

AWS re:Invent

DECEMBER 2 – 6, 2024 | LAS VEGAS, NV

UNST
RUCT
URED

AIM223 - S

SPONSORED BY UNSTRUCTURED

Refine RAG performance: Use Unstructured for enhanced data ingestion



Brian Raymond

Founder and CEO
Unstructured



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

U N S T

R U C T

U R E D

Company Snapshot



Hi. We're Unstructured.

We help organizations transform their complex, unstructured data, like PDFs, PowerPoints, .html files, and more, into formats compatible with large language models (LLMs) so employees can chat with their internal data.



>10,000

paying customers

>16M

product downloads

>60,000

organizations using Unstructured

\$65M

raised since founding in 2022



Problem to Solve

We get ALL your data RAG-ready.

Unstructured makes it effortless to use human-generated data in conjunction with LLMs.

Out-of-the-box LLMs have three problems

- 1 They hallucinate
- 2 They know nothing about your organization
- 3 They're frozen in time

The path to mitigating these problems starts with grounding LLMs on your organization's data.



What's New This Year

2022 → **Memory**

2023 → **RAG**

2024 → **Agents**

Agents Are Mainstream



UPDATED 13:51 EDT / SEPTEMBER 19 2024



At Dreamforce, tech leaders see adoption of AI agents as next chapter of change for enterprise computing

BY MARK ALBERTSON

Hot Take

Models are getting boring.

Agents are getting interesting.

But nothing works without **good data.**

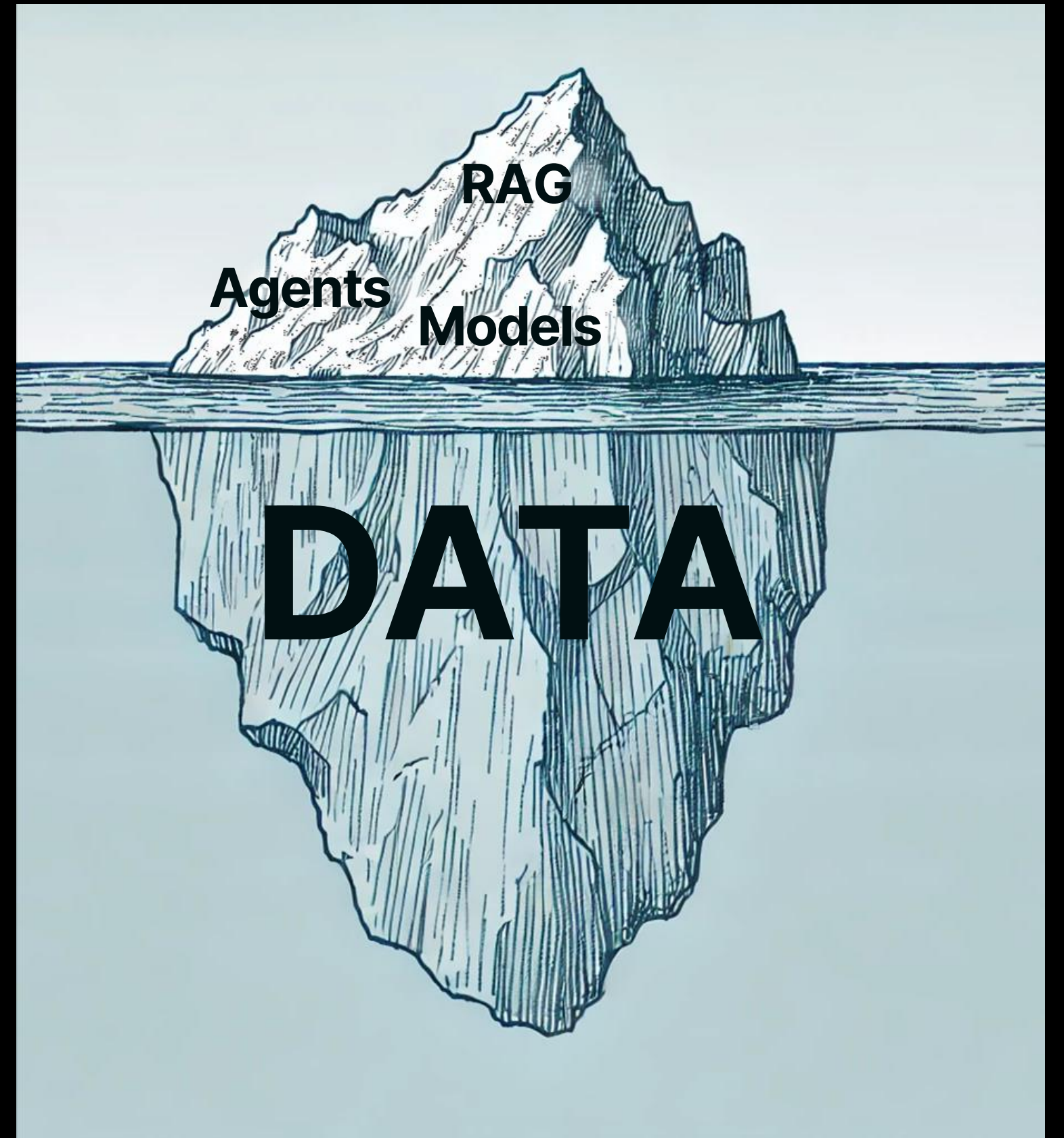
The screenshot shows a mobile view of a Fast Company article. At the top, there is a navigation bar with a menu icon, a search icon, and the 'FAST COMPANY' logo. The article title is 'In the AI era, data is gold.' in a large, bold, black font. Below the title, the date '07-01-2024' and the category 'ENGINES OF GROWTH' are displayed. The main image is a stylized tree with glowing yellow nodes and circuit-like branches against a dark background. Below the image, there is a caption: '[Source image: Vladimir Kazakov/Getty Images]'. The author information is 'BY STEVEN MELENDEZ 6 MINUTE READ'. The first paragraph of the article reads: 'As businesses scramble to take advantage of artificial intelligence, they're finding data makes all the difference in using AI effectively.' Below this, there is a link icon and the start of a second paragraph: 'A recent report from Amazon Web Services found small and medium-sized businesses that have already integrated data analysis into their...'

Data, Data, Data

It's the data that makes them work.

No matter how advanced your AI system is—whether it's an agent, RAG, or a model—its power is directly tied to the quality of data behind it.

Data is what fuels the success of every AI workflow.



Data Players Popping Up




Unlocking the Data

What we're doing about it.

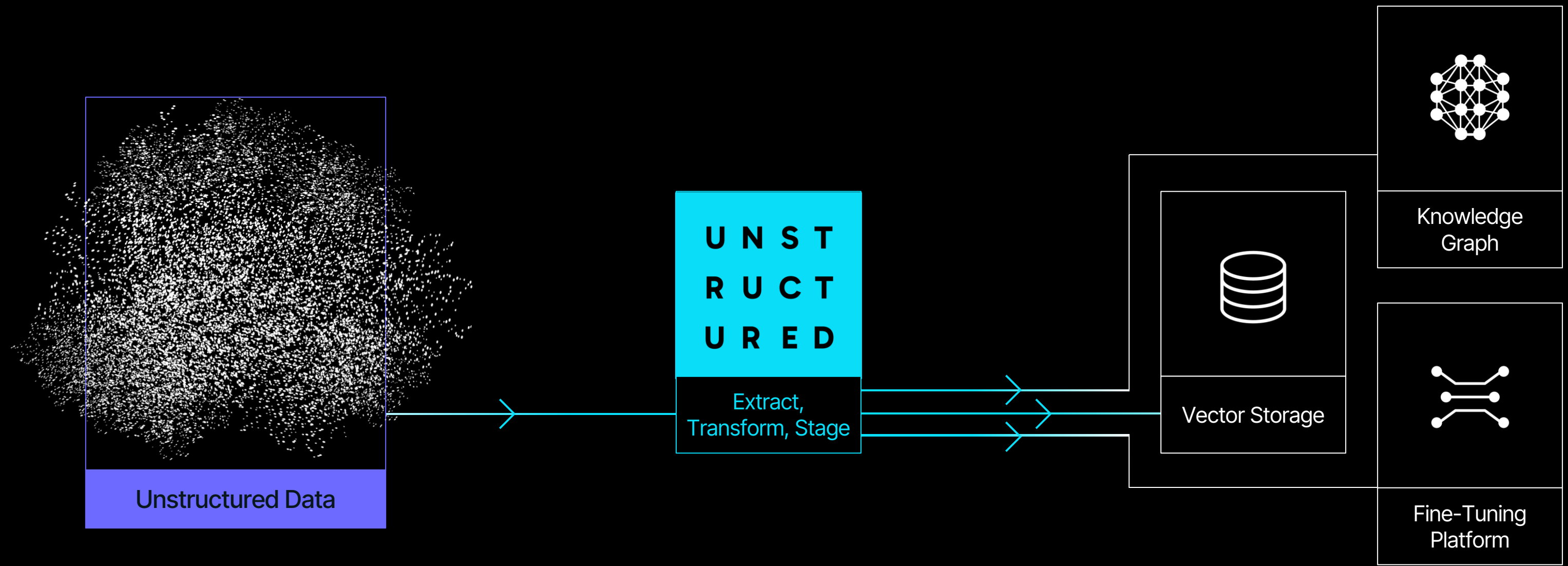
Your Internal Data = Production AI.

Without the right data, the efficacy of fine-tuning, RAG, and agents plummets. Their true potential is only unlocked when they're fueled by high-quality, proprietary data.



U N S T
R U C T
U R E D

Where we fit in the tech stack.



Superior Transformation

Input

Classification

Runs object detection algorithm and classifies each element.

Processing

Output

Yosemite Valley Overview

Yosemite Valley, located in the Sierra Nevada mountains of California, has a rich history that intertwines natural wonders and human activities. Formed over millions of years by glacial erosion, the valley is renowned for its dramatic granite cliffs, waterfalls, and diverse ecosystems. The Ahwahnechee people, a Miwok-speaking tribe, were the valley's original inhabitants, living there for thousands of years before European contact. They called the valley "Ahwahneechee," meaning "big mouth," and thrived by hunting, fishing, and trading. In 1851, during the Mariposa Indian War, the first documented European Americans entered the valley. The Mariposa Battalion, led by Major James W. W. Seargeant, sought to suppress the native population, naming the valley "Yosemite," a mispronunciation of the native word for "grizzly bear." Yosemite's beauty captured the imagination of early settlers and explorers. Photographers like Ansel Adams and naturalists like John Muir advocated for its preservation. Their efforts culminated in President Abraham Lincoln signing the Yosemite Grant in 1864, protecting the valley and the Mariposa Grove of giant sequoias, making it the first instance of land being set aside for preservation and public use. In 1890, Yosemite National Park was established, expanding protection to surrounding wilderness areas. Today, Yosemite Valley remains a testament to natural beauty and conservation, attracting millions of visitors annually who seek to experience its iconic landscapes.




Image: Yosemite Valley

| Year | Visitors |
|------|-----------|
| 2021 | 3,287,595 |
| 2020 | 2,268,313 |
| 2019 | 4,422,861 |
| 2018 | 4,009,436 |

These figures illustrate annual visitation trends, with a notable decrease in 2020 due to the COVID-19 pandemic. Yosemite's peak visitation usually occurs from May through October, with July and August being the busiest months, each accounting for approximately 16% of annual visitors (NPS) (National Park) (NPS).

PDF

UNSTRUCTURED

High-Resolution

Yosemite Valley Overview

Yosemite Valley, located in the Sierra Nevada mountains of California, has a rich history that intertwines natural wonders and human activities. Formed over millions of years by glacial erosion, the valley is renowned for its dramatic granite cliffs, waterfalls, and diverse ecosystems. The Ahwahnechee people, a Miwok-speaking tribe, were the valley's original inhabitants, living there for thousands of years before European contact. They called the valley "Ahwahneechee," meaning "big mouth," and thrived by hunting, fishing, and trading. In 1851, during the Mariposa Indian War, the first documented European Americans entered the valley. The Mariposa Battalion, led by Major James W. W. Seargeant, sought to suppress the native population, naming the valley "Yosemite," a mispronunciation of the native word for "grizzly bear." Yosemite's beauty captured the imagination of early settlers and explorers. Photographers like Ansel Adams and naturalists like John Muir advocated for its preservation. Their efforts culminated in President Abraham Lincoln signing the Yosemite Grant in 1864, protecting the valley and the Mariposa Grove of giant sequoias, making it the first instance of land being set aside for preservation and public use. In 1890, Yosemite National Park was established, expanding protection to surrounding wilderness areas. Today, Yosemite Valley remains a testament to natural beauty and conservation, attracting millions of visitors annually who seek to experience its iconic landscapes.

Text



Photo

| Year | Visitors |
|------|-----------|
| 2021 | 3,287,595 |
| 2020 | 2,268,313 |
| 2019 | 4,422,861 |
| 2018 | 4,009,436 |

These figures illustrate annual visitation trends, with a notable decrease in 2020 due to the COVID-19 pandemic. Yosemite's peak visitation usually occurs from May through October, with July and August being the busiest months, each accounting for approximately 16% of annual visitors (NPS) (National Park) (NPS).

Table

UNSTRUCTURED

OCR / Post-Processing



Multimodal LLM to Generate Summary of Image

UNSTRUCTURED

Table Processing Model

JSON [;]

Text

JSON [;]

Text

JSON [;]

Table

Tech Explosion

**There is a major tech
explosion happening.**

Tech Explosion

**The proliferation
of model choice is
incredible.**

Tech Explosion

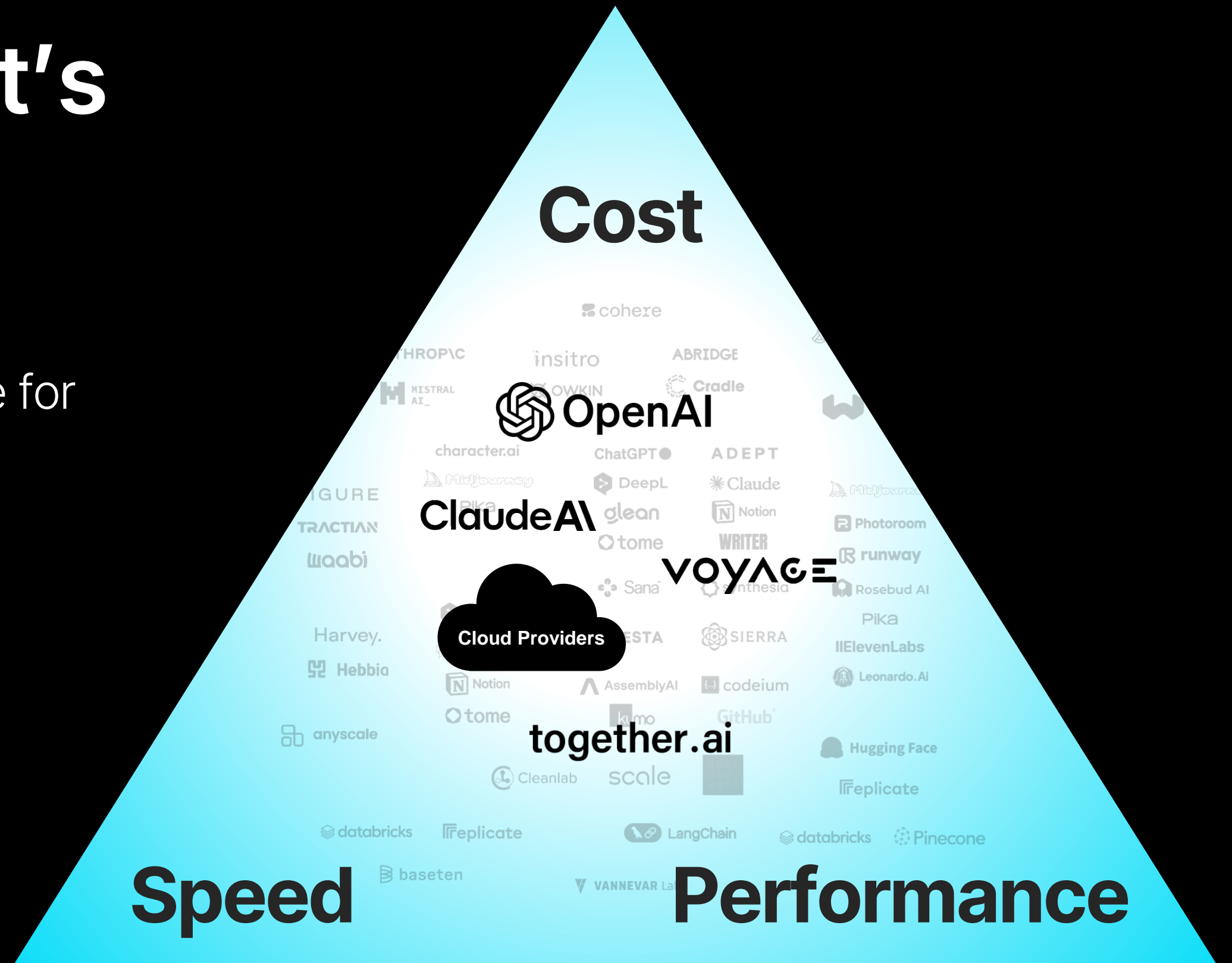
That's why we allow users to tailor their data pipelines to meet their needs.

No one size fits all here.

Optimize

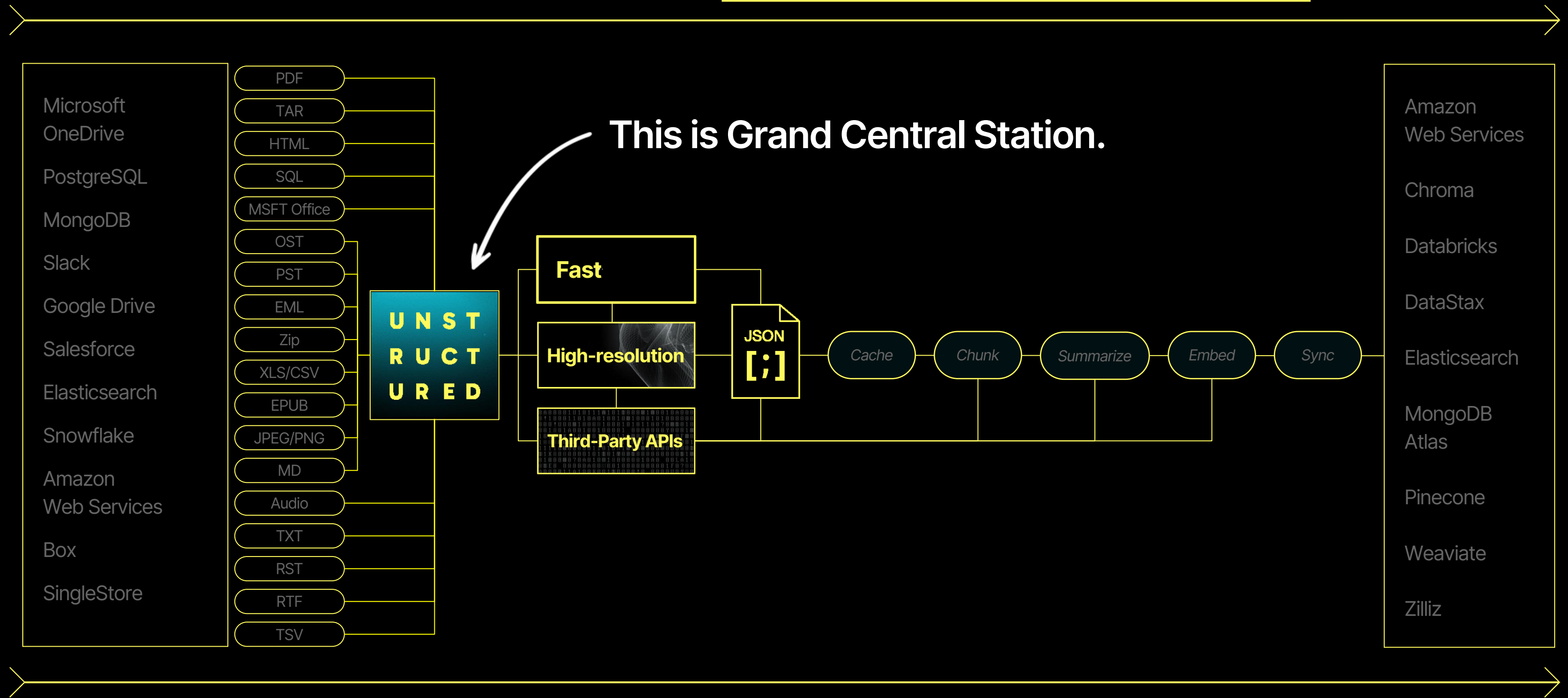
You choose what's best for you.

And optimize your own experience for cost, speed, and performance.



Effortless Orchestration

Think of Unstructured as Grand Central Station

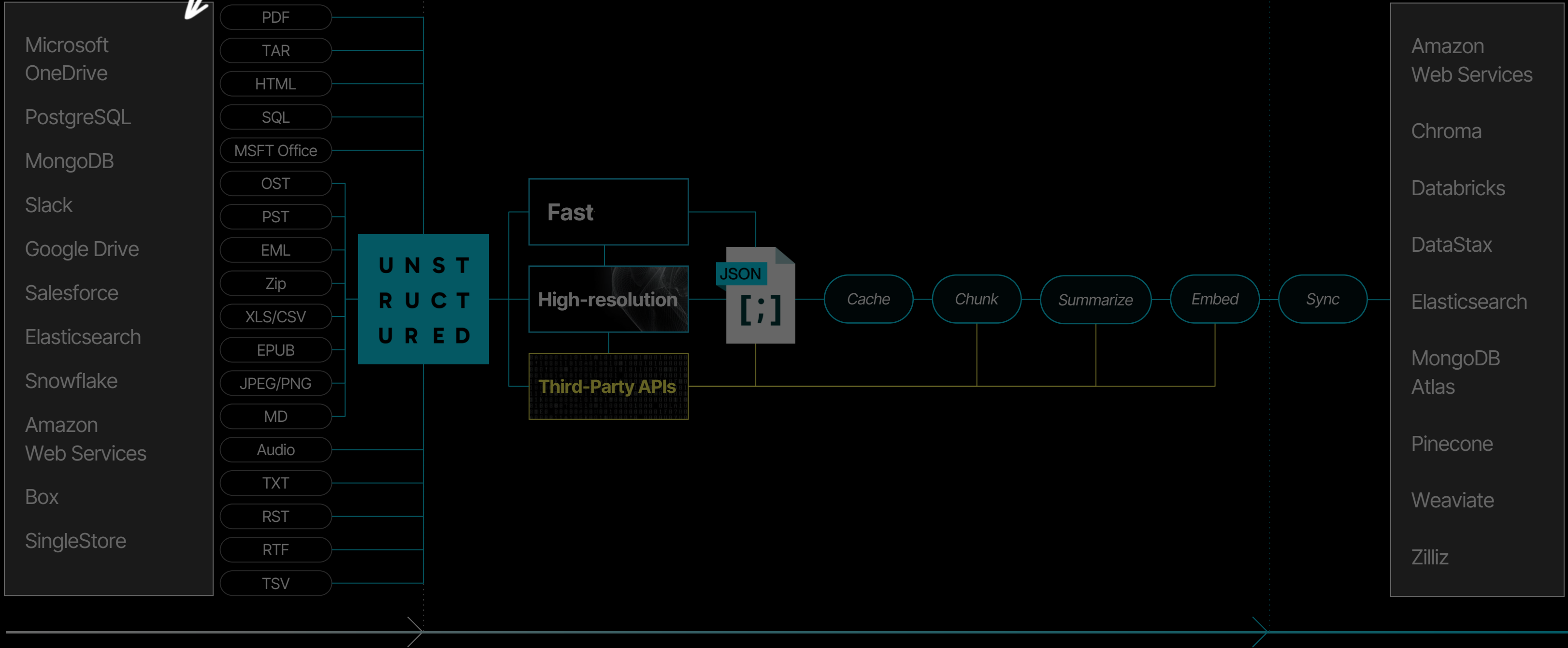


Effortless Orchestration

Connect

But what about keeping these connectors up to date?

Write / Sync



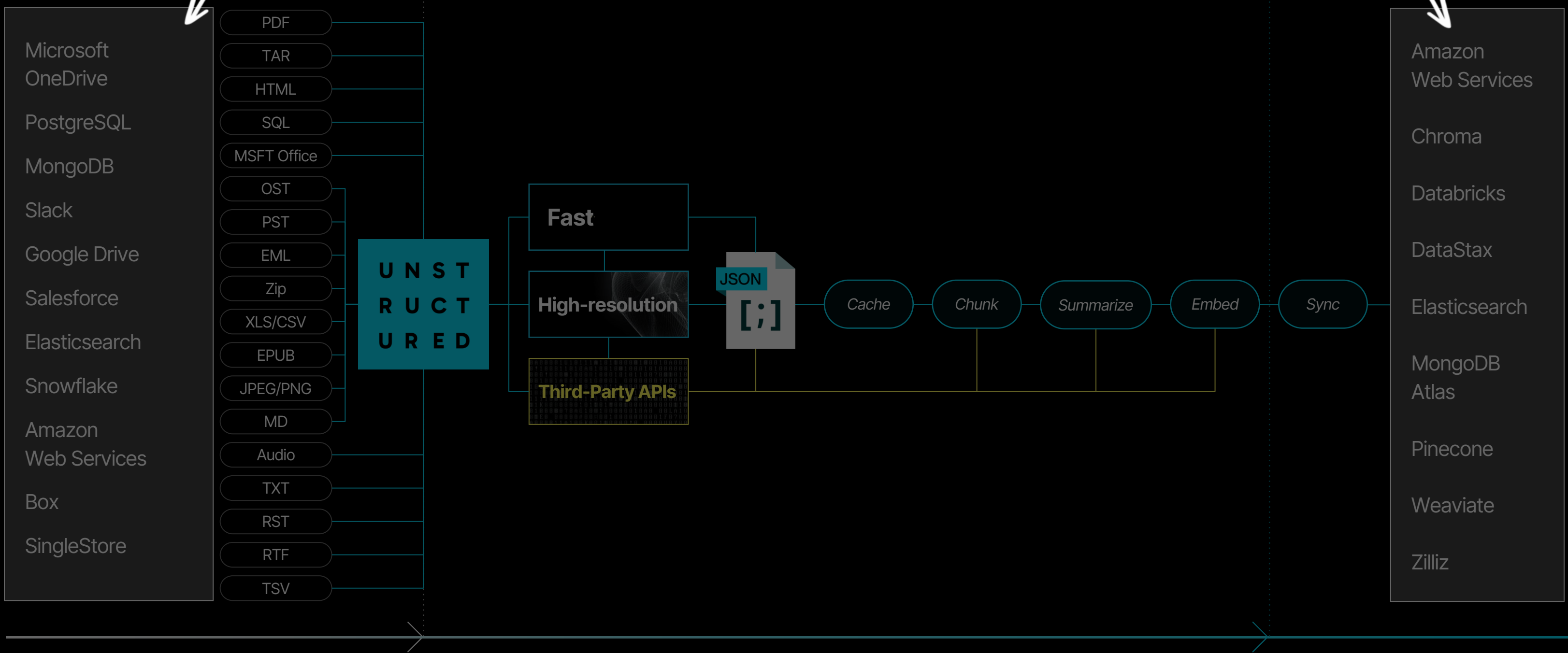
Effortless Orchestration

Connect

But what about keeping these connectors up to date?

Or these connectors?

Write / Sync



Effortless Orchestration

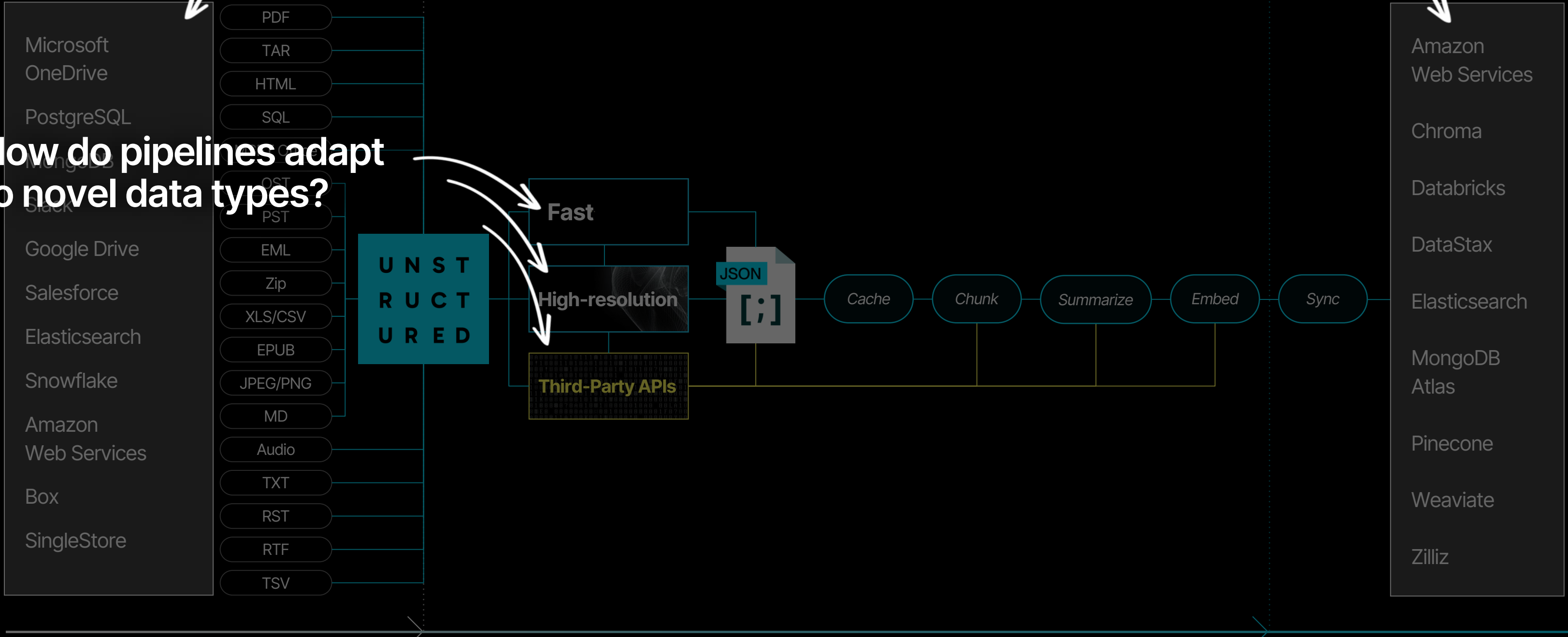
Connect

But what about keeping these connectors up to date?

Or these connectors?

Write / Sync

How do pipelines adapt to novel data types?

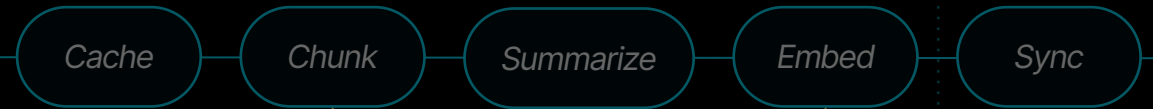


Effortless Orchestration

Connect

- Microsoft OneDrive
- PostgreSQL
- MongoDB
- Slack
- Google Drive
- Salesforce
- Elasticsearch
- Snowflake
- Amazon Web Services

- PDF
- TAR
- HTML
- SQL
- DOCX
- PST
- EML
- Zip
- XLS/CSV
- EPUB
- JPEG/PNG
- MD
- Audio
- TXT
- RST



- Amazon Web Services
- Chroma
- Databricks
- DataStax
- Elasticsearch
- MongoDB Atlas
- Pinecone
- Weaviate
- Zilliz

Write / Sync

Or these connectors?

But what about keeping these connectors up to date?

How do pipelines adapt to novel data types?

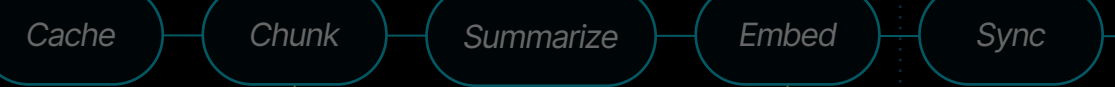
How will you manage prompts and uptime for third-party dependencies?

Effortless Orchestration

Connect

- Microsoft OneDrive
- PostgreSQL
- MongoDB
- Slack
- Google Drive
- Salesforce
- Elasticsearch
- Snowflake
- Amazon Web Services

- PDF
- TAR
- HTML
- SQL
- DOC
- PST
- EML
- Zip
- XLS/CSV
- Epub
- JPEG/PNG
- MD
- Audio
- TXT
- RST



Write / Sync

- Amazon Web Services
- Chroma
- Databricks
- DataStax
- Elasticsearch
- MongoDB Atlas
- Pinecone
- Weaviate
- Zilliz

But what about keeping these connectors up to date?

Or these connectors?

How do pipelines adapt to novel data types?

How will you manage prompts and uptime for third-party dependencies?

How expensive and performant is your chunking strategy?

Effortless Orchestration

Connect

But what about keeping these connectors up to date?

Or these connectors?

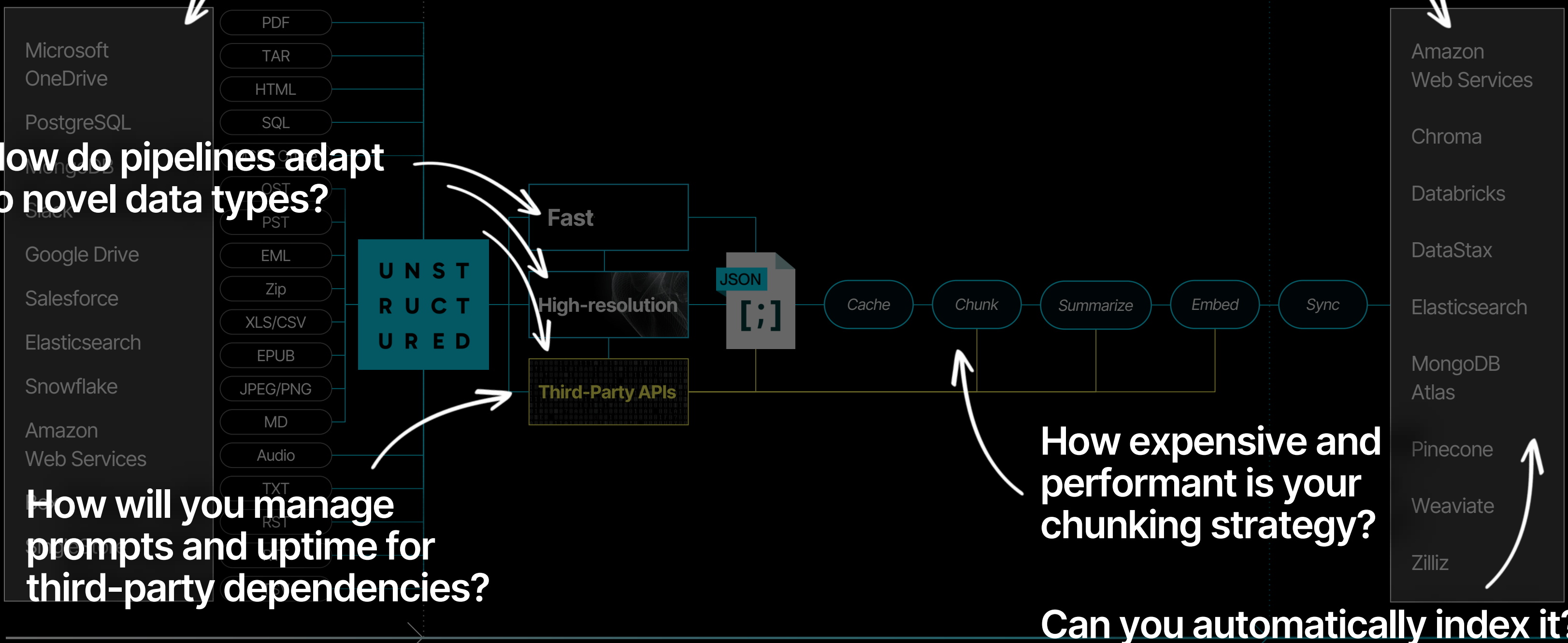
Write / Sync

How do pipelines adapt to novel data types?

How will you manage prompts and uptime for third-party dependencies?

How expensive and performant is your chunking strategy?

Can you automatically index it?

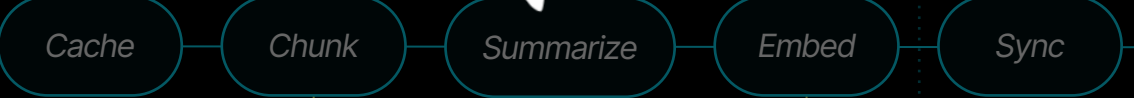
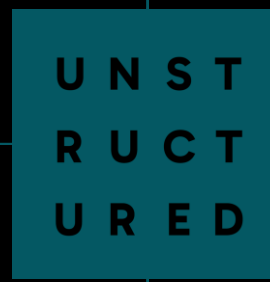


Effortless Orchestration

Connect

- Microsoft OneDrive
- PostgreSQL
- MongoDB
- Slack
- Google Drive
- Salesforce
- Elasticsearch
- Snowflake
- Amazon Web Services

- PDF
- TAR
- HTML
- SQL
- DOCX
- OST
- PST
- EML
- Zip
- XLS/CSV
- EPUB
- JPEG/PNG
- MD
- Audio
- TXT
- RST



Write / Sync

- Amazon Web Services
- Chroma
- Databricks
- DataStax
- Elasticsearch
- MongoDB Atlas
- Pinecone
- Weaviate
- Zilliz

But what about keeping these connectors up to date?

Or these connectors?

How do pipelines adapt to novel data types?

Can your prompts generate summaries and structured data for GraphRAG?

How will you manage prompts and uptime for third-party dependencies?

How expensive and performant is your chunking strategy?

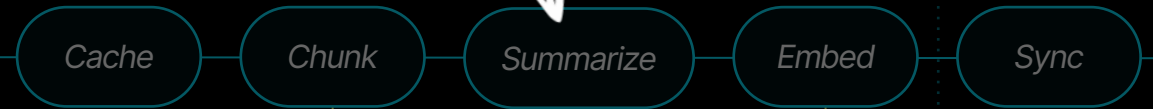
Can you automatically index it?

Effortless Orchestration

Connect

- Microsoft OneDrive
- PostgreSQL
- MongoDB
- Slack
- Google Drive
- Salesforce
- Elasticsearch
- Snowflake
- Amazon Web Services

- PDF
- TAR
- HTML
- SQL
- DOCX
- OST
- PST
- EML
- Zip
- XLS/CSV
- EPUB
- JPEG/PNG
- MD
- Audio
- TXT
- RST



- Amazon Web Services
- Chroma
- Databricks
- DataStax
- Elasticsearch
- MongoDB Atlas
- Pinecone
- Weaviate
- Zilliz

Write / Sync

But what about keeping these connectors up to date?

Or these connectors?

How do pipelines adapt to novel data types?

How will you manage prompts and uptime for third-party dependencies?

Can your prompts generate summaries and structured data for GraphRAG?

Are you preserving reading order?

How expensive and performant is your chunking strategy?

Can you automatically index it?

Effortless Orchestration

Connect

- Microsoft OneDrive
- PostgreSQL
- MongoDB
- Slack
- Google Drive
- Salesforce
- Elasticsearch
- Snowflake
- Amazon Web Services

- PDF
- TAR
- HTML
- SQL
- DOCX
- OST
- PST
- EML
- Zip
- XLS/CSV
- EPUB
- JPEG/PNG
- MD
- Audio
- TXT
- RST

UNSTRUCTURED

Fast
High-resolution
Third-Party APIs

JSON [;]

Cache -> Chunk -> Summarize -> Embed -> Sync

- Amazon Web Services
- Chroma
- Databricks
- DataStax
- Elasticsearch
- MongoDB Atlas
- Pinecone
- Weaviate
- Zilliz

Write / Sync

But what about keeping these connectors up to date?

Or these connectors?

How do pipelines adapt to novel data types?

How will you manage prompts and uptime for third-party dependencies?

Are you preserving reading order?

Can your prompts generate summaries and structured data for GraphRAG?

Do you have auto-recommenders?

How expensive and performant is your chunking strategy?

Can you automatically index it?

Effortless Orchestration

Connect

But what about keeping these connectors up to date?

Or these connectors?

Write / Sync

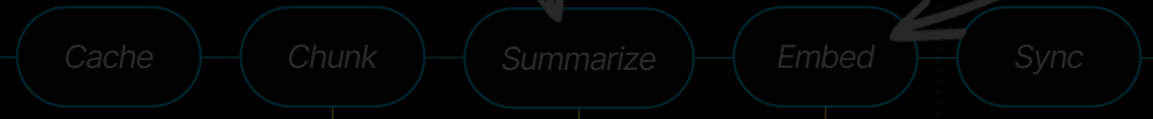
- Microsoft OneDrive
- PostgreSQL
- Google Drive
- Salesforce
- Elasticsearch
- Snowflake
- Amazon Web Services

- PDF
- TAR
- HTML
- SQL
- Zip
- XLS/CSV
- Epub
- JPEG/PNG
- MD
- Audio
- TXT
- RS1



High-resolution
Third-Party APIs

JSON [;]



- Amazon Web Services
- Chroma
- Elasticsearch
- MongoDB Atlas
- Pinecone
- Weaviate
- Zilliz

This is orchestration.

How do pipelines adapt to novel data types?

Can your prompts generate summaries and structured data for GraphRAG?

Do you have auto-recommenders?

How will you manage prompts and uptime for third-party dependencies?

Are you preserving reading order?

How expensive and performant is your chunking strategy?

Can you automatically index it?

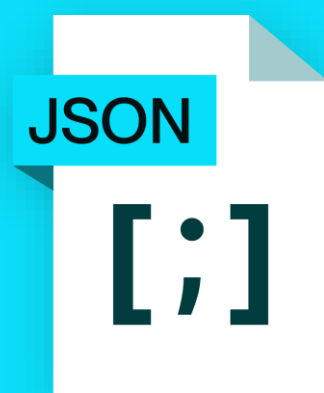
Platform Demo



Why Unstructured

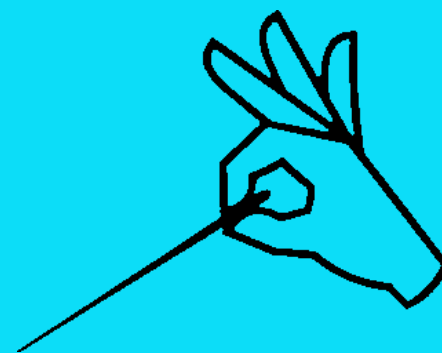
Our special sauce.

Industry-Leading
Transformation



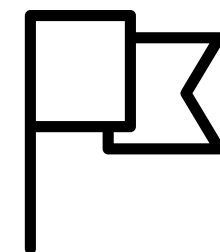
+

Effortless ETL
Orchestration



=

Developers' Choice
for ETL for LLMs



U N S T
R U C T
U R E D

**Whatever it is,
we can structure it.**



Thank you!



<https://unstructured.io>



Please complete the session survey in the mobile app

