ETL 2.0

It's not just for data engineers anymore

#OReillySACon
@rmoff

Photo by rmoff

Photo by Khai Sze Ong on Unsplash


Photo by Rohit Tandon on Unsplash


Photo by Theodore Moore on Unsplash

Photo by Cristian Grecu on Unsplash

It used to be so simple

Photo by Patrick Fore on Unsplash

Photo by Filippo Mazzone on Unsplash

Photo by Tom Barrett on Unsplash

Photo by Kvalli on Unsplash


Batches and Buckets

Applications

Respond

→ an order was placed!

Analytics

Tell Us What Happened

→ how many orders were placed

Photo by Deva Darshan from Pexels


Photo by NASA on Unsplash

Events

Photo by Mark Kamalov on Unsplash

An event is both:

- Notification
- State transfer

A Customer Experience

A Sensor Reading

Databases

The Stream/Table Duality

Stream

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

Time

Table

Account ID	Balance
12345	€50

Account ID	Balance
12345	€75

Account ID	Balance
12345	€15

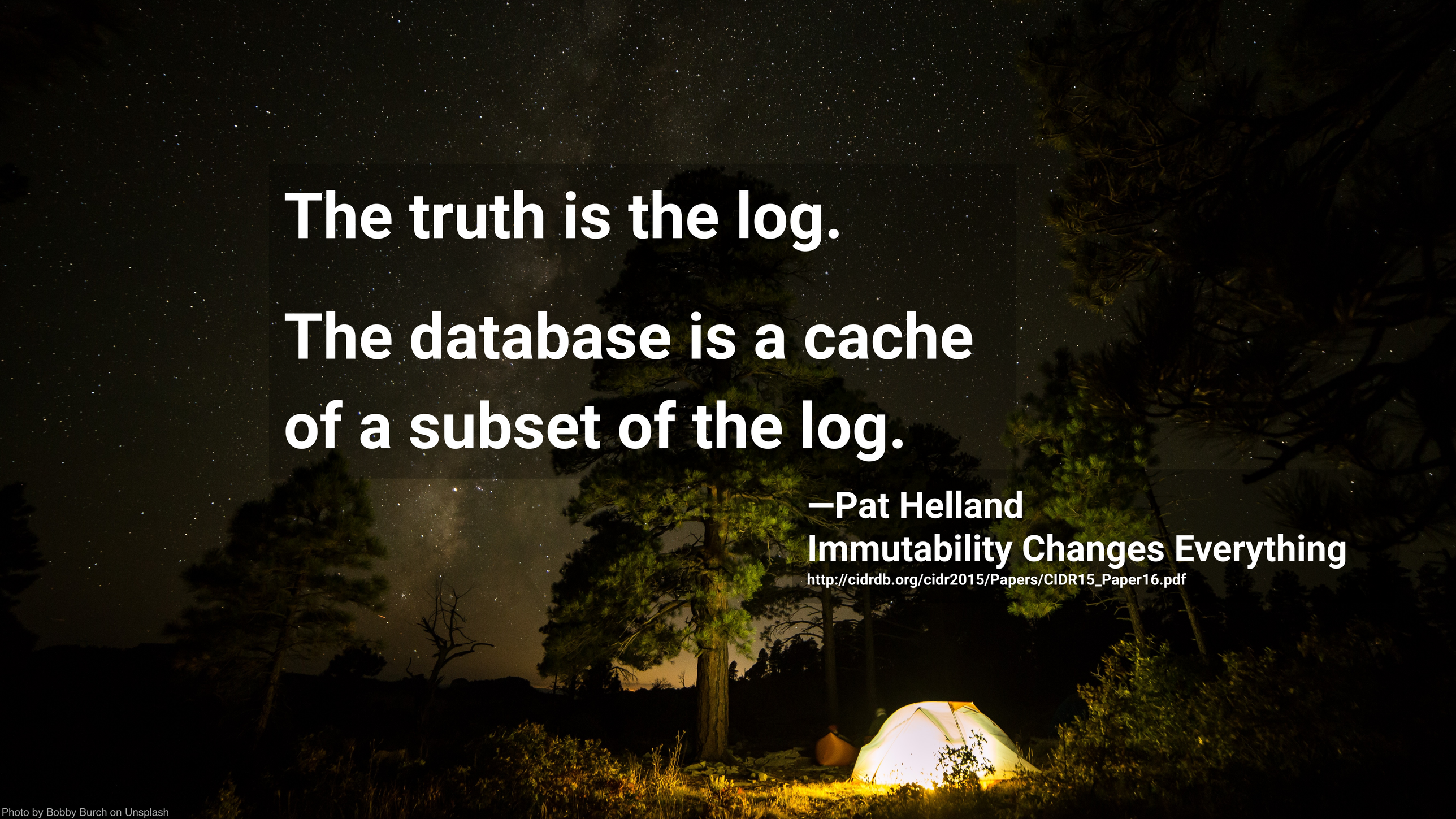
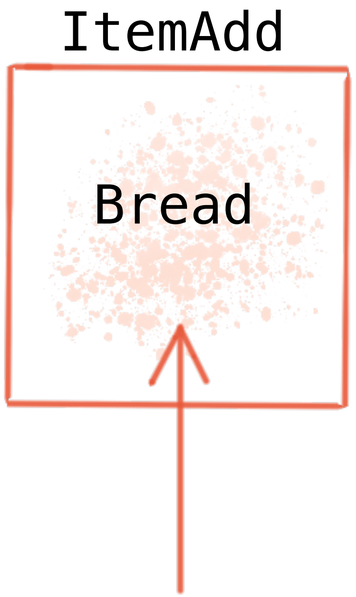
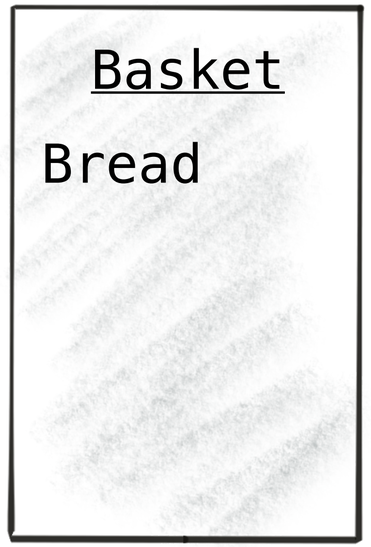


Photo by Bobby Burch on Unsplash

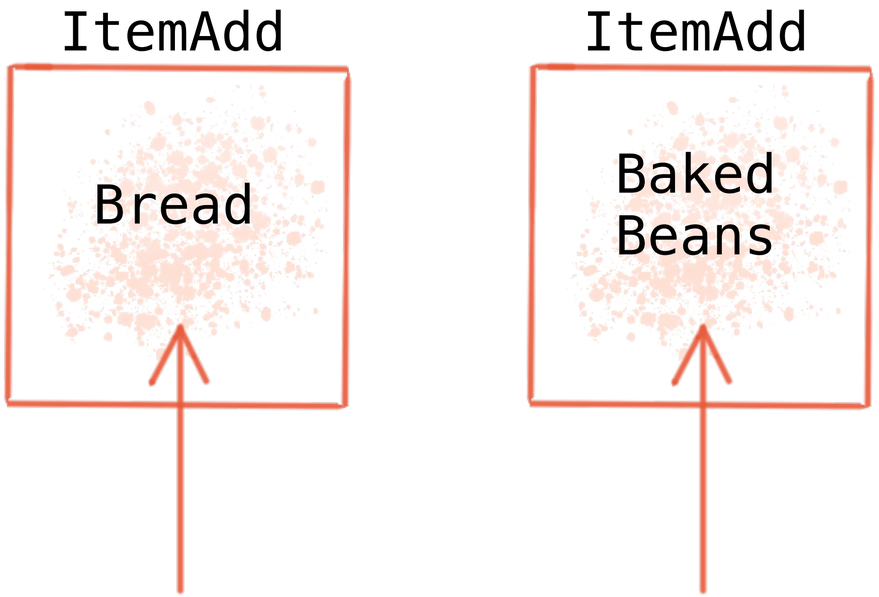
Events

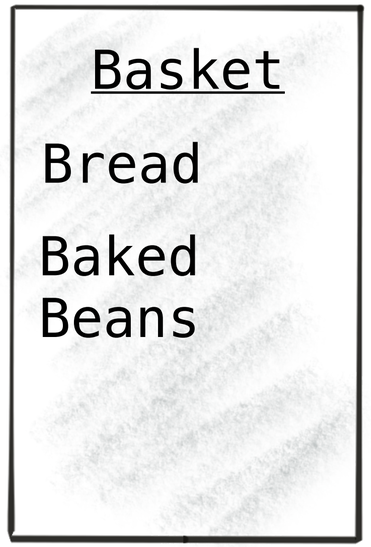
Events



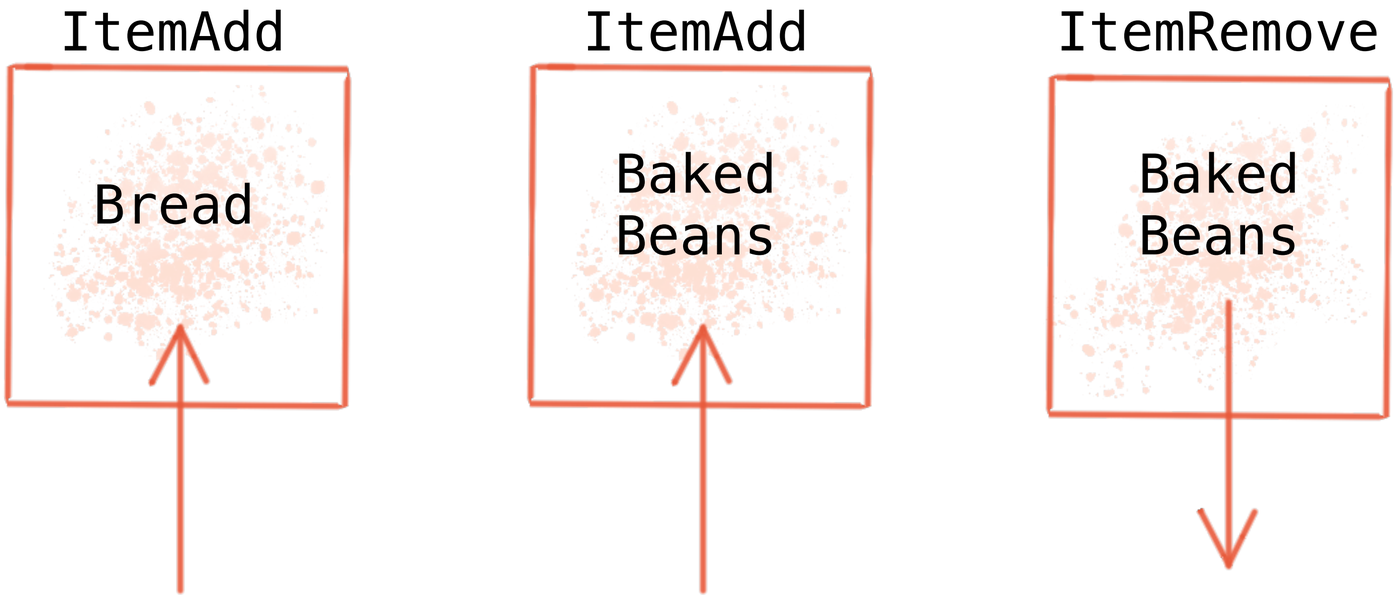


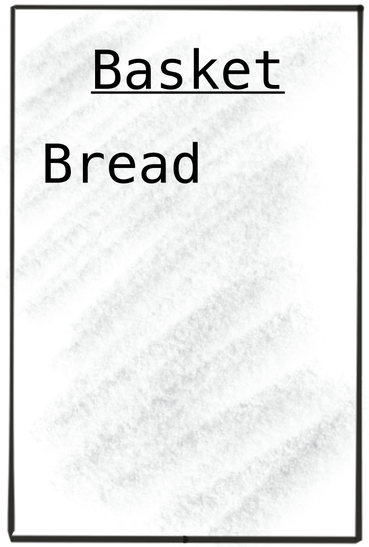
Events


ItemAdd
Bread
ItemAdd
Baked
Beans

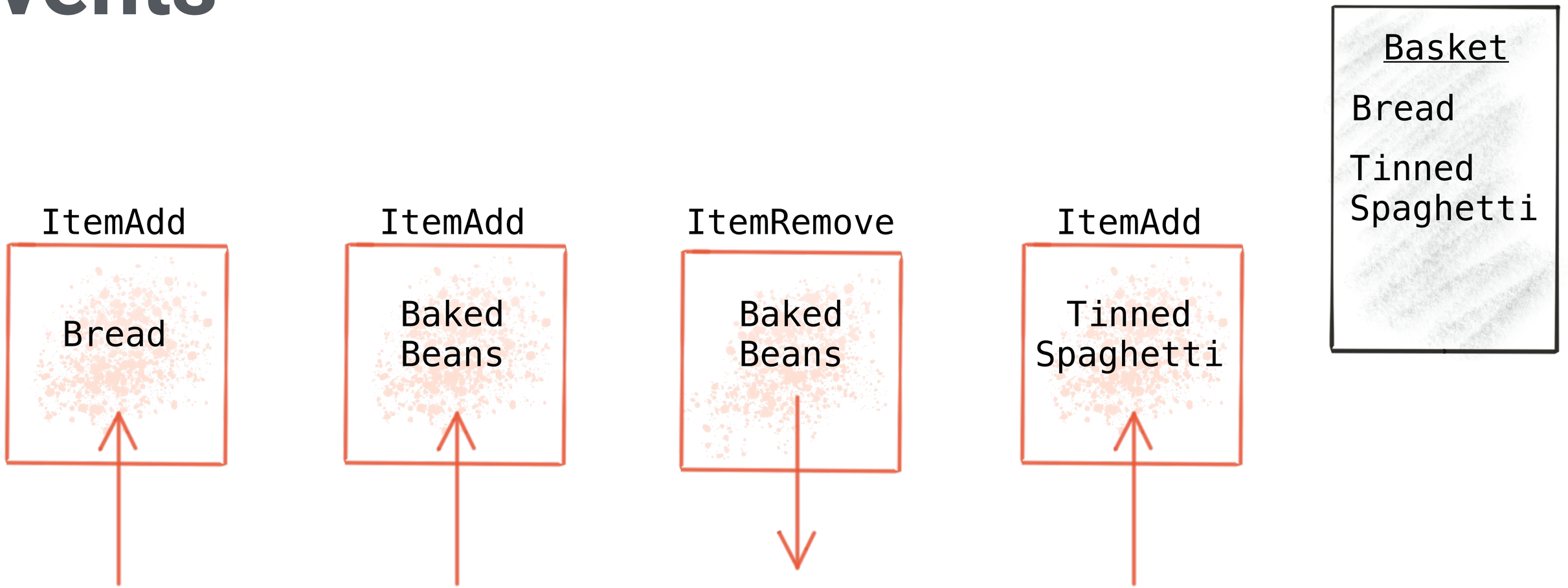

Basket
Bread
Baked
Beans

Events

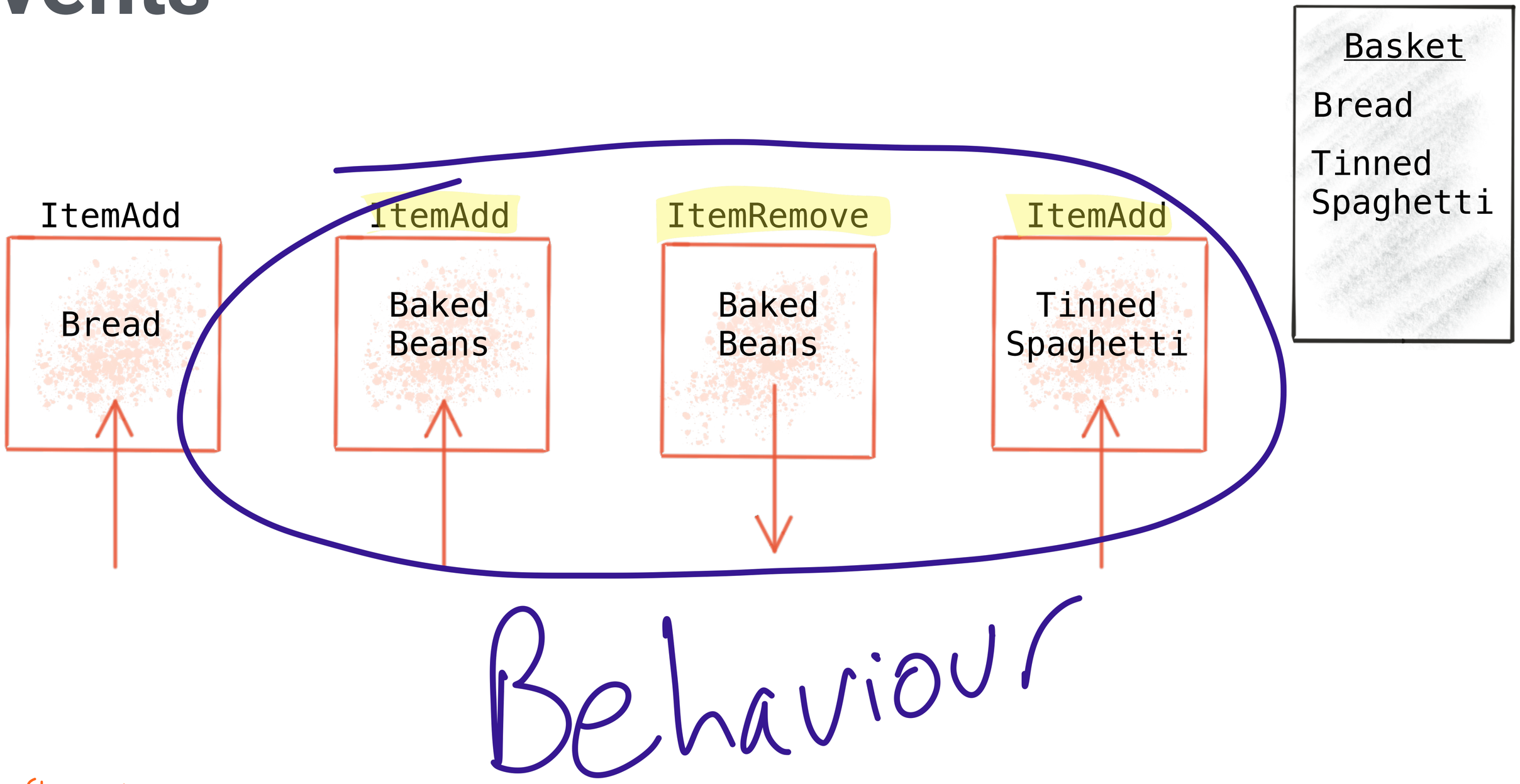

ItemAdd
Bread
ItemAdd
Baked
Beans
ItemRemove
Baked
Beans


Basket
Bread

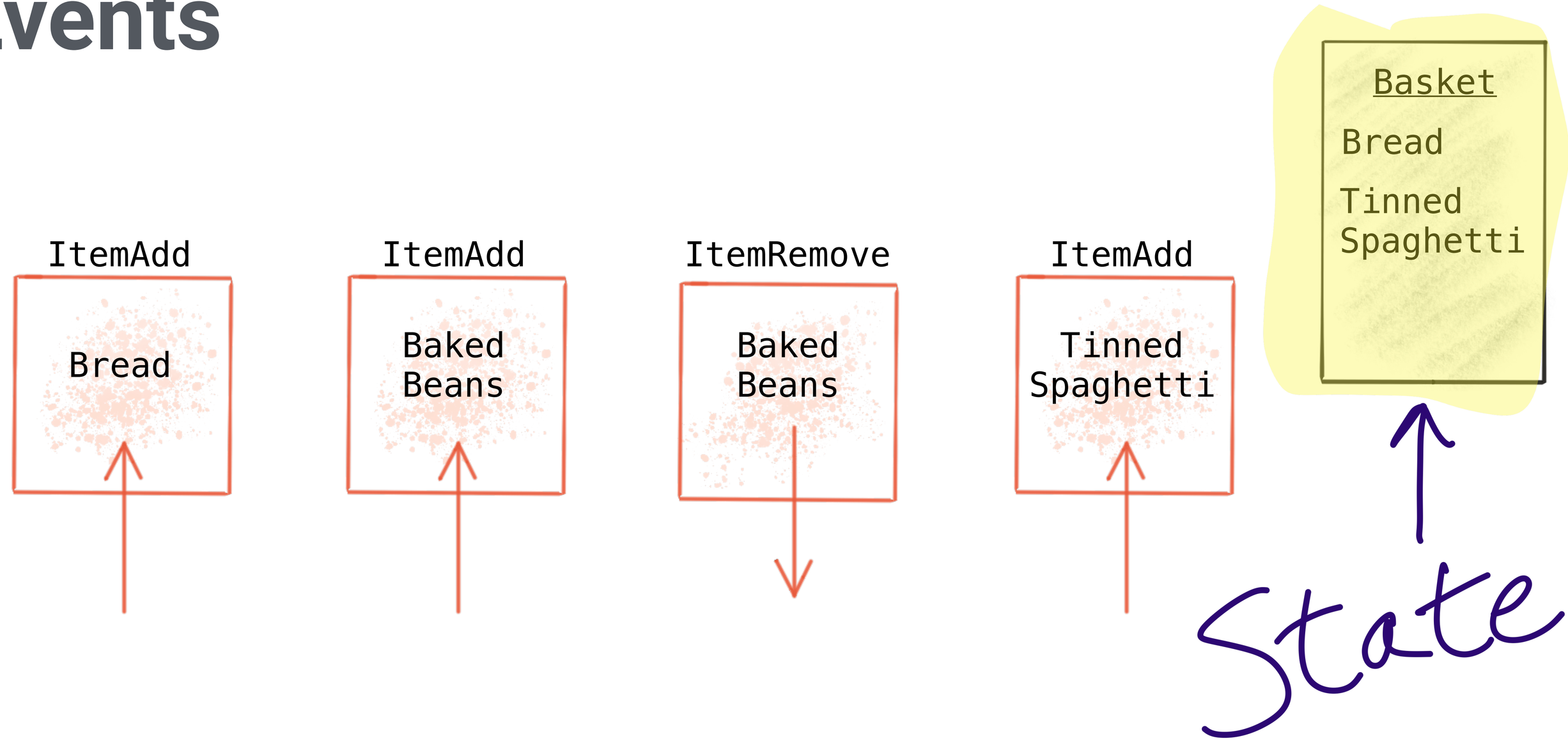
Events


ItemAdd
Bread
ItemAdd
Baked
Beans
ItemRemove
Baked
Beans
ItemAdd
Tinned
Spaghetti
Basket
Bread
Tinned
Spaghetti

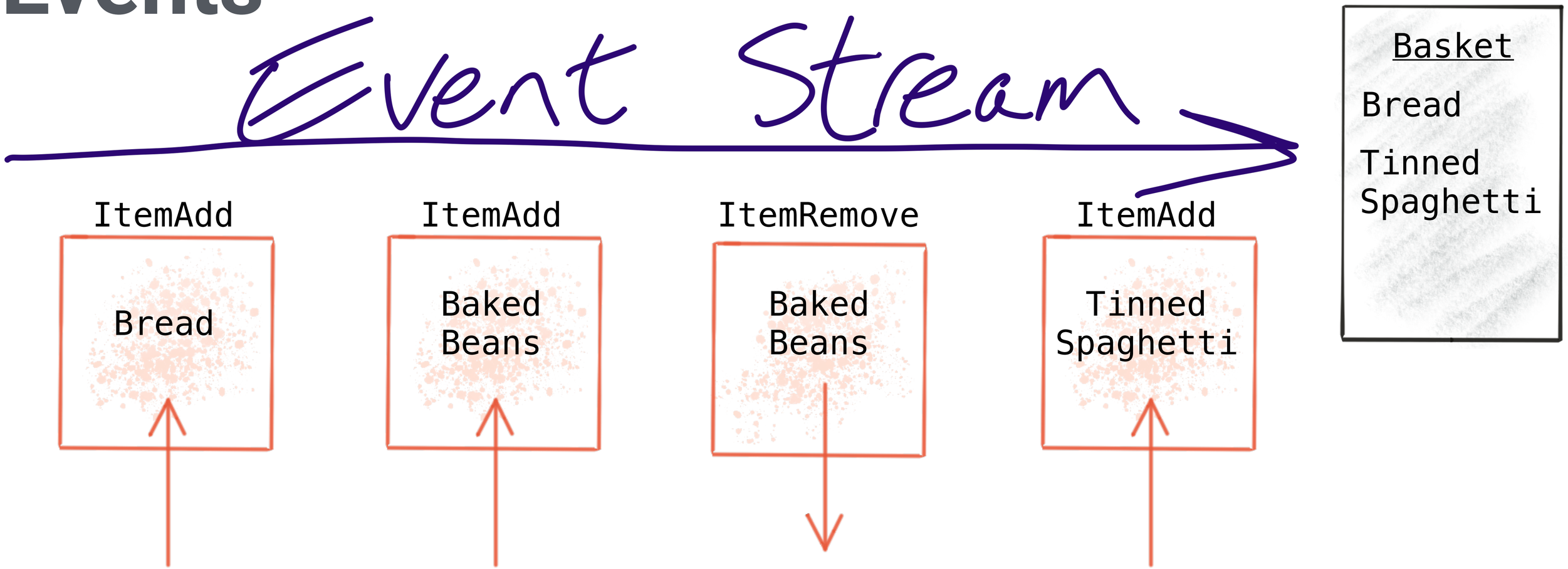
Events


Basket
Bread
Tinned Spaghetti
ItemAdd
Bread
ItemAdd
Baked Beans
ItemRemove
Baked Beans
ItemAdd
Tinned Spaghetti
Behaviour

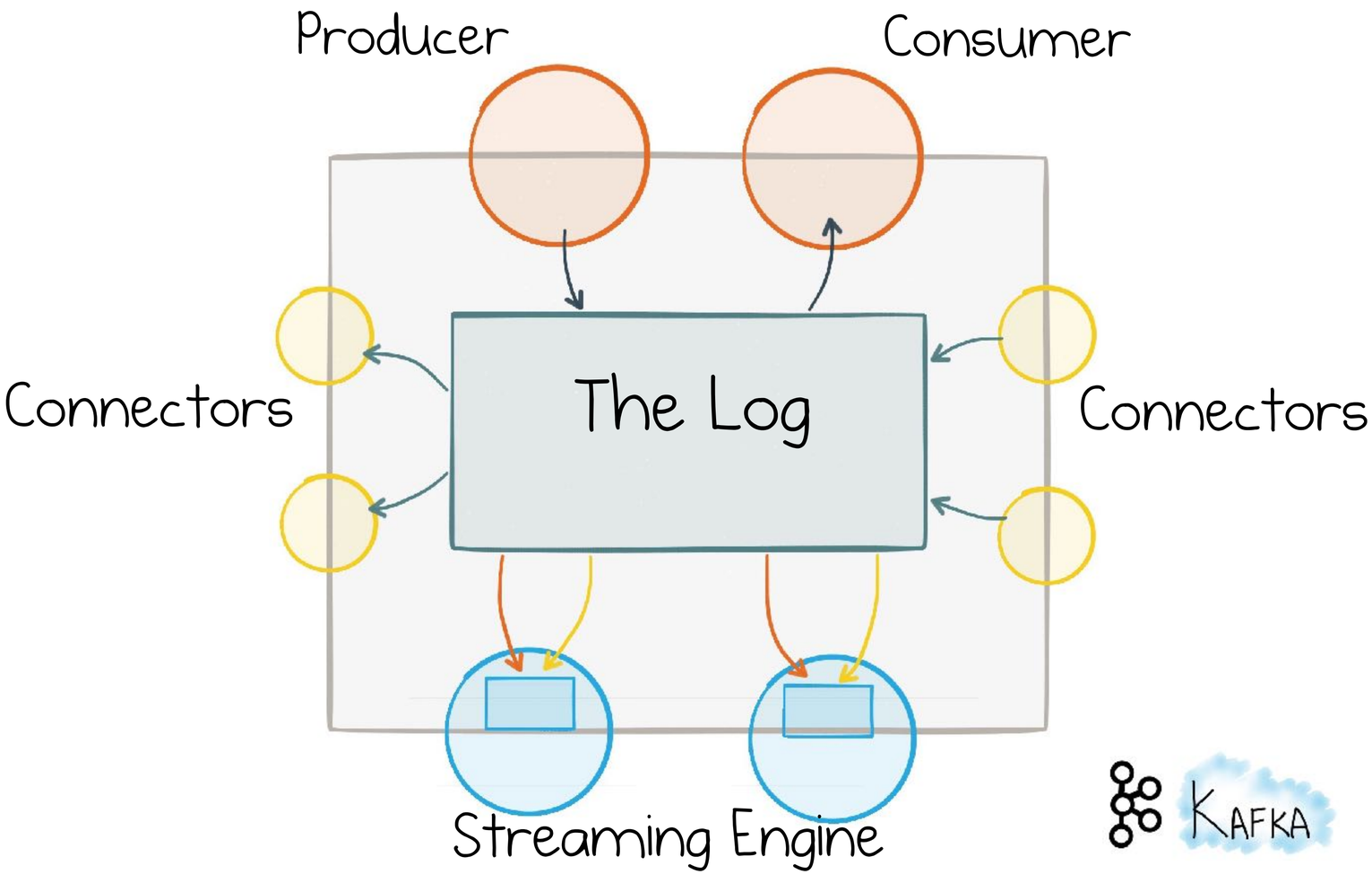
Events


ItemAdd
Bread
ItemAdd
Baked Beans
ItemRemove
Baked Beans
ItemAdd
Tinned Spaghetti
Basket
Bread
Tinned Spaghetti
State

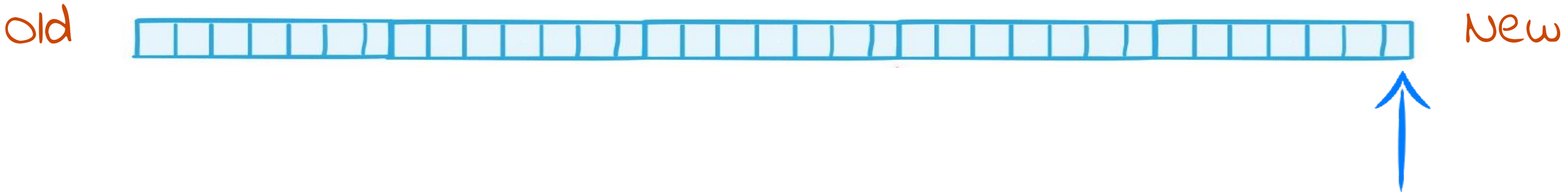
Events


Event Stream
ItemAdd
Bread
ItemAdd
Baked Beans
ItemRemove
Baked Beans
ItemAdd
Tinned Spaghetti
Basket
Bread
Tinned Spaghetti

What is an Event Streaming Platform?

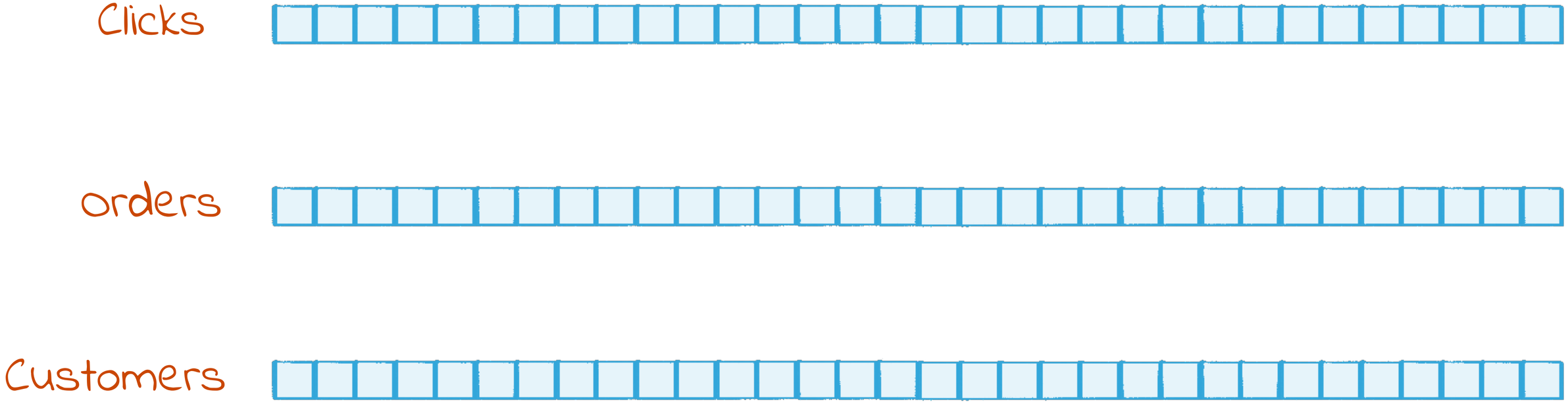


Immutable Event Log



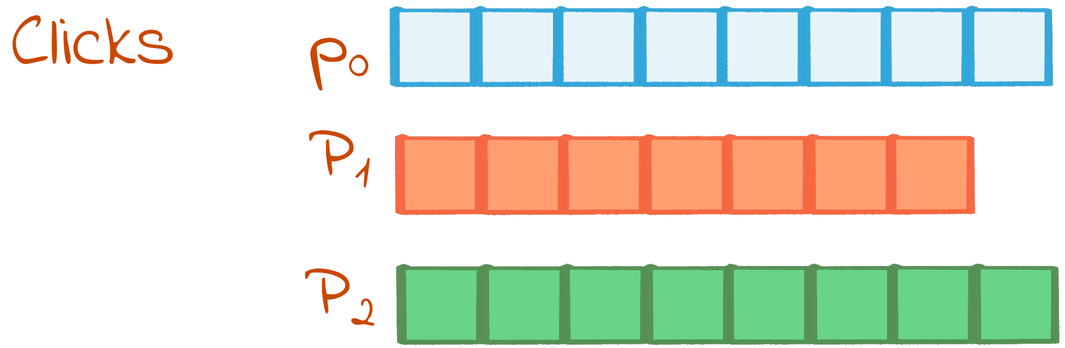
Messages are added at the end of the log

Topics



Topics are similar in concept to tables in a database

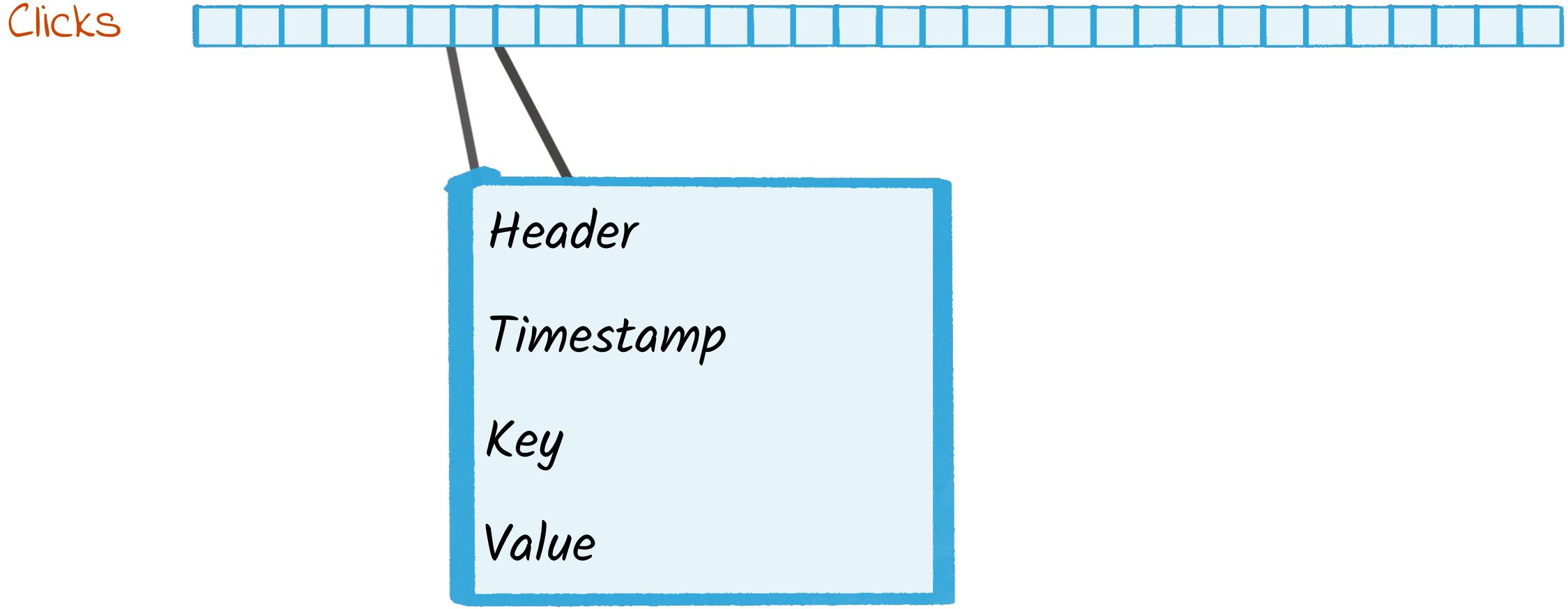
Partitions



Messages are guaranteed to be strictly ordered within a partition

Messages are just K/V bytes

plus headers + timestamp



Messages are just K/V bytes

With great power comes great responsibility

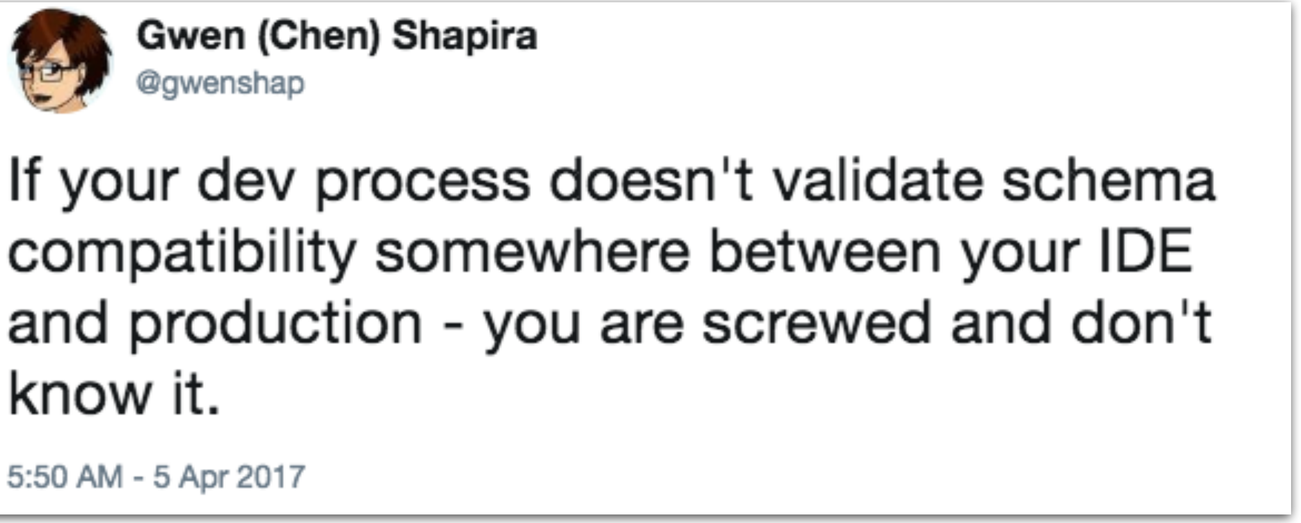
Avro

-> Confluent Schema Registry

Protobuf

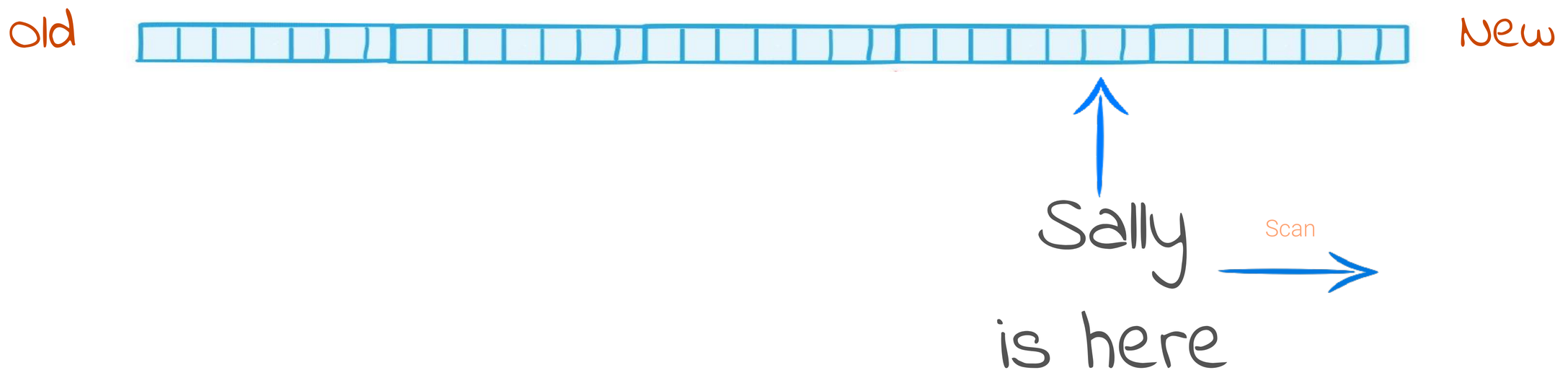
JSON

CSV

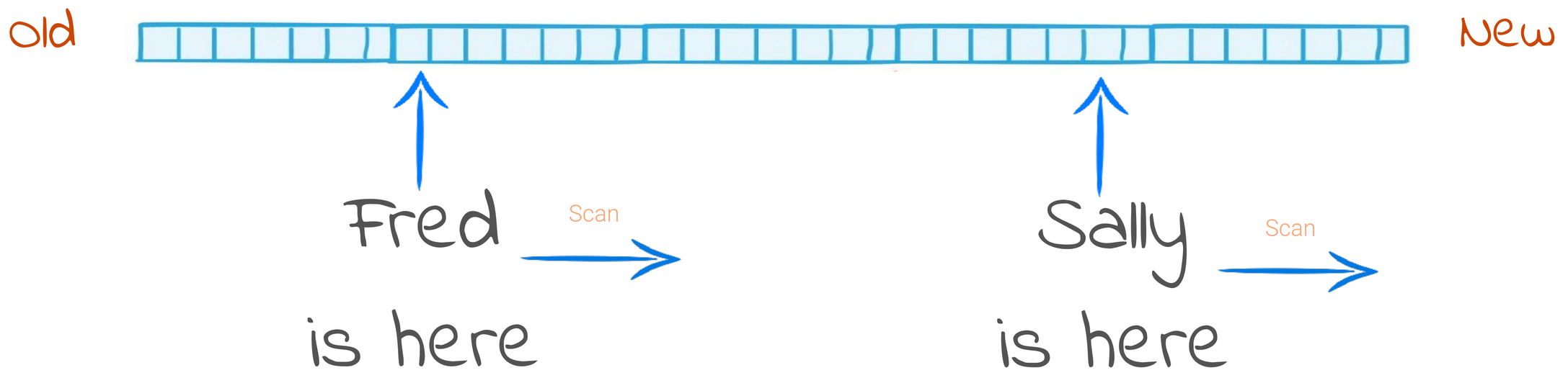


https://qconnewyork.com/system/files/presentation-slides/qcon_17_-_schemas_and_apis.pdf

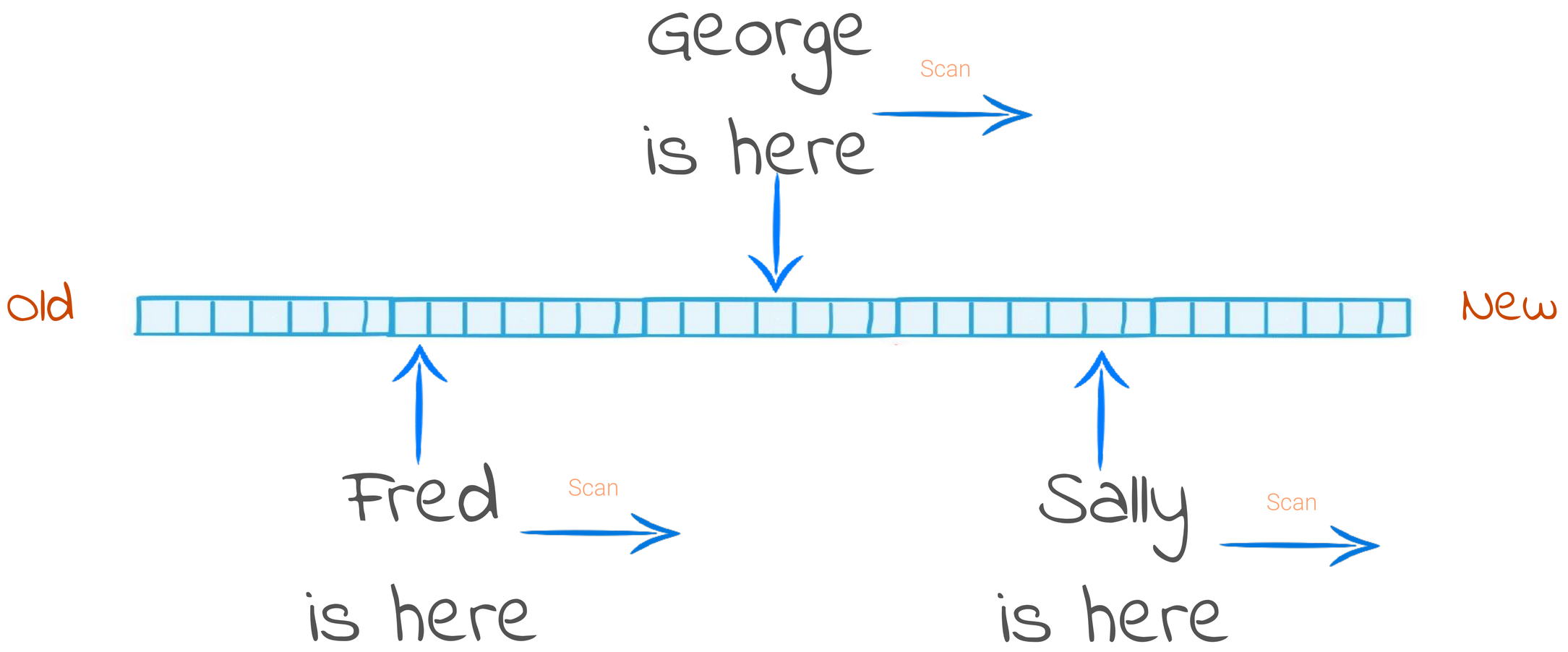
Consumers have a position all of their own



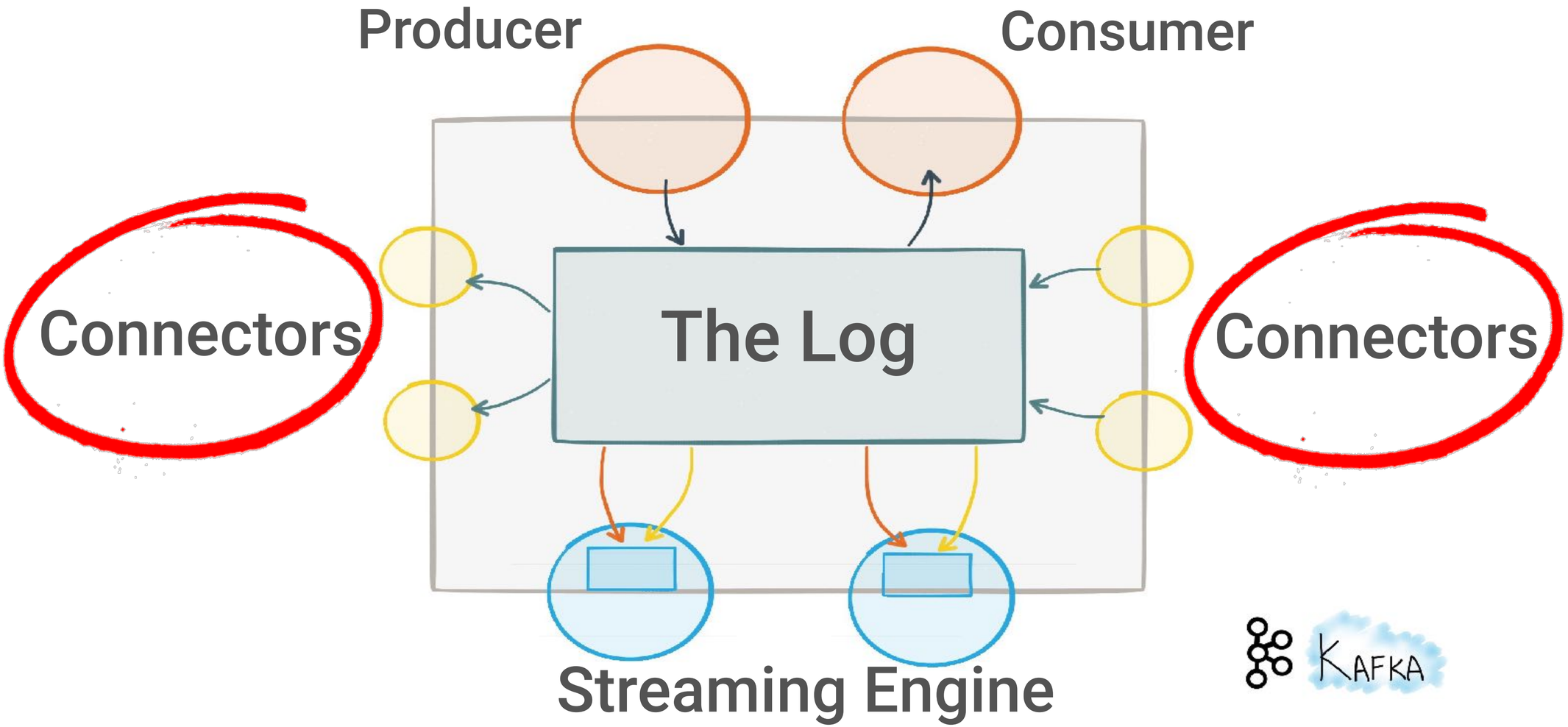
Consumers have a position all of their own



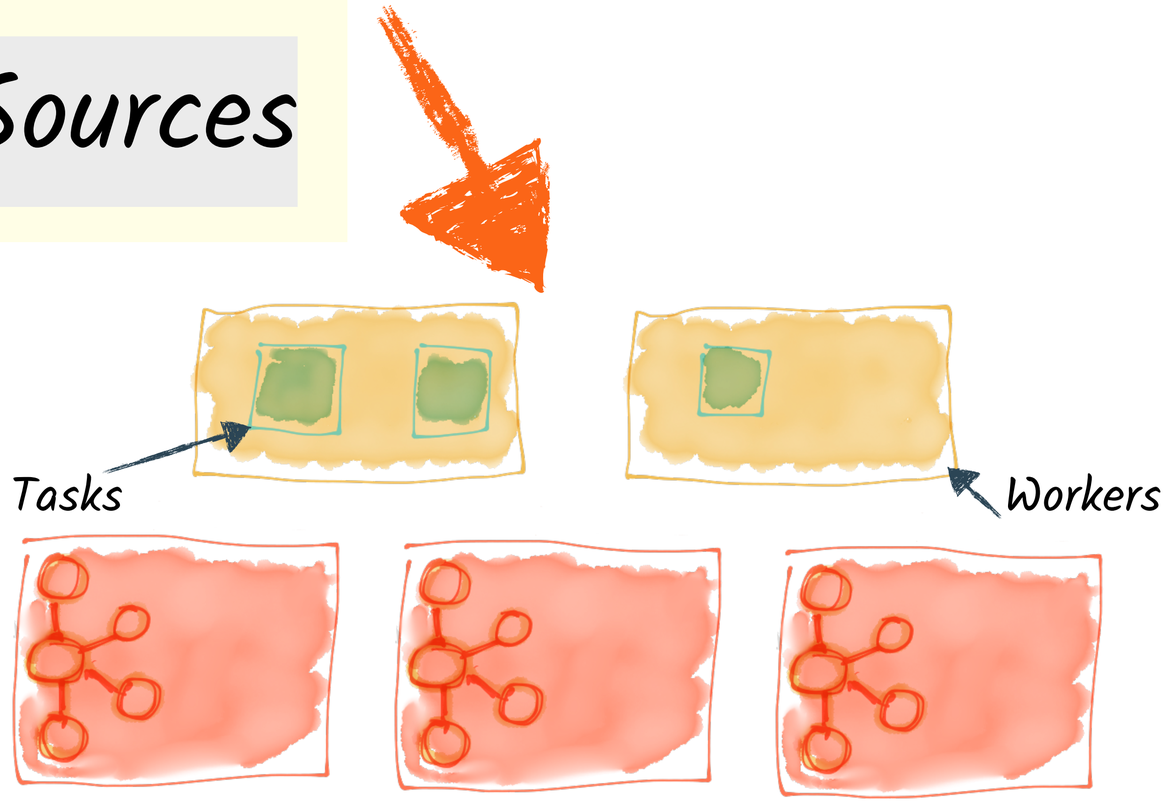
Consumers have a position all of their own



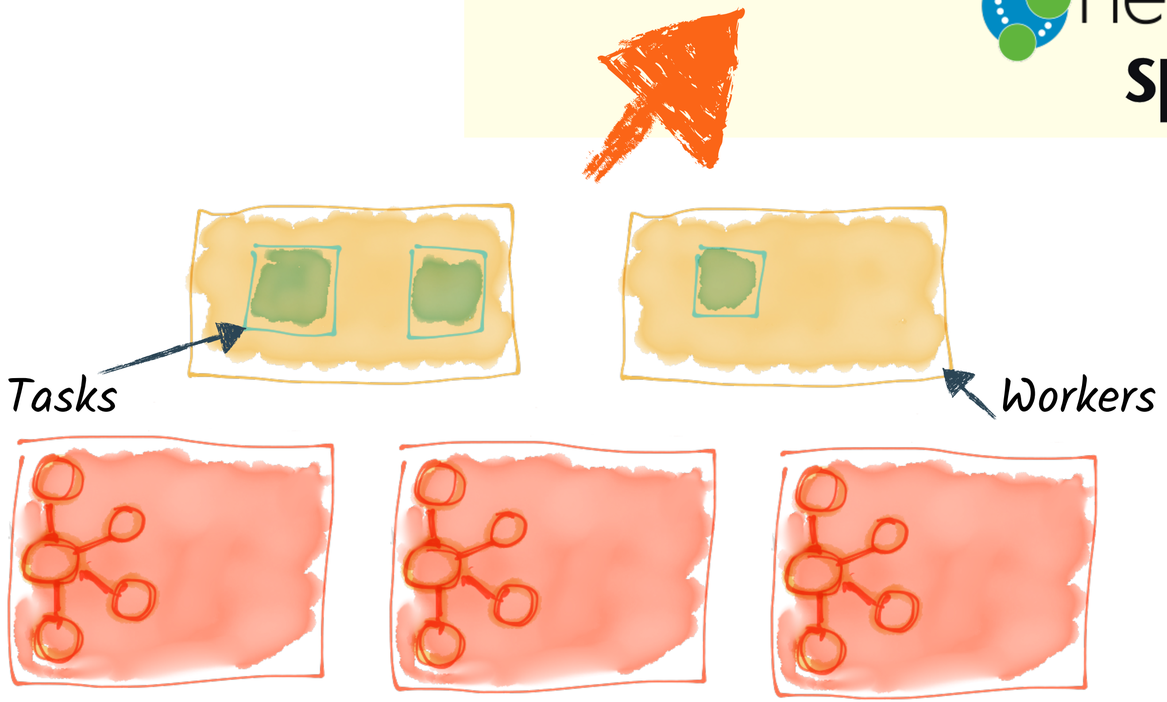
The Connect API



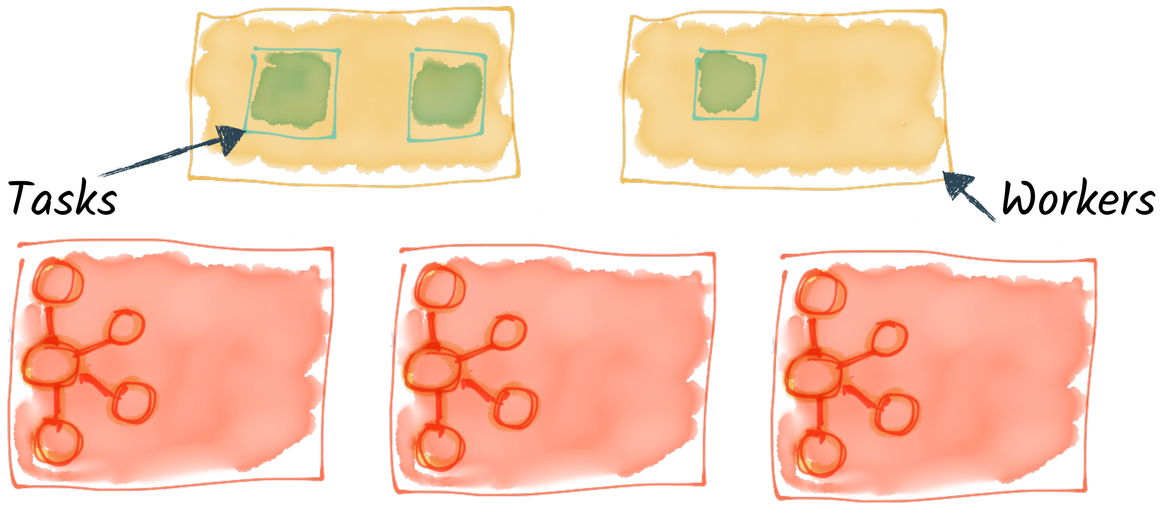
Streaming Integration with Kafka Connect



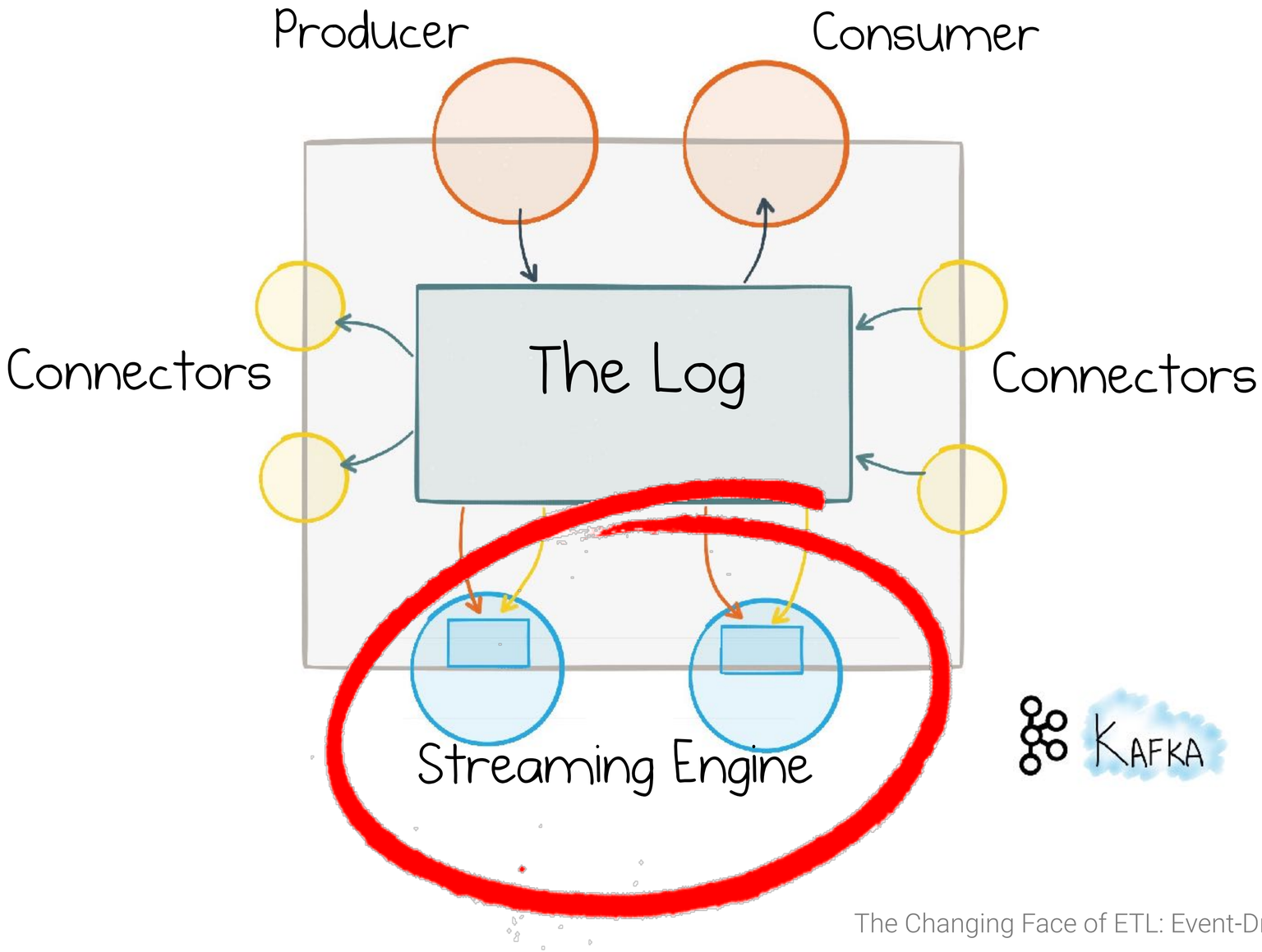
Streaming Integration with Kafka Connect



Streaming Integration with Kafka Connect

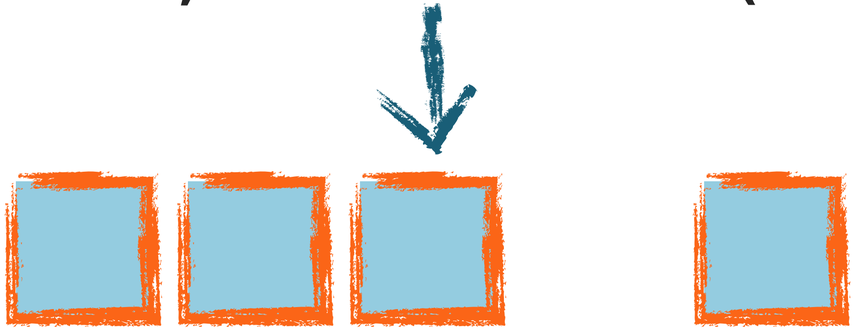


Stream Processing in Kafka



Kafka Streams API

```
final StreamsBuilder builder = new StreamsBuilder()
    .stream("orders", Consumed.with(stringSerde, ordersSerde))
    .filter( (key, order) -> order.getStatus().equals("COMPLETE") )
    .to("complete_orders", Produced.with(stringSerde, ordersSerde));
```



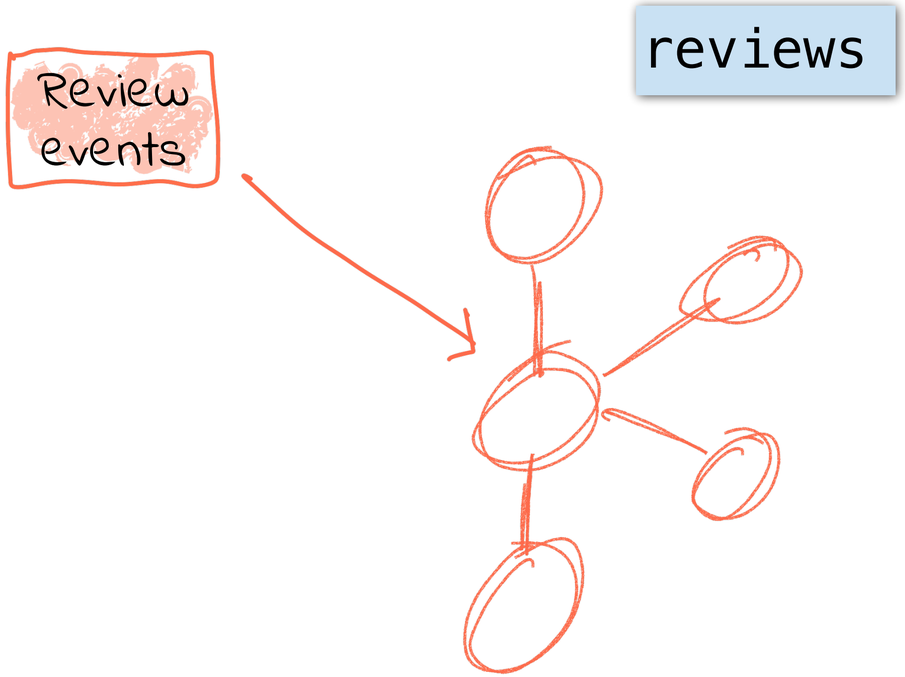
Stream Processing with KSQL

```
CREATE STREAM completedOrders AS
SELECT *
FROM orders
WHERE status='COMPLETE';
```

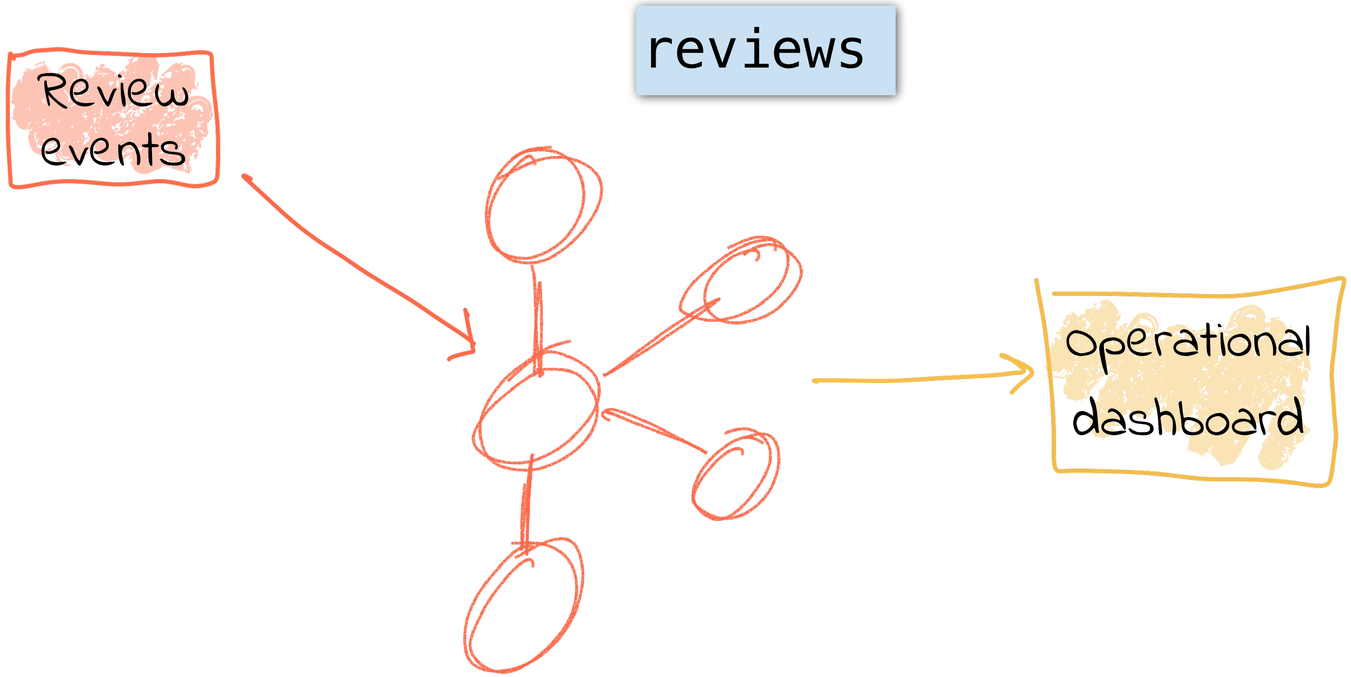
This is Something New

Photo by Ashkan from Modern Affiatus on Unsplash

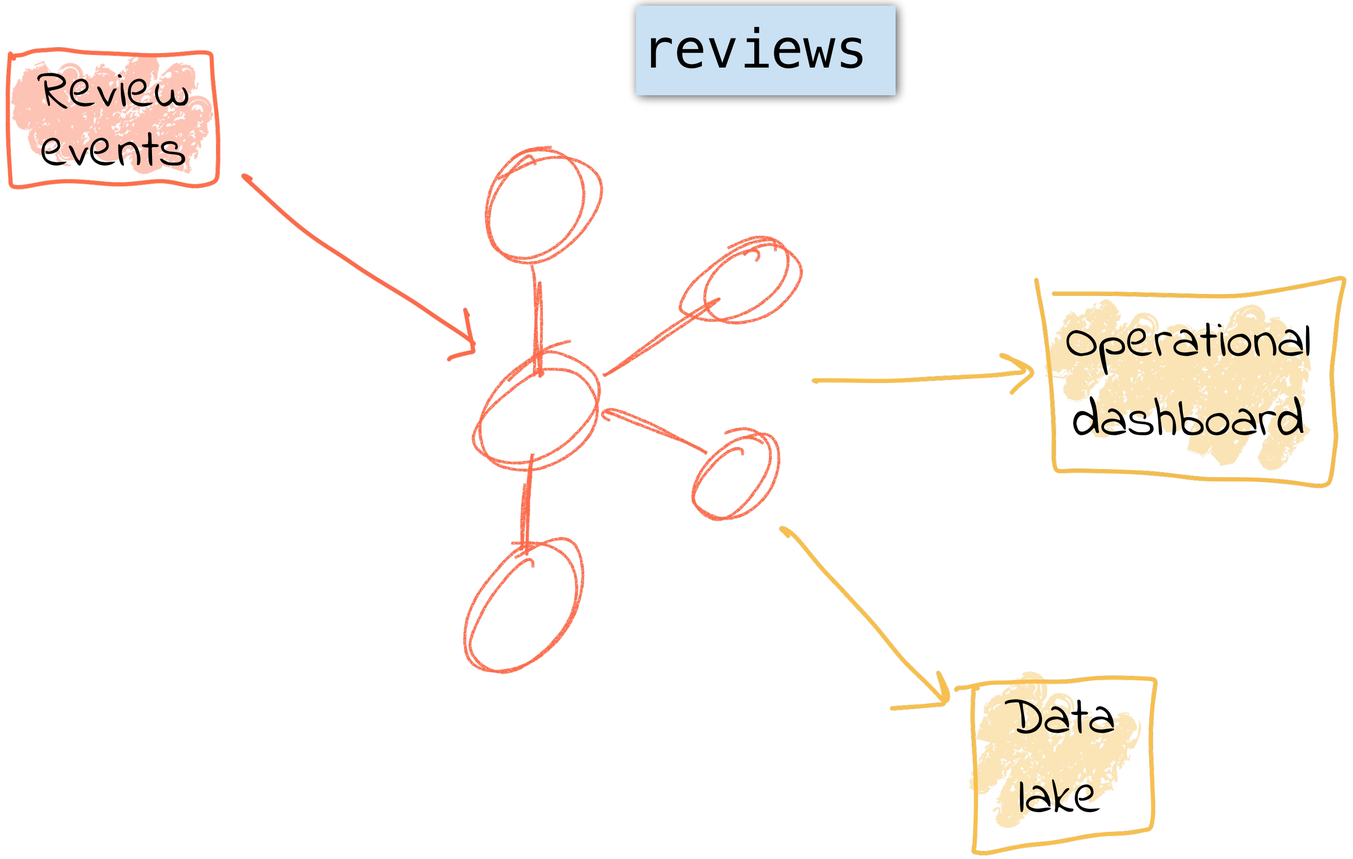
Events in Action



Events in Action

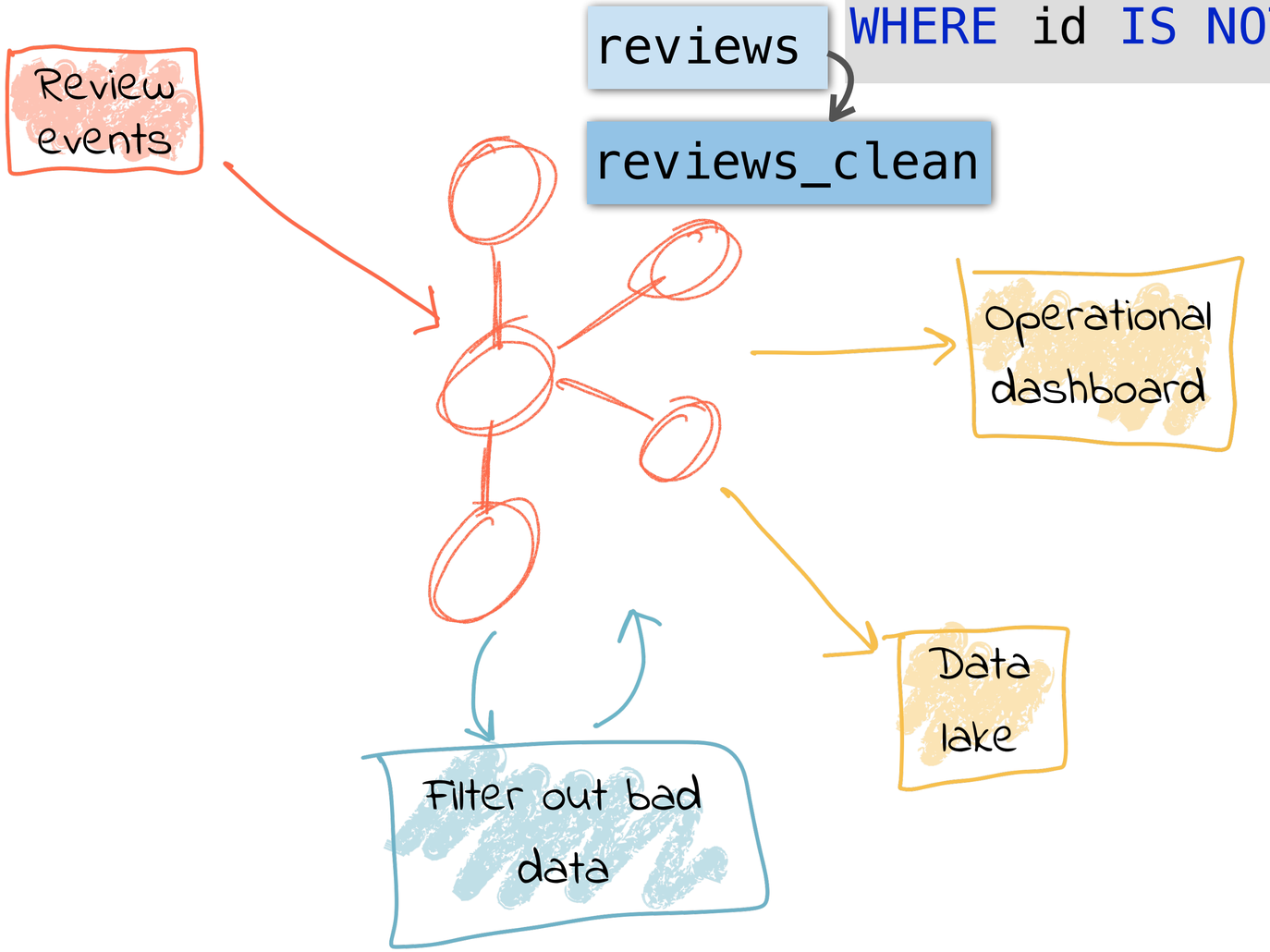


Events in Action

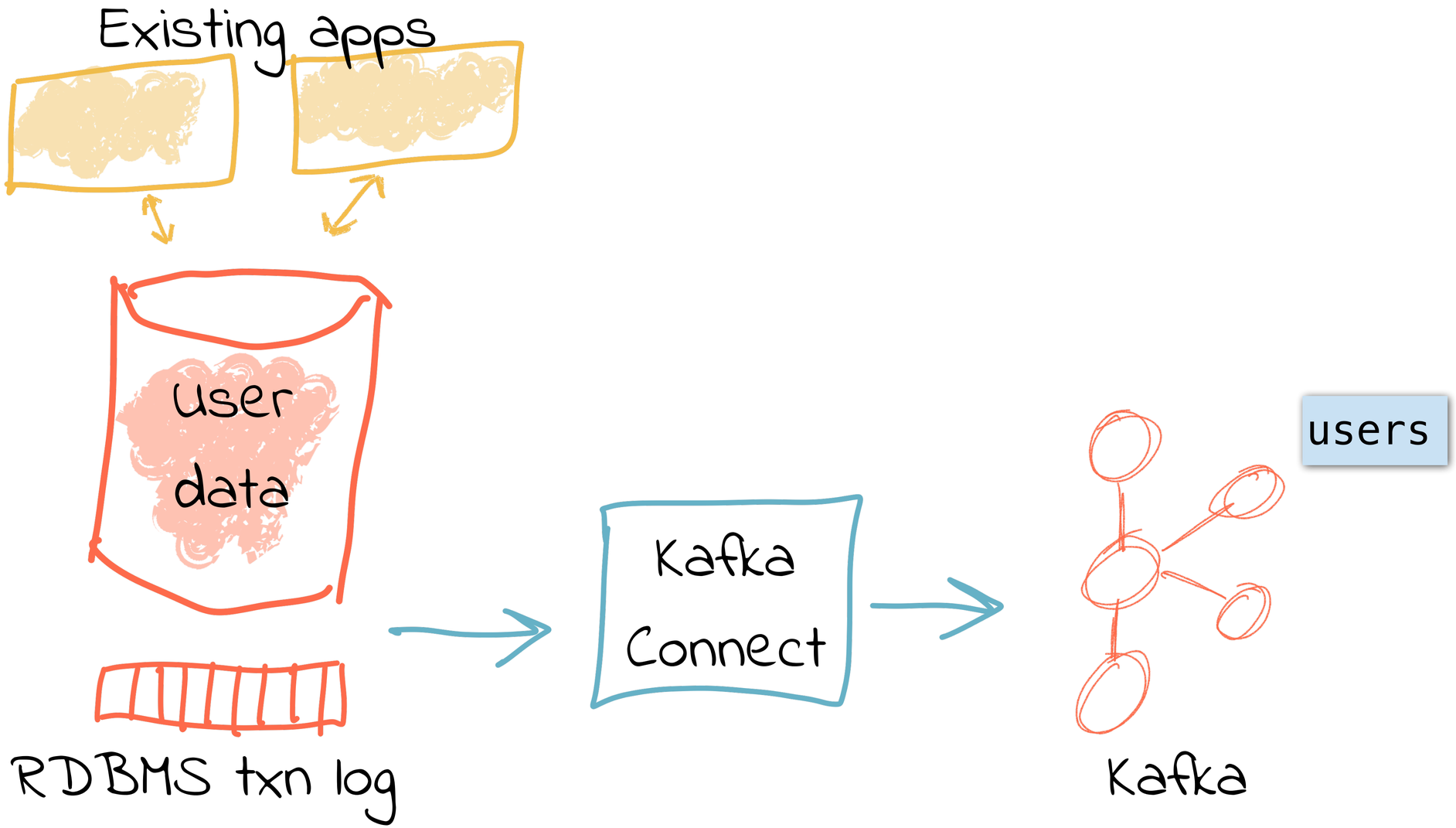


Events in Action

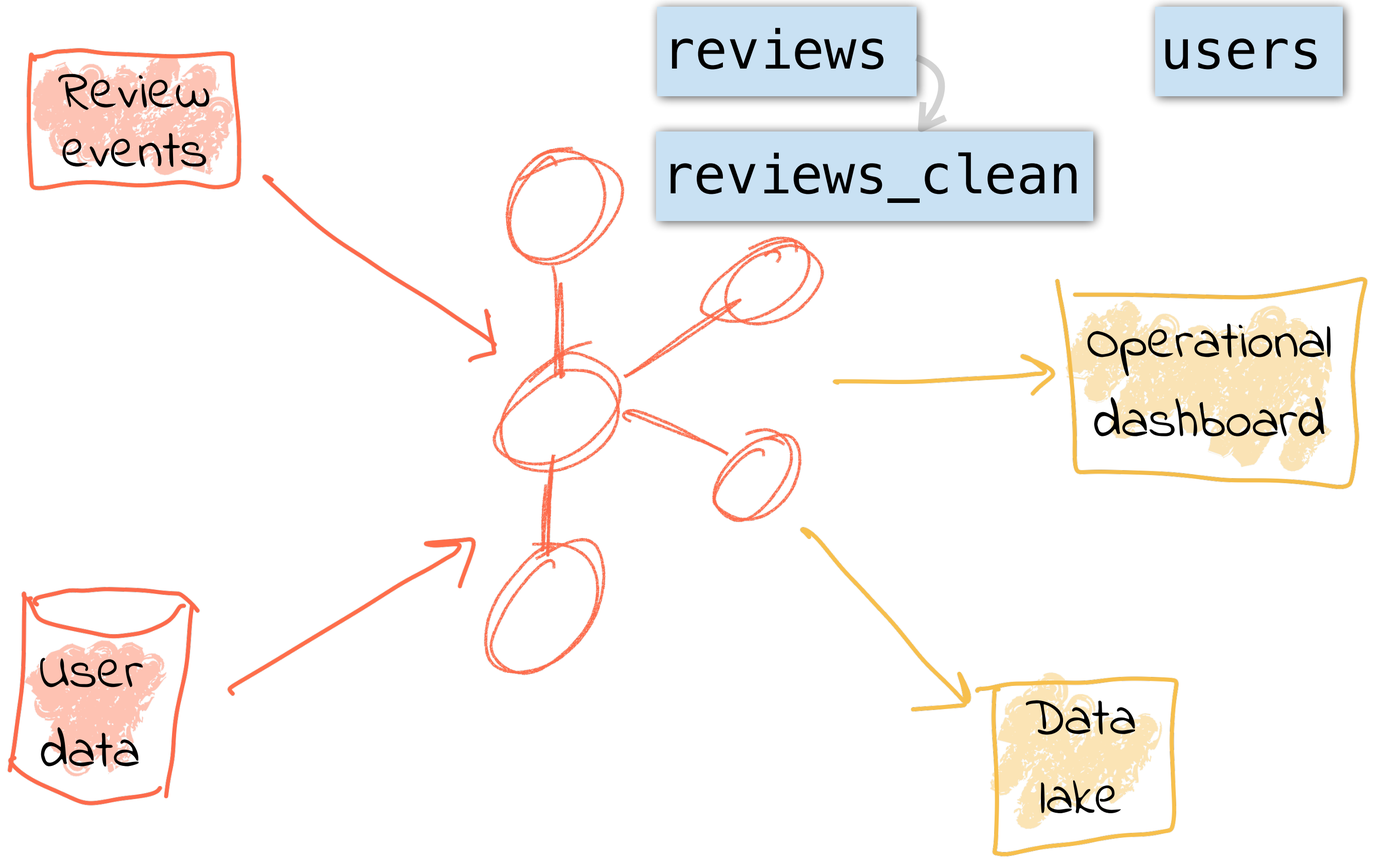
```
CREATE STREAM reviews_clean AS
SELECT * FROM reviews
WHERE id IS NOT NULL;
```



Events in Action

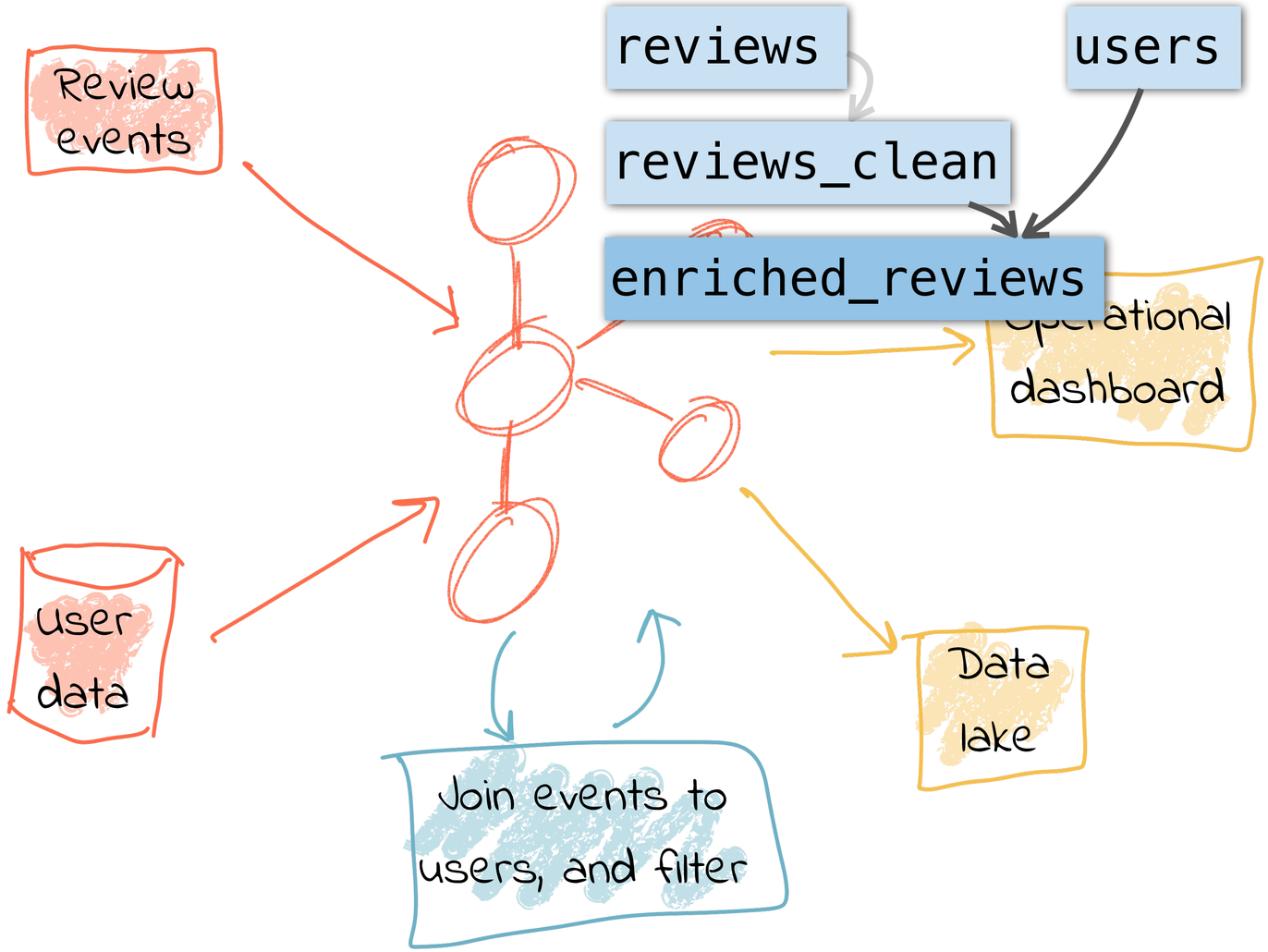


Events in Action

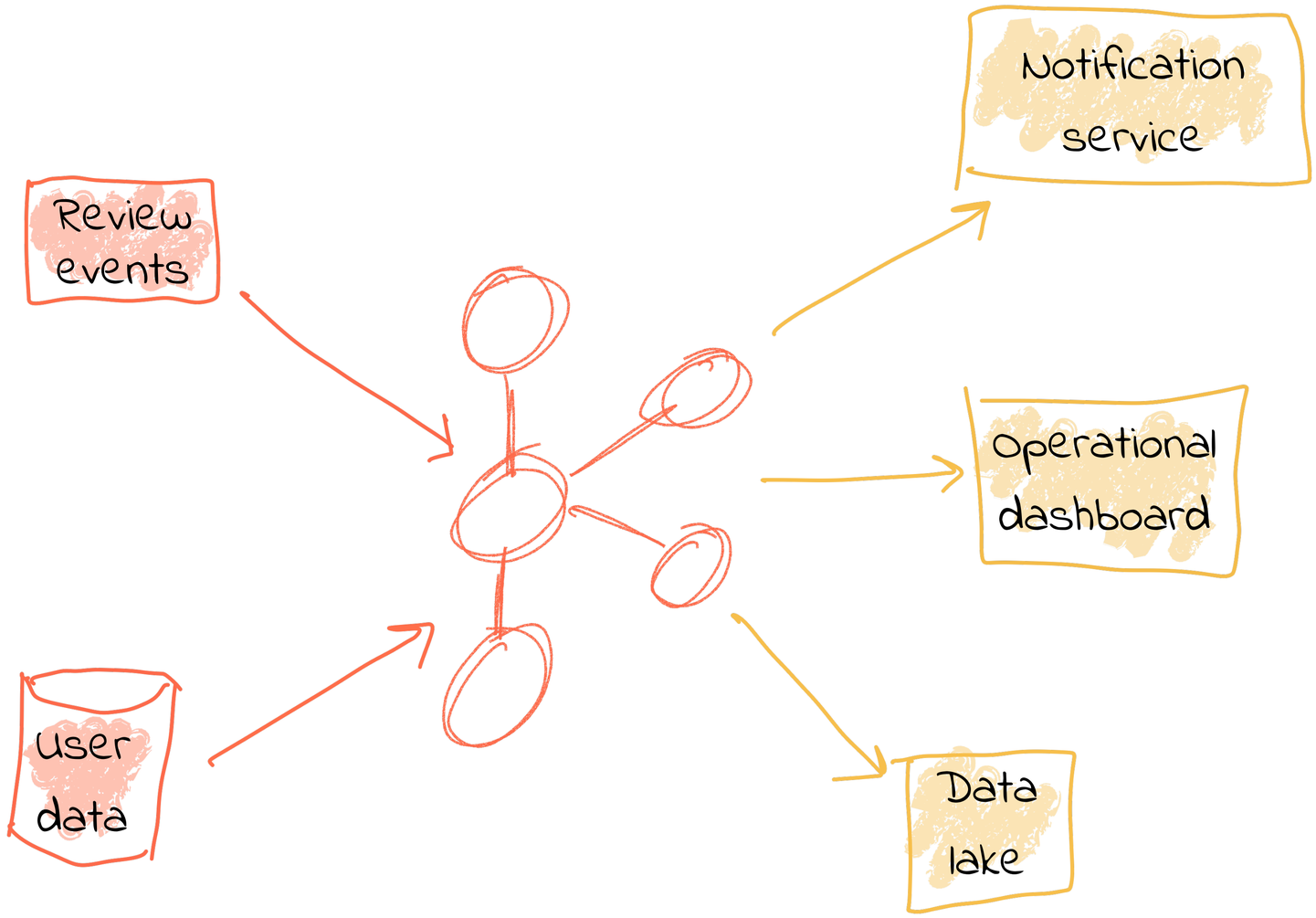


Events in Action

```
CREATE STREAM enriched_reviews AS
SELECT * FROM reviews_clean r
         INNER JOIN users u
         ON r.userid=u.userid;
```

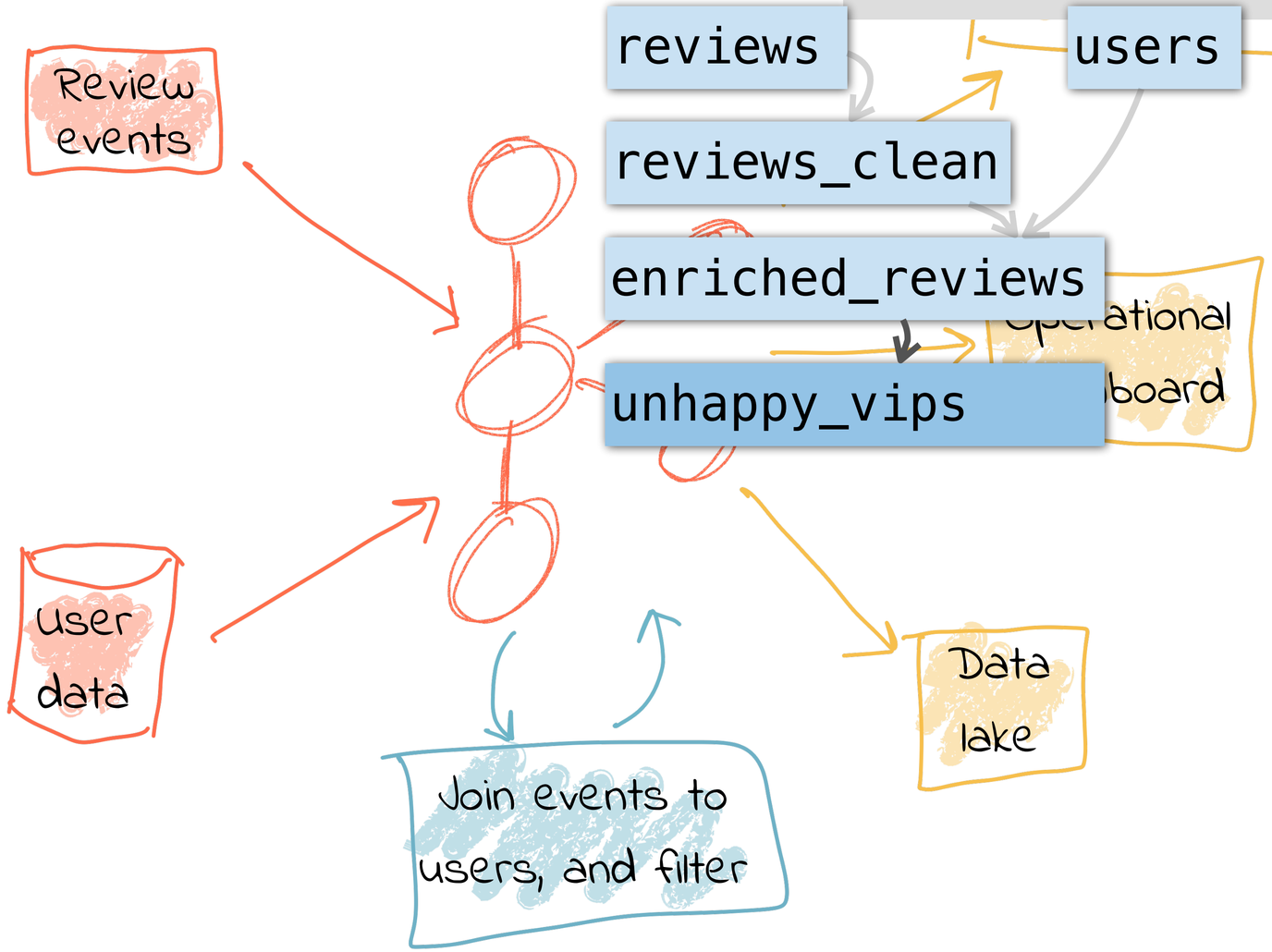


Events in Action



Events in Action

```
CREATE STREAM unhappy_vips AS
SELECT * FROM enriched_reviews
WHERE rating < 3
AND   status = 'Platinum';
```



The Power of an Event-Driven Architecture

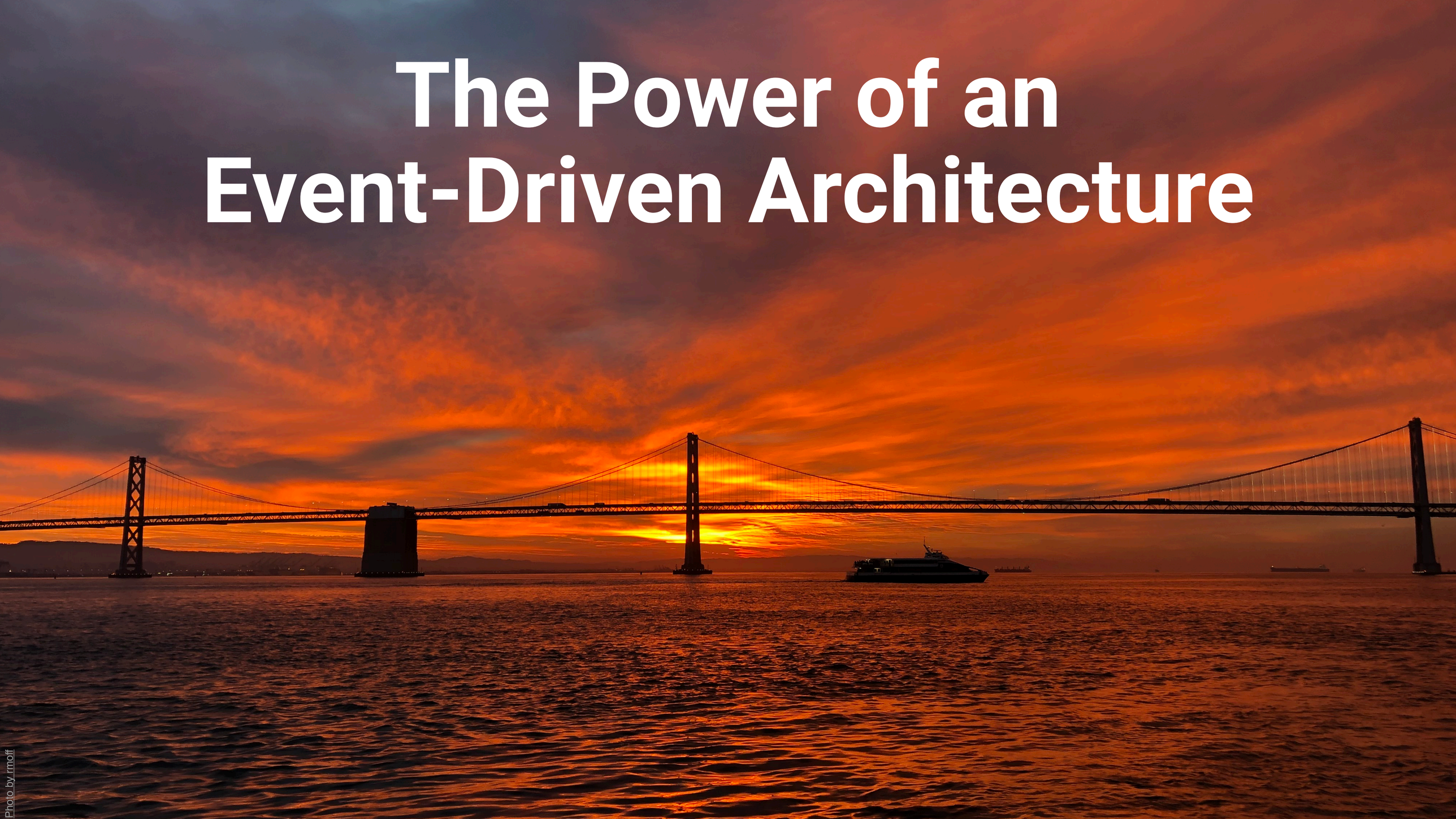
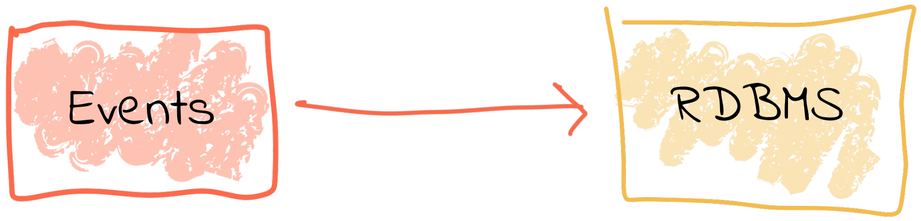
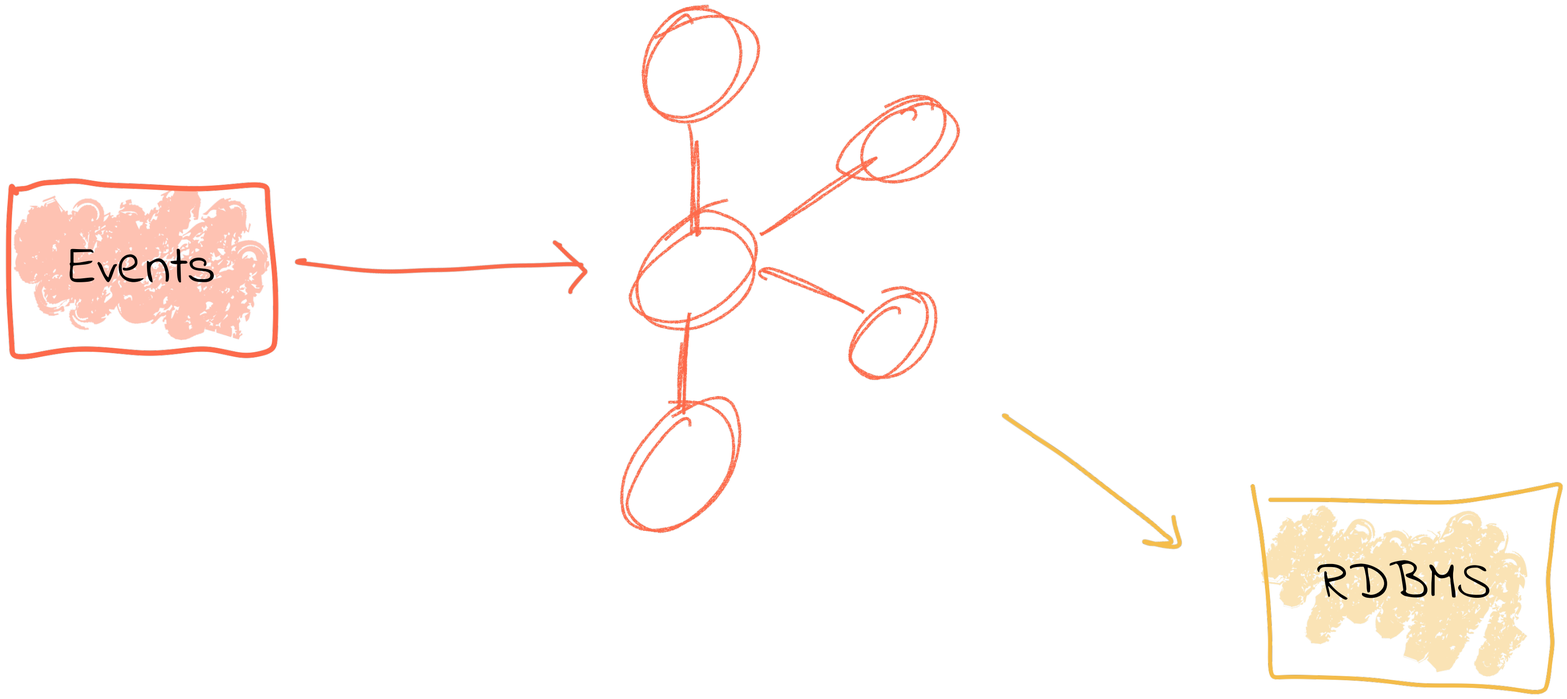


Photo by rmoff

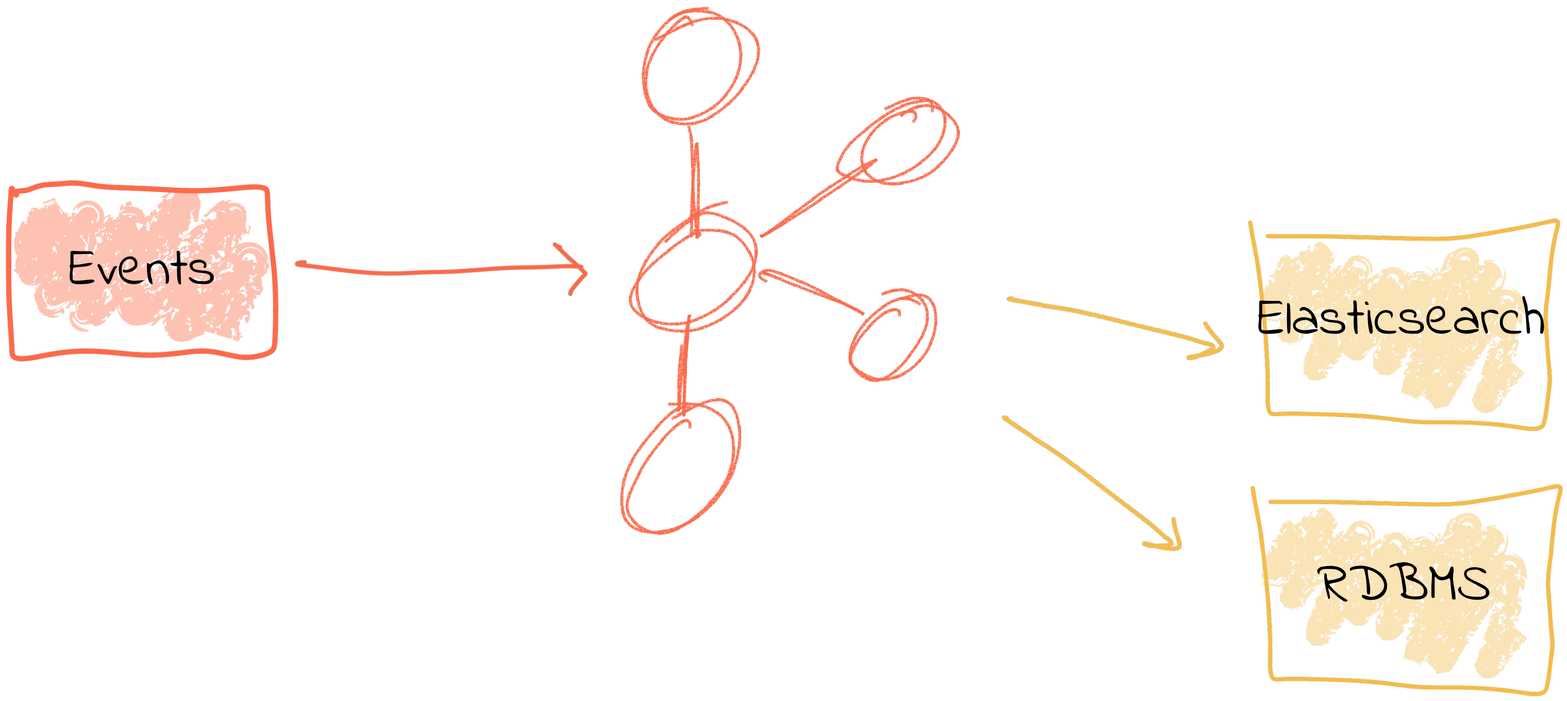
Not Everything is a Nail



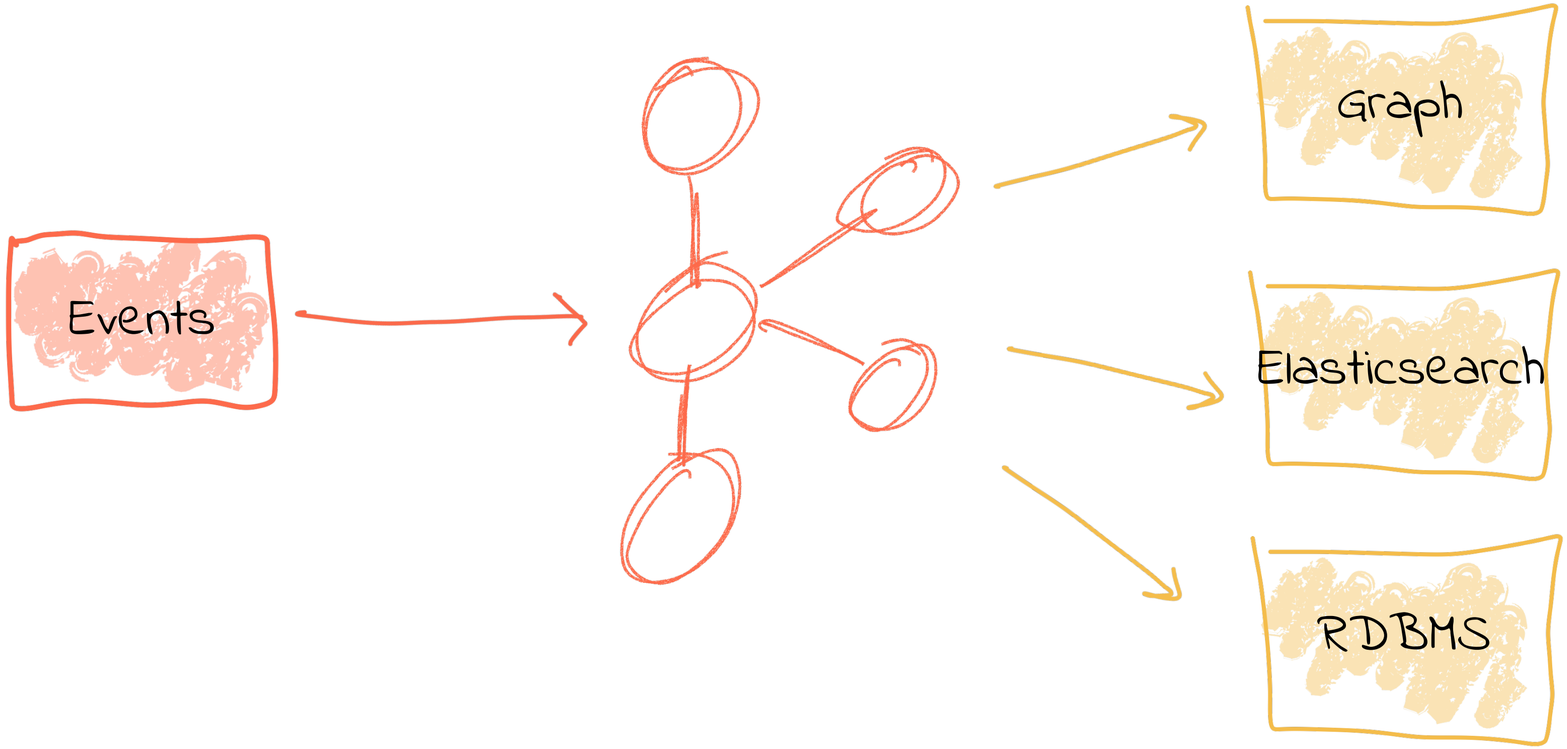
Not Everything is a Nail



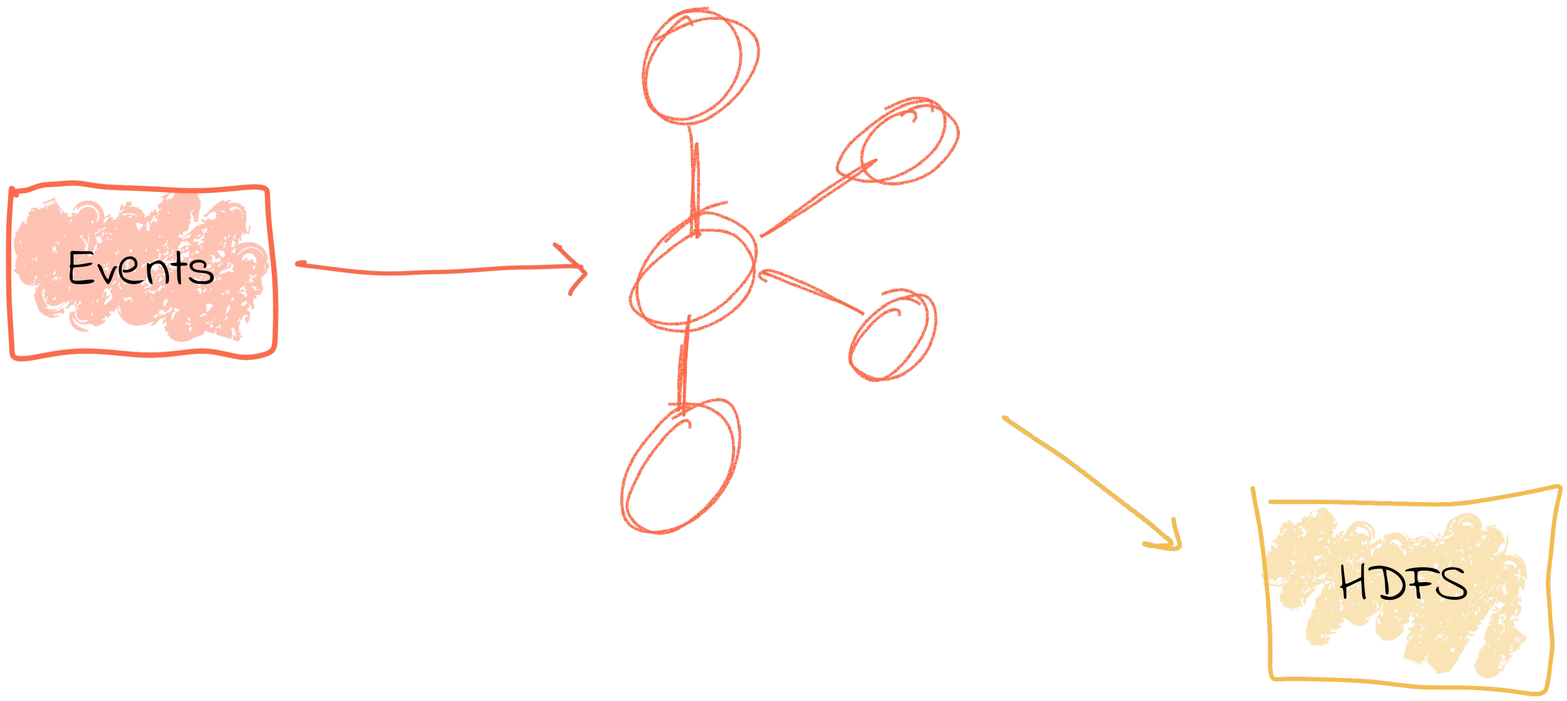
Not Everything is a Nail



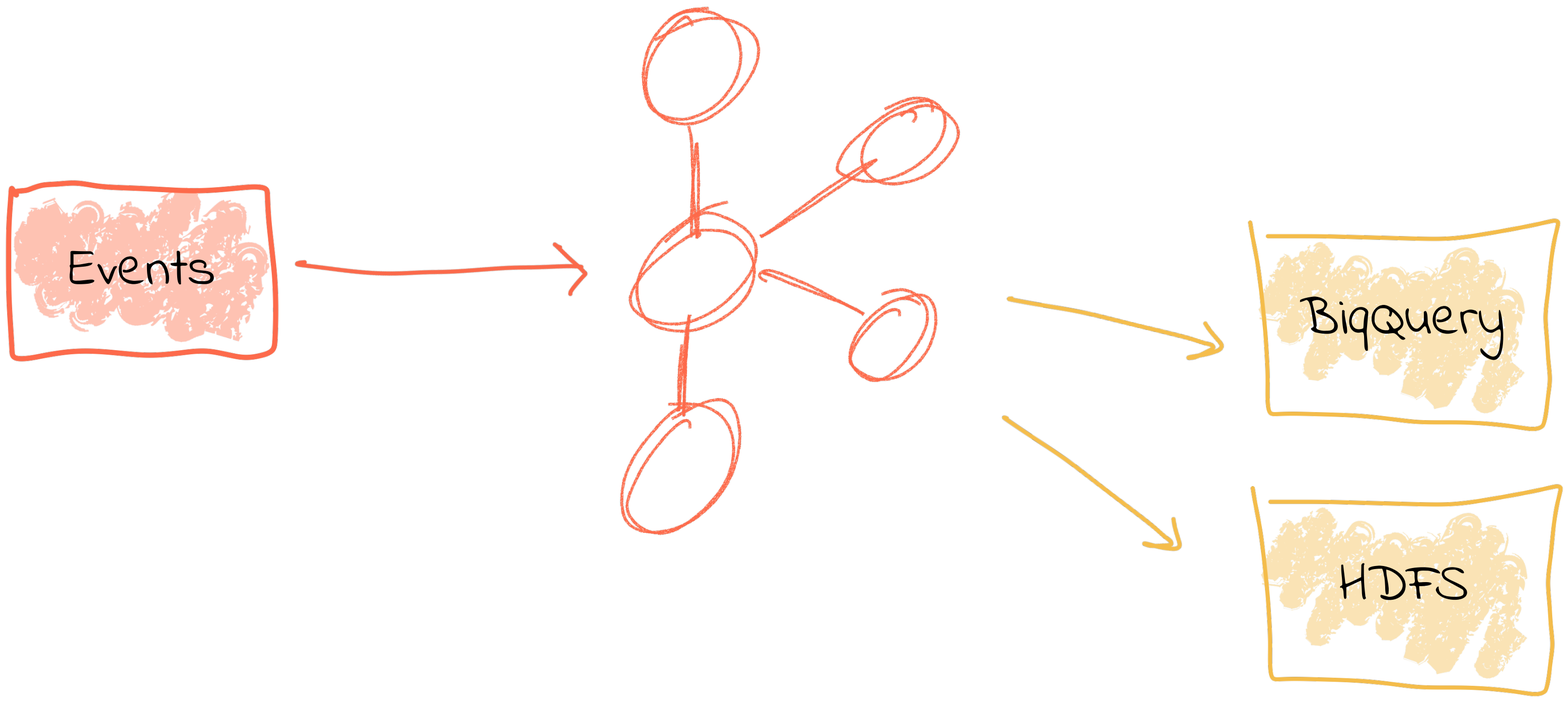
Not Everything is a Nail



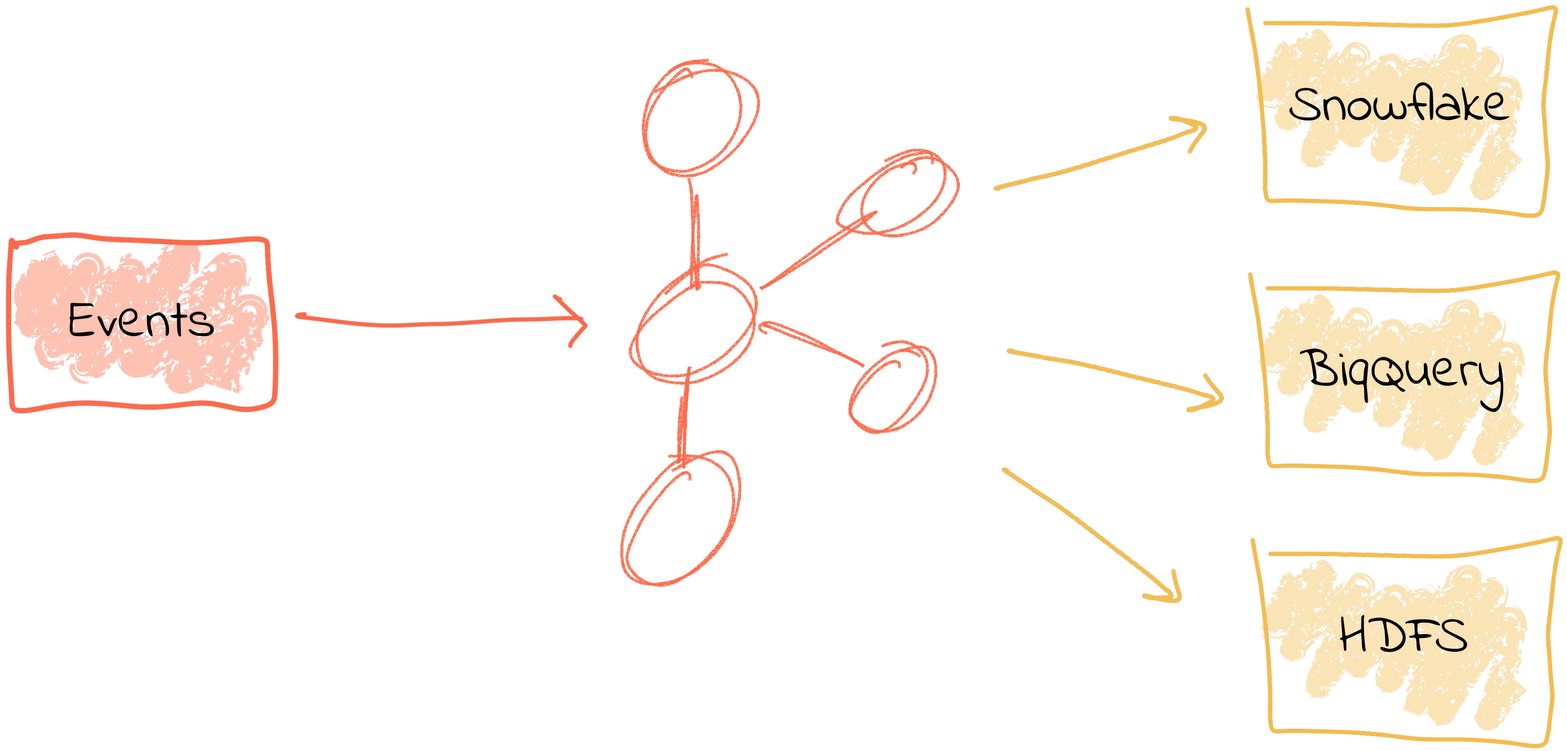
Side-by-Side Tech Evaluation



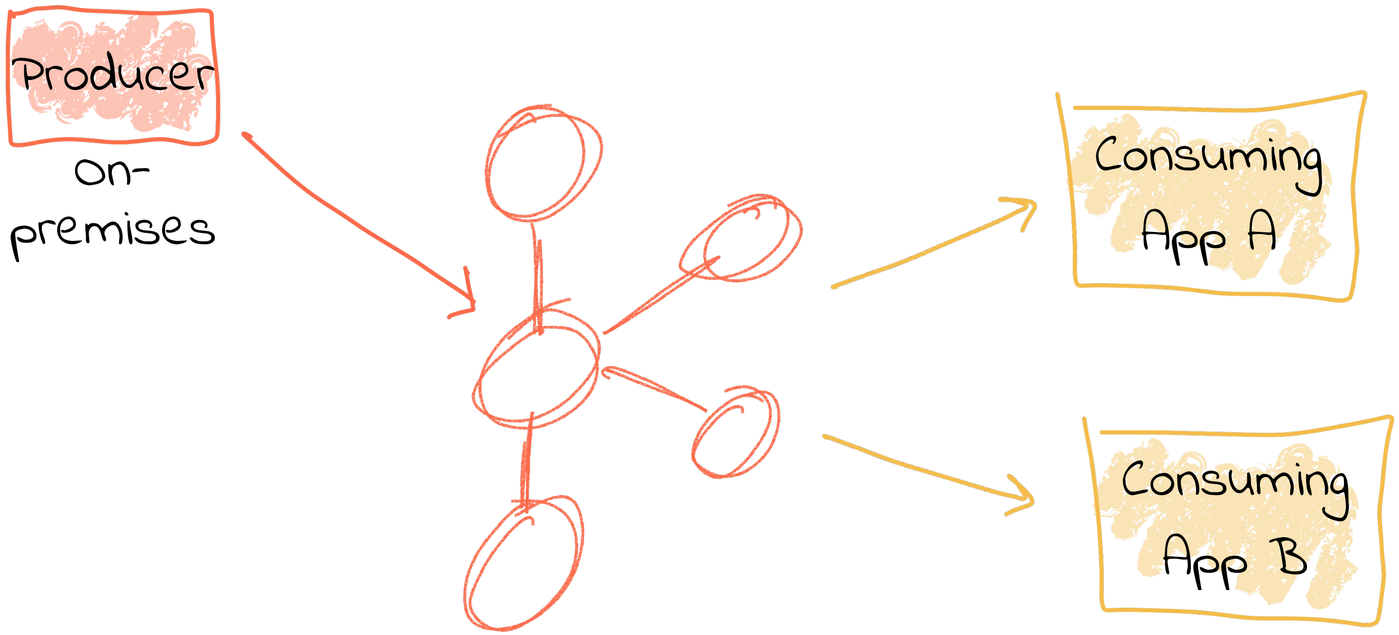
Side-by-Side Tech Evaluation



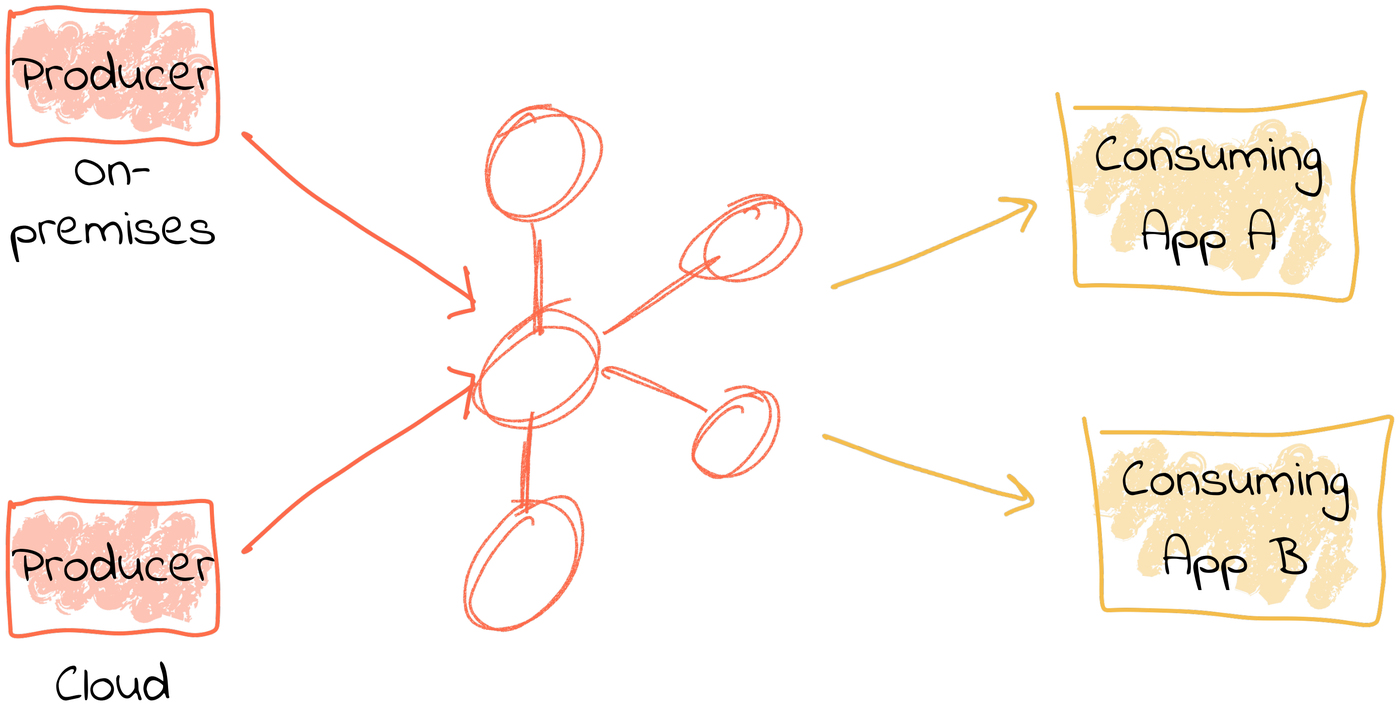
Side-by-Side Tech Evaluation



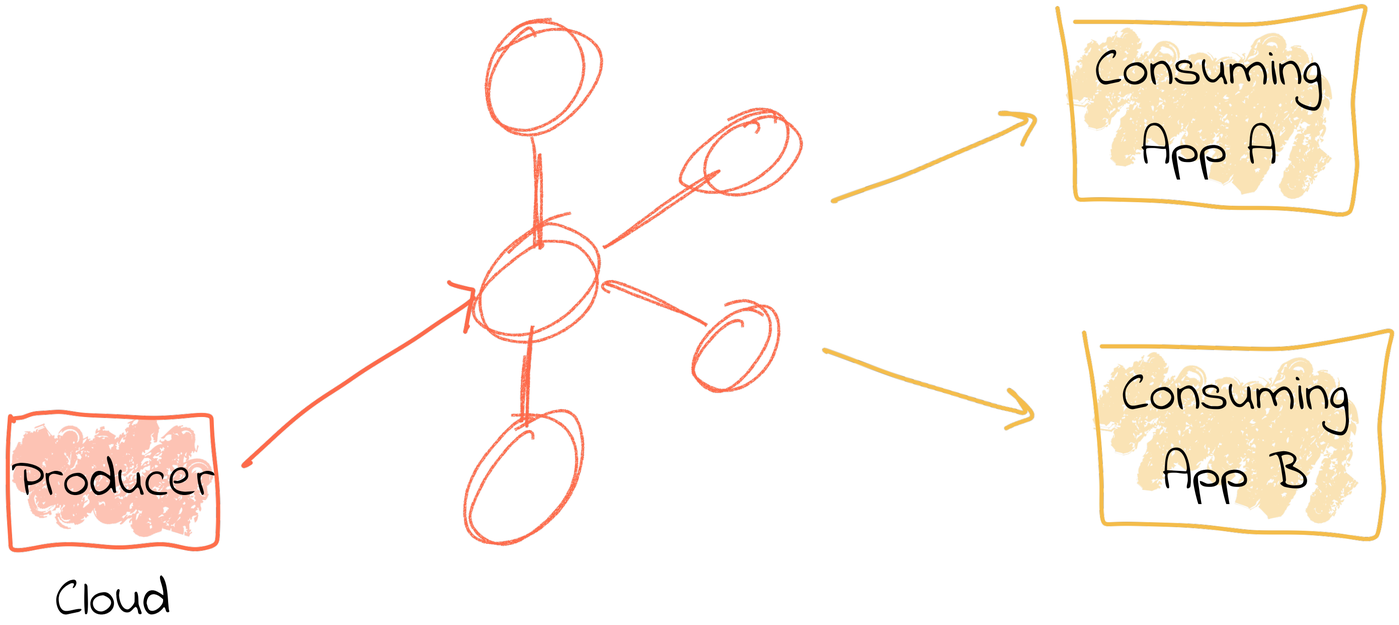
Evolve Data Sources



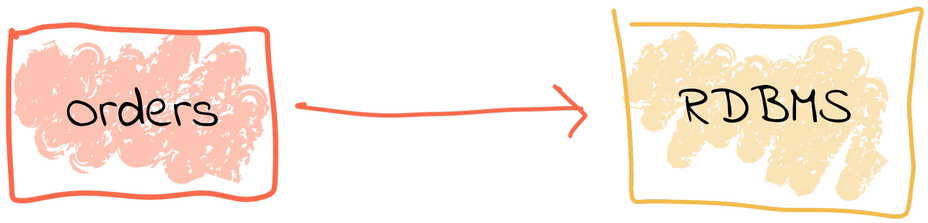
Evolve Data Sources



Evolve Data Sources

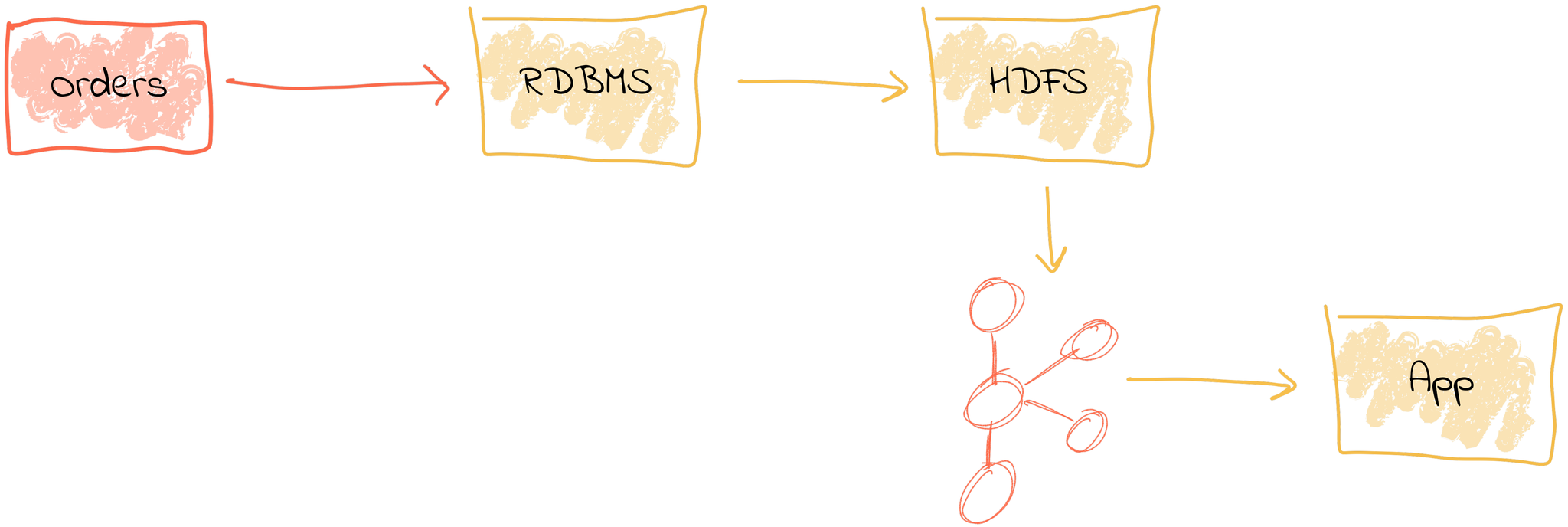


Tight Coupling != Flexible

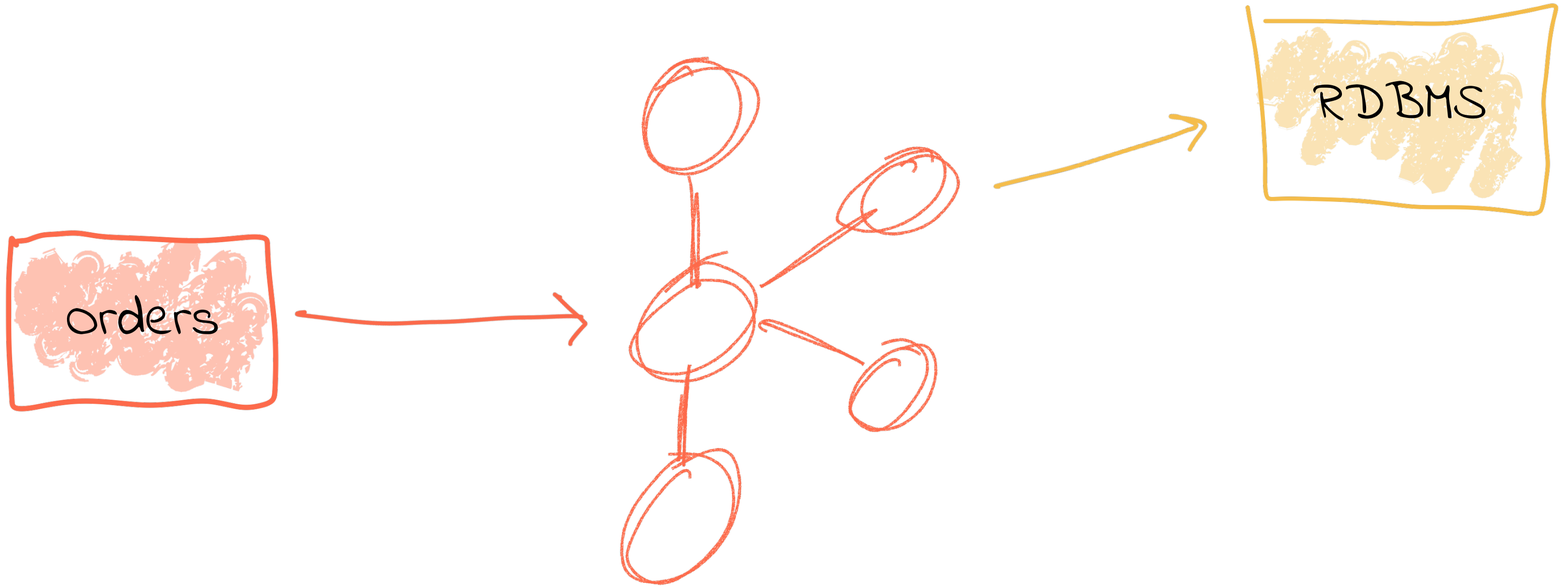


Tight Coupling != Flexible

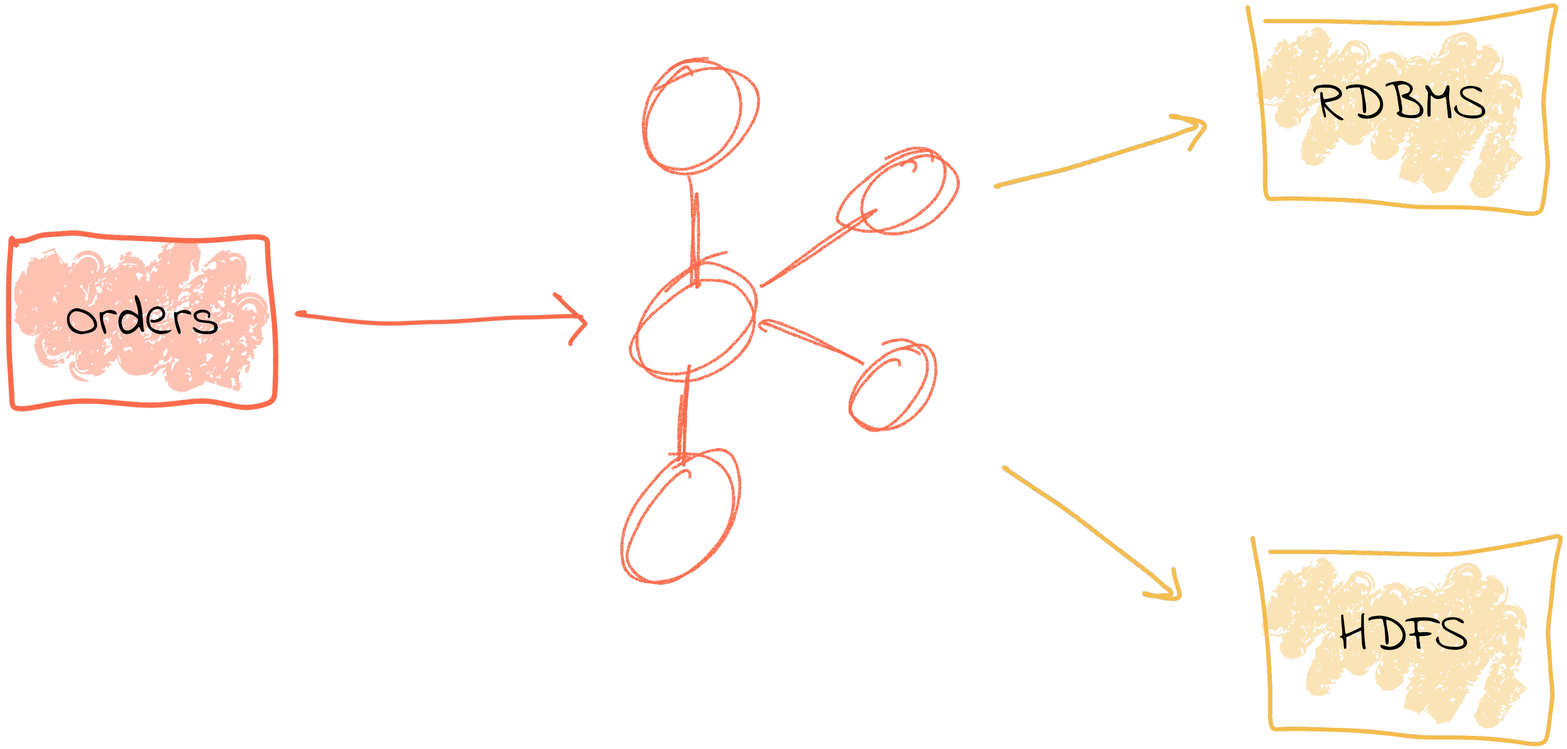
Tight Coupling != Flexible



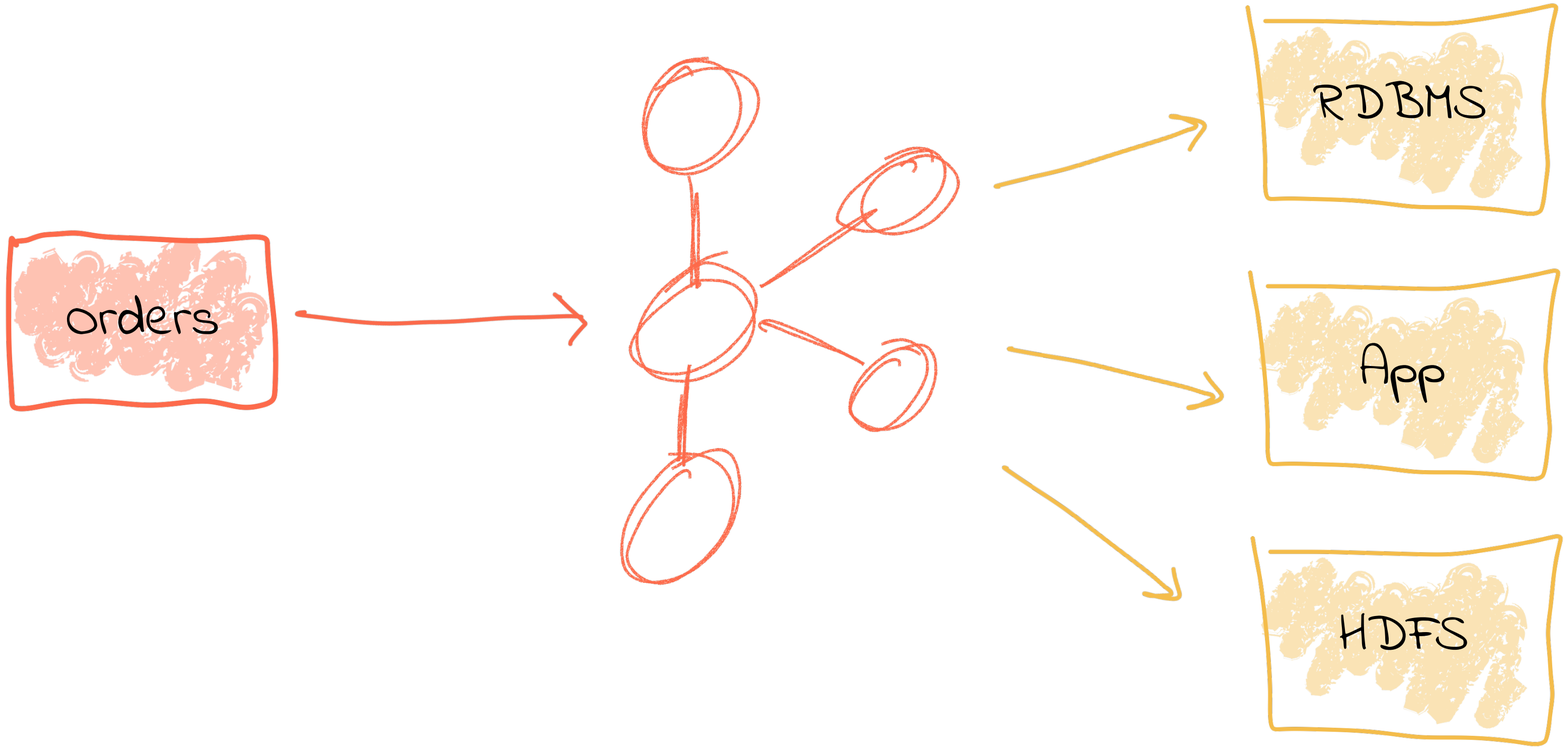
Loose Coupling == Freedom to Evolve



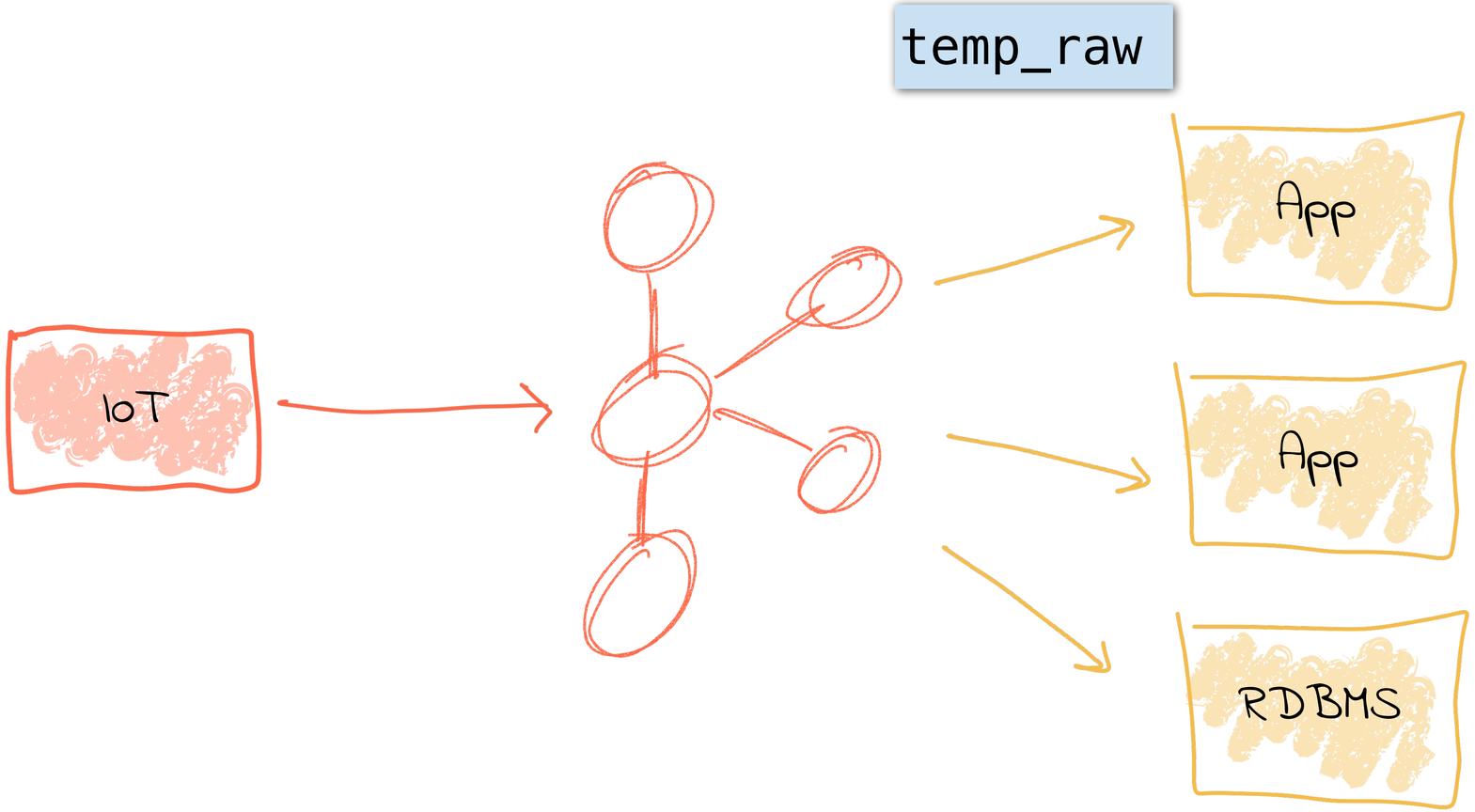
Loose Coupling == Freedom to Evolve



Loose Coupling == Freedom to Evolve

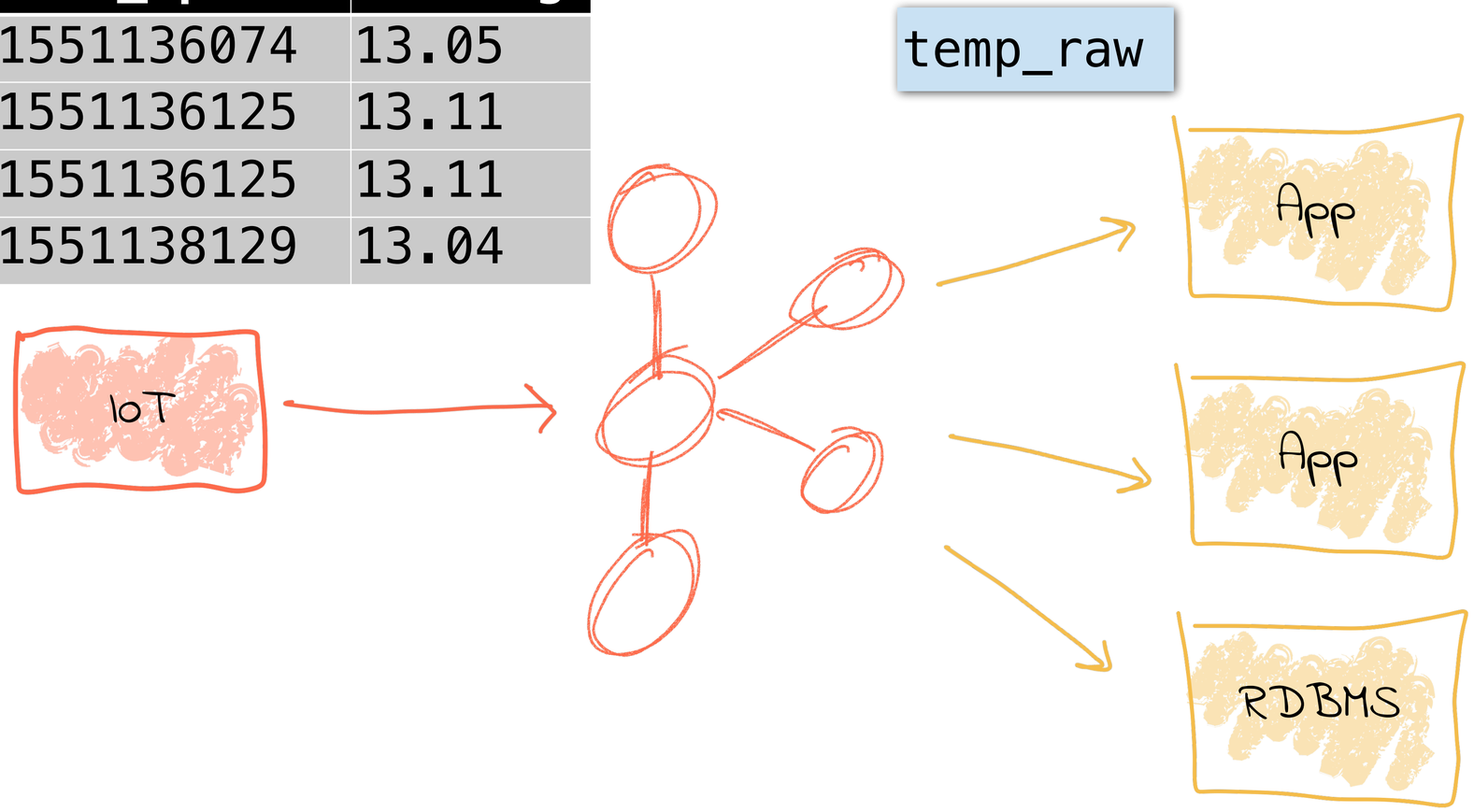


Transform Once, Use Many: Data Cleansing



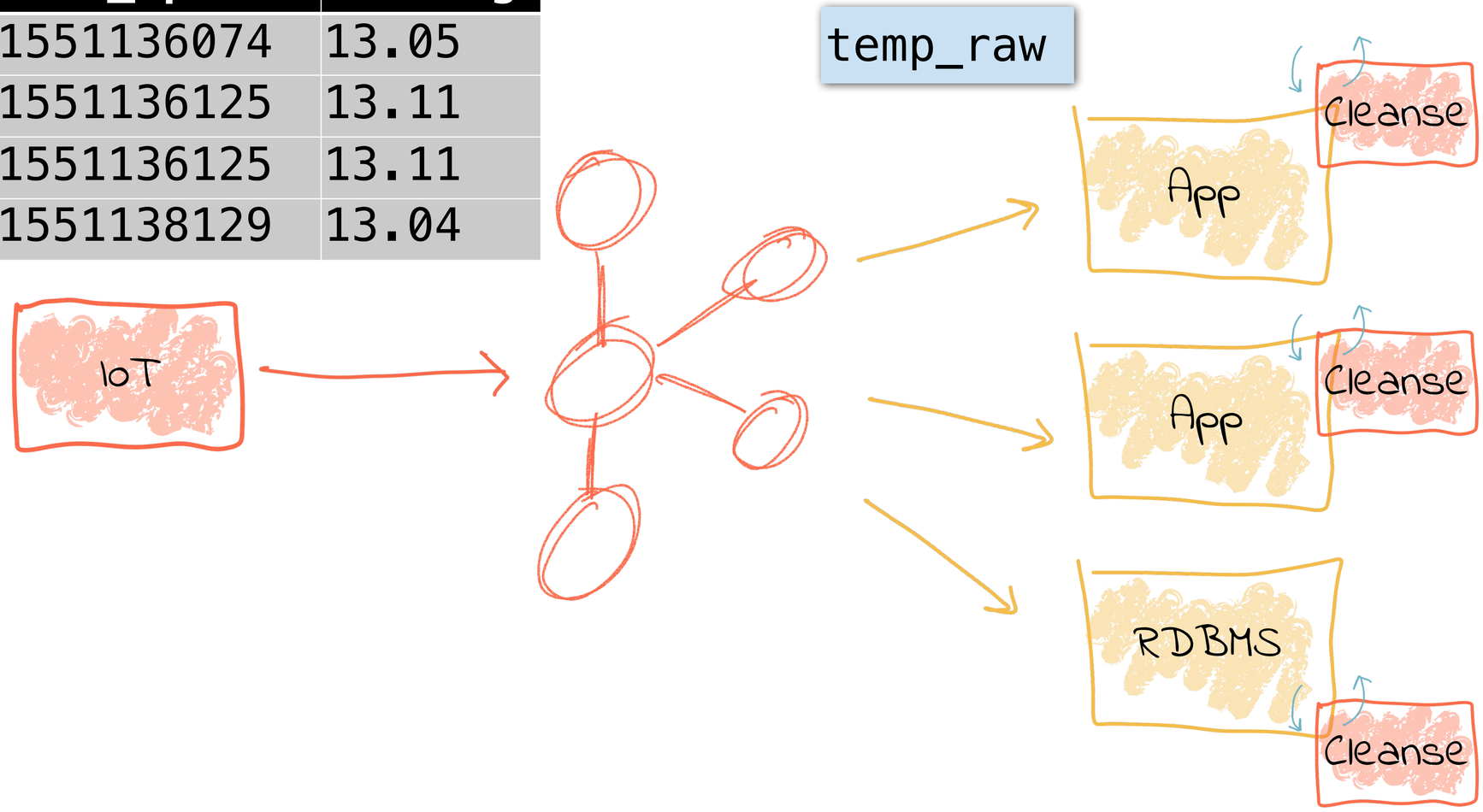
Transform Once, Use Many: Data Cleansing

sensor_id	time_epoch	reading
42	1551136074	13.05
42	1551136125	13.11
	1551136125	13.11
42	1551138129	13.04



Transform Once, Use Many: Data Cleansing

sensor_id	time_epoch	reading
42	1551136074	13.05
42	1551136125	13.11
	1551136125	13.11
42	1551138129	13.04

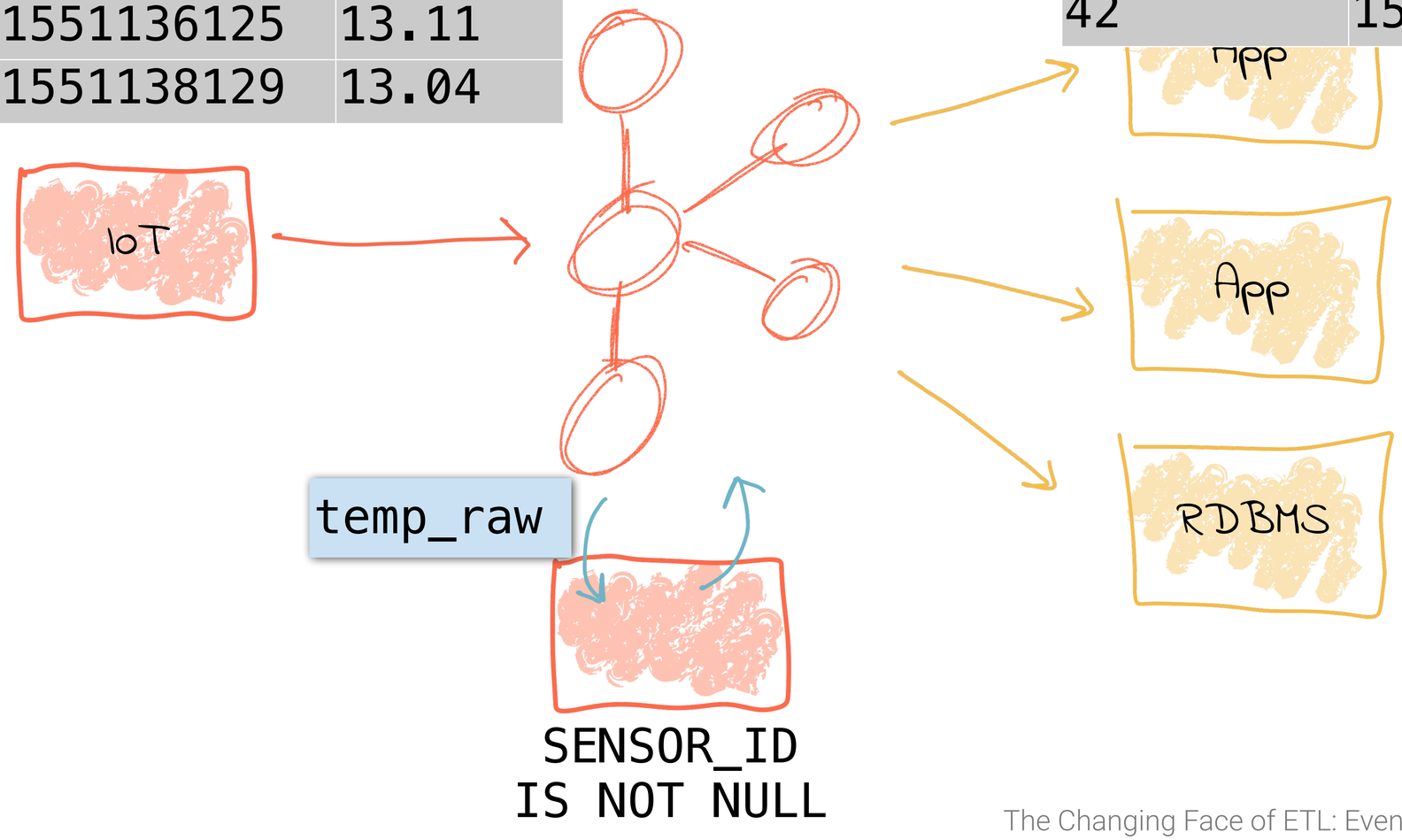


Transform Once, Use Many: Data Cleansing

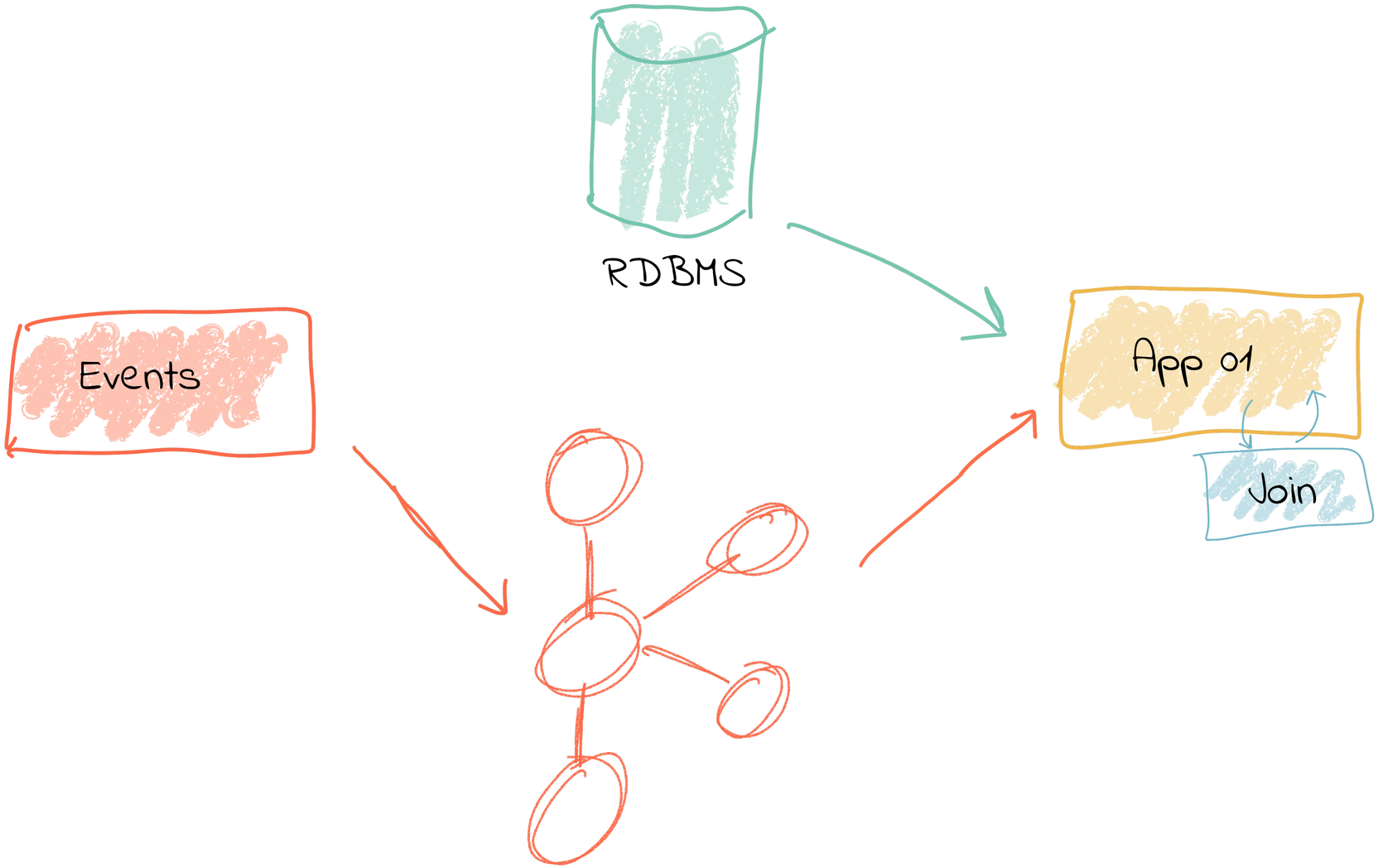
sensor_id	time_epoch	reading
42	1551136074	13.05
42	1551136125	13.11
	1551136125	13.11
42	1551138129	13.04

temp_clean

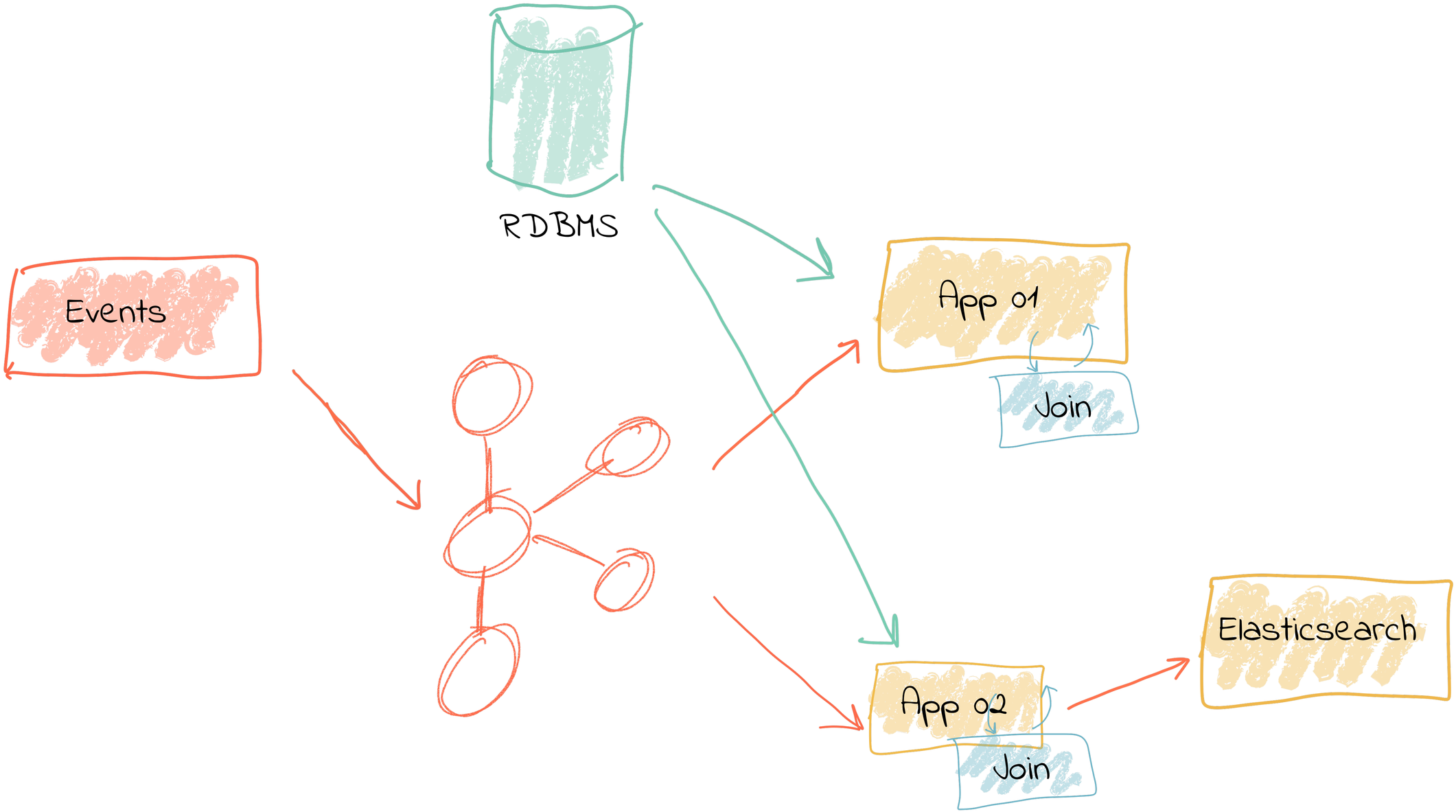
sensor_id	time_epoch	reading
42	1551136074	13.05
42	1551136125	13.11
42	1551138129	13.04



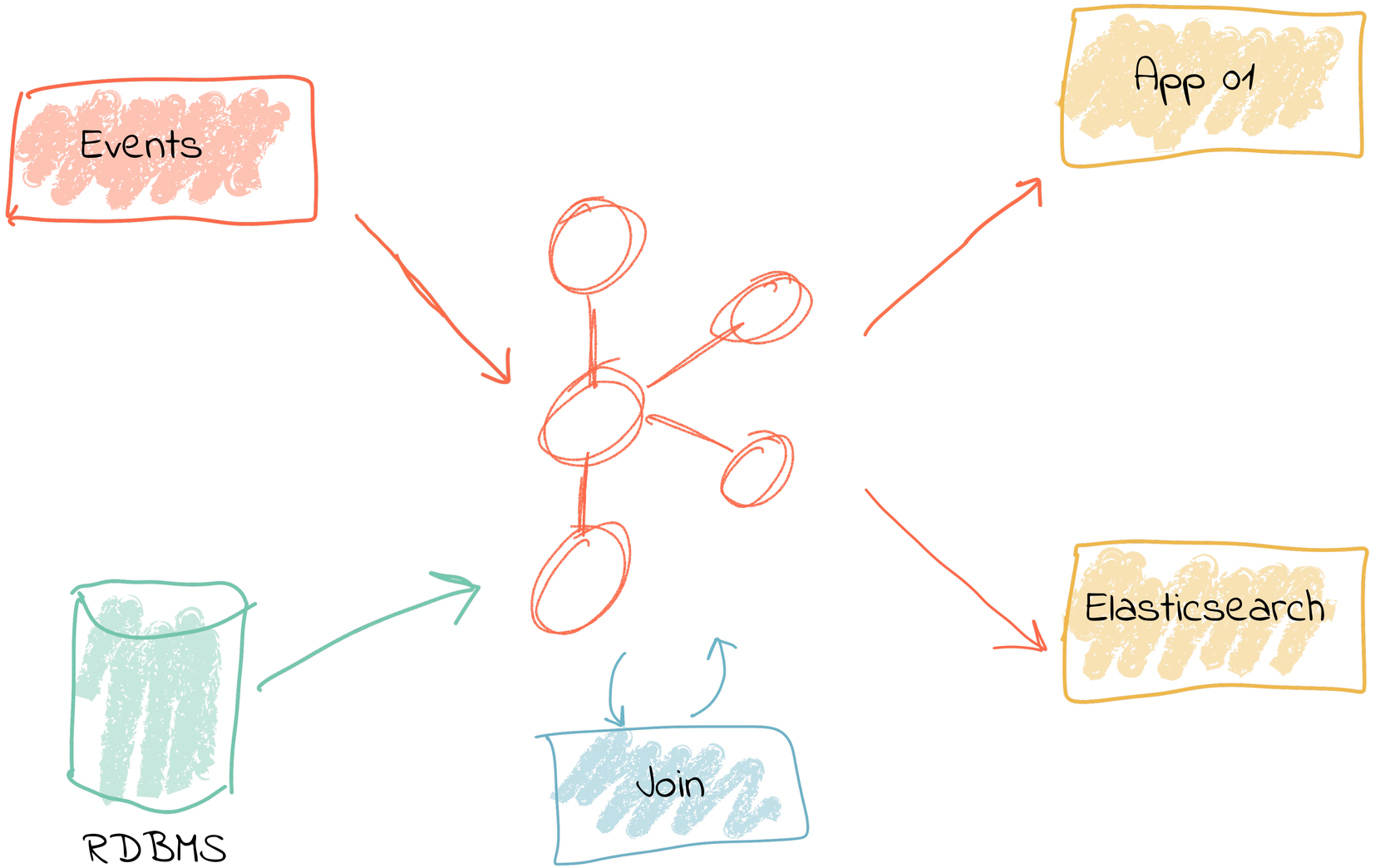
Transform Once, Use Many: Data Enrichment



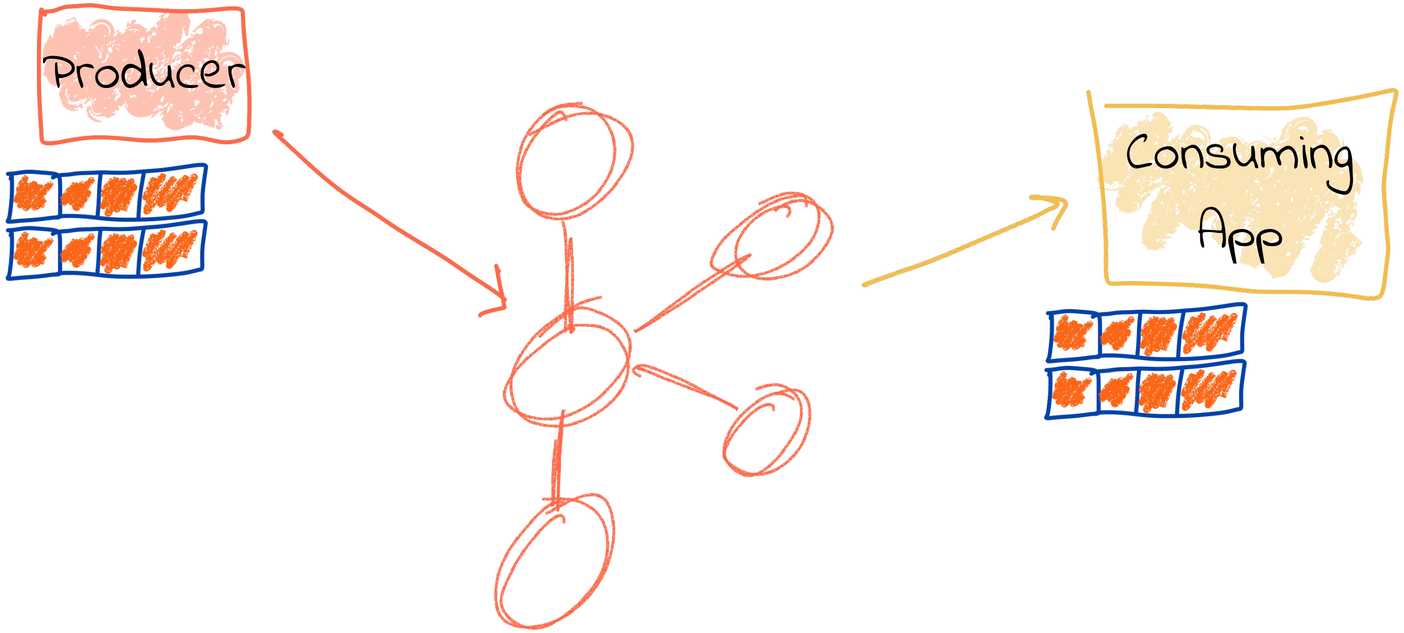
Transform Once, Use Many: Data Enrichment



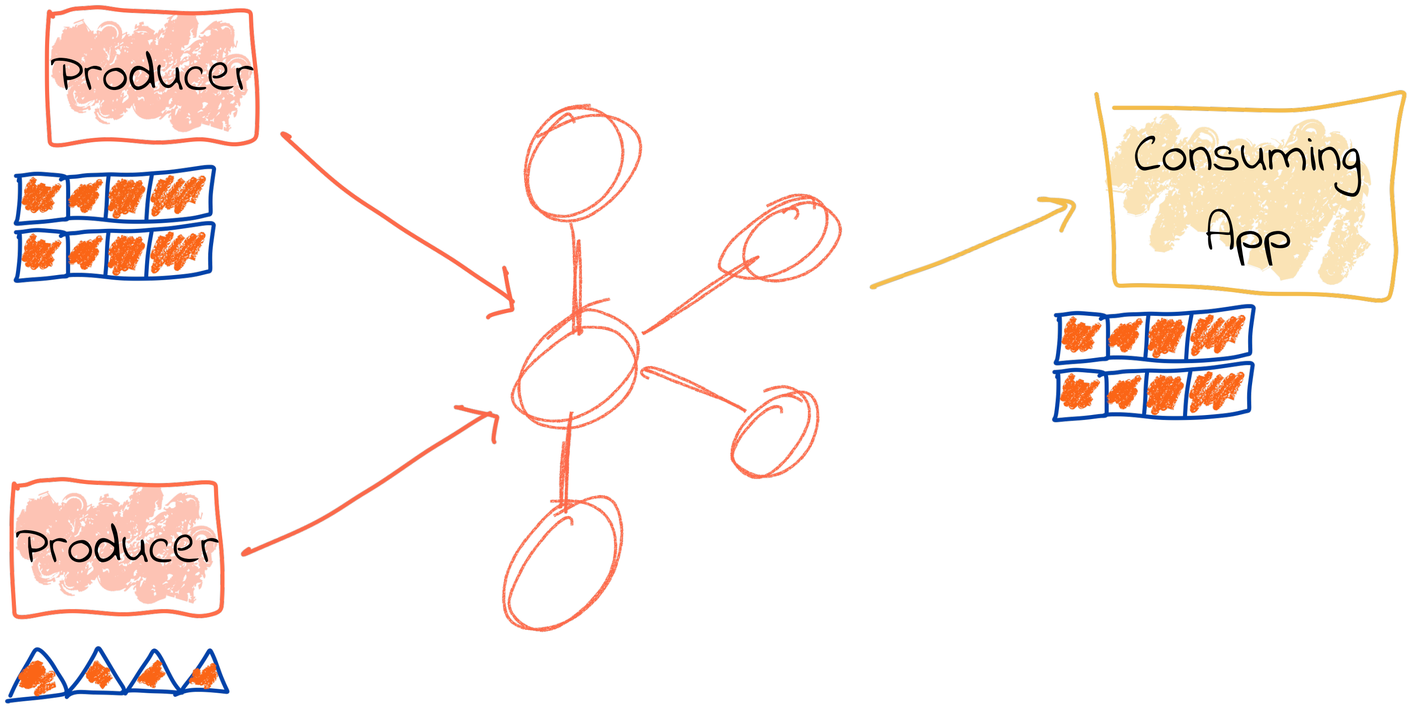
Transform Once, Use Many: Data Enrichment



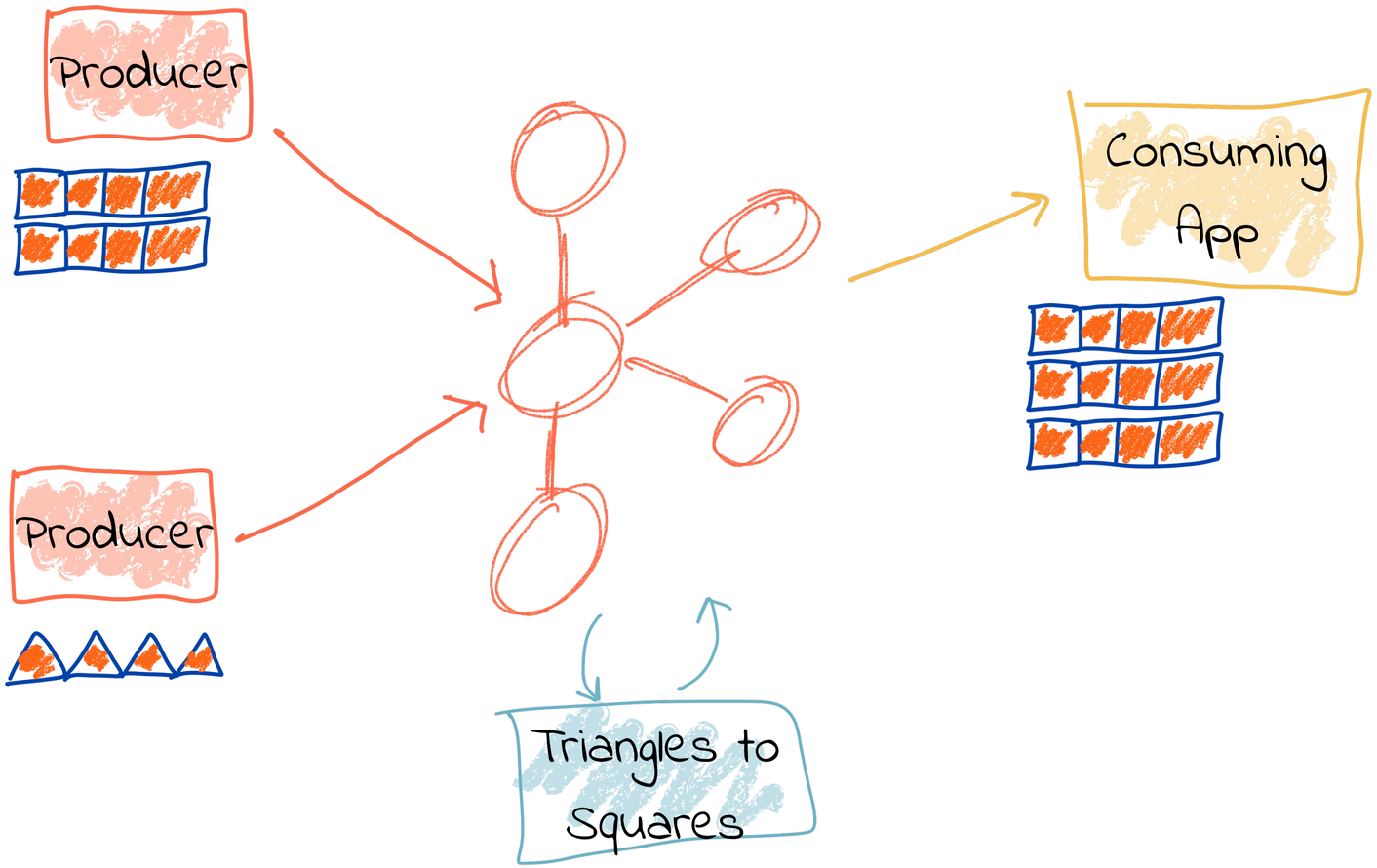
Message Payload Compatibility



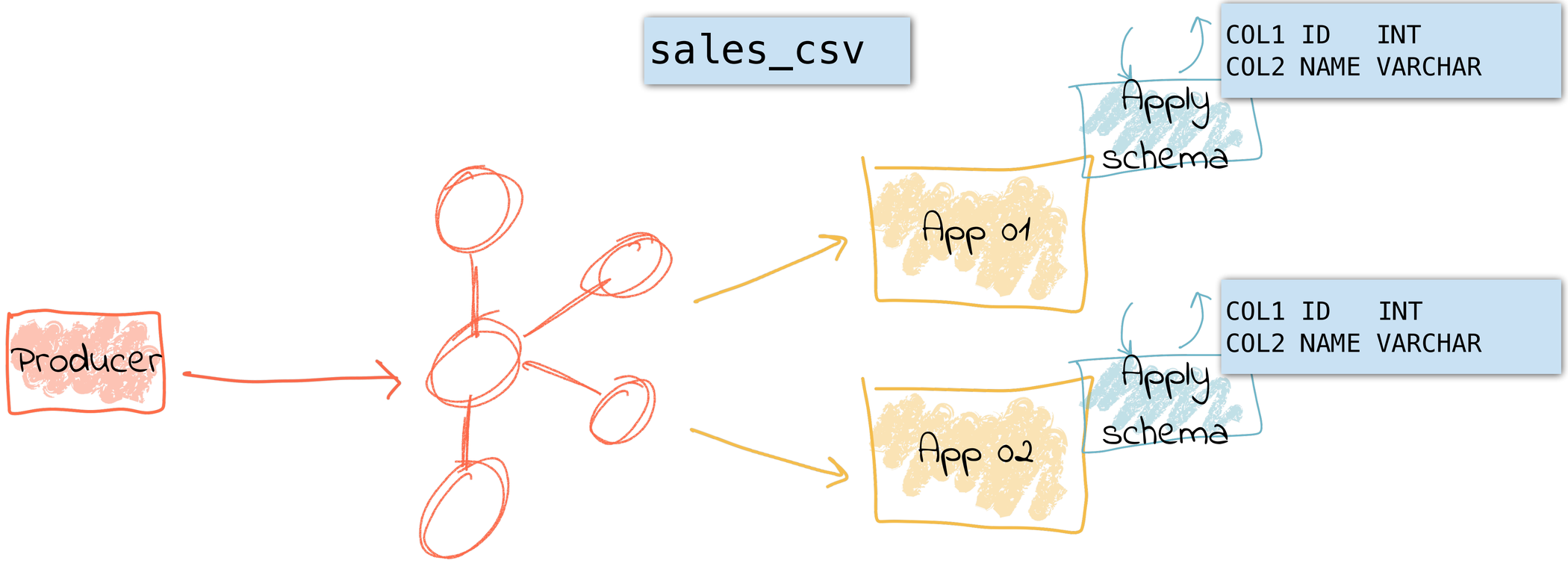
Message Payload Compatibility



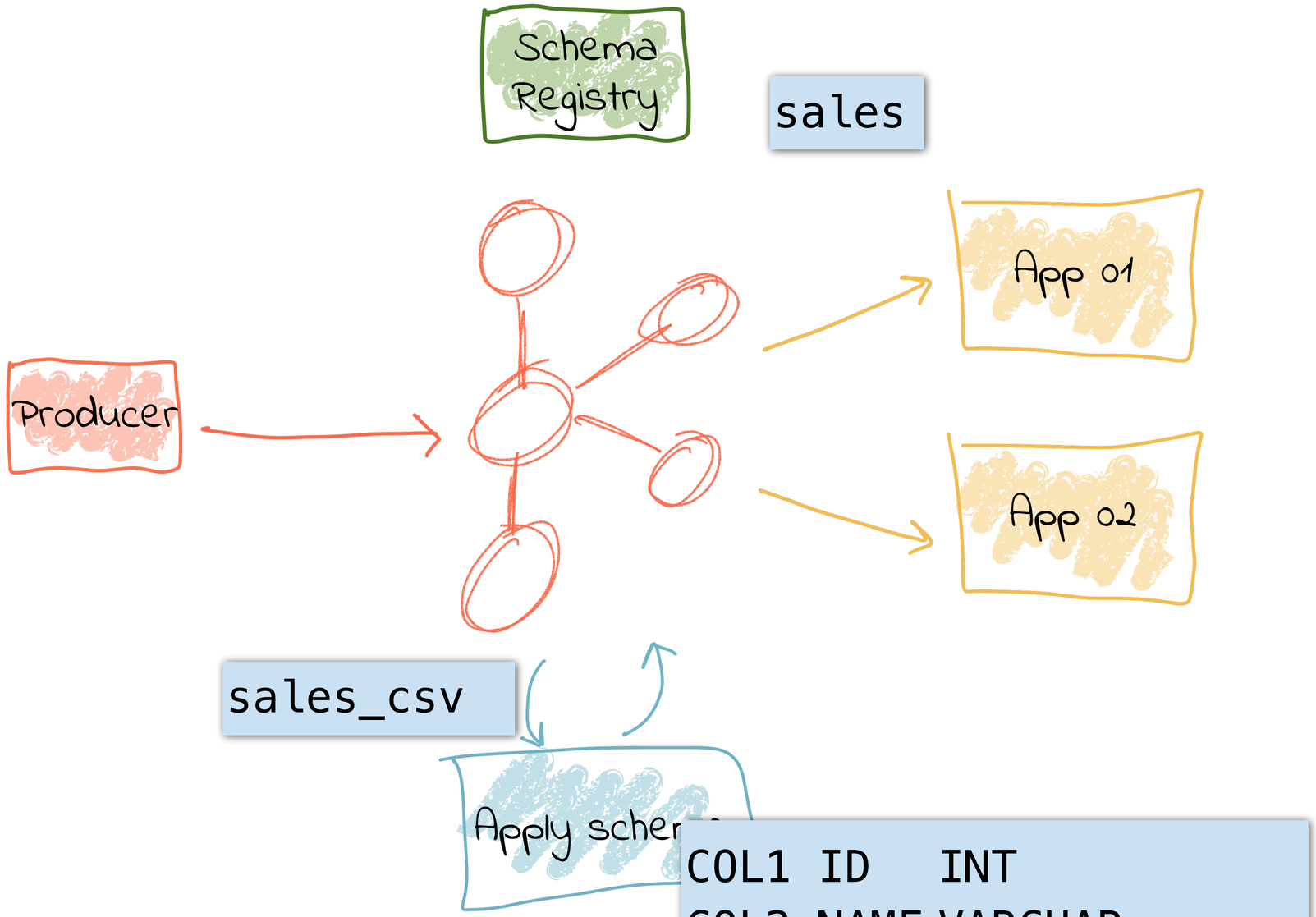
Message Payload Compatibility



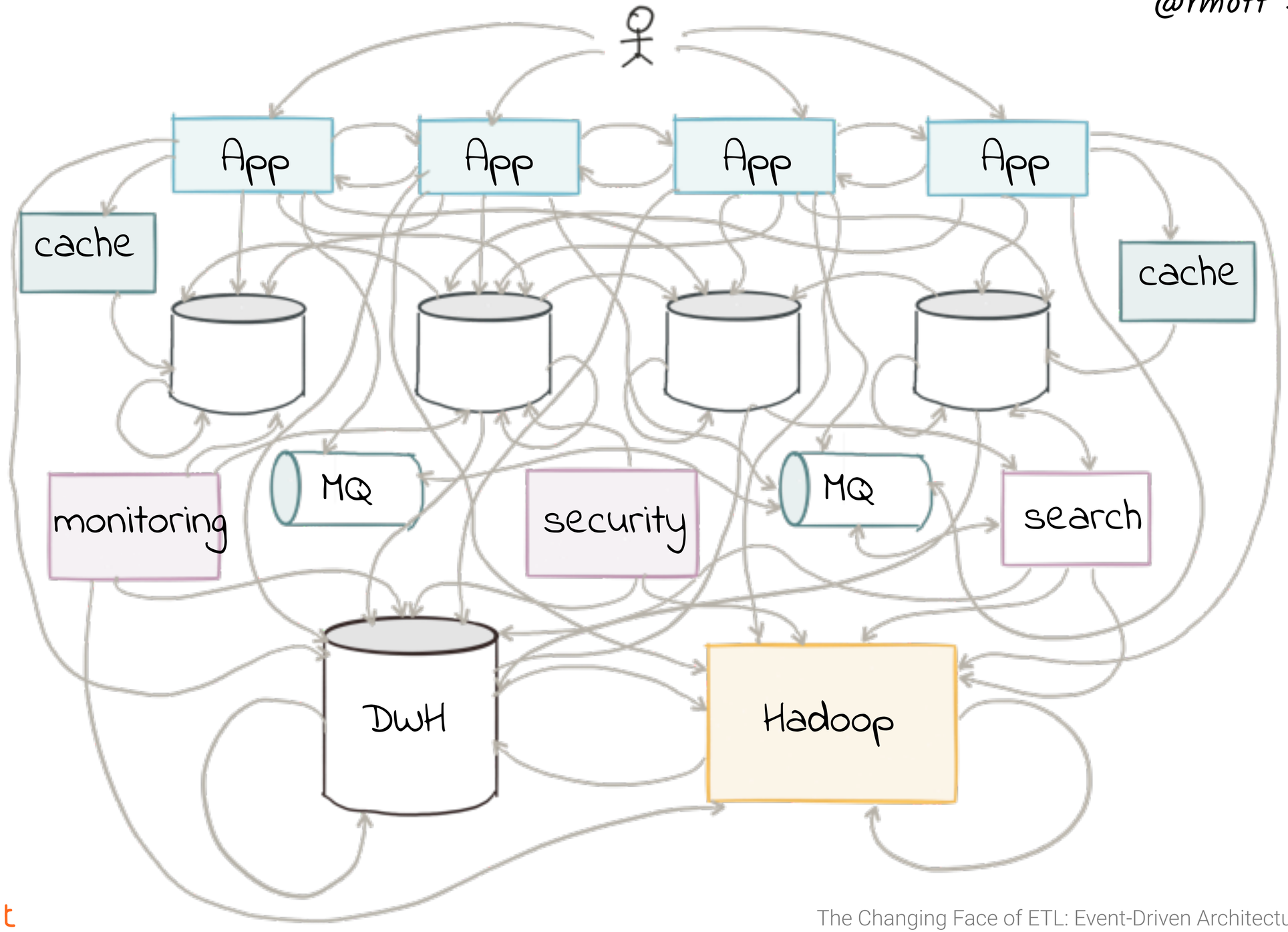
Build Resilient Pipelines with Schemas

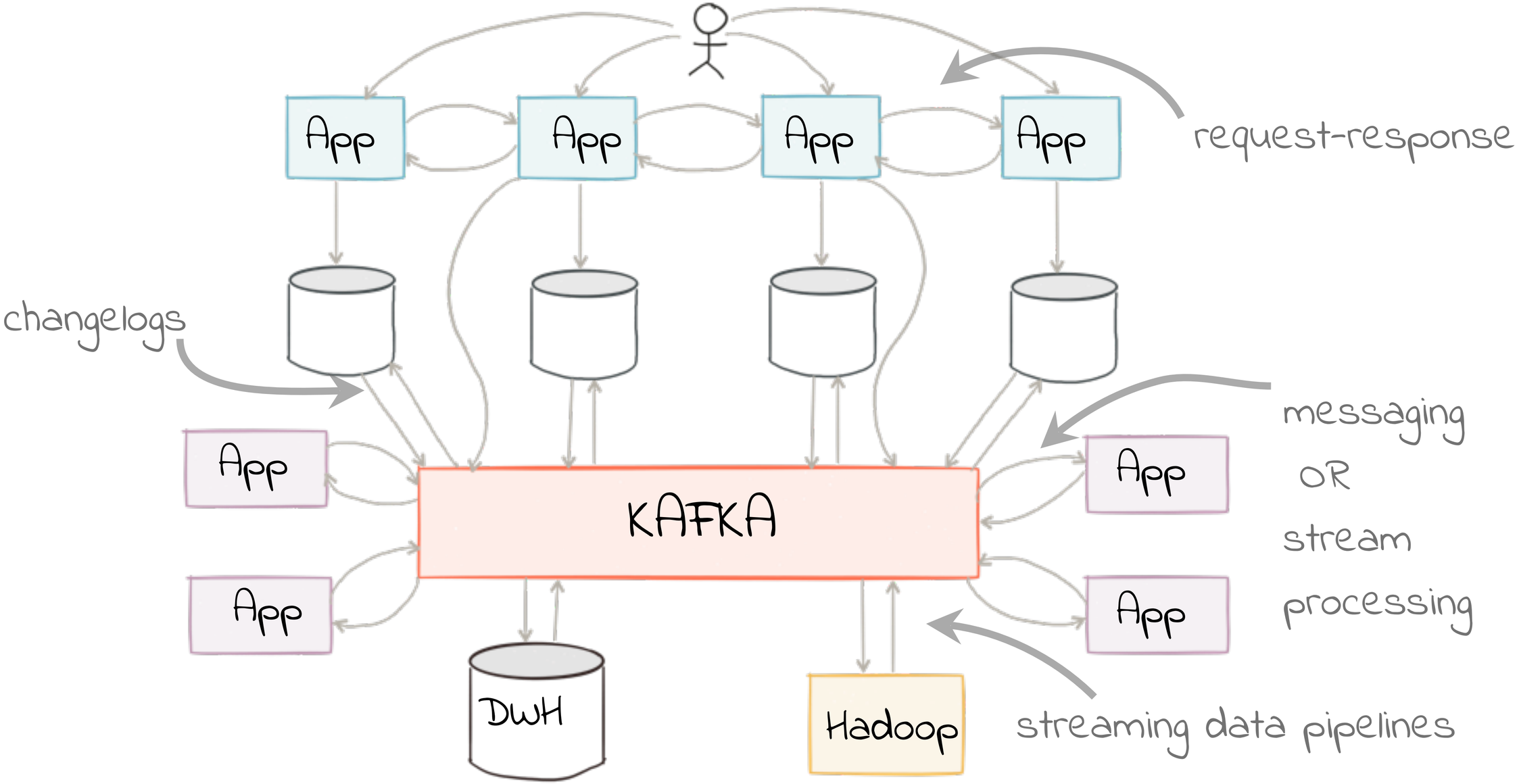


Build Resilient Pipelines with Schemas




Say NO to brittle pipelines
Photo by rmoff


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


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

Events model the real world

Photo by rmoff

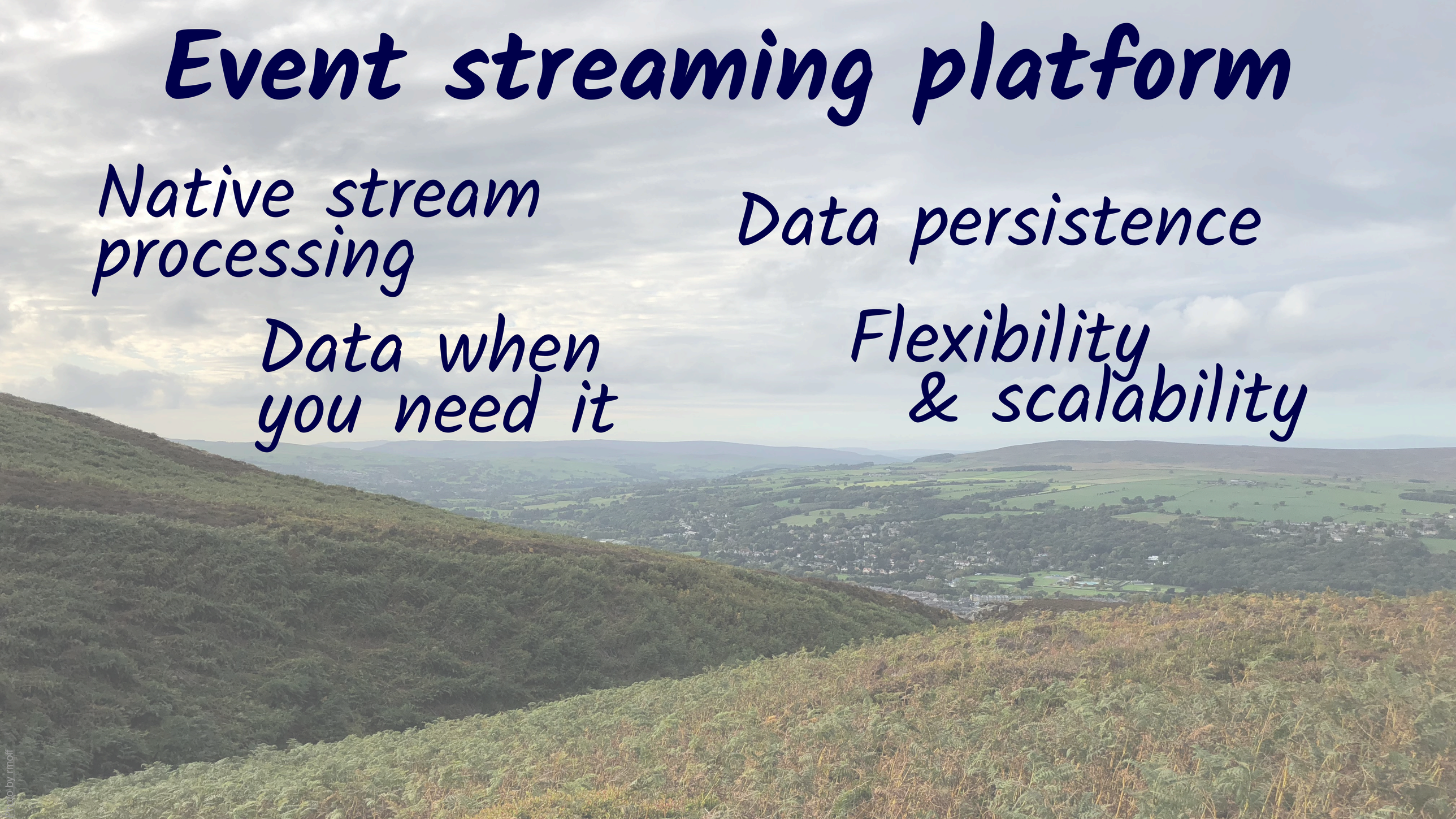
Event streaming platform

Native stream processing

Data persistence

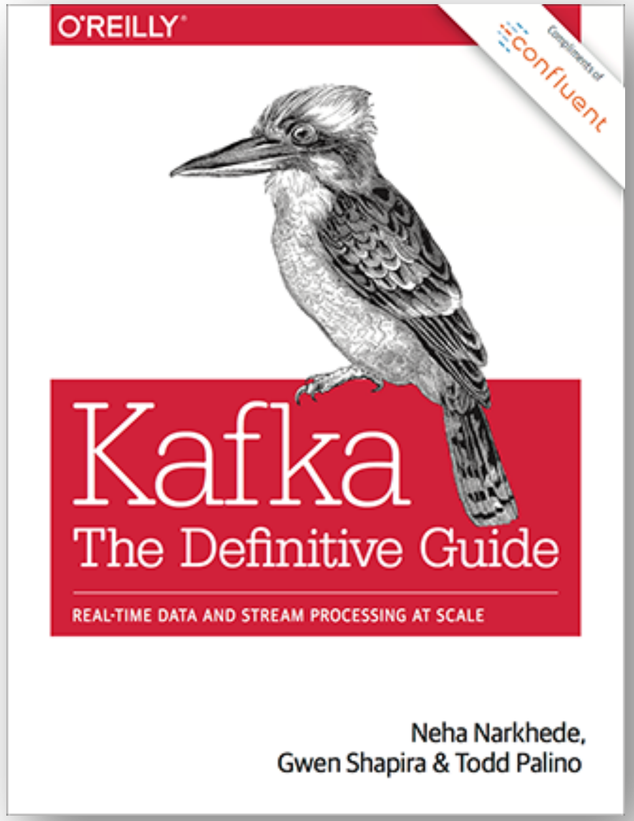
Data when you need it

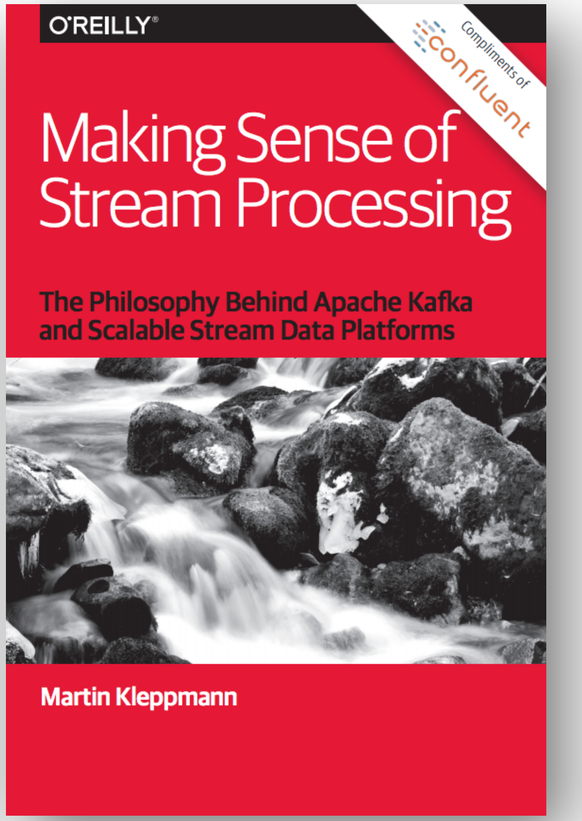
Flexibility & scalability

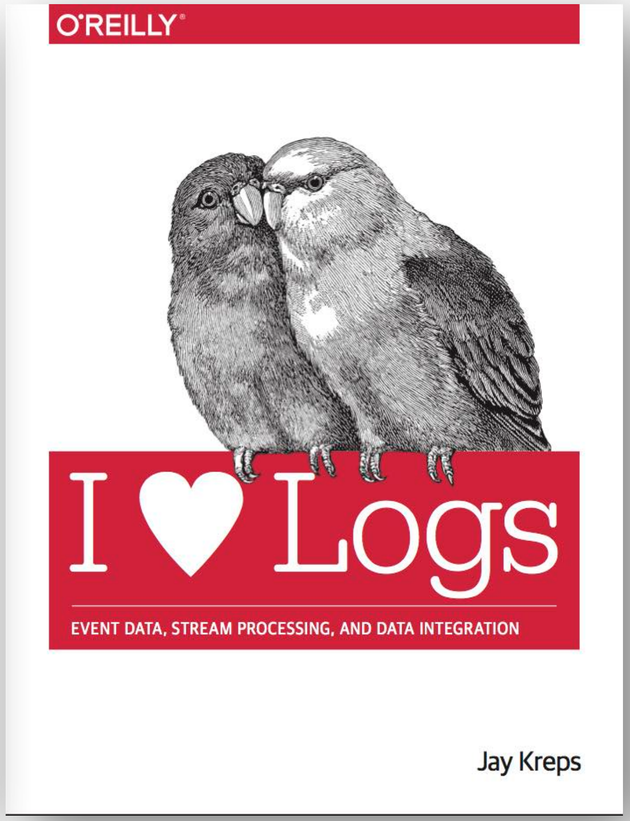


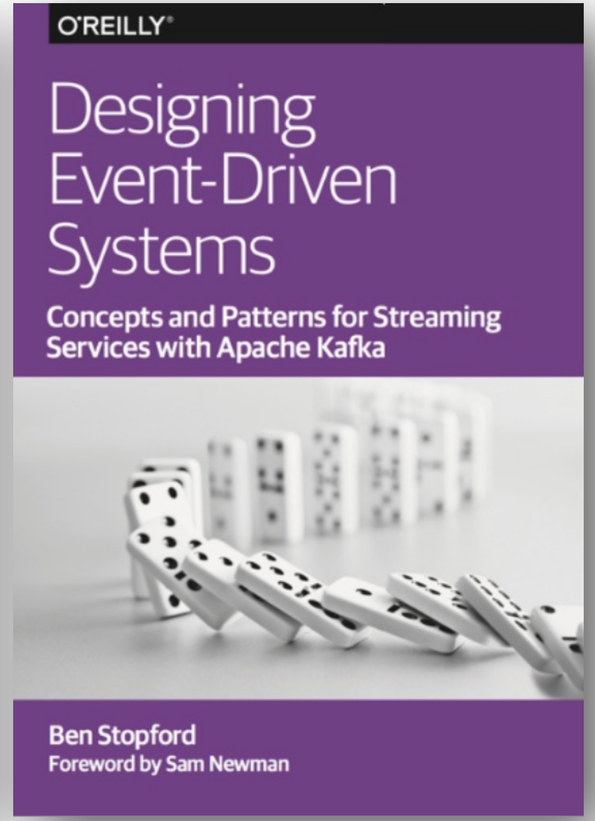
Fully Managed Kafka as a Service

http://cnfl.io/book-bundle










@rmoff
#OReillySACon
Photo by rmoff
https://talks.rmoff.net
http://cnfl.io/slack