



Intelligent Content Extraction at Scale

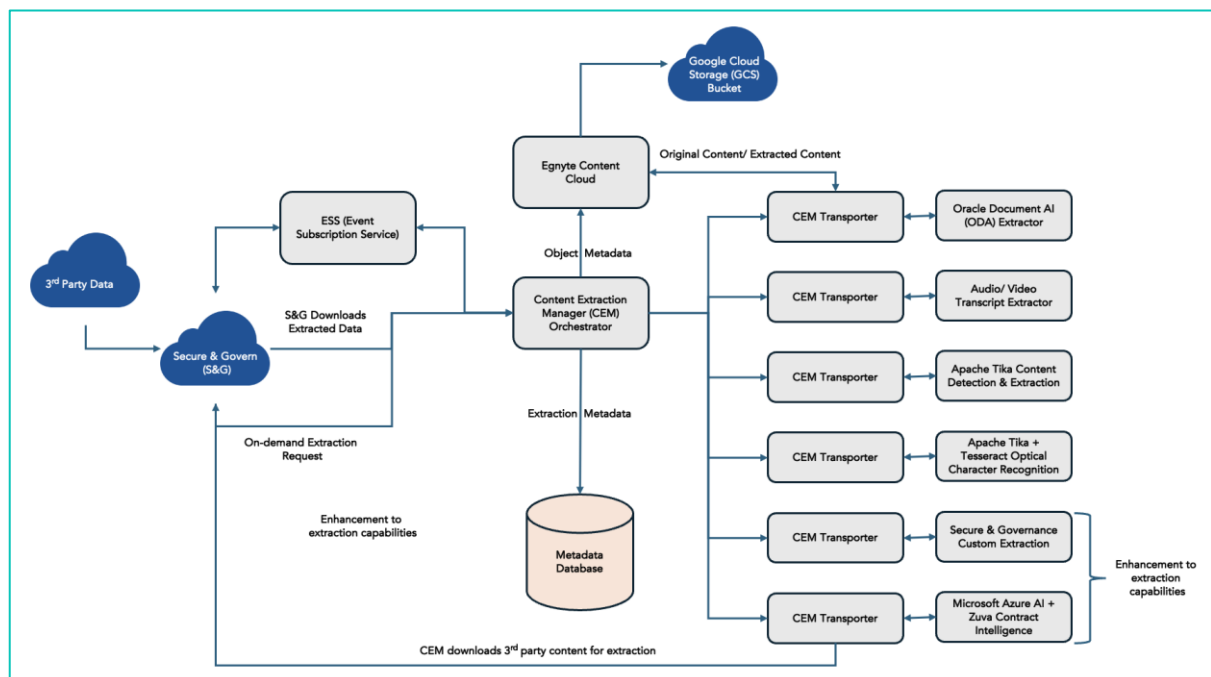
June 2026

Introduction

Organizations today manage massive volumes of unstructured content, from scanned contracts and recorded meetings to archived emails, image-heavy reports, and third-party data feeds. But content management alone is no longer enough. To drive faster decisions, organizations need to add structure to that content by extracting structured metadata from raw files, making it searchable, classifiable, and ready for AI-powered workflows.

Egnyte's secure content platform makes extraction the foundation of content intelligence. Through advanced extraction pipelines and AI-powered orchestration, Egnyte transforms raw files into structured, searchable, metadata-enriched content—turning static archives into a knowledge base that's ready for search, classification, compliance, and AI-powered workflows. The more content you bring in, the more useful the platform becomes.

Egnyte's Content Extraction Manager (CEM) addresses this at scale. Through intelligent extraction pipelines and AI-powered orchestration, CEM automatically identifies content types, routes files to the right extraction engine, and enriches each file with structured metadata. The result is a searchable, governed content repository where AI assistants and agents can surface accurate, relevant information without manual tagging or human triage.



Egnyte's CEM Architecture

Architecture Overview

The system has four main layers: Data Ingestion, Secure & Govern, Orchestration and Specialized Extraction. Each layer handles a specific job and connects with the next layer through clear, structured processes.

Layer 1: Data Ingestion

The process starts with third-party data, which is content coming from outside the platform. This can include files, documents, media, and structured data from vendors, partners, or connected business systems. The data enters as raw, unprocessed content and is sent to the platform hub for intake and coordination.

Layer 2: Secure & Govern

Secure & Governance (S&G) acts as central cloud hub for managing incoming third-party content. It tracks content changes through event notifications, supports on-demand extraction requests, and retrieves processed data for search, classification, compliance, and broader business workflows across the platform.

Layer 3: Orchestration—Content Extraction Manager (CEM)

CEM is the central system that manages the extraction process. It receives requests from S&G, identifies the content type, chooses the right extraction method, and sends the job to the correct specialized extractor.

It also tracks every job by logging metadata for auditing, reporting, and compliance, while updating Egnyte's storage system with processing details. CEM can directly pull content when needed and uses transport systems to efficiently route tasks between components, allowing the system to scale smoothly.

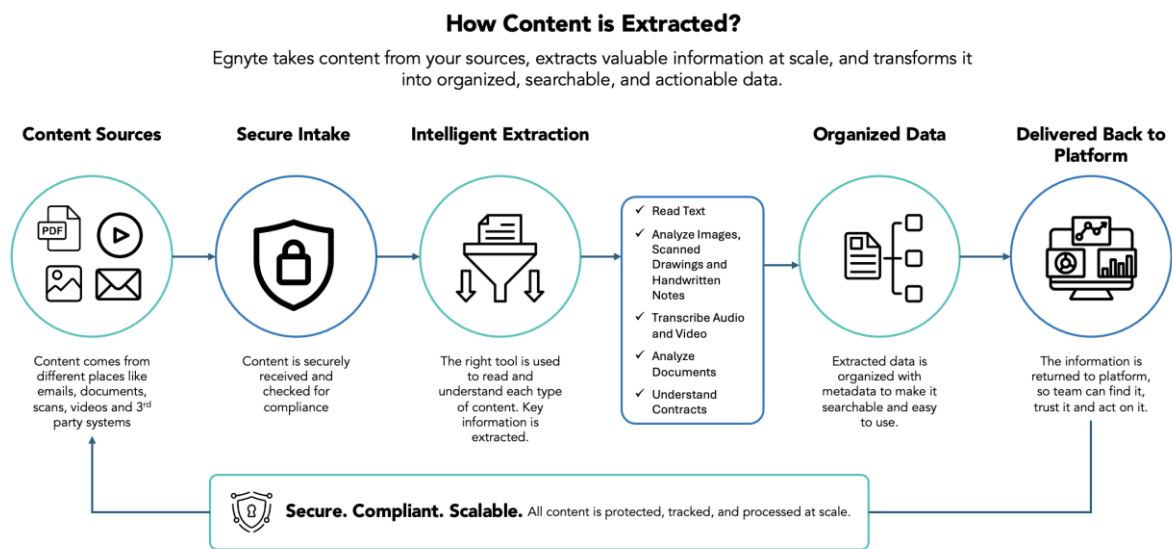
Layer 4: Specialized Extractors

The extraction layer consists of six specialized engines, each optimized for a different content type or extraction modality.

Extractor	Full Name / Technology	What It Does
ODA Extractor	Oracle Document AI Extractor	Uses AI to understand and extract information from structured documents like forms, invoices, and contracts.
Audio/Video Transcript Extractor	Speech-to-Text Transcription Engine	Converts spoken words from audio and video files into searchable text transcripts.

Tika Extractor	Apache Tika Content Detection & Extraction	Extracts text and metadata from many file types, including PDFs, Office files, and web content.
Tika Tesseract OCR Extractor	Apache Tika + Tesseract OCR	Extracts text from scanned documents, image-based PDFs, and photos using optical character recognition.
S&G Custom Extractor	S&G Custom Extraction Module	Handles Egnyte-specific content and governance needs that standard tools may not cover.
External Extractor (Azure / Zuva)	Microsoft Azure AI + Zuva Contract Intelligence	Uses advanced external AI tools for deeper analysis, including legal document review, clause detection, and entity extraction.

The S&G Custom Extractor and External Extractor provide advanced capabilities that go beyond basic text extraction, adding deeper intelligence and specialized content understanding.



Extraction at Scale

Use Cases

AEC: Automated Drawing and Specification Indexing

AEC teams manage thousands of drawings, RFIs, and specification documents throughout a project. When metadata is applied manually, inconsistencies quickly creep in. Sheet

numbers are missed, revision codes don't match, and discipline tags vary from person to person. As a result, teams struggle to find the right files, version control becomes unreliable, and connected systems like Procore and Autodesk are populated with inconsistent data.

With Egnyte's Bulk Extraction, metadata is captured automatically as documents are uploaded. The system identifies document types and extracts key details such as sheet numbers, revision codes, disciplines, and CSI sections without manual effort. Teams get consistent, searchable project data from the start, helping improve document management, streamline workflows, and keep downstream systems aligned.

Media: Automated Asset Classification and Metadata Extraction

As media libraries grow, finding and managing content becomes increasingly difficult. Inconsistent file names and manual tagging make assets harder to locate, create gaps in rights management, and slow production workflows, especially when ingest volumes increase.

Egnyte's Bulk Extraction automatically classifies incoming media assets and extracts important production details such as episode numbers, scene references, talent names, rights expiration dates, and vendor information. Assets become searchable and organized the moment they arrive, helping production teams accelerate content delivery, improve compliance, and maintain accurate metadata at scale.

Legal: Automated Contract Analysis

Legal teams rely on manual reviews to identify critical contract terms hidden within lengthy agreements. This process is time-consuming, difficult to scale, and increases the risk of missing important dates, obligations, or exposure.

Egnyte's Bulk Extraction analyzes contracts as they're uploaded, extracting key terms such as renewal dates, termination clauses, liability caps, governing jurisdictions, and counterparty details. The extracted data becomes immediately searchable and actionable, enabling faster contract management, automated retention policies, legal holds, and better visibility into contractual risk, without requiring teams to review every document manually.

Conclusion

Egnyte's platform transforms unstructured enterprise content into structured, searchable, AI-ready data. By combining large-scale extraction with intelligent metadata enrichment and governed orchestration, Egnyte enables organizations to unlock value from content previously buried in raw files.

Automated metadata extraction improves operational speed, search accuracy, and compliance while reducing manual effort. Instead of simply storing files, Egnyte helps enterprises make content fully accessible to AI assistants, agents, and users, at any scale and across any content type.