

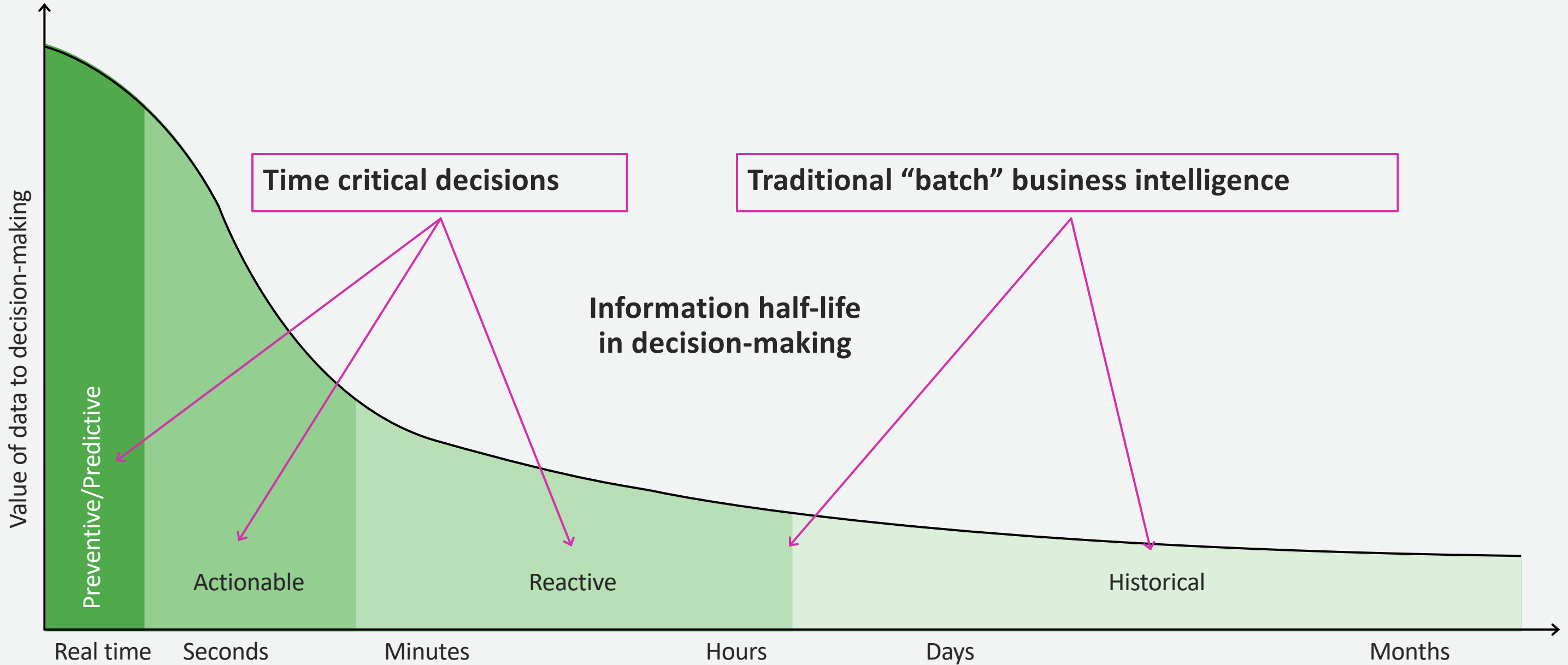


Real-Time Analytics with Kinesis

Immersion Day

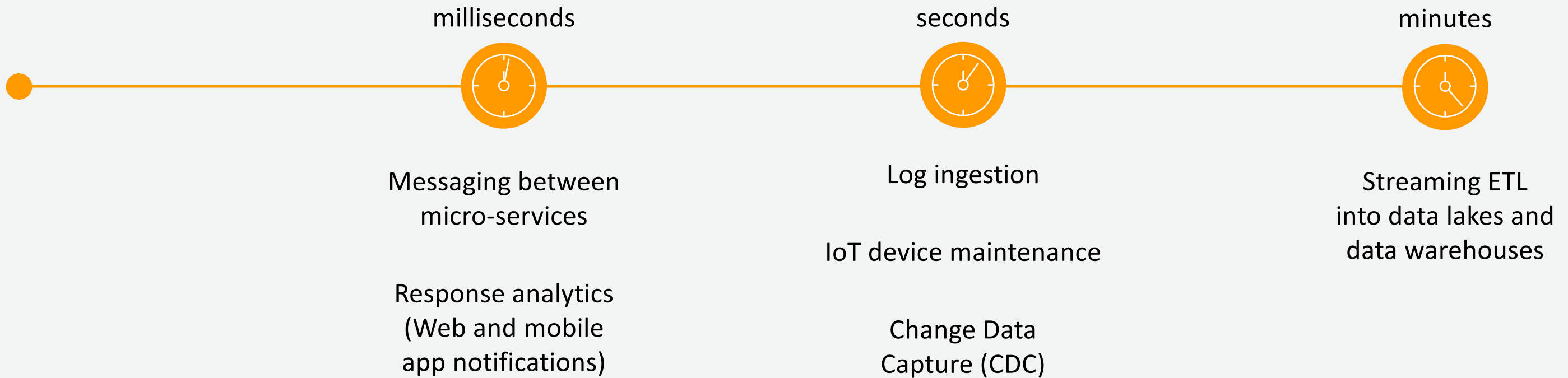
Data Streaming and Processing Overview

Why Streaming Data?



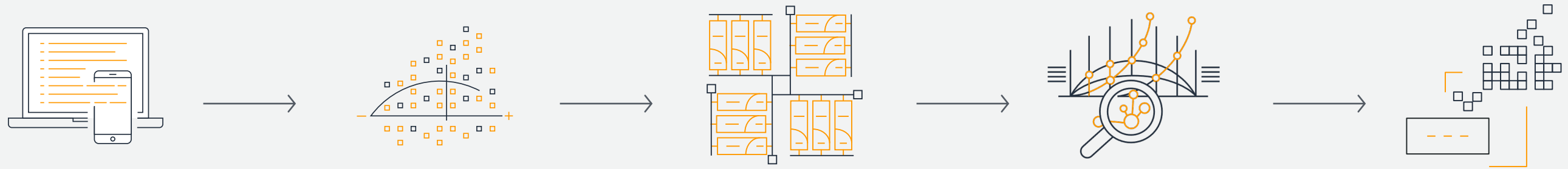
Source: Perishable insights, Mike Gualtieri, Forrester

Common real-time analytics use cases



Enabling real-time analytics

Data streaming technology enables a customer to ingest, process and analyze high volumes of high velocity data from a variety of sources **in real time**



Source

Devices and or applications that produce real-time data at high velocity.

Stream ingestion

Data from tens of thousands of data sources can be written to a single stream.

Stream storage

Data is stored in the order it was received for a set duration of time, and can be replayed indefinitely during this time.

Stream processing

Records are read in the order they are produced enabling real-time analytics or streaming ETL.

Destination

Data lake (most common)
Database (less common)

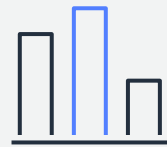
Challenges of Data Streaming



Difficult to setup



Tricky to scale



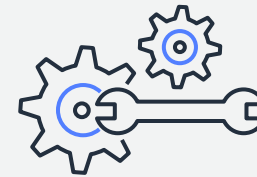
Hard to achieve high availability



Integration requires development



Error prone and complex to manage



Expensive to maintain

Streaming real-time data with AWS

Easily collect, process and analyze data streams in real time

Easy to use

Elastic

High availability
and durability

Seamless integration
with AWS services

Fully managed

Pay for what you use

Real-time Streaming on AWS

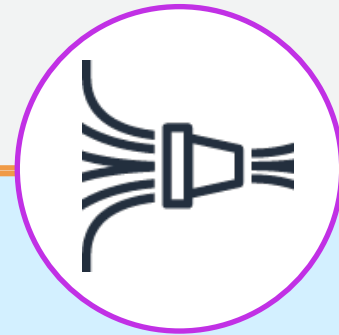
Easily collect, process, and analyze video and data streams in real time

**Kinesis
Data Streams**



Collect and store data streams for analytics

**Kinesis
Data Firehose**



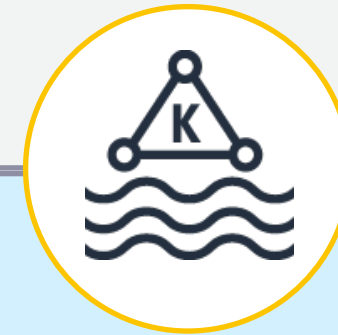
Load data streams into AWS data stores

**Kinesis
Data Analytics**



Analyze data streams with SQL or Java

**Amazon Managed
Streaming for Apache
Kafka**



Collect and store data streams for analytics

**Kinesis
Video Streams**

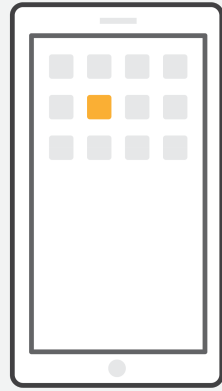


Capture and store video streams for analytics

Sources



Devices and or applications that produce real-time data at high velocity



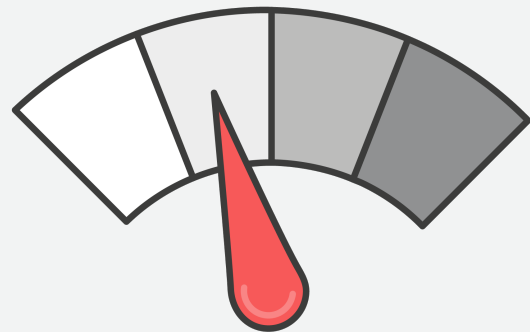
Mobile Apps



Web Clickstream

```
[Wed Oct 11 14:32:52  
2018] [error] [client  
127.0.0.1] client  
denied by server  
configuration:  
/export/home/live/ap/ht  
docs/test
```

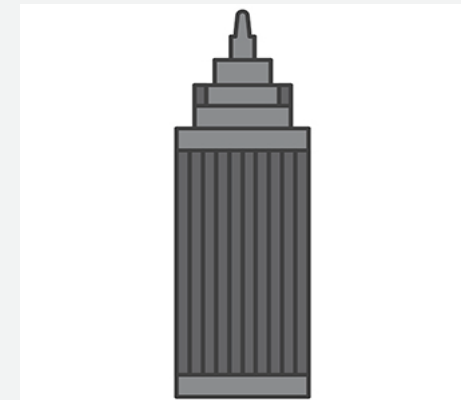
Application Logs



Metering Records



IoT Sensors







Smart Buildings

Stream Ingestion



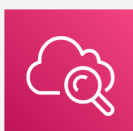
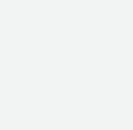


Data from tens of thousands of data sources can be written to a single stream




AWS Toolkits/Libraries

- AWS SDK 
- Kinesis Producer Library 
- AWS Mobile SDK 
- Kinesis Agent 

AWS Service Integrations

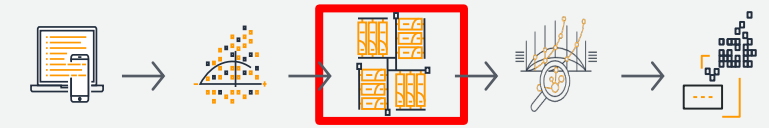
- AWS IoT 
- Amazon CloudWatch Logs 
- Amazon CloudWatch Events 
- Amazon Database Migration Service * 

3rd Party Offerings

- LOG4J 
- Flume 
- Fluentd 

* Amazon DMS supports 8 on-premise databases, 1 Azure database, 5 RDS/Aurora database types, and S3

Stream Storage



Data is stored in the order it was received for a set duration of time, and can be replayed indefinitely during this time.

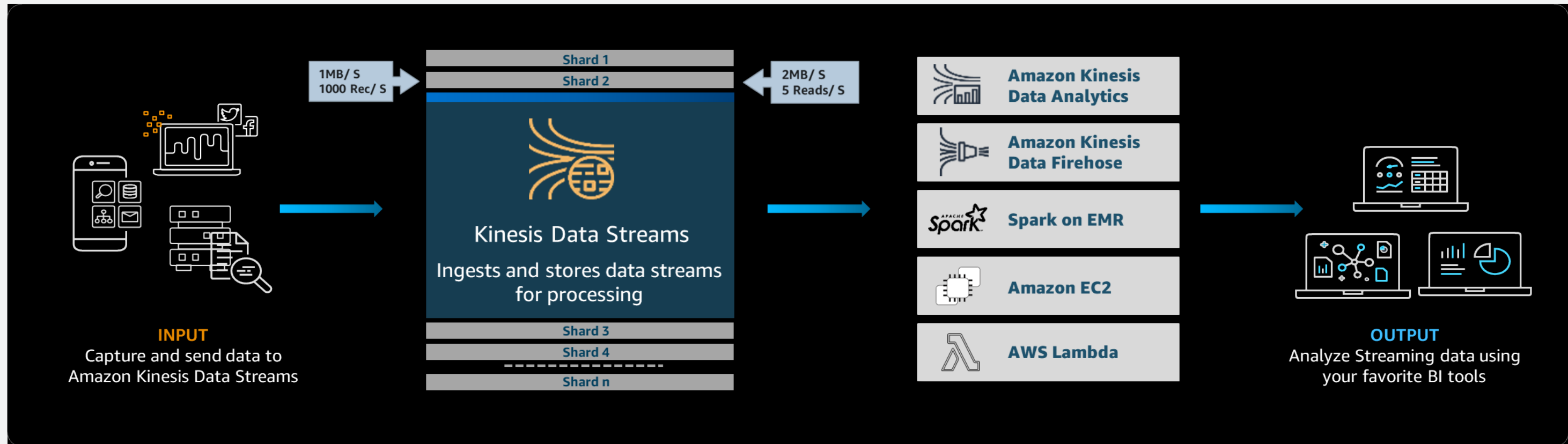
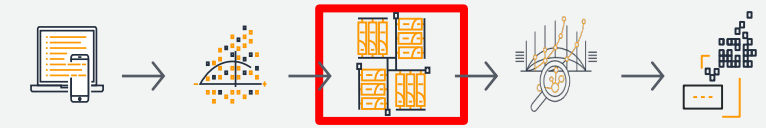


**Amazon
Kinesis Data
Streams**



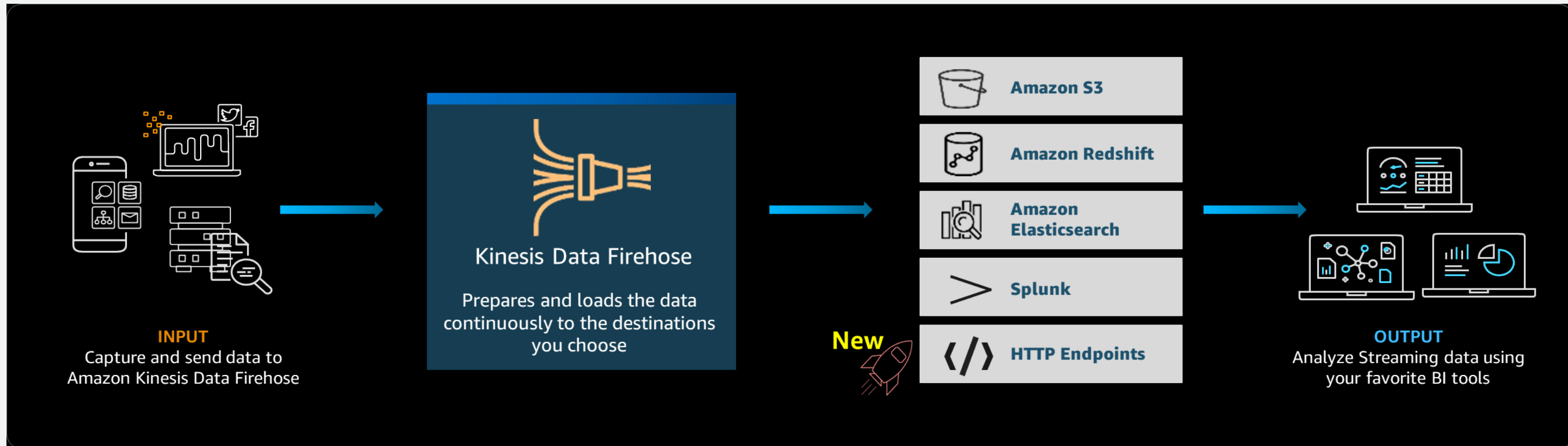
**Amazon
Managed
Streaming for
Apache Kafka**

Amazon Kinesis Data Streams



- Easy administration and low cost
- Real-time, elastic performance
- Secure, durable storage
- Available to multiple real-time analytics applications
- Average latency of 200ms with one standard consumer
- Enhanced Fan Out with SubscribeToShard API offers typical average latency of 70 ms

Amazon Kinesis Data Firehose



- Zero administration and seamless elasticity
- Direct-to-data store integration
- Serverless continuous data transformations

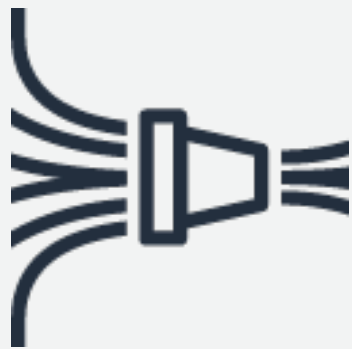
- Near real time
- Data format conversion to Parquet/ ORC
- Deliver data directly to Datadog, Sumo Logic, New Relic and MongoDB

Amazon Kinesis – Streams vs Firehose



Kinesis Data
Streams

Amazon Kinesis Data Streams is for use cases that require custom processing, per incoming record, with sub-1 second processing latency, and a choice of stream processing frameworks



Kinesis Data
Firehose


Amazon Kinesis Data Firehose is for use cases that require zero administration, ability to use existing analytics tools based on Amazon S3, Amazon Redshift, and Amazon ES, and a data latency of 60 seconds or higher

Stream Processing




Records are read in the order they are produced enabling real-time analytics or streaming ETL

Kinesis



SQL/
Java



Amazon Kinesis
Data Analytics



Kinesis Client Library
+
Connector Library

AWS Services



AWS Lambda

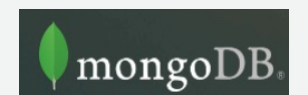


Amazon EMR

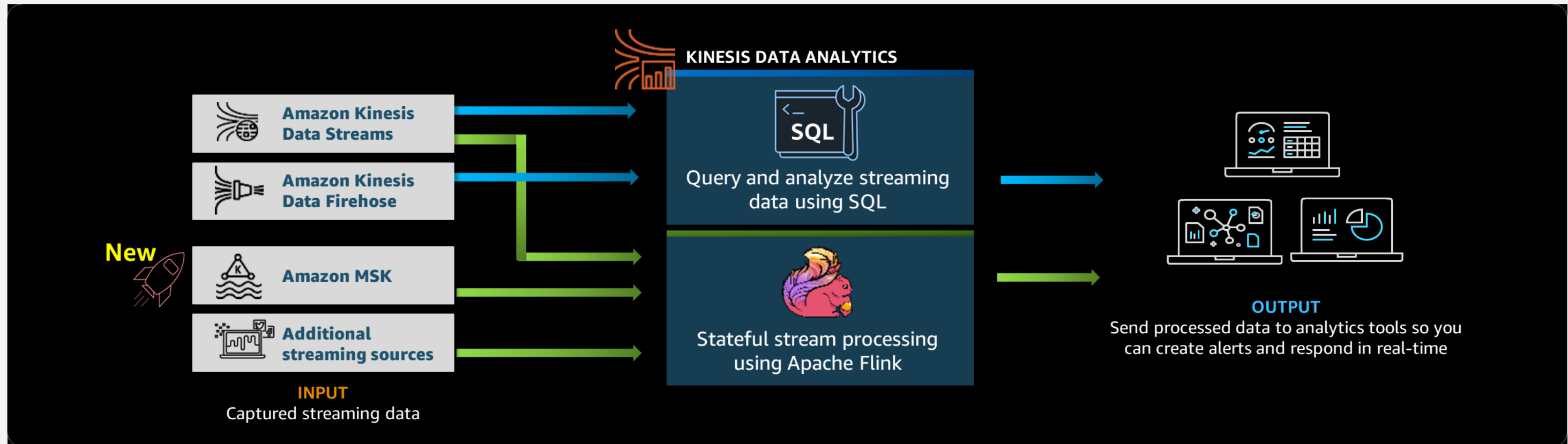
3rd party



Apache Spark



Amazon Kinesis Data Analytics



- Interact with streaming data in real-time using SQL or integrated Apache Flink applications
- Build fully managed and elastic stream processing applications

KDA SQL for simple and fast use cases



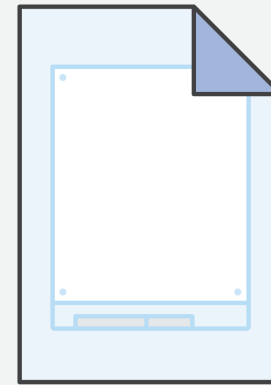
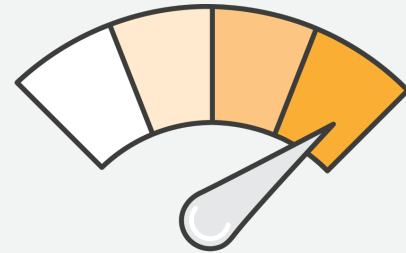
- Sub-second end to end processing latencies
- SQL steps can be chained together in serial or parallel steps
- Build applications with one or hundreds of queries
- Pre-built functions include everything from sum and count distinct to machine learning algorithms
- Aggregations run continuously using window operators



KDA Java for sophisticated applications



Utilizes Apache Flink, a framework and distributed engine for stateful processing of data streams



Simple programming

Easy to use and flexible APIs make building apps fast

High performance

In-memory computing provides low latency & high throughput

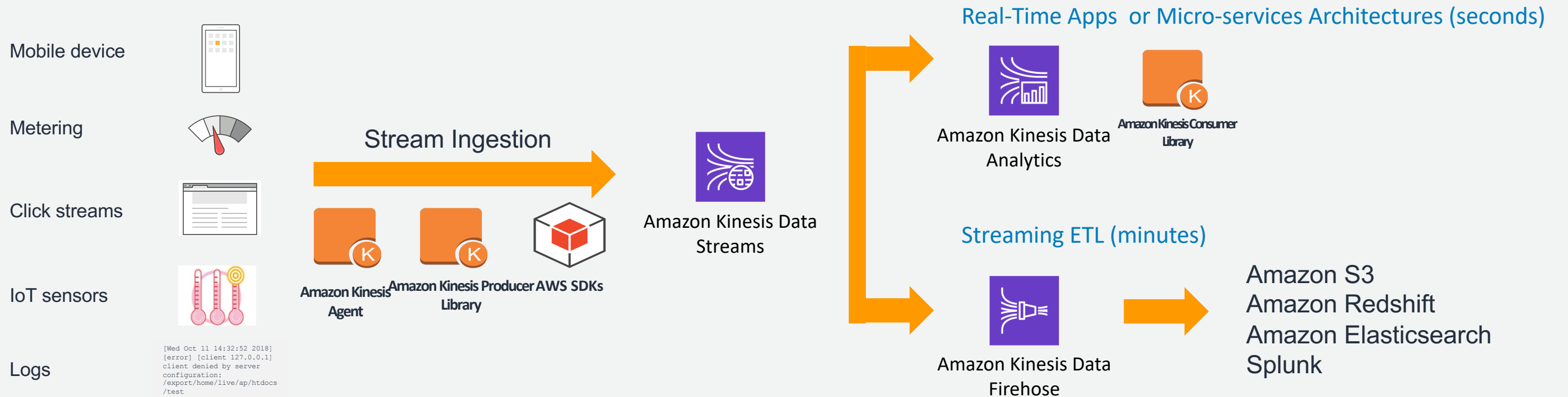
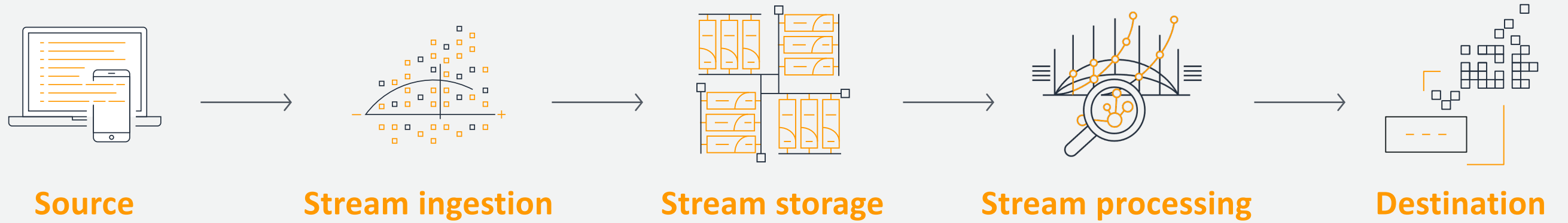
Stateful Processing

Durable application state saves

Strong data integrity

Exactly-once processing and consistent state

An Example Architecture



Thank you!