Building Data Pipelines with Open Source Components and Services

Heikki Nousiainen

Open Source Summit Japan 2018

Agenda

1. Introduction
2. Motivation
3. Challenges
4. Open Source
5. Data pipelines - Old and New
6. Build or Buy
7. Conclusions
8. Q & A

This presentation was created by Aiven Ltd - https://aiven.io.
Product and vendor logos used for identification purposes only.

Introduction

Speaker

- Heikki Nousiainen
- CTO, co-founder @ Aiven, a cloud DBaaS company
- Previously: software architect - cloud transformation, distributed systems
- Open source user and fan since 1995

@hnousiainen

UpCloud

DigitalOcean

Microsoft Azure

Google Cloud Platform

amazon web services

Aiven

- Independent Database as a Service provider
- Based in Helsinki and Boston
- 8 database systems available in 70+ regions around the world

https://aiven.io

@aiven_io

Motivation

Motivation

You've all heard it,

Data is your most valuable asset

"Data is the new oil"

Data is disrupting every industry

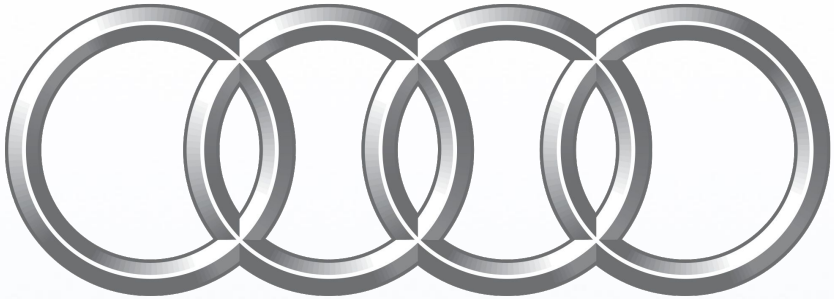
...but let's take a look at some actual uses

Motivation

Car-as-a-Sensor

Traffic and road condition detection & routing

Vacant parking space locator



Motivation

Welding Management

- Procedures
- Qualification verification
- 100% traceability

Motivation

Home / commercial automation

Smart Locks & Entry Controls

Environment sensing and management, lighting

Motivation

Predicting performance and proactively preventing downtime.

Pay-per-use models with SLA.

Fuel consumption management.

Challenges

Challenges

Area

- **Liveliness**
- Volume & Velocity
- Data/system lifespan
- Changing business requirements

Requirements

- Low latency / Real-time eventing
  - Interactive usage
  - Environmental awareness
  - Routing decisions
- Batch
  - Analytics
  - Reporting
  - Research

Challenges

Area

- Liveliness
- **Volume & Velocity**
- Data/system lifespan
- Changing business requirements

Requirements

- Billions of messages and terabytes of data 24/7
- 2013, 787 Dreamliner, 1TB data per flight. 150 units / year.
- 2018, Audi Concept, 4TB data per day per car. 2M units / year.

Challenges

Area

- Liveliness
- Volume & Velocity
- **Data/system lifespan**
- Changing business requirements

Requirements

- Production systems have long lifespans
  - Car ~15-20 years from design to disposal
  - Sea vessel 25-30 years
- Collected and consumed data differ
- Software / hardware upgrades

Challenges

Area

- Liveliness
- Volume & Velocity
- Data/system lifespan
- **Changing business requirements**

Requirements

- New services derived from data
- New sources/sinks for data
- Discontinued systems
- New experiments
- Legal landscape changes
- New/disbanded teams
- Acquisitions / integrations

Open Source

Open Source

- Open Source product development pace trumps that of any private undertaking
  - A lot of data management innovation happens in Open Source
  - Open Source is quick to absorb innovation from any source
  - The pragmatic evolutionary development cycles are efficient in improving quality

- Using Open Source guarantees continued access to business critical data
  - Avoid lock-in to a single vendor
  - Even with the 10 - 20 year lifespans

Data Pipelines - Old and New

Common components of a data pipeline

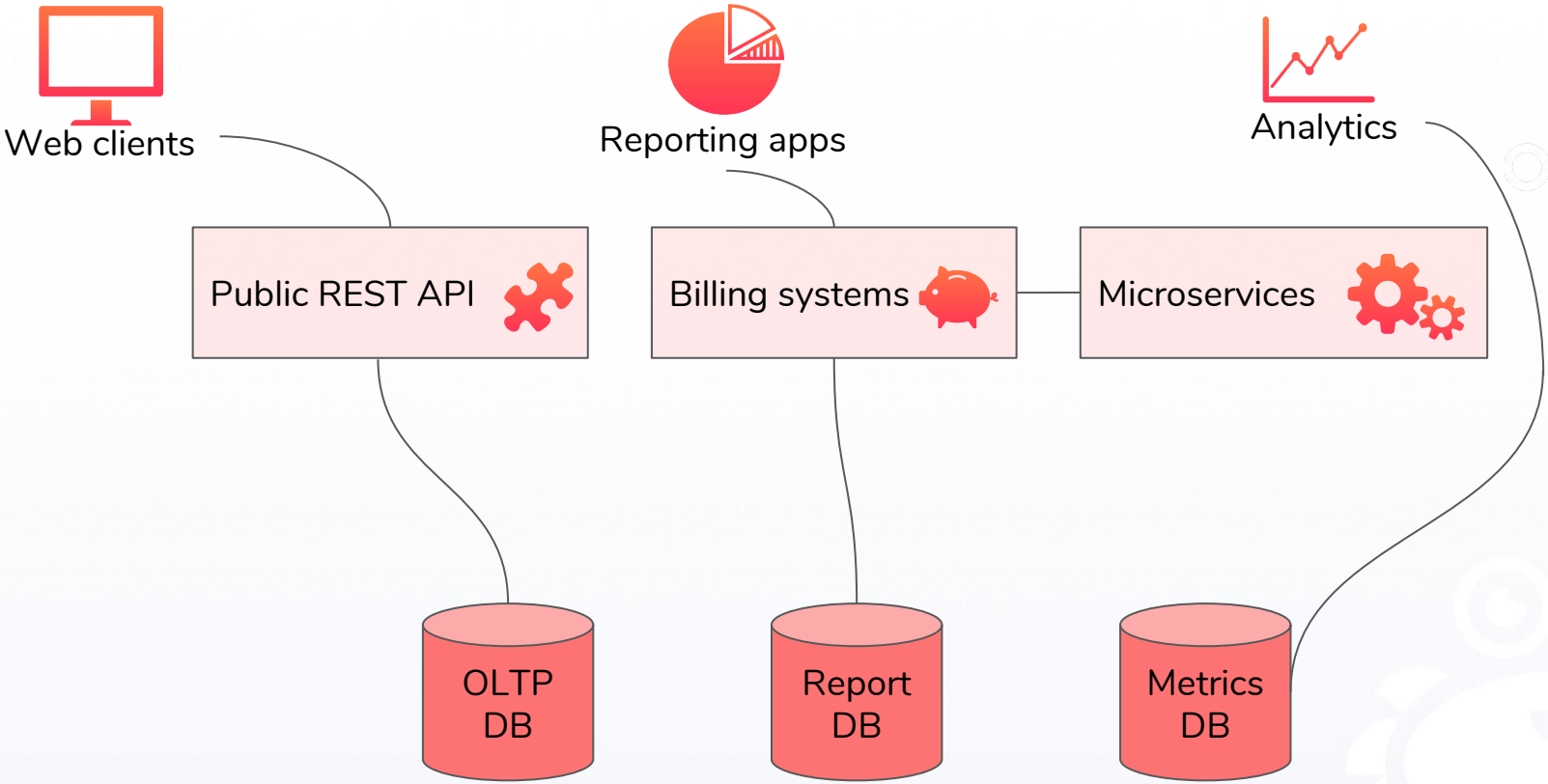
Typical parts of a data pipeline

- Data ingestion
- Filtering & Enrichment
- Routing
- Processing
- Querying / Visualization / Reporting
- Data warehousing
- Reprocessing capabilities

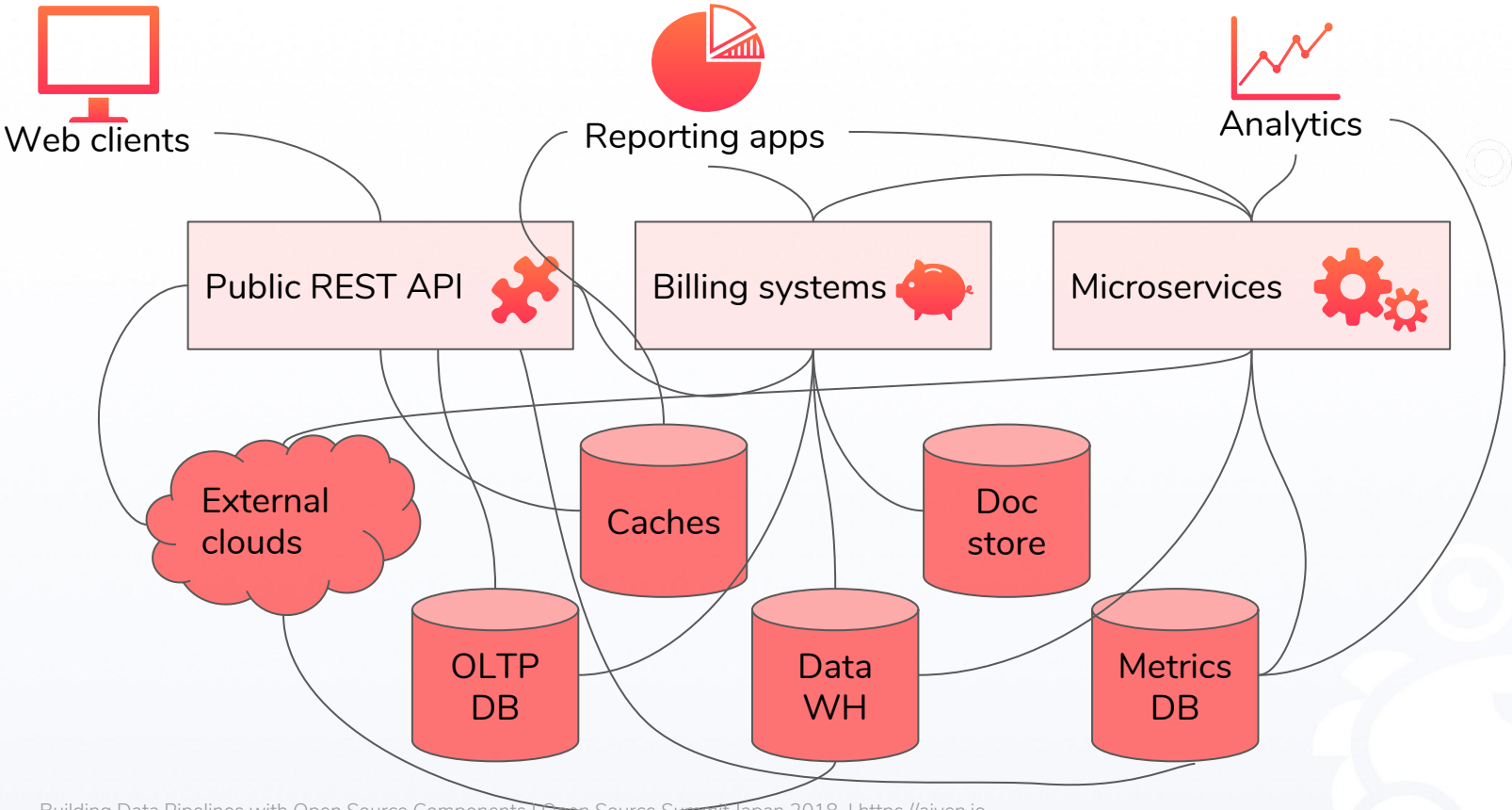
Typical requirements

- Scalability
  - Billions of messages and terabytes of data 24/7
- Availability and redundancy
  - Across physical locations
- Latency
  - Real-time / batch
- Adaptability / Platform support

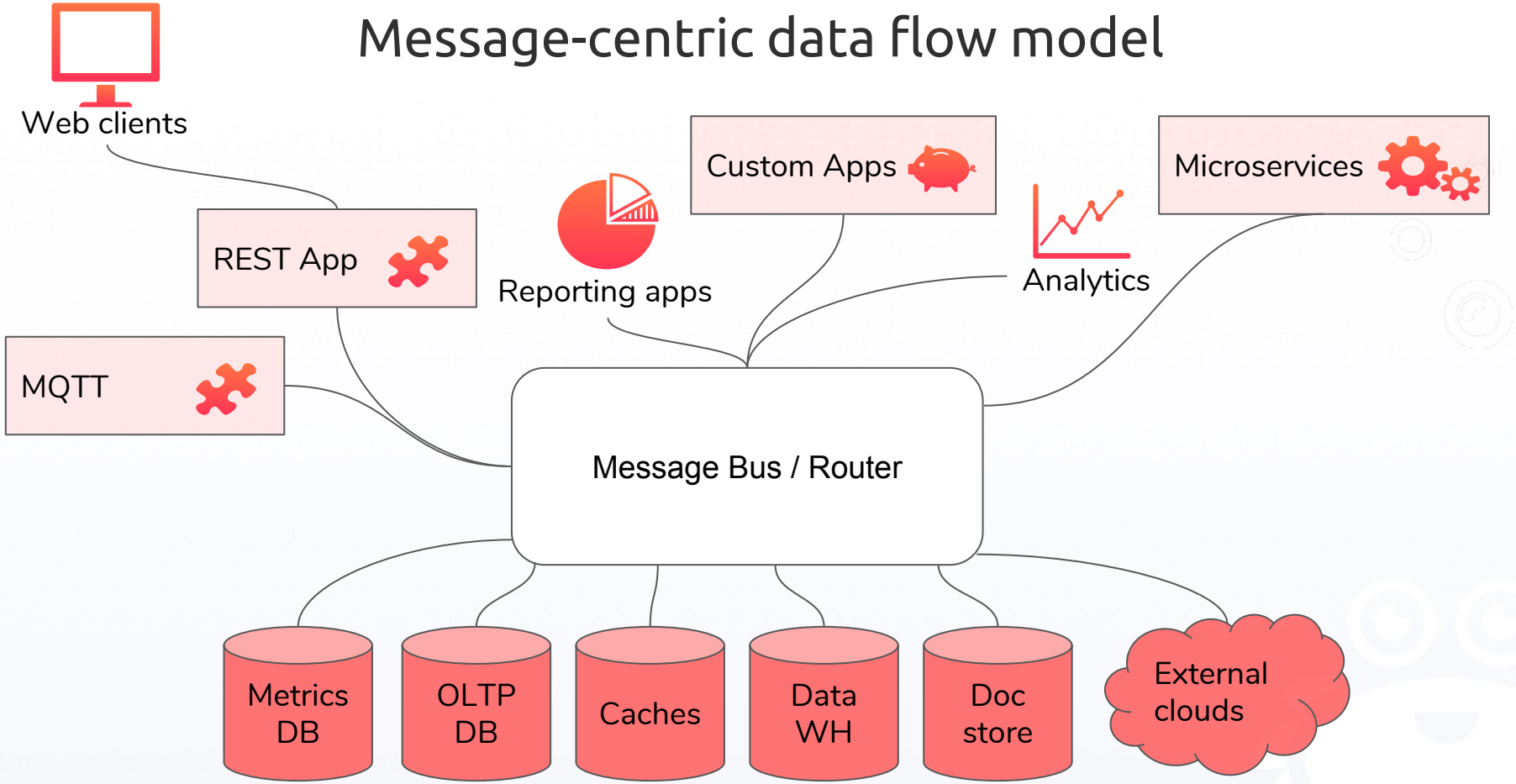
“Traditional” data flow model



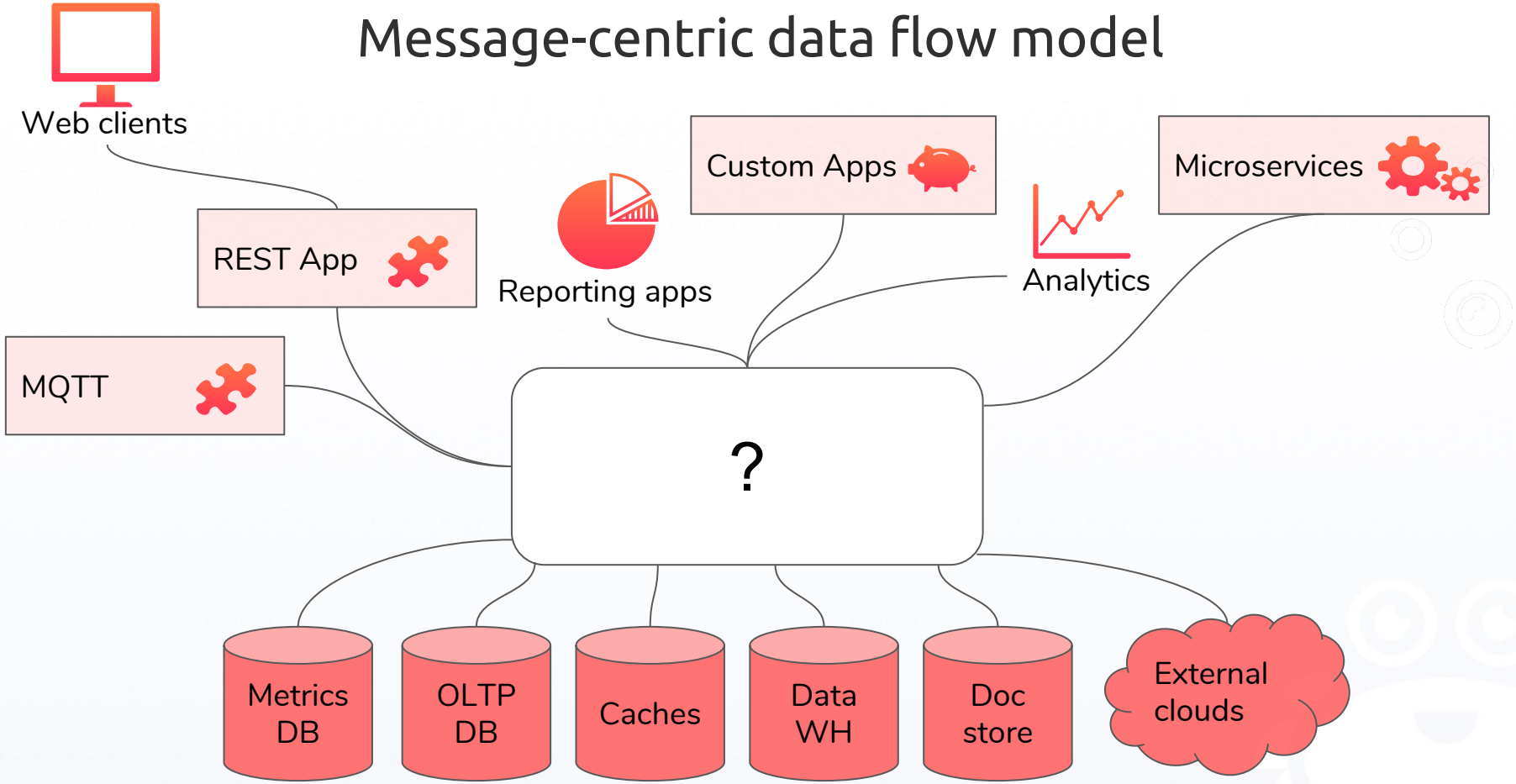
“Traditional” data flow model



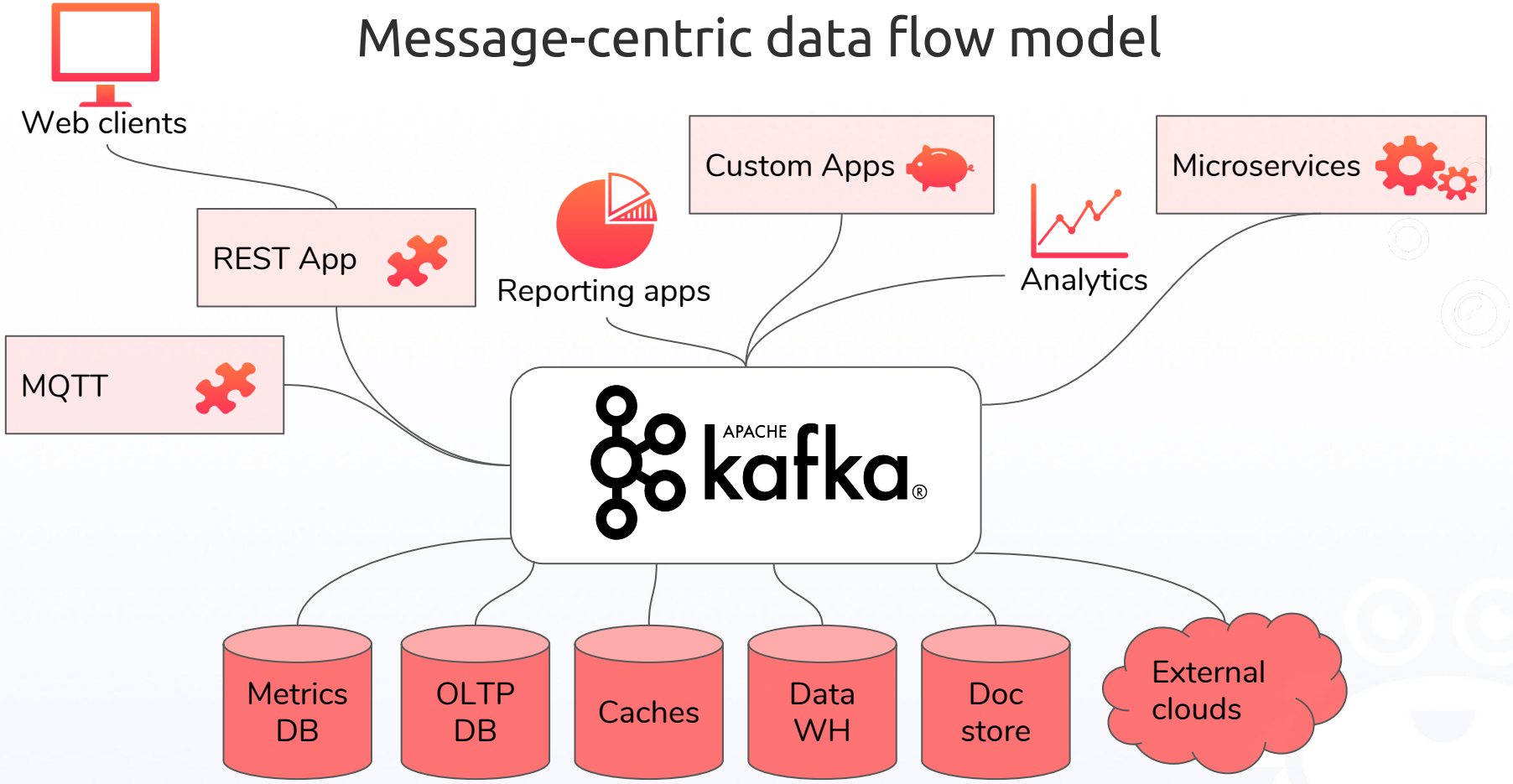
Message-centric data flow model



Message-centric data flow model



Message-centric data flow model



Apache Kafka

Apache Kafka is an open source stream processing platform.

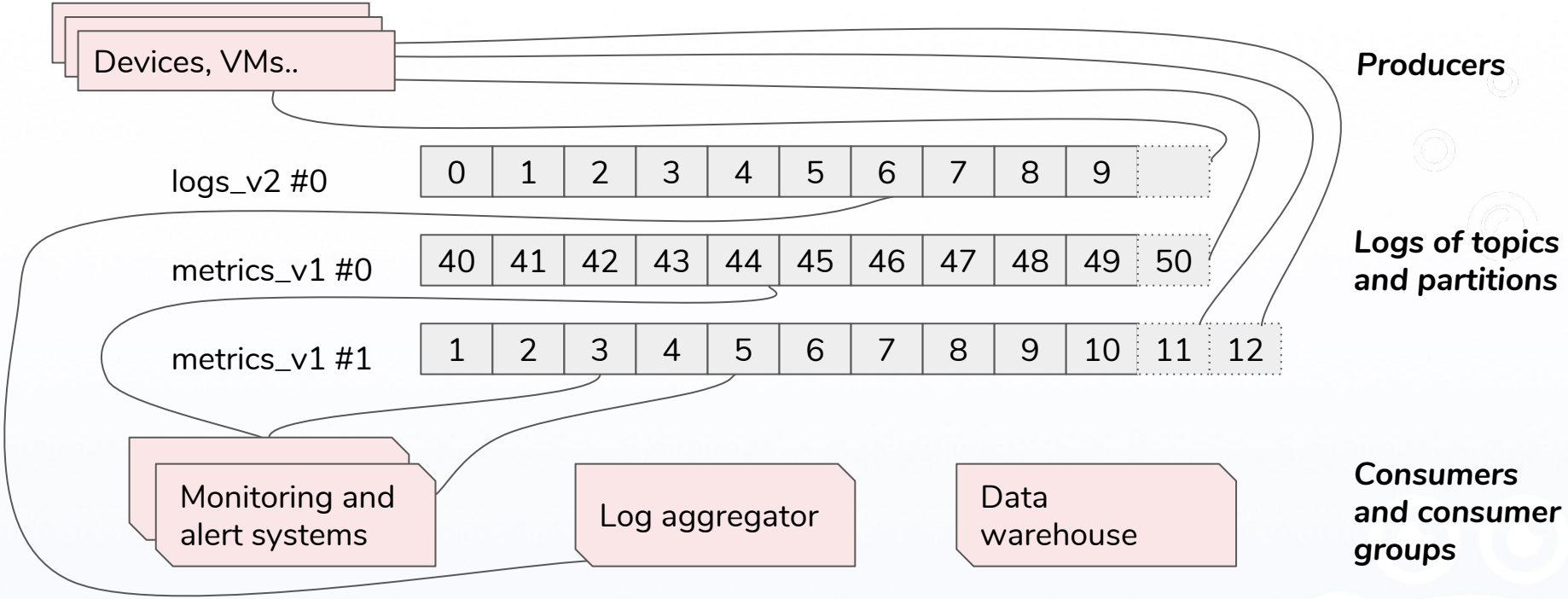
> "The project aims to provide a unified, high-throughput, low-latency platform for handling real-time data feeds."

Originally developed by LinkedIn, open sourced in 2011, now a top-level Apache project. Nowadays used by e.g. New York Times, Pinterest, Zalando, Airbnb, Shopify, Spotify and many others for event streaming. [See https://kafka.apache.org/powered-by for more.]

Kafka excels as a centerpiece for event delivery, where a range of applications can produce and consume real-time event streams.

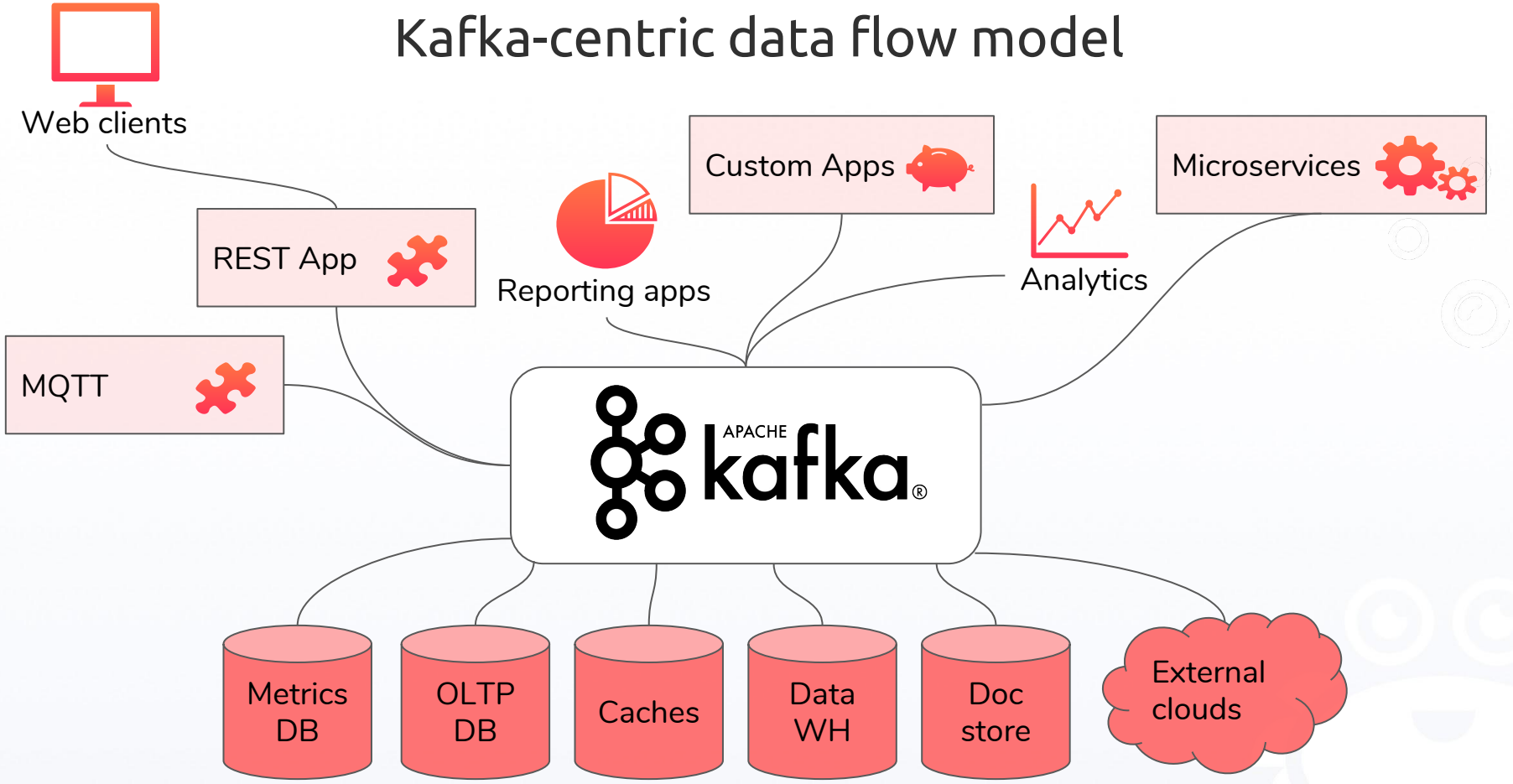
https://kafka.apache.org/

Kafka concepts



A key abstraction in Kafka is its commit log, where each consumer maintains maintains its own position in the log. This allows clean decoupling of the producing and consuming processes.

Kafka-centric data flow model



Kafka Connect

Framework for importing data from other systems and services - Sources - to Kafka and exporting to other services and systems - Sinks.

The framework makes it easy to create and share connectors in Open Source.

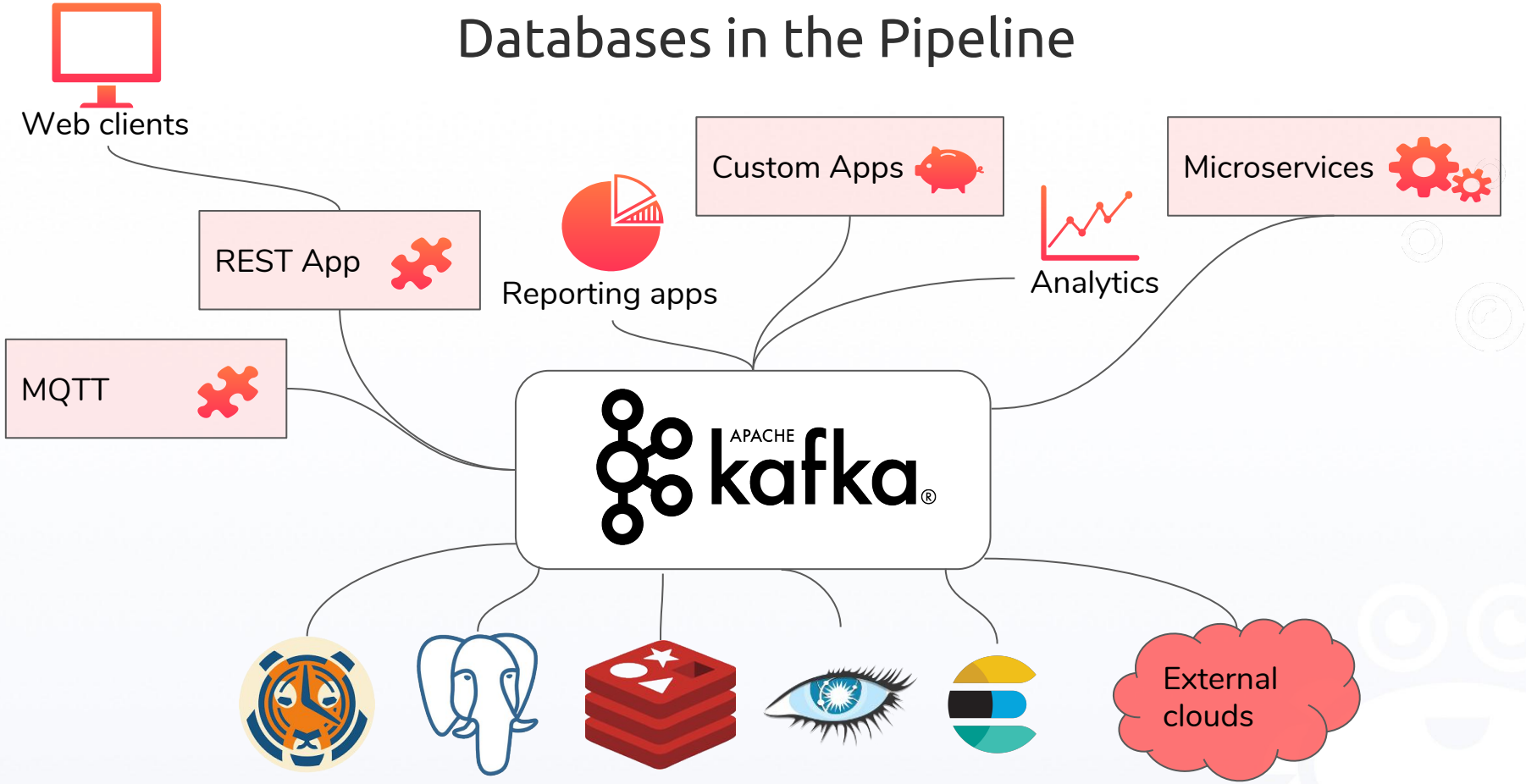
A host of connectors are available:

BigQuery, Cassandra, DynamoDB, Elasticsearch, Github, IOTHub / Azure, JDBC, JMS, Kinesis, PubSub / Google, MQTT, MySQL CDC, PostgreSQL CDC, RabbitMQ, Redshift, Redis, SalesForce, SAP, Solr, Splunk, SQS, Syslog, Twitter, Vertica

Databases in the Pipeline

- Specialized Open Source database technologies available for different use cases
- Consider the same requirements as for the streaming platform:
  - Access patterns: transactional, relational
  - Scalability
  - Reliability
  - Adaptability / Platform support / SDKs & Libraries
  - Available competencies

Databases in the Pipeline



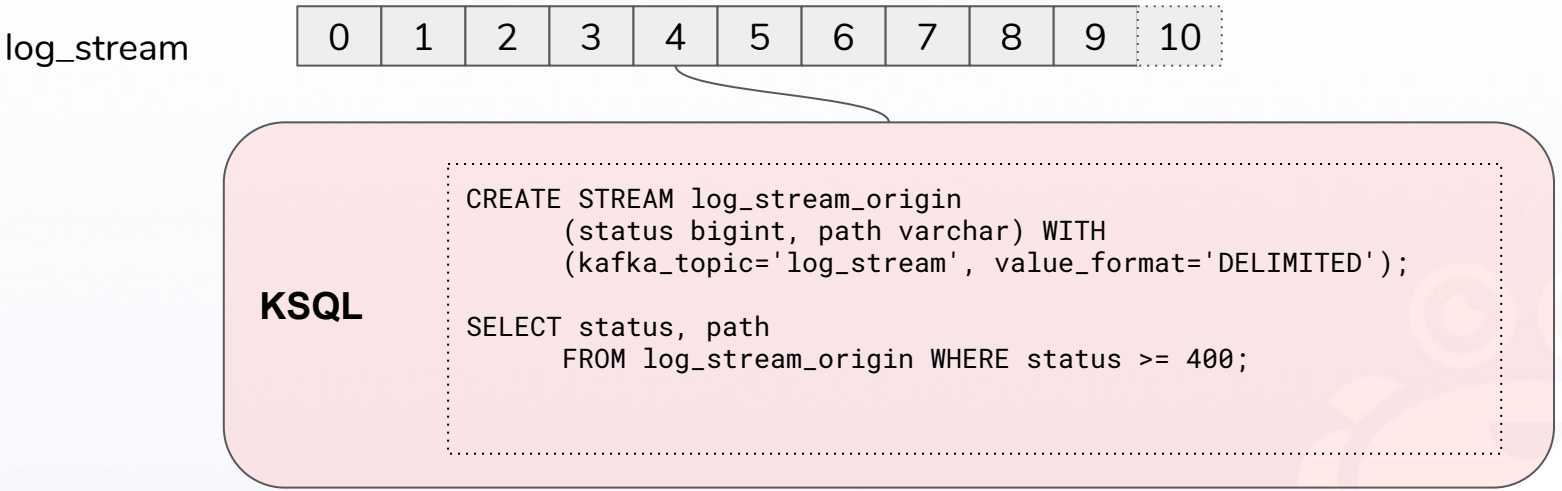
Kafka Streams

- Kafka Streams is an SDK / library for building application that process data in real-time
- DSL for defining streams & processing steps
- Supports abstraction for stream of events, but also tables and state.
- Stateless and stateful transformations

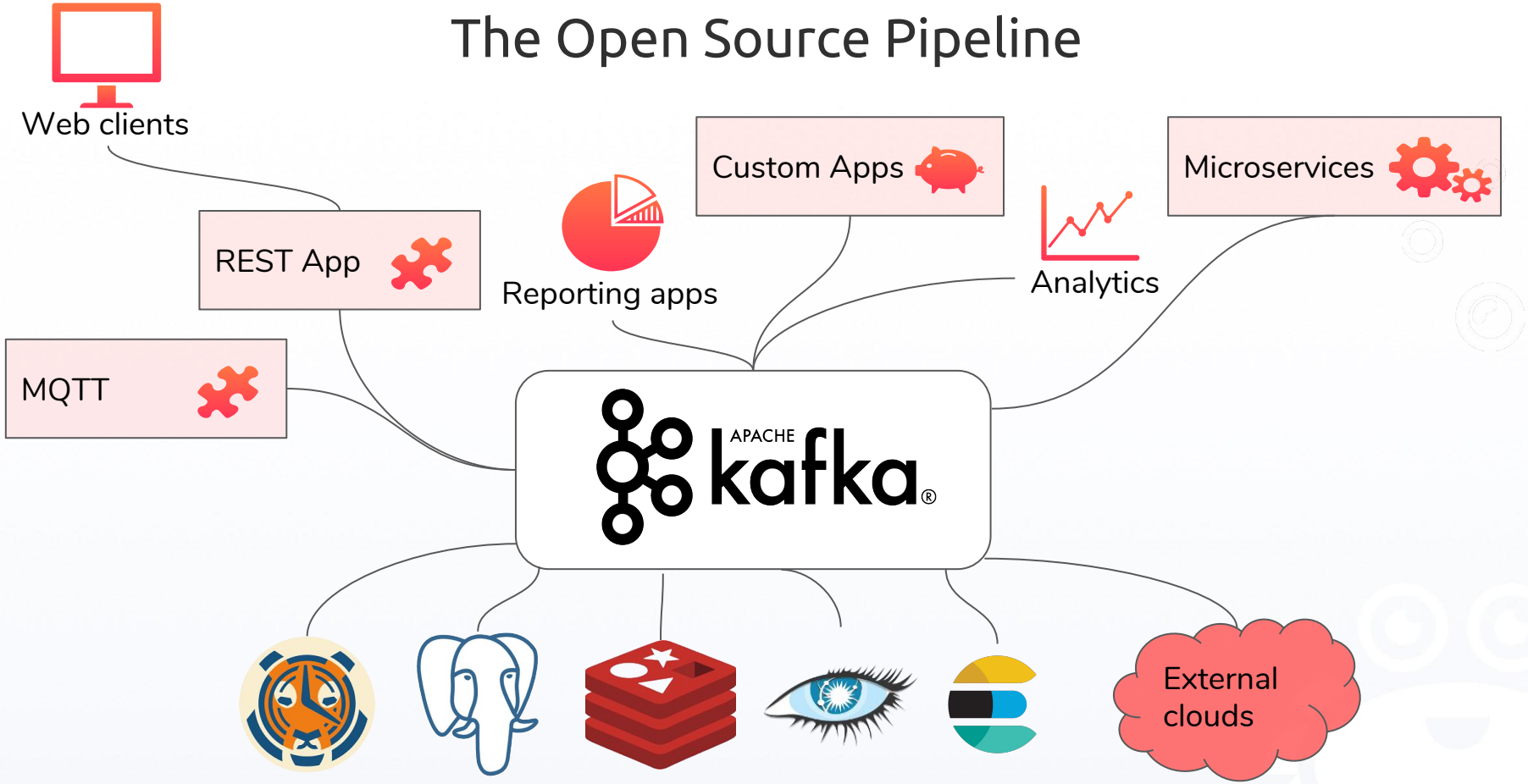
```
KStream<String, String> textLines = builder.stream("InputLinesTopic");
KTable<String, Long> wordCounts = textLines
    .flatMapValues(textLine -> Arrays.asList(textLine.toLowerCase().split("\\W+")))
    .groupBy((key, word) -> word)
    .count("Counts");
wordCounts.to(Serdes.String(), Serdes.Long(), "WordsWithCountsTopic");
```

KSQL: SQL engine for Kafka

- KSQL allows performing continuous queries and transformation using SQL syntax
- Standalone service using Kafka APIs typically running as its own cluster next to Kafka



The Open Source Pipeline



Build or Buy

Data management is hard to do well

- Management of stateful systems requires specialized personnel and 24/7 response capability.
- Failures are difficult to predict and can have extremely high impact on business.

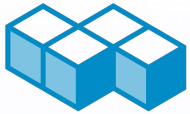
- Managed clouds services allow users to stay focused on their core business without worrying about software infrastructure.
- Open Source solutions allow one to move between in-house and managed offering.

Managed Open Source Solutions

aiven

confluent

instaclustr



COMPOSE
AN IBM COMPANY

Conclusions

Summary

- A lot of hype around data, but it's still real deal and to be taken seriously
- The challenges with data management are both technical and temporal
- Open Source is the best bet to meet the data management challenges
- Kafka as the central component of a data pipeline helps clean up messy architectures
- A host of good Open Source database solutions can help to meet the data storage and access needs
- You can leverage a host of managed service providers or build your own capability
- With Open Source, you have the option to revisit that choice at any time

aiven

Thanks!

https://aiven.io

@aiven_io

@hnousiainen