



10 Best Practices for Data Engineers

Clarke Patterson | Head of Product Marketing

A poll...

I've been a data engineer for:

1. Just starting
2. Less than a year
3. More than a year
4. I'm an expert!



What does a data engineer do, exactly?

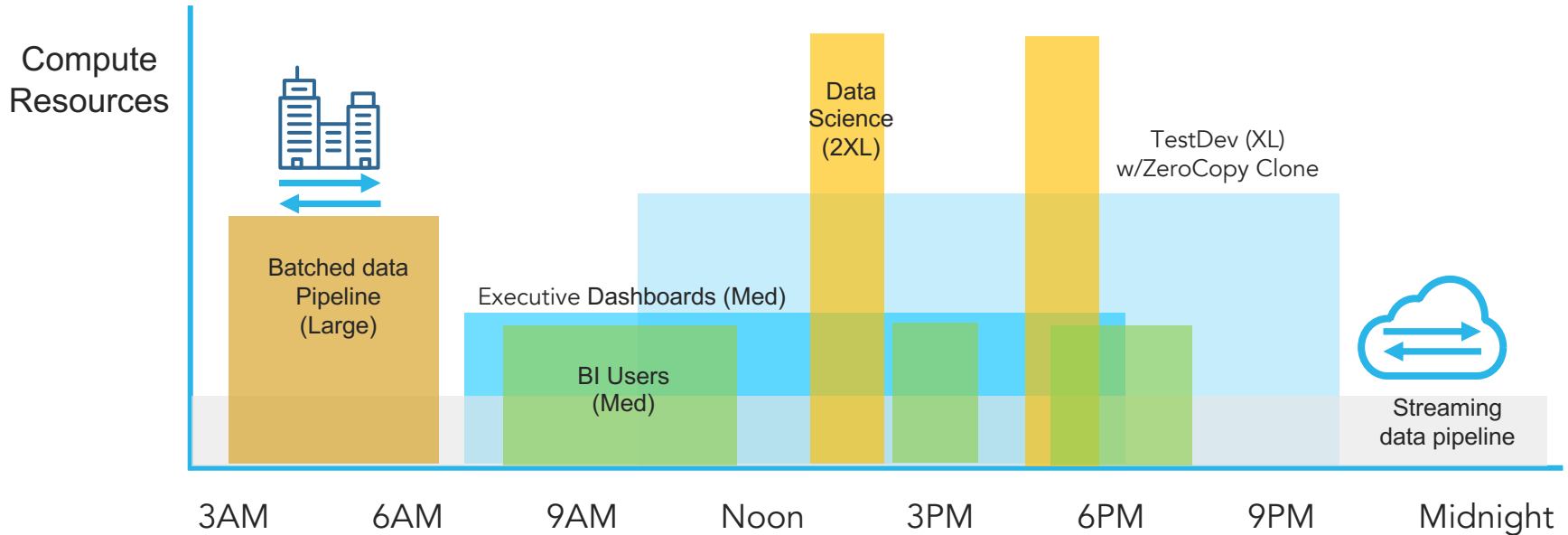
- **Build data pipelines** to collect data and **move it into storage**;
- **Prepare the data** as part of an ETL or ELT process;
- **Stitch the data together** with scripting languages;
- Work with the DBA to **construct data stores**;
- Ensure the data is **ready for use**;
- Use frameworks and microservices to **serve data**.



#1:
**Enable your pipeline to
handle concurrent
workloads**

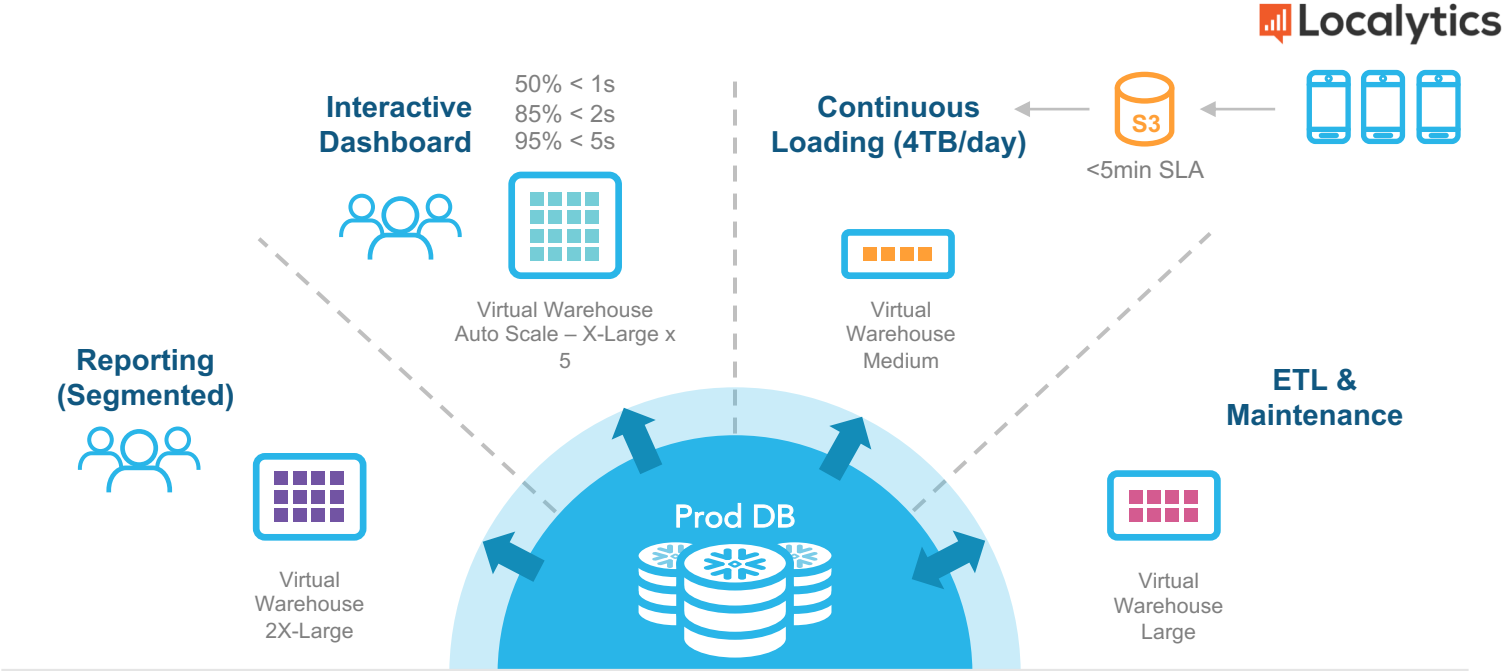


OPERATE ISOLATED WORKLOADS, CONCURRENTLY



Customize for business needs
No contention – No data copying required
Any number of workloads

REAL-WORLD USE CASE



4 trillion rows
3+ petabyte raw data
8x compression ratio
25M micro partitions

#2:
**Tap into existing skills to
get the job done**



You have a choice...



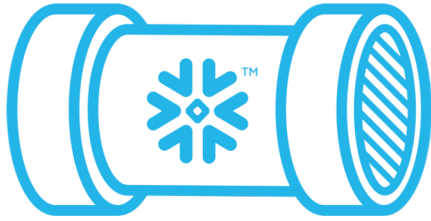
> SQL



#3:
**Use data streaming
instead of batch
ingestion****



Automate ingest when and where possible



Snowpipe



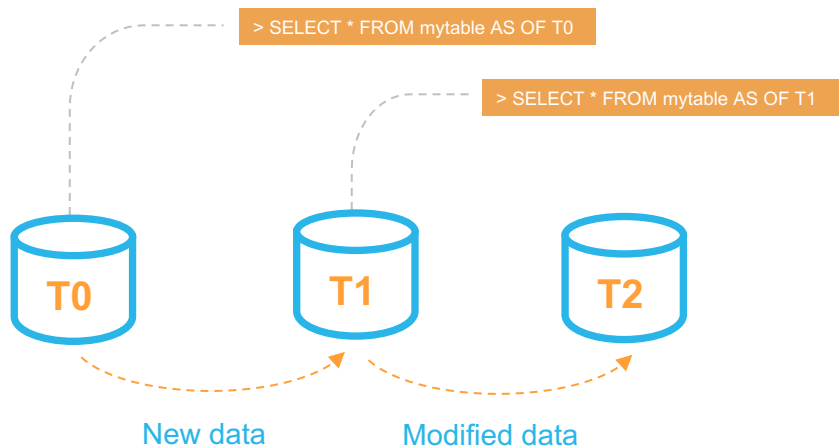
kafka



#4:
**Streamline &
operationalize pipeline
development**



“TIME TRAVEL” FOR DATA



Previous versions of data automatically retained

Retention period selected by customer

Accessed via SQL extensions

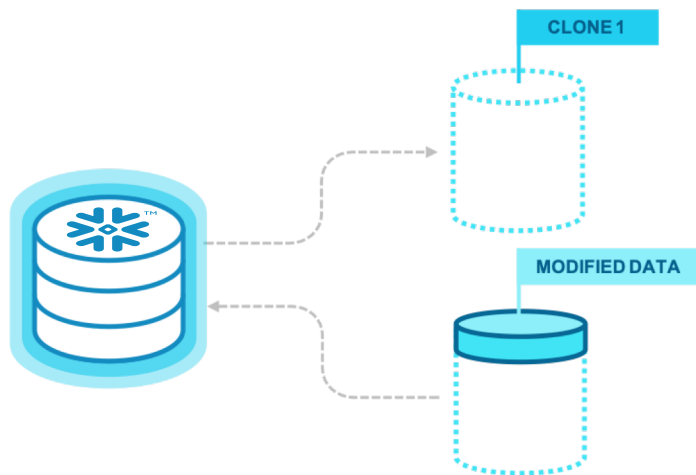
AS OF for selection

CLONE to recreate

UNDROP recovers from accidental deletion



ZERO-COPY DATA CLONING



Instant data cloning operations

Databases, schema, tables, etc

Metadata-only operation

Modified data stored as new blocks

Unmodified data stored only once

No data copying required, no cost!

Instant test/dev environments

Test code on your entire production dataset

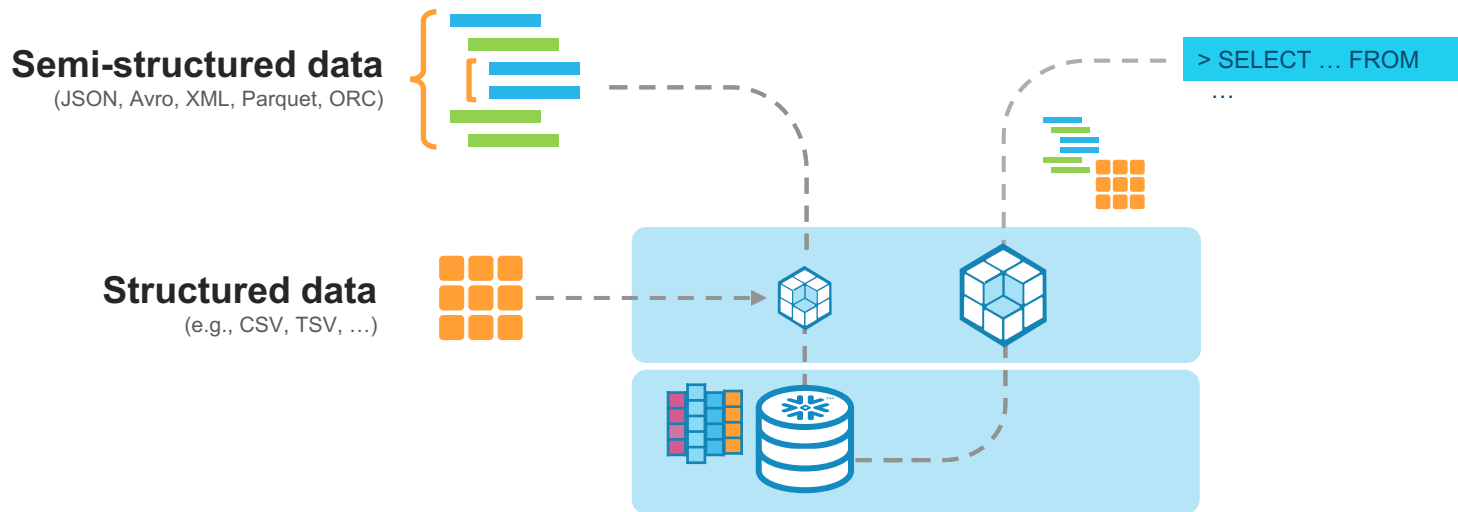
Swap tables into production when ready



#5:
**Invest in tools with built-
in connectivity**



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



Platform



BI/Analytics

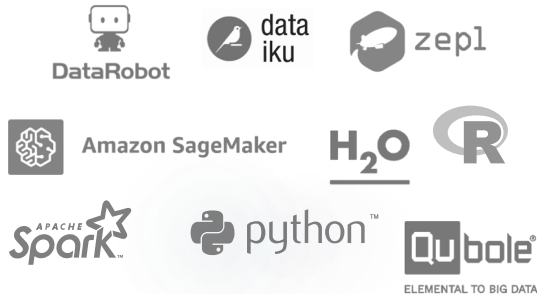


ETL



EVER EXPANDING ECOSYSTEM

Data Science



Services



A poll...

What data do you primarily work with today?

1. Structured data
2. JSON
3. XML
4. AVRO
5. All of the above!



#6: Incorporate extensibility



Embrace each others differences!



#7: Enable data sharing in your pipelines



TRADITIONAL METHODS FOR SHARING DATA

1 Data Transfer



- Data extraction, transformation (for providers and consumers)
- Loading into FTP, Cloud Bucket
- Repeat for every update

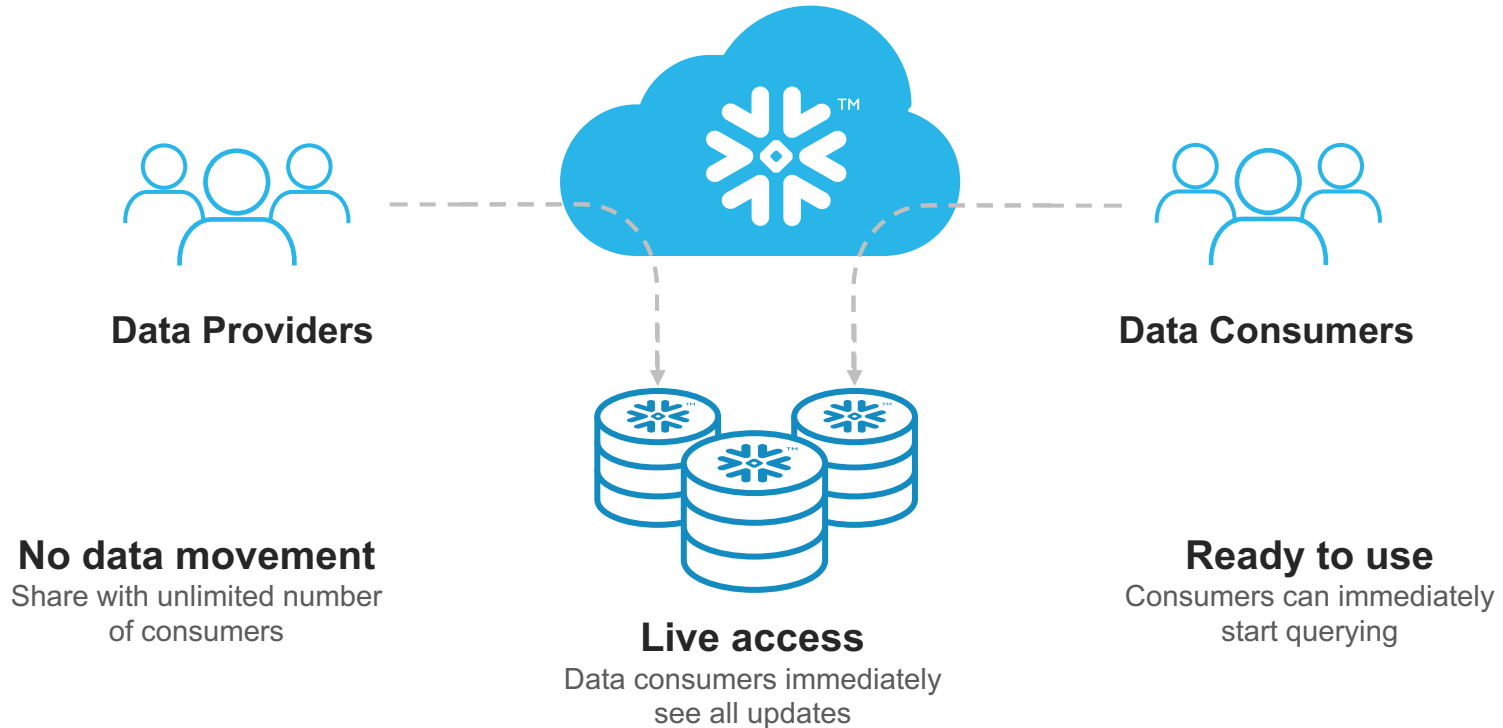
2 API

- Development
- Maintenance
- Support

-
- Time Intensive
 - Sub-par User Experience
 - Costly
 - Raises Security Concerns
 - Limited scalability



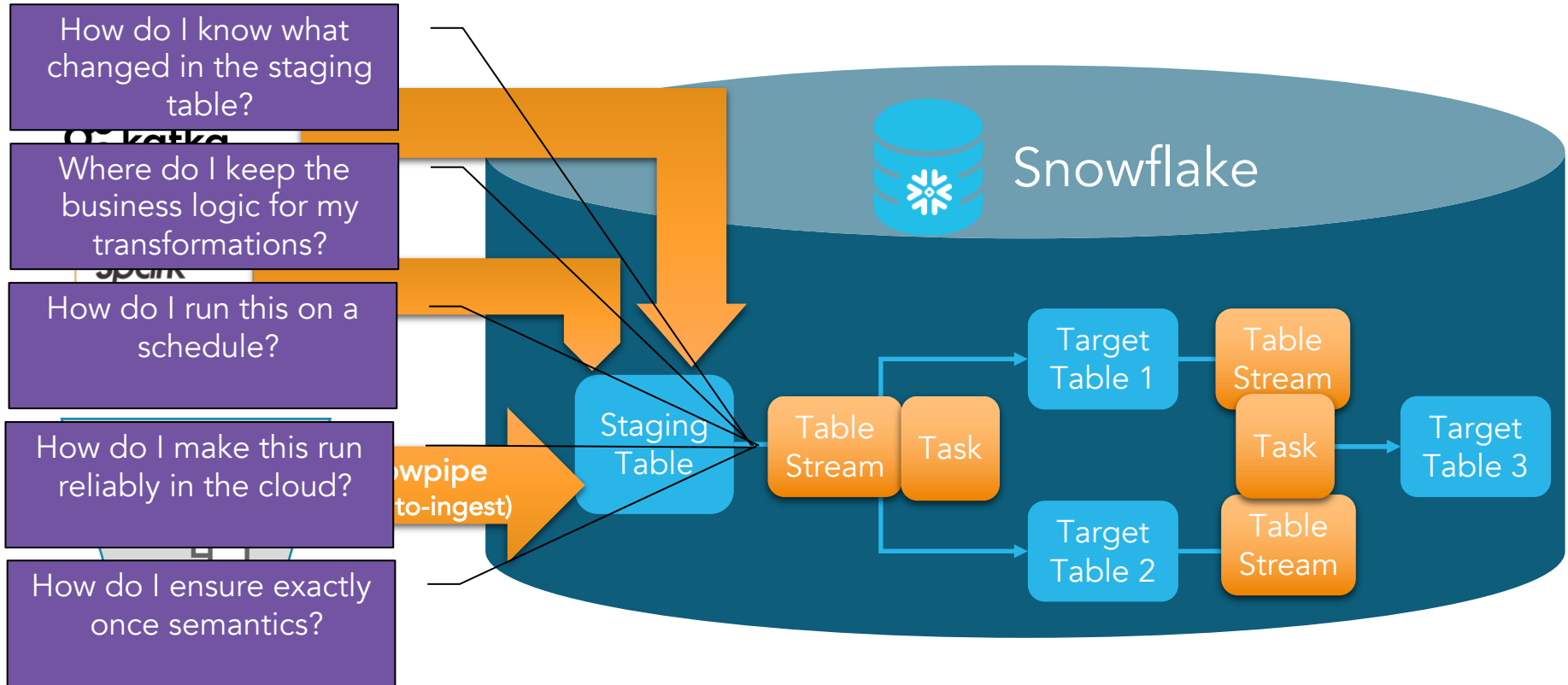
A BETTER WAY TO SHARE DATA



#8: Choose the right tool for data wrangling

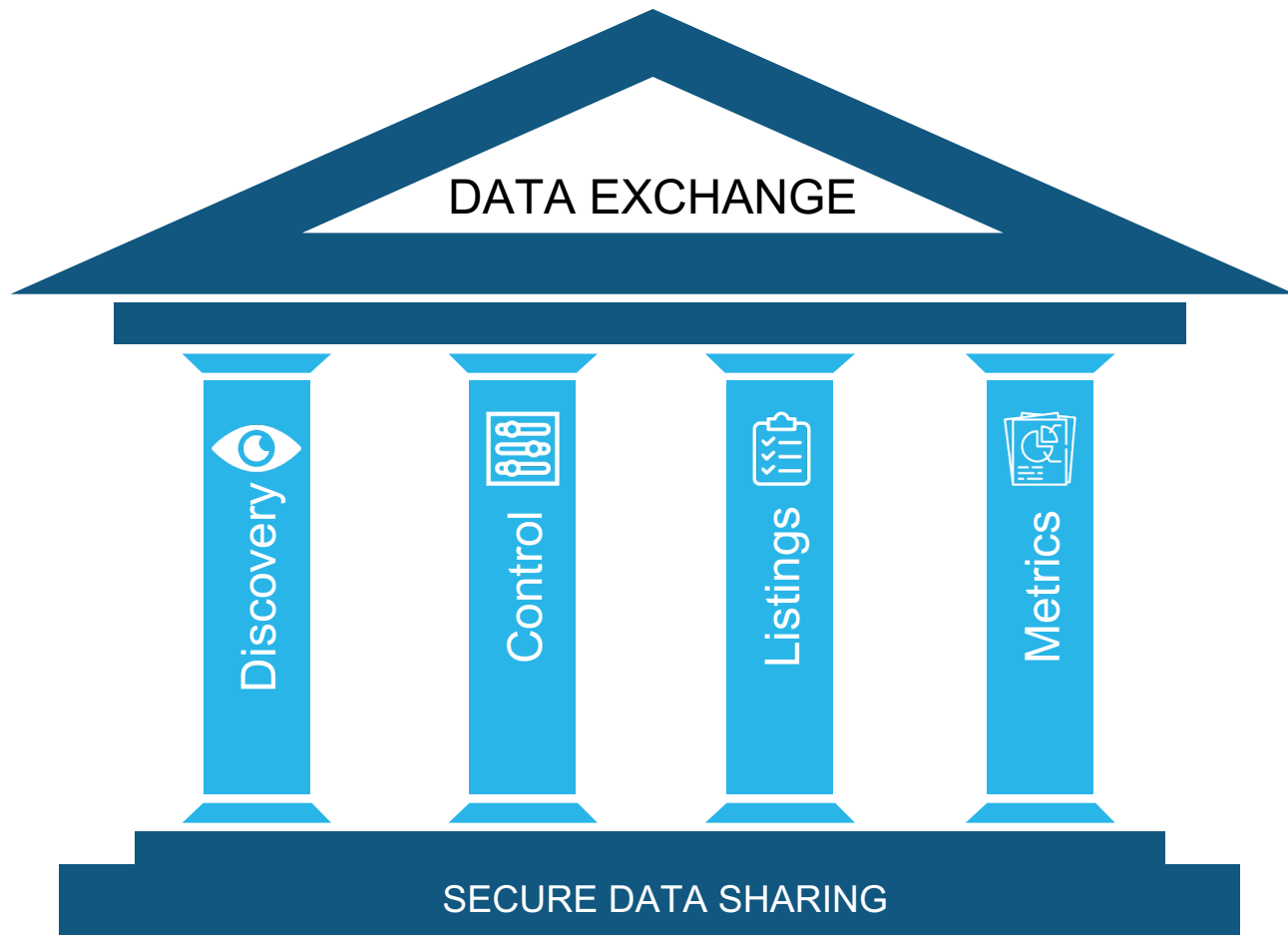


CONTINUOUS DATA PIPELINES



#9:
**Build data cataloging into
your engineering
strategy**

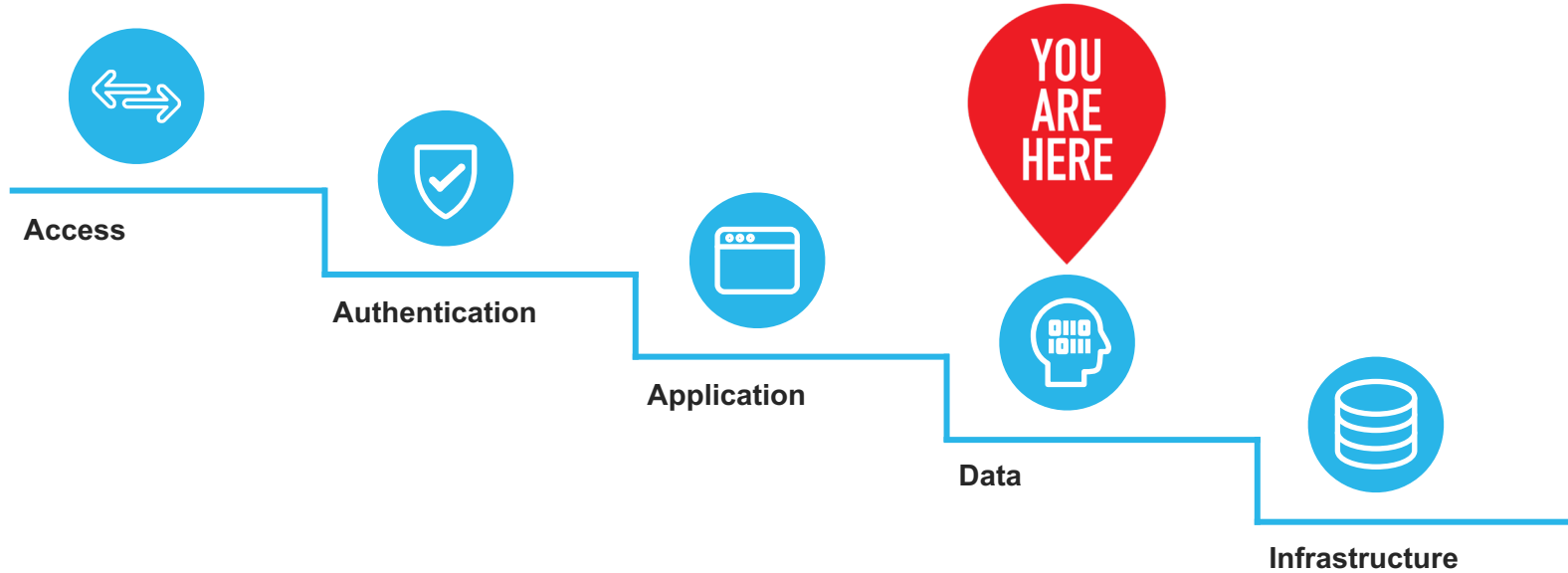




#10:
**Rely on data owners to
set security policy**



SECURITY IS A TEAM SPORT



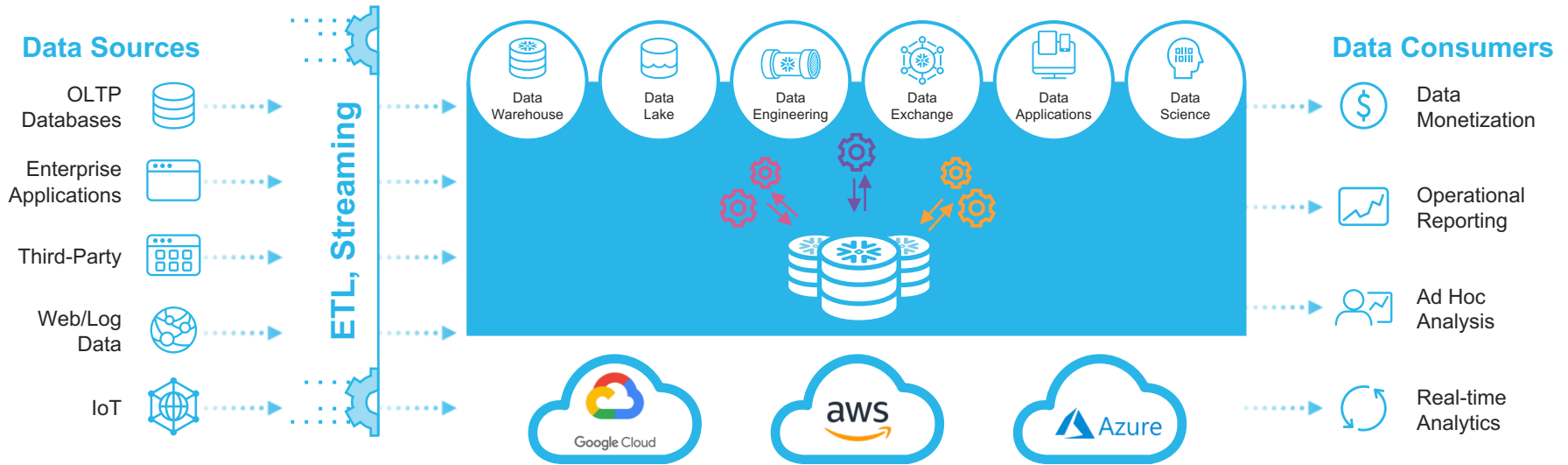
How does Snowflake help?



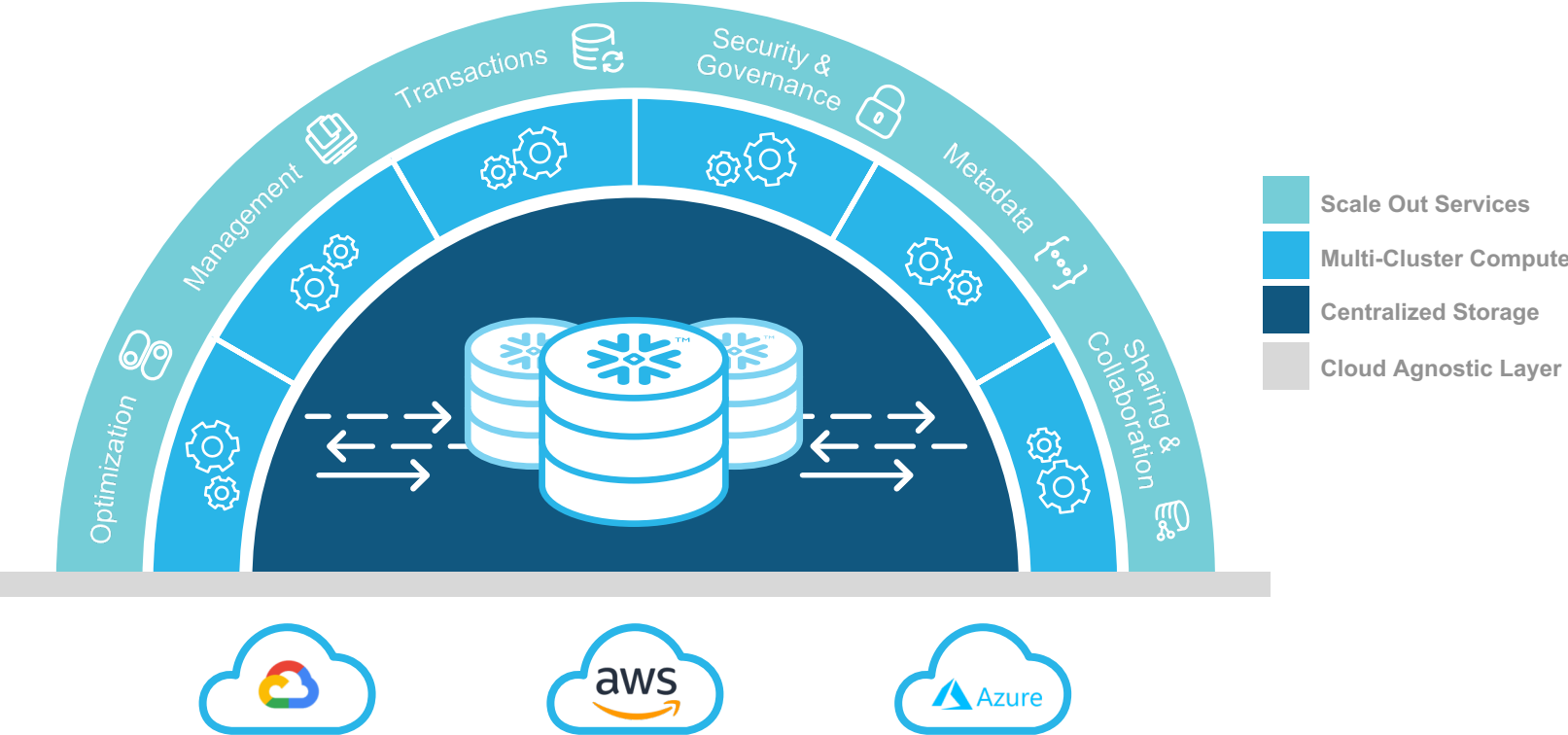
MODERN DATA ARCHITECTURE WITH SNOWFLAKE



MODERN DATA ARCHITECTURE WITH SNOWFLAKE



SNOWFLAKE ARCHITECTURE



SNOWFLAKE DATA EXCHANGE

The screenshot displays the Snowflake Data Exchange interface. At the top, a navigation bar includes icons for Databases, Shares, Data Exchange (active), Warehouses, Worksheets, History, Account, Partner Connect, Help, and Notifications. Below this is a blue banner with the text "DATA EXCHANGE THE WORLD'S DATA + YOURS" and icons for INSTANT, LIVE, and SECURE data exchange methods.

On the left, a "Categories" sidebar lists: All, Business, Cyber, Demographics, Financial, Government, Marketing, Transportation, and Weather. The main content area is for the "BEIERGROUP" user, featuring a search bar and a "Data Exchange" section. Under "Providers", there are tabs for "My Memberships" and "Explore", followed by a row of provider logos: Experian, QL2, Heap, Wunderman, Spring Serve, SafeGraph, and Int Sight. The "Highlights" section has tabs for "Recently Added" and "Highest Rated", and displays three featured data products:

- Heap**: Web & Mobile Behavioral Data. Description: Heap automatically captures every user action without manual event tagging. Tags: Daily update, Human Resource.
- Envirotics**: NA Demographic, Wealth and Household Spending Data. Description: Functions to do Market Basket Analysis (MBA) to understand what items are bought together by customers. Tags: Quarterly update, Demographics.
- Beier Group**: Customer Data Platform. Description: Satisfy advanced business requirements with the most flexible and full-featured customer data platform on the market. Tags: Daily update, Marketing.

SNOWFLAKE DELIVERS PERFORMANCE AT SCALE



650k users
across **3300+**
customers

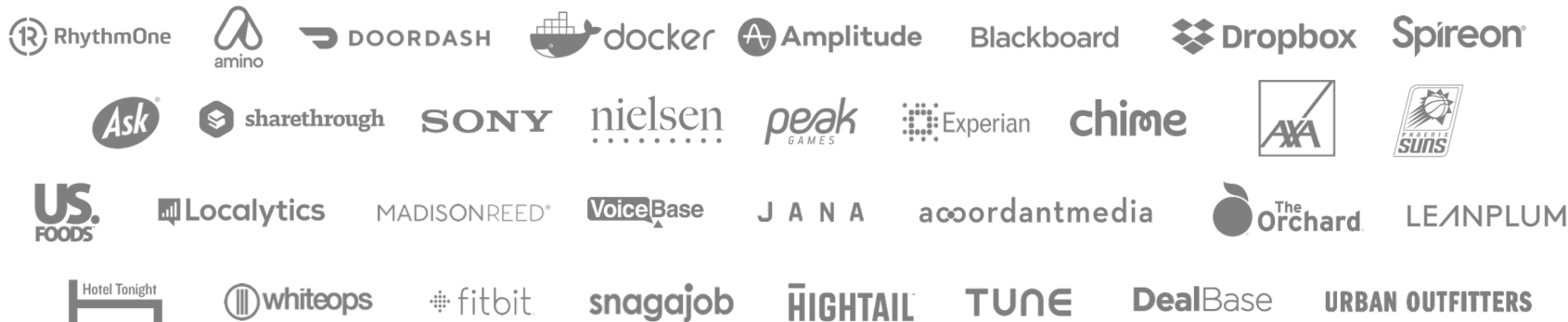
Over **145PB**
under
management

>56PB
scanned
daily

355M jobs
run per
day



PROVEN BY OVER 3400 CUSTOMERS



Questions?





THANK YOU

