AWS Certified solutions associate

Compute

EC2

Requires managing infrastructure, while Fargate, containers and kubernetes don't.

Multi-AZ uses synchronous replication ensuring almost no RPO (recovery point objective). Read replicas take longer.

EBS can persist independently from the life of an instance. For EBS up to 64k IOPS, chooose Nitro-based EC2 instance. Other instances guarantee up to 32k IOPS only.

If the instance has data stored in RAM and the instance has to be shut down for some time, enable hibernation and hibernate the instance before shutdown. Snapshotting the instance won't help because RAM contents are reloaded.

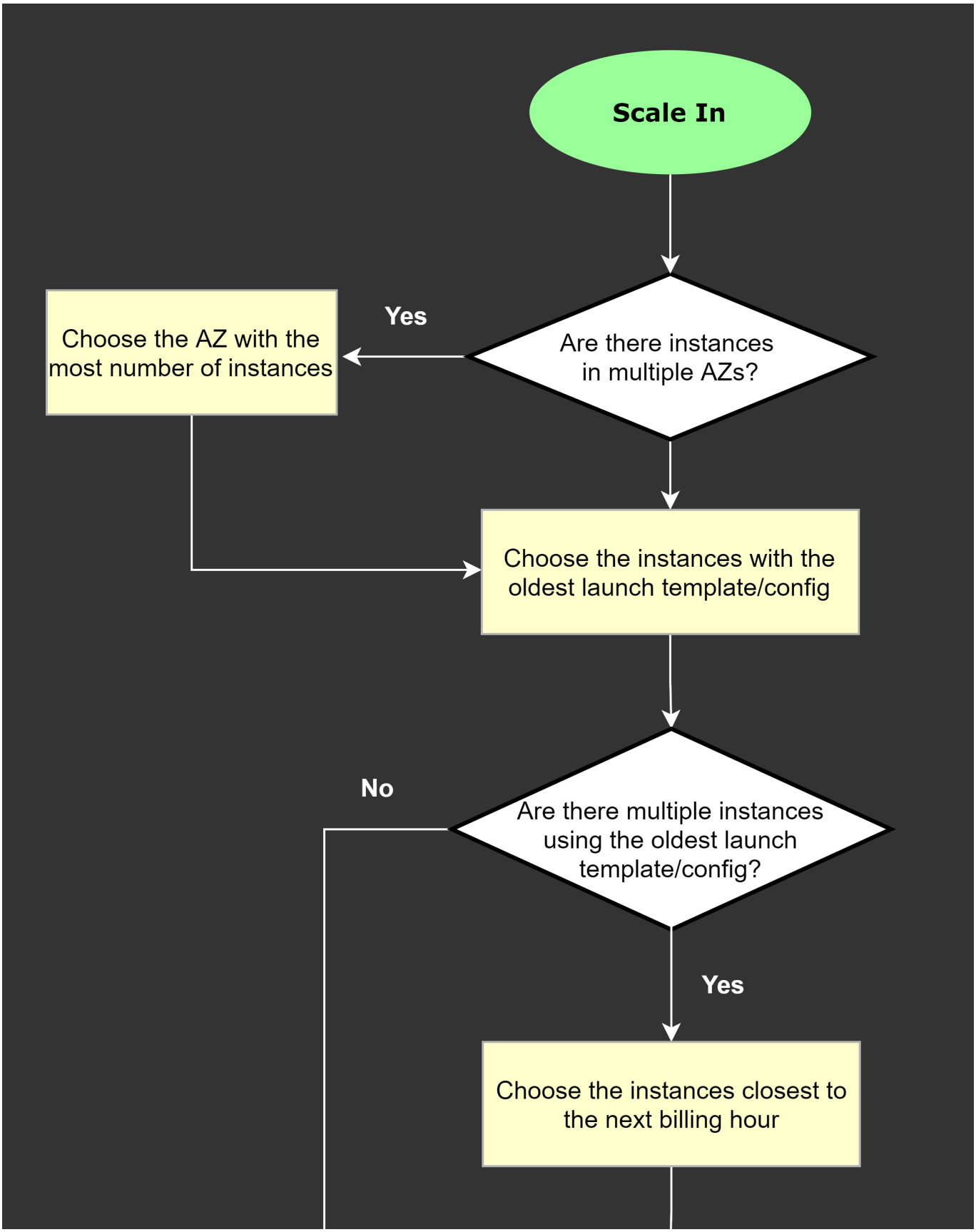
- Standard reserved instance: more discount, can't exchange instances but can change Availability Zone, scope, network platform, or instance size.
- Convertible reserved instance: flexibility to change families, OS types and tenancies

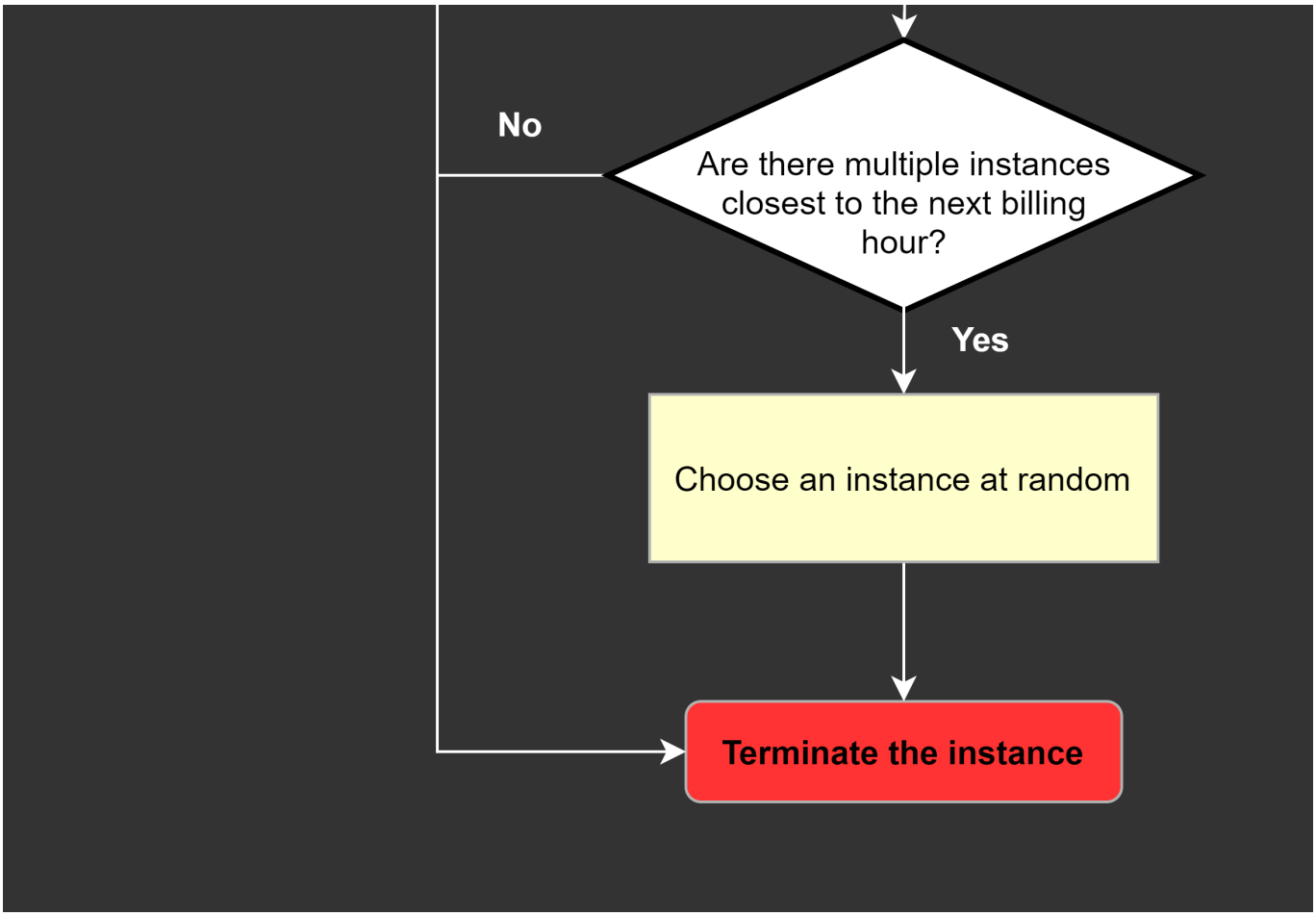
Autoscaling

Simple scaling: you have to wait for the cooldown period to complete before initiating additional scaling activities.

When downscaling, Autoscaling first deletes the instance launched from the oldest launch configuration.

Default termination policy in the picture below:





Target tracking policy: allows you to increase/decrease the current capacity of the group based on a target value for a specific metric, which helps with overprovisioning of resources. The scaling policy adds/removes capacity as required to keep the metric as close to the specified target value

For highly available instances, deploy in at least 2 AZ. If we need to have at least 2 instances running, we need 2 AZ and 2 instances in each. Worst case, one AZ fails but we still have 2 instances running. If we only had 1 instance in the unaffected AZ, Autoscaling would deploy the second instance but it would take some time, for a while there would be only 1 instance. Which we don't want. Max. capacity is 6, 3 in each AZ.

For predictable load changes, e.g. when expecting a load at 9AM when people come to work, you can configure a scheduled scaling policy to perform scaling at specified times.

ECS

ECS tasks can be run on CloudWatch events, e.g. when a file is uploaded to a certain S3 bucket using a S3 PUT operation. You can also declare a reduced number of ECS tasks whenever a file is deleted on the S3 bucket using the DELETE operation.

First, create an Event **rule** for S3 that watches for object-level operations (PUT, DELETE). For object-level operations, it is required to create a CloudTrail trail first. On the Targets section, select the "ECS task" and input the needed values such as the cluster name, task definition and the task count. You need two rules – one for the scale-up and another for the scale-down of the ECS task count.

To insert sensitive data into containers, you can store it in Secrets Manager secrests or SYstem Manager Parameter Store parameters. Then you can reference them in the container definition. This feature is supported by tasks using both the EC2 and Fargate launch types.

- Use the `secrets` container definition parameter to inject sensitive data as environment variables
- Use the `secretOptions` container definition parameter to inject sensitive data in the log configuration of a container

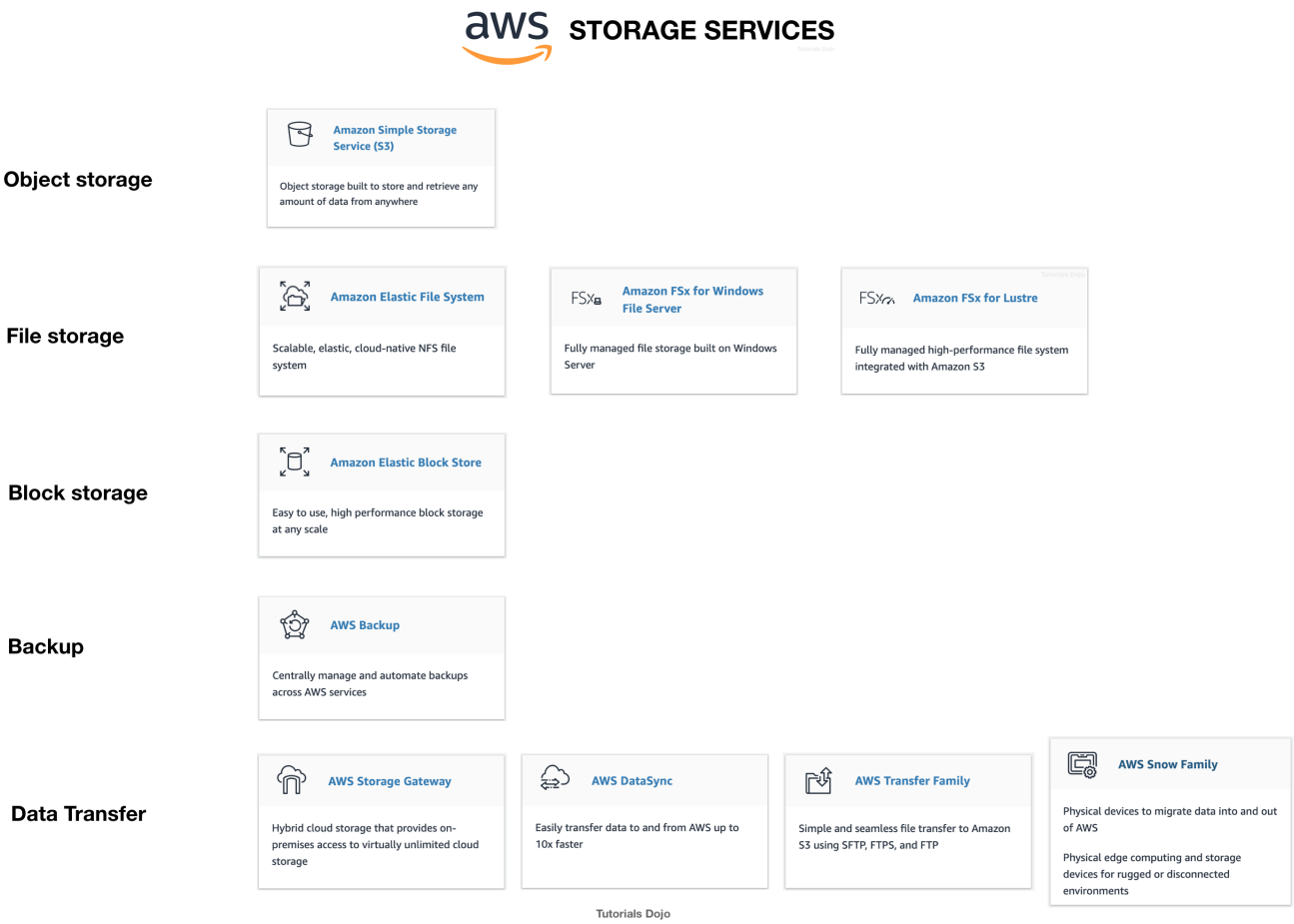
They can also handle bursts in traffic but it takes minutes to set up new containers.

Lambda

They can be used to handle bursts of traffic in seconds in serverless applications.

For sensitive info in env variables, create a new KMS key and use it to enable encryption helpers that leverage on KMS to store and encrypt the sensitive info. Lambda encrypts the environment variables in the function by default, but the info is still visible to other users who have access to the Lambda console. Lambda uses a default KMS key to encrypt the variables, which is usually accessible by other users.

X. Storage



- Amazon FSx For Lustre: high-performing *parallel* file system for fast processing of workloads
- Amazon FSx For Windows File Server: fully managed Microsoft Windows filesystem with support for SMB protocol, Windows NFTS, Microsoft Active Directory integrations
- AWS Storage gateway: integrate the on-premises network to AWS but doesn't migrate apps. If using a fileshare in Storage Gateway, the on-premises systems are still kept. Hybrid storage solutions. Enables Active Directory users to deploy storage on their workstations as a drive. mounted as a disk for on-premises desktop computers
- EFS: only supports Linux workloads

X.Y. S3

To aggregate data in buckets in many locations, use Transfer Acceleration in the destination bucket and upload the collected data using Multipart upload. Uploading the data to the closest S3 bucket and setting up cross-region replication and copying objects to destination bucket works as well, but it takes longer.

Amazon Macie is a ML-powered service that monitors and detects usage patterns on S3 data, it can detect anomalies, risk of unauthorized access or inadvertent data leaks. It can recognize PII (personally identifiable info) or IP.

CORS (cross-origin resource sharing) allows webapps loaded in one domain to interact with resources in a different domain. For instance, to add JavaScript to the webapp.

How to give access to paying subscribers to a specific files? Use Signed cookies to control who can access private files in the CloudFront distribution by modifying the app to check whether a user should have access or not. For members, send the required `SetCookie` headers to the viewer which will unlock the content only to them.

Signed URLs are when you want to restrict access to individual files, or when users are using a client (e.g. HTTP client) that doesn't support cookies. With signed cookies, you can provide access to multiple restricted files but from a restricted group of users.

Use pre-signed URLs to access specific objects.

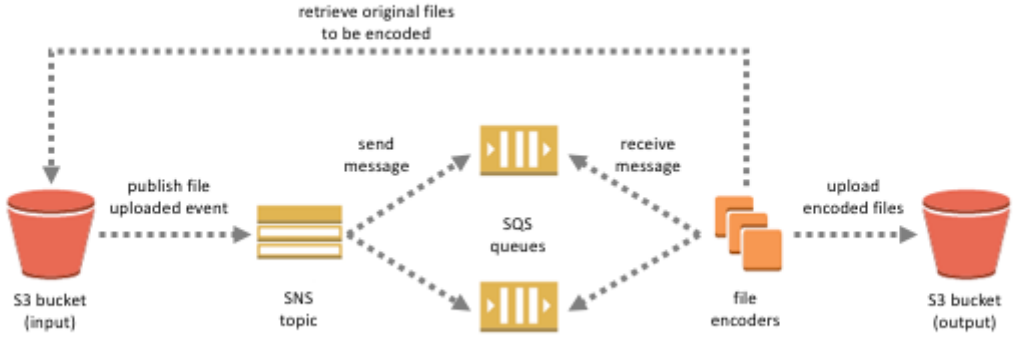
For data archiving, S3 Glacier (also Glacier deep archive). One Zone-IA is for infrequent access. With lifefycle policy, you can specify that the data is moved to another storage class (like for archiving).

> S3 notification feature

The S3 notification feature can send notifications when certain events happen in a bucket. S3 event notifications are designed to be delivered at least once and to one destination only. You cannot attach two or more SNS topics or SQS queues for S3 event notification. Therefore, you must send the event notification to Amazon SNS.

First add notification config saying which event should S3 publish and the destination for the notification, which can be the following:

- SNS: the fanout scenario is when a message published to an SNS topic is replicated and pushed to multiple endpoints (SQS, HTTP(s), Lambda). This allows for parallel asynchronous processing
    - you can create a topic and use two Amazon SQS queues to subscribe to the topic. If Amazon SNS receives an event notification, it will publish the message to both subscribers
    - For example, you can develop an application that publishes a message to an SNS topic whenever an order is placed for a product. Then, SQS queues that are subscribed to the SNS topic receive identical notifications for the new order. An Amazon Elastic Compute Cloud (Amazon EC2) server instance attached to one of the SQS queues can handle the processing or fulfillment of the order. And you can attach another Amazon EC2 server instance to a data warehouse for analysis of all orders received



X.Y. DynamoDB

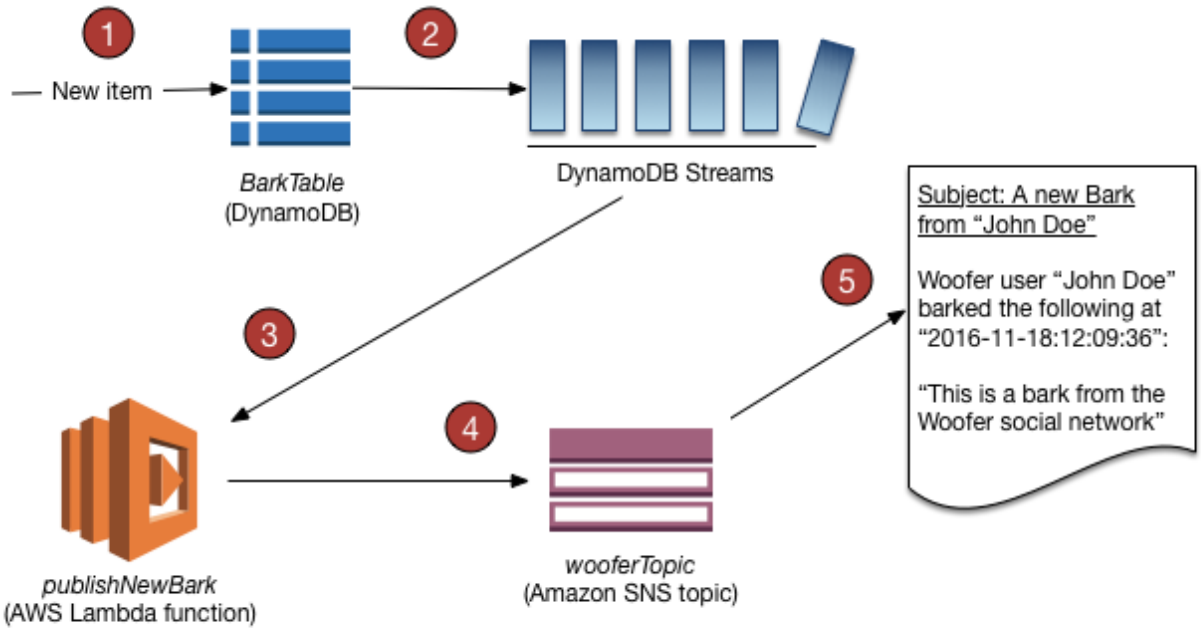
DynamoDB is a NoSQL db that can handle frequent schema changes and doesn't have downtime with schema changes.

DynamoDB table in on-demand capacity mode should be used for unpredictable traffic. Pricing is pay-per-request for read and write requests.

Changes in a table (create, update, delete) can be sent to a DynamoDB stream. This stream can be read b Lambda.

DynamoDB and ElastiCache provide high performance storage of key-value pairs.

DynamoDB Stream is an ordered flow of info about changes to items in a table. When enabling stream on a table, DynamoDB captures info about every modification to data items in the table. A stream record contains info about data modifications. You can configure the stream so that the record captures additioanl info, e.g. the before/after images of modified items. You can set up triggers so that a specific change in a table triggers a Lambda function.



Partition keys

To improve performance by distributing the workload evenly and using the provisioned throughput efficiently, use partition keys with high-cardinality attributes, which have a large number of dinstinct values for each item.

The partition key portion of a table's primary key determines the logical partitions in which a table's data is stored. This in turn affects the underlying physical partitions. Provisioned I/O capacity for the table is divided evenly among these physical partitions. Therefore a partition key design that doesn't distribute I/O requests evenly can create "hot" partitions that result in throttling and use your provisioned I/O capacity inefficiently. You should also design the partition-key system efficiently, with the info in the previous paragraph.

X.Y. Aurora

Relational db, supports dynamic storage scaling and can conduct table joins. Automatically scales to accomodate data growth.

For certain Aurora tasks, different instances perform different roles. The primare instance handles all data definition language (DDL) and data manipulation language (DML) statements. Up to 15 Aurora Replicas handle read-only query traffic. Using endpoints, you can ap each connection to the appropriate instance based on your use case. The custom endpoint provides load-balanced database connections based on criteria other than the read-only or read-write capability of the DB instances.

An Aurora serverless DB cluster is a DB cluster that automatically starts up, shuts down and scales up/down based on the app's needs. It's a simple, cost-effective option for infrequent, intermittent sporadic or unpredictable workloads. You can create a db endpoint without specifying the DB instance class size, you only set the min and max capacity. The endpoint connects to a proxy fleet that routes the workload to a fleet of resources that are automatically scaled.

Non serverless clusters use the *provisioned* db engine mode.

Aurora Global db is designed for globally distributed apps. It supports storage-based replication (RPO) with a latency of <1s. If there's an unplanned outage, one of the secondary regions you assigned can be promoted to read and write capabilities (RPO) in <1min.

X.Y. RDS

To monitor RDS, enable Enhanced Monitoring in RDS. By default these logs are stored in CloudWatch for 30 days

If an instance of RDS in 1 AZ fails very often, enable Multi-AZ deployment, which has synchronous replication. Making a snapshot allows a backup, but it doesn't provide immediate availability in case of AZ failure.

X.Y. EBS

Features	**SSD**	**HDD**
Best for	small, random I/O operations	large, sequential I/O operations

Features	SSD	HDD
Can be used as a bootable volume?	Yes	No
Suitable use cases	Transactional workloads, sustained IOPS performance, large db workloads	large streaming workloads, big data, data warehouses, throughput-oriented storage for large volumes of data that's infrequently accessed
Cost	Moderate/high	Low

- General purpose
- Provisioned IOPS: I/O intensive
- Throughput optimized: large operations (large data)

Messaging

SQS

SQS polling is not real time. If we receive empty messages when polling, enable long polling: set ReceiveMessageWaitTimeSeconds to higher than 0. In long polling, SQS waits until a message is available before sending a response to a ReceiveMessage request.

SNS works real-time. Lambda is a valid subscriber. EventBridge is not a valid SNS destination

MQ

MQ is used for migrating messaging services to the cloud quickly and easily.

Managed message broker service for Apache ActiveMQ that makes it easy to set up and operate message brokers in the cloud and hybrid architecture. The user case is when migraing to a managed message broker to automate software administration and maintenance, without having to re-write existing applications.

Security

- GuardDuty: threat detection service that monitors for malicious activity and unauthorized behavior in your AWS accounts and workloads
- Inspector: automated security assessment service that helps improve the security and compliance of apps deployed on AWS
- Shield: detect and mitigate DDoS attacks
    - Shield standard: network and transport layer protection
    - Shield advanced: additional detection and mitigation. Near real-time visibility into attacks, integration with WAF
- WAF: blocks common attack patterns such as SQL injection or cross-site scripting
    - AWS Firewall manager: simplify WAF administration and maintenance tasks across multiple accounts and resources
    - If there are external requests to a website from multiple systems with IPs that constantly change, create a rate-based rule in WAF and associate the web ACL to an ALB

Access management

If a company is using Active Directory in their on-premise system, AWS Directory Service AD connector for easier integration. If the roles on-prem are already assigned using groups, in AWS use IAM roles or use Microsoft AD federation service.

Use IAM users only when creating new credentials, if the company already has then on-premises, they can be imported some other way.

IAM groups is a collection os IAM users that lets you specify permissions for multiple users.

To manage AWS resources centrally, use AWS organizations and AWS RAM (resource access manager) which enables you to share resources with any account or within organizations. You can share AWS Transit Gateways, Subnets, AWS License Manager configurations, and Amazon Route 53 Resolver rules resources with RAM.

IAM doesn't allow MFA.

MySQL and PostgreSQL dbs instance can be authenticated with IAM DB authentication and then you only need an authentication token to access it.

For temporary credentials, use Single-Sign-on: users can log in from their on-prem AD or LDAP directory: set up a federation proxy or an identity provider and use AWS STS (security token service) to generate temporary tokens.

Redis

If users need to authenticate, use Redis AUTH by creating a new Redis Cluster with bot hthe `--transit-encryption-enabled` and `--auth-token` parameters enabled. The second parameters asks users for a password.

CloudFront

To block access for certain coutries, use CloudFront geo restriction.

To reduce delay around the world, use Lambda@Edge which allows Lambda functions to execute the authentication process in AWS locations closer ot the users.

If users around the world have HTTP 504 errors, set up an origin failover by creating an origin group with 2 origins: specify one as the primary origin, the other as secondary origin which CloudFront automatically switches to when the primary origin returns specific HTTP status code failure updates.

You can also deploy the app to multiple AWS regions and set up a Route53 record with latency routing policy to route incoming traffic to the region that provides the best latency, but this has more costs.

Networking

To check all healthy instances, use multivalue answer routing policy to help distribute DNS responses across multiple resources. For example, use multivalue answer routing when you want to associate your routing records with a Route 53 health check.

AWS Site-to-Site VPN: to connect on-prem and AWS, cheap option with limited bandwidth and limited traffic

Security groups are stateful, everything is blocked by default. the security group specifies what's allowed.

To give access only to the IP: 110.238.98.71 via SSH, make a security group inbound rule: protocol TCP, Port range - 22, source 110.238.98.71/32. /32 denotes one IP address, /0 denotes the entire network. SSH protocol uses TCP and port 22.

VPC endpoints for Amazon S3 provide secure connections to S3 buckets that do not require a gateway or NAT instances.

The online application must be in public subnets to allow access from clients' browsers. The database cluster must be in private subnets to meet the requirement that there be no access from the Internet. A NAT Gateway is gives private subnets access to the Internet. **NAT Gateways must be deployed in public subnets.** For resouces in various AZs, to improve resilieny, create one NAT Gateway per AZ and configure routing so that resources use the NAT in their AZ.

An ALB sends requests to healthy instances only. It performs periodic health checks on targets in a target group, if an instance fails the health check a configurable amount of times it'll be marked as unhealthy and won't receive traffic until it passes another health check.

An ENI (elastic network interface) is a logical networking component in a VPC that represents a virtual network card. It includes a primary private IPv4 address, 1+ secondary private IPv4 addresses, 1 Elastic IPv4 per private IPv4 address, 1 public IPv4, 1+ IPv6.

Encryption

If the encryption keys must be stored on premises, use SSE-C (server-side, customer provided keys) but in this case the key is sent to AWS as part of the request. or use client-side encryption to provide at-rest encryption.

If the master key and the unencrypted data can't be sent to AWS, we need client-side encryption.

KMS with CMK in a custom key store and storing the non-extractble key material in CloudHSM: allows you to have full control of the encryption of the created key and audit key usage in CloudTrail.

CloudTrail

Logging system to track all changes in all regions: set up a new CloudTrail trail in a new S3 bucket using CLI and pass the `--is-multi-region-trail` and `--include-global-service-events` parameters, then encrypt log files using KMS encryption.

API Gateway

Amazon API Gateway provides throttling at multiple levels including global and by a service call. Throttling limits can be set for standard rates and bursts. For example, API owners can set a rate limit of 1,000 requests per second for a specific method in their REST APIs, and also configure Amazon API Gateway to handle a burst of 2,000 requests per second for a few seconds

API Gateway can scale using AWS Edge locations, but for bursts of API, you need to configure **throttling limits**. Any request over the limit will receive a 429 HTTP response.

CloudWatch

CloudWatch by default monitors CPU, network and disk read activity on EC2 instances. To get memory utilization, need to have a custom metric.

Install the CloudWatch agent in the EC2 instances that gathers all the metrics (memory usage for ex.). View the custom metrics in the CloudWatch console.