AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

ANT309

Simplifying Amazon S3 analytics with Amazon Kinesis Data Firehose

Mostafa Mansour
Principal Product Manager, Kinesis Data Firehose
AWS

Radhika Ravirala
Principal Product Manager, Streaming Services
AWS

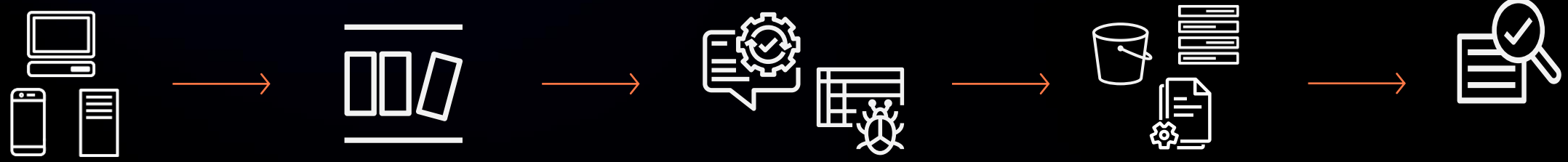
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What to expect in this session

- Empower Amazon Simple Storage Service (Amazon S3) streaming analytics with Amazon Kinesis Data Firehose
- Kinesis Data Firehose dynamic partitioning demo and use cases
- Resources on how to start

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data streaming workflow



Source

Devices and or applications that produce real-time data at high velocity

Stream ingestion

Data from tens of thousands of data sources can be written to a single stream

Stream processing

Records are read in the order they are produced, enabling real-time analytics or streaming ETL

Stream destination

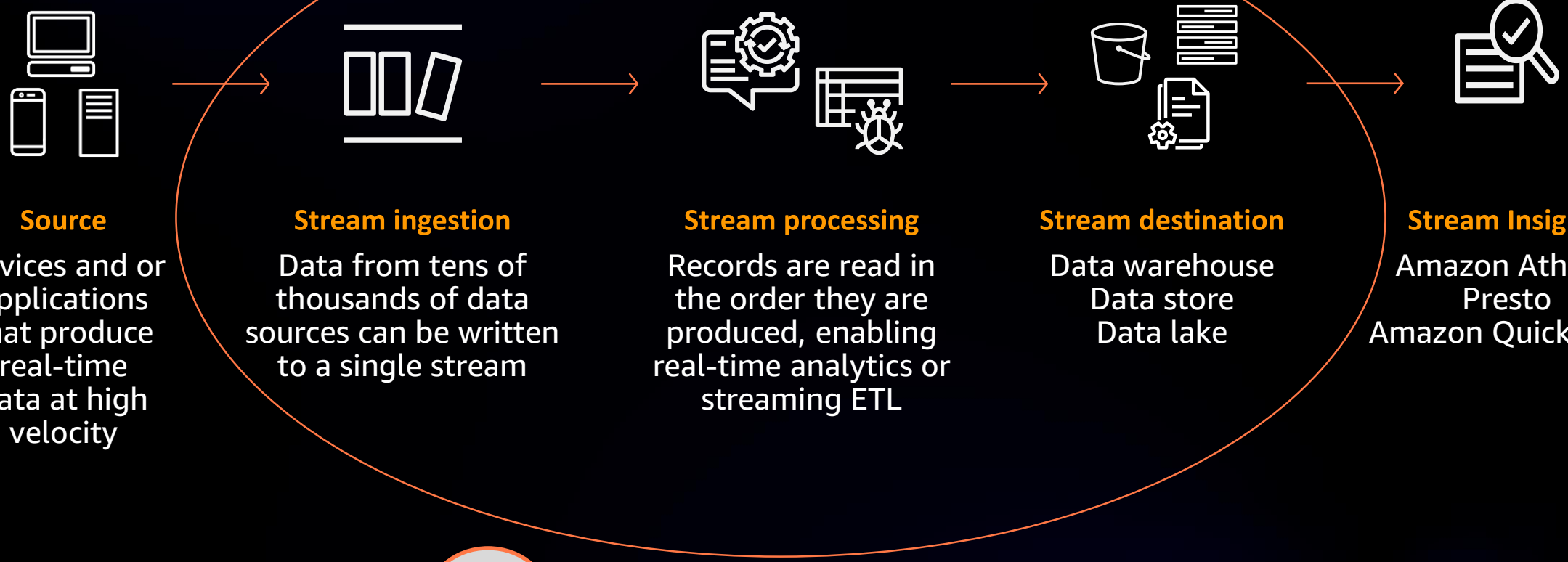
Data warehouse
Data store
Data lake

Stream Insights

Amazon Athena
Presto
Amazon QuickSight

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Data streaming workflow



Source

Devices and or applications that produce real-time data at high velocity

Stream ingestion

Data from tens of thousands of data sources can be written to a single stream

Stream processing

Records are read in the order they are produced, enabling real-time analytics or streaming ETL

Stream destination

Data warehouse
Data store
Data lake

Stream Insights

Amazon Athena
Presto
Amazon QuickSight



Kinesis Data Firehose

Near real-time streaming ETL
For any data, at any scale, and at low-cost

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Kinesis Data Firehose benefits

Easy to use	Elastic scale	Serverless built-in transformation	Integrated with AWS services
With just a few clicks, create your delivery stream with your desired destination	Firehose automatically provisions, manages, and scales resources with no on-going administration required	No code required for data transformation, like Apache Parquet and dynamic partitioning to Amazon S3	30+ fully integrated sources and destinations (for example, Amazon S3, Amazon Redshift, and AWS Partners)

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Kinesis Data Firehose

MANAGED, NEAR REAL-TIME "STREAMING ETL" FOR AMAZON S3

Extract
Sources (20+)

- AWS SDK - Direct Put
- Amazon Kinesis Data Streams
- Amazon Kinesis Agent
- Open-source agents (for example, Fluent Bit)
- 20+ AWS services (for example, Amazon CloudWatch, AWS IoT Greengrass, and so on)

Transform

- Lambda Custom Transformation
- Built-in Transformation
  - Parquet, ORC
  - S3 Dynamic Partitioning
  - Delimiters
  - Compression
  - Encryption
  - Buffer and batch

Load
Outputs (15+)

- Amazon S3
- Amazon OpenSearch Service
- Amazon Redshift
- Amazon API Gateway
- Splunk
- 6+ HTTP endpoint partners (for example, Datadog, MongoDB, and so on)

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Kinesis Data Firehose

MANAGED, NEAR REAL-TIME "STREAMING ETL" FOR AMAZON S3

Extract
Sources (20+)

- AWS SDK - Direct Put
- Amazon Kinesis Data Streams
- Amazon Kinesis Agent
- Open-source agents (for example, Fluent Bit)
- 20+ AWS services (for example, Amazon CloudWatch, AWS IoT Greengrass, and so on)

Transform

- Lambda Custom Transformation
- Built-in Transformation

New

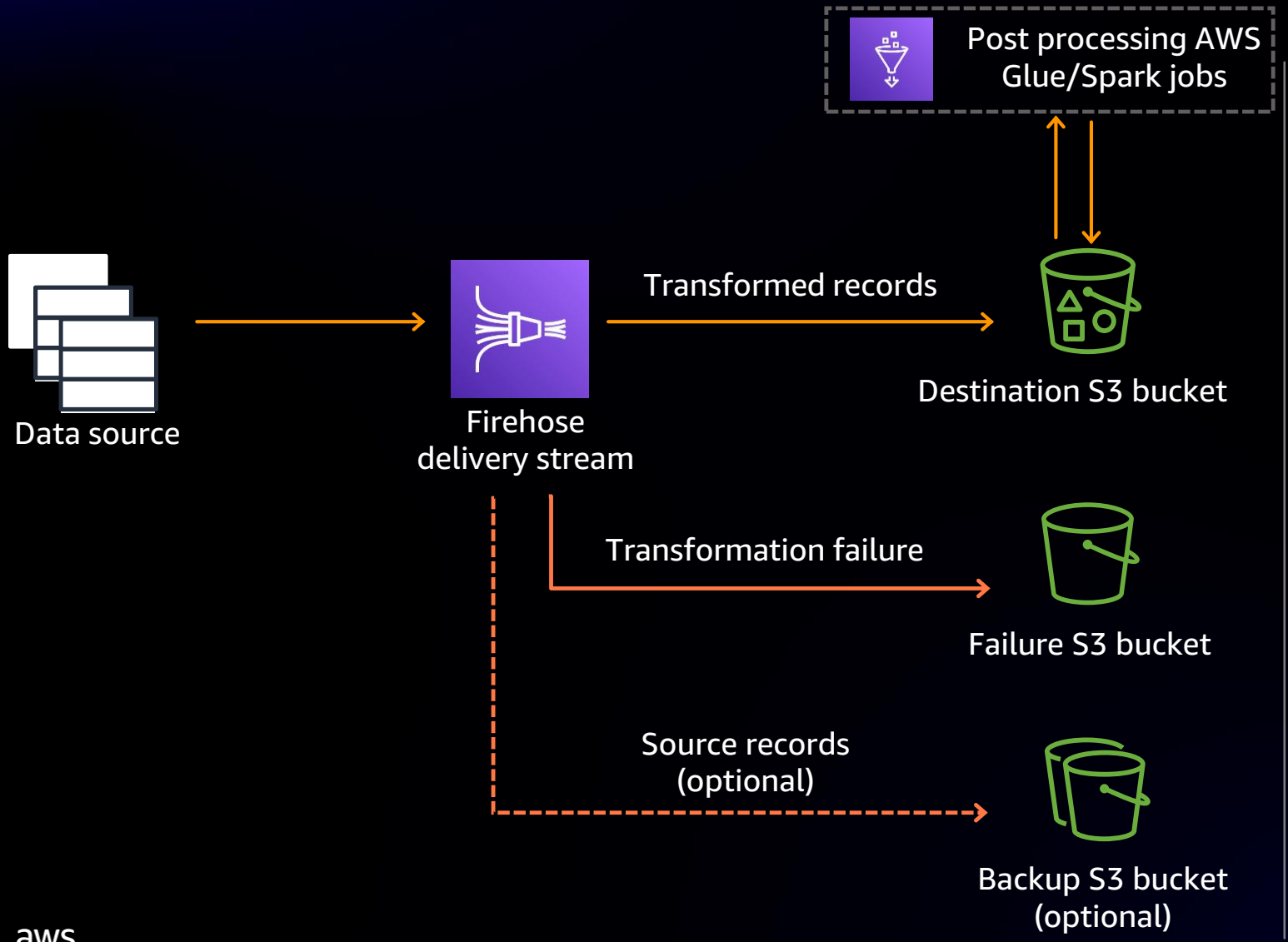
- Parquet, ORC
- S3 Dynamic Partitioning
- Delimiters
- Compression
- Encryption
- Buffer and batch

Load
Outputs (15+)

- Amazon S3
- Amazon OpenSearch Service
- Amazon Redshift
- Amazon API Gateway
- Splunk
- 6+ HTTP endpoint partners (for example, Datadog, MongoDB, and so on)

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Firehose delivery stream Amazon S3 architecture

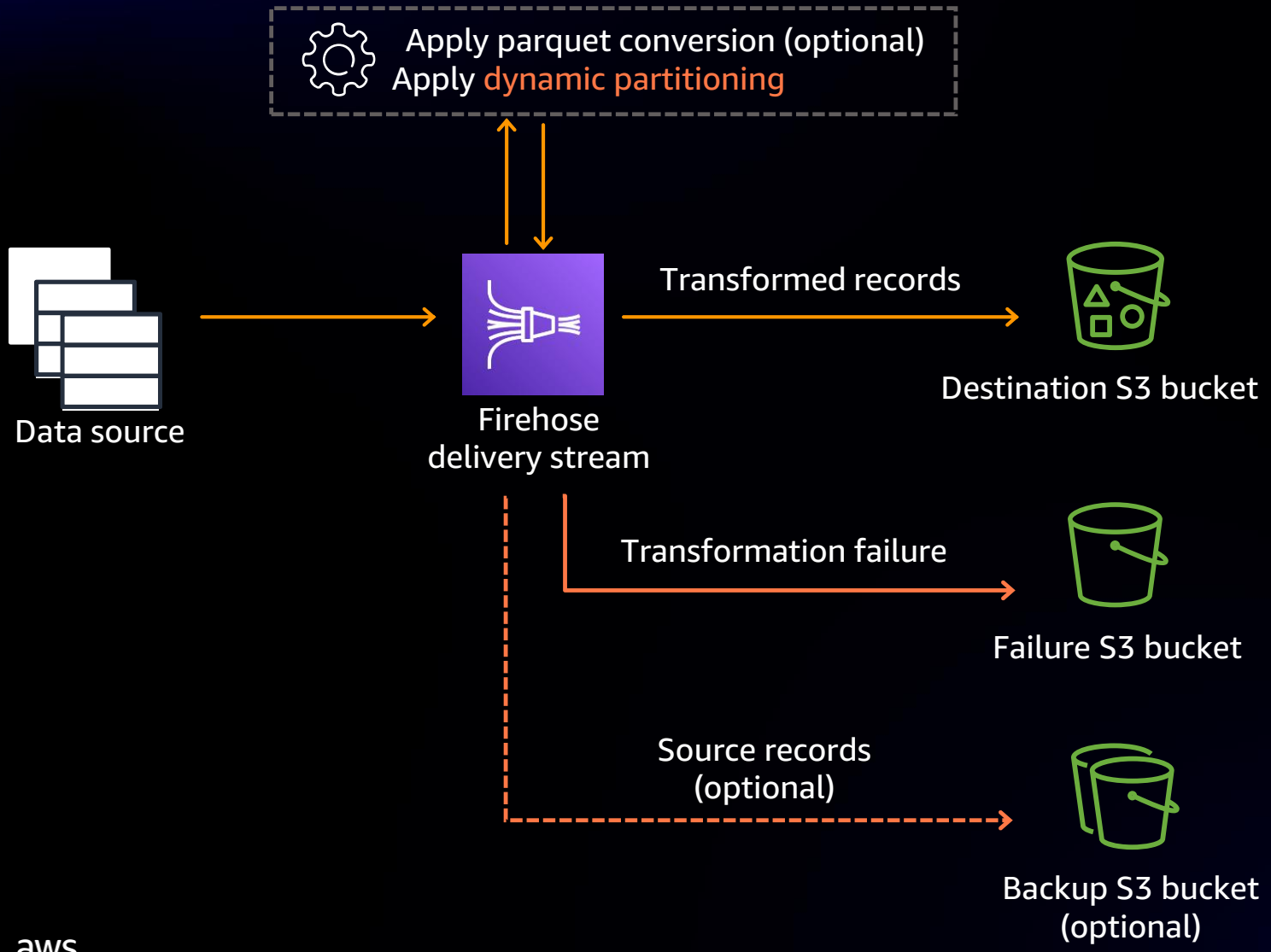


Before dynamic partitioning

- Implement additional ETL jobs using AWS Glue or Amazon EMR to repartition your data after it lands in Amazon S3

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Firehose delivery stream Amazon S3 architecture



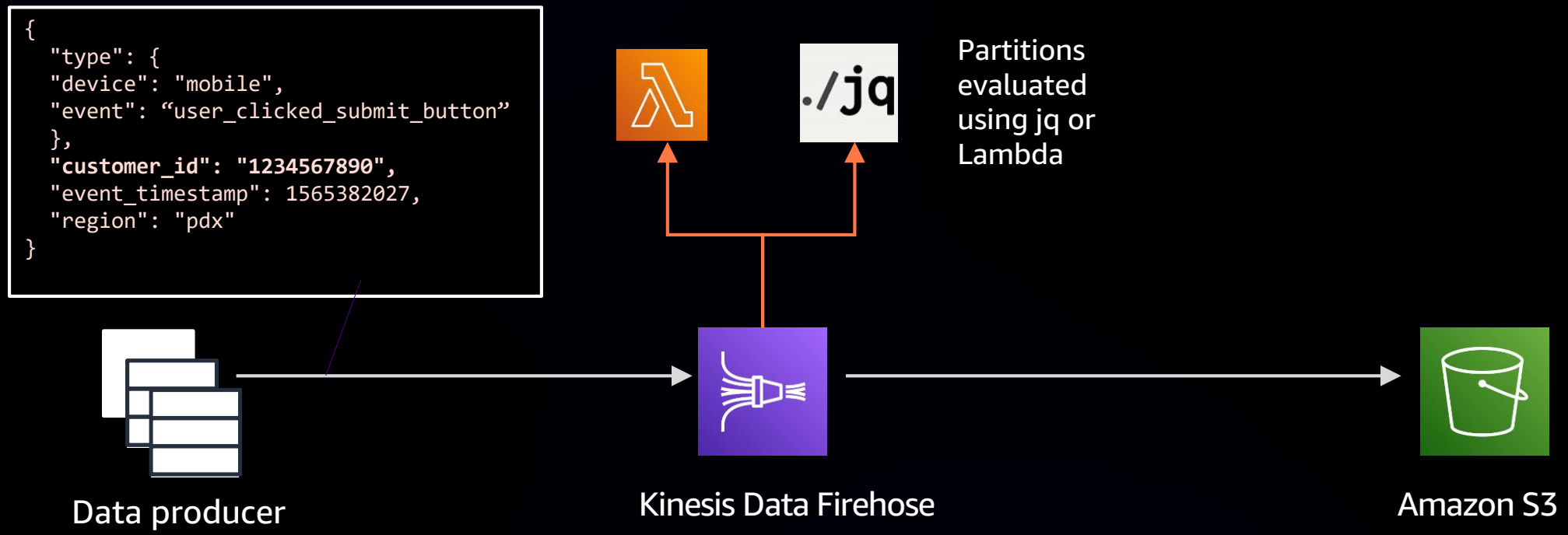
With dynamic partitioning

- Eliminate Amazon S3 post processing needed for Athena, Amazon Redshift Spectrum, Presto, and Snowflake
- In-transit data partitioning based on information parsed from each record
- Organize streaming data into partitions (buckets) on Amazon S3
- Reduce time-to-insight
- Reduce cost of analytics queries

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

“In-transit” streaming data partitioning

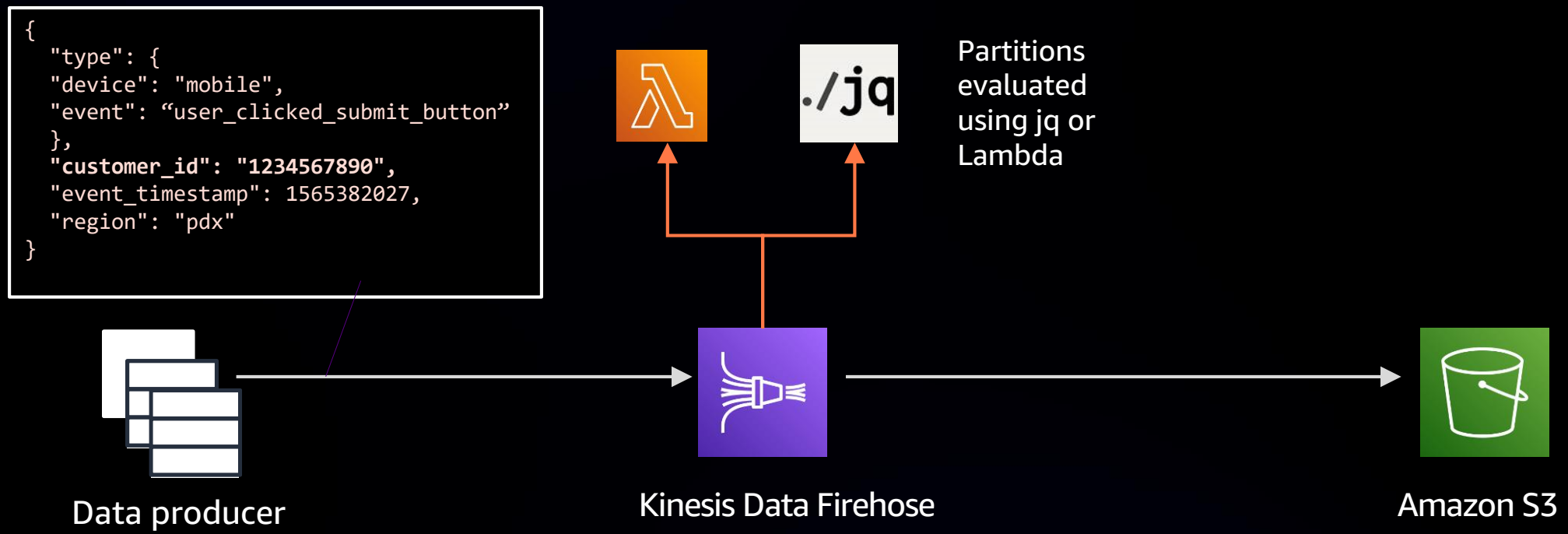
JQ EXAMPLE



© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

“In-transit” streaming data partitioning

JQ EXAMPLE



1

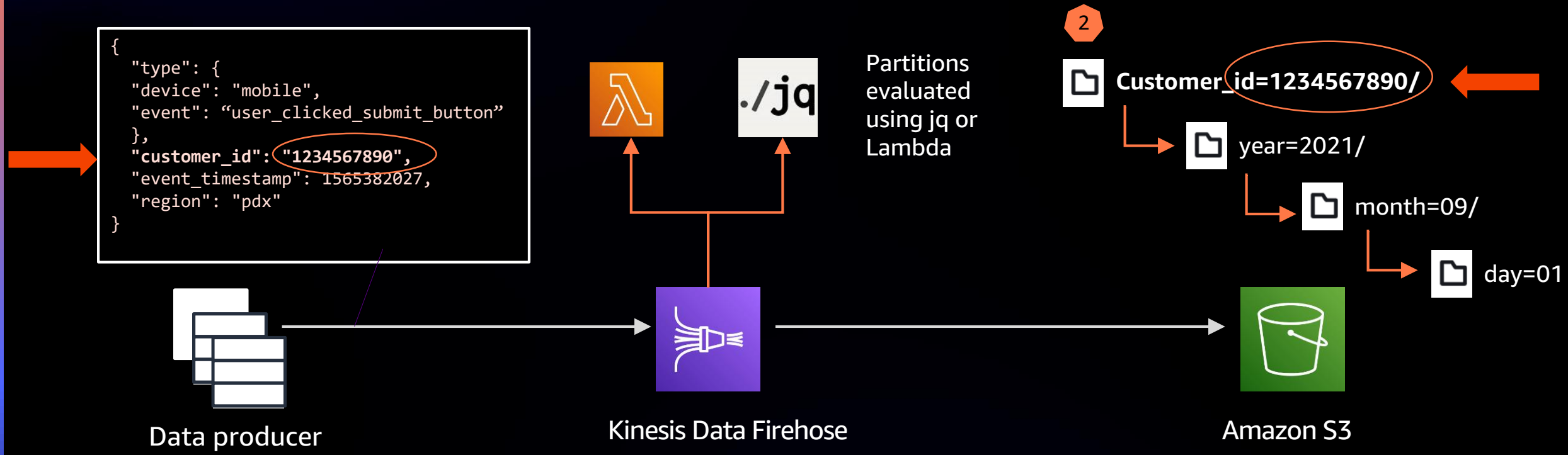
./jq

Parameter	JQ
customer_id	.customer_id
year	.event_timestamp | strftime(“%Y”)
month	.event_timestamp | strftime(“%m”)
day	.event_timestamp | strftime(“%d”)

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

“In-transit” streaming data partitioning

JQ EXAMPLE



1

./jq

Parameter	JQ
customer_id	.customer_id
year	.event_timestamp | strftime(“%Y”)
month	.event_timestamp | strftime(“%m”)
day	.event_timestamp | strftime(“%d”)

aws

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Dynamic partitioning examples

Transaction ID or customer ID	→ Incoming data of e-commerce transactions or points-of-sale
CDC schema or table ID	→ Preparing data to be sent from Amazon S3 to Amazon Redshift (or any warehouse solution) based on the right table ID or right schema
Type of device	→ Incoming IoT data with different devices: Android, iOS, Fire TV, and so on
Log type	→ Incoming data of Amazon CloudWatch Logs: log info, log error, log warning, and so on
API version	→ Incoming data is generated from two API versions; each version should go to a different Amazon S3 partition: Version 1, Version 2

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

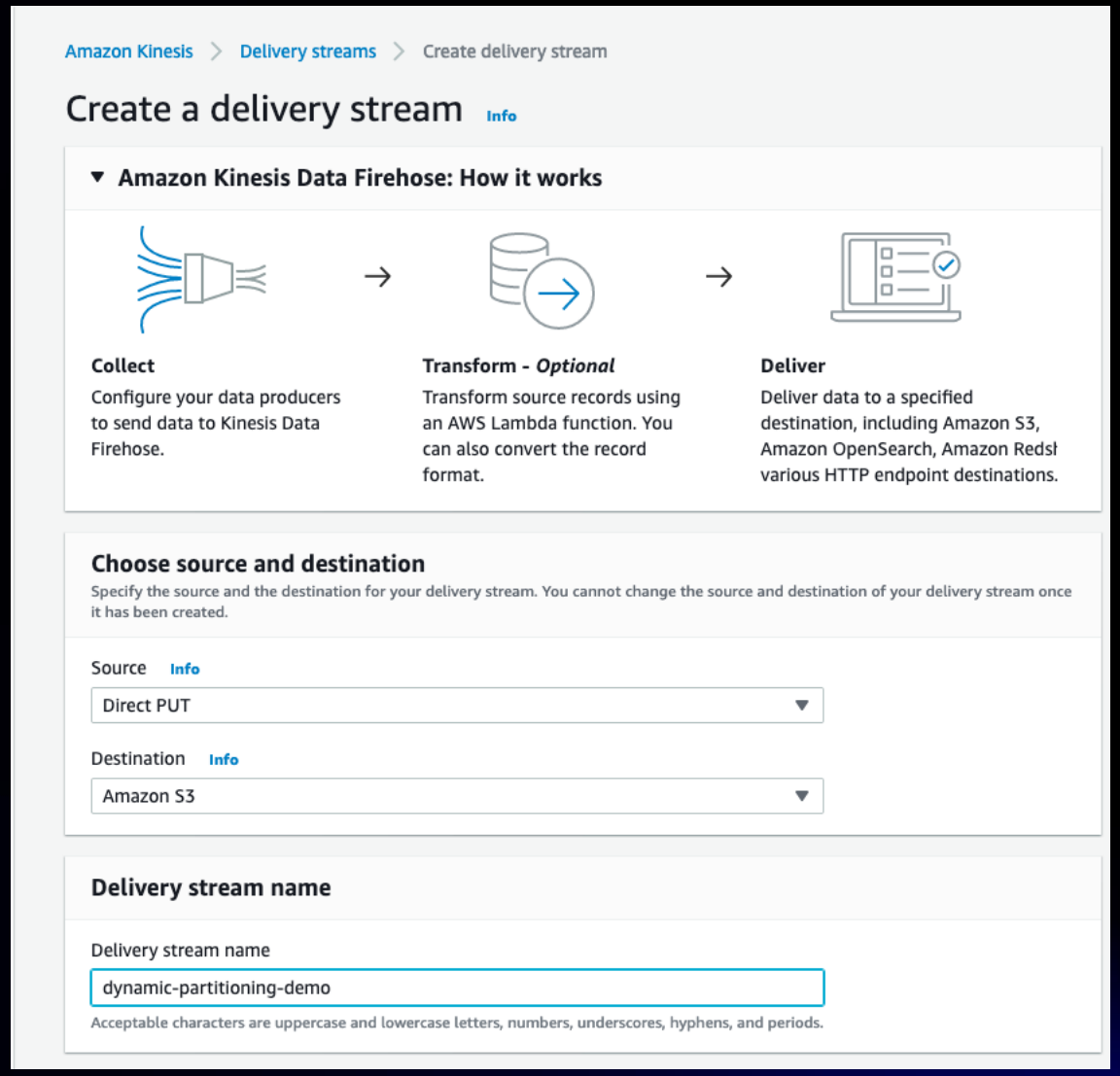
Console walkthrough

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Dynamic partitioning: Flow

1

Create a Firehose Delivery Stream

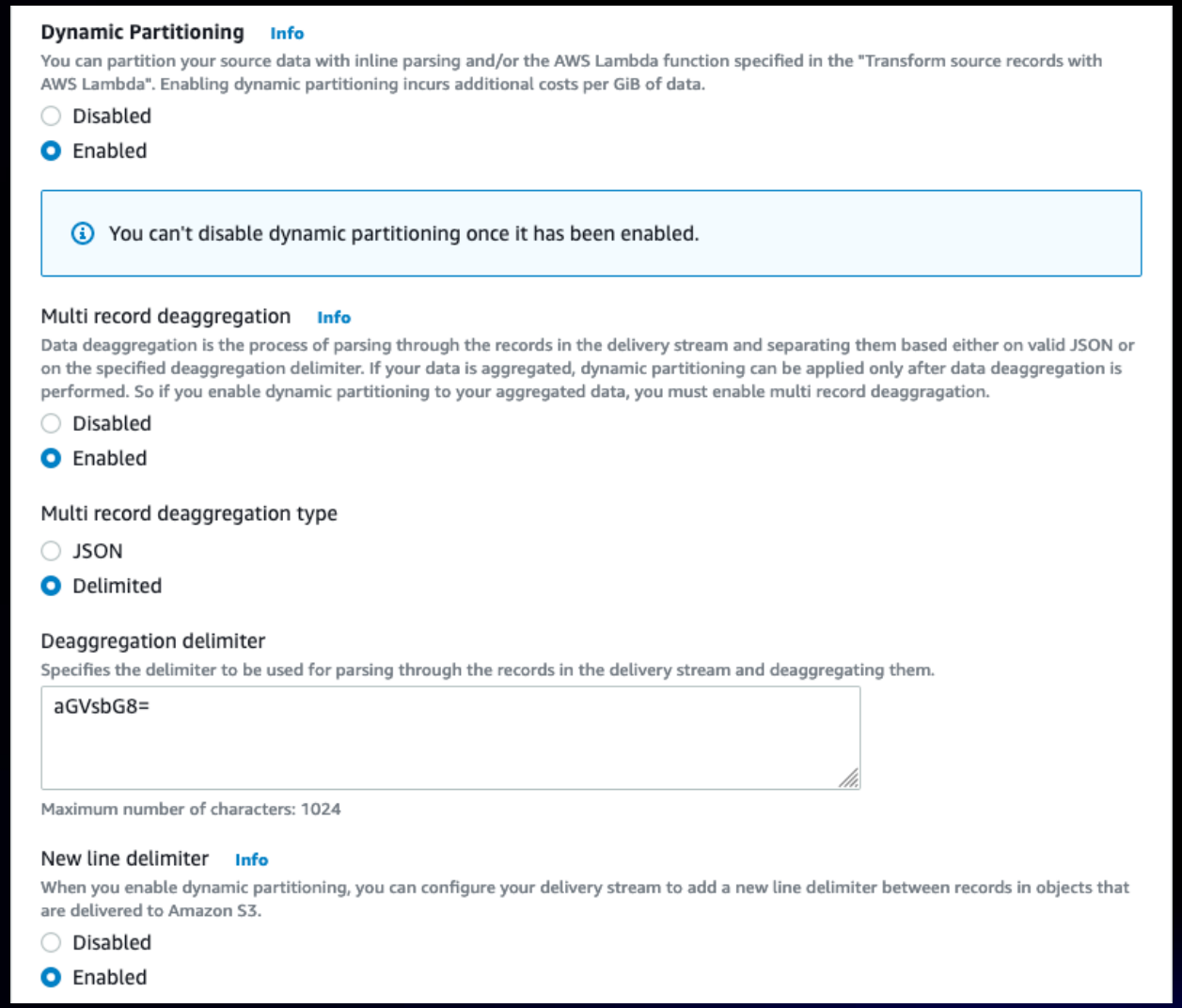


© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Dynamic partitioning: Flow

2

```
Enable Dynamic Partitioning
Enable Multi record de-aggregation
-  New line
-  Json (default)
-  Custom Delimiter
```



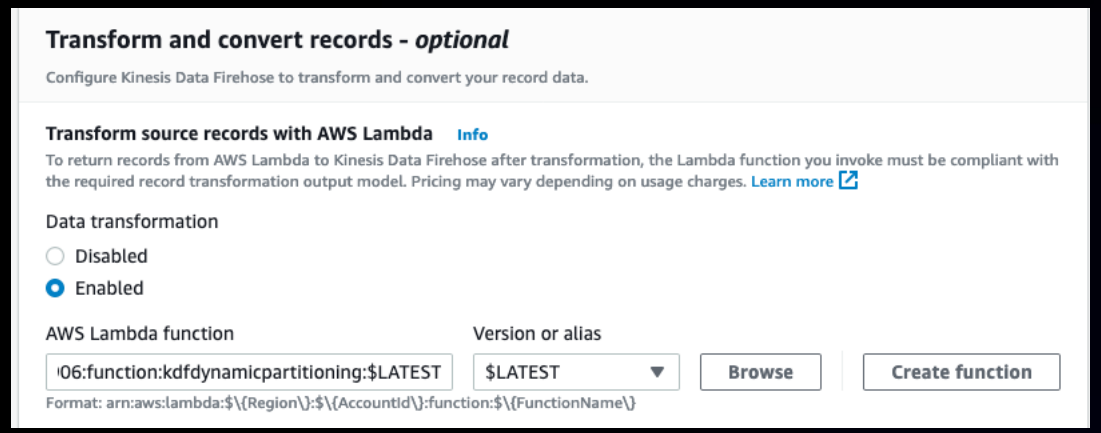
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

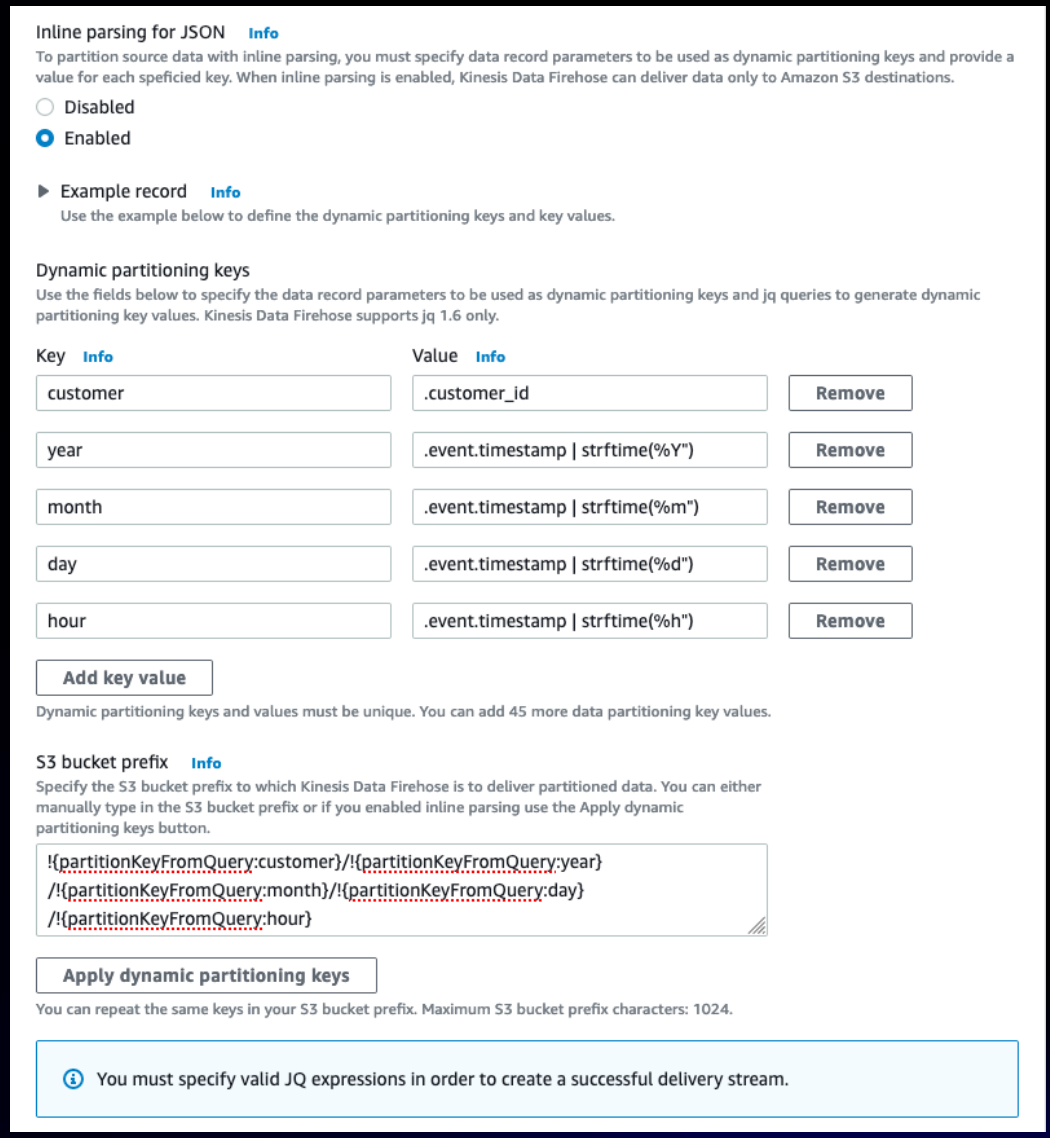
Dynamic partitioning: Flow

3

Build prefix using JQ only

Or build prefix using Lambda





© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Resources

- Kinesis Data Firehose workshop and dynamic partitioning tutorial
  - https://go.aws/3xeZiUX
- Kinesis Data Firehose dynamic partitioning blogpost
  - https://go.aws/3r04YB2
- Kinesis Data Firehose dynamic partitioning documentation
  - https://go.aws/3r7IQEP
- Kinesis Data Firehose dynamic partitioning video
  - https://bit.ly/3FELF4i

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.