Building Kafka-based Microservices with Akka Streams and Kafka Streams

Strata Data NYC 2018

Boris Lublinsky and Dean Wampler, Lightbend

boris.lublinsky@lightbend.com
dean.wampler@lightbend.com

Lightbend

Outline

- Overview of streaming architectures
    - Kafka, Spark, Flink, Akka Streams, Kafka Streams
- Running example: Serving machine learning models
- Streaming in a microservice context
    - Akka Streams
    - Kafka Streams
- Wrap up

But first, introductions…

If you have not done this already, download the tutorial from GitHub

https://github.com/lightbend/kafka-with-akka-streams-kafka-streams-tutorial

These slides are in the presentation folder

Why Streaming?

"We live as streams, but we have a tendency to think in batch. Batch might be faster (simpler), but the reality is streams"

— Fabio Yamada, Kafka Mailing List

About Streaming Architectures

Why Kafka, Spark, Flink, Akka Streams, and Kafka Streams?

O'REILLY

Fast Data Architectures for Streaming Applications

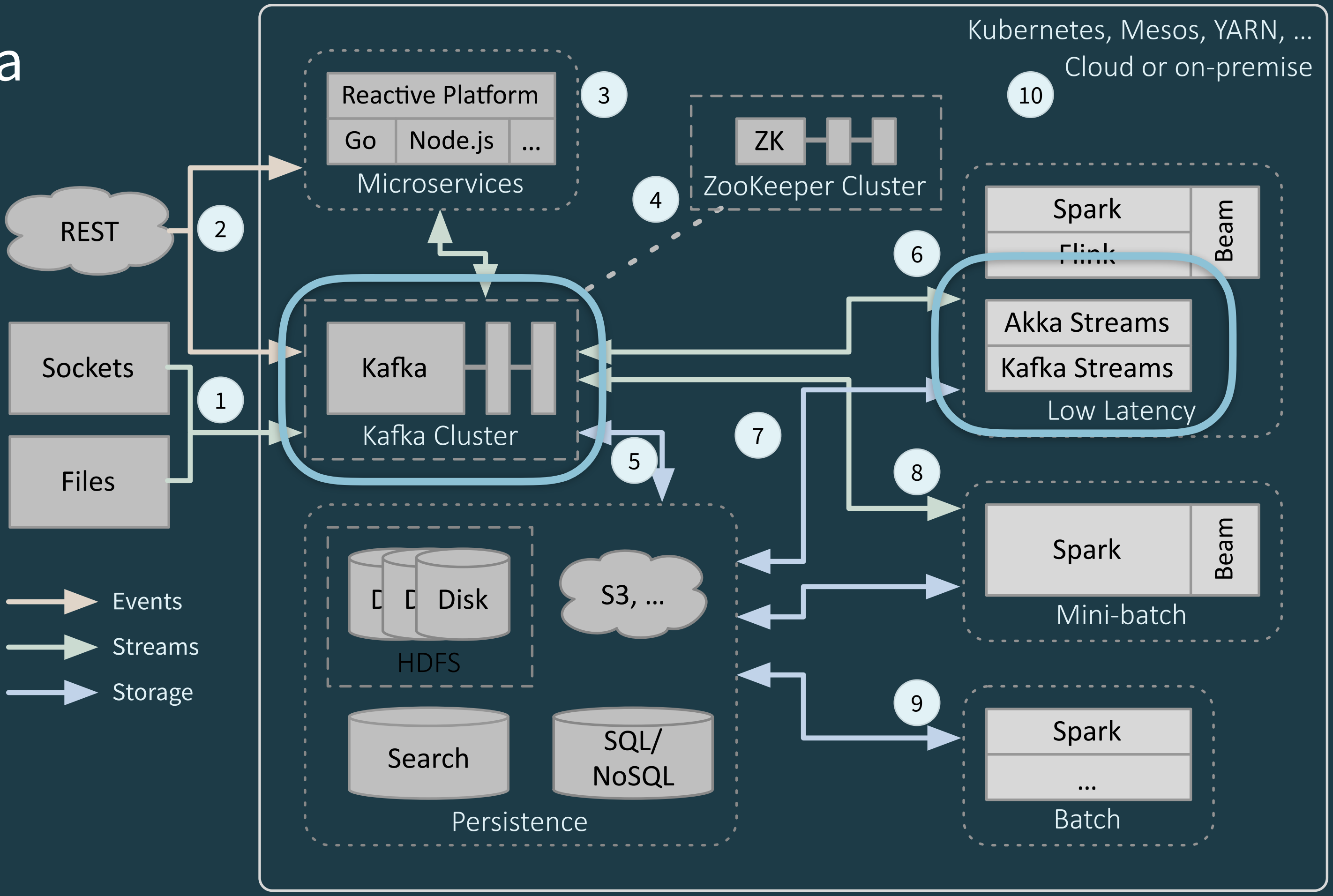
Second edition coming in October!

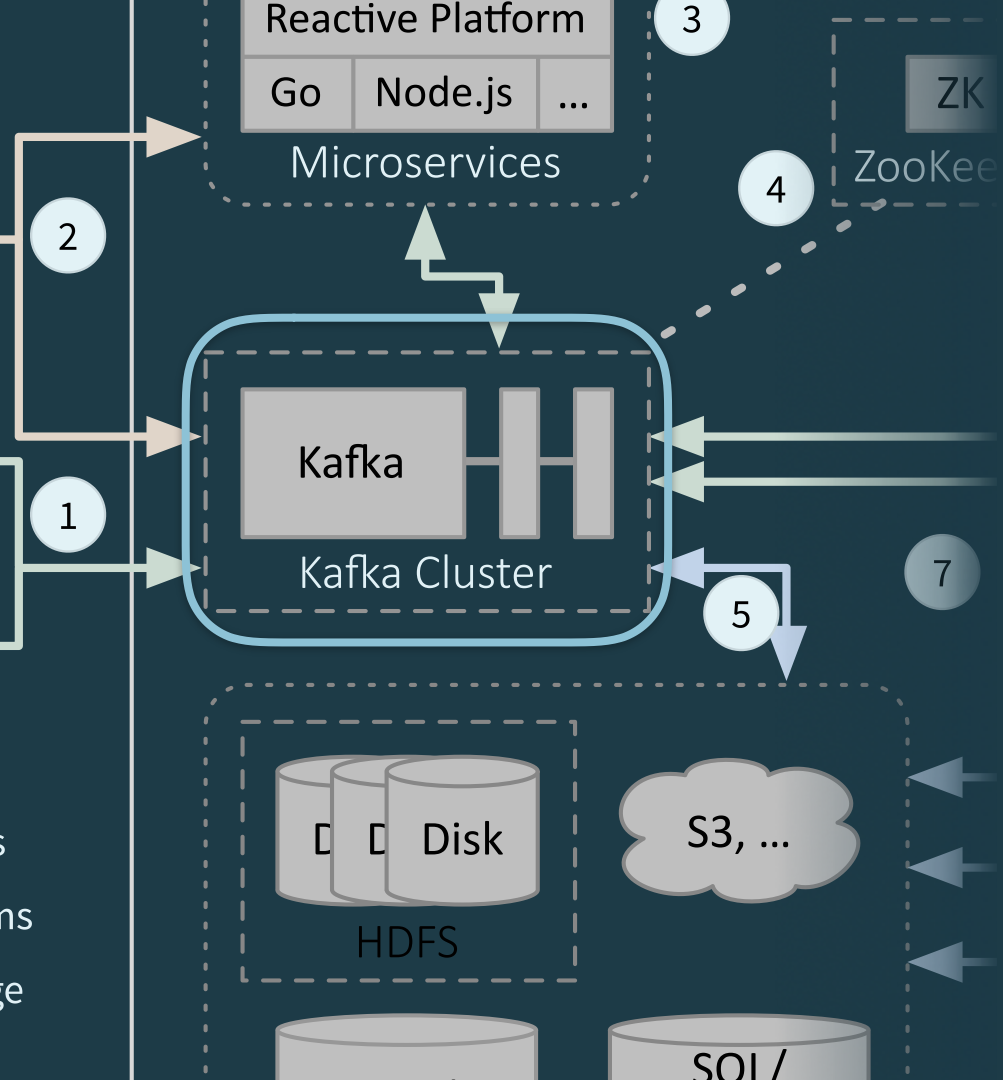
By Dean Wampler, Ph. D., VP of Fast Data Engineering

Get Your Free Copy

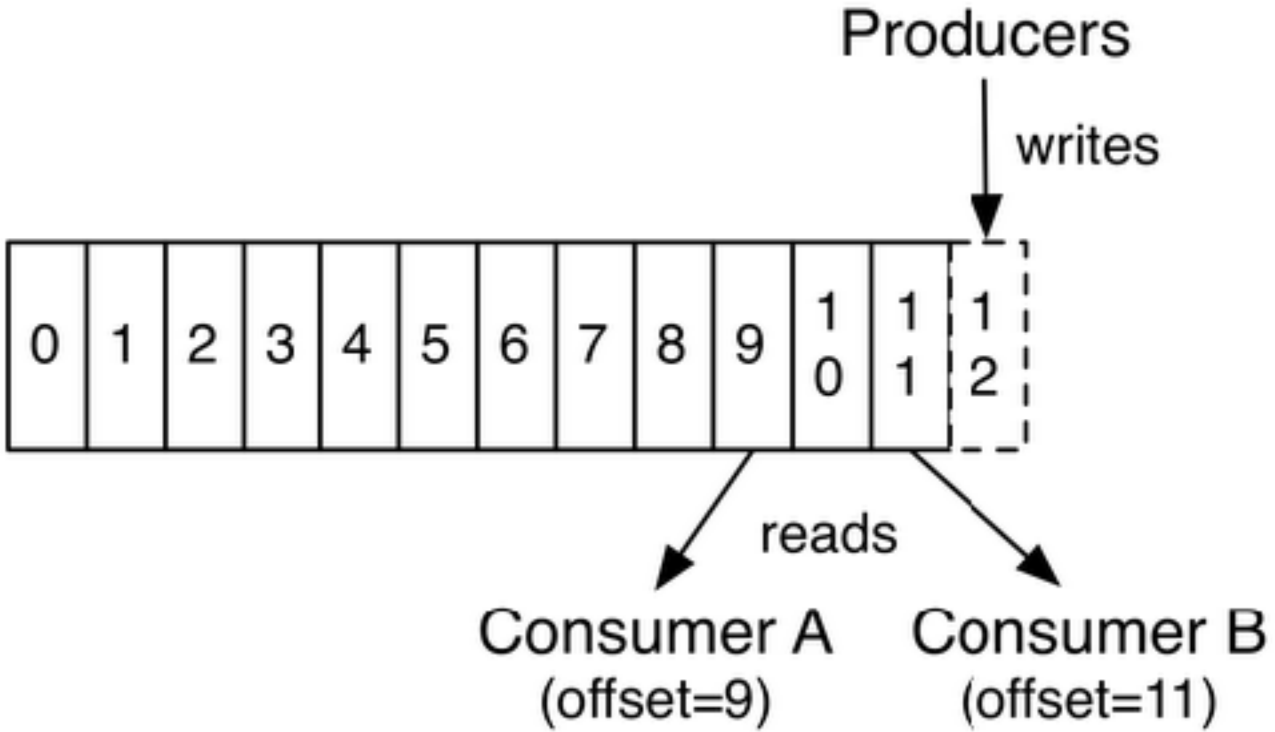
Today's focus:

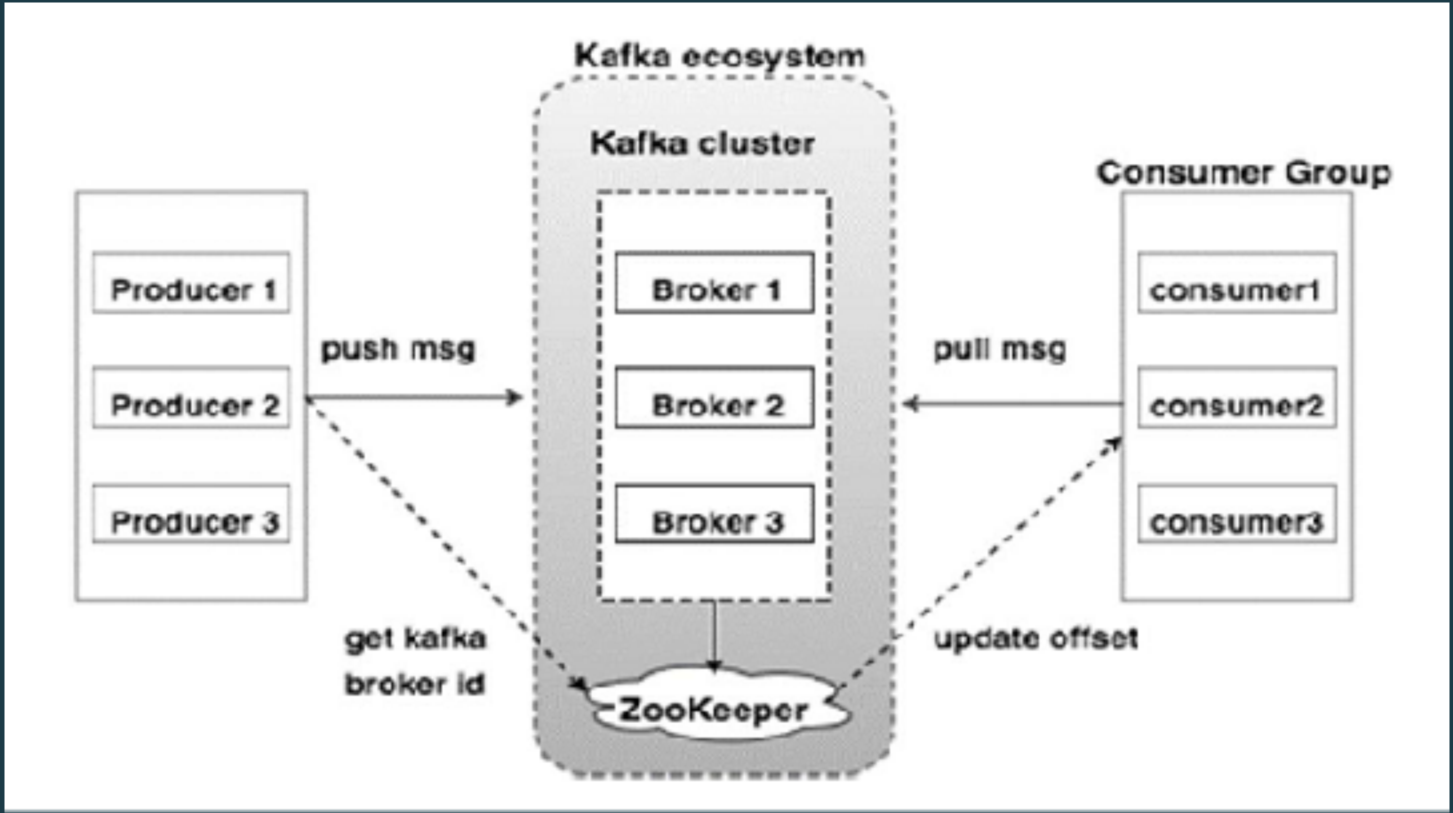
- Kafka - the data backplane
- Akka Streams and Kafka Streams - streaming microservices



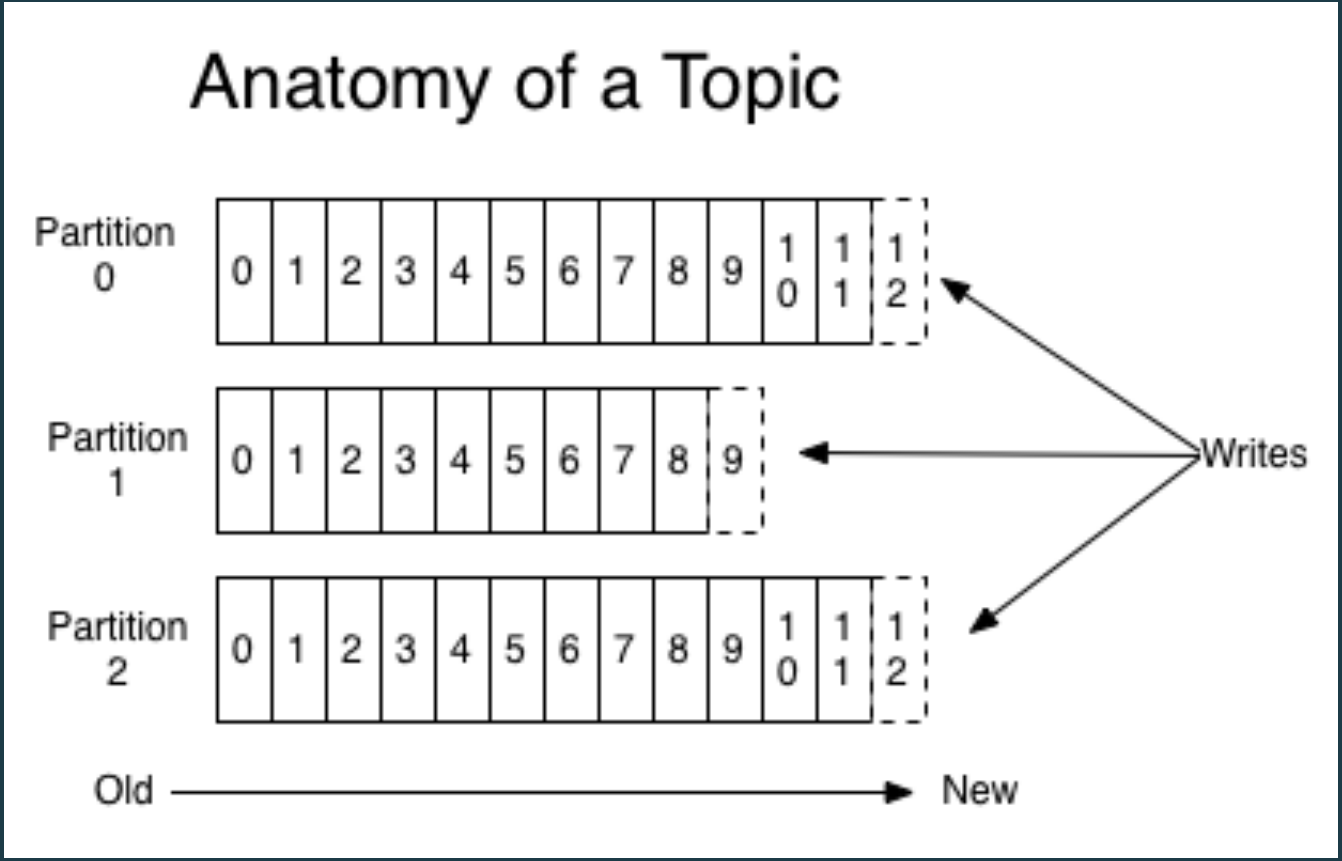

Reactive Platform
Go
Node.js
...
Microservices
3
ZK
ZooKee
4
2
Kafka
Kafka Cluster
1
7
5
Disk
HDFS
S3, ...
SQL/

Why Kafka?

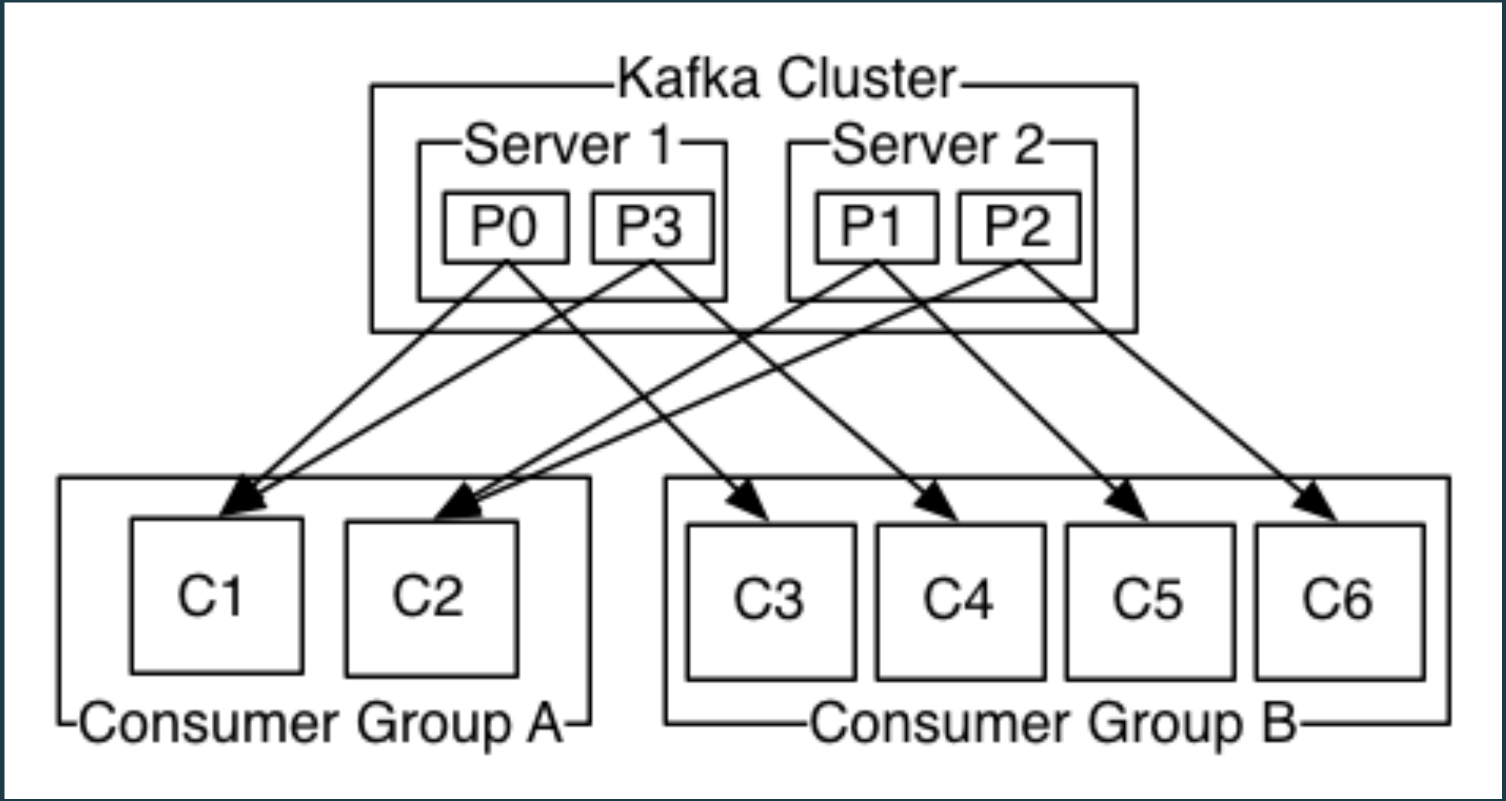

Producers
writes
0
1
2
3
4
5
6
7
8
9
10
11
12
reads
Consumer A
(offset=9)
Consumer B
(offset=11)


Kafka ecosystem
Kafka cluster
Consumer Group
Producer 1
Producer 2
Producer 3
push msg
Broker 1
Broker 2
Broker 3
pull msg
consumer1
consumer2
consumer3
get kafka
broker id
ZooKeeper
update offset

A Topic and Its Partitions



Consumer Groups



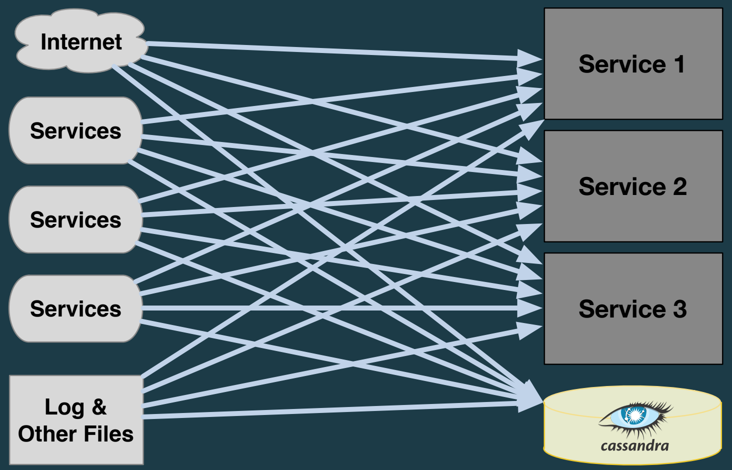
Kafka Producers and Consumers

Code time

1. Project overview
2. Explore and run the *client* project
    - Creates in-memory (“embedded”) Kafka instance and our topics
    - Pumps data into them

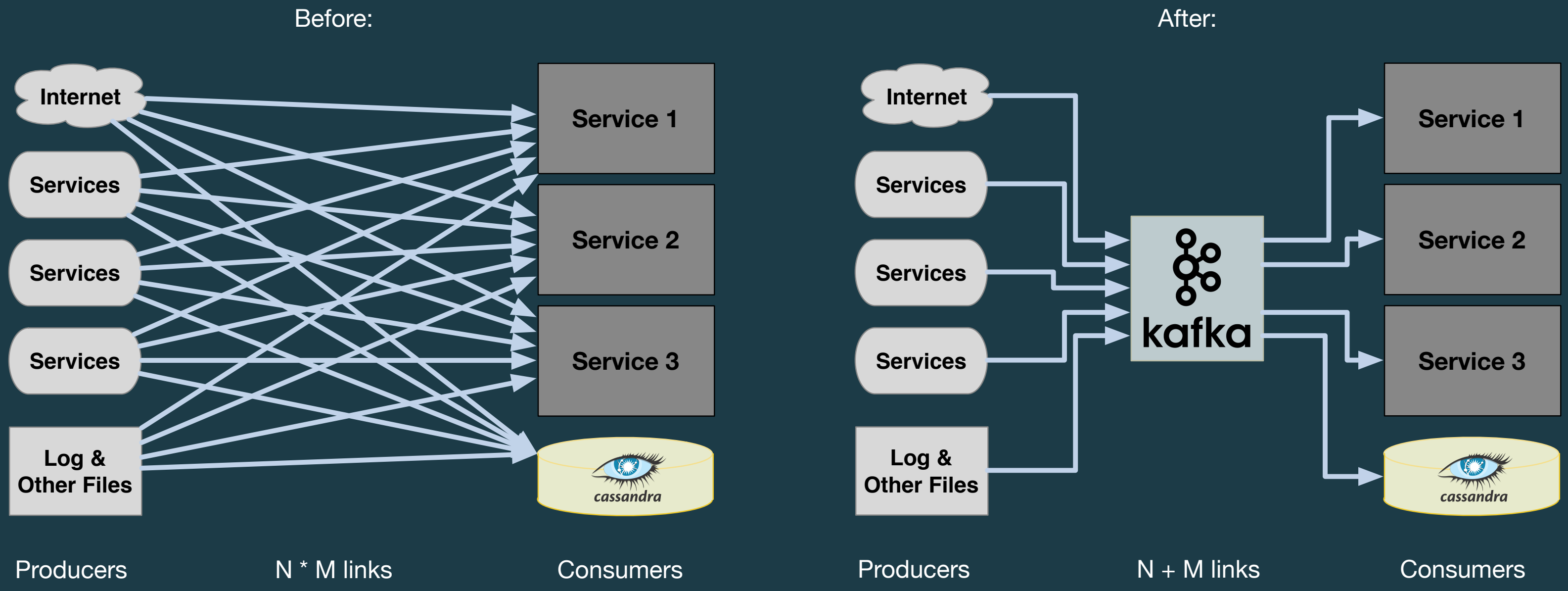
Architecture Benefits of Kafka

Before:



Producers N * M links Consumers

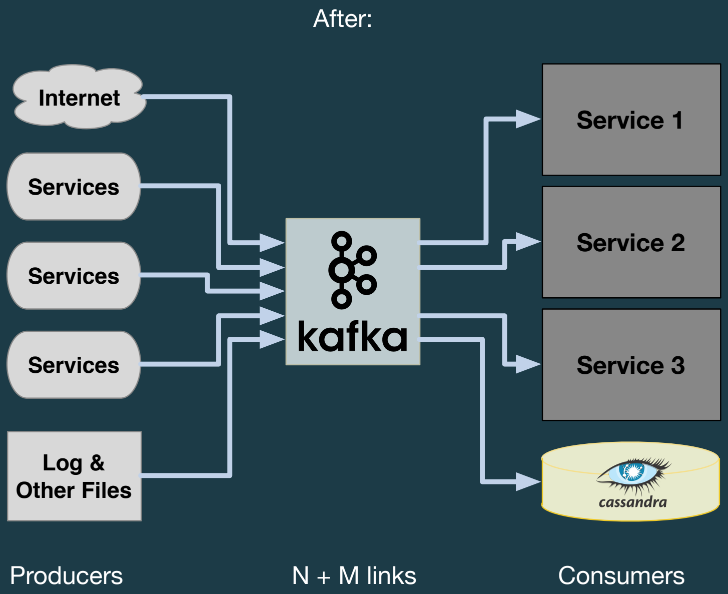
Architecture Benefits of Kafka



Architecture Benefits of Kafka

Kafka:

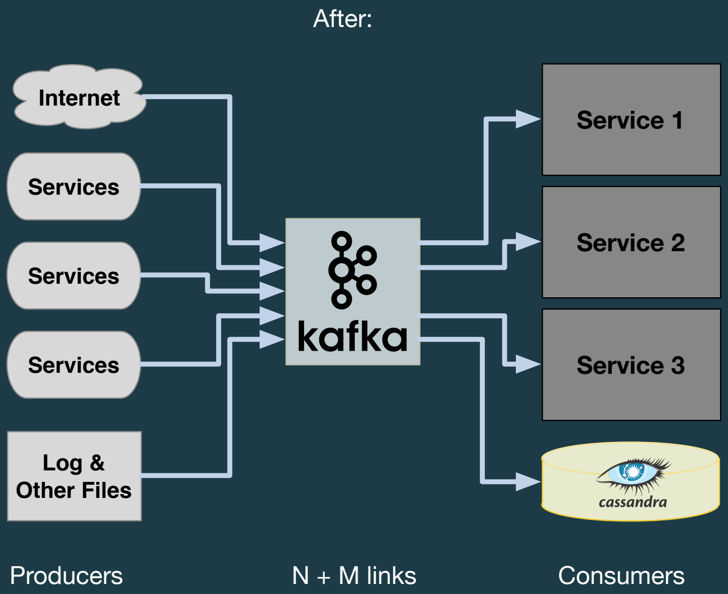
- Simplify dependencies between services
    - Improved data consistency
- Minimize data transmissions
- Reduce data loss when a service crashes



Architecture Benefits of Kafka

Kafka:

- M producers, N consumers
  - Improved extensibility
- Simplicity of one “API” for communication



Kafka Message size considerations

- Should I use Kafka for all messages?
  - Best Kafka performance is with messages size in the order of a few KB. Larger messages put heavy load on brokers and is very inefficient. It is inefficient on producers and consumers as well.
- What if my messages are very large?
  - Consider using messaging by reference - store a message in S3, HDFS, etc and send the reference to the location via Kafka

Message compatibility for Kafka

- Is it okay if messages have different **schemas**?
  - If so, handled at run time ("dynamic typing") or design time ("static typing")?
- How is message type determined?
  - Registry or repository?
  - Embedding in Kafka headers?

Message versioning

- What happens if a Producer needs to create a new message version that's incompatible with previous versions?
  - Topic versioning similar to endpoint versioning used by services.
  - Should you start new services instead?

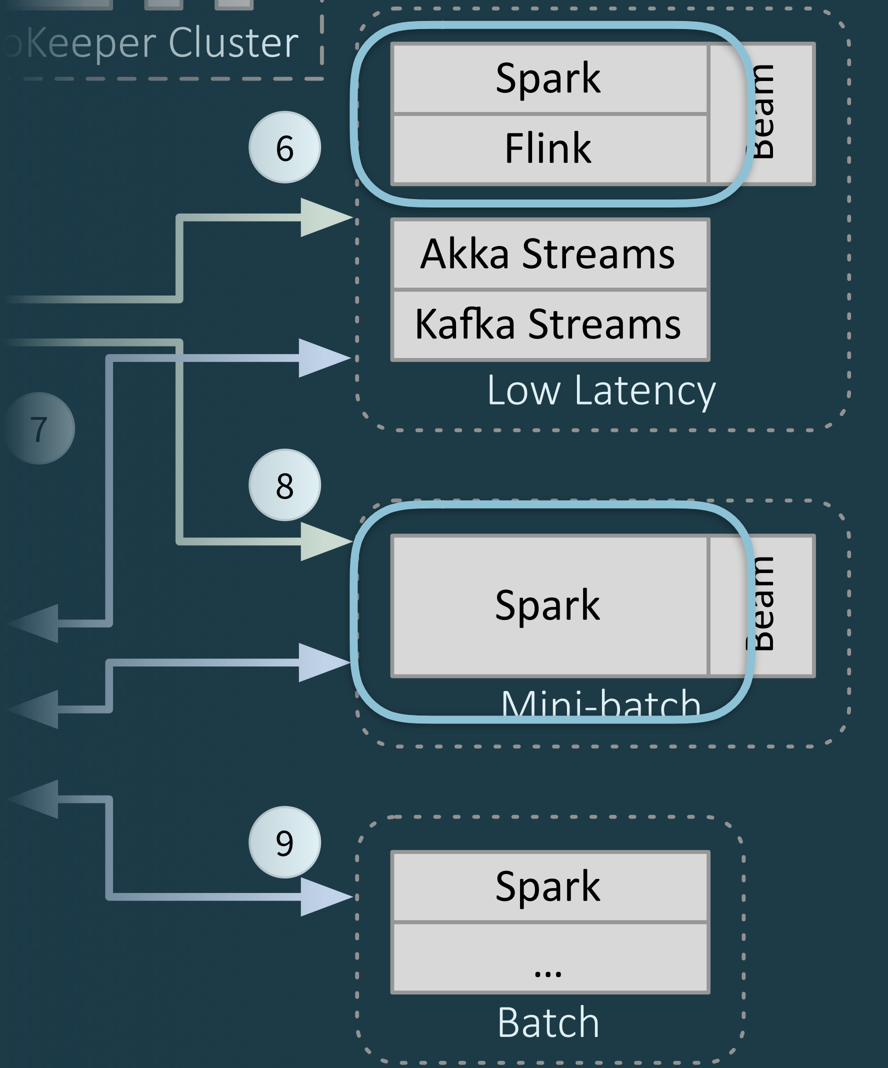
Streaming Architectures

Two options:

- Stream processing engines
- Streaming libraries

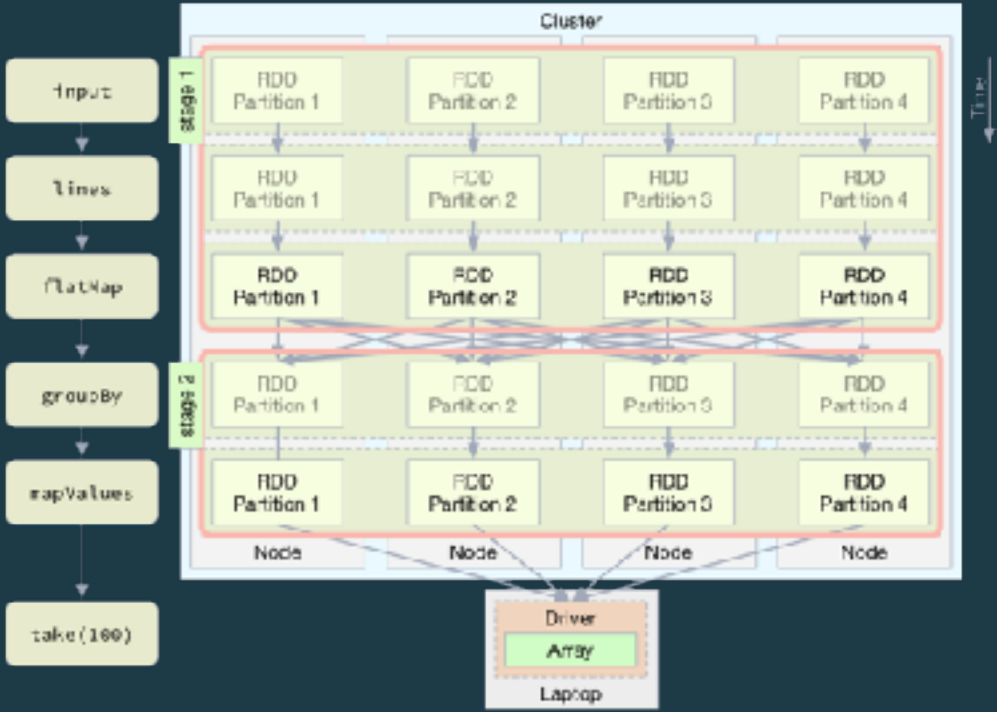
Streaming Engines:

Spark, Flink - services to which you submit work. Large scale, automatic data partitioning.



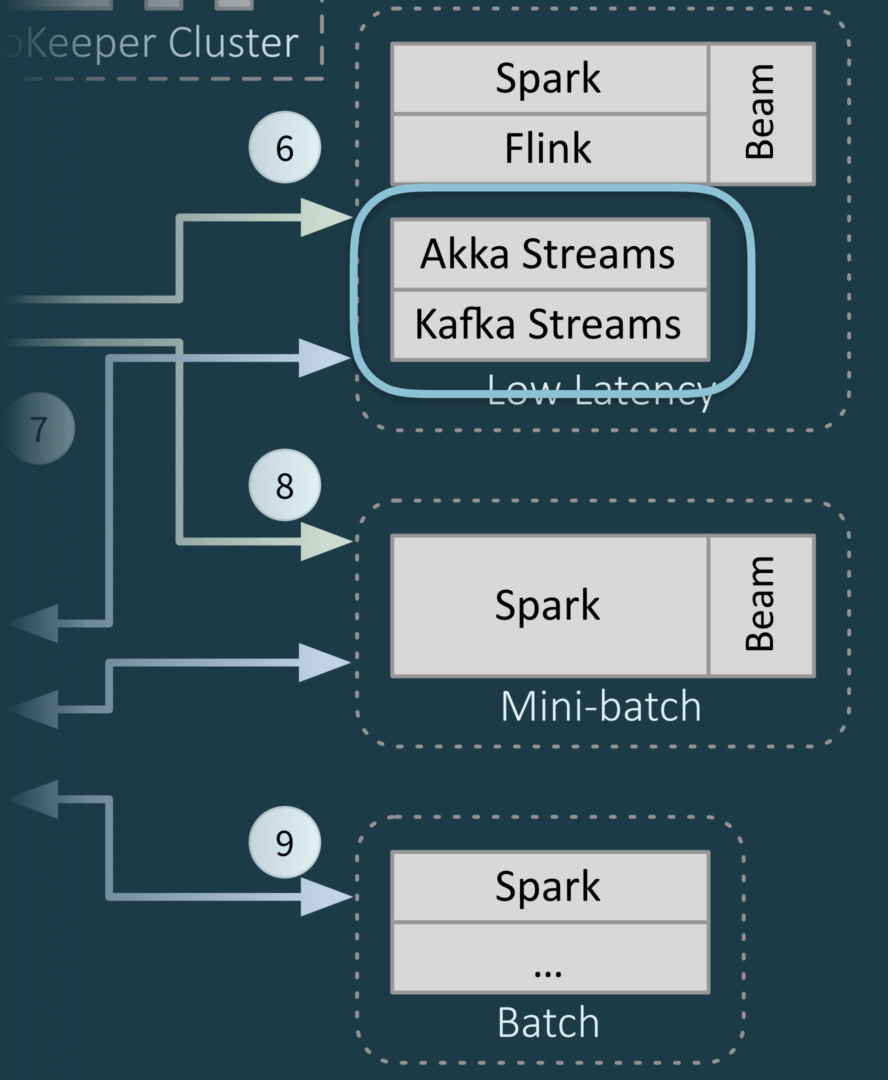
Streaming Engines:

Spark, Flink - services to which you submit work. Large scale, automatic data partitioning.



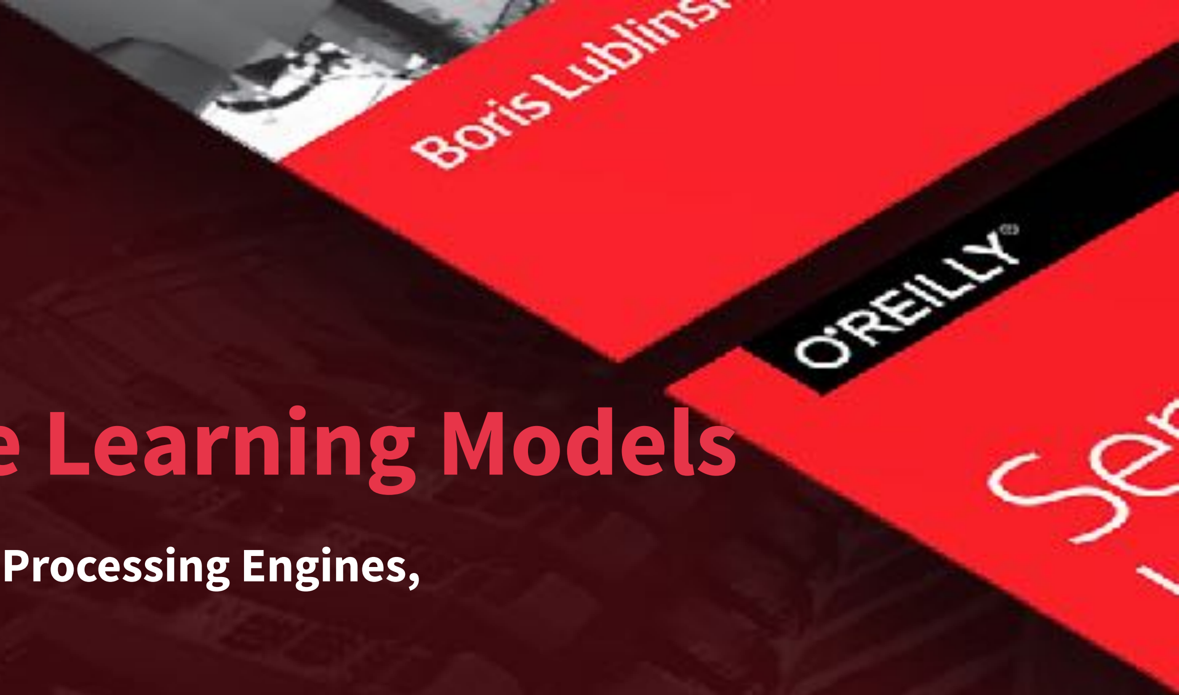
Streaming Libraries:

Akka Streams, Kafka Streams - libraries for “data-centric micro services”. Smaller scale, but great flexibility.



Machine Learning and Model Serving: A Quick Introduction

O'REILLY



Serving Machine Learning Models

A Guide to Architecture, Stream Processing Engines, and Frameworks

By Boris Lublinsky, Fast Data Platform Architect

Get Your Free Copy

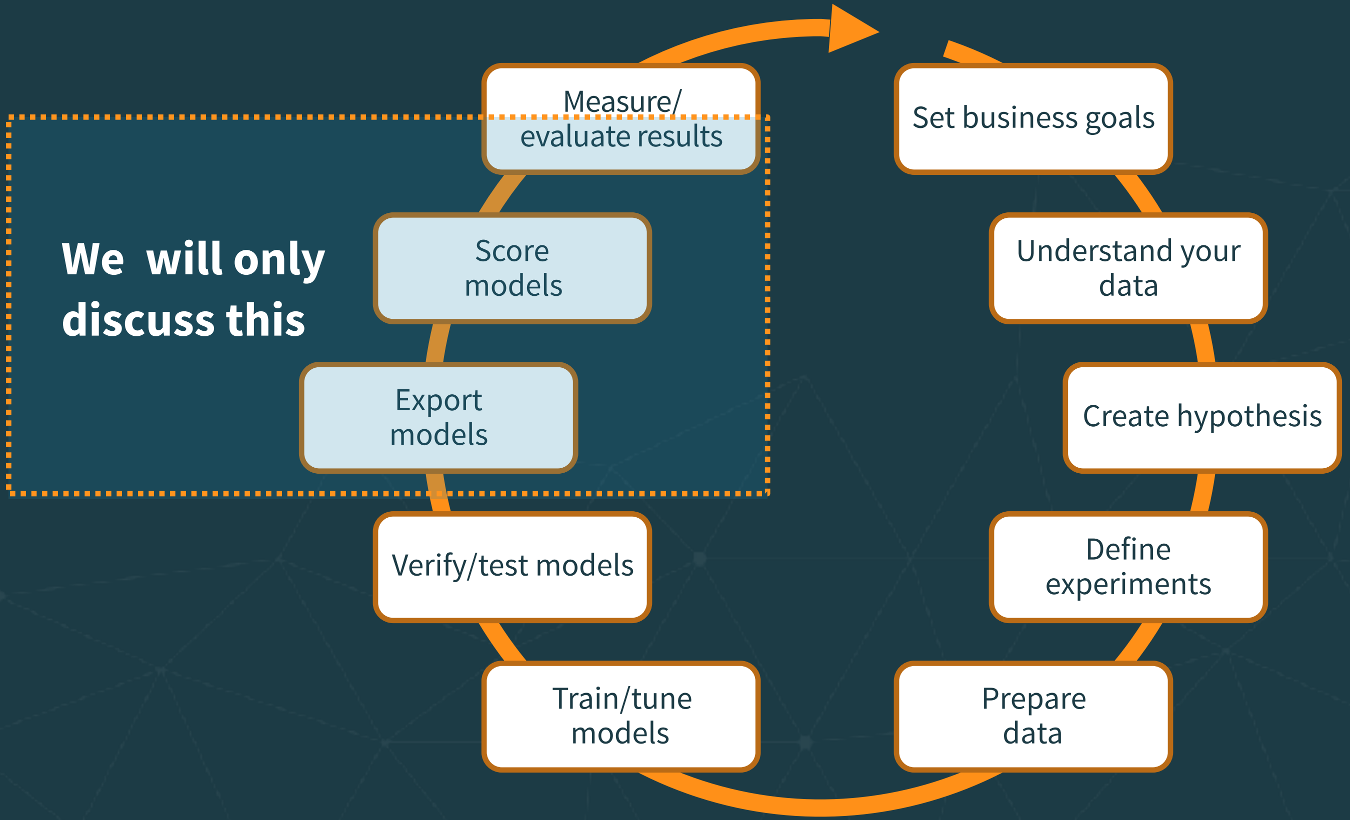
ML Is Simple

Data Magic Happiness

Maybe Not

Even If There Are Instructions

The Reality



Lightbend

What Is The Model?

A model is a function transforming inputs to outputs - $y = f(x)$

for example:

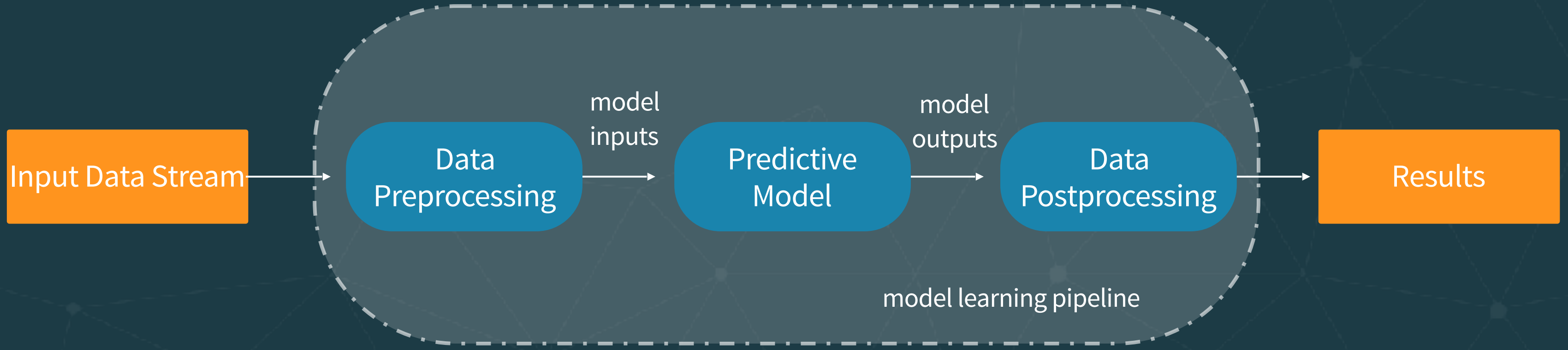
Linear regression: $y = a_c + a_1 * x1 + \ldots + a_n * x_n$

Neural network: $f(x) = K(\sum_i w_i g_i(x))$

Such a definition of the model allows for an easy implementation of model’s composition. From the implementation point of view it is just function composition

Model Learning Pipeline

UC Berkeley AMPLab introduced machine learning pipelines as a graph defining the complete chain of data transformation.



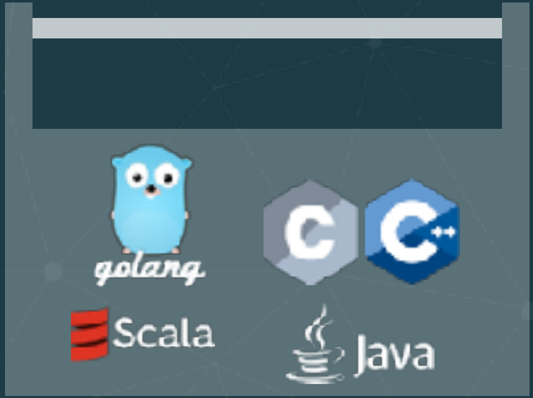
Traditional Approach to Model Serving

- Model is code
- This code has to be saved and then somehow imported into model serving

Why is this problematic?

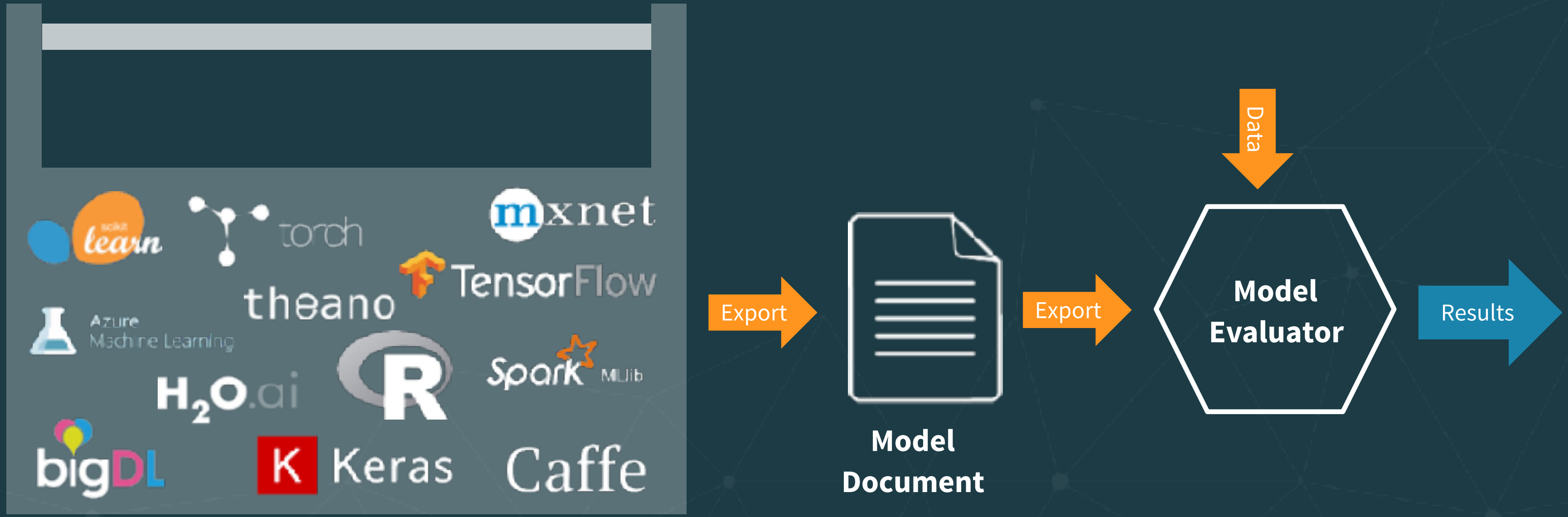
Impedance Mismatch

Continually expanding
Data Scientist toolbox



Defined Software
Engineer toolbox

Alternative - Model As Data



Standards PMML Predictive Model Markup Language

Portable
Format for
Analytics **(PFA)**

TensorFlow

Exporting Model As Data With PMML

There are already a lot of export options

Spark MLlib https://github.com/jpmml/jpmml-sparkml

scikit learn https://github.com/jpmml/jpmml-sklearn

R https://github.com/jpmml/jpmml-r

TensorFlow https://github.com/jpmml/jpmml-tensorflow

Evaluating PMML Model

There are also a few PMML evaluators

Java https://github.com/jpmml/jpmml-evaluator

python https://github.com/opendatagroup/augustus

Exporting Model As Data With Tensorflow

- Tensorflow execution is based on Tensors and Graphs
- Tensors are defined as multilinear functions which consist of various vector variables
- A computational graph is a series of Tensorflow operations arranged into graph of nodes
- Tensorflow supports exporting graphs in the form of binary protocol buffers
- There are two different export format - optimized graph and a new format - saved model

Evaluating Tensorflow Model

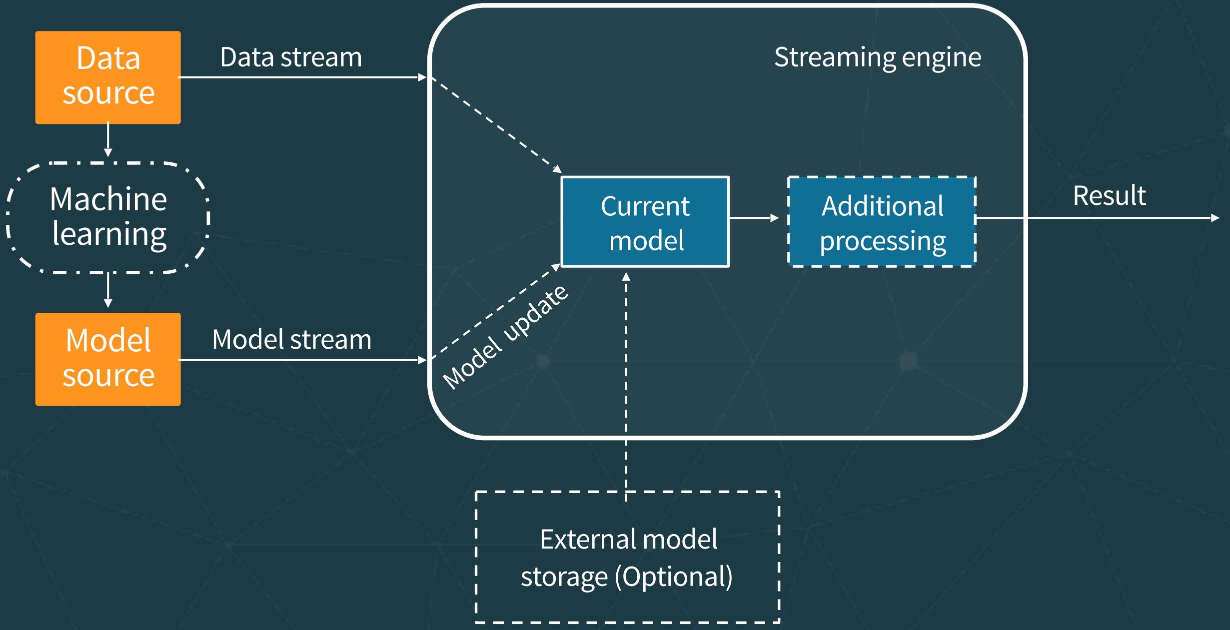
- Tensorflow is implemented in C++ with a Python interface.
- In order to simplify Tensorflow usage from Java, in 2017 Google introduced Tensorflow Java API.
- Tensorflow Java API supports importing an exported model and allows to use it for scoring.

Additional Considerations – Model Lifecycle

- Models tend to change
- Update frequencies vary greatly – from hourly to quarterly/yearly
- Model version tracking
- Model release practices
- Model update process

The Solution

A streaming system allowing to update models without interruption of execution (dynamically controlled stream).



Model Representation (Protobufs)

```
// On the wire
syntax = "proto3";
// Description of the trained model.
message ModelDescriptor {
  string name = 1;     // Model name
  string description = 2;     // Human readable
  string dataType = 3;  // Data type for which this model is applied.
  enum ModelType {    // Model type
    TENSORFLOW  = 0;
    TENSORFLOWSAVED = 2;
    PMML = 2;
  };
  ModelType modeltype = 4;
  oneof MessageContent {
    // Byte array containing the model
    bytes data = 5;
    string location = 6;
  }
}
```

Model Representation (Scala)

```
trait Model {
  def score(input : Any) : Any
  def cleanup() : Unit
  def toBytes() : Array[Byte]
  def getType : Long
}

def ModelFactoryl {
  def create(input : ModelDescriptor) : Model
  def restore(bytes : Array[Byte]) : Model
}
```

Side Note: Monitoring

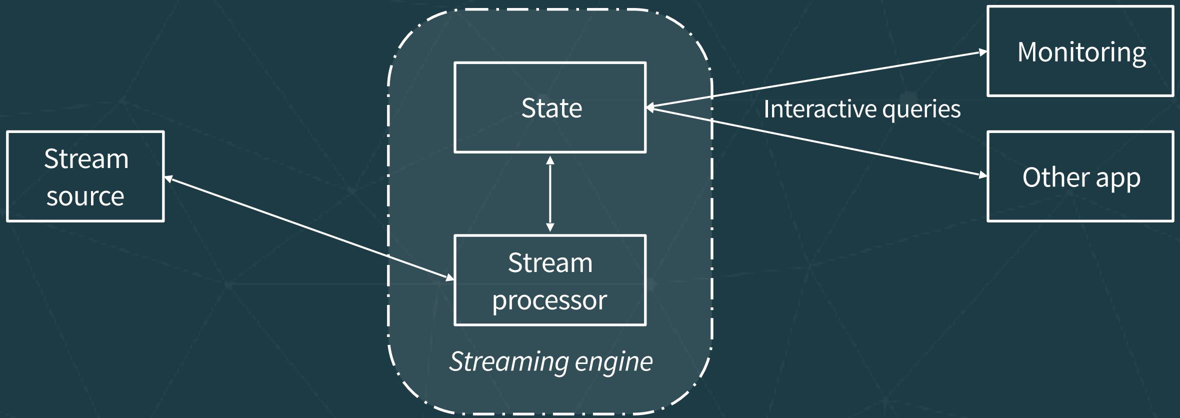
Model monitoring should provide information about usage, behavior, performance and lifecycle of the deployed models

```
case class ModelToServeStats(
name: String,                                   //  Model name
      description: String,                      // Model descriptor
      modelType: ModelDescriptor.ModelType,     // Model type
      since : Long,                             // Start time of model usage
      var usage : Long = 0,                     // Number of servings
      var duration : Double = 0.0,              // Time spent on serving
      var min : Long = Long.MaxValue,           // Min serving time
      var max : Long = Long.MinValue            // Max serving time
)
```

Queryable State

Queryable state: ad hoc query of the state in the stream. Different than the normal data flow.

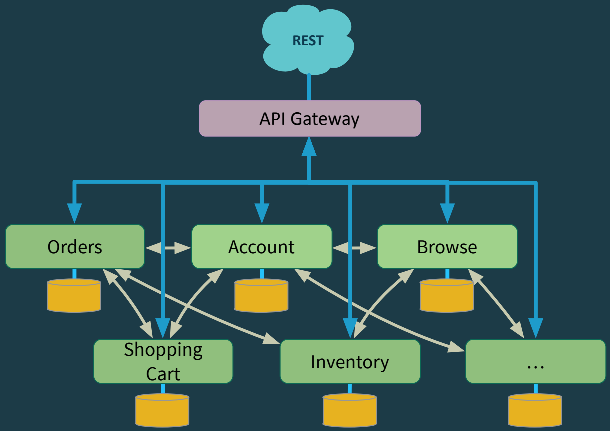
Treats the stream processing layer as a lightweight embedded *database. Directly query the current state* of a stream processing application. No need to materialize that state to a database, etc. first.



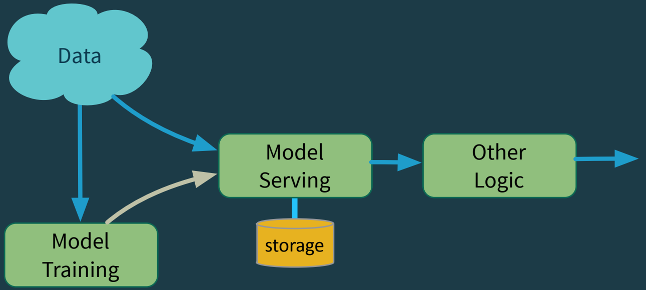
Microservice All the Things!

A Spectrum of Microservices

Event-driven µ-services



"Record-centric" µ-services

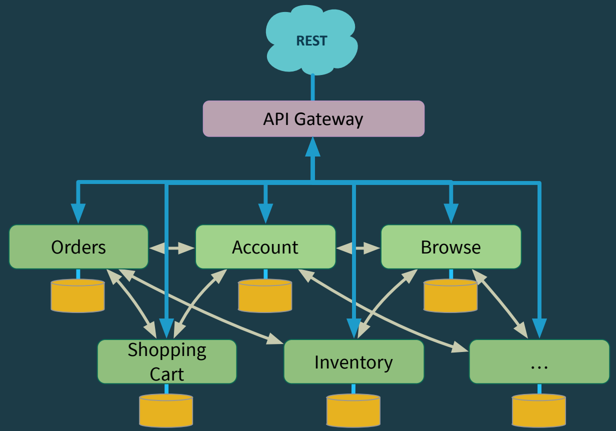


Events

Records

A Spectrum of Microservices

Event-driven µ-services



Akka emerged from the left-hand side of the spectrum, the world of highly *Reactive* microservices.

Akka Streams pushes to the right, more data-centric.

Events

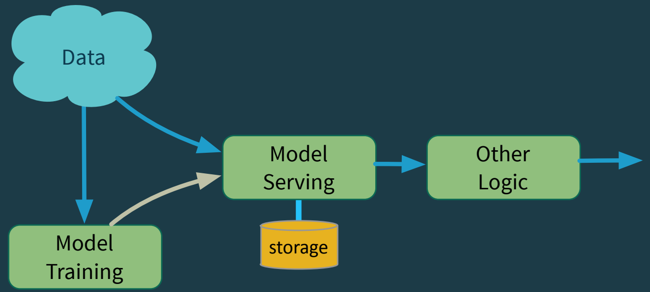
Records

A Spectrum of Microservices

Emerged from the right-hand side.

Kafka Streams pushes to the left, supporting many event-processing scenarios.

"Record-centric" μ-services



Events

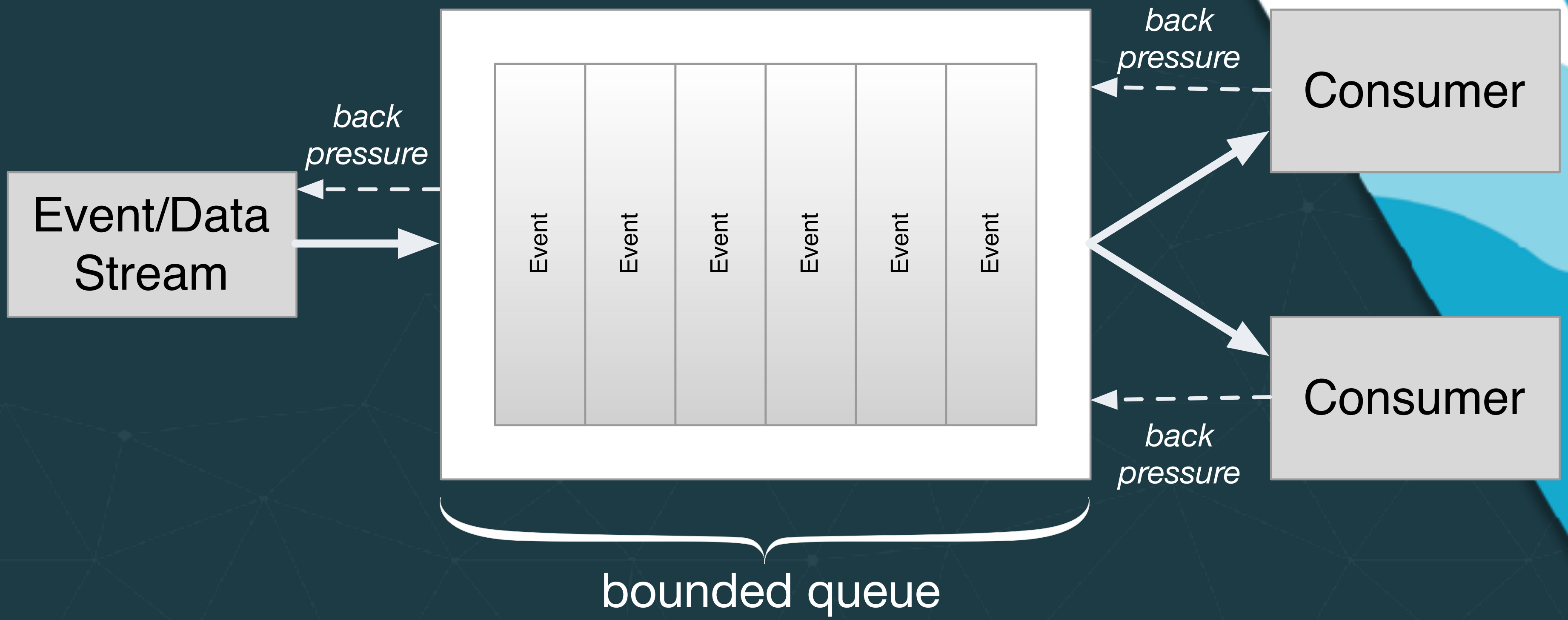
Records

Akka Streams

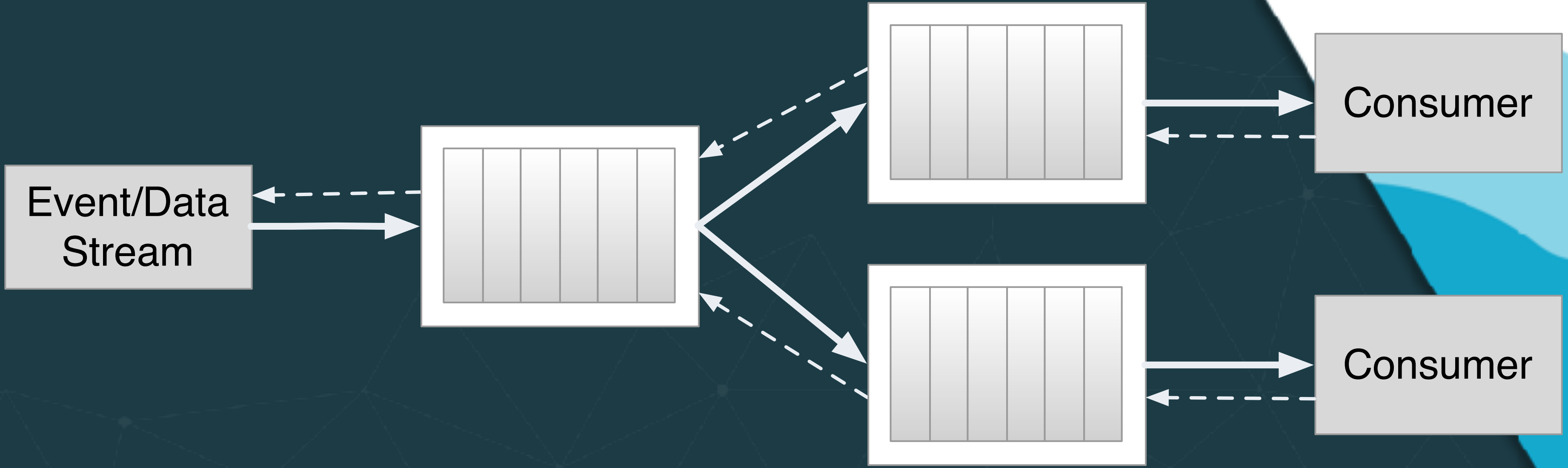
akka streams

- A *library*
- Implements Reactive Streams.
  - http://www.reactive-streams.org/
  - *Back pressure* for flow control

akka streams



akka streams



... and they compose

akka streams

- Part of the Akka ecosystem
  - Akka Actors, Akka Cluster, Akka HTTP, Akka Persistence, …
  - Alpakka - rich connection library
    - like Camel, but implements Reactive Streams
  - Commercial support from Lightbend

akka streams

- A very simple example to get the “gist”…

```scala
import akka.stream._
import akka.stream.scaladsl._
import akka.NotUsed
import akka.actor.ActorSystem
import scala.concurrent._
import scala.concurrent.duration._

implicit val system = ActorSystem("QuickStart")
implicit val materializer = ActorMaterializer()

val source: Source[Int, NotUsed] = Source(1 to 10)
val factorials = source.scan(BigInt(1)) ( (acc, next) => acc * next )
factorials.runWith(Sink.foreach(println))
```

```scala
import akka.stream._
import akka.stream.scaladsl._
import akka.NotUsed
import akka.actor.ActorSystem
import scala.concurrent._
import scala.concurrent.duration._

implicit val system = ActorSystem("QuickStart")
implicit val materializer = ActorMaterializer()

val source: Source[Int, NotUsed] = Source(1 to 10)
val factorials = source.scan(BigInt(1)) ( (acc, next) => acc * next )
factorials.runWith(Sink.foreach(println))
```

Imports!

```scala
import akka.stream._
import akka.stream.scaladsl._
import akka.NotUsed
import akka.actor.ActorSystem
import scala.concurrent._
import scala.concurrent.duration._

implicit val system = ActorSystem("QuickStart")
implicit val materializer = ActorMaterializer()

val source: Source[Int, NotUsed] = Source(1 to 10)
val factorials = source.scan(BigInt(1)) ( (acc, next) => acc * next )
factorials.runWith(Sink.foreach(println))
```

Initialize and specify now the stream is "materialized"

```scala
import akka.stream._
import akka.stream.scaladsl._
import akka.NotUsed
import akka.actor.ActorSystem
import scala.concurrent._
import scala.concurrent.duration._

implicit val system = ActorSystem("QuickStart")
implicit val materializer = ActorMaterializer()

val source: Source[Int, NotUsed] = Source(1 to 10)
val factorials = source.scan(BigInt(1)) ( (acc, next) => acc * next )
factorials.runWith(Sink.foreach(println))
```

Create a Source of Ints. Second type represents a hook used for “materialization” - not used here

```scala
import akka.stream._
import akka.stream.scaladsl._
import akka.NotUsed
import akka.actor.ActorSystem
import scala.concurrent._
import scala.concurrent.duration._

implicit val system = ActorSystem("QuickStart")
implicit val materializer = ActorMaterializer()

val source: Source[Int, NotUsed] = Source(1 to 10)
val factorials = source.scan(BigInt(1)) ( (acc, next) => acc * next )
factorials.runWith(Sink.foreach(println))
```

Scan the Source and compute factorials, with a seed of 1, of type BigInt

```scala
import akka.stream._
import akka.stream.scaladsl._
import akka.NotUsed
import akka.actor.ActorSystem
import scala.concurrent._
import scala.concurrent.duration._

implicit val system = ActorSystem("QuickStart")
implicit val materializer = ActorMaterializer()

val source: Source[Int, NotUsed] = Source(1 to 10)
val factorials = source.scan(BigInt(1)) ( (acc, next) => acc * next )
factorials.runWith(Sink.foreach(println))
```

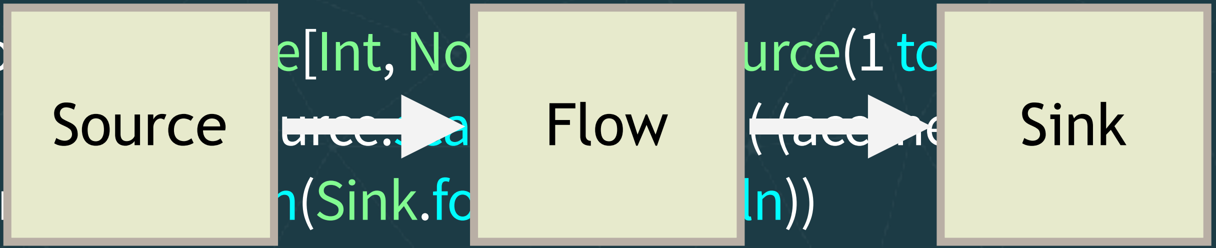
Output to a Sink,
and run it

```scala
import akka.stream._
import akka.stream.scaladsl._
import akka.NotUsed
import akka.actor.ActorSystem
import scala.concurrent._
import scala.concurrent.duration._

implicit val system = ActorSystem("QuickStart")
implicit val materializer = ActorMaterializer()
```

A source, flow, and sink constitute a graph

```scala
val so e[Int, No urce(1 to
val fa urce.sca ((acc, ne next )
factor n(Sink.fo ln))
```



akka streams

- This example is included in the project:
  - akkaStreamsModelServer/simple-akka-streams-example.sc
- To run it (showing the different prompt!):

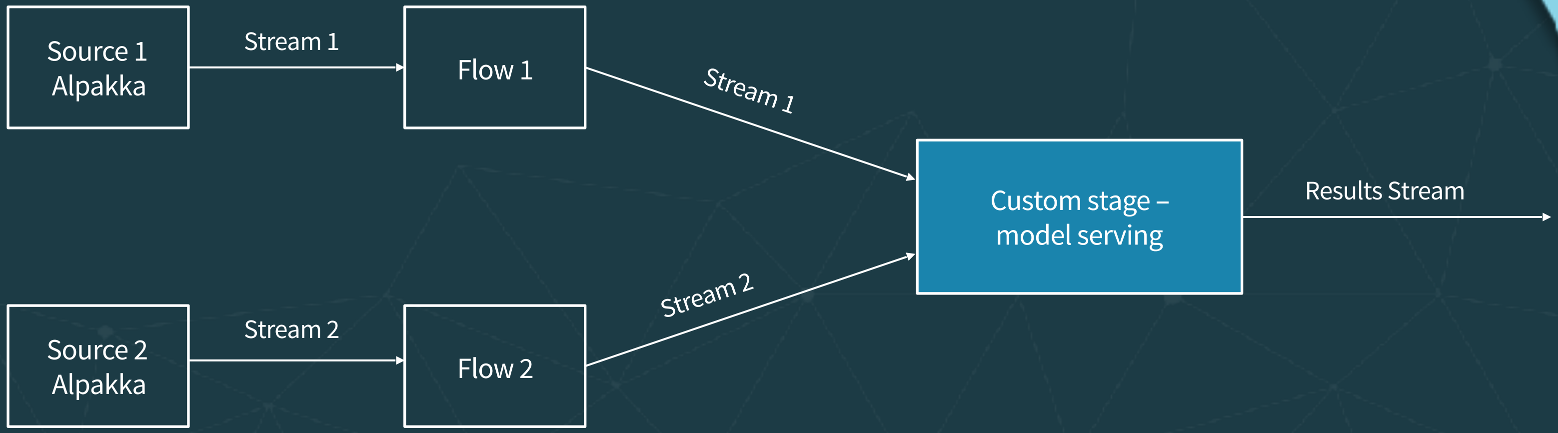
```
$ sbt
sbt:akkaKafkaTutorial> project akkaStreamsModelServer
sbt:akkaStreamsModelServer> console
scala> :load akkaStreamsModelServer/simple-akka-streams-example.sc
```

Implementations

- How do we integrate model serving (or any other new capability) into an Akka Streams app? We'll look at two approaches:
  - Implement a *Custom Stage*. Once implemented, you use it like any other "step" in the Akka Streams app.
  - Make asynchronous calls to Akka Actors to do anything you want…

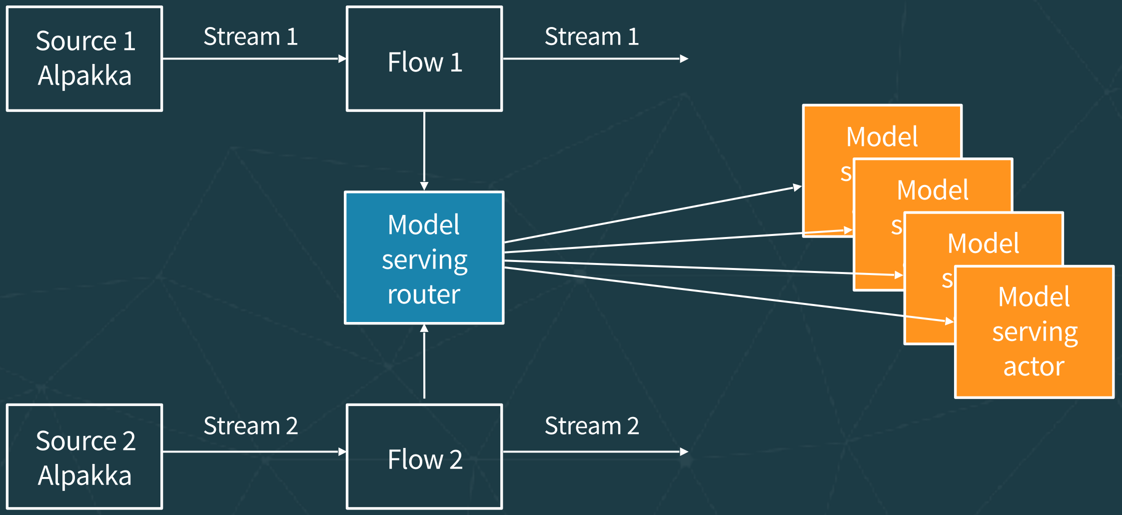
Using Custom Stage

Create a custom stage, a fully type-safe way to encapsulate new functionality. Like adding a new “operator”.



Using Akka Actors

Use a router actor to forward requests to the actor(s) responsible for processing requests for a specific model type. Clone for scalability!!



Akka Streams Example

Code time

1. Run the *client* project (if not already running)
2. Explore and run *akkaStreamsModelServer* project
    1. Use the `c` or `custom` (or default) command-line argument for the *custom stage*
    2. Use the `a` or `actor` command-line argument for the *actor model server*
    3. Use `–h` or `––help` for help

Akka Streams Example

Check Queryable state

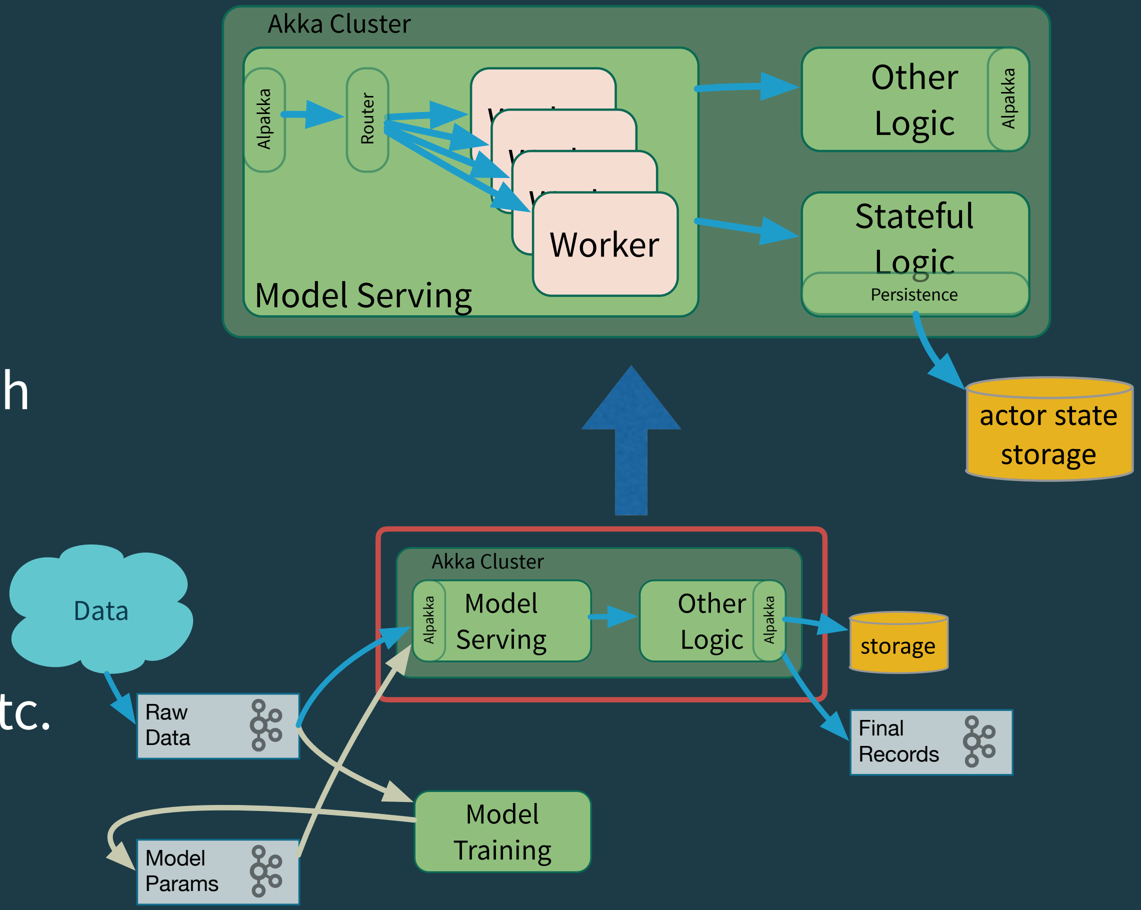
- For custom stage go to
  http://localhost:5500/state
- For actor-based implementation go to:
  http://localhost:5500/models
  http://localhost:5500/state/wine

Exercises!

- We’ve prepared some exercises. We’ll return to them after discussing Kafka Streams.
- To find them, search for `// Exercise` comments in the code base.

Other Production Concerns

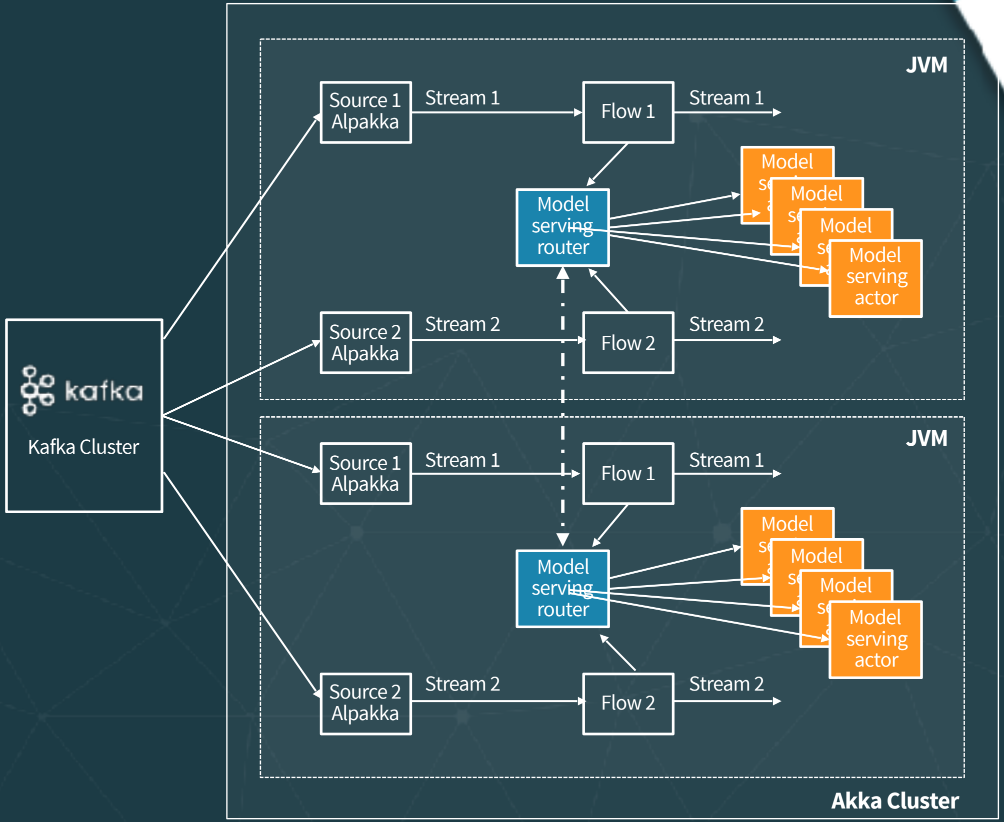
- Scale scoring with workers and routers, across a cluster
- Persist actor state with Akka Persistence
- Connect to *almost* anything with Alpakka
- *Lightbend Enterprise Suite*
  - for production monitoring, etc.



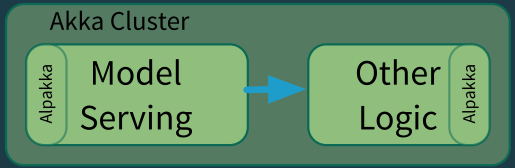
Using Akka Cluster

Two levels of scalability:

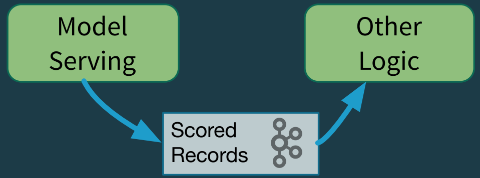
- Kafka partitioned topic allow to scale listeners according to the amount of partitions.
- Akka cluster sharing allows to split model serving actors across clusters.



Go Direct or Through Kafka?



vs.

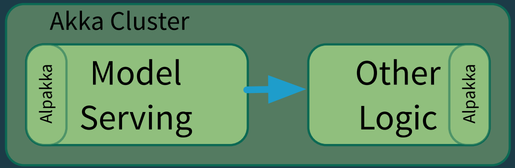


?

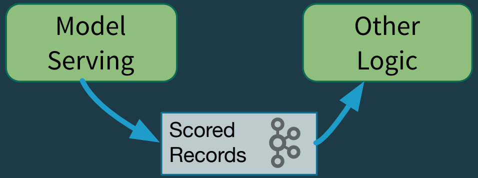
- Extremely low latency
- Minimal I/O and memory overhead
- No marshaling overhead (maybe…)

- Higher latency (including queue depth)
- Higher I/O and processing (marshaling) overhead
- Better potential reusability

Go Direct or Through Kafka?



vs.



?

- *Reactive Streams* back pressure
- Direct coupling between sender and receiver, but indirectly through an ActorRef

- Very deep buffer (partition limited by disk size)
- Strong decoupling - M producers, N consumers, completely disconnected

Kafka Streams

Kafka Streams

- Important stream-processing concepts, e.g.,
  - Distinguish between *event time* and *processing time*
  - Windowing support.
  - For more on these concepts, see
    - Dean's O'Reilly report ;)
    - Talks, blog posts, & book by Tyler Akidau

Kafka Streams

- KStream - per-record transformations
- KTable - key/value store of supplemental data
    - Efficient management of application state

Kafka Streams

- Low overhead
- Read from and write to Kafka topics, memory
  - Could use Kafka Connect for other sources and sinks
- Load balance and scale based on partitioning of topics
- Built-in support for Queryable State

Kafka Streams

- Two types of APIs:
  - Processor Topology API
    - Compare to Apache Storm
  - DSL based on collection transformations
    - Compare to Spark, Flink, Scala collections.

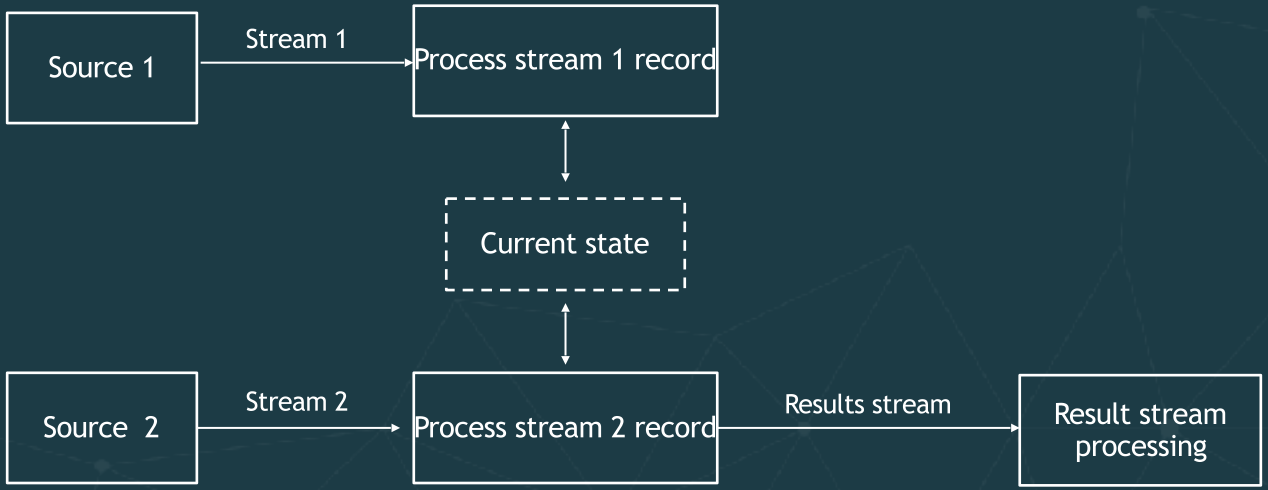
Kafka Streams

- Started with a Java API
- Lightbend donated a Scala API to Kafka
  - https://github.com/apache/kafka/tree/trunk/streams/streams-scala
  - See also our convenience tools for distributed, queryable state: https://github.com/lightbend/kafka-streams-query
- SQL - yes, but requires a specialized application (i.e., not a library like in Spark or Flink)

Kafka Streams

- Ideally suited for:
  - ETL -> KStreams
  - State -> KTable
  - Joins, including Stream and Table joins
  - “Effectively once” semantics
- Commercial support from Confluent, Lightbend, Hadoop vendors, and others

Model Serving With Kafka Streams



Lightbend

State Store Options We'll Explore

- "Naive", in memory store (no durability!)
  - Also uses the KS [Processor Topology API]
- Built-in key/value store provided by Kafka Streams
  - Uses the KS [DSL]
- Custom store
  - Also uses the DSL

Model Serving With Kafka Streams

Code time

1. Run the *client* project (if not already running)
2. Explore and run *kafkaStreamsModelServer* project
   1. Use the `c` or `custom` (or default) command-line argument for the *custom state store*
   2. Use the `s` or `standard` command-line argument for the KS built-in *standard store*
   3. Use the `m` or `memory` command-line argument for the *in-memory store*
   4. Use `–h` or `––help` for help

Model Serving With Kafka Streams

Check Queryable state

- For in Memory implementation
  - http://localhost:8888/state/value
- For build in Standard Store
  - http://localhost:8888/state/instances
  - http://localhost:8888/state/value
- For Custom store
  - http://localhost:8888/state/instances
  - http://localhost:8888/state/value

Additional architectural concerns for model serving

- Model tracking
- Speculative model execution

Model tracking - Motivation

- You update your model periodically
- You score a particular record **R** with model version **N**
- Later, you audit the data and wonder why **R** was scored the way it was

- You can't answer the question unless you know which model version was actually used for **R**

Model tracking

- Need to remember models - a model repository
- Basic info for the model:
  - Name
  - Version (or other unique ID)
  - Creation date
  - Quality metric
  - Definition
  - …

Model tracking

- You also need to augment the records with the model ID, as well as the score.
  - Input Record

…	…	…	…

  - Output Record with Score, model version ID

…	…	…	…	score	ID

Speculative execution

According to Wikipedia speculative execution is:

- an **optimization** technique
- The **system** performs work that may not be needed, before it’s known if it will be needed
- So, if and when we discover it IS needed, we don’t have to wait
- Or, results are discarded if not needed.

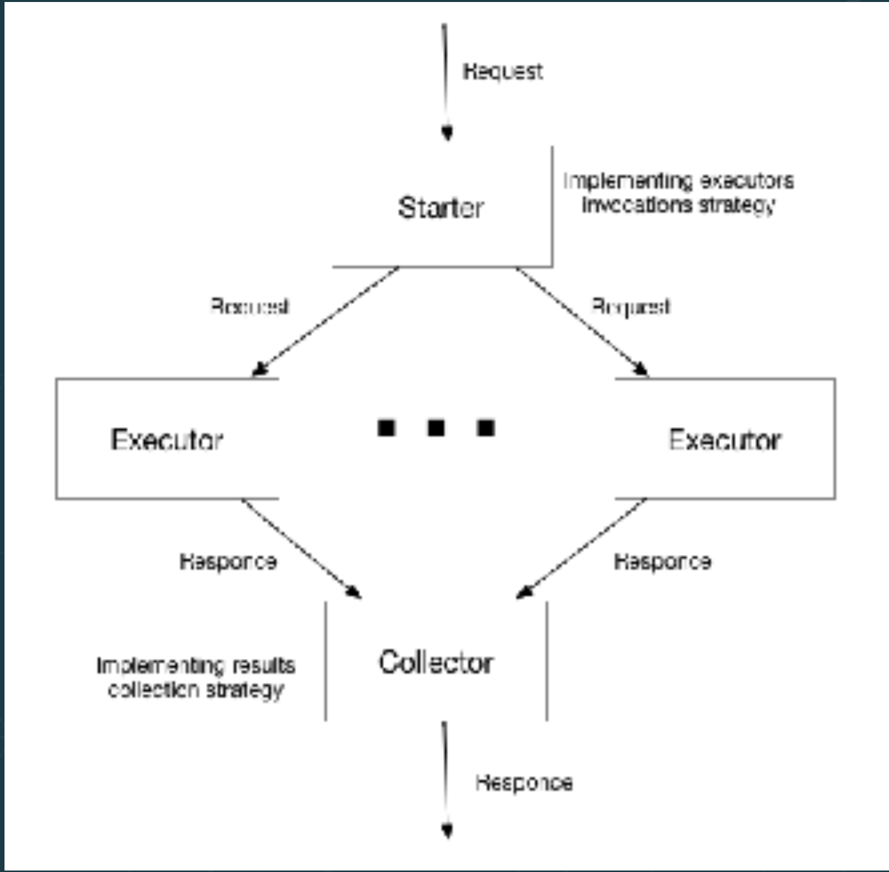
Speculative execution

- Provides more **concurrency** if extra **resources** are available.
- Used for:
  - **branch prediction** in **pipelined processors**,
  - value prediction for exploiting value locality,
  - prefetching **memory** and **files**,
  - etc.

    - Why not use it with machine learning??

General Architecture for speculative execution

- Starter (proxy) controlling parallelism and invocation strategy.
- Parallel execution by identical executors
- Collector responsible for bringing results from multiple executors together



Applicability for model serving

- Used to guarantee execution time
  - Several models:
    - A smart model, but takes time $T1$
    - A "less smart", but fast model with a fixed upper-limit on execution time, $T2 \ll T1$
  - If timeout ($T > T2$) occurs before smart finishes, return the less accurate result

  - Why is $T > T2$ required??
  - …

Applicability for model serving

- …
- Consensus based model serving
  - If we have 3 or more models, score with all of them and return the majority result
- …

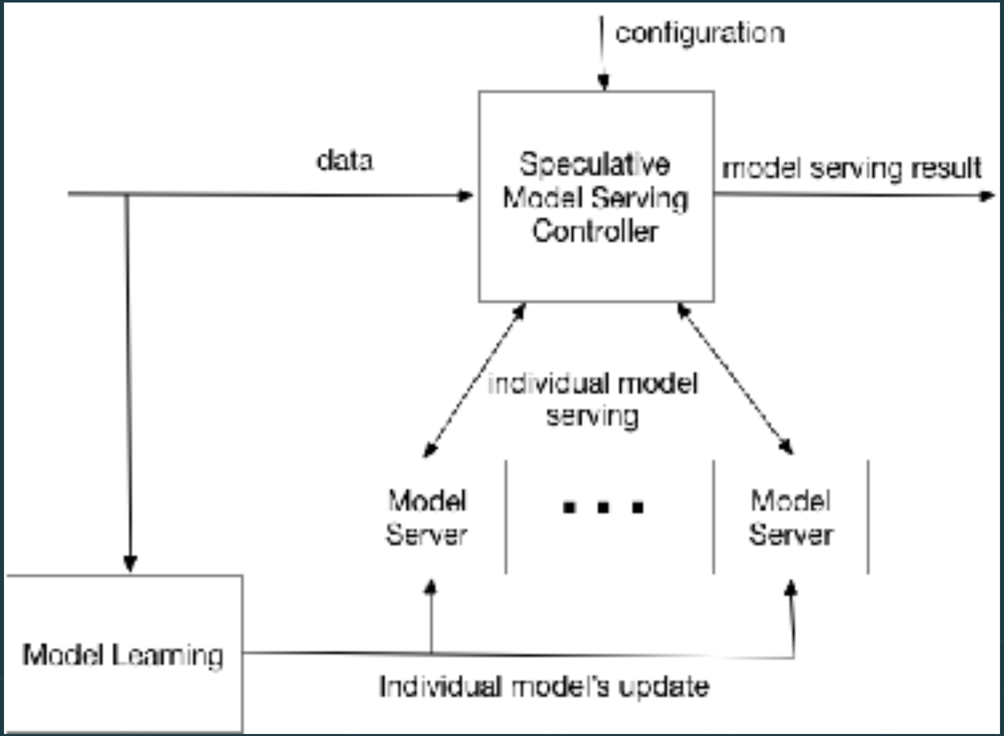
Lightbend

Applicability for model serving

- …
- Quality based model serving.
  - If we have a quality metric, pick the result with the best result.

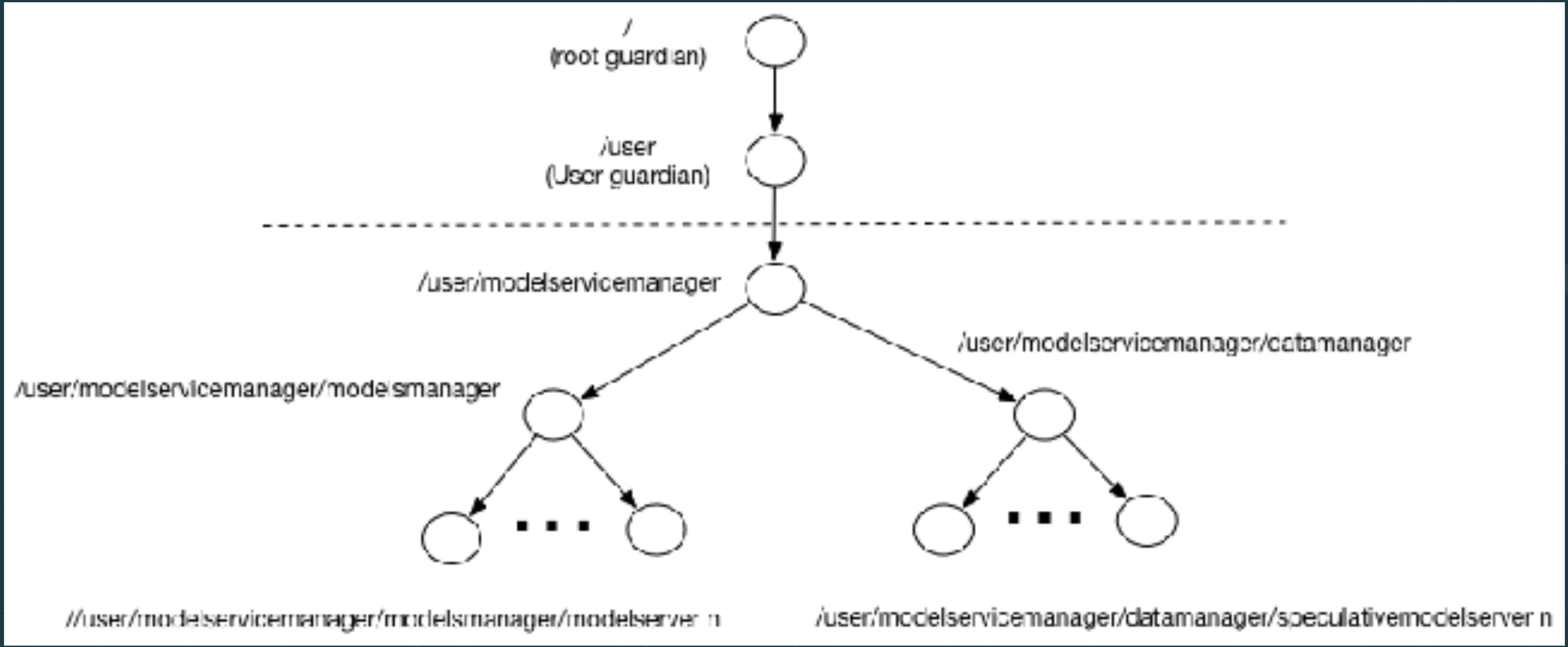
- Of course, you can combine these techniques.

Architecture



https://developer.lightbend.com/blog/2018-05-24-speculative-model-serving/index.html

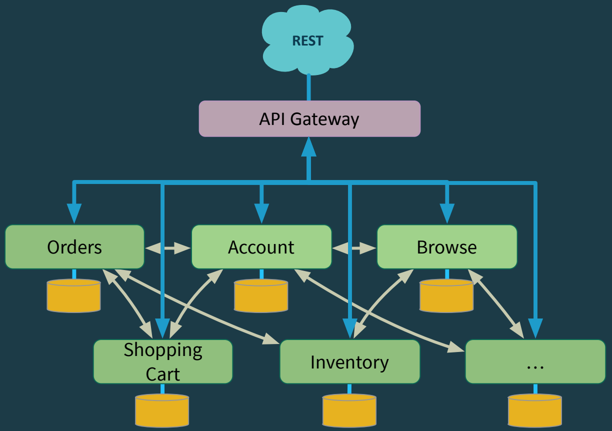
Actors


/
(root guardian)
/user
(User guardian)
/user/modelservicemanager
/user/modelservicemanager/datamanager
/user/modelservicemanager/modelsmanager
/user/modelservicemanager/datamanager/speculativemodelserver n

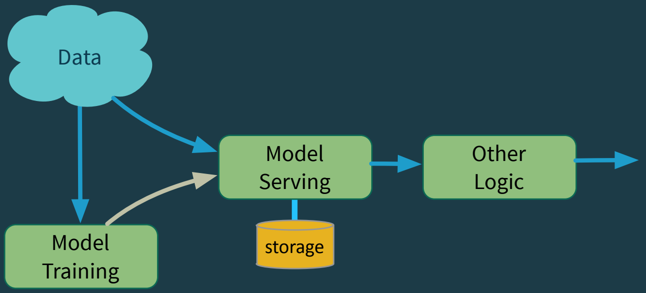
Wrapping Up

To Wrap Up

Event-driven µ-services



"Record-centric" µ-services



Events

Records

In Our Remaining Time Today… (1/2)

1. Explore the code we didn't discuss (there is a lot ;)
    1. Study the different model serving techniques
    2. Study the "model" subproject
    3. Look at how the following are implemented
        1. queryable state
        2. embedded web servers
        3. use of Akka Persistence
        4. model serialization
2. …

In Our Remaining Time Today... (2/2)

1. …
2. Try the exercises - search for `// Exercise` in the code
3. Ask us for help on anything…
4. Visit lightbend.com/fast-data-platform
5. Profit!!

Other Lightbend Talks

- Other Lightbend sessions this week:
  - Gerard Maas, Processing fast data with Apache Spark: A tale of two APIs, Wednesday, 11:20-12:00, 1E 07/08
  - Dean, Executive Briefing: What you need to know about fast data, Thursday, 2:00-2:40, 1E 14.
  - .. and visit the Lightbend booth!!

Thanks for coming!

Questions?

lightbend.com/products/fast-data-platform
boris.lublinsky@lightbend.com
dean.wampler@lightbend.com