Snowflake, *the* Cloud Data Platform (2023)

It's been one year since my last article describing Snowflake the Data Cloud, see bit.ly/3xO2B7Y. Let's investigate how Snowflake has scaled since the announcements at Snowflake Summit July 2022!

Within this article we'll highlight the major Summit announcements from last year and other features added to Snowflake since 2022, for the core elements please refer to the bitly link above. If you compare the above diagram to the diagram from a year before you'll notice subtle additions to the platform but still based around the three key elements, you will need to utilise to scale on Snowflake the data cloud:

- **Data Objects** to store your data (internal, external OLAP and OLTP (***NEW***)),
- **Compute** to do work with your data, and
- **Roles** to combine storage and compute, every action/object must have role owner

All other elements and features in Snowflake build on these 3 core elements. Individually **they do nothing** but together you can define the constraints and scale to build almost anything you desire on the data cloud.

Has what's in the box changed?
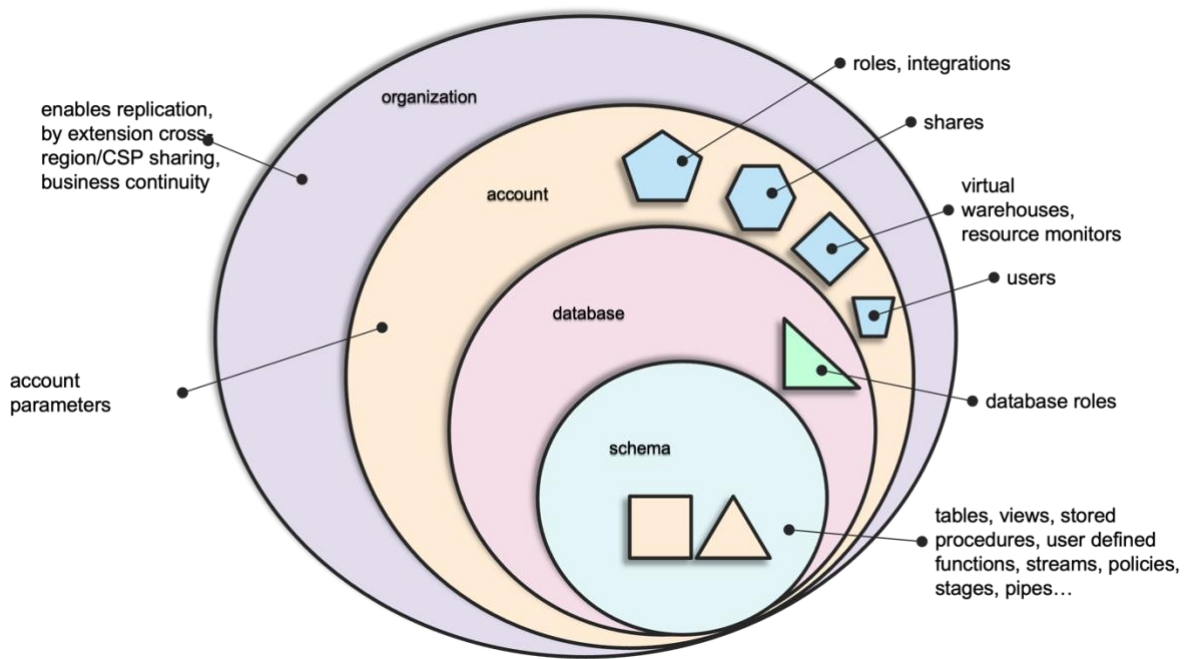
Figure 1 Snowflake Object Hierarchy

To a large extent, no. But we do have the first database object that is not a schema object, the **database role**. *Why do we need it? See Native Apps below!*
Let's dive into Snowflake's capabilities! Note that many of these have existed pre-'22 but have matured with new features and sub-capabilities!

Data Architecture

We see many customers that want to build a **data mesh**, Snowflake is **not** a Data Mesh platform itself, but Snowflake has the technology to support your Data Mesh **initiatives** with its built-in features, starting with…

1. Data Governance

Figure 2 Data Governance 2023… and growing

Know Your Data – is to understand, classify, and track data and its usage

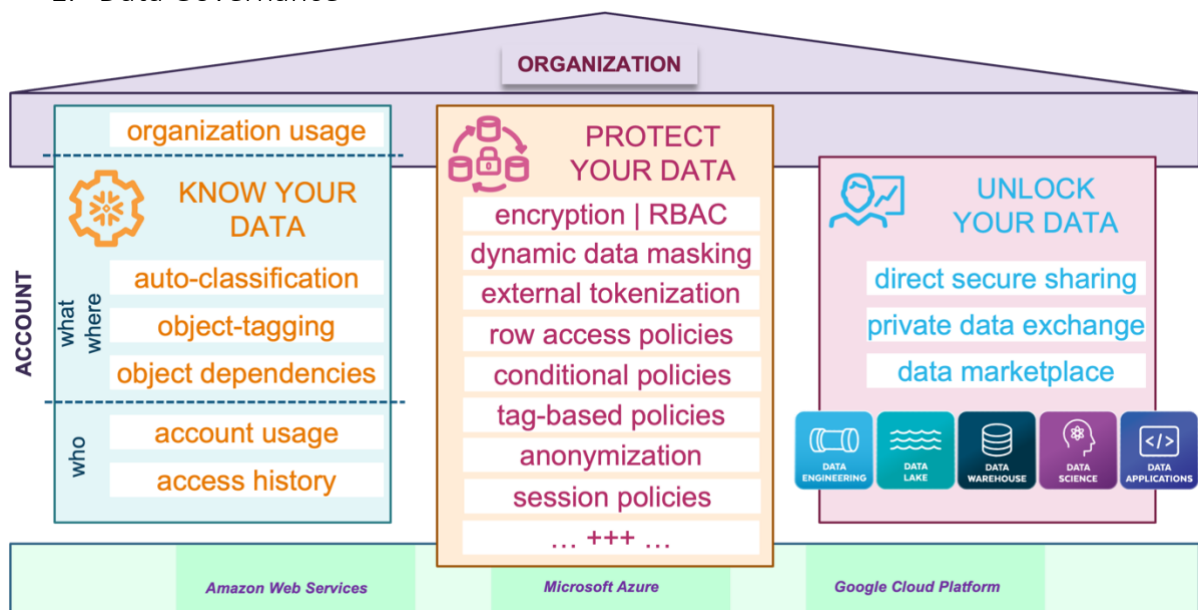- account_usage (and read_account_usage) – contains account metadata and usage metrics for up to one year.
- access_history – contains metadata about which Snowflake roles accessed data artefacts, what objects were created based on that access (including column lineage) and what policy was in place when that data was accessed.
- object_dependencies – to detect what the impact would be if changes are made to a Snowflake object
- Object tagging –tagging Snowflake objects with key-value pairs such resources belonging to a business domain.
- Auto-classification – using Snowflake's own trained classification model to tag content with a semantic and privacy categories
- organization_usage – tracking your cluster of Snowflake accounts under your organization, including billing data

Protect your data – to secure sensitive data with native encryption and your own custom policy-based access controls

- Encryption – world class end-to-end and key-management
- Role-based access control (RBAC) – industry best practice for managing access to your data.
- Dynamic data masking (DDM) – built-in column level security policies
- External tokenization – similar to DDM, but utilizing an external partner through external functions to mask data in place
- Row access policies – dynamically isolate rows of a data object based on your RBAC,
- Conditional policies – applying DDM based on a column value in the *same* table.
- Tag-based policies – applying tags from database or below and all objects in the hierarchy inherit the policies applied based on data type.
- Anonymization – applying k-anonymity or data hierarchies to anonymize data.
- Session policies – designed to control session timeouts
- +++ - *more to come!!!*

2. Domain-oriented architecture

When you decide on a decentralised, hybrid or governed approach to your Data Mesh, then consider the following.

- Using Snowflake object tags and RBAC to tag ownership of Snowflake account resources under a domain. This can be extended with Snowflake's budgets and resource group objects to alert and suspend credit consumption within the Snowflake account; in addition to Resource Monitors.
- If each domain has their own Snowflake account, then also consider managing access and data product sharing in a private data exchange.
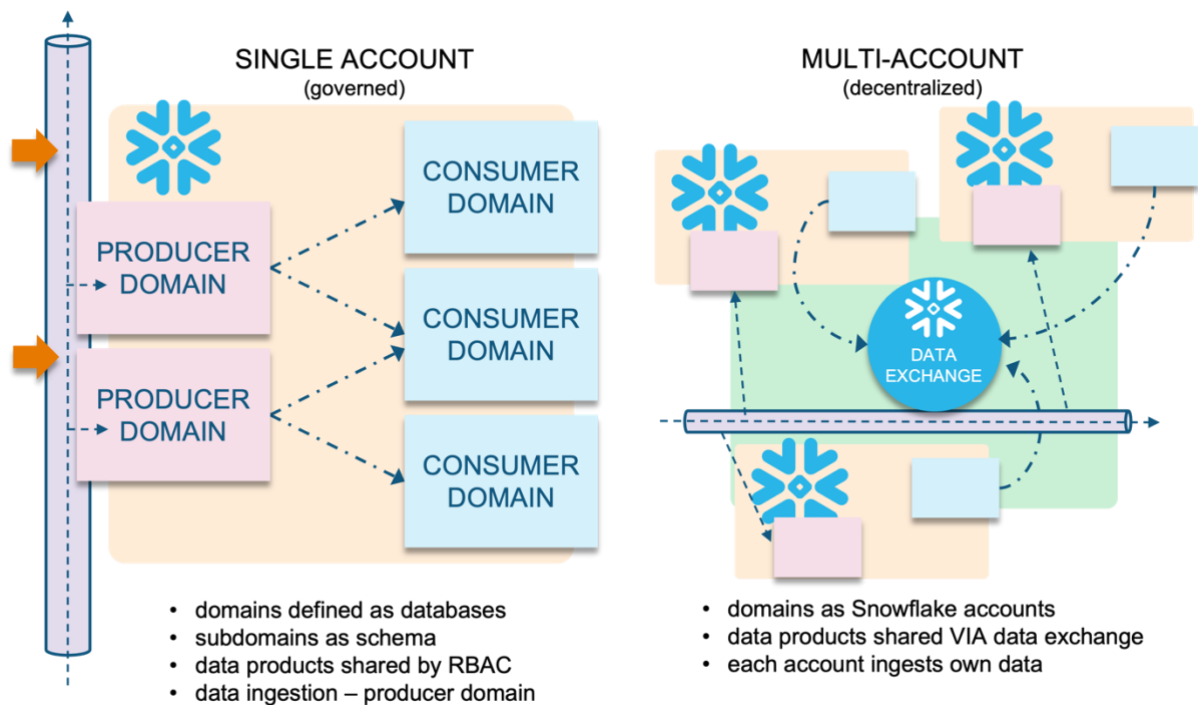
Figure 3 Considerations for your Data Mesh on Snowflake

- Streaming ingestion *(aka dumb pipes)* as new business events are streamed, Snowflake's new Streaming API for Snowpipe enables rowset ingestion into a Snowflake account. This enables you to insert directly into Snowflake tables without the need to stage the data first using Snowpipe + Kafka.

3. Business Continuity and Disaster Recovery

Snowflake inherits the guarantees of the cloud service provider (AWS, Azure, GCP) for fault tolerance, durability, and scalability; you may, however, have the requirement to support multi-region redundancy in the form of business continuity and replication over and above what Snowflake already provides by default. Yes, Snowflake supports business continuity too in the case if a CSP's region goes down, you may require the assurance your business can continue operations without disruption and with Snowflake's replication and failover technology that is possible. Utilizing Snowflake's organizations capability, you can asynchronously replicate your Snowflake account(s) and/or database(s) in one region to another and failover your Snowflake supported connectors to the secondary Snowflake account with minimal disruption. Snowflake supports:

- Database replication at the database level-*only* (excludes privileges, account parameters). These include replication groups, streams, materialised views, tags, passwords, policies. Database replication can be configured to be cross-region or within the same region and replication and failover can be configured to multiple target accounts within an organization.
- Account replication replicates the entire account.
- Client redirection to enable redirecting your client connections (ex. JDBC, ODBC, Snowpark, SnowSQL, Python Connector, more…) to Snowflake accounts in different regions.

Keep in mind your jurisdictional responsibilities, every customer is different. Snowflake provides the technology to make cross-region replication possible, but it is up to you to ensure you're doing so within your regulatory and legal boundaries!

4. Data Model

More Data Vault? As we discovered in the previous article, Data Vault 2.0 is an INSERT only data model and it suits Snowflake traditional tables very well. We have released a series on tips and techniques on how to deploy your Data Vault on Snowflake

Data Vault on Snowflake Techniques
1. Immutable Store, Virtual End Dates, [bit.ly/3LWR8qR](bit.ly/3LWR8qR)
2. Snowsight Dashboards for Data Vault, bit.ly/3QPhM7T
3. Point-in-Time (PIT) Constructs and Join Trees, bit.ly/3wfCBAG
4. Querying Really Big Satellite Tables, bit.ly/3BcuRkE
5. Streams and Tasks on Views, bit.ly/3Smfi0u
6. Conditional Multi-Table INSERT, and Where to Use It, bit.ly/3C9YPHw
7. Row Access Policies and Multi-Tenancy, bit.ly/3CwfOmf
8. Hub Locking on Snowflake, bit.ly/3WbxsVo
9. Out-of-Sequence Data, bit.ly/3WsRjzD
10. Virtual Warehouses & Charge Back, [bit.ly/3EsV2F7](bit.ly/3EsV2F7)
11. Handling Semi-Structured Data, [bit.ly/3UmB4Sn](bit.ly/3UmB4Sn)

Now if you want to explore Data Vault on Data Mesh, look no further then this blog series!
1. Data Vault and Domain Driven Design, bit.ly/3KMPSGS
2. Data Vault as a Product, bit.ly/3TYHrfY
3. Data Vault and Domain Oriented Architecture, bit.ly/3qmUeLz
4. Data Vault semantics & ontologies, [bit.ly/3RNH9GD](bit.ly/3RNH9GD)
5. Data Vault and Analytics Maturity, bit.ly/3SrPqQt

A data vault on Snowflake is possible but you can also bring your preferred data model style and data architecture to Snowflake. The above is to highlight the many built-in features and techniques you should consider when bringing your data model (even if it's not a data vault)!

Data Science

"…the moment you need to move data from one architecture to another you're losing business value …"- certainly there is a cost to moving data from one data platform to another simply because a data platform may not have the capabilities you need to derive that value.

Snowpark (portmanteau of Snowflake and Apache Spark) is Snowflake's answer to bringing the data science (and advanced ETL/ELT) workload to the Snowflake platform. Designed to resemble Apache Spark in interface (there is no Apache Spark within Snowflake) but without pulling the data *out of* Snowflake, Snowpark offers the opportunity to run your data science/engineering workload to the data already accessible within Snowflake.
No additional fees, follow the setup guide and start developing in **Java**, **Scala,** or **Python** in addition to SQL. Here is a demo on using dbt with Snowpark.
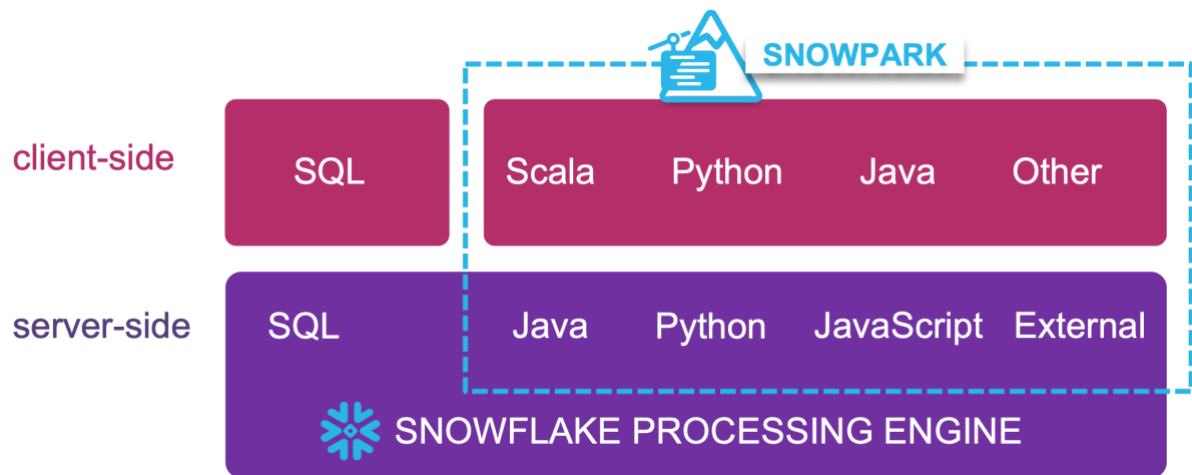
Figure 4 Generally Available, Snowpark for advanced ELT/ETL (beyond just SQL) and Data Science workloads

All the above languages are in General Availability support and include (but **not** limited to):

- [High-memory](#) virtual warehouses for running machine learning workloads
- [Python UDF Batch API](#) for batches of inputs
- [Snowpark API](#) library itself
- Secure [Anaconda](#) libraries
- [Unstructured data support](#), bring your images, PDFs, and other blob objects into Snowflake to run your data science workloads on!
- *and more…*

New supported table types
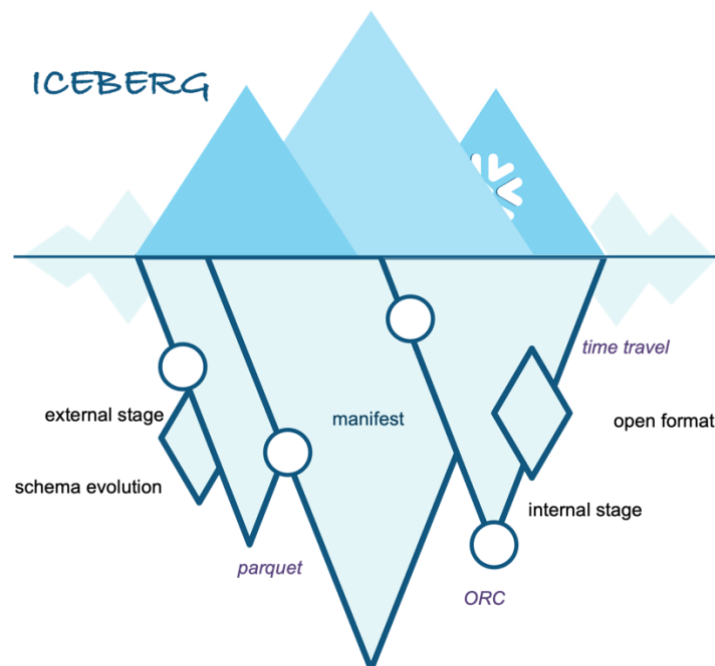Expanding Snowflake's storage options to

Apache Iceberg

Figure 5 Apache Iceberg, supported on Snowflake

As the Big Data revolution evolves, new analytical table data formats have arrived to improve performance at open-source scale. One such innovation is Netflix's [Apache Iceberg](#)

table format offering performant analytical value at big data scale with ACID. Snowflake supports this table as a read only table in external stage and as a native table in internal stage.

Within Snowflake we still prefer that if your data hasn't already been committed to Iceberg that you use Snowflake traditional tables. Snowflake traditional tables for OLAP is what the Snowflake platform is based on, from optimized caching to time-travel (+Fail Safe) and all the already supported optimization like Search Optimization Service (SOS).

Hybrid Tables

Announced at Summit '22, Snowflake's UNISTORE workload is Snowflake's focus on bringing OLTP database functionality to the platform and the first artifact is Snowflake's Hybrid Table. A single table type that combines both OLTP (row-optimised) and OLAP (column-optimised) into a single table to bring you a unique HTAP experience. Hybrid tables will honour table **constraints** such as **primary keys**, **foreign keys**, **table uniqueness**. It will also inherit all the functionality of traditional Snowflake tables *(ex. Time-Travel)* whilst being able to join to any other table type within Snowflake.

Snowflake's traditional tables are batch-oriented, column optimised tables with response/update times in the 100s of milliseconds, Hybrid tables are first row-optimised with response/update times in the 10s of milliseconds.

Use cases for Hybrid tables at this stage include:

- Data application support, for those apps requiring millisecond response time supported by row-optimised databases
- Application state, for those control frameworks needing rapid response times to able to track and control your ELT/ETL pipelines.
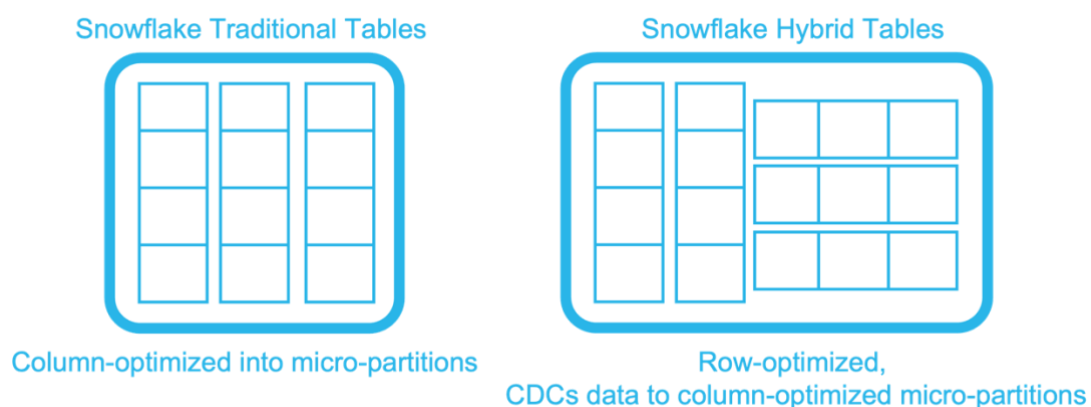- Feature stores, for online inferencing
- *and more…*

Figure 6 Snowflake's new Hybrid table

It is still early into Snowflake's OLTP journey, expect more use cases in the coming quarters as Snowflake brings more and more features (reasons) to keep your data in the Data Cloud!

Dynamic tables

Also announced at Summit '22 (as materialized tables), dynamic tables bring something different to Snowflake customers.

Back in 2020 Snowflake had released a set of features called "Streams & Tasks" that streamlined your data pipelines by explicitly defining a Snowflake object called a "stream" (offset) on a table, a share, or a view. You could then define another internal object called a

"task" to periodically execute your DML to process new or changed data as defined by your stream offsets.

Dynamic tables are the **declarative** form of streams & tasks, you simply define a table definition with syntax similar to an SQL CTAS operation but with a *lag* option and Snowflake dynamically defines stream offsets on the dependent table objects as defined by your SQL syntax and dynamically defines tasks based on that lag (currently as low as a minute but in the future this will go down to the seconds!).
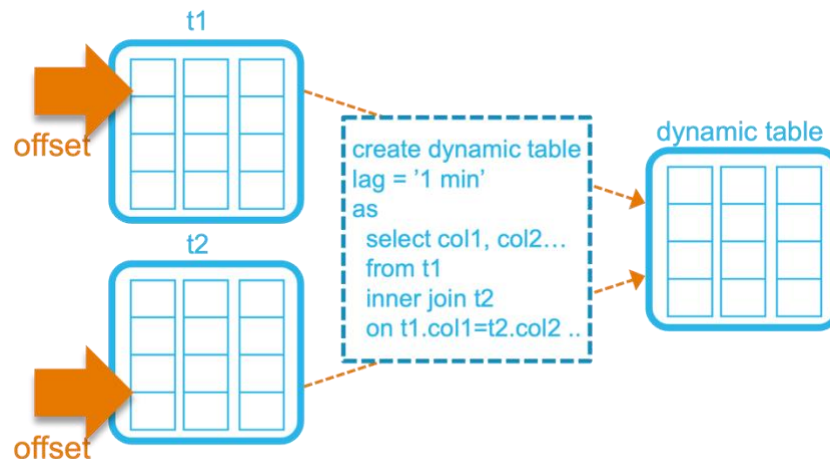

Figure 7 Dynamic Table DAG

NOTE: A dynamic table's SQL definition can include other Dynamic tables!

New storage locations

Not quite a new data type, however important to mention support for those use cases where you simply *cannot* load your data into the cloud. Snowflake announced support for on-premises data access to data that lives on S3-compatible (s3compat) data storage locations. At the time of writing, the S3 compatible technology partners are:

- Pure Storage, distributed file system and object storage technology
- Dell, storage and cloud technology behemoth
- MiniIO, multi-cloud storage technology
- and any technology partner that conforms to AWS's S3 API.

Data applications and Streamlit

Need automation? We have an app for that! (*Powered by Snowflake*)

Snowflake's data cloud is rapidly expanding into an application *ecosystem*. And there are predominantly three Snowflake application frameworks:

- As a **managed app**, your three-tiered (UX/presentation tier + business logic/processing tier + data tier) app architecture is supported by the app builder's own Snowflake account and optionally the cloud service provider (CSP). The app builder ingests the customer's data and pays for storage and compute costs. The app builder's customer primarily interacts with the application tier and may be granted access to data via data sharing, extracts, or APIs; you decide as the app builder, you're managing the app. Data *rests* in the builder's app – *customer trusts app builder with their data*
- As a **connected** app, depending on the app design, the app builder or their customer manages data ingestion, processes data in Snowflake or CSP and both builder and customer pay for storage and compute. The app builder's customer manages their

own raw data in their Snowflake account, and the app builder manages the app but the data *rests* with the customer's environment – *customer data does not leave the customer account.*

- As a **native** app, all three tiers are supported within Snowflake. The app builder is limited to Snowflake accounts only (no CSP). The builder or their customer manages data ingestion, but all processing happens in the customer's account and therefore the customer pays for all Snowflake storage and compute. Data *rests* with the customer's Snowflake account.

Native apps are made available through the Snowflake marketplace

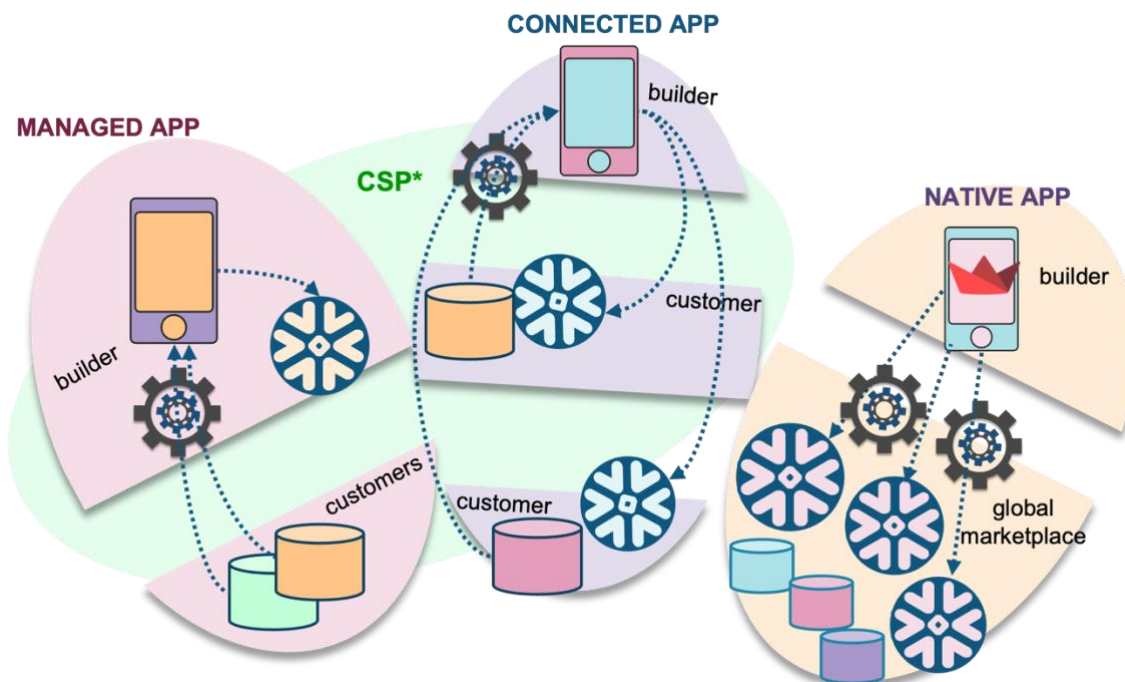
Figure 8 Managed, Connected and Native Apps (not mutually exclusive deployment models)

What is the value of this ecosystem?

Collaboration: supporting new ways to bring your app to your customers, making your apps discoverable in the Snowflake marketplace, no need to move data (where applicable), and use native Snowflake's native sharing and governance capabilities.

To complement and streamline collaboration, Snowflake acquired Streamlit almost a year ago. In addition to the frameworks suggested above, Streamlit offers a way of rapidly developing and sharing your own analytics internally as Streamlit apps. No need to learn frameworks like Flask or Django, simply build and deploy using simple Python syntax which you can now develop within Snowsight itself or consider using Snowflake's new Visual Code Studio extension!
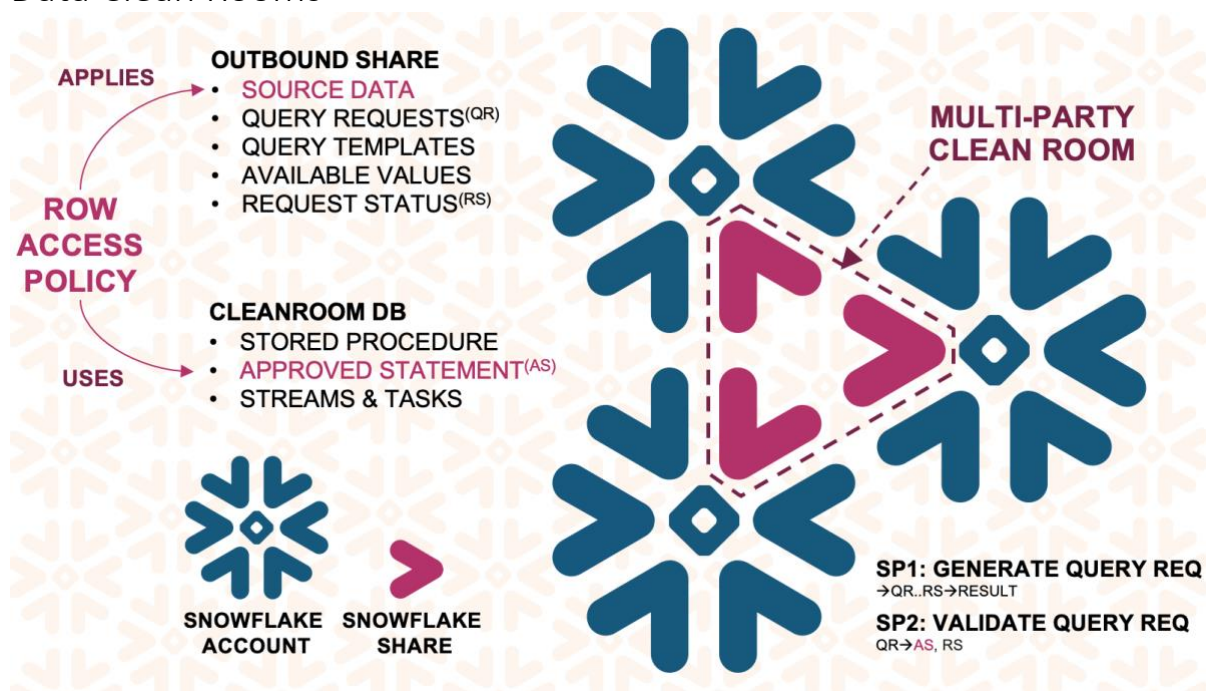
Connecters

Data Clean Rooms

Figure 9 Data Clean Room framework on Snowflake

Data Clean rooms have gone to the next level! Taking advantage of Snowflake's unique Data Sharing architecture, we have seen use cases expand from the classic publisher-advertiser original model to **machine learning** models running on partner data without revealing personally identifiable information and without moving any data! It's not magic, it's Snowflake! *Now imagine a Data Clean Room using Snowflake's Native App Framework…*

Salesforce + Snowflake

Realtime data sharing for Salesforce Customer Data Platform (Genie) using Snowflake technology, you read that right. Snowflake data sharing technology will in partnership with Salesforce allow you to open up Snowflake bound data within your Salesforce Cloud! Think big!

Figure 10 Genie-us… so many puns I could not help myself!

Other notable Snowflake ventures

- Tecton, Domino, Hyperfinity – Feature Store technology for your MLOps workflows
- Dataiku, DataRobot – partnering on our Data Science story
- Panther, Hunters, Securonix, Material – adding to our new Cybersecurity workflow
- dbt – imagine building Dynamic Tables in dbt!!!
- Immuta, Collibra and Alation for managing your Data Governance
- Habu, OpenAP – focussing on our Data Clean Room story
- DataOps, Hex, Overlay – doing more with your data pipelines!
- Sigma, Thoughspot, Robling – for your BI needs!
- And more from service and technology partner network!
- As well as Snowflake partner connect setup!

Industry focus

Technology is merely an enabler for a corporation's value streams, Snowflake has the technology down and as a technology provider have added value engineering into their portfolio of services with solutions and guidance on:

- Financial Services (inc. Insurance) Data Cloud – *ex. Next Best Action, ESG Portfolio Construction, Quant Research, Fraud Detection, Claims Management, Blockchain Analysis, and more…*
- Advertising, Media & Entertainment Data Cloud – *ex. Accelerating advertising revenue, Subscriber acquisition and retention, Telemetry, Recommendation engine, and more…*
- Healthcare & Life Sciences Data Cloud – *ex. Patient 360, Operation efficiency, Pandemic / Crisis Response Data Hub, Patient Journey 350 Analytics, Unstructured Data Analytics, and more…*
- Retail & Commercial Package Goods (CPG) Data Cloud – *ex. Customer Satisfaction (NPS score), Product Recommendation Engine, Demand Forecasting, Forecast Returns, On Shelf Availability, and more…*
- Public Sector Data Cloud – i.e. Government & Education

- [Manufacturing](#) Data Cloud – *ex. Inventory Optimization, Cycle Time analytics, Multi-Enterprise Demand Forecasting, Factory Predictive Maintenance, Supply Chain Risk, Returns Logistics, and more…*
- [Telecommunications](#) Data Cloud – *ex. Churn, Customer Pathing, Roaming, Segmentation, Social Network Analysis and Recency Frequency Monetary modelling, Householding, and more…*

We could never list every use case here, but you get the idea! Snowflake is more than just a Data Cloud; it is a Business **Application Ecosystem**!

Where to go from here?

You can play before you pay on Snowflake,

- Sign up for a trial account with $400 worth of credits from the get-go!
- You can get set up through the partner portal on their tools using Snowflake to ingest, transform and consume data.
- You can also visit our **[QuickStarts](#)** online documentation with step-by-step instructions on how to set up the above features.
- You could also get in touch with our Powered by Snowflake enablement teams to help you design your apps to run on the data cloud.

Aside from the product, Snowflake also employs skilled **Solution Architects** with the accumulative delivery experience of over a hundred years (we've been hiring a lot of talent!). Well versed in analytical competency and putting customers first through our corporate values to [think big](#); our globally accessible **Professional Services** (PS) Solution Architects help customers deliver best practices and patterns that work in the cloud.

And we have options!

- Hire us for a week to get you going on your business case (PS [QuickStart](#))
- Hire us for a week to assess your existing Snowflake deployment ([Snowflake 360](#) & [Best Practices](#) & [Security Consultation](#)).
- Make us a part of the furniture by having one of us on site for 10/20/40 hours a week from six months to a year, you decide! (**Resident Solution Architect**).

You could also attend some of our many marketing led Snowflake+Customer initiatives (ex. [Snowflake Build](#)) to bring together other customers, prospects, partners and our professional solution architects and sales engineers to help you make up your mind!

Attend a [Data for Breakfast](#) at a major city near you.

PS also offer [Migration Services](#) to streamline your move to the data cloud off legacy data platforms!

See you at Summit! https://www.snowflake.com/summit/

By the way, we turn 11 this year – look at all that has been achieved in these short years, imagine what even the next year will bring to your business use cases!

The views expressed in this article are that of my own, you should test implementation performance before committing to this implementation. The author provides no guarantees in this regard.