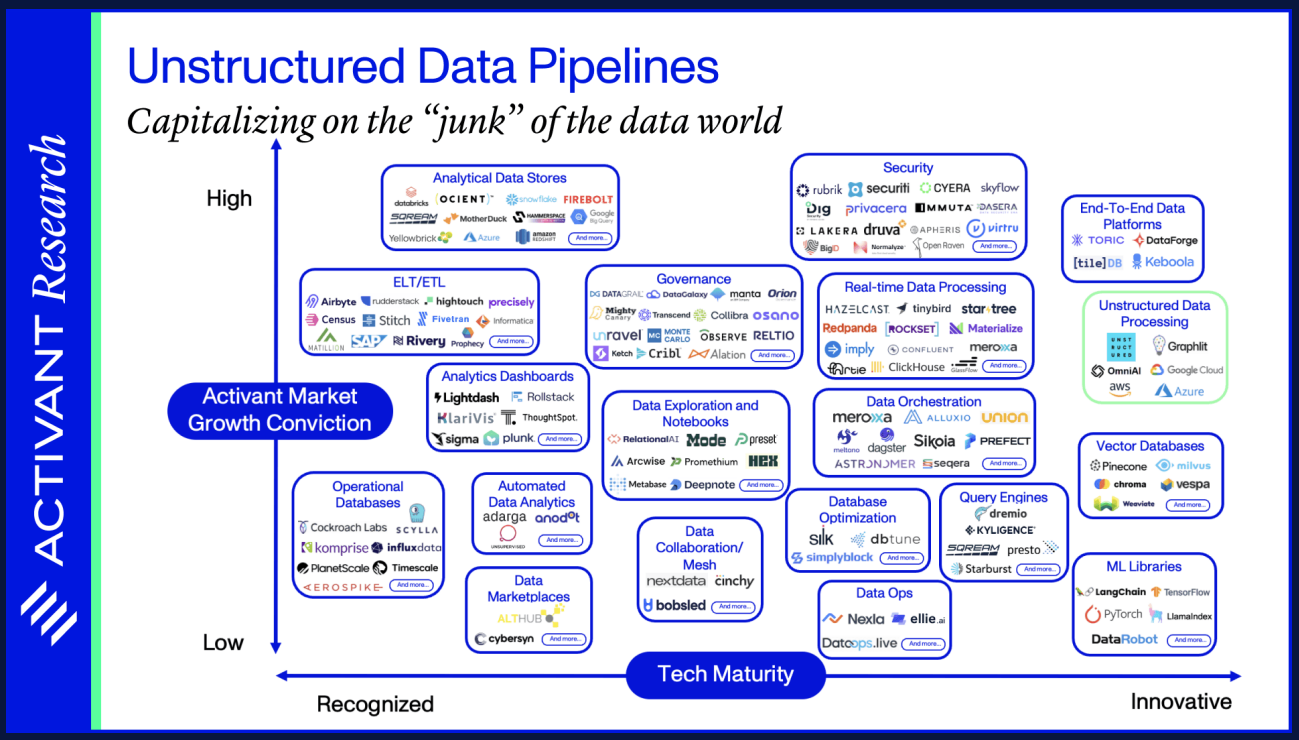
ACTIVANT RESEARCH

Unstructured Data Pipelines

Capitalizing on the "junk" of the data world

Jono Vickery



Q2 2024

Humans share knowledge through writing, speaking, and demonstrating ideas to one another. Within companies this knowledge is shared in the form of hand-written notes, PDF documents, scanned images, Slack messages, and collaborative documents such as Google Docs or Notion. These disparate sources collectively encapsulate the company's proverbial brain. IDG reported that the volume of internal company data is growing by as much as 63% a month from an average of 400 different sources.[1] And only half of that data is currently being used to extract value.[2] Tapping into this rich source of information has been a challenge because computers need neat rows and columns to make sense of it, but that's changing.

Advances in Artificial Intelligence (AI), particularly in the fields of Natural Language Processing (NLP) and Large Language Models (LLMs), have closed this gap. Deloitte found in a 2019 survey that organizations making use of unstructured data were **24% more likely to exceed their business goals.**[3] The benefits today are undoubtedly many times greater and the demand for solutions that can prepare unstructured data for AI processing has exploded. It remains to be seen whether the winners will be point solutions integrating with existing tools or new data pipelines designed to accommodate unstructured data.[4, 5, 6, 7]

Capitalizing on the "junk" of the data world

Unstructured data was typically disregarded by enterprises as "junk" amidst the orderly world of structured data. But, companies can capitalize on the latent value within the masses of unstructured data if the right AI tools are implemented. Recent advancements have enabled material improvements in key areas such as customer service, workforce scheduling, demand forecasting, and knowledge management.

1. **Customer service:** NLP can be used to optimize call centers by analyzing human speech and text to extract sentiment and engage in complex human-like conversations with customers. Klarna, for example, uses AI-enabled assistants to handle two-thirds of its customer service chats more accurately **than human agents, delivering service improvements and a $40mn saving**. If we extrapolate this experience and assume that 50% of global customer service chats could be handled by AI in this manner, the industry could save ~$30bn annually.[4]
2. **Workforce scheduling:** Managing a workforce is complex and requires juggling employee availability, skill sets, preferences, and performance standards. Machine Learning (ML) models collate these factors from unstructured data and historical patterns to optimize cost savings, improve employee satisfaction, and reduce human bias and error. ML models continuously learn and adjust in real-time to match demands, a task that would be near impossible for a human to keep up with in large corporations.
3. **Demand forecasting:** Many demand forecasting solutions lean heavily on regression models, occasionally enhanced with seasonal overlays. Forecasting models need to be far more sophisticated than these parametric solutions to effectively synchronize across stakeholders that include sales, marketing, finance, and external third-party collaborators. Using NLP to

analyze customer sentiment from reviews is more beneficial than relying solely on the rating of a product to forecast sales.

4. **Knowledge management:** Searching for and gathering data is tedious and time consuming.[8] Machines can optimize these time- and labor-intensive tasks while also synthesizing and summarizing information, enhancing content with relevant links and insights, identifying gaps, and revealing untapped potential or weaknesses in internal knowledge and processes.

These use cases illustrate the market opportunity for tools that leverage unstructured data. So why have we waited so long to use it and why is organization-wide adoption only 27%?[9]

The sticking points of unstructured data

Unstructured data processing is complicated by the fragmented nature of the sources, difficulty searching the data and extracting accurate information, as well as privacy and security concerns.

1. **Fragmentation:** 90% of internal company data is unstructured and needs to be identified, located, and centralized before it can be processed. Fragmented data limits knowledge sharing and data observability.
2. **Searchability:** Conventional search methods discover data based on file names or metadata (if tagging has taken place). Locating key information at scale within a vast pool of unstructured data requires advanced search capabilities able to locate specific components within documents.
3. **Accuracy:** Extracting accurate information from diverse sources is difficult. Optical Character Recognition (OCR) and NLP can be used to extract structured or semi-structured data from unstructured data, but even these advanced methods are fallible, and errors can creep in with significant downstream consequences. An invoice automation tool that misreads the placement of a decimal in the invoice amount could result in the final payment being off by a significant amount.
4. **Privacy and security:** Companies need to be cognizant of the privacy and security of data if it is destined for LLMs. Unstructured data, including invoices, customer complaints, and health records, is governed by regulations and restrictions. Concerns arise when model vendors use private data to train new models in LLMs and Generative AI. Private data can potentially become an AI-generated response for an unknown future user.

Enterprises need infrastructure that can deal with these complexities. Many companies have existing data pipelines built for structured data processing that can be augmented to process unstructured data. The alternative is to completely overhaul the conventional data stack and implement a new unstructured data pipeline. The result needs to be unified, governed, secure, and accessible. It is likely well worth the effort and expense though as 92% of organizations believe that achieving this would benefit innovation and reduce costs.[10]

Keep the old and plug in the new

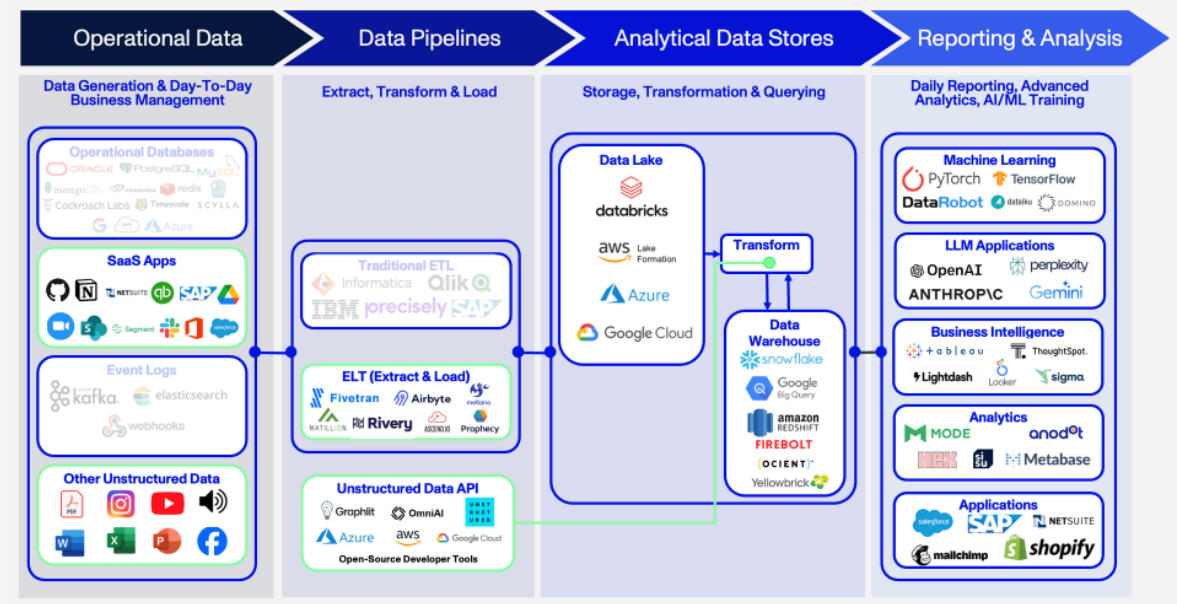
Conventional data pipeline infrastructure and transformation tools were assembled to process structured data. ETL (Extract, Transform, Load) pipelines emerged in the 1970s when the cost of data storage was prohibitively high, and the only viable option was to load transformed data into the warehouses and maintain a single copy of a structured dataset. The cost of data storage has fallen by 99.99% since the 1980s[11] and today companies typically move and dump all available raw data into data lakes, deferring transformation until later. This shift has modularized systems by removing dependency between loading and the transformation process. Relocating the "T" to the end of the pipeline has resulted in the ELT (Extract, Load, Transform) models popularized in recent years.

This modularity has been instrumental in augmenting the conventional data stack to allow the processing of unstructured data and the creation of a “modern data stack”. The simple exhibit on the next page shows how the transformation step addresses many of the stumbling blocks encountered when dealing with unstructured data. Simple API calls can be integrated to accommodate the transformation of previously unused data sources like SaaS applications, PDF and Word documents, or social media content. API providers such as Unstructured, Graphlit, Google Cloud, Microsoft Azure, or AWS can operate inside of the datastore focusing on the “T” in ELTs – leaving the “E” and “L” components to remain in their current forms. This means that:

1. companies can leverage sunk costs,
2. structured and unstructured data analysis can be unified,
3. existing vendors can give access to significant technology that newer vendors might be behind on – such as privacy and security, rate limiting, and pagination.

ACTIVANT *Value Chain*

Modern Data Stack for Unstructured Data



This presents a straightforward solution for those with established data pipelines, but it may not be that simple. In 2023, 40% of organizations reported difficulties integrating unstructured data technology with existing tools.[12] For new LLM-centric companies or new projects in existing companies, there is potential for a new data pipeline focused solely on unstructured data and the critical tasks needed to get it LLM-ready.

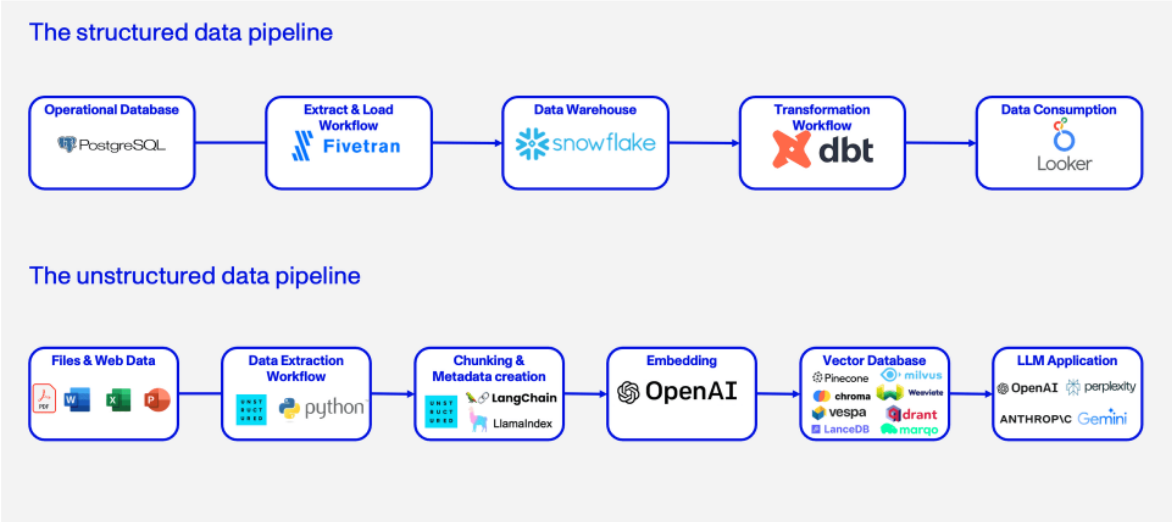
> Unstructured data ingestion and preprocessing is a critical bottleneck hindering organizations' ability to move LLM architectures from prototype to production. The challenges faced are 1) transforming any unstructured data into a standardized format ready for LLMs, and 2) creating value during preprocessing to help models find exactly the data they need, and nothing more, at inference time."
>
> **Brian Raymond**
> Founder & CEO, Unstructured Technologies.

Out with the old and in with the new?

LLM models require inputs in a very particular form. Transformation tools need to perform *extraction (which incorporates cleanup, normalization, and classification), chunking, metadata creation, embedding,* and *loading into vector databases*. These additional transformation requirements are crucial but not yet sufficiently catered for by the structured data pipeline market. As shown in the exhibit on the next page, unstructured data for a Retrieval Augmented Generation (RAG)model needs to be processed from its raw form in a very different way to structured data. Verizon uses an approach like this to feed the RAG model underpinning a chatbot that provides field service workers with a Q&A-type function that draws on operating manuals stored as PDF documents and HTML web pages. The unstructured data from these sources needs to be *extracted* and accurately converted to JSON, a kind of semi-structured text. That text is then *chunked* – broken up into smaller pieces to enhance the semantic understanding of an LLM – and each chunk converted to *vector embeddings*, a numerical representation enabling the machine to comprehend text relationships. These embeddings are then loaded into a *vector database*. This final step requires employing an advanced LLM, like GPT-4, in conjunction with the vector database to power the chatbot.

ACTIVANT *Value Chain*

The Structured vs Unstructured Pipeline

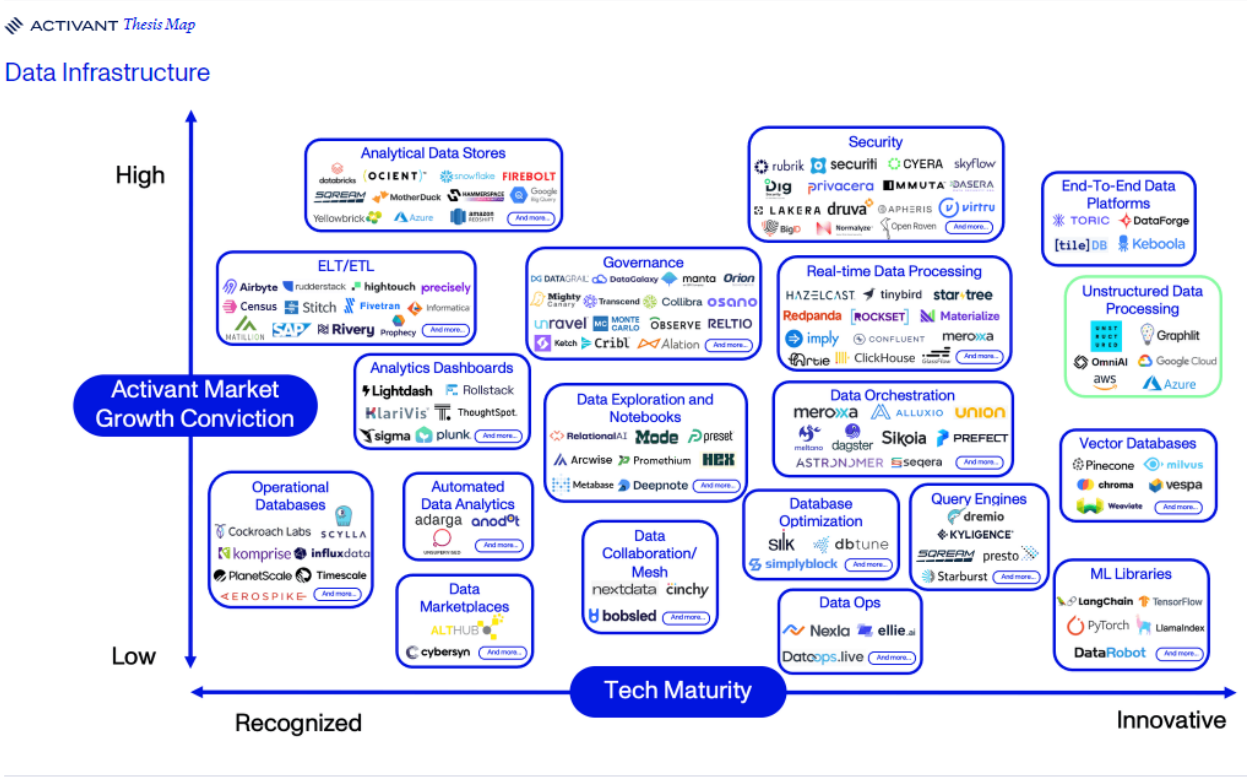


A vendor that makes these capabilities the core to their tools could lead the shift away from a solitary API call – a mere point-solution within the existing data stack – to becoming the pivotal enablers of a new data pipeline.

Unstructured Data in the Greater Scheme

Enterprises are ramping up their infrastructure to keep up with shifting AI trends and unstructured data is becoming increasingly more valuable with each advancement. It remains to be seen

whether the winners of unstructured data transformation will be point-solutions or new data pipeline tools. Either way, unstructured data processing tools show great promise in the overall landscape of data infrastructure despite the current small number of contenders. These innovation leaders are outlined in green in the Activant Data Infrastructure Thesis Map which arrays emerging solutions on technical maturity – from recognized to innovative solutions – and Activant's market growth conviction.



Leaders of the Pack

It is always difficult to forecast winners in rapidly changing markets but, based on our initial work in the space, we like the solutions provided by the following players:

1. **Unstructured**, with transformation capabilities that require no pre-training and have no file type limitations, providing 90%+ accuracy.[13] The company is deeply embedded in the LLM ecosystem, offering an API tool that provides a plug-and-play solution for more than 40 upstream and downstream locations. They have had 7.5mn downloads in the last twelve months with 50k company users. Their new platform product aimed at automating retrieval, transformation, and staging of unstructured data for LLMs is in the works and they have recently partnered with LangChain to provide RAG support offerings.
2. **Graphlit**, an *API-first platform, that* delivers a singular solution to a currently segmented process and simplifies the task of turning raw unstructured data into usable content for LLMs, allowing users to build multimodal RAG applications out of the box. Additional features enable

users to engage with their data by using their RAG-as-a-service tool, performing semantics searches, and automating content generation.

3. **LangChain** and **LlamaIndex**, which are grouped separately as developer-focused Machine Learning (ML) libraries, widely used to build modular Gen AI products while integrating with major providers such as AWS, OpenAI, and Microsoft through API frameworks.
4. **AWS, Google Cloud,** and **Microsoft Azure**, which all have extensive capabilities in this area and are well-positioned to take a larger market share if current customers choose to keep their pipeline tools consolidated with an existing provider. Startups need to compete strategically against the advantages of these platforms. Two drawbacks of these tools are that 1) users are typically required to pull together multiple products to achieve what one powerful tool otherwise could, and 2) they are unable to handle more than a handful of file types.

Conclusion

AI has unlocked the potential for enterprises to harness and process vast amounts of unstructured data, offering significant benefits to business performance. To fully capitalize on this opportunity, data architecture must continue to evolve, and sophisticated solutions are required to transform and leverage unstructured data.

If you have a different view or are building in this space, we'd love to hear from you. Please reach out!

Endnotes

1 Matillion and IDG, Data Growth is Real, and 3 Other Key Findings, 2022

2 IDC, Untapped Value: What Every Executive Needs to Know About Unstructured Data, 2023

3 Deloitte, Analytics and AI-driven enterprises, 2019

4 Klarna, Klarna AI assistant handles two-thirds of customer service chats in its first month, 2024

5 CWA Union, The United States Call Center Worker and Consumer Protection Act, 2024

6 Tres Commas, How Big The Global Call Center Industry Actually Is In Statistics And Numbers, 2021

7 Activant Analysis

8 Microsoft, Work Trend Index Annual Report: Will AI Fix Work?, 2023

9 Ibid

10 Ibid

11 Fivetran, The Ultimate Guide to Data Integration, 2021

12 IDC, Untapped Value: What Every Executive Needs to Know About Unstructured Data, 2023

13 Accuracy is based on broad user feedback