Beyond RAG – going from search to analytics on unstructured data

Mehul A. Shah, CEO

About Us

Make it easy to use GenAI for unstructured data analysis at scale

Founding team from AWS, Google, and Meta

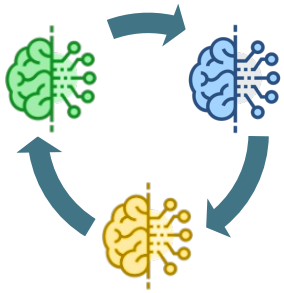
OpenSearch Software Foundation, founding member & TSC

Funded by top-tier investors: 8VC and Factory

Introduction

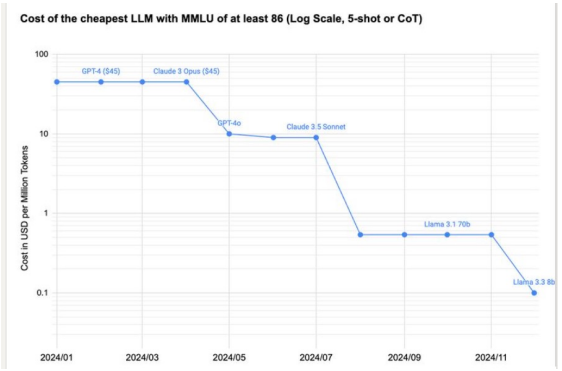
90% of enterprise data is unstructured and holds a gold mine of untapped information

For 30+ years, search has been our most effective tool.



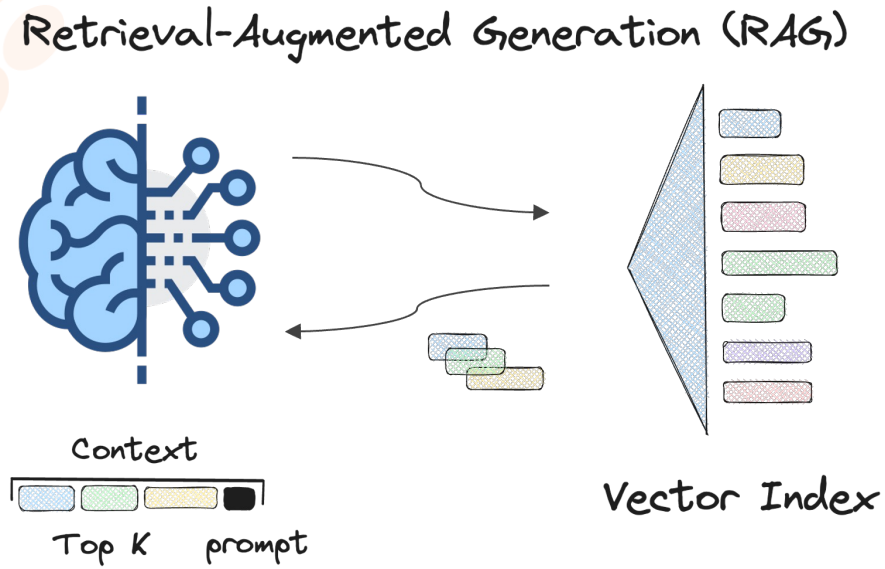
GenAI give SOTA results *without* an AI team

- Summarization
- Entity extraction
- Sentiment analysis



GenAI models are rapidly getting faster (10x/yr) and cheaper (50x-80x/yr)

RAG: Stick an LLM at the end of the search pipeline



- LLM is the new compute, and its context is like the new main memory.
- RAG applications do search and then pass the top-k results into an LLM
- Provides a natural language interface, but answers are limited by the context

Do we need to do more?

Enterprise use cases need to go beyond question answering

Financial Research

Financial Analyst

Get market trend insights from earnings calls, filings, and investment memos.

Summarize the yearly revenue growth and outlook of companies whose CEO recently changed?

Primary Market Research

Marketing Analyst

Analyze performance of marketing campaigns from thousands of hours of customer interviews.

Provide the top reasons PCPs prescribed medication X versus specialists?

Legal Research

Legal Analyst

Analyze thousands of legal judgments, filings, and memos.

Find all precedent of companies that violated the Section X rule and calculate the total sanctions levied.

Beyond RAG: "Deep Analytics" patterns for complex analyses across document collections

RAG

Hunt and peck

"What was the capital expenditure for GOOG in Q4 2024?"

Deep Analytics

Sweep and harvest

"From board decks and investment memos, summarize the capital outlays of top companies by ARR in the construction engineering market."

LLM-Extract ARR, market → Filter on construction engineering → Sort by ARR → LLM-Summarize Capital outlays

Integrate and synthesize

"For all the high-risk loan apps, lookup their market segment and key competitors."

Aryn is building a new category of AI-data systems ...

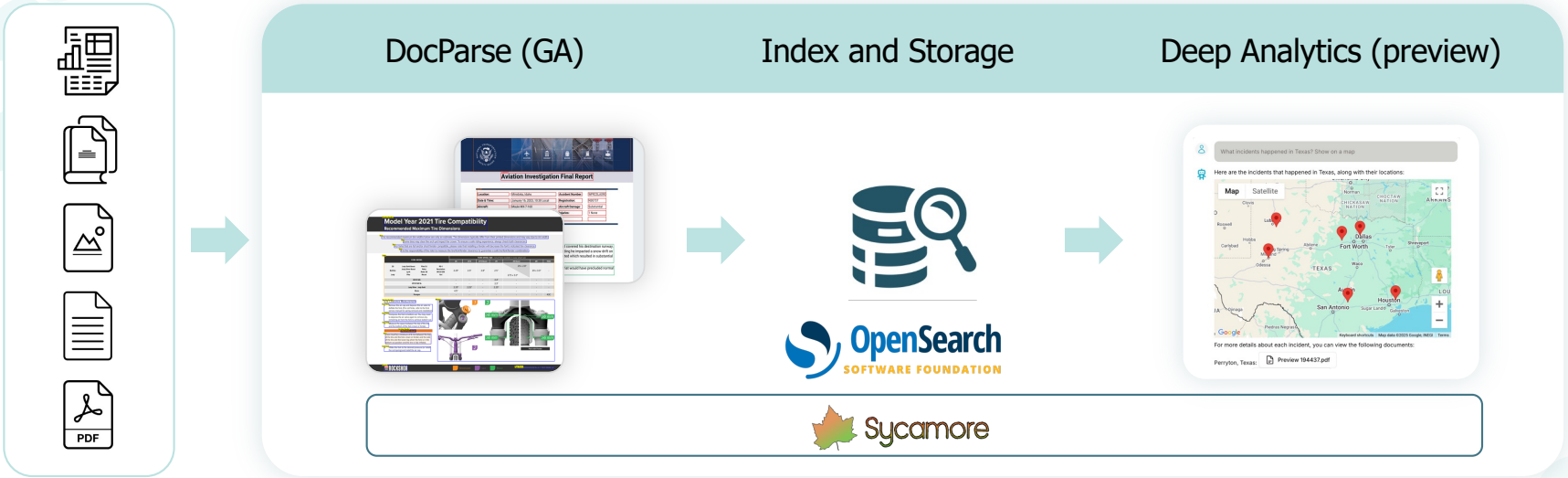
Agentic Unstructured Data Warehouse

Enterprise AI platform for "Deep Analytics" on complex documents and structured data

1. ELT - transformation
2. Ad-hoc analytics
3. Reporting

→ Now in public preview!

The Aryn Platform



Any format

AI-powered Document ETL

Extract metadata accurately at scale
- **3x** more accurate
- **5x** cheaper
- **10x** faster

Core distributed infrastructure

OpenSearch – index and store

Sycamore (Apache v2) – scalable AI-powered dataflow processing

Agentic reasoning on complex, unstructured data

Talk outline

Introduction

See it in action

Deep dive: Ingestion and Deep Analytics

Parting thoughts

Talk outline

Introduction

See it in action

Deep dive: Ingestion and Deep Analytics

Parting thoughts

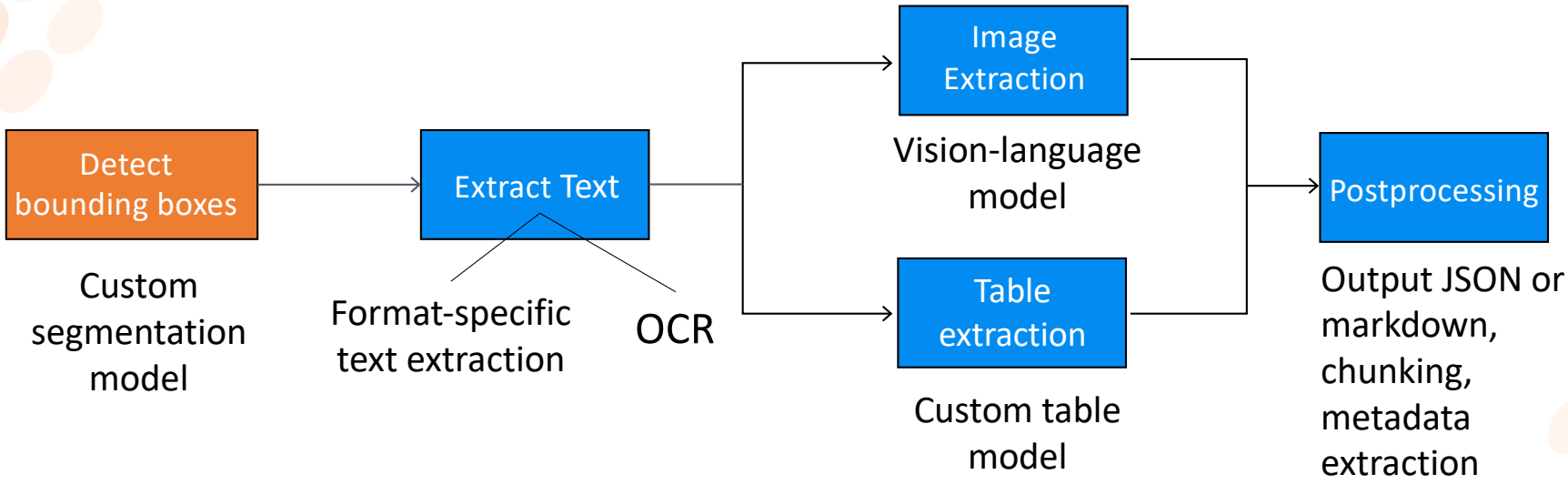
Key Challenges

1. Parse complex unstructured documents (pdfs, docx, ppts, etc.)

2. Build scalable and expressive ETL pipelines to enrich and index unstructured documents

3. Convert natural language tasks into execution plans, run them at scale, and make it easy to verify

Aryn DocParse: Decompose and Conquer



Uses custom segmentation and table model based on Deformable DETR architecture

580K+ Downloads on Hugging Face

High quality metadata extraction and planning give better results

Best in class segmentation and labeling

DocLayNet Competition	mAP	mAR	Aryn vs. others (mAP, mAR)
Aryn DocParse	**0.640**	**0.747**	1.0x, 1.0x
Amazon Textract	0.423	0.507	**1.5x, 1.5x better**
Unstructured	0.347	0.505	**1.8x, 1.5x better**
Azure Document Intelligence	0.266	0.457	**2.4x, 1.6x better**

Customer support assistant

Basic RAG	w/ alternative ETL	w/ Aryn
55% recall @10	81% recall @10	**97%** recall @10

FinanceBench

Basic RAG	w/ Aryn ETL	w/ Aryn planner
39% correct	61% correct	**77%** correct

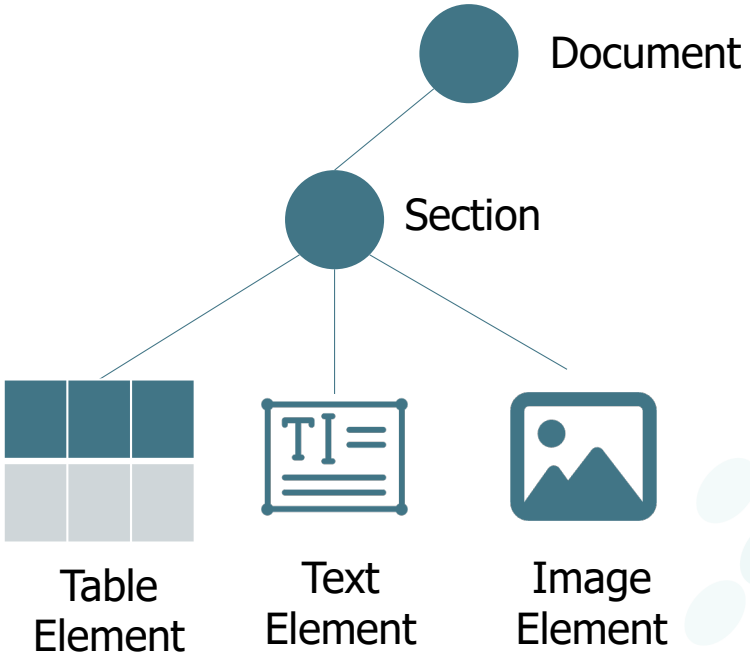
Sycamore

- Scalable agentic dataflow system for unstructured documents
- Developers write “Spark”-like programs to transform, enrich, and analyze their document collections in Python
- Open source (Apache-v2) license –

github

Document Model

- Documents are represented as trees
- Each node in the tree has arbitrary key-value properties
- The leaves of the tree are elements that correspond to pieces of the document
- Documents are grouped into DocSets which are distributed collections analogous to DataFrames



Sycamore Operators

Structured

Common dataflow operations

- queryDatabase
- map
- filter
- groupBy
- aggregate

Semantic

LLM-powered operations

- queryVectorDB
- llmFilter
- llmExtract
- llmGroupBy
- llmSummarize

Sample Sycamore Script

```
schema = {
  "us_state": "string",
  "probable_cause": "string",
  "weather_related": "bool"
}

extractor = OpenAIPropertyExtractor(
                    "gpt-4o", schema=schema)
ds = context.read.binary("/path/to/ntsb_data")
       .partition(DocParse())
       .llmExtract(extractor)

ds = ds.explode()
       .embed(OpenAIEmbedder("text-embedding-3-small"))
```

Declare the schema to extract

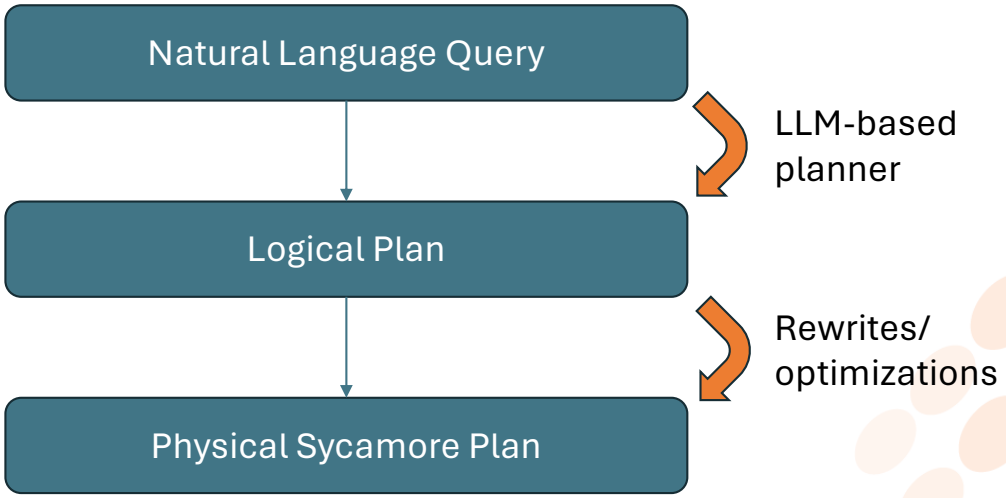
Use OpenAI to extract the fields for each document.

Compute vector embeddings for the text portions of the docs.

Deep Analytics Overview

- Luna – Query planner takes natural language and converts it into an execution plan.

- Sycamore is used to execute the queries.
  - Learned from Spark that ETL and analytics share many operators!



Luna Planner Prompt

Preamble	Task Description and instructions	`You are an agent that translates a question into a query plan…`
Schema	Information about metadata fields	`{“name:” “location”, “type”: “str”, “examples”: [“San Francisco, CA”, …], “description”: “Location where incident occurred”}`
Operators	Logical operators and their inputs and outputs	`LlmFilter. LlmFilter uses a Large Language Model (LLM) to filter database records based on the value of a field…`
Examples	Few-shot examples of questions to plans	Which incidents occurred in CA when then wind was > 4 knots ↳ `QueryDatabase` → `LLMFilter`
Question	Actual question submitted by the user.	"Which three aircraft types were involved in the most accidents?"

Sample Benchmark

- We built a small benchmark of 30 questions based on 100 NTSB incident reports.
- Questions combine metadata lookup and LLM-based extraction.
- Sample questions
  - How many incidents were there by state?
  - What fraction of incidents that resulted in substantial damage were due to engine problems?
  - Which incidents occurred in July involving birds?

- **Early Results**: 20/30 Correct
- Common errors:
  - **Counting errors (6 cases)**

    The LLM makes off-by-one errors due to things like incidents involving two aircraft.
  - **Filter errors (3 cases)**

    The LLM is too generous about whether a given document should pass the filter.
  - **Query interpretation (1 case)**

    The LLM misinterpreted “aircraft manufacturer” as “aircraft type”
- Can be fixed with better prompting and human feedback.

Talk outline

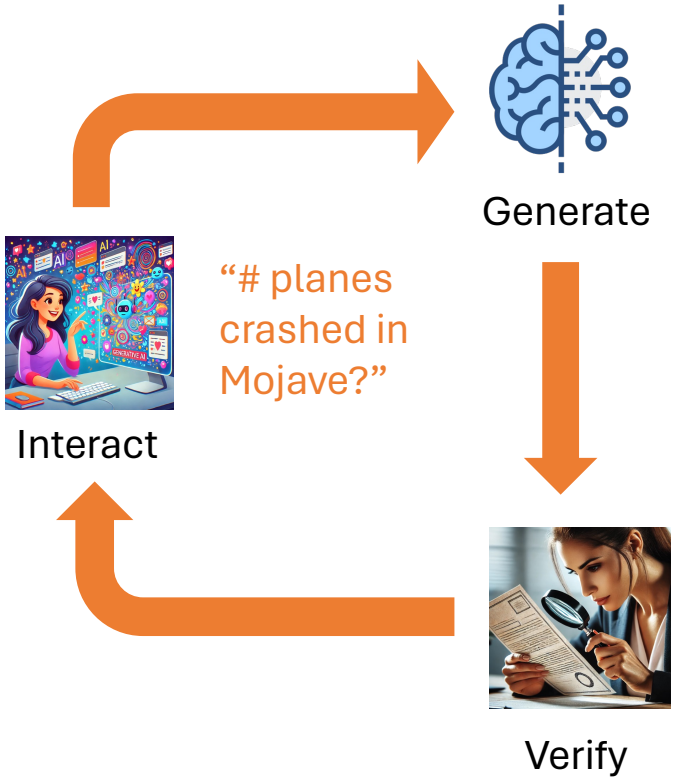
Introduction

See it in action

Deep dive: Ingestion and Deep Analytics

Parting thoughts

Explainability and iteration are essential for Deep Analytics



"AI are incredible guessing machines"

co-pilot – code

Luna – query plans

"21 planes"

"AI output independently verified"

co-pilot – humans, testing

Luna – inspect plans and data, follow-up builds trust

Conclusion

GenAI unlocks a lot of batch AI use cases.

- Users want so much more than RAG.

Deep Analytics uses agentic reasoning and a mix of DB-style structured and LLM semantic operations.

- There's room for optimizations.

Verifiability is critical; just scratched the surface.

- Human in the loop is needed, there's much room for better explainability.

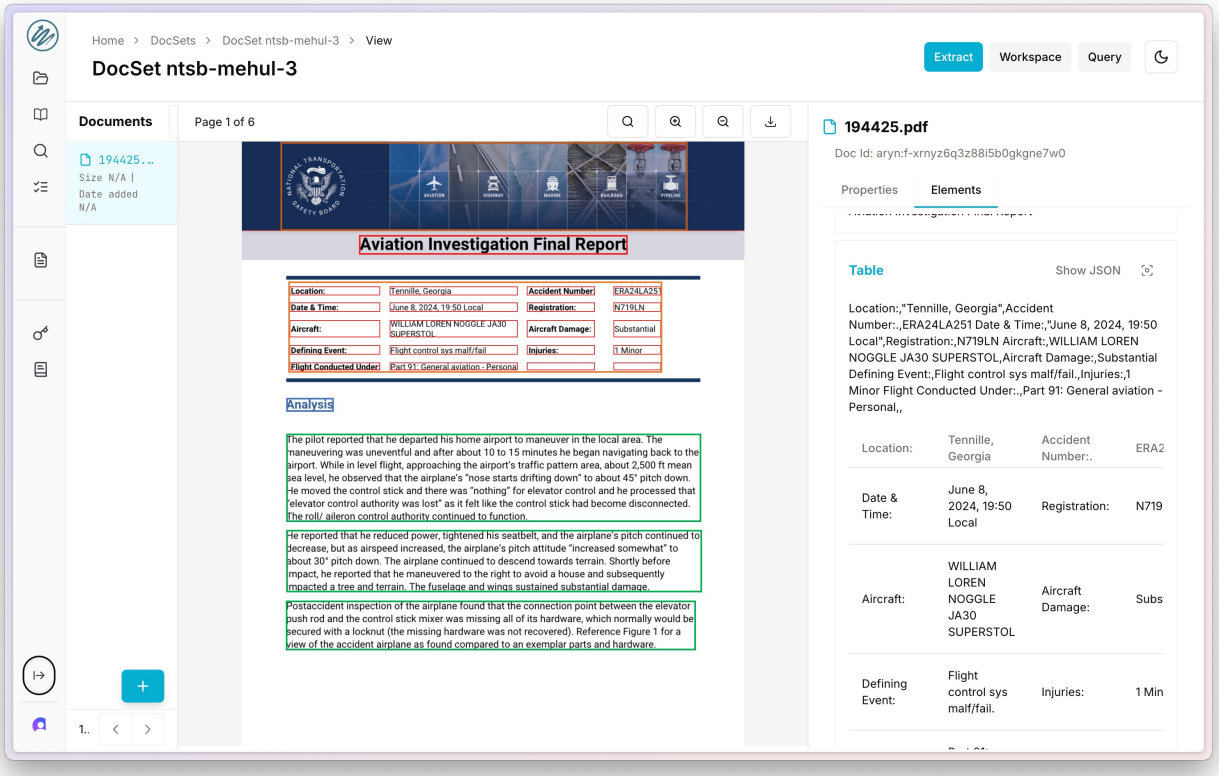
Try it out!

Aryn.ai

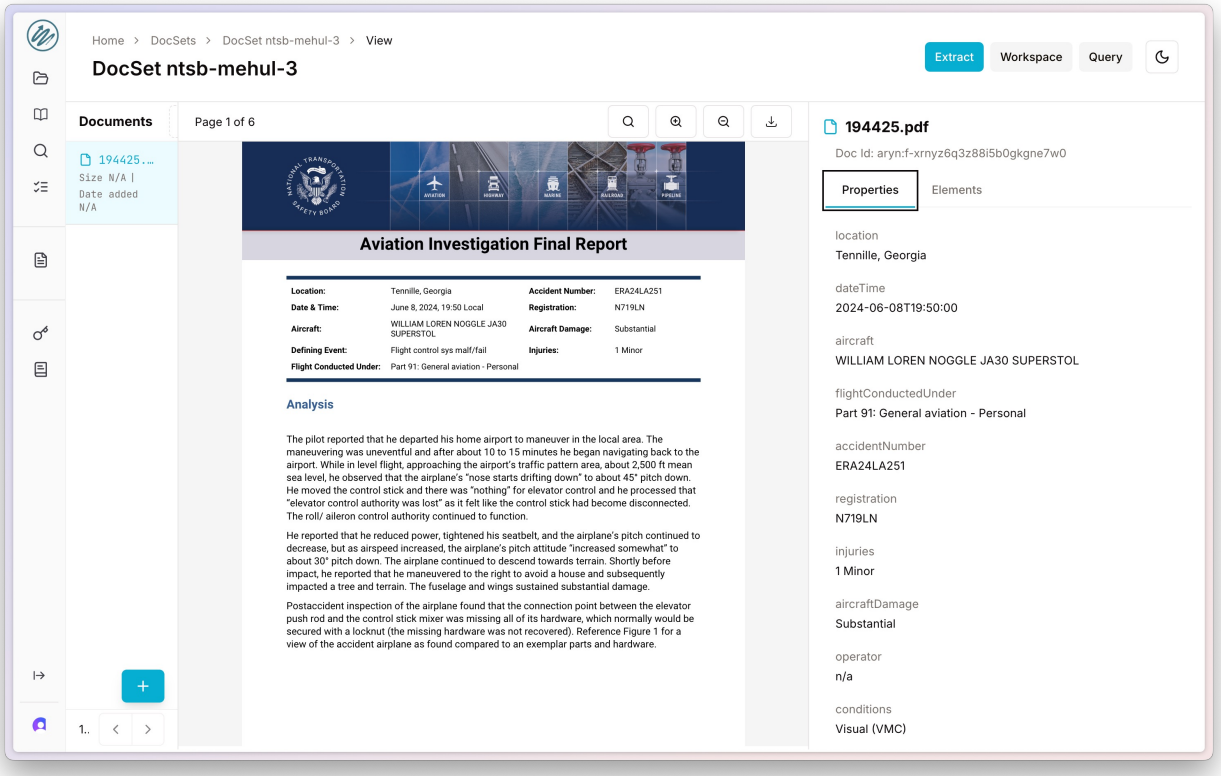
github

Additional slides

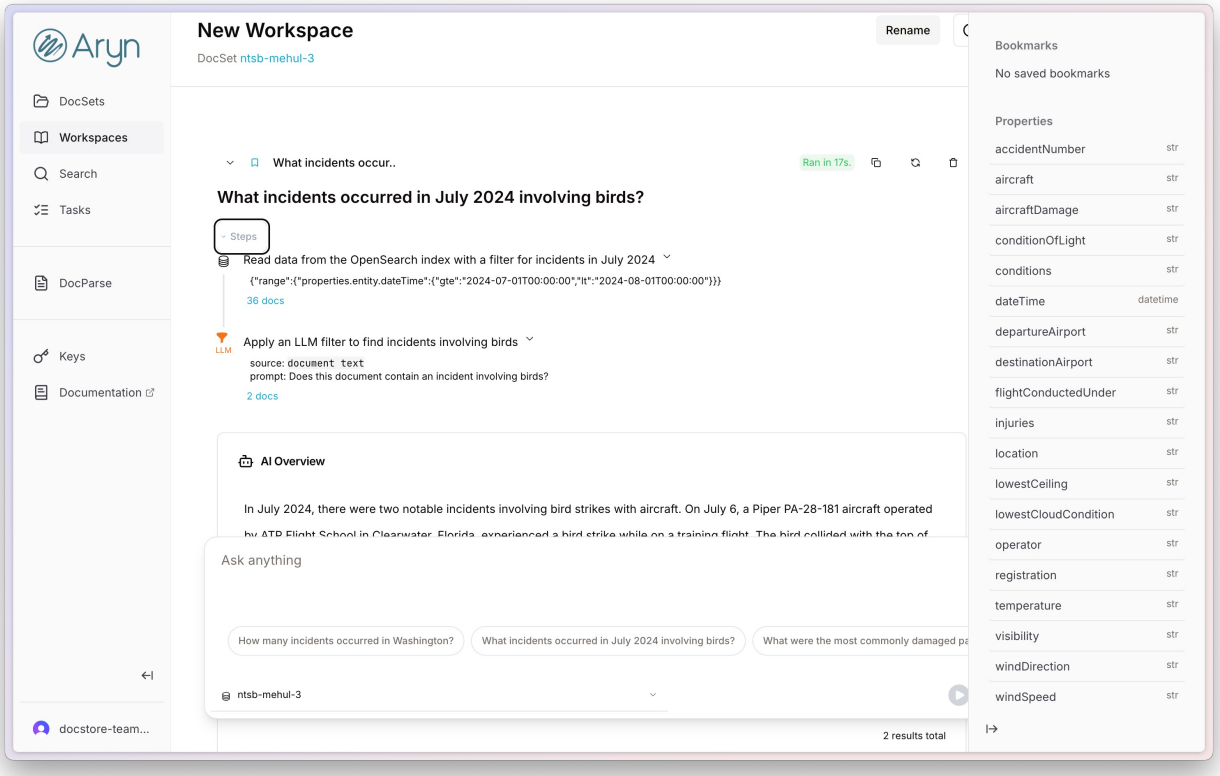
Aryn DocParse



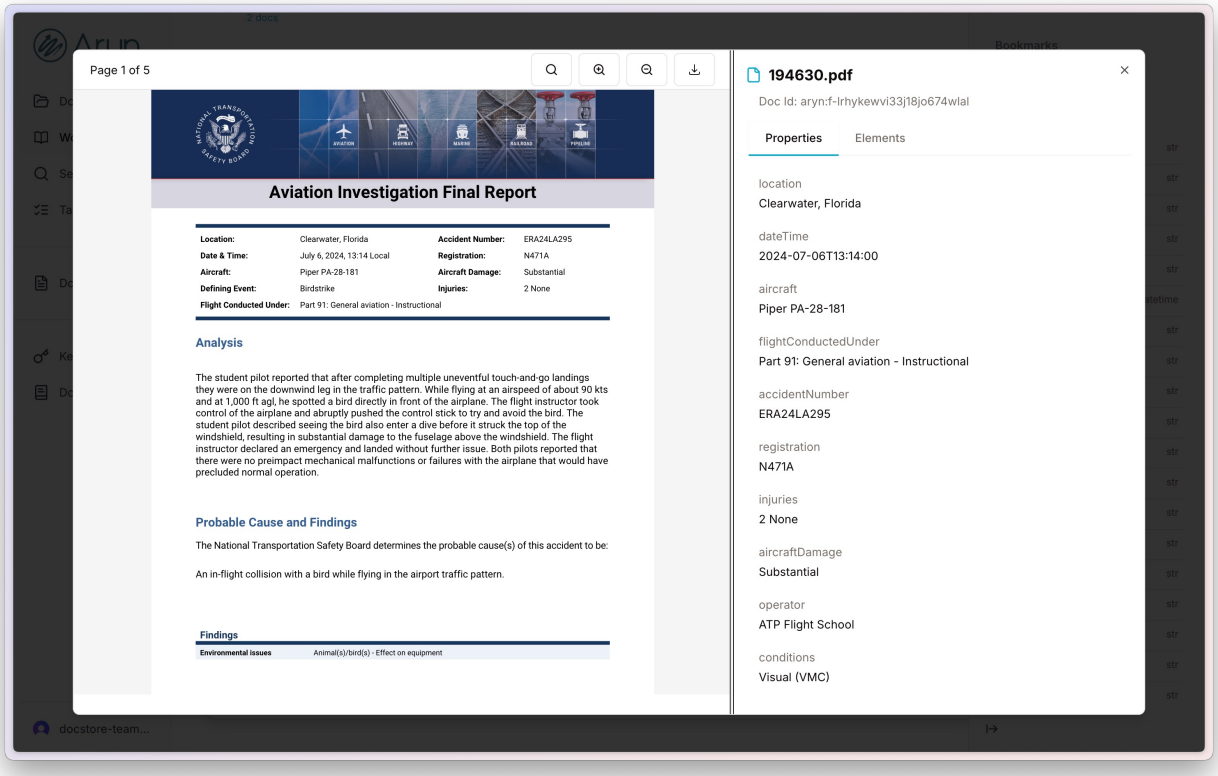
Aryn DocParse



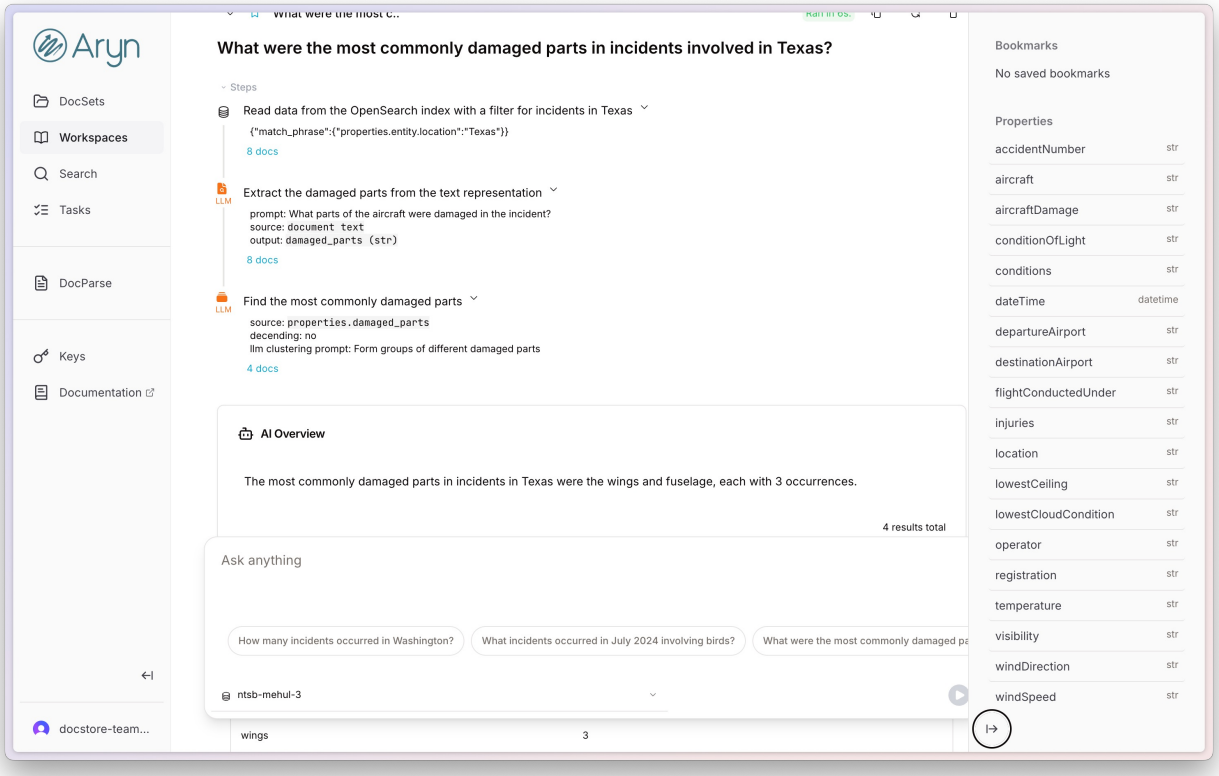
Deep Analytics



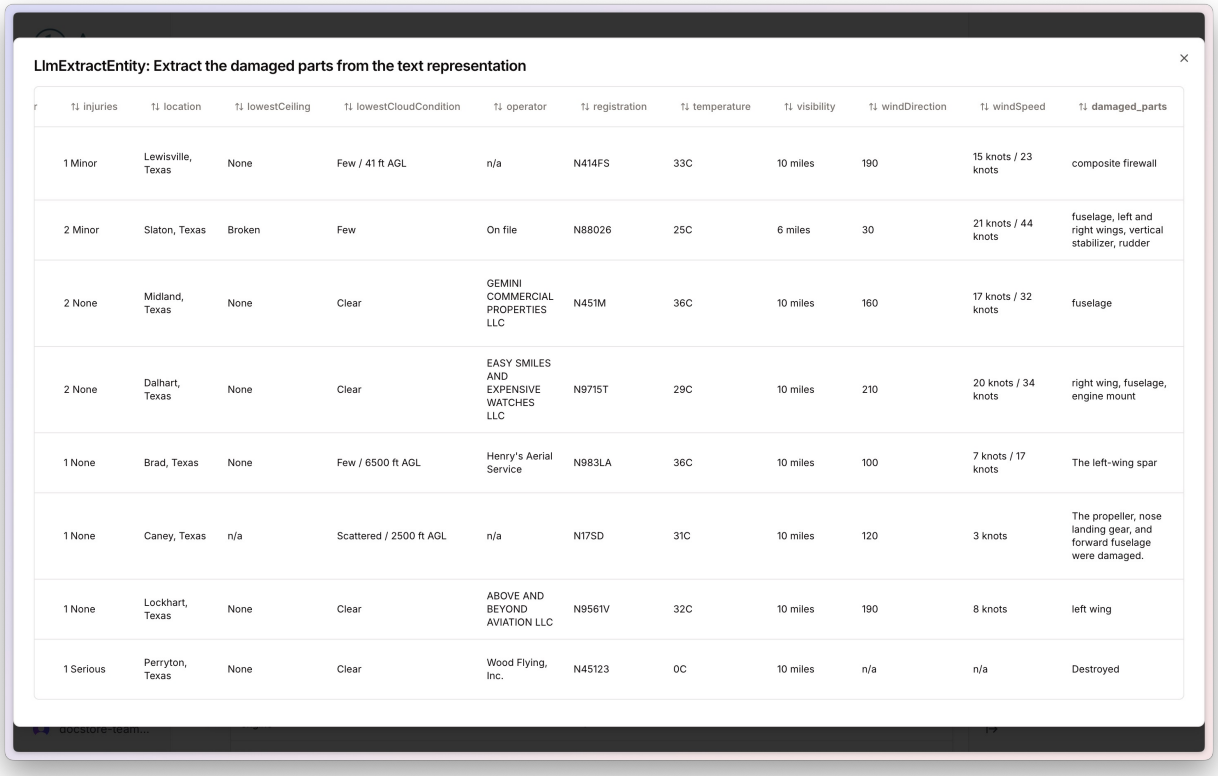
Deep Analytics



Deep Analytics



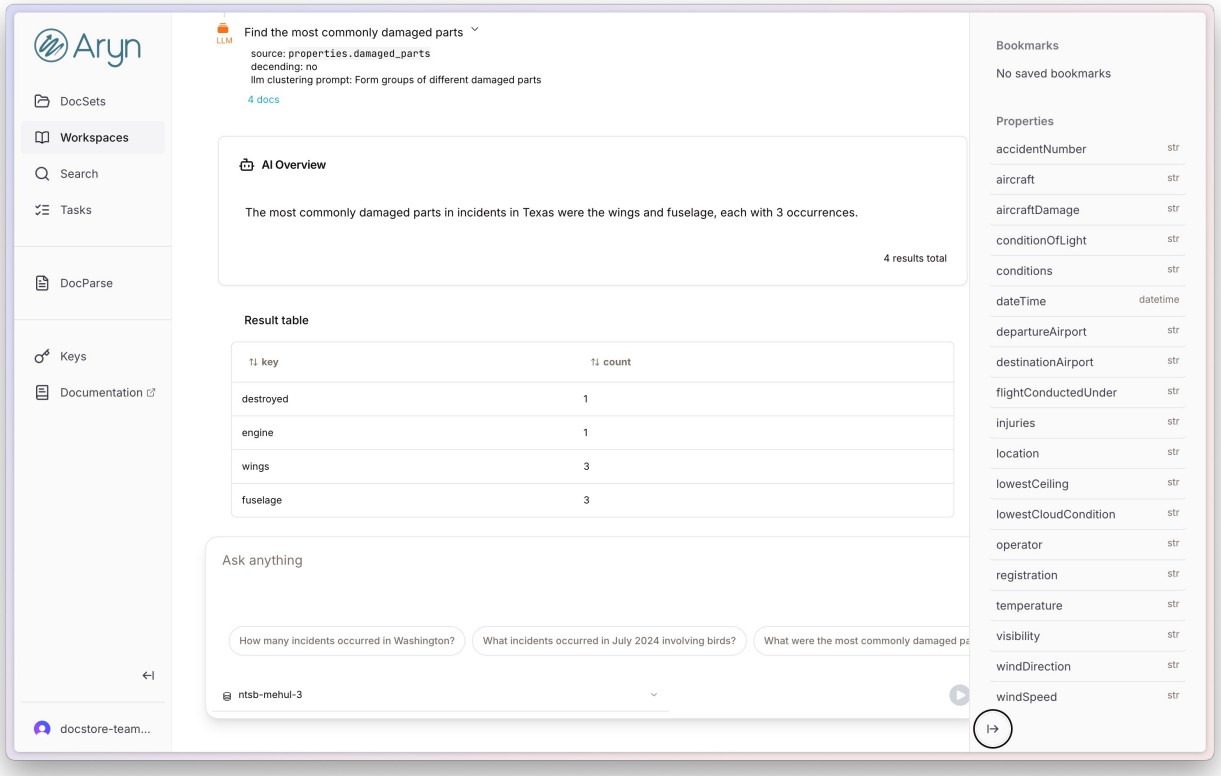
Deep Analytics



LlmExtractEntity: Extract the damaged parts from the text representation

injuries	location	lowestCeiling	lowestCloudCondition	operator	registration	temperature	visibility	windDirection	windSpeed	damaged_parts
1 Minor	Lewisville, Texas	None	Few / 41 ft AGL	n/a	N414FS	33C	10 miles	190	15 knots / 23 knots	composite firewall
2 Minor	Slaton, Texas	Broken	Few	On file	N88026	25C	6 miles	30	21 knots / 44 knots	fuselage, left and right wings, vertical stabilizer, rudder
2 None	Midland, Texas	None	Clear	GEMINI COMMERCIAL PROPERTIES LLC	N451M	36C	10 miles	160	17 knots / 32 knots	fuselage
2 None	Dalhart, Texas	None	Clear	EASY SMILES AND EXPENSIVE WATCHES LLC	N9715T	29C	10 miles	210	20 knots / 34 knots	right wing, fuselage, engine mount
1 None	Brad, Texas	None	Few / 6500 ft AGL	Henry's Aerial Service	N983LA	36C	10 miles	100	7 knots / 17 knots	The left-wing spar
1 None	Caney, Texas	n/a	Scattered / 2500 ft AGL	n/a	N17SD	31C	10 miles	120	3 knots	The propeller, nose landing gear, and forward fuselage were damaged.
1 None	Lockhart, Texas	None	Clear	ABOVE AND BEYOND AVIATION LLC	N9561V	32C	10 miles	190	8 knots	left wing
1 Serious	Perryton, Texas	None	Clear	Wood Flying, Inc.	N45123	0C	10 miles	n/a	n/a	Destroyed

Deep Analytics



We make it **easy** to unleash the power of AI on all your **unstructured data** and get **accuracy at scale**

Innumerable use cases for unstructured data analytics with tangible value

Enterprise GenAI platforms

Legal memos: discover precedent

Industry research: investment thesis

Offering memos: real estate investment

Technical docs: customer support

Interview transcripts: primary market research

Document automation

KYC onboarding: risk assessment

Insurance claims: detect fraud

Invoice and receipts: cost and error reduction

Clinical trial applications: compliance