New Innovations from Amazon Kinesis for Real-Time Analytics

Allan MacInnis, Solutions Architect

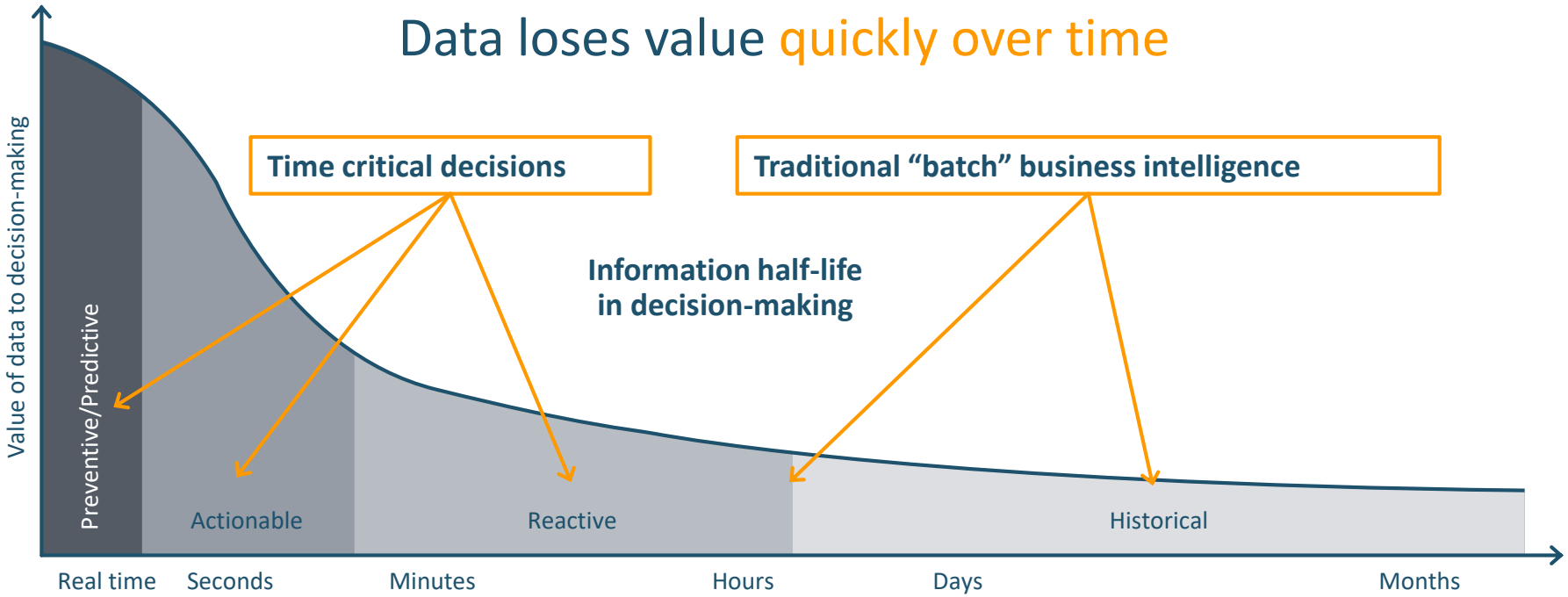
September 20, 2018

aws

What to expect

- Streaming data overview
- Brief Kinesis services overview
- New features review:
    - Kinesis Data Streams: Enhanced fan-out
    - Kinesis Data Firehose: Delivery to Splunk
    - Kinesis Data Firehose: Data format conversion

Timely Decisions Require New Data in Minutes



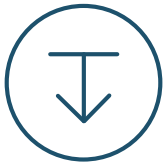
Source: Perishable insights, Mike Gualtieri, Forrester

aws

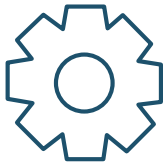
Stream New Data in Seconds

Get actionable insights quickly

Streaming



Ingest data as it's generated



Process data on the fly

Real-time analytics/ML, alerts, actions

aws

Most Common Uses of Streaming

Smart Home
Smart City

Industrial
Automation

Log
Analytics

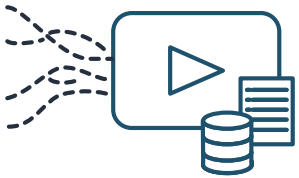
Data
Lakes

IoT
Analytics

aws

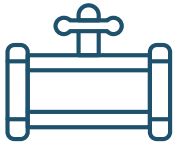
Streaming with Amazon Kinesis

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



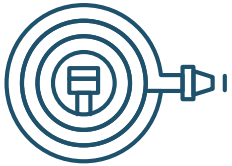
Kinesis Video Streams

Capture, process, and store video streams



Kinesis Data Streams

Capture, process, and store data streams



Kinesis Data Firehose

Load data streams into data stores

Kinesis Data Analytics

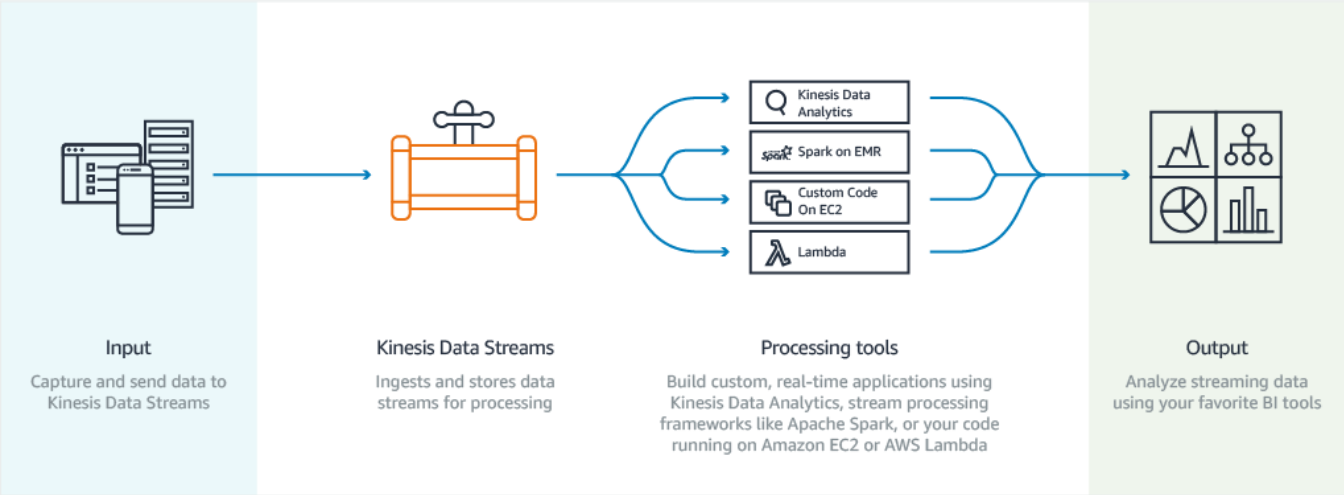
Analyze data streams with SQL

aws

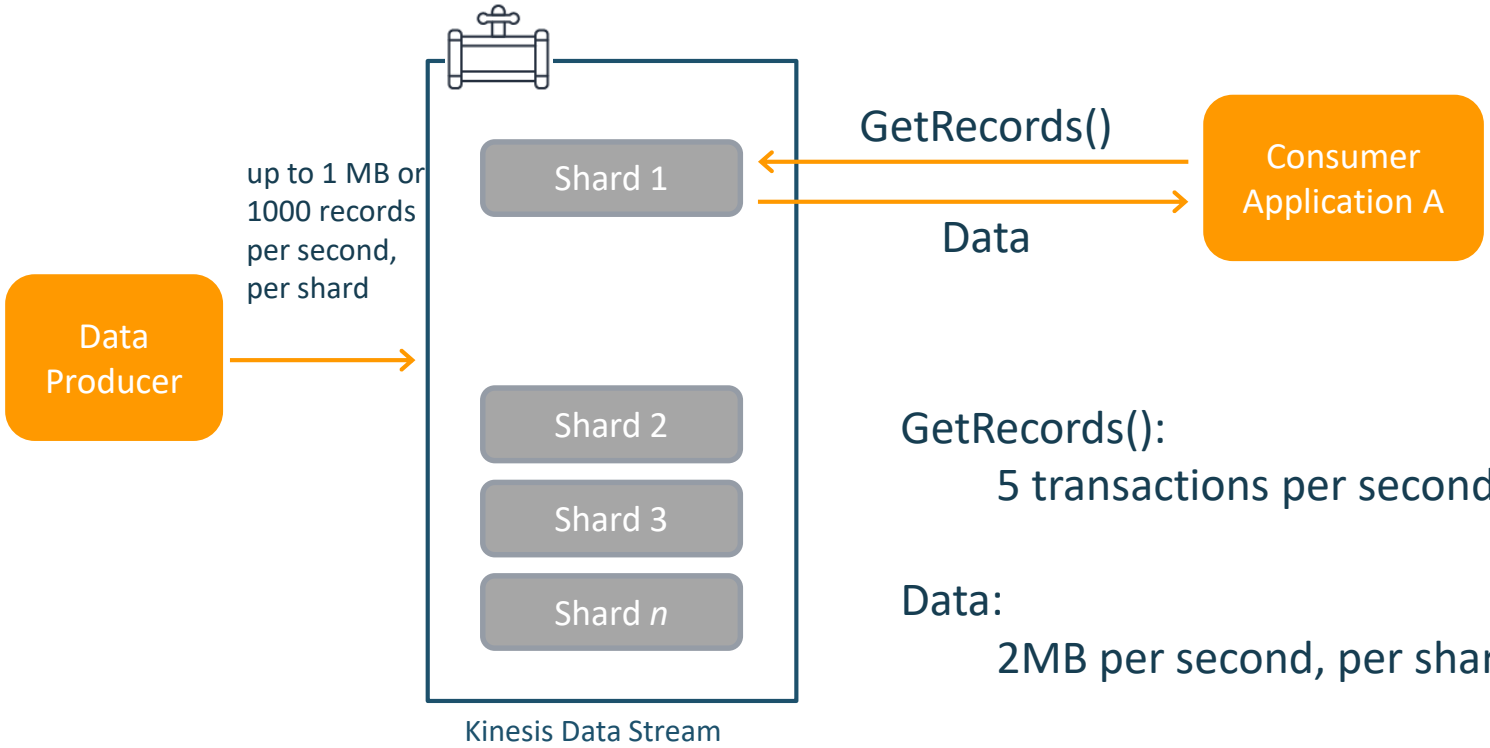
Kinesis Data Streams

Enhanced fan-out

Amazon Kinesis Data Streams



Kinesis Data Streams: Standard egress limitations



With only one consumer application, records can be retrieved every 200 ms.

GetRecords():

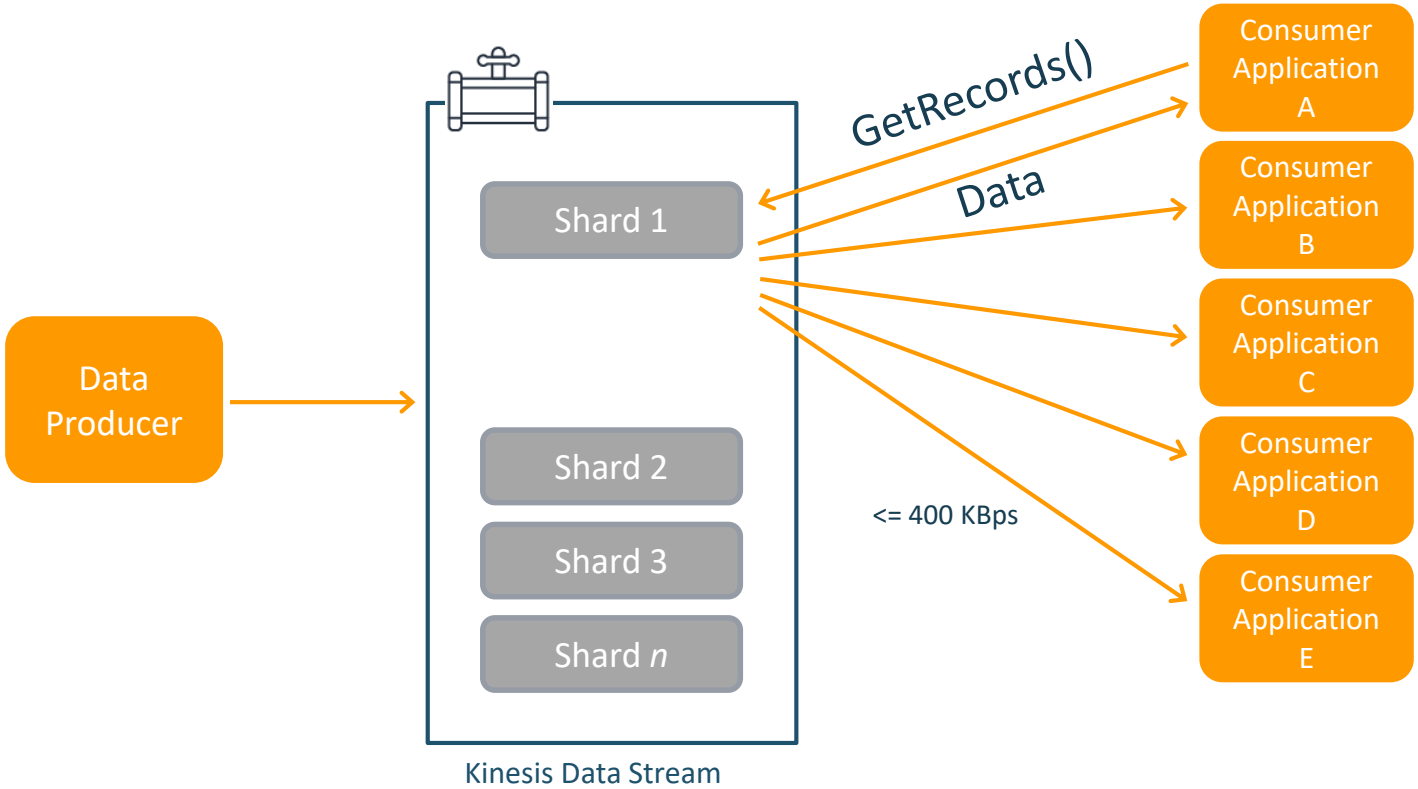
5 transactions per second, per shard

Data:

2MB per second, per shard

aws

Kinesis Data Streams: Standard egress limitations



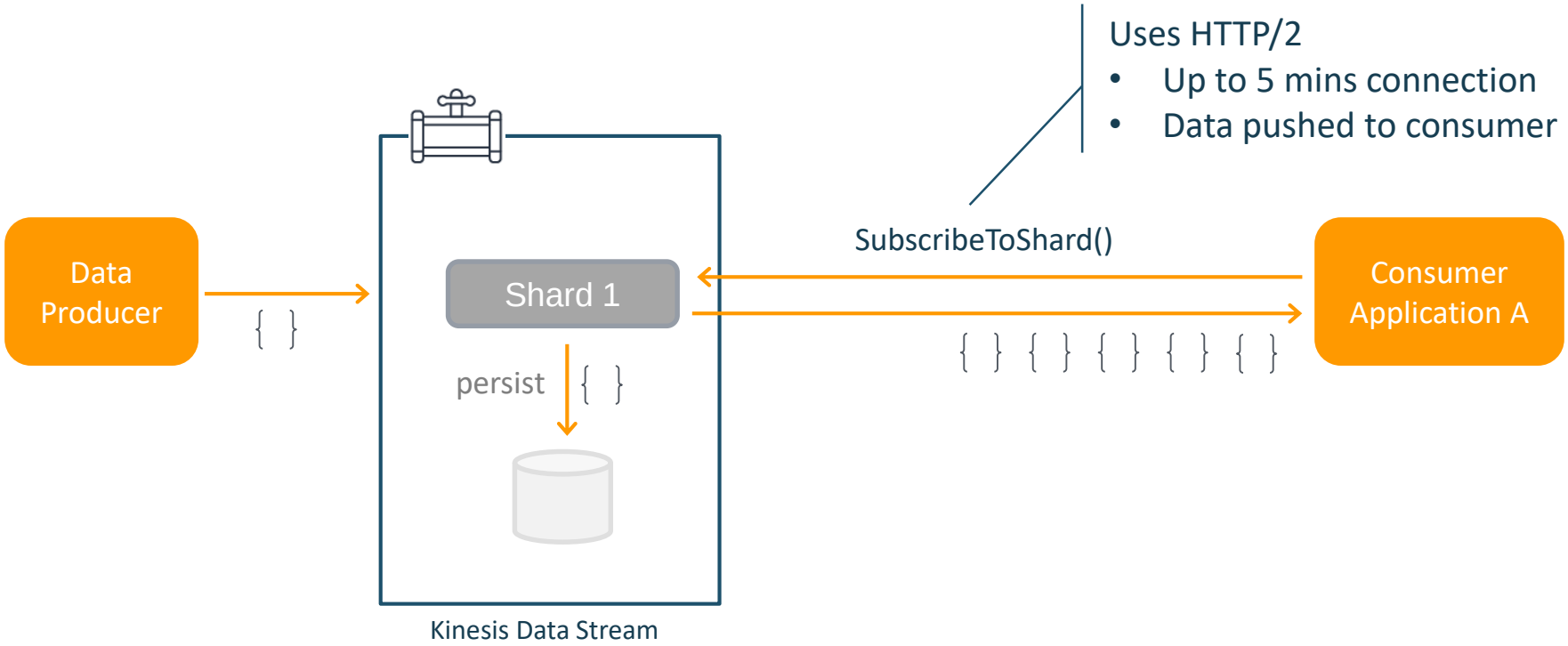
With more consumer applications, propagation delay increases.

For example, with 5 consumer applications, each can only retrieve records once per second, and less than 400 KBps.

aws

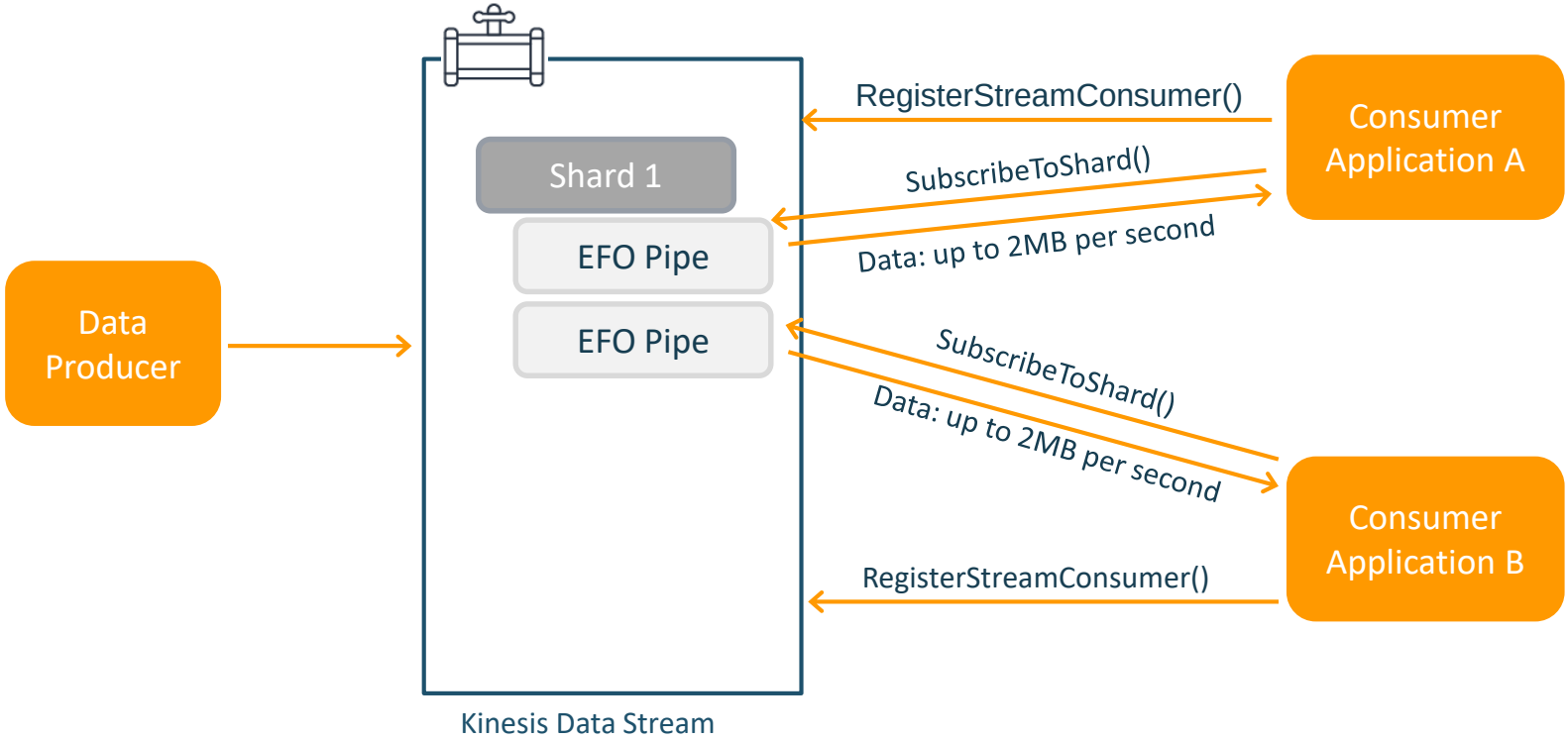
Kinesis Data Streams: Enhanced fan-out

Polling no longer necessary. Messages are pushed to the consumer as they arrive.

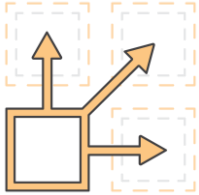


Kinesis Data Streams: Enhanced fan-out

Each consumer application gets dedicated 2MB per second egress, per shard



Kinesis Data Streams: Enhanced fan-out



When to use enhanced fan-out:

- Multiple consumer applications for the same Kinesis Data Stream
    - Default limit of 5 registered consuming applications. More can be supported with a service limit increase request.
- Low-latency requirements for data processing
    - Messages are typically delivered to a consumer in less than 70 ms

Developing EFO Applications:

- Use Kinesis Client Library (KCL) 2.0

Kinesis Data Streams: Enhanced fan-out

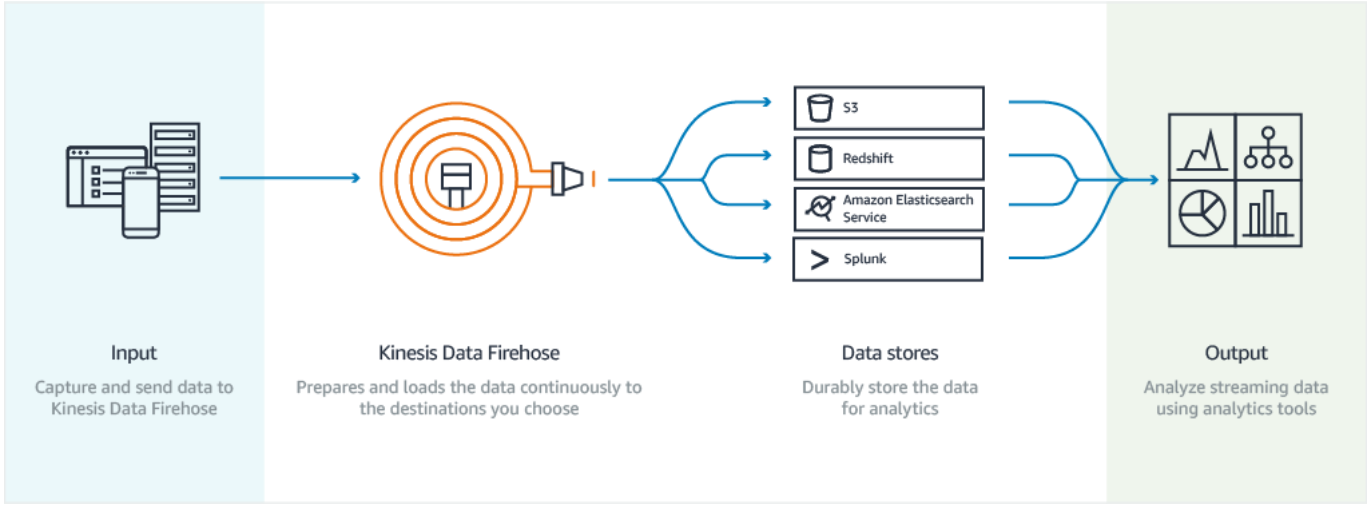
Pricing (US East - Ohio):

- $0.013 per GB of data retrieved
- $0.015 per consumer-shard hour

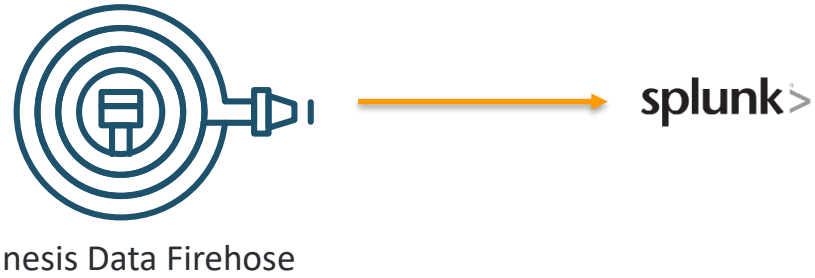
Kinesis Data Firehose

Splunk delivery

Kinesis Data Firehose



Kinesis Data Firehose: Splunk delivery

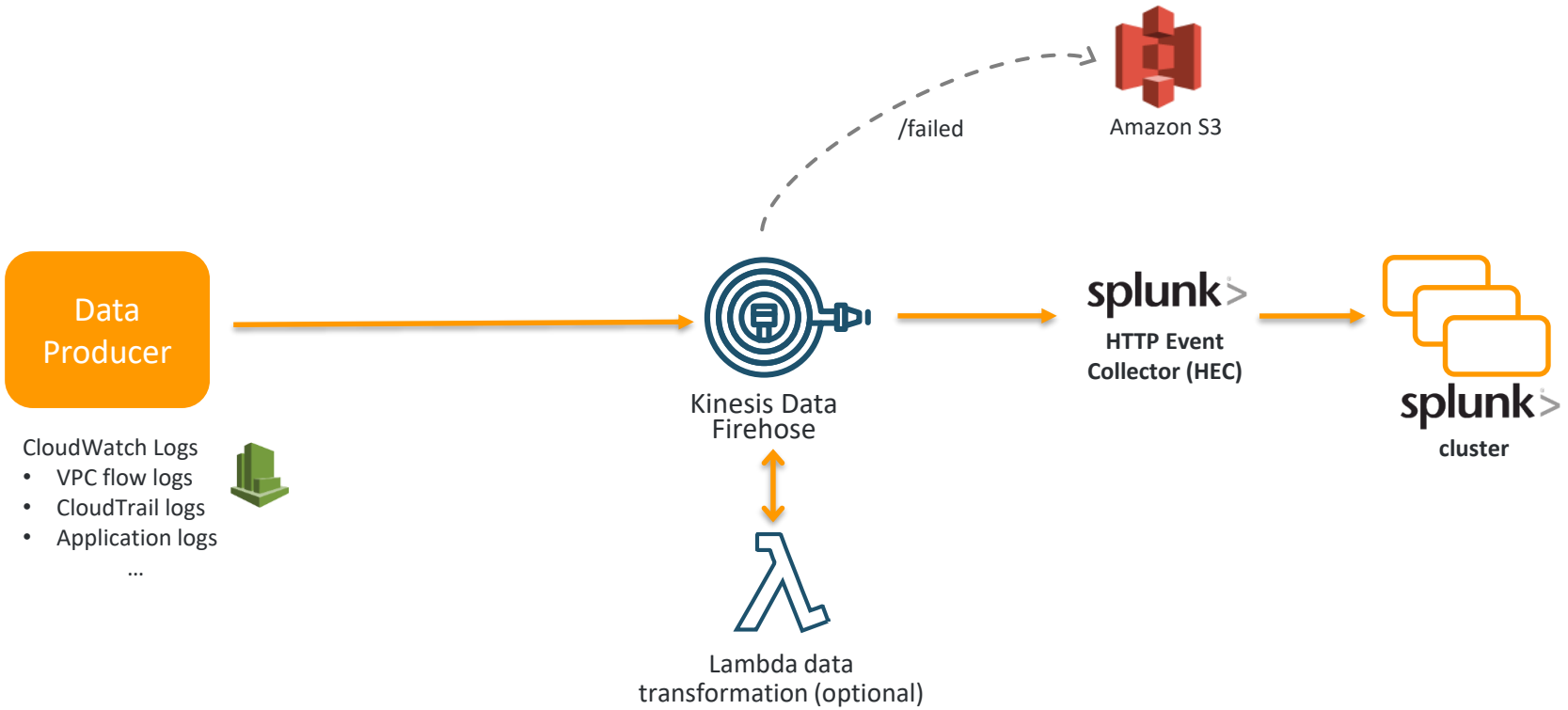


Kinesis Data Firehose

- Deliver to your own Splunk index

aws

Kinesis Data Firehose: Splunk delivery



aws

Kinesis Data Firehose: Splunk delivery

Splunk destination

Firehose accesses Splunk instances through an endpoint and an authentication token. Generate the endpoint and authentication token by enabling the HTTP Event Collector (HEC) on your Splunk instances. Learn more

To grant Firehose access to an op-premises data center and Splunk instance, ensure proper network configurations.

Splunk cluster endpoint* https:// example.splunk.com:8080

Splunk endpoint type

- (•) Raw endpoint
  Capable of parsing most common log formats. View supported log formats.
- ( ) Event endpoint
  Requires specific JSON formatting. Use the Firehose data transformation feature to properly format source data.

Authentication token* *Specify authentication token*

[ ] Show token

HEC acknowledgement timeout* 180 seconds

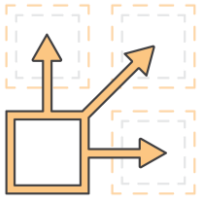
Specify a timeout duration from 180 - 600 seconds

Retry duration* 300 seconds

Specify a retry duration from 0 - 7200 seconds

aws

Kinesis Data Firehose: Splunk delivery



When to use Splunk delivery:

- Delivering to Splunk cluster: either EC2 or Splunk cloud
  - Easily ingest AWS logs from CloudWatch logs (i.e. VPC flow logs, CloudTrail logs) in near real-time for analysis
  - Stream and analyze other logs (application logs, security logs, error logs) in near real-time.

Kinesis Data Firehose

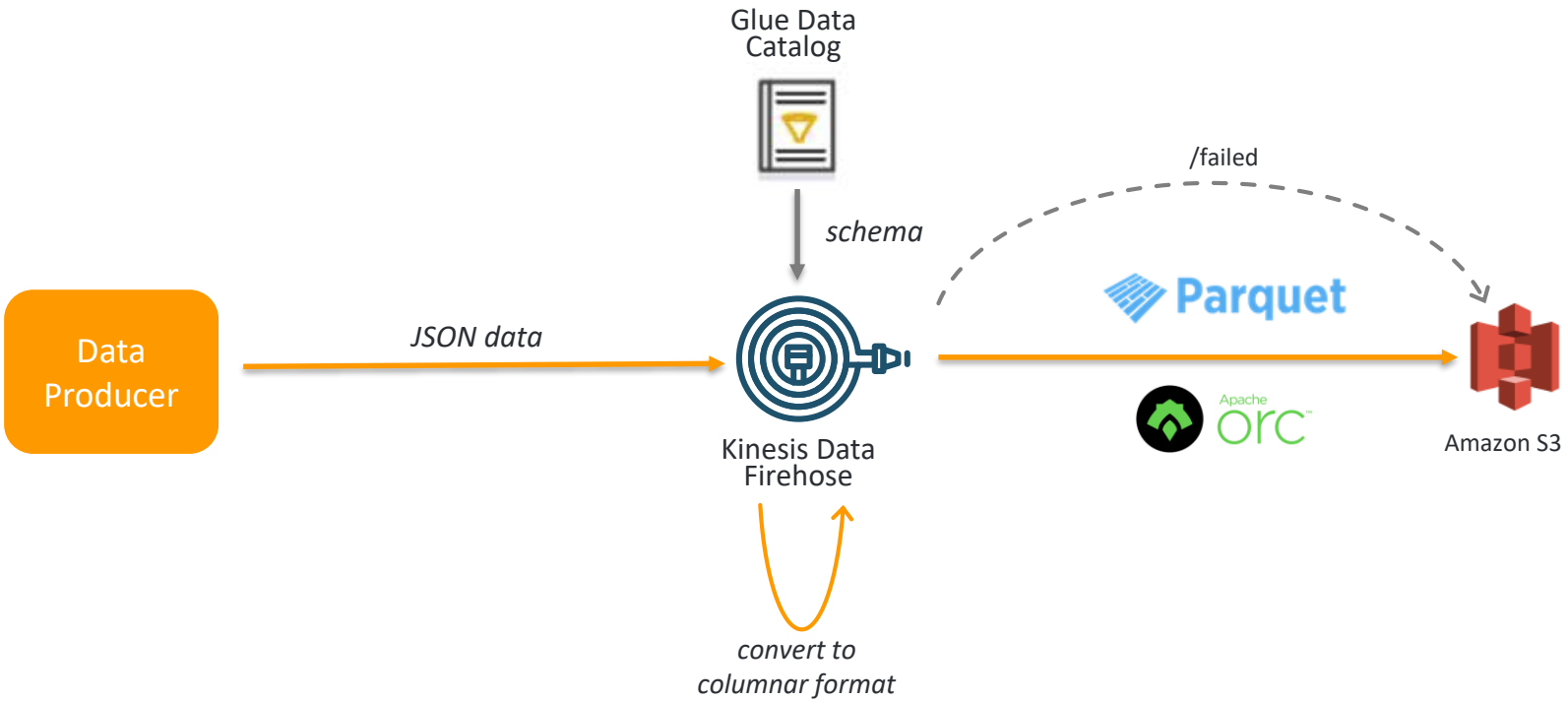
Record format conversion

Kinesis Data Firehose: Record format conversion

- Convert to Parquet or ORC before delivery to S3
- Compresses file sizes and optimal format for Hadoop usage

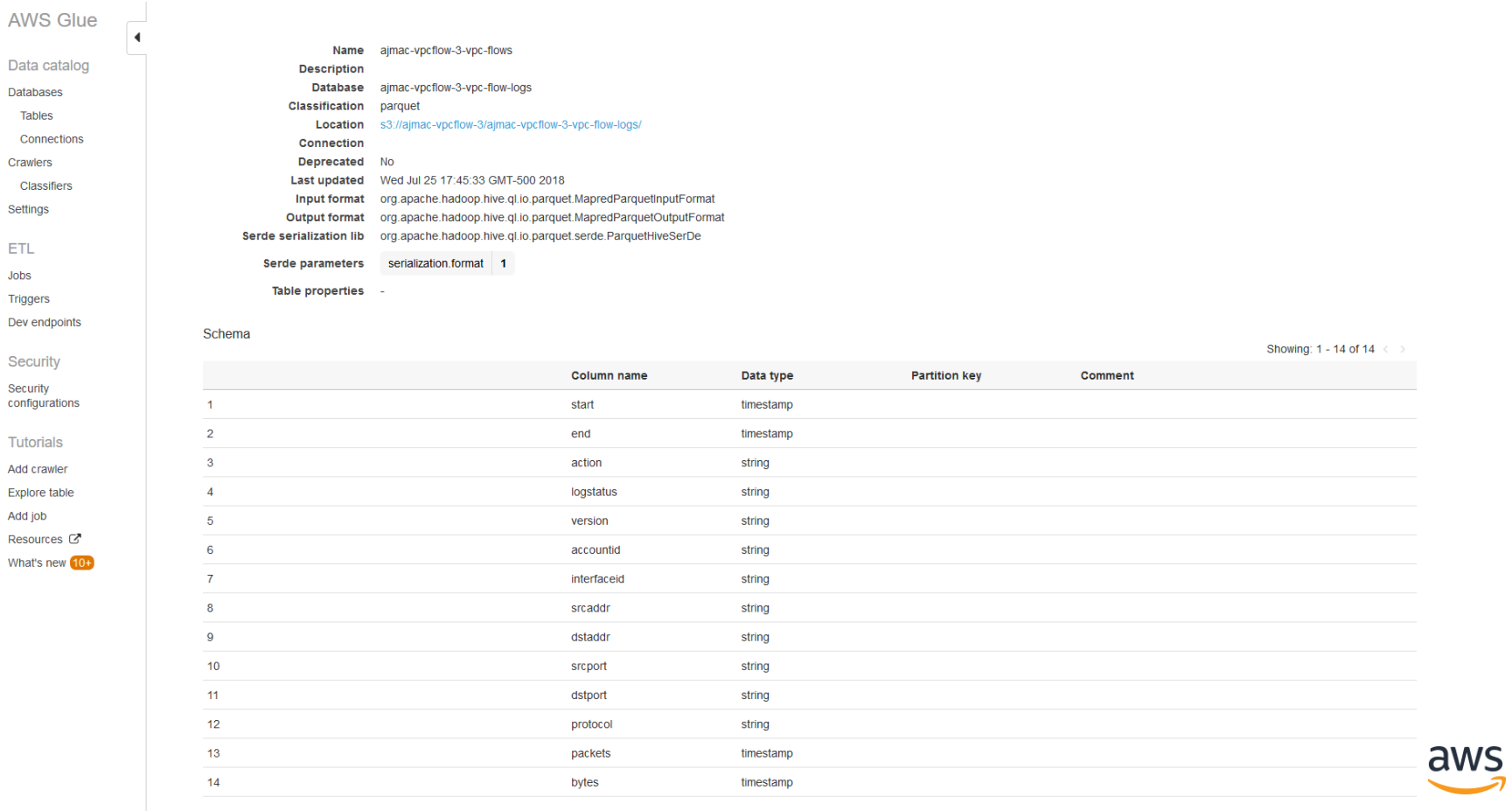
aws

Kinesis Data Firehose: Record format conversion

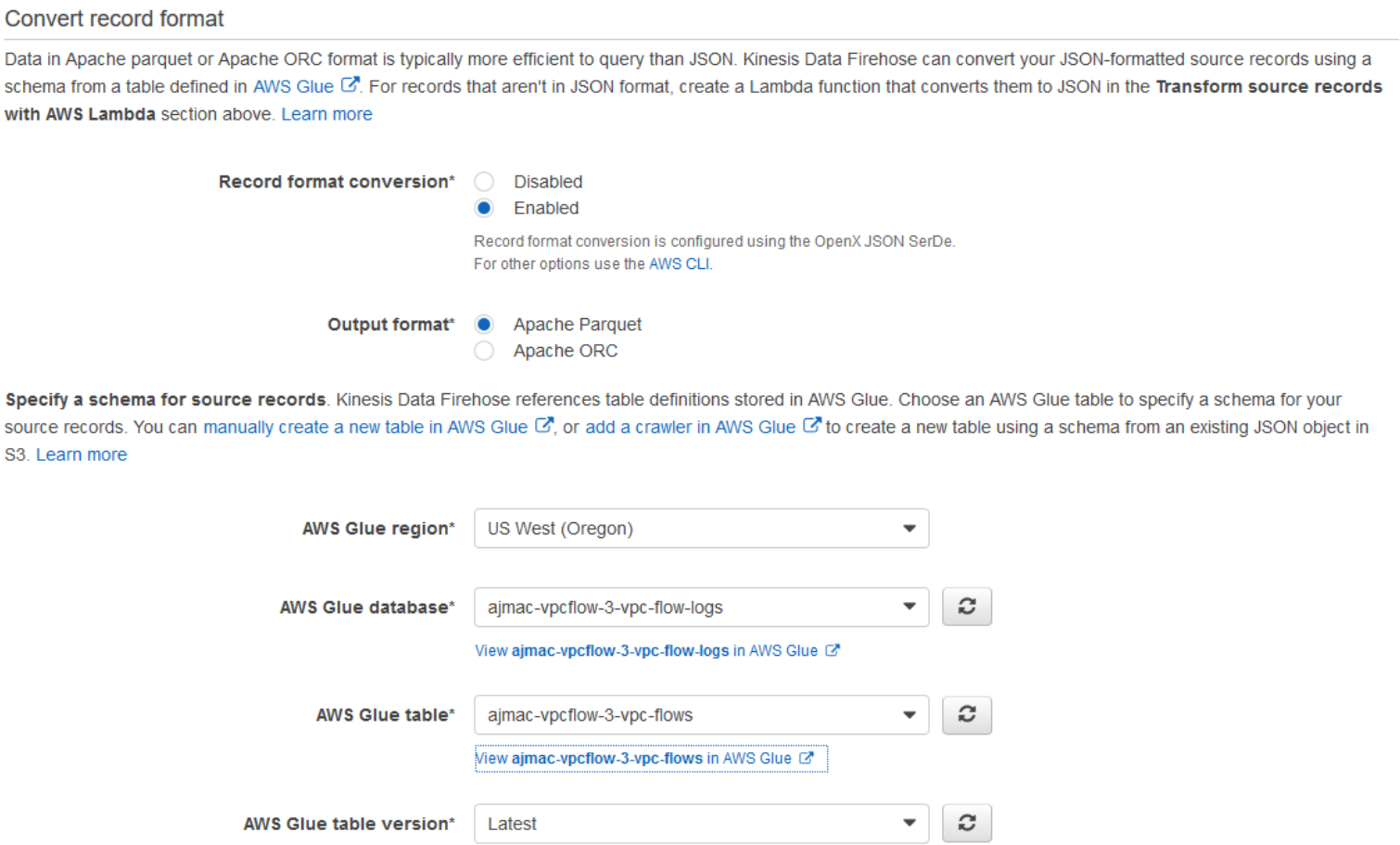


aws

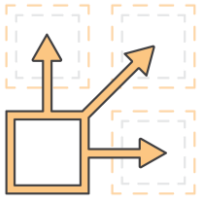
Kinesis Data Firehose: Record format conversion



Kinesis Data Firehose: Record format conversion



Kinesis Data Firehose: Record format conversion



When to use record format conversion:

- Delivering to S3 for use by Hadoop systems
  - Columnar formats are more efficient.
  - Compression leads to lower storage costs (vs. uncompressed data).

aws

Kinesis Data Firehose: Record format conversion

Pricing (US East - Ohio):

- $0.018 per GB of data converted

Learn More

Learn More

More information about Amazon Kinesis: http://aws.amazon.com/kinesis

Kinesis Data Streams enhanced fan-out:

- Documentation: Using Consumers with Enhanced Fan-out
- Blog post: Amazon Kinesis Data Streams Adds Enhanced Fan-out and HTTP/2 for Faster Streaming
- Develop an EFO consumer: Kinesis Client Library (KCL) 2.0

Kinesis Data Firehose data formation conversion:

- Documentation: Converting Your Input Record Format in Kinesis Data Firehose
- Blog post: Analyze Apache Parquet optimized data using Amazon Kinesis Data Firehose, Amazon Athena, and Amazon Redshift

Kinesis Data Firehose deliver to Splunk:

- Documentation: Streaming Real-time Data to Splunk
- Blog post: Power data ingestion into Splunk using Amazon Kinesis Data Firehose

aws

Thank you!