Unstructured.io

Mitigating the Risks of Bias & Generative AI Limitations with Gold-Standard Data

BENGAL Lightning Talk

Jonathan Nadzam, Public Sector

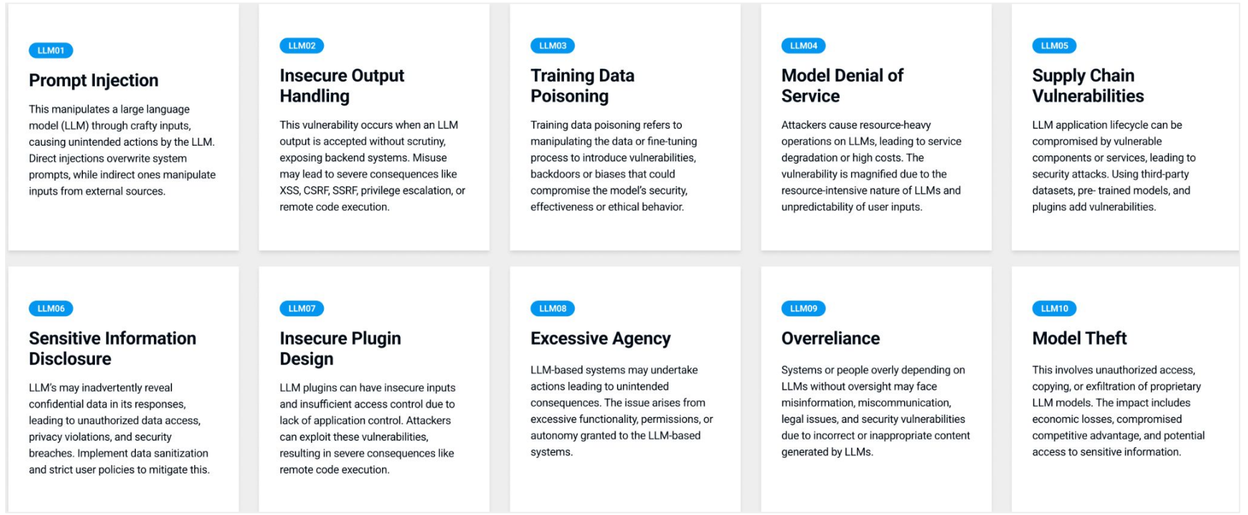
UNSTRUCTURED

UNSTRUCTURED

BENGAL

LLM Threats, Biases & Vulnerabilities

- For safe & effective use of Large Language Models (LLMs) in IC applications, critical to understand and mitigate against LLM threat vectors and vulnerabilities
- When developing LLM applications for Government use, high-quality data plays a crucial role in combating these threats and weaknesses in various stages of development.



OWASP Top 10 for LLM, https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-slides-v1_0.pdf

BENGAL

LLMs need recent, relevant, and validated data to mitigate against weaknesses

Problem 1: Nearly all LLMs have been trained on the same corpus of internet data

Solution 1: Pre-train or fine tune on **proprietary data**

Problem 2: LLMs are “frozen in time”

Solution 2: Make **new data** available to the model in a vector database

Problem 3: LLMs hallucinate

Solution 3: Force LLMs to focus on **validated data**

MITIGATE

BIAS

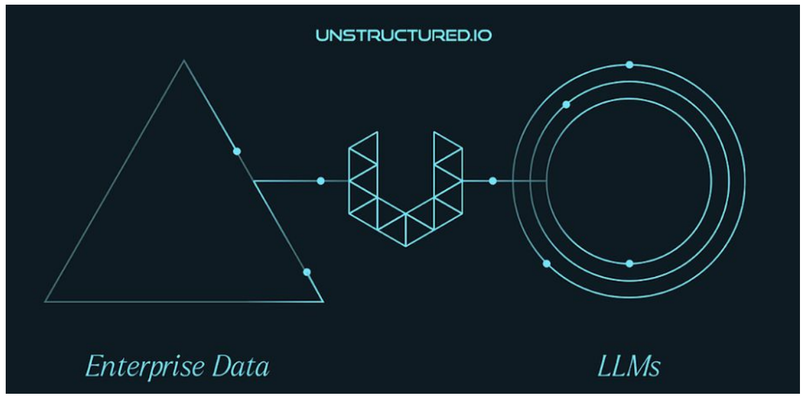
HALLUCINATIONS

BENGAL

High-Quality, Organizational-Specific Data for Safe LLM Use

With Gold-Standard, Organization Specific Data:

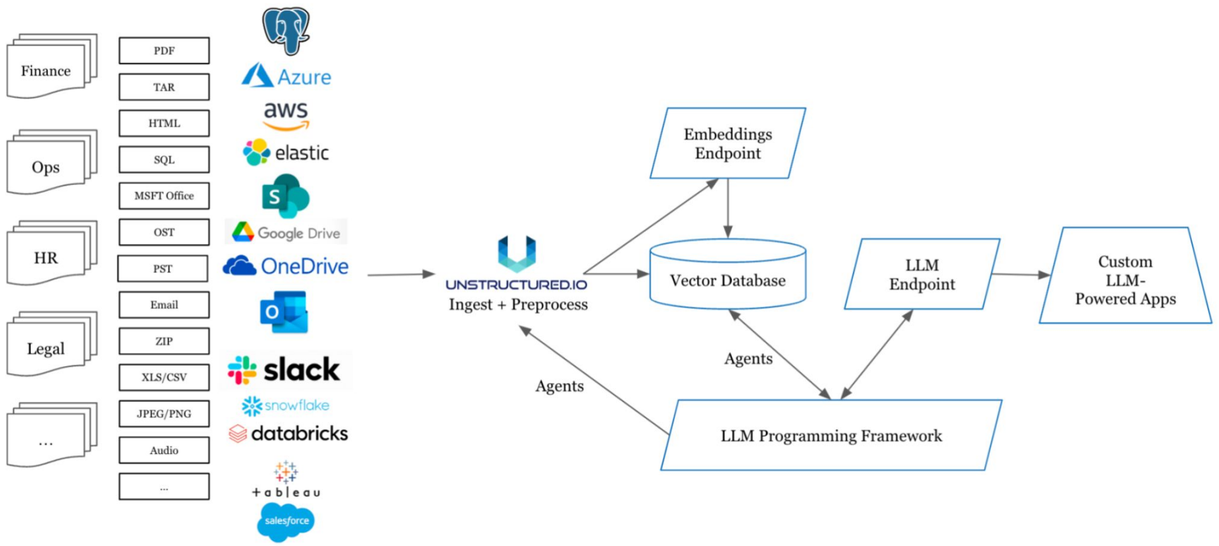
- Reduce Hallucinations/Erroneous Outputs
- Mitigate Bias
- Verify Sourcing
- Control Access
- Reduce Toxic Outputs



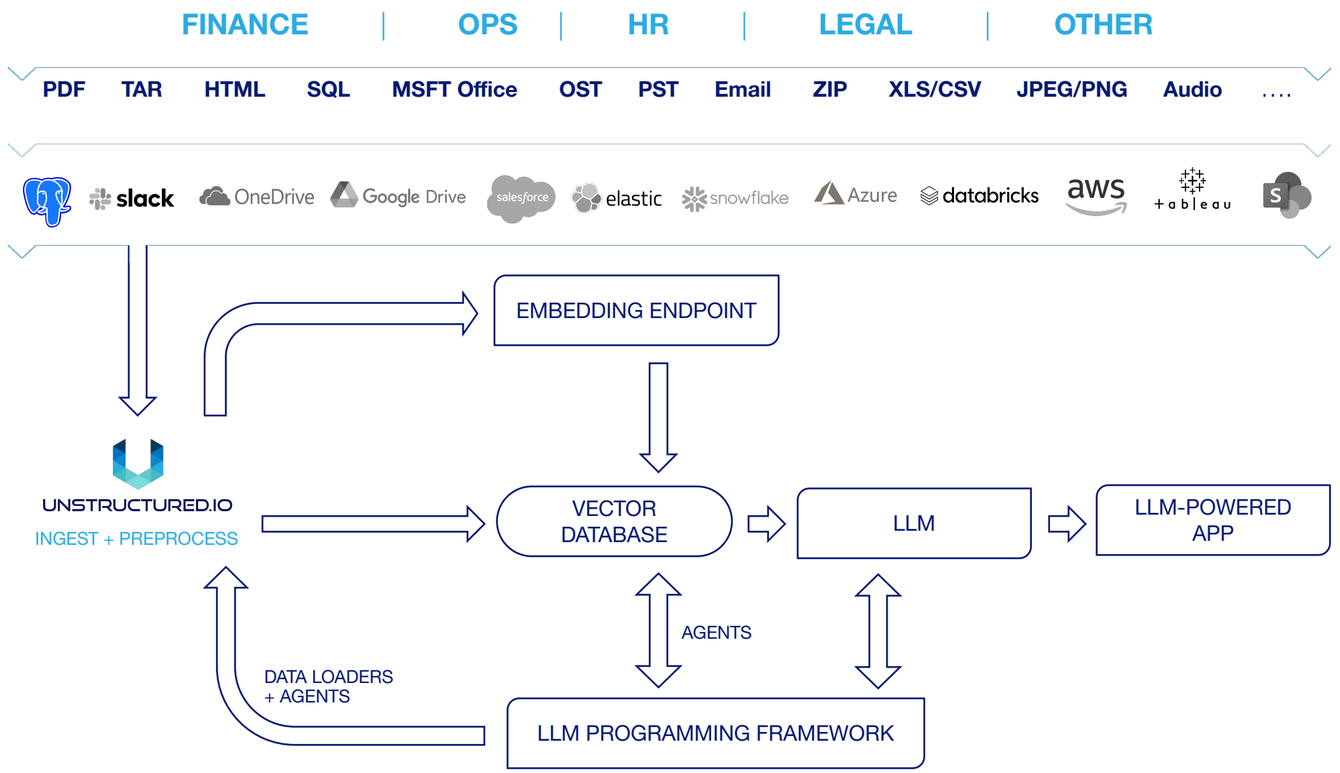
BENGAL

Emerging LLM Tech Stack: Retrieval Augmented Generation

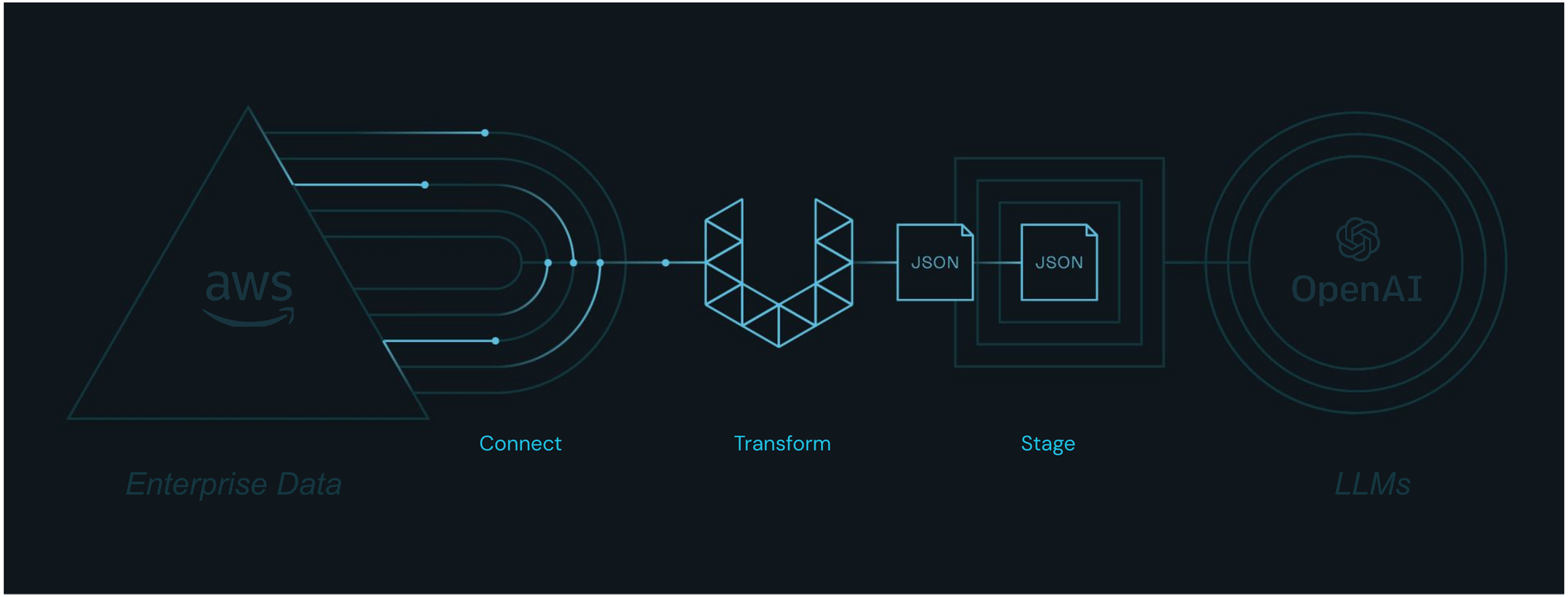
- **Retrieval Augmented Generation (RAG):** an AI framework that synergistically combines the capabilities of LLMs with an integrated information retrieval system, leveraging external databases to deliver more precise, contextually relevant, and up-to-date responses & limit hallucinations



Where Unstructured Fits In

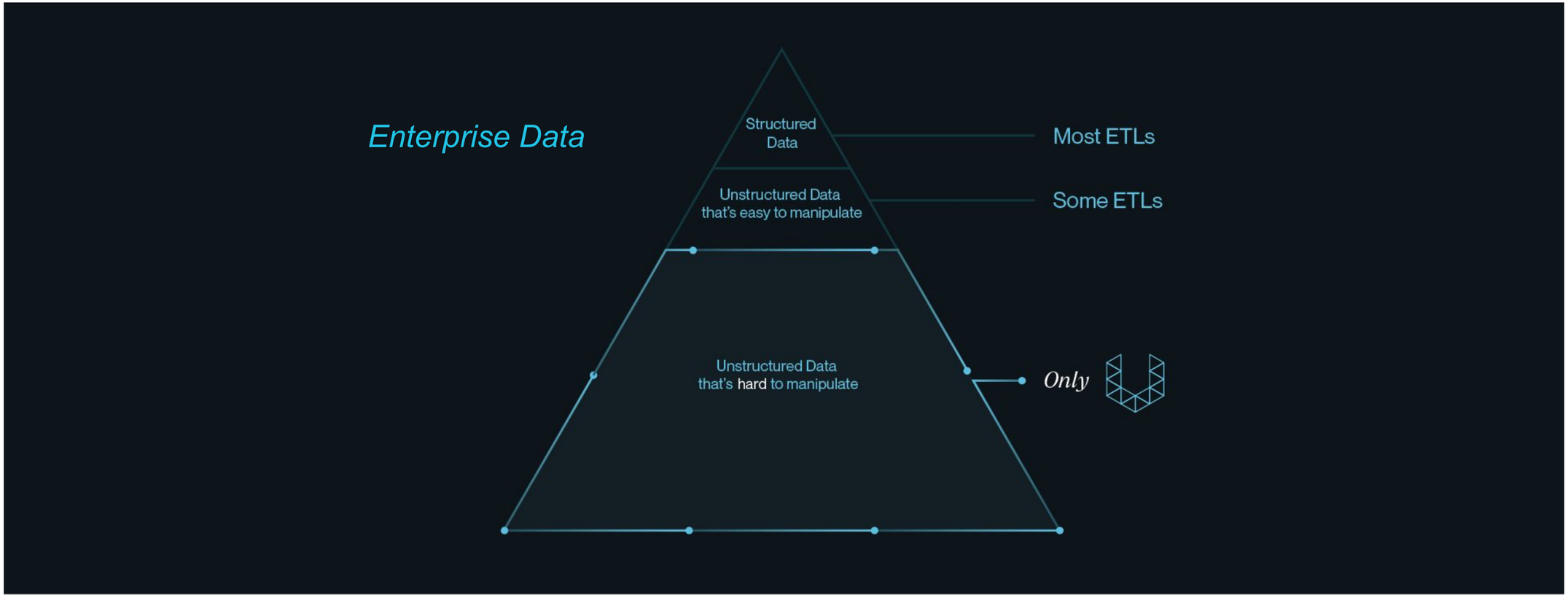


We connect natural language data to LLMs



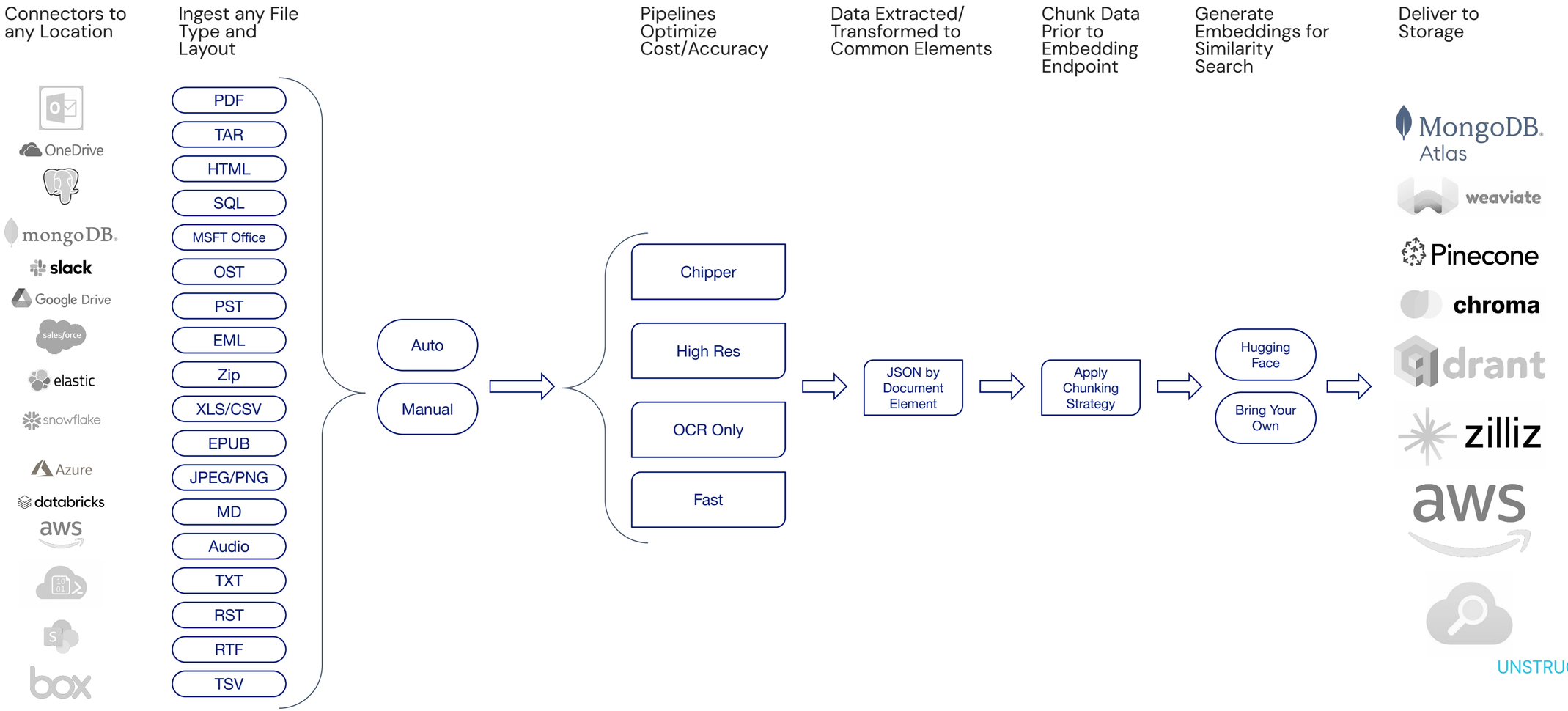
How We're Different

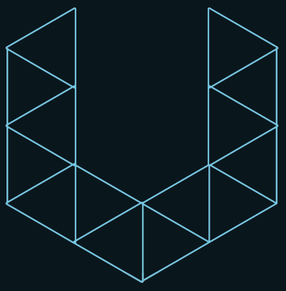
Any document. Any file type. Any layout.



In Detail

Logical Architecture:





UNSTRUCTURED