amazon kinesis

What is amazon kinesis

Amazon Kinesis is a cloud-based service provided by Amazon Web Services (AWS) that enables real-time streaming and processing of large amounts of data. It is designed to collect, process, and analyze data from various sources such as websites, mobile devices, IoT sensors, and log files. With Amazon Kinesis, you can ingest, store, and process data streams in real time, making it suitable for applications that require immediate analysis and response.

Amazon kinesis Data Stream

That enables you to collect and process large amounts of streaming data in real time. It is a core component of the Amazon Kinesis platform and provides a scalable and durable solution for handling data streams

Data Streams: A data stream is the fundamental entity in Amazon Kinesis Data Streams. It represents an ordered sequence of data records. Each data record consists of a payload (the actual data) and an associated sequence number. You can think of a data stream as a pipeline through which you can continuously ingest and process streaming data.

Shards: Data streams are divided into shards, which are the units of scalability and parallelism in Amazon Kinesis Data Streams. Each shard is an independent unit that allows for concurrent processing of data records. The number of shards determines the overall capacity of a data stream. You can dynamically adjust the number of shards to scale the throughput of your stream.

Producers: Producers are the entities that send data records to a data stream. They can be applications or devices generating data in real time. Producers use the Amazon Kinesis Data Streams API or AWS SDKs to push data records into the stream. data stream can have a maximum size of 1Mb are 1000 messages for sec. example for producers mobile , computers , log devices , stream devices

Consumers: Consumers are the entities that retrieve and process data records from a data stream. They can be applications or services that perform real-time analytics, generate insights, or store the data in other systems. Consumers use the Amazon Kinesis Data Streams API or AWS SDKs to read data records from the stream.

example for consumers lambda fucntion , kinesis data analysis , kinesis firehouse , s3bucket etc .. data stream can have a maximum size of 2mb sec / 2000 messages for sec

Retention Period: Amazon Kinesis Data Streams stores data records for a configurable retention period, which can range from 24 hours to 7 days. After the retention period, the data records are automatically deleted from the stream.

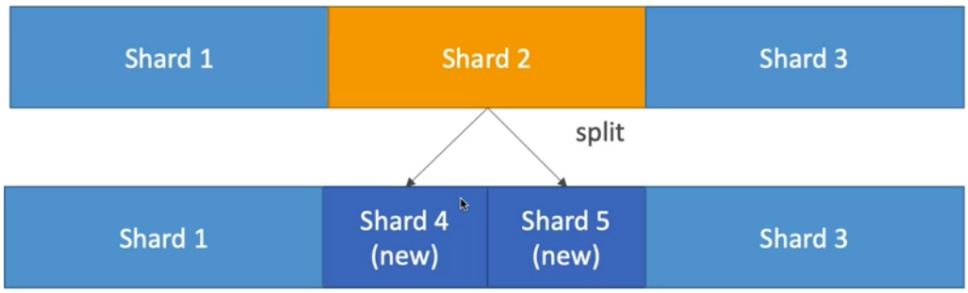
you can also modifie the maximum value of a stream's retention period is 8760 hours (365 days)

Data Record Size: Each data record in a data stream can have a maximum size of 1 MiB (before base64 encoding).

Enhanced Fan-Out: Enhanced Fan-Out is a feature of Amazon Kinesis Data Streams that allows you to build applications that can consume data from a stream with low latency and high throughput. It enables multiple consumers to read from the same shard concurrently, providing a scalable solution for real-time data processing.

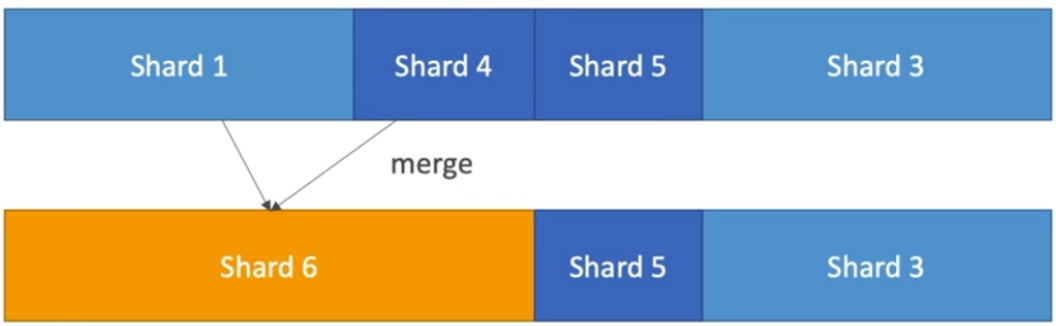
Kinesis Operations – Adding Shards

- Also called "Shard Splitting"
- Can be used to increase the Stream capacity (1 MB/s data in per shard)
- Can be used to divide a "hot shard"
- The old shard is closed and will be deleted once the data is expired



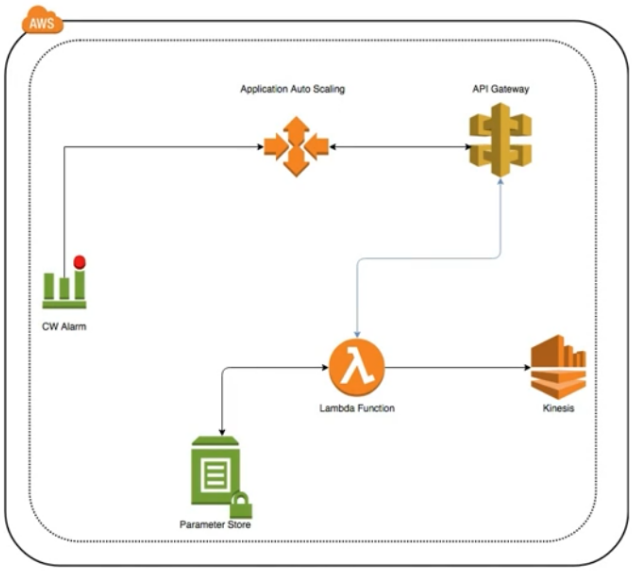
Kinesis Operations – Merging Shards

- Decrease the Stream capacity and save costs
- Can be used to group two shards with low traffic
- Old shards are closed and deleted based on data expiration



Kinesis Operations – Auto Scaling

- Auto Scaling is not a native feature of Kinesis
- The API call to change the number of shards is UpdateShardCount
- We can implement Auto Scaling with AWS Lambda
- See: https://aws.amazon.com/blogs/big-data/scaling-amazon-kinesis-data-streams-with-aws-application-auto-scaling/



Scaling Policies: You define scaling policies that specify the target utilization level for the stream, such as a percentage of shard utilization or a specific metric like GetRecords.IteratorAgeMilliseconds.

CloudWatch Alarms: CloudWatch Alarms are used to monitor the metrics associated with the scaling policies. These alarms track the stream's performance and trigger scaling actions when certain thresholds are breached.

Scaling Actions: When a CloudWatch Alarm is triggered, an associated scaling action is performed. Scaling actions can either add shards to the stream to handle increased data rates or merge shards to reduce the number of shards during low traffic periods.

Scaling Limits: Auto scaling is subject to certain limits imposed by AWS. The maximum number of shards you can have in a stream is determined by the shard limit for the region and account.

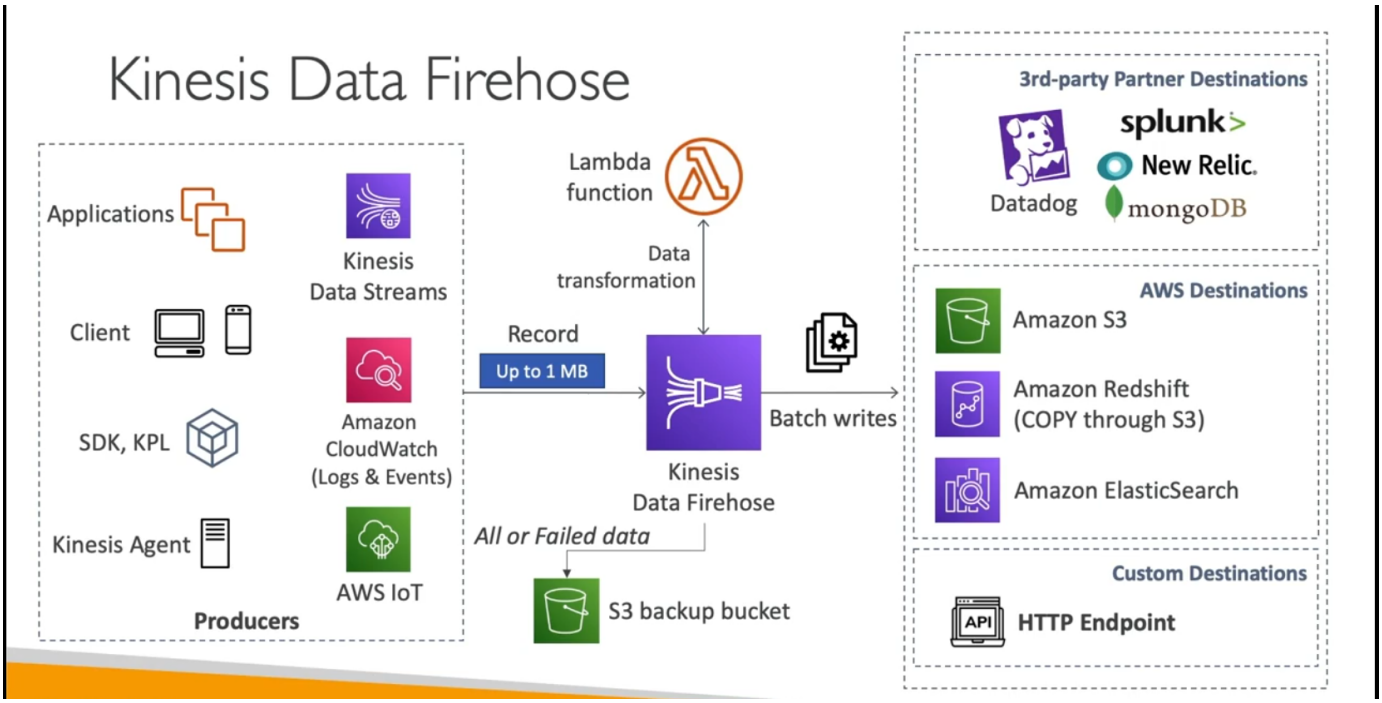
```
#this code requries how many shards you requried

Required Shards = ceil(Estimated Data Rate / Maximum Data Throughput per Shard)
```

kinesis producer put_record partition_key must be unique then no duplicates

```
import boto3
import json
import uuid

# Create a Kinesis client
kinesis_client = boto3.client('kinesis', region_name='us-east-1')
```

```python
# Define the name of the data stream
stream_name = 'your-stream-name'

# Create a sample data record
data = {
    'sensor_id': '001',
    'temperature': 25.5,
    'humidity': 70.2
}

partition_key = str(uuid.uuid4())


# Convert data to JSON
data_json = json.dumps(data)

# Send the data record to the Kinesis stream
response = kinesis_client.put_record(
    StreamName=stream_name,
    Data=data_json,
    PartitionKey=partition-key
)

print('Data sent to Kinesis stream:', response['SequenceNumber'])
```

Kinesis data fire house



above the pic to show how Kinesis data fire house works it

Data Delivery: Kinesis Data Firehose can deliver streaming data to several destinations, including Amazon S3, Amazon Redshift, Amazon Elasticsearch Service, and Splunk. It automatically handles the buffering, compression, and delivery of data to the specified destination.

Data Transformation: You can configure Kinesis Data Firehose to transform the incoming data before delivering it to the destination. It supports data transformation using AWS Lambda functions or Apache Hive-style SQL queries. This allows you to convert data formats, filter data, or enrich it with additional information.

Reliability and Scalability: Kinesis Data Firehose is a fully managed service that takes care of handling incoming data spikes and scaling resources accordingly. It ensures reliable and durable data delivery with built-in retry mechanisms and error handling.

Monitoring and Management: Kinesis Data Firehose integrates with AWS CloudWatch, allowing you to monitor and gain insights into the performance and operational metrics of your delivery streams. It provides visibility into data delivery rates, delivery success, and potential errors.

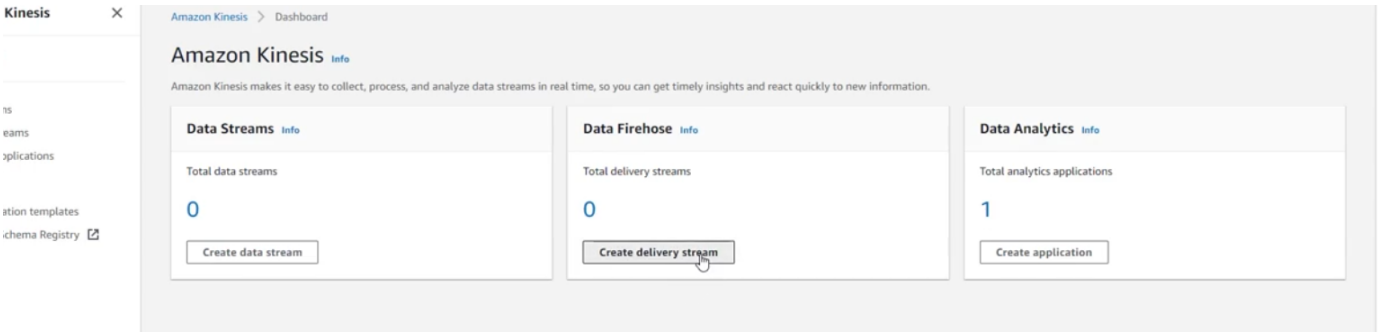
Data Compression and Encryption: Kinesis Data Firehose supports data compression to reduce storage costs and improve network efficiency. It also provides options for encrypting the data at rest using AWS Key Management Service (KMS) for enhanced security.

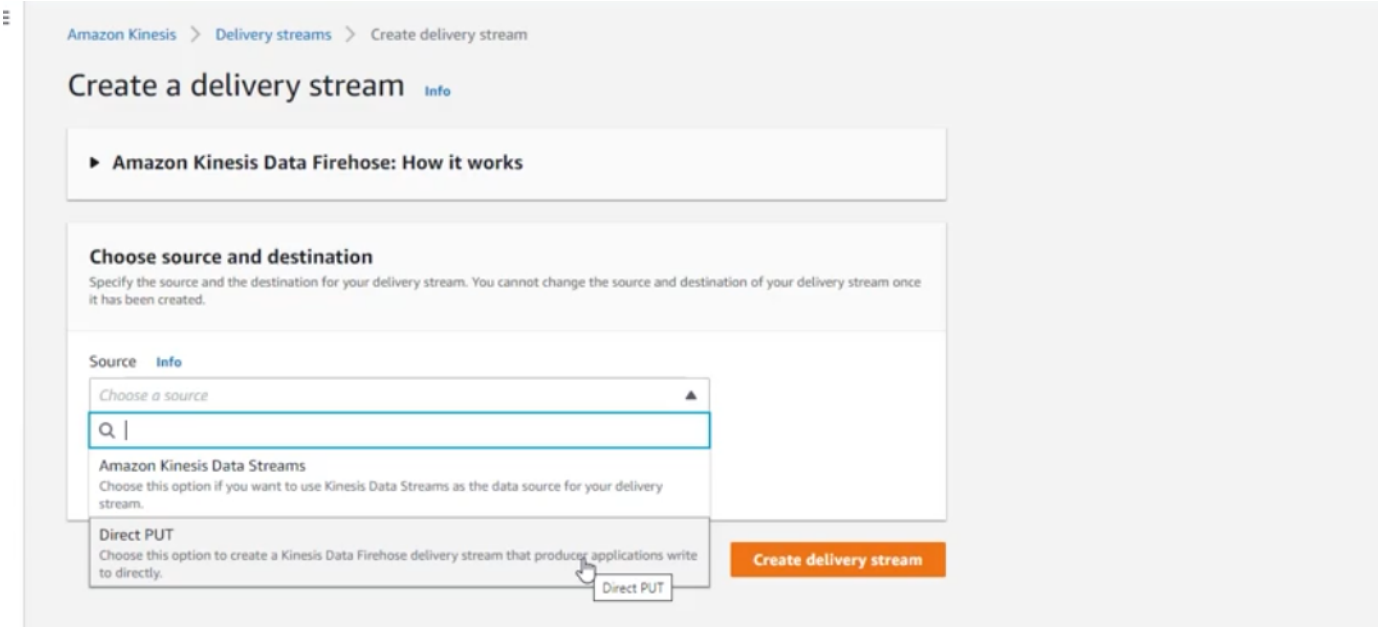
Data Conversion: Kinesis Data Firehose can automatically convert data formats during delivery. For example, it can convert JSON data to Apache Parquet or Apache ORC format for efficient storage and analytics in data lakes.

we can also flesh the data using set buffer size and buffer time

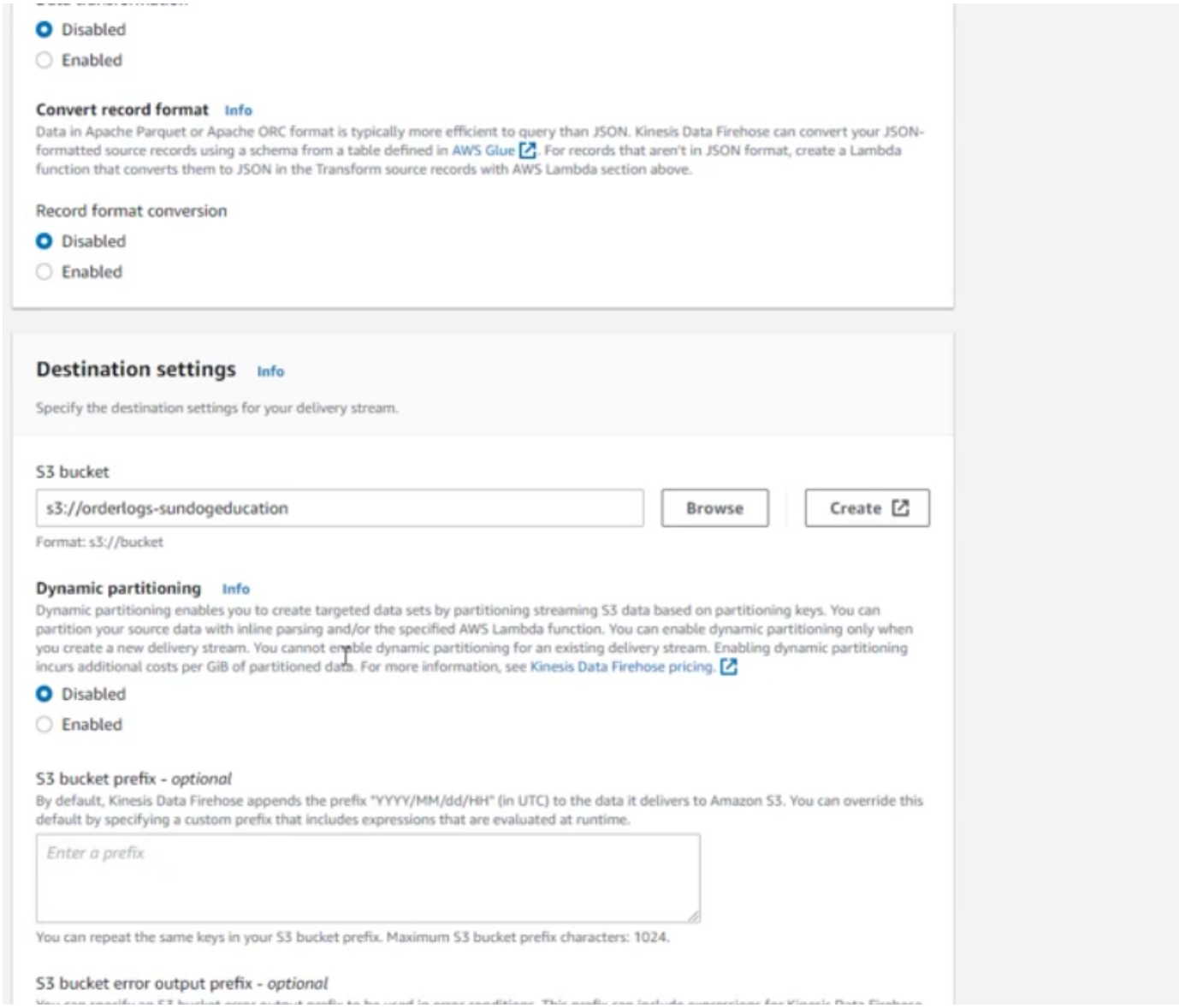
example : Set Buffer Size (ex 32mb) ; set Buffer time (2 minutes)

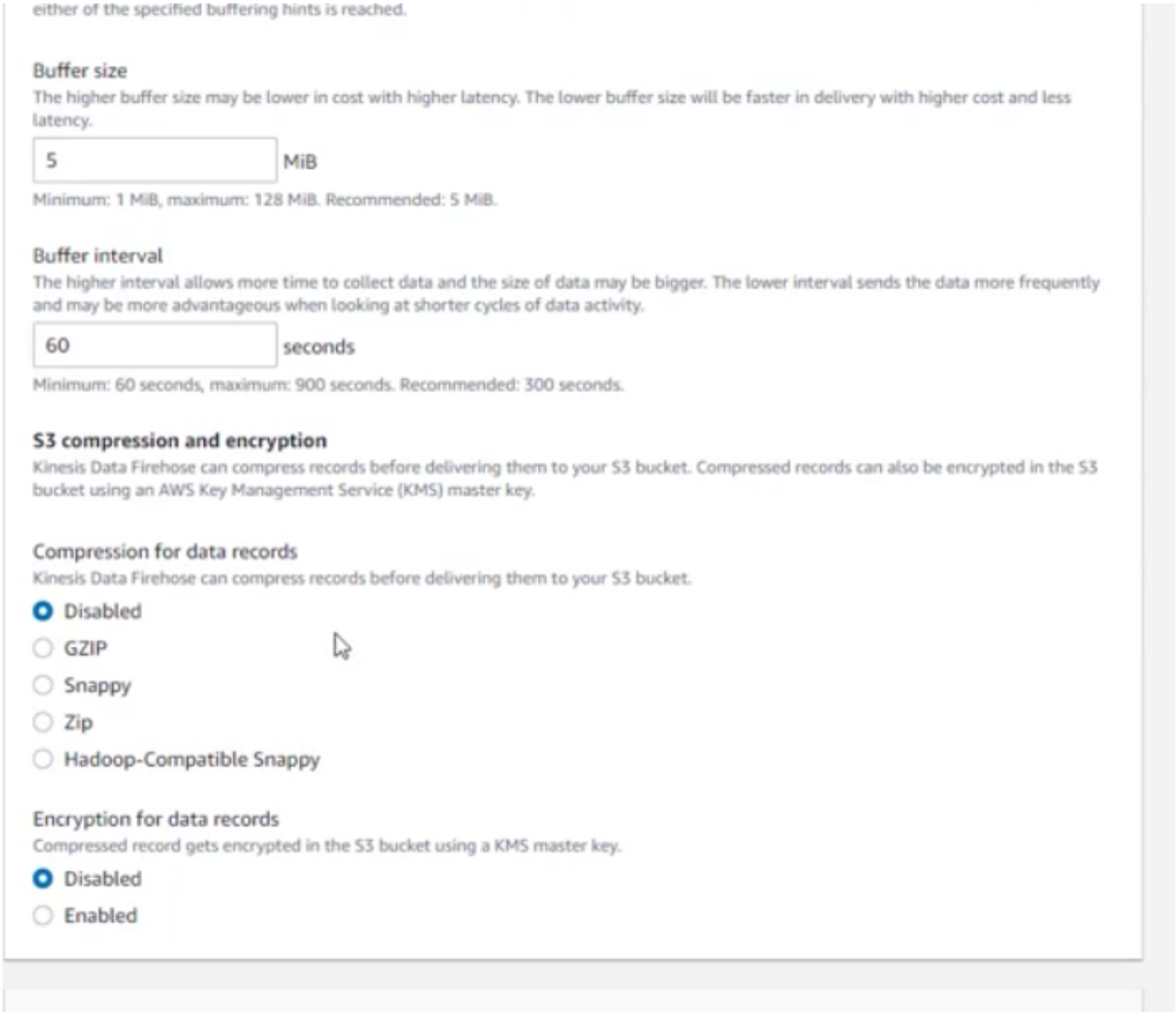
create a fire house real time





there are much more options to set add lambda function to modifies data are directly send data to s3 bucket





same as create a kinesis stream to process the data in real but you need to delete after it completed either it will charged