

Beyond RAG – going from search to analytics on unstructured data

Mehul A. Shah, CEO



About Us

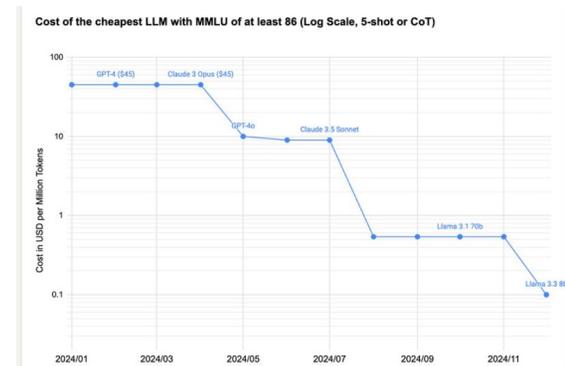
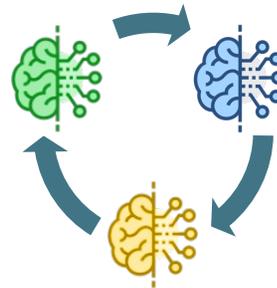
Make it easy to use GenAI for unstructured data analysis at scale

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.

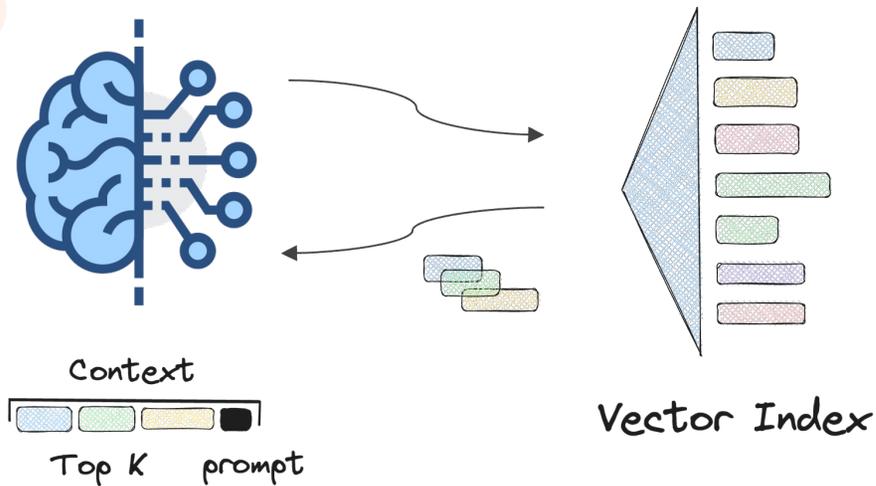
GenAI give SOTA results *without* an AI team

- 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

Retrieval-Augmented Generation (RAG)



- 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

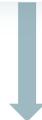
Hunt and peck



“What was the capital expenditure for GOOG in Q4 2024?”

Deep Analytics

Sweep and harvest

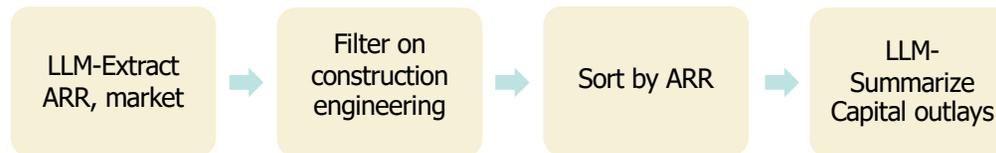


“From board decks and investment memos, summarize the capital outlays of top companies by ARR in the construction engineering market.”

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

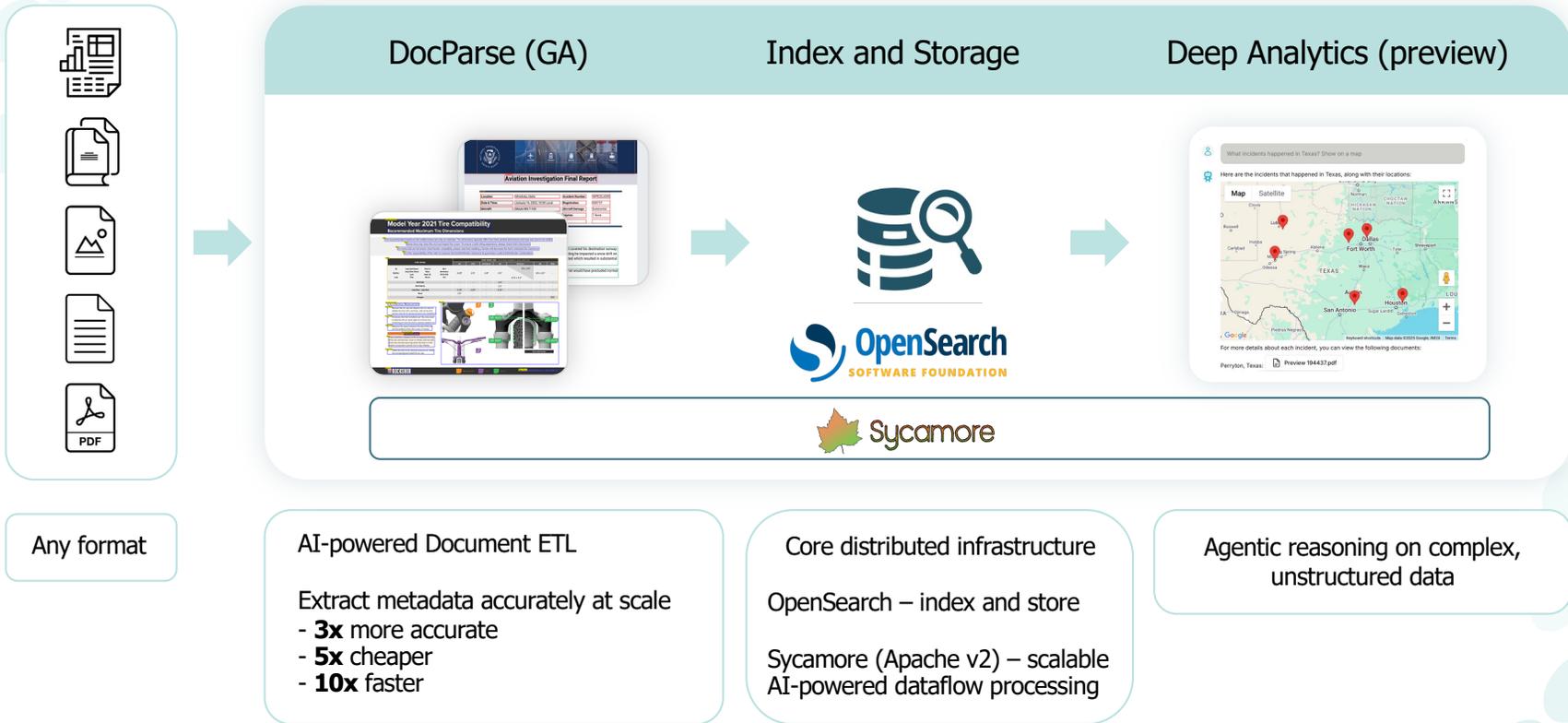
Enterprise AI platform for “Deep Analytics” on
complex documents and structured data

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

→ Now in public
preview!



The Aryn Platform





Talk outline

Introduction

See it in action

Deep dive: Ingestion and Deep Analytics

Parting thoughts



Talk outline

Introduction

See it in action

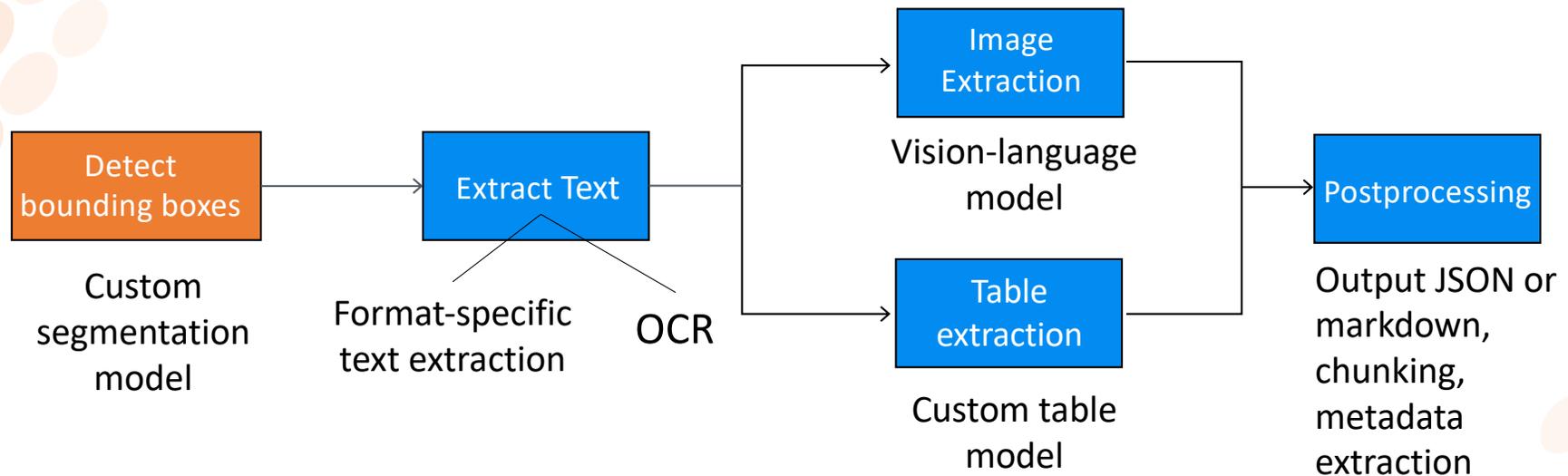
Deep dive: Ingestion and Deep Analytics

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

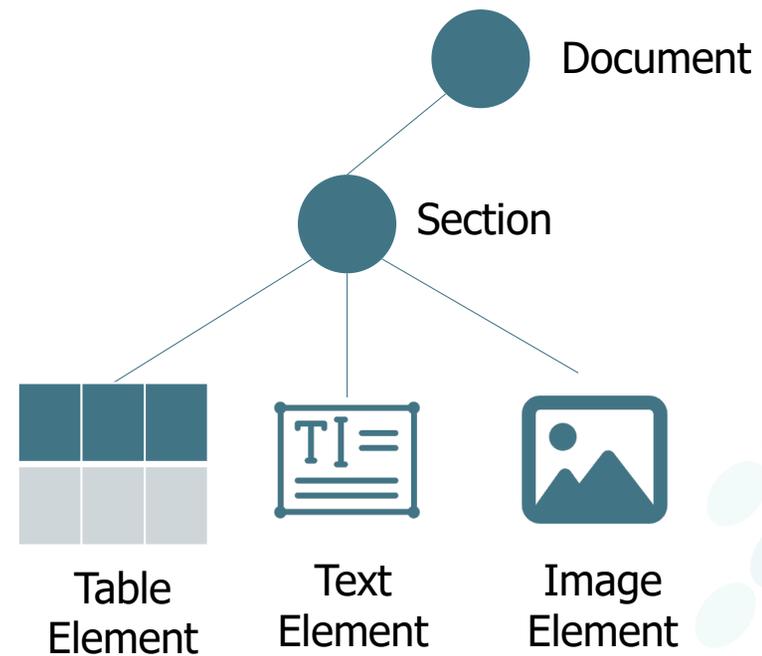
- Scalable agentic dataflow system for unstructured documents
- Developers write “Spark”-like programs to transform, enrich, and analyze their document collections in Python
- Open source (Apache-v2) license –



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"  
}
```

Declare the schema to extract

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

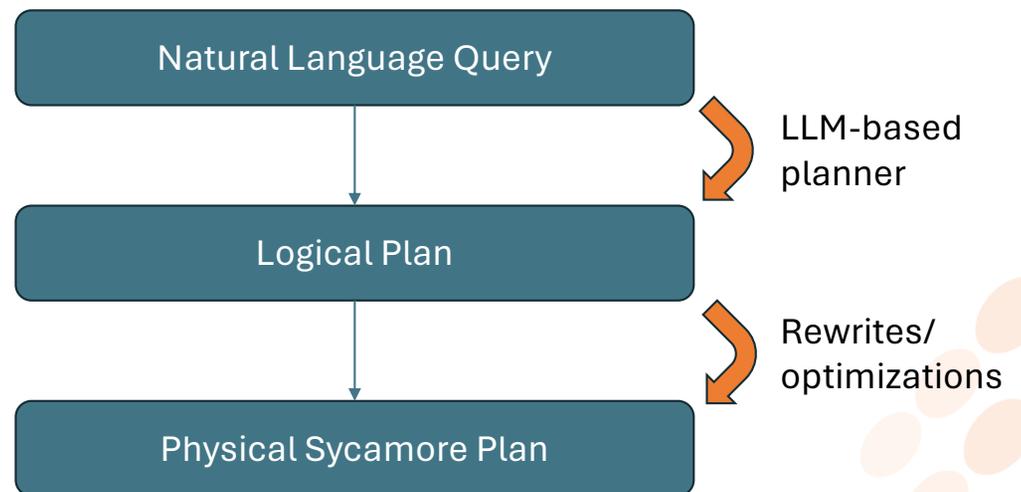
Use OpenAI to extract the fields for each document.

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

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	<pre>{ "name": "location", "type": "str", "examples": ["San Francisco, CA", ...], "description": "Location where incident occurred" }</pre>
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	<p>Which incidents occurred in CA when then wind was > 4 knots</p> <pre>└─ QueryDatabase → LLMFilter</pre>
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

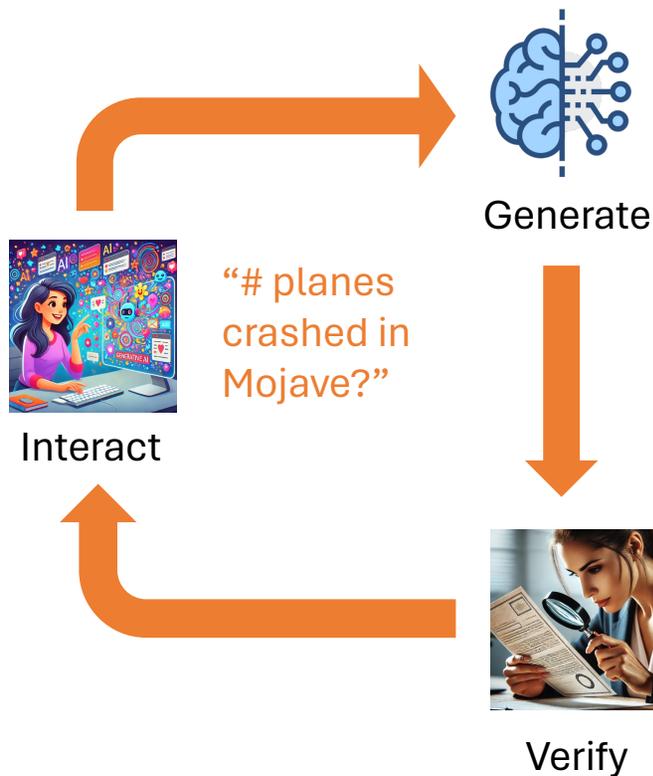
Introduction

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

Home > DocSets > DocSet ntsb-mehul-3 > View
Extract Workspace Query

DocSet ntsb-mehul-3

Documents Page 1 of 6

194425...
Size N/A |
Date added N/A

Aviation Investigation Final Report

Location:	Tennille, Georgia	Accident Number:	ERA24LA25
Date & Time:	June 8, 2024, 19:50 Local	Registration:	N719LN
Aircraft:	WILLIAM LOREN NOGGLE JA30 SUPERSTOL	Aircraft Damage:	Substantial
Defining Event:	Flight control sys malffail	Injuries:	1 Minor
Flight Conducted Under:	Part 91: General aviation - Personal		

Analysis

The pilot reported that he departed his home airport to maneuver in the local area. The maneuvering was uneventful and after about 10 to 15 minutes he began navigating back to the airport. While in level flight, approaching the airport's traffic pattern area, about 2,500 ft mean sea level, he observed that the airplane's "nose starts drifting down" to about 45° pitch down. He moved the control stick and there was "nothing" for elevator control and he processed that "elevator control authority was lost" as it felt like the control stick had become disconnected. The roll/aileron control authority continued to function.

He reported that he reduced power, tightened his seatbelt, and the airplane's pitch continued to decrease, but as airspeed increased, the airplane's pitch attitude "increased somewhat" to about 30° pitch down. The airplane continued to descend towards terrain. Shortly before impact, he reported that he maneuvered to the right to avoid a house and subsequently impacted a tree and terrain. The fuselage and wings sustained substantial damage.

Postaccident inspection of the airplane found that the connection point between the elevator push rod and the control stick mixer was missing all of its hardware, which normally would be secured with a locknut (the missing hardware was not recovered). Reference Figure 1 for a view of the accident airplane as found compared to an exemplar parts and hardware.

194425.pdf

Doc Id: aryn:f-xrnyz6q3z88i5b0gkgn7w0

Properties Elements

Table Show JSON

Location: "Tennille, Georgia", Accident Number: ERA24LA251 Date & Time: "June 8, 2024, 19:50 Local", Registration: N719LN Aircraft: WILLIAM LOREN NOGGLE JA30 SUPERSTOL, Aircraft Damage: Substantial Defining Event: Flight control sys malffail, Injuries: 1 Minor Flight Conducted Under: Part 91: General aviation - Personal,

Location:	Tennille, Georgia	Accident Number:	ERA2
Date & Time:	June 8, 2024, 19:50 Local	Registration:	N719
Aircraft:	WILLIAM LOREN NOGGLE JA30 SUPERSTOL	Aircraft Damage:	Subs
Defining Event:	Flight control sys malffail.	Injuries:	1 Min

Aryn DocParse

Home > DocSets > DocSet ntsb-mehul-3 > View
Extract Workspace Query

Documents

194425...
Size N/A |
Date added
N/A

Page 1 of 6



Aviation Investigation Final Report

Location:	Tennille, Georgia	Accident Number:	ERA24LA251
Date & Time:	June 8, 2024, 19:50 Local	Registration:	N719LN
Aircraft:	WILLIAM LOREN NOGGLE JA30 SUPERSTOL	Aircraft Damage:	Substantial
Defining Event:	Flight control sys malff/fail	Injuries:	1 Minor
Flight Conducted Under:	Part 91: General aviation - Personal		

Analysis

The pilot reported that he departed his home airport to maneuver in the local area. The maneuvering was uneventful and after about 10 to 15 minutes he began navigating back to the airport. While in level flight, approaching the airport's traffic pattern area, about 2,500 ft mean sea level, he observed that the airplane's "nose starts drifting down" to about 45° pitch down. He moved the control stick and there was "nothing" for elevator control and he processed that "elevator control authority was lost" as it felt like the control stick had become disconnected. The roll/ aileron control authority continued to function.

He reported that he reduced power, tightened his seatbelt, and the airplane's pitch continued to decrease, but as airspeed increased, the airplane's pitch attitude "increased somewhat" to about 30° pitch down. The airplane continued to descend towards terrain. Shortly before impact, he reported that he maneuvered to the right to avoid a house and subsequently impacted a tree and terrain. The fuselage and wings sustained substantial damage.

Postaccident inspection of the airplane found that the connection point between the elevator push rod and the control stick mixer was missing all of its hardware, which normally would be secured with a locknut (the missing hardware was not recovered). Reference Figure 1 for a view of the accident airplane as found compared to an exemplar parts and hardware.

194425.pdf

Doc Id: aryn:f-xrnyz6q3z88i5b0gkne7w0

Properties Elements

location
Tennille, Georgia

dateTime
2024-06-08T19:50:00

aircraft
WILLIAM LOREN NOGGLE JA30 SUPERSTOL

flightConductedUnder
Part 91: General aviation - Personal

accidentNumber
ERA24LA251

registration
N719LN

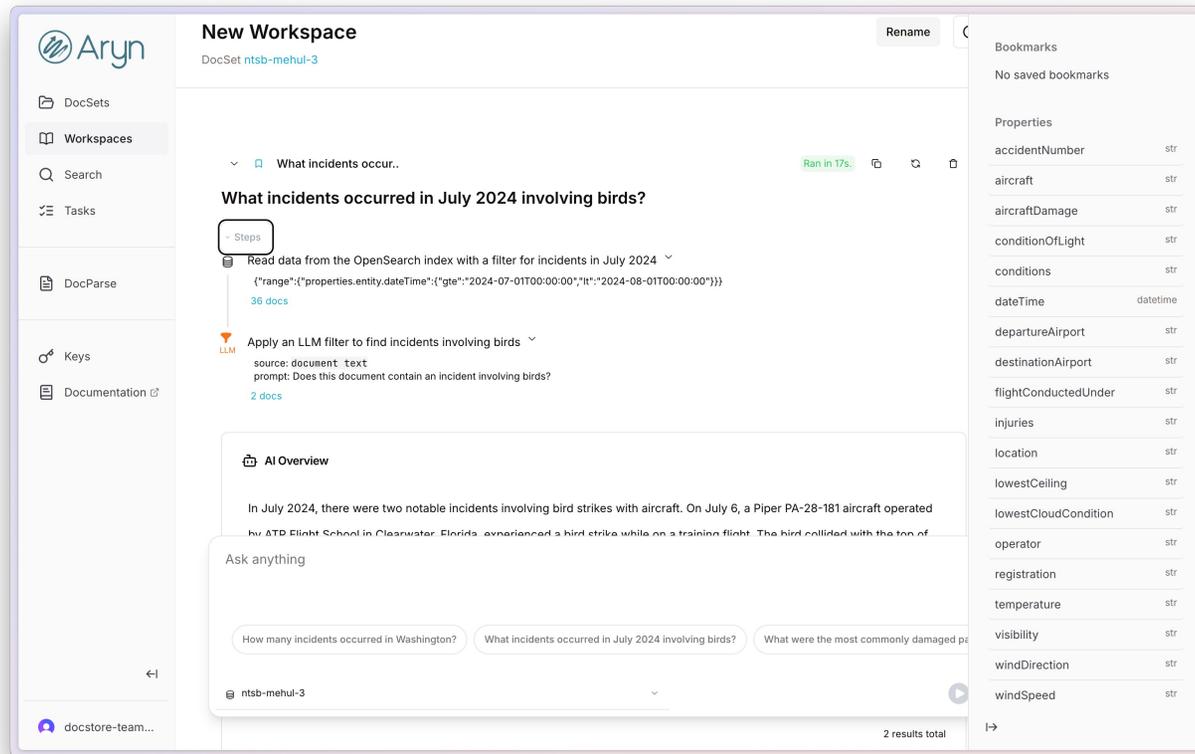
injuries
1 Minor

aircraftDamage
Substantial

operator
n/a

conditions
Visual (VMC)

Deep Analytics



New Workspace DocSet ntsb-mehul-3

What incidents occur.. Ran in 17s.

What incidents occurred in July 2024 involving birds?

Steps

- Read data from the OpenSearch index with a filter for incidents in July 2024
`{ "range": { "properties.entity.dateTime": { "gte": "2024-07-01T00:00:00", "lt": "2024-08-01T00:00:00" } } }`
36 docs
- Apply an LLM filter to find incidents involving birds
source: document_text
prompt: Does this document contain an incident involving birds?
2 docs

AI Overview

In July 2024, there were two notable incidents involving bird strikes with aircraft. On July 6, a Piper PA-28-181 aircraft operated by ATP Flight School in Clearwater, Florida, experienced a bird strike while on a training flight. The bird collided with the top of

Ask anything

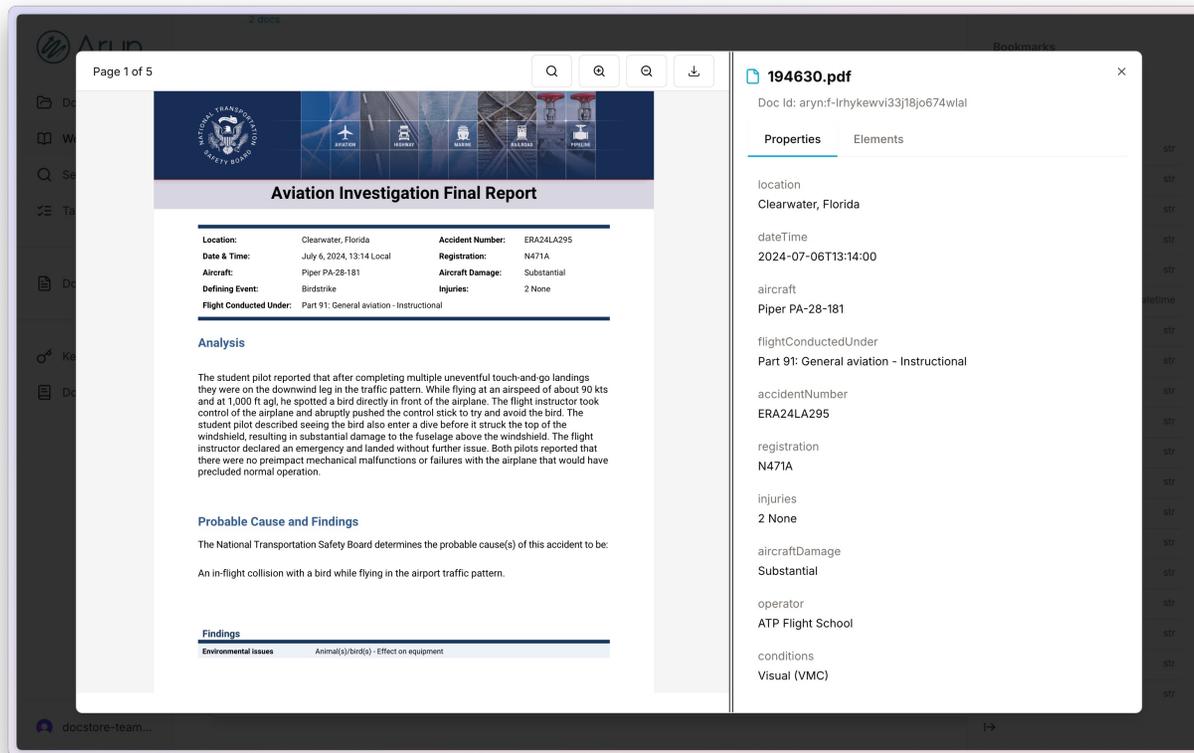
How many incidents occurred in Washington? What incidents occurred in July 2024 involving birds? What were the most commonly damaged parts?

nts-mehul-3

2 results total

Properties	
accidentNumber	str
aircraft	str
aircraftDamage	str
conditionOfLight	str
conditions	str
dateTime	datetime
departureAirport	str
destinationAirport	str
flightConductedUnder	str
injuries	str
location	str
lowestCeiling	str
lowestCloudCondition	str
operator	str
registration	str
temperature	str
visibility	str
windDirection	str
windSpeed	str

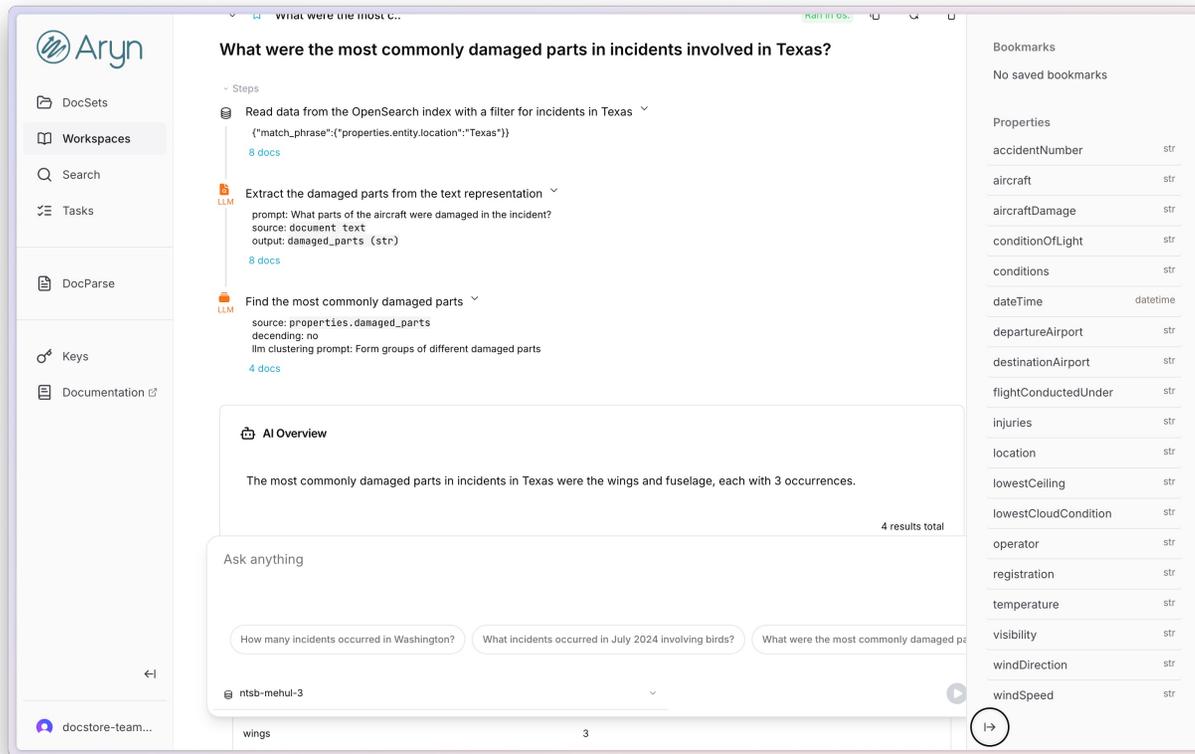
Deep Analytics



The screenshot displays a document viewer interface. The main document is titled "Aviation Investigation Final Report" and is the first page of a five-page document. The report details an accident that occurred in Clearwater, Florida, on July 6, 2024, involving a Piper PA-28-181 aircraft. The accident number is ERA24LA295, and the registration is N471A. The defining event was a birdstrike, resulting in substantial damage to the aircraft and no injuries. The probable cause was an in-flight collision with a bird while flying in the airport traffic pattern. The findings section notes environmental issues related to animal(s)/bird(s) and their effect on equipment.

On the right side of the viewer, a metadata panel for the PDF file "194630.pdf" is visible. It lists various properties such as location (Clearwater, Florida), date and time (2024-07-06T13:14:00), aircraft (Piper PA-28-181), flight conducted under (Part 91: General aviation - Instructional), accident number (ERA24LA295), registration (N471A), injuries (2 None), aircraft damage (Substantial), operator (ATP Flight School), and conditions (Visual (VMC)).

Deep Analytics



What were the most commonly damaged parts in incidents involved in Texas?

Steps

- Read data from the OpenSearch index with a filter for incidents in Texas
 - { "match_phrase": { "properties.entity.location": "Texas" } }
 - 8 docs
- Extract the damaged parts from the text representation
 - prompt: What parts of the aircraft were damaged in the incident?
 - source: document.text
 - output: damaged_parts (str)
 - 8 docs
- Find the most commonly damaged parts
 - source: properties.damaged_parts
 - decending: no
 - llm clustering prompt: Form groups of different damaged parts
 - 4 docs

AI Overview

The most commonly damaged parts in incidents in Texas were the wings and fuselage, each with 3 occurrences.

4 results total

Ask anything

How many incidents occurred in Washington? What incidents occurred in July 2024 involving birds? What were the most commonly damaged parts?

ntsb-mehul-3

wings	3
-------	---

Properties

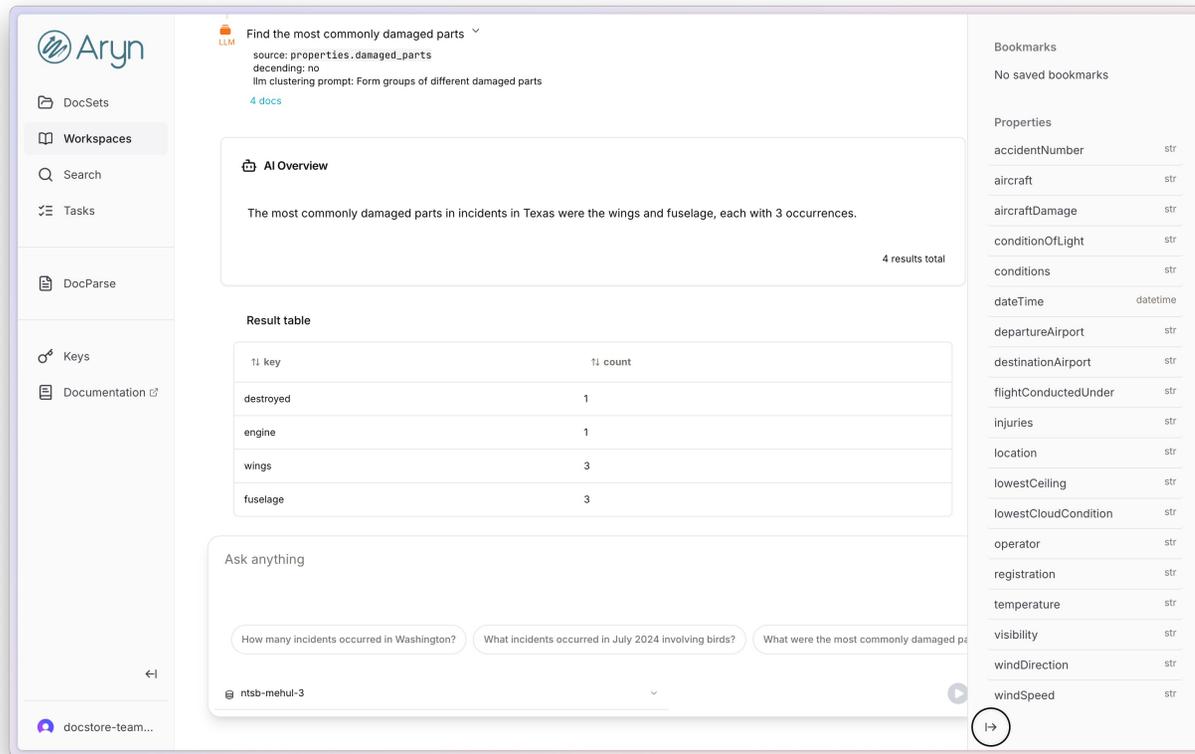
accidentNumber	str
aircraft	str
aircraftDamage	str
conditionOfLight	str
conditions	str
dateTime	datetime
departureAirport	str
destinationAirport	str
flightConductedUnder	str
injuries	str
location	str
lowestCeiling	str
lowestCloudCondition	str
operator	str
registration	str
temperature	str
visibility	str
windDirection	str
windSpeed	str

Deep Analytics

LlmExtractEntity: Extract the damaged parts from the text representation

	1: injuries	1: location	1: lowestCeiling	1: lowestCloudCondition	1: operator	1: registration	1: temperature	1: visibility	1: windDirection	1: windSpeed	1: 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	N971ST	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



Find the most commonly damaged parts

source: properties.damaged_parts
 descending: no
 llm clustering prompt: Form groups of different damaged parts
 4 docs

AI Overview

The most commonly damaged parts in incidents in Texas were the wings and fuselage, each with 3 occurrences.

4 results total

Result table

key	count
destroyed	1
engine	1
wings	3
fuselage	3

Ask anything

How many incidents occurred in Washington? What incidents occurred in July 2024 involving birds? What were the most commonly damaged parts?

ntsb-mehul-3

Properties

- accidentNumber str
- aircraft str
- aircraftDamage str
- conditionOfLight str
- conditions str
- dateTime datetime
- departureAirport str
- destinationAirport str
- flightConductedUnder str
- injuries str
- location str
- lowestCeiling str
- lowestCloudCondition str
- operator str
- registration str
- temperature str
- visibility str
- windDirection str
- windSpeed str

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