EBOOK

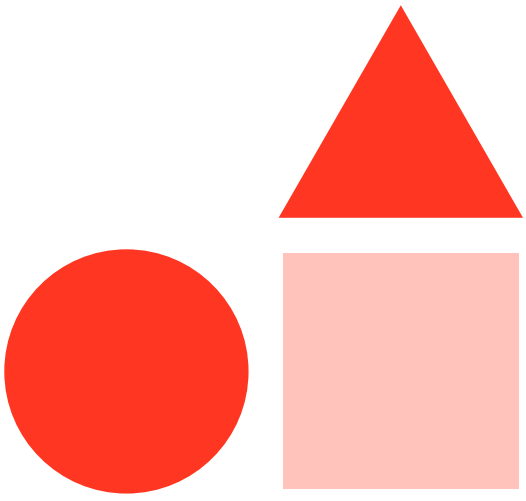
databricks × Google Cloud

How your legacy data can secure your AI future

A new approach to data intelligence

databricks
×
Google Cloud

Contents



Introduction

Artificial intelligence is redrawing the software landscape, and it is those companies with rich reserves of data that are best placed to navigate it.

But they can only do that if they are able to make that data accessible to today's cutting-edge AI tools and large language models.

The catch is that legacy data infrastructures were often built for the big data era and are disparate, complex, and often still on-prem. This is compounded when organizations grow through mergers and acquisitions, bringing on board additional data stacks and teams. This introduces more complexity, as different teams may have developed distinct approaches to data management—sometimes even resorting to shadow IT to achieve their goals.

As a result, many organizations have fragmented data estates, making it difficult to navigate and manage their assets. This can also lead to elevated costs, raise significant concerns about data reliability and security, and contribute to vendor lock-in for critical data and applications.

Many organizations have fragmented data estates... this can lead to elevated costs, raise significant concerns about data reliability and security, and contribute to vendor lock-in

How this impacts AI projects

A fragmented data estate, while already challenging for traditional large-scale projects, becomes a significant hurdle for the dynamic data and AI workflows that modern, AI-infused applications demand. Even more so, when it comes to the new wave of agentic AI use cases. Failing to address this can lead to AI initiatives struggling to reach full production or achieve their intended scale.

Building a robust AI strategy hinges on unifying disparate data and streamlining workflows. This integration is essential, as AI systems require diverse data formats, from structured to unstructured, to unlock their full capabilities.

AI systems require diverse data formats, from structured and unstructured

Large parts of older data estates will run as batch operations. This might be suitable for daily or weekly updates to traditional business intelligence systems. However, when it comes to modern AI systems operating in real time, preserving data integrity is imperative.

Further, CTOs face growing pressure around compliance and security when it comes to their organization's data. Rushed implementations can present data leak worries, while LLMs and AI applications represent a larger attack surface. Concerns over data and AI sovereignty compound the complexities around the movement of data and models.

This is before they face the challenge of evaluating the models, applications, and agents they are producing. Ongoing, iterative evaluation is critical from the onset, both when models and applications are being developed, and once they go live.

And fears over vendor lock-in mean both CTOs and data professionals worry they will be unable to fully exploit either their own data or upcoming innovations in AI tooling or models. This could mean losing their edge or ceding ground to their competition.

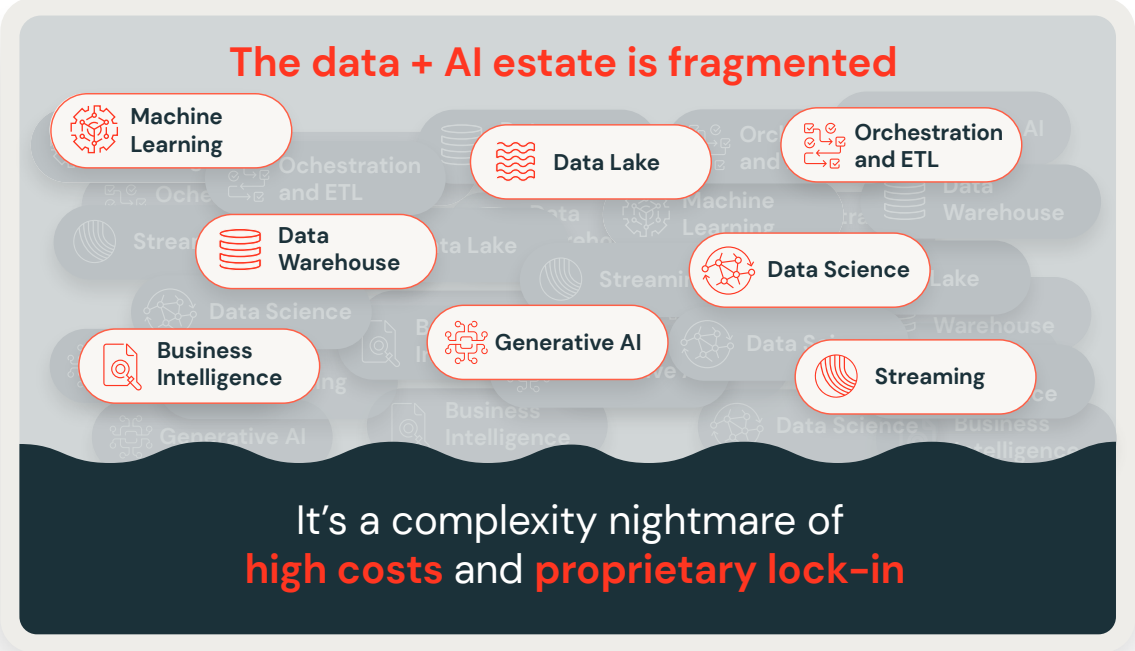
The impact is stark. Research by RAND highlighted five major reasons for AI project failure, two of which were unequivocally data related:

- The lack of necessary data to train an AI model.
- The lack of "adequate infrastructure to manage their data and deploy completed AI models.

Gartner recently predicted that 40 percent of agentic AI projects will be cancelled by the end of 2027, with hype blinding organizations to the "real cost and complexity" of deploying agents at scale. The key to successful implementation is for organizations to rethink workflows with agentic AI from the ground up, the organization advised.

Gartner says the key to successful implementation is for organizations to rethink workflows with agentic AI from the ground up

The fragmented data landscape



What is a data platform approach?

The rise of AI means that traditional data stacks, often fragmented with multiple overlapping systems, need to evolve to meet its demanding requirements. A unified platform approach provides the essential foundation to optimize for AI's capabilities and unlock its full potential.

Forrester, for example, defines a platform as "a product that supports the creation and/or delivery of other products."

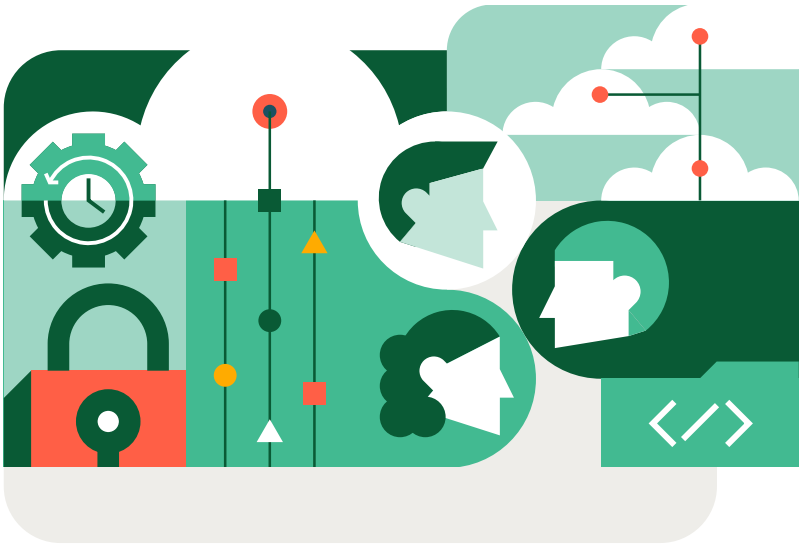
In the case of data platforms, this suggests an integrated approach that spans data, analytics, and AI, on a resilient secure cloud platform where appropriate.

This centralizes management, security, and governance, with data resources made available through a catalog. This provides a map of what data an organization has, and an overview of who is accessing a given piece of data.

It also equips data professionals with the tools, data, and models they need to build data pipelines and get projects and applications into production. Non-data specialists and business users also get access to data resources and tooling where appropriate.

Additionally, it provides CIOs and senior executives with the assurance that their data and applications are secure, compliant, and scalable while TCO is manageable and predictable.

Databricks redefined what a data platform can be with its whitepaper detailing the lakehouse architecture back in 2020.



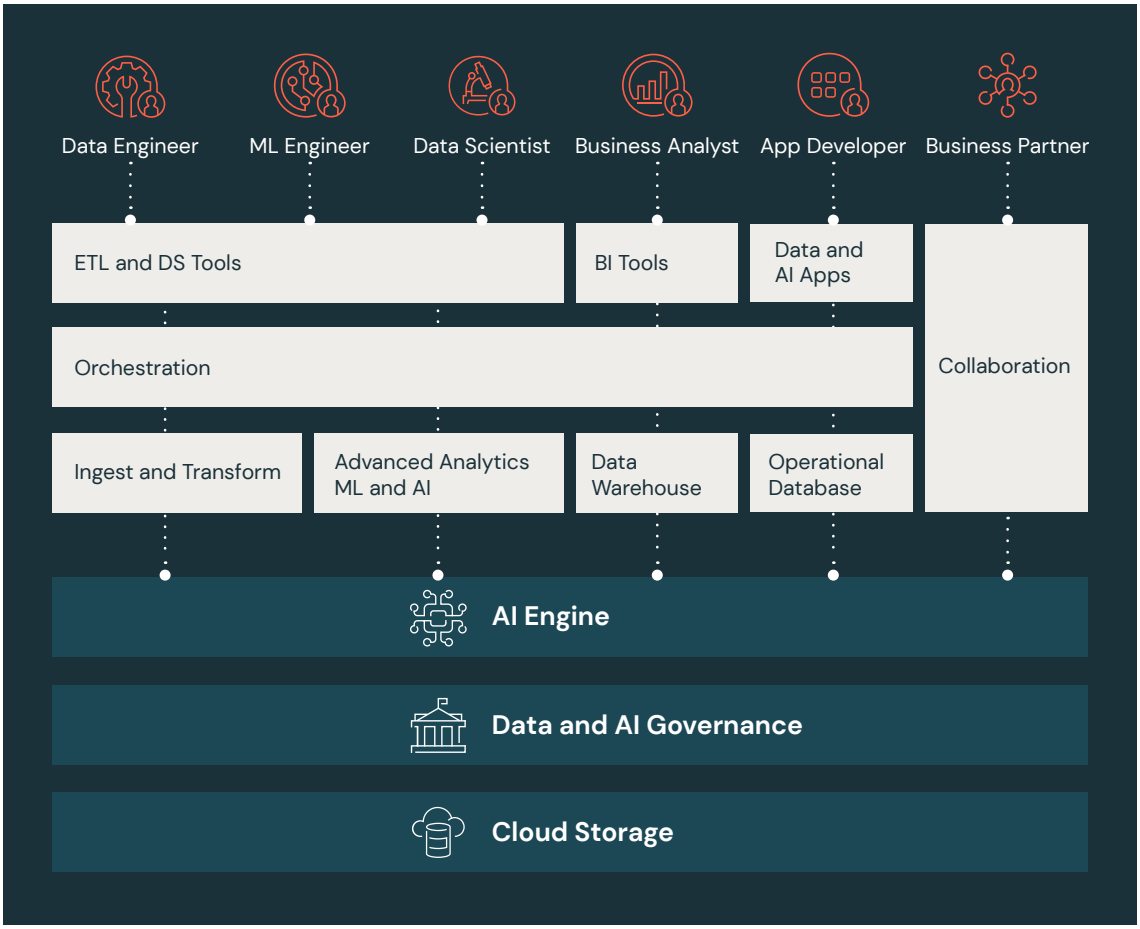
Databricks redefined what a data platform can be

This tackled the contemporary problems around data lakes and warehouses. These included:

- **Multiple rounds of ETL operations** – from operational systems into lakes and again into warehouses – which led to reliability problems and overburdened data engineers.
- **Data stagnation**, as data specialists waited for information to update, and data lock-in due to proprietary formats.
- **Limited support** for increasingly advanced analytics and machine learning, something that has become only more pressing with advances in AI in recent years.
- **The cost of ownership** challenges these increasingly complex systems presented, especially as data moved onto and around the cloud, attracting pay per use, ingress and egress fees, and ballooning storage costs as data is generated and duplicated.

Databricks proposed an architecture that combined the low-cost storage and open formats of data lakes and warehouses with the powerful management and optimization features of traditional analytical DBMS systems. By maintaining the separation of compute and storage, the architecture provided a good fit with cloud environments.

The key components of a data platform



Design Principles of a future proof data platform

When it first detailed the lakehouse architecture, Databricks noted that "machine learning APIs are quickly evolving." That evolution has accelerated to a previously unthinkable degree, with the technology world awestruck by generative AI and more recently, agentic AI, epitomized by Google's Gemini family of Large Language Models.

These innovations have further increased the pressure on traditional data estates, raising the stakes for data platforms. It is even more important that technology leaders and data specialists think carefully about how they ensure their data platforms are not just ready for today's AI powered world, but that to come.

Governance

AI, particularly agentic AI, presents novel challenges when it comes to governance and security.

The need for data to underpin the development of AI applications and training of models is a given. That data may take many forms, from traditional structured data, unstructured formats, and metadata.

This means data engineers and application developers will be drawing data from multiple sources. This data will then find its way into multiple projects, notebooks, and dashboards, and eventually, applications and agents.

However, it could also find its way into unexpected places, such as large language models.

Agentic AI poses particular concerns. Autonomous agents might access multiple data stores repeatedly, or deal with real-time data streams. Tracking that usage and ensuring it meets compliance and legal requirements throughout the agent lifecycle is essential. The challenges around audit logging, monitoring, and lineage are all amplified by the transition to agentic AI, nowhere more so than in highly regulated industries such as finance and healthcare.

Governance then becomes more than simply controlling initial access to specific data resources. It must extend beyond the confines of the traditional data catalog, to all the places data might be located or put to use, giving full observability throughout the lifecycle of projects.

Agentic AI poses particular concerns

Openness

The lakehouse architecture is predicated on open data formats, to avoid the danger of lock-in, ensure data storage and movements costs are manageable, and to maximize the range of use cases organizations' data can be leveraged for.

Openness becomes even more important when it comes to gen AI and agentic AI. Enterprises want to exploit the full array of cutting-edge large language models. Agent creation will require data, but agents themselves will need ongoing access to data and will themselves change or create data.

Open formats like Delta Lake and Iceberg are central to enabling this.

Likewise, customers should consider how open their underlying cloud infrastructure really is. There can be a world of difference between a cloud provider that tolerates open formats, and one that actively encourages and nurtures openness.

More broadly, demands for data across the organization and across functions will only increase. Business users will increasingly need direct self-serve access to data. An open approach makes this much easier to deliver and govern.

Organizations must keep on top of the rapid evolution of AI models and tooling. They cannot afford to let restrictive data formats put a brake on their adoption of the next innovation.

Optionality

Data teams and their business counterparts have a dizzying array of AI tools and models to choose from, with new advances and innovations appearing almost daily.

No organization wants to miss out on innovation, or the chance to better serve their customers or gain a business edge on their competition.

So, it is critical that they retain optionality over the tools they have access to, and where they run them. Data teams must have access to the best tools, models, and data, for a given project.

But optionality also means having the ability to keep data and projects in the optimal location. Sovereignty or compliance concerns might dictate whether data is held on prem, or in a given cloud region.

Likewise, latency concerns will become a factor as applications operate in real time. Some applications, such as inference, might be better handled at the edge, closer to customers.

This will have implications for resilience and cost, as projects scale.

Automation

AI is not just an outcome for data engineers. It will become increasingly central to data systems and workflows themselves. AI and natural language promise to simplify the curation and querying of enterprise data assets, allowing a broader range of professionals and functions to access them.

AI and natural language will simplify the curation and querying of enterprise data

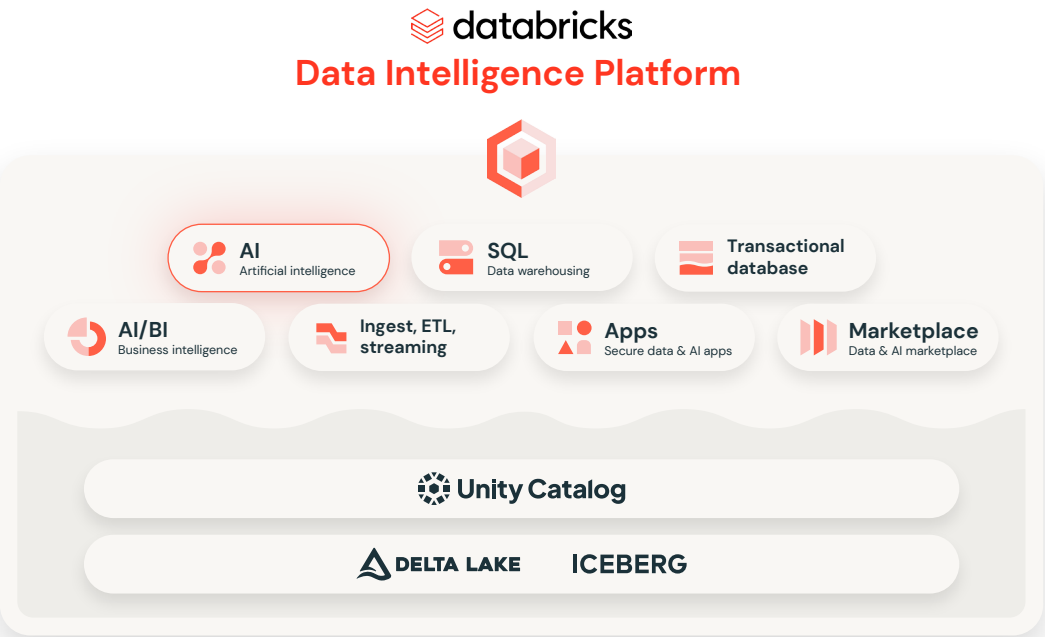
At a deeper level, artificial intelligence should take over much of the heavyweight data engineering that will still be necessary to build and manage the organization's data infrastructure, particularly the production of the business-ready gold tables needed for advanced analytics.

This will open up data resources beyond just a highly skilled cadre of data specialists. Those data specialists will, in turn, be able to focus on higher value or particularly challenging work.

Convergence

Traditionally, there was a sharp divide between operational data and analytics data. But modern AI systems need access to the full range of data assets. Agentic AI applications will demand deep analytics capabilities to make sense of the vast amounts of data within an enterprise. They will also need access to more structured assets, such as the live transactional data needed to make decisions in financial organizations, or recommendations for ecommerce. That means a transactional database engine becomes an essential part of the data intelligence platform and must meet the same requirements around governance, openness, and optionality.

What a future proof data platform looks like



How Databricks and Google deliver the optimal data platform

Google Cloud and Databricks offer the full range of data, analytics and AI tooling needed by data teams in particular, and enterprises in general.

Databricks on Google Cloud combines Databricks' lakehouse architecture and data tools with Google Cloud's leading-edge infrastructure and cutting-edge AI and data tooling, as well as Gemini, Google's multimodal portfolio of AI models, all in a unified platform for analytics and AI workloads.

Databricks' Data Intelligence Platform takes a lakehouse approach, providing a unified foundation for data and governance, and tunes itself to an organization's unique data. This includes using natural language to help both data professionals and non-specialists navigate, discover and query data, and put it to work.

Its centralized governance, via the Unity Catalog, doesn't stop at data access and discovery, but extends out through the full data lifecycle, including AI agent data use. It offers agent and model evaluation based on the Mosaic AI Agent Framework Agent Evaluation, which automatically collates and grades feedback to inform further tuning and experimentation.

Databricks' approach is open from the outset, focused on formats such as Delta, Apache Iceberg, Hudi and Parquet.

Enterprises benefit from deep integration with Google's cloud-native data tools

databricks × Google Cloud

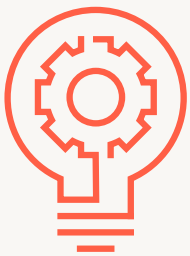
With Databricks on Google Cloud, the possibilities are extended further. The collaboration ensures seamless integration between Databricks and Google Cloud's infrastructure and services, including Google Cloud Storage.

Google Cloud is equally committed to openness, both in its own tooling and by supporting open data formats.

Google Cloud also delivers distinct infrastructure advantages as well, thanks to its custom Axion processors. This ARM-based architecture uses up to 60 percent less energy than other architectures on Google Cloud. It delivers a performance boost on Databricks of up to 50 percent compared to other instances while boosting ROI for every dollar spent by 50 to 60 percent.

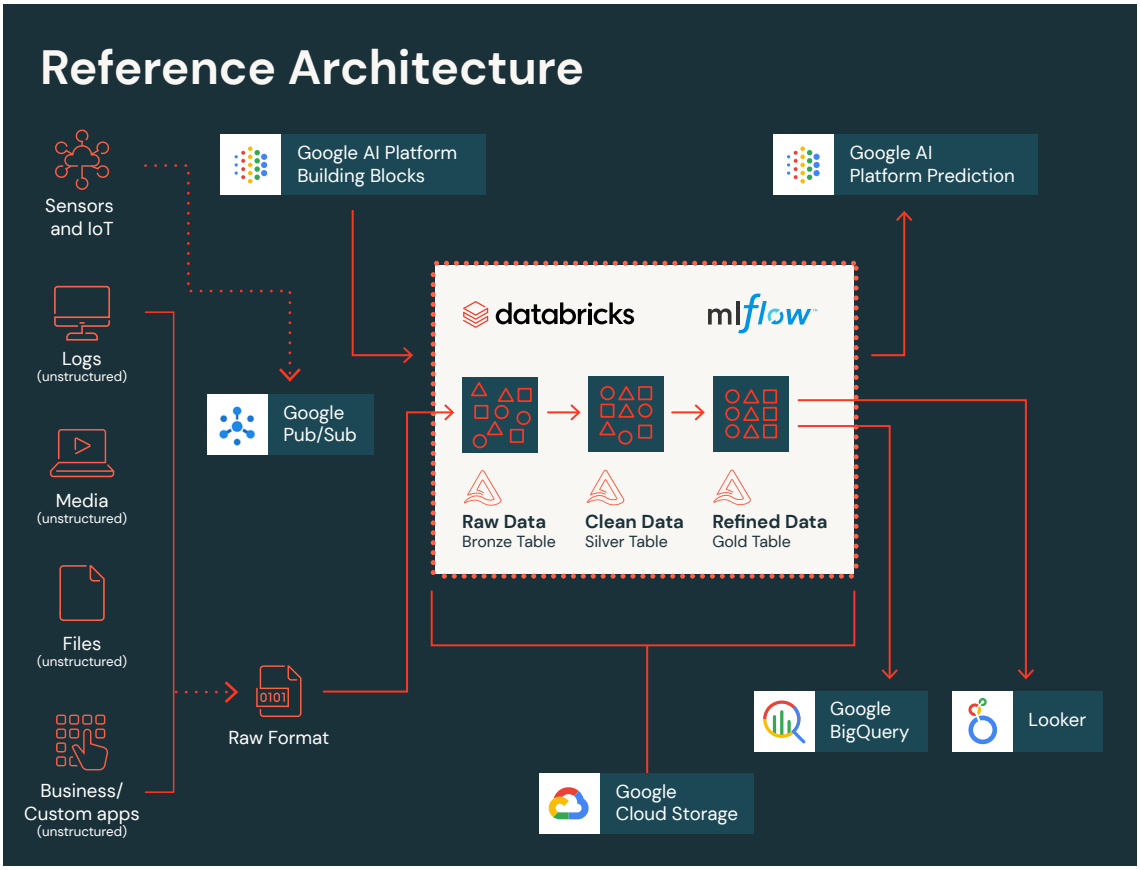
Enterprises also benefit from deep integration with Google Cloud's cloud-native data tools – including BigQuery, Vertex AI, and Looker. At the same time, they can quickly adopt Gemini, Google's family of advanced multimodal models that span the full range of tasks and platforms, with responsible development and deployment at its core.

This means teams can innovate quickly without the risk of ongoing lock-in, both at a tooling and data or infrastructure level. So, enterprises can use their data to build state-of-the-art AI applications and services today and take advantage of new innovations as they break cover.



Teams can innovate quickly without the risk of lock-in

The key components offered by the Databricks/Google data platform



Takeaway

The pace of AI innovation is speeding up. Enterprises are racing to stay on top of AI developments and prepare for whatever comes next.

Traditional data approaches built for a more static world are just not able to keep up.

A data platform approach helps enterprises get today's applications into production, quickly and securely.

But the onset of agentic AI is upping the stakes, raising new questions about governance, data access, and TCO.

Databricks on Google Cloud allows businesses to reconfigure their data operations for this new world, focusing on governance, optionality, and openness, so they can adopt new AI innovations and models quickly, securely and cost effectively.

Get Started with a Free 14-day Trial of Databricks on Google Cloud.

START YOUR FREE TRIAL