TABLEAU CONFERENCE

#TC18

# Real-Time Insights for IoT and Applications

**Max Zuckerman**

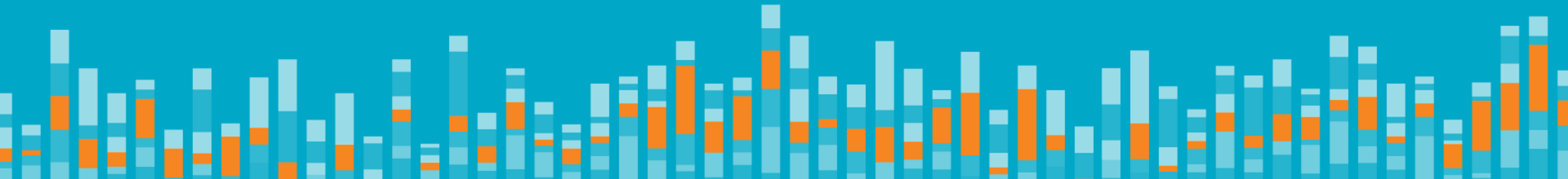
Sr. Director, Solutions Sales

Alooma

TABLEAU CONFERENCE

# First of all, let’s talk cloud.

But I already use the cloud...



# Expand Who Can Use the Data

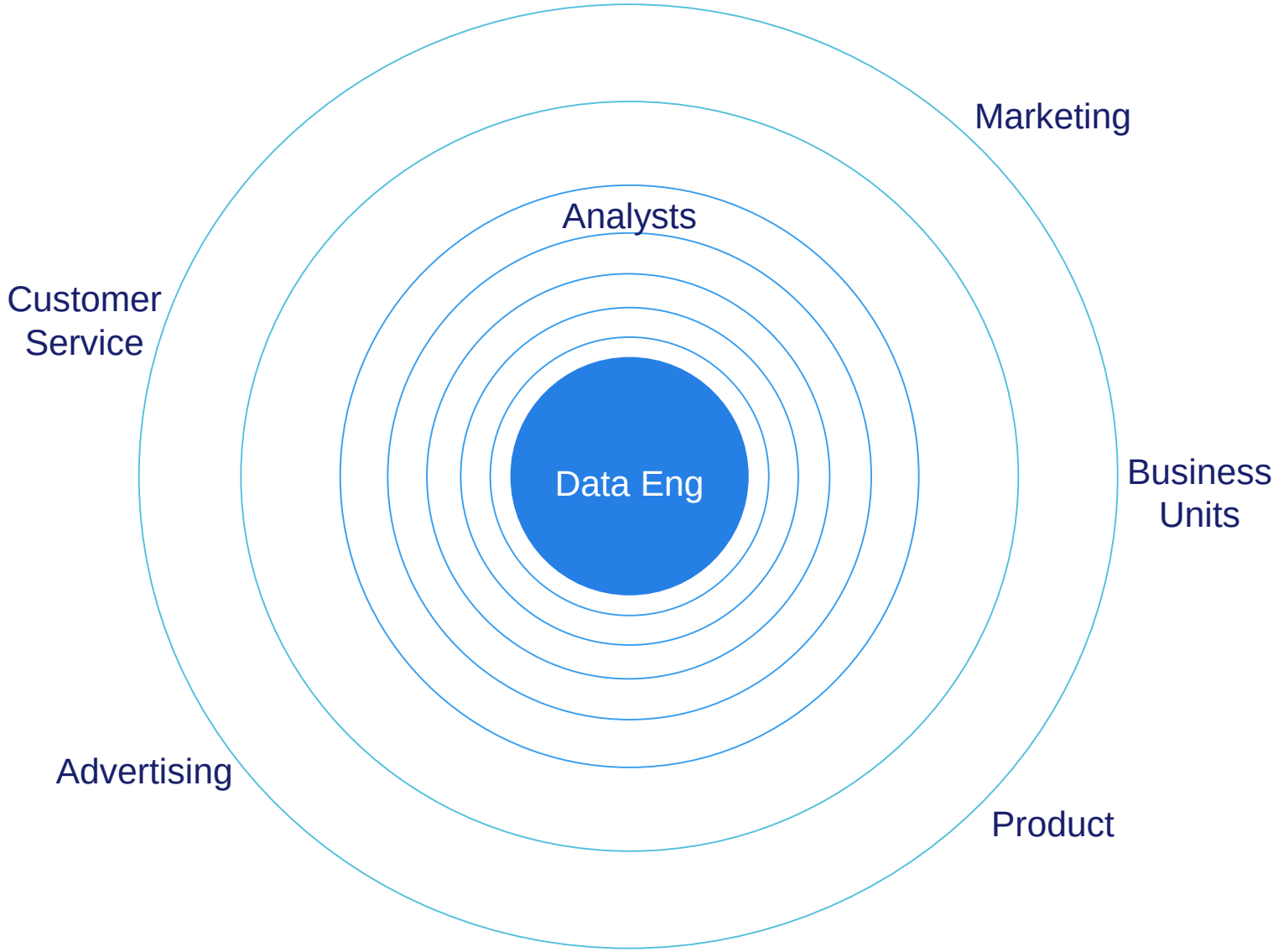
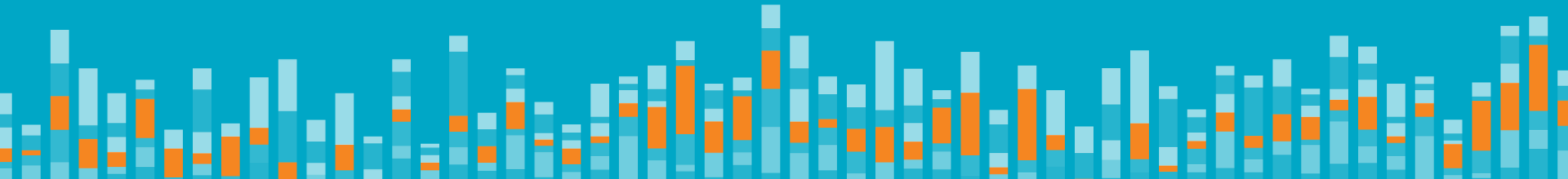


TABLEAU CONFERENCE

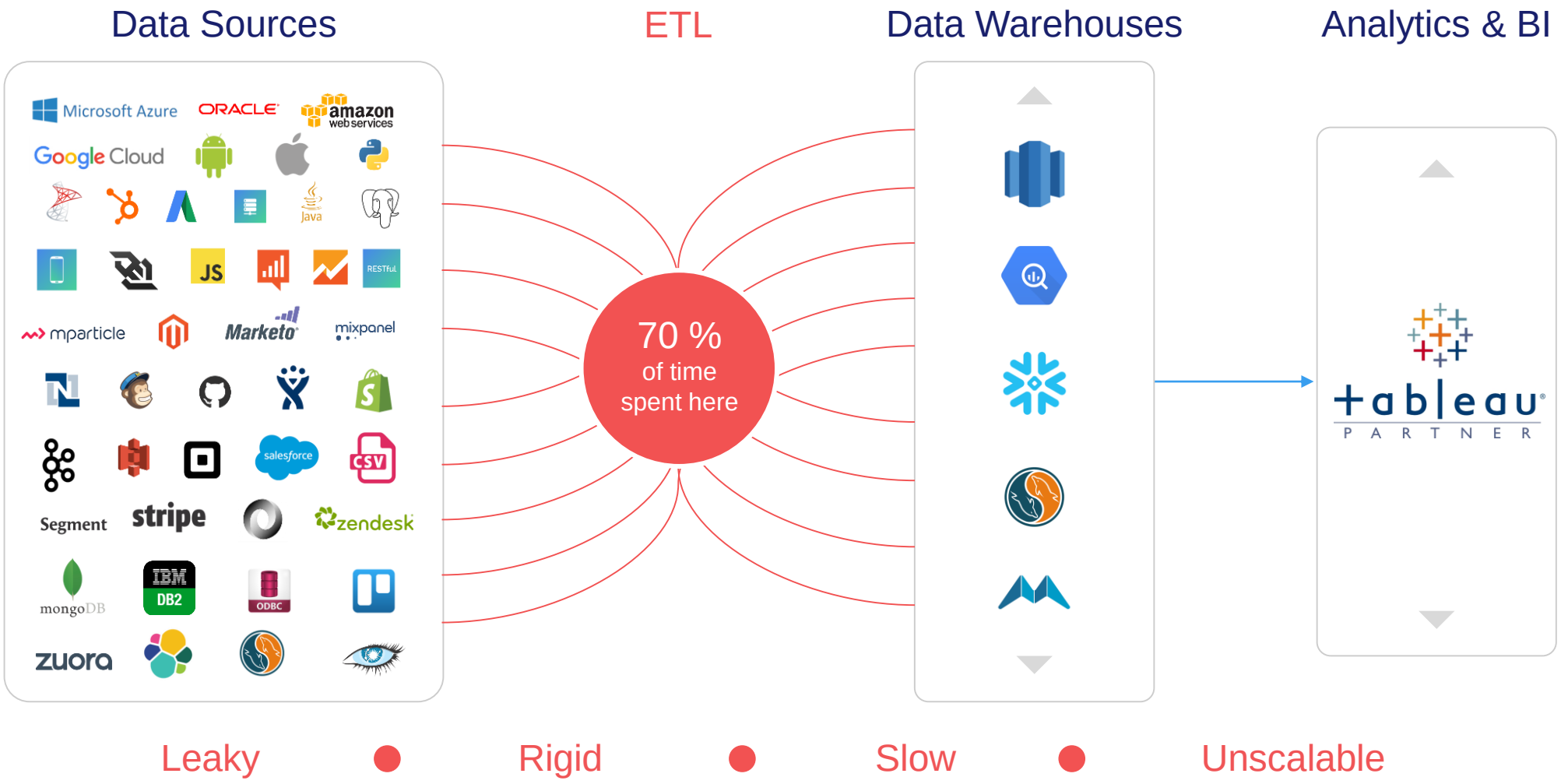
# OK, so how do I get my data there?

I'll just write a simple script…

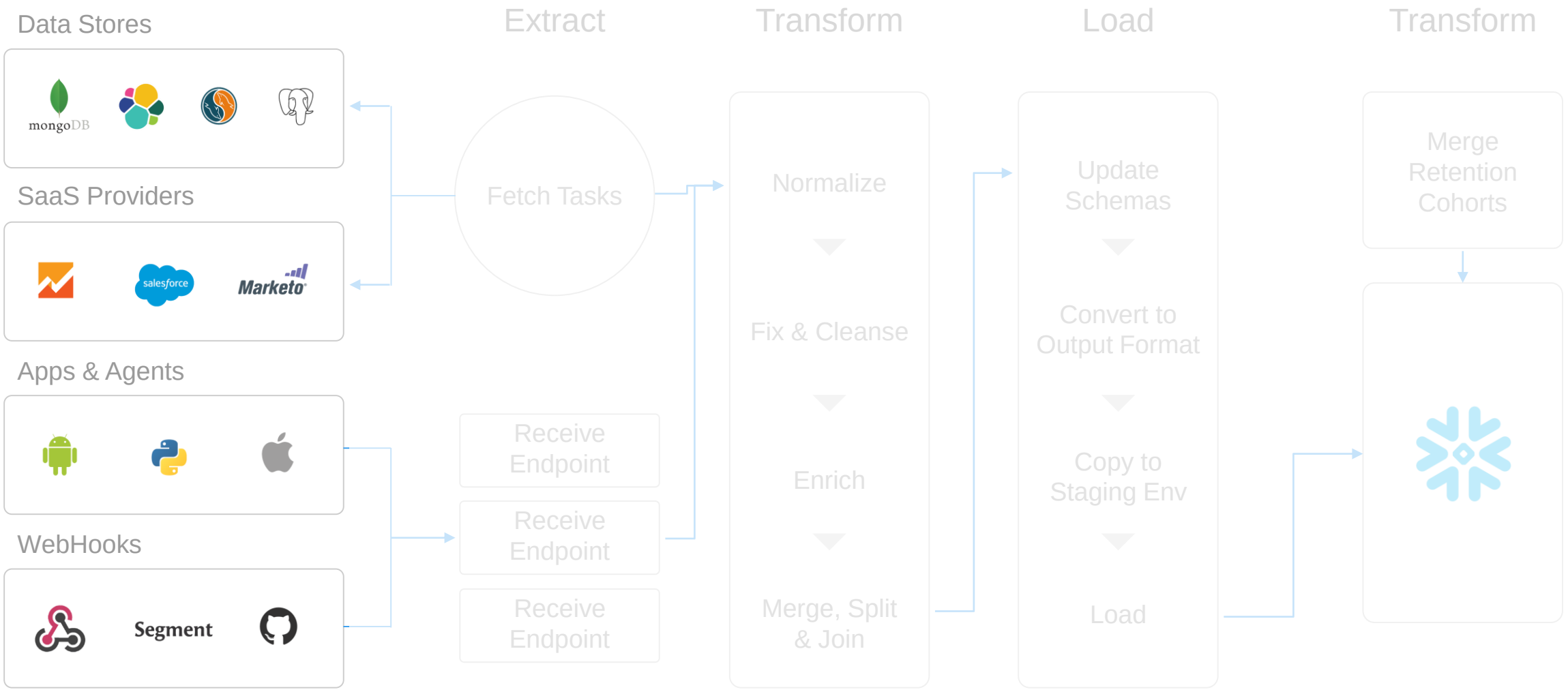


# Getting Data to the Cloud

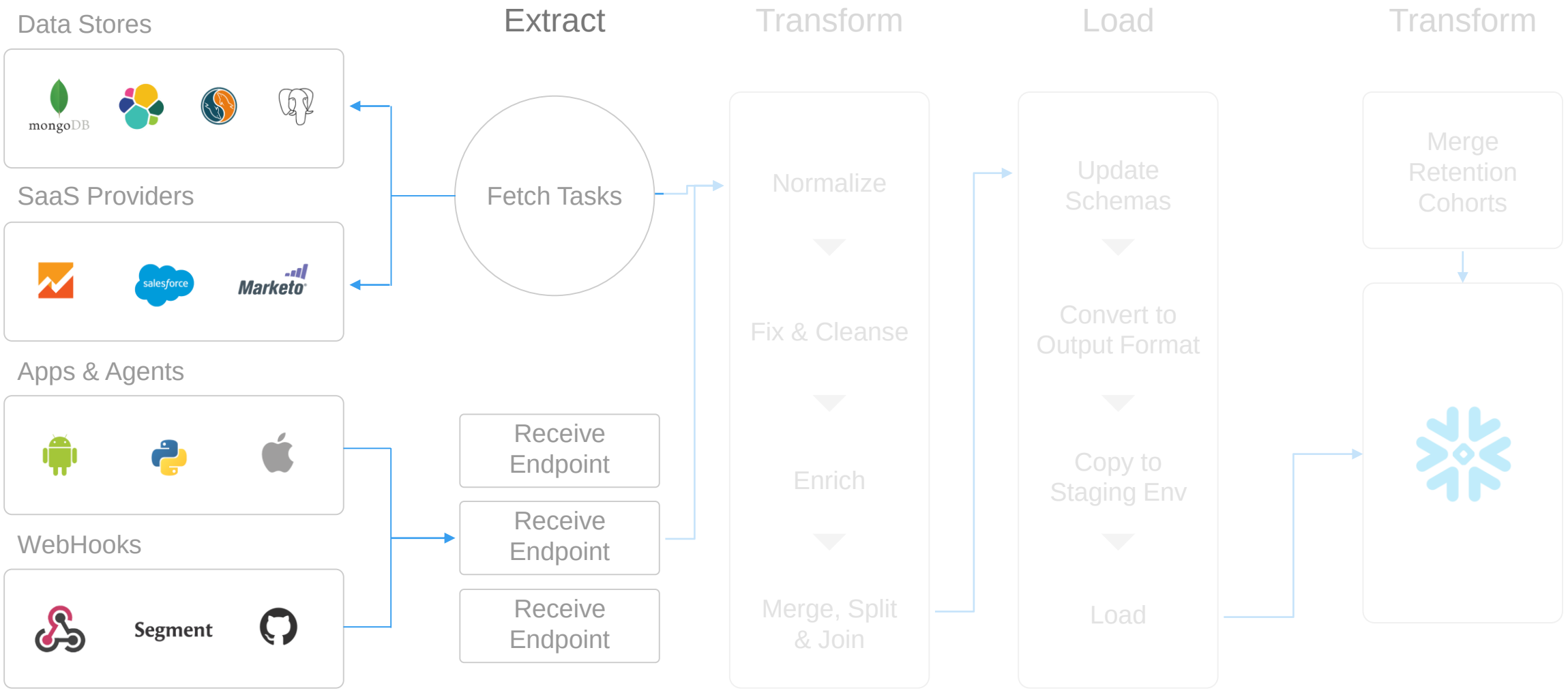




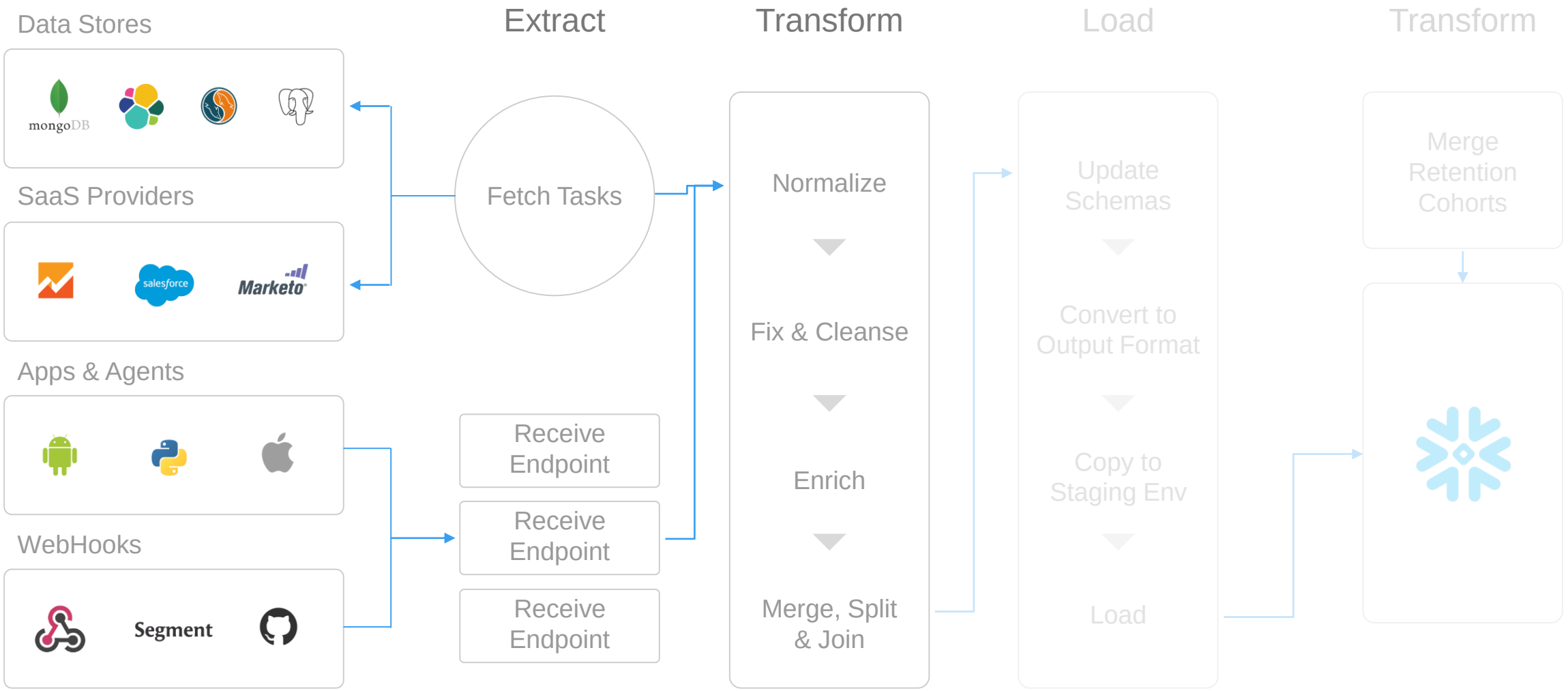
# Basic Architecture of a Unified Data Pipeline



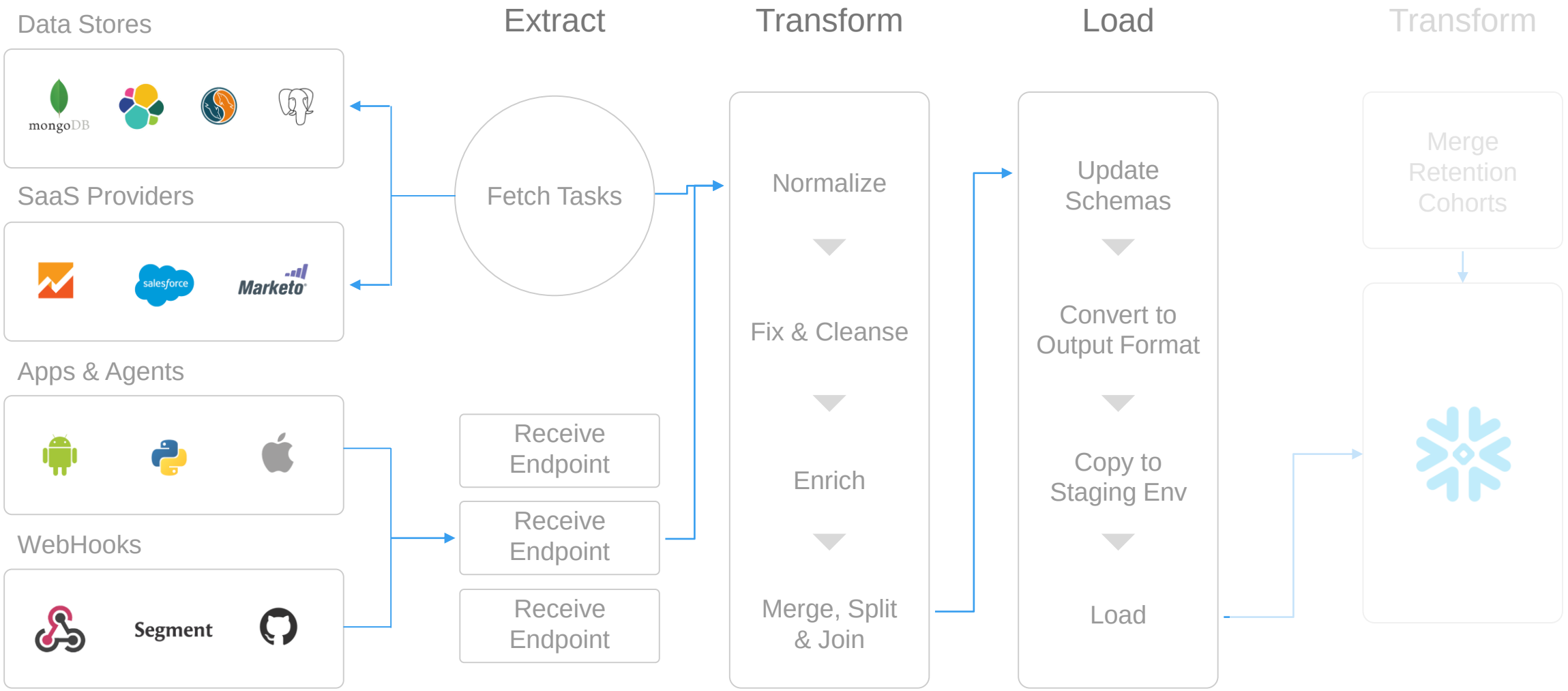
# Basic Architecture of a Unified Data Pipeline



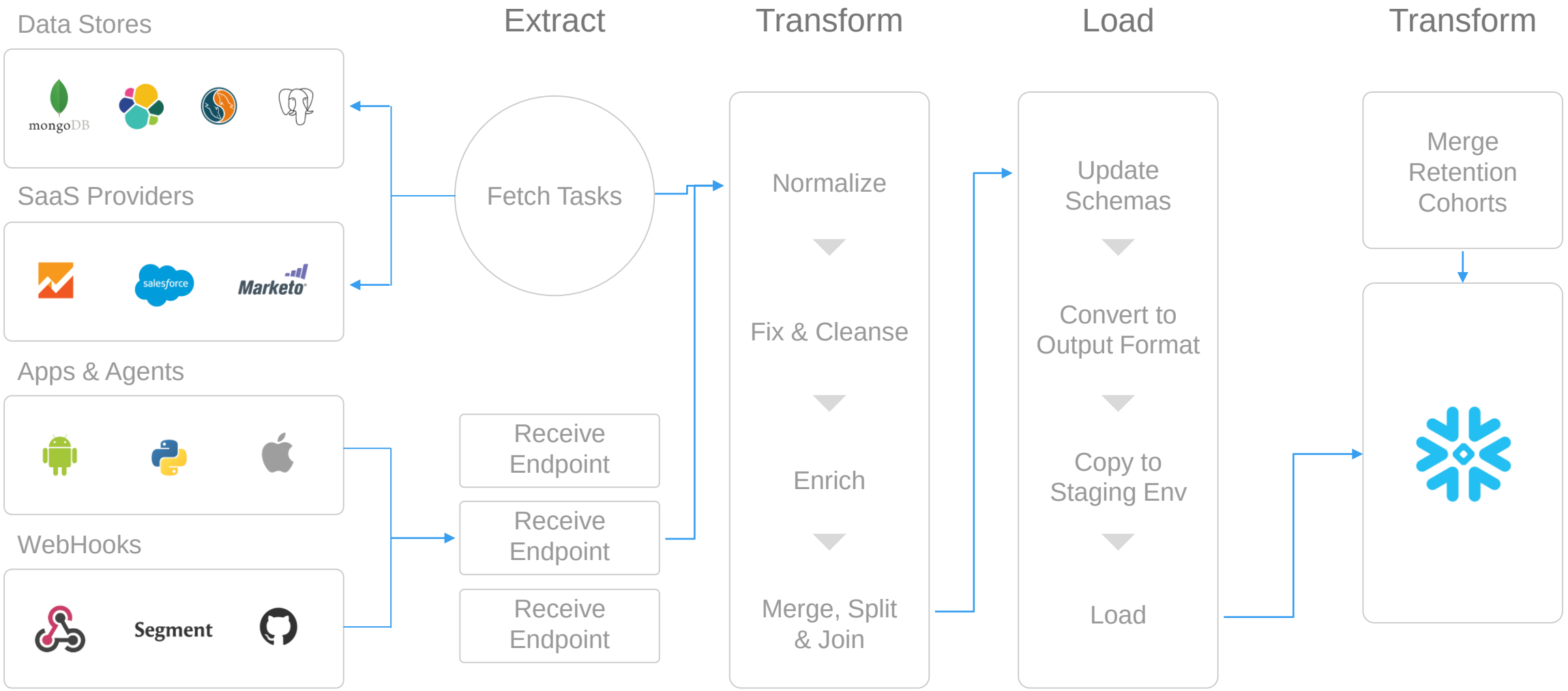
# Basic Architecture of a Unified Data Pipeline



# Basic Architecture of a Unified Data Pipeline



# Basic Architecture of a Unified Data Pipeline



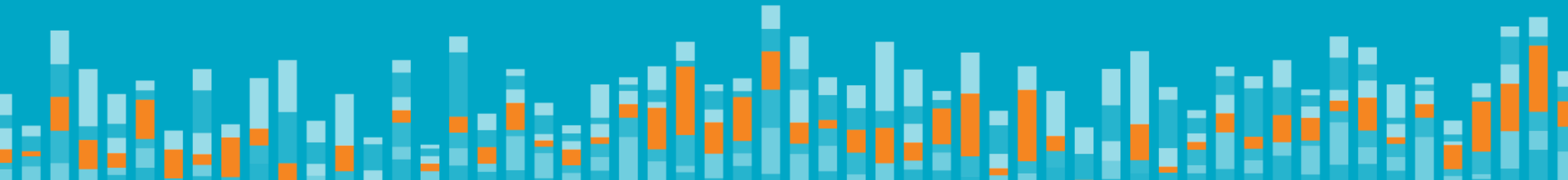
# Main Challenge of a Data Pipeline is Integrity

TABLEAU CONFERENCE

TABLEAU CONFERENCE

# Let’s tackle dynamic schemas

Because the only constant, is change.



# **Challenge:** Schemas Always Change

TABLEAU CONFERENCE

New tables,
New columns

NoSQL ==
No schema

New features,
New events

Backend apps require monitoring

Weekly
A/B tests

New Marketing tools

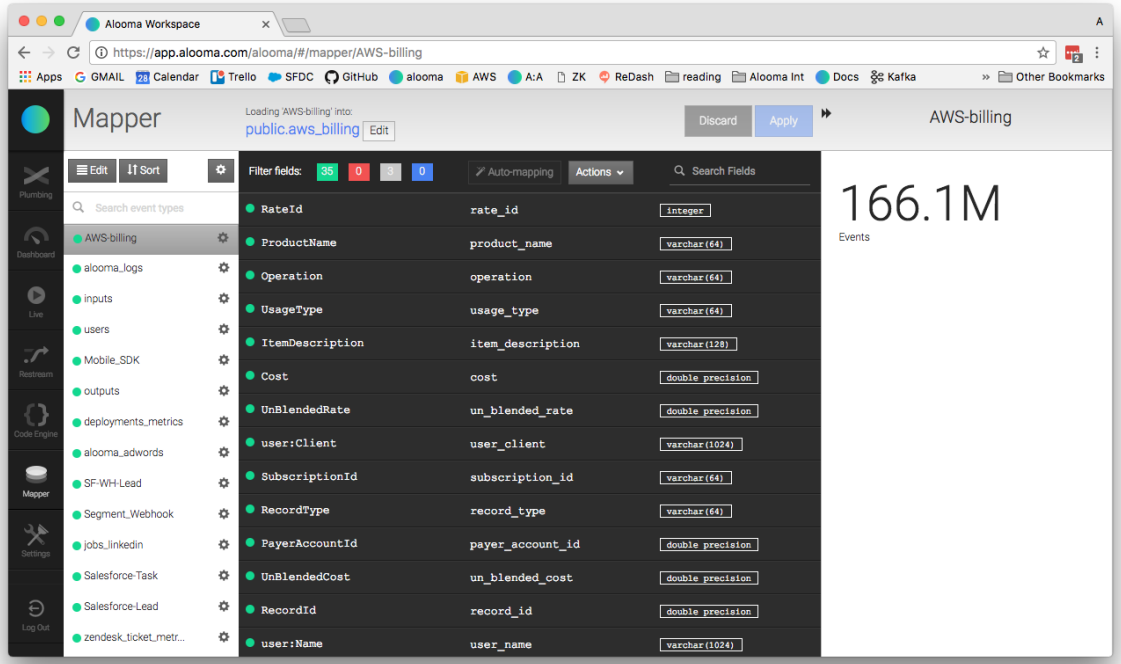
Salesforce data

# Partial Solution: “Guesstimate”

TABLEAU CONFERENCE

- Very early on we built our *Automapper:*
  - Periodic process: Detect Schema Changes
    → Modify Schemas on data warehouse
    → Update schema mappings

- Detecting schema changes was based on sampling the events, and guessing the data types of the different values



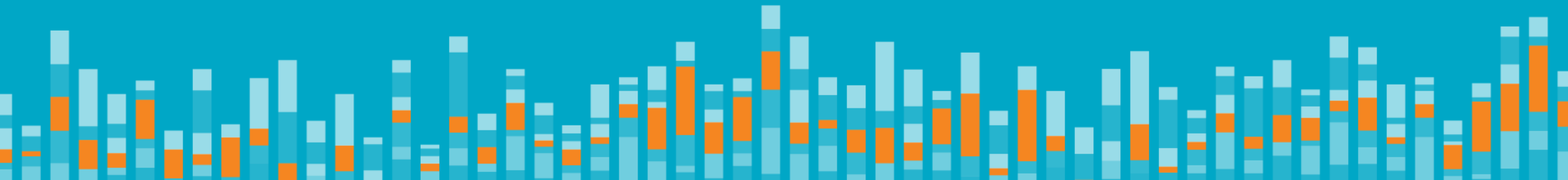
# More Complete Solution:
# “Guesstimate” + Schema Import

TABLEAU CONFERENCE

- Import schemas from any data source that supports it
  - Added an import_schema(schema_id) function to our **inputs**:
    MySQL, Postgres, MSSQL, Oracle, Mongo, Salesforce, Google Adwords, ...
  - Added a schema_id metadata field to all **events**
  - New Automapper:
    Detect Schema Change
    → Try to extract schema_id and import the updated schema
    → Translate source schema to data warehouse schema
    → Modify Schemas on data warehouse
    → Update schema mappings

TABLEAU CONFERENCE

# OK, but what about errors?



# **Challenge:** The Pipeline May Fail at Any Stage

TABLEAU CONFERENCE

Failed Value Conversions
37%

Data Warehouse Errors
19%

Missing Required Values
11%

Changed Schemas
11%

Custom Logic Failures
9%

Strings to Integers
9%

Overly long strings
5%

** Statistics taken from 1B events in our Restream Queues*

# **Solution:** Store All Failed Events and Failure Details

TABLEAU CONFERENCE

- Build a central, “catch-all,” event store that can also save error details
- Program all pipeline components to write to it in case of error
- Always store the original, unmodified event to allow simple re-processing
- Design a querying interface to enable easy classifying, grouping, and prioritizing of errors

- Beware of:
  - pipeline loops
  - multiple pipeline destinations
  - complex processing (event splitting / joining / enrichment)

# Implementation V1

TABLEAU CONFERENCE

- Kafka queue to store all failed events (in original format)
- Metadata on each failed event:
  - Failure reason
  - # of failures (to avoid loops)
- Notification service which aggregates failure notifications by error type, similar parameters
- Redis to store samples of failed events

# Implementation V2

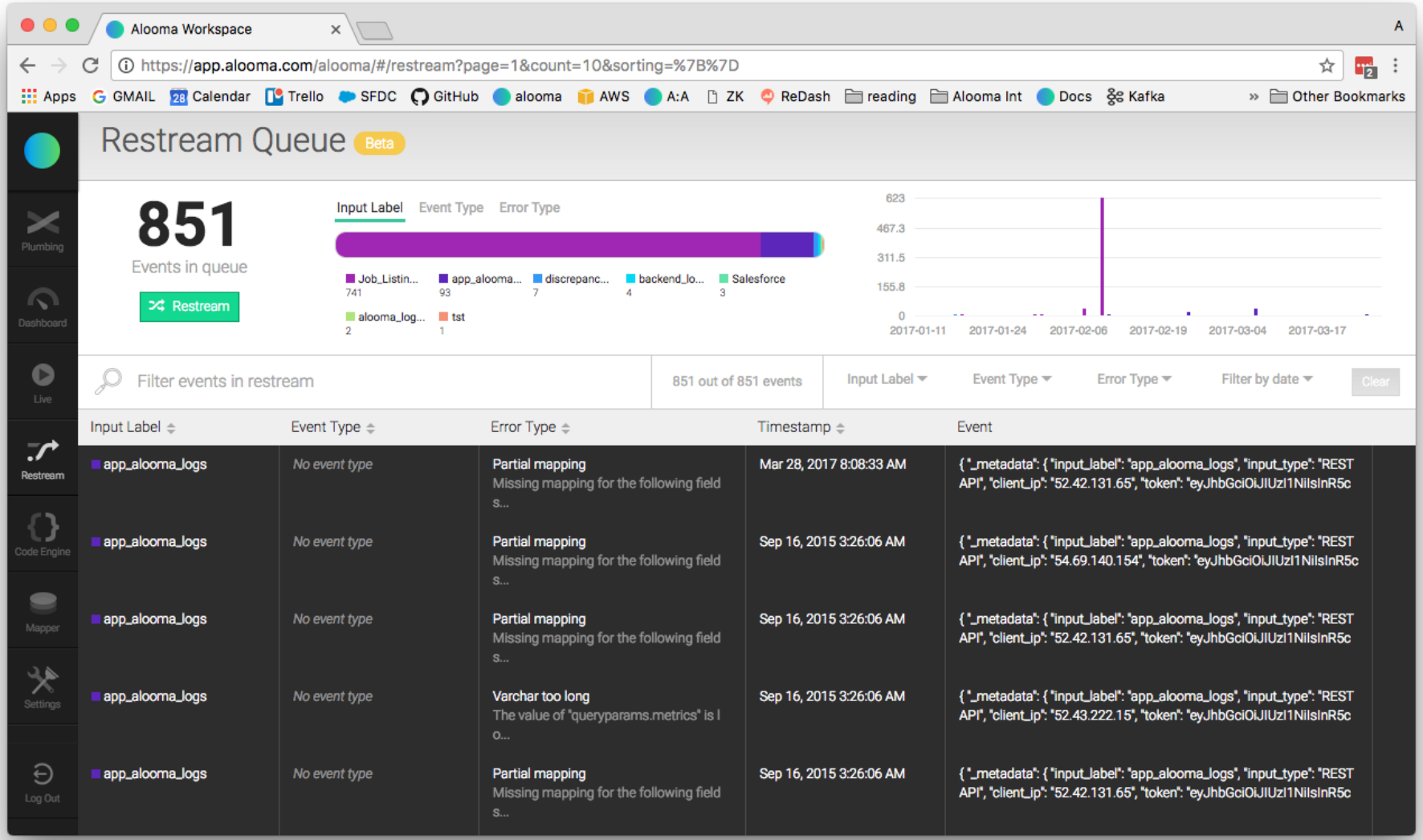
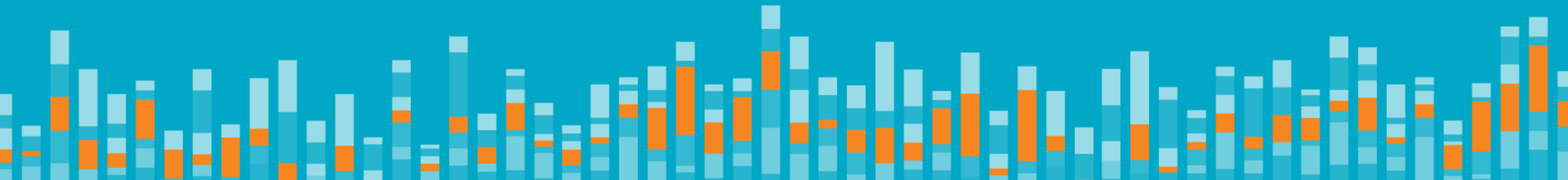


TABLEAU CONFERENCE

# Uh oh, won’t this mean duplicates?

Or accidentally overwriting newer data with old data?



# **Challenge:** Transmit All Events Once, and *Only* Once


# **Solution:** Idempotency

TABLEAU CONFERENCE

**idempotent** | ˌīdemˈpōt(ə)nt ˈēdemˌpōt(ə)nt | *Mathematics*

adjective

denoting an element of a set that is unchanged in value when multiplied or otherwise operated on by itself.

- In simple words: a duplicate event will overwrite itself
- In practice, this is still difficult:
  - What identifies an event? Its content ? Its ID ?
  - Not all data warehouses support idempotency
  - Every component in the pipeline may fail and resend its last few events upon resuming

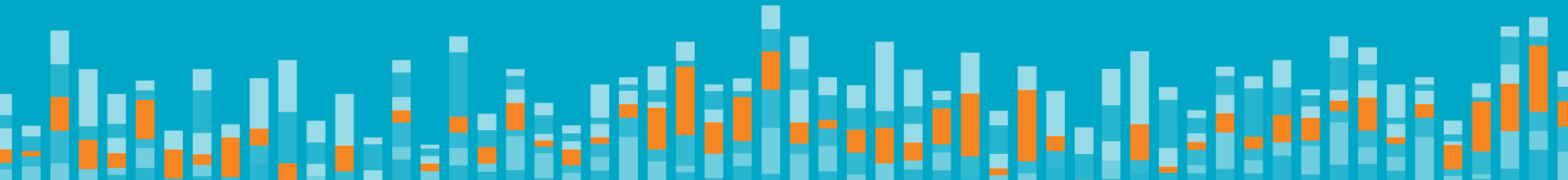
# **Implementation:** Track IDs and Batches End-to-End

TABLEAU CONFERENCE

- Identifying events & duplicates
  - Generate an event ID & batch ID as early as possible (at the data source)
  - Where possible, make the IDs sequential
- Idempotency with cloud data warehouses
  - Primary key constraints are not validated on cloud data warehouses, so overwriting is not an option
  - We must keep track of loaded events (or batches) in a separate table and:
    - Verify writing to both tables is atomic (using a transaction)
    - Check whether events have been written before writing every new batch
  - Expand this to encompass your whole pipeline, either end-to-end or step-by-step

TABLEAU CONFERENCE

# So does all of this work?



“**Alooma just works.** We're able to leverage the platform's real-time architecture to **iterate on solutions in minutes rather than days**, giving us the agility our business demands.”

“

Alooma makes it easy for us to join data from across all marketing channels **without draining our internal engineering resources** on untested solutions.

SONY

“

Alooma gives our marketing organization **real-time data analytics pulled from across all app and ad sources** without draining our internal engineering resources or sacrificing our security practices.

STRAVA

TABLEAU CONFERENCE

“

Alooma provides us a scalable, robust and flexible infrastructure needed to **accelerate our real-time analytics across the entire company**.

TABLEAU CONFERENCE

# Can I see this in action?

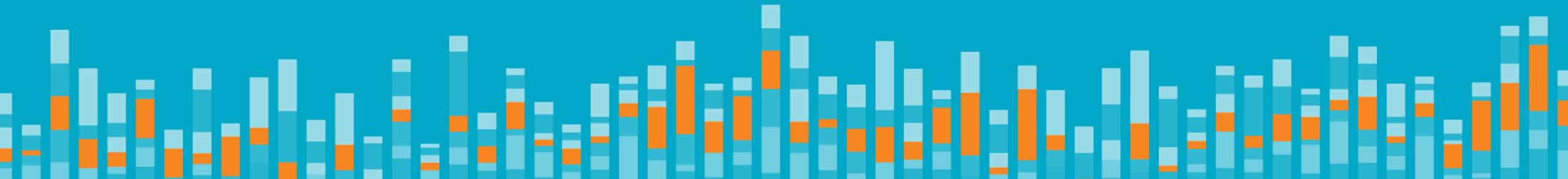
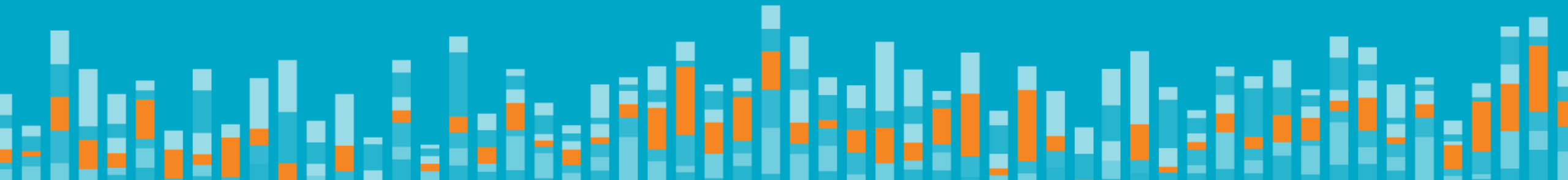


TABLEAU CONFERENCE

# Closing Thoughts



# Closing Thoughts

TABLEAU CONFERENCE

- By leveraging a cloud based data warehouse and ETL platform, you dramatically improve data usage across the organization with more reliable data
- 99.999% data integrity requires design and tested implementation
- When you plan your next analytics platform:
  - Plan for a highly scalable, simple to manage data warehouse
  - Include a schema repository & automated mapping tasks
  - Integrate sources with end-to-end idempotency in mind
  - Design a catch-all error queue & an iterative manual fixing process
- Come say hi and **catch a live demo at our booth: #417**

TABLEAU CONFERENCE

# Questions?

max@alooma.com – booth #417