AWS

MACHINE LEARNING

CERTIFICATION

DOMAIN #1: DATA ENGINEERING (20% EXAM)

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

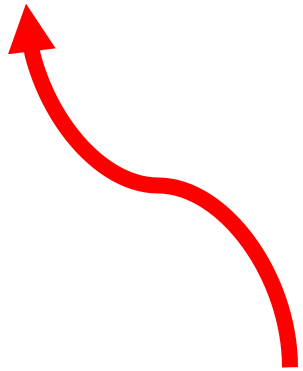
DOMAIN #1: WHERE ARE WE NOW!!?

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)



WE ARE HERE!

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

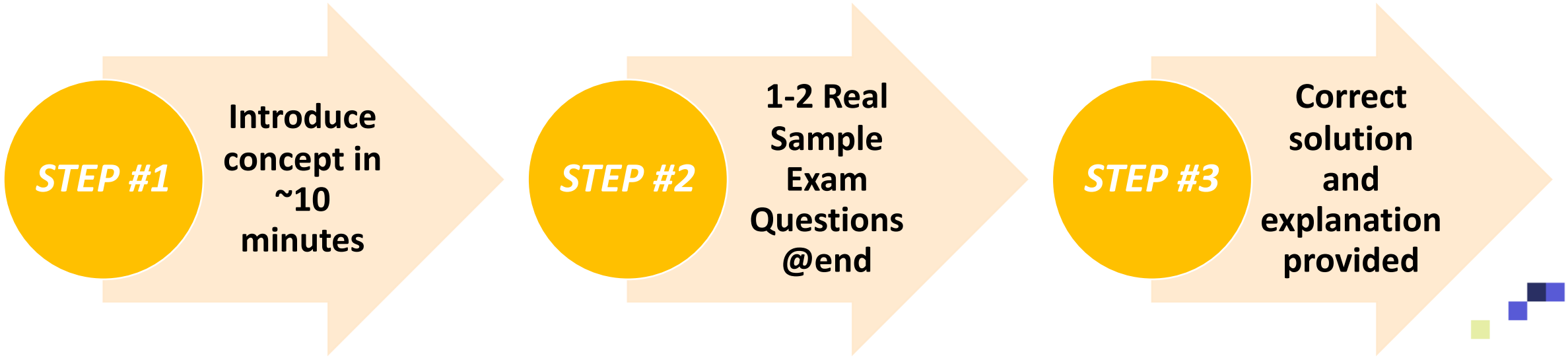
- Kinesis Overview
- Kinesis Video Streams
- Kinesis Data Streams
- Kinesis Firehose
- Kinesis Analytics and Random Cut Forest

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/

AMAZON S3

WHAT IS AMAZON S3?

- Amazon Simple Storage Service (Amazon S3) is a storage service that allows enterprises/individuals to store and protect any amount of data.
- Amazon S3 offers numerous enhanced features such as:
  - (1) Scalability
  - (2) Data availability
  - (3) Security
  - (4) Performance
- Amazon S3 is extremely easy to use and allows enterprises to organize their data and configure finely-tuned access controls.
- Amazon S3 extremely durable to 99.999999999% (11 9's).

Photo Credit: https://commons.wikimedia.org/wiki/File:AWS_Simple_Icons_AWS_Cloud.svg
Photo Credit: https://commons.wikimedia.org/wiki/File:AWS_Simple_Icons_Storage_Amazon_S3.svg

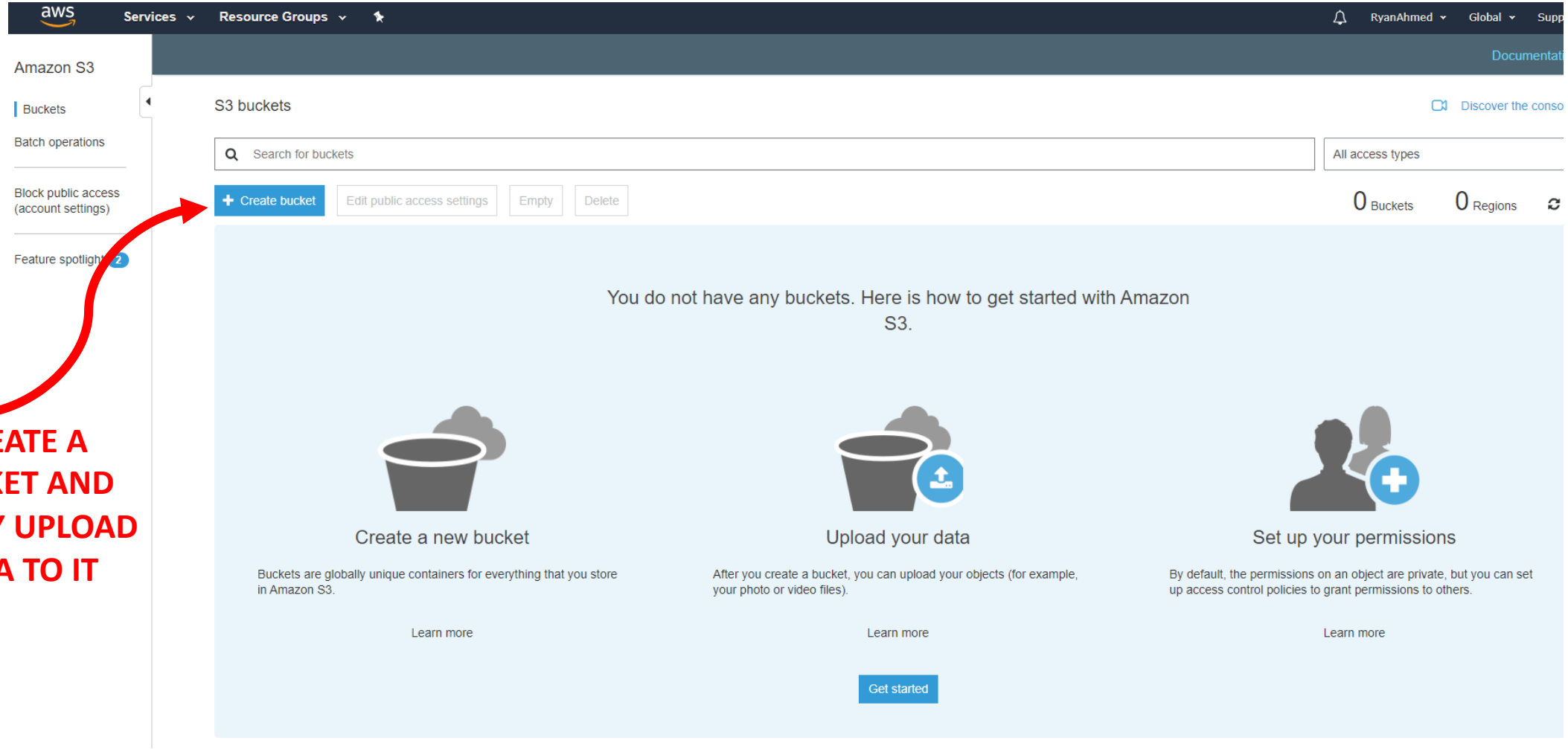
WHAT IS AMAZON S3? CONTINUED

- Amazon Simple Storage Service (Amazon S3) is built to be extremely simple and robust.
- Amazon S3 allows customers to store data in buckets or directories (much like folders).
- A bucket is a container for objects stored in Amazon S3. Every object is contained in a bucket.
- Each of the buckets will have **global unique name**.
- You can store an infinite amount of data in a bucket in which each object can contain up to 5 TB of data.
- For example, if we have an object **images/mycat.jpg** is stored in the **mitchsteve** bucket, use can use the following URL to access it:

http://mitchsteve.s3.amazonaws.com/images/mycat.jpg

Photo Credit: https://commons.wikimedia.org/wiki/File:AWS_Simple_Icons_AWS_Cloud.svg
Photo Credit: https://commons.wikimedia.org/wiki/File:AWS_Simple_Icons_Storage_Amazon_S3.svg

WHAT IS AMAZON S3?



CREATE A BUCKET AND SIMPLY UPLOAD DATA TO IT

OBJECT TAGS

- You can use object Tags (key value pairs) to categorize storage.
- Here's an example of an object tag:
  - Project = Machinelearning
  - Classification = confidential
  - PHI = True
- The maximum number of tags per object is 10.
- Objects tags are important for:
  - Granting or denying permission (for example: read only user access).
  - Managing object lifecycle by creating a lifecycle rule based on associated tags.
  - Performing analytics.

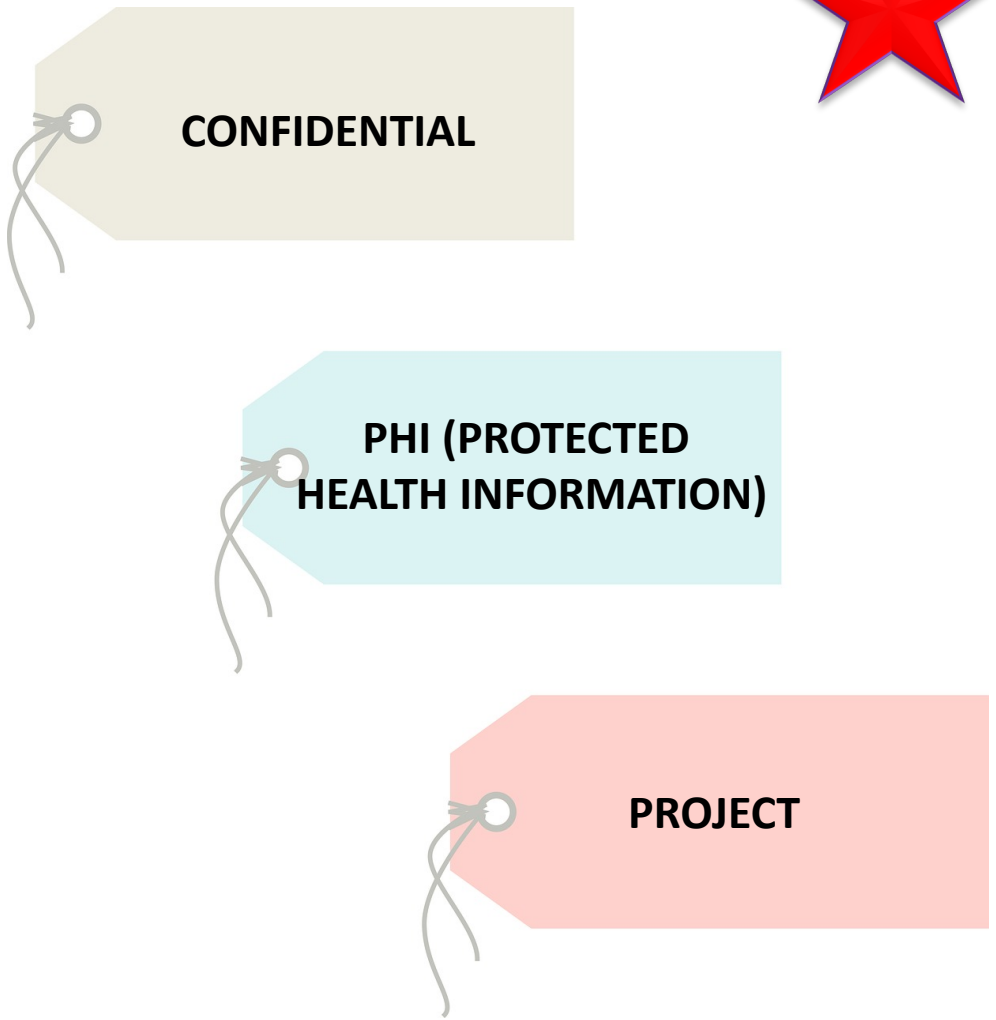


Photo Credit: https://www.needpix.com/photo/1265094/tag-label-price-hang-tag-ticket-string-retail-business-promotion

AWS S3 DATA LAKE

- Amazon S3 offers a great, easy to use solution to create a data lake because its highly scalable.
- Amazon S3 allows enterprises and individuals to increase their storage from gigabytes to petabytes in minutes.
- It is cost effective and pay-per-use model is very efficient. You do not need to buy or administer any hardware.
- Amazon S3 is extremely durable with 99.999999999%.
- Offers easy to use access controls so users can grant and deny fine grained access.
- Amazon S3 works with common formats such as: CSV, Parquet, ORC, Avro, Protobuf and JSON.
- AWS S3 hosts the data that machine learning services such as SageMaker will use for model training and testing.

AWS S3 DATA LAKE: ENABLING FEATURES

COMPUTE AND DATA PROCESSING ARE DECOUPLED

- Storage and compute are coupled in most data warehousing solutions (Hadoop) increasing complexity and cost.
- Amazon S3 offers a cost effective solution to store any data of any size in its native format.
- Users can launch Amazon Elastic Compute (EC2) cloud instances to access and process data.

CENTRALIZED DATA ARCHITECTURE

- Traditional solutions forces enterprises to have multiple copies of data distributed across multiple processing platforms.
- S3 overcomes that and offers a “multi-tenant” environment.
- Include one copy of the data and many users can apply their tools on the same data.

INTEGRATION WITH CLUSTERLESS/SERVERLESS AWS SERVICES

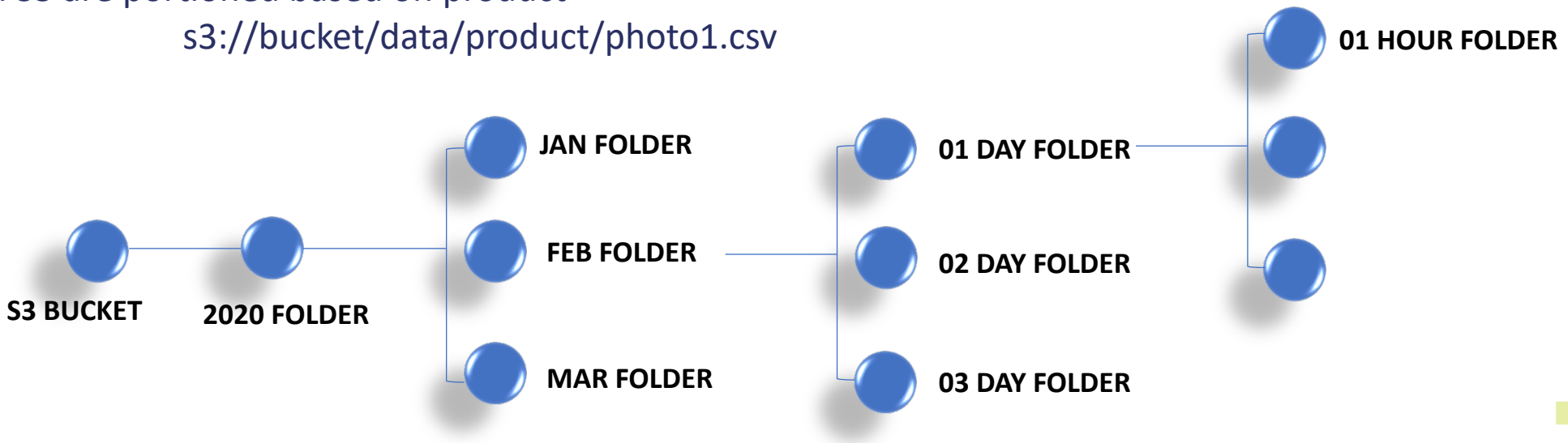
- Amazon S3 is seamlessly integrated with other services to query/process data such as Athena, Redshift, Rekognition, AWS Glue.
- Amazon S3 integrates with AWS Lambda serverless computing to run code without provisioning or managing servers.

STANDARDIZED APIS

- Amazon S3 RESTful APIs easy to and work well with Apache Hadoop and many analytics tools.
- Users can leverage their skills in certain analytics tool but using data available in Amazon S3.

AWS S3 DATA PARTITIONING

- AWS S3 data partitioning is critical when querying data because it could dramatically reduce the cost required for scanning.
- Partitioning is important when data query in Amazon Athena an Redshift ($5/TB scanned).
- Data could be partitioned by time, product or customized by user.
- Examples:
  - Folders on S3 are partitioned based year/month/day/hour
    s3://bucket/data/year/month/day/hour/photo1.csv
  - Folders on S3 are portioned based on product
    s3://bucket/data/product/photo1.csv

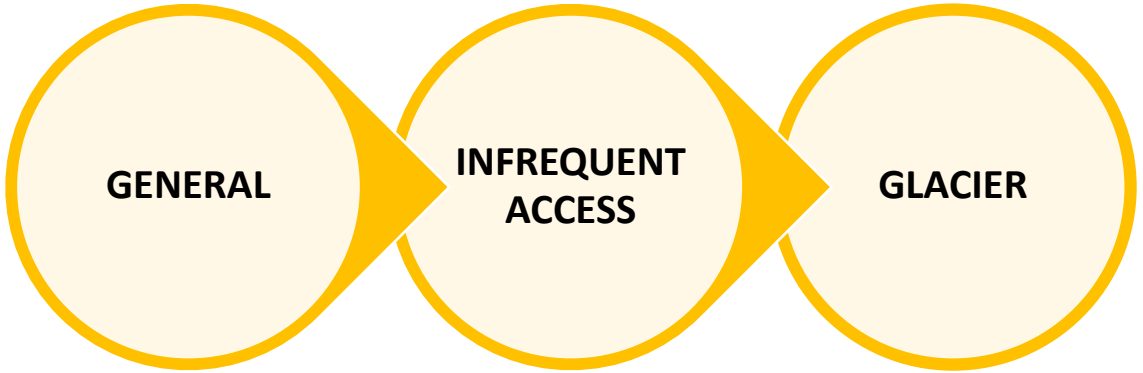


AMAZON S3 STORAGE TIERS AND LIFECYCLES

AWS S3 STORAGE TIERS

- Amazon S3 offers a range of storage classes:
  - S3 Standard: works well with storage that is general purpose and frequently accessed
  - S3 Intelligent-Tiering: works well with data that has varying access patterns
  - S3 Standard-Infrequent Access (Standard-IA): for long-lived but less frequently accessed data
  - S3 One Zone-Infrequent Access (One Zone-IA): for long-lived, but less frequently accessed data
  - Amazon S3 Glacier (S3 Glacier) and Amazon S3 Glacier Deep Archive (S3 Glacier Deep Archive): works well for long-term archived data.

Amazon S3 offers the ability to change the storage tiers throughout the data lifecycle by setting a lifecycle policy.



AWS S3 STORAGE TIERS SUMMARY

	FREQUENTLY ACCESSED	CHANGING PATTERN	INFREQUENTLY ACCESSED		ARCHIVED	
	S3 Standard	**S3 Intelligent-Tiering***	**S3 Standard-IA**	**S3 One Zone-IA†**	**S3 Glacier**	**S3 Glacier Deep Archive**
Designed for durability	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)
Designed for availability	99.99%	99.9%	99.9%	99.5%	99.99%	99.99%
Availability Zones	≥3	≥3	≥3	1	≥3	≥3
Minimum storage duration charge	N/A	30 days	30 days	30 days	90 days	180 days
Retrieval fee	N/A	N/A	per GB retrieved	per GB retrieved	per GB retrieved	per GB retrieved
First byte latency	milliseconds	millseconds	milliseconds	milliseconds	select minutes or hours	select hours
Storage type	Object	Object	Object	Object	Object	Object
Lifecycle transitions	Yes	Yes	Yes	Yes	Yes	Yes

AWS S3 LIFECYCLES

- In order to reduce cost, you can configure a lifecycle policy that S3 could apply to group of objects.
- Two types of actions are present:

    - **Transition actions:** policy that governs when an object transitions from one storage class to another.
        - transition object from STANDARD_IA storage class 30 days creation
        - Transition objects to the GLACIER storage class after 1 year.

    - **Expiration actions:** policy that governs when objects expire (deleted by S3).

- You can set the data to be archived once uploaded (Glacier storage class) for some cases such as:
    - Data that need to be retained for compliance purposes
    - Healthcare records
    - Long-term database backups

HOW TO CONFIGURE AWS S3 LIFECYCLES?

- Lifecycles are managed via a lifecycle configuration file (XML file).
- XML files consists of rules/polices that S3 can apply to objects.
- For each bucket, Amazon S3 keeps the configuration file attached to it.

AMAZON S3 ENCRYPTION

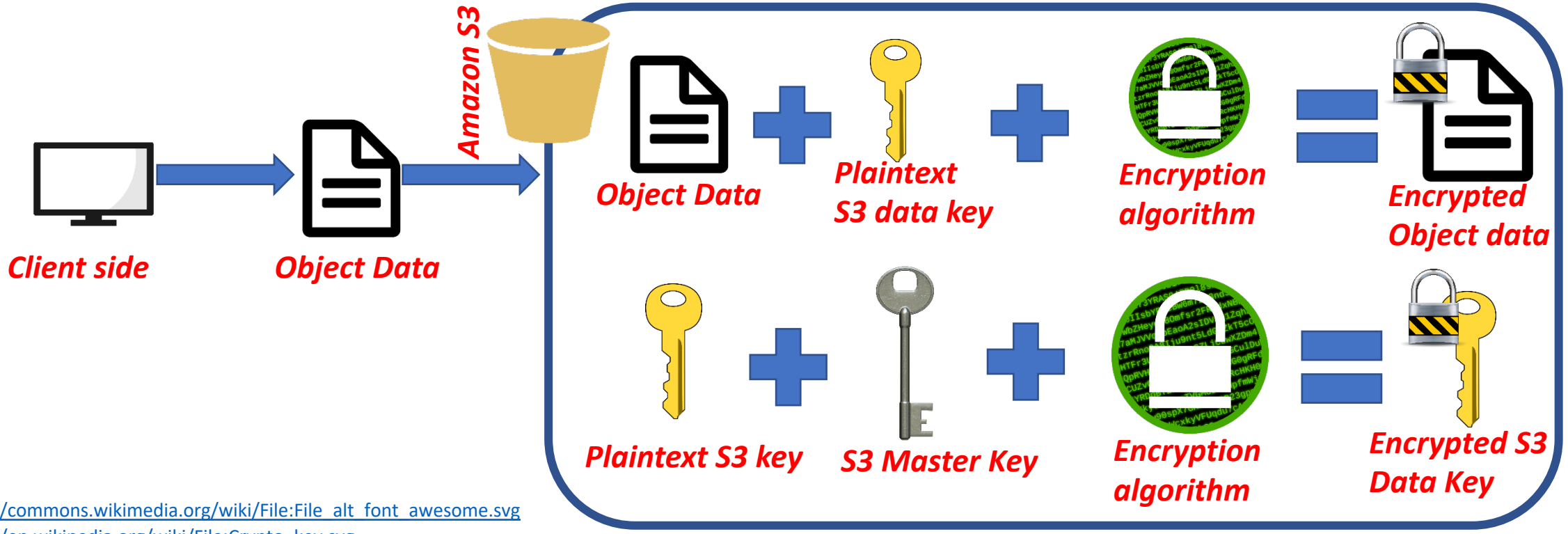
S3 ENCRYPTION

- Amazon S3 focuses primarily on data security so data encryption is an important feature of AWS S3.
- You can set a default encryption which can allow for default encryption settings to your created S3 bucket.
- With server-side encryption, Amazon **S3 encrypts an object before saving it to disk and decrypts it the object is downloaded**.
- Here are the options for encryption:
  - Amazon S3-managed keys (SSE-S3) – *common with ML services*
  - Customer master keys (CMKs) stored in AWS Key Management Service (AWS KMS). This allows for extra security – *common with ML services*.
  - SSE-C: users can manage their own keys to perform data encryption.
  - Client side encryption

Photo Credit: https://pixabay.com/vectors/computer-encrypt-encryption-1294045/

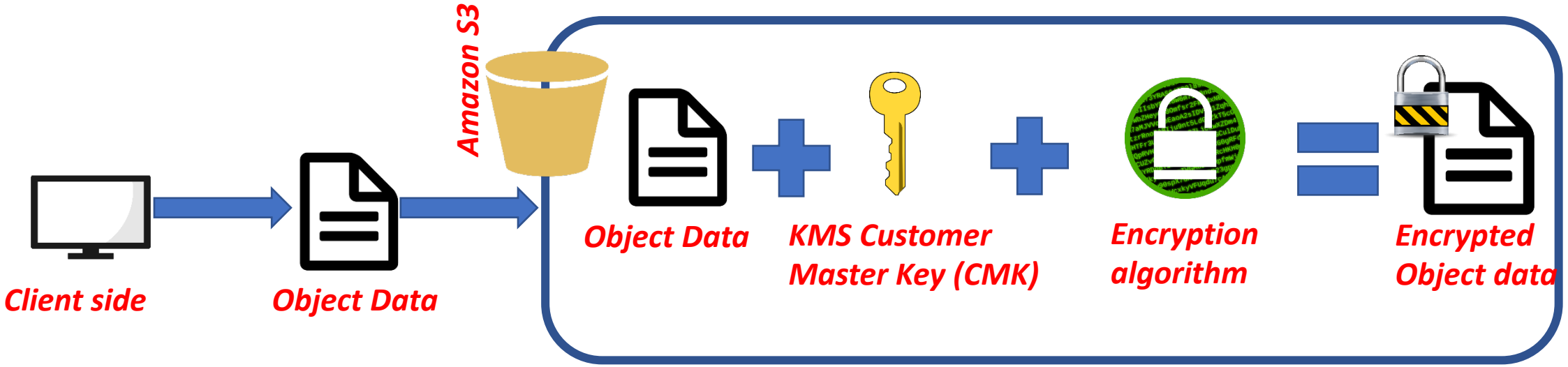
AMAZON S3 MANAGED KEYS (SSE-S3)

- Client uploads objects to S3 and select encryption method SSE-S3.
- S3 generates a plaintext key and encrypts object.
- Encrypted object is stored in S3.
- Amazon S3 master key then encrypts the plaintext key so now the key is encrypted as well and stored in S3.



https://commons.wikimedia.org/wiki/File:File_alt_font_awesome.svg
https://en.wikipedia.org/wiki/File:Crypto_key.svg
Photo Credit: https://pixabay.com/vectors/computer-encrypt-encryption-1294045/
https://www.needpix.com/photo/98552/padlock-encrypt-encrypted-lock-locked-protected-resistant-safe-secure

AWS SSE-KMS: AWS KEY MANAGEMENT SERVICE



https://commons.wikimedia.org/wiki/File:File_alt_font_awesome.svg
https://en.wikipedia.org/wiki/File:Crypto_key.svg
Photo Credit: https://pixabay.com/vectors/computer-encrypt-encryption-1294045/
https://www.needpix.com/photo/98552/padlock-encrypt-encrypted-lock-locked-protected-resistant-safe-secure

AMAZON S3 SECURITY

AWS S3 SECURITY

- Amazon S3 ensures the highest level of security to its customers.
- Amazon S3 follows a ***shared security model*** as follows:

  - **Security of the cloud:**
    - AWS ensures the protection of the infrastructure.
    - All services offered by AWS are very secure.
    - Security is being regularly audited by third party to ensure compliance.

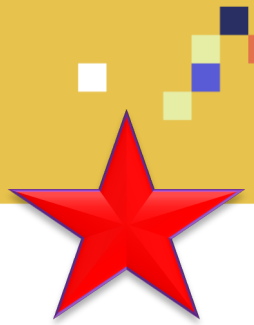
  - **Security in the cloud:**
    - Users of AWS are responsible for their own data sensitivity and organization requirements.

AWS S3 SECURITY: DATA PROTECTION IN S3

- Amazon S3 ensures data protection and durability by:
  1. Adding redundancy storage over multiple devices across many facilities/regions
  2. S3 ensure that any corrupted files are detected and repaired
  3. Using versioning: which allow users to retrieve several versions of the same object (S3 automatically retrieves the most up to data version.

- Amazon S3 has the following unique characteristics:
  - Ensures 99.999999999% durability and 99.99% availability
  - Can sustain data loss in two locations at the same time

AWS S3 SECURITY: BEST PRACTICES



Server-side encryption

In transit data encryption

Amazon Macie

Perform auditing to Amazon S3 buckets

AWS security advisories monitoring

MACIE IS A SECURITY SERVICE TO AUTOMATICALLY FIND SENSITIVE DATA AND ALERT USERS FOR ANOMALIES

AWS S3 SECURITY: RESOURCE Vs. USER BASED POLICES

- S3 buckets are private by default, only the owner can access the bucket.
- Owner can allows access to others by creating an access policy.
- There are generally two types of polices in AWS to grant permission to resources in AWS S3:
  - **Resource based:** these are access polices that you attach to your buckets.
    - ❖ Example: bucket polices and Access Control List (ACL)
  - **User based:** polices assigned to specific users in the account
    - ❖ Example: Identity and Access Management (IAM) polices
- Both polices use JSON-based access policy language.

AWS S3 SECURITY: ACCES CONTROL LIST (ACL)

- Access control lists (ACLs) belongs to the resource-based access policy.
- ACL could be used to allow permission to other AWS accounts to read/write objects
- Scenario: if a bucket owner let another AWS account upload object to the bucket. The AWS account that owns the object can only allow ACLs to this object.

AWS S3 SECURITY: S3 BLOCK PUBLIC ACCESS

- In order to manage public access to your Amazon S3, amazon has a feature to “**block public access**” settings for buckets.
- Any newly created buckets does not allow public access by default.
- Users can change this and allow for public access if they want to.
- Amazon S3 block public settings overrides any created polices and permissions to block public access to resources/buckets (centralized control).
- If any request is made from anywhere to access a specific bucket, Amazon S3 checks “**block public access**” settings and these settings has the ultimate power (overrides any other polices).
- If “**block public access”** settings prevents access to the request, access request will be denied.

AMAZON S3 SECURITY – PART #2

AWS S3 SECURITY: LOGGING AND MONITORING IN AMAZON S3

- Monitoring is important to keep track of the health of AWS services/resources to ensure availability and performance.
- AWS provides many tools to monitor S3 resources as follows:

AWS S3 SECURITY: LOGGING AND MONITORING IN AMAZON S3

Amazon CloudWatch Alarms

- Used to send an alarm once a certain threshold is exceeded for multiple number of cycles.
- The alarm is sent to AWS autoscaling policy.

AWS CloudTrail Logs

- CloudTrail is used to track activity made by users, roles, or on AWS service.
- CloudTrail provides a record of past requests, IP addresses, timing of the request…etc.

Amazon S3 Access Logs

- Access logs are important to ensure security and to conduct access audits.
- Access logs are used to record/track requests made to buckets.

AWS Trusted Advisor

- Amazon offers trusted advisory service that makes recommendations on how to improve the systems performance and close any security gaps.
- Trusted Advisor provides the following:
    - Check that amazon S3 buckets have proper configuration.
    - Check amazon S3 buckets that have permissions set to “open access”.
    - Checks Amazon S3 buckets that did not enable versioning.

AWS S3 SECURITY: ADDITIONAL SECURITY (IMPORTANT)

- **Networking - VPC Endpoint Gateway:**
  - Amazon Virtual Private Cloud (VPC) allows users to create an AWS resources inside a virtual Network.
  - This means that the traffic will never leave or go through the public internet and will stay inside the VPC for maximum security.
  - A VPC endpoint will route requests to Amazon S3 and back to the VPC.

- **Tagging**
  - You can use tagging in tandem with bucket polices and IAM policies to ensure secruty.
  - Tag example: sensitive = TRUE

ADDITIONAL NOTES

DATA WAREHOUSE Vs. DATABASE

	Data Warehouse	Database
Used for?	Used for data analytics	Used for Transaction processing
Sources of data?	Data gathered from multiple sources	Data collected as-is from one source
Data writing frequency?	Bulk write operations per fixed schedule (every day)	Many write operations as data becomes available
Storage optimized for?	Optimized for high-speed query in column format	Optimized for high throughout write operations to a single row

https://aws.amazon.com/data-warehouse/

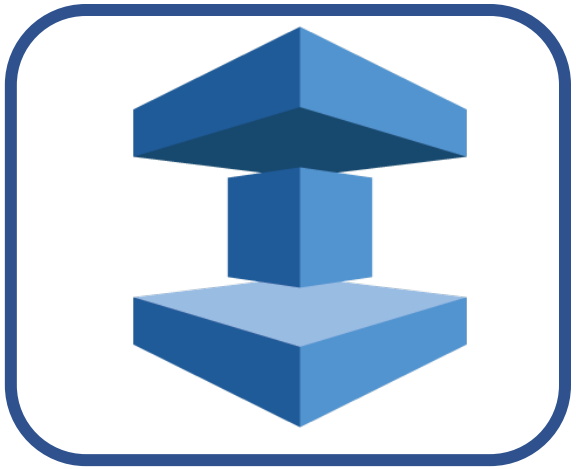
AMAZON ELASTICSEARCH

- Elasticsearch is an open source distributed search and analytics engine.
- Elasticsearch works with various types of data such as numerical, text, structured and unstructured.
- Elasticsearch performs data ingestion, enrichment, storage, analysis, and visualization.
- Tools are known as ELK Stack (Elasticsearch, Logstash, and Kibana).
- Amazon Elasticsearch Service is a fully managed service that allows for Elasticsearch deployment easily and securely.
- Amazon Elasticsearch is cost effective with zero upfront cost.
- Elasticsearch could be used for clickstream analytics, data indexing.
- Watch Video: https://aws.amazon.com/elasticsearch-service/

AMAZON ELASTICSEARCH

AMAZON ELASTICACHE

- Amazon ElastiCache is an in-memory data store and cache.
- Amazon ElasiCache is extremely fast and designed specifically for demanding applications that need sub-millisecond response times.
- Amazon ElastiCache is designed for Gaming, Internet of things, and Healthcare applications.
- Elasticache is used for data intensive apps by retrieving data from high throughput and low latency in-memory data stores.



AMAZON ELASTICACHE