Read https://www.selikoff.net/2020/09/05/studying-for-aws-associate-architect-in-the-time-of-covid-19/ before reading my study notes (also these notes probably have errors in them; they are just what I used)

General

- # edge locations > #availability zones > # regions

	Basic	**Developer**	**Business**	**Enterprise**
Description	Free	Email during bus hours	24x7	Must spend 15K/month
Contacts		1	Unlimited	Tech acct mgr

IAM (Identity and Access Management)

Group	Collection of users
Roles	Global, assign to AWS resources
Policy	JSON, effect (allow, deny), action (ex: *) and resource

S3 (Simple Storage Service)

- Object based, global namespace, up to 100 buckets/account
- 5TB/file, multi part recommended 100MB, required 5GB
- 11 9's durability

	Availability	**Notes**
S3	99.99%	Survives 2 data center failures

S3 IA	99.9%	Lower storage fee, pay for retrieval
S3 Intelligent Tiering	varies	Automatically moves data
S3 One Zone IA	99.5%	Only stored in one availability zone Similar to old RRS (reduced redundancy storage)
Glacier	99.99% (SLA 99.9%)	Minutes- hours to retrieve (expedited/standard/bulk)
Glacier Deep Archive	99.99% (SLA 99.9%)	12 hours to retrieve data
Snowball		Physical device, petabyte scale (50TB or 80 TB), for S3 import/export
Snowball Edge		100TB, has compute too (ex: lambda). Can support local workloads remotely/offline. Can cluster them

URL formats

- Path style - s3.region.amazonaws.com/bucket/path (deprecated)
- Static hosting – your domain or bucket.s3-website.region.amazonaws.com
- Virtual style – bucket.s3.region.amazonaws.com/path
- Legacy – bucket.s3.amazonaws.com (discouraged)

Read after write consistency	Creation
Eventual consistency	Updates/deletes

- 30 days in S3 before move to another type. 90 days minimum Glacier
- Cross region replication
    - High availability/disaster recovery
    - Automatically replicates each time write object
    - Must have versioning on source/target
    - Existing files not replicated unless changed
    - Delete markers not replicated nor deleting individual versions of specific files
- Transfer acceleration – uses CloudFront to transfer long distances
- Versioning + can set up multifactor for deletes
- Costs: storage, requests/data retrievals, data transfer, management/replication
- Can create tag on bucket, log on individual files, lock objects in lock
- Bucket not public by default
- Control access with bucket ACL (object level) or bucket policy (bucket level)
- Encryption In Transit – SSL/TLS
- Encryption at Rest
    - S3 Managed Keys – SSE-S3 – Amazon fully manages
    - Server Side Encryption with Customer Provided Keys – SSE-C

    - AWS Key Management Services – SSE-KMS – customer and Amazon manage keys. Audit trail
- Versioning can be suspended but not disabled
- Stores all objects. Uses delete marker if delete
- Lifecycle rules - can add rules for current and previous versions of objects or both. Ex: transition to storage class X after Y days. Can also expire objects after Z days
- S3 Object Lock/Glacier Vault Lock
    - WORM (write once, read many)
    - Can prevent deletion for fixed time or indefinitely
    - Governance mode – need special perms to delete version or alter lock settings
    - Compliance mode – nobody can delete version or alter lock settings. Not even root user
    - Can additionally place legal hold on a version
- Multipart uploads – recommended if over 100MB. Required over 5GB
- S3 Byte Range Fetches – Parallelizes download. Can redownload just one part
- S3 Select and Glacier Select - use SQL to get subset of data
- Can share S3 buckets across account using
    - Bucket policies/IAM – programmatic access only, applies for whole bucket
    - Using Bucket ACLs/IAM – programmatic access only, individual objects
    - Cross account IAM roles – programmatic and console access
- Can change storage class when replicate
- Transfer Acceleration – URL to upload to edge location. Can go to UI and compare speed to upload directly vs different regions edge location
- DataSync – agent on your server, accesses file system, encrypts and sends it to AWS S3 or EFS
- CloudFront – CDN (Content Delivery Network)
    - Edge location – where cached
    - Origin – S3, EC2, ELB or Route 53
    - Distribution – CDN – collection of edge locations
    - Read/write
    - Cached for TTL (time to live) in seconds. Charge to clear cached objects
    - Web or RTMP (media streaming)
    - Can restrict access and require signed URLs/cookies
- Signed URLs/cookies
    - Signed URLs – individual files
    - Signed cookie - for multiple files
    - Attach policy – URL expiration, IP address, which accounts can create
    - Use OAI (origin access identity) so users must go through CloudFront
    - CloudFront Signed URL – does not have to be EC2, caching, managed by root user, can filter by data, path, IP address
    - S3 Signed URL – have same permissions as IAM user who creates, limited lifetime. Users must be able to access S3 directly

- Storage Gateway
    - physical or VM – files (nfs or smb), volumes (hard drives), tape
    - cached (local cache) or stored (all local with async aws bckup)

EC2 (Elastic Compute Cloud)

- Compute, scale up/down (kills in AZ with most, oldest one or one closest to next billing hour)

On-demand	Fixed rate	New/spiky apps
Reserved	1 or 3 year contract Standard or Convertible (change EC2 type) Scheduled if need just certain window	Steady state
Spot	Bid. Charged if terminate. Decide in 2 minutes if price rises.	Flexible start/end, low prices, urgent capacity
Dedicated Hosts	Physical server	Regulatory requirements, server bound licenses

- Choose AMI, instance type, # instances, VPC, subnet (availability zone), shutdown behavior, optional bootstrap scripts
- Can encrypt any volume including root on creation
- Termination protection off by default
- Root EBS volume deleted when instance terminated by default
- Spot fleets – collection of spot (and on demand). Tries to maintain capacity/re-launch. Strategies: capacityOptimized, diversified (across pools), lowestPrice (default), instancePoolsToUseCount (with lowest prize)
- Can run bootstrap/startup script in bash. Ex: copy files from s3
- Metadata (about instance) vs user data (bootstrap script)
- Placement groups, must stop instance to move between groups
    - Clustered – grouping with zone, good if need low latency/high throughput
    - Spread – distinct hardware, good if small number critical instances that must be kept separate. Can span availability zones
    - Partitioned – each subgroup (of multiple instances) has own racks
- EFS (Elastic File System) – can share volume across EC2 (unlike EBS), NFS, pay for storage used, up to petabytes, stored in multiple availability zones, read after write consistency, Linux file based storage only
- Amazon FSx
    - Managed Windows Server, SMB (Windows Server Message Block)
    - Lustre file system – higher throughput/performance/big data. Can store on S3
- HPC (high performance computing)
    - Data transfer – Snowball/snowmobile, AWS DataSync (from S3/filesystem), Direct Connect (to on-prem)

- Computer/networking – GPU/CPU optimized EC2, Spot instances/fleets, cluster placement groups, enhanced networking, elastic network adapters, elastic fabric adapters
- Storage – EBS provisioned IOPS, instance store, S2, EFS, Amazon FSx for Luster
- Orchestration/automation – AWS Batch, AWS Parallel Cluster

EBS (Elastic Block Store)

- For use with EC2 (same availability zone as EC2), like virtual hard drive
- Snapshot of volume are on S3. Incremental (only changed blocks)
- Automatically replicated within availability zone

General Purpose SSD (gp2)	Most workloads, transaction, small/random operations
Provisioned IOPS SSD (io1)	Highest performance SSD, mission-critical apps, databases
Throughput Optimized HDD (st1)	Low cost, infrequently accessed, throughput intensive, big data, large/sequential operations, not bootable. Storage optimized
Cold HDD (cs1)	Lowest cost, infrequently accessed, file servers, not bootable
EBS Magnetic (Standard)	Previous generation

- Root volume has snapshot name referenced
- Should stop instance to snapshot root device but not required
- Turn snapshot into AMI to move to different availability zone. Then launch new EC2 instance from AMI
- Can change volume size including size and storage type while running
- Can select AMI based on region, OS, 32/64 bit, launch permissions
- Storage
  - Instance store – temporary, from S3 template
  - EBS – root device from AMI. By default, deleted on termination, but can configure to keep
- ENI (elastic network interface) – virtual network card, 1+ private IPv4, 1 elastic IPv4 per private IP, 1 public IPv4, 1+ IPv6, 1+ security groups, low cost, separate network
- EN (enhanced networking) – virtualization to provide network capabilities. More powerful than ENI. Higher bandwidth (10-100GB/sec), lower CPU, when want good network performance. Better than Virtual Function (VF) because higher bandwidth. SR-IOV (single root i/o virtualization)
- EFA (elastic fabric adapter) – network device for high performance computer and machine learning
- Now, can create root volumes when provision EC2, can also encrypt by launching encrypted root instance
- Can only share unencrypted snapshots

- Can hibernate on demand/reserved instances up to 60 days if encrypted and less than 150GB

Database

RDS	Aurora (serverless), MariaDB, MySQL, Oracle, PostgresSQL, SQL Server
Aurora	• MySQL/Postgres compatible • 3+ availability zones, 2 copies per zone • Up to 15 read replicas
Features	Cannot access OS, up to 35 day backup
Multi-AZ	Disaster recovery, synchronous replication, auto failover
Read Replicas	Up to 5 for performance, async replication, any region, can promote to own db
DynamoDB	• NoSQL, low latency, 3 data centers • Eventually consistent reads (default) or strongly consistent reads (all writes prior to read) • Transactions – Two underlying reads/writes for prepare/commit. Up 25 items or 4MB
DAX (DynamoDB Accelerator)	In-memory write through cache, 10X performance
DMS (Db migration service)	On prem/AWS, different dbs via schema conversion tool
RedShift	Business Intelligence, data warehousing Leader (connections), compute nodes Data compressed by column 1 day retention, can increase to 35 Spectrum can query S3 files
EMR (elastic map reduce)	Three times faster than Spark. Nodes in cluster Master node (track status), core node (run tasks/store data), task node (optional compute only)

- ElastiCache
    - In-memory cache – Memcached(multi-threading, horizontal scaling), Redis (advanced data type, multi-az, backup restore)
    - Caching – CloudFront, API Gateway, ElastiCache, DAX (DynamoDB Accelerator)
- OLAP – online analytics processing – ex: big data
- OLTP – online transaction processing – ex: straightforward web
- When create cluster, can set up master node to write log to S3 at 5 minute intervals

Security

- CORS – server can relax same origin policy

- Directory Services
    - SSO with EC2 in domain
    - 2 domain controllers (DC) - different availability zone
    - AWS Managed AD – in cloud + AD Trust – extend to on-prem
    - Simple AD – standalone, small (up to 500 users) or large (up to 5k users)
    - AD Connector – proxy for on-prem AD
    - Cloud Directory – organize data across dimensions – ex: org chart
    - Cognito User Pools – for SaaS, sign in/up for web/mobile, use social media SSO
- Cloud HSM – hardware, FIPS 140-2Level3
- Secrets manager – similar to System Manager Parameter Store, but charged per secret, can rotate/randomize secret, can apply RDS key/db

CloudWatch

- Performance/monitoring tool
- Every 5 minutes by default. Or detailed monitoring every minute
- Host level metrics - CPU, network, disk, status check
- Can set up custom metrics – ex: disk space utilization, disk swap utilization, memory utilization, page file utilization and log collection
- Set billing alarm – can be static value or anomaly detection
- Can set up dashboards, alarms (when threshold met), events (state changes)
- Cloudwatch Events can trigger ECS tasks
- Can install agent

Route 53

- IPv4 – 32 bits. IPv6 – 128 bits
- Amazon has own domain registrar, can buy domain there. Can take up to 3 days to register
- SOA (Start of Authority) Record – name of server, admin, current version, TTL
- NS - name server records – direct top level domain server to content NDS server (ex: .com)
- TTL - time to live in seconds, how long cached on server/user's machine. Default is 48 hours
- CNAME - canonical name, resolve one name to another. Ex: for ELB
- A record - map DNS name to IP
- Alias record (route 53 specific) – maps one name to another, must use for zone apex/naked domain (ex: google.com), choose over CNAME
- MX record – for mail
- PTR – loop up name when know IP
- ELB uses DNS name, not IP

- Record sets are groups of IPs

Simple	Returns random IP
Weighted	Adds up weights and returns based on ratio/%. Ok if not 100%
Latency	Fastest response
Failover	Active/passive
Geolocation	Based on source IP
Geoproximity	Traffic flow only, by location. Bias expands/contracts regional size

WAF (Web application firewall)

- Layer 7 aware (network)
- Monitor http/https requests going to CloudFront, ALB or API Gateway
- Can control access to content – ex: IP address, query string params, country, strings, length, malicious content
- Can allow/block requests except for those specified or count/monitor matches

AWS Organization

- Multiple accounts
- Root for billing only, don't deploy services directly
- Organizational units have own accounts, but billed together. Create OUs in UI
- Service control policy (SCP) can limit what AWS services an OU can use or on accounts directly
- Rate determined across accounts so pay lower rate combined
- Linked accounts stay independent for access

Failover

Pilot light	Minimal and can create rest
Warm	Scaled down version always running

VPC (Virtual private cloud)

- Flow: internet gateway to virtual private gateway -> route table -> network ACL -> security group within subnet
- Allowed 5 per region
- Default VPC – so can deploy fast, auto connect to internet, each EC2 instance has a public and private IP
- Custom VPC – private by default
- Peering – connect VPC to another directly using private IPs. Act as if on same network. Can peer with other Amazon accounts/regions. Not transitive
-

Security Group	**NACL**

Stateful	Stateless (create outbound if want)
Allow rules only	Allow and deny rules
Default: no incoming/all outgoing	Default: allow all in default VPC, deny all in custom VPC
Evaluate all rules before decide	Evaluate rules lowest to highest
Type (HTTP/SSL/TCP), port, CIDR	Type (HTTP/SSL/etc), protocol (TCP), port, source IP
For instance in VPC	For VPC/subnet
	Good for blocking specific IPs

- If use ALB, can block IP at ALB.
- EC2 – can have host based firewall to block IP
- Subnet can be private or public
    - IPs can be unique with subnet
    - Within single availability zone
    - One NACL at any time (but NACL can associate with multiple subnets)
    - Changes take effect immediately
    - Private IPs reserved
        - 10.0.0.0-10.255.255.255 (10/8)
        - 172.16.0.0-172.31.255.255 (172/16)
        - 192.168.0.0-192.168.255.255 (192.168/16)
- Address space is /16 - /28 for subnet (and /16 or smaller for VPC). However, 5 IPs are reserved In the block – first four and last one
- Route table
    - For subnets to talk to each other
    - Register which subnets can use internet gatweway
    - Can have multiple route tables (ex: one for public routes)
- When create VPC, get route table/NAC/security group, but not subnet/internet gateway
- Availability zones are randomized so “a1” doesn’t mean same thing to all accounts
- NAT – Network Address Translation, for outbound internet communication
    - NAT Gateway better than NAT instance
    - NAT Instance – single EC2 instance, must disable SrcDestCheck to make it a gateway, must be in public subnet, must be route from private subnet to NAT instance, increase size if bottleneck, can do HA with Autoscaling Groups, behind security group
    - NAT Gateway – HA, redundant inside availability zone, cannot span AZ, preferred by enterprise, scales/patches automatically, not associate with security group, automatically gets public IP, have to update rout tables, create multiple gateways so not dependent on one AZ
- Load balancer – need two public subnets
- VPC Flow logs – capture data about IP traffic and storing in CloudWatch. VPC, subnet or network interface level, Peered VPC must be in same account. Can tag. Cannot change config once create

- Bastion host
    - Special purpose computer to withstand attacks
    - Only contains one app (ex proxy server)
    - In public subnet
    - SSH/RDP (remote desktop protocol) to the bastion host. Also called jump box
    - Forwards to private subnet
    - Needs network load balance since IP based
- Internet gateway
    - One per VPC
    - Retains private, but not public, IP on restart
    - Egress only internet gateway - Ipv6 traffic in VPC can exit, nothing can enter
- Internet: 0.0.0.0 IP V4 and ::/0 – IP V6
- Direct Connect – connect on-prem to AWS, good for high throughput, stable/reliable secure connection. To create:
    - Public virtual interface
    - Customer gateway
    - Virtual private gateway & attach to VPC
    - VPN connection and set on customer gateway
- Global Accelerator – directs to optimal endpoints, comes with two static APIs
- VPC endpoint – virtual device, connect from EC2 to AWS services without gateway, connect within Amazon network. Interface (ENI with private IP) and Gateway (for S3 and DynamoDB)
- PrivateLink – connect to another VPC, doesn't require peering, uses Network Load Balancer on service VPC and ENI on customer VPC. Better when would have needed to peer to dozens of VPC
- Transit Gateway – hub in hub & spoke. Simulates transitive peering. Supports IP multicast, can use across multiple accounts
- VPN CloudHub – Virtual Private Gateway to connect on-prem site
- Costs –
    - Free - to connect into VPC, private IP with VPC
    - Paid – between availability zones, public IP (costs more)

Load Balancer

ALB (app)	HTTP, layer 7, can see inside requests
NLB (network)	TCP, layer 4, fastest
CLB (classic)	Legacy, TCP/HTTP, cannot see inside requests
Scaling	• Groups of EC2 instances • Configuration templates – launch template or launch config for EC2 instances. Instructions on what to launch ○ Scaling options: on demand by policy/usage (most popular), same # of instances at all times, manually, scheduled, predictive

- Includes health check to ping – InService or OutOfService
- Use DNS name, not IP address
- X-Forwarded-For gives original IPv4 address of user
- Classic load balancer does sticky sessions at EC2 level. ALB does them at target group level
- Cross zone load balancing – Route 53 sends to different availability zones. Load balancer can send to another zone if busy/has less capacity. All instances have same amount of traffic.
- Path pattern – send to different target group based on substring on url

Other services

CloudTrail	For auditing - records management and API calls
Kinesis	• Streams – producers/consumers, shards, retained one day • Firehose – delivery streams/records/destinations, no retention
AWS Shield	DDoS
WAF (Web App firewall)	Filter HTTP requests ex: SQL injection, XSS
SQS (Simple queue service)	• Pull based from queue • Standard queue – message 1+ times, roughly in order • FIFO – message 1 time, exactly in order • Visibility timeout – another readers sees if not processed, default 12 hours • Long polling – returns when message/timeout • Message retention of 2 weeks
SWF (Simple workflow)	• Coordinate work in tasks (run once) run up to a year • Domain – group of workflow • Workflow starters -> deciders -> activity workers • Executions up to a year
Step Functions	• Serverless, microservice orchestration
SNS (Simple Notification Service)	• Push based using topics • To mobile/SMS/SQS/HTTP
Athena	Serverless, Query S3 with SQL. Supports JSON (not XML)
Macie	ML/NLP to discover PII in S3 or suspicious CloudTrail activity
Elastic Transcoder	Converts media files
Elastic Beanstalk	Compute Service, UI based deployment
Lambda	• Node.JS, Java, Python, C#, PowerShell • Trigger from SQS, SNS, S3, Dynamo, API Gateway, etc • Pay by request, RAM and duration • If fails, retries. If fails again, sends to SQS dead letter queue or SNS topic

SAM (Serverless Application Model	CloudFormation extension
Xray	Distributed debugging
API Gateway	• Entry point for EC2, lambda, dynamo or web app • Exposes HTTPS for REST, multiple versions of API • Caching, scaling, request/response transformation • Track/control usage by API key • Stage – deployment
Cognito	• Web Identity Federation –temp token after SSO • Identify Broker – gets JWT (JSON web token) • Sign up/sign in, guest access, mobile device data sync • User pools – directories • Identity pools – provides temporary credentials
ECS (Elastic Container Service)	• Container orchestration, clusters of EC2/Fargate • Task definition – defines app – like Dockerfile • Container Definition – inside task definition • Task – single running copy of containers • Service – scaling range • Registry – storage
Fargate	• Serverless container engine
EKS (Elastics Kubernetes Service)	• Kubernetes, pods
Trusted Advisor	• Real time guidance for best practices
Inspector	• Security recommendations

Http Codes

- 200 – Good/success
- 504 – Gateway timed out (app not responding)

CIDR – cidr.xyz website helps visualize

- /32 – one IP
- /16 – two sets of 255
- /8 – three sets of 255
- /0 – whole subnet

Protocols

- HTTP/S – 8080/443
- SSH - 22
- FTP/TCP – 21