Amazon Kinesis and Amazon MSK Overview
Analyzing real-time data streams on AWS

Sven Hansen
Solutions Architect – AWS
svenhans@amazon.com
Agenda

Why real-time analytics and data streaming?
Amazon Kinesis overview
Amazon MSK overview
Demo
Data is produced from a large variety of sources

- Mobile apps
- Web clickstream
- Application logs
- Metering records
- IoT sensors
- Smart buildings
The value of data diminishes over time

Source: Perishable insights, Mike Gualtieri, Forrester
Enabling real-time analytics: Considerations

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.
Challenges of data streaming

- Difficult to setup
- Hard to achieve high availability
- Error prone and complex to manage
- Tricky to scale
- Integration requires development
- Expensive to maintain
# Streaming real-time data with AWS

Easily collect, process and analyze data streams in real time

<table>
<thead>
<tr>
<th>Easy to use</th>
<th>Elastic</th>
</tr>
</thead>
<tbody>
<tr>
<td>High availability and durability</td>
<td>Seamless integration with AWS services</td>
</tr>
<tr>
<td>Fully managed</td>
<td>Pay for what you use</td>
</tr>
</tbody>
</table>
Streaming data with AWS
Easily collect, process, and analyze data streams in real time

- Capture and store data streams
- Analyze data streams in real time
- ETL streaming into streams, data lakes and warehouses
- 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**.
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
- 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 \textit{in real time}. 

Source $\rightarrow$ Stream ingestion $\rightarrow$ Stream storage $\rightarrow$ Stream processing $\rightarrow$ Destination
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

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

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

AWS Services
- AWS Lambda
- Amazon EMR

3rd party
- Apache Spark
- Anodot
- Datadog
- MongoDB
- Quibole
- Databricks
- VoltDB
- MemSQL
- Splunk

© 2019, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
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
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
An Example Architecture

- **Source**
  - Mobile device
  - Metering
  - Click streams
  - IoT sensors
  - Logs

- **Stream Ingestion**
  - Amazon Kinesis Agent
  - Amazon Kinesis Producer Library
  - AWS SDKs

- **Stream Storage**
  - Amazon Kinesis Data Streams

- **Stream Processing**
  - Real-Time Applications (seconds)
  - Streaming ETL (minutes)
  - Amazon Kinesis Data Analytics
  - Amazon Kinesis Consumer Library
  - Amazon S3
  - Amazon Redshift
  - Amazon Elasticsearch
  - Splunk

- **Destination**
Apache Managed Streaming for Kafka
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
Demo
Questions?