

# Executive Briefing: What You Need to Know about Fast Data

Dean Wampler, Ph.D.  
[dean@lightbend.com](mailto:dean@lightbend.com)  
[@deanwampler](https://twitter.com/deanwampler)  
[polyglotprogramming.com/talks](https://polyglotprogramming.com/talks)



Based on  
this report

[lightbend.com/fast-data-platform](http://lightbend.com/fast-data-platform)  
2nd edition coming in October!

# Fast Data Architectures for Streaming Applications

Getting Answers Now from  
Data Sets that Never End

A black and white photograph showing a rocky riverbed. Water is flowing rapidly over and around the stones, creating white foam and ripples. The scene conveys a sense of constant motion and flow.

Dean Wampler



# What We'll Discuss

- 
- Why streaming? Why now?
  - How to choose technologies
  - The impact streaming will have on your organization

What We'll Discuss



# Why Streaming?

- 
- New opportunities that require streaming
  - Upgrading batch applications for competitive advantage

# Why Streaming?



## Similar IoT Architectures

# Fast Data Use Cases

### Predictive Analytics

Apply ML models to large volumes of device data to pre-empt failures / outages



**Hewlett Packard  
Enterprise**

### IoT

Real-time consumer and industrial Device and Supply Chain management at scale



### Real-time Personalization

Real-time marketing based on behavior, location, inventory levels, product promotions, etc.

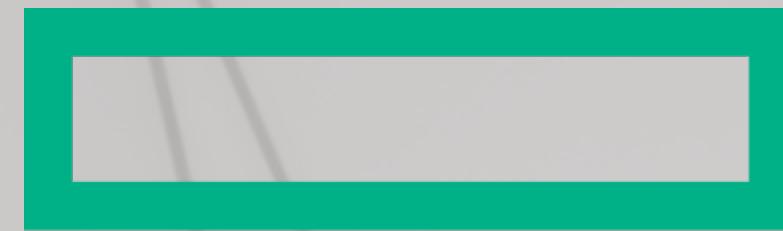


**RoyalCaribbean  
INTERNATIONAL**®

### Real-time Financial Processes

Drive better business outcomes through real-time risk, fraud detection, compliance, audit, governance, etc.





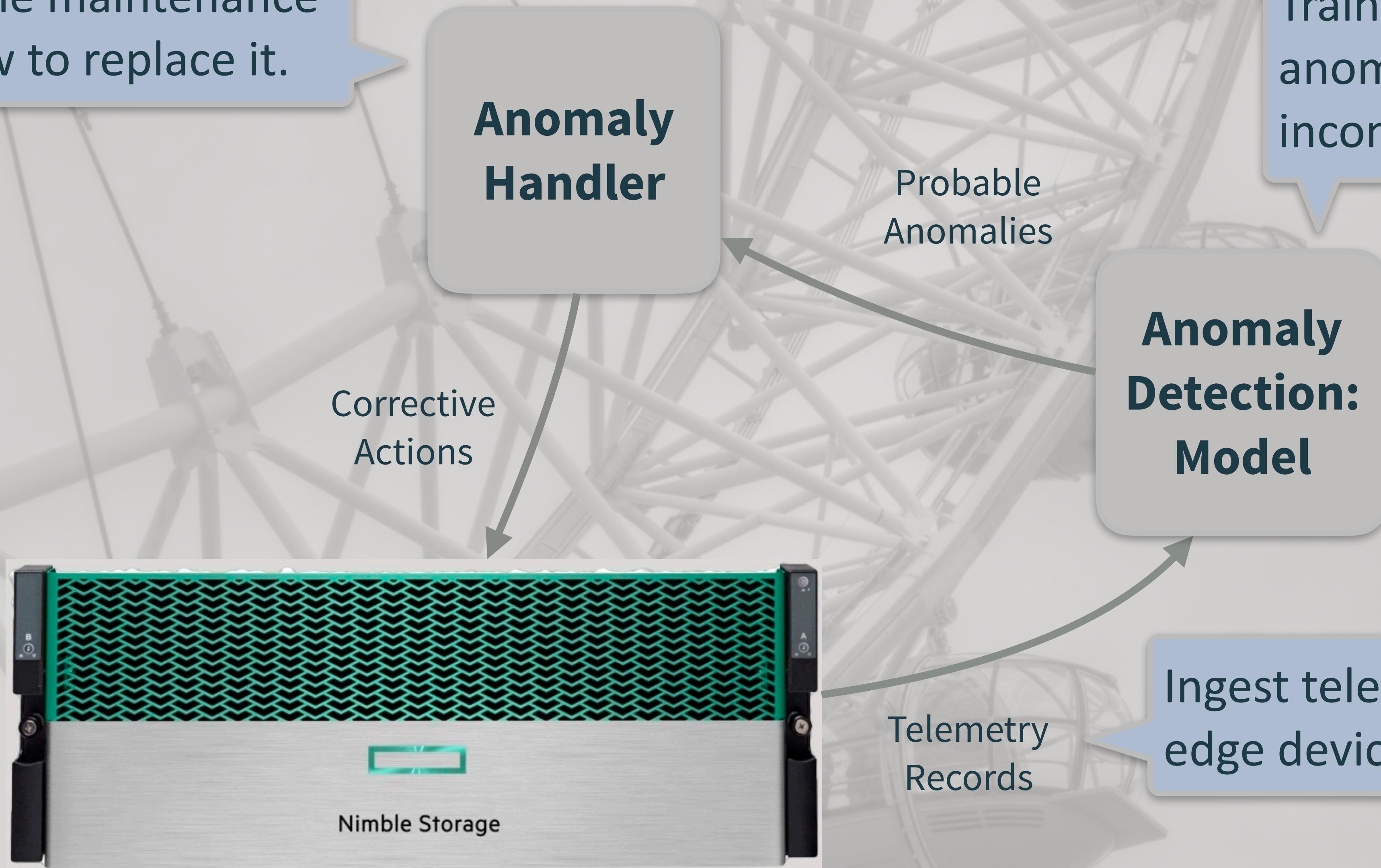
# Predictive Analytics

## Hewlett Packard Enterprise

- ML models applied to device telemetry to detect anomalies
- Preemptive maintenance prevents potential failures that would impact users

# Predictive Analytics - Core Idea

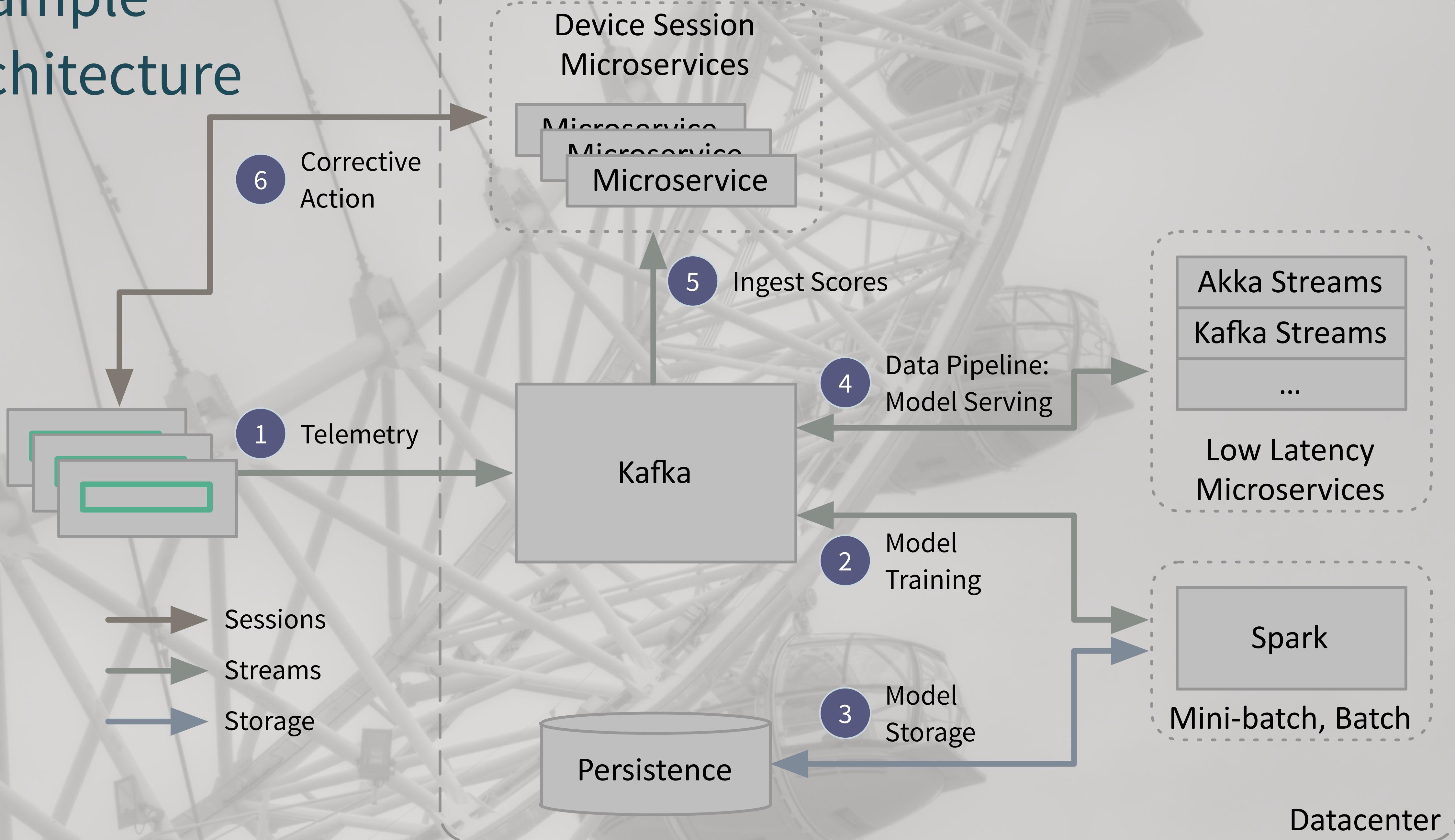
Handle anomaly: move activity off component, schedule maintenance window to replace it.



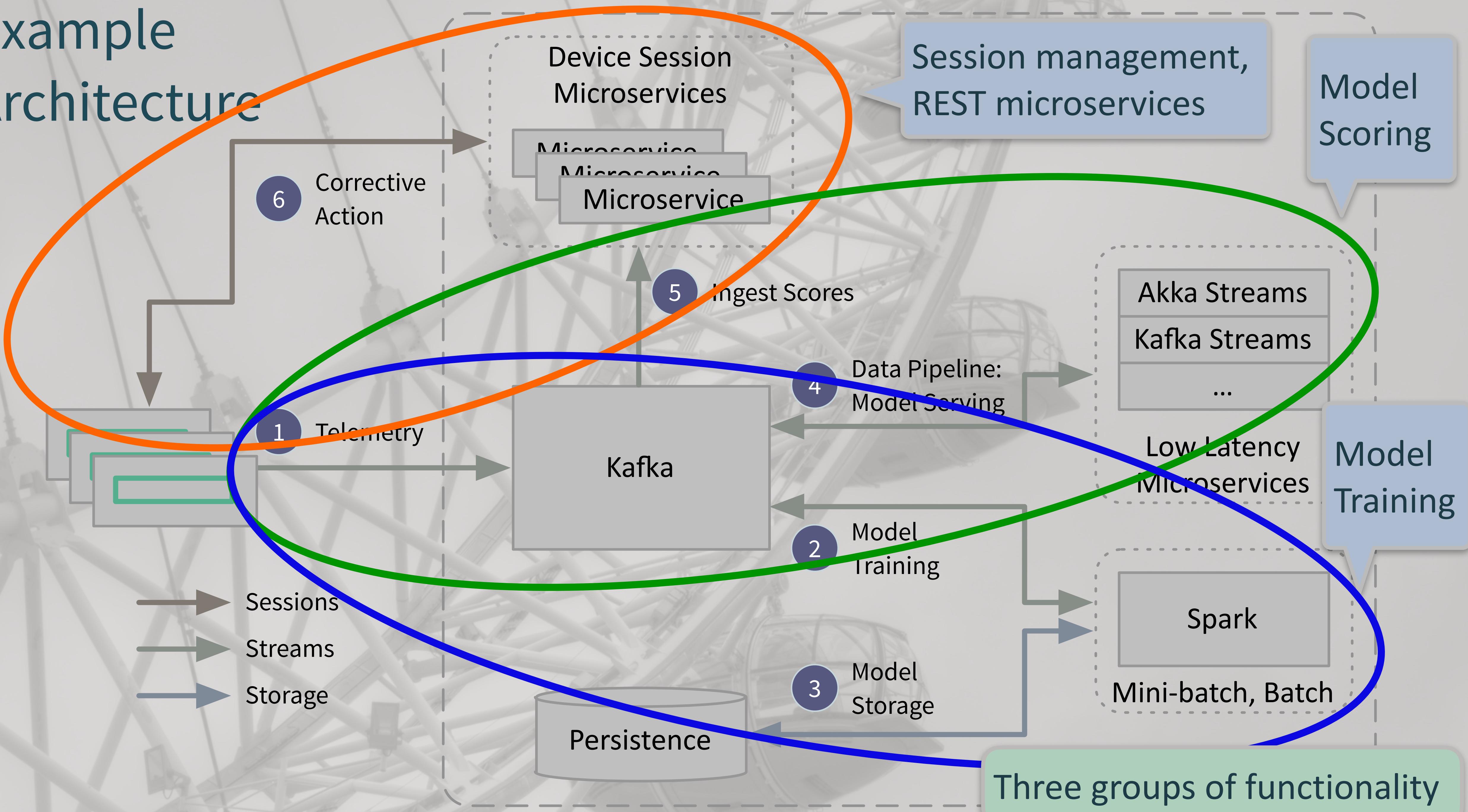
Train models to look for anomalies... and score incoming telemetry.

Ingest telemetry from edge devices.

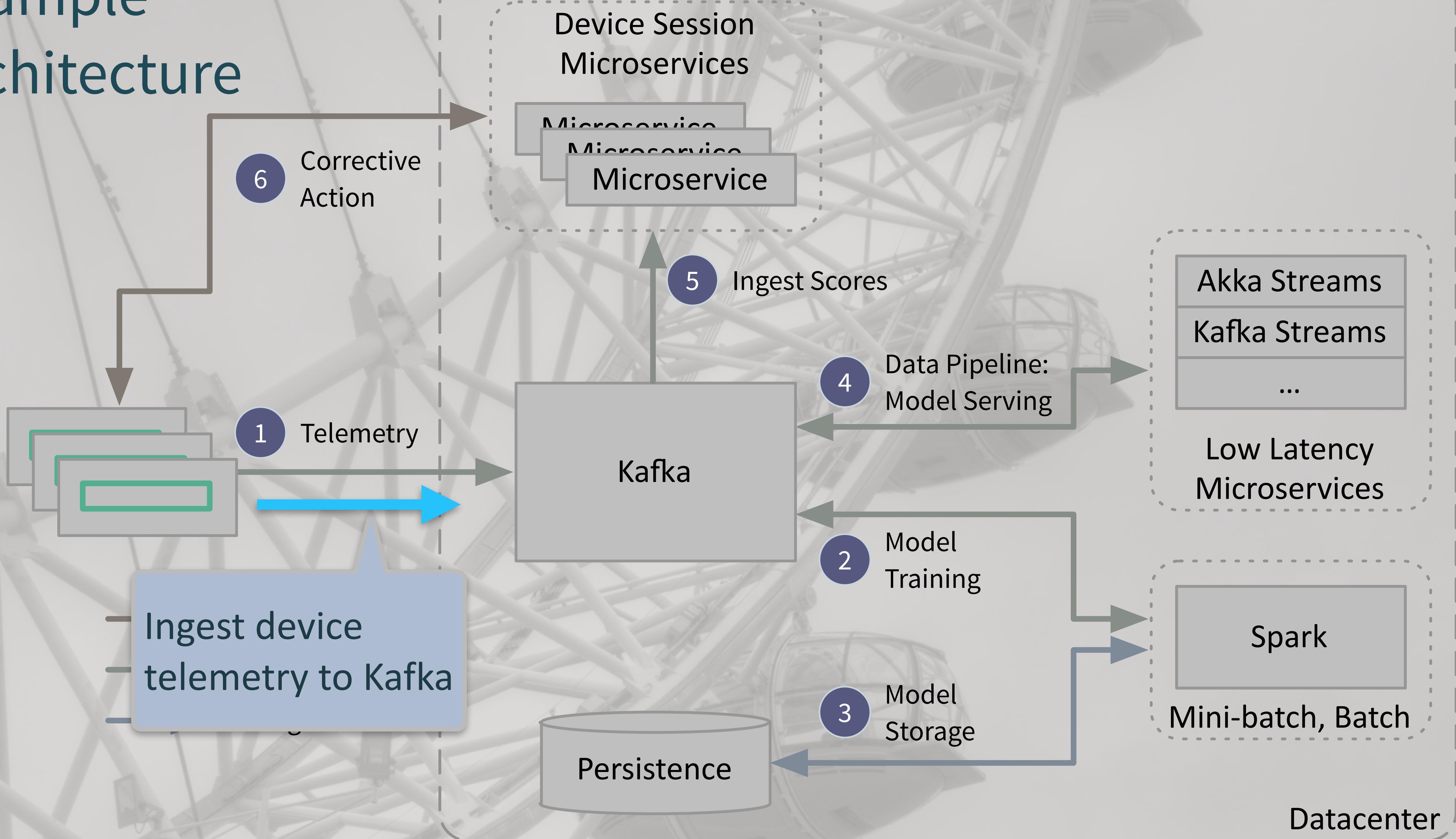
# Example Architecture



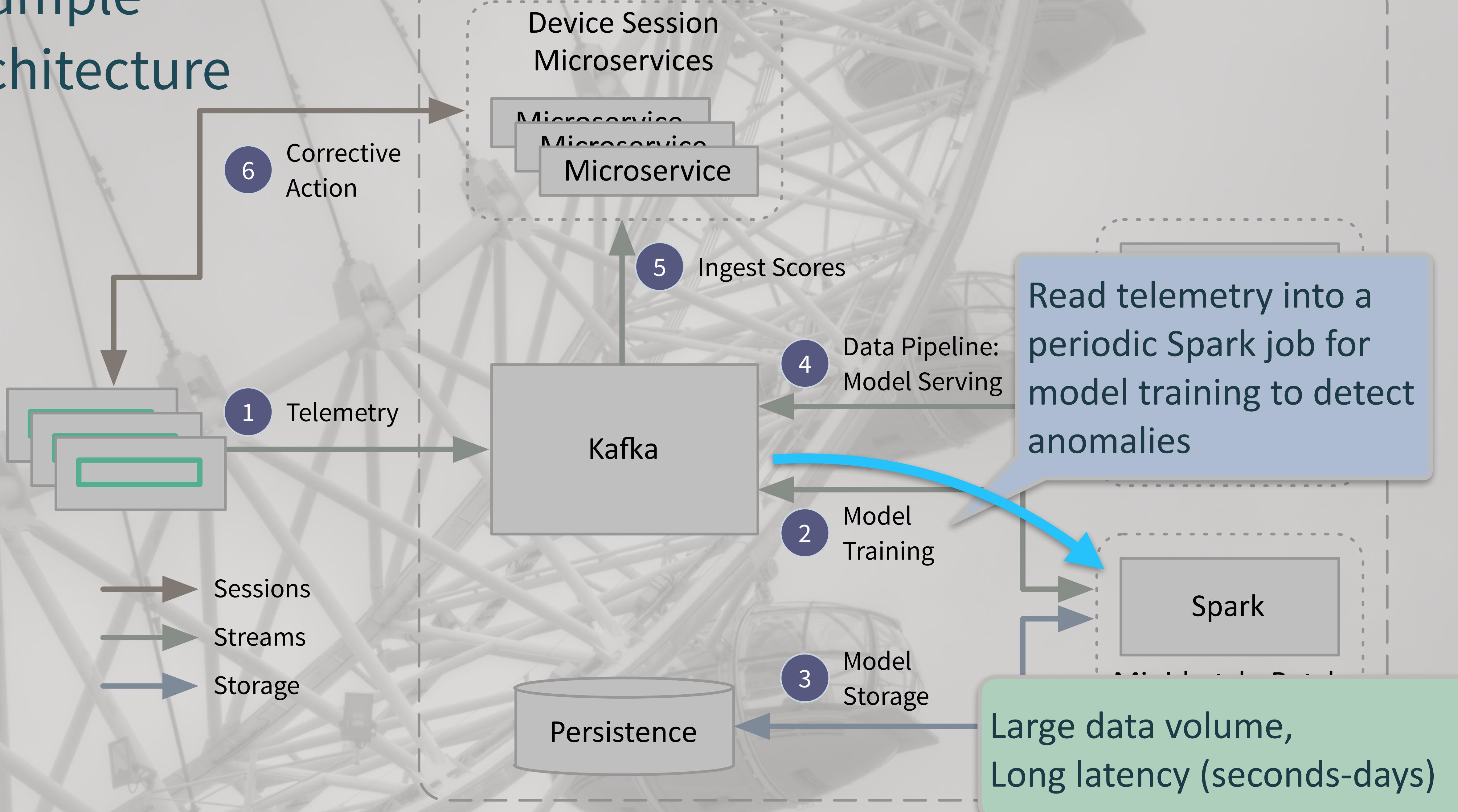
# Example Architecture



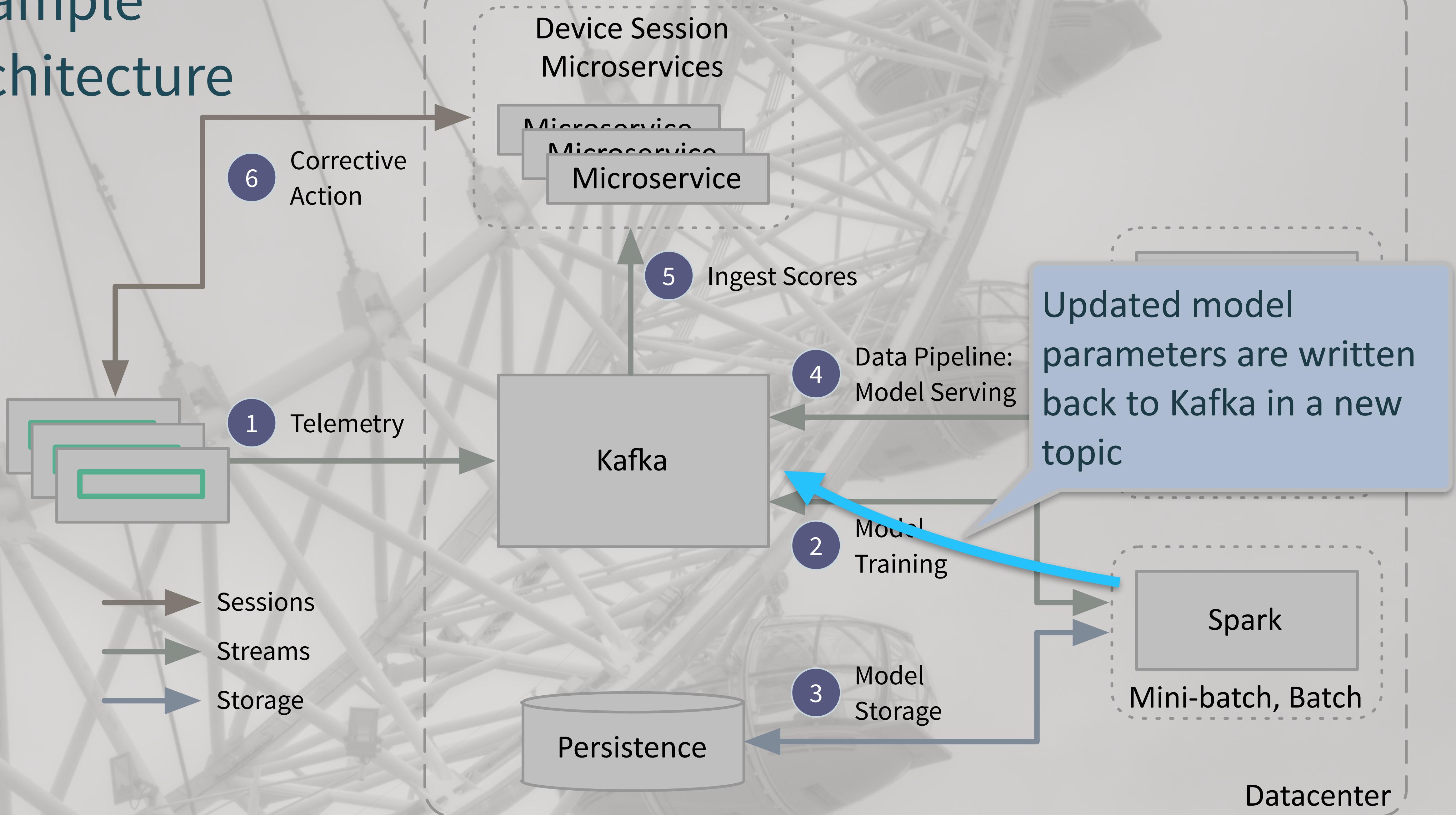
# Example Architecture



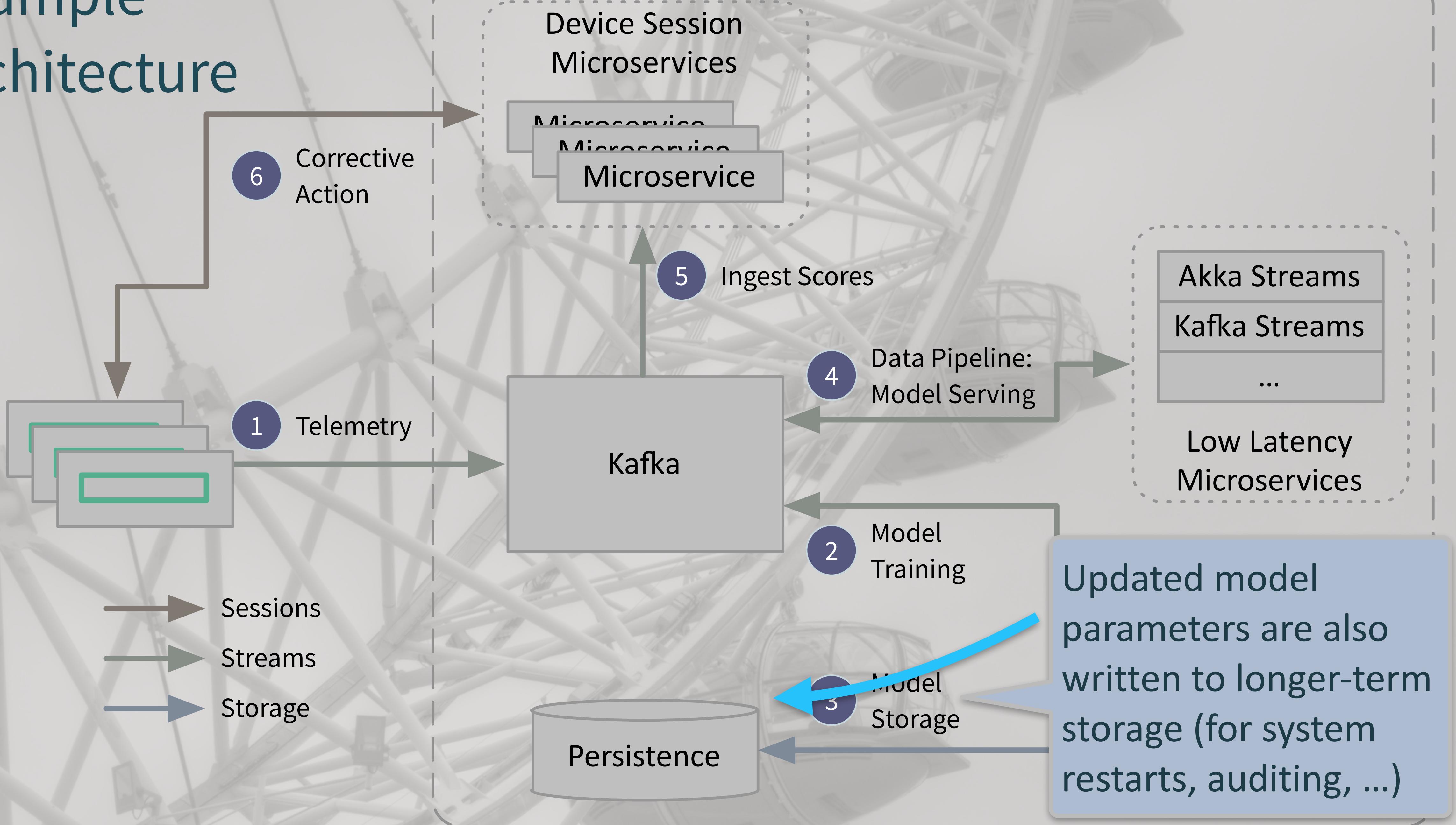
# Example Architecture



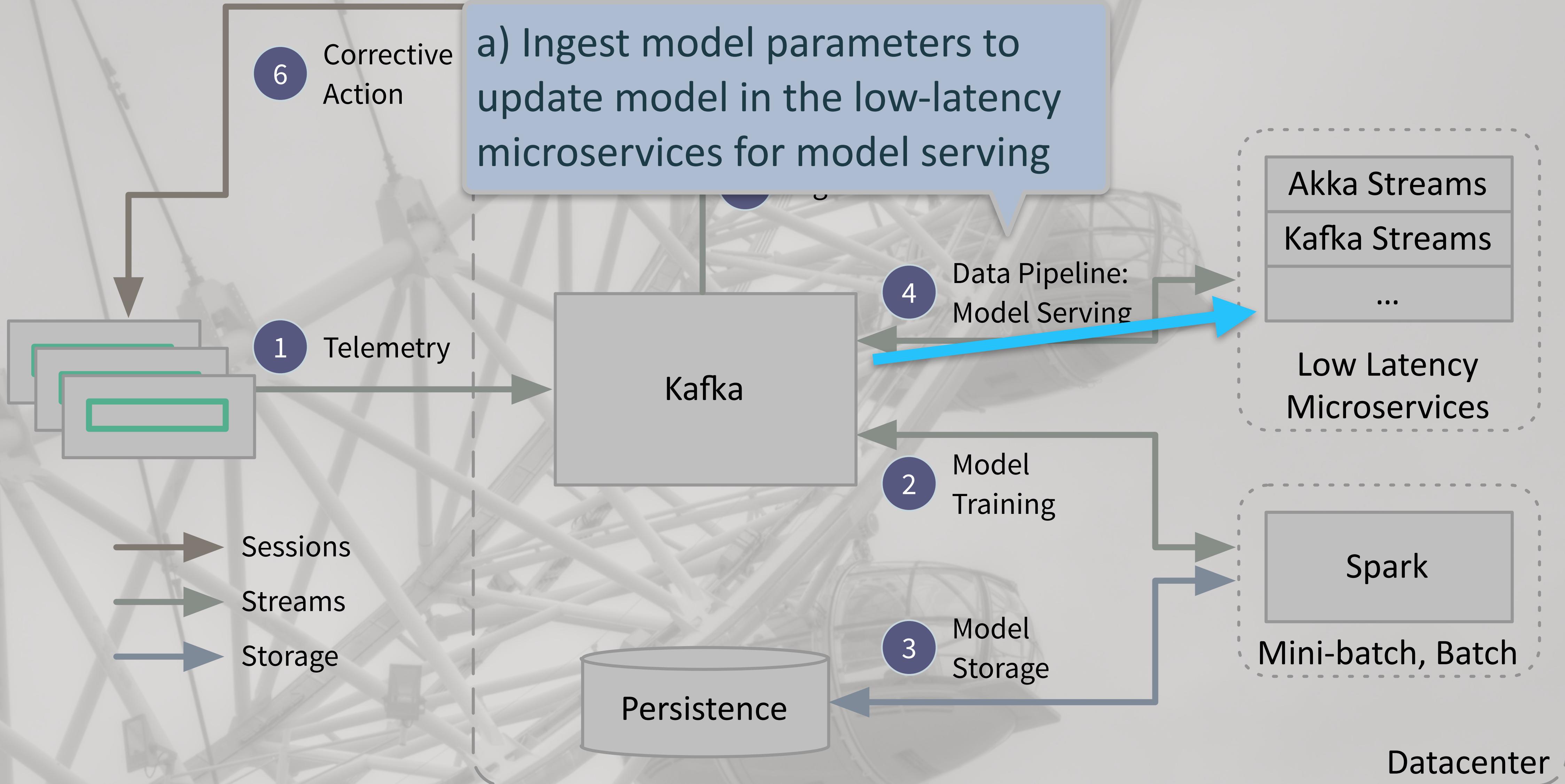
# Example Architecture



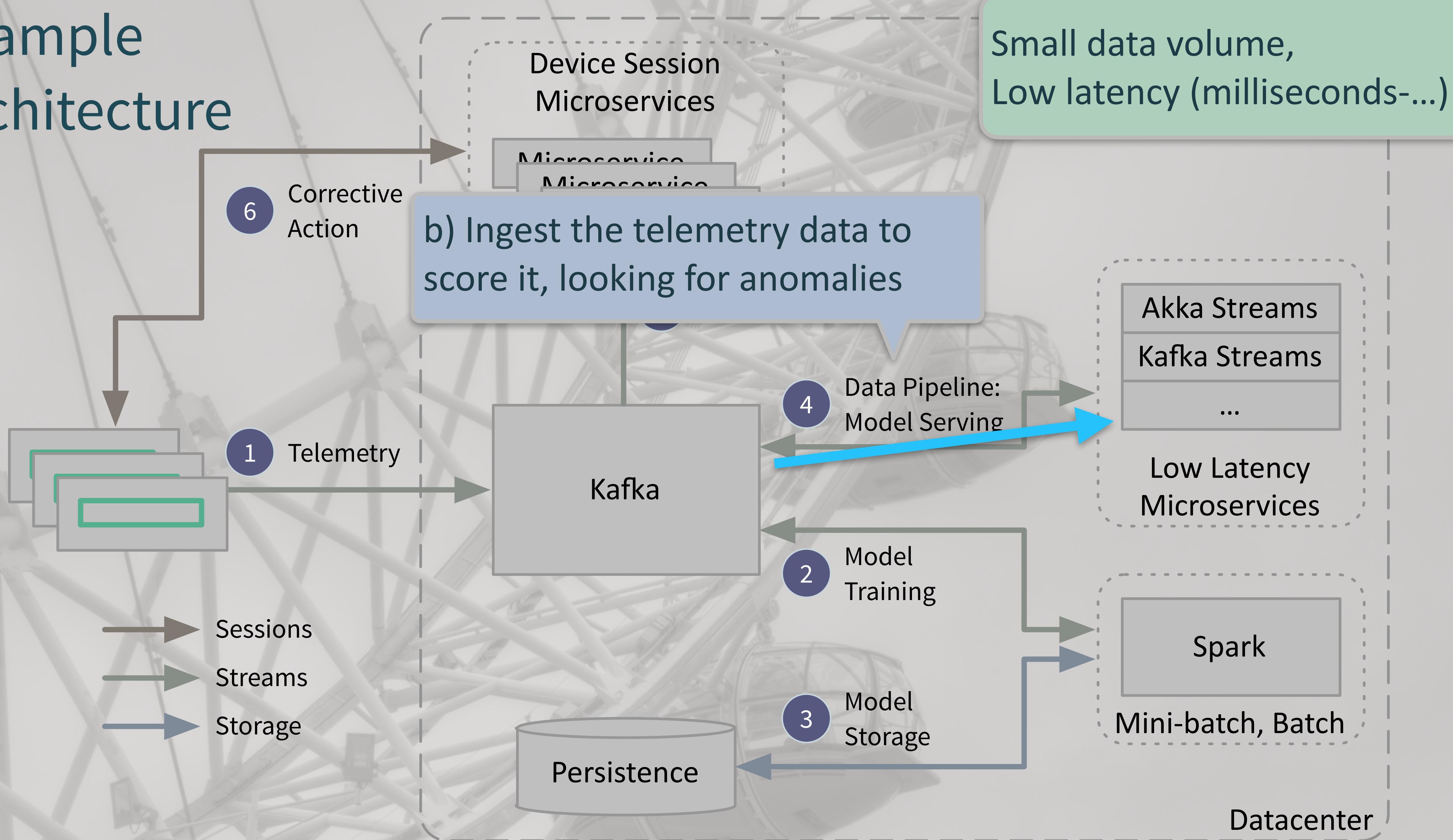
# Example Architecture



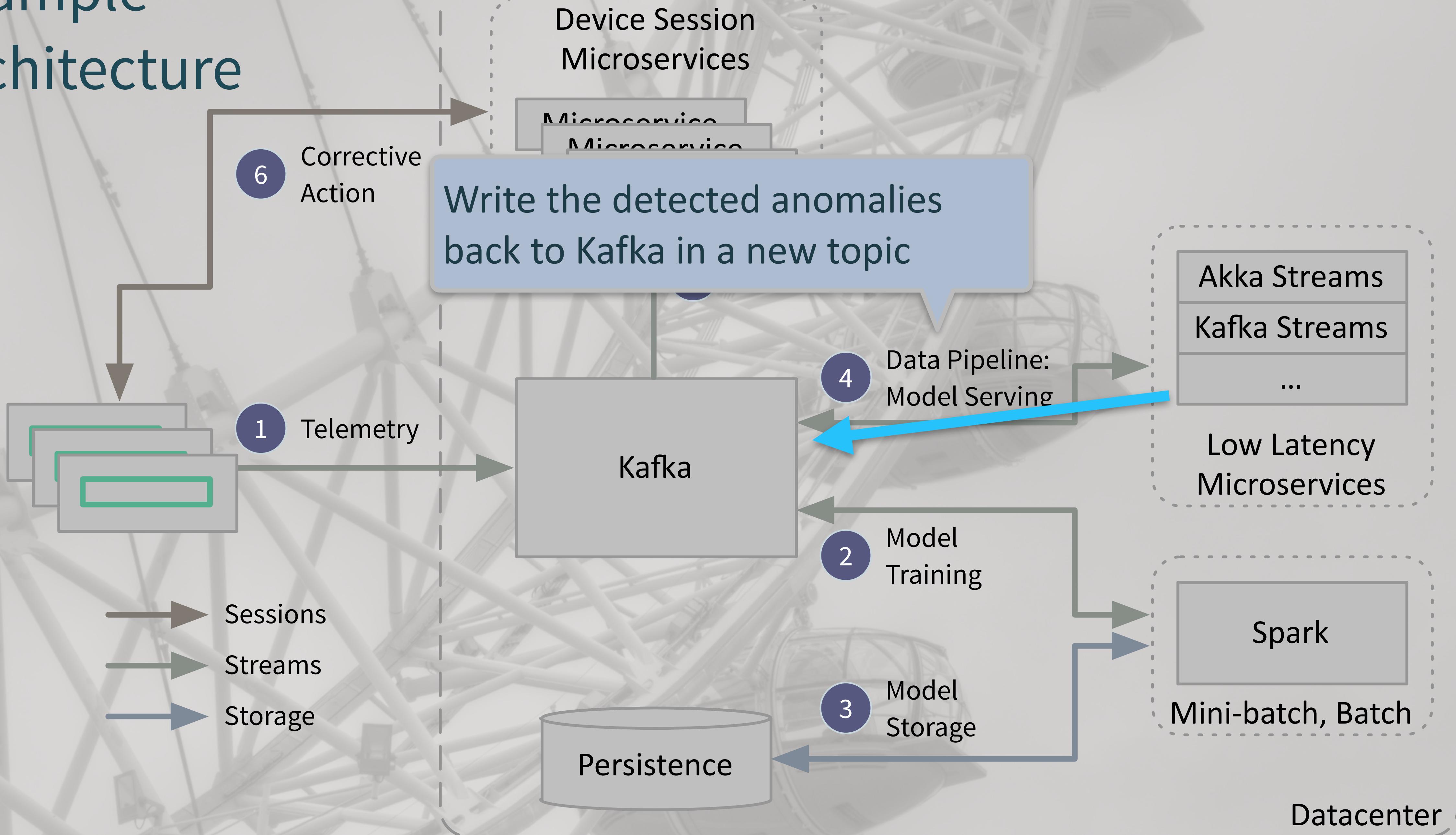
# Example Architecture



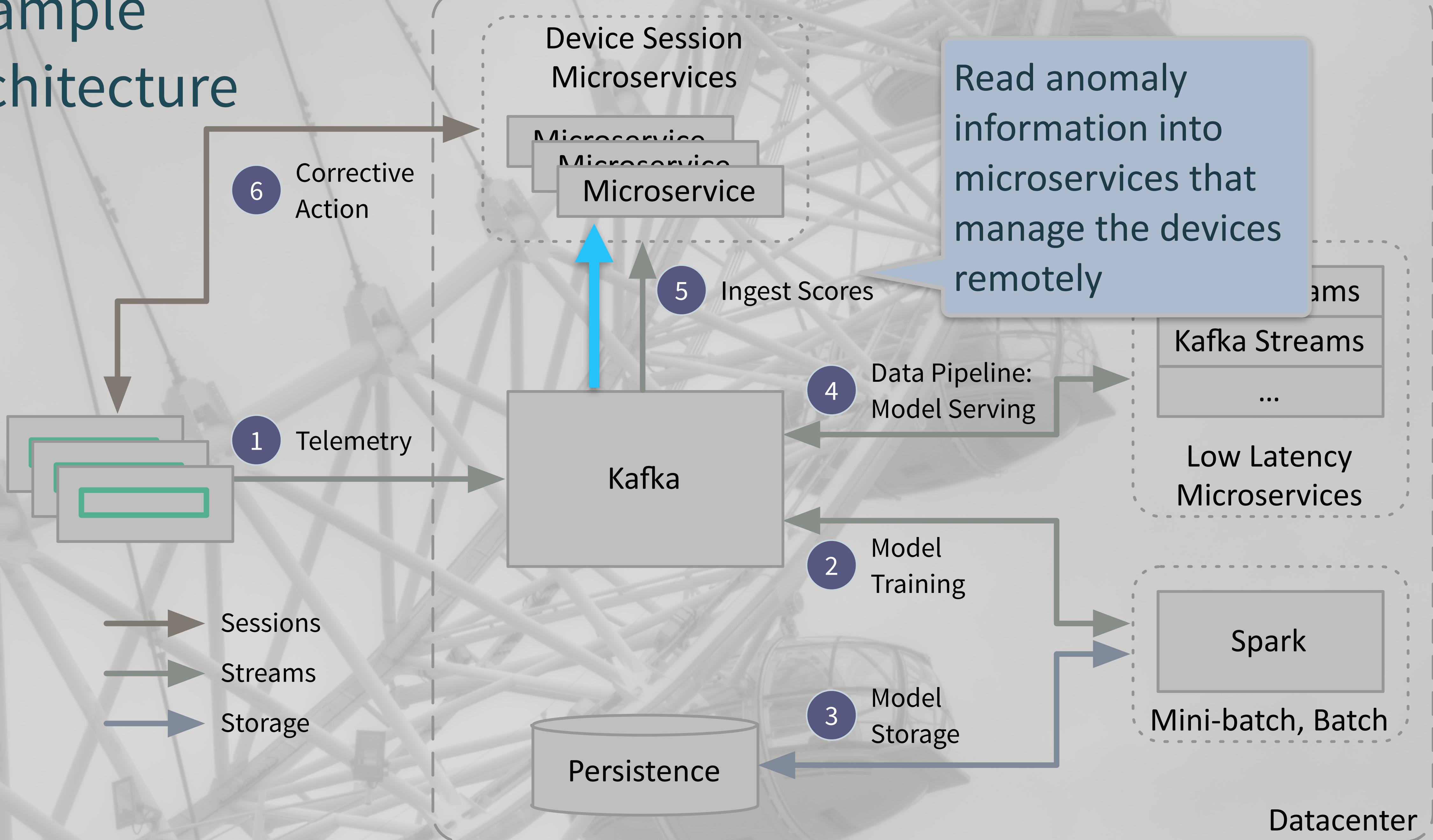
# Example Architecture



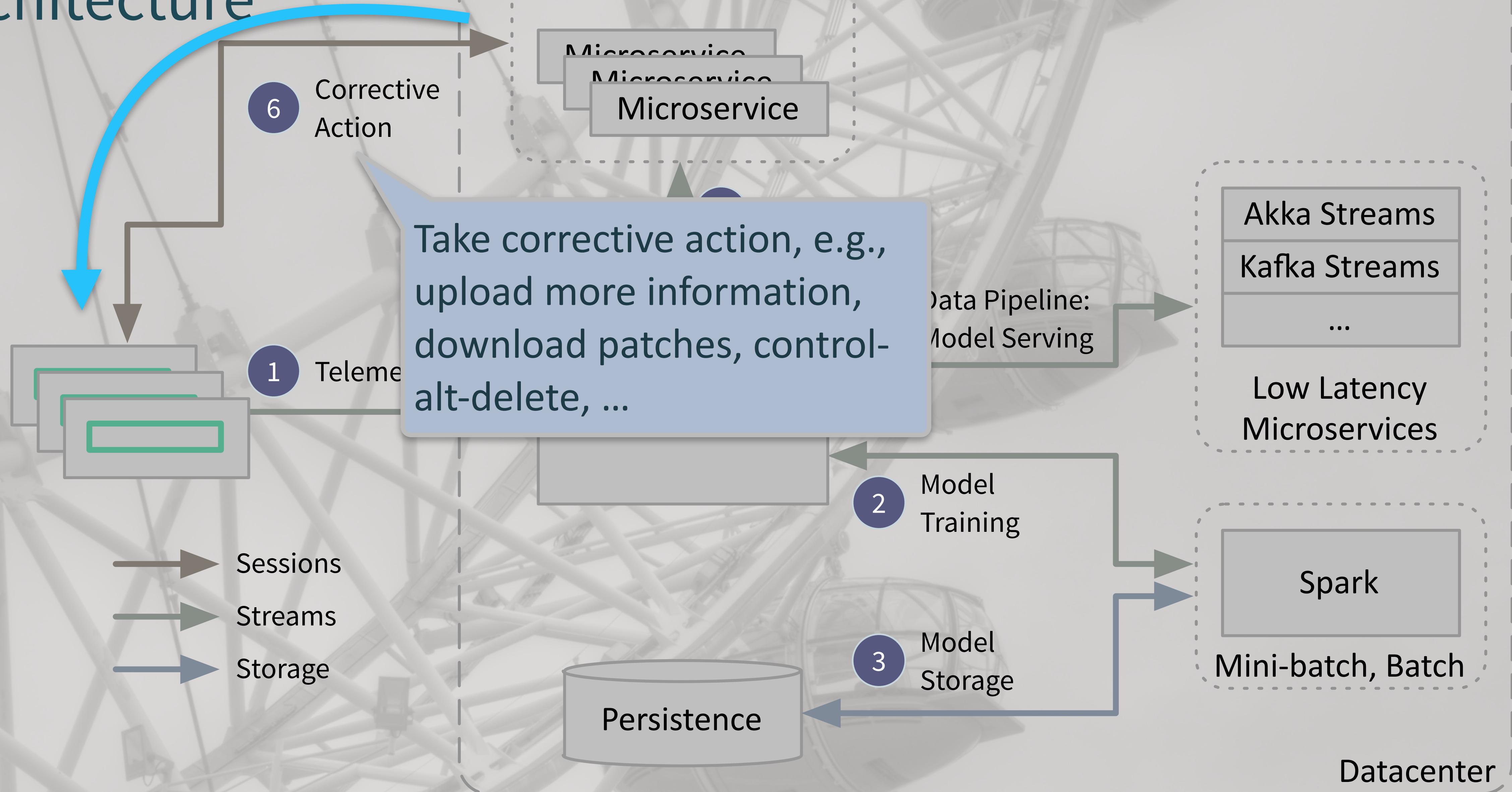
# Example Architecture



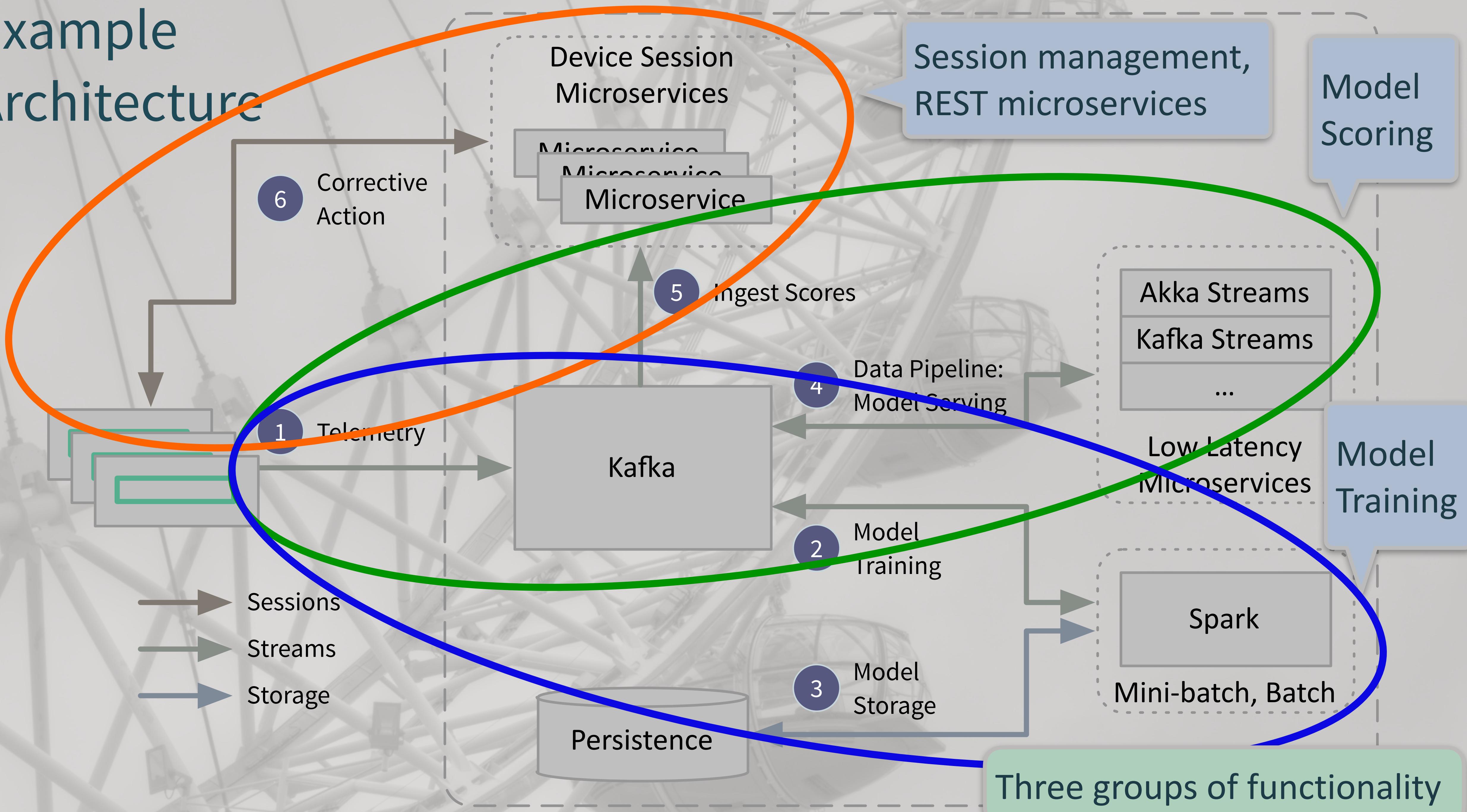
# Example Architecture



# Example Architecture

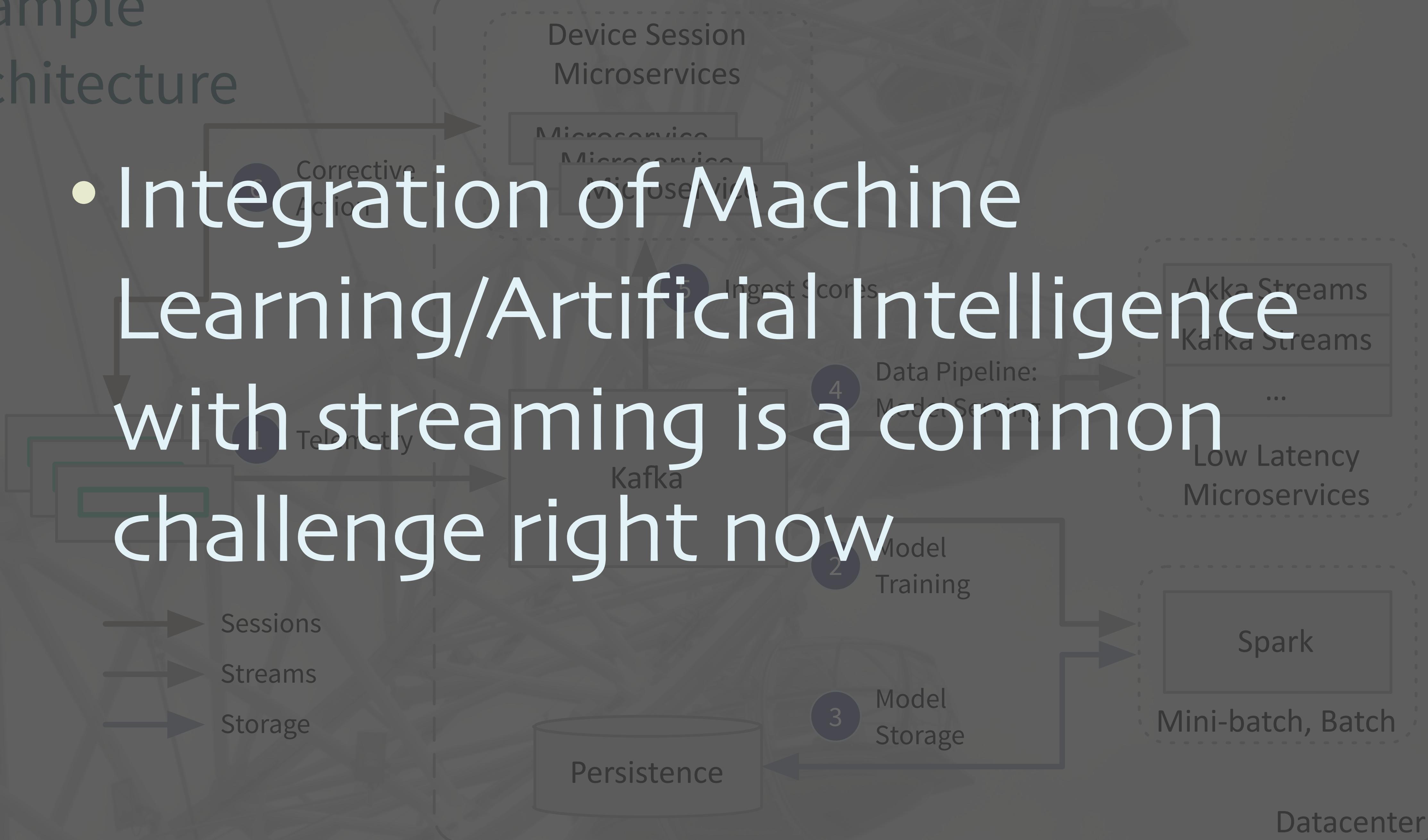


# Example Architecture



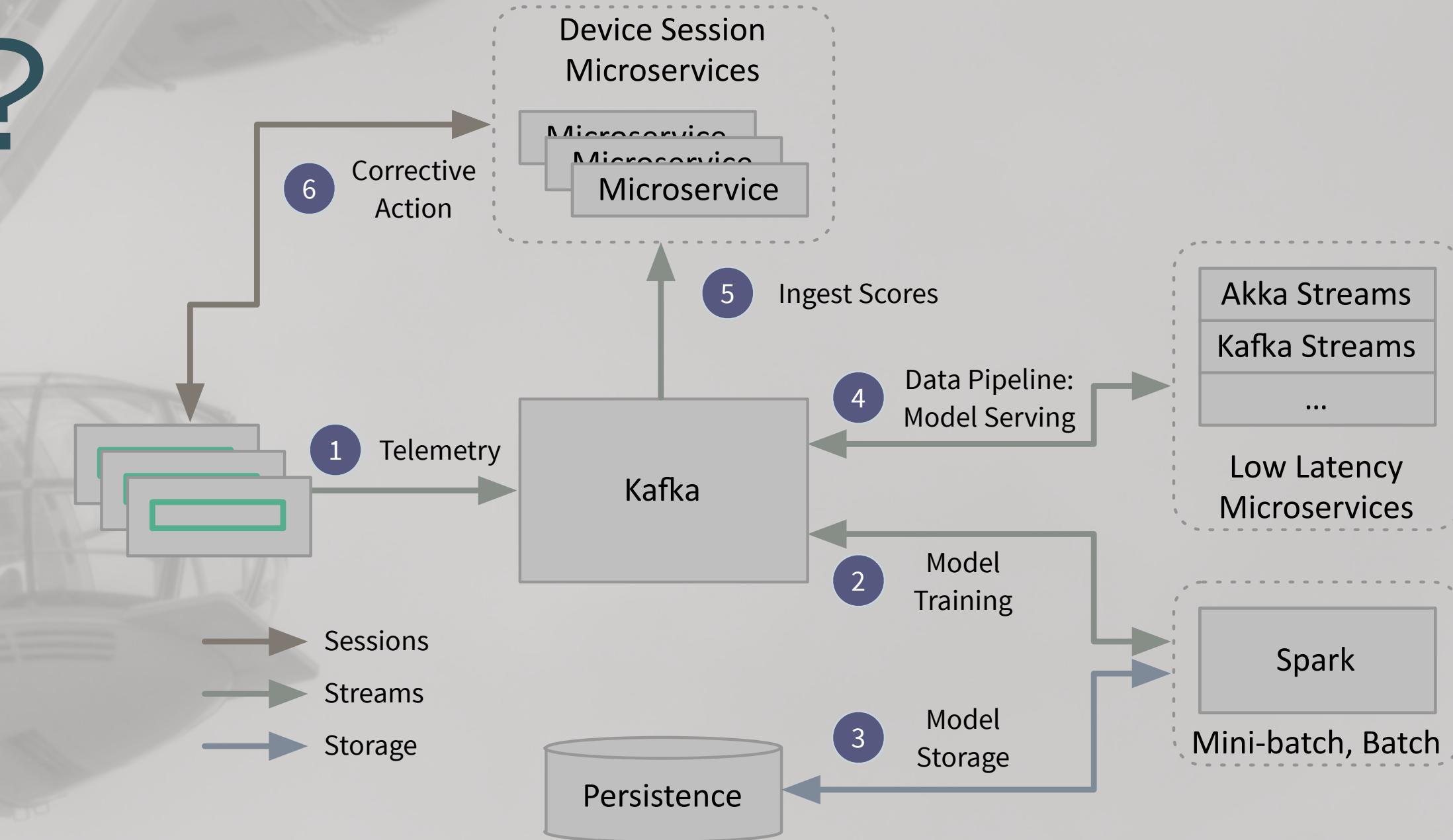
# Example Architecture

- Integration of Machine Learning/Artificial Intelligence with streaming is a common challenge right now



# Challenges

- Network overhead for telemetry ingestion too high?
- Model serving latency too long?
- Datacenter unavailable?
- Idea: Serve models on the device!

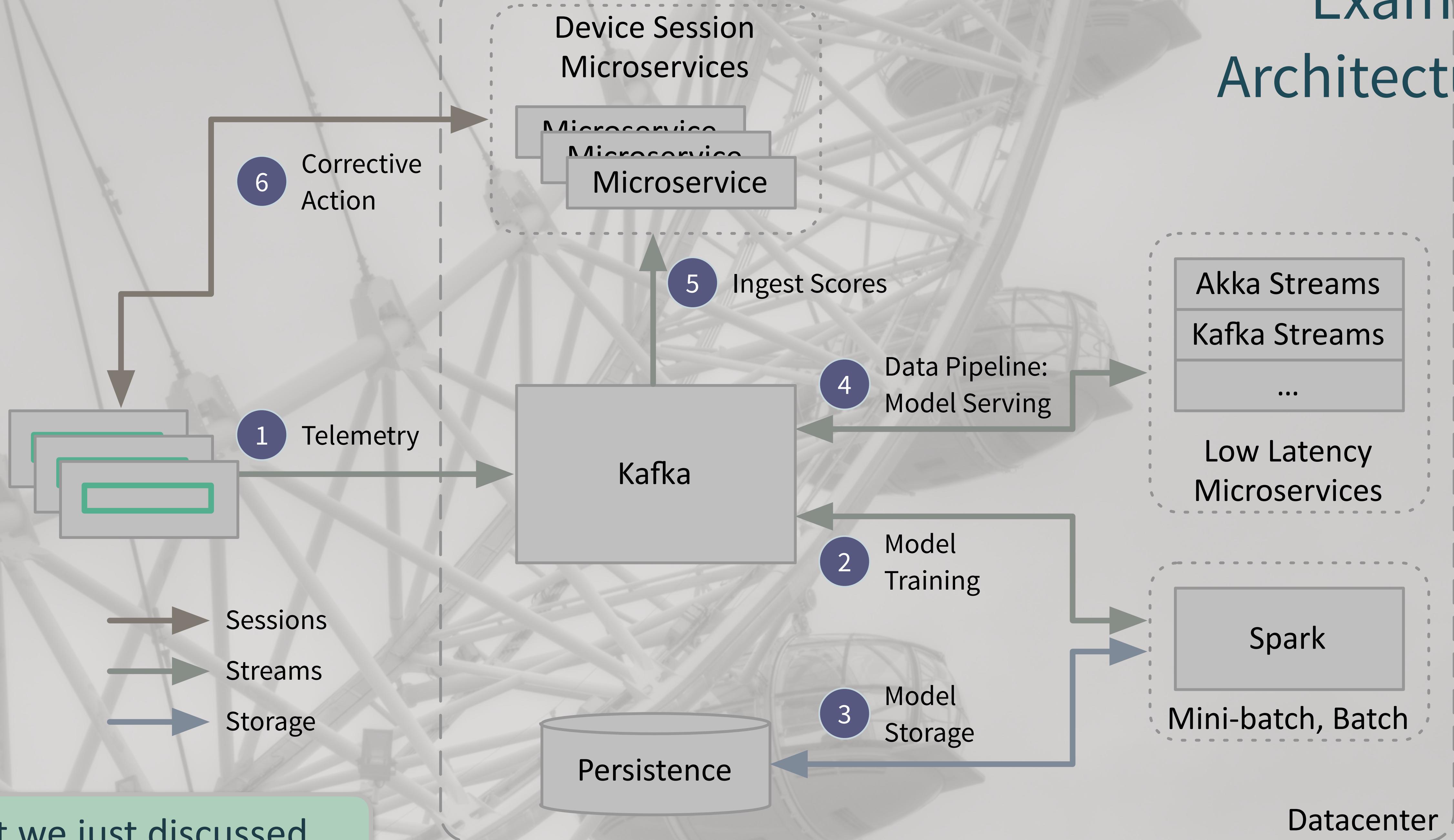




# Internet of Things

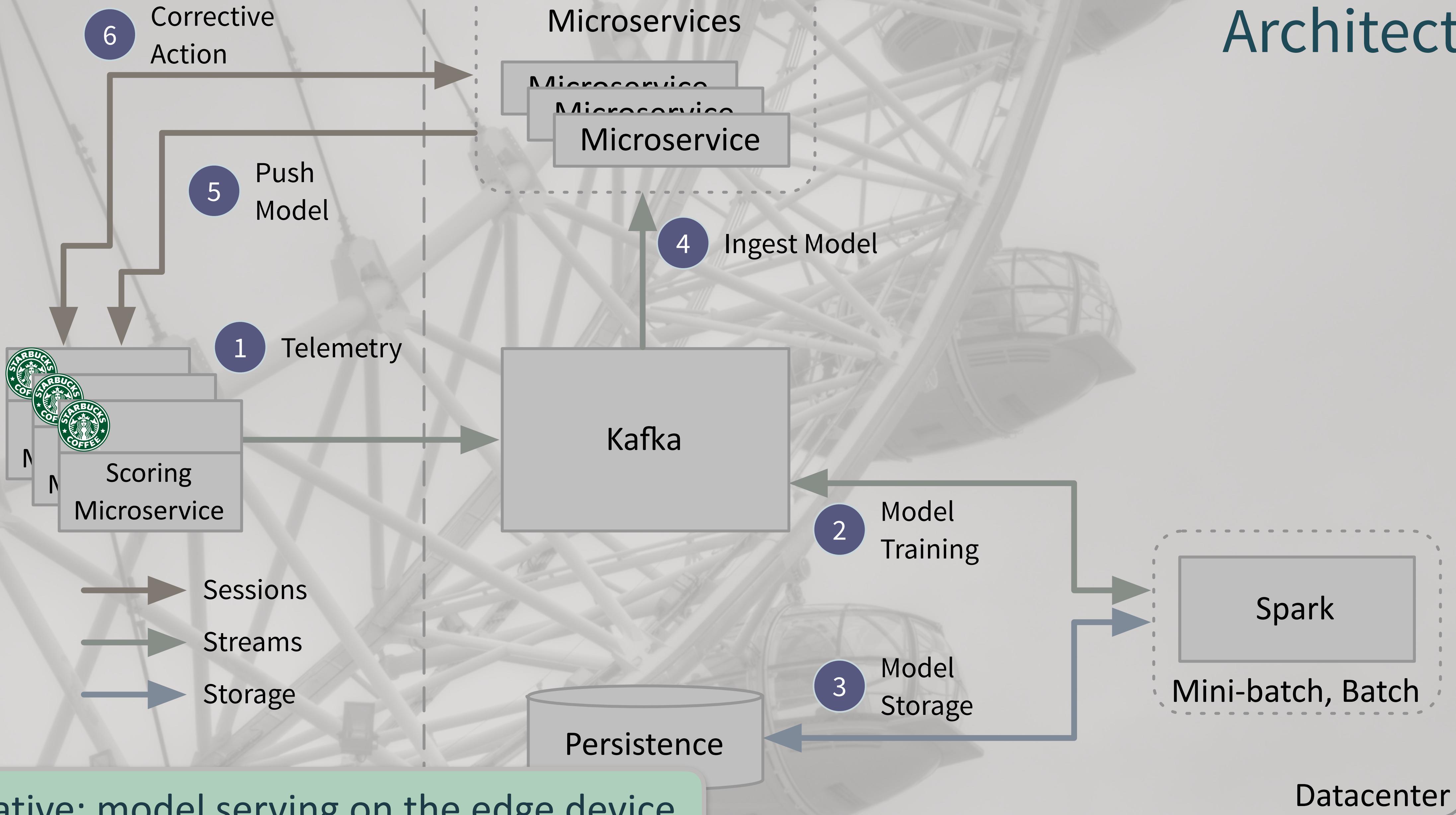
- Real-time consumer and industrial device and supply chain management at scale

# Example Architecture

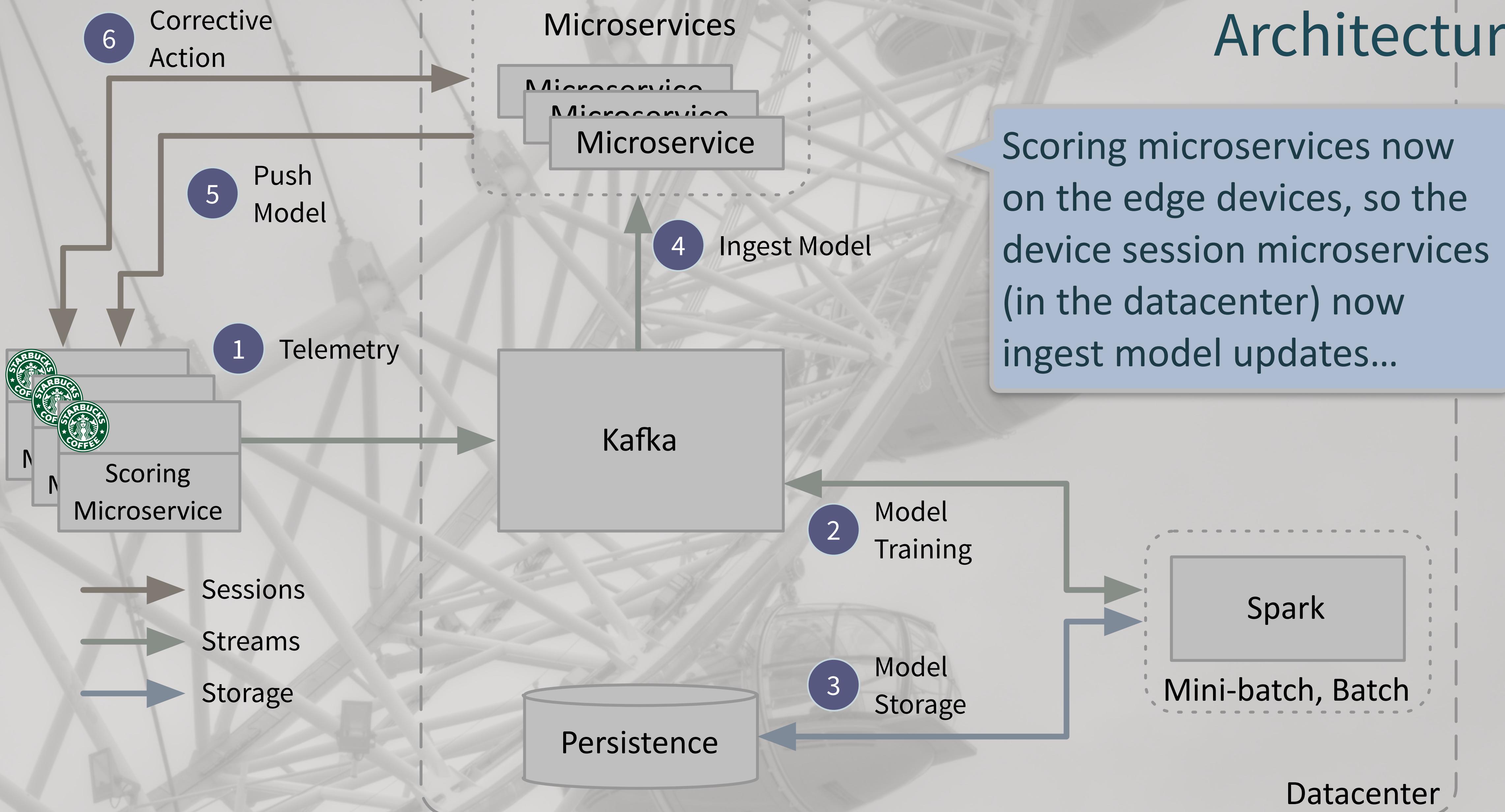


What we just discussed...

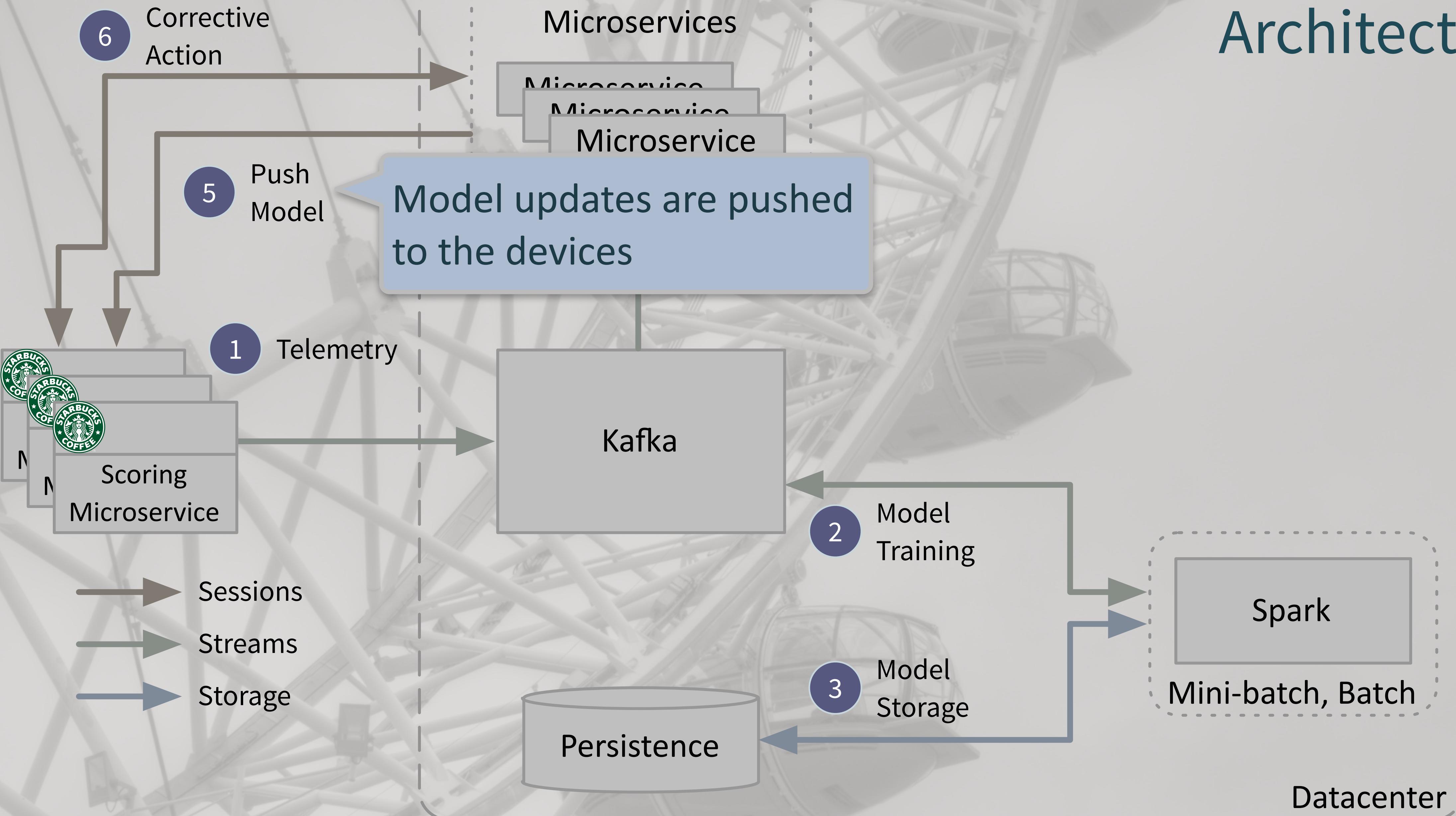
# Edge Scoring Example Architecture



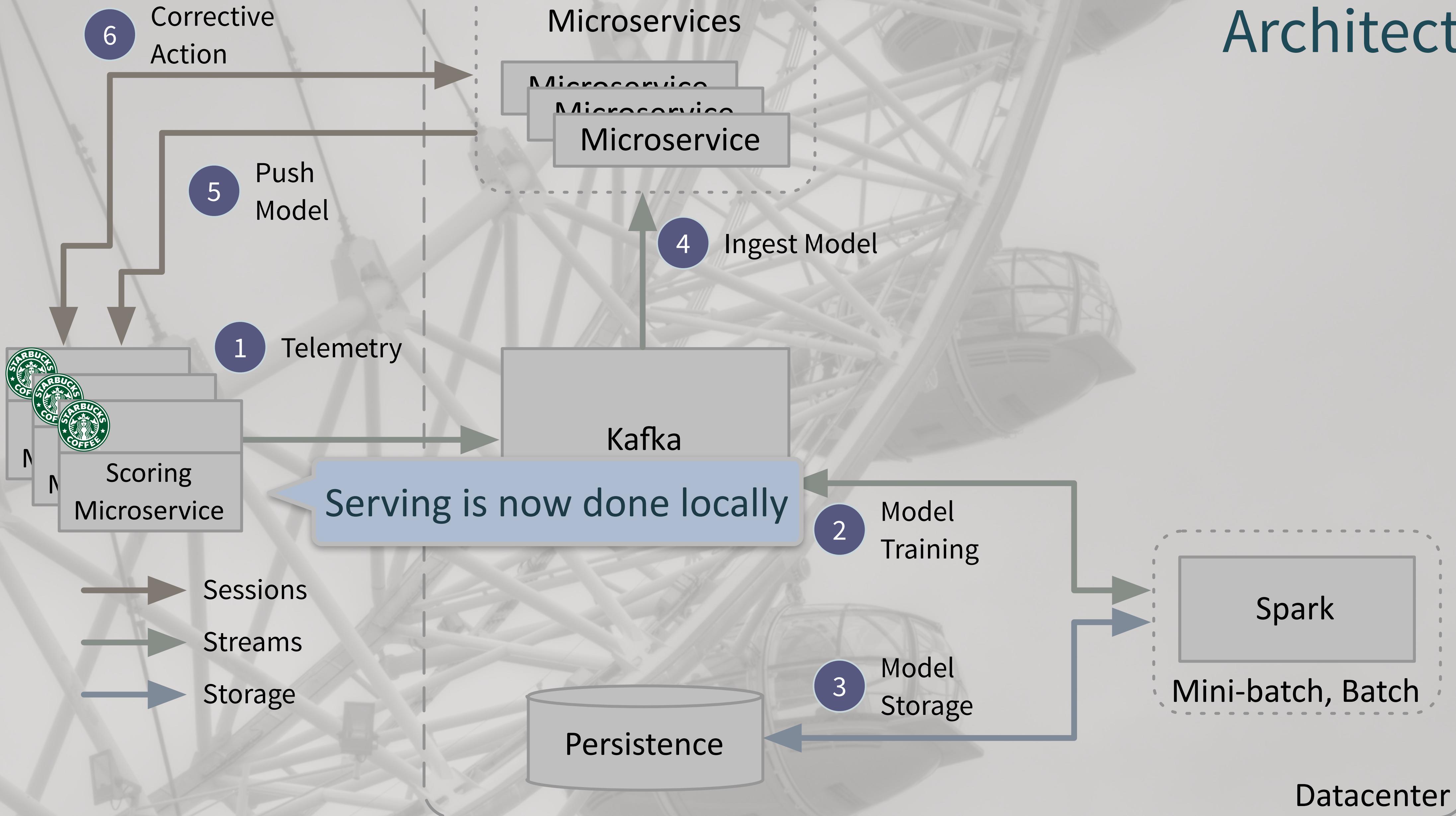
# Edge Scoring Example Architecture



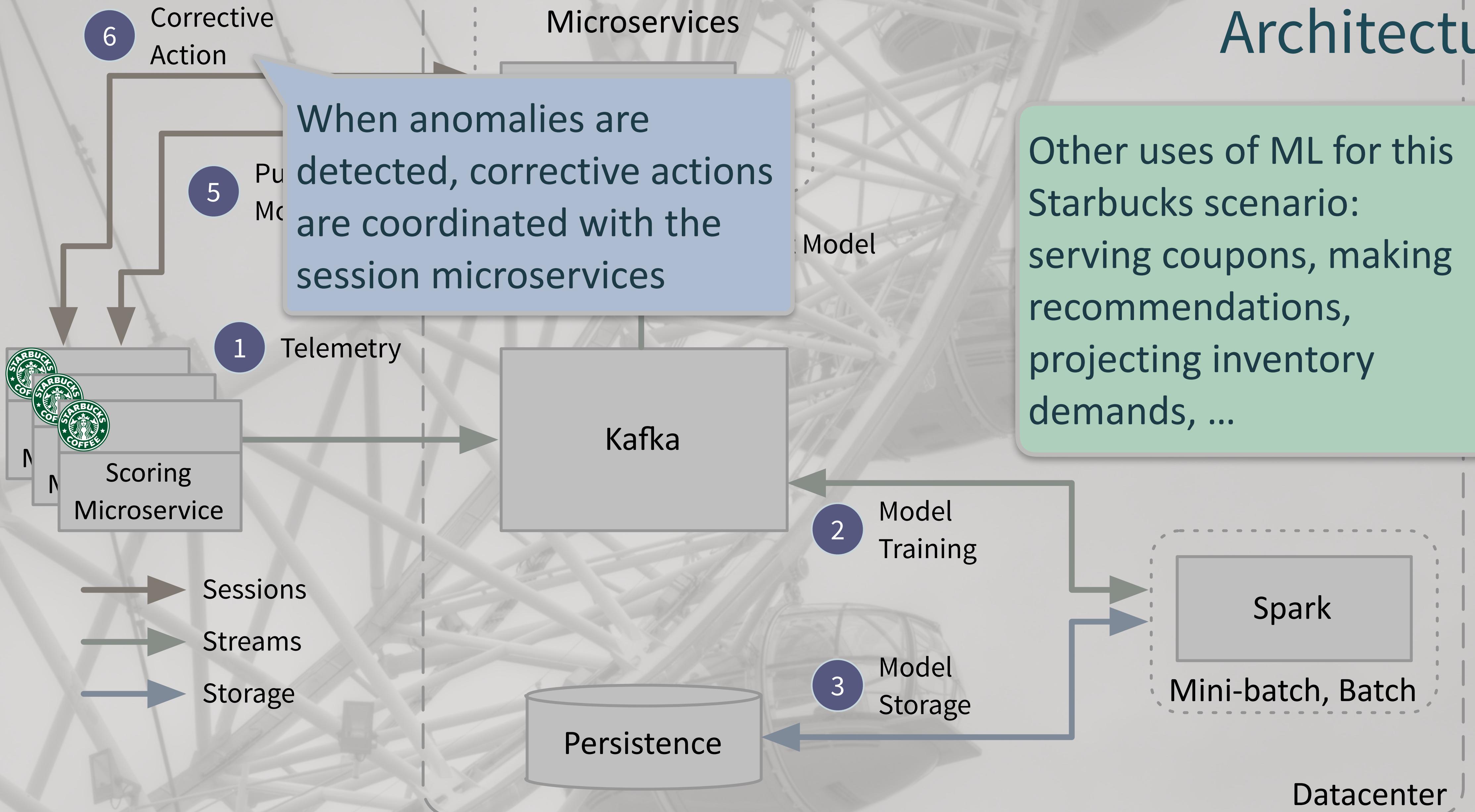
# Edge Scoring Example Architecture



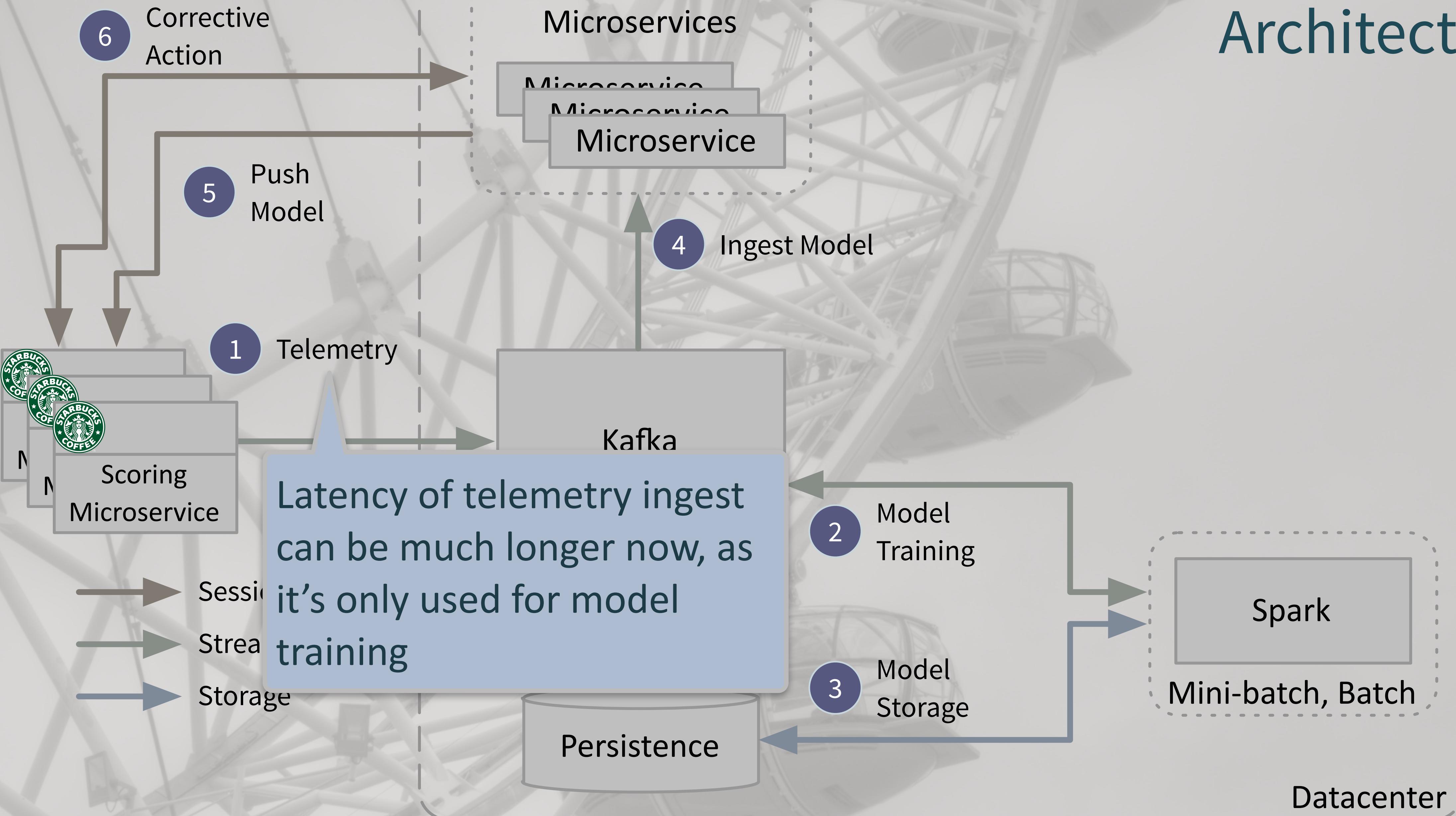
# Edge Scoring Example Architecture



# Edge Scoring Example Architecture



# Edge Scoring Example Architecture



# Fas

Batch changed to streaming  
for competitive advantage

## Predictive Analytics

Apply ML models to large volumes of device data to pre-empt failures / outages



Hewlett Packard  
Enterprise

## IoT

Real-time consumer and industrial Device and Supply Chain management at scale



## Real-time Personalization

Real-time marketing based on behavior, location, inventory levels, product promotions, etc.



RoyalCaribbean  
INTERNATIONAL®

## Real-time Financial Processes

Drive better business outcomes through real-time risk, fraud detection, compliance, audit, governance, etc.





# Technology Choices

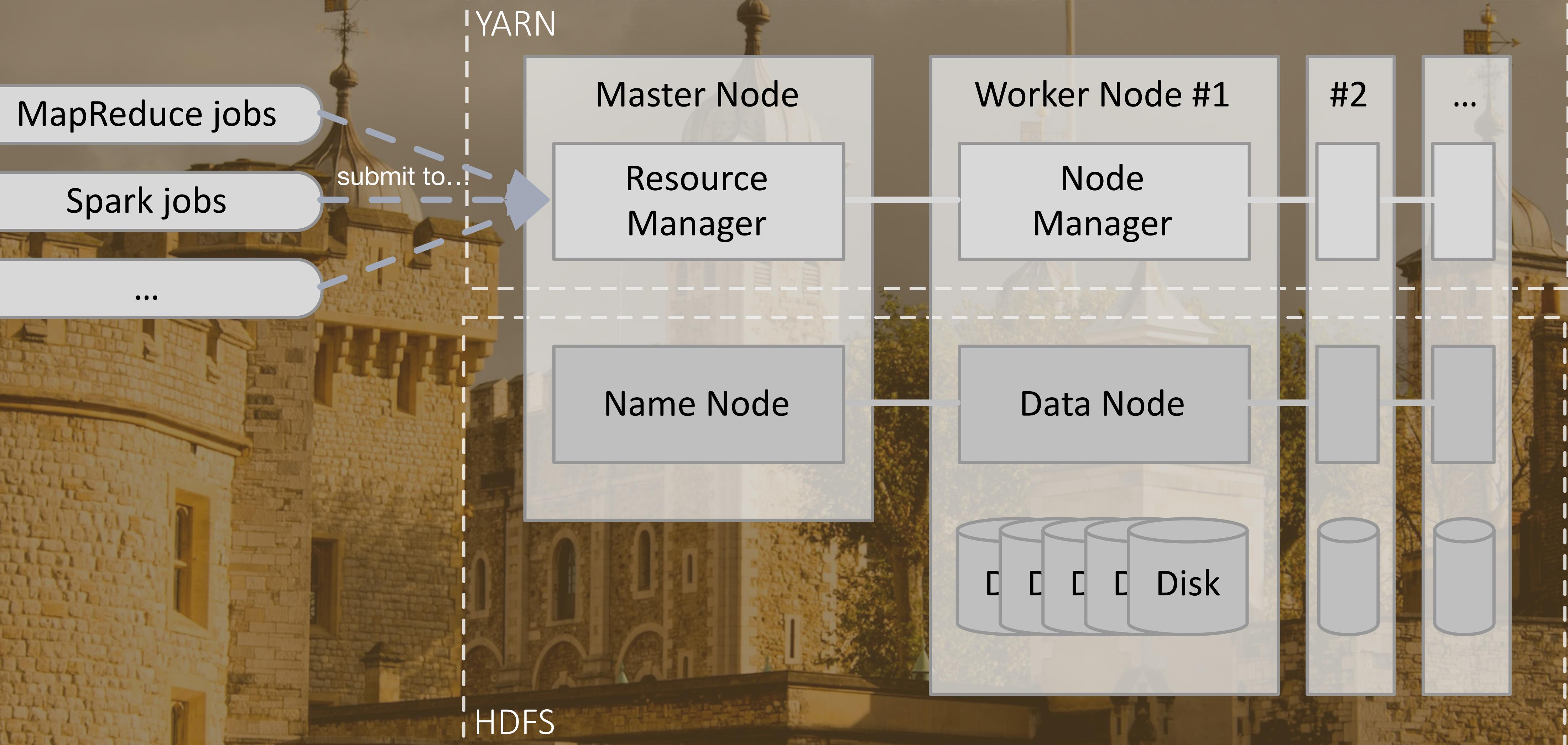
- 
- The background of the slide features a composite image of two iconic London landmarks. On the left, the dome and spire of St. Paul's Cathedral are visible against a cloudy sky. On the right, the modern, curved metal structure of the Millennium Bridge spans the River Thames, with several people walking across it. In the foreground, there's a dark, semi-transparent overlay.
- More than “faster” Hadoop...
  - New architectures that merge data processing with microservices

# Technology Choices

# Recall Hadoop...



- 
- A photograph of a historic stone building, likely a university or institutional complex, featuring multiple towers and domes. A Union Jack flag flies from a pole on top of one of the towers. The building is made of light-colored stone and has several arched windows and doorways. The sky is overcast.
- Data warehouse replacement
  - Historical analysis
  - Interactive exploration
  - Offline training of machine learning models
  - ...



# Resource Management

Compute

MapReduce jobs

Spark jobs

...

submit o...

YARN

Master Node

Resource Manager

Worker Node #1

Node Manager

Worker Node #2

Node Manager

Worker Node #3

Node Manager

Name Node

Data Node

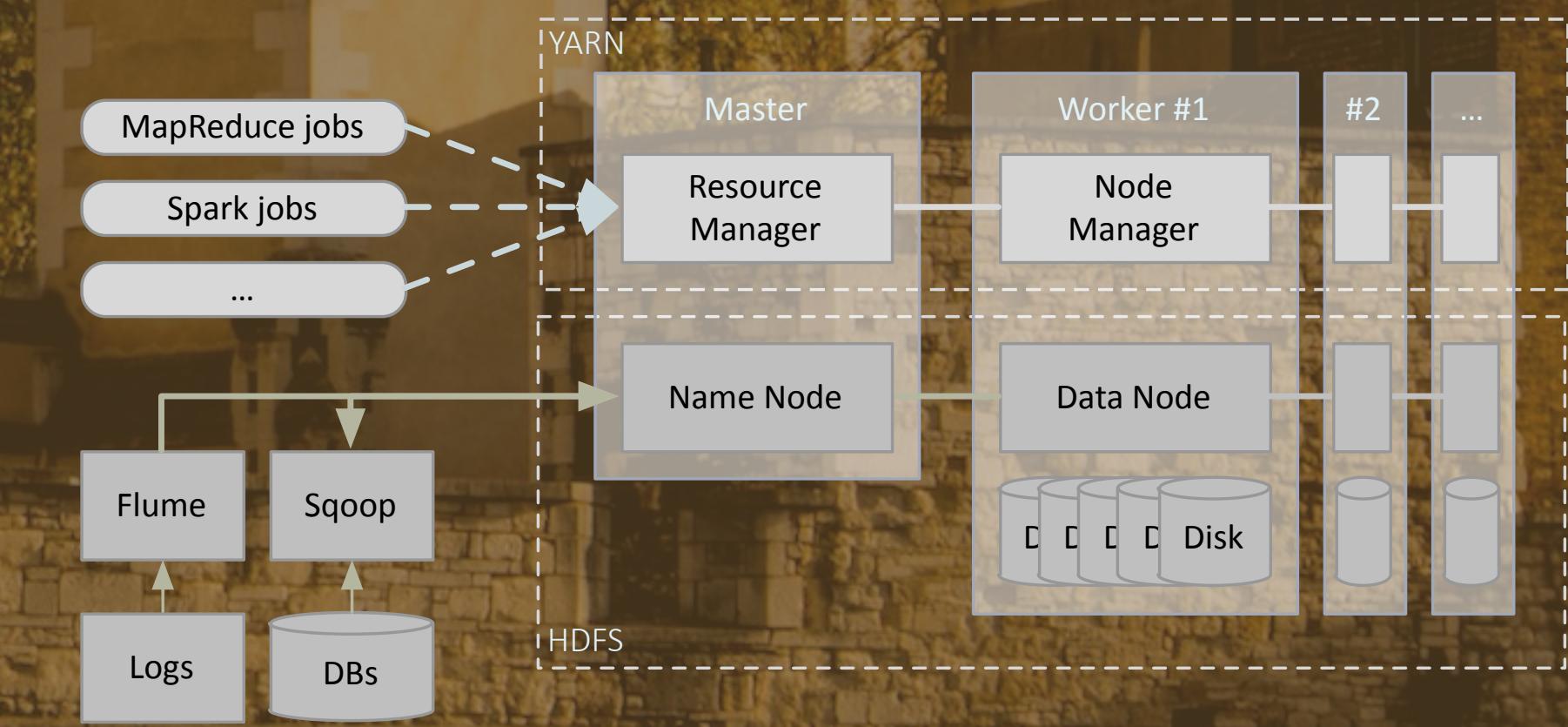
Disk

HDFS

Storage

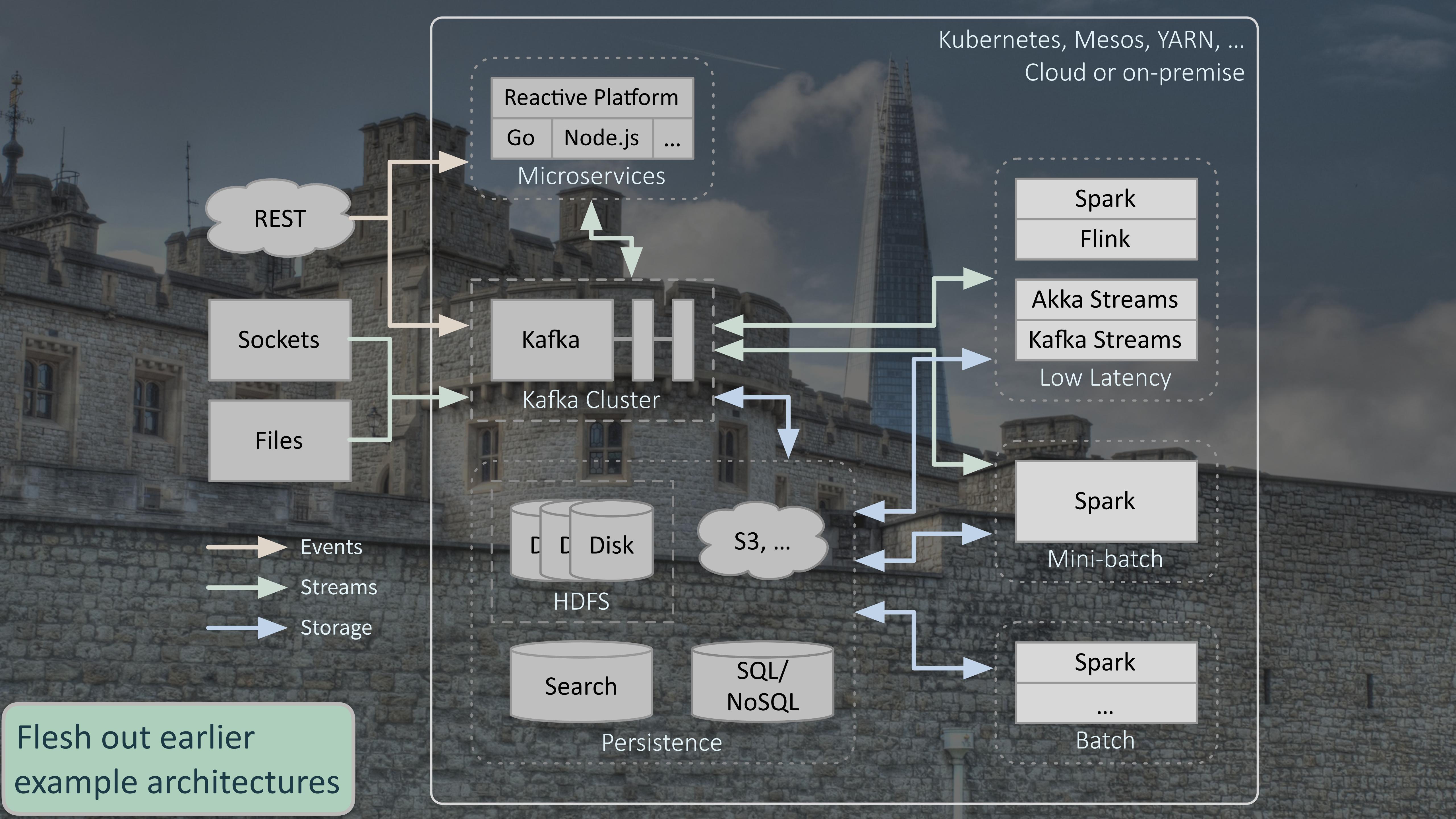
Optimized for storing lots of data *at rest*, with subsequent processing, but not optimized for data *in motion*.

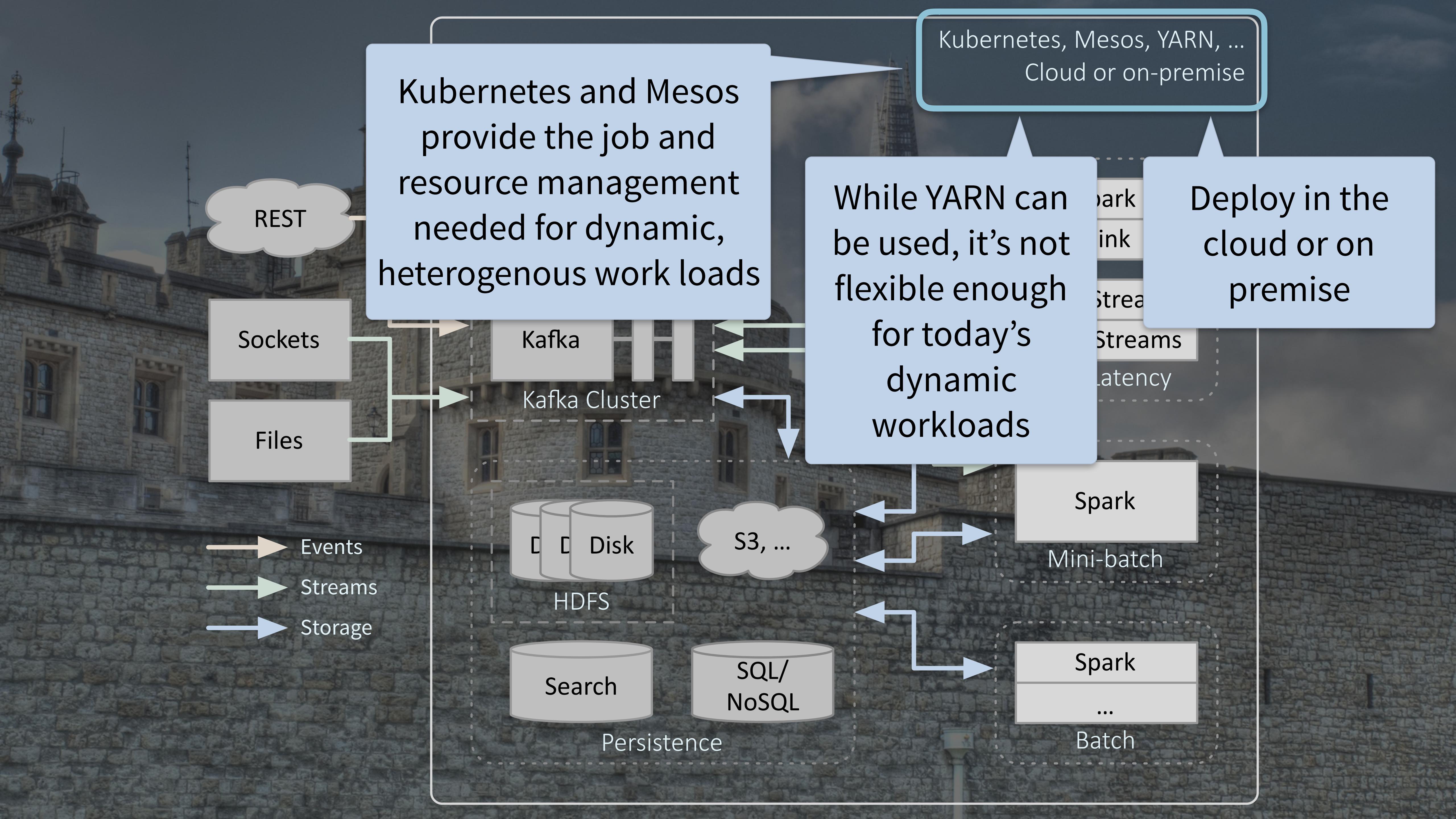
- Hadoop is ideal for batch and interactive apps
- ... but also constrained by that model

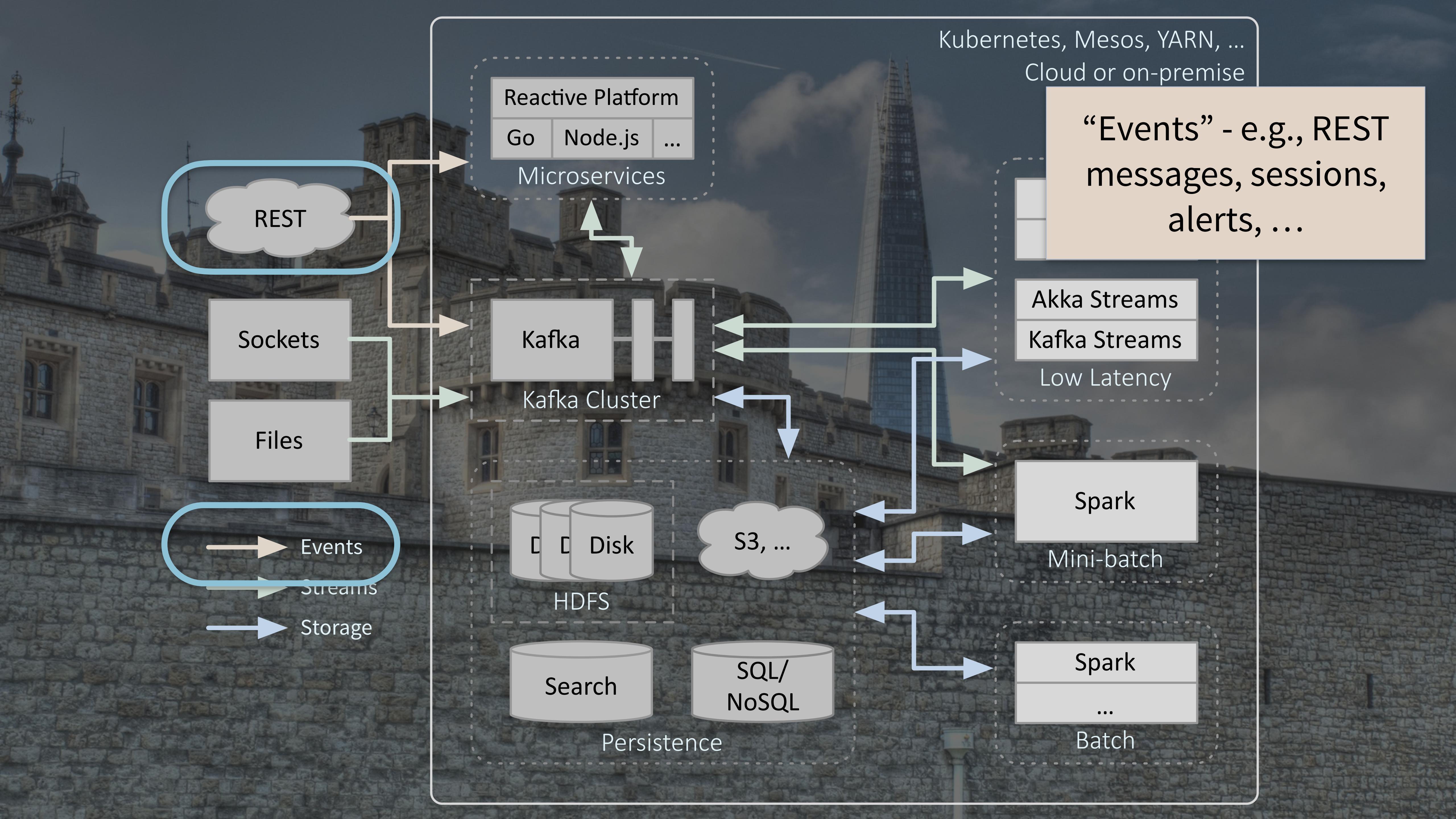


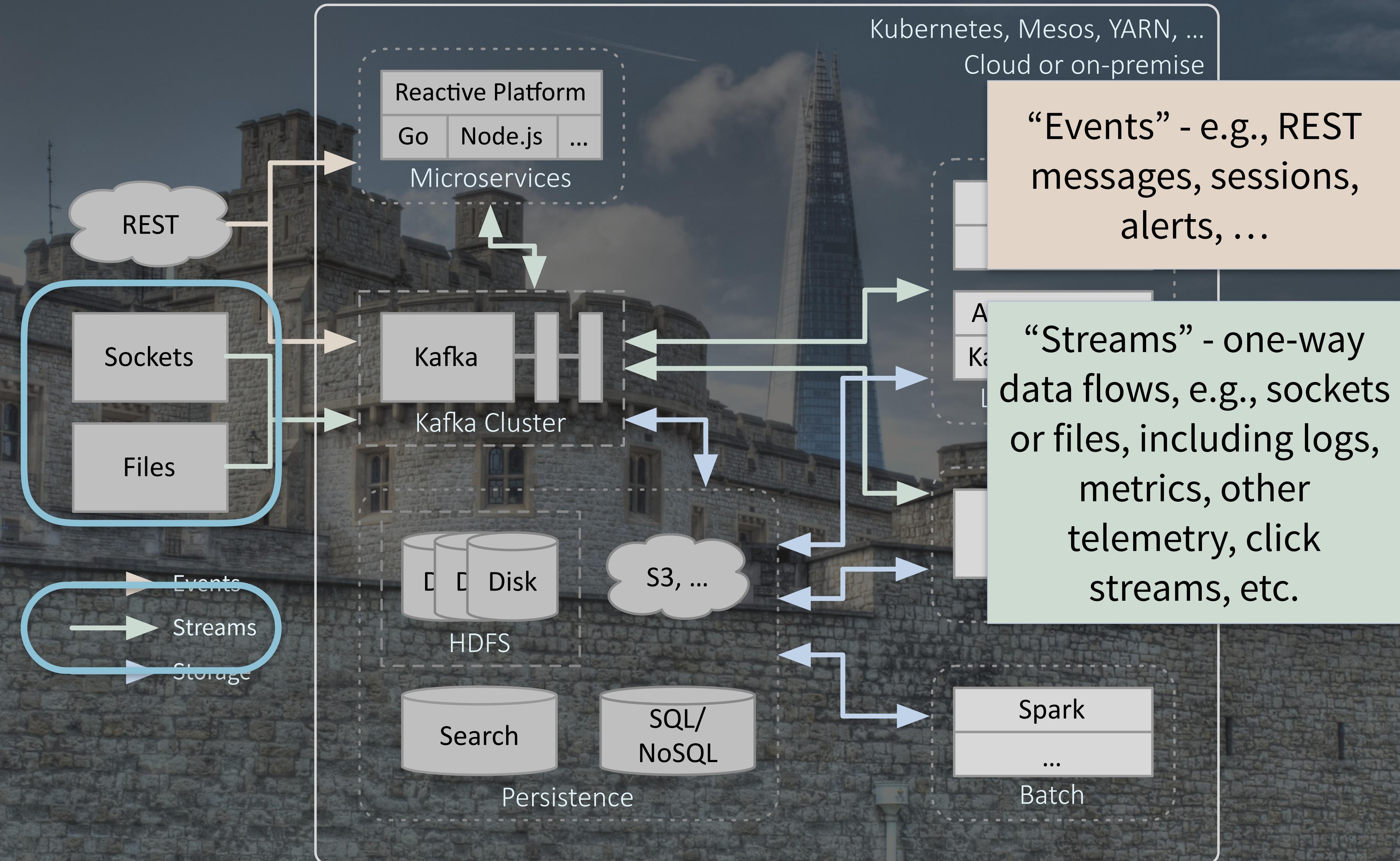


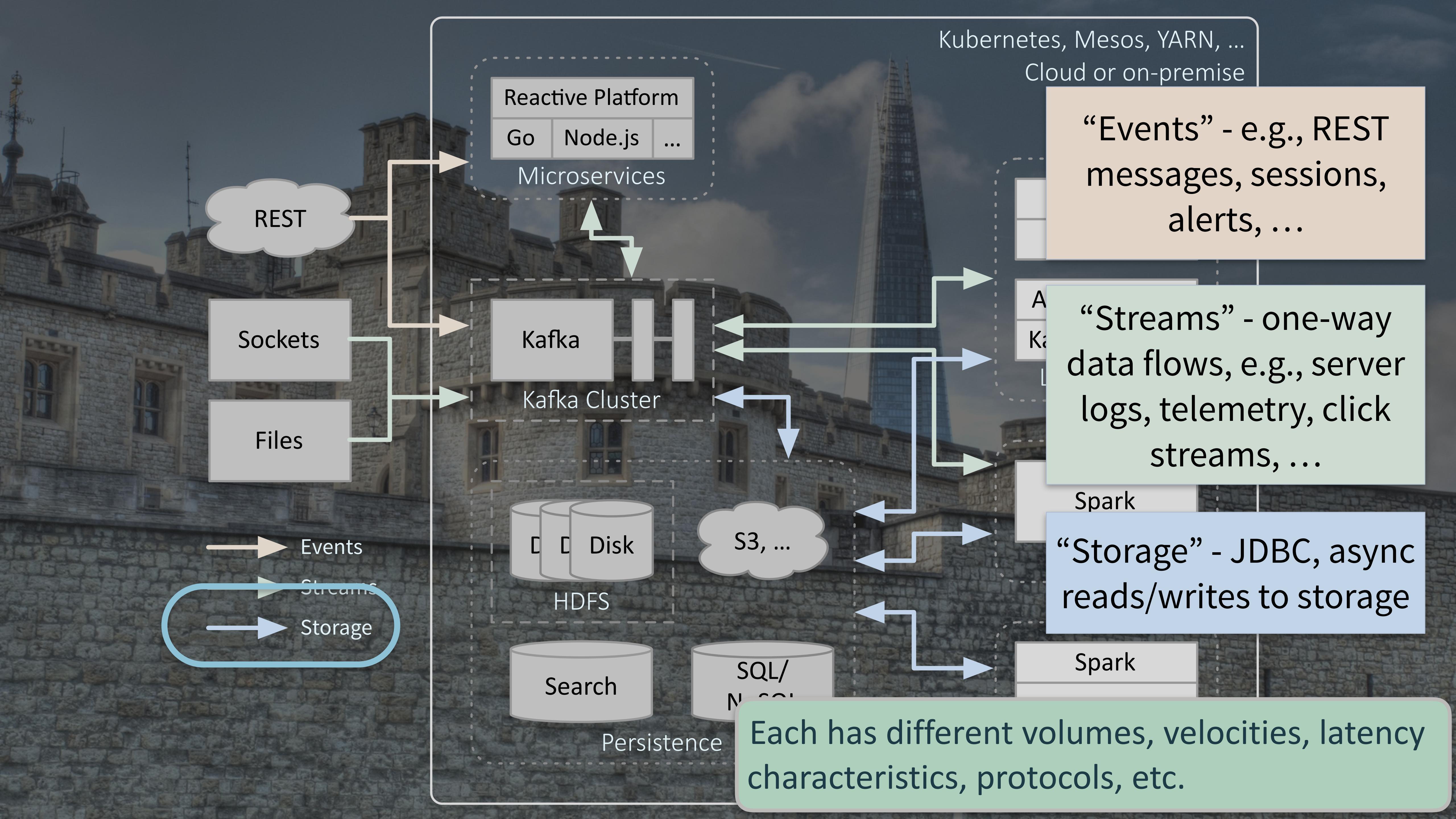
# New Fast Data Architecture

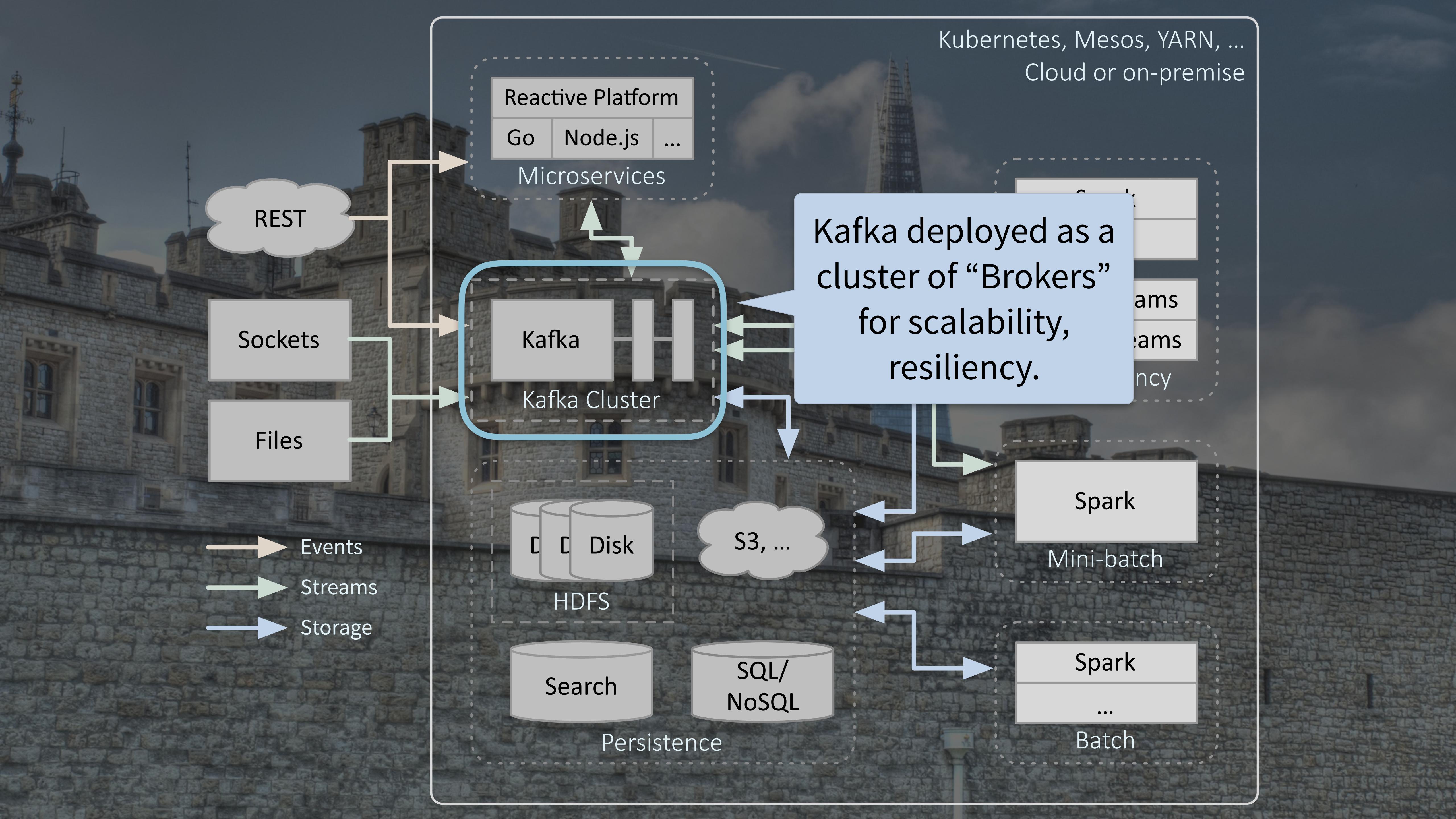


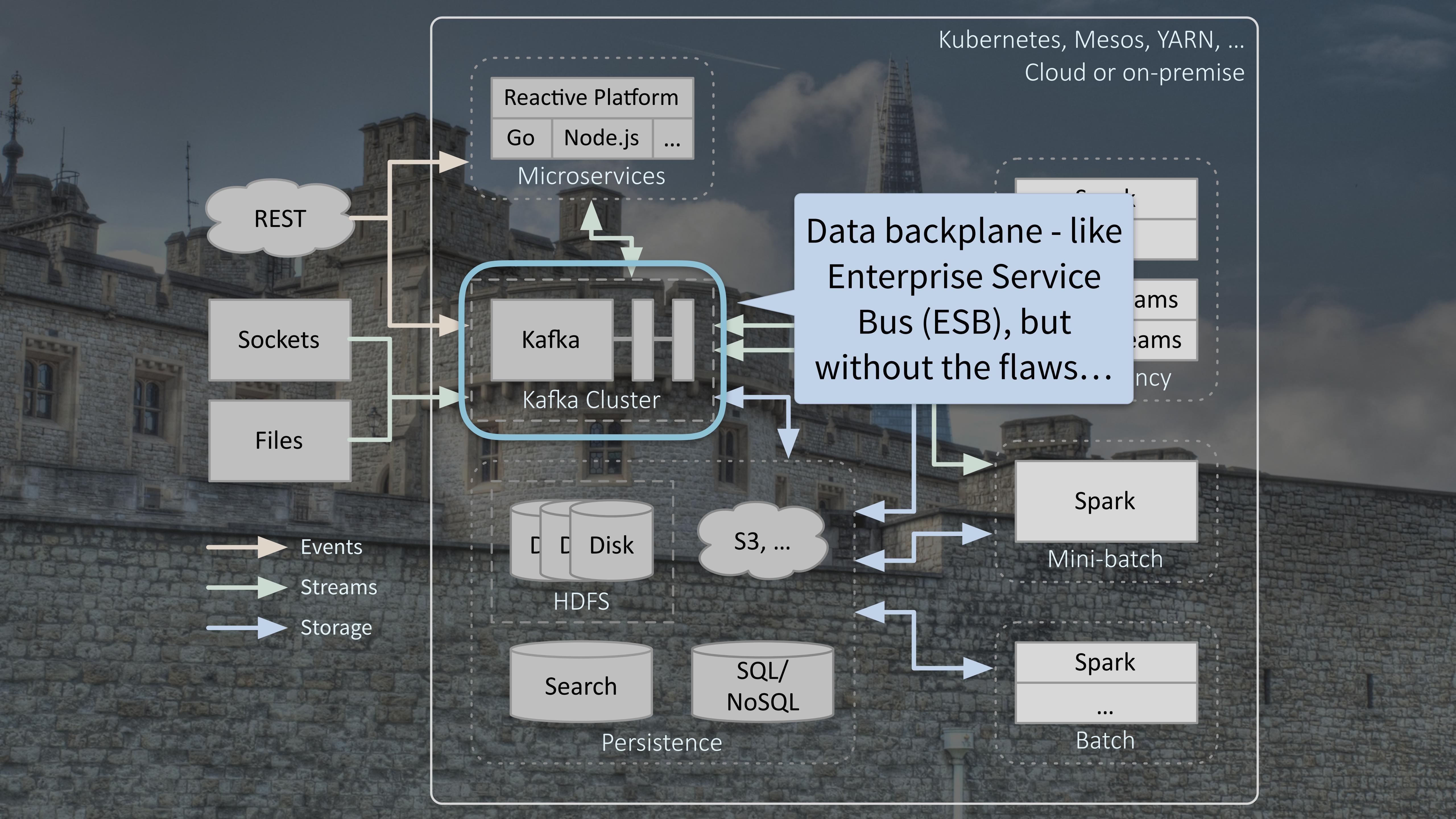








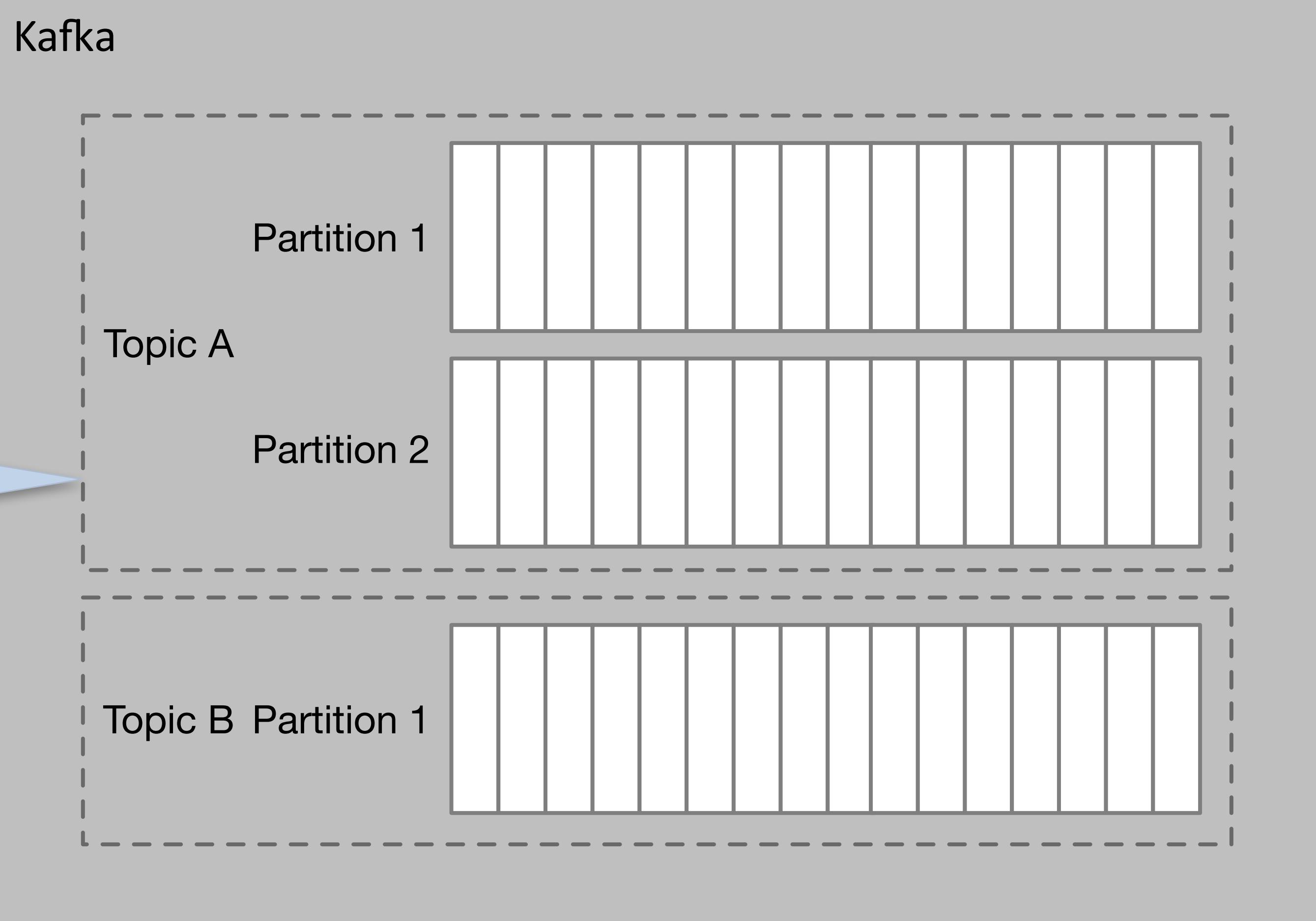




# Why Kafka?

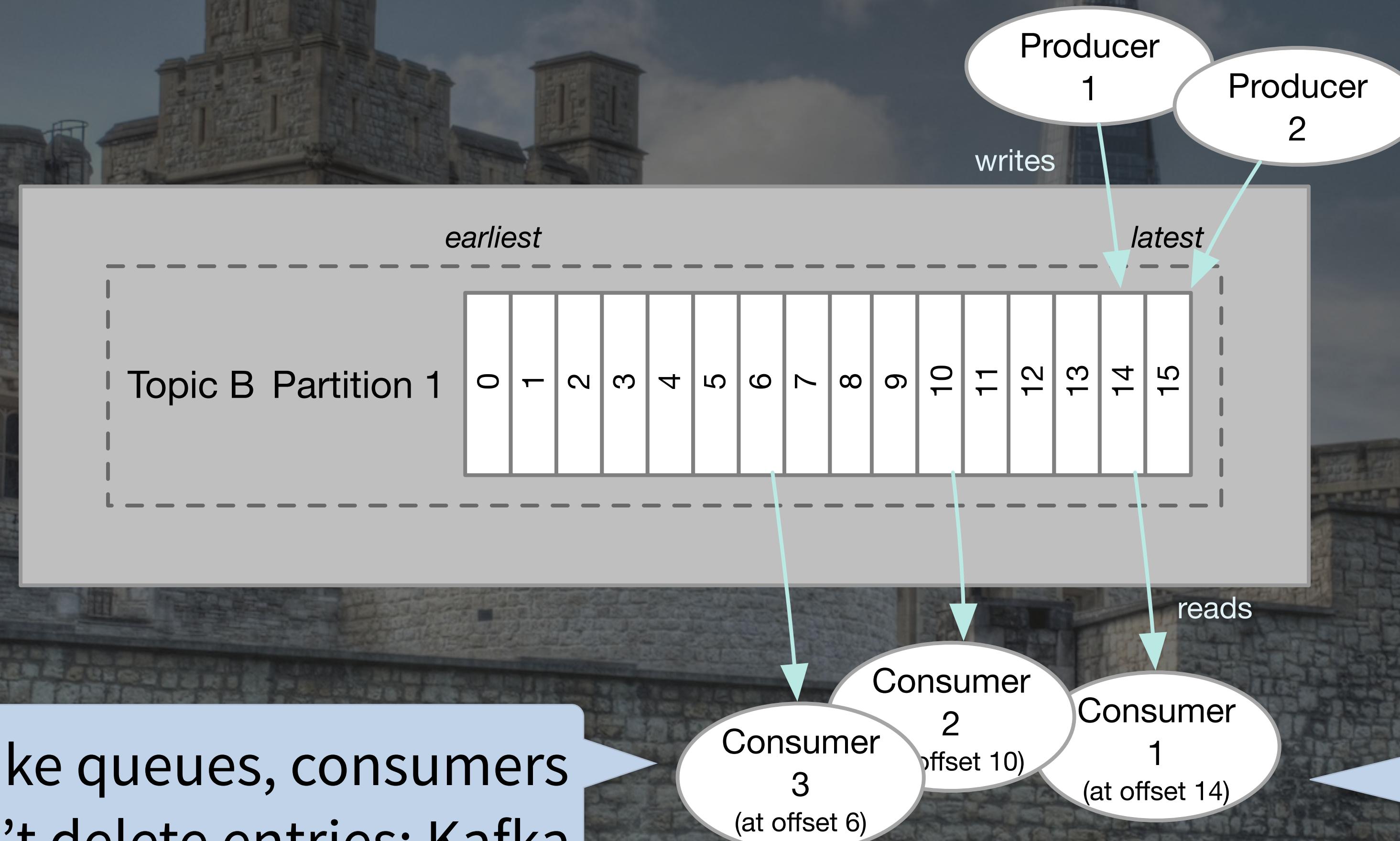
Organized into topics

Topics are partitioned, replicated, and distributed



# Why Kafka?

Logs, not queues!

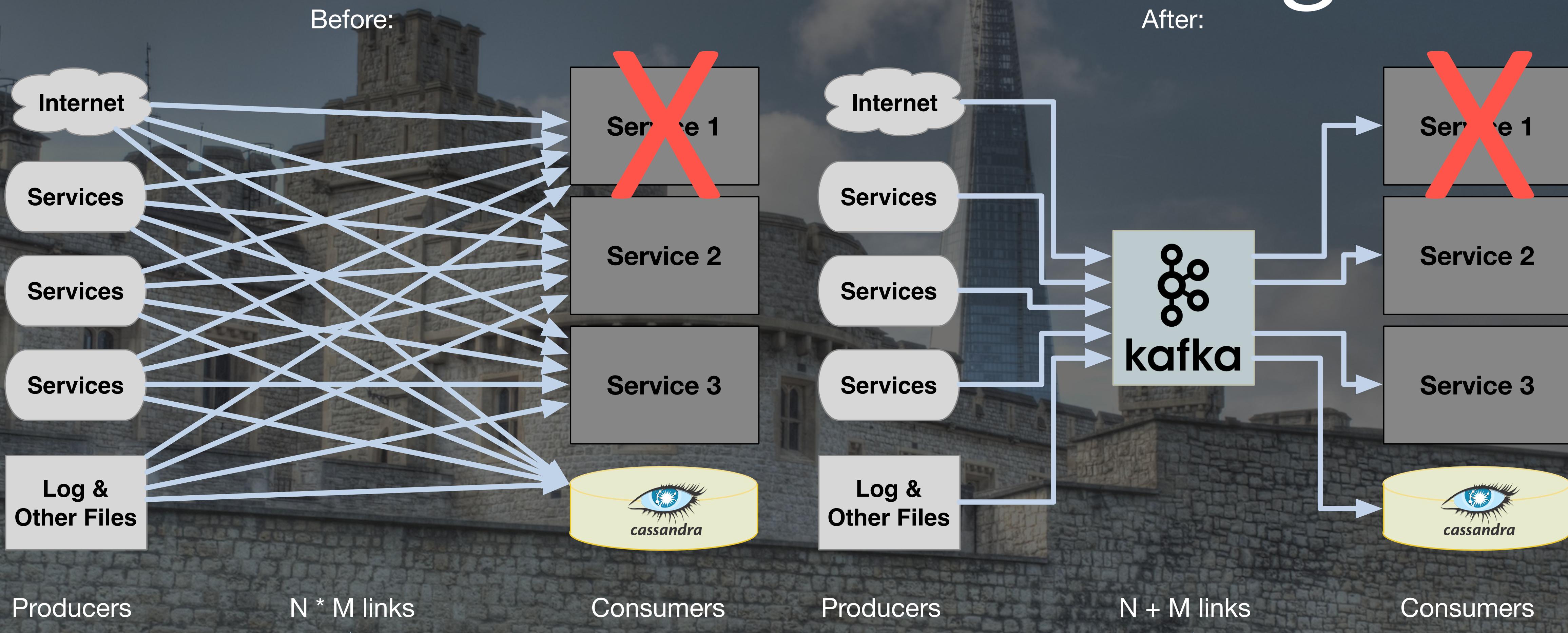


Unlike queues, consumers don't delete entries; Kafka manages their lifecycles

N Consumers,  
who start  
reading where  
they want

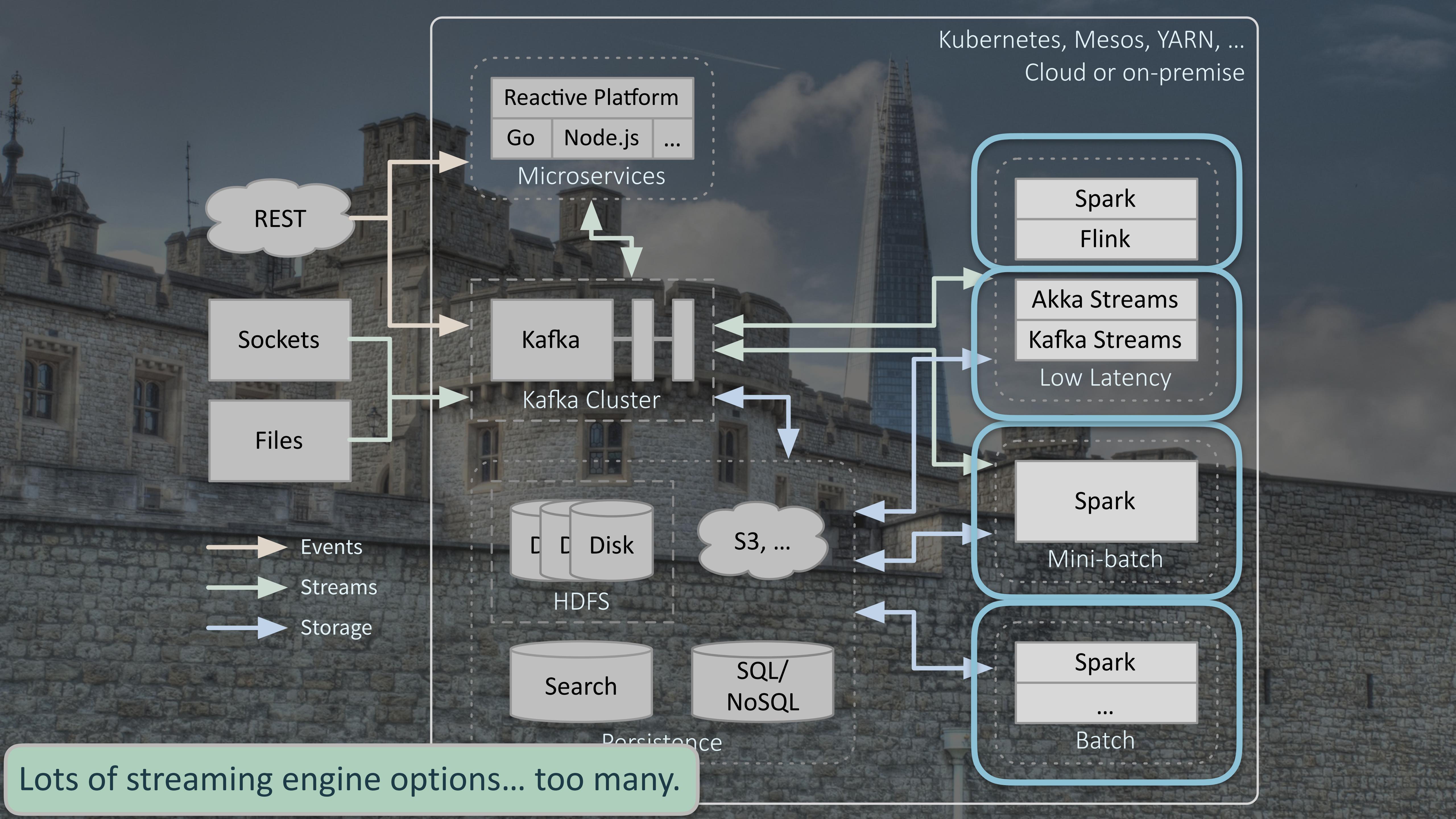
M Producers

# Using Kafka



Messy and fragile;  
what if “Service 1”  
goes down?

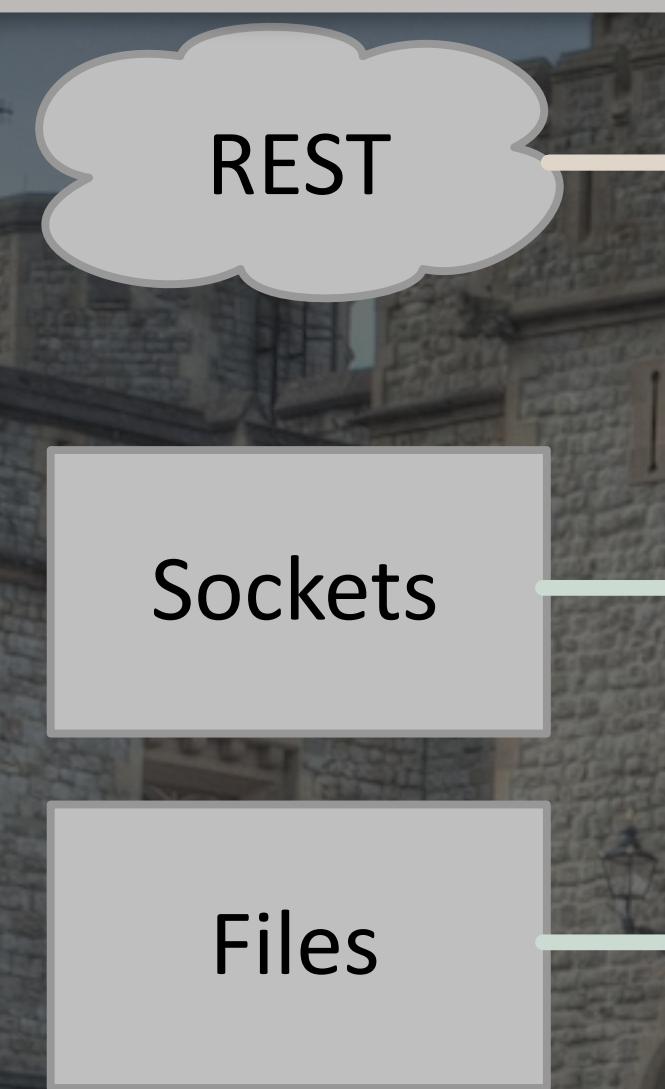
Simpler and more  
robust! Loss of Service  
1 means no data loss.



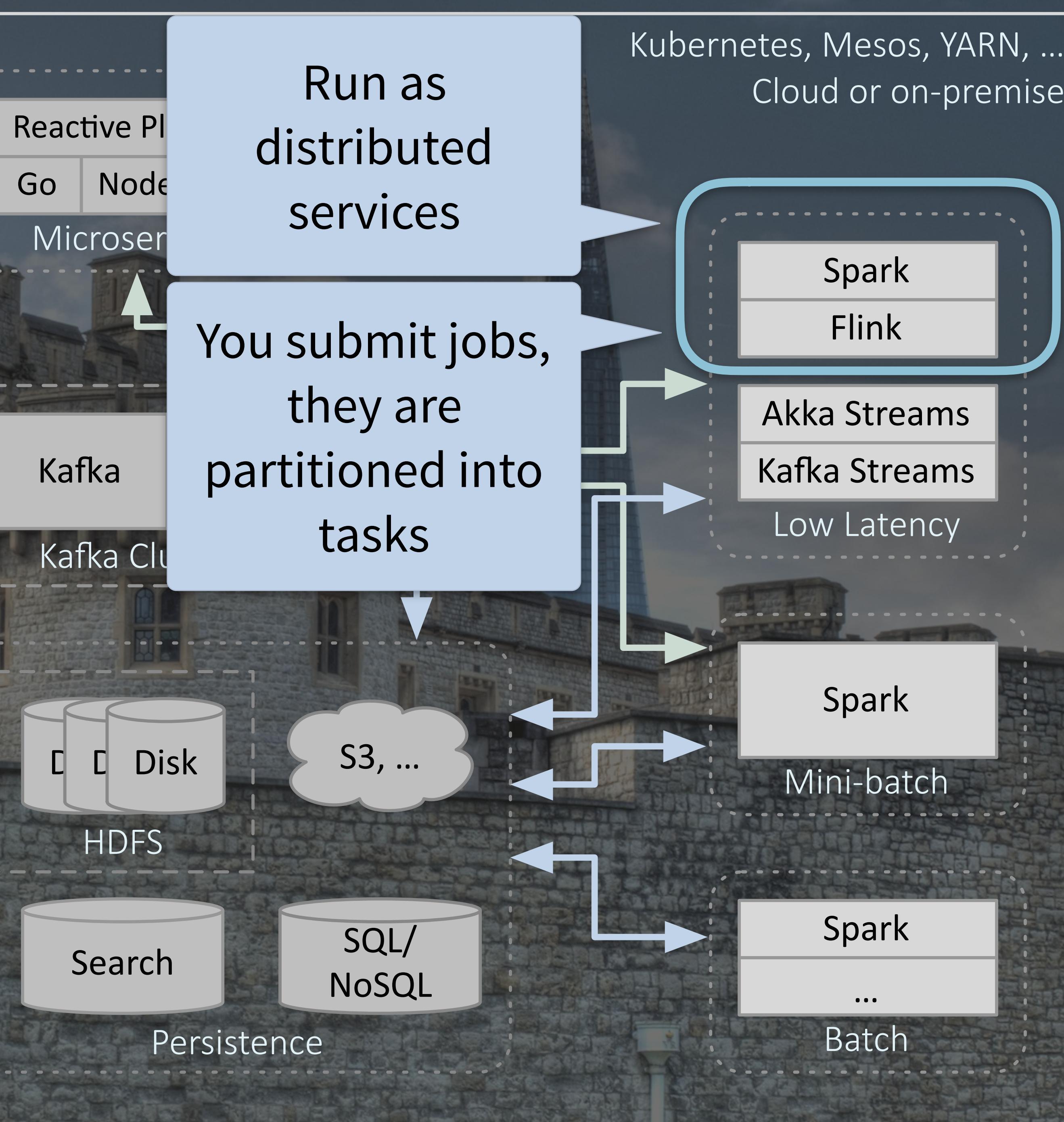
# How do you choose?

- Latency: how low?
- Volume per unit time: how high?
- Data processing: which kinds?
- Build, deploy, and manage services: what are your preferences?

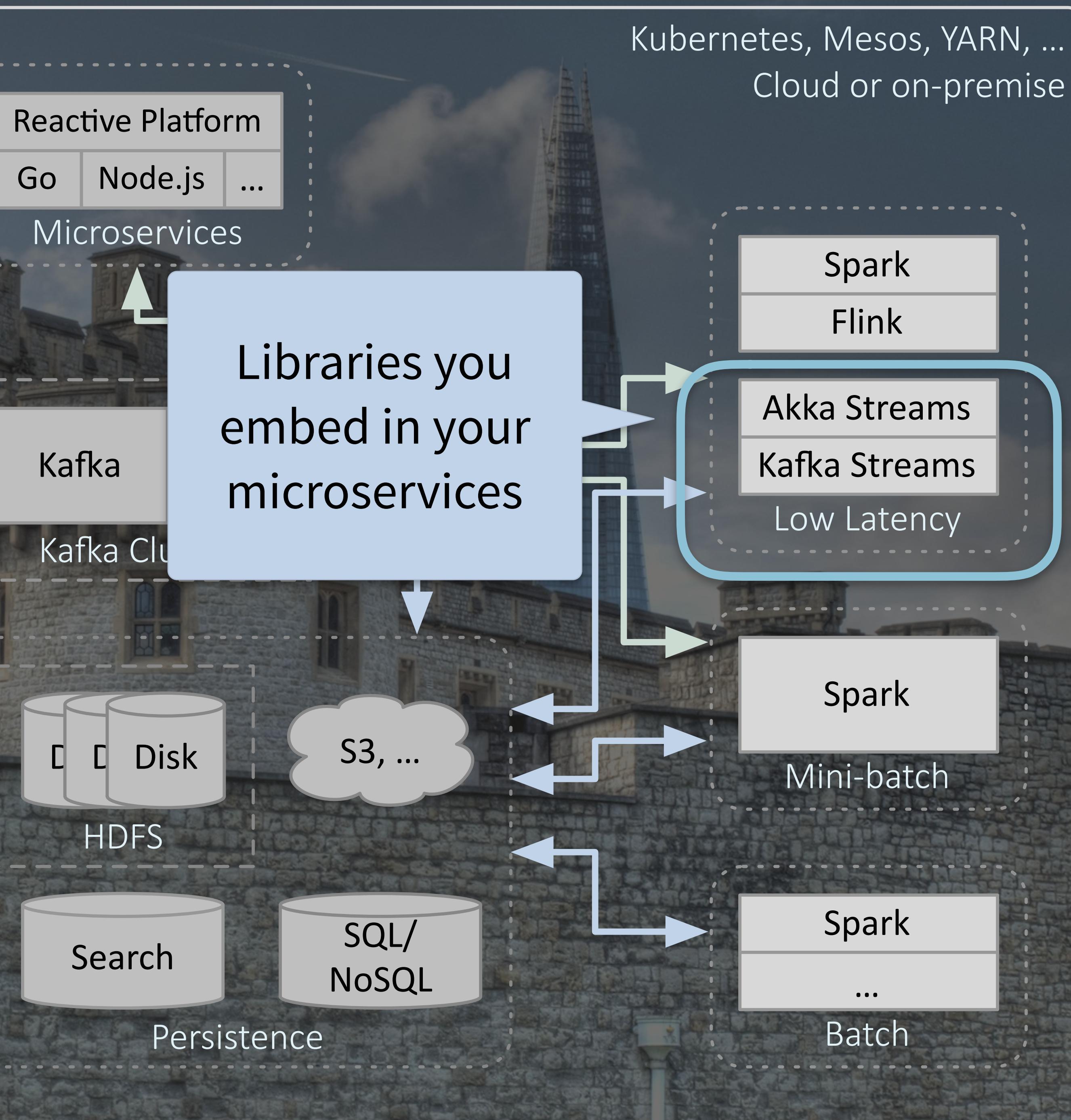
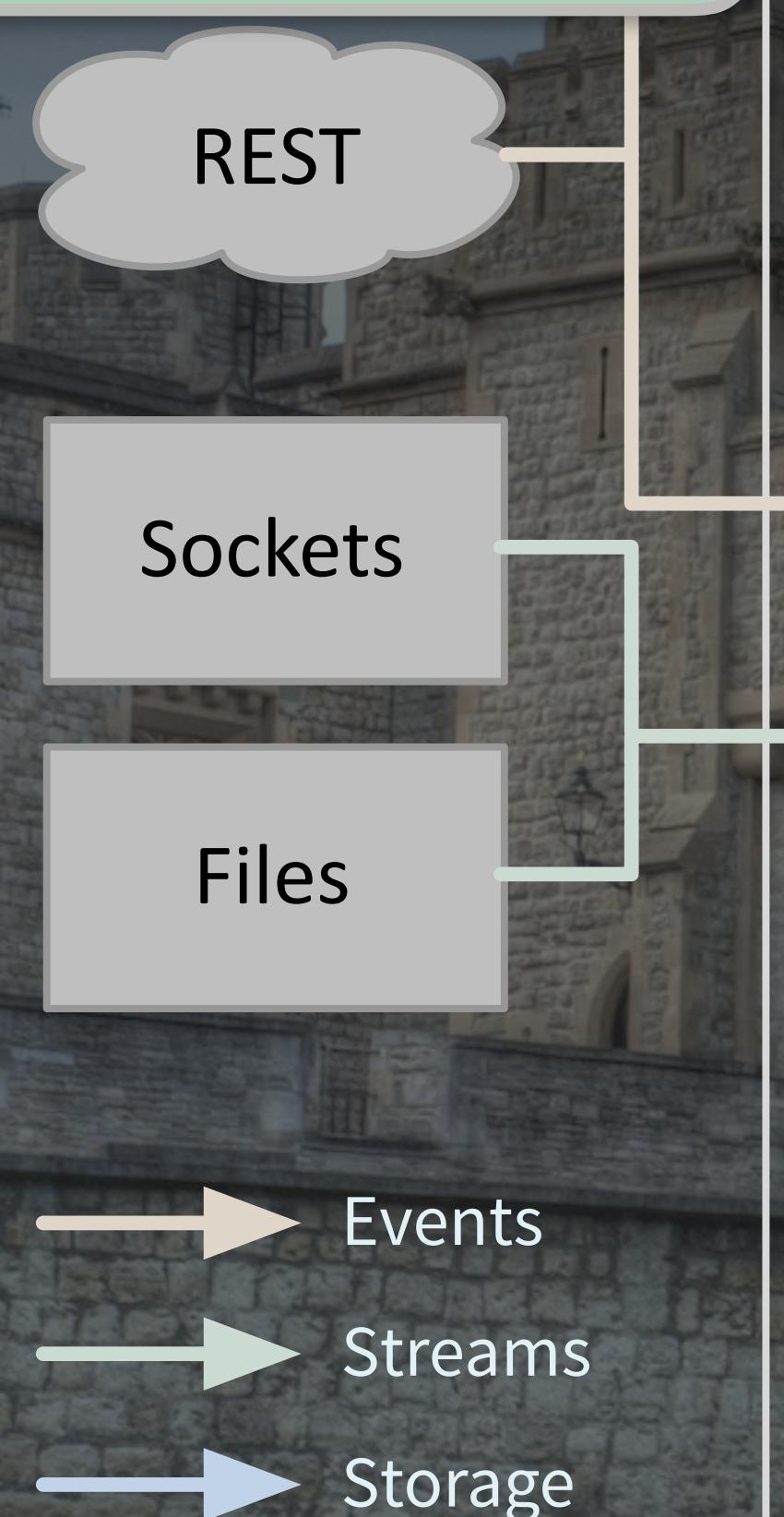
The streaming engines form two groups:

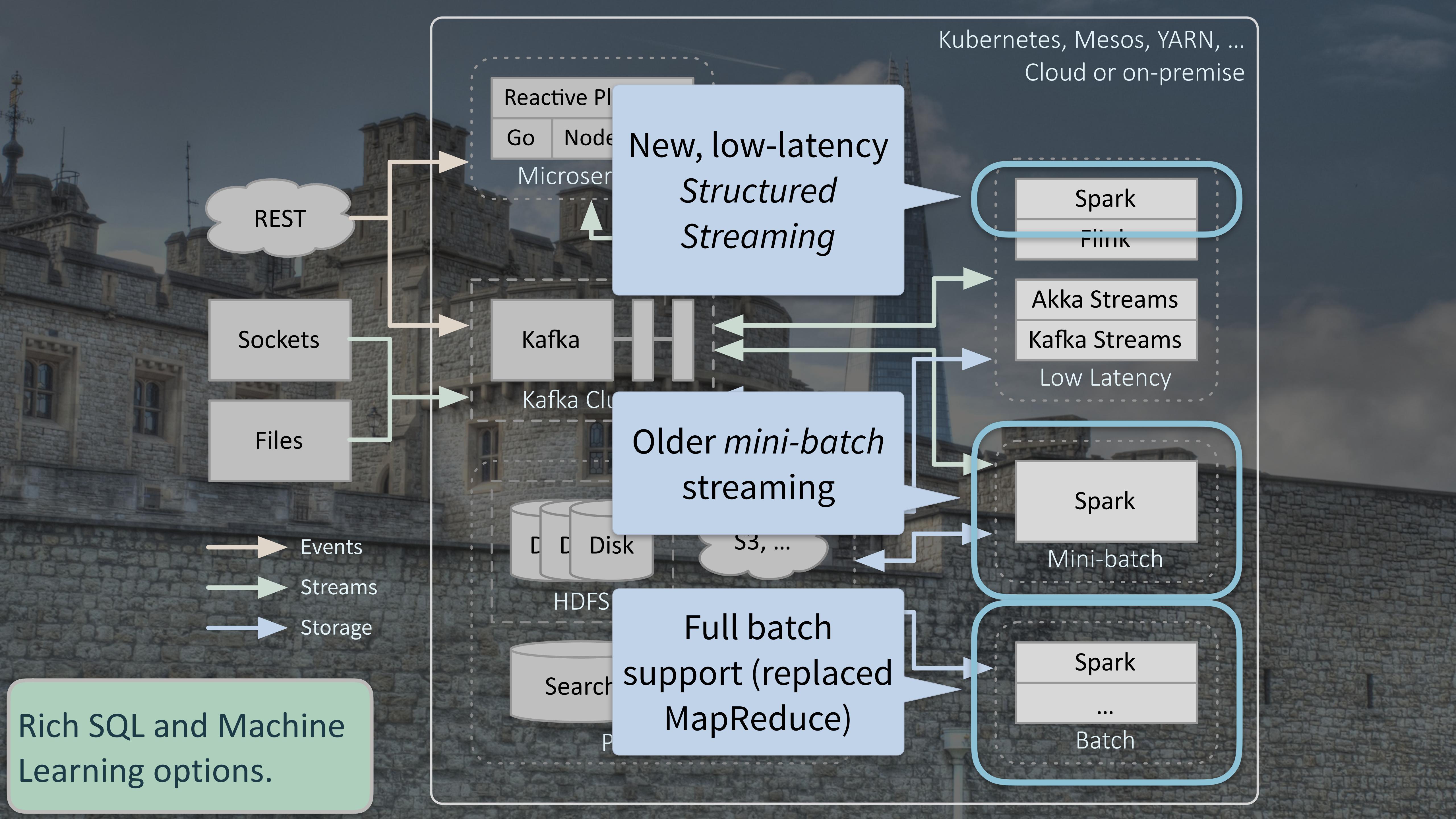


Events  
Streams  
Storage



The streaming engines form two groups:





Rich SQL and Machine  
Learning options.

New, low-latency  
Structured  
Streaming

Older *mini-batch*  
streaming

Full batch  
support (replaced  
MapReduce)

Kubernetes, Mesos, YARN, ...  
Cloud or on-premise

Spark  
Flink

Akka Streams  
Kafka Streams

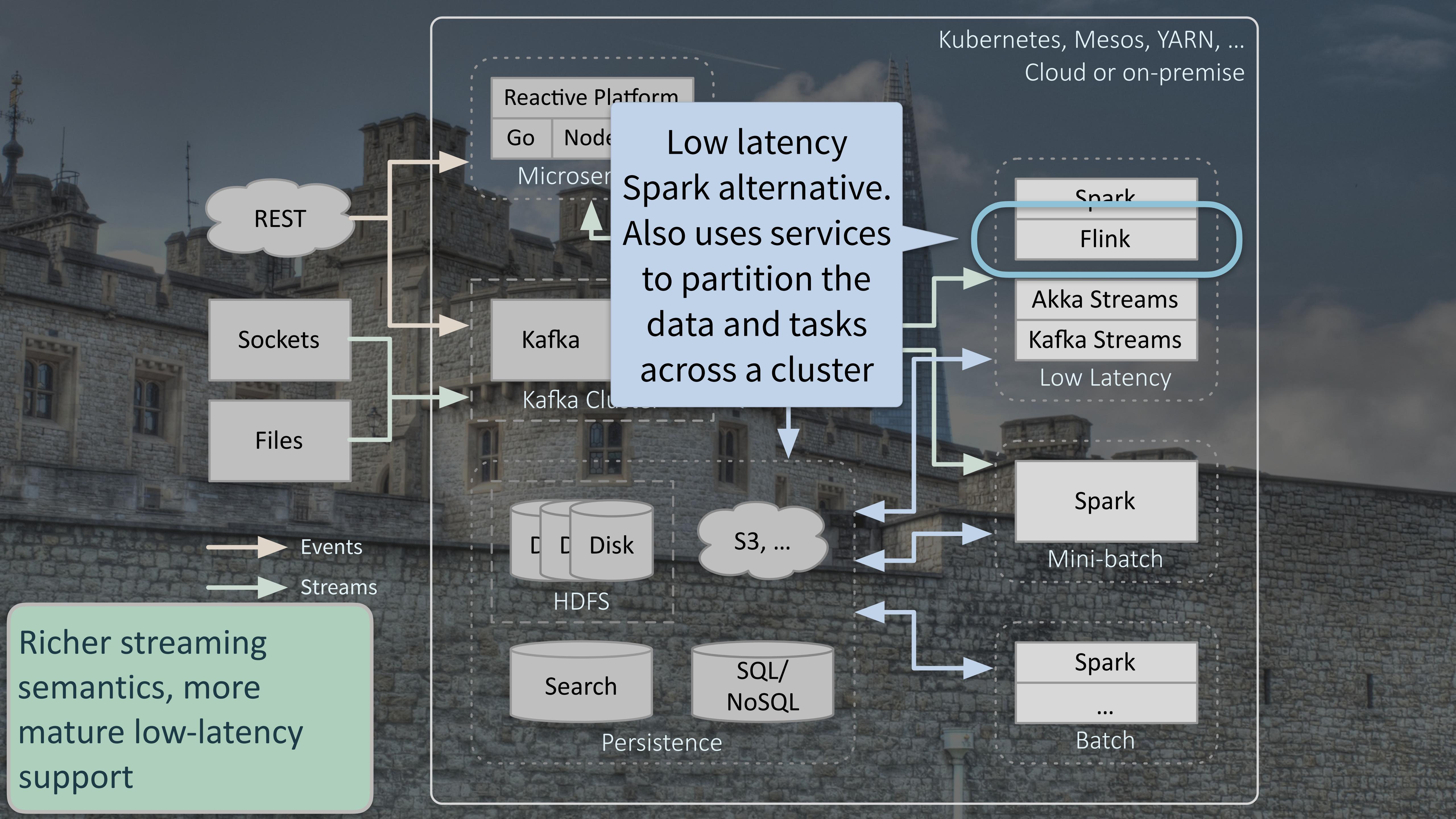
Low Latency

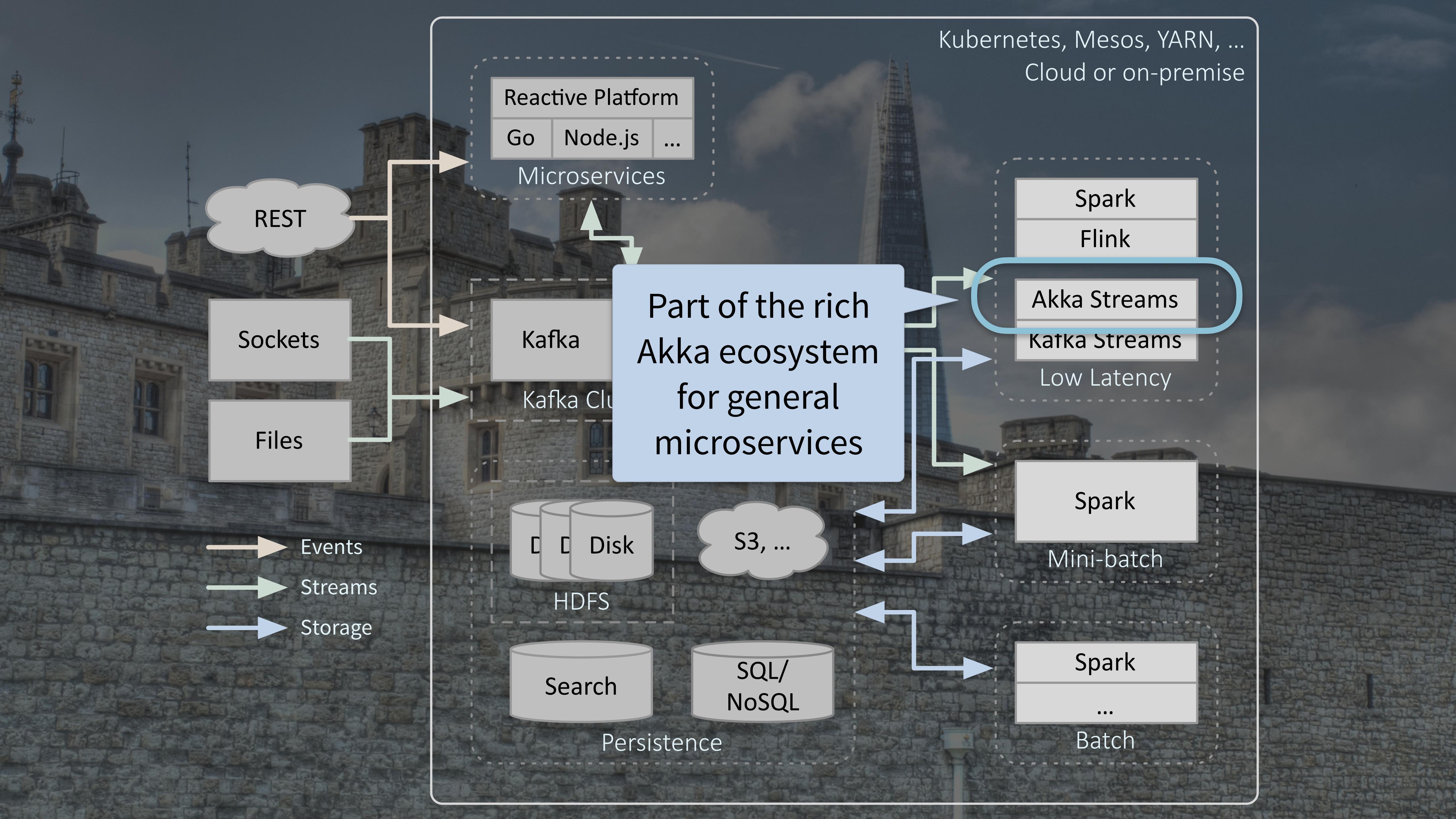
Spark

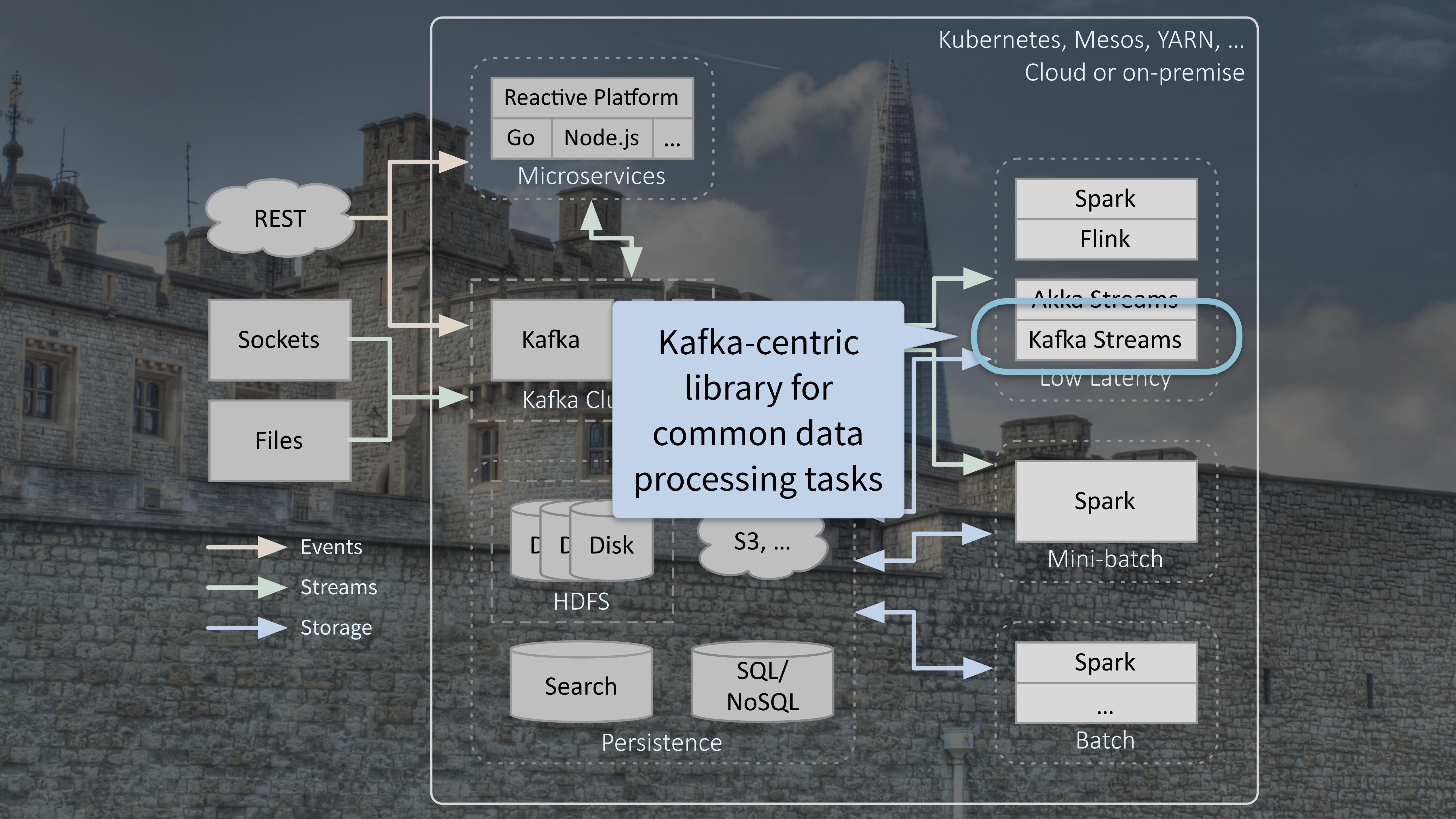
Mini-batch

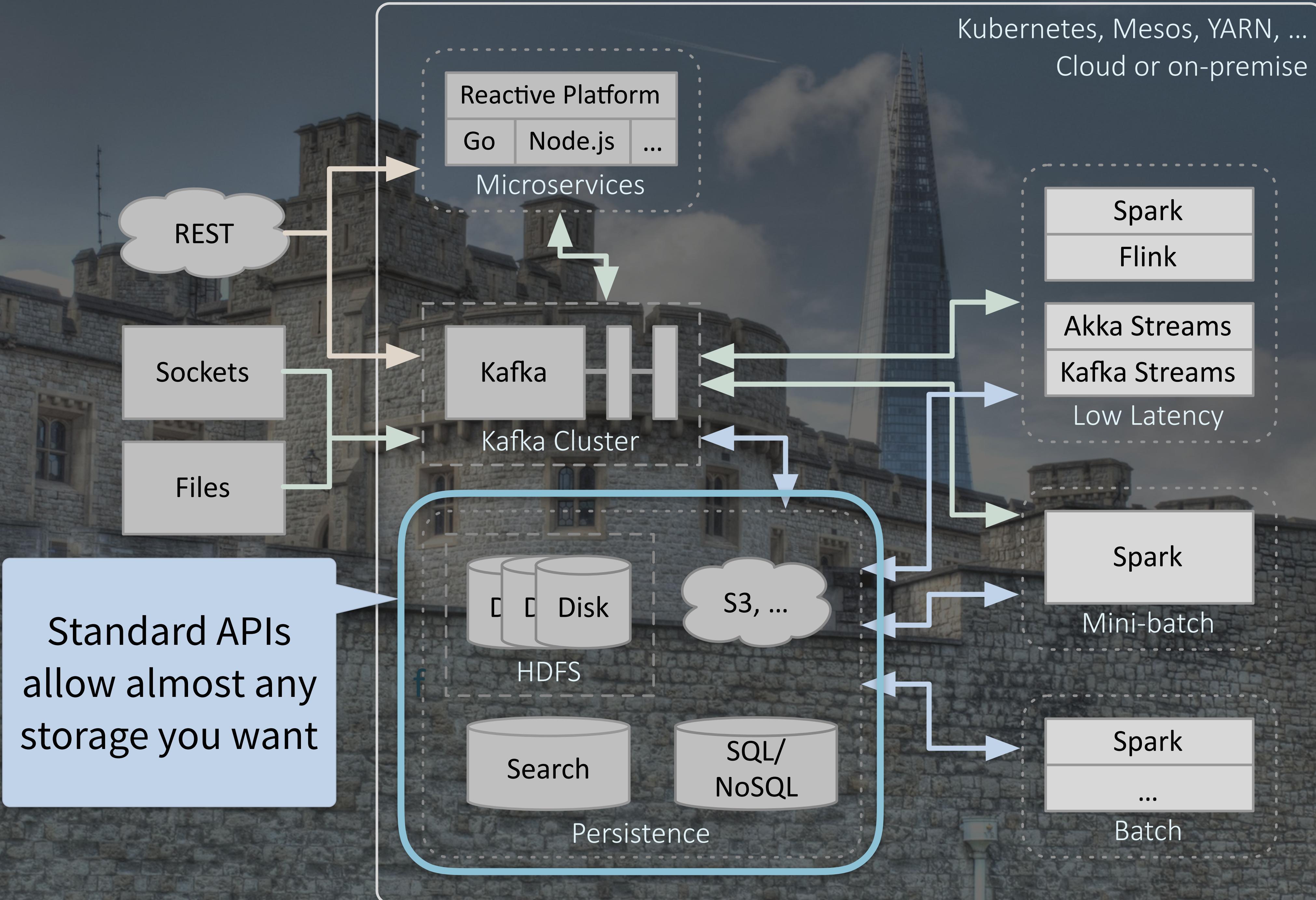
Spark

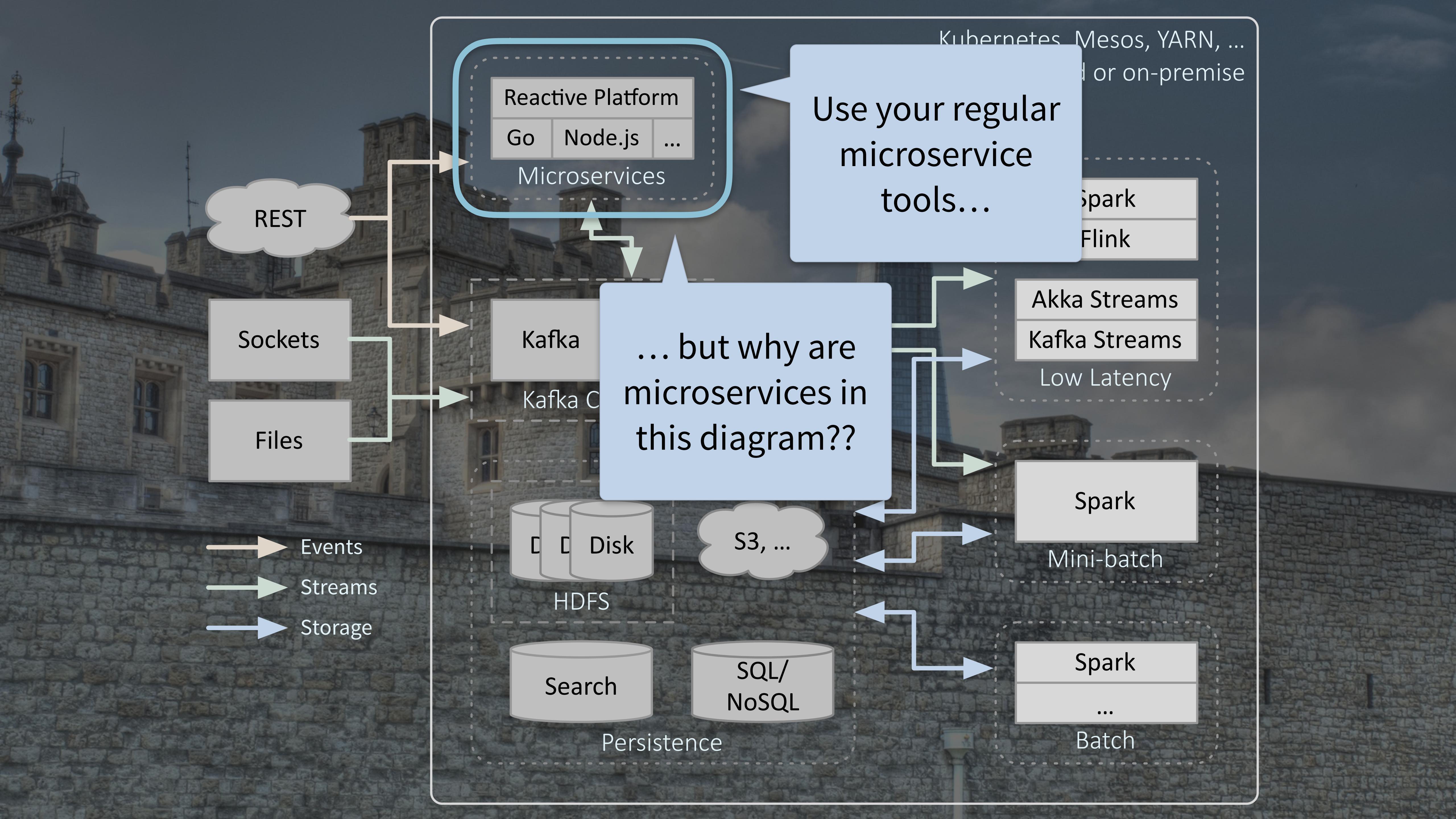
...  
Batch











# Why Microservices in Fast Data?

1. The trend is to run everything in big clusters using Kubernetes or Mesos
  - In the cloud or on-premise

# Why Microservices in Fast Data?

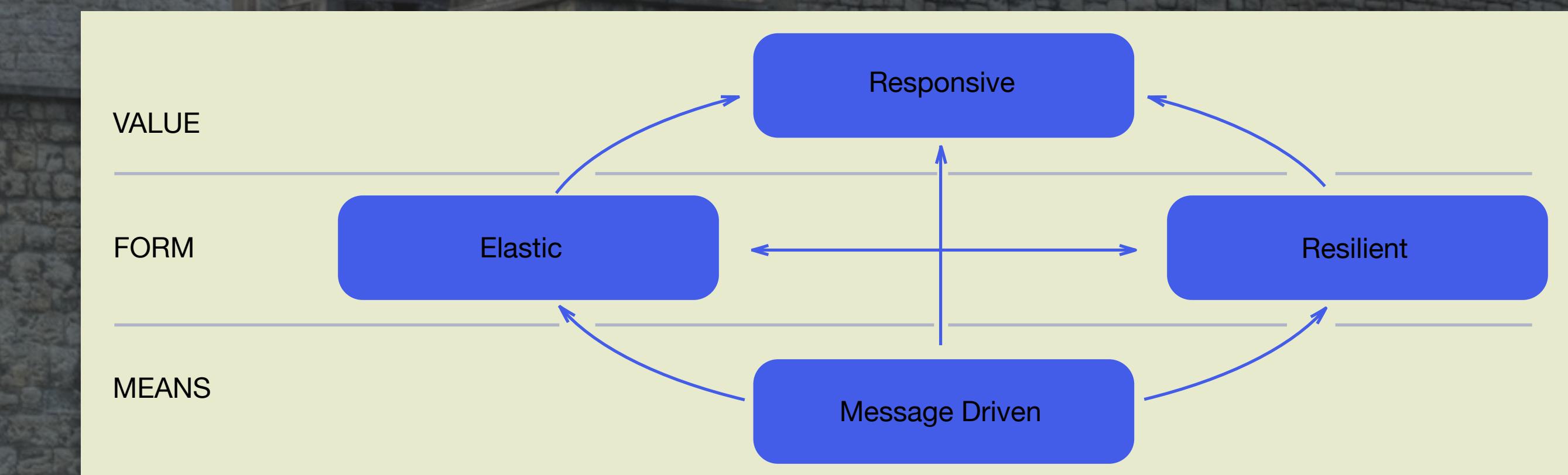
2. If streaming gives you information faster...

- ...you'll want quick access to it in your other services!

# Why Microservices in Fast Data?

## 3. Streaming raises the bar on data services

- Compared to batch services, long-running streaming services must be more:
- Scalable
- Resilient
- Flexible



# Why Microservices in Fast Data?

4. This leads to our last major point...



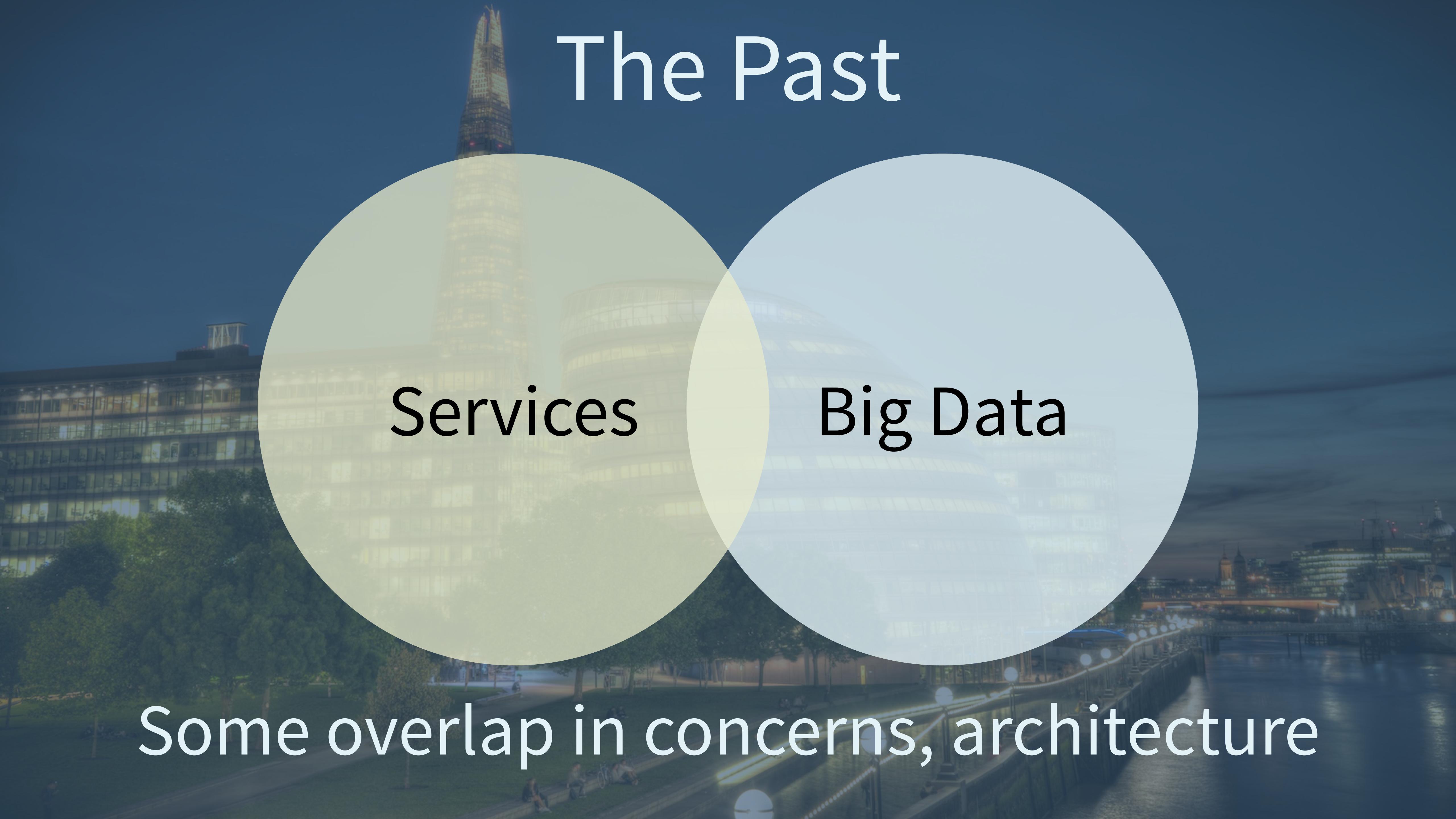
# Organizational Impact



# Organizational Impact

- Data scientists have to understand production issues
- Data engineers have to become good at highly-available microservices
- Microservice engineers have to become good at data

# The Past

A Venn diagram is overlaid on a photograph of a city skyline at dusk. The background shows the Shard skyscraper, the London Eye, and other buildings along the River Thames. The left circle is yellow and contains the word "Services". The right circle is light blue and contains the words "Big Data". The two circles overlap in the center.

Services

Big Data

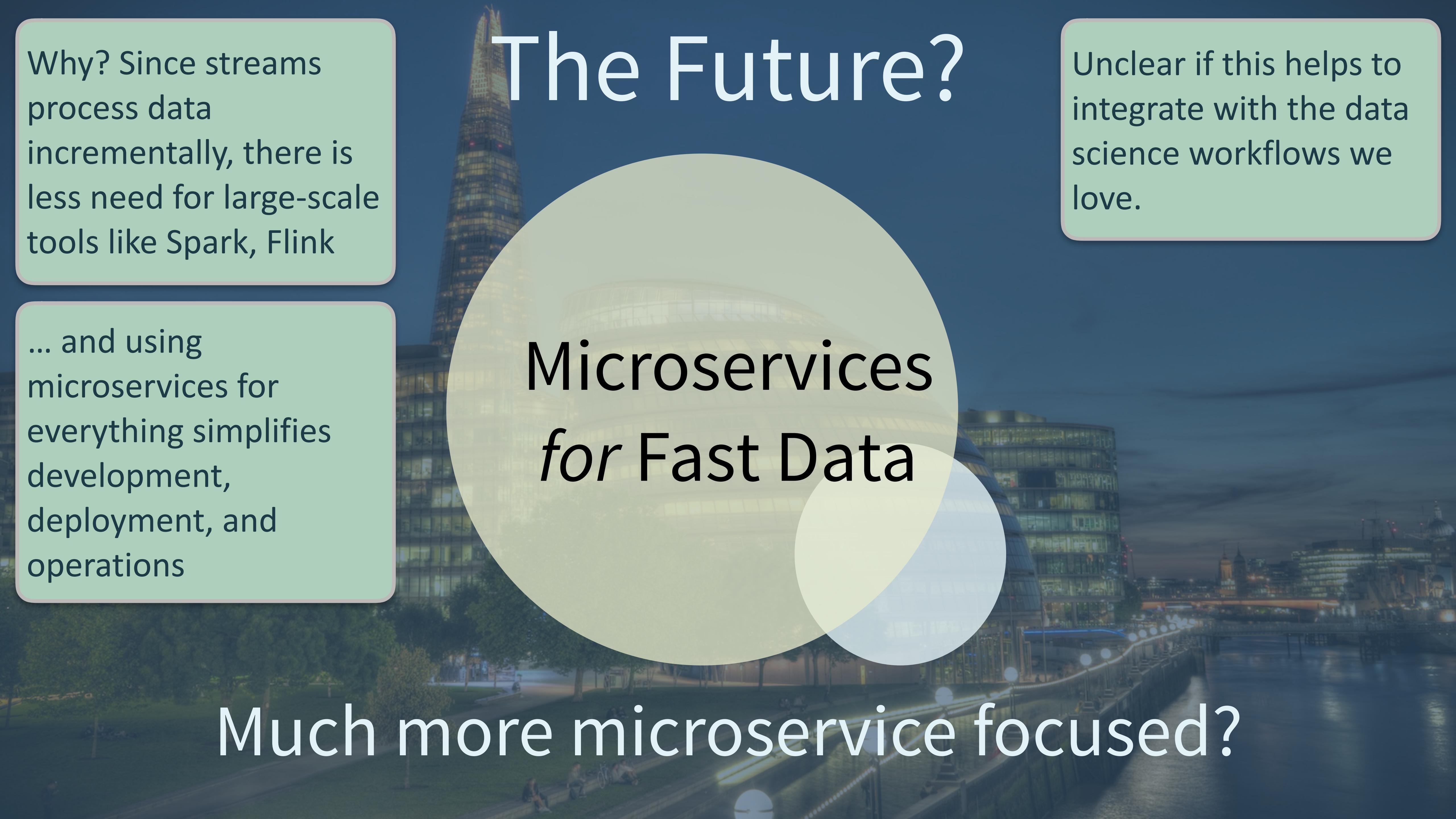
Some overlap in concerns, architecture

# The Present



Microservices  
& Fast Data

Much more overlap



Why? Since streams process data incrementally, there is less need for large-scale tools like Spark, Flink

... and using microservices for everything simplifies development, deployment, and operations

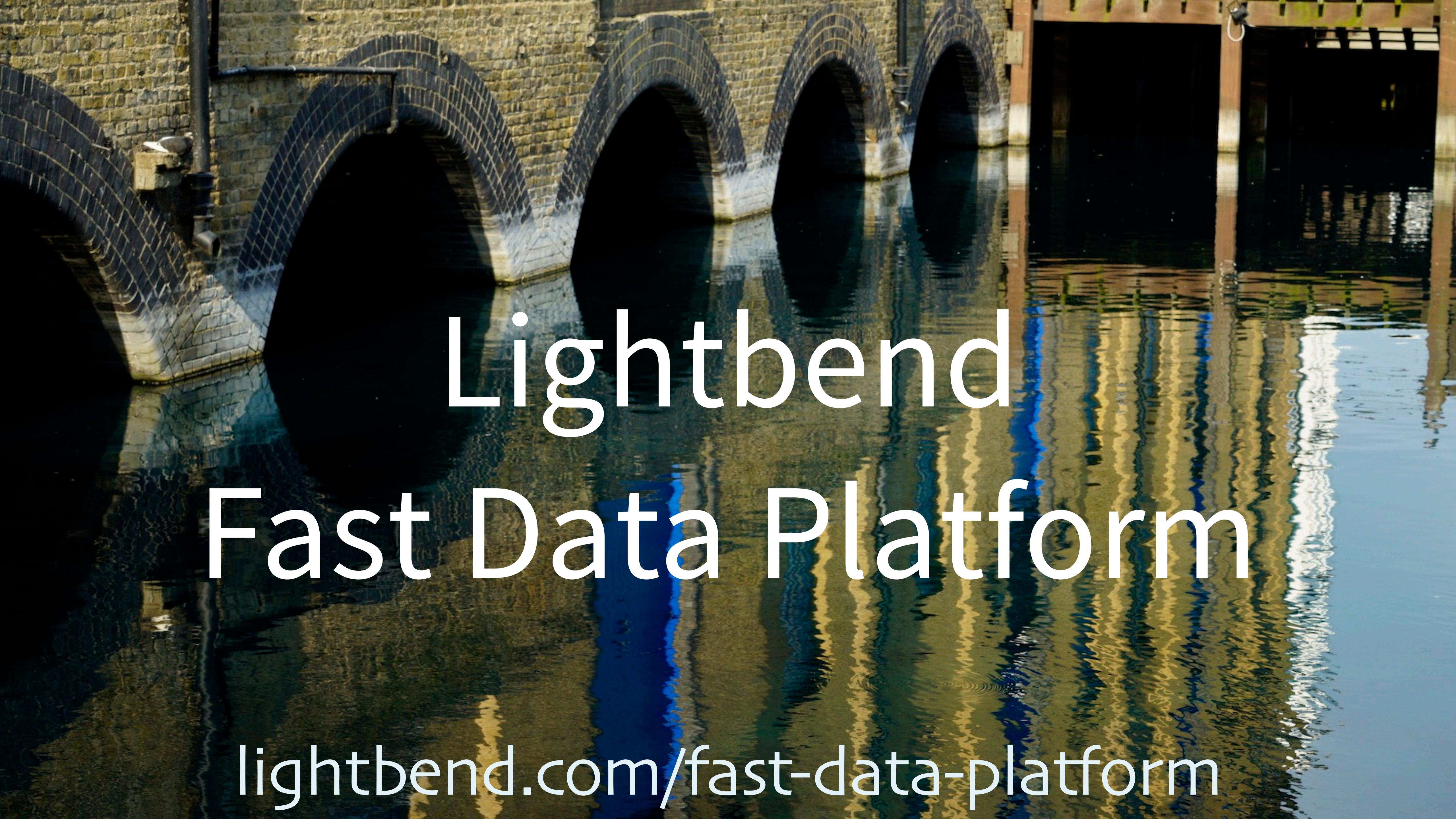
# The Future?



Microservices  
for Fast Data

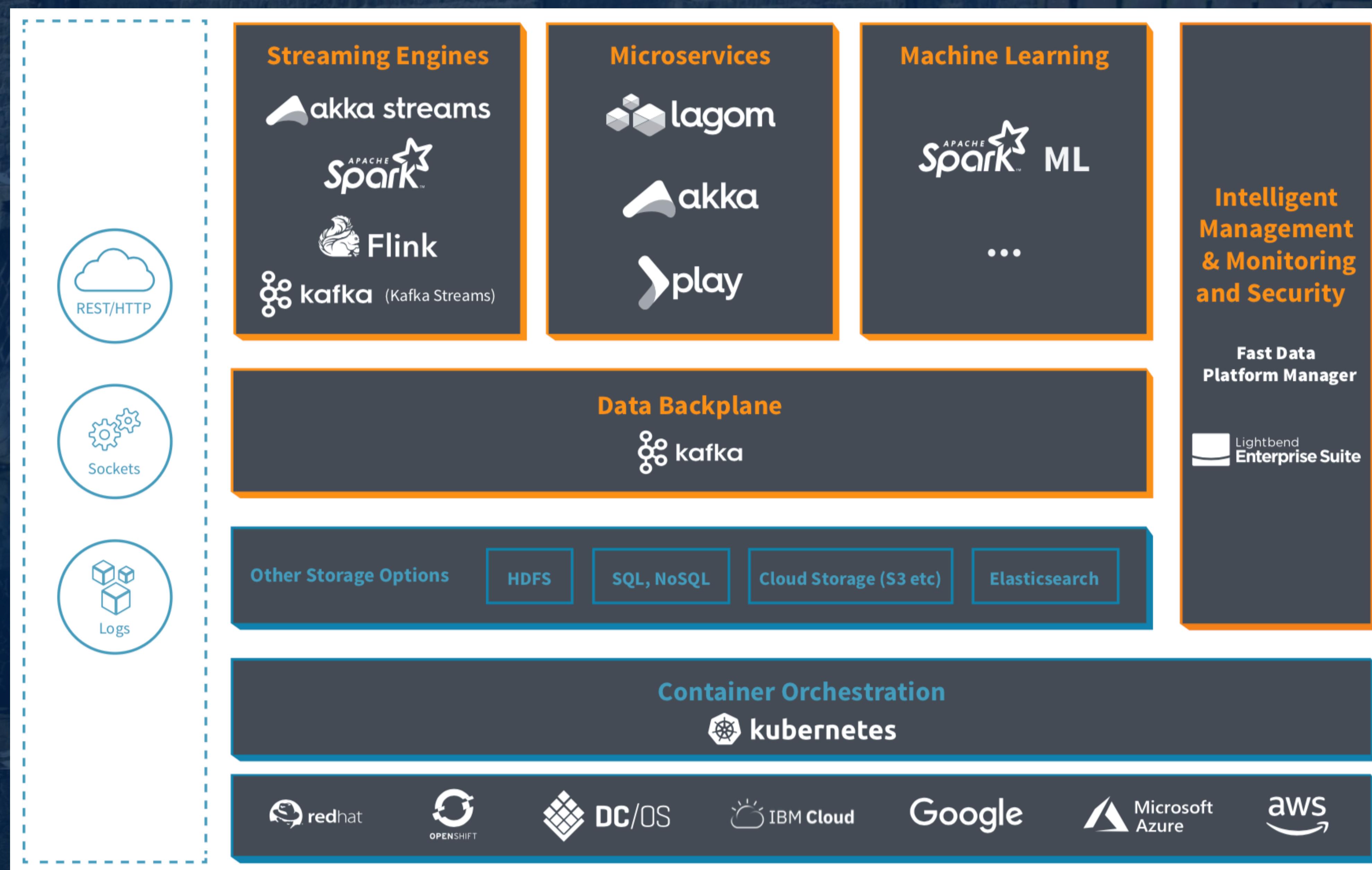
Much more microservice focused?

Unclear if this helps to integrate with the data science workflows we love.

A photograph of a multi-arched stone bridge reflected perfectly in the still water below. The bridge's arches create a rhythmic pattern of light and shadow on the surface. A single bird is perched on a small ledge on the left side of the bridge. The overall scene is peaceful and symmetrical.

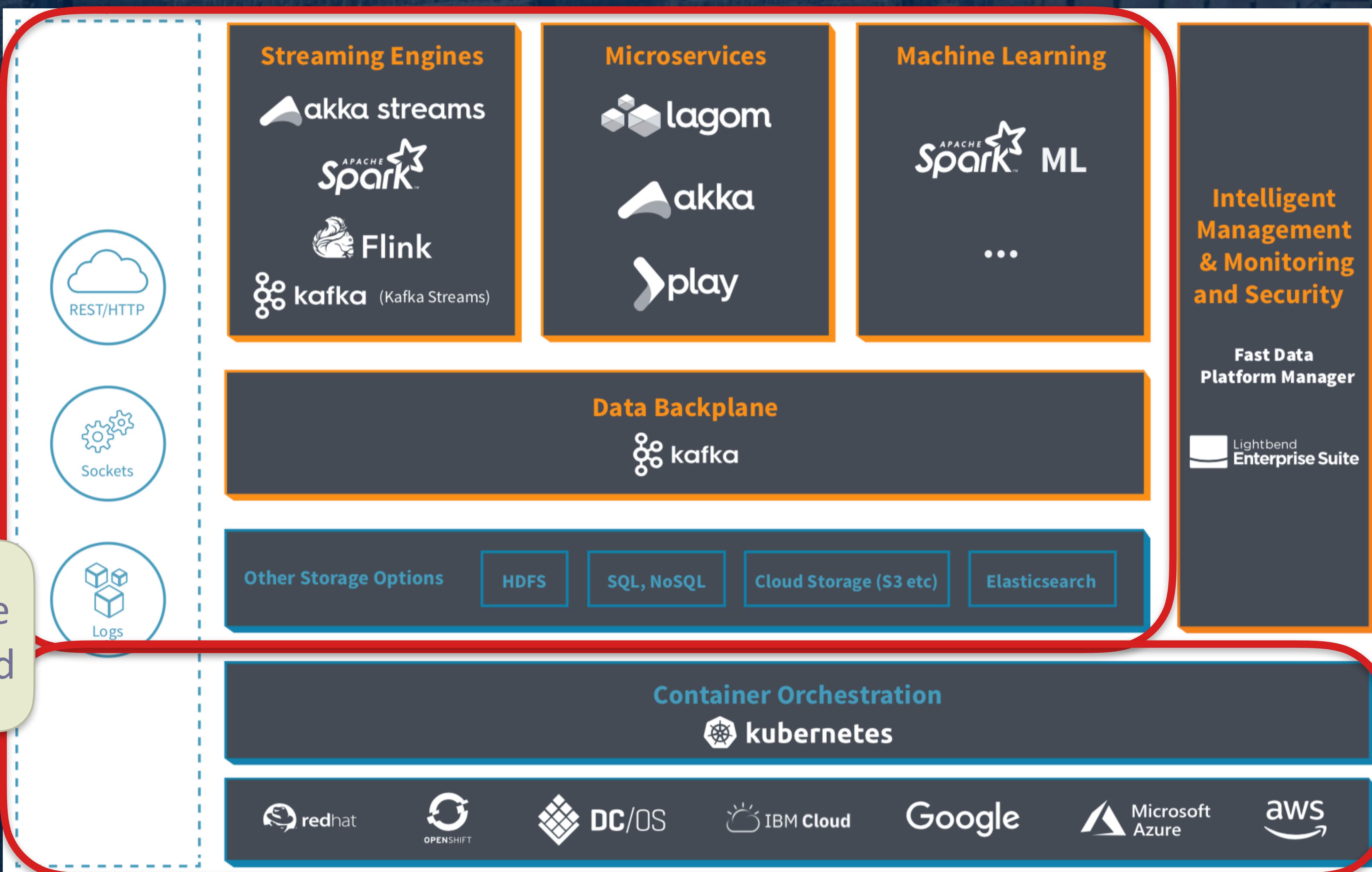
# Lightbend Fast Data Platform

[lightbend.com/fast-data-platform](http://lightbend.com/fast-data-platform)

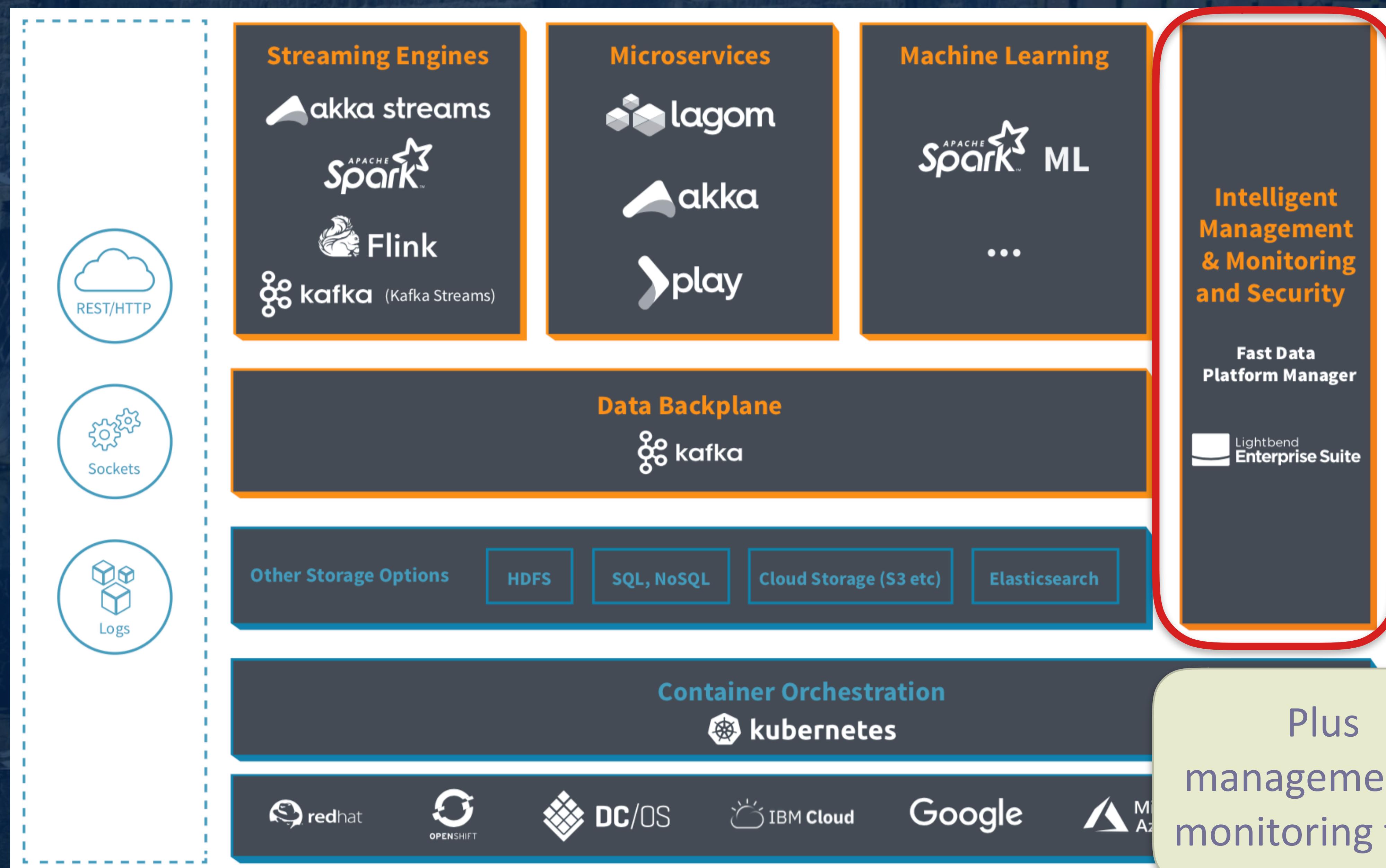


[lightbend.com/fast-data-platform](http://lightbend.com/fast-data-platform)

What we  
discussed

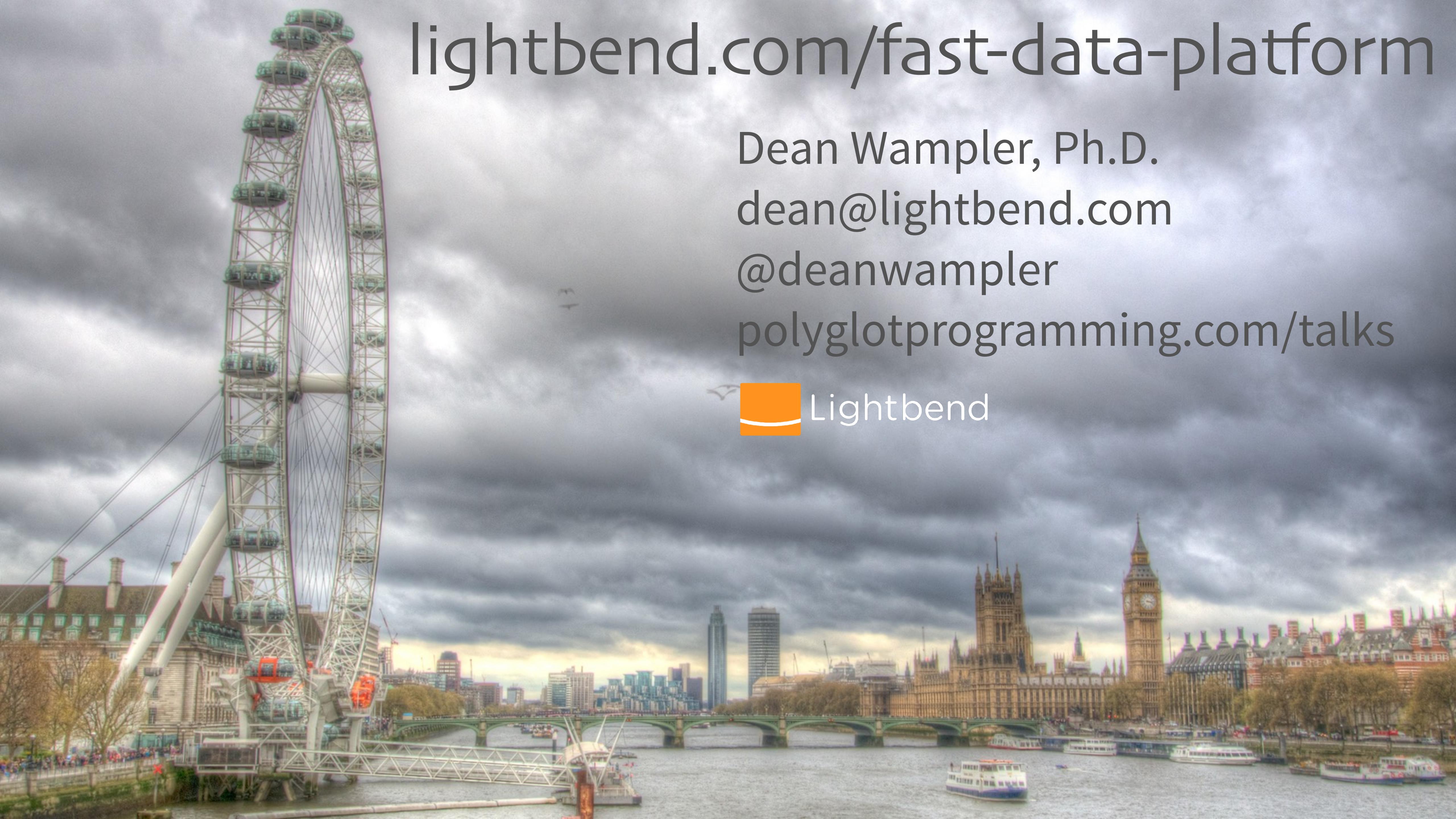


[lightbend.com/fast-data-platform](http://lightbend.com/fast-data-platform)



Plus  
management &  
monitoring tools

[lightbend.com/fast-data-platform](http://lightbend.com/fast-data-platform)

A wide-angle photograph of the London skyline under a dramatic, cloudy sky. On the left, the London Eye Ferris wheel is prominent. In the center-right, the Elizabeth Tower (Big Ben) and the Palace of Westminster are visible. The River Thames flows in the foreground, with several boats and bridges like Westminster Bridge across it.

# [lightbend.com/fast-data-platform](http://lightbend.com/fast-data-platform)

Dean Wampler, Ph.D.  
[dean@lightbend.com](mailto:dean@lightbend.com)  
[@deanwampler](https://twitter.com/deanwampler)  
[polyglotprogramming.com/talks](https://polyglotprogramming.com/talks)

