Unstructured.io Capabilities Statement

Executive Summary

Unstructured automates the preprocessing of varied, messy natural document types into clean, structured data for use in downstream AI applications, including Large Language Models (LLMs). Unstructured has built the first preprocessing platform that makes it easy to swiftly transform raw files (e.g., PDF/XML/HTML etc. → JSON) into data ready for consumption in downstream AI/ML services (e.g. data annotation, model training, production pipelines). For the Intelligence Community (IC) to be able to fully leverage the power of AI/ML, they must be able to build machine learning pipelines (including LLMs) that are built on organization-specific, gold-standard data. Our preprocessing capability delivers an astounding ~99% time savings over existing tools, a critical improvement considering data scientists typically spend approximately 80% of their time on data preprocessing. Unstructured can be leveraged for:

Key Value of Unstructured's Data Preprocessing Platform

1. Accelerates the Deployment of AI/ML (e.g. LLM) Solutions on Enterprise Data and Customer-Managed Networks
2. Provides Canonical Schema for Unstructured Data Across Enterprise
3. Eases Integration of Novel Data
4. Deploys as a Modular Rest API "Microservice"

- Fine-tuning
- Pre-training
- Retrieval Augmented Generation (RAG)
- Traditional Extract, Transform, Load (ETL)

The Emerging Large Language Model (LLM) Tech Stack

Since the fall of 2022, with the advent of OpenAI's ChatGPT, there has been a surge in interest among enterprise organizations–including the IC–in leveraging LLMs. However, the integration of tools like ChatGPT presents several challenges for security and performance-conscious enterprises. These include the inability to transmit sensitive data to an external API, the constraint of not being able to prompt LLMs about topics beyond a certain date due to their time-bound nature, and the fact that LLMs are trained on a generic corpus of internet data rather than defense-specific data. Additionally, there is a lack of trust in the outputs of generative models owing to their propensity to "hallucinate" or generate false information. To address these existing challenges, the emerging LLM tech stack–which Unstructured plays a critical role in–enables enterprises to deploy LLM pipelines that leverage recent, relevant, and validated gold-standard enterprise data.

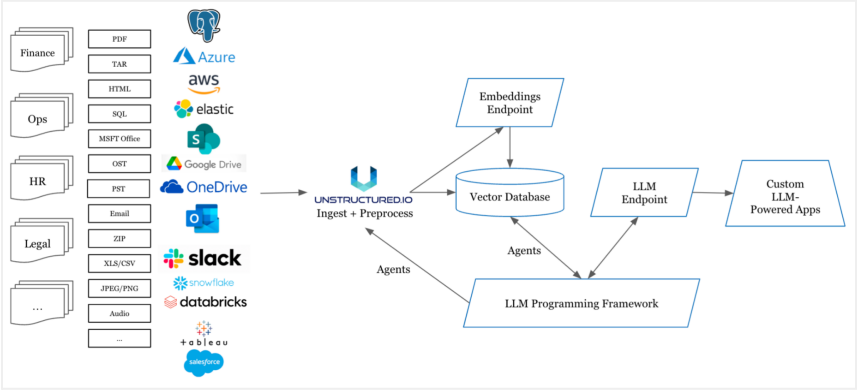


Figure 1: The new LLM tech stack.

Figure 1 depicts the architecture of the emerging LLM tech stack, primarily focused on Retrieval Augmented Generation-powered systems. First, the system ingests raw documents such as PDFs, Word Documents, PowerPoints, and emails using Unstructured. After preprocessing raw documents with Unstructured, the system embeds the raw text as vectors using an LLM and stores the vectors alongside the source text in a vector database. Importantly, Unstructured also extracts metadata from the source document—for example, filename and page number—and stores it in the vector database to provide traceability for the LLM response. With the documents now in the vector database, a user can query the system. After the user provides the query, an LLM programming framework like LangChain fetches relevant documents from the vector database, formats a prompt, and submits the prompt to an LLM endpoint. To ensure the LLM has not hallucinated an answer, the system runs a second query to the vector database using the LLM response and only provides the response to the user if the system can find a source document sufficiently similar to the response. The system can provide the user the document from the second query as the source, while the metadata that Unstructured extracts serves as a citation. This tech stack constrains an LLM to ensure that it operates solely on the basis of gold-standard enterprise data.

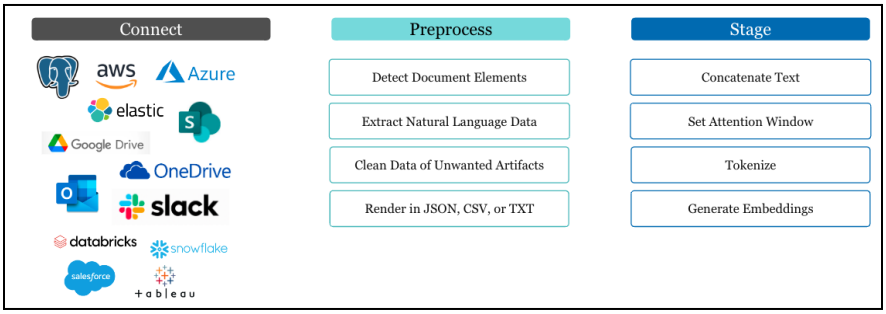
Unstructured's Capabilities

The reality of all natural language AI applications is that they require clean data from messy file types, such as PDFs and HTML. In the past, most applications have had custom pipelines built for each file type. These pipelines are brittle and frequently break down when there are minor changes to the file format. To address this issue, Unstructured has developed a toolkit to enable automated processing of varied raw natural language documents into clean, structured data for use in downstream AI applications.

How it Works

Unstructured allows organizations to ingest diverse file types from their preferred data storage platforms and quickly and easily convert that messy, natural language data into clean, structured data.. Unstructured's capabilities can be broken down into the following key segments:

1. **Data Ingest:** Connect with upstream data storage platforms and ingest any file type and layout.
2. **Pre-processing:**
    a. ***Partitioning***: Segregating data into smaller, manageable segments or partitions.
    b. ***Cleaning:*** Removing anomalies, filling missing values, and eliminating any irrelevant or erroneous information.
3. **Staging**: staging outputs for downstream tasks, such as integration into vector databases.



Data Ingest

The ability to connect with various data sources is crucial in data processing. Unstructured's upstream connectors make data ingestion easy, allowing you to seamlessly integrate your preprocessing pipeline with your preferred data storage platforms. The upstream connectors allow you to batch-process all your unstructured documents and store the resulting structured outputs locally on your file system. They ensure that your data is accessible, up to date, and usable for any downstream task. Currently[1], Unstructured supports the following pre-built connectors, each designed to integrate seamlessly with different data sources:

Airtable,	Discord,	Google Drive,	Reddit,
Azure,	Dropbox,	Jira,	S3,
Biomed,	Elasticsearch,	Local,	Salesforce,
Box,	Github,	Notion,	Sharepoint,
Confluence,	Gitlab,	One Drive,	Slack,
Delta Table,	Google Cloud Storage,	Outlook,	Wikipedia

Unlike intelligent document processing capabilities which require standard, unchanging forms for performance, Unstructured addresses the challenge of varied document types as well as supporting an extensive range of file formats. Supported file types include:

Category	**Output**
Plaintext	.eml, .html, .json, .md, .msg, .rst, .rtf, .txt, .xml
Images	.jpeg, .png
Documents	.csv, .doc, .docx, .epub, .odt, .pdf, .ppt, .pptx, .tsv, .xlsx

Pre-Processing

Before the core analysis, raw data often requires significant preprocessing. Preprocessing ensures data integrity and can significantly influence the outcomes of subsequent tasks. Unstructured features intelligent document routing that sends documents to custom extraction pipelines that optimize cost and accuracy depending on the document. Unstructured offers a variety of different ways to preprocess documents; available options include:
1) **Auto:** Chooses partition strategy based on document characteristics and other settings.
4) **Fast:** Leverages traditional NLP extraction techniques to quickly pull all the text elements.
3) **Optical Character Recognition (OCR)**: Leverages OCR to extract text from image based files.
2) **High-Resolution:** Identifies the layout of the document using state-of-the-art detection and segmentation algorithms. Uses document layout to gain additional information about document elements.

Unstructured's offering is powered by document understanding models. Document understanding techniques use an encoder-decoder pipeline that leverages the power of both

[1] As of 10/17/2023. Refer to Unstructured's documentation: https://unstructured-io.github.io/unstructured/source_connectors.html

computer vision and natural language processing methods. **Chipper** is our in-house image-to-text model based on transformer-based Visual Document Understanding (VDU) models. As Unstructured's foundation model, **Chipper** enables organizations to send thousands of heterogeneous files to a single endpoint for transformation.

Partitioning

The goal of document partitioning is to read in a source document, split the document into sections, categorize those sections, and extract the text associated with those sections. Unstructured takes an unstructured document and extracts structured content by breaking it into elements like `Title`, `NarrativeText`, `Abstract`, and `ListItem`, enabling users to decide what content they'd like to keep for their particular application. Depending on the document type, Unstructured uses different methods for partitioning a document.

Metadata Capabilities

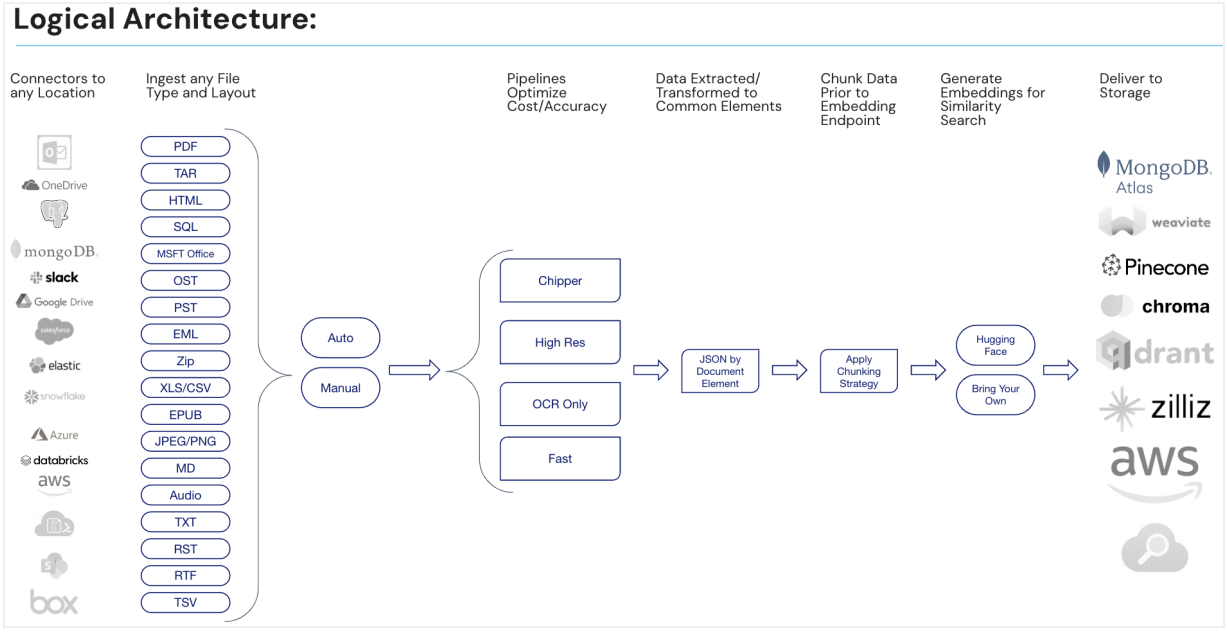
- Can track any metadata that is contained in the document all the way to storage (such as Vector Database), including critical data like RBAC information.
- 25 element ontology and can easily drop classifiers in for custom pipelines

Cleaning

Data cleaning is a critical step in data preprocessing as clean and reliable data enhances the performance and accuracy of machine learning models, enabling them to generate more meaningful insights and make more reliable predictions.

Staging

Unstructured offers capabilities to tailor data so that when data leaves the platform, it is fully ready to be utilized in downstream AI/ML services and applications, including by an LLM and a programming framework (e.g. LangChain). These tools include concatenation, tokenization, chunking, data schema modification, embeddings endpoints, and more.



How to Leverage Unstructured

Unstructured offers multiple ways for organizations to leverage our capabilities for uses ranging from prototype to production:

- Open source Python library, containers, and free API
- Paid API within AWS and Azure Marketplaces
    - Containerized for easy deployment
    - Keeps proprietary data from leaving the client's networks
- Enterprise Platform (launching end of year 2023)
    - Supports enterprises for full-services including scheduled task execution, upstream and downstream data source connectors, self-service UI, SLAs, authentication and permissioning, SOC 2/GovCloud, etc.

Additionally, for government customers, Unstructured has partnered with Second Front Systems to make our unclassified, cloud-hosted platform available on SIPR cloud networks via their Game Warden platform.

Unstructured's Government Customers

Delivering value for the following customers:
- US Air Force Testing Center (AFTC),
- US Space Force Space Systems Command TAP Lab, and
- US Army Special Operations Command