Amazon Kinesis Immersion Day

Harsha Tadiparthi, Specialist DB & Analytics Solutions Architect
David Bayard, Specialist DB & Analytics Solutions Architect

Agenda

Amazon Kinesis Overview

Kinesis Capabilities

- Kinesis Data Streams & Lab
- Kinesis Data Firehose & Lab
- Kinesis Data Analytics & Lab

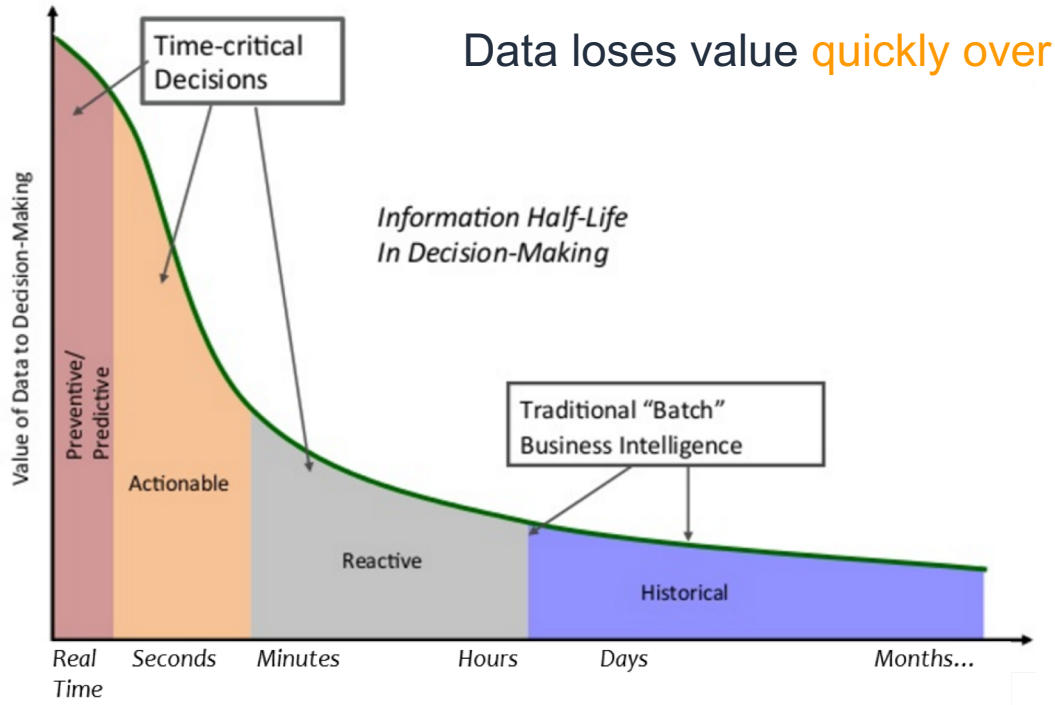
Customer Examples

aws

Lab Preparation - Housekeeping

- Send email to David Bayard dbayard@amazon.com in next 5 minutes

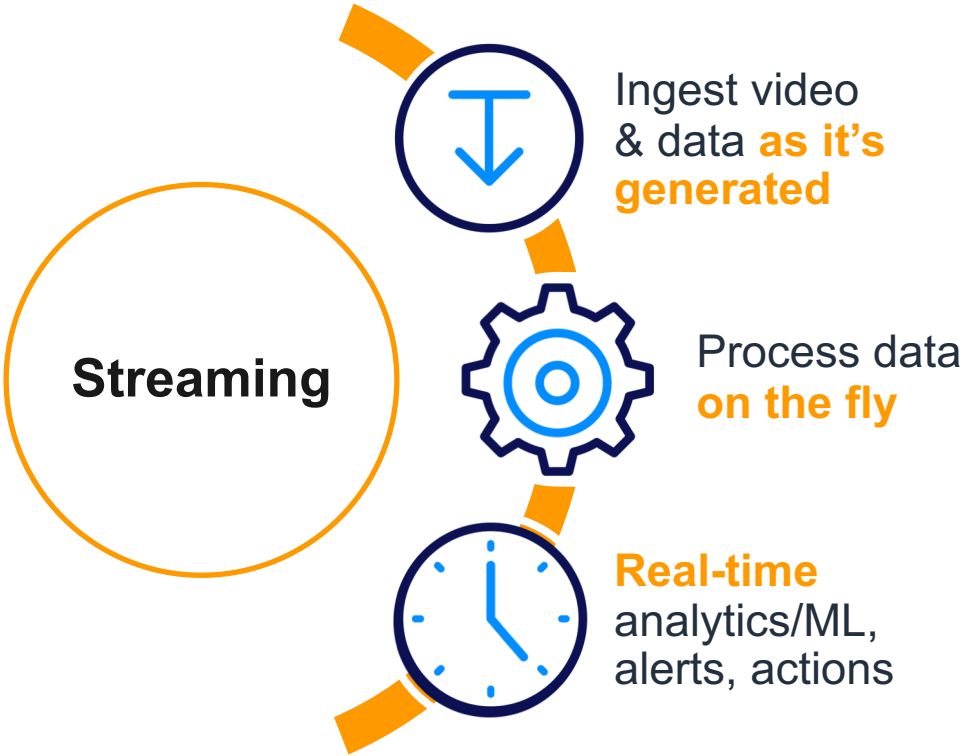
Timely Decisions Require New Data in Minutes



Source: Perishable insights, Mike Gualtieri, Forrester

Stream New Data in Seconds

Get actionable insights quickly

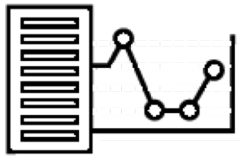


aws

Most Common Uses of Streaming

Security Monitoring

Industrial Automation



Log Analytics

Data Lakes

IoT Device Monitoring

Streaming with Amazon Kinesis

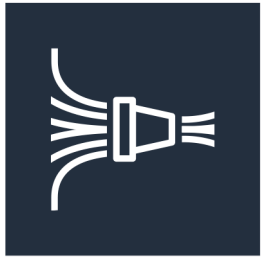
Easily collect, process, and analyze video and data streams in real time

Amazon Kinesis Video Streams

Capture, process, and store video streams

Amazon Kinesis Data Streams

Capture, process, and store data streams



Amazon Kinesis Data Firehose

Load data streams into AWS data stores

Amazon Kinesis Data Analytics

Analyze data streams in real-time

aws

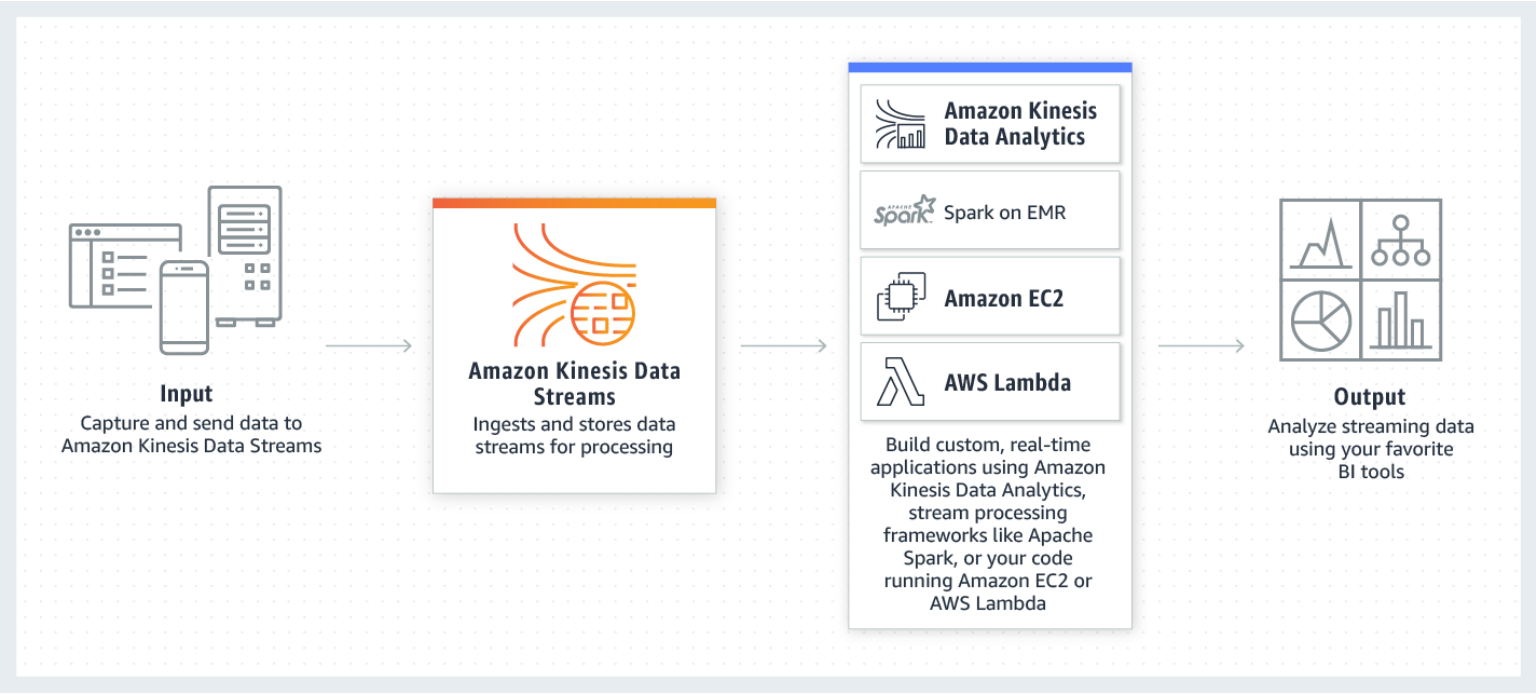
Amazon Kinesis

Data Streams

Deep Dive

aws

Amazon Kinesis Data Streams



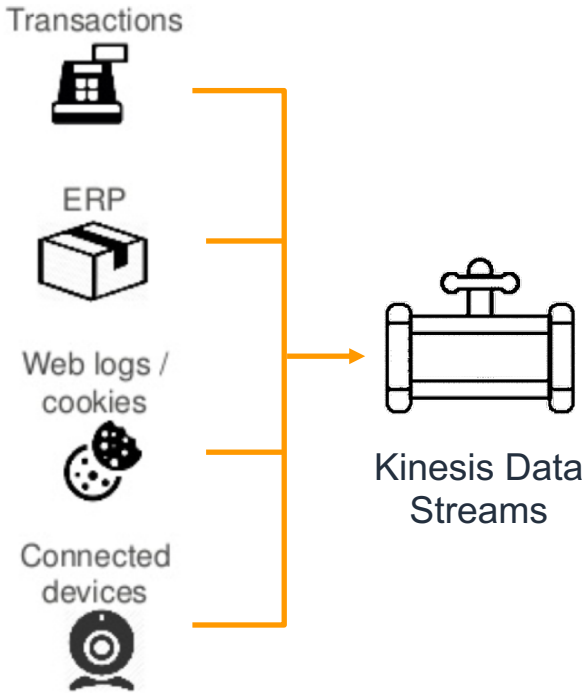
Key Benefits

- **Easy administration:** Simply create a new stream, and set the desired level of capacity with shards. Scale to match your data throughput rate and volume

- Perform continual processing on streaming big data using Kinesis Client Library (KCL), Kinesis Analytics, Apache Spark/Storm, AWS Lambda, and more

- **Low cost:** Cost-efficient for workloads of any scale

Data Ingestion From a Variety of Sources



AWS SDKs
- Publish directly from application code via APIs
- AWS Mobile SDK
- Managed AWS sources: CloudWatch Logs, AWS IoT, Kinesis Data Analytics and more
- RDS Aurora via Lambda

Kinesis Agent
- Monitors log files and forwards lines as messages to Kinesis Data Streams

Kinesis Producer Library (KPL)
- Background process aggregates and batches messages

3rd party and open source
- Log4j appender
- Apache Kafka
- Flume, fluentd, and more…

aws

Supports Wide Range of Processing Tools

Kinesis Client Library (KCL)

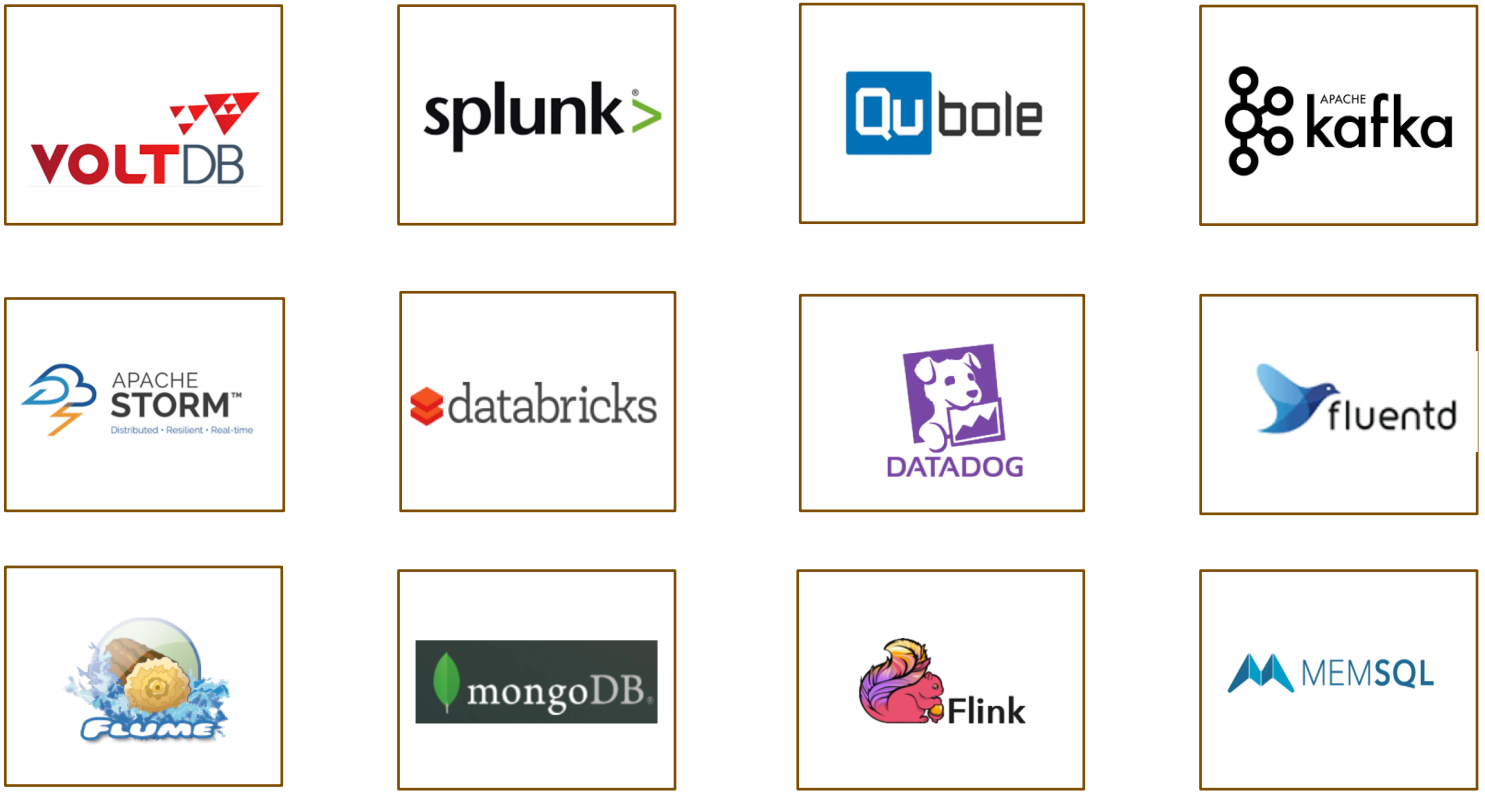
Kinesis Data Analytics

Apache Spark/Storm/Flink

AWS Lambda

Others: Anodot, VoltDB, MemSQL and more

Kinesis Data Streams 3rd Party Connectors



aws

Kinesis Data Streams Producers and Consumers

Producers

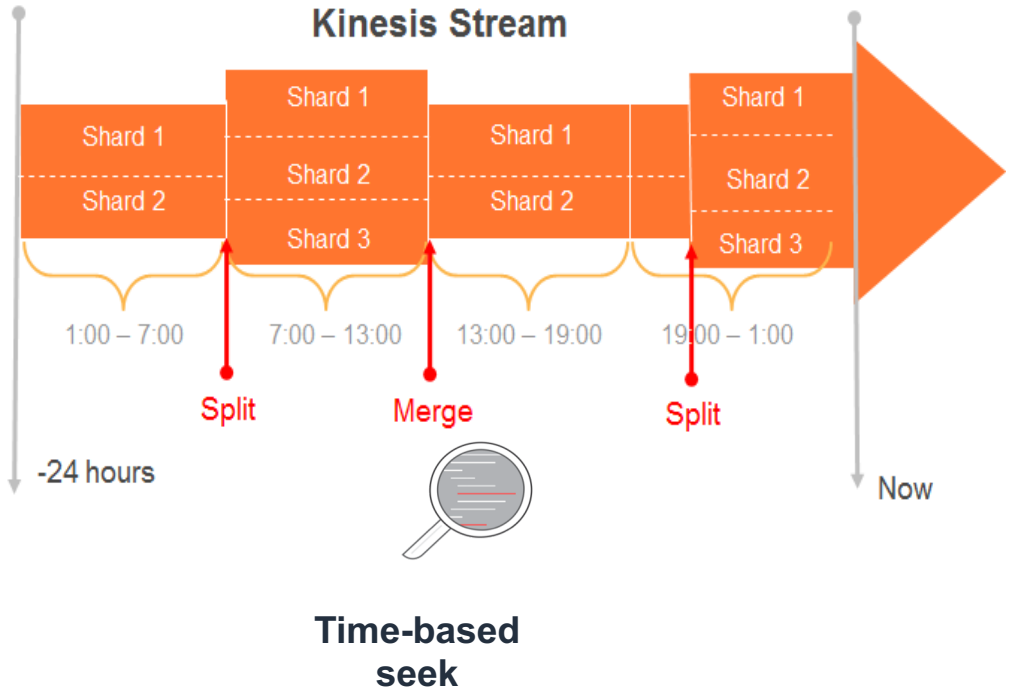
- AWS SDK
- Kinesis Producer Library
- AWS Mobile SDK
- LOG4J
- Flume
- Fluentd
- Kinesis Agent
- Apache Kafka

Consumers

- Get* APIs
- Kinesis Client Library + Connector Library
- AWS Lambda
- Amazon EMR
- Apache Storm
- Apache Spark

aws

Managed Ability to Capture & Store Data



Time-based seek

- Data streams are made of **Shards**
- Each Shard ingests data up to 1MB/sec, and up to 1000 TPS
- Each Shard emits up to 2 MB/sec
- All data is stored for **24 hours – 7 days**
- **Scale** Kinesis data streams by splitting or merging Shards
- **Replay** data inside of 24Hr -7days Window

aws

Security and Compliance

Supports VPC Endpoint powered by AWS PrivateLink

Supports server-side encryption and client-side encryption

Using SSL and HTTPS

Integrated with AWS Identity and Access Management (IAM)

FedRAMP, HIPAA, Soc

Cost-Effective

Pay-as-you-go pricing

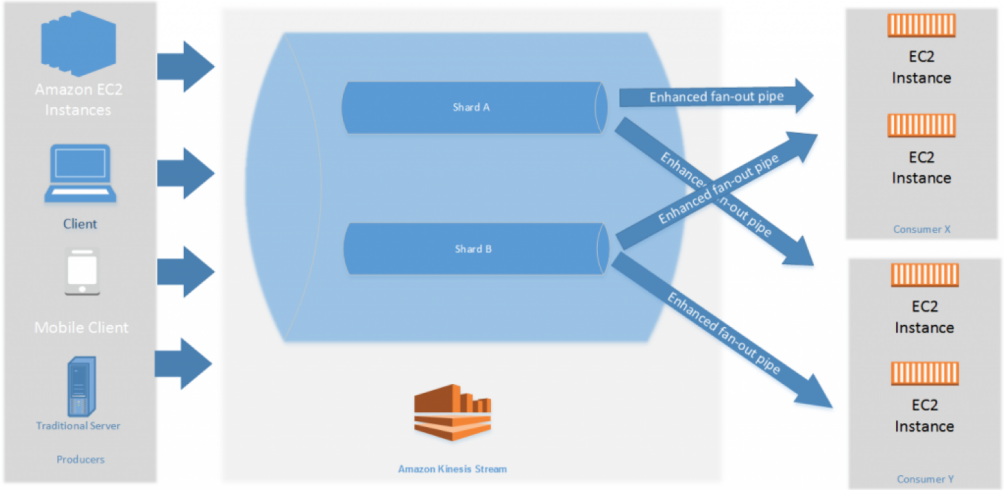
No upfront cost and no minimum fees

Based on two dimensions:

- Shard-Hour: $0.015
- PUT Payload Units (25K), per million units: $0.014

Extended data retention, per shard hour: $0.020

Enhanced Fan-Out and HTTP/2 support for Faster Streaming



HTTP/2 to allow <100 ms delivery

Enhanced Fan-out allow multiple consumers, each at 2MB/second, independently.

aws

Lab1 – Ingest with Kinesis Data Streams

Amazon Kinesis

Data Firehose

Deep Dive

Stream your data into Amazon S3, Amazon Redshift, Amazon Elasticsearch Service, and Splunk

aws

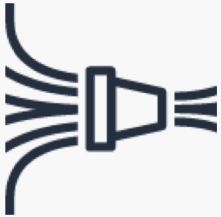
Amazon Kinesis Data Firehose

- Zero administration and seamless elasticity
- Direct-to-data store integration
- Serverless continuous data transformations
- Near real-time
- Data format conversion to Parquet/ ORC

Amazon Kinesis – Streams vs Firehose

Kinesis Data Streams

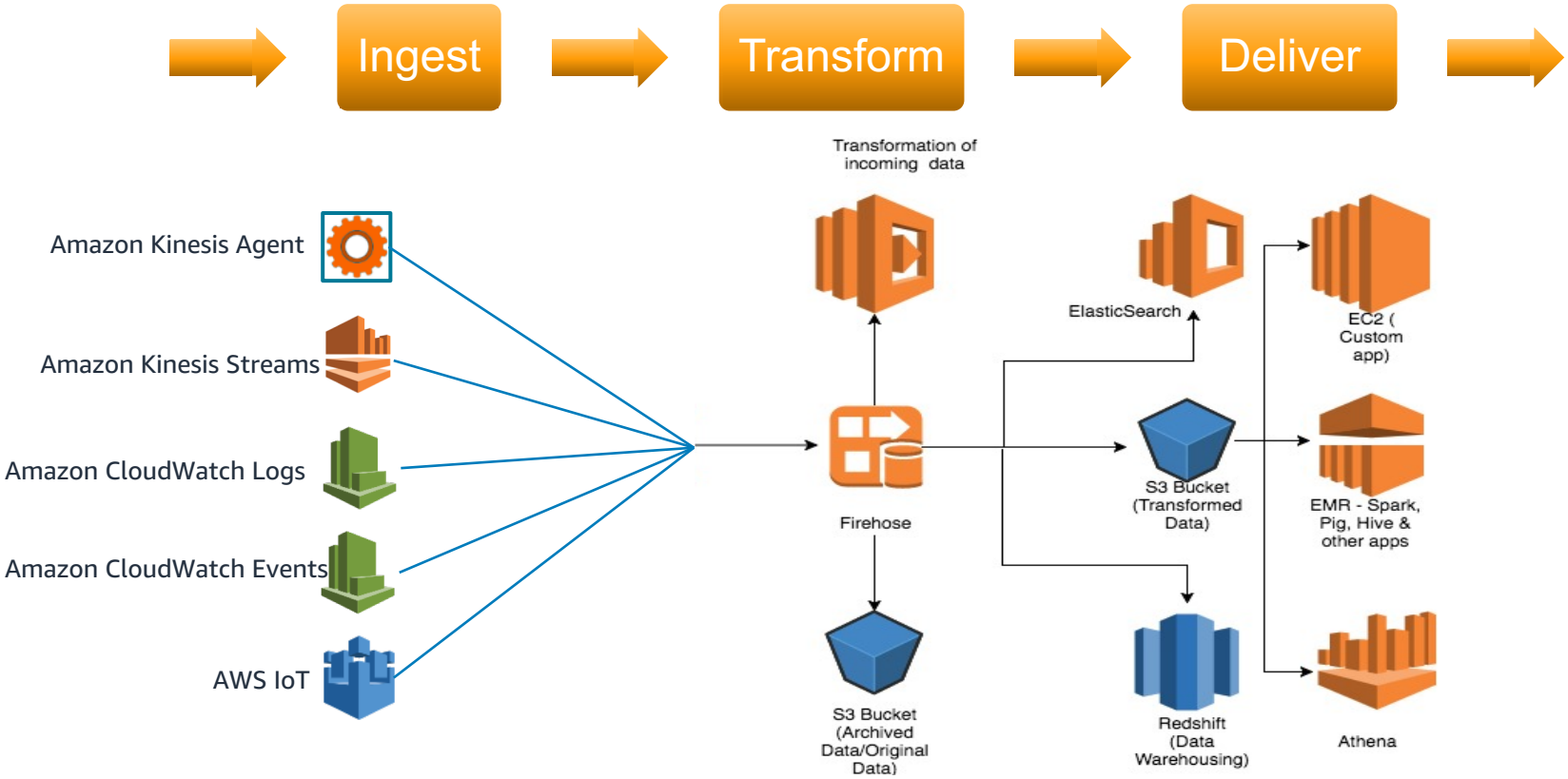
Amazon Kinesis Data Streams is for use cases that require custom processing, per incoming record, with sub-1 second processing latency, and a choice of stream processing frameworks



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

Kinesis Data Firehose – How it Works



Key Features

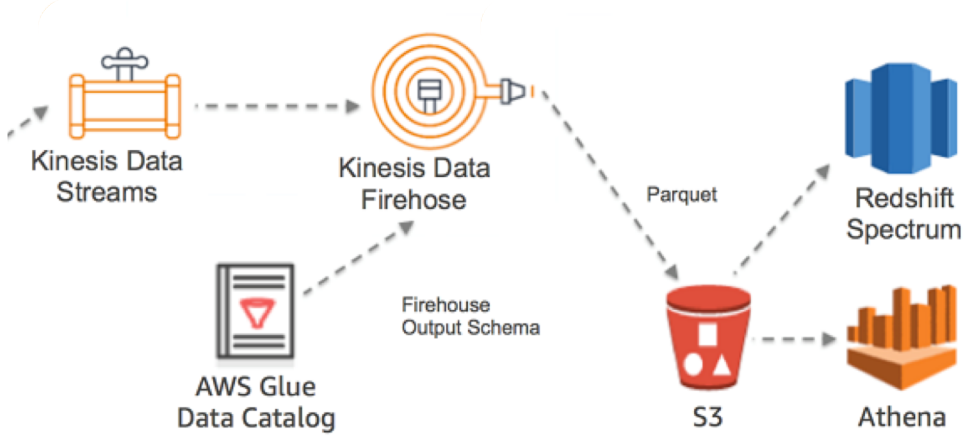
- Data durability:
  - Data backup to S3 upon delivery or transformation failure
  - 3X data replication in delivery stream for high data durability
- Up to 24 hours data retention in delivery stream to absorb backpressure from destinations
- Serverless architecture with no on-going management overhead

Key Features (continued…)

- Direct source integration with Kinesis Data Streams, CloudWatch Logs, CloudWatch Events, and AWS IoT for easy data ingestion
- Serverless data transformation before loading to destinations
- Configurable buffer size, encryption, compression, and format conversion help optimize destination cost and performance
- Metrics and logging for easy monitoring and troubleshooting

Pre-Built Data Conversions

Save space and enable faster queries compared to row-oriented formats like JSON.



- Convert the format of your input data from JSON to columnar data format **Apache Parquet** or **Apache ORC** before storing the data in Amazon S3.
- Works in conjunction to the transform features to convert other format to JSON before the data conversion

aws

Lab2 –Process Data using a Lambda function and send to Kinesis Data Firehose

Amazon Kinesis

Data Analytics

Deep Dive

Analyzing data streams with standard SQL

aws

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

aws

Example Use Case – Streaming ETL for IoT

Transform, aggregate, and filter streaming data from IoT devices

aws

Kinesis Data Analytics for SQL

Kinesis Data Analytics Applications

Connect to streaming source

Easily write SQL code to process streaming data

Continuously deliver SQL results

Connect to Streaming Data Sources

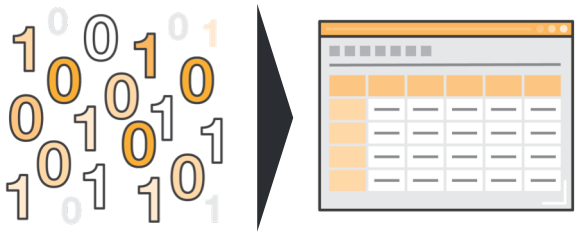
Easily **connect** to Kinesis Data streams and Kinesis Data Firehose delivery streams

Amazon Kinesis Data Firehose

Amazon Kinesis Data Streams

Automatic schema discovery which works for CSV and JSON data

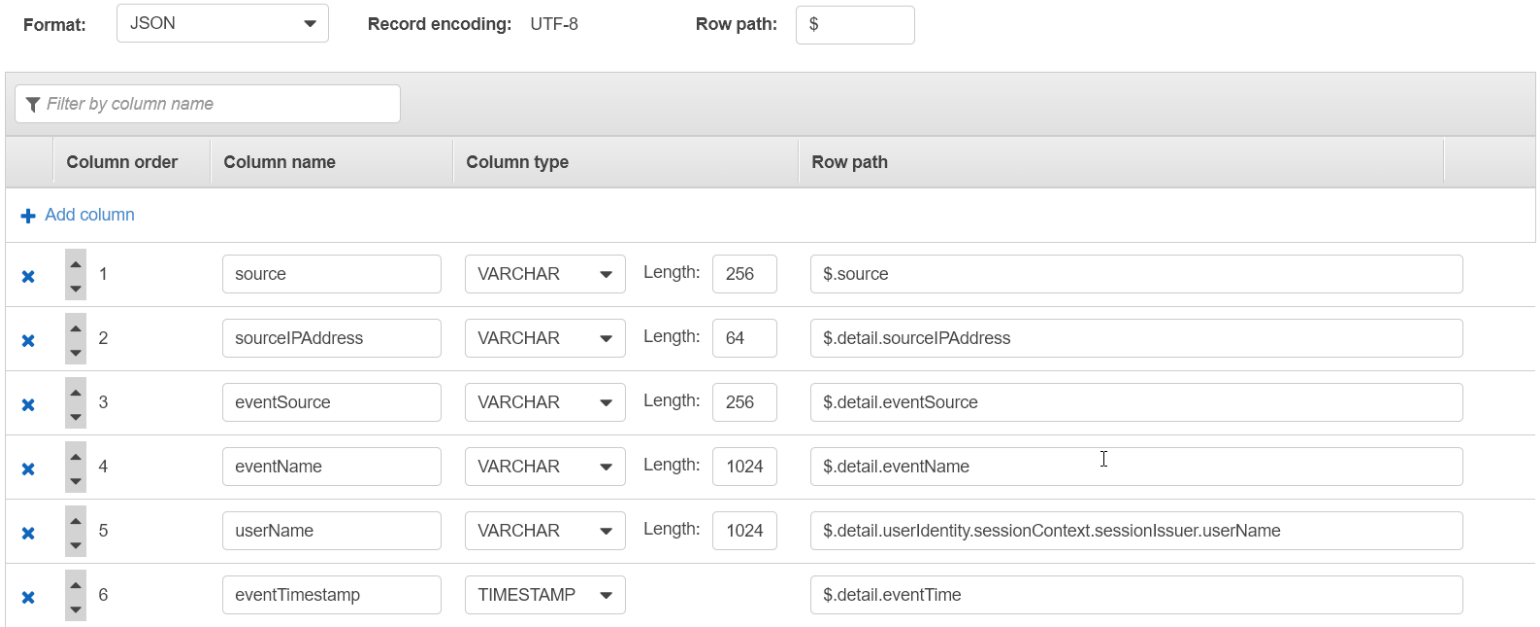
Supports multiple event types, arbitrary object nesting, single level of array nesting



aws

Pre-process Data Streams Using Schema Editor

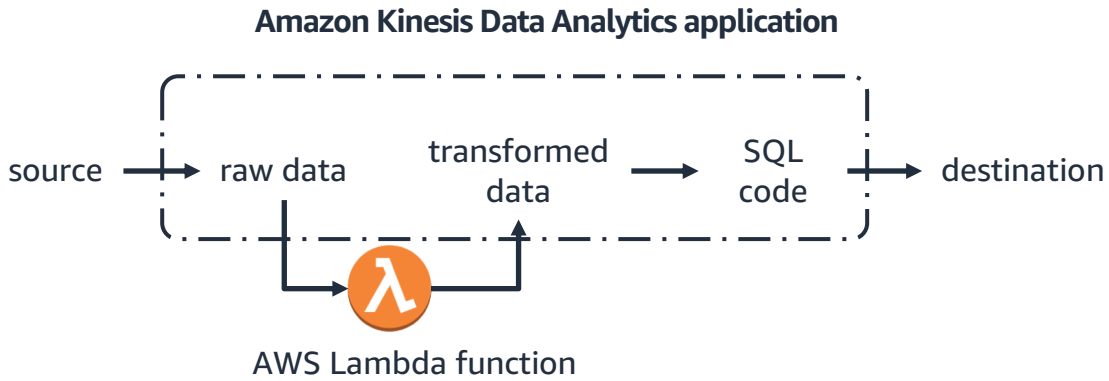
Schema editor provides fine grained control of mapping to SQL columns



Pre-process Data Streams Using AWS Lambda

Built-in AWS Lambda integration provides flexible pre-processing ahead of SQL code for:

- Normalizing 10s to 100s of different event types
- Converting other data formats (AVRO, Protobuf, ZIP) to JSON and CSV
- Custom enrichment from database tables or API calls



Easily Write SQL code to Process Data Streams

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

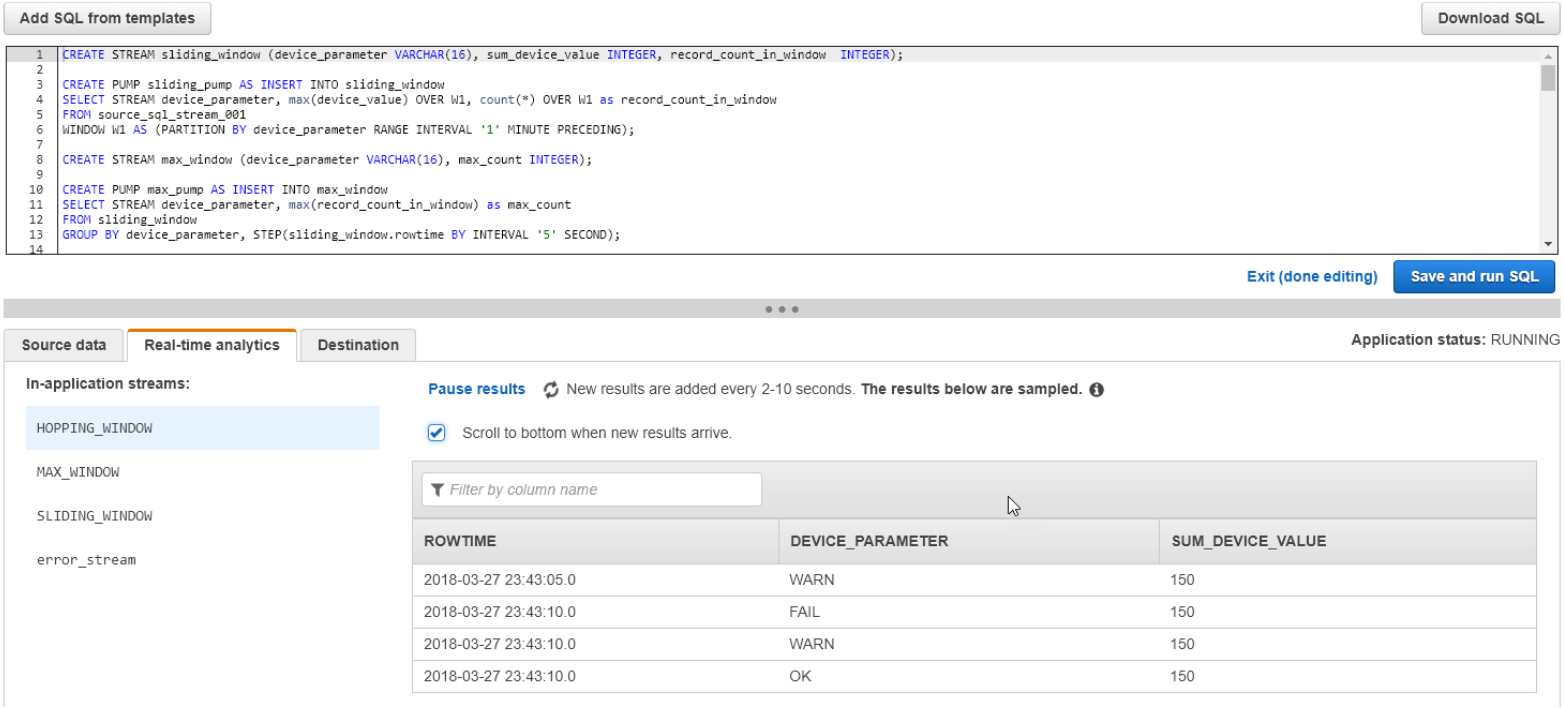
Pre-built functions include everything from sum and count distinct to machine learning algorithms

Aggregations run continuously using window operators

aws

Interactive SQL Editor

Fast, iterative development with SQL templates in console to get started



Writing Streaming SQL

Streams (in memory tables)

```
CREATE STREAM calls_per_ip_stream(
    eventTimeStamp TIMESTAMP,
    computationType VARCHAR(256),
    category VARCHAR(1024),
    subCategory VARCHAR(1024),
    unit VARCHAR(256),
    unitValue BIGINT
);
```


aws

Writing Streaming SQL

Pumps (continuous query)

```
CREATE OR REPLACE PUMP calls_per_ip_pump AS
INSERT INTO calls_per_ip_stream
SELECT STREAM "eventTimestamp",
    COUNT(*),
    "sourceIPAddress"
FROM source_sql_stream_001 ctrail
GROUP BY "sourceIPAddress",
    STEP(ctrail.ROWTIME BY INTERVAL '1' MINUTE),
    STEP(ctrail."eventTimestamp" BY INTERVAL '1' MINUTE);
```

Kinesis Data Analytics for Java

Java applications in Kinesis Data Analytics

1

Build Java applications in your IDE of choice using open source including Apache Flink

2

Upload your application code to Kinesis Data Analytics

3

Run your application in a fully managed and elastic service

aws

Apache Flink

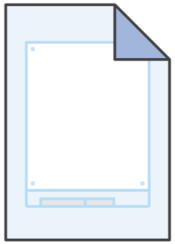
Framework and distributed engine for stateful processing of data streams

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

aws

Apache Flink supports over 25 operators

Example Operators	Typically usage
Map, FlatMap, Filter, Iterative	Basic transformations
Key By, Split, Shuffle, Custom Partition	Change logical or physical structure of the stream
Window, Reduce, Fold, Sum, Min, Max	Analytics and aggregations
Join, Union, coGroup,	Combine multiple data streams

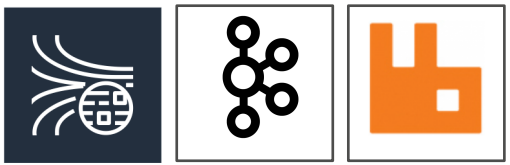
… and much, much more.

Extensible integrations with AWS services

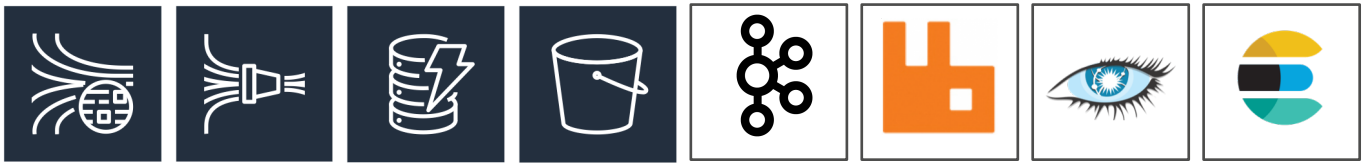
Easily add sources and sinks to an application

Build custom connectors for other data sources and sinks

Example Sources



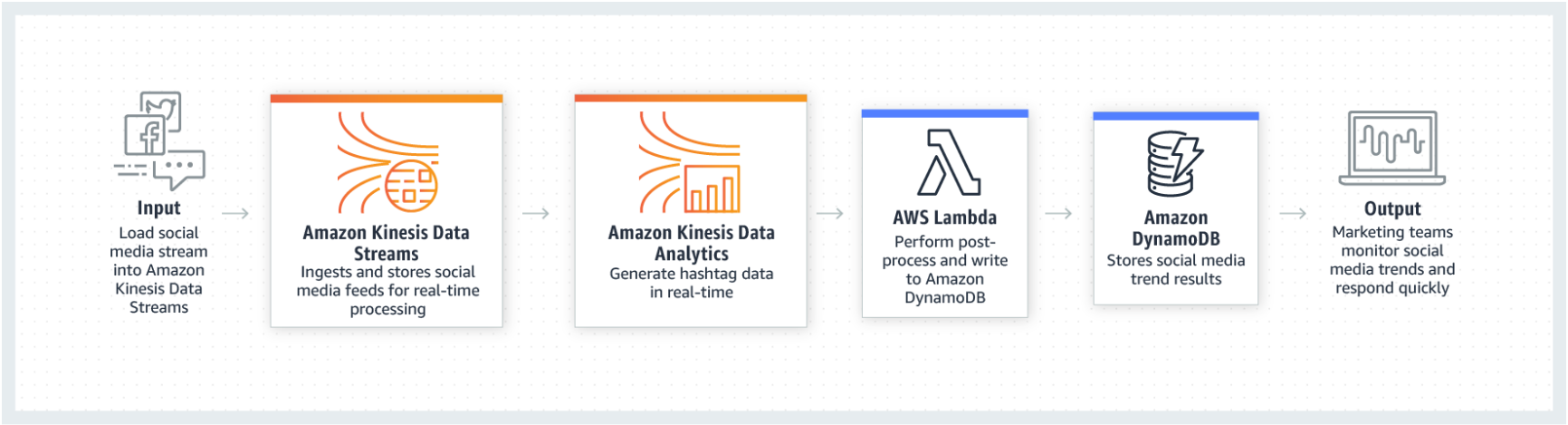
Example Destinations (Sinks)



aws

Lab3 – Clean, Aggregate, and Enrich Events with Kinesis Data Analytics SQL

Build Real-Time Applications



NETFLIX

Netflix uses Amazon Kinesis to monitor the communications between all of its applications so it can detect and fix issues quickly, ensuring high service uptime and availability to its customers.

aws

Thomson Reuters: Real-Time Dashboards

> “Using Amazon Kinesis, our solution delivers new events to user dashboards in less than 10 seconds.”
>
> **Anders Fritz**
> Senior Manager, Product Innovation

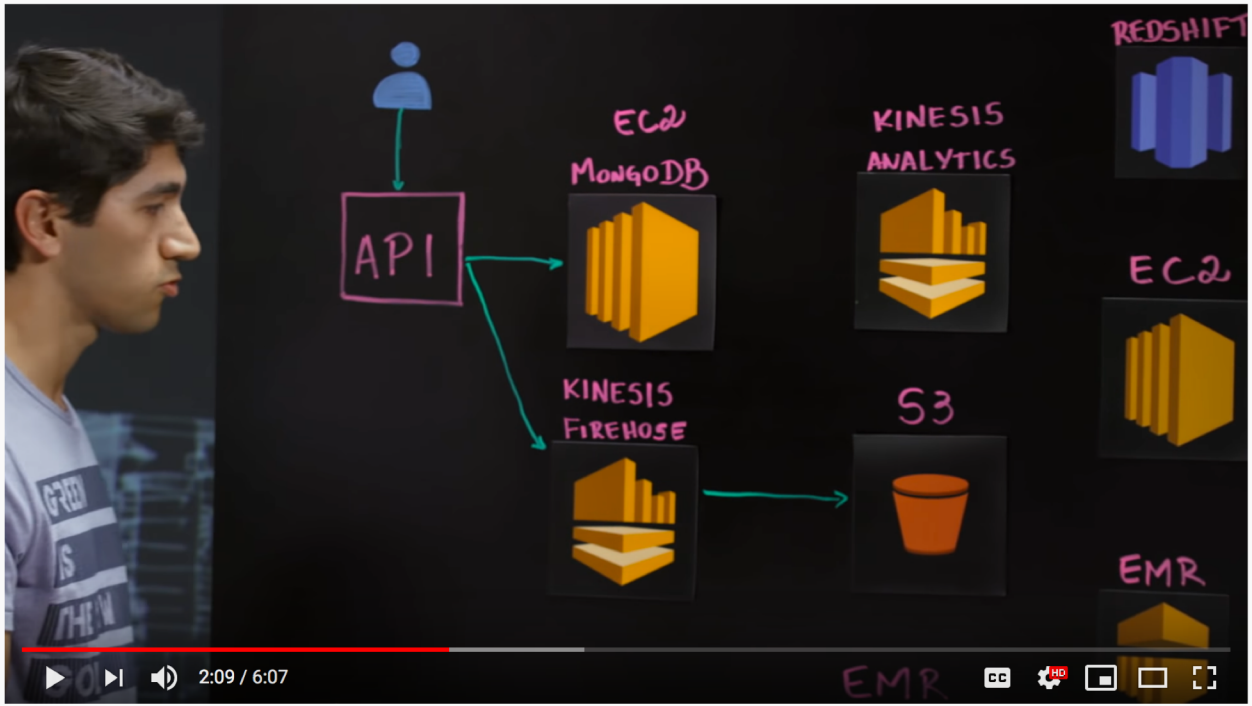
THOMSON REUTERS

Thomson Reuters provides professionals with the intelligence, technology, and human expertise they need to find trusted answers.

- Ability to process up to 4,000 events per second, anticipated to scale to 10,000 within one year
- Data pipeline accommodates twofold to threefold traffic increases during breaking news
- No data loss or downtime since launch, thanks to robust failover architecture
- Near-real-time availability of analytics data
- Simultaneous streaming and batching of data in one solution

aws

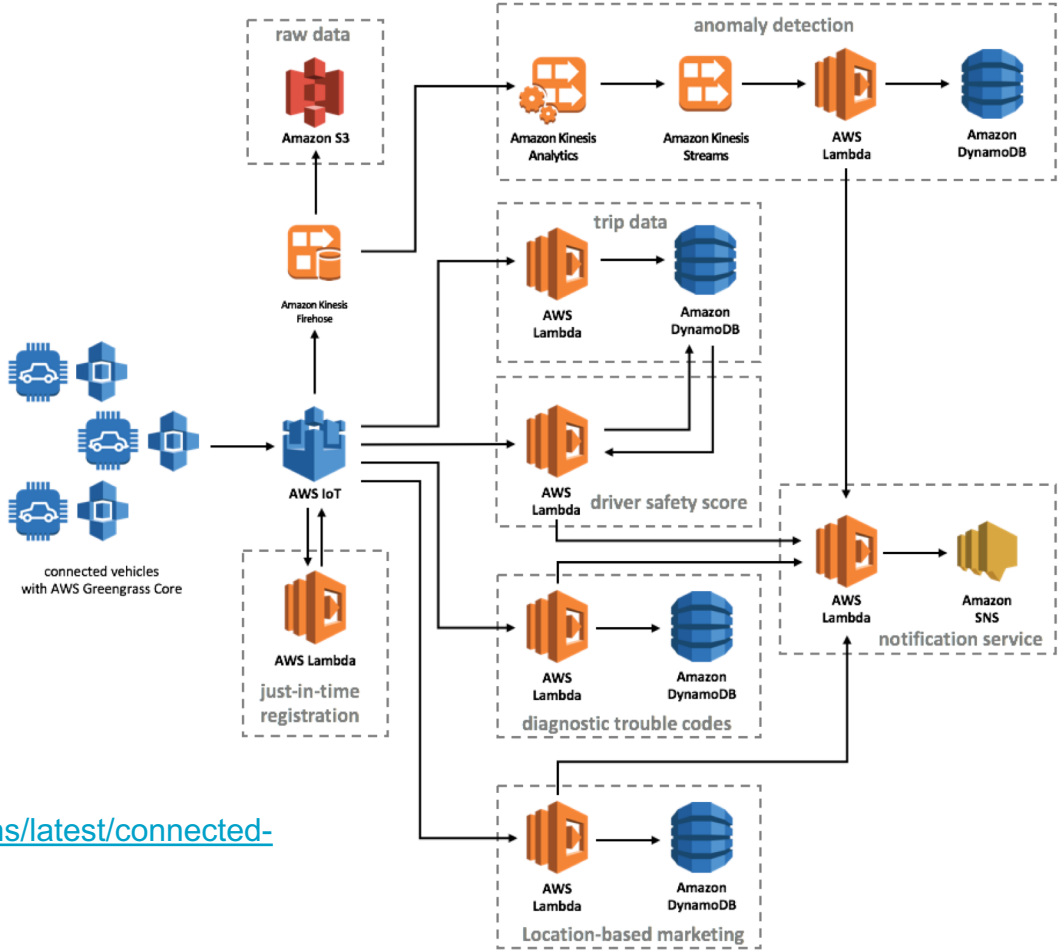
Example: Risk Processing Score



Diebold Nixdorf: Processing Risk Score in Less Than 1 Second on AWS Through a Lambda Architecture

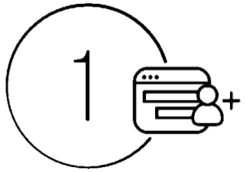
https://www.youtube.com/watch?v=7HXTEewn5bE&did=ta_card&trk=ta_card

Example: Connected Car

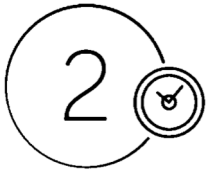


https://docs.aws.amazon.com/solutions/latest/connected-vehicle-solution/architecture.html

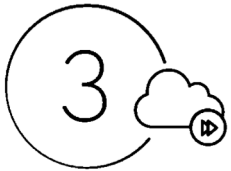
Next steps…



Put into practice what you learned in this session



Review use cases with AWS technical experts



Partner with AWS SME’s to white board and review solution design

THANK YOU!!!