Notes for AWS Certified Solutions Architect Associate

I recently got the AWS solutions architect associate certificate in July 2019, and wanted to share my notes with anyone who might benefit from it. The path I followed was

- Go through the ACloudGuru course.
- Attempt the Whizlabs practice tests. After each test, note down the concepts I had difficulties with.
- Attempt the practice tests by Jon Bonso at Udemy. Again, after each test, note down the concepts I had difficulties with.

So you should go through the notes only after you have done a course that explains the basics, such as the one from ACloudGuru. Also, full disclosure, the links to the above courses are referral ones. So if these notes helped you and you're planning to buy the courses or practices tests, please consider going through the links when you're buying.

Note — You can also check out this blog post where I describe my preparation strategy in detail.

Contents (19 Key Concepts)

- Storage Gateway
- IAM, Cognito and Directory Services
- KMS and CloudHSM
- Kinesis
- EMR
- Misc

1. Well-Architected Framework

The five pillars are —

1. Operational Excellence
2. Security
3. Reliability
4. Performance Efficiency
5. Cost Optimization

Operational Excellence

Design Principles

- Perform operations as code
- Annotate documents
- Make frequent, small, reversible changes
- Refine operations procedures frequently
- Anticipate failure
- Learn from all operational failures

Best Practices

- Prepare
- Operate
- Evolve

Key AWS Service — AWS CloudFormation.

Security

Design Principles

- Implement a strong identity foundations
- Enable traceability
- Apply security at all layers
- Automate security best practices
- Protect data in transit and at rest

- Keep people away from data
- Prepare for security events

Best Practices

- Identity and Access Management
- Detective Controls
- Infrastructure Protection
- Data Protection
- Incident Response

Key AWS Service — AWS Identity and Access Management (IAM).

Reliability

Design Principles

- Test recovery procedures
- Automatically recover from failure
- Scale horizontally to increase aggregate system availability
- Stop guessing capacity
- Manage change in automation

Best Practices

- Foundations
- Change Management
- Failure Management

Key AWS Service — Amazon CloudWatch.

Performance Efficiency

Design Principles

- Democratize advanced technologies
- Go global in minutes
- Use serverless architecture
- Experiment more often

- Mechanical sympathy

Best Practices

- Selection
  - Compute
  - Storage
  - Database
  - Network
- Review
- Monitoring
- Tradeoffs

Key AWS Service — Amazon CloudWatch.

Cost Optimization

Design Principles

- Adopt a consumption model
- Measure overall efficiency
- Stop spending money on data center operations
- Analyze and attribute expenditure
- Use managed and application level services to reduce cost of ownership

Best Practices

- Expenditure Awareness
- Cost-Effective Resources
- Matching Supply and Demand
- Optimizing Over Time

Key AWS Service — Cost Explorer.

2. Route53

Main functions of Route53 —

1. Register domain names.
2. Route internet traffic to the resources for your domain.
3. Check the health of your resources.

It's not used to *distribute* traffic.

CNAME vs ALIAS —

- For routing to S3 bucket // Elastic load balancer use A record with ALIAS.
- For routing to RDS instance use CNAME with NO ALIAS // without ALIAS.

ALIAS only supports the following services —

- API Gateway
- VPC interface endpoint
- CloudFront distribution
- Elastic Beanstalk environment
- ELB load balancer
- S3 bucket that is configured as a static website
- Another Route 53 record in the same hosted zone

Route53 does not directly log to S3 bucket, we can forward that from Cloudwatch, but can't do it directly.

Types of Route53 health checks —

1. Health checks that monitor an endpoint. This can be on-premise too.
2. Health checks that monitor other health checks.
3. Health checks that monitor Cloudwatch alarms.

Multivalue answer routing policy responds with upto 8 healthy records selected at random.

Weighted routing policy is a good fit for blue-green deployments.

3. S3

In a newly created S3 bucket, everything // every additional option is turned off by default. Also, no bucket policy exists.

S3 bucket properties are —

1. Versioning
2. Server access logging
3. Static website hosting
4. Object level logging // Essentially CloudTrail
5. Transfer acceleration
6. Events

Object level properties—

Metadata and Storage class are object level properties. All object level properties are

1. Storage class
2. Encryption
3. Metadata
4. Tags
5. Object lock

DELETE operation does not keep a copy unless you have versioning enabled. From the docs

The DELETE operation removes the null version (if there is one) of an object and inserts a delete marker, which becomes the current version of the object. If there isn't a null version, Amazon S3 does not remove any objects.

S3 is a managed service. It can't be part of a VPC.

S3 object metadata—

1. System metadata
2. User-defined metadata

User defined metadatas must start with `x-amz-meta`.

When you enable logging on a bucket, the console both enables logging on the source bucket and adds a grant in the target bucket's access control list (ACL) granting write permission to the Log Delivery Group.

S3 bucket endpoints formats —

1. http://bucket.s3.amazonaws.com
2. http://bucket.s3.aws-region.amazonaws.com
3. http://bucket.s3-aws-region.amazonaws.com
4. http://s3.amazonaws.com/bucket
5. http://s3.aws-region.amazonaws.com/bucket
6. http://s3-aws-region.amazonaws.com/bucket

Update — AWS will stop supporting the URL path format for buckets created after September 30, 2020. Read this for details.

Object sizes — S3 can store objects of size 0 bytes to 5 TB. A single PUT can transfer 5 GB max. For files larger than 100MB, multipart upload is recommended.

Cross-region replication requires that versioning be enabled on both the source bucket and the destination bucket.

AWS Glacier archive retrieval options —

- Expedited: Costly, 1-5 minutes.
- Standard: Default, 3-5 hours.
- Bulk: Cheapest, 5-12 hours.

To increase performance, we can prefix each object name with a hash key along with the current date. But, according to the new S3 performance announcement, this is not needed anymore.

Increasing performance in S3 —

- If workload is mainly GET requests, integrate Cloudfront with S3.
- If workload consists of PUT requests, use S3 transfer acceleration.

In the CORS configuration, the exact URLs must be added, with the correct protocol, i.e. http vs https.

S3 does not support `OPTIONS`, `CONNECT` and `TRACE` methods.

S3 encryptions —

- SSE-S3: Data and master keys managed by S3.
- SSE-C: The user manages the encryption keys.
- SSE-KMS: AWS manages the data key, the user manages the master key.

To make sure that S3 objects are only accessible from Cloudfront, create an Origin Access Identity (OAI) for Cloudfront and grant access to the objects to that OAI.

We can create event notification in S3 to invoke lambda function.

Customer managed S3 encryption workflow —

Generate a data key using Customer managed CMK. Encrypt data using data key and delete data key. Store encrypted data key and data in S3 buckets. For decryption, use CMK to decrypt data key into plain text and then decrypt data using plain text data key.

AWS S3 performance —

- 3,500 requests per second to add data
- 5,500 requests per second to retrieve data

Provisioned capacity should be used when we want to guarantee the availability of fast expedited retrieval from S3 Glacier within minutes.

For S3 static website hosting, the default provided URL is https://bucket-name.s3-website-aws-region.amazonaws.com.

S3 server side encryption uses AES 256.

S3 event notification targets —

- SQS
- SNS
- Lambda

An 80 TB Snowball appliance and 100 TB Snowball Edge appliance only have 72 TB and 83 TB of usable capacity respectively.

For static website hosting with S3, the name of the bucket must be the same as the domain or subdomain name.

Preventing accidental deletion of S3 objects —

- Enable versioning
- Enable MFA delete

4. RDS, Redshift and ElastiCache

Amazon Redshift Enhanced VPC Routing provides VPC resources the access to Redshift.

Amazon ElastiCache offers fully managed Redis and Memcached.

Cross-region replication can be setup for Redshift Clusters.

Redshift encryption —

- Using AWS KMS to encrypt the underlying data.
- Using S3 and its encryption.

RDS data size limits —

- Aurora: 64 TB
- Others: 16 TB.

During automated backup, Amazon RDS performs a storage volume snapshot of entire Database instance. Also, it captures transaction logs every 5 minutes.

AWS RDS is a service whereas AWS Aurora is a database engine.

For Redshift, spot instances are not an option.

Encryption of RDS — Have to enable it on database creation. Also, not all instance classes support encryption, we have to choose one which supports it.

To enable multi-region replication of RDS, we have to use Read Replicas. Multi-AZ is not the solution here.

RDS Read Replicas are synced asynchronously, so it can have replication lag.

Redshift automated snapshot retention period — 1 day to 35 days.

We can't use auto-scaling with RDS. To improve performance, we should look to sharding instead. Starting from June 20, we can use auto-scaling with RDS instances.

We configure RDS engine configurations using parameter groups.

To use REDIS AUTH with ElastiCache, in-transit encryption must be enabled for clusters.

For RDS, Enhanced Monitoring gathers its metrics from an agent on the instance.

In case of a failover, Amazon RDS flips the canonical name record (CNAME) for your DB instance to point at the standby.

Aurora endpoints, by default —

- A reader endpoint. It load balances all read traffic between instances.
- A cluster endpoint. For write operations.

We can create additional custom endpoints that load balance based on specified criteria.

With Redshift Spectrum, we can run complex queries on data stored in S3.

We can use WLM in the parameter group configuration of Redshift to define number of query queues and how queries are routed to those queues.

The memory and processor usage by each process in an RDS instance can not be monitored by Cloudwatch, we have to use RDS Enhanced Monitoring for that. Because Cloudwatch monitors the hypervisor, not the individual instances.

IAM DB authentication can be used with MySQL and PostgreSQL. With this, you don't need to use a password when you connect to a DB instance. Instead, you use an authentication token.

5. EC2 and EBS

Instance store — You cannot add instance store volume to an instance after it's launched. Not all EC2 instance types support instance store volume.

Persistence — Instance store persists during reboots, but not stop or terminate. EBS volumes however persists accross reboot, stop, and terminate.

EBS volume types —

1. General purpose SSD. For web applications // most use cases.
2. Provisioned IOPS SSD. For critical high performing databases.
3. Throughput optimized HDD. For Big Data.
4. Cold HDD. For infrequently accessed data.

Also, to note, HDDs cannot be boot volumes.

We can use Amazon Data Lifecycle Manager to automate taking backups // snapshots of EBS volumes, and protect them from accidental or unwanted deletion.

EBS-optimized EC2 instances provide additional, dedicated capacity for EBS IO. Helps squeeze out the last ounce of performance.

Encrypted EBS volumes are not supported on all instance types.

To get more performance out of EBS volumes —

1. Use a more modern Linux Kernel.
2. Use RAID 0.

VolumeRemainingSize is not an Cloudwatch metric for EBS volumes.

EBS volume types —

- For throughput, Throughput optimized HDD.
- For large number of transaction, i.e. IOPS, Provisioned IOPS SSD.

By default, EBS volumes are automatically replicated within their availability zone, and offers a significant high availability.

AWS Cloudwatch Logs can be used to monitor and store logs from EC2 instances. The instance needs awslogs log driver installed to be able to send logs to CloudWatch. We don't need any database or S3 for storage.

Cloudwatch logs agent is more efficient than AWS SSM Agent.

With EC2 dedicated hosts we have control over number of cores, not anywhere else.

Placement groups —

- Cluster
- Spread. Maximum number of instances in an AZ is 7.
- Partitioned

The console does not support placement groups, have to do it from CLI.

Cluster Placement groups have very low inter-note latency.

Hibernation of EC2 instances —

- When EC2 instance is hibernated and brought back up, the public IP4 address is renewed. All the other IP addresses are retained.
- When EC2 instance is in hibernate, you are only charged for elastic IP address and EBS storage space.

Default Cloudwatch metrics —

- CPU utilization
- Disk reads and writes
- Network in and out

Custom metrics —

- Memory utilization
- Disk swap utilization
- Disk space utilization
- Page file utilization
- Log collection

Reserved Instances that are terminated are still billed until the end of their term according to their payment option.

Upon stopping and starting an EC2 instance —

- Elastic IP address is disassociated from the instance if it is an EC2-Classic instance. Otherwise, if it is an EC2-VPC instance, the Elastic IP address remains associated.
- The underlying physical host is possibly changed.

EBS is lower-latency than EFS.

The maximum ratio of provisioned IOPS to requested volume size (in GiB) is 50:1.

For new accounts, Amazon has a soft limit of 20 EC2 instances per region, which can be removed by contacting Amazon.

You can attach a network interface (ENI) to an EC2 instance in the following ways —

1. When it's running. Hot attach.
2. When it's stopped. Warm attach.
3. When the instance is being launched. Cold attach.

EBS snapshots are more efficient and cost-effective solution compared to disk mirroring using RAID1.

EBS volumes can only be attached to an EC2 instance in the same Availability Zone.

EBS snapshot creation — In usual scenarios EBS volume snapshots can be created at the same time it's in usage. But when using RAID configurations, there are additional complexities and we should stop every IO operation and flush the cache before taking a snapshot.

Cloudwatch alarm actions can automatically start, stop or reboot EC2 instances based on alarms.

With scheduled reserved instances, we can plan out our future usage and get reserved instances in those planned time-frame only.

Throughput optimized HDD vs Cold HDD — Throughput optimized is used for frequently accessed data, whereas Cold HDD is used for infrequently accessed data. Also the later is more cost-effective.

RAID0 vs RAID1 —

- RAID1 is used for mirroring, high-availability and redundancy.
- RAID0 is used for higher performance, it can combine multiple disk drives together.

Larger EC2 instances have higher disk data throughput. This can be used in conjunction with RAID 0 to improve EBS performance.

EFS

EFS supports cross availability zone mounting, but it is not recommended. The recommended approach is creating a mount point in each availability zone.

You can mount an EFS file system in only one VPC at a time. If you want to access it or mount it from another VPC, you have to create a VPC peering connection. You should note that all of these must be within the same region.

NFS port 2049 for EFS.

Encryption

1. Encryption at rest must be specified at the creation of file system. If you want to modify it later on, create a new EFS file system with encryption enabled and copy the data over.
2. Encryption at transit is supported by EFS // NFS, and must be enabled from the client side. It simply uses SSL to encrypt the connection.

Performance mode

1. General purpose must be used for most purposes, it has low latency, so ideal for web applications.
2. Max IO is ideal for big data and parallel connection and processing from a large number of hosts. It has higher latency but large throughput.

Throughput mode

1. Bursting is ideal for arbitrary large amount of data, because it scales properly.
2. But for cases with high throughput to storage ratio, such as common in web applications, provisioned mode is better.

6. ELB and Autoscaling

Patching an AMI for an auto scaling group, the procedure is —

1. Create an image out of the main patched EC2 instance
2. Create a new launch configuration with new AMI ID
3. Update auto scaling group with new launch configuration ID.

Note that AMI ID is set during creation of launch configuration and cannot be modified, so we have to create a new launch configuration.

Default metric types for a load balancer —

1. Request count per target.
2. Average CPU utilization.
3. Network in.
4. Network out.

Monitoring Application Load Balancers —

1. Cloudwatch metrics
2. Access logs
3. Request tracing
4. Cloudtrail logs.

Adding lifecycle hooks to ASGs put instances in wait state before termination. During this wait state, we can perform custom activities. Default wait period is 1 hour.

ASG Dynamic Scaling Policies —

- Target tracking scaling. The preferred one to use, this should be the first one we should consider.
- Step scaling
- Simple scaling

If you are scaling based on a utilization metric that increases or decreases proportionally to the number of instances in an Auto Scaling group, we recommend that you use target tracking scaling policies. Otherwise, we recommend that you use step scaling policies.

The ELB service does not consume an IP address, it's the nodes that consume one IP address each.

Auto-scaling ensures —

- Fault tolerance
- Availability

ELBs can manage traffic within a region and not between regions.

For unstable scaling behavior, that is scaling multiple times frequently, the following things can be done —

- Increasing auto-scaling cooldown timer value would give scaling activity sufficient time to stabilize.
- Modify the cloudwatch alarm period that triggers scaling activity.

Default cooldown period is 300 seconds.

Port based routing is supported by Application Load Balancer.

Network Load Balancer can be used to terminate TLS connections. For this, NLB uses a security policy which consists of protocols and ciphers. The certificate used can be provided by AWS Certificate Manager.

Connection draining enables the load balancer to complete in-flight requests made to instances that are de-registering or unhealthy.

ASG termination policy —

1. Oldest launch configuration.
2. Closest to next billing hour.
3. Random.

Load balancer does not create or terminate instances, that's done by auto scaling group.

7. SQS

Consumers must delete an SQS message manually after it has done processing the message. To delete a message, use the ReceiptHandle of a message, not the MessageId.

Incoming messages can trigger a lambda function.

We can use dead letter queues to isolate messages that can't be processed right now.

SQS does not encrypt messages by default.

Default visibility timeout for SQS is 30 seconds.

Each FIFO Queue uses —

- Message Deduplication ID
- Message Group ID. Message Group ID helps preserve order.

For application with identical message bodies, use unique deduplication ID, while for unique message bodies, use content-based deduplication ID.

Both the default and maximum batch size for `ReceiveMessage` call of SQS is 10.

Reducing SQS API calls —

- Use long polling.
- Send `DeleteMessage` requests in batch using `DeleteMessageBatch`. Other batch actions are SendMessageBatch and `ChangeMessageVisibilityBatch`.

Message retention period in SQS — 1 minute to 14 days. The default is 4 days.

Limit on number of inflight messages — 120,000 for standard queue and 20,000 for FIFO queue.

8. SNS

Available protocols for AWS SNS —

- HTTP // HTTPS
- Email
- Email-JSON
- SQS
- Application
- Lambda
- SMS

We can add filter policies to individual subscribers in an SNS topic.

SNS message attributes are —

- Name
- Type
- Value

With Amazon SNS, there is a possibility of the client receiving duplicate messages.

9. API Gateway

API Gateway can integrate with any HTTP based operations available on the public internet, as well as other AWS services.

Integration types —

- Lambda function, can be from another AWS account as well.
- HTTP
- Mock
- AWS Service
- VPC Link

For connecting API Gateway to a set of services hosted in an on-premise network, we can use

1. DirectConnect to connect the private network to AWS directly.
2. Then use VPCLink to set up API Gateway connection.

API Gateway Throttling —

- Burst limit refers to the first millisecond.
- Steady-state limit refers to an one second interval.

Throttling behaviors —

- If an user exceeds the burst limit but not the steady-state limit, the rest of the requests are throttled over the one second steady-state interval.
- If an user exceeds the steady-state limit, AWS returns `429 Too Many Requests` error.

When it comes to throttling settings, you can override stage settings on an individual method within the stage. That is, there is an option for method level throttling to override stage level throttling.

Access control mechanisms for API Gateway —

- Resource policies
- AWS IAM roles and policies
- CORS or Cross-origin resource sharing
- Lambda authorizers
- Amazon Cognito user pools

- Client side SSL certs
- Usage plans

API Gateway automatically protects the backend systems from DDoS attack.

Cache properties and settings —

- Cache status
- Flush entire cache
- Enable API cache
- Cache capacity
- Encrypt cache data
- Cache TTL
- Require authorization
- Handle unauthorized requests

Monitoring API Gateway usage — we can use CloudWatch or Access logging. Access logging logs who accessed the API and how the caller accessed the API, CloudWatch does not include this data.

Protect backend systems behind API gateway from traffic spikes —

- Enable throttling.
- Enable result caching.

10. Lambda

Lambda functions can be run within a private VPC.

Lambda can read events from —

- Amazon Kinesis
- Amazon DynamoDB
- Amazon Simple Queue Service

Services that can invoke Lambda functions —

- Elastic Load Balancing (Application Load Balancer)
- Amazon Cognito
- Amazon Lex
- Amazon Alexa
- Amazon API Gateway
- Amazon CloudFront (Lambda@Edge)
- Amazon Kinesis Data Firehose
- Amazon Simple Storage Service
- Amazon Simple Notification Service
- Amazon Simple Email Service
- AWS CloudFormation
- Amazon CloudWatch Logs
- Amazon CloudWatch Events
- AWS CodeCommit
- AWS Config

AWS CodePipeline and AWS OpsWorks can't invoke lambda functions.

For failures we can configure lambda to send non-processed payloads to SQS Dead letter queue. Then we can configure SNS to send a notification if we want. Lambda does not have an in-built mechanism for notification upon failure.

A policy on a role defines which API actions can be made on the target, it does not define whether the source can access the target or not.

Each lambda function has an ephemeral storage of 512 MB in the `tmp` directory.

AWS CloudWatch rule can be configured to trigger a lambda function. While configuration, the following can be used as input to the target lambda function —

- Matched event
- Part of the matched event
- Constant (JSON text)

The following CloudFront events can trigger lambda function —

- Viewer request
- Viewer response
- Origin request
- Origin response

Lambda function update has eventual consistency. Which means, for a brief window of less than a minute, it may execute either the old version or the new version.

We can use alias versions to point to another version. This can enable easier upgradation from the viewpoint of a consumer.

Limits —

- Function memory allocation: 128 MB to 3008 MB, in 64 MB increments.
- Function timeout: 900 seconds.
- Deployment package: 50 MB * 5 layers.
- `tmp` directory storage: 512 MB.

To grant cross-account permission to a function, we have to modify the function policy, not the execution role policy.

The console doesn't support directly modifying permissions in a function policy. You have to do it from the CLI or SDK.

If we run lambda functions inside a VPN, they use subnet IPs or ENIs. There should be sufficient ones otherwise it will get throttled.

ENI capacity = Projected peak concurrent executions * (Memory in GB / 3 GB).

The lambda console provides encryption and decryption helpers for encryption of environment variables.

By default, the a KMS default service key is used for encryption, which makes the information visible to anyone who has access to the lambda console. For further restriction, create a custom KMS key and use that to encrypt.

CloudWatch metrics for Lambda —

- Invocations
- Errors
- Dead Letter Error
- Duration
- Throttles
- IteratorAge
- ConcurrentExecutions
- UnreservedConcurrentExecutions

We can get the function version within the function using —

- `getFunctionVersion` from the Context object.
- `AWS_LAMBDA_FUNCTION_VERSION` environment variable.

Lambda Retry upon Failure Behavior —

- Event sources that aren't stream-based
    - Synchronous invocation — Returns error with status code 200. Includes FunctionError field and X-Amz-Function-Error header.
    - Asynchronous invocation — Retry twice, then sent to Dead Letter Queue.
- Poll-based and stream-based event source (Kinesis or DynamoDB) — Lambda keeps retrying until the data expires. The exception is blocking, this ensures the data are processed in order.
- Poll-based but not stream-based event source (SQS) — On unsuccessful processing or if the function times out of the message, it is returned to the queue, and ready for further reprocessing after the visibility timeout period. If the function errors out, it is sent to Dead Letter Queue.

Lambda traffic shifting —

- Canary
- Linear
- All at once

11. VPC

We cannot route traffic to a NAT gateway or VPC gateway endpoints through a VPC peering connection, a VPN connection, or AWS Direct Connect. A NAT gateway or VPC gateway endpoints cannot be used by resources on the other side of these connections. Conversely, a NAT gateway // VPC gateway endpoints cannot send traffic over VPC endpoints, AWS VPN connections, Direct Connect or VPC Peering connections either.

Every route table contains a local route for communication within the VPC over IPv4. We cannot modify or delete these routes.

VPC endpoints always take precedence over NAT Gateways or Internet Gateways.

Network ACL rules are evaluated in order, starting with the lowest numbered rule. As soon as a rule matches, it is applied regardless of any higher numbered rule that may contradict it.

SSH connections are between port 22 of the host and an ephemeral port of the client. In fact, this is true for any TCP service.

Security groups are stateful, this means any connection initiated successfully will be completed.

We can create S3 proxy server for enabling use cases where S3 has to be accessed privately through VPN connection, AWS Direct Connect or VPC peering.

AWS reserves 5 IPs for every subnet, not for every VPC.

Instances in custom VPCs don't get public DNS hosts by default, we have to set the `enableDnsHostnames` attribute to true. The `enableDnsSupport` is to be set to true too, but that is done by default.

We can set a custom route table as the main route table.

We can add secondary CIDR ranges to an existing VPC. When a secondary CIDR block is added to a VPC, a route for that block with target as "local" is automatically added to the route table.

VPC peering connection route contains Target as `pcx-xxxxxx`. VPN connection // Direct Connect connection route contains Target as `vgw-xxxxxx`.

VPN is established over a Virtual Private Gateway.

There are two types of VPC Endpoints —

- Gateway endpoints support only S3 and DynamoDB.
- Interface endpoints (Powered by PrivateLink) supports Amazon ECR and many other services.

Difference between DirectConnect and VPN — DirectConnect does not involve the Internet, while VPN does.

AWS Direct Connect doesn't encrypt in transit data, while VPN does.

To establish a VPN connection, we need —

- A public IP address on the customer gateway for the on-premise network.
- A virtual private gateway attached to the VPC.

To setup AWS VPN CloudHub —

- Each regional site should have non overlapping IP prefixes.
- BGP ASN should be unique at each site.
- If BGP ASN are not unique, additional ALLOW-INs will be required.

The allowed block size in VPC is between a /16 netmask (65,536 IP addresses) and /28 netmask (16 IP addresses).

The following VPC peering connection configurations are not supported —

1. Overlapping CIDR Blocks
2. Transitive Peering
3. Edge to Edge Routing Through a Gateway or Private Connection

We can move part of our on-premise address space to AWS. This is called BYOIP. For this, we have to acquire a ROA, Root Origin Authorization from the the regional internet registry and submit it to Amazon.

12. DynamoDB

AWS DynamoDB is durable, ACID compliant, can go through multiple schema changes, and changes to the database does not result in any database downtime.

DynamoDB Global Tables can be used to deploy a multi region, multi AZ, fully managed database solution.

We can create secondary indexes for DynamoDB tables. Always choose DynamoDB when possible.

DynamoDB streams can be used to monitor changes made to a database, and they can trigger lambda functions.

We can turn on autoscaling for DynamoDB.

For write heavy use cases in DynamoDB, use partition keys with large number of distinct values.

DynamoDB Accelerator, DAX is an in-memory cache for DynamoDB that reduces response time from milliseconds to microseconds.

13. ECS

Launch types —

- Fargate
- EC2

All types of instances, i.e. on-demand, spot and reserved can be used with ECS.

Docker containers and ECS are particularly suited for batch job workloads as they can get embarassingly parallel.

Amazon ECS enables you to inject sensitive data into your containers by storing your sensitive data in either —

- AWS Secrets Manager secrets
- AWS Systems Manager Parameter Store parameters

14. Elastic Beanstalk

AWS Elastic Beanstalk can be used to create —

- Web application using DB
- Capacity provisioning and load balancing of websites
- Long running worker process
- Static website

It should not be used to create tasks which are run once or on a nightly basis, because the infrastructure is provisioned and will be running 24/7.

Elastic Beanstalk can be used to host Docker containers.

15. Storage Gateway

AWS Storage Gateways—

1. File gateway
2. Volume gateway: Cached volumes
3. Volume gateway: Stored volumes
4. Tape gateway

16. IAM, Cognito and Directory Services

Amazon Cognito has two authentication methods, independent of one another —

- Sign in via third party federation
- Cognito user pools

AWS Directory Service options —

- AWS Managed Microsoft AD
- AD Connector
- Simple AD
- Amazon Cloud Directory
- Amazon Cognito

There is no default policy ever, anywhere. When permissions are checked, roles and policies are considered together, and in the default case there is no policy, so only the role is considered.

We can configure IAM policies that allows access to specific tags.

Connecting AWS SSO to On-Premise Active Directory —

- Two-way trust relationship: Preferred. Users can do everything from both portals.
- AD connector: SSO does not cache user credentials. Users can't reset password from SSO portal, have to do it from on-premise portal.

For two-step verification, SSO sends code to registered email. It can set to be either —

- Always-on
- Context-aware

Cross-account IAM roles allow customers to securely grant access to AWS resources in their account to a third party.

If our identity store is not compatible with SAML, we can develop a custom application on-premise and use it with STS.

Microsoft Active Directory supports SAML.

17. KMS and CloudHSM

KMS master keys are region specific.

CloudHSM backup procedure — Ephemeral backup key (EBK) is used to encrypt data and Persistent backup key (PBK) is used to encrypt EBK before saving it to an S3 bucket in the same region as that of AWS CloudHSM cluster.

With AWS CoudHSM, we can control the entire lifecycle around the keys.

AWS KMS API can be used to encrypt data.

18. Kinesis

Kinesis stream data retention period — 24 hours (default) to 168 hours.

For Kinesis, we have to use VPC Interface Endpoint, powered by AWS PrivateLink.

Amazon Kinesis Scaling Utility is a less cost-effective solution compared to doing it with Cloudwatch alarms + API Gateway + Lambda function.

Kinesis data streams store the data, by default for 24 hours and upto 7 days. Whereas Kinesis Firehose stream the data directly into either —

- S3
- Redshift
- Amazon Elasticsearch Service
- Splunk

Kinesis — If ShardIterator expires immediately and data is lost, we have to increase the write capacity assigned to the Shard table.

19. EMR

AWS EMR — AWS Elastic MapReduce, Hadoop based big data analytics.

AWS EMR is preferred for processing log files.

EMR can use spot instances as underlying nodes.

We can access the underlying EC2 instances in AWS EMR cluster.

Misc

AWS STS — The policy of the temporary credentials generated by STS are defined by the intersection of your IAM user policies and the policy that you pass as argument.

AWS VM Import // Export can be used to transfer virtual machines from local infrastructure to AWS and vice-versa.

AWS Trusted Advisor is a resource that helps users with cost management, performance and security.

We can create a CloudTrail log across all regions.

CloudFormation Drift Detection can be used to detect changes in the environment. Drift Detection only checks property values which are explicitly set by stack templates or by specifying template parameters. It does not determine drift for property values which are set by default.

AWS Server Migration Service (SMS) is an agentless service which makes it easier and faster for you to migrate thousands of on-premise workloads to AWS.

AWS Athena is a managed service which can be used to make interactive search queries to S3 data.

Amazon Inspector is a security assessment service, which helps improve security and compliance of applications.

AWS Opsworks is a configuration management service for Chef and Puppet. With Opsworks Stacks, we can model our application as a stack containing different layers.

By default, CloudTrail logs are encrypted using S3 server-side encryption (SSE). We can also choose to encrypt with AWS KMS.

Amazon ECS for Kubernetes (EKS) exists, it's a managed service.

Changes to CloudTrail global service event logs can only be done via the CLI or the SDKs, not the console.

For CloudFront query string forwarding, the parameter names and values used are case sensitive.

AWS Polly — Lexicons are specific to a region. For a single text appearing multiple times, we can create alias using multiple Lexicons.

Amazon Quicksight is a managed service for creating dashboards with data visualization.

AWS Athena pricing is based upon per query and amount of data scanned in each query. To reduce price —

- Partition data based on different parameters so that amount of data scanned gets reduced.
- Create separate workgroups based upon user groups.

AWS CloudSearch helps us add search to our website or application. Like Elasticsearch.

AWS Glue is a fully managed ETL service for data. It keeps a track of processed data using Job Bookmark. Enabling Job Bookmark will help to scan only changes since last bookmark and prevent processing of whole data again.

AWS X-Ray — Helps debug and analyze microservices architecture.

Reducing cost with AWS X-Ray — Sampling at a lower rate.

Amazon WorkDocs has a poweruser facility, which on enabling restricts sharing of documents to that user only.

AWS Data Pipeline can automate the movement and transformation of data for data-driven workflows. For example, transferring older data to S3 from DynamoDB.

Disaster recovery solutions —

- Backup and Restore. Cheapest.
- Pilot Light
- Warm Standby
- Multi-Site
- Multiple AWS Regions. Costliest.

With AWS Config, we can get a snapshot of the current configuration of our AWS account.

For queue based processing, scaling EC2 instances based on the size of the queue is a preferred architecture.

It's best practice to launch Amazon RDS instance outside an Elastic Beanstalk environment.

AWS Athena is simpler and requires less effort to set up than AWS Quicksight.

RI Coverage Budget reports number of instances that are part of Reserved Instance. For an organisation using default IAM policy, each member account owner needs to create a budget policy for individual accounts and not by master account.

Consolidated Billing in AWS Organisations combines usage from all accounts and billing is generated based upon total usage. Services like EC2 and S3 have volume pricing tiers where with more usage volume the overall charge decreases.

To automatically trigger CodePipeline with changes in source S3 bucket, use CloudWatch Events rule and CloudTrail trail.

Amazon Data Lifecycle Manager can be used for creation, retention and deletion of EBS snapshots.

With AWS Organizations, we can centrally manage policies across multiple AWS accounts. With Service Control Policies (SCPs), we can ensure security policies are in place.

AWS WAF is a web application firewall.

In AWS Managed Blockchain network, the format for resource endpoint is — `ResourceID.MemberID.NetworkID.managedblockchain.us-east-1.amazonaws.com:PortNumber`.

When you want to keep your expenditure within a budget, use AWS Budgets, not AWS Cost Explorer.

Cloudwatch monitoring schemes —

- Basic. 5 minutes.
- Detailed. 1 minute.
- Custom. Can be down to 1 second.

Transferring data from an EC2 instance to Amazon S3, Amazon Glacier, Amazon DynamoDB, Amazon SES, Amazon SQS, or Amazon SimpleDB in the same AWS Region has no cost at all.

We can use signed URLs and signed cookies with Cloudfront to protect resources.

Amazon MQ is a message queue which supports industry standard messaging protocols.

Slower login time and 504 errors in front of Cloudfront can be optimized by —

- Lambda @ Edge.
- Setting up an Origin Failover Policy.

AWS Shield is a service that protects resources against DDoS attacks to EC2, ELB, Cloudfront and Route53.

AWS IoT Core is a managed service that lets IoT devices connect and interact with AWS applications and resources.

The following storage have encryption at rest by default —

- AWS Glacier
- Storage Gateway in S3

Perfect Forward Secrecy is supported by —

- Cloudfront
- Elastic Load Balancing

Enabling multiple domains to serve HTTPS over same IP address —- Generate an SSL cert with AWS Certificate Manager and create a Cloudfront distribution. Associate cert with distribution and enable Server Name Indication (SNI).

Classic Load Balancer does not support SNI, we have to use Application Load Balancer or Cloudfront.

The following services enable us to run SQL queries directly against S3 data —

- AWS Athena
- Redshift Spectrum
- S3 Select

By default, each workflow execution can run for a maximum of 1 year in Amazon SWF.

In AWS SWF, a decision task tells the decider the state of the workflow execution.

Third party SSL cert can be imported into —

- AWS Certificate Manager
- IAM Certificate Store