snowflake®

Databases and the Cloud: Opportunities and Challenges

FG DB Spring Symposium

March 24, 2022

HPI, Potsdam, Germany

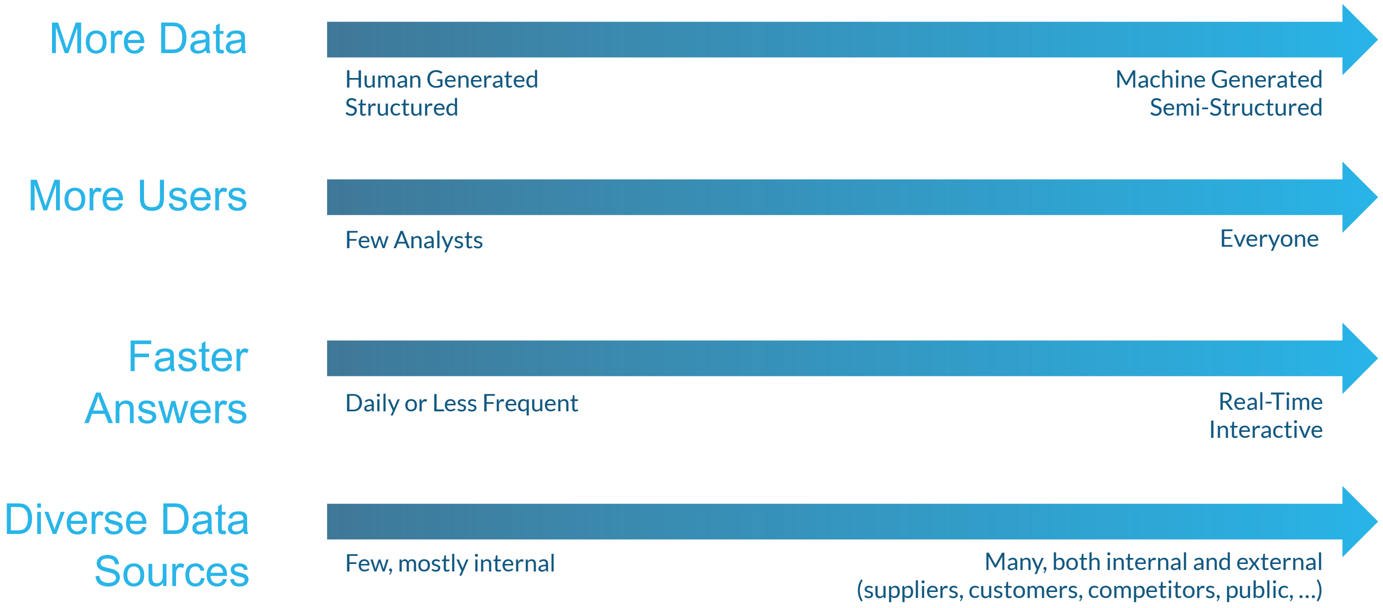
MARCIN ZUKOWSKI – Snowflake Co-Founder

SNOWFLAKE

- Snowflake Inc.
  - Founded 2012, 4000+ employees
  - Sept 2020: Largest software IPO ever
- Cloud Data Platform
  - Data warehousing and other big data tasks
  - SaaS, cloud-native
- Data Cloud
  - Global network for data access

WHY SNOWFLAKE?

No Good Solution to Tackle Modern Data Challenges



More Data

Human Generated
Structured

Machine Generated
Semi-Structured

More Users

Few Analysts

Everyone

Faster Answers

Daily or Less Frequent

Real-Time
Interactive

Diverse Data Sources

Few, mostly internal

Many, both internal and external
(suppliers, customers, competitors, public, …)

WHY CLOUD?

Resources

Hardware and services

Infinitely* elastic

Pay for what you use

SAAS model

Always available

Continuously improving

Hiding complexity

Elastic costs

Ecosystem

Global "meeting place"

Organizations, data and processes

Boundaries purely virtual*

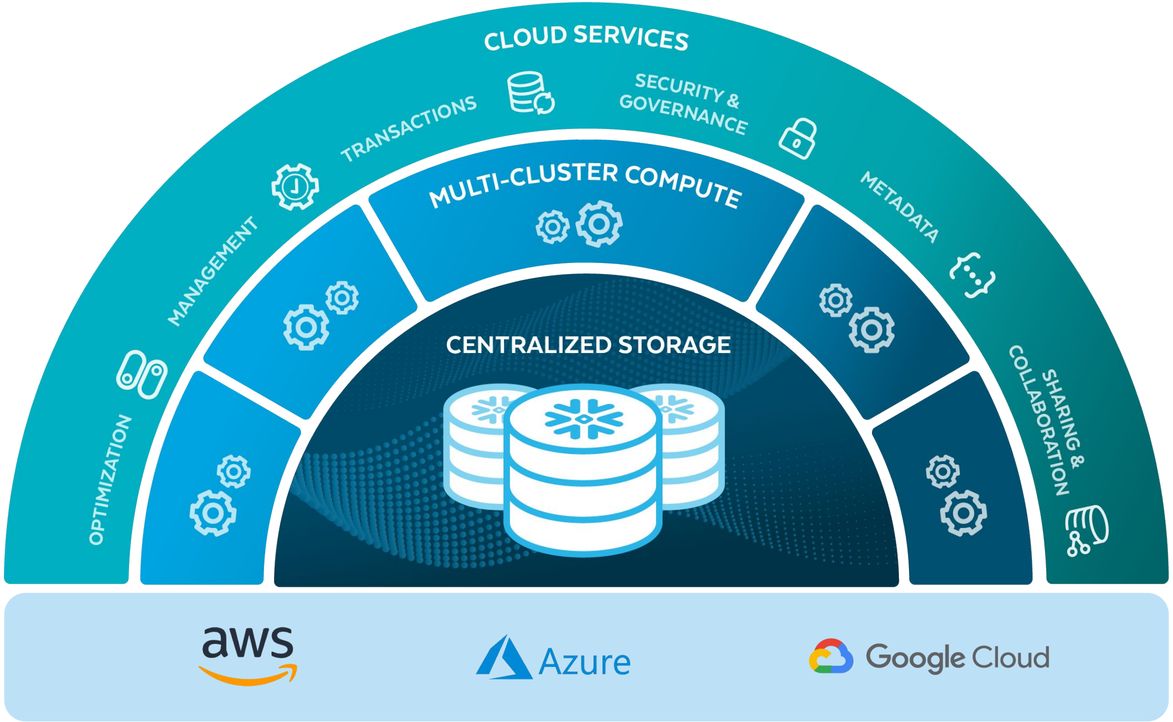
SNOWFLAKE DESIGN MOTIVATION

- Want cost-efficient storage
  → Use S3
- S3 has no updates
  → Immutable data units
- S3 is slow
  → Columnar, data skipping, caching, compression
- Want elastic compute
  → Stateless workers

SNOWFLAKE DESIGN MOTIVATION (cont.)

- Data in S3, elastic workers
  → Need coordination
- S3 is bad for state
  → Need a metadata store
- Barriers are logical
  → Data sharing is possible

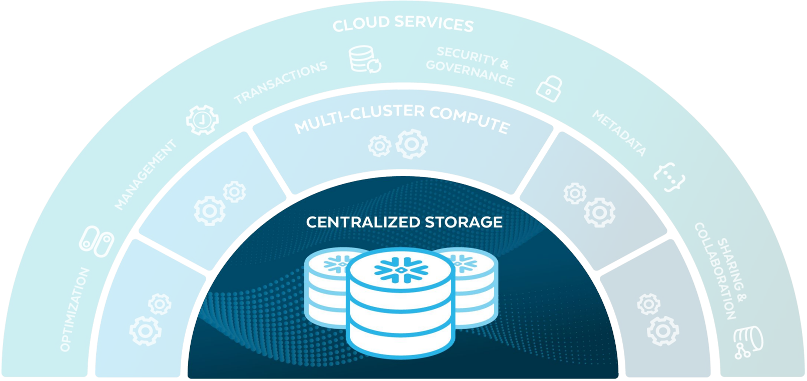
SNOWFLAKE ARCHITECTURE
(one cloud region)



CENTRALIZED STORAGE

- Store all your data: relational, JSON, XML, GEO ...
- Pre-indexed for fast access
- One copy for all users

- Infinitely* elastic
- Cost effective

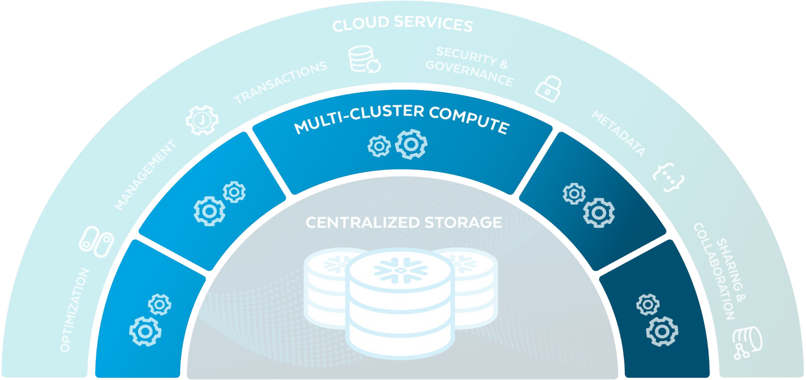


MULTI-CLUSTER COMPUTE

- High performance SQL processing

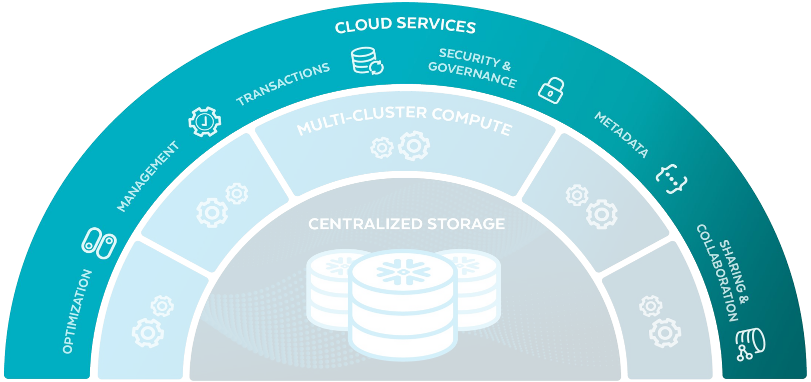
- Private clusters for different users
- Instantly available
- Infinitely* elastic
- Pay for what you use

- **Decoupled from storage**



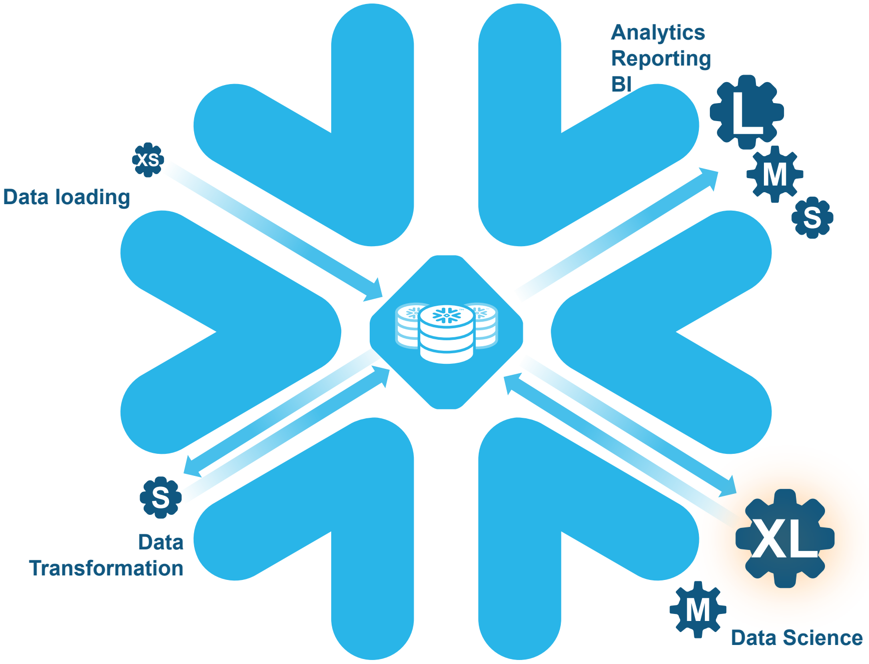
CLOUD SERVICES

- Shared layer for all users
- Pure SAAS experience
  - Always on
  - Frequent (transparent) upgrades
- Fully managed
- Includes persistent state ("big metadata")



SNOWFLAKE FEATURES

MULTI-DIMENSIONAL ELASTICITY



- Elastic scaling for storage
  - Low-cost, fully replicated, secure and resilient
- Elastic scaling for compute
  - Virtual warehouses scale to support workload needs
- Elastic scaling for concurrency
  - Scale concurrency using independent virtual warehouses or with multi-cluster warehouses
- Both up and down

MULTIPLE WORKLOADS

DATA WAREHOUSE

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

DATA ENGINEERING

Streaming Ingest
Tasks
Table Streams
External Functions
Data Pipelines

DATA LAKE

Semi-structured Data
Unstructured Data
External Tables

DATA SCIENCE

Java/Scala/Python
Data Frames

DATA APPLICATIONS

Rest APIs
Real-time

FULLY MANAGED

Infrastructure

Initial Setup

Upgrading

Patching

Capacity Planning

Storage

Security

Physical Design

Partitioning

Indexing

Ordering

Vacuuming

Data Collaboration

Loading

Moving

Transforming

Copying

Securing

Query Tuning

Statistic Collection

Memory Management

Parallelism

Query Plan Hinting

Workload Management

Availability

Setup High availability

Handle Hardware Faults

Manage Backups

FULLY MANAGED

Infrastructure

Initial Setup

Upgrading

Patching

Capacity Planning

Storage

Security

Physical Design

Partitioning

Indexing

Ordering

Vacuuming

Data Collaboration

Loading

Moving

Transforming

Copying

Securing

Query Tuning

Statistic Collection

Memory Management

Parallelism

Query Plan Hinting

Workload Management

Availability

Setup High Availability

Handle Hardware Faults

Manage Backups

Simply load/share data and run queries

SCIENCE FICTION ;)

- Cloning

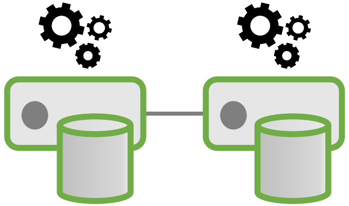
```
CREATE DATABASE my_copy CLONE production;
```

- Time Travel

```
SELECT * FROM users AT (OFFSET => -3600*2)
WHERE id NOT IN (SELECT id FROM users);
```

POWER OF ELASTICITY

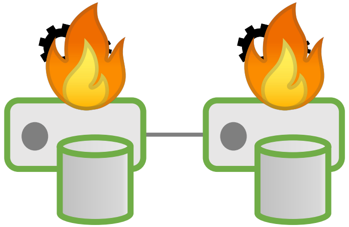
ON PREMISE: SIZED FOR AVERAGE USAGE



Regular day

Users	IT	Finance
😀	😀	😀

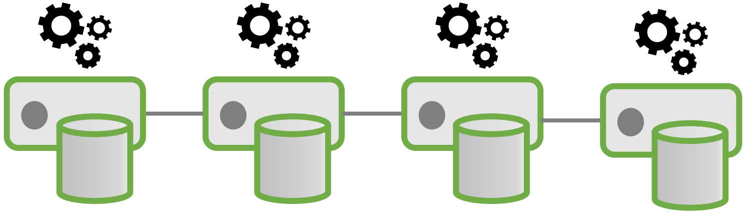
ON PREMISE: SIZED FOR AVERAGE USAGE



Monday morning

Users	IT	Finance
😡	😕	😀

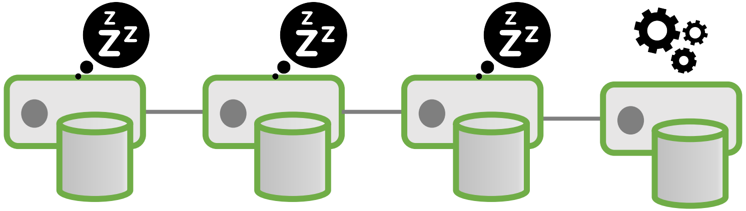
ON PREMISE: SIZED FOR PEAK USAGE



Monday morning

Users	IT	Finance
😀	😀	🙁

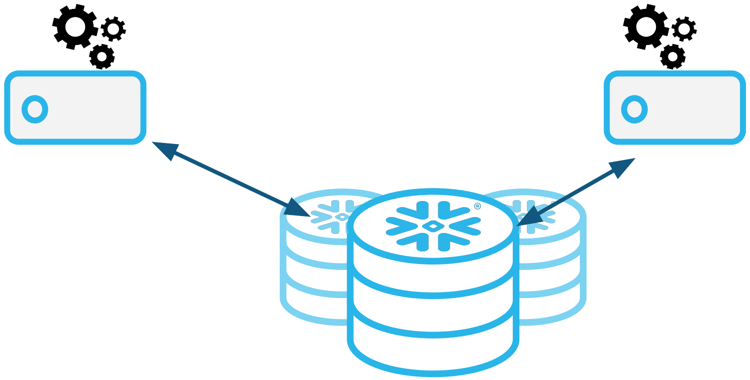
ON PREMISE: SIZED FOR PEAK USAGE



Regular day

Users	IT	Finance
😀	😕	😡

SNOWFLAKE



Regular day

Users

😀

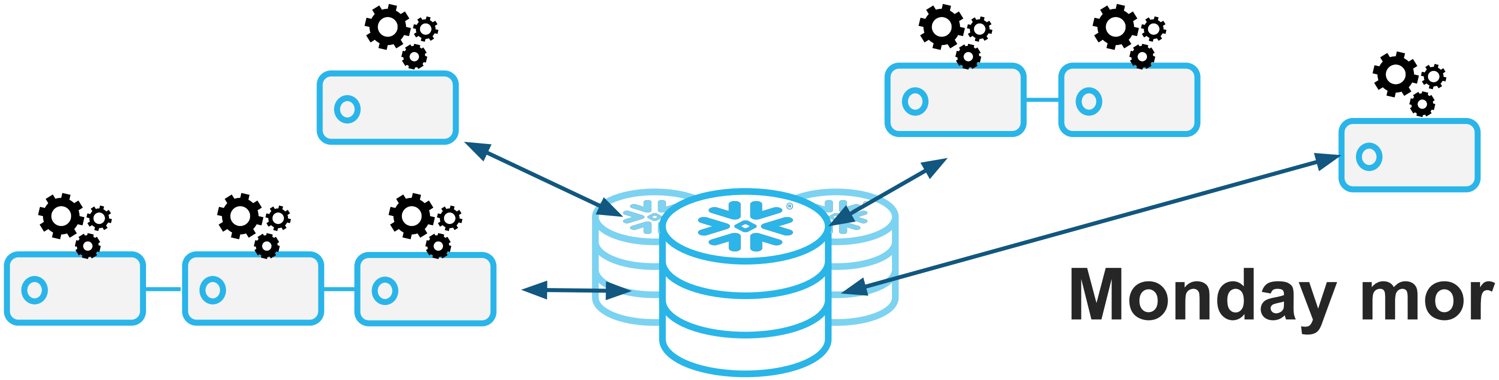
IT

😀

Finance

😀

SNOWFLAKE



Monday morning

Users	IT	Finance
😍	😀	🙁

SNOWFLAKE

Sunday evening

Users	IT	Finance
😴	😀	😍

POWER OF SAAS

CUSTOMER BENEFITS

- Simplicity
  - "It just works" - minimal config
  - Always-on and up-to-date
  - Automate administration tasks
- Pushing technical complexity down
  - Allows previously unachievable systems
- Reduced investment risk
  - Easier to test
  - Pay for what provides value

PROVIDER BENEFITS

- Easier customer acquisition
- Economics of scale
    - Shared costs between customers
    - Simplified maintenance
- Usage-driven development
    - Full insight into customer activities
    - Determine problems and opportunities
    - "Test in production"
- Single "live" system version

ALIGNING INCENTIVES

Usage-based pricing - new model

Performance improvements ?

Snowflake - "Put customer first"

Net revenue retention rate: 178%

CHALLENGES

CLOUD IS A DIFFERENT BEAST

- Performance unpredictability
  - Hardware and services
- Increased failure rates
- Black-box infrastructure
- Multi-cloud challenges

BUILDING ENTERPRISE SAAS IS HARD!

- Need to cover a broad range of "abilities"
  - Stability, availability…
  - Monitoring, manageability, audit…
  - Security, certifications…
- Working on a "live" system
  - Continuous updates, rollbacks
- Handling scale
  - 10x increase every ~2-3 years
- It's hard to make something truly simple to use

THE DATA CLOUD

WHY CLOUD?

Resources

✓ CHECK

Hardware and services

Infinitely* elastic

Pay for what you use

SAAS model

Always available

Continuously improving

Hiding complexity

Elastic costs

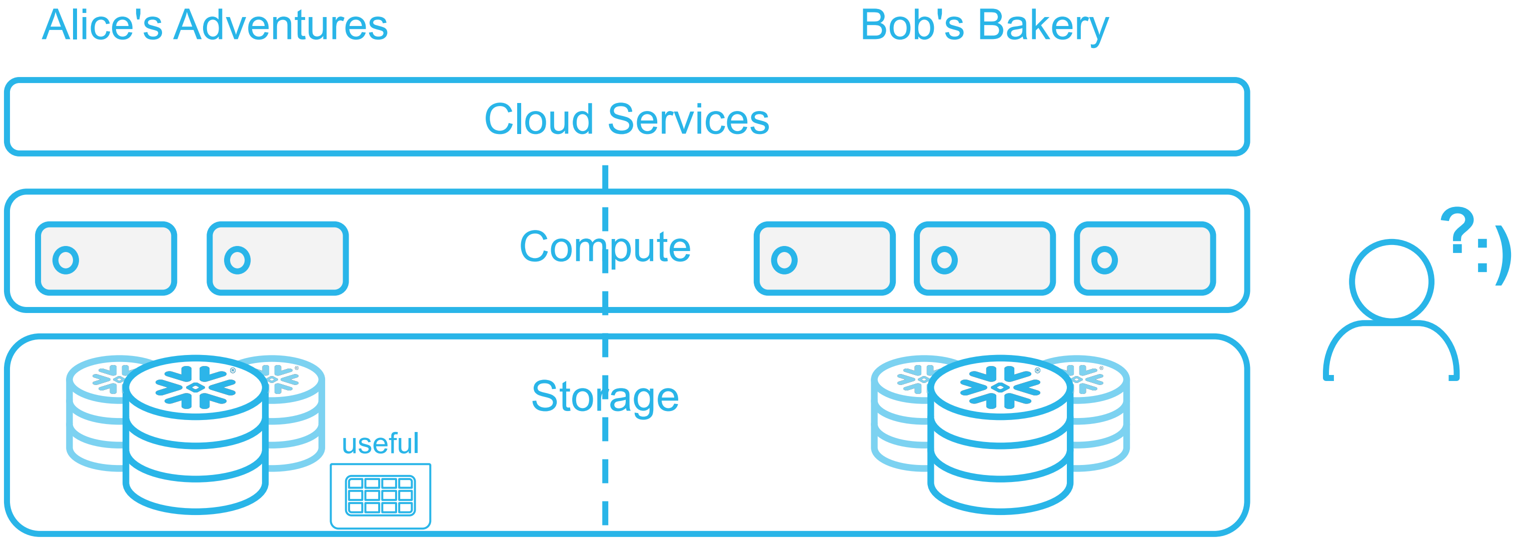
Ecosystem

Global "meeting place"

Organizations, data and processes

Boundaries purely virtual*

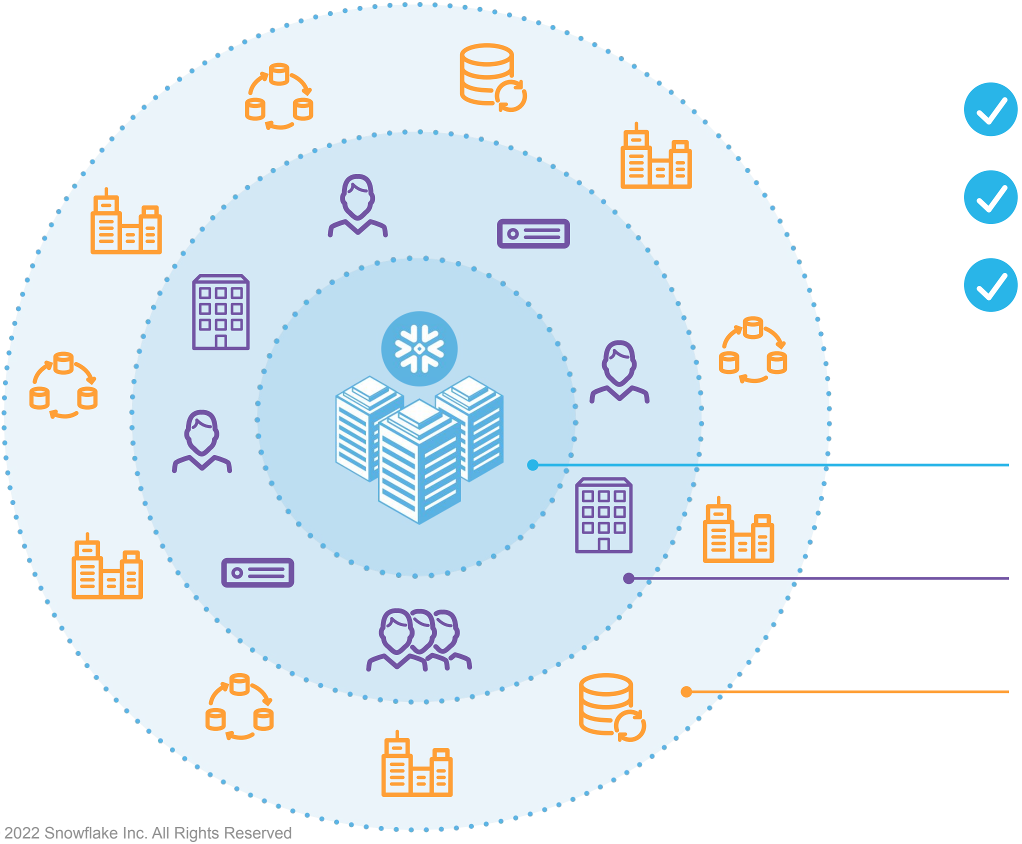
DATA SHARING



```
CREATE SHARE public_data;
GRANT SELECT ON TABLE useful
    TO SHARE public_data;
ALTER SHARE public_data
    ADD ACCOUNTS = BB;
```

```
CREATE DATABASE AA_data
    FROM SHARE AA.public_data;
SELECT * FROM AA_data.useful;
```

ACCESS TO ALL DATA



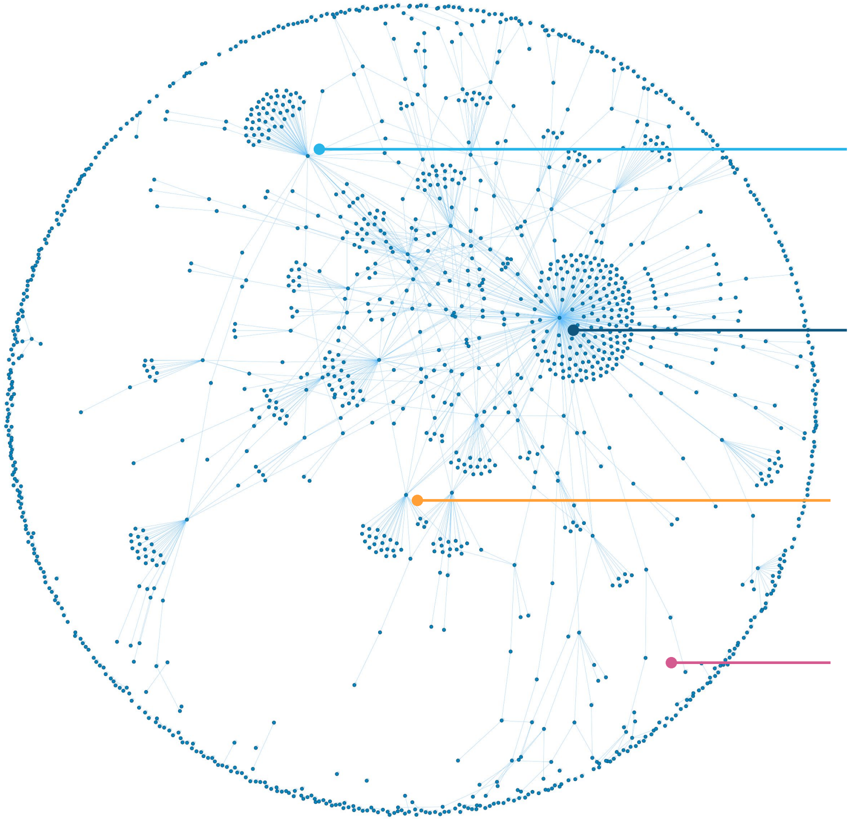
All of Your Organization’s Data, on One Platform

Your Ecosystem - Partners, Suppliers, Customers

Snowflake Data Marketplace - Industry Datasets, Data Services, Applications

- ✓ **Secure**
- ✓ **Nearly unlimited scale**
- ✓ **No Copying or Moving**

COLLABORATION NETWORK TODAY



A software company shares terabytes of data with hundreds of customers

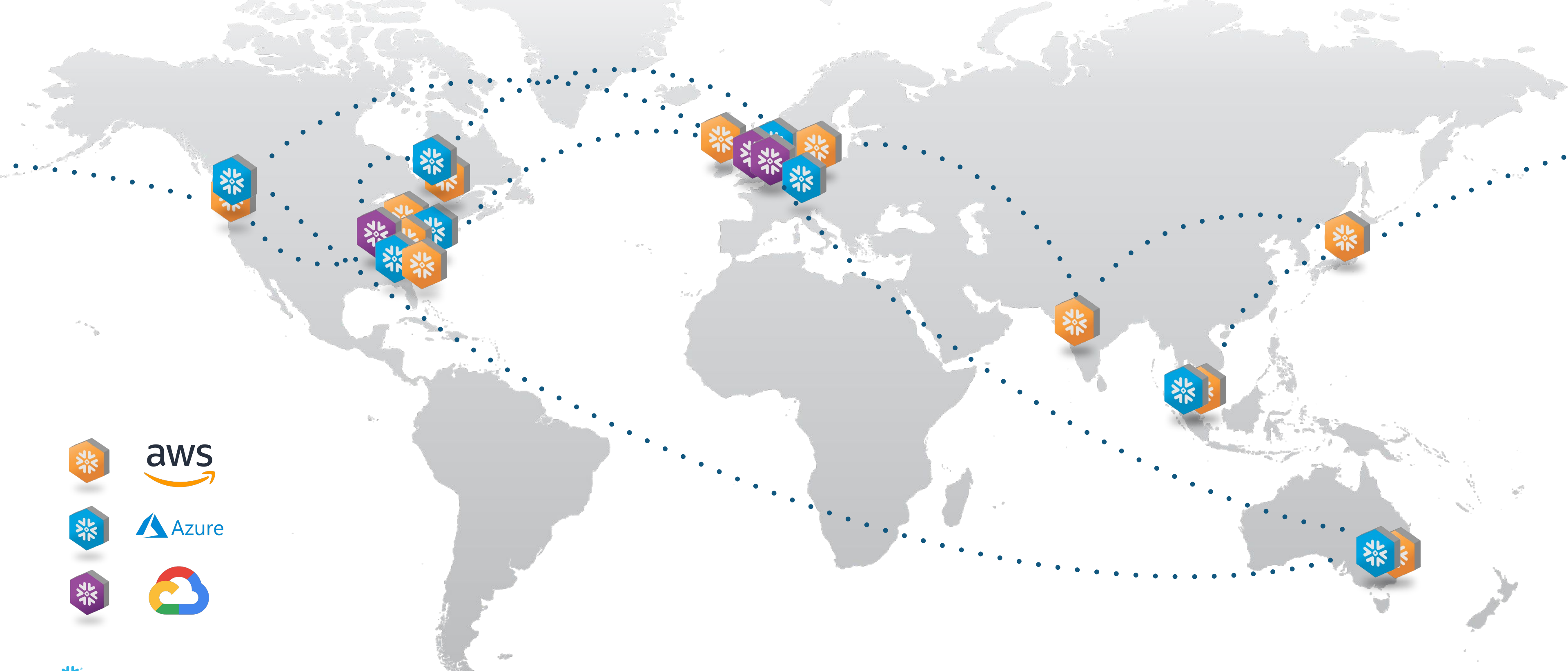
COVID-19 data is available live on Snowflake Data Marketplace from a US State, and other organizations

Today's financial data is available immediately without data pipelines

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

ONE SINGLE DATA CLOUD

23 Data Cloud Regions (10 countries, 3 clouds)



SNOWFLAKE TODAY

All major
cloud vendors

5900+
active customers

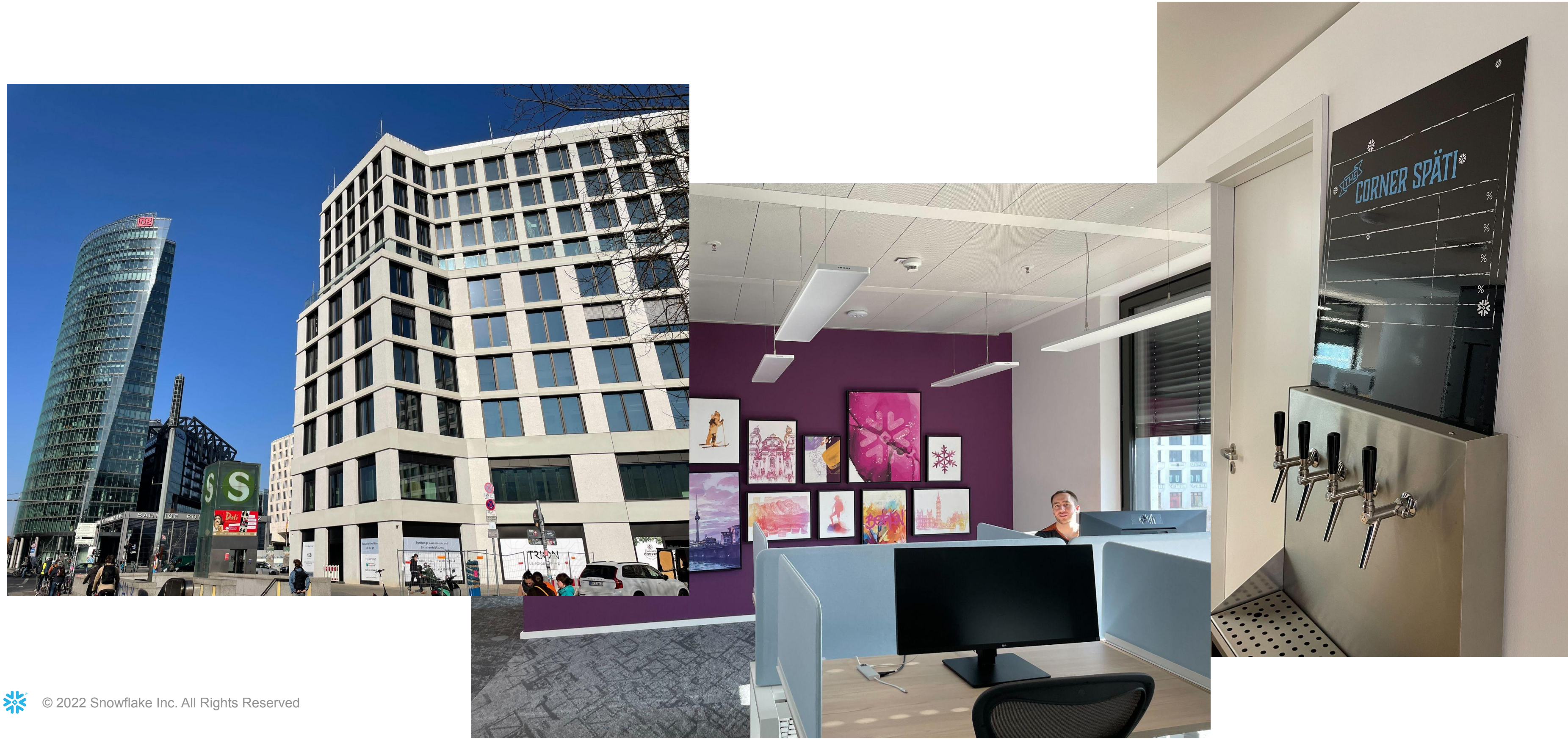
> 1000 Data
Marketplace
listings

100s of PB storage (compressed)
Biggest table ~100TN rows

>1B queries daily

NPS - 68
Industry average - 21

WE'RE IN BERLIN!



CLOUD AND DB RESEARCH

CHALLENGES

- Many research areas hard to apply in cloud
    - Modern / exotic hardware
    - Software plugins / accelerators
    - Open source ?
- Hard to build a user-ready SAAS product
    - A lot of non-DB complexity
    - Making things "just work"

MORE CHALLENGES

- Reduced visibility into platform
  - Infastructure - black box
  - More levels of abstractions
- Reduced visibility into customers

- Databases are getting commoditized
  - Aurora Serverless V2 - enough for most

OPPORTUNITIES

- Unique platform properties
  - Heterogenous resources (hard- and software)
  - Embrace infinite elasticity
  - Optimize for cost

- Cloud vendors evolve fast
  - Software: EC2 → Containers → Lambda
  - Hardware: x86 → ARM (→ RISC-V ?)

MORE OPPORTUNITIES

- Energy efficiency critical
- Databases at global scale
- Making data collaboration better
- Cloud as a shared platform for researchers?

CONCLUSION

Cloud is the new normal

We all need to adapt (a bit)

Opportunities are endless

snowflake®

THANK YOU!