SNOWFLAKE DATA CLOUD

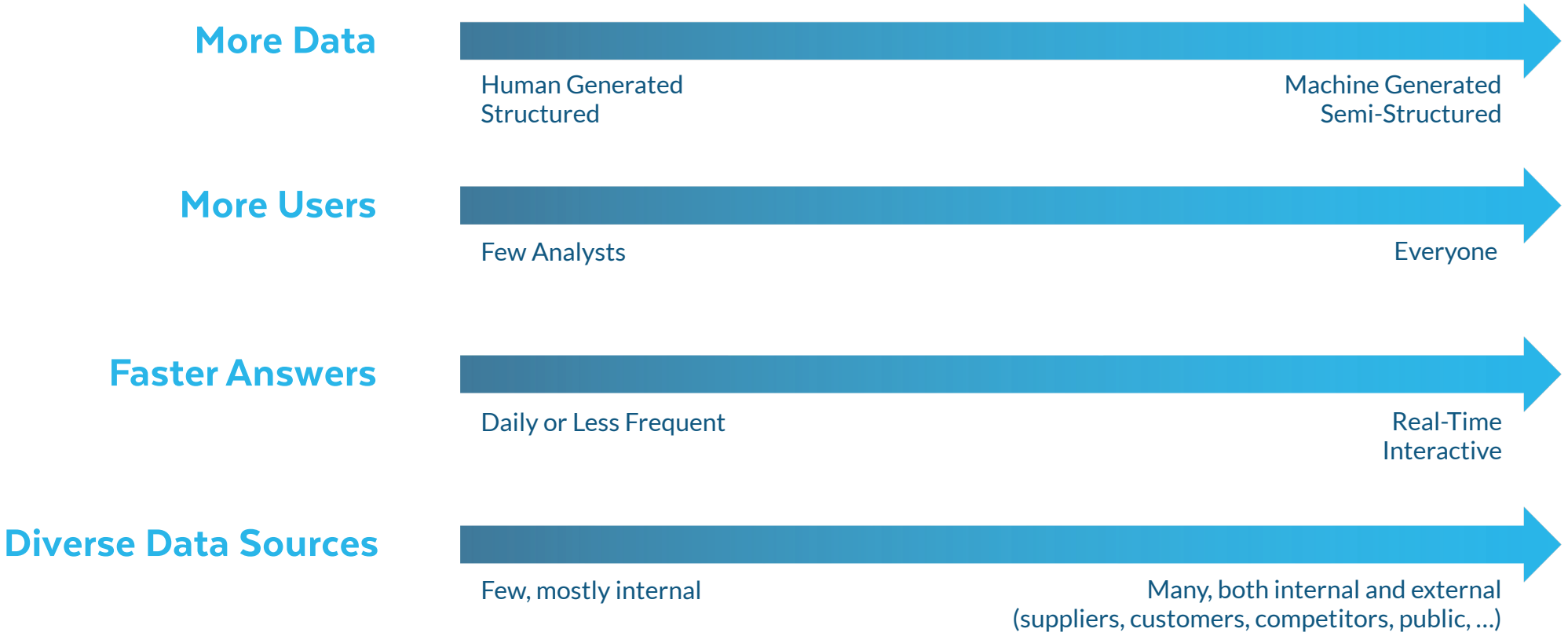
SCALING ANALYTICS AND BEYOND

HTPS, October 2022

Thierry Cruanes, Benoit Dageville

WHY SNOWFLAKE DATA CLOUD?

No good solution to tackle modern data challenges



WHY THEN?

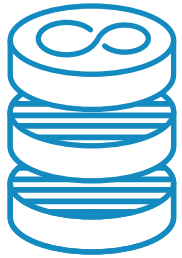
OUR (THEN OPTIMISTIC) VIEW OF THE CLOUD…

- Cloud object store made storage dirt cheap and reliable
- Flat network offered uniform and scalable bandwidth
- Single core performance stalled
- Data warehouse and analytic workload are mostly CPU bound

Design for abundance and not scarcity of resources

THE DREAM ANALYTIC SYSTEM

(CIRCA 2012)



Unlimited and Instant Scaling

- No data silos
- 10x faster for the same price, no over provisioning

+

Store all your data

- Structured and semi-structured
- Petabyte scale at very low cost

+

Extreme simplicity

- No management tasks, offered as a service
- Fast out-of-box with no tuning knobs

+

No compromises full fledge Data Warehouse

- Full support for ACID transactions with read consistency
- ANSI SQL, RBAC

BUILT AROUND THREE PILLARS

Multi-cluster shared data Architecture

- Leverage cloud elasticity and pay only what you use
- Instant scale
- Performance isolation

Immutable Scalable Storage

- Fast response time at scale
- Fine grain vertical and horizontal pruning on any data
- Automatically applied to any data (structured and semi-structured)

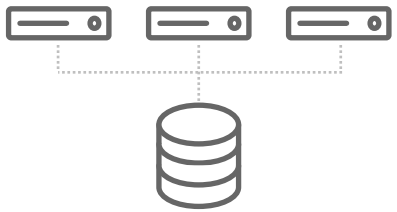
Multi-Tenant Service

- Self-tuning, self-healing
- Transparent upgrade
- Service architecture designed for availability, durability and security

HOW DOES SNOWFLAKE HANDLE SCALE?

AN ARCHITECTURE BUILT FOR THE CLOUD

Traditional Architectures



Shared-disk

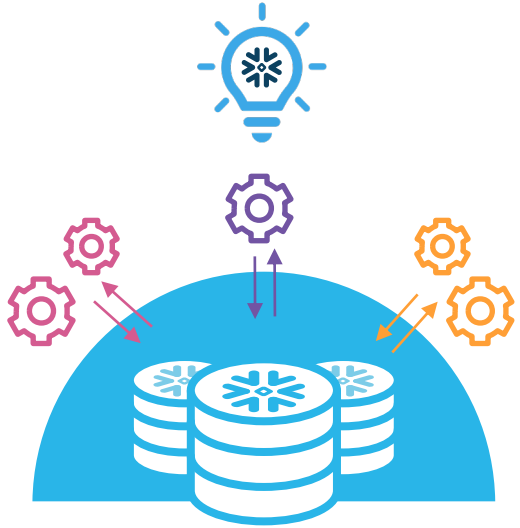
Shared storage

Single cluster

Shared-nothing

Decentralized, local storage

Single cluster

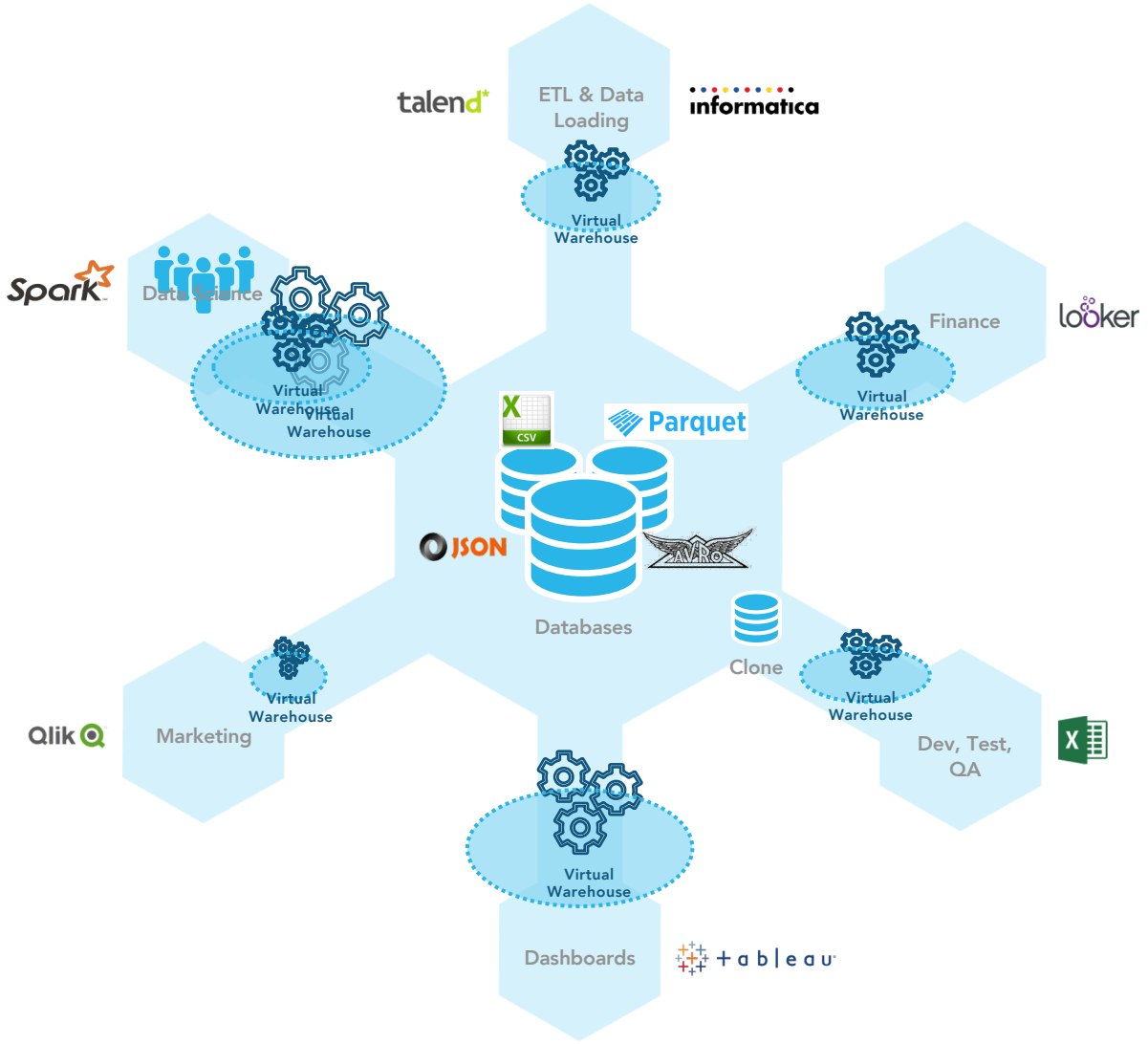


Multi-cluster, shared data

- Centralized, scale-out storage
- Multiple, independent compute clusters
- Independent control plan and routing

MULTI-CLUSTER, SHARED DATA ARCHITECTURE

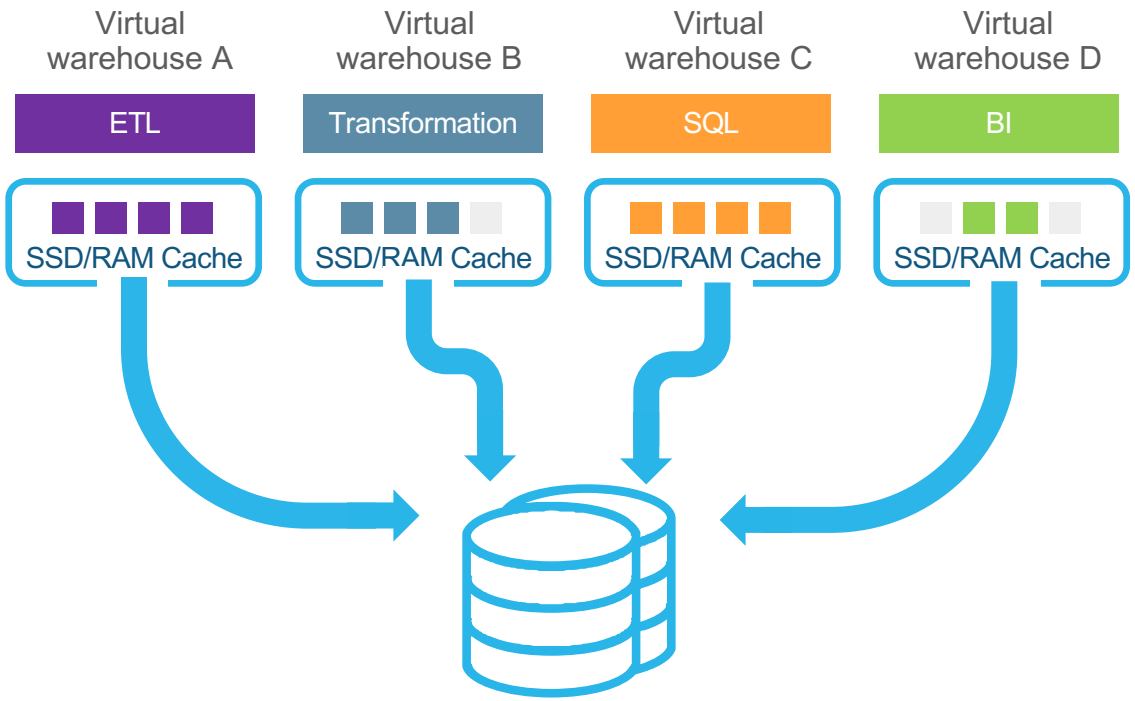
- **No data silos**
  Storage decoupled from compute
- **Any data**
  Native for structured & semi-structured
- **Unlimited scalability**
  Along many dimensions
- **Low cost**
  Compute on demand
- **Instantly cloning**
  Isolate prod from dev & qa
- **Highly available**
  11 9's durability, 4 9's availability



VIRTUAL WAREHOUSE

How to allow concurrent workloads run without impacting each other?

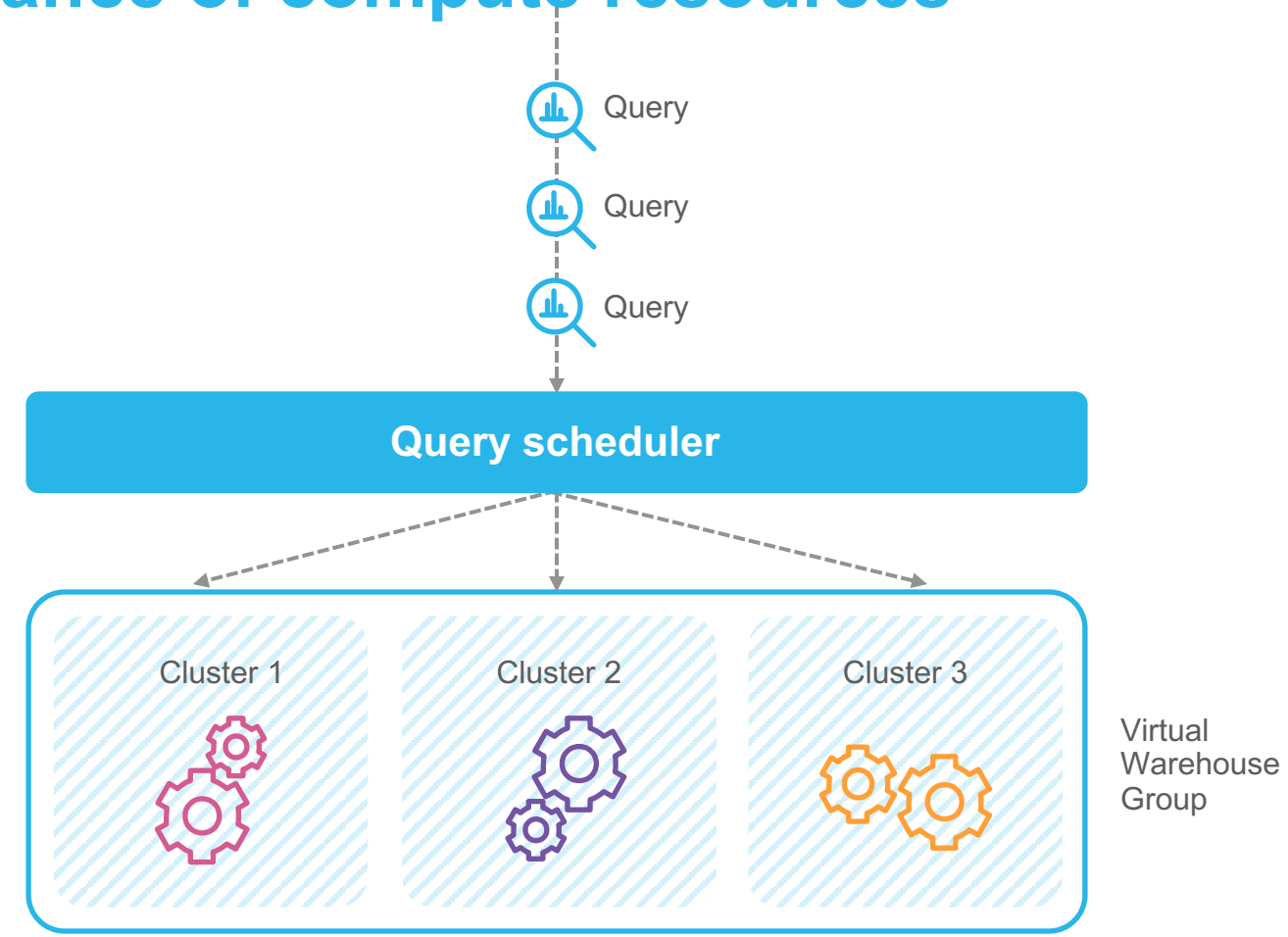
- One or more MPP compute cluster
- Unit of fault and performance isolation
- Use multiple warehouses to segregate workload
- Resizable on the fly
- Able to access data in any database
- Transparently caches data accessed
- Transaction manager synchronizes data access
- Automatic suspend when idle and resume when needed



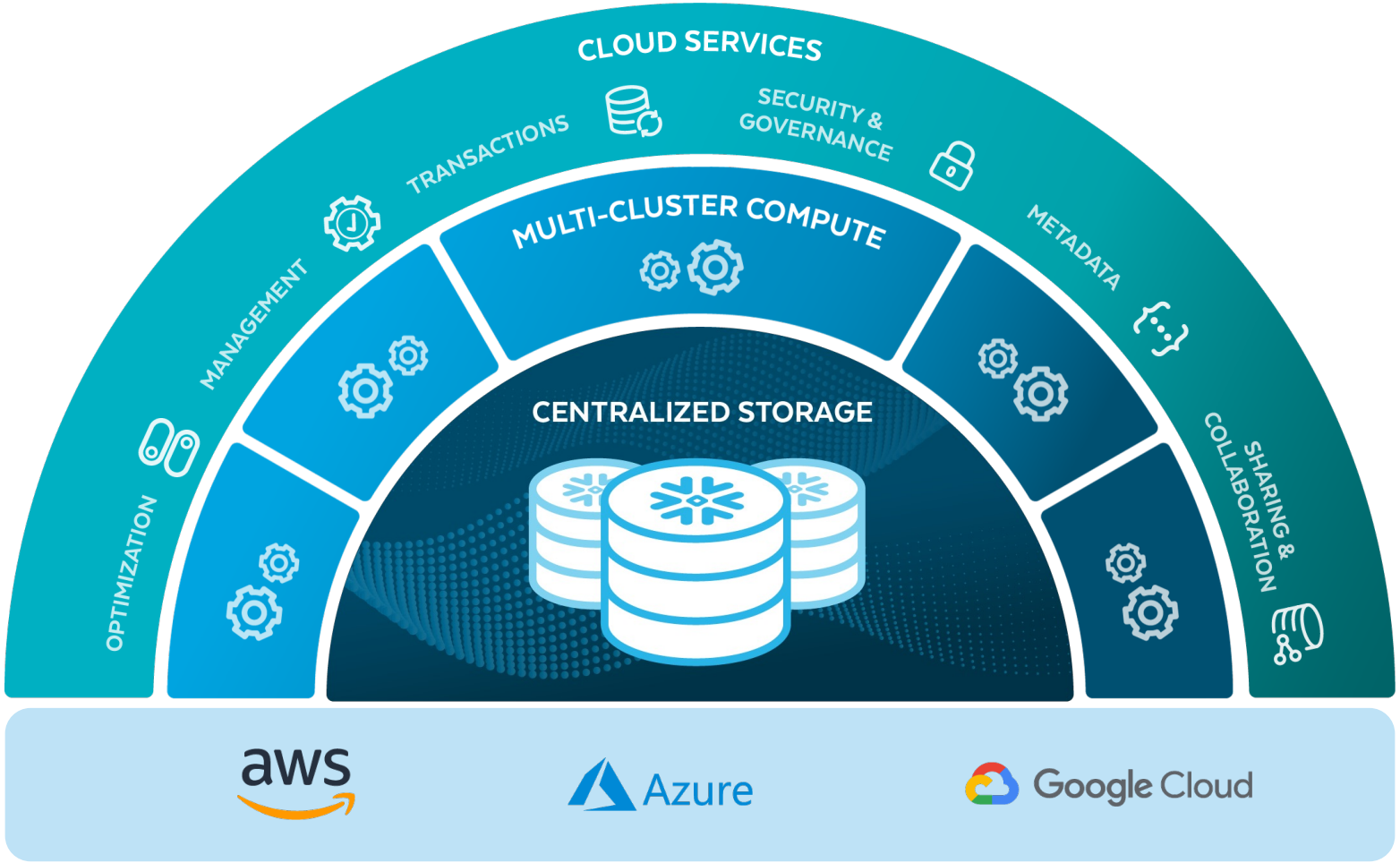
MULTI-CLUSTER WAREHOUSE

orchestrate abundance of compute resources

- Automatically scales compute resources based on concurrent usage
- Single virtual warehouse of multiple compute clusters
- Queries are load balanced across the clusters in a virtual warehouse
- Split across availability zones for high availability



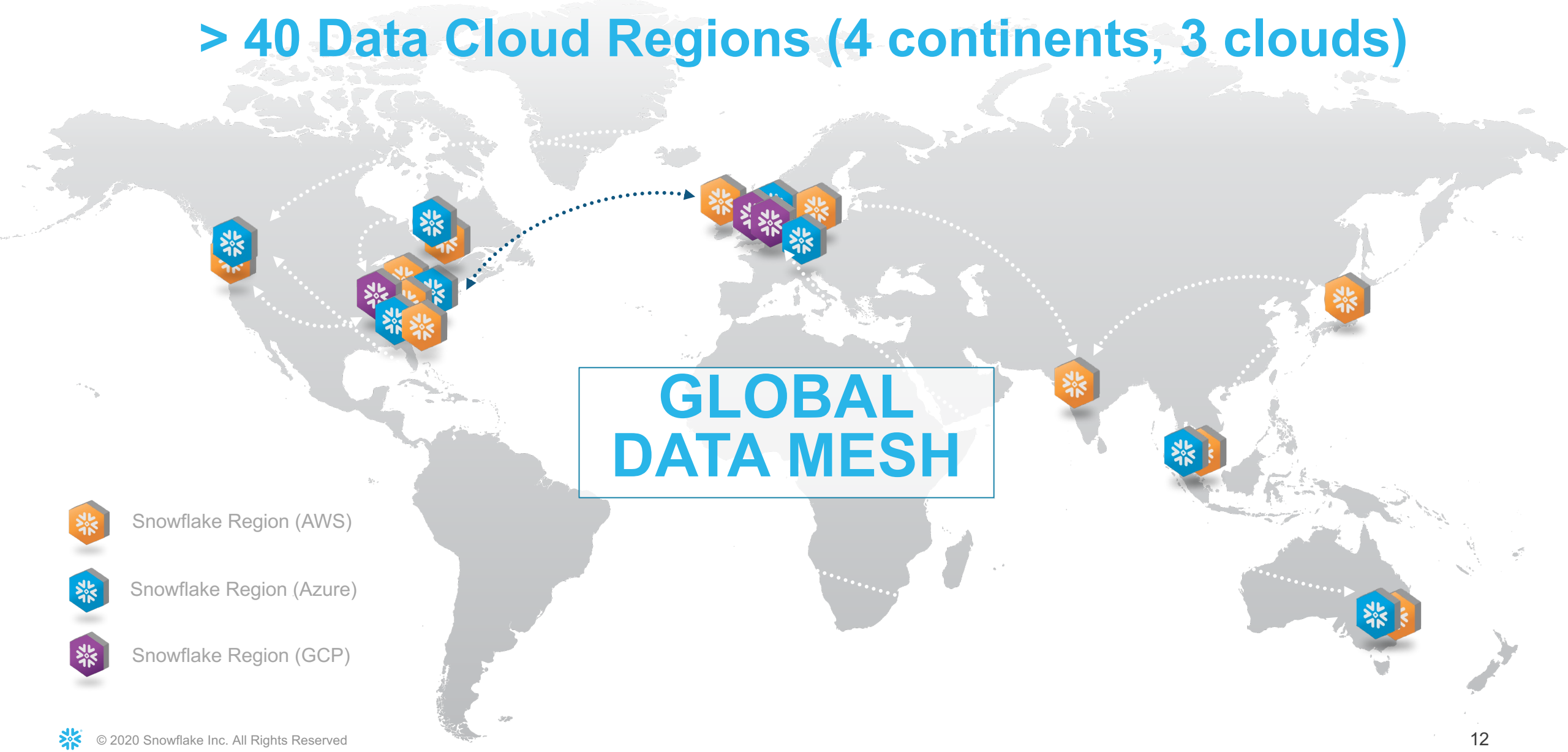
SNOWFLAKE DATA CLOUD REGION



© 2020 Snowflake Inc. All Rights Reserved

SNOWFLAKE DATA CLOUD

> 40 Data Cloud Regions (4 continents, 3 clouds)



© 2020 Snowflake Inc. All Rights Reserved

EMBRACE MORE WORKLOADS

SOFTWARE CONTROLS HARDWARE

Workload Scaling

- Increased Concurrency
- Quasi infinite resources for a given job
- Decrease wasteful resources

Workload isolation

- Performance isolation
- Failure Isolation
- Security and Sandboxing
- Transparent upgrade

Workload Specialization

- Adapt hardware to software requirements
- Support wildly disparate needs
- More cpu, more memory, specialized hardware, etc…

SNOWFLAKE DATA CLOUD

One Integrated Platform Supporting Multiple Workload Types

© 2020 Snowflake Inc. All Rights Reserved

SNOWFLAKE DATA CLOUD

One Integrated Platform Supporting Multiple Workload Types

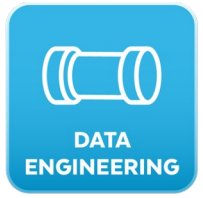
Complete SQL
ACID
Low-latency
High-concurrency
UDFs, UDTs
Data Governance
Stored Procedures

© 2020 Snowflake Inc. All Rights Reserved

SNOWFLAKE DATA CLOUD

One Integrated Platform Supporting Multiple Workload Types

Complete SQL
ACID
Low-latency
High-concurrency
UDFs, UDTs
Data Governance
Stored Procedures



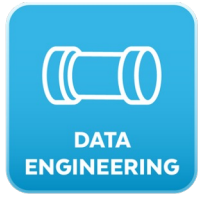
Streaming Ingest
Tasks
Table Streams
External Functions
Data Pipelines

© 2020 Snowflake Inc. All Rights Reserved

SNOWFLAKE DATA CLOUD

One Integrated Platform Supporting Multiple Workload Types

Complete SQL
ACID
Low-latency
High-concurrency
UDFs, UDTs
Data Governance
Stored Procedures



Streaming Ingest
Tasks
Table Streams
External Functions
Data Pipelines

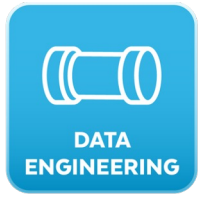
Semi-structured
Data
Unstructured Data
External Tables
Open File Format

© 2020 Snowflake Inc. All Rights Reserved

SNOWFLAKE DATA CLOUD

One Integrated Platform Supporting Multiple Workload Types

Complete SQL
ACID
Low-latency
High-concurrency
UDFs, UDTs
Data Governance
Stored Procedures



Streaming Ingest
Tasks
Table Streams
External Functions
Data Pipelines

Semi-structured
Data
Unstructured Data
External Tables
Open File Format

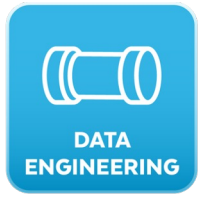
Java/Scala/Python
Data Frames

© 2020 Snowflake Inc. All Rights Reserved

SNOWFLAKE DATA CLOUD

One Integrated Platform Supporting Multiple Workload Types

Complete SQL
ACID
Low-latency
High-concurrency
UDFs, UDTs
Data Governance
Stored Procedures



Streaming Ingest
Tasks
Table Streams
External Functions
Data Pipelines

Semi-structured
Data
Unstructured Data
External Tables
Open File Format

Java/Scala/Python
Data Frames

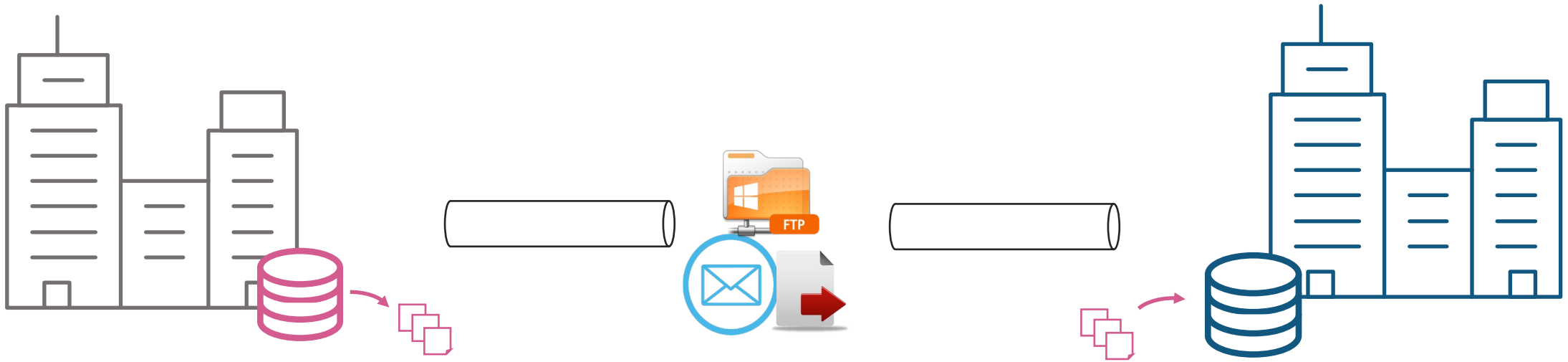
Rest APIs
Programmable
HTAP

© 2020 Snowflake Inc. All Rights Reserved

DATA APPLICATION
=
SHARED DATA +
SHARED CODE

DATA COLLABORATION

Traditional Way



Data providers

1. Export data to files
2. Publish schema
3. Stage files for transport

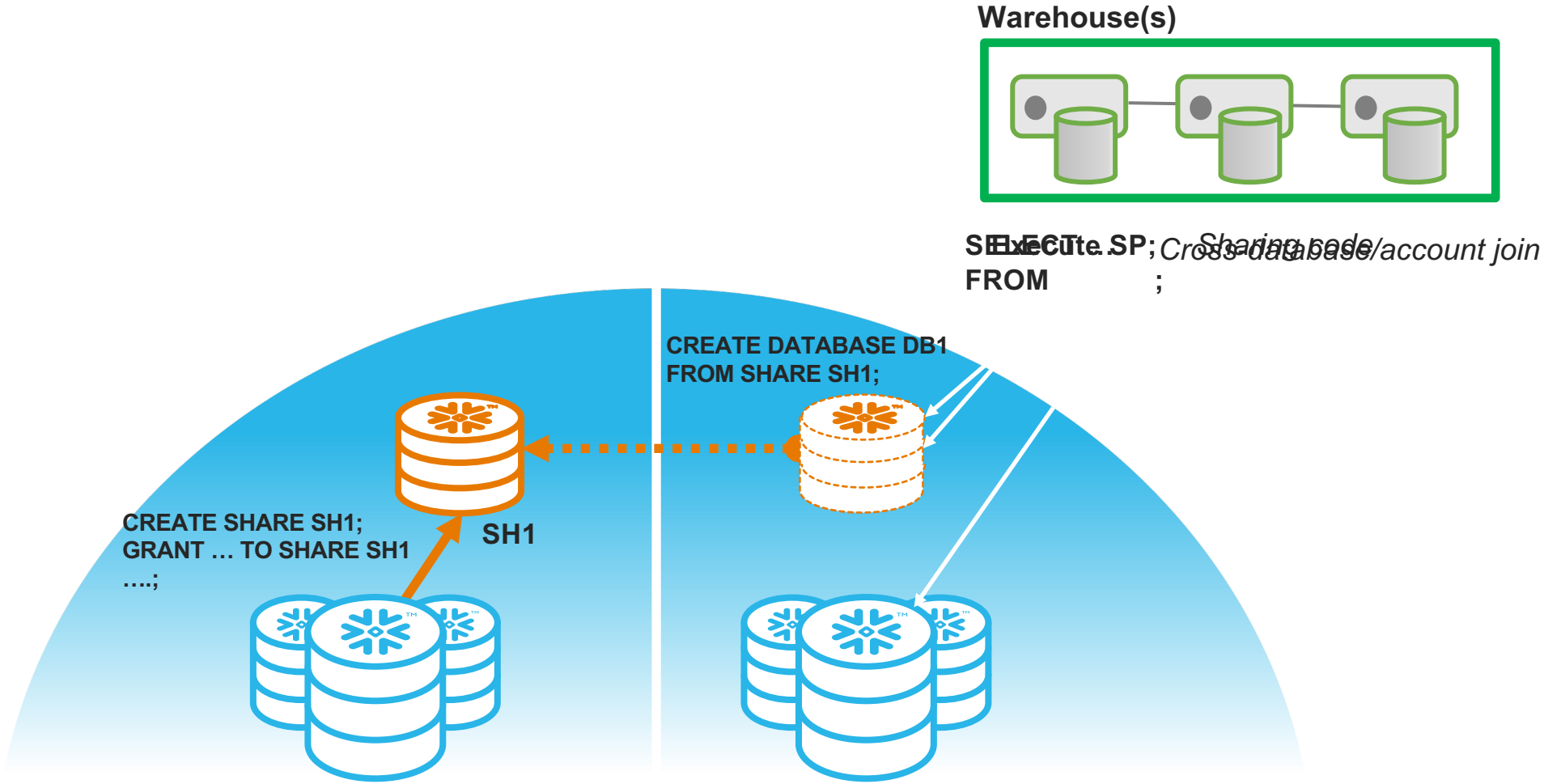
- Redundant
- Inflexible
- Inefficient
- Insecure
- Expensive

Data consumers

1. Additional infrastructure
2. Forced to recreate data structure
3. Delayed updates to data

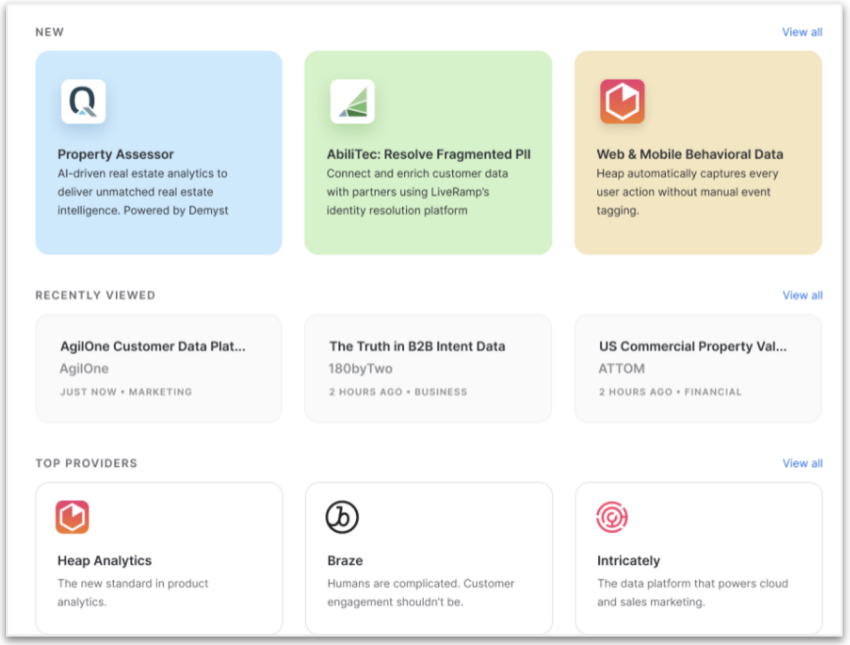
© 2017 Snowflake Computing Inc. All Rights Reserved.

SNOWFLAKE DATABASE SHARING



SNOWFLAKE DATA MARKETPLACE

READY TO USE DATABASES FROM MULTIPLE PROVIDERS

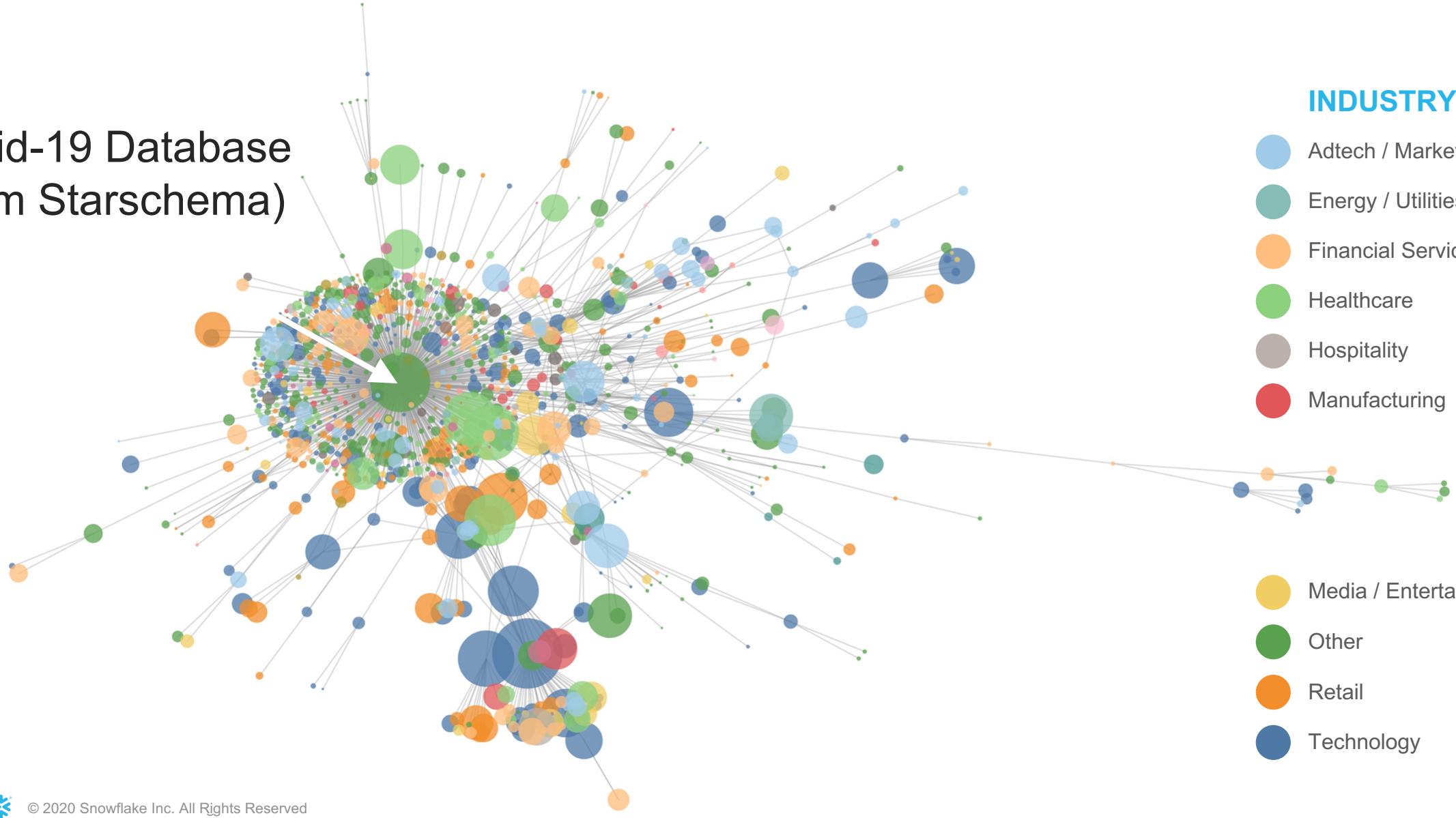


- Live, ready-to-query data; no copying or moving
- Only data marketplace with personalized data
- Globally available, across clouds

Financial | Marketing | Demographic | Macroeconomic | Government | Healthcare | Business

CONNECTED THROUGH DATA

Covid-19 Database
(from Starschema)



© 2020 Snowflake Inc. All Rights Reserved

BUT SHARING DATA IS NOT ENOUGH

Understanding data is expensive because semantic is too low-level

Making use of the data requires deep expertise

Build, deploy and distribute global application is hard

Both Code and Data are confidential

Data Application

DATA APPLICATION

CONNECT THROUGH THE PLATFORM DATA AND EXPERTISE

Simple

- Single system to manage application state
- Analytical and Transactional processing
- Streaming, transformation and replication

Trusted Data Enclave

- Share data securely
- Execute trusted application code
- Integrated and auditable authentication and authorization

Distribute, Monetize and Support

- Provider control application and data lifecycle
- Consumer controls resources
- Logs, events and metrics

STILL IN THE FIRST INNING OF THAT BRAVE NEW WORLD

STILL NEED TO UNDERSTAND

- Impact of the Cloud platform on system architecture
- Impact of application generated data
- Impact of internalizing data applications

LESSONS LEARNED

BUILDING A LARGE-SCALE GLOBAL SERVICE

Lessons Learned

Way harder than anticipated…

- Customers expect at least 3+ 9's of availability, 24x7
- At large scale, anything will happen. ➔ need to proactively anticipate and defend
- Everything needs to be fully automated and fully adaptive
- As much as possible self-managed versus dev-ops automation
- Keeping up with exponential growth ➔ scale cloud services and removing bottlenecks
- Weekly release without introducing (visible) regressions

© 2020 Snowflake Inc. All Rights Reserved

BUILDING A LARGE-SCALE GLOBAL SERVICE

Lessons Learned

Way harder than anticipated…

- Customers expect at least 3+ 9's of availability, 24x7
- At large scale, anything will happen. ➔ need to proactively anticipate and defend
- Everything needs to be fully automated and fully adaptive
- As much as possible self-managed versus dev-ops automation
- Keeping up with exponential growth ➔ scale cloud services and removing bottlenecks
- Weekly release without introducing (visible) regressions

… but so much faster development cycles

- We have built a top-notch and feature rich platform in only few years!
- Weekly release worldwide with single version to maintain
- Virtuous cycle – data driven development to identify and prioritize feature development
- Snowflake platform is extensively instrumented ➔ we generate many terabytes of service data daily
- SQL is used throughout the platform, from the control plan to processing logs, metrics, incidents

© 2020 Snowflake Inc. All Rights Reserved