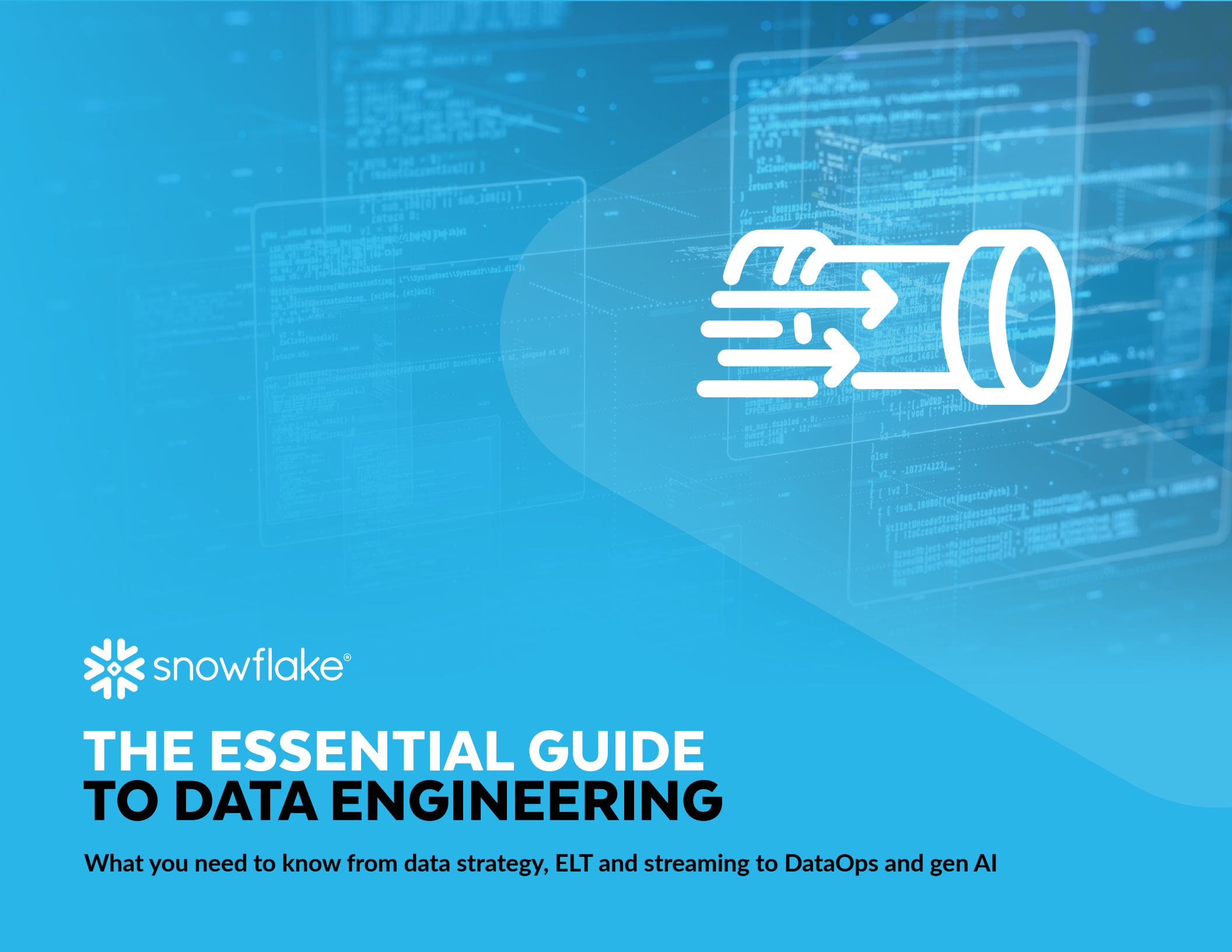


snowflake®

THE ESSENTIAL GUIDE TO DATA ENGINEERING

What you need to know from data strategy, ELT and streaming to DataOps and gen AI

TABLE OF CONTENTS

The vast majority of data — some 80% to 90% — is considered unstructured or semi-structured, including text documents, emails, chat sessions, video files, social media posts, app and web metrics, equipment logs, IoT sensor data and more.

INTRODUCTION

The world now generates far more data each day than ever before. And despite the trend that the volume of information produced by humans and machines continues to double approximately every two years, our appetite for data remains unsated.

Making use of that data hasn't always been easy. The vast majority of data — some 80% to 90% — is considered unstructured or semi-structured, including text documents, emails, chat sessions, video files, social media posts, app and web metrics, equipment logs, IoT sensor data and more. The rise of generative AI (gen AI) pictures a state-of-the-art reality of what may have seemed impossible in the past, creating even more demand for this type of data. However, preparing and processing this data has not been easy or straightforward, yet it is proving increasingly crucial as machine learning (ML) and artificial intelligence drive more business decisions.

Data is the fuel that powers artificial intelligence. Generally, the more high-quality data you can feed into a machine learning model, the more accurate its outputs are likely to be. But before enterprises can use AI to derive insights from the petabytes of data they collect, they need to break down internal data silos, organize the data, determine whether it contains sensitive or regulated information, establish strong governance and access policies around it, and determine how to securely store and share it across the organization.

Most of all, enterprises need an effective data strategy and an efficient data management platform, and they need skilled data engineers to implement and maintain them.

WHAT IS A DATA STRATEGY AND WHAT ROLE DO DATA ENGINEERS PLAY?

Simply put, a data strategy is a long-term plan that defines how an organization collects, processes, uses, governs and stores information. The goal of a data strategy is to make relevant information available to stakeholders across the organization so they can use it to drive decisions. That requires enterprises to actively break down information silos between business units and adopt uniform policies in regard to data types, storage architectures and workloads. As nearly every organization collects some type of personally identifiable, proprietary or regulated information, creating strong, unified governance policies is also crucial.

In an ideal scenario, an enterprise's data strategy ensures that business users have the information they need to make decisions when they need it. It allows for seamless data sharing between internal and external partners, with proper safeguards in place to prevent data leaks or unauthorized access. It enables better performance when accessing data and greater resilience during business disruptions, likely leading to overall lower total cost of ownership (TCO).

The advent of gen AI and the need to collect and prepare training data for LLMs makes having an effective data strategy one of the top priorities of any modern organization. That, in turn, requires a renewed emphasis on the importance of data engineering.

WHAT DATA ENGINEERS DO

Business users want to ask questions of their data and get back useful answers in return. That basic need is made possible by the practice of data engineering.

Data engineering encompasses the tools and expertise needed to facilitate the steady flow of data. Data engineers make data "production-ready" by getting it into a usable form, typically accessible via a centralized facility or cloud data platform. They make this possible using a variety of practices and techniques, including the following:

- **Collecting and ingesting:** Data engineers are skilled at wrangling data from hundreds of sources, including relational databases, enterprise applications, SaaS providers and cloud storage vendors. An increasing amount of this data is captured in real time, streaming in from IoT devices, machine logs, clickstream monitors, messaging systems and more.
- **Transforming and standardizing:** Data is inherently messy, incomplete and inconsistent. Engineers often need to standardize data by converting different file types to a universal format, cleanse it by removing inconsistencies and inaccuracies, map it by combining elements from multiple data models, and augment and enrich it with business logic or by pulling data from other sources to fill in missing pieces.
- **Loading into a centralized repository:** Along with ingesting and transforming the data, engineers need to load it into a storage solution that can scale as the volume of data increases. Such repositories and processing needs formed the different architecture options including data lakes (raw data stored in native formats), data warehouses (data organized around a set of logical rules, or schema), data lakehouses that combine elements of both, or other cloud data platforms. Depending on the type of architecture, different data engineering practices may apply.

Together, these three core functions form a data pipeline that automates the flow of information from data sources to the repositories and makes it available to enterprise applications, where business users can query it and analyze the results.

HOW THE CLOUD IMPACTS DATA ENGINEERING

Storing data in a centralized cloud repository offers several key advantages over more siloed approaches. The cloud offers virtually unlimited storage capacity, with near-infinite elasticity and scalability. This allows companies of any size to deploy a large number of concurrent, high-performance workloads within a centralized platform.

The ability to scale storage capacity allows businesses to retain more types of data and in greater volumes. Over the past decade, the prevailing sentiment among large enterprises has been to "capture data first, ask questions of it later." Organizations can now retain years of historical data, in the hopes of identifying patterns and gaining insights using AI.

The cloud has also made computing much more scalable and flexible. Organizations can scale the compute power virtually unlimitedly if done right. This also includes serverless computing, in which cloud service providers automatically provision, scale and manage the infrastructure required to host your data and run your business applications.

This has also given rise to highly efficient methods of application development and operations (DevOps). A serverless environment allows developers to bring products to market faster via containers (software packages that contain everything needed to run an application), microservices (applications built as modular components or services), and continuous integration/continuous delivery (CI/CD) processes. Cloud-native services allow organizations to provision storage and computing independently, based on the needs of each workload.

Storing data in a cloud-native repository makes it possible to simultaneously load and query data without degrading performance. A modern cloud data platform can seamlessly replicate data across multiple regions and clouds to enhance business continuity. It also allows enterprises to apply uniform governance policies to data across the organization, while making it easier to securely share data with internal and external stakeholders.

The ability to deploy and manage data pipelines in a cloud environment is now a primary requirement for data engineers, who must be fluent in cloud-native technologies such as Apache Kafka, Docker and Kubernetes, among others.

HOW AI IS CHANGING DATA ENGINEERING

The emergence of gen AI has focused even more attention on how enterprises collect, manage and govern data.

Well-documented problems with gen AI regarding biased outputs and nonfactual statements (known as hallucinations) bring additional scrutiny to the provenance, accuracy and inclusivity of training data. Organizations that want to train LLMs using internal data need to ensure that their data is both available in sufficient quantities and can be relied upon. The job of ensuring data quantity and data quality falls largely on the shoulders of data engineers.

Because AI-based tools such as fraud and intrusion detection rely on real-time data feeds, data engineers play a crucial role in streamlining these systems to ensure a consistent and reliable flow of information. Expertise in SQL, Python, Java or Scala is increasingly seen as a fundamental requirement for the job.

Gen AI tools can also be used to generate synthetic data that mimics the properties of real-world data. This can be used by data engineers as a proxy for information that's missing or incomplete, allowing data scientists and business intelligence analysts to conduct more comprehensive analysis of a particular data set.

> ***The emergence of generative AI has focused even more attention on how enterprises collect, manage and govern data.***

Data pipelines will be one of the primary ways enterprises will look to when feeding LLMs. Data engineers who understand how to plug existing data pipelines into their organization's LLMs will play a crucial role in how effectively an enterprise implements gen AI technology.

INSIDE THE DATA ENGINEERING PROCESS

The most important responsibility of data engineers is to build data pipelines, using the techniques described above, to better enable business leaders to make data-driven decisions. There are two well-established procedures for making this happen.

DELIVERING AND SHARING DATA THE MODERN WAY

For data to be useful, it has to be shared. But data sharing must be enabled subject to governance policies that establish appropriate guardrails around privacy, security and regulatory compliance. Here, too, data engineering and data platforms play a vital role.

In the past, data was typically delivered to stakeholders via application programming interfaces (APIs), file transfer protocols (FTP) and even email. This process was fraught with potential governance nightmares, including a much higher risk of data leaks, the sharing of outdated or inaccurate information, and unauthorized individuals gaining access to sensitive files.

Modern data-sharing technologies avoid many of these issues by granting users access to read-only versions of data stored in their original locations. With granular-level access control, data is shared rather than copied, and no additional cloud storage is required. In this fashion, organizations can avoid data movement, eliminate costly extract-transform-load (ETL) processes, and ensure that users are always able to access the most current version of data. This also eliminates the need for data engineers to set up FTP locations or configure APIs.

Cloud-based data platforms are ideal for this type of seamless, secure data sharing, allowing organizations to deliver analytics-ready data on demand to the people who need it.

ETL VS. ELT: WHICH IS BETTER?

Traditionally, data pipelines have been created using a practice known as extract-transform-load (ETL). With ETL, engineers extract information from multiple sources, prepare (or transform) the data to fit the requirements of the data repository, and then load it into a database or data warehouse. Because the data needs to be transformed first using a separate processing engine, it involves more data movement, transfer and manipulation, adding both time and cost.

Engineers building more modern data pipelines will extract data, load it in its native formats to a data repository, then transform it as needed — a technique known as extract-load-transform (ELT). Though they sound nearly identical, these processes differ in several key ways.

- **ELT is ideal for unstructured data:** ETL only works with data that can be transformed into a tabular format — columns and rows. ELT is virtually unlimited in the types of data it can handle.
- **ELT is much faster:** Because data engineers don't have to standardize data formats or establish schema, the data is ingested more quickly and easily.
- **ELT is more flexible:** Data is simply loaded in raw form. Analysts can decide later how they want to transform the data, or if transformation is even needed. This avoids unnecessary data movement and wasted computing resources.
- **ELT is preferable for delivering real-time data:** In any scenario where time delays can be costly — such as stock trading or analyzing possible cyberattacks — faster data ingestion and processing is essential.
- **ELT is better for advanced analytics:** Data scientists commonly load data into a data repository and then combine it with another data source, or use it to train predictive models. Maintaining the data in a raw (or less processed) state allows data scientists to keep their options open.

While ETL is still commonly used with legacy relational databases and data warehouses, ELT has become the process of choice for cloud-based data platforms due to its flexibility, speed and scalability.

ENTERING THE DATAOPS ERA

Good data engineering requires a unique mindset as much as a specific skill set. A talent for data modeling and design are just as important as familiarity with basic ELT techniques. Just as the adoption of agile software development techniques led to the new IT discipline known as DevOps, a similar blending of data operations expertise is now being bundled into a practice called DataOps.

DataOps brings together data engineers, data scientists, business analysts and other stakeholders to apply agile best practices to the data lifecycle, from data preparation to reporting. This approach is being used to automate critical data engineering activities and orchestrate hand-offs throughout the data management cycle, ensuring a continuous loop among planning, building and testing data delivery systems as well as their operation, monitoring and improvement.

For example, as business analysts update their queries, worksheets and schema, data teams are expected to track these changes, ensure proper version control, and make sure these analytic assets still fulfill their original purposes.

DataOps uses good data governance as the foundation for everything else. Key aspects of good DataOps practices include:

- **Lineage tracing:** Organizations need to know where their data originated, who has access to it, and how that data has changed over time. In addition to being an essential element of auditing and compliance practices, lineage tracing helps data engineers identify and debug errors. And because data scientists are increasingly being asked to explain how the ML models they built generate certain predictions, lineage tracing is a key component of AI explainability.
- **Data quality:** Low-quality data can result in faulty business decisions, potentially leading to increased costs, missed market opportunities, lost revenue and hefty fines for compliance failures. DataOps teams must implement and monitor strong governance frameworks that maintain data quality as well as data security and change management processes.
- **Data cataloging:** A data store is only as good as its organizing principles. Data catalogs collect metadata that helps analysts and other users find the data they need and evaluate its relevance for intended uses.
- **Access privileges:** All data governance policies must clearly define who can access, manipulate or change every element of data within a repository. Any information that falls under regulatory oversight — such as personally identifiable information (PII), financial or health data — may need to be masked or tokenized to comply with privacy regulations. Some cloud platforms can mask this data automatically and only reveal it to authorized users.
- **Change management:** Data is constantly in flux, often as a result of individuals making changes to it. Organizations need to maintain a comprehensive audit trail recording which changes were made, who made them, when they occurred and which applications those changes may impact. This can reduce the potential for unauthorized alterations and errors.

For DataOps to be effective, data engineers and key business stakeholders need to work hand in hand to ensure the right data pipelines are being created, optimal data structures and sources are being deployed, analysts are getting the information they need in a form that is useful, data is being properly governed, and that all of these elements fit into a comprehensive data strategy as defined by the organization's top data decision-makers.

THE DATA MANAGEMENT PLATFORM IS KEY

The continued growth of unstructured data and the increasing urgency to train and build enterprise-specific LLMs require a new approach to data management. The data platform an enterprise chooses will have a critical impact on its success.

Today's enterprises want to democratize the use of data throughout the organization, optimize basic data management activities, and support a broad range of data and analytics workloads. Most importantly, they need a single source of truth they can rely upon. A properly architected cloud data platform enables all this and more.

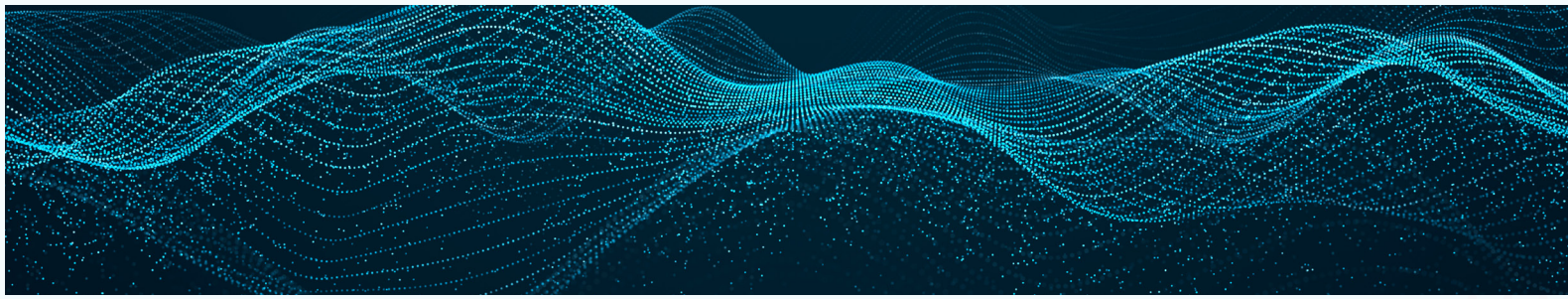
Cloud data platforms should include scalable pipeline services that can ingest streaming and batch data. They enable a wide variety of concurrent workloads, including data warehouses, data lakes, data pipelines, and sharing, as well as facilitating business intelligence, AI/ML and applications.

Consolidating data into a single source of truth makes it easier to access, analyze and share with other stakeholders across the enterprise. Cloud-based solutions allow data engineers to shift their focus from managing infrastructure to managing data, elevating their value to the organization as a whole.

UNLOCKING INNOVATION WITH SNOWFLAKE

The Snowflake AI Data Cloud offers a comprehensive platform for serving your immediate business needs while also unlocking future AI innovation. Here are five reasons how:

1. **No silos:** Every organization struggles with data silos that have developed over time across different business units, workloads, architectures, languages and tools. Employees are continually asked to master new tools and systems in order to collaborate on projects. Snowflake's AI Data Cloud supports a full spectrum of data formats, architectures, programming languages and use cases. Users have the ability to work with SQL, Python, Java, Scala or custom language through container services to transform their data in ways that make sense to them, with the ability to adapt as those needs change over time.
2. **Data security and governance:** No organization ever wants to lose control over the security of its data — a major barrier to enterprise adoption of publicly available LLMs. The Snowflake platform provides automated and continuous compliance monitoring systems that address customers' most sophisticated and complex security requirements. Whether your raw data is being leveraged for analytics, ingested into an LLM, or harnessed to build bespoke apps, Snowflake provides data governance and security capabilities.
3. **Lower TCO with managed services:** Manually managing data across a complex hybrid IT environment invariably involves trade-offs between cost and performance. Snowflake offers scalable performance cost effectively by relieving enterprises of the burden of manually managing clusters, infrastructure including scaling storage and compute power to meet current demand. It requires almost no maintenance, with near-instant elasticity for scaling up or down as needed, while offering consumption-based pricing with transparency and predictability. There are more options with serverless computing in Snowflake.
4. **Easy sharing:** Building multiple data pipelines and moving data around in order to share it is expensive, inefficient and prone to errors. Modern enterprises require access to live, ready-to-query data, free from worry about data spillage or cloud outages. The Snowflake AI Data Cloud addresses both issues, allowing near-instant read-only data access to authorized users along with automatic replication and failover across cloud providers. This reduces complexity while minimizing risk.
5. **Custom apps:** Enterprise applications are never one-size-fits-all. Snowflake allows teams to build custom interactive apps in minutes, then securely share these lightweight apps and data securely with colleagues, partners and external customers.



BUILD POWERFUL STREAMING AND BATCH DATA PIPELINES IN SQL OR PYTHON

Today, organizations hope to use AI to drive new innovation, better productivity and deeper competitive differentiation. But many still lack the data engineering expertise, strategy and platform necessary to transform those hopes into reality. There's no need to try to do everything by yourself. Snowflake can provide the data foundation that is a prerequisite for implementing AI in the enterprise.

To learn more, visit **Snowflake for Data Engineering**.

ABOUT SNOWFLAKE

Snowflake makes enterprise AI easy, efficient and trusted. Thousands of companies around the globe, including hundreds of the world's largest, use Snowflake's AI Data Cloud to share data, build applications, and power their business with AI. The era of enterprise AI is here.

Learn more at **snowflake.com** (NYSE: SNOW)