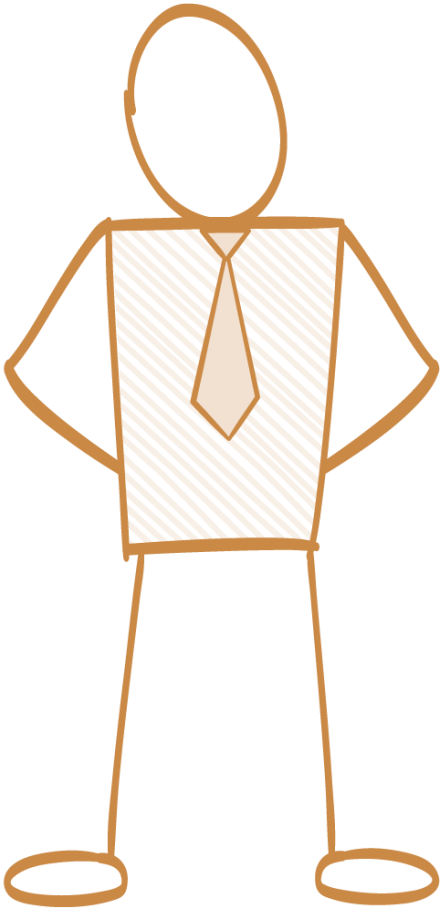
Architecting on AWS
Student Guide
Version 3.1

100-ARC-31-EN-SG



Module 11 : Overview of Serverless Services

AWS Solutions Architect Associate

Lambda

AWS Lambda Introduction

amazon web services

AWS Lambda

Run code **without provisioning or managing** servers.
Servers automatically start and stop when needed.
Serverless Functions. Pay per invocation.

amazon web services

Introduction to AWS Lambda

AWS Lambda is a compute service that lets you run code **without** **provisioning** or **managing** servers.

Lambda executes your code only when needed and scales automatically to a few to a 1000 lambda functions concurrently is seconds.

.

You pay only for the compute time you consume - there is no charge when your code is not running.

Lambda is **Cheap**

Lambda is **Serverless**

Lambda **Scales Automatically**

Natively supports 7 runtimes languages:

1. Ruby
2. Python
3. Java
4. Go
5. Powershell
6. NodeJs
7. C#

You can also create your own **custom runtime** environments







AWS Solutions Architect Associate

Lambda

Lambda Use Cases

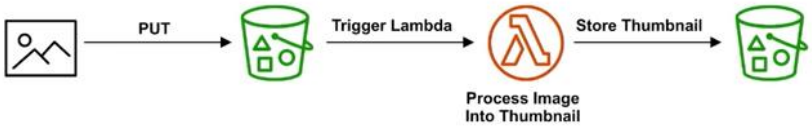
amazon web services

AWS Lambda – Use Cases

Lambda is commonly used to **glue different services together** so the use cases are endless.

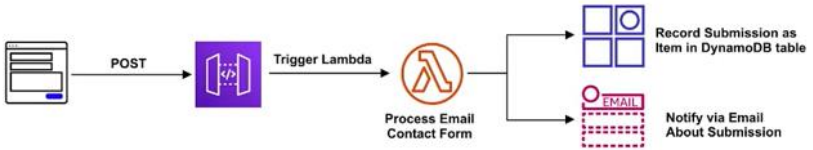
Processing Thumbnails

A web-service allows users to upload their profile photo. They are stored in an S3 bucket. We can setup an Event Trigger which will invoke a Lambda which will process the Profile Photo into a Thumbnail and the store it back in the bucket.



Contact Email Form

A company has a contact email form which submits form data via API Gateway Endpoint. That endpoint triggers a lambda which validates the form data and if valid will save the submission in DynamoDB and send and email notification via SNS to the company



AWS Solutions Architect Associate

Lambda

Lambda Triggers

amazon web services

AWS Lambda – Triggers

Lambdas can be **invoked** via the AWS SDK or or trigger from other AWS Services.
(This is not a complete list)

API Gateway
api application-services aws serverless

AWS IoT
aws devices iot

Alexa Skills Kit
alexa iot

Alexa Smart Home
alexa iot

Application Load Balancer
aws load-balancing

CloudFront
aws cdn edge

CloudWatch Events
aws events management-tools

CloudWatch Logs
aws logging management-tools

CodeCommit
aws developer-tools git

Cognito Sync Trigger
authentication aws identity mobile-services sync

DynamoDB
aws database nosql

Kinesis
analytics aws streaming

S3
aws storage

SNS
aws messaging notifications pub-sub push

SQS
aws queue

amazon web services

AWS Lambda – Triggers

Partner event sources (powered by Amazon EventBridge)

Datadog
Datadog is the essential monitoring platform for cloud applications.

OneLogin
OneLogin, the leader in Unified Access Management, connects people with technology through a simple and secure login, empowering organizations to access the world.

PagerDuty
PagerDuty helps AWS users automatically turn any signal into the right insight and action.

Saviynt
Saviynt enables enterprises to secure applications, data and infrastructure in a single platform for cloud and enterprise.

Segment
Segment provides the customer data infrastructure that businesses use to put their customers first.

SignalFx
SignalFx, the only real-time cloud monitoring platform for infrastructure, microservices, and applications, collects and analyzes metrics and traces across every component in your cloud environment.

SugarCRM
SugarCRM enables businesses to create extraordinary customer relationships with the most empowering, adaptable and affordable customer relationship management (CRM) solution on the market.

Whispir
Whispir is a cloud based platform that automates, personalises and layers communications using smart workflow technology.

Zendesk
Zendesk makes better customer service experiences for agents, admins, and customers.

AWS Solutions Architect Associate

Lambda

Lambda Pricing

amazon web services

AWS Lambda – Pricing

First **1 million requests** per month are free.
There-after **$0.20** per additional 1 million requests

400,000 GB seconds free per month
Thereafter **$0.0000166667** for every GB second

**This price will vary on the amount of memory you allocate

128MB of Memory **X** 30M executed per month **X** 200ms run time per invocation = **$5.83**

AWS Solutions Architect Associate

Lambda

Lambda Interface

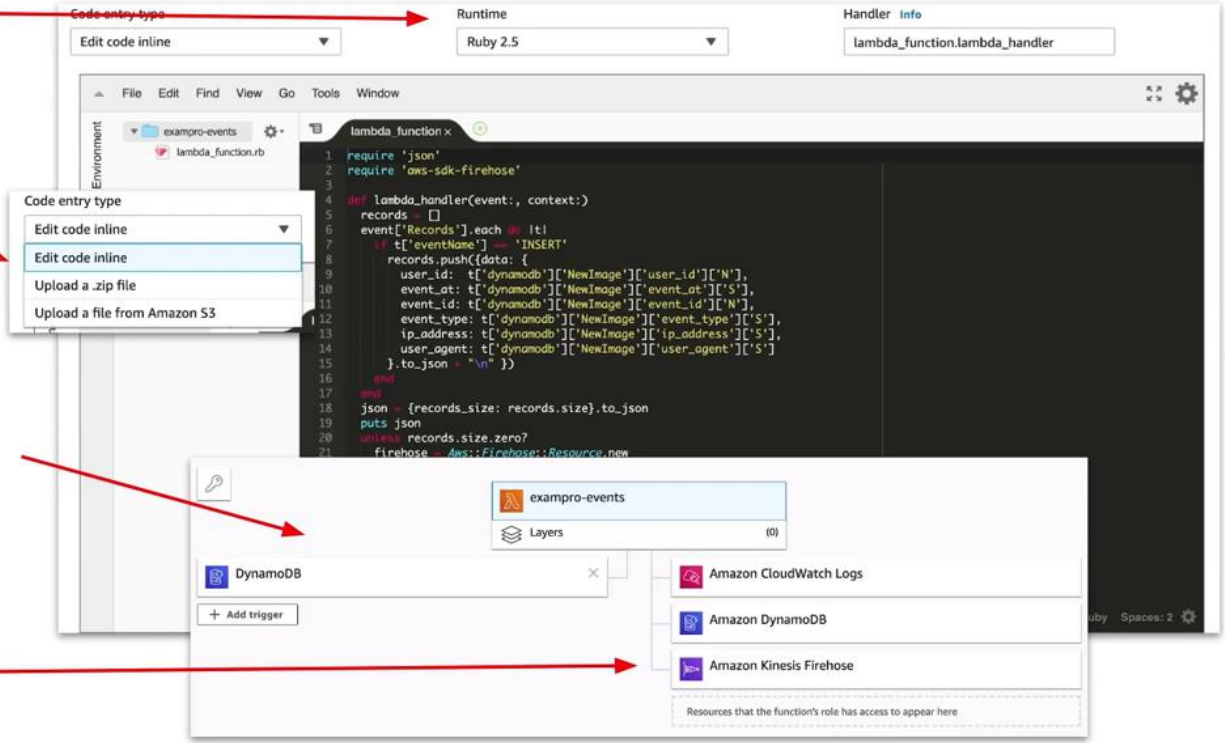
AWS Lambda - Interface

You choose you runtime

You upload Your code

You choose your triggers

You grant permissions for outputs via an IAM Role



AWS Solutions Architect Associate

Lambda

Lambda Defaults and Limits

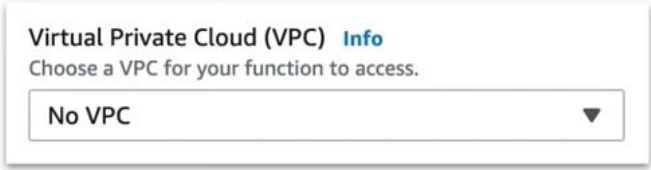
amazon web services

AWS Lambda – Defaults and Limits

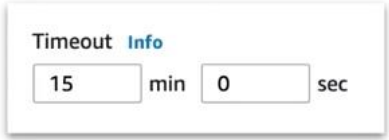
By Default you can have 1000 Lambda running concurrently (Ask AWS Support for Limit Increase)

/tmp directory can contain up to **500MB**

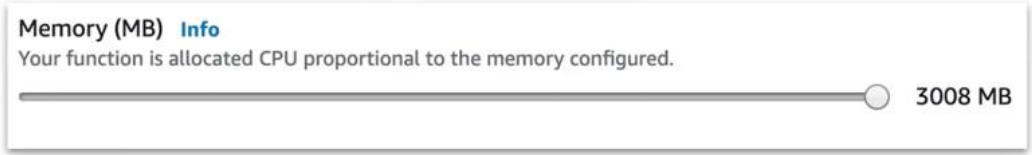
By Default Lambda run in No VPC. You can set them to by in your own VPC but your lambda will lose internet access



You can set timeout to be a maximum of **15 minutes**



Memory can be set between **128MB** to a Maximum of **3008MB** at an increment of **64MB**



amazon web services

AWS Solutions Architect Associate

Lambda

Lambda Cold Starts

AWS Lambda – Cold Starts

AWS has servers preconfigured (just sitting around turned off) for your runtime environment.
When a Lambda is invoked these servers need to be turned on and your code needs to be copied over.

During the time there will be a delay when the function will initially run which is called a **Cold Start**

If the same Lambda is invoked and the server is still running it will use that server again, so there will be little to delay to running that function. This what we call a **Warm Server**

Serverless functions are **cheap** but everything comes with a trade off. Serverless functions Cold Starts can **cause delays in the User Experience**. If your web-application relies on being very responsive, than you want to reconsider Serverless functions.

There are strategies around Cold Starts such as **Pre Warming** which keep servers continuously running. Cloud Providers are always looking for ways to reduce cold starts.

AWS Solutions Architect Associate

Lambda

Lambda Cheat Sheet

amazon web services

Lambda CheatSheet

- **Lambda's** are serverless **functions.** You upload your code and it runs without you managing or provisioning any servers.
- Lambda is **serverless**. You don't need to worry about underlying architecture
- Lambda is a good fit for short running tasks where you don't need to customize the os environment. If you need long running tasks (> 15mins) and a custom OS environment than consider using **Fargate**
- There are **7 runtime language environments** officially supported by Lambda: **Ruby, Python, Java, NodeJs, C#, Powershell** and **Go**
- You pay per invocation (The **duration** and the amount of **memory** used) rounded up to the nearest 100 milliseconds and you based on amount of requests. First 1M requests per month are free
- You can adjust the duration timeout for up to **15 mins** and memory up to **3008 MB**
- You can trigger Lambdas from the SDK or multiple AWS services eg. S3, API Gateway, DynamoDB
- Lambdas by default run in No VPC. To interact with some services you need to have your Lambda in the same VPC eg. RDS
- Lambda can scale to **1000 of concurrent functions** in seconds. (1000 is the default, you can increase with AWS Service Limit Increase)
- Lambdas have **Cold Starts.** If a function has not been recently been execute there will be a delay

AWS Solutions Architect Associate

API Gateway

API Gateway Introduction

amazon web services

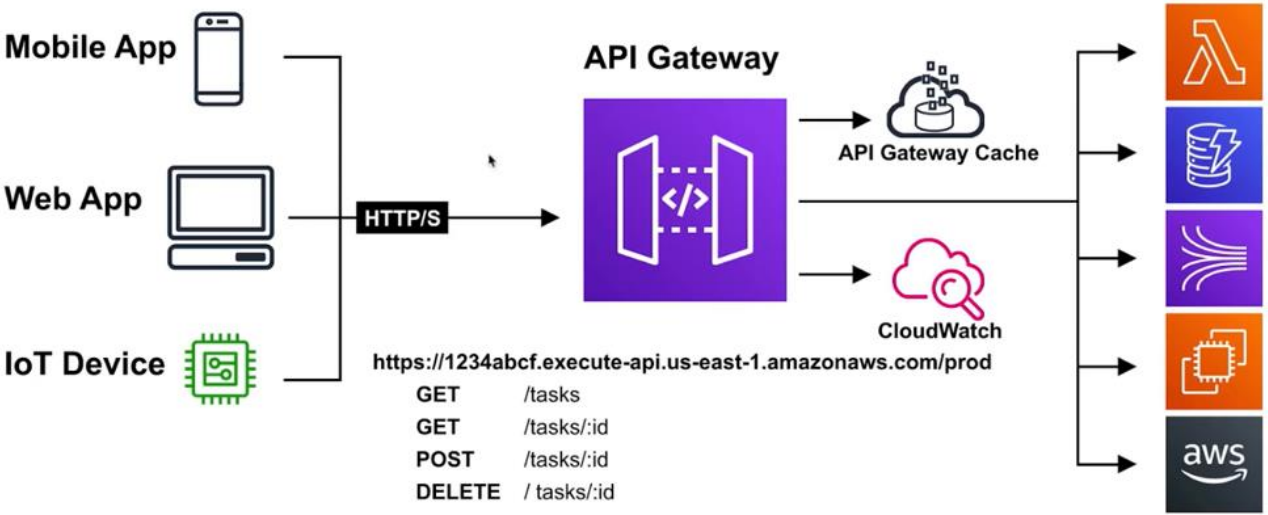
API Gateway

Fully managed service to create, publish, maintain, monitor, and secure APIs at any scale.

Introduction to API Gateway

API Gateway is a solution for **creating secure APIs** in your cloud environment at **any scale.**

Create APIs that act as a **front door** for applications to **access data, business logic,** or **functionality from back-end services.**



AWS Solutions Architect Associate

API Gateway

Key Features

amazon web services

API Gateway – Key Features

API Gateway handles all the tasks involved in accepting and processing **up to hundreds of thousands of concurrent API calls**, including traffic management, authorization, and monitoring.

Allows you to **track** and **control any usage** of the API. **Throttle** requests to help **prevent attacks**.

Expose HTTPS endpoints to define a **RESTful API.**

Highly scalable (**happens automatically**) and cost effective.

Send each API endpoint to a different target.

Maintain **Multiple Versions** of your API.

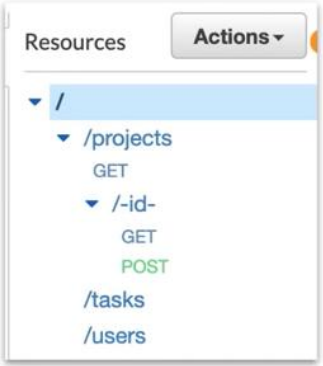
AWS Solutions Architect Associate

API Gateway

API Gateway Configuration

amazon web services

API Gateway – Configuration

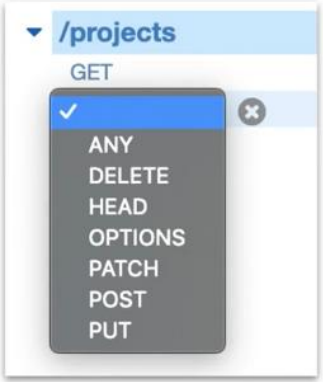


Resources

When you create an API you need to also create multiple **Resources.**

Resources are the urls you define eg. /projects

Resources can have child resources eg. /projects**/-id-/edit**



Methods

You need to define **Methods** on Resources

You can define multiple **Methods** on a Resource

Methods allow you to make API calls that resource url with that protocol

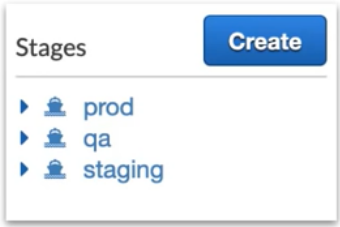
eg.

GET /projects/-id-

POST /projects/-id-

amazon web services

API Gateway – Configuration



Stages

In order to use your API you need to Deploy it to Stages

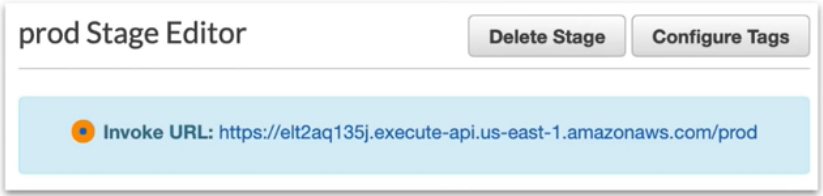
Stages are versions of your API

Invoke URL

For each stage AWS provides you a Invoke URL

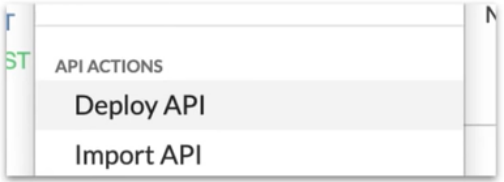
This is where you'll make your API calls.

It is possible to use a custom domain for your Invoke URL



Deploy API

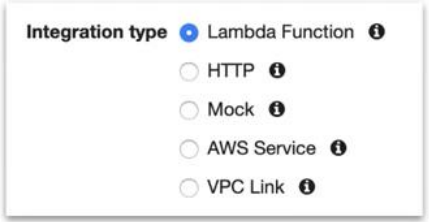
Everytime you make a change to your API you need to Deploy it via the **Deploy API** action. When you deploy you choose the stage

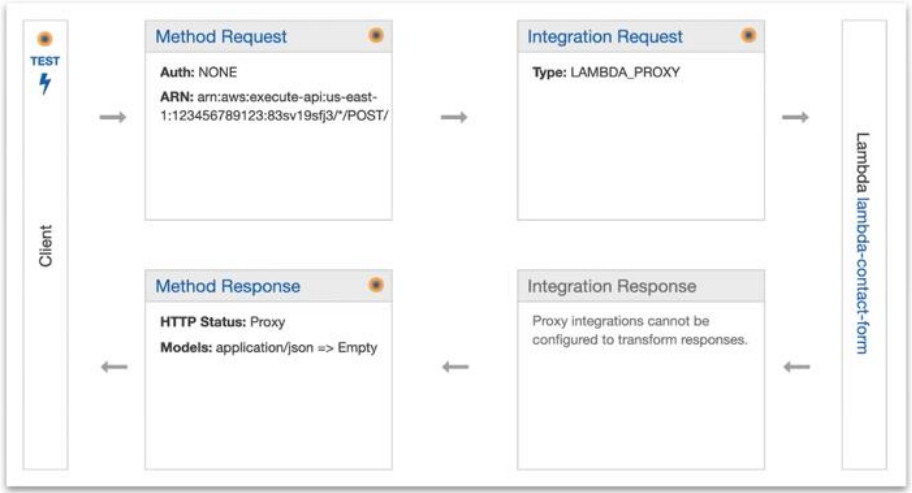


API Gateway – Configuration

When you create an API Method on a resource you need to choose the Integration type.

The most common Integration type is **Lambda**





You have fine tune control over the **Request** and **Response** for the Method Execution.

AWS Solutions Architect Associate

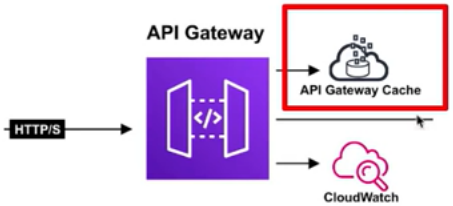
API Gateway

API Gateway Caching

amazon web services

API Gateway – Caching

API Caching can be enabled to cache your endpoints response to API calls.



- When enabled on a stage, API Gateway caches responses from your endpoint for a specified **time-to-live (TTL)** period.
- API Gateway responds to requests by **looking up the response from the cache**. (instead of making a request to the endpoint)

Reduces the number of calls made to your endpoint.

Improves latency of the requests made to your API.

AWS Solutions Architect Associate

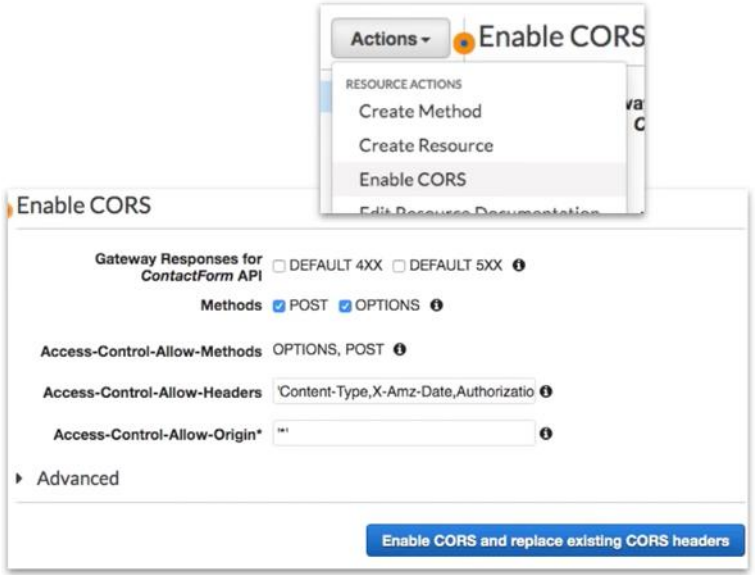
API Gateway

Cross-Origin Resource Sharing (CORS)

amazon web services

API Gateway – CORS

Cross-Origin Resource Sharing (CORS) is a way that the server at the other end (not client code in the browser) can relax a same-origin policy.



- Allows restricted resources (ie Fonts) on a webpage to be requested from a **different domain than the initial resource** that it came from.
- Should always be enabled if using Javascript/AJAX that uses multiple domains with an API gateway.

CORS is **always** enforced by the client.

AWS Solutions Architect Associate

API Gateway

Same Origin Policy

amazon web services

API Gateway – Same Origin Policy

Same Origin Policy is a concept in the application security model, where a web browser permits **scripts contained in a first web page,** to access **data in a second webpage.**

- Same Origin Policies are used to help prevent Cross-Site Scripting (XSS) attacks.
- They only works if both web pages have the same origin
- They are enforced at the web browser level
- They ignore tools such as Postman or Curl

AWS Solutions Architect Associate

API Gateway

API Gateway Cheat Sheet

amazon web services

API Gateway CheatSheet

- API Gateway is a solution for creating secure APIs in your cloud environment at any scale.
- Create APIs that act as a front door for applications to access data, business logic, or functionality from back-end services.
- API Gateway throttles api endpoints at **10,000** requests per second (can be increase via service request through AWS support)
- **Stages** allow you to have multiple published versions of your API eg. prod, staging, QA
- Each Stage has an **Invoke URL** which is the endpoint you use to interact with your API
- You can use a custom domain for your Invoke URL eg. api.exampro.co
- You need to publish your API via Deploy API. You choose which Stage you want to publish your API
- Resources are your URLs eg. /projects
- Resources can have child resources eg. /projects**/-id-/edit**
- You defined multiple Methods on your Resources eg GET, POST, DELETE
- CORS issues are common with API Gateway, CORS can be enabled on all or individual endpoints
- Caching improves latency and reduces the amount of calls made to your endpoint
- Same Origin Policies help to prevent XSS attacks
- Same Origin Policies ignore tools like postman or curl
- CORS is always enforced by the client.
- You can require Authorization to your API via AWS Cognito or a custom Lambda.

AWS Solutions Architect Associate

ElastiCache

ElastiCache Introduction

amazon web services

ElastiCache

Managed caching service which
either runs Redis or Memcached

What is In-Memory Data Store?

Caching

Caching is the process of storing data in a cache. A cache is a **temporary storage** area. Caches are optimized for fast retrieval with the trade off that data is not durable.

In-Memory Data Store

When data is stored In-Memory (think of RAM). The trade off is high volatility (low durability, risk of data loss) but **access** to data is **very fast**.

Introduction to ElastiCache

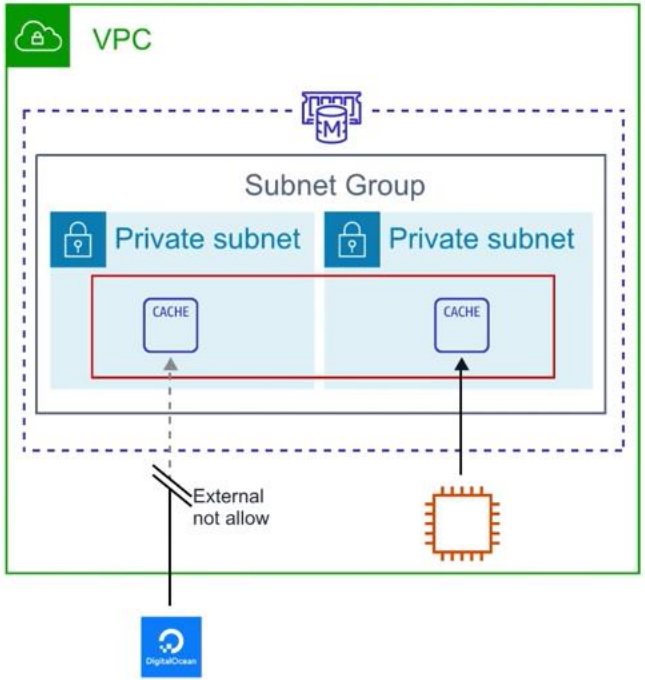
Deploy, run, and scale **popular open source compatible** **in-memory** data stores.

Frequently identical queries are stored in the cache.

ElastiCache is only accessible to resource operating with the same VPC to ensure low latency

ElastiCache supports 2 open-source caching engines:

1. Memcached
2. Redis



AWS Solutions Architect Associate

ElastiCache

Caching Comparison

amazon web services

ElastiCache – Caching Comparison

Memcached is generally preferred for caching HTML fragments. Memachce is a simple key/value store. The trade off it to being simple is that its very fast

Redis can perform many different kinds of operations on your data. It's very good for leaderboards, keep track of unread notification data. It's very fast, but **arguably** not as fast as Memcached.

Don't google "Memcache vs Redis" unless you want to read endless arguments as if people are arguing "Kirk vs Picard"

	Memcached	Redis
Sub-millisecond latency	Yes	Yes
Developer ease of use	Yes	Yes
Data partitioning	Yes	Yes
Support for a broad set of programming languages	Yes	Yes
Advanced data structures	—	Yes
Multithreaded architecture	Yes	—
Snapshots	—	Yes
Replication	—	Yes
Transactions	—	Yes
Pub/Sub	—	Yes
Lua scripting	—	Yes
Geospatial support	—	Yes

AWS Solutions Architect Associate

ElastiCache

ElastiCache Cheat Sheet

amazon web services

ElastiCache *CheatSheet*

- **ElastiCache** is a managed **in-memory** caching service
- ElastiCache can launch either **Memcached** or **Redis**
- **Memcached** is a simple key / value store preferred for caching HTML fragments and is arguably faster than Redis
- **Redis** has richer data types and operations. Great for leaderboard, geospatial data or keeping track of unread notifications.
- A cache is a **temporary storage** area.
- Most frequently identical queries are stored in the cache
- Resources only **within the same VPC** may connect to ElastiCache to ensure low latencies.

AWS Solutions Architect Associate

Kinesis

Kinesis Introduction

amazon web services

Amazon Kinesis

Scalable and durable **real-time** **data streaming** service
To ingest, and analyze data in real-time from multiple source

amazon web services

Introduction to Kinesis

Amazon Kinesis is the AWS fully managed solution for **collecting**, **processing**, and **analyzing streaming data** in the cloud.

When you need **"real-time"** think Kinesis.

Streaming Data Examples

- Stock Prices
- Game Data (as the player plays)
- Social Network Data
- Geospatial Data
- Click Stream Data

There are **4** different types of Kinesis Streams

- **Kinesis Data Streams**
- **Kinesis Firehose Delivery Streams**
- **Kinesis Data Analytics**
- **Kinesis Video Analytics**

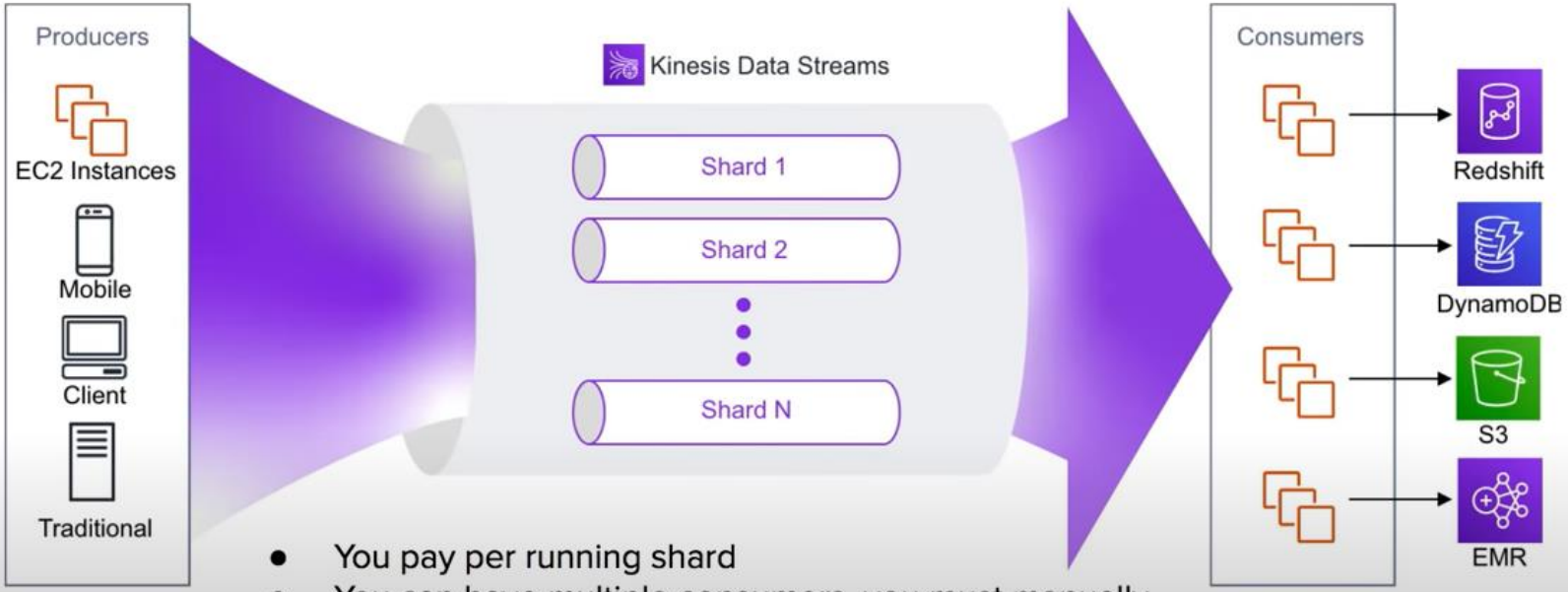
AWS Solutions Architect Associate

Kinesis

Kinesis Data Streams

amazon web services

Kinesis – Data Streams



- You pay per running shard
- You can have multiple consumers, you must manually configure your consumers.
- Data can be persist from **24 hours (default)** to 168 hours before it disappears from the stream

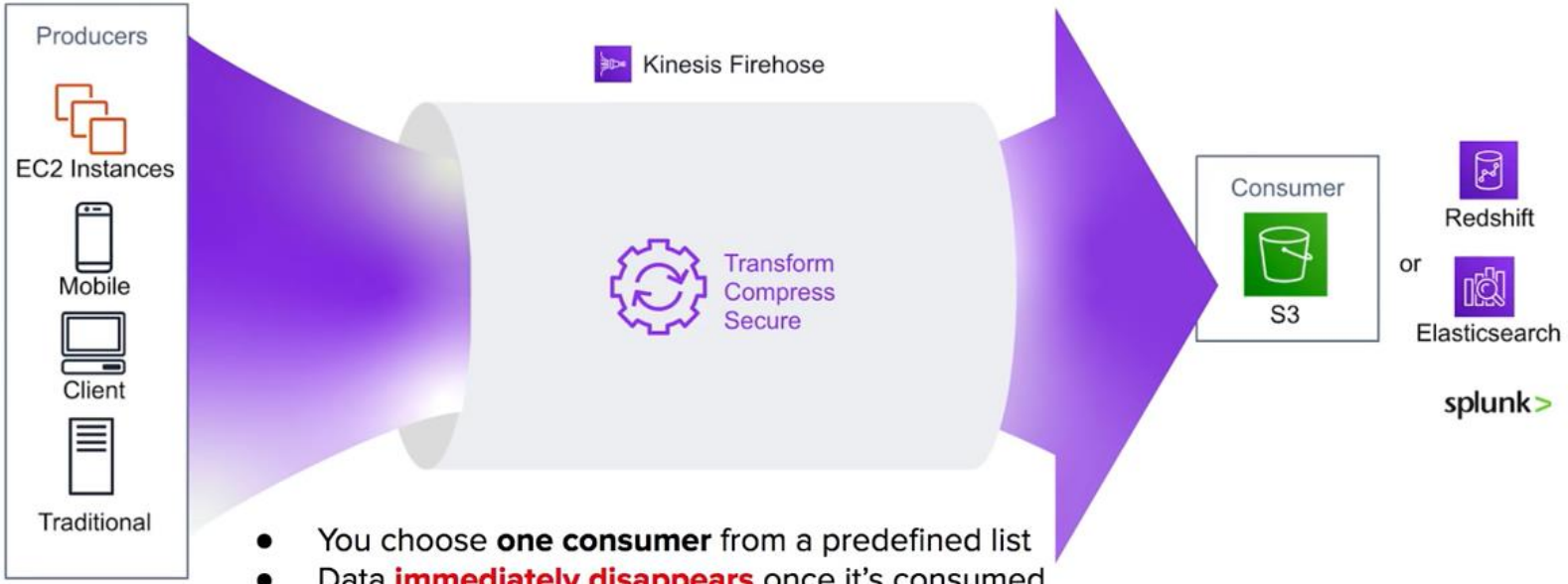
AWS Solutions Architect Associate

Kinesis

Kinesis Data Firehose

amazon web services

Kinesis – Firehose Delivery Stream



- You choose **one consumer** from a predefined list
- Data **immediately disappears** once it's consumed
- You can convert incoming data to other to a few files formats, compress and secure data.
- You pay only for data that is ingested

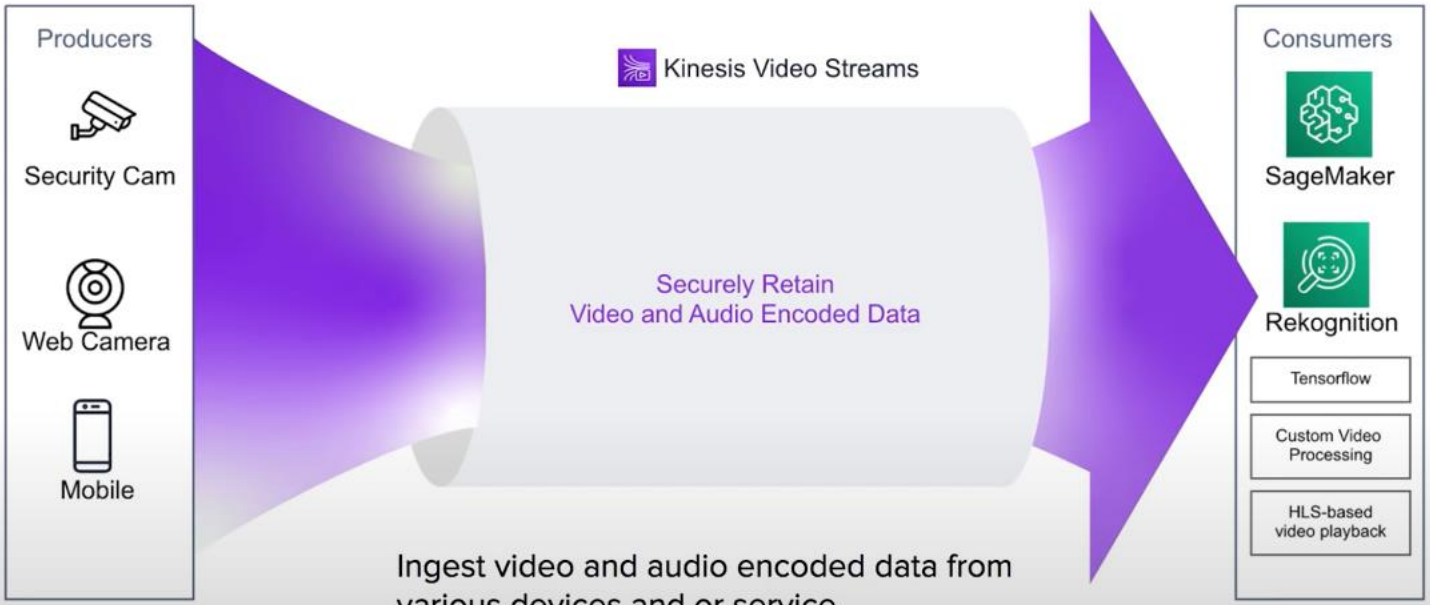
AWS Solutions Architect Associate

Kinesis

Kinesis Video Streams

amazon web services

Kinesis – Video Streams



Ingest video and audio encoded data from various devices and or service.
Output video data to to ML or video processing services

AWS Solutions Architect Associate

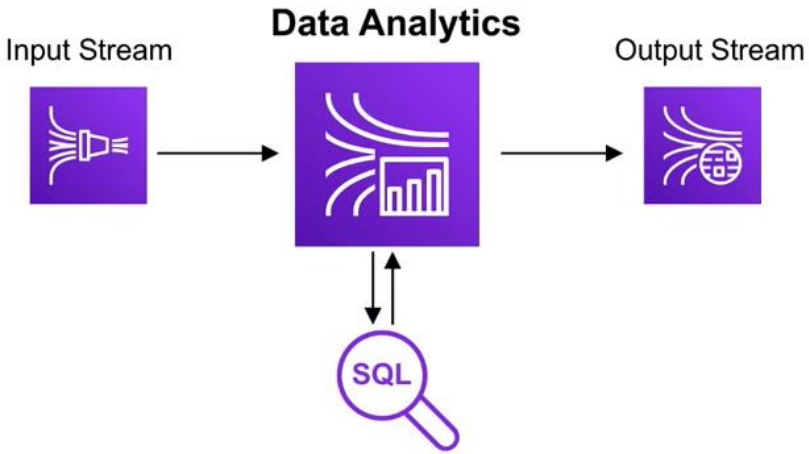
Kinesis

Kinesis Data Analytics

amazon web services

Kinesis – Data Analytics

You can specific Firehose or Data Streams as an input and an output.
Data that pass through Data Analytics is run through **custom SQL**
you provide and the results are then output.
This allows for real-time analytics of your data.



AWS Solutions Architect Associate

Kinesis

Kinesis Cheat Sheet

amazon web services

Kinesis *CheatSheet*

- **Amazon Kinesis** is the AWS solution for **collecting**, **processing**, and **analyzing** **streaming data** in the cloud. When you need **"real-time"** *think* Kinesis.
  Kinesis Data Streams Per per running shard, data can persist within the stream, data is ordered and every consumer keep its own position. Consumers have to be manually added (coded), Data persists for **24 hours (default)** to **168 hours**
- **Kinesis Firehose** - Pay for only the data ingested, data **immediately disappears** once processed. Consumer of choice is from a predefined set of services: S3, Redshift, Elasticsearch or Splunk
- **Kinesis Data Analytics** - allows you to perform **queries in real-time**. Needs a Kinesis Data Streams/Firehose as the input and output.
- **Kinesis Video Analytics** securely ingests and stores video and audio encoded data to consumers such as SageMaker, Rekognition or other services to apply Machine learning and video processing.
- KPL (Kinesis Producer Library) is a Java library to write data to a stream
- You can write data to stream using AWS SDK, but KPL is more efficient