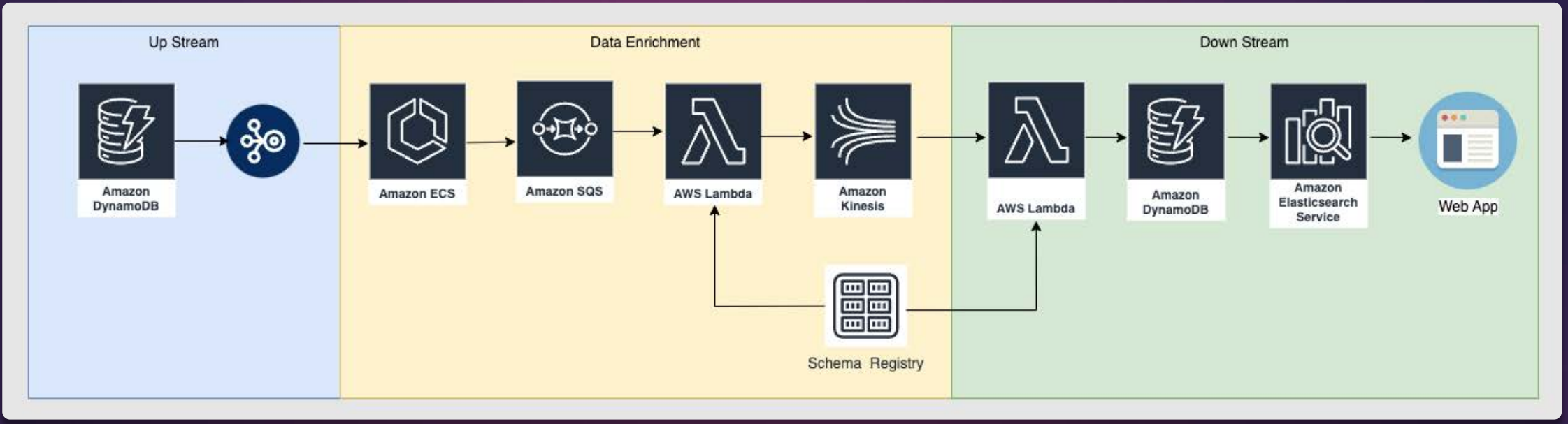
Best Practices - AWS Streaming Data Pipelines

AKSHAY JAIN

Use Case


Up Stream
Data Enrichment
Down Stream
Amazon DynamoDB
Amazon ECS
Amazon SQS
AWS Lambda
Amazon Kinesis
AWS Lambda
Amazon DynamoDB
Amazon Elasticsearch Service
Web App
Schema Registry

Requirements

950M events per day

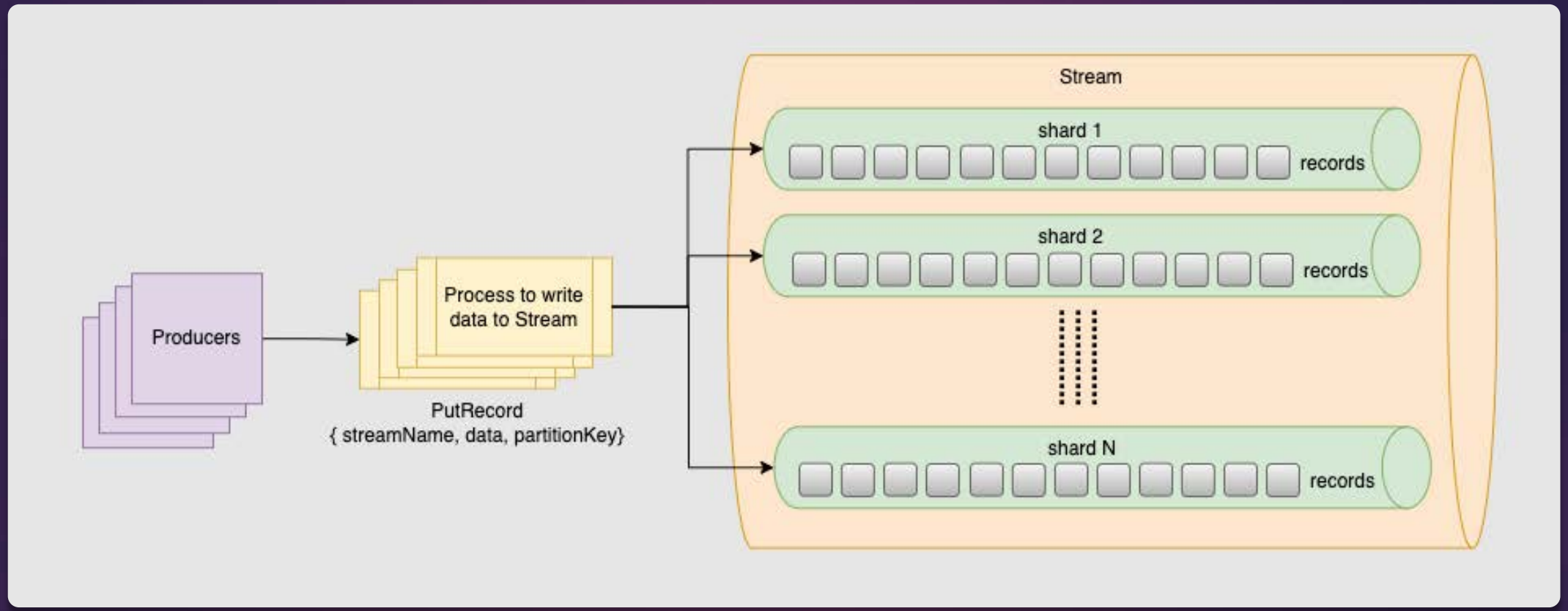
Data Enrichment

Support for schema evolution

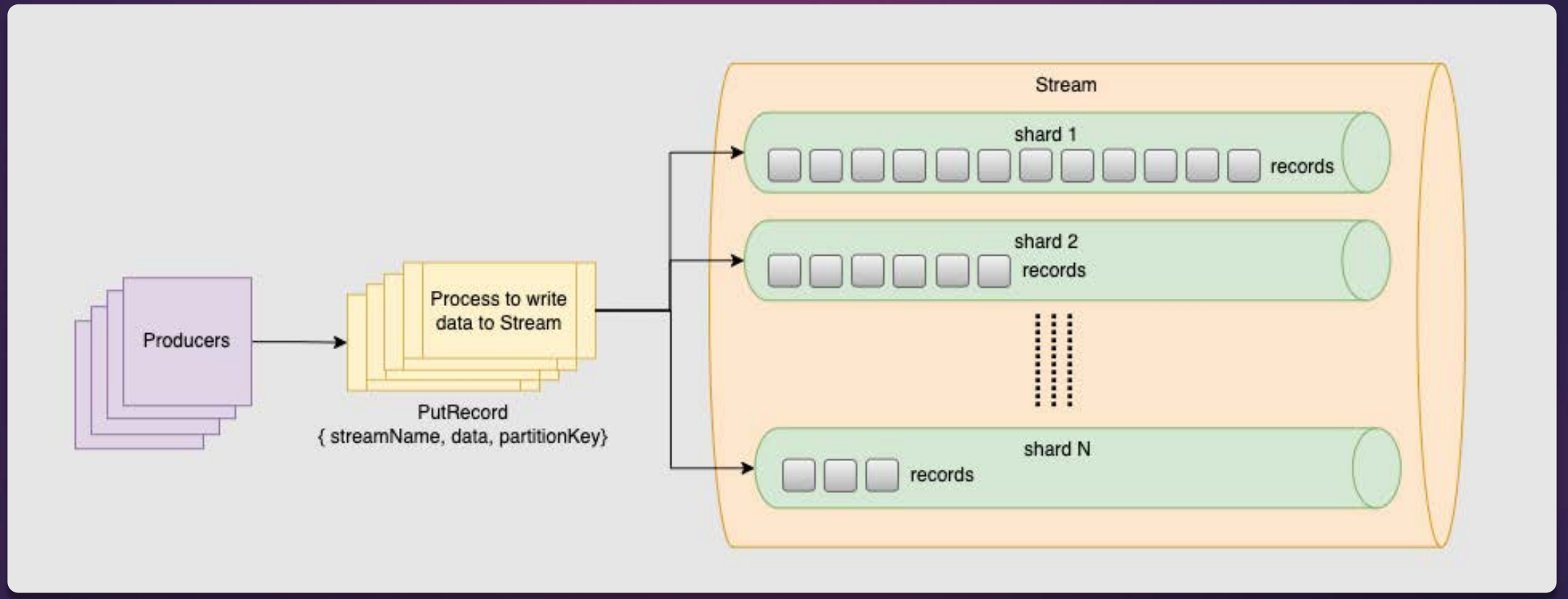
Continuous data validations

Enable logging & lineage tracking

Shard Data Distribution



Shard Data Distribution



How to check shard data distribution ?

- To get shard-level metrics, you will need to use the EnableEnhancedMonitoring API to turn on enhanced granularity for a stream and enable enhanced monitoring for the following metrics:
  - IncomingBytes
  - IncomingRecords
  - OutgoingBytes
  - OutgoingRecords
  - WriteProvisionedThroughputExceeded
  - ReadProvisionedThroughputExceeded
  - IteratorAgeMilliseconds
- On the consumer side, you can use custom logging. For each record batch processed in your IRecordProcessor implementation, you can count the incoming data counts for each shard.
- You can customize producer, and log PutRecordResponses. It returns "your data is placed under XXX shard" for each Put call.

Shard Data Distribution with Partition Key

Assign each record a unique partition key using hash key

```
import json
from boto import kinesis

kinesis_client = kinesis.connect_to_region("YOUR_AWS_REGION")
test_record = {'this': 'is', 'a': 'test'}

kinesis_client.put_record(
  "YOUR_KINESIS_DATA_STREAM",
  json.dumps(test_record), # put_record expects a string
  str(hash(test_record['this']) # partition key
)
```

Shard Data Distribution with Explicit Hash Key

Assign ExplicitHashKey to each record. Use list shards method to get shards information and dump message in the shards based on random distribution.

Process Large Records

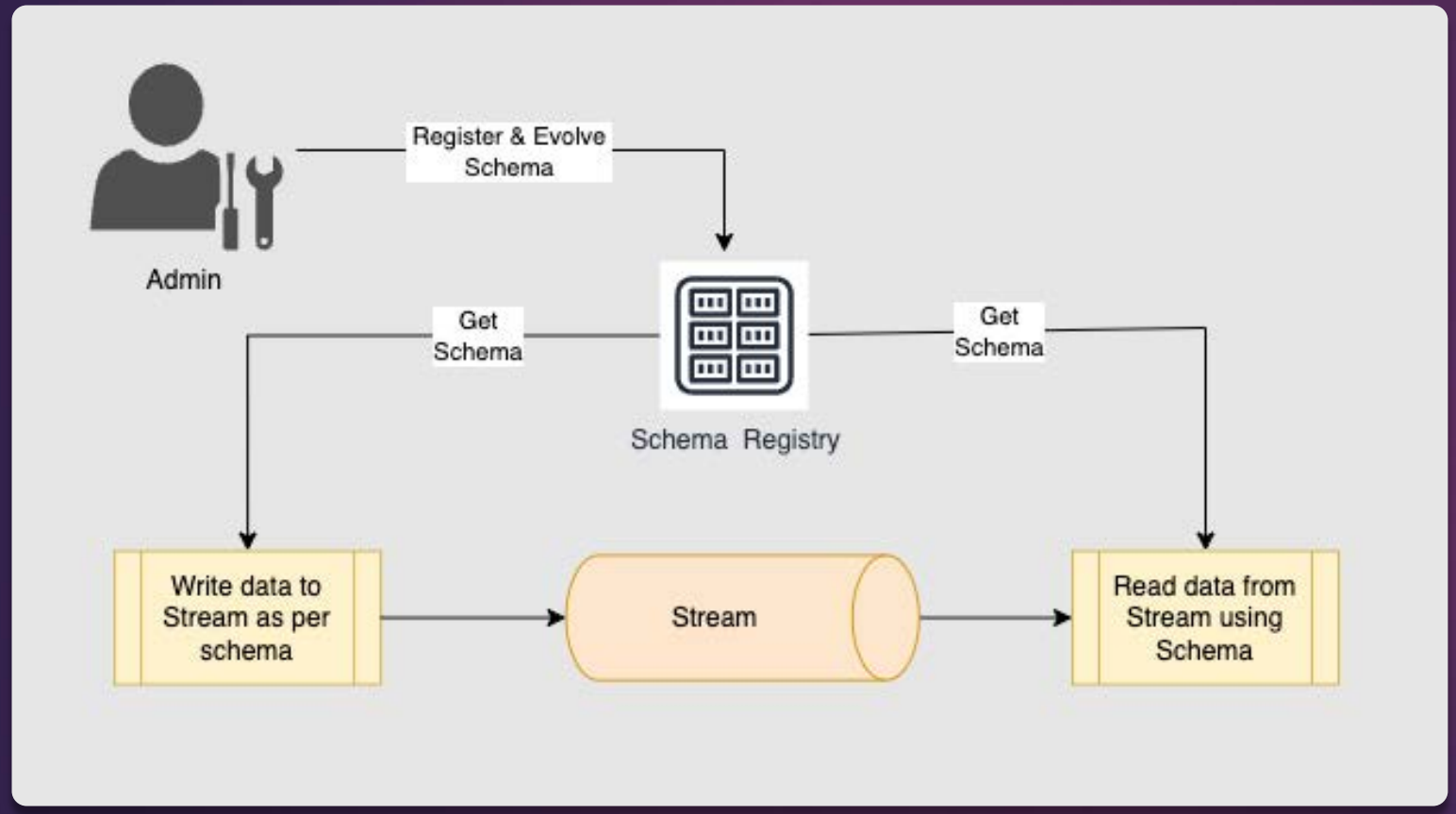
Each shard, which holds data, can handle writing up to 1 MB per second.

Common solutions

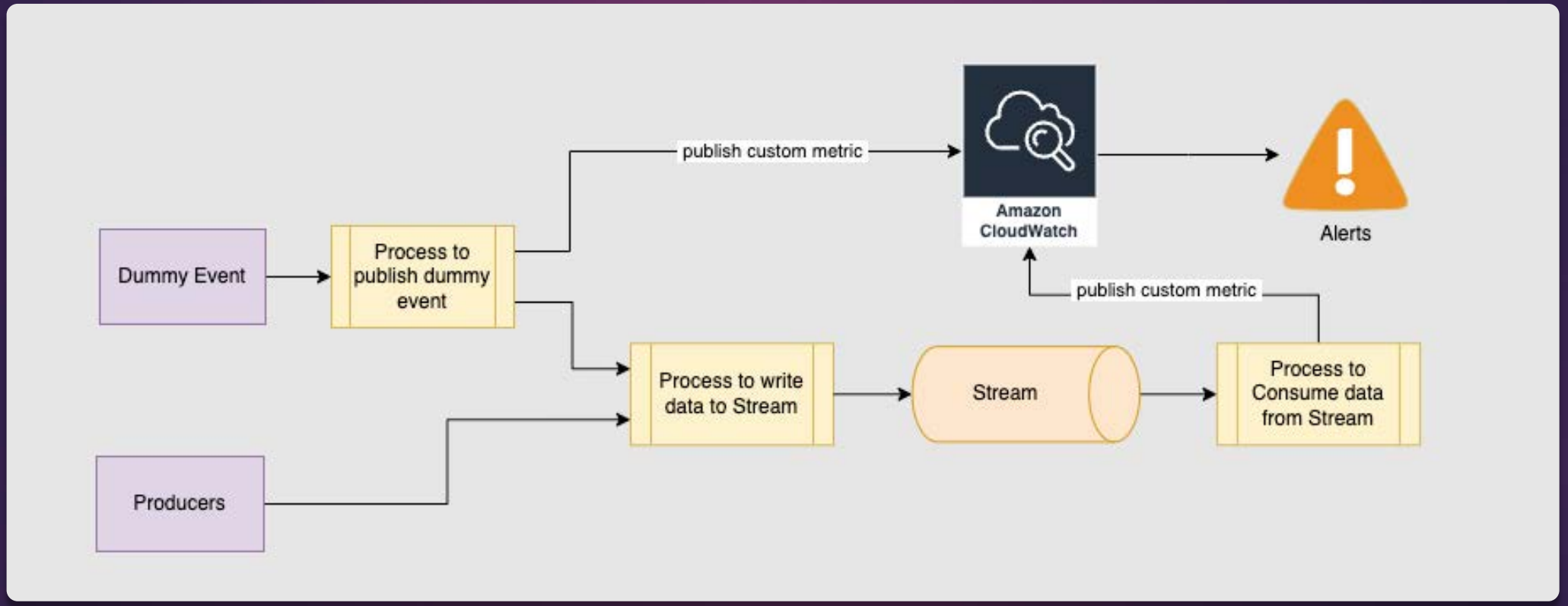
Compress your large records using algorithms such as GZIP, Snappy, LZ4 or choose compressed format such as AVRO

Store large records in Amazon S3 / DB with a reference in Kinesis Data Streams

Schema Registry



Validations



Thank You...!