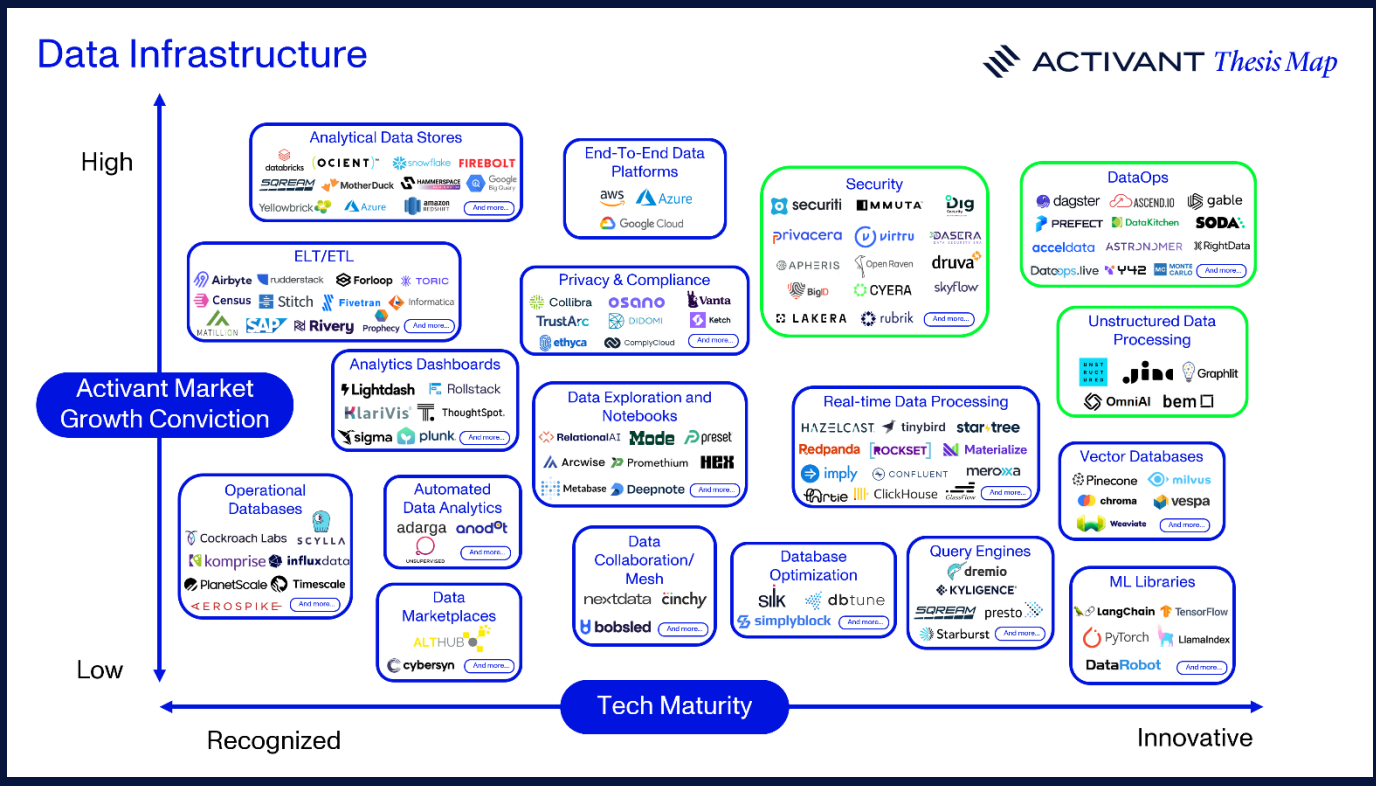
ACTIVANT RESEARCH

AI's Undervalued Asset

Data's pivotal role in the next era of Gen AI deployment



Nina Matthews, Jonathan Vickery

Q1 2025

2024 delivered one newsworthy AI-related headline after another. Software powerhouse OpenAI secured a historic $6.6 billion in October '24, the largest venture capital round at the time.[1] NVIDIA, the leader in GPU compute, saw its CEO Jensen Huang reach celebrity status – dubbed the "Taylor Swift of tech" by Meta's Mark Zuckerberg – while its stock surged over 200% over the year.[2] Google DeepMind's Demis Hassabis and John Jumper were co-awarded the Nobel Prize in Chemistry for AlphaFold, an AI breakthrough that solved a 50-year-old challenge in protein structure prediction.[3] And of course, ChatGPT became as ubiquitous as the internet itself, embedding AI firmly into daily life and redefining how industries envision their futures.

While only 5% of enterprises had deployed AI into live production as of 2023, that number is expected to climb to 80% by 2026, driven in no small part by the flood of success stories.[4] Despite the spotlight falling primarily on compute and software, AI requires three equally critical foundational layers to drive its capabilities and growth:

1. **Compute Infrastructure:** Advanced hardware from semiconductor leaders like TSMC and GPU designers such as NVIDIA, supported by cloud providers such as Microsoft Azure, AWS, and Google Cloud
2. **Software and Foundation Models**: Cutting-edge algorithms, frameworks like TensorFlow and PyTorch, and foundational models developed by companies like OpenAI and Meta aid the learning and adaptive abilities of the models
3. **Data**: The lifeblood of models, enabling them to generate insights and make decisions, facilitated by companies like Snowflake and Databricks specializing in data infrastructure

When we initially drafted this article in November '24 – with plans for an early 2025 release – data still played the role of the neglected sibling. But the VC world moves fast and waits for no man (or publication schedule), and by December '24 Databricks had announced a record-shattering $10 billion funding round, surpassing OpenAI as the largest in history.[5]

This dramatic shift in how the industry values data validates the initial premise of this article: the data layer is critical to AI's success and has some catching up to do. Despite Databricks' milestone, this is just the beginning and there's a great deal more progress to come.

Compute and Software Still Dominate AI Investment

As of 2024, major compute companies boasted a combined market capitalization of $12.7 trillion, with software following at $4.2 trillion.[6,7] By comparison, the combined market capitalization of the largest data companies sat at just $600 billion.[8] Compute was three times larger than software, but a shocking 21 times larger than data, creating a significant imbalance among the three layers. Venture capital told a similar story. In 2024, compute attracted $15 billion in VC funding and software drew $56 billion, yet to date data infrastructure remains unlisted as a formal category in major VC platforms like Pitchbook and Dealroom.[9,10] As captured below, the data layer is behind and underfunded compared to its peers.

Why Data Has Lagged – And Why That's Changing

Given the nascent stage of Gen AI, the focus on scaling compute and software made sense. It allowed us to advance from GPT-1's 117 million parameters to GPT-4's estimated trillions, transforming from basic text-only capabilities to complex multimodal reasoning and extended contextual understanding.[11] Most models are available through APIs or platforms like Azure Cloud, but companies that want to train their own models can gain a competitive edge just by securing GPU capacity.

The intense emphasis on compute and software, however, still relegates data to an afterthought, largely because managing it is inherently complex and costly. In fact, 72% of enterprises cite data management as a major hurdle to AI adoption as information is often dispersed across multiple systems, formats, and geographies.[12] The fragmentation makes it difficult to ensure consistency, maintain quality, and control access. Even among high-performing Gen AI users*, 70% identify data issues as a significant barrier to unlocking value.[13] Gartner reinforces this, projecting that data-related challenges will cause more than 30% of Gen AI projects to fail.[14] With the average enterprise expecting an 80% increase in new data volumes over the next three years, the problem is set to worsen.[15]

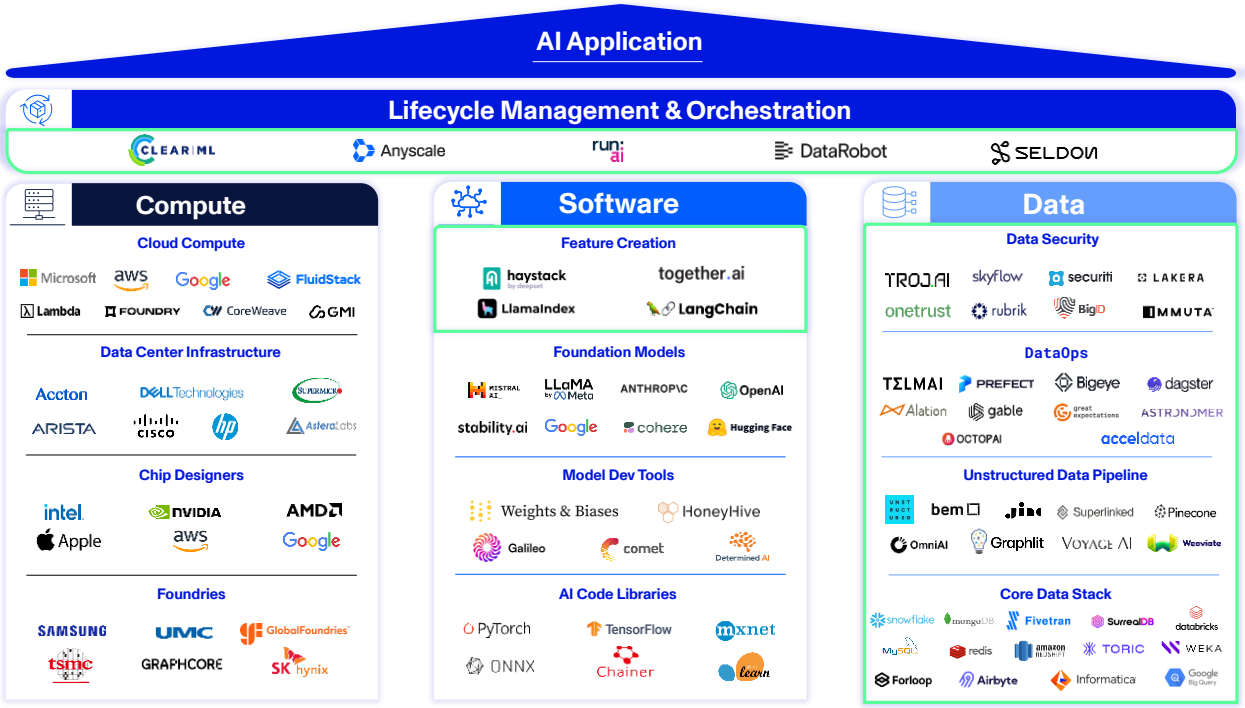
* Defined by McKinsey as those attributing at least 11% of their 2023 EBIT to Gen AI tools.

Data Is the Key Differentiator

The industry is transitioning from experimentation to mass deployment but advancements in compute and foundation models are beginning to plateau, delivering diminishing returns from scaling resources. OpenAI's upcoming Orion model, for example, is reportedly showing only moderate improvements over its predecessor GPT-4, signaling that simply increasing computational power is no longer enough.[16] As investment continues to surge, compute and software – with market opportunities of $426 billion and $185 billion respectively – are becoming increasingly democratized, if not fully commoditized.[17,18] When that happens, companies will be challenged to differentiate their AI deployments. The answer, we believe, lies in data.

> "The importance of data... I underrated it dramatically I thought it was just scale. A lot of proof points have happened internally at Cohere that have just totally transformed my understanding of what matters in building this technology. The quality of data... A single bad example amongst billions - it's so sensitive. It is a bit surreal how sensitive the models are to their data. Everyone underrates it."
>
> **Aidan Gomez, Co-Founder & CEO, Cohere**
>
> As discussed on 20VC with Harry Stebbings.

If we think of the foundational layers as pillars, building an application becomes akin to constructing a house. A strong foundation is essential for durability, but if one of the three pillars is weaker than the others then the structure risks faltering under pressure, or, in the case of AI, failing to produce reliable outputs. In our graphic below, we outline the architecture of a typical AI application, supported by the three pillars: compute, software, and data.



Note: This is by no means an exhaustive list.

Tools are scaffolded progressively from the bottom up, moving from fundamental requirements to increasingly advanced capabilities. In the compute pillar, companies like NVIDIA, Cisco, Microsoft Azure, and Google Cloud have already built strong foundations. But, as we shift our focus to software and data, the areas highlighted in green are what we believe to be key opportunities for differentiation in the coming years.

While the software stack is largely mature, encompassing established AI code libraries, model development tools, and foundational models, there remains substantial potential in deploying technologies such that enhance the features set of the average model to include capabilities such as Retrieval-Augmented Generation (RAG). RAG enhances the accuracy and relevance of AI-generated text by grounding it in information retrieved from an external knowledge source.

The success of RAG, however, relies heavily on the strength of the data pillar. Without robust infrastructure to support efficient data storage, processing, and accessibility, RAG's full capabilities cannot be achieved. This dependence on data highlights a paradox: despite being essential for reliable Gen AI, data has historically received far less investment than compute and software.

What's the Opportunity?

The prevailing underinvestment in data presents both a challenge and a significant opportunity. Increasing investment in data can benefit model trainers and AI users alike.

For model trainers, the focus should shift to prioritizing data quality over compute to improve performance. As compute and software advancements begin to level off, richer, more complex datasets will be essential for driving the next wave of breakthroughs. Future progress will depend less on scaling computational power and more on the caliber of the training data. Domain-specific data presents a significant opportunity here, enabling models to deliver tailored, high-accuracy results for specialized industries.

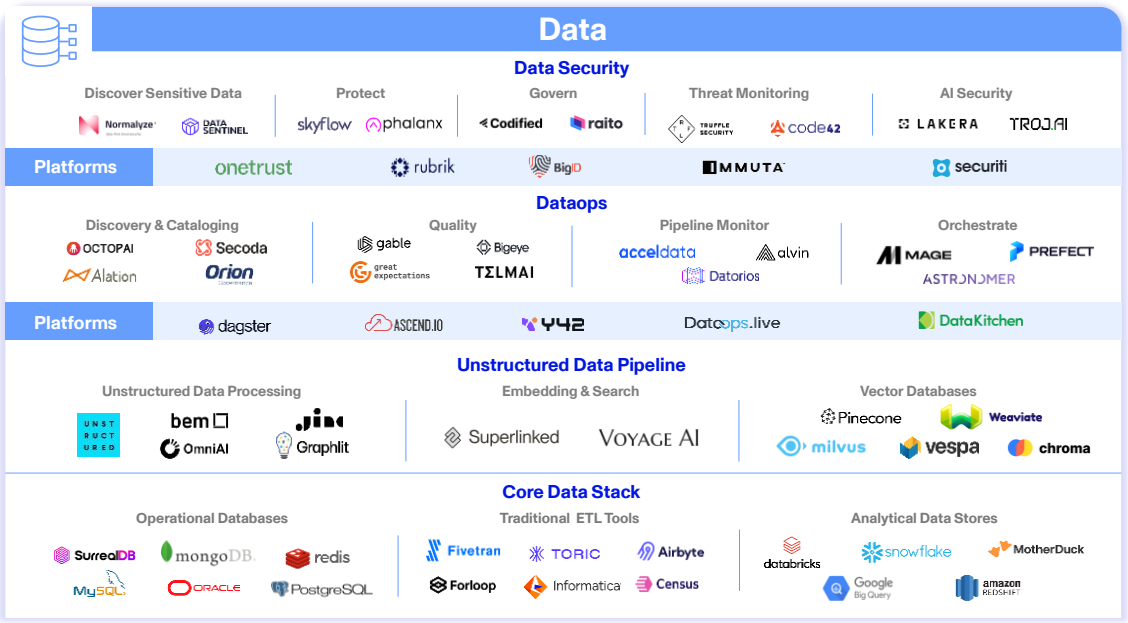
Enterprises adopting AI must recognize proprietary data as a competitive edge. Off-the-shelf models won't meet specific business needs and will no longer serve as differentiators. By customizing models with their own data, companies can achieve more precise and context-enriched results. Organizations need to invest in finetuning existing models and implementing techniques like RAG to significantly boost performance for focused use cases.

How Will We Achieve This?

Most companies already have core elements in place, including operational databases, traditional ETL tools, and analytical data stores, supported by established providers like MongoDB, Airbyte, Snowflake, and of course Databricks. The next step, however, will require advanced tools that can handle both structured and unstructured data at scale, with greater precision and stronger security. Over the past year, Activant has researched the evolving landscape of data infrastructure. Consolidating our findings, we present a closer look at the foundational building blocks necessary to strengthen AI's data pillar, as illustrated in the visual below.

A Closer Look at The **Data Pillar of AI**

ACTIVANT VALUE CHAIN



Note: This is by no means an exhaustive list.

Tackling Unstructured Data

Given that approximately 94% of global data is unstructured, the ability to process and manage this data is crucial, particularly for companies aiming to leverage RAG models on proprietary datasets.[19] Modern applications increasingly depend on **unstructured data**, highlighting the need for tools like Unstructured, Jina AI, and Superlinked. These tools excel at breaking down content such as text, images, and videos into manageable components, transforming them into meaningful vector embeddings stored in databases like Pinecone or Weaviate.

Managing Complex Data Pipelines with DataOps

Models access both structured and unstructured data from diverse sources, therefore maintaining control and reliability is critical for data engineers. **DataOps** tools like Octopai facilitate discovery, categorization, and tracking, generating detailed data lineage to improve transparency and break down data silos. Telmai ensures rigorous quality checks across all data types, guaranteeing that data is accurate, consistent, and properly formatted for AI applications. Organizations could save an average of $12.9 million annually by using such tools to improve data quality and accuracy.[20]

Upholding smooth data flow and system transparency is equally important. Solutions like Acceldata provide real-time pipeline monitoring, quickly identifying and resolving performance bottlenecks before they escalate into larger issues while pipeline orchestrators such as Prefect automate task execution and coordinate complex workflows. Platforms like Dagster or Ascend unify these capabilities into a single interface, simplifying pipeline management and enhancing collaboration. These solutions not only streamline operations but also empower organizations to scale their AI initiatives effectively. Users of these tools have reported increases up to 10x in pipeline throughput and significant reductions in processing times.[21]

Stepping up Data Security

There is a clear need for increased **data security** measures. In 2022 alone, the average enterprise was targeted by more than 1,000 cyber-attacks each week – double the rate of 2020.[22] Scattered data and excessive access permissions create vulnerabilities and AI adoption further expands the attack surface.

Companies need effective solutions to identify sensitive data and determine where protection and governance are required. Vendors such as Normalyze, Skyflow, and Codified excel at discovering and securing critical information. For monitoring threats, tools like Code42 are indispensable, while solutions like Lakera provide targeted protection against vulnerabilities introduced by AI. Much like DataOps tools, these data security functions can be streamlined and centrally managed through platforms like Securiti.

By embracing these advanced data tools, over and above the core data stack, companies can transform data from a bottleneck into a catalyst for innovation.

Looking Forward: The Next Phase of AI

Compute and software have propelled AI to where it is today and will continue to play a significant role, but without equal investment in data, we're approaching a ceiling. Databricks' $10 billion funding round was a long-overdue acknowledgment of data's critical importance. However, the gap in funding between data, compute, and software is still significant. Companies that recognize this imbalance and act quickly won't just stay relevant – they have the opportunity to redefine the AI landscape by building data-centric ecosystems that set new industry benchmarks. This shift will usher in a new phase of AI deployment, set to drive the data infrastructure market toward its $288 billion potential.[23]

End Notes

[1] OpenAI, New funding to scale the benefits of AI, 2024

[2] BBC News, Why is Nvidia boss the "Taylor Swift of tech"?, 2024

[3] Google DeepMind, Demis Hassabis and John Jumper were awarded Nobel Prize in Chemistry, 2024

[4] Gartner, Hype Cycle for Generative AI, 2023

[5] Crunchbase News, Databricks Raises $10B In 2024's Largest Venture Funding Deal, 2024

[6] Activant Analysis, Total Market Capitalization of Leading Compute and Cloud Service Providers, 2025

[7] Activant Analysis, Total Market Capitalization of Top AI Model & Software Companies, 2025

[8] Activant Analysis, Total Market Capitalization of Leading Data and Analytics Companies, 2025

[9] Dealroom.co, Industries fundings heatmap, 2025

[10] TechCrunch, Generative AI funding reached new heights in 2024, 2025

[11] OpenAI, Better language models and their implications, 2019

[12] McKinsey, The data dividend: Fueling generative AI, 2023

[13] McKinsey, The state of AI in early 2024: Gen AI adoption spikes and starts to generate value, 2024

[14] Gartner, Gen AI Seminar, 2024

[15] IDC, IDC WW Global Datasphere Structured & Unstructured Data Forecast, 2022 - 2026

[16] Bloomberg, OpenAI, Google and Anthropic Are Struggling to Build More Advanced AI, 2024

[17] AMD, Advanced AI, 2023

[18] Nomura, Anchor Report: Global Markets Research, 2023

[19] IDC, IDC WW GlobalDatasphere Structured & Unstructured Data Forecast, 2022 - 2026

[20] Gartner, How to Improve Your Data Quality, 2021

[21] Enterprise Strategy Group, Analyzing the Economic Impact of Ascend Data Automation, 2023

[22] Wall Street Journal, Why even big tech companies keep getting hacked – and what they plan to do about it, 2022

[23] IDC, Worldwide Big Data and Analytics Software Market, 2023