Streaming Overview

Kafka Immersion Day @ REA

Day 1

Timothy Downs, Vikas Bajaj

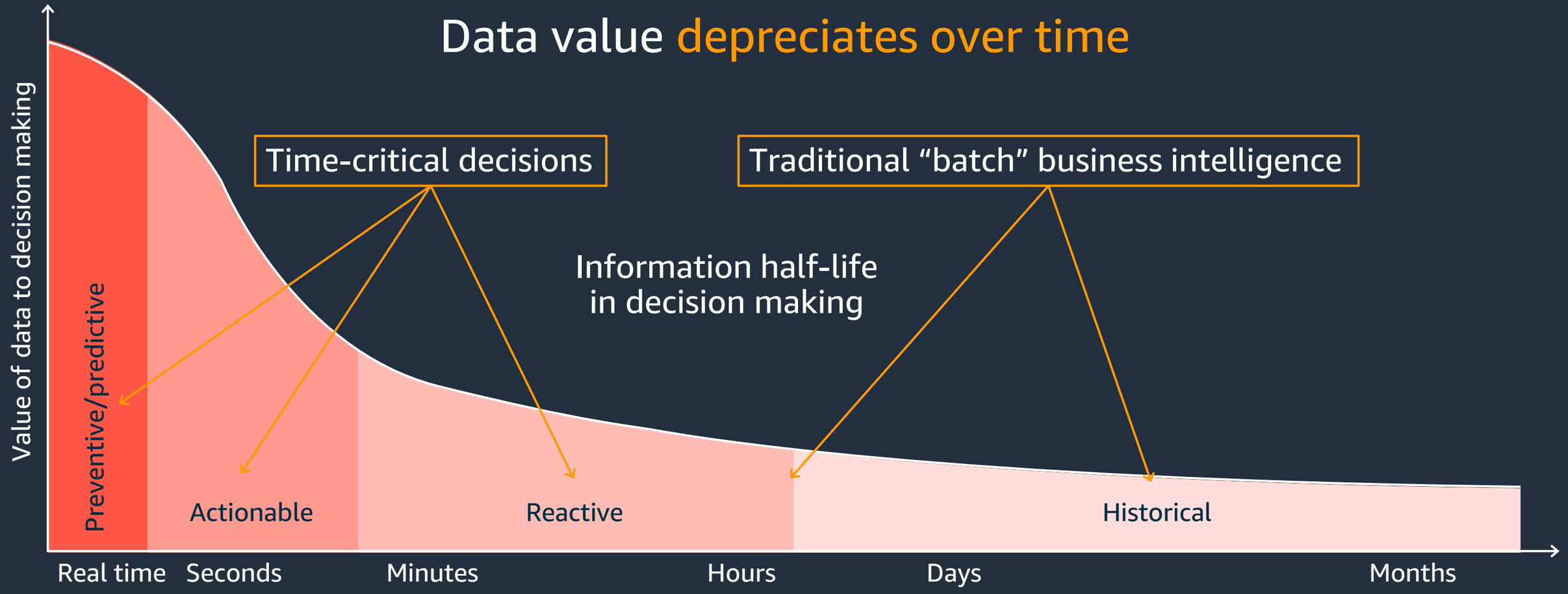
9/7/21

Agenda

- Why streaming
- What is streaming
- Streaming ecosystem
- AWS Services in the Streaming space
- Brief introduction of the AWS streaming services
- Common architectures
- Customer example

aws

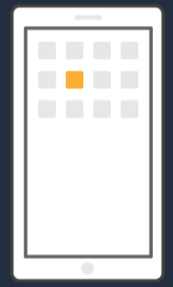
Timely decisions require new data in minutes



Source: Perishable insights, Mike Gualtieri, Forrester

aws

Data is produced continuously from a large variety of sources



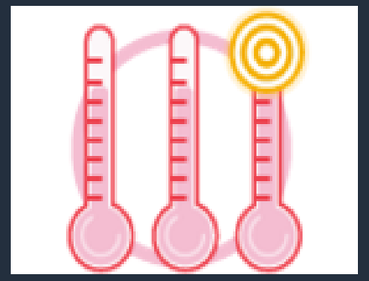
Mobile apps

Web clickstream

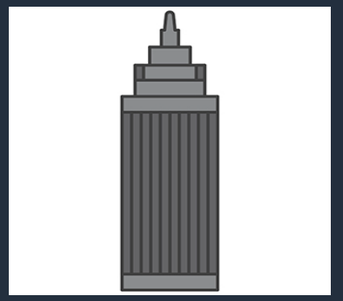
```
[Wed Oct 11 14:32:52
2018] [error] [client
127.0.0.1] client denied
by server configuration:
/export/home/live/ap/htdo
cs/test
```

Application logs

Metering records



IoT sensors



Smart buildings

aws

Why capture and process data in real time?

Businesses can more easily

- react to customers
- react to threats
- react to environmental changes
- make timely decisions
- innovate down the road

aws

The foundation of real-time analytics

1. Data is produced, delivered, and processed in milliseconds
2. Data is retained, enabling parallel and independent processing
3. Data must be captured and processed in the order it was generated

aws

Technology fit for real-time analytics

Technology	Lag	Parallel processing	Ordering guarantees
Relational database	Asynchronous	Yes	No
Message Queues	Milliseconds	No	No
Hadoop	Hours or days	Yes	No
Commit log	Milliseconds	Yes	Yes

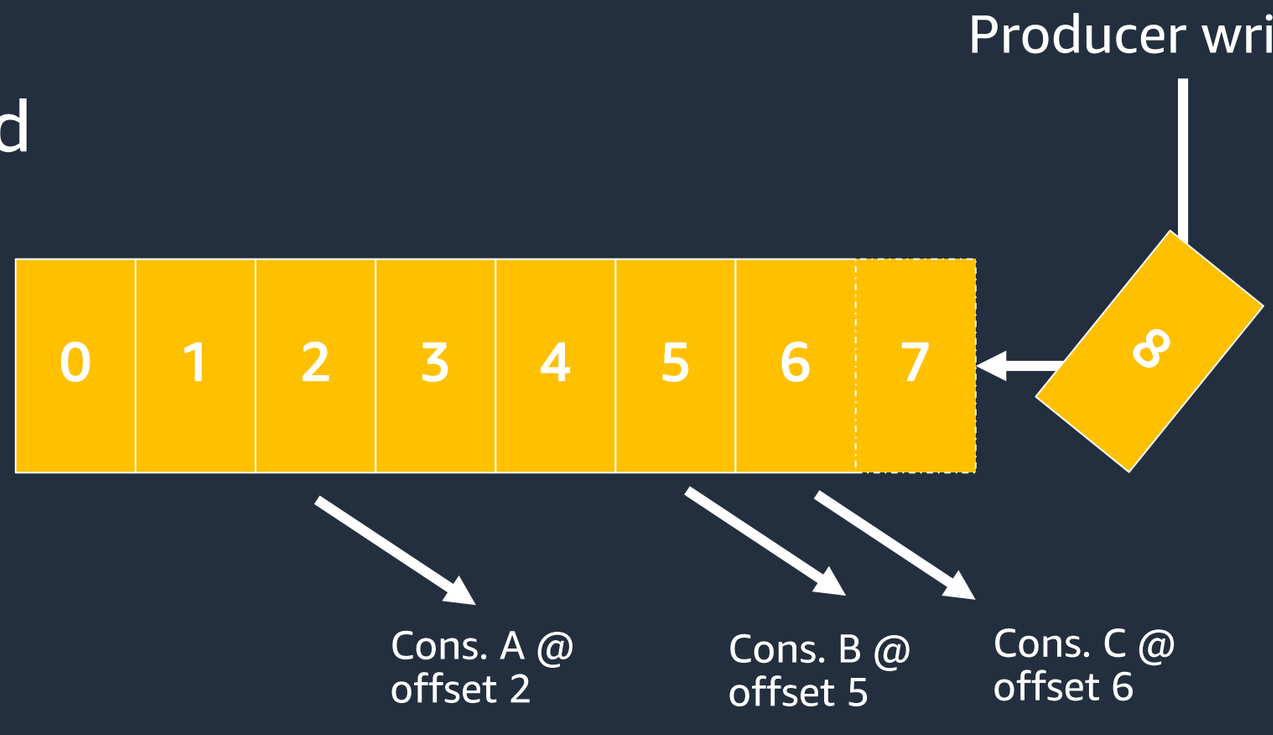
aws

The commit log is the construct for real-time data

Producers and consumers are decoupled

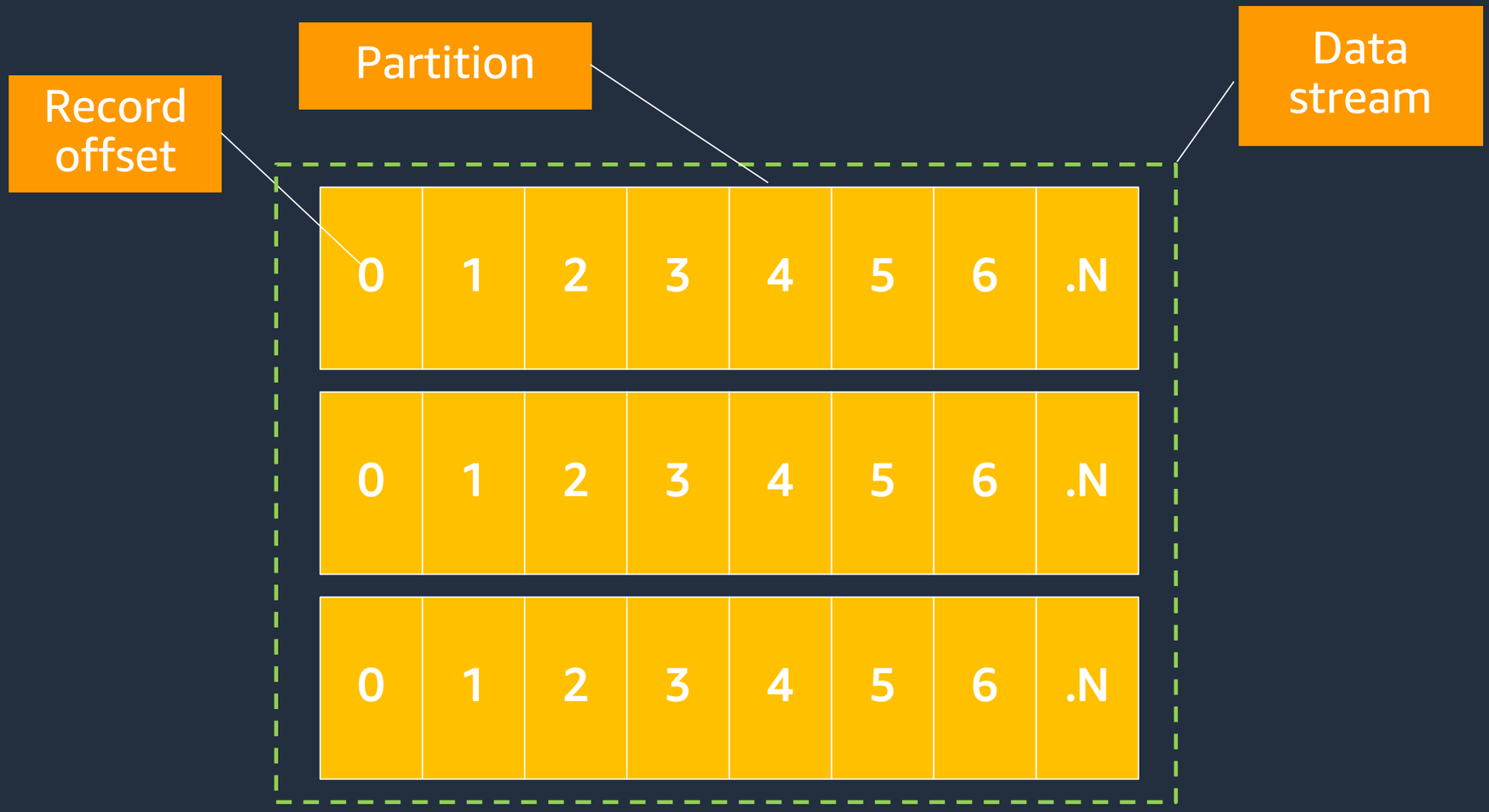
Data is retained, not destroyed

Write order is preserved



aws

What is a data stream?



A data stream is a logical grouping of at least 1 commit log (aka a “partition” or a “shard”)

aws

Current trends

Data streams are replacing message queues

Data streams are replacing Hadoop workflows

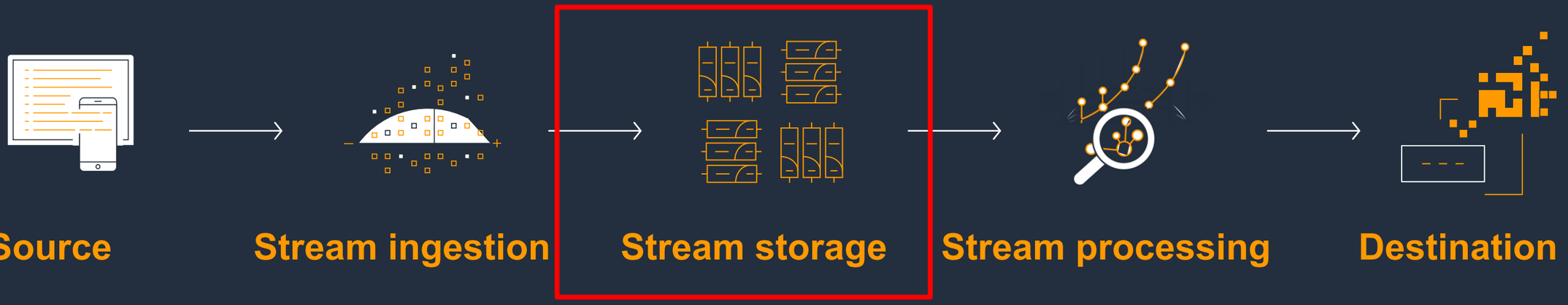
Data streams are the spinal cord for microservices

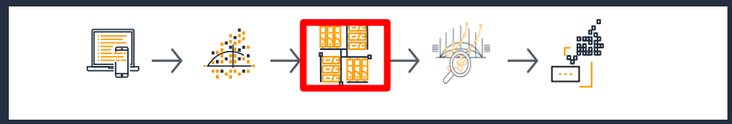
Relational database analytics are happening on change streams

aws

Enabling real-time analytics

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





Stream storage

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

Amazon Kinesis Data Streams

Amazon Managed Streaming for Apache Kafka

Amazon Kinesis Data Firehose

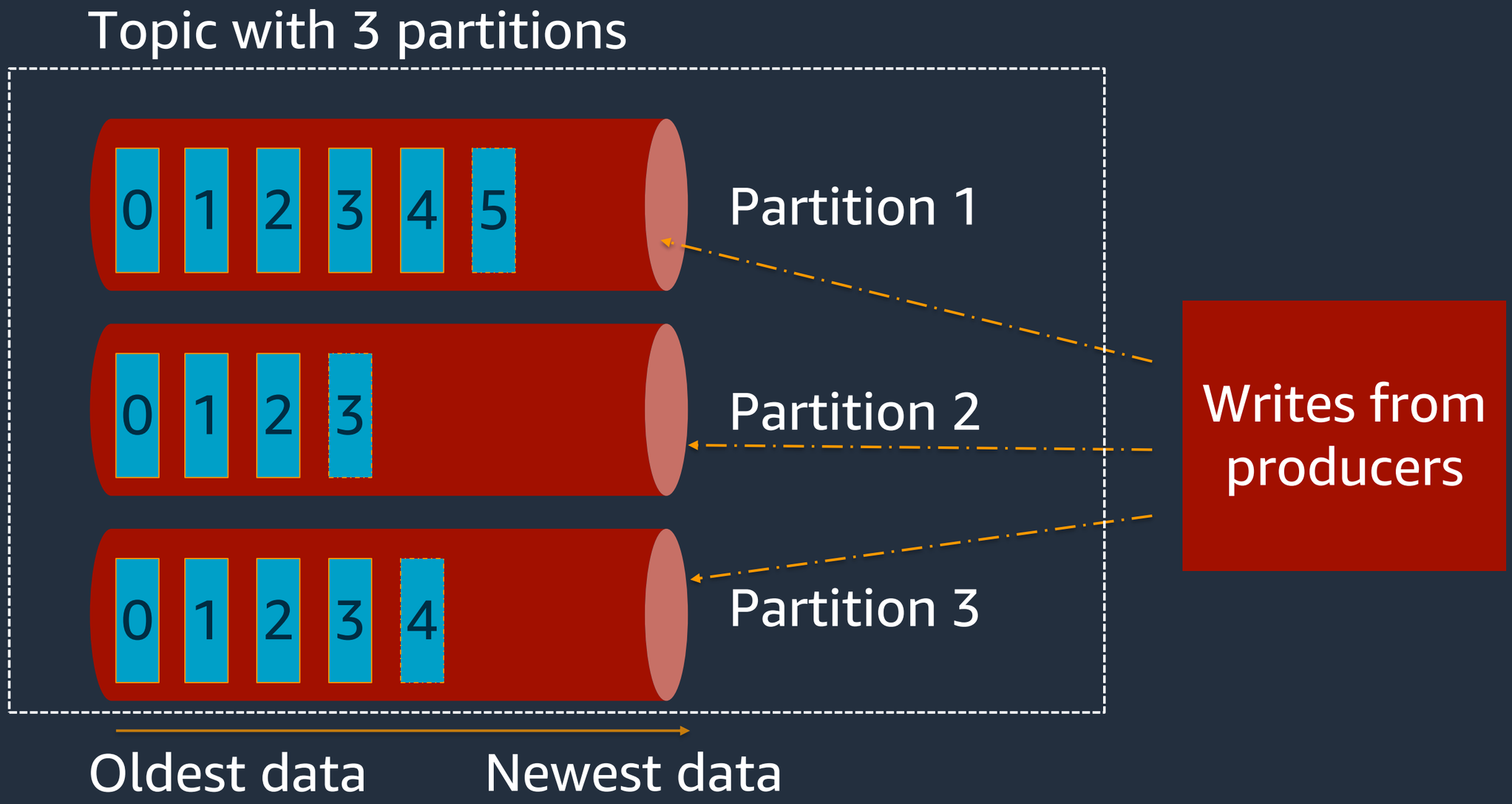
aws

Enabling real-time analytics

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

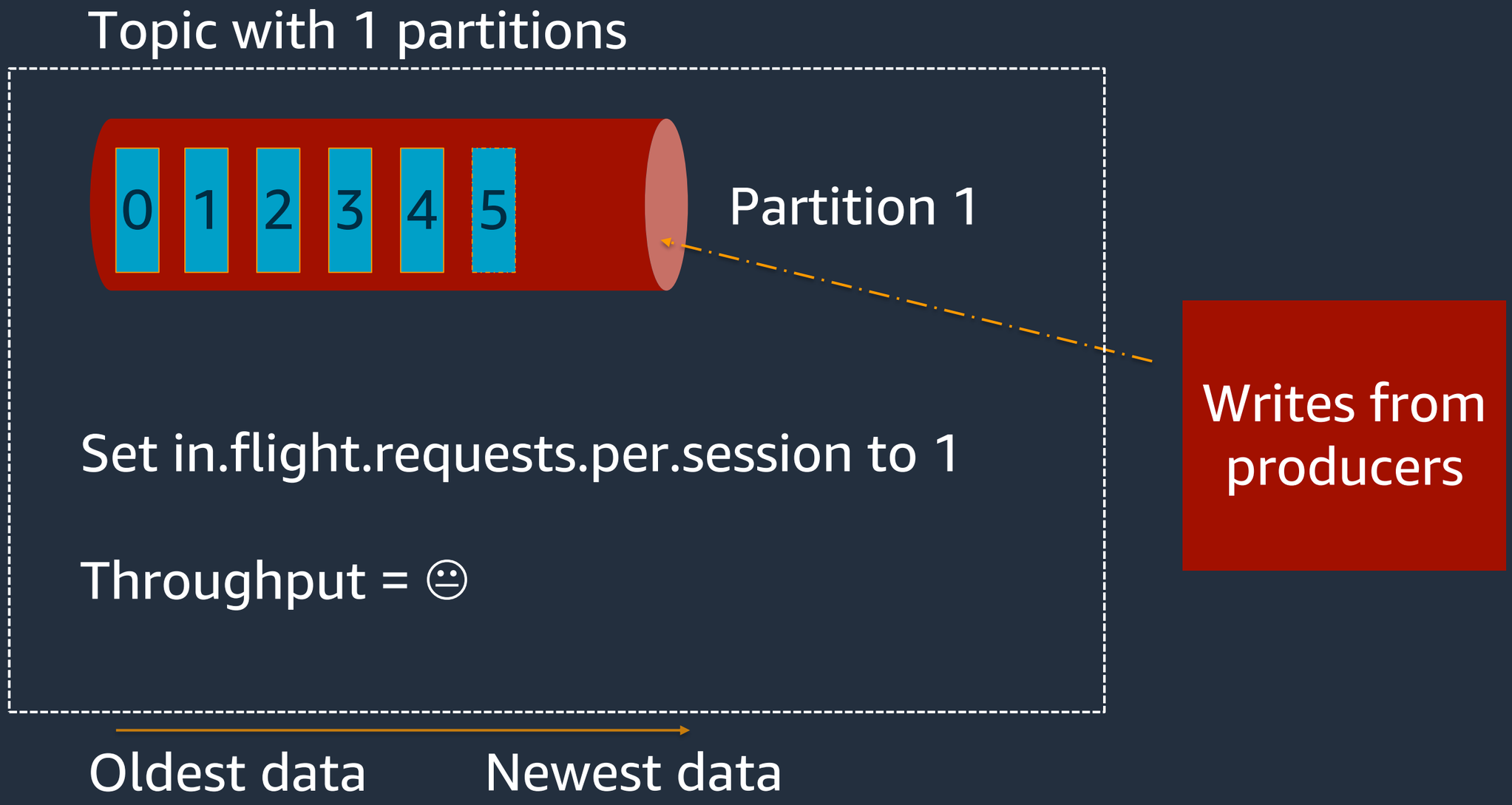
aws

Apache Kafka Ordering & Partitions



aws

Apache Kafka In-Order Delivery



aws