Amazon Kinesis and Amazon MSK Overview

Analyzing real-time data streams on AWS

Robert Ray
Solutions Architect

August 21, 2019

Agenda

Recap

Amazon Kinesis overview

Amazon MSK overview

Amazon Kinesis vs. Amazon MSK

Who are real-time analytics candidates?

Companies that have a high volume, velocity and variety of data and:

- Analyze and react in real time to events affecting business
- Ingest continuous data into a data lake for near-real-time analysis

Common real-time analytics use cases

Messaging between micro-services

Response analytics (Web and mobile app notifications)

Log ingestion

IoT device maintenance

Change Data Capture (CDC)

Streaming ETL into data lakes and data warehouses

Challenges of data streaming

Difficult to setup

Tricky to scale

Hard to achieve high availability

Integration requires development

Error prone and complex to manage

Expensive to maintain

aws

Streaming real-time data with AWS

Easily collect, process and analyze data streams in real time

- Easy to use
- Elastic
- High availability and durability
- Seamless integration with AWS services
- Fully managed
- Pay for what you use

Enabling real-time analytics

Data streaming technology enables a customer to ingest, process and analyze high volumes of high velocity data from a variety of sources in real time

Streaming data with AWS

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

Amazon
Kinesis Data
Streams

Capture and store data streams

Amazon
Kinesis Data
Analytics

Analyze data streams in real time

Amazon
Kinesis Data
Firehose

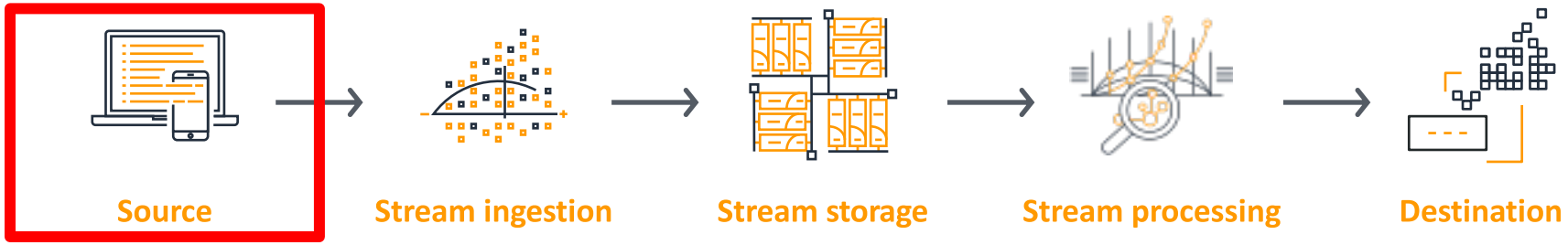
ETL streaming into streams, data lakes and warehouses

Amazon Managed
Streaming for Kafka

Capture and store data streams

Enabling real-time analytics

Data streaming technology enables a customer to ingest, process and analyze high volumes of high velocity data from a variety of sources in real time



Source

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

Mobile Apps

Web Clickstream

```
[Wed Oct 11 14:32:52
2018] [error] [client
127.0.0.1] client
denied by server
configuration:
/export/home/live/ap/h
tdocs/test
```

Application Logs

Metering Records

IoT Sensors

Smart Buildings

Enabling real-time analytics

Data streaming technology enables a customer to ingest, process and analyze high volumes of high velocity data from a variety of sources in real time

Stream Ingestion

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

AWS Toolkits/Libraries

- AWS SDK
- Kinesis Producer Library
- AWS Mobile SDK (iOS, fire)
- Kinesis Agent

AWS Service Integrations

- AWS IoT
- Amazon CloudWatch Logs
- Amazon CloudWatch Events
- Amazon Database Migration Service*

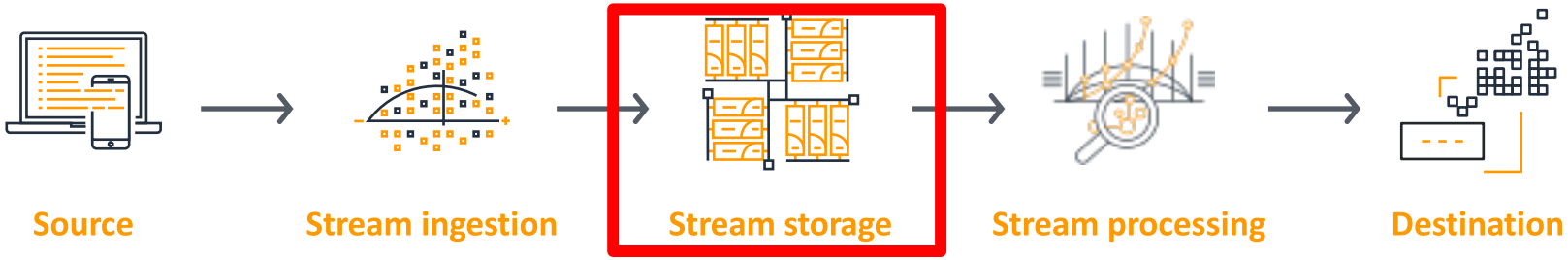
3rd Party Offerings

- LOG4J
- Flume
- Fluentd

* Amazon DMS includes 8 on-premise databases, 1 Azure database, 5 RDS/Aurora database types, and S3

Enabling real-time analytics

Data streaming technology enables a customer to ingest, process and analyze high volumes of high velocity data from a variety of sources in real time



Stream Storage

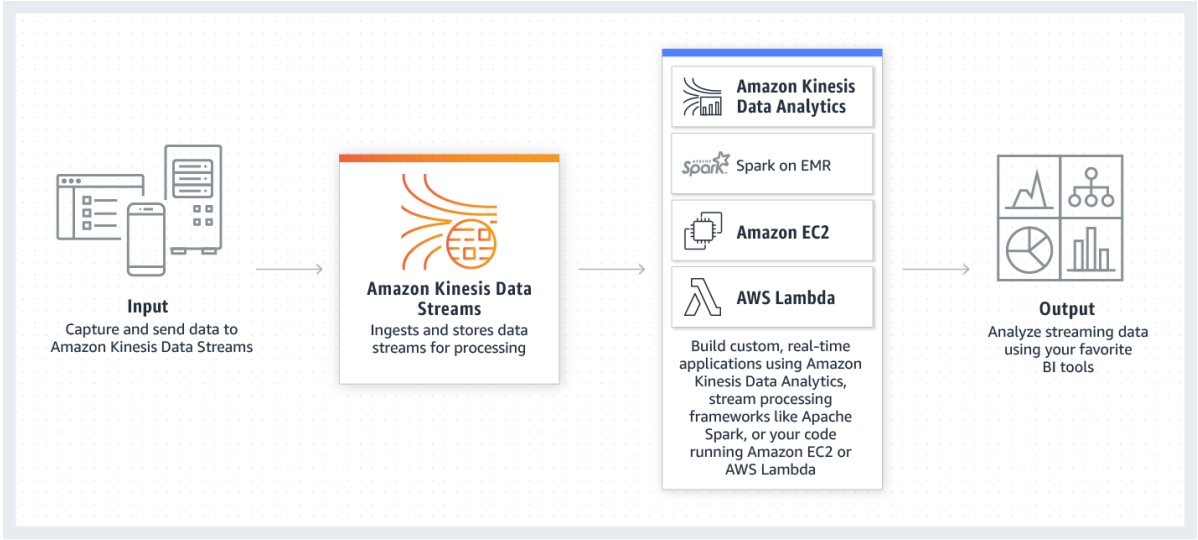
Data is stored in the order it was received for a set duration of time, and can be replayed indefinitely during this time.

Amazon Kinesis Data Streams

Amazon Managed Streaming for Kafka

aws

Amazon Kinesis Data Streams



- Easy administration and low cost
- Real-time, elastic performance
- Secure, durable storage
- Available to multiple real-time analytics applications

Easy to Use

- Create a new stream in seconds
- Set desired level of capacity with shards
- Set up elastic scaling to match data throughput rate & volume

aws

Enabling real-time analytics

Data streaming technology enables a customer to ingest, process and analyze high volumes of high velocity data from a variety of sources in real time

Stream Processing

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

Kinesis

Amazon Kinesis Data Analytics

SQL

Flink

Kinesis Client Library
+
Connector Library

AWS Services

AWS Lambda

Amazon EMR

3rd party

Apache Spark

Apache Storm

Datadog

mongoDB

Anodot

databricks

Qubole

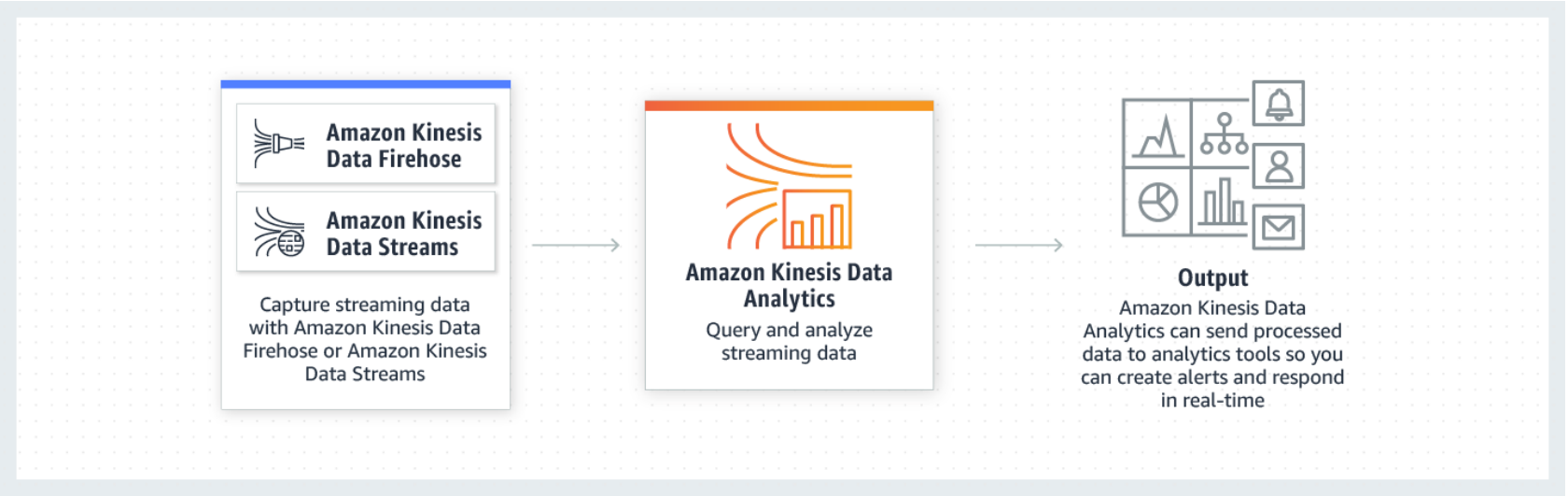
VoltDB

MemSQL

splunk>

aws

Amazon Kinesis Data Analytics



- Interact with streaming data in real-time using SQL or integrated Java applications
- Build fully managed and elastic stream processing applications

Kinesis Data Analytics

Connect to streaming source

Easily write SQL or Java code to process streaming data

Continuously deliver results

KDA for SQL for simple and fast use cases

- Sub-second end to end processing latencies
- SQL steps can be chained together in serial or parallel steps
- Build applications with one or hundreds of queries
- Pre-built functions include everything from sum and count distinct to machine learning algorithms
- Aggregations run continuously using window operators

aws

KDA for Java for sophisticated applications

Utilizes Apache Flink, a Framework and distributed engine for stateful processing of data streams

Simple programming

Easy to use and flexible APIs make building apps fast

High performance

In-memory computing provides low latency & high throughput

Stateful Processing

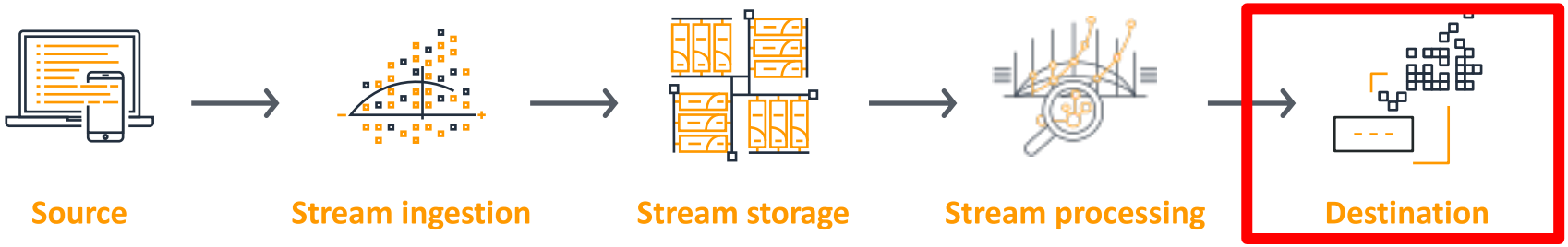
Durable application state saves

Strong data integrity

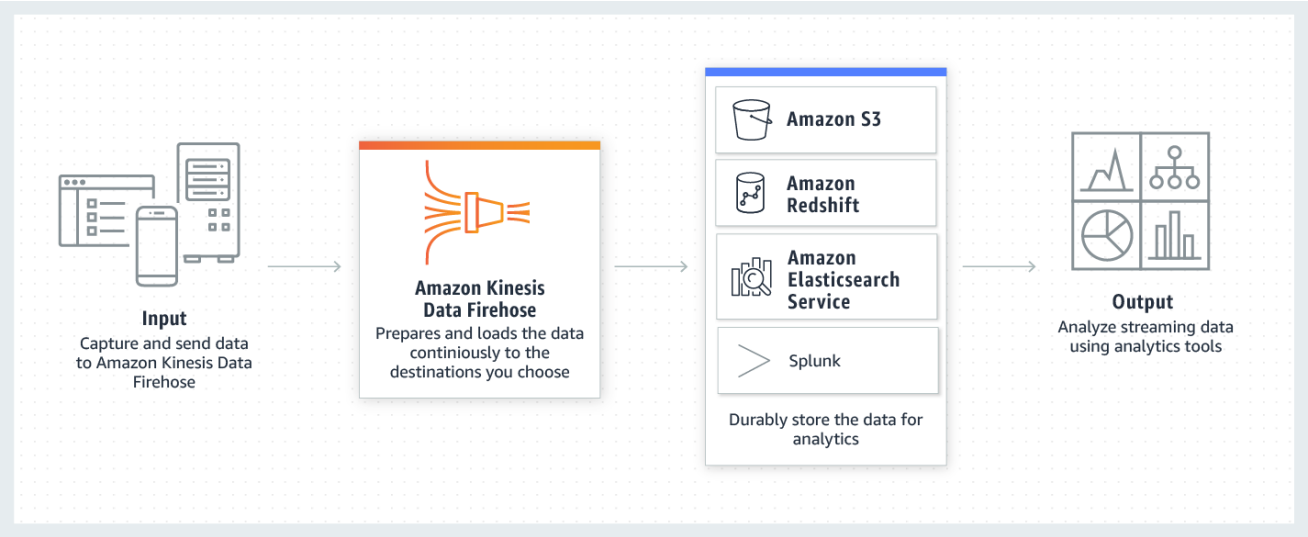
Exactly-once processing and consistent state

Enabling real-time analytics

Data streaming technology enables a customer to ingest, process and analyze high volumes of high velocity data from a variety of sources in real time



Amazon Kinesis Data Firehose

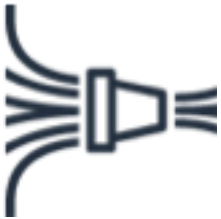


- Zero administration and seamless elasticity
- Direct-to-data store integration
- Serverless continuous data transformations
- Near real-time

Amazon Kinesis - Firehose vs. Streams

Kinesis Data Streams

Amazon Kinesis Data Streams is for use cases that require custom processing, per incoming record, with sub-1 second processing latency, and a choice of stream processing frameworks



Kinesis Data Firehose

Amazon Kinesis Data Firehose is for use cases that require zero administration, ability to use existing analytics tools based on Amazon S3, Amazon Redshift, and Amazon ES, and a data latency of 60 seconds or higher

An Example Architecture

Source | Stream ingestion | Stream storage | Stream processing | Destination

Mobile device

Metering

Click streams

IoT sensors

Logs

[Wed Oct 11 14:32:52 2018] [error] [client 127.0.0.1] client denied by server configuration: /export/home/live/ap/htdocs/test

Stream Ingestion

Amazon Kinesis Agent

Amazon Kinesis Producer Library

AWS SDKs

Amazon Kinesis Data Streams

Real-Time Applications (seconds)

Amazon Kinesis Data Analytics

Amazon Kinesis Consumer Library

Streaming ETL (minutes)

Amazon Kinesis Data Firehose

Amazon S3
Amazon Redshift
Amazon Elasticsearch
Splunk

aws

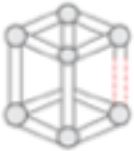
Apache Kafka

Challenges operating Apache Kafka

Difficult to setup

Tricky to scale

Hard to achieve high availability



AWS integrations = development

No console, no visible metrics

$$f(kafka_{usage}) = \sum_{n=1}^{\infty}(SRE)$$

aws

Apache Managed Streaming for Kafka

aws

Getting started with Amazon MSK is easy

- Fully compatible with Apache Kafka v1.1.1 and v2.1.0
- AWS Management Console and AWS API for provisioning
    - **Clusters are setup automatically**
    - **Provision Apache Kafka brokers and storage**
    - **Create and tear down clusters on-demand**
- Deeply integrated with AWS services

aws

Amazon Managed Streaming for Kafka

Comparing Amazon MSK with Amazon Kinesis Data Streams

aws

Comparing Amazon Kinesis Data Streams to MSK

Amazon Kinesis Data Streams

- Streams and shards
- AWS API experience
- Throughput provisioning model
- Seamless scaling
- Typically lower costs
- Deep AWS integrations

Amazon MSK

- Topics and partitions
- Open-source compatibility
- Strong third-party tooling
- Cluster provisioning model
- Apache Kafka scaling isn't seamless to clients
- Raw performance

aws

Using Kafka with Kinesis

Download the Kafka-Kinesis Connector Library