

Industry	Primary Use Case
Financial Services	Normalizing regulatory filings, research documents, and market reports into queryable assets.
Healthcare	Transforming clinical documentation and vast research literature into structured, searchable data twins.
Manufacturing	Converting complex technical specifications and quality records into actionable operational data.
Legal	Automating the analysis of contracts and case documentation while maintaining strict chain of custody.

Table 3: *TopicLake Versatility*

Figure 3: *On-Premise Advantages*

Conclusion

The transformation of unstructured enterprise data into compute-ready formats is a critical foundation for successful AI deployments. As this paper demonstrates, on-premises data preparation delivers measurable advantages across three dimensions that matter most to enterprise organizations: data sovereignty, ensuring complete control over document processing and model selection; higher token throughput that outperforms cloud-dependent alternatives; and energy efficiency, reducing operational costs while supporting sustainability objectives. The TopicLake Insights architecture, combining topic-oriented data organization with enriched compute-ready documents, enables organizations to normalize diverse content types into a unified format that serves AI applications, business intelligence platforms, and operational systems simultaneously, transforming unstructured data into a strategic utility.

For business leaders evaluating AI infrastructure investments, the choice of data preparation architecture will determine long-term success more than any single model or application decision. The Dell PowerEdge platform, paired with Broadcom high-performance networking, provides the computational density, memory capacity, and throughput characteristics required to process enterprise document corpora at scale. This infrastructure enables organizations to maintain a chain of custody from source documents through AI consumption, eliminate third-party data exposure, and achieve predictable performance without the latency variability and escalating costs associated with cloud processing. Signal65 analysis confirms that purpose-built on-premises infrastructure delivers superior token generation rates while consuming less energy per document processed, accelerating ROI while reducing operational risk.

Compute-ready data will define enterprise AI readiness. Organizations that invest in robust data preparation infrastructure today, combining the processing power of Dell PowerEdge servers, the network efficiency of Broadcom connectivity, and the architectural sophistication of topic-oriented data lakes, will lay the foundation for adopting emerging AI capabilities as they mature. Rather than retrofitting data pipelines to meet each new application requirement, these enterprises will operate from a position of preparedness, with clean, governed, and accessible data assets ready to power the next generation of AI innovation. The competitive advantage belongs to organizations that recognize this fundamental truth: AI systems are only as valuable as the data they consume.

Acknowledgments

The authors would like to thank Gadget Software for demonstrating TopicLake and Dell Technologies for providing access to the PowerEdge XE7745 hardware platform and technical expertise. Special recognition goes to the engineering teams at Broadcom for their collaboration on network optimization strategies and to the open-source community behind vLLM for their contributions to this field.

Important Information About this Report

CONTRIBUTORS

Matthew Goldensohn
Performance | Signal65

Brian Martin
AI Data Center Performance | Signal65

PUBLISHER

Ryan Shrout
President and GM | Signal65

INQUIRIES

Contact us if you would like to discuss this report and Signal65 will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "Signal65." Non-press and non-analysts must receive prior written permission by Signal65 for any citations.

LICENSING

DISCLOSURES

Signal65 provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

ABOUT SIGNAL65

Signal65 is a leading research organization specializing in enterprise AI infrastructure optimization and deployment strategies. Our lab focuses on evaluating and optimizing AI hardware and software solutions for real-world enterprise applications, with particular expertise in large language models, retrieval-augmented generation systems, and distributed AI architectures.

For more information, visit signal65.com
or contact research@signal65.com

IN PARTNERSHIP WITH

signal65

CONTACT INFORMATION
Signal65 I signal65.com