snowflake®

ESSENTIAL GUIDE

LAKEHOUSE ANALYTICS AND AI

Designing enterprise analytics for the new era of AI

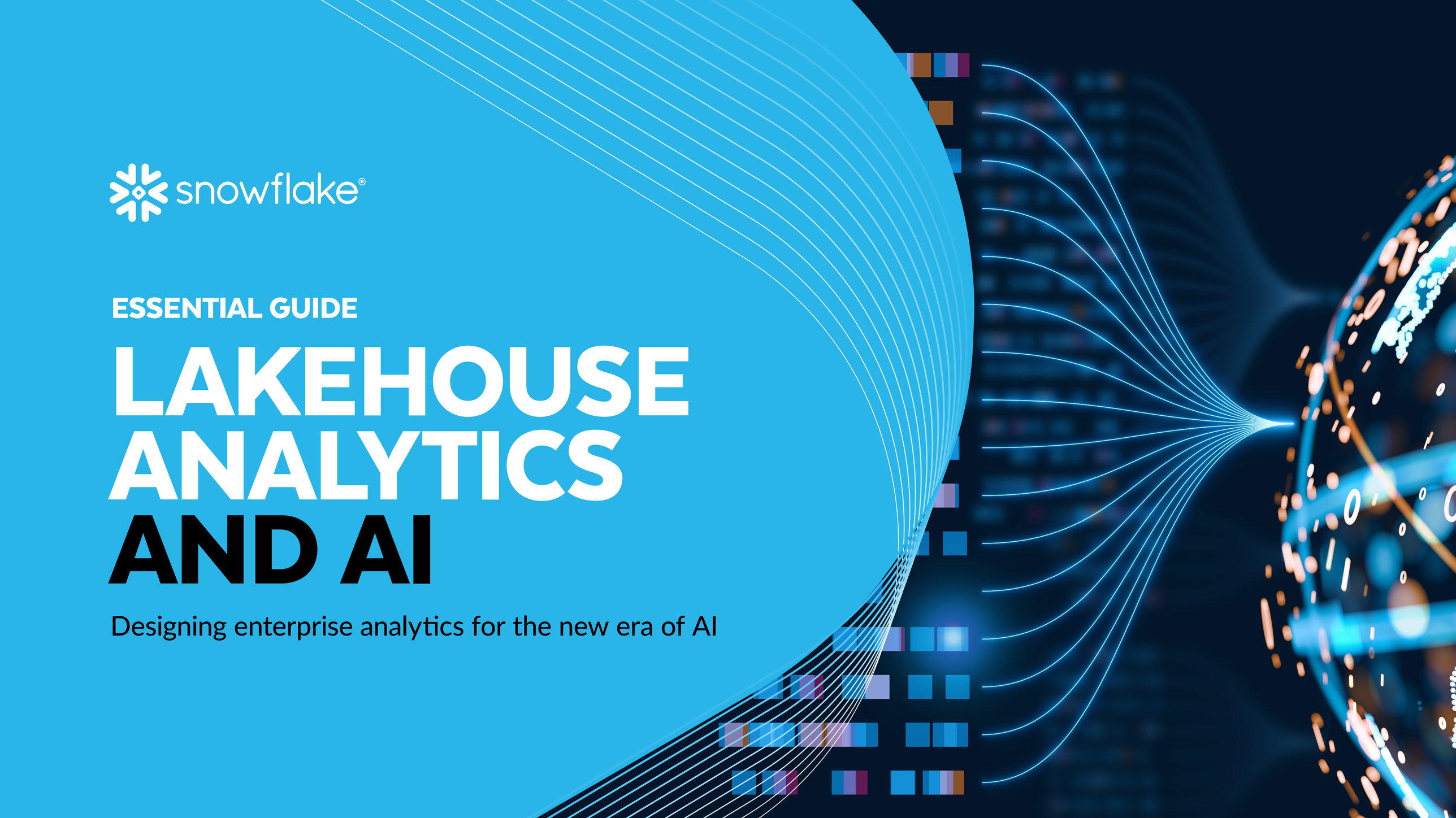


TABLE OF CONTENTS

THE IMPERATIVE OF LAKEHOUSE ANALYTICS IN THE AGE OF AI

For data leaders responsible for shaping their organization's future — architects, CIOs and CDOs — the strategic challenge is no longer just about managing data. It's about unifying a vast and varied data estate to power today's business intelligence and AI-driven innovation.

The answer for many forward-thinking organizations with an in-house data engineering team is the open lakehouse. This modern architectural approach promises to deliver the best of two worlds: the performance and governance of a traditional data warehouse combined with the flexibility and scale of a data lake.

> **A lakehouse architecture untangles storage, catalog and compute, providing the flexibility to choose the right tools for each team.**

A lakehouse architecture untangles storage, catalog and compute, providing the flexibility to choose the right tools for each team. The emergence of Apache Iceberg™ as the leading vendor-neutral and interoperable open table format has accelerated this trend by making it easier to bring tools to your data, rather than data to your tools. The result is greater data democratization by empowering organizations to rapidly adopt new tools, drive faster innovation, and scale analytics and AI initiatives all via a single copy of data and without being locked into specific vendors or complex architectures.

But adopting a lakehouse architecture is only the first step. To truly unlock its potential, you must power it with an analytics platform that can meet the demands of the AI era. This guide introduces a strategic framework for evaluating a lakehouse analytics solution that delivers on the promise of openness without compromising on performance, security or reliability. It is designed to help you make sense of shifting requirements, understand what a world-class solution looks like, and frame a productive conversation with your internal stakeholders.

THE OPEN LAKEHOUSE: STORAGE, CATALOG AND COMPUTE

At its core, a lakehouse architecture is defined by the separation of three key components: storage, catalog and compute. Understanding how these layers interact when standardizing on Iceberg tables is fundamental to building a flexible and powerful data foundation.

The storage layer

The storage layer is the foundation of the lakehouse, using low-cost, highly scalable cloud object storage (like Amazon S3, Google Cloud Storage or Azure Data Lake Storage) to hold all data — structured, semistructured and unstructured — in its raw or transformed state. Apache Iceberg™ is the leading open table format for semistructured and structured data, delivering critical capabilities like schema evolution, partitioning and transaction management. Its broad support across engines and tools provides the essential foundational flexibility that allows you to select the right catalog and compute layers for your lakehouse architecture.

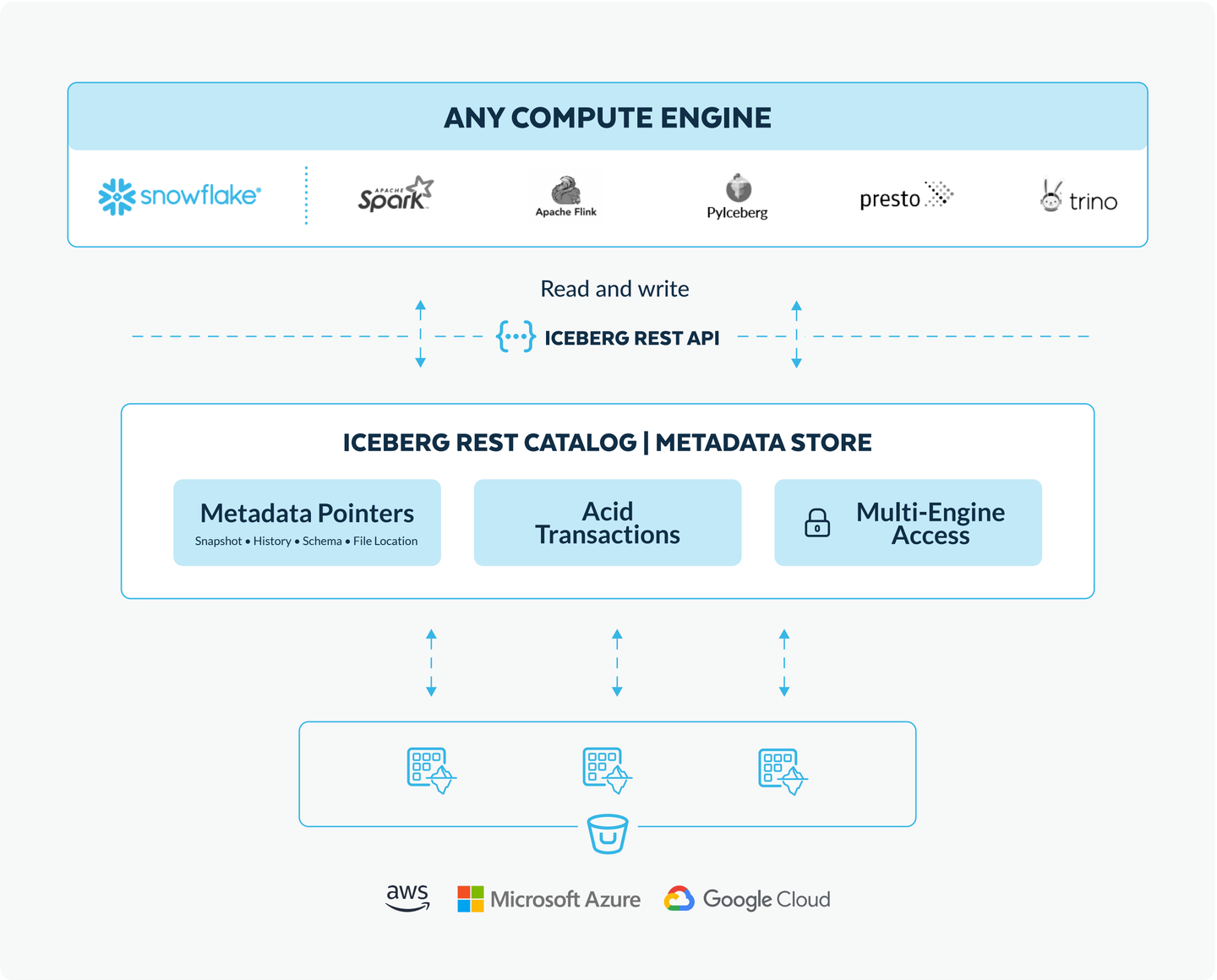
The Iceberg catalog layer

An Iceberg catalog serves as the metastore and the authoritative source of truth for all data in your table layer. Instead of storing all table metadata internally, the catalog maintains a pointer to the current metadata file for each table. This file contains the complete snapshot history, schema, partition spec and file manifests describing where the actual data files are stored. By updating this pointer atomically — using a compare-and-swap operation — the catalog enables ACID transactions (atomicity, consistency, isolation, durability), providing data reliability and preventing corruption across concurrent operations. For broad interoperability, catalogs should implement the Iceberg REST Catalog Specification, a standard API that lets any compliant engine (such as Snowflake, Trino, Spark or Flink) interact consistently with your Iceberg tables. Choosing a catalog that adheres to this specification is essential for maintaining a single, governed copy of data.

The compute layer

This is where the work happens. The compute layer consists of one or more engines — for SQL analytics, data engineering or model training and inference — that query and process data from the storage layer by interacting with the catalog. For any read operation, the engine accesses the catalog to get the table's current metadata file, which it uses to plan the query and fetch data files directly from object storage. For write operations, the engine writes new data files, creates new metadata, and then uses an atomic "compare-and-swap" operation to ask the catalog to update the pointer, preserving transactional integrity. This clear separation of duties — where the engine handles the heavy lifting and the catalog provides consistency — is what enables multiple compute engines to operate efficiently on the same copy of data, allowing organizations to choose the best engine for each workload and independently optimize for their team's performance, reliability and ease of use.

While these decoupled components offer incredible flexibility, they also introduce critical choices. The success of your lakehouse depends not just on the components themselves, but on the architectural principles you use to unite them into a cohesive, high-performing data practice.



ARCHITECTING A RESILIENT LAKEHOUSE ANALYTICS AND AI PRACTICE

Lakehouse analytics and AI — the intelligent computational analysis of data or statistics — enables organizations to discover, interpret and communicate meaningful patterns found in their data stored in their data lakes. While the specific applications are vast, analytical endeavors typically fall into four primary categories:

- **Descriptive analytics** explain what has already happened, providing insights into past events and performance.
- **Predictive analytics** forecast future trends and outcomes, enabling proactive decision-making.
- **Prescriptive analytics** recommend specific actions to optimize results, guiding organizations toward desired objectives.
- **Agentic AI** automates complex decision-making and actions, learning and adapting to achieve specific goals autonomously.

Building an effective lakehouse analytics and AI practice requires more than just choosing the right open table format; it demands a holistic approach to your data architecture. To deliver lasting value, your data centralization and democratization strategy must be built on four key pillars that address scalability, accessibility, security and collaboration.

Scale your team's roadmap, not your infrastructure headcount

Your AI data platform should be a catalyst for innovation, not a drain on resources. A successful lakehouse analytics and AI practice must be powered by a single, intelligent platform that supports the full range of your team's skills and workloads. It should reliably handle complex use cases — from customer segmentation, to geospatial analysis to agentic workflows— without requiring different solutions for different jobs. Instead of requiring manual tuning, ongoing maintenance and deep expertise, it should deliver high performance out of the box, freeing your data engineers and architects to focus on innovating more in the AI era.

Unify data across regions, clouds, formats and catalogs

Business doesn't happen in a single cloud or region, or even catalog or table format. While the ideal goal is to have all data centralized, the reality is that you need to unlock value today while successfully navigating different regulatory requirements. By selecting a lakehouse analytics and AI platform that can integrate data across regions, clouds, formats and catalogs, you accelerate your time to value while minimizing migrations or the need to manage architecture on an ongoing basis. This eliminates data silos created by cloud-specific tools and removes the need for complex, costly data replication and federation strategies, providing a consistently secure experience for all users, everywhere.

> **Building an effective lakehouse analytics and AI practice requires a holistic approach to your data architecture.**

Secure your data with performant, granular controls

Centralizing your data in a lakehouse architecture requires a heightened level of security and governance. These capabilities must be woven into the fabric of your lakehouse solutions, not bolted on as an afterthought. While open standards like the Iceberg REST Catalog API provide a baseline of control for access at the namespace and table level, they often lack standardized mechanisms to centralize advanced governance features. This forces a fragmented security model where you must evaluate each engine or platform's policies for fine-grained access, row- and column-level security, and data masking. Therefore, it's recommended you evaluate a platform's ability to provide these capabilities natively. World-class data isolation, role-based access controls (RBAC), and the ability to tag both data attributes and principal attributes should be integrated, not custom-built. You can evaluate a platform's native support by ensuring these capabilities work across regions and clouds, do not require complex logic, and their implementation does not degrade query performance. A modern platform provides these critical features out of the box, ensuring that protecting your data doesn't come at the cost of performance.

Democratize access to high-quality data for advanced analytics and AI

The true power of a unified data estate is realized when you can securely share, access and activate data across teams, in near real time. Your lakehouse analytics and AI practice should make it simple to share governed data with any stakeholder — be it internal teams, external customers or third-party partners — without creating copies, managing complex ETL pipelines or compromising security and governance policies. The ability to easily and securely share data transforms it from a static asset into dynamic, collaborative data products that connect your entire business ecosystem to drive critical decision-making, power AI and potentially open new monetization opportunities.

> **Security and governance must be woven into the fabric of your lakehouse solutions, not bolted on as an afterthought.**

COMMON PITFALLS OF TRADITIONAL LAKEHOUSE SOLUTIONS

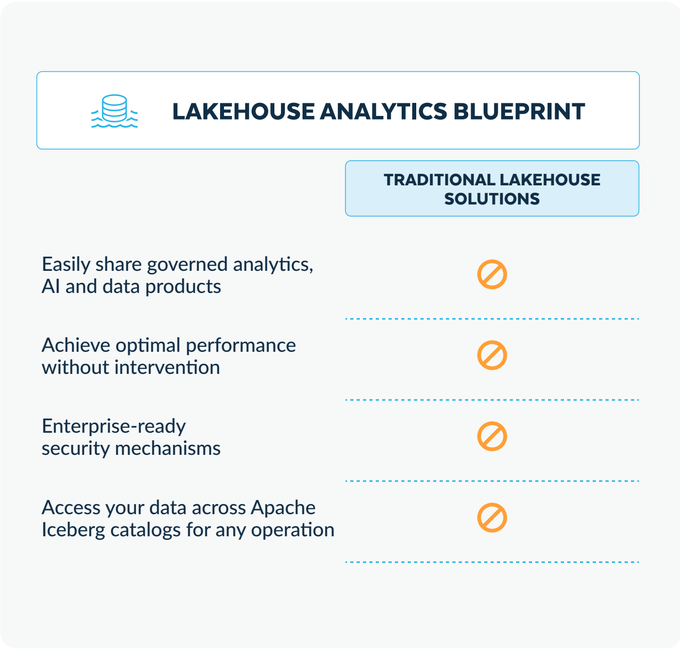
The allure of the open lakehouse is its promise of ultimate flexibility. In an ideal world, your data scientists could use their preferred Python libraries, your analysts could connect with their BI tool of choice, and your data engineers could build pipelines with best-of-breed technologies — all operating on the same, consistent and governed data. This vision eliminates data silos, reduces data movement and accelerates innovation.

However, many organizations find that the reality of powering analytics on a lakehouse is fraught with unexpected complexity. The path is lined with pitfalls that can undermine the very goals you set out to achieve.

Performance bottlenecks and the migration treadmill

Many lakehouse analytics solutions create a frustrating cycle of performance tuning and eventual migration. There are engine-only solutions that, while flexible, cannot deliver the reliability and concurrency needed for enterprise workloads, requiring labor-intensive cluster management and maintenance. This leads to a predictable pattern: Teams battle aborted queries, miss service-level agreements (SLAs) for critical dashboards, and spend nights and weekends manually tuning performance. When the engine inevitably hits its limits, the only option is a costly and disruptive migration to a new solution, starting the cycle all over again. What was conceived as a cost-efficient solution turns into a blocker to scale your team's impact.

Other platforms attempt to solve this by stitching together a complex web of specialized engines: one for SQL, another for streaming, a third for machine learning. While this appears comprehensive, it introduces massive operational overhead and performance challenges of its own. Data teams are forced to manage complex integrations, becoming infrastructure managers rather than data engineers as they wrestle with different security models, varying cost profiles and the constant need to use multiple engines. This fragmentation doesn't break the migration cycle; it just makes it more granular, forcing teams to constantly evaluate and replace individual components.



The hidden productivity tax of fragmented security

Traditional lakehouse solutions often fail to meet the needs of modern data architectures, providing data isolation that is porous and only works at a coarse, group level when it's more secure to isolate by granular, hierarchical and team roles. Additionally, implementing data masking requires complex logic that must be processed each time the query runs. This approach creates a significant "productivity tax," slowing down your team with cumbersome data security solutions — consuming valuable time that should be spent on innovation. Furthermore, these solutions often lack the ability to tag both data attributes (e.g., PII) and principals attributes (users or roles), making it nearly impossible to efficiently manage and audit security policies at scale.

Ultimately, security itself becomes a performance and roadmap blocker, preventing data teams from focusing on what truly matters: innovation to power business decisions and AI.

> Secure data collaboration turns into a friction point, not an enabler, of your goals.

The cross-cloud barrier

In a multi-cloud world, many analytics engines create new silos. Solutions from cloud service providers are often designed to work only within their own ecosystems, while other platforms require additional implementation and ongoing maintenance to make sure there's a consistent cross-cloud, cross-region experience. In many cases, legacy solutions opt for "federation" or "data virtualization" as shortcuts to delivering on cross-region and cross-cloud data estates — however, these approaches add significant cost, degrade performance and can introduce security vulnerabilities. This prevents a truly unified view of your data and forces complex integrations that could lead to costly data egress charges to bridge the gaps between clouds and regions.

The impossible complexity of data sharing

Traditional approaches like FTP and APIs introduce manual effort, slow data delivery and lack a central governance model, but many modern data sharing solutions are not much better. Even those based on open source protocols require providers to use a specific, vendor-managed server and catalog to define and manage shares. This creates a technical dependency on that platform for secure, governed sharing, thereby introducing operational complexity and a new lock-in. While the protocol may allow recipients to access data with various tools, the burden of managing egress costs and ensuring consistent performance often falls on the data provider, making it an expensive and operationally complex solution not ready for enterprise-scale B2B collaboration.

Furthermore, advanced security features like fine-grained controls and policy enforcement may not extend consistently to external consumers, forcing teams to build manual workarounds while increasing the risk of data leakage. Ultimately, secure data collaboration turns into a friction point, not an enabler, of your goals.

> Data teams are forced to manage complex integrations, becoming infrastructure managers rather than data engineers.

Inadequate support for agentic AI workflows

Legacy solutions often lack the integrated tools, scalable infrastructure, and near-time processing capabilities necessary to build, train, and deploy sophisticated agentic AI models, hindering autonomous decision-making and action.

In all scenarios, the result is the same: Your most valuable technical talent is diverted from innovation to integration and maintenance, slowing down your ability to power business decisions and limiting capacity for generative and agentic AI development. Instead of building data products and delivering insights and AI, your team is forced to manage a fragile collection of point solutions or stitching together tools that should work together. The promise of flexibility gives way to the reality of complexity and fragmentation.

SNOWFLAKE FOR LAKEHOUSE ANALYTICS AND AI

The evaluation principles outlined above are not just theoretical; they are the blueprint for a high-performing, enterprise-grade lakehouse analytics practice. The Snowflake AI Data Cloud was built to deliver on these principles, providing a unified platform that embraces openness without sacrificing the performance, governance and ease of use that modern data teams demand.

A single engine that just works

Snowflake provides a single, elastic performance engine for all your workloads that is deeply committed to open standards. With robust, native support for Apache Iceberg, you can bring the power of Snowflake's engine to your existing lakehouse data without additional tuning or data movement. The engine's multi-cluster compute architecture eliminates resource contention — allowing your teams to run BI, AI and data engineering jobs concurrently with world-class performance. It adaptively optimizes jobs, ensuring even the most complex queries are performing — ultimately empowering data teams, regardless of experience, to deliver impact from Day 1.

Unified security and governance

With Snowflake, you can define and enforce security and governance policies in one place, and it all works natively and across clouds and regions. These policies are applied consistently to all data, whether it's stored by Snowflake or in your own storage, enabling a secure data estate. Leverage role-based access controls from the account to the row-and-column level. You can define roles based on type of data access (ex: ReadOnly), functions (FinancialReporter) and technical capabilities (ETLDeveloper, etc.), giving you an enterprise-grade solution that provides flexibility and scalability while preserving clear audit trails that separate identity from function. For securing objects, you can tag data attributes as well as principal attributes, and data-masking is natively integrated into the query engine, making it easy to implement with only SQL-based statements and nearly zero impact on performance. In Snowflake, security and governance are designed to enable your innovation, not hinder it.

Unify your entire data estate without data movement

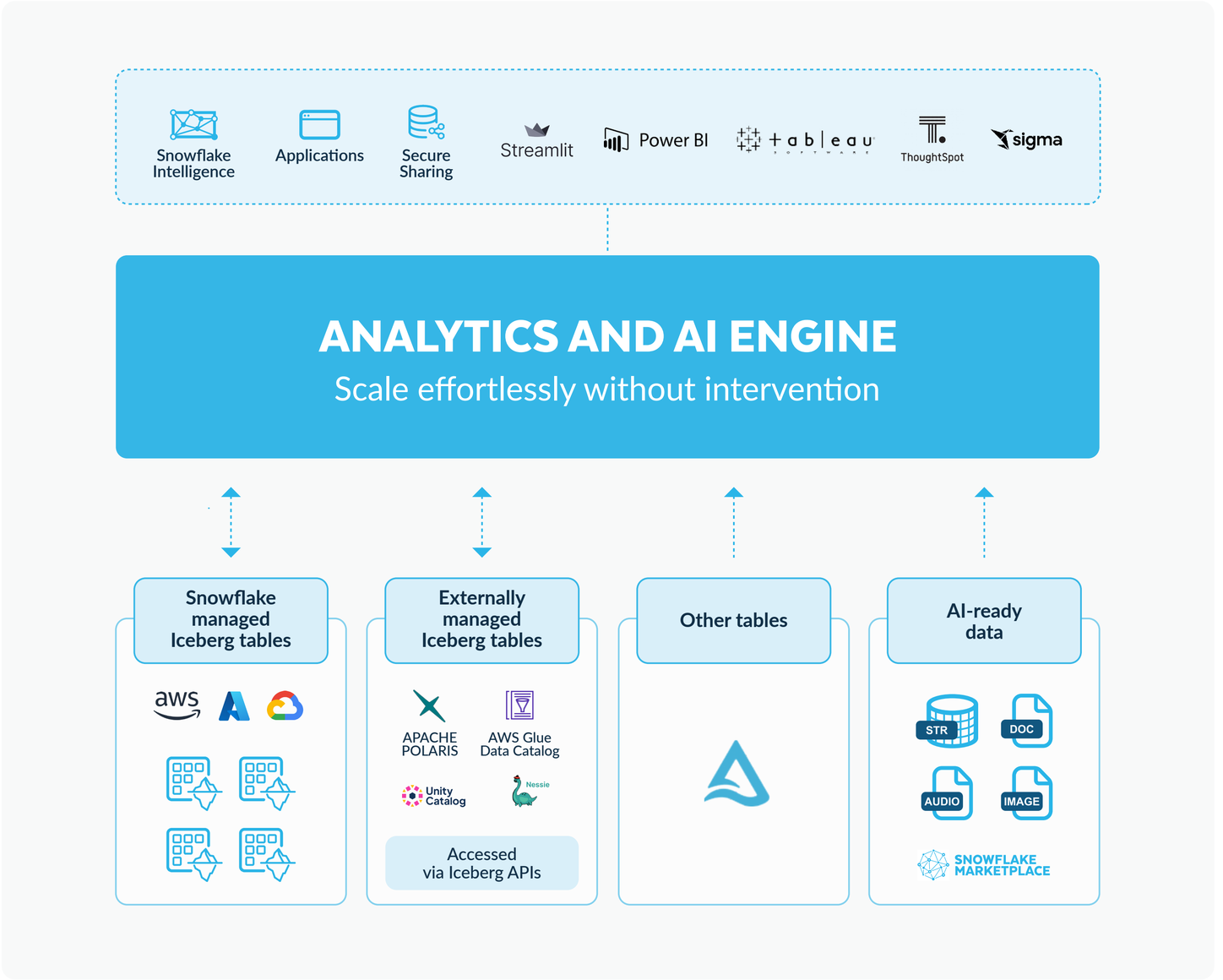
Snowflake was designed for a multi-cloud world, allowing you to build a central, governed pane of glass from which you can read from and write to any Iceberg table, regardless of its catalog. For enterprises with heterogeneous data, you can connect to any catalog that has implemented the Iceberg REST Catalog specification — including AWS Glue Data Catalog and Unity, among others — for immediate table discovery and activation. You can even transform existing Parquet and Delta Lake tables into Iceberg tables without copying or moving the underlying data. Combined, these solutions enable you to truly centralize data without costly migrations and ultimately allow you to accelerate time-to-value without leaving a byte behind.

Securely share AI-ready data products with ease

Snowflake's industry-leading secure data sharing enables you to collaborate on governed data with partners and customers, eliminating the need for complex ETL or data copies. It fully supports Apache Iceberg. Crucially, the solution preserves your security and governance policies, enabling your data to remain protected and giving your data teams peace of mind. This transforms your data into secure, collaborative data products that can be shared across clouds and regions, allowing you to connect your business ecosystem, drive critical decision-making, and even open new monetization opportunities by listing your data products on Snowflake Marketplace.

> **Snowflake enables you to truly centralize data without costly migrations and ultimately allow you to accelerate time-to-value without leaving a byte behind.**

By aligning with these core principles, Snowflake provides a path to build a better enterprise lakehouse analytics practice — one that empowers your teams to innovate more.



CHARTING YOUR COURSE: A PRACTICAL TRANSITION STRATEGY

Transitioning to a modern lakehouse analytics and AI strategy is as much evolutionary as it is revolutionary. The goal is to build a solution that delivers long-term stability and scalability, reducing the need for costly and disruptive migrations down the road.

Start by conducting an assessment of your current architecture. Where are your teams spending the most time? Are you struggling with reliability and performance? Are your governance and security processes fragmented? How will your business demands evolve over the next several years? An honest appraisal of these pain points will build the business case for change.

As you plan your transition, consider the following:

- **Embrace a phased approach**: You don't need to move everything at once. Start with a high-value workload where the pain is most acute. A successful proof of concept will build momentum and demonstrate the value of a unified platform.
- **Involve stakeholders early and foster a culture of change:** Bring business, data and IT stakeholders into the process from the outset. Open dialogue and shared goals help align priorities, anticipate user requirements and smooth the adoption journey, leading to higher buy-in and faster value realization.
- **Prioritize ease of use:** In today's business environment, complexity is the enemy of efficiency. Your team's time is your most valuable asset. Prioritize platforms that simplify, rather than complicate, your architecture. Empower your current team to innovate and deliver value without sacrificing the performance and reliability your business demands.
- **Choose solutions with interoperability and future-readiness:** Prioritize technologies that natively integrate with your existing systems and data sources, and are designed to support evolving needs across multiple clouds and regions. This helps your investment adapt to future business and technology trends, minimizing technical debt and vendor lock-in.
- **Iterate, measure and optimize continuously:** Set clear success metrics for each phase and regularly review progress. What does "good" look like? Use feedback and real-world results to guide incremental improvements, so your new lakehouse platform continues to deliver value as your organization grows.

A successful transition is not just about adopting new technology; it's about adopting a more agile way of working with data.

LAKEHOUSE ANALYTICS BLUEPRINT

	SNOWFLAKE DIFFERENCE	TRADITIONAL LAKEHOUSE SOLUTIONS
Easily share governed analytics, AI and data products	✓	⊘
Achieve optimal performance without intervention	✓	⊘
Enterprise-ready security mechanisms	✓	⊘
Access your data across Apache Iceberg catalogs for any operation	✓	⊘

CONCLUSION: FROM DATA TO IMPACT

The success of a lakehouse analytics and AI practice is not measured by the volume of data it reads, but by the business impact it generates. The ultimate goal is to activate your data — to make it securely and reliably available for the analytics and AI initiatives that drive your business forward.

Achieving this requires more than just open formats and flexible storage. It needs a powerful, unified platform that eliminates the trade-offs between performance and openness, governance and flexibility. By evaluating solutions against the core principles of a single performant engine, unified governance, seamless collaboration and agentic capabilities, you can build a data foundation that is not only ready for today's challenges but also can extend to tomorrow's opportunities.

Don't let traditional complexities or fragmented solutions keep your teams from delivering meaningful business impact. The new era of enterprise lakehouse analytics and AI is here, and the time to architect for impact is now.

Let's chat.

Snowflake is the platform for the AI era, making it easy for enterprises to innovate faster and get more value from data. More than 12,000 customers around the globe, including hundreds of the world's largest companies, use Snowflake's AI Data Cloud to build, use and share data, applications and AI. With Snowflake, data and AI are transformative for everyone.

Learn more at **snowflake.com**

(NYSE: SNOW)