AWS

MACHINE LEARNING

CERTIFICATION

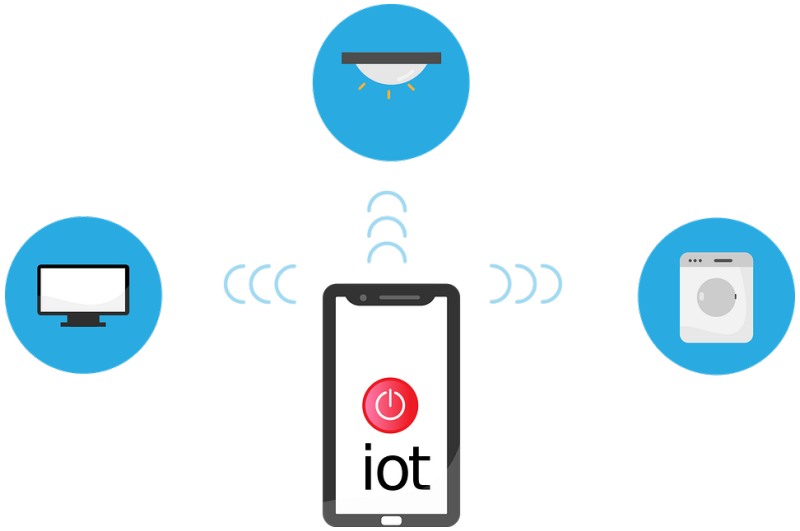
DOMAIN #1: DATA ENGINEERING (20% EXAM)

DATA STREAMING

HOW TO INGEST AND ANALYZE STREAMING DATA?

- Streaming Data can come from so many sources such as clickstreams, IOT devices, stock data..etc.
- Collecting, structuring and analysing this data is critical for companies to gain customers insights and set their marketing and product strategies.
- Data could arrive in real-time and gaining valuable insights from it in real-time as well is crucial.

STOCK DATA



INTERNET OF THINGS (IOT) DEVICES

Photo Credit: https://pixabay.com/illustrations/internet-of-things-iot-network-3671222/
Photo Credit: https://www.pexels.com/photo/blue-and-yellow-graph-on-stock-market-monitor-159888/

AWS ML CERTIFICATION EXAM DOMAINS

Domain	% of Examination
Domain 1: Data Engineering	20%
Domain 2: Exploratory Data Analysis	24%
Domain 3: Modeling	36%
Domain 4: Machine Learning Implementation and Operations	20%
TOTAL	100%

Source: https://d1.awsstatic.com/training-and-certification/docs-ml/AWS%20Certified%20Machine%20Learning%20-%20Specialty_Exam%20Guide%20(1).pdf

DOMAIN #1: WHERE ARE WE NOW!!?

SECTION #1: INTRODUCTION, DATA/ML LINGO, AWS DATA STORAGE

- What is Machine Learning and Artificial Intelligence?
- What is Amazon Web Services (AWS)?
- Artificial Intelligence and Machine learning Lingo (data types, Labeled vs. unlabeled, sagemaker groundtruth)
- structured vs. unstructured and database vs. data lake vs. data storage
- AWS Data Storage (Redshift, RDS, S3, DynamoDB)

SECTION #2: AMAZON S3

- Amazon S3 in Depth (partitions, tags)
- Amazon S3 Storage Tiers and Lifecycles
- Amazon S3 Encryption and Security
- Amazon S3 Encryption and Security – Part #2 (ACL, CloudWatch, CloudTrail, VPC)
- Additional Notes (Elastisearch, ElastiCache, and Database vs. data warehouse)

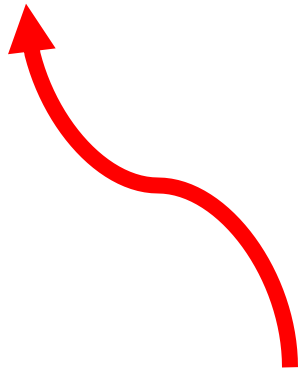
DOMAIN #1 OVERVIEW:

SECTION #3: AWS DATA MIGRATION, GLUE, PIPELINE, STEP AND BATCH

- AWS Glue (crawlers, features, built-in transformations etc)
- AWS Data pipeline
- AWS Data Migration Service (DMS)
- AWS Batch
- Step Function

SECTION #4: DATA STREAMING & KINESIS

- Kinesis Overview
- Kinesis Video Streams
- Kinesis Data Streams
- Kinesis Firehose
- Kinesis Analytics and Random Cut Forest



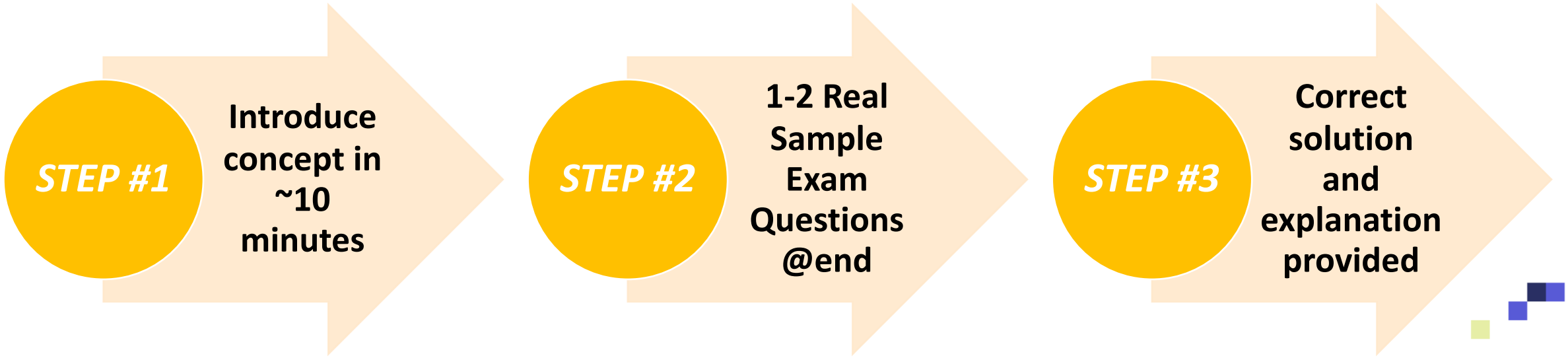
WE ARE HERE!

LECTURE DESIGN

- We know how hard it is to study for an exam especially if you have a busy schedule.
- This course is designed to be extremely on point and optimized to pass the exam.

No boring content. Zero unnecessary information.

- Here's the lecture structure that we will follow:



RECALL OUR MINI CHALLENGE AND PRIZE!

- For those of you who will successfully complete the entire section and watch the videos till the end, they will receive a valuable prize!

https://pixabay.com/illustrations/trophy-win-prize-transparent-1414791/

AWS KINESIS OVERVIEW

AWS KINESIS

- Amazon Kinesis enable enterprises/individuals to consume and analyze streaming data in real-time.
- Kinesis is cost optimized.
- This feature is critical for some enterprises who want to get information quickly such as stock traders.
- Real-time data is consumed and processed by Kinesis in Realtime such as video, audio, application logs, website clickstreams.
- Data analytics and machine learning techniques could be applied to this real-time data to get valuable quick insights.
- Kinesis is a managed alternative to Apache Kafka.

Real-time

Consume and process data in real-time to get valuable insights in minutes.

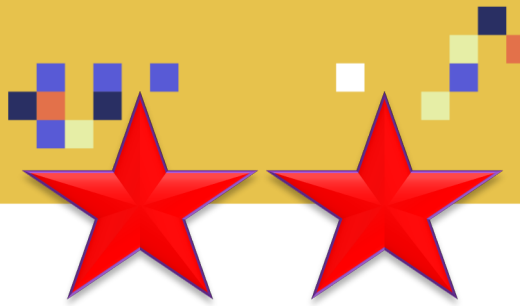
Scalable

Process extremely large amount of data at great speed.

Fully managed

Fully managed service without the need to configure servers or compute clusters.

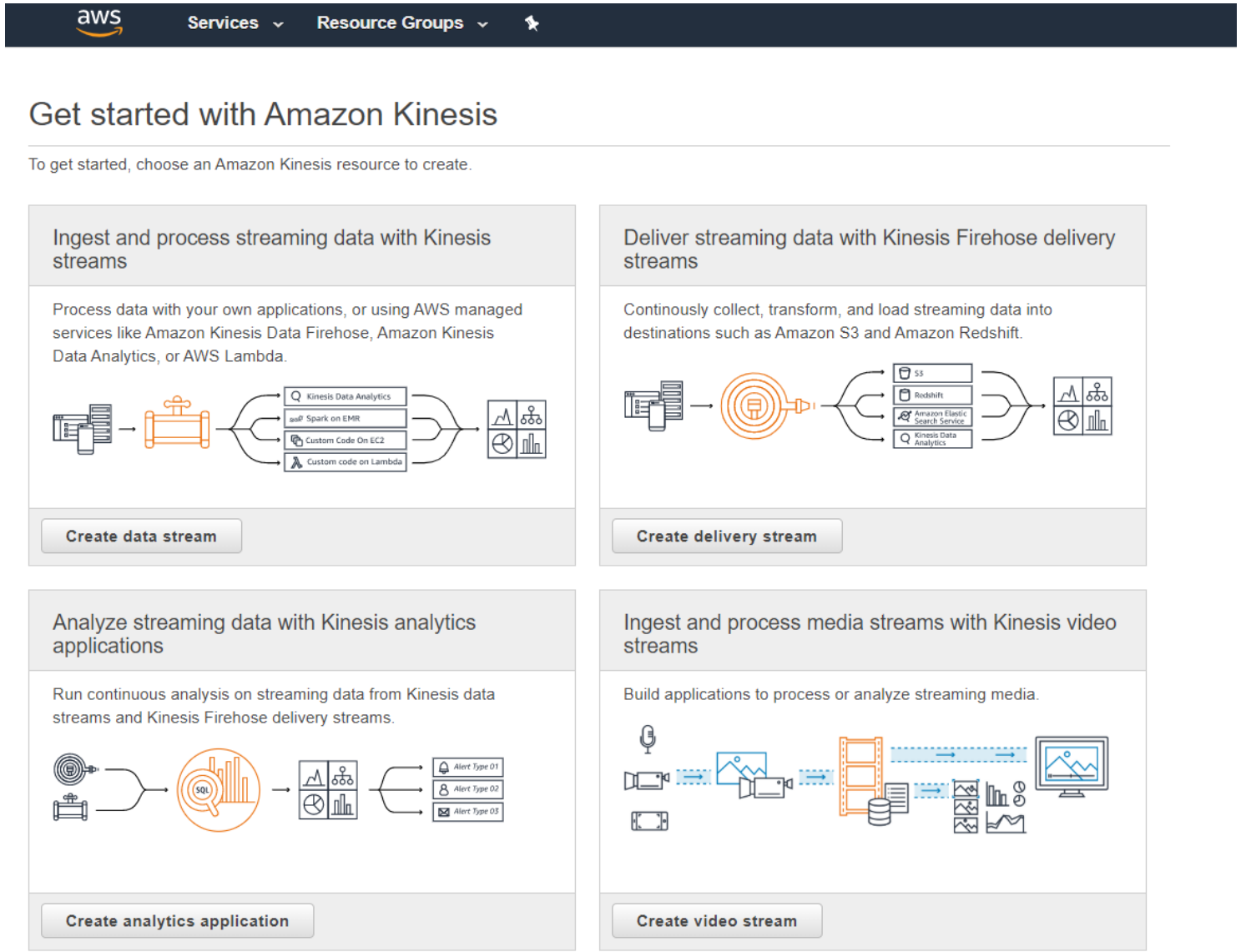
AWS KINESIS



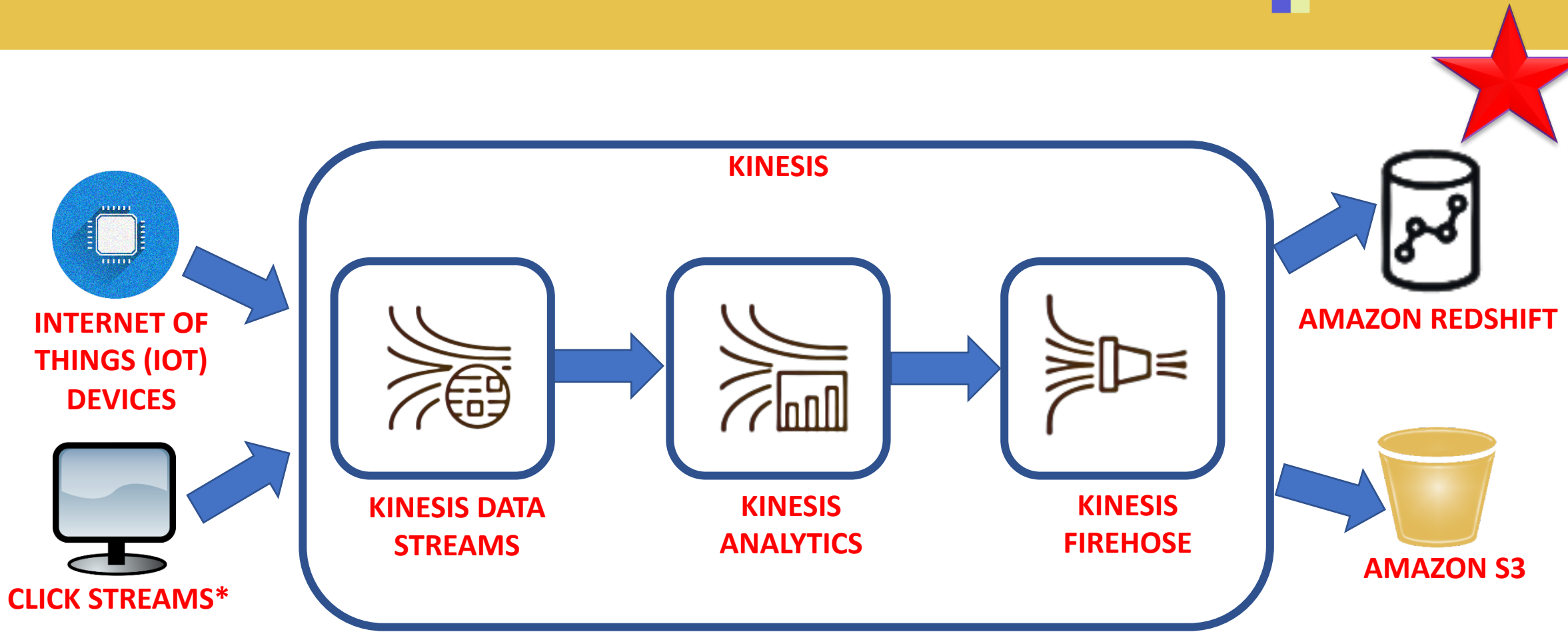
1. Amazon Kinesis Video Streams	• Video streaming service from connected devices to AWS. • Data is processed using machine learning (ML)/data analytics.
2. Amazon Kinesis Data Streams	• Real-time data streaming service. • Capture gigabytes of data per second from hundreds of thousands of sources.
3. Amazon Kinesis Data Firehose	• Near real-time service for capturing and loading data streams into AWS. • Near real-time analytics with existing business intelligence tools.
4. Amazon Kinesis Data Analytics	• Data Analytics and processing service with SQL or Java. • No need to know programming or setup any frameworks.

AWS KINESIS

https://aws.amazon.com/kinesis



AWS KINESIS



** **clickstream** is a summary of customers activity including websites, clicks, time spent on website..*

Photo Credit: https://pixabay.com/vectors/cpu-processor-computer-electronics-2103856/
Photo Credit: https://commons.wikimedia.org/wiki/File:AWS_Simple_Icons_Storage_Amazon_S3.svg
Photo Credit: https://freesvg.org/computer-flat-monitor-symbol-vector-illustration

AWS KINESIS VIDEO STREAMS – PART #1

1. KINESIS VIDEO STREAMS

- Amazon Kinesis Video Streams allows for seamless video streaming from millions of devices.
- Feed this video stream to computer vision/Deep Learning algorithms (ex: face detection).
- Kinesis video streams is extremely elastic and automatically scales for any number of devices.
- As always, pay per use model and data is kept for 1 hour and up to 10 years.
- Kinesis video streams is secure since it durably encrypts the video streams and store it.
- It is super easy to use with powerful APIs
- Connect Kinesis video streams to OpenCV, Amazon Rekognition, Apache MXNet, TensorFlow.
- You can setup Kinesis video streams quickly and easily from the AWS main console. Then, on your device (mobile phone), download Kinesis video streams SDK, and voila! You have access to secure data to store and run analytics on!
- You can integrate with Kinesis video streams with AWS DeepLens and Realtime Streaming Protocol (RTSP) camera

Photo Credit: https://pixabay.com/vectors/surveillance-camera-security-video-147831/
Photo Credit: https://commons.wikimedia.org/wiki/File:Mobile-Smartphone-icon.png

1. KINESIS VIDEO STREAMS: UNIQUE FEATURES

VIDEO STREAMING FROM NUMEROUS EDGE DEVICES

- Kinesis video streams offer easy to use SDKs to allow easy integration with edge device.
- Data could be ingested from cameras, mobile phones, satellites, LiDARs.

DEVELOP COMPUTER VISION-BASED APPS

- Kinesis video streams allow for seamless integration with Amazon Rekognition.
- It can also be integrated with other frameworks such as TensorFlow, MXNet, OpenCV.

EASILY PLAYBACK VIDEOS

- Kinesis video streams offer HTTP live streaming (HLS) service. You can play recorded videos or play live stream on any browser.

1. KINESIS VIDEO STREAMS: UNIQUE FEATURES

ELEVATED SECURITY

- Amazon Kinesis Video Streams automatically encrypt data both in transit and at rest using:
  - **AT REST:** using AWS Key Management Service (KMS)
  - **IN TRANSIT:** Using industry-standard Transport Layer Security (TLS) protocol.
- Access management using AWS Identity and Access Management (IAM).

HIGH DURABILITY

- Amazon Kinesis Video Streams relies on Amazon S3 for storage so you're in good hands!

ZERO HASSLE/NO SERVERS TO MANAGE

- No maintenance, software update management and hassle! Amazon Kinesis Video Streams manages the entire infrastructure.

AWS KINESIS VIDEO STREAMS – PART #2

1. KINESIS VIDEO STREAMS: HOW DOES IT WORK?

VIDEO SOURCES

Using Kinesis video streams SDK, multiple devices can stream videos in real-time

KINESIS VIDEO STREAMS

Durably and securely ingests and stores video streams in Realtime

AMAZON REKOGNITION VIDEO

MXNET AND TENSORFLOW

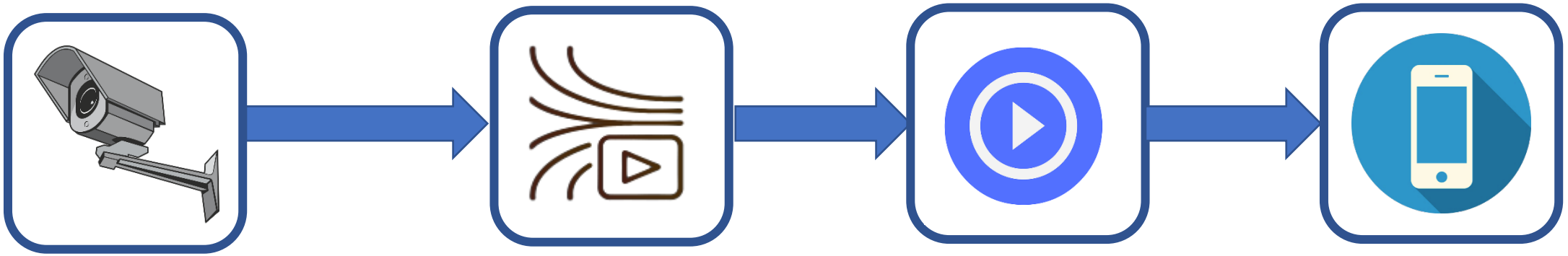
HLS-BASED VIDEO PLAYBACK

AMAZON SAGEMAKER

Access video streams and run intelligent Realtime algorithms (face detection)

1. KINESIS VIDEO STREAMS: USE CASE #1

- You can use Amazon Kinesis video streams to develop a smart home
- By easily installing video cameras such as baby video cameras, outside home surveillance cameras, and pet cameras.
- You can then ingest these video streams and record it, play it live on your device, or even run deep learning algorithms to detect certain people.
- Happy House Problem cannot be made easier! Only smiling people are allowed in the house! https://www.kaggle.com/iarunava/happy-house-dataset



PET MONITORING CAMERA

Using Kinesis video streams SDK, stream pet video

KINESIS VIDEO STREAMS

Durably and securely ingests and stores video streams in Realtime

VIDEO PROCESSING APP

Play back video using HTP Live Streaming (HLS)

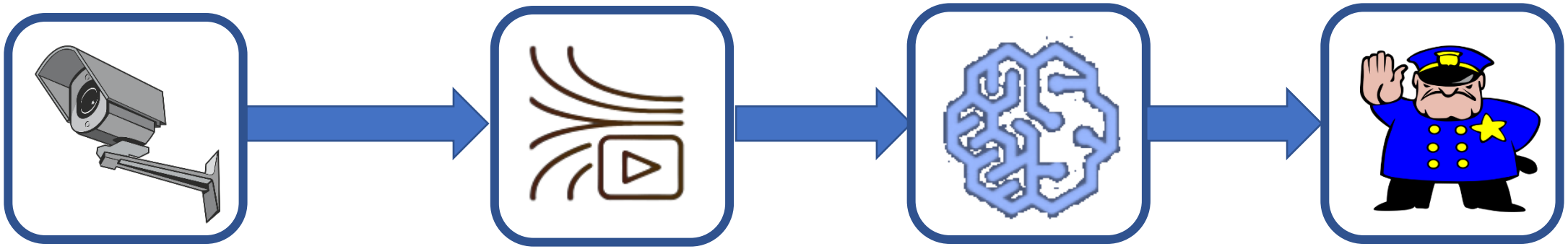
SMART PHONE

Watch your favourite pet and send Deep Learning based alarms

Photo Credit: https://pixabay.com/vectors/surveillance-camera-security-video-147831/
Photo Credit: https://commons.wikimedia.org/wiki/File:Mobile-Smartphone-icon.png
https://pixabay.com/illustrations/icon-play-video-movie-youtube-1968245/

1. KINESIS VIDEO STREAMS: USE CASE #2

- Governments can use Amazon Kinesis to develop smart cities by reducing crime rates and solving traffic!
- By installing cameras in traffic lights and shopping malls.



TRAFFIC LIGHT MONITORING CAMERA

Using Kinesis video streams SDK, stream traffic light video

KINESIS VIDEO STREAMS

Durably and securely ingests and stores video streams in Realtime

AMAZON AI

Run deep learning algorithm to detect car plates

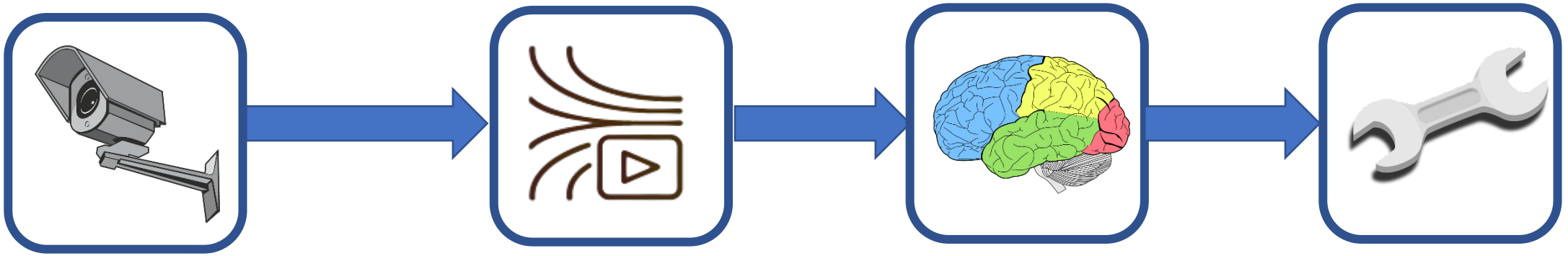
POLICE

Police receives Realtime notifications

Photo Credit: https://pixabay.com/vectors/surveillance-camera-security-video-147831/
Photo Credit: https://commons.wikimedia.org/wiki/File:Mobile-Smartphone-icon.png
https://pixabay.com/illustrations/icon-play-video-movie-youtube-1968245/
https://publicdomainvectors.org/en/free-clipart/Police-man-vector-drawing/7211.html

1. KINESIS VIDEO STREAMS: USE CASE #3

- You can use Amazon Kinesis video streams to develop a smart preventative maintenance system in factories.
- For example: you can use Amazon Kinesis Video Streams to ingest data from radars, Lidars and cameras and then run AI/ML algorithms using TensorFlow/MXNet to predict when a valve would fail. This is crucial to reduce downtime.



VALVE MONITORING CAMERA

Stream video of valve from camera @manufacturing facility

KINESIS VIDEO STREAMS

Durably and securely ingests and stores video streams in Realtime

AI/ML FRAMEWORK

Predict whether the valve will fail or not

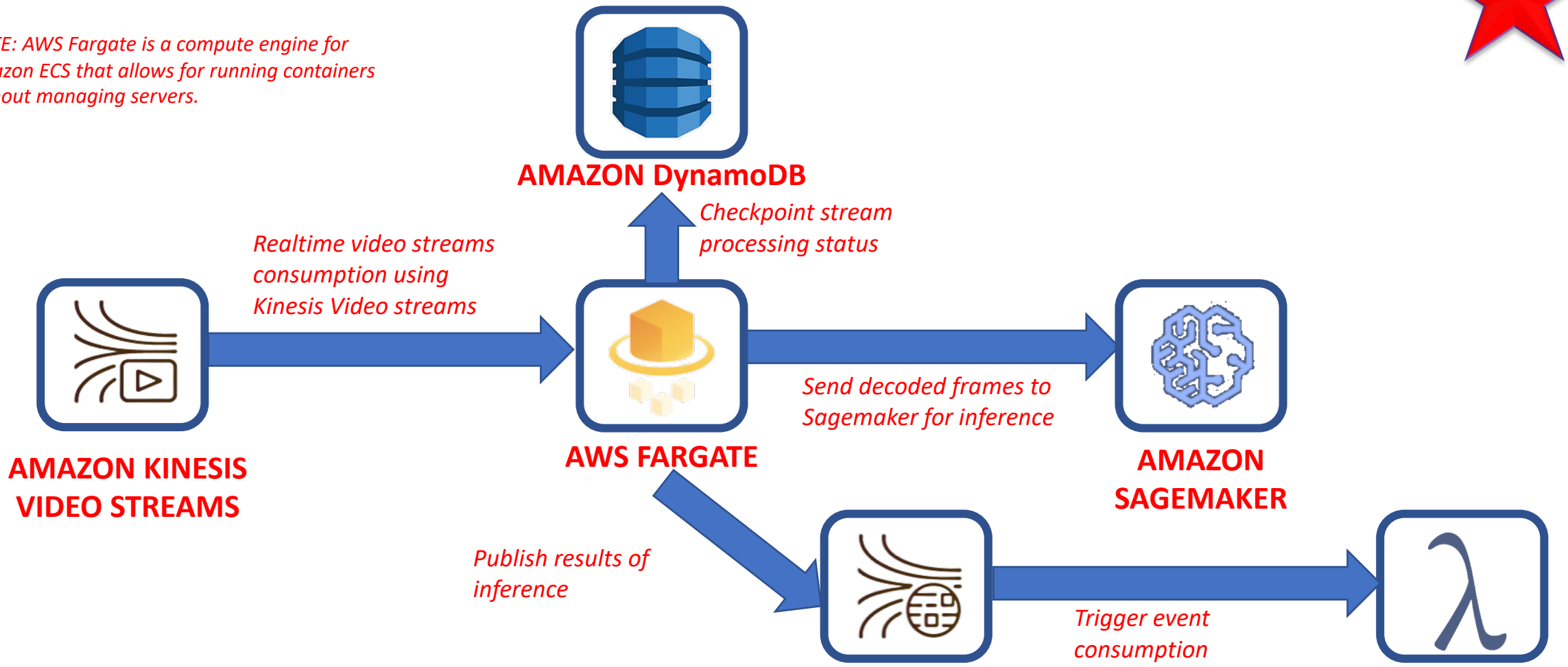
OUTPUT

Schedule preventative maintenance

Photo Credit: https://pixabay.com/vectors/surveillance-camera-security-video-147831/
Photo Credit: https://pixabay.com/vectors/wrench-tools-repair-fix-toolbox-3459519/
Photo Credit: https://pixabay.com/illustrations/brain-lobes-neurology-human-body-1007686/

1. KINESIS VIDEO STREAMS: SAGEMAKER, DATASTREAMS, AWS LAMBDA, DYNAMODB INTEGRATION

NOTE: AWS Fargate is a compute engine for Amazon ECS that allows for running containers without managing servers.



Source: https://aws.amazon.com/blogs/machine-learning/analyze-live-video-at-scale-in-real-time-using-amazon-kinesis-video-streams-and-amazon-sagemaker/
Photo Credit: https://en.wikipedia.org/wiki/File:Lambda-letter-lowercase-symbol-Garamond.svg

AWS KINESIS DATA STREAMS – PART #1

2. KINESIS DATA STREAMS

- Amazon Kinesis Data Streams (KDS) is streaming service that works seamlessly in Realtime.
- You can ingest millions of data coming from various sources such as click streams, IOT, and several devices.
- Realtime analytics can be conducted on the data and displayed on a dashboard.

KINESIS DATA STREAMS

Photo Credit: https://www.needpix.com/photo/1539231/nyse-newyorkstockexchange-floor-business-commerce-trading-economy-people-newyork

2. KINESIS DATA STREAMS

WORKS IN REAL-TIME

- Within 70 ms, let your streaming data ready for analysis (Amazon S3, AWS lambda) from multiple sources.

HIGH DATA DURABILITY

- Kinesis data streams ensures data durability by (1) data replication over 2 data centers, (2) 7 days data storage.

ELEVATED SECURITY

- Kinesis data streams offer elevated security by: (1) allowing KDS-enabled data encryption, (2) Amazon Virtual Private Cloud (VPC) privately accessed network, (3) server-side encryption and AWS KMS master keys.

SIMPLE

- Super easy and simple to build powerful and reliable streaming service by (1) integrating KDS with Kinesis Firehose, Kinesis data analytics, AWS lambda, (2) leveraging AWS SDK, Kinesis Client Library (KCL), connectors, and agents.

HIGH ELASTICITY

- Kinesis data streams is very elastic and can be scaled easily (scale from thousands to millions of PUT records per second)

OPTIMIZED (REDUCED) COST

- Zero infrastructure/upfront cost. Pay per use model.

2. KINESIS DATA STREAMS: HOW IT WORKS?

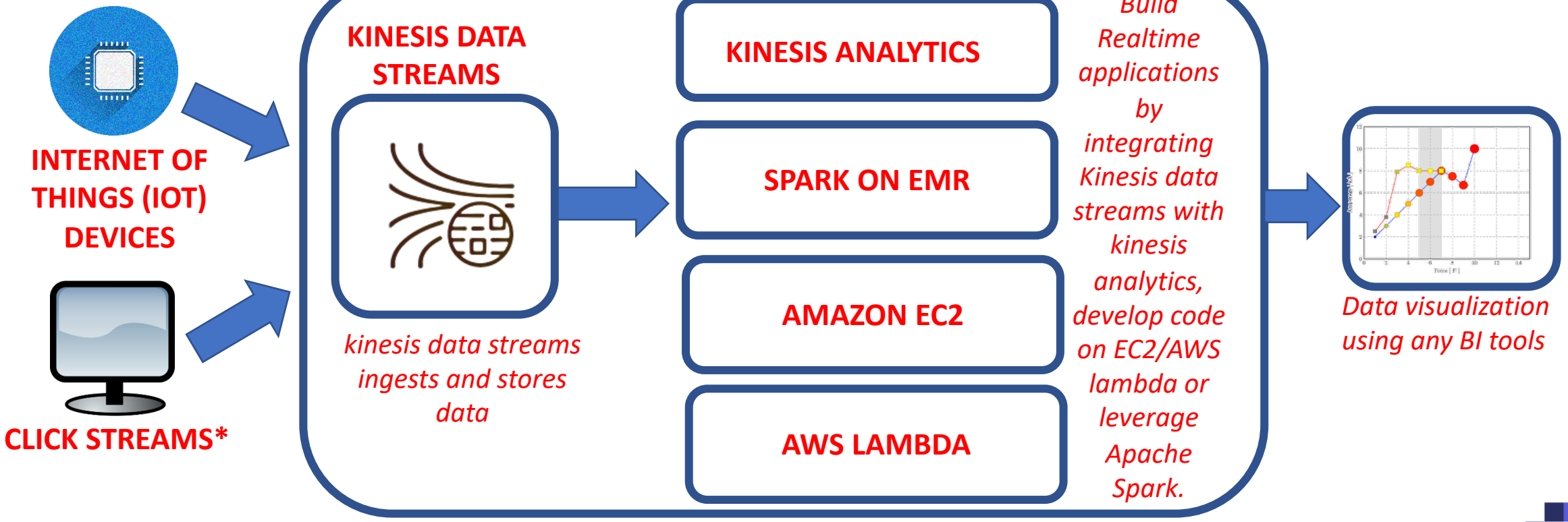
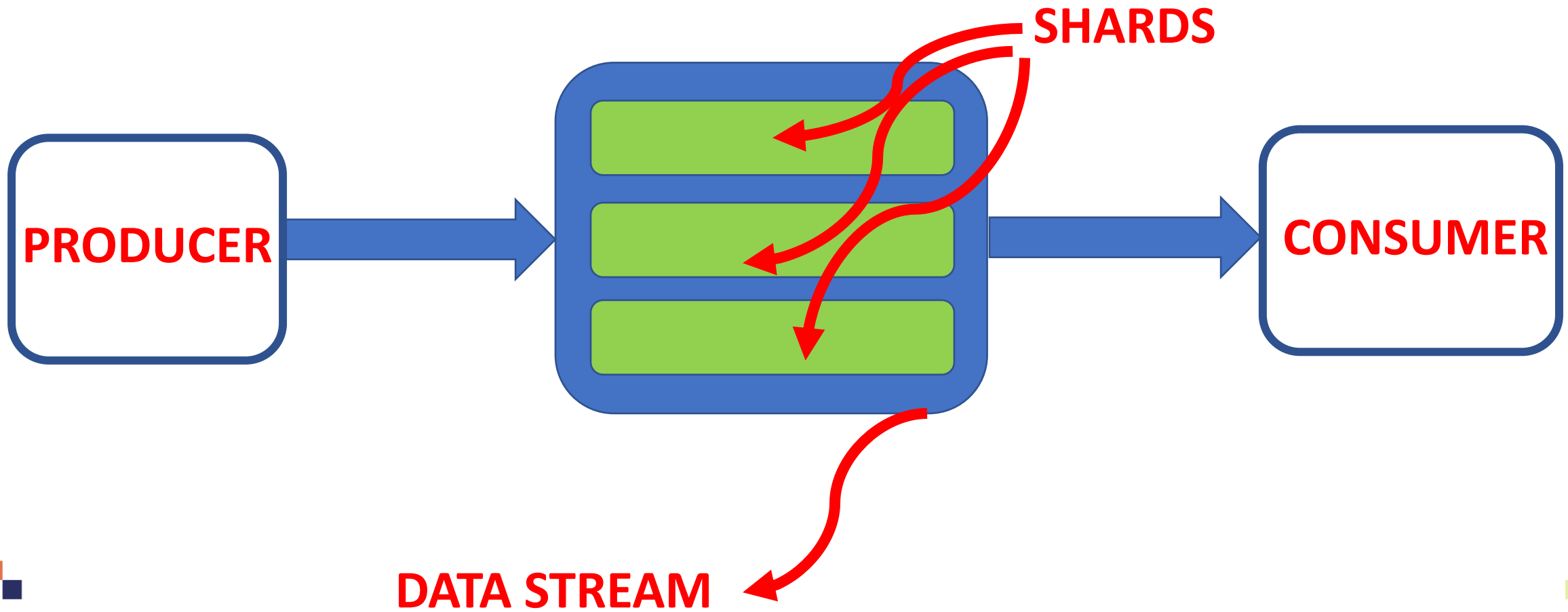


Photo Credit: https://pixabay.com/vectors/cpu-processor-computer-electronics-2103856/
Photo Credit: https://commons.wikimedia.org/wiki/File:AWS_Simple_Icons_Storage_Amazon_S3.svg
Photo Credit: https://freesvg.org/computer-flat-monitor-symbol-vector-illustration

AWS KINESIS DATA STREAMS – PART #2

2. KINESIS DATA STREAMS: DEFINTIONS



2. KINESIS DATA STREAMS: DEFINTIONS

Data Producer:

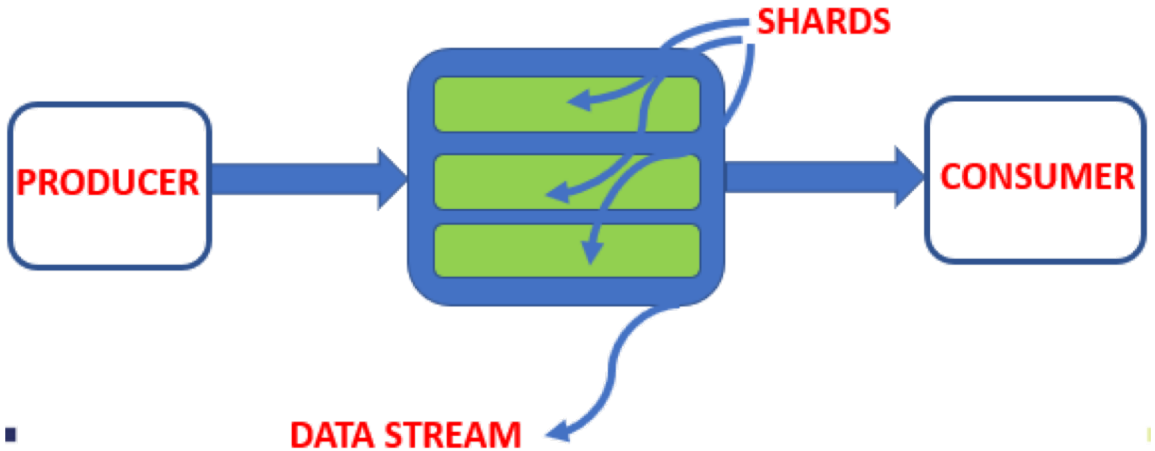
- Applications that are generating data to be ingested by Kinesis data stream.
- For each record, producers associate partition keys.
- Partition keys are important to map records to shards.

Data Consumer:

- Consists of Kinesis applications receiving data from shards in stream.
- Real-time performance allow for spot analytics.

Data Stream:

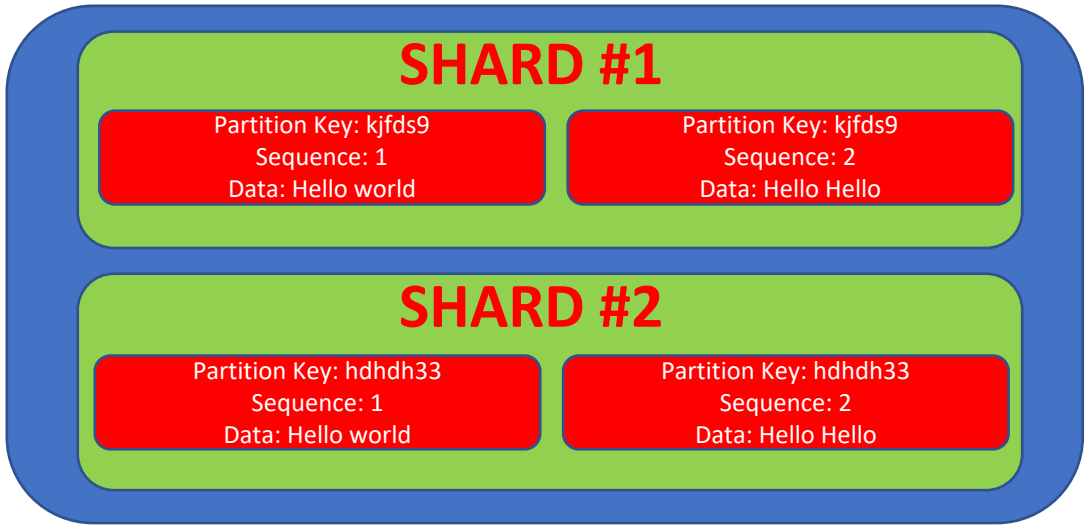
- A group of shards form a data stream.
- Data is retained for 24 hours by default.
- Data could be retained up to 7 days (upon request).
- Data is immutable; once ingested in KDS, cannot be deleted

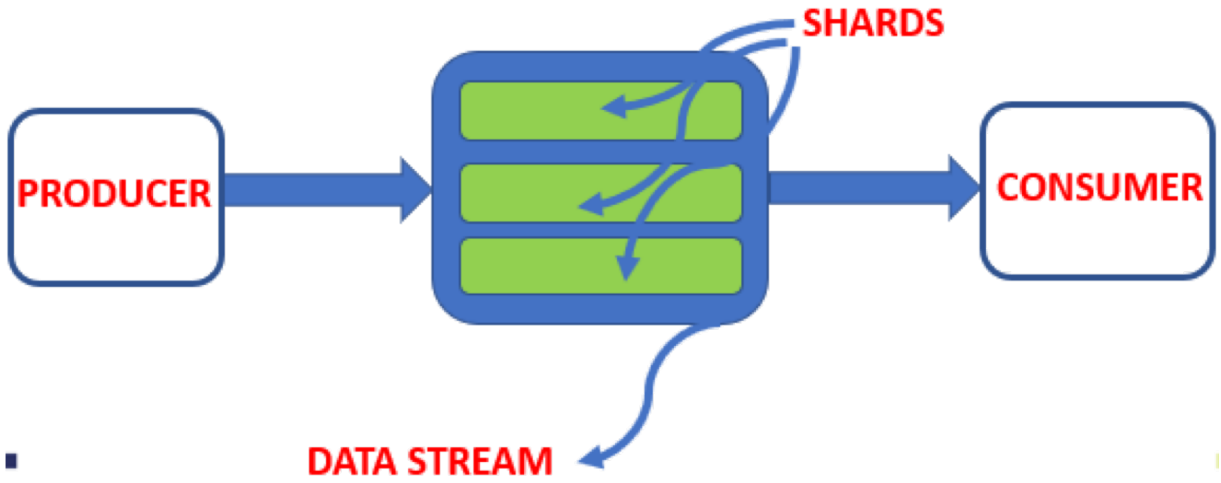


2. KINESIS DATA STREAMS: DEFINTIONS

Shard:

- A shard is the base throughput unit
- Shards are portioned in advance which require capacity planning.
- Each shard = 1000 data records per second (or 1MB/sec).
- Shards could be added via AWS console, auto scaling, UpdateShardCount API, trigger automatic scaling via AWS Lambda.
- Default limit is 500 shards but unlimited number could be requested.
- Data record consists of a unit of captured data including: (1) Sequence number, (2) partition key, and (3) payload (limited to 1 MB).





2. KINESIS DATA STREAMS: INTERACTION

KINESIS PRODUCER LIBRARY (KPL)

- The KCL allows for data writing to a Kinesis Data Streams.
- KPL simplifies producer application development

KINESIS CLIENT LIBRARY (KCL)

- The KCL allows for data consumption from Kinesis Data Streams.
- It acts as an intermediary between processing record code and Kinesis Data Streams.
- KCL handles the complex tasks of managing instances.

KINESIS SDK API

- Kinesis Data Streams API allows for managing streams, resharding, and putting and getting records.
- It offers a low level interface with Kinesis data streams.

- *KPL can incur an additional processing delay compared to AWS SDK API.*
- *For time sensitive applications, it is recommended to use AWS SDK directly.*

2. KINESIS DATA STREAMS: SHARDS

Amazon Kinesis

Dashboard

Data Streams

Data Firehose

Data Analytics

Video Streams

External resources

What's new

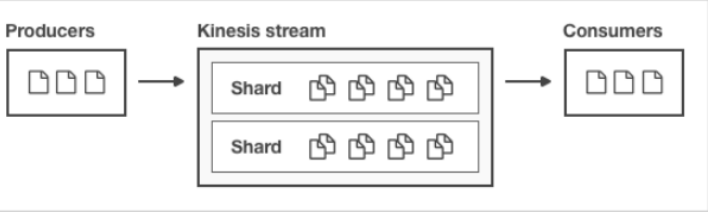
Create Kinesis stream

Kinesis stream name* Trial_Stream

Acceptable characters are uppercase and lowercase letters, numbers, underscores, hyphens, and periods.

Shards

A shard is a unit of throughput capacity. Each shard ingests up to 1MB/sec and 1000 records/sec, and emits up to 2MB/sec. To accommodate for higher or lower throughput, the number of shards can be modified after the Kinesis stream is created using the API. Learn more



▼Estimate the number of shards you'll need

Shard calculator

Average record size 1000 KB

Record size is an integer between 1 and 1024

Max records written 10 per second

(Number of records per second) x (Number of producers)

Number of consumer applications 10

Estimated shards 49 **Use this value**

Number of shards*

You can provision up to 500 more shards before hitting your account limit of 500.

AWS KINESIS FIREHOSE

3. KINESIS FIREHOSE: OVERVIEW

- Amazon Kinesis Data Firehose allows for easy and cost effectively way to load streaming data into data lakes.
- Amazon kinesis firehose allows for near real-time analytics.
- Using Kinesis data firehose, data could be ingested, transformed, encrypted and loaded into:
  - Amazon S3
  - Amazon Redshift
  - Amazon Elasticsearch Service
  - Splunk
- Kinesis data firehose does not require upfront cost or infrastructure management.
- It follows pay per use model.
- It automatically scales to match any throughput of data.
- Data Transformation through AWS Lambda (CSV => JSON) and supports compression with Amazon S3 as destination
- To optimize cost, Amazon Kinesis firehose allows for converting the incoming data into columnar formats such as Apache Parquet and Apache ORC, before feeding it into Amazon S3.
- Great Video: https://aws.amazon.com/kinesis/data-firehose/

3. KINESIS FIREHOSE: UNIQUE FEATURES

EXTREMELY EASY TO SETUP AND USE

- In minutes, you can capture, transform, and load streaming data from multiple sources. Integrated with AWS data lakes and data stores
- All scaling, sharding, servers maintenance end management are taken care by AWS.

SEAMLESS INTEGRATION WITH AWS STORAGE

- Amazon Kinesis Data Firehose is easily integrated with Amazon S3, Amazon Redshift, and Amazon Elasticsearch Service.

SERVERLESS DATA TRANSFORMATION

- Without the need to create data processing pipeline (no servers), you can use Kinesis Data Firehose to convert raw data into any format before storing it into S3.

3. KINESIS FIREHOSE: UNIQUE FEATURES

NEAR REAL-TIME PERFORMANCE

- Amazon Kinesis Data Firehose captures and loads data into S3, Redshift and ElasticSearch within 60 secs data (near real-time).

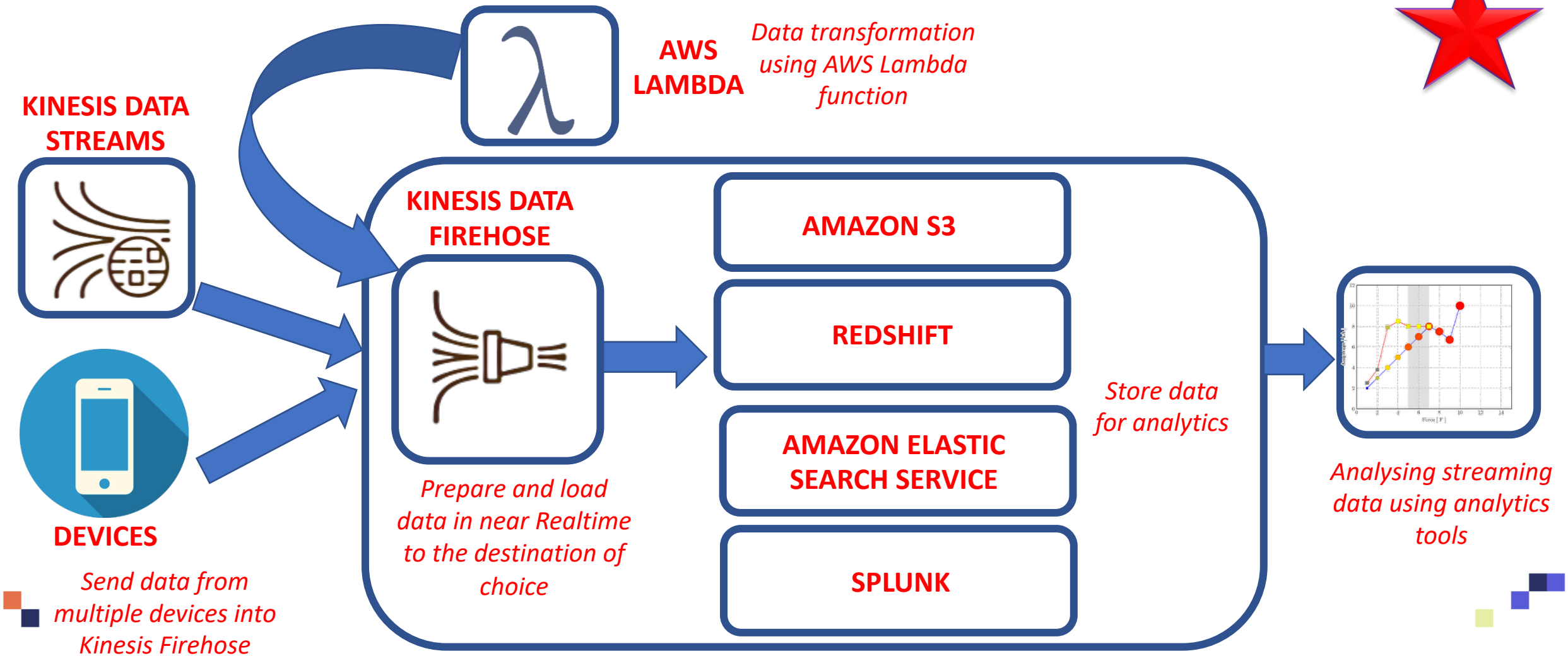
NO MAINTENANCE AND SERVERS ADMINSTRATION

- Fully managed service with automatic scaling.

PAY PER USE

- No upfront cost and pay per use model.

3. KINESIS FIREHOSE: HOW DOES IT WORK?



3. KINESIS FIREHOSE VS. STREAMS

	KINESIS DATA STREAMS	KINESIS FIREHOSE
USE/SERVICE	SCALABLE REAL-TIME STREAMING SERVICE	DATA TRANSFER SERVICE TO LOAD STREAMING DATA INTO S3, REDSHIFT AND ELASTICSEARCH
LATENCY	REAL TIME (~200 MS FOR CLASSIC AND 70 MS FOR ENHANCED FAN OUT)	NEAR REAL TIME (60 SECS)
STORAGE	FROM 1 TO 7 DAYS	NO DATA STORAGE
SCALING	REQUIRE SHARDS ADMINSTRATION	AUTO SCALING
SERVICE MANAGEMENT	MANAGED SERVICE EXCEPT FOR SHARDS CONFIGURATION	FULLY MANAGED

AWS KINESIS DATA ANALYTICS – PART #1

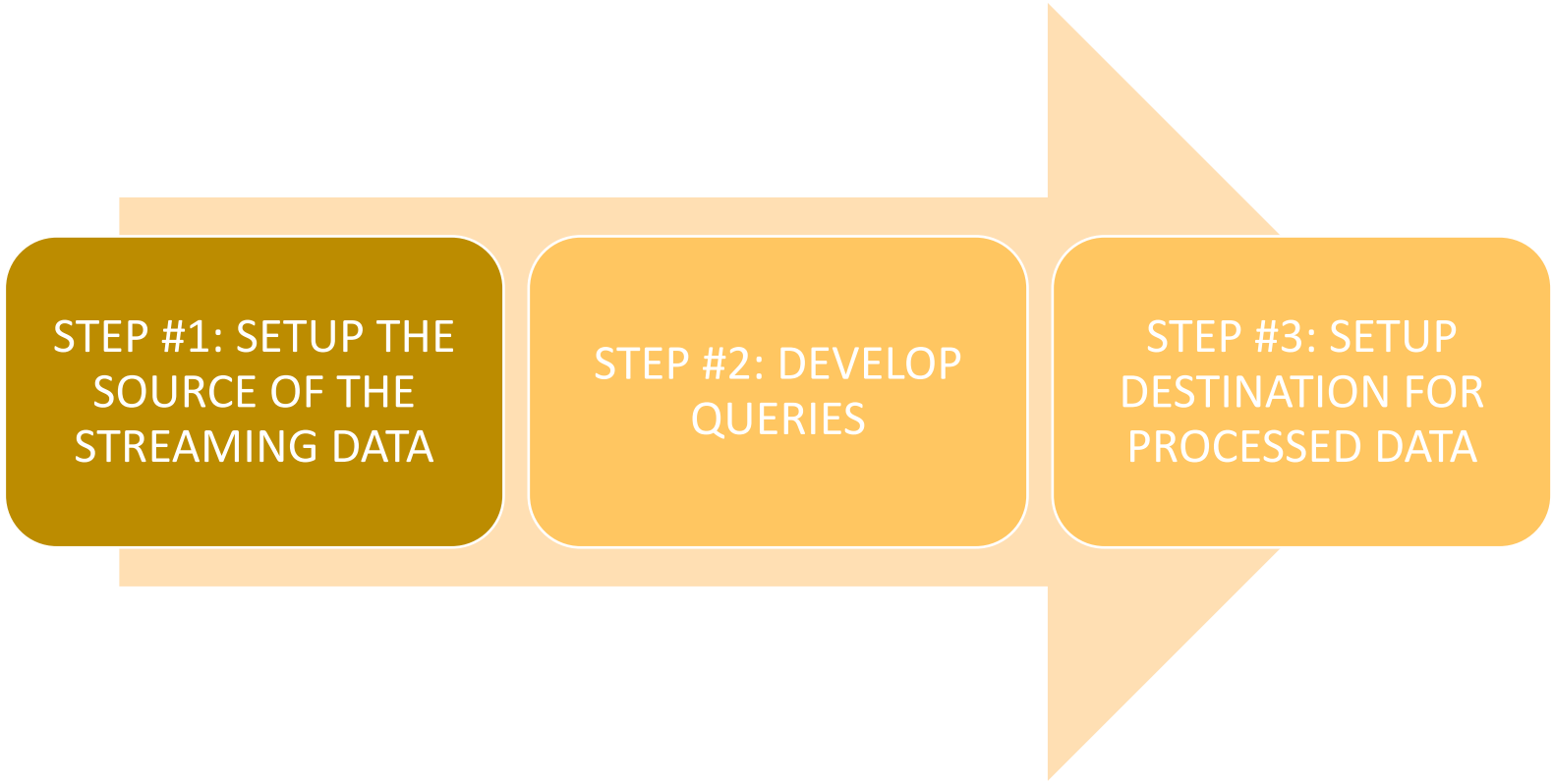
4. KINESIS DATA ANALYTICS: OVERVIEW

- Amazon Kinesis Data Analytics is a powerful tool to perform analysis on streaming data and get important insights.
- Amazon Kinesis data analytics run in Realtime and is fully automatically scalable based on incoming data throughput.
- Zero setup and upfront cost. Pay per use model.
- Available for SQL and JAVA developers:
  - **For SQL users/developers:**
    - Developers can query streaming data by leveraging readily available templates.
    - Developers can select a template for specific analytics task, edit code using SQL editor.

  - **For JAVA developers:**
    - Developers can leverage open source Java libraries to perform data analysis in Real-time
    - Java Library includes over 25 pre-built operators to perform streaming data aggregation, filtering and transformation.

4. KINESIS DATA ANALYTICS: BUILD IN 3 STEPS

- Continuously, Kinesis data analytics will process the data and send the results to destination.
- Amazon Kinesis Data Analytics streaming applications could be easily built using the three steps shown below:



4. KINESIS DATA ANALYTICS: BENEFITS

POWERFUL PERFOMRANCE

- Realtime performance with minimum latency.

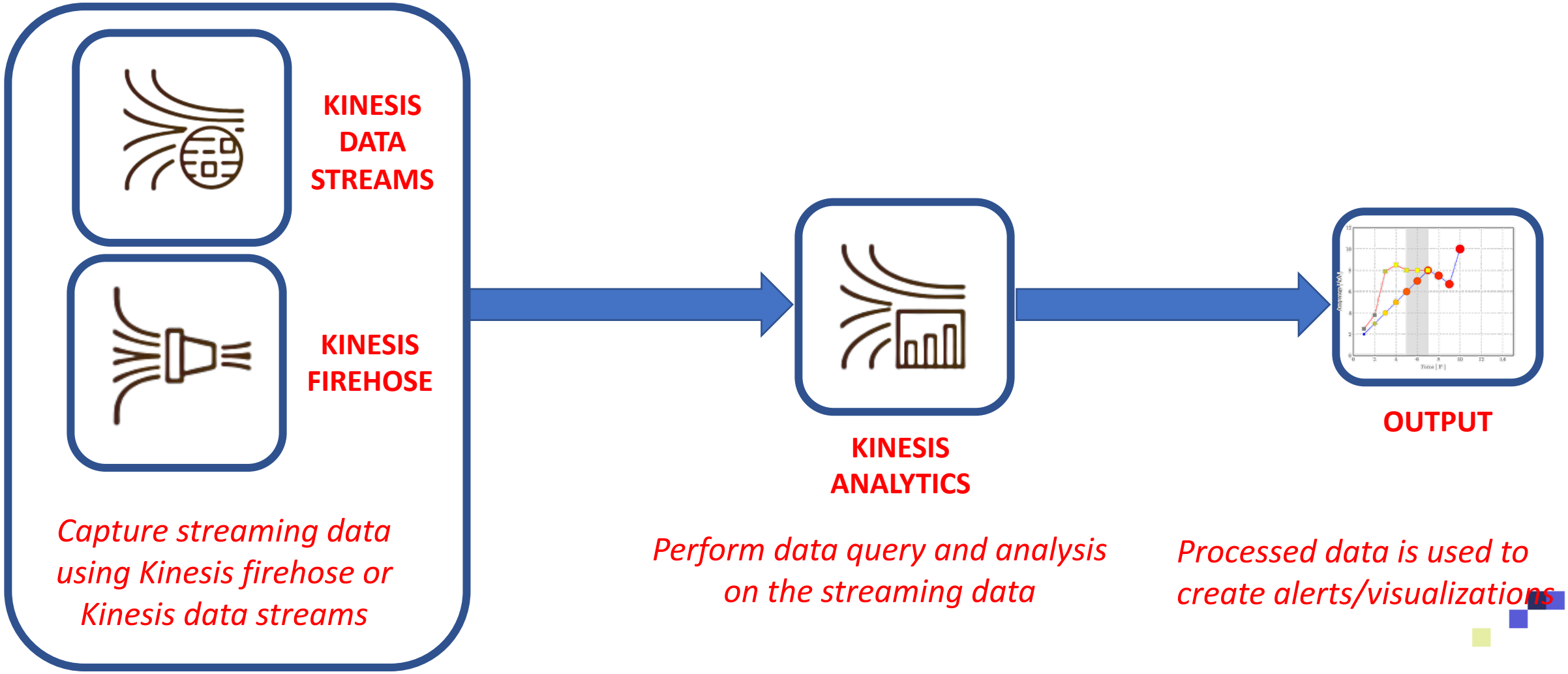
SERVERLESS

- Zero upfront cost or servers/infrastructure to manage.
- Automatic scaling to ensure best performance.

PAY PER USE

- No commitment and pay-per-use model offers a cost effective solution.

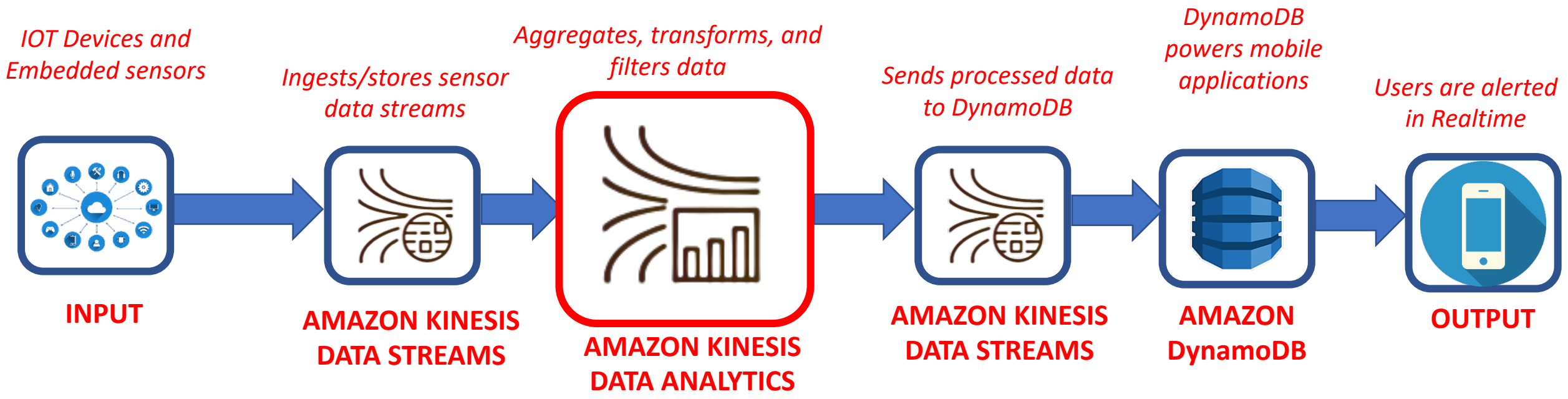
4. KINESIS DATA ANALYTICS: HOW IT WORKS?



AWS KINESIS DATA ANALYTICS – PART #2

4. KINESIS DATA ANALYTICS: USE CASE #1: STREAMING ETL FOR IOT

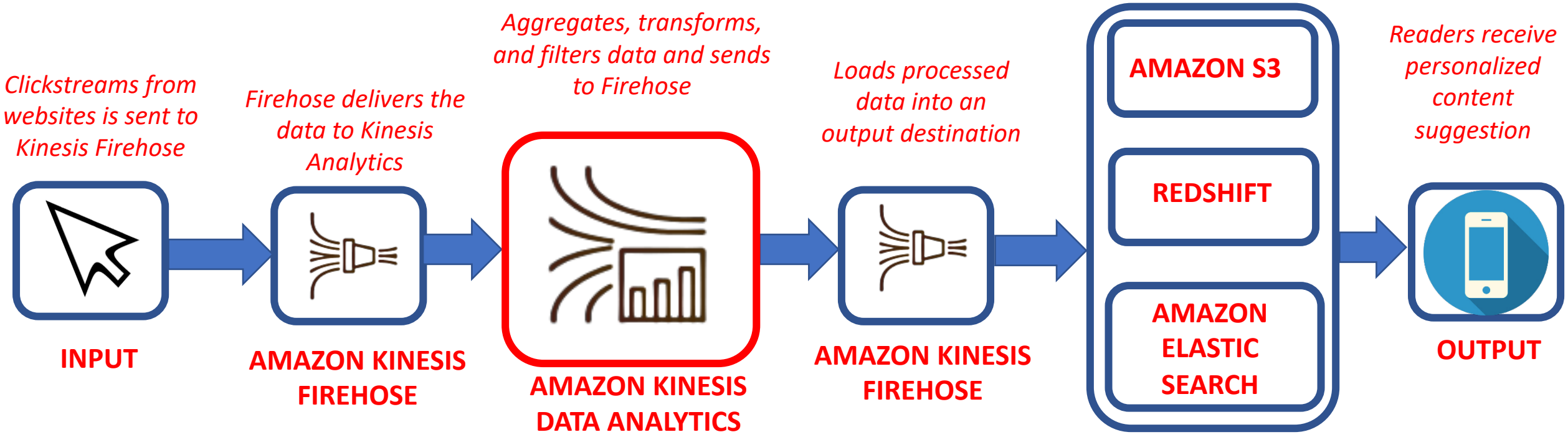
- Amazon Kinesis Data Analytics could be used to transform and filter streaming data from IOT devices such as sensors and then send real-time alerts when a variable exceeds a certain limit.



https://pixabay.com/illustrations/iot-internet-of-things-network-3337536/
Photo Credit: https://commons.wikimedia.org/wiki/File:Mobile-Smartphone-icon.png

4. KINESIS DATA ANALYTICS: USE CASE #2: REALTIME LOG ANALYTICS WITH SQL

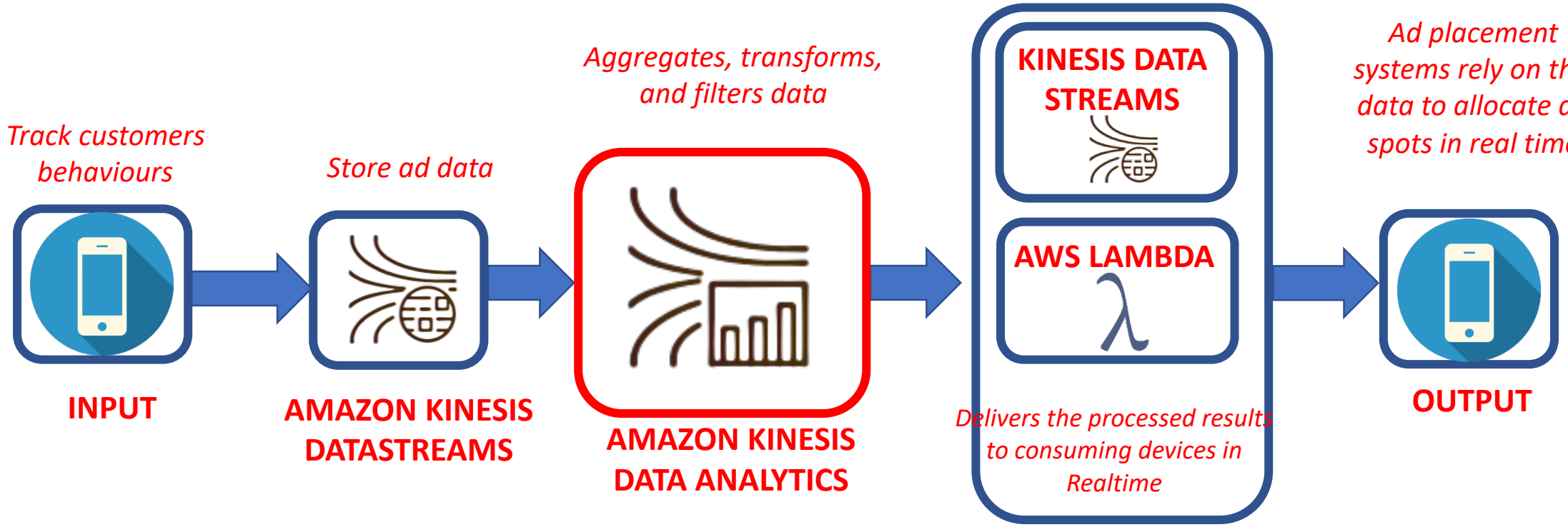
- Amazon Kinesis Data Analytics could be used to calculate metrics sent from users clickstreams (what did users click? How long did they stay on website?) and then present users with personalized content suggestions and targeted ads.



https://pixabay.com/illustrations/iot-internet-of-things-network-3337536/
Photo Credit: https://commons.wikimedia.org/wiki/File:Mobile-Smartphone-icon.png

4. KINESIS DATA ANALYTICS: USE CASE #3: DIGITAL MARKETING WITH SQL

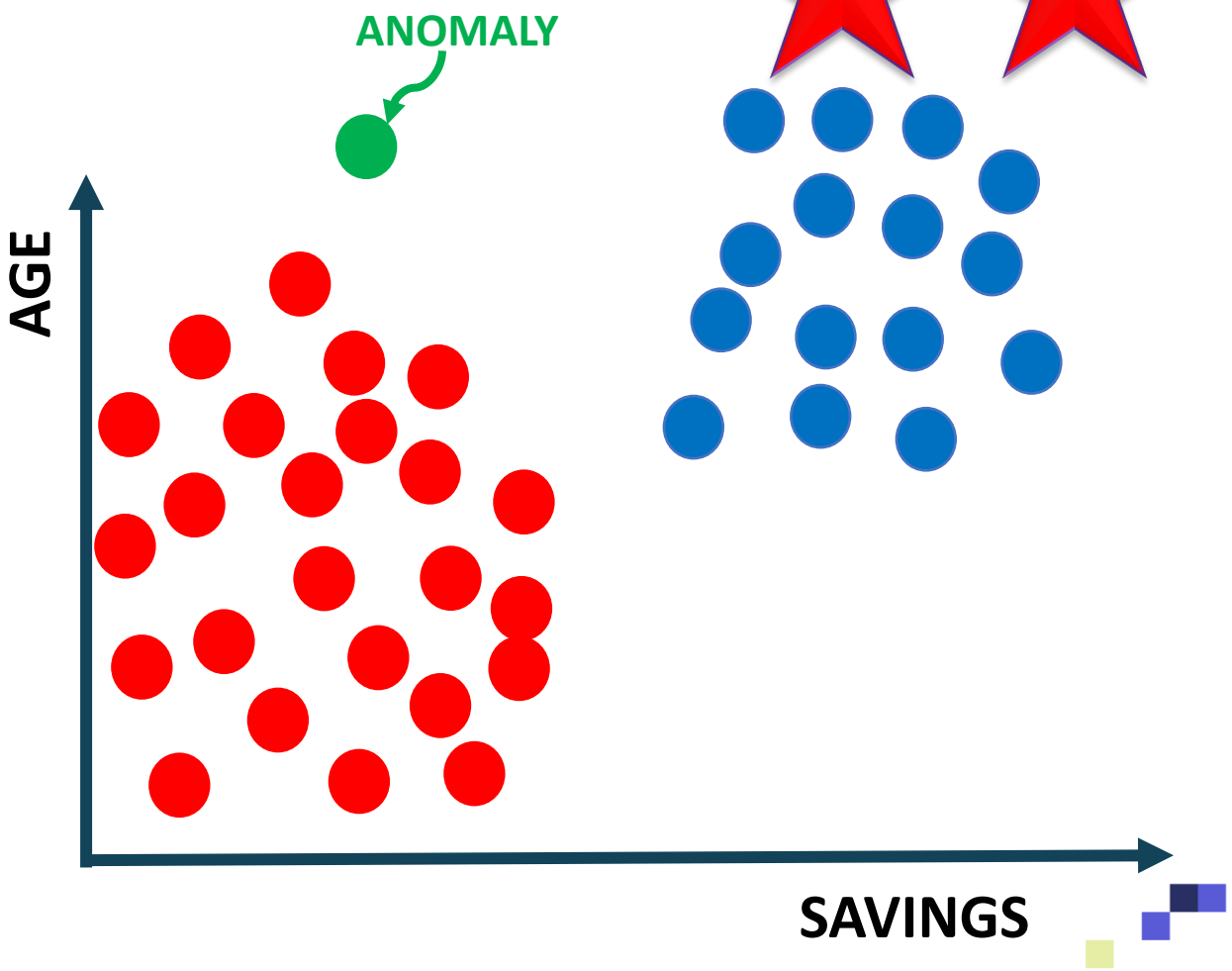
- Amazon Kinesis Data Analytics could be used to perform data transformation in Realtime to offer an optimized digital marketing solutions.



https://pixabay.com/illustrations/iot-internet-of-things-network-3337536/
Photo Credit: https://commons.wikimedia.org/wiki/File:Mobile-Smartphone-icon.png

4. KINESIS DATA ANALYTICS: MACHINE LEARNING RANDOM CUT FOREST

- Amazon Kinesis Data Analytics includes a function named RANDOM_CUT_FOREST that is used for anomaly detection.
- The function works by assigning an anomaly score to each data record.
- Anomalies are data points that diverge from the rest of the properly structured data.
- Anomalies could be spikes or breaks in the dataset and could be detected from the regular structured dataset.
- Anomaly detection and removal is crucial in machine learning because adding anomalies will unnecessarily increase the complexity of the model.



4. KINESIS DATA ANALYTICS: MACHINE LEARNING HOTSPOTS

- Amazon Kinesis Data Analytics includes a function named HOTSPOTS
- HOTSPOTS can be utilized to locate to identify dense areas in the data (activity in some regions that might be higher than the norm).
- By identifying “hot” regions, you can then focus the attention on these areas and take actions accordingly.

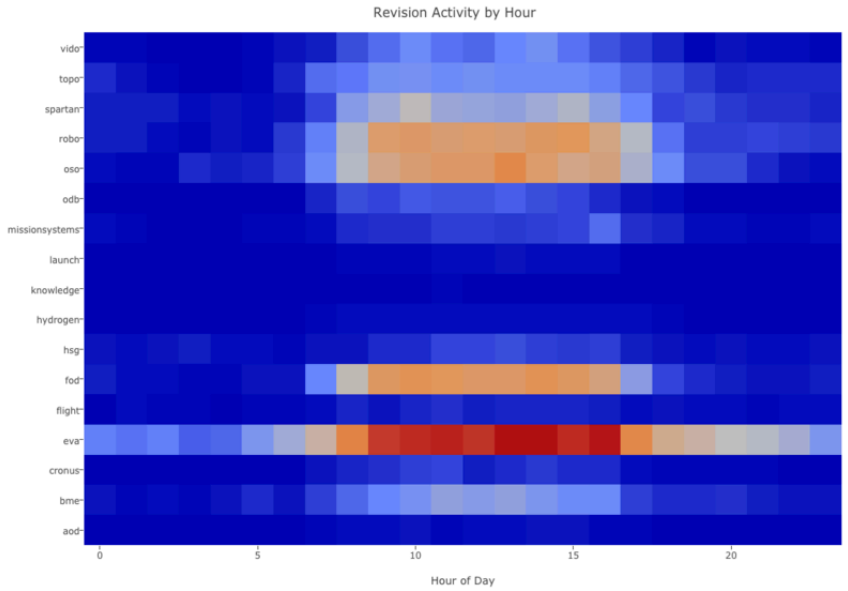


Photo Credit: https://commons.wikimedia.org/wiki/File:Heatmap_of_revision_activity_by_hour.png

KINESIS SUMMARY

KINESIS SERVICE	USE CASE
Kinesis Data Streams	For real-time streaming of data and to gain real-time insights using ML applications
Kinesis Firehose	For streaming data in near real time and storing them into S3 or redshift
Kinesis Video Streams	For streaming live videos in real time
Kinesis Analytics	To run SQL queries on streaming data and generate graphs and gain valuable metrics.