MONGODB /
WORLD

Streaming ETL on the Shoulders of Giants

Scott L'Hommedieu, MongoDB
llamadew

Hans-Peter Grahsl, NETCONOMY
hpgrahsl

Streaming ETL on the Shoulders of Giants

Why ETL is important

How we can “ETL better”

Let’s see (some use cases) + a DEMO!

Speed & Agility

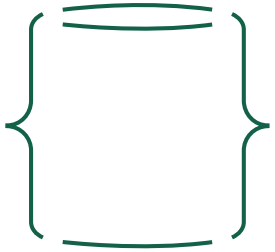
A Top 5 Tech Risk*

For businesses to stay relevant they must deliver value at a breakneck pace and be constantly seeking new sources of value.

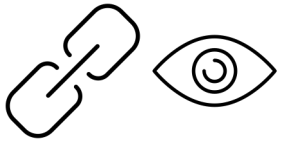
*google "top tech risks"

Managing, Processing and Analyzing Data

We use Data



To unlock insights



And drive value

But, historic ETL is painful

An **antipattern** for Speed and Agility

ETL = **Batch**(Error Prone , Brittle, Slow)

Solving the pain of ETL through Streaming Data

Speed and Agility

ETL = **DataStream** (Resilient, Loosely Coupled, Realtime)

Streaming ETL on the Shoulders of Giants

Why ETL is important

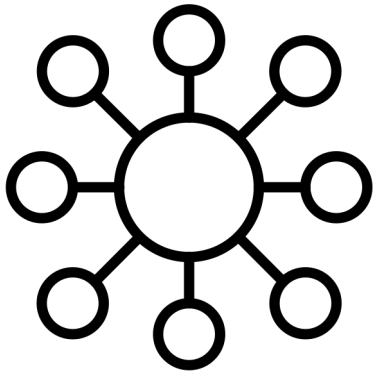
How we can “ETL” better

Let’s see (some use cases) + a DEMO!

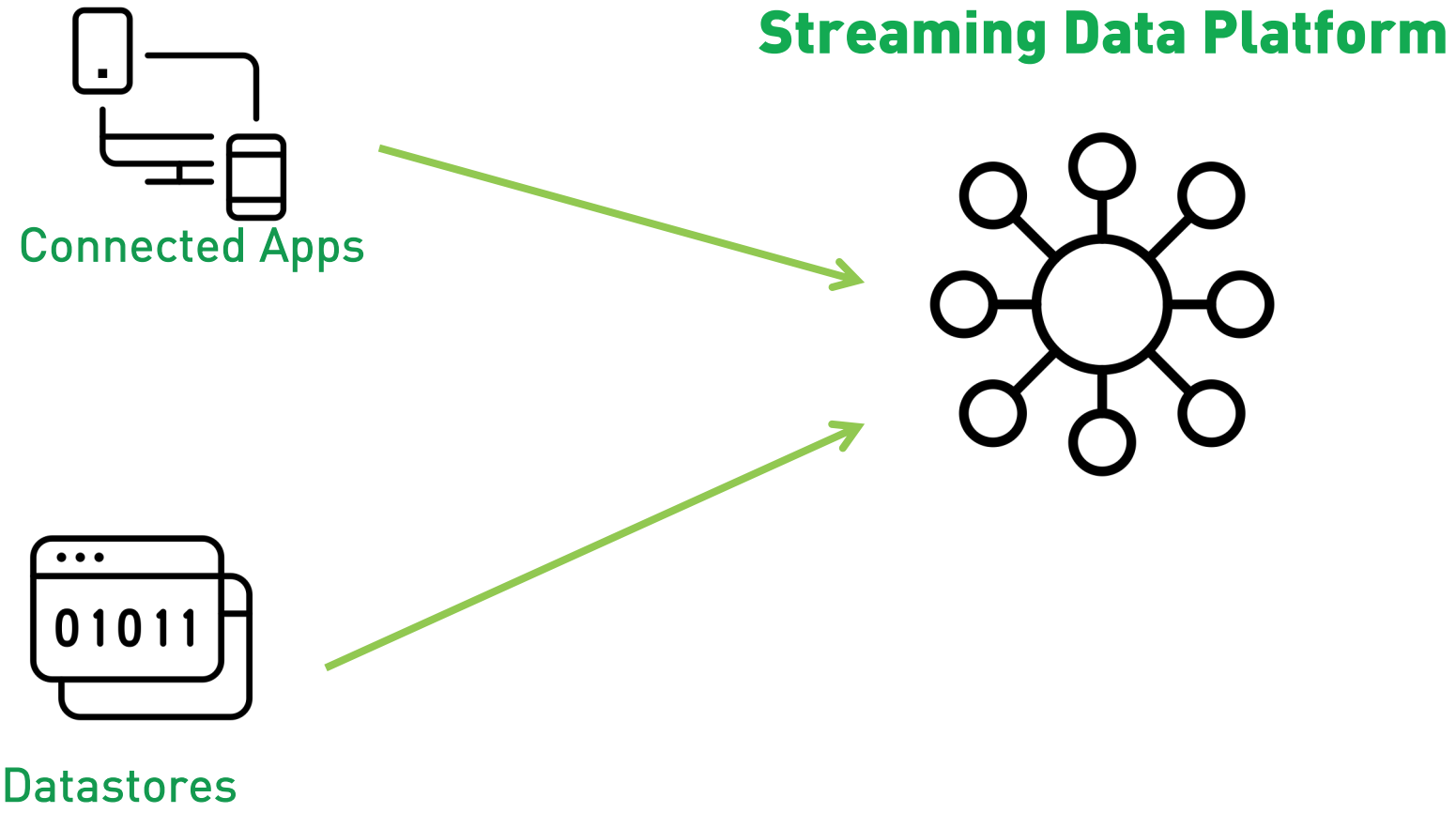
Architecture of a Modern Data Platform

Architecture of a Modern Data Platform

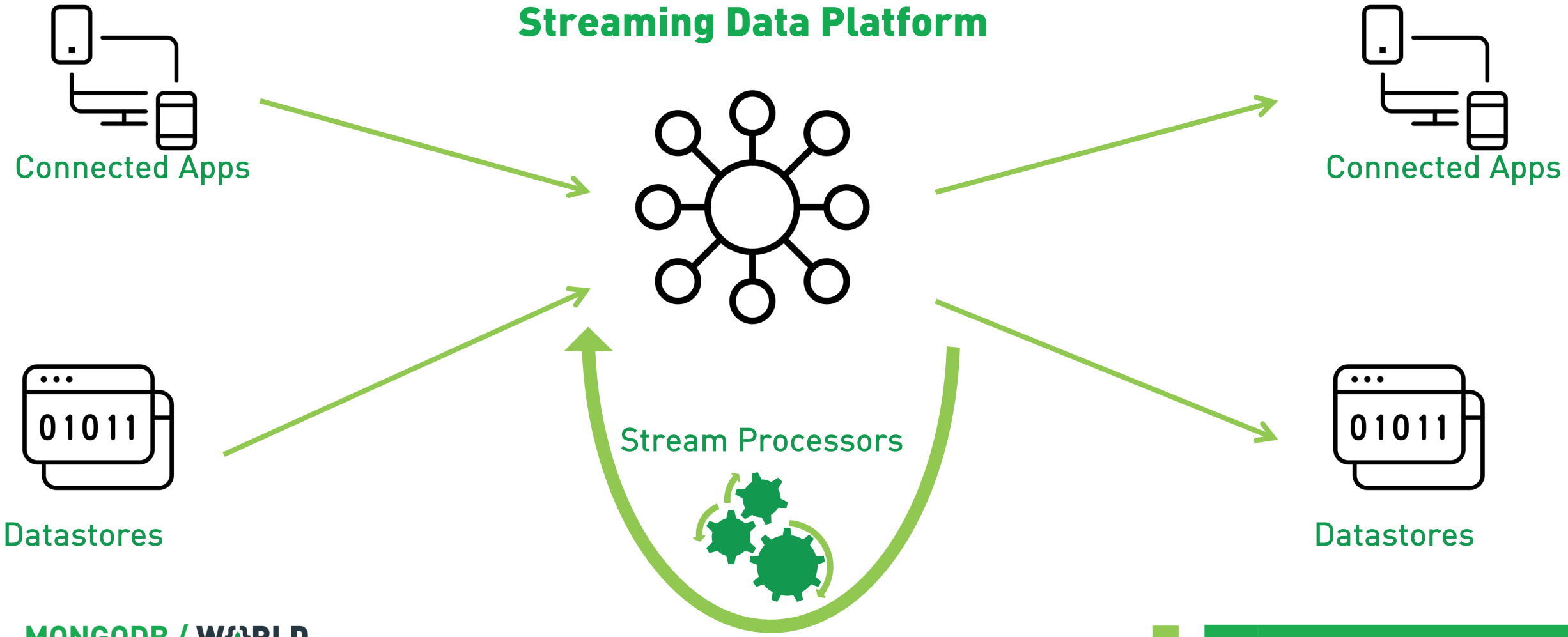
Streaming Data Platform



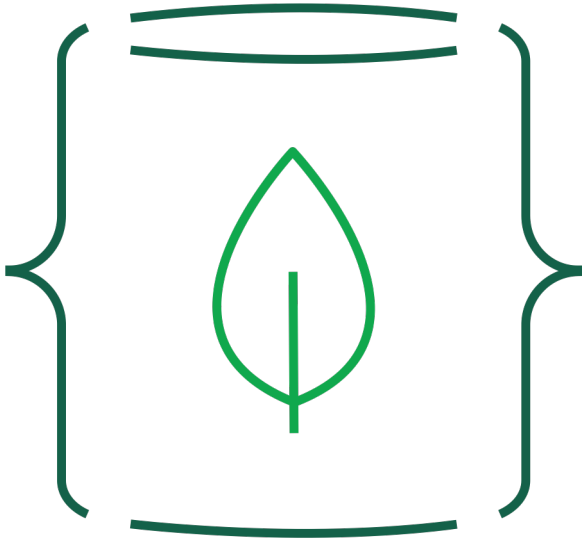
Architecture of a Modern Data Platform



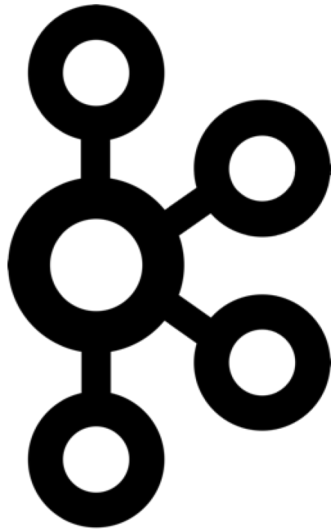
Architecture of a Modern Data Platform



On the shoulders of Giants

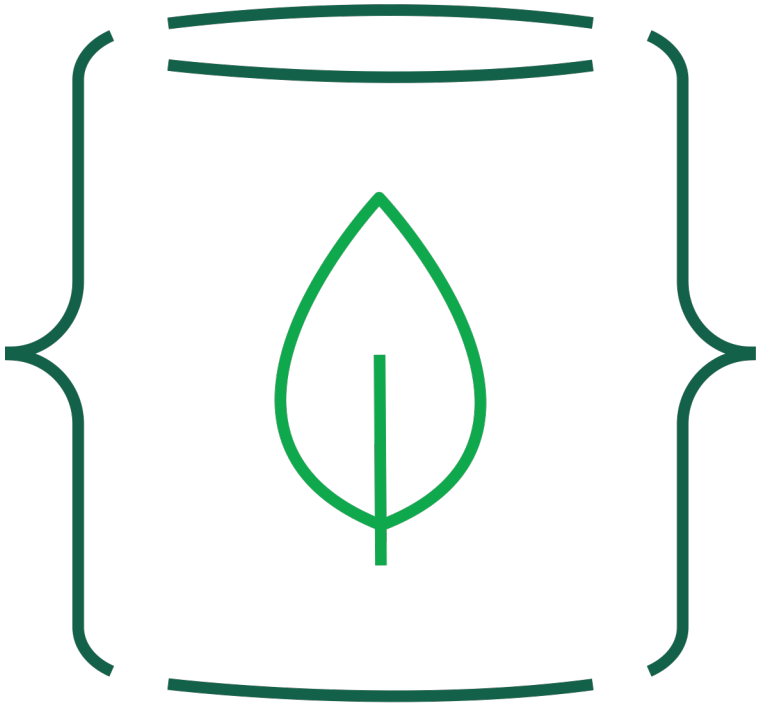


MongoDB



Kafka

Modern Data Platform

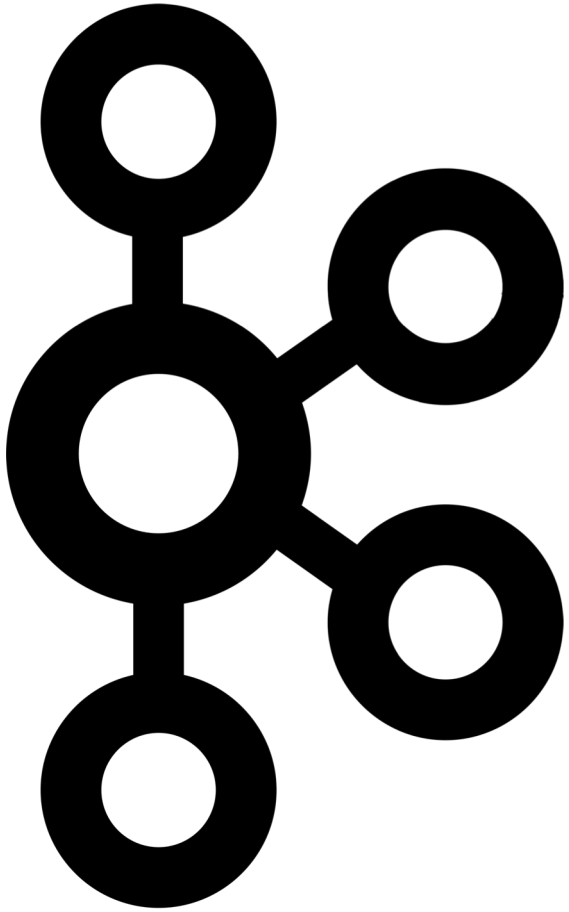


Modern Data Platform

- Doc Model
- Run Anywhere
- Distributed and Scalable
- Resilient and Performant

Apache Kafka 101

Streaming Platform



Streaming Platform

- distributed
- horizontally scalable
- highly fault-tolerant

What is Streaming?

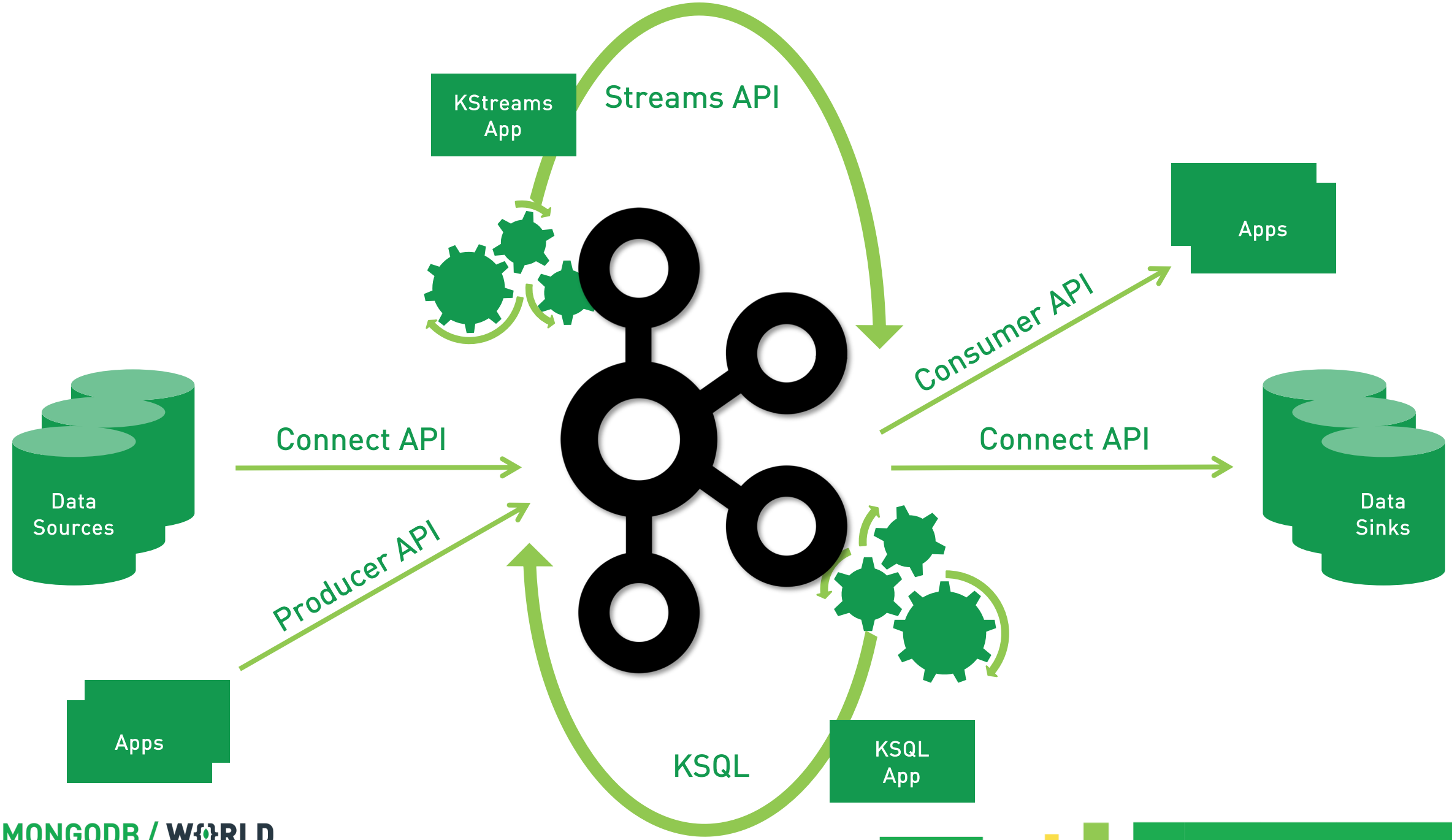
"a type of data processing that is designed with infinite data sets in mind"

–Tyler Akidau

"...everything that happens in a company – every customer interaction, every API request, every database change – can be represented as real-time stream that anything else can tap into, process or react to."

"...Kafka and the whole category of **stream processing** represents a **fundamental paradigm shift** in how the digital part of a company is built, how data is used, and **how applications are built.** This is actually a pretty rare thing..."

– Jay Kreps


KStreams App
Streams API
Apps
Consumer API
Data Sources
Connect API
Connect API
Data Sinks
Producer API
Apps
KSQL
KSQL App

Kafka APIs in a Nutshell...

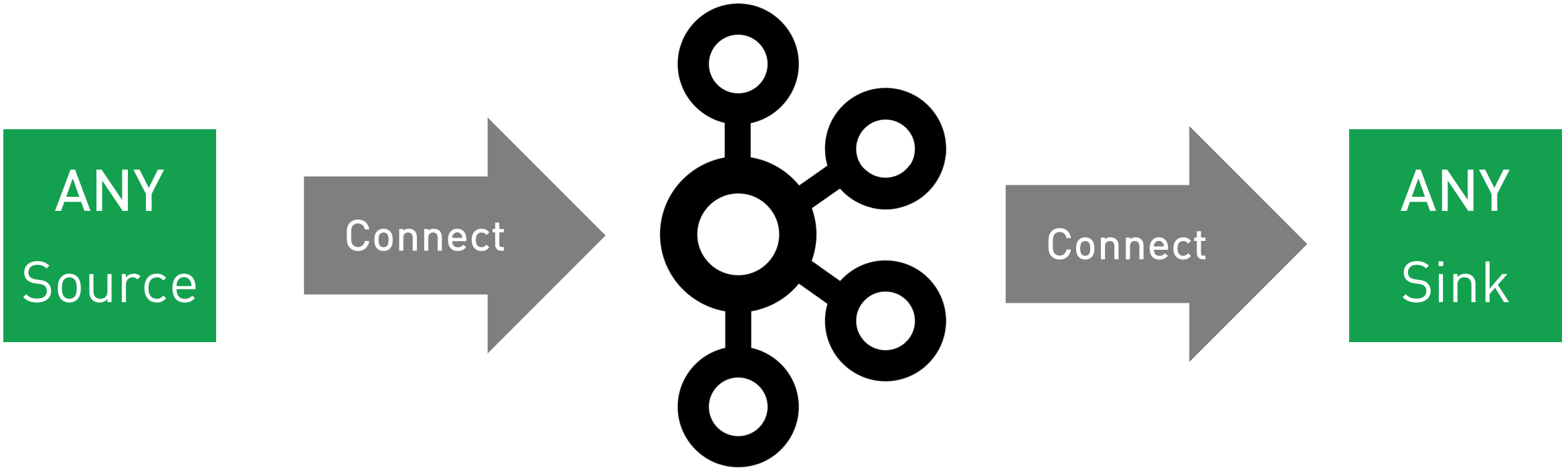
- Producer & Consumer API
  → publish-subscribe scenarios

- Connect API
  → streaming data integration scenarios

- Streams API & KSQL
  → code or SQL-based streaming scenarios

A bit more about Kafka Connect ...

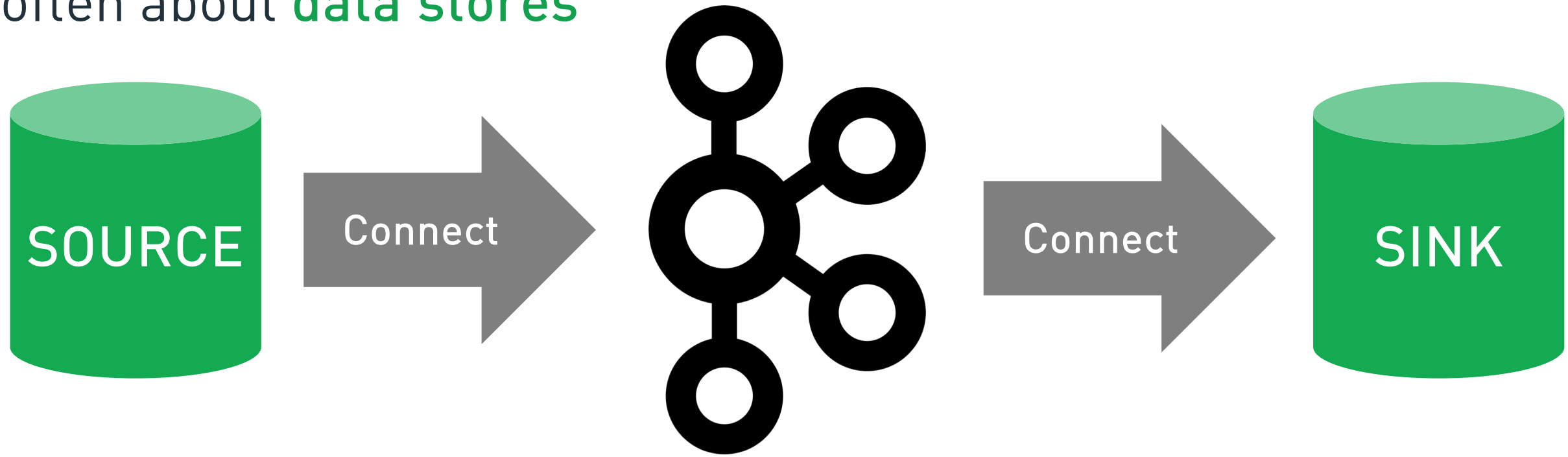
Kafka Connect Basics



ANY → e.g. file systems, **data stores,** REST endpoints, …

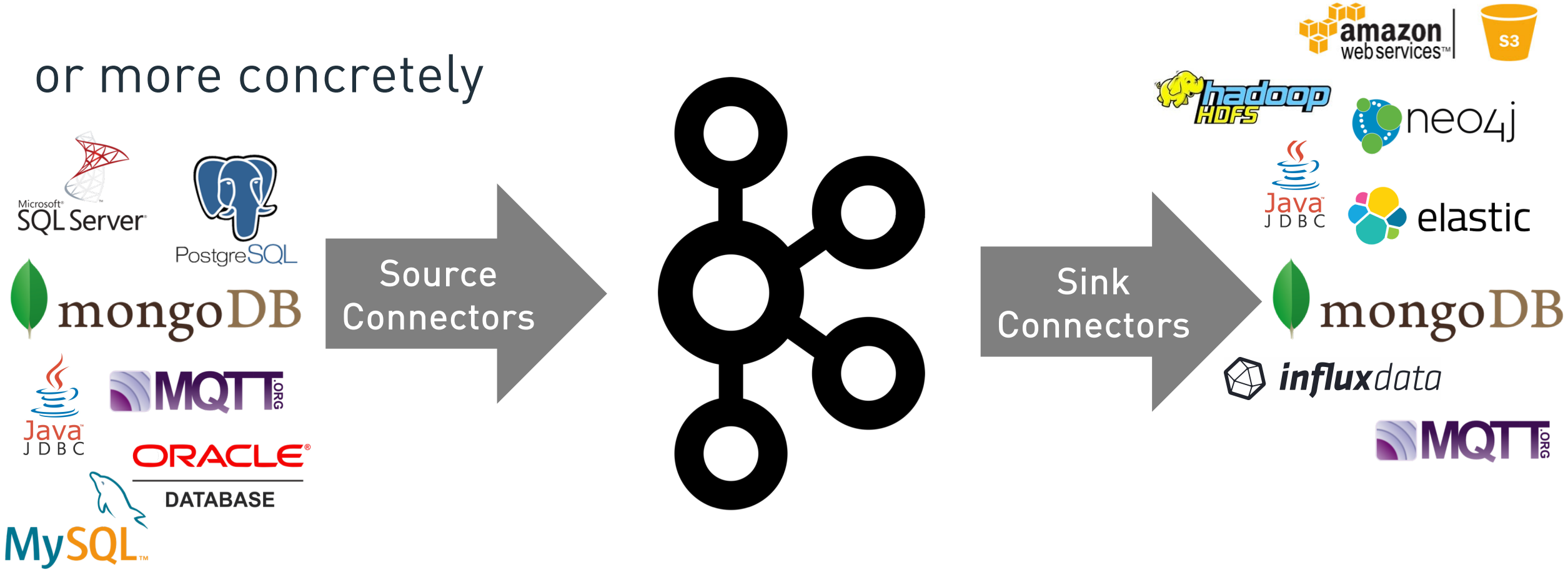
Kafka Connect Basics

often about data stores



Kafka Connect Basics

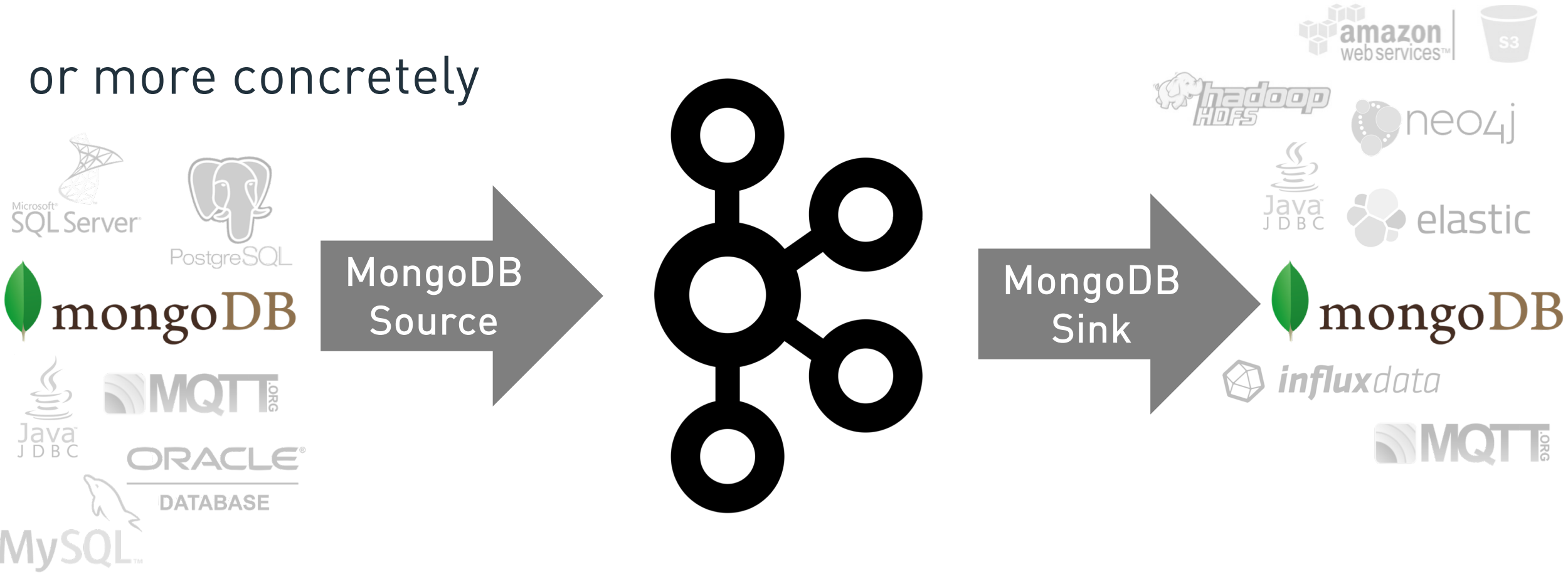
or more concretely



https://hub.confluent.io → many many more

Kafka Connect Basics

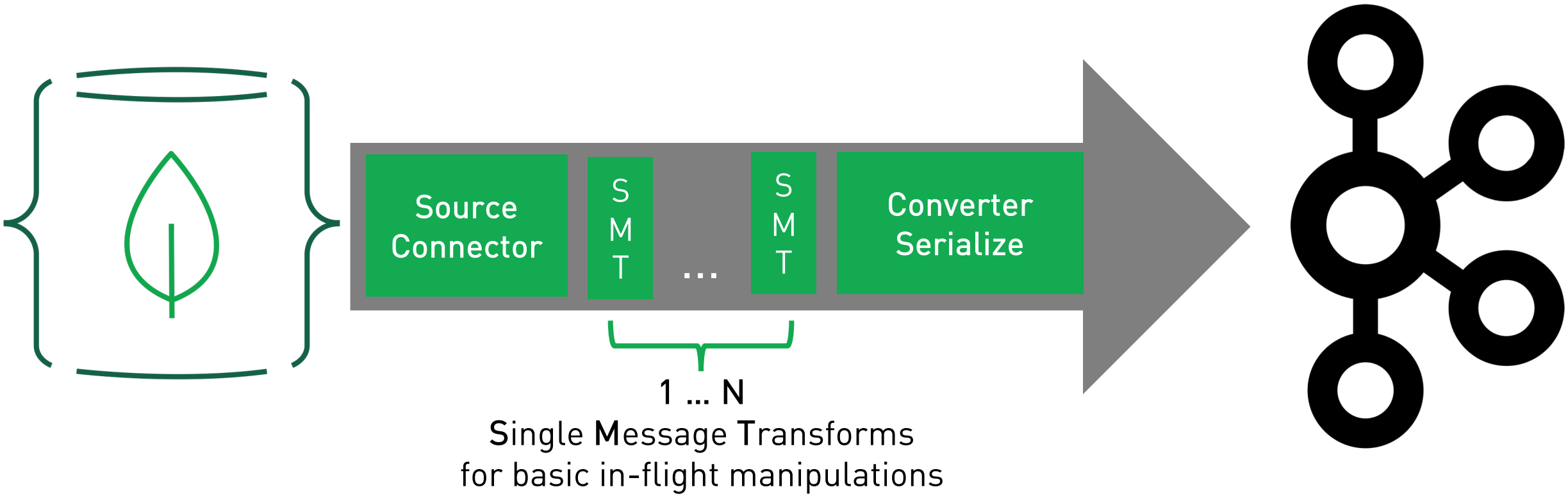
or more concretely



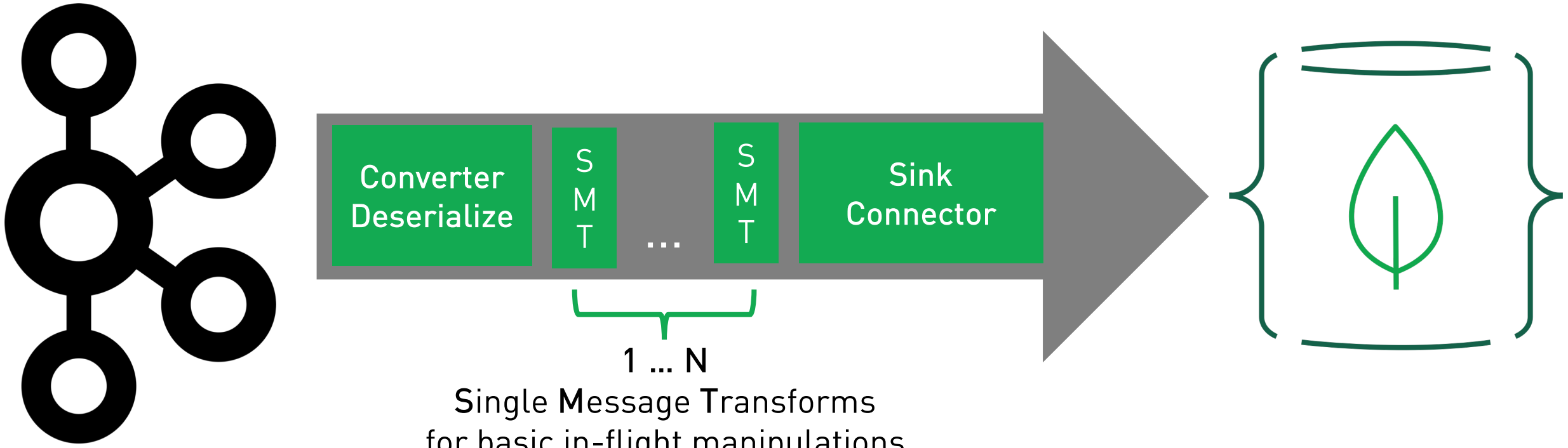
https://hub.confluent.io → many many more

How do connectors operate?

Kafka Source Connectors



Kafka Sink Connectors



Announcing ...

MongoDB Connector for Apache Kafka

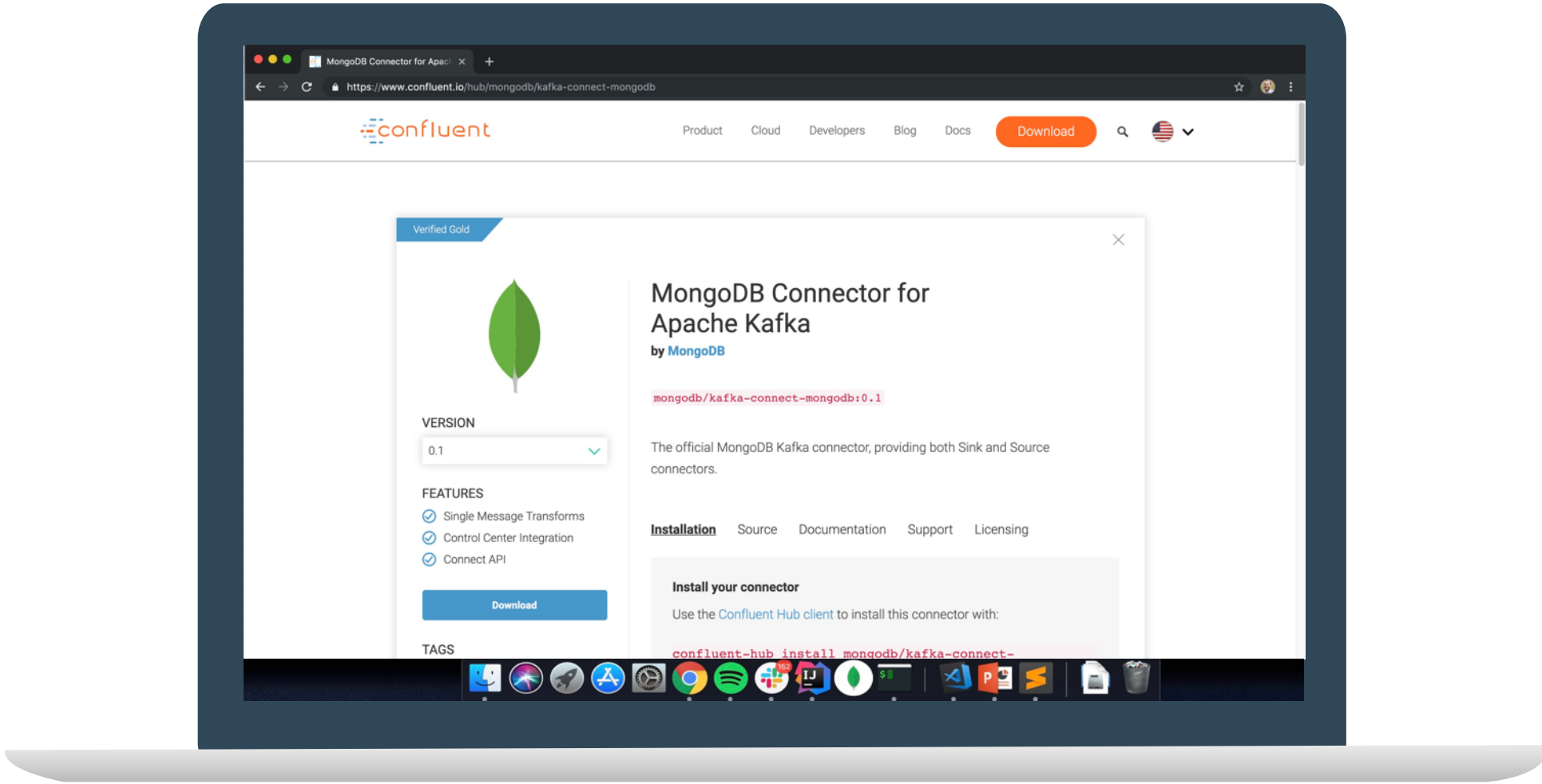
Supported by MongoDB

Verified Gold by confluent

Available on the Confluent Hub: https://www.confluent.io/hub/mongodb/kafka-connect-mongodb

MongoDB Connector for Apache Kafka

Available on the Confluent Hub: https://www.confluent.io/hub/mongodb/kafka-connect-mongodb



Streaming ETL on the Shoulders of Giants

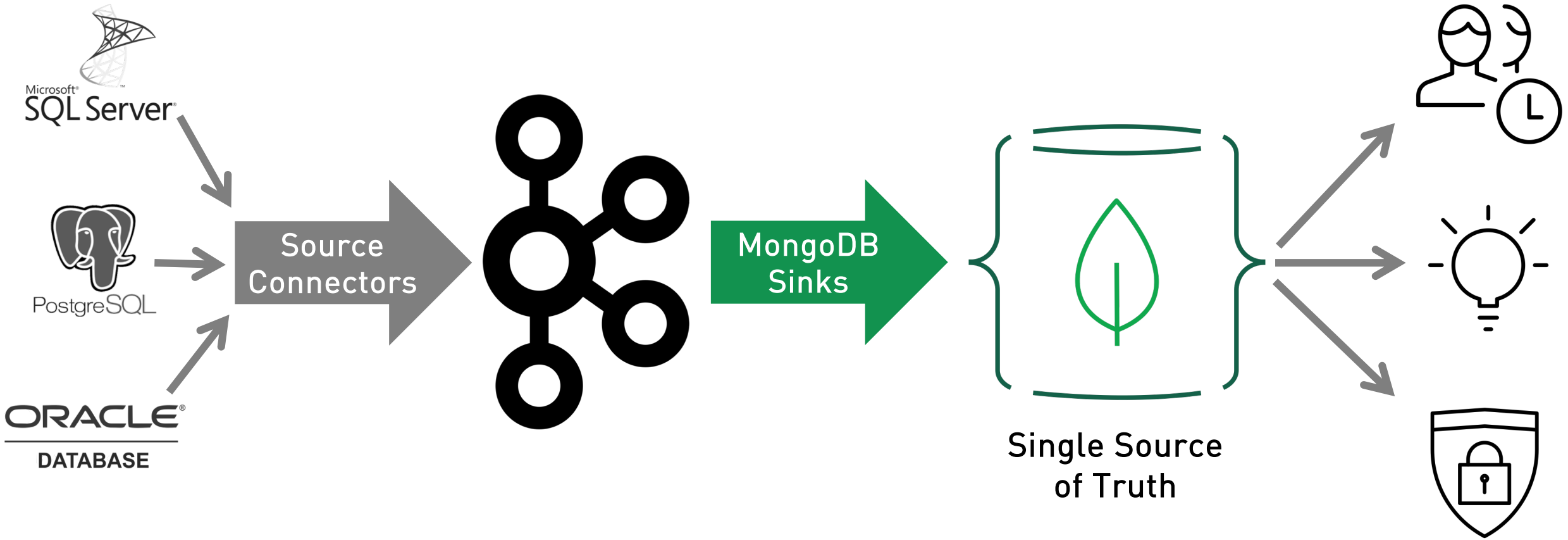
Why ETL is important

How we can “ETL better”

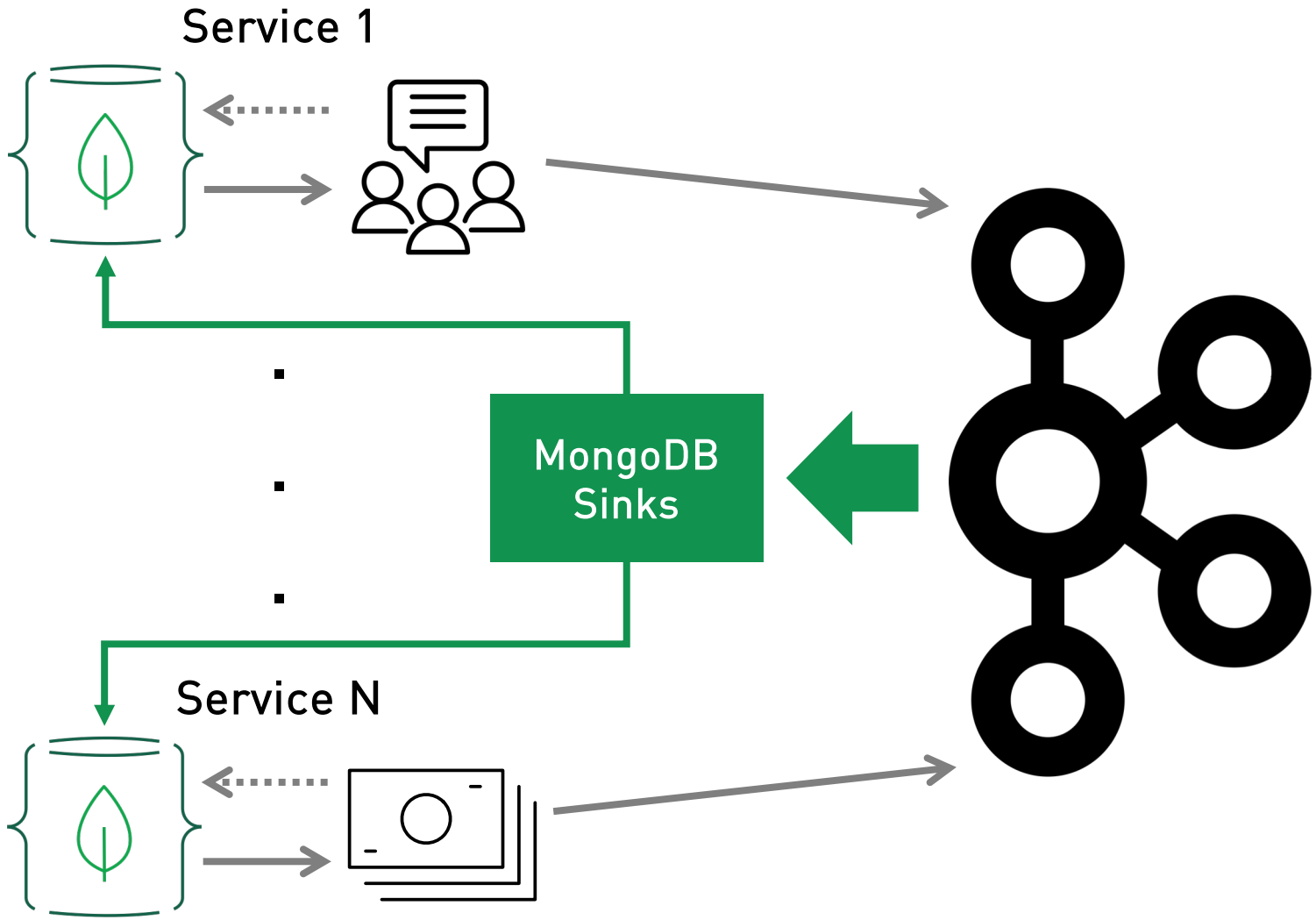
Let’s see (some use cases) + a DEMO!

Streaming ETL Use Cases

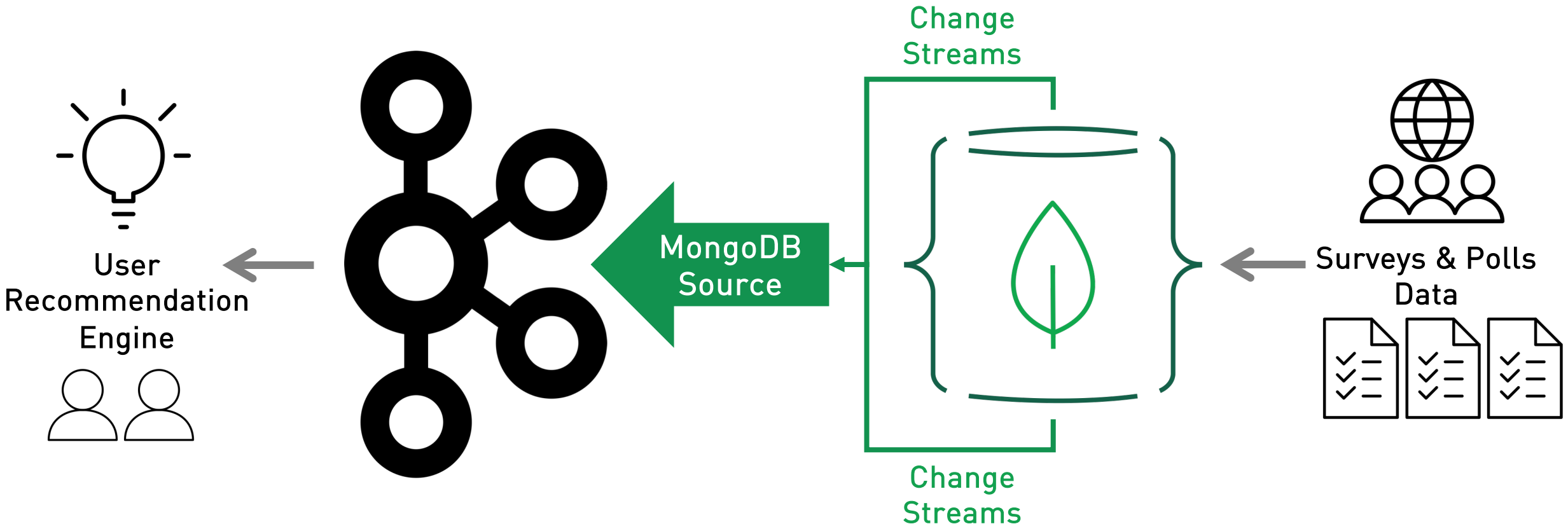
Single Customer View for eCommerce



Data Synchronization between Microservices

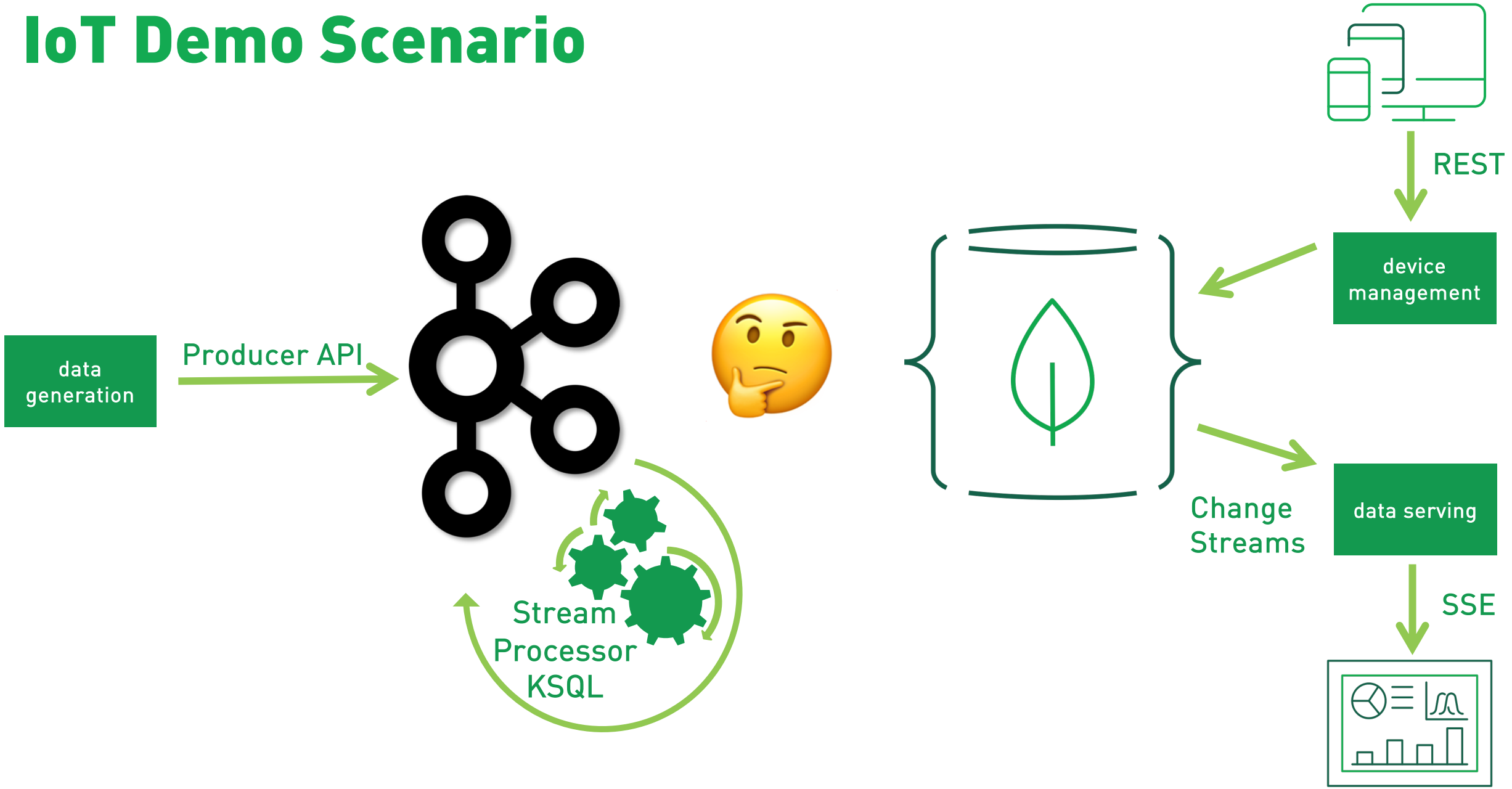


Recommendation Engine for Opinion Mining

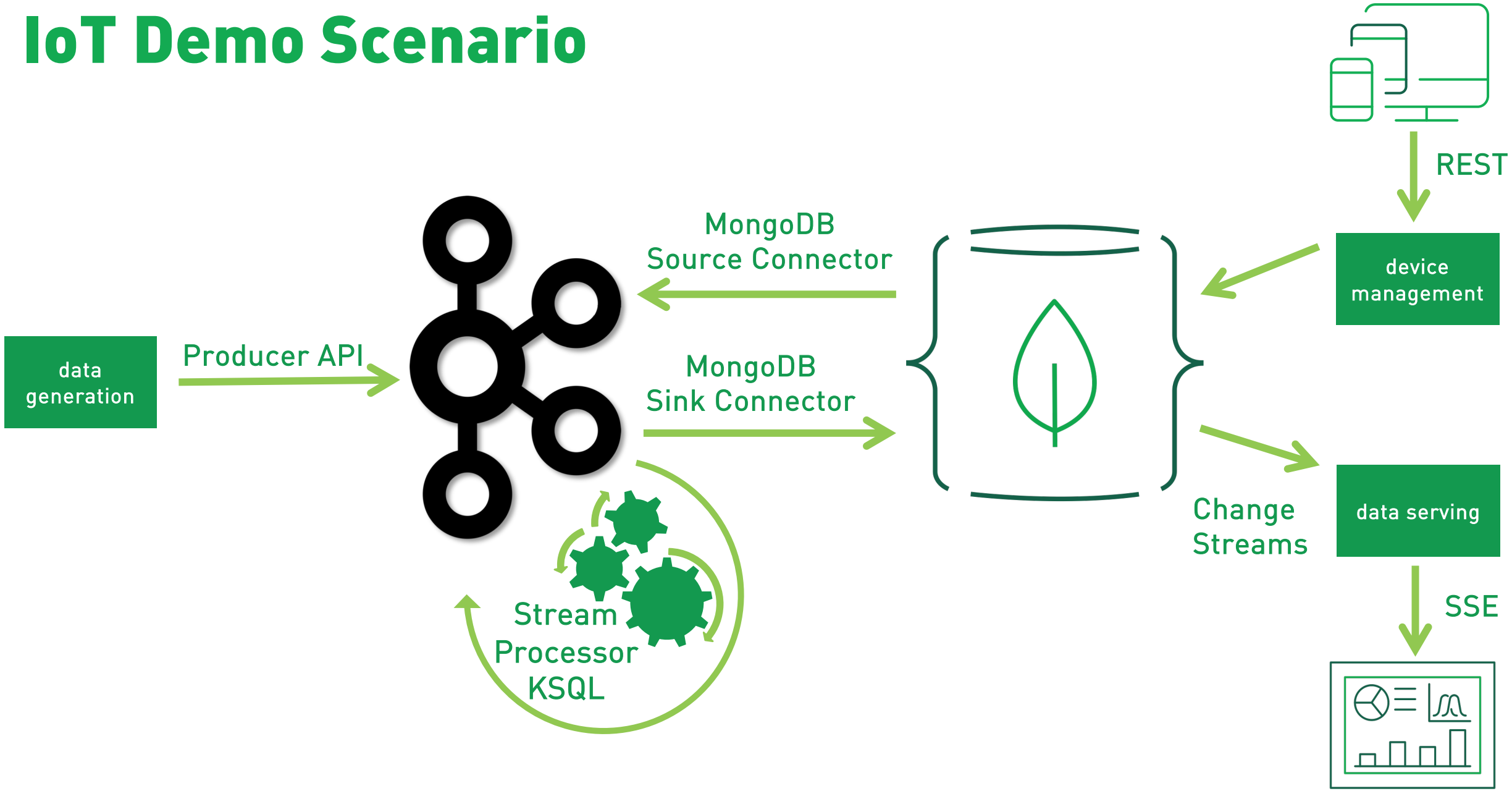


IoT Demo Scenario in Action

IoT Demo Scenario



IoT Demo Scenario



That’s all folks!

THANK YOU

MONGODB /
WORLD