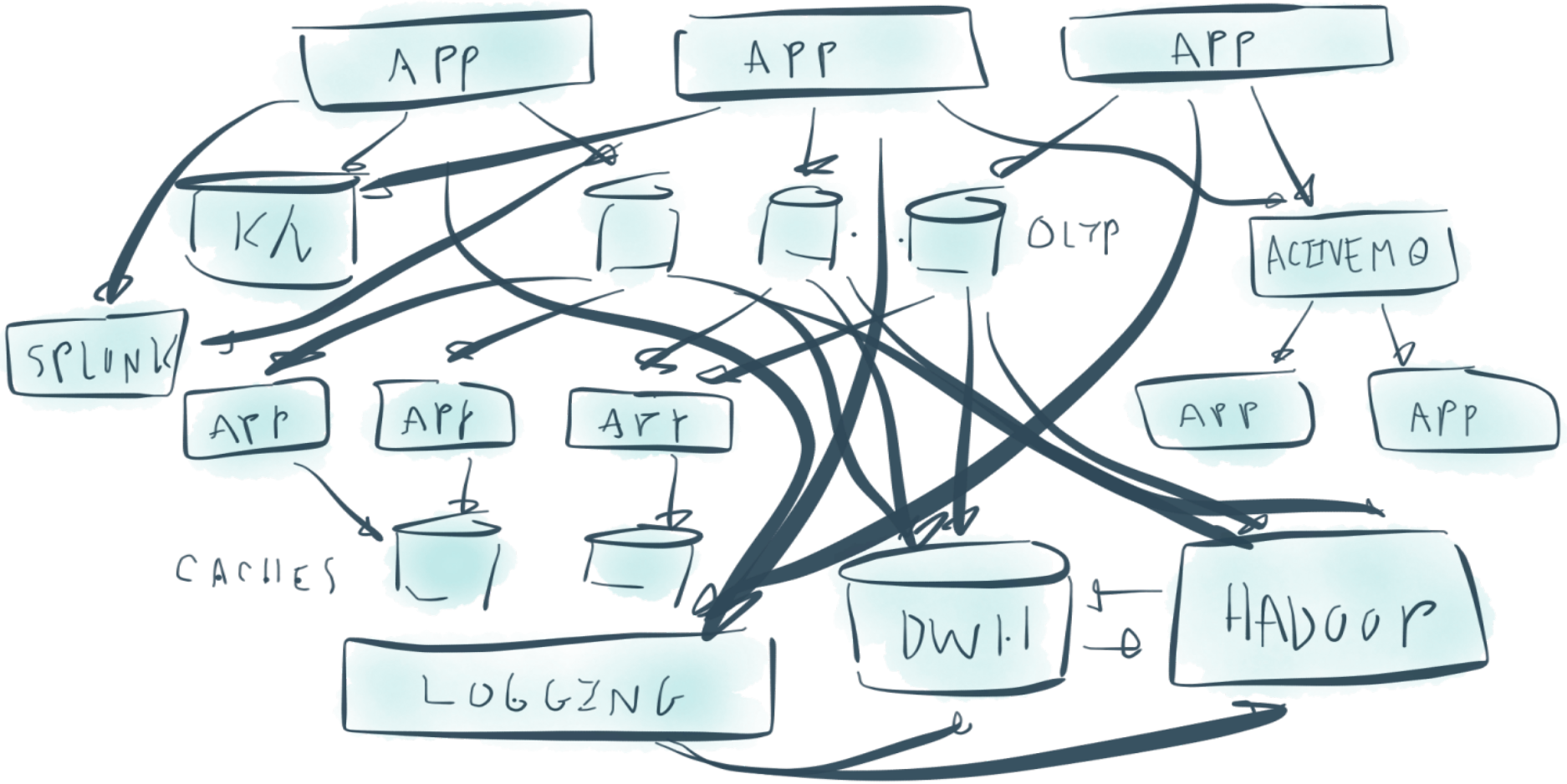
Real-time Data Integration at Scale with Kafka Connect

Robin Moffatt

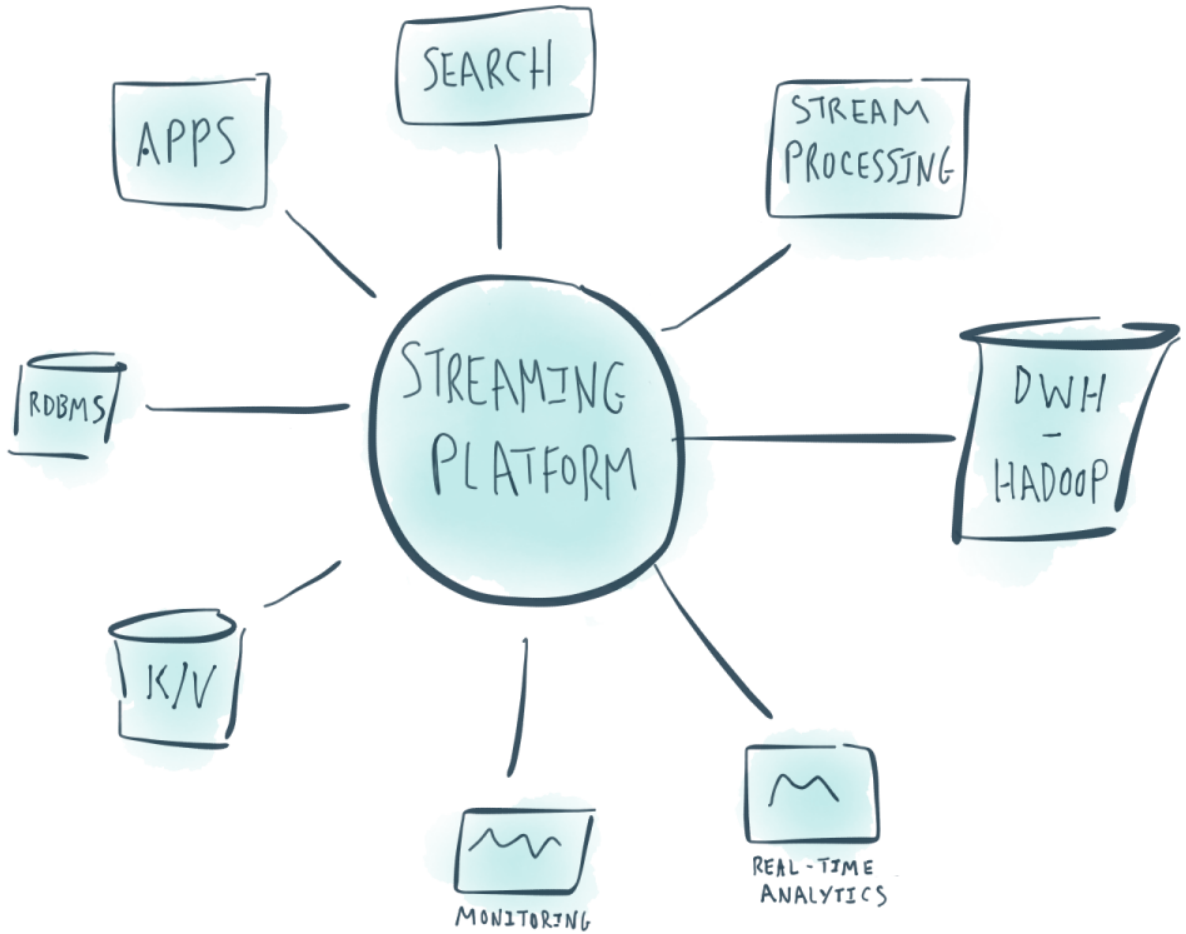
Partner Technology Evangelist, EMEA @ Confluent

@rmoff robin@confluent.io

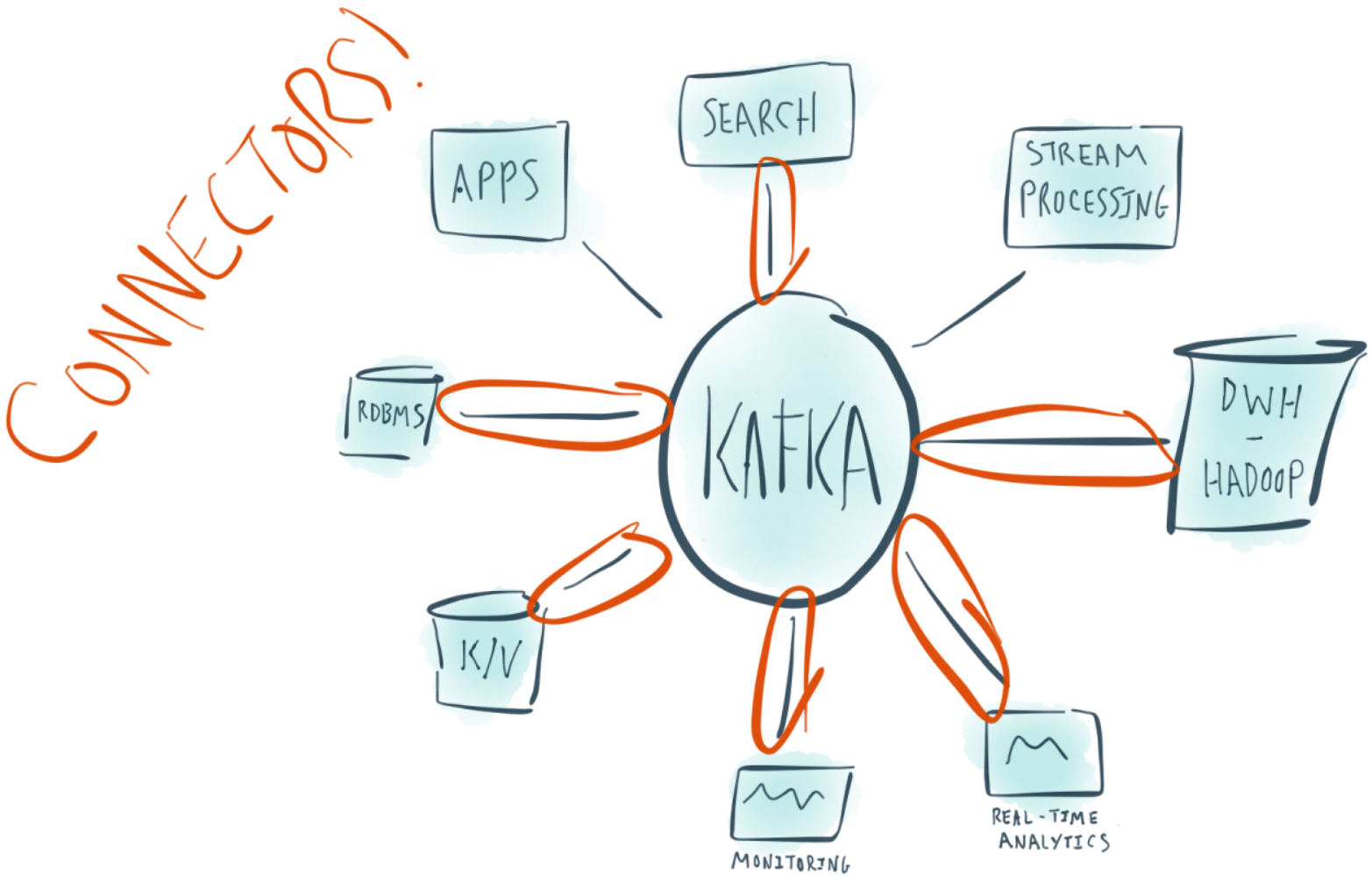
A GIANT MESS


APP
APP
APP
K/V
OLTP
ACTIVEMQ
SPLUNK
APP
APP
APP
APP
APP
CACHES
DW1.1
HADOOP
LOGGING

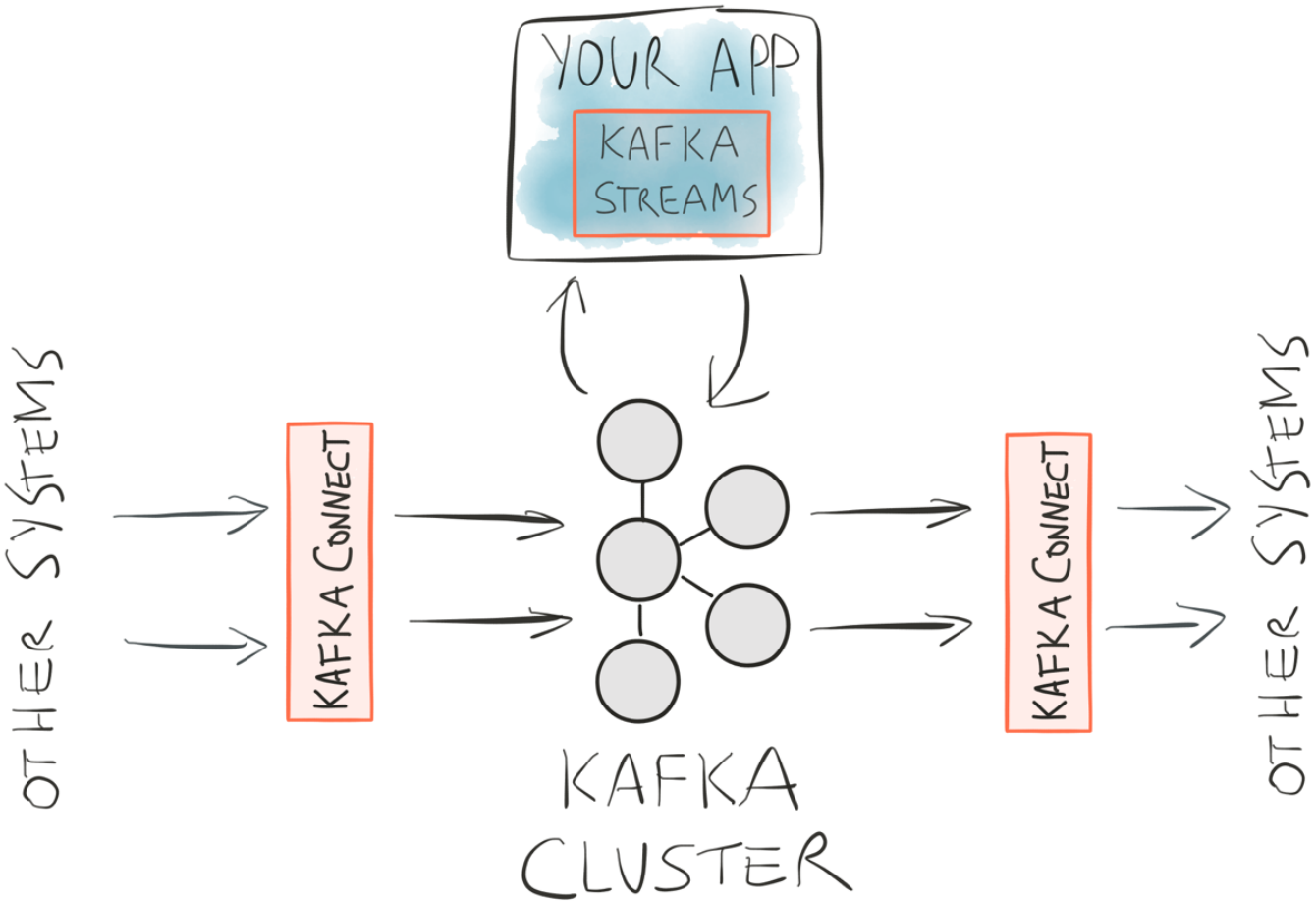
STREAMING PLATFORM



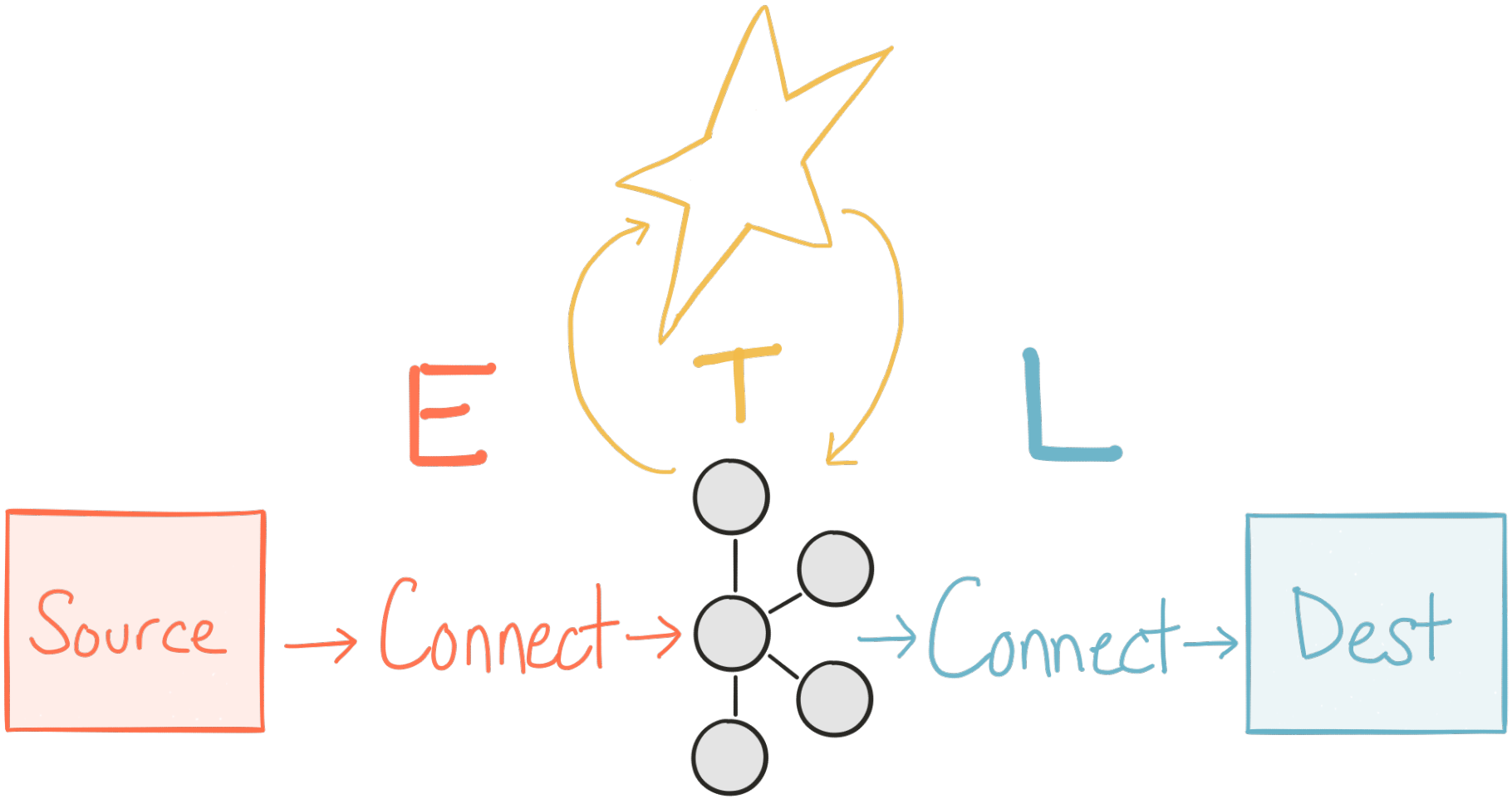
STREAMING PLATFORM

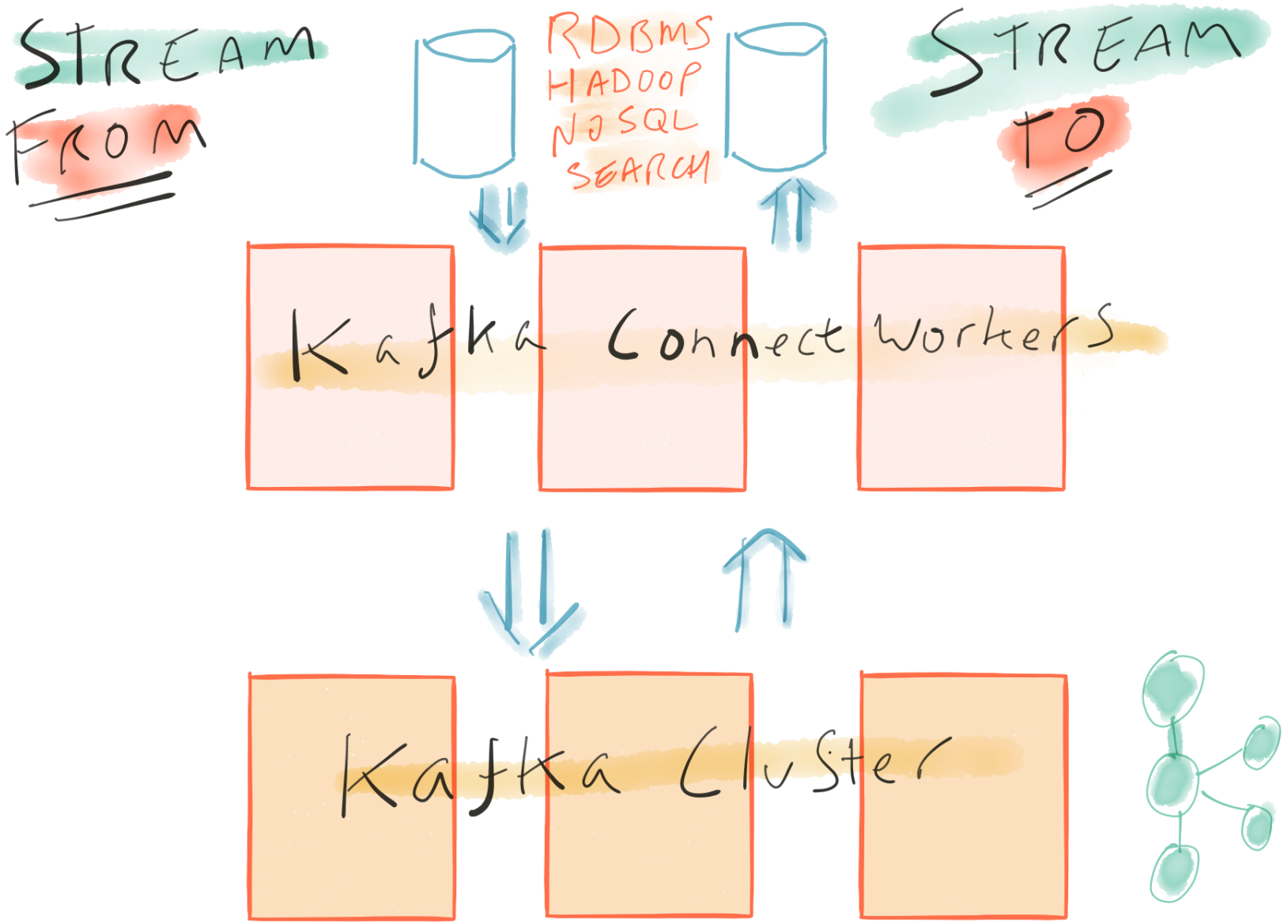


Kafka Connect in the Apache Kafka ecosystem



Kafka Connect : Separation of Concerns

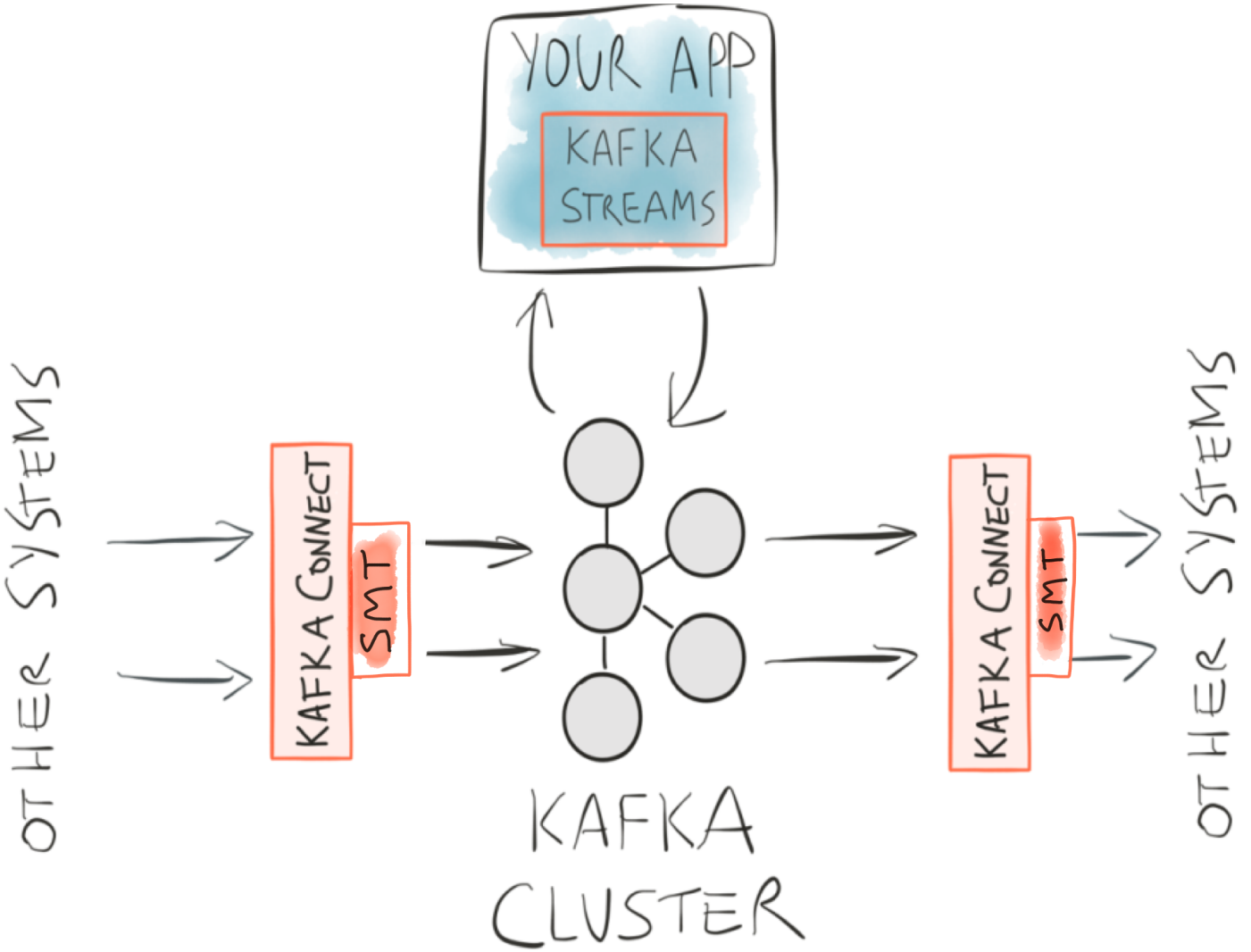




Build Data Pipelines

Single Message Transform (SMT) -- Extract, **TRANSFORM**, Load…

- Modify events ***before storing*** in Kafka:
  - Mask/drop sensitive information
  - Set partitioning key
  - Store lineage
- Modify events ***going out*** of Kafka:
  - Route high priority events to faster data stores
  - Direct events to different ElasticSearch indexes
  - Cast data types to match destinatior



Kafka Connect API Library of Connectors

Databases

Datastore/File Store

Analytics

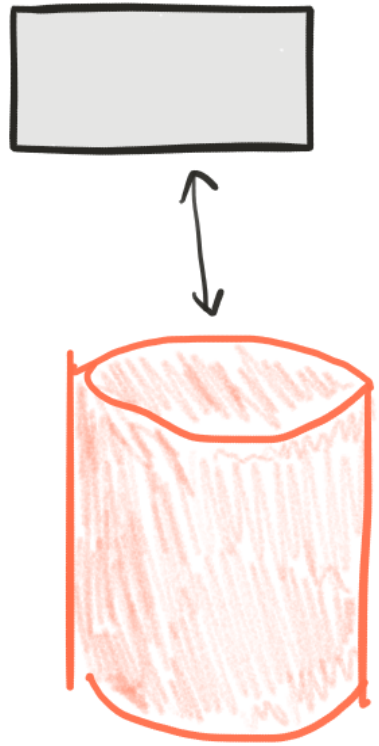
Applications / Other

confluent

https://www.confluent.io/product/connectors/

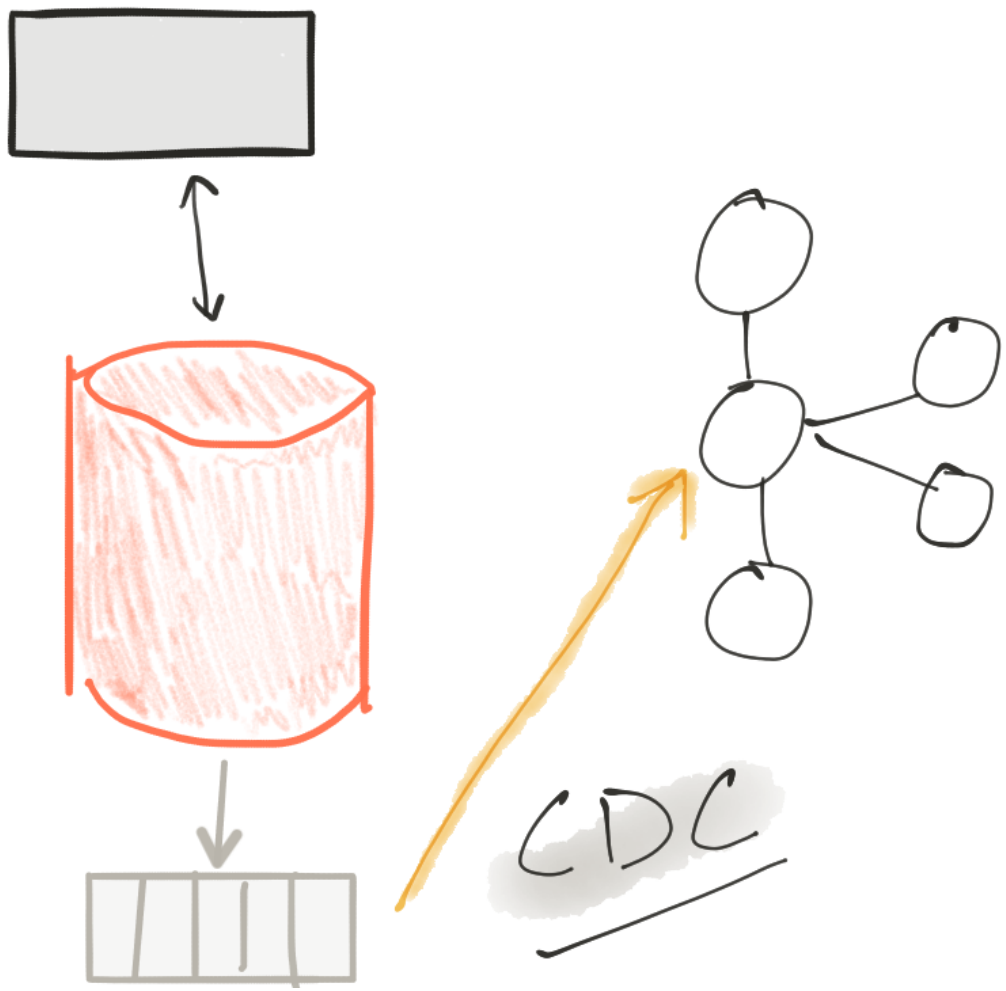
Streaming Application Data to Kafka

- Applications are rich source of events
- Modifying applications is not always possible or desirable
  - And what if the data gets changed within the database or by other apps?
- JDBC is one option for extracting data

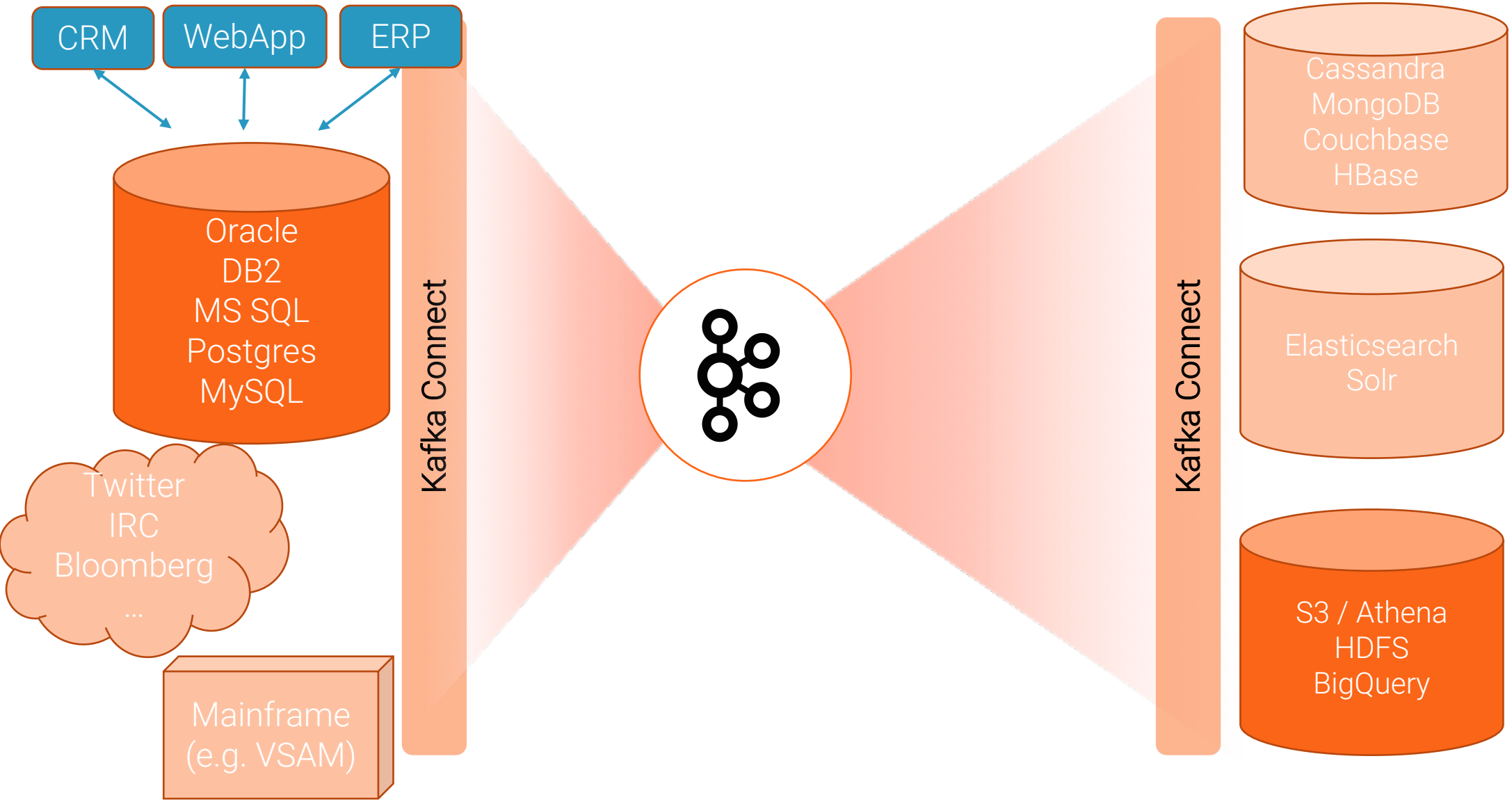


Liberate Application Data into Kafka with CDC

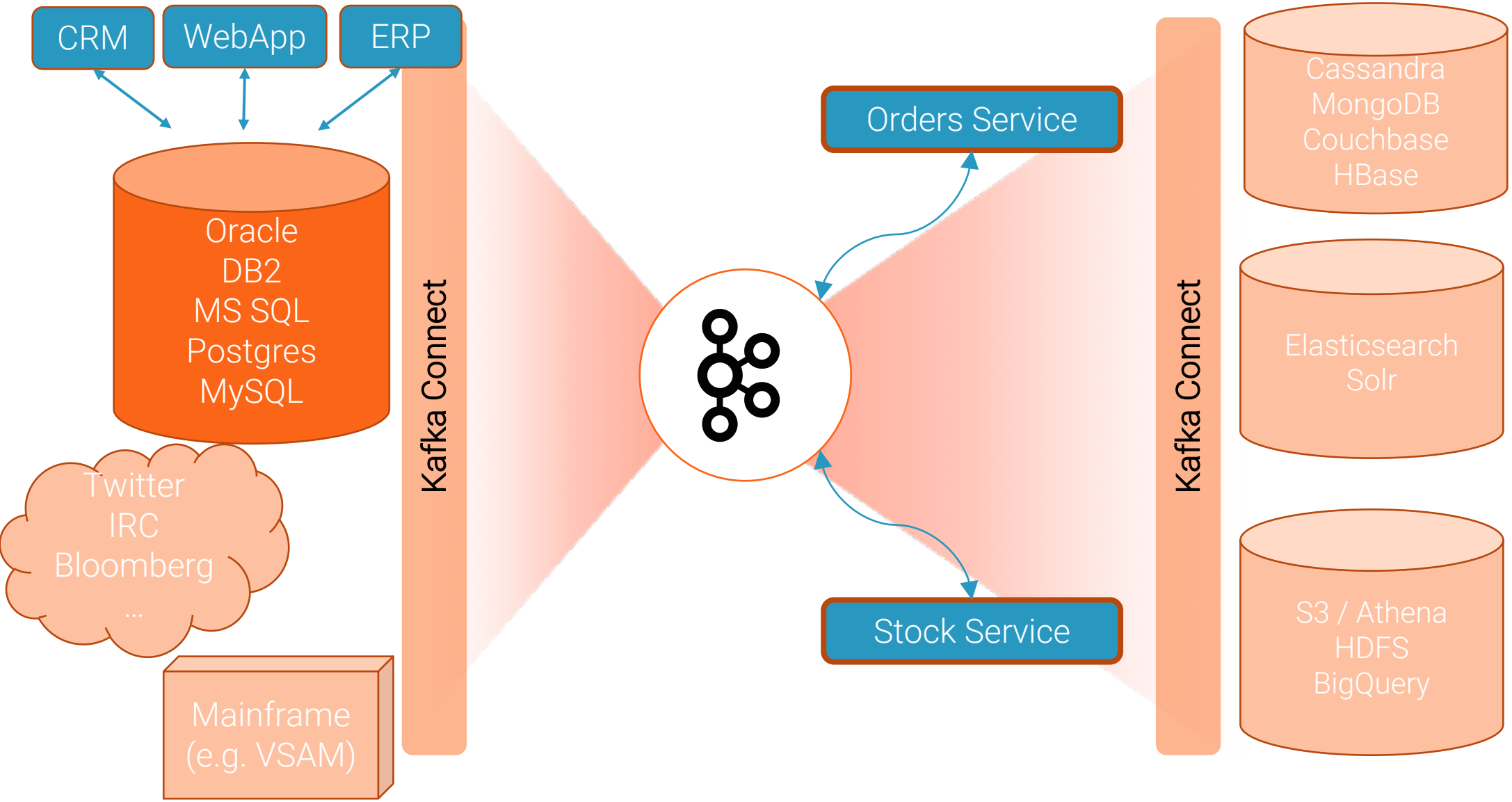
- Relational databases use transaction logs to ensure Durability of data
- Change-Data-Capture (CDC) mines the log to get raw events from the database
- CDC tools that integrate with Kafka Connect include:
  - Debezium
  - DBVisit
  - GoldenGate
  - Attunity
  - + more



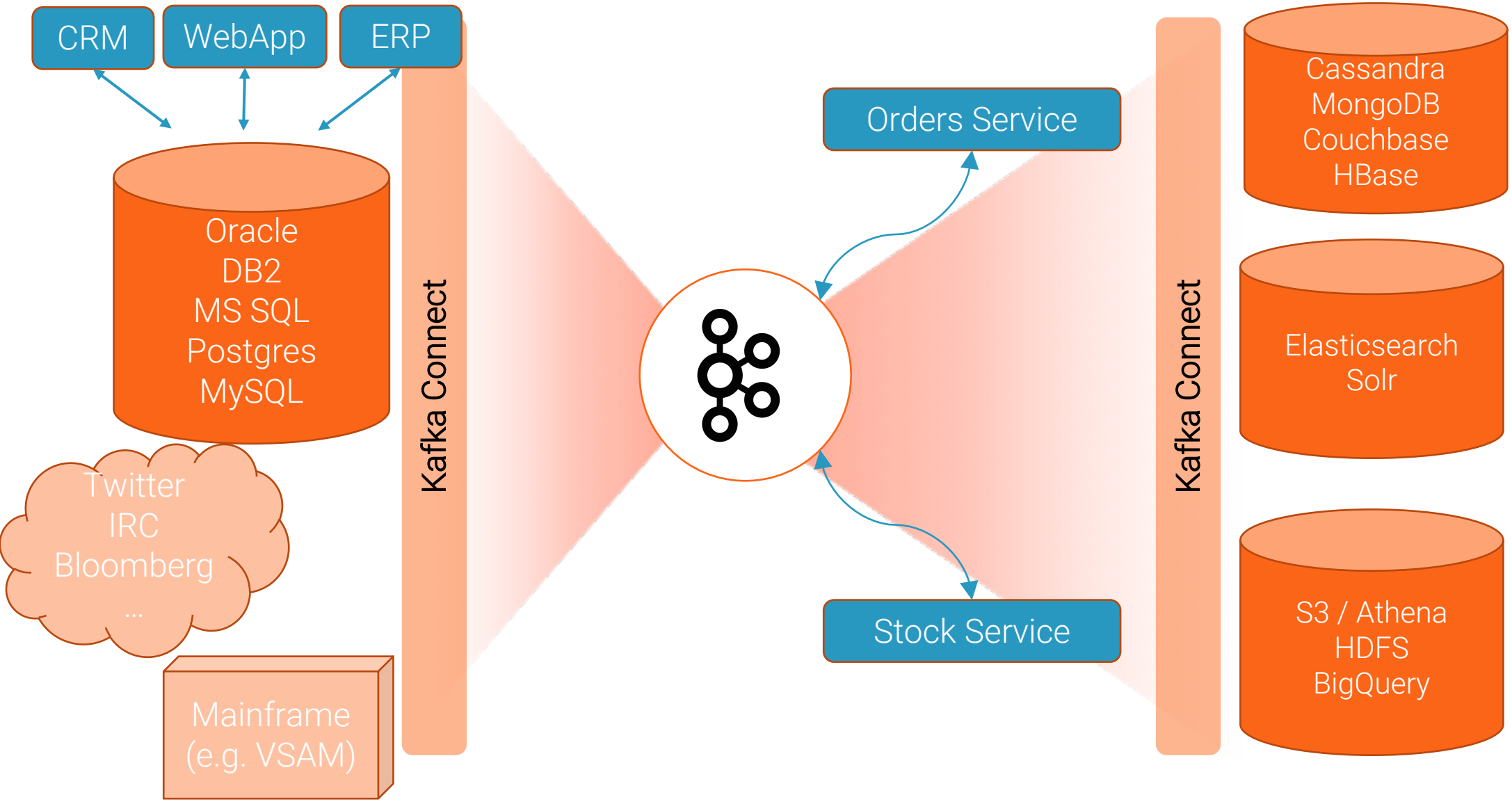
Common Patterns – Data Integration into Data Lake for batch analytics



Common Patterns – Event-Driven microservices



Common Patterns – Event-Driven microservices & audit/search/storage

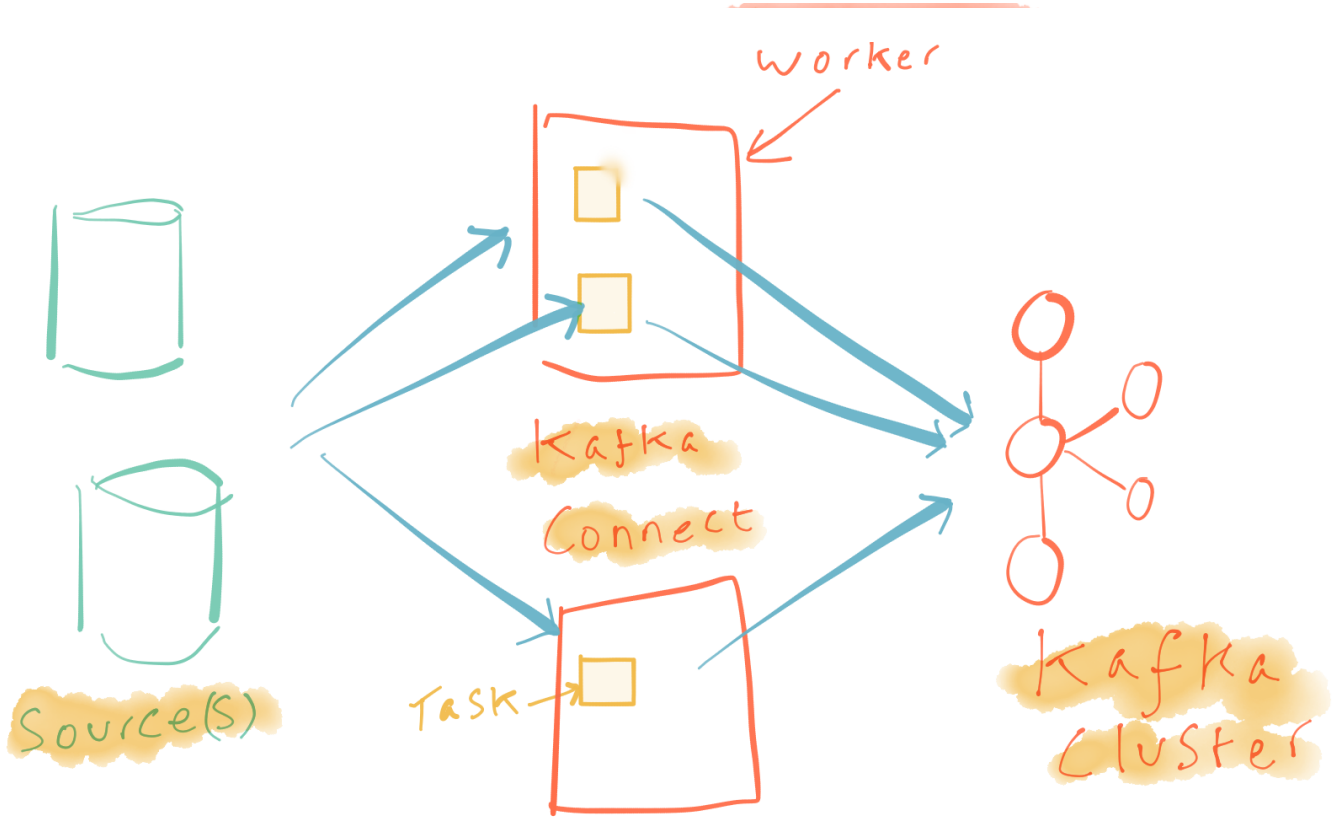


The Numerous Benefits of Kafka Connect

- Restart capabilities (offset management)
- Distributed workers
  - Parallelism (for throughput)
  - Load balancing
  - Fault tolerance
- Schema preservation
- Data serialisation
- Centralised management and configuration

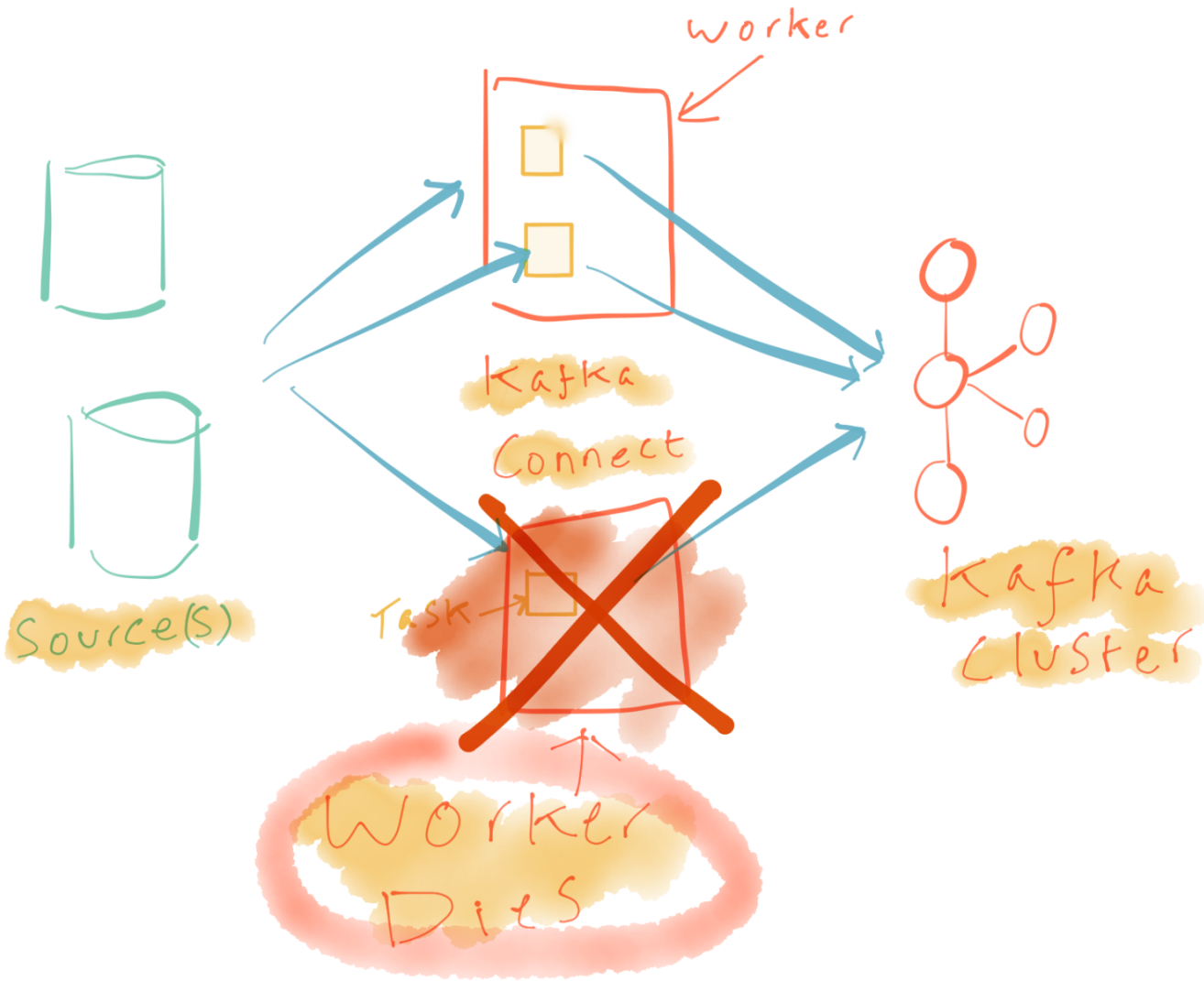
Kafka Connect – under the covers

- Each Kafka Connect node is a **worker**
- Each worker executes one or more **tasks**
- Tasks do the actual work of pulling data from sources / landing it to sinks
- Kafka Connect manages the distribution and execution of tasks
- Parallelism, fault-tolerance, load balancing all **handled automatically**



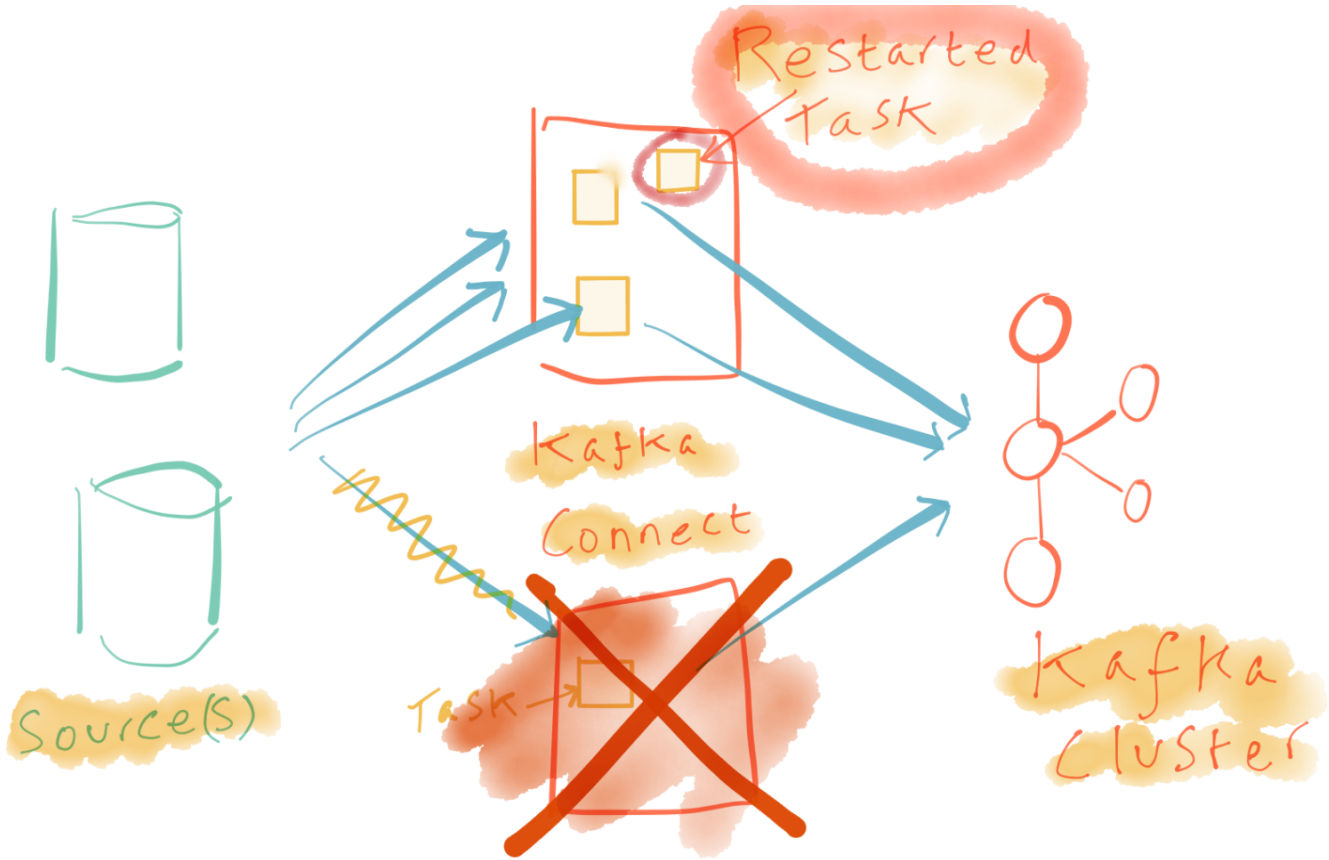
Kafka Connect – under the covers

- Each Kafka Connect node is a **worker**
- Each worker executes one or more **tasks**
- Tasks do the actual work of pulling data from sources / landing it to sinks
- Kafka Connect manages the distribution and execution of tasks
- Parallelism, fault-tolerance, load balancing all **handled automatically**



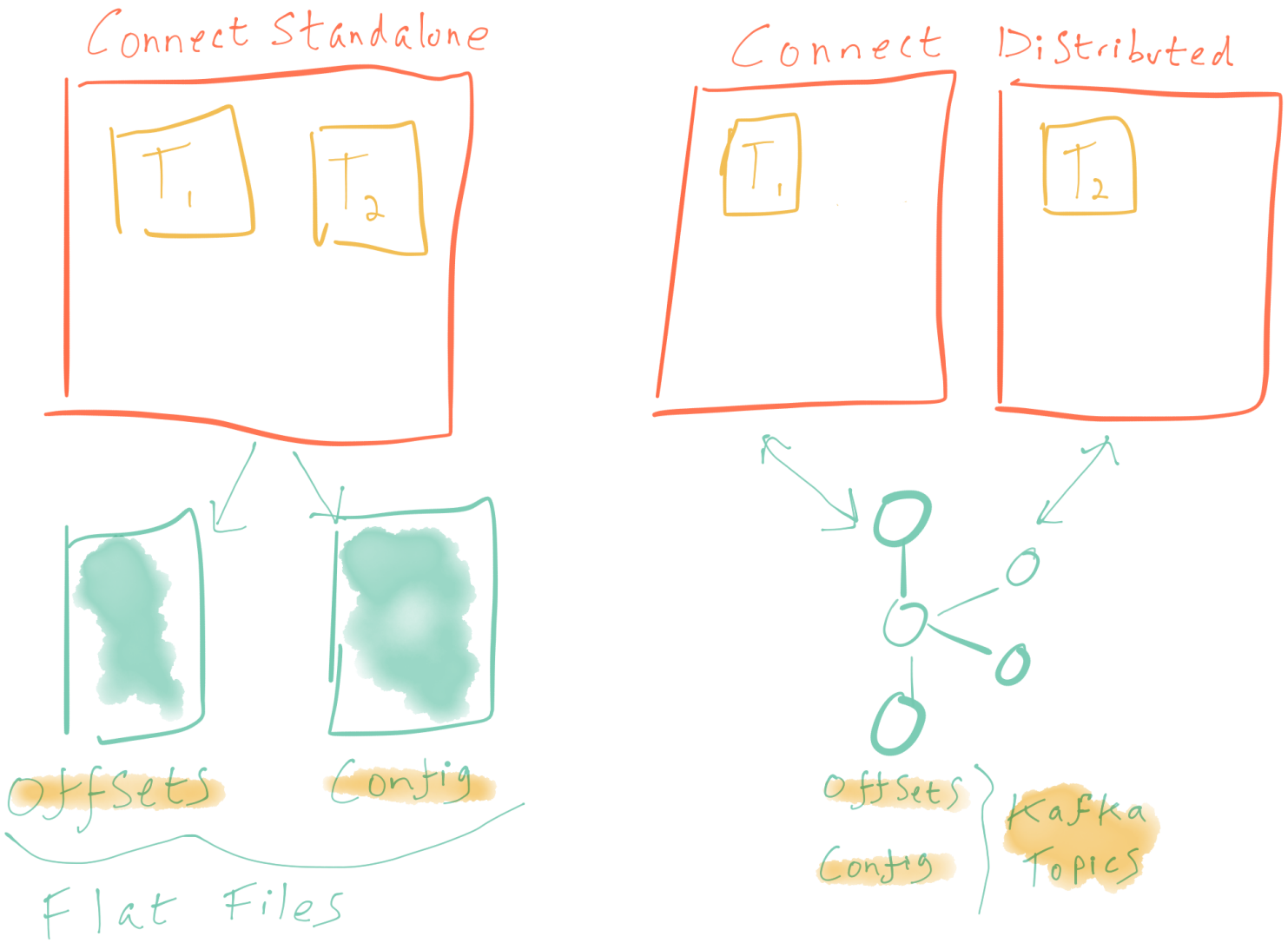
Kafka Connect – under the covers

- Each Kafka Connect node is a **worker**
- Each worker executes one or more **tasks**
- Tasks do the actual work of pulling data from sources / landing it to sinks
- Kafka Connect manages the distribution and execution of tasks
- Parallelism, fault-tolerance, load balancing all **handled automatically**



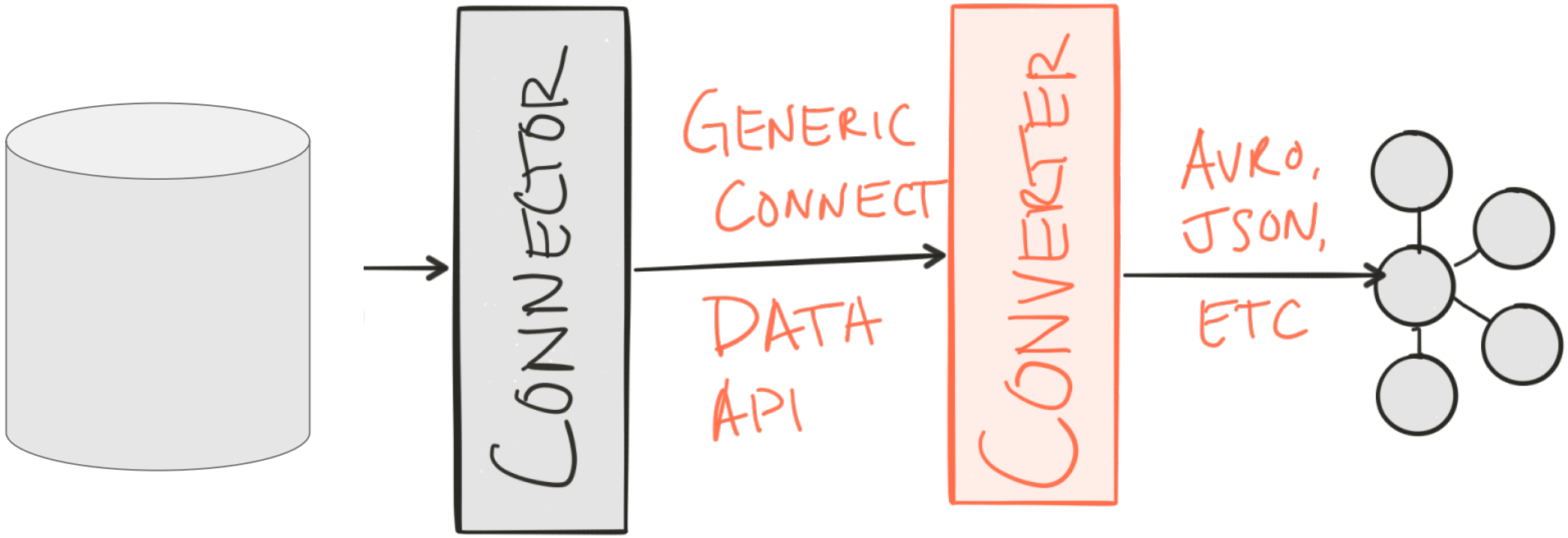
Kafka Connect – Standalone vs Distributed

- Kafka Connect has two modes **standalone** or **distributed**
- Distributed - Scaleout & fault tolerance easy – just add more workers
  - Can run on one node!
- Standalone - Useful for where data source is machine-specific (e.g. single-node log files)



Kafka Connect - Converters

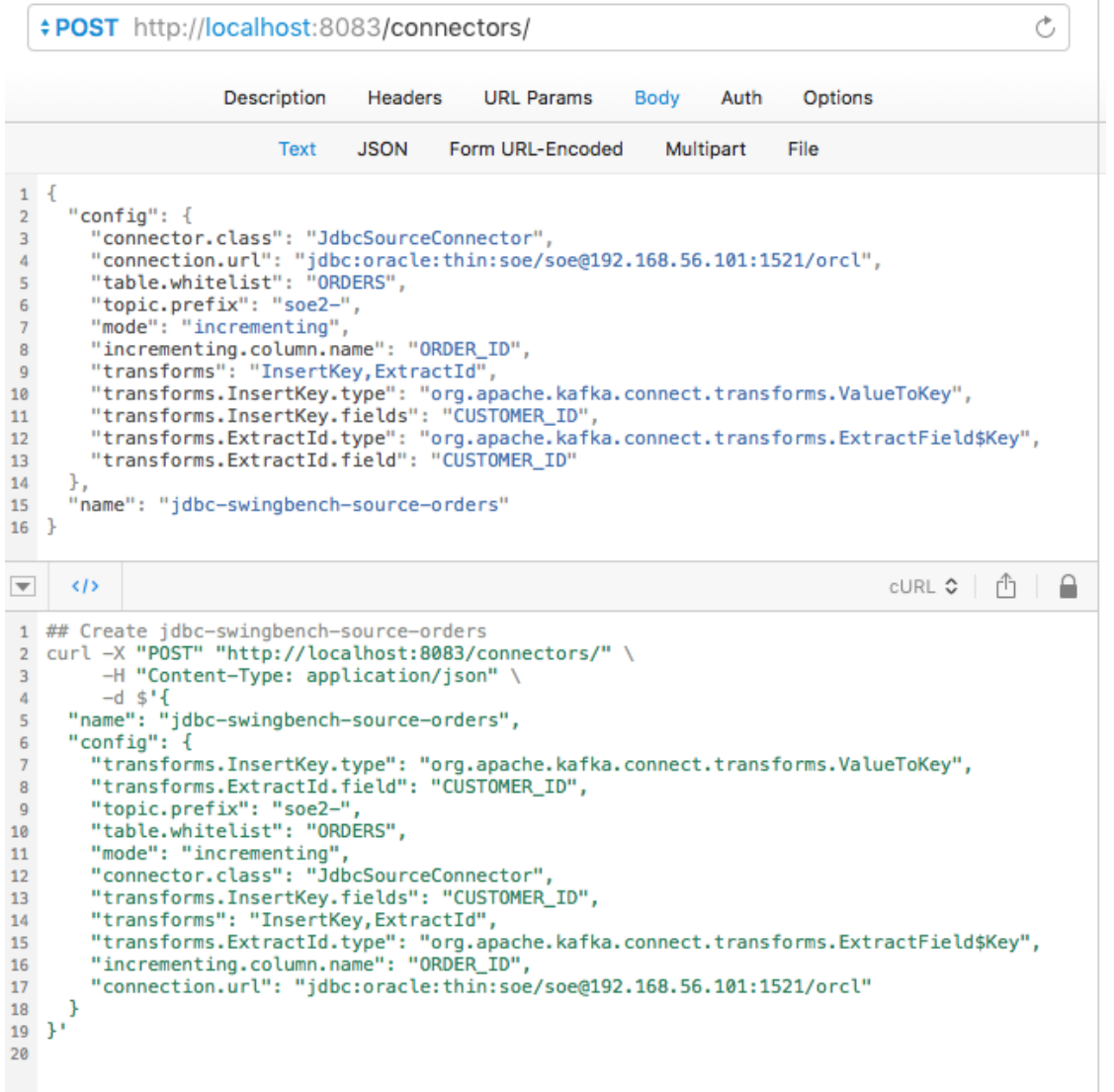
- Data from source system is in its own format (e.g. RecordSet from JDBC)
- Kafka Connect's Converters provide reusable functionality to serialise data into JSON or Avro
  - The Confluent **Schema Registry** is used to stores schemas of ingested data



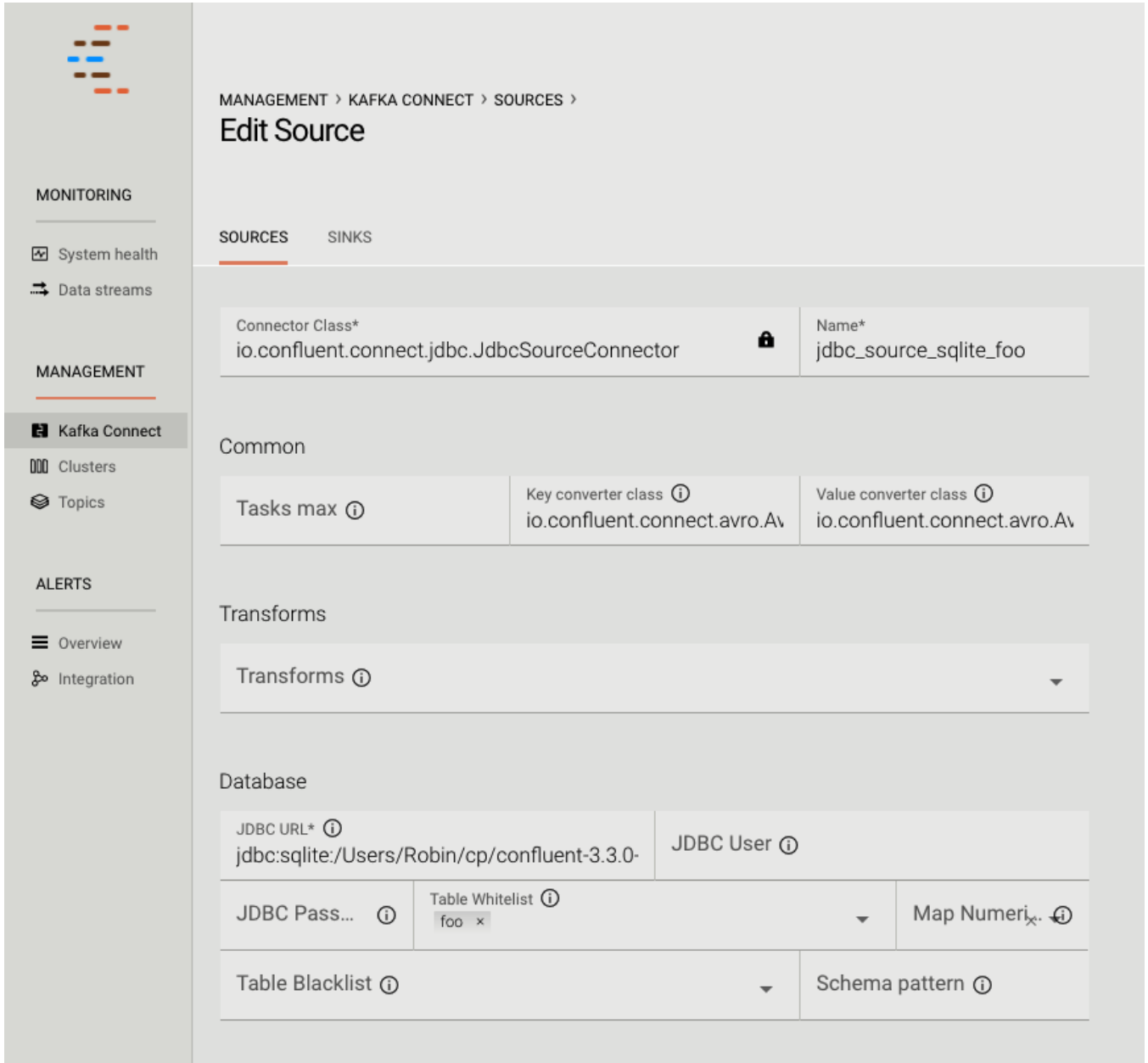
http://docs.confluent.io/current/connect/concepts.html#converters

Configuring Kafka Connect - REST API

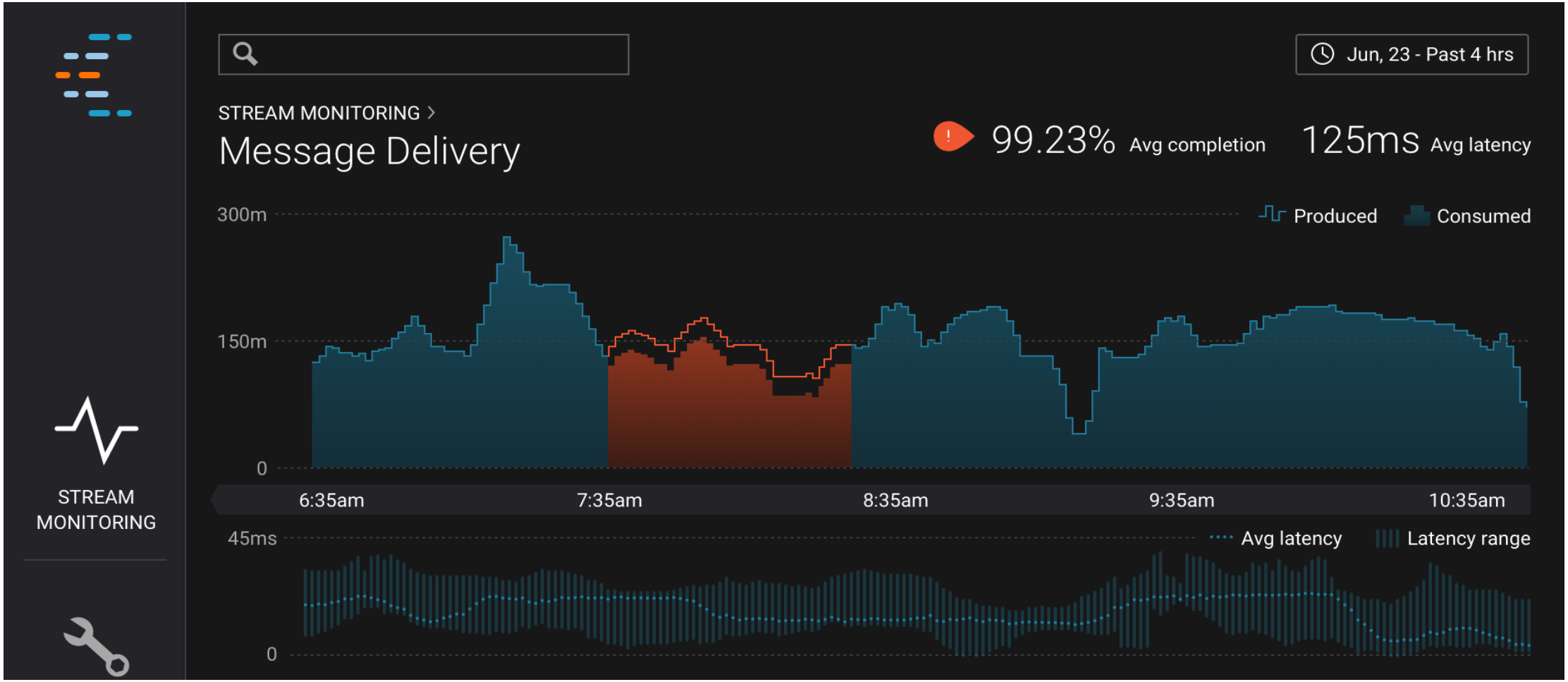
- Configure & control Kafka Connect through REST API
  - Validate connector configuration
  - Create connectors
  - List available plugins
  - Query connector & task state
  - Pause, resume, restart connectors + tasks
- Configuration is persisted through a Kafka topic
- Reference : http://docs.confluent.io/current/connect/restapi.html



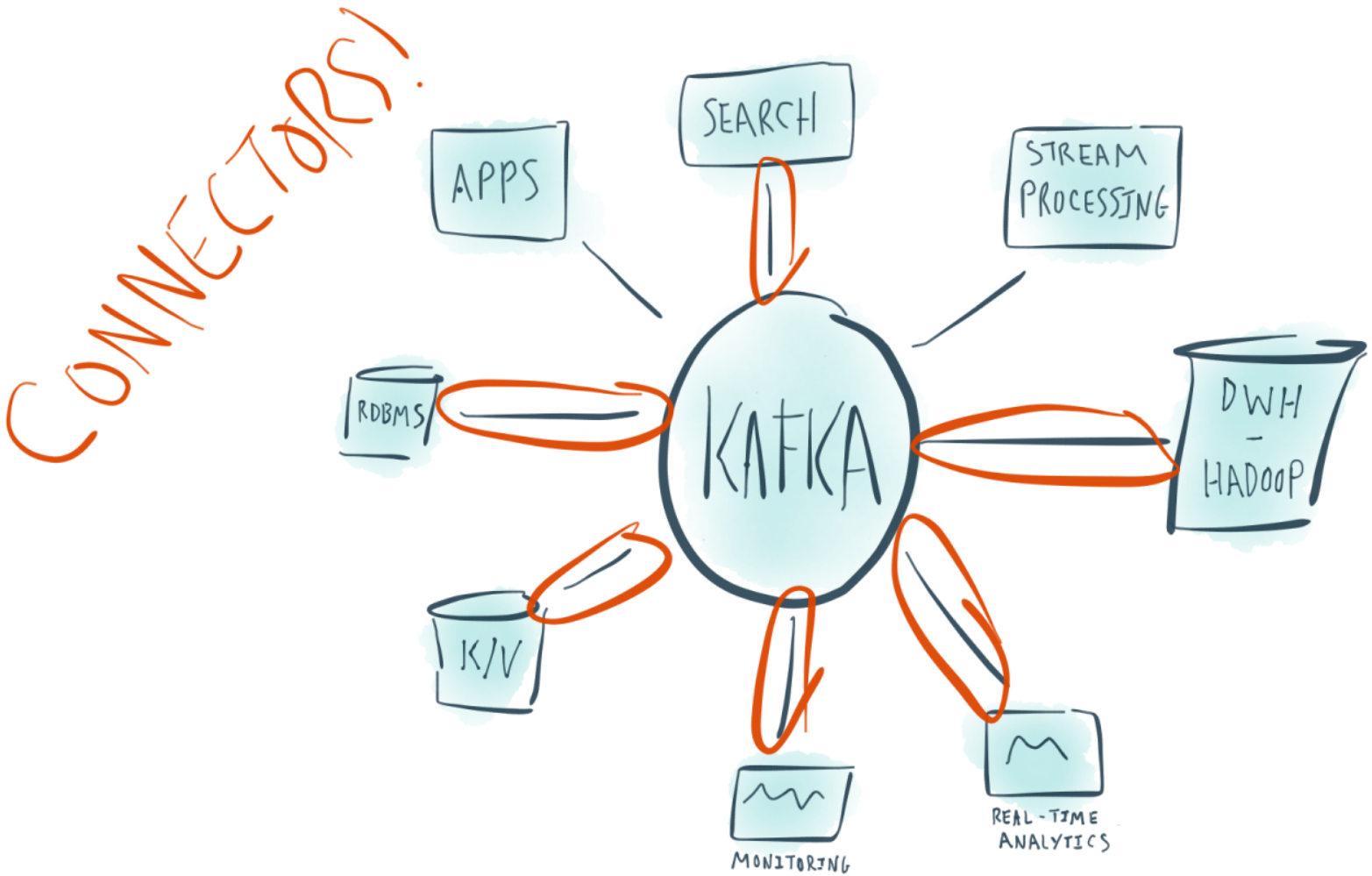
Configure Kafka Connect with Confluent Control Center



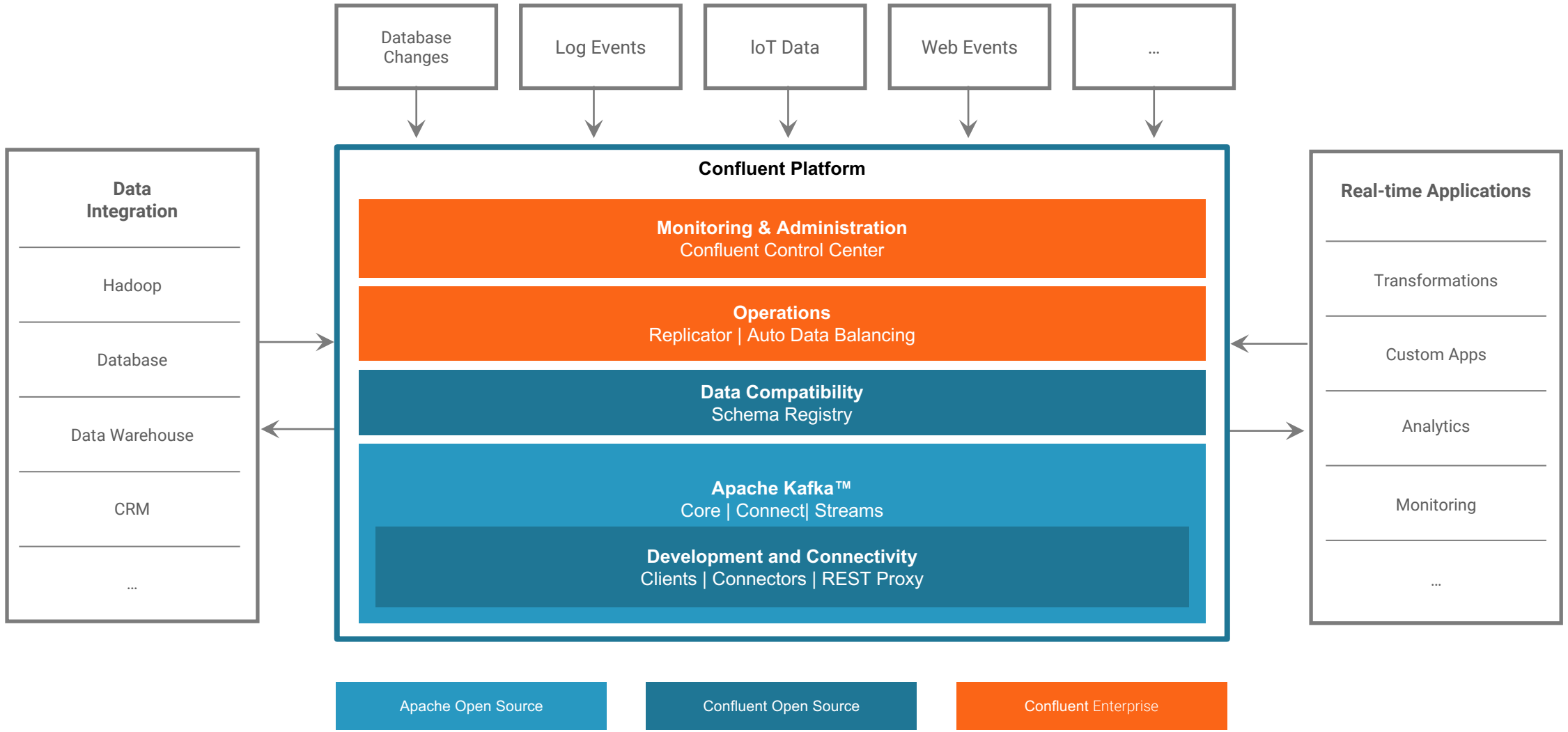
Monitor Your Data Pipeline from End to End with Confluent Control Center



STREAMING PLATFORM

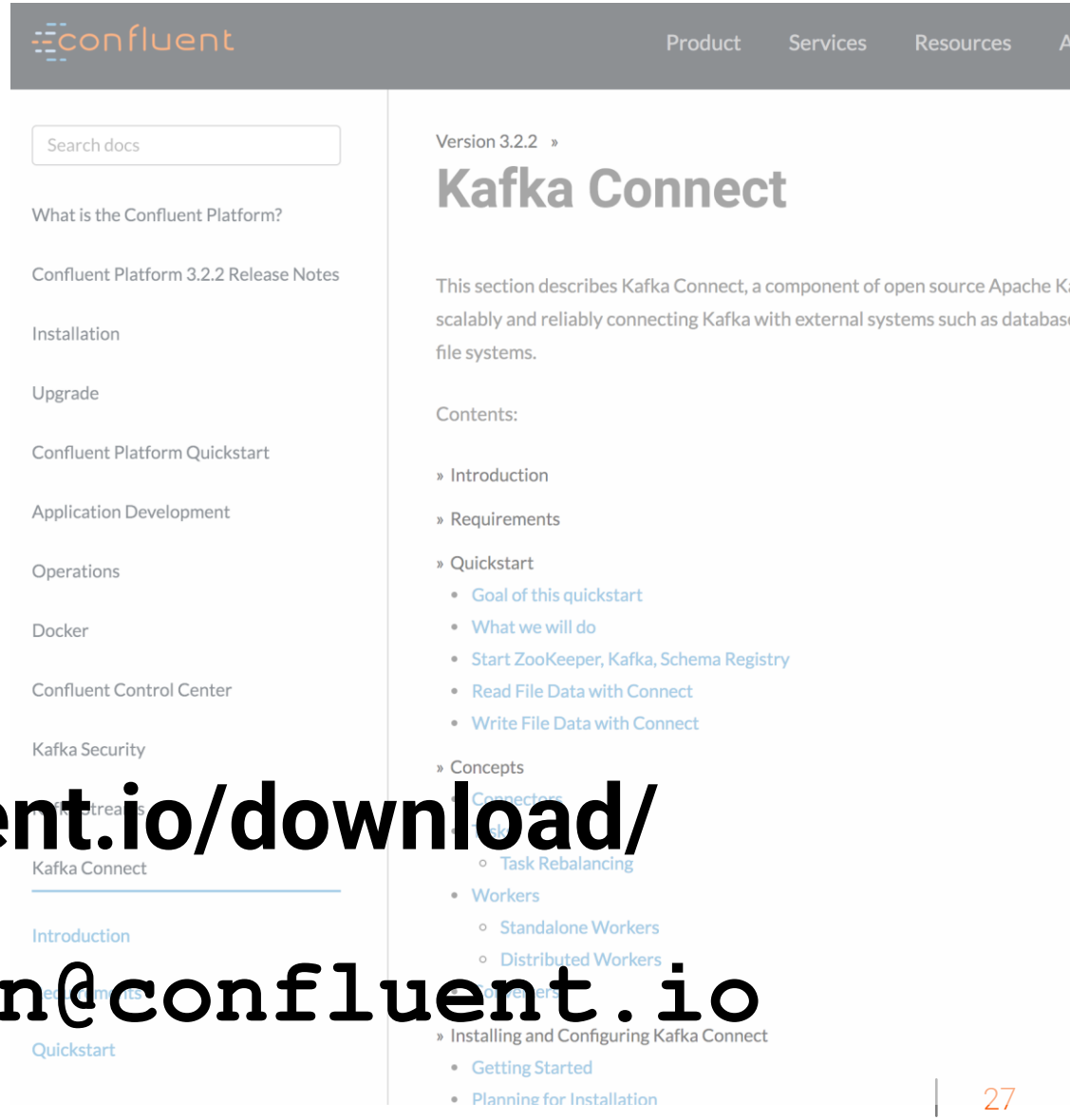


Confluent: a Streaming Platform based on Apache Kafka™



Kafka Connect – Getting Started

- Docs : http://docs.confluent.io/current/connect/
- Includes **Quickstart** and full Connect documentation including Architecture + Internals
- Official Confluent Platform Docker images available
  - http://docs.confluent.io/current/cp-docker-images/docs/quickstart.html#kafka-connect
- List of connectors
  - https://www.confluent.io/product/connectors/
  - Also search on github https://github.com/search?q=kafka-connect



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

@rmoff robin@confluent.io