

New Innovations from Amazon Kinesis for Real-Time Analytics

Allan MacInnis, Solutions Architect

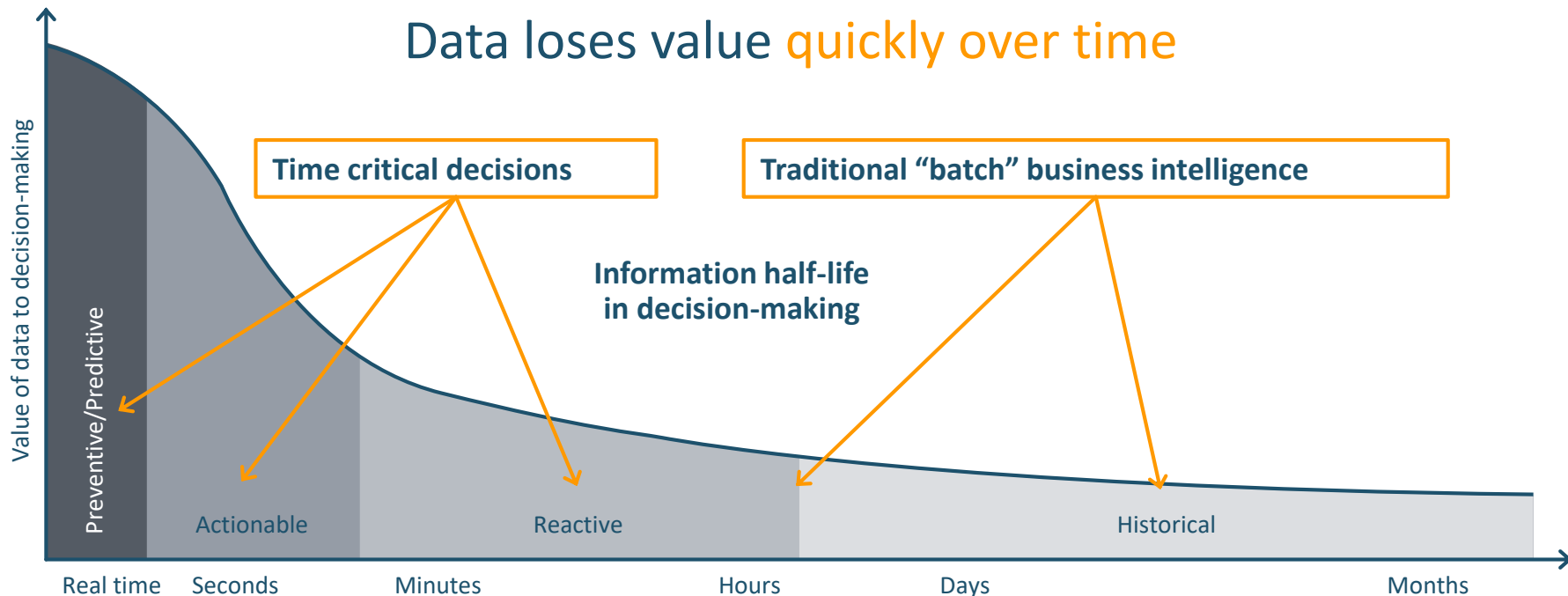
September 20, 2018

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

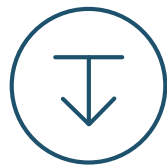


Source: Perishable insights, Mike Gualtieri, Forrester

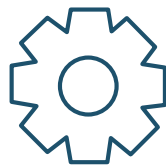
Stream New Data in Seconds

Get actionable insights quickly

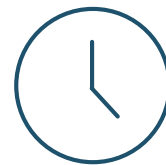
Streaming



Ingest data **as it's**
generated



Process data
on the fly



Real-time
analytics/ML, alerts,
actions

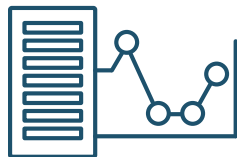
Most Common Uses of Streaming



Smart Home
Smart City



Industrial
Automation



Log
Analytics



Data
Lakes



IoT
Analytics

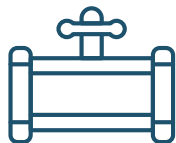
Streaming with Amazon Kinesis

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



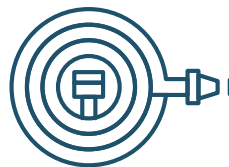
Kinesis Video Streams

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

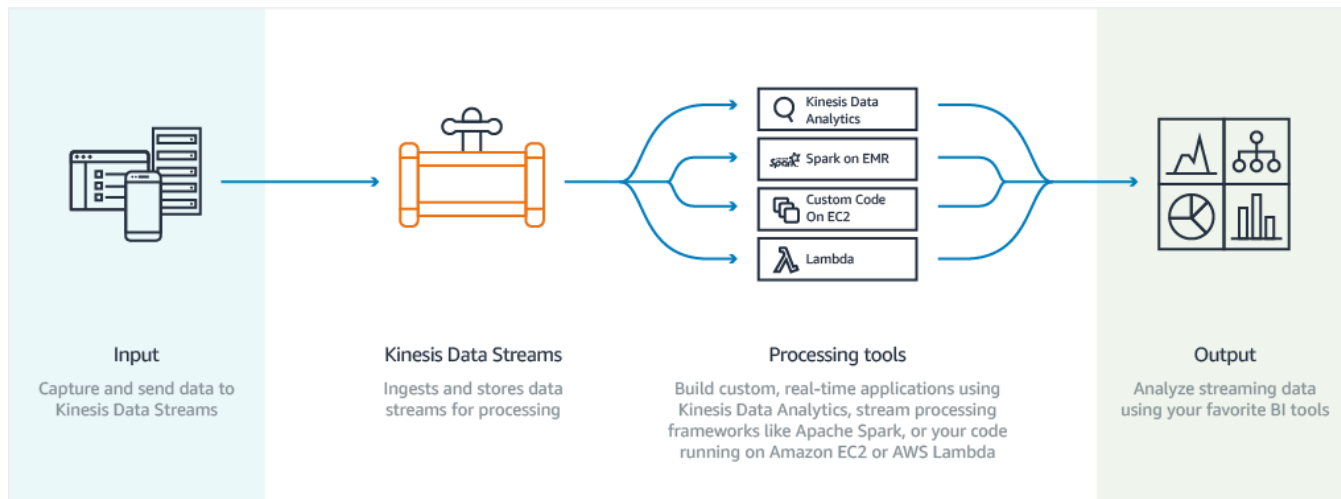
Analyze data streams with SQL

Kinesis Data Streams

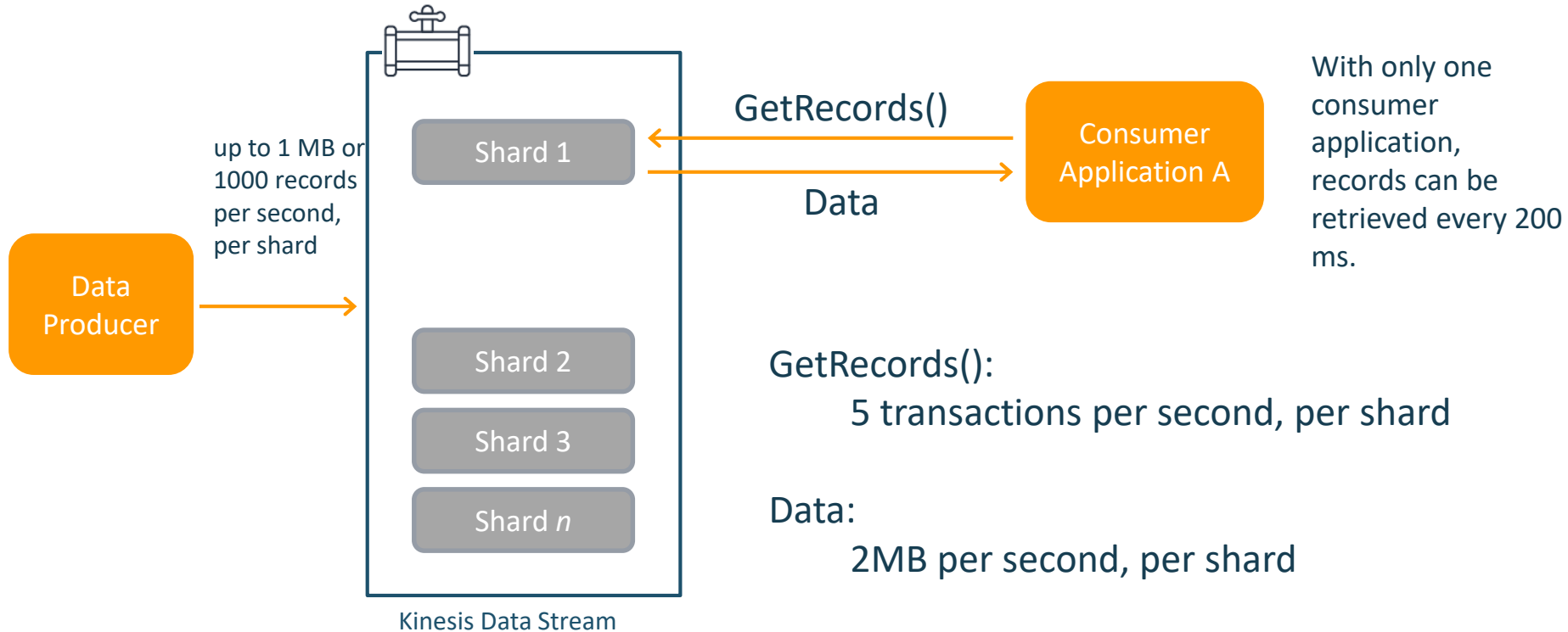
Enhanced fan-out



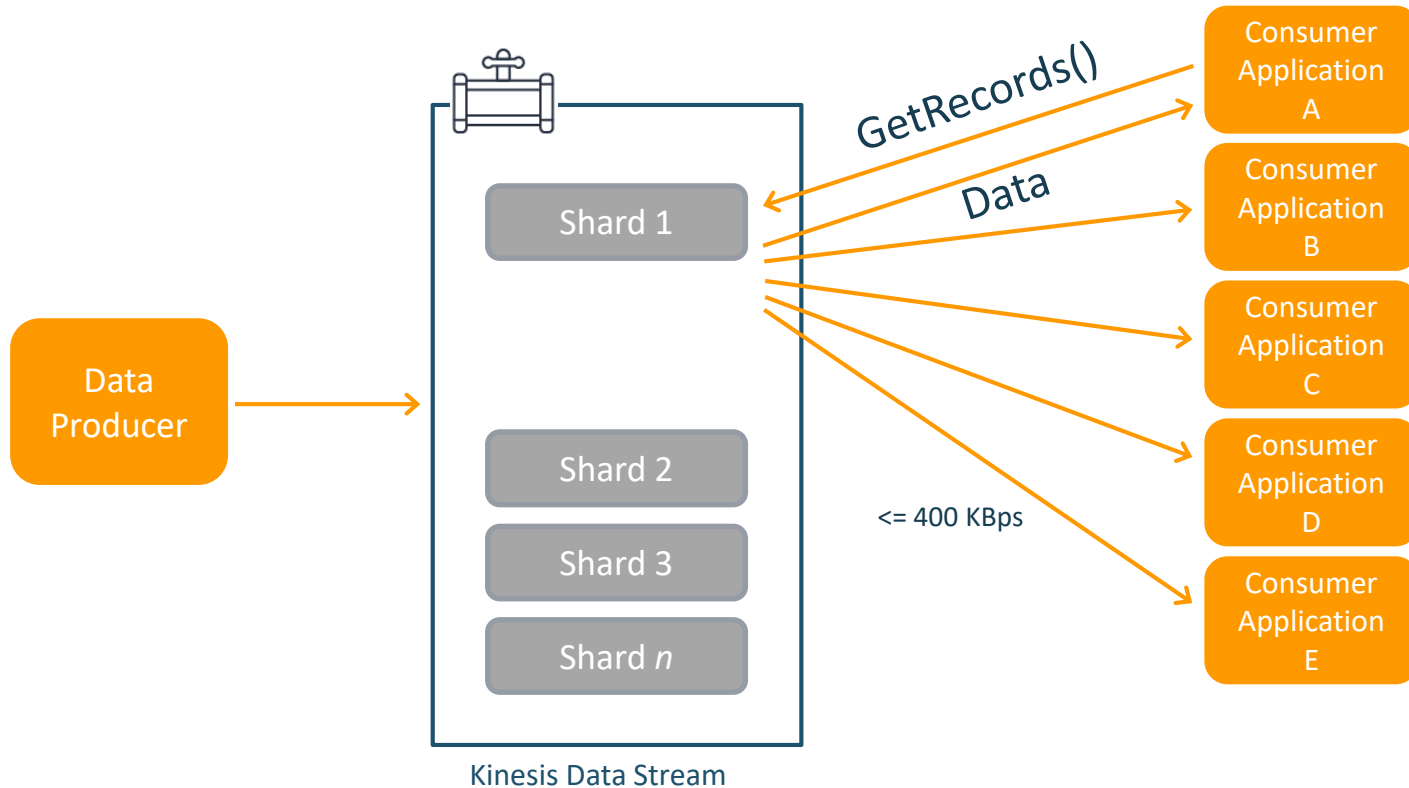
Amazon Kinesis Data Streams



Kinesis Data Streams: Standard egress limitations



Kinesis Data Streams: Standard egress limitations

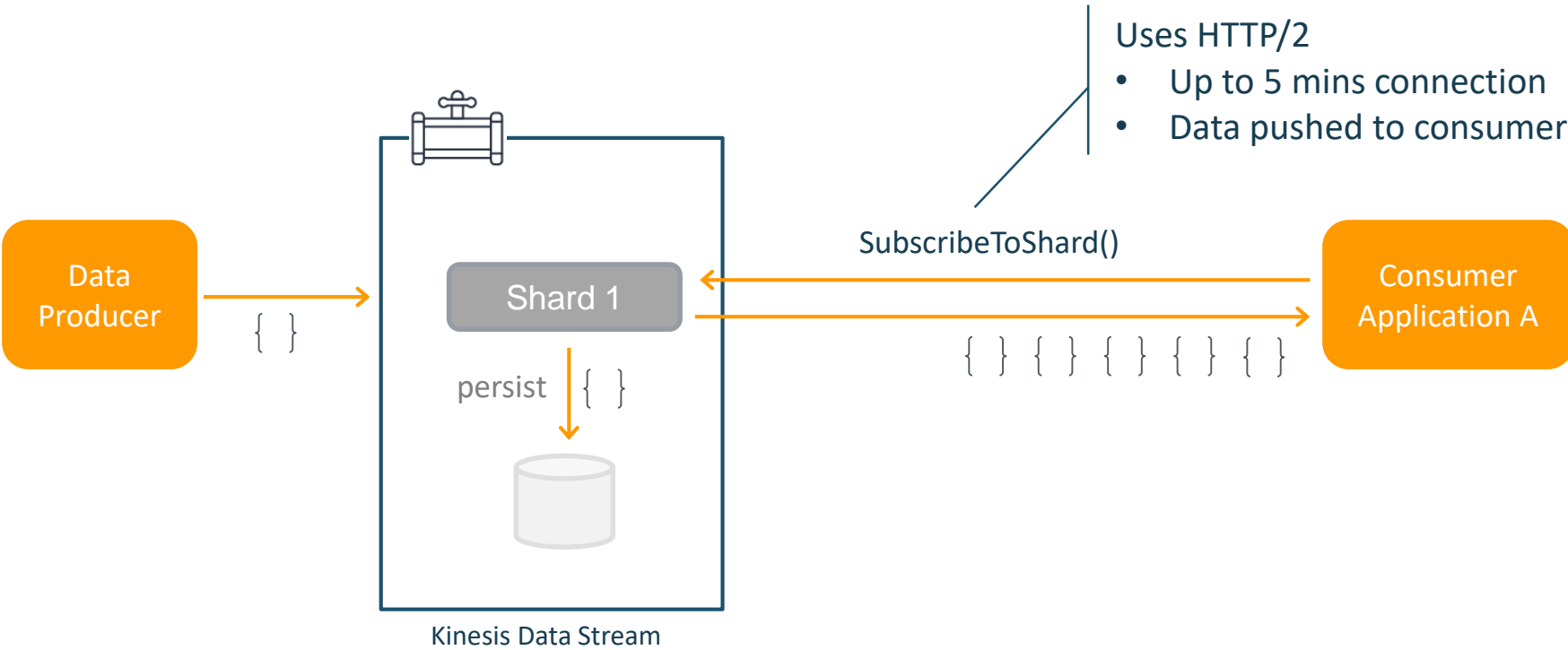


With more consumer applications, propagation delay increases.

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

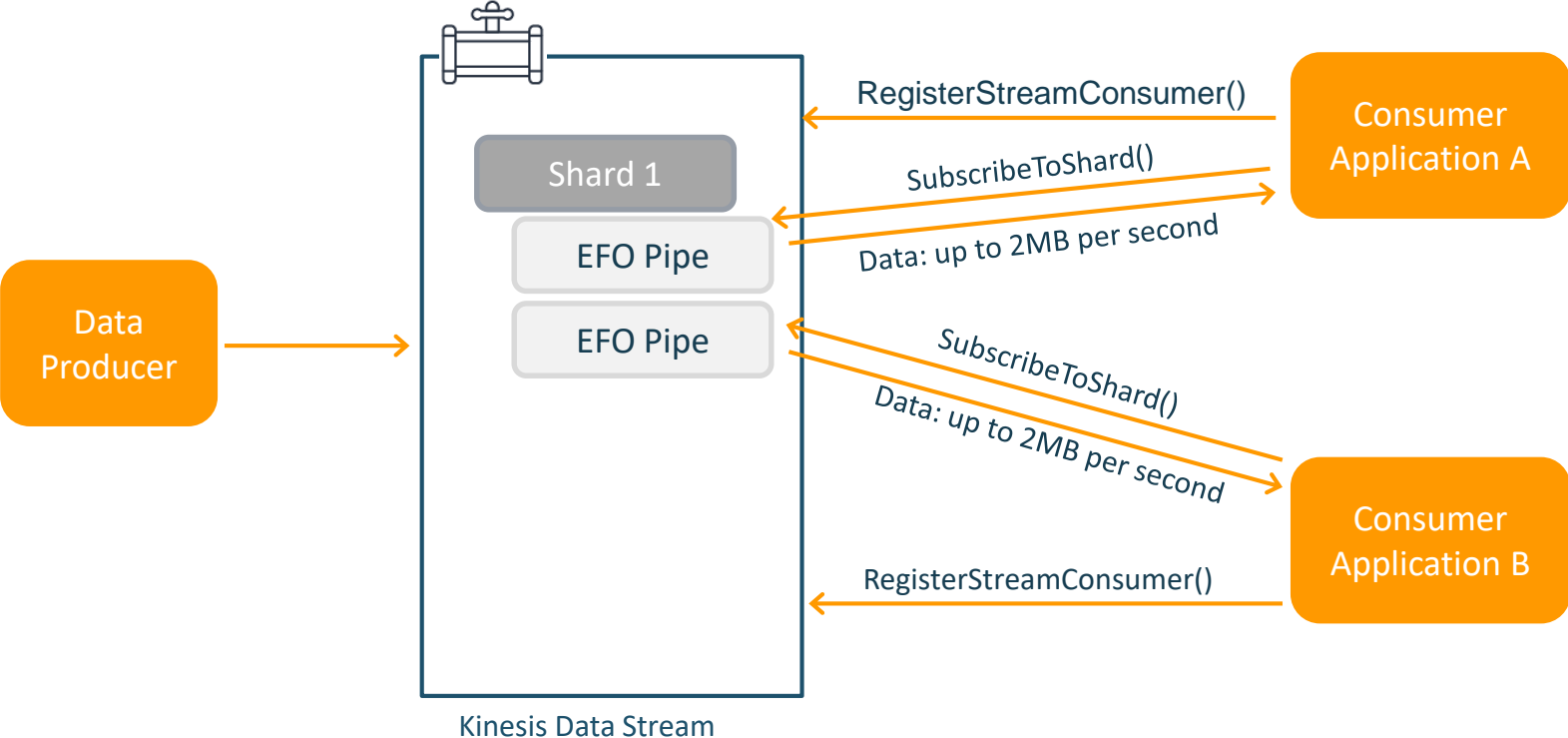
Kinesis Data Streams: Enhanced fan-out

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

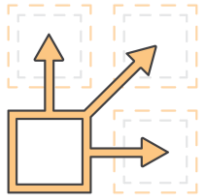


Kinesis Data Streams: Enhanced fan-out

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):

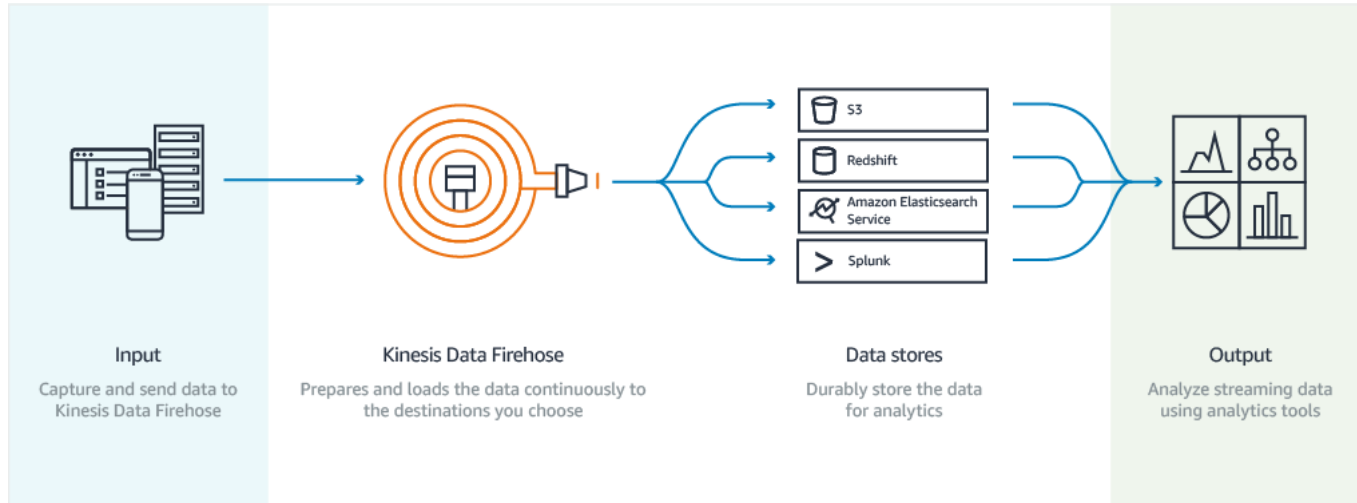
- \$0.013 per GB of data retrieved
- \$0.015 per consumer-shard hour

Kinesis Data Firehose

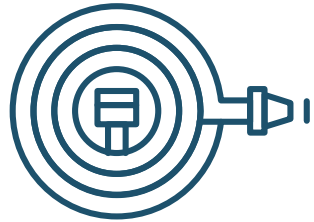
Splunk delivery



Kinesis Data Firehose



Kinesis Data Firehose: Splunk delivery



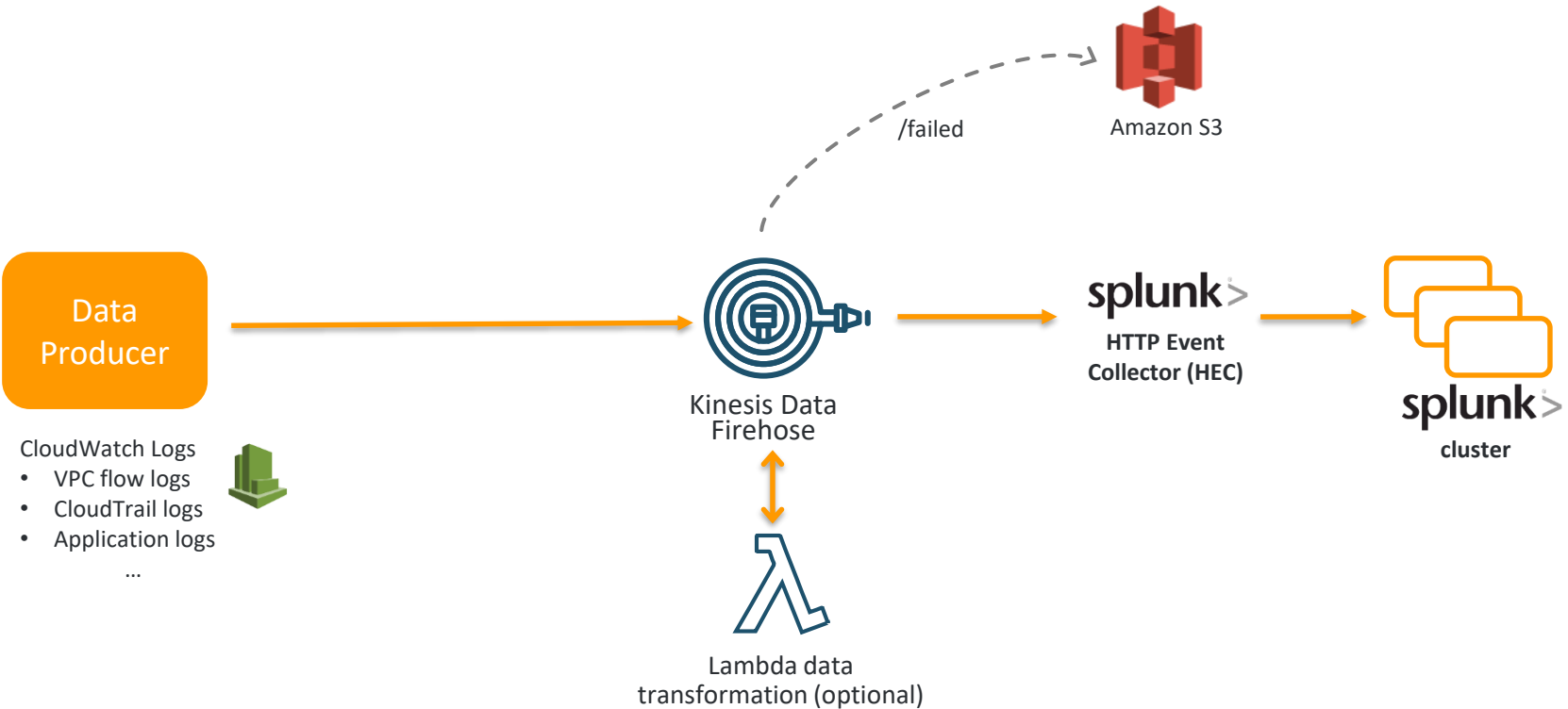
Kinesis Data Firehose



splunk >

- Deliver to your own Splunk index

Kinesis Data Firehose: Splunk delivery



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://

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*



Show token

HEC acknowledgement timeout*

seconds

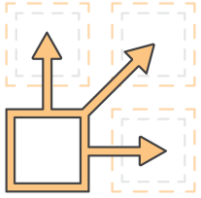
Specify a timeout duration from 180 - 600 seconds

Retry duration*

seconds

Specify a retry duration from 0 - 7200 seconds

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

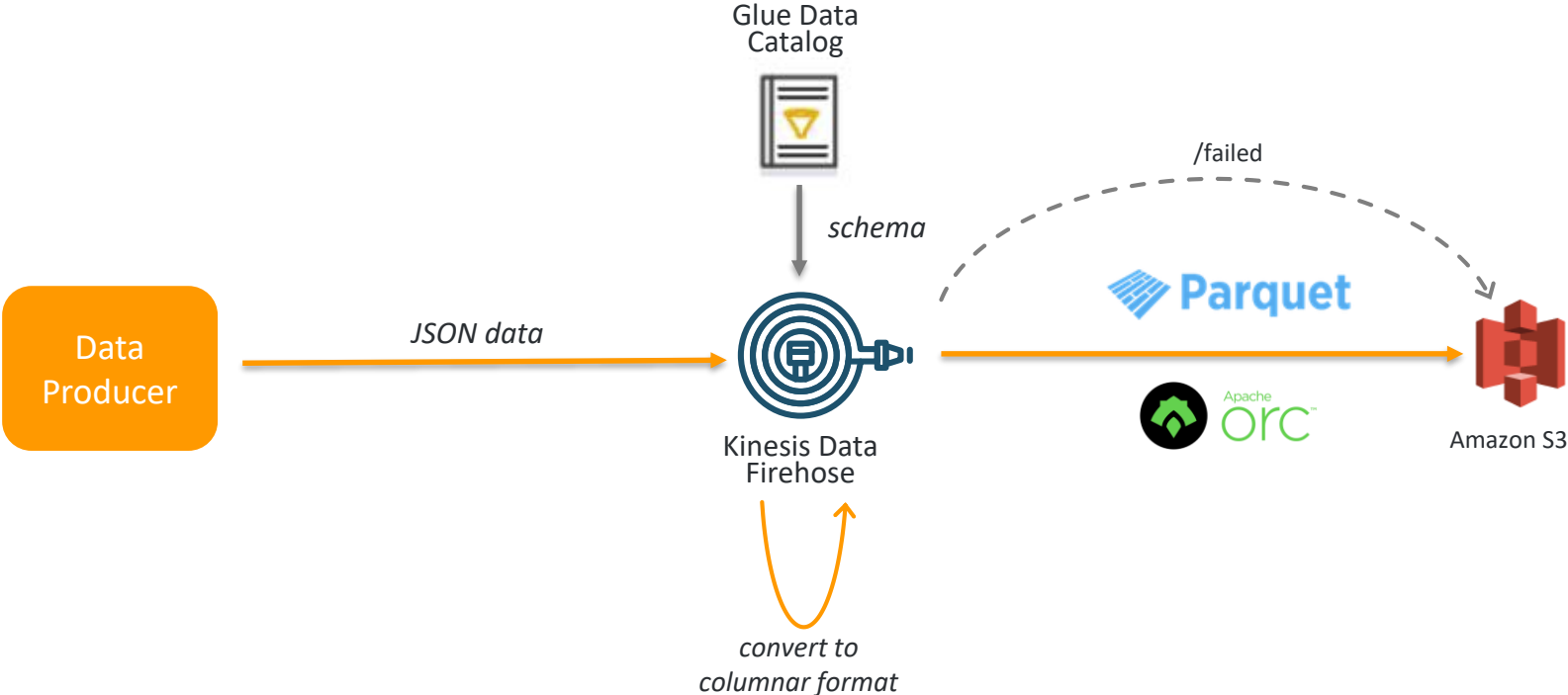


Kinesis Data Firehose: Record format conversion



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

Kinesis Data Firehose: Record format conversion



Kinesis Data Firehose: Record format conversion

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

Settings

ETL

Jobs

Triggers

Dev endpoints

Security

Security configurations

Tutorials

Add crawler

Explore table

Add job

Resources [↗](#)

What's new **10+**

Name	ajmac-vpctflow-3-vpc-flows		
Description			
Database	ajmac-vpctflow-3-vpc-flow-logs		
Classification	parquet		
Location	s3://ajmac-vpctflow-3/ajmac-vpctflow-3-vpc-flow-logs/		
Connection			
Deprecated	No		
Last updated	Wed Jul 25 17:45:33 GMT-500 2018		
Input format	org.apache.hadoop.hive.q1.io.parquet.MapredParquetInputFormat		
Output format	org.apache.hadoop.hive.q1.io.parquet.MapredParquetOutputFormat		
Serde serialization lib	org.apache.hadoop.hive.q1.io.parquet.serde.ParquetHiveSerDe		
Serde parameters	<table><tr><td>serialization.format</td><td>1</td></tr></table>	serialization.format	1
serialization.format	1		
Table properties	-		

Schema

Showing: 1 - 14 of 14 < >

	Column name	Data type	Partition key	Comment
1	start	timestamp		
2	end	timestamp		
3	action	string		
4	logstatus	string		
5	version	string		
6	accountid	string		
7	interfaceid	string		
8	srcaddr	string		
9	dstaddr	string		
10	srcport	string		
11	dstport	string		
12	protocol	string		
13	packets	timestamp		
14	bytes	timestamp		



Kinesis Data Firehose: Record format conversion

Convert record format

Data in Apache parquet or Apache ORC format is typically more efficient to query than JSON. Kinesis Data Firehose can convert your JSON-formatted source records using a schema from a table defined in [AWS Glue](#). For records that aren't in JSON format, create a Lambda function that converts them to JSON in the **Transform source records with AWS Lambda** section above. [Learn more](#)

Record format conversion* Disabled
 Enabled

Record format conversion is configured using the OpenX JSON SerDe.
For other options use the [AWS CLI](#).

Output format* Apache Parquet
 Apache ORC

Specify a schema for source records. Kinesis Data Firehose references table definitions stored in AWS Glue. Choose an AWS Glue table to specify a schema for your source records. You can [manually create a new table in AWS Glue](#), or [add a crawler in AWS Glue](#) to create a new table using a schema from an existing JSON object in S3. [Learn more](#)

AWS Glue region*

AWS Glue database*



[View ajmac-vpcflow-3-vpc-flow-logs in AWS Glue](#)

AWS Glue table*

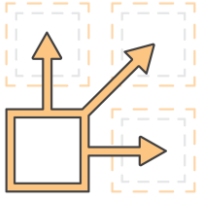


[View ajmac-vpcflow-3-vpc-flows in AWS Glue](#)

AWS Glue table version*



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).

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 format 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](#)

Thank you!