CONFLUENT

Serverless Kafka and Spark in a Multi-Cloud Lakehouse Architecture

Kai Waehner

Field CTO

kai.waehner@confluent.io
linkedin.com/in/kaiwaehner
@KaiWaehner
confluent.io
kai-waehner.de

KAI WAEHNER

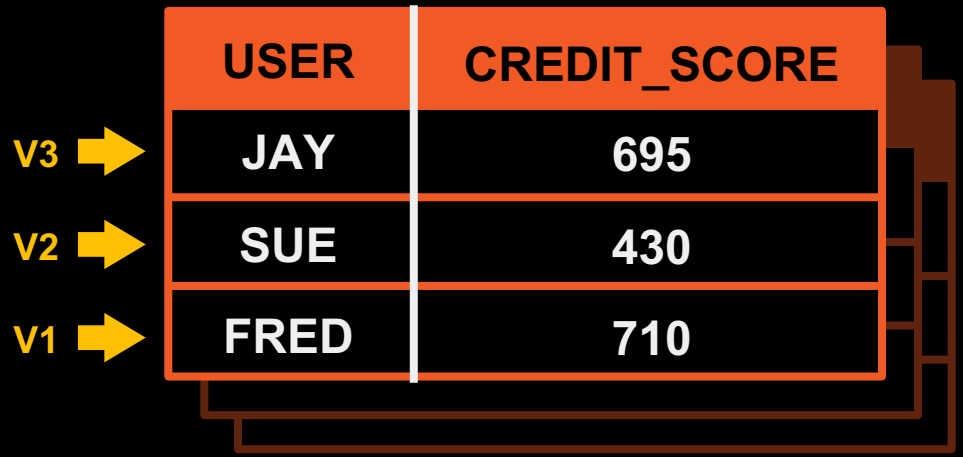
Agenda

- Data Analytics at Rest
- Data Streaming in Motion
- Lakehouse: Data Streaming + Analytics
- A Lakehouse Example: Intelligent Connected Cars
- Cloud-Native vs. Serverless Infrastructure
- Central vs. Hybrid and Global Data Mesh

Agenda

- **Data Analytics at Rest**
- Data Streaming in Motion
- Lakehouse: Data Streaming + Analytics
- A Lakehouse Example: Intelligent Connected Cars
- Cloud-Native vs. Serverless Infrastructure
- Central vs. Hybrid and Global Data Mesh

Storage at Rest



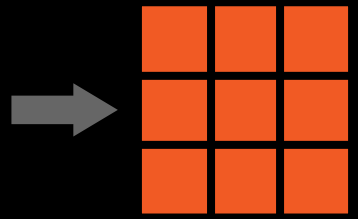
KAI WAEHNER

Analytics at Rest

Active Query:

```
SELECT * FROM
DB_TABLE
```

Passive Data:



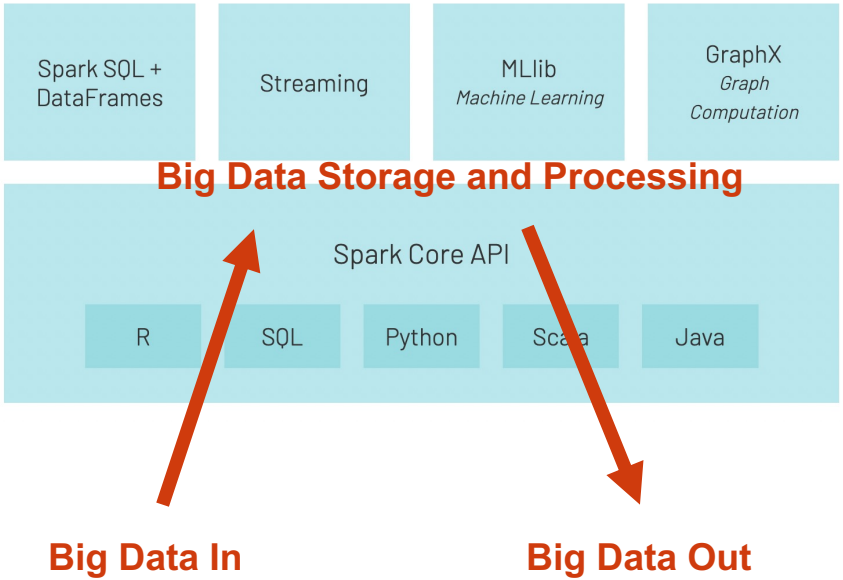
DB Table

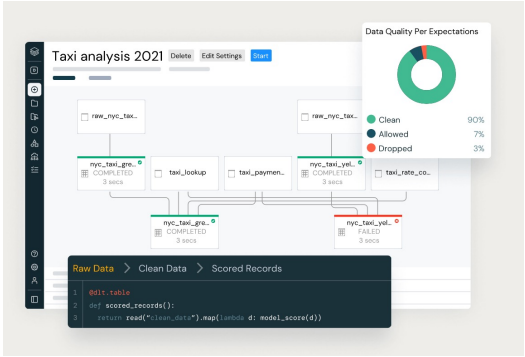
KAI WAEHNER

Use Cases for Data at Rest

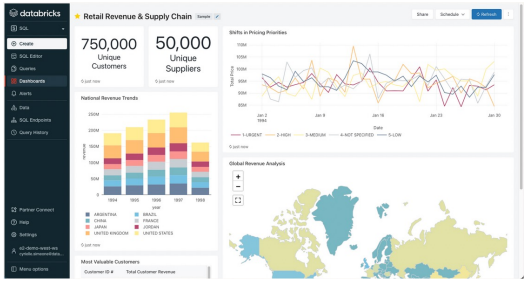
- Reporting
- Business Intelligence
- Data Engineering
- Big Data Analytics
- Machine Learning

Apache Spark – The De Facto Standard for Big Data at Rest



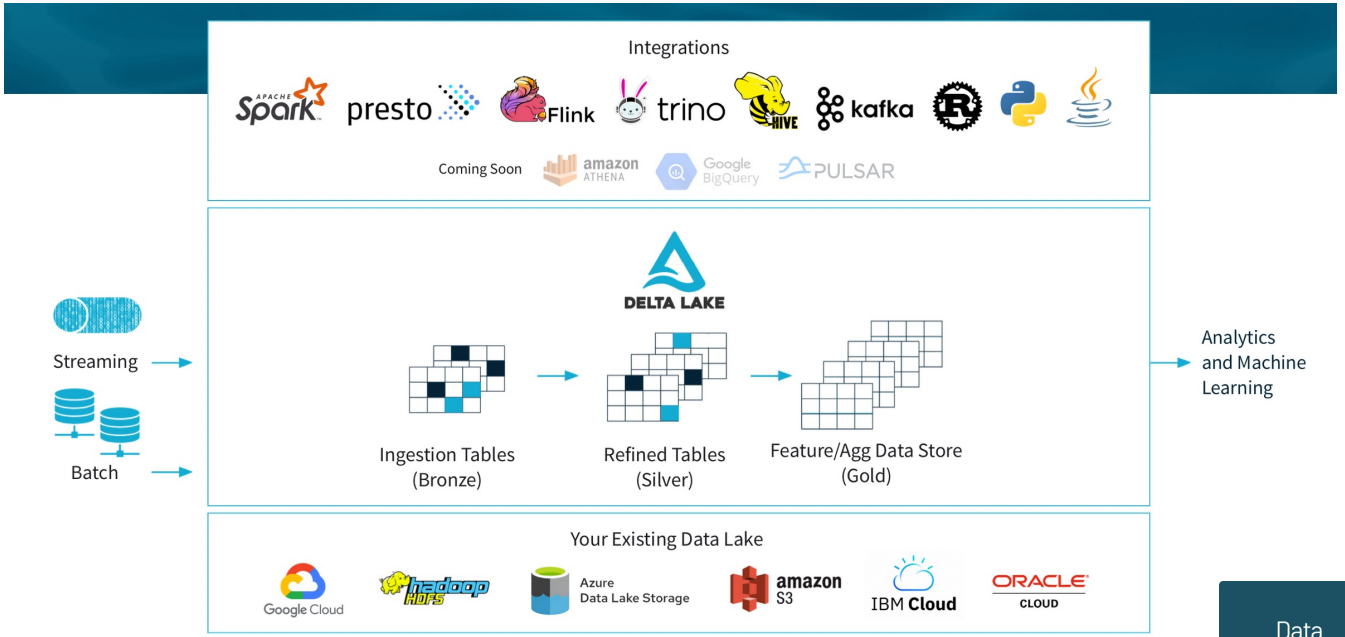


From Historical Data to Insights



Delta Lake

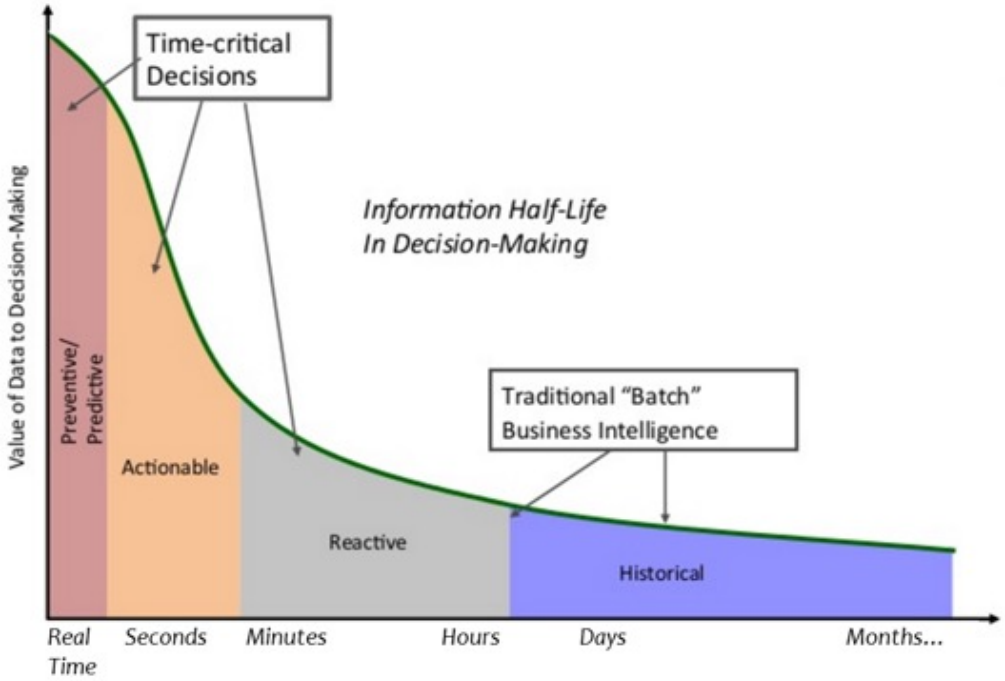
Open-source storage framework and open format for data analytics



Agenda

- Data Analytics at Rest
- **Data Streaming in Motion**
- Lakehouse: Data Streaming + Analytics
- A Lakehouse Example: Intelligent Connected Cars
- Cloud-Native vs. Serverless Infrastructure
- Central vs. Hybrid and Global Data Mesh

Real-time Data beats Slow Data.



Source: Perishable insights, Mike Gualtieri, Forrester

Real-time Data beats Slow Data.

Transportation

Real-time sensor diagnostics

Driver-rider match

ETA updates

Insurance

Claim processing

Fraud detection

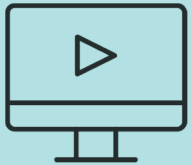
Omnichannel quote processing

Retail

Real-time inventory

Real-time POS reporting

Personalization



Entertainment

Real-time recommendations

Personalized news feed

In-app purchases

Data at Rest

Active Query:

```
SELECT * FROM
DB_TABLE
```

Passive Data:

DB Table

Data in Motion

Active Data:

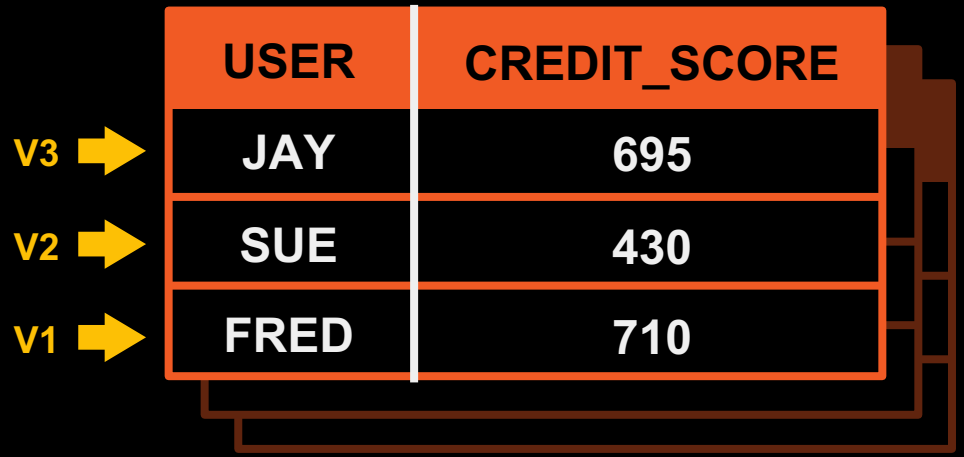
Event Stream

Passive Query:

```
CREATE TABLE T
AS SELECT * FROM
EVENT_STREAM
```

KAI WAEHNER

Tables at Rest



Streams in Motion

USER	PAYMENTS
JAY	42
SUE	18
FRED	65
...	...

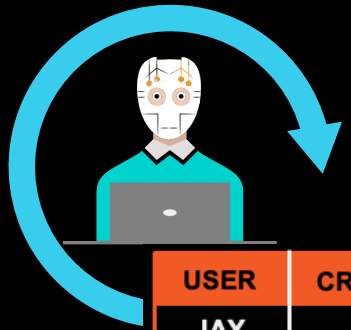
KAI WAEHNER

Data Streaming = Data at Rest + Data in Motion

USER	PAYMENTS
JAY	42
SUE	18
FRED	65
...	...

Payments Stream

```
CREATE TABLE credit_scores AS
  SELECT user, updateScore(p.amount)…
```

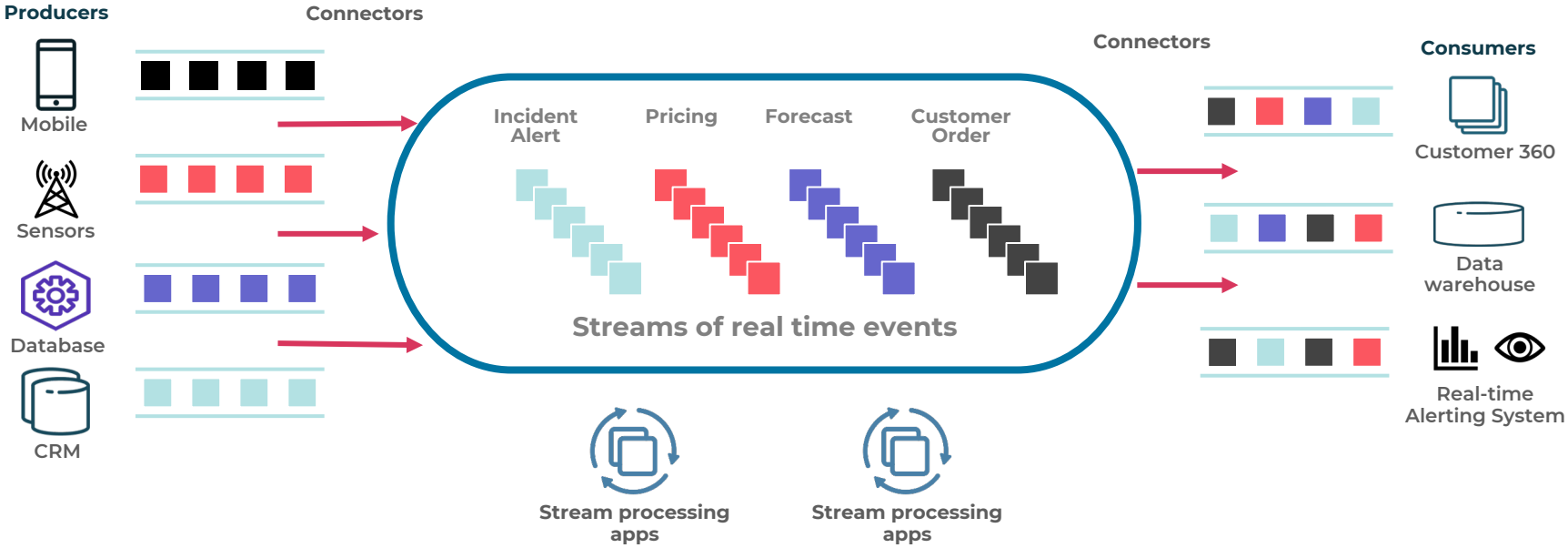


USER	CREDIT_SCORE
JAY	695
SUE	430
FRED	710

Credit Score Stream

KAI WAEHNER

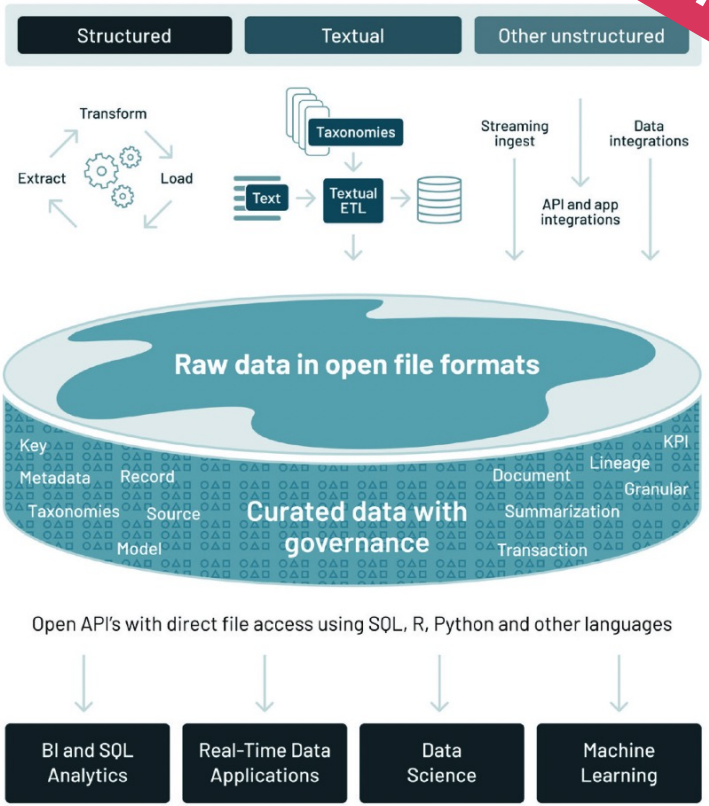
Apache Kafka – The De Facto Standard for Data in Motion



Agenda

- Data Analytics at Rest
- Data Streaming in Motion
- **Lakehouse: Data Streaming + Analytics**
- A Lakehouse Example: Intelligent Connected Cars
- Cloud-Native vs. Serverless Infrastructure
- Central vs. Hybrid and Global Data Mesh

Data Lakehouse

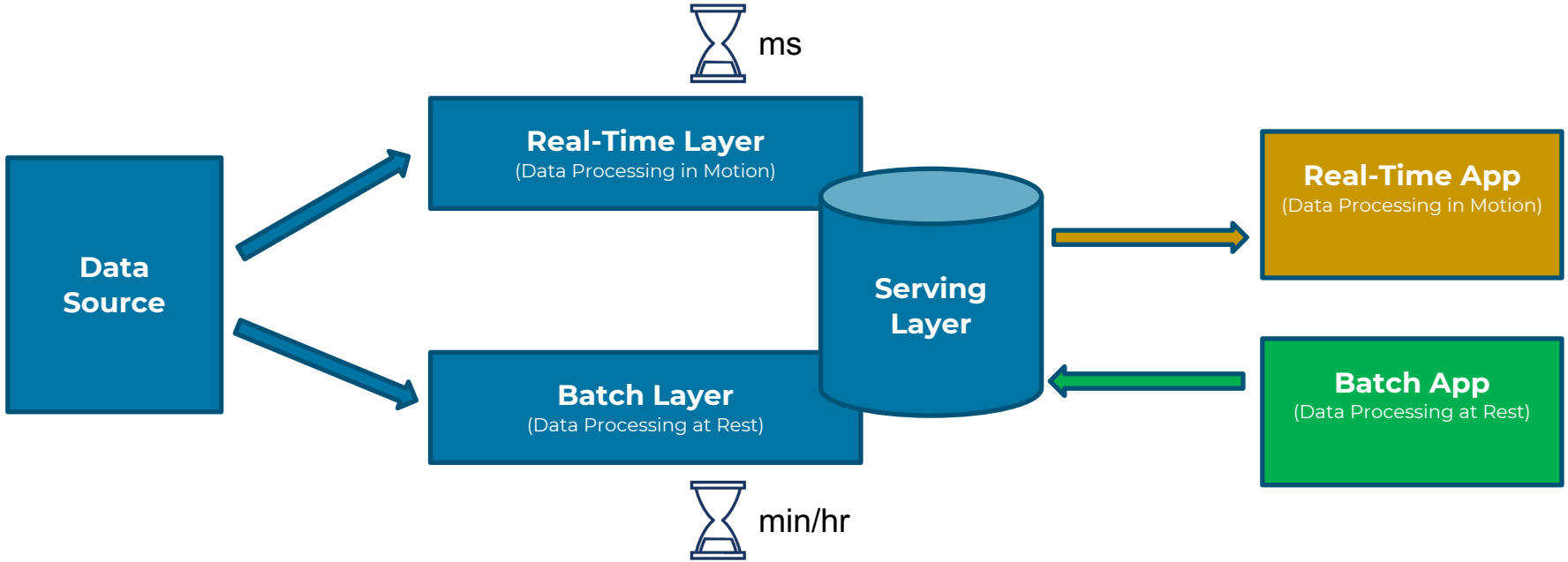


Lakehouse is a logical view, not physical!

Building the Data Lakehouse
Author: Bill Inmon

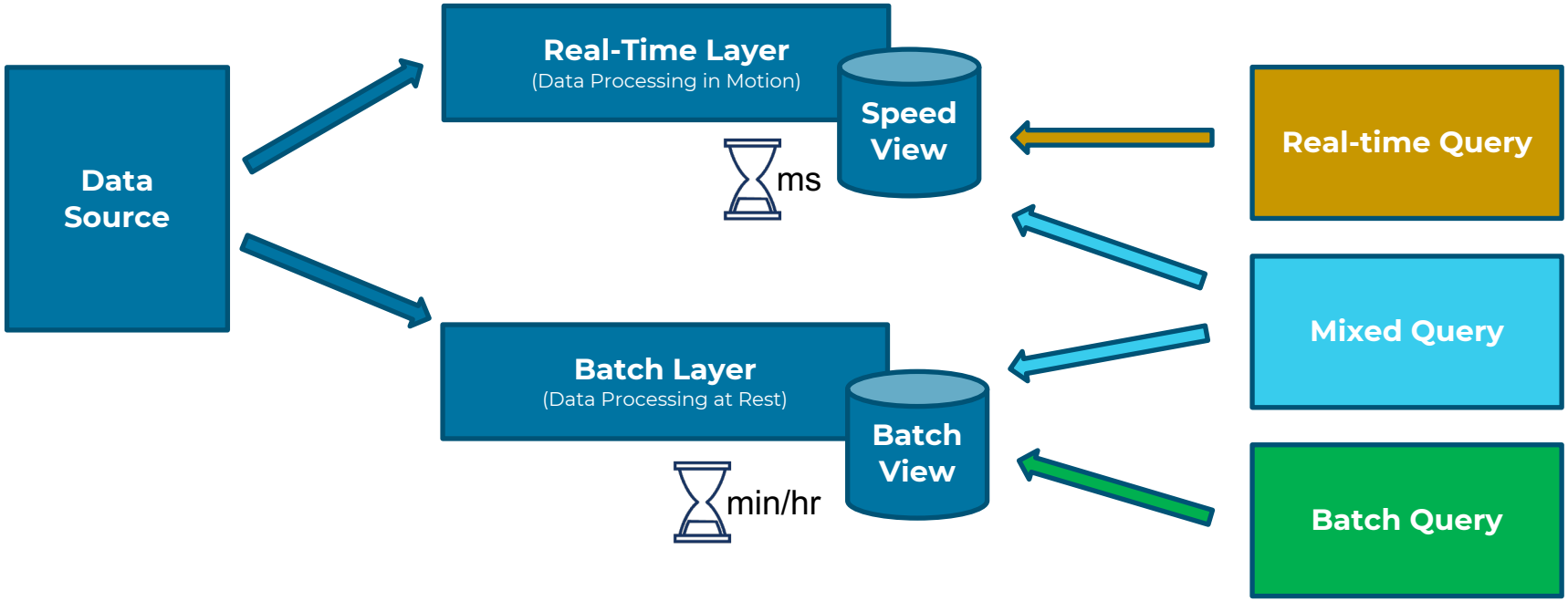
Lambda Architecture

Option 1: Unified serving layer



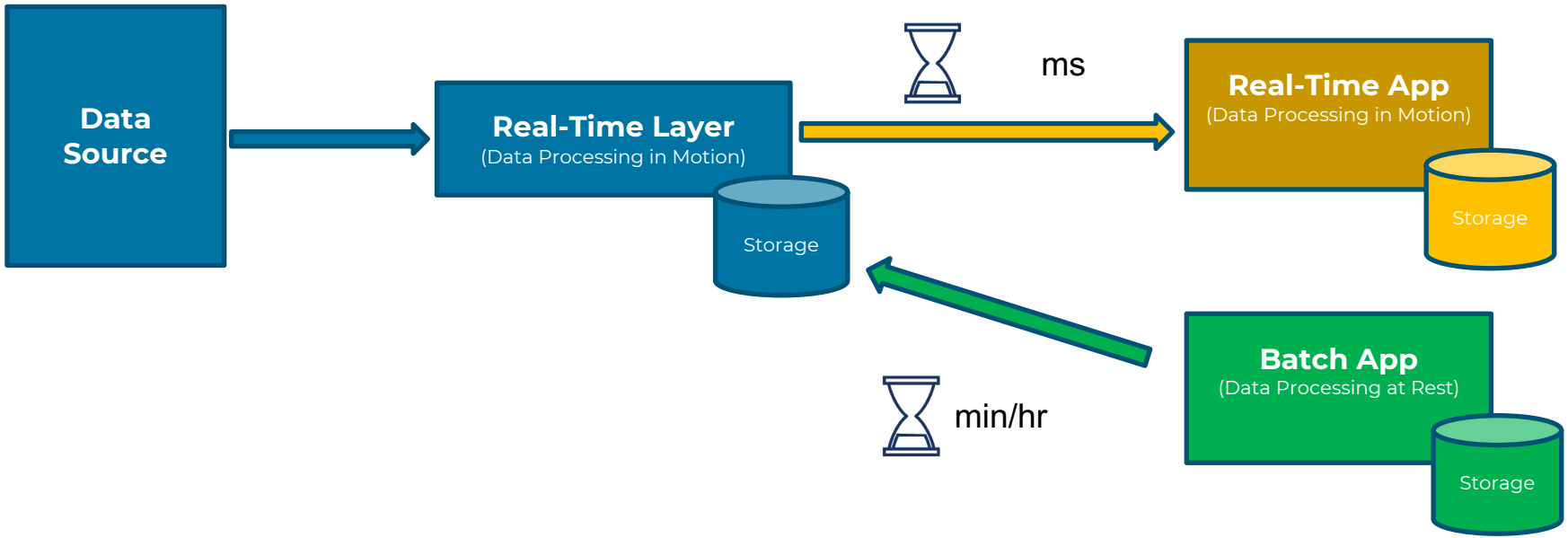
Lambda Architecture

Option 2: Separate serving layers



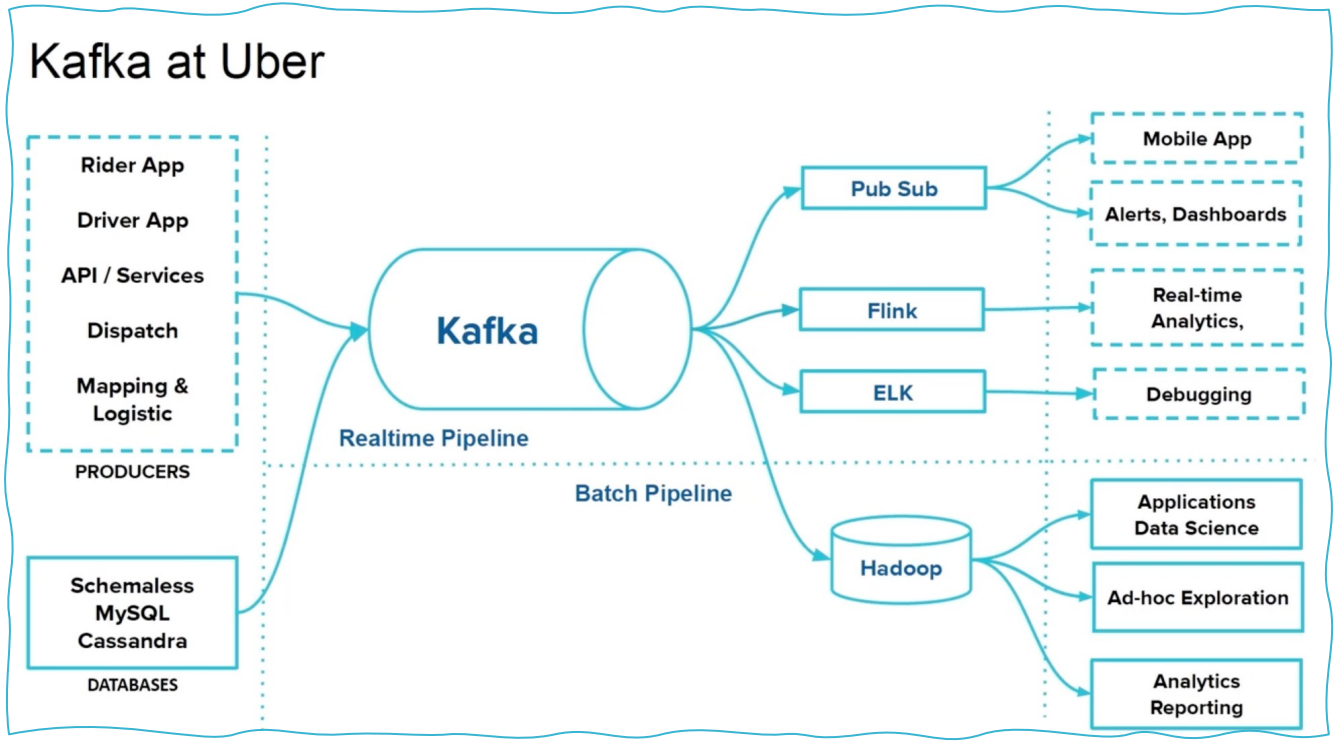
Kappa Architecture

One pipeline for real-time and batch consumers

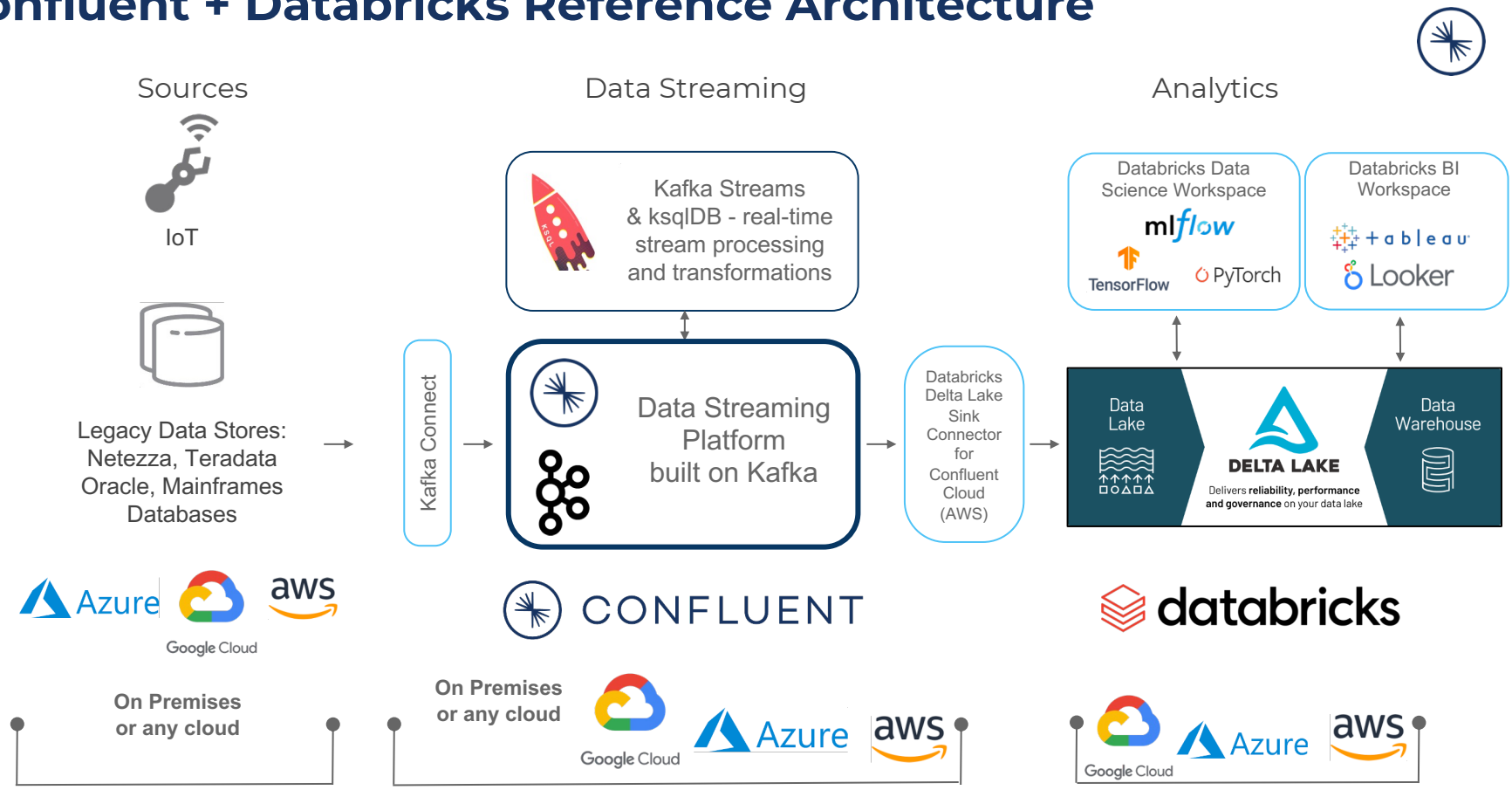


Kappa @ Uber

Uber



Confluent + Databricks Reference Architecture



Agenda

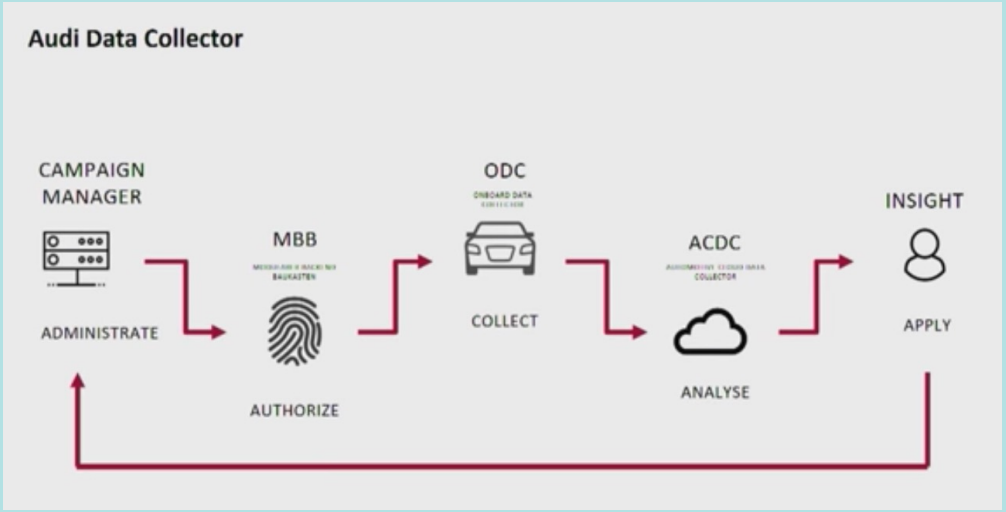
- Data Analytics at Rest
- Data Streaming in Motion
- Lakehouse: Data Streaming + Analytics
- **A Lakehouse Example: Intelligent Connected Cars**
- Cloud-Native vs. Serverless Infrastructure
- Central vs. Hybrid and Global Data Mesh

Connected Car Infrastructure at Audi

Autonomous vehicles like 'Jack' stream roughly

4 terabytes

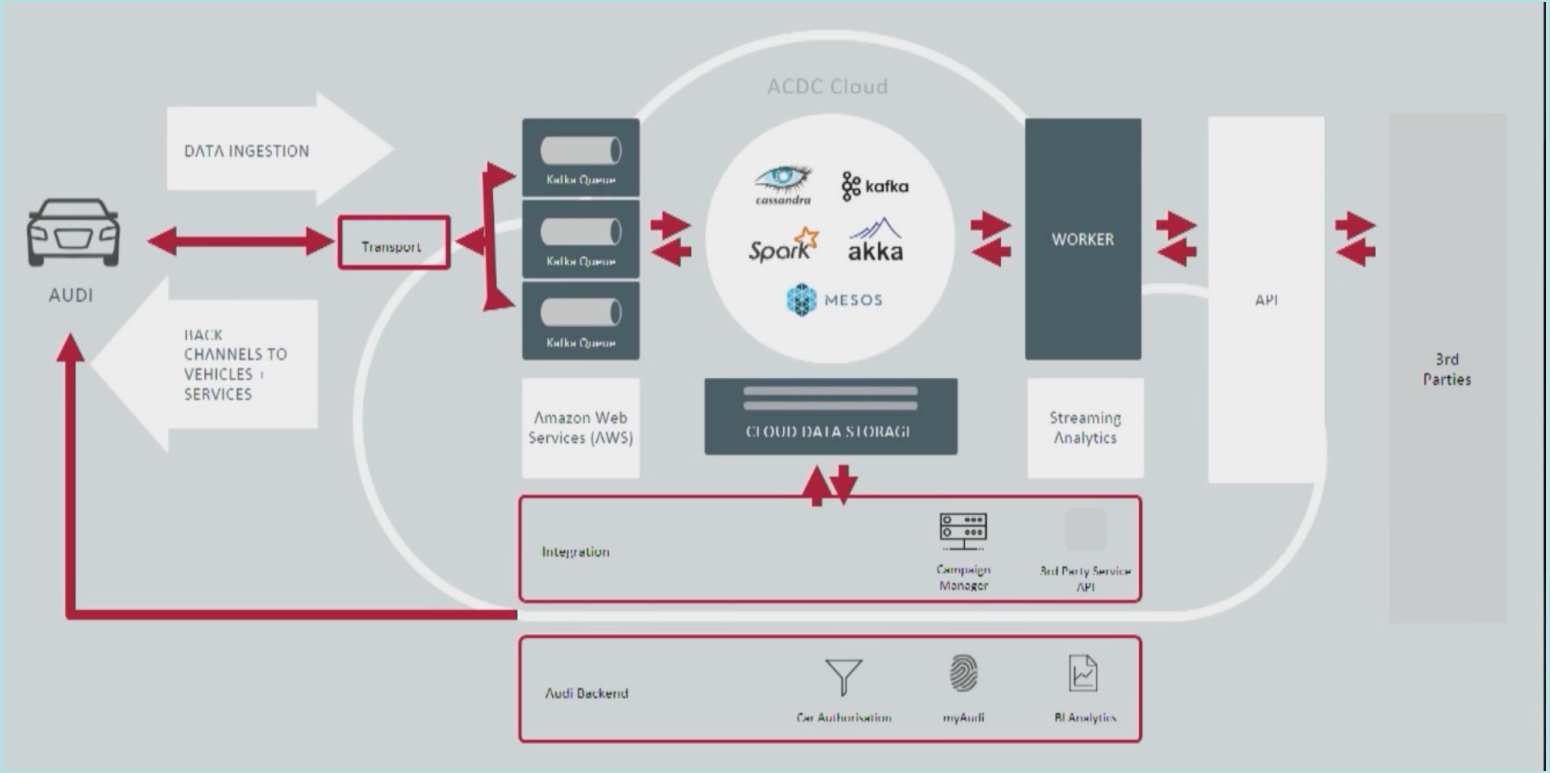
of data a day.



- Real Time Data Analysis
- Swarm Intelligence
- Collaboration with Partners
- Predictive AI
- ...

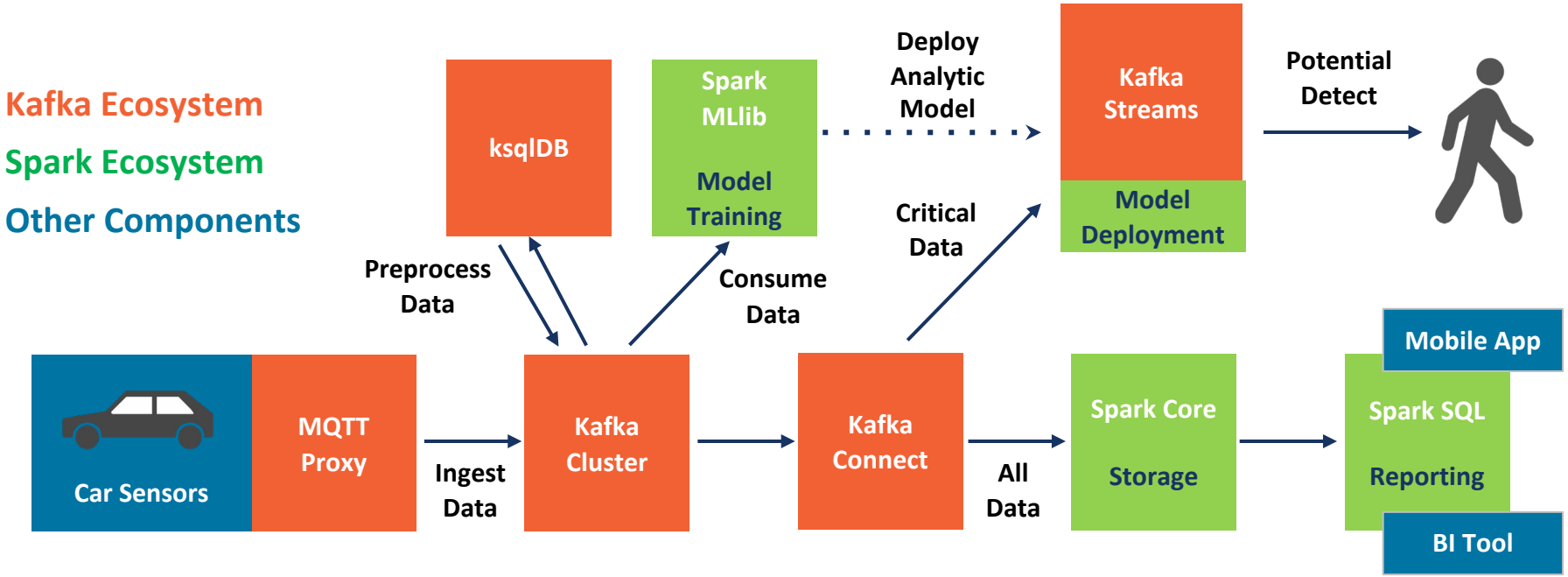
https://www.youtube.com/watch?v=vGLKi3TMJv8

Connected Car Infrastructure at Audi

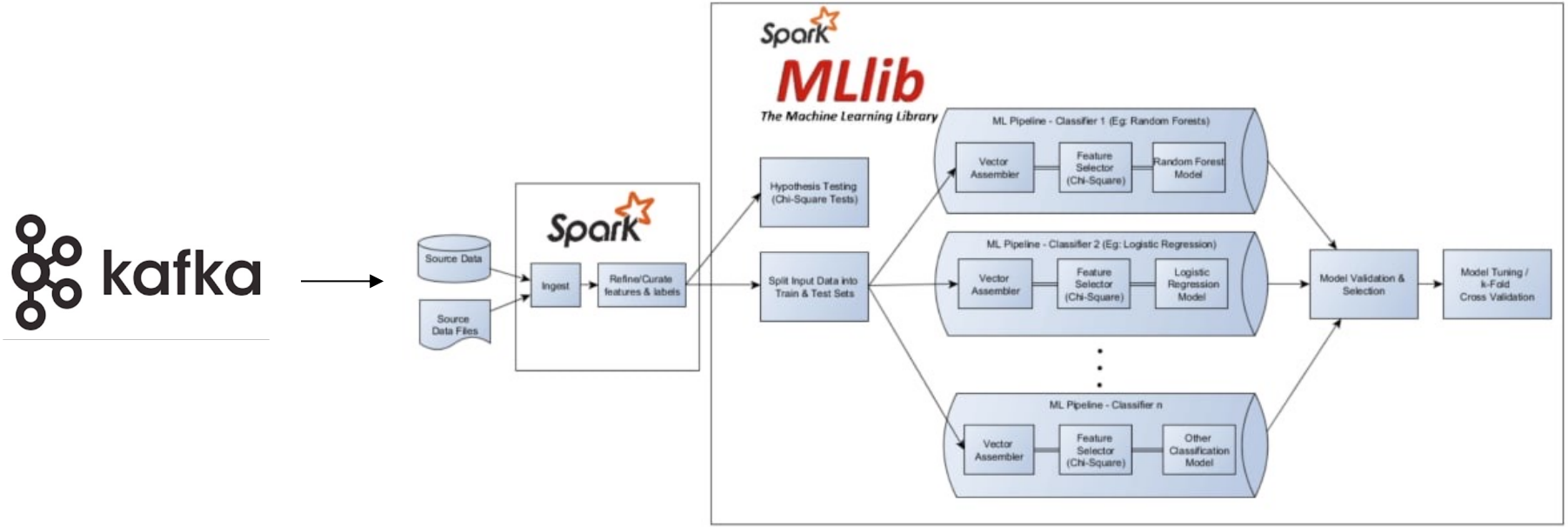


https://www.youtube.com/watch?v=vGLKi3TMJv8

Kappa Architecture for a Lakehouse with Kafka and Spark

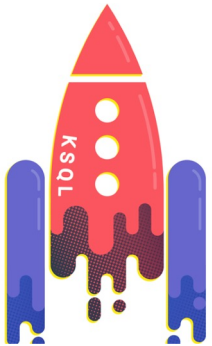


Machine Learning Model Training
with Spark MLlib



https://dev.to/siddhantpatro/spark-mllib-for-big-data-and-machine-learning-330j

Model Deployment with
Apache Kafka, ksqlDB and Spark MLlib



“CREATE STREAM AnomalyDetection AS

SELECT sensor_id, detectAnomaly(sensor_values)

FROM car_engine;“

User Defined Function (UDF)

APACHE Spark™ MLlib

Stream Processing with Kafka or Spark?

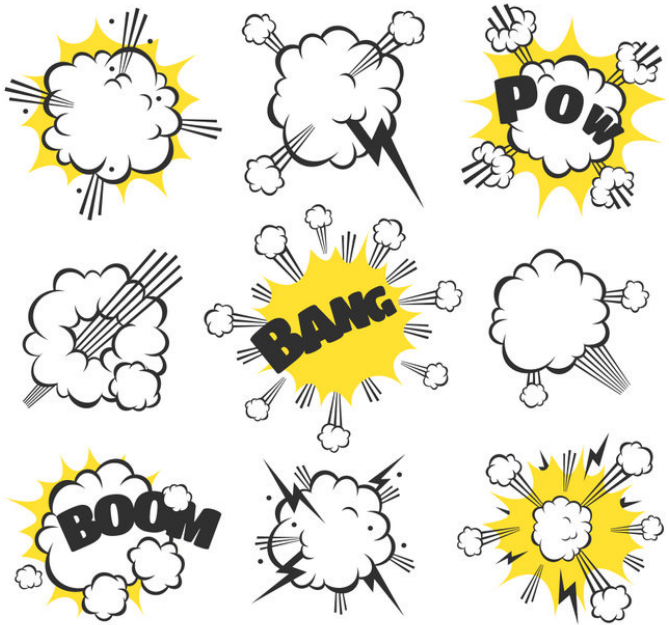
Kafka Streams / ksqlDB

Component of the data streaming infrastructure

Low latency

Focus on 24/7 operations

Lightweight, decoupled microservices



Spark Streaming

Component of the data analytics infrastructure

Strong integration with the rest of the Spark ecossytem

Stream and batch

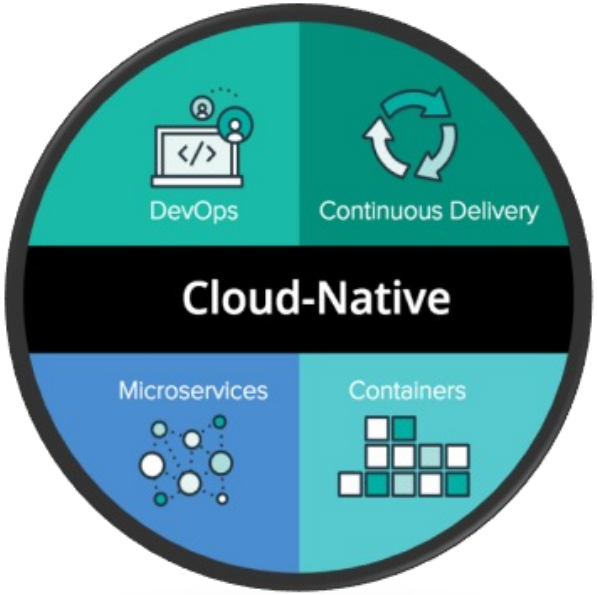
Machine Learning “embedded”

Agenda

- Data Analytics at Rest
- Data Streaming in Motion
- Lakehouse: Data Streaming + Analytics
- A Lakehouse Example: Intelligent Connected Cars
- **Cloud-Native vs. Serverless Infrastructure**
- Central vs. Hybrid and Global Data Mesh

Cloud-Native Deployment

→ Elastic Infrastructure and Faster Time-to-Market



kubernetes

docker

RED HAT OPENSHIFT

CLOUDFOUNDRY

RANCHER

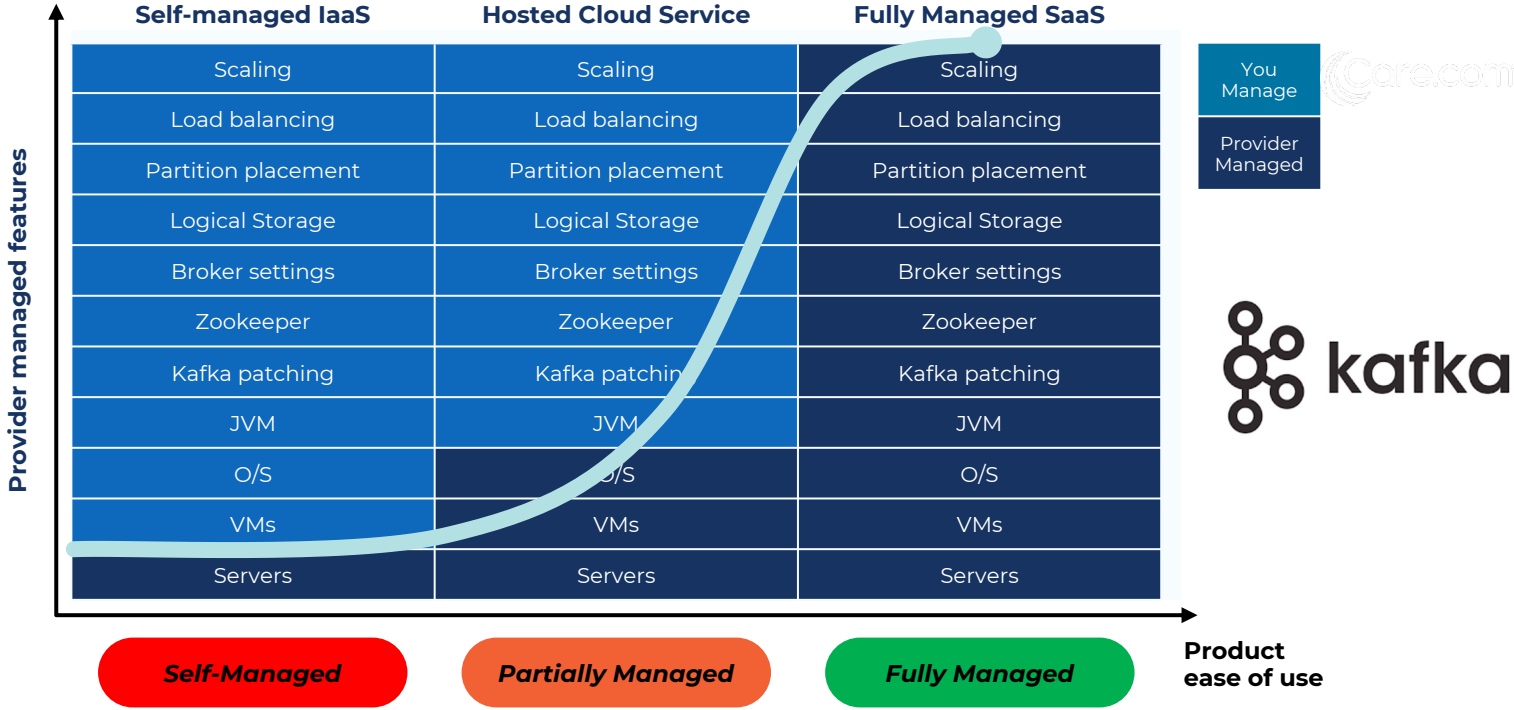
D2IQ

Amazon EKS

Google Kubernetes Engine

Azure Kubernetes Service (AKS)

What is a (truly) fully-managed SaaS?



Agenda

- Data Analytics at Rest
- Data Streaming in Motion
- Lakehouse: Data Streaming + Analytics
- A Lakehouse Example: Intelligent Connected Cars
- Cloud-Native vs. Serverless Infrastructure
- **Central vs. Hybrid and Global Data Mesh**

AWS Cloud Outage hit Disney World Visitors...

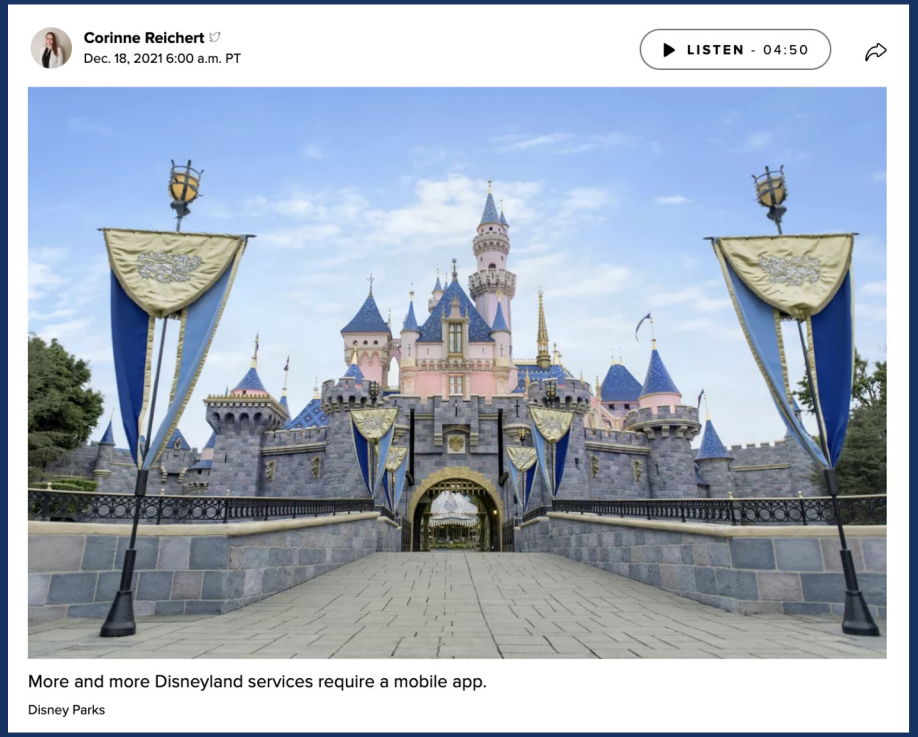
Disney parks were already facing heat from fans. Then an AWS outage came along

As Disney increasingly leans on apps for almost every facet of guest experience, tech problems have a wide-reaching impact on expensive days in the theme parks.

Not even Disney's vaunted magic could save its Disneyland park app from a widespread Amazon Web Services outage temporarily wrecking the day for its guests this week. But for fans of "the happiest place on Earth," this was just the latest in a string of problems.

Disney has been increasingly pushing its theme park guests to use their mobile devices to do everything from ordering food to accessing tickets and park reservations. It has also put a new paid version of its FastPass system, now re-branded Genie Plus, into the app. That means outages, including one that hit Walt Disney World last week, can bring enjoyment in the parks to a screeching halt.

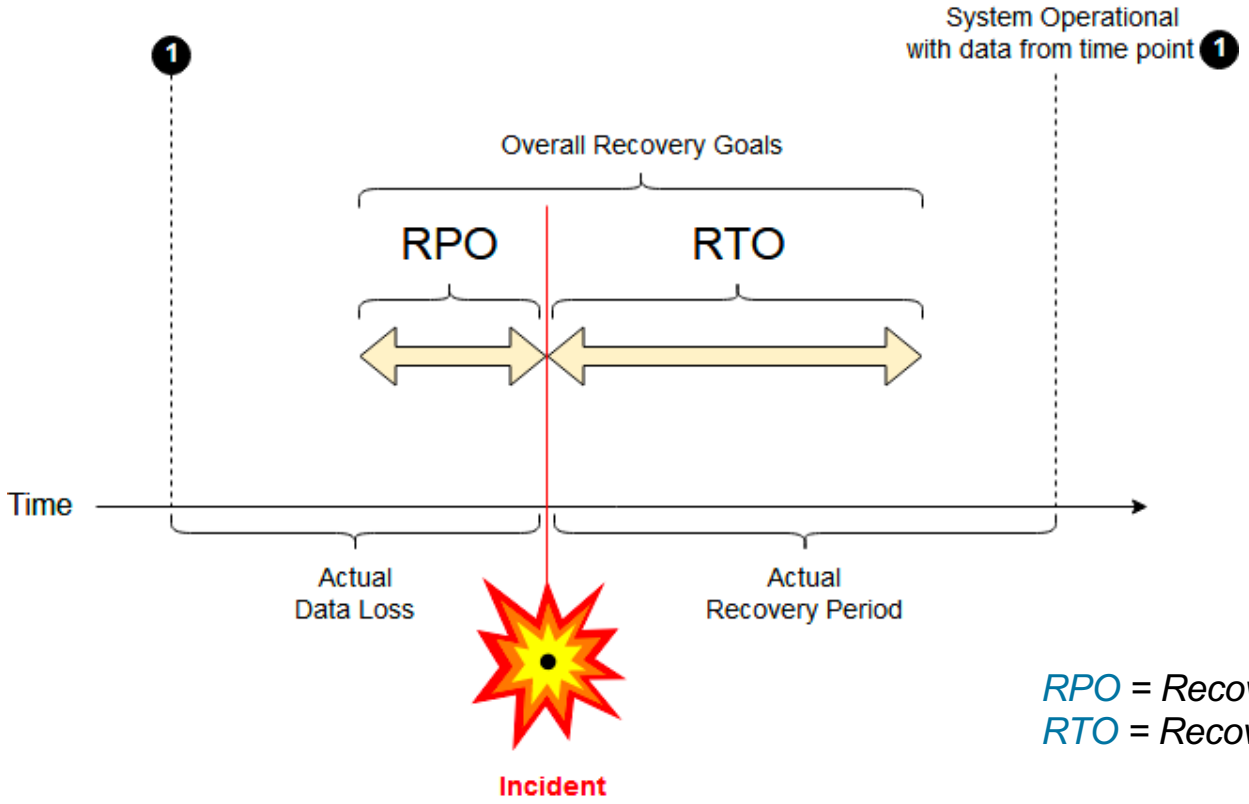
https://www.cnet.com/tech/services-and-software/disney-parks-were-already-facing-heat-from-fans-then-an-aws-outage-came-along/



More and more Disneyland services require a mobile app.

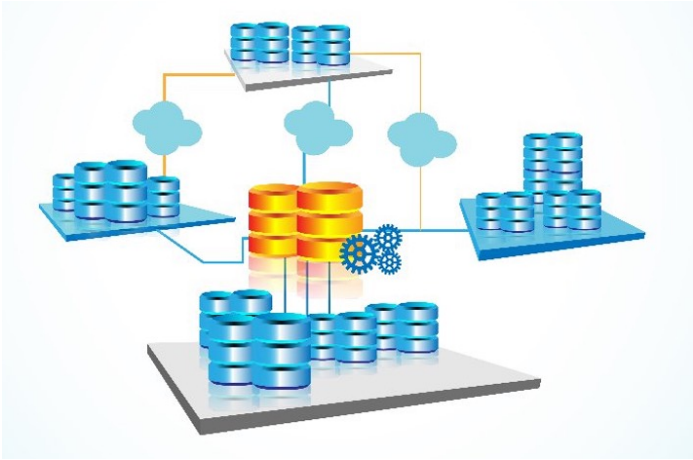
Disney Parks

Disaster Recovery – RPO and RTO

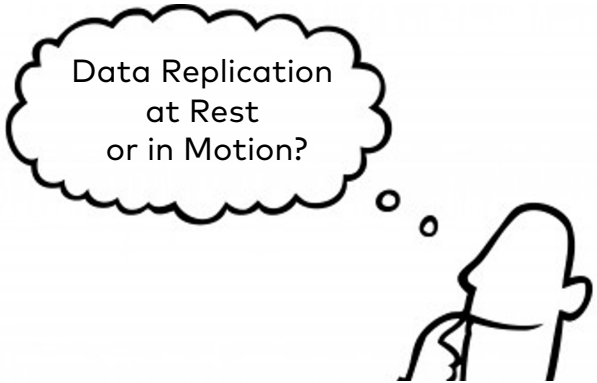


RPO = *Recovery Point Objective*
RTO = *Recovery Time Objective*

Use Cases for Hybrid and Multi-Cloud Data Lakehouses



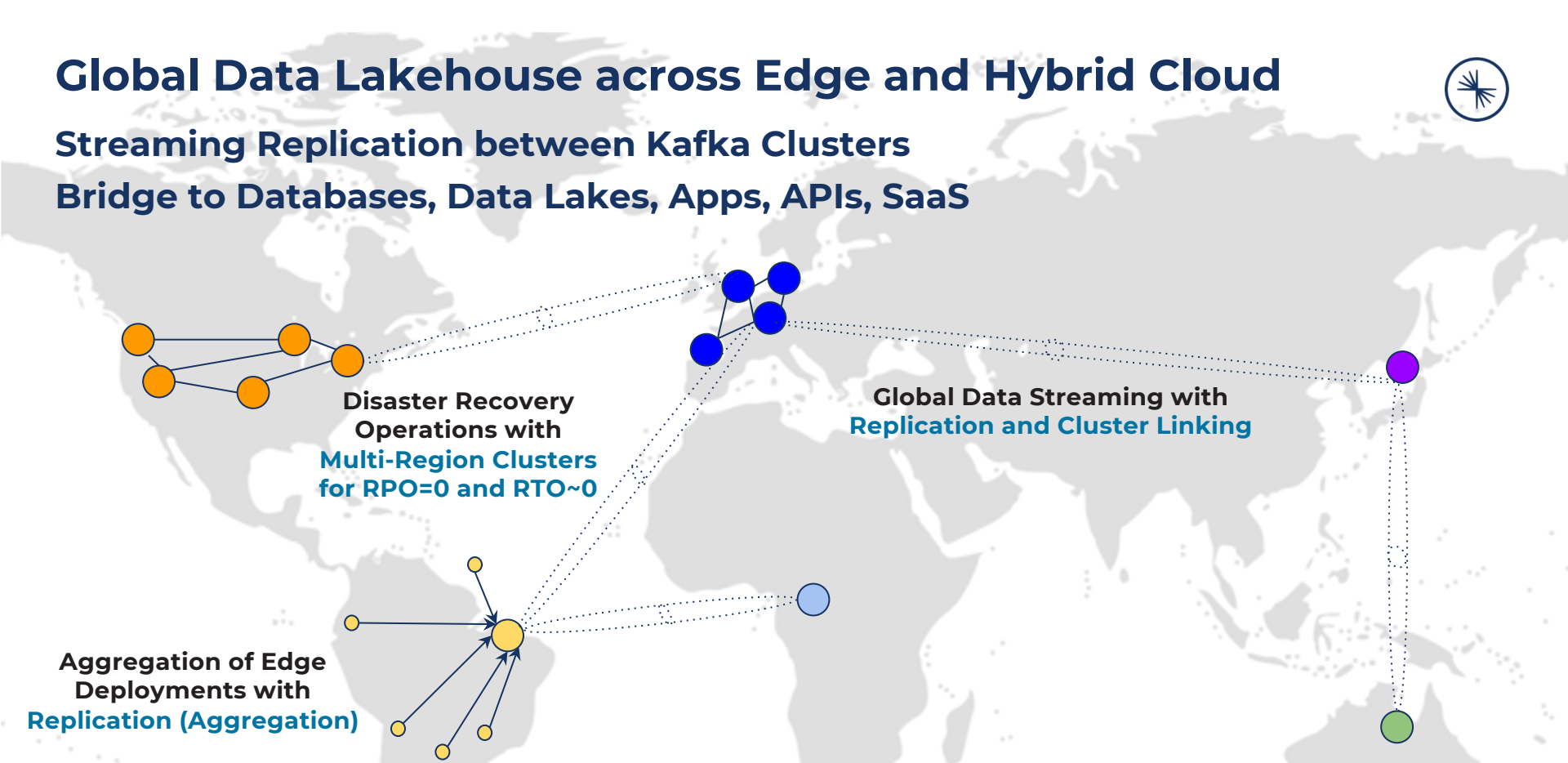
- **Disaster Recovery and High Availability**: Create a disaster recovery cluster, and fail over to it during an outage.
- **Global and Multi-Cloud Replication**: Move and aggregate data across regions and clouds.
- **Data Sharing**: Share data with other teams, lines-of-business, or organizations.
- **Data Migration**: Migrate data and workloads from one cluster to another (like from legacy on-premise data warehouse to cloud-native data lakehouse).



Global Data Lakehouse across Edge and Hybrid Cloud

Streaming Replication between Kafka Clusters

Bridge to Databases, Data Lakes, Apps, APIs, SaaS

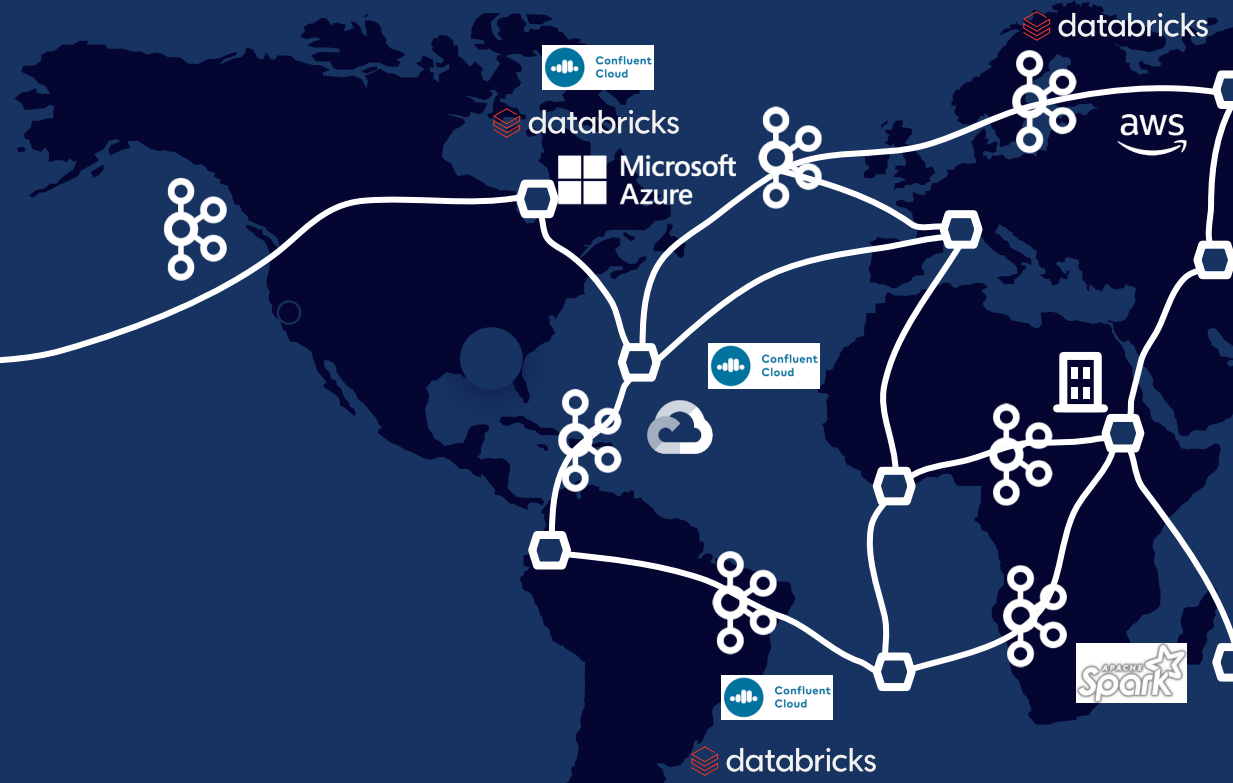


Copyright 2021, Confluent, Inc. All rights reserved. This document may not be reproduced in any manner without the express written permission of Confluent, Inc.

A data mesh for decentralized data products

Independent Data Products for Reporting, Analytics, Data Streaming

For instance: A KSQL microservice



CONFLUENT

Questions? Feedback? Let's connect!

Kai Waehner

Field CTO

kai.waehner@confluent.io
@KaiWaehner
confluent.io
kai-waehner.de
linkedin.com/in/kaiwaehner

KAI WAEHNER