

Blockify for Data Orchestration Engine

Dramatically improve DOE RAG Pipelines with Blockify

Abstract

This document describes how Blockify integrated with Dell Data Orchestration Engine (DOE) and Data Search Engine (DSE) can dramatically improve RAG pipelines.

Technical White Paper

Notes, cautions, and warnings

 **NOTE:** A NOTE indicates important information that helps you make better use of your product.

 **CAUTION:** A CAUTION indicates either potential damage to hardware or loss of data and tells you how to avoid the problem.

 **WARNING:** A WARNING indicates a potential for property damage, personal injury, or death.

Chapter 1: Chapter placeholder..... 4

 Executive summary..... 4

 Audience.....4

 Prerequisites..... 4

 Terminology.....4

 We value your feedback.....5

 The problem: why naive RAG fails in production..... 5

 How Blockify works in Data Orchestration Engine.....6

 Solution summary..... 6

 Pipeline stages.....6

 Deployment architecture.....7

 The IdeaBlock format.....7

 Deployment footprint.....8

 Conclusion..... 8

 Benefits of Blockify..... 8

 References..... 8

 Dell Technologies AI Solutions..... 8

 Blockify.....8

Chapter placeholder

Executive summary

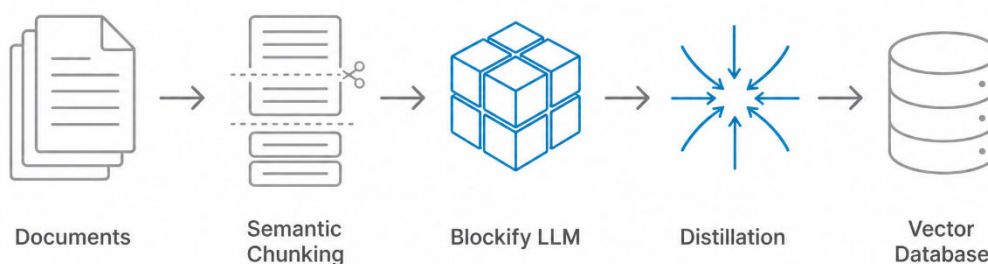


Figure 1. RAG with Blockify

Blockify is a patented data optimization platform that transforms unstructured enterprise content into compact, governed knowledge units called IdeaBlocks. Integrated with Dell's Data Orchestration Engine, Data Search Engine, and Data Analytics Engine, Blockify delivers up to 78X aggregate AI accuracy improvement and ~3X token efficiency while reducing dataset size to ~2.5% of original — all running 100% on-prem within Dell enterprise infrastructure. No data leaves the customer environment.

Audience

The intended audience for this document includes data owners, data scientists, data engineers, and AI/ML software developers who want to build and deploy RAG pipelines with DOE and Blockify.

Prerequisites

You must have developer account access to DOE. For support on DOE, contact the Dell Technologies Solutions team by [email](#).

Your organization and project spaces must be already set up on DOE, and you must have an active Developer/Project Owner account linked to your org and project space(s).

Terminology

The following table provides definitions for some of the terms that are used in this document.

Table 1. Terminology

Term	Definition
Dell AI Data Platform (Dell AIDP)	The Dell AI Data Platform is a comprehensive solution designed to support the entire data lifecycle for your analytics and AI applications. It provides a unified, modern approach to managing complex data workflows at-scale, empowering your teams to turn raw data into strategic assets and accelerate innovation.
Data Orchestration Engine (DOE)	Dell AI Data Orchestration Engine provides a unified platform for data, models, agents, pipelines, and GenAI workflows.
Data Search Engine (DSE)	The Dell Data Search Engine is a specialized, AI-driven search and vector database solution developed in collaboration with Elastic, designed to help enterprises find, analyze, and use unstructured data for AI applications.
Blockify	Blockify is a patented data optimization platform from Iternal that transforms unstructured enterprise content into compact, governed knowledge units called IdeaBlocks.

This document may contain language from third-party content that is not under Dell Technologies control and is not consistent with current guidelines for Dell Technologies own content. When such third-party content is updated by the relevant third parties, this document will be revised accordingly.

This document may contain language that is not consistent with Dell Technologies’ current guidelines. Dell Technologies plans to update the document over subsequent future releases to revise the language accordingly.

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on the solution and the solution documentation. Contact the Dell Technologies Solutions team by [email](#).

Author: John Byron Hanby, IV (Iternal.ai), Sunil Upalekar (Dell)

Contributors: Gaurav Chawla, Vrashank Jain

The problem: why naive RAG fails in production

Traditional Retrieval-Augmented Generation pipelines follow a "dump-and-chunk" approach: documents are sliced into fixed-size fragments, embedded, and stored in a vector database. This breaks down at enterprise scale:

- **Version conflicts** — old and new content coexist in the same index
- **Semantic fragmentation** — fixed-length chunking splits ideas mid-sentence
- **Duplication bloat** — average 15:1 content duplication across enterprise corpora according to Gartner and IDC
- **Permission leaks** — document-level governance can't tag classified content at chunk level
- **Hallucinations** — LLMs blend conflicting chunks into authoritative-sounding wrong answers

Gartner predicts 60% of AI projects will be abandoned through 2026 due to data-quality issues.

The model isn't the problem — the data feeding it is.

How Blockify works in Data Orchestration Engine

Solution summary

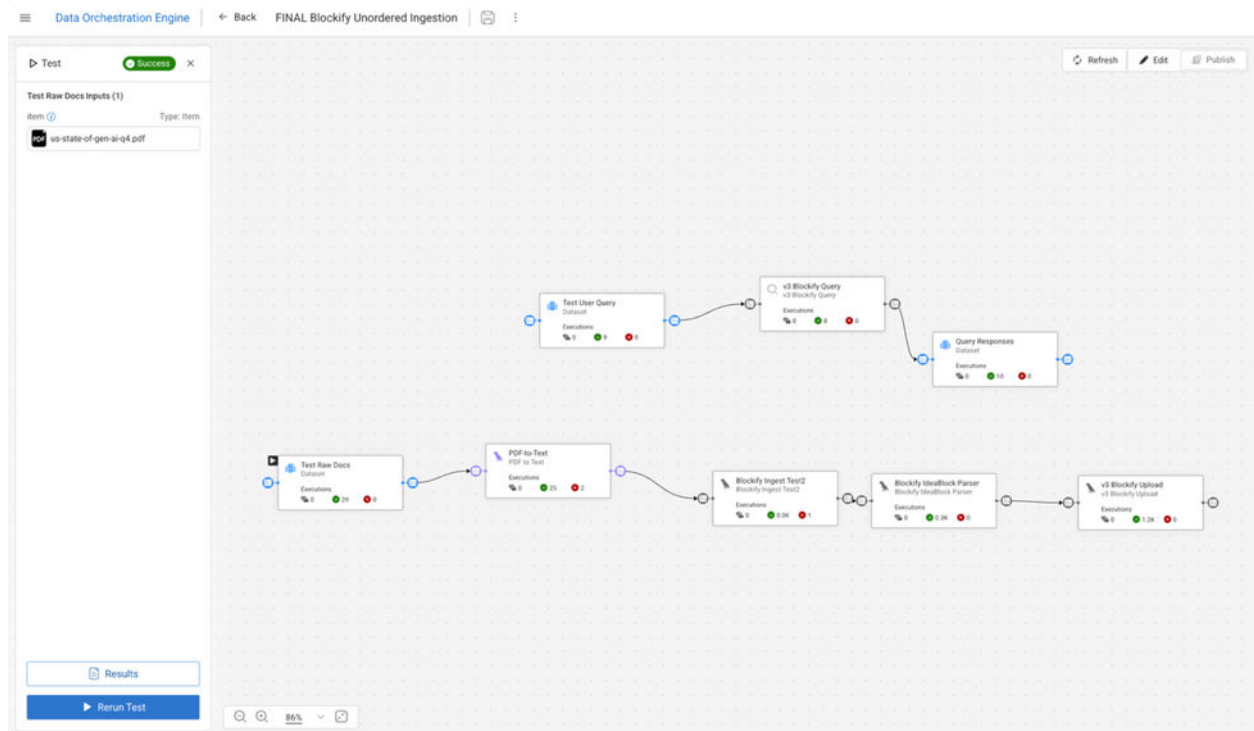


Figure 2. Dell Data Orchestration Engine dashboard

The Blockify Unified Service deploys as a Docker container through the Data Orchestration Engine Hub and connects to any on-prem Data Search Engine database. The pipeline runs entirely within the Dell tenant.

Pipeline stages

1. **Ingestion:** Raw documents (PDF, DOCX, PPTX, HTML) are uploaded via Data Orchestration Engine and parsed into clean text
2. **Semantic Chunking:** Content-aware splitting at natural boundaries, not fixed character windows
3. **IdeaBlock Generation:** A fine-tuned Blockify LLM (running in VLLM or NVIDIA NIM) converts chunks into structured XML knowledge units containing name, critical_question, trusted_answer, tags, entities, and keywords
4. **Distillation & Deduplication:** Data Search Engine-backed clustering merges duplicate blocks, collapsing 1,000 versions of a mission statement into 2-3 canonical blocks
5. **Retrieval:** User queries return metadata-rich, deduplicated results with similarity scores and governance tags.

Deployment architecture

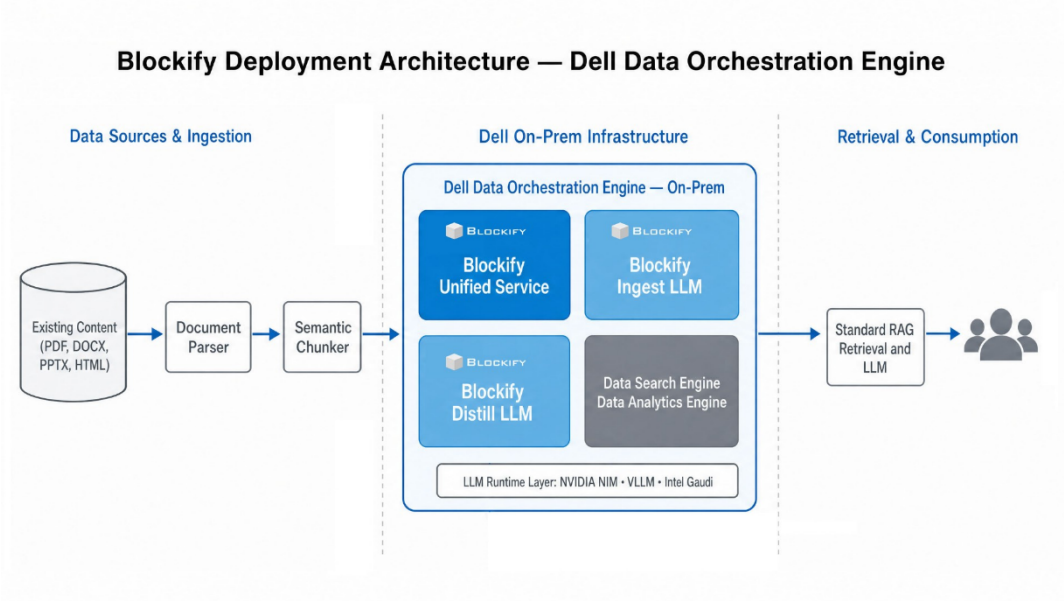


Figure 3. Blockify deployment architecture on DOE

The IdeaBlock format

Each IdeaBlock is a self-contained XML knowledge unit. The Q&A structure mirrors how users query AI systems, creating a natural bridge between user intent and stored knowledge.

 BLOCKIFY; INTERNAL; IDEABLOCKS; RAG

Blockify's Data Optimization Platform

What is Blockify?

Blockify is Internal's patented data optimization platform that transforms enterprise content into governed IdeaBlocks, delivering 78X accuracy improvement and 3X fewer tokens compared to traditional RAG.

Blockify

Internal

Data Optimization

IdeaBlocks

RAG

78X

3X

Figure 4. Blockify IdeaBlocks

```
<ideablock>
  <name>Blockify Overview</name>
  <critical_question>What is Blockify?</critical_question>
  <trusted_answer>Blockify converts unstructured content into compact, deduplicated XML knowledge units optimized for RAG.</trusted_answer>
  <tags>RAG, DATA_OPTIMIZATION</tags>
  <entity>
    <entity_name>BLOCKIFY</entity_name>
    <entity_type>PRODUCT</entity_type>
  </entity>
  <keywords>blockify, ideablock, RAG, distillation</keywords>
</ideablock>
```

Deployment footprint

1. Blockify components ship as Docker containers managed through Data Orchestration Engine Hub
1. **Blockify Unified Service:** Core ingestion and distillation engine
2. **Blockify Ingest LLM:** Fine-tuned model for IdeaBlock generation
3. **Blockify Distill LLM:** Specialized model for cluster merging
4. **Data Search Engine:** On-prem database for IdeaBlock storage and metadata
5. **Data Analytics Engine:** On-prem data governance and analytics

Conclusion

Benefits of Blockify

References

Dell Technologies AI Solutions

- [Dell Technologies AI Solutions](#)
- [Dell Technologies Info Hub for AI](#)
- [Dell AI Data Platform](#)

Blockify

- [Blockify on AIDP](#)
- [Blockify study](#)