confluent.io/ksql

Apache Kafka and KSQL in Action : Let's Build a Streaming Data Pipeline!

USENIX Large Installation System Administration Conference (LISA)
October 31 2018

@rmoff robin@confluent.io

$ whoami

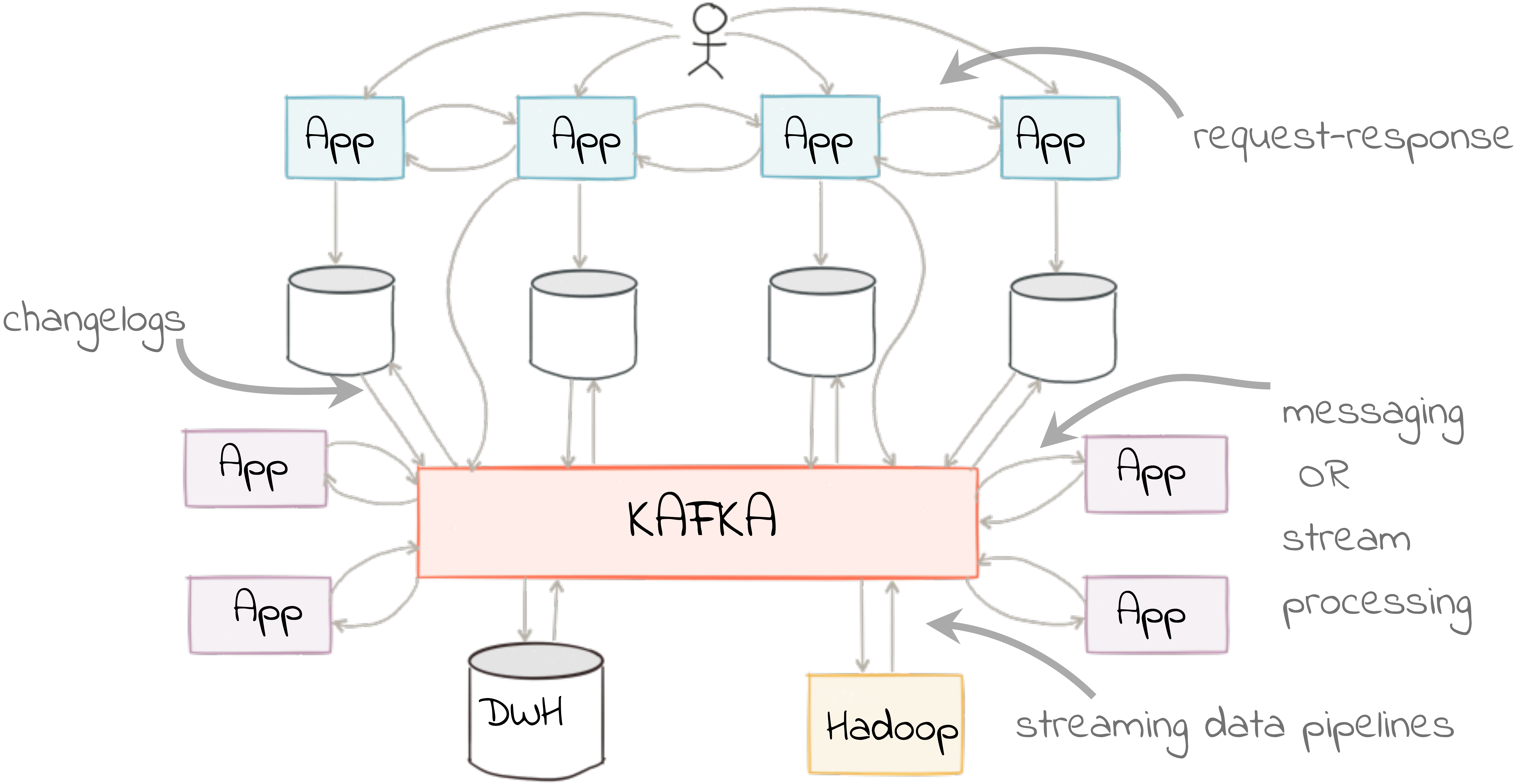
- Developer Advocate @ Confluent
- Working in data & analytics since 2001
- Oracle Developer Champion
- Blogging : http://rmoff.net & http://cnfl.io/rmoff
- Twitter:
  - Geek stuff
  - Beer & Fried Breakfasts

@rmoff

https://speakerdeck.com/rmoff/

Kafka

Kafka is a Streaming Platform

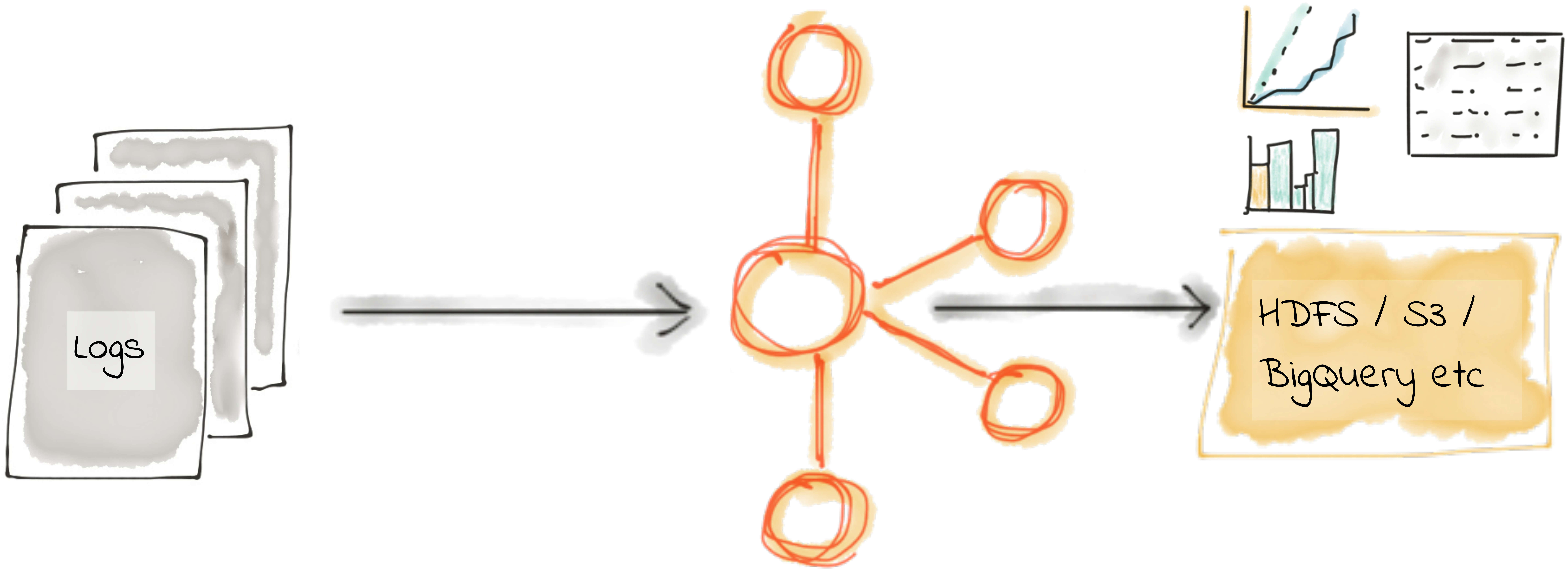


Streaming is not just for realtime

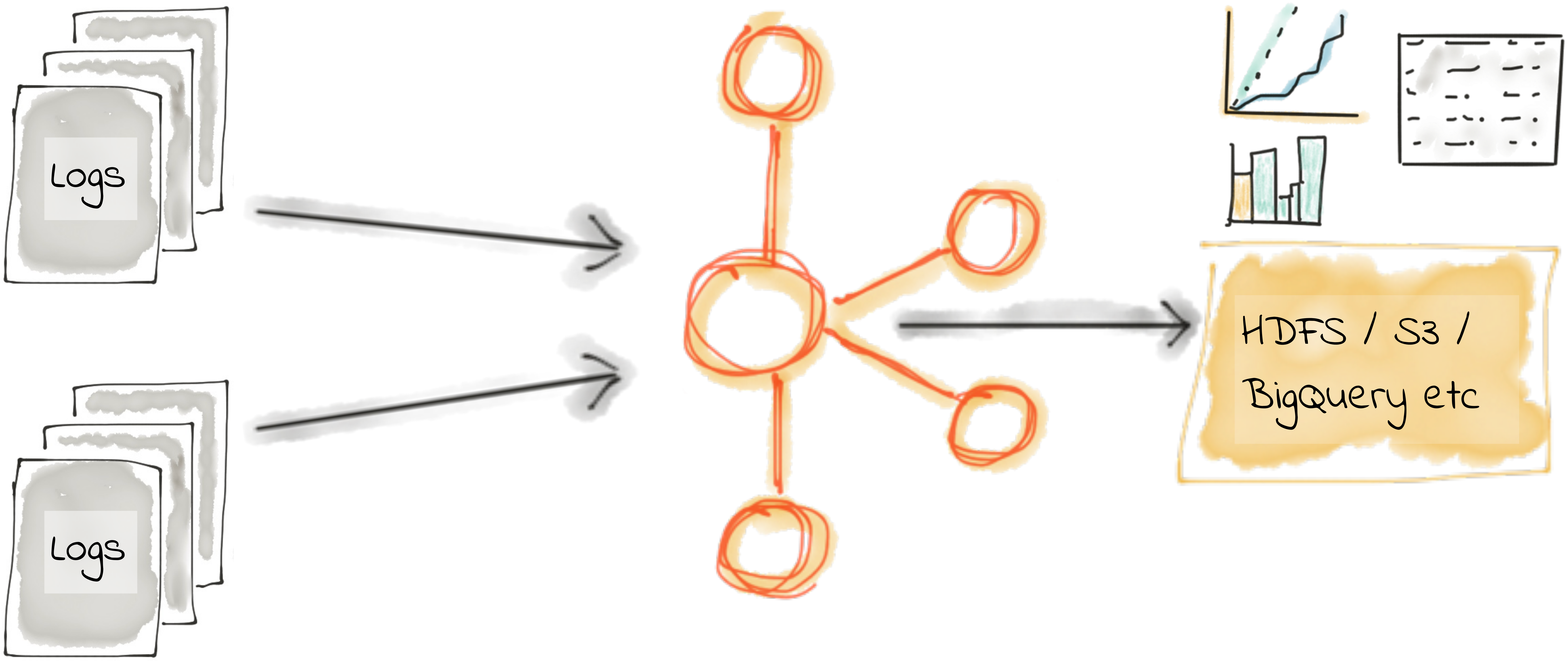
Streaming is for everyone

All data is events

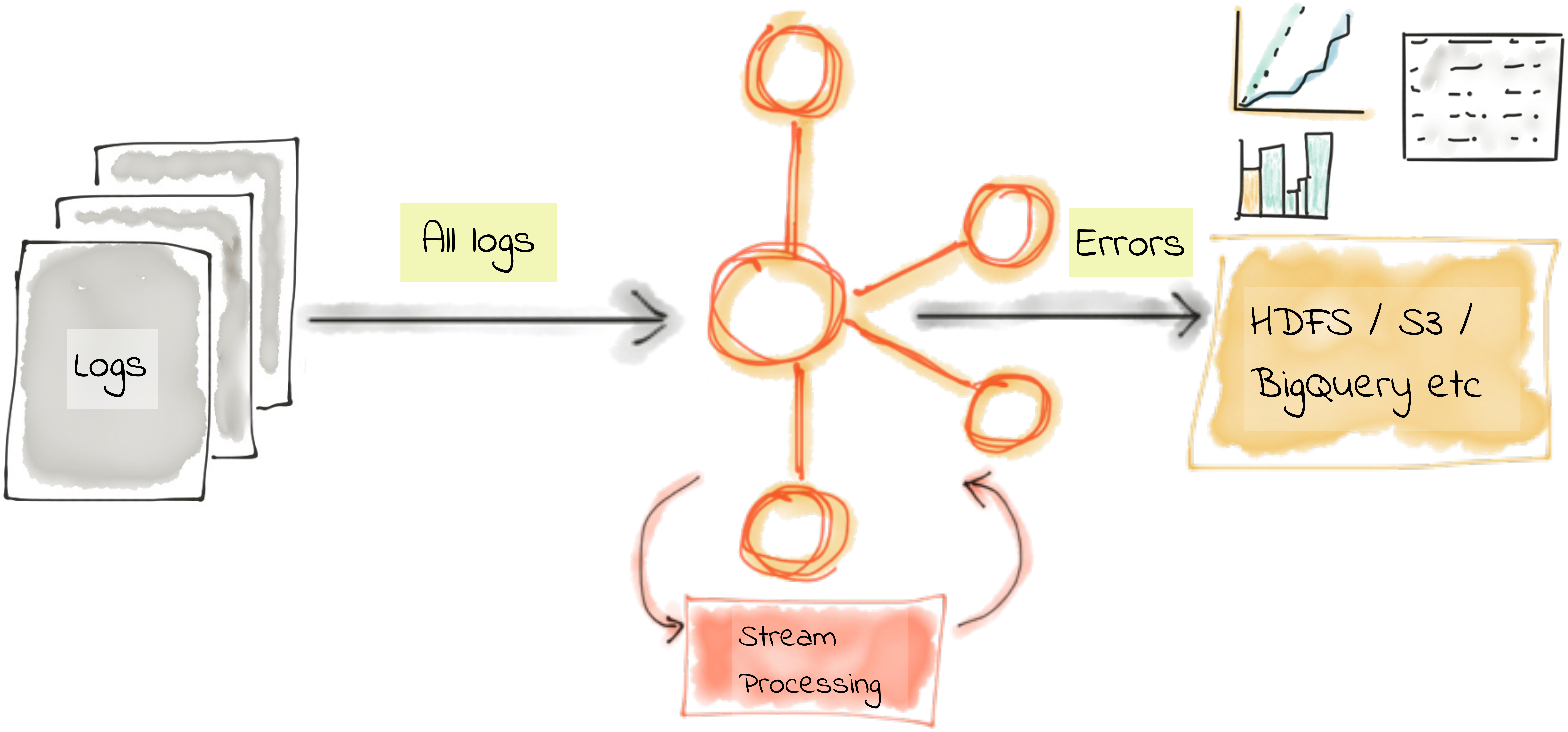
A Dumb Pipeline



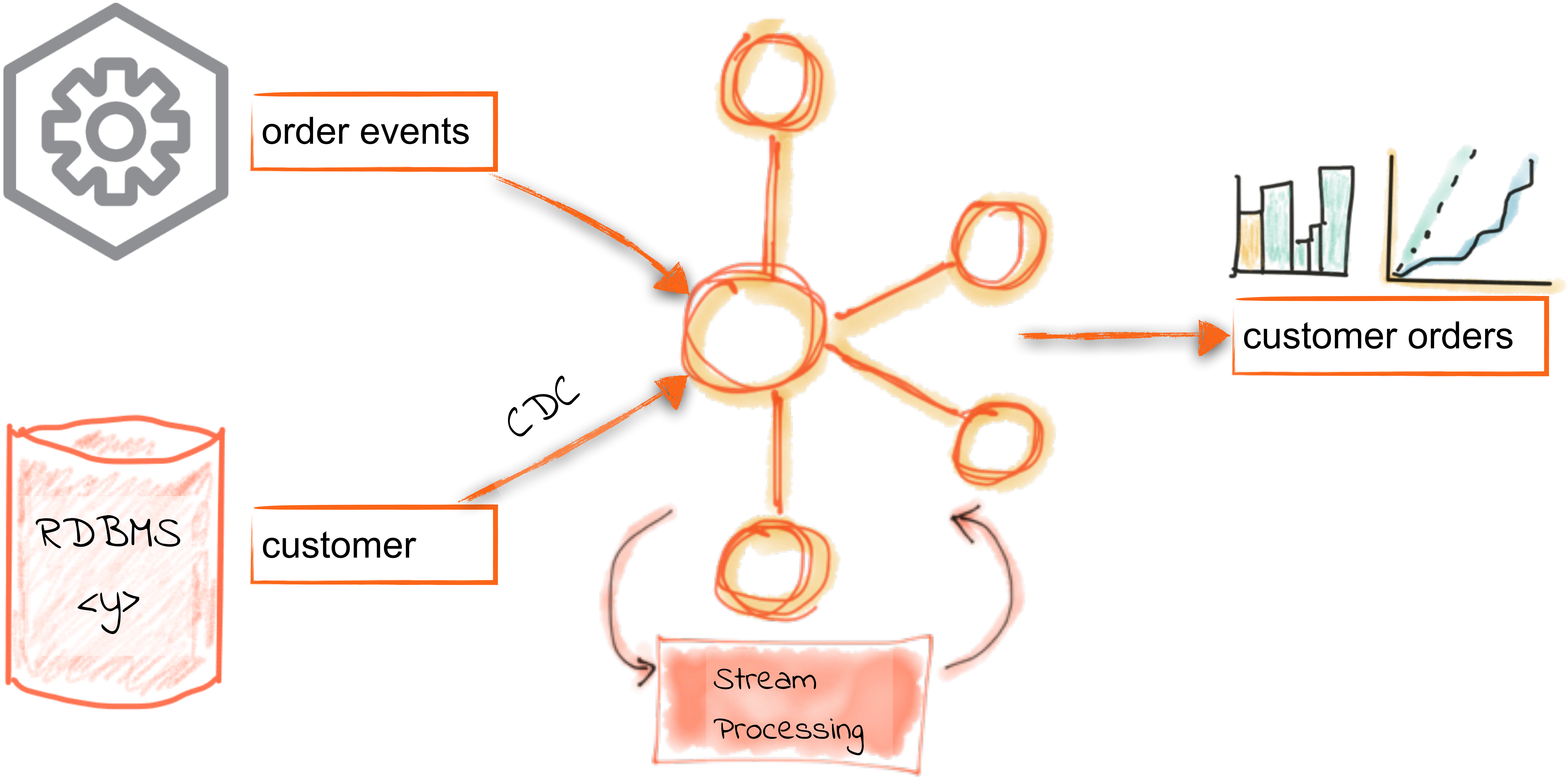
A Dumb Pipeline



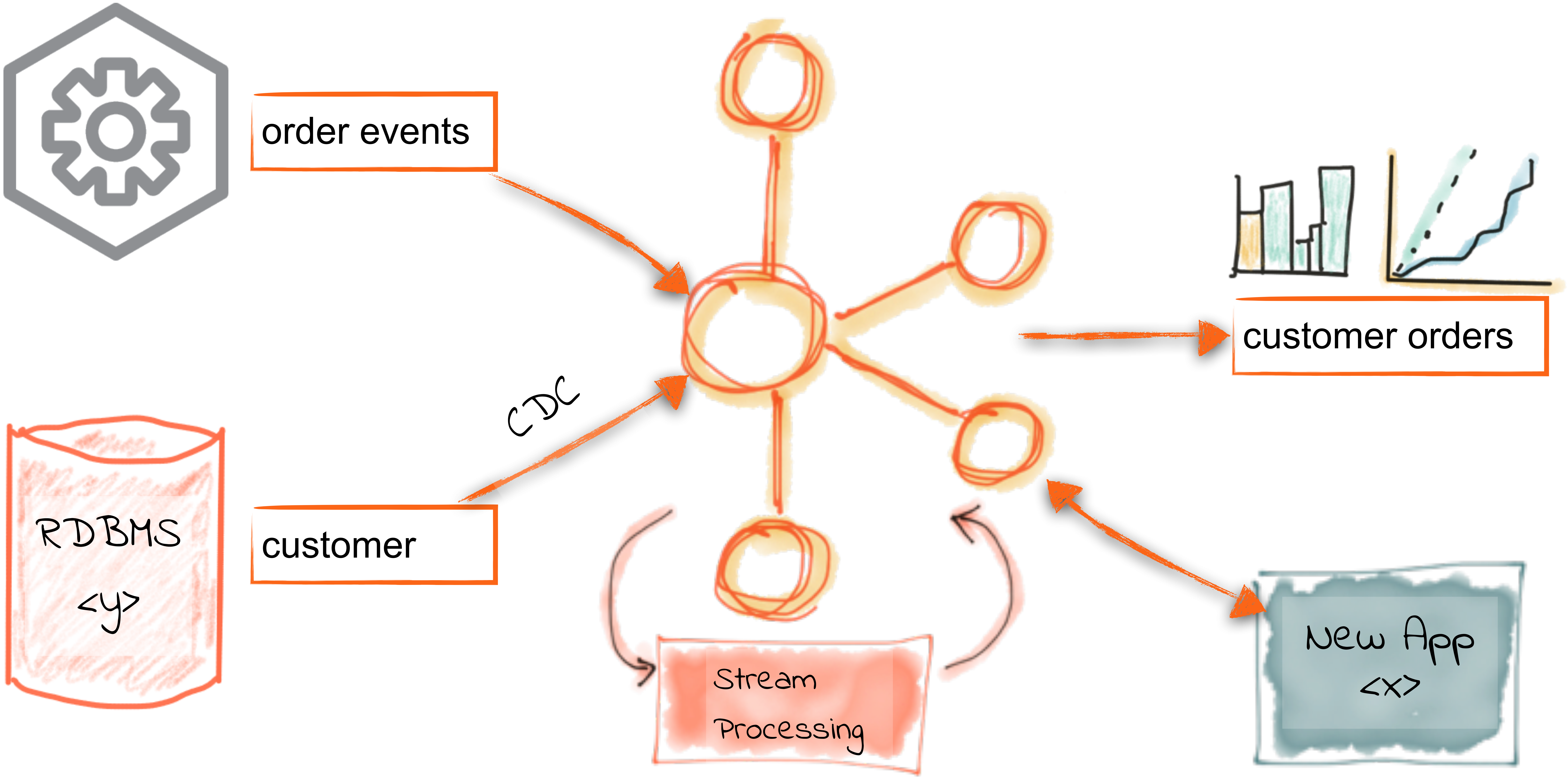
Stream Processing with Apache Kafka and KSQL



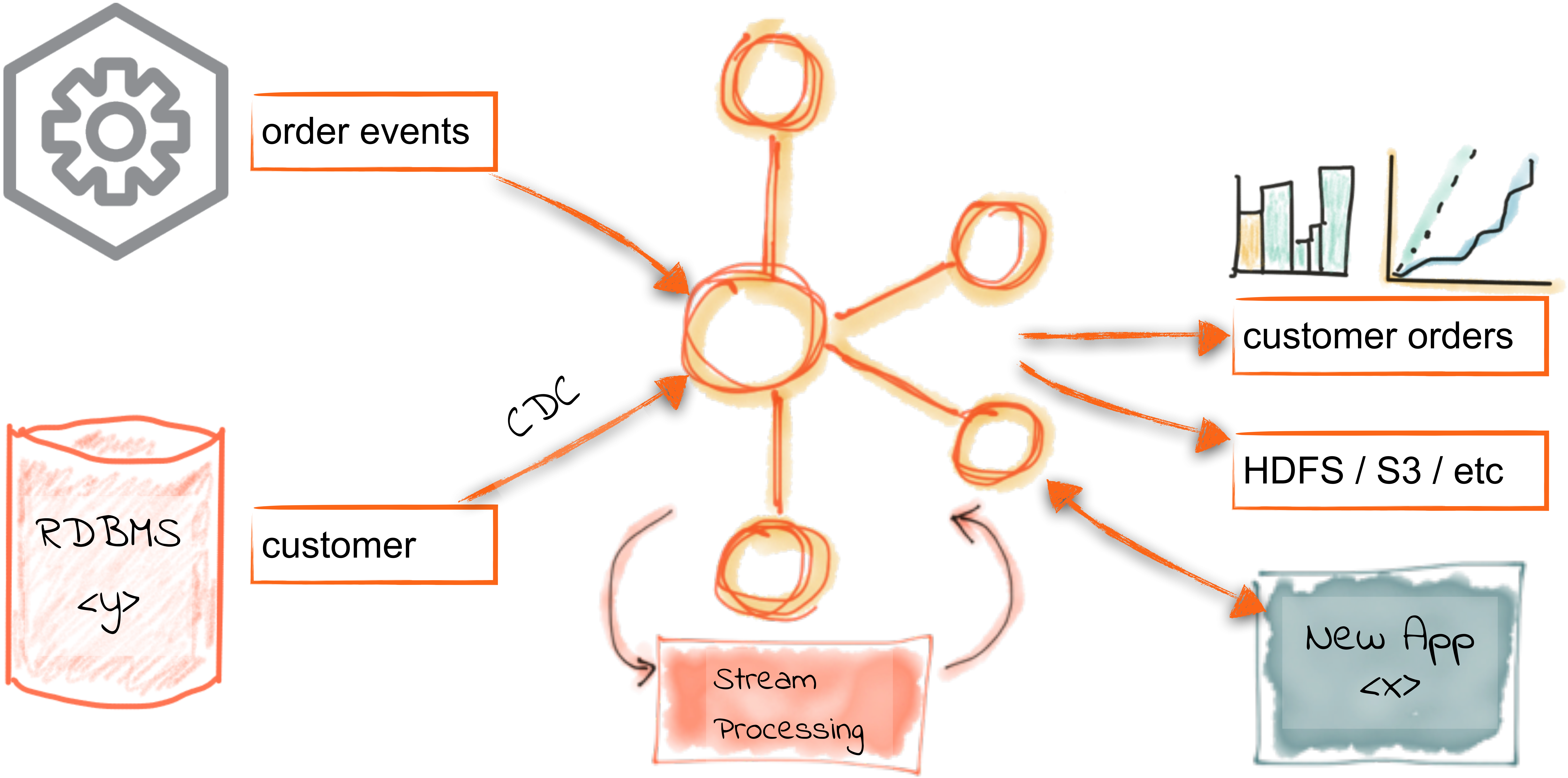
Real-time Event Stream Enrichment



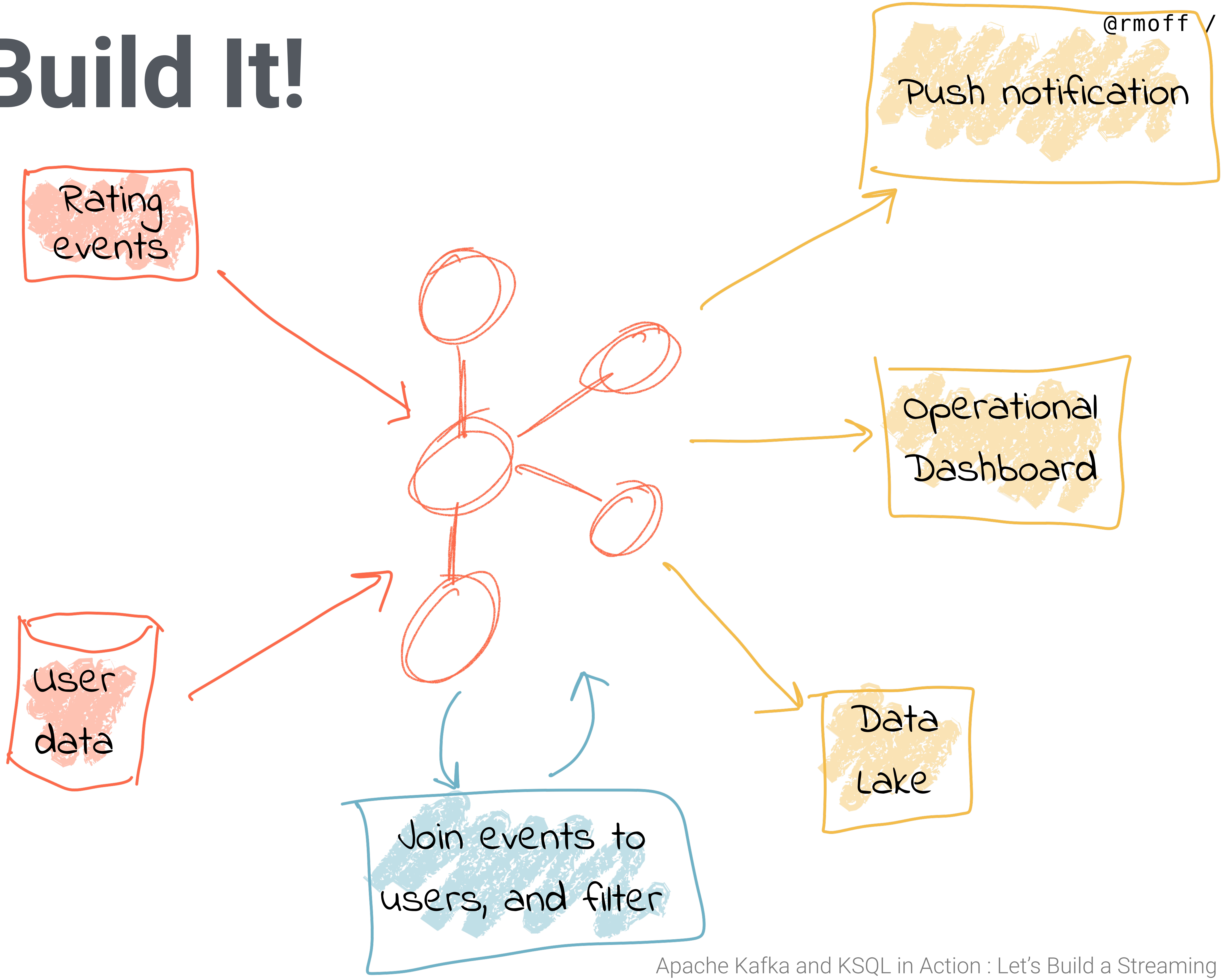
Transform Once, Use Many



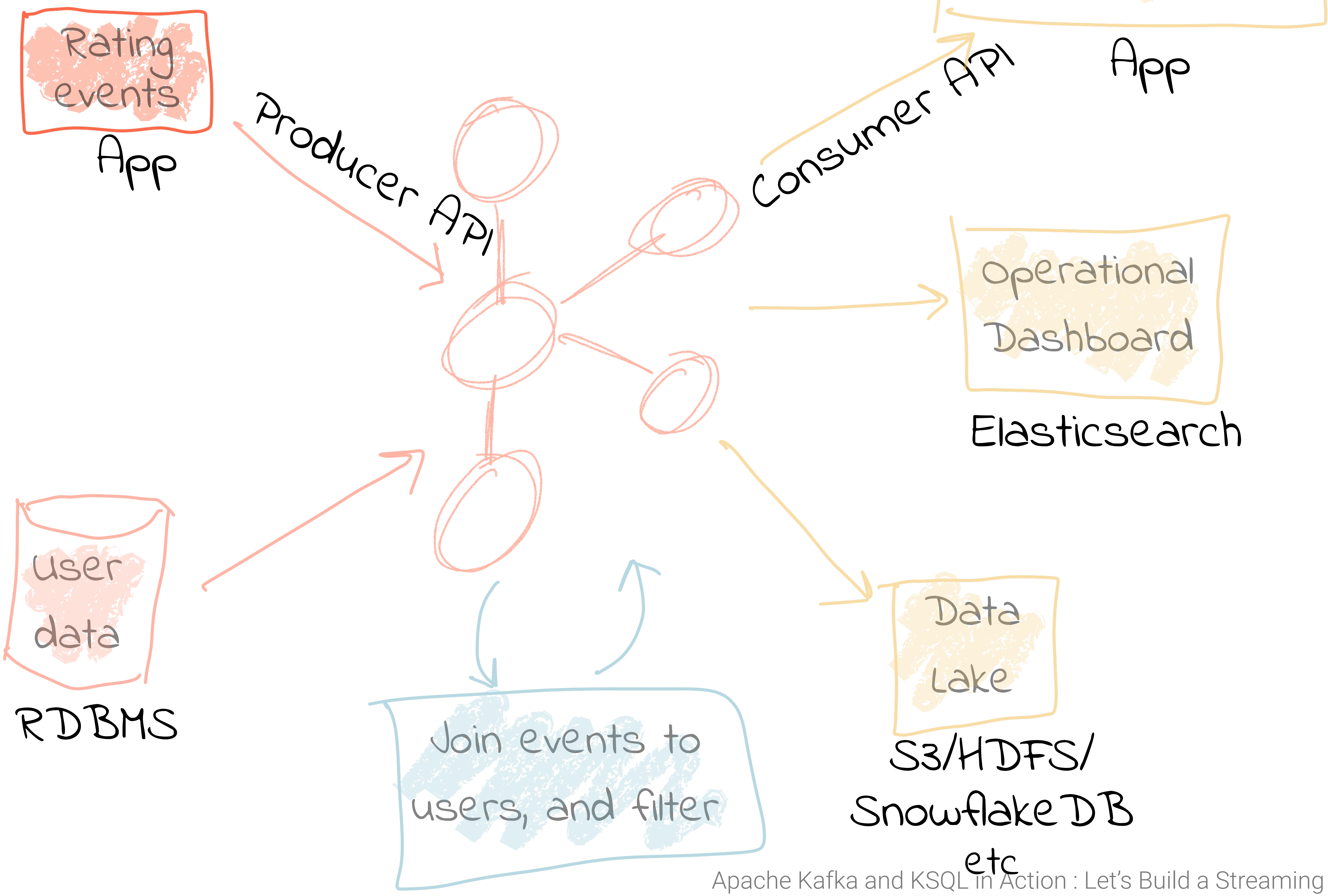
Transform Once, Use Many



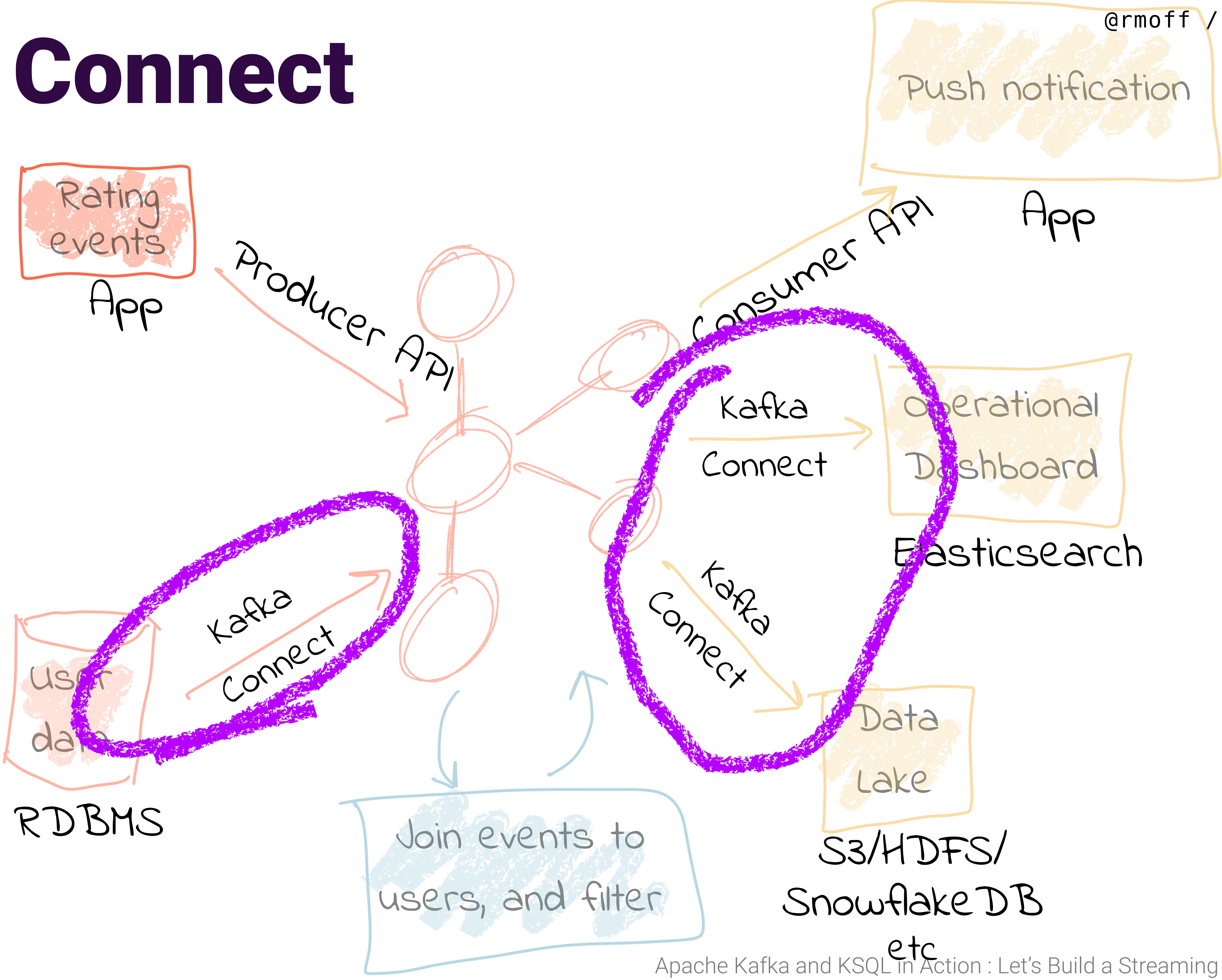
Let's Build It!



Let's Build It!



Kafka Connect



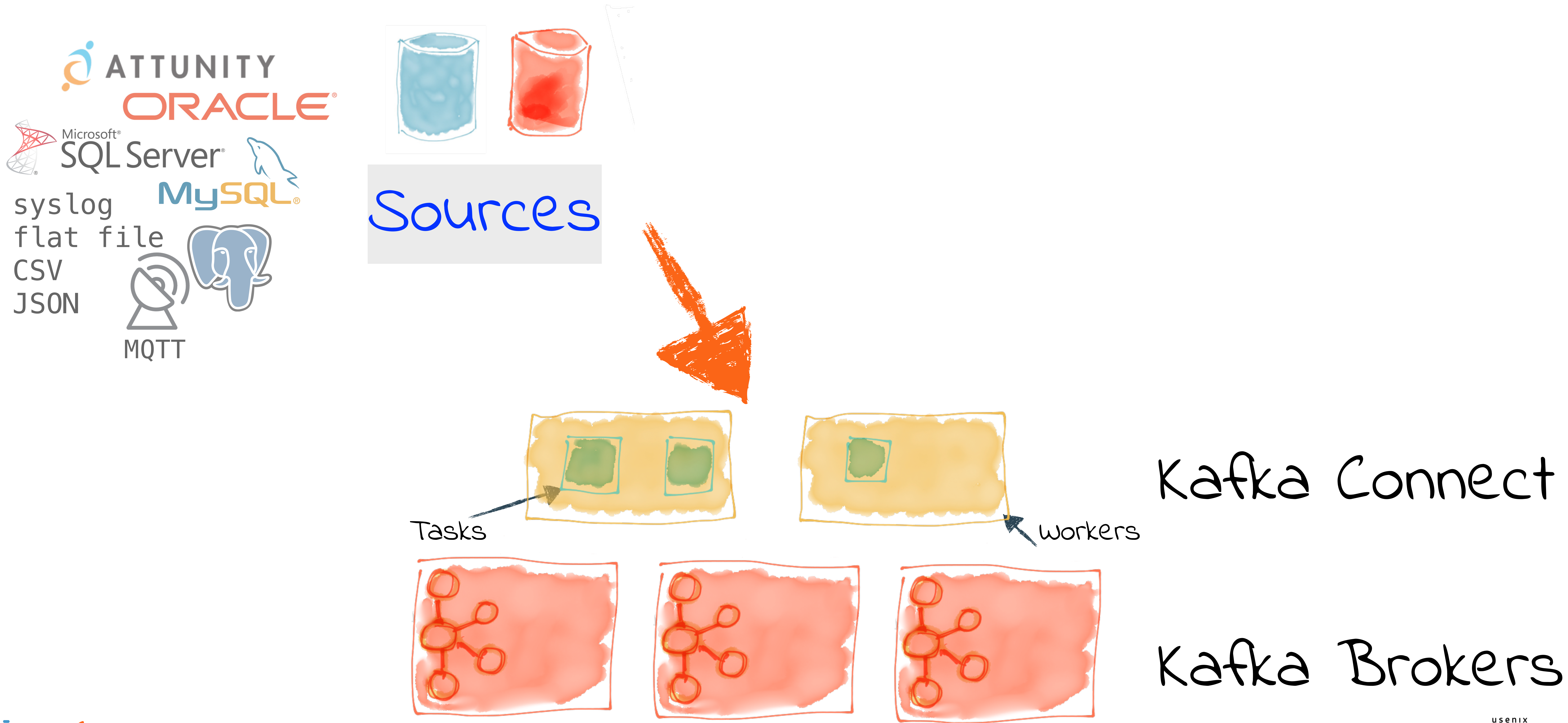
Kafka Connect

An API of Apache Kafka, providing reliable and scalable integration of Kafka with other systems – no coding required.

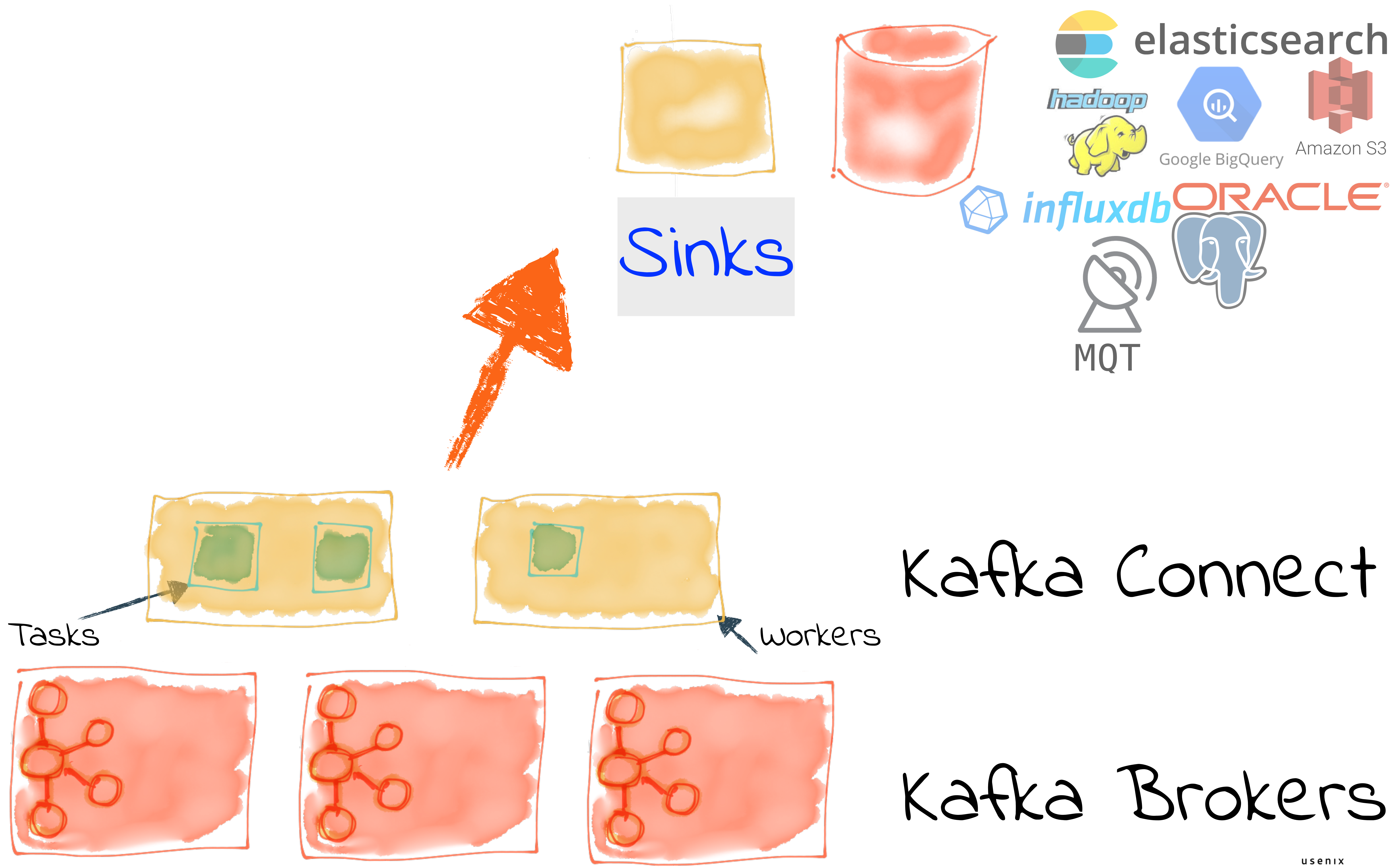
```
{
  "connector.class":
    "io.confluent.connect.jdbc.JdbcSourceConnector",
  "connection.url":
    "jdbc:mysql://localhost:3306/demo?user=rmoff&password=foo",
  "table.whitelist":
    "sales,orders,customers"
}
```

https://docs.confluent.io/current/connect/

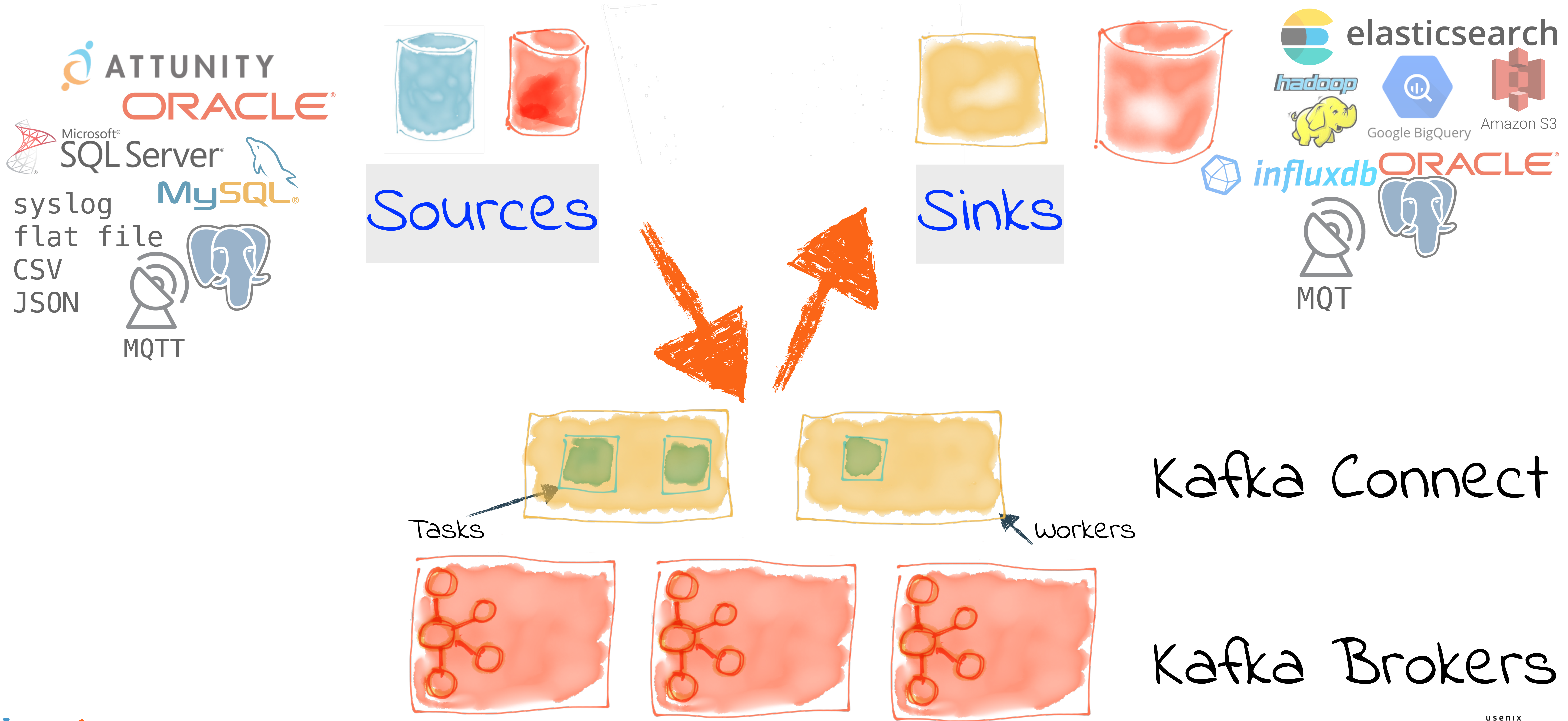
Streaming Integration with Kafka Connect



Streaming Integration with Kafka Connect

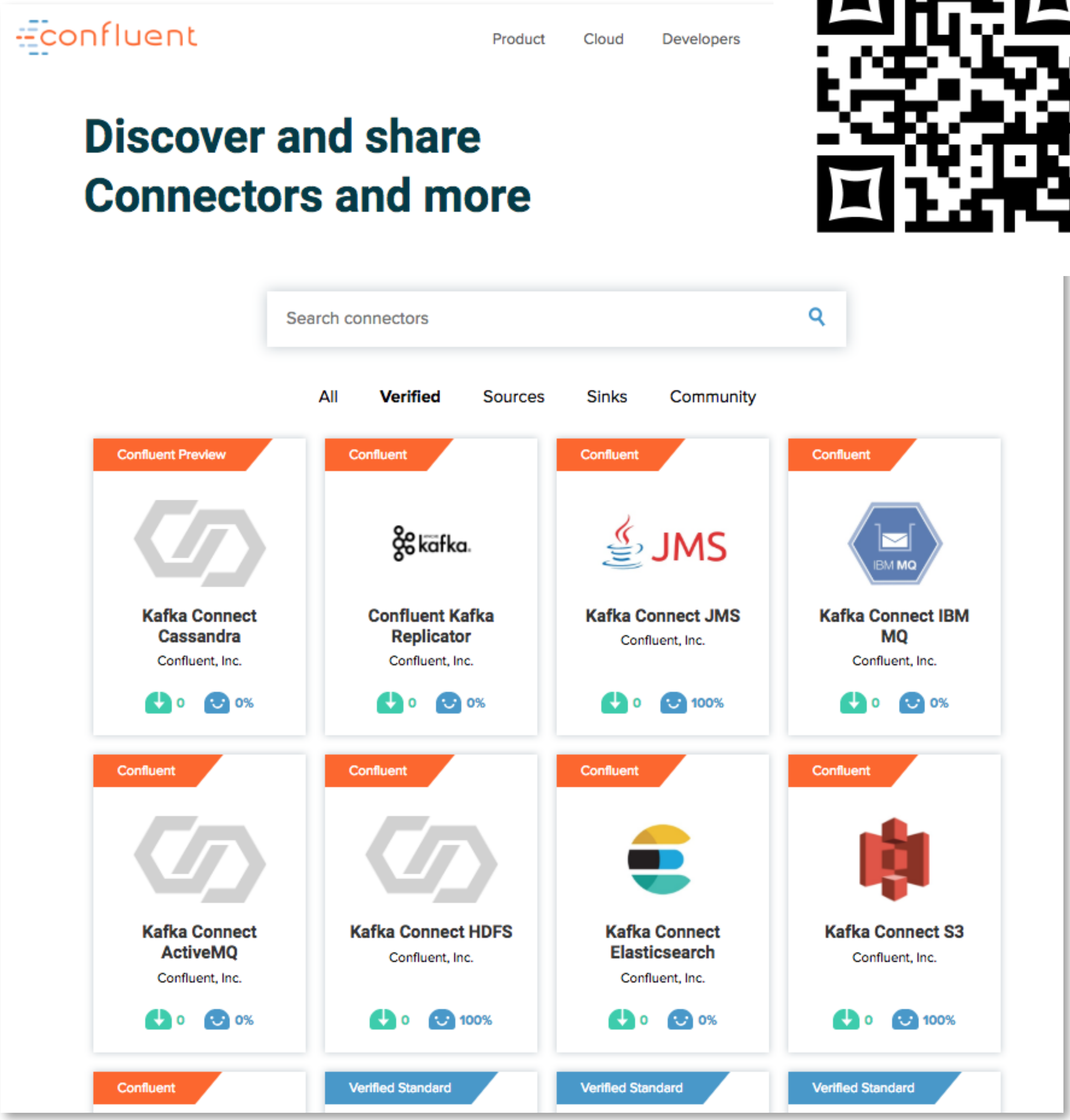


Streaming Integration with Kafka Connect



Confluent Hub

- One-stop place to discover and download :
  - Connectors
  - Transformations
  - Converters



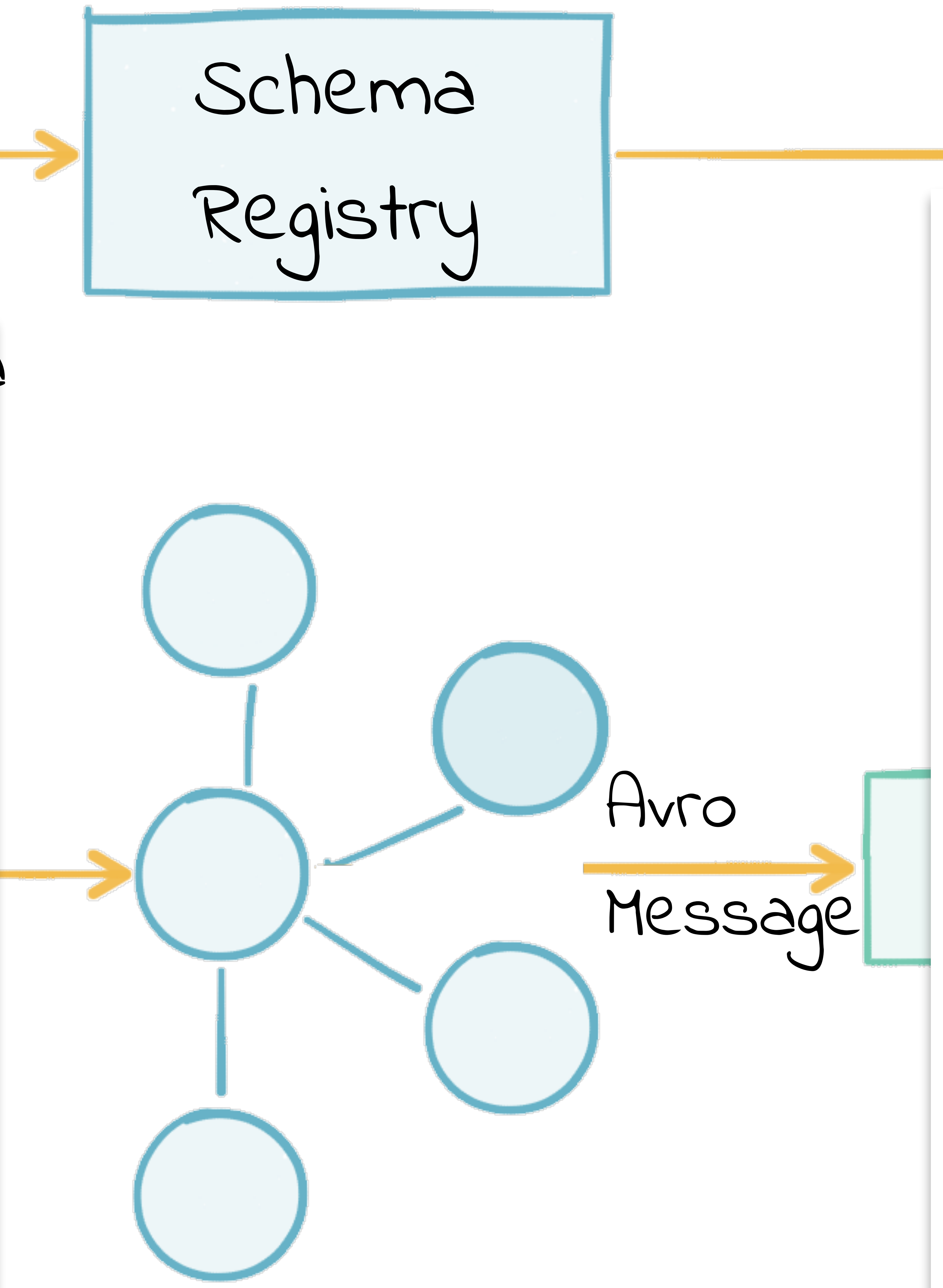
hub.confluent.io

Kafka Connect + Schema Registry = WIN

Kafka Connect + Schema Registry = WIN

Avro

```
mysql> describe rental;
+---------------+------------
| Field         | Type
+---------------+------------
| rental_id     | int(11)
| rental_date   | datetime
| inventory_id  | mediumint(8
| customer_id   | smallint(5)
| return_date   | datetime
| staff_id      | tinyint(3)
| last_update   | timestamp
+---------------+------------
```



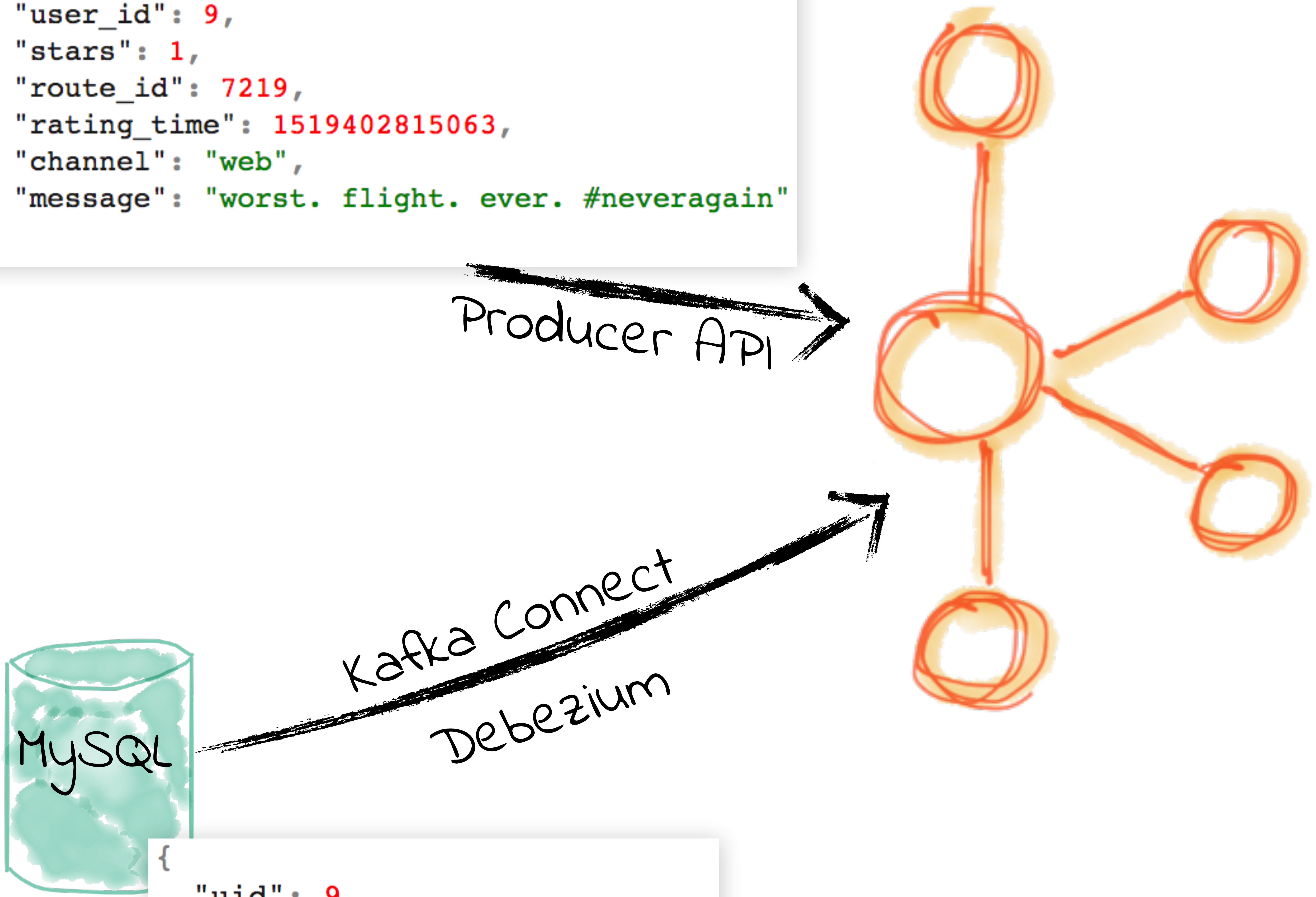

```
"mappings": {
  "type.name=kafka-connect": {
    "dynamic_templates": [⟺],
    "properties": {
      "customer_id": {
        "type": "integer"
      },
      "inventory_id": {
        "type": "long"
      },
      "last_update": {
        "type": "date"
      },
      "rental_date": {
        "type": "date"
      },
      "rental_id": {
        "type": "integer"
      },
      "return_date": {
        "type": "date"
      },
```

Kafka → Elasticsearch

```
curl -X "POST" "http://kafka-connect-cp:18083/connectors/" \
     -H "Content-Type: application/json" \
     -d '{
    "name": "es_sink_lisa18",
    "config": {
    "connector.class":
"io.confluent.connect.elasticsearch.ElasticsearchSinkConnector",
    "key.converter": "org.apache.kafka.connect.storage.StringConverter",
    "value.converter": "org.apache.kafka.connect.json.JsonConverter",
    "value.converter.schemas.enable": false,
    "topics": "lisa18",
    "key.ignore": "true",
    "schema.ignore": "true",
    "type.name": "type.name=kafkaconnect",
    "connection.url": "http://elasticsearch:9200"
  }
}'
```

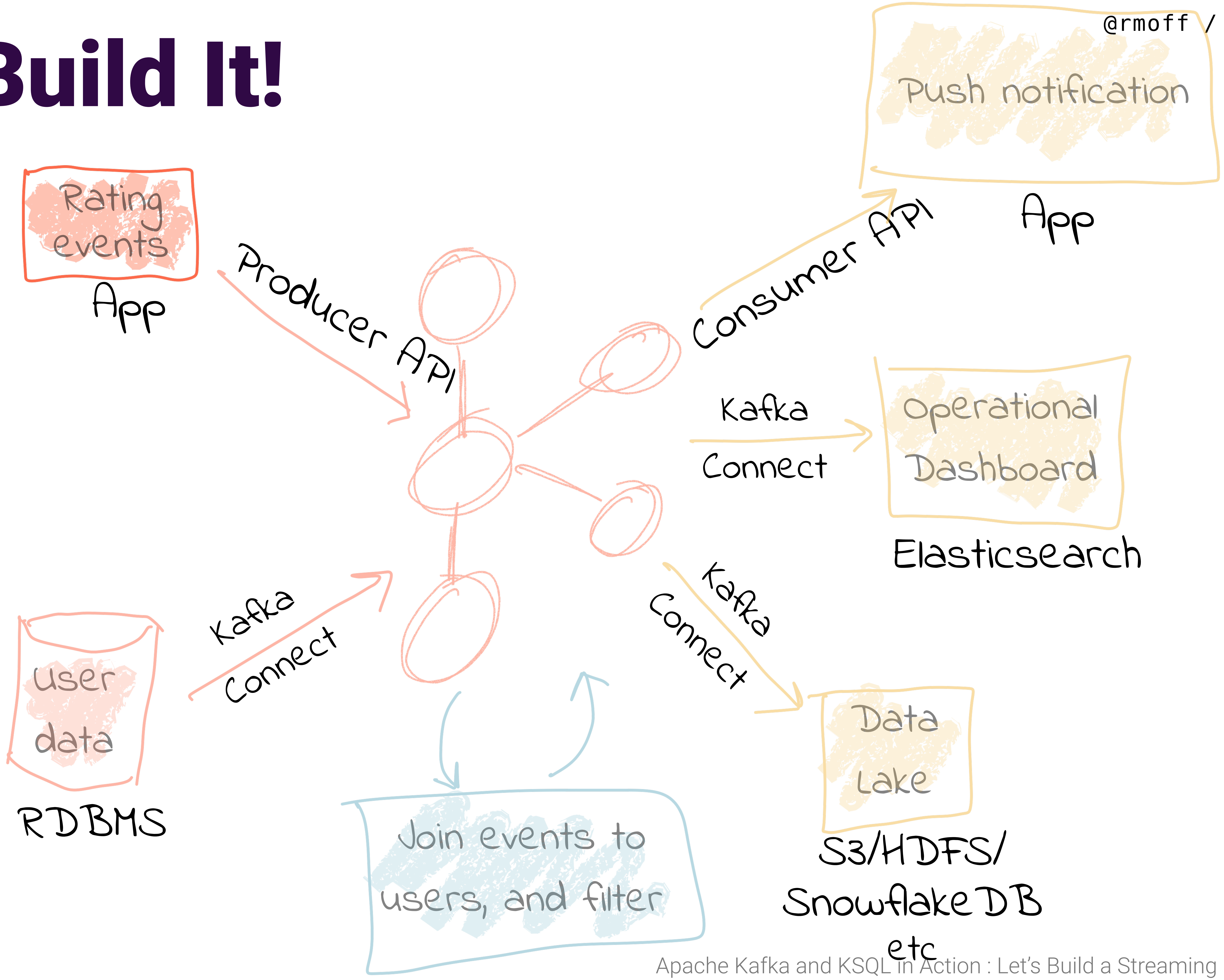
Demo Time!


```
{
  "rating_id": 2133,
  "user_id": 9,
  "stars": 1,
  "route_id": 7219,
  "rating_time": 1519402815063,
  "channel": "web",
  "message": "worst. flight. ever. #neveragain"
}
```

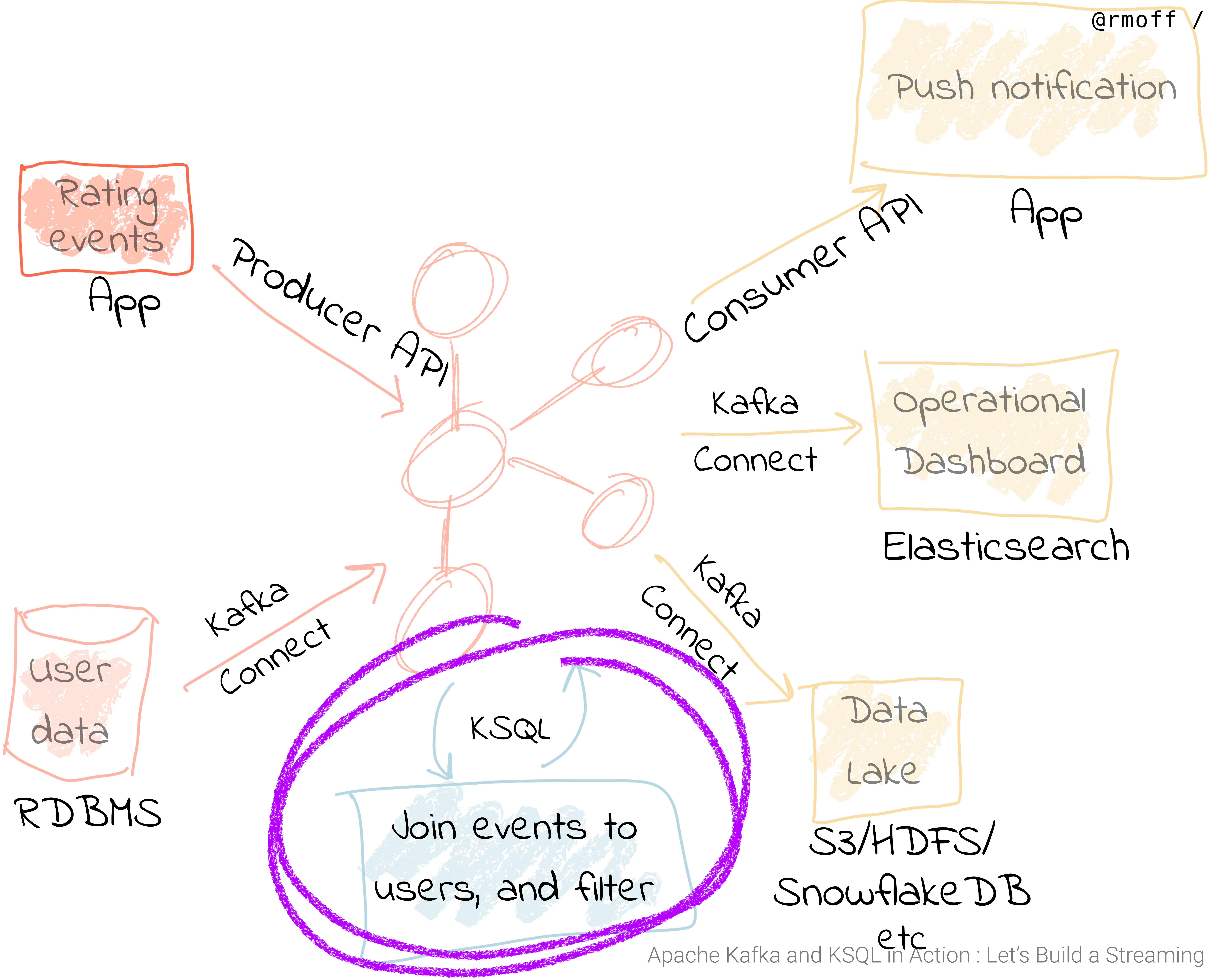




```
{
  "uid": 9,
  "name": "Neha",
  "locale": "en_US",
  "address_city": "Palo Alto",
  "elite": "P"
}
```

Let's Build It!



KSQL

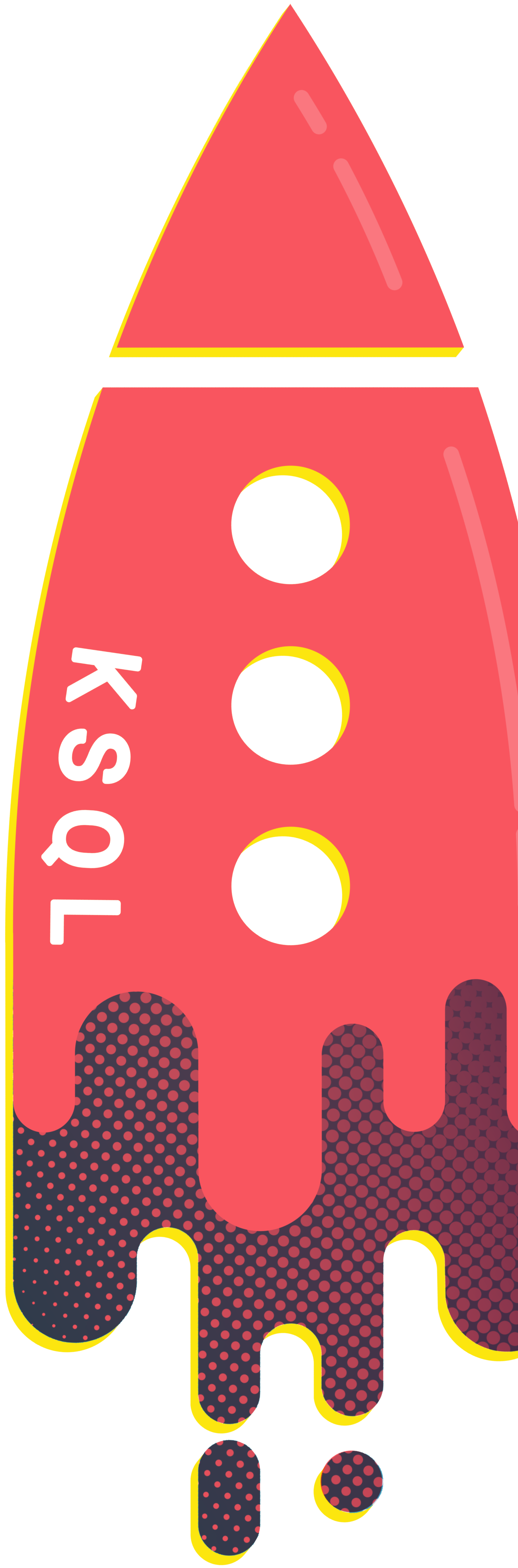

Push notification
App
Rating events
App
Producer API
Consumer API
Kafka Connect
Operational Dashboard
Elasticsearch
User data
RDBMS
Kafka Connect
KSQL
Join events to users, and filter
Kafka Connect
Data Lake
S3/HDFS/
SnowflakeDB
etc

KSQL

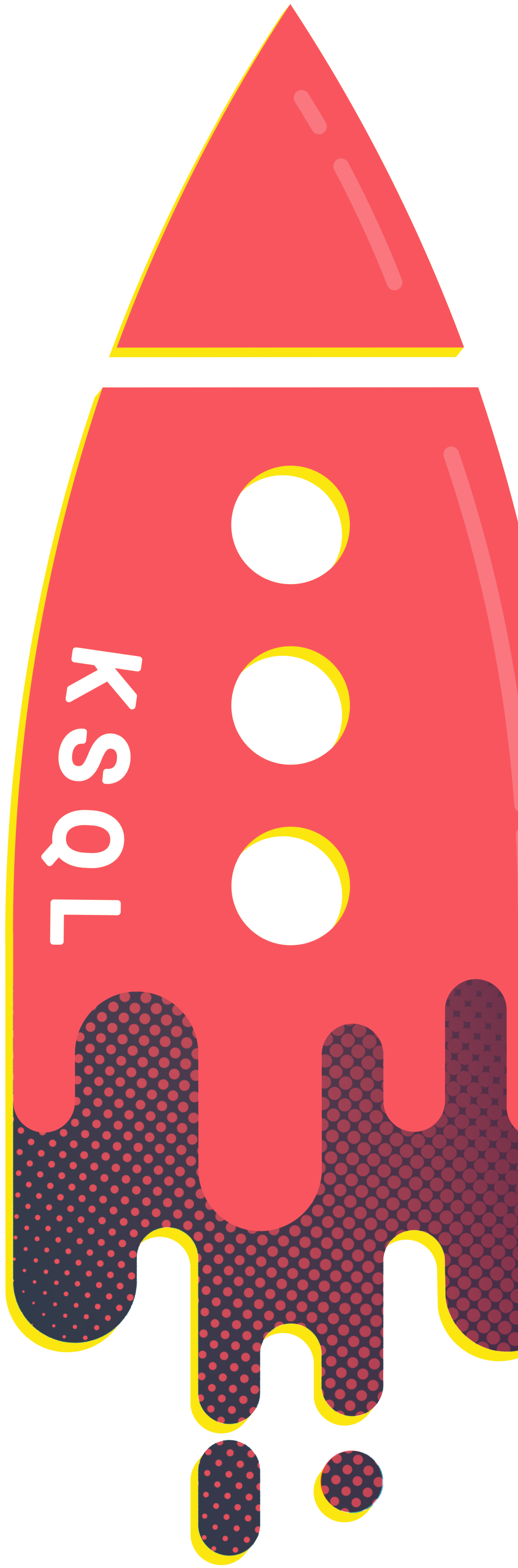
is a

Declarative

Stream Processing Language



KSQL
is the
Streaming
SQL Engine
for
Apache Kafka



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 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';
```

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 Data Transformation

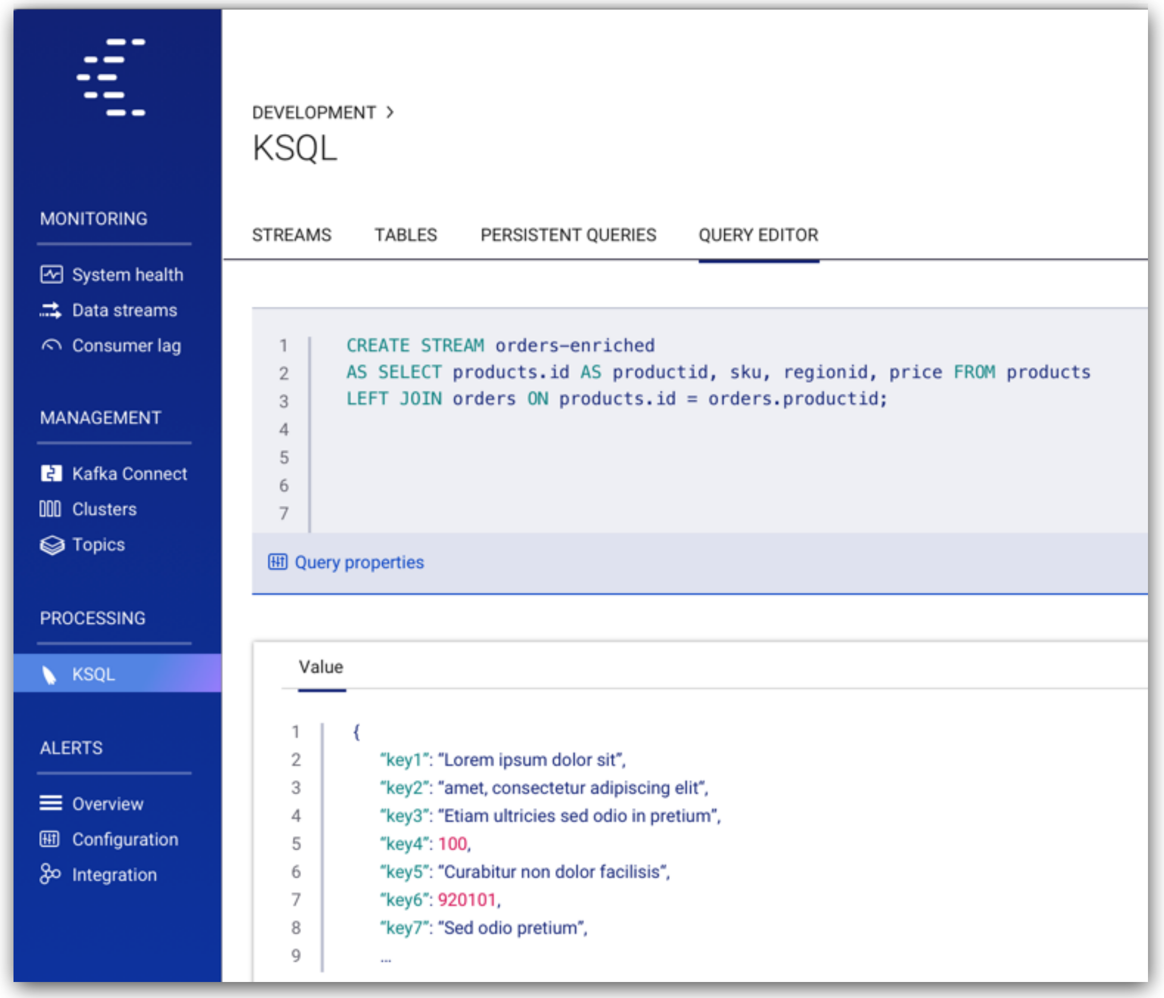
Make simple derivations of existing topics from the command line

```
CREATE STREAM pageviews
  WITH (PARTITIONS=4,
        VALUE_FORMAT='AVRO') AS
  SELECT * FROM pageviews_json;
```

KSQL in Development and Production

Interactive KSQL
for development and testing

Headless KSQL
for Production



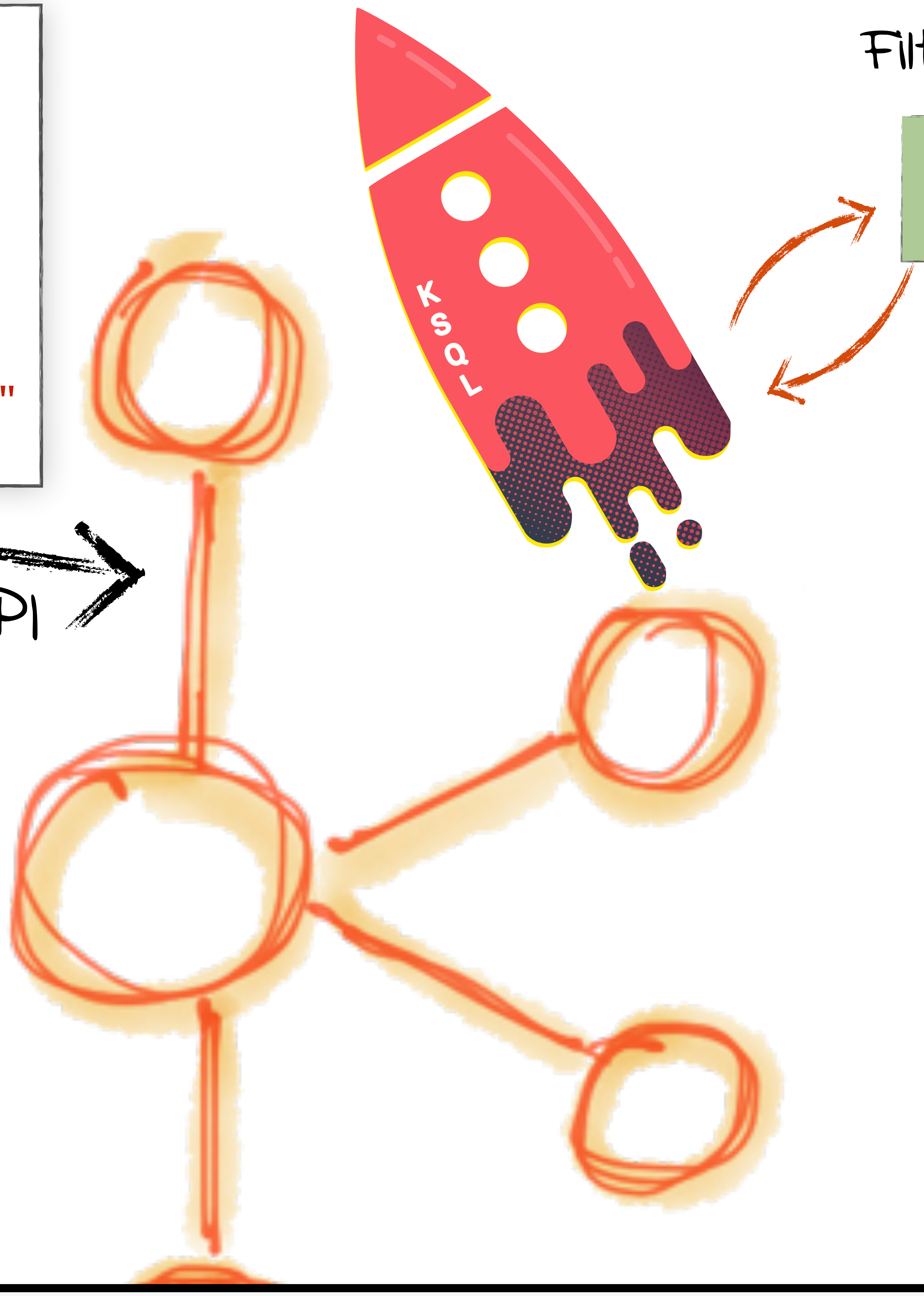
"Hmm, let me try out this idea..."

Demo Time!


{
"rating_id": 2133,
"user_id": 9,
"stars": 1,
"route_id": 7219,
"rating_time": 1519402815063,
"channel": "web",
"message": "worst. flight. ever. #neveragain"
}
Producer API
Kafka Connect
kibana
80
Fire 01 (Red) - Count
60
Robin's iPhone - Count
46
Fire 02 (Yellow) - Count
31
Study light - Count
26
Robin's Kindle - Count
Elasticsearch
MySQL
Kafka Connect
Debezium
{
"uid": 9,
"name": "Neha",
"locale": "en_US",
"address_city": "Palo Alto",
"elite": "P"
}
KSQL

```
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304105213,
 "channel": "web",
 "message": "worst. flight. ever. #neveragain"
}
```


Producer API



Filter all ratings where STARS<3

POOR_RATINGS

```
CREATE STREAM POOR_RATINGS AS
SELECT * FROM ratings WHERE STARS <3
```


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

The Stream Table 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

Read more: https://cnfl.io/stream-table-duality

The truth is the log.

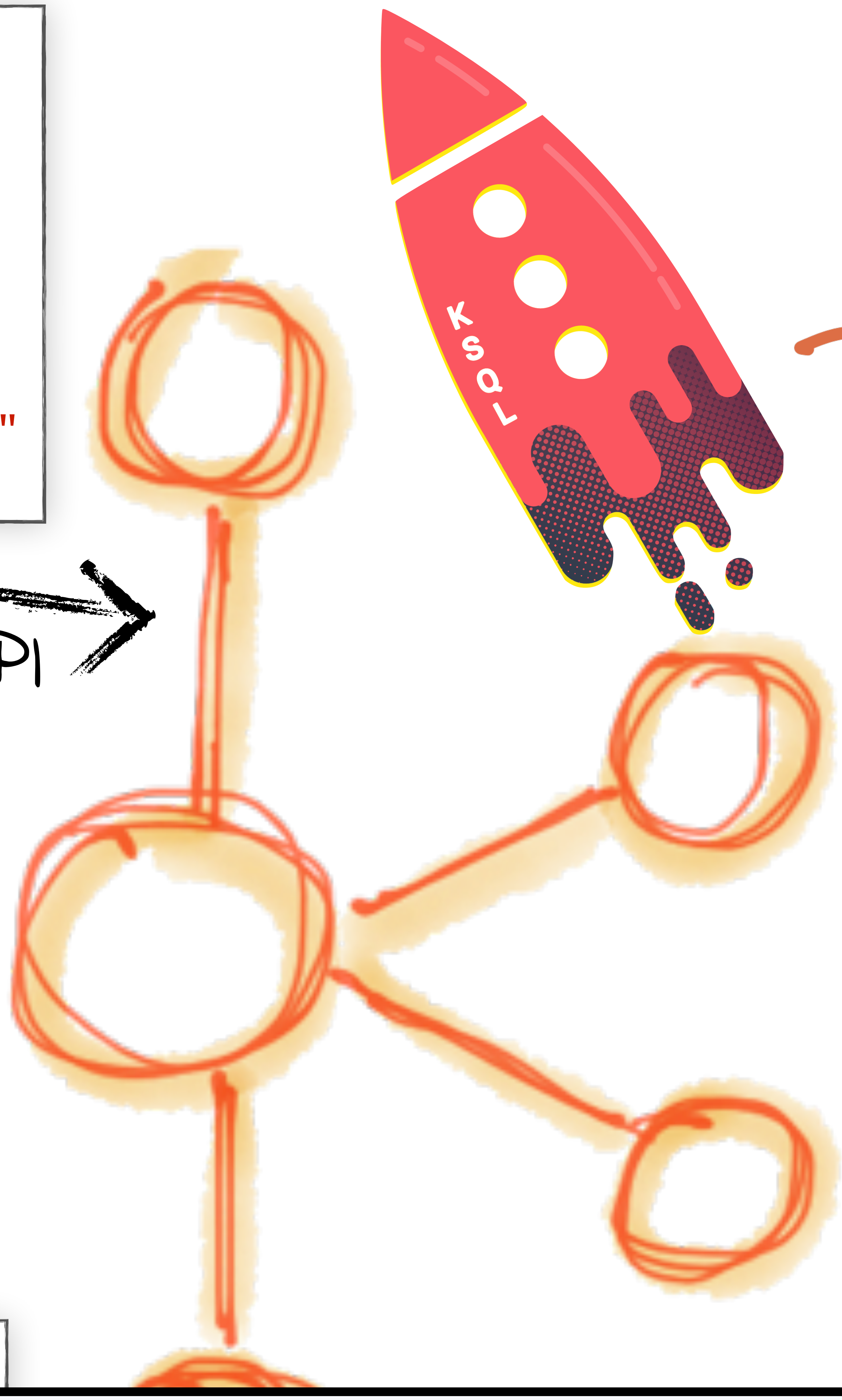
The database is a cache of a subset of the log.

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

Photo by Bobby Burch on Unsplash

```
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304105213,
 "channel": "web",
 "message": "worst. flight. ever. #neveragain"
}
```

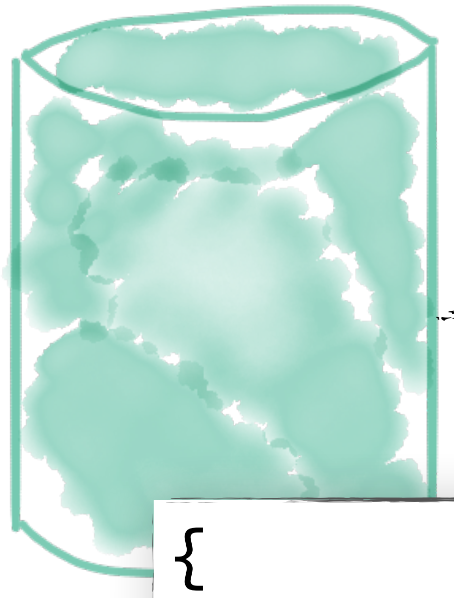

Producer API



Join each rating to customer data

RATINGS_WITH_CUSTOMER_DATA

Kafka Connect

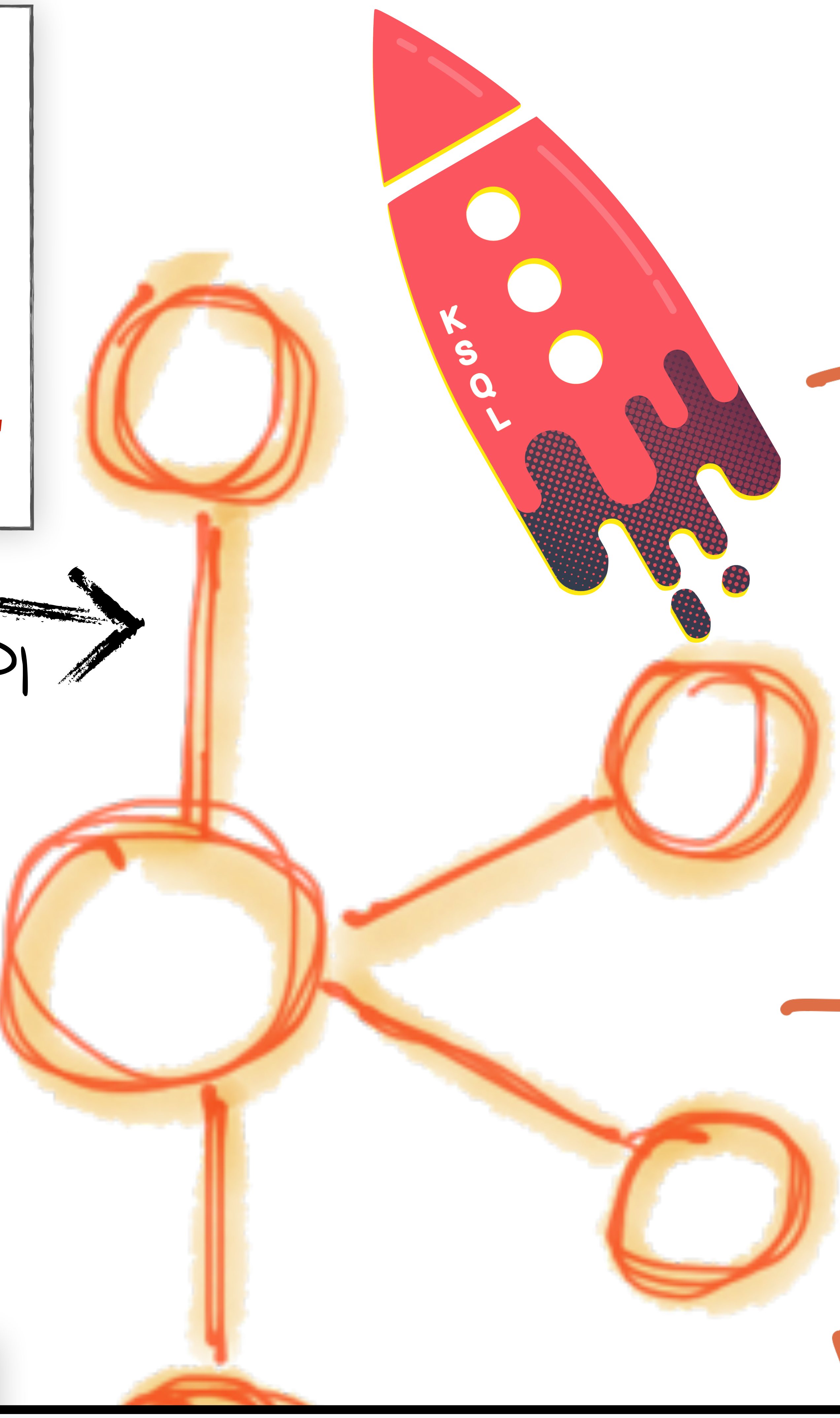



```
{
  "id": 3,
  "first_name
  "last_name"
  "email": "m
  "gender": "
  "club_statu
  "comments":
}
```


```sql
CREATE STREAM RATINGS_WITH_CUSTOMER_DATA AS
SELECT * FROM RATINGS LEFT JOIN CUSTOMERS
ON R.ID=C.ID;
```

```
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304105213,
 "channel": "web",
 "message": "worst. flight. ever. #neveragain"
}
```


Producer API

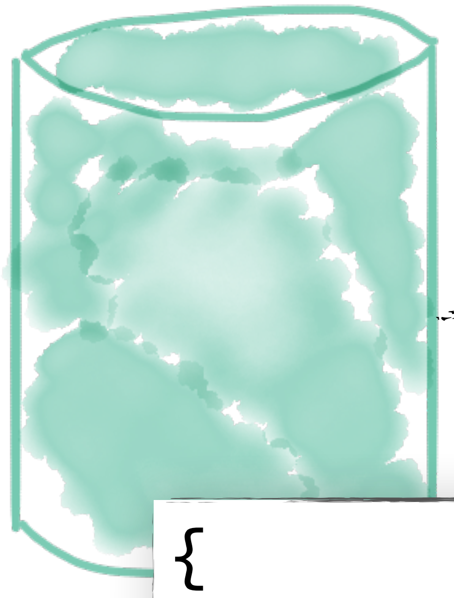


Join each rating to customer data

RATINGS_WITH_CUSTOMER_DATA

Filter for just PLATINUM customers

UNHAPPY_PLATINUM_CUSTOMERS

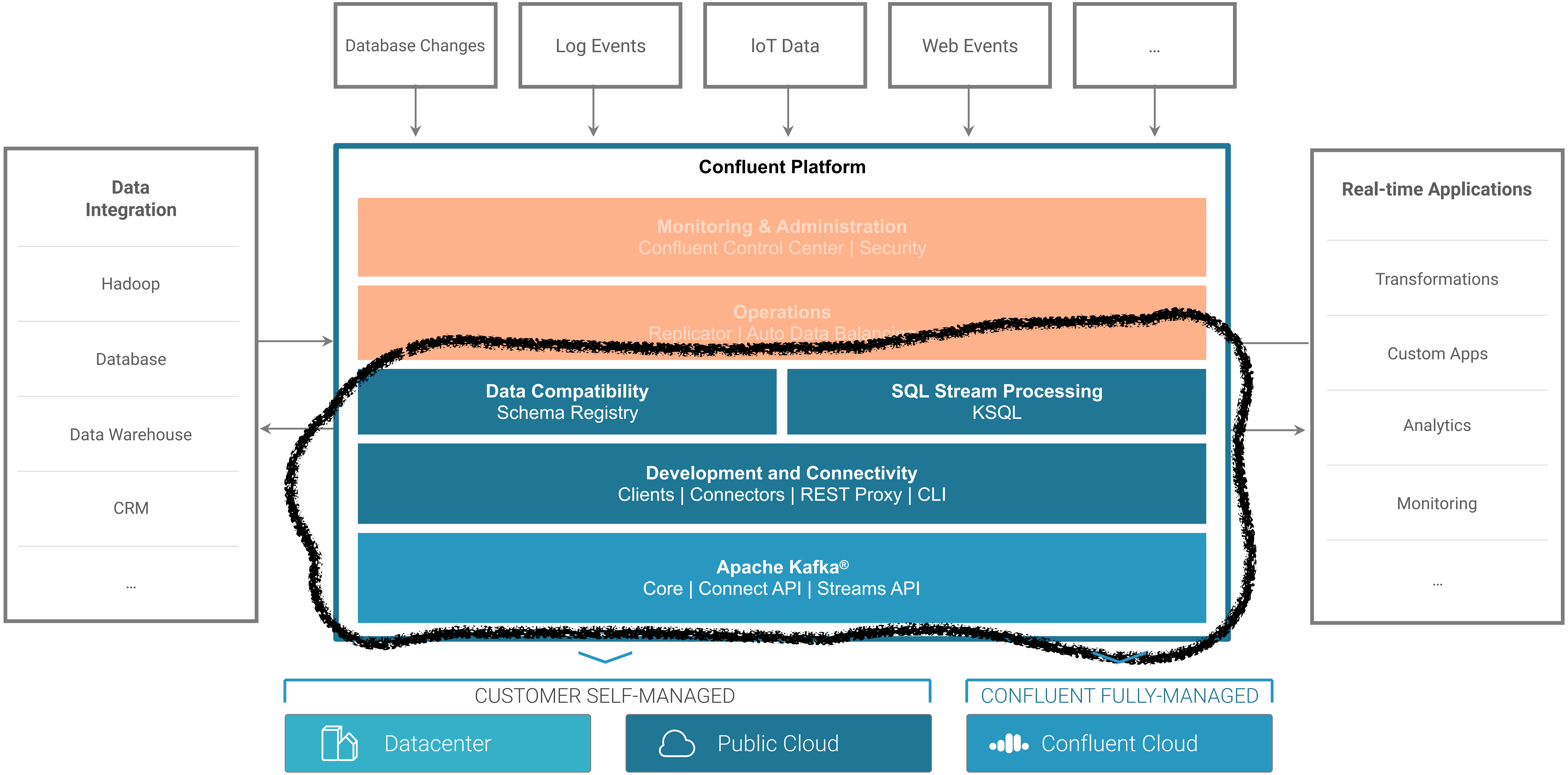


Kafka Connect


```
{
  "id": 3,
  "first_name
  "last_name"
  "email": "m
  "gender": "
  "club_statu
  "comments":
}
```


```
CREATE STREAM UNHAPPY_PLATINUM_CUSTOMERS AS
SELECT * FROM RATINGS_WITH_CUSTOMER_DATA
WHERE  STARS < 3
```

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

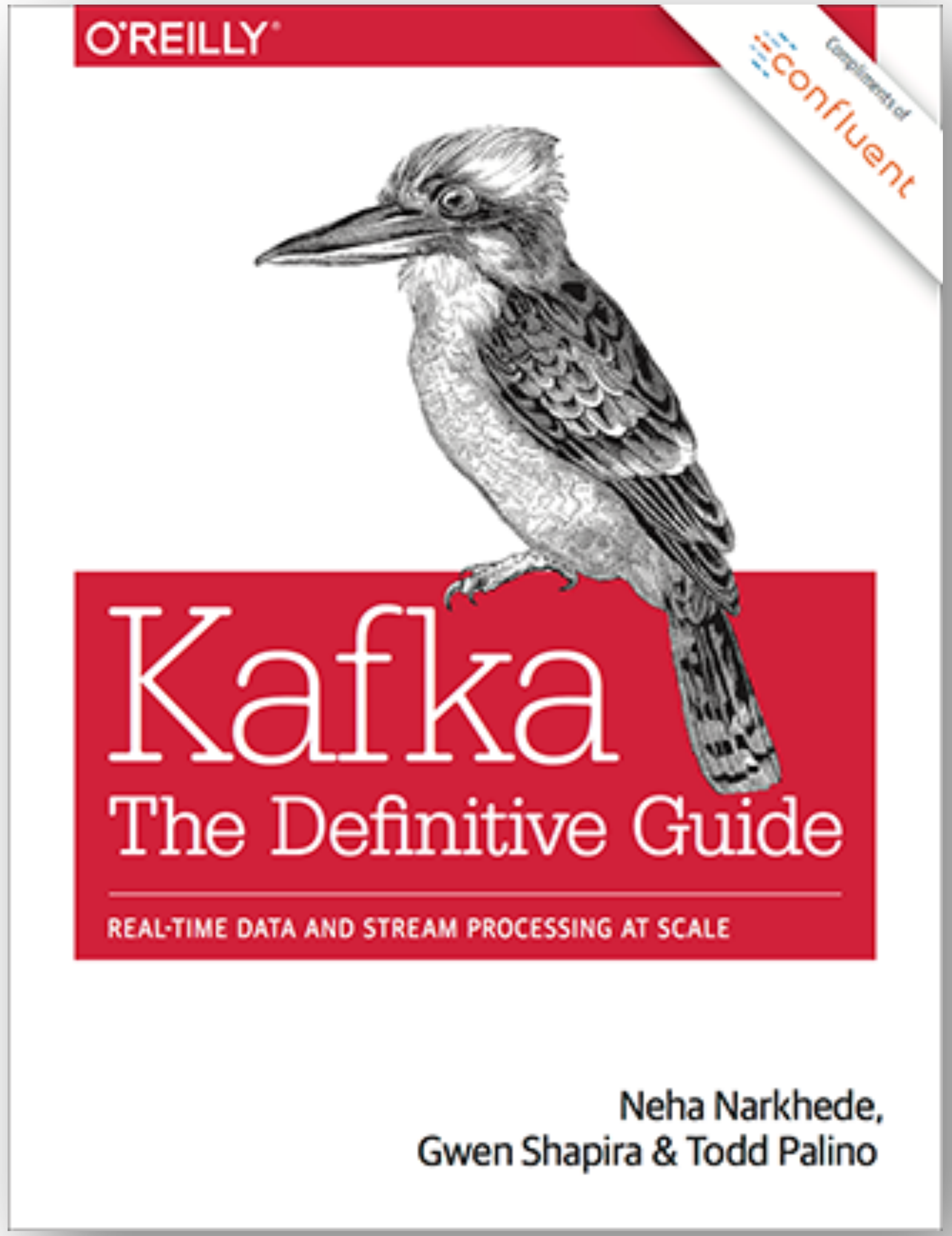


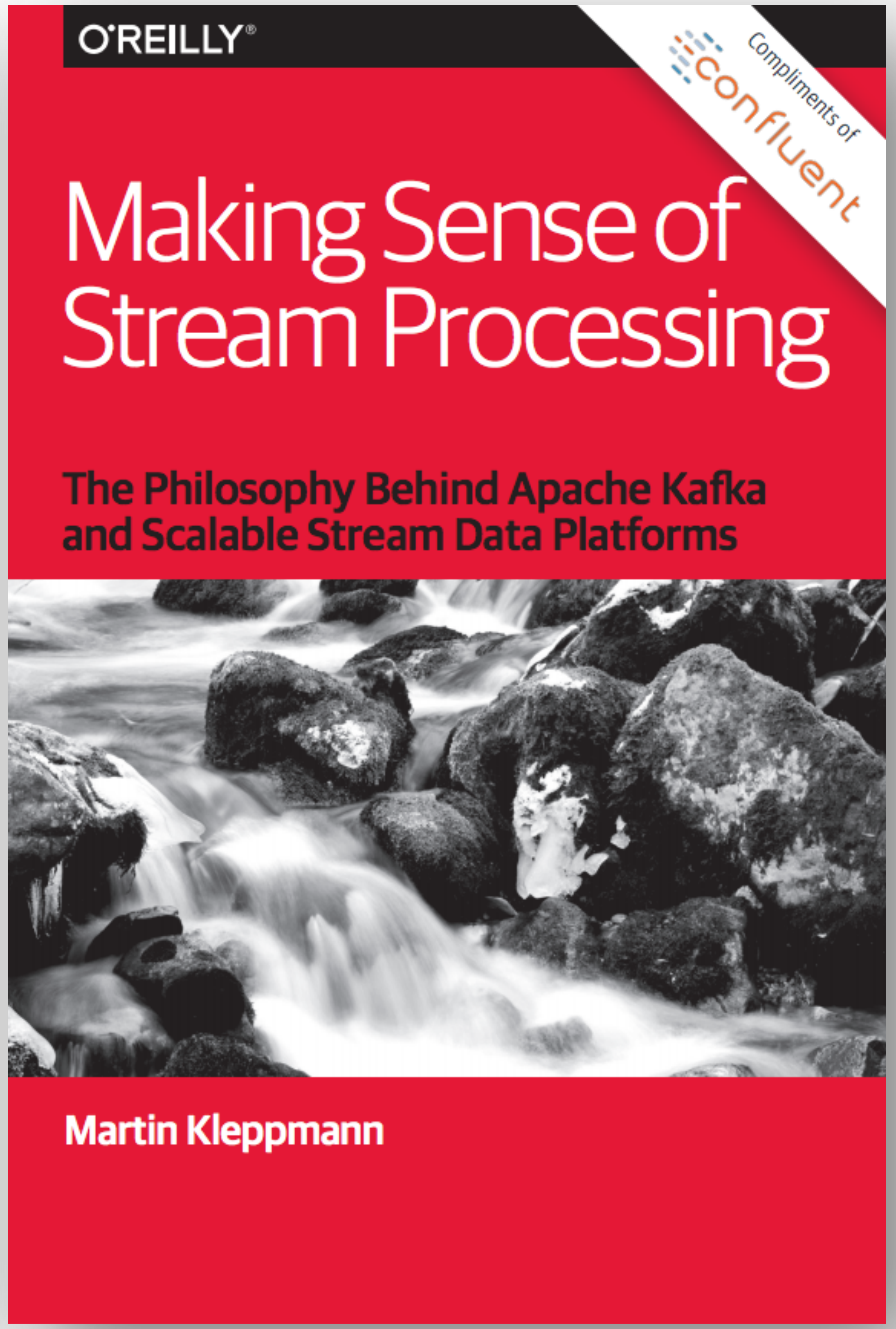
If you remember one thing… (or three)

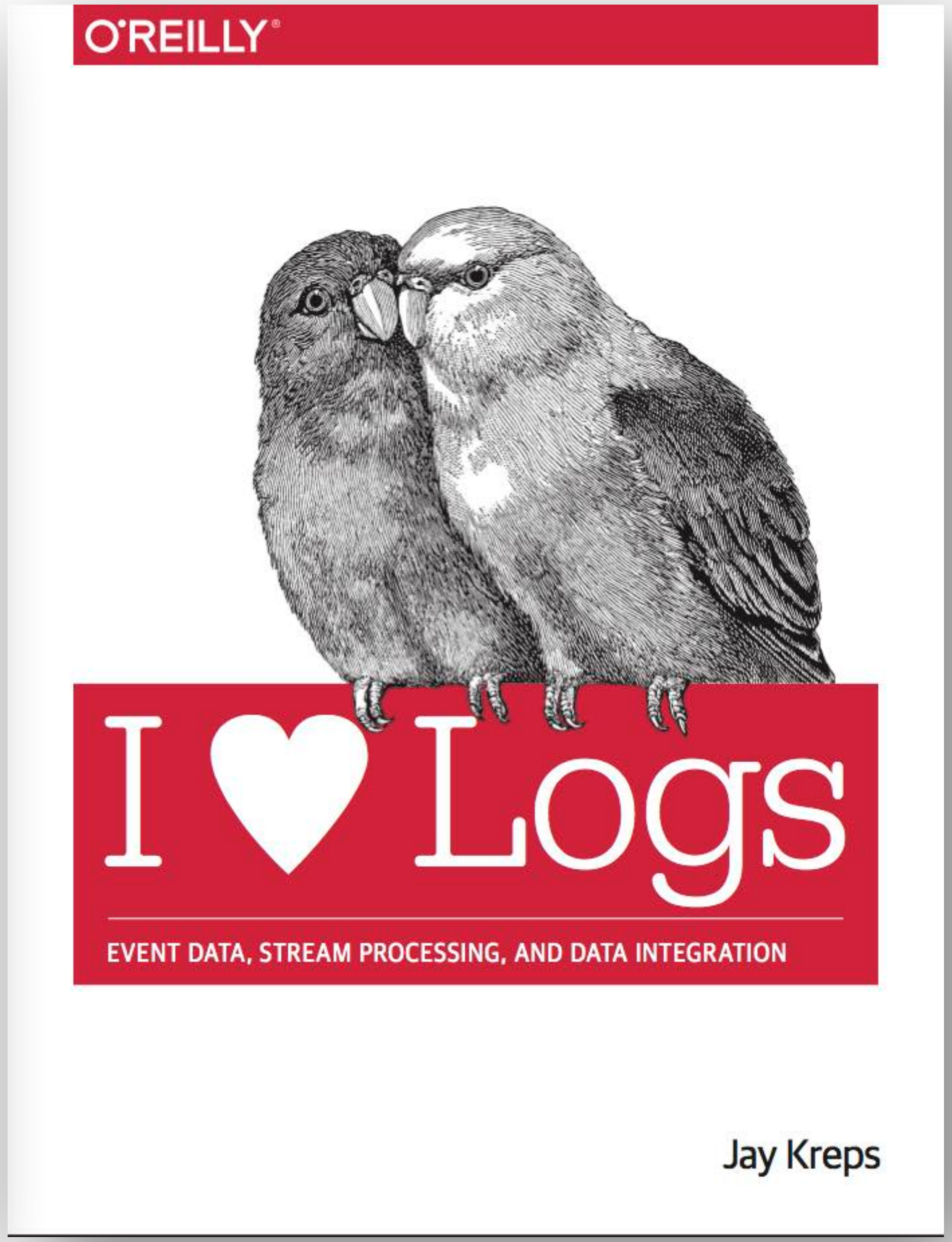
- **Kafka Connect**
  - *Integration* between Kafka and other data stores
- **Kafka**
  - Provides *stream processing* natively
- **KSQL**
  - Build stream processing apps with *just SQL*

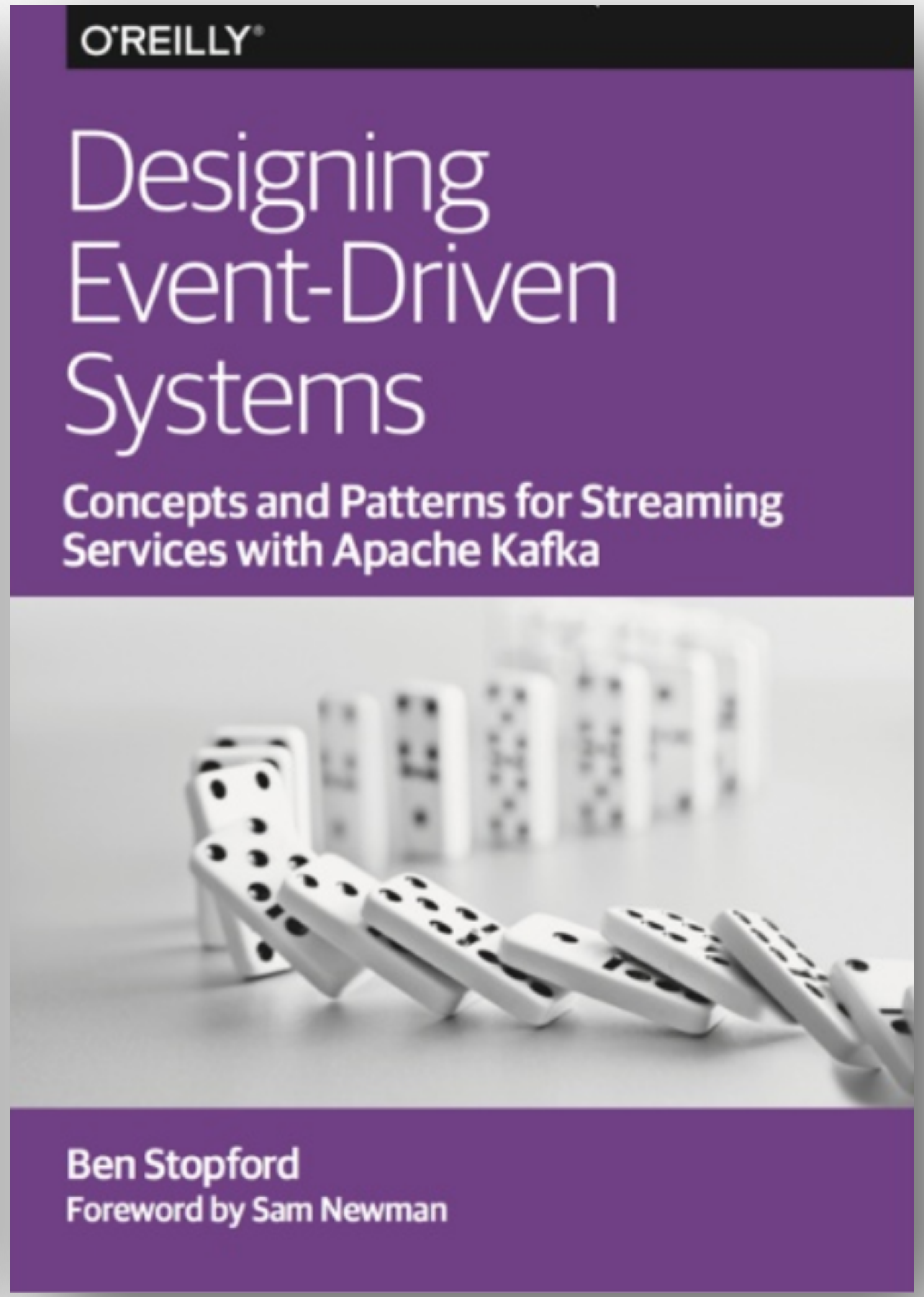
Free Books!

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









Try it out!

https://cnfl.io/kafka-ksql-elastic

https://www.confluent.io/ksql

http://cnfl.io/slack

@rmoff

robin@confluent.io

Useful links

- Embrace the Anarchy : Apache Kafka's Role in Modern Data Architectures Recording & Slides
- Look Ma, no Code! Building Streaming Data Pipelines with Apache Kafka and KSQL
- Steps to Building a Streaming ETL Pipeline with Apache Kafka and KSQL Recording & Slides
- https://www.confluent.io/blog/ksql-in-action-real-time-streaming-etl-from-oracle-transactional-data
- https://github.com/confluentinc/ksql/

#EOF