

# 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)

O'REILLY®

# Fast Data Architectures for Streaming Applications

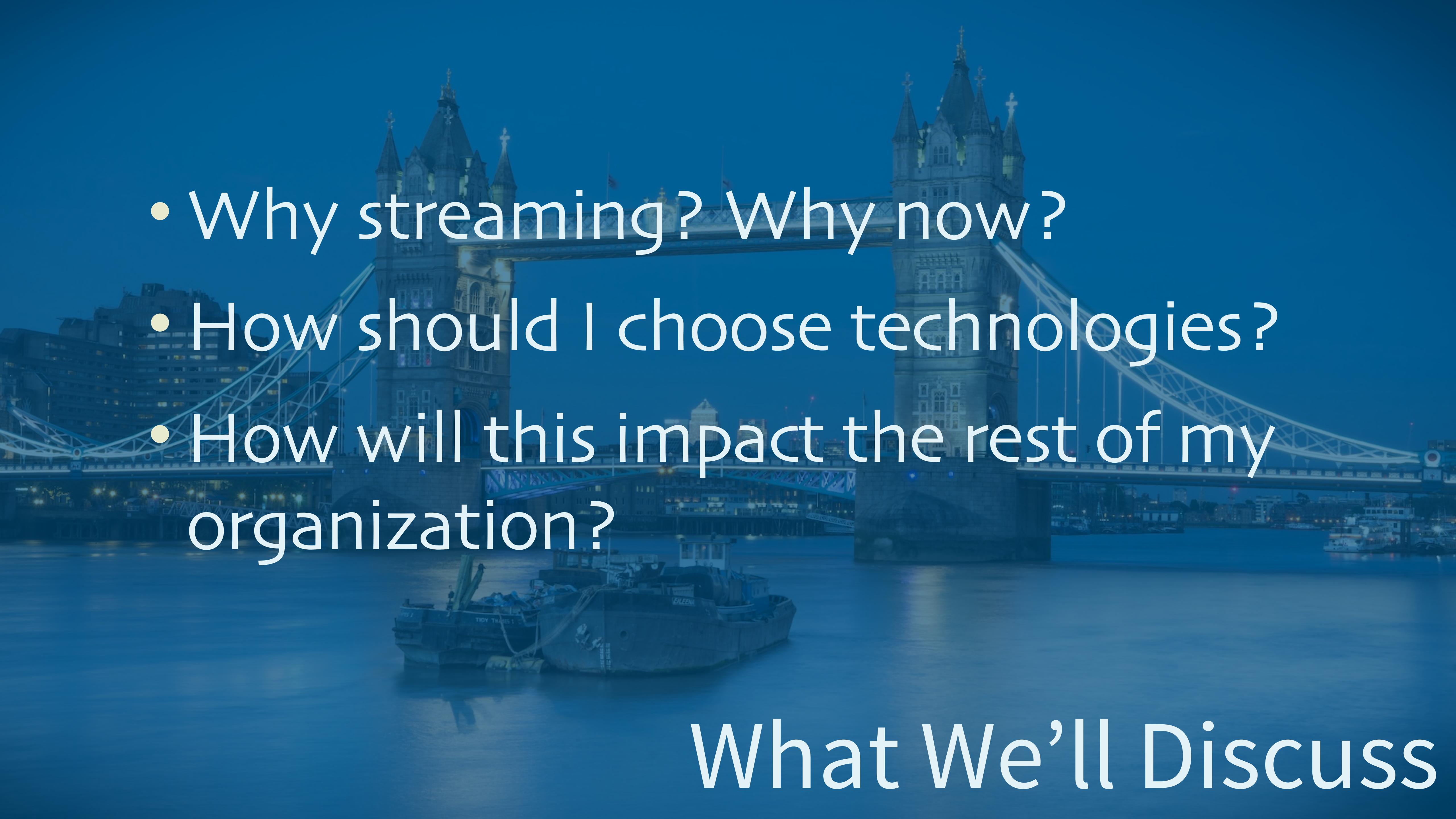
Getting Answers Now from  
Data Sets that Never End

Dean Wampler

Compliments of  
 Lightbend



# What We'll Discuss

- 
- Why streaming? Why now?
  - How should I choose technologies?
  - How will this impact the rest of my organization?

What We'll Discuss



# Why Streaming?

- 
- A large, stylized blue whale sculpture is the central focus of a city fountain. The whale is depicted in a dynamic, leaping pose, with its body curved and its tail pointing upwards. Water is spraying from its blowhole and along its back. In the background, there are classical buildings, including one with a prominent clock tower and another with a sign that reads "CANADIAN PACIFIC". People are visible around the fountain and on the surrounding stone walls.
- New opportunities that require streaming
  - Upgrading batch applications for competitive advantage

# Why Streaming?



# Fast Data Use Cases

## Predictive Analytics

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



## 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.



## Real-time Financial Processes

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

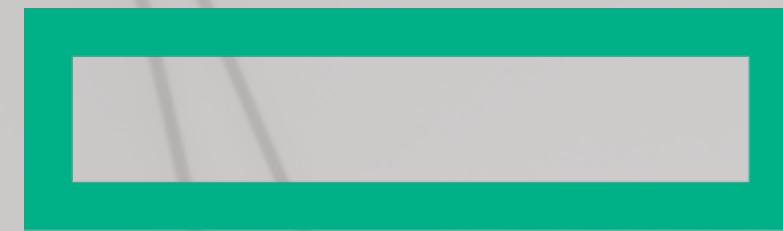


## Legacy Modernization

Accelerate decision making processes and optimize infrastructure costs by moving from batch to streaming



More at: <https://www.lightbend.com/customers>



# Predictive Analytics

## Hewlett Packard Enterprise

- ML models applied to device telemetry to detect anomalies
- Preemptive maintenance prevents user impacts from potential failures

# Core Idea

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

## Anomaly Handler

Corrective Actions

Probable Anomalies

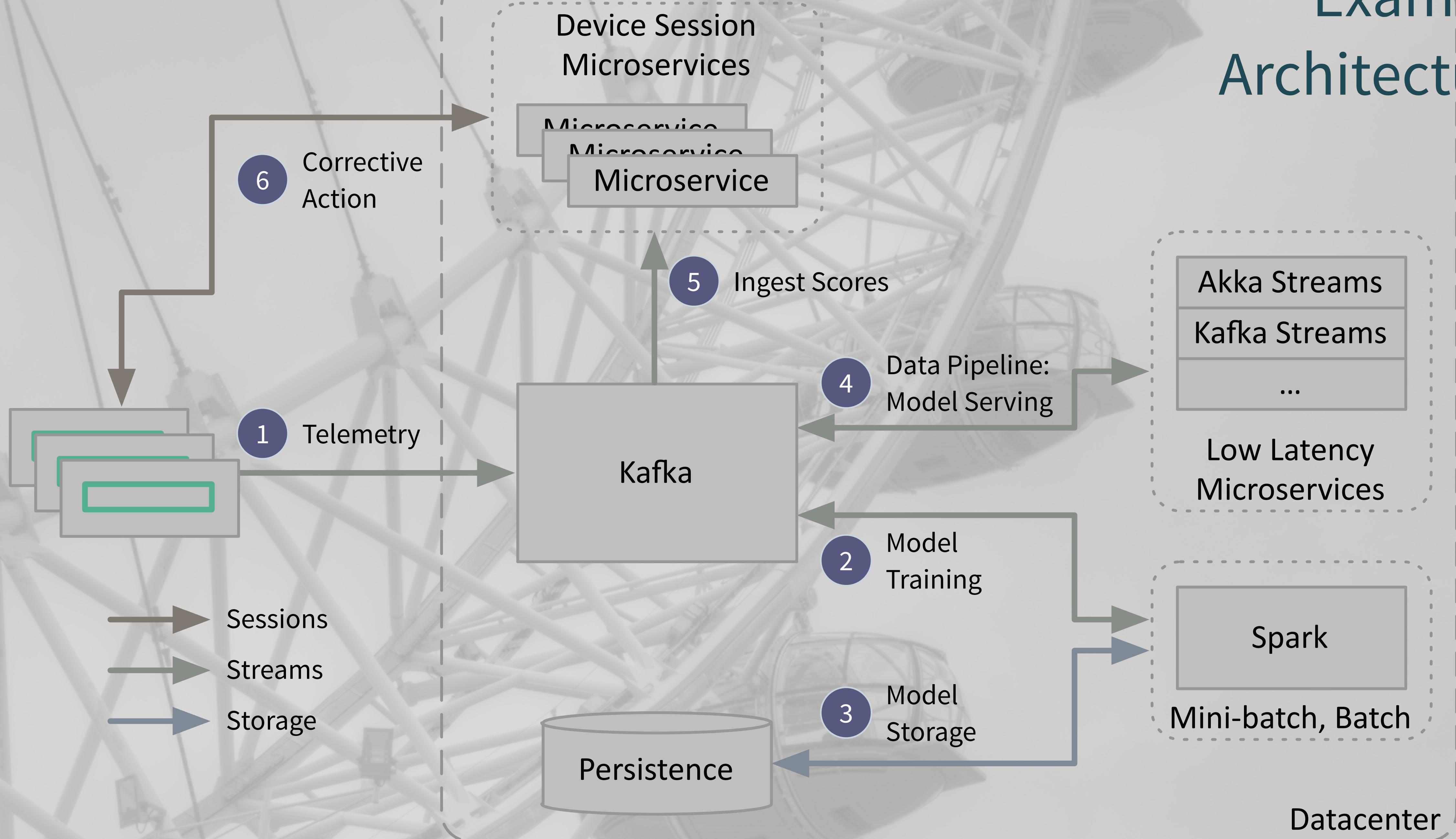
## Anomaly Detection: Model

Ingest telemetry from edge devices.

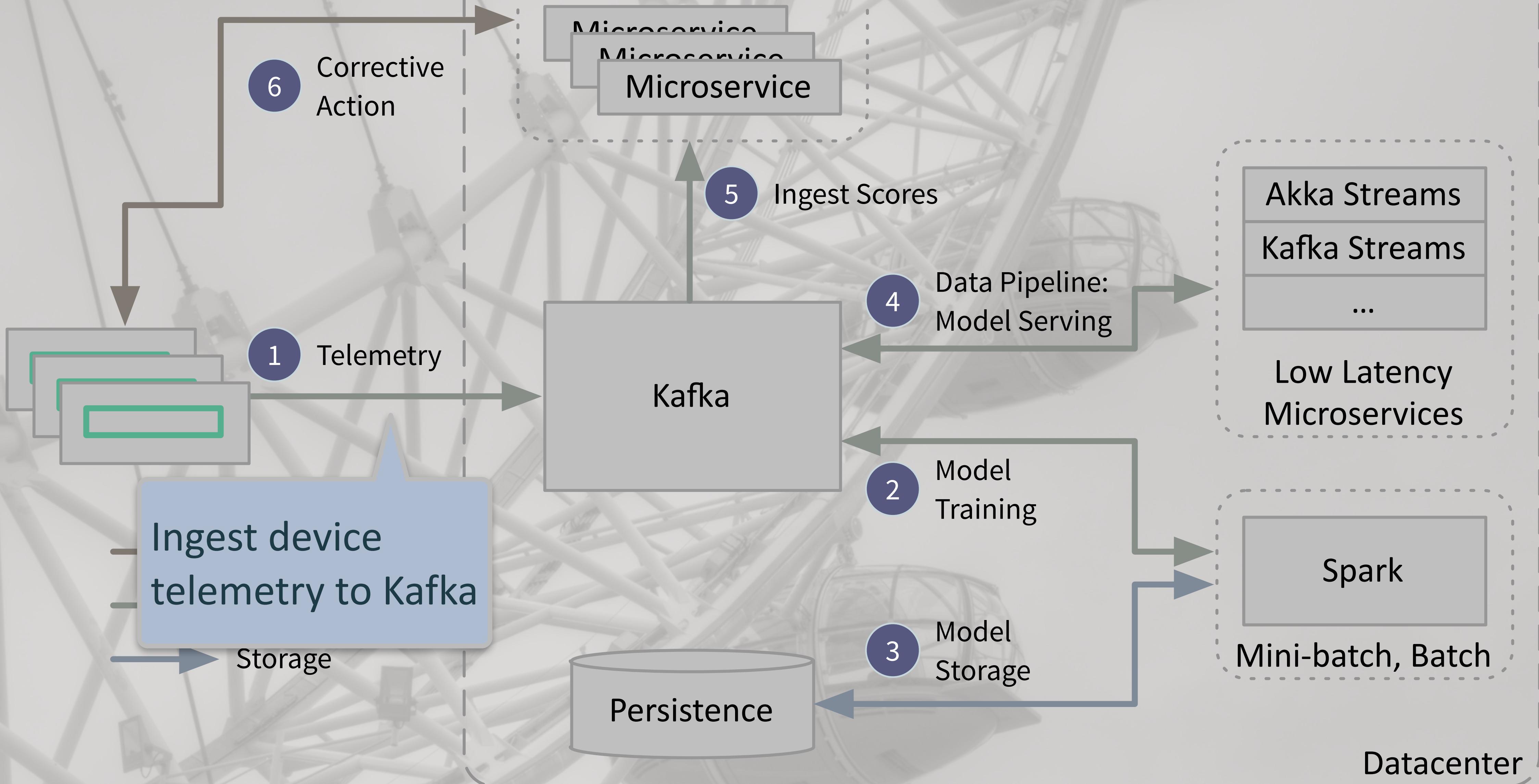
Nimble Storage



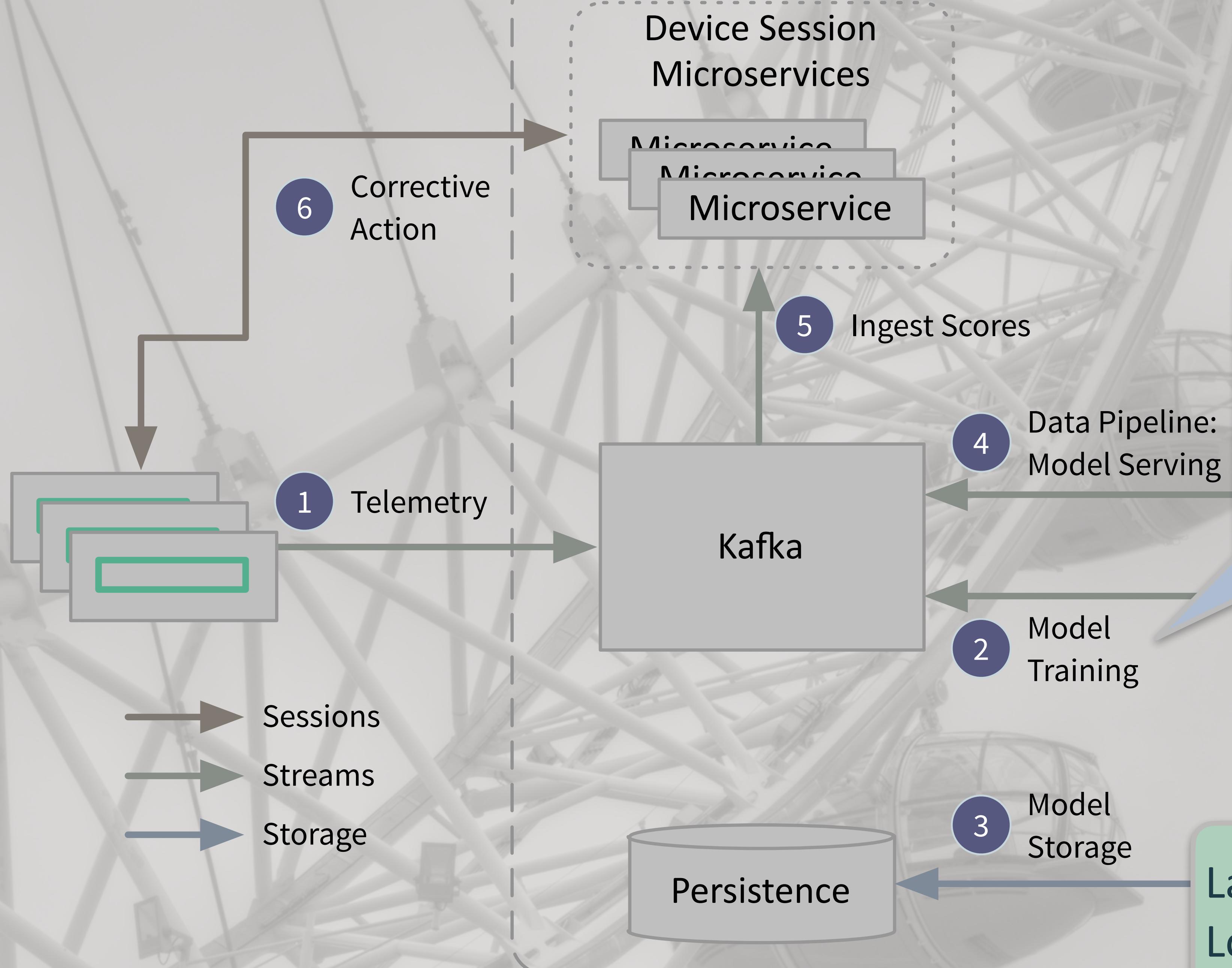
# Example Architecture



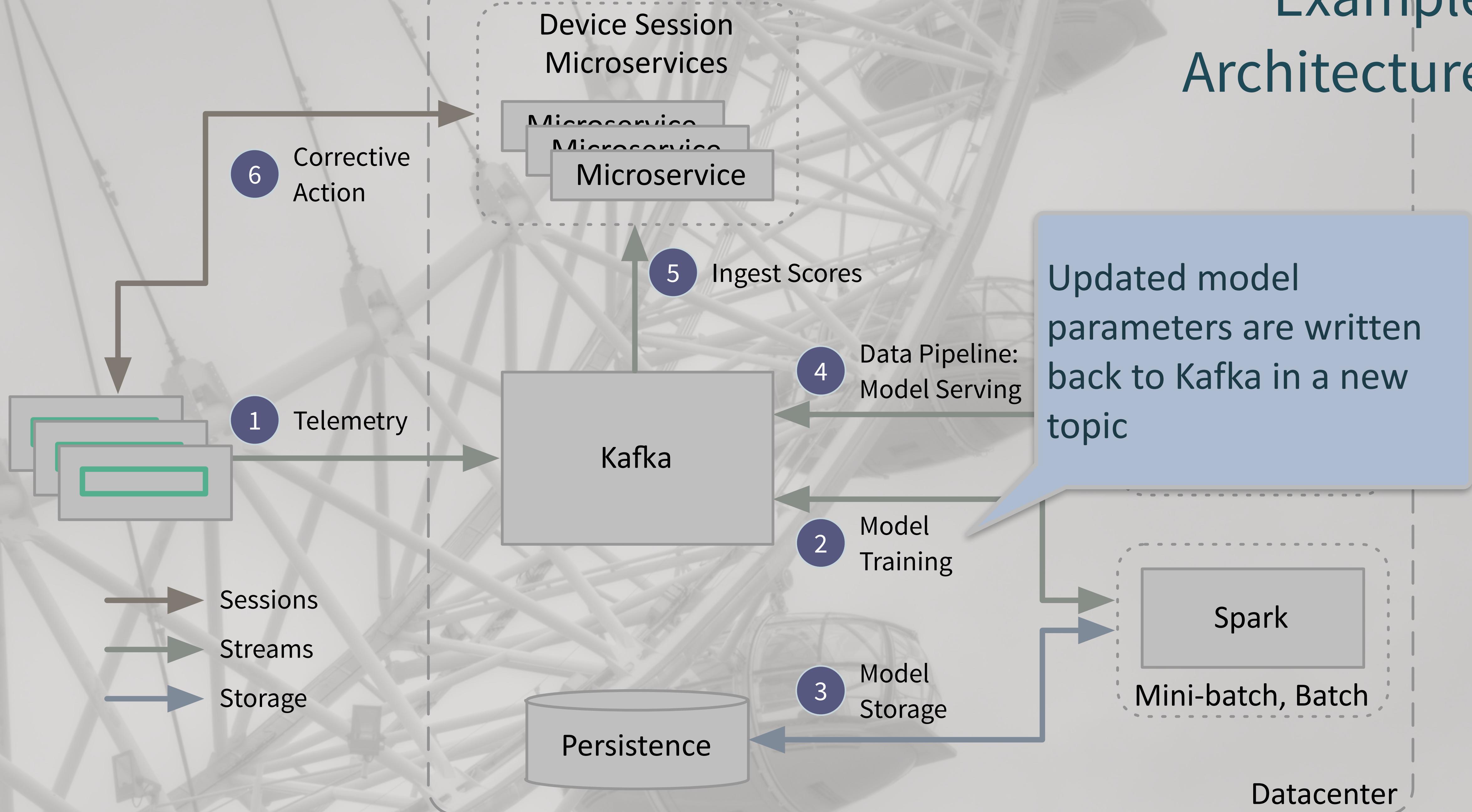
# Example Architecture



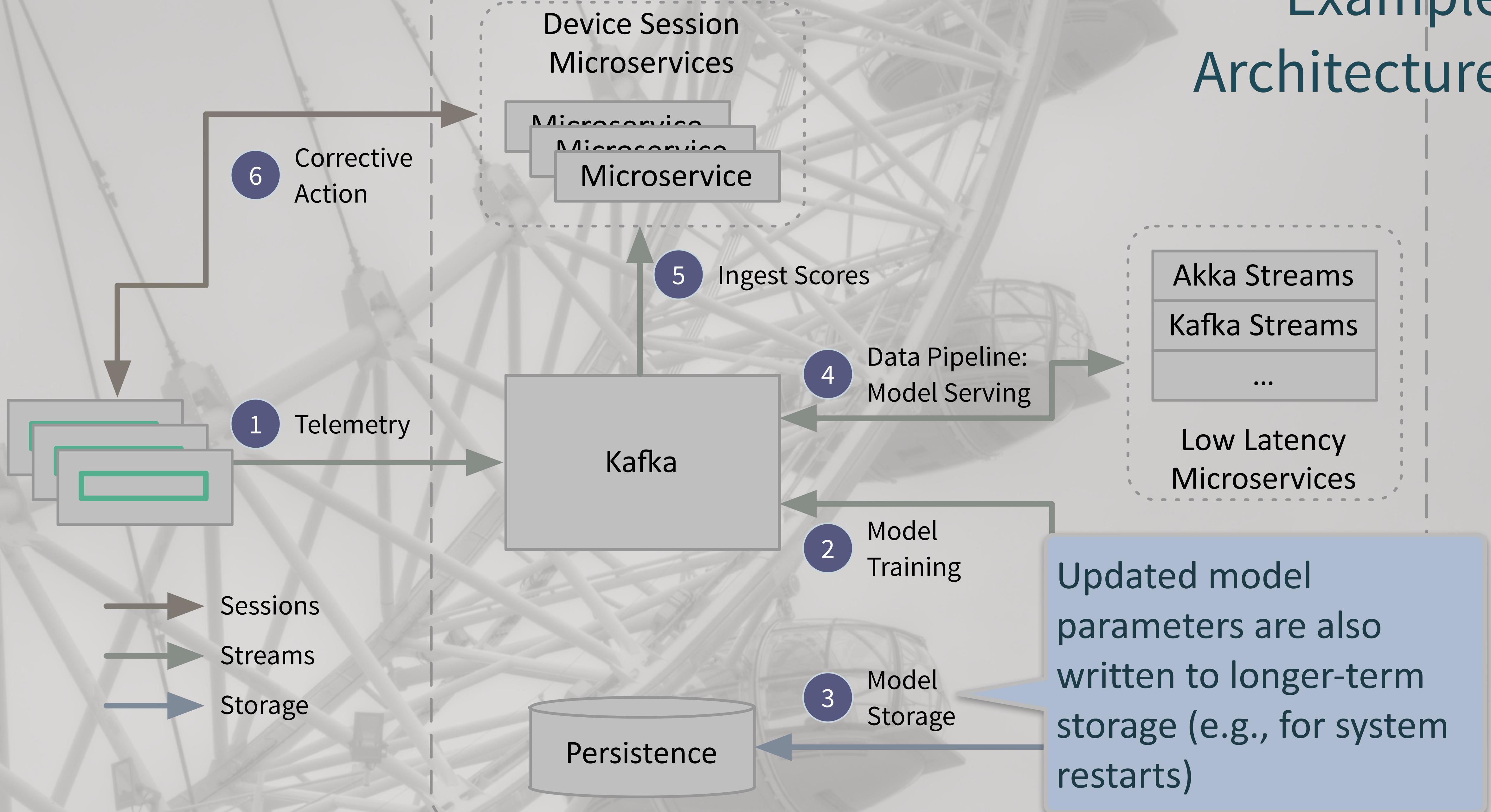
# Example Architecture



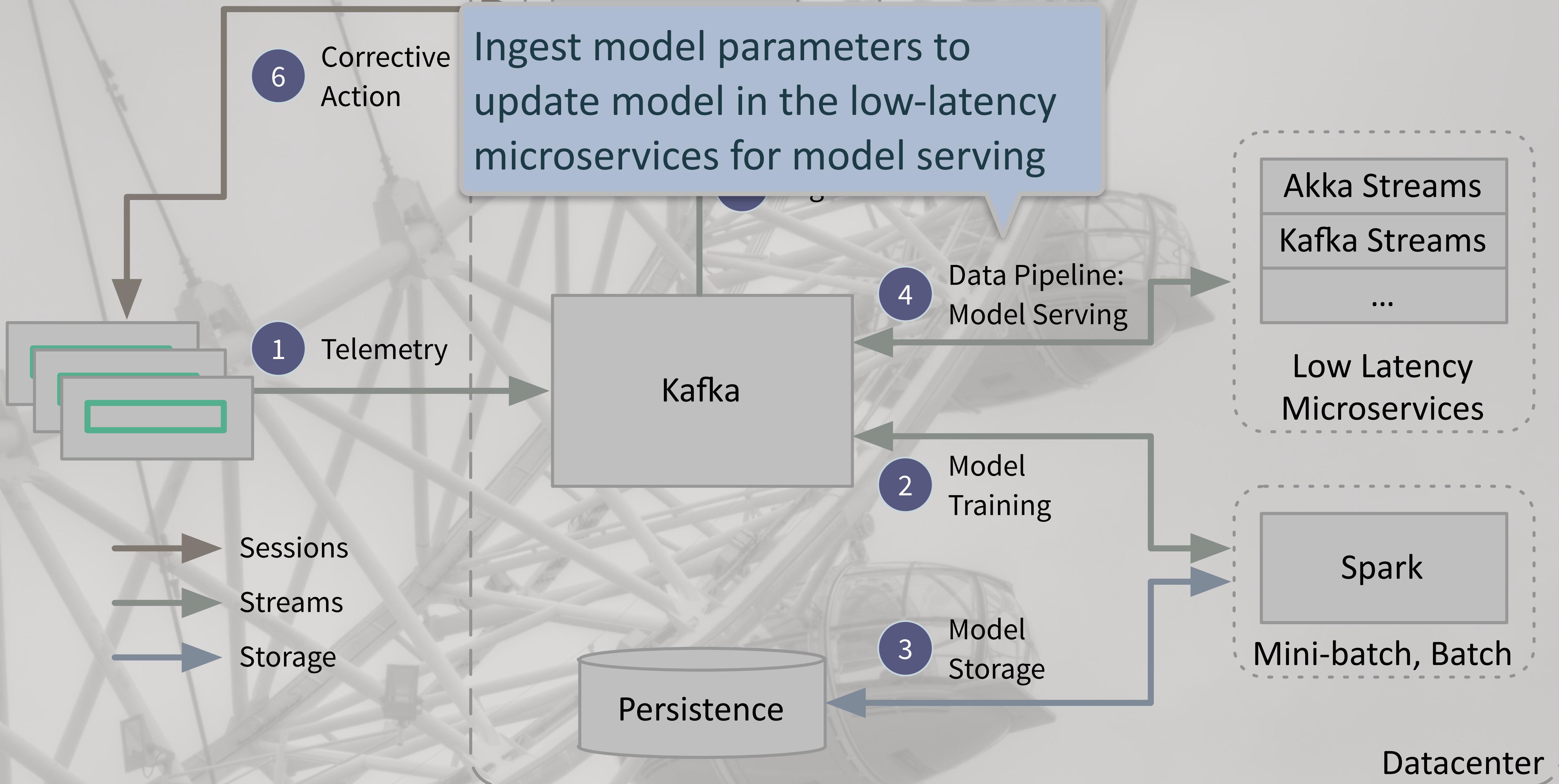
# Example Architecture



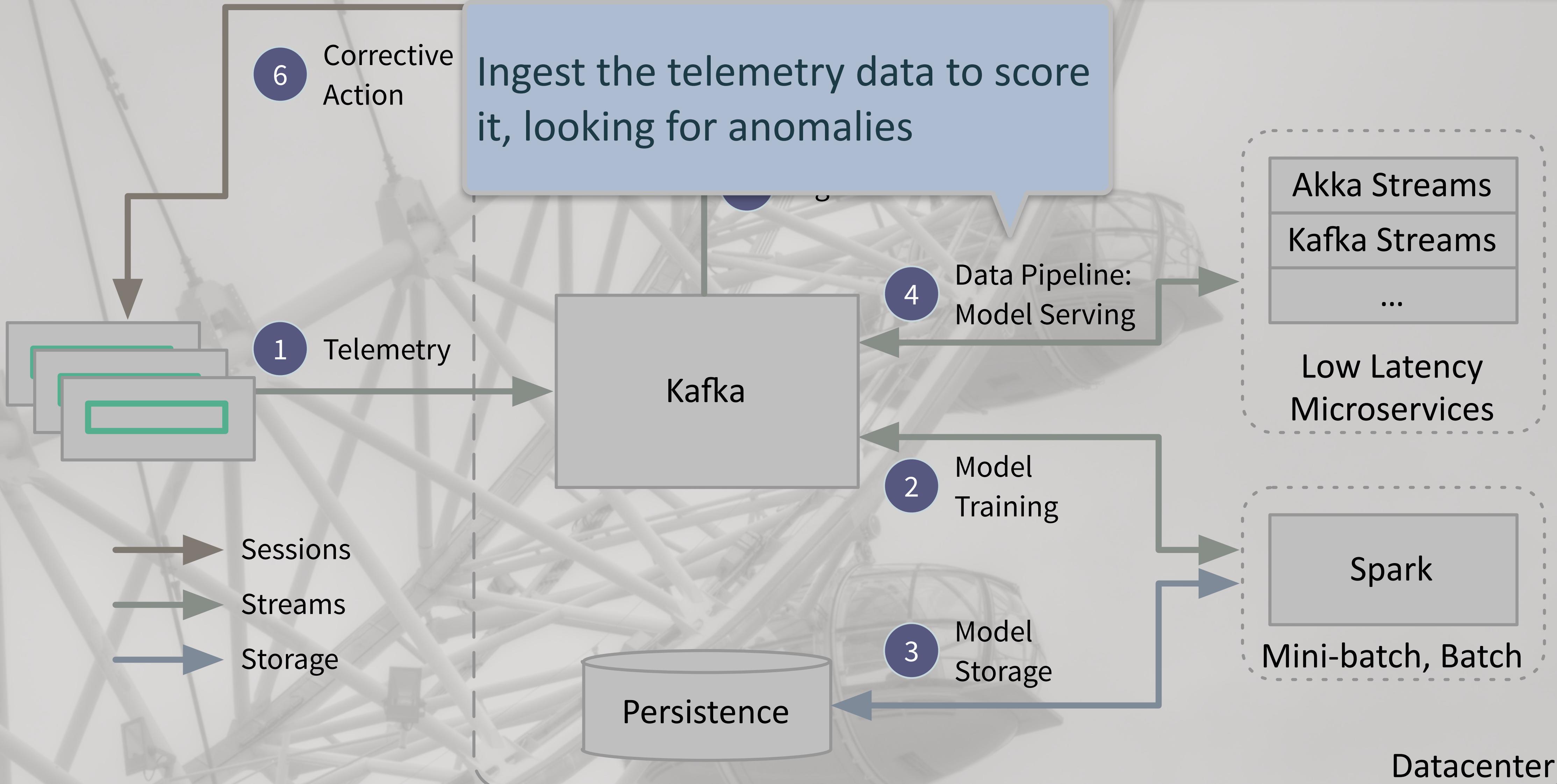
# Example Architecture



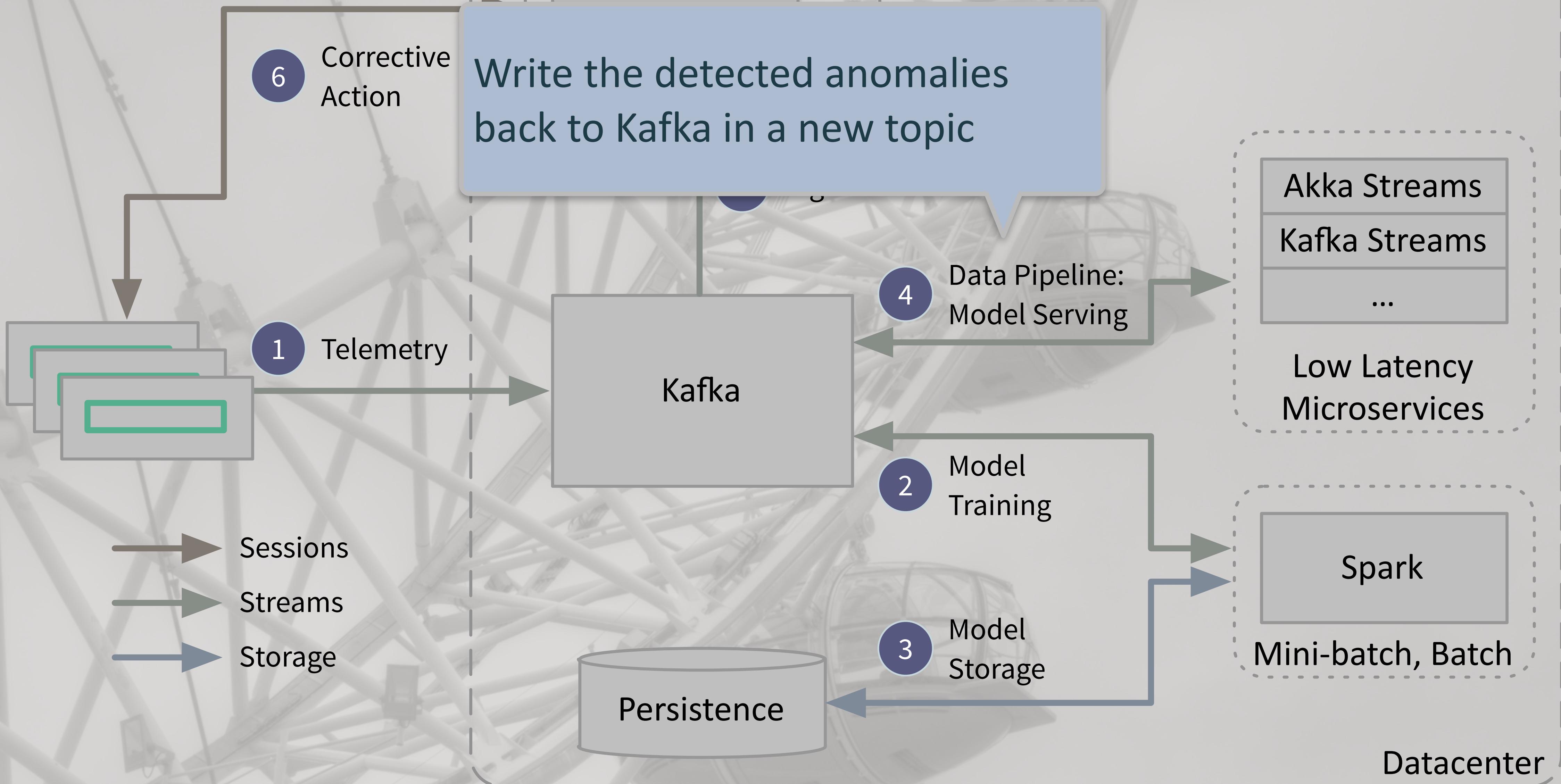
# Example Architecture



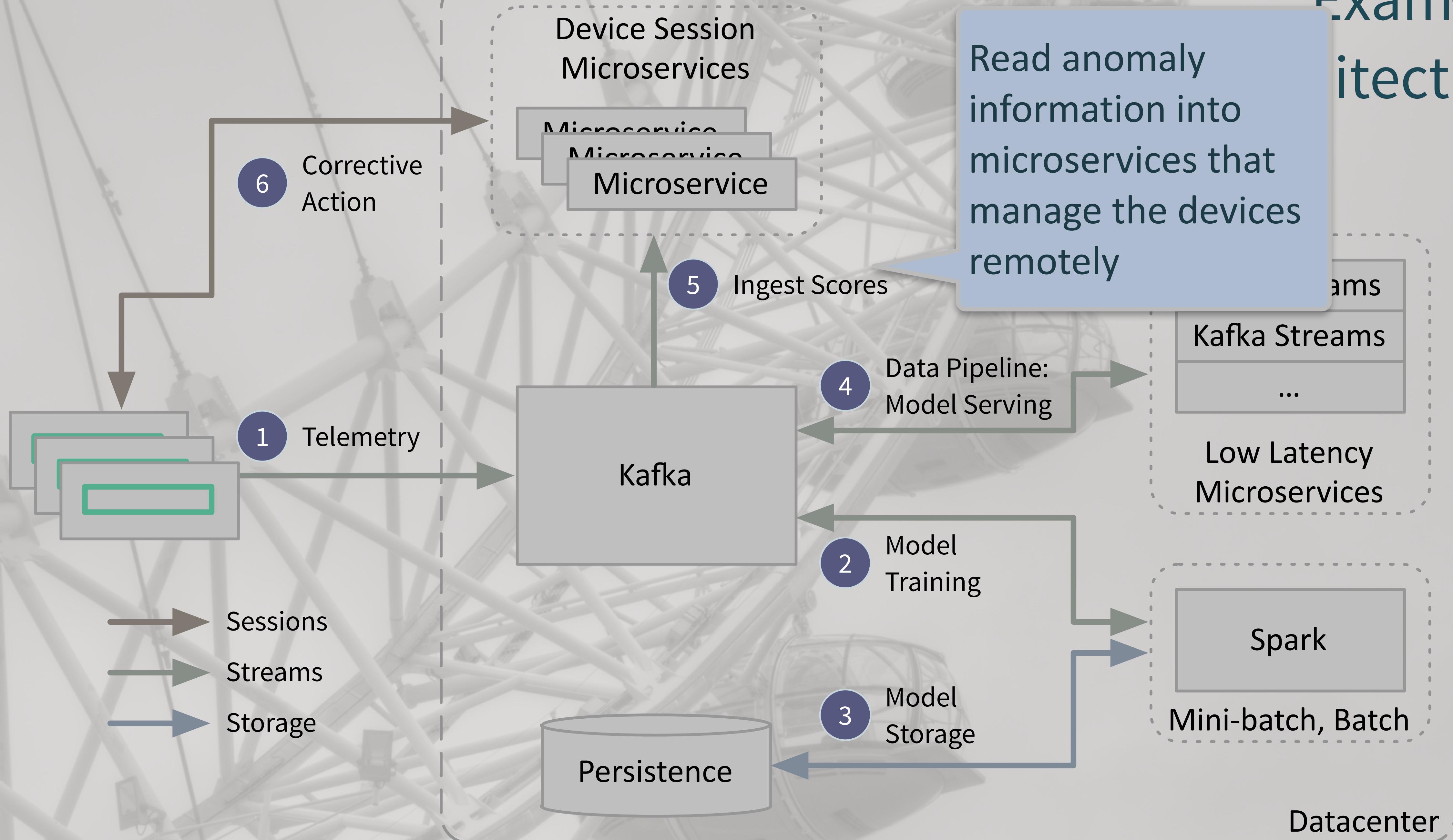
Small data volume,  
Low latency (milliseconds-...)



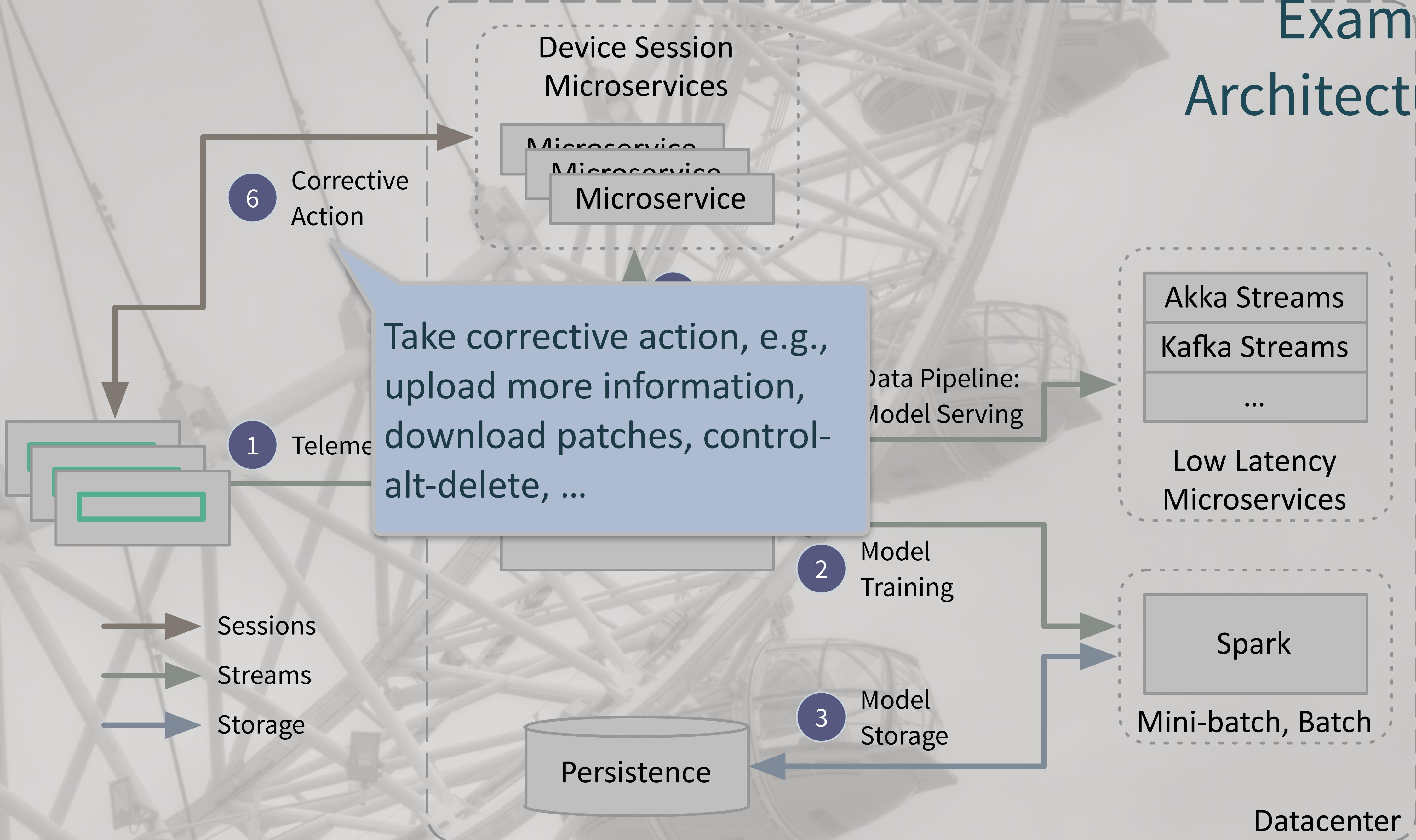
# Example Architecture



# Example architecture

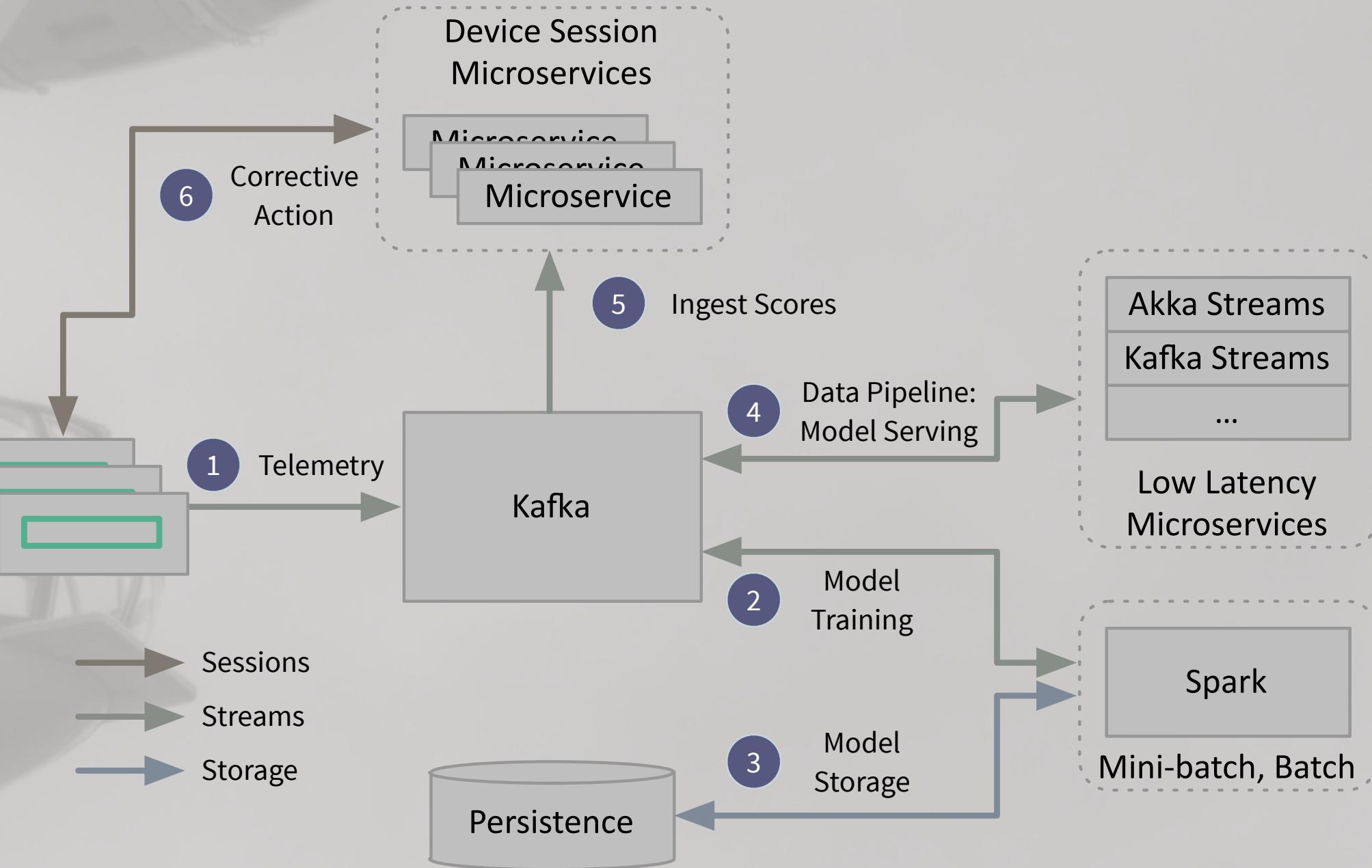


# Example Architecture



# Challenges

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

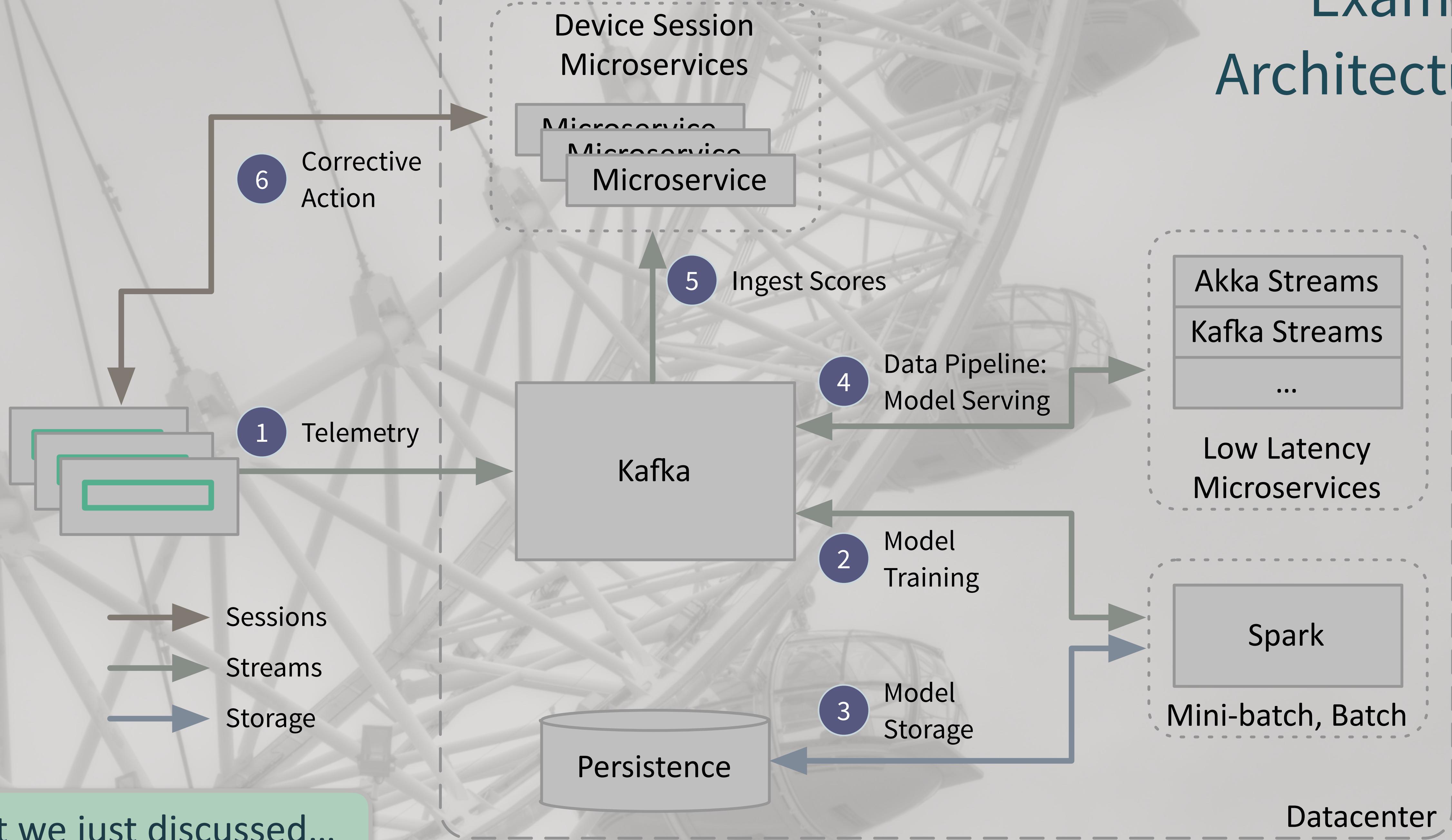




# Internet of Things

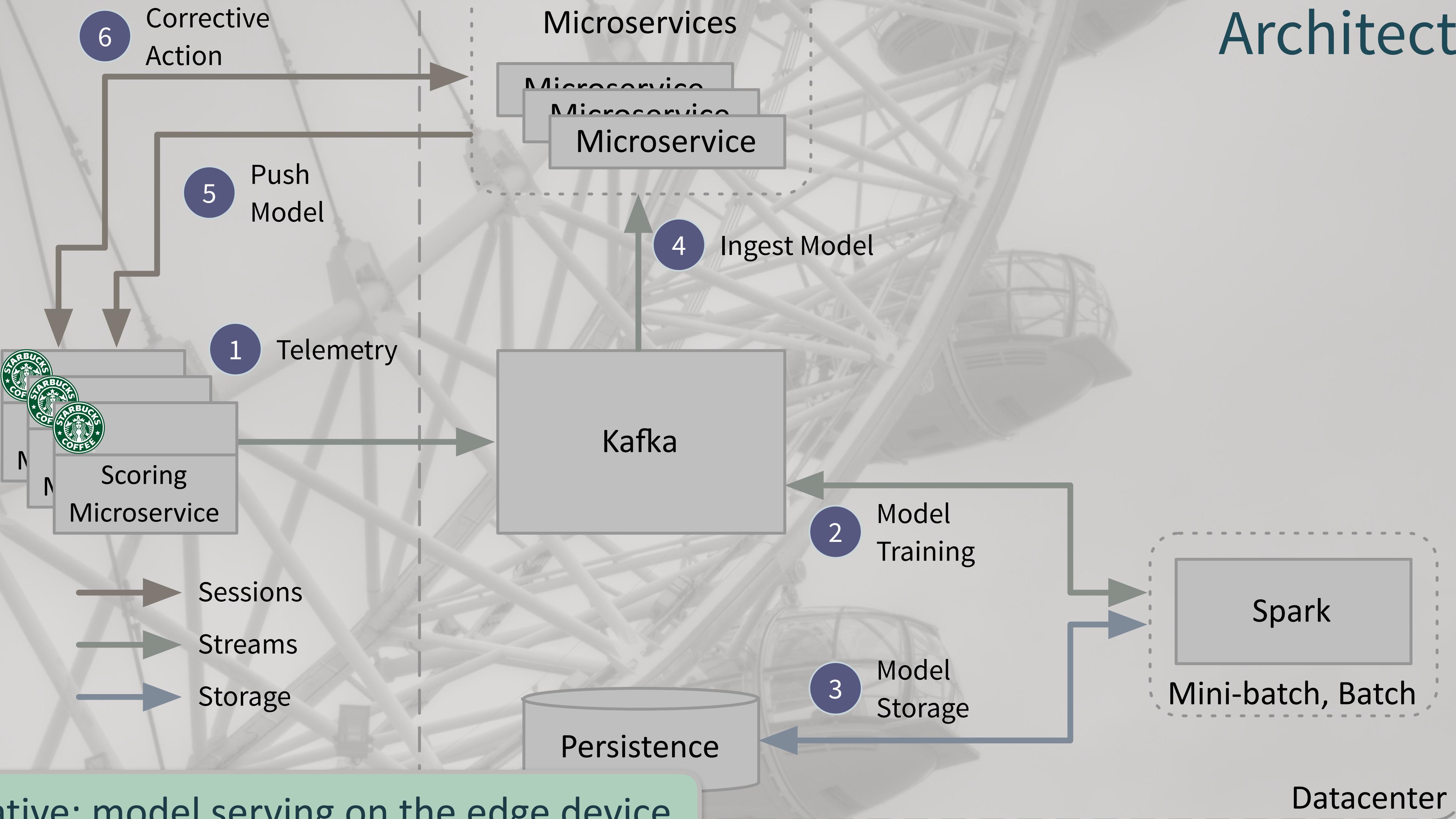
- Real-time consumer and industrial device and supply chain management at scale

# Example Architecture

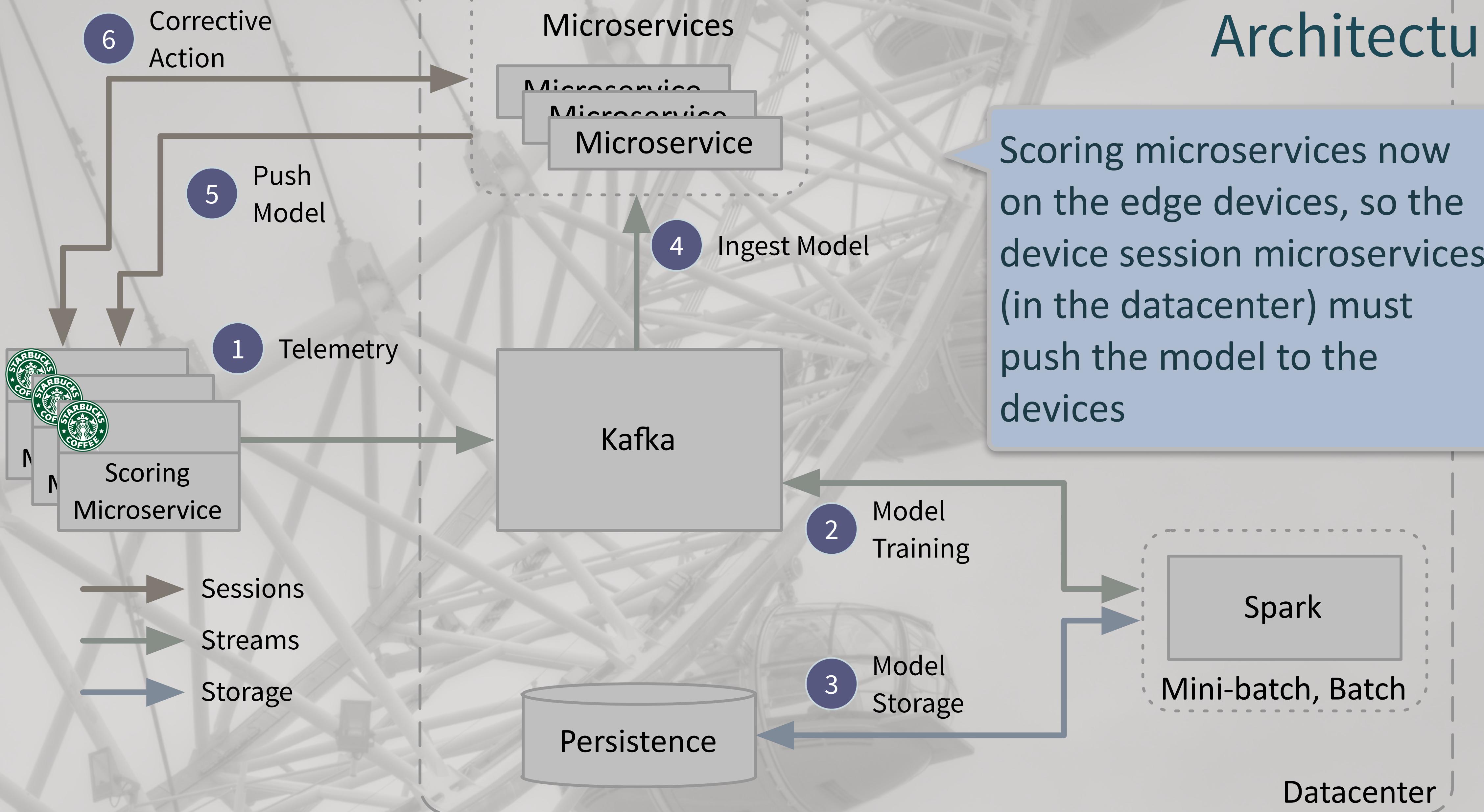


What we just discussed...

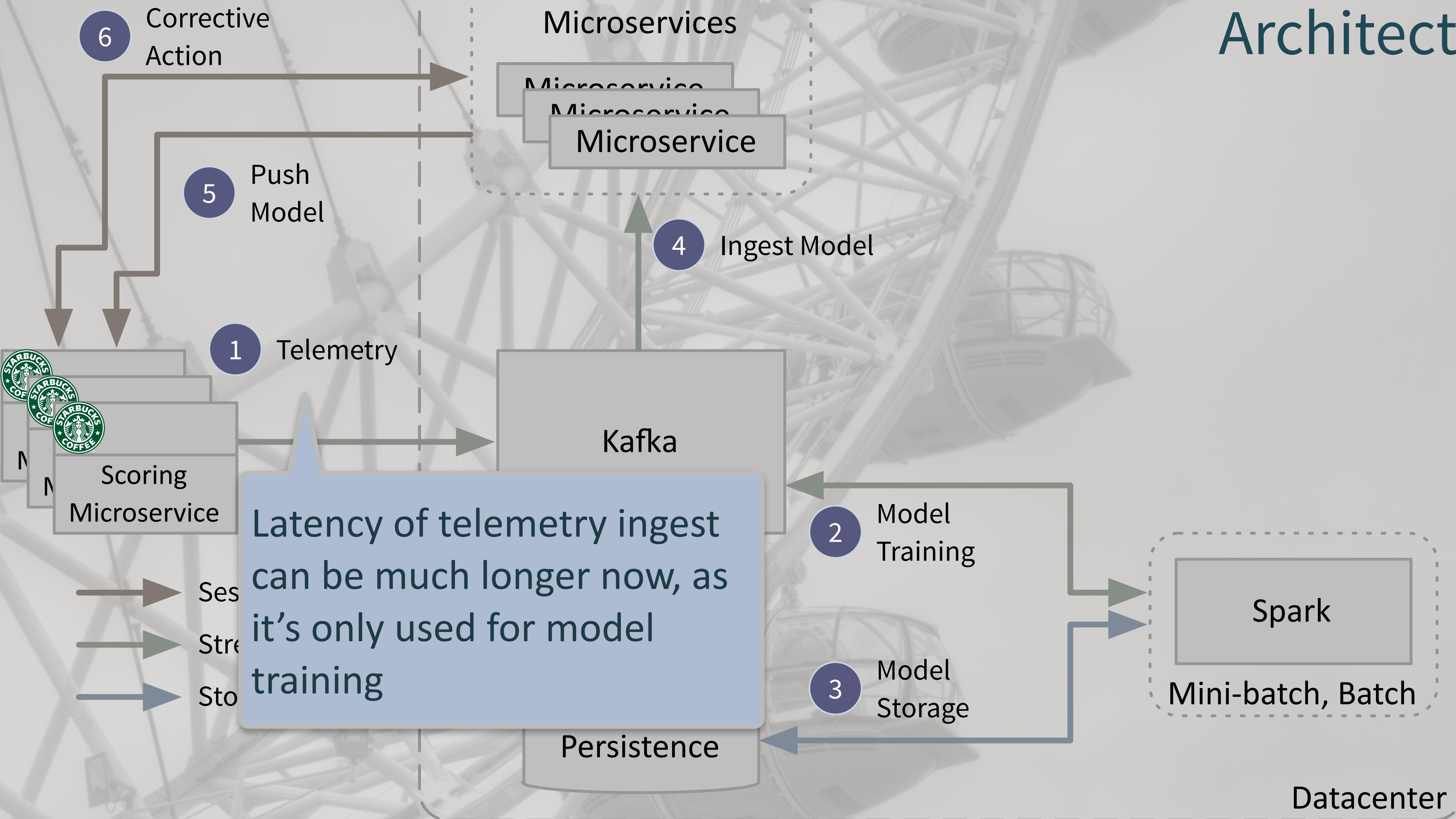
# Edge Scoring Example Architecture



# Edge Scoring Example Architecture



# Edge Scoring Example Architecture



# Real-time Personalization



- Model users to provide real-time marketing based on behavior, location, inventory levels, product promotions, etc.

A few more examples. Similar architectures...

# Real-time Financial



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

# Legacy Modernization



- Accelerate decision making processes and optimize infrastructure costs by moving from batch to streaming
- Hadoop replacement



# Technology Choices

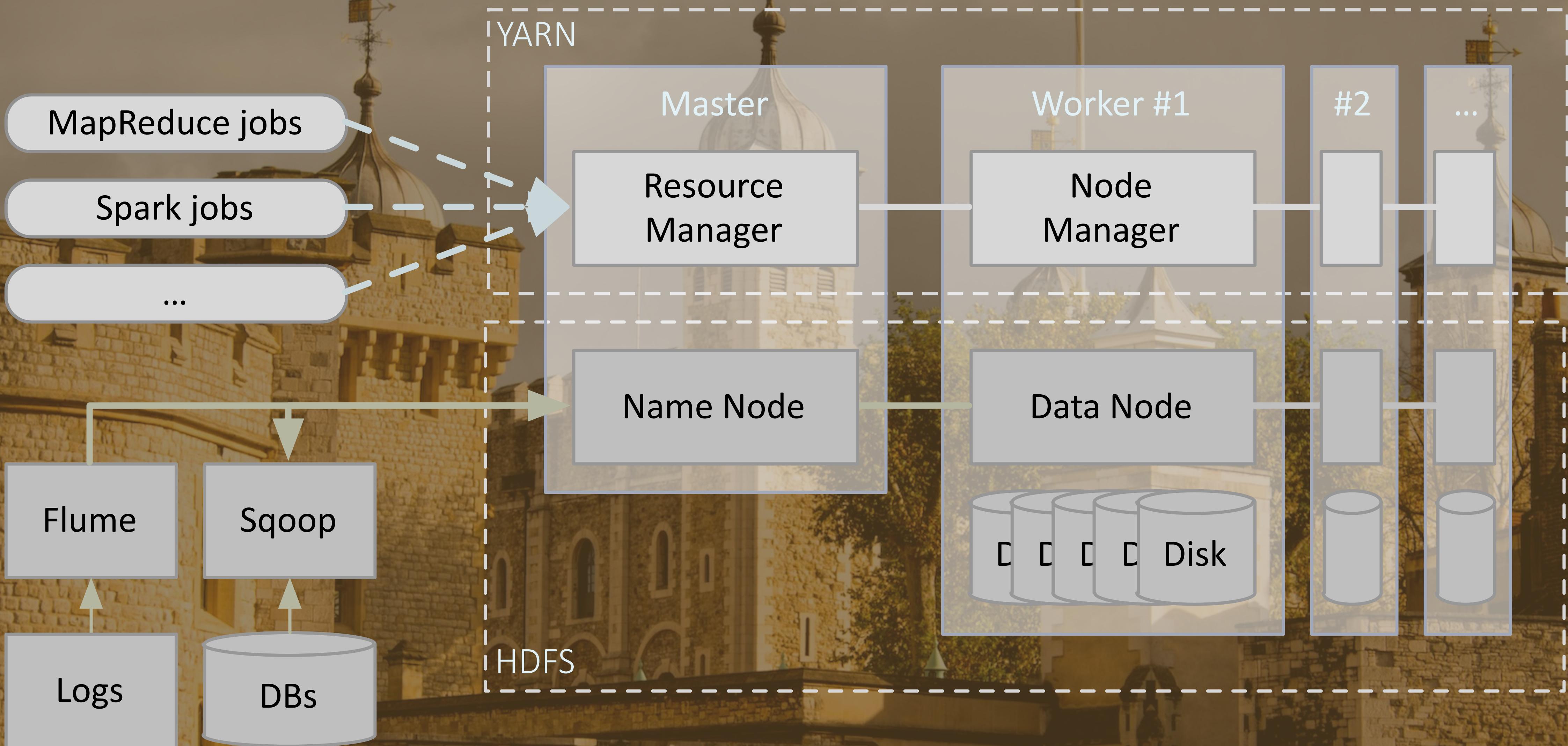
# Technology Choices

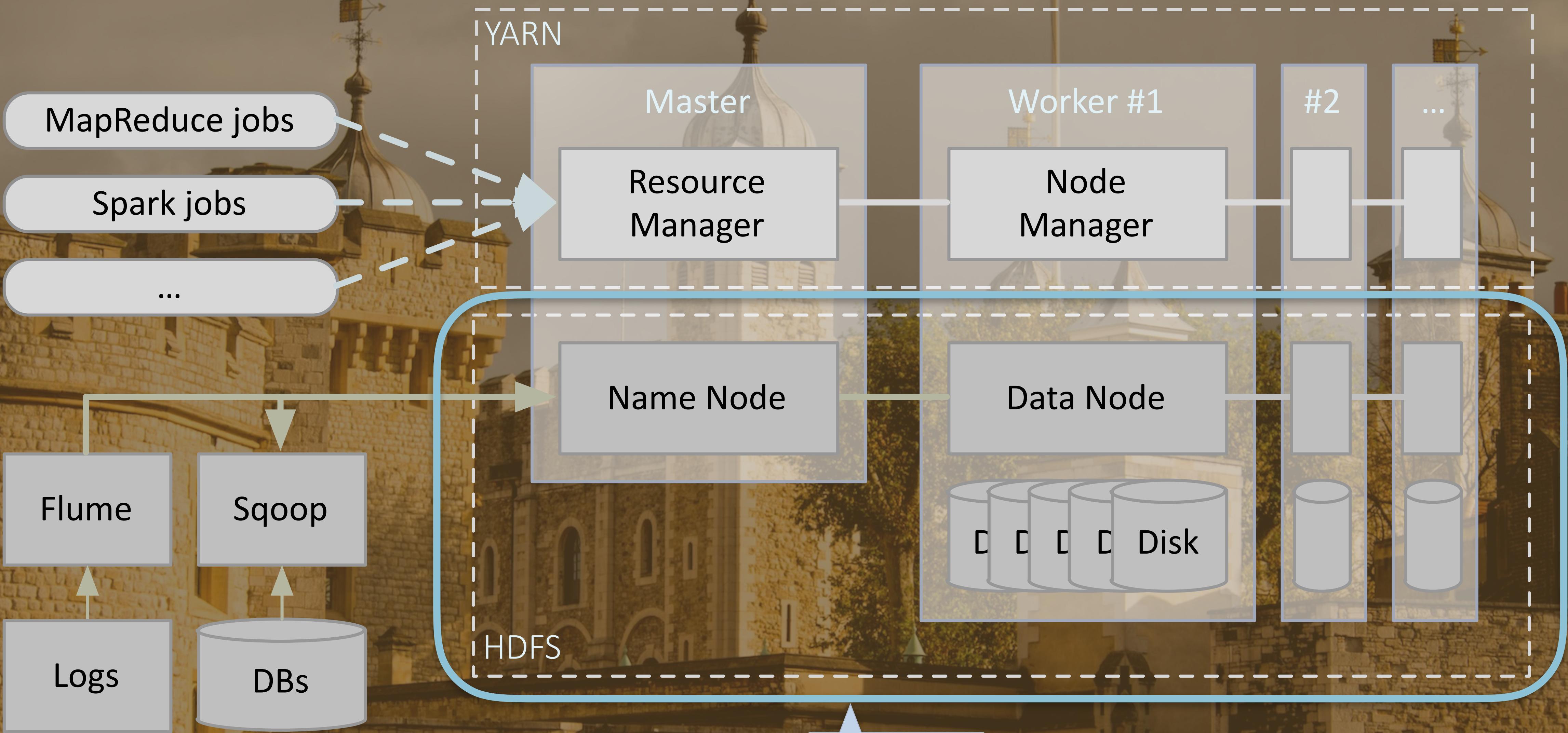
- More than “tweaking” Hadoop...
- New architectures that merge data processing with microservices

# Recall Hadoop...



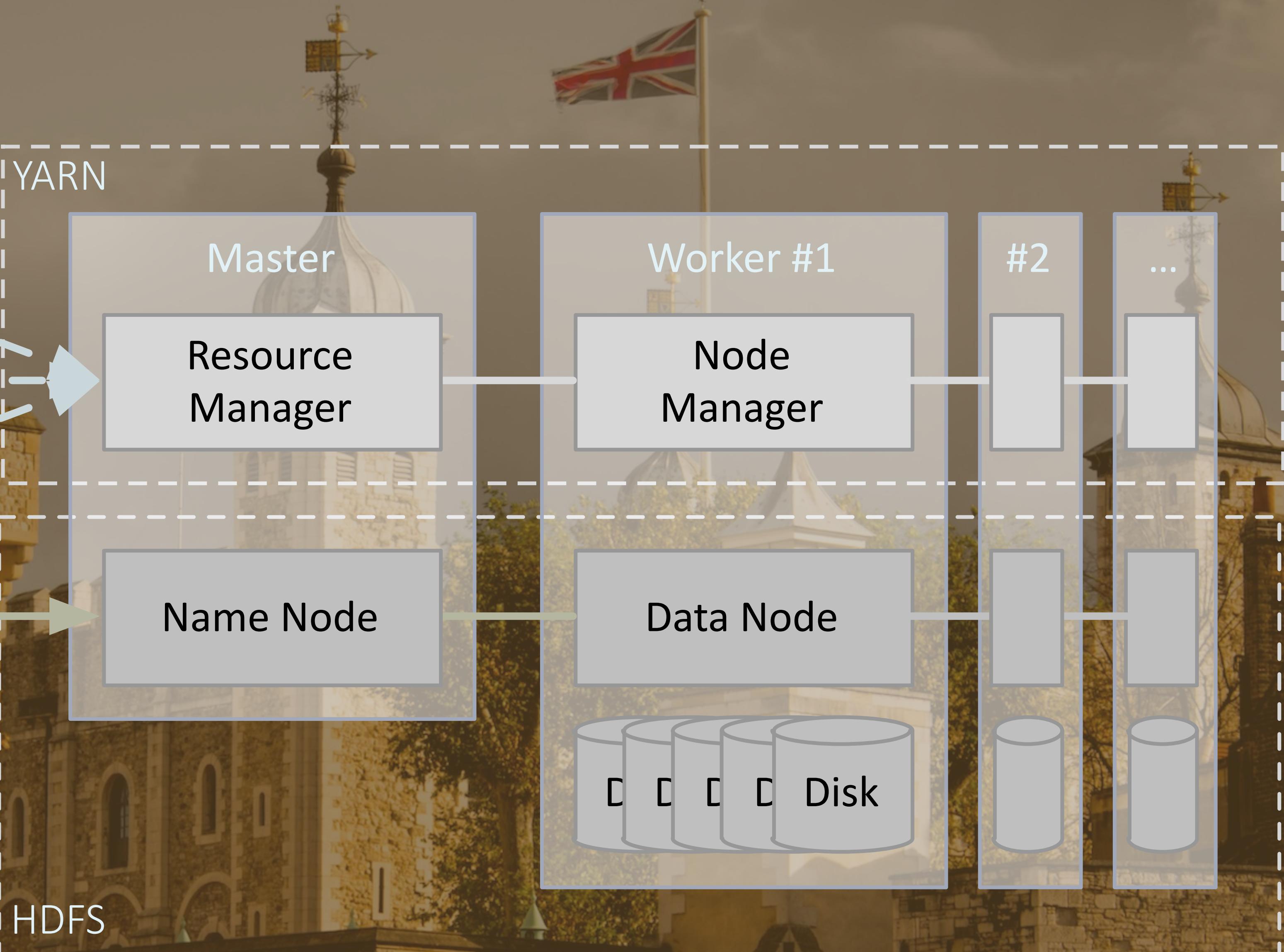
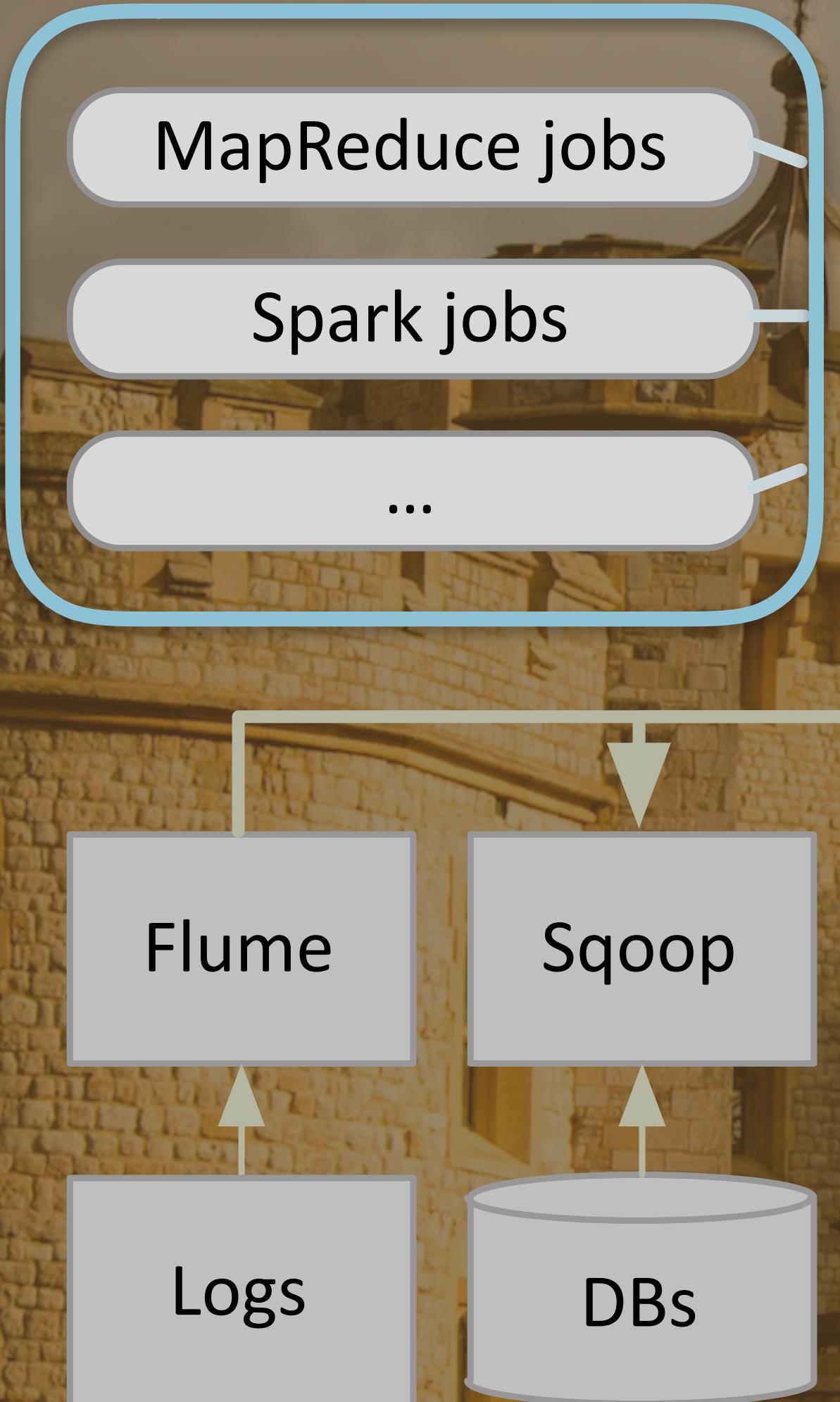
- 
- A photograph of a historic stone building, likely a university or institutional complex, featuring multiple towers and domes. A Union Jack flag flies from a pole on top of one of the towers. The building is made of light-colored stone and has several arched windows and doorways. The sky is overcast.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





Storage

# Compute



# Resource Management

MapReduce jobs

Spark jobs

...

Flume

Sqoop

Logs

DBs

YARN

Master

Resource  
Manager

Worker #1

Node  
Manager

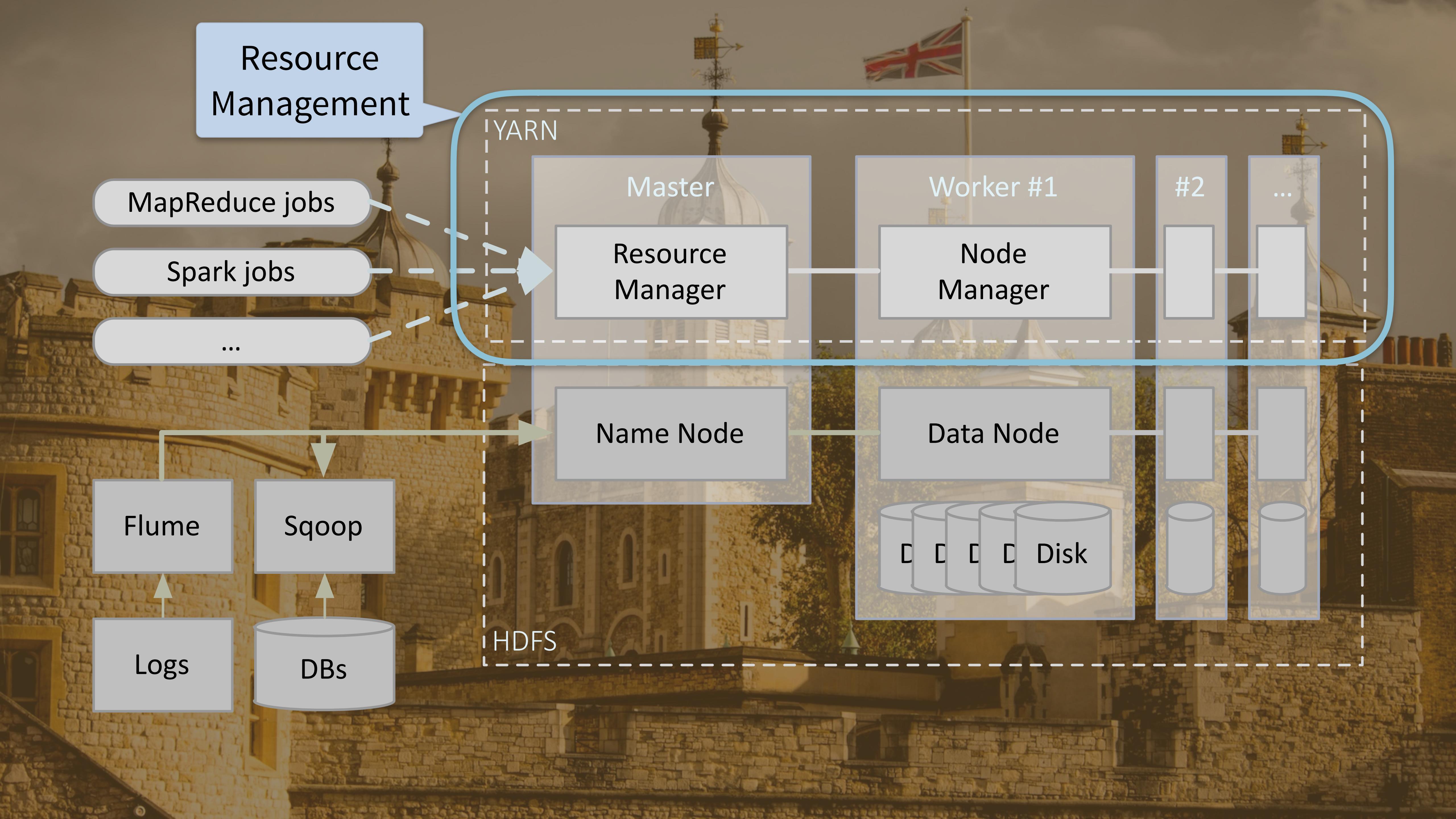
#2

...

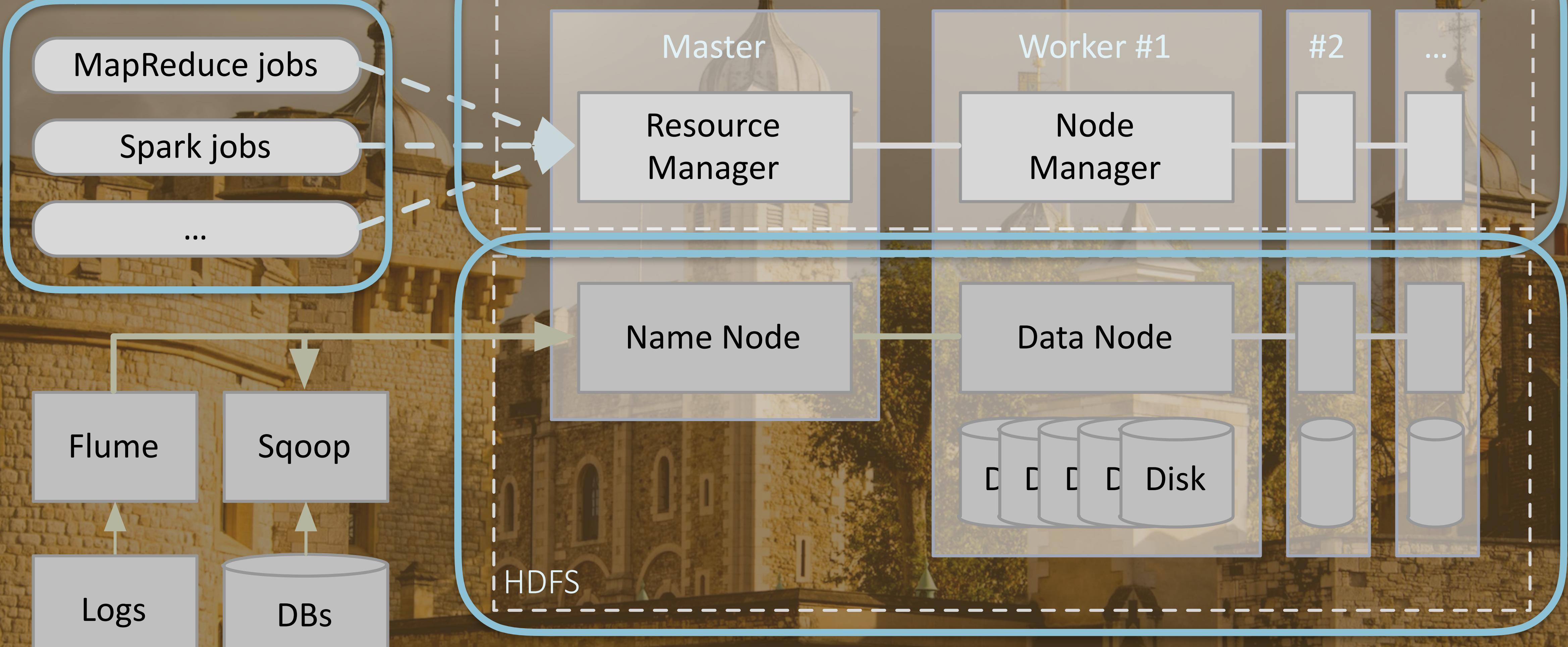
HDFS

Name Node

Data Node



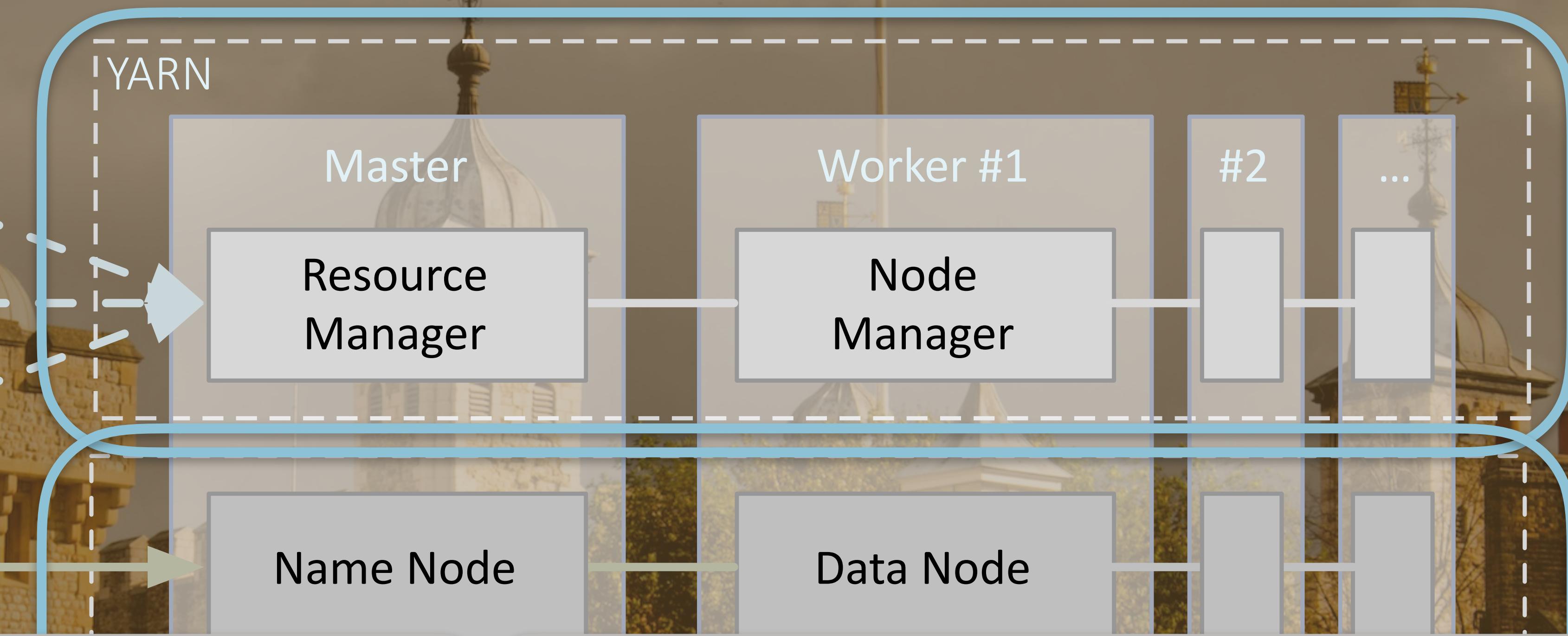
# Database Deconstructed!



The three major components inside your RDBMS black box are compute, storage, and resource management.

# Database Deconstructed!

MapReduce jobs  
Spark jobs  
...

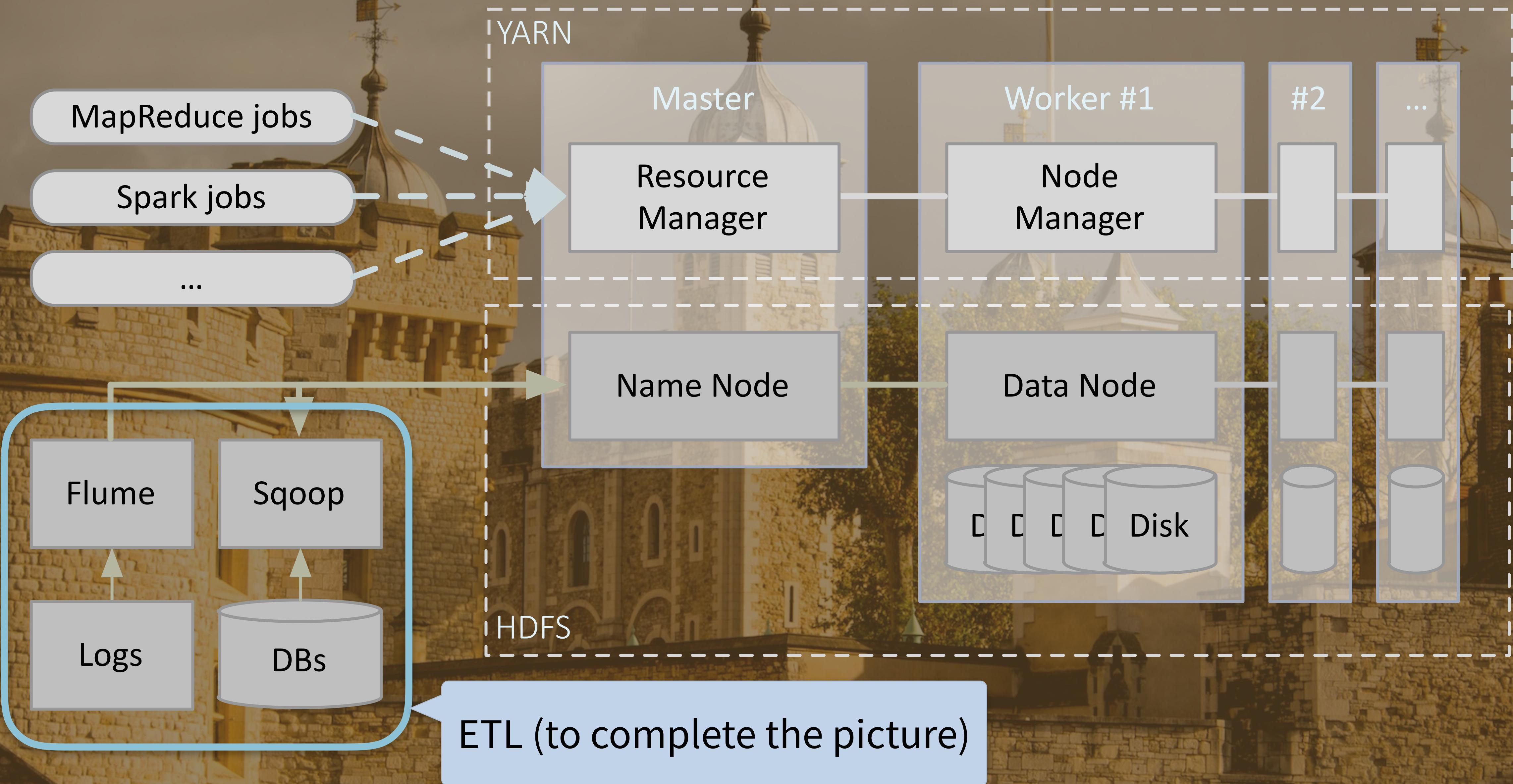


## Good:

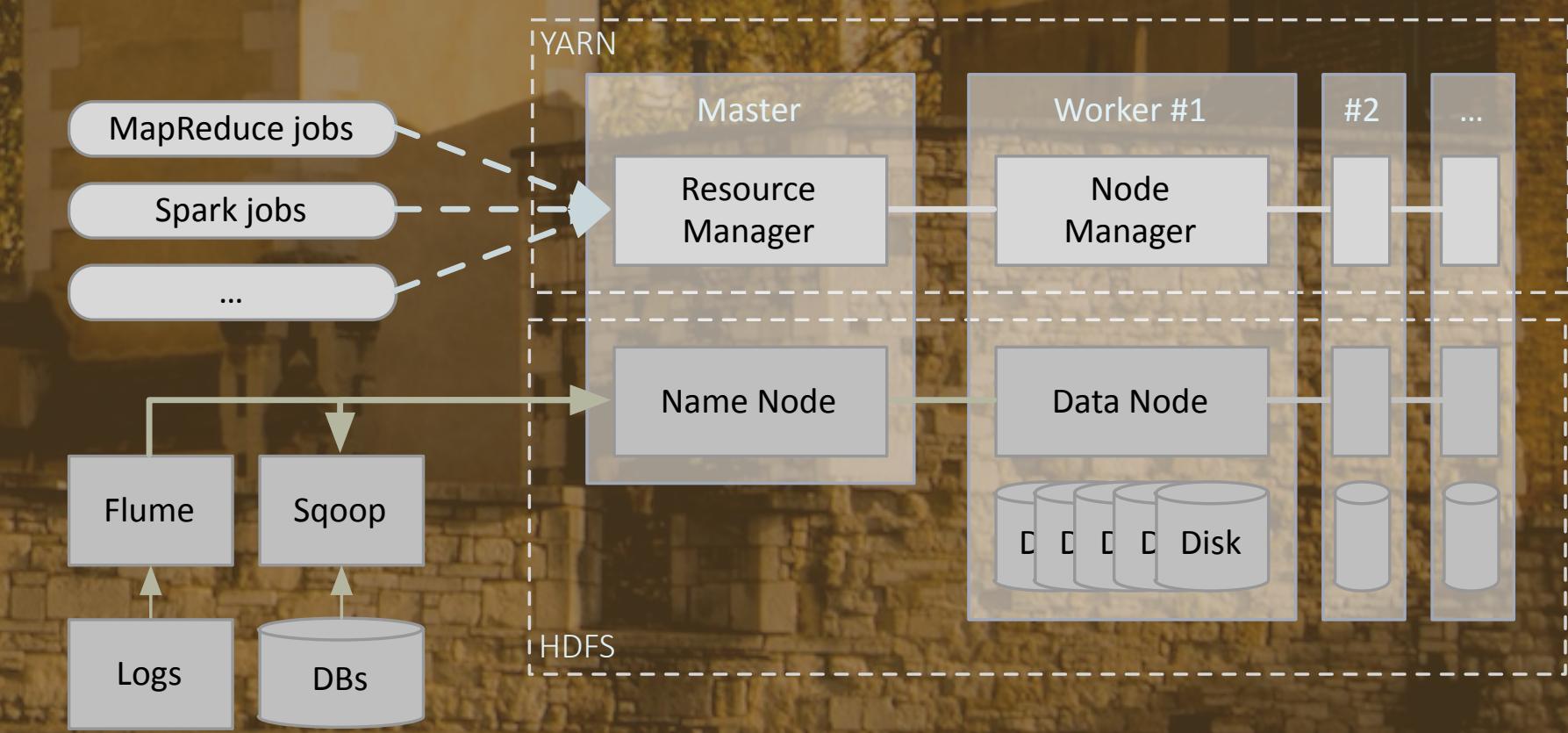
Lots more flexibility for each component, allowing Hadoop to be used for a wider class of needs than a traditional database can cover, including better scalability, cost containment, more data models and formats, processing beyond SQL, etc.

## Bad:

Now it's your responsibility to integrate these pieces and make them work well together.

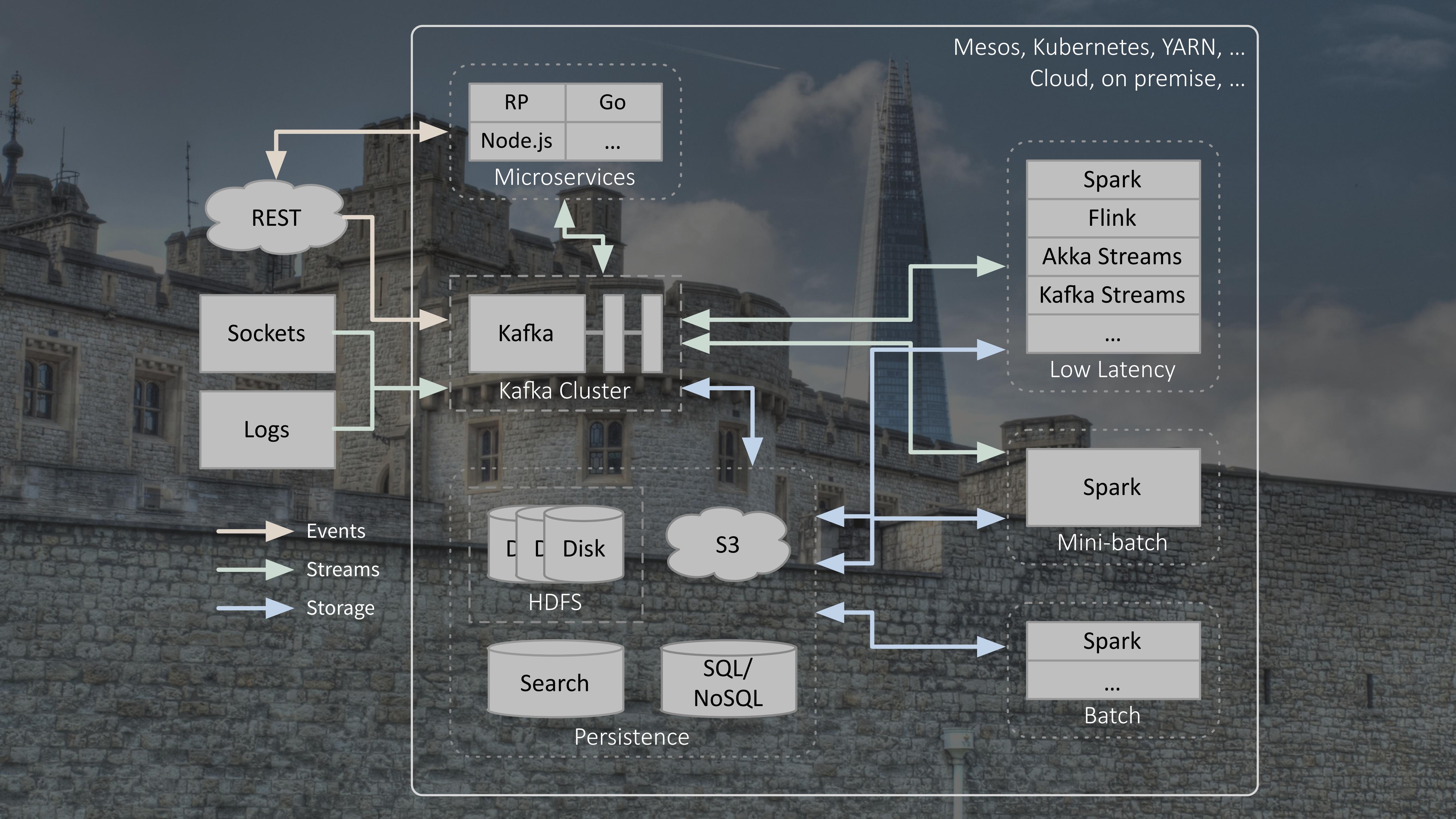


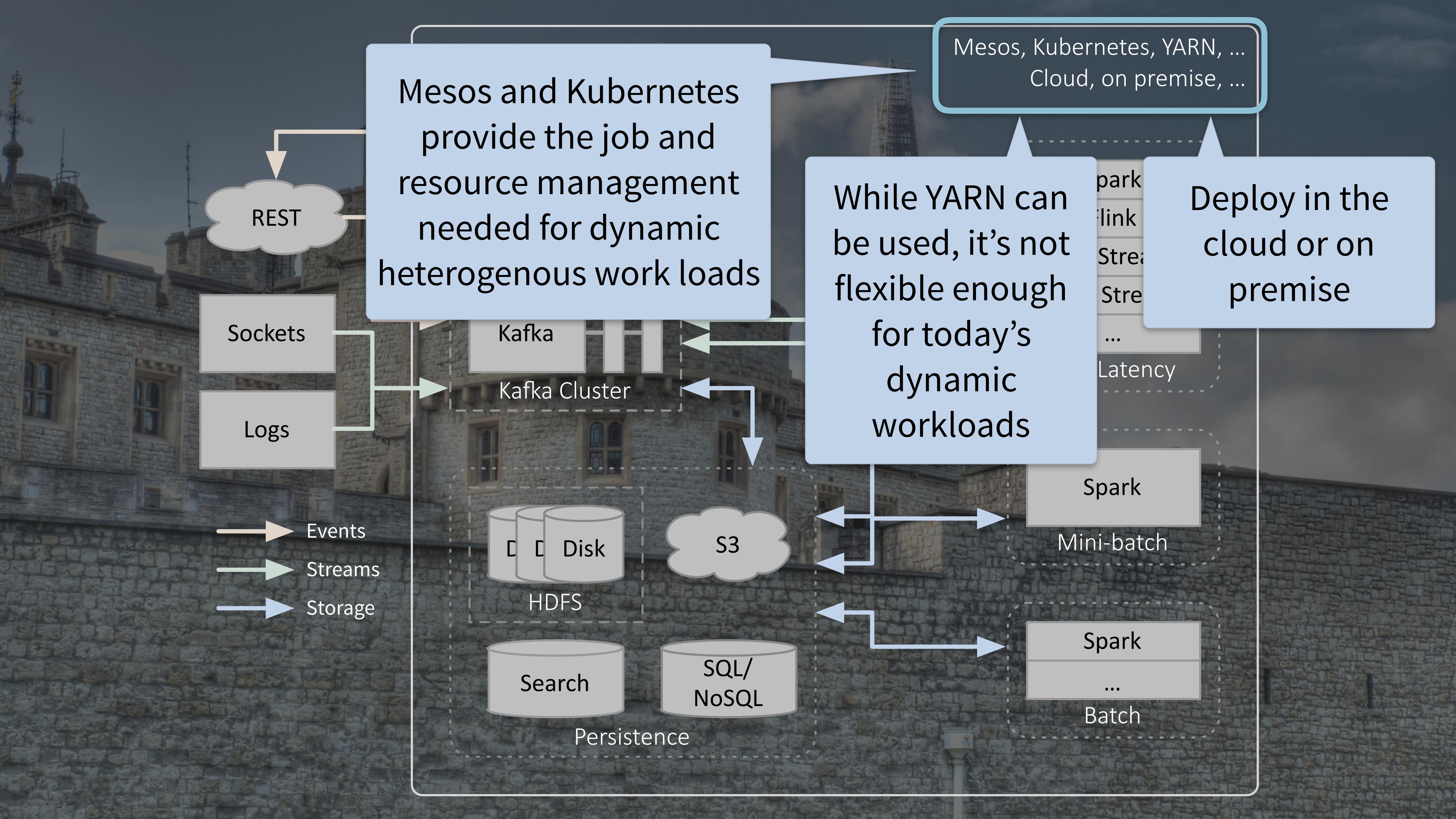
- 10000 meter view:
  - Hadoop is the database deconstructed and reimagined
  - Ideal for batch and interactive apps
  - ... but also constrained by that model





# New Fast Data Architecture





Mesos and Kubernetes provide the job and resource management needed for dynamic heterogenous work loads

While YARN can be used, it's not flexible enough for today's dynamic workloads

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

Deploy in the cloud or on premise

Sockets

Logs

Kafka

Kafka Cluster

Spark

Mini-batch

Spark

Batch

S3

Disk

HDFS

Search

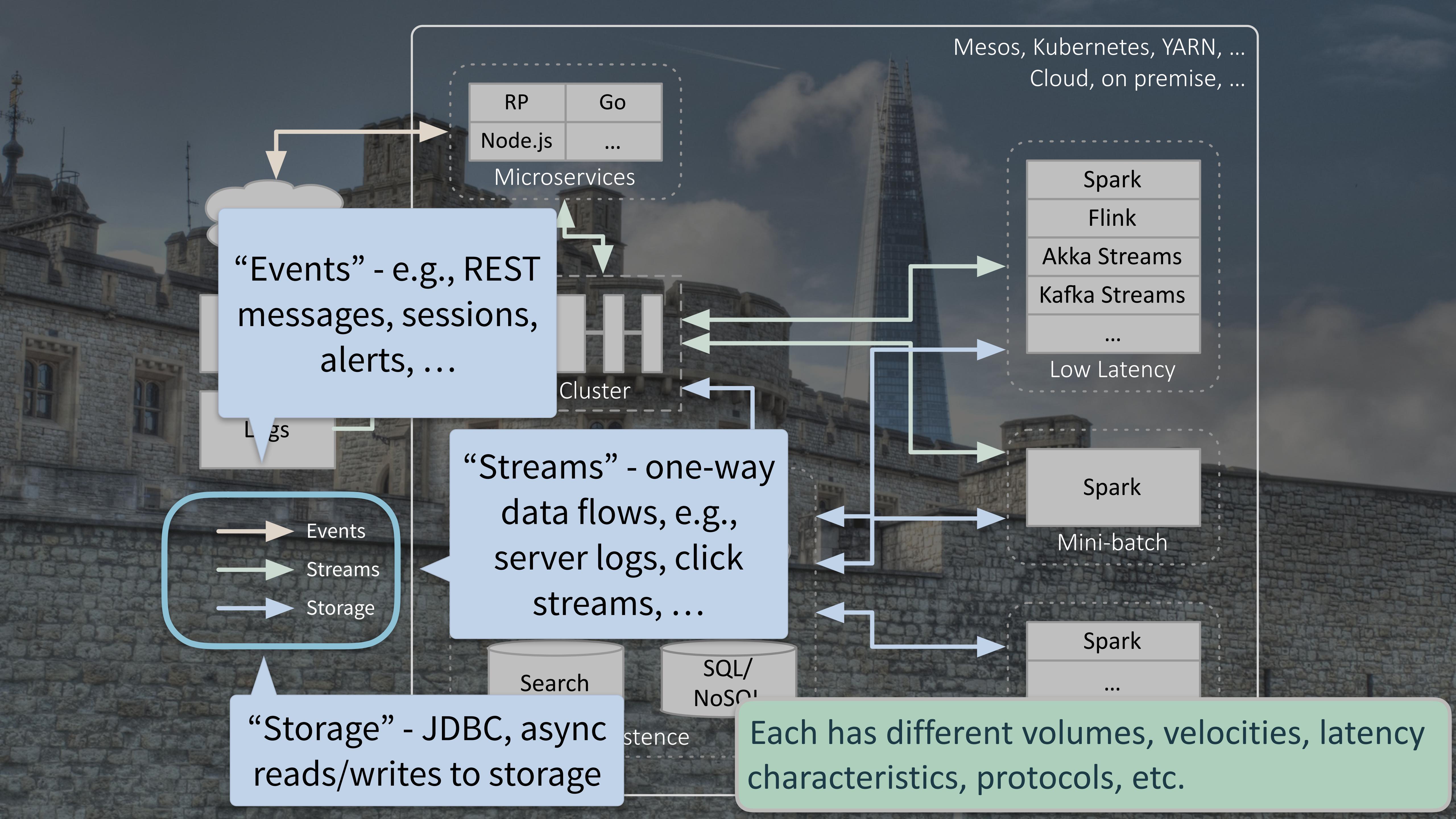
SQL/  
NoSQL

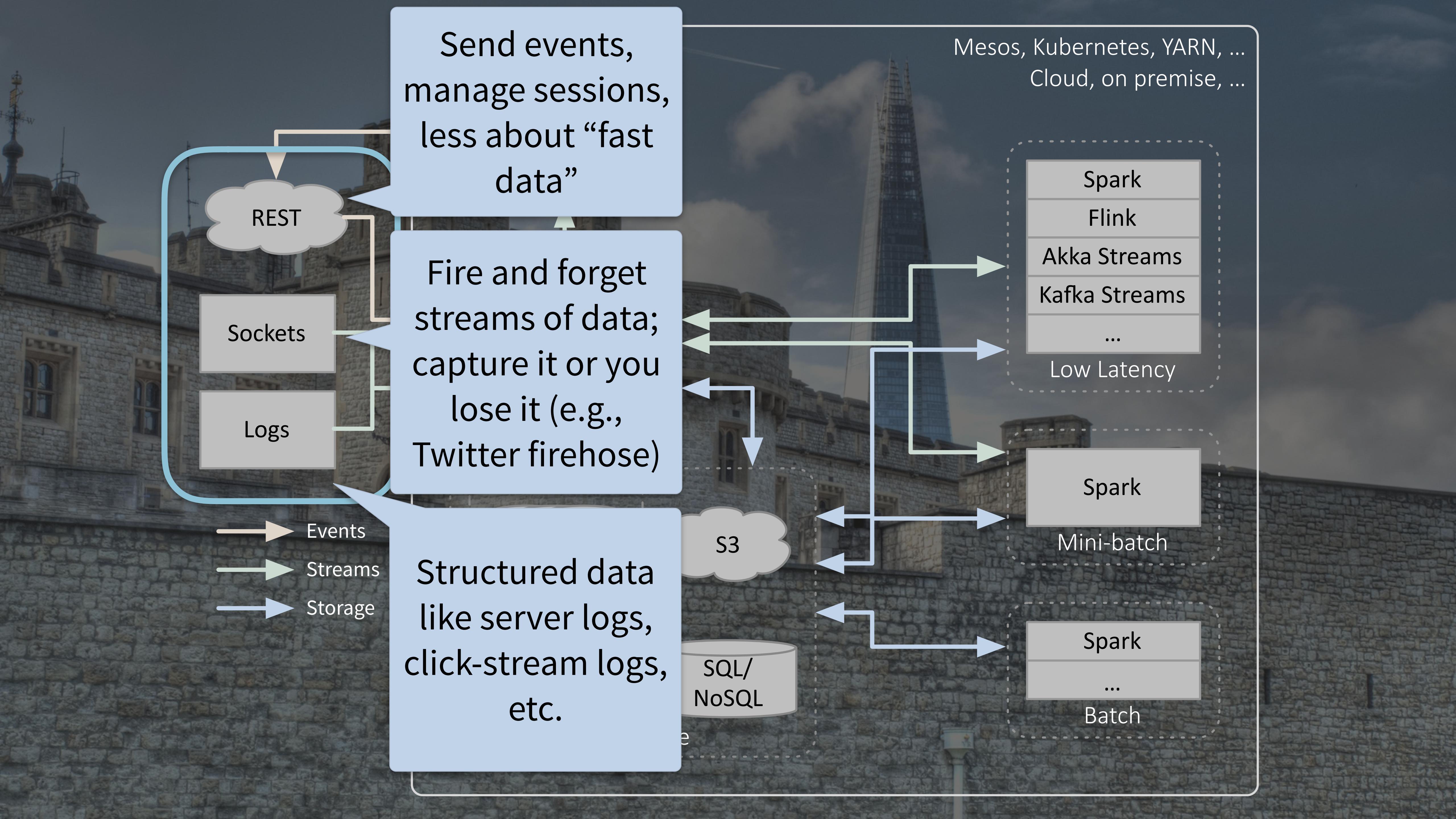
Persistence

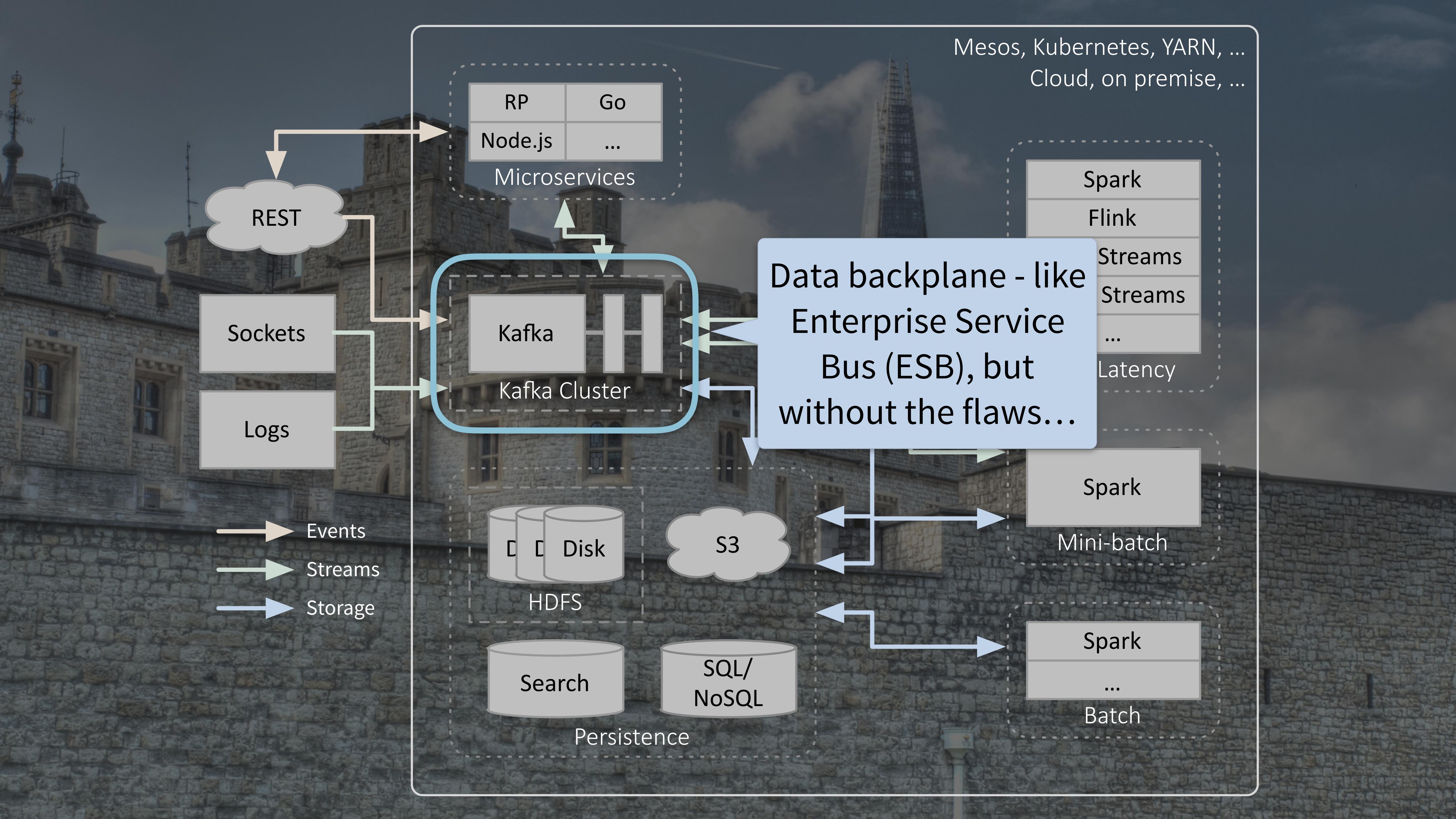
Events

Streams

Storage

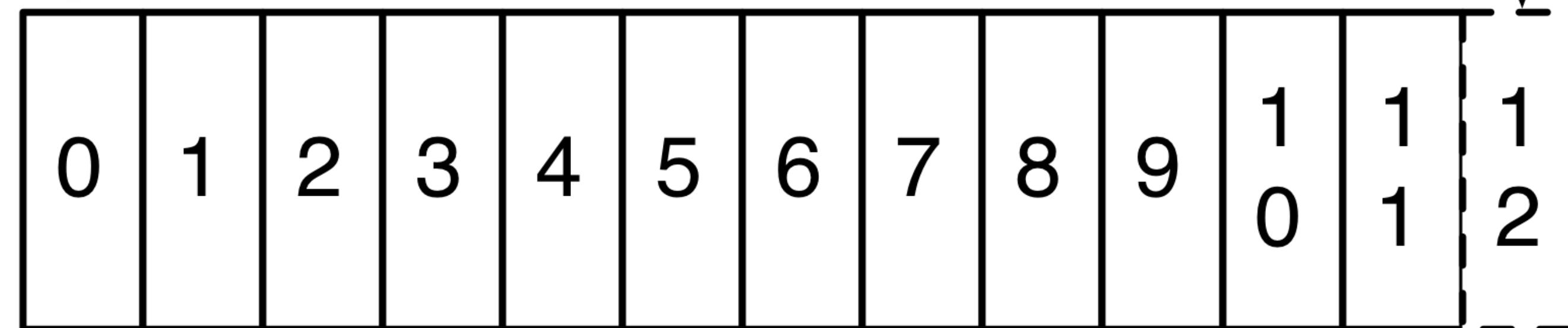






# Why Kafka?

Organized into topics

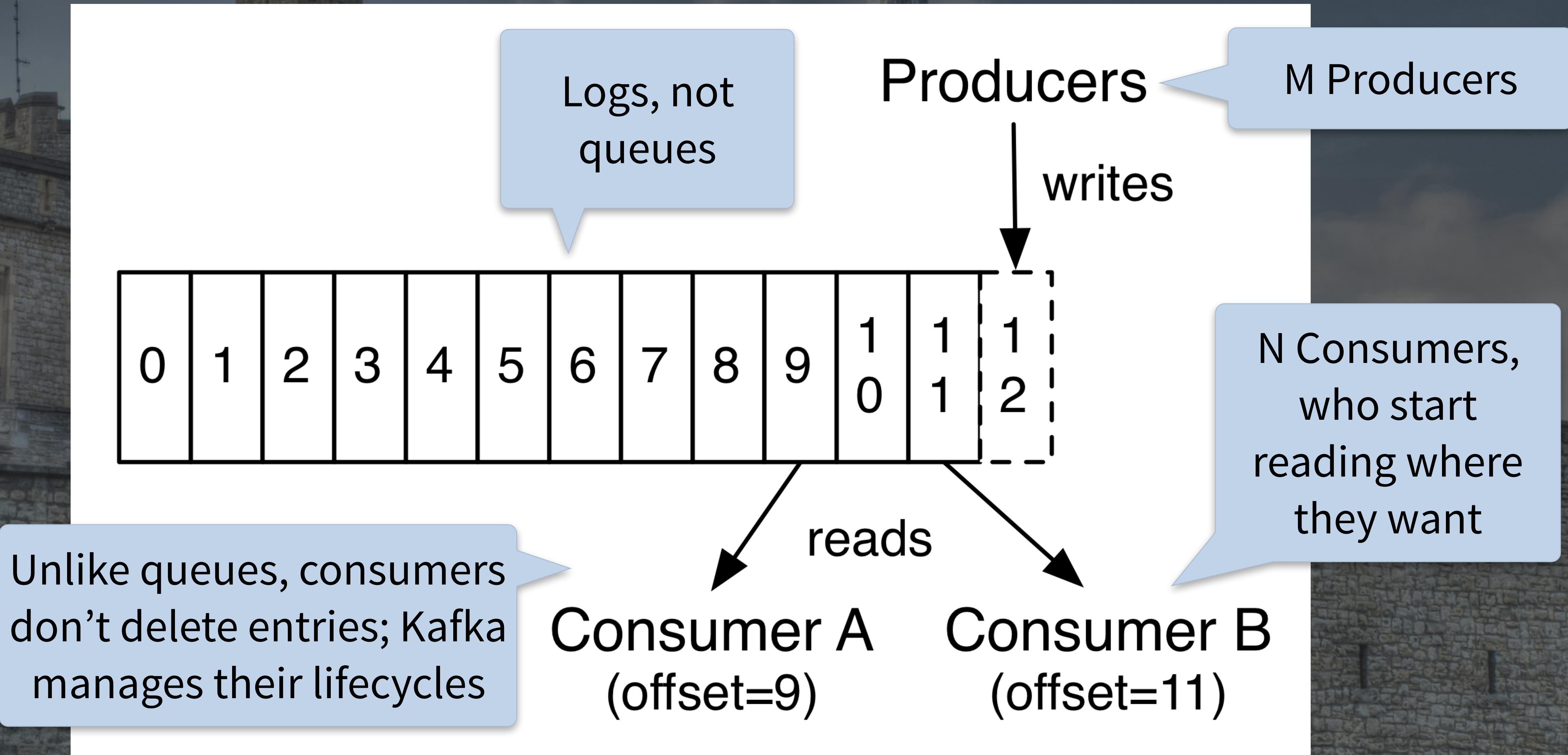


Topics are partitioned,  
replicated, and  
distributed

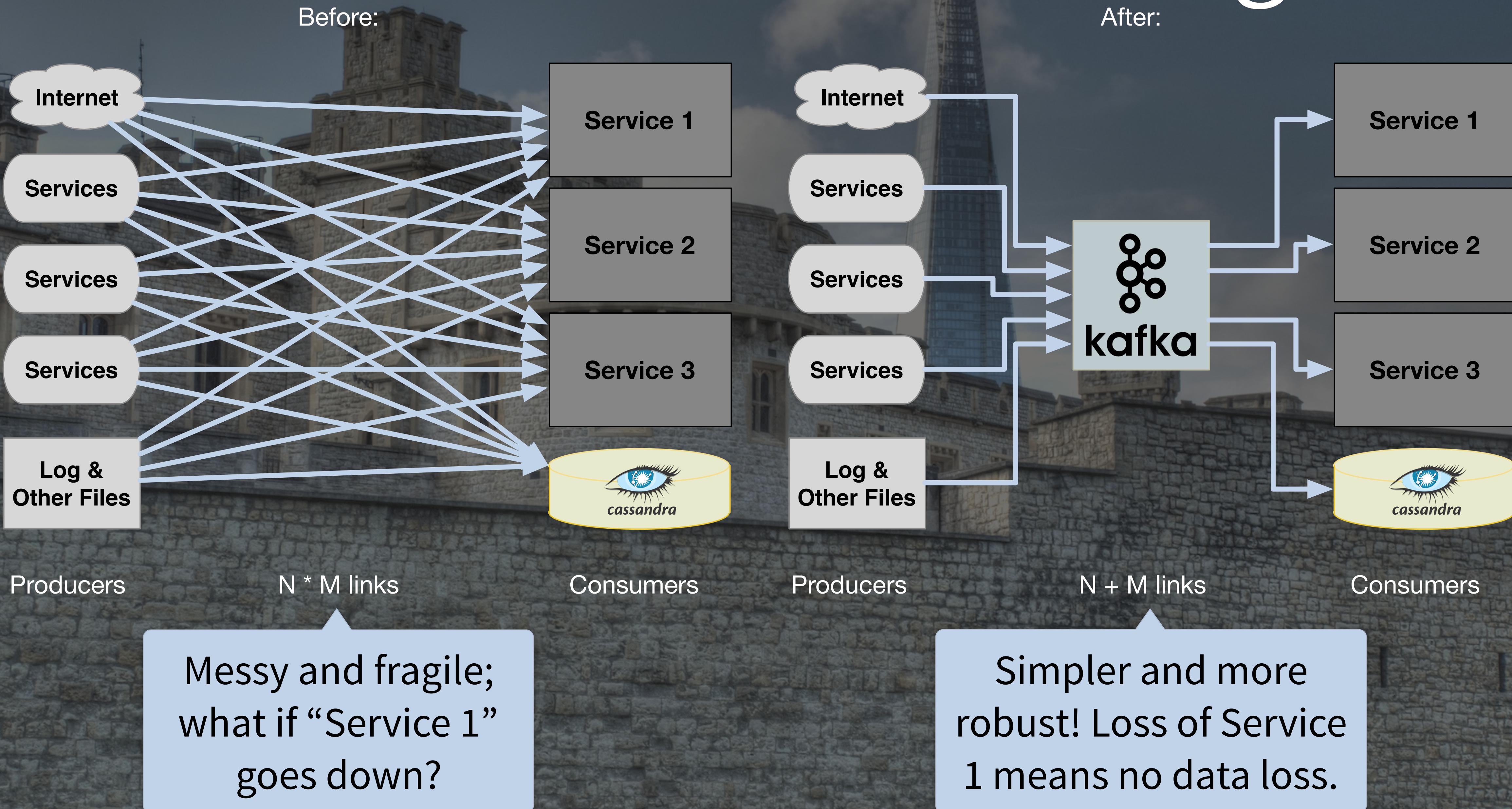
Consumer A  
(offset=9)

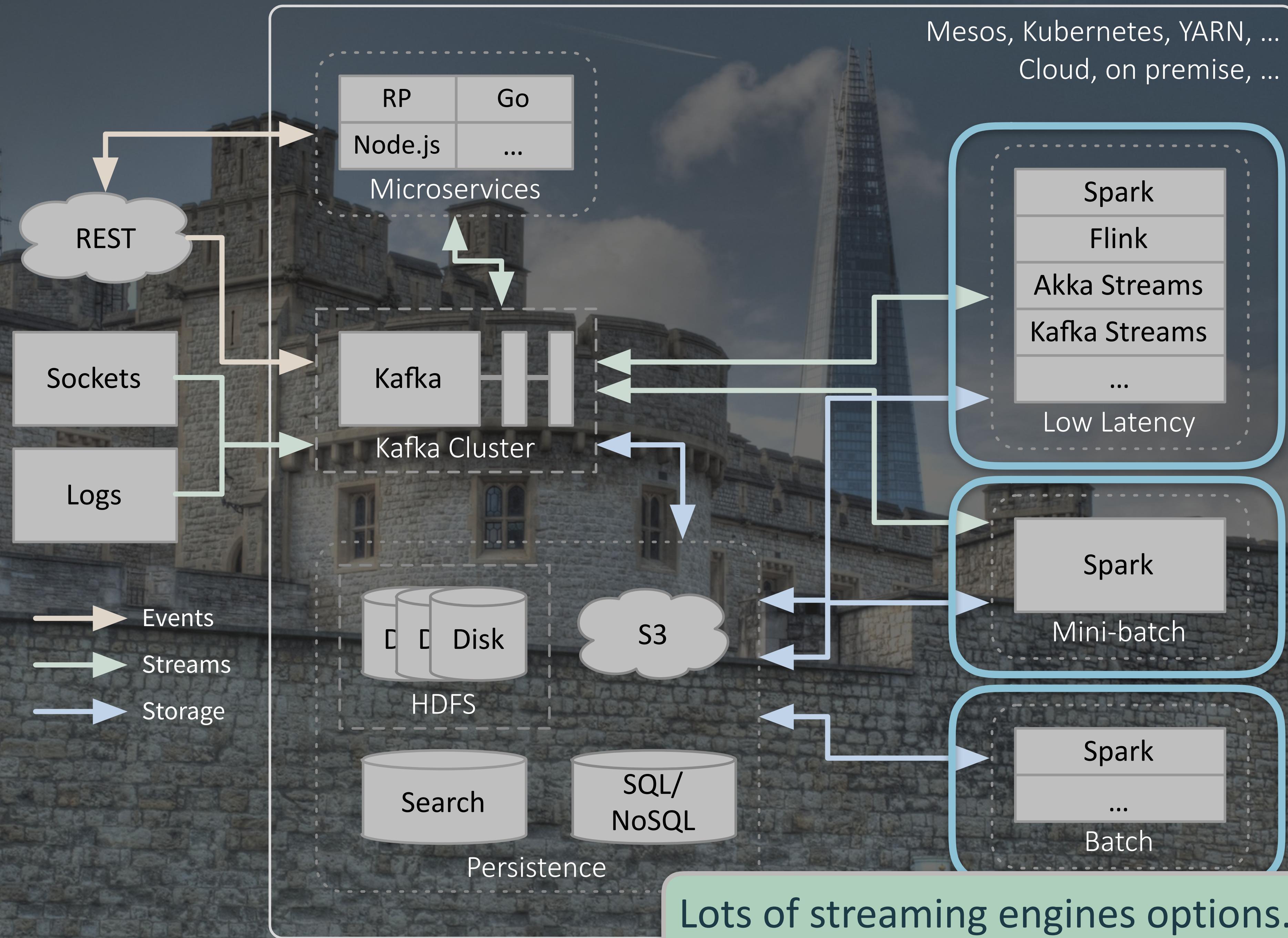
Consumer B  
(offset=11)

# Why Kafka?



# Using Kafka

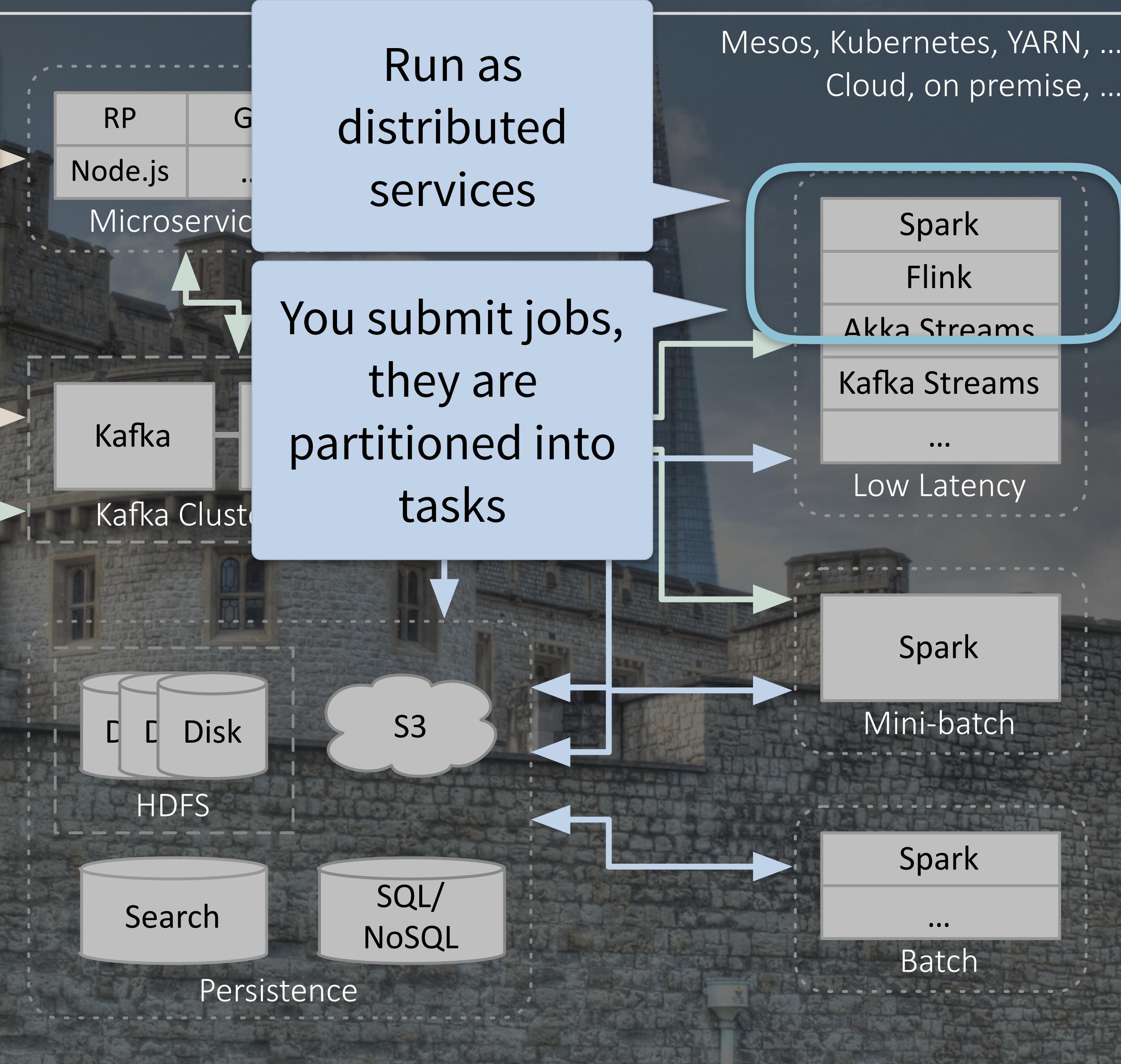
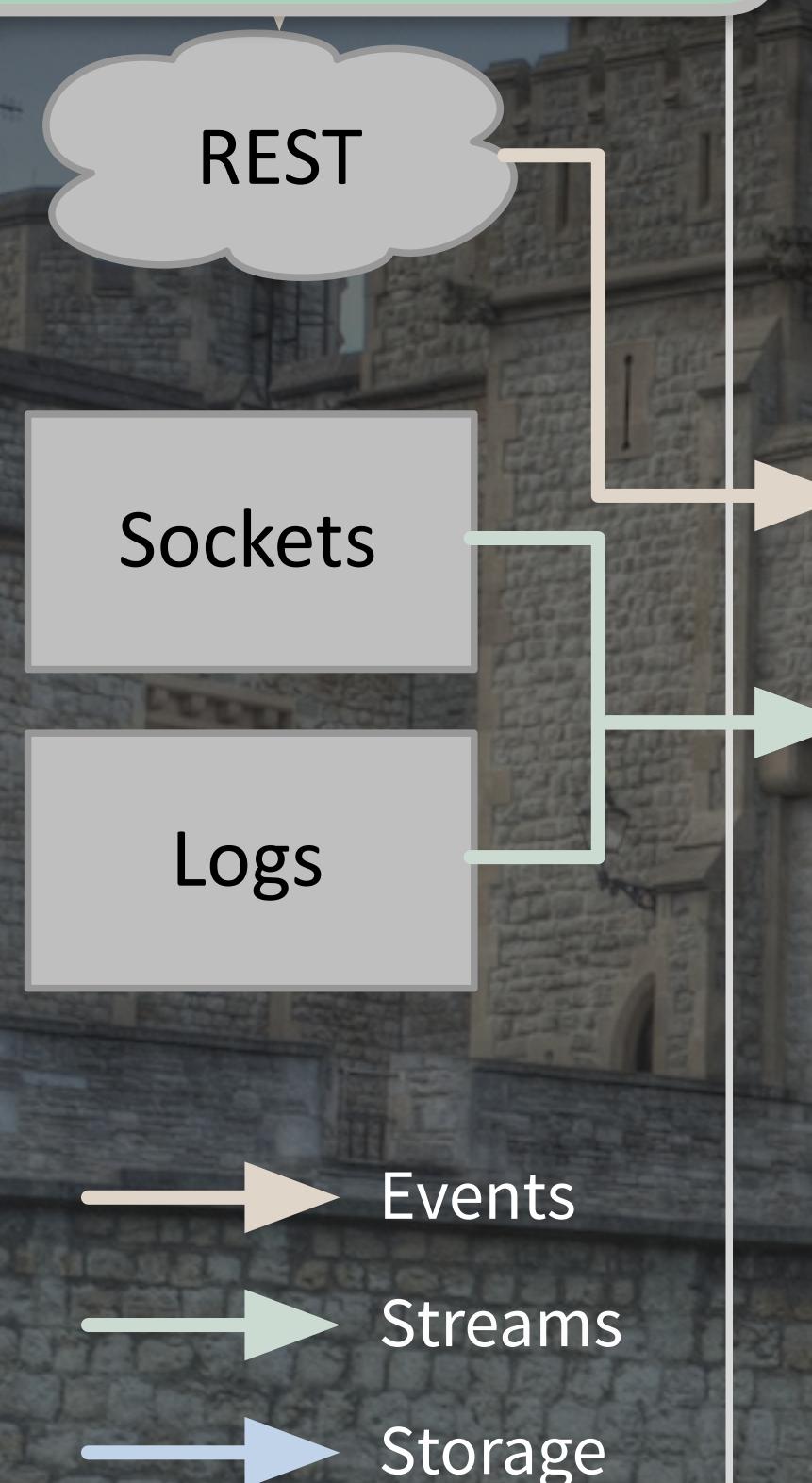




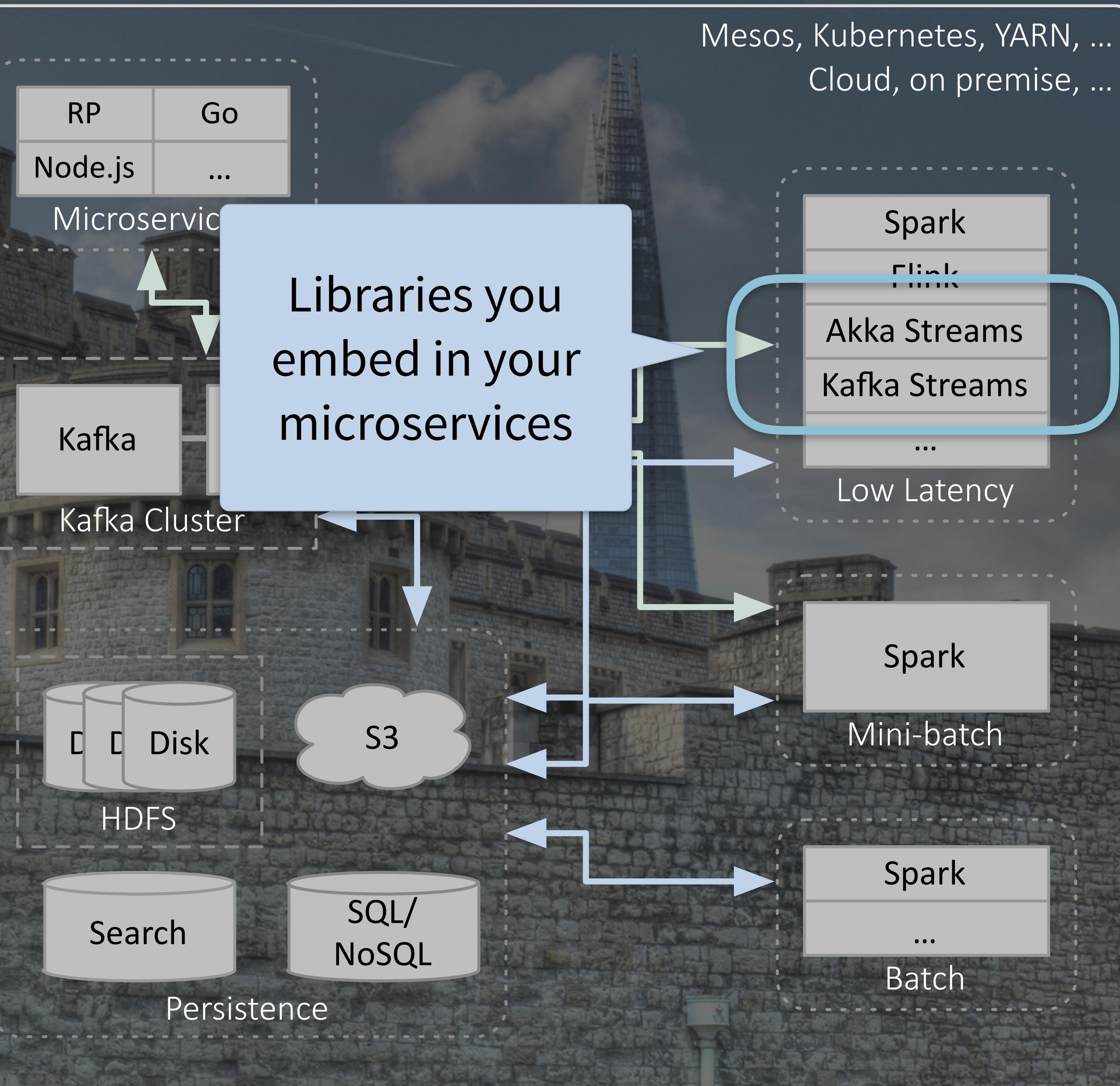
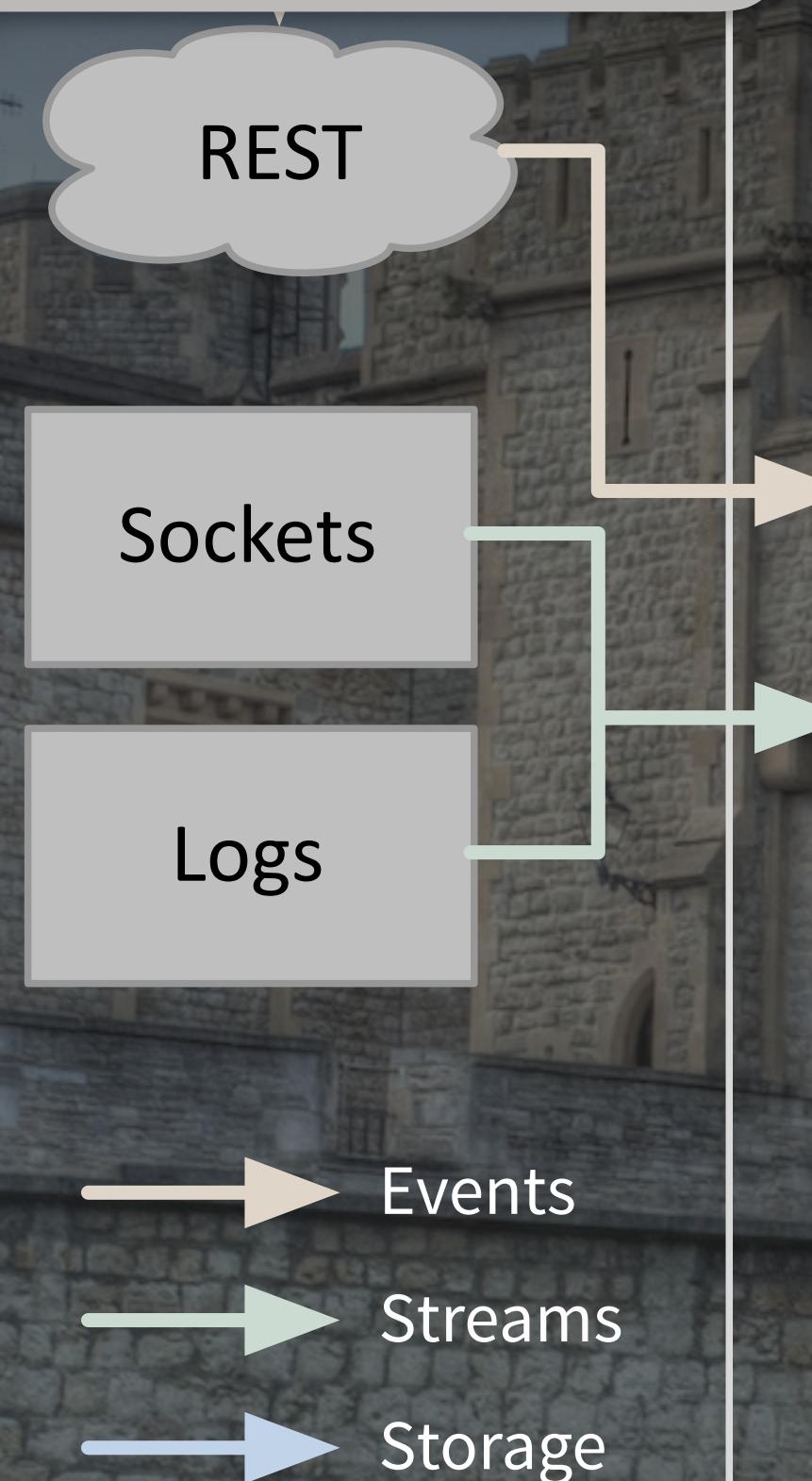
# How do you choose?

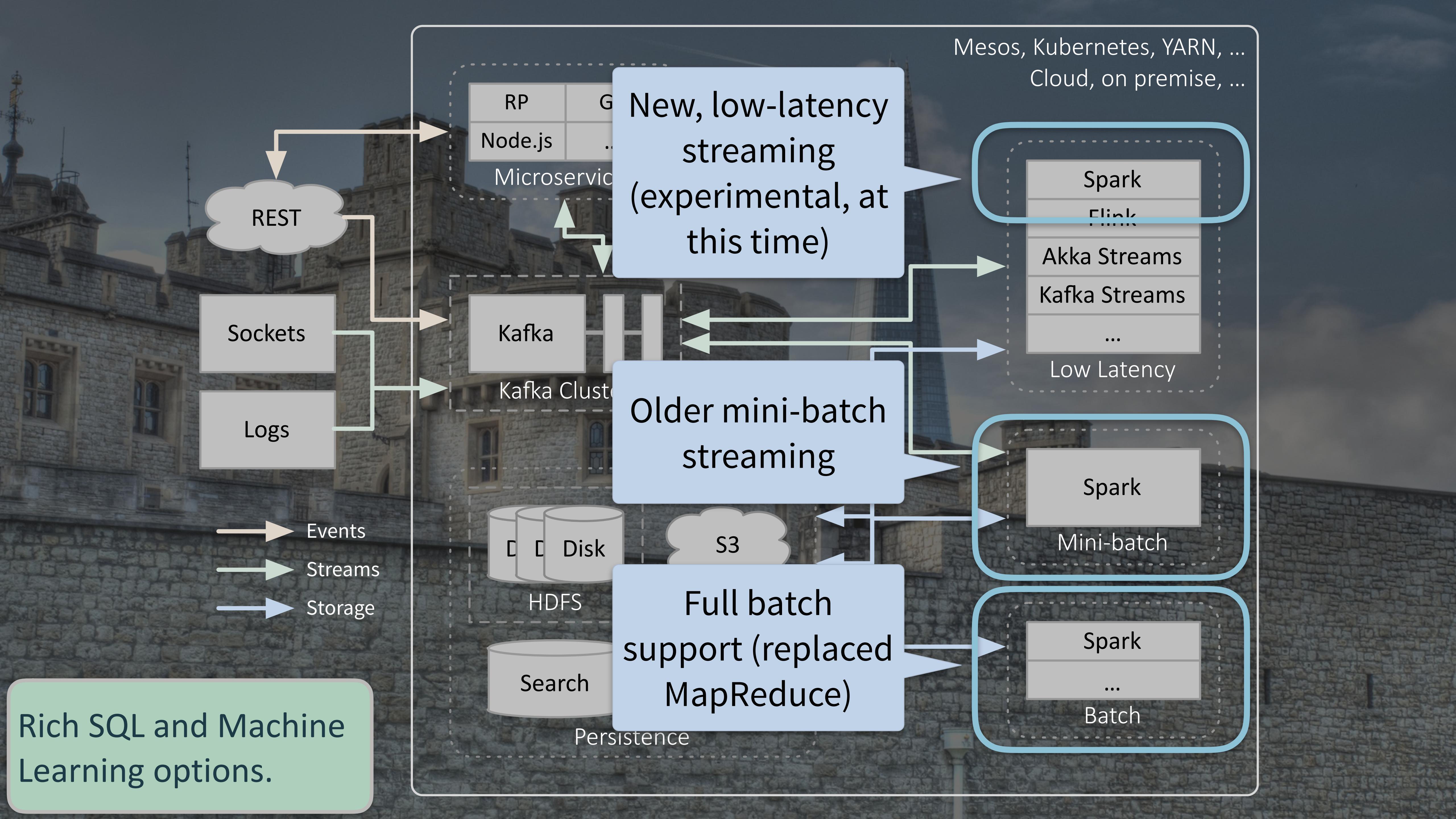
- Latency? How low?
- Volume: How high?
- Which kinds of data processing?
- How do you want to build, deploy, and manage these services?

The streaming engines form two groups:



The streaming engines form two groups:





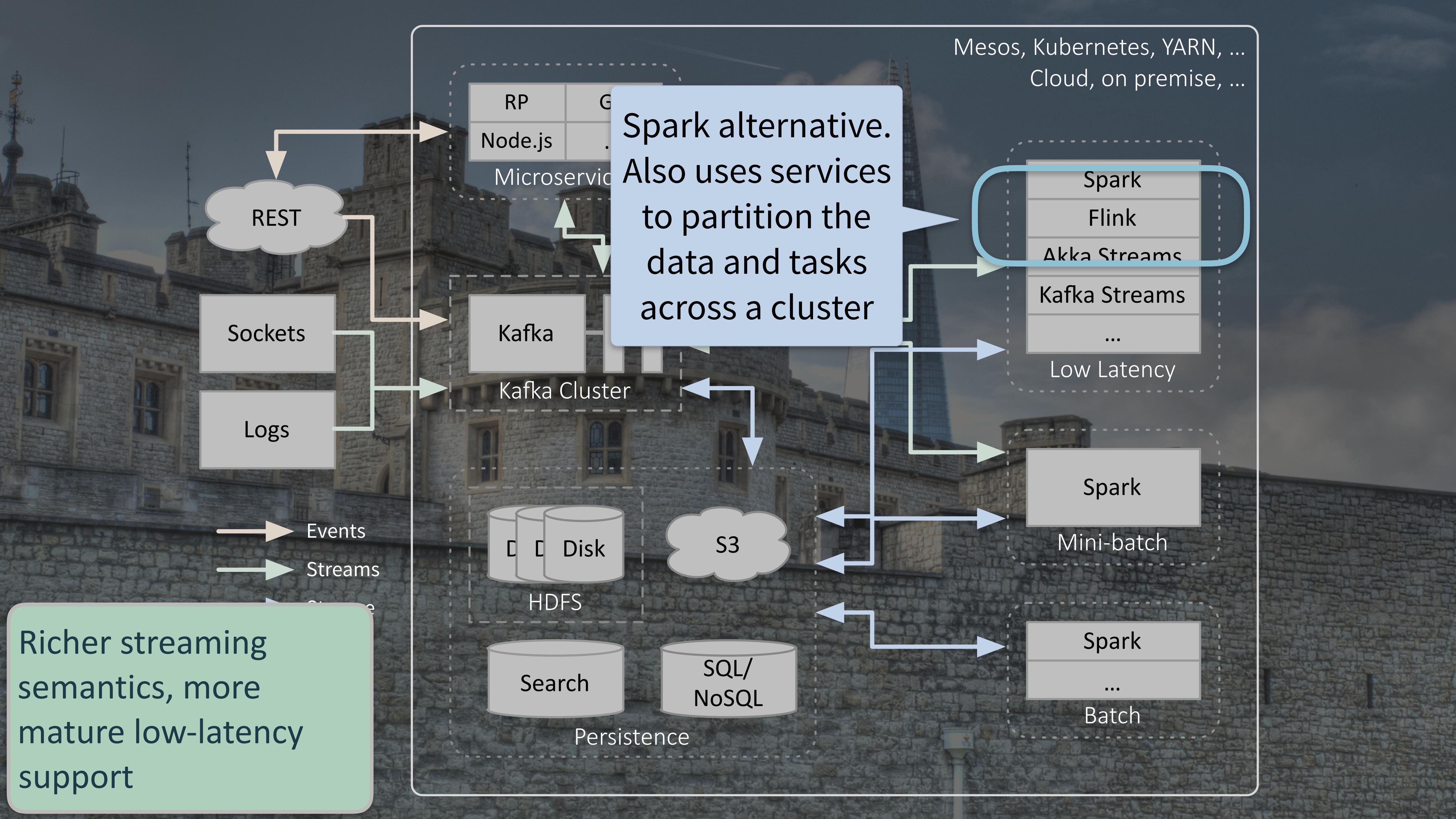
Mesos, Kubernetes, YARN, ...  
Cloud, on premise, ...

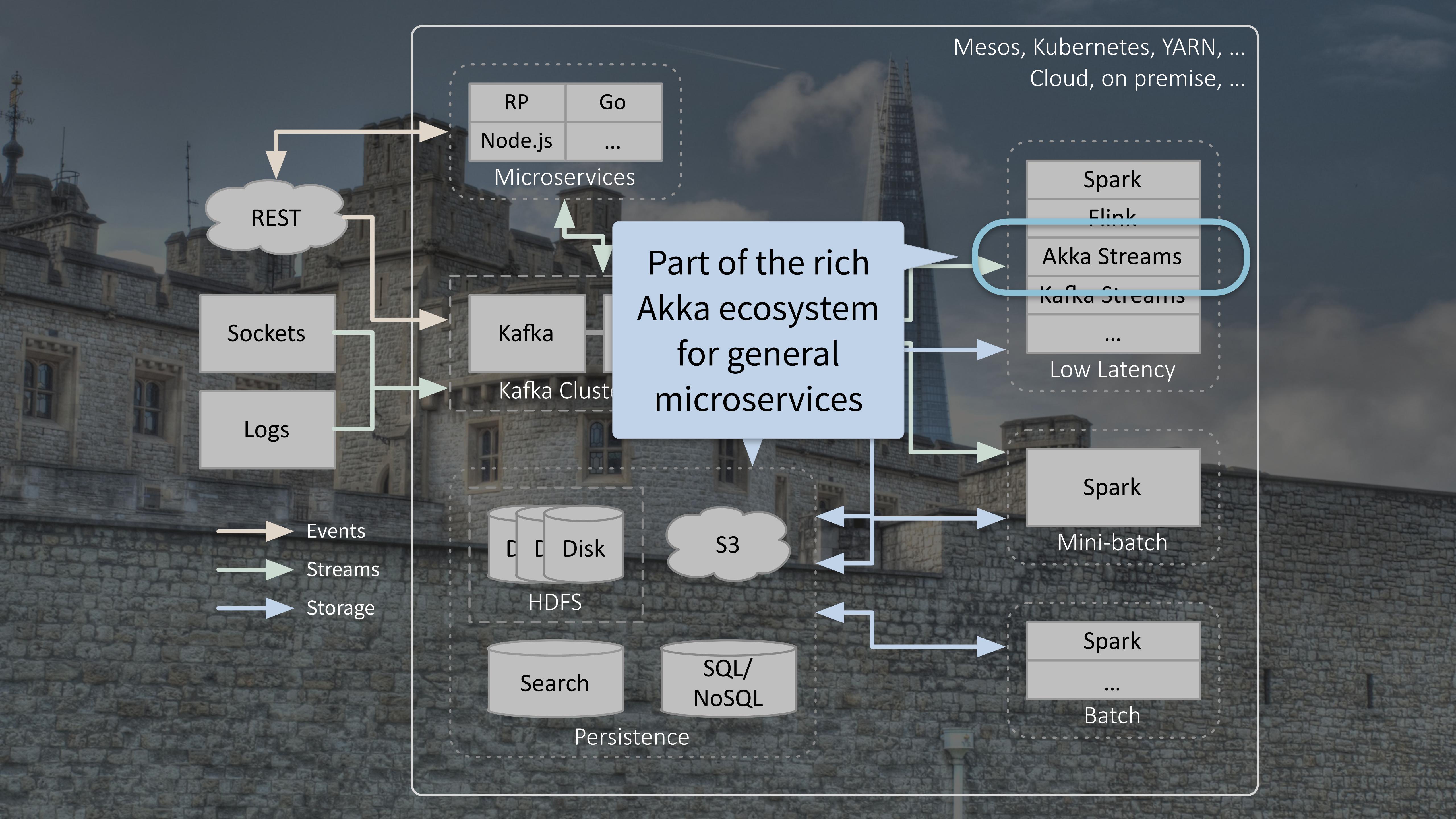
New, low-latency  
streaming  
(experimental, at  
this time)

Older mini-batch  
streaming

Full batch  
support (replaced  
MapReduce)

Rich SQL and Machine  
Learning options.





Part of the rich  
Akka ecosystem  
for general  
microservices

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

Spark

Elink

Akka Streams

Kafka Streams

...

Low Latency

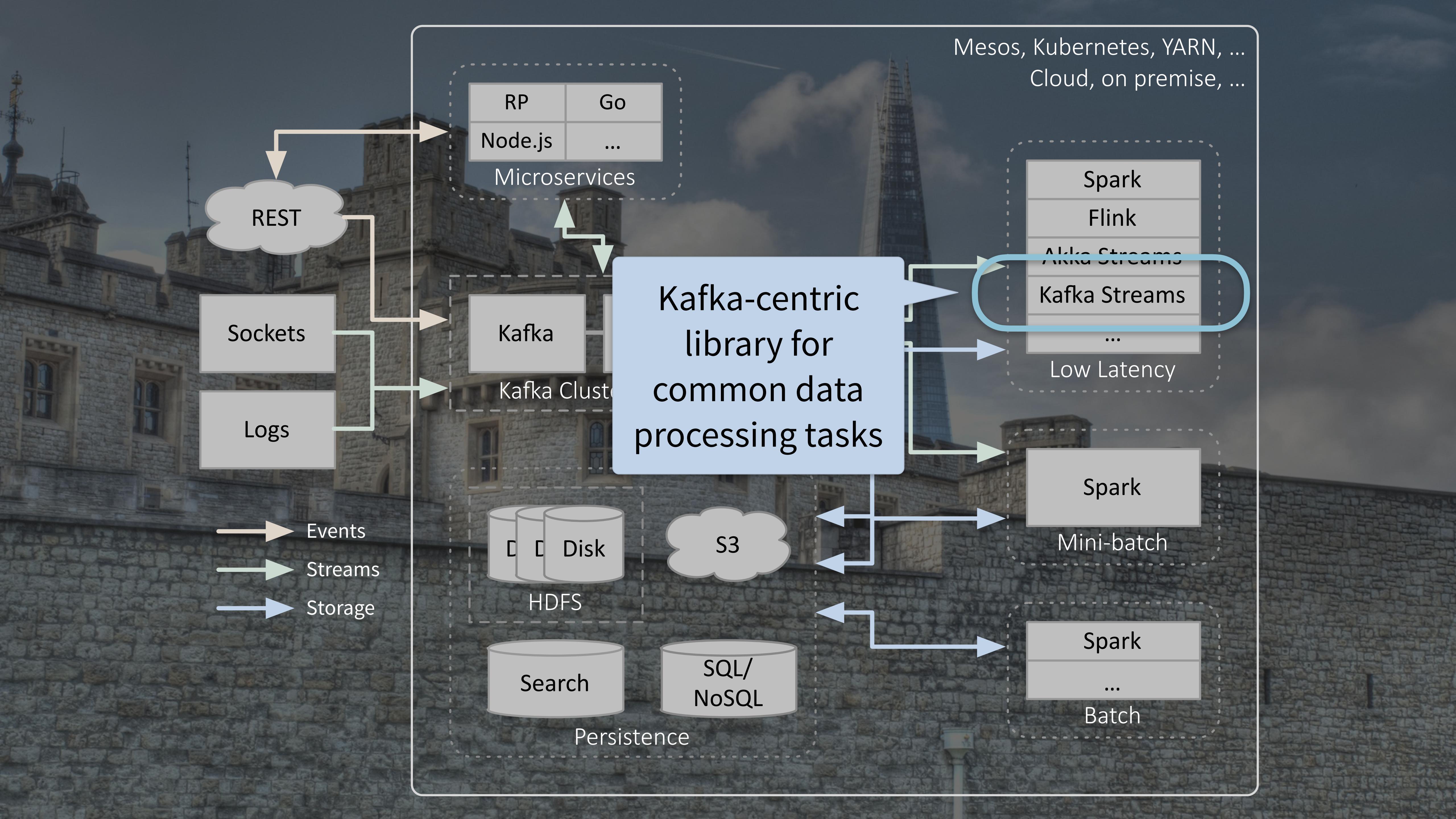
Spark

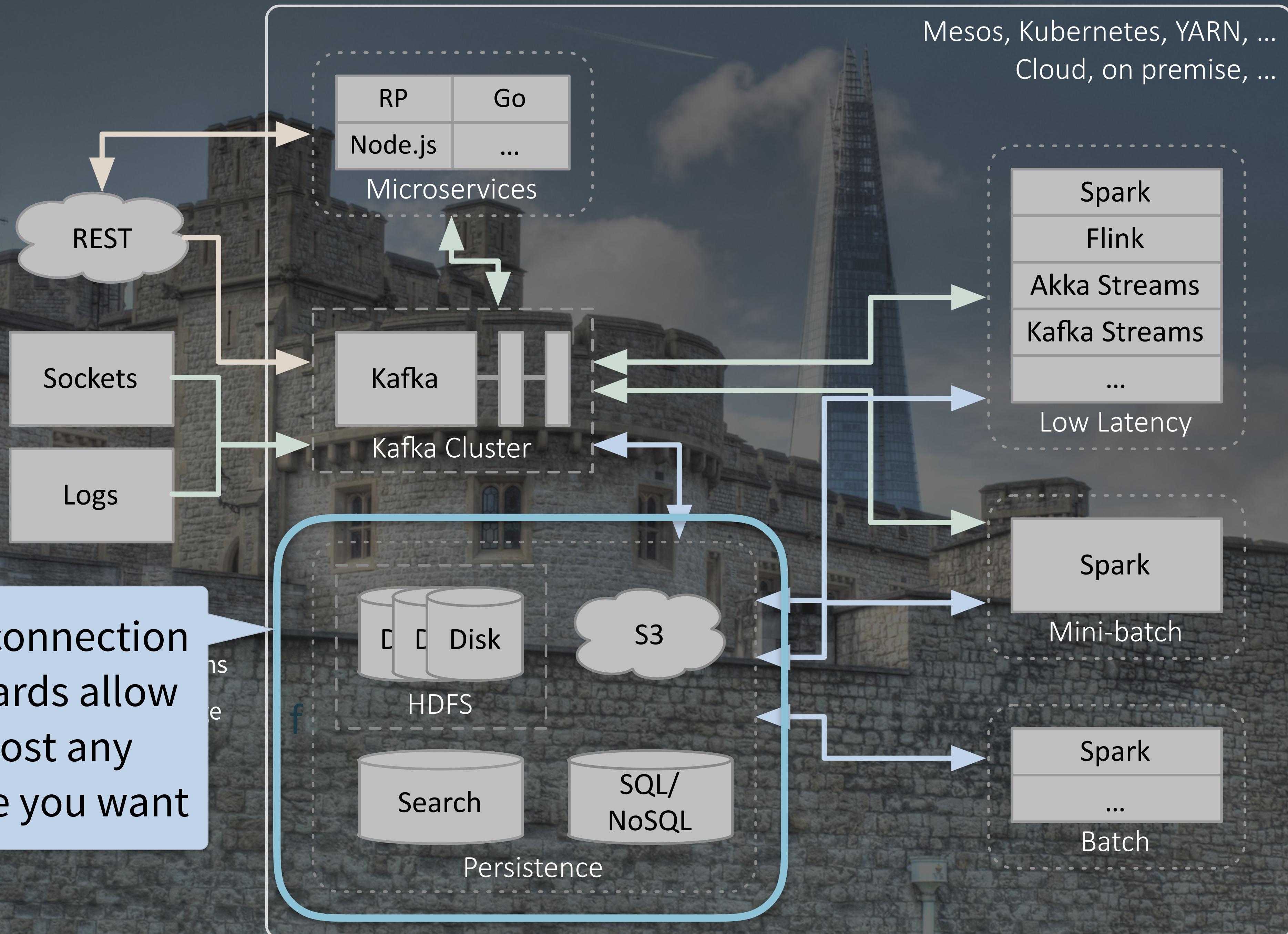
Mini-batch

Spark

...

Batch





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

Spark  
Flink  
Akka Streams  
Kafka Streams  
...

Low Latency

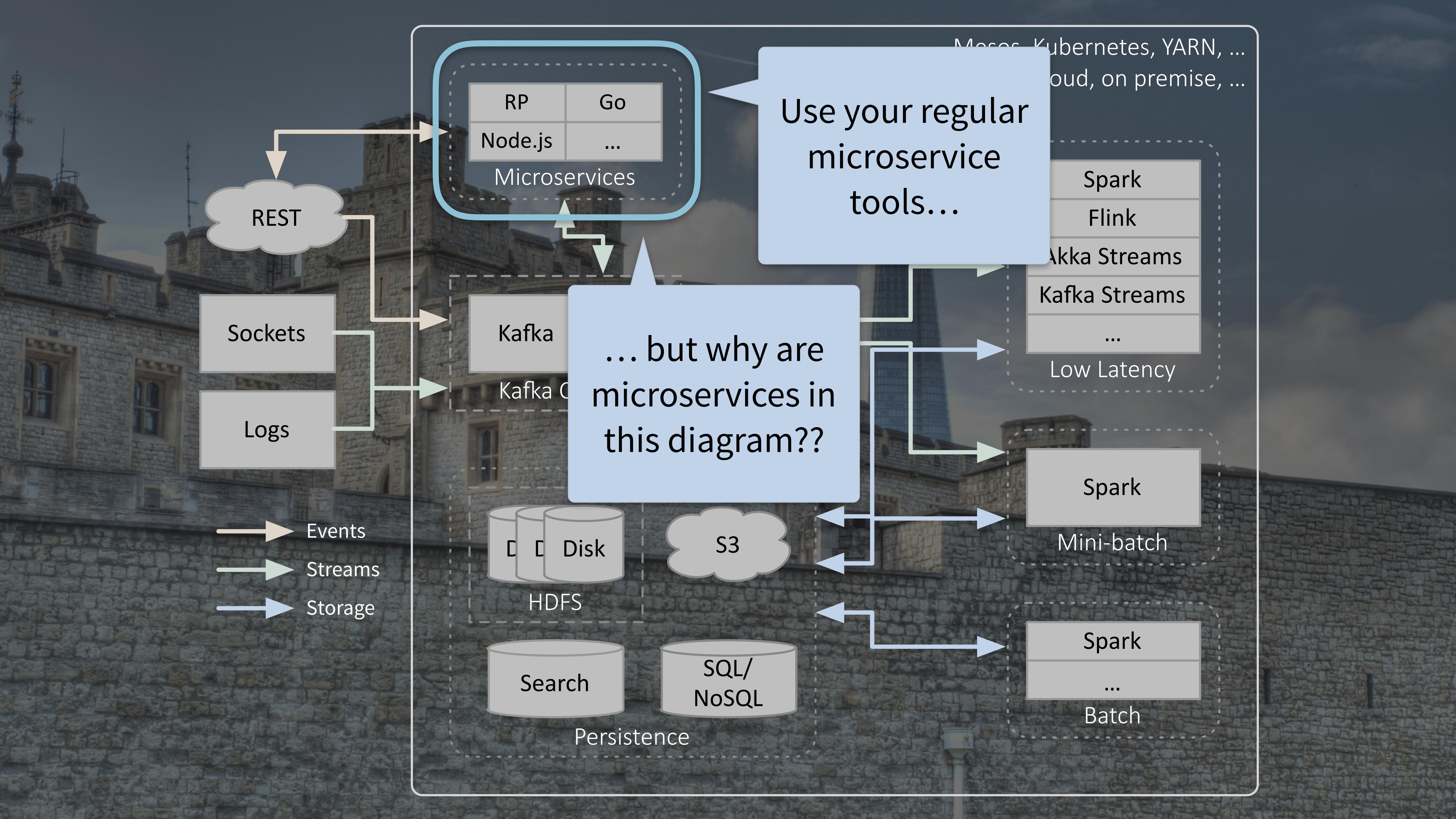
Spark

Mini-batch

Spark

Batch

Open connection  
standards allow  
almost any  
storage you want



# Why Microservices in Fast Data?

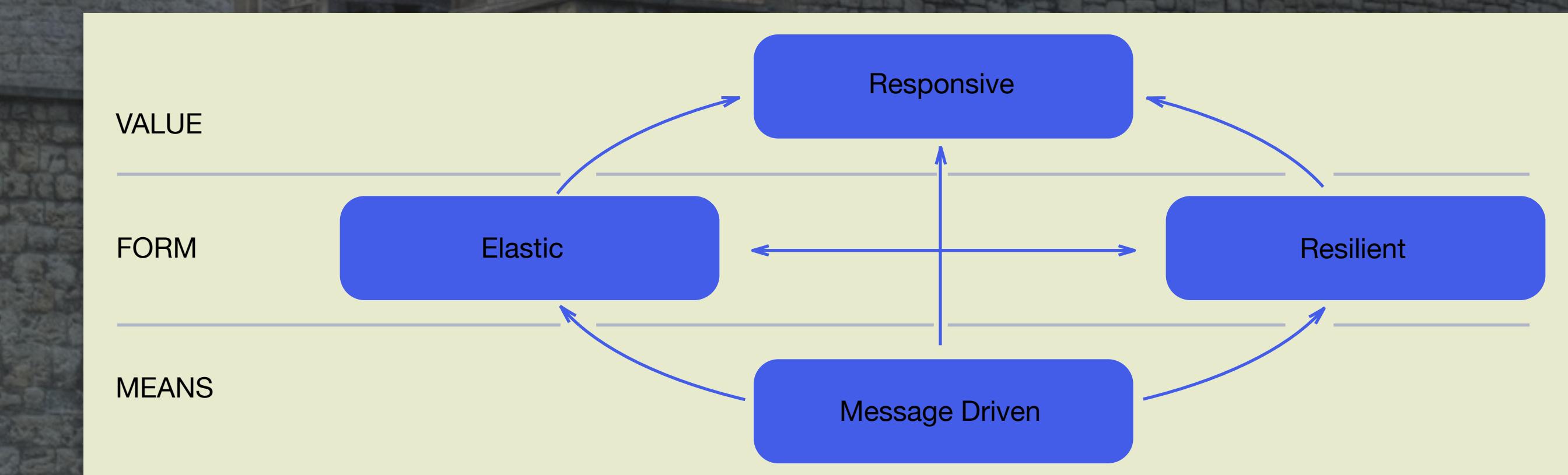
- The trend is to run everything in big clusters using Kubernetes or Mesos
  - In the cloud or on-premise

# Why Microservices in Fast Data?

- If streaming gives you information faster...
  - ... you'll want quick access to it in your other services!

# Why Microservices in Fast Data?

- 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?

- This leads to our last major point...

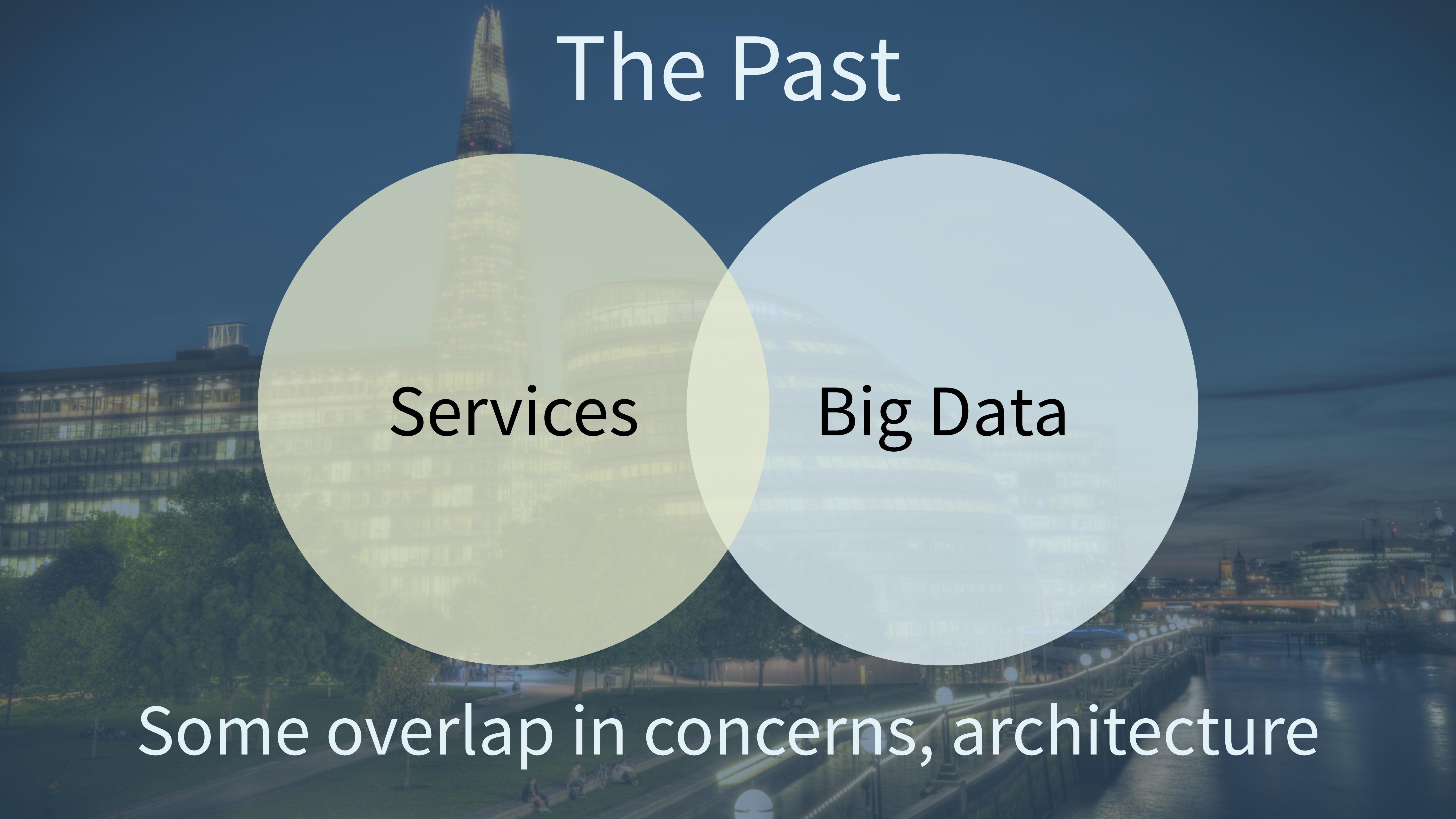
# Organizational Impact



# Organizational Impact

- 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

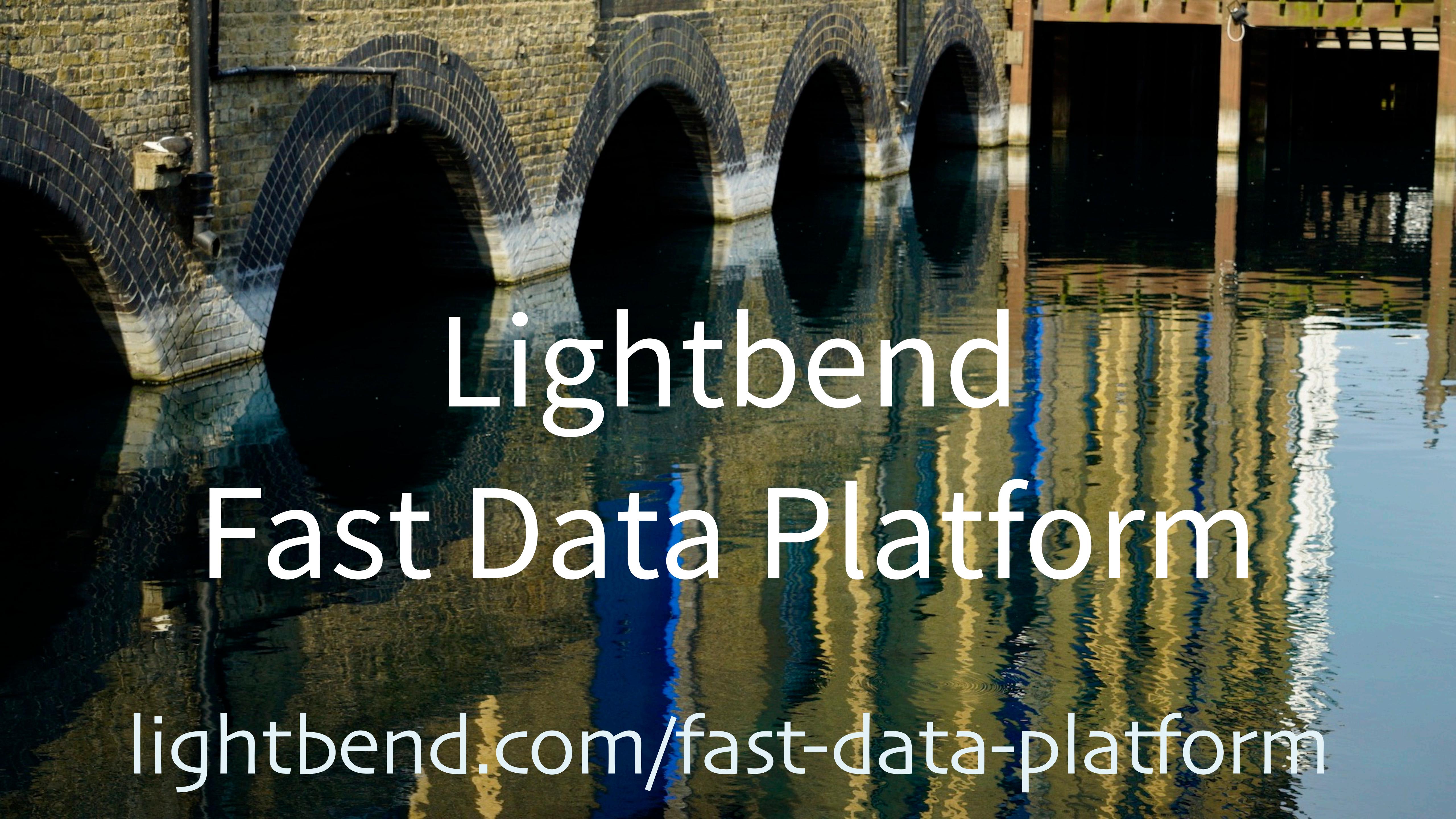
Much more overlap

# The Future?



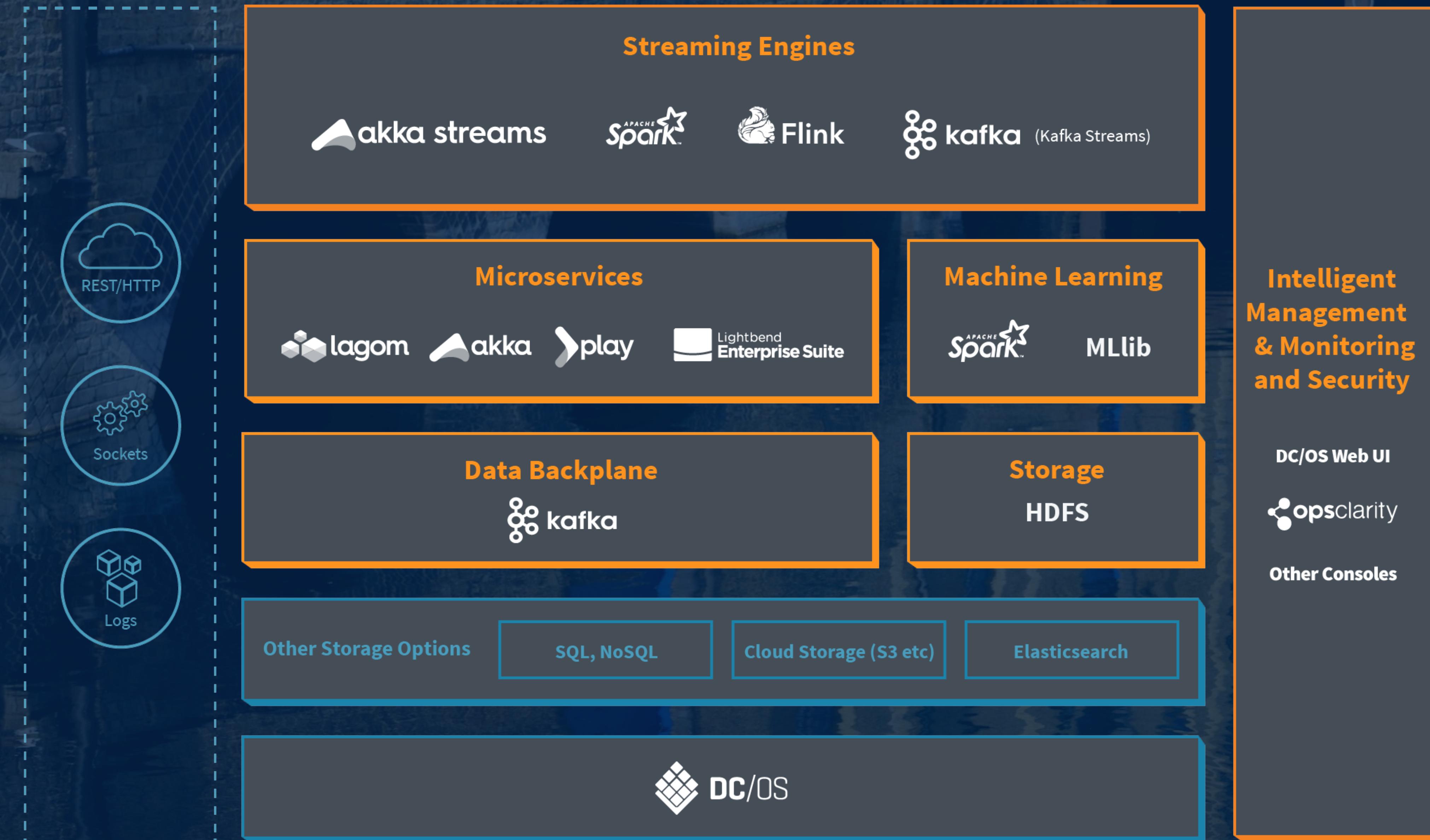
Microservices  
*for* Fast Data

Much more microservice focused?

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)



## Streaming Engines



## Microservices



## Machine Learning



## Data Backplane



## Storage



## Other Storage Options

SQL, NoSQL

Cloud Storage (S3 etc)

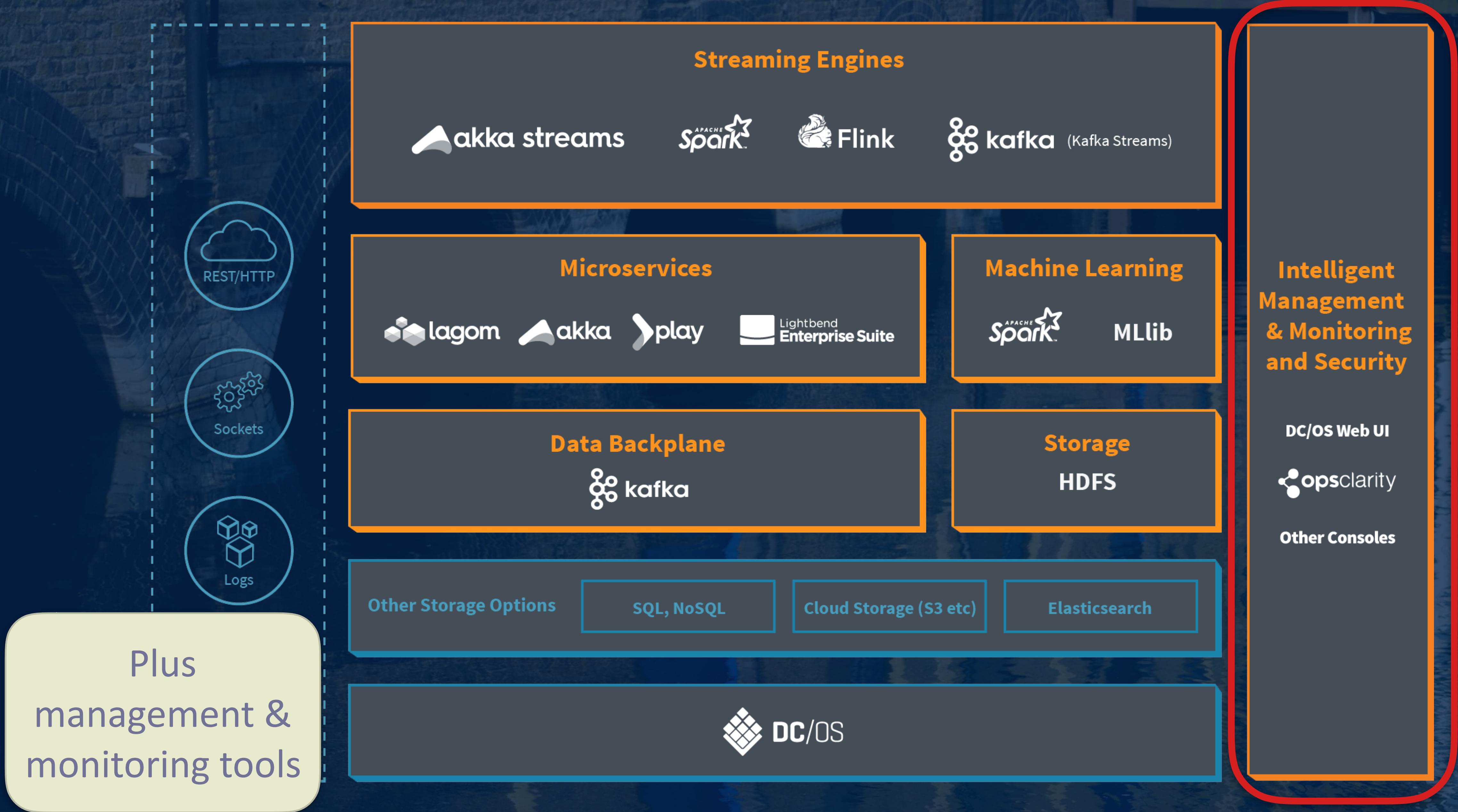
Elasticsearch

Intelligent  
Management  
& Monitoring  
and Security

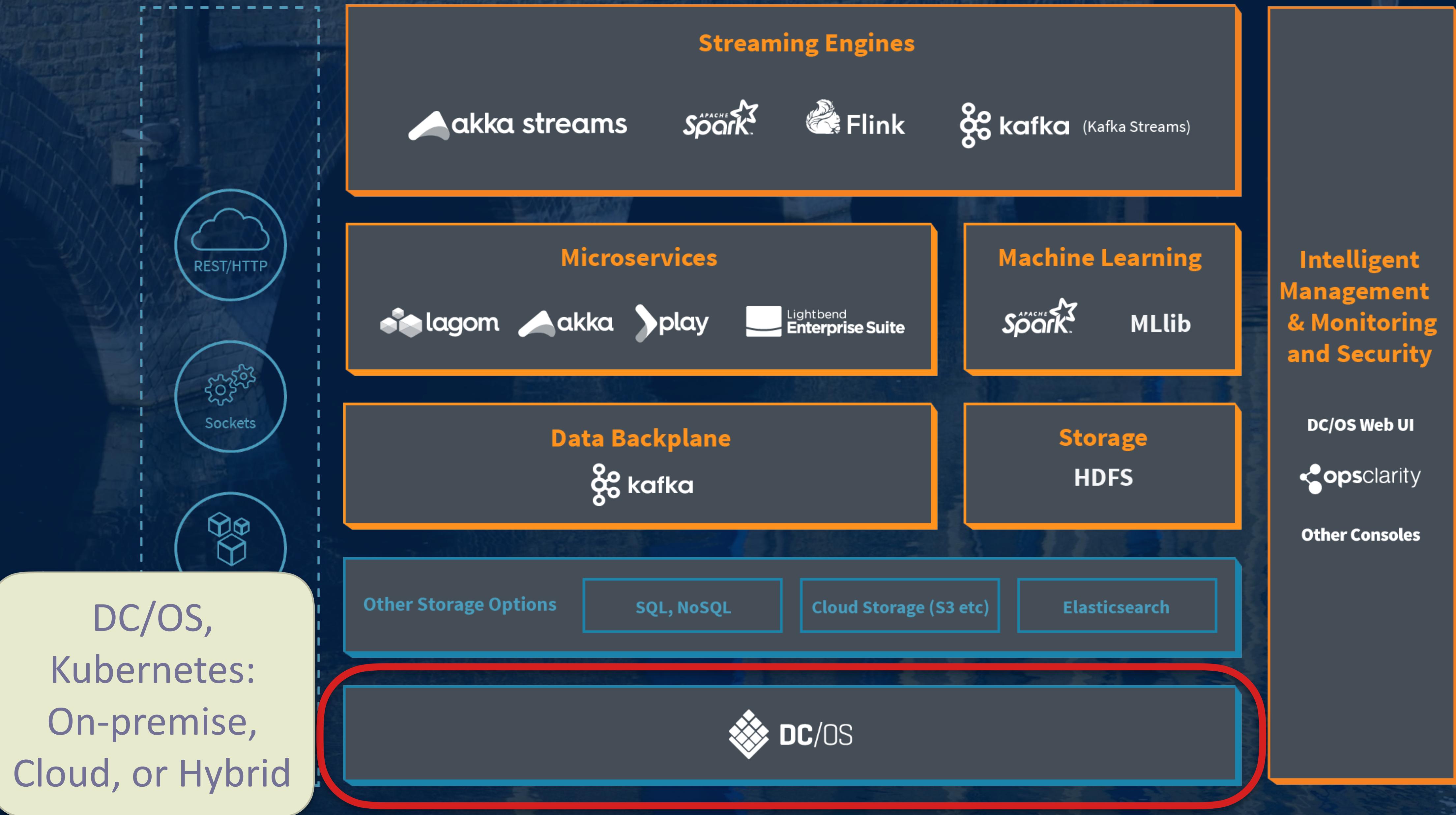
DC/os Web UI

opsclarity

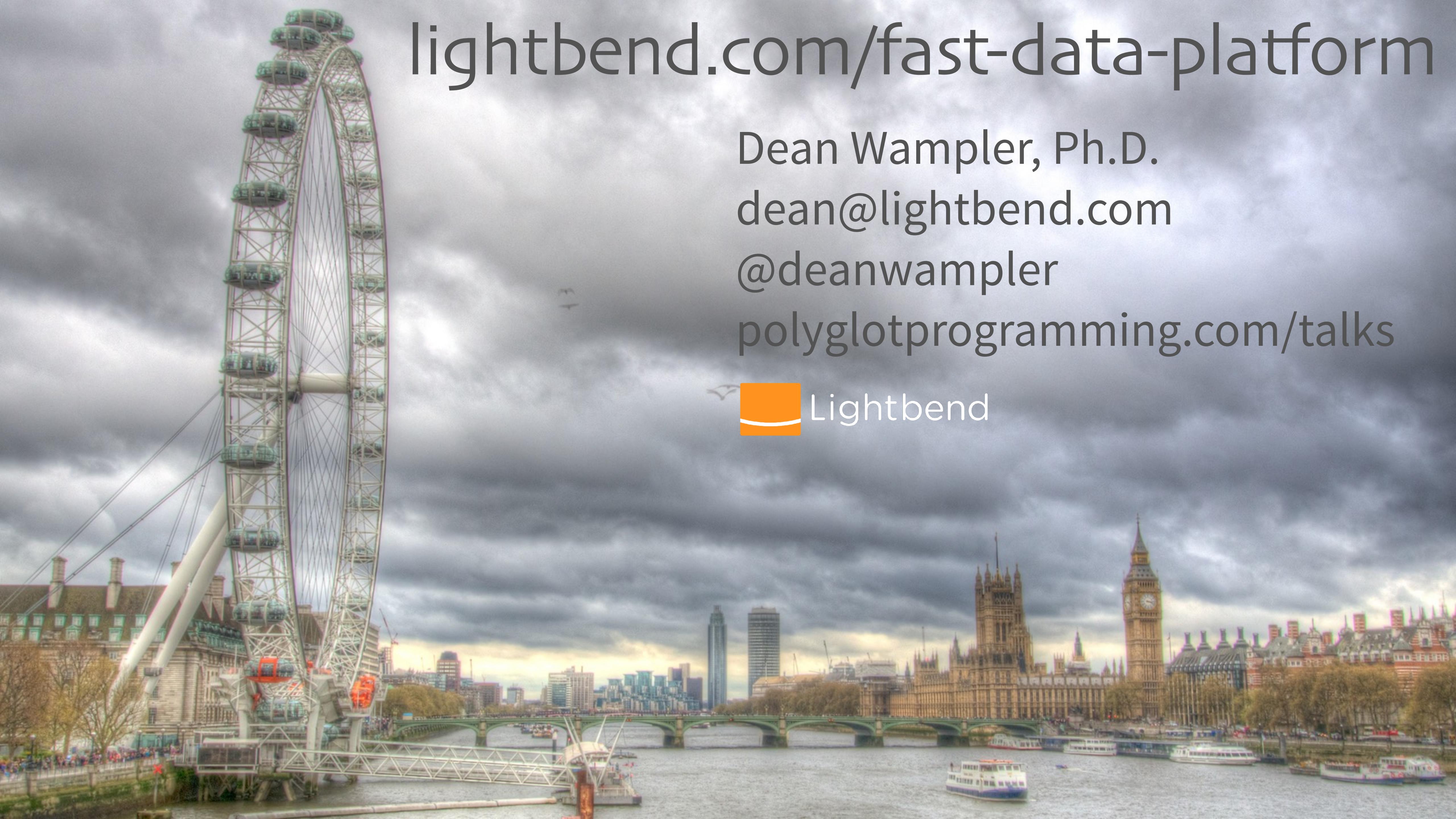
Other Consoles



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



[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)

