

WHITEPAPER



Snowflake Fusion

Blending the Future
with Modernization

Author

C. Kalyana Chakravarthy



Contents

01. Compulsive modernization	3
02. Architectural breakthrough	4
• Decoupling compute and storage	4
• Storage	5
• Compute	6
• Metadata services	7
03. Data acquisition	8
• Ability to connect diverse source systems	8
04. In-transit processing	9
• Data engineering	9
• Data science	10
05. Data synthesis	11
06. Conclusion	12

01

Compulsive modernization

Compulsive modernization is the insatiable desire to change and grow.

Today, customers are spoiled for choices, affecting their buying behavior. The unprecedented pace of technological advancements is also rapidly changing how and where they make decisions. All this gives rise to market volatility, where an incorrect decision could end a business. The pandemic has proven that right product at the right time is of prime importance. For example, when everything has come to a stall in Covid lock downs, many businesses have opened doors in the name of door-step delivery. That was the need of the hour like never before.

Data, as we know, is the at the core of every business. Perspective of data lens has changed from being a by-product, to an asset, to today's monetizable product. This demanded and witnessed numerous break throughs in infrastructure, architecture, storage and data processing. Most of them were incremental addressing an issue and opening the door for another. Until the emergence of cloud data warehouses and decoupling of compute and storage, legacy system's limitations remained the same in form or the other.

Cloud platforms like Snowflake offer cutting-edge data solutions that were either impossible or extremely expensive earlier. This PoV will focus on some of the ground-breaking technological breakthroughs and divide them into four categories within the progressive data lifecycle.

02

Architectural breakthrough:

Decoupling compute and storage

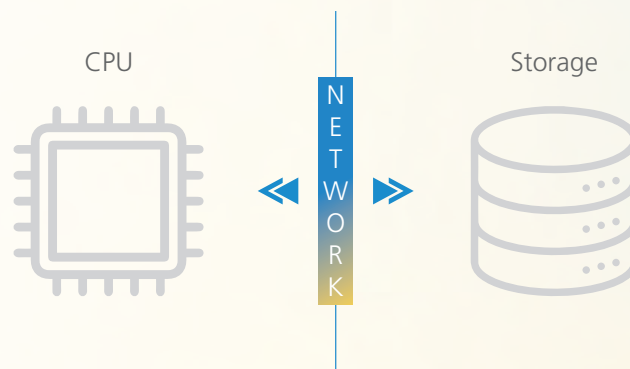


Fig 1: Decoupling compute and storage

This ground-breaking innovation has completely transformed the infrastructure landscape. It has shattered conventional beliefs that storage and compute must be purchased together, creating unprecedented freedom for businesses. Now, businesses can start their cloud journey with minimal Capital Expenditure (Capex) and adapt their Operational Expenditure (Opex) thoughtfully.

The unique separation of compute and storage, forming an independent modus operandi, has enabled the realization of extraordinary features and capabilities.

Storage

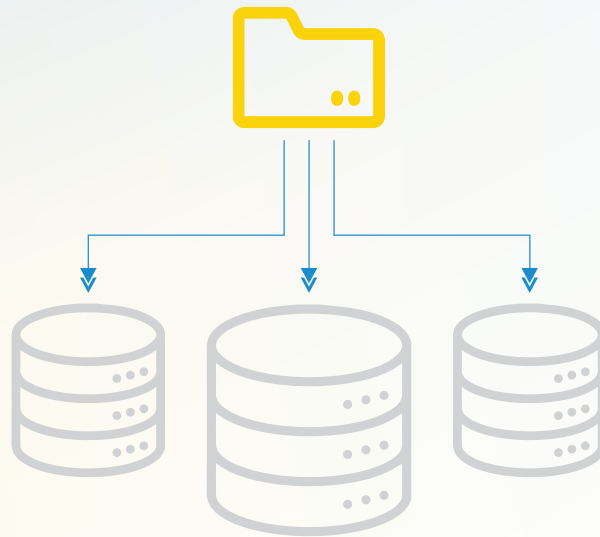


Fig 2: Object storage

The unconventional nature of object storage, being immutable, raises concerns regarding database updation. Updates executed on the database result in new micro-partitions being added to the system, effectively cascading as inserts. Snowflake has ingeniously transformed this consequence into an extraordinary feature called Time Travel, which forms the core of its comprehensive data protection capabilities. Furthermore, Snowflake maintains Atomicity, Consistency, Isolation, and Durability (ACID) compliance by implementing Snapshot Isolation using Multi-Version Concurrency Control (MVCC).

This approach ensures that Snowflake can seamlessly ingest all forms of data without compromising the system's simplicity, transaction control, security, and reliability.

Compute

Snowflake leverages the cloud offerings similarly to storage in terms of compute resources. These resources, known as virtual warehouses, are ephemeral and billed per second for the duration they are active and executing tasks. Virtual warehouses provide the flexibility to scale up or down and spin in or out as per the specific needs of the workload.

Additionally, Snowflake implements a feature called File Stealing within a warehouse. This means that if some nodes are still reading files from storage while others have completed their tasks, the pending files in the queue will be taken over by the available nodes. This eliminates runtime execution skew, if present, among the warehouse nodes.

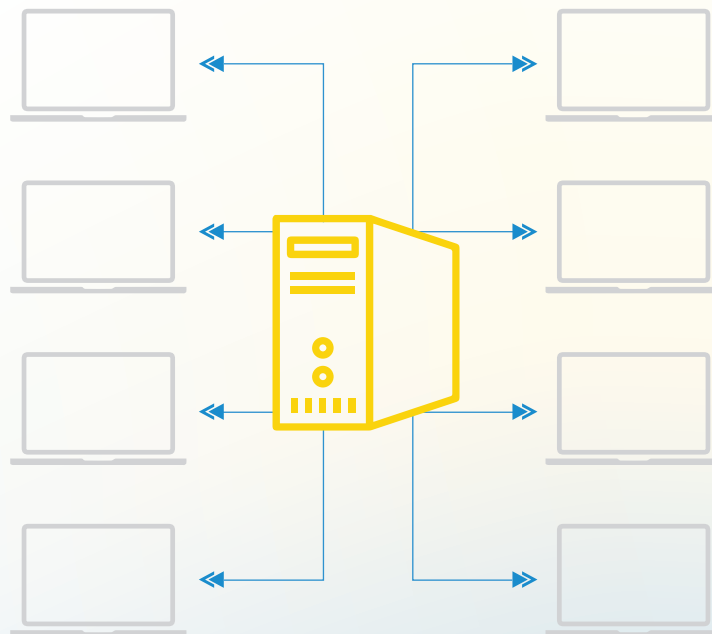


Fig 3: Compute clusters

Metadata services

The metadata services act as the system's brain, driving its intelligence. Most optimizations occur within this layer, which enhances performance. This includes lightning-fast query caching, comprehensive metadata collection for every object, and warehouse caching in the compute layer. These capabilities work harmoniously to retrieve query results at an optimal cost.

The **robust transactional metadata system** forms the foundation for a range of advanced features, such as clustering, materialized views, serverless functionality, Smart Optimization Service (SOS), query acceleration, etc. These features are made possible by the reliability and power of the metadata system, which helps deliver exceptional performance and enhanced functionalities.

03

Data acquisition

Ability to connect diverse source systems

An analytical system should be capable of seamlessly connecting with multiple source systems. Snowflake's ecosystem began with the essential support of Java Database Connectivity (JDBC)/Open Database Connectivity (ODBC) connectors and has significantly expanded to include custom connectors and drivers, integrating new data sources as they evolve.

Snowflake's remarkable ability to connect to various systems, such as the Hive MetaStore (HMS), which houses a vast amount of data, and external cloud platforms, enables it to read data directly from these sources using metadata-rich external tables. This eliminates the need for traditional data migration or ingestion processes into Snowflake from these cloud providers, ultimately reducing the time-to-market value and accelerating data accessibility.

This ever-growing partner ecosystem strengthened Snowflake's existence leading to the rightful market invasion.

04

In-transit processing

Once data is ingested into Snowflake, it transforms seamlessly from one layer to another. In traditional systems, accomplishing this transformation typically requires third-party Extract, Transform, and Load (ETL) tools or custom scripts. However, this approach can be inefficient as it involves moving data in and out of the system, leading to potential performance issues and data handling complexities.

Data engineering

Data engineering is a vital bridge between data acquisition and AI/ML. Snowflake manages structured and semi-structured data at scale as a unified data platform, delivering unparalleled performance.

Snowflake has evolved with advanced features like:

- 01** Streams for Change Data Capture (CDC)
- 02** External functions for additional code usability
- 03** Snow pipe streaming for micro batching
- 04** Stored procedures for extended SQL
- 05** Kafka connector for real-time loads
- 06** Code sharing for monetization
- 07** Tasks for scheduling
- 08** Bulk loading for batch loads

These features ensure the availability of a robust, fully-grown solution, regardless of whether you are using a Lambda or Kappa architecture.

In addition, Snowpark provides extensive programmability features to Snowflake. Snowpark's code is seamlessly translated to SQL and executed within Snowflake, leveraging its virtual warehouses. Unlike previous solutions, there is no need to move data around as the **code goes directly to the data**, eliminating the need to extract and process data outside the system.

Data science

Snowflake initially emerged as a cloud data warehouse, offering exceptional features as discussed above. However, for data science and AI/ML purposes, there was a need for third-party applications or tools to connect with Snowflake.

To address this, Snowpark was introduced, enabling the efficient execution of AI/ML models. This breakthrough has expanded the possibilities for various AI/ML use cases. Python can now be utilized within Snowpark to develop complex programming solutions, ensuring efficient outcomes.

Furthermore, with the acquisition of Streamlit, **prototyping** becomes seamless, allowing the creation of native applications directly within the Snowflake environment.

As a result, Snowflake has evolved from being a cloud data warehouse to a comprehensive and mature cloud data platform, offering solutions across **data engineering and data science domains**.

05

Data synthesis

Data is a **multifaceted** entity. In its pure form, multiple teams can utilize data for various use cases. However, data sharing was neither common nor easily achievable in the past, as it involved a series of cumbersome steps in-between. This process often took time, resulting in the data recipient receiving outdated information by the time it reached them.

This tedious and inflexible data-sharing process hindered the true purpose of sharing data between parties. To address this challenge, Snowflake introduced the data sharing feature, which enables the seamless sharing of live data with consumers in the simplest manner possible. This innovative approach has revolutionized the way we perceive data. It has transformed data from being a **by-product** to an **asset** and further to a **product** that can be monetized.

Additionally, Snowflake's inherent masking ability helps create **data-clean rooms**, which can be securely shared with consumers. Over time, data sharing has evolved into data exchange and marketplace, introducing a culture of effective **monetization**.

06

Conclusion

With its multiple ground-breaking features, Snowflake continuously evolves and establishes itself as a transformative global solution. Undoubtedly, Snowflake possesses the potential to serve as an architecture-supporting data fabric.

With its unique architecture, Snowflake enables seamless migration, integration, and analysis of data at any scale. Snowflake provides unparalleled flexibility, cost-efficiency, and performance by decoupling storage and compute. Its advanced features, including automatic scaling, near-zero maintenance, and robust security, empower businesses to unlock valuable insights from their data. Snowflake's comprehensive data modernization capabilities position organizations to thrive in the era of data-driven decision-making, innovation, and growth.

Contents



C. Kalyana Chakravarthy

C. Kalyana Chakravarthy is a Snowflake Consultant. He is a data enthusiast with over a decade of experience in data warehousing, big data, and analytics. He worked on ETL tools like Abinitio and Informatica, and reporting tools like SAP BO. He is certified in Snowpro Core and is a Snowpro Advanced Architect and Snowpro Advanced Data Analyst.

LTIMindtree is a global technology consulting and digital solutions company that enables enterprises across industries to reimagine business models, accelerate innovation, and maximize growth by harnessing digital technologies. As a digital transformation partner to more than 700 clients, LTIMindtree brings extensive domain and technology expertise to help drive superior competitive differentiation, customer experiences, and business outcomes in a converging world. Powered by 82,000+ talented and entrepreneurial professionals across more than 30 countries, LTIMindtree — a Larsen & Toubro Group company — combines the industry-acclaimed strengths of erstwhile Larsen and Toubro Infotech and Mindtree in solving the most complex business challenges and delivering transformation at scale. For more information, please visit <https://www.ltimindtree.com/>