



Session 1- Day1- Snowpro Core certification Snowflake Deep Dive

Santosh Ubale & Robby Thomas GSI Sales Engineer 26 July 2021

AGENDA

Day1

- SnowPro Core Certification exam overview
- Snowflake Data Cloud
- Architecture Overview
- Zero-copy Cloning
- Time Travel
- SnowPro Core Certification exam Sample Questions
- Snowflake hands lab on over view
- Q&A

➤ SnowPro Core Certification exam overview



REQUIRED- SNOWPRO CORE CERTIFICATION

SnowPro Certification is the objective measure of Snowflake expertise. It is an industry certification, administered via a 3rd party. Different than an accreditation.

- [SnowPro landing page](#)
- [About SnowPro \(Partner Academy\)](#)
- Review the [Study Guide](#) to prep

[Course Overview:](#) SnowPro Core Certification

- Administered online by a 3rd party (*Webassessor*)
- 100 Multiple Choice Questions
- 2 hours
- \$175/person/exam
- [SnowPro FAQ](#)

The **4-Day Fundamentals** training is the primary course designed to set you up to pass the exam. It is not required but recommended.

Refer to [Snowflake Training Class Schedule](#) for information on registering for this course.



Services Partner Requirements

	Requirement	Registered	Select	Premier	Elite
Financial	Annual Program Fee		\$2,500	\$5,000	\$7,500
Training & Certification	Snowflake Sales Pro Accreditation: For Sales account execs	2	4	8	
	Snowflake Tech Sales Pro Accreditation: For Sales Engineers	2	4	8	
	Certification: SnowPro Core	2	4	6	
	Certification: SnowPro Advanced	1	2	2	



SNOWPRO CORE CERTIFICATION



Target Audience

Data Analyst

Data Engineers

Data Scientist

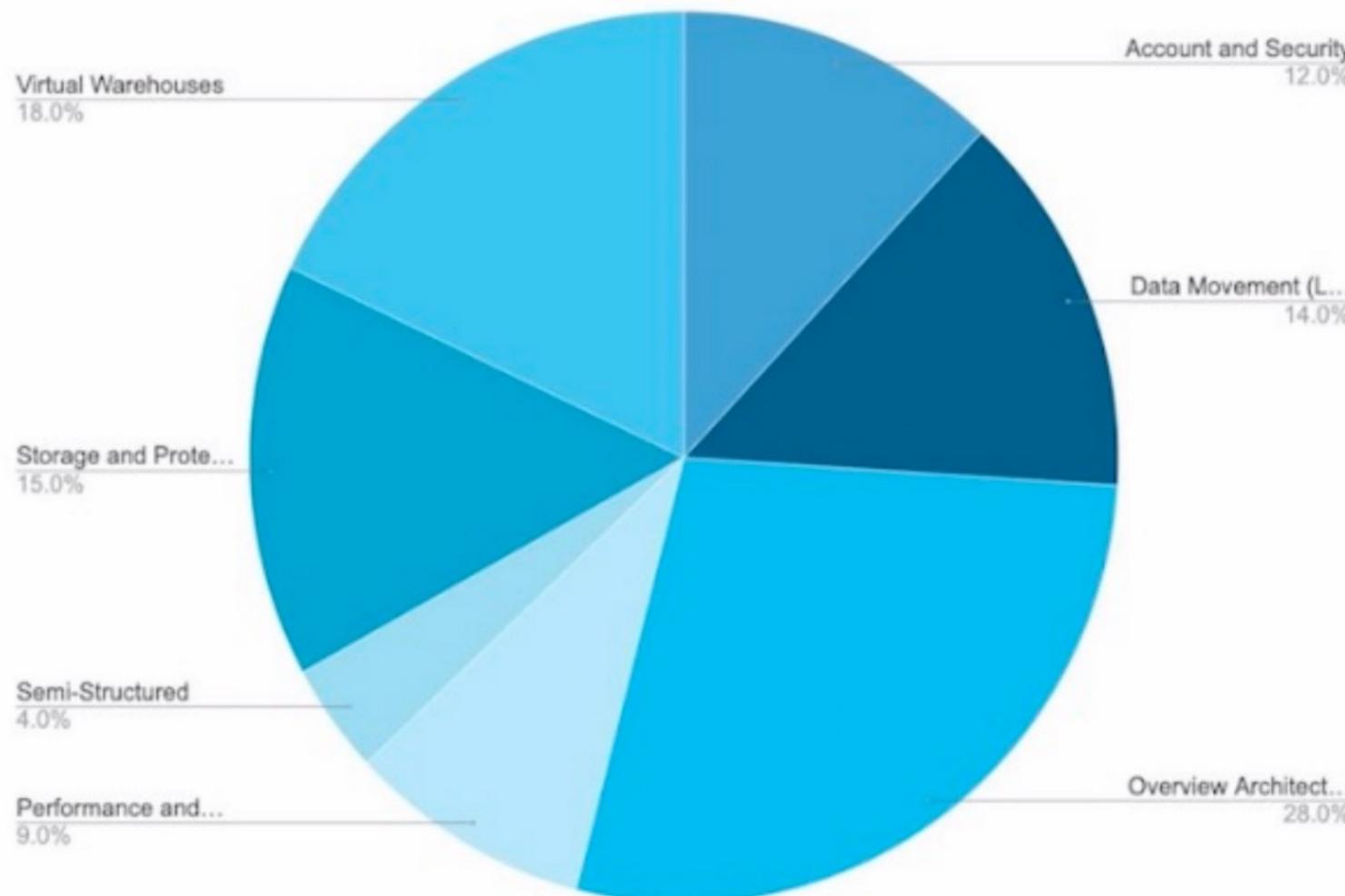
Database Architect

Database Administ

Solution Architects



SNOWPRO CORE CERTIFICATION



➤ Introduction To Snowflake & Data Cloud



OUR MISSION & STORY

Enable every organisation to be data-driven



Founded in 2012
by Oracle industry
veterans



\$1.4B funding from
leading investors on a
\$12.4B valuation



4,000+ active customers, “The
Cloud Data Platform”
launched and IPO



2015 general availability of
“The Data Warehouse Built for
the Cloud”



Gartner Data Management
Solutions for Analytics
leader



500m jobs per day



250+ PB total storage,
biggest table 68TN rows



1.2k data providers



NPS - 71, industry average

21

DATA CLOUD
Content Vector & Network Effects

CLOUD DATA PLATFORM
Workload & User Expansion

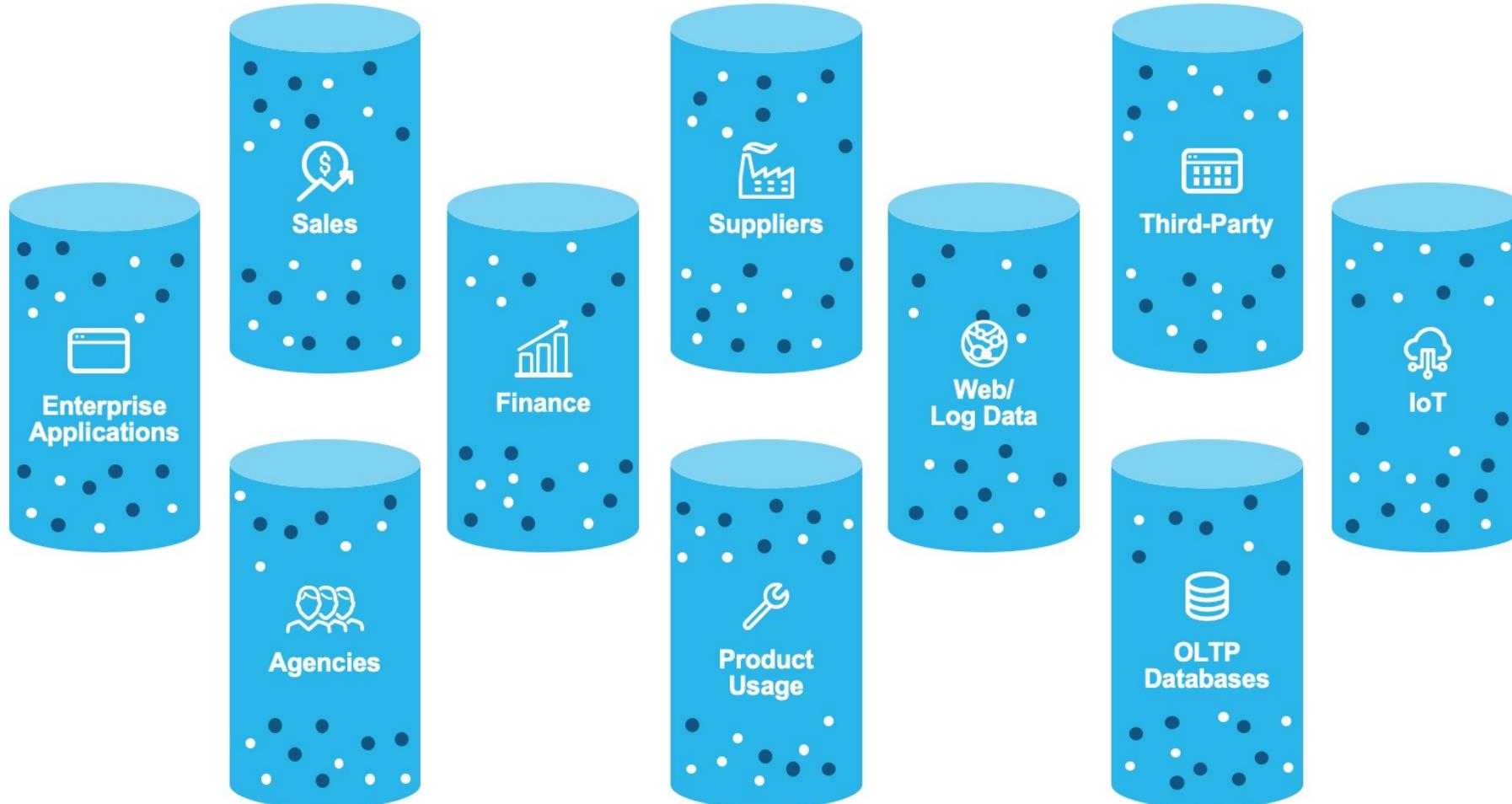
CLOUD DATA WAREHOUSE
Superior Performance

**CLOUD NATIVE
ARCHITECTURE**

2014 . . . 2019 . . . 2020

RISE OF THE DATA CLOUD

DATA SILOS PREVENT VALUE REALIZATION

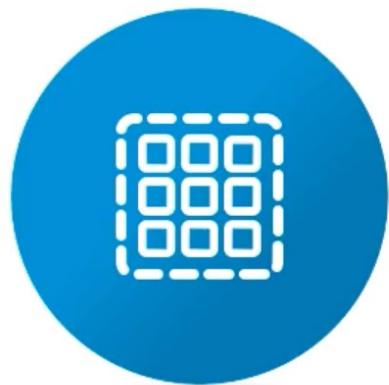


87%

Decision makers want to
expand their ability
to use external data¹

¹ "The Insights Professional's Guide To External Data Sourcing" Forrester, 2020

REQUIREMENTS OF A CLOUD DATA PLATFORM



One Platform
One Copy of Data,
Many Workloads



Secure &
Governed Access
to All Data

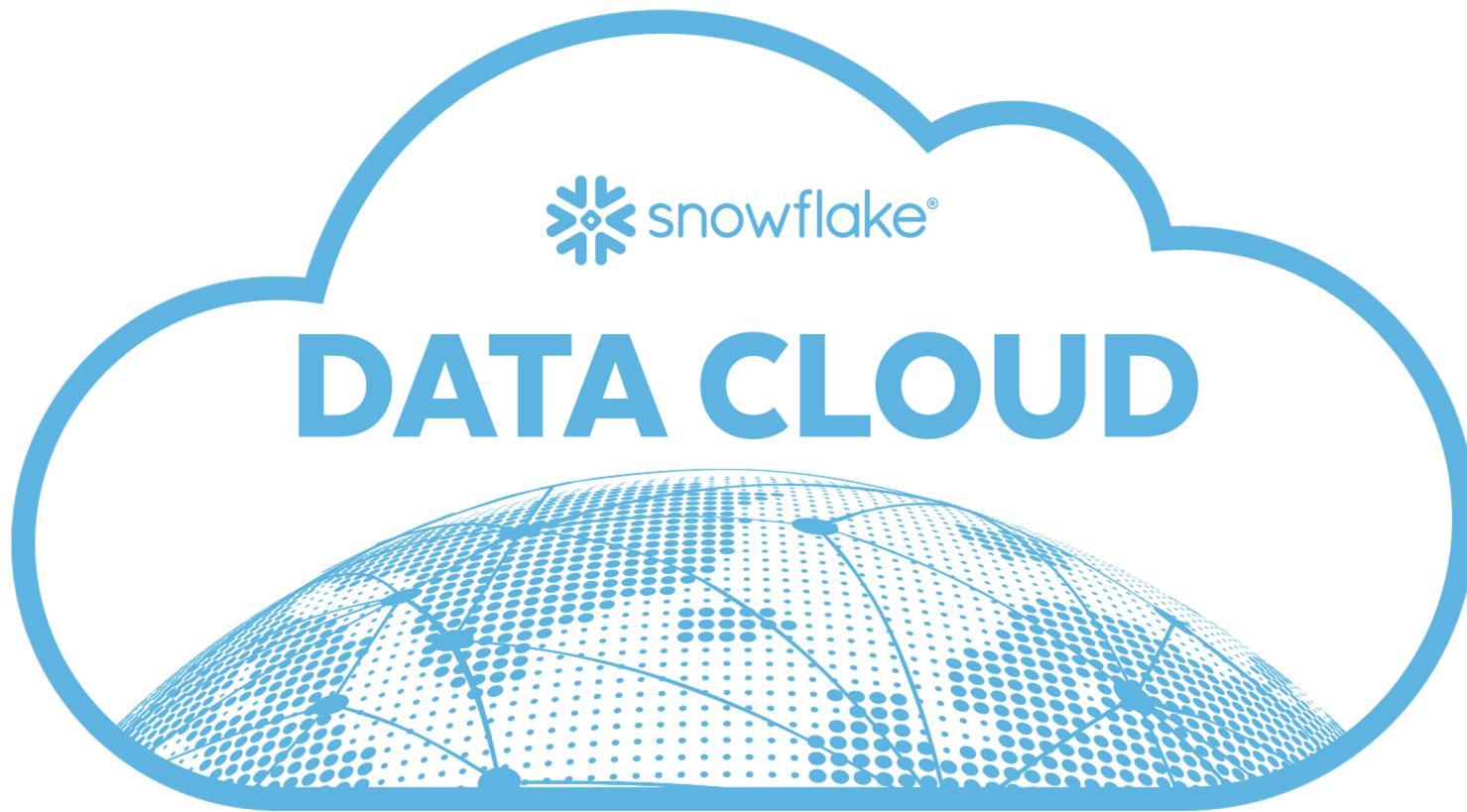


Near-zero
Maintenance,
as a Service



Unlimited
Performance
and Scale

WHAT IS THE DATA CLOUD?



The Data Cloud = Platform plus Data

THE SOLUTION:

The Data Cloud is a global network where thousands of organizations mobilize data with near-unlimited scale, concurrency, and performance.

Snowflake's platform is the engine that powers and provides access to the Data Cloud.

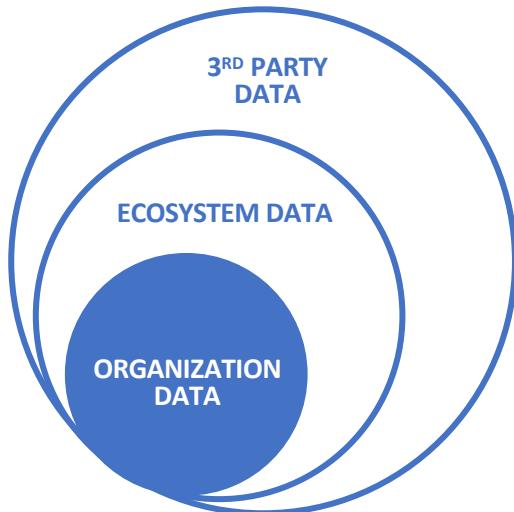
CORE BENEFITS:

Access | Governance | Action



BENEFITS OF THE DATA CLOUD

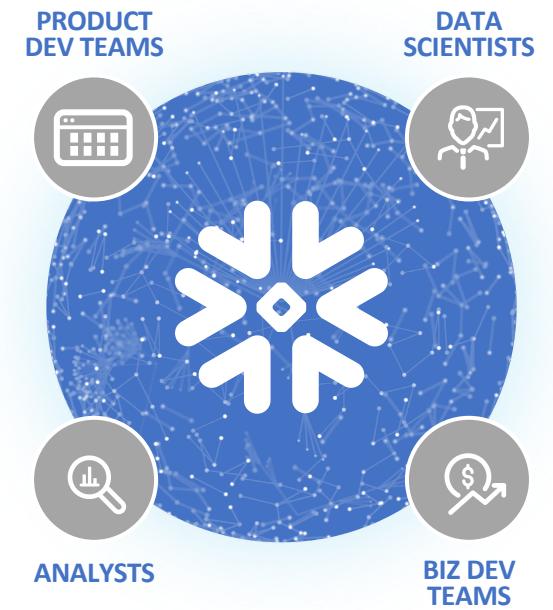
ACCESS



GOVERNANCE



ACTION



ACCESS



- ✓ Any Format
- ✓ Near-Unlimited Scale
- ✓ Sharing Without Copying or Moving

All of Your Organization's Data, on One Platform

Your Ecosystem - Partners, Suppliers, Customers

Snowflake Data Marketplace - Industry Datasets, Data Services, Applications

GOVERNANCE

Know Your Data

Understand and classify data across your entire ecosystem



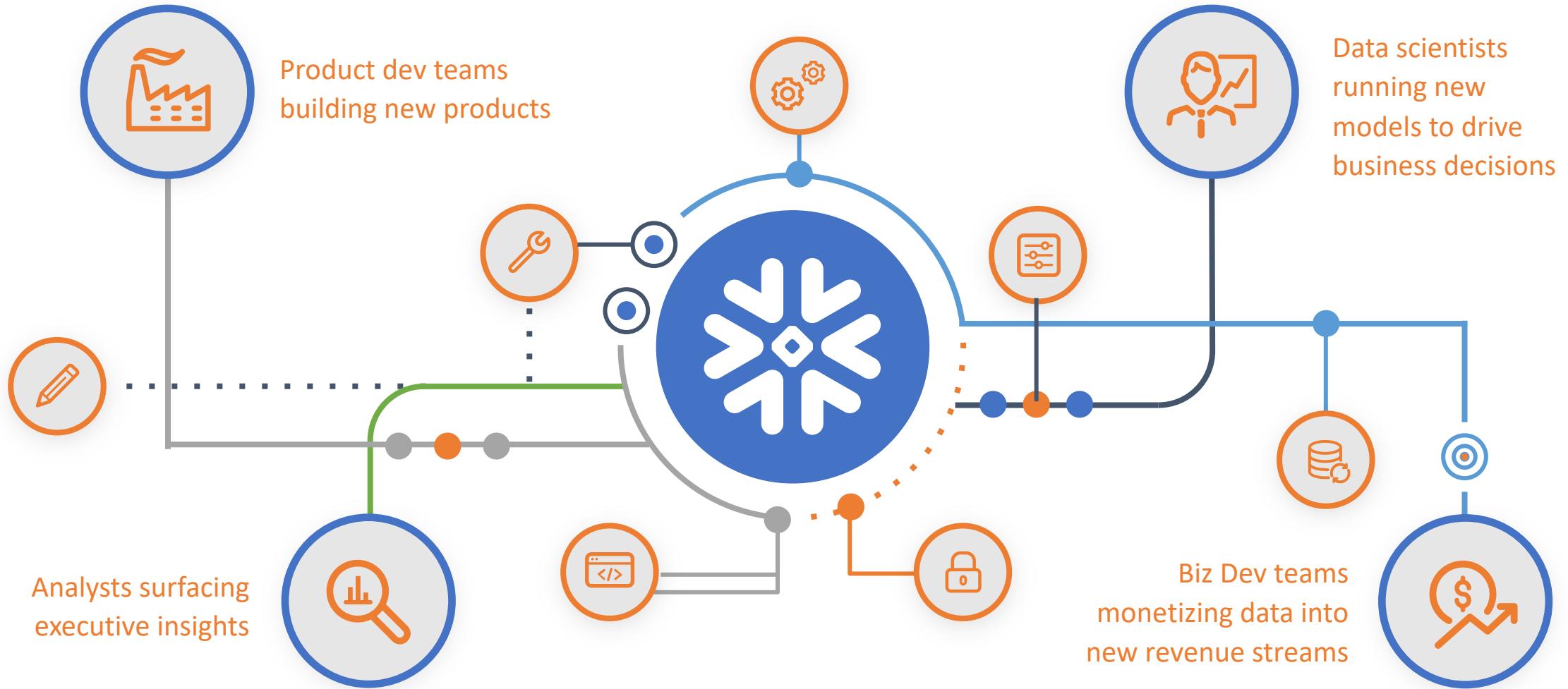
Unify Security & Governance

Simplify governance across workloads with centralized controls

Control your Data

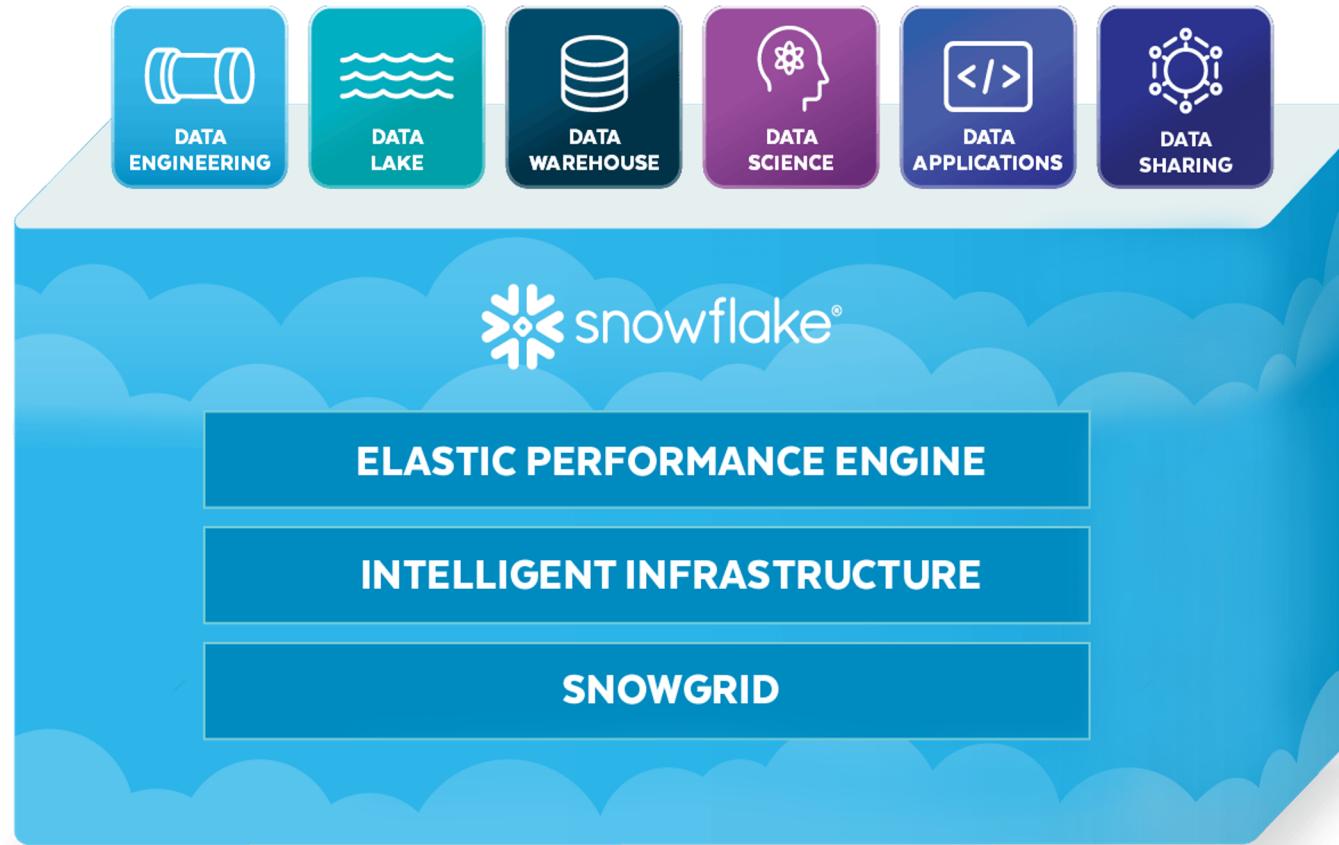
Implement flexible governance and security policies, that don't hinder innovation

ACTION



SNOWFLAKE PLATFORM

Under the hood



Google Cloud



aws

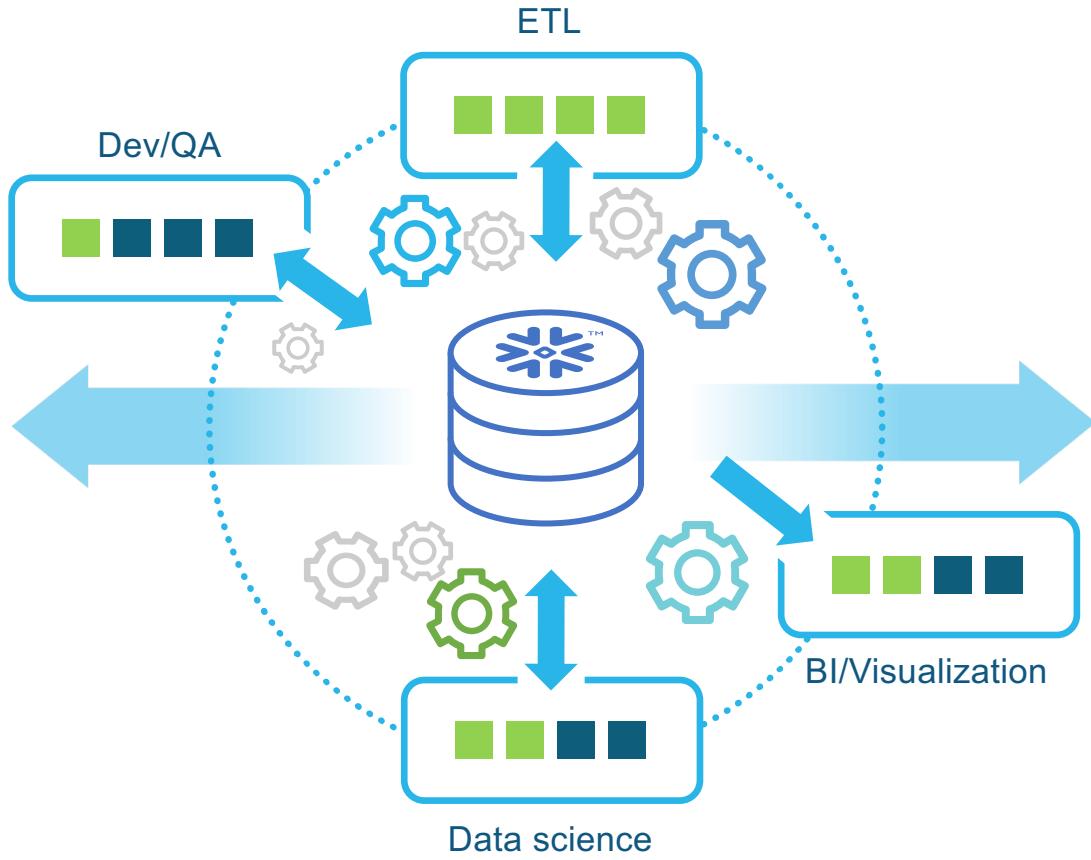


Azure



© 2020 Snowflake Inc. All Rights Reserved

ELASTIC PERFORMANCE ENGINE



One engine for every workload

Simplify your architecture. Power complex pipelines, analytics, data science, interactive applications, and more.

Leading performance and concurrency

Fast, reliable performance every time with no tuning or contention. Instantly and cost-efficiently scale to any amount of users, jobs, or data.

Support any user or skillset

Get the accessibility of SQL, with the flexibility to support Java, Scala, and more. Run external tools directly for extended capabilities.



INTELLIGENT INFRASTRUCTURE



Snowflake Managed

MAINTENANCE & TUNING

MULTI-CLUSTER COMPUTE RESOURCES

ADMINISTRATION

NETWORKING & ENCRYPTION

DATA MANAGEMENT

CENTRALIZED STORAGE

Automated and fully managed for you

Focus on what matters. Fully managed with automations that encrypt data, control access, and eliminate manual maintenance and troubleshooting.

High availability, high reliability

Automate complex replication and failover cross-clouds and cross-regions. Stay up-and-running no matter what happens.

Optimized costs for all data

Usage-based model paired with patented compression and fine-grained controls to right-size costs. Continual improvements for new efficiencies.

SNOWGRID



Snowflake Regions



AWS



Azure



GCP

Maintain global business continuity

Eliminate disruptions, deliver better experiences, and comply with changing regulations through unique cross-cloud, cross-region connectivity.

Share data with no ETL or silos

Remove the barriers to data, regardless of cloud, region, workload, or organizational domains. Get instant access and distribution through a single copy of data.

Cross-cloud governance controls

Simplify governance at scale with flexible policies that follow the data for consistent enforcement across users and workloads.

Tap into the extended ecosystem

Enrich insights with a network of third-party data. Discover and run new functions for extended workflows.



© 2020 Snowflake Inc. All Rights Reserved

SNOWGRID UNLOCKS DATA SHARING

Traditional Methods

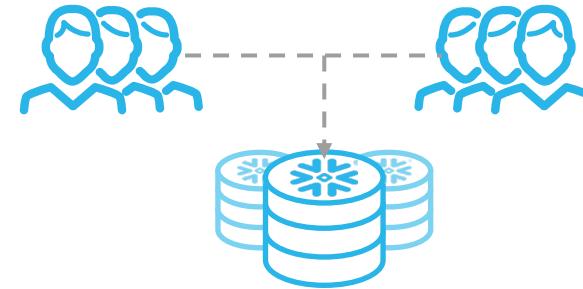
FTP | APIs | ETL | Cloud buckets



- ✖ Copy and move data
- ✖ Data is delayed (Not real time)
- ✖ Costly to manage and maintain
- ✖ Unsecure, once data is moved
- ✖ Error prone; pipelines break

Snowflake

Secure Data Sharing



- ✓ Single copy of live data, no delays
- ✓ No costs of moving, copying, ingestion
- ✓ No more data lake silos
- ✓ Privacy compliant
- ✓ Governed, revocable access

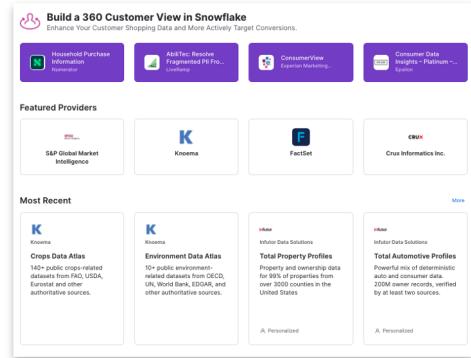
SHARE AND COLLABORATE IN THE DATA CLOUD

DISCOVER AND BE DISCOVERED IN THE DATA CLOUD

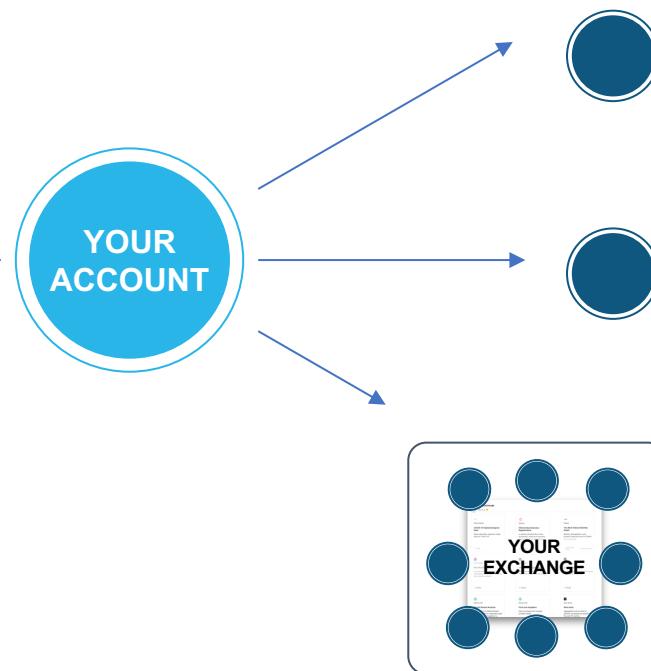
Access data and services from 150+ providers

SNOWFLAKE DATA MARKETPLACE

Market and deliver your products to customers



SHARE ACROSS YOUR BUSINESS ECOSYSTEM



DIRECT SHARE

Share with other Snowflake customers

READER ACCOUNTS

Share with companies not yet on Snowflake

DATA EXCHANGE

Administer group sharing and data discovery across business units



CONNECT TO THE MOST RELEVANT CONTENT



SNOWFLAKE DATA MARKETPLACE

Discover and be discovered with data and services from 150+ providers across 16+ categories.

SNOWFLAKE CUSTOMERS

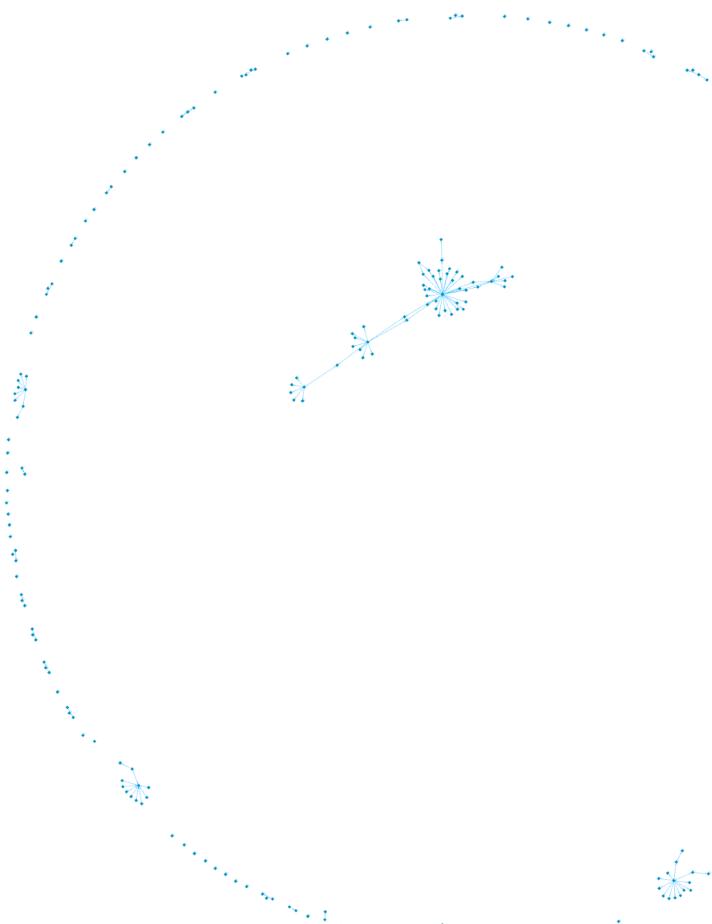
Thousands of companies share data with suppliers, partners, or other business units.

POWERED BY SNOWFLAKE APPLICATIONS

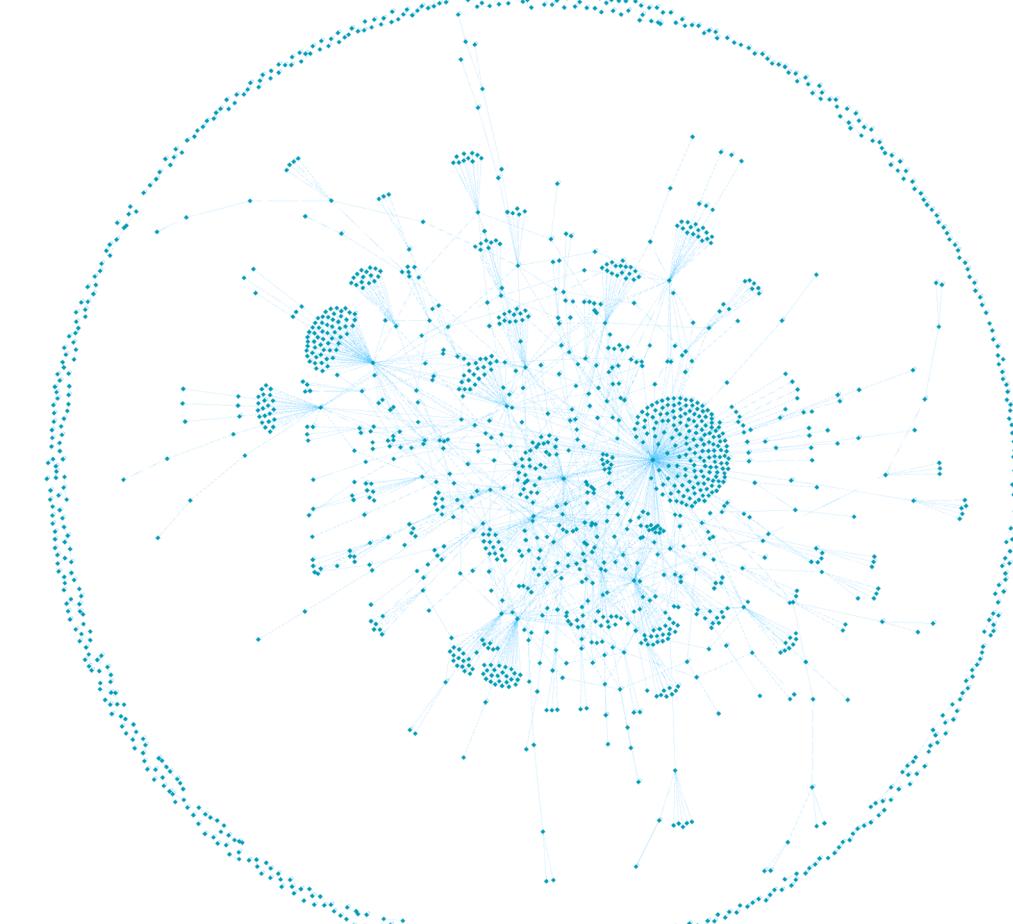
Hundreds of applications that businesses rely on run in the Data Cloud.



DATA CLOUD GROWTH



April 2020



April 2021

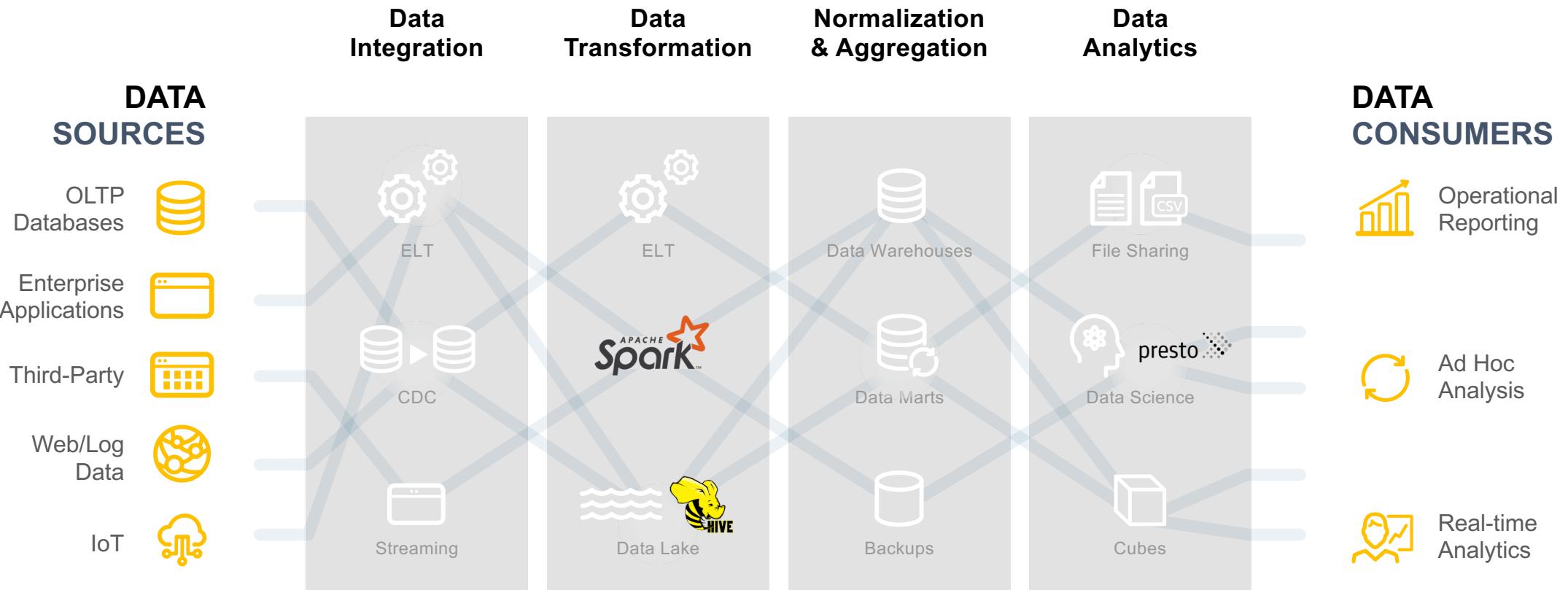


© 2020 Snowflake Inc. All Rights Reserved

*Visualizations based on actual Data Cloud sharing activity as of April 30, 2020 and April 30, 2021

TRADITIONAL DATA ARCHITECTURE

Complex, Costly, and Constrained



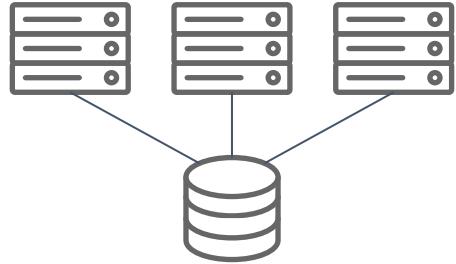


➤ ARCHITECTURE REVIEW



NEXT EVOLUTION OF DATA WAREHOUSE ARCHITECTURE COMPUTE & STORAGE SEPARATED

Traditional architectures

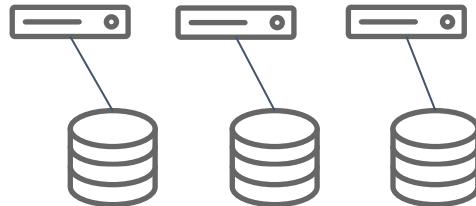


“Shared-disk”

Shared storage

Single Server

SMP



“Shared-nothing”

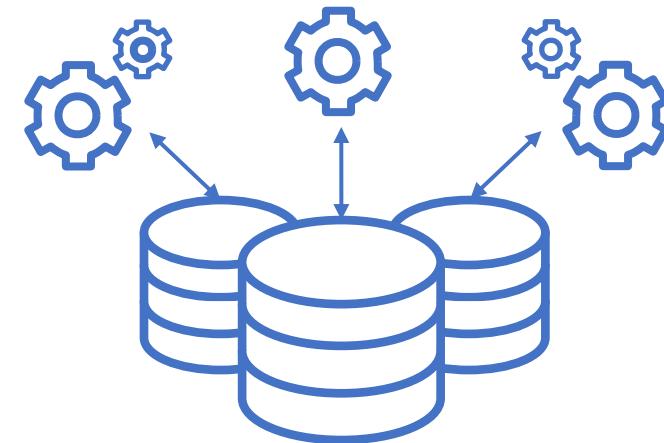
Decentralized, local storage

Single cluster

MPP APPLIANCE

Teradata, Netezza, Vertica,
GreenPlum, Redshift (ParAccel)

Data Cloud



MCSD

“Multi-cluster, shared data”

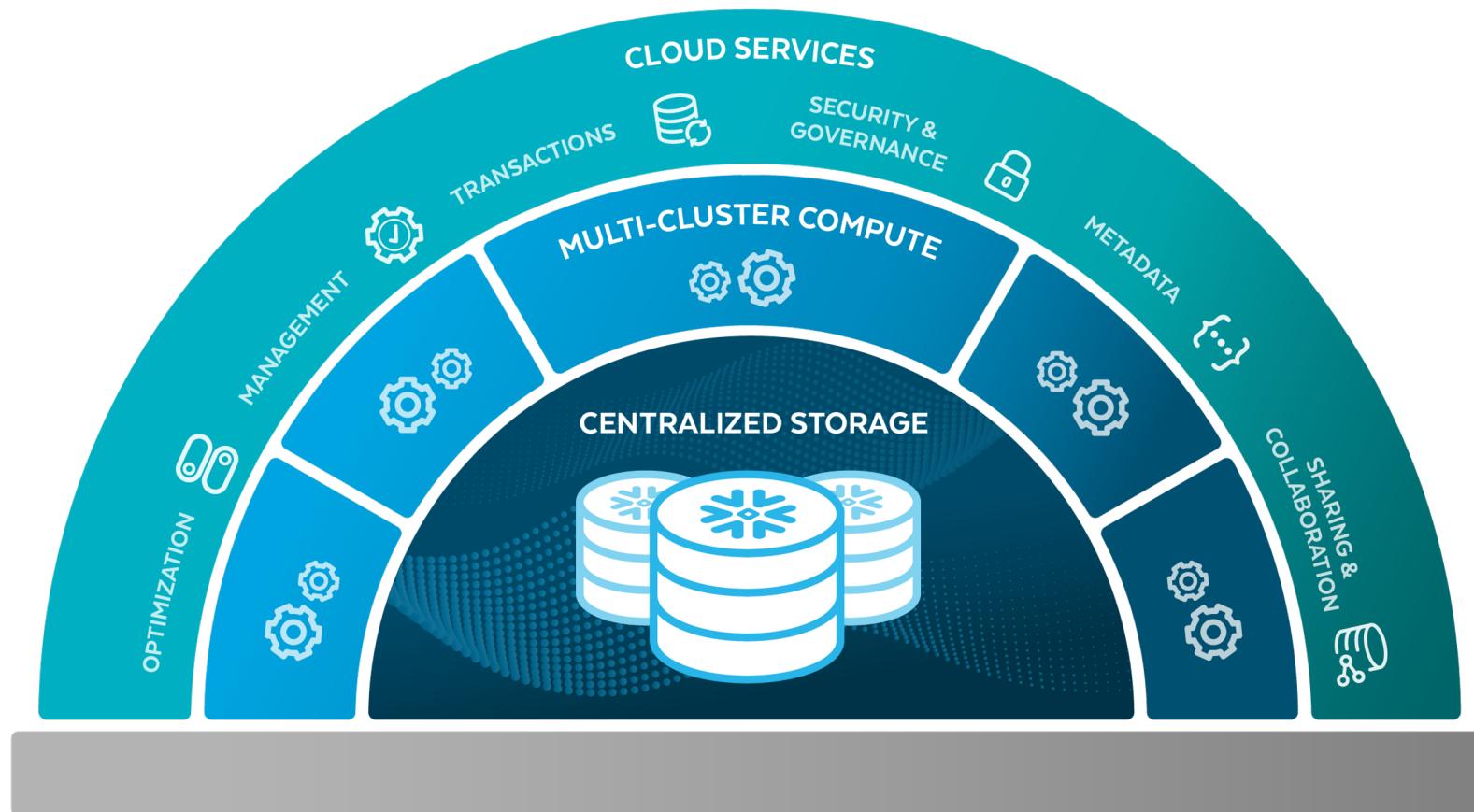
Centralized, elastic storage

Multiple MPP, independent compute clusters

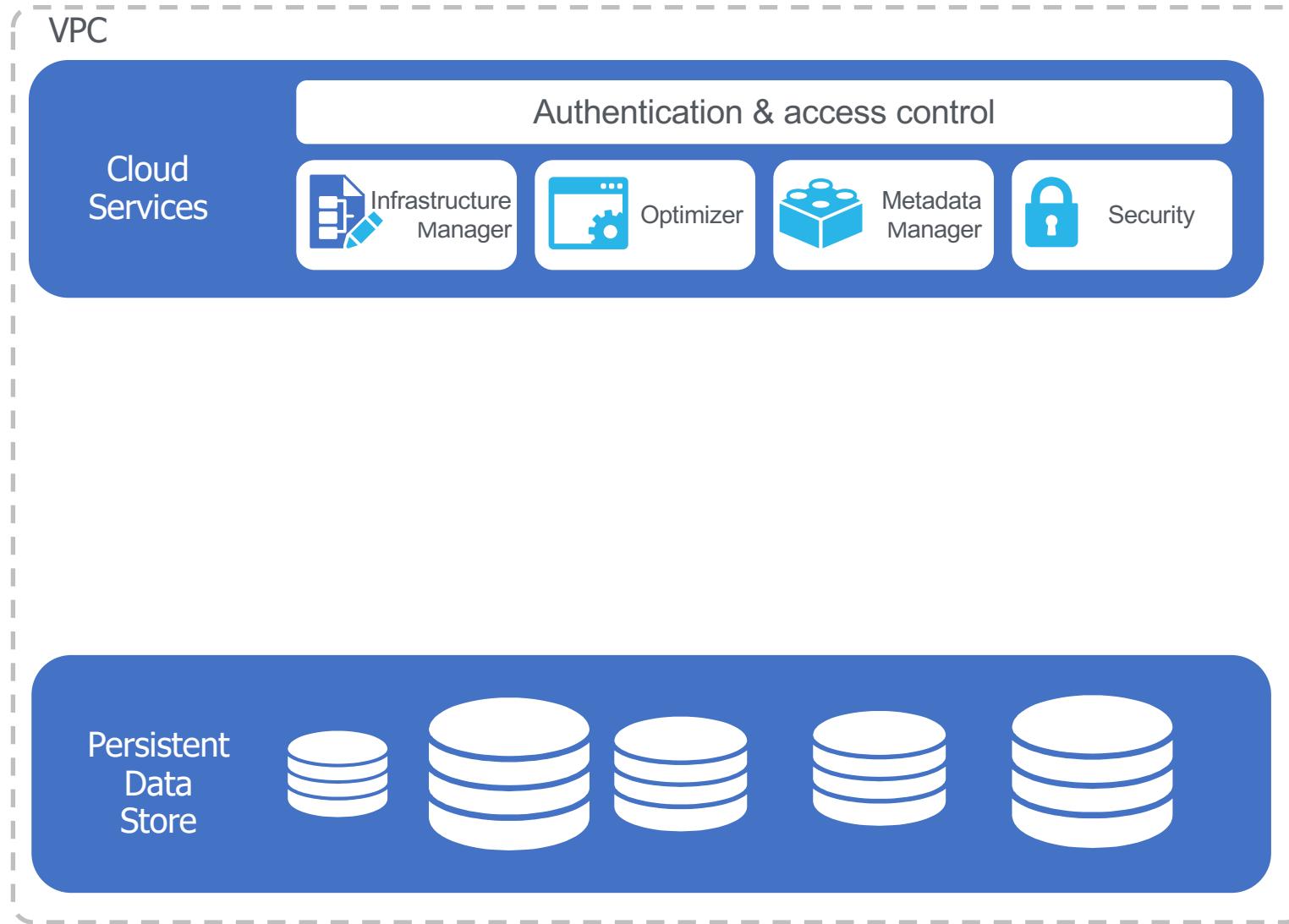
UNLIMITED Storage and Compute



SNOWFLAKE ARCHITECTURE



SNOWFLAKE ARCHITECTURE



Cloud Services + Serverless

Compute not running on a Virtual Warehouse

- SHOW commands
- Query result cache

Serverless

- Search optimization service
- Snowpipe
- Auto-clustering
- Materialized view maintenance
- Cross-cloud replication



SNOWFLAKE ARCHITECTURE

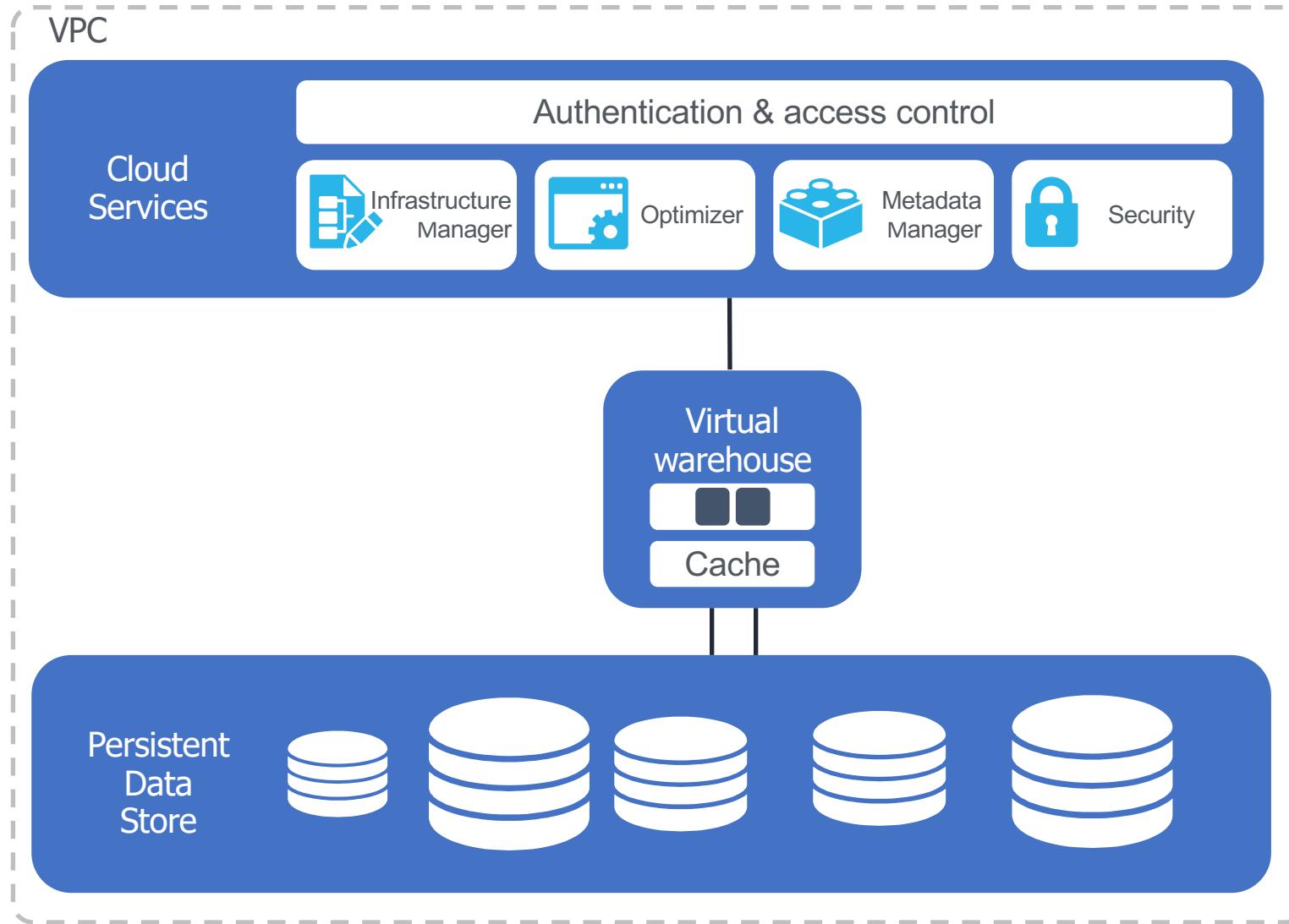


Centralized Storage

- Pass through pricing
- Up to 5x compression
- Native support or semi-structured data
- Zero-copy cloning
- Time Travel



SNOWFLAKE ARCHITECTURE

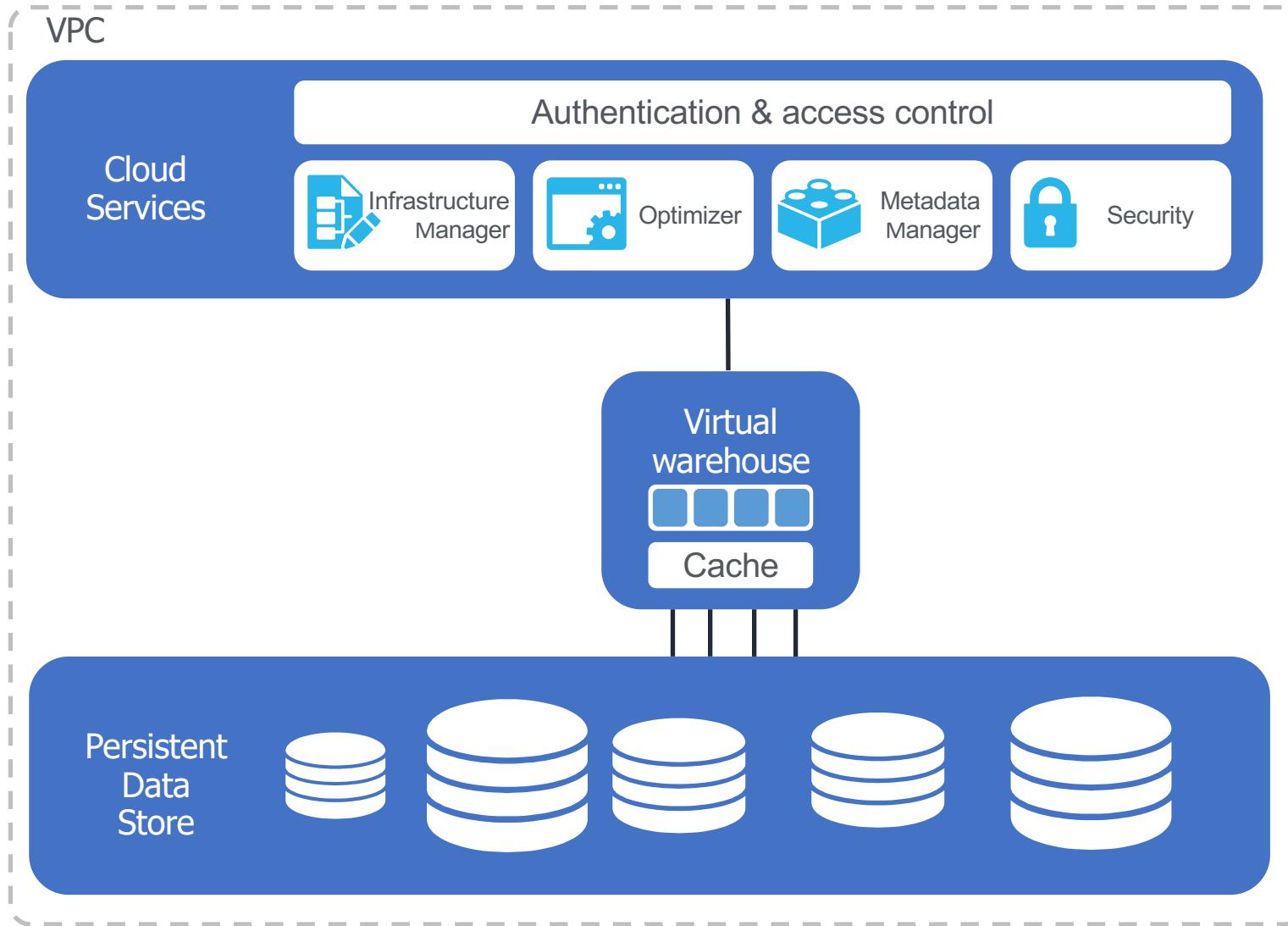


Independently Scalable Compute

- Elastic (turn on instantly)



SNOWFLAKE ARCHITECTURE

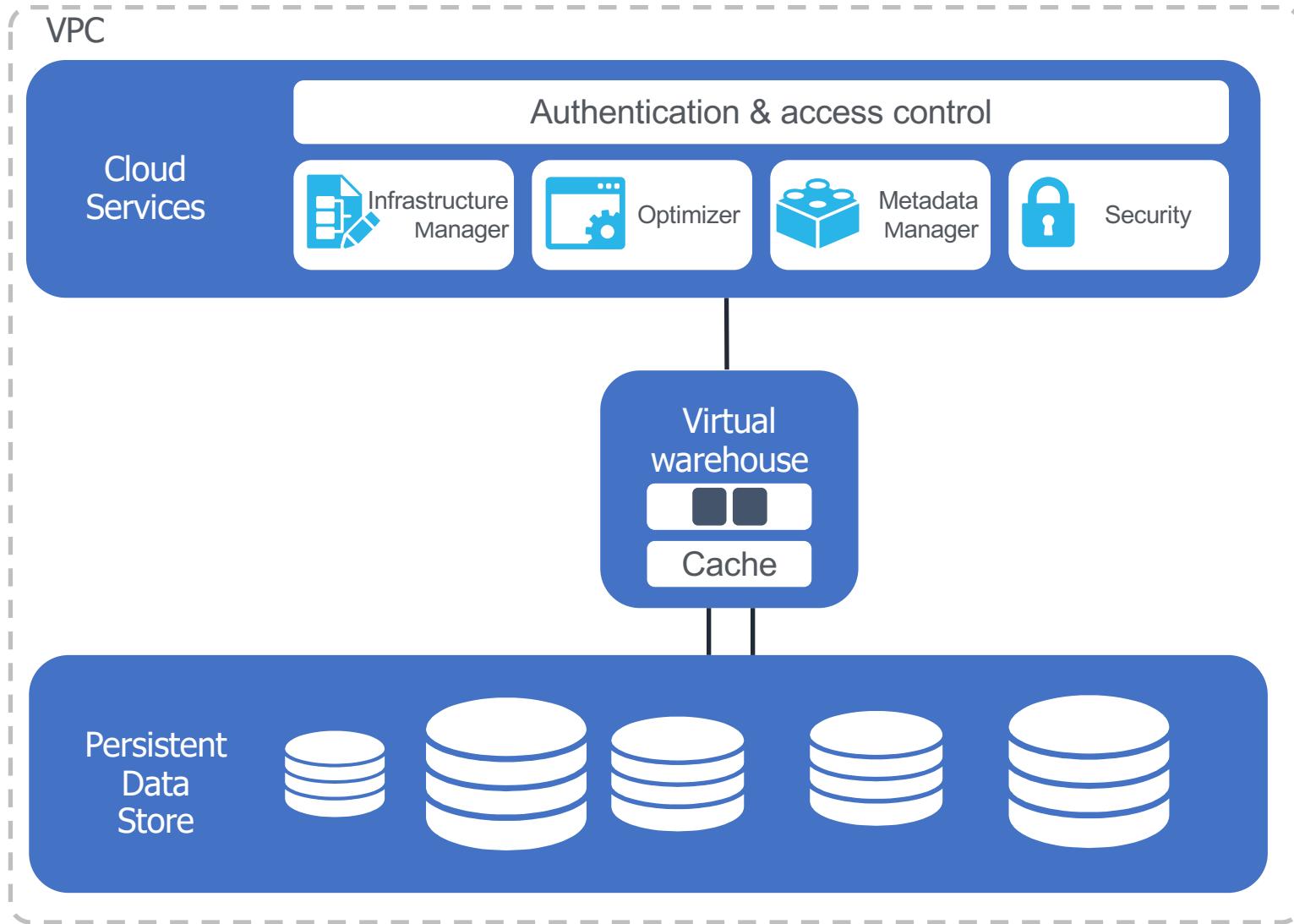


Independently Scalable Compute

- Elastic (turn on instantly)
- Scale up (complex queries)



SNOWFLAKE ARCHITECTURE

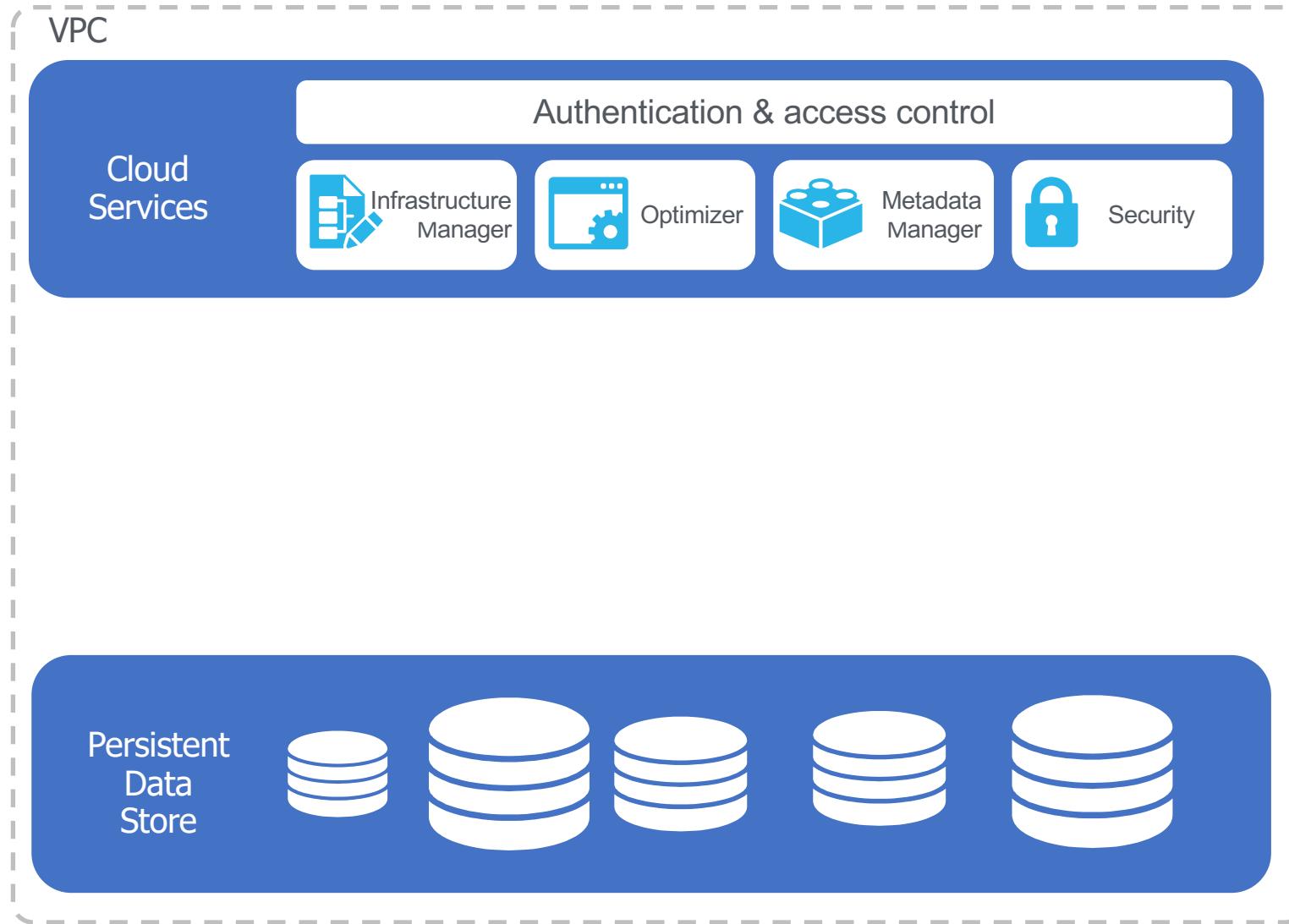


Independently Scalable Compute

- Elastic (turn on instantly)
- Scale up (complex queries)



SNOWFLAKE ARCHITECTURE

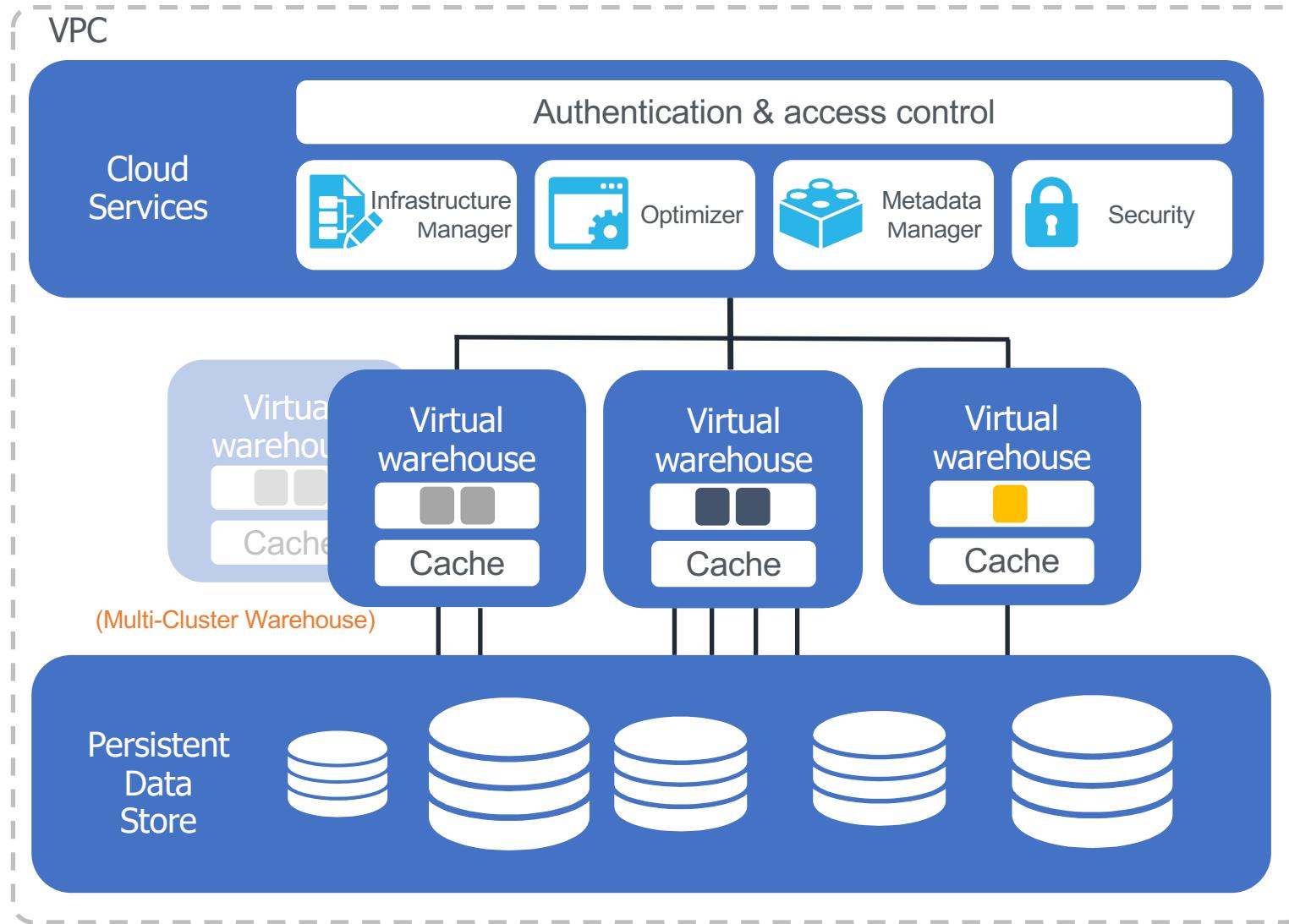


Independently Scalable Compute

- Elastic (turn on instantly)
- Scale up (complex queries)
- Suspend



SNOWFLAKE ARCHITECTURE

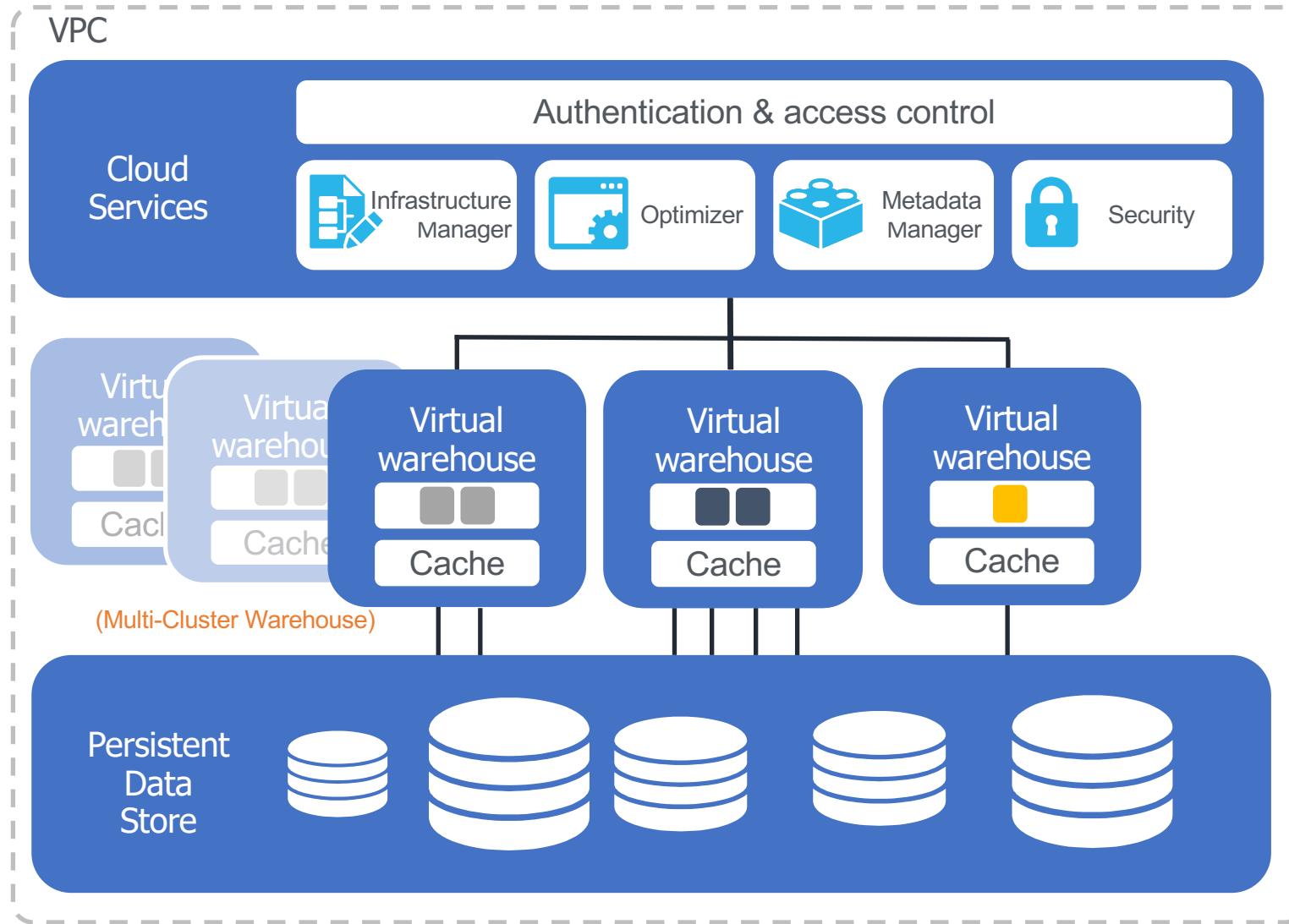


Independently Scalable Compute

- Elastic (turn on instantly)
- Scale up (complex queries)
- Suspend
- Scale out (concurrency)



SNOWFLAKE ARCHITECTURE

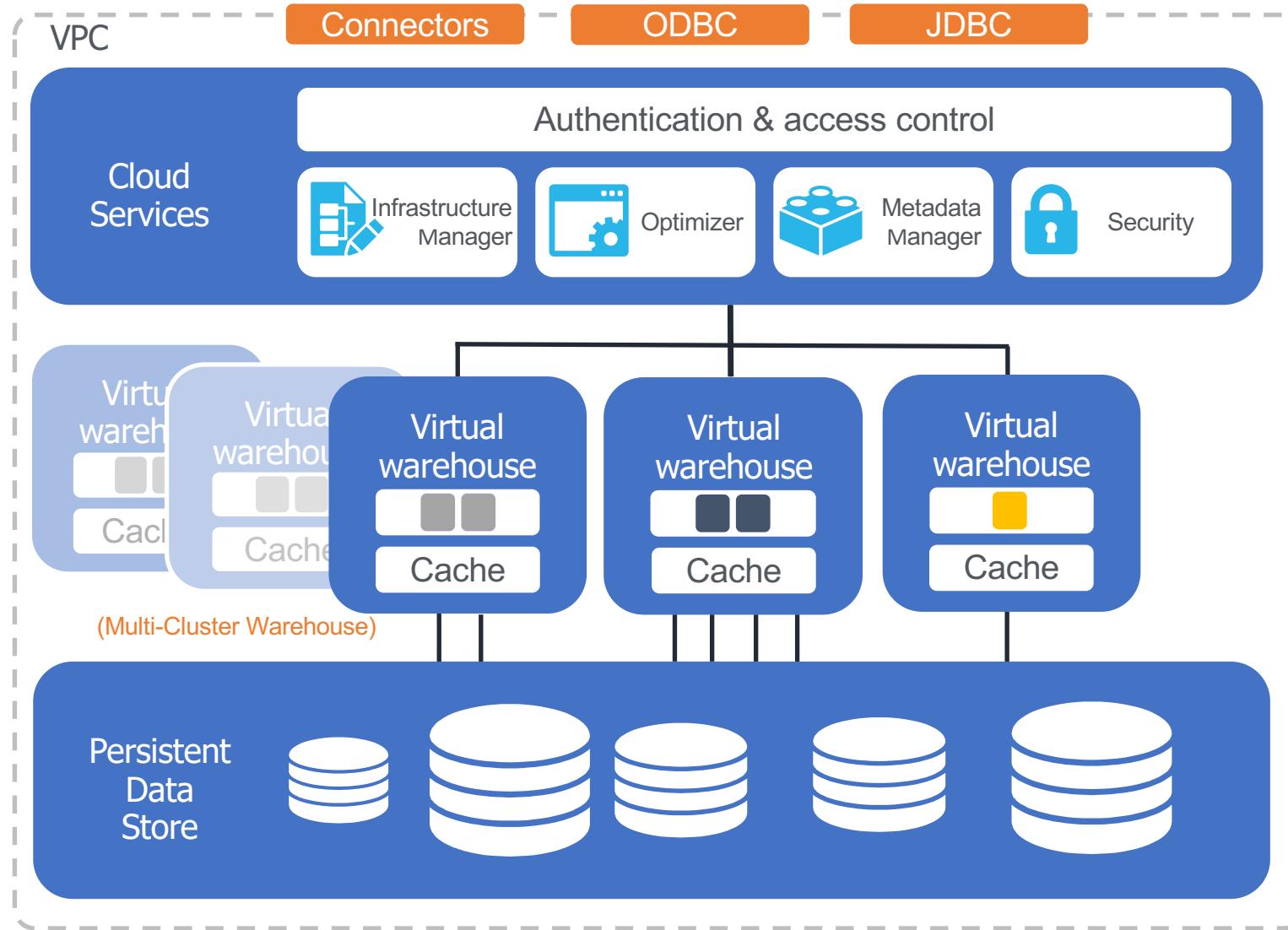


Independently Scalable Compute

- Elastic (turn on instantly)
- Scale up (complex queries)
- Suspend
- Scale out (concurrency)



SNOWFLAKE ARCHITECTURE



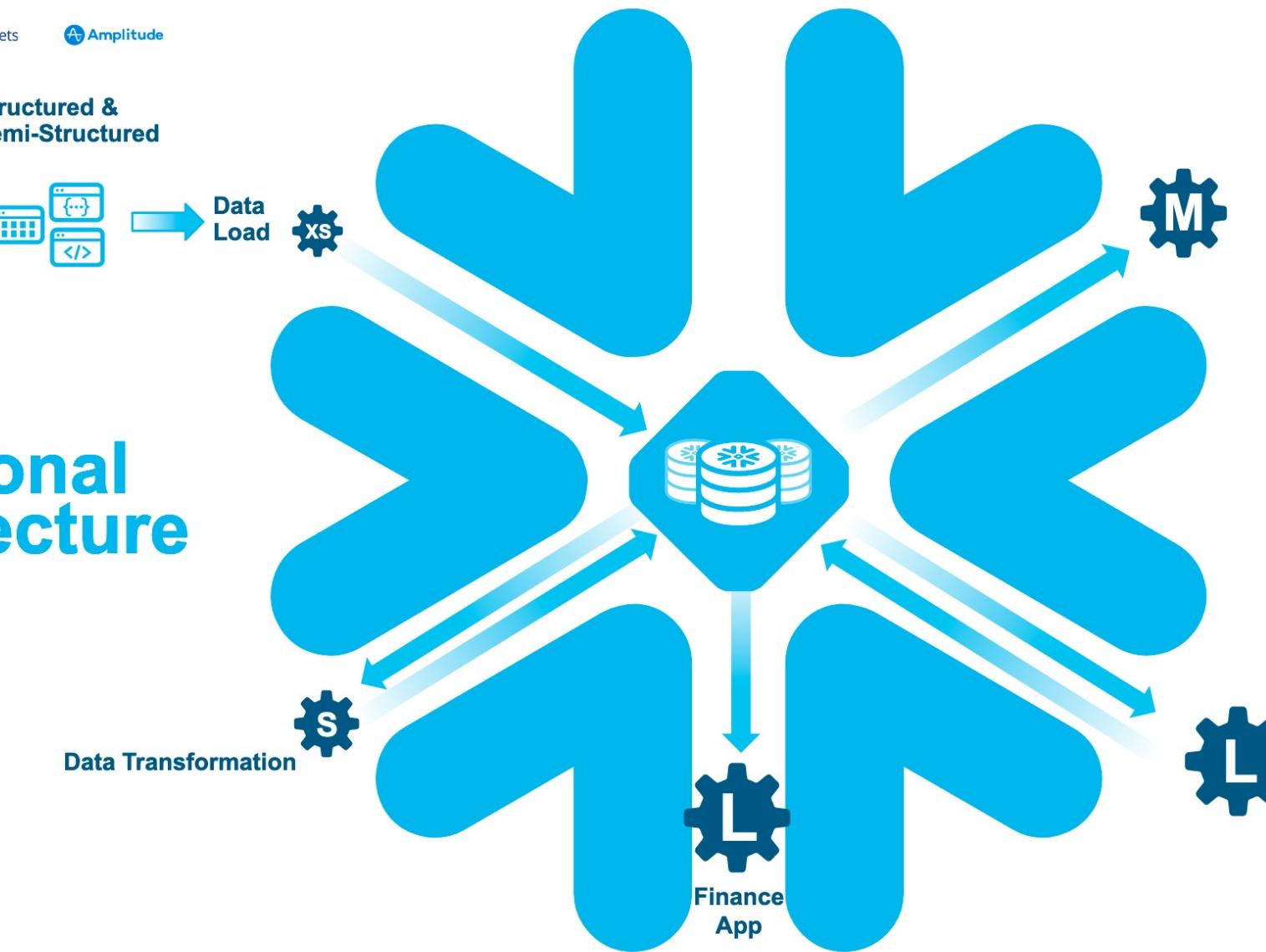
Independently Scalable Compute

- Elastic (turn on instantly)
- Scale up (complex queries)
- Suspend
- Scale out (concurrency)
- Per second billing



PAY FOR WHAT YOU ACTUALLY USE... DOWN TO THE SECOND

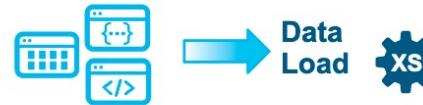
Functional Architecture



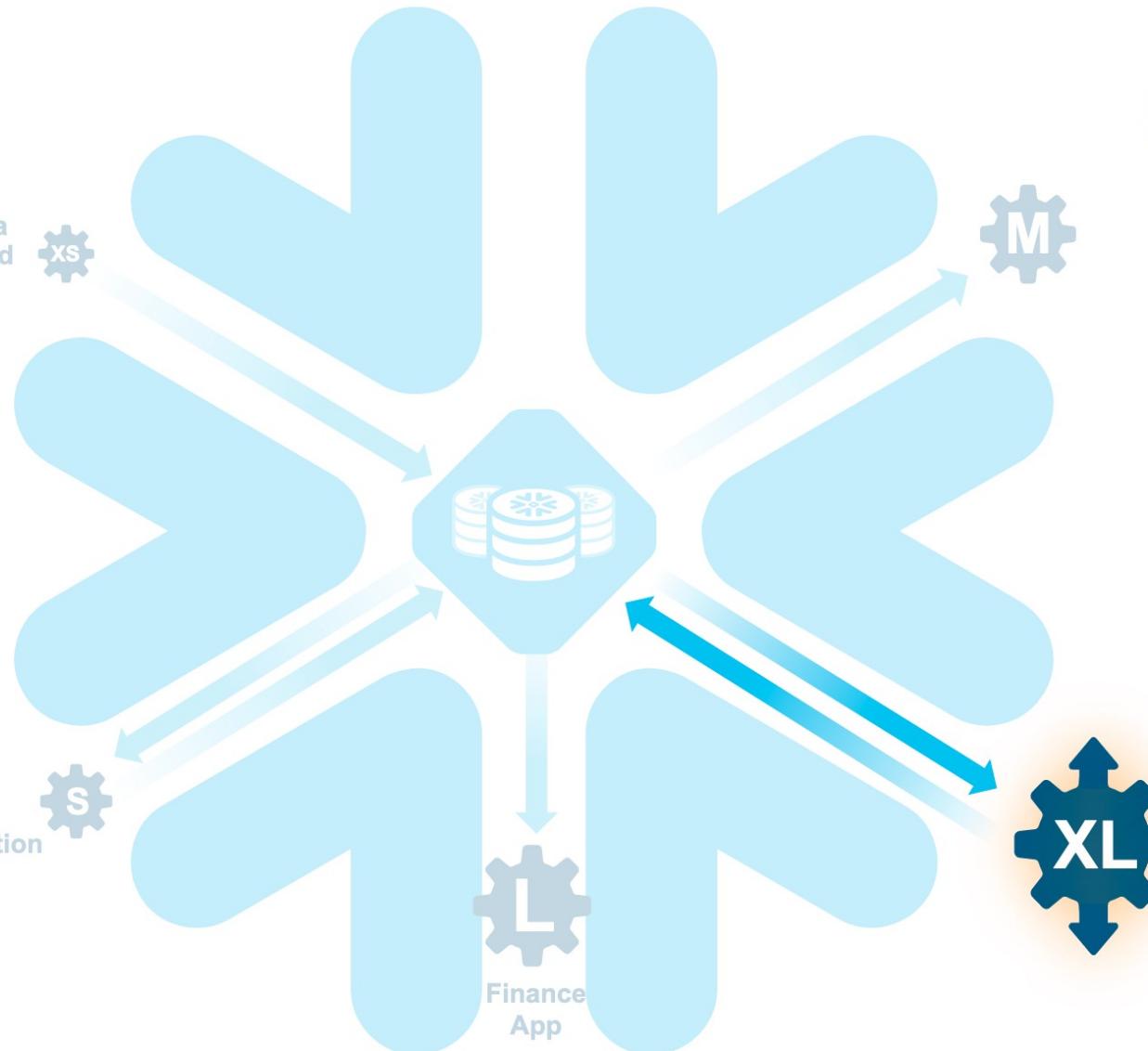
Marketing Analytics / Reporting / BI

Data Science

Structured &
Semi-Structured



Functional Architecture

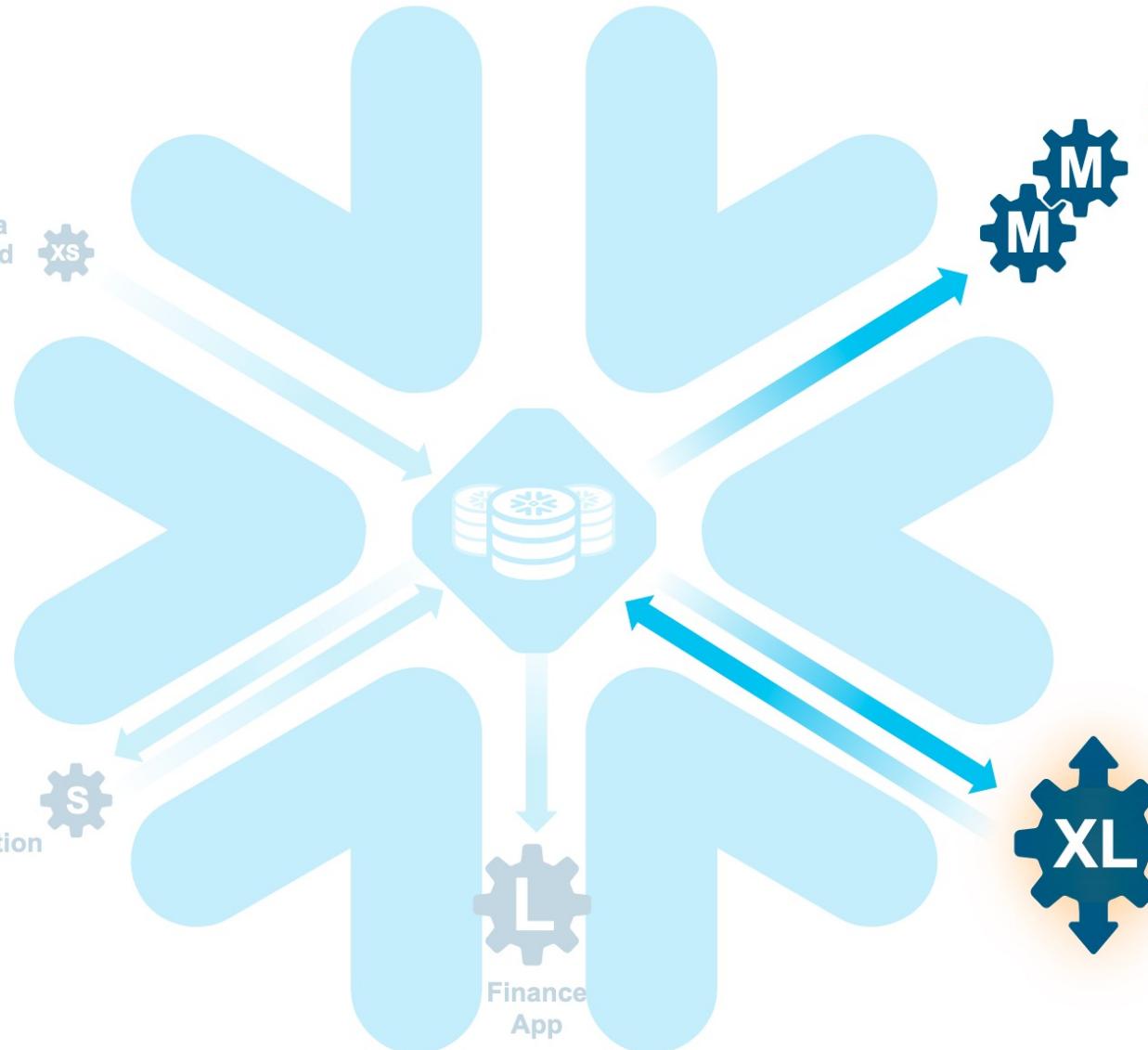
Structured & Semi-Structured

Functional Architecture

**Marketing
Analytics / Reporting / BI****Data Science**

Structured & Semi-Structured

Functional Architecture

**Marketing
Analytics / Reporting / BI****Data Science**

Structured & Semi-Structured**Marketing Analytics / Reporting / BI****Data Science**

Structured & Semi-Structured



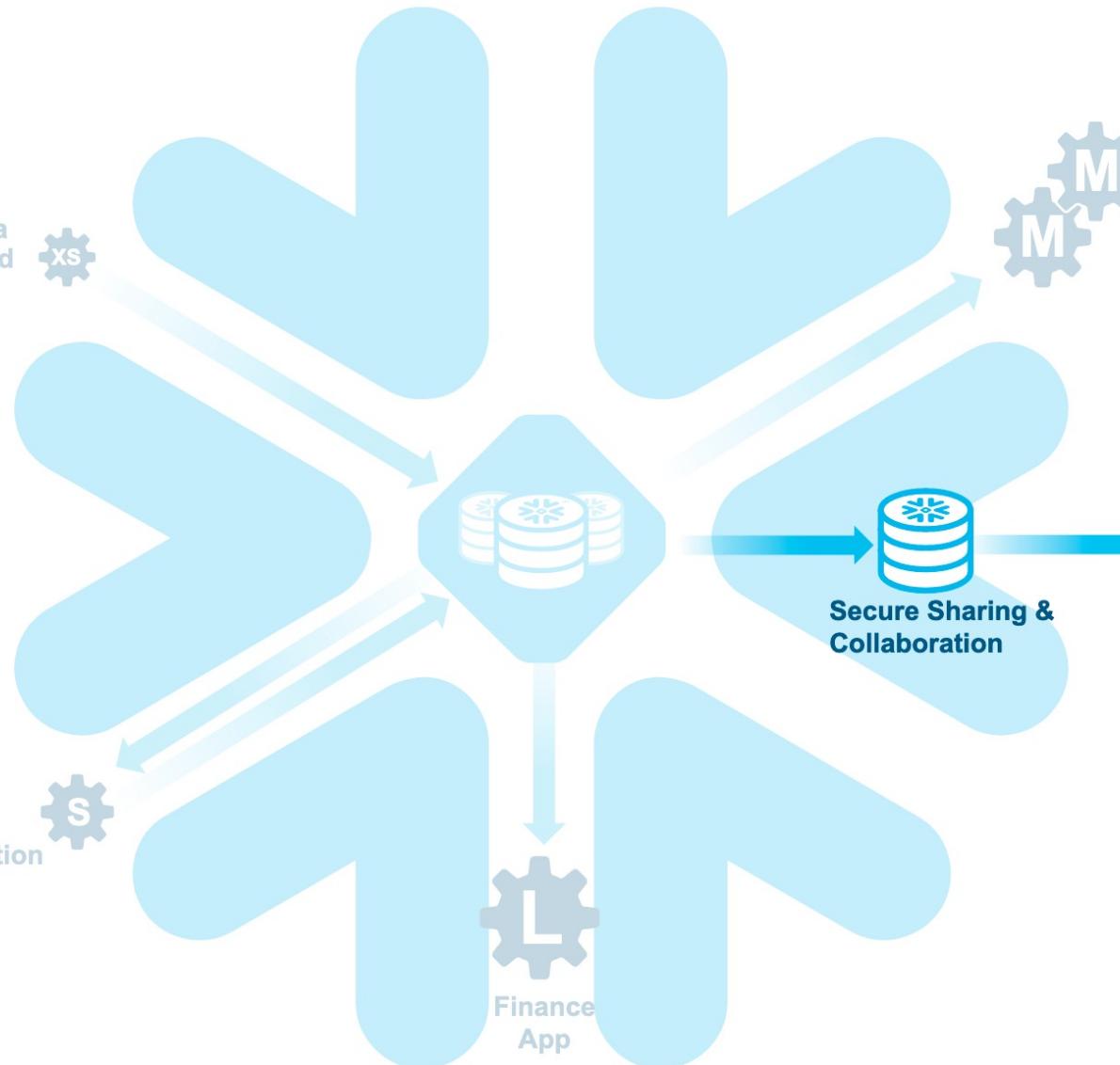
Data Load



Functional Architecture

Data Transformation

Finance
App



Marketing Analytics / Reporting / BI



Data Exchange

Your Employees



Structured &
Semi-Structured

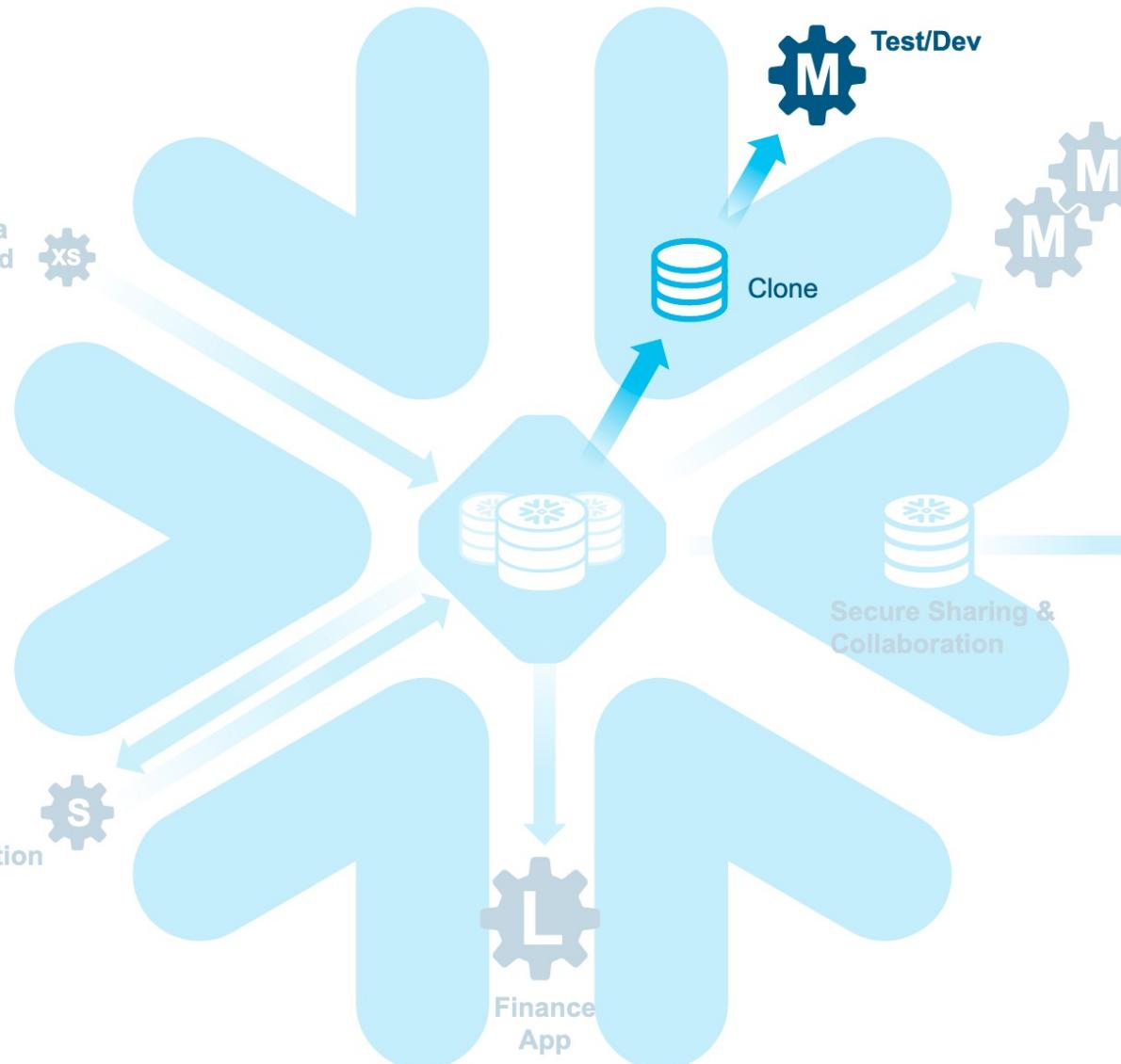


Data Load



Functional Architecture

Data Transformation



Marketing
Analytics / Reporting / BI

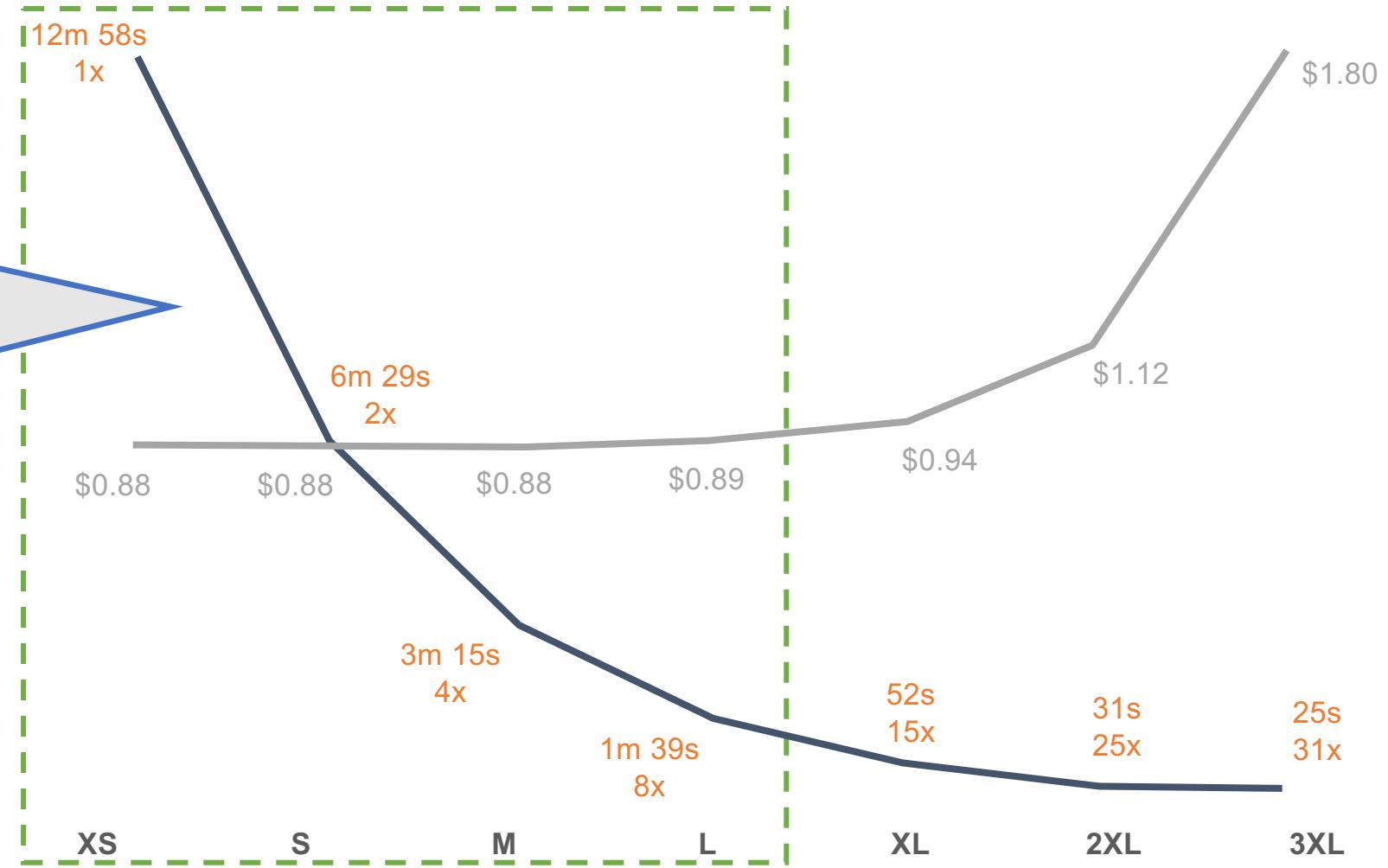
Your Business
Ecosystem

Data Exchange

SCALE UP - LOADING 1BN RECORDS

Doubling the number of servers halves the run time
but you pay per-server, per second of compute
so you get your answer
8x faster for the same cost

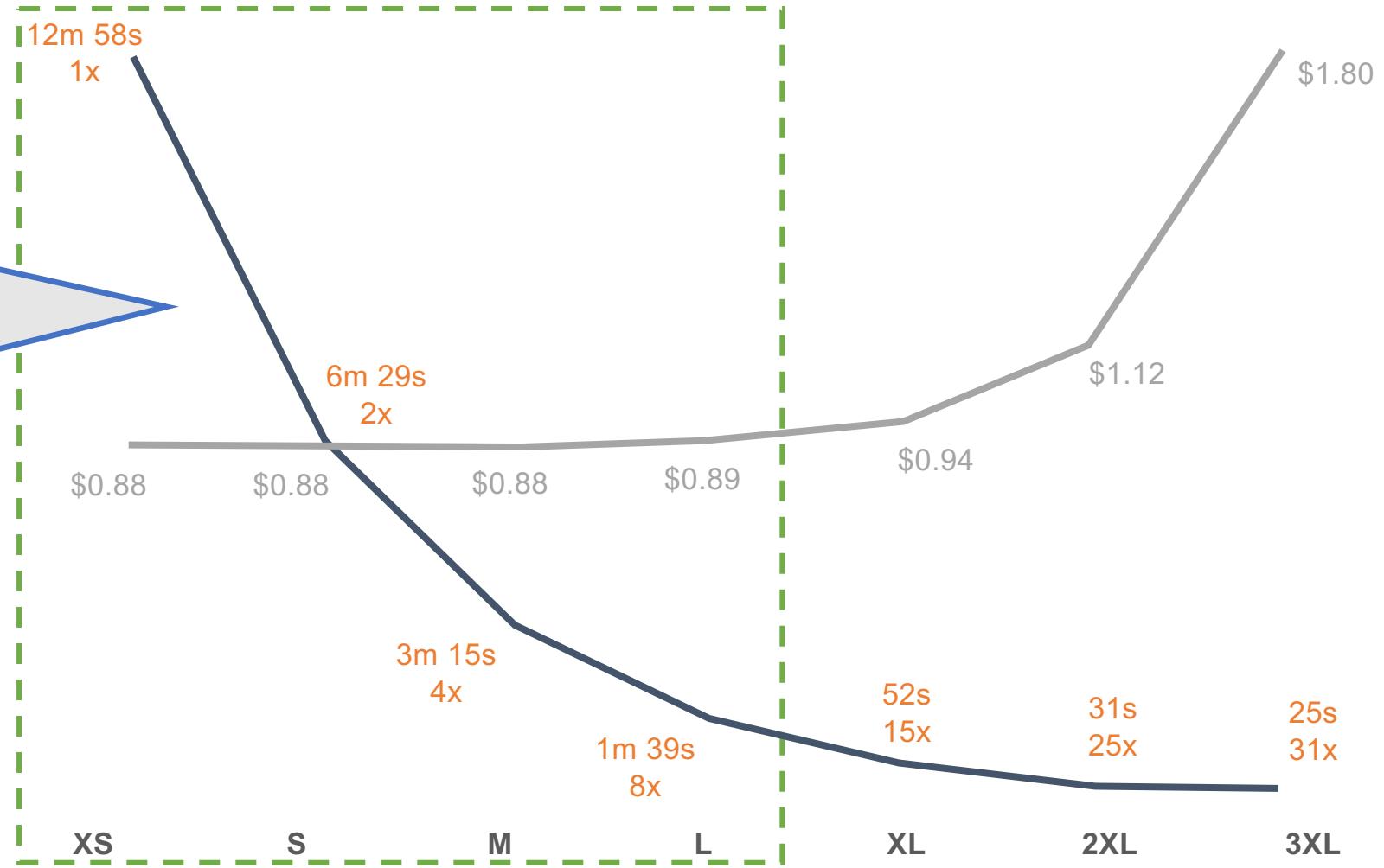
— Cost
— Secs



SCALE UP - LOADING 1BN RECORDS

Doubling the number of servers halves the run time
but you pay per-server, per second of compute
so you get your answer
8x faster for the same cost

— Cost
— Secs



COMPUTE LAYER

- Known internally as "Execution Platform" ("XP")
- A "compute cluster" is logical collection of 1+ VMs grouped together in a cluster—all VMs that are interconnected in a "mesh" network
- Size of compute cluster (number of nodes) is determined by the configured size of Virtual Warehouse that it exists within

Warehouse Size	Servers / Cluster	Credits / Hour	Credits / Second
X-Small	1	1	0.0003
Small	2	2	0.0006
Medium	4	4	0.0012
Large	8	8	0.0024
X-Large	16	16	0.0048
2X-Large	32	32	0.0096
3X-Large	64	64	0.0192
4X-Large	128	128	0.0384



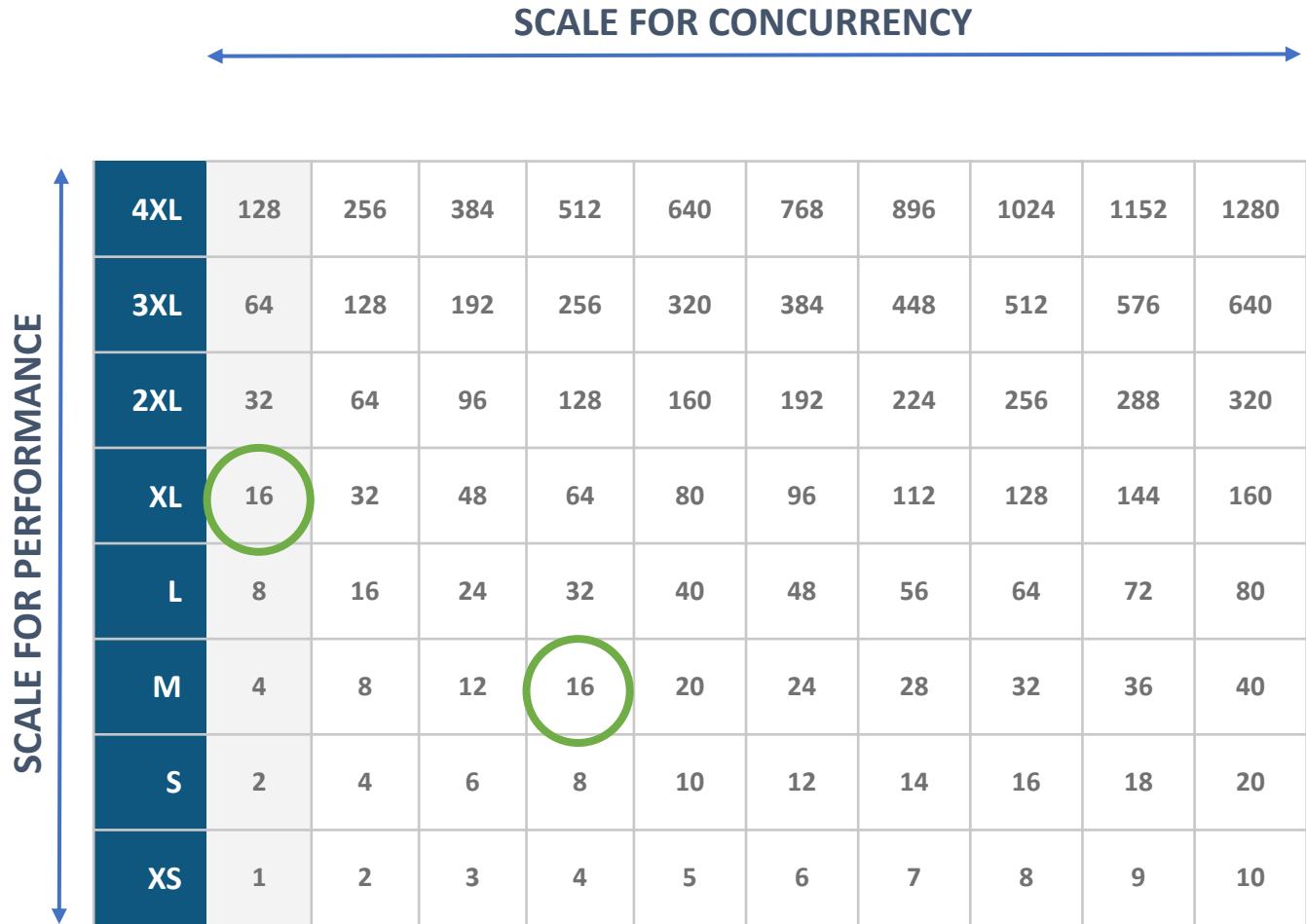
VIRTUAL WAREHOUSE SIZING

These figures indicate the number of credits consumed for an hour's worth of compute

Scale each Warehouse independently for each workload type to provide additional compute power

Automatically scale concurrent homogenized workloads with multi-cluster warehouses

While both an XL with 1 cluster and a Medium with 4 clusters both will utilize 16 credits in 1 hour, they will behave differently and are equipped for different types of workloads.



VIRTUAL WAREHOUSE MANAGEMENT

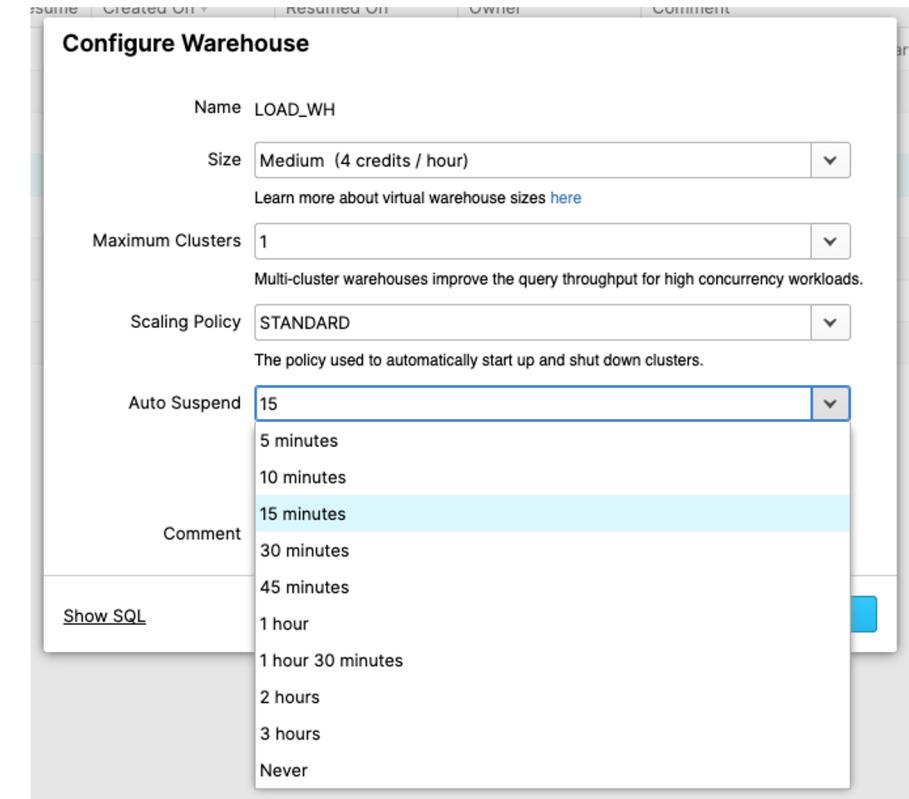
Assign unique warehouses to unique workloads

Start testing query workload with an X-SMALL, and test scaling-up.



Auto-suspend should be set for all warehouse. Time setting should be dependent on caching requirement.

Leverage MCW when it makes sense.



Warehouse Design Consideration

How are Credits Charged for Warehouses?

Credit charges are calculated based on:

- ✓ The number of servers per cluster (determined by warehouse size).
- ✓ The number of clusters (if using multi-cluster warehouses).
- ✓ The length of time each server in each cluster runs.

How Does Query Composition Impact Warehouse Processing?

The number of servers required to process a query depends on the size and complexity of the query.

- ✓ The overall size of the tables being queried has more impact than the number of rows.
- ✓ Query filtering using predicates has an impact on processing, as does the number of joins/tables in the query.

How Does Warehouse Caching Impact Queries?

Each warehouse, when running, maintains a cache of table data accessed as queries are processed by the warehouse. This enables improved performance for subsequent queries if they are able to read from the cache instead of from the table(s) in the query. The size of the cache is determined by the number of servers in the warehouse

- ✓ This cache is dropped when the warehouse is suspended
- ✓ Consider the trade-off between saving credits by suspending a warehouse versus maintaining the cache of data from previous queries to help with performance.

Scaling Up vs Scaling Out

Snowflake supports two ways to scale warehouses:

- ✓ Scale up by resizing a warehouse.
- ✓ Scale out by adding clusters to a warehouse (requires Snowflake Enterprise Edition or higher).

Warehouse Resizing Improves Performance

Resizing a warehouse generally improves query performance, particularly for larger, more complex queries.

- ✓ Larger is not necessarily faster; for smaller, basic queries that are already executing quickly,
- ✓ There is a trade-off with regards to saving credits versus maintaining the server cache.

Multi-cluster Warehouses Improve Concurrency

Multi-cluster warehouses are designed specifically for handling queuing and performance issues related to large numbers of concurrent users and/or queries.

- ✓ Multi-cluster warehouses should be configured to run in Auto-scale mode, which enables Snowflake to automatically start and stop clusters as needed.
- ✓ When choosing the minimum and maximum number of clusters for a multi-cluster warehouse



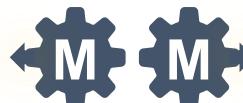
ALL TOGETHER - SCALE, ELASTICITY, COST



All three examples contain the
Same amount of work.



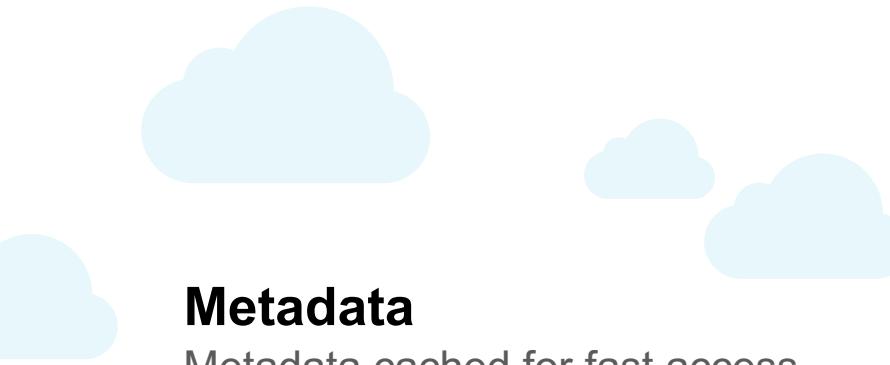
Using scale up and scale out, total
run-time is significantly reduced.



You pay per server, per second
so they **all cost the same.**

Time

ADAPTIVE CACHING



Metadata

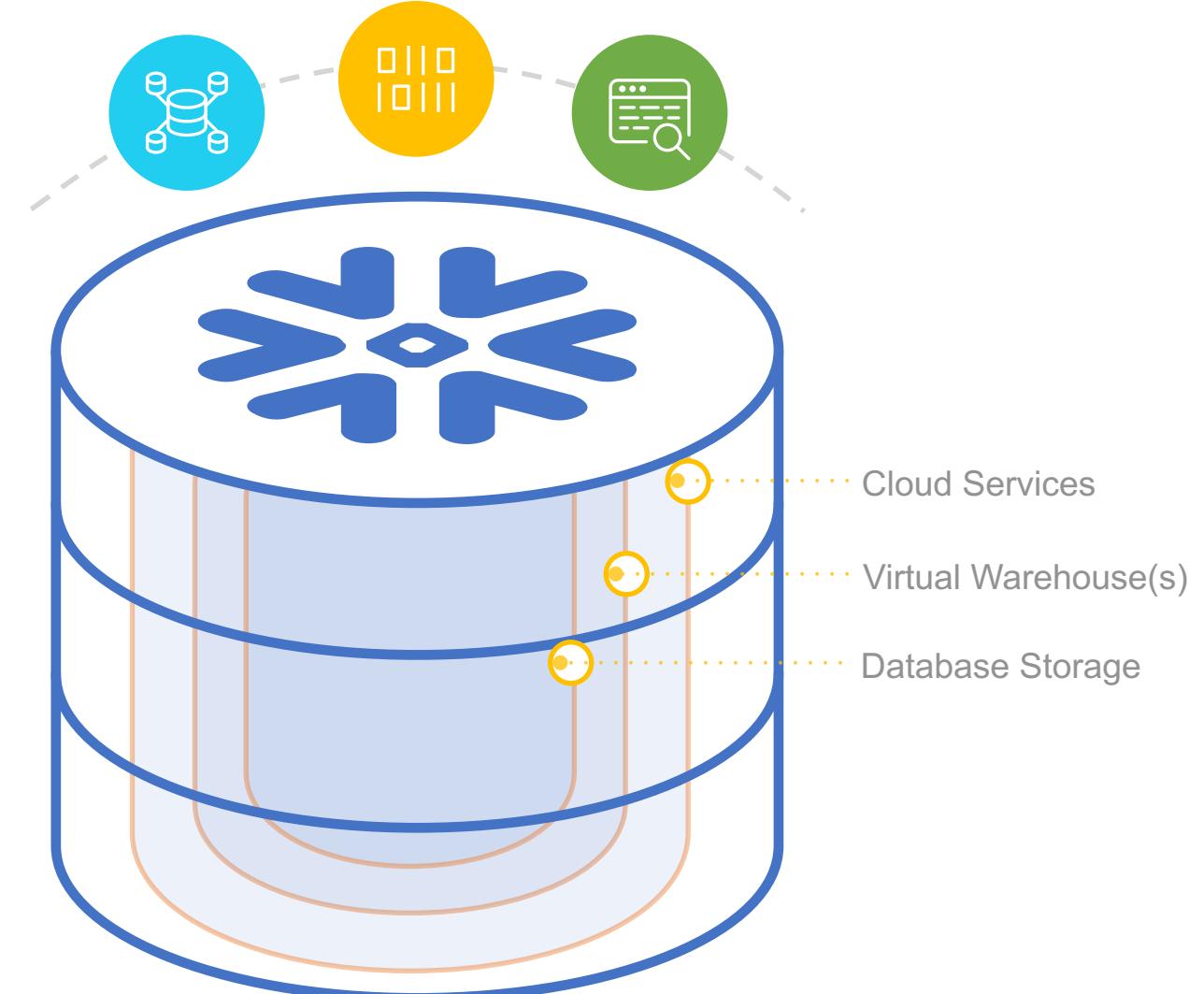
Metadata cached for fast access during query planning

Data

Active working set transparently cached on virtual warehouse SSD

Query results

Results sets cached for reuse without requiring compute (e.g., static dashboard queries)





STORAGE



ISSUES WITH TRADITIONAL PARTITIONING

- ❑ Static partitioning needs to be **defined upfront**
- ❑ Only improves queries on **partitioning columns**
- ❑ Size: can be **too big or too small**
 - ❑ Data has to be uniformly distributed over the partitions
- ❑ **Skew**: the wrong number of partitions, or an unbalanced distribution of data, can have a huge impact on performance
 - ❑ e.g., some dates may have many more observations / much more data than others



SNOWFLAKE'S AUTOMATIC MICRO-PARTITIONING

Tables divided horizontally in smaller units

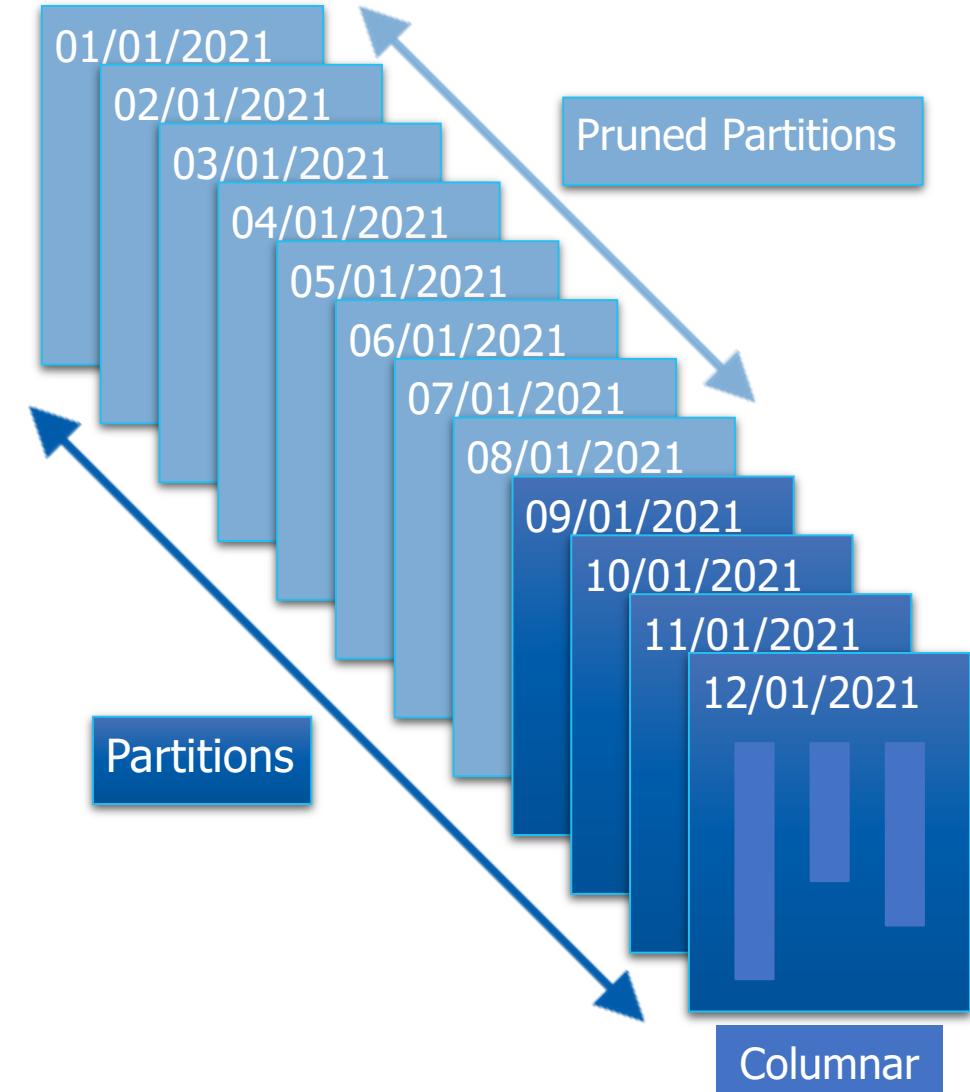
- Natural ingestion order maintains correlations between columns
- Few MBs per partition

Statistics per column stored in metadata

- Range of values
- Number of distinct values
- Number of NULLs

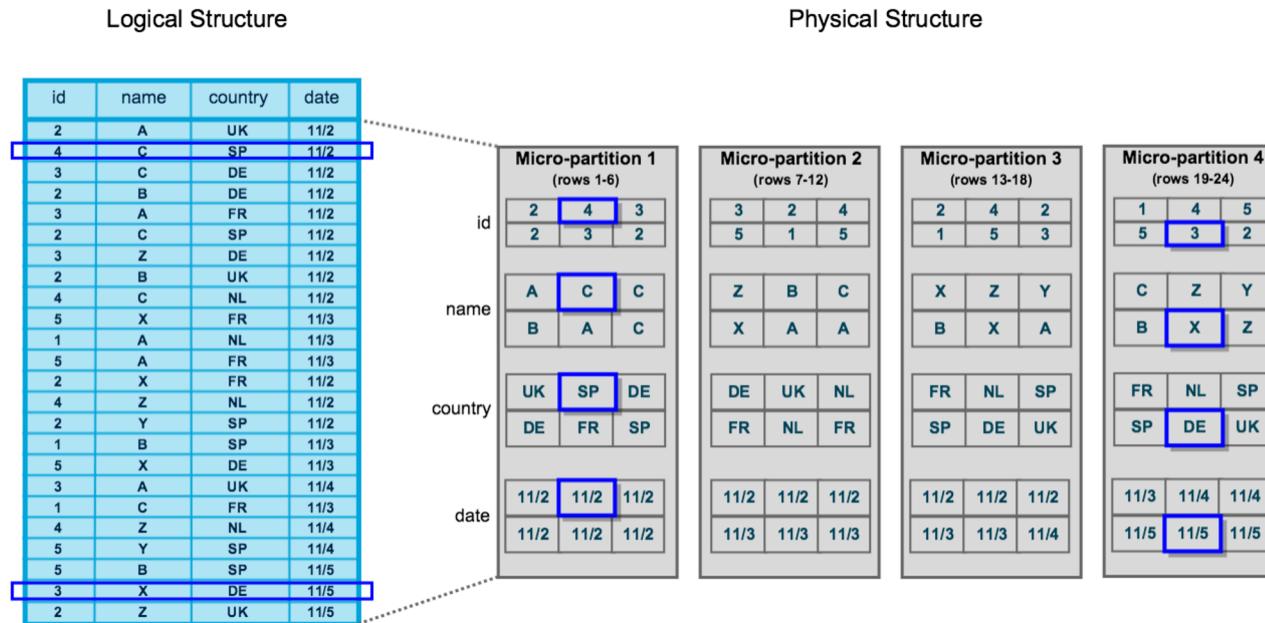
Advantages

- Optimizer uses metadata for fine-grained pruning
- Helps limiting memory footprint
- Help data shuffling



MICRO-PARTITIONING FORMAT: FDN

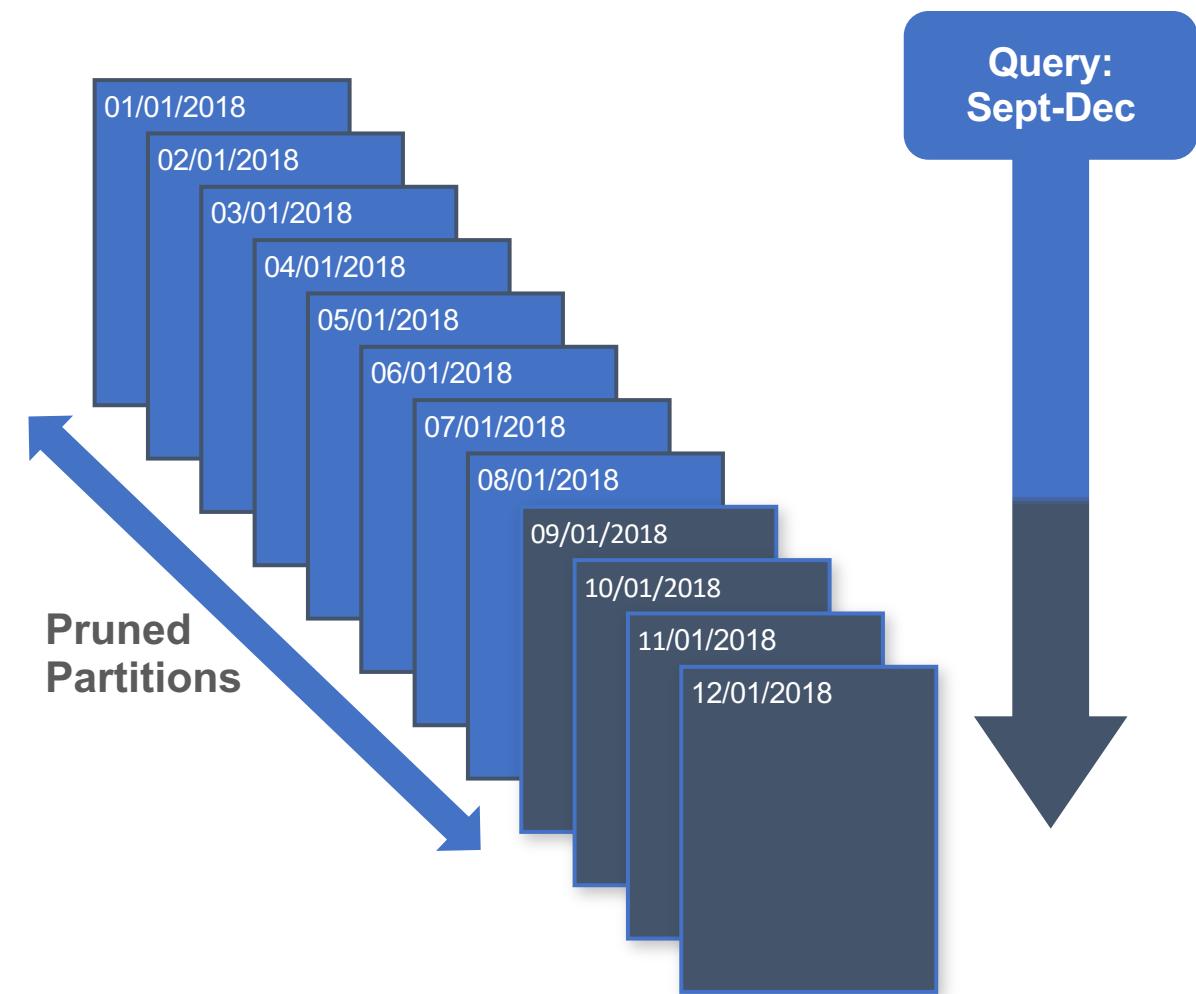
- Micro-partition files are in a **Snowflake proprietary format**: FDN
- Both structured and semi-structured data share the same proprietary FDN format
- **Since FDNs are completely immutable**
 - New FDNs are created to reflect the new final state of the data when it changed by inserts/updates/deletes (DML)
 - Old FDNs are retained and versioned in metadata (basis for Time Travel capability)



QUERY PRUNING WITH DEFAULT CLUSTERING

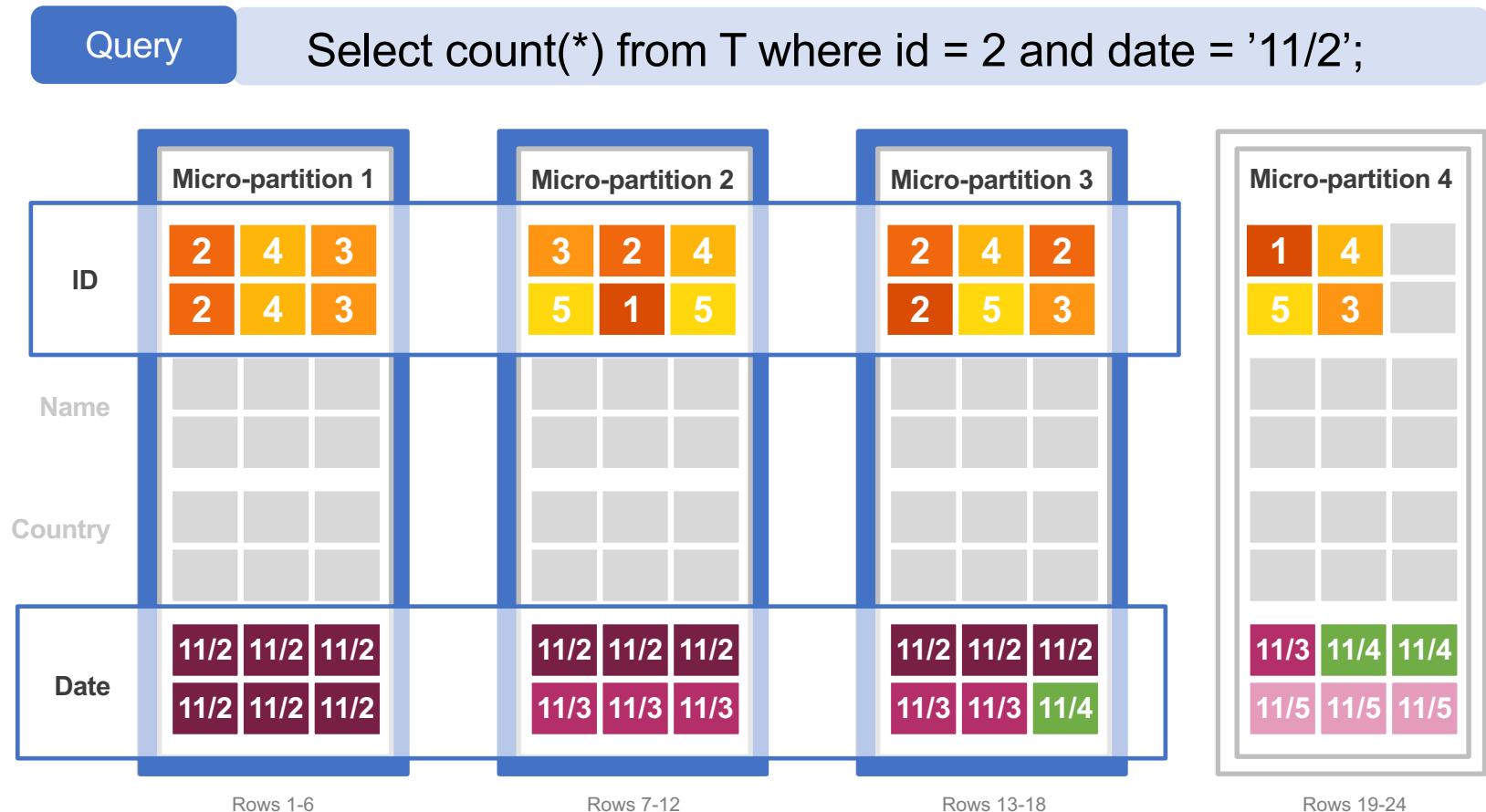
Select ... from orders where
date between
'09-01-2018' and '12-01-2018'

Select ... from orders where
product = 'iPhone'



DEFAULT CLUSTERED TABLE

ID	Name	Country	Date
2	A	UK	11/2
4	C	SP	11/2
3	C	DE	11/2
2	B	DE	11/2
3	A	FR	11/2
2	C	SP	11/2
3	Z	DE	11/2
2	B	UK	11/2
4	C	NL	11/2
5	X	FR	11/3
1	A	NL	11/3
5	A	FR	11/3
2	X	FR	11/2
4	Z	NL	11/2
2	Y	SP	11/2
2	B	SP	11/3
5	X	DE	11/3
3	A	UK	11/4
1	C	FR	11/3
4	Z	NL	11/4
5	Y	SP	11/4
5	B	SP	11/5
3	X	DE	11/5
2	Z	UK	11/5

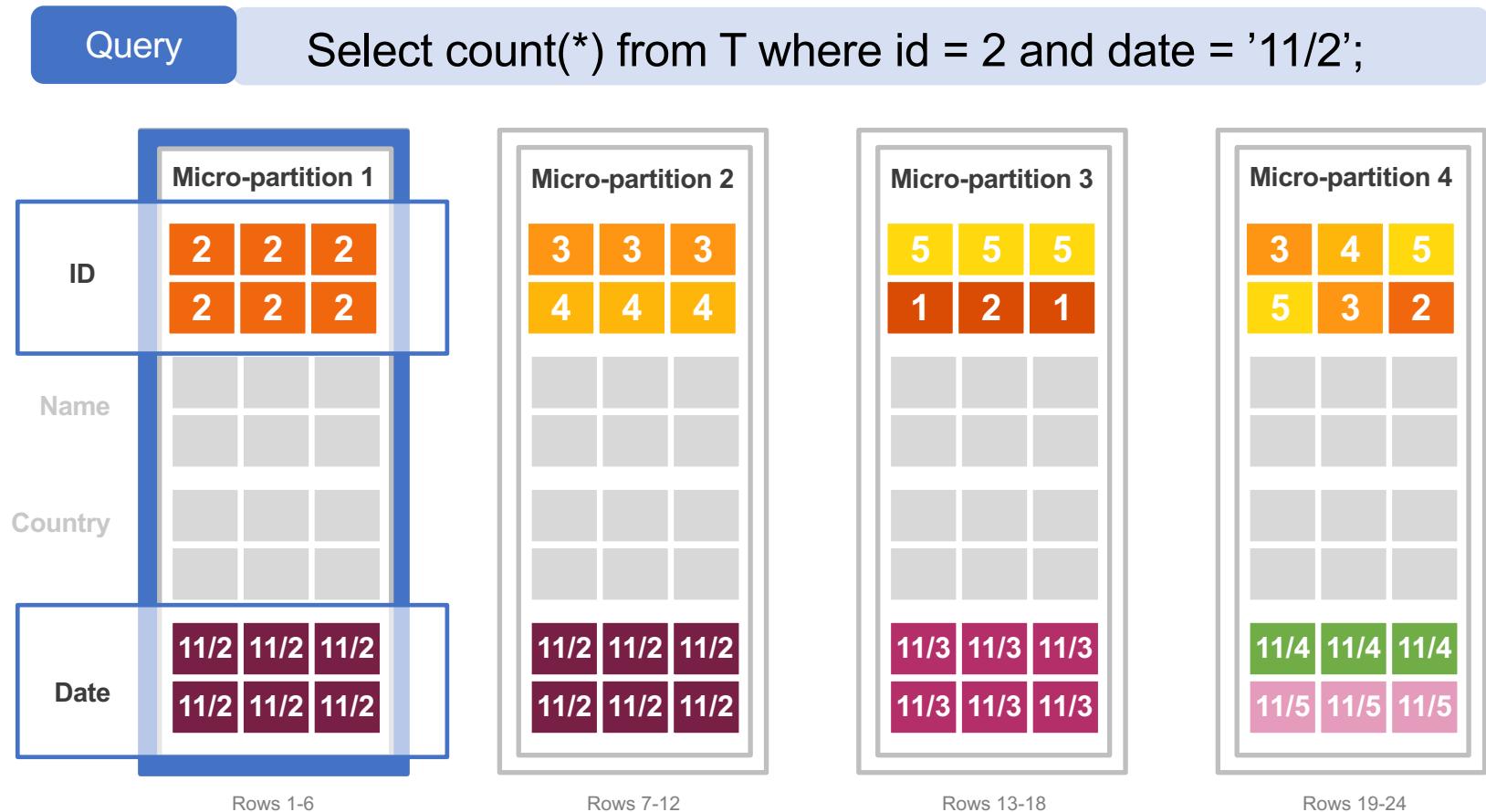


- Natural ordering by date
- Scans 3 partitions



EXPLICITLY CLUSTERED TABLE

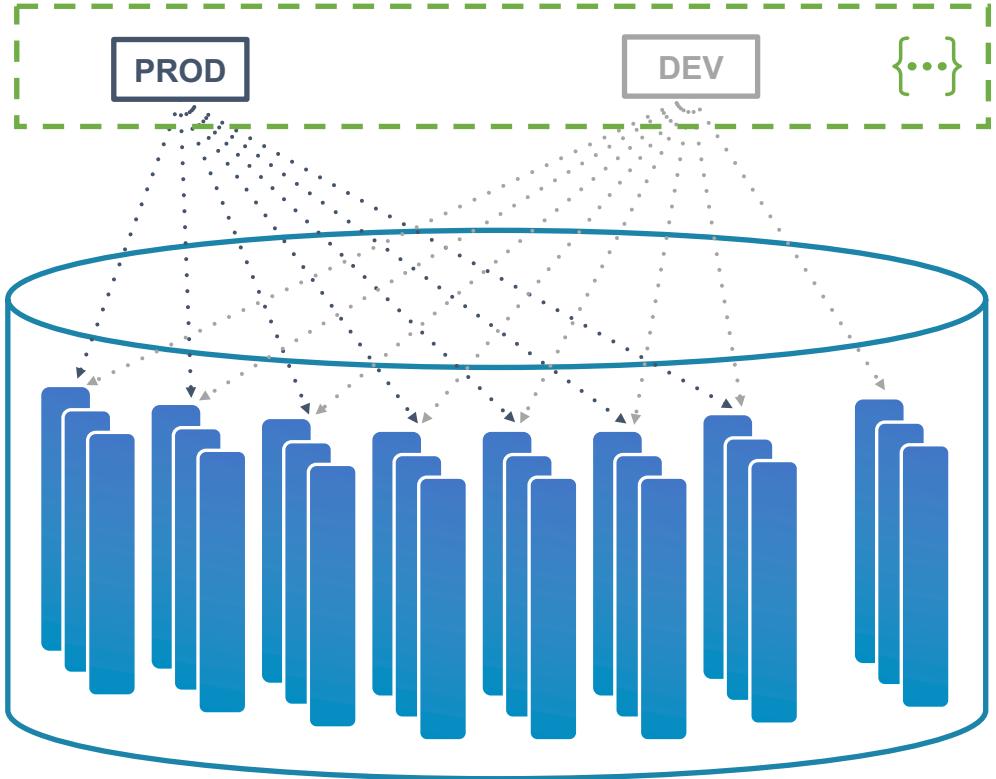
ID	Name	Country	Date
2	A	UK	11/2
4	C	SP	11/2
3	C	DE	11/2
2	B	DE	11/2
3	A	FR	11/2
2	C	SP	11/2
3	Z	DE	11/2
2	B	UK	11/2
4	C	NL	11/2
5	X	FR	11/3
1	A	NL	11/3
5	A	FR	11/3
2	X	FR	11/2
4	Z	NL	11/2
2	Y	SP	11/2
2	B	SP	11/3
5	X	DE	11/3
3	A	UK	11/4
1	C	FR	11/3
4	Z	NL	11/4
5	Y	SP	11/4
5	B	SP	11/5
3	X	DE	11/5
2	Z	UK	11/5



- Clustering keys (date, id)
- Scans only 1 partition



ZERO-COPY CLONING



The Metadata layer keeps track of every micro-partition file in every customer database.

Creating a **DEV** environment usually means copying the **PROD** database

Limited to subset of full Prod

Up to 2x storage requirement

Periodic refreshes

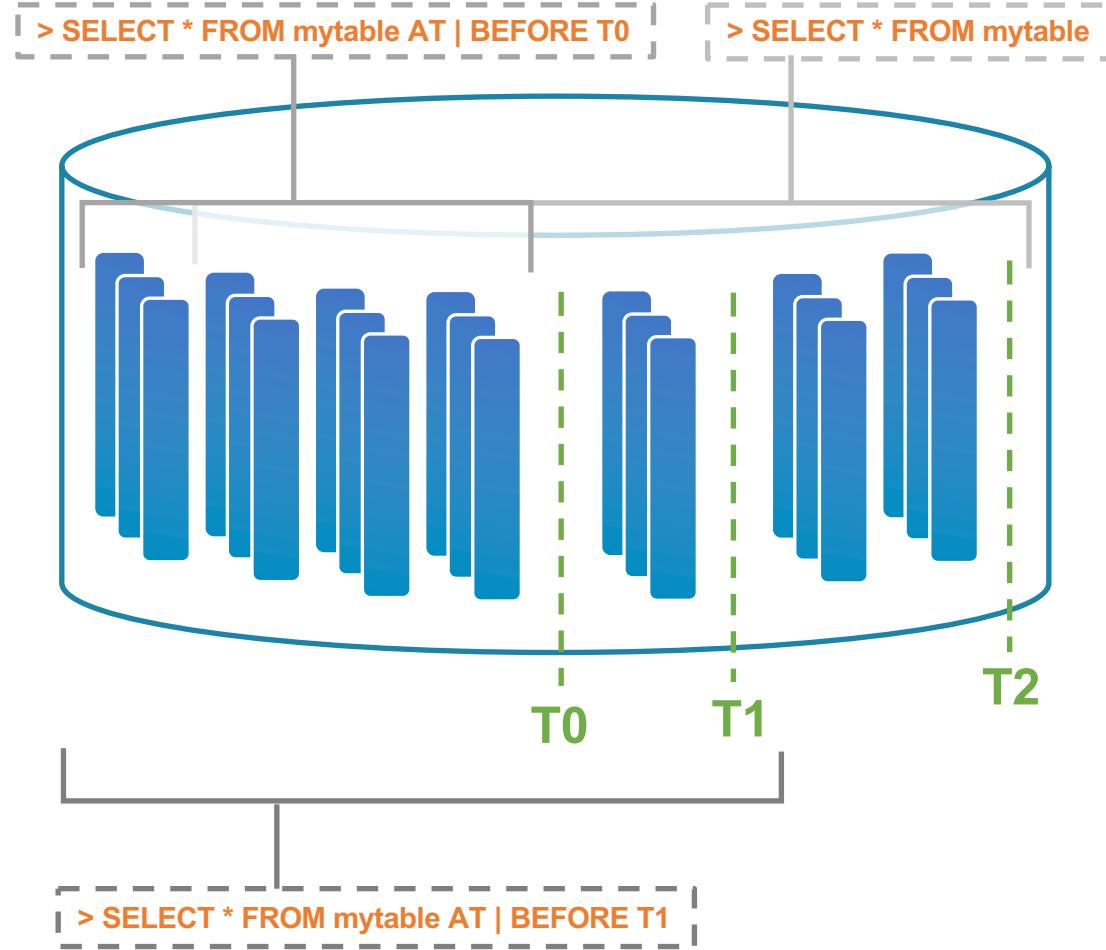
Snowflake Zero-Copy Clones

Simply “point” to the same files

Consumes zero additional storage

Changes to either DB are isolated

TIME TRAVEL



T0 – Initial state of database

T1 – update myTable set colX = Y where...

T2 – ELT job loads new data

Previous versions of data automatically retained

AT | BEFORE [timestamp | statement | offset]

CLONE AT | BEFORE to recreate a prior version

UNDROP recovers from accidental deletion

Accessed via SQL extensions

AT | BEFORE [timestamp | statement | offset]

CLONE AT | BEFORE to recreate a prior version

UNDROP recovers from accidental deletion

PROTECTING YOUR DATA IN SNOWFLAKE

End-to-End Encryption

Always-encrypted client communications, plus integration with cloud provider private networking



Fully Encrypted Storage

Data at rest is always encrypted while handled by the Snowflake drivers and systems



Strong Authentication

Built in multi-factor, integration with your federated SSO, easy user management



Full Auditing

Track every login, every transaction, every data transfer, and export to your security tools



Role-Based Access Control

All objects, actions, and even compute usage can be controlled with roles



Recovery

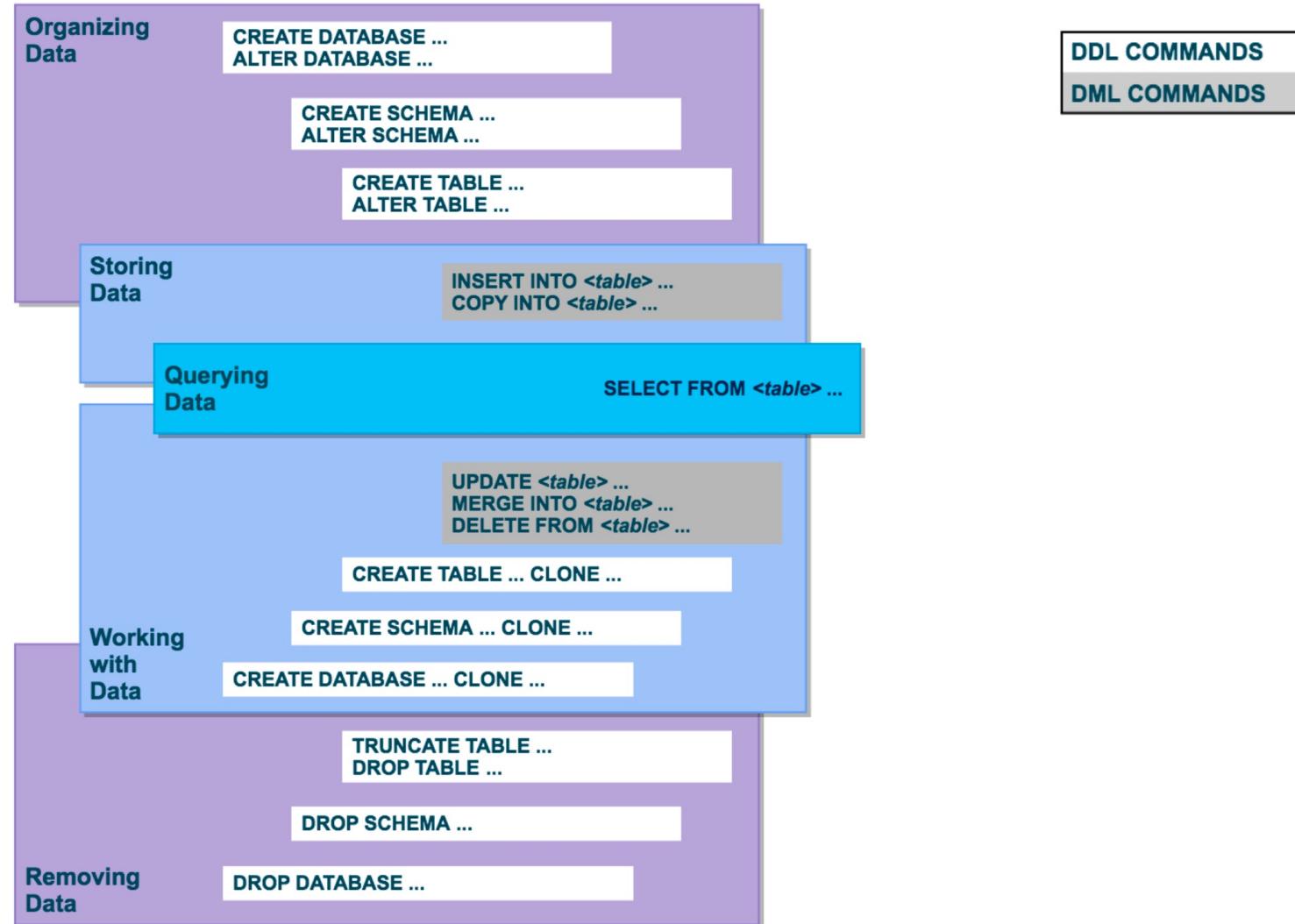
We give you options to ensure your data can be recovered in case of an accident or worse



[Snowflake Security Product Documentation](#)



Snowflake Data Life Cycle



Sample Questions

Question 1 of 30

100 pts

A

Which of the following is not a layer in Snowflake's Architecture?

- Storage
- Cloud Services
- Data Catalog
- Virtual Warehouses

Question 4 of 30

100 pts

A

What technique does Snowflake use to limit the number of micro-partitions scanned by each query?

- Indexing
- Pruning
- Map Reduce
- B-Tree

Question 3 of 30

100 pts

A

Which of the following are true of Multi-Cluster warehouses? Select all that apply.

- Adds clusters automatically based on query activity
- Sizes each cluster optimally based on the queries
- Scales down when query activity slows
- Multi-cluster warehouses will never auto-suspend



Question 6 of 30

100 pts

A

True or false: Virtual Warehouses cannot be resized while queries are running.

- TRUE
- FALSE

Sample Questions

Question 9 of 30

100 pts

A

True or false: The warehouse cache may be reset if a running warehouse is suspended and then resumed.

- TRUE
- FALSE

Question 12 of 30

100 pts

A

True or false: Snowflake only works with cloud-based tools.

- TRUE
- FALSE

Question 17 of 30

100 pts

A

Snowflake supports which of the following file formats for data loading? Select all that apply.

- Parquet
- ORC
- CSV
- PDF

Question 18 of 30

100 pts

A

Which of the following are options when creating a Virtual Warehouse?

- Auto-suspend
- Storage size
- Auto-resume
- Cache size



Sample Questions

Question 21 of 30

100 pts

A

True or false: A table in Snowflake can only be queried using the Virtual Warehouse used to load the data.

- TRUE
- FALSE

Question 25 of 30

100 pts

A

Which of the following terms best describes Snowflake's database architecture?

- Columnar shared nothing
- Shared disk
- Multi-cluster shared data
- Cloud-native shared memory

Question 22 of 30

100 pts

A

True or false: Snowflake offers tools to extract data from source systems.

- TRUE
- FALSE

Question 29 of 30

100 pts

A

Which of the following are types of caching used by Snowflake? Select all that apply.

- Warehouse caching
- Index Caching
- Metadata caching
- Query result caching



