Deloitte.

snowflake®

Unifying Enterprise Data for Generative AI

The next era of artificial intelligence (AI) has arrived, but many organizations may not be ready for it. By making it easy to interact with AI models using natural language, the excitement around Generative AI is warranted. Data is the core differentiator of Generative AI strategies, and the enterprise ability to move from experimentation to levering these capabilities cost-effectively and at scale likely depends on the organization's data maturity and the diversity or sprawl of enabling platforms.

A challenge for many organizations is that the path to their current data state has injected fragmentation across the tech stack, with ETL, data storage, and AI workloads running on different platforms. This has a significant impact on cost and efficiency, and when it comes to Generative AI, the fragmented data state may limit the enterprise's ability to derive value from proprietary data in a secure environment and deploy it at scale.

Today, many organizations are rushing to identify value-driving, differentiating Generative AI use cases. For those ambitions to grow into reality, organizations face a pressing need to resolve the data fragmentation issue and adopt platforms that can consolidate compute and storage to fuel Generative AI applications. To understand the remedy and identify the way forward, consider the path that has led organizations to their current data state.

The history of fragmentation in corporate analytics

Corporate analytics has a genesis, in part, in military organizations. Staff support organizations supporting various commands or services gathered data to produce intelligence that helped frontline actors in their mission. This approach was followed in the business realm, with support organizations within the enterprise gathering data and creating intelligence close to the point of action within business units. As the complexity and technology requirements for data and AI grew, it became less feasible for individual staff organizations with disparate units to manage the scale, complexity, and cost of this function. Corporate analytics were born to facilitate synergies and scale across business unit data and analytic needs through a shared service.

When analytics moved to shared service IT, data platforms were re-platformed to traditional single node, vertically scalable relational database management systems (RDBMS), which was the standard for the online transactional processing (OLTP) databases that IT managed. These databases had several limitations that made them less suitable for analytics, including: limitations on vertical scale; the inability to process, sort, and aggregate large volumes; optimizers suitable for OLTP necessitating multiple index; and use of OLTP standard shared storage that resulted in contention with unrelated applications in the data center.

Gen 2 enterprise data warehouses came to prominence in part to solve limitations of OLTP RDBMs. Data warehouse systems enabled horizontal scaling, handled mixed workloads, and were able to consolidate data gathering and storage from thousands of OLTP RDBMS to tens of enterprise data warehouses (EDW) that enabled greater scale in analytics. Data warehouse systems were powerful but soon became cost-prohibitive in many cases, as the expenses around managing and using parallel systems for warehousing and analytics grew, even as the systems struggled to handle the near-exponential growth of data over those several decades.

Apache Hadoop was introduced as a Gen 3 system in part to solve the cost challenge of EDWs. Hadoop systems were horizontally scalable and ran on commodity hardware, in contrast to Gen 1 OLTP platforms and Gen 2 data warehouses that ran on highly optimized proprietary hardware. While Hadoop was cheap for bulk data processing, its complexity and architecture made it less suitable for high cardinality dynamic tasks, like reporting and ad hoc querying. As a result, Hadoop was used for bulk data processing while the EDW was still used for high cardinality end-user ad hoc interaction and reporting.

On the one hand, this offloaded the bulk of ETL onto a commodity hardware-based architecture while maintaining ad hoc and formal reporting on EDWs. On the other hand, this split the data ecosystem in two, with enterprise data warehouses struggling to scale alongside dozens of Hadoop instances. While the cost of purchasing EDW went down, the complexity and cost of data in the organization started to ramp up exponentially.

The arrival of Apache Spark and the capacity to work with machine learning (ML) on top of Spark fueled further ecosystem fragmentation, with distinct ETL complexes, numerous data warehouses, and Apache Spark-based complexes for AI and ML. The net result is that a significant portion of capital and operational budget for many organizations is spent managing these disparate platforms and the exchanges between them. With Generative AI-enabled edge applications, there is even more pressure to consolidate the data estate.

Recognizing that this is generally how many organizations have reached their current data predicament, the question becomes, where do we go from here?

Consolidating data platforms with Snowpark Container Services

The arrival of modern data platforms where compute and storage were separate initiated a new phase of consolidation, including bringing multiple data warehouses into the same environment. The next step is unifying the entire data ecosystem, bringing together ETL, storage, reporting, AI, and Generative AI into one ecosystem with no need to copy data.

One offering is Snowflake's Snowpark Container Services (SPCS), which facilitates bringing ETL, reporting, AI applications, AI models, and data products to the enterprise data. Container systems can achieve low cost for low-value data distillation, as well as high-value ML, while traditional data warehouse workloads run on the same storage layer. The Snowpark runtime option allows developers to deploy and scale workloads, relying on infrastructure managed by Snowflake while also accessing configurable hardware (e.g., GPUs).

One of the benefits of unifying the enterprise data ecosystem is that it simplifies the task of managing services and tools. Rather than independently assembling a container registry, management service, and compute service (alongside managing tools for working with the data), enterprises can run proprietary data products and third-party or Snowflake Native Apps in the Snowflake environment. This allows developers to explore and build sophisticated, data-intensive applications, including Generative AI deployments. Indeed, Snowpark Container Services offer GPU-accelerated model training, allowing the enterprise to leverage open-source models and fine-tune them on enterprise data in the Snowflake environment.

The reality is that Generative AI is not simply another incremental step forward in the trajectory of AI. It holds promise as a differentiating, disruptive technology that enterprises are looking to train and deploy models and seize a first-mover advantage. The challenge is that if the data ecosystem is fragmented, it can inject significant cost and complexity when attempting to scale a use case. Proof of concept (POC) projects unconstrained by the realities of technical debt and code writing may evidence the viability of a use case, but simply scaling an unconstrained POC without the necessary consolidated architecture will limit Generative AI value in terms of effectiveness and cost.

There is an opportunity today to acknowledge the limits imposed by a fragmented data ecosystem and consolidate the backend to prepare the enterprise for a future with Generative AI that scales. Just as important, consolidating data platforms for storage and compute supports AI governance, risk mitigation, and data security. To be sure, change is hard, and transforming the enterprise's data estate can introduce complexity, uncertainty, and risk. Deloitte can help you modernize data and applications in a way that is fast, efficient, and secure. We offer rich experience and subject matter experience in the end-to-end complexities of data migration, consolidation, and management, and our clients seek our knowledge and services across cybersecurity, compliance, risk management, and AI.

Areas for collaboration on the Generative AI journey

Looking to the opportunities afforded by Snowpark Container Services, Deloitte is a trusted advisor to help your enterprise reshape and consolidate data platforms. Deloitte has one of the largest Snowflake practices among professional services firms, and we were named Snowflake's 2023 Partner of the Year for a third year in a row. There are five key areas organizations can focus on to leverage the depth of capabilities from Snowflake and Deloitte to jumpstart their journey.

1. Explore the art of the possible with Generative AI. Take time to explore the spectrum of Deloitte and Snowflake's technology relationships and investigate how their offerings align with your Generative AI vision.

2. Set up a Deloitte- and Snowflake-led Generative AI fluency/training series, offered for a small group or for as many as 15,000 learners, to help accelerate learning and inform your point of view in this fast-evolving space.

3. Work with Deloitte and Snowflake to formalize a data strategy, including exploring use case prioritization with a focus on when to build versus buy in light of feasibility, cost, time, and value. A formal data strategy also incorporates an examination of technology/vendor options (Platform-as-a-Service, Software-as-a-Service, Infrastructure-as-a-Service, or a hybrid model) with Deloitte, Snowflake, and a cloud provider of your choice.

4. Start your value realization journey by delivering value to business or operating units while also proactively modernizing and simplifying the underlying architecture to enable value, scale, and an optimal price point. In addition, consider the risks that can emerge in Generative AI deployments, such as by using Deloitte's Trustworthy Generative AI framework, and develop the mitigation tactics to address things like bias, security, accountability, and transparency. This prepares the organization to account for trust, ethics, and risk mitigation when scaling use cases.

5. Design and build for efficient operations by including PlatformOps, AI/ML/LLM Ops, and Data/App operations. Inability to efficiently scale is one of the most persistent problems in today's experiment-oriented world. Designing for operations when the value vectors and scale factors are unclear is difficult, but there are methods and techniques that support efficient and flexible design that evolves and scales.

Importantly, time is of the essence. Many businesses are making investments in Generative AI with the ambition to be first to market, and the unification of data platforms is an essential component for a competitive edge. With Deloitte and Snowflake, you can access the data capabilities you need to confidently embrace this new era of Generative AI.

Ready to get started?

Please get in touch! Deloitte is eager to learn about your priorities and help you chart your path to a modern data environment with Snowflake.

Matt Wallbrown
Snowflake Lead Alliance Partner
Deloitte Consulting LLP
mwallbrown@deloitte.com

Rupesh Dandekar
Chief Technical Officer for Snowflake
Deloitte Consulting LLP
rudandekar@deloitte.com

Goutham Belliappa
Managing Director | AI & Data
Deloitte Consulting LLP
gbelliappa@deloitte.com

Anthony Ciarlo
Alliances Relationship Executive
Deloitte Consulting LLP
aciarlo@deloitte.com

About Deloitte

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited ("DTTL"), its global network of member firms, and their related entities (collectively, the "Deloitte organization"). DTTL (also referred to as "Deloitte Global") and each of its member firms and related entities are legally separate and independent entities, which cannot obligate or bind each other in respect of third parties. DTTL and each DTTL member firm and related entity is liable only for its own acts and omissions, and not those of each other. DTTL does not provide services to clients. Please see http://www.deloitte.com/about to learn more.

This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor. Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.