
香港
Data Pipeline with Kafka
Dr. Mole T.Y. WONG @ HK OSCON 2018
2018 / 06 / 16 - 17

whoami

Why

深入了解用戶行為，洞悉可行的改善方法

Understand our users.

Provide actionable insights.

How

以數據驅動產品方向

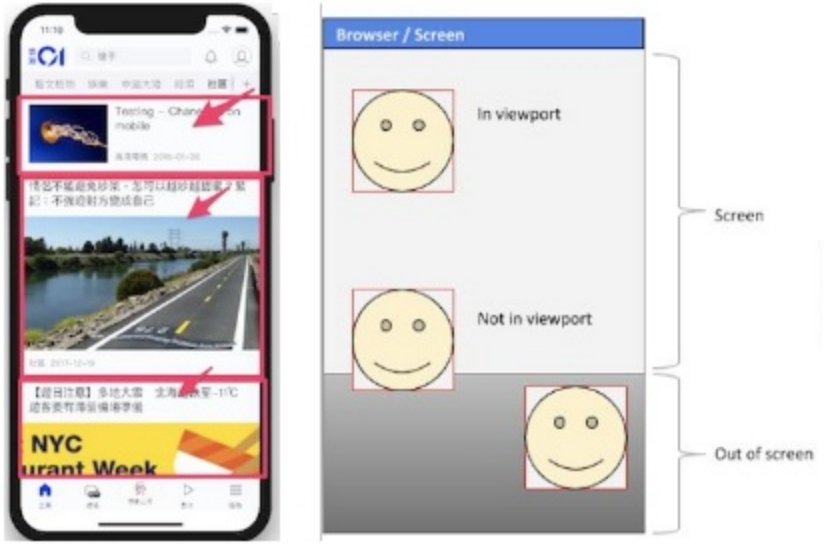
Data driven: steer our product direction.

What

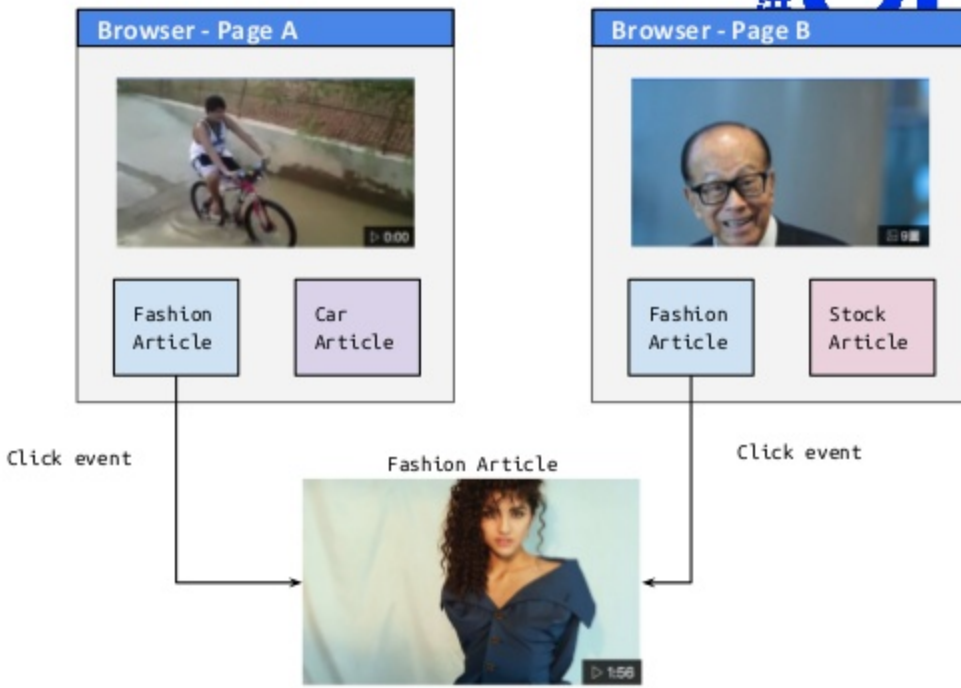
數據：定義、收集、處理、洞見

Data: definition, ingress, process, insight.


SERVER LOG?
GIVE ME FRONT END EVENTS
imgflip.com



Click-Through Rate VS Pageview



Traffic Source Analysis

Data-Driven Product Development

User Reading History

NLP Content-based Clustering

Collaborative filtering
Image source: wikipedia

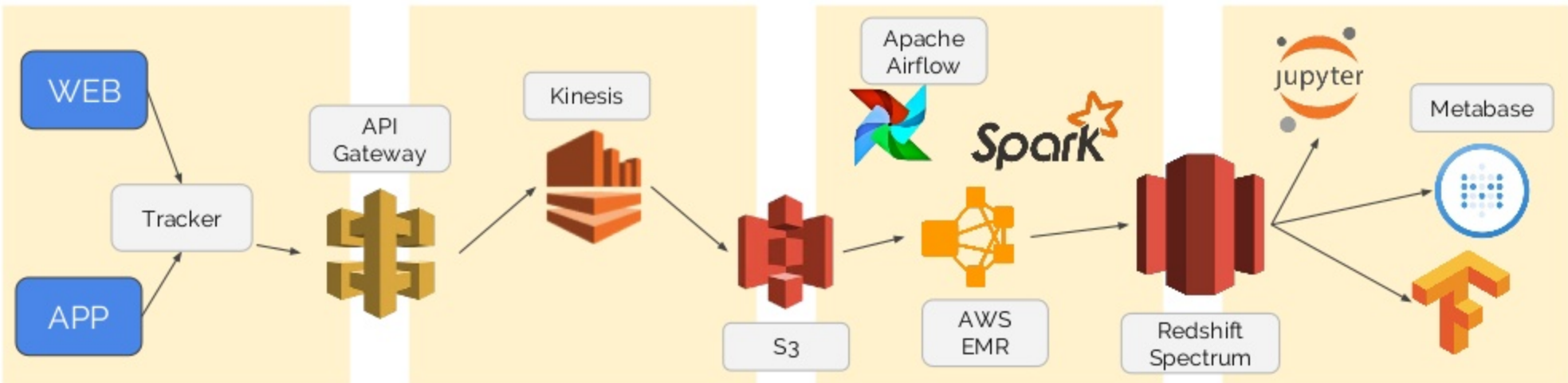
Personalized Recommendation Feed

Machine Learning Products

Outline

- Data pipeline - what is it?
- Kafka - roles in a data pipeline
- Other use cases of Kafka

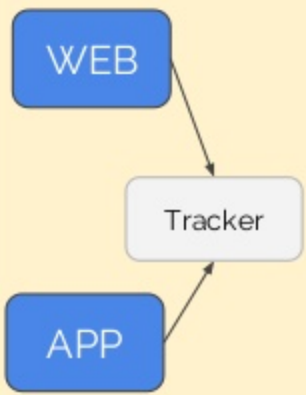
Typical Data Pipeline Setup



Data Ingress

JS Library (WEB)
Native Library (APP)

Google Analytics
Mixpanel
Matomo (Piwik)



Data Tracker

- **Nature**
  - Lightweight
  - Programmable
- **Capability**
  - Page view / Screen view
  - Custom events
  - Device identification
  - Session management

Different Aspects of a Data Pipeline

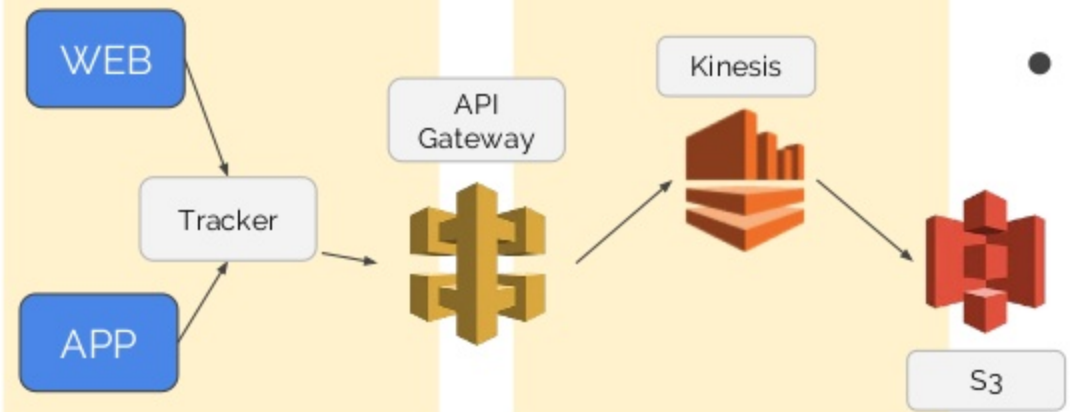
Data Ingress

JS Library (WEB)
Native Library (APP)

Google Analytics
Mixpanel
Matomo (Piwik)

Infrastructure

- AWS Kinesis
- Google Pub/Sub
- Apache Kafka



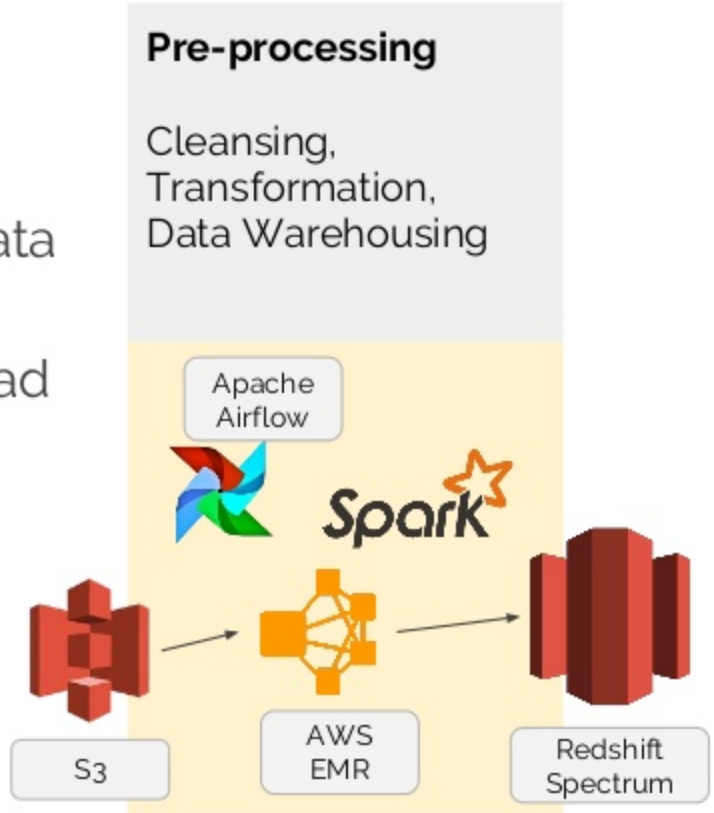
Data Infrastructure

- **Main Roles**
  - Buffering
  - Routing
  - Writing
- **Characteristics**
  - Multiple producers
  - Multiple consumers
  - Batch / Real-time

Different Aspects of a Data Pipeline

Pre-processing

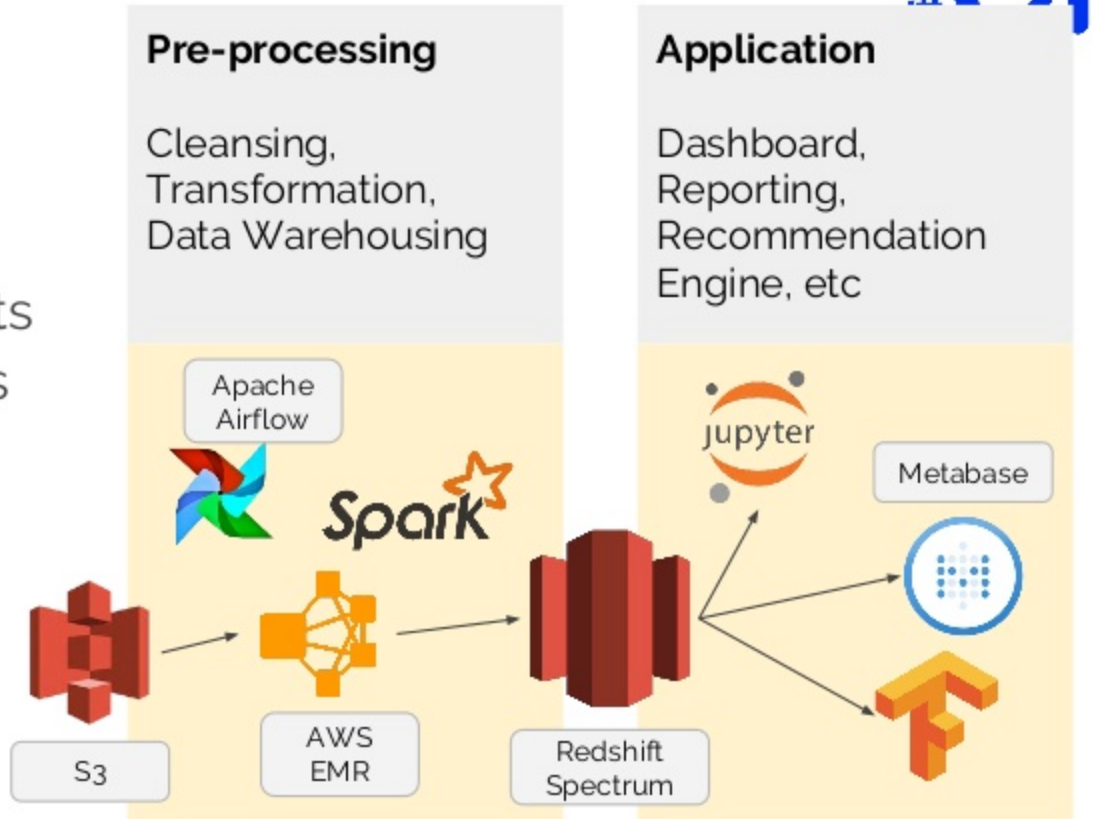
- **Main Roles**
  - Avoid direct querying raw data
  - Cleansing
  - ETL - Extract, Transform, Load
  - Scheduling
- **Characteristics**
  - Defining data sets
  - Time-frame-based queries



Different Aspects of a Data Pipeline

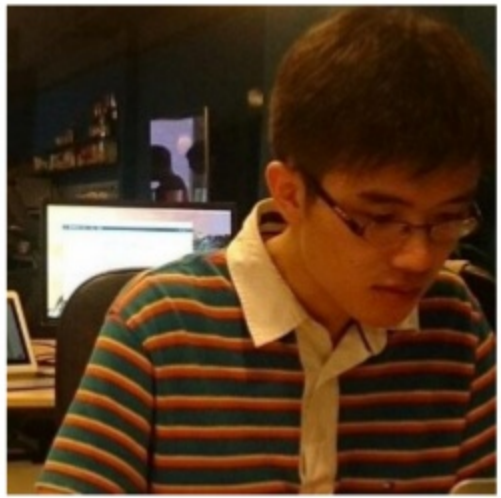
Application

- **Main Roles**
  - KPI VS Exploration
  - Operators VS Data Scientists
  - Planned VS Ad-hoc queries
- **Characteristics**
  - Production-grade data
  - Fast is a must



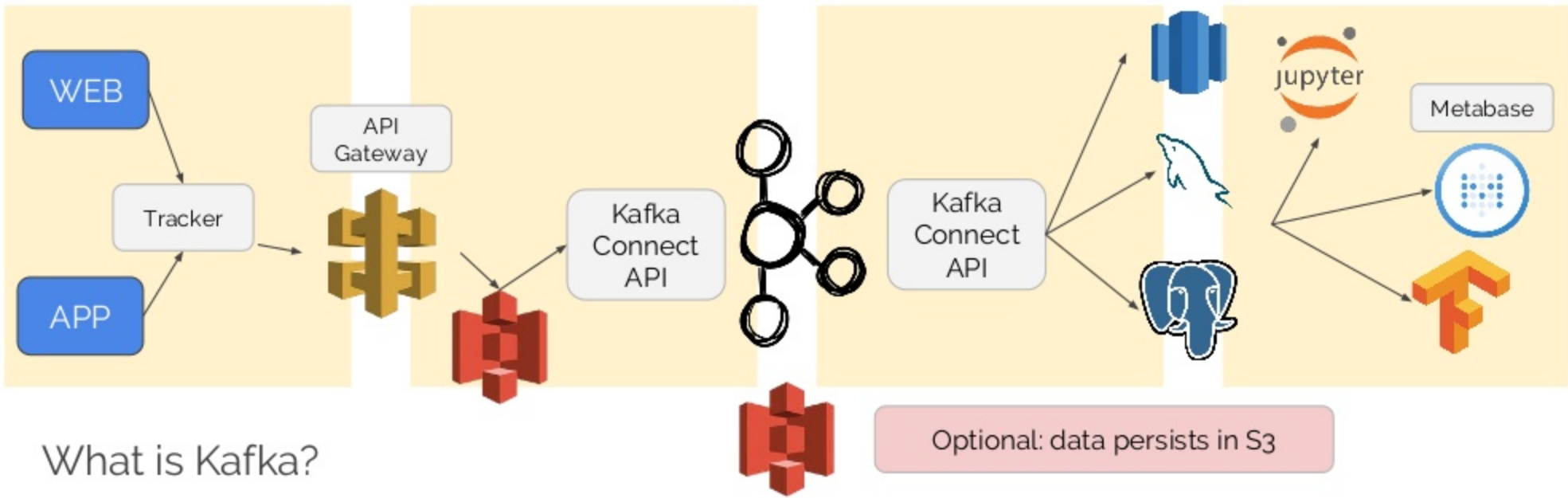
Different Aspects of a Data Pipeline

APACHE kafka®
A distributed streaming platform



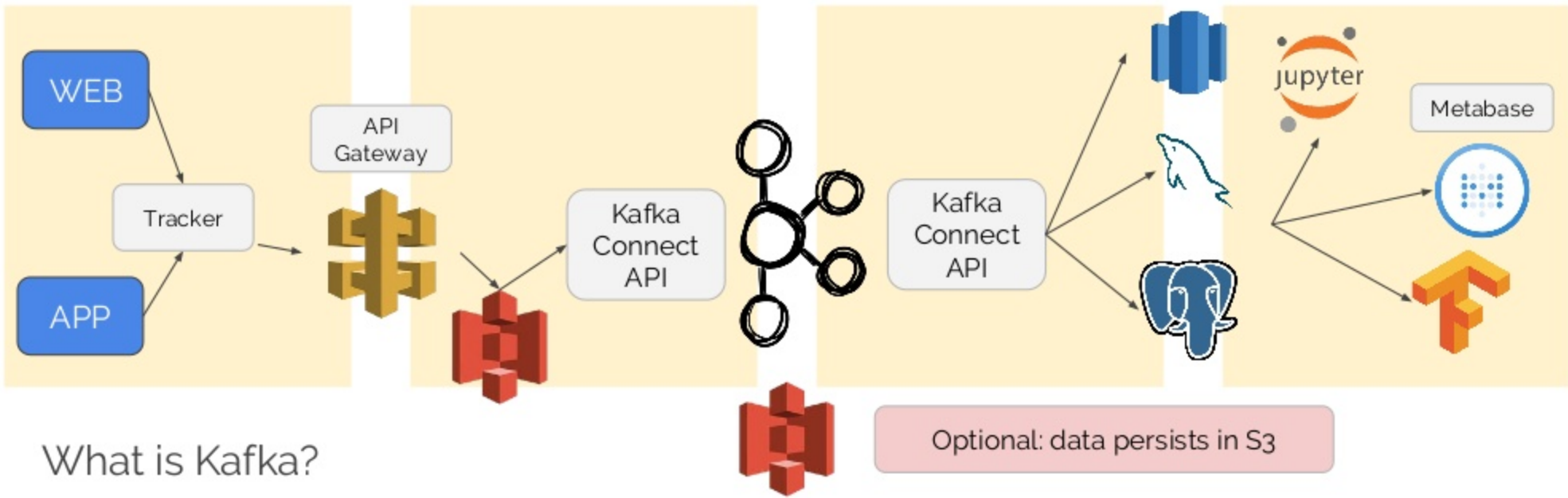
What is Kafka? https://kafka.apache.org/ Main Contributor: Gene NG

Data Pipeline with Kafka



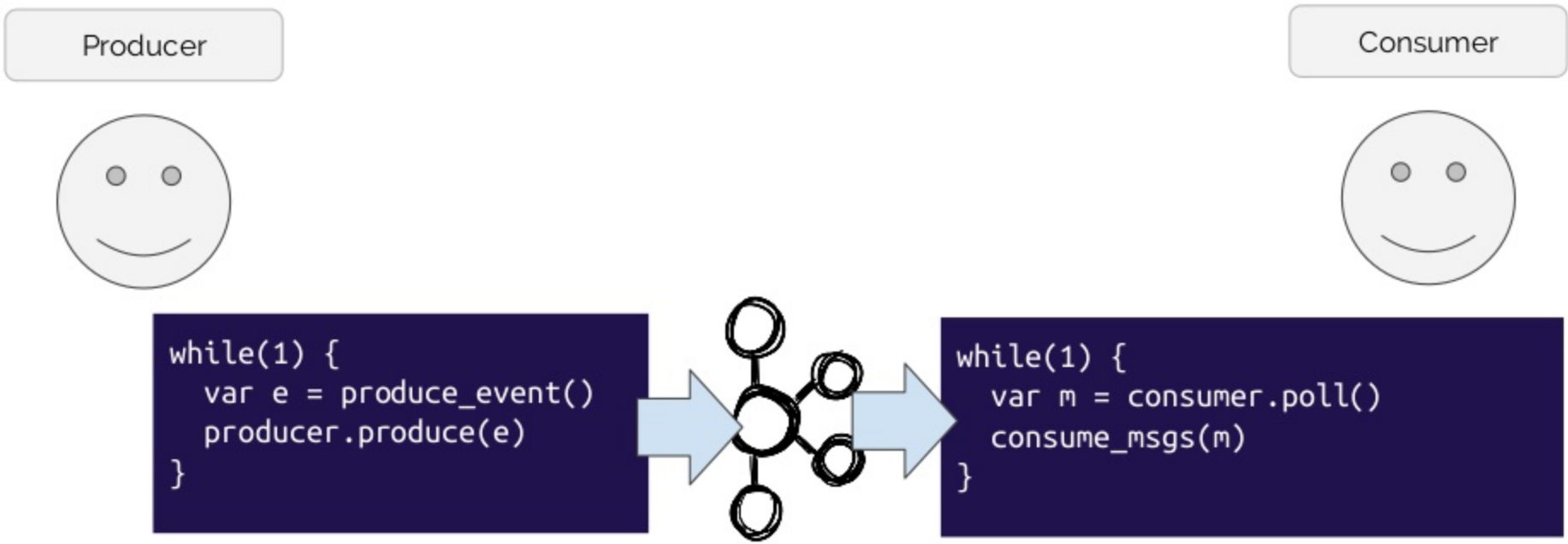
What is Kafka?

Data Pipeline with Kafka



What is Kafka?

Basics: Producer-Consumer Model



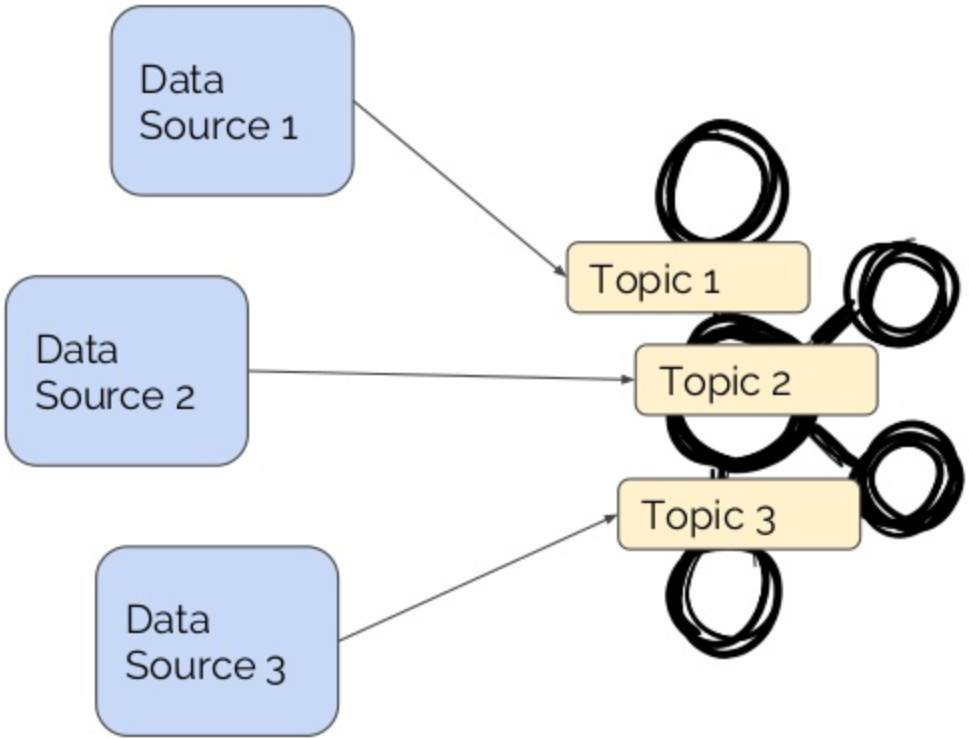
What is Kafka - terminology

Connect API

- For database / data source
- Wrapped consumer & producer code
- Nice thing: config file only!

Connect API - common connectors

JDBC - MySQL, PgSQL	S3
HDFS	ElasticSearch



Data Topic Model

- One-to-one (most common)

Feature

- Autonomous
  - Loads data from sources whenever changes occur
- Storage
  - Writes data to the hosted HDD
  - Optional: sync data to S3

Kafka Connect

```
name=test-source-sqlite-jdbc-autoincrement
connector.class=io.confluent.connect.jdbc.JdbcSourceConnector
tasks.max=1
connection.url=jdbc:sqlite:test.db
mode=incrementing
incrementing.column.name=id
topic.prefix=test-sqlite-jdbc-
```

Kafka Connect - Source Property File

Source: https://github.com/confluentinc/kafka-connect-jdbc/blob/master/config/source-quickstart-sqlite.properties

```
1  name=test-source-sqlite-jdbc-autoincrement
2  connector.class=io.confluent.connec
3  tasks.max=1
4  connection.url=jdbc:sqlite:test.db
5  mode=incrementing
6  incrementing.column.name=id
7  topic.prefix=test-sqlite-jdbc-
```

Topic naming convention

- Prefix, and
- DB table name

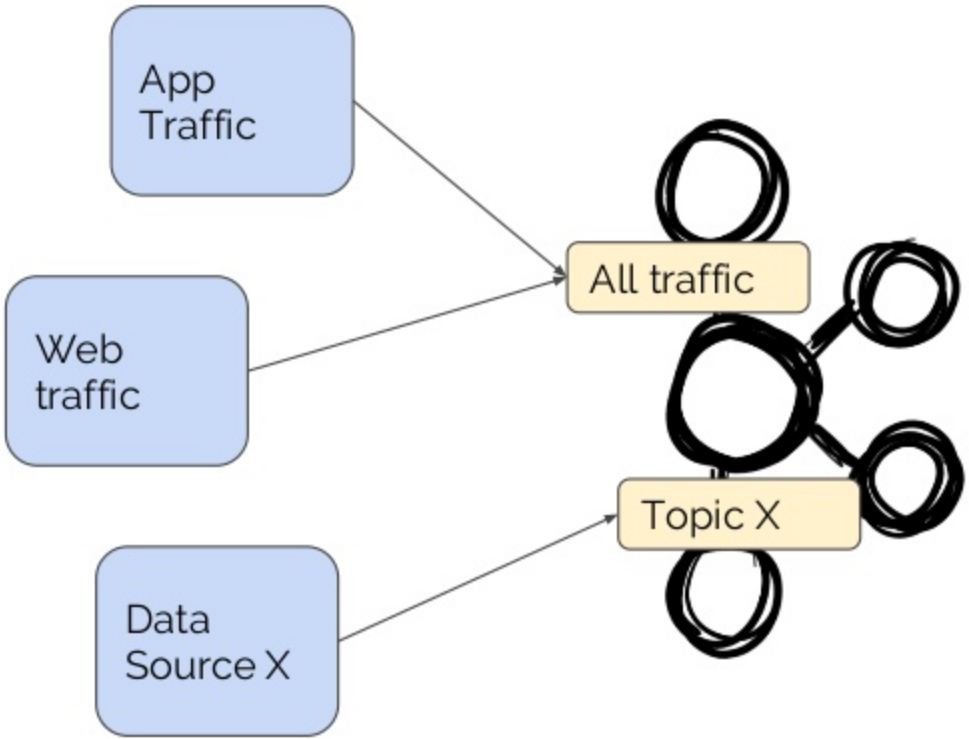
How it works:

- Each table implies one topic.

Kafka Connect - Source Property File

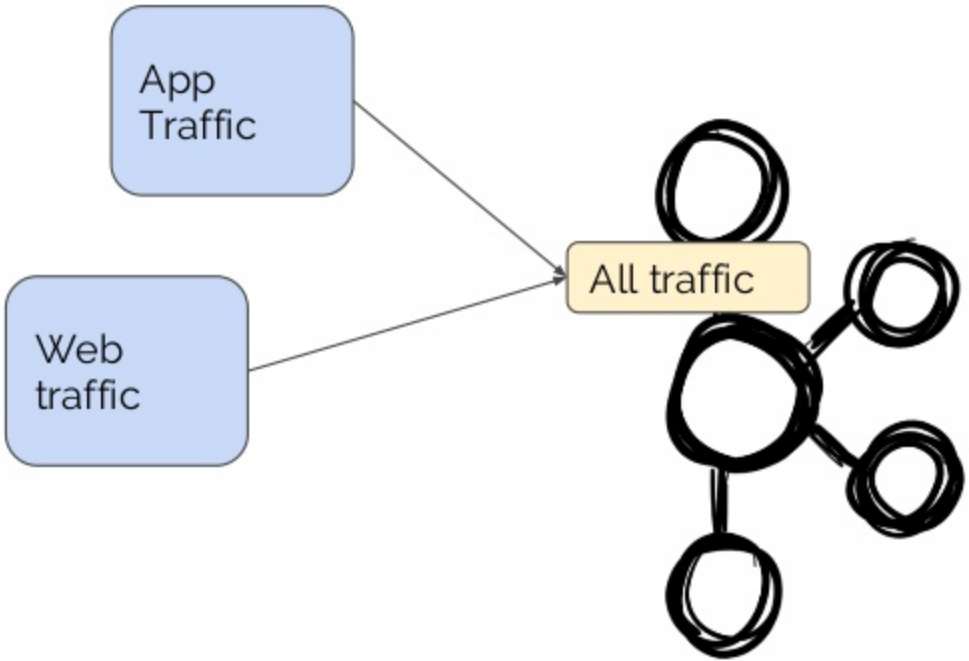
Source: https://github.com/confluentinc/kafka-connect-jdbc/blob/master/config/source-quickstart-sqlite.properties

Data Topic Model



- One-to-one (most common)
- Many-to-one

Kafka Connect

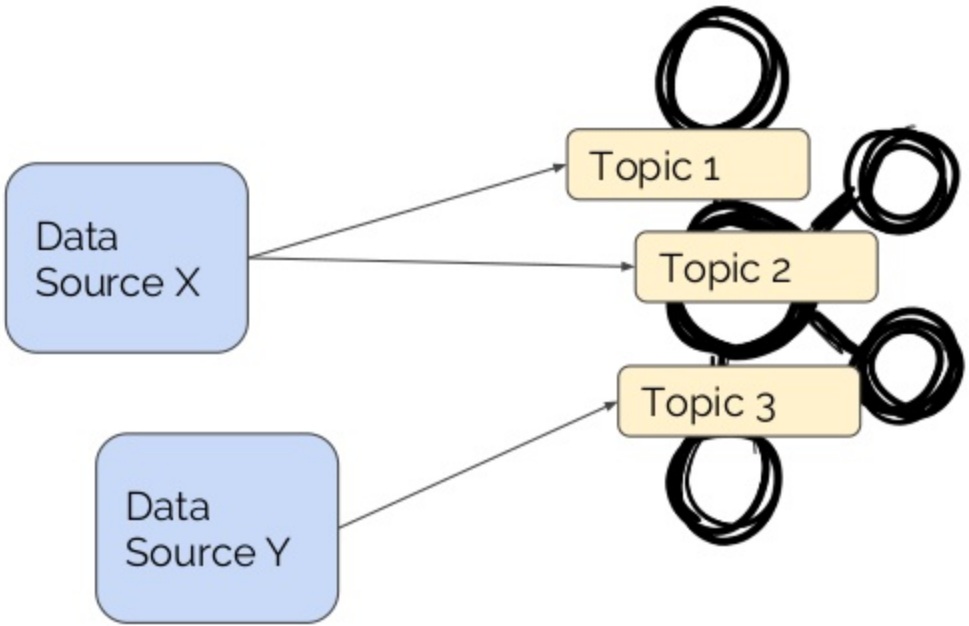


Schema-less

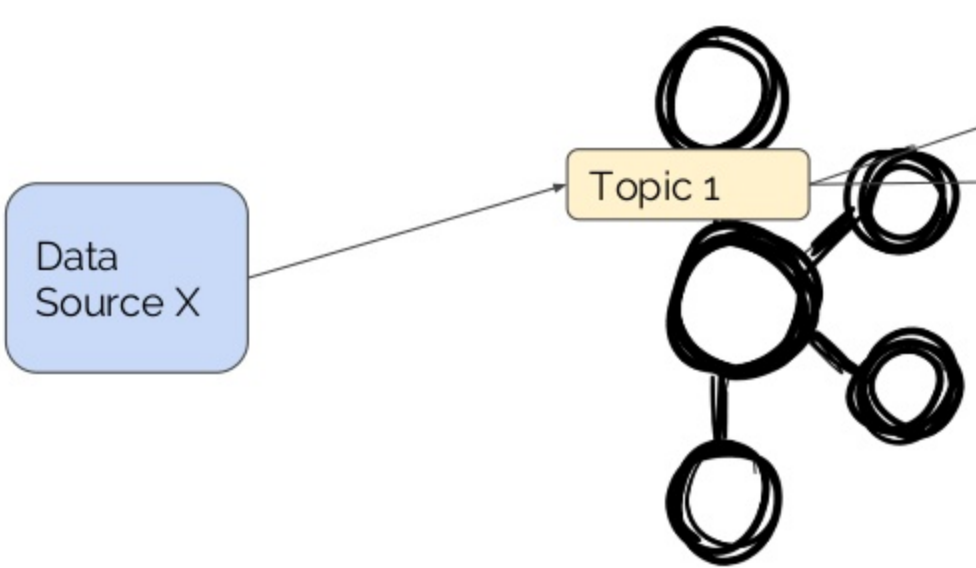
- Practically, you can write any types of data to the topic
- Most common choice is Avro

Btw, Avro is an open-source library for schema specification and data serialization.

Data Topic Model



- One-to-one (most common)
- Many-to-one
- One-to-many (most rare)



A more practical approach

- Use the same truth / data
- Consume by multiple guys!

Takeaway Messages

- Producers and consumers are actors
  - Push data to or pull data from Kafka

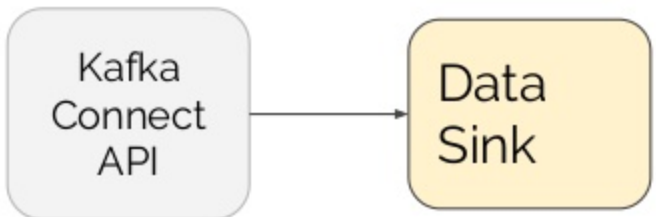
- Connect API automates the above actions
  - Work nicely with databases

Data Pipeline Use Cases

Kafka Internal - consumer's state				
Consumer	Topic	Current Topic Position	Your last-read position	Lag behind by
hello_world	foobar	1080	1000	80

Kafka keeps track on consumer's state:

- A consumer can always resume work-in-progress
- New consumer can start fresh!



Kafka as a data pipeline - data resiliency

```
$ /usr/bin/kafka-consumer-groups --zookeeper zk01.example.com:2181 --describe --group

GROUP    TOPIC  PARTITION  CURRENT-OFFSET  LOG-END-OFFSET  LAG      OWNER
flume    t1     0          1               3               2        test-consumer-group_
```

Source:
https://www.cloudera.com/documentation/kafka/latest/topics/kafka_command_line.html

JAVA
Or
Scala

```
.filter()
.map()
.reduce()
```

T

Kafka
Stream
API

Function	Use Case
`filter()`	Cleansing
`map()`	Transformation
`reduce()`	Aggregation

E

Data
Source

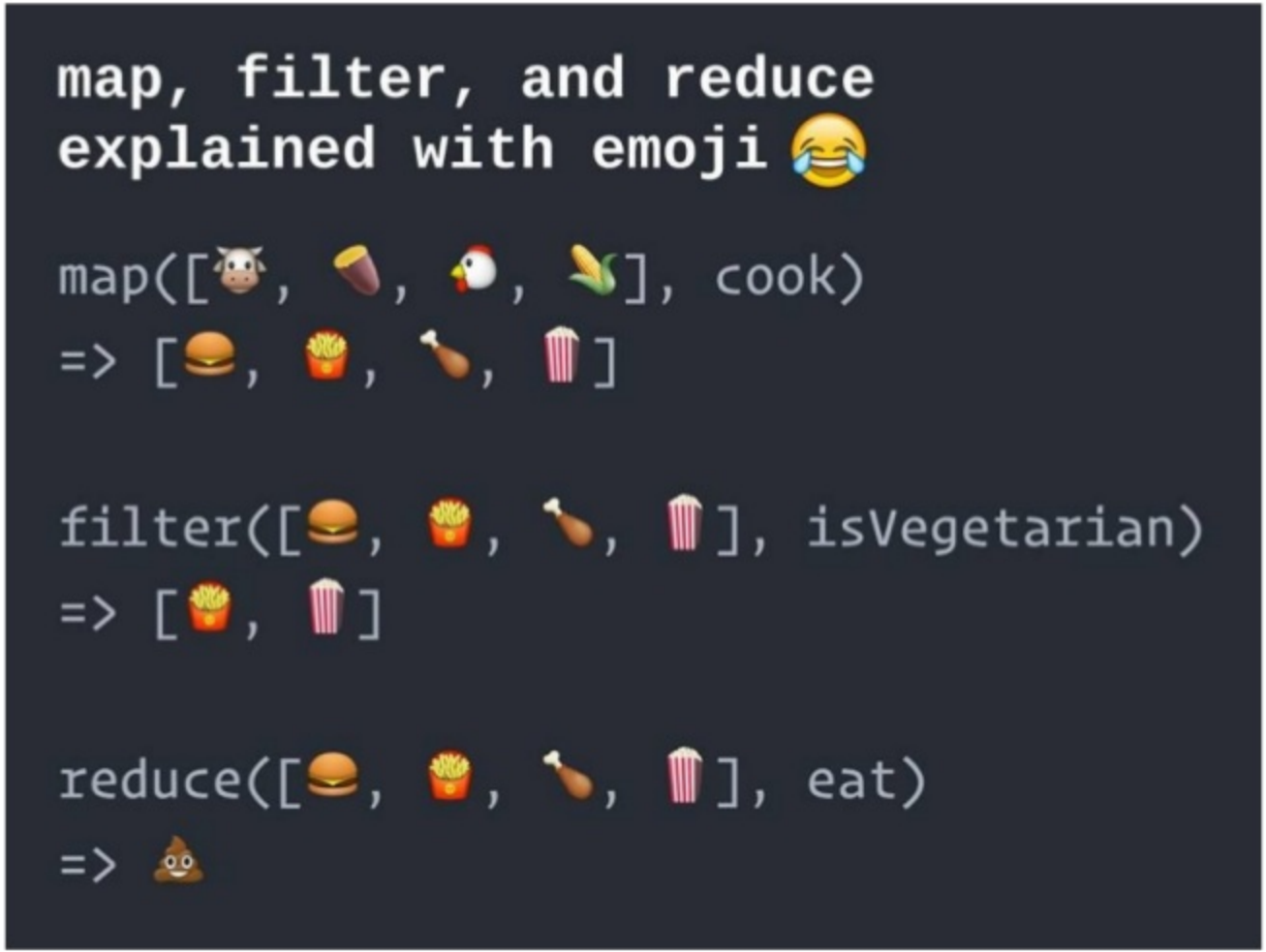
Kafka
Connect
API

L

Kafka
Connect
API

Data
Sink

Kafka as a data pipeline - Replace ETL



Source: https://i.redd.it/yf7rw3pjiapx.jpg