Executive Briefing: What You Need to Know about Fast Data

Dean Wampler, Ph.D.
dean@lightbend.com
@deanwampler
polyglotprogramming.com/talks

Lightbend

Based on this report

lightbend.com/fast-data-platform

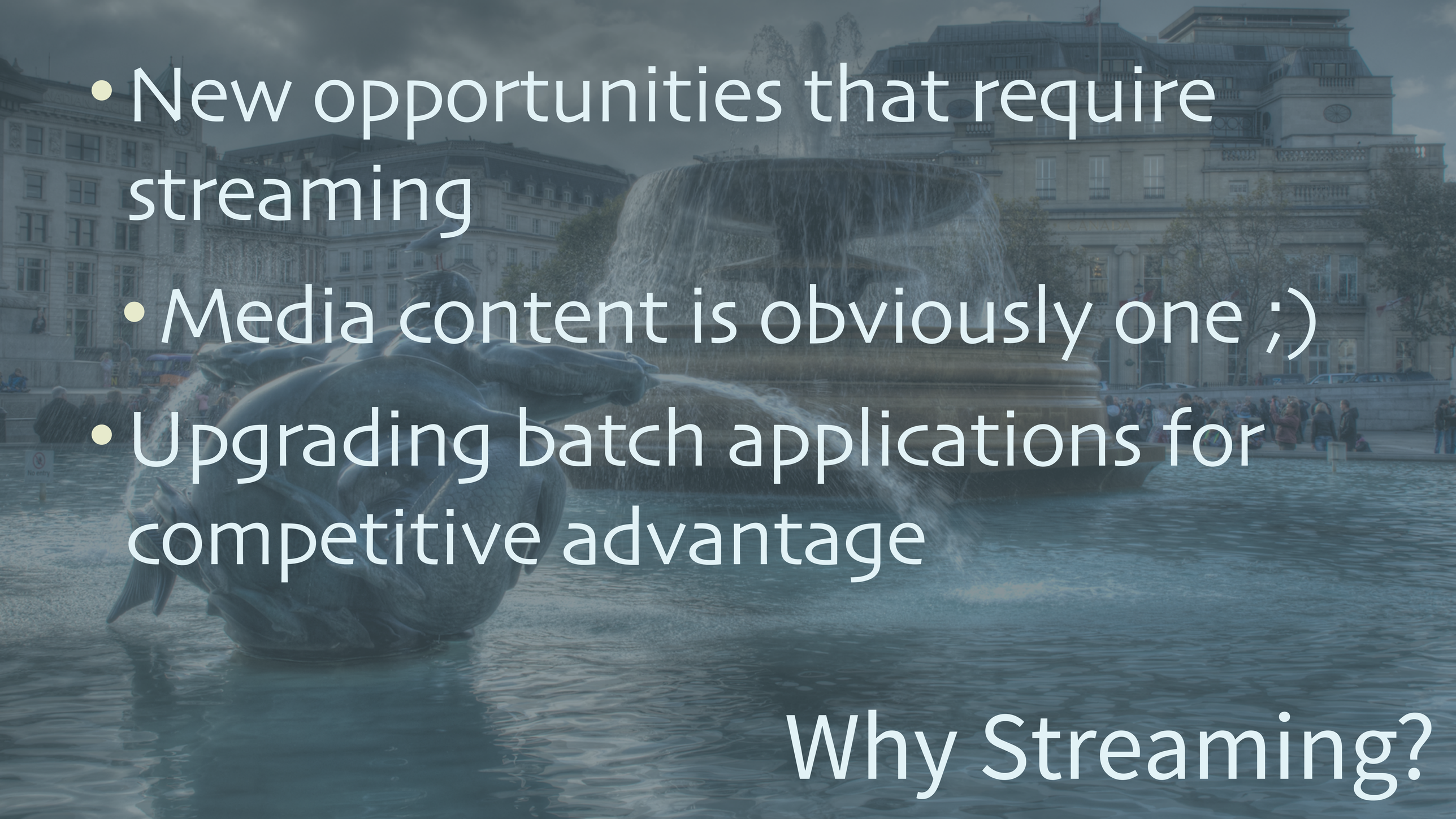
2nd edition coming in October!

What We'll Discuss

What We'll Discuss

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

Why Streaming?



- New opportunities that require streaming
  - Media content is obviously one ;)
- Upgrading batch applications for competitive advantage

Why Streaming?

Fast Data Use Cases

Similar IoT Architectures

Predictive Analytics	IoT	Real-time Personalization	Real-time Financial Processes
Apply ML models to large volumes of device data to pre-empt failures / outages	Real-time consumer and industrial Device and Supply Chain management at scale	Real-time marketing based on behavior, location, inventory levels, product promotions, etc.	Drive better business outcomes through real-time risk, fraud detection, compliance, audit, governance, etc.
Hewlett Packard Enterprise	STARBUCKS COFFEE	RoyalCaribbean INTERNATIONAL	CapitalOne

https://www.lightbend.com/customers

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.

Anomaly Handler

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

Probable Anomalies

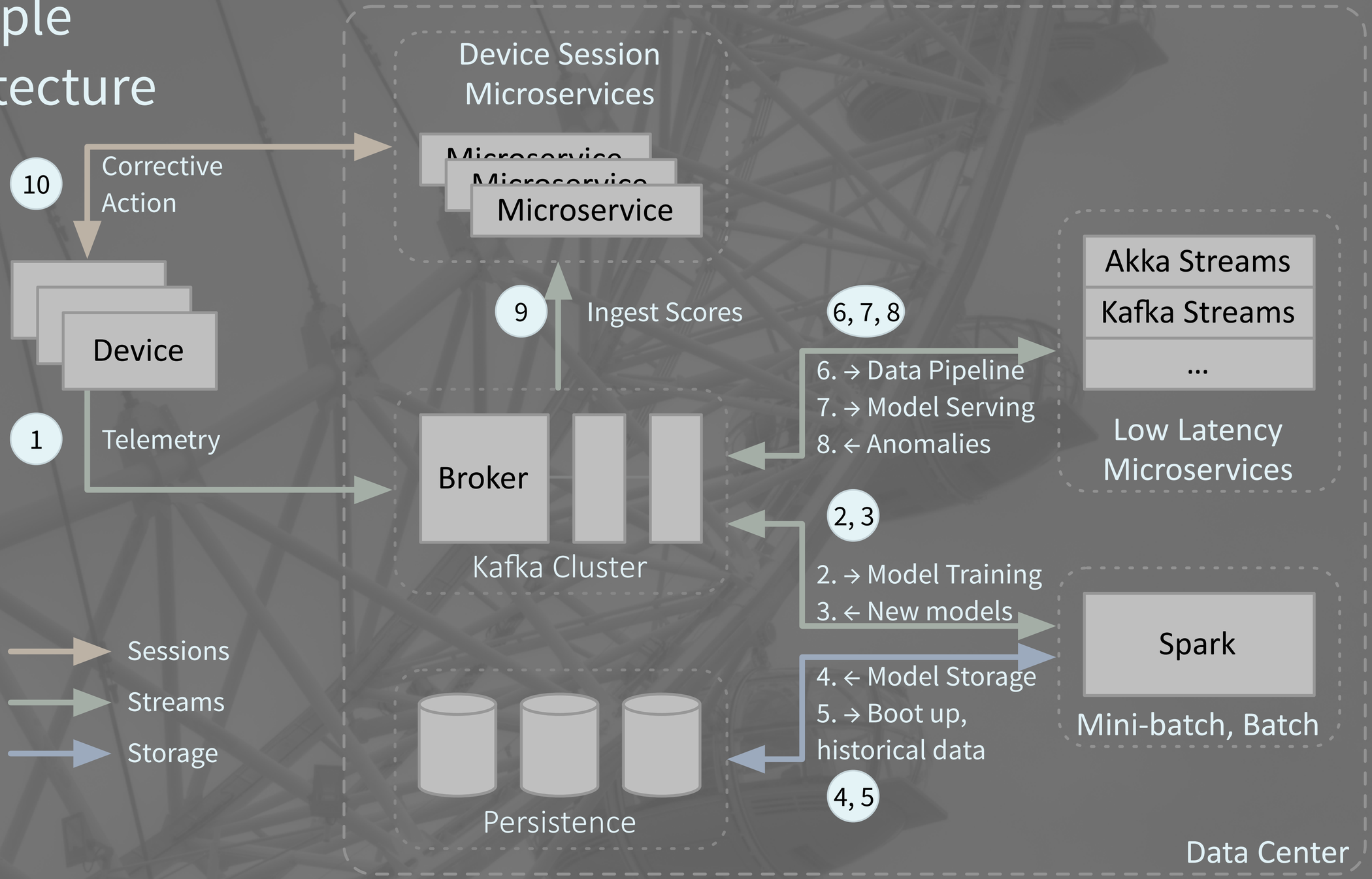
Anomaly Detection: Model

Corrective Actions

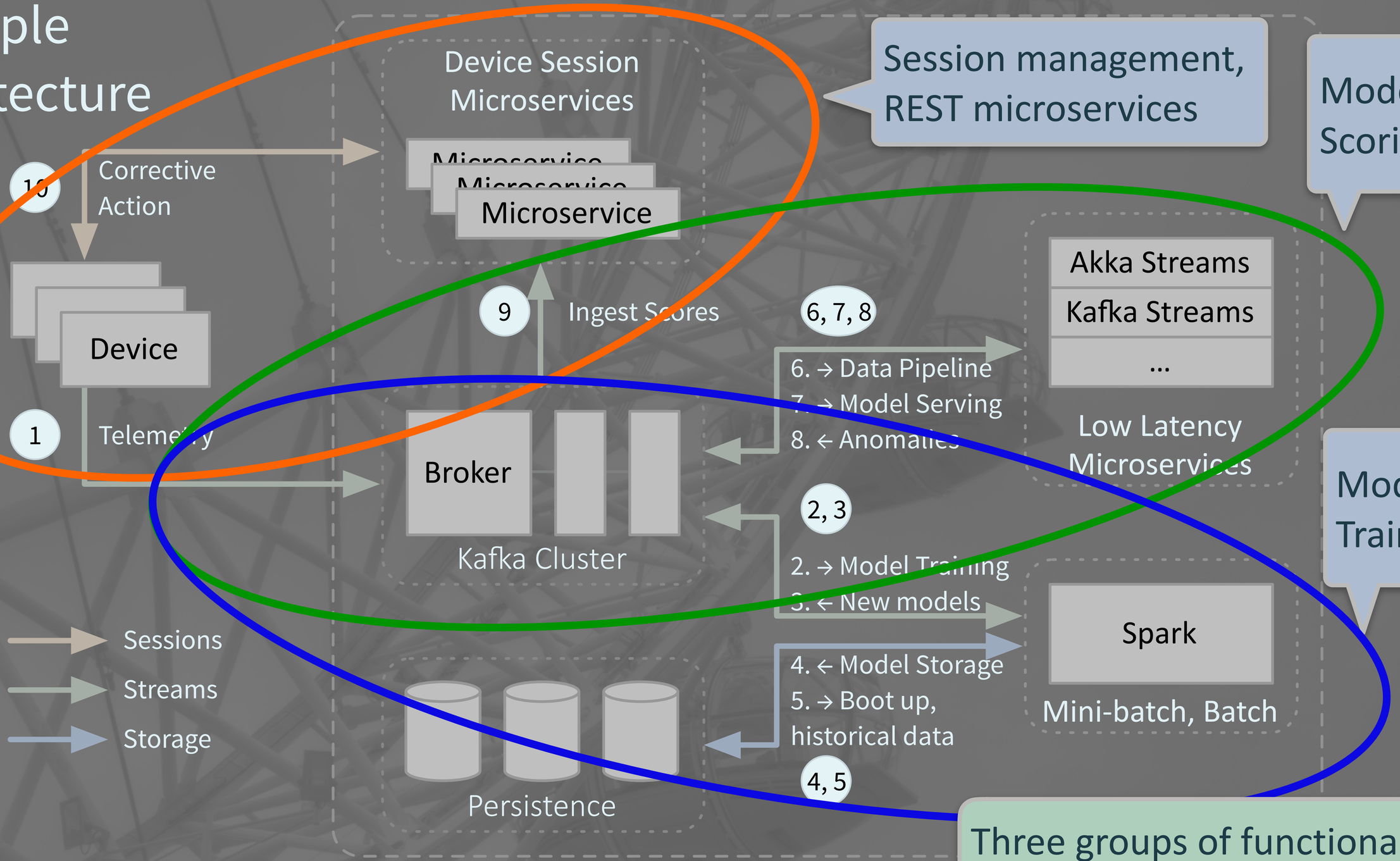
Telemetry Records

Ingest telemetry from edge devices.

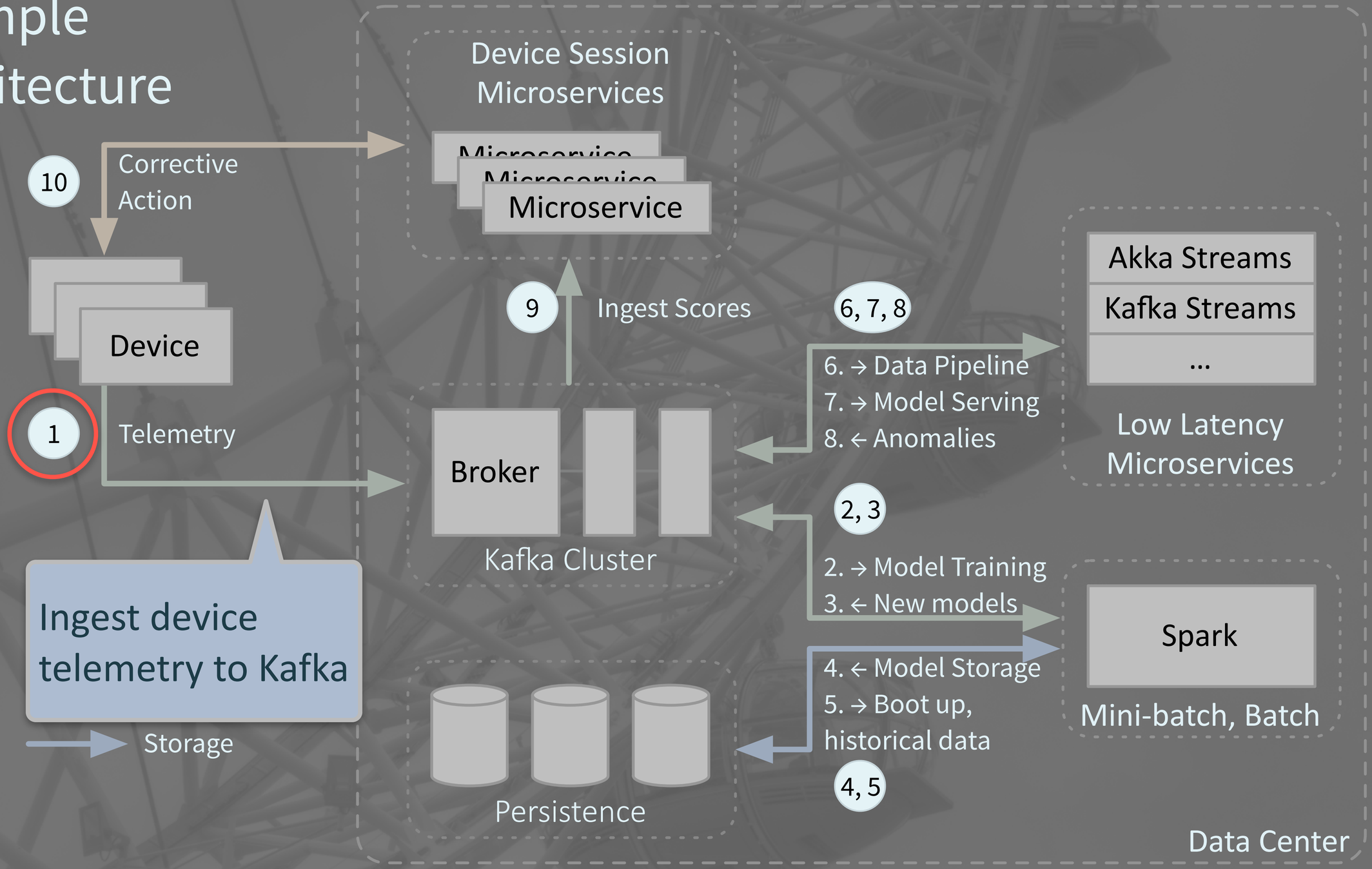
Example Architecture



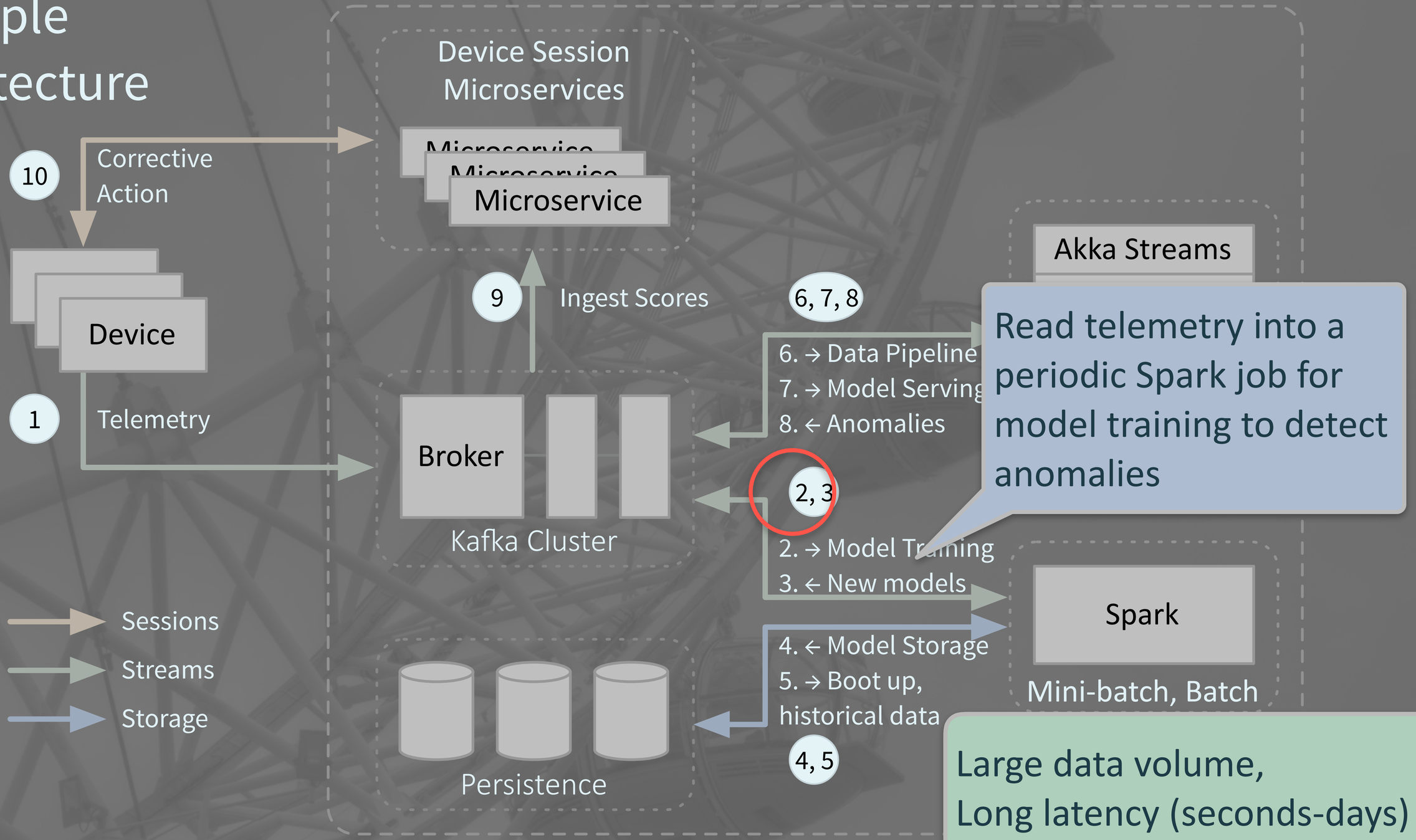
Example Architecture



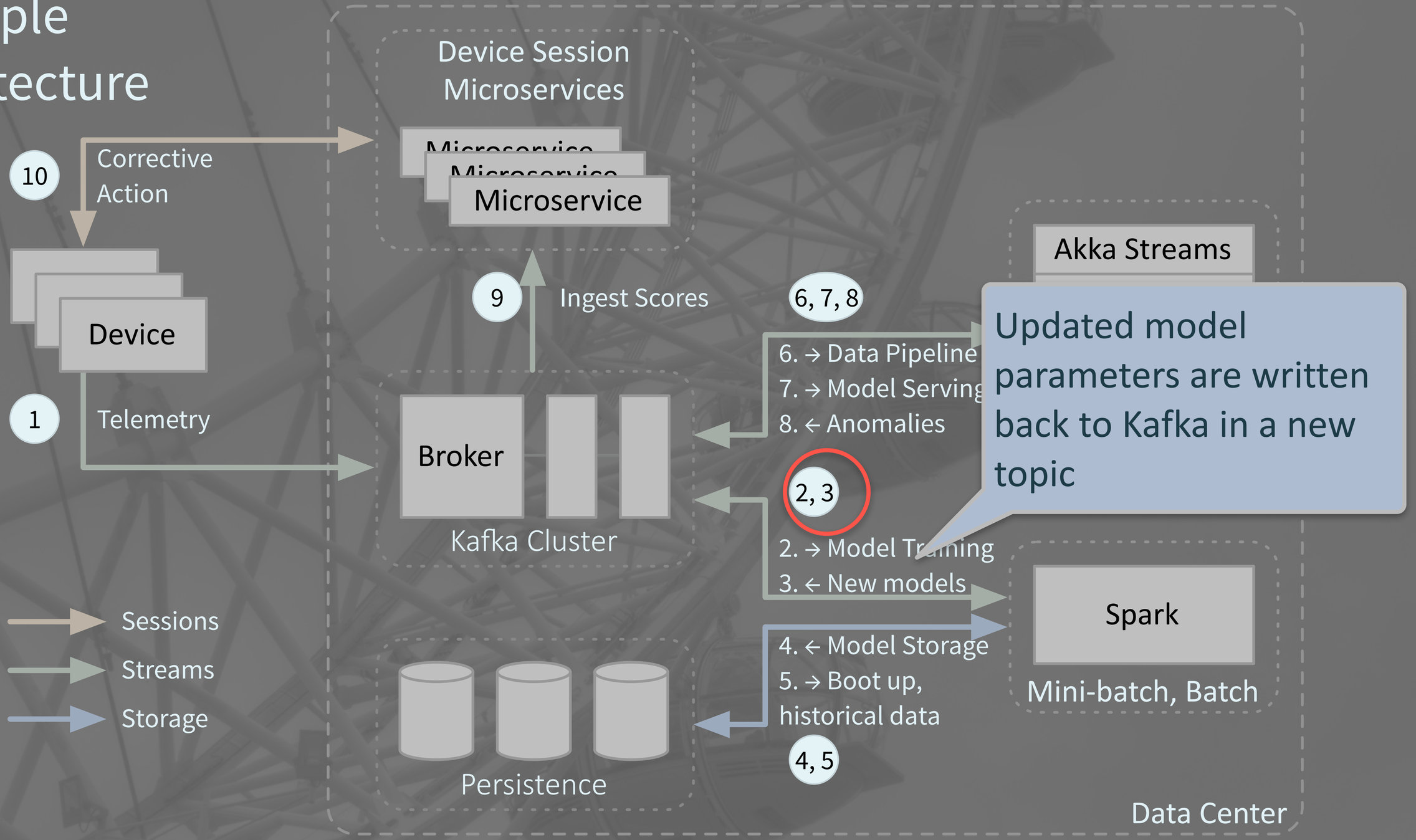
Example Architecture



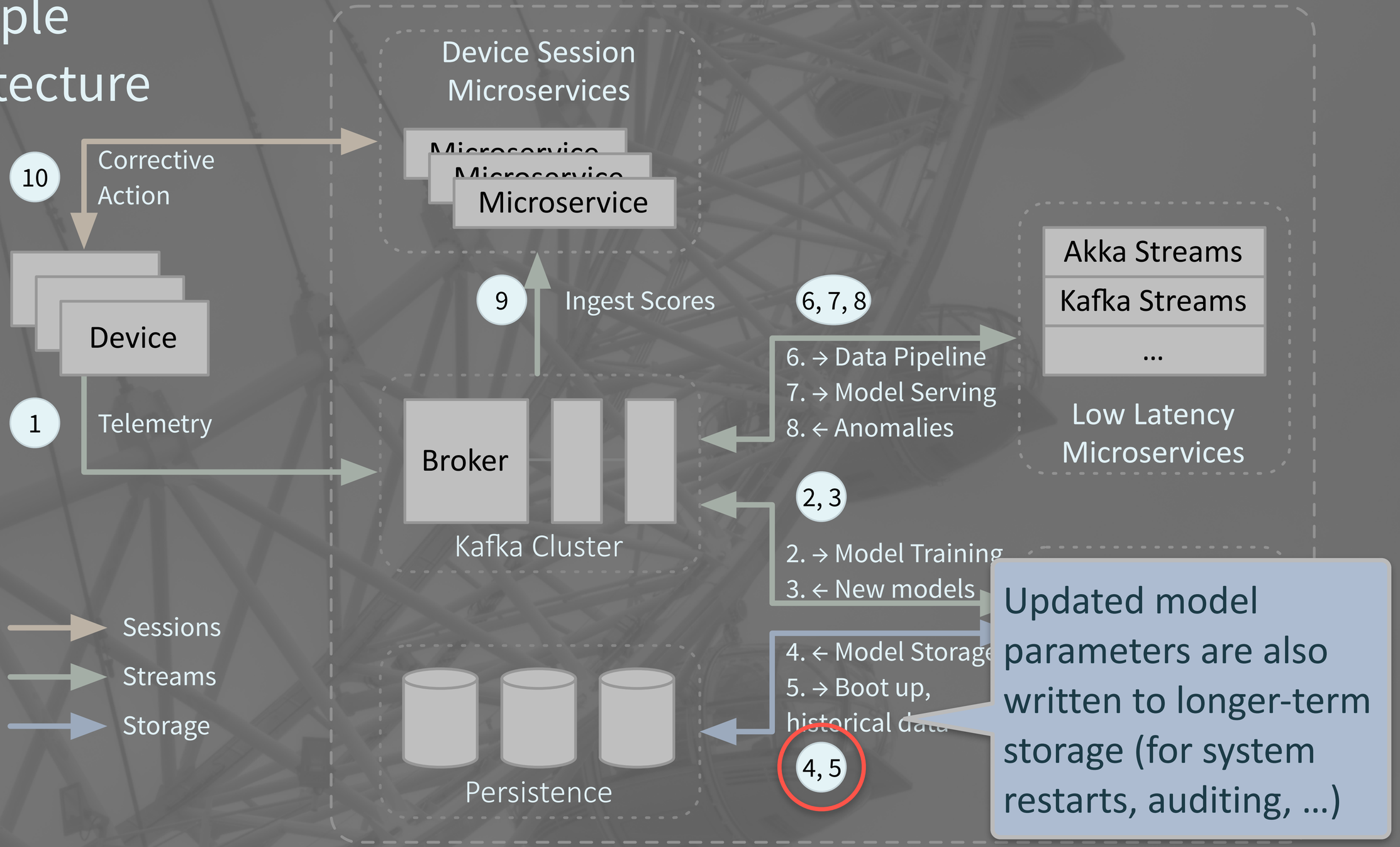
Example Architecture



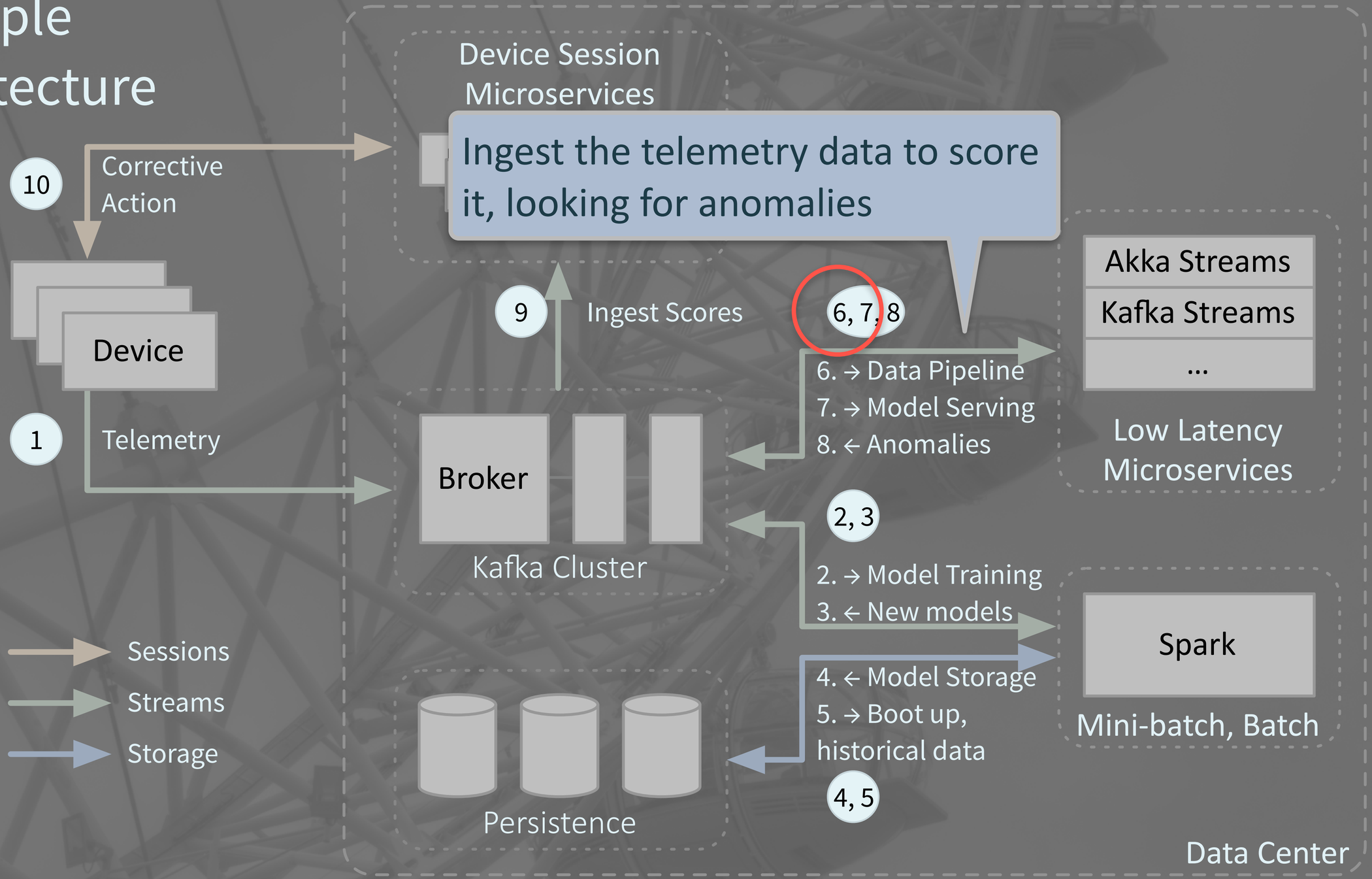
Example Architecture



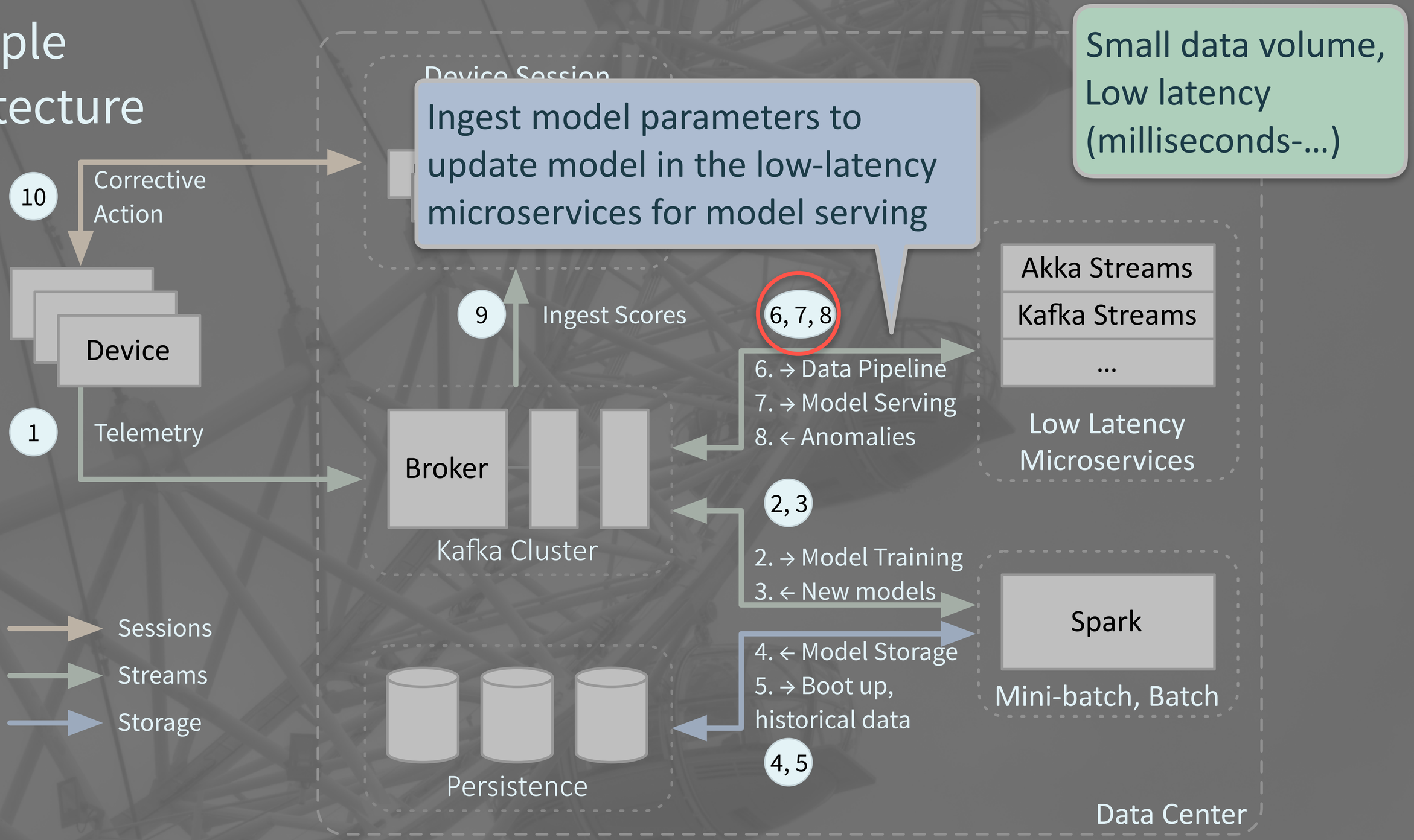
Example Architecture



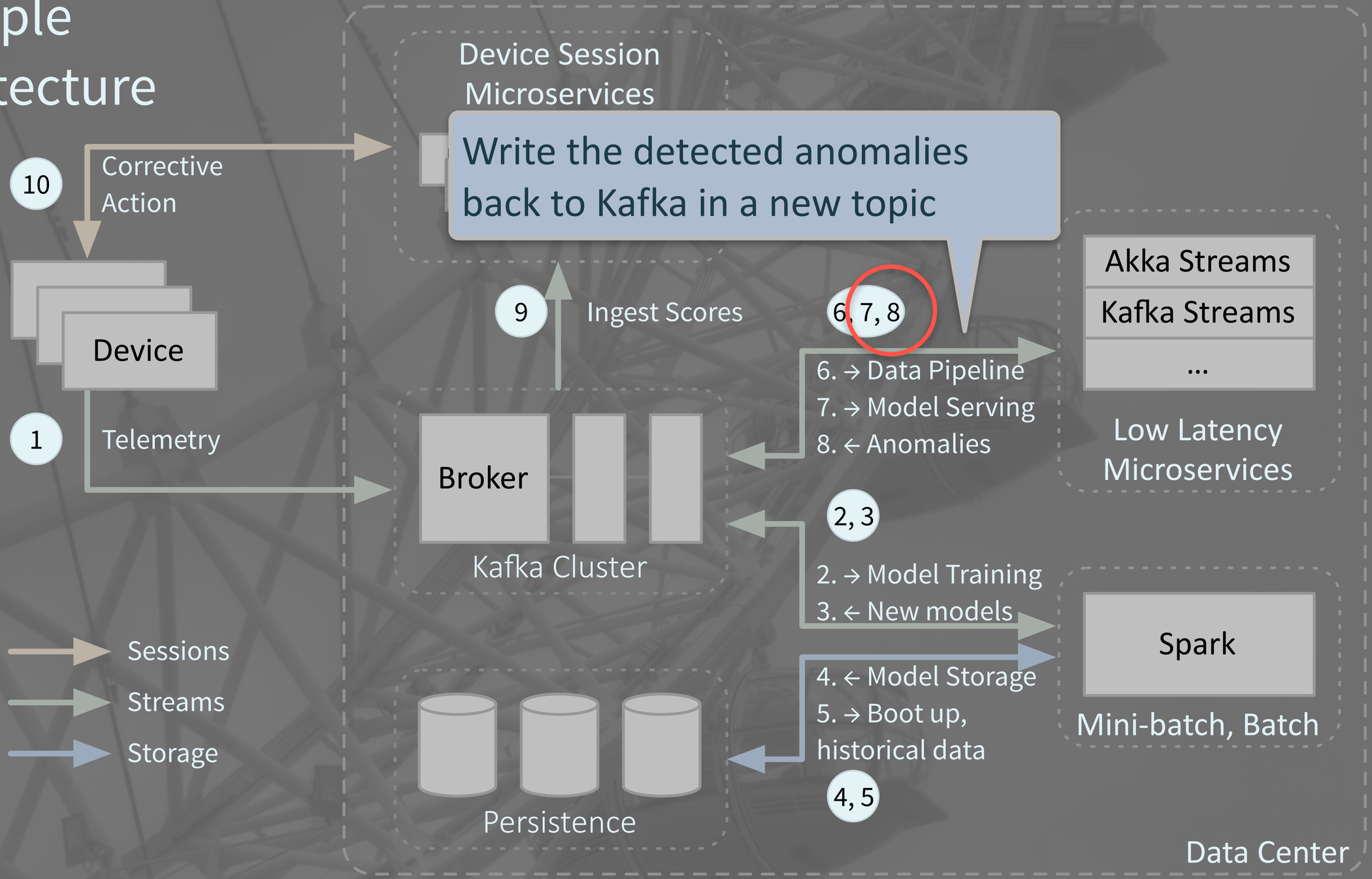
Example Architecture



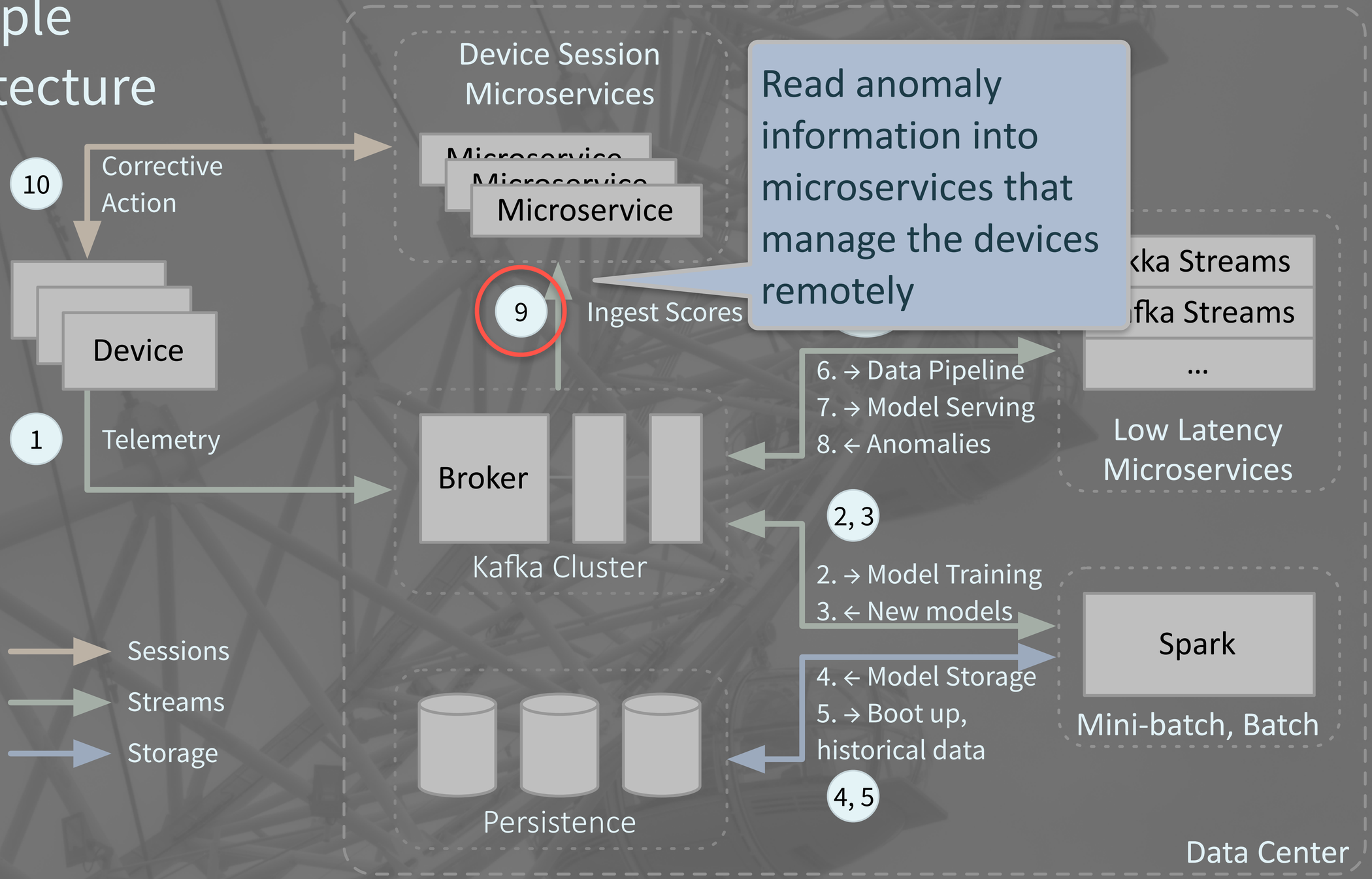
Example Architecture



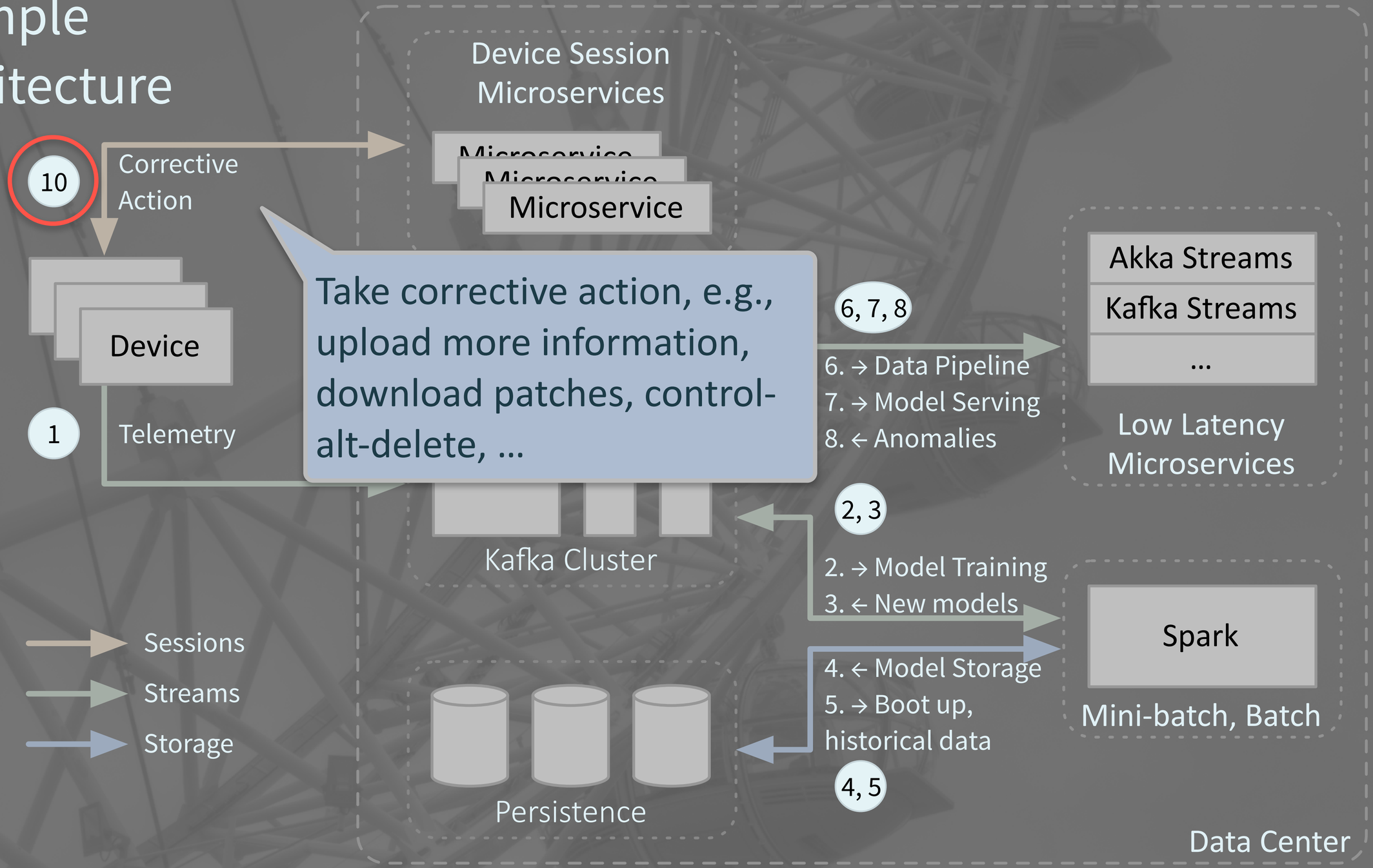
Example Architecture



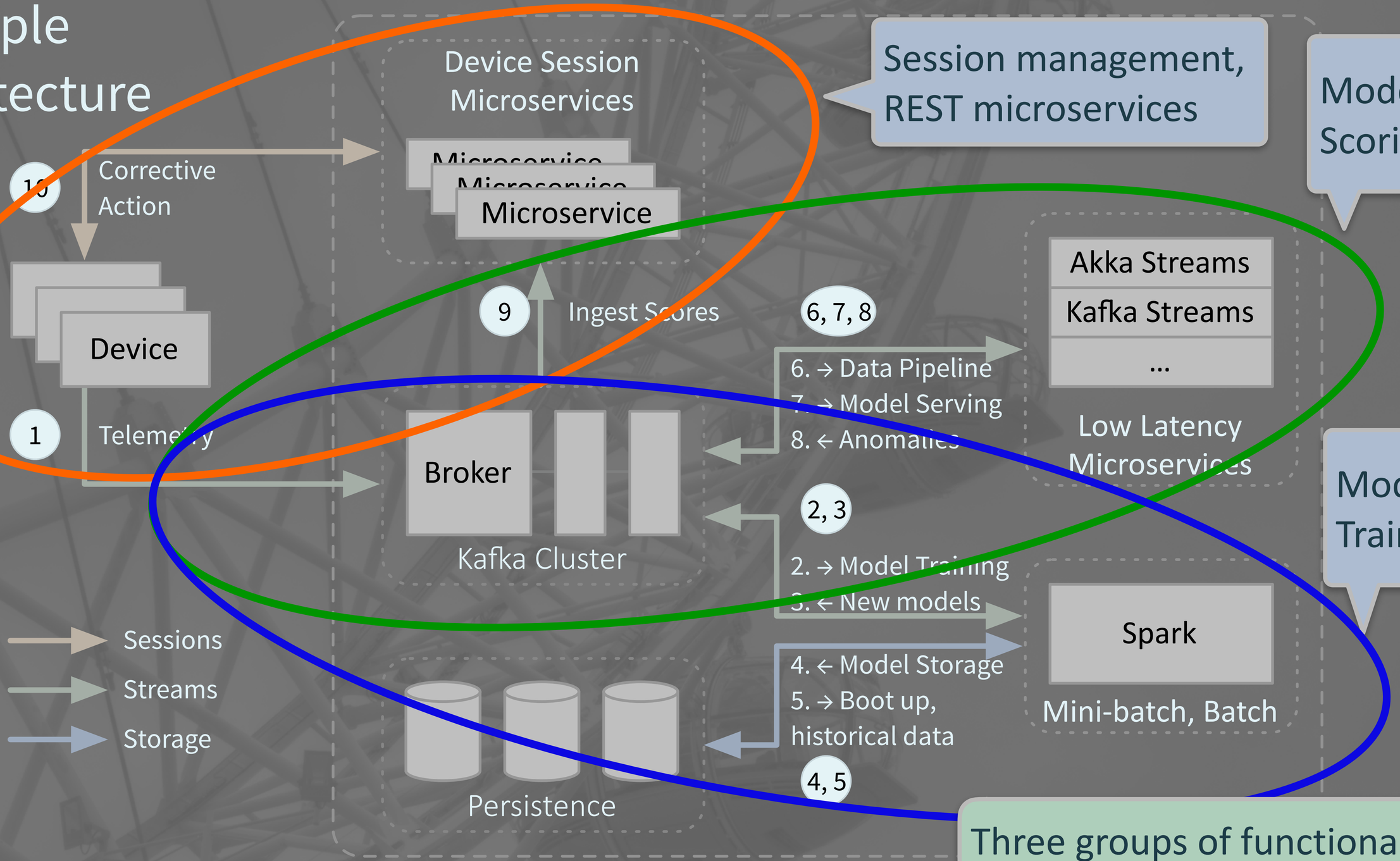
Example Architecture



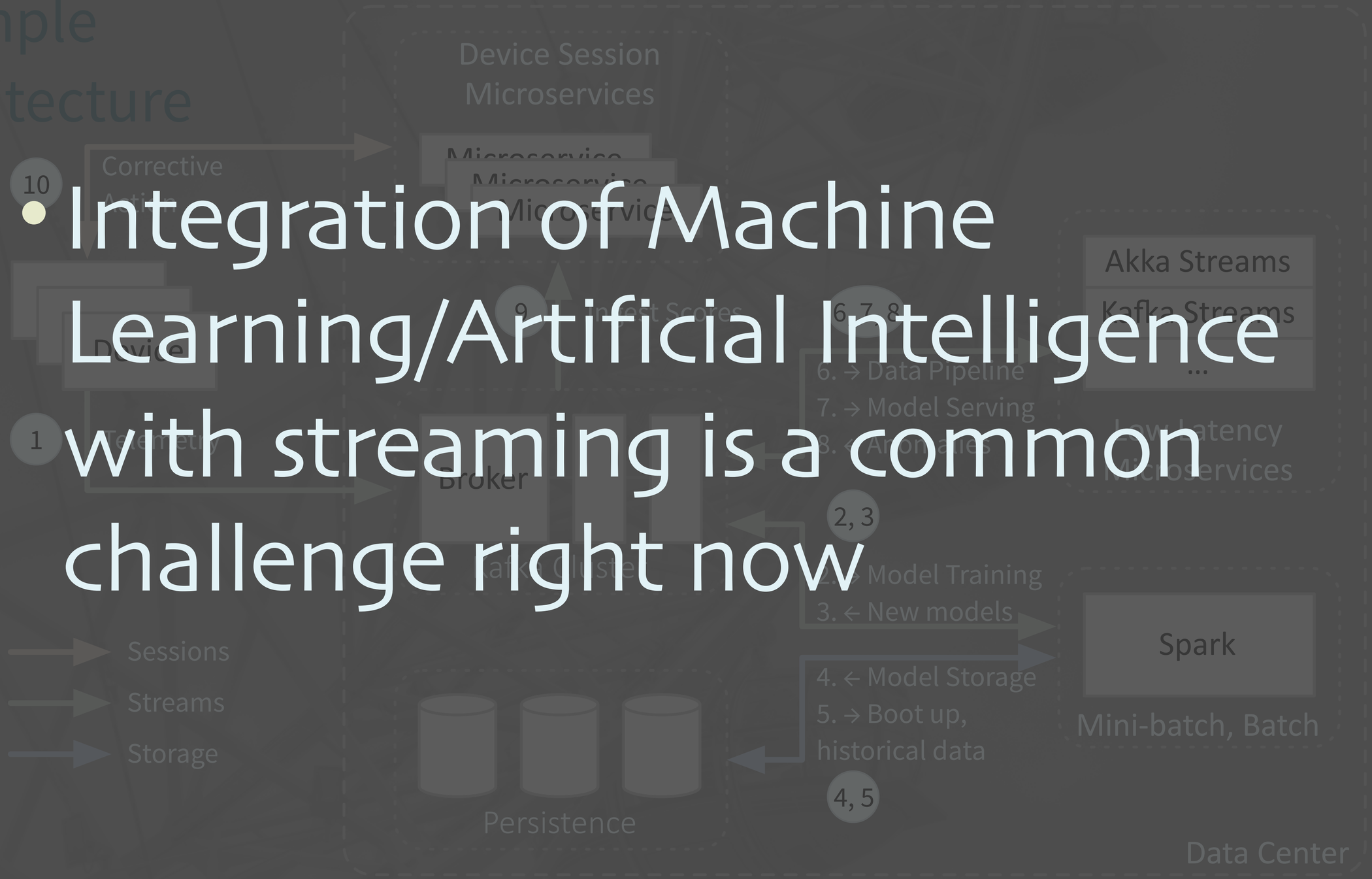
Example Architecture



Example Architecture



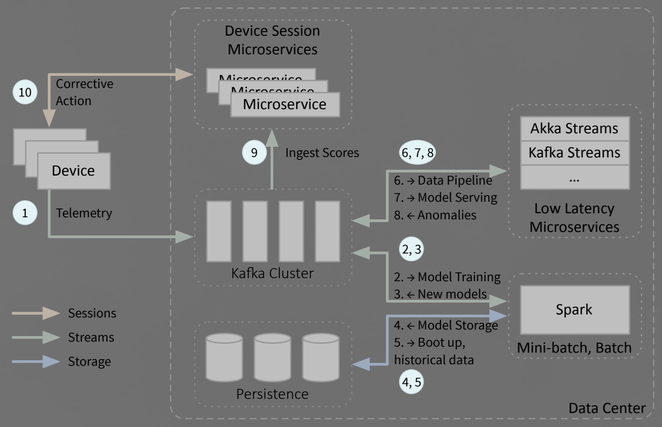
Example Architecture



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

Challenges

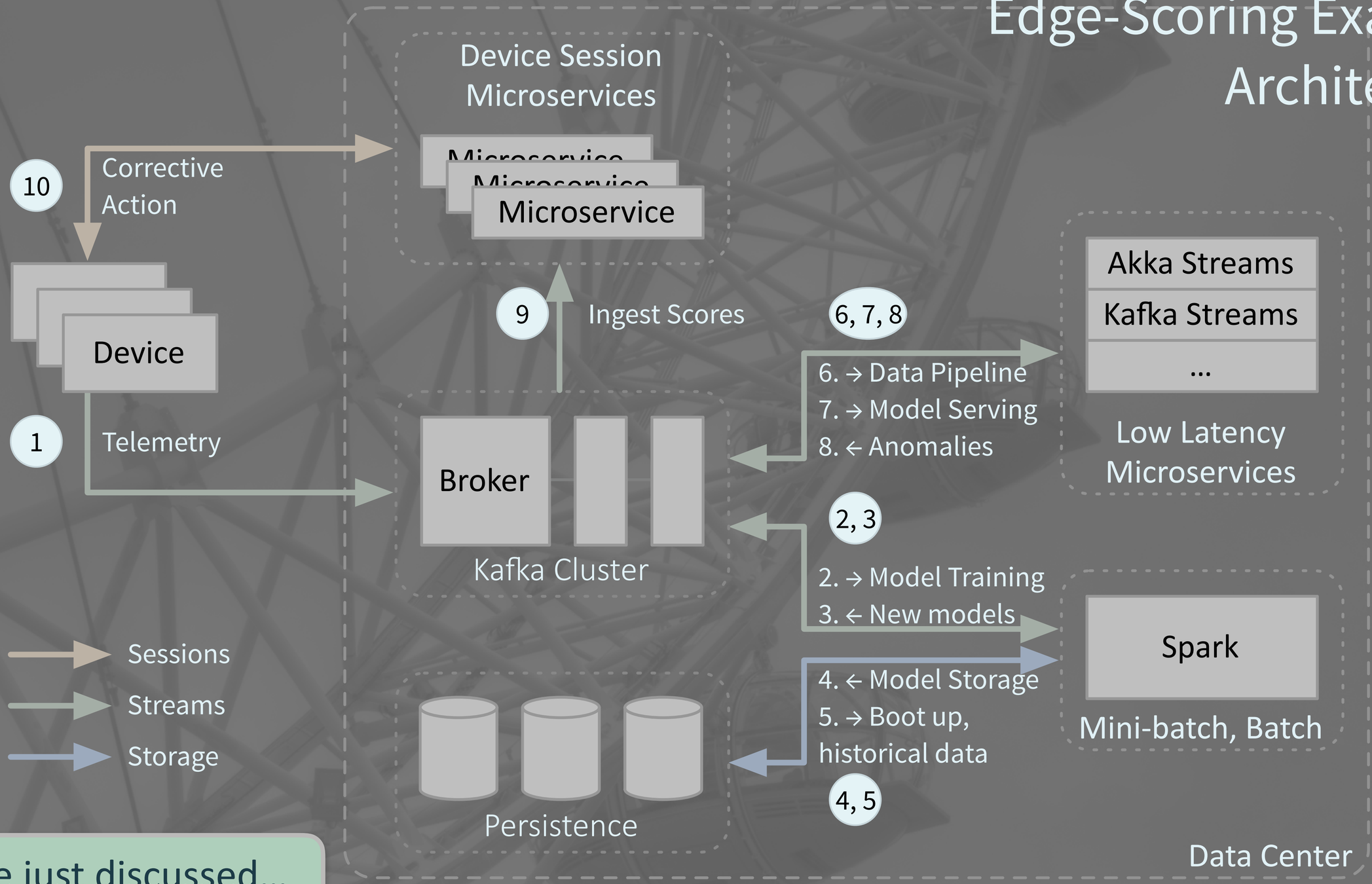
- Network overhead for telemetry ingestion too high?
- Model serving latency too long?
- Datacenter unavailable?
- Idea: Serve models on the device!



Internet of Things

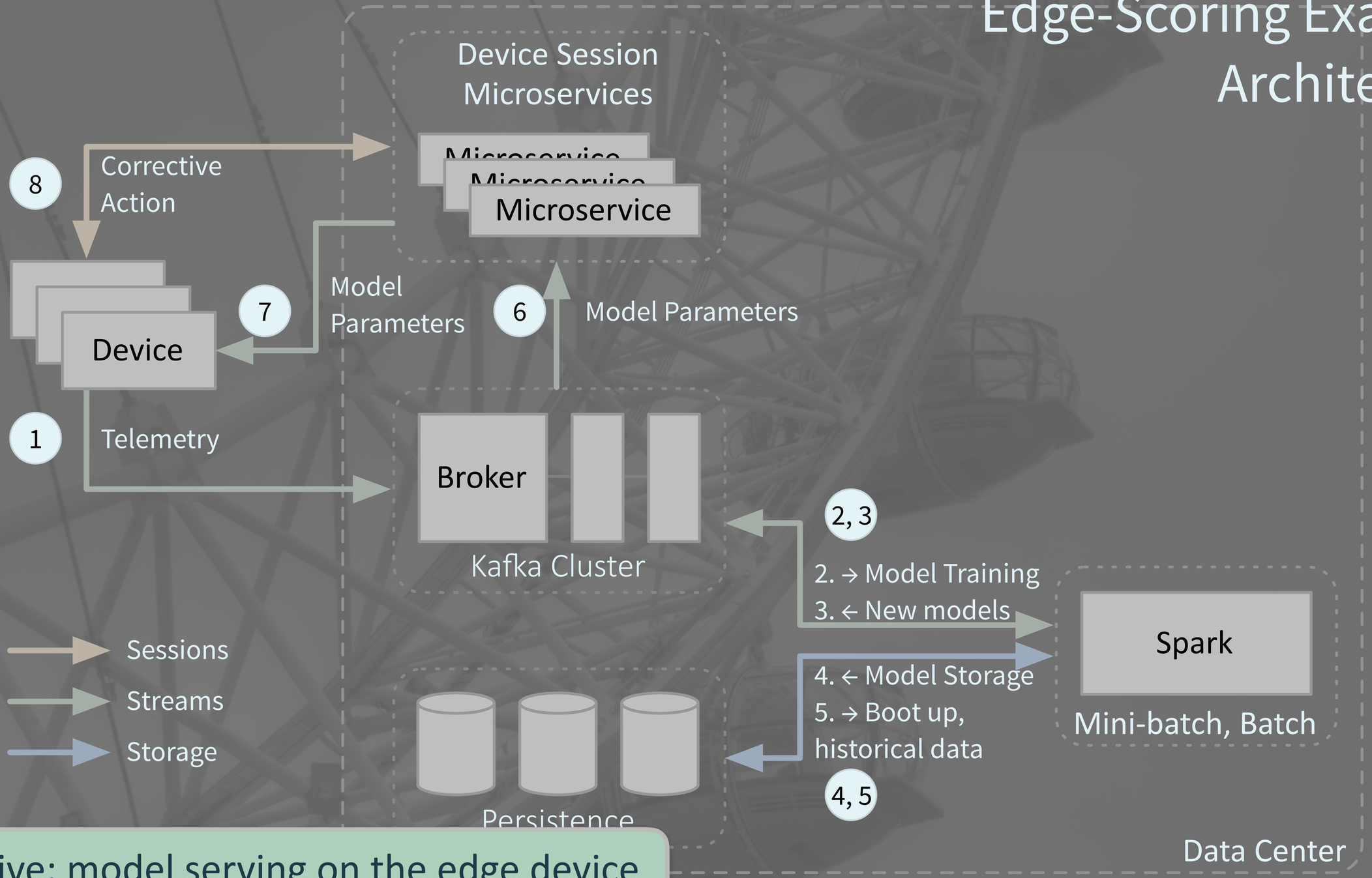
- Real-time consumer and industrial device and supply chain management at scale

Edge-Scoring Example Architecture



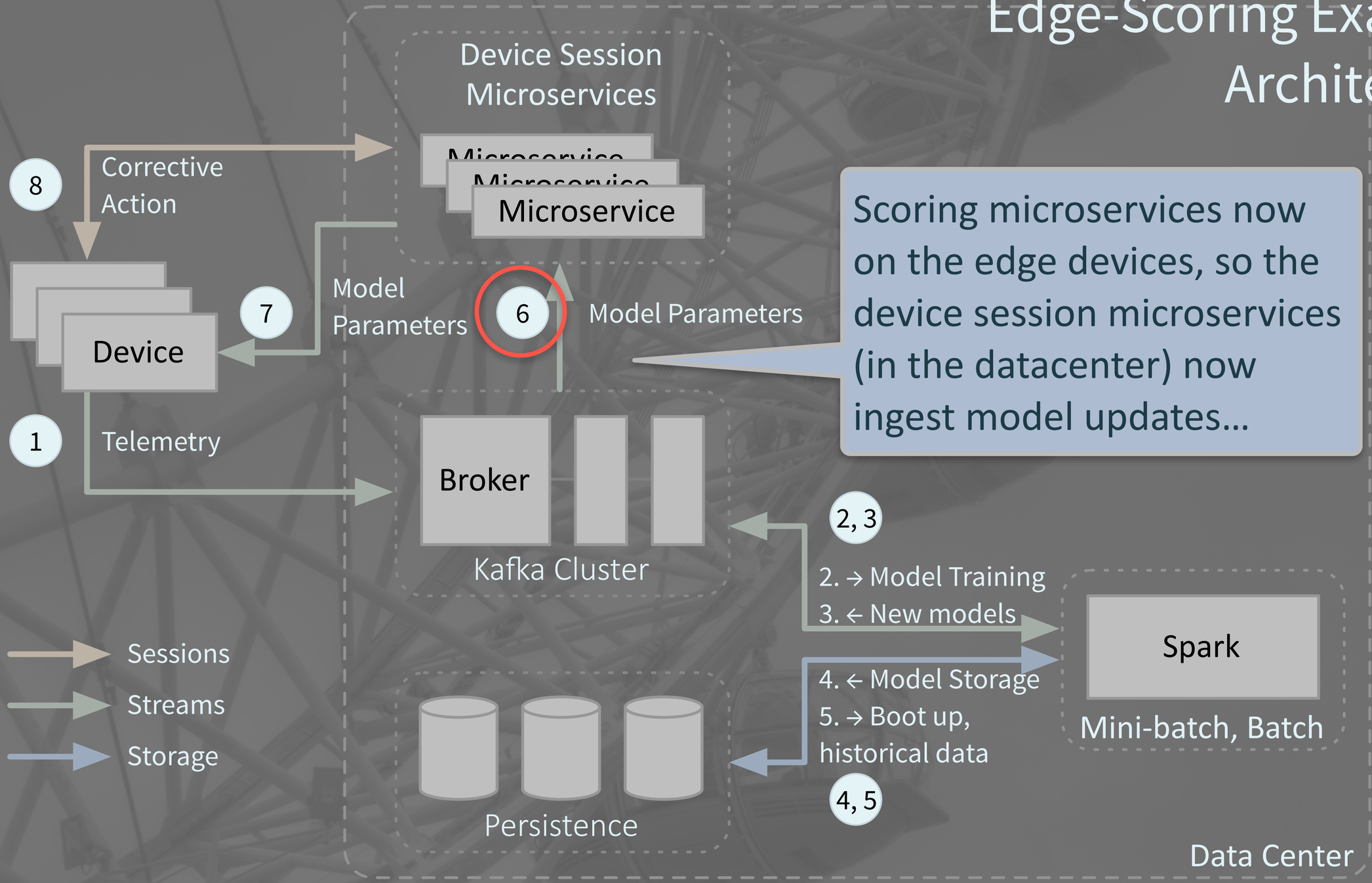
What we just discussed...

Edge-Scoring Example Architecture



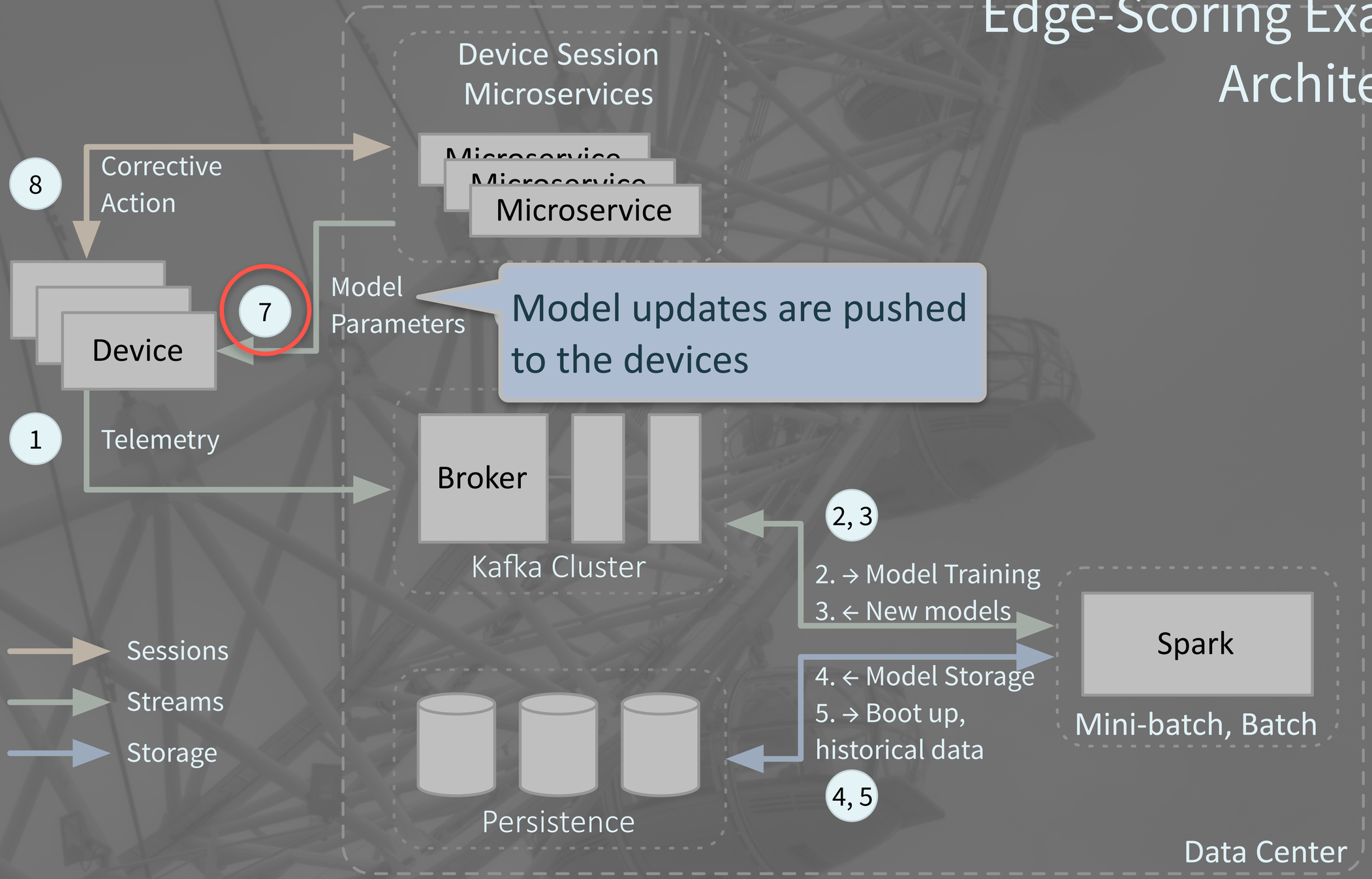
Alternative: model serving on the edge device

Edge-Scoring Example Architecture

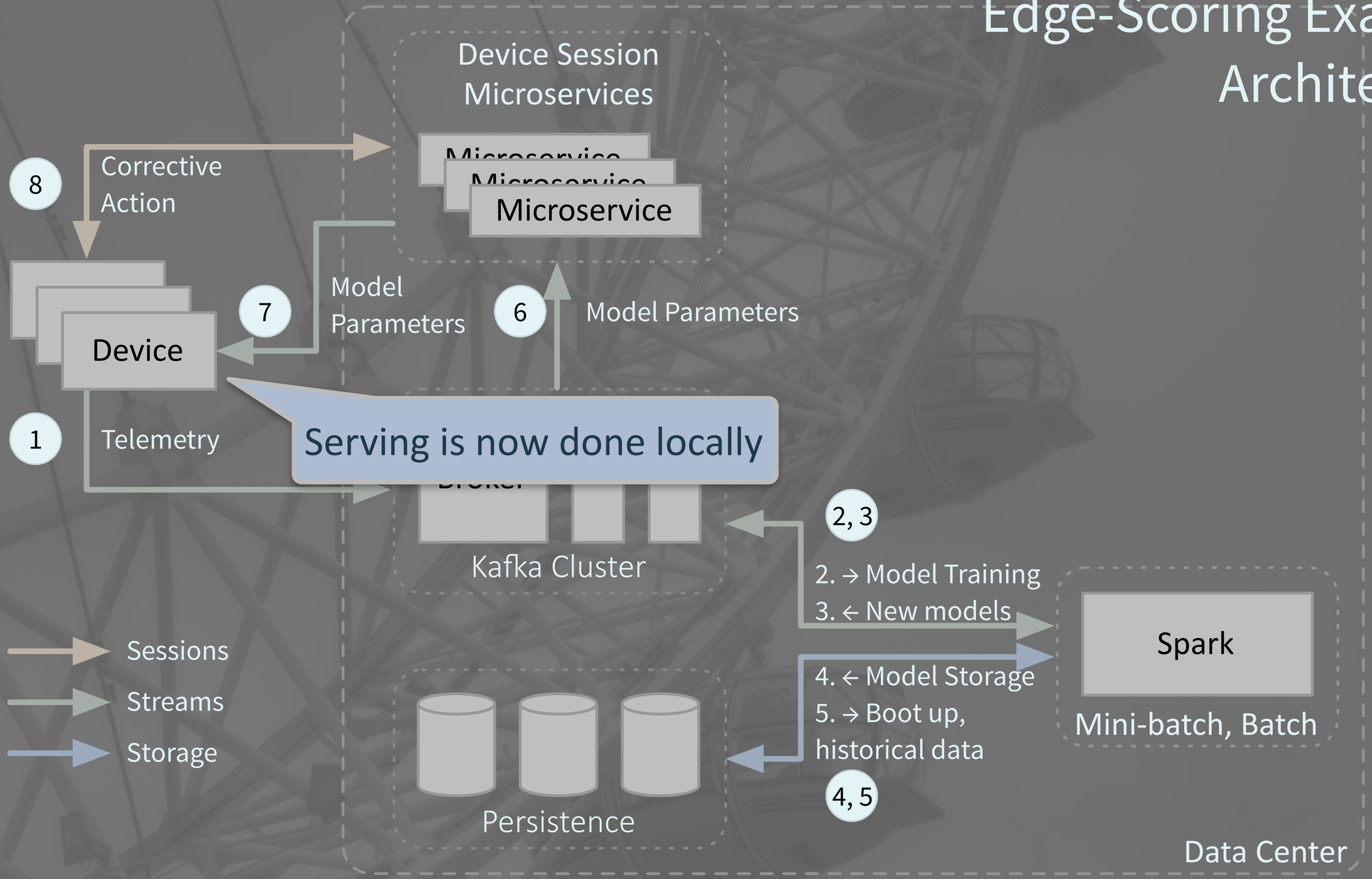


Scoring microservices now on the edge devices, so the device session microservices (in the datacenter) now ingest model updates…

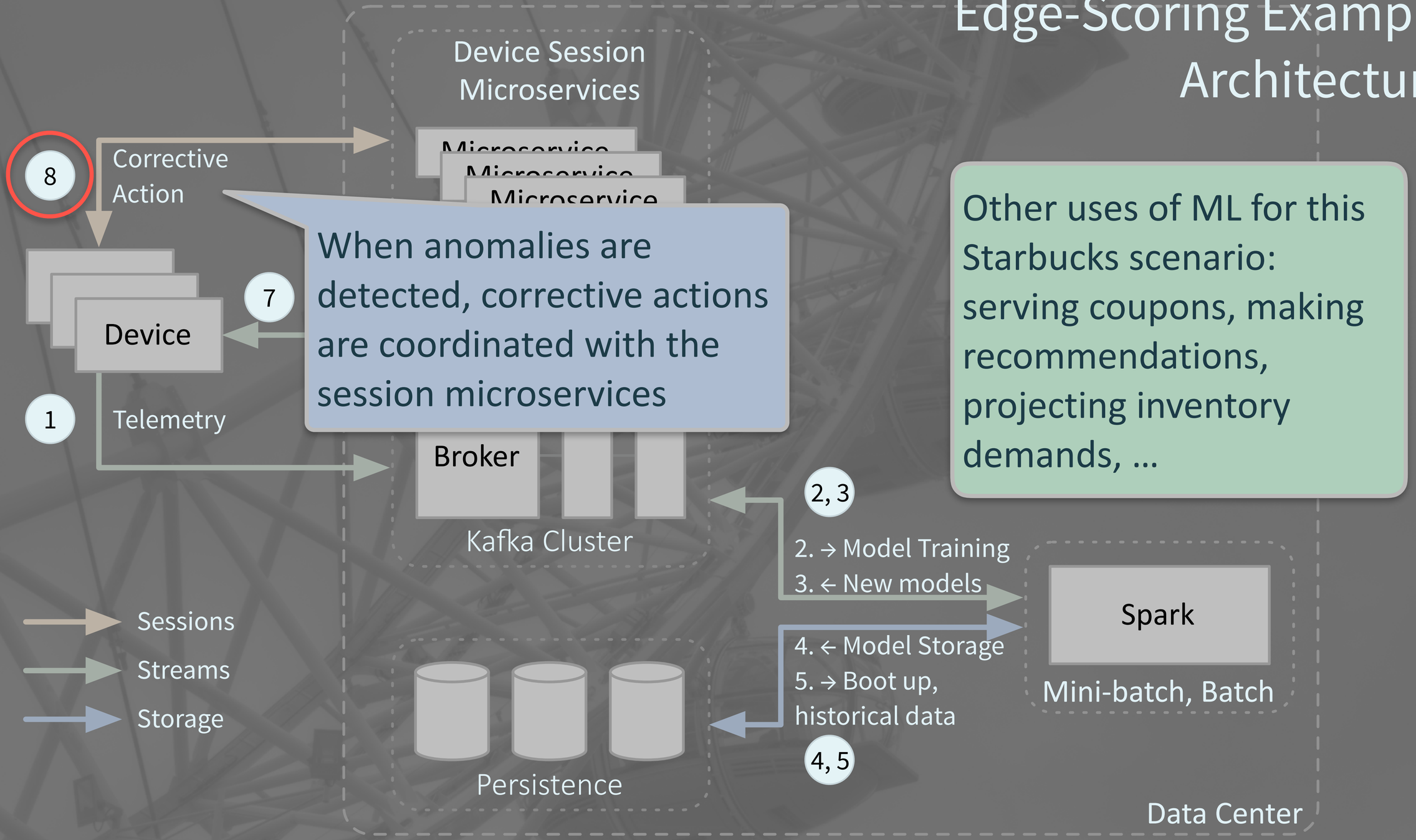
Edge-Scoring Example Architecture



Edge-Scoring Example Architecture

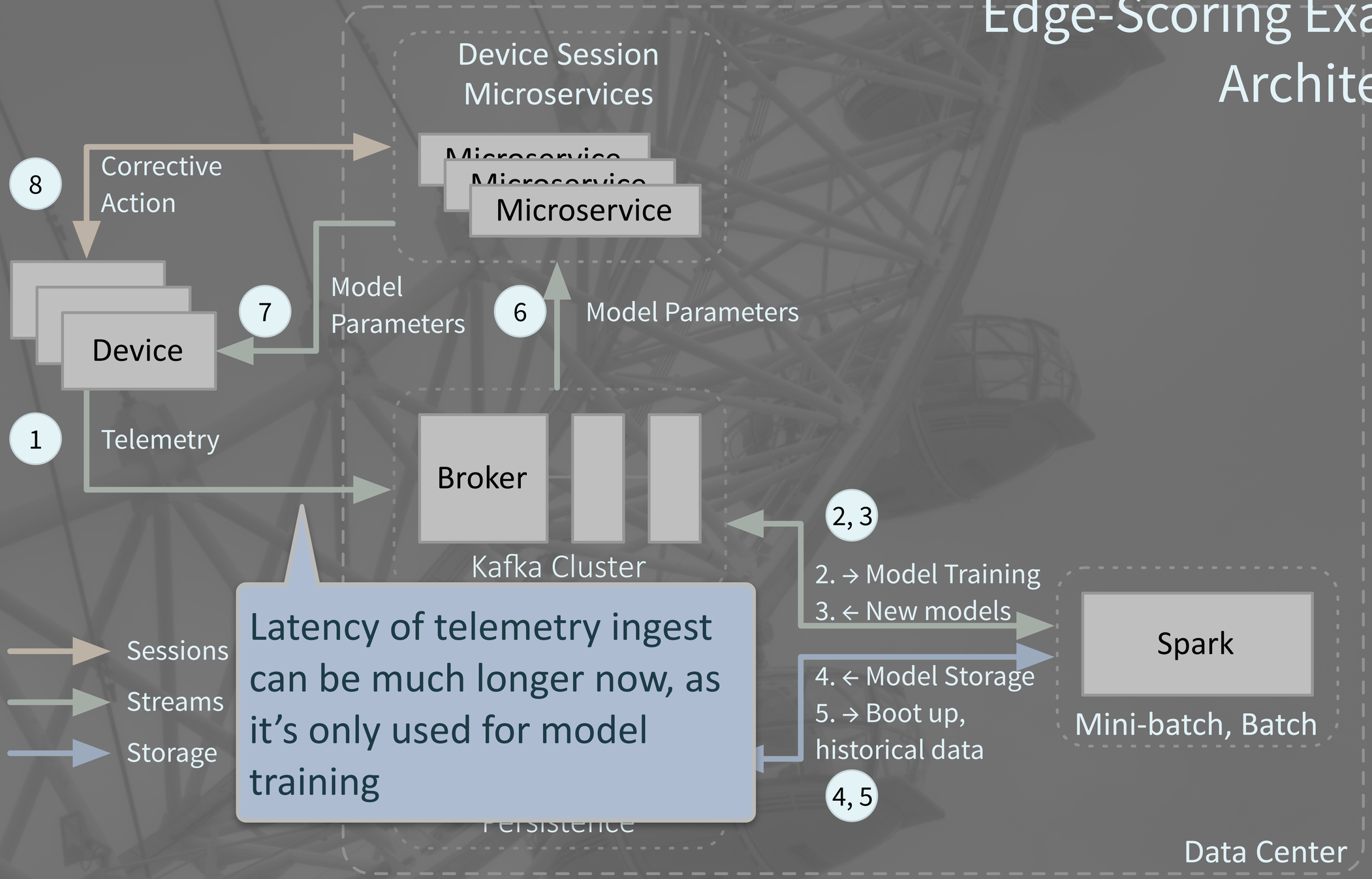


Edge-Scoring Example Architecture

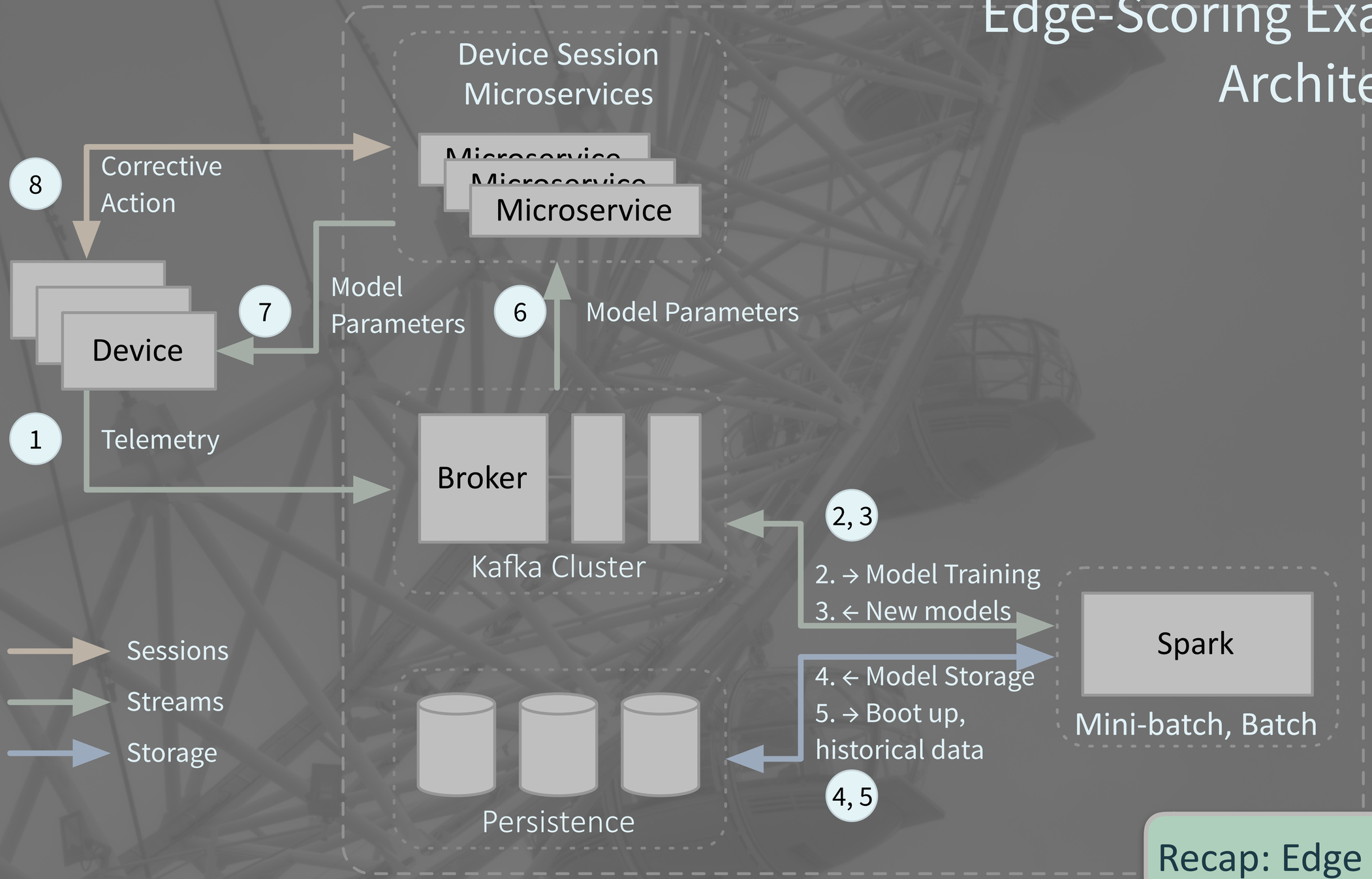


Other uses of ML for this Starbucks scenario: serving coupons, making recommendations, projecting inventory demands, …

Edge-Scoring Example Architecture



Edge-Scoring Example Architecture



Recap: Edge Serving

Fas ases

Batch changed to streaming for competitive advantage

Predictive Analytics	IoT	Real-time Personalization	Real-time Financial Processes
Apply ML models to large volumes of device data to pre-empt failures / outages	Real-time consumer and industrial Device and Supply Chain management at scale	Real-time marketing based on behavior, location, inventory levels, product promotions, etc.	Drive better business outcomes through real-time risk, fraud detection, compliance, audit, governance, etc.
Hewlett Packard Enterprise	STARBUCKS COFFEE	RoyalCaribbean INTERNATIONAL	CapitalOne

https://www.lightbend.com/customers

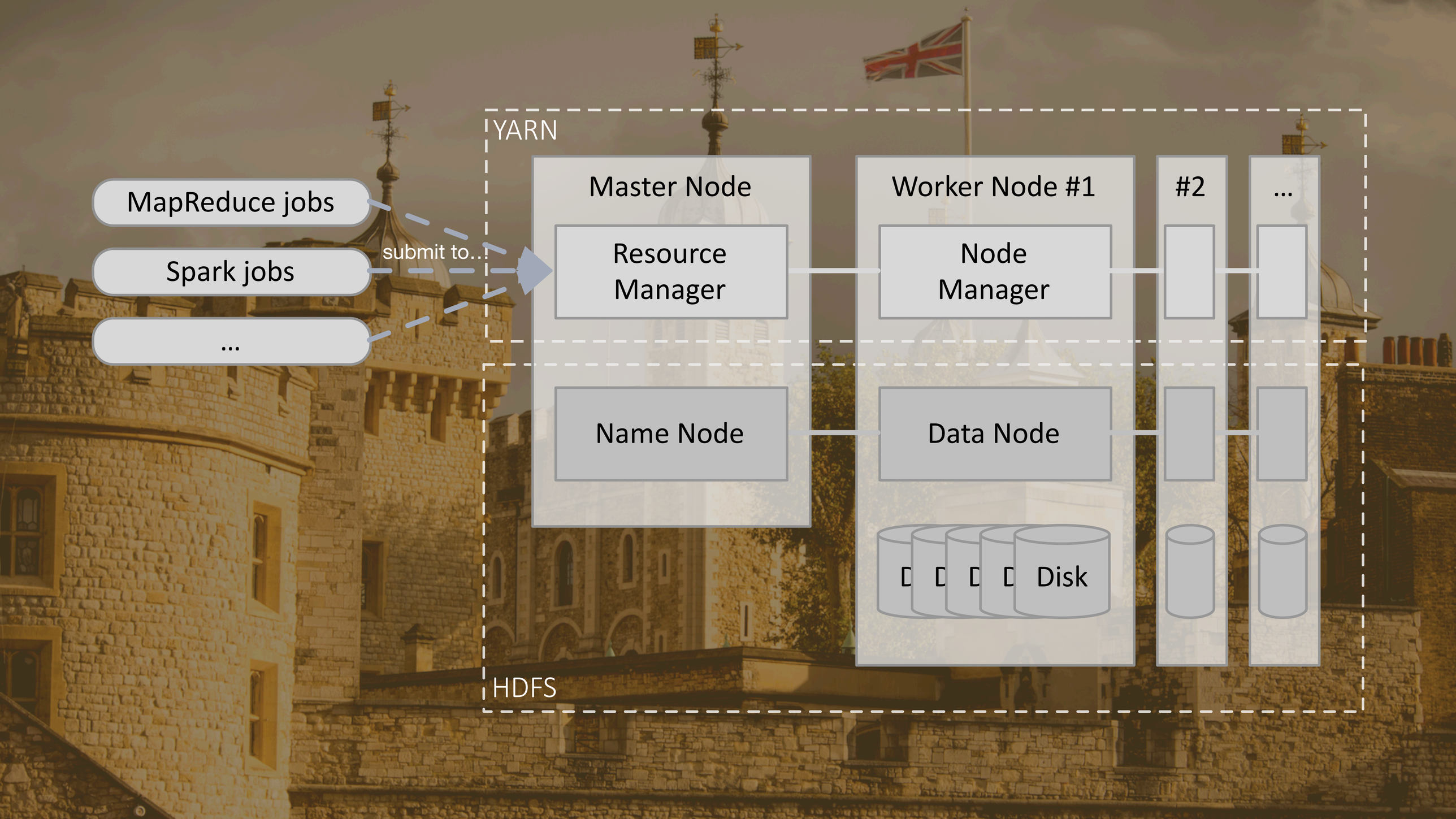
Technology Choices

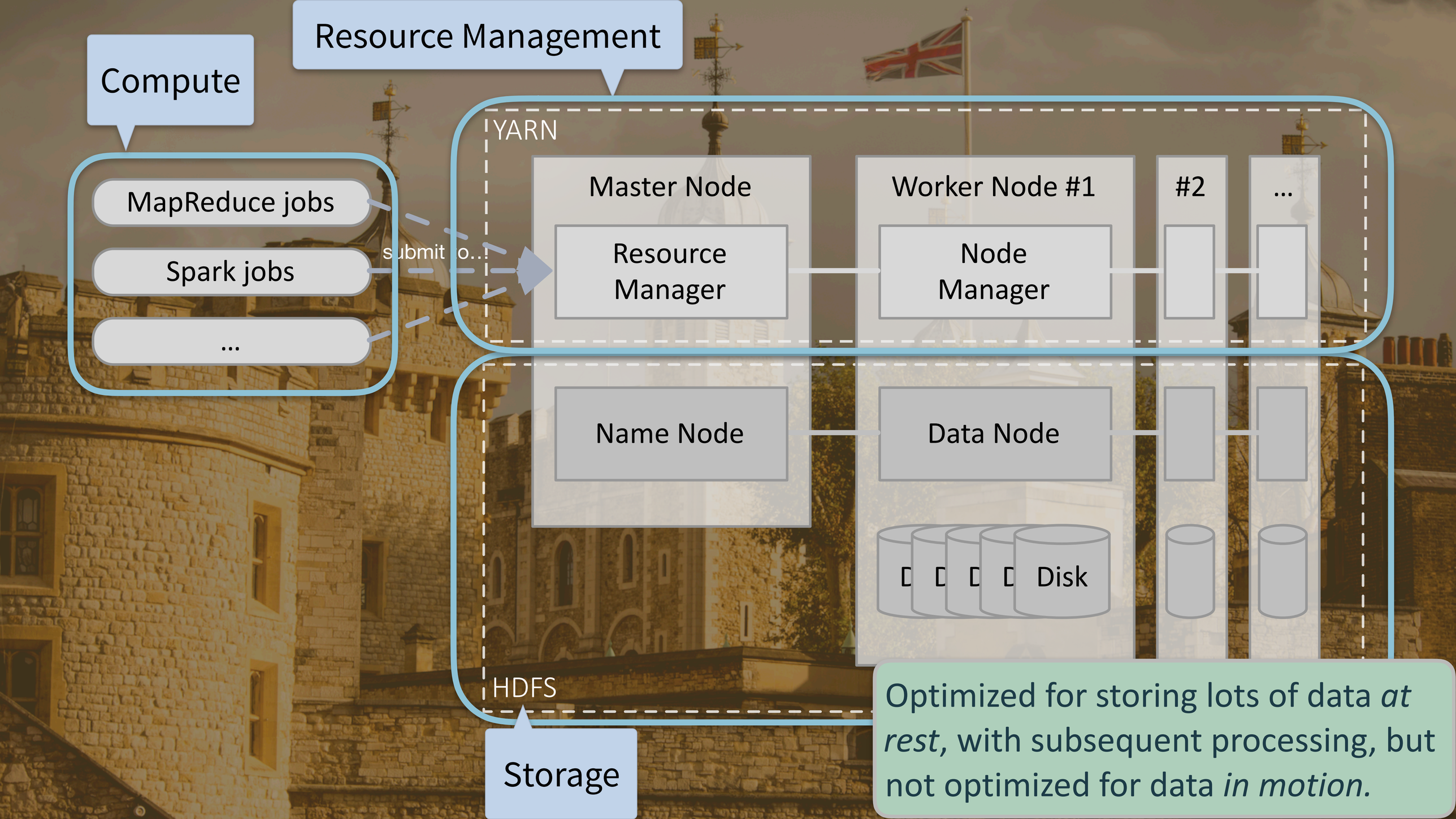
- More than "faster" Hadoop…
- New architectures that merge data processing with microservices

Technology Choices

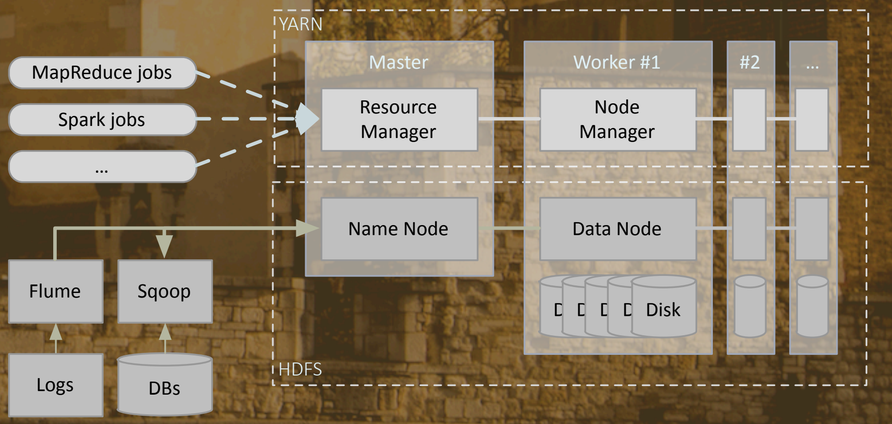
Recall Hadoop…

- Data warehouse replacement
- Historical analysis
- Interactive exploration
- Offline training of machine learning models
- …

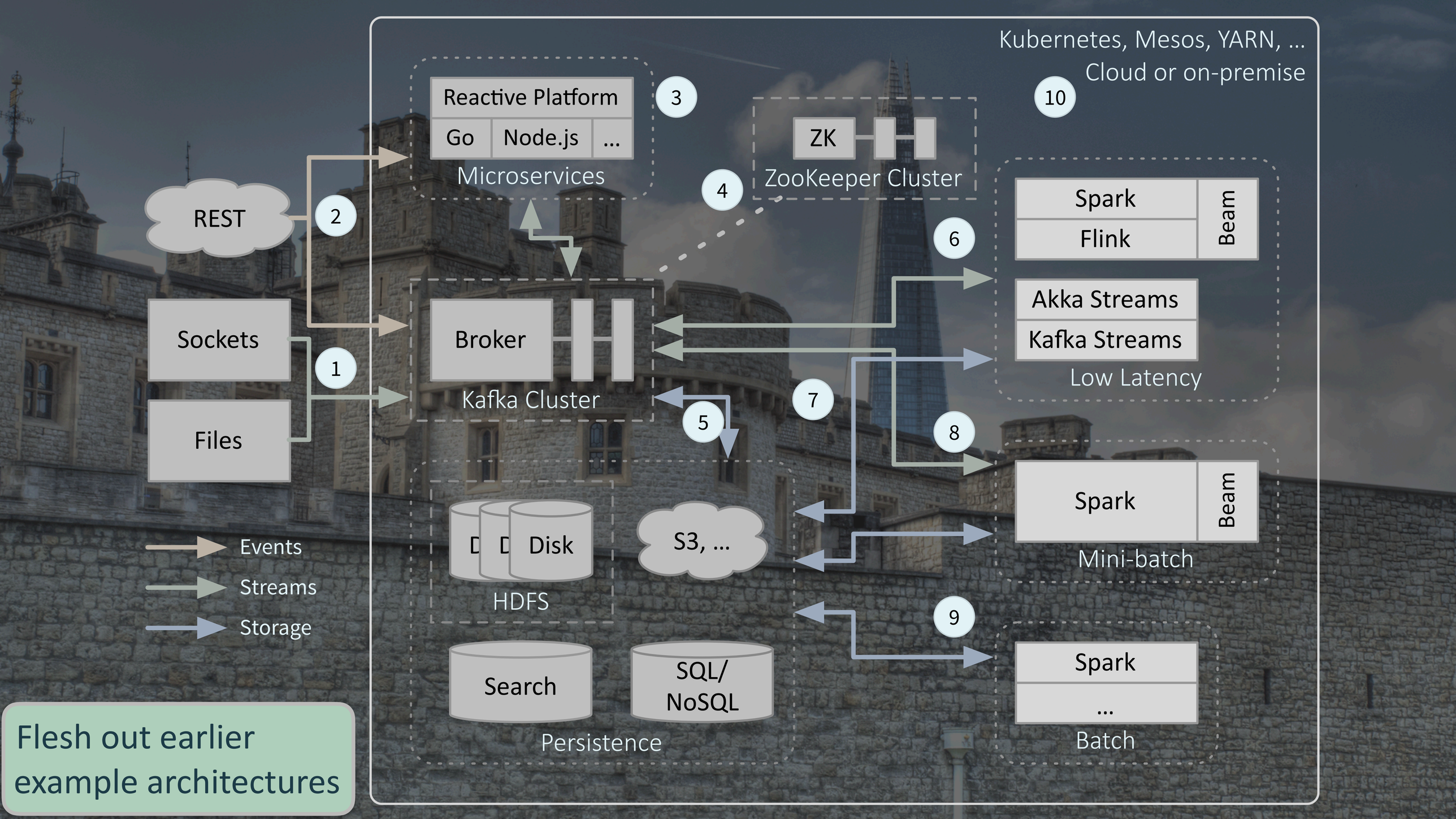

YARN
Master Node
Worker Node #1
#2
…
MapReduce jobs
Spark jobs
…
submit to…
Resource Manager
Node Manager
Name Node
Data Node
Disk
HDFS

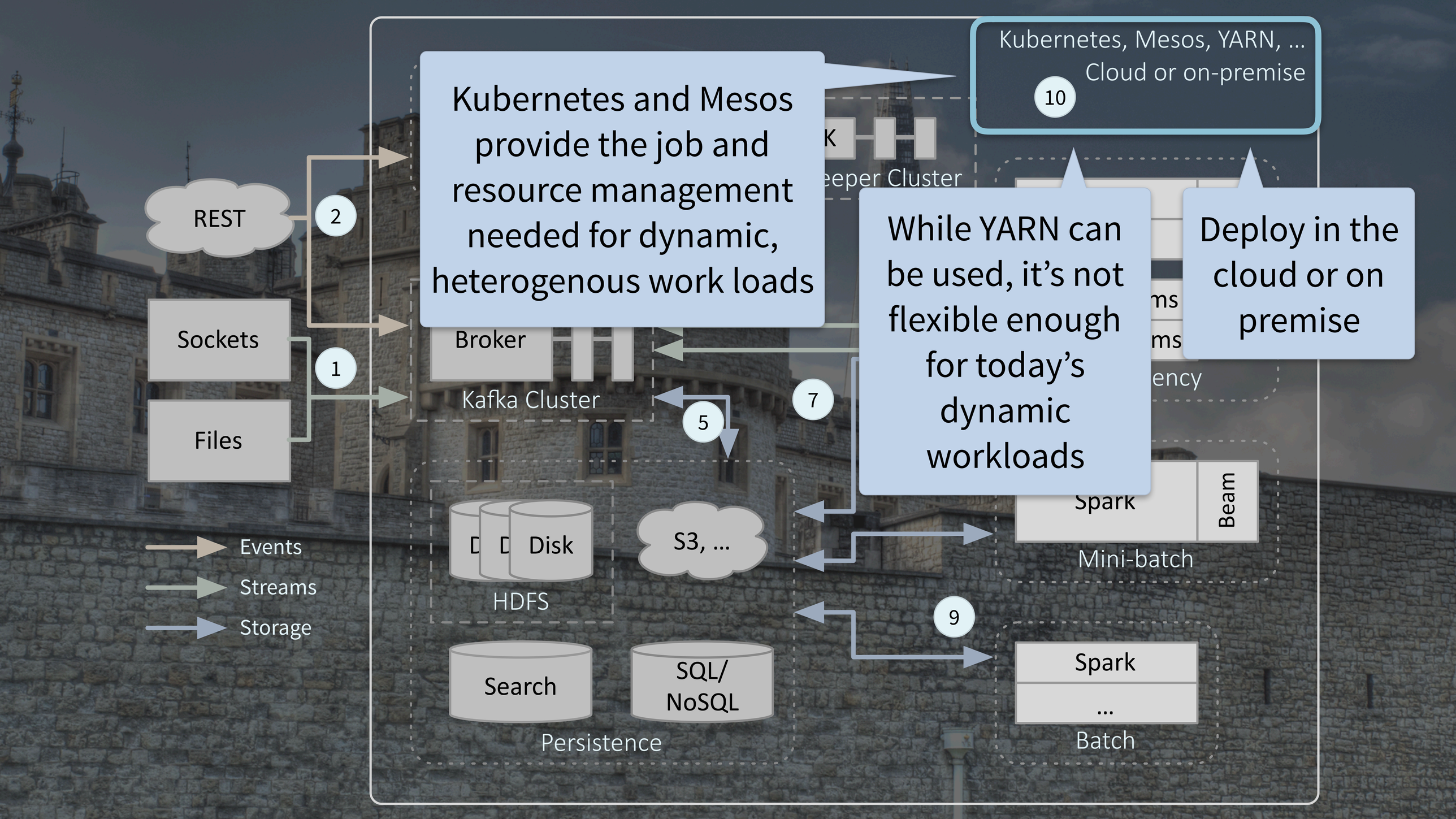

Resource Management
Compute
YARN
MapReduce jobs
Spark jobs
...
submit to...
Master Node
Worker Node #1
#2
...
Resource Manager
Node Manager
Name Node
Data Node
Disk
HDFS
Storage
Optimized for storing lots of data *at rest*, with subsequent processing, but not optimized for data *in motion.*

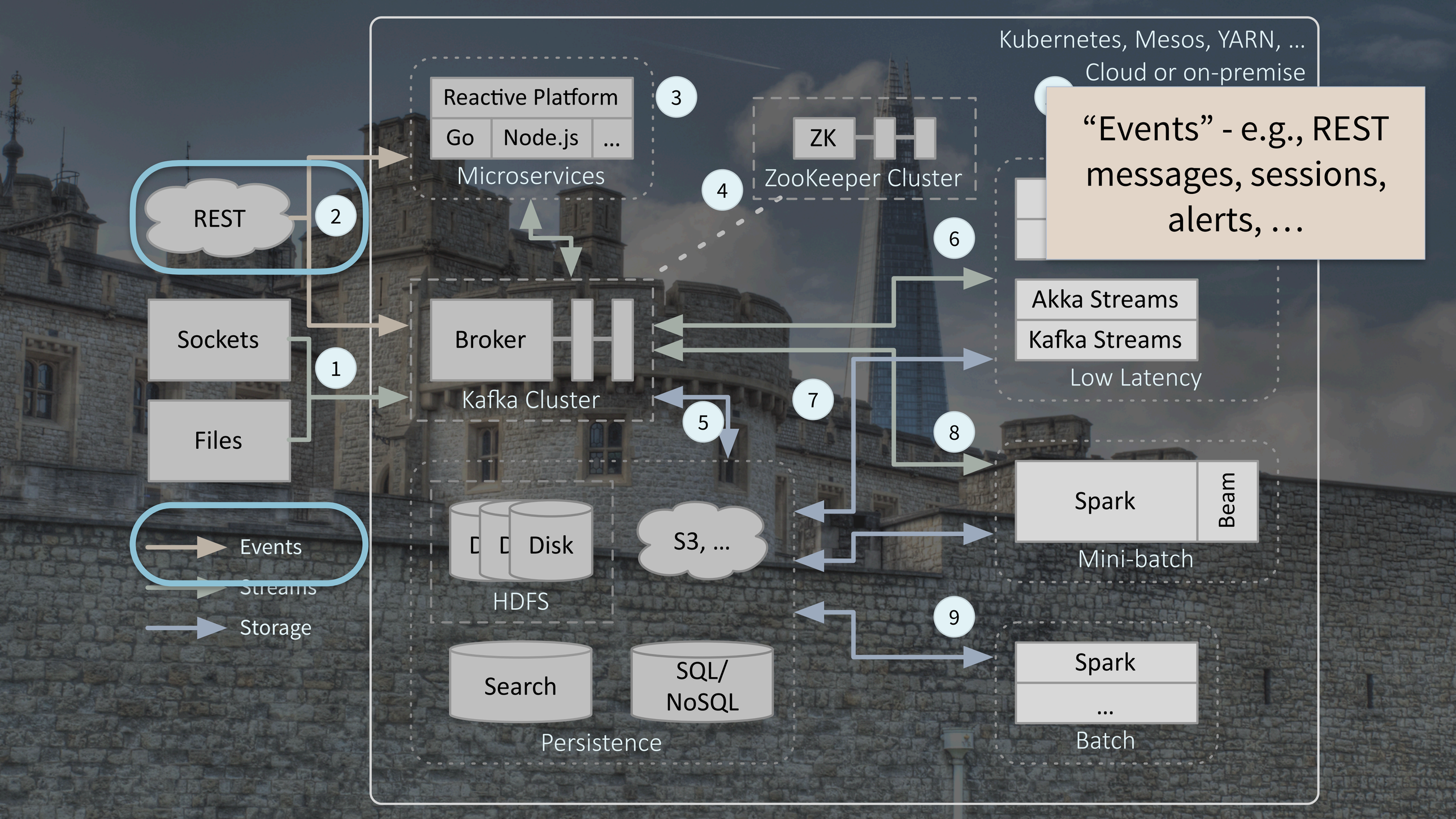
- Hadoop is ideal for batch and interactive apps
- … but also constrained by that model

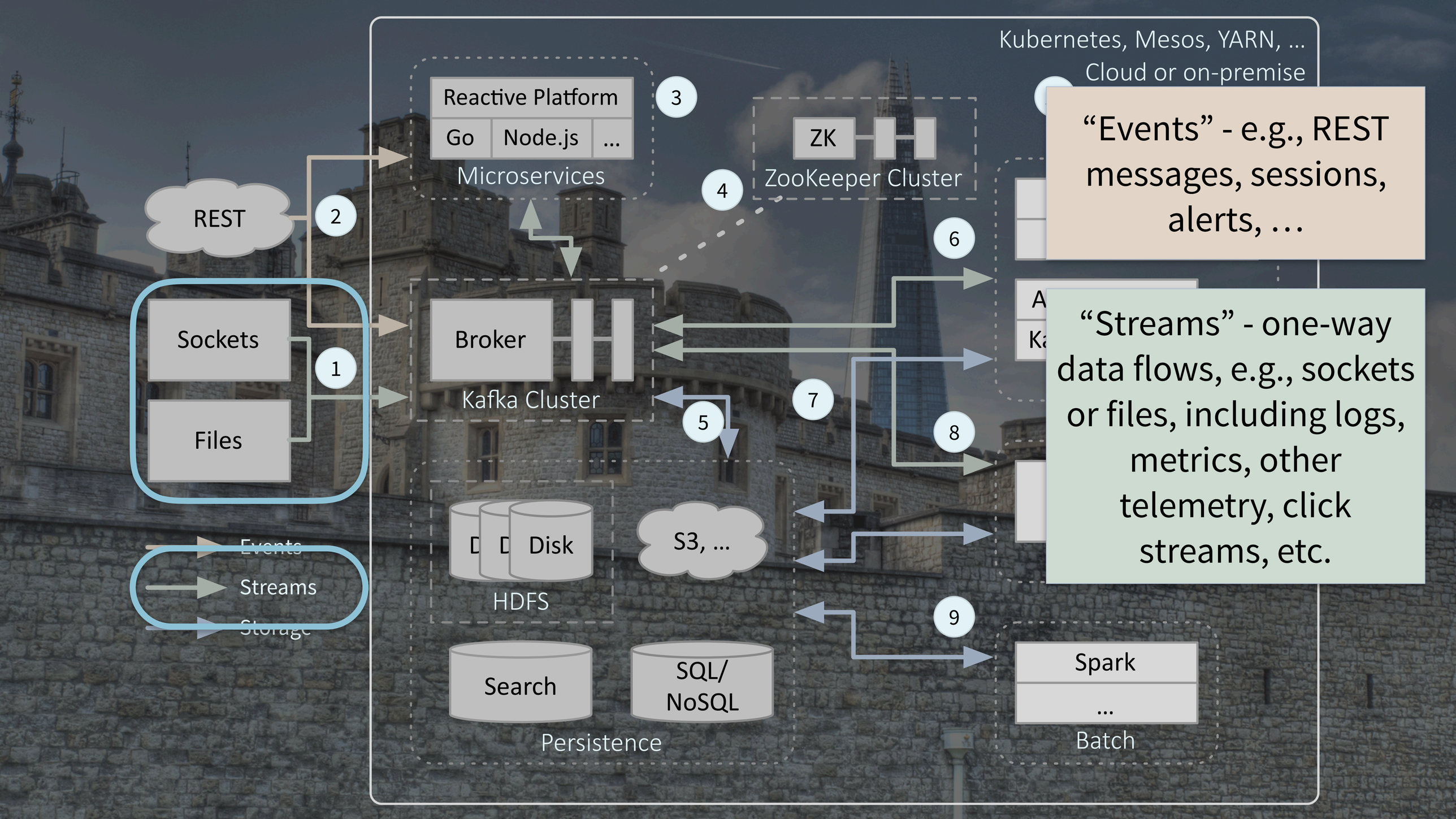


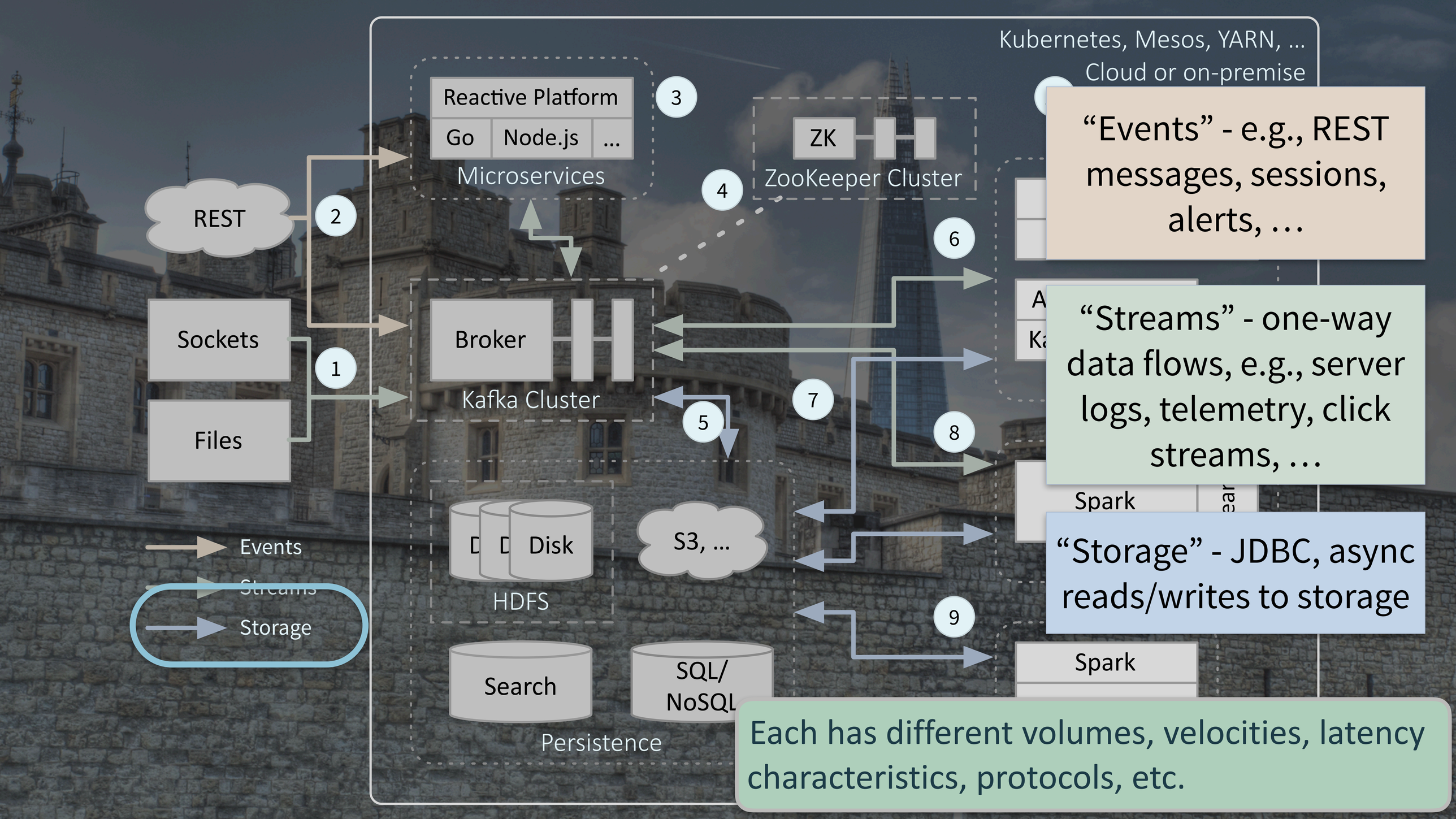
New Fast Data Architecture

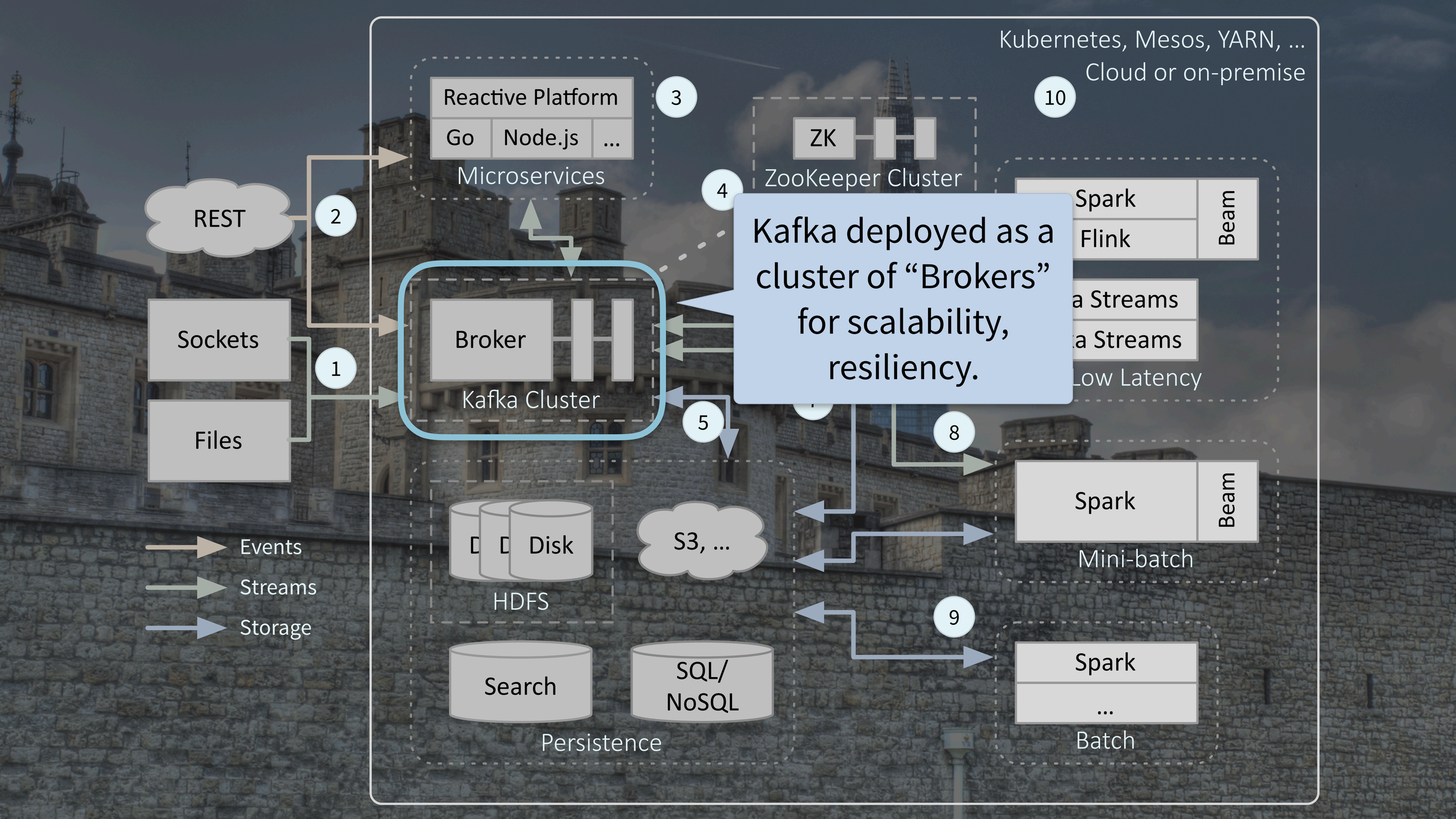

Kubernetes, Mesos, YARN, ...
Cloud or on-premise
10
Reactive Platform
Go
Node.js
...
Microservices
3
ZK
ZooKeeper Cluster
4
REST
2
Sockets
1
Files
Broker
Kafka Cluster
5
6
Spark
Flink
Beam
Akka Streams
Kafka Streams
Low Latency
7
8
Spark
Beam
Mini-batch
Disk
HDFS
S3, ...
9
Spark
...
Batch
Search
SQL/
NoSQL
Persistence
Events
Streams
Storage
Flesh out earlier
example architectures

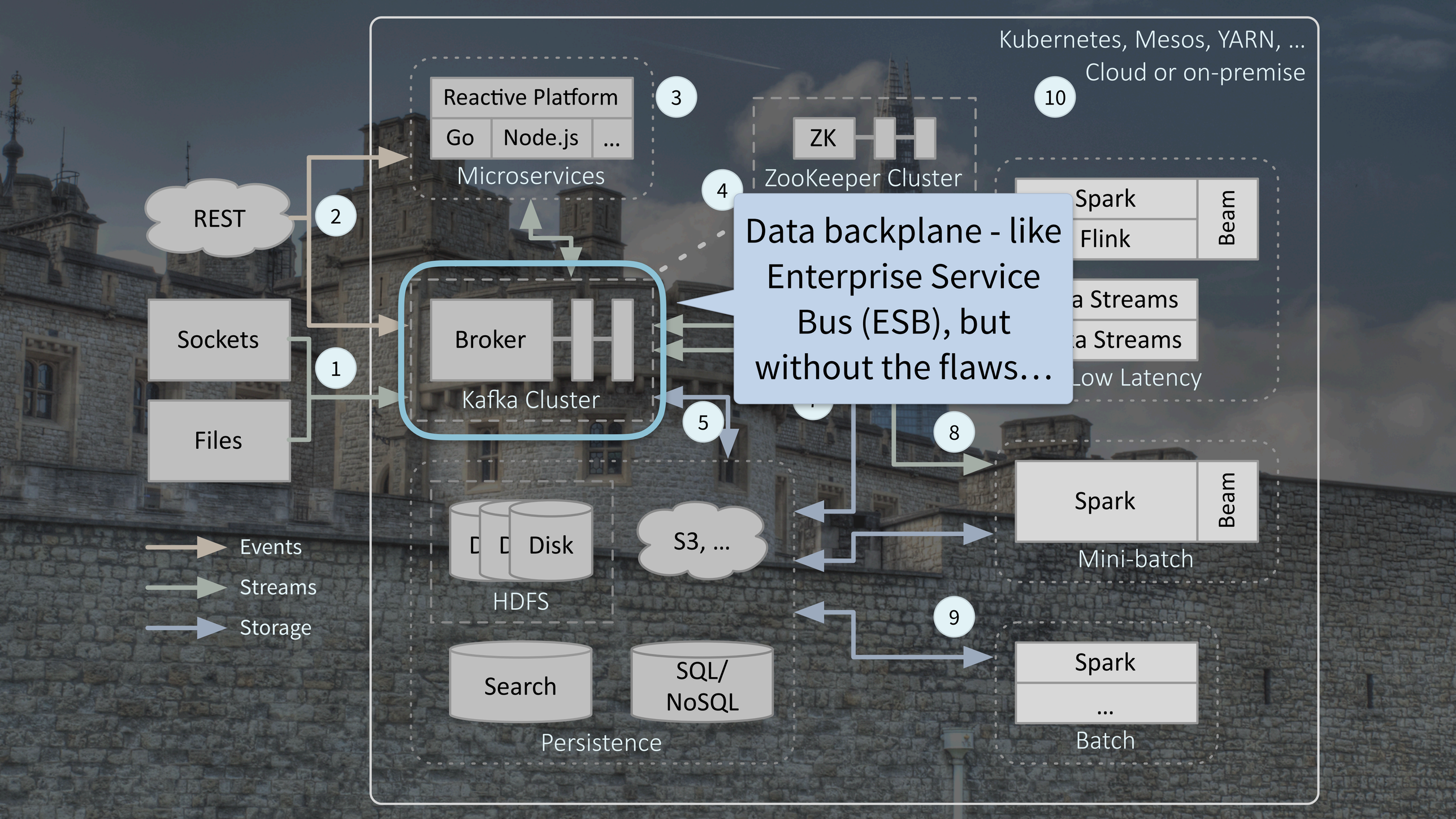

Kubernetes, Mesos, YARN, ...
Cloud or on-premise
10
Kubernetes and Mesos provide the job and resource management needed for dynamic, heterogenous work loads
While YARN can be used, it's not flexible enough for today's dynamic workloads
Deploy in the cloud or on premise
REST
2
Sockets
1
Files
Broker
Kafka Cluster
5
7
Disk
HDFS
S3, ...
Search
SQL/ NoSQL
Persistence
Spark
Beam
Mini-batch
9
Spark
...
Batch
Events
Streams
Storage


Kubernetes, Mesos, YARN, …
Cloud or on-premise
Reactive Platform
Go
Node.js
…
Microservices
3
ZK
ZooKeeper Cluster
4
REST
2
Sockets
1
Files
Broker
Kafka Cluster
5
6
7
8
9
Akka Streams
Kafka Streams
Low Latency
Spark
Beam
Mini-batch
Disk
HDFS
S3, …
Search
SQL/
NoSQL
Persistence
Spark
…
Batch
Events
Streams
Storage
“Events” - e.g., REST messages, sessions, alerts, …


Kubernetes, Mesos, YARN, …
Cloud or on-premise
Reactive Platform
Go
Node.js
…
Microservices
3
ZK
ZooKeeper Cluster
4
REST
2
Sockets
Files
1
Broker
Kafka Cluster
5
6
7
8
Disk
HDFS
S3, …
Search
SQL/
NoSQL
Persistence
9
Spark
…
Batch
Events
Streams
Storage
“Events” - e.g., REST messages, sessions, alerts, …
“Streams” - one-way data flows, e.g., sockets or files, including logs, metrics, other telemetry, click streams, etc.


Kubernetes, Mesos, YARN, …
Cloud or on-premise
Reactive Platform
Go
Node.js
…
Microservices
3
ZK
ZooKeeper Cluster
4
REST
2
Sockets
1
Files
Broker
Kafka Cluster
5
6
7
8
9
Spark
Spark
Disk
HDFS
S3, …
Search
SQL/
NoSQL
Persistence
Events
Streams
Storage
“Events” - e.g., REST messages, sessions, alerts, …
“Streams” - one-way data flows, e.g., server logs, telemetry, click streams, …
“Storage” - JDBC, async reads/writes to storage
Each has different volumes, velocities, latency characteristics, protocols, etc.

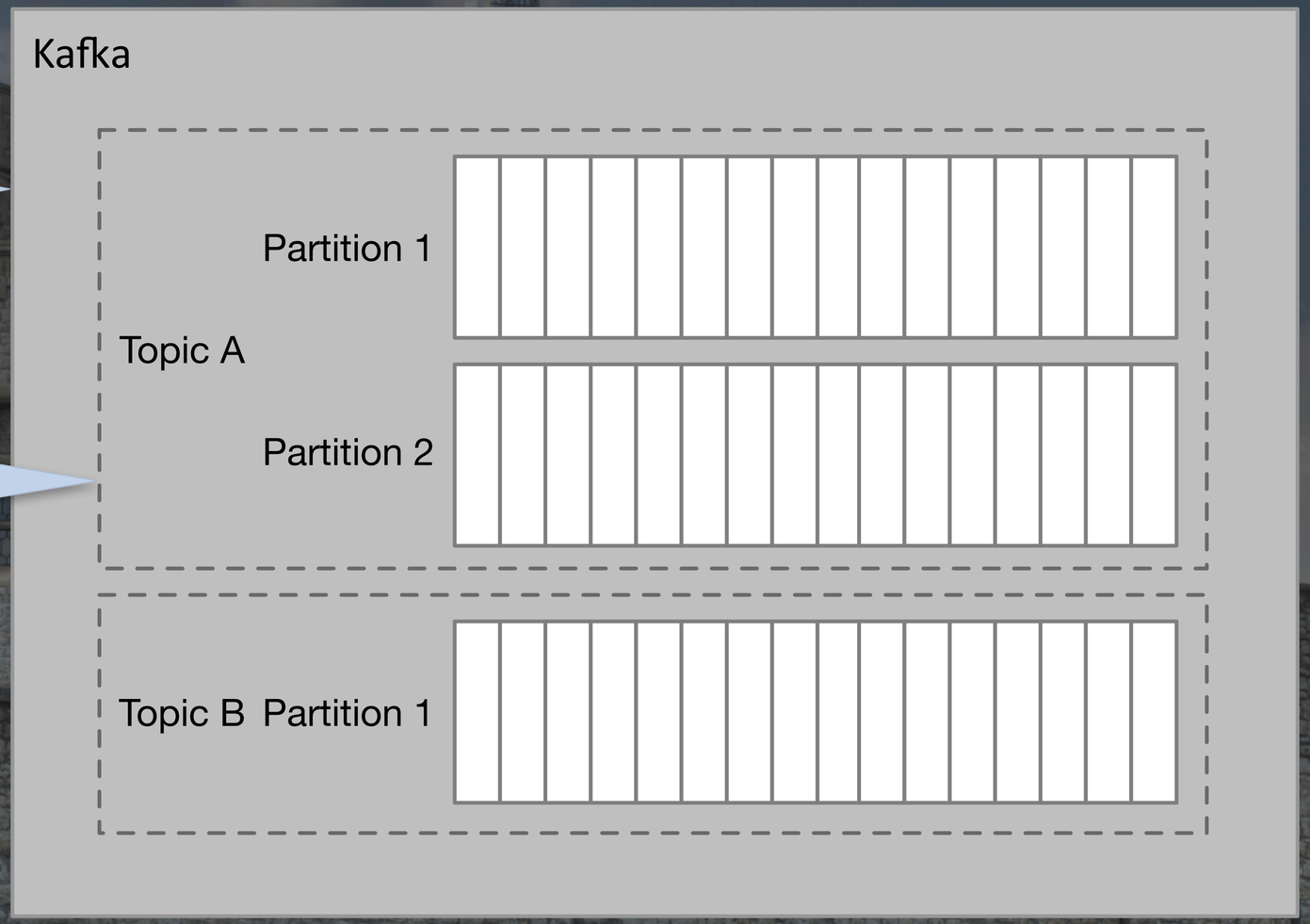

Kubernetes, Mesos, YARN, ...
Cloud or on-premise
Reactive Platform
Go
Node.js
...
Microservices
3
ZK
ZooKeeper Cluster
10
4
REST
2
Sockets
1
Files
Broker
Kafka Cluster
Kafka deployed as a cluster of “Brokers” for scalability, resiliency.
5
Spark
Flink
Beam
a Streams
a Streams
Low Latency
8
Spark
Beam
Mini-batch
Disk
HDFS
S3, ...
9
Spark
...
Batch
Search
SQL/
NoSQL
Persistence
Events
Streams
Storage


Kubernetes, Mesos, YARN, ...
Cloud or on-premise
Reactive Platform
Go
Node.js
...
Microservices
3
10
ZK
ZooKeeper Cluster
4
REST
2
Sockets
1
Files
Broker
Kafka Cluster
Data backplane - like Enterprise Service Bus (ESB), but without the flaws…
5
Spark
Flink
Beam
a Streams
a Streams
Low Latency
8
Spark
Beam
Mini-batch
Disk
HDFS
S3, ...
9
Spark
...
Batch
Search
SQL/
NoSQL
Persistence
Events
Streams
Storage

Why Kafka?

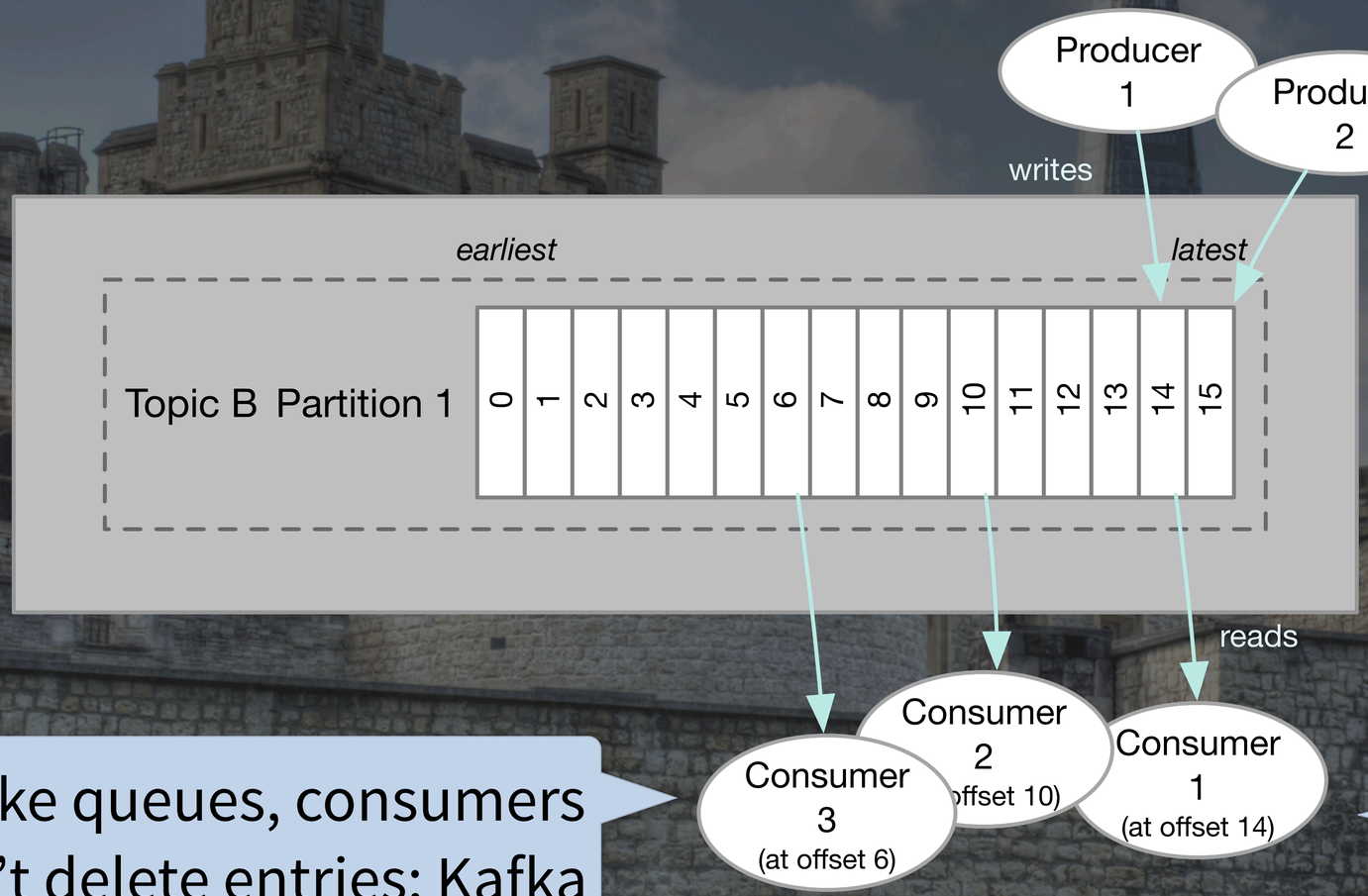
Organized into topics

Topics are partitioned, replicated, and distributed



Why Kafka?

Logs, *not* queues!

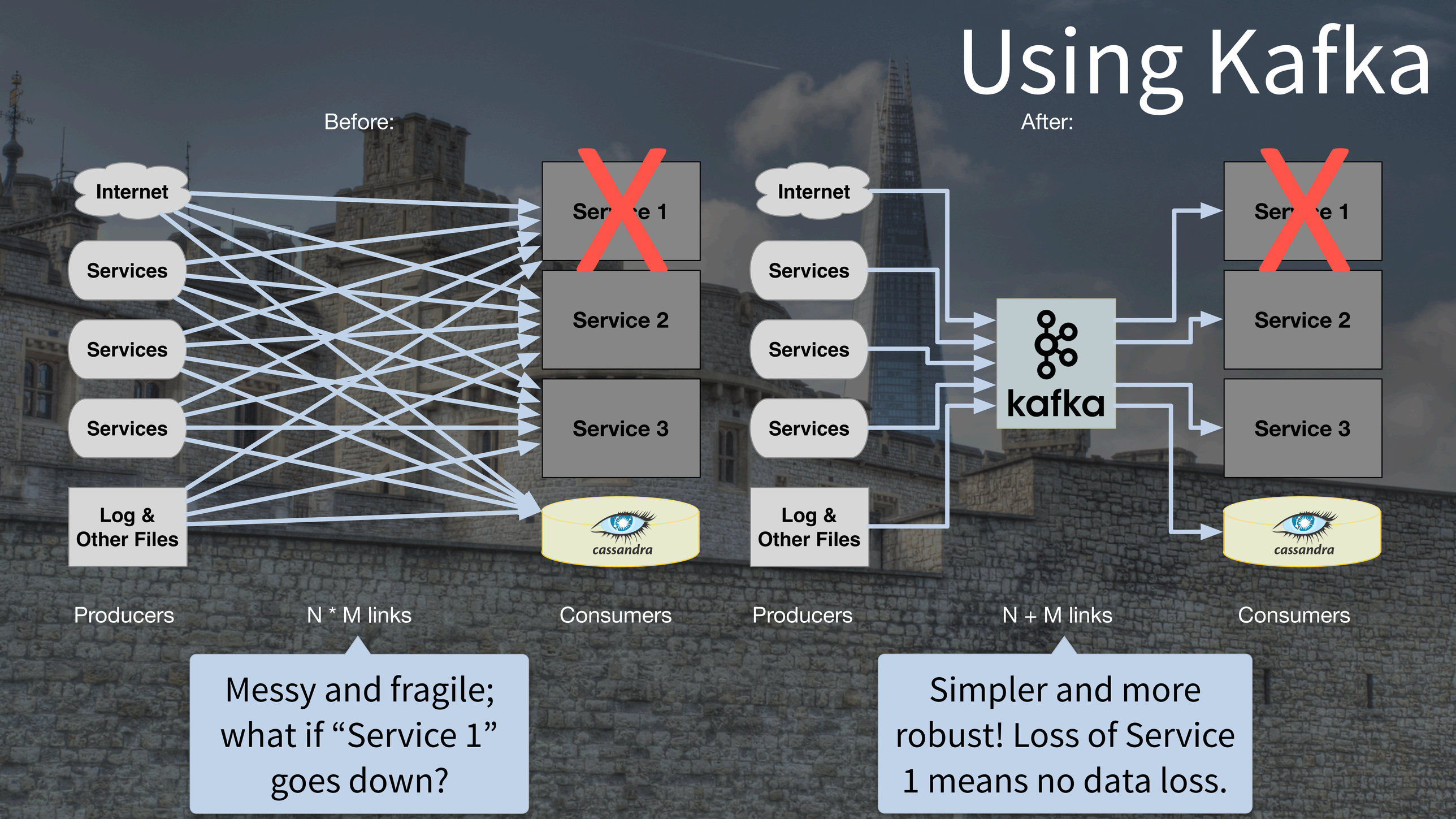


M Producers

N Consumers, who start reading where they want

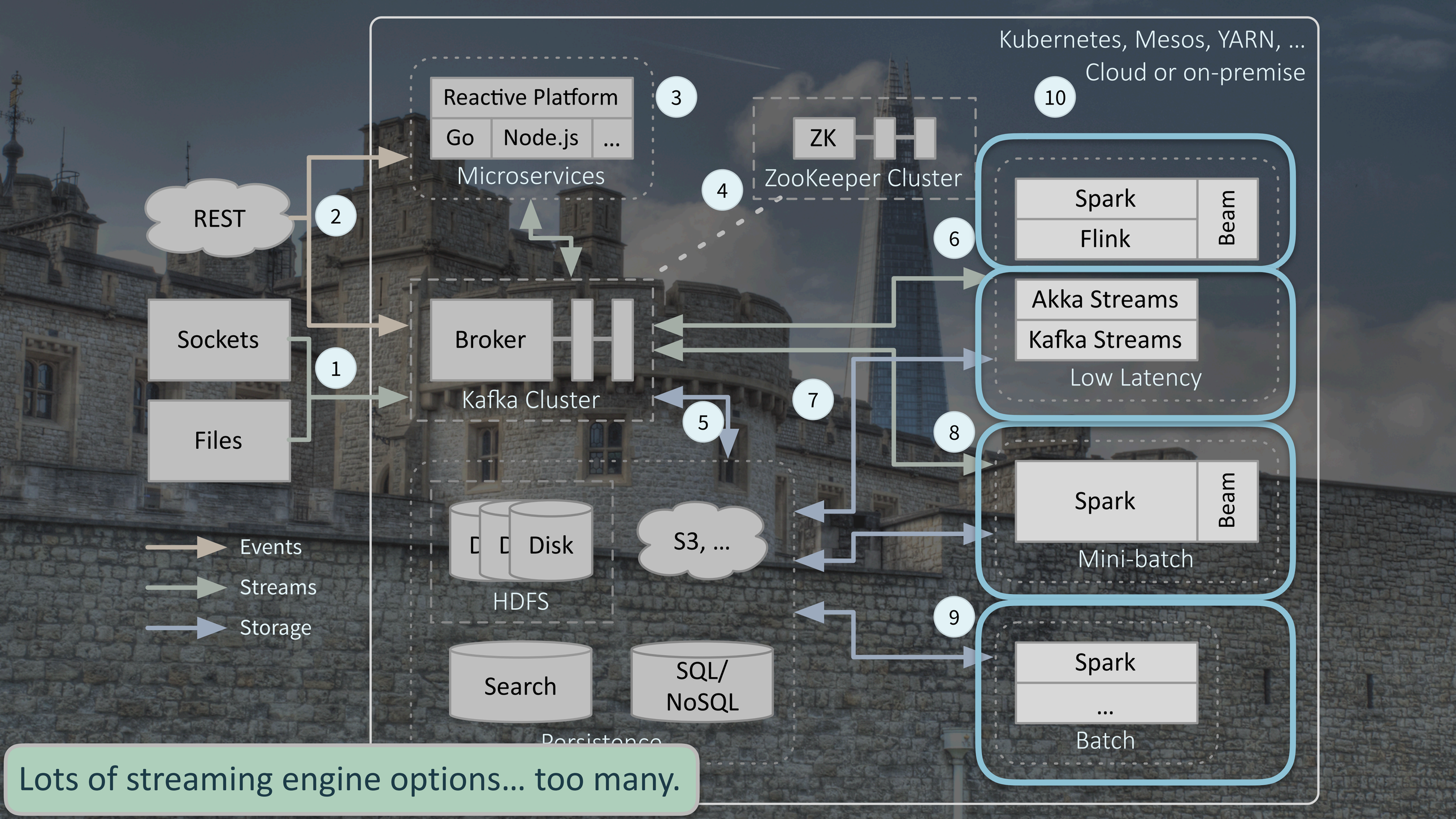
Unlike queues, consumers don't delete entries; Kafka manages their lifecycles

Using Kafka



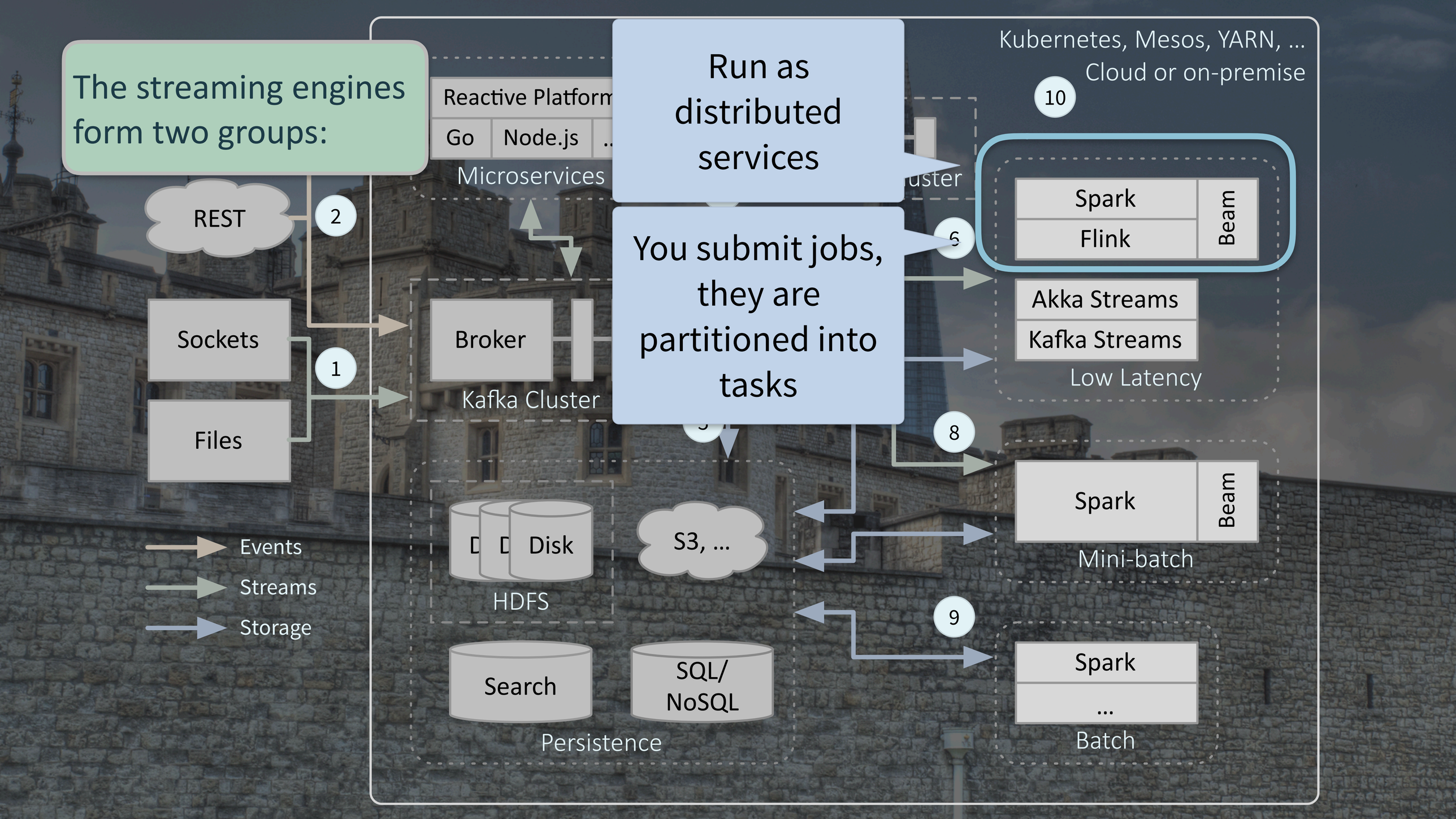
Messy and fragile; what if “Service 1” goes down?

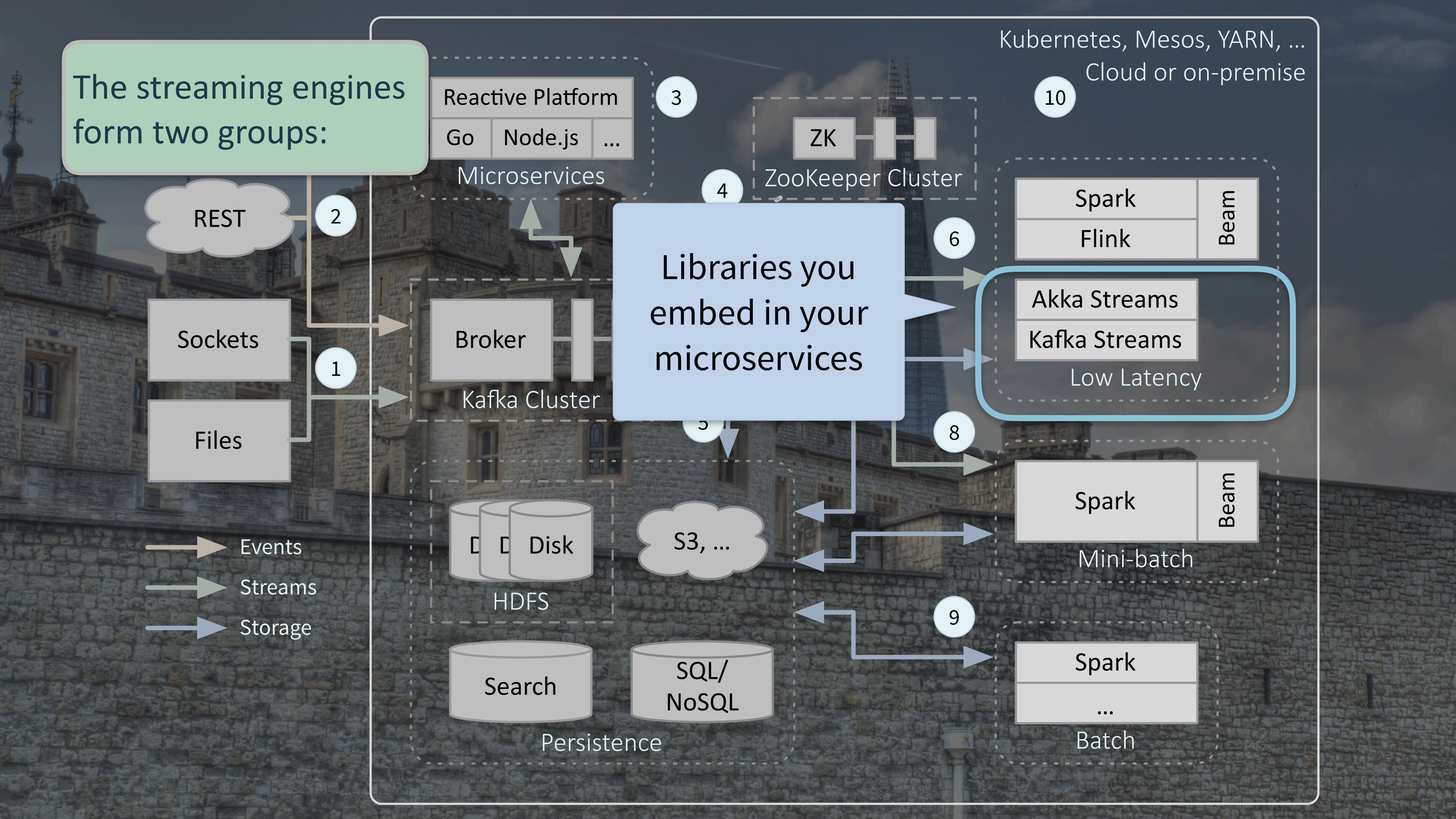
Simpler and more robust! Loss of Service 1 means no data loss.

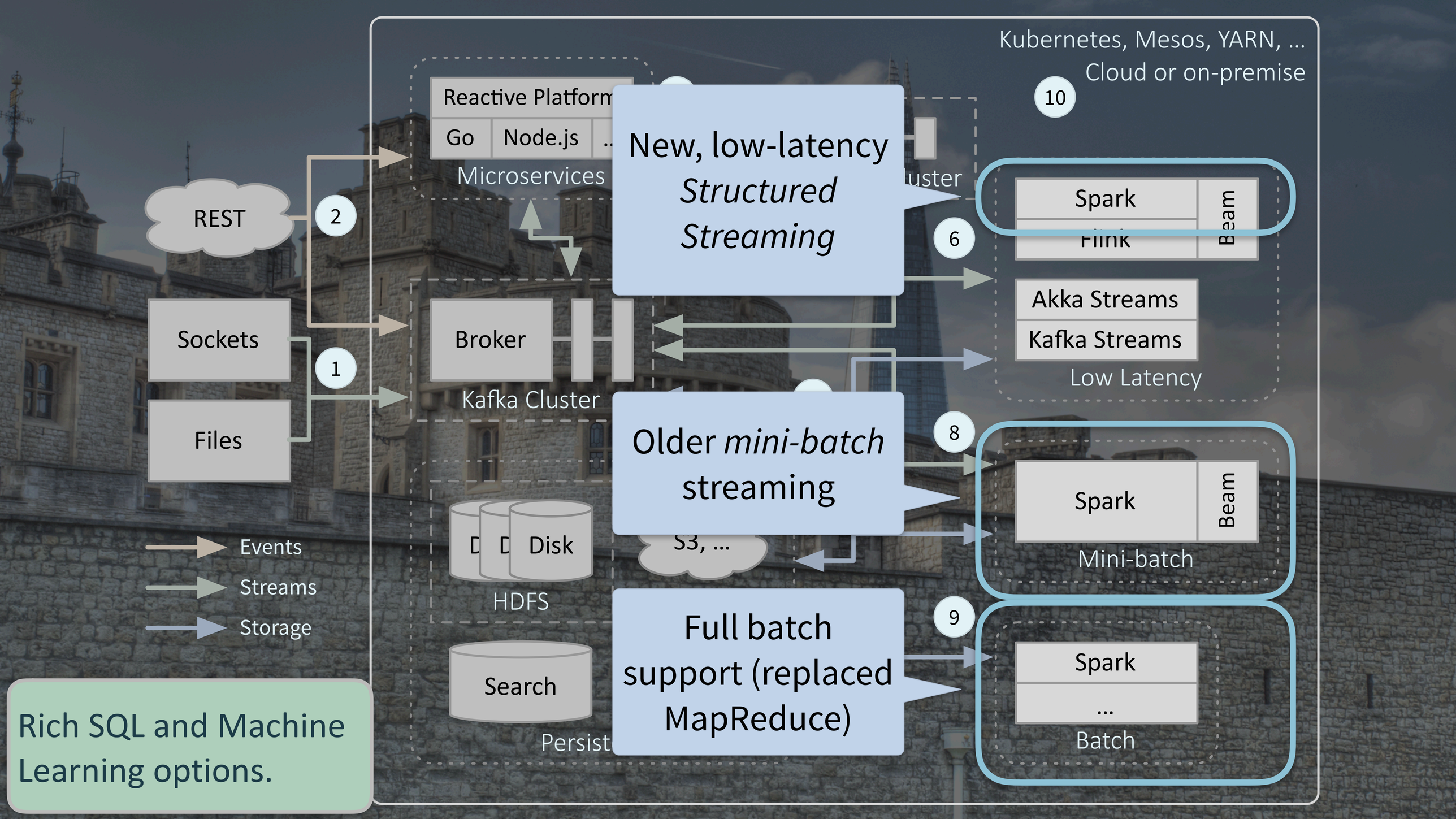

Kubernetes, Mesos, YARN, ...
Cloud or on-premise
Reactive Platform
Go
Node.js
...
Microservices
3
ZK
ZooKeeper Cluster
4
10
REST
2
Sockets
1
Files
Broker
Kafka Cluster
5
7
6
Spark
Flink
Beam
Akka Streams
Kafka Streams
Low Latency
8
Spark
Beam
Mini-batch
9
Spark
...
Batch
Disk
HDFS
S3, ...
Search
SQL/
NoSQL
Persistence
Events
Streams
Storage
Lots of streaming engine options... too many.

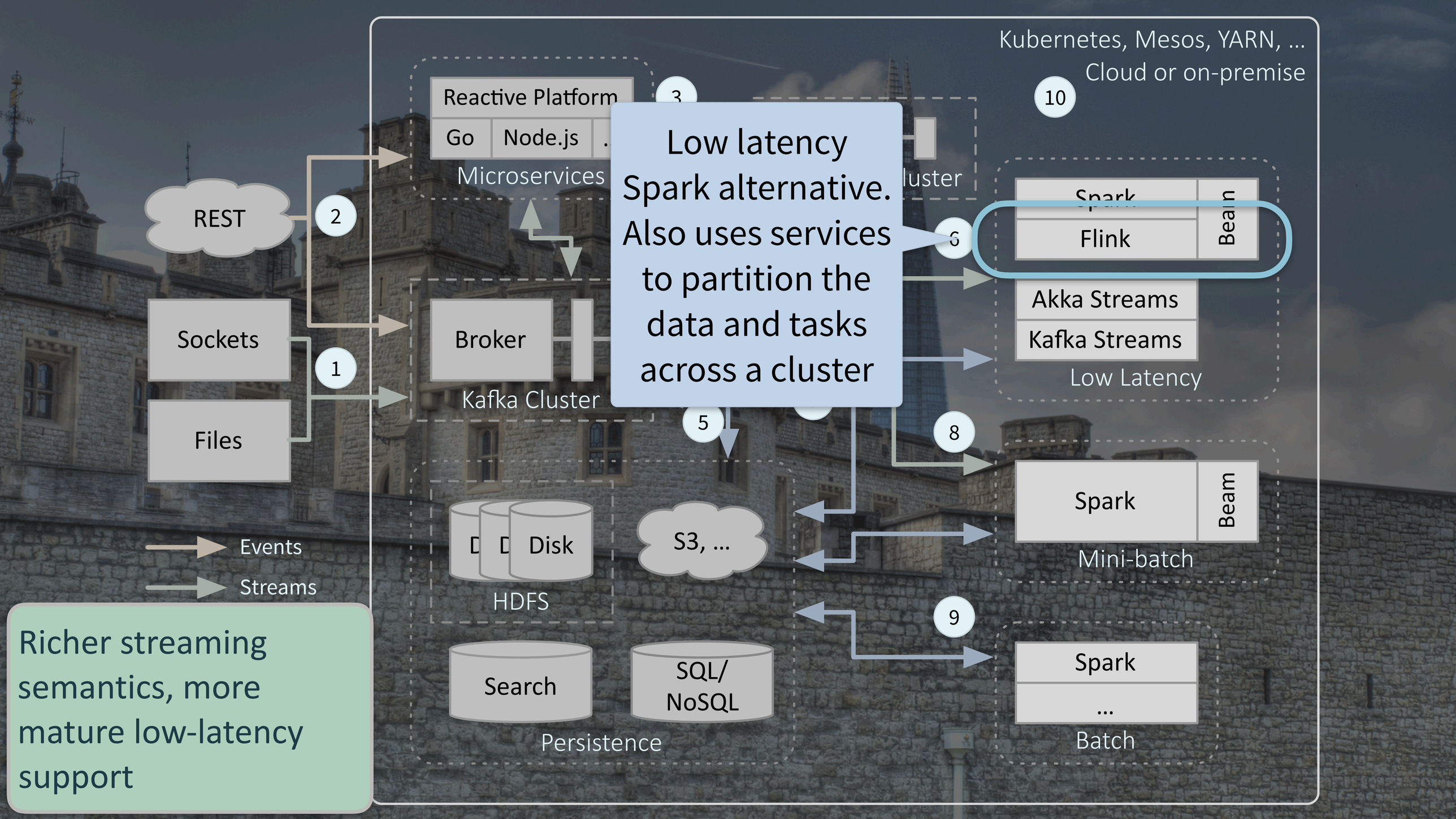
How do you choose?

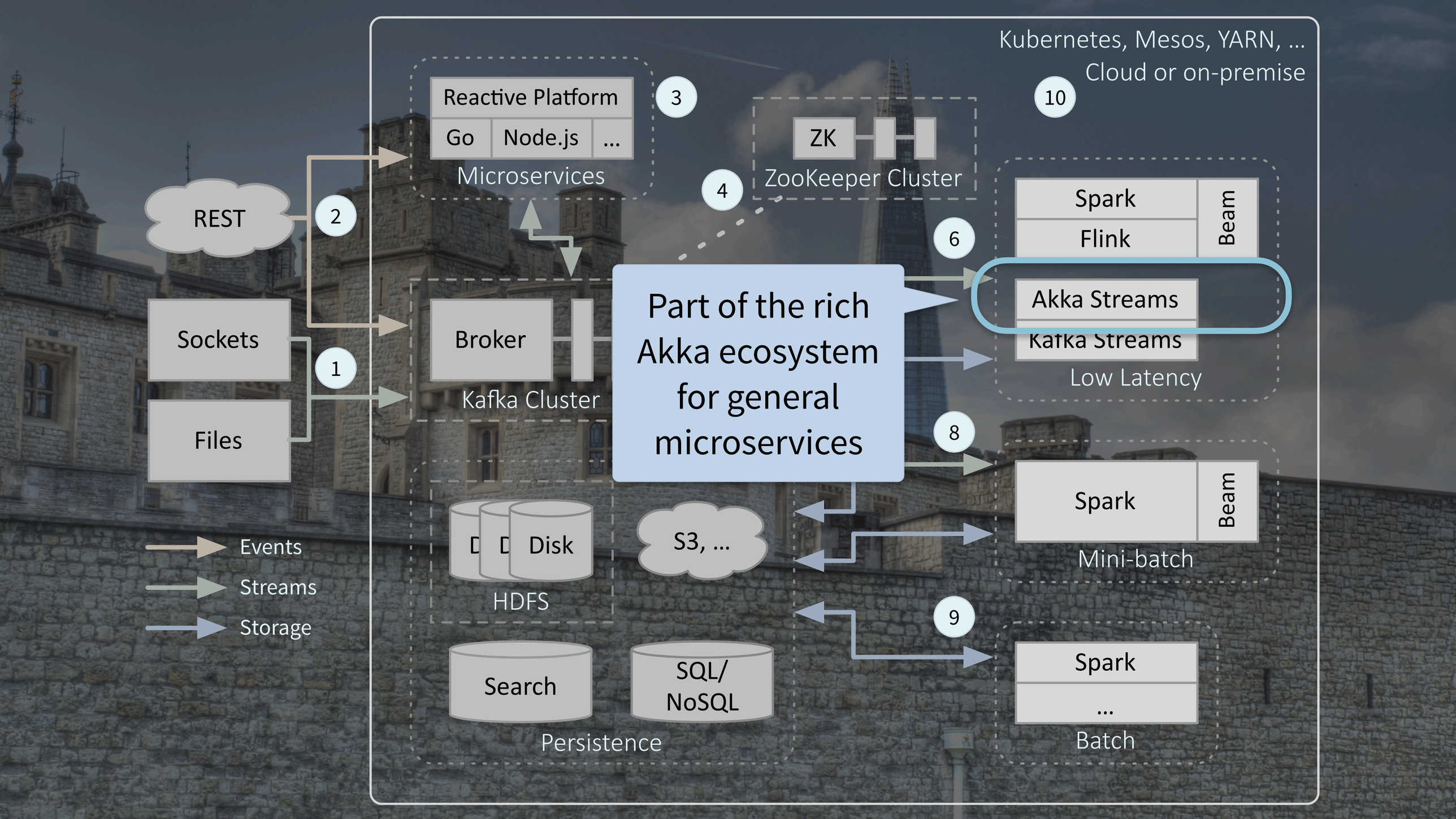
- Latency: how low?
- Volume per unit time: how high?
- Data processing: which kinds?
- Build, deploy, and manage services: what are your preferences?

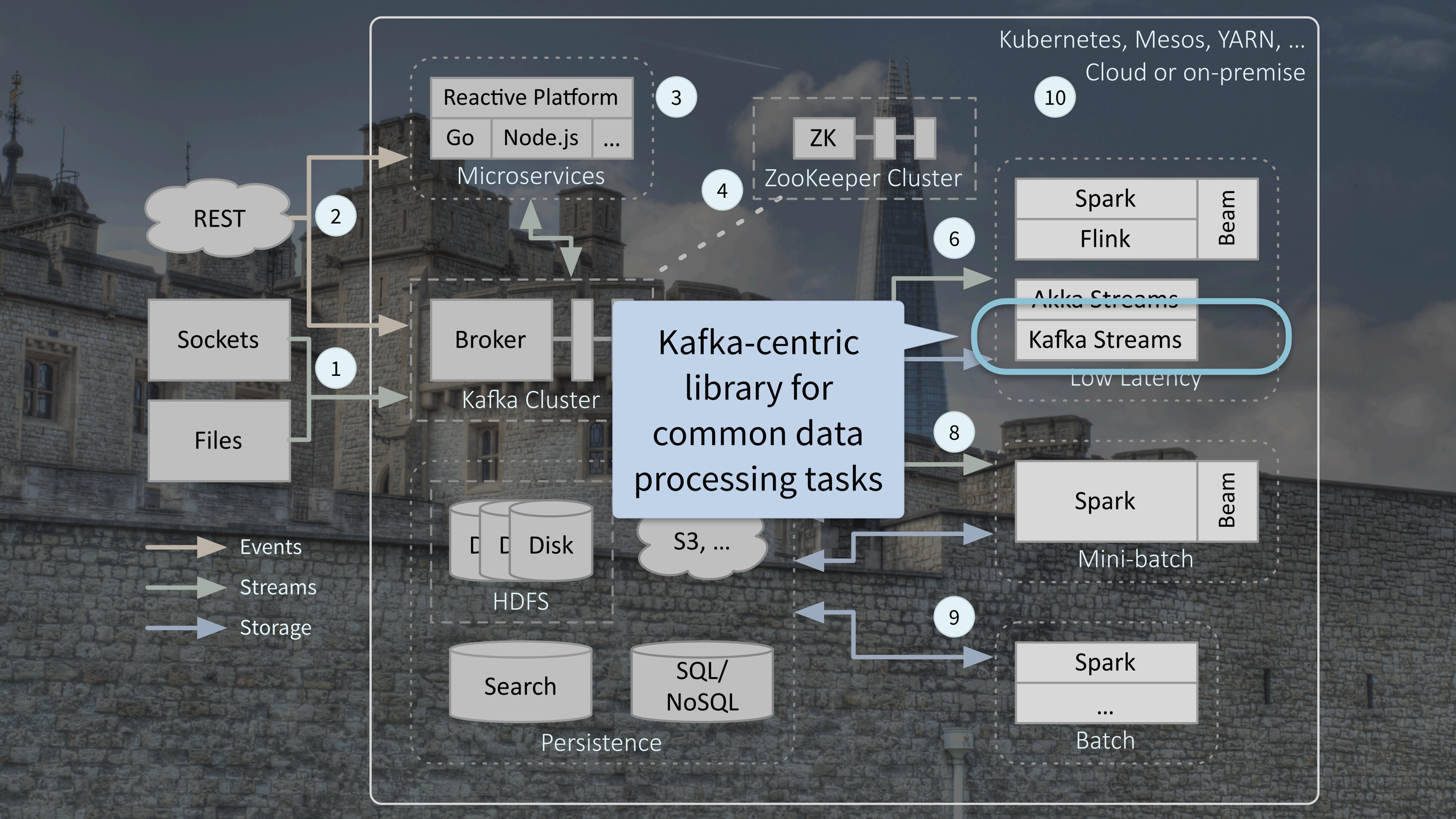

The streaming engines form two groups:
Run as distributed services
You submit jobs, they are partitioned into tasks
Kubernetes, Mesos, YARN, ...
Cloud or on-premise
Reactive Platform
Go
Node.js
..
Microservices
REST
2
Sockets
1
Files
Broker
Kafka Cluster
10
Spark
Flink
Beam
6
Akka Streams
Kafka Streams
Low Latency
8
Spark
Beam
Mini-batch
Disk
HDFS
S3, ...
9
Spark
...
Batch
Search
SQL/
NoSQL
Persistence
Events
Streams
Storage

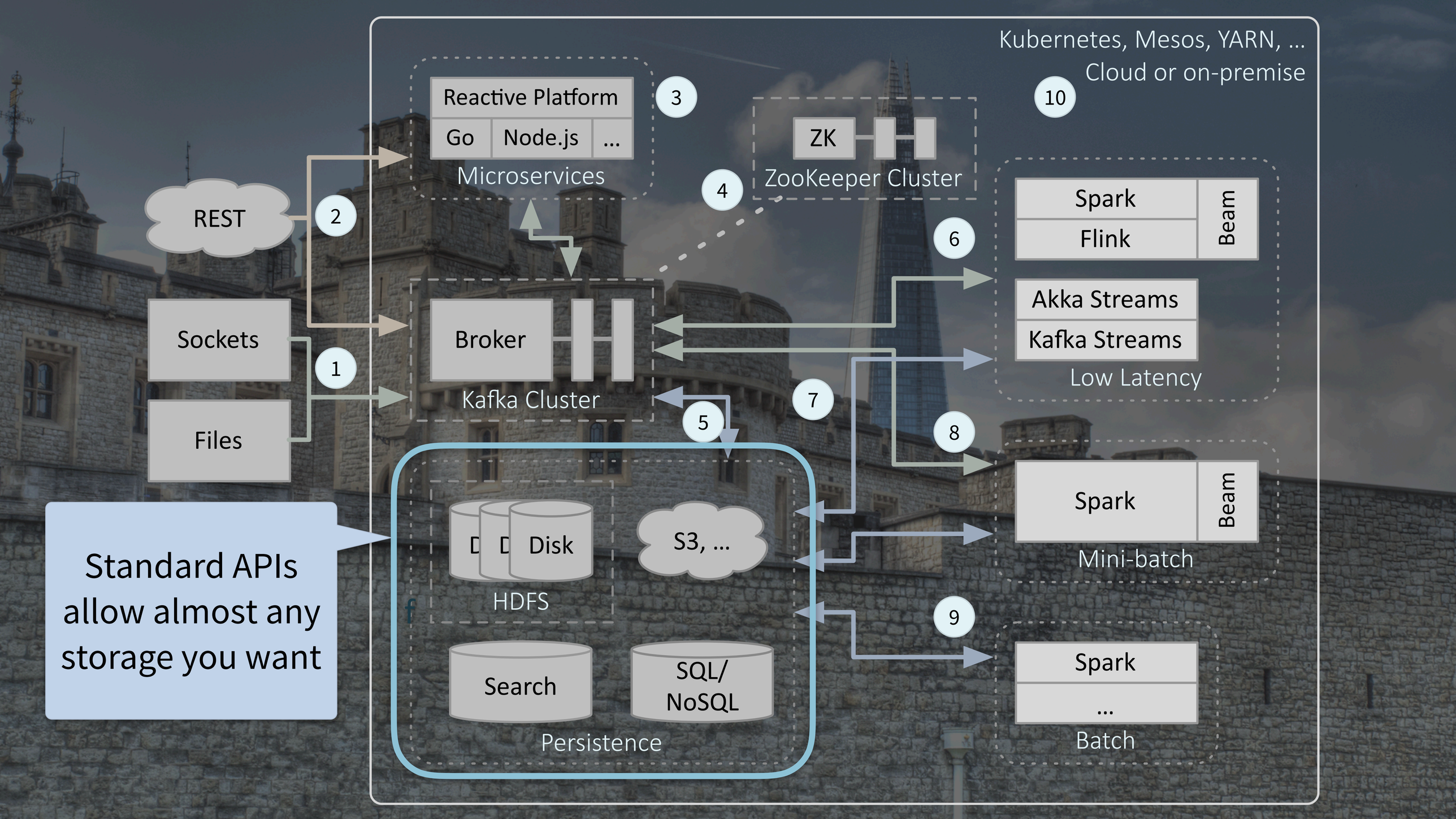

The streaming engines form two groups:
Kubernetes, Mesos, YARN, ...
Cloud or on-premise
Reactive Platform
Go
Node.js
...
Microservices
3
ZK
ZooKeeper Cluster
10
4
REST
2
Sockets
1
Files
Broker
Kafka Cluster
Libraries you embed in your microservices
6
Spark
Flink
Beam
Akka Streams
Kafka Streams
Low Latency
8
Spark
Beam
Mini-batch
Disk
HDFS
S3, ...
Search
SQL/ NoSQL
Persistence
9
Spark
...
Batch
Events
Streams
Storage

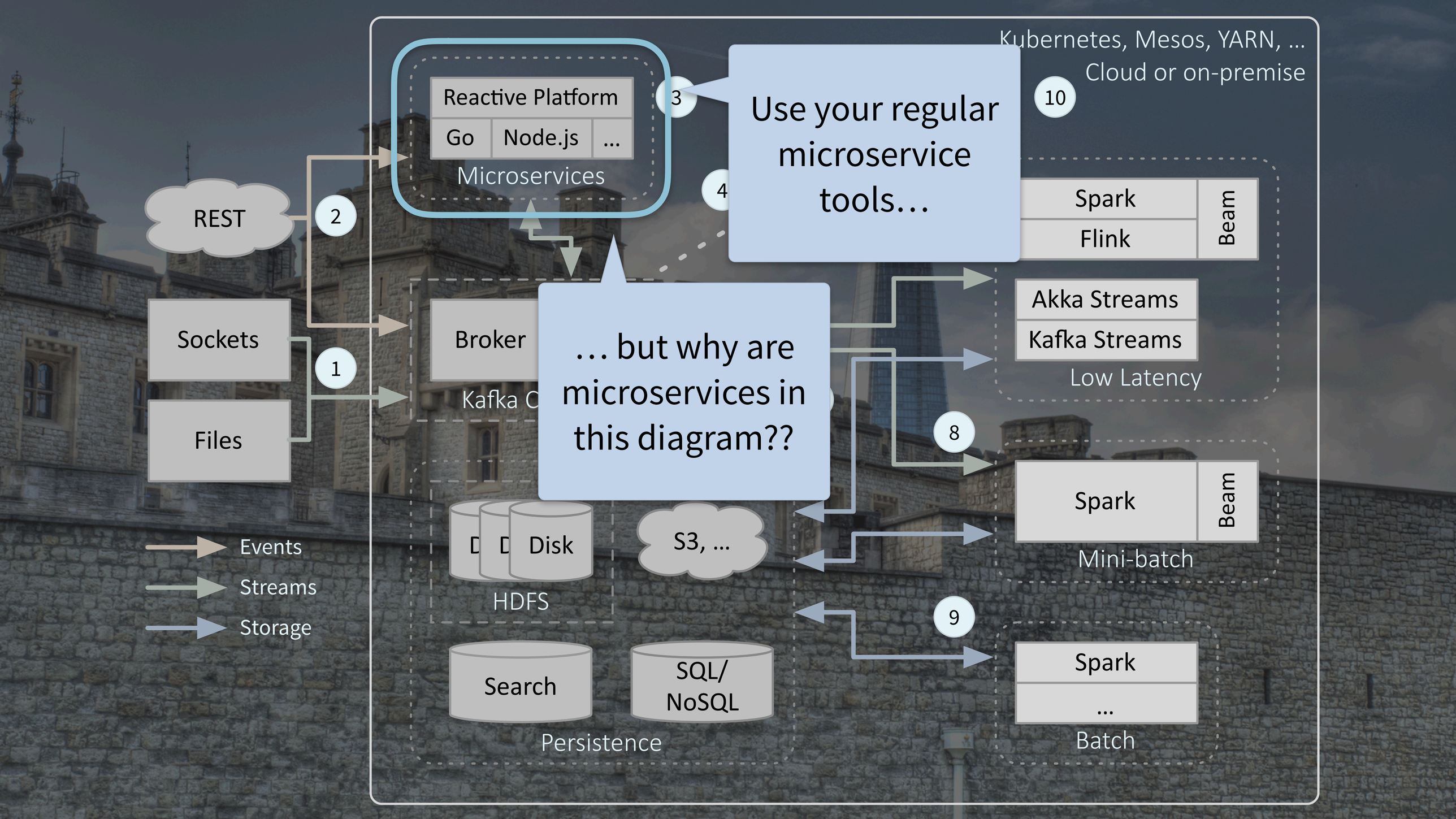

Kubernetes, Mesos, YARN, ...
Cloud or on-premise
10
Reactive Platform
Go
Node.js
..
Microservices
REST
2
Sockets
1
Files
Broker
Kafka Cluster
New, low-latency
Structured
Streaming
luster
6
Spark
Flink
Beam
Akka Streams
Kafka Streams
Low Latency
Older *mini-batch*
streaming
8
Spark
Beam
Mini-batch
D
D
Disk
HDFS
S3, ...
Full batch
support (replaced
MapReduce)
9
Spark
...
Batch
Search
Persist
Events
Streams
Storage
Rich SQL and Machine
Learning options.


Kubernetes, Mesos, YARN, ...
Cloud or on-premise
Reactive Platform
Go
Node.js
Microservices
REST
Sockets
Files
Broker
Kafka Cluster
Low latency Spark alternative. Also uses services to partition the data and tasks across a cluster
Spark
Flink
Beam
Akka Streams
Kafka Streams
Low Latency
Spark
Beam
Mini-batch
Spark
...
Batch
Disk
HDFS
S3, ...
Search
SQL/ NoSQL
Persistence
Events
Streams
1
2
3
5
6
8
9
10
Richer streaming semantics, more mature low-latency support


Kubernetes, Mesos, YARN, ...
Cloud or on-premise
10
Reactive Platform
Go
Node.js
...
Microservices
3
ZK
ZooKeeper Cluster
4
REST
2
Sockets
1
Files
Broker
Kafka Cluster
Part of the rich Akka ecosystem for general microservices
6
Spark
Flink
Beam
Akka Streams
Kafka Streams
Low Latency
8
Spark
Beam
Mini-batch
Disk
HDFS
S3, ...
9
Spark
...
Batch
Search
SQL/ NoSQL
Persistence
Events
Streams
Storage


Kubernetes, Mesos, YARN, ...
Cloud or on-premise
Reactive Platform
Go
Node.js
...
Microservices
3
ZK
ZooKeeper Cluster
10
4
REST
2
Sockets
1
Files
Broker
Kafka Cluster
Kafka-centric library for common data processing tasks
6
Spark
Flink
Beam
Akka Streams
Kafka Streams
Low Latency
8
Spark
Beam
Mini-batch
Disk
HDFS
S3, ...
9
Spark
...
Batch
Search
SQL/ NoSQL
Persistence
Events
Streams
Storage


Kubernetes, Mesos, YARN, ...
Cloud or on-premise
10
Reactive Platform
Go
Node.js
...
Microservices
3
ZK
ZooKeeper Cluster
4
REST
2
Sockets
1
Files
Broker
Kafka Cluster
5
6
Spark
Flink
Beam
Akka Streams
Kafka Streams
Low Latency
7
8
Spark
Beam
Mini-batch
9
Spark
...
Batch
Disk
HDFS
S3, ...
Search
SQL/
NoSQL
Persistence
Standard APIs
allow almost any
storage you want


Kubernetes, Mesos, YARN, ...
Cloud or on-premise
Reactive Platform
Go
Node.js
...
Microservices
3
10
Use your regular microservice tools…
4
REST
2
Sockets
1
Files
Broker
Kafka C
… but why are microservices in this diagram??
Spark
Flink
Beam
Akka Streams
Kafka Streams
Low Latency
8
Spark
Beam
Mini-batch
Disk
S3, ...
HDFS
9
Search
SQL/
NoSQL
Persistence
Spark
...
Batch
Events
Streams
Storage

Why Microservices in Fast Data?

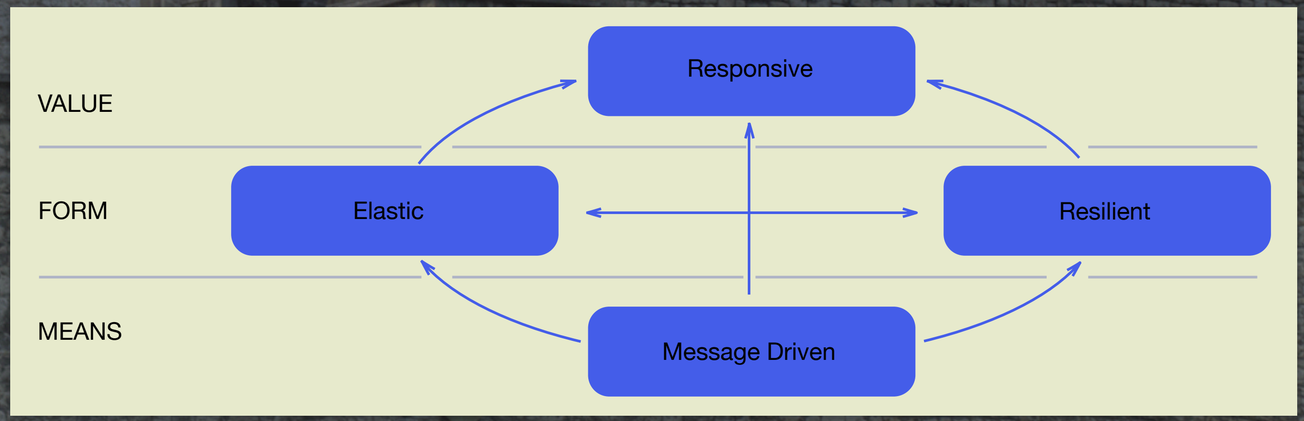
1. The trend is to run everything in big clusters using Kubernetes or Mesos
   - In the cloud or on-premise

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



https://www.reactivemanifesto.org/

Why Microservices in Fast Data?

4. This leads to our last major point…

Organizational Impact

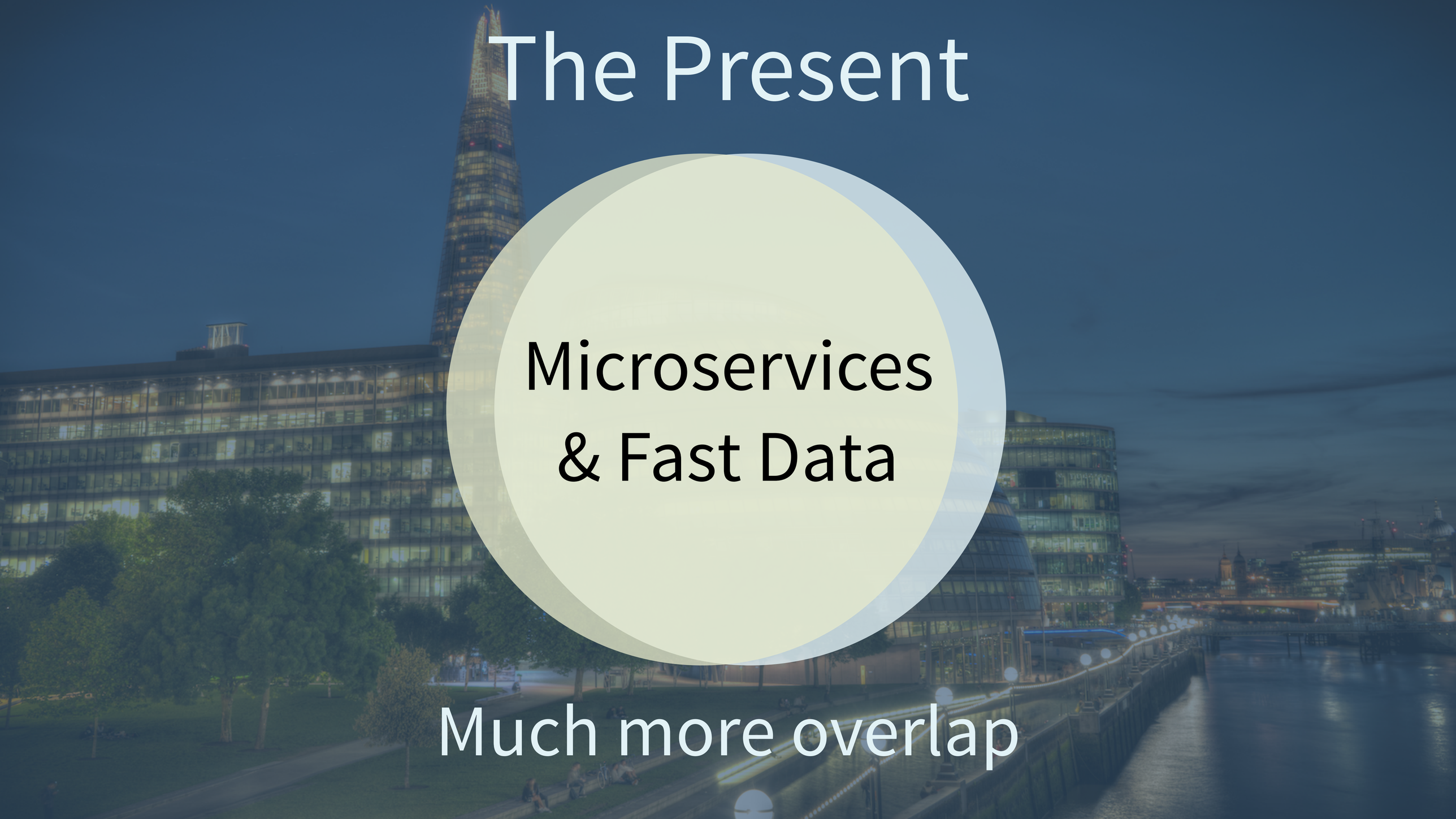
Organizational Impact

- 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

Some overlap in concerns, architecture

The Present



Much more overlap

The Future?

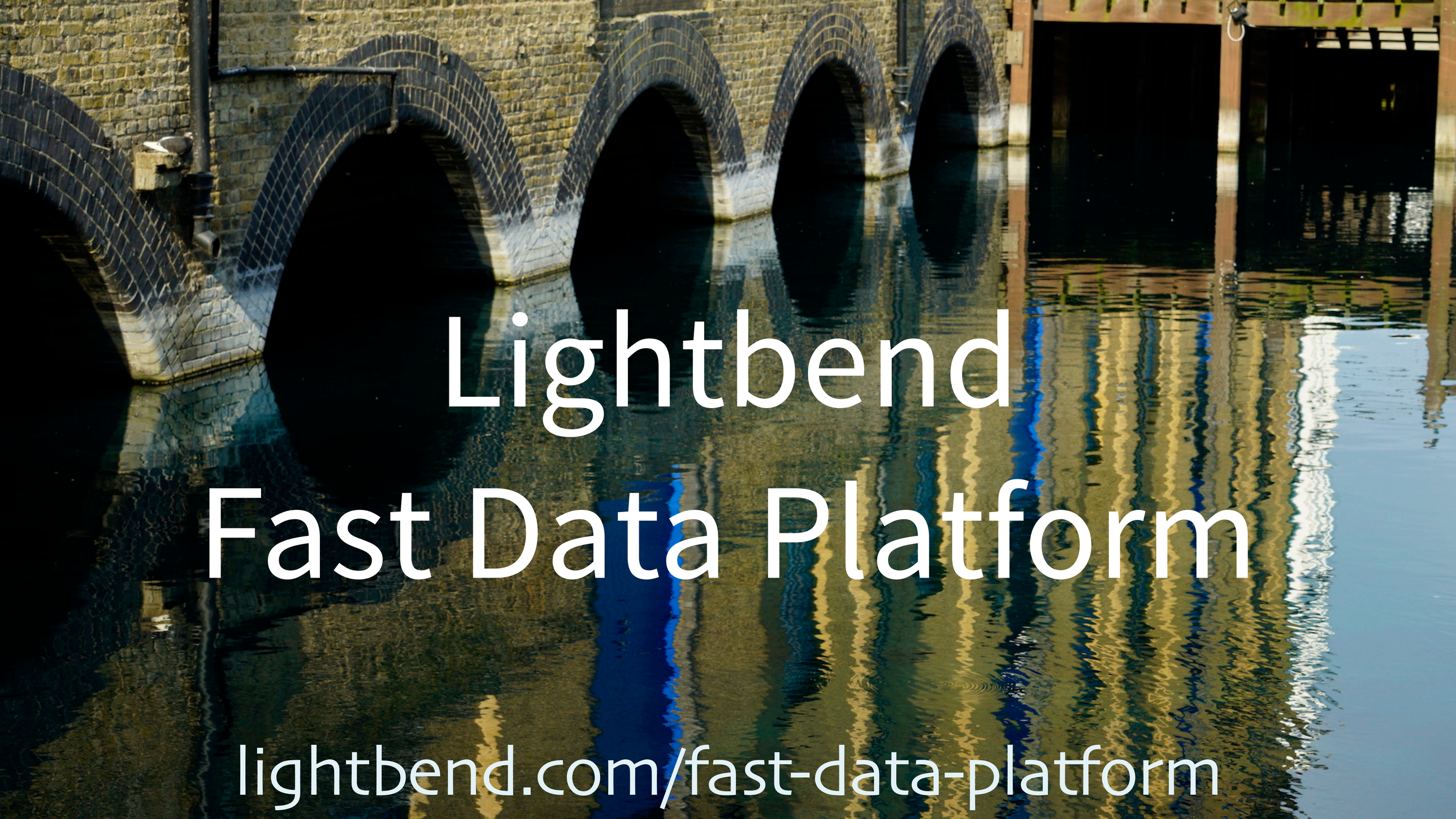
Why? Since streams process data incrementally, there is less need for large-scale tools like Spark, Flink

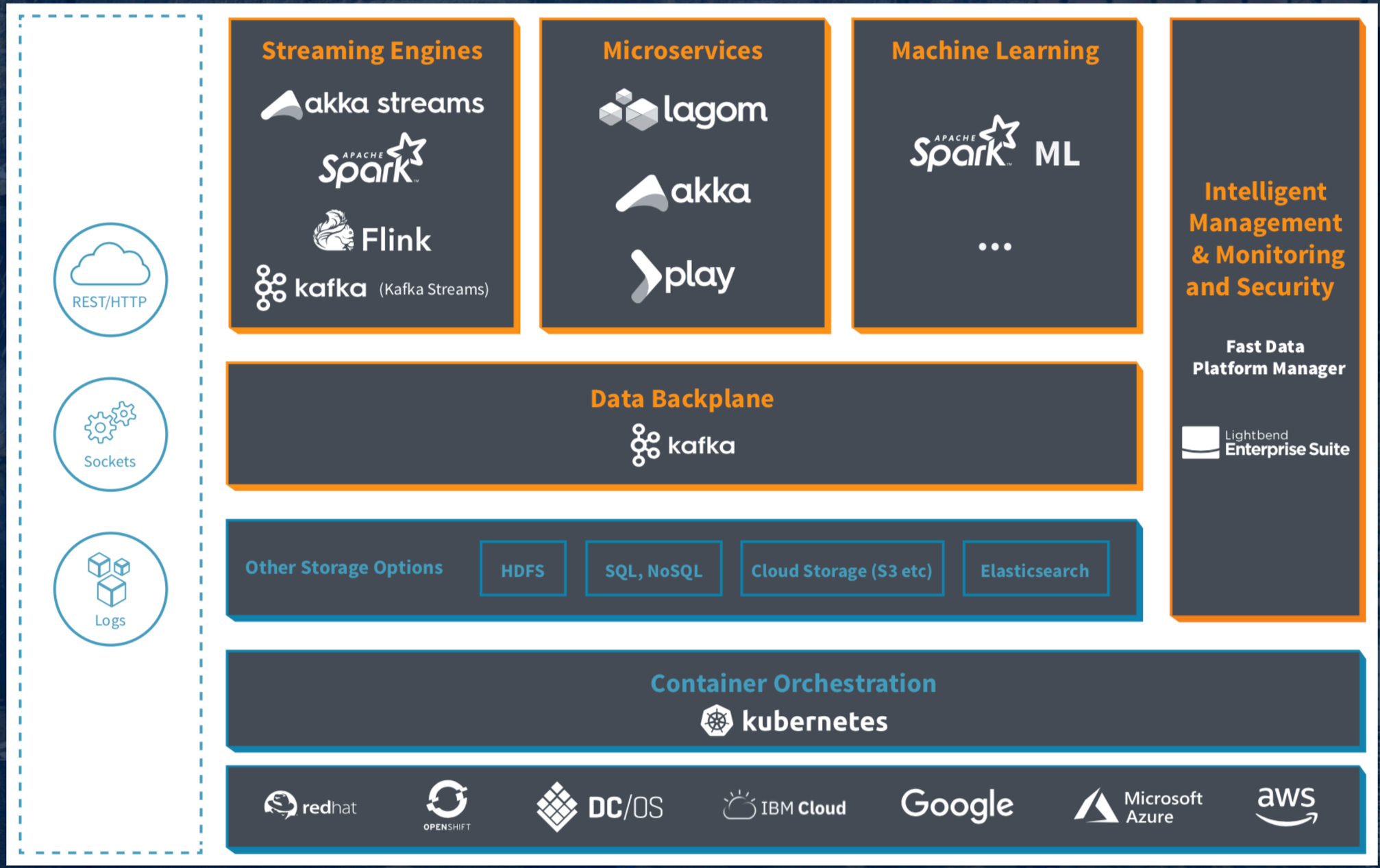
… and using microservices for everything simplifies development, deployment, and operations

Microservices *for* Fast Data

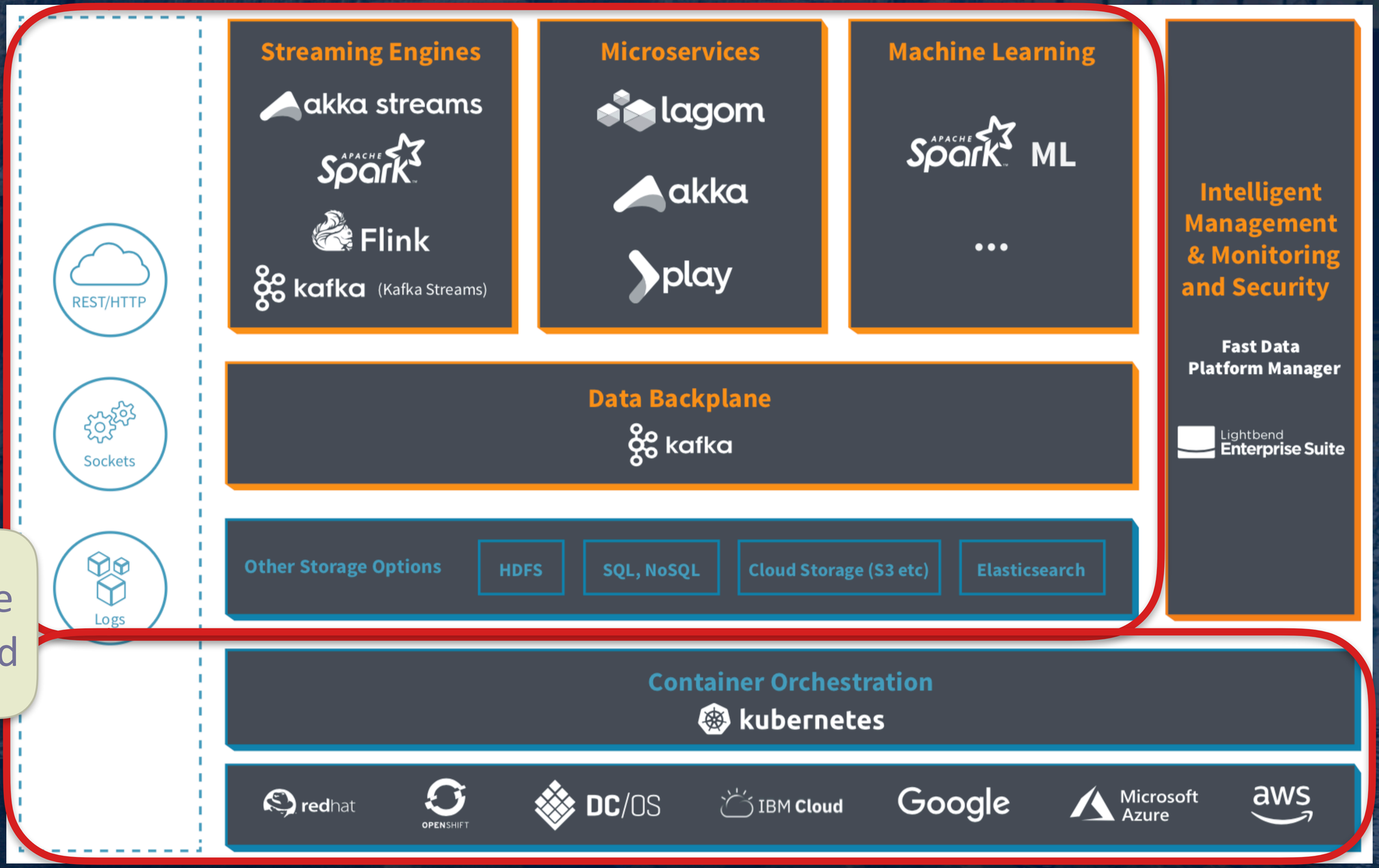
Unclear if this helps to integrate with the data science workflows we love.

Much more microservice focused?


Lightbend
Fast Data Platform
lightbend.com/fast-data-platform

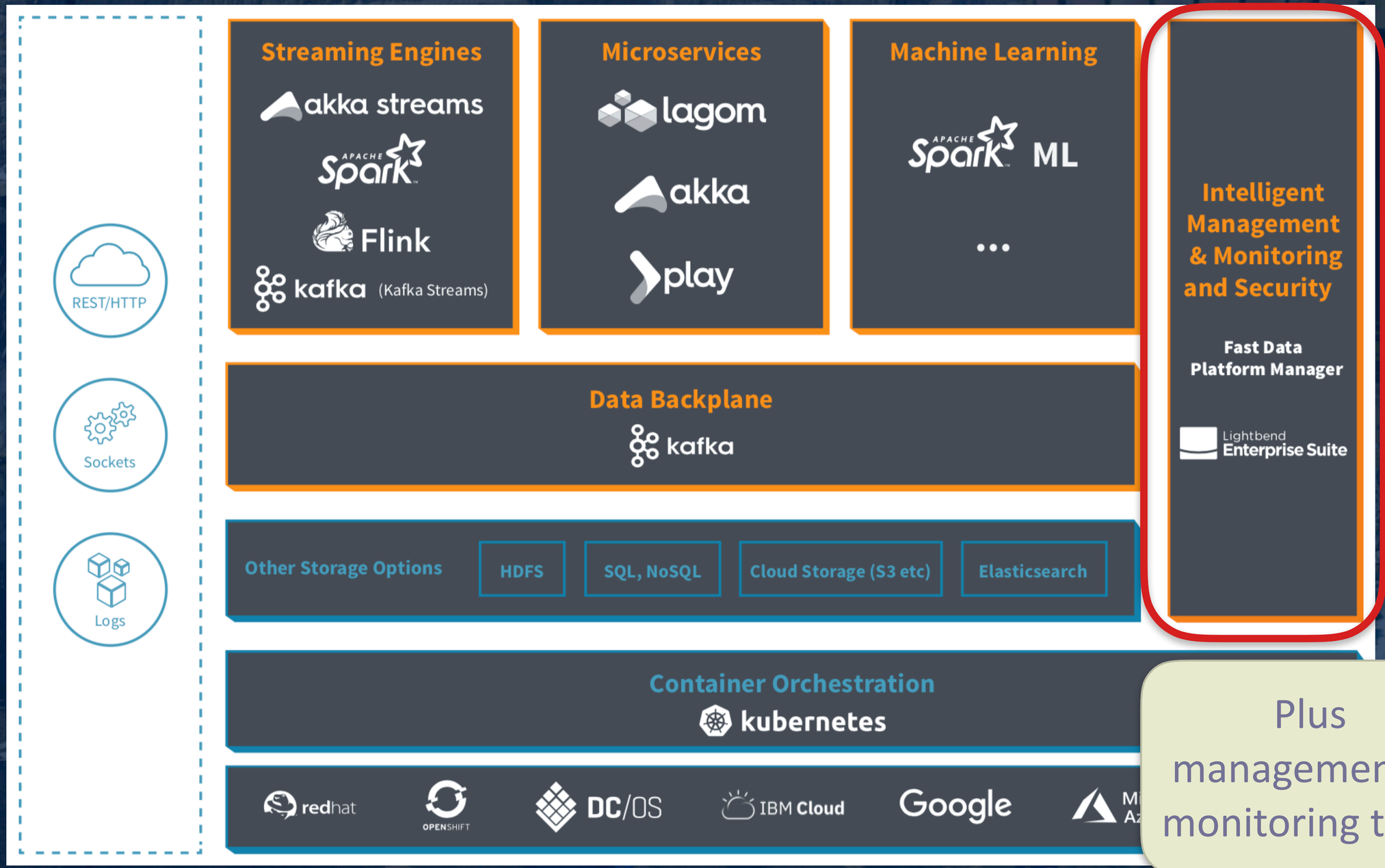


lightbend.com/fast-data-platform



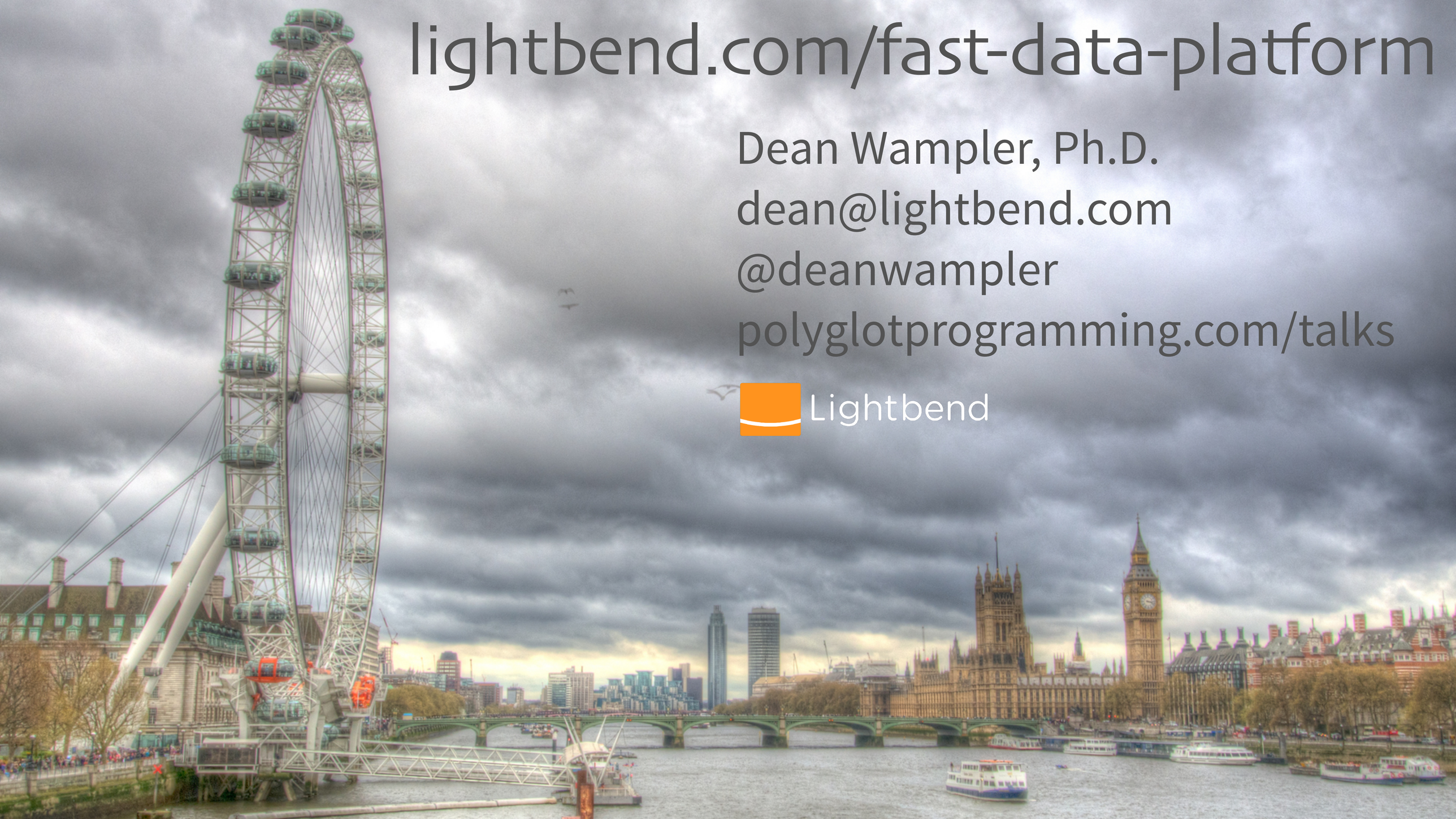
What we discussed

lightbend.com/fast-data-platform



Plus management & monitoring tools

lightbend.com/fast-data-platform



lightbend.com/fast-data-platform

Dean Wampler, Ph.D.
dean@lightbend.com
@deanwampler
polyglotprogramming.com/talks

Lightbend