BUILDING REALTIME DATA PIPELINES WITH KAFKA CONNECT AND SPARK STREAMING

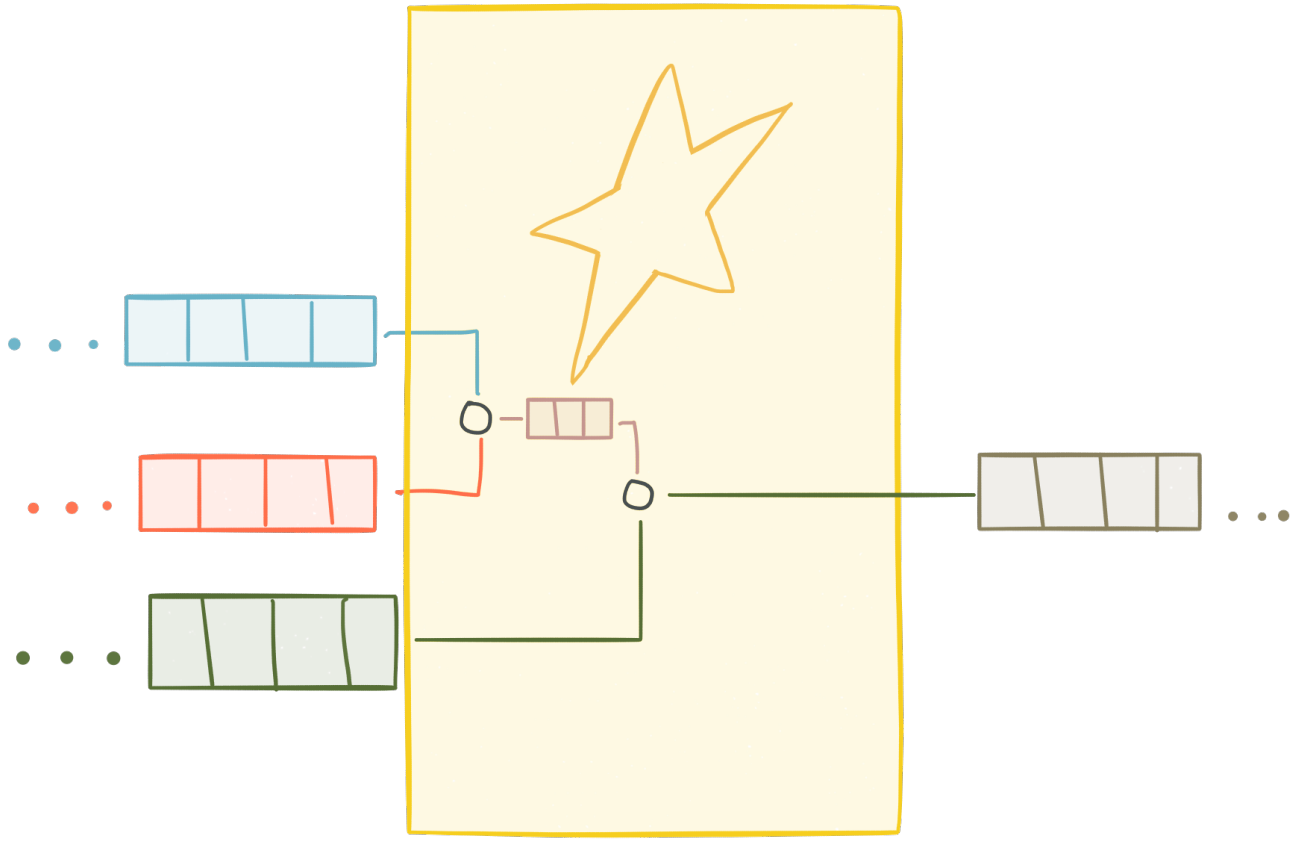
Ewen Cheslack-Postava
Confluent

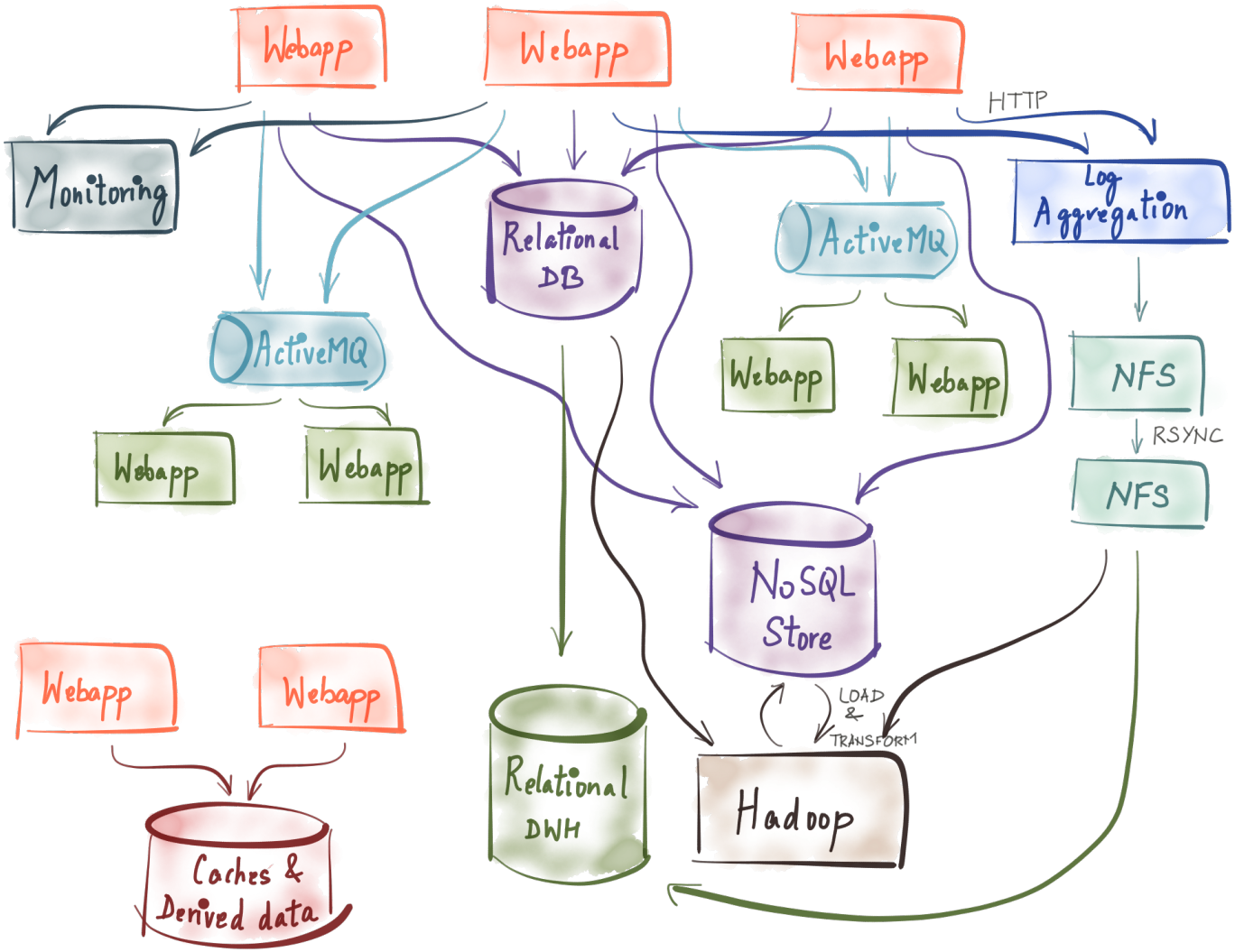
About Me

- Engineer @ Confluent
- Kafka Committer
- Kafka Connect Lead


monetdb
HIVE
VOLTDB
elasticsearch
ORACLE
DATABASE
APACHE
HBASE
OrientDB
KUDU
AEROSPIKE
Spark
OMNITURE
riak
cassandra
accumulo
Solr
NUODB
memsql
Couchbase
neo4j
RethinkDB
logstash
LOG
ZOO
mongoDB
Scalaris
Flink
ArangoDB
APACHE
GIRAPH
SPARK
SUMMIT
EAST 2017

Stream Processing





ANTI PATTERNS

1. One-off tools
2. Kitchen sink tools
3. Stream processing frameworks

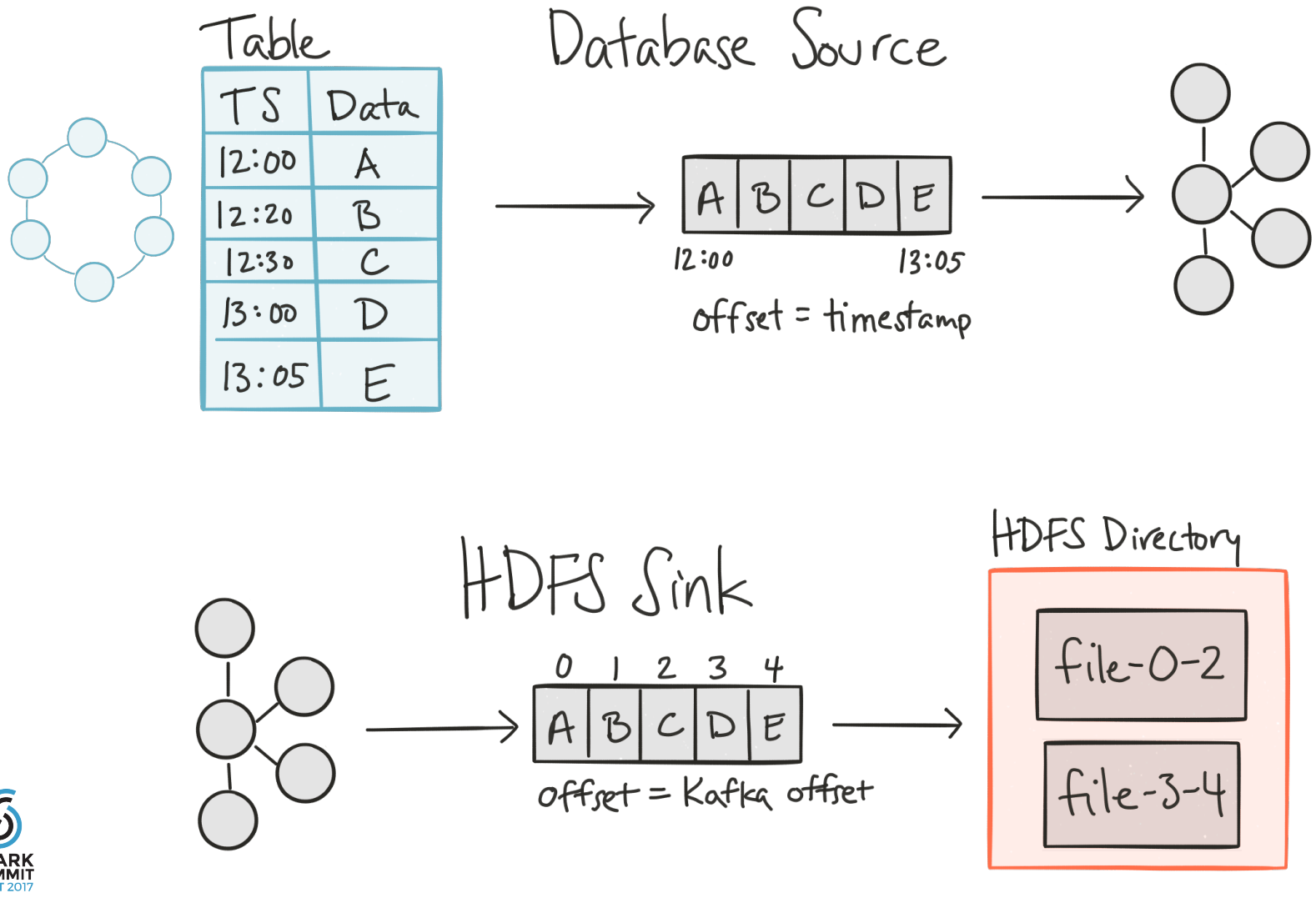

Ad-hoc
?
E,T,&L

Introducing Kafka Connect

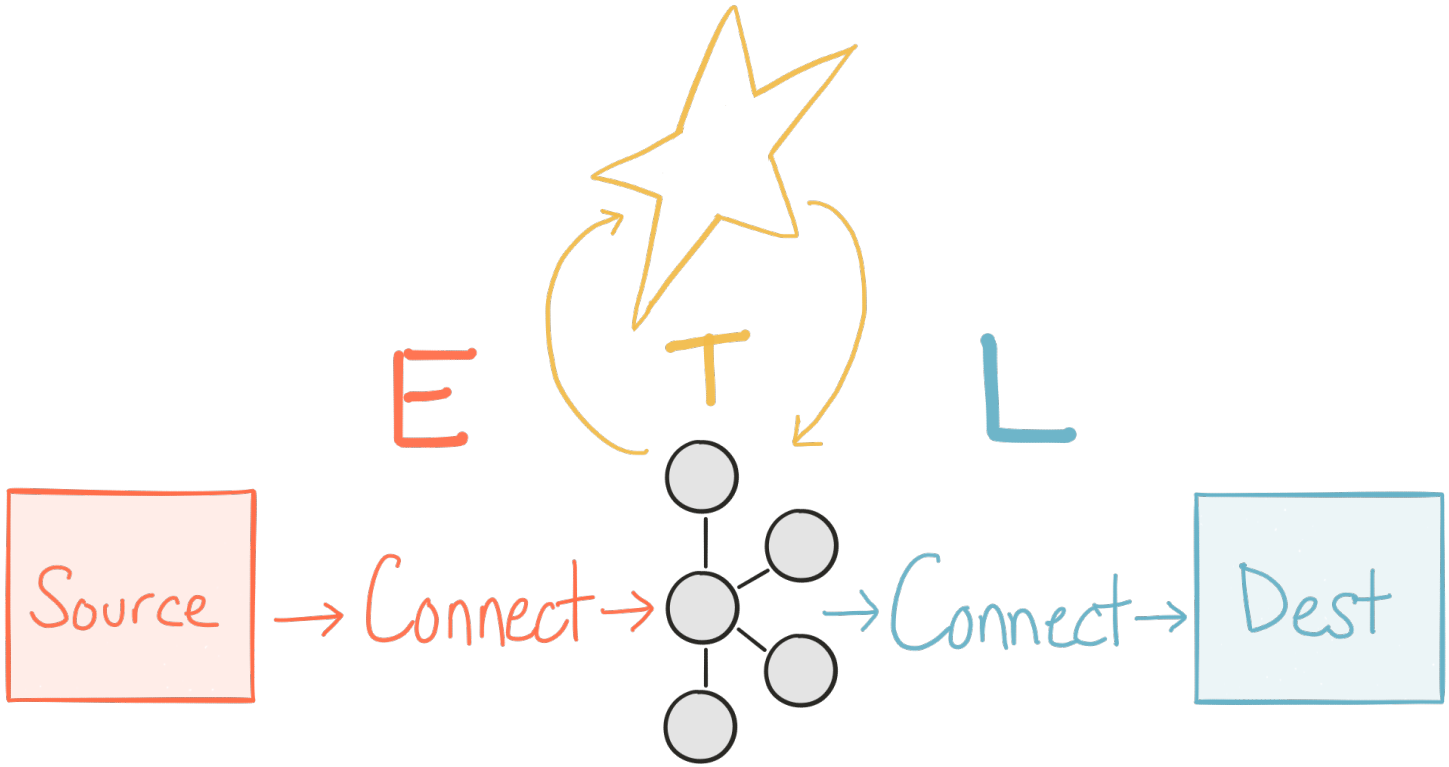
Large-scale streaming data import/export for Kafka

GOALS

1. Focus on copying
2. Batteries included
3. Standardize
4. Parallelism
5. Scale



Separation of Concerns

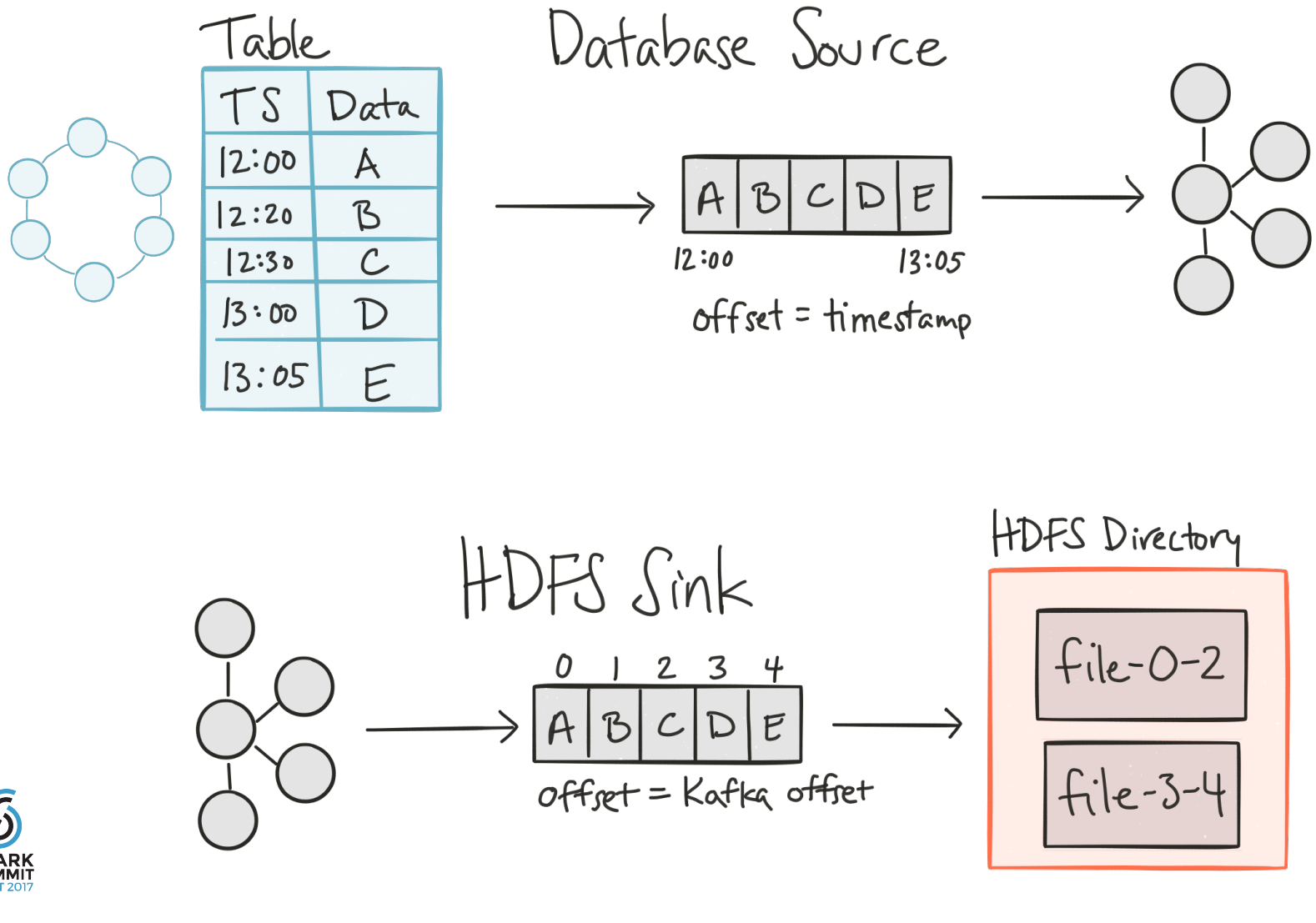


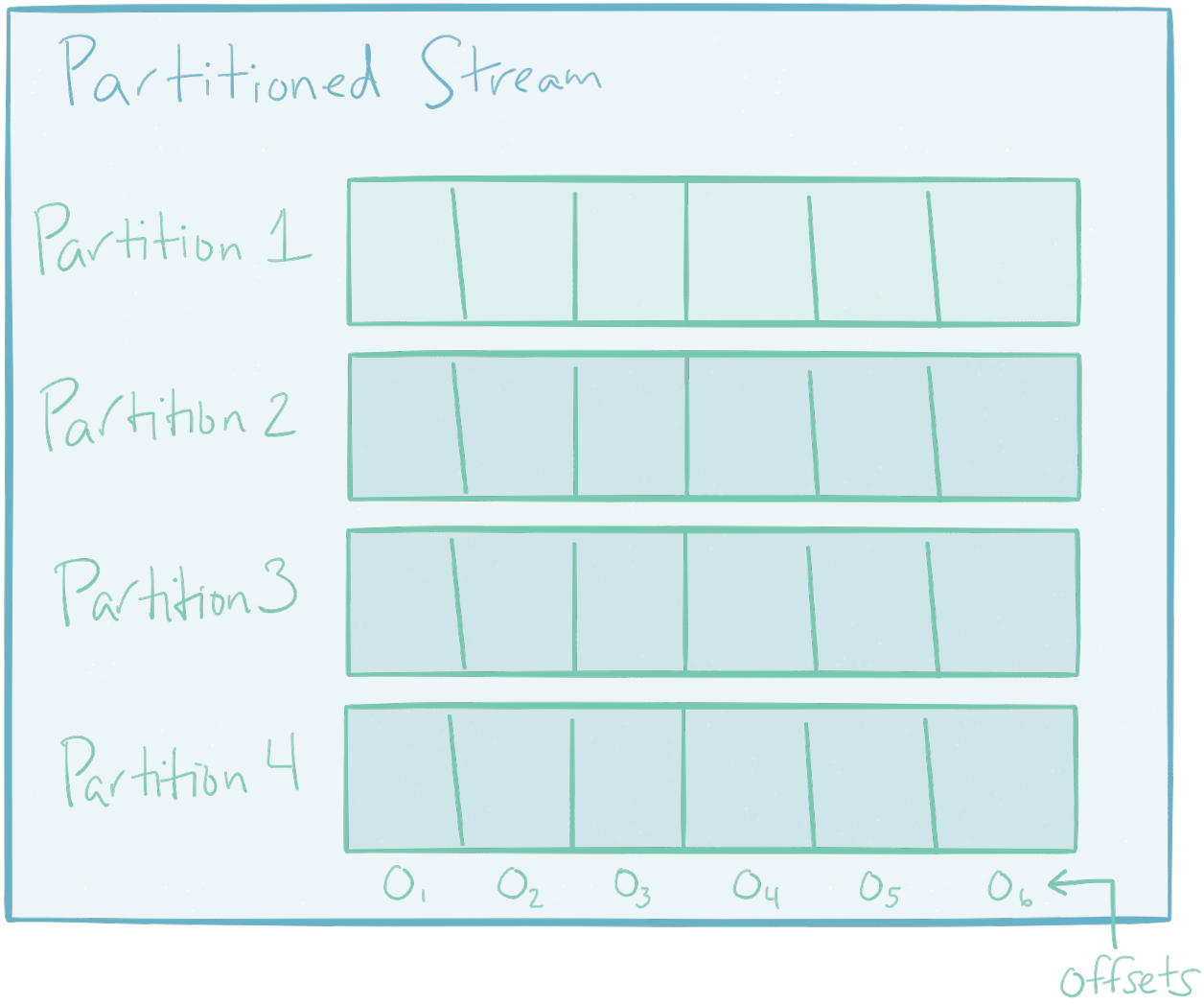
SPARK STREAMING & KAFKA

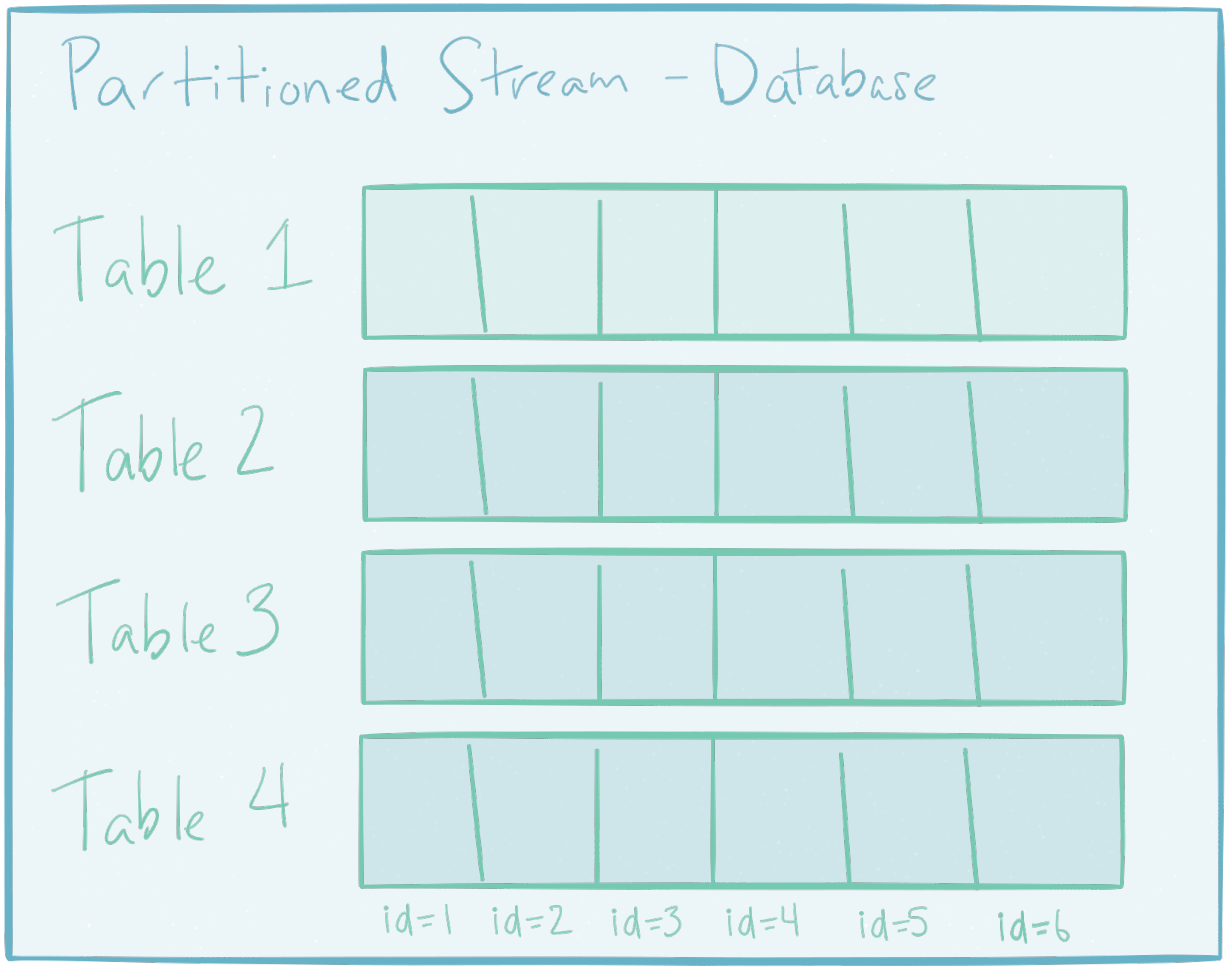
INPUT: DIRECT KAFKA STREAMS

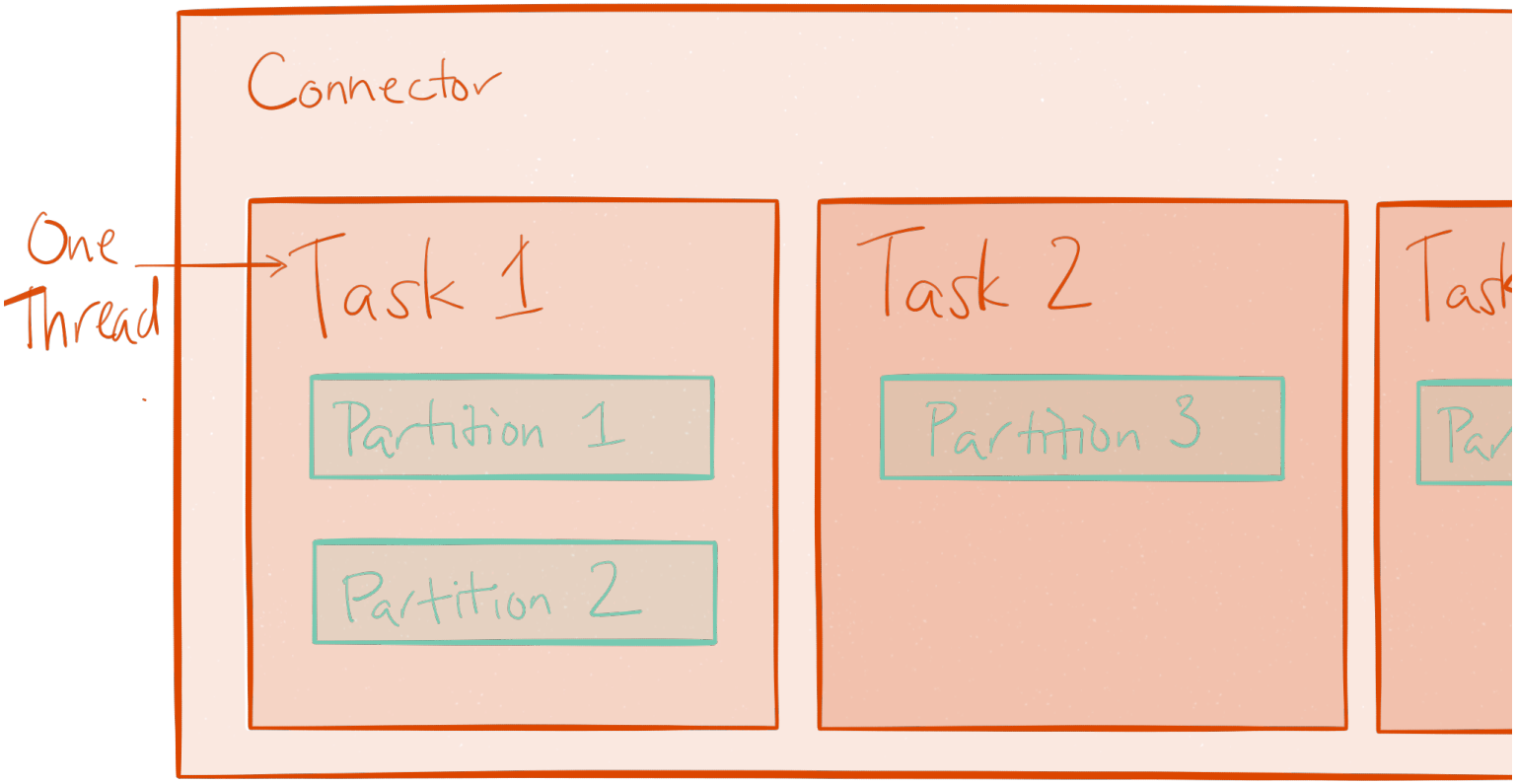
- OLD CONSUMER (0.8.2.1+) - SPARK 2.0+
- NEW CONSUMER (0.10.0.0+) - EXPERIMENTAL

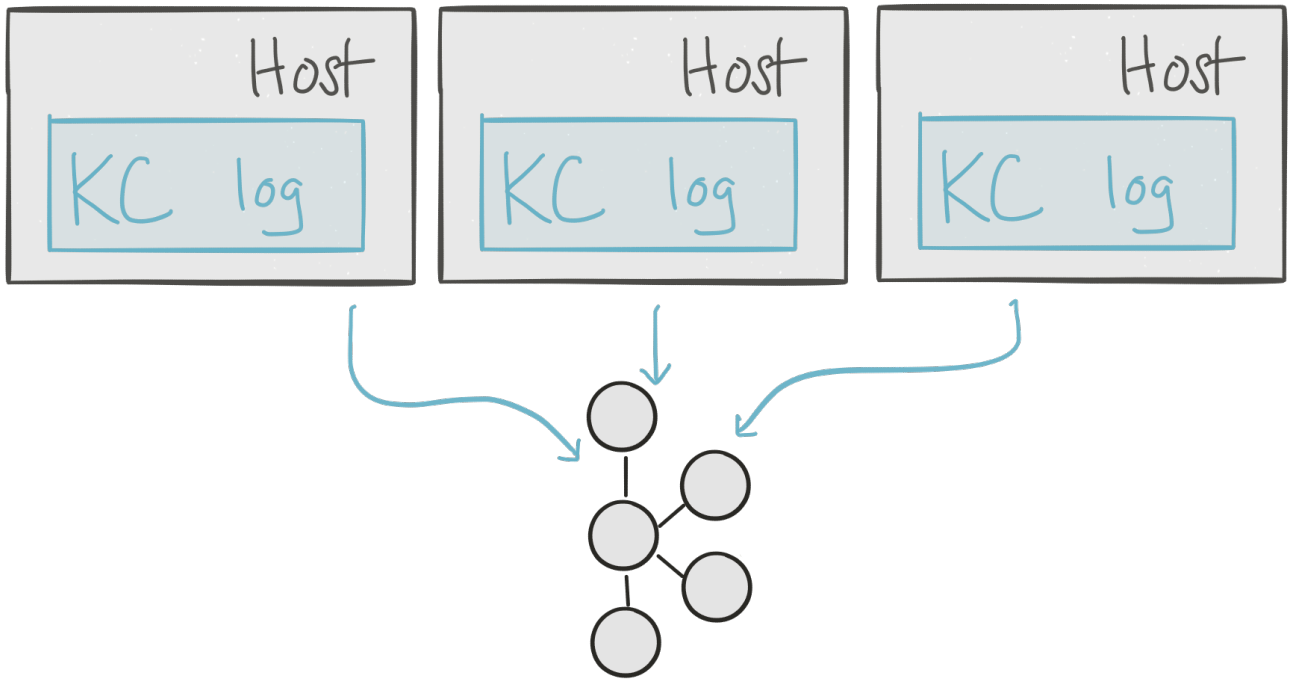
OUTPUT: SPARK KAFKA WRITER

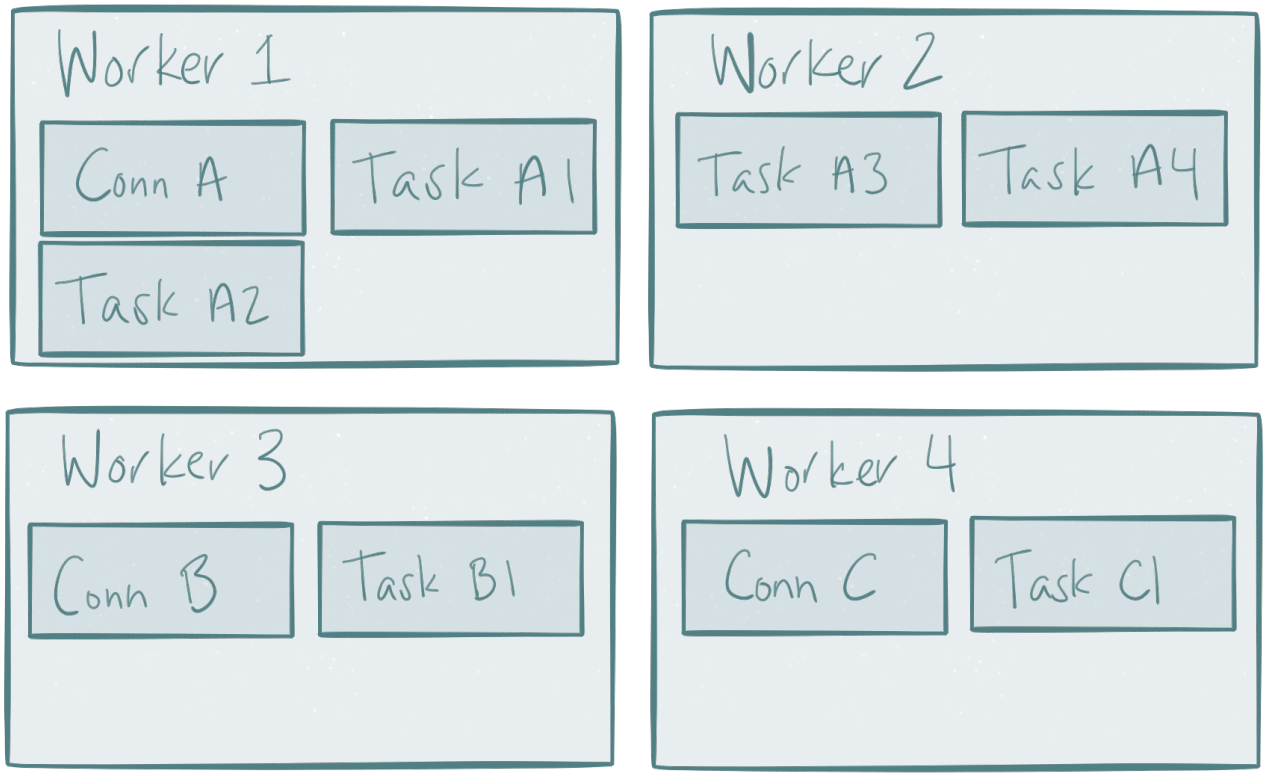


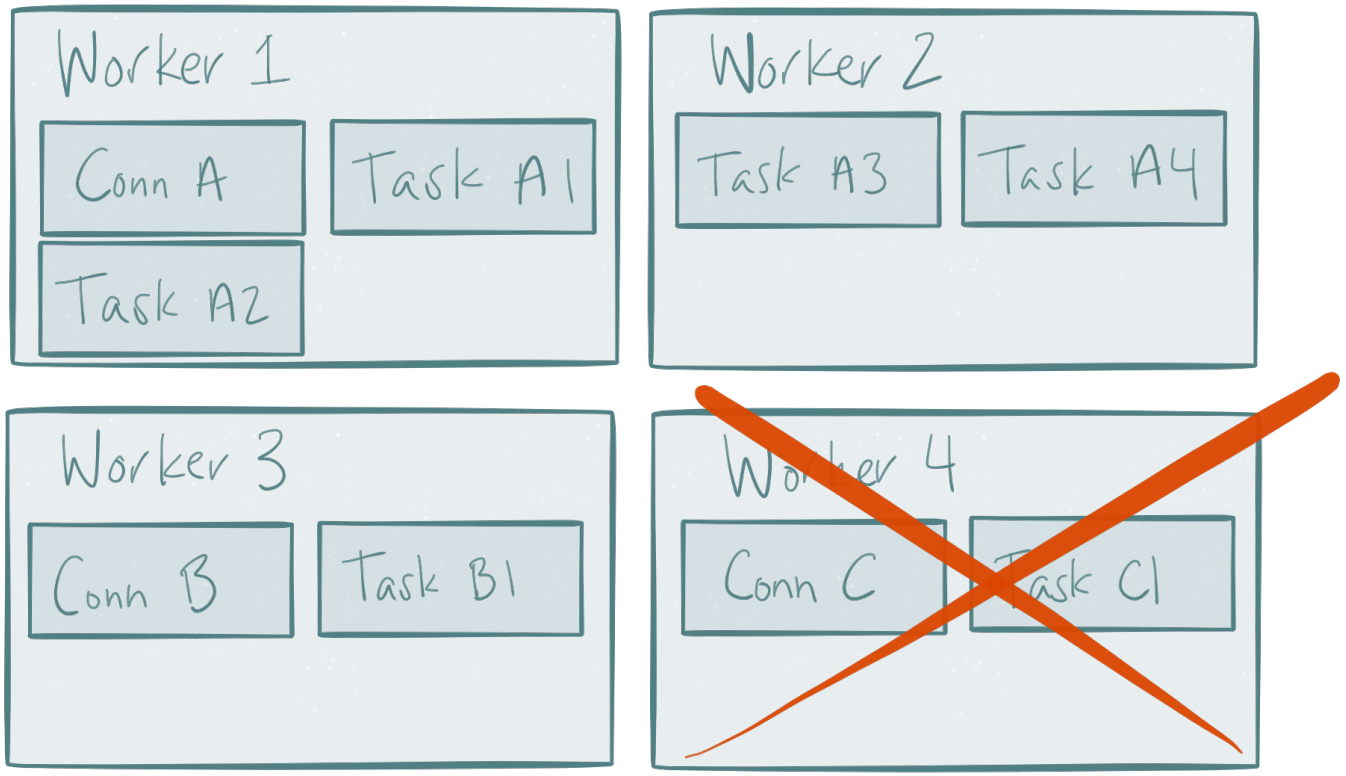

Partitioned Stream
Partition 1
Partition 2
Partition 3
Partition 4
O1
O2
O3
O4
O5
O6
offsets

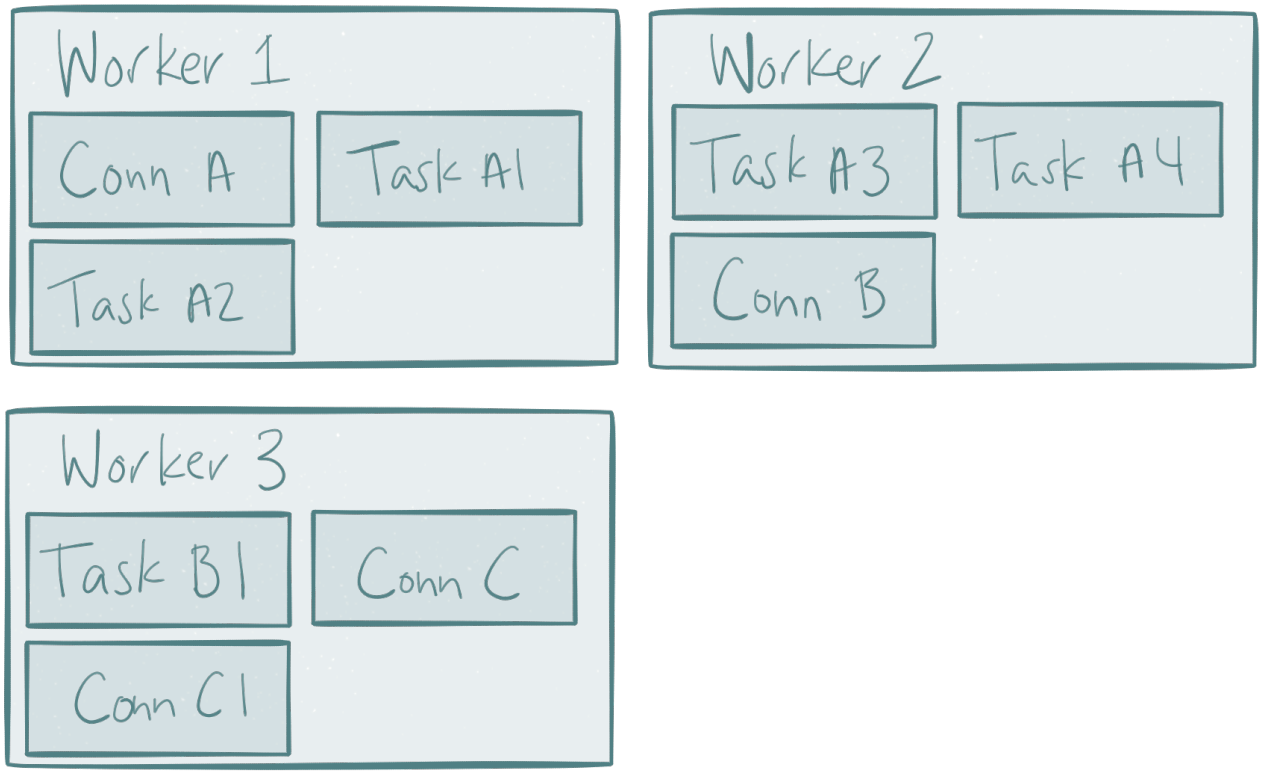



Connector
One Thread
Task 1
Partition 1
Partition 2
Task 2
Partition 3

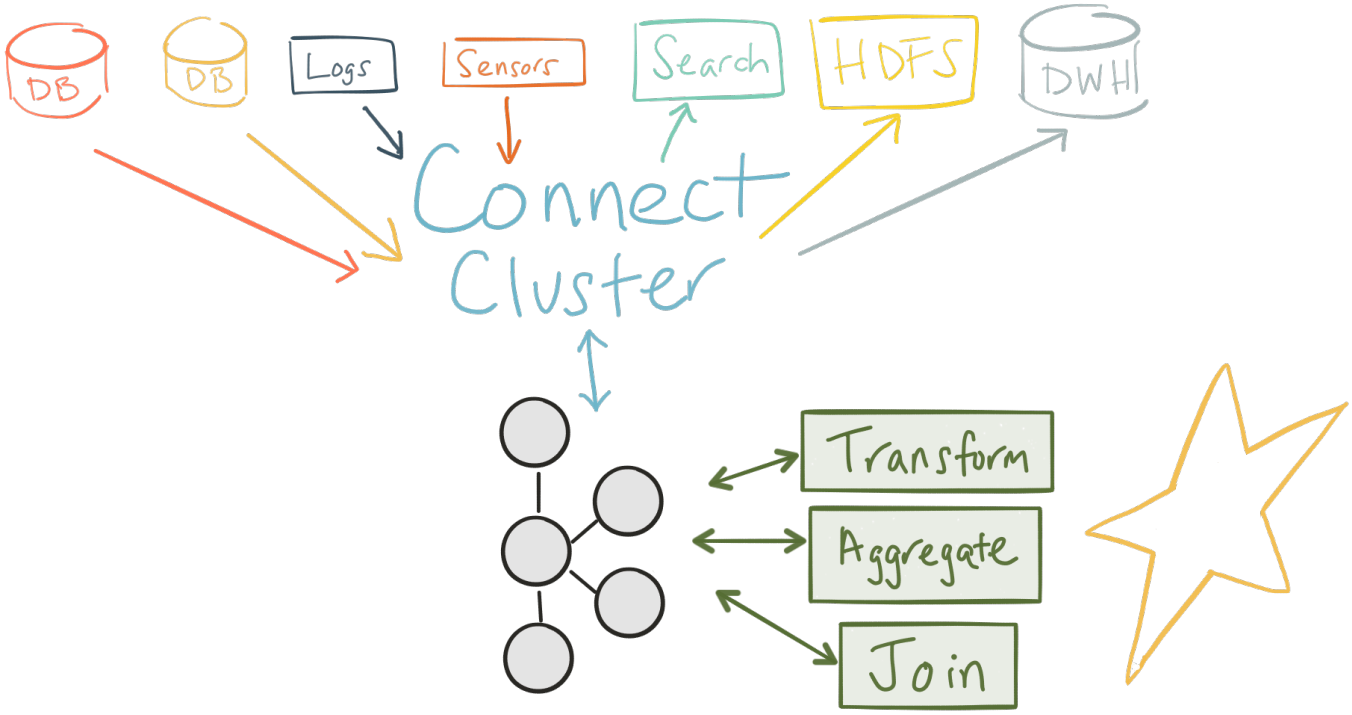

Host
KC log
Host
KC log
Host
KC log


Worker 1
Conn A
Task A1
Task A2
Worker 2
Task A3
Task A4
Worker 3
Conn B
Task B1
Worker 4
Conn C
Task C1


Worker 1
Conn A
Task A1
Task A2
Worker 2
Task A3
Task A4
Worker 3
Conn B
Task B1
Conn C
Task C1


Worker 1
Conn A
Task A1
Task A2
Worker 2
Task A3
Task A4
Conn B
Worker 3
Task B1
Conn C
Conn C1

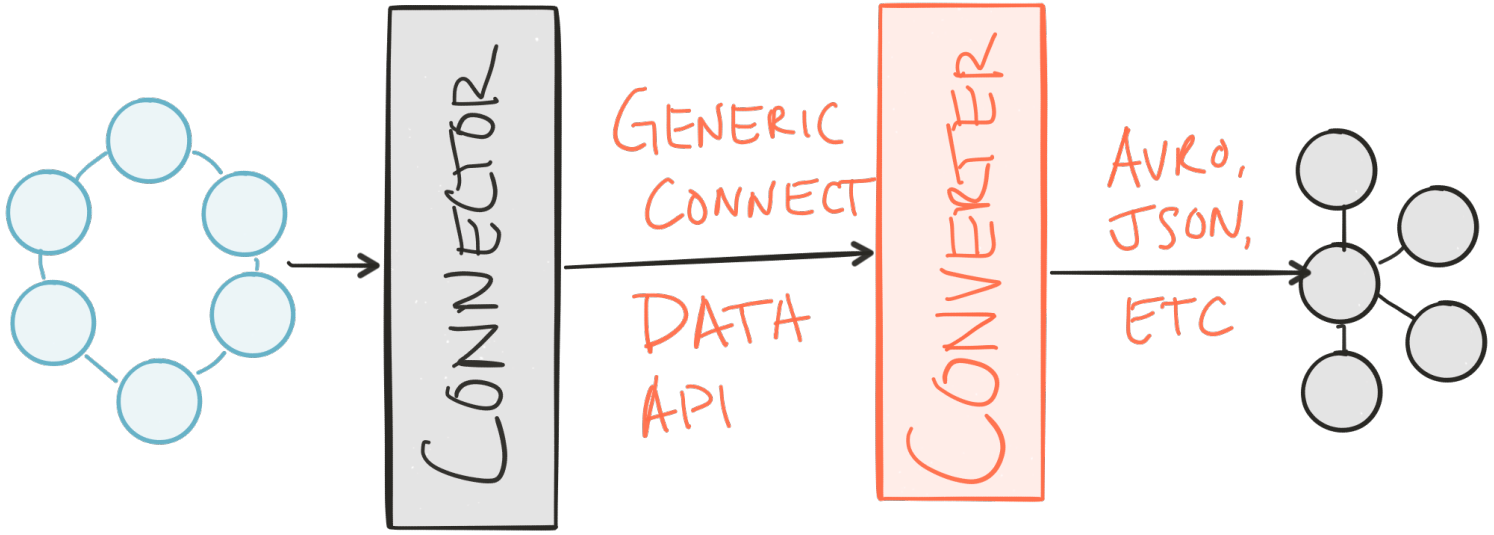
Data Integration as a Service



DELIVERY GUARANTEES

- FRAMEWORK MANAGED OFFSETS
- AT LEAST ONCE DEFAULT
- EXACTLY ONCE (W/ CONNECTOR SUPPORT)

CONVERTERS


CONNECTOR
GENERIC
CONNECT
DATA
API
CONVERTER
AVRO,
JSON,
ETC

Spark & Kafka Connect

More Spark Streaming Sources + Sinks

Reduce Adoption Friction

Leverage Kafka As De Facto Streaming Storage Engine

COMING SOON...

IMPROVED KAFKA CLIENT COMPATIBILITY

EXACTLY ONCE SEMANTICS

Certified Connectors

Certified Connectors have been developed by vendors and/or Confluent utilizing the Kafka Connect framework. These Connectors have met criteria for code development best practices, schema registry integration, security, and documentation.

CONNECTOR	TAGS	DEVELOPER	SUPPORT
HDFS (Sink)	HDFS, Hadoop, Hive	Confluent	Confluent
JDBC (Source)	JDBC, MySQL	Confluent	Confluent
Elastic Search (Sink)	search, Elastic, log, analytics	Confluent	Confluent
DataStax (Sink)	Cassandra, DataStax	Data Mountaineer	Data Mountaineer
Attunity (Source)	CDC	Attunity	Attunity
Couchbase (Source)	Couchbase, NoSQL	Couchbase	Couchbase
GoldenGate (Source)	CDC, Oracle	Oracle	Community
JustOne (Sink)	Postgress	JustOne	JustOne
Striim (Source)	CDC, MS SQLServer	Striim	Striim
Syncsort DMX	DB2, IMS, VSAM, CICS	Syncsort	Syncsort

Want to build a connector?

Interested Open Source Developers and Vendors can get started with the following resources:

> Kafka Connect Overview
> Kafka Connect Developers Guide

Already built a connector?

If you have a connector that you'd like to see added to our list, let us know at: confluent-platform@googlegroups.com

If your connector is well done, we'll send you a free T-shirt and may even list it here.

Contact Us

For software vendors seeking to build a

THANK YOU

@ewencp

@confluentinc

http://confluent.io/download/

http://connectors.confluent.io