WEBINAR | THURSDAY DEC 6TH, 9:00 AM PT / 5:00 PM UTC

Hands On With Spark: Creating A Fast Data Pipeline With Structured Streaming And Spark Streaming

Gerard Maas,
Senior Engineer at Lightbend, Inc.

Lightbend

Gerard Maas

Señor SW Engineer

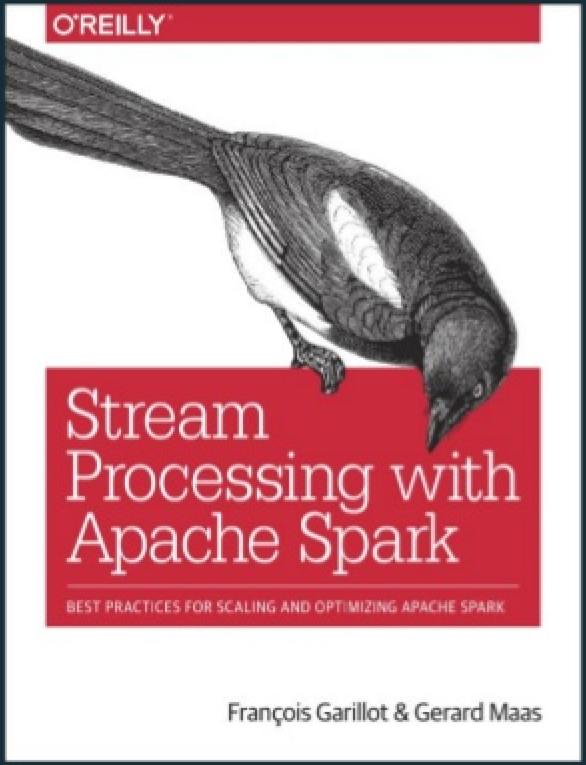
Lightbend

@maasg

https://github.com/maasg

https://www.linkedin.com/in/gerardmaas/

https://stackoverflow.com/users/764040/maasg

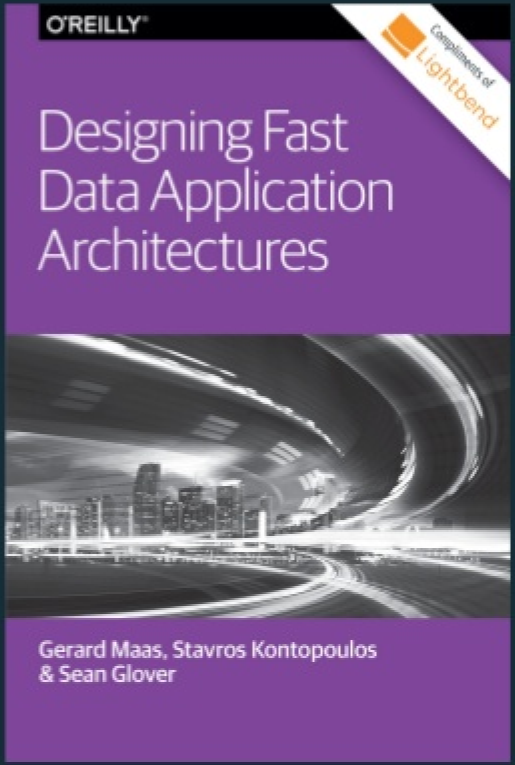


O'REILLY

Stream Processing with Apache Spark

BEST PRACTICES FOR SCALING AND OPTIMIZING APACHE SPARK

François Garillot & Gerard Maas



O'REILLY

Compliments of Lightbend

Designing Fast Data Application Architectures

Gerard Maas, Stavros Kontopoulos & Sean Glover

Agenda

Creating a Fast Data Pipeline with Structured Streaming and Spark Streaming

Data Pipelines

Data Pipelines

Data

Kafka

Processor

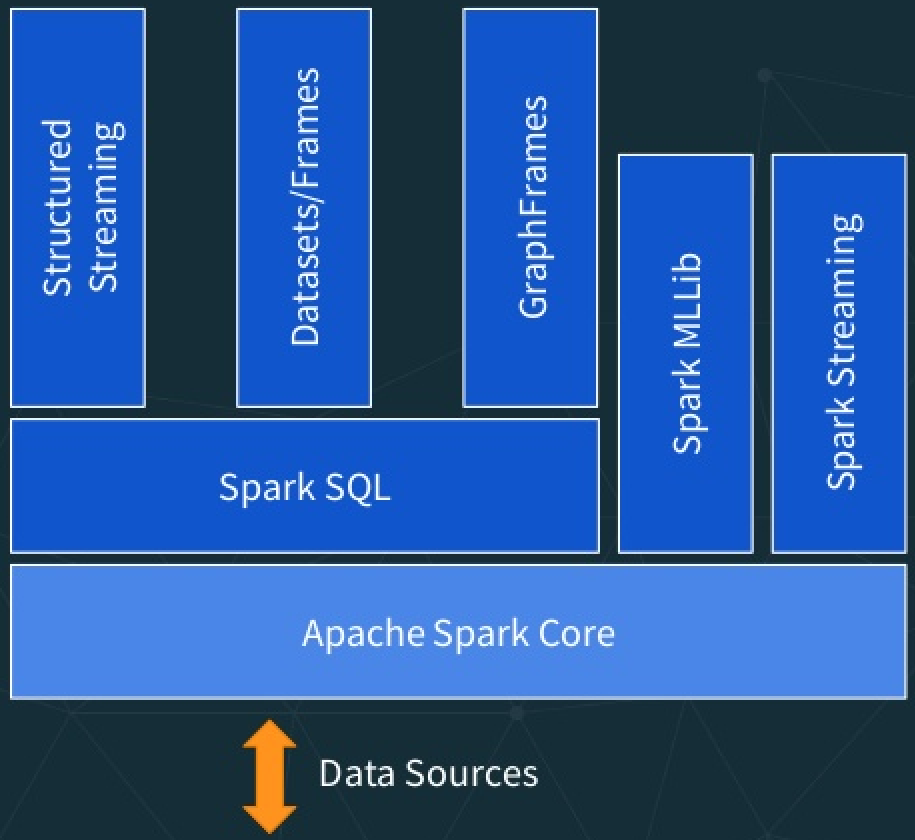
Data Pipelines

- Create Composable Streaming Applications
- Using the Best Tool for the Job
- Generating a Network Effect

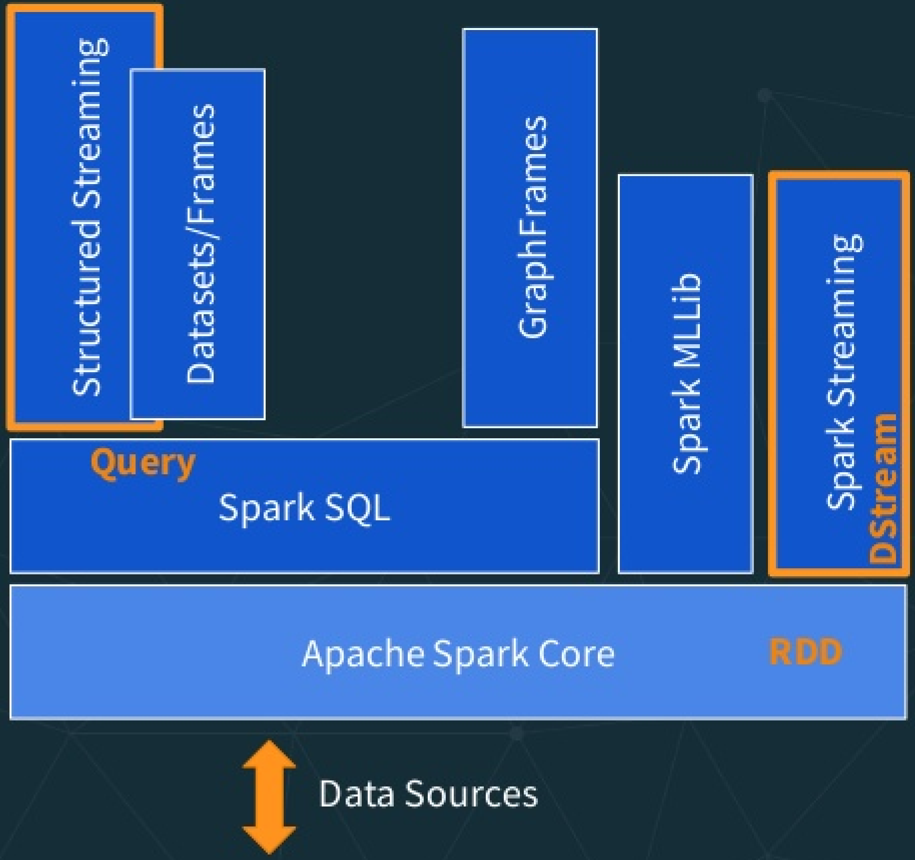
Agenda

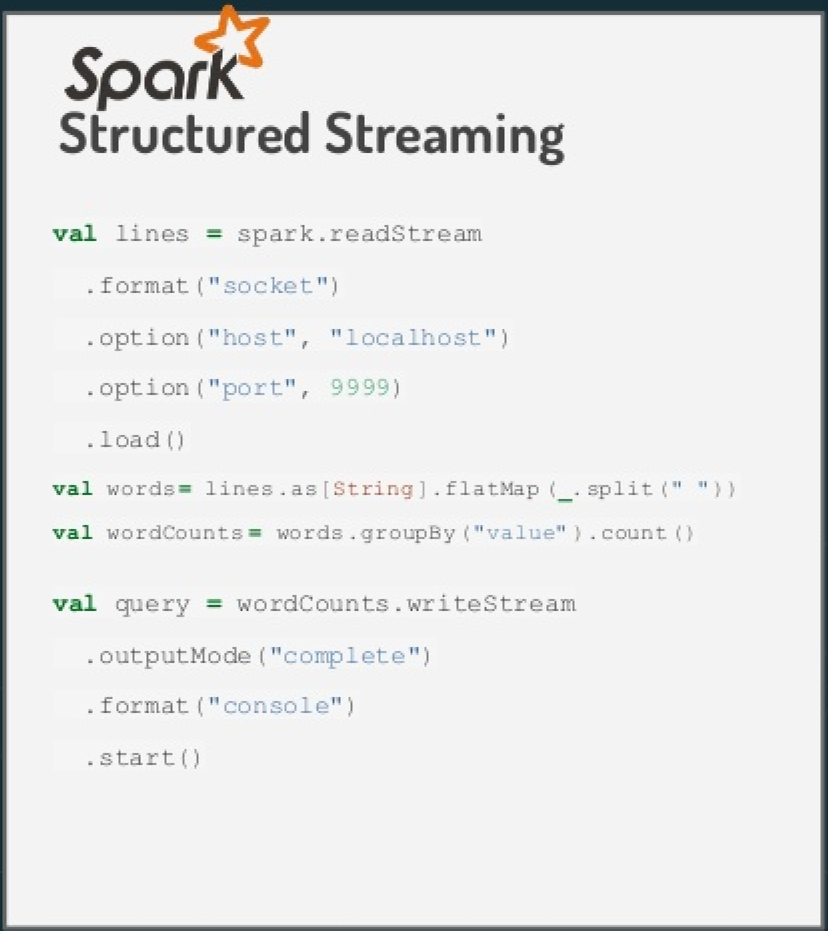
Creating a Fast Data Pipeline with **Structured Streaming** and **Spark Streaming**

Scala Java python R



Scala Java python R

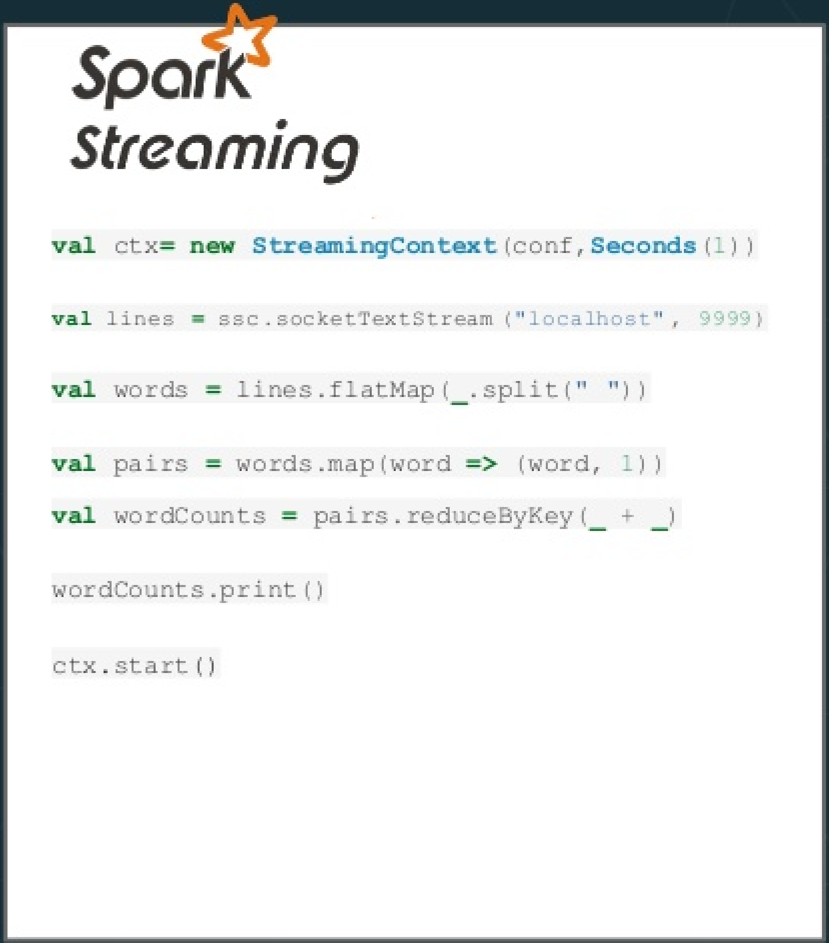




Spark Structured Streaming

```
val lines = spark.readStream
  .format("socket")
  .option("host", "localhost")
  .option("port", 9999)
  .load()
val words= lines.as[String].flatMap(_.split(" "))
val wordCounts= words.groupBy("value").count()

val query = wordCounts.writeStream
  .outputMode("complete")
  .format("console")
  .start()
```



Spark Streaming

```
val ctx= new StreamingContext(conf,Seconds(1))

val lines = ssc.socketTextStream("localhost", 9999)

val words = lines.flatMap(_.split(" "))

val pairs = words.map(word => (word, 1))
val wordCounts = pairs.reduceByKey(_ + _)

wordCounts.print()

ctx.start()
```

Lightbend

https://spark.apache.org/docs/latest/streaming-programming-guide.html
https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html

@maasg

	Structured Streaming	Spark Streaming
Time	Abstract (Processing Time, Event Time)	Fixed to microbatch Streaming Interval
Execution	Fixed Micro batch, Best Effort MB, Continuous (NRT)	Fixed Micro batch
Abstraction	DataFrames/Dataset	DStream, RDD

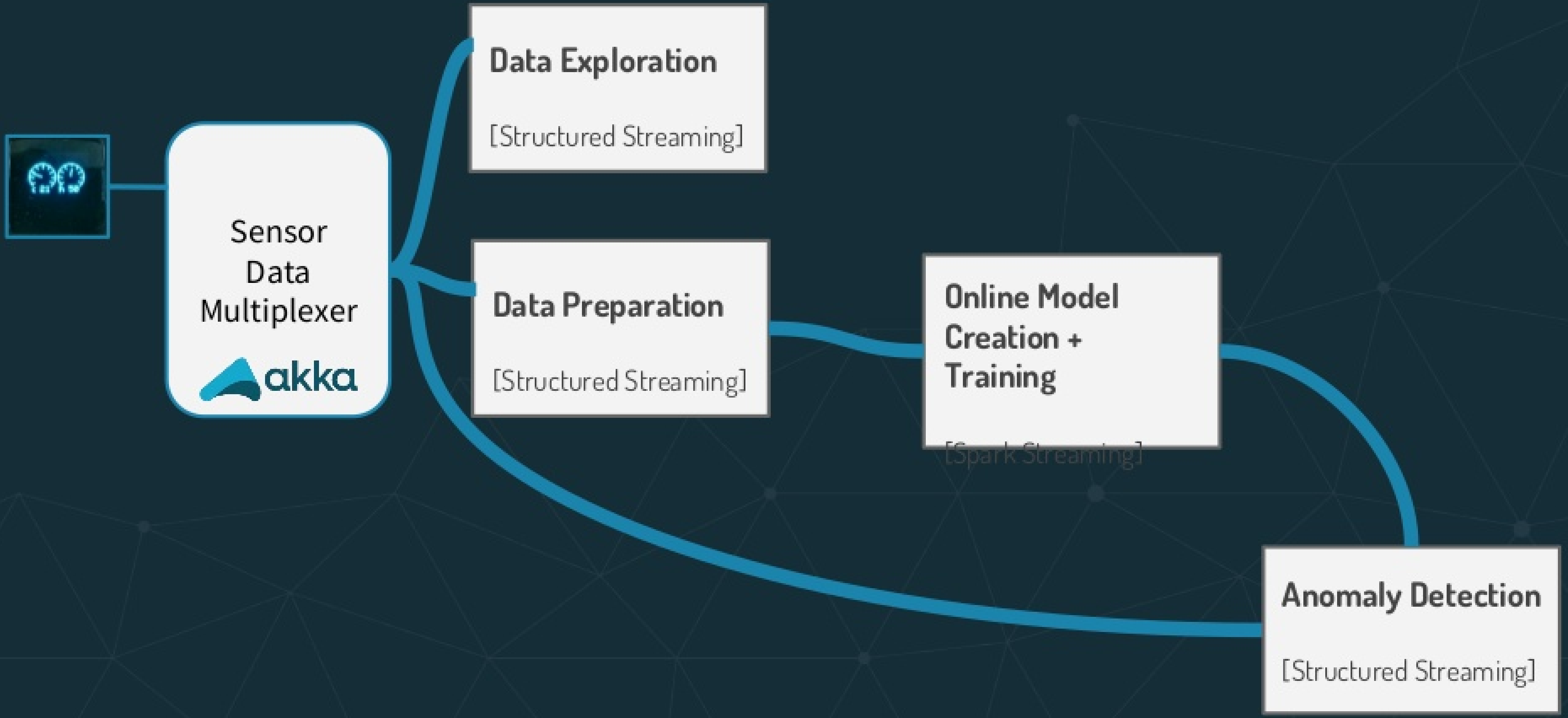
★ Access to the scheduler

Agenda

Hands On with Spark:

Creating a Fast Data Pipeline with Structured Streaming and Spark Streaming

Sensor Anomaly Detection Pipeline



Sensor Data Multiplexer

akka

Data Exploration

[Structured Streaming]

Data Preparation

[Structured Streaming]

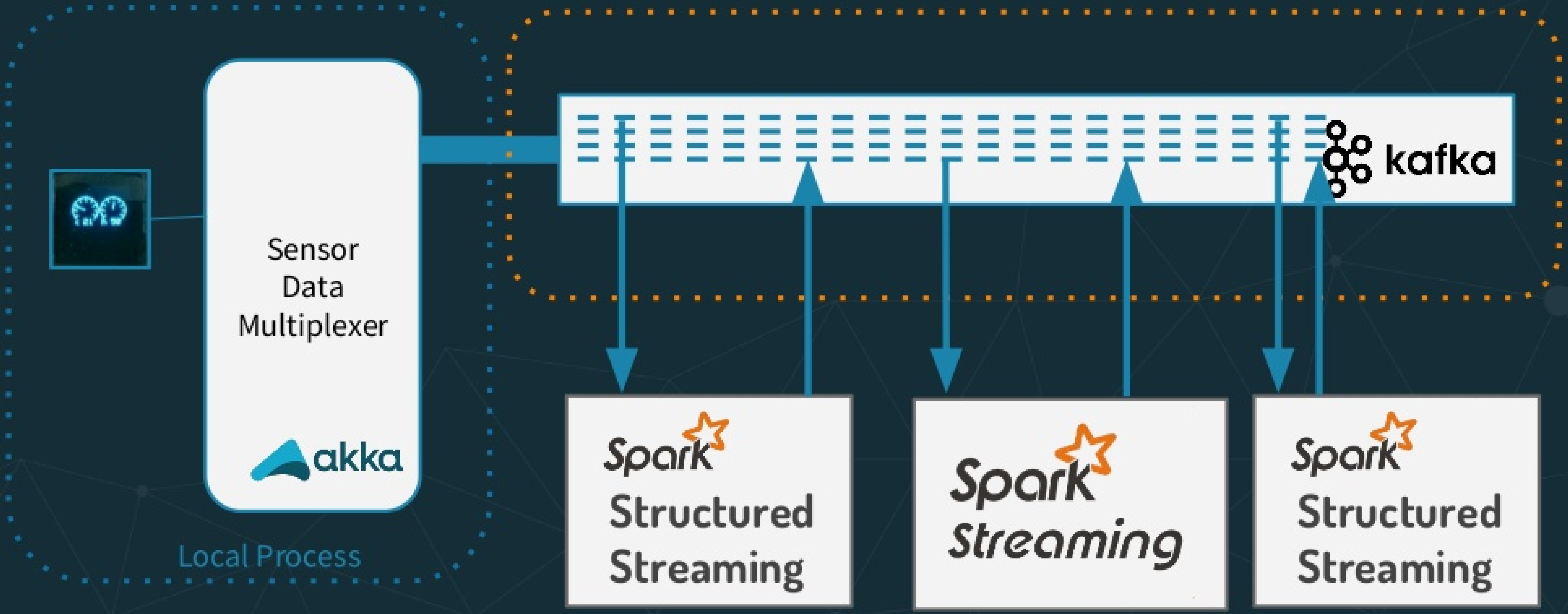
Online Model Creation + Training

[Spark Streaming]

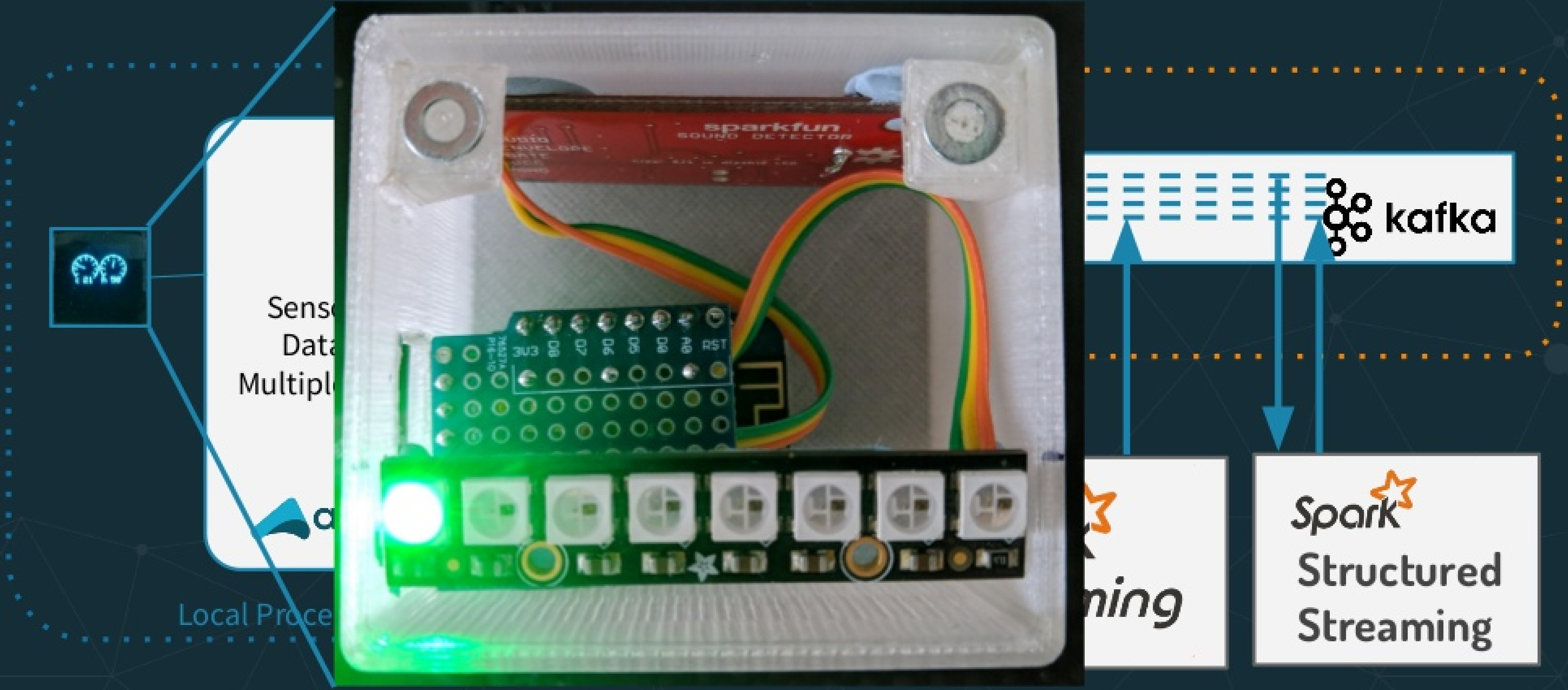
Anomaly Detection

[Structured Streaming]

Sensor Anomaly Detection

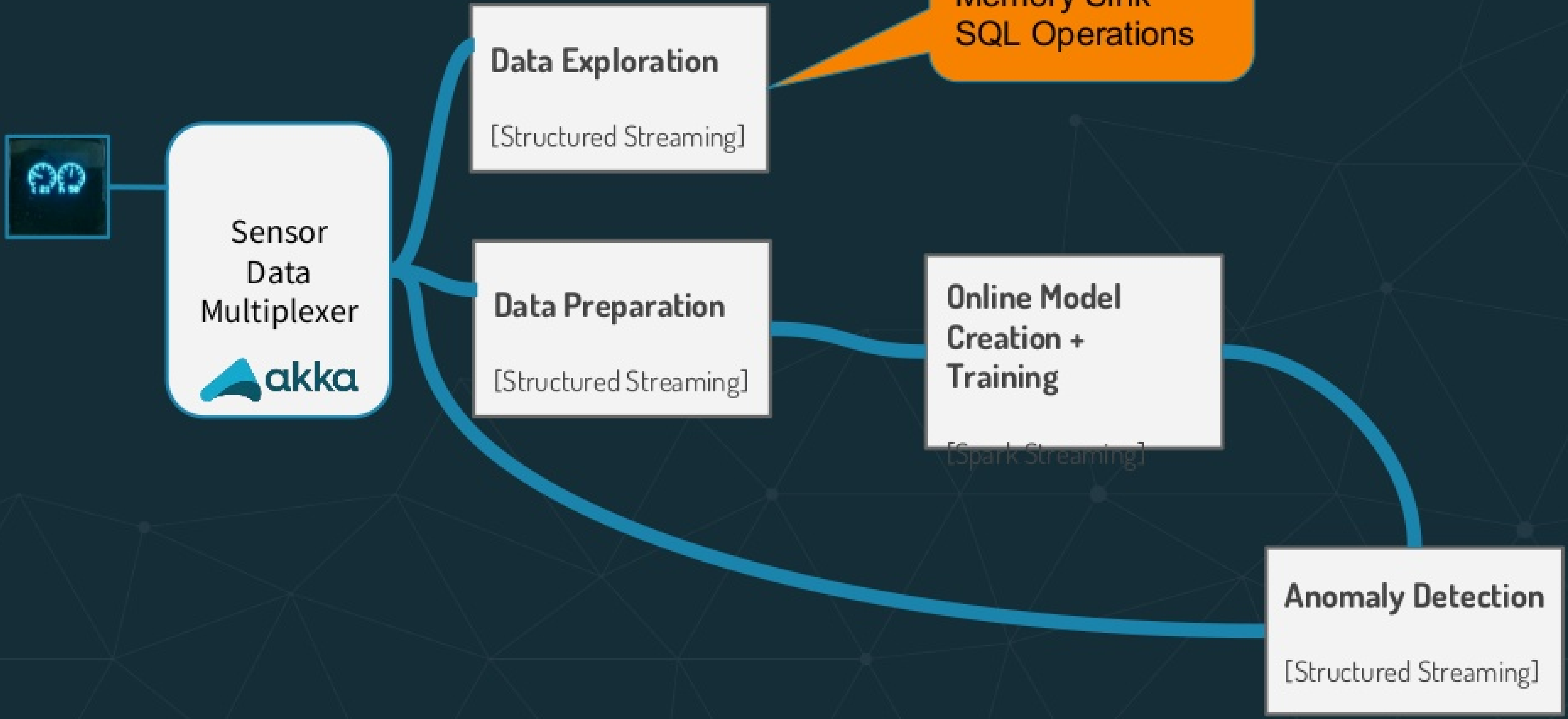


Sensor Anomaly Detection



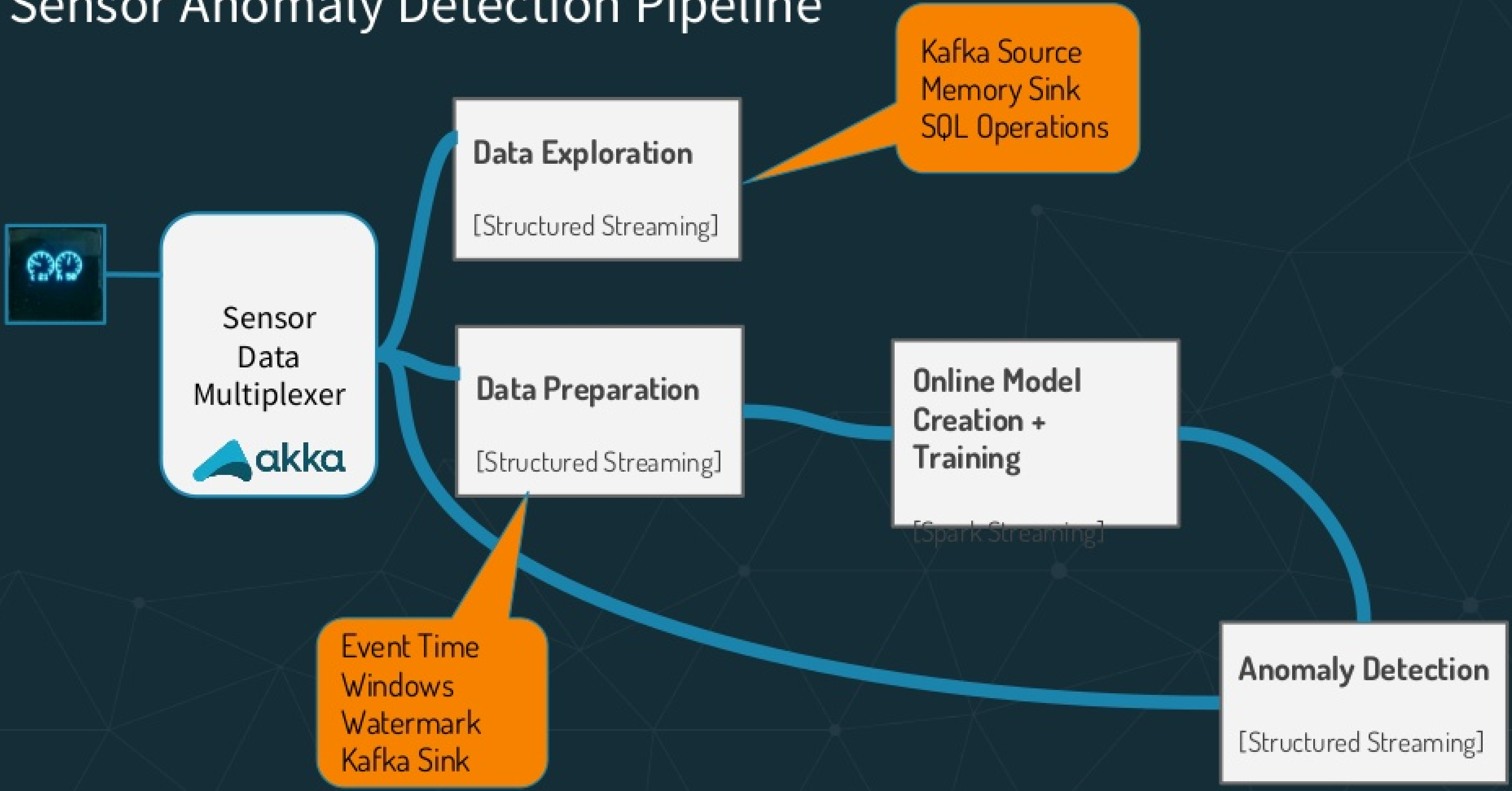
● Live

Sensor Anomaly Detection Pipeline



Kafka Source
Memory Sink
SQL Operations

Sensor Anomaly Detection Pipeline



Sensor Data Multiplexer

akka

Data Exploration

[Structured Streaming]

Kafka Source
Memory Sink
SQL Operations

Data Preparation

[Structured Streaming]

Event Time
Windows
Watermark
Kafka Sink

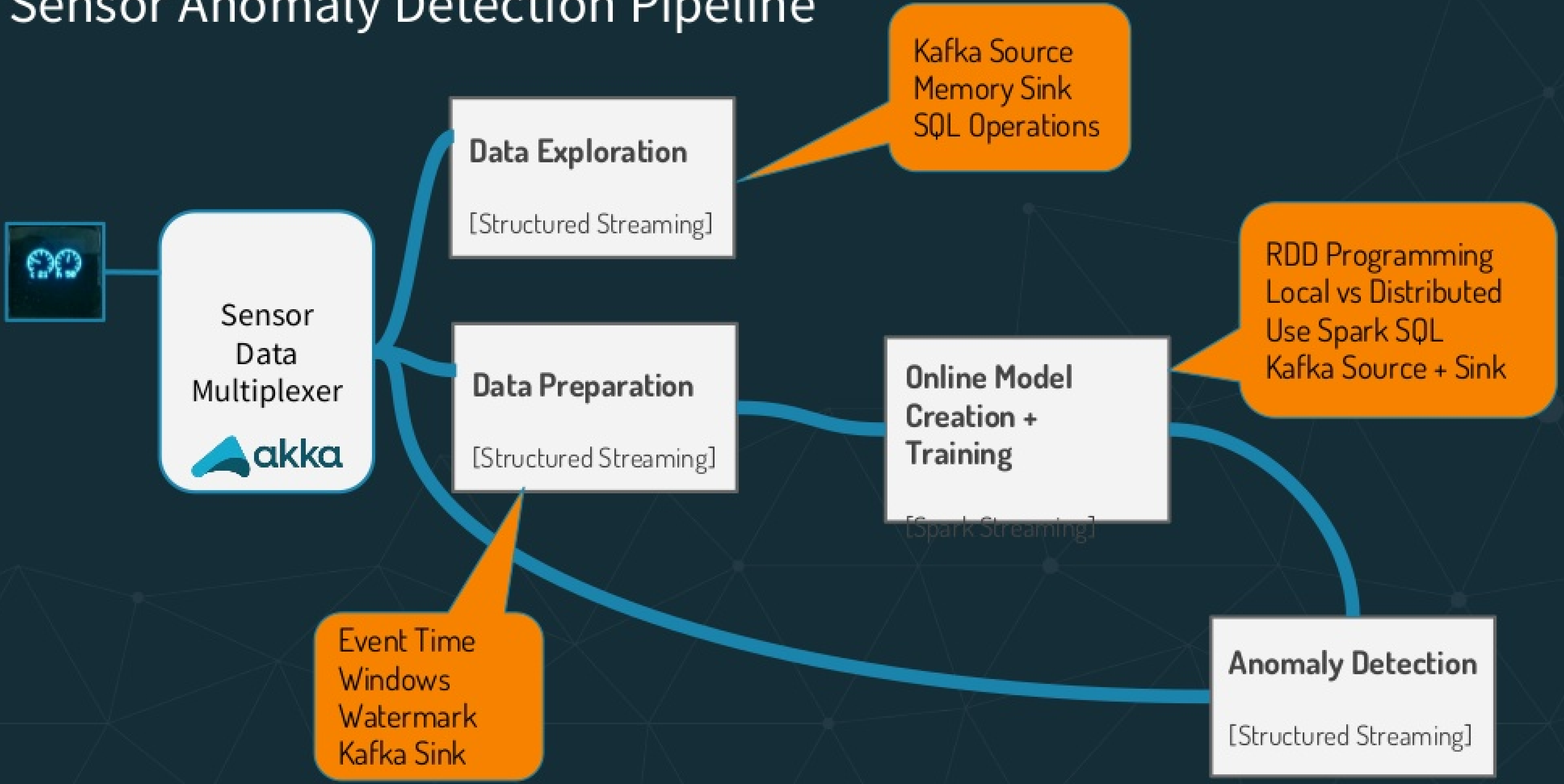
Online Model Creation + Training

[Spark Streaming]

Anomaly Detection

[Structured Streaming]

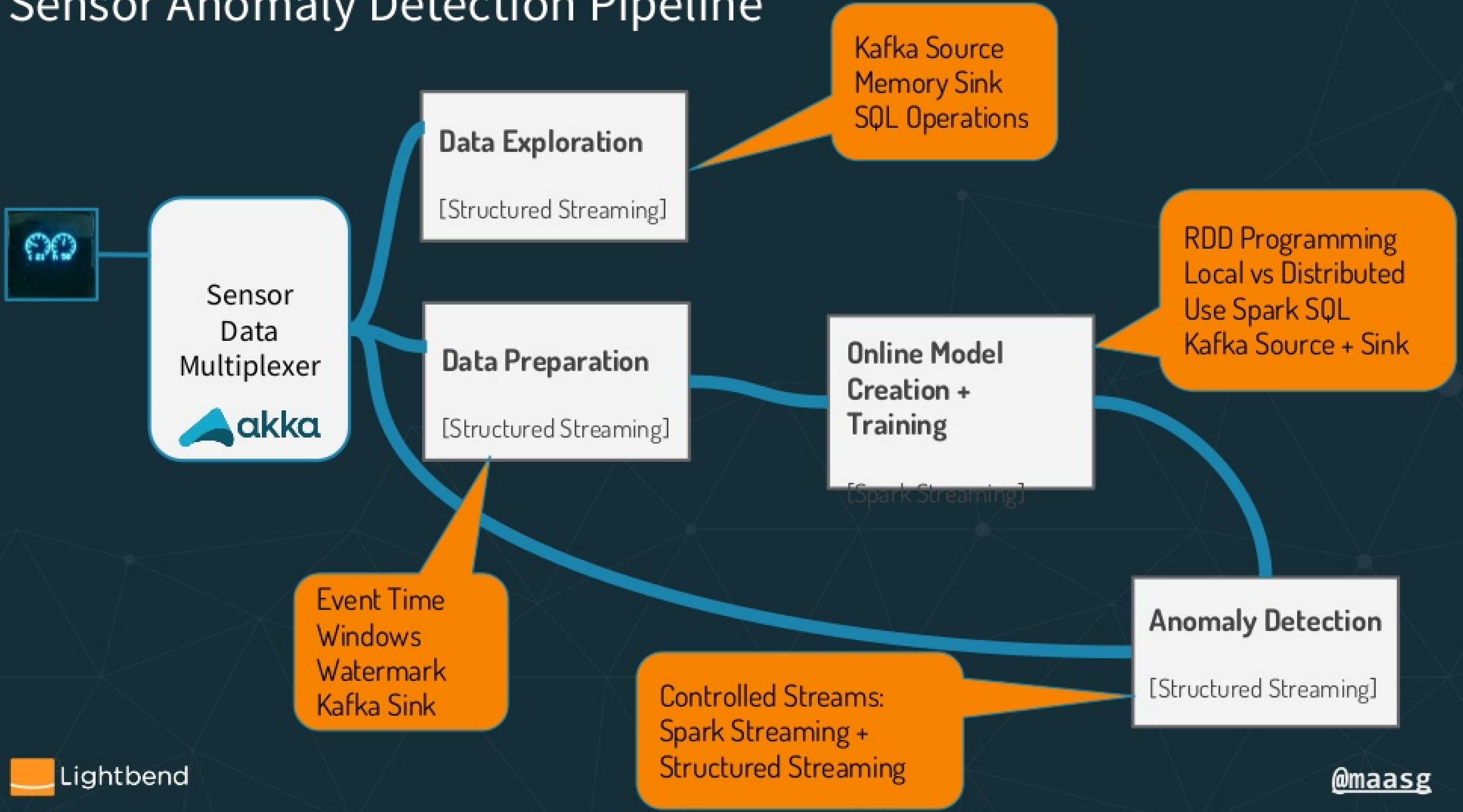
Sensor Anomaly Detection Pipeline



Lightbend

@maasg

Sensor Anomaly Detection Pipeline



Sensor Data Multiplexer

akka

Data Exploration

[Structured Streaming]

Kafka Source
Memory Sink
SQL Operations

Data Preparation

[Structured Streaming]

Event Time
Windows
Watermark
Kafka Sink

Online Model Creation + Training

[Spark Streaming]

RDD Programming
Local vs Distributed
Use Spark SQL
Kafka Source + Sink

Anomaly Detection

[Structured Streaming]

Controlled Streams:
Spark Streaming +
Structured Streaming

	Structured Streaming	Spark Streaming
Time	Abstract (Processing Time, Event Time)	Fixed to microbatch Streaming Interval
Execution	Fixed Micro batch, Best Effort MB, Continuous (NRT)	Fixed Micro batch
Abstraction	DataFrames/Dataset	DStream, RDD

★ Access to the scheduler

Resources

Notebooks used today:

https://github.com/maasg/spark-notebooks/tree/master/streaming-anomaly-detection

Pipelines:

https://www.reactivesummit.org/2018/schedule/taking-the-pain-out-of-deploying-streaming-applications

Structured Streaming + Spark Streaming:

https://www.reactivesummit.org/2018/schedule/processing-fast-data-with-apache-spark-the-tale-of-two-streaming-apis

Fast Data:

https://www.lightbend.com/products/fast-data-platform

Gerard Maas

Señor SW Engineer

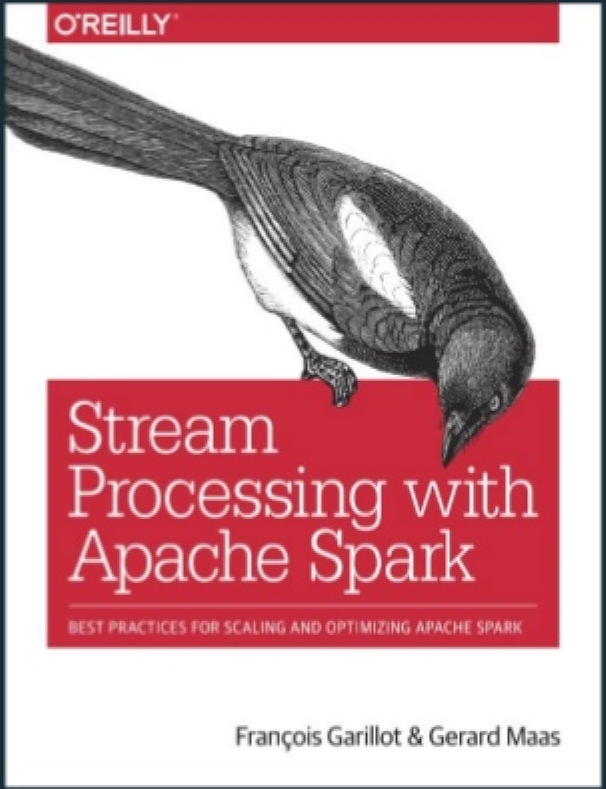
Lightbend

@maasg

https://github.com/maasg

https://www.linkedin.com/in/gerardmaas/

https://stackoverflow.com/users/764040/maasg



O'REILLY

Stream Processing with Apache Spark

BEST PRACTICES FOR SCALING AND OPTIMIZING APACHE SPARK

François Garillot & Gerard Maas

Your turn for…

Questions?

Thank You!

Lightbend