

Enabling Federal AI Operations through Enterprise-Scale Data Ingestion with Blockify® and Intel Gaudi 3 at Sterling



Executive Summary

Challenge: Sterling, a Major Federal Systems Integrator serving DoD and Intelligence Community customers, needed to demonstrate enterprise-scale AI data processing capabilities for classified and controlled environments—where data governance, classification marking preservation, and air-gapped operation are non-negotiable requirements.

Solution: Intel and Iternal Technologies deployed Blockify® on Intel Gaudi 3 accelerators within Sterling's Solution Center, creating a reference architecture for federal customers that demonstrates how to transform millions of pages of unstructured intelligence, technical manuals, and contract documentation into AI-optimized, classification-aware knowledge bases.

Results:

- 78x improvement in LLM accuracy vs. traditional RAG pipelines
- Processing of 500,000 pages reduced from 17+ days (Xeon) to 8.5 hours (Gaudi 3)—a 48x throughput improvement
- 17.8 million pages per month processing capacity on a single Gaudi 3 core
- 100% on-premise operation with classification marking preservation for SECRET, TS/SCI, and NATO-restricted content
- Seamless integration with Red Hat OpenShift AI on existing Kubernetes infrastructure

Technology Stack: Intel Gaudi 3 (Data Center Ingestion) + Intel Core Ultra AI PCs (Edge Inference) | Blockify® + AirgapAI™ + Red Hat OpenShift AI

The Federal AI Challenge: Classification-Aware Intelligence at Scale

AI is rapidly becoming a strategic asset for defense and intelligence operations. To maintain mission advantage, analysts, operators, and warfighters must access the right information, at the right time, in a safe, secure, and trusted manner—while being absolutely certain every insight is factually correct and compliant with stringent classification requirements.

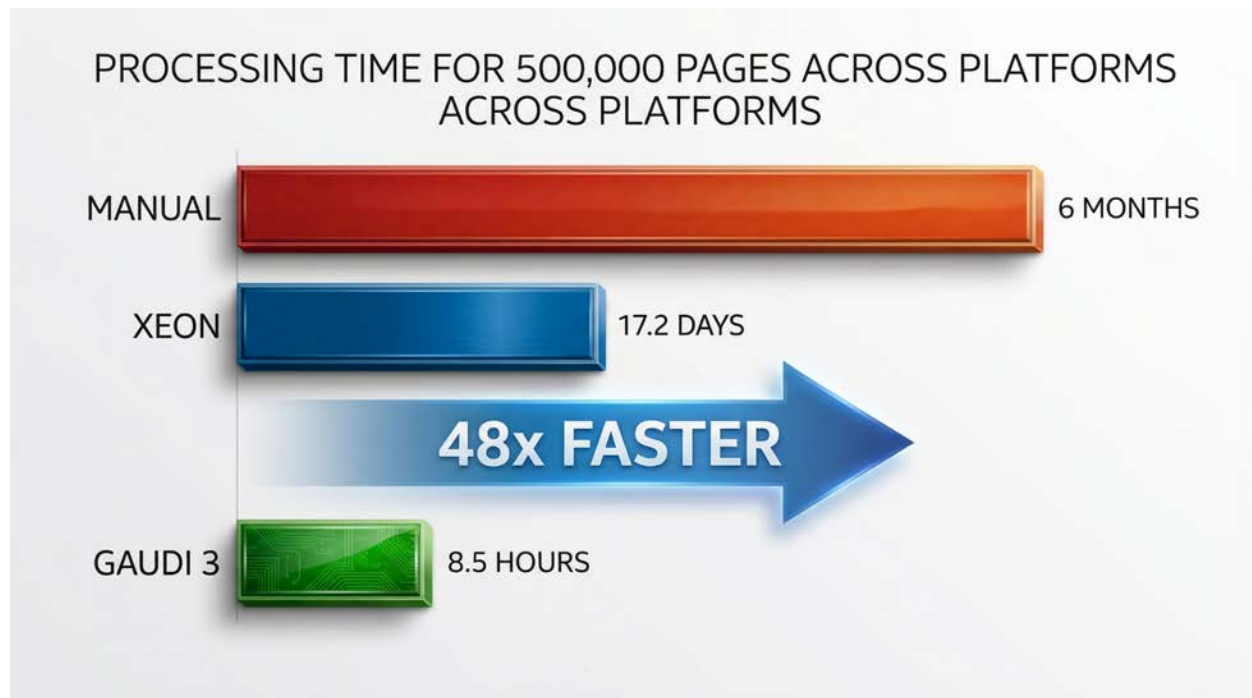
For federal systems integrators like Sterling—serving the Department of Defense, Intelligence Community, and federal agencies—the challenge extends beyond technical capability. Their customers operate in environments where:

- **Classification markings must be preserved** - SECRET, TS/SCI, FOUO, and NATO-restricted content cannot be commingled or mishandled
- **Air-gapped operation is mandatory** - Data cannot leave controlled facilities or traverse external networks
- **Role-Based Access Control (RBAC)** - Different clearance levels require different data access, even within the same AI system
- **Scale meets security** - Millions of legacy documents require processing without compromising operational security

Together, Intel and Internal Technologies have partnered to provide a unique solution that addresses Sterling's federal customer requirements, deploying Blockify® on Intel Gaudi 3 to demonstrate enterprise-scale AI data ingestion with full classification governance.

Benchmark Results: Gaudi 3 Performance for Federal Workloads

Sterling's Solution Center deployment demonstrates the transformative throughput of Intel Gaudi 3 for federal AI operations. The benchmark processed a representative corpus of 500,000 pages of technical documentation through Blockify®'s data optimization pipeline.



Intel Gaudi 3 (Production Deployment):

Leveraging Intel Gaudi 3 with OpenVINO and VLLM-Havana, Blockify® processing completes in approximately **8.5 hours** for 500,000 pages. The ability to process **17.8 million pages of text per month** on a single Gaudi 3 core demonstrates the enterprise-scale efficiency required for federal data center operations.

Metric	Value
Total Processing Time	≈ 30,750 seconds (8.5 hours)

Total Pages	500,000 pages (125 million words)
Tokens per Second	≈ 5,420 tokens/second
Words per Second	≈ 4,065 words/second
Monthly Capacity	17.8 million pages

Intel Xeon 6767P (Comparison Baseline):

For organizations beginning their AI journey or requiring non-accelerated processing, Intel Xeon with OpenVINO provides a capable baseline with significant processing capacity.

Metric	Value
Total Processing Time	≈ 1,488,095 seconds (413 hours / 17.2 days)
Total Pages	500,000 pages (125 million words)
Tokens per Second	≈ 112 tokens/second
Words per Second	≈ 84 words/second
Monthly Capacity	622,000 pages

Throughput Comparison:

Platform	Time for 500K Pages	Speed vs. Manual	Gaudi 3 Advantage
Manual Review	~6 months	Baseline	528x faster
Intel Xeon 6767P	17.2 days	10x	48x faster
Intel Gaudi 3	8.5 hours	528x	Reference

Accuracy Improvements:

Beyond speed, Blockify®'s approach increased the precision of vector searches and RAG models, virtually eliminating hallucinations and improving LLM accuracy by approximately **78× (7,800 percent)** compared with a traditional RAG pipeline. For intelligence operations where a single inaccurate fact can compromise mission success, this accuracy improvement is mission-critical.

Why Classification-Aware AI Matters for Federal Operations

Federal AI deployments face unique challenges that commercial solutions cannot address. Traditional approaches to RAG (Retrieval-Augmented Generation) and AI knowledge management fail in classified environments:



Traditional AI Pain Points for Federal Customers:

- **Classification commingling** - Standard RAG pipelines cannot distinguish between UNCLASSIFIED and SECRET content, creating spillage risk
- **No provenance tracking** - When AI returns information, analysts cannot verify the classification of source documents
- **RBAC impossibility** - Traditional vector databases lack the metadata structure to enforce need-to-know restrictions
- **Coalition operations complexity** - NATO and Five Eyes partners require different data access even within the same operational context
- **Declassification workflows** - Legacy content requires systematic review as classification expires

Risk Exposure: Without an effective approach to content lifecycle and classification governance, organizations risk relying on an AI that will never be fully accurate. It's a case of garbage in, garbage out—and if you can't trust your AI to continually be correct, you can't trust it at all because the mistakes are unpredictable. For intelligence operations, unpredictable AI errors aren't just inconvenient—they're potential mission failures.

Solution Summary: Blockify® on Gaudi 3 with OpenShift AI

Blockify® Data Ingestion on Intel Gaudi 3

Blockify® is Iternal's patented data ingestion and optimization system that improves AI accuracy by up to 78x (7,800%). Through intelligent chunking and taxonomy creation, Blockify® transforms unstructured documents into optimally structured data for LLM processing—ensuring each chunk contains the right context for accurate extraction while preserving critical metadata including classification markings.

Sterling's Deployment Architecture:



Key Technical Components:

- **Intel Gaudi 3** - Enterprise AI accelerator for data center processing
- **VLLM-Havana** - Intel's optimized fork for Gaudi inference
- **Red Hat OpenShift AI** - Kubernetes-native AI/ML platform
- **N8N Workflow Automation** - Open-source pipeline orchestration
- **Llama 8B Fine-Tuned Model** - Blockify® optimized for structured output

AirgapAI™ Edge Inference on Intel Core Ultra

For field operations, SCIF deployments, and disconnected environments, AirgapAI™ delivers 100% offline AI capabilities on Intel Core Ultra AI PCs:

Intel Core Ultra processors with integrated NPU and OpenVINO optimization deliver enterprise AI capabilities directly on employee devices:

- Up to 4.9x faster inference and 112x faster first-token latency with OpenVINO
- 100% local operation with no network dependency
- Sustained AI workloads without impacting system performance
- Edge deployment for field operations and remote environments



The Blockify® Workflow: Classification-Preserving Data Transformation

The Blockify® workflow steps include:

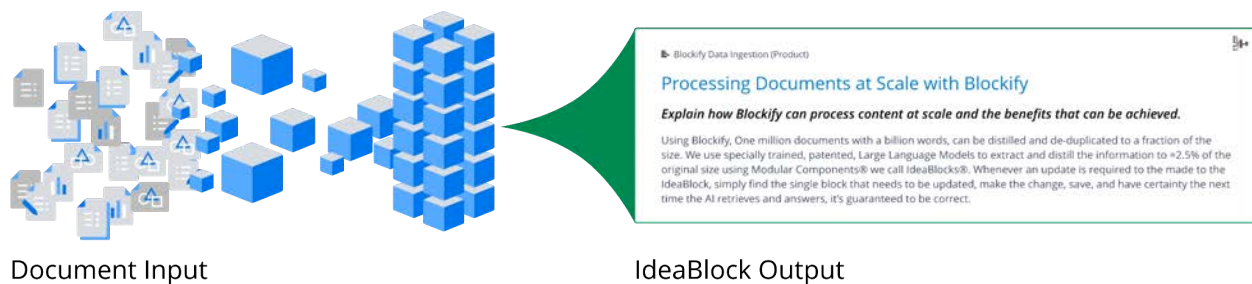
1. **Chunking the Text:** Source documents divided into smaller content chunks based on a proprietary algorithm. Chunks passed into specially configured LLM, outputting modular blocks with robust taxonomy for reuse—including classification metadata preserved from source documents.
2. **Embeddings with Governance Tags:** Content blocks converted into embeddings (vector representations) capturing unique context and structure,

enabling content-aware retrieval within AirgapAI™. Critical classification markings injected as structured XML tags for downstream RBAC enforcement.

3. **Retrieval and Response Generation:** Based on user queries and clearance level, the system retrieves relevant content from the Classification-Aware Retrieval Database for accurate, contextually relevant responses that respect need-to-know restrictions.

Structured Output with Classification Preservation:

Blockify® outputs structured XML blocks that preserve classification metadata throughout the AI pipeline:



None

```
<ideablock>
  <name>Blockify's Performance Improvement</name>
  <critical_question>What performance improvement does Blockify
offer?</critical_question>
  <trusted_answer>Up to 78X aggregate enterprise performance
improvement through intelligent data distillation and semantic
optimization.</trusted_answer>
  <tags>IMPORTANT, INFORM (WITHOUT EMOTION), COMPANY FOCUS, PRODUCT
FOCUS, PERFORMANCE, EFFICIENCY, INNOVATION, DATA, GROWTH</tags>
  <entity>
    <entity_name>BLOCKIFY</entity_name>
    <entity_type>PRODUCT</entity_type>
  </entity>
  <keywords>Blockify, Performance Improvement, 78X, Data
Distillation, Semantic Optimization</keywords>
</ideablock>
```

This structured approach enables:

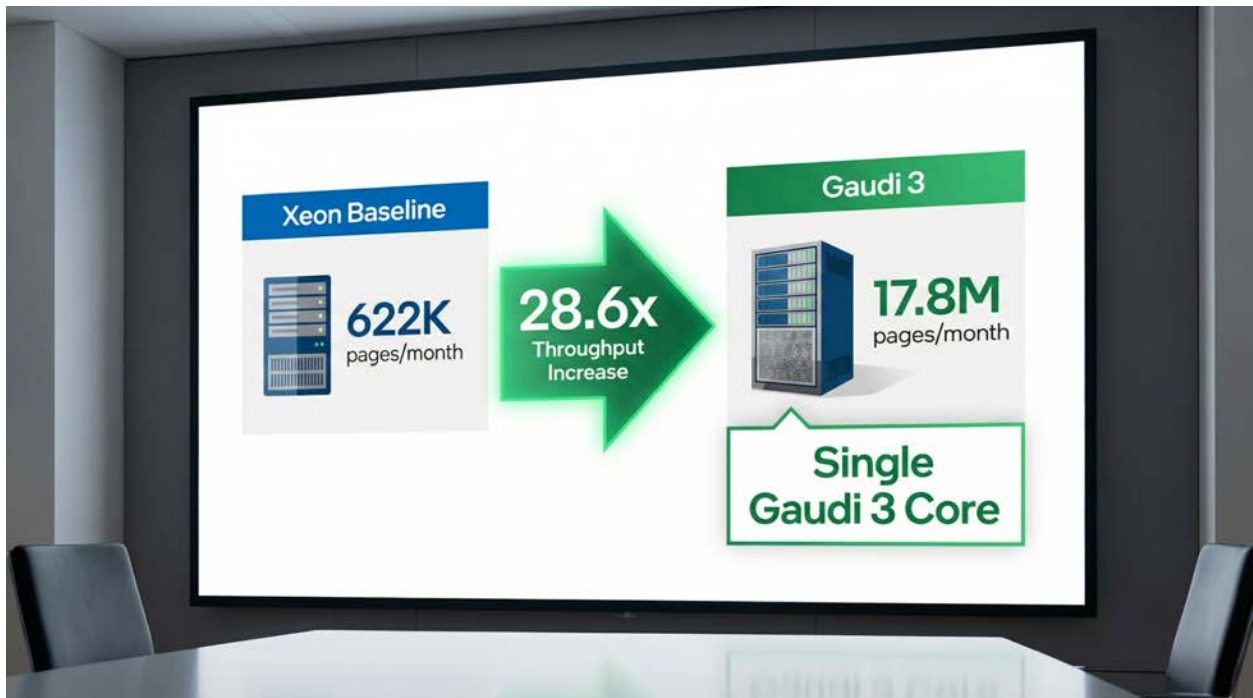
- **RBAC enforcement** - Query systems filter results by user clearance
- **Coalition operations** - Different data subsets for Five Eyes vs. NATO vs. US-only
- **Audit compliance** - Complete provenance from source document to AI response
- **Declassification automation** - Systematic review as classification dates expire

Business Use Cases: Federal AI Operations

- **Intelligence Analysis Acceleration:** Analysts query millions of intelligence reports, cables, and assessments through natural language—receiving accurate, classification-appropriate responses in seconds rather than hours of manual research.
- **Technical Manual Q&A for Maintenance Operations:** Maintenance personnel access procedures, specifications, and troubleshooting guidance for complex weapons systems through AI-powered retrieval, reducing downtime and ensuring procedural compliance.
- **Contract and Acquisition Document Analysis:** Program offices extract key terms, compliance requirements, and deliverables from thousands of contract documents—accelerating acquisition cycles and reducing oversight burden.
- **Coalition Operations Support:** NATO and allied operations receive AI-powered intelligence support with automatic filtering based on release markings—preventing inadvertent disclosure while enabling collaboration.
- **CDRL and Deliverable Generation:** Automatically draft Contract Data Requirements Lists, technical reports, and program documentation using approved organizational content—ensuring consistency and compliance.
- **Security Classification Review:** Systematically review legacy content for declassification eligibility based on date triggers and content analysis—transforming months of manual review into automated workflows.

The Single Source of Truth: Distillation for Federal Knowledge Bases

Blockify®'s patented approach creates a single source of truth. Distilling technical manuals, intelligence reports, and policy documents down to 2.5% of the original size allows for easy content lifecycle management and virtually eliminates hallucinations.



Why Distillation Matters for Federal Operations:

The Problem: A typical DoD program generates thousands of documents over its lifecycle—requirements, specifications, test reports, modifications, and updates. Information is repeated, contradicted, and superseded across versions. When analysts or AI systems query this corpus:

- **Version conflicts** - Which specification is current?
- **Contradictory information** - Earlier documents contain superseded facts
- **Redundancy overhead** - Same facts repeated across hundreds of documents
- **Update propagation** - Changes in one document require updates across many

The Solution: Blockify® distillation consolidates redundant information, identifies conflicts, and creates authoritative blocks that represent the current state of knowledge:

Original Corpus	After Blockify® Distillation
500,000 pages	~12,500 pages (2.5%)

~350,000 unique blocks	~8,750 distilled blocks
Redundancy: 96%	Redundancy: <1%
Review time: Months	Review time: Days

For federal customers managing decades of legacy documentation, this distillation creates manageable, reviewable knowledge bases that humans can actually oversee and approve—ensuring the AI only accesses verified, current information.

Maximizing AI Investment with The AI Strategy Blueprint

Technology alone doesn't guarantee AI success—85% of enterprise AI investments fail to deliver expected results. Iternal enhances its technology deployment with *The AI Strategy Blueprint*, a complete framework for AI transformation written by Iternal's CEO based on seven years of enterprise deployment experience.

The Blueprint addresses the 70% of AI success that depends on people and processes:

- **5x revenue gains** compared to competitors stuck in pilot programs
- **Governance frameworks** that enable speed rather than create barriers
- **Crawl-walk-run deployment strategies** that prove value before scaling risk
- **ROI quantification methodologies** that satisfy CFO scrutiny

This strategic foundation ensures Sterling's federal customers extract maximum value from their Blockify® and AirgapAI™ deployments rather than joining the majority of failed AI initiatives.

Accelerating Adoption with Iternal AI Academy

The fastest path to AI ROI is workforce capability. Iternal enhances its technology with the AI Academy (edu.italnal.ai)—comprehensive AI fluency training that accelerates how quickly analysts, operators, and program staff can leverage AI tools:

- **610+ courses** spanning every major industry and job role

- **Three proficiency tiers** - Beginner, Advanced, and Expert tracks
- **Real-time AI feedback** - Prompts scored and critiqued for immediate improvement
- **Verified certificates** - Independently verifiable credentials demonstrating competency
- **Universal application** - Skills transfer to any AI platform

By combining foundational education with interactive practice, AI Academy ensures Sterling's federal customers can fully leverage their technology investment from day one.

Immediate Productivity with 2,800+ Quick Start Workflows

AirgapAI™'s technology is enhanced with more than 2,800 pre-configured Quick Start Workflows—enabling non-technical users across federal agencies to achieve immediate productivity gains without a learning curve.

Pre-built workflows for federal operations include:

- Intelligence report summarization and key finding extraction
- Contract clause analysis for compliance verification
- Technical manual Q&A for maintenance operations
- Program status briefing generation
- Proposal response drafting for competitive acquisitions
- Security classification review assistance

Custom workflow generation: Users can describe their work process in natural language, and AirgapAI™ automatically creates a custom workflow tailored to their specific mission needs—no technical expertise required.

These workflows transform AirgapAI™ from a powerful tool into an immediate productivity multiplier for federal operations.

Maximizing ROI with AI Assist Bundle

For organizations seeking to maximize their return on AI investment, the AI Assist Bundle enhances Blockify® and AirgapAI™ with comprehensive end-to-end services:

- **Full deployment and integration** - Expert setup and configuration

- **Training and education** - Workforce enablement programs
- **Tier 1, 2, and 3 support** - Phone and chat support for all user levels
- **Ongoing optimization** - Continuous improvement and workflow refinement

The AI Assist Bundle transforms technology deployment into measurable mission outcomes—cost reduction, productivity gains, and operational advantage—at a fraction of the cost of cloud-based alternatives while keeping all data 100% on-premise.

What This Means for You: Scalable AI Inference for Federal Operations

By uniting Intel Gaudi 3 with OpenVINO and Iternal Technologies' Blockify® and AirgapAI™, federal organizations can harness AI inference at enterprise scale—turning decades of unstructured intelligence, technical manuals, and contract documentation into actionable insights without ever exposing classified information to third-party clouds.

Together, Intel and Iternal help Federal Organizations:

- **Automate Data Ingestion & Curation:** Process millions of pages of legacy documentation while slashing inaccuracies with Blockify®'s 78x accuracy improvement
- **Preserve Classification Integrity:** Ensure every AI response respects classification markings, RBAC policies, and need-to-know restrictions
- **Accelerate Decision-Making:** Surface trusted answers in seconds with OpenVINO-optimized inference on Gaudi 3
- **Protect Classified Information:** Keep all data within organizational boundaries with 100% air-gapped operation

Enhancing Technology with Complete AI Enablement:

- **Maximize Investment:** The AI Strategy Blueprint ensures your deployment succeeds where 85% of AI initiatives fail
- **Accelerate Adoption:** AI Academy training builds workforce capability for faster time-to-value
- **Immediate Productivity:** 2,800+ Quick Start Workflows eliminate the learning curve
- **White-Glove Support:** Optional AI Assist Bundle provides end-to-end deployment services



AI has moved from conceptual to operational in federal environments. Whether accelerating intelligence analysis, enabling coalition operations, or transforming technical documentation into AI-powered knowledge bases, Intel and Iternal deliver the secure, accurate, and scalable framework federal systems integrators like Sterling need to maintain mission advantage.