3.25.25 AI & ML Data

Roadmap: Data 3.0 in the Lakehouse Era

We're moving beyond the modern data stack as a new architectural revolution sweeps the infrastructure landscape, ushering in unprecedented interoperability for enterprises in an AI age.

Contributors

 Janelle Teng

 Lauri Moore

Enterprise data infrastructure is simultaneously a call and response to every technological shift — it both enables new products and businesses, while simultaneously evolving to support the demands created by these same innovations.

Over the last fifty years, we've progressed from traditional on-premise data warehouses to cloud-native data warehouses and data lakes. Today, we're at an exciting inflection point for the landscape as we're evolving quickly past the modern data stack (which was the premise of our Data Infrastructure roadmap from 2021) due to multiple catalysts that are ushering in a Data 3.0 era.

For one, as we noted last year, AI's proliferation has led to profound changes within the AI infrastructure landscape. But in the midst of this major technological shift, another tectonic transformation is afoot. The very core of enterprise data infrastructure is being reimagined due to the impact of a revolutionary architectural paradigm—the data lakehouse—which supports multiple use cases, from analytics to AI workloads, in a powerful, interoperable platform.

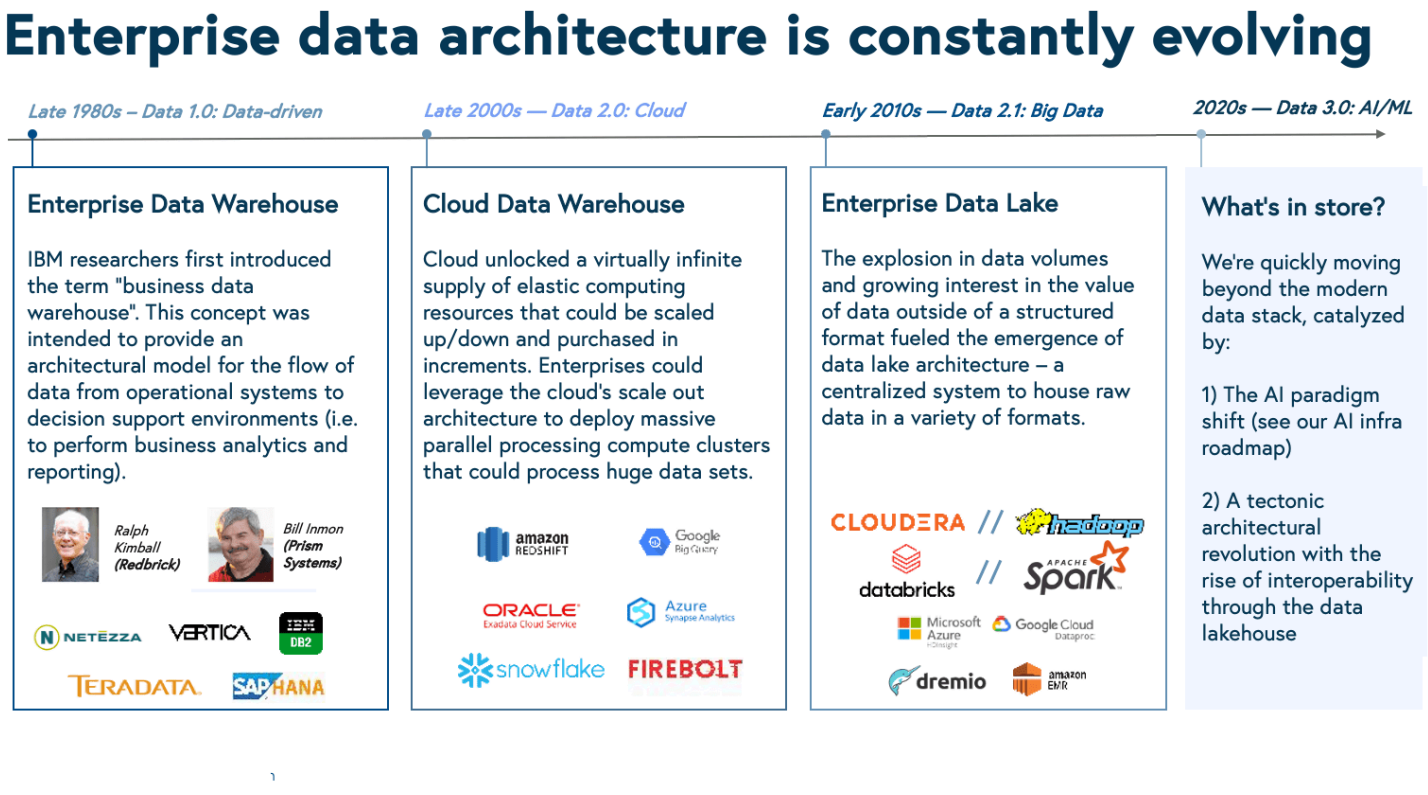
The lakehouse paradigm doesn't just represent a marginal improvement to the architectures that came before it. Rather, it is a radical transformation that will bring forth an era of unprecedented interoperability and set the stage for the next wave of multi-billion-dollar data infrastructure giants to emerge.

The lead-up to the lakehouse revolution

Enterprise investment in data infrastructure has nearly doubled in just five years—growing from $180 billion in 2019 to $350 billion in 2024. Hyperscalers like AWS, Microsoft Azure, and Google Cloud are pouring trillions into AI-optimized infrastructure, while major players like Databricks have reached profitability, signaling a mature yet still rapidly evolving market.

Yet, while the space has progressed significantly, enterprises are recognizing that the current cohort of data infrastructure architectures cannot keep pace with the AI revolution's scale, speed, and complexity. This is fueling a burgeoning demand for a new architectural standard to rise.

Enterprise data architecture is a call and response to innovation



By the end of the 2010s, as we transitioned from Data 1.0 to Data 2.0, two dominant architectural paradigms had emerged: the cloud data warehouse and the data lake. While powerful, each of these have limitations (both in terms of capabilities as well as user experience). These shortfalls are now being critically exposed as the current explosion of AI-powered applications and techniques pushes these architectures beyond their limits:

AI applications and workloads require new abilities and capacities that many legacy infrastructure technologies have yet to support:

- In an AI-first reality, enterprises need infrastructure that supports both structured, semi-structured, and unstructured data, with high performance, scalability, and governance.
  - This is an evolution from previous paradigms where unstructured data was largely untapped or viewed as more of a "supporting cast" without a real usage plan. Enterprises have now internalized the increasing importance of this type of data.
  - Furthermore, unstructured data is being generated at unprecedented speed in the AI era. Just think of the growing tsunami of AI-generated text and visual content that has been created by ChatGPT's 100M monthly active users.
  - The line between data types continues to blur as AI-powered products are hungry for access to all types of data in order to strengthen capabilities.
- AI workloads also often require real-time, multimodal, and composable data processing which traditional architecture is not purpose-built for.
- New agentic technology is pushing on the logic frontier for reasoning.
- There is an increasing need for interoperability between modern data stack infrastructure (such as "traditional" databases) and new AI-infrastructure tools (such as vector databases).

In this new context, the demands from enterprise customers have evolved as well:

Data redundancy (often from copying data from data lake into data warehouse) is an unnecessary expense that many enterprises are seeking to resolve.

Governance is becoming more challenging for various reasons:

- More data silos are being added, which is not just a cost issue, but also poses a governance concern.
- Data that was previously mostly used internally is now more likely to be used in product through AI models, expanding the "walled gardens" of what data governance was previously focused on.
- Most AI/ML workflows consume directly from the data lake which, unlike warehouses, usually does not have baked-in governance

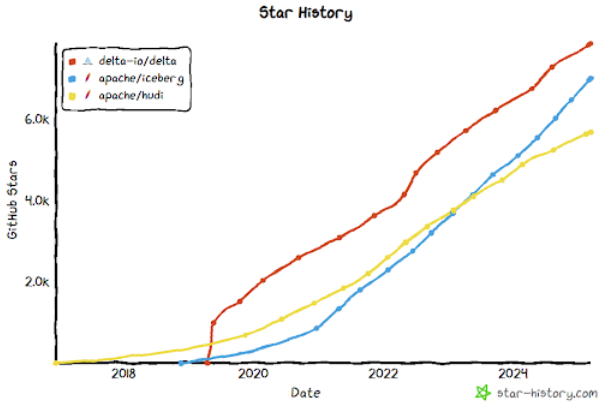
Organizations are frustrated by vendor lock-in. This has always been a recurring concern even through prior paradigm shifts, but has been further emphasized in the current AI era in part due to high costs of existing services (many priced during ZIRP years) and also because traditional infrastructure providers may not be adapting to AI workloads fast enough.

In all the commotion and excitement around AI, it has been increasingly noisy for AI application developers and machine learning engineers to recognize new solutions that are the best fit for their needs.

Each chapter of data infrastructure not only introduces new capabilities, but also reveals limitations that invite future waves of innovation and technological compounding. The dynamically changing tech and customer landscape compelled technologists to begin experimenting to find a viable successor to the warehouse and data lake, given the clear need for a higher standard of technology that could naturally capture the best of both worlds in a single architecture.

Chatter around the unbundling of the data warehouse (aka the Deconstructured Database) began in as early as the late 2010s. New innovations around open table formats (such as Delta Lake, Iceberg, and Hudi), and associated open standards for storage (such as Parquet, ORC) and data access (such as Arrow) began emerging during this time.

Modern open table formats are unlocking the full potential of the lakehouse



It would be remiss of us not to recognize the groundbreaking generation of open table formats (OTFs) including Delta Lake, Iceberg, and Hudi, that has unleashed the power of the lakehouse. These formats are an abstraction of underlying files, presenting them as a single "table" that users can then access through an API. These formats enable advanced capabilities, such as:

ACID compliance. Ensuring data integrity by preventing corruption, inconsistencies, and incomplete transactions:

- **Atomicity:** Transactions are fully completed or not executed at all.
- **Consistency:** Data changes follow strict integrity constraints.
- **Isolation:** Multiple transactions run concurrently without conflicts.
- **Durability:** Successful transactions persist even after system failures.

Supporting both batch and streaming pipelines. Allowing real-time and large-scale processing.

Allowing for schema and partition evolution. Enabling seamless adjustments to data structures without downtime.

Time travel functionality. Letting users roll back to previous states for debugging and auditing.

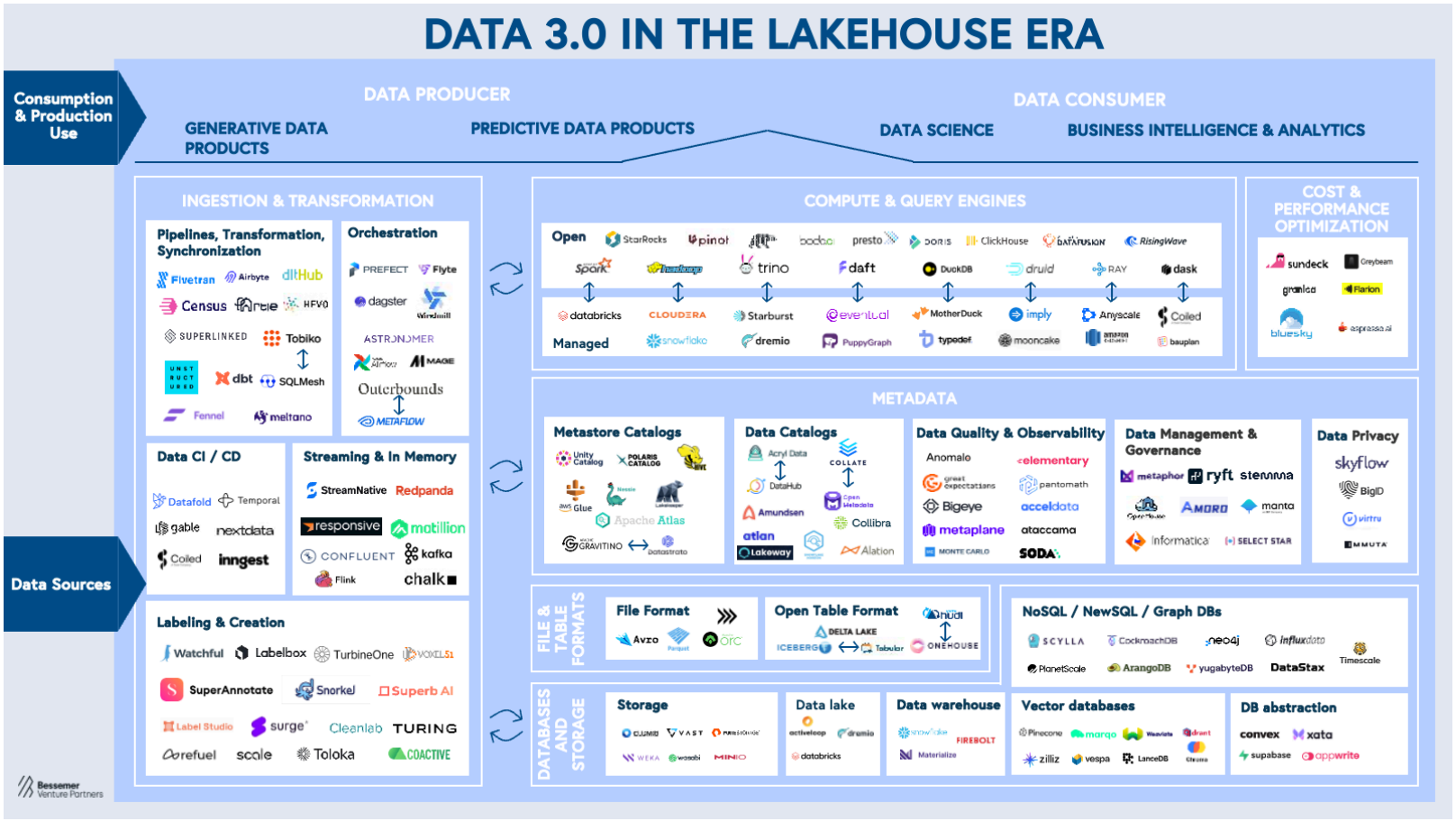
Scalable metadata management. Improving performance and query efficiency.

Here, we'd like to give recognition to Iceberg for having a phenomenal year in 2024, including Tabular being acquired by Databricks in a blockbuster deal.

At the tip of the "Iceberg" for innovation in this new architectural paradigm

By the early 2020s, these efforts ultimately led to the birth of a new architectural concept—the data lakehouse. The lakehouse paradigm isn't just an upgrade—it's a foundational shift that rewires how data infrastructure operates. By combining the refined power of data warehouses with the flexibility of data lakes, it unlocks a level of interoperability we've never seen before. This isn't just about better performance; it's about enabling entirely new capabilities in AI-driven applications, real-time analytics, and enterprise intelligence.

This phenomenon has now taken the industry by storm, with major market players, enterprises, and startups such as Bessemer portfolio company TRM Labs, all racing to embrace this new paradigm. As the market coalesces around the lakehouse shift, the opportunity to back category-defining infrastructure companies has never been greater. We see a massive opportunity for bold founders to both pioneer brand new categories within the lakehouse stack and also disrupt existing infrastructure categories.



Note: We recommend digesting our Data 3.0 roadmap in conjunction with our AI Infrastructure roadmap since we believe that Data and AI are two sides of the same coin.

With the right mix of technical innovation and go-to-market strategy, the post-modern-data-stack Data 3.0 ecosystem is primed for the rise of net new multi-billion-dollar infrastructure platforms. **We have four major theses around how this space will develop in the coming years.**

Thesis 1: AI-native ingestion and transformation will power real-time, scalable, and intelligent data pipelines

Data pipelines, extract transform load (ETL) tools, and orchestrators are undergoing a fundamental shift in Data 3.0. Increasingly, these tools are serving user-facing production use cases, as opposed to internal deployments with a higher tolerance for failure scenarios, drawing on dynamic data stores, all of which need to be observed, secured and governed.

As AI-driven data workflows increase in scale and become more complex, modern data stack tools such as drag-and-drop ETL solutions are too brittle, expensive, and inefficient, for dealing with the higher volume and scale of pipeline and orchestration approaches. Instead, products like Prefect, Windmill and dltHub are challenging the status quo with code-native data transport, pipeline orchestration, scheduling, and monitoring—all of which are critical for task execution and managing production-scale data pipelines. A code-native approach is also key in an AI era as such tools can become building blocks and components to turn manual workflows into agentic ones. Others on the transformation side like Tobiko (from the creators of SQLMesh) allow users to move data, automate SQL queries, track metrics, and view lineage and dependencies within its interface.

Furthermore, traditional data infrastructure requires users to manually manage context across different queries and transformations, leading to inefficient workflows and inconsistent results. While tedious in a previous paradigm, this becomes impractical if not impossible in the context of AI workloads. Bessemer portfolio company Anthropic recently introduced Model Context Protocol (MCP), which significantly advances in how AI systems integrate with the Data 3.0 ecosystem. MCP does this by providing a standardized framework for context-aware AI interactions that preserves relationships between queries, transformations, and output—while maintaining governance and security standards. This could unlock a future full of agentic innovations built on top of code-native infrastructure for use cases such as automating repetitive, low-context work.

While batch engines have dominated previous architectural paradigms and will continue to play an important role in the ecosystem, we also anticipate that data processing will progressively "shift left" to be closer to the time of action. Apache Kafka has served dutifully as the foundational messaging backbone for ML architectures, enabling reliable data movement between applications and AI systems. Its distributed log architecture provides the durability, scalability, and exactly-once processing guarantees essential for mission-critical AI operations. Apache Flink complements this ecosystem by providing advanced stateful stream processing capabilities needed for complex ML pipelines, with its unified batch and streaming model. This model has proved particularly valuable for continuous model training and inference workloads, but may be too complex to manage directly for most enterprises.

As organizations mature their AI capabilities, we'll see the emergence of specialized stream processing patterns optimized for model serving and continuous learning pipelines that blend historical batch data with real-time signals for more accurate predictions and recommendations. Companies like Chalk provide real-time inference as a managed platform, which is especially critical for companies that need to make instant, data-rich decisions, say for credit approvals.

This evolution toward more streaming-first architectures represents a paradigm shift in how enterprises architect their data platforms to support AI-driven decision making at scale and we see a future of stronger co-existence between different styles of engines (see Thesis 4).

Thesis 2: We need "data about the data" — the metadata layer will move front and center

The metadata layer is emerging in importance as a strategic frontier in lakehouse architecture, serving as the core layer that governs how data is understood, discovered, accessed, and optimized. While metadata tools have long existed in different chapters of enterprise data infrastructure, we believe that a new generation of metadata tools will be greatly elevated in significance and impact:

> In previous architectural paradigms, metadata products were important as a "reflection of truth" as data updated into this layer indirectly and asynchronously. In the Data 3.0 era, metadata products are now seen as the direct "source of truth" of schemas and definitions.
>
> The AI era has brought shifts in underlying file structures, blending of warehouse and raw data, and a importantly a growing need to actively manage metadata through actions (not just in a read-only context)

Unsurprisingly, everyone from startups to incumbents are racing to establish standards in the metadata layer. As we've highlighted earlier, a pioneering cohort of modern open table formats has emerged to define this frontier. Large cloud and data infrastructure players also recognize metadata as a strategic priority – they are investing aggressively in new innovations, launching proprietary catalog products, and open-sourcing key components in this layer to drive further compute adoption.

Within this layer, a new generation of lakehouse-native data catalogs such as Datastrato (by the creators of Gravitino) and Vakamo (by the creators of Lakekeeper), have sprouted up since data catalogs are essential in the lakehouse paradigm.

Catalogs provide an organized inventory of data assets, enabling efficient discovery, governance, and optimization. Governance and compliance is also much more integrated and precise in the Data 3.0 era since the metadata layer is now the "source of truth". Companies may be asked at a moment's notice to produce lineage and access records not just for employees, but agents. Platforms like Acryl Data (creators of DataHub) provide a unified data catalog for discovery and lineage while further enabling control and visibility around human, or agent data access across both lakehouse and traditional architectures.

Furthermore, the metadata layer is now critical for taking action and orchestration. Features like caching and data versioning are increasingly vital for AI workloads, meaning metadata management is becoming a core enabler of high-performance, efficient, AI-native data infrastructure. Solutions across open- and closed-source initiatives such as OpenHouse, Apache Amoro, and Bessemer portfolio company Ryft are positioned to provide a "control plane" for enterprises to handle data management challenges effectively.

Related to actionable metadata, another interesting theme we are witnessing within this layer is the emergence of new optimization tools. Optimization is not necessarily a new phenomenon within the data infrastructure landscape as enterprise customers are always mindful of cost and latency concerns across their stack. However, from Flarion.io to Greybeam, we're now seeing teams innovate on primitives outside of the underlying storage layer – allowing organizations to make data processing in the AI era more cost-, time-, and resource-efficient.

Thesis 3: The landscape of compute and query engines will be reshaped

Although known to be an "established" infrastructure category, the compute and query layer is undergoing major transformation in Data 3.0, driven by the rise of lakehouse architectures, the AI revolution, and growing demand for interoperability. In Thesis 1, we've already painted a picture of a future where there could be stronger co-existence between batch and streaming engines. This represents just one potential area of transformation.

While giants like Snowflake and Databricks currently dominate this market—each generating over $3 billion in annual revenue—we expect best-in-breed, AI-native startups to gain ground once a previously unheard-of level of interoperability is unlocked. Lakehouses are greatly reducing vendor lock-in, reducing reliance on monolithic systems and enabling unbundled, best-in-class solutions. This shift is fueling the rise of solutions like DuckDB for local development, ClickHouse and Druid for real-time analytics, and Daft and typedef to optimize AI-driven workloads effortlessly. There are even entrants such as Mooncake and Bauplan building Iceberg-opinionated platforms from the ground up.

This transformation is not just about incremental improvements—it represents a fundamental rethinking of how data is processed, optimized, and queried in AI-driven environments. While Spark and Ray remain the lingua franca of the AI engineering world, next generation compute frameworks that optimize for AI-first workloads are beginning to sprout from different angles, including the emergence of AI-optimized query engines, federated compute platforms, and Iceberg-native solutions.

Thesis 4: The line between Data Engineering & Software Engineering will get even blurrier

The Data 3.0 era will continue to see the breakdown of traditional boundaries between software engineering and data engineering disciplines. While software engineers have traditionally focused on application development and data engineers on data pipeline construction, data-intensive AI applications require data engineers to work in production environments and software engineers to work with data. AI is already pushing developers toward full-stack proficiency and organizations increasingly demand AI-stack developers—"AI Engineer" is the fastest growing job title in the US and 14 other countries, according to the 2025 LinkedIn Work Change Report.

Companies like dbt Labs have helped pioneer the democratization of data-driven development by bringing software engineering best practices—such as version control, testing, and CI/CD—to data workflows. Companies like Gable have created developer-friendly interfaces for data pipeline orchestration that abstract away infrastructure complexity while maintaining the flexibility software engineers expect. The need for closer to real-time data processing has similarly transformed how engineering teams approach data systems, with companies like Temporal and Inngest developing frameworks that bring reliability and observability to distributed data workflows, allowing engineers to handle complex failure scenarios with application-like constructs.

Historically, open source in data technologies lagged behind general software engineering, but this is no longer true. GitHub's 2024 State of the Octoverse report shows that contributions to data-focused repositories grew at a 37% CAGR since 2020, compared to 18% for general software repositories. Enterprise adoption of open source has also gone mainstream, with McKinsey reporting that the majority of enterprises are now using open source technologies for some AI or data initiatives—driven by cost-savings (in part thanks to simpler implementations), ease of security and compliance evaluations, and of course developer preference.

Furthermore, in the current AI wave, many enterprises are re-evaluating their technology choices on the basis of where they can receive the best LLM assistance. This has boosted the adoption of popular open source technologies that are well-understood by the LLMs since closed source technologies have less training data available and thus more difficult to get Copilot-like support.

Since software is driven by AI and AI is driven by data, the assembly line model of engineering no longer makes sense in 2025. Teams will depend on agile and open tools to collaborate in this new era.

We're at the precipice of Data 3.0 in the Lakehouse era

As we enter a new era of unprecedented interoperability, we are excited to support founders at the forefront of innovating in this new architectural paradigm—whether it's transforming legacy categories or pioneering new ones. At Bessemer, we believe this evolution represents one of the most compelling investment opportunities in enterprise infrastructure today.

We've spent over a decade partnering with category-defining data and AI infrastructure companies—including Auth0, HashiCorp, Imply, Twilio, and Zapier—so we recognize the unique challenges and opportunities data infrastructure founders face in building foundational technology. We are proud to support the next generation of data infrastructure startups with:

Access to renown advisors. Operating and technical advisors including experts such as Adam Fitzgerald (Head of Developer Relations at HashiCorp), Emilio Escobar (CISO at Datadog), Mike Gozzo (Chief Product and Technology Officer at Ada), Ryan Sepassi (Staff Software Engineer at Google AI), Solmaz Shahalizadeh (former Head of Data at Shopify), Talha Tariq (CIO & CSO at HashiCorp), Barak Turovsky (Chief AI Officer at General Motors)), and Tony Rodoni (former EVP at Salesforce).

Credit programs for savings. Unique access and credit programs to 100+ vendors across compute, cloud, software, and model APIs with aggregate potential savings of over $2M.

A welcome into the Bessemer community networks. Connect with Bessemer's 300+ active portfolio companies through our partnerships directory for business development and strategic partnerships.

Invite to exclusive events. Invite-only events, briefing sessions, and presentation opportunities with leading academics and business leaders.

Networking groups. Join us for community-specific learning and networking groups for functional leaders at infrastructure startups. For instance:

Exclusive Events for Tech & Engineering Leaders: Quarterly Council sessions, private networking dinners, tactical workshops, and specialized programming.

CTO Roundtables: Curated, small-group virtual discussions connecting over 400 CTOs and senior technical leaders across the Bessemer portfolio. These exclusive sessions provide a high-value forum to exchange ideas, gain fresh insights, and strengthen peer connections.

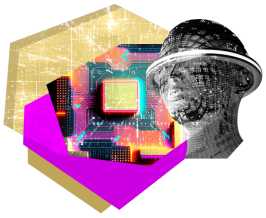
Extensive PR & Comms support. Communications expertise to advance PR goals, reach target audiences, and deliver press impact.

Access to our talent network. Find your next A-players with our talent network for infrastructure startups to build your GTM team and beyond.

We're eager to partner with and support the next generation of data infrastructure leaders; please reach out to Janelle Teng and Lauri Moore for further discussion.

Special thanks to Solmaz Shahalizadeh (BVP Operating Advisor & Former Head of Data at Shopify), Will Gaviria Rojas (Co-founder & Field CTO of Coactive AI), Renu Tewari (former Sr. Dir Data Infrastructure at LinkedIn), Yossi Reitblat (Co-founder and CEO of Ryft), and Wes McKinney (Co-Founder of Voltron Data & Principal Architect at Posit) for their feedback.

Recommended Articles



AI & ML

Roadmap: AI Infrastructure

Developer

How developer platforms can triumph over the three most common product-led growth roadblocks

AI & ML

Roadmap: Data Infrastructure