AWS Solution Architect Associate Exam Notes

For more information on AWS, visit aws.amazon.com

Description

Notes and information that were collected while studying and prepping for the AWS SA Associate Exam.

Topic	Answer
Exam Time:	80 Minutes
No. Questions:	60 Questions
Question Types:	Scenario and Multiple Choice
Passing Score:	~ 70%
Validity Period:	2 years
Renewal Exam:	1/2 price off

General

> Amazon History:

- 2003 - Chris Pinkman and Benjamin Black presented a paper on what Amazon's internal infrastructure should look like and suggested sellingit as a service
- 2004 - SQS the first AWS service launched
- 2006 - Official AWS Launch
- 2007 - 180K devs on platform
- 2010 - Amazon.com moved to AWS PlatformRegion is geographical area which consists of at least 2 AZ's
- 2012 - First Re-Invent conference in Las Vegas
- 2013 - Certifications Launched
- 2014 - AWS commited to achieve 100% renewable energy usage for its global footprint
- 2015 - AWS broke out it's revenue, 6 Billion USD per annum and growing close to 90% year after year

Global Infrastructure:

Regions Vs. Availability Zones:
A Region is geographical area which consists of at least 2 Availability Zone's or AZ's. An AZ is simply a data center.

- 14 Regions with 38 Availability Zones
- Projected to spin up 4 additional regions, with 9 additional AZ's over the next year
- Edge locations are CDN endpoints for CloudFront. Currently there are over 50 edge locations. Not the same as an AZ
- Maximum Response time for Business is 1 hour
- Services such as CloudFormation, Elastic Beanstalk, Autoscaling, and OpsWorks are free however resources generated by these services are not free

Tags:

- Key/Value pairs attached to AWS resources
- Metadata (data about data)
- Sometimes can be inherited (Auto-scaling, CloudFormation, Elastic Beanstalk can create other resources)
- Resource Groups make it easy to group your resources using the tags that are assigned to them
- You can group resources that share one or more tags
- Resource groups contain info such as region, name, health checks
- Contains specific info such as pub/private IP's in EC2 instances, Port configs for ELB, Database engine in RDS

IP Address Info:

- Query meta-data:
  - curl http://169.254.169.254/latest/meta-data/
  - get http://169.254.169.254/latest/meta-data/
- META DATA only there is not USER DATA
- AWS uses the first 4, and last IP addresses of a subnet:
  - x.x.x.0 - Network Address
  - x.x.x.1 - Gateway Address
  - x.x.x.2 - DNS Address
  - x.x.x.3 - Future Allocation Address
  - x.x.x.255 - Broadcast Address

Consolidated Billing:

- Accounts roll for customers:
  - Paying account is independent, can not access resources of the other accounts
  - Linked accounts are independent from one another
  - Currently there is a limit of 20 linked accounts for consolidated billing (soft limit)
  - One bill per AWS account
  - Easy to track charges and allocate costs between linked accounts
  - Volume pricing discount
  - Resources across all linked accounts are tallied, and billing is applied collectively to allow bigger discounts

Active Directory:

- Provides single sign-on to the AWS console, which authenticates directly off of your Active Directory infrastructure
- Uses Secure Assertive Markup Language (SAML) authentication responses.
- Behind the scenes, sign-In's use the AssumeRoleWithSAML API to request temporary security credentials and then constructs a sign-in URL for the AWS Management Console
- Browser will then receive the sign-in URL and will be redirected to the console
- You always authenticate against AD first, and then are granted security credentials that allow you to log into the AWS console

> Best Practices:

- Business Benefits of Cloud:
  - Almost 0 upfront infrastructure investment
  - Just in time Infrastructure
  - More efficient resource utilization
  - Usage based costing
  - Reduced time to market
- Technical Benefits of Cloud:
  - Automation - Scriptable infrastructure
  - Auto-Scaling
  - Proactive Scaling
  - More efficient development life cycle
  - Improved testability
  - DR and Business Continuity
  - Overflow the traffic to the cloud
- Design for Failure
  - Rule of thumb: Be a pessimist when designing architectures in the cloud
  - Assume things will fail, always design implement and deploy for automated recovery from failure
  - Assume your hardware will fail
  - Assume outages will occur
  - Assume that some disaster will strike your application
  - Assume that you will be slammed with more than the expected number of requests per second
  - Assume that with time your application software will fail too
- Decouple your components:
  - Think SQS
  - Build components that do not have tight dependencies on each other so that if one component dies, fails, sleeps, or becomes busy, the other components are built so they can continue to work as if no failure is happening. Build each component as a black box

Service Limits:

Each service has the default limits defined, to see the official AWS documentation on service limits, check here

Networking:

> **VPC (Virtual Private Cloud):**

Lets you provision a logically isolated section of the AWS Cloud where you can launch AWS resources in a virtual network that you define. You have complete control over your virtual networking IP ranges

a virtual network that you define. You have complete control over your virtual networking, IP ranges, creation of subnets and configuration of route tables and network gateways.

- Virtual data center in the cloud
- Allowed up to 5 VPCs in each AWS region by default. This limit can be increased with a support ticket request
- All subnets in default VPC have an Internet gateway attached
- Multiple IGW's can be created, but only a single IGW can be attached to a VPC.. No exceptions
- Again, You can only have 1 Internet gateway per VPC
- Each EC2 instance has both a public and private IP address
- If you delete the default VPC, the only way to get it back is to submit a support ticket
- This answer is correct for the current iteration of tests, however AWS has now crated a mechanism in the console that allows you to recreate a default VPC
- By default when you create a VPC, a default main routing table automatically gets created as well.
- Subnets are always mapped to a single AZ
- Subnets can not be mapped to multiple AZ's
- /16 is the largest CIDR block available when provisioning an IP space for a VPC
- /28 is the smallest CIDR block available when provisioning an IP space for a VPC
- Amazon uses 3 of the available IP addresses in a newly created subnet
    - x.x.x.0 - Always subnet network address and is never usable
    - x.x.x.1 - Reserved by AWS for the VPC router
    - x.x.x.2 - Reserved by AWS for subnet DNS
    - x.x.x.3 - Reserved by AWS for future use
    - x.x.x.255 - Always subnet broadcast address and is never usable.
- 169.254.169.253 - Amazon DNS
- By default all traffic between subnets is allowed
- By default not all subnets have access to the Internet. Either an Internet Gateway or NAT gateway is required for private subnets
- A security group can stretch across different AZ's
- Security Groups are stateful (Don't need to open inbound and outbound, if inbound is allowed, outbound is auto allowed)
- Network Access Control Lists (NACLs) are stateless (Must define both inbound and outbound rules)
- You can also create Hardware Virtual Private Network (VPN) connection between your corporate data center and your VPC and leverage the AWS cloud as an extension of your corporate data center
- VPC Flow Logs:
- VPC Flow Logs is a feature that enables the user to capture information about the IP traffic going to and from network interfaces in your VPC
- Flow log data is stored using Cloudwatch Logs
- When Flow log data is collected it can be viewed and its data can be retrieved within Cloudwatch
- Flow logs can be created at 3 different levels, VPC, Subnet and Network Interface levels
- Flow logs via Cloudwatch can be configured to stream to services such as Elasticache, or Lambda
- You cannot enable flow logs for VPC's that are peered with your VPC unless the peer VPC is in your account
- You cannot tag a flow log
- After you have created a flow log, you cannot change its configuration, for example you cannot associate a different role with the flow log
- Not all traffic is monitored:
    - Traffic generated by instances when they contact Route53 is not monitored or logged
    - If you use your own DNS server, then all traffic to that DNS server is logged
    - Traffic generated by a Windows instance for Windows license activation is not monitored or logged
    - Traffic to and from the metadata service (169.254.169.254) is not monitored or logged
    - DHCP traffic is not monitored or logged
    - Traffic to the reserved IP address for the default VPC router is not monitored or logged
- Network Address Translation (NAT) Instances:
    - When creating a NAT instance, disable Source/Destination checks on the instance or you could encounter issues
    - NAT instances must be in a public subnet
    - There must be a route out of the private subnet to the NAT instance in order for it to work
    - The amount of traffic that NAT instances support depend on the size of the NAT instance. If bottlenecked, increase the instance size

    - If you are experiencing any sort of bottleneck issues with a NAT instance, then increase the instance size
    - HA can be achieved by using Auto-scaling groups, or multiple subnets in different AZ's with a scripted fail-over procedure
    - NAT instances are always behind a security group
- Network Address Translation (NAT) Gateway:
    - NAT Gateways scale automatically up to 10Gbps
    - There is no need to patch NAT gateways as the AMI is handled by AWS
    - NAT gateways are automatically assigned a public IP address
    - When a new NAT gateway has been created, remember to update your route table
    - No need to assign a security group, NAT gateways are not associated with security groups
    - Preferred in the Enterprise
    - No need to disable Source/Destination checks
    - More secure than a NAT instance
- Network Access Control Lists (NACLS):
    - NACL's are stateless, meaning both inbound and outbound rules must be configured for traditional request/response model
    - Numbered list of rules that are evaluated in order starting at the lowest numbered rule first to determine what traffic is allowed in or out depending on what subnet is associated with the rule
    - The highest rule number is 32766
    - Start with rules starting at 100 so you can insert rules if needed
    - NACL's have separate inbound and outbound rules, and each rule can either allow or deny traffic
    - The Default NACL will allow ALL traffic in and out by default
    - Custom NACL's by default will deny all inbound and outbound traffic until allow rules are added
    - You must assign a NACL to each subnet, if a subnet is not associated with a NACL, it will allow no traffic in or out
    - NACL rules are stateless, established in does not create outbound rule automatically
    - You can only assign a single subnet to a single NACL
    - When you associate a NACL with a subnet, any previous associations are removed
    - You can associate a single NACL with multiple subnets
    - Each subnet in your VPC must be associated with a NACL. If you don't explicitly associate a subnet with an ACL, the subnet automatically gets associated with the default ACL
    - You can block IP addresses using NACLs not Security Groups
- VPC Peering:
    - Connection between two VPCs that enables you to route traffic between them using private IP addresses via a direct network route
    - Instances in either VPC can communicate with each other as if they are within the same network
    - You can create VPC peering connections between your own VPCs or with a VPC in another account within a SINGLE REGION
    - AWS uses existing infrastructure of a VPC to create a VPC peering connection. It is not a gateway nor a VPN, and does not rely on separate hardware
    - There is NO single point of failure for communication nor any bandwidth bottleneck
    - There is no transitive peering between VPC peers (Can't go through 1 VPC to get to another)
    - Hub and spoke configuration model (1 to 1)
    - Be mindful of IPs in each VPC, if multiple VPCs have the same IP blocks, they will not be able to communicate
    - You can peer VPC's with other AWS accounts as well as with other VPCs in the same account
- VPC Endpoints:
- Allows internal resources such as EC2 instances to reach various AWS services without having to traverse the public internet to get to the service
- When you use an endpoint, the source IP address from your instances in your affected subnets for access the AWS service in the same region will use private IP address's instead of public IP address's
- When configuring VPC endpoints, existing connections from your affected subnets to the AWS service that use public IP address's may be dropped

Resource or Operation	Default Limit	Comments
VPCs per region:	5	The limit for Internet gateways per region is directly correlated to this one. Increasing this limit will increase the limit on Internet gateways per region by the same amount.
Subnets per VPC:	200	
Internet gateways per region:	5	This limit is directly correlated with the limit on VPCs per region. You cannot increase this limit individually; the only way to increase this limit is to increase the limit on VPCs per region. Only one Internet gateway can be attached to a VPC at a time.
Customer gateways per region:	50	
VPN connections per region:	50	
VPN connections per VPC (per virtual private gateway):	10	
Route tables per VPC:	5	Including the main route table. You can associate one route table to one or more subnets in a VPC.
Routes per route table (non-propagated routes):	50	This is the limit for the number of non-propagated entries per route table. You can submit a request for an increase of up to a maximum of 100; however, network performance may be impacted.
BGP advertised routes per route table (propagated routes):	5	You can have up to 100 propagated routes per route table; however, the total number of propagated and non-propagated entries per route table cannot exceed 100. For example, if you have 50 non-propagated entries (the default limit for this type of entry), you can only have 50 propagated entries. This limit cannot be increased. If you require more than 100 prefixes, advertise a default route.
Elastic IP addresses per region for each AWS account:	5	This is the limit for the number of VPC Elastic IP addresses you can allocate within a region. This is a separate limit from the Amazon EC2 Elastic IP address limit.
Security groups per VPC:	500	
Inbound or outbound rules per security group:	50	You can have 50 inbound and 50 outbound rules per security group (giving a total of 100 combined inbound and outbound rules). If you need to increase or decrease this limit, you can contact AWS Support — a limit change applies to both inbound and outbound rules. However, the multiple of the limit for inbound or outbound rules per security group and the limit for security groups per network interface cannot exceed 250. For example, if you want to increase the limit to 100, we decrease your number of security groups per network interface to 2.

Security groups per network interface:	5	If you need to increase or decrease this limit, you can contact AWS Support. The maximum is 16. The multiple of the limit for security groups per network interface and the limit for rules per security group cannot exceed 250. For example, if you want 10 security groups per network interface, we decrease your number of rules per security group to 25.
Network interfaces per instance:	N/A	This limit varies by instance type. For more information, see Private IP Addresses Per ENI Per Instance Type.
Network interfaces per region:	350	This limit is the greater of either the default limit (350) or your On-Demand instance limit multiplied by 5. The default limit for On-Demand instances is 20. If your On-Demand instance limit is below 70, the default limit of 350 applies. You can increase the number of network interfaces per region by contacting AWS Support, or by increasing your On-Demand instance limit.
Network ACLs per VPC:	200	You can associate one network ACL to one or more subnets in a VPC. This limit is not the same as the number of rules per network ACL.
Rules per network ACL:	20	This is the one-way limit for a single network ACL, where the limit for ingress rules is 20, and the limit for egress rules is 20. This limit can be increased upon request up to a maximum if 40; however, network performance may be impacted due to the increased workload to process the additional rules.
Active VPC peering connections per VPC:	50	If you need to increase this limit, contact AWS Support . The maximum limit is 125 peering connections per VPC. The number of entries per route table should be increased accordingly; however, network performance may be impacted.
Outstanding VPC peering connection requests:	25	This is the limit for the number of outstanding VPC peering connection requests that you've requested from your account.
Expiry time for an unaccepted VPC peering connection request:	1 week (168 hrs)	
VPC endpoints per region:	20	The maximum limit is 255 endpoints per VPC, regardless of your endpoint limit per region.
Flow logs per single eni, single subnet, or single VPC in a region:	2	You can effectively have 6 flow logs per network interface if you create 2 flow logs for the subnet, and 2 flow logs for the VPC in which your network interface resides. This limit cannot be increased.
NAT gateways per Availability Zone:	5	A NAT gateway in the pending, active, or deleting state counts against your limit.

For additional information about VPC Limits, see Limits in Amazon VPC

> **Direct Connect:**

Direct Connect:

AWS Direct Connect lets you establish a dedicated network connection between your network and one of the AWS Direct Connect locations. Using industry standard 802.1q VLANs.

- DX or Direct Connect makes it easy to establish a dedicated network connection from your premises to AWS
- Using DX, you can establish private connectivity between AWs and your data center, office or collocation environment
- Requires a dedicated line such as MPLS, or other circuit ran from tel-co.
- From this line, you would have a cross connect from your on-premises device direct to AWS data centers
- Using DX, can reduce network costs, increase bandwidth throughput and provide a more consistent network experience then internet based connections
- Lets you establish a dedicated network connection between your network and one of the AWS DX locations
- Uses industry standard 802.1Q VLANs
- Dedicated connections can be partitioned into multiple virtual interfaces
- Same connection can be used to access public resources such as objects stored in S3 using public IP's and private resources such as EC2 instances running in a VPC using private IP's, all while maintaining network separation between the public and private environments
- Virtual interfaces can be reconfigured at any time to meet changing needs
- Offers more bandwidth and a more consistent network experience over using VPN based solutions
- VPC VPN connections utilize IPSec to establish encrypted network connectivity between your intranet and your AWS VPC over the internet
- VPN connections can be configured in minutes and are a good solution if you have an immediate need
- DX does NOT involve the internet, instead, it uses dedicated private network connections between your intranet and AWS VPC

Resource or Operation	Default Limit	Comments
Virtual interfaces per AWS Direct Connect connection:	50	
Active AWS Direct Connect connections per region per account:	50	
Routes per Border Gateway Protocol (BGP) session:	100	This limit cannot be increased.

Route 53:

Amazon Route 53 is a highly available and scalable cloud Domain Name System (DNS) web service.

- ELBs do not have a pre-defined IPv4 address. You resolve them using a DNS name
- The Apex domain record MUST be an A record or an alias
- Aliases map AWS resources to zone records
- Alias records you are not charged for, CNAME records you are charged for
- Always chose an alias record, over a CNAME record, as alias records are free, and can be mapped to a domain apex record where CNAMES cannot
- Limit of 50 Domain Names can be managed in Route53. This limit can be raised by support.
- Route 53 Routing Policies:
    - Simple
        - Default routing policy when you create a new record set
        - Most common when you have a single resource that performs given function for your domain

        - Route53 will respond to DNS queries that are only in the record set.
        - No Intelligence is built into the response
    - Weighted
        - Let you split traffic based on different weights defined
        - 1 AZ can be set to 90%, and another can be set to 10% for example
    - Latency
        - Allows you to route your traffic based on the lowest network latency for your end user. (Which region will give them the fastest response time)
        - Create a latency resource record set in each region that hosts your website
        - When Route53 receives a query for your site, it selects the latency resource for the region that gives the user the lowest latency
    - Fail-over
        - Used when you want to create an active/passive set up
        - Route53 will monitor the health of your primary site using a health check
        - Health check monitors the health of your endpoints
    - Geo-location
        - Lets you choose where your traffic will be sent based on the geographic location of your users
        - Good if you want all queries from Europe to be routed to a fleet of EC2 instances in one of the EU regions
        - Servers in these locations could have all prices and language set to EU standards for example

Resource or Operation	Default Limit
Hosted zones:	500
Domains:	50
Resource record sets per hosted zone:	10,000
Reusable delegation sets:	100
Hosted zones that can use the same reusable delegation set:	100
Amazon VPCs that you can associate with a private hosted zone:	100
Health checks:	50
Traffic policies:	50
Policy records:	5

For additional information about Route 53 Limits, see Limits in Amazon Route 53

Compute:

> ### EC2 (Elastic Compute Cloud):

Elastic Compute Cloud - Backbone of AWS, provides re-sizable compute capacity in the cloud. Reduces the time required to obtain and boot new server instances to minutes allowing you to quickly scale capacity, both up and down, as your computing requirements change.

- Once an Instance has been launched with instance store storage, you can not attach additional instance store volumes after the instance is launched, only EBS volumes
- When using an instance store volume, you can not stop the instance (the option to do so will not be available, as the instance moves to another host and and would cause complete data loss)
- When using ephemeral storage, an underlying host failure will result in data loss
- You can reboot both instance types (w/ephemeral and EBS volumes) and will not lose data, but

again, an ephemeral volume based instance can NOT be stopped
- By default both Root volumes will be deleted on termination, however you can tell AWS to keep the root device volume on a new instance during launch
- You can poll an instances meta-data by using curl http://169.254.169.254/latest/meta-data/
- You can get an instance's IP address by using curl http://169.254.169.254/latest/meta-data/public-ipv4
- No such thing as user-data, remember its always meta-data not user-data
- Can not encrypt root volumes, but you can encrypt any additional volumes that are added and attached to an EC2 instance.
- You can have up to 10 tags per EC2 instance
- AWS does not recommend ever putting RAID 5's on EBS
- When configuring a launch configuration for an auto-scaling group, the Health Check Grace Period is the period of time to ignore health checks while instances or auto-scaled instances are added and booting.
- Termination protection is turned off by default, you must turn it on
- Roles:
    - You can only assign an EC2 role to an instance on create. You can not assign a role after the instance has been created and/or is running
    - You can change the permissions on a role post creation, but can NOT assign a new role to an existing instance
    - Role permissions can be changed, but not swapped
    - Roles are more secure then storing your access key and secret key on individual EC2 instances
    - Roles are easier to manager, You can assign a role, and change permissions on that role at any time which take effect immediately
    - Roles can only be assigned when that EC2 instance is being provisioned
    - Roles are universal, you can use them in any region
- Instance sizing:
    - T2 - Lowest Cost General Purpose - Web/Small DBs
    - M4 - General Purpose - App Servers
    - M3 - General Purpose - App servers
    - C4 - Compute Optimized - CPU Intensive Apps/DBs
    - C3 - Compute Optimized - CPU Intensive Apps/DBs
    - R3 - Memory Optimized - Memory Intensive Apps/DBs
    - G2 - Graphics / General Purpose - Video Encoding/Machine Learning/3D App Streaming
    - I2 - High Speed Storage - NoSQL DBs, Data Warehousing
    - D2 - Dense Storage - Fileservers/Data Warehousing/Hadoop
    - D - Density
    - I - IOPS
    - R - RAM
    - T - Cheap General Purpose
    - M - Main General Purpose
    - C - Compute
    - G - Graphics
- Storage Types:
    - Instance Store (Ephemeral):
        - Also referred to as ephemeral storage and is not persistent
        - Instances using instance store storage can not be stopped. If they are, data loss would result
        - If there is an issue with the underlying host and your instance needs to be moved, or is lost, Data is also lost
        - Instance store volumes cannot be detached and reattached to other instances; They exist only for the life of that instance
        - Best used for scratch storage, storage that can be lost at any time with no bad ramifications, such as a cache store
    - EBS (Elastic Block Storage):
        - Elastic Block Storage is persistent storage that can be used to procure storage to EC2 instances.
        - You can NOT mount 1 EBS volume to multiple EC2 instances instead you must use EFS
        - Default action for EBS volumes is for the root EBS volume to be deleted when the instance is terminated
        - By default, ROOT volumes will be deleted on termination, however with EBS volumes

only, you can tell AWS to keep the root device volume
        - EBS backed instances can be stopped, you will NOT lose any data
        - EBS volumes can be detached and reattached to other EC2 instances 3 Types of available EBS volumes can be provisioned and attached to an EC2 instance:
            - General Purpose SSD (GP2):
                - General Purpose up to 10K IOPS
                - 99.999% availability
                - Ratio of 3 IOPS per GB with up to 10K IOPS and ability to burst
                - Up to 3K IOPS for short periods for volumes under 1GB
            - Provisioned IOPS SSD (I01)
                - Designed for I/O intensive applications such as large relational or No-SQL DBs.
                - Use if need more than 10K IOPS
            - Magnetic (Standard)
                - Lowest cost per GB
                - Ideal for workloads where data is accessed infrequently and apps where the lowest cost storage is important.
                - Ideal for fileservers
        - Encryption:
            - Root Volumes cannot be encrypted by default, you need a 3rd party utility
            - Other volumes added to an instance can be encrypted.
- AMIs:
    - AMI's are simply snapshots of a root volume and is stored in S3
    - AMI's are regional. You can only launch an AMI from the region in which it was stored
    - You can copy AMI's to other regions using the console, CLI or Amazon EC2 API
    - Provides information required to launch a VM in the cloud
    - Template for the root volume for the instance (OS, Apps, etc)
    - Permissions that control which AWS accounts can use the AMI to launch instances
    - When you create an AMI, by default its marked private. You have to manually change the permissions to make the image public or share images with individual accounts
    - Block device mapping that specifies volumes to attach to the instance when its launched
    - Hardware Virtual Machines (HVM) AMI's Available
    - Paravirtual (PV) AMI's Available
    - You can select an AMI based on:
        - Region
        - OS
        - Architecture (32 vs. 64 bit)
        - Launch Permissions
        - Storage for the root device (Instance Store Vs. EBS)
- Security Groups:
    - Act like virtual firewalls for the associated EC2 instance
    - If you edit a security group, it takes effect immediately.
    - You can not set any deny rules in security groups, you can only set allow rules
    - There is an implicit deny any any at the end of the security group rules
    - You don't need outbound rules for any inbound request. Rules are stateful meaning that any request allowed in, is automatically allowed out
    - You can have any number of EC2 instances associated with a security group
- Snapshots:
    - You can take a snapshot of a volume, this will store that volumes snapshot on S3
    - Snapshots are point in time copies of volumes
    - The first snapshot will be a full snapshot of the volume and can take a little time to create
    - Snapshots are incremental, which means that only the blocks that have changes since your last snapshot are moved to S3
    - Snapshots of encrypted volumes are encrypted automatically
    - Volumes restored from encrypted snapshots are encrypted automatically
    - You can share snapshots but only if they are not encrypted
    - Snapshots can be shared with other AWS accounts or made public in the market place again as long as they are NOT encrypted
    - If you are making a snapshot of a root volume, you should stop the instance before taking the snapshot
- RAID Volumes:
    - If you take a snapshot, the snapshot excludes data held in the cache by applications or OS. This tends to not be an issue on a single volume, however multiple volumes in a RAID array,

can cause a problem due to interdependencies of the array
    - Take an application consistent snapshot
        - Stop the application from writing to disk
        - Flush all caches to the disk
    - Snapshot of RAID array --> 3 Methods:
        - Freeze the file system
        - Unmount the RAID Array
        - Shutdown the EC2 instance --> Take Snapshot --> Turn it back on
- Placement Groups:
    - A logical group of instance in a single AZ
    - Using placement groups enables applications to participate in a low latency, 10Gbps network
    - Placement groups are recommended for applications that benefit from low network latency, high network throughput or both
    - A placement group can't span multiple AZ's so it is a SPoF.
    - Then name you specify for a placement group must be unique within your AWS account
    - Only certain types of instances can be launched in a placement group. Computer Optimized, GPU, Memory Optimized, and Storage Optimized.
    - AWS recommends that you use the same instance family and same instance size within the instance group.
    - You can't merge placement groups
    - You can't move an existing instance into a placement group
    - You can create an AMI from your existing instance and then launch a new instance from the AMI into a placement group
- Pricing Models:
    - On Demand:
        - Pay fixed rate by the hour with no commitment
        - Users that want the low cost and flexibility of EC2
        - Apps with short term, spiky or unpredictable workloads that cannot be interrupted
        - Apps being developed or tested on EC2 for the first time
    - Reserved:
        - Provide capacity reservation and offer significant discount on the hourly charge for an instance (1-3 year terms)
        - Applications have steady state, or predictable usage
        - Apps that require reserved capacity
        - Users able to make upfront payments to reduce their total computing costs even further.
    - Spot:
        - Bid whatever price you want for instance capacity by the hour
        - When your bid price is greater than or equal to the spot price, your instance will boot
        - When the spot price is greater than your bid price, your instance will terminate with an hours notice.
        - Applications have flexible start and end times
        - Apps that are only feasible at very low compute prices
        - Users with urgent computing needs for large amounts of additional capacity
        - If the spot instance is terminated by Amazon EC2, you will not be changed for a partial hour of usage
        - If you terminate the instance yourself you WILL be charged for any partial hours of usage.

Resource or Operation	Default Limit
Elastic IP addresses for EC2-Classic:	5
Security groups for EC2-Classic per instance:	500
Rules per security group for EC2-Classic:	100
Key pairs:	5000

Key pairs:	5000
On-Demand instances:	Varies based on instance type
Spot Instances:	Varies based on instance type
Reserved Instances:	20 instance reservations per Availability Zone, per month
Dedicated Hosts:	Up to 2 Dedicated Hosts per instance family, per region can be allocated
AMI Copies:	Destination regions are limited to 50 concurrent AMI copies at a time, with no more than 25 of those coming from a single source region.
Throttle on the emails that can be sent :	Throttle applied
Tags per EC2 instance:	10

ELB (Elastic Block Storage Limits)

Resource or Operation	Default Limit
Number of EBS volumes:	5000
Number of EBS snapshots:	10,000
Total volume storage of General Purpose SSD (gp2) volumes:	20 TiB
Total volume storage of Provisioned IOPS SSD (io1) volumes:	20 TiB
Total volume storage of Throughput Optimized HDD (st1):	20 TiB
Total volume storage of Cold HDD (sc1):	20 TiB
Total volume storage of Magnetic volumes:	20 TiB
Total provisioned IOPS:	40,000

For additional information about EC2 Limits, see Limits in Amazon EC2

ELB (Elastic Load Balancer)

Elastic Load Balancing offers two types of load balancers that both feature high availability, automatic scaling, and robust security. These include the Classic Load Balancer that routes traffic based on either application or network level information, and the Application Load Balancer that routes traffic based on advanced application level information that includes the content of the request.

- When configuring ELB health checks, bear in mind that you may want to create a file like healthcheck.html or point the ping path of the health check to the main index file in your application
- Remember the health check interval is how often a health check will occur
- Your Healthy/Unhealthy thresholds are how many times either will check before marking the origin either healthy or unhealthy
    - Health Check Interval: 10 seconds
    - Unhealthy Threshold: 2
    - Healthy Threshold: 3
    - This means that if the health check interval occurs twice without success, then the source will be marked as unhealthy. This is 2 checks @ 10 seconds per check, so basically after 20 seconds the origin will be marked unhealthy
    - Likewise, if the healthy threshold is marked at 3, then it would be 3 x health check interval or 10 seconds being 30 seconds. After 30 seconds with 3 consecutive success checks, the origin will be marked as healthy.

- Enable Cross-Zone Load Balancing will distribute load across all back-end instances, even if they exist in different AZ's
- ELBs are NEVER given public IP Addresses, only a public DNS name
- ELBs can be In Service or Out of Service depending on health check results
- Charged by the hour and on a per GB basis of usage
- Must be configured with at least one listener
- A listener must be configured with a protocol and a port for front end (client to ELB connection), as well as a protocol and port for backed end (ELB to instances connection)
- ELBs support HTTP, HTTPS, TCP, and SSL (Secure TCP)
- ELBs support all ports (1-65535)
- ELBs do not support multiple SSL certificates
- Classic ELBs support the following ports:
    - 25 (SMTP)
    - 80 (HTTP)
    - 443 (HTTPS)
    - 465 (SMTPS)
    - 587 (SMTPS)
    - 1024-65535
- HTTP Error Codes:
    - 200 - The request has succeeded
    - 3xx - Redirection
    - 4xx - Client Error (404 not found)
    - 5xx - Server Error

Application Load Balancer Limit	Default Limit
Load balancers per region:	20
Target groups per region:	50
Listeners per load balancer:	10
Targets per load balancer:	1000
Subnets per Availability Zone per load balancer:	1
Security groups per load balancer:	5
Rules per load balancer (excluding defaults:	10
No. of times a target can be registered per LB:	100
Load balancers per target group:	1
Targets per target group :	1000

Classic Load Balancer Limit	Default Limit
Load balancers per region:	20
Listeners per load balancer:	100
Subnets per Availability Zone per load balancer:	1
Security groups per load balancer:	5

Load Balancers per Region Limit NOTE:
This limit includes both your Application load balancers and your Classic load balancers. This limit can be increased upon request.

ECS (Elastic Container Service):

Amazon EC2 Container Service (ECS) is a highly scalable, high performance container management service that supports Docker containers and allows you to easily run applications on a managed cluster of Amazon EC2 instances.

- Not covered as exam topic currently

Resource or Operation	Default Limit
Number of clusters per region per account:	1000
Number of container instances per cluster:	1000
Number of services per cluster:	500

For additional information about Elastic Container Service Limits, see Limits in Amazon ECS

Elastic Beanstalk:

AWS Elastic Beanstalk is an easy-to-use service for deploying and scaling web applications and services developed with Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker on familiar servers such as Apache, Nginx, Passenger, and IIS.

- Not covered as exam topic currently

Resource or Operation	Default Limit
Applications:	1000
Application Versions:	1000
Environments:	500

Lambda:

Compute service that runs your code in response to events and automatically manages the underlying compute infrastructure resources for you.

- Serverless processing
- AWS lambda can automatically run code in response to modifications to objects in S3 buckets, messages arriving in Amazon Kinesis streams, table updates in DynamoDB, API call logs created by CloudTrail, and custom events from mobile applications, web applications, or other web services
- Lambda runs your code on high-availability compute infrastructure and performs all of the administration of the compute resources including server and operating system maintenance, capacity provisioning and automatic scaling, node and security patch deployment, and code monitoring and logging.. All you need to do is supply the code.
- Supports NodeJs, Python 2.x, Java
- 99.99% availability for both the service itself and the functions it operates.
- First 1 million requests are free
- 0.20 per 1 million requests thereafter

- Duration is calculated from the time your code begins executing until it returns or otherwise terminates, rounded up to the nearest 100 ms
- The price depends on the amount of memory you allocate to your function. You are charged 0.00001667 for every GB-second used
- Free Tier gives you 1 Million free requests per month, and 400K GB-Seconds of compute time per month
- The memory size you choose for your functions, determines how long they can run in the free tier
- The lambda free tier does not automatically expire at the end of your 12 month AWS free tier term, but is available to both existing and new AWS customers indefinitely
- Functions can be ran in response to HTTP requests using API Gateway or API calls made using AWS SDKs

Resource or Operation	Default Limit
Concurrent requests safety throttle per account:	100

For additional information about Lambda Limits, see Limits in Amazon Lambda

Storage:

> ## S3 (Simple Storage Service):

Amazon Simple Storage Service (Amazon S3), provides developers and IT teams with secure, durable, highly-scalable cloud storage. Amazon S3 is easy to use object storage, with a simple web service interface to store and retrieve any amount of data from anywhere on the web.

- Object based storage only for files, can not install OS or applications
- Data is spread across multiple devices and multiple facilities
- Can loose 2 facilities and still have access to your files
- Files can be between 1 byte and 5TB, and has no storage limit
- Files are stored flatly in buckets, Folders don't really exist, but are part of the file name
- S3 bucket names have a universal name-space, meaning each bucket name must be globally unique
- S3 Stores data in alphabetical order (lexigraphical order)
- S3 URL structures are region/amazon.aws.com/bucketname (https://s3-eu-west-1.amazonaws.com/myawesomebucket)
- Read after write consistency for PUTS of new objects (As soon as you write an object, it is immediately available)
- Eventual consistency for overwrite PUTS and DELETES. (Updating or deleting an object could take time to propagate)
- S3 is basically a key value store and consists of the following:
    - Key - Name of the object
    - Value - Data made up of bytes
    - Version ID (important for versioning)
    - Meta-data - Data about what you are storing
    - ACLs - Permissions for stored objects
- Amazon guarantees 99.99% availability for the S3 platform
- Amazon guarantees 99.999999999% durability for S3 information (11 x 9's)
- Tiered storage, and life-cycle management available
- Versioning is available but must be enabled. It is off by default
- Offers encryption, and allows you to secure the data using ACLs
- S3 charges for storage, requests, and data transfer
- Bucket names must be all lowercase, however in US-Standard if creating with the CLI tool, it will allow capital letters
- The transfers tab shows uploads, downloads, permission changes, storage class changes, etc..
- When you upload a file to S3, by default it is set private
- You can transfer files up to 5GB using PUT requests

- You can transfer files up to 5GB using PUT requests
- You can setup access control to control your buckets access by using bucket policies or ACLs
- Change the storage class under the Properties tab when an object is selected
- S3 buckets can be configured to create access logs which logs all requests to the S3 bucket
- S3 Events include SNS, or SQS events or Lambda functions. Lambda is location specific, not available in South Korea
- All storage tiers have SSL support, millisecond first byte latency, and support life-cycle management policies.
- Storage Tiers:
    - Standard S3:
        - Stored redundantly across multiple devices in multiple facilities
        - Designed to sustain the loss of 2 facilities concurrently
        - 11-9's durability, 99.99% availability
    - S3-IA (Infrequently Accessed):
        - For data that is accessed less frequently, but requires rapid access when needed
        - Lower fee than S3, but you are charged a retrieval fee
        - Also designed to sustain the loss of 2 facilities concurrently
        - 11-9's durability, 99.99% availability
    - Reduced Redundancy Storage (RSS):
        - Use for data such as thumbnails or data that could be regenerated
        - Costs less than Standard S3
        - Designed to provide 99.99% durability and 99.99% availability of objects over a year
        - Designed to sustain the loss of a single facility
    - Glacier:
        - Very cheap, Stores data for as little as $0.01 per gigabyte, per month
        - Optimized for data that is infrequently accessed. Used for archival only
        - It takes 3-5 hours to restore access to files from Glacier
- Versioning and Cross-Region Replication (CRR):
    - Versioning must be enabled in order to take advantage of Cross-Region Replication
    - Versioning resides under Cross Region Replication tab
    - Once Versioning is turned on, it can not be turned off, it can only be suspended
    - If you truly wanted versioning off, you would have to create a new bucket and move your objects
    - When versioning is enabled, you will see a slider tab at the top of the console that will enable you to hide/show all versions of files in the bucket
    - If a file is deleted for example, you need to slide this tab to show in order to see previous versions of the file
    - With versioning enabled, if you delete a file, S3 creates a delete marker for that file, which tells the console to not display the file any longer
    - In order to restore a deleted file you simply delete the delete marker file, and the file will then be displayed again in the bucket
    - To move back to a previous version of a file including a deleted file, simply delete the newest version of the file or the delete marker, and the previous version will be displayed
    - Versioning does store multiple copies of the same file. So in the example of taking a 1MB file, and uploading it. Currently your storage usage would be 1MB. Now if you update the file with small tweeks, so that content changes, but the size remains the same, and upload it. With the version tab on hide, you will see only the single updated file, however if you select show on the slider, you will see that both the original 1MB file exists as well as the updated 1MB file, so your total S3 usage is now 2MB not 1MB
    - Versioning does NOT support de-duplication or any similar technology currently
    - For Cross Region Replication (CRR), as long as versioning is enabled, clicking on the tab will now give you the ability to suspend versioning, and enable cross region replication
    - Cross Region Replication (CRR) has to be enabled on both the source and destination buckets in the selected regions
    - Destination bucket must be created and again globally unique (can be created right from the versioning tab, in the CRR configuration section via button)
    - You have the ability to select a separate storage class for any Cross Region Replication destination bucket
    - CRR does NOT replicate existing objects, only future objects meaning that only objects stored post turning the feature on will be replicated
    - Any object that already exists at the time of turning CRR on, will NOT be automatically replicated
    - Versioning integrates with life-cycle management and also supports MFA delete capability. This will use MFA to provide additional security against object deletion

This will use MFA to provide additional security against object deletion

- Life-cycle Management:
  - When clicking on Life-cycle, and adding a rule, a rule can be applied to either the entire bucket or a single 'folder' in a bucket
  - Rules can be set to move objects to either separate storage tiers or delete them all together
  - Can be applied to current version and previous versions
  - If multiple actions are selected for example transition from STD to IA storage 30 days after upload, and then Archive 60 days after upload is also selected, once an object is uploaded, 30 days later the object will be moved to IA storage. 30 days after that the object will be moved to glacier.
  - Calculates based on UPLOAD date not Action data
  - Transition from STD to IA storage class requires MINIMUM of 30 days. You can not select or set any data range less than 30 days
  - Archive to Glacier can be set at a minimum of 1 day If STD->IA is NOT set
  - If STD->IA IS set, then you will have to wait a minimum of 60 days to archive the object because the minimum for STD->IA is 30 days, and the transition to glacier then takes an additional 30 days
  - When you enable versioning, there will be 2 sections in life-cycle management tab. 1 for the current version of an object, and another for previous versions
  - Minimum file size for IA storage is 128K for an object
  - Can set policy to permanently delete an object after a given time frame
  - If versioning is enabled, then the object must be set to expire, before it can be permanently deleted
  - Can not move objects to Reduced Redundancy using life-cycle policies
- S3 Transfer Acceleration:
  - Utilizes the CloudFront Edge Network to accelerate your uploads to S3
  - Instead of uploading directly to your S3 bucket, you can use a distinct URL to upload directly to an edge location which will then transfer the file to S3
  - This topic is covered in AWS Solutions Architect Study Guide
  - There is a test utility available that will test uploading direct to S3 vs through Transfer Acceleration, which will show the upload speed from different global locations
  - Turning on and using Transfer Acceleration will incur an additional fee
- 2 types of encryption available:
  - In transit:
    - Uses SSL/TLS to encrypt the transfer of the object
  - At Rest (AES 256):
    - Server Side: S3 Manged Keys (SSE-S3)
    - Server Side: AWS Key Management Service, Managed Keys (SSE-KMS)
    - Server Side: Encryption with Customer provided Keys (SSE-C)
    - Client Side Encryption
- Pricing (What your charged for when using S3):
  - Storage used
  - Number of Requests
  - Data Transfer

Resource or Operation	Default Limit
Buckets per account:	100
Largest files size you can transfer with PUT request:	5GB
Minimum file size:	1 byte
Maximum file size:	5 TB

For additional information about S3 Limits, see Limits in Amazon S3

> ## CloudFront:

Amazon CloudFront is a global content delivery network (CDN) service that accelerates delivery of your websites, APIs, video content or other web assets.

- Edge Location is the location where content will be cached, separate from an AWS Region/AZ
- Origin is the origin of all files, can be S3, EC2 instance, a ELB, or Route53
- Distribution is the name given to the CDN which consists a collection of edge locations
- Web Distributions are used for websites
- RTMP - (Real-Time Messaging Protocol) used for streaming media typically around adobe flash files
- Edge locations can be R/W and will accept a PUT request on an edge location, which then will replicate the file back to the origin
- Objects are cached for the life of the TTL (24 hours by default)
- You can clear objects from edge locations, but you will be charged
- When enabling cloudfront from an S3 origin, you have the option to restrict bucket access; this will disable the direct link to the file in the S3 bucket, and ensure that the content is only served from cloudfront
- The path pattern uses regular expressions
- You can restrict access to your distributions using signed URLS
- You can assign Web Application Firewall rules to your distributions
- Distribution URLs are going to be non-pretty names such as random_characters.cloudfront.com; you can create a CNAME that points to the cloudfront name to make the URL user friendly
- You can restrict content based on geographical locations in the behaviors tab
- You can create custom error pages via the error pages tab
- Purging content is handled in the Invalidations tab

Resource or Operation	Default Limit
Data transfer rate per distribution:	40 Gbps
Requests per second per distribution:	100,000
Web distributions per account:	200
RTMP distributions per account:	100
Alternate domain names (CNAMEs) per distribution:	100
Origins per distribution:	25
Cache behaviors per distribution:	25
White-listed headers per cache behavior:	10
White-listed cookies per cache behavior:	10
SSL certificates per account when serving HTTPS requests using dedicated IP addresses (no limit when serving HTTPS requests using SNI):	2
Custom headers that you can have Amazon CloudFront forward to the origin:	10 name–value pairs

For additional information about CloudFront Limits, see Limits in Amazon CloudFront

EFS (Elastic File System):

File storage service for EC2 instances. Its easy to use and provides a simple interface that allows you to create and configure file systems quickly and easily. With EFS storage capacity is elastic, growing and shrinking automatically as you add and remove files so your applications have the storage they need, when they need it.

- Think NFS, only without a set storage limit
- Supports NFSv4, and you only pay for the storage you use

- Supports NFSv4, and you only pay for the storage you use
- Billing rate is 30 cents per GB
- Can scale to exabytes
- Can support thousands of concurrent NFS connections
- Data is stored across multiple AZ within a region
- Block based storage.
- Can be shared with multiple instances
- Read after Write Consistency
- You must ensure that instances that will mount EFS are in the same security group as the EFS allocation. If they are not, you can modify the security groups, and add them to the same security group that was used to launch the EFS storage

Resource or Operation	Default Limit
Total throughput per file system:	3 GB/s for all connected clients

For additional information about EFS Limits, see Limits in Amazon EFS

Snowball:

Snowball is a petabyte-scale data transport solution that uses secure appliances to transfer large amounts of data into and out of the AWS cloud.

- Appliance allows local import using AWS internal network

Resource or Operation	Default Limit	Comments
Snowball:	1	If you need to increase this limit, contact AWS Support.

Storage Gateway:

The AWS Storage Gateway is a service connecting an on-premises software appliance with cloud-based storage to provide seamless and secure integration between an organization's on-premises IT environment and AWS's storage infrastructure. On-premise virtual appliance that can be downloaded and used to cache S3 locally at a customers site

- Replicates data to and from AWS platform
- Gateway Cached Volumes:
  - Entire dataset is stored on S3 and the most frequently accessed data is cached on-site
  - These volumes minimize the need to scale your on-prem storage infrastructure while providing your applications with low-latency access to their frequently accessed data
  - Can create storage volumes up to 32TBs in size and mount them as iSCSI devices from your on-premises application servers.
  - Data written to these volumes is stored in S3, with only a cache of recently written and recently read data stored locally on your on-premises storage hardware.
- Gateway Stored Volumes:
  - Store your primary data locally while asynchronously backing up that data to AWS
  - Provide low-latency access to their entire datasets, while providing durable, off-site backups.
  - Can create storage volumes up to 1TB in size and mount them as iSCSI devices from your on-premises application servers.
  - Data written to your gateway stored volumes is stored on your on-prem storage hardware, and asynchronously backed up to S3 in the form of EBS snapshots.
- Gateway Virtual Tape Library (VTL):
  - Used for backup and uses popular backup applications like NetBackup, Backup Exec and

Veam

- Pricing:
  - You pay for what you use
  - Has 4 pricing components:
    - Gateway usage (per gateway per month)
    - Snapshot storage usage (per GB per month)
    - Volume storage usage (Per GB per month)
    - Data transfer out (Per GB per month)
- Import Export:
  - Import Export Disk - Import to EBS, S3, Glacier but only export to S3
  - Pay for what you use
  - Has 3 pricing components:
    - Per device fee
    - Data load time charge per data-loading-hour
    - Possible return shipping charges for expedited shipping or shipping to destinations not local to the Import/Export region

Databases:

RDS (Relational Database Service):

Amazon Relational Database Service (Amazon RDS) makes it easy to set up, operate, and scale a relational database in the cloud. Amazon RDS provides you six familiar database engines to choose from, including Amazon Aurora, Oracle, Microsoft SQL Server, PostgreSQL, MySQL and MariaDB.

- Traditional relational databases that include tables, rows, fields
- On-Line Transaction Processing (OLTP) type DB
- You can copy a snapshot to another region if you want to have your database available in another region
- You scale your DB by taking a snapshot and doing a restore to a larger sized tier
- RDS maximum size for a MS SQL Server DB with SQL Server Express Edition is 10GB per DB
- Supported RDS Platforms:
  - MS SQL Server
  - Oracle
  - MySQL Server
  - PostgreSQL
  - Aurora
  - MariaDB
- When a backup is restored, the restore will always be a new RDS instance, with a new DNS name
- Backup types:
  - Automated backups
    - Allows you to recover your database to any point in time within a retention period
    - Retention periods can be between 1 and 35 days
    - Takes a full daily snapshot and will also store transaction logs through the day
    - When you do a recovery, AWS will choose the most recent daily backup and then apply transaction logs
    - Allows you to do a point in time recover down to a second within the retention period
    - Enabled by default
    - Backup data is stored in S3
    - You get free storage space equal to the size of your database.
    - Taken within a defined window
    - During the backup, storage I/0 may be suspended and you may experience extended latency
  - Database snapshots
    - User initiated from the console
    - Stored even after you delete the original RDS instance unlike automatic backups
- Encryption:
- Encryption at rest is supported for MySQL, Oracle, SQL Server, PostgreSQL, and MariaDB

- Encryption is done using the AWS Key Management Service (KMS)
- Once your RDS instance is encrypted the data stored at rest in the underlaying storage is encrypted, as are its automated backups, read replicas and snapshots
- To use RDS encryption, create a new DB instance with encryption enabled and migrate your data to it
- Encrypting an existing DB instance is not supported
- Multi-AZ:
    - Allows you to have an exact copy of your production database in another AZ
    - AWS handles the replication for you, so when your prod database is written to, the write will automatically be synchronized to the stand-by DB
    - In the event of DB maintenance, instance failure or AZ failure, RDS will automatically fail-over to the standby so that database operations can resume quickly without Admin intervention.
    - In a fail-over scenario, the same DNS name is used to connect to the secondary instance, There is no need to reconfigure your application
    - Multi AZ configurations are used for HA/DR only, and is not used for improving performance
    - To scale for performance you need to set up read replicas
    - Available for SQL Server, Oracle, MySQL, PostGreSQL, and Aurora
- Read Replica's:
    - Uses asynchronous replication, from the primary instance to other instances that can be read from
    - You can have up to 5 read replicas of your main database
    - Allow you to have a read only copy of your prod database
    - Used primarily for very read-heavy database workloads
    - SQL Server and Oracle are not supported
    - Used for scaling not DR
    - Must have automatic backups setup
    - You can have read replicas of read replicas (but could incur latency as its daisy chained)
    - Each read replica will have its own DNS endpoint
    - You cannot have read replicas that have Multi-AZ
    - You can create read replicas of Multi-AZ source databases however
    - Read Replicas can be promoted to be their own databases, however this breaks replication
    - Read Replicas in a second region for MySQL and MariaDB, not for PostgreSQL
    - Read Replicas can be bigger than the primary source DB from a resource perspective
- Aurora:
    - MySQL compatible relational database engine that combines speed and availability of high end commercial databases with the simplicity and cost-effectiveness of open source databases
    - Provides up to 5 times better performance than MySQL at a price point 1/10th of a commercial database while delivering similar performance and availability
    - Starts with 10GB, scales in 10GB increments up to 64TB (Storage Auto scaling)
    - Compute resources can scale up to 32 vCPUs and 244 GB of memory
    - Maintains 2 copies of your data contained in each availability zone, with minimum of 3 AZs. 6 copies of your data
    - Designed to transparently handle the loss of up to two copies of data without affecting the DB write availability and up to 3 copies without affecting read availability
    - Designed to handle loss of up to 2 copies without affecting DB write availability
    - Designed to handle loss of up to 3 copies without affecting DB read availability
    - Self healing storage, data blocks and disks are continuously scanned for errors and repaired automatically
    - 2 Types of replicas available:
        - Aurora Replicas - Separate aurora DB, can have up to 15 replicas
        - MySQL read replicas, can have up to 5
    - If a failure occurs of the primary database, a fail-over will happen automatically to an aurora replica, but will NOT auto fail over to a MySQL read replica.
    - Only available in certain regions, not all

Resource or Operation	Default Limit
Clusters:	40

Cluster parameter groups:	50
DB Instances:	40
Event subscriptions:	20
Manual snapshots:	50
Manual cluster snapshots:	50
Option groups:	20
Parameter groups:	50
Read replicas per master:	5
Aurora only read replicas per master:	15
Reserved instances (purchased per month):	40
Rules per security group:	20
Security groups:	25
Security groups (VPC):	5
Subnet groups:	20
Subnets per subnet group:	20
Tags per resource:	50
Total storage for all DB instances:	100 TB

DynamoDB (No-SQL):

Fast and flexible NoSQL DB service for all apps that need consistent, single-digit millisecond latency at any scale. It is a fully managed database and supports both document and key-value data models. Its flexible data model and reliable performance make it a great fit for mobile, web, gaming, ad-tech, IoT, and many other applications.

- Non Relational DB (No-SQL), comprised of collections (tables), of documents (rows), with each document consisting of key/value pairs (fields)
- Document oriented DB
- Offers push button scaling, meaning that you can scale your db on the fly without any downtime
- RDS is not so easy, you usually have to use a bigger instance size or add read replicas
- Stored on SSD Storage
- Spread across 3 geographically distinct data centers
- Eventual Consistent Reads (Default)
    - Consistency across all copies of data is usually reached within 1 second
    - Repeating a read after a short time should return updated data
    - Best Read Performance
- Strongly Consistent Reads
    - Returns a result that reflects all writes that received a successful response prior to the read
- Structure:
    - Tables
    - Items (Think rows in a traditional table)
    - Attributes (Think columns of data in a table)
- Provisioned throughput capacity
- Write throughput 0.0065 per hour for every 10 units
- Read throughput 0.0065 per hour for every 50 units
- First 25 GB of storage is free
- Storage costs of 25 cents per additional GB per Month
- Can be expensive for writes, but really really cheap for reads

- Can be expensive for writes, but really really cheap for reads
- The combined key/value size must not exceed 400 KB for any given document

US East (N. Virginia) Region	Default Limit
Maximum capacity units per table or global secondary index:	40,000 read capacity units and 40,000 write capacity units
Maximum capacity units per account:	80,000 read capacity units and 80,000 write capacity units

All Region Resource or Operation	Default Limit
Maximum capacity units per table or global secondary index:	10,000 read capacity units and 10,000 write capacity units
Maximum capacity units per account:	20,000 read capacity units and 20,000 write capacity units
Maximum number of tables:	256

For additional information about DynamoDB Limits, see Limits in Amazon DynamoDB

Elasticache:

Amazon ElastiCache is a web service that makes it easy to deploy, operate, and scale an in-memory data store or cache in the cloud.

- Can be used for DB caching in conjunction with services like RDS
- Web service that makes it easy to deploy, operate, and scale in memory cache in the cloud
- Improves the performance of web applications by allowing you to retrieve information from fast, managed in-memory caches, instead of relying entirely on slower disk based databases
- Improves application performance by storing critical pieces of data in memory for low-latency access
- Cached information may include the results of I/O intensive database queries or the results of computationally intensive calculations
- Supports 2 open-source in-memory caching engines:
    - Memcached:
        - Widely adopted memory object caching system
        - Elasticache is protocol complaint with memcached, so popular tools that you use today with existing memcached environments will work seamlessly with the service
        - No Multi AZ support
    - Redis:
        - Popular open-source in-memory key-value store that supports data structures such as sorted sets and lists
        - Elasticache supports Master/Slave replication and Multi-AZ which can be used to achieve cross AZ redundancy
        - Good choice if your db is read heavy and not prone to frequent changing

All Region Resource or Operation	Default Limit	Description
Nodes per region:	50	The maximum number of nodes across all clusters in a region.
Nodes per cluster (Memcached):	20	The maximum number of nodes in an individual Memcached cluster.

Nodes per cluster (Redis):	1	The maximum number of nodes in an individual Redis cluster.
Clusters per replication group (Redis):	6	The maximum number of clusters in a Redis replication group. One is the read/write primary. All others are read only replicas.
Parameter groups per region:	20	The maximum number of parameters groups you can create in a region.
Security groups per region:	50	The maximum number of security groups you can create in a region.
Subnet groups per region:	50	The maximum number of subnet groups you can create in a region.
Subnets per subnet group:	20	The maximum number of subnets you can define for a subnet group.

Redshift:

Fast and powerful, fully managed, petabyte-scale data warehouse service in the cloud. Customers can start small for just 25 cents per hour with no commitments or upfront costs and scale to a petabyte or more for 1000 per TB per year. Less than a tenth of most other data warehousing solutions.

- Used for data warehousing / business intelligence
- Uses 1024KB/1MB block size for its columnar storage
- Tools like Cognos, Jaspersoft, SQL Server Reporting Services, Oracle Hyperion, SAP NetWeaver
- Used to pull in very large and complex data sets
- Used by management to do queries on data such as current current performance vs target
- 10 times faster than traditional RDS
- Massively Parallel Processing (MPP)
- Automatically distributes data and query load across all nodes
- Currently only available in 1 AZ at a time
- Can restore snapshots to new AZ's in the event of an outage
- 2 types of transactions:
    - On-line Transaction Processing (OLTP) - Standard transaction driven database insert/retrieval -Pulls up a row of data such as Name, Date etc..
    - On-line Analytics Processing (OLAP) - Pulls up a row of data such as Name, Date etc..
        - Uses different type of architecture both from a DB and infrastructure layer
        - Pull in data from multiple queries, gathering tons of information depending on what type of report is required
- Start with Single Node (160GB)
- Multi-node configurations available:
    - Leader Node - Manages client connections and receives queries
    - Compute Node - Store data and perform queries and computations
    - Can have up to 128 compute nodes
- Columnar data storage:
    - Instead of storing data as a series of rows, redshift organizes data by column.
    - Unlike row-based systems, which are ideal for transaction processing, Column-based systems are ideal for data warehousing and analytics where queries often involve aggregates performed over large data sets.
    - Only columns involved in the queries are processed and columnar data is stored sequentially on the storage media
    - Column-based systems require far fewer I/Os, greatly improving query performance
- Advanced compression:
    - Columnar data stores can be compressed much more than row-based data stores because similar data is stored sequentially on disk
    - Redshift employs multiple compression techniques and can often achieve significant compression relative to traditional relational data stores

    - Does not require indexes or materialized views so uses less space than traditional relational db systems
    - Automatically samples your data and selects the most appropriate compression scheme
- Priced on 3 things
    - Total number of hours you run across your compute nodes for the billing period
    - You are billed for 1 unit per node per hour, so 3-node cluster running an entire month would incur 2,160 instance hours
    - You will not be charged for leader node hours, only compute nodes will incur charges
    - Charged on backups
    - Charged for data transfers (only within VPC not outside)
- Security:
    - Encrypted in transit using SSL
    - Encrypted at rest using AES-256 encryption
    - Takes care of key management by default
    - Manage your own keys through Hardware Security Module (HSM)
    - AWS Key Management Service

Resource or Operation	Default Limit
Nodes per cluster:	101
Nodes per cluster:	200
Reserved Nodes:	200
Snapshots:	20
Parameter Groups:	20
Security Groups:	20
Subnet Groups:	20
Subnets per Subnet Group:	20
Event Subscriptions:	20

For additional information about Redshift Limits, see Limits in Amazon Redshift

DMS (Database Migration Service):

AWS Database Migration Service helps you migrate databases to AWS easily and securely. The source database remains fully operational during the migration, minimizing downtime to applications that rely on the database. The AWS Database Migration Service can migrate your data to and from most widely used commercial and open-source databases. The service supports homogenous migrations such as Oracle to Oracle, as well as heterogeneous migrations between different database platforms, such as Oracle to Amazon Aurora or Microsoft SQL Server to MySQL.

- Allows migration of your production DB platforms to AWS or between services like MySQL -> PostgreSQL
- Once started, AWS manages all the complexities of the migration process like data type transformation, compression, and parallel transfer for faster transfer, while ensuring that data changes to the source database that occur during the migration process are automatically replicated to the target
- AWS schema conversion tool automatically converts the source DB schema and a majority of the custom code, including views, stored procedures and functions to a format compatible with the target DB

Analytics:

EMR (Elastic Map Reduce):

Amazon EMR is a web service that makes it easy to quickly and cost-effectively process vast amounts of data.

- Not covered as exam topic currently

Resource or Operation	Default Limit
Replication instances:	20
Total amount of storage:	6 TB
Replication subnet groups:	20
Subnets per replication subnet group:	20
Endpoints:	20
Tasks:	200
Endpoints per instance:	20

Data Pipeline:

AWS Data Pipeline is a web service that helps you reliably process and move data between different AWS compute and storage services, as well as on-premise data sources, at specified intervals. With AWS Data Pipeline, you can regularly access your data where it's stored, transform and process it at scale, and efficiently transfer the results to AWS services such as Amazon S3, Amazon RDS, Amazon DynamoDB, and Amazon Elastic MapReduce (EMR).

- Not covered as exam topic currently

Resource or Operation	Default Limit	Adjustable
Number of pipelines:	100	Yes
Number of objects per pipeline:	100	Yes
Number of active instances per object:	5	Yes
Number of fields per object:	50	No
Number of UTF8 bytes per field name or identifier:	256	No
Number of UTF8 bytes per field:	10,240	No
Number of UTF8 bytes per object:	15,360 (including field names)	No
Rate of creation of a instance from an object:	1 per 5 minutes	No
Retries of a pipeline activity:	5 per task	No
Minimum delay between retry attempts:	2 minutes	No
Minimum scheduling interval:	15 minutes	No

Maximum number of roll-ups into a single object:	32	No
Maximum number of EC2 instances per Ec2Resource object:	1	No

For additional information about Data Pipelines Service Limits, see Limits in Amazon DataPiplelines

Elastic Search:

Amazon Elasticsearch Service is a managed service that makes it easy to deploy, operate, and scale Elasticsearch in the AWS Cloud.

- Not covered as exam topic currently

Resource or Operation	Default Limit
Number of Amazon ES instances per cluster:	20

Kinesis:

Kinesis is a fully managed service for real time processing of streaming data at massive scale.

- Streaming Data platform
- If any questions reference streaming, think Kinesis
- Used to consume big data
- Stream large amounts of social media, news feeds, logs, etc in the cloud
- Process large amounts of data
- Elastic Map Reduce is for big data processing
- Business intelligence and reporting, would be derived from redshift

Resource or Operation	Default Limit	Notes
Delivery streams per region:	20	
Delivery stream capacity:	2,000 transactions/second 5,000 records/second 5 MB/second	The three capacity limits scale proportionally. For example, if you increase the throughput limit to 10MB/second, the other limits increase to 4,000 transactions/sec and 10,000 records/sec.
Shards per region:	US EAST, US WEST, EU: 50 All other supported regions: 25	

For additional information about Kinesis Limits, see Firehose limits in Amazon Kinesis and Stream limits in Amazon Kinesis

Machine Learning:

Amazon Machine Learning is a service that makes it easy for developers of all skill levels to use machine learning technology. Amazon Machine Learning provides visualization tools and wizards that guide you through the process of creating machine learning (ML) models without having to learn complex ML algorithms and technology.

- Not covered as exam topic currently

Resource or Operation	Default Limit
Data file size:	100 GB
Batch prediction input size:	1 TB
Batch prediction input (number of records):	100 Million
Number of variables in a data file (schema):	1000
Recipe complexity (number of processed output variables):	10,000
Transactions Per Second for each real-time prediction endpoint:	200
Total Transactions Per Second for all real-time prediction endpoints:	10,000
Total RAM for all real-time prediction endpoints:	10 GB
Number of simultaneous jobs:	5
Longest run time for any job:	7 days
Number of classes for multiclass ML models:	100
ML model size:	2 GB

Data File Size Note:
The size of your data files is limited to ensure that jobs finish in a timely manner. Jobs that have been running for more than seven days will be automatically terminated, resulting in a FAILED status.

For additional information about Machine Learning Limits, see Limits in Amazon ML

Quick Sight:

Amazon QuickSight is a very fast, cloud-powered business intelligence (BI) service that makes it easy for all employees to build visualizations, perform ad-hoc analysis, and quickly get business insights from their data.

- Not covered as exam topic currently

Security and Identity:

IAM (Identity and Access Management):

AWS Identity and Access Management (IAM) enables you to securely control access to AWS services and resources for your users.

- This topic is covered in AWS Solutions Architect Study Guide
- Allows for centralized control and shared access to your AWS Account and/or AWS services
- By default when you create a user, they have NO permissions to do anything
- Root account has full admin access upon account creation
- Not region specific, can be shared between all regions
- Granular permission sets for AWS resources
- Includes Federation Integration which taps into Active Directory, Facebook, Linkedin, etc. for authentication
- Multi-factor authentication support
- Allows configuration of temporary access for users, devices and services
- Set up and manage password policy and password rotation policy for IAM users
- Integration with many different AWS services
- Supports PCI DSS compliance
- Access can be applied to:
    - Users - End users (people)
    - Groups - Collection of users under one set of permissions
    - Roles - Assigned to AWS resources, specifying what the resource (such as EC2) is allowed to access on another resource (S3)
    - Policies - Document that defines one or more permissions
- Policies can be applied to users, groups and roles
- You can assign up to 10 policies to a single group
- Policy documents must have a version, and a statement in the body; The statement must consist of Effects (Allow, Deny), Actions(Which action to allow/deny such a * for all actions), and Resources (affected resources such as * for all resources)
- All resources can share the same policy document
- There are 3 different types of roles:
    - Service Roles
    - Cross account access roles
        - Used when you have multiple AWS accounts and another AWS account must interact with the current AWS account
    - Identity provider access roles
        - Roles for facebook or similar Identity providers
- In order for a new IAM user to be able to log into the console, the user must have a password set
- By default a new users access is only accomplished through the use of the access key/secret access key
- If the users password is a generated password, it also will only be shown at the time of creation.
- Customizable Console Sign-in link can be configured on the main IAM page (aws.yourdomain.com)
- Customizable Console Sign-in links must be globally unique. If a sign in link name is already taken, you must choose an alternative
- Root account is email address that you used to register your account
- Recommended that root account is not used for login, and should be secured with Multi-factor Authentication (MFA)
- Can create Access Keys/ Secret Access Keys to allow IAM users (or service accounts) to be used with AWS CLI or API calls
- Access Key ID is equivalent to a user-name, Secret Access Key is equivalent to a password
- When creating a user's credentials, you can only see/download the credentials at the time of creation not after.
- Access Keys can be retired, and new ones can be created in the event that secret access keys are lost
- To create a user password, once the users have been created, choose the user you want to set the password for and from the User Actions drop list, click manage password. Here you can opt to create a generated or custom password. If generated, there is an option to force the user to set a custom password on next login. Once a generated password has been issued, you can see the password which is the same as the access keys. Its shown once only
- Click on Policies from the left side menu and choose the policies that you want to apply to your users. When you pick a policy that you want applied to a user, select the policy, and then from the top Policy Actions drop menu, choose attach and select the user that you want to assign the policy to

Resource or Operation	Default Limit
Groups per account:	100
Instance profiles:	100
Roles:	250
Server Certificates:	20
Users:	5000
Number of policies allowed to attach to a single group:	10

For additional information about IAM Limits, see Limits in IAM entities and objects

Directory Service:

AWS Directory Service makes it easy to setup and run Microsoft Active Directory (AD) in the AWS cloud, or connect your AWS resources with an existing on-premises Microsoft Active Directory.

- Not covered as exam topic currently

Resource or Operation	Default Limit
Simple AD directories:	10
AD Connector directories:	10
Manual snapshots:	5 per Simple AD

Inspector:

Amazon Inspector is an automated agent based security assessment service that helps improve the security and compliance of applications deployed on AWS.

- Allows customers to install agents on EC2 instances and inspect the instance for security vulnerabilities
- Not covered as exam topic currently

Resource or Operation	Default Limit
Running agents:	500
Assessment runs:	50,000
Assessment templates:	500
Assessment targets:	50

For additional information about Inspector Limits, see Limits in Amazon Inspector

WAF (Web Application Firewall):

AWS WAF is a web application firewall that helps protect your web applications from common web exploits that could affect application availability, compromise security, or consume excessive resources.

- Allows customers to secure their cloud infrastructure
- Not covered as exam topic currently

Resource or Operation	Default Limit
Web ACLs per account:	10
Rules per account:	50
Conditions per account:	50

For additional information about Web Application Firewall Service Limits, see Limits in Amazon WAF

Cloud HSM (Hardware Security Module):

The AWS CloudHSM service helps you meet corporate, contractual and regulatory compliance requirements for data security by using dedicated Hardware Security Module (HSM) appliances within the AWS cloud.

- Allows customers to secure their cloud infrastructure
- Not covered as exam topic currently

Resource or Operation	Default Limit
HSM appliances:	3
High-availability partition groups:	20
Clients:	800

KMS (Key Management Service):

AWS Key Management Service (KMS) is a managed service that makes it easy for you to create and control the encryption keys used to encrypt your data, and uses Hardware Security Modules (HSMs) to protect the security of your keys.

- Not covered as exam topic currently

Resource or Operation	Default Limit
Customer Master Keys (CMKs):	1000
Aliases:	1100
Grants per CMK:	2500
Grants for a given principal per CMK:	30
Requests per second:	Varies by API operation

KMS Note:
All limits in the preceding table apply per region and per AWS account.

For additional information about Key Management Service Limits, see Limits in Amazon KMS

Management Tools:

CloudWatch:

Amazon CloudWatch is a monitoring service for AWS cloud resources and the applications you run on AWS.

- By default all EC2 instances will have basic monitoring, which is a 5 minute poll
- If you want detailed CloudWatch monitoring, you get more graphs at a 1 minute poll interval
- Standard monitoring is on by default (5 min intervals)
- Detailed monitoring is on a 1 minute interval
- Detailed monitoring does cost 3.50 per instance per month
- CPU/Disk/Network In/Status metrics are available
- RAM is a host level metric and not available on a per instance basis
- Events can trigger Lambda functions or SNS events based on criteria, which helps you to respond to state changes within your AWS resources.
- Logs help you to aggregate, monitor, and store log data
- Logs can go down to the application level but requires an agent to be installed.
- Alarms can be set against any metrics that are available, and will perform an alert/notification and an action when the alarm criteria is met
- CloudWatch is used for performance monitoring, not auditing, that is what CloudTrail is for
- You can create dashboards with custom widgets to keep track of what is happening in your environment

CloudWatch Resource Limit	Default Limit	Comments
DescribeAlarms:	3 transactions per second (TPS)	The max number of operation requests you can make per second without being throttled.
GetMetricStatistics:	400 transactions per second (TPS)	The max number of operation requests you can make per second without being throttled.
ListMetrics:	25 transactions per second (TPS)	The max number of operation requests you can make per second without being throttled.
PutMetricAlarm:	3 transactions per second (TPS)	The max number of operation requests you can make per second without being throttled.
PutMetricData:	150 transactions per second (TPS)	The max number of operation requests you can make per second without being throttled.

CloudWatch Event Resource Limit	Default Limit
Rules	50 per account

CloudWatch Logs Resource Limit	Default Limit	Comments
CreateLogGroup:	500 log groups/account/region	If you exceed your log group limit, you get a ResourceLimitExceeded exception.
DescribeLogStreams:	5 transactions per second (TPS)/account/region	If you experience frequent throttling, you can request a limit increase.
FilterLogEvents:	5 transactions per second (TPS)/account/region	This limit can be changed only in special circumstances.
GetLogEvents:	5 transactions per second (TPS)/account/region	We recommend subscriptions if you are continuously processing new data. If you need historical data, we recommend exporting your data to Amazon S3. This limit can be changed only in special circumstances.

Cloud Trail:

AWS CloudTrail is a web service that records AWS API calls for your account and delivers log files to you.

- Provides way for customers to audit access to what people are doing on the platform in your account
- Not covered as exam topic currently

OpsWorks:

AWS OpsWorks is a configuration management service that helps you configure and operate applications of all shapes and sizes using Chef.

- Configuration management service which uses Chef in the background
- Consists of recipes to maintain a consistent state
- Look for term chef, recipes or cook books think OpsWorks

Resource Limit	Default Limit
Stacks:	40
Layers per stack:	40
Instances per stack:	40
Apps per stack:	40

Config:

AWS Config is a fully managed service that provides you with an AWS resource inventory, configuration history, and configuration change notifications to enable security and governance.

- Provides customer with configuration history, change notifications, and inventory
- Can perform tasks such as ensuring that all EBS volumes are encrypted etc..

Service Catalog:

AWS Service Catalog allows organizations to create and manage catalogs of IT services that are approved for use on AWS.

- Create and manage catalogs of services you are allowed to use on AWS
- Not covered as exam topic currently

Resource or Operation	Default Limit
Portfolios:	25 per account
Users, groups, and roles:	25 per portfolio
Products:	25 per portfolio, 25 total per account
Product versions:	50 per product
Constraints:	25 per product per portfolio
Tags:	3 per product, 3 per portfolio, 10 per stack
Stacks:	200 (AWS CloudFormation limit)

Trusted Advisor:

An on-line resource to help you reduce cost, increase performance, and improve security by optimizing your AWS environment, Trusted Advisor provides real time guidance to help you provision your resources following AWS best practices.

- Automated service that scans customer environment and offers advise on how how to save money, lock down resources, and reports security vulnerabilities

Application Services:

API Gateway:

Amazon API Gateway is a fully managed service that makes it easy for developers to create, publish, maintain, monitor, and secure APIs at any scale.

- Create, maintain, publish, monitor, and secure API endpoints at any scale
- Can be used as an HTTP interface for Lambda functions

Resource or Operation	Default Limit
APIs per account:	60
API keys per account:	500

Client certificates per account:	60
Throttle Rate:	1K requests per second (rps) with a burst limit of 2K rps
Usage plans per account:	300
Custom authorizers per API:	10
Resources per API:	300
Stages per API:	10

For additional information about API Gateway Limits, see Limits in Amazon API Gateway

AppStream:

Amazon AppStream enables you to stream your existing Windows applications from the cloud, reaching more users on more devices, without code modifications.

- AWS version of XenApp
- Steam Windows apps from the cloud
- Not covered as exam topic currently

Resource or Operation	Default Limit
Concurrent streaming sessions per account:	5
Concurrent streaming application deployments using the interactive wizard:	2
streaming applications in the Building, Active, or Error states:	3

For additional information about AppStream Service Limits, see Limits in Amazon AppStream

CloudSearch:

Amazon CloudSearch is a managed service in the AWS Cloud that makes it simple and cost-effective to set up, manage, and scale a search solution for your website or application.

- Makes it simple to manage and scale search across your entire application
- Not covered as exam topic currently

Resource or Operation	Default Limit
Partitions:	10
Search instances:	50

For additional information about Cloud Search Service Limits, see Limits in Amazon CloudSearch

Elastic Transcoder:

Amazon Elastic Transcoder is media transcoding in the cloud. It is designed to be a highly scalable, easy to use and a cost effective way for developers and businesses to convert (or “transcode”) media files from their source format into versions that will playback on devices like smart phones, tablets and PCs.

- Media transcoder in the cloud
- Convert media files from their original source format to different formats that will play on smart phones, tablets, PC's etc.
- Provides transcoding presets for popular output formats, which means you don't need to know or guess with which settings work best on which devices
- Pay based on the minutes that you transcode and the resolution at which you transcode.

Resource or Operation	Default Limit	US-EAST (VA) , US-WEST(Oregon), EU (Ireland)	All Others
Pipelines per region:	4		
User-defined presets:	50		
Max no. of jobs processed simultaneously by each pipeline:	N/A	20	12

For additional information about ElasticTranscoder Limits, see Limits in Amazon ElasticTranscoder

SES (Simple E-Mail Service):

Amazon Simple Email Service (Amazon SES) is a cost-effective email service built on the reliable and scalable infrastructure that Amazon.com developed to serve its own customer base. With Amazon SES, you can send and receive email with no required minimum commitments – you pay as you go, and you only pay for what you use.

- Not covered as exam topic currently

Resource or Operation	Default Limit
Daily sending quota:	200 messages per 24 hour period
Maximum send rate:	1 EMail per second
Recipient address verification:	All recipient addresses must be verified

Maximum Send Rate:
The rate at which Amazon SES accepts your messages might be less than the maximum send rate.

For additional information about Simple E-Mail Service Limits, see Limits in Amazon SES

SQS (Simple Queue Service):

Web service that gives you access to a message queue that can be used to store messages while waiting for a computer to process them. SQS is a distributed queue system that enables applications to quickly and reliably queue messages that one component of the application generates to be consumed by another component. A queue is a temp repository for messages that are awaiting processing.

- Used to allow customers the ability to decouple infrastructure components
- Very first service AWS released. Even older then EC2
- Messages can contain up to 256 KB of text in any format
- Acts as a buffer between the component producing and saving data, and the component receiving and processing the data

and processing the data
- Ensures deliver of each message at least once and supports multiple readers and writers interacting with the same queue
- A single queue can be used simultaneously by many distributed application components, with no need for those components to coordinate or communicate with each other
- Will always be available and deliver messages
- Does not guarantee FIFO delivery of messages
- Messages can be delivered multiple times and in any order
- FIFO is not supported
- If sequential processing is a requirement, sequencing information can be placed in each message so that message order can be preserved
- SQS always asynchronously PULLs messages from the queue
- Retention period of 14 days
- 12 hour visibility timeout by default
- If you find that the default visibility timeout period (12 hours) is insufficient to fully process and delete the message, the visibility timeout can be extended using the ChangeMessageVisibility action
- If the ChangeMessageVisibility action is specified to set an extended timeout period, SQS restarts the timeout period using the new value
- Engineered to provide delivery of all messages at least one
- Default short polling will return messages immediately if messages exist in the queue
- Long polling is a way to retrieve messages from a queue as soon as they are available; long polling requests don't return a response until a message arrives in the queue
- Maximum long poll time out is 20 seconds
- 256kb message sizes (originally 64kb)
- Billed for 64kb chunks
- First million messages free, then $.50 per additional million thereafter
- Single request can have from 1 to 10 messages, up to a max payload of 256KB
- Each 64KB chunk of payload is billed as 1 request. If you send a single API request with a 256KB payload, you will be billed for 4 requests (256/64 KB chunks)
- "Decouple" = SQS on exam
- Auto-scaling supported
- Message prioritization is not supported
- Process:
    - Component 1 sends a message to the queue
    - Component 2 retrieves the message from the queue and starts the visibility timeout period
    - Visibility timer only starts when the message is picked up from the queue
    - Component 2 processes the message and then deletes it from the queue during the visibility timeout period
    - If the visibility timeout period expires, the message will stay in the queue and not be deleted
    - The process is only complete when the queue receives the command to delete the message from the queue

For additional information about SQS Limits, see Limits in Amazon SQS

SWF (Simple Workflow Service)

Simple Workflow Service is a web service that makes it easy to coordinate work across distributed application components. Enabled for a range of uses such as media processing, web back ends, business process work-flows, and analytics pipelines, all to be designed as a coordination of tasks. Tasks represent invocations of various processing steps in an application which can be performed by code, API calls, human action and scripts.

- Build, run and scale background jobs or tasks that have sequential steps
- Way to process human oriented tasks using a framework
- SQS has a retention period of 14 days, vs SWF has up to a 1 year for work-flow executions
- Workflow retention is always shown in seconds (3.1536E+07 seconds)
- "Task could take a month" = SWF, as SQS only has a 14 day retention
- Presents a task-oriented API, whereas SQS offers a message-oriented API
- Ensures a teaks is assigned only once and is never duplicated; SQS duplicate messages are

allowed, and must be handled
- Keeps track of all tasks and events in an application, SQS would need an implementation of a custom application-level tracking mechanism
- A collection of work-flows is referred to as a domain
- Domains isolate a set of types, executions, and task lists from others within the same account
- You can register a domain by using the AWS console or using the RegisterDomain action in the SWF API
- Domain parameters are specified in JSON format
- SWF Actors:
    - Workflow starters - An application that can initiate a Workflow
    - Decider's - Control the flow or coordination of activity tasks such as concurrency, or scheduling in a work-flow execution; If something has finished in a work-flow (or fails), a decider decides what to do next
    - Activity Workers - Programs that interact with SWF to get tasks, process received tasks, and return the results
- Brokers the interactions between workers and the decider; Allows the decider to get consistent views into the progress of tasks and to initiate new tasks in an ongoing manner
- Stores tasks, assigns them to workers when they are ready and monitors their progress
- Ensures that a task is assigned only once and is never duplicated
- Maintains the application state durably, workers and decider's don't have to keep track of the execution state, and can run independently, with the ability to scale quickly

For additional information about SWF Limits, see Limits in Amazon SWF

Developer Tools:

CodeCommit:

AWS CodeCommit is a fully-managed source control service that makes it easy for companies to host secure and highly scalable private Git repositories.

- AWS implementation of GIT
- Not covered as exam topic currently

Resource or Operation	Default Limit
Number of repositories per account:	1000

For additional information about Code Commit Service Limits, see Limits in Amazon CodeCommit

CodeDeploy:

AWS CodeDeploy is a service that automates code deployments to any instance, including Amazon EC2 instances and instances running on-premises.

- Automate code deployments, AWS CI/CD service
- Not covered as exam topic currently

Resource or Operation	Default Limit
Number of applications under an account in a single region:	40

Number of concurrent deployments under an account:	10
Number of deployment groups associated with a single application:	50
Number of instances in a single deployment:	50

For additional information about Code Deploy Service Limits, see Limits in Amazon CodeDeploy

CodePipeline:

AWS CodePipeline is a continuous delivery service for fast and reliable application updates. CodePipeline builds, tests, and deploys your code every time there is a code change, based on the release process models you define.

- Build, test, and deploy code based on commits
- Not covered as exam topic currently

Resource or Operation	Default Limit
Number of pipelines per AWS account:	20
Number of stages in a pipeline:	Minimum of 2, maximum of 10
Number of actions in a stage:	Minimum of 1, maximum of 20
Number of parallel actions in a stage:	5
Number of sequential actions in a stage:	5
Number of custom actions per AWS account:	20
Maximum number of revisions running across all pipelines:	20
Maximum size of source artifacts:	500 megabytes (MB)
Maximum number of times an action can be run per month:	1,000 per calendar month

For additional information about Code Pipelines Service Limits, see Limits in Amazon CodePipelines

Service Limit Changes:
It may take up to two weeks to process requests for a limit increase.

Mobile Services:

Mobile Hub:

AWS Mobile Hub lets you easily add and configure features for your mobile apps, including user authentication, data storage, backend logic, push notifications, content delivery, and analytics. After you build your app, AWS Mobile Hub gives you easy access to testing on real devices, as well as analytics dashboards to track usage of your app – all from a single, integrated console.

- Build, run, and test usage of your mobile applications
- Not covered as exam topic currently

Cognito:

Amazon Cognito lets you easily add user sign-up and sign-in to your mobile and web apps. With Amazon Cognito, you also have the options to authenticate users through social identity providers such as Facebook, Twitter, or Amazon, with SAML identity solutions, or by using your own identity system. In addition, Amazon Cognito enables you to save data locally on users devices, allowing your applications to work even when the devices are offline. You can then synchronize data across users devices so that their app experience remains consistent regardless of the device they use.

- Save mobile data like game states or preferences
- Not covered as exam topic currently

Device Farm:

AWS Device Farm is an app testing service that lets you test and interact with your Android, iOS, and web apps on many devices at once, or reproduce issues on a device in real time. View video, screenshots, logs, and performance data to pinpoint and fix issues before shipping your app.

- Enables customers to test their mobile applications against real smart phones in the cloud
- Not covered as exam topic currently

Resource or Operation	Default Limit
App file size you can upload:	4 GB
Number of devices AWS Device Farm can test during a run:	5 which can be increased to 1K upon request
Number of devices you can include in a test run:	None
Number of runs you can schedule:	None
Duration of a remote access session:	60 Minutes

Mobile Analytics:

With Amazon Mobile Analytics, you can measure app usage and app revenue. By tracking key trends such as new vs. returning users, app revenue, user retention, and custom in-app behavior events, you can make data-driven decisions to increase engagement and monetization for your app.

- Measure mobile application usage, revenue and track new/returning users, etc..
- Not covered as exam topic currently

SNS (Simple Notification Service):

Simple Notification Service is a web service that makes it easy to set up, operate, and send notifications from the cloud. It provides developers with a highly scalable, flexible, and cost-effective capability to publish messages from an application and immediately deliver them to subscribers or other applications.

- Web service that allows customers to setup, operate, and send notifications from the cloud

- Can push to Apple, Google, FireOS, and Windows devices, as well as Android devices in China with Baidu cloud push
- Follows the publish-subscribe (pub-sub) messaging paradigm, with notifications being delivered to clients using a push mechanism that eliminates the need to poll for updates
- Can deliver notifications by SMS, email, SQS queues, or any HTTP endpoint
- SNS notifications can be used to trigger lambda functions
- When a message is published to an SNS topic that has a lambda function subscribed to it, the function is invoked with the payload of the published message. The lambda function would receive the message payload as an input parameter, and can manipulate the info in the message, publish the message to other SNS topics or send the message to other AWS services
- Allows you to group multiple recipients using topics
- Topics are access points for allowing recipients to dynamically subscribe for copies of the notification
- One topic can support deliveries to multiple endpoint types, for example, IOS, Android, and SMS recipients can be grouped together
- When message is published, SNS delivers appropriately formatted copies of your message to each subscriber
- Email notifications will be JSON formated not XML
- Subscriptions have to be confirmed
- Subscription expire after 3 days if they are not confirmed
- TTL is the number of seconds since the message was published
- If the message is not delivered within the TTL time, then the message will expire
- To prevent messages from being lost, all messages published to SNS are stored redundantly across multiple AZ's
- Instantaneous, PUSH based delivery (No Polling) --> SQS requires polling
- Simple API and easy integration with applications
- Flexible message deliver over multiple transport protocols
- Inexpensive, pay as you go model
- Web based AWS management console offers simplicity of point and click interface
- $.50 per million SNS requests
- $.06 per 100,000 notification deliveries over HTTP
- $0.75 per 100 notifications over SMS
- $2.00 per 100,000 notification deliveries over email
- Can be used in conjunction with SQS to fan a single message out to multiple SQS queues
- Remember:
    - SNS - PUSH
    - SQS - PULL (poll)
- Subscribers:
    - HTTP
    - HTTPS
    - Email
    - Email-JSON
    - SQS
    - Application
    - Lambda
    - Messages can be customized for each of the available protocols

Resource or Operation	Default Limit
Topics :	100,000
Account spend threshold for SMS:	50 USD
Delivery rate for promotional SMS messages:	20 Messages per second
Delivery rate for transactional SMS messages:	20 Messages per second

Enterprise Applications:

Workspaces:

Amazon WorkSpaces is a fully managed, secure desktop computing service which runs on the AWS cloud. Amazon WorkSpaces allows you to easily provision cloud-based virtual desktops and provide your users access to the documents, applications, and resources they need from any supported device, including Windows and Mac computers, Chromebooks, iPads, Fire tablets, and Android tablets.

- Virtual Desktop Infrastructure (VDI) that provides a bundle of compute resources, storage space, and software application access that allow a user to interact with just as a traditional desktop
- Users can connect to a WorkSpace from any supported device (PC, Mac, Chrome-book, iPad, Kindle Fire, or Android) using a free Workspace Client application
- Can be integrated into Active Directory using federated services
- Runs Windows 7 provided by Windows Server 2008 R2
- Users can personalize their workspace with their favorite settings for items such as wallpaper, icons, shortcuts, etc. This can be locked down by an administrator
- By default you will be given local admin access so you can install your own applications
- Workspaces are persistent
- All data on the D:\ is backed up every 12 hours

Resource or Operation	Default Limit	Comments
WorkSpaces:	5	To prevent denial of service attacks, accounts new to the Amazon WorkSpaces service are limited to five WorkSpaces.

For additional information about Workspaces Limits, see Limits in Amazon WorkSpaces

WorkDocs:

Amazon WorkDocs is a fully managed, secure enterprise storage and sharing service with strong administrative controls and feedback capabilities that improve user productivity.

- AWS version of Dropbox for the enterprise
- Not covered as exam topic currently

WorkMail:

Amazon WorkMail is a secure, managed business email and calendar service with support for existing desktop and mobile email clients.

- AWS version of Exchange Server for E-mail Services
- Not covered as exam topic currently

Internet of Things:

IoT (Internet of Things):

AWS IoT is a managed cloud platform that lets connected devices easily and securely interact with cloud applications and other devices. AWS IoT can support billions of devices and trillions of messages, and can process and route those messages to AWS endpoints and to other devices reliably and securely.

- Not covered as exam topic currently

Resource or Operation	Default Limit
Topic length limit:	The topic passed to the message broker when publishing a message cannot exceed 256 bytes encoded in UTF-8.
Restricted topic prefix:	Topics beginning with '$' are considered reserved and are not supported for publishing and subscribing except when working with the Thing Shadows service.
Maximum number of slashes in topic and topic filter:	A topic provided while publishing a message or a topic filter provided while subscribing can have no more than eight forward slashes (/).
Client ID size limit:	128 bytes encoded in UTF-8.
Restricted client ID prefix:	'$' is reserved for internally generated client IDs.
Message size limit:	The payload for every publish message is limited to 128 KB. The AWS IoT service will reject messages larger than this size.
Throughput per connection:	AWS IoT limits the ingress and egress rate on each client connection to 512 KB/s. Data sent or received at a higher rate will be throttled to this throughput.
Maximum subscriptions per subscribe call:	A single subscribe call is limited to request a maximum of eight subscriptions.
Subscriptions per session:	The message broker limits each client session to subscribe to up to 50 subscriptions. A subscribe request that pushes the total number of subscriptions past 50 will result in the connection being disconnected.
Connection inactivity (keep-alive) limits:	By default, an MQTT client connection is disconnected after 30 minutes of inactivity. When the client sends a PUBLISH, SUBSCRIBE, PING, or PUBACK message, the inactivity timer is reset. A client can request a shorter keep-alive interval by specifying a keep-alive value between 5-1,200 seconds in the MQTT CONNECT message sent to the server. If a keep-alive value is specified, the server will disconnect the client if it does not receive a PUBLISH, SUBSCRIBE, PINGREQ, or PUBACK message within a period 1.5 times the requested interval. The keep-alive timer starts after the sender sends a CONNACK. If a client sends a keep-alive value of zero, the default keep-alive behavior will remain in place. If a client request a keep-alive shorter than 5 seconds, the server will treat the client as though it requested a keep-alive interval of 5 seconds. The keep-alive timer begins immediately after the server returns a CONNACK to the client. There may be a brief delay between the client's sending of a CONNECT message and the start of keep-alive behavior.
Maximum inbound unacknowledged messages:	The message broker allows 100 in-flight unacknowledged messages (limit is across all messages requiring ACK). When this limit is reached, no new messages will be accepted until an ACK is returned by the server.
Maximum outbound unacknowledged messages:	The message broker only allows 100 in-flight unacknowledged messages (limit is across all messages requiring ACK). When this limit is reached, no new messages will be sent to the client until the client acknowledges the in-flight messages.
Maximum retry interval for	If a connected client is unable to receive an ACK on a QoS 1 message for one hour, the message broker will drop the message. The client may be unable to receive the

delivering QoS 1 messages:	message if it has 100 in-flight messages, it is being throttled due to large payloads, or other errors.
WebSocket connection duration:	WebSocket connections are limited to 24 hours. If the limit is exceeded, the WebSocket connection will automatically be closed when an attempt is made to send a message by the client or server. If you need to maintain an active WebSocket connection for longer than 5 minutes, simply close and re-open the WebSocket connection from the client side before the 5 minutes elapses.
IoT rules per AWS account	1000

The following limits apply to thing shadows:

Resource or Operation	Default Limit
Maximum size of a JSON state document:	The maximum size of a JSON state document is 8 KB.
Maximum number of JSON objects per AWS account:	There is no limit on the number of JSON objects per AWS account.
Shadow lifetime:	A thing shadow is deleted by AWS IoT if it has not been updated or retrieved in more than 1 year.
Maximum number of in-flight, unacknowledged messages:	The Thing Shadows service supports up to 10 in-flight unacknowledged messages. When this limit is reached, all new shadow requests will be rejected with a 429 error code.
Maximum depth of JSON device state documents:	The maximum number of levels in the "desired" or "reported" section of the JSON device state document is 5.

The following limits apply to security:

Resource or Operation	Default Limit
Policies that can be applied to an AWS IoT certificate:	10
Number of versions of a named policy:	5
Policy document size limit:	2048 characters

Throttling Limits:

Resource or Operation	Default Limit
AcceptCertificateTransfer:	10
AttachThingPrincipal:	15
CancelCertificateTransfer:	10
CreateCertificateFromCsr:	15
CreatePolicy:	10
CreatePolicyVersion:	10
CreateThing:	15

DeleteCertificate:	10
DeleteCACertificate:	10
DeletePolicy:	10
DeletePolicyVersion:	10
DeleteThing:	10
DescribeCertificate:	10
DescribeCACertificate:	10
DescribeThing:	10
DetachThingPrincipal:	10
DetachPrincipalPolicy:	15
DeleteRegistrationCode:	10
GetPolicy:	10
GetPolicyVersion:	15
GetRegistrationCode:	10
ListCertificates:	10
ListCertificatesByCA:	10
ListPolicies:	10
ListPolicyVersions:	10
ListPrincipalPolicies:	15
ListPrincipalThings:	10
ListThings:	10
ListThingPrincipals:	10
RegisterCertificate:	10
RegisterCACertificate:	10
RejectCertificateTransfer:	10
SetDefaultPolicyVersion:	10
TransferCertificate:	10
UpdateCertificate:	10
UpdateCACertificate:	10
UpdateThing:	10

Well Architected Framework:

- Consists of 4 pillars:
    - Security
        - Apply security at all layers

- Enable Traceability
- Automate response to security events
- Focus on securing your system
- Automate security best practices
- Encrypt your data both in transit and at rest using ELB, EBS, S3 and RDS
- Use IAM and MFA for privilege management
- Security in the cloud has 4 areas:
    - Data Protection
        - Organize and classify your data into segments such as public, available only to org/dept/user
        - Implement a least privilege access system so people can only access what they need
        - Encrypt everything where possible, whether it be at rest or in transit
        - Customers maintain full control of your data
        - AWS makes it easy to manage keys using KMS or KMS-C
        - Detailed logging is available that contains important content such as file access and changes
        - Designed storage systems for exceptional resiliency.
        - S3 is designed for 11 nines durability. If you store 10K objects on S3, you can on average expect to incur a loss of a single object once every 10,000,000 years.
        - Versioning which can protect against accidental overwrites, deletes, and similar harm
        - AWS never initiates the movement of data between regions. Content placed in a region will remain in that region, unless manually moved.
    - Privilege Management
        - Ensures that only authorized and authenticated users are able to access your resources
        - Mechanisms in place such as ACLs, Role based access controls, Password management such as password rotation policies
    - Infrastructure Protection
        - How do you protect your data center
        - RFID controls
        - Security
        - Lockable cabinets
        - CCTV
        - Amazon handles all of the physical, really customer is responsible for VPC protection.
        - Enforce network and host level boundary protection
        - Enforce the integrity of the OS, updates, patches, and anti-virus
    - Detective Controls
        - Detect or identify a security breach, tools available to help with this are:
            - CloudTrail
            - CloudWatch
            - AWS Config
            - S3
            - Glacier

- Reliability
    - Ability of a system to recover from a service or infrastructure outage/disruptions
    - Ability to dynamically acquire computing resources to meet demand
    - Test recovery procedures
    - Automatically recover from failure
    - Scale horizontally to increase aggregate system availability
    - Stop guessing capacity
    - Consists of 3 areas:
        - Foundations:
            - Make sure you have the prerequisite foundations in place
            - Consider the size of communication links between HQ and data centers
            - Mis-provisioning connections could result in 3-6 upgrade time-frames
            - AWS handles most of the foundations for you. The cloud is designed to be essentially limitless meaning that AWs handles the networking, and compute requirements themselves. They set service limits to limit accidental spin up of too many resources.

        - Change Management:
            - Be aware of how change affects a system so you can plan pro-actively around it.
            - Monitoring allows you to detect any changes to your environment and react.
            - Traditionally change control is done manually and carefully co-ordinated with auditing
            - CloudWatch can be configured to monitor your environment and services such as auto-scaling, to automate change in response to changes in your prod environment.
        - Failure Management:
            - Always architect your system with the assumption that failure will occur
            - Become aware of these failures, how they occurred, how to respond to them and then plan on how to prevent them in the future.
    - Performance Efficiency:
        - Focuses on how to use computing resources efficiently to meet requirements
        - How to maintain that efficiency as demand changes and technology evolves
        - Democratize advanced technologies (Consume as service vs setup and maintain)
        - Go Global in minutes
        - Use server-less architectures
        - Experiment more often
        - Consists of 4 areas:
            - Compute:
                - Choose the right kind of server
                - AWS servers are virtualized and at the click of a button you can change server types
                - You can even switch to running with no servers, and use Lambda
            - Storage:
                - Optimal storage solutions for your environment depend on access methods (block, file or object), patterns of access, throughput, frequency of access, frequency of update, availabilty constraints, and durability constraints.
                - S3 has 11x9's durability and cross region replication
                - EBS has different mediums such as magnetic, SSD, or provisioned IOPS SSD
                - Can easily switch between different mediums
            - Databases:
                - Optimal database solution depends on number of factors, do you need database consistency, high availability, No-SQL, DR, Relational tables?
                - Lots of options, RDS, DynamoDB, Redshift, etc..
            - Space Time Trade off:
                - Using services such as RDS to add read replicas reduces the load of your database and creates multiple copies of the data to help lower latency
                - Can use Direct Connect to provide predictable latency between HQ and AWS
                - Use the global infrastructure to have copies of environment in regions closest to where your customer base is located.
                - Caching services such as Elasticache or CloudFront to reduce latency
    - Cost Optimization
        - Reduce cost to minimum and use those saving for other parts of your business
        - Allows you pay the lowest price possible while still achieving your business objectives
        - Transparently attribute expenditure
        - Use managed services to reduce the cost of ownership
        - Trade capital expense for operating expense
        - Benefit from economies of scale (AWS buys servers by the thousands)
        - Stop spending money on data center operations
        - Design Principles:
            - Stop guessing your capacity needs
            - Test systems at production scale
            - Lower the risk of architecture change
            - Automate to make architectural experimentation easier
            - Allow for evolutionary architectures
        - Comprised of 4 different areas:

- Matched Supply and demand
  - Align supply with demand
  - Don't over or under provision, instead expand as demand grows
  - Auto-scaling or lambda execute or respond when a request comes in
  - Services such as CloudWatch can help you keep track as to what your demand is.
- Cost-Effective resources
  - Use correct instance type
  - Well architected system will use the most cost efficient resources to reach the end business goal
- Expenditure awareness
  - No longer need to get quotes for physical servers, choosing a supplier, have resources delivered, installed, manufactured, etc..
  - Can provision things within seconds
  - Be aware of what each team is spending and where is crucial to any well architected system
  - Use cost allocation tags to track this, billing alerts as well as consolidated billing.
- Optimizing over time
  - A service that you chose yesterday man not be the best service to be using today
  - Constantly re-evaluate your existing architecture
  - Subscribe to the AWS blog
  - Use Trusted Advisor

White Paper Review:

- 6 Advantages of Cloud
  - Trade capital expense for variable expense
  - Benefit from massive economies of scale
  - Stop guessing about capacity
  - Increase speed and agility
  - Stop spending money running and maintaining data centers
  - Go Global in minutes
- 14 Regions, each with different number of AZ's
- Storage devices uses DoD 5220.22-M or NIST 800-88 to destroy data when a device has reached the end of its useful life. All decommissioned magnetic storage devices are degaussed and physically destroyed in accordance with industry standard practices
- VPC provides a private subnet within the cloud and the ability to use an IPsec VPN to provide an encrypted tunnel between the VPC and your data center
- AWS prod is segregated from the AWS Corporate network by means of a complex set of network security / segregation devices
- Provides protection against DDoS, Man in the middle attacks, IP spoofing, Port Scanning, and Packet Sniffing by other tenants
- AWS has a host based firewall infrastructure that will not permit an instance to send traffic with a source IP or MAC address other than its own, which prevents IP Spoofing
- Unauthorized port scans by EC2 customers are a violation of the Acceptable use policy
- You may request permission to conduct vulnerability scans as required to meet your specific compliance requirements
- Any pre-approved vulnerability scans must be limited to your own instances and must not violate the Acceptable use policy; You MUST request a vulnerability scan in advance
- Password for root or IAM user accounts into the console should be protected by MFA
- Use access keys to access AWS APIs (using AWS SDK, CLI, REST/Query APIs)
- Use SSH Key Paris to login to EC2 instances, or CloudFront signed URLS
- Use x.509 Certs to tighten security of your applications/cloudfront via HTTPS
- Trusted Advisor inspects your environment and makes recommendations when opportunities exist to save money, improve system performance, or close security gaps
- Different instances running on the same physical machine are isolated from each other via the Xen hypervisor
- AWS firewall resides within the hypervisor layer, between the physical network and the the instances virtual interface.

- ALL packets must pass through this layer. Any instance's neighbors have no more access to the instance than any other host on the Internet and can be treated as if they are separate hosts
- Physical RAM is separated using similar mechanisms
- Customer instances have no access to raw disk devices, but instead are presented with virtualized disks
- AWS proprietary disk virtualization layer automatically resets every block of storage used by the customer, so that one customers data is never unintentionally exposed to another
- Memory allocated to guests is scrubbed (set to 0) by the hypervisor when it is unallocated to a guest
- Memory is not returned to the pool of free memory available for new allocations until th memory scrub process has completed
- Virtual instances are completely controlled by you, the customer. You have full root access or administrative control over accounts, services, and applications. AWS does not have any access rights to any instance or guest OS
- EC2 provides a complete firewall solution. The inbound firewall is configured in a default deny any any mode and EC2 customers must explicitly open the ports needed to allow inbound traffic
- Encryption of sensitive data is generally a good practice and AWS provides the ability to encrypt EBS volumes and their snapshots with AES-256. The encryption occurs on the servers that host the EC2 instances and EBS storage
- EBS encryption feature is only available on EC2's more powerful instance types (M3, C3, R3, G2)
- SSL termination on ELB is supported and recommended
- X- forwarded for headers enabled, passes real IP from LB's to web servers
- You can procure rack space within the facility housing the AWS direct connect location and deploy your equipment nearby. Once deployed, you can connect to this equipment to AWS direct connect using cross-connect
- Using 802.1q VLANs dedicated connections can be partitioned into multiple virtual interfaces. This allows you to use the connection to access public resources such as objects stored in S3 using public IP address space and private resources such as EC2 instances running within the VPC private IP space, while maintaining network separation between public and private environments
- AWS management re-evaluates the strategic business plan at least bi-annually
- AWS security regularly scans all Internet facing service endpoint IP addresses for vulnerabilities. These do NOT include customer instances
- External vulnerability threat assessments are performed regularly by independent security firms, and their findings are passed to management
- Data Center Security:
    - State of the art electronic surveillance and MF access control
    - Staffed 24x7 by security guards
    - Access is authorized on a least privilege basis
- Compliance:
    - SOC 1/SSAE 16/ISAE 3402 (formally SAS 70 Type II)
    - SOC2
    - SOC3
    - FISMA, DIACAP, and FedRAMP
    - PCI DSS Level 1
    - ISO 27001
    - ISO 9001
    - HIPAA
    - Cloud Security Alliance (CSA)
    - Motion Picture Association of America (MPAA)
    - ITAR
    - FIPS 140-2
    - DSS 1.0
- Data Security:
    - Shared security model
        - AWS:
            - Responsible for securing the underlying infrastructure
            - Responsible for protecting the global infrastructure that runs all of the services offered on the AWS cloud.
            - Infrastructure comprised of hardware, software, networking, and facilities that run AWS services
            - Responsible for the security configuration of its products that are considered managed services, such as DynamoDB, RDS, Redshift, Elastic MapReduce, lambda, and Workspaces.

. .

- User:
  - Responsible for anything put on the cloud
  - EC2, VPC, S3 security configuration and management tasks
  - Account Management (MFA, SSL, TLS, CloudTrail API/User activity logging)