Capgemini

Snowflake & Capgemini, powering data & AI at scale

Technology without compromise

Table of Content

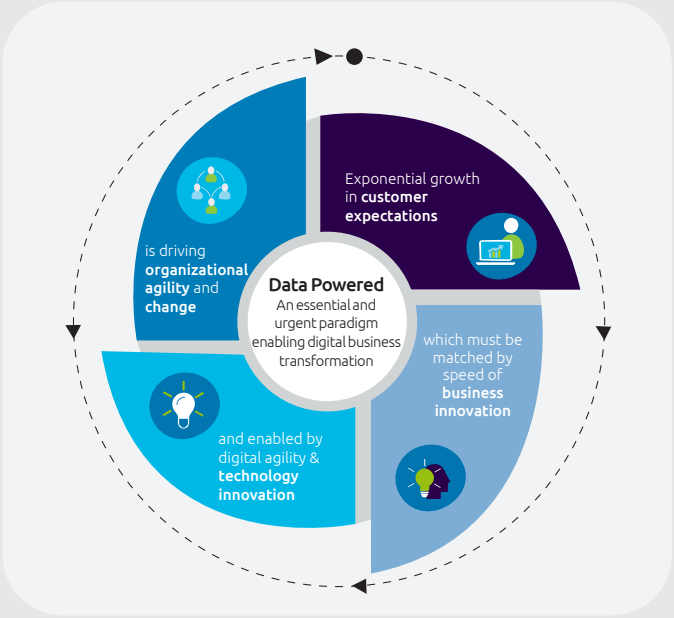
1. Understanding the **data challenges** of organizations

1.1 The speed of change is the new norm!

The average lifespan of an S&P 500 company is now less than 20 years, down from 60 years in the 1950s. This calls for a radical change in strategies and business models. The lines between different industries are blurred as cross-sector ecosystems of partners build new business models that were once unthinkable. Telcos are building banks. Online retailers are building physical stores.

The speed of change and business innovation required to meet customer expectations is becoming the new norm. Organizations need to be able to adapt quickly, inventing and reinventing themselves to survive disruption. This has never been highlighted more clearly than the disruption of the COVID-19 pandemic.

Secure, trusted, ethical managed data is powering organizations' ability to innovate and impacting their bottom line. To achieve this, organizations need to have met expectations and challenges coming from three directions.



Firstly, customer expectations

Customers expect organizations to embrace ethical AI, which in turn requires ethical data management. Ethical data management is the cornerstone upon which customer trust and loyalty are built.

Secondly, from data regulation

Increased global and local data regulation requires mature data privacy, quality and lifecycle management. Remember that not all data is governed equally, but it all needs to be governed appropriately.

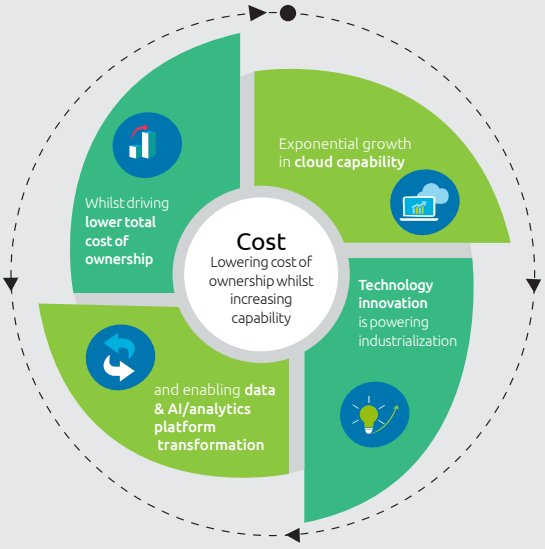
Finally, data powered operational excellence is driving change and complexity

Organizations need to be able to deliver value from the explosion of data available today, which demands agile management and governance while ensuring appropriate levels of data security and privacy. Additonally, an increased exploitation of AI & cognitive computing is both demanding & supporting better data quality & data orchestration.

1.2 Embracing change and the technology revolution

Technology is not new as a disruptive force, but it is accelerating change like never before. Emerging technologies continue to evolve, but many businesses are still in the dark as to how and where they might be integrated. Investment in Silicon Valley shows no sign of slowing down, with billions going into the development of experimental and potentially game-changing technologies.

IT departments are facing a tidal wave of new demand and technology emblement at a time when there is a drive to take cost out of the IT organization. There is exponential growth in cloud technology, which is powering technology innovation. The objective is to harness these capabilities to deliver on future demand while reducing the total cost of ownership.



Governance

- Data governance
- Uniform data taxonomies
- Data model
- Metadata management and data lineage
- Data ownership
- Data dictionary
- Aggregation methods
- Process automation
- Flexible aggregation
- Data security

Adaptability

- Flexible aggregation
- Flexible reporting
- Data availability
- Responsiveness

Reporting accuracy and capability

- End-user computing
- Flexible reporting
- Data quality management
- Reconciliation and sign-off
- Adjustments
- Risk models and aggregations.

1.3 What are the common challenges organizations have with their data estate?

There is an ever-increasing demand for data and AI capabilities to enable business innovation, while there is an increased focus on data governance, regulations, and ethical AI.

We are also facing the challenges of managing and innovating on a data and reporting ecosystem of technology which has been developed over the last couple of decades.

Legacy data and BI challenges

- Lack of business autonomy
- Excessive time to market (insights)
- Barriers to data and insights innovation
- Approach to governance and delivery is not aligned with business risk appetite
- Shadow IT delivering insights and reporting, agile but not scalable
- Duplication of effort increasing TCO for data and insights.

Early big data adoption challenges

- AI and big data pilot and Proof of Value fatigue, fast innovation, but no path to industrialization
- Data swamps, failing to deliver business benefits due to lack of data quality and governance

Common challenges

- Siloed implementations and data repositories
- No path from innovation to industrialization
- Poor data quality and governance (GDPR risk)
- Disparate worlds of big data and operations analytics.

The technology landscape of today is **unable to meet** the agility and innovation required by the **business and customer demands** of today and tomorrow.

1.4 Changing the data fabric of the organization

Data foundations power business decisioning, transformation, and innovation.

The democratization of secure, trusted, ethically managed data:

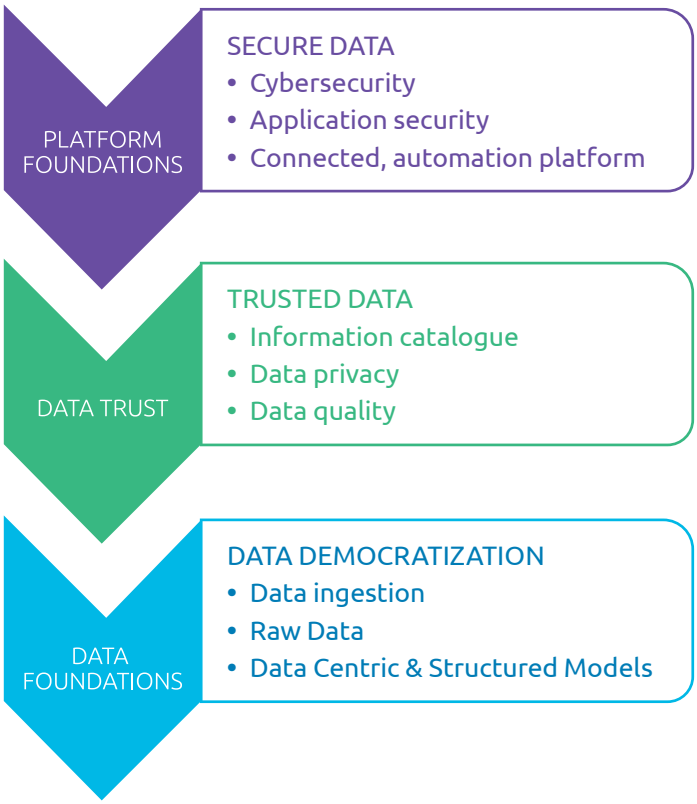
- Acting as an accelerator to innovation and industrialization, enabling more extensive use of agile methods
- Act as the single version of the data truth to support innovation and industrialization
- Though all data is not governed equally, ensuring all data is governed appropriately.

Empowering business and accelerating time to market

- Creating a data asset that supports business self-service, data science, and shadow IT
- Technology-enabled scalability, cross self-service, shadow IT, data science, and IT industrialized solutions.

Lowering the TCO, reducing IT debt

- Reducing IT debt generated by silo solutions that do not scale.
- Centralizing do once, use multiple times tasks, which increases the quality while lowering the total cost of solutions.



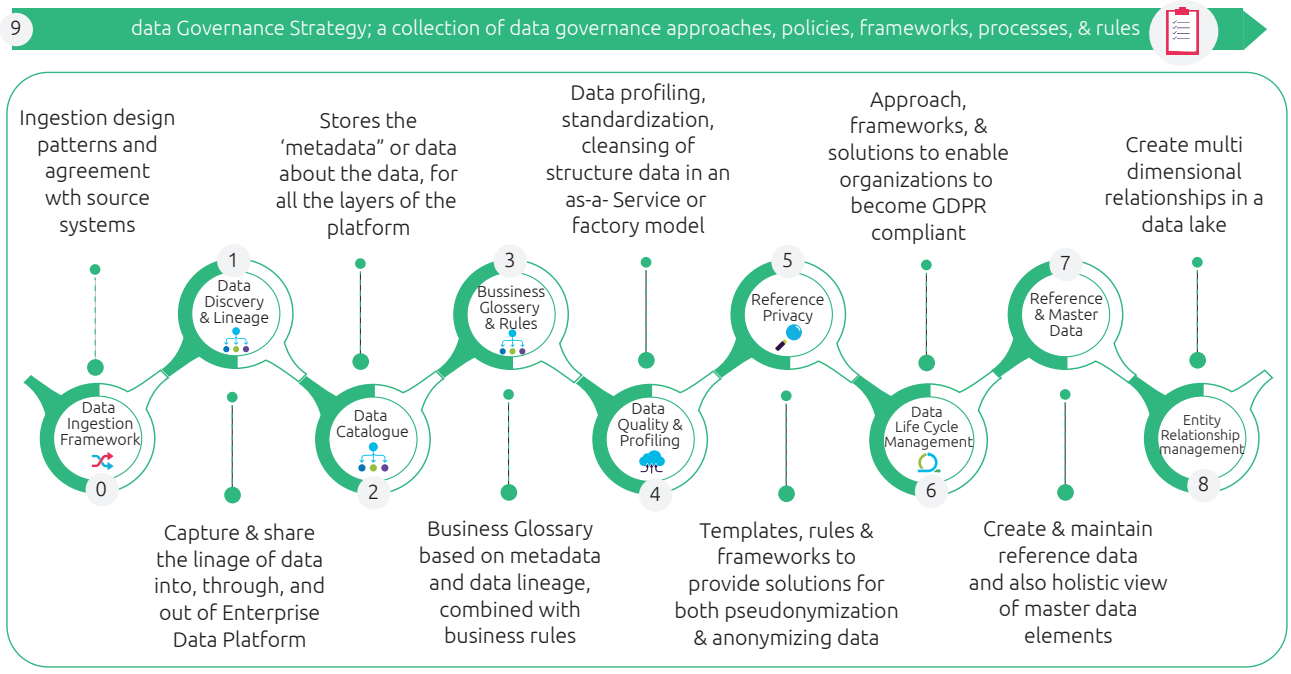
1.5 Putting data governance at the core of the data & AI/ analytics platform

Create a data asset and avoiding the data swamp

Data and AI/analytics platforms can become focused on the detail of collecting and storing high volumes of data in a data lake, losing sight of the bigger picture that data is in service of business insight, through the transformation of data into actionable information, creating a data asset.

Create a data asset

1. **The ingestion of data into the lake will require governance and security compliance to avoid creating a data swamp. Creating a data asset for the business:**
   - The definition of the incoming data from a business-use perspective **(business catalogue)**
   - Documentation of the **metadata, context, lineage,** and **frequency** of the incoming data
   - Data file storage structure and security model
   - **Security** level classification (public, internal, sensitive, restricted) of the incoming data
   - Documentation of creation, usage, privacy, regulatory, and encryption business rules that apply to the incoming data.
2. **Business ownership of data and management of the data:**
   - The **data owner** (sponsor) of the ingested data
   - The **data steward(s)** charged with managing the health of data items
   - Continuous measurement of the **data quality** as it resides in the data lake.
3. **It is recommended that the policies and processes for the consumption of the data from the lake are established to:**
   - Publish and maintain a data catalog and business catalogue to all stakeholders
   - Configure and manage access to data in the lake
   - Monitor PII, GDPR and regulatory compliance of usage of the data.



2. How can **Snowflake power** the data fabric of an organization?

2.1 Summarizing the challenge

Organizations are looking to:

- Increase the governance, quality and ethical management of data
- Increase business trust in data held by the organizations
- Increase solution quality and reliability
- Enable business self-service and innovation of secure data
- Remove the challenges and constraints of the siloed legacy estate while reducing the total cost of ownership.

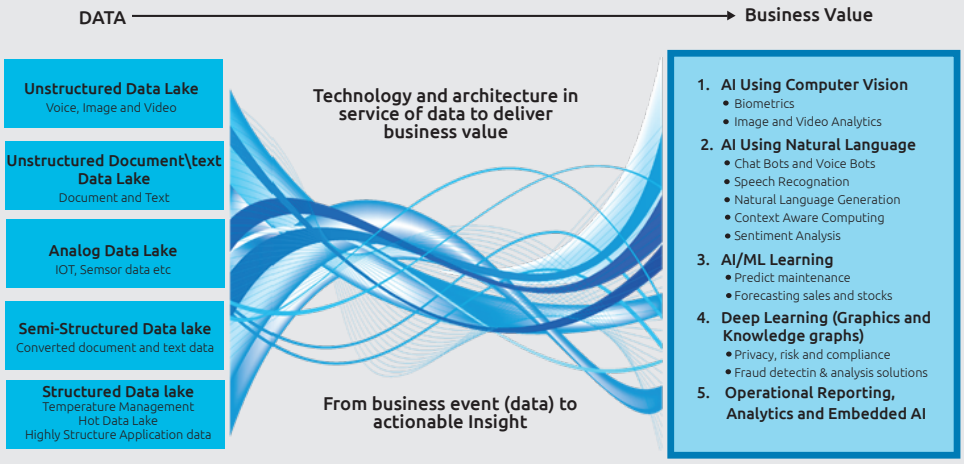
For many organizations around the globe, the move to the cloud is seen as a critical step in achieving these objectives. Organizations are moving to multi-cloud offerings with an increased need to share data across clouds, with secure data exchange.

How does Snowflake support these requirements whilst lowering the cost of ownership?

2.2 Introducing Snowflake

The Snowflake vision is to use cloud technology to leverage elastic compute and storage to create a cost-effective Data and Analytics Platform which solves the challenges with outdated on-premise EDW solutions, but also bringing together the siloed worlds of structured EDW and semi-structured data lakes capability into a single cloud data platform.

If we view the data foundations of a modern enterprise data and AI/Analytics platform as being constructed from five logical data lakes, see the diagram below:



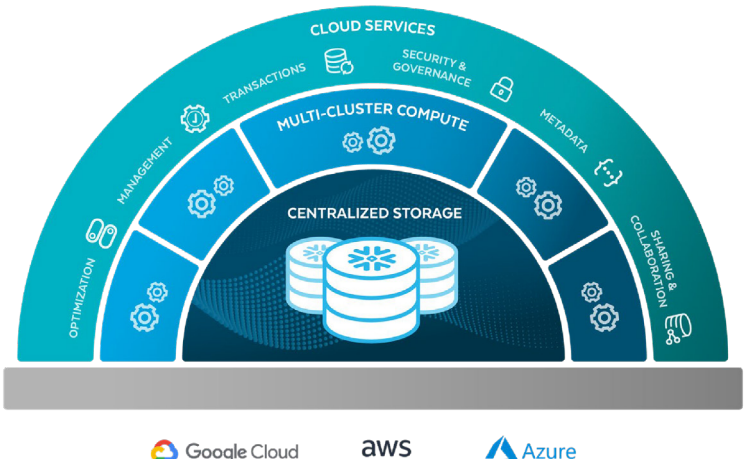
Snowflake supports the analog data lake, semi-structured lake and the structured data lake. The approach is to:

- Consolidate omnichannel data
- Focus on insights from data, not infrastructure
- Provide automatic performance tuning
- Be able to deliver sub-second response for repeated queries while lowering the cost of ownership
- Deliver secure and trusted data.

In technical language – Snowflake's architecture is a hybrid of traditional shared-disk and shared-nothing database architectures. Similar to shared-disk architectures, Snowflake uses a central data repository for persisted data that is accessible from all compute nodes in the platform. But like shared-nothing architectures, Snowflake processes queries using MPP (massively parallel processing) compute clusters where each node in the cluster stores a portion of the entire data set locally. This approach offers the data management simplicity of a shared-disk architecture, but the performance and scale-out benefits of a shared-nothing architecture.

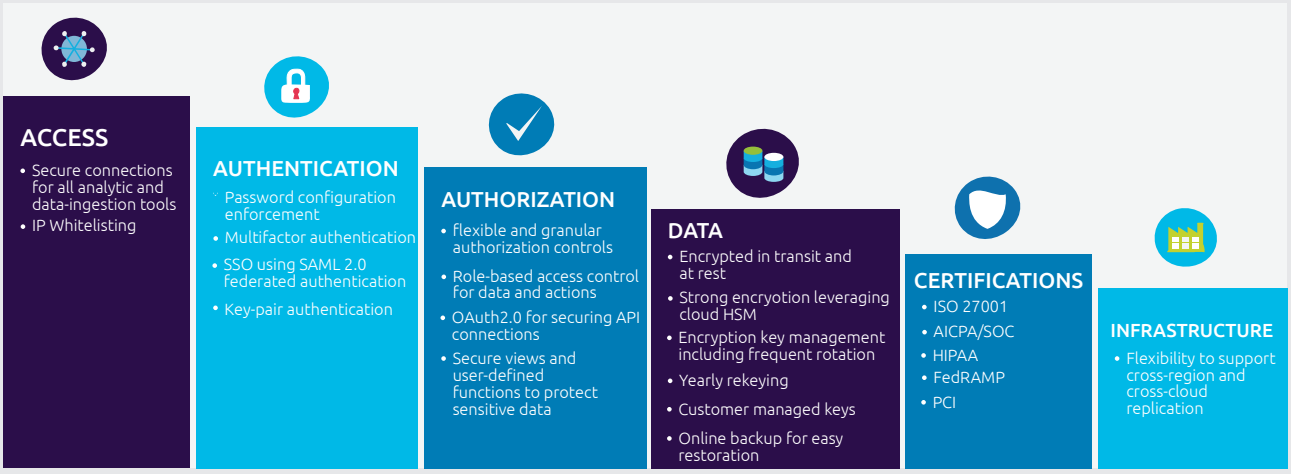
The Snowflake cloud data platform uses a new SQL database engine with a unique architecture designed for the cloud. To the user, Snowflake has many similarities to other enterprise data platforms. The architecture offers three layers: supporting database storage, query processing, and cloud services.

Snowflake manages structured data while using low-cost cloud storage, depending on the host cloud environment, AWS S3 buckets, Azure Blob storage, and Google Cloud Storage.



2.3 Security and compliance

The Snowflake cloud data platform is built on a multilayered security foundation that includes encryption, access control, network monitoring, and physical security measures, in conjunction with comprehensive monitoring, alerts, and cybersecurity practices. Every aspect of the platform is geared toward protecting your data, both in transit and at rest.



These control mechanisms can be applied to all database objects, including tables, schemas, and any virtual extensions to a data set. Multi-factor authentication procedures can issue a secondary verification, such as a one-time security code sent to a user's mobile phone. However, all this comprehensive security doesn't stand in the way of usability. Single sign-on procedures and federated authentication technologies make it easy for people to log into a data platform, data lake, or another analytic service directly from within those applications.

2.4 Governance

Remember that not all data is equal, but all data needs to be governed appropriately.

Robust transaction management and data partitioning

Guaranteeing the integrity of database transactions is essential in many industries. This is especially important for organizations that handle time-sensitive data, such as financial services companies that conduct monetary transactions and manufacturing firms that run real-time production processes. Snowflake enforces database atomicity, consistency, isolation, and durability (ACID) properties to guarantee transaction consistency even in the event of unforeseen errors, power failures, and other mishaps.

Adherence to ACID properties ensures database accuracy and transactional integrity, while micro-partitions - contiguous improves database performance. As data is loaded, Snowflake transparently divides it into micro-partitions contiguous units of storage organized in a columnar fashion. This structure allows for extremely granular pruning of very large tables, which can be composed of millions of micro-partitions to significantly improve query performance.

Metadata management forcing accuracy and consistency

A robust metadata service spans the entire system. Queries are compiled within the services layer, and metadata is used to determine the micro-partition columns that need to be scanned. This makes it possible to track where data is coming from, who touched that data, and how various data sets relate to one another.

External tables store file-level metadata such as file paths, version identifiers, and partitioning information. This enables multiple workloads to query one single copy of the data with transactional consistency.

2.5 Simplicity instead of silos

Snowflake's multi-cluster, shared data architecture provides virtually unlimited concurrency and performance on a single copy of the data. This eliminates the need for multiple data marts to materialize the data for end-user consumption and allows the data marts and EDW to be combined into one repository. This simplifies the architecture, creates a single version of the truth, and reduces the cost of ownership.

Snowflake is increasingly used as a replacement for NoSQL data lakes for simpler, faster data exploitation and consumption. This new, modern cloud data platform combines data lake, EDW and data marts is an SQL based solution that offers virtually unlimited storage and compute. The access to data lake and EDW and Data Lake can be logically separated and controlled via user access allowing users to seamlessly move between the two.

**Data Engineering, Data Lake, Data Warehouse, Data Science, Data Applications and Data Exchange.*

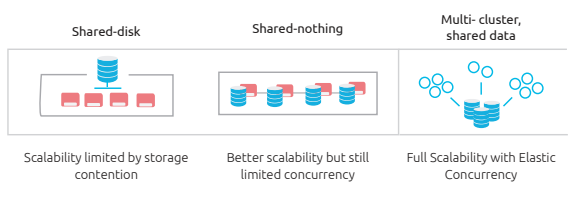
2.6 Scalability & performance – Petabytes with decoupled limitless compute

Snowflake is designed to scale up, but more importantly, scale down. Compute is charged per second by compute cluster (with a minimum of 60 seconds) and storage is charged based on the data stored. Snowflake uses cheap cloud storage from each of the cloud hyper-scalers AWS S3 buckets, Azure Blob storage, and Google Cloud Storage.

Scalability

The virtual warehouse (compute resource) can be sized with different level of compute. The virtual warehouse can also be scaled up and down on the fly while queries are running independently of other virtual warehouses. The new sizes take effect when the next query starts.

Virtual warehouses have role-based security, which can include limits to the compute a user can access.



Performance

Snowflake is an MPP system, scaling up involves adding nodes to an individual cluster. Scale-up in Snowflake is used to improve query run time but doesn't improve concurrency. As with other MPP solutions snowflake uses a queueing process for queries waiting for computing resources.

To achieve greater concurrency in Snowflake, scale-out is required. This is achieved by creating additional independent clusters (virtual warehouses). Snowflake also supports the option of auto-scaling multi-cluser warehouses.

2.7 Workload isolation and cost management

Fast performance is not enough to support the data and analytics requirements of organizations today. The platform needs to be delivered in a cost-optimal manner, with supporting functionality which lowers the total cost of ownership.

The separation of compute and storage allows Snowflake to implement a robust workload isolation capability, complex schema support, resources, and pricing controls to enable the environment to be managed to support performance and cost SLAs with confidence.

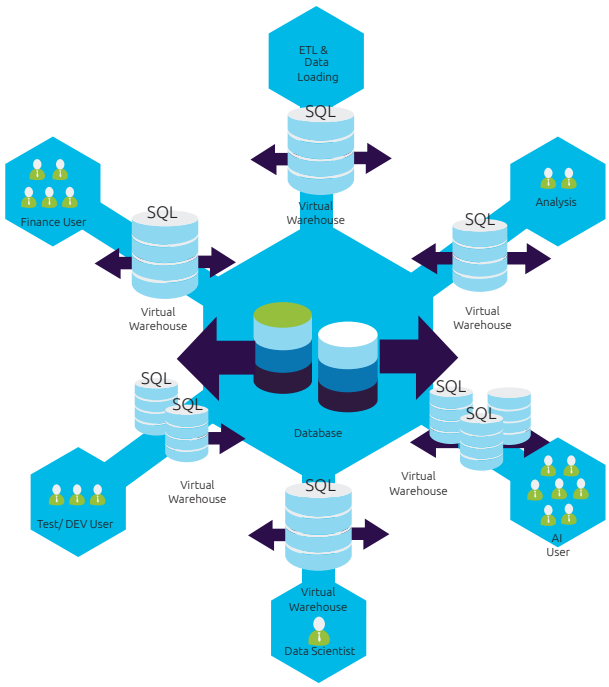
Each virtual warehouse is an independent compute cluster that does not share compute resources with other virtual warehouses. As a result, each virtual warehouse has no impact on the performance of other virtual warehouses.

This allows the sizing, scalability and concurrency for each virtual warehouse to be defined in line with the demands of the workloads being processed.

Key cost optimization functionality includes auto suspend and auto-resume of virtual warehouses.

For complete scale down functionality - when a virtual warehouse isn't in use you aren't paying for it.

Storage charges in Snowflake are a simple "pass-through" of the cloud vendor's storage charges.



2.8 Data ingestion

A common challenge is keeping pace with IoT scale and concurrency demands. This is resolved by Snowflake's independent compute clusters, which enable near real-time data loading and reporting without contention or additional management.

The Snowflake cloud data platform includes a serverless ingestion service called Snowpipe that asynchronously loads data into your cloud storage environment. Standard connectors and adapters allow organizations to easily ingest event streams from Kafka and other messaging systems, while Snowflake streams and tasks make it easy to schedule data loads for SQL jobs.

The platform automatically transforms data into the type and shape required for each target table. For example, Apache Kafka connector lets you continuously stream JSON records for storage and analysis.

- Real-time streaming of structured and semi-structured
- Ability to scale to support high volume and velocity of IoT data
- Ability to support change data capture ingestion.

Snowflake's native SQL querying of large semi-structured data sets eliminated time-consuming data parsing projects, accelerating the team's ability to innovate and build industrialized dashboards and analytics.

2.9 Data marketplace (discoverable yet governable data assets)

Snowflake's Data Marketplace offers an alternative to traditional data sharing methods, which require an organization to share physical copies of data with the data consumers. This approach results in static versions of data being shared, which require frequent updates. In addition, this process is cumbersome, costly, risky, and can lack the ability to secure sensitive information or prevent data breaches.

Snowflake's data sharing capability enables authorized members of a cloud ecosystem to tap into live, read-only versions of the data. This allows organizations to easily and securely share subsets of your data, as well as receive shared data in a secure and governed way. For example:

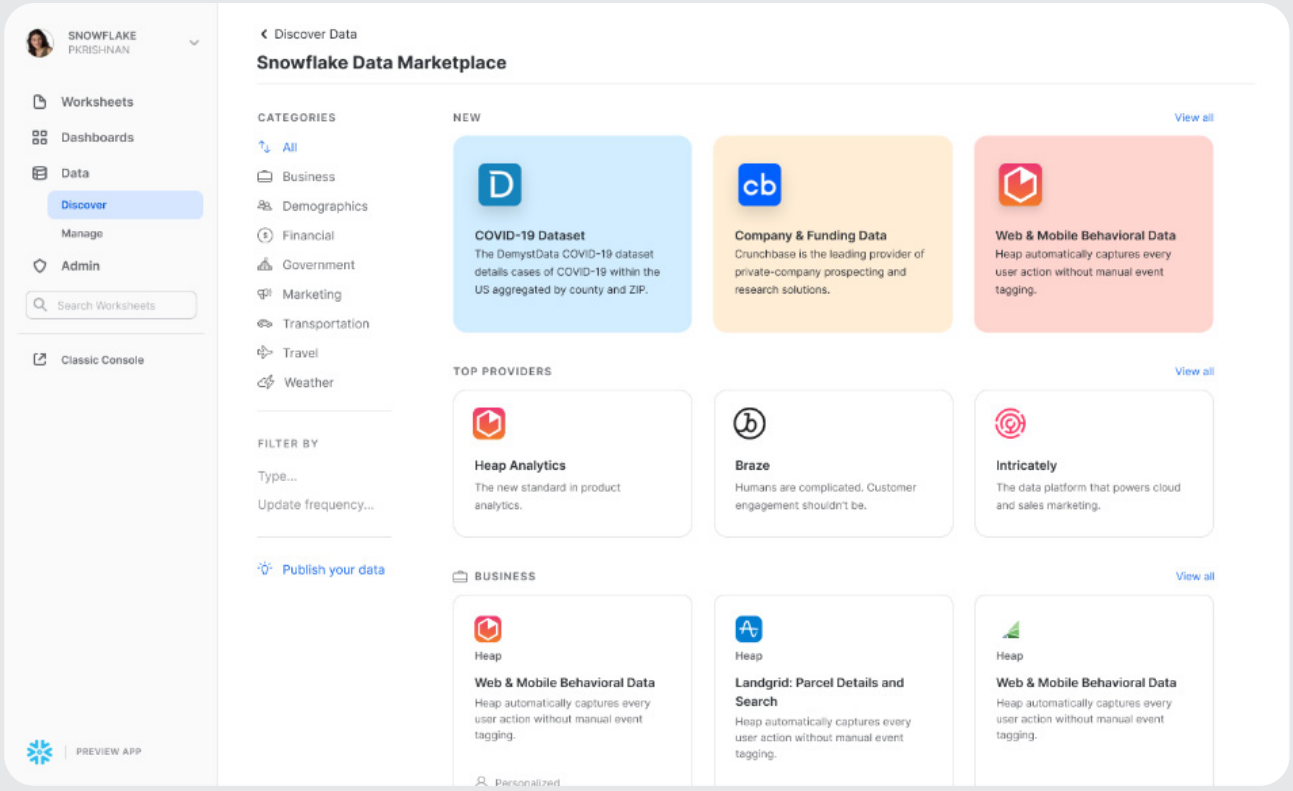
- Leverage public data sets from the Snowflake Data Marketplace and combine those data sets with your own data to gain even deeper insights and make data-powered decisions.
- Create and host customer's data in a secure environment, giving their users the ability to discover and securely access shared data directly from their Snowflake account.
- Create customer's own data exchange and invite their employees, subsidiaries, partners, customers, and others to securely access their data sets without having to move, copy, or transfer that data.

Forrester Research reports that 90% of global data and analytics decision-makers are making it a priority to improve the use of data and analytics insights in their business.

1) Data Marketplace – Snowflake Data Marketplace interface

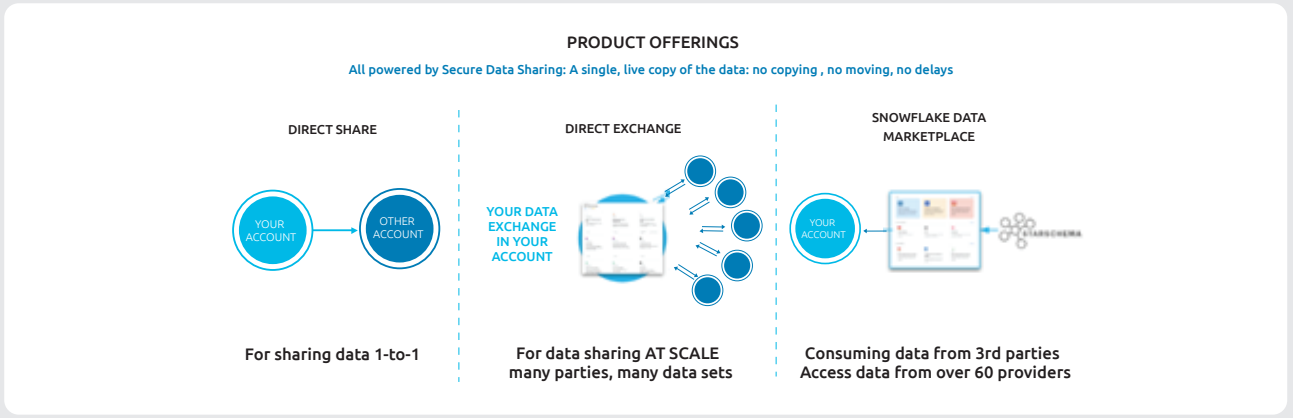
- Snowflake facilitates the publishing and distribution of data published by a data provider. As there's no ETL and data movement, the data is always live. There are also no storage fees for data consumers.
- Data consumers can instantly discover, query, join, and enrich data from the marketplace with their own data inside their Snowflake instant. They have full access to live, granular, and historical data without manual intervention.
- Provides data listings with public data like census data.
- Enables data consumers to connect with data providers and seamlessly discover access and generate insights from the provider's data.
- Allows data providers a mechanism to promote their data services across Snowflake users and create new revenue streams.

1) Data Marketplace – Snowflake Data Marketplace interface *(Continues)*



2) Data exchange

- The organization controls and manages the publishing process as well as who can access and discover data.
- The centralized location from which different internal teams in the org. for example analytics and data science teams can access secure data, thereby improving both data integrity and security and leading to "shortened time to insights."



2.10 Cross-cloud

Organizations are moving to the cloud in multi-cloud strategies, acknowledging that innovation does not reside in one cloud provider, systems integrator or technology vendor embrace the reality of hybrid multi-cloud, multi-vendor approach. There are a number of challenges:

- Cloud silos are created as soon as data exists in a public cloud. Because each major cloud provider created a unique offering with proprietary APIs for data management, there's no easy way to copy or share data from cloud to cloud.
- Cloud services work best when users are in close proximity. As a result, geography plays a role in creating data silos by region, especially for organizations that operate in multiple locations (regions, countries, and continents).
- Data portability is a problem for all organizations, including those that use open source technologies and open data formats. Today, there's no easy way to lift multiple petabytes of data to change clouds, open-source or otherwise.

Snowflakes cross-cloud capability allows organizations to securely share data across regions and cloud accounts while adhering to the same rules of data sharing. Note: data exists locally in a single source where it's accessed rather than moved.

This functionality powers the ability to:

- Analyze all data for decision making, no matter where the data is located.
- Ensure business continuity and disaster recovery through cross-cloud replication.

Cross-cloud capability delivers the unified data management platform needed to enable secure data sharing, fully execute multi-cloud strategies, and provide organizations with a single source of truth. By enabling data to move freely, cross-cloud capability delivers on the promise of multi-cloud strategies.

The strengths of Snowflake at a glance

1) It is an analytical hybrid columnar MPP that is a departure from shared disk and shared-nothing architectures.
2) Compute and storage are decoupled, and metadata is managed separately. Shutdown all compute, and you still have your data.
3) Supports structured and semi-structured data, with native support for JSON through SQL making it one of the best platforms for processing JSON data. Eliminate pre-processing or transformation of semi-structured data.
4) Easily ingest data from multiple sources, social media, traditional media, sensor data, customer profile, product data, etc.
5) Run multiple workloads across shared data all on one platform, without compromise.
6) Explore patterns and relations between data types previously not relatable.
7) Use robust ANSI SQL familiar to millions of users and business intelligence/data-mining tools. Fully leverage SQL skills and expertise.
8) Scale-out and scale0-in is very fast without overheads of data redistribution and need for quiet time on a cluster.
9) No need for indexing, statistics generation, rebalancing cluster or other administrative overheads. A leading cloud data platform to manage high concurrent workloads, meeting SLA.
10) Secure, trusted, data democratization through public and private data sharing, and cross-cloud capabilities

One Platform
One Copy of Data,
Many Workloads

Secure &
Governed Access
to All Data

Near- zero
Maintenance
as a Service

Unlimited
Performance
and Scale

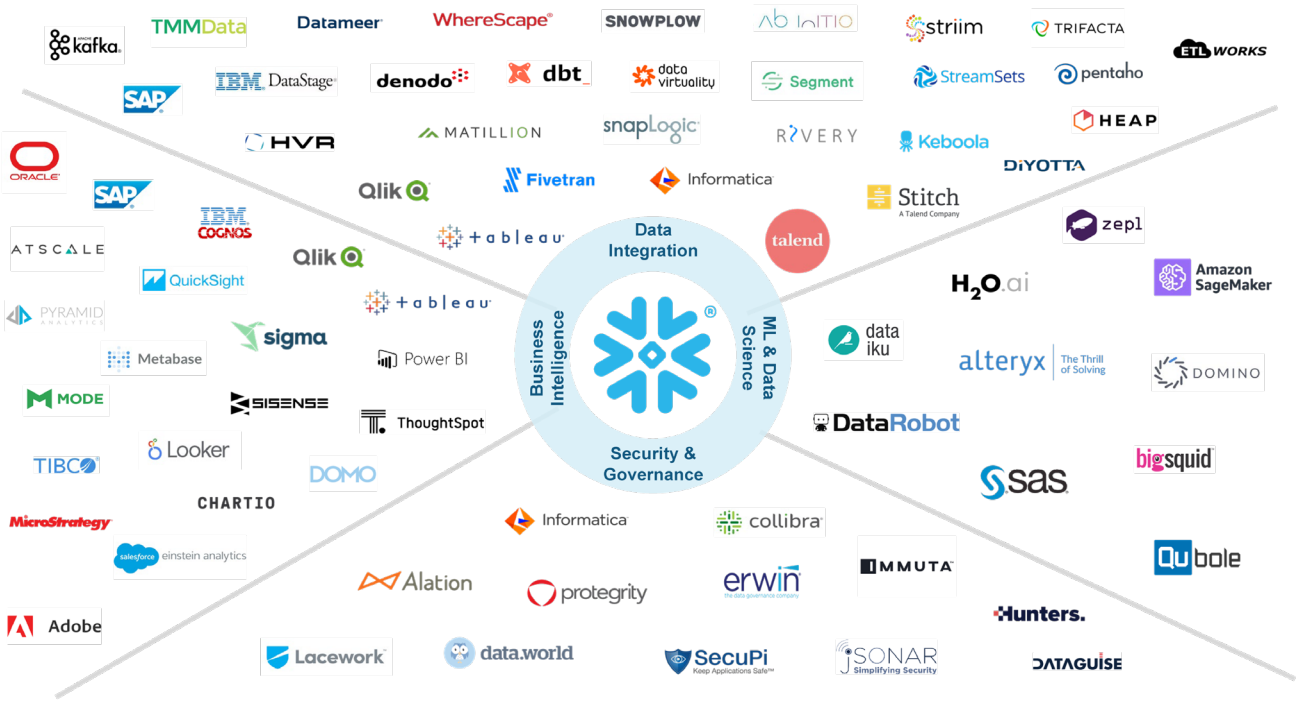
3. Snowflake value proposition

3.1 Snowflake is a solution where you can have your cake and eat it

In conclusion, what does Snowflake offer

	The organizational Ask		With added benefit (no compromise)
1	The ability to have a multi-cloud strategy	+	with an integrated data fabric across hyper-scalers and regions
2	The ability to leverage low-cost elastic cloud storage	+	with the governance, metadata and security of a role/ column-based EDW
3	The ability to bring the data lake, EDW and semi-structure data into an integrated data fabric	+	without the complexity in tooling and skills of a big data platform
4	The ability to leverage the best in market BI and AI/ ML tooling	+	being able to leverage the years of incumbent SQL skills in your organization
5	The sophistication of scale-up and scale- down compute to manage costs, with the ability to isolate workloads and commit CPU based on the criticality of activity	+	without the need to develop and support the DevOps tooling to support automated scaling.
6	Ability to ingest semi-structured, structured and high-velocity, real-time data (e.g. IoT)	+	whilst simplifying the ingested process and eliminate pre-processing or transformation of semi-structured data.
7	Out of the box functionality which lowers the cost of development and increases solution quality, e.g. time travel, change data capture, complex joins and star schemas, etc.	+	avoiding custom coding and ongoing support costs
8	The ability to share data securely within and outside the organization, across regions and cloud hyper-scaler AWS, Azure and GCP	+	Unlocking the ability to monetize data
9	Single governance and security model across cloud regions and hyper-scaler	+	Avoiding complex duplication of data security
10	Supporting multi could data and solution replication with the ability to failover across regions and hyper-scaler	+	Creating a new level of solution and data resilience, whilst simplifying the data fabric across cloud providers.

It is true to avoid cloud lock-in you are committing to Snowflake, but also a culture of integration within the innovation marketplace that is exploding in the cloud.



3.2 Continued innovation and expansion of partner relationships 2020

Snowflake continues to innovate and partner to grow the capability of the product to meet customer demands, below is a summary of the 2020 announcements:

CORE PLATFORM FEATURES

- Snowsight
- Transparent usage of materialized views
- Snowflake-Salesforce partnership
- New and larger sizes of compute clusters
- Search optimization service
- SQL stored procedures
- Geospatial data
- Dynamic data masking.

EXTENSIBLE DATA PIPELINES

- Partition-aware exports
- External functions
- Java functions.

DATA CLOUD CONTENT

- Snowflake-Salesforce partnership
- Einstein Analytics Direct Data for Snowflake
- Data Exchange
- Snowflake Data Marketplace.

About Capgemini

Capgemini is a global leader in consulting, digital transformation, technology and engineering services. The Group is at the forefront of innovation to address the entire breadth of clients' opportunities in the evolving world of cloud, digital and platforms. Building on its strong 50-year+ heritage and deep industry-specific expertise, Capgemini enables organizations to realize their business ambitions through an array of services from strategy to operations. Capgemini is driven by the conviction that the business value of technology comes from and through people. Today, it is a multicultural company of 270,000 team members in almost 50 countries. With Altran, the Group reported 2019 combined revenues of €17billion.

Learn more about us at

www.capgemini.com

Learn more about Snowflake at

snowflake.com

Fiona Critchley
Global I&D Portfolio Lead Data Foundations
& Data Estate Modernisation

People matter, results count.