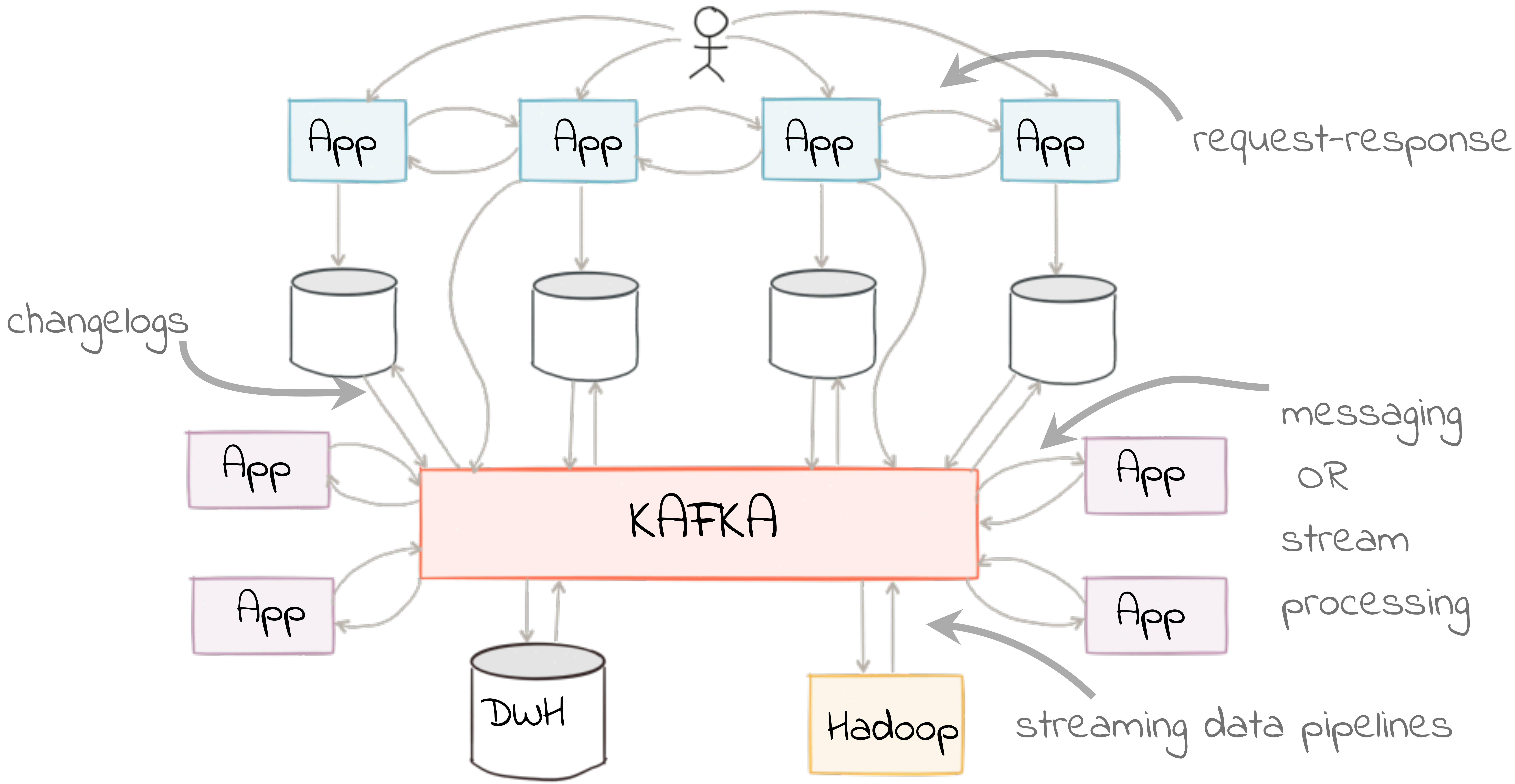
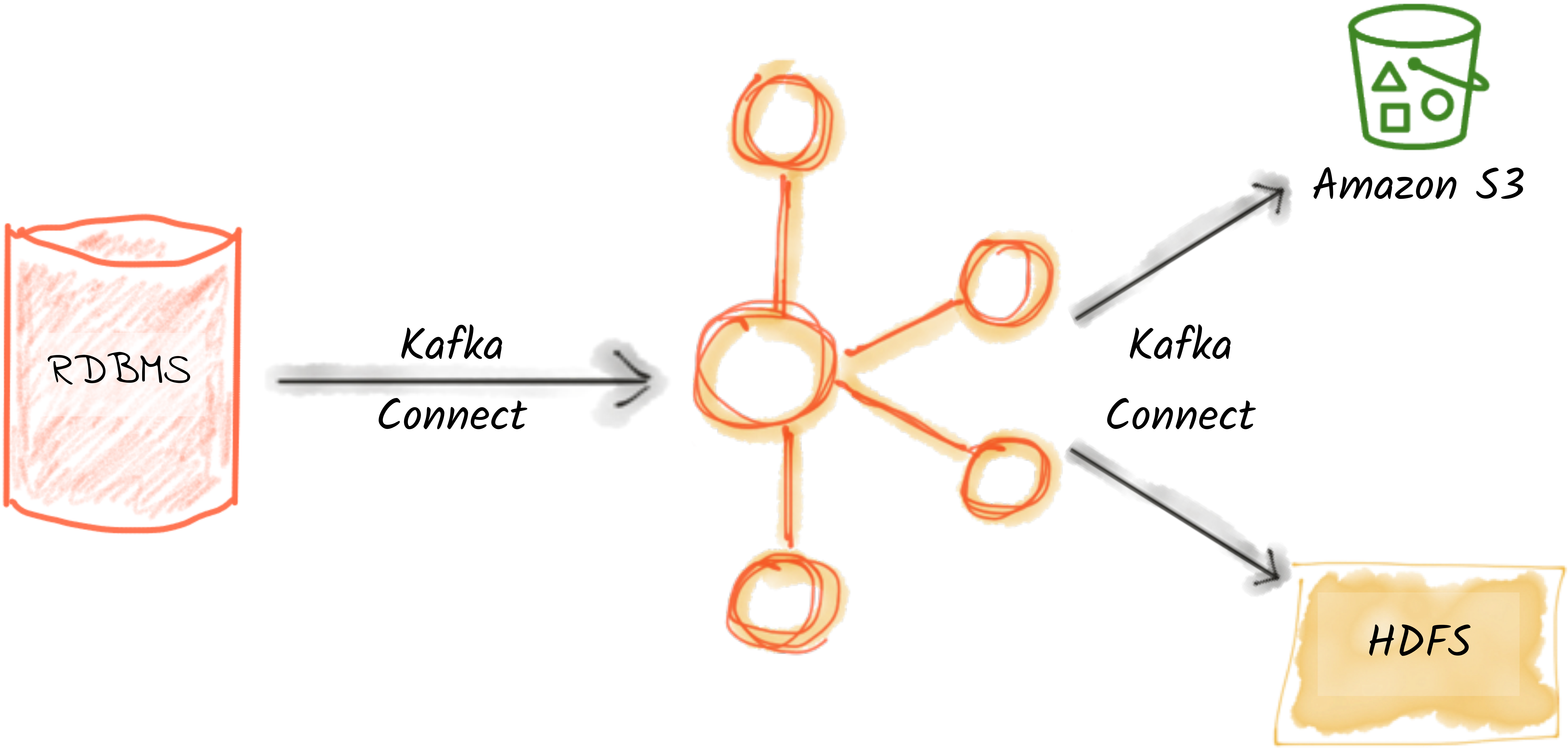
Apache Kafka and ksqlDB in Action: Let's Build a Streaming Data Pipeline!

@rmoff #KafkaMeetup

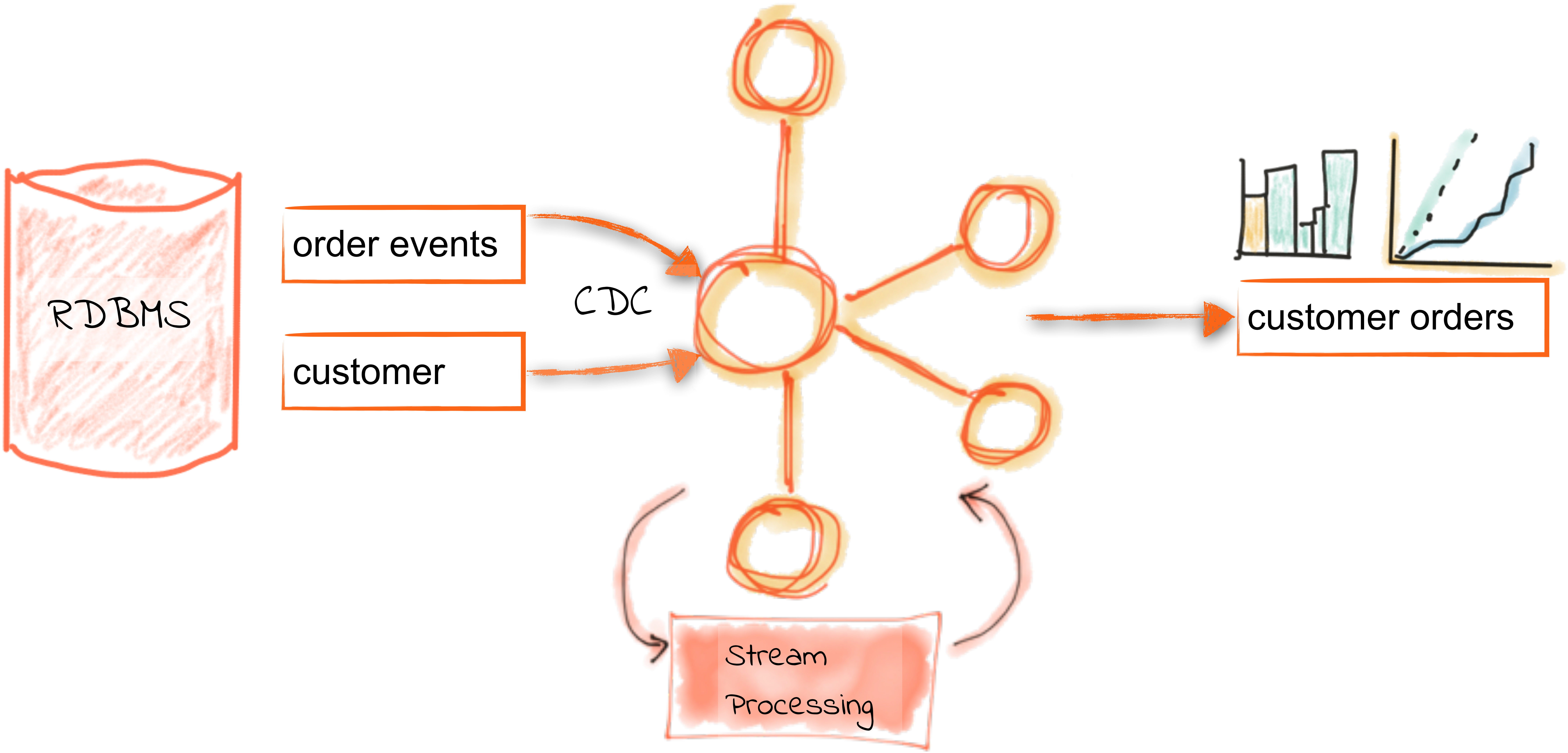
Kafka is an Event Streaming Platform



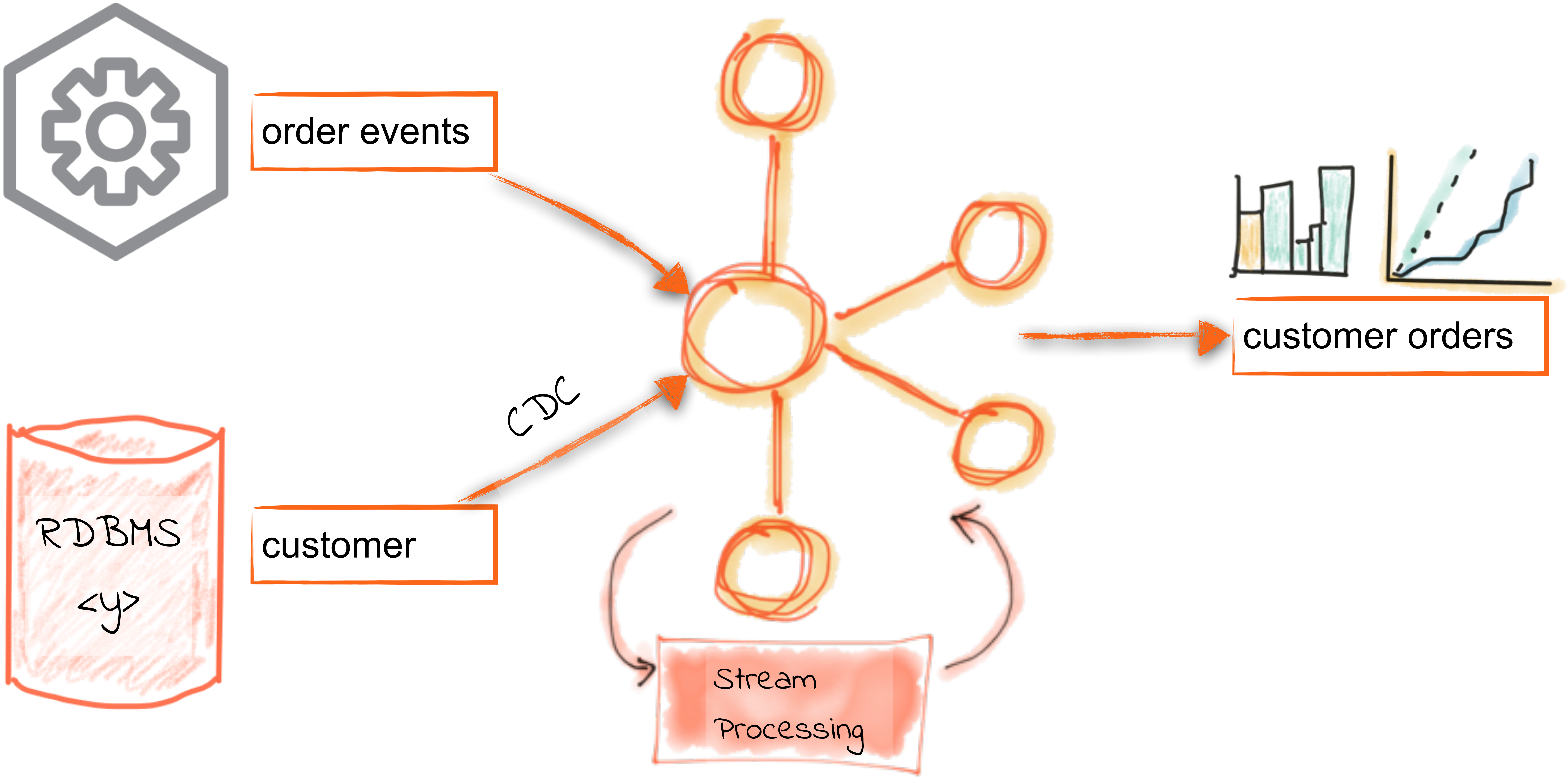
Database Offload



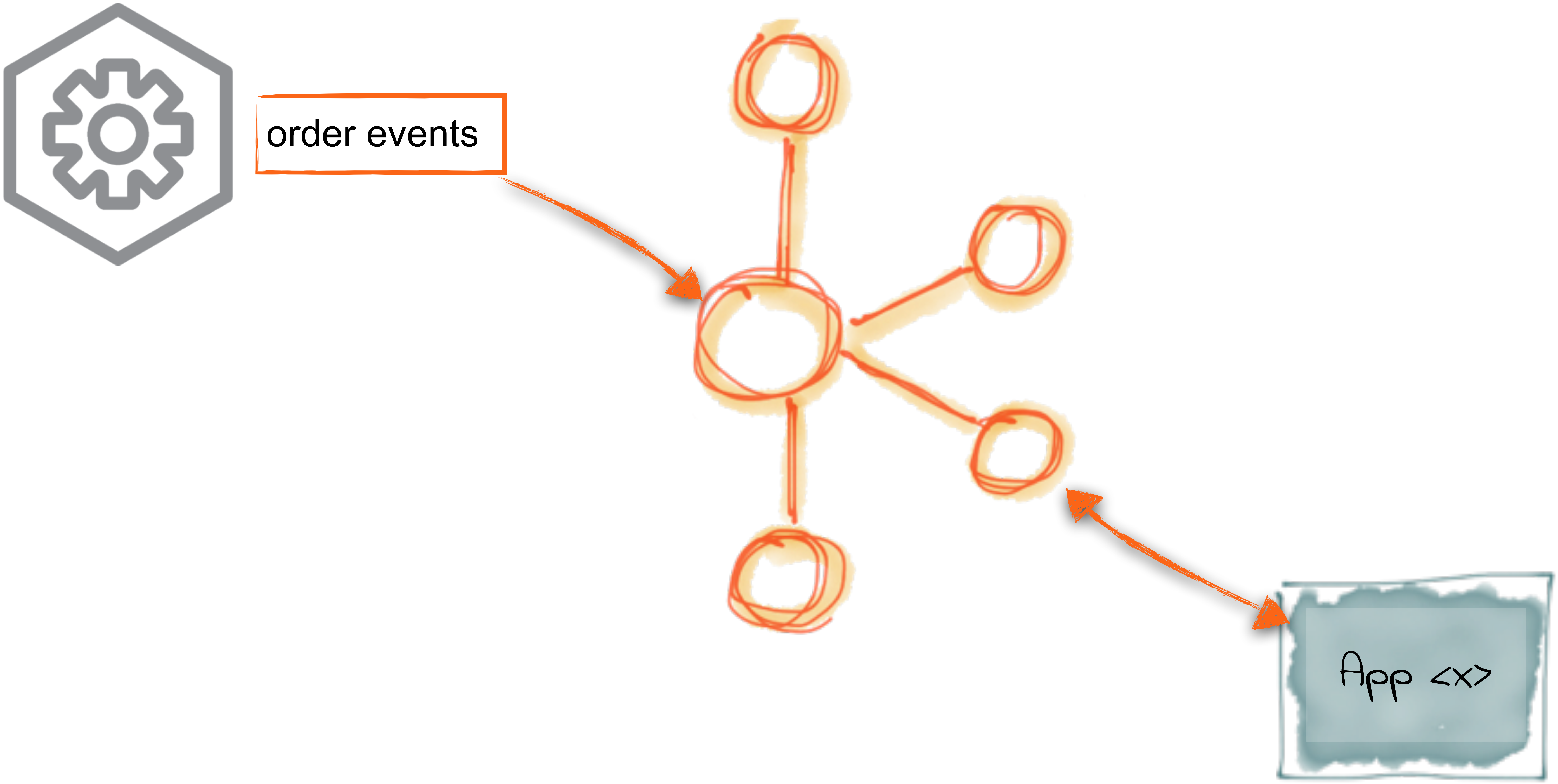
Stream Processing with Apache Kafka and ksqlDB



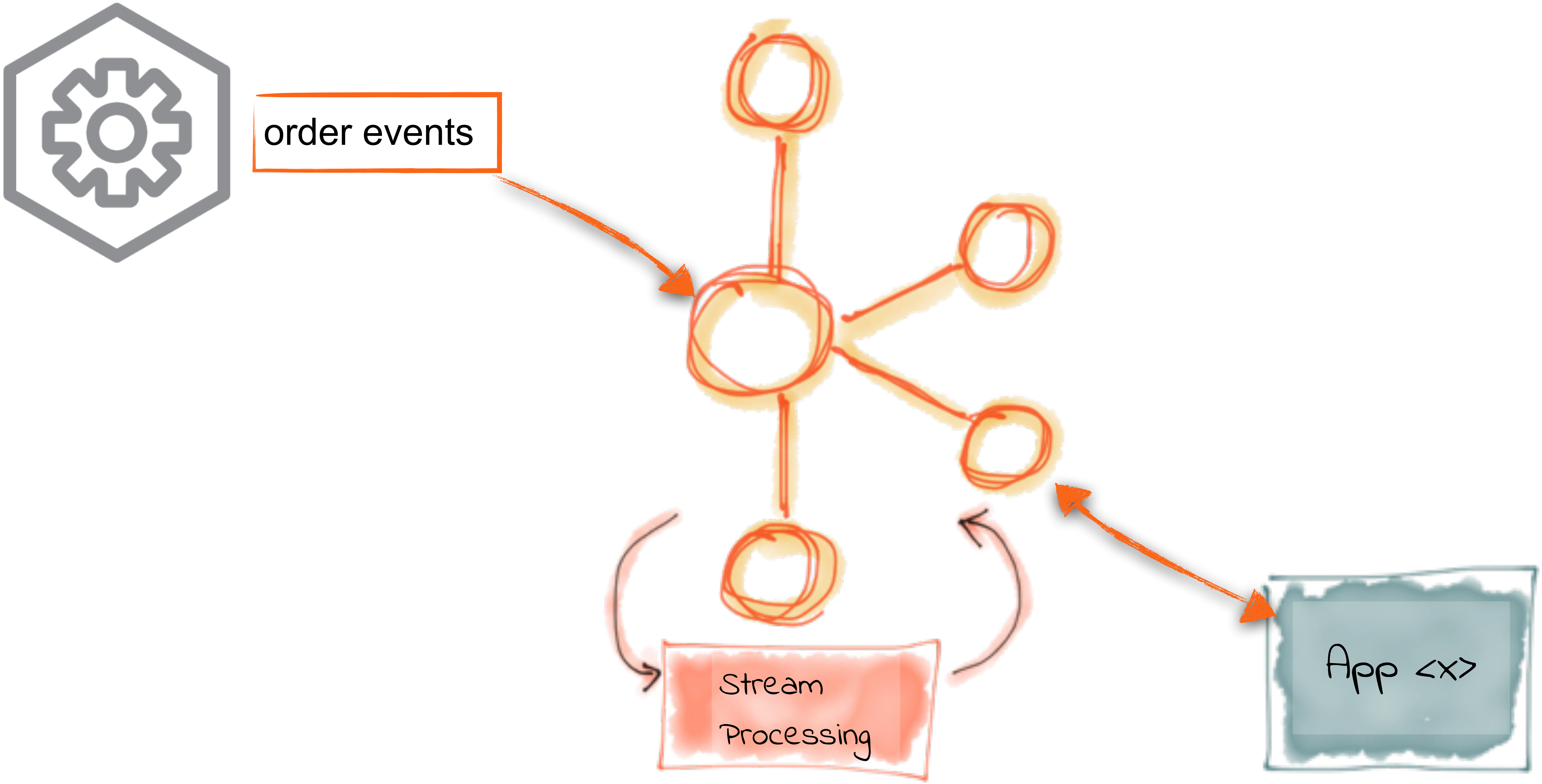
Real-time Event Stream Enrichment



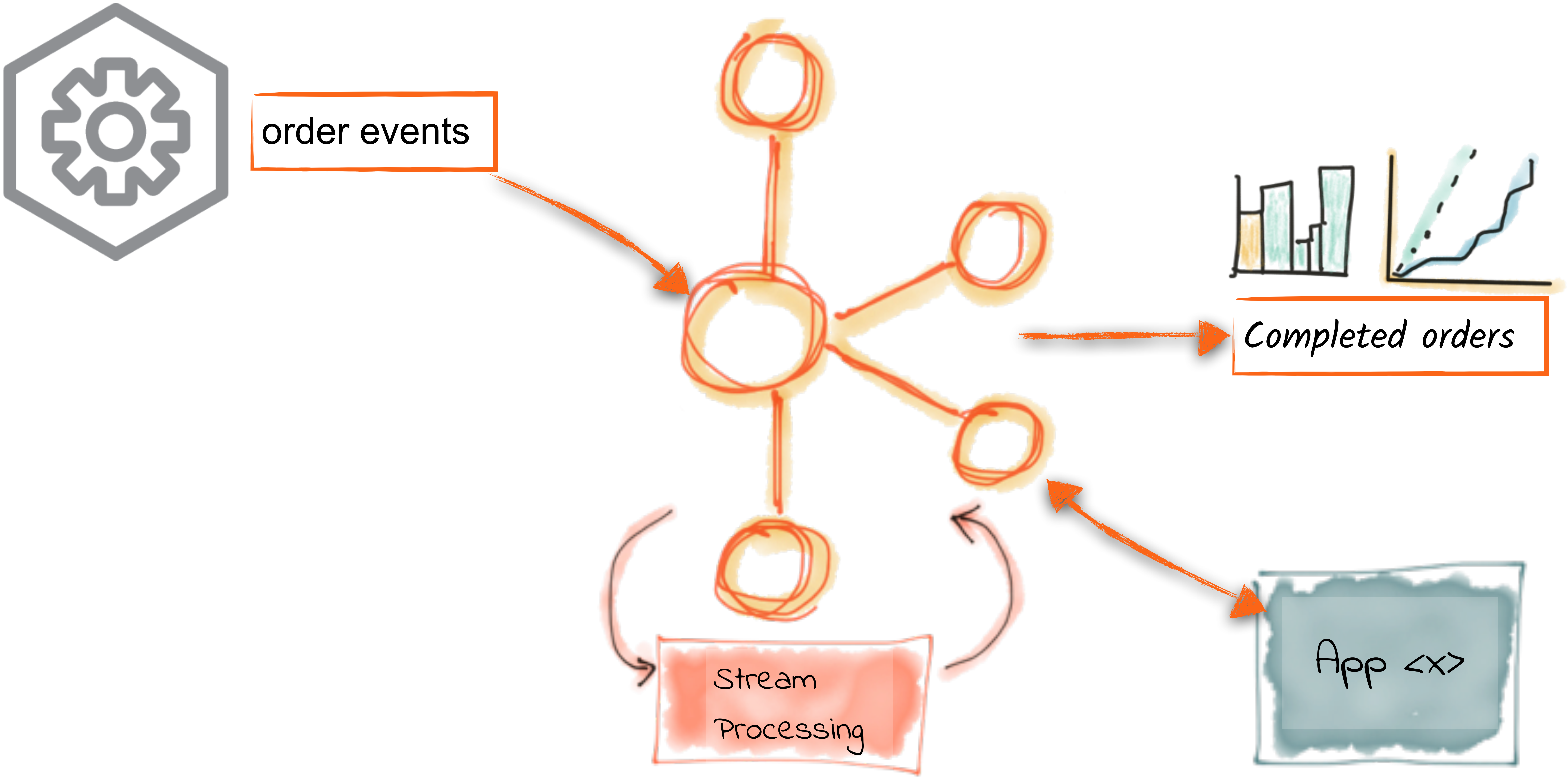
Message Bus



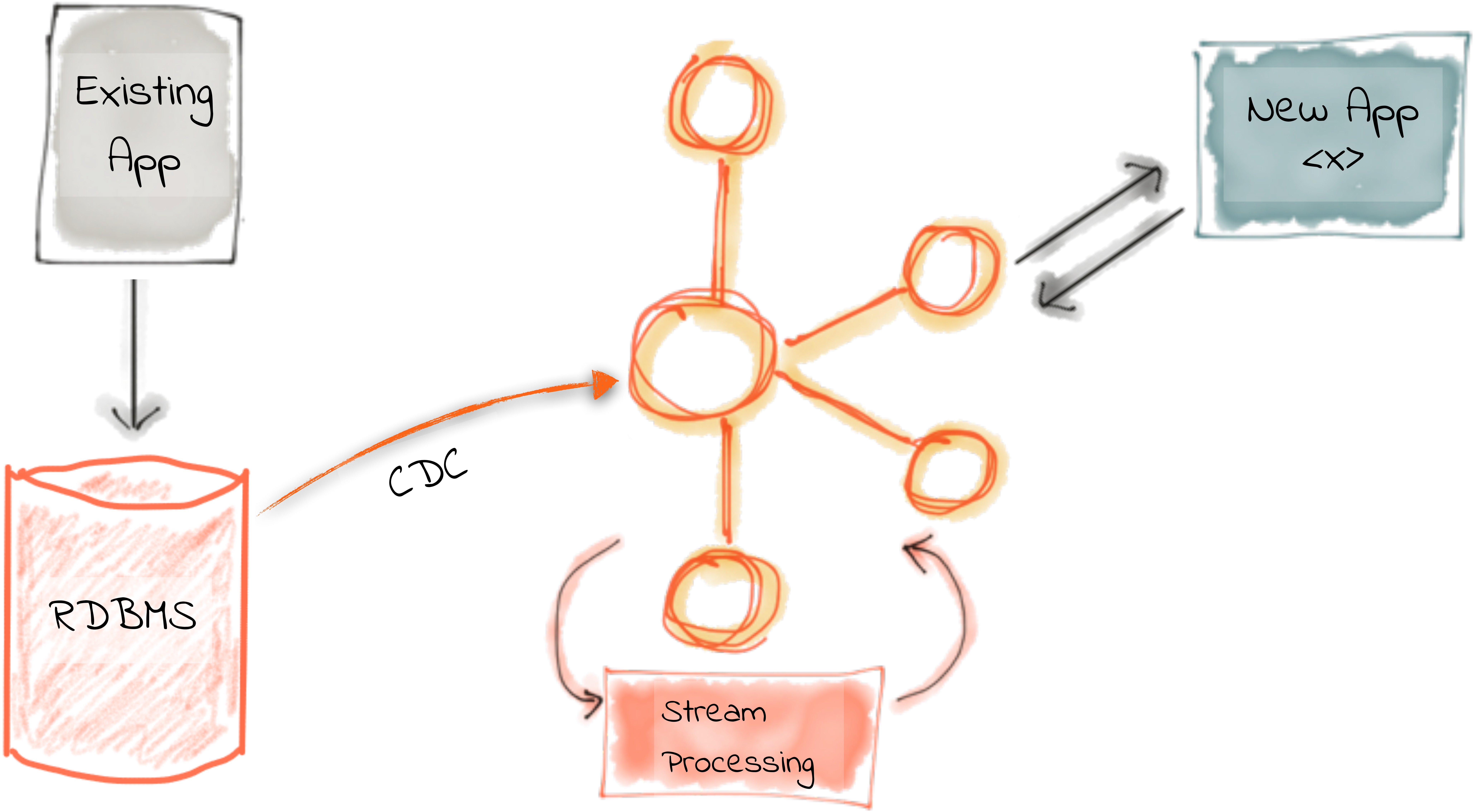
Stream Processing



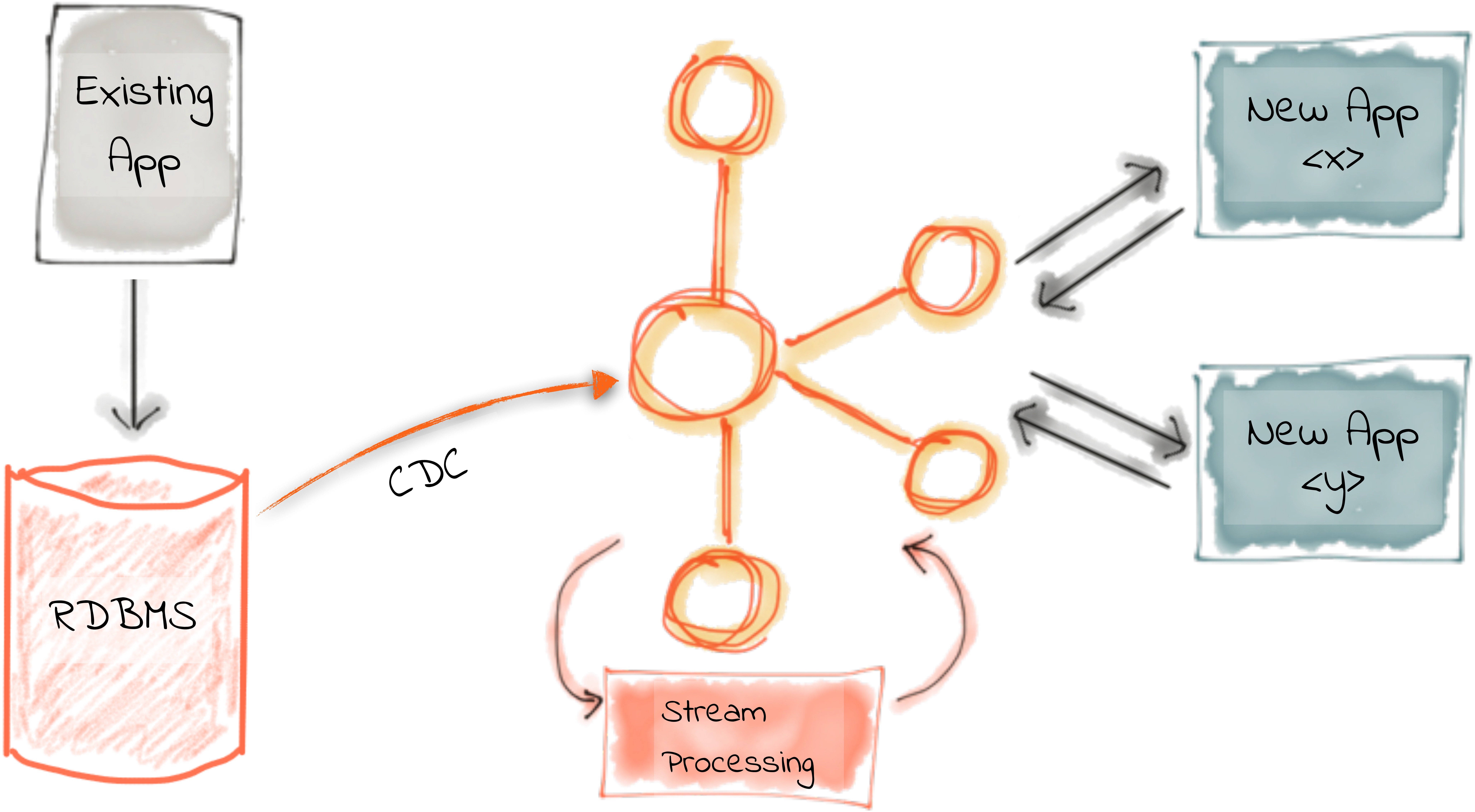
Transform Once, Use Many



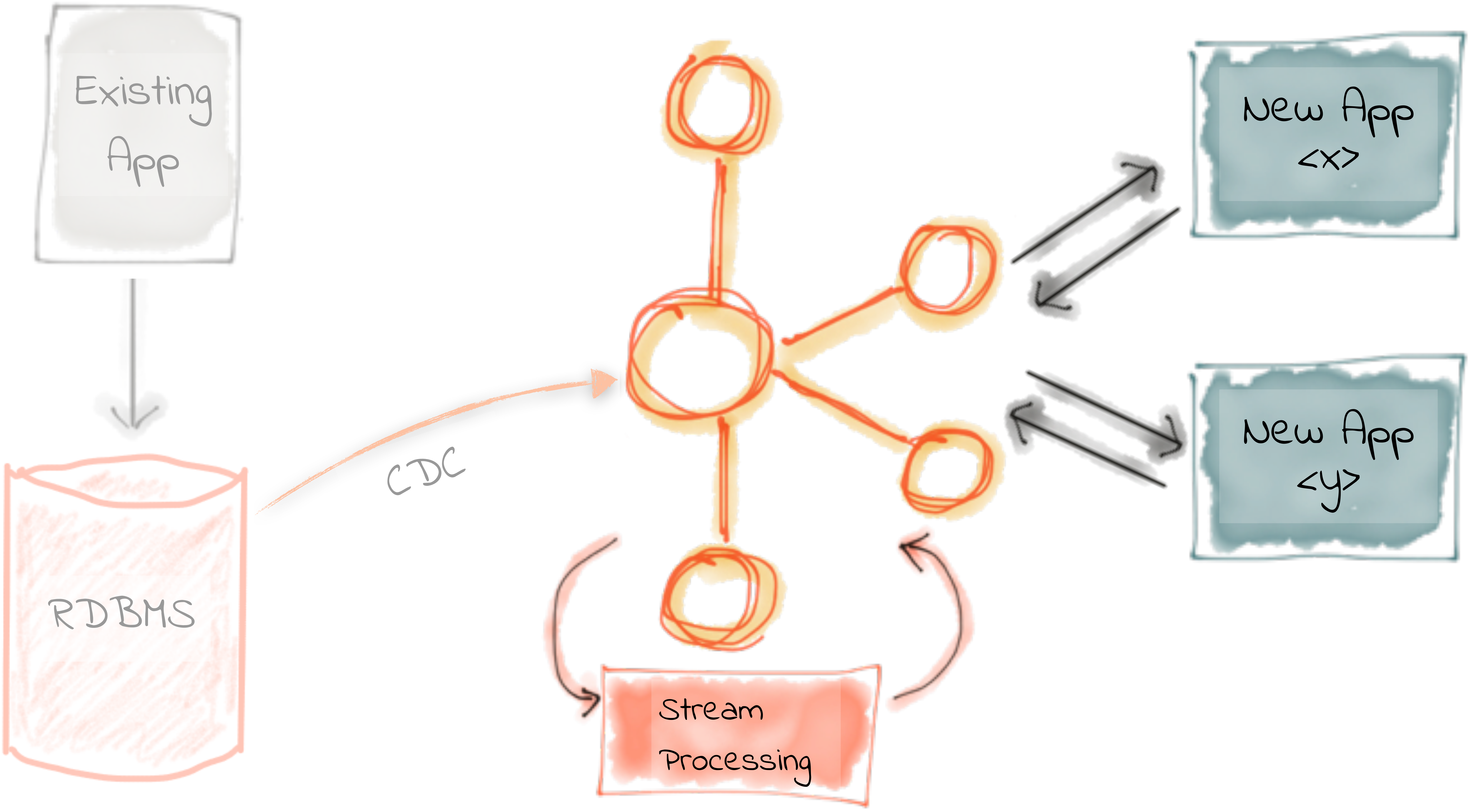
Evolve processing from old systems to new



Evolve processing from old systems to new



Evolve processing from old systems to new



Building a streaming data pipeline

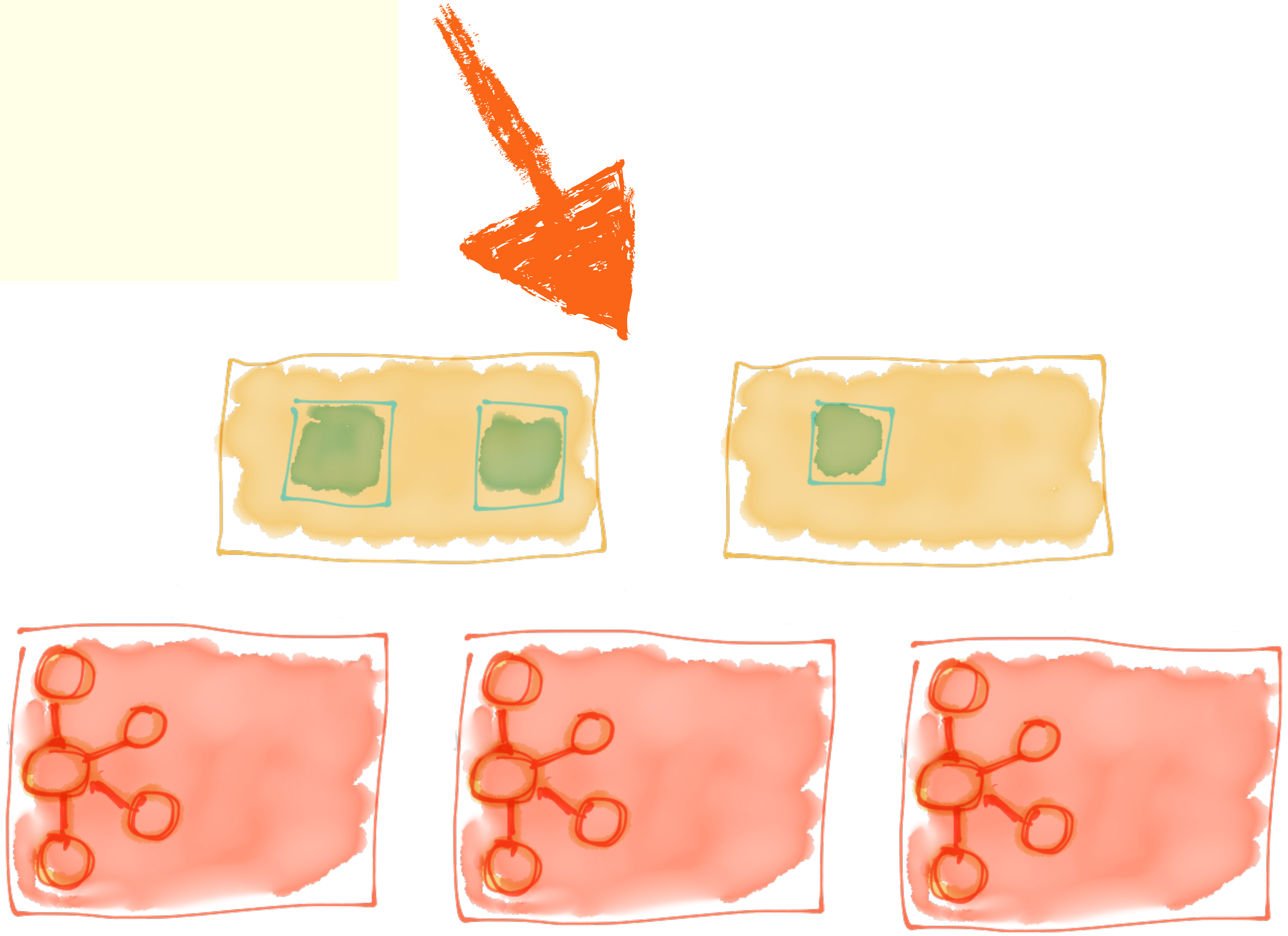
Integration + Stream Processing

Integration

Streaming Integration with Kafka Connect

JSON
MQTT.ORG
MySQL
ORACLE
syslog
CSV
salesforce
http://
Microsoft SQL Server

Sources

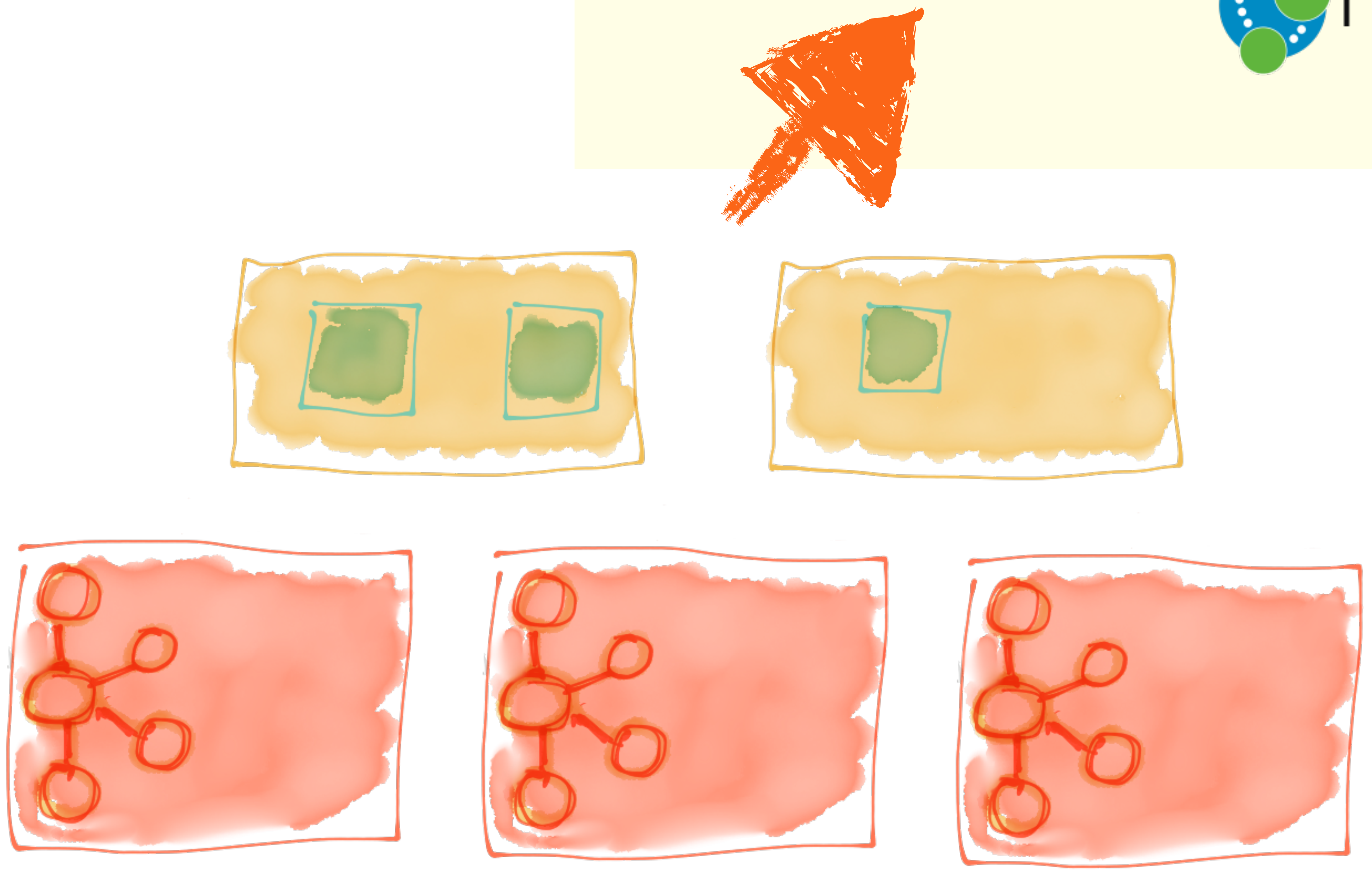


Kafka Connect

Kafka Brokers

Streaming Integration with Kafka Connect

Sinks



Kafka Connect

Kafka Brokers

Streaming Integration with Kafka Connect

Look Ma, No Code!

```
{
  "connector.class":
    "io.confluent.connect.jdbc.JdbcSourceConnector",
  "connection.url":
    "jdbc:mysql://asgard:3306/demo",
  "table.whitelist":
    "sales,orders,customers"
}
```

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

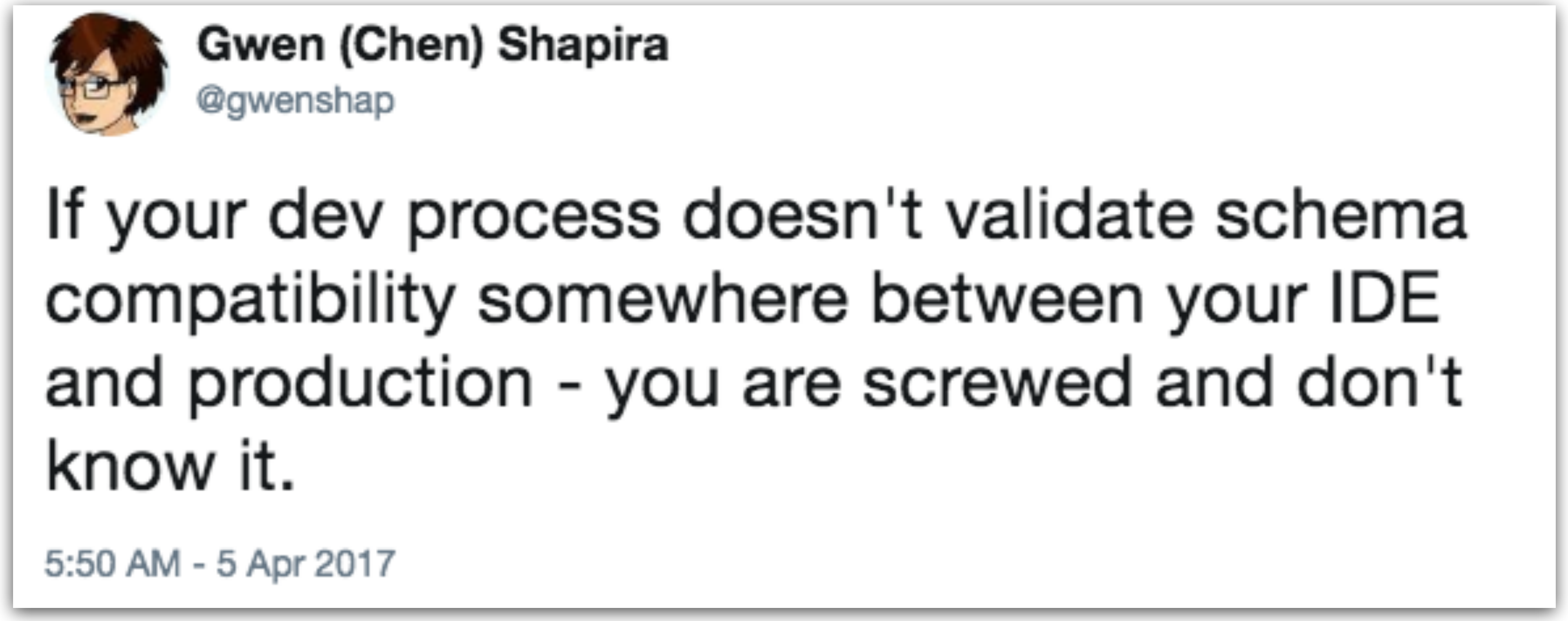
Serialisation & Schemas

Avro -> Confluent Schema Registry

Protobuf

JSON

CSV


Gwen (Chen) Shapira
@gwenshap

If your dev process doesn't validate schema compatibility somewhere between your IDE and production - you are screwed and don't know it.

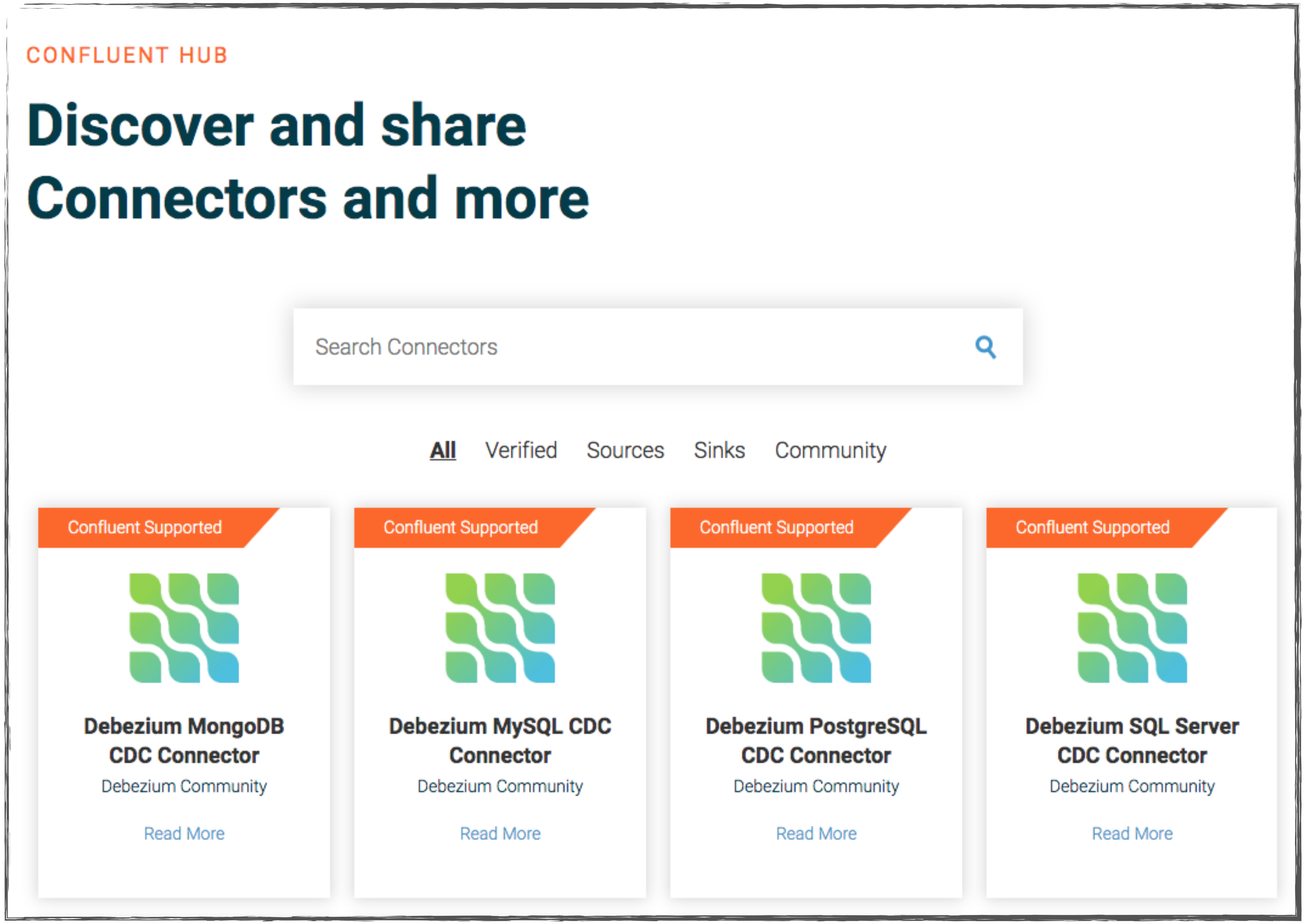
5:50 AM - 5 Apr 2017

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

The Confluent Schema Registry

Extensible

Confluent Hub



hub.confluent.io

Integration + Stream Processing

Stream Processing

```json
{
    "reading_ts": "2020-02-14T12:19:27Z",
    "sensor_id": "aa-101",
    "production_line": "w01",
    "widget_type": "acme94",
    "temp_celcius": 23,
    "widget_weight_g": 100
}
```


```
SELECT *
FROM WIDGETS
WHERE WEIGHT_G > 120
```


```
{
    "reading_ts": "2020-02-14T12:19:27Z",
    "sensor_id": "aa-101",
    "production_line": "w01",
    "widget_type": "acme94",
    "temp_celcius": 23,
    "widget_weight_g": 100
}
```


```
SELECT COUNT(*)
FROM WIDGETS
GROUP BY PRODUCTION_LINE
```

```
SELECT AVG(TEMP_CELCIUS) AS TEMP
FROM WIDGETS
GROUP BY SENSOR_ID
HAVING TEMP>20
```

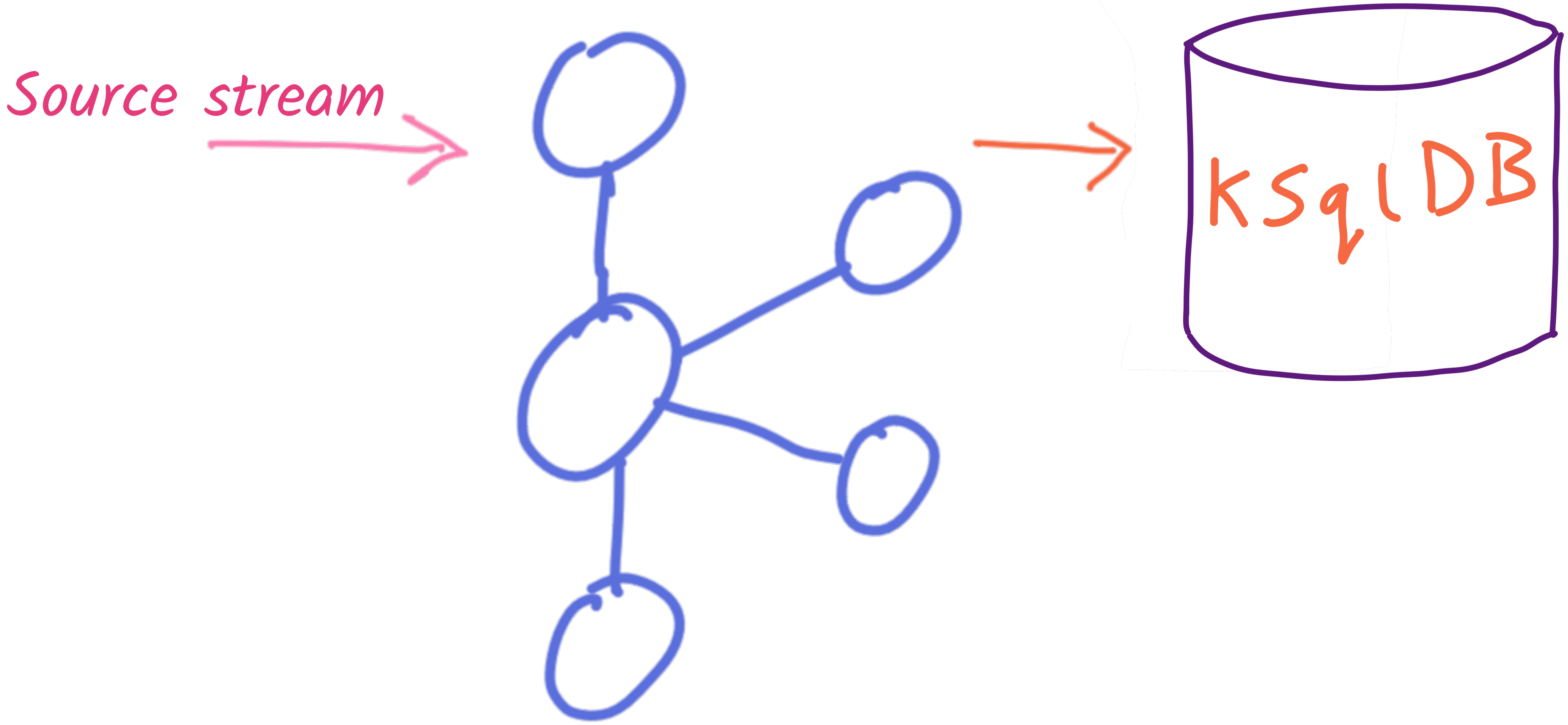
Object store,
data warehouse,
RDBMS

```
CREATE SINK CONNECTOR dw WITH (
'connector.class' = 'S3Connector',
'topics' = 'widgets'
…);
```

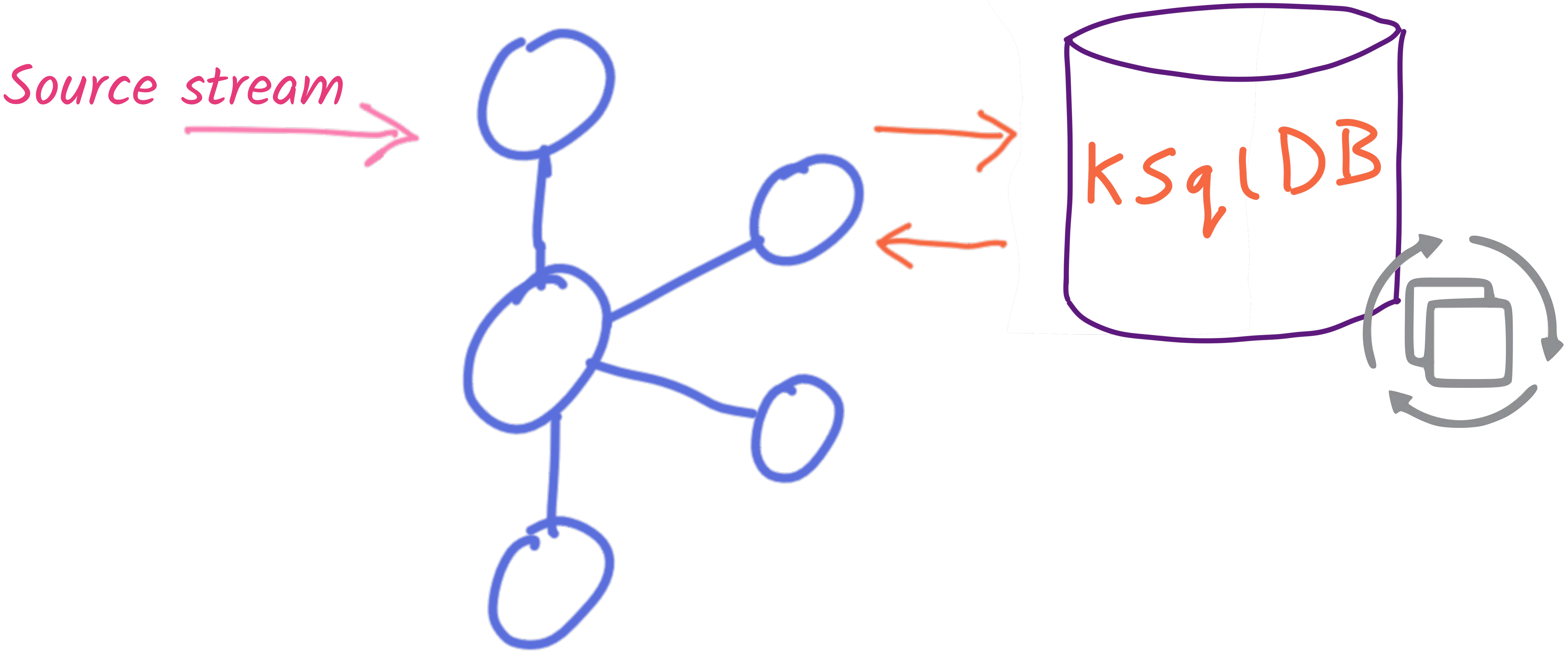
ksqlDB

The event streaming database purpose-built for stream processing applications.

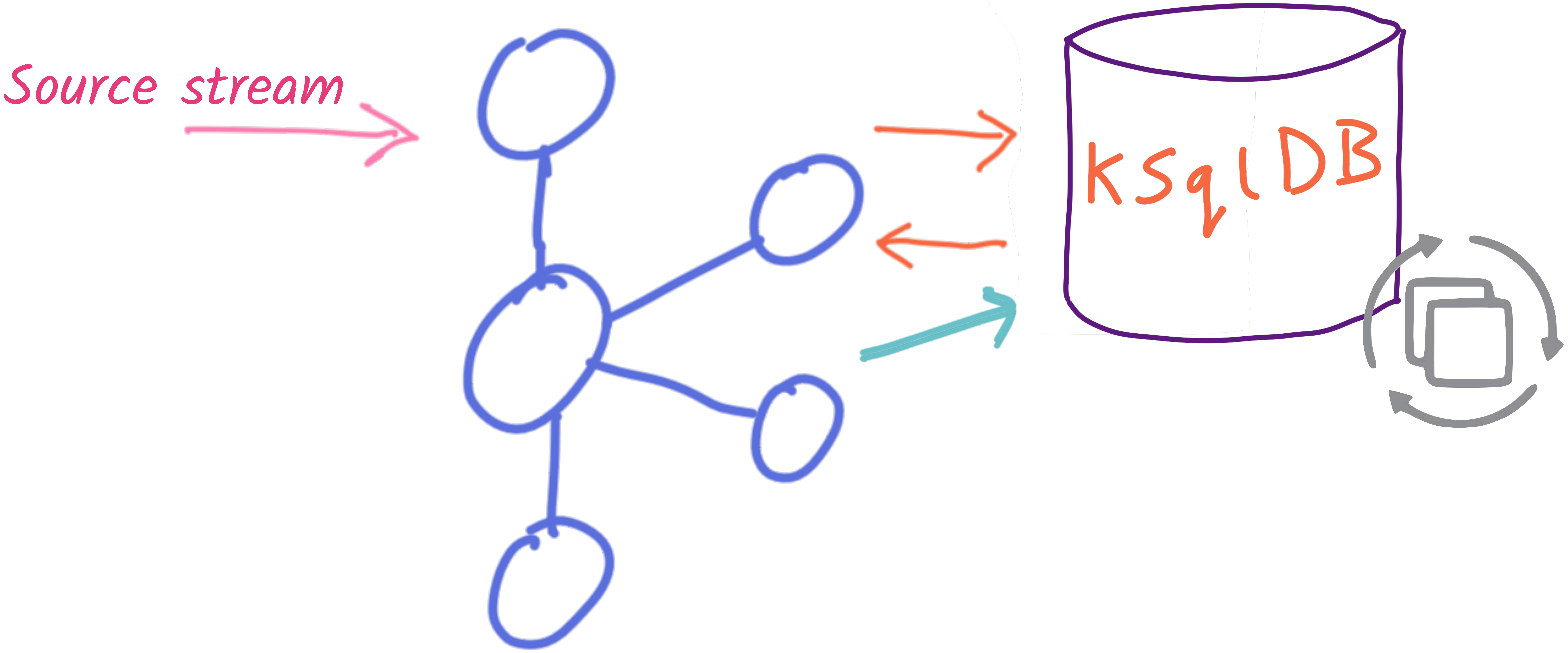
Stream Processing with ksqlDB



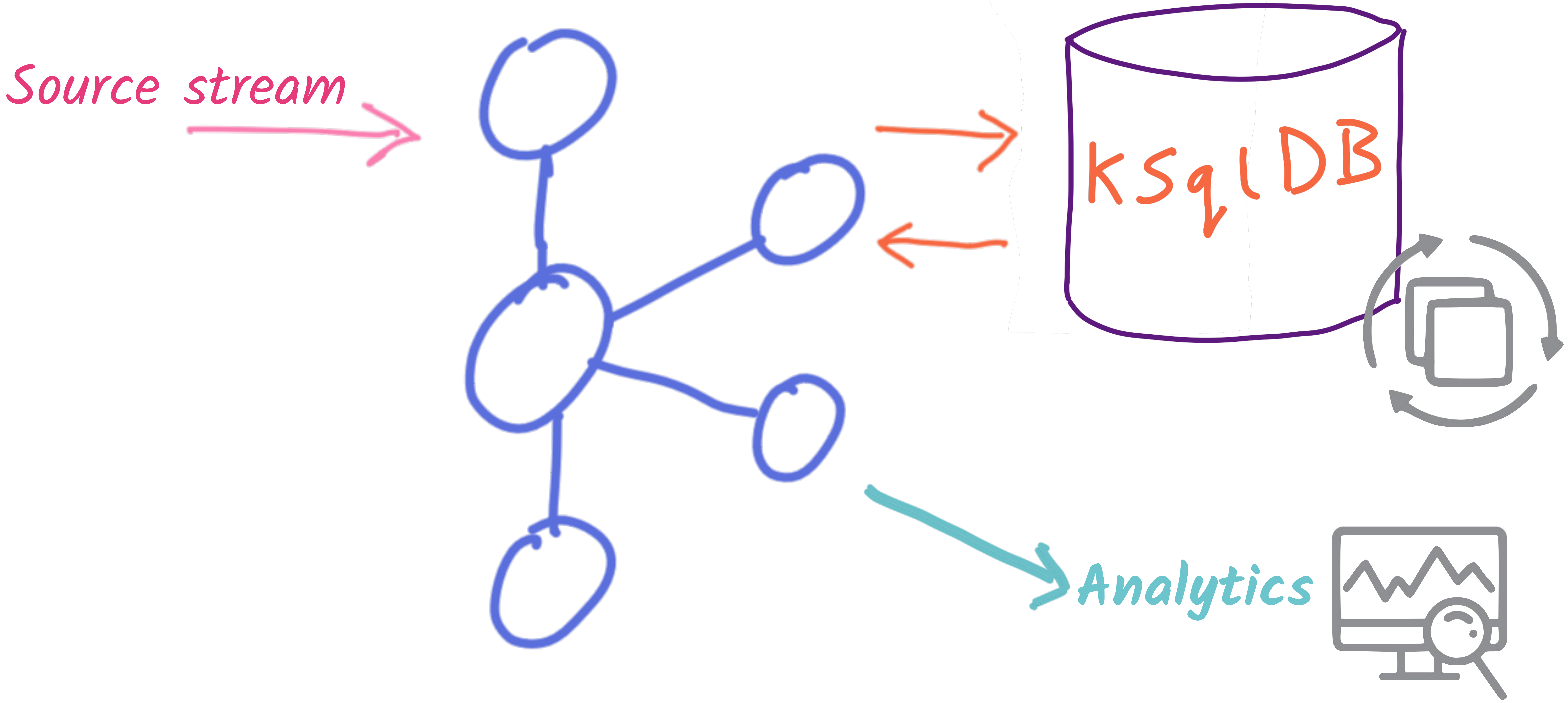
Stream Processing with ksqlDB



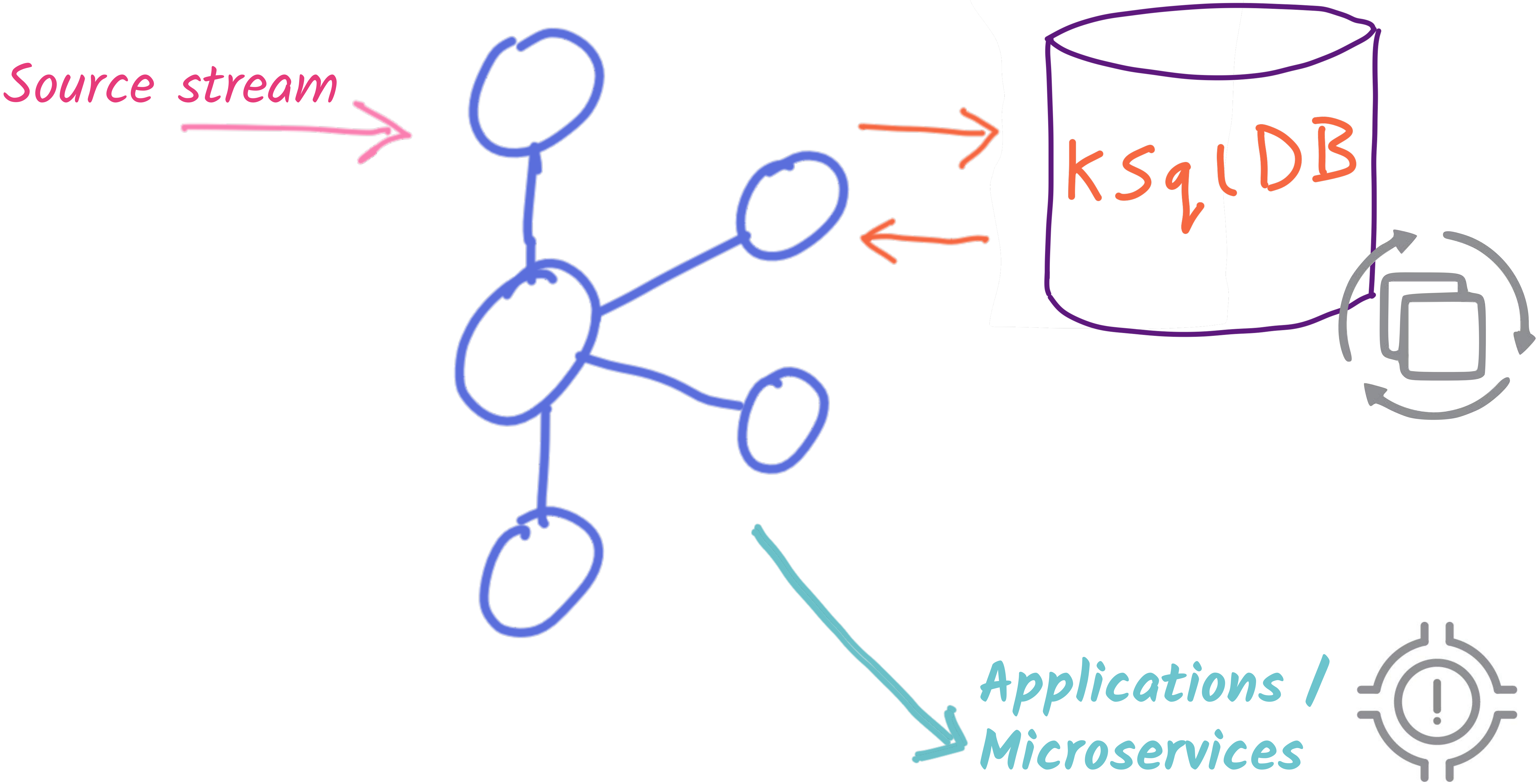
Stream Processing with ksqlDB



Stream Processing with ksqlDB

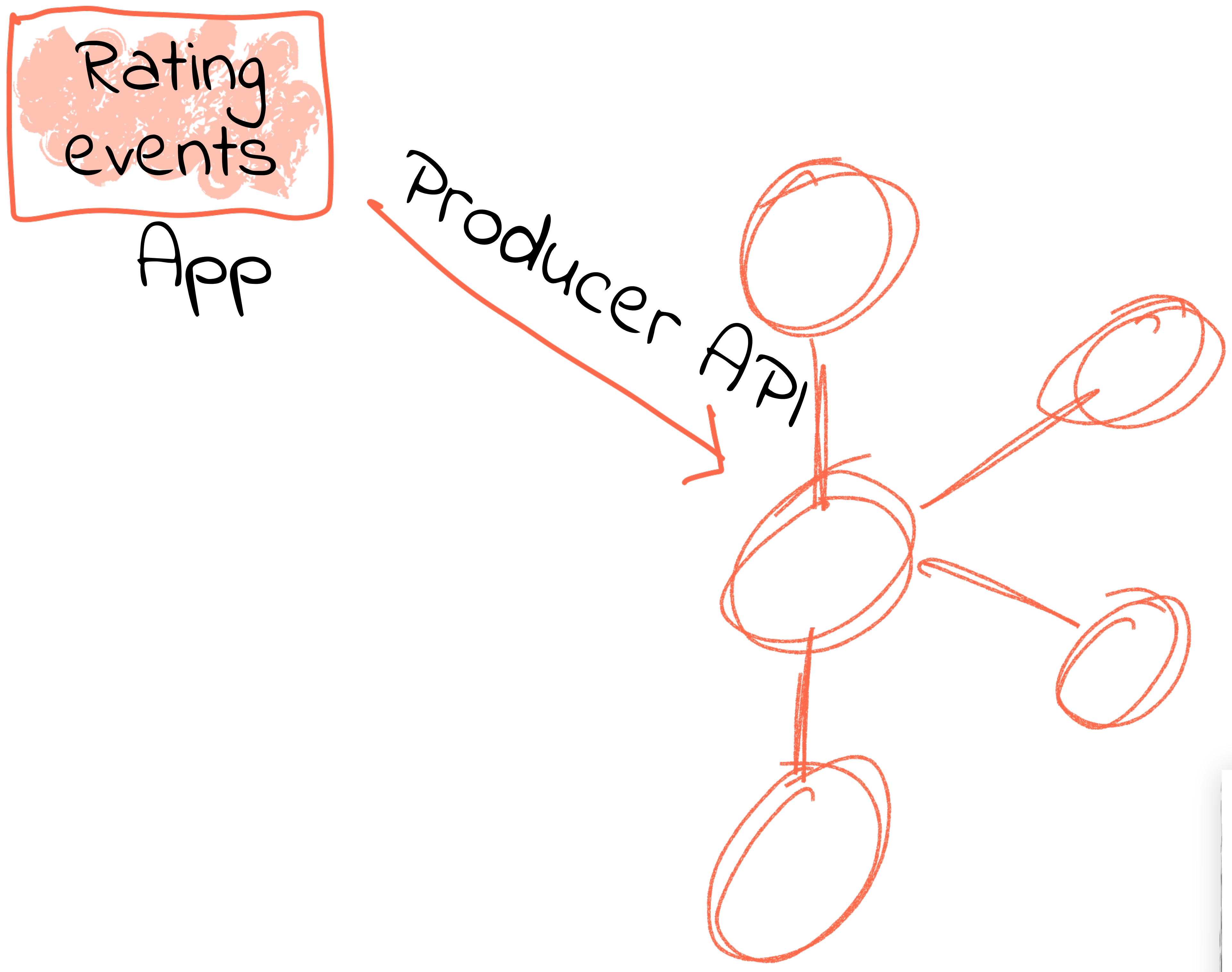


Stream Processing with ksqlDB



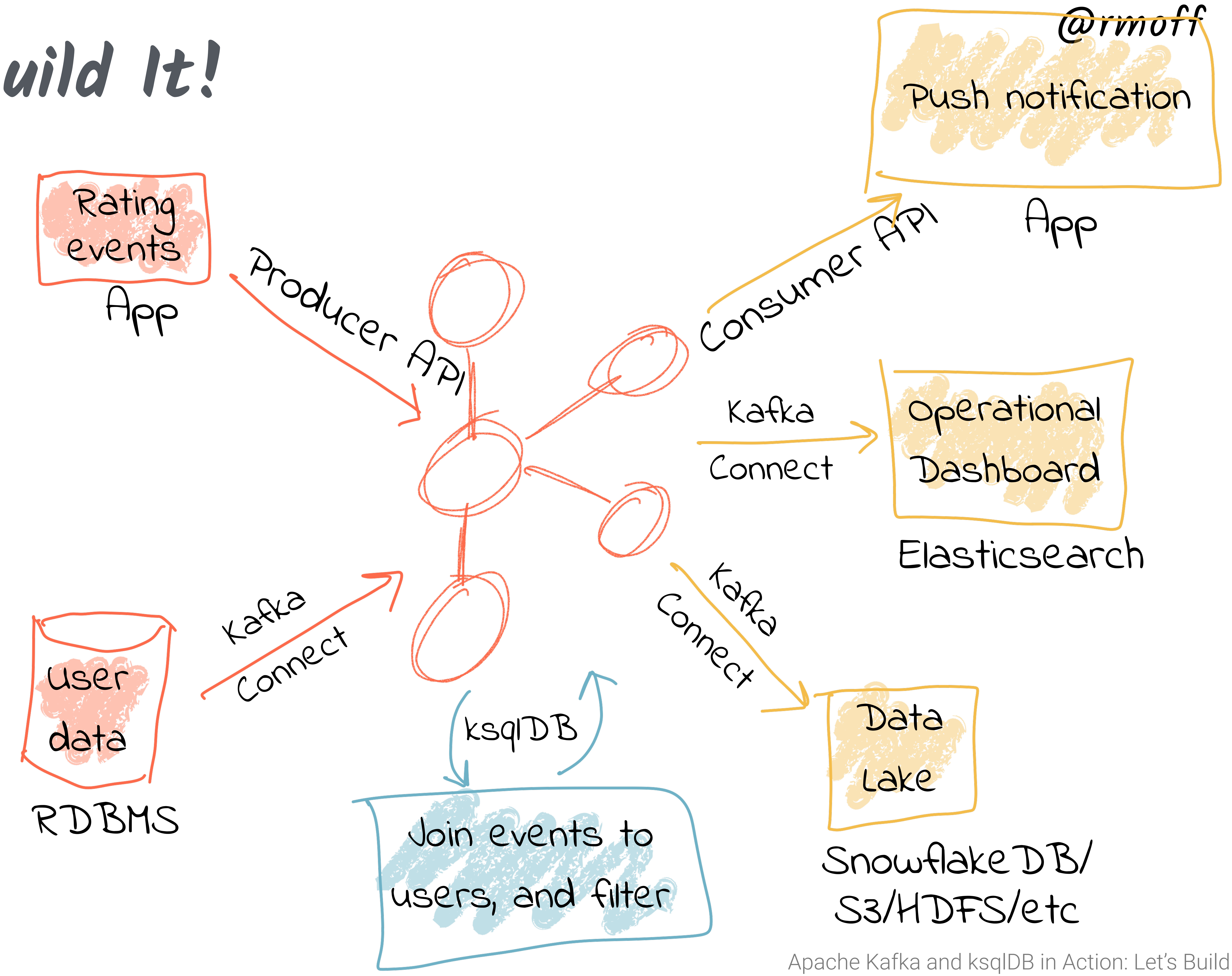
Stream Processing with ksqlDB

Let's Build It!

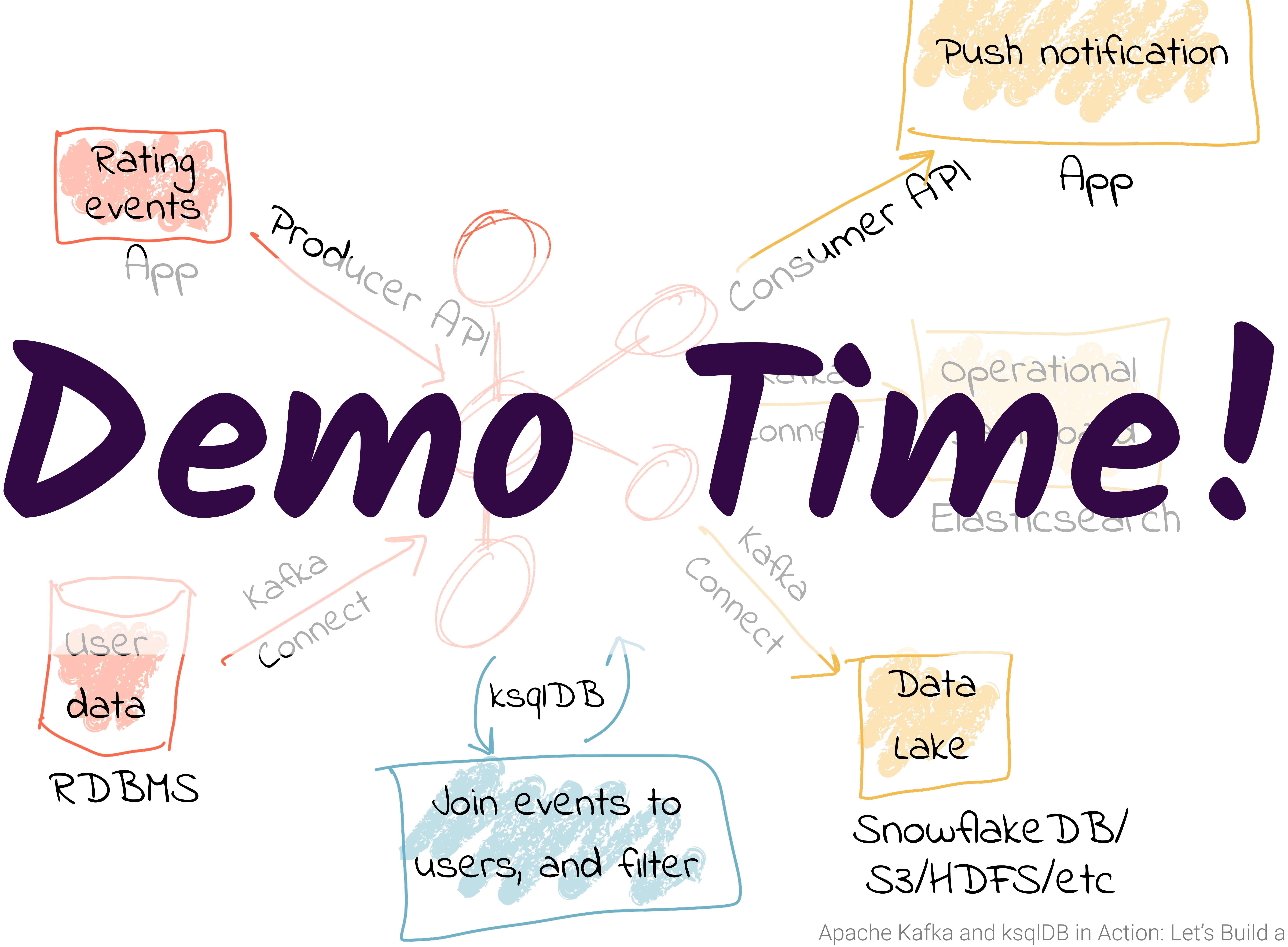


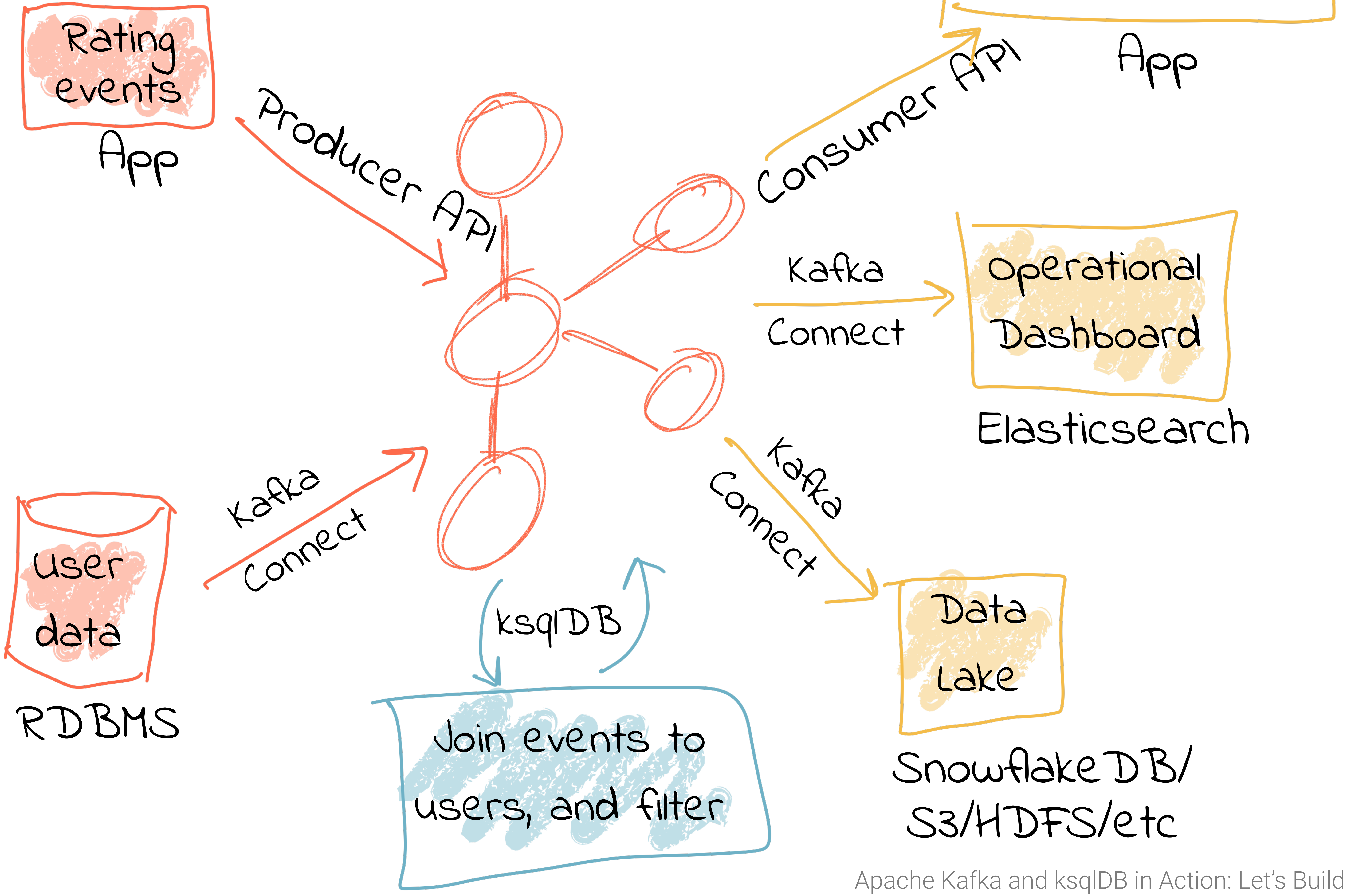

```json
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304105213,
 "channel": "web",
 "message": "worst. flight. ever. #neveragain"
}
```

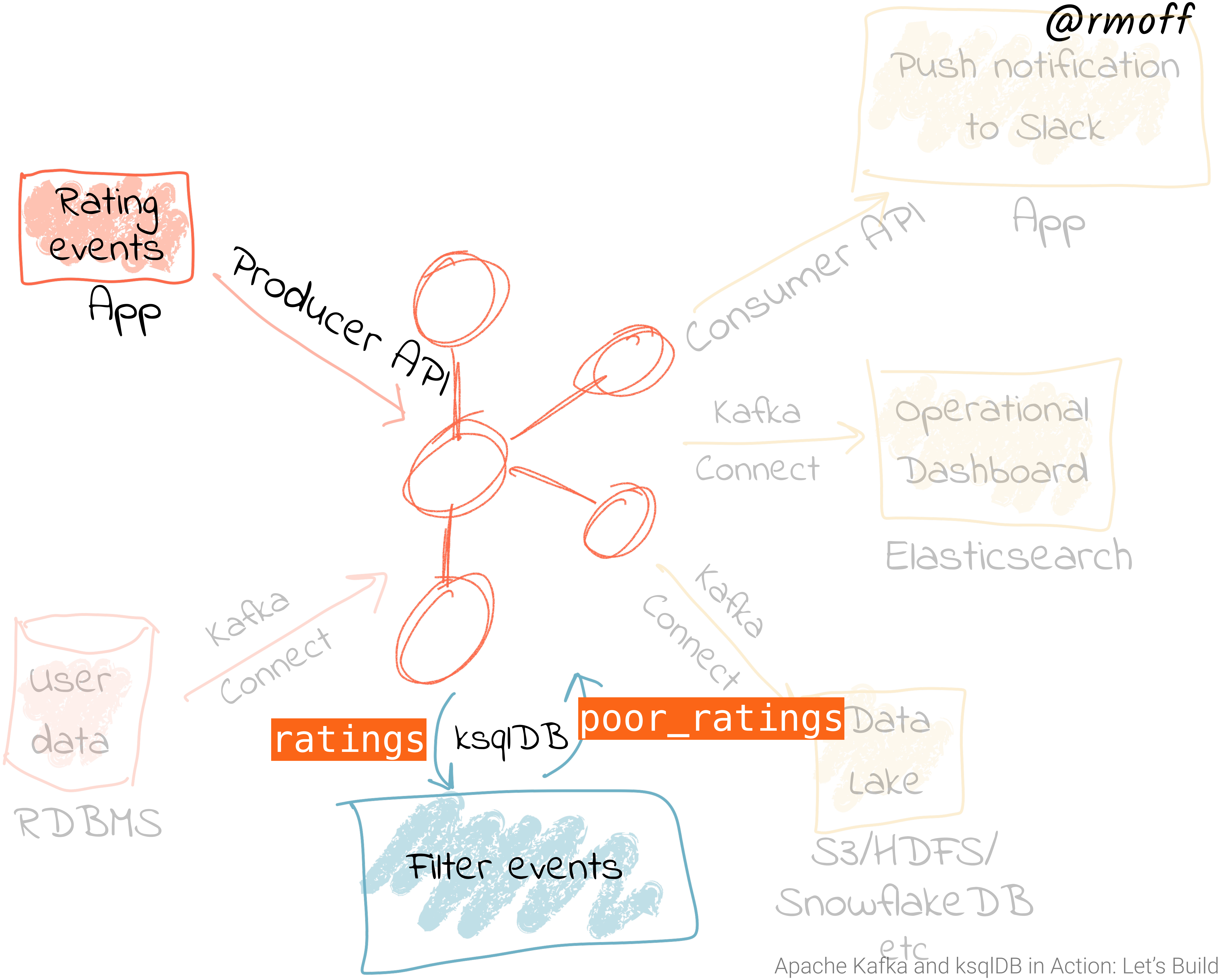
Let's Build It!



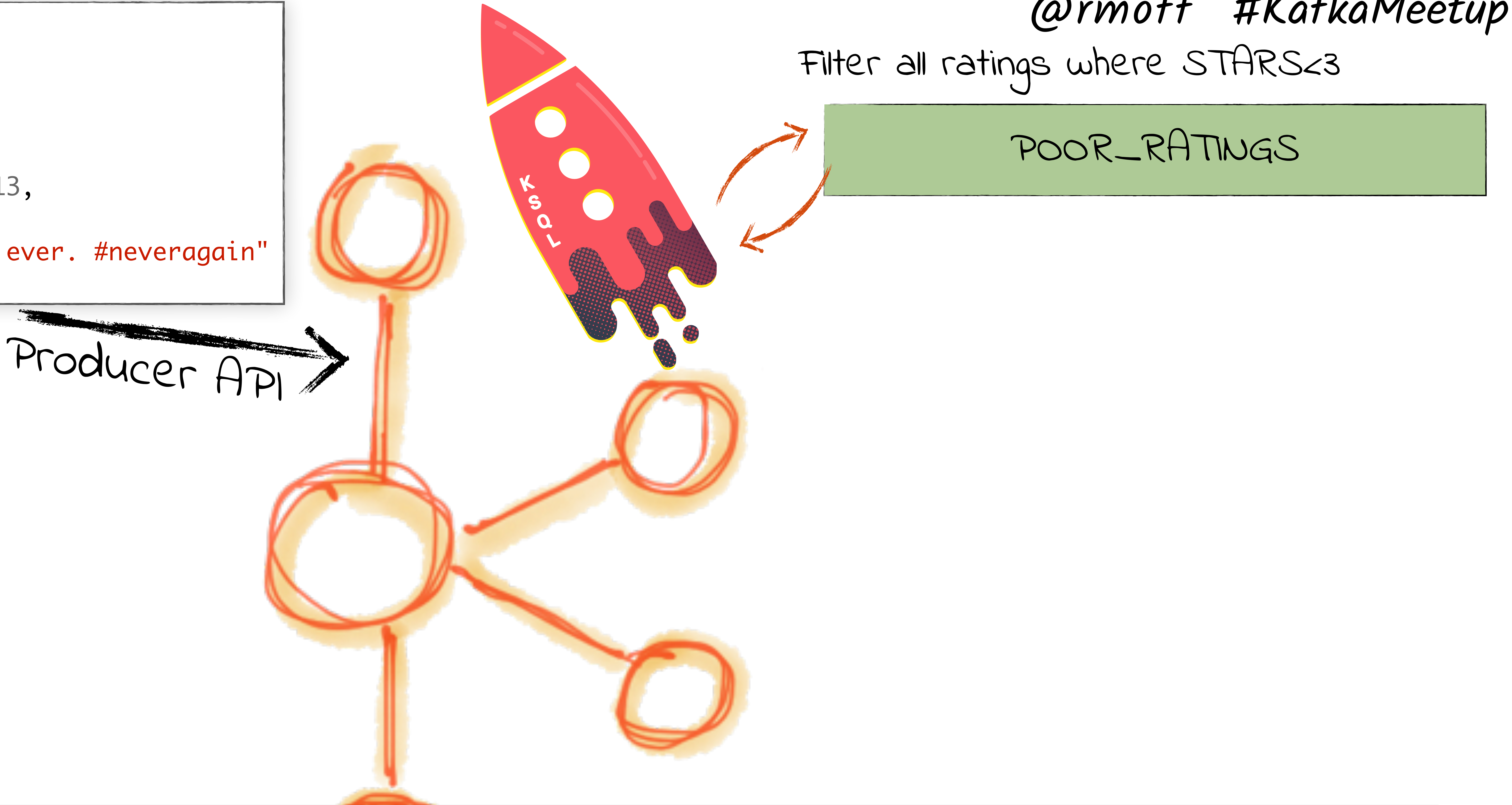
Demo Time!


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


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

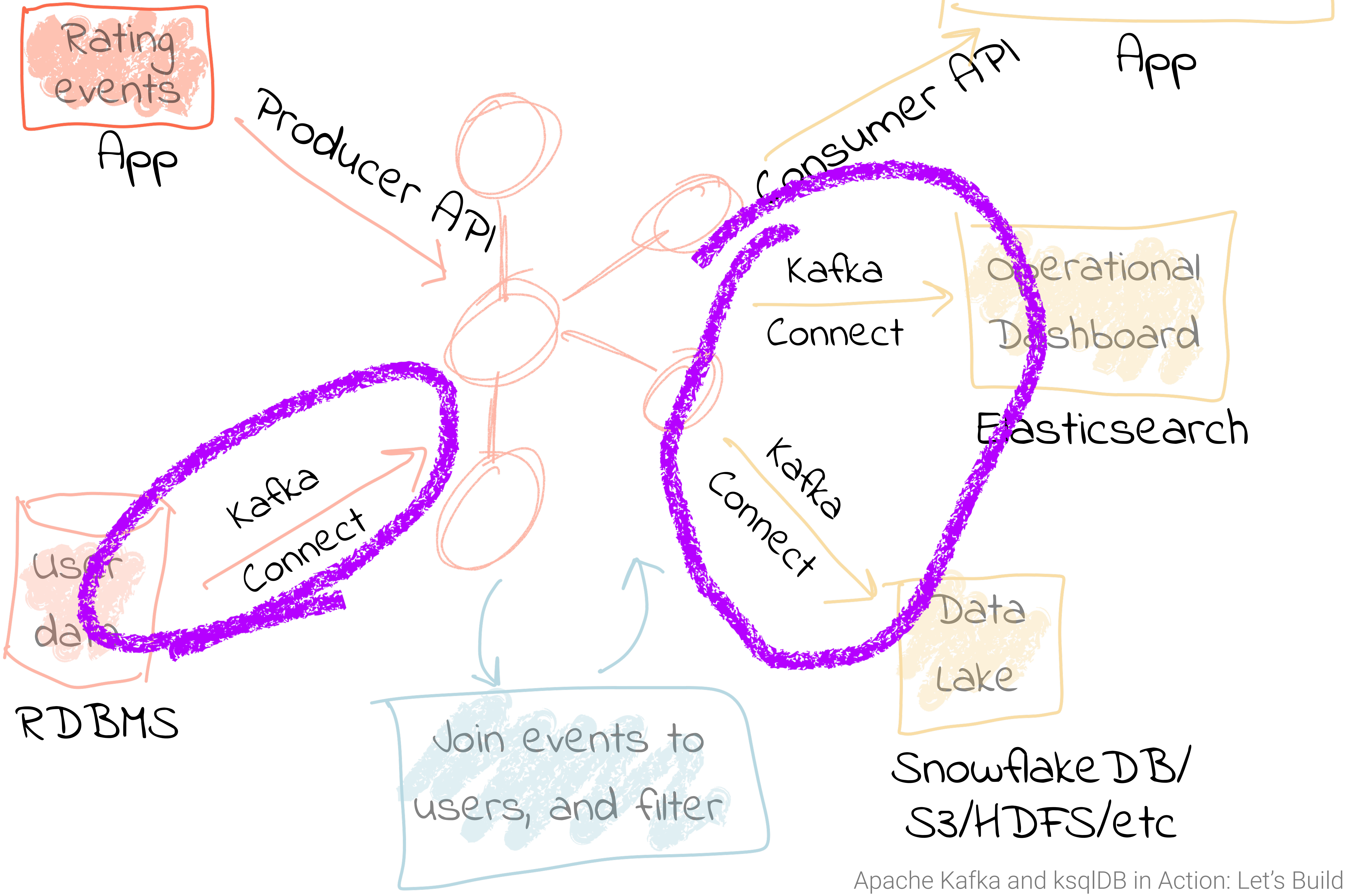

Rating events
App
Producer API
Consumer API
Push notification to Slack
App
Kafka Connect
Operational Dashboard
Elasticsearch
Kafka Connect
Data Lake
S3/HDFS/ SnowflakeDB etc
User data
RDBMS
Kafka Connect
ratings
ksqlDB
poor_ratings
Filter events

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

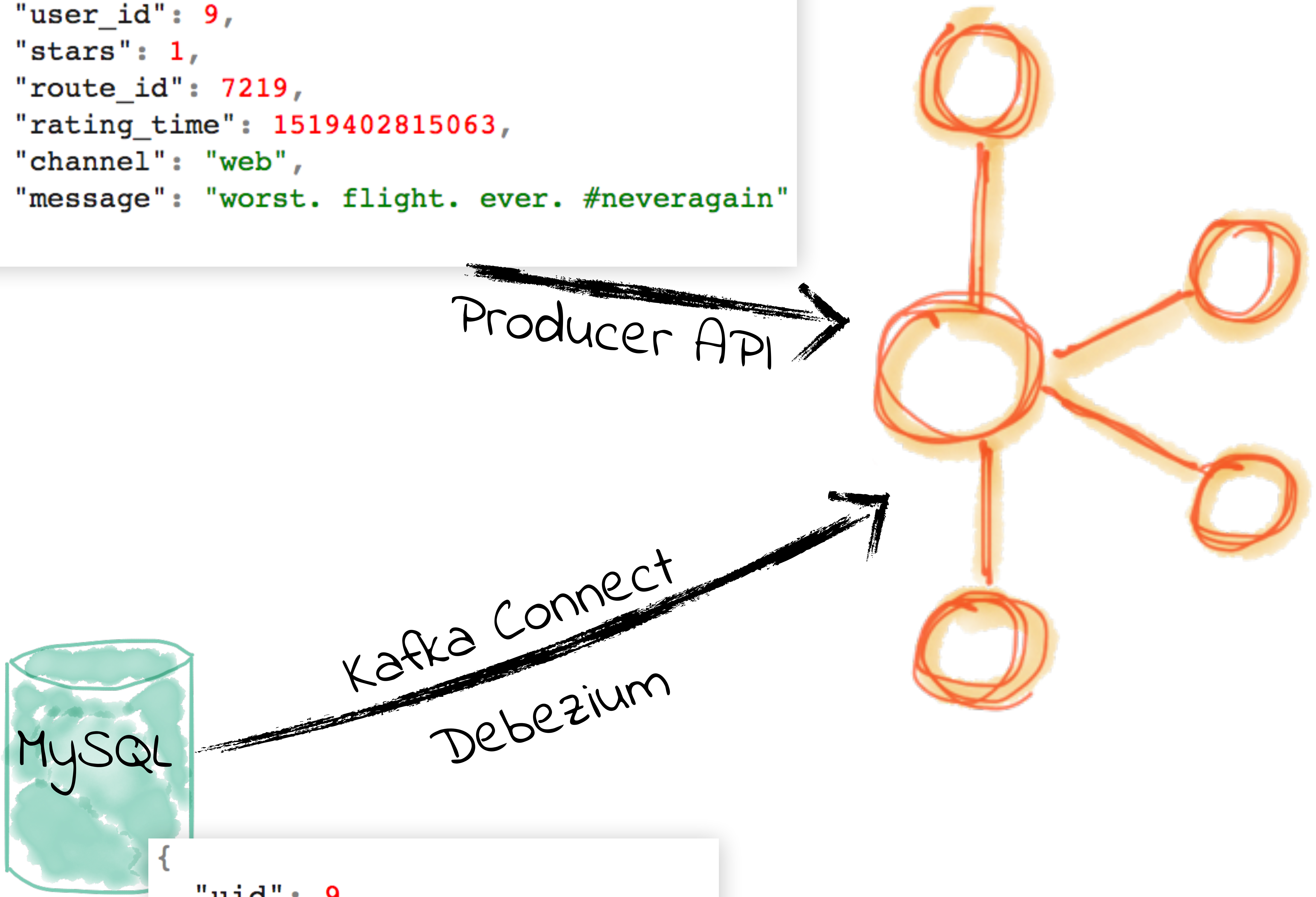




```sql
CREATE STREAM POOR_RATINGS AS
SELECT * FROM ratings WHERE STARS <3
```

Kafka Connect



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


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

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

The truth is the log.

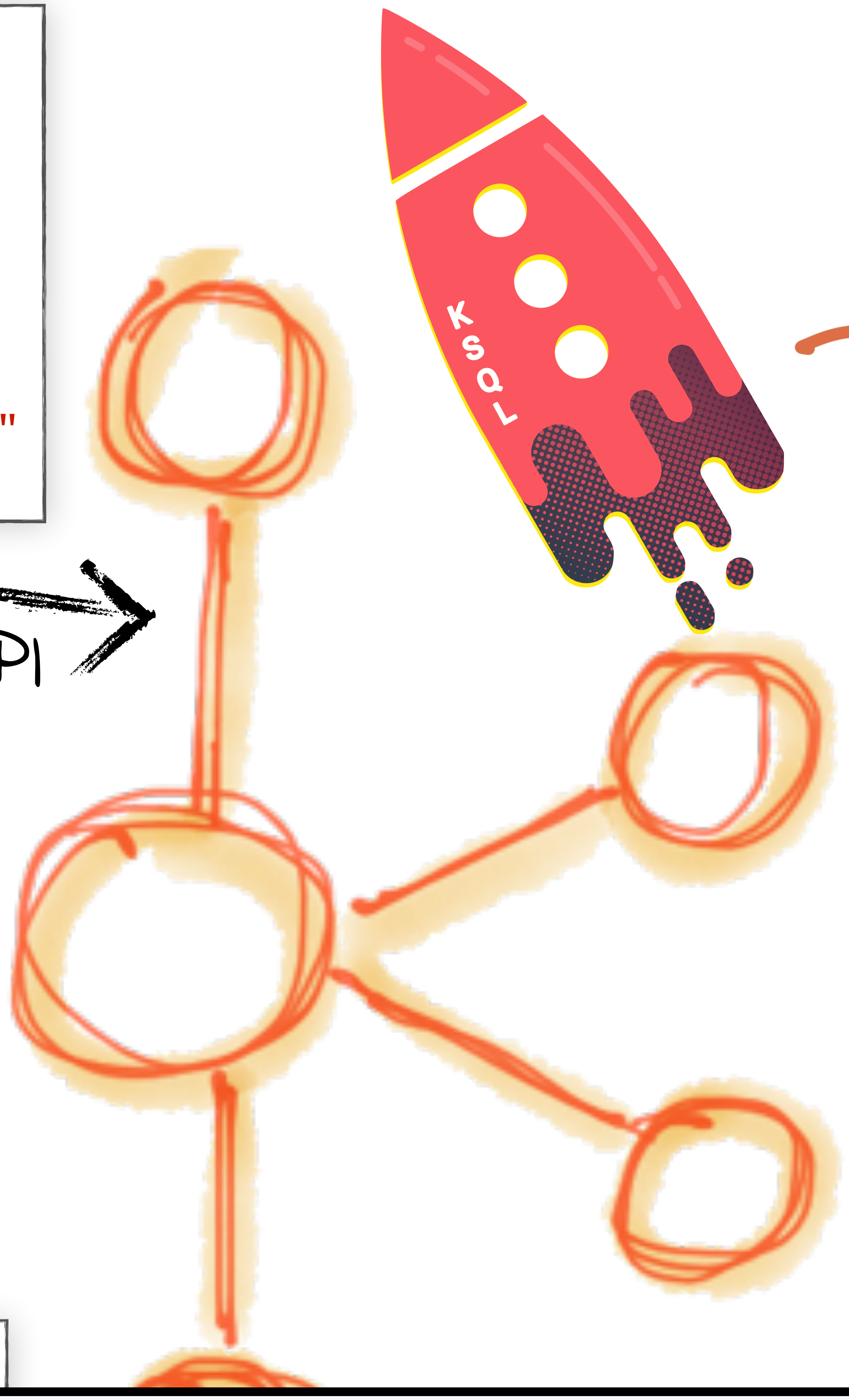
The database is a cache of a subset of the log.

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

Photo by Bobby Burch on Unsplash

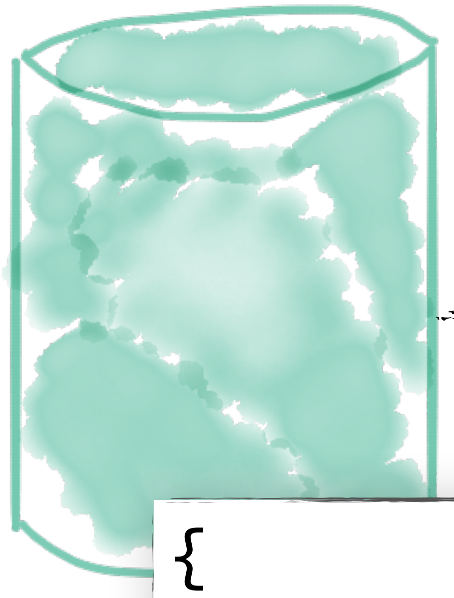
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304105213,
 "channel": "web",
 "message": "worst. flight. ever. #neveragain"
}

Producer API



Join each rating to customer data

RATINGS_WITH_CUSTOMER_DATA



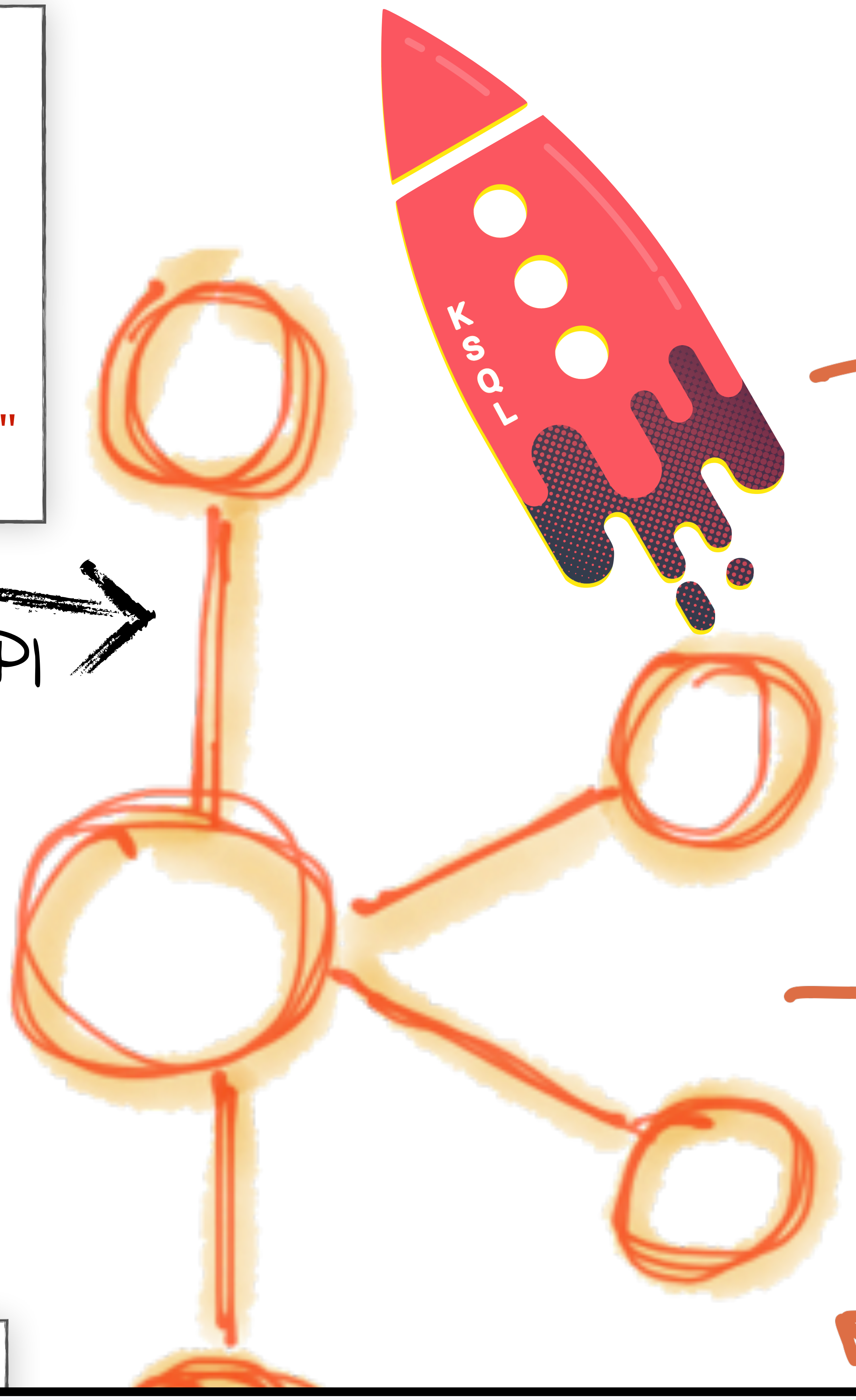
Kafka Connect

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

```
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
```

```
{
 "rating_id": 5313,
 "user_id": 3,
 "stars": 4,
 "route_id": 6975,
 "rating_time": 1519304
 "channel": "web",
 "message": "worst. fli
}
```


```sql
CREATE TABLE RATINGS_BY_CLUB_STATUS AS
SELECT    CLUB_STATUS, COUNT(*)
FROM      RATINGS_WITH_CUSTOMER_DATA
          WINDOW TUMBLING (SIZE 1 MINUTES)
GROUP BY CLUB_STATUS;
```

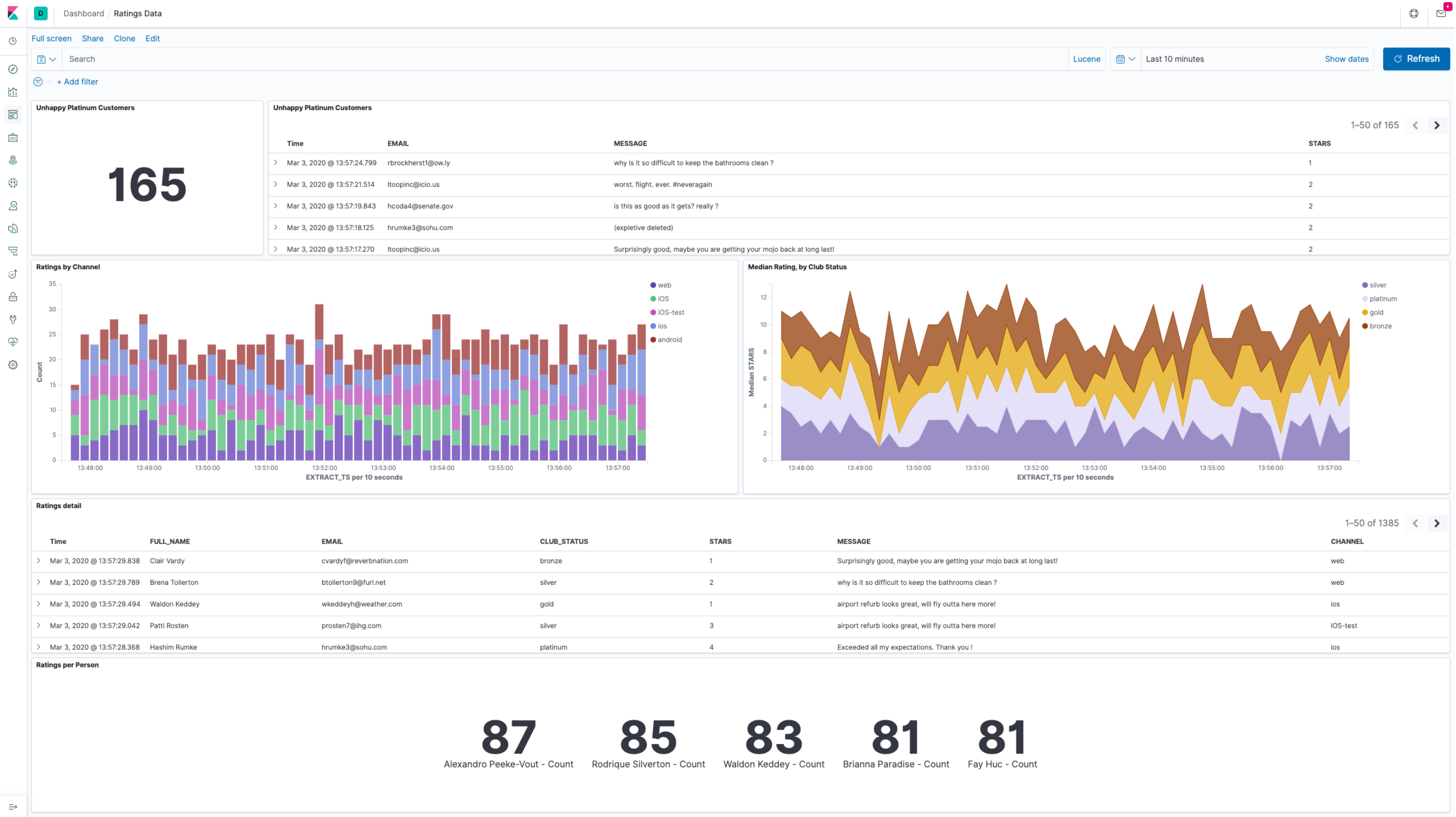
Kafka Connect

Aggregate per-minute by CLUB_STATUS

RATINGS_BY_CLUB_STATUS_1MIN


```
{
  "id": 3,
  "first_name": "Merilyn",
  "last_name": "Doughartie",
  "email": "mdoughartie1@dedecms.com",
  "gender": "Female",
  "club_status": "platinum",
  "comments": "none"
}
```

Kafka Connect → Elasticsearch



Want to learn more?

CTAs, not CATs
(sorry, not sorry)

kafka-summit.org

KAFKA SUMMIT 2020

LONDON

APRIL 27 & 28

Moffatt30

*30% OFF**

*Standard Priced Conference pass

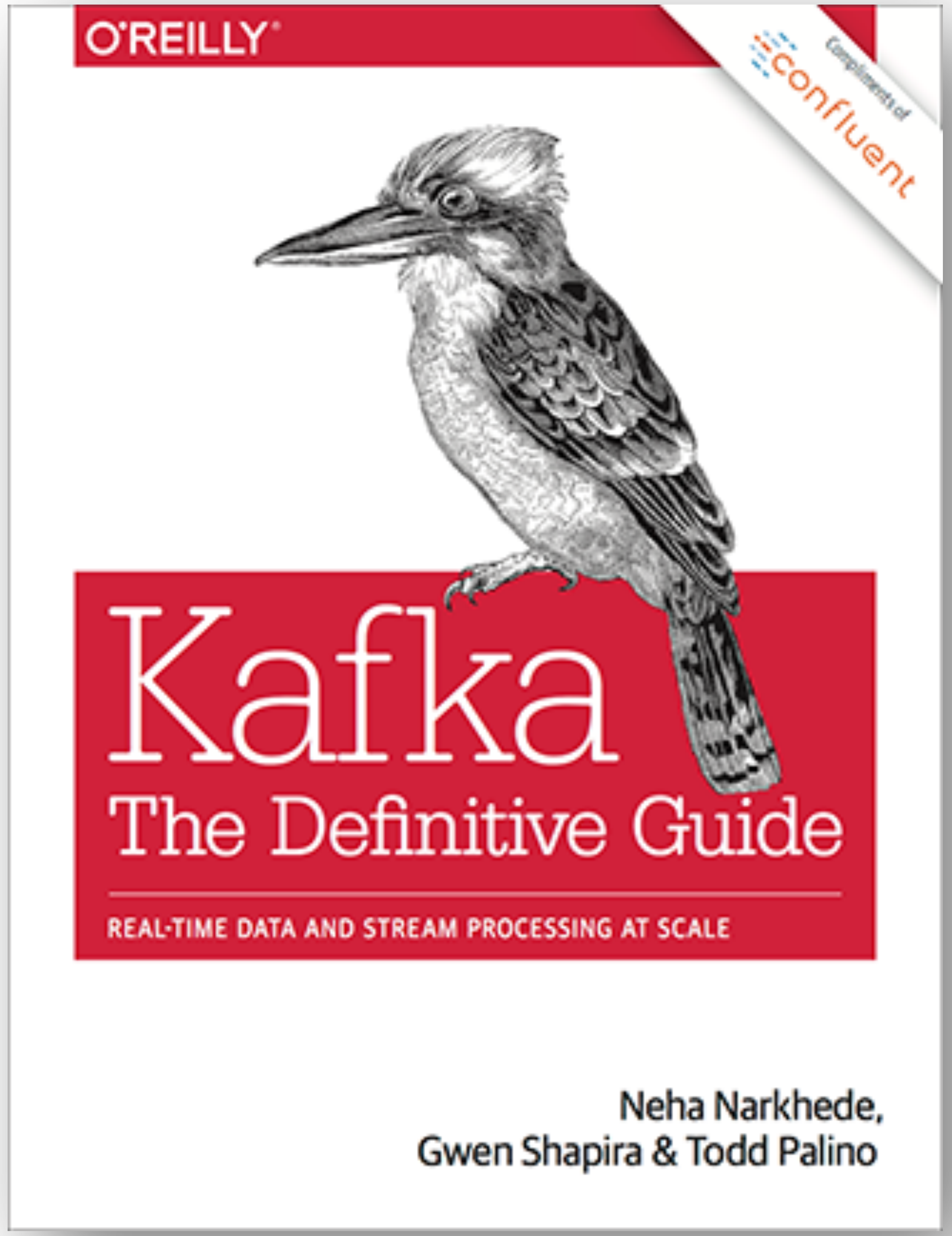

Learn Kafka.
Start building with Apache Kafka at Confluent Developer.
developer.confluent.io

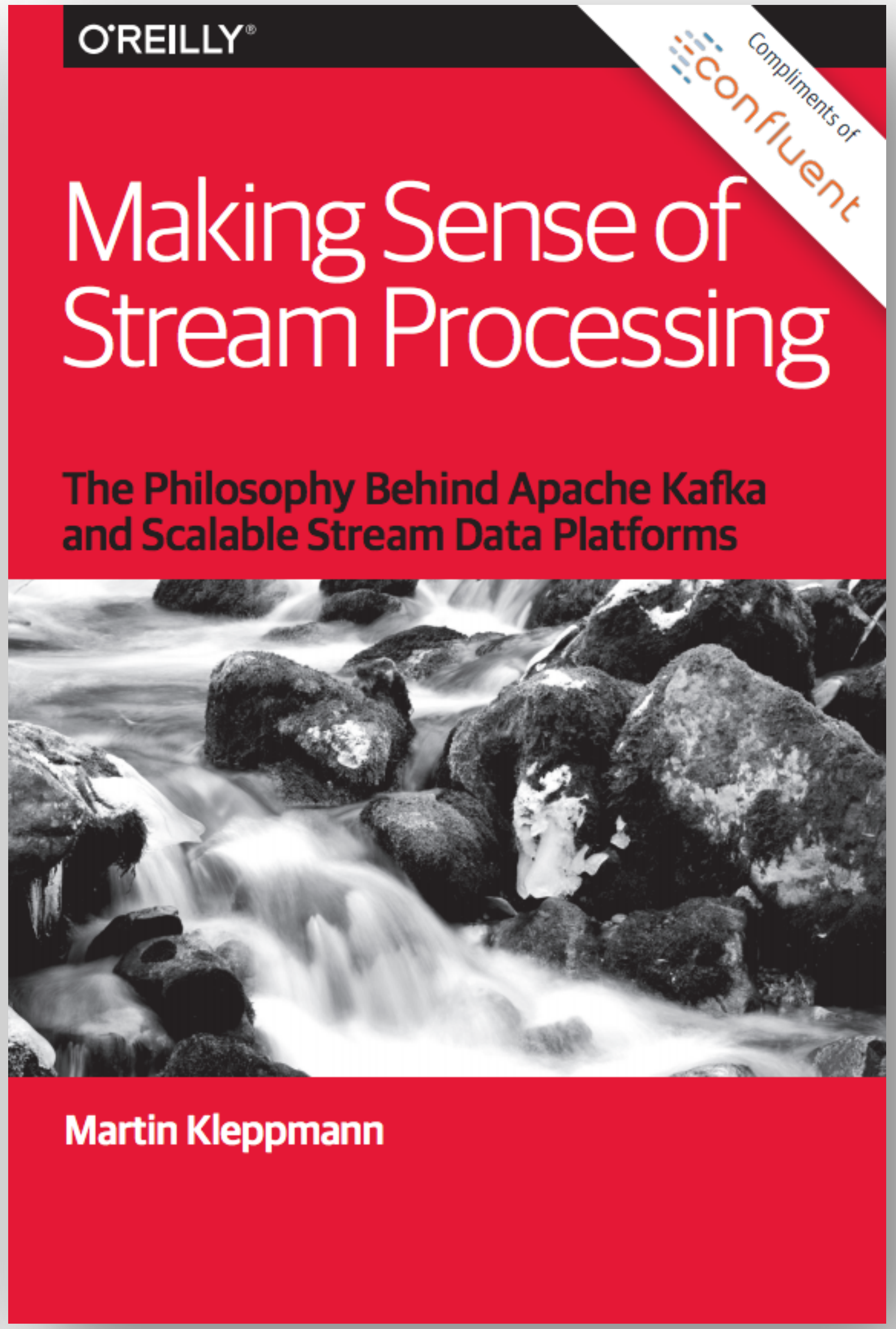
Confluent Community Slack group

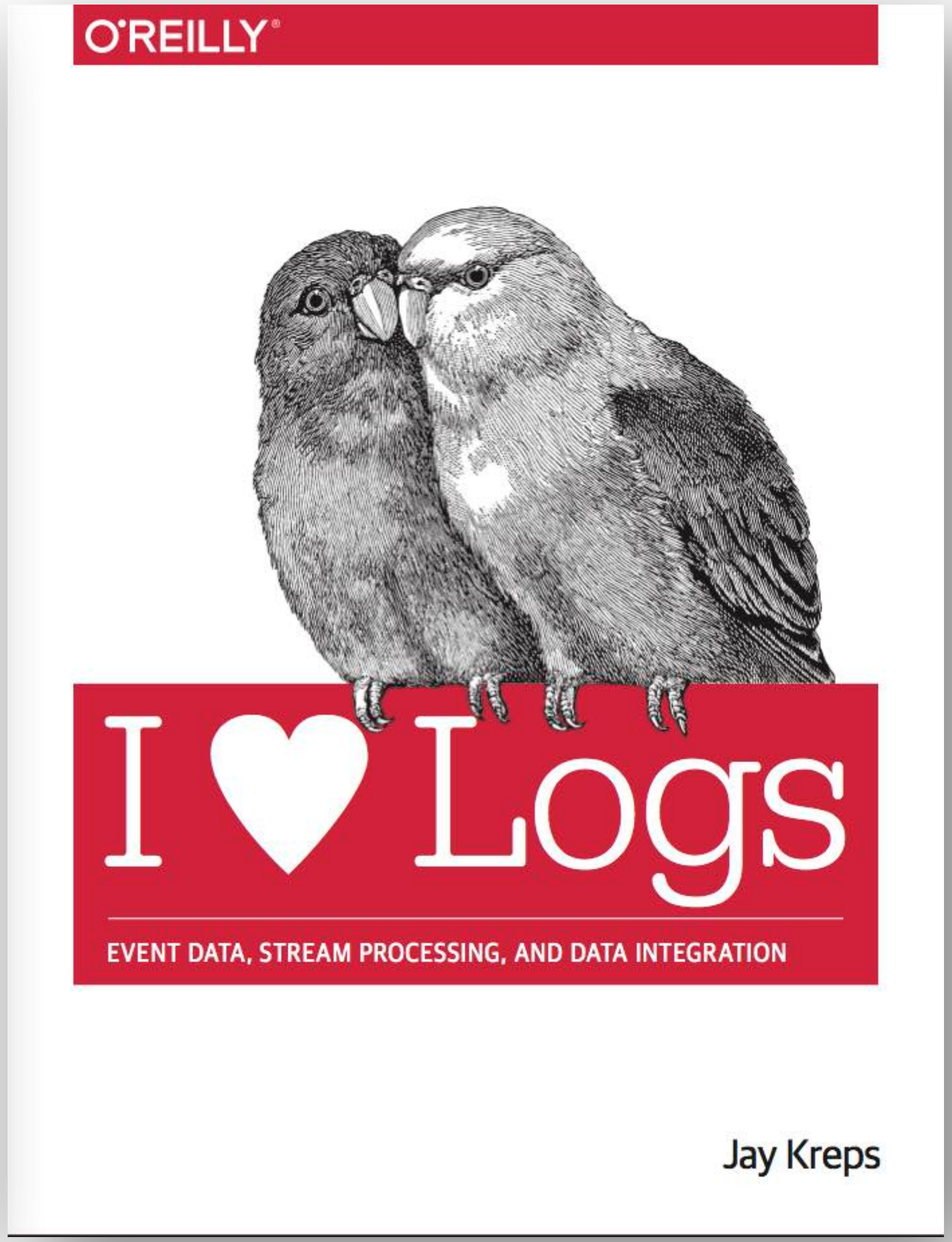
cnfl.io/slack

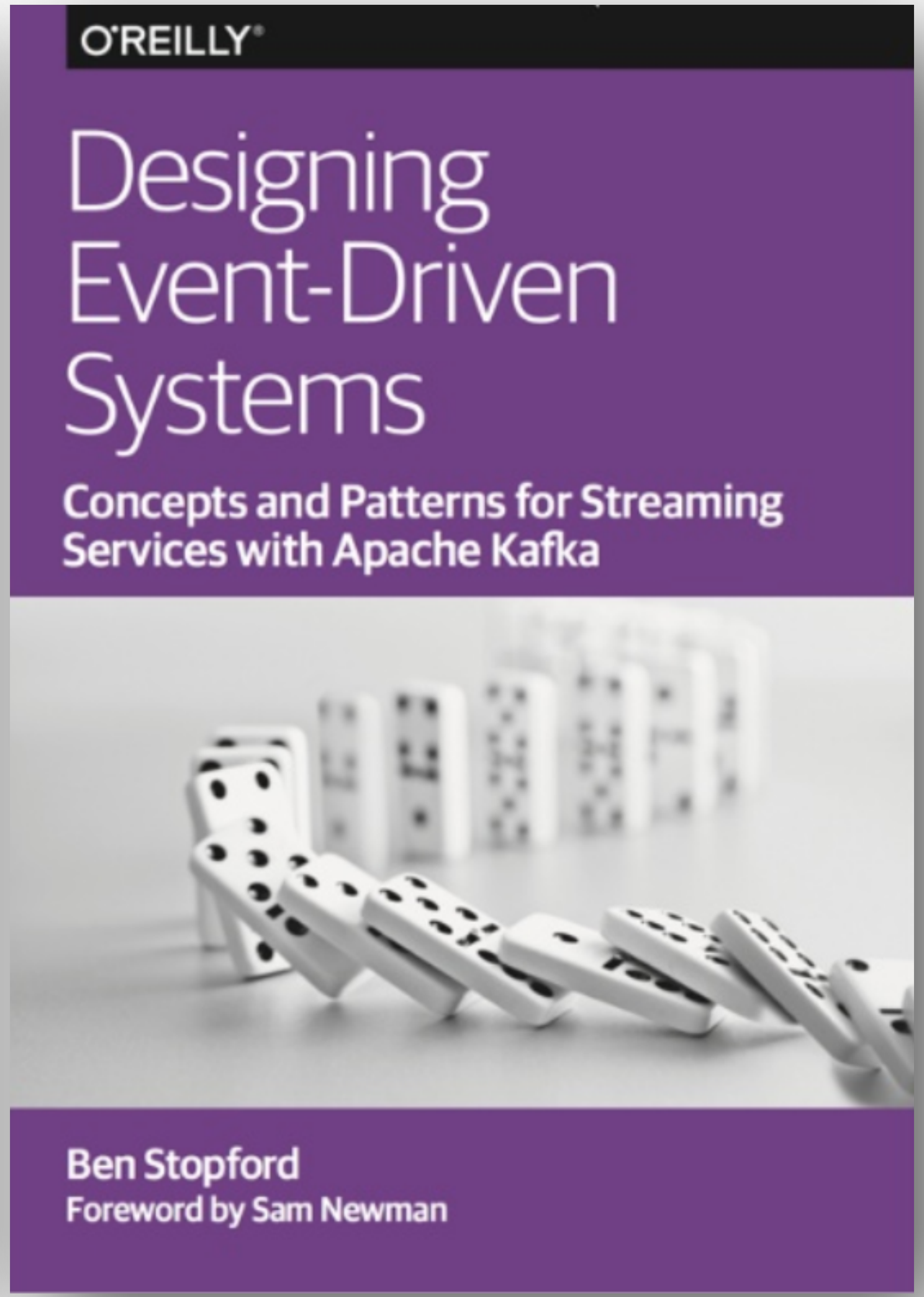
Free Books!

https://rmoff.dev/ldn-mar20









#EOF

Related Talks

- The Changing Face of ETL:
  Event-Driven Architectures for Data Engineers
  - 📖 Slides
  - 📽 Recording

- ATM Fraud detection with Kafka and ksqlDB
  - 📖 Slides
  - 👾 Code
  - 📽 Recording

- Embrace the Anarchy: Apache Kafka's Role in Modern Data Architectures
  - 📖 Slides
  - 📽 Recording
- Apache Kafka and ksqlDB in Action : Let's Build a Streaming Data Pipeline!
  - 📖 Slides
  - 👾 Code
  - 📽 Recording

- No More Silos: Integrating Databases and Apache Kafka
  - 📖 Slides
  - 👾 Code (MySQL)
  - 👾 Code (Oracle)
  - 📽 Recording