Embrace the Anarchy :

Apache Kafka's Role in Modern Data Architectures

Robin Moffatt / Confluent

Photo by Jaak Horn on Unsplash

$ whoami

- Developer Advocate @ Confluent
- Working in data & analytics since 2001
- Oracle Developer Champion
- Blogging : http://rmoff.net & http://cnfl.io/rmoff
- Twitter: **@rmoff**
  - Geek stuff
  - Beer & Fried Breakfasts

https://speakerdeck.com/rmoff/

STREAM PROCESSING

We ❤ syslogs: Real-time syslog Processing with Apache Kafka and KSQL—Part 1: Filtering

syslog is one of th
Built into operating
devices like IP ca
metadata such as
target is simply a l
may log or process

As a high-performa
centralised ingesti

Robin Moffatt 🍻🏃🥓
@rmoff

#FullEnglish ift.tt/2v7cLfE

“

Apache Kafka is a Streaming Platform

“why do we need a streaming platform?

“

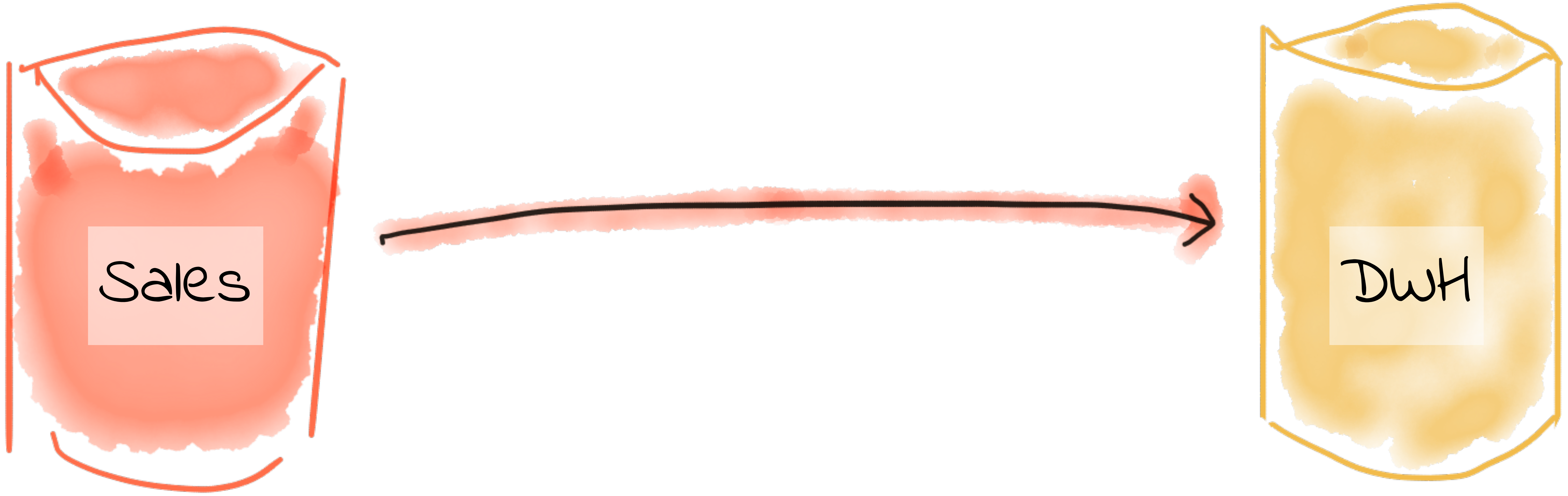
one of the reasons:

Decoupling

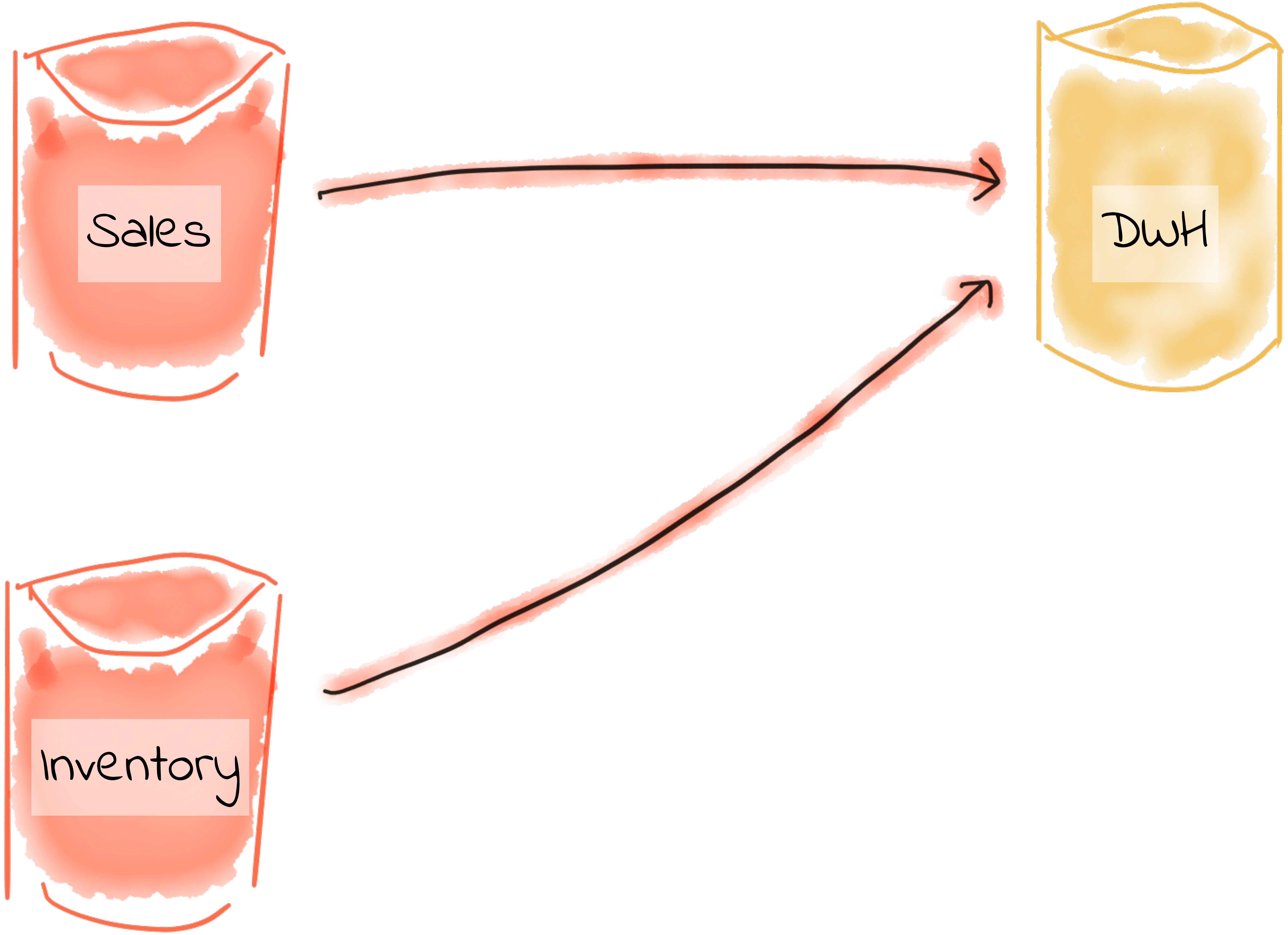
“

A case in point…Analytics

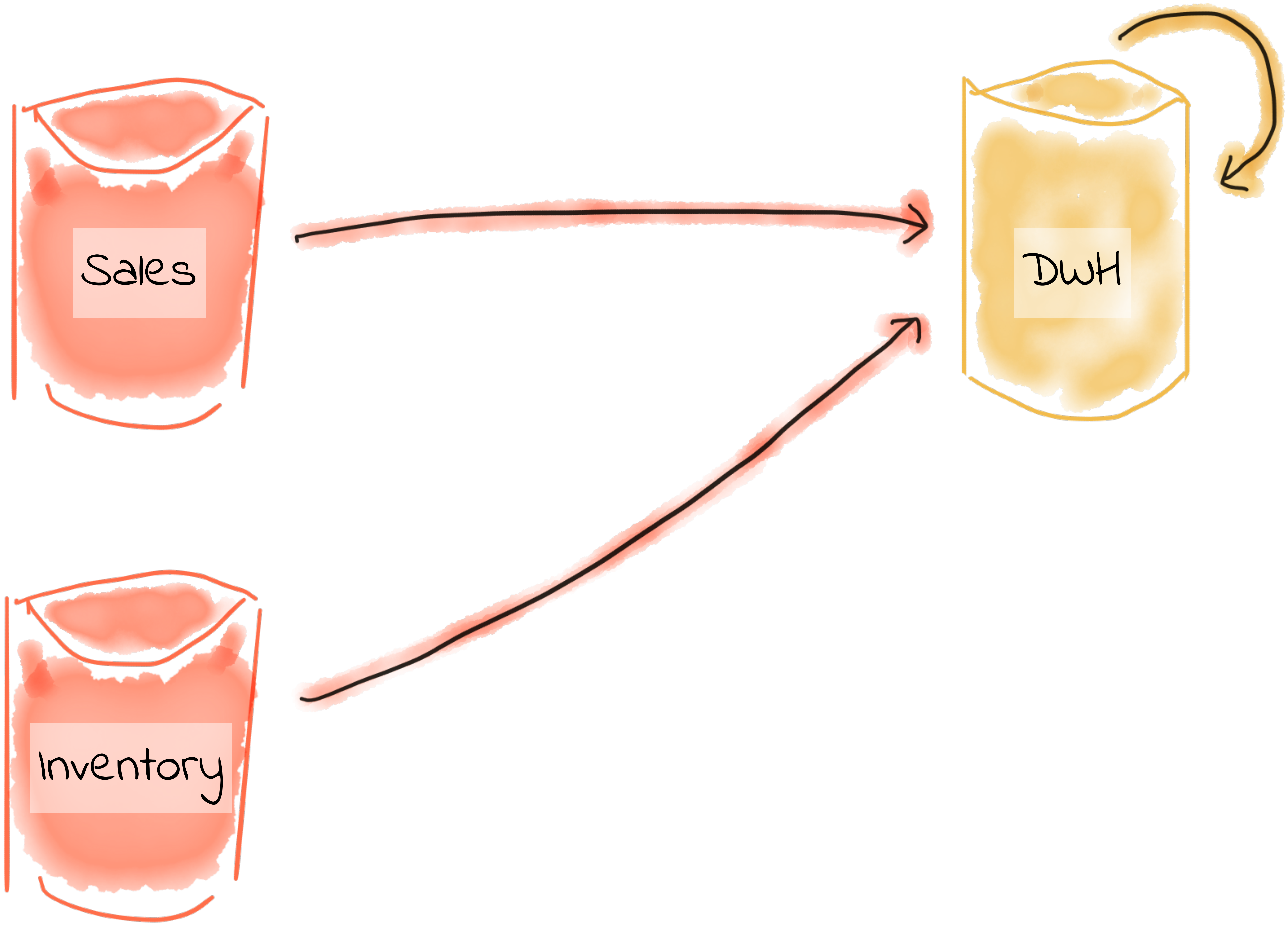
Analytics—In the beginning…



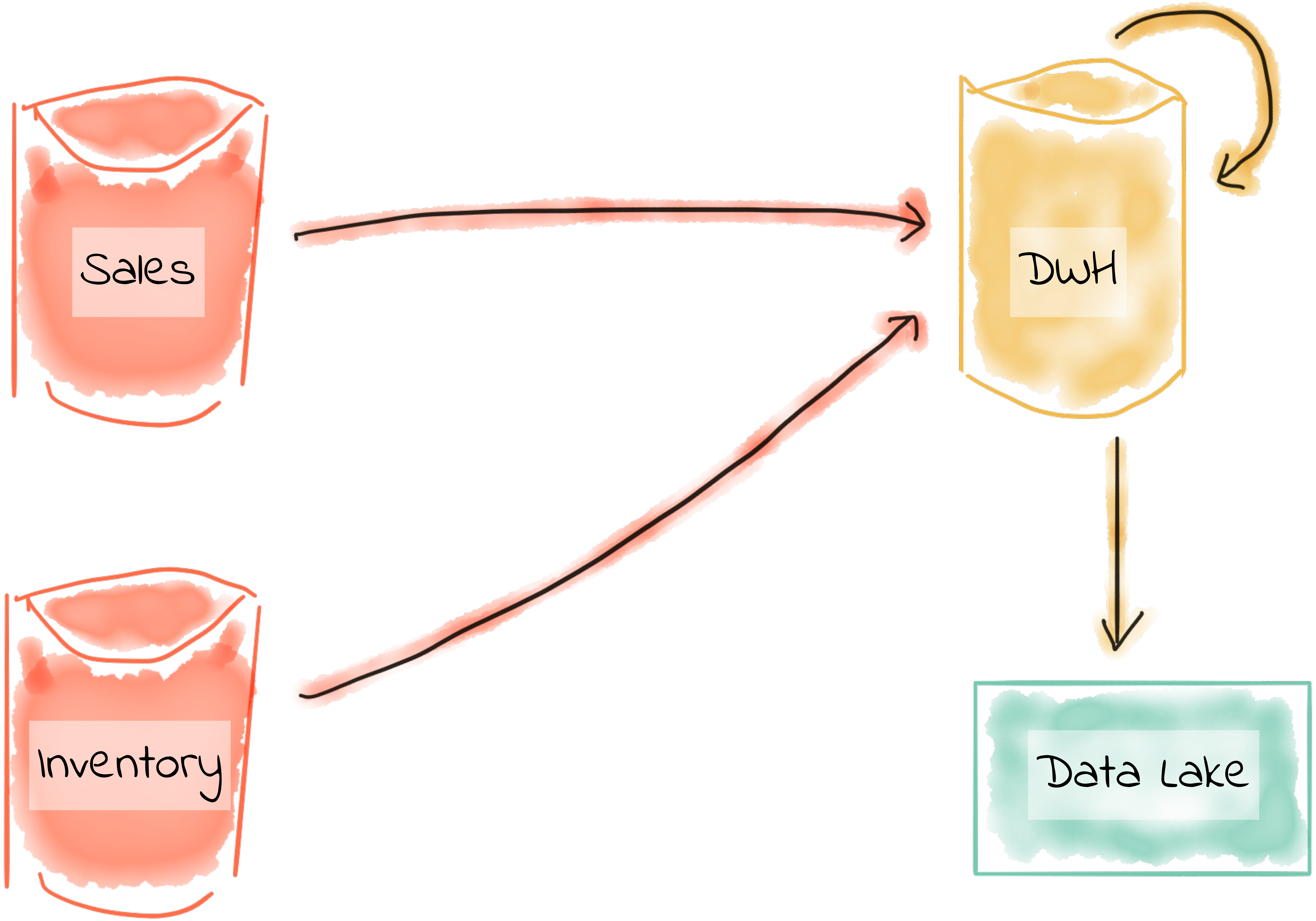
And then there were more data sources…



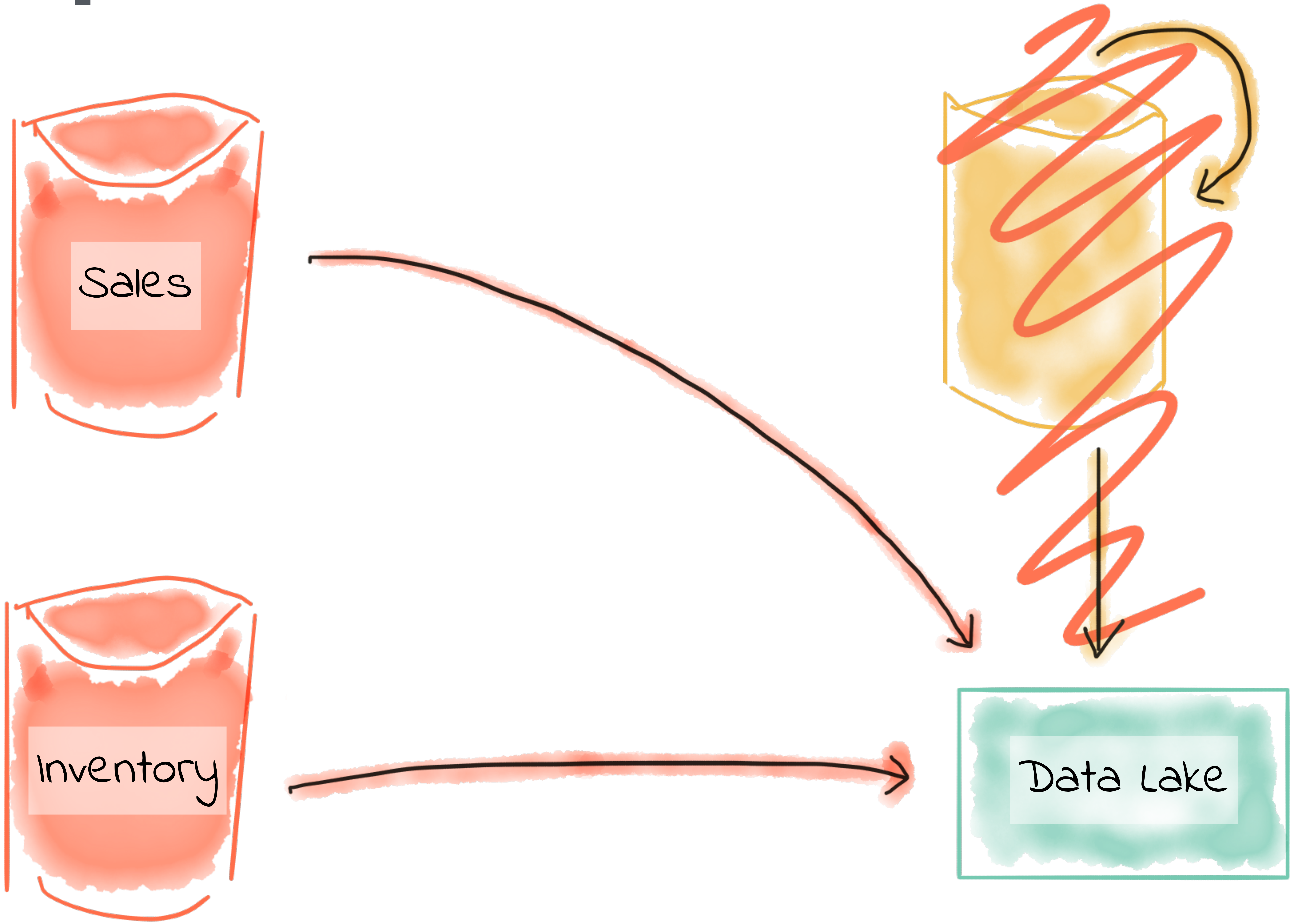
Batch Transformations … (ETL / ELT)



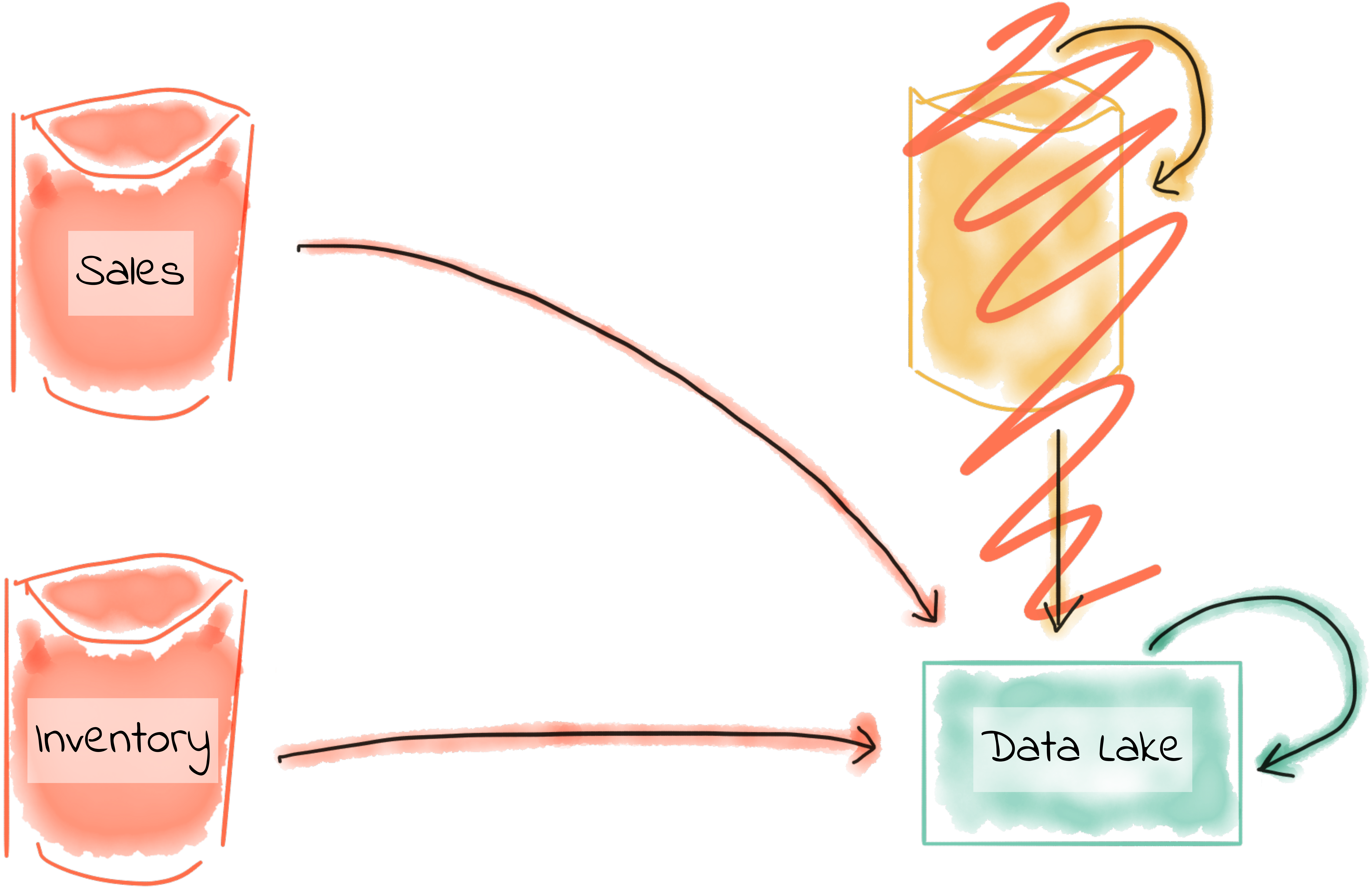
Add a Data Lake…



...or Replace the Data Warehouse



Still need to do Batch transformations…



Want your data anytime ➡ SOON ?

Batch is Latency built in by Design

The World has Changed

Internet of Things

Microservices

Mobile

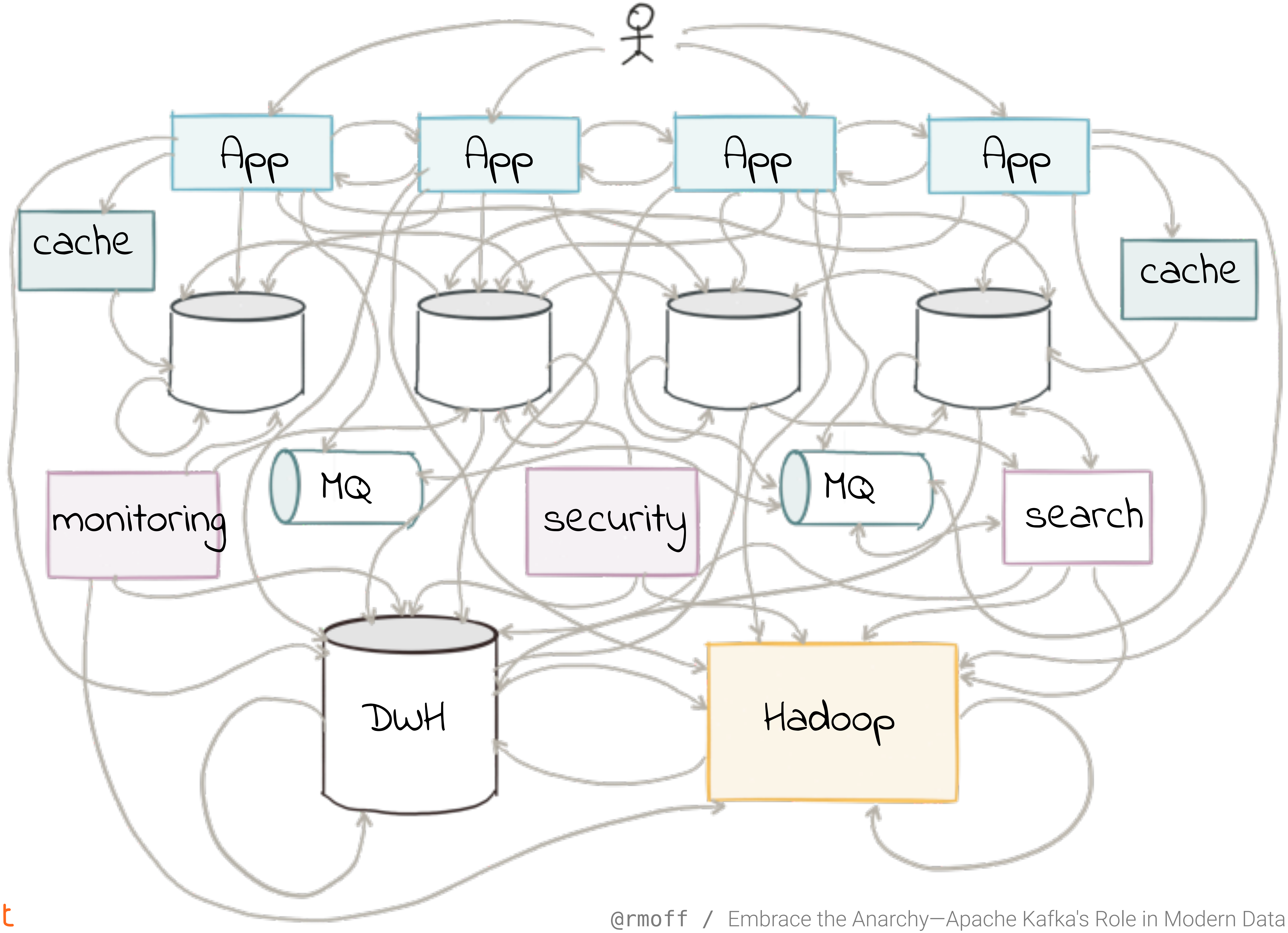
Machine Learning

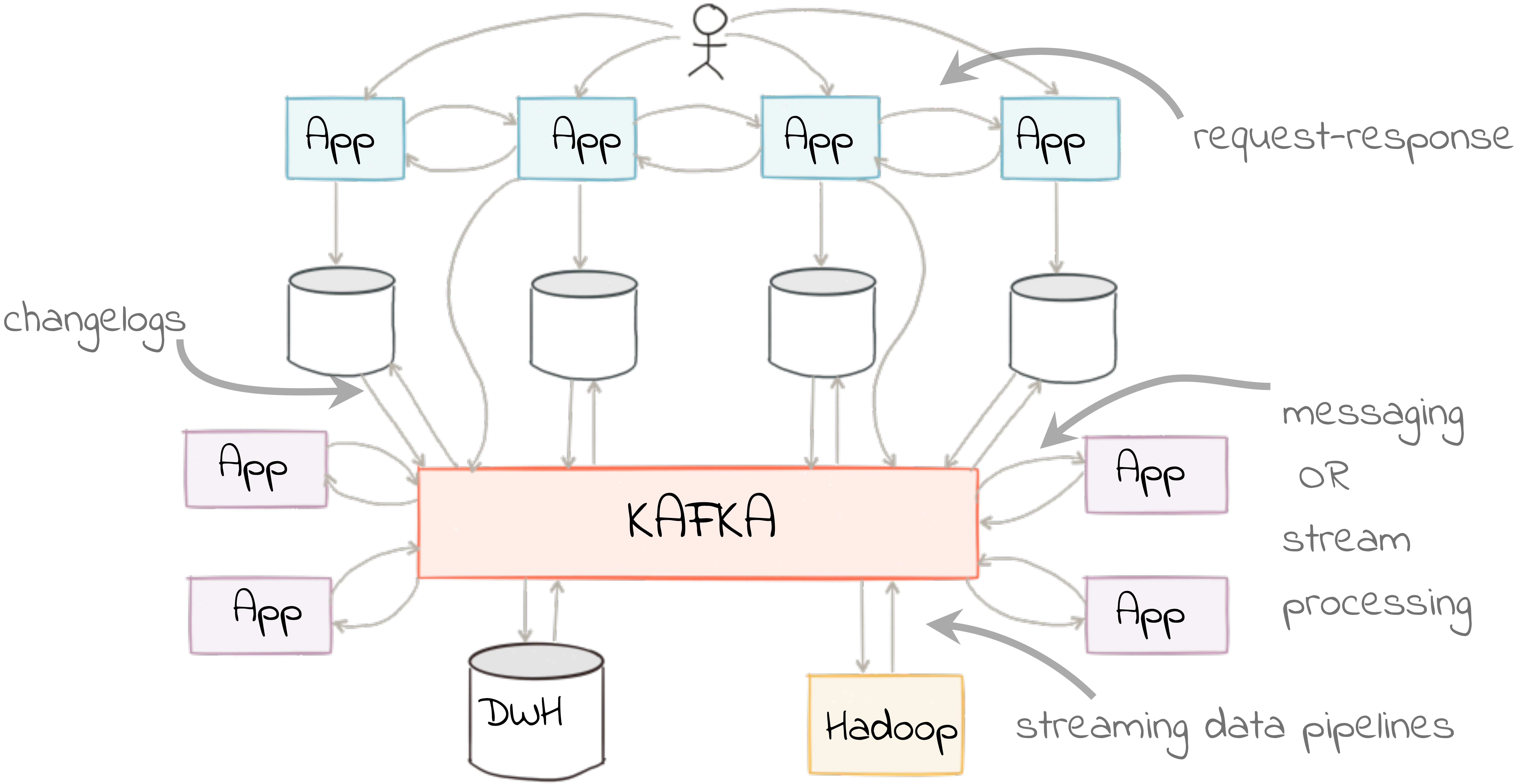
Photo by Denys Nevozhai on Unsplash

Lots of new technologies

(whether you like it or not)

Photo by Rosie Fraser on Unsplash


App
App
App
App
cache
cache
MQ
MQ
monitoring
security
search
DWH
Hadoop


App
App
App
App
request-response
changelogs
App
KAFKA
App
messaging
OR
stream
processing
App
App
DWH
Hadoop
streaming data pipelines

> “Apache Kafka is a Streaming Platform

confluent

Three Lenses

What is Apache Kafka?

01

Messaging
Done Right

02

Scalable Streaming
Data Pipelines

03

Foundation for
Stream Processing

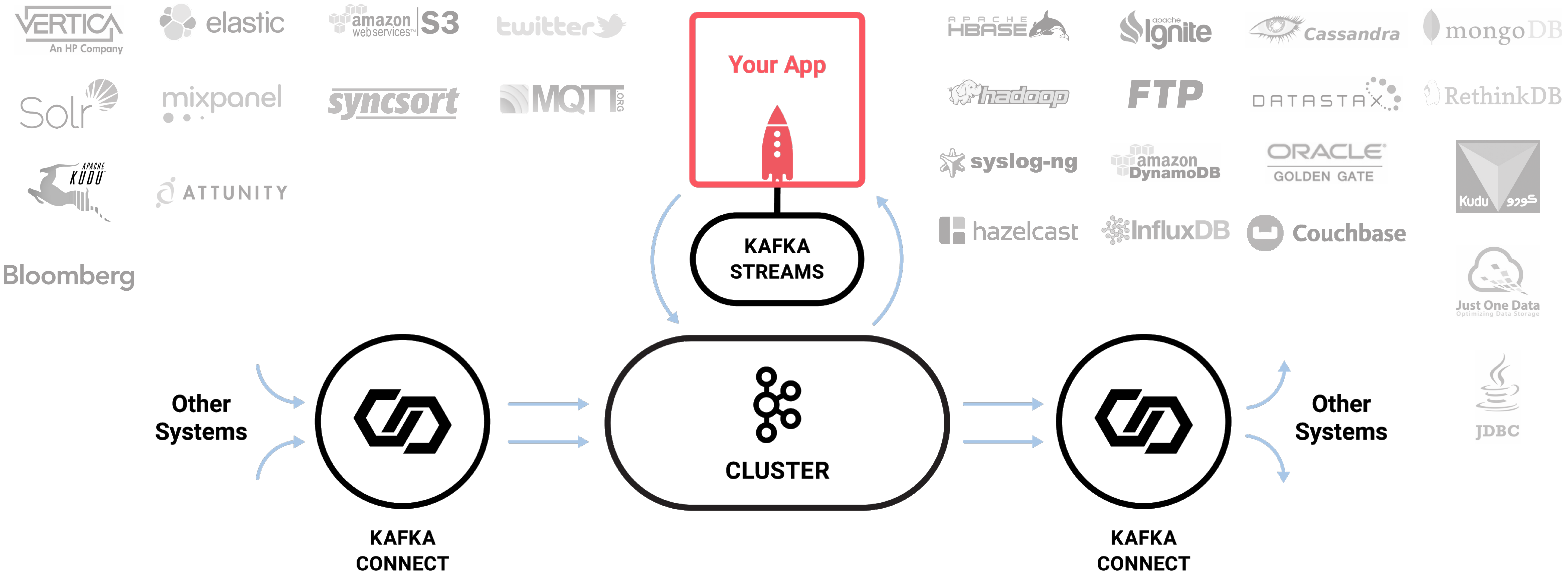
Lens 1: Messaging Done Right

Scalability

True Storage

Real-Time Processing

Lens 2: Scalable Streaming Data Pipelines



Lens 3: Foundation for Stream Processing

KSQL

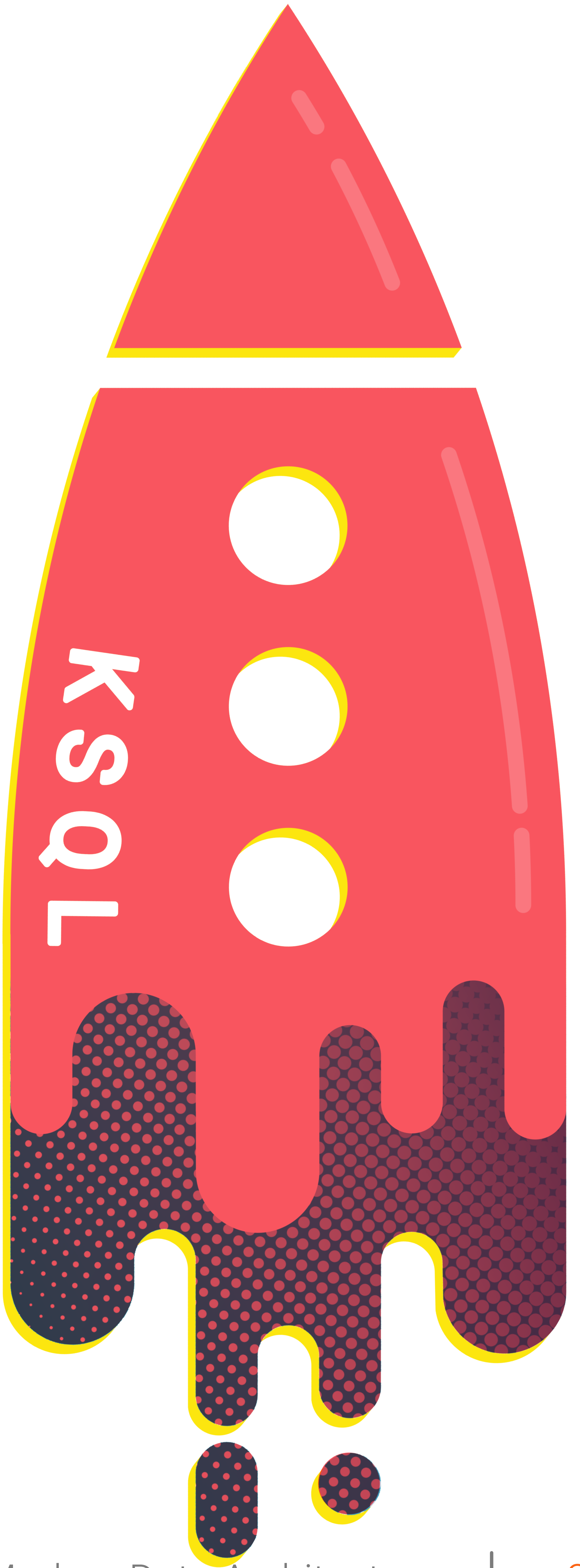
is the

Streaming

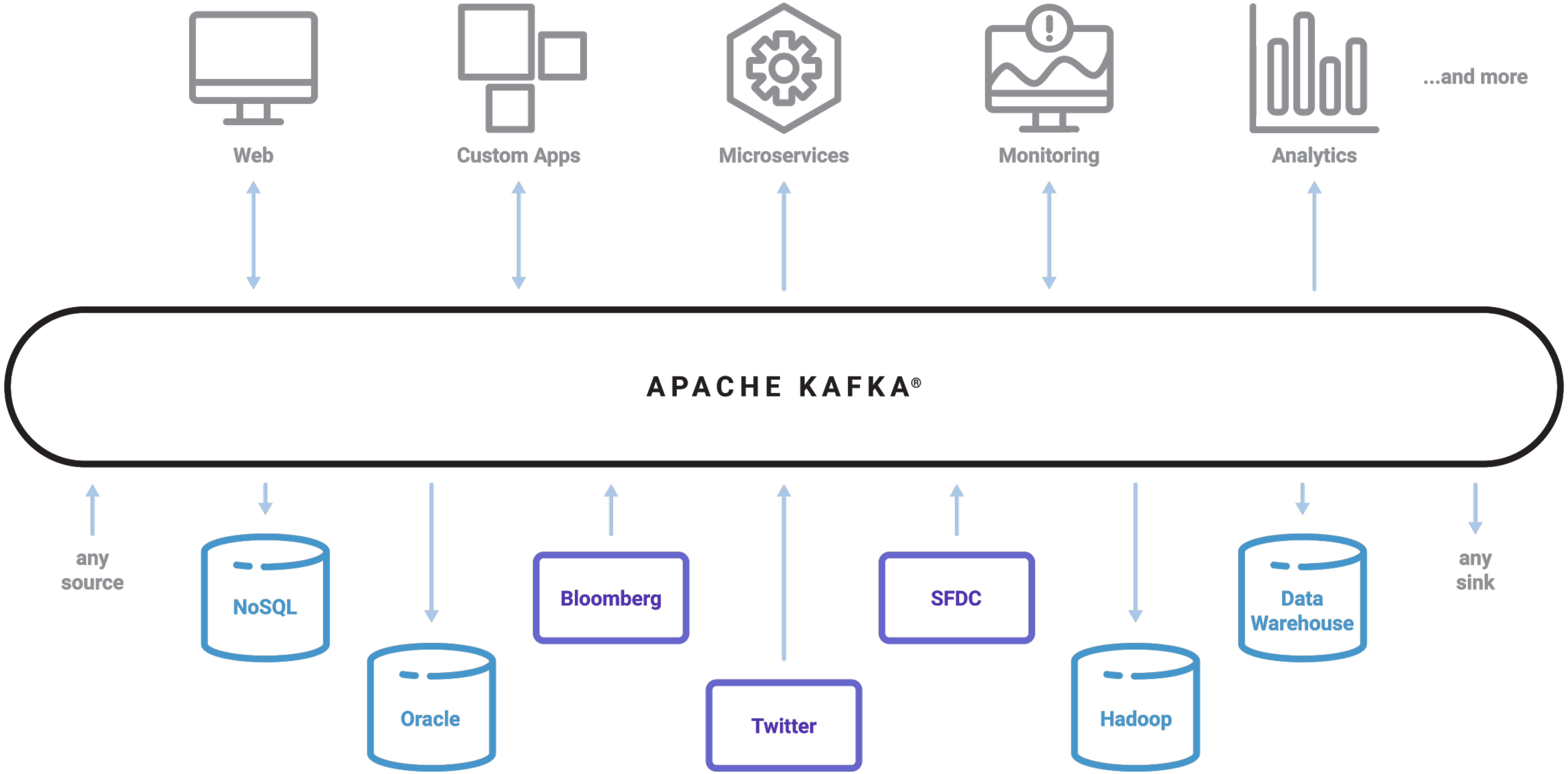
SQL Engine

for

Apache Kafka



The Streaming Platform

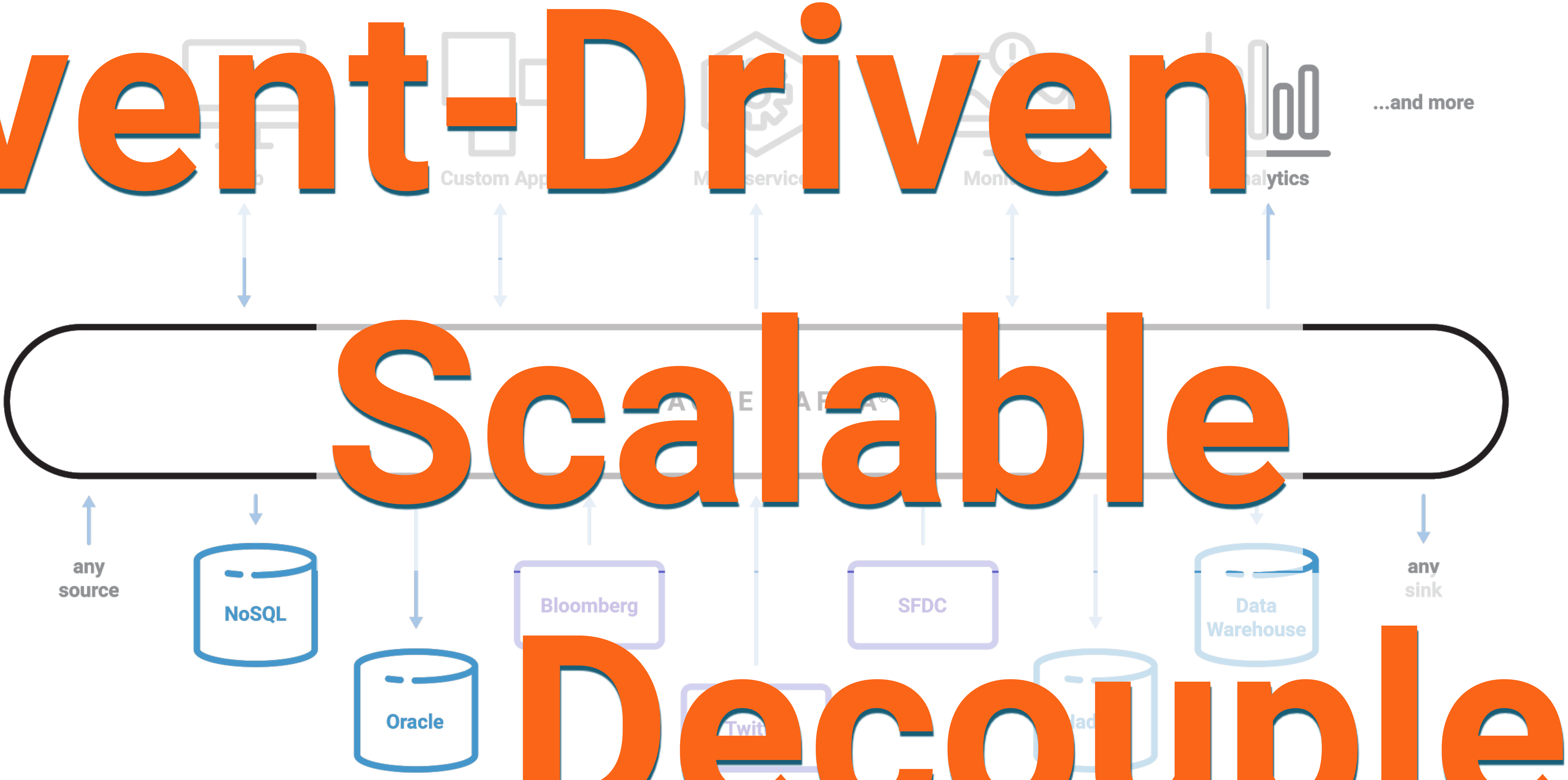


The Streaming Platform

Event-Driven

Scalable

Decoupled



> “Bold claim: all your data is event streams

A Customer Experience

A Sale

A Sensor Reading

An Application Log Entry

Databases

Do you think that’s a table you are querying?

The Table Stream Duality

Stream

Time

Account ID	Amount
12345	+ €50
12345	+ €25
12345	-€60

Table

Account ID	Balance
12345	€50

Account ID	Balance
12345	€75

Account ID	Balance
12345	€15

> The truth is the log.
>
> The database is a cache of a subset of the log.

—Pat Helland
Immutability Changes Everything
http://cidrdb.org/cidr2015/Papers/CIDR15_Paper16.pdf

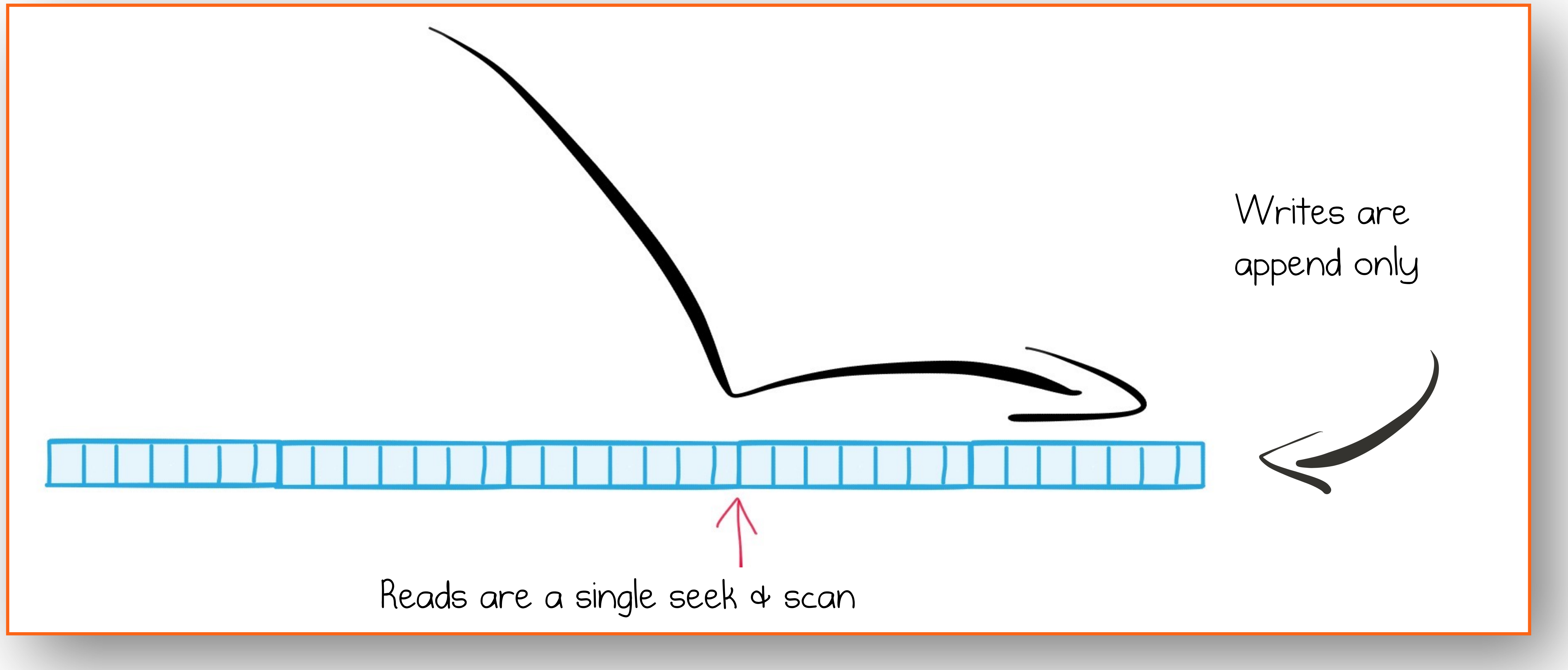
Photo by Bobby Burch on Unsplash

A Brief Look at Kafka's Technology

Apache Kafka

Kafka

A **Distributed Commit Log**. Publish and subscribe to streams of records. Highly scalable, high throughput. Supports transactions. Persisted data. Stream processing.



Producer & Consumer APIs

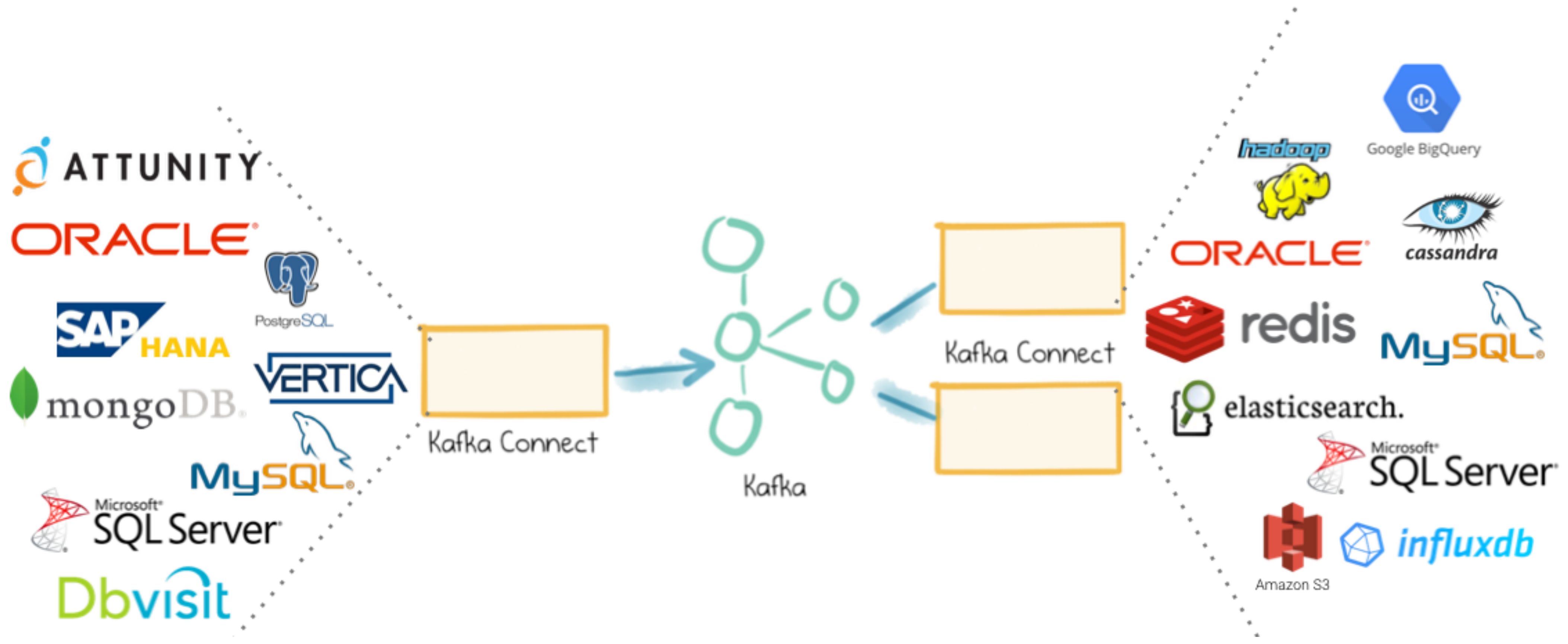
Open-source client libraries for numerous languages, to directly integrate with your applications.

python
Go
Microsoft .NET
Scala
Java

Apache Kafka

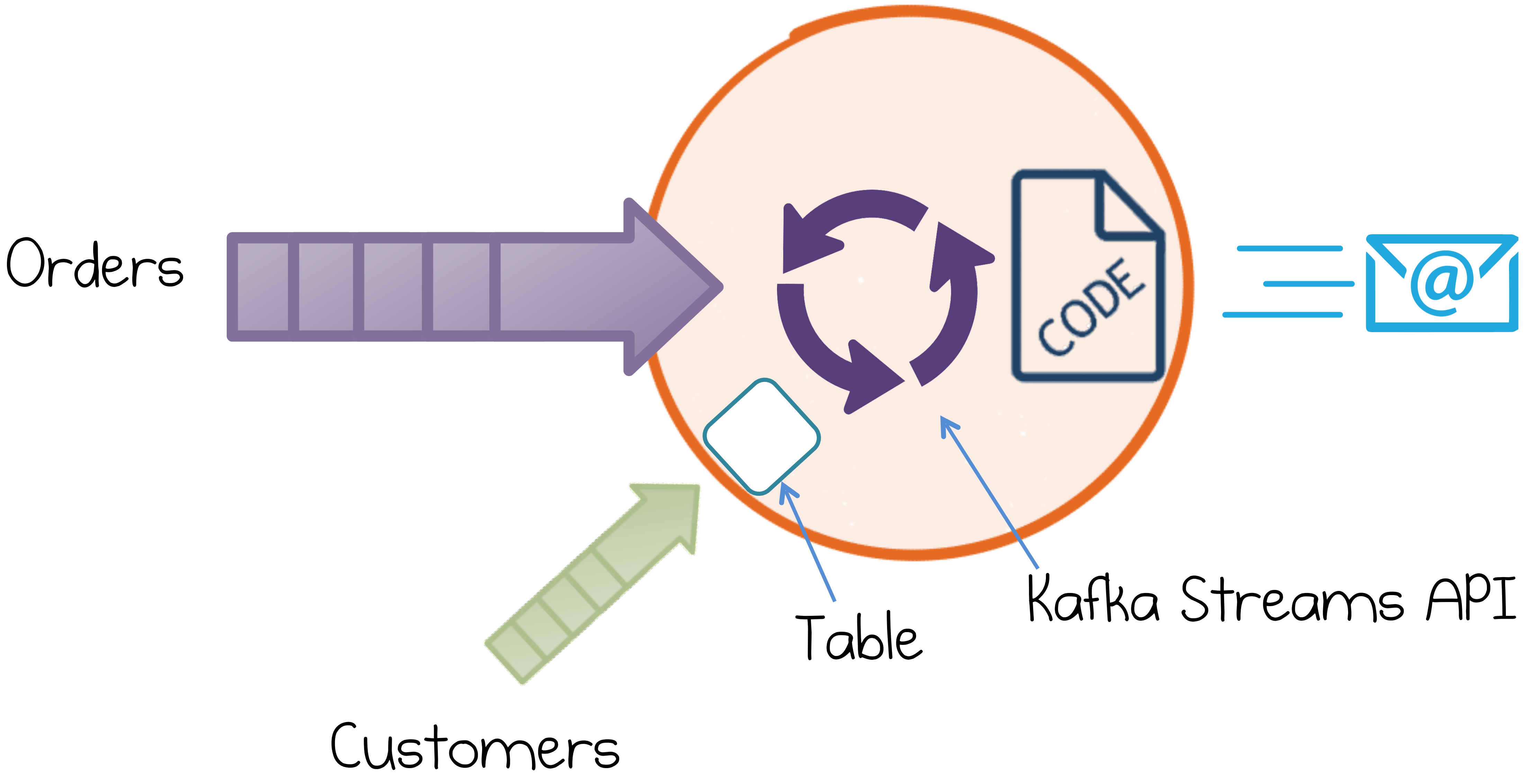
Kafka Connect API

Reliable and scalable integration of Kafka with other systems – no coding required.



Kafka Streams API

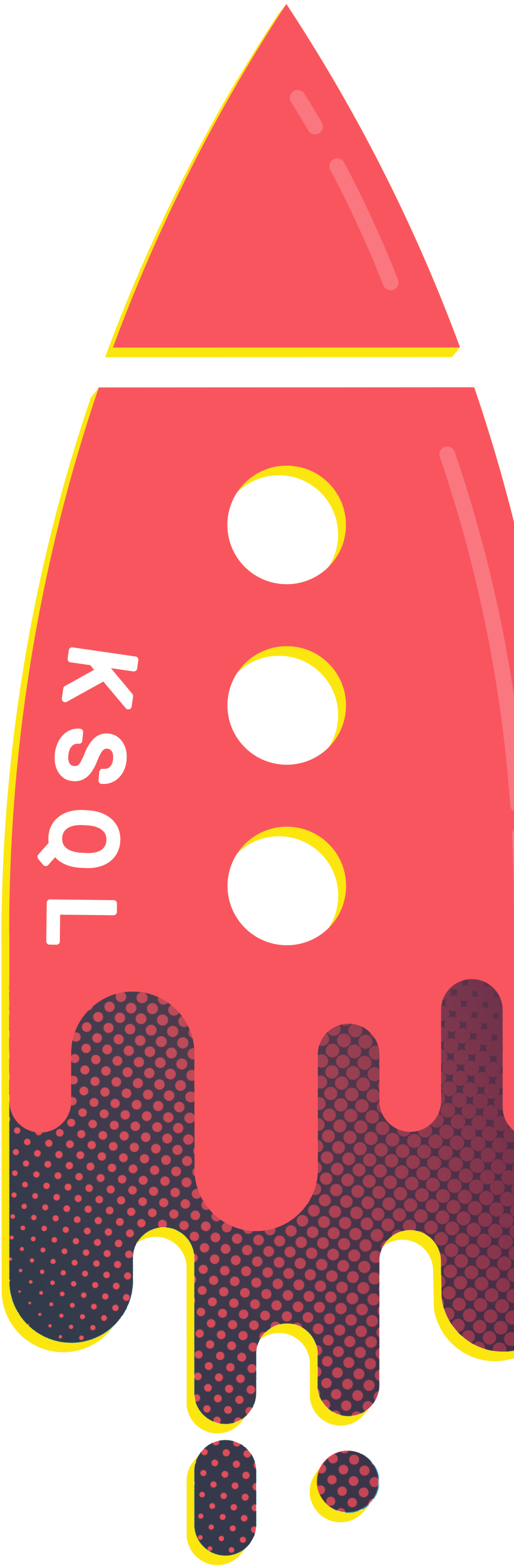
Write standard Java applications & microservices to process your data in real-time



KSQL

is a

Declarative Stream Processing Language



KSQL

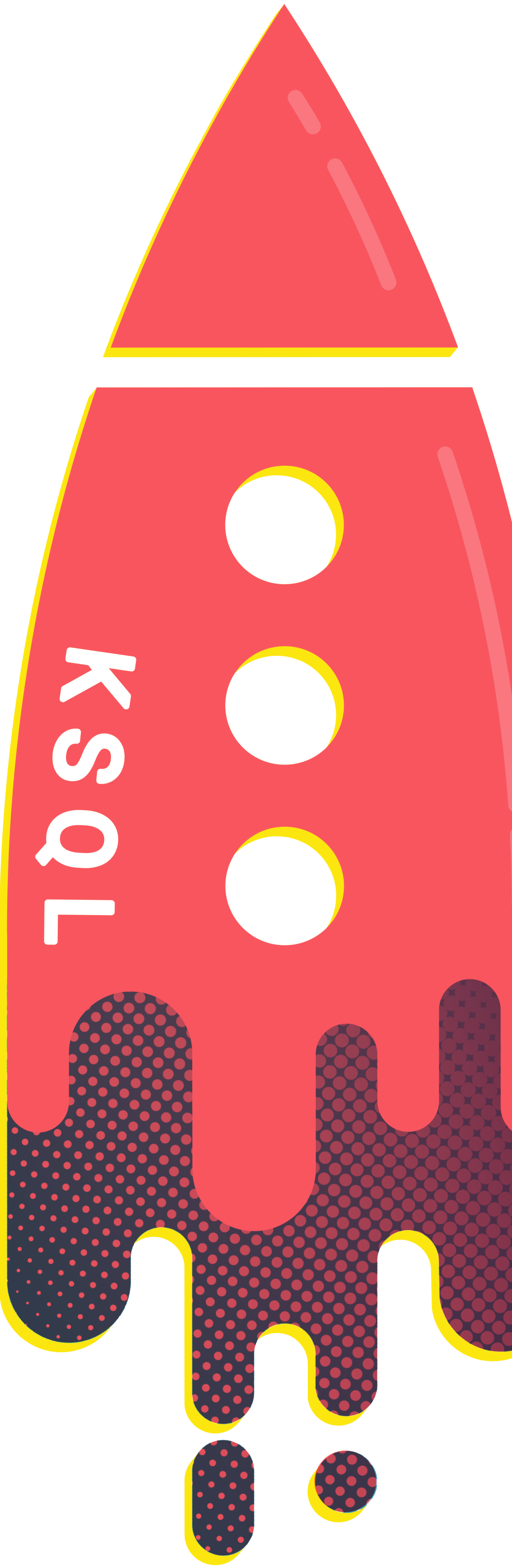
is the

Streaming

SQL Engine

for

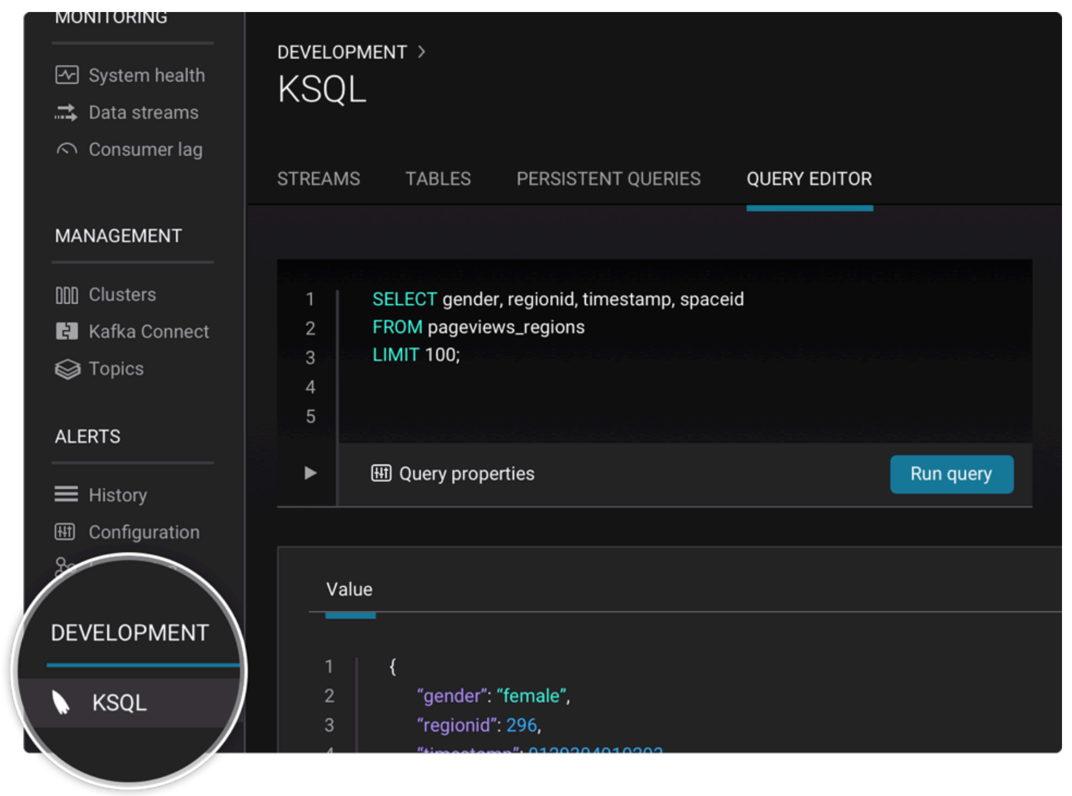
Apache Kafka



KSQL in Development and Production

Interactive KSQL
for development and testing

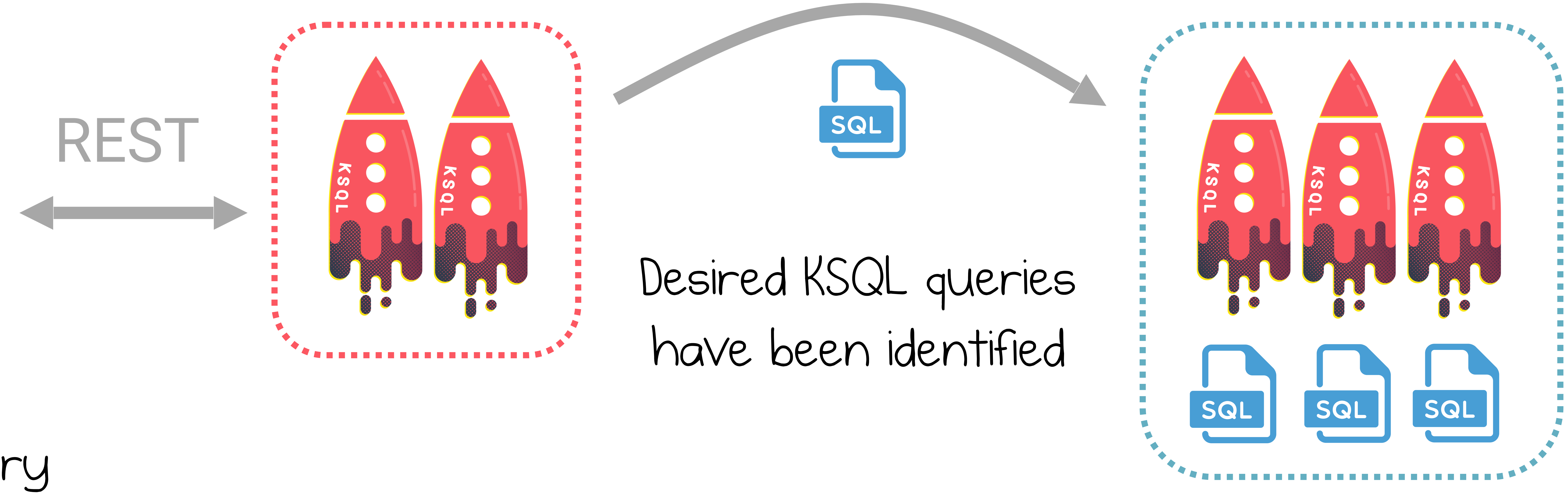
Headless KSQL
for Production



DEVELOPMENT >
KSQL

```
SELECT gender, regionid, timestamp, spaceid
FROM pageviews_regions
LIMIT 100;
```

```
{
    "gender": "female",
    "regionid": 296,
```



Desired KSQL queries have been identified

"Hmm, let me try out this idea…"

KSQL for Real-Time Monitoring

- Log data monitoring, tracking and alerting
- syslog data
- Sensor / IoT data

```
CREATE STREAM SYSLOG_INVALID_USERS AS
SELECT HOST, MESSAGE
FROM SYSLOG
WHERE MESSAGE LIKE '%Invalid user%';
```

http://cnfl.io/syslogs-filtering / http://cnfl.io/syslog-alerting

KSQL for Anomaly Detection

Identifying patterns or anomalies in real-time data, surfaced in milliseconds

```
CREATE TABLE possible_fraud AS
  SELECT card_number, count(*)
    FROM authorization_attempts
    WINDOW TUMBLING (SIZE 5 SECONDS)
    GROUP BY card_number
    HAVING count(*) > 3;
```

KSQL for Streaming ETL

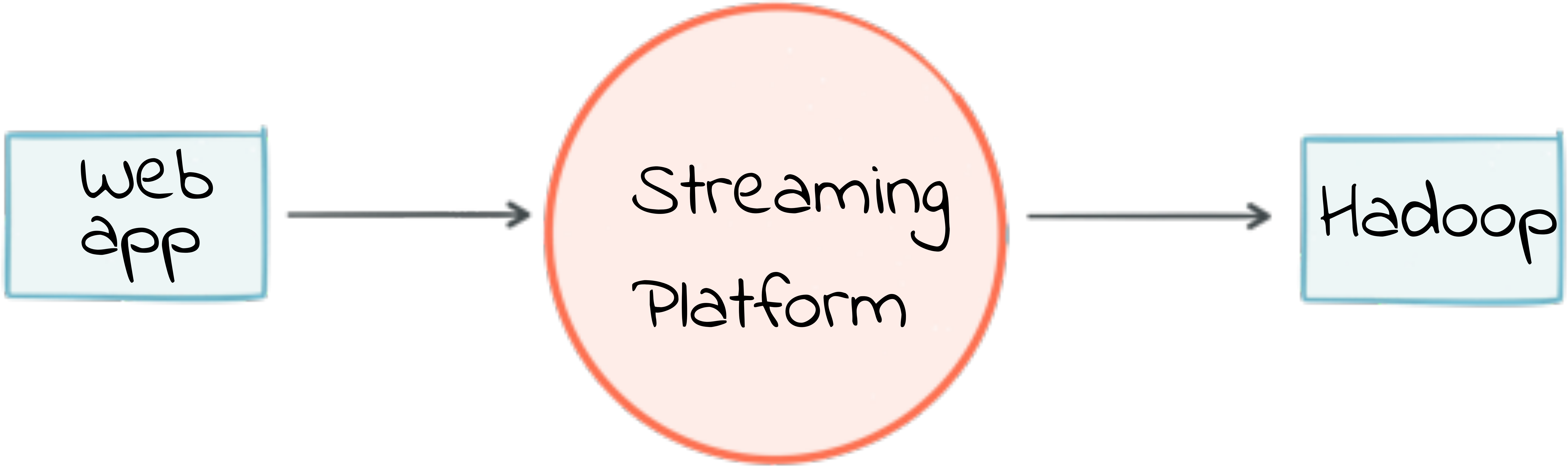
Joining, filtering, and aggregating streams of event data

```
CREATE STREAM vip_actions AS
  SELECT userid, page, action
  FROM clickstream c
  LEFT JOIN users u
    ON c.userid = u.user_id
  WHERE u.level = 'Platinum';
```

“

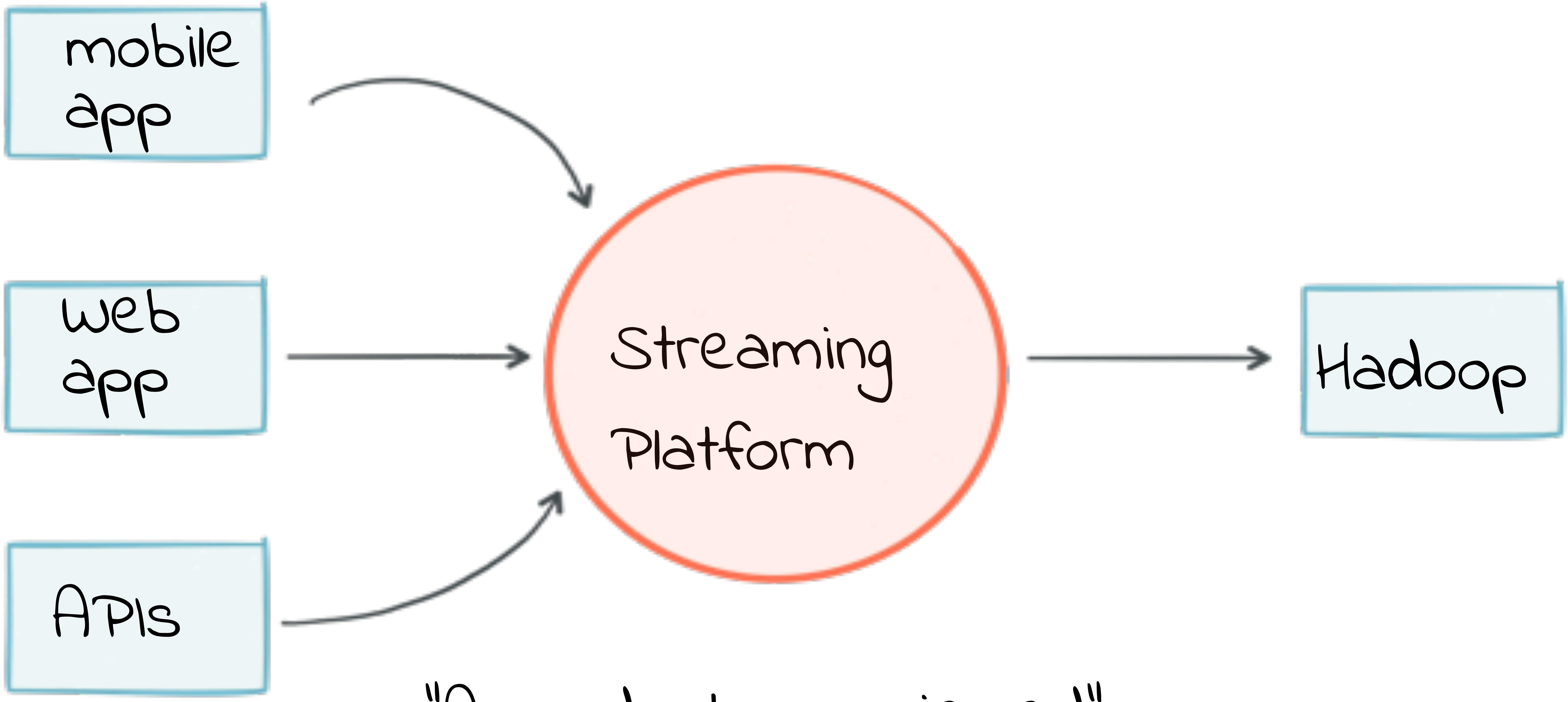
What Problems does Kafka Solve?

Event-Centric Thinking



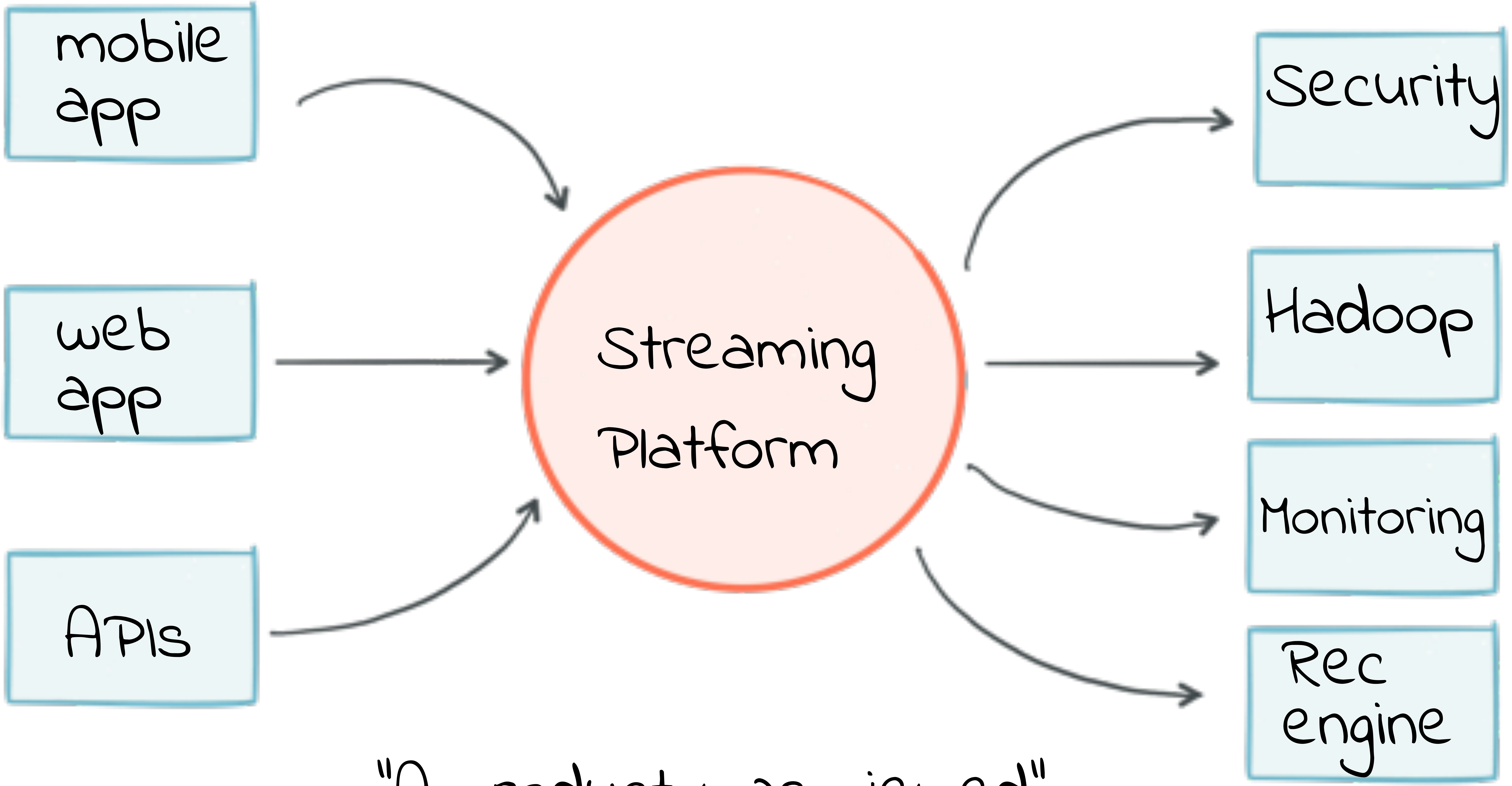
"A product was viewed"

Event-Centric Thinking



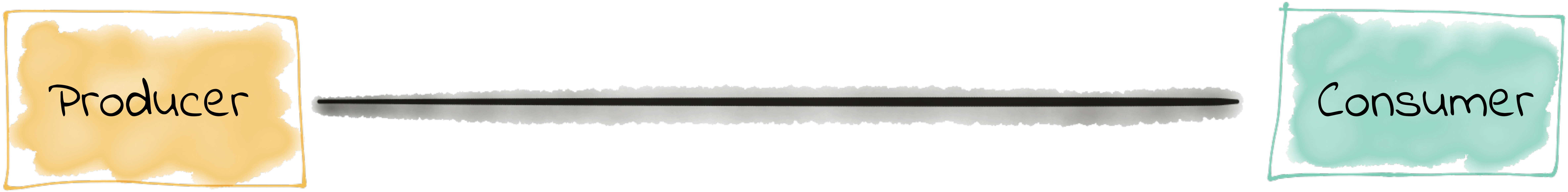
"A product was viewed"

Event-Centric Thinking

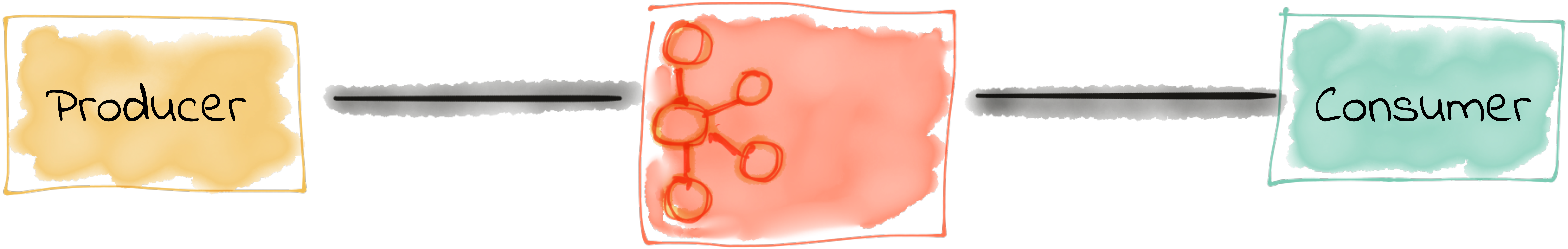


"A product was viewed"

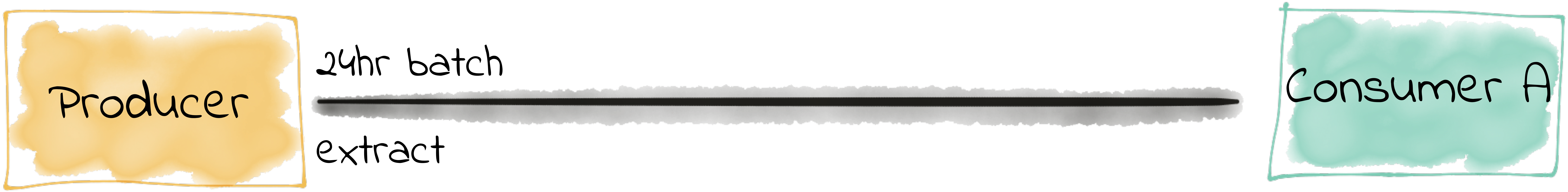
System Availability and Event Buffering



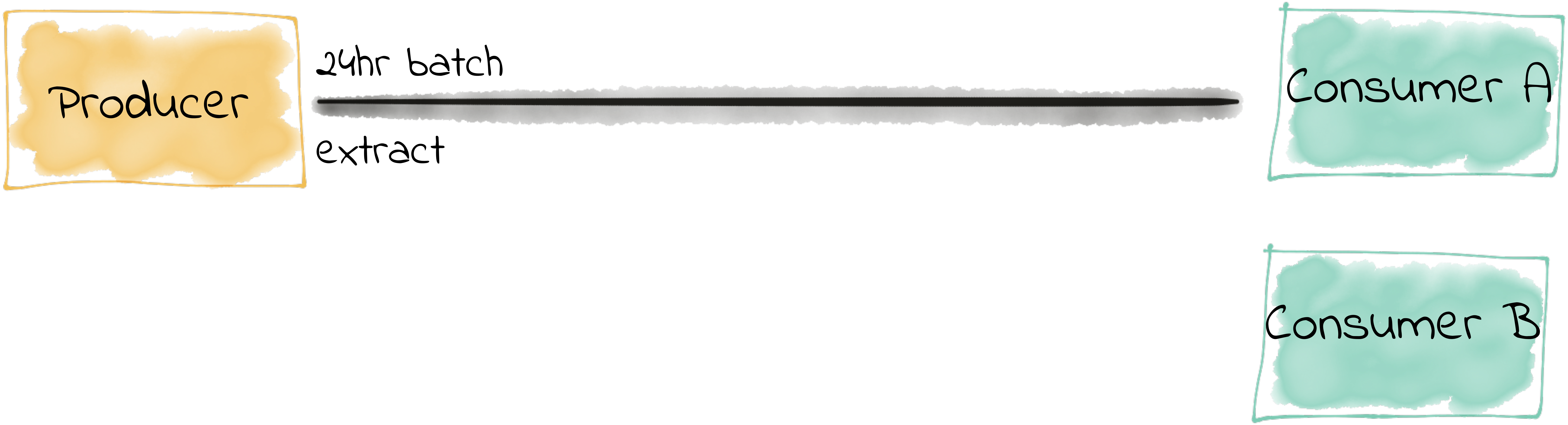
System Availability and Event Buffering



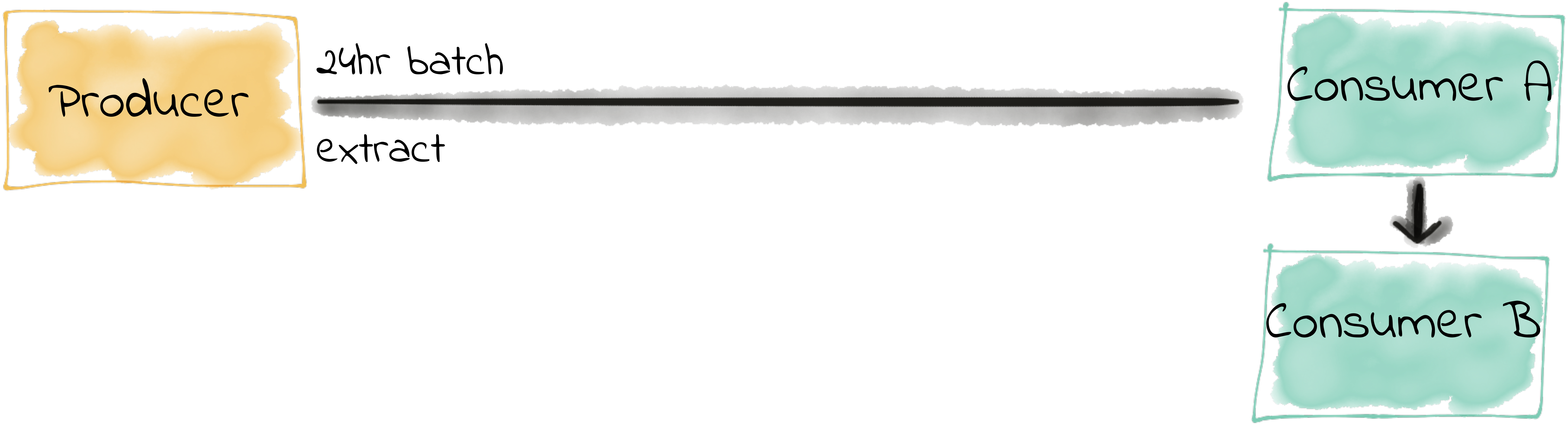
Varying Latency Requirements / Batch vs Stream



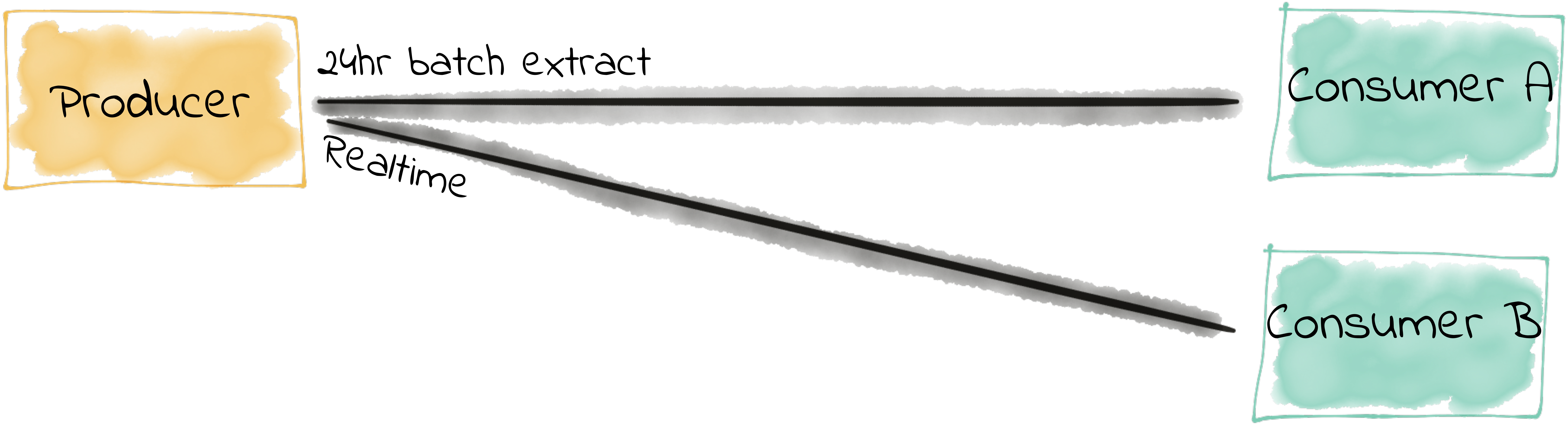
Varying Latency Requirements / Batch vs Stream



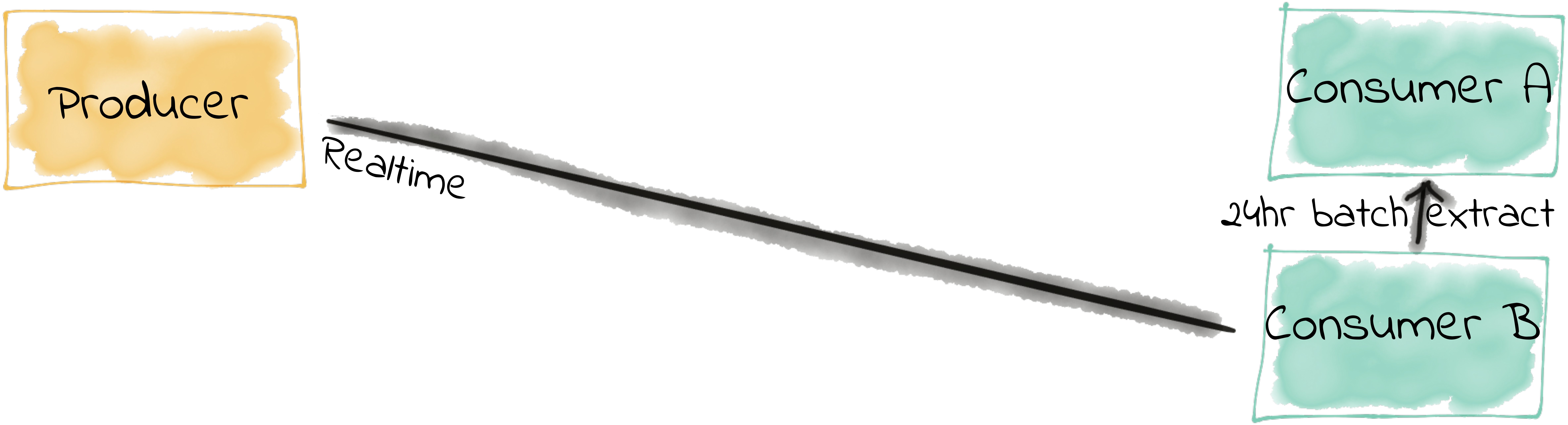
Varying Latency Requirements / Batch vs Stream



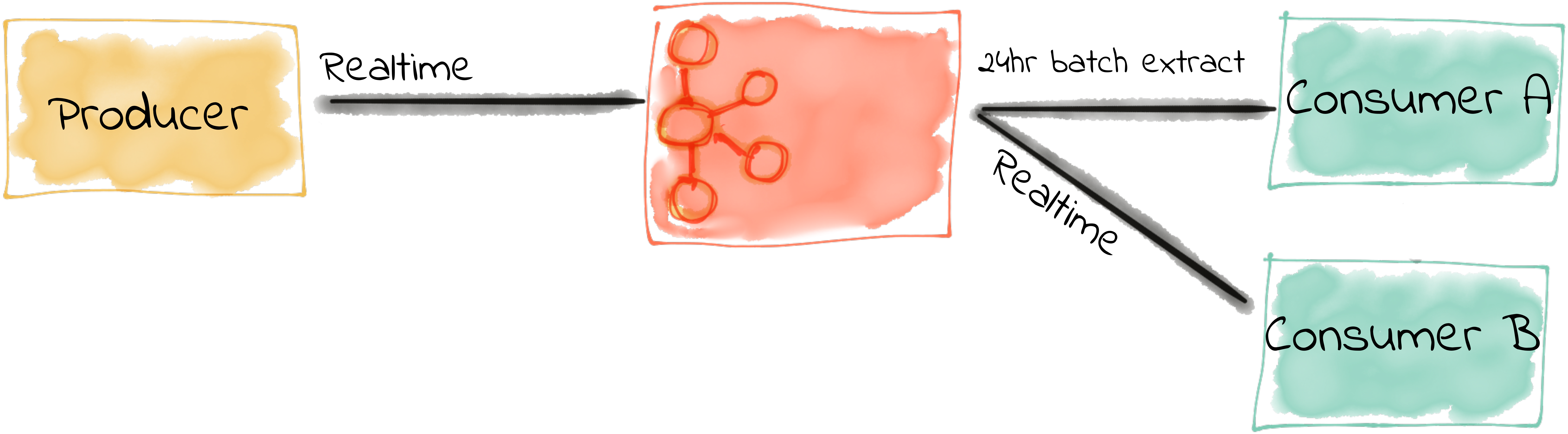
Varying Latency Requirements / Batch vs Stream



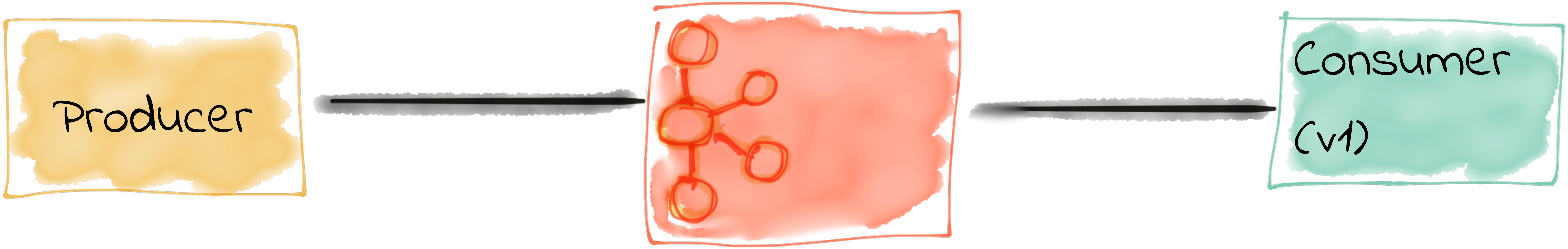
Varying Latency Requirements / Batch vs Stream



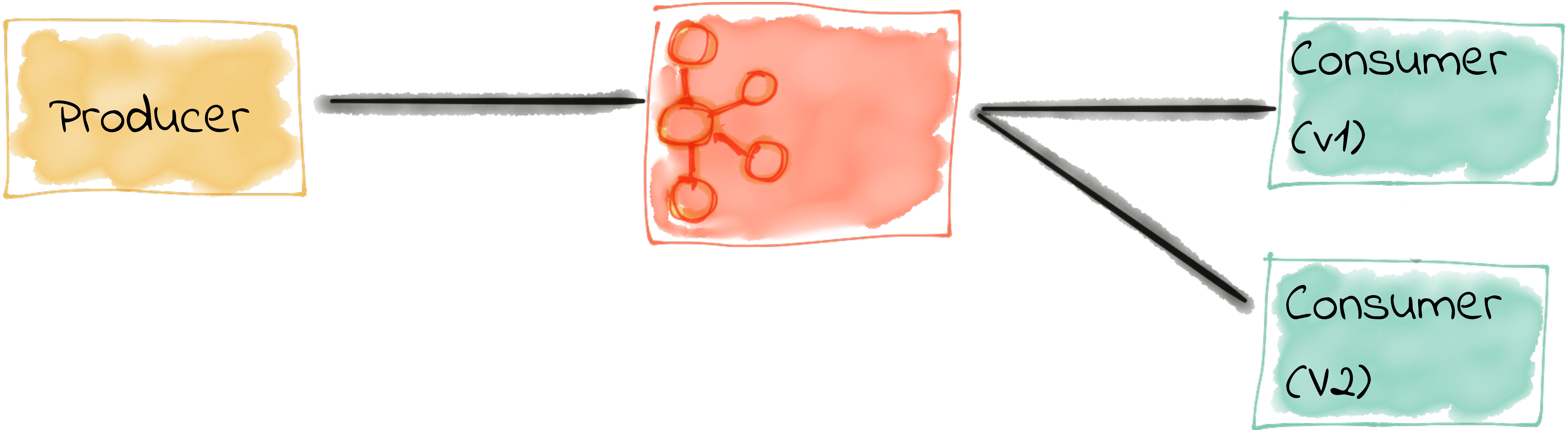
Varying Latency Requirements / Batch vs Stream



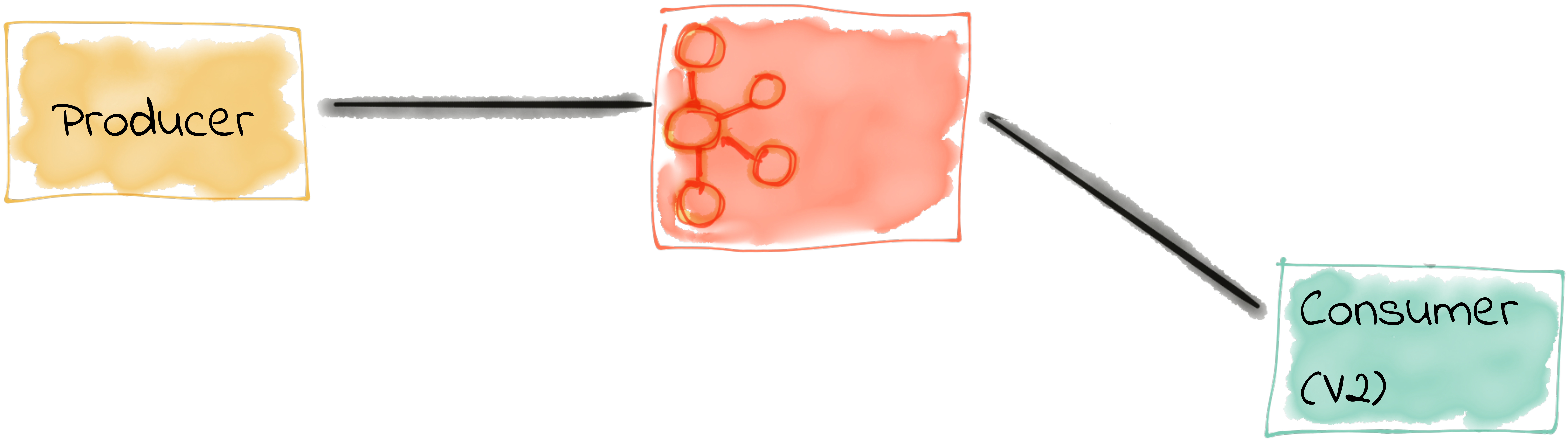
Technology & Code/Algo version Changes



Technology & Code/Algo version Changes



Technology & Code/Algo version Changes



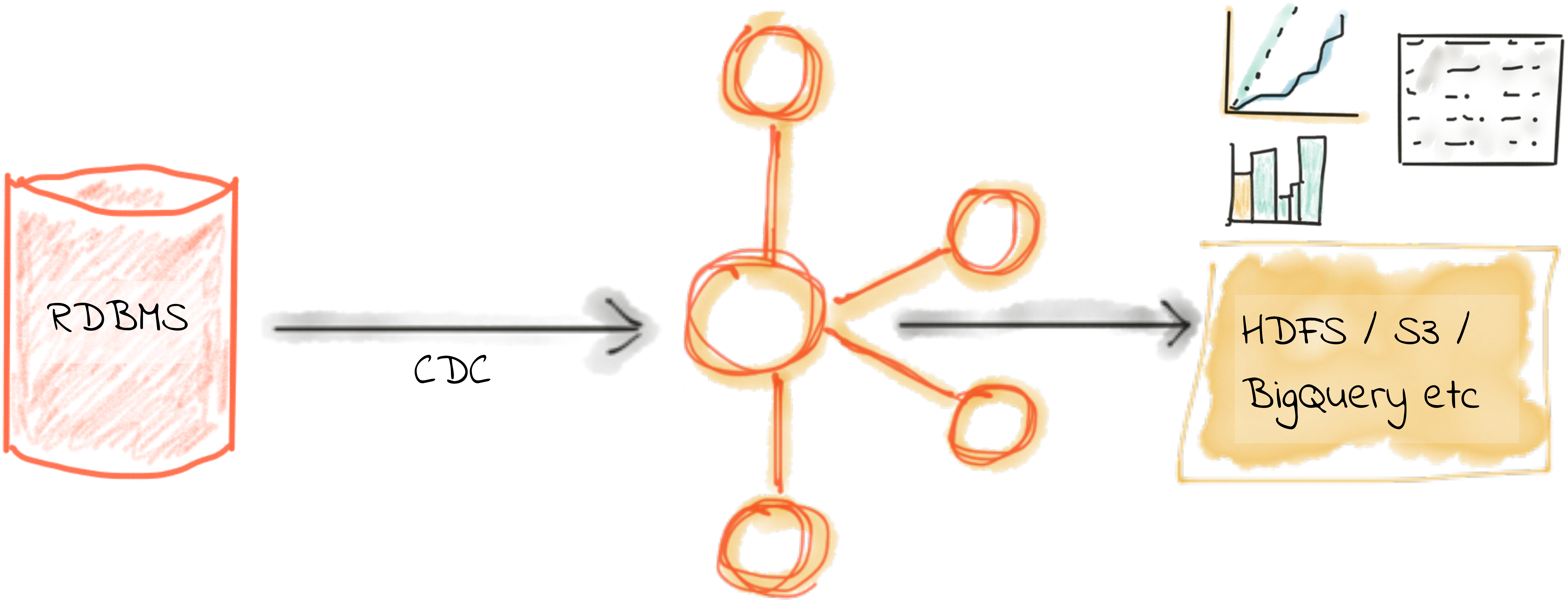
“Architectural Patterns with Apache Kafka

Photo by Christopher Burns on Unsplash

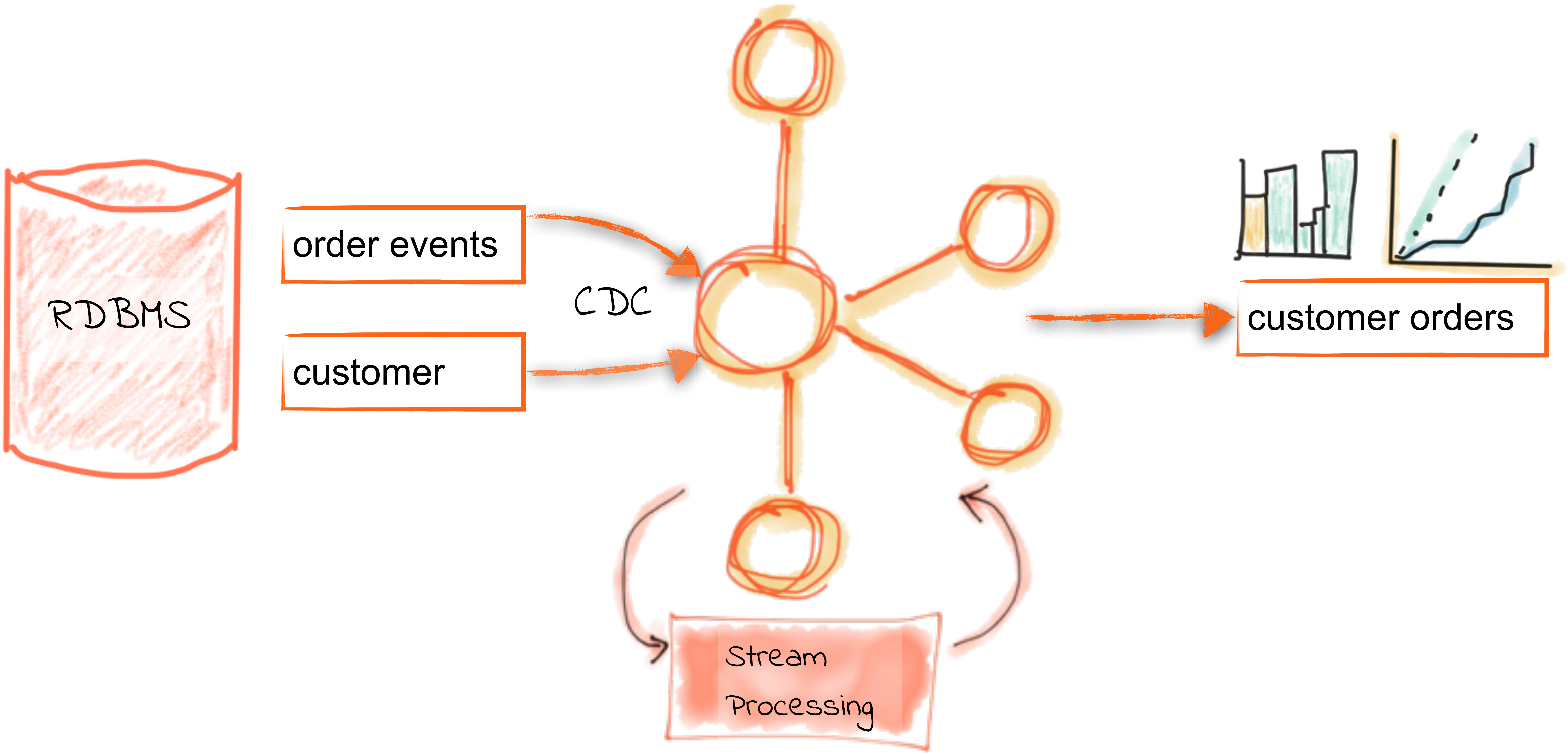
Building for the Future

Tightly-coupled = Inflexible

Analytics - Database Offload



Stream Processing with Apache Kafka and KSQL

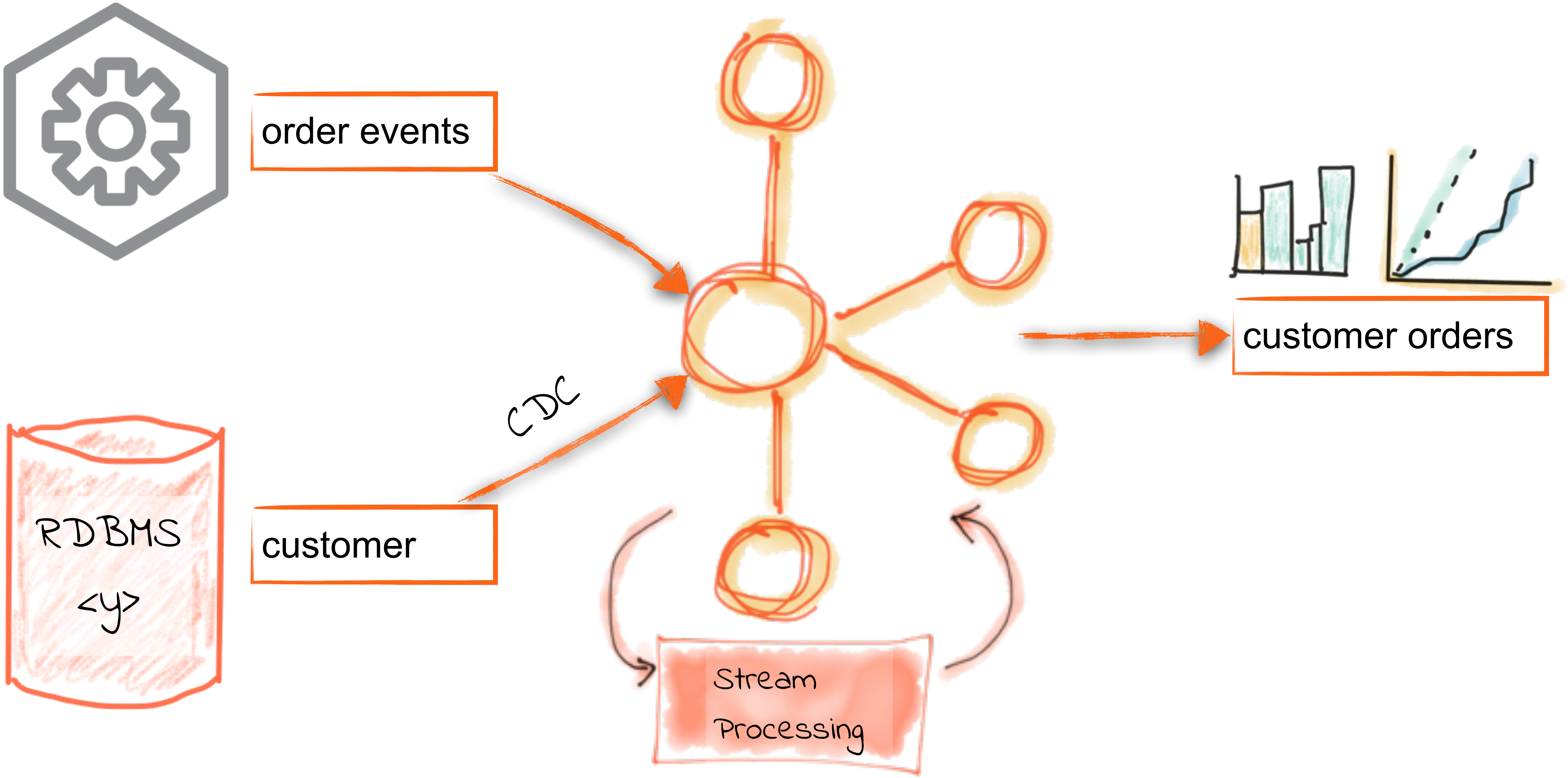


order events

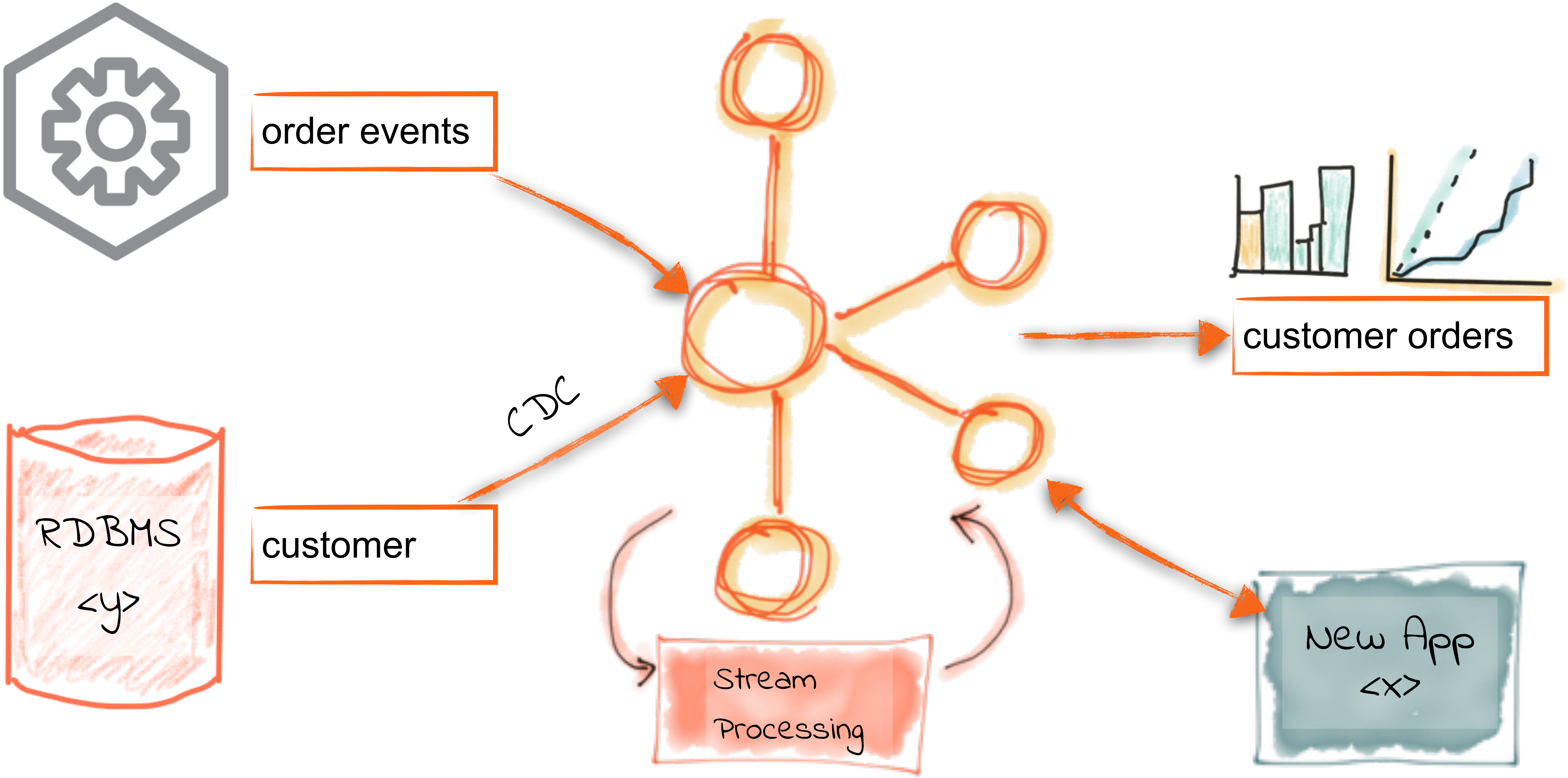
customer

customer orders

Real-time Event Stream Enrichment



Transform Once, Use Many

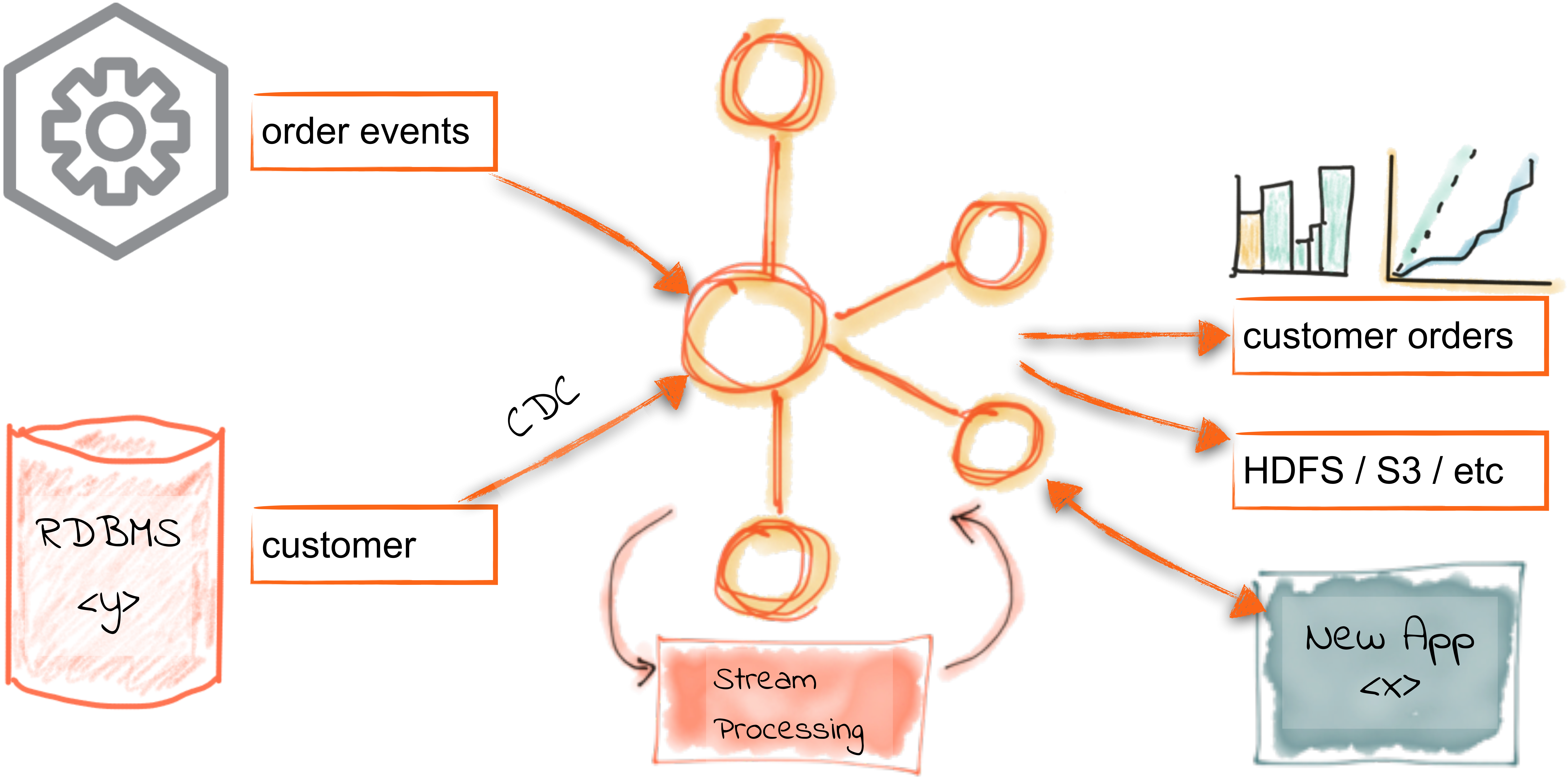


order events

customer

customer orders

Transform Once, Use Many



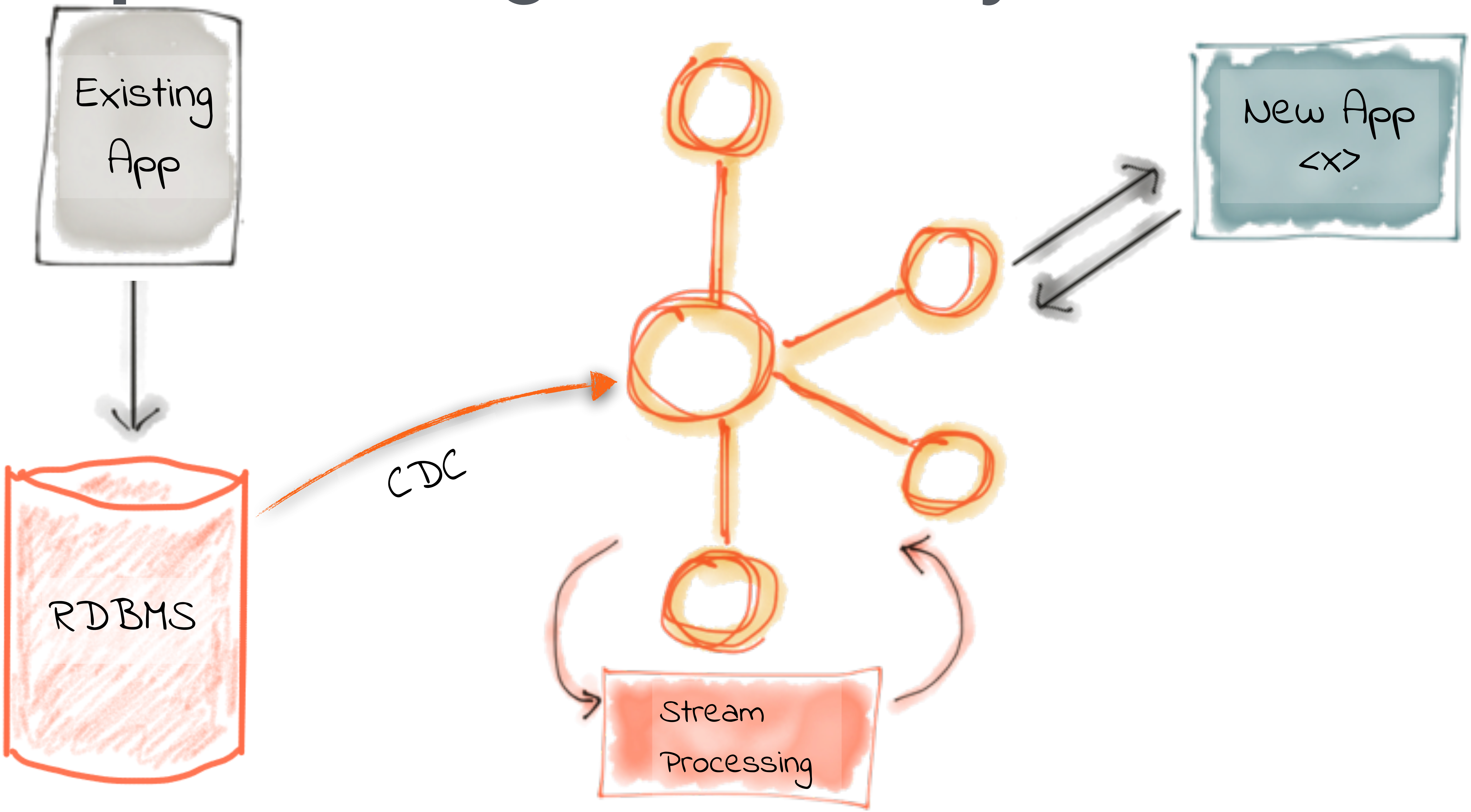
order events

customer

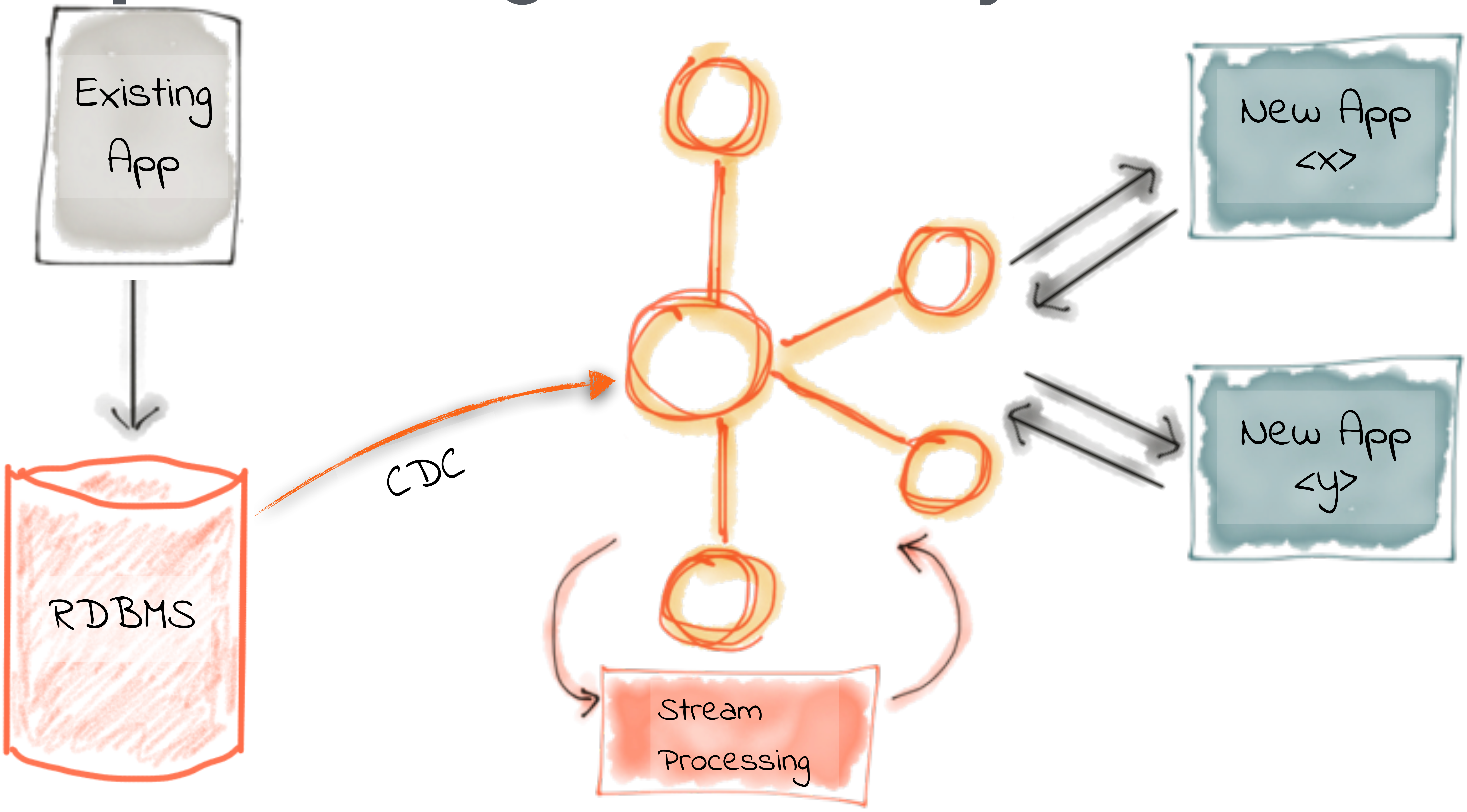
customer orders

HDFS / S3 / etc

Evolve processing from old systems to new



Evolve processing from old systems to new

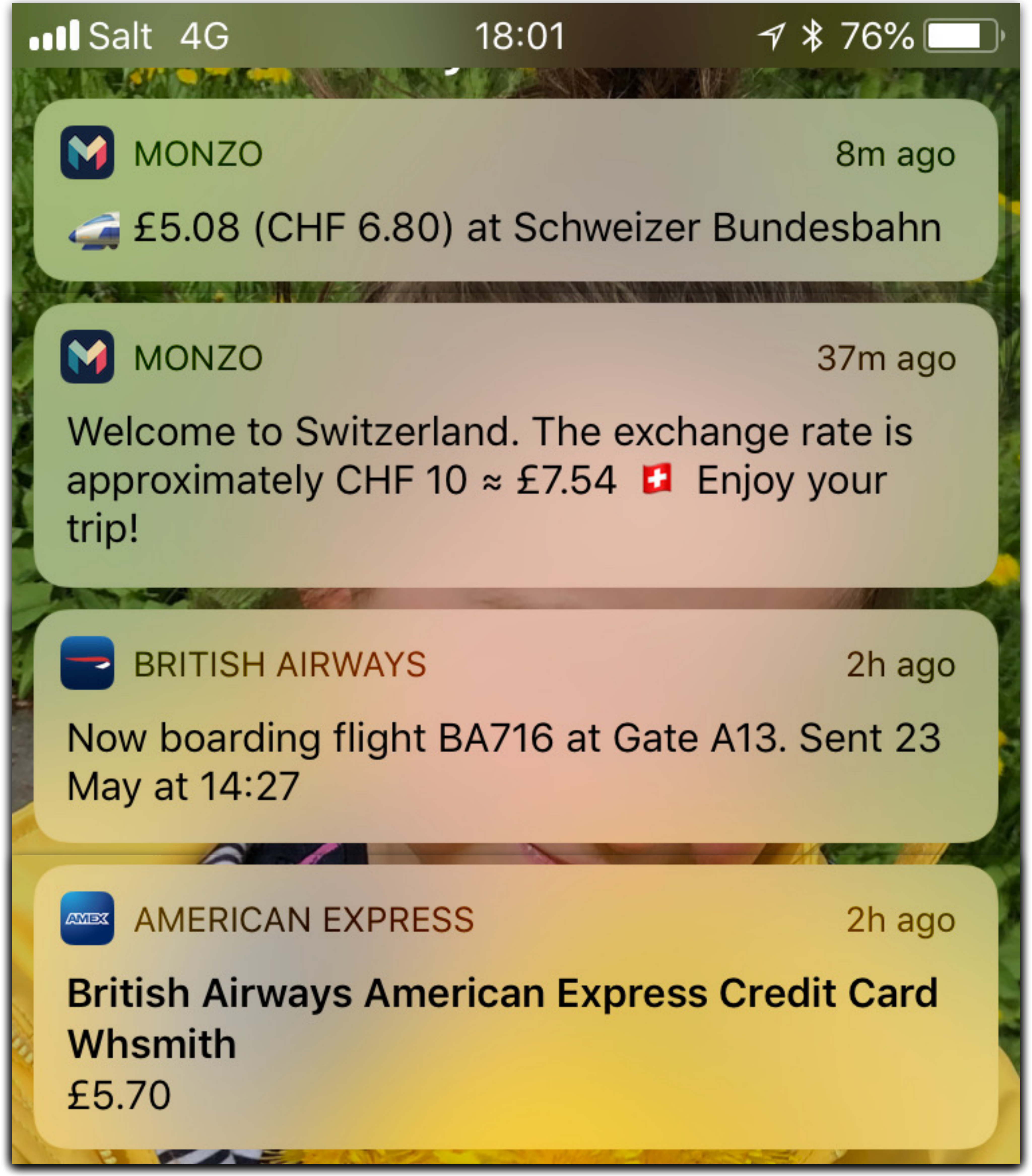


Want your data anytime SOON ?

You say that like "latency" is a synonym for "evil"

Batch is Latency built in by Design

It's all about the Events!



Salt 4G 18:01 76%

MONZO 8m ago

£5.08 (CHF 6.80) at Schweizer Bundesbahn

MONZO 37m ago

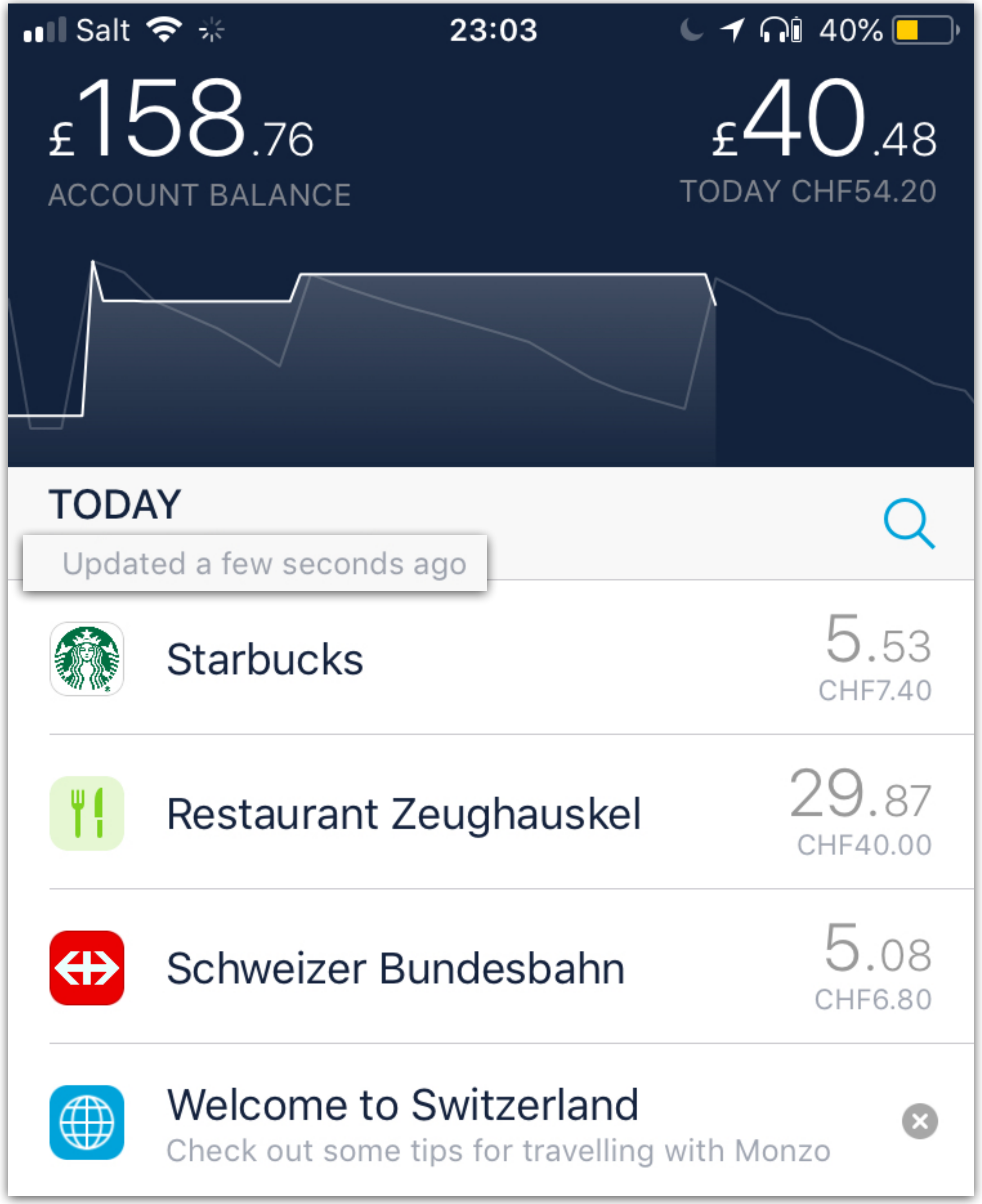
Welcome to Switzerland. The exchange rate is approximately CHF 10 ≈ £7.54 🇨🇭 Enjoy your trip!

BRITISH AIRWAYS 2h ago

Now boarding flight BA716 at Gate A13. Sent 23 May at 14:27

AMERICAN EXPRESS 2h ago

British Airways American Express Credit Card Whsmith
£5.70



Salt 23:03 40%

£158.76
ACCOUNT BALANCE

£40.48
TODAY CHF54.20

TODAY

Updated a few seconds ago

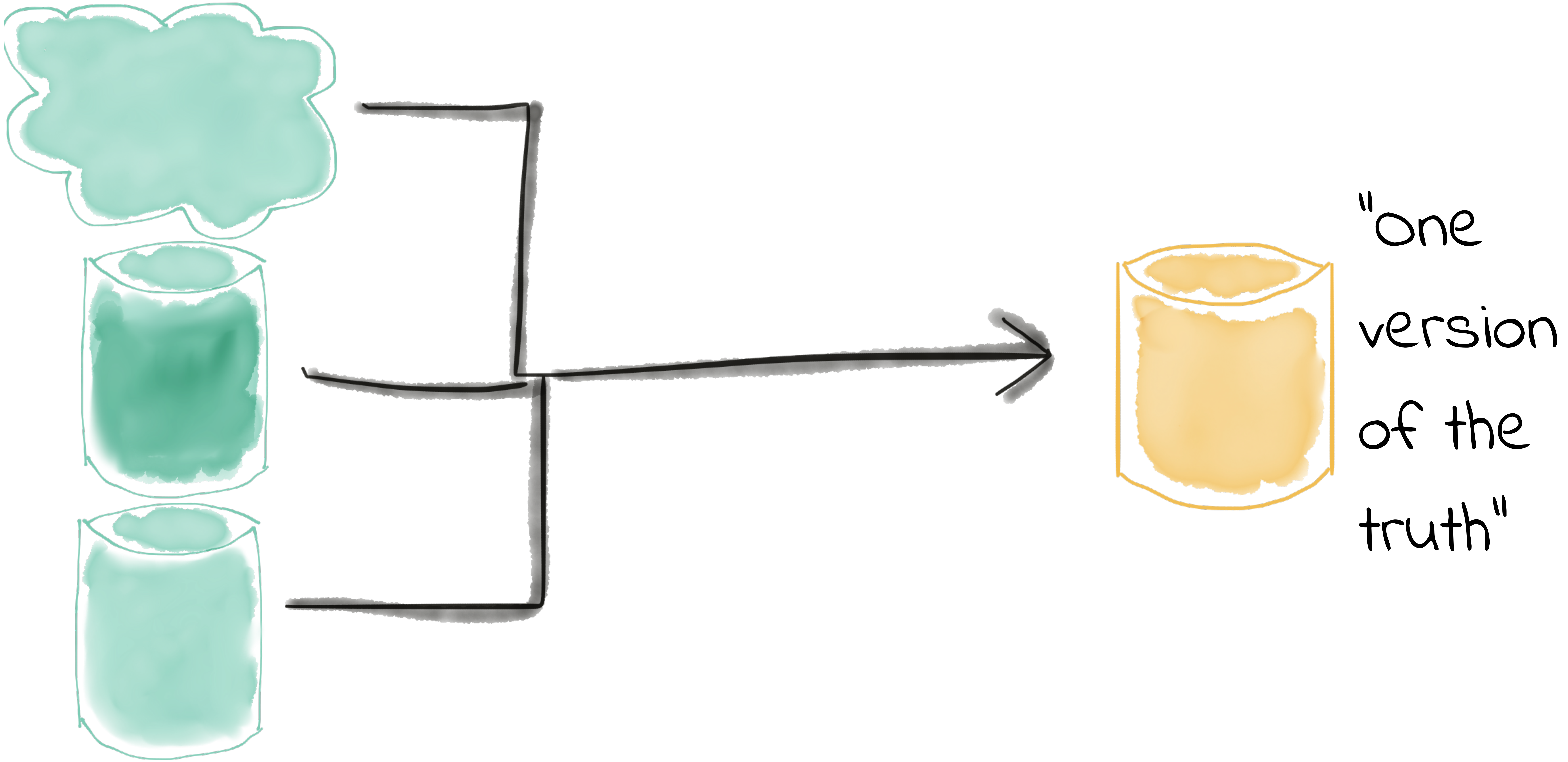
Starbucks	5.53 CHF7.40
Restaurant Zeughauskel	29.87 CHF40.00
Schweizer Bundesbahn	5.08 CHF6.80

Welcome to Switzerland
Check out some tips for travelling with Monzo

“

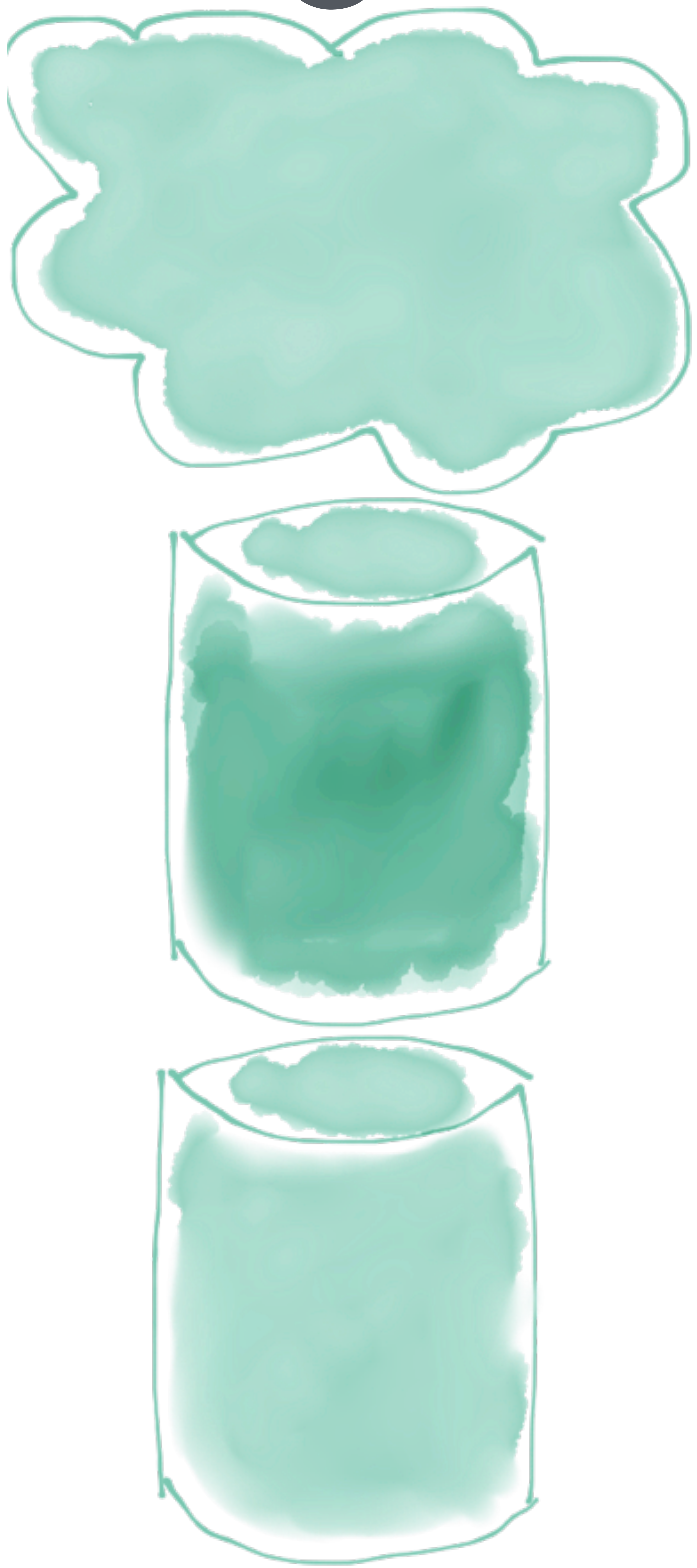
So…Analytics and Kafka

The Vision!



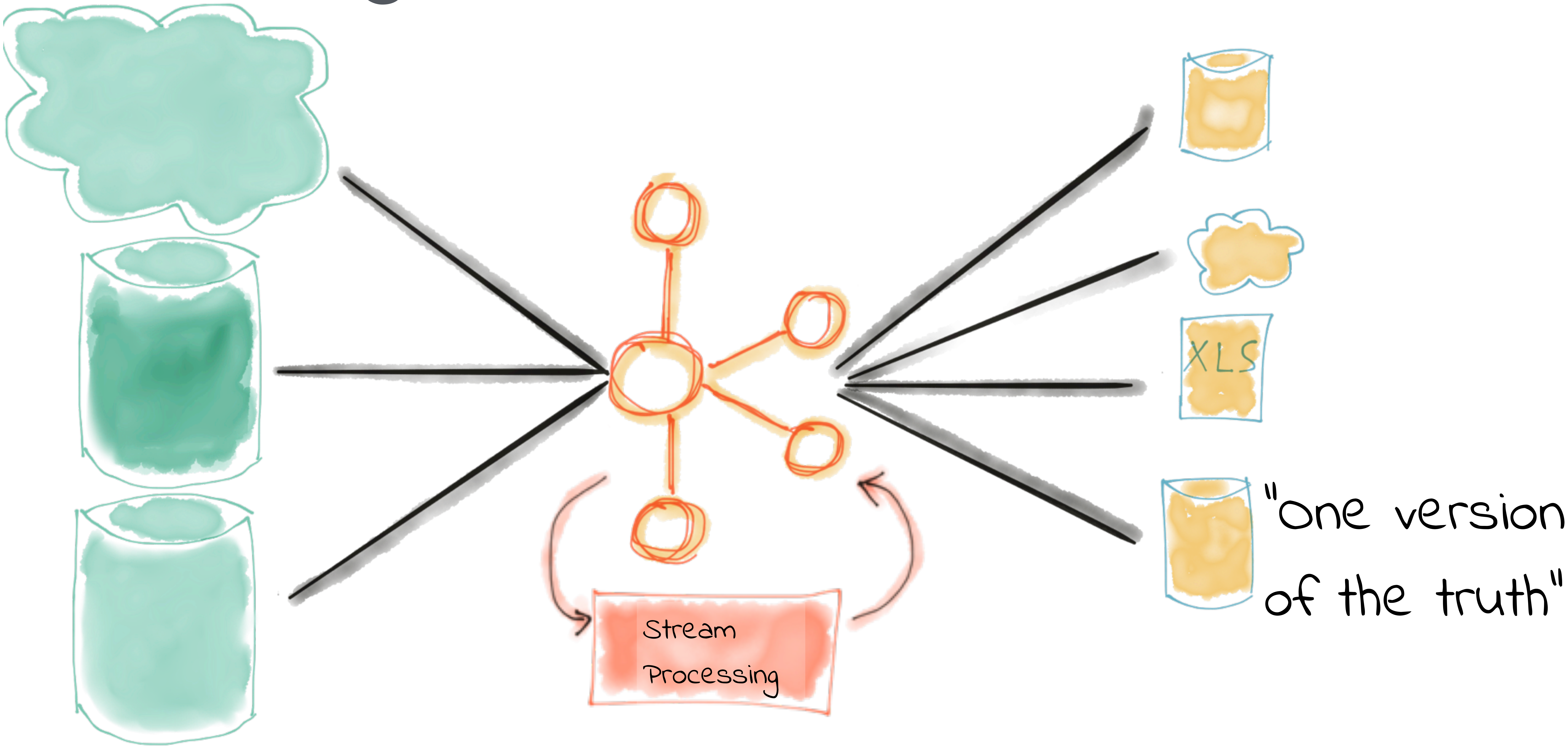
The Reality...

Pragmatism is…

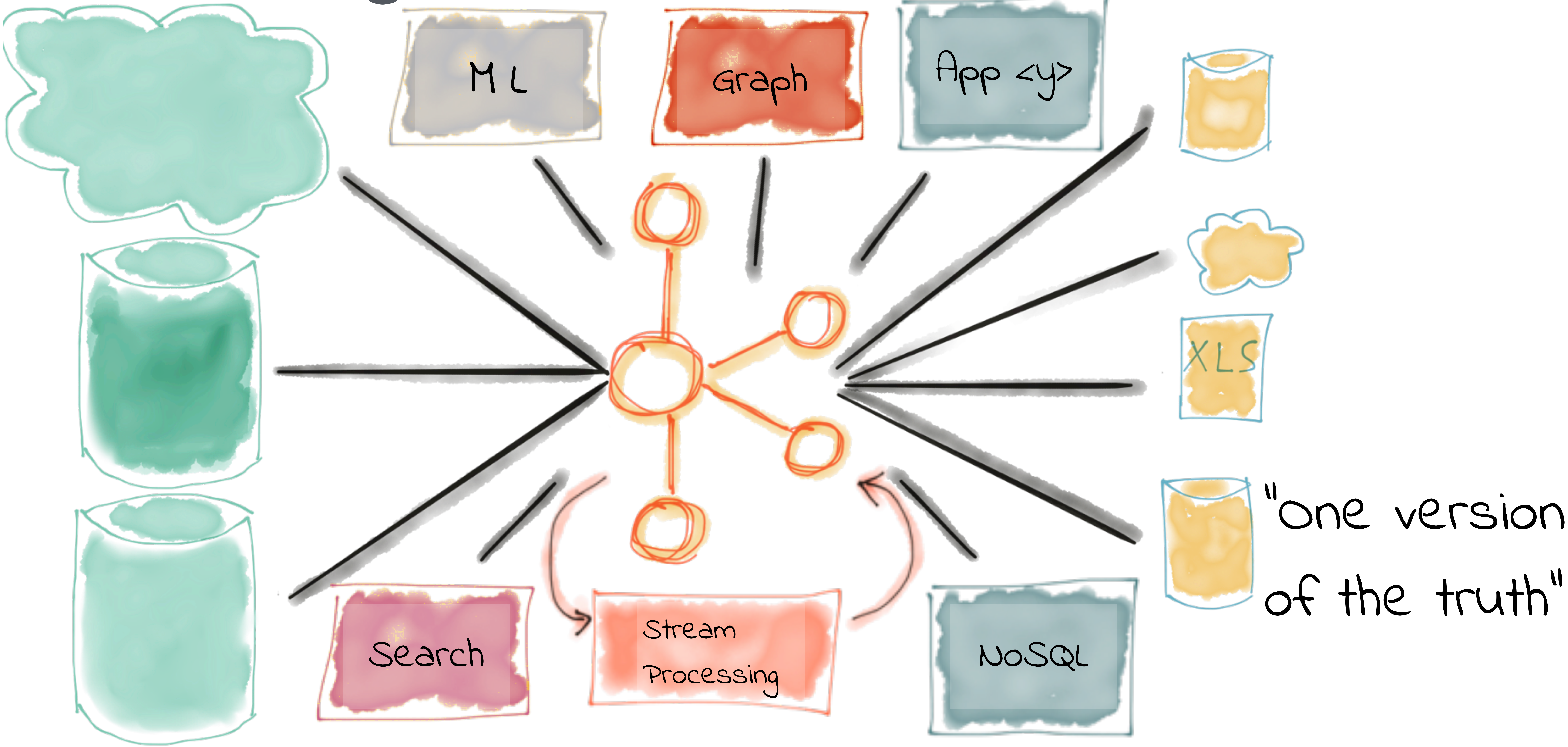


"one version of the truth"

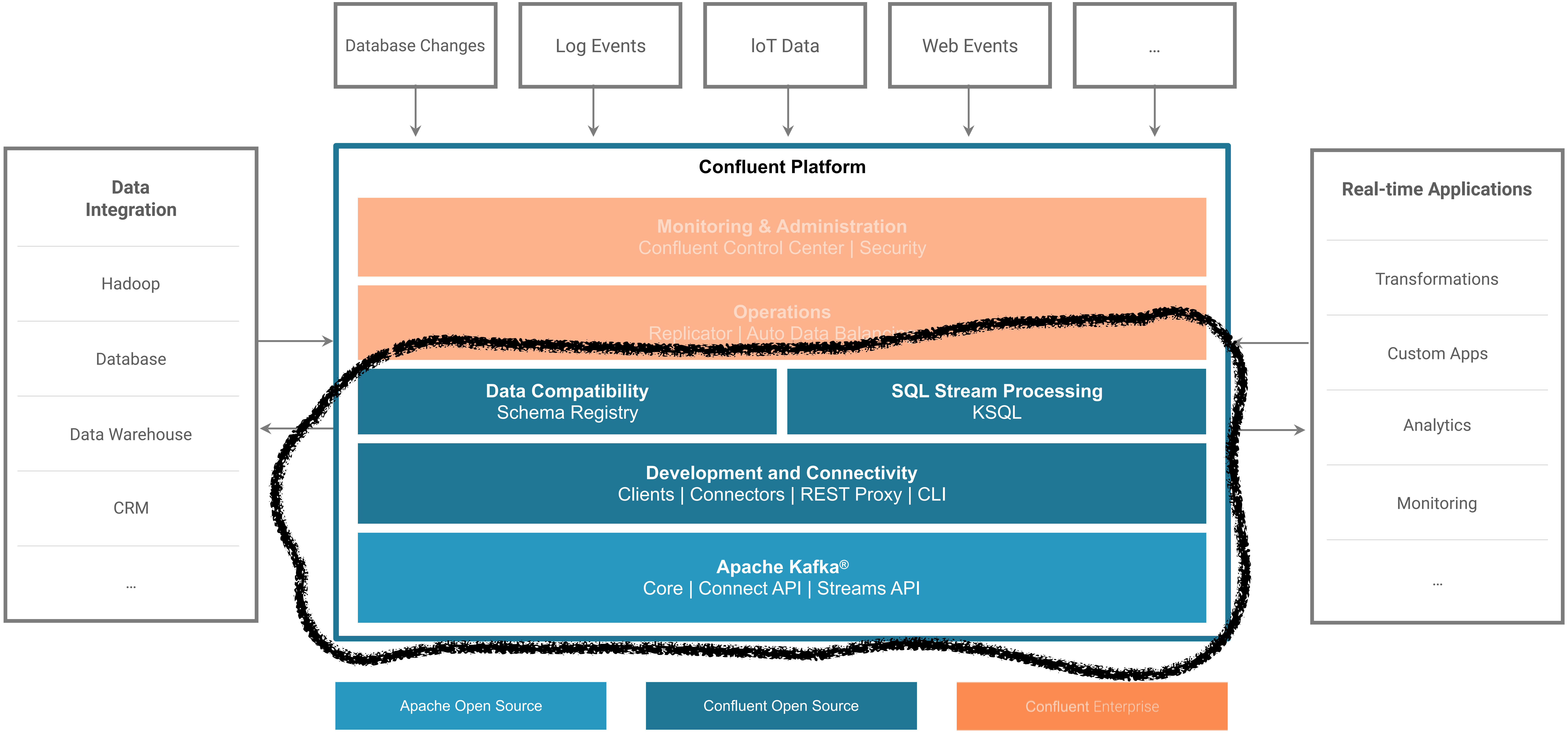
Streaming Platform



Streaming Platform



Confluent Open Source :
Apache Kafka with a bunch of cool stuff! For free!



Database Changes

Log Events

IoT Data

Web Events

...

Data Integration

Hadoop

Database

Data Warehouse

CRM

...

Confluent Platform

Monitoring & Administration
Confluent Control Center | Security

Operations
Replicator | Auto Data Balancing

Data Compatibility
Schema Registry

SQL Stream Processing
KSQL

Development and Connectivity
Clients | Connectors | REST Proxy | CLI

Apache Kafka®
Core | Connect API | Streams API

Real-time Applications

Transformations

Custom Apps

Analytics

Monitoring

...

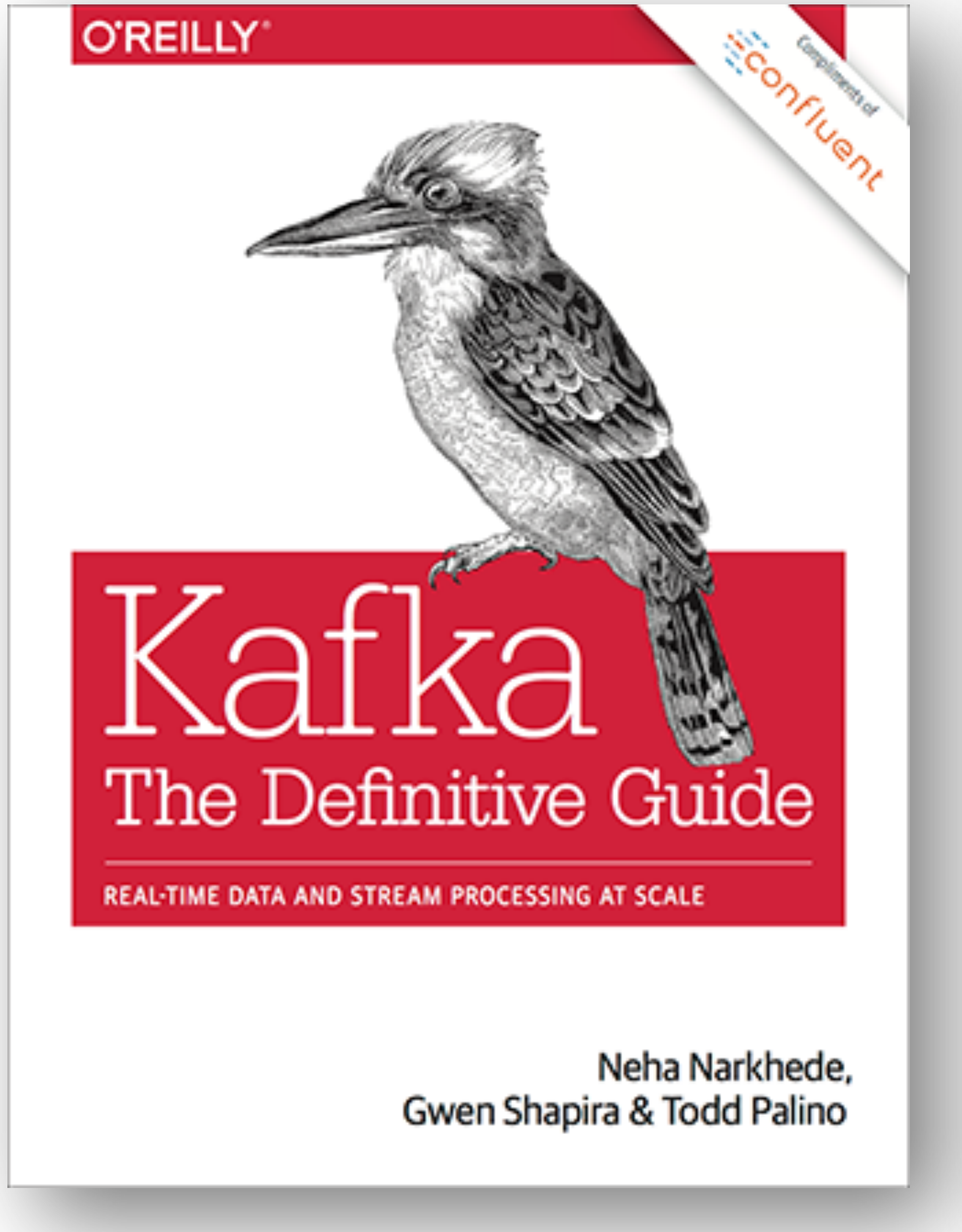
Apache Open Source

Confluent Open Source

Confluent Enterprise

Free Books!

https://www.confluent.io/apache-kafka-stream-processing-book-bundle



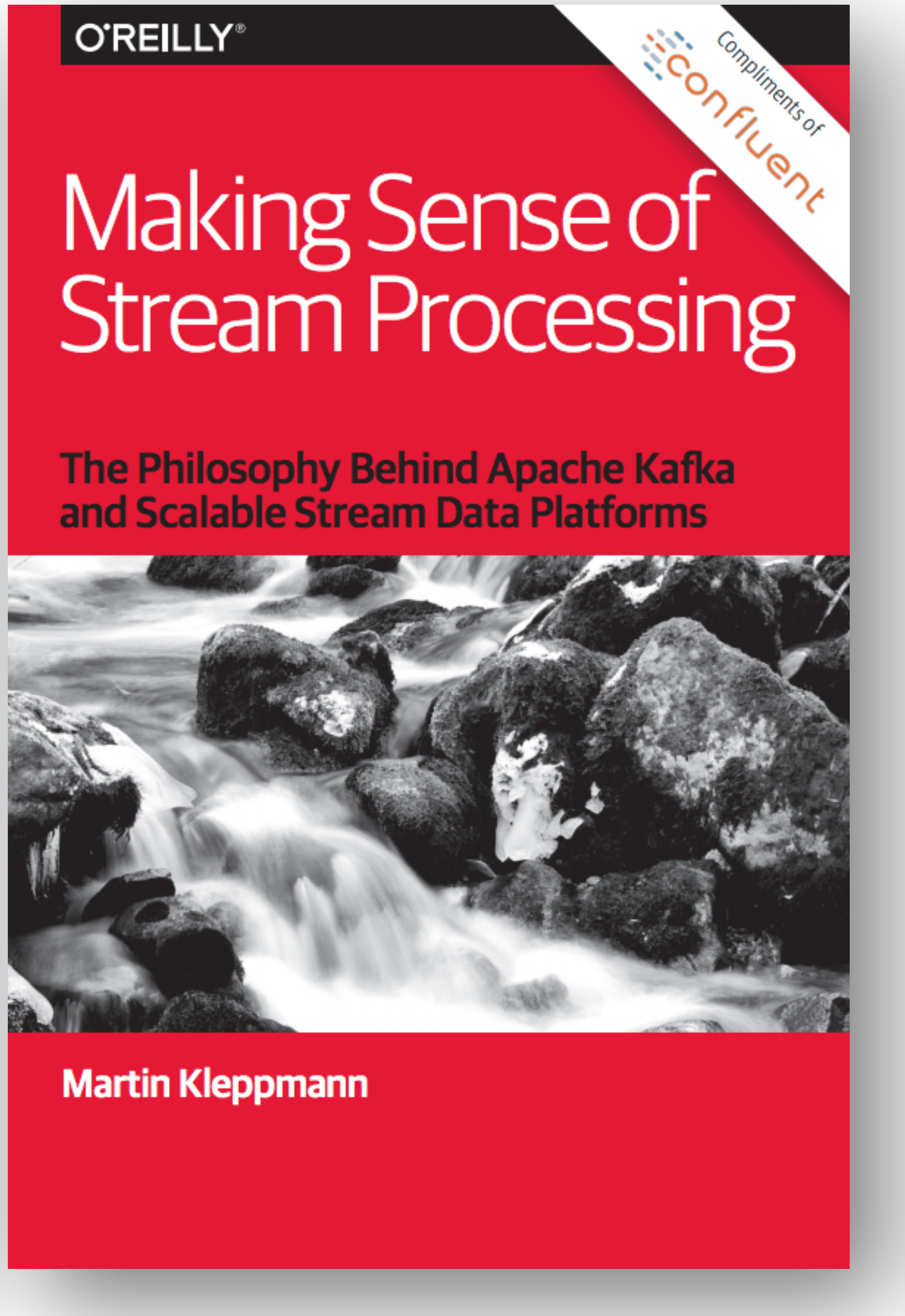
O'REILLY

Compliments of confluent

Kafka
The Definitive Guide

REAL-TIME DATA AND STREAM PROCESSING AT SCALE

Neha Narkhede,
Gwen Shapira & Todd Palino



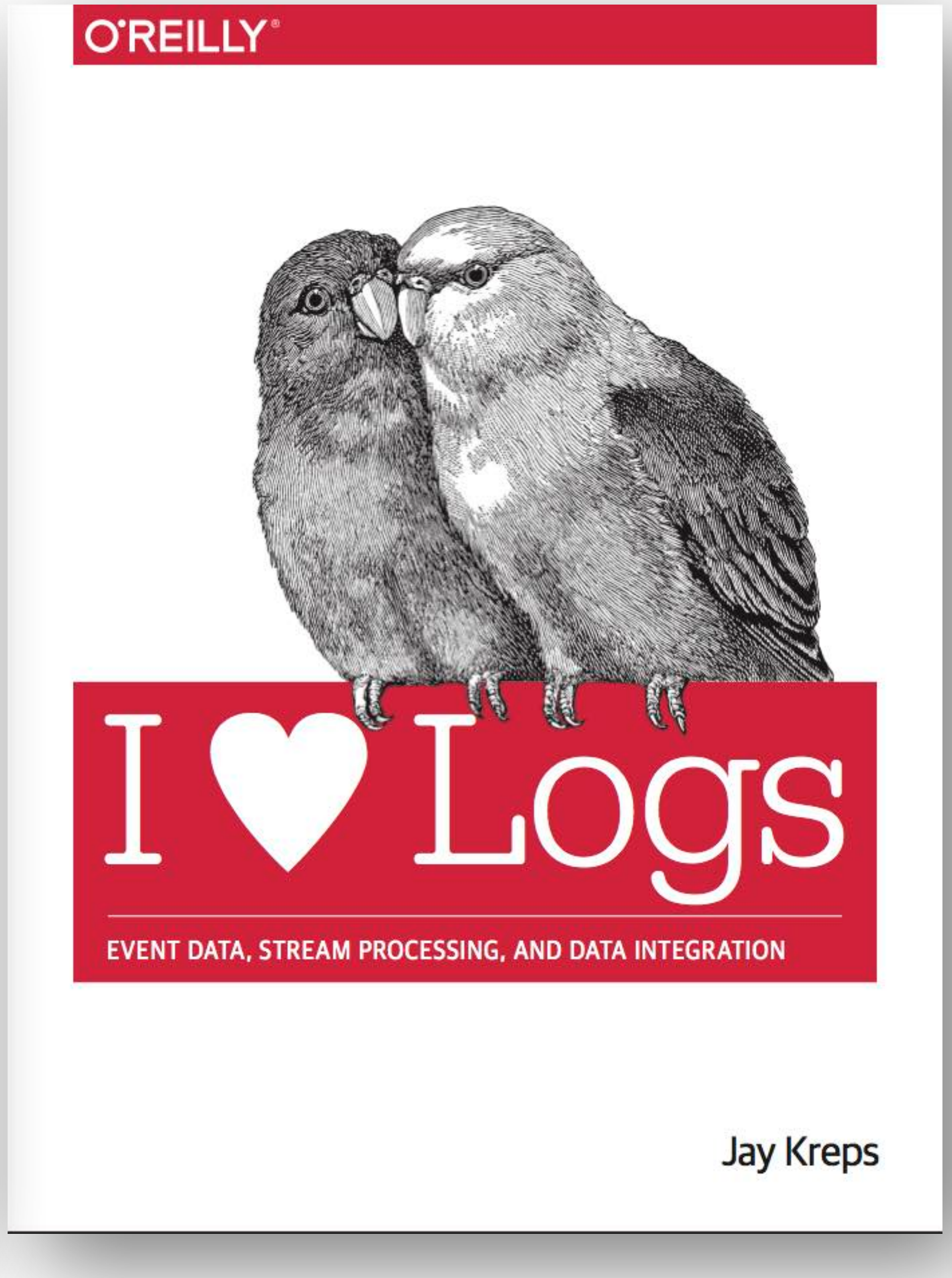
O'REILLY

Compliments of confluent

Making Sense of Stream Processing

The Philosophy Behind Apache Kafka and Scalable Stream Data Platforms

Martin Kleppmann

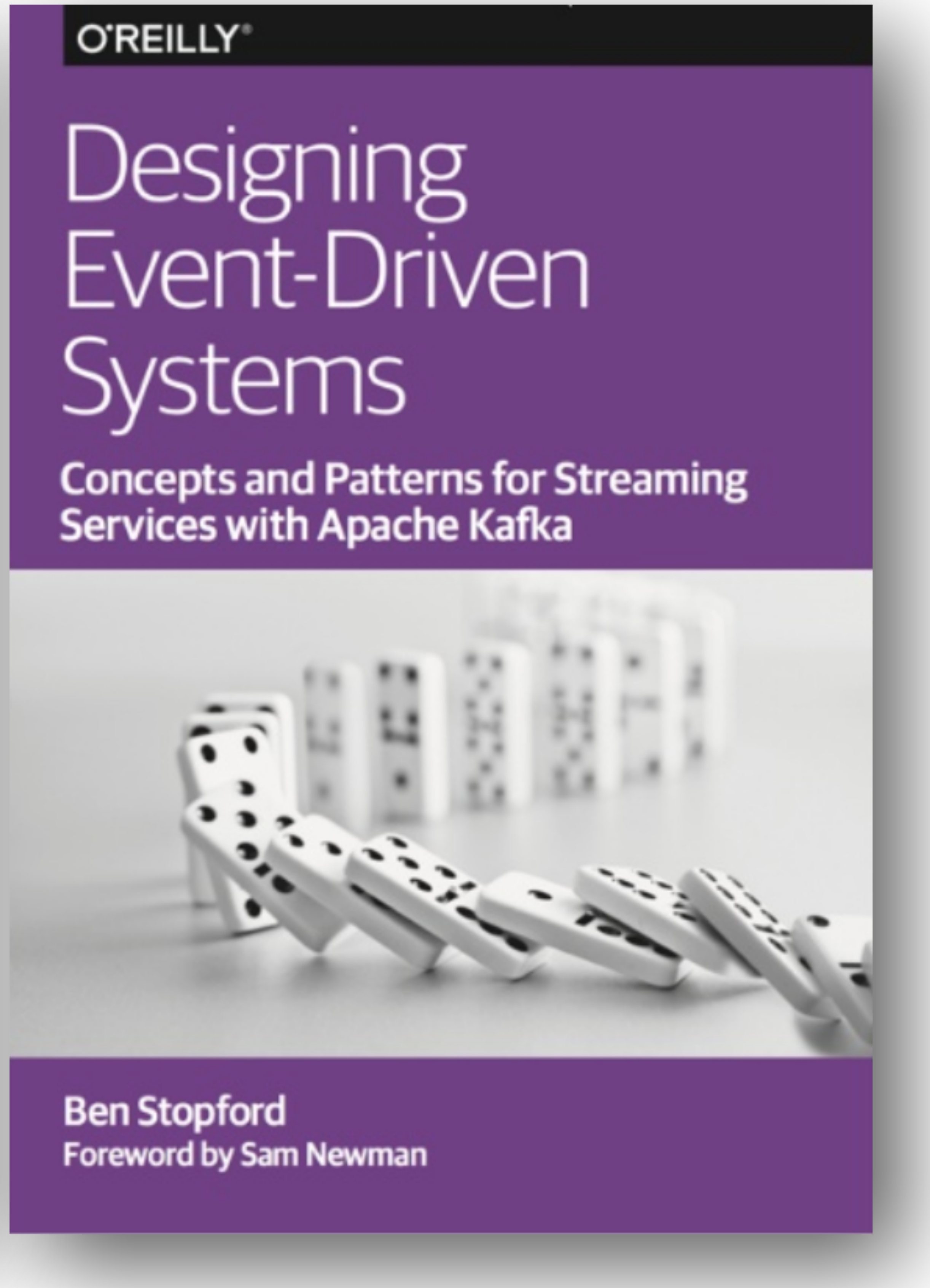


O'REILLY

I ♥ Logs

EVENT DATA, STREAM PROCESSING, AND DATA INTEGRATION

Jay Kreps



O'REILLY

Designing Event-Driven Systems

Concepts and Patterns for Streaming Services with Apache Kafka

Ben Stopford
Foreword by Sam Newman

Confluent Streaming Event, Munich

CONFLUENT
STREAMING
EVENT

München
09.Oktober 2018

Jetzt anmelden

http://cnfl.io/streaming-event-munich

https://www.confluent.io/download/

http://cnfl.io/slack

@rmoff
robin@confluent.io

#EOF