snowflake

Data Warehousing Reinvented for the Cloud World

Benoit Dageville

Snowflake?

Startup founded in August 2012 with the ambition to build a data warehouse for the cloud

- Located downtown San Mateo
- 90+ employees, about 40 engineers, 25 core developers (and hiring…)
- GA in June 2015
- Snowflake service currently hosted on Amazon public cloud
- Stores multiple petabyte of raw data
- Runs million SQL statements every day

Why Cloud?

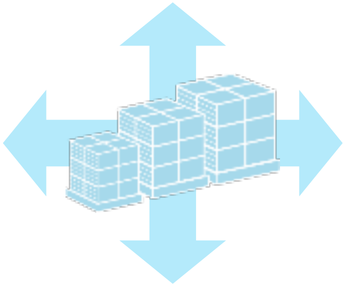
- Cloud is an amazing platform for building distributed systems
  - Access to unlimited compute power and storage capacity
    - Provision a fleet of servers in few minutes
    - Blob storage service allows to cheaply store petabyte data
    - Pay what you use model
  - Provide multi-datacenter availability
  - Efficient access from anywhere

- Data democratization
  - Enables Software as a Service
  - Self-service, no need for complex IT organization and infrastructure
  - Virtuous circle measured in days, not years

Our Vision for a Cloud Data Warehouse

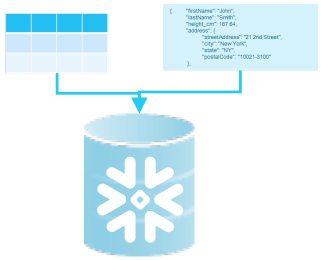
Data warehouse as a service

No infrastructure to manage, no knobs to tune



Multidimensional elasticity

On-demand scalability data, queries, users



```
{       "firstName": "John",
        "lastName": "Smith",
        "height_cm": 167.64,
        "address": {
                "streetAddress": "21 2nd Street",
                "city": "New York",
                "state": "NY",
                "postalCode": "10021-3100"
        },
```

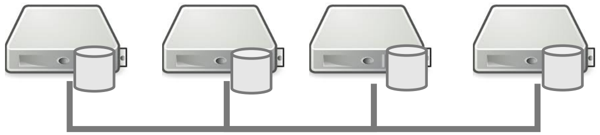
All business data

Native support for structured + semi-structured data

So Many Challenges…

- Multi-tenant service – one single database for the world
  - No scalability bottleneck
  - Resource isolation
- True Elasticity
  - Online resize without any negative impact
  - Provision hundred servers and use them for only an hour
  - Shutoff compute when done
- Extreme availability
  - Resilient to infrastructure failures (node, cluster, full data center)
  - Protect against any type of data loss
  - No downtime for software/hardware upgrades
- Efficient support of schemaless semi-structured data
  - Data pruning, columnar storage, vectorized execution
  - Petabyte volume

Shared-nothing Architecture?

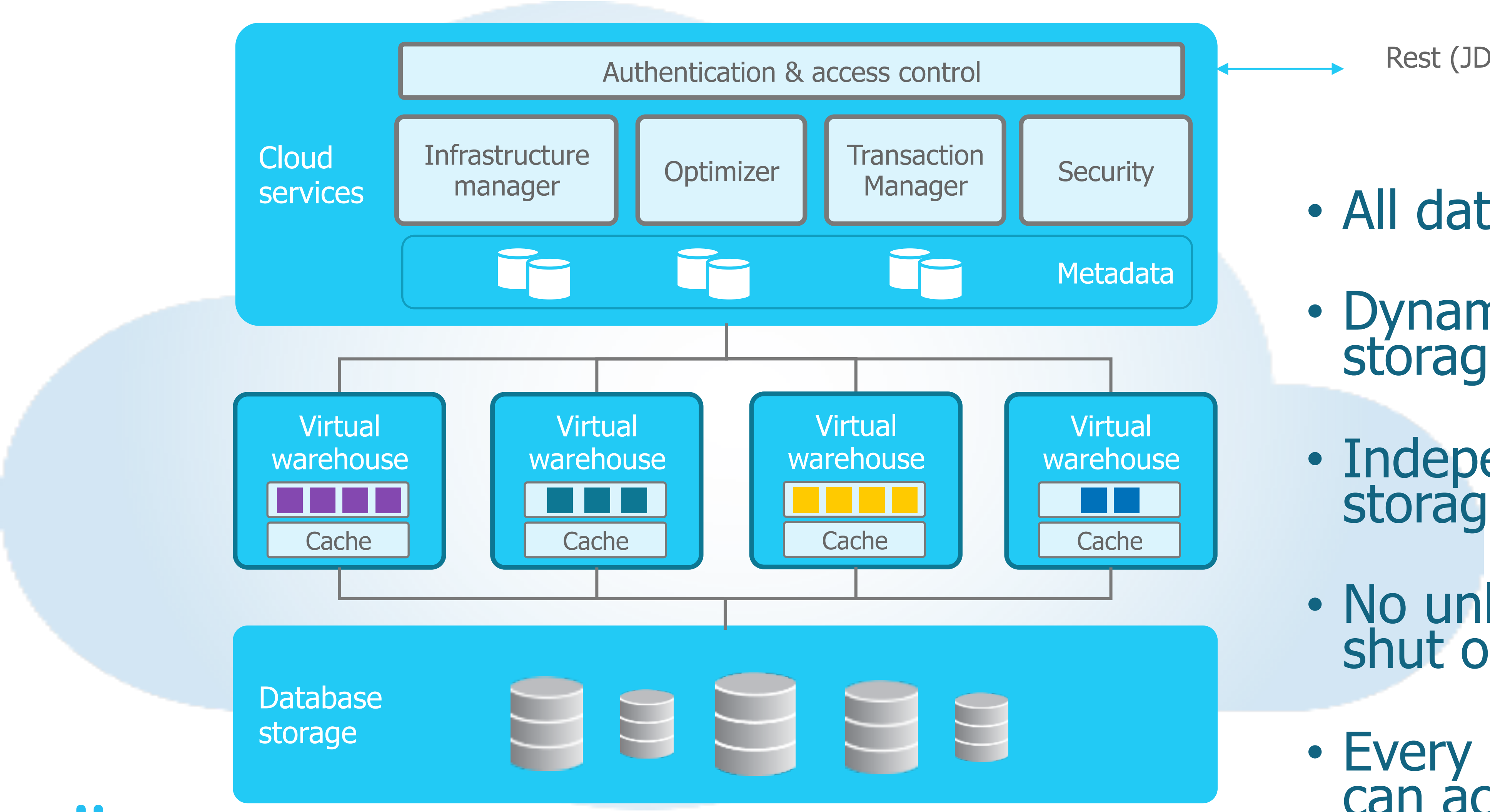


Shared-nothing architecture is not a good fit for cloud

- **Not elastic:** resizing compute cluster requires redistributing data
- **Cannot pay as you go:** shutting off compute cluster requires unloading data
- **Limited scalability:** poor multi-user scalability
- **Limited availability:** simultaneous node failures will cause downtime and data loss

➔ We need a new architecture for cloud

Snowflake
Multi-cluster Shared-data Architecture

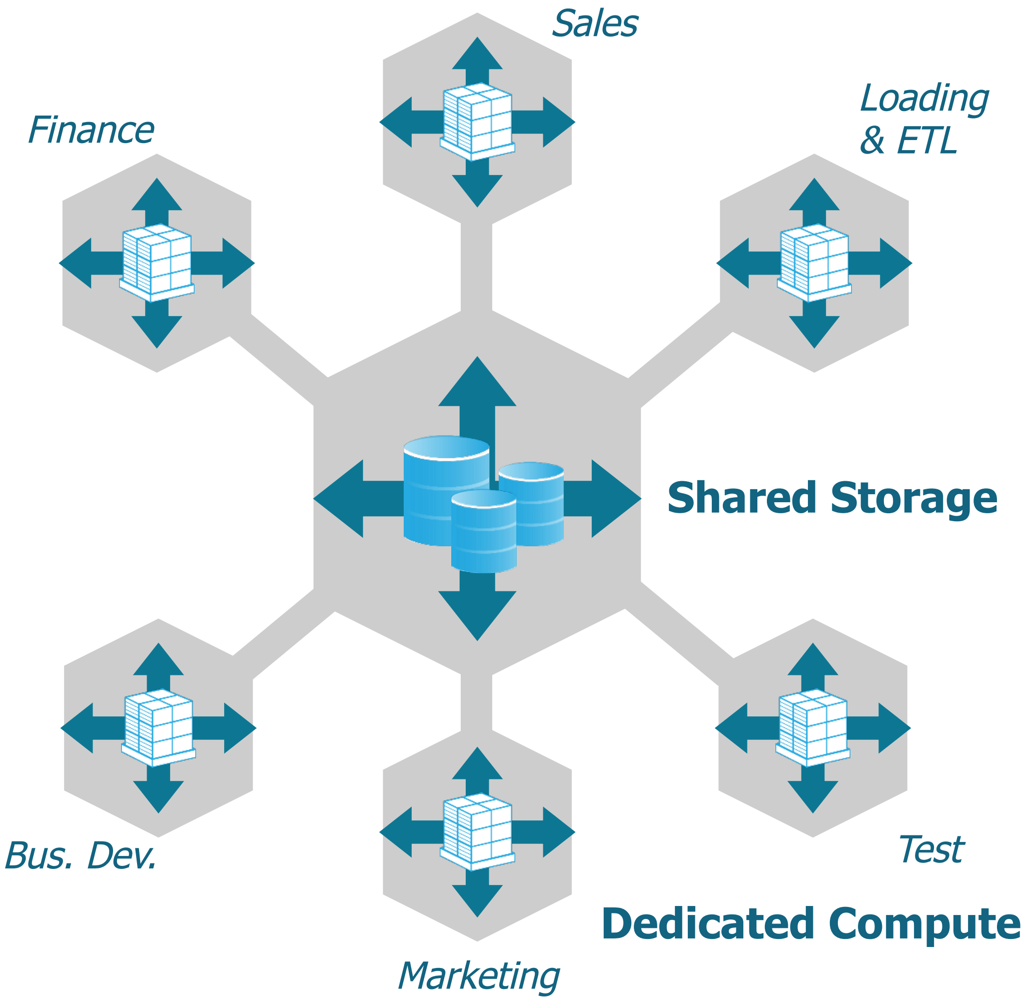


- All data in one place
- Dynamically combine storage and compute
- Independently size storage and compute
- No unload / reload to shut off compute
- Every compute cluster can access any data

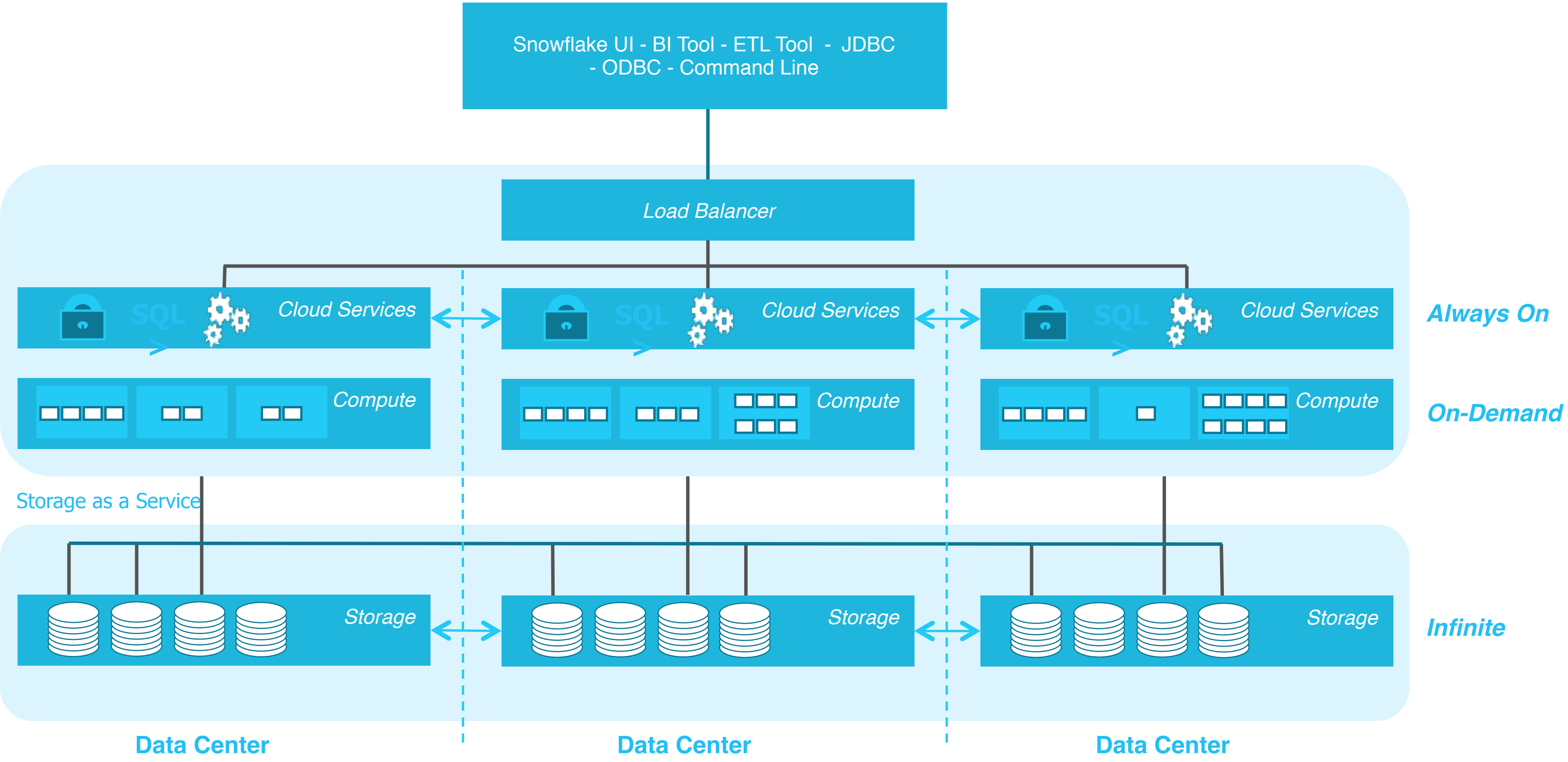
snowflake

Multi-cluster Shared-data Architecture

- Query performance is isolated to each Virtual Warehouse
- Cluster configuration customized for workload
- Unlimited scalability
- Eliminates need to physically move and copy data
  - Data Marts, Cubes, and Test/Dev
- Enables billing by department

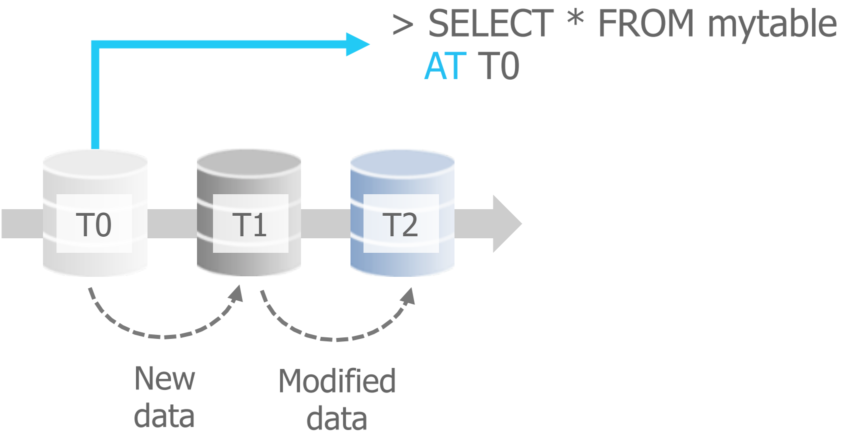


Extreme Availability

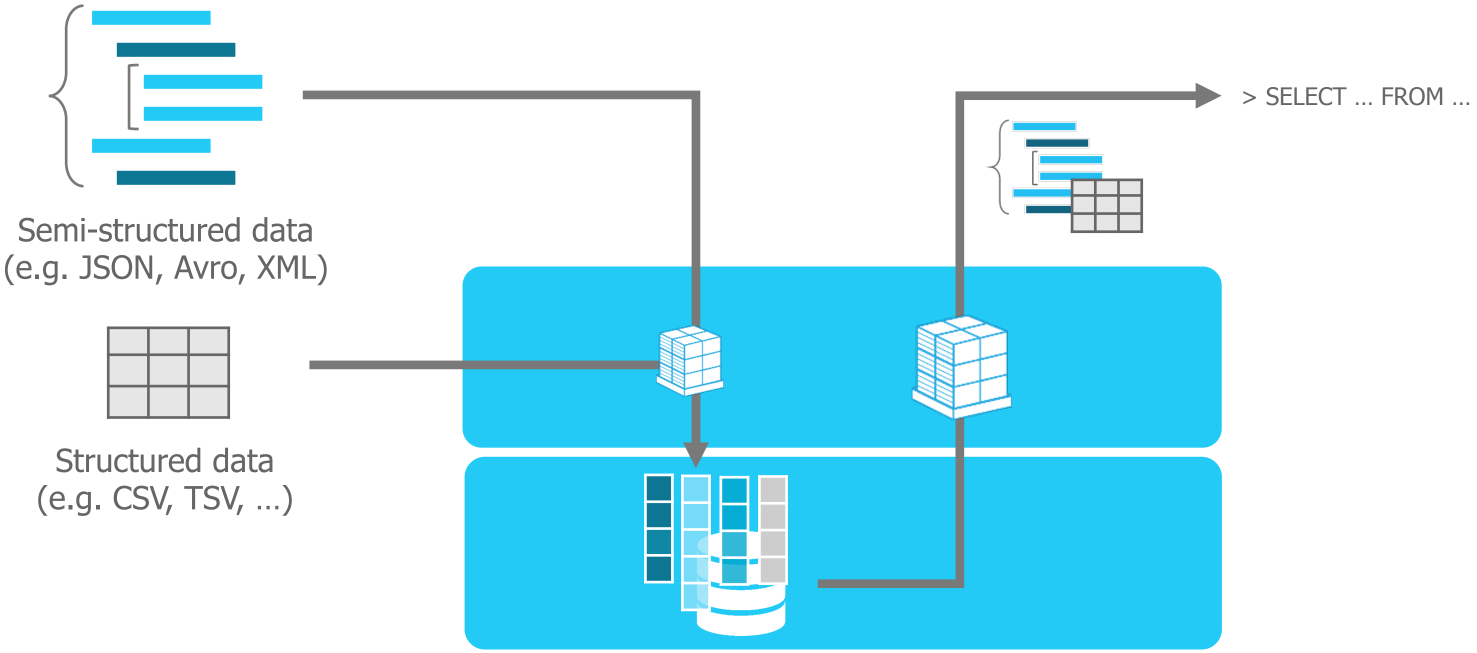


“Time travel” data recovery

- Previous versions of data automatically retained
  - Retention period selected by customer
- Accessed via SQL extensions
  - UNDROP recovers from accidental deletion
  - SELECT AT for point-in-time selection
  - CLONE AT to recreate past versions



Relational database extended to semi-structured data



Storage optimization

- Transparent discovery and storage optimization of repeated elements

Query optimization

- Full database optimization for queries on semi-structured data

Q & A