

Executive Briefing:

What it takes to use machine

learning in fast data pipelines

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

Data Streaming, in General

lbnd.io/fast-data-ebook

O'REILLY®

Compliments of
Lightbend

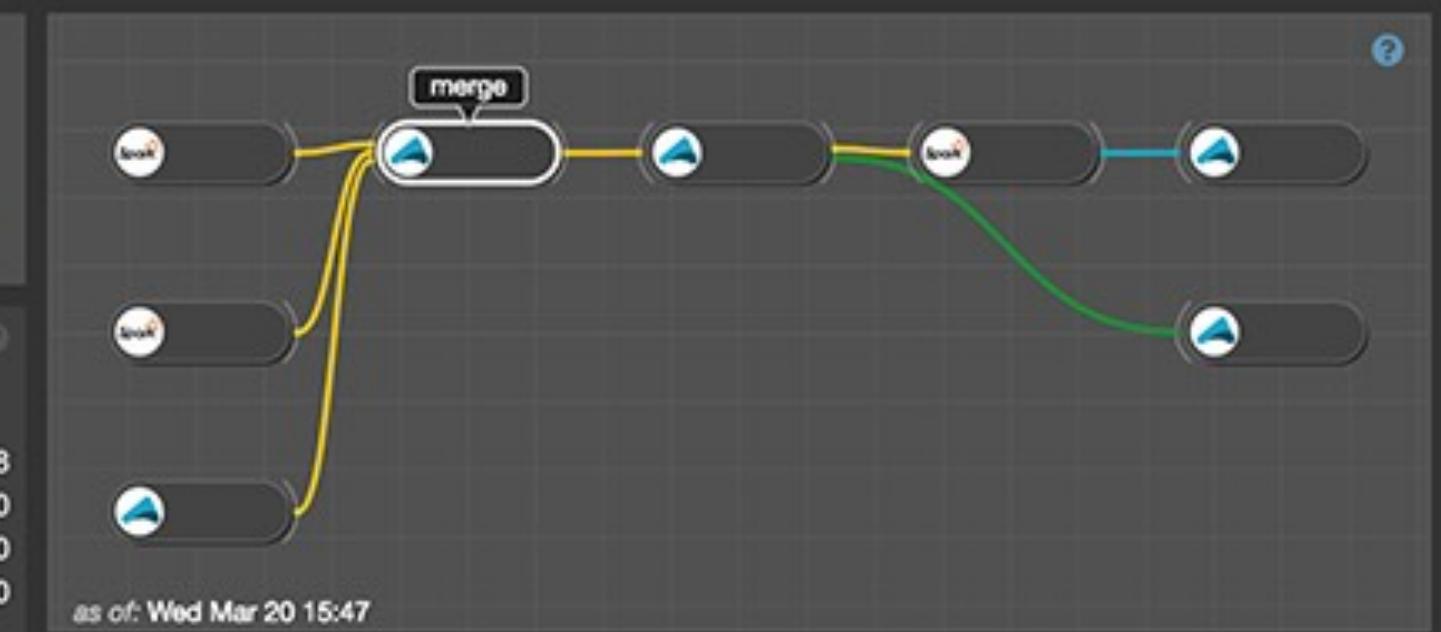
Fast Data Architectures for Streaming Applications

Getting Answers Now from
Data Sets That Never End

2nd
Edition



Dean Wampler, PhD



Application Details

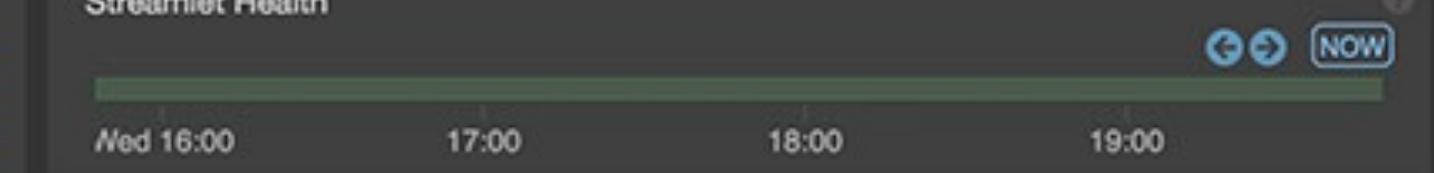
Streamlet Current Health

Healthy	8
Warning	0
Critical	0
Unknown	0

Streamlet Health Events

cdr-validator	---
cdr-aggregator	---
merge	---
console-egress	---
error-egress	---
cdr-generator1	---
cdr-generator2	---
cdr-ingress	---

Streamlet Health



Selection: merge

Diagnostics

Ports

Streamlet Monitors

SORT BY First unhealthy ▾

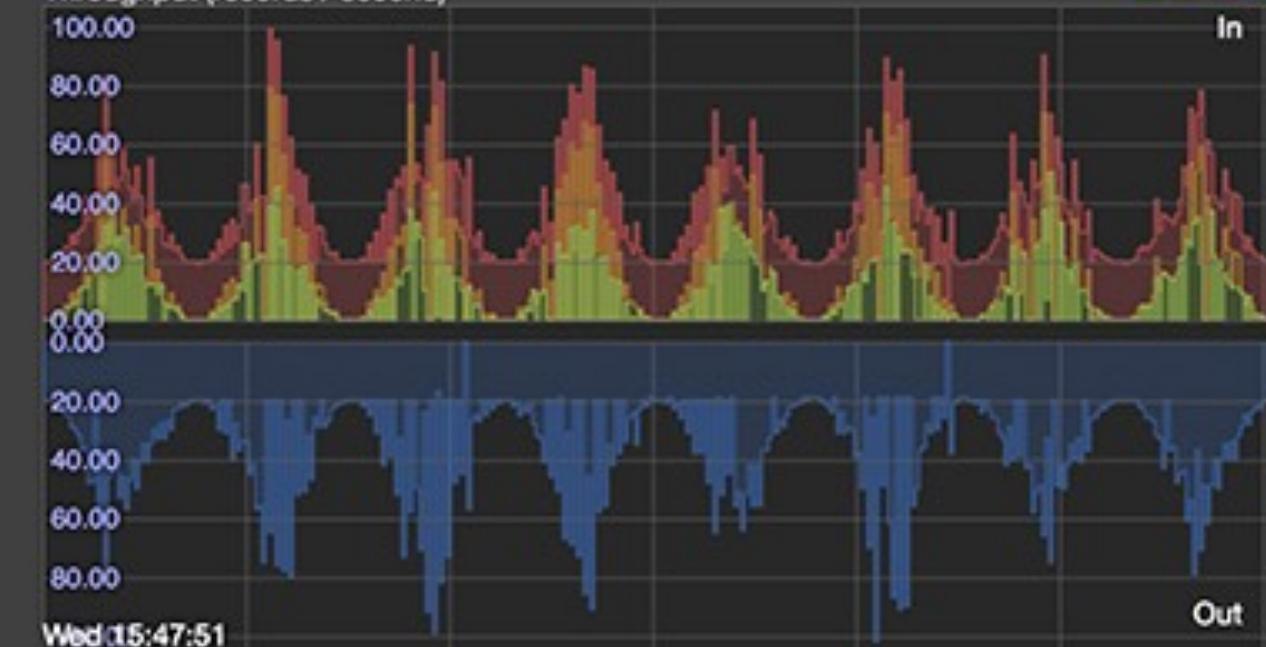
kafka_consumer_lag

kafka_consumer_throughput

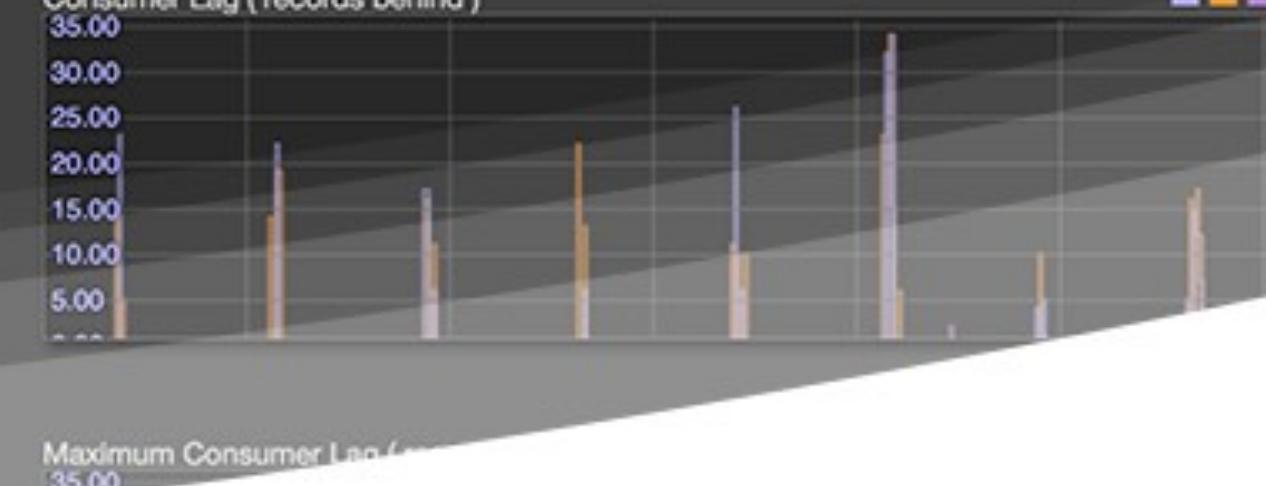
kafka_producer_throughput

Metrics

Throughput (records / second)



Consumer Lag (records behind)



lightbend.com/lightbend-pipelines-demo



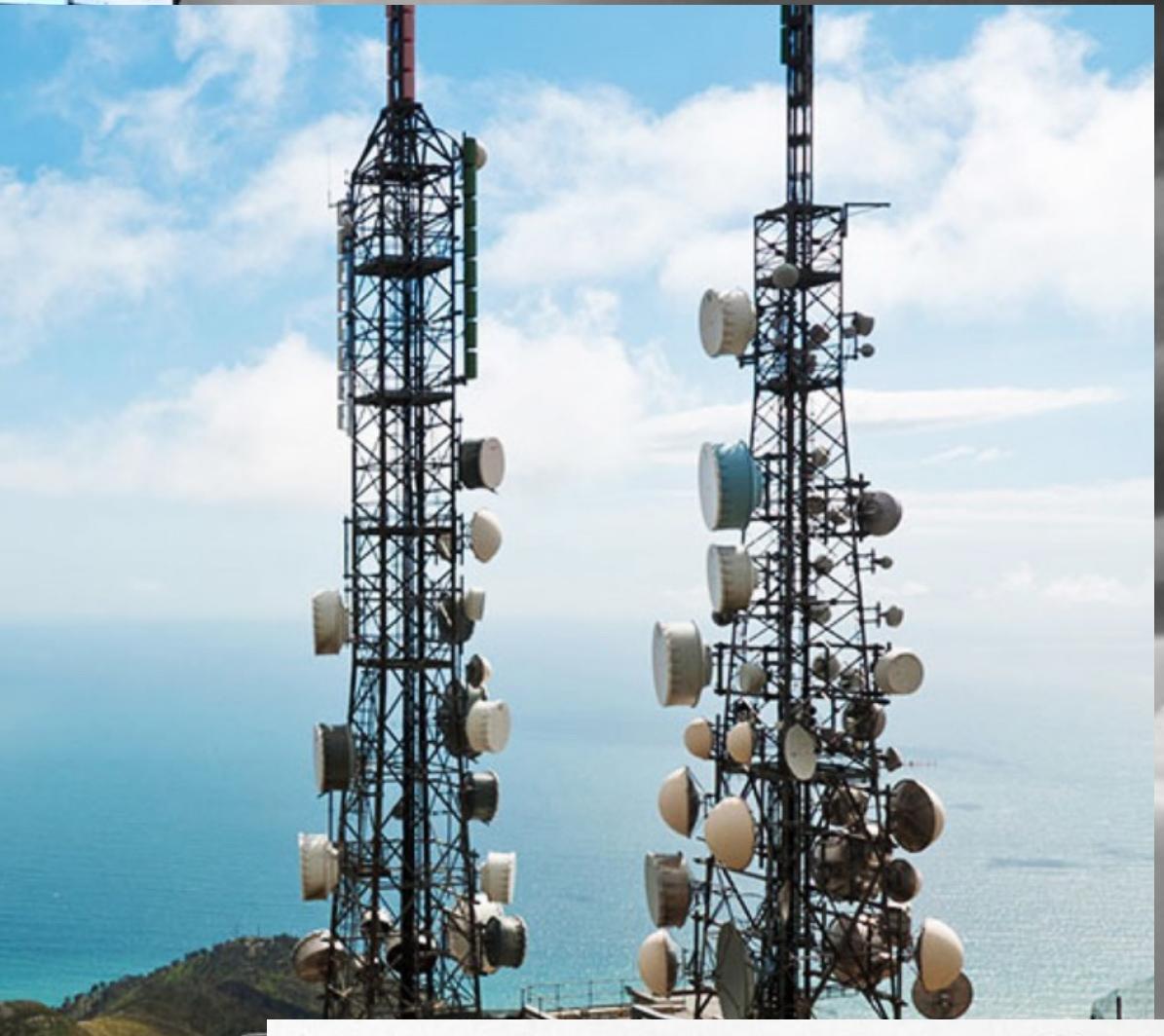
What We'll Discuss

- Batch vs. streaming... and why
- Data science vs. data engineering
- Serving models in production
- CI/CD Systems for ML
- Example architecture
- Updating Models in Production



Batch vs. streaming... and why

Telecom



Finance



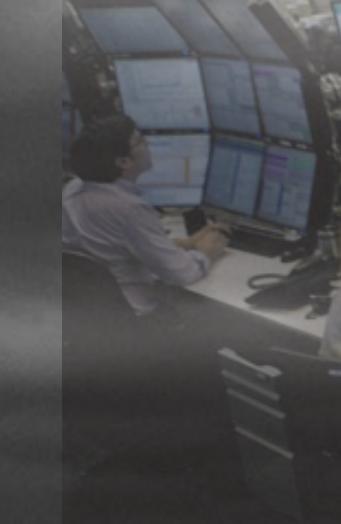
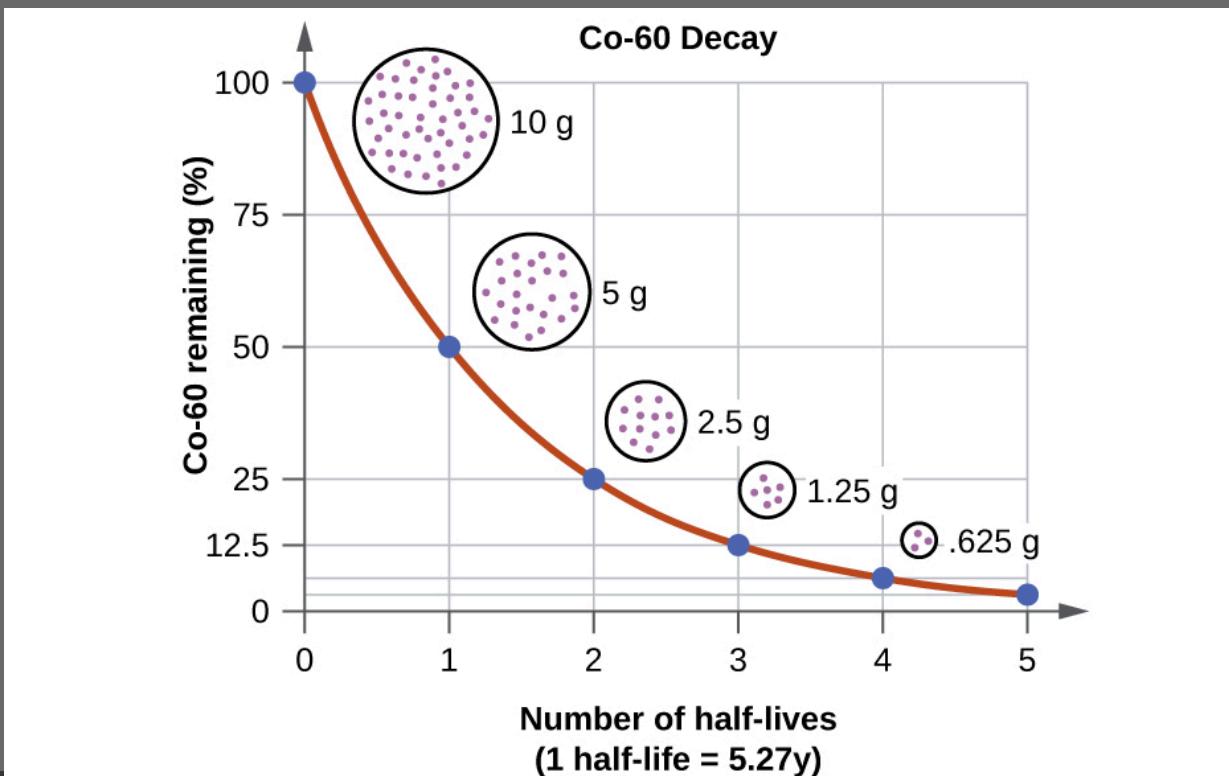
Energy

... and IoT

State of the art phone!



Information value has a half life; it decays with time





Data Science vs. Data Engineering



Data Science toolbox



Software
Engineering toolbox

Data Scientists

- Comfortable with uncertainty
- Less process oriented
 - Iterative, experimental

Data Engineers

- Uncomfortable with uncertainty
- Process oriented
 - Agile Manifesto
 - ... which does not mention data!

<https://derwen.ai/s/6fqt>

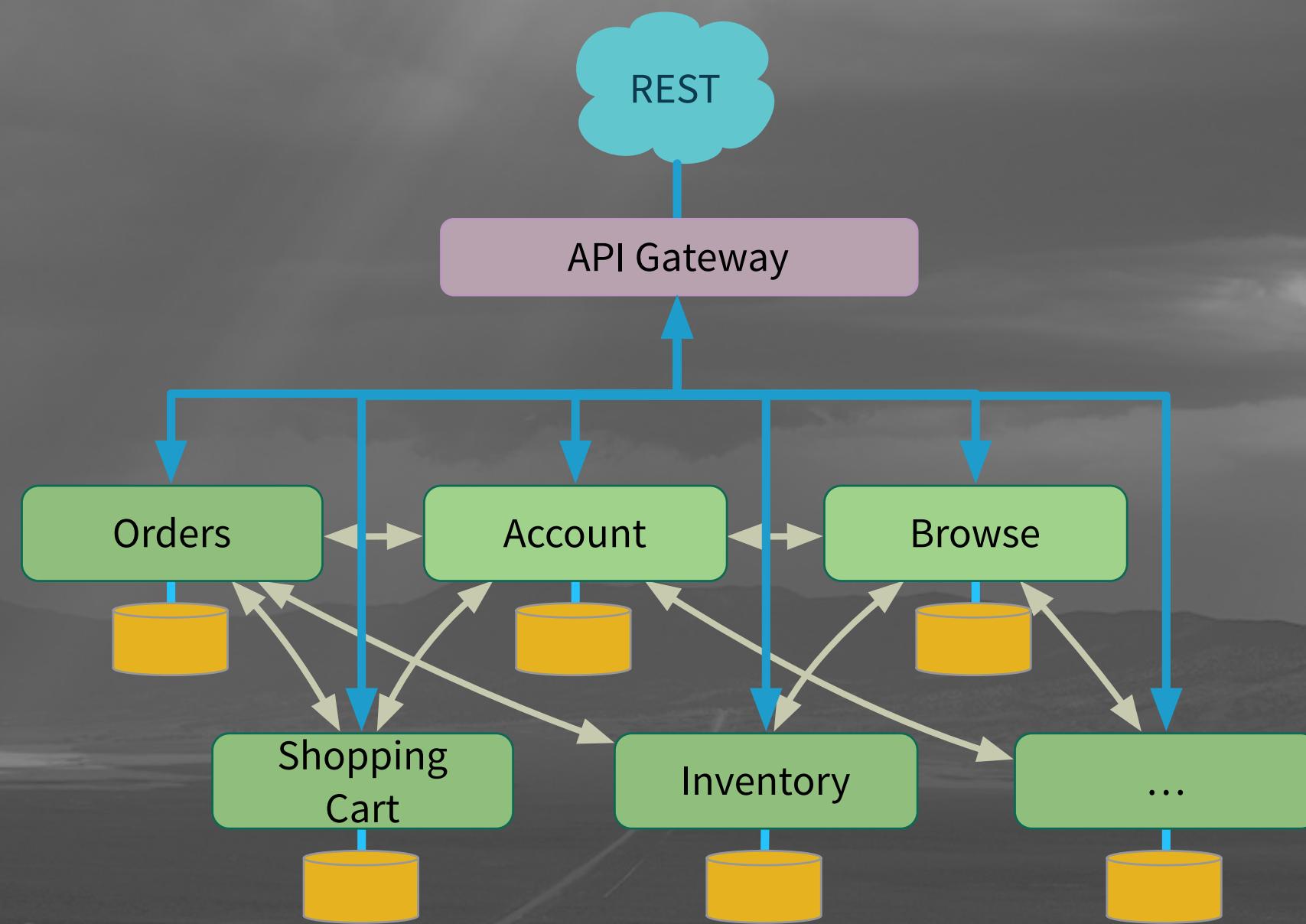


Streaming Imposes New Requirements

If you run something long enough, all rare problems eventually happen!

- Reliability - fault and “surprise” tolerant
- Availability - “always on”
- Low latency - for some definition of “low”
- Scalability - up and down
- Adaptability - ideally without restarts

Reminds Me of Microservices



Serving Models in Production



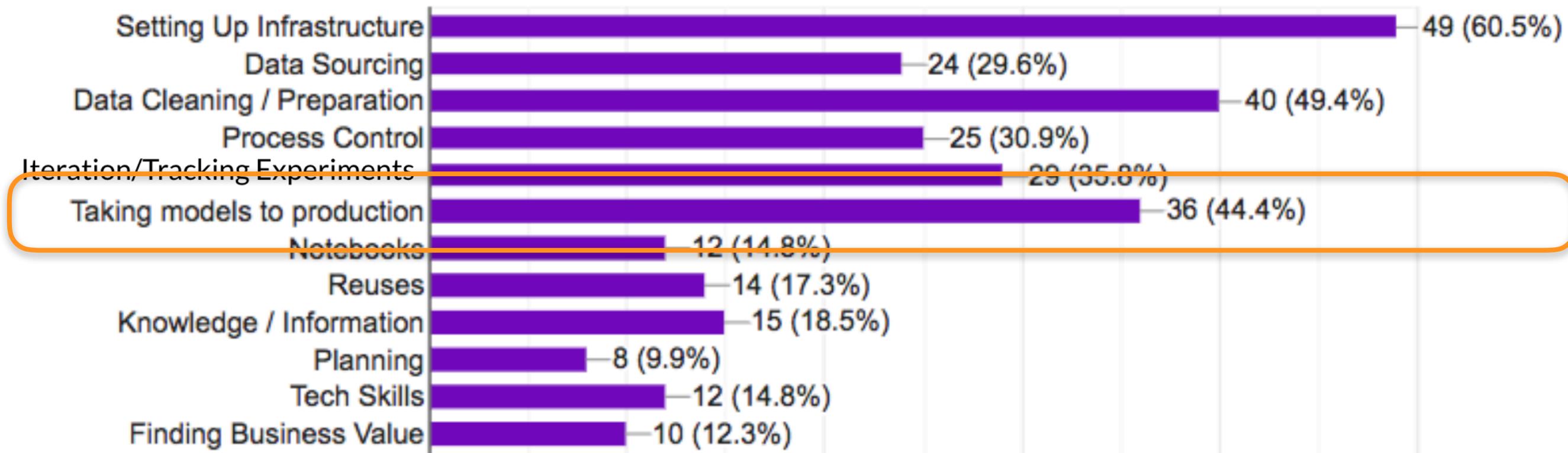
Lightbend

@deanwampler

A Recent Kubeflow User Survey

What are the major pain points in your ML workflows today? Check all that apply.

81 responses



Kuberflow User Survey - 1Q2019

Lack of Tool/Process Integration

- ~60% worry about missed opportunities
- ~50% worry about loss of data team productivity
- ~45% worry about slow time-to-market
- ~40% worry about customer dissatisfaction

Can You Answer this Question?

- Why did the model reject that loan application?

First, which model was it?

- Which version of the model was used?
- How was it trained?
- When was this model deployed?
- ...

CI/CD Process Required (1/3)

- Version control - for models and code
- Automation - builds, tests, other quality checks, artifact management & delivery
- Necessary for reproducibility

CI/CD Processes Required (2/3)

- Auditing
 - Which model used to score this record?
 - Which records used to train this model?
 - Who accessed this model and when?

CI/CD Processes Required (2/3)

- Auditing
 - Which model used to score this record?
 - Which records used to train this model?
 - Who accessed this model and when?

Models Are Data

CI/CD Processes Required (2/3)

- Auditing
 - Which model used to score this record?
 - Which records used to train this model?
 - Who accessed this model and when?

GDPR - What if a customer asks you to delete their data? Do you also delete the models trained with that data?

CI/CD Processes Required (3/3)

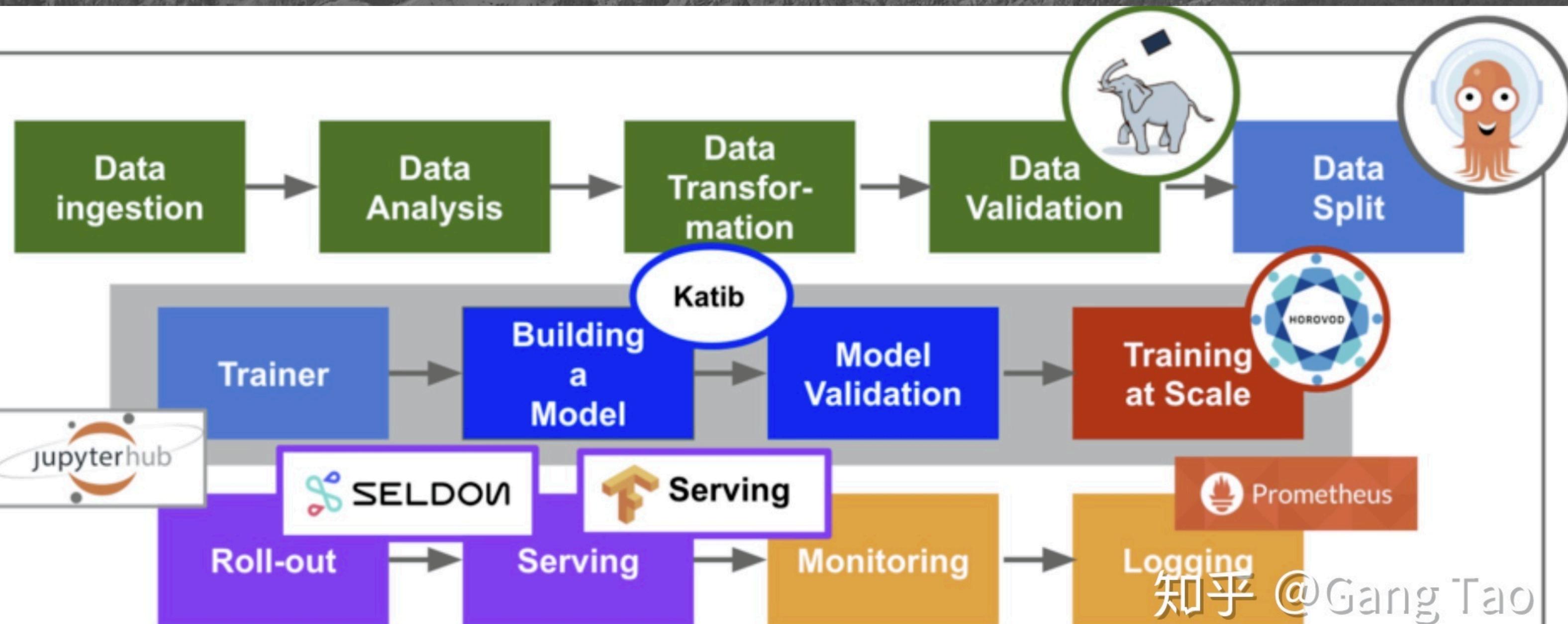
- Monitoring
 - Resource utilization changes?
 - Quality metrics:
 - Match performance during training?
 - Concept drift?



CI/CD Systems for ML?

- Kubeflow - for Kubernetes
- SageMaker - for AWS users
- MLFlow - from the Spark community
- ... plus emerging vendors

Systems - Kubeflow

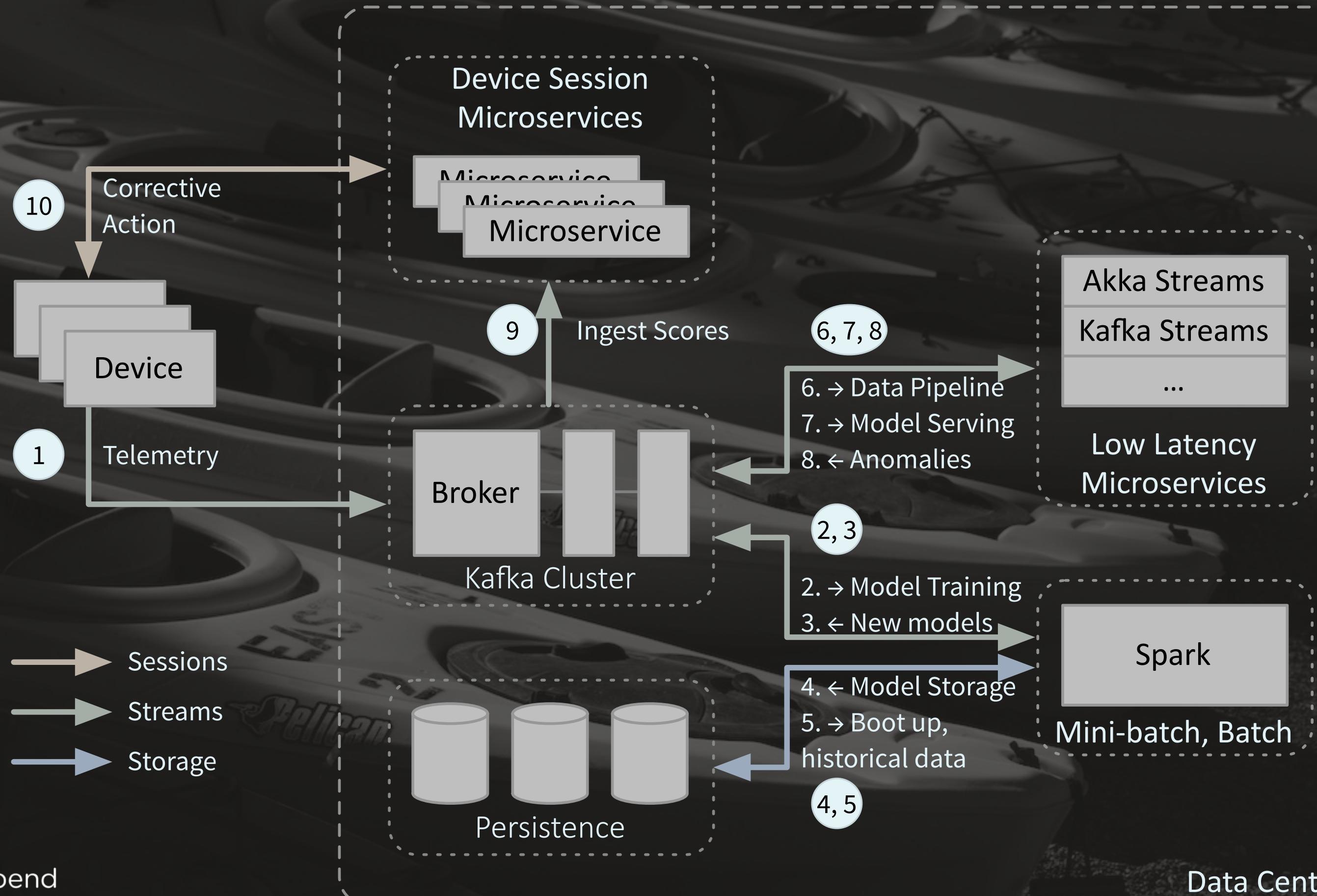


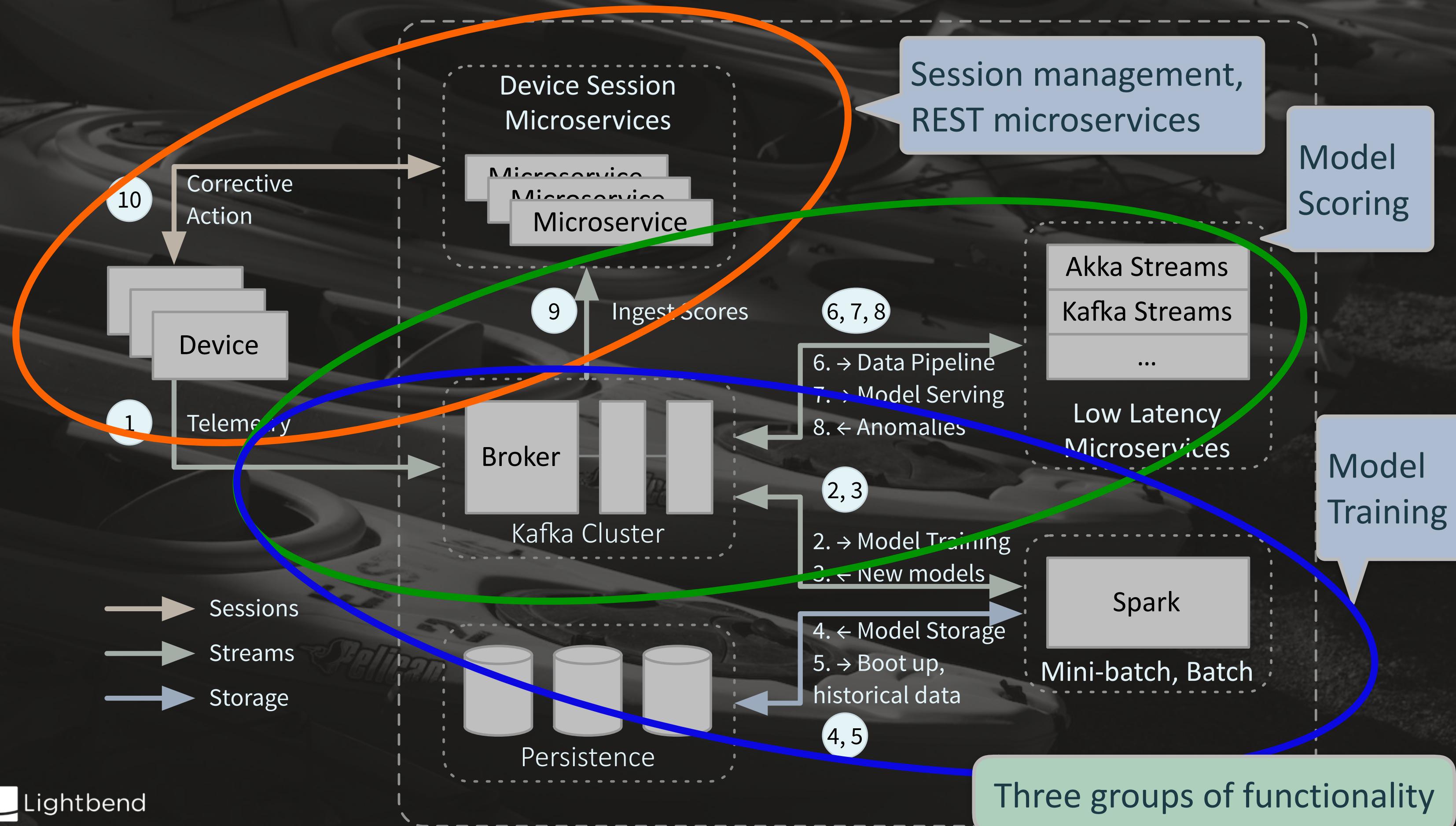
An Example

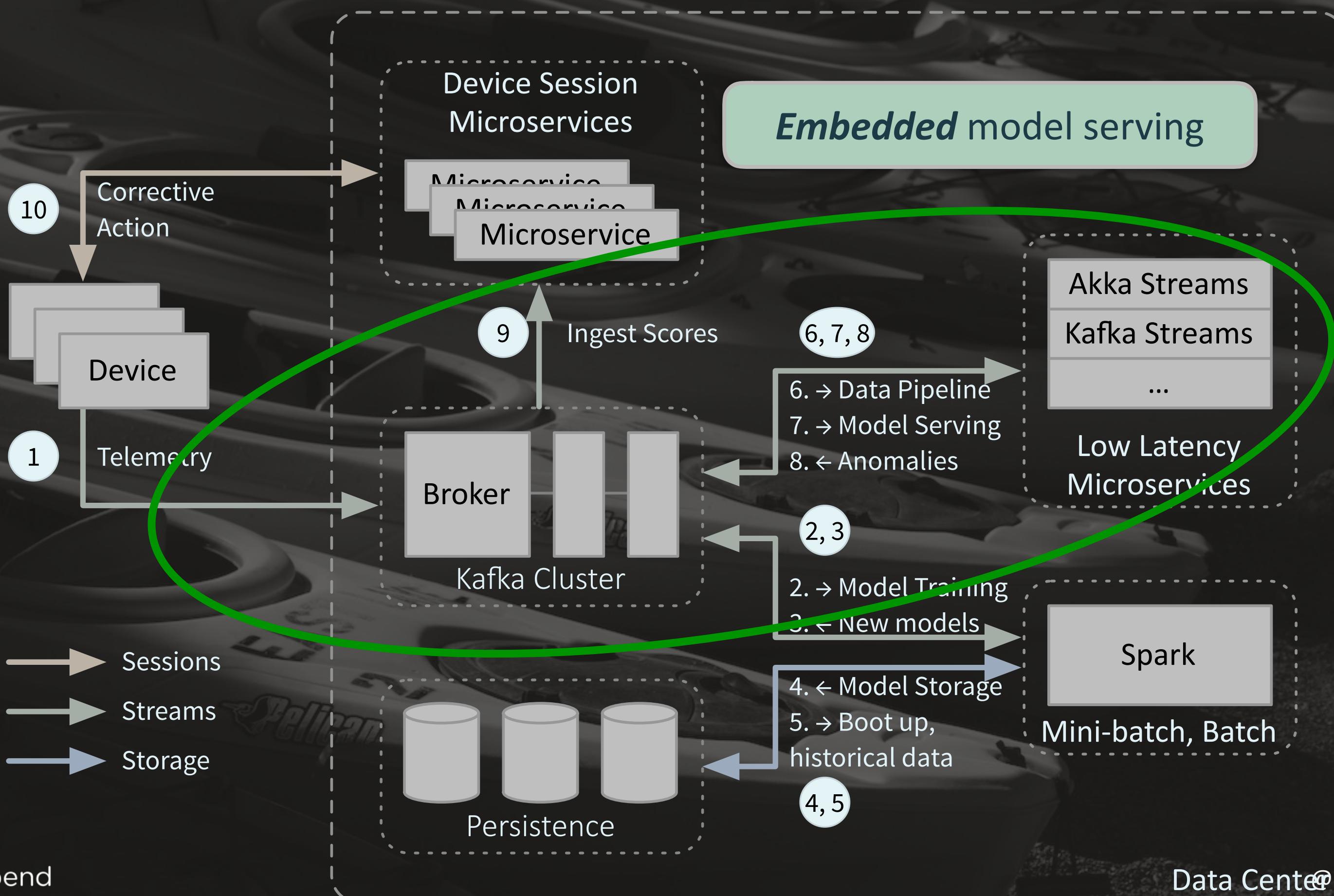


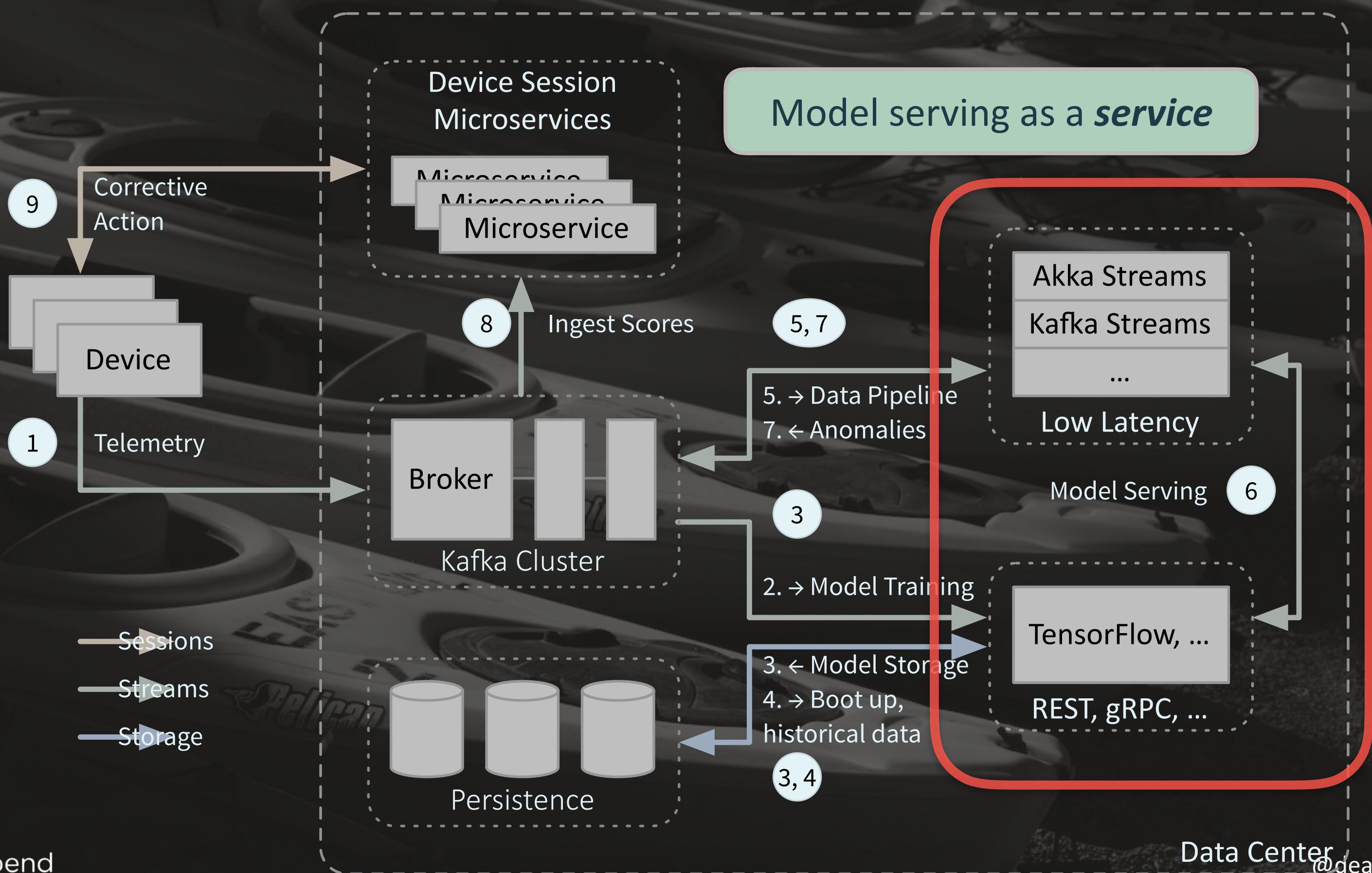
An Example

Timely Information
Integrated with
Your APPS

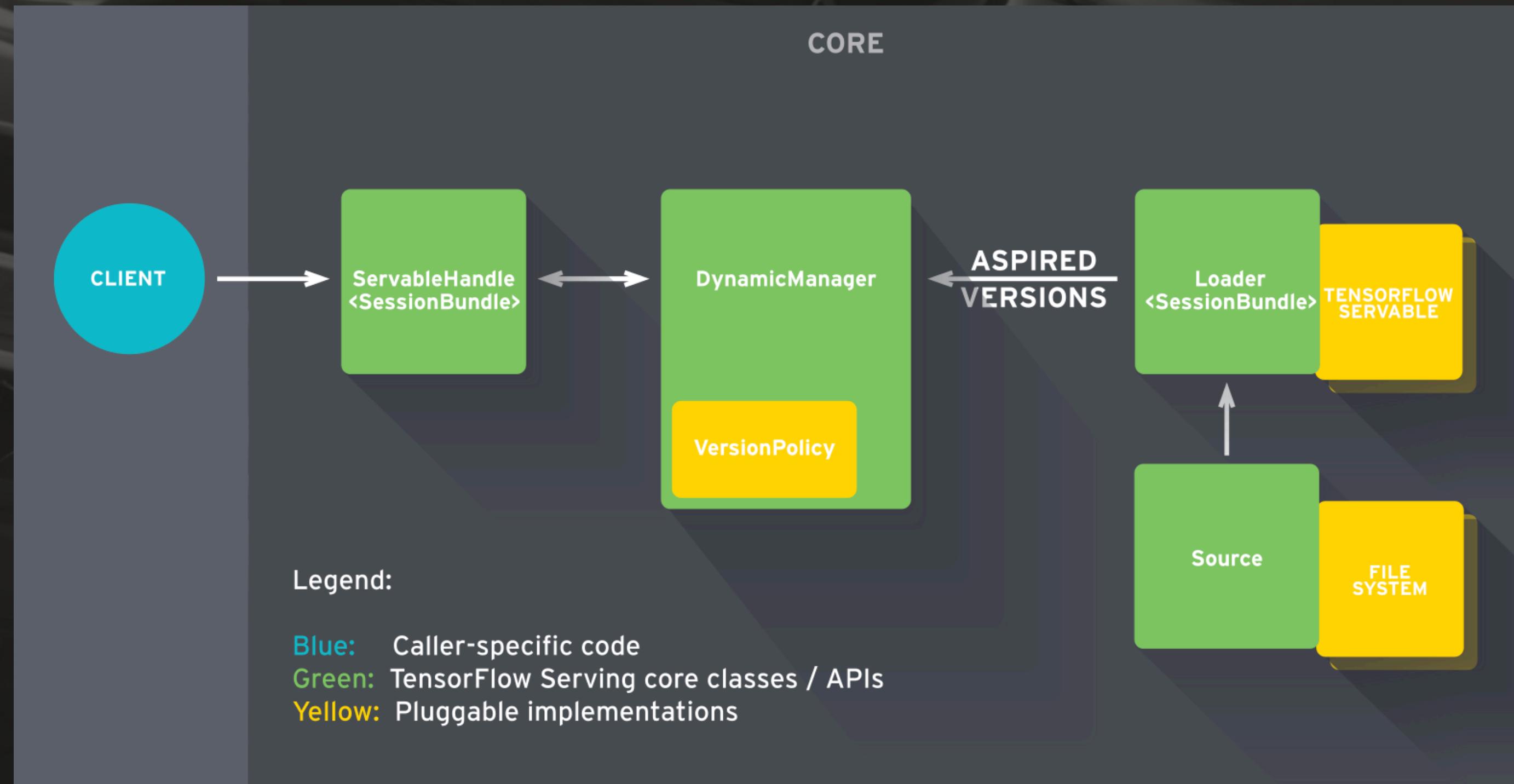






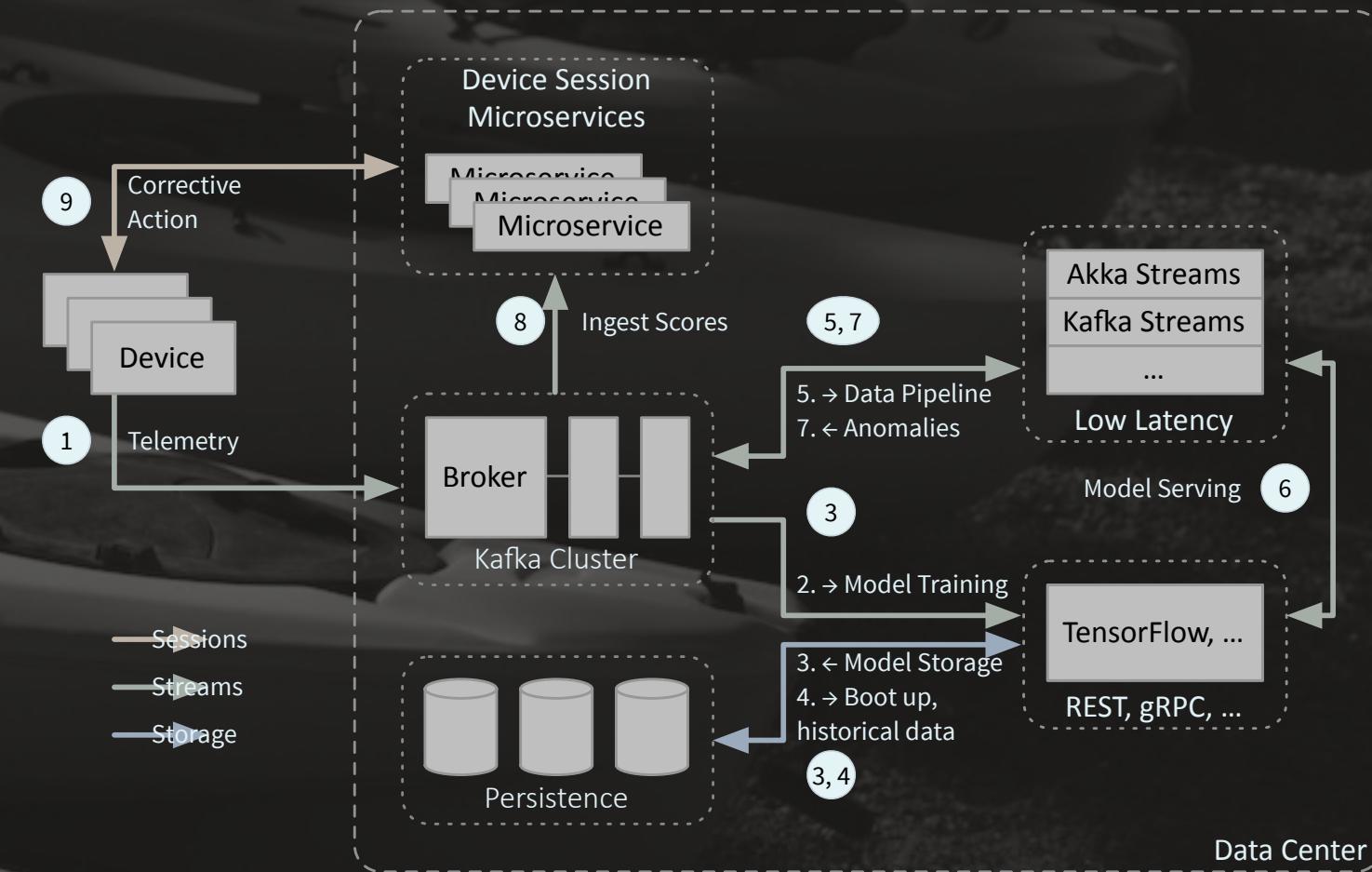


TensorFlow Serving



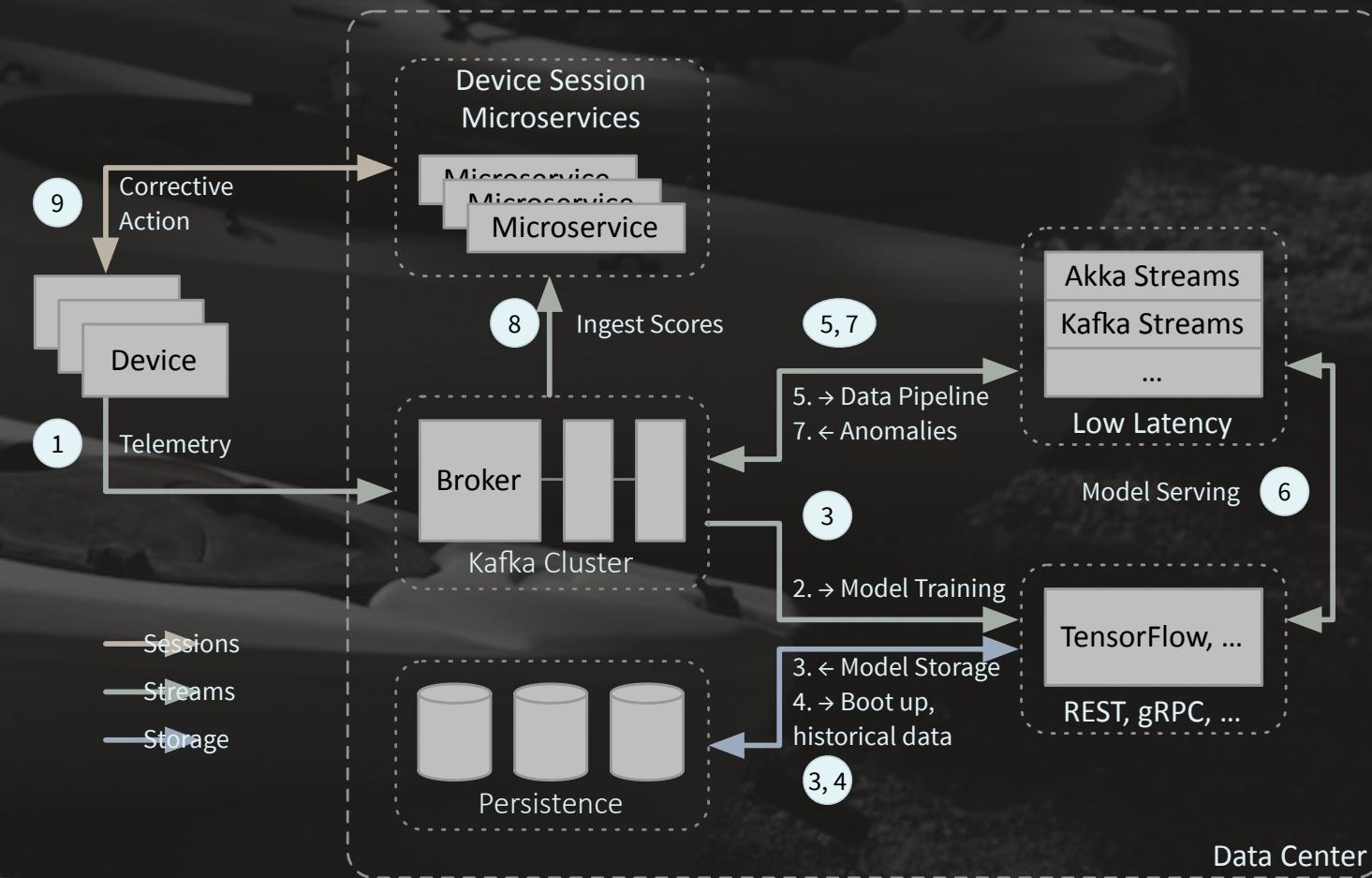
Model Serving as a Service

- Pros:
 - A familiar integration pattern
 - Decouples “concerns”: AI tools, scalability, upgradability, ...



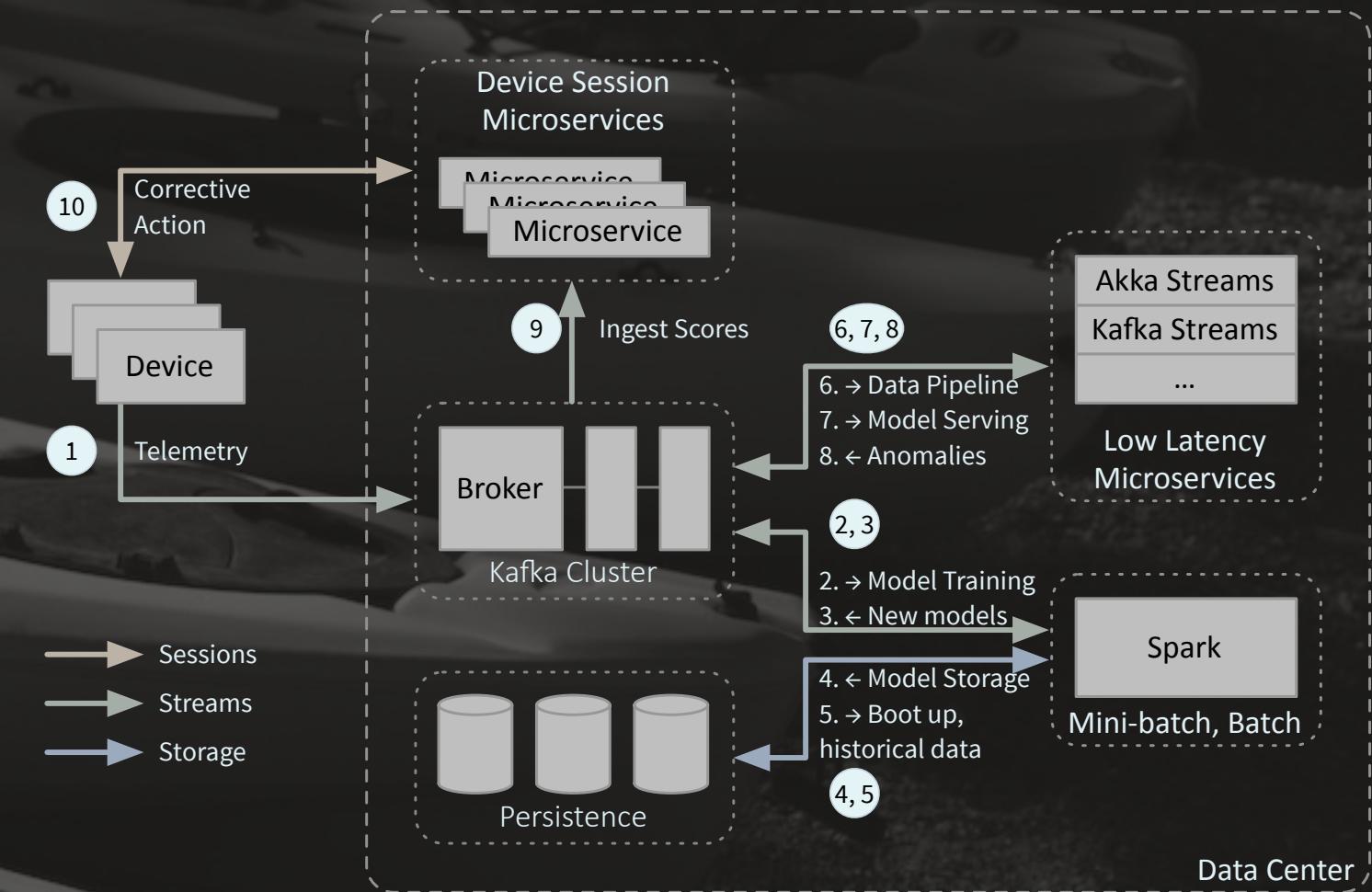
Model Serving as a Service

- Cons:
 - Overhead of invocation, e.g., REST
 - ML Pipeline becomes a unique production work flow



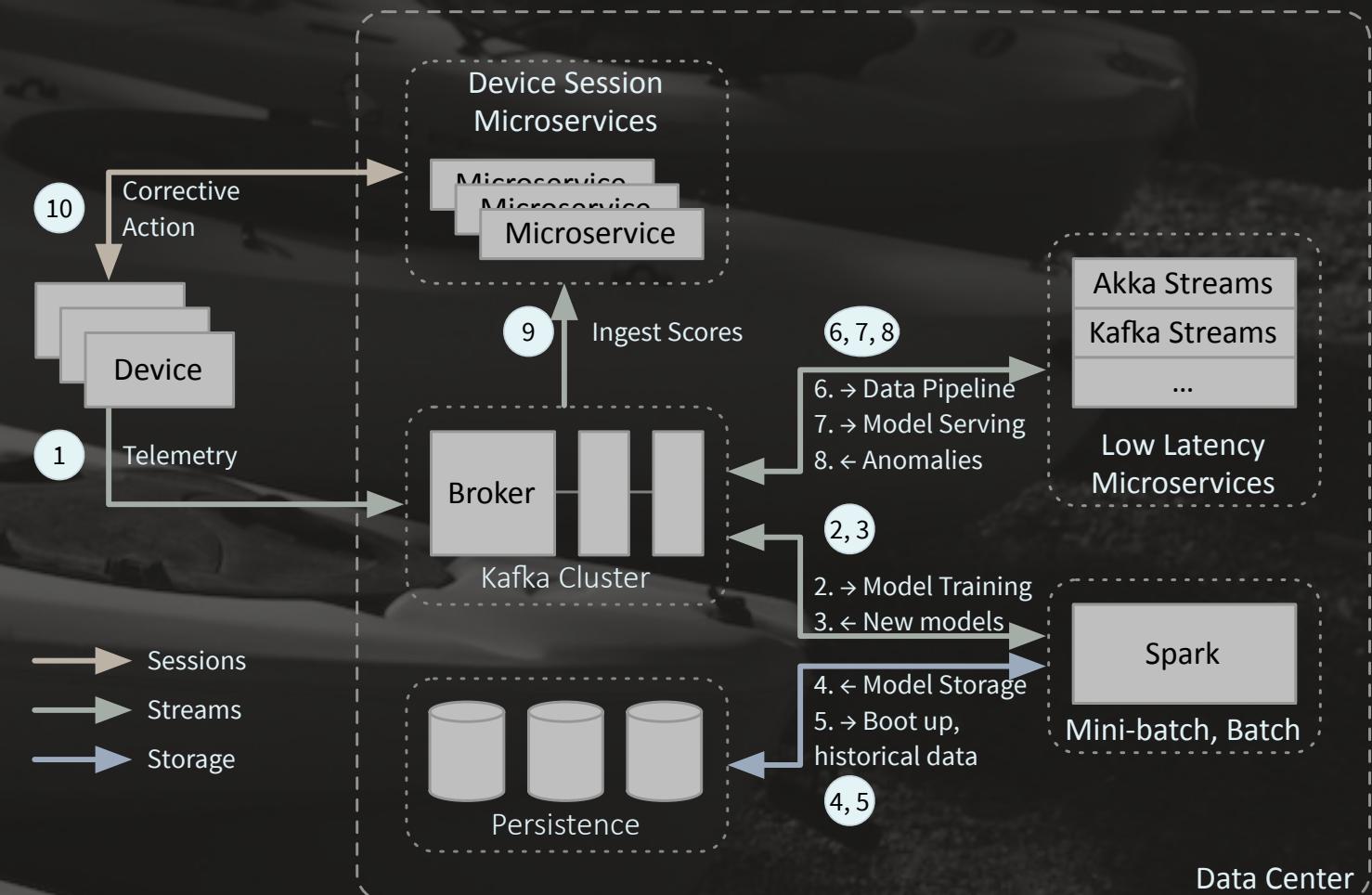
Embedded Model Serving

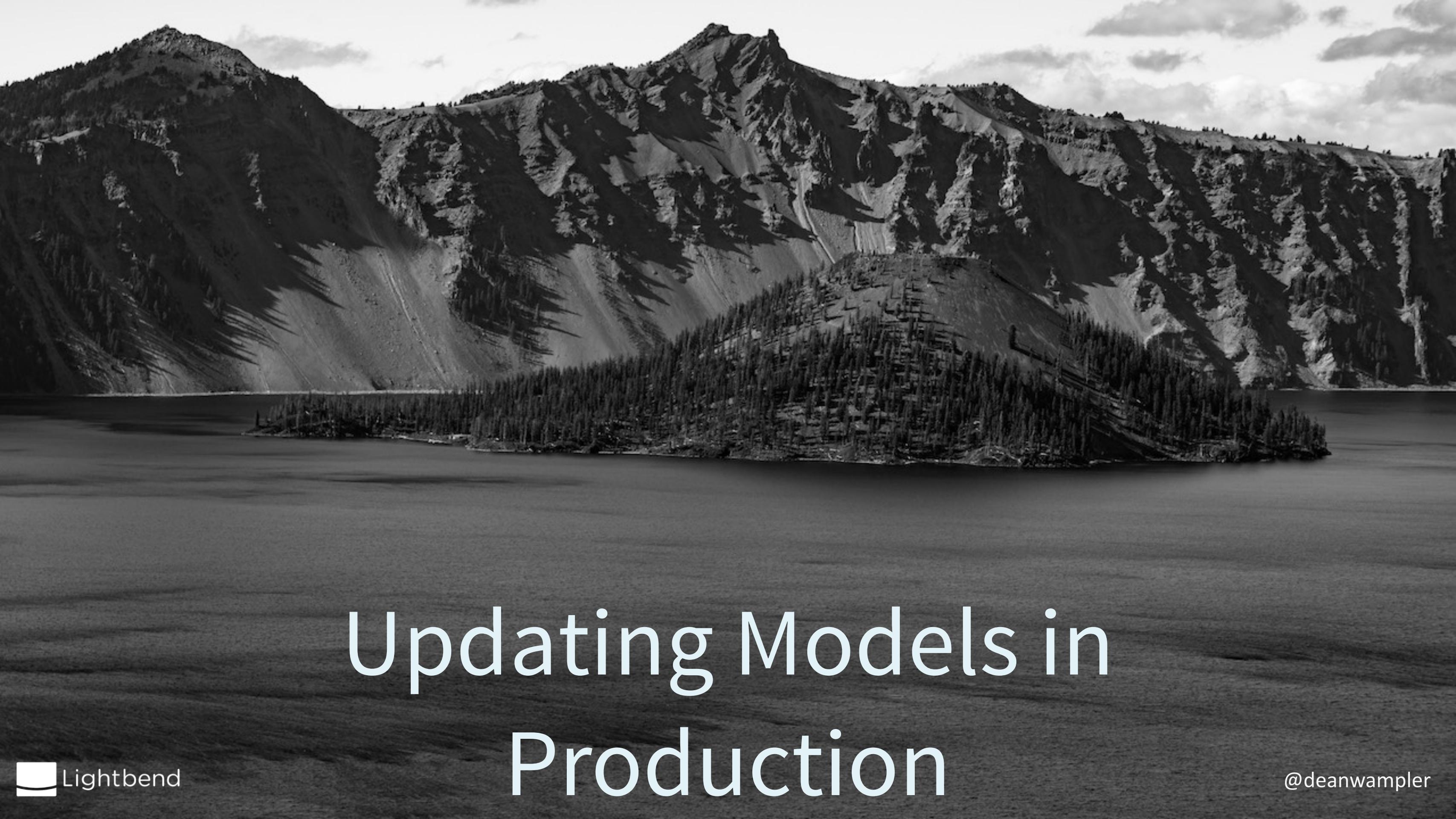
- Pros:
 - Lowest scoring overhead - interprocess communication only used for model updates
 - Performance tuning focuses on one system, the data pipeline



Embedded Model Serving

- Cons:
 - Model parameters must be serialized
 - Model serving library must be “compatible” with training system:
 - Algorithms, quality



A black and white landscape photograph showing a range of mountains in the background, their peaks partially obscured by clouds. In the middle ground, a large body of water, possibly a lake or reservoir, stretches across the frame. The foreground is dark and textured, suggesting a rocky shoreline or a close-up of the water's surface.

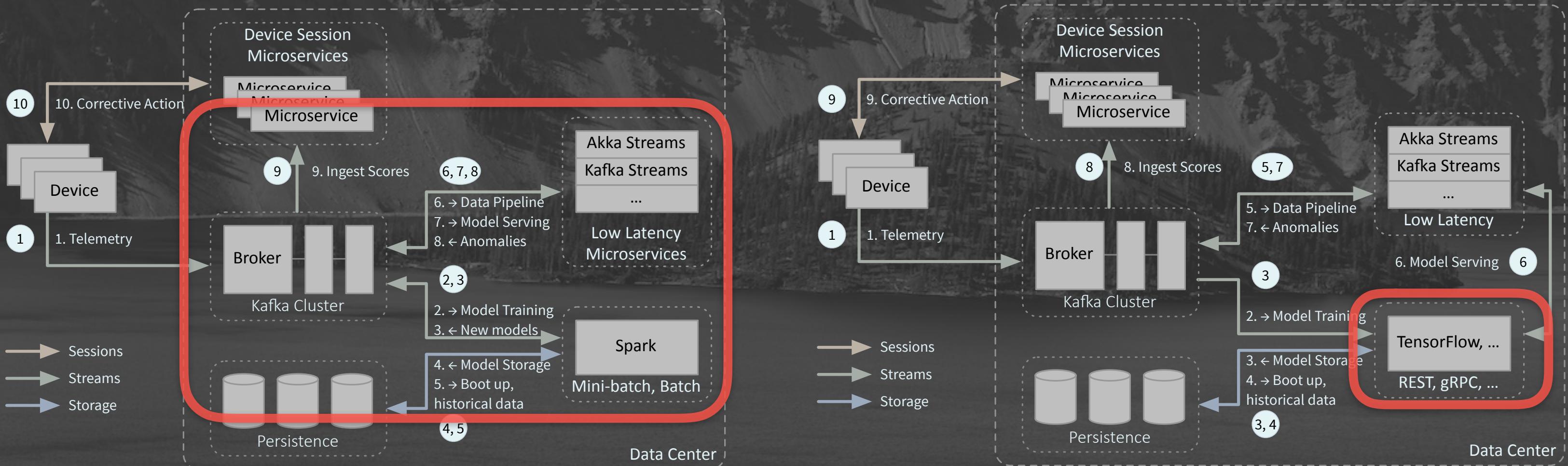
Updating Models in Production

Model Updates

- Concept Drift - models grow stale
 - They have a half life, too
 - So, periodically retrain, then serve the new model, ideally without downtime

Retraining Considerations

- How do you measure model quality?
- What's the trade-off between model performance vs. retraining cost?
- How far back in the data set do you go when training?



Complex to update
embedded models!

Model updates can be
straightforward



Dusty Milky Way

Mars last Summer



Lightroom

@deanwampler

References

- Ideas:
 - [Ben Lorica on 9 AI Trends](#)
 - [Paco Nathan's Data Governance Talk](#)

You can get these slides with the links here:
polyglotprogramming.com/talks

References

- Ideas:
 - O'Reilly Radar: [Data, AI, others](#)
 - [distill.pub](#)
 - [The Algorithm](#)
 - [The Gradient](#)

References

- A few research papers, etc.
- Incremental training
- an example
- Continual learning
- Explainability

References

- Kubeflow
- MLFlow
- DVC
- AWS SageMaker
- Fiddler (explainable AI)

References

- General Information about Stream Processing
- [My O'Reilly Report on Architectures](#)
- [Streaming Systems Book](#)
- [Stream Processing with Apache Spark](#)
- [Designing Data-Intensive APPS book](#)

References

- Other Talks
 - [Strata Talk on ML in a Streaming Context](#)
 - [Stream All the Things! \(video\)](#)
 - [Streaming Microservices with Akka Streams and Kafka Streams \(video\)](#)

References

- Tutorials
 - [Model serving in streams](#)
 - [Stream processing with Kafka and microservices](#)

Please Provide Feedback!

Cyberconflict: A new era of war, sabotage, and fear

[See passes & pricing](#)

David Sanger (The New York Times)
9:55am-10:10am Wednesday, March 27, 2019
Location: Ballroom
Secondary topics: Security and Privacy

Rate This Session

We're living in a new era of constant sabotage, misinformation, and fear, in which everyone is a target, and you're often the collateral damage in a growing conflict among states. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. Moving from the White House Situation Room to the dens of Chinese, Russian, North Korean, and Iranian hackers to the boardrooms of Silicon Valley, David reveals a world coming face-to-face with the perils of technological revolution—a conflict that the United States helped start when it began using cyberweapons against Iranian nuclear plants and North Korean missile launches. But now we find ourselves in a conflict we're uncertain how to control, as our adversaries exploit vulnerabilities in our hyperconnected nation and we struggle to figure out how to deter these complex, short-of-war attacks.

David Sanger
The New York Times

David E. Sanger is the national security correspondent for the *New York Times* as well as a national security and political contributor for CNN and a frequent guest on *CBS This Morning*, *Face the Nation*, and many PBS shows.



✓ Attending [Notes](#) [Remove](#)

Cyberconflict: A new era of war, sabotage, and fear

⌚ 9:55 AM - 10:10 AM, Wed, Mar 27, 2019

Speakers

 **David Sanger**
National Security Correspondent
The New York Times

📍 Ballroom

Keynotes

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

SESSION EVALUATION

Session page on conference website



Application Details

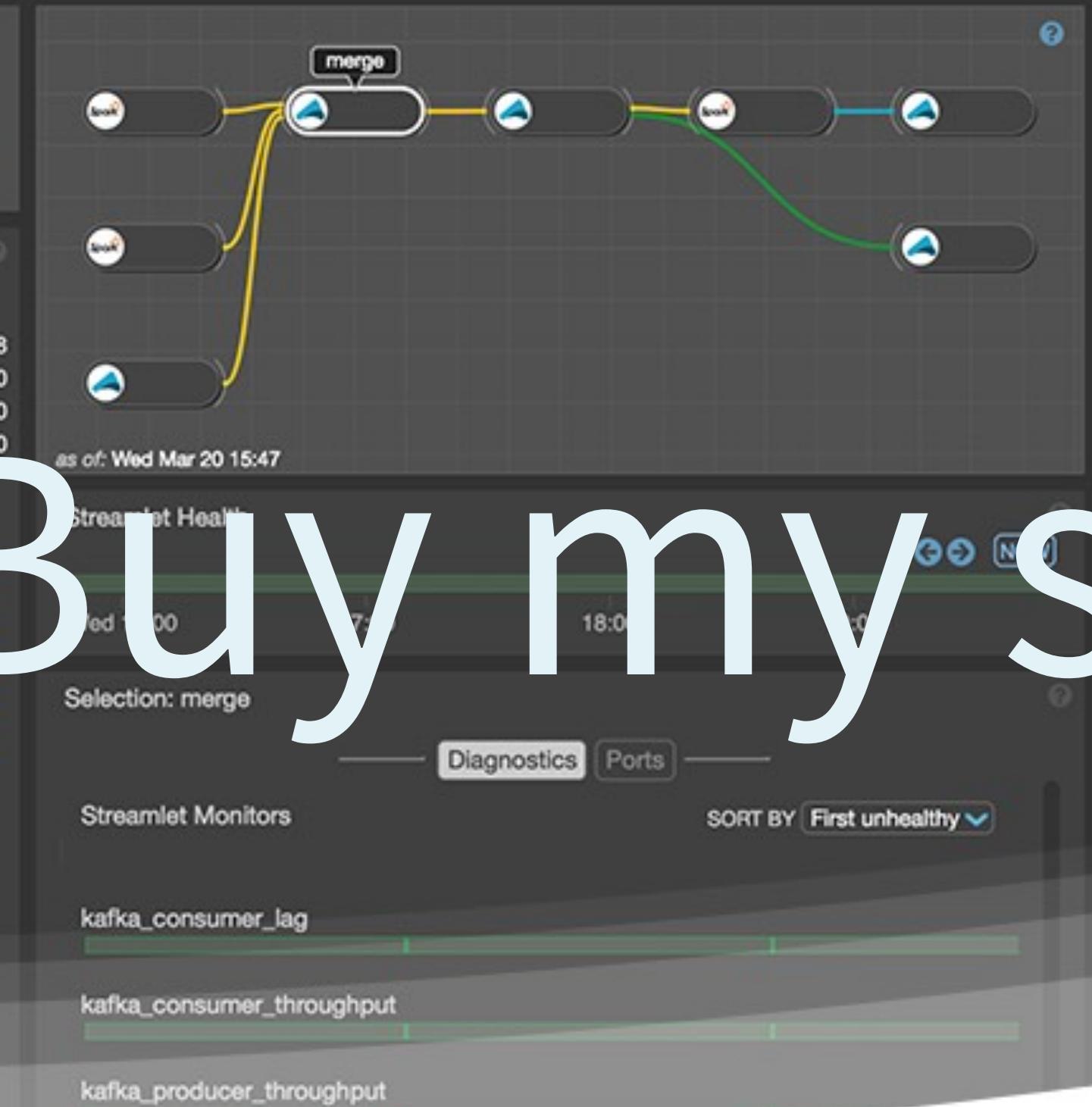
Streamlet Current Health

Healthy	8
Warning	0
Critical	0
Unknown	0

Streamlet Health Events

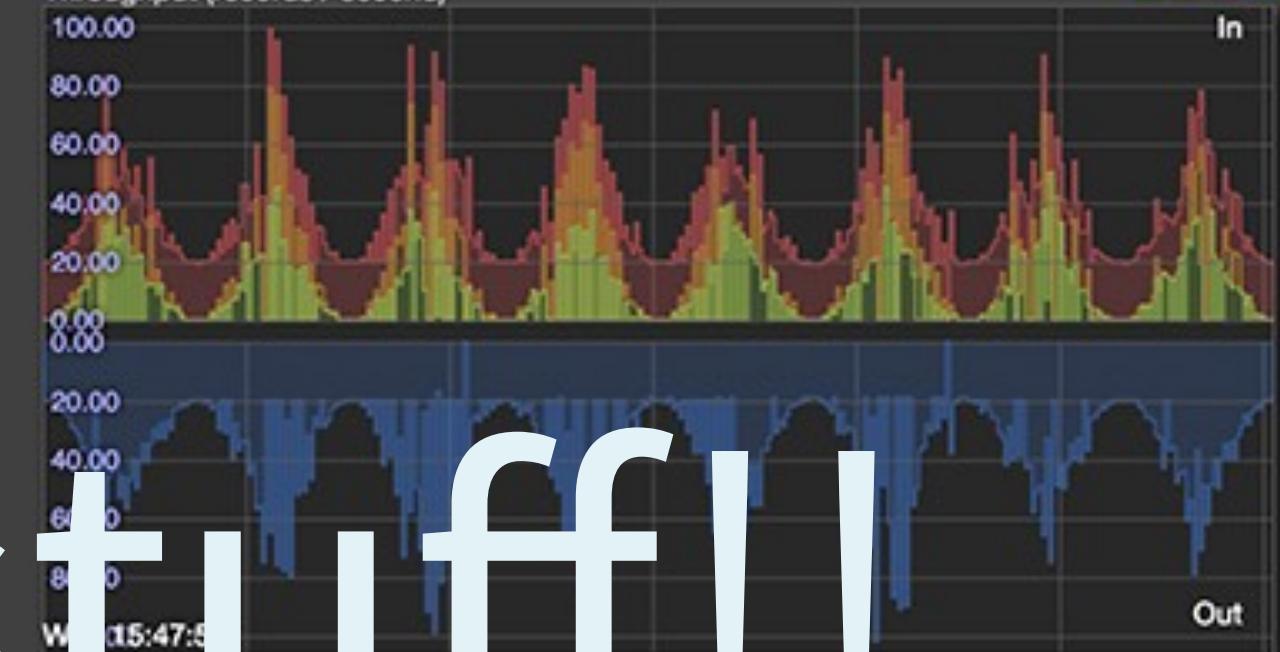
cdr-validator
cdr-aggregator
merge
console-egress
error-egress
cdr-generator1
cdr-generator2
cdr-ingress

Buy my stuff!!

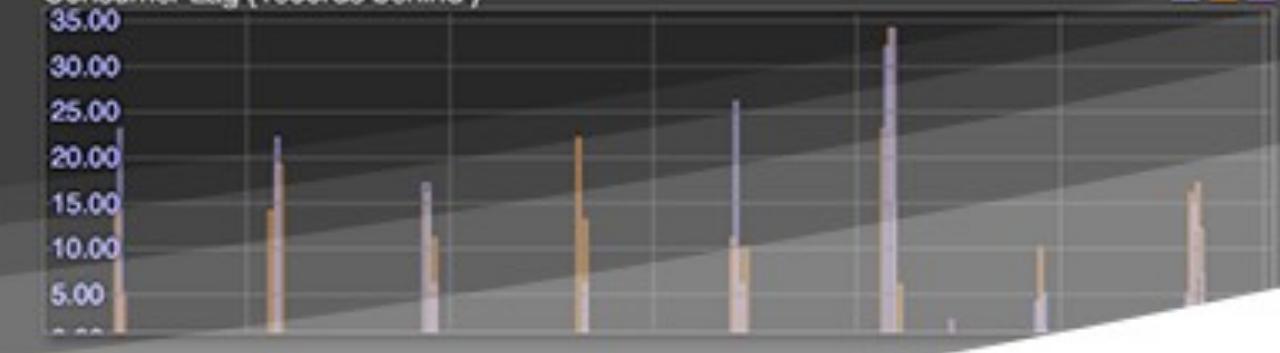


Metrics

Throughput (records / second)



Consumer Lag (records behind)



Maximum Consumer Lag (records)



lightbend.com/lightbend-pipelines-demo



Questions?

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