

Streaming ETL for Data Lakes using Amazon Kinesis Firehose

Ryan Nienhuis, Sr. Product Manager, Amazon Kinesis

4/19/2017

Agenda

- Data Lake Overview
- Kinesis Firehose Overview
- Demo and Walkthrough
- Q & A

Data Lake

Data Lake Capabilities

- Collecting and storing any type of data, at any scale and at low costs
- Securing and protecting all of data stored in the central repository
- Searching and finding the relevant data in the central repository
- Quickly and easily performing new types of data analysis on datasets
- Querying the data by defining the data's structure at the time of use (schema on read)

Data Lake Capabilities

- **Collecting and storing any type of data, at any scale and at low costs**
- Securing and protecting all of data stored in the central repository
- Searching and finding the relevant data in the central repository
- Quickly and easily performing new types of data analysis on datasets
- Querying the data by defining the data's structure at the time of use (schema on read)

Data Lake Storage

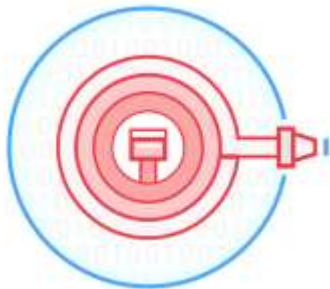


Amazon S3

highly scalable and durable object storage

**for any type of data, at any scale
and at low costs**

Data Lake Ingestion

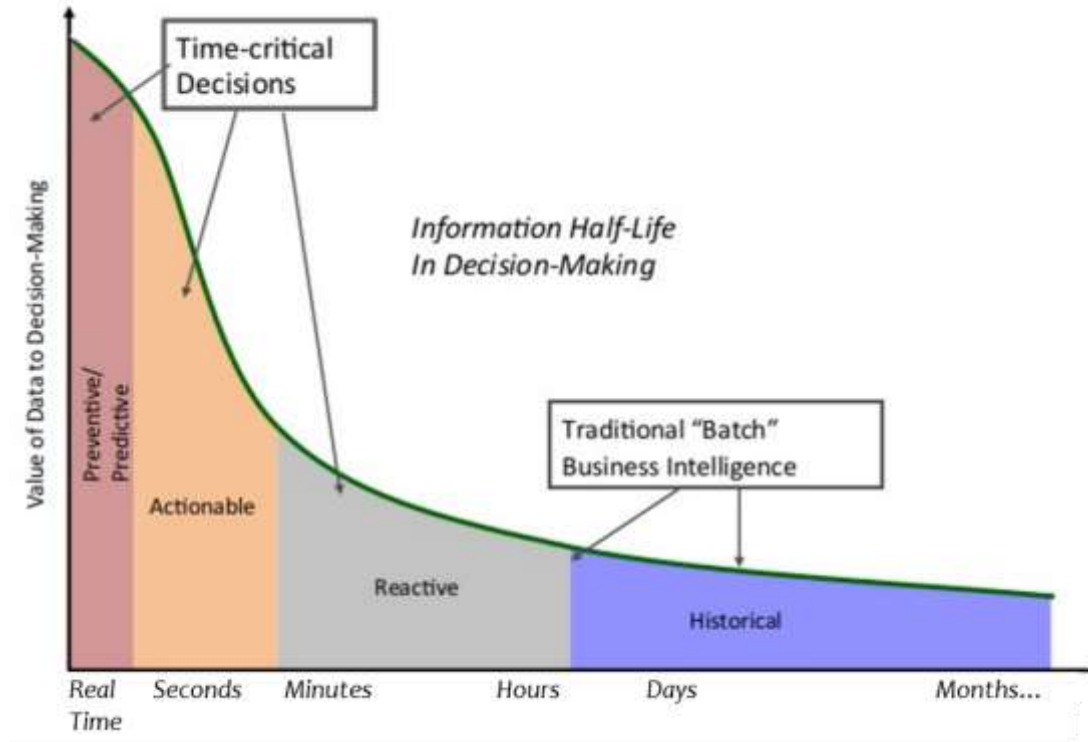


Kinesis Firehose

Real-time streaming data ETL

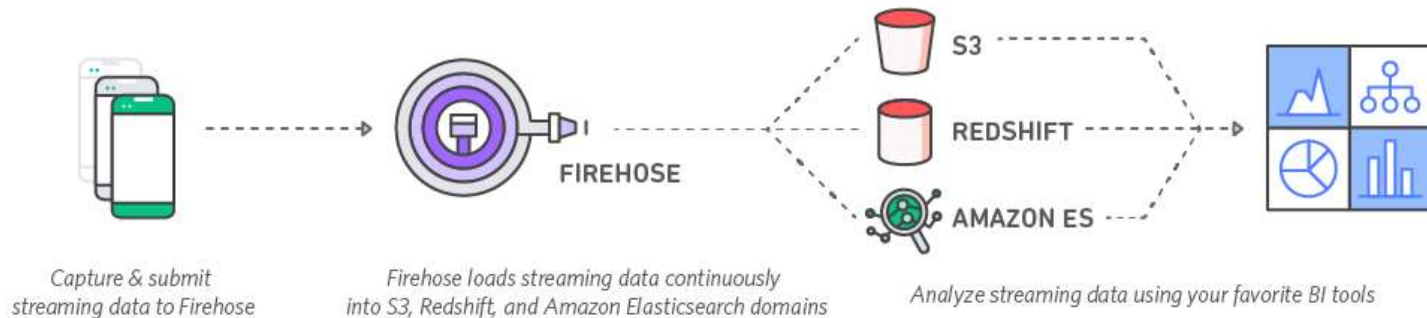
**for any type of data, at any scale
and at low costs**

Time Value of Money Data

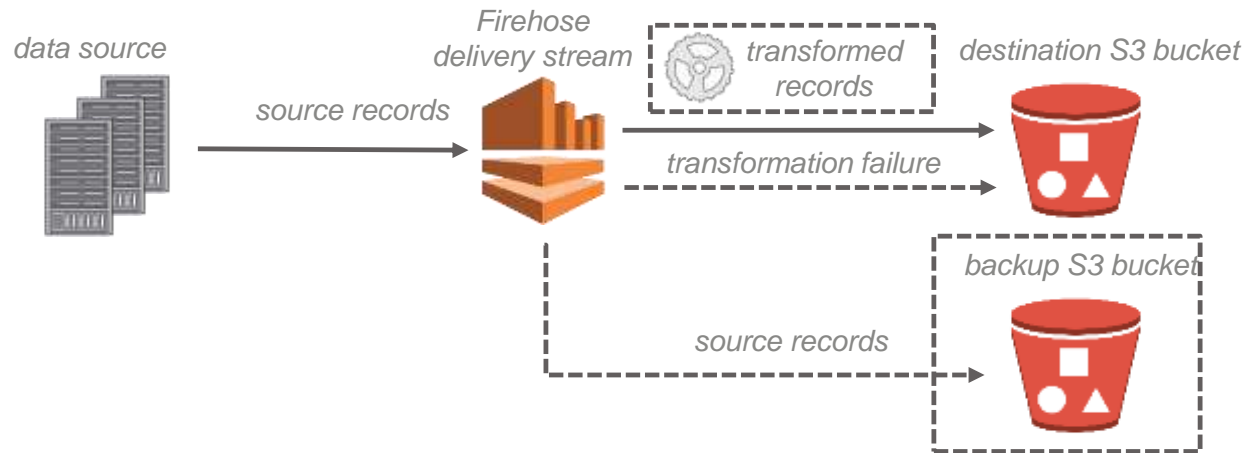


Kinesis Firehose

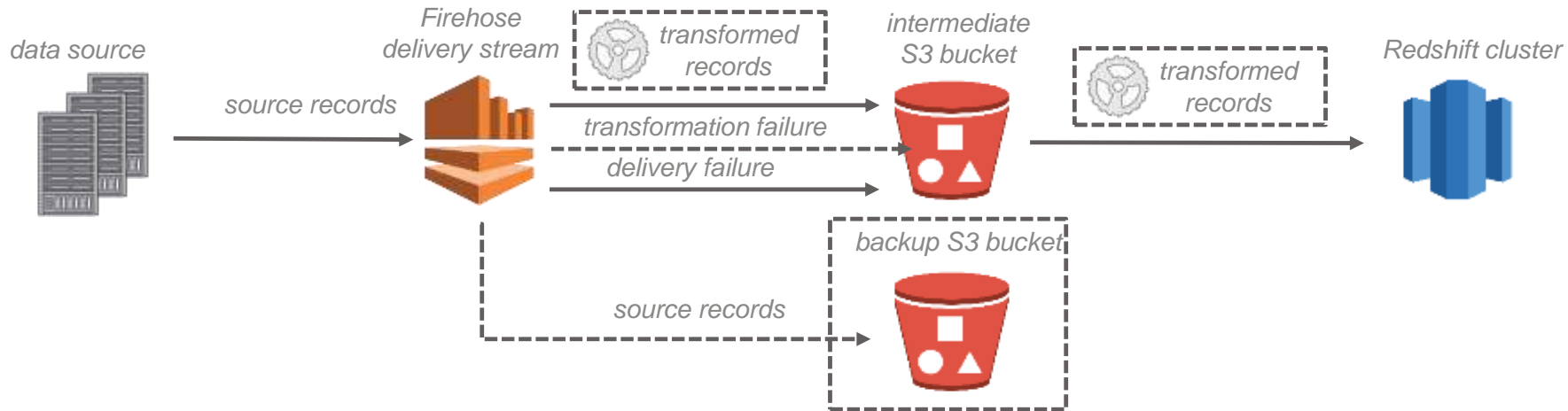
Kinesis Firehose



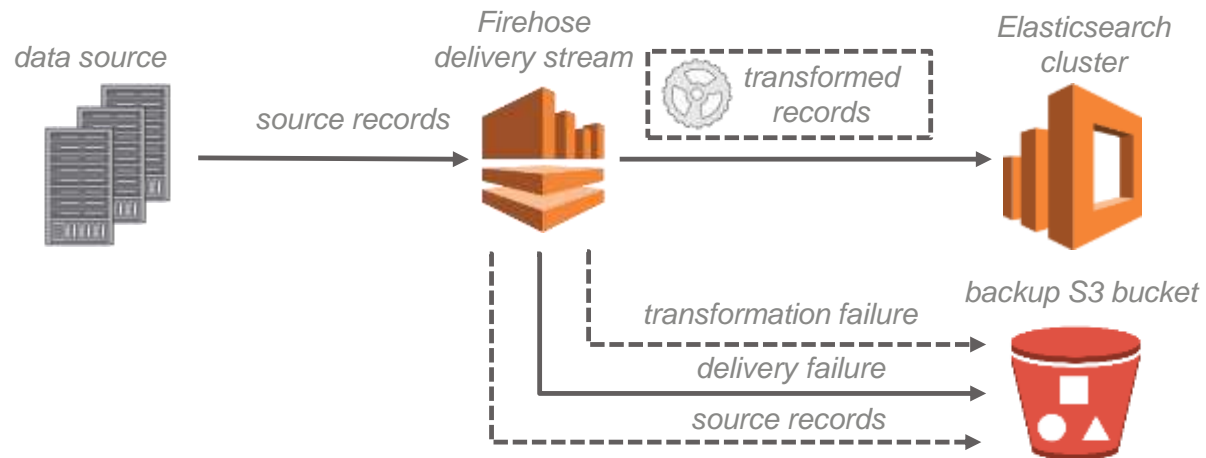
Streaming ETL to S3



Streaming ETL to Redshift



Streaming ETL to Elasticsearch



Demo and Walkthrough

Step 1 Set Up Firehose Delivery Stream and **Configure Data Transformation**

Destination

Create Delivery Stream

Step 1: Destination

Step 2: Configuration

Step 3: Review

Destination

Select the destination where your streaming data will be delivered.

Destination*

Amazon S3 

Delivery stream name*

datalakedemo 

S3 Bucket

S3 bucket*

datalaketutorial  

S3 prefix

formattedlogs/ 

*Required

Cancel

Next

Configuration

Create Delivery Stream

[Step 1: Destination](#)

Step 2: Configuration

[Step 3: Review](#)

Configuration



Configure buffer, compression, logging and IAM role options for your delivery stream.

Data transformation with AWS Lambda

AWS Lambda lets you run code without provisioning or managing servers. Firehose can invoke your Lambda function to transform data before delivering it to destinations. [Learn more](#).

Data transformation* ☐ Disable
☒ Enable

Your Lambda function must be compliant with the record transformation and status model required by Firehose in order to return transformed records from Lambda to Firehose. [Learn more](#)

► (Optional) Create a new Lambda function

Lambda function* datalakefunction ↻

Lambda function version* Latest

Lambda function description An Amazon Kinesis Firehose stream processor that converts input records from Apache Common Log format to JSON.

Lambda function runtime nodejs6.10

Lambda function timeout 3 minutes ↻

Configuring your function timeout to be 1 minute or longer ensures that data transformation will be complete before the function times out.

[View/edit function in Lambda](#)

Configuration

Source record backup

Enabling a source record backup ensures that source records can be recovered if data transformation does not produce the desired results.

[Learn more](#)

Source record backup* ☐ Disabled

☒ Enabled

Source record backup bucket* datalaketutorial ▼

Source record backup prefix rawlogs/

S3 Buffer

Firehose buffers incoming data before delivering to your S3 bucket. You can configure buffer size and buffer interval. The first satisfied condition will trigger the data delivery to your S3 bucket.

Buffer size* 1

Buffer size can range from 1MB to 128MB in 1MB increments.

Buffer interval* 60

Buffer interval can range from 60s to 900s in 1 second increments.

S3 Compression and Encryption

Firehose can compress and encrypt the data before delivering to your S3 bucket.

Data compression UNCOMPRESSED ⓘ

Data encryption No Encryption ⓘ

Error Logging

Firehose can log data delivery errors to CloudWatch Logs. If enabled, a CloudWatch Log Group and corresponding Log Stream(s) are created on your behalf. [Learn more](#).

☒ Enable ☐ Disable

Review

Create Delivery Stream

[Step 1: Destination](#)

[Step 2: Configuration](#)

| Step 3: Review

Review



Review your destination and configuration before creating your delivery stream.

Destination

Edit

Destination	Amazon S3
Delivery stream name	datalakedemo
S3 bucket	datalaketutorial
S3 prefix	formattedlogs/

Configuration

Edit

Data transformation	Enabled
Lambda function	↗ datalakefunction
Lambda function version	Latest
Lambda function description	An Amazon Kinesis Firehose stream processor that converts input records from Apache Common Log format to JSON.
Lambda function runtime	nodejs6.10

Step 2 Send Data to Firehose Delivery Stream

Sample Data

219.134.32.117 - - [16/Feb/2017:09:38:20 -0800] "GET /wp-content HTTP/1.1" 200 4521
"-" "Mozilla/5.0 (compatible; MSIE 10.0; Windows NT 6.1; Trident/5.1; .NET CLR
3.8.23015.5)"

95.169.41.62 - - [16/Feb/2017:09:38:20 -0800] "PUT /app/main/posts HTTP/1.1" 200
3883 "-" "Mozilla/5.0 (Windows NT 6.2; Trident/7.0; rv:11.0) like Gecko"

221.147.191.247 - - [16/Feb/2017:09:38:20 -0800] "GET /explore HTTP/1.1" 200 6579 "-"
"Mozilla/5.0 (Windows; U; Windows NT 5.1) AppleWebKit/538.0.1 (KHTML, like Gecko)
Chrome/38.0.895.0 Safari/538.0.1"

179.96.123.130 - - [16/Feb/2017:09:38:20 -0800] "GET /list HTTP/1.1" 200 560 "-"
"Mozilla/5.0 (Windows NT 6.3; Win64; x64; rv:5.4) Gecko/20100101 Firefox/5.4.6"

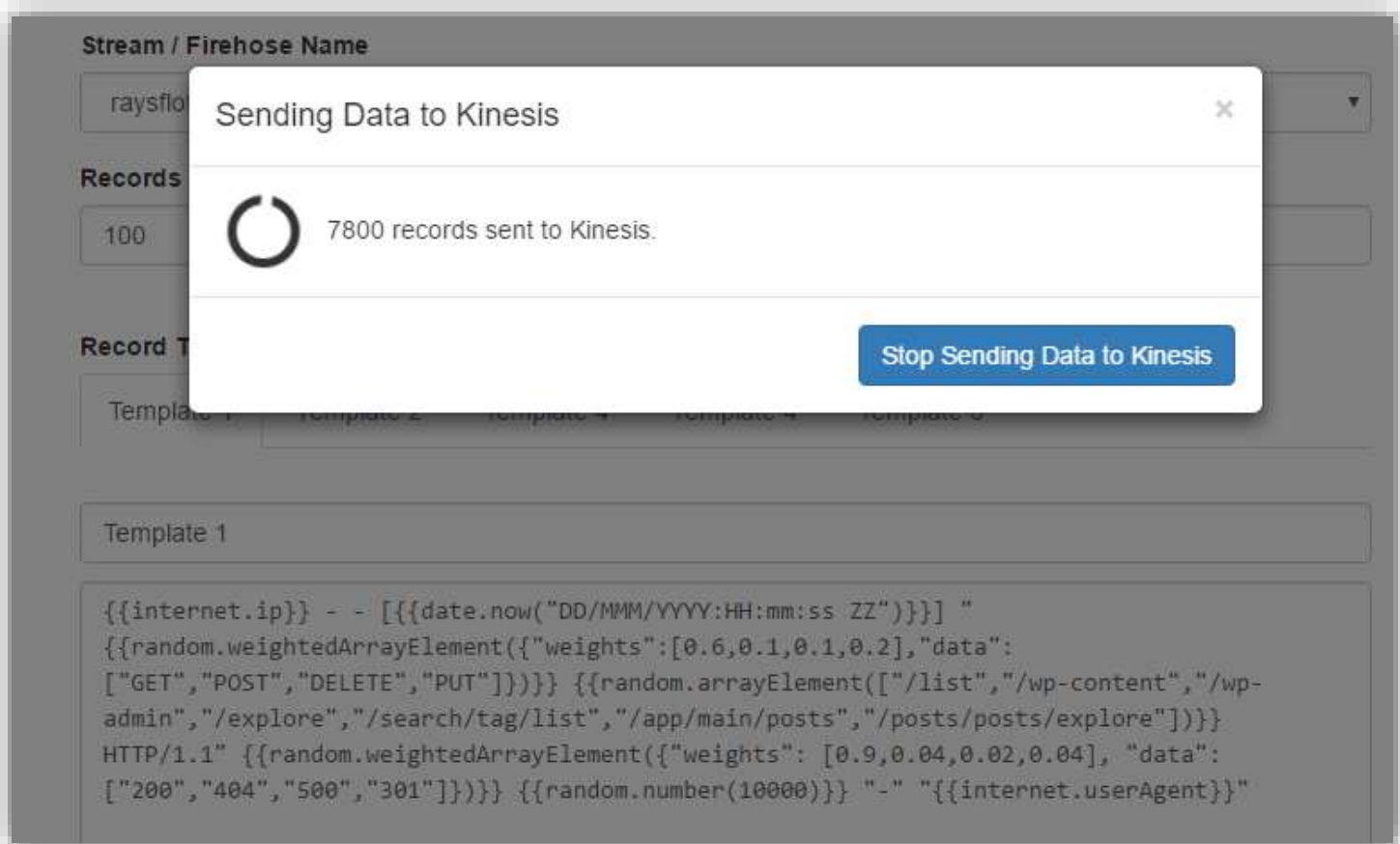
132.119.12.76 - - [16/Feb/2017:09:38:20 -0800] "PUT /explore HTTP/1.1" 200 3131 "-"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_0 rv:5.0; AZ) AppleWebKit/535.1.0
(KHTML, like Gecko) Version/4.0.3 Safari/535.1.0"

74.113.56.92 - - [16/Feb/2017:09:38:20 -0800] "DELETE /app/main/posts HTTP/1.1" 200
7069 "-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_7_9) AppleWebKit/532.1.0
(KHTML, like Gecko) Chrome/15.0.877.0 Safari/532.1.0"

After Data Transformation

```
{"host":"26.56.11.130","ident":"-","authuser":"-","request":"GET /wp-content  
HTTP/1.1","response":200,"bytes":4582,"verb":"GET","@timestamp":"2017-04-  
04T11:32:29.000Z","timezone":"-0700","@timestamp_utc":"2017-04-04T18:32:29.000Z"}  
{"host":"180.153.215.216","ident":"-","authuser":"-","request":"PUT /search/tag/list  
HTTP/1.1","response":200,"bytes":1461,"verb":"PUT","@timestamp":"2017-04-  
04T11:32:29.000Z","timezone":"-0700","@timestamp_utc":"2017-04-04T18:32:29.000Z"}  
{"host":"155.233.163.37","ident":"-","authuser":"-","request":"GET /explore  
HTTP/1.1","response":500,"bytes":326,"verb":"GET","@timestamp":"2017-04-  
04T11:32:29.000Z","timezone":"-0700","@timestamp_utc":"2017-04-04T18:32:29.000Z"}  
{"host":"189.176.106.5","ident":"-","authuser":"-","request":"POST /search/tag/list  
HTTP/1.1","response":200,"bytes":3059,"verb":"POST","@timestamp":"2017-04-  
04T11:32:29.000Z","timezone":"-0700","@timestamp_utc":"2017-04-04T18:32:29.000Z"}
```

Send Data



The screenshot shows the AWS Kinesis console interface. A modal dialog box titled "Sending Data to Kinesis" is centered on the screen. The dialog has a close button (X) in the top right corner. Below the title bar, there is a circular progress indicator and the text "7800 records sent to Kinesis." At the bottom right of the dialog is a blue button labeled "Stop Sending Data to Kinesis".

Stream / Firehose Name
raysflo

Records
100

Record Template
Template 1 Template 2 Template 3 Template 4 Template 5

Template 1

```
{{internet.ip}} - - [{{date.now("DD/MMM/YYYY:HH:mm:ss ZZ")}}] "  
{{random.weightedArrayElement({"weights": [0.6, 0.1, 0.1, 0.2], "data":  
["GET", "POST", "DELETE", "PUT"]})}} {{random.arrayElement(["/list", "/wp-content", "/wp-  
admin", "/explore", "/search/tag/list", "/app/main/posts", "/posts/posts/explore"])}}  
HTTP/1.1" {{random.weightedArrayElement({"weights": [0.9, 0.04, 0.02, 0.04], "data":  
["200", "404", "500", "301"]})}} {{random.number(10000)}} "-" "{{internet.userAgent}}"
```

Step 3 Check Results in S3

Path: /



File	Size	Type	Last Modified
 formattedlogs/			
 rawlogs/			

Step 4 Monitor Streaming Data Pipeline

Monitor with CloudWatch Metrics

Use the tabs below to view, edit and monitor your delivery stream.

Details

Monitoring

S3 Logs

Delete Delivery Stream

Delivery Stream metrics

Time range: Last 1 hour

Period: 1 minute(s)



Go to [CloudWatch](#) for a complete list of Firehose metrics. [Learn more.](#)

All graphs are displayed in UTC time zone.

IncomingBytes (Sum)



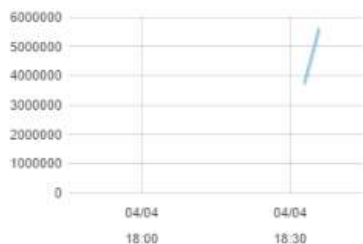
IncomingRecords (Sum)



DeliveryToS3 DataFreshness (Maximum)



DeliveryToS3 Bytes (Sum)



DeliveryToS3 Records (Sum)



DeliveryToS3 Success (Average)



Monitor with CloudWatch Logs

Delivery Streams > datalakedemo

▸ Test with demo data

Use the tabs below to view, edit and monitor your delivery stream.

Details

Monitoring

S3 Logs

Delete Delivery Stream

Log Group /aws/kinesisfirehose/datalakedemo

Log Stream S3Delivery



Time	Destination	Message	Error Code	Version
2017-04-04T11:53:16.16T-0700D	arn:aws:s3:::sydney-ray-bucket-test	The specified bucket does not exist. Create the bucket or use a different bucket name that does exist.	S3.NoSuchBucket	2

« < Viewing 1 - 1 of 1 items > »

Firehose Pricing

Pricing

Pricing by Region

Region:

US East (N. Virginia) ▾

	Data Ingested, per GB
First 500 TB/Month	\$0.029
Next 1.5 PB/Month	\$0.025
Next 3 PB/Month	\$0.020
Over 5 PB/Month	Contact Us

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS is subject to Japanese Consumption Tax. [Learn more](#).

Q & A

Thank you!