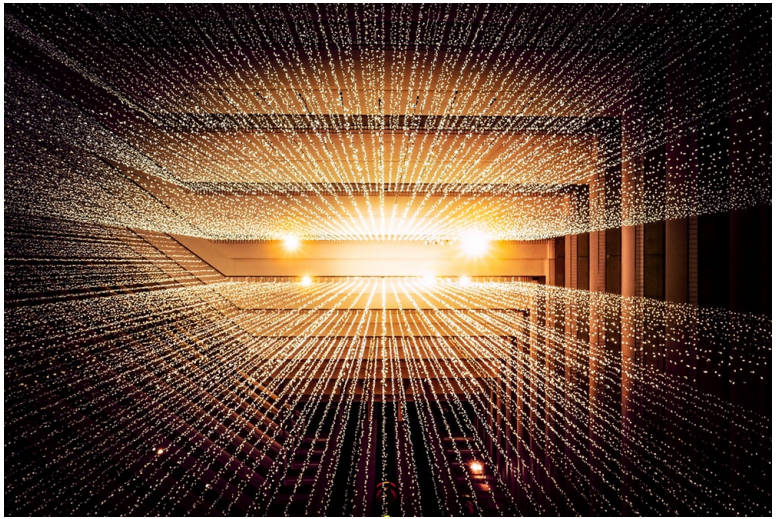
Streaming data pipelines for real-time analytics – Are you ready?

Low latency and low cost at any scale

Sai Maddali, Senior Product Manager, Amazon Kinesis

Table of contents

- Why data streaming?
- 5 Kinesis super powers
- Use cases to get started
- Questions

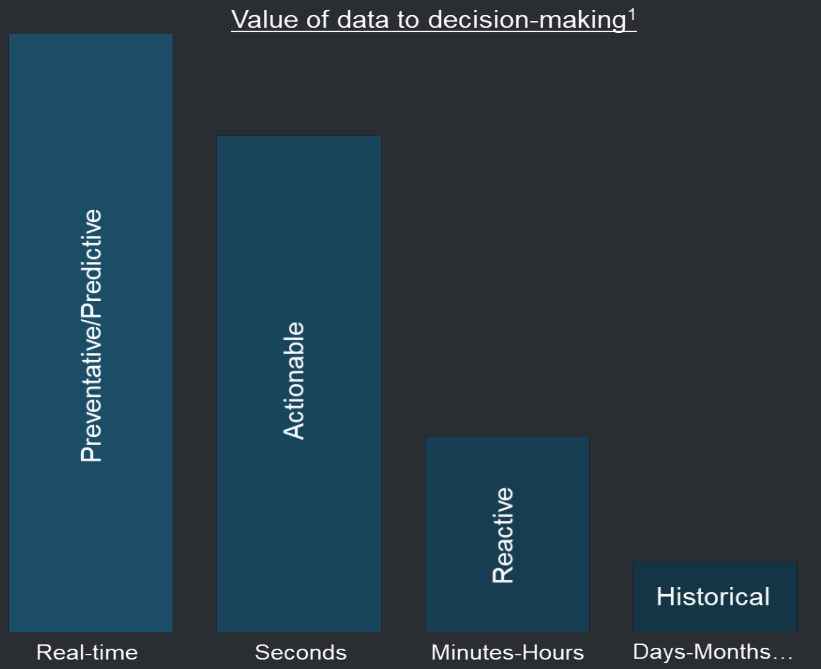


aws

Why data streaming?

Timely decisions require new data in minutes

Data have a short shelf life of actionability[1]. AWS lets you act on that data as fast as the market dictates.



58%

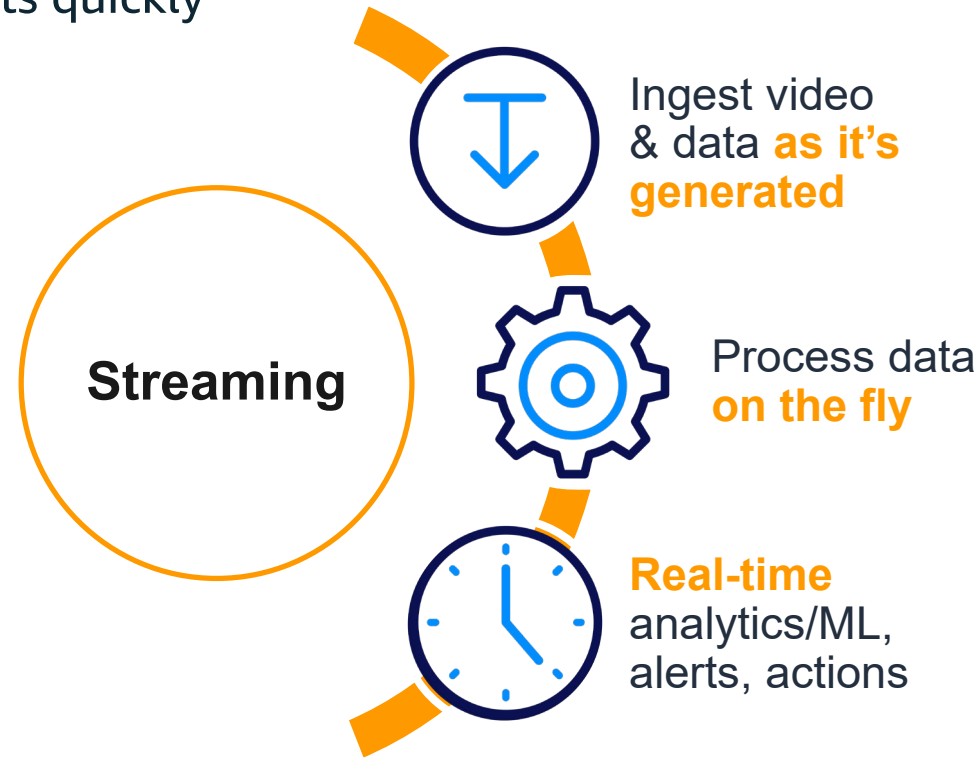
Percentage of those surveyed who say their organization saw **significant increases in customer retention and loyalty** as a result of using real-time customer analytics[2]

75%

Percentage of enterprise organizations surveyed that believe **untimely data inhibited business opportunities**[3]

Stream new data in seconds

Get actionable insights quickly



aws

Epic Games continually improves Fortnite for 250+ million players globally

Challenge:

They needed a way to process and analyze over **100PB** of data (**125M events/min**) ingested from game clients and game servers to understand and adapt to player engagement.

Solution:

Epic Games turned to AWS for an Amazon S3 data lake in combination with Amazon EMR, Amazon EC2, and Amazon Kinesis.

Result:

The data provides a constant feedback loop for designers, and an up to the minute analysis of gamer satisfaction to drive gamer engagement.

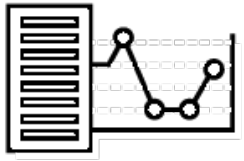
Amazon S3 Amazon EMR Amazon Kinesis Amazon EC2

aws

Most common uses of streaming

Security Monitoring

Industrial Automation



Log Analytics

Data Lakes

Microservices communication

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 it 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)

Analytics services

Database (least common)

aws

Amazon Kinesis: Real-time streaming on AWS

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

Amazon Kinesis Data Streams

Collect and store data streams for analytics

Amazon Kinesis Data Firehose

Load data streams into data stores

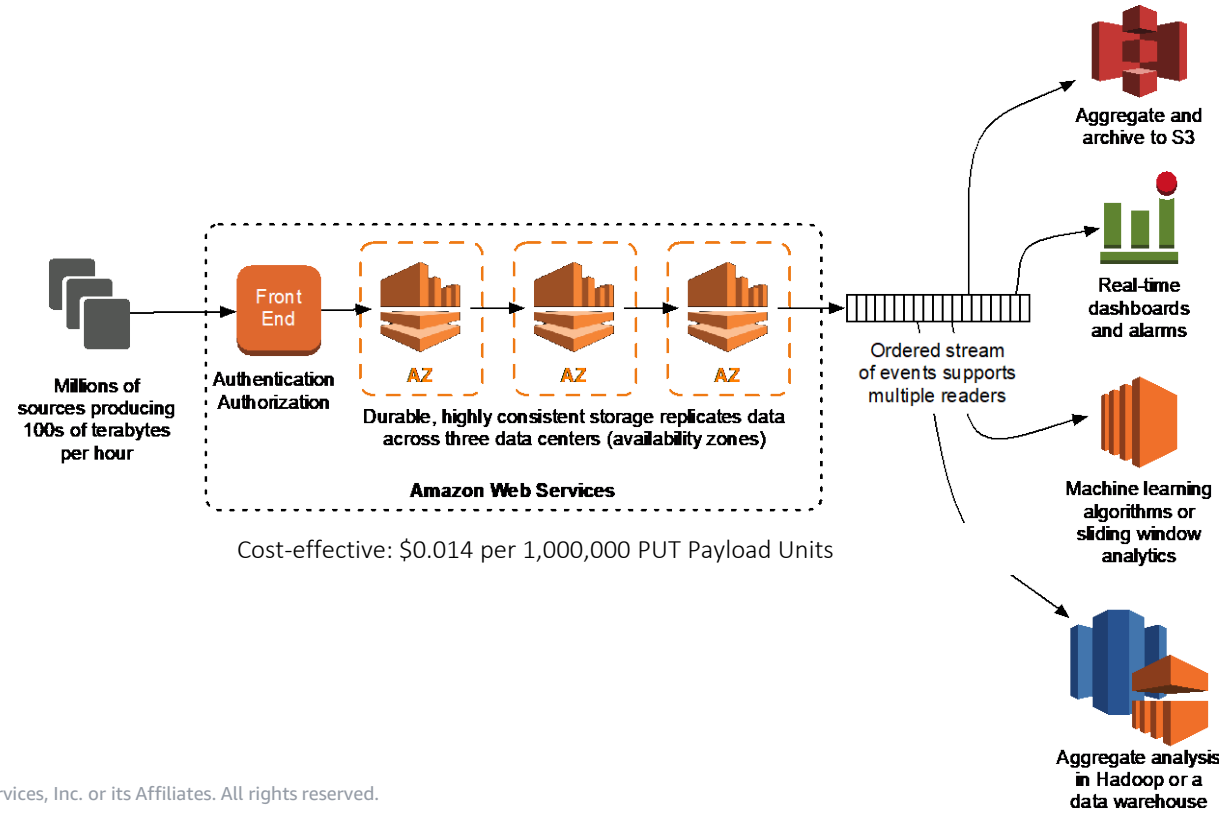
Amazon Kinesis Data Analytics

Analyze data streams with SQL or Java (Apache Flink)

aws

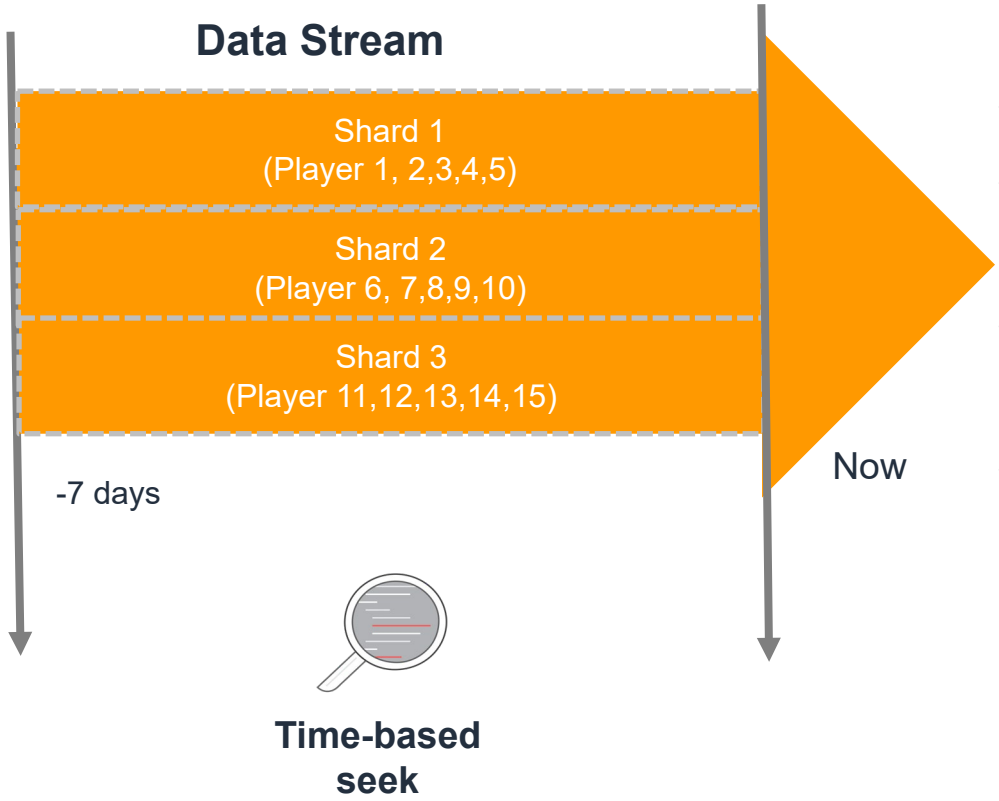
Amazon Kinesis Data Streams – How it works

Fully managed service for real-time processing of streaming data



Cost-effective: $0.014 per 1,000,000 PUT Payload Units

Managed ability to capture and store data



- Data streams are made of **Shards**
- **Shard** – Unit of throughput and parallelism
- **Partition key -** Business key that enables automatic data mapping into shards.
- **Iterator -** Able to seek at any point in the stream to read data

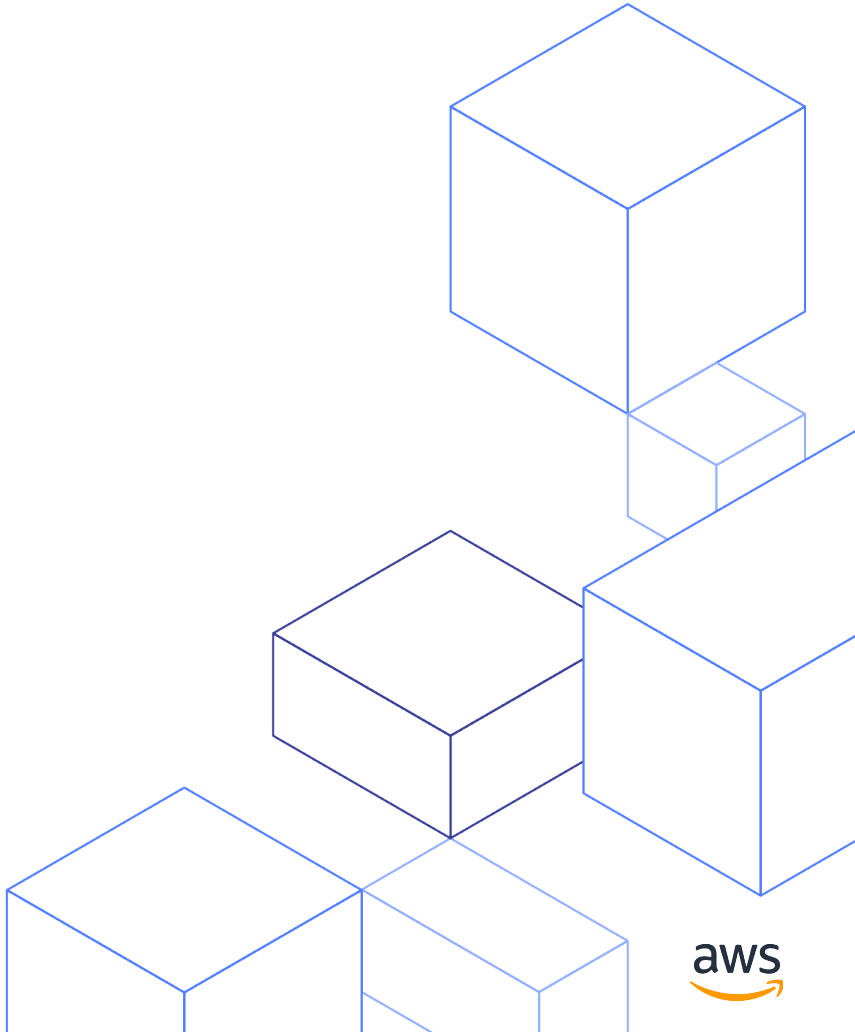
aws

5 super powers of Amazon Kinesis Data Streams

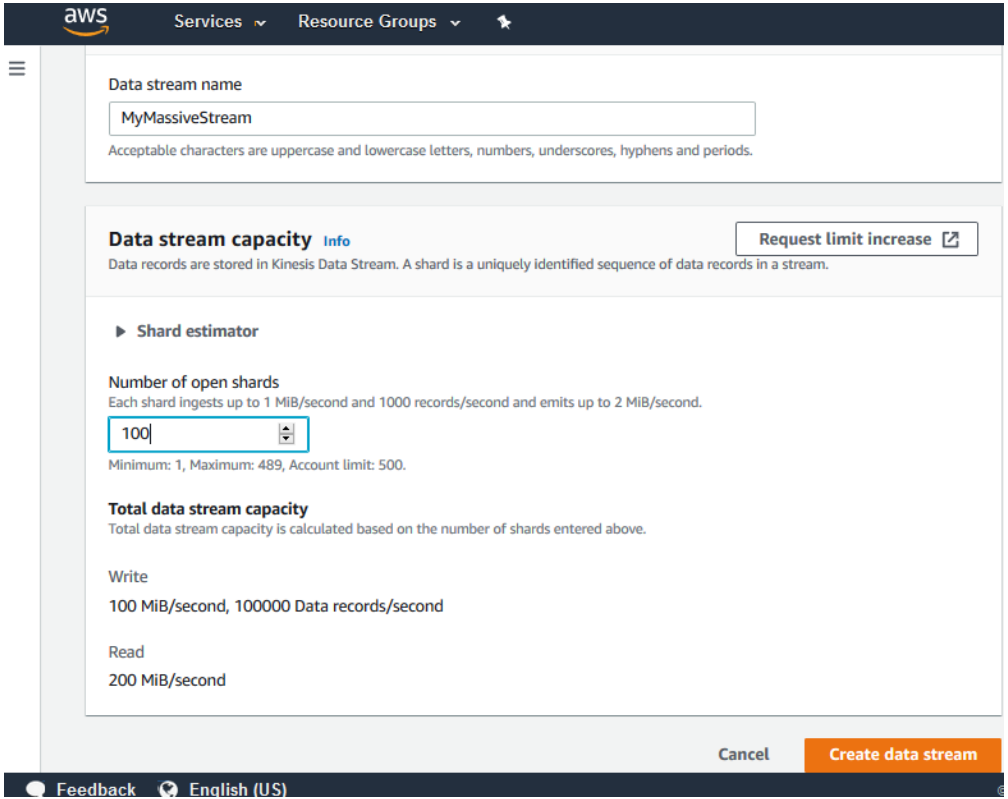
1. Easy to get started
2. Easy to operate
3. Massive scale
4. Low-latency
5. Low cost

aws

Easy to get started



Get started in minutes with a few clicks



- Guaranteed throughput, making it easy to size the workload
- Create a data stream with the necessary throughput in minutes

aws

Integrate with existing systems to ingest data

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

AWS toolkits/libraries

- AWS SDK
- Amazon Kinesis Producer Library
- AWS Mobile SDK — iOS, Android, Kindle Fire
- Amazon Kinesis Agent

AWS service integrations

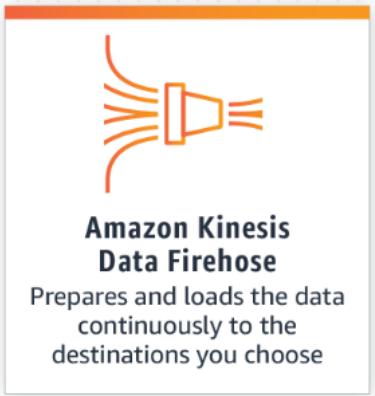
- AWS IoT Core
- Amazon CloudWatch Logs
- Amazon CloudWatch Events
- AWS Database Migration Service (AWS DMS)*

Third-party offerings

- Log4j
- Flume
- Fluentd

*AWS DMS includes eight on-premises databases, one Azure database, five Amazon RDS/Amazon Aurora database types, and Amazon Simple Storage Service (Amazon S3)

Deliver data to different destinations in few clicks



Amazon Kinesis Data Firehose

Deliver data to destinations such as Amazon S3, Amazon Redshift, Amazon Elasticsearch Service, and generic **HTTP endpoints** so you can use existing analytics tools

aws

Write data processing applications quickly

Use AWS Lambda to quickly process streaming data

Get actionable insights from streaming data in real time

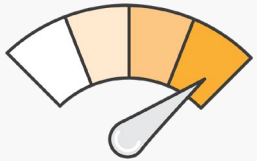
aws

Kinesis Data Analytics for Java for sophisticated applications

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

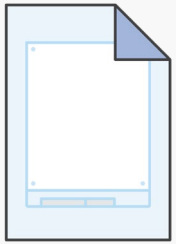
Simple programming

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

aws

Native AWS Integrations

AWS Identity and Access Management

AWS Key Management Service

Amazon VPC

AWS PrivateLink

Amazon API Gateway

Amazon CloudWatch

Amazon EventBridge

Amazon Pinpoint

Amazon Redshift

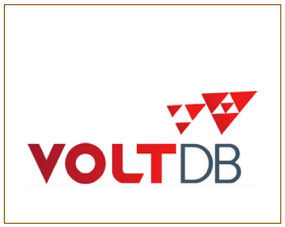
Amazon Quantum Ledger Database

AWS Database Migration Service

Amazon Elasticsearch Service

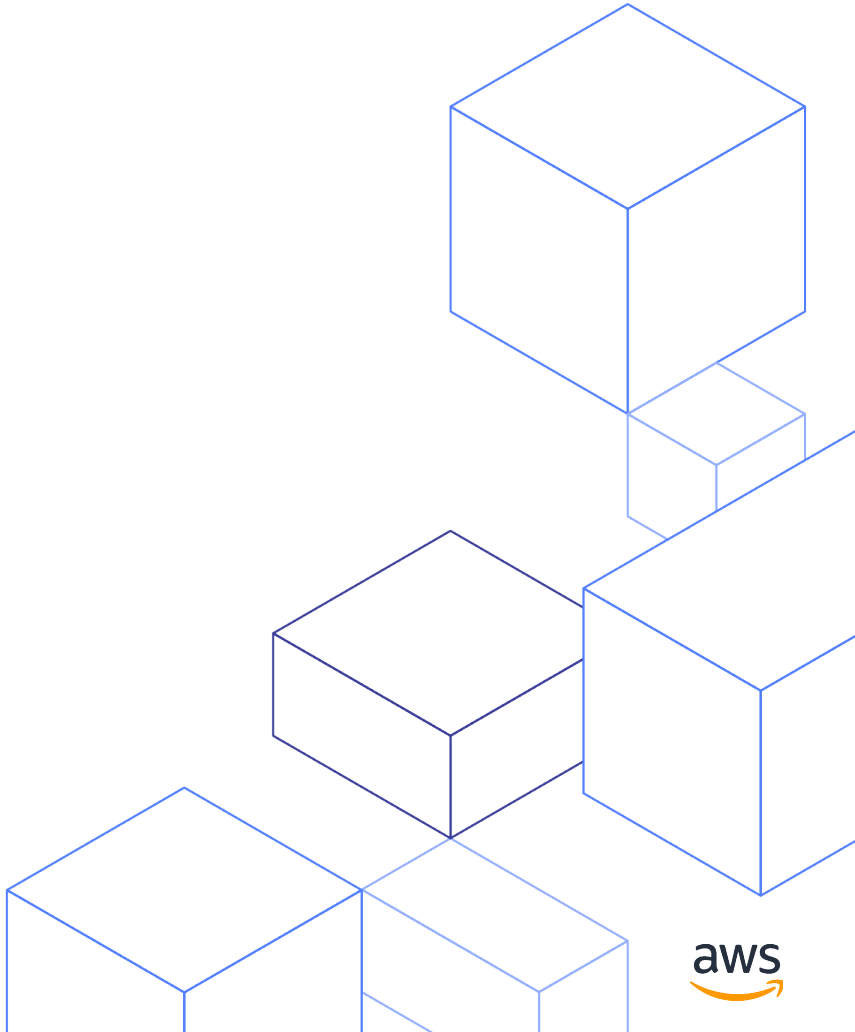
aws

3rd Party Connectors

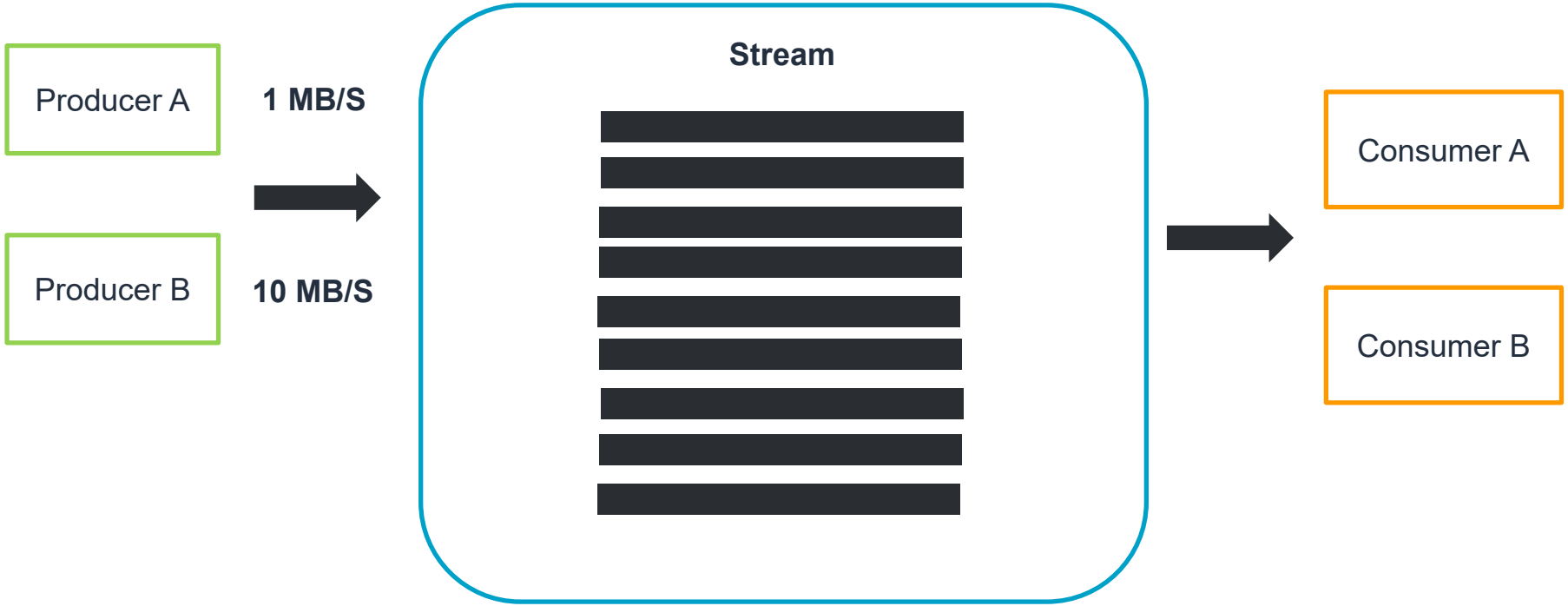


aws

Easy to operate

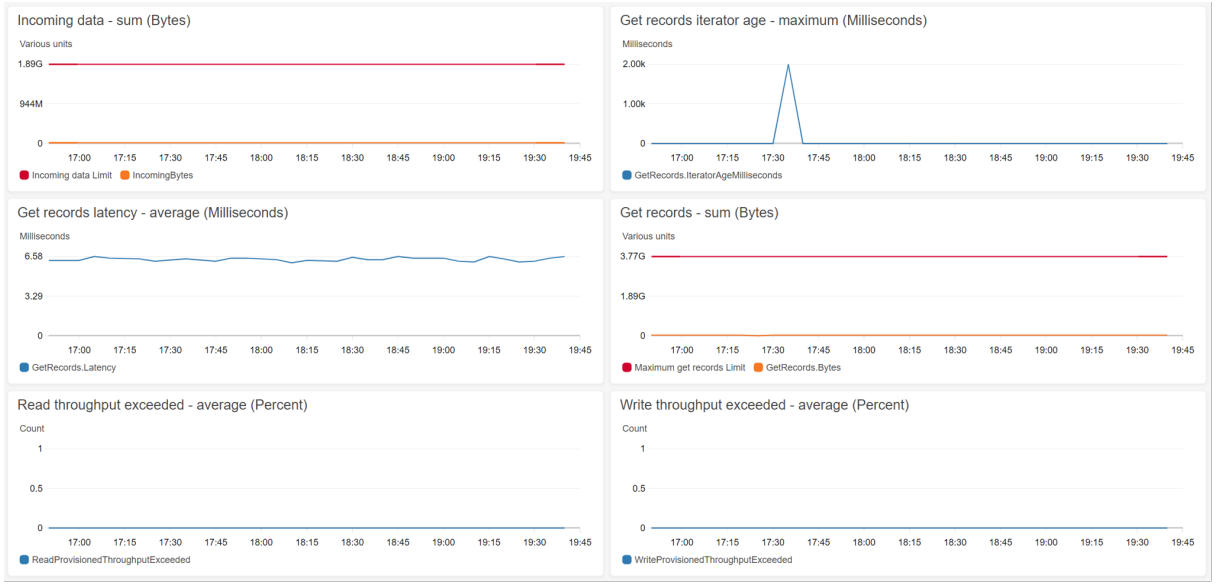


Seamless and non-disruptive scaling

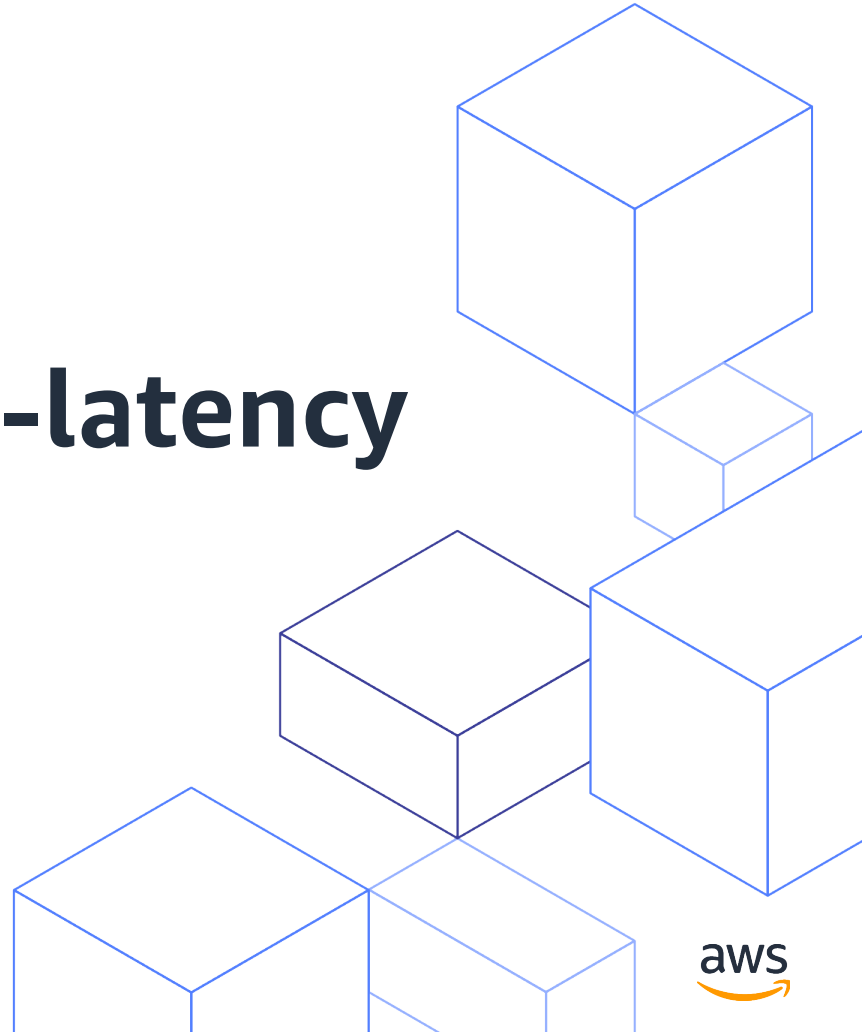


Low operations overhead to manage large streams

- Quickly identify and remediate issues using integration with Amazon CloudWatch
- Automate capacity management using CloudWatch and AWS Auto-scaling

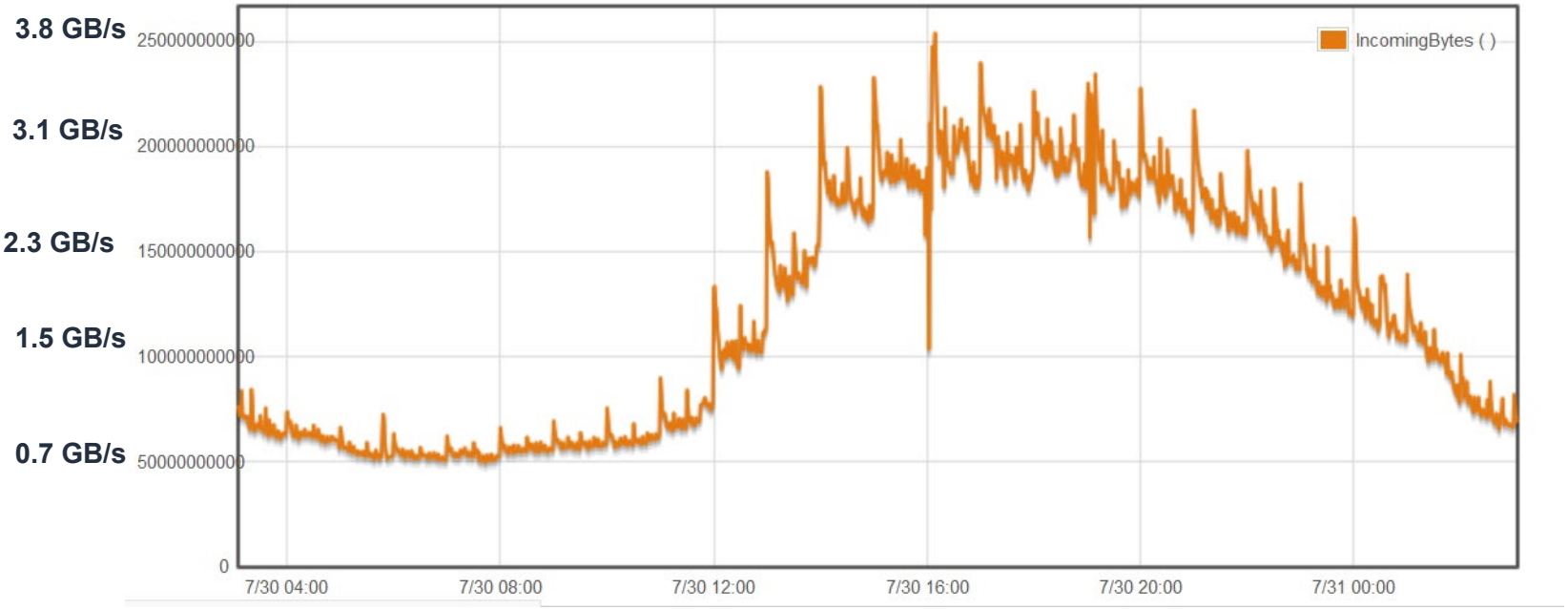


Massive scale & low-latency



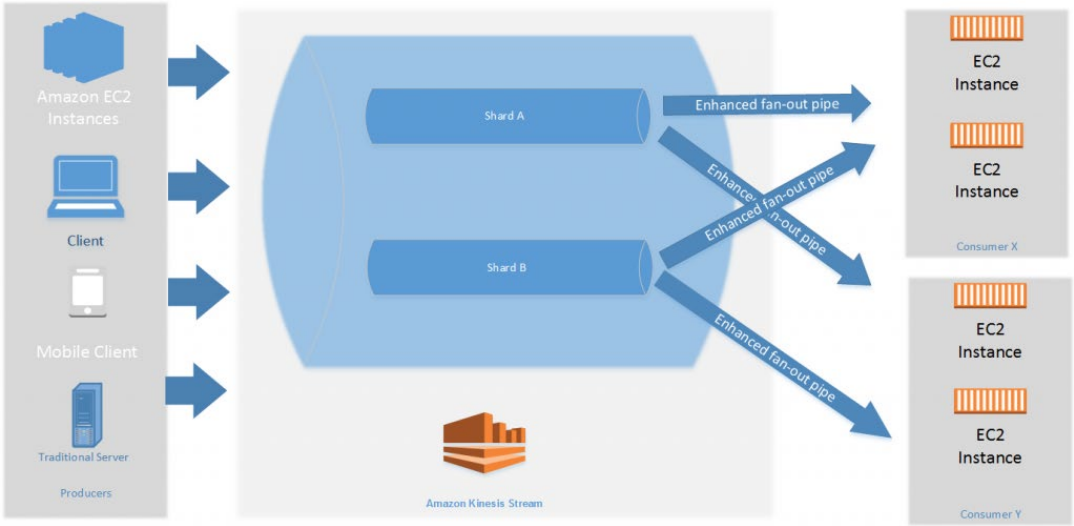
Supports massive scale

You can grow your data stream to support any throughput



aws

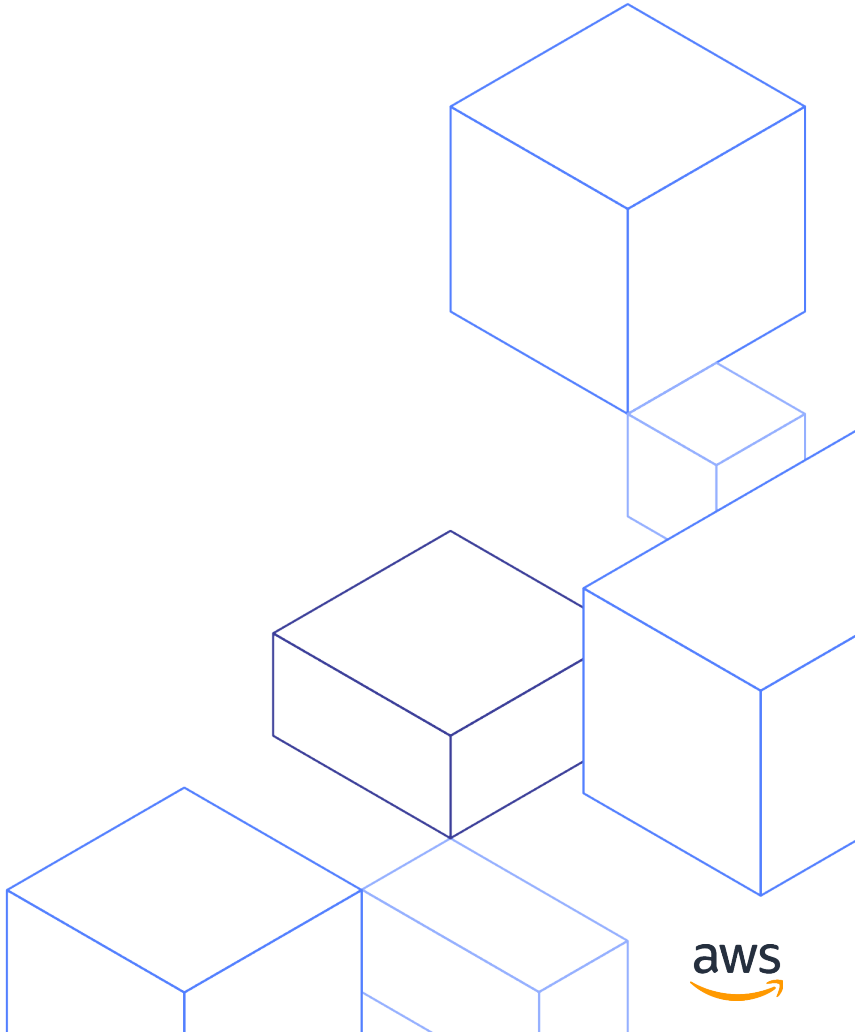
Low-latency and high fan out



- Add 20 consumers
- HTTP/2 to allow <100 ms delivery
- Enhanced Fan Out allows multiple consumers, each at 2MB/second, independently

aws

Low cost

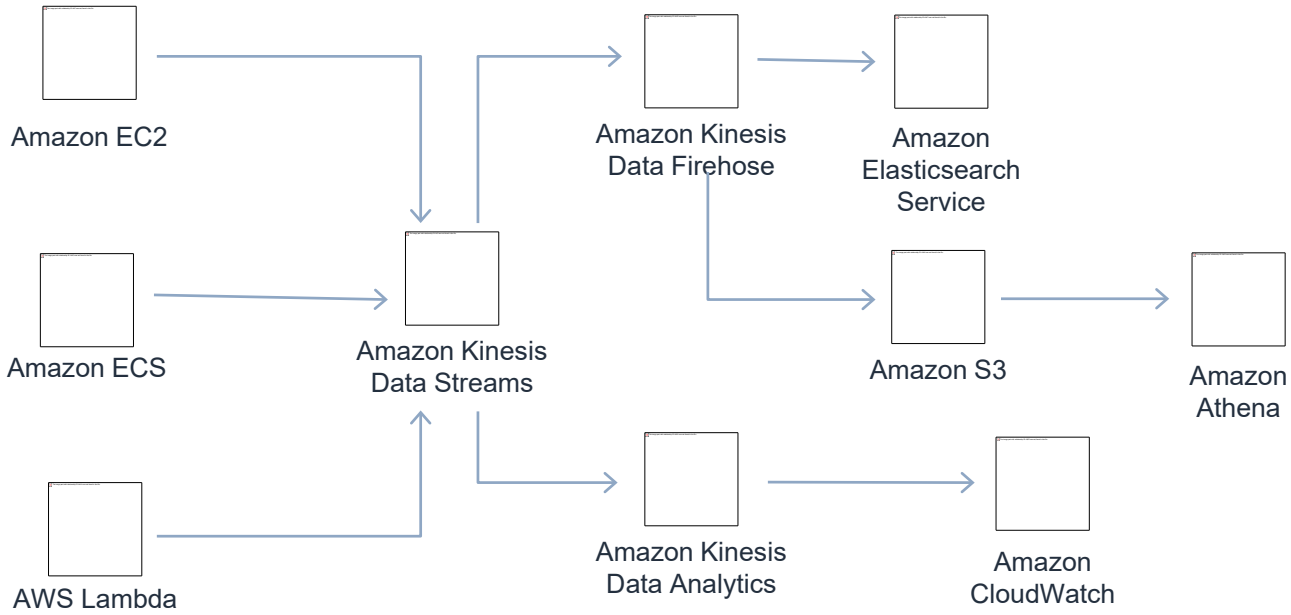


Cost-Effective

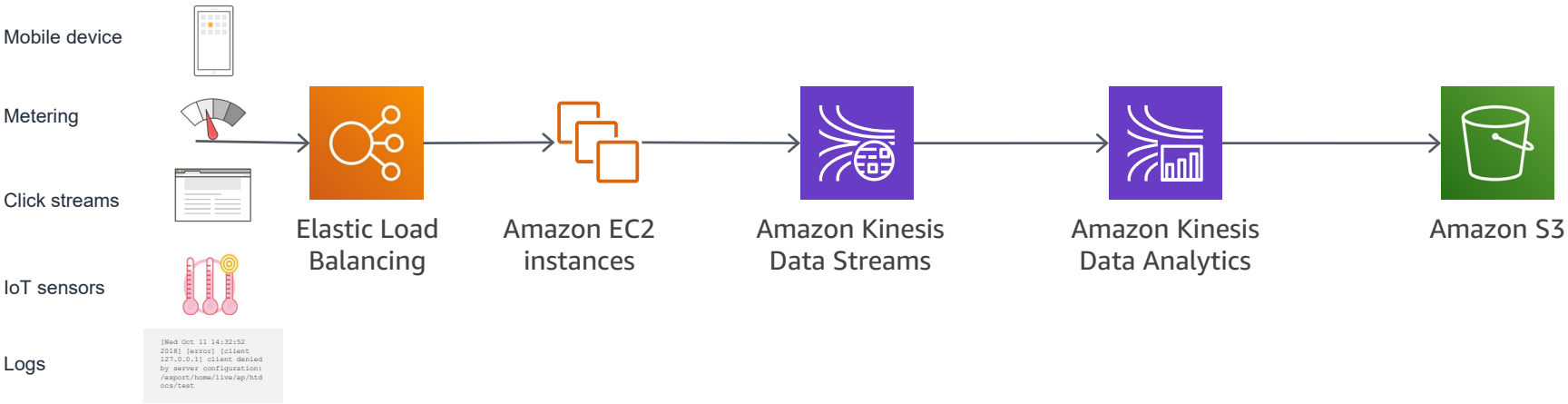
- Pay-as-you-go pricing
- No upfront cost and no minimum fees
- Based on two dimensions:
  - Shard-Hour: $0.015
  - PUT Payload Units (25K), per million units: $0.014
- Granular scaling that enables you to balance capacity and costs

Typical data streaming use cases

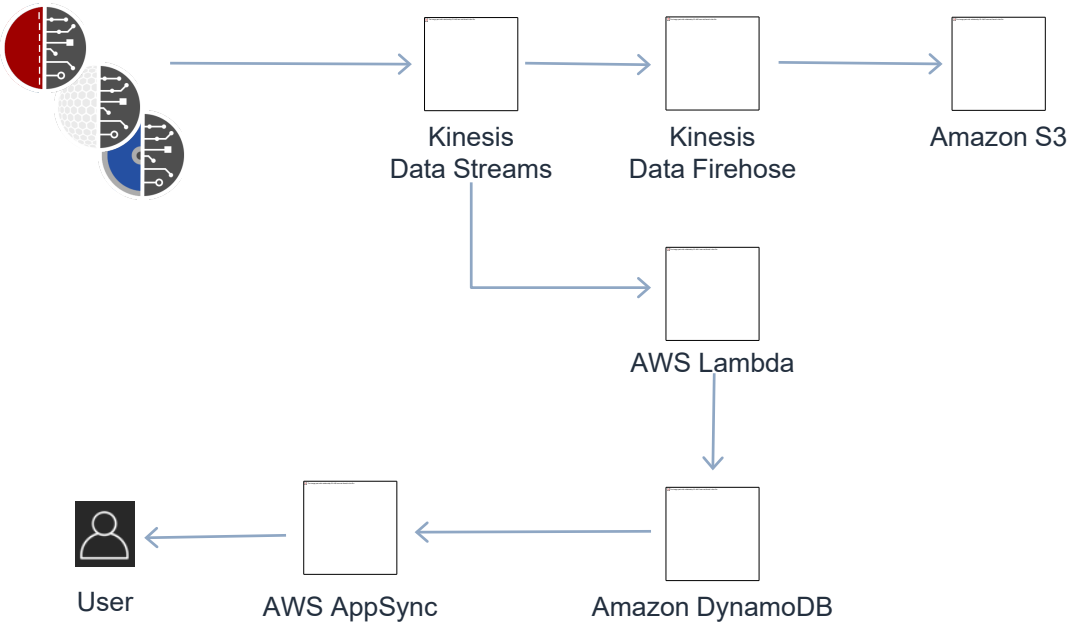
Log ingestion to process Terabytes of data in real time



Streaming ETL to your data lake



IoT sensor data collected, ingested, and analyzed



Recap

- Data streaming opens up possibilities of speed
- Amazon Kinesis makes it easy to build and scale streaming applications at low-cost
- Leverage solution guides and Data Labs to get started

Next steps

Learn more about Amazon Kinesis:
aws.amazon.com/kinesis

Get started with Amazon Kinesis:
aws.amazon.com/kinesis/getting-started