**Review**

**SECURITY**

Models:

* CIA: Confidentiality (MFA/IAM), Avialbility (Auto-scaling/Mutly-AZ), Integrity (IAM/BucketPolicies/CertificateManager)
* AAA: Authentication (IAM), Authorization (Policies), Accounting (CloudTrail)
* Non-repudiation (can’t deny that you did something)

[Amazon.com](http://amazon.com/) network is completely separated/isolated from the AWS network.

Explicit Denies on policies, ALWAYS take precendence over any Allow.

**AWS CONFIG**

AWS Config enables you to assess, audit, and evaluate the configurations of your AWS resources. It continuously monitors and records your AWS resource configurations and allows you to automate the evaluation of recorded configurations against desired configurations.

As an example, if you would like to be notified if any of the systems administrators in your organization creates a security group with SSH open to the world, you can create a rule in AWS Config to notify you using SNS if this happens.

**KMS**

AWS KMS can be found within IAM, n the Encryption Keys section. It cannot be found just as any other regular service from the console.

When you create an Encryption Key, you can create it as a KMS Key or an External Key. A KMS Key has Origin Key Material, which means that the key material is provided by AWS. Also, a KMS cannot be exported.

The *External Key* requires to have key material to be imported after the Key has been created. You then need to download the *wrapping key* and *import token*, and generate the *Encrypted key material* and the *import token* by means of (for example) Open SSL. Once you have those, you can import the key into IAM and complete the generation of the Key. If you use the files for creating a given Key, you won’t be able to use the same Hash files for creating a different key.

You cannot enable Automatic Key Rotation for a CMK (Customer Master Key) with imported key material. The CMKs need to be rotated manually because it wil require to have a new CMK with new key material to be manipulated by the KMS Administrator and then to replace the previous CMK with the new one.

A CMK can be re-created, since AWS will have enough information on their end to do so, but you will need to provide the original key material that you used when creating the CMK.

When you create a Key, you may allow an external account to use that key by specifying the ARN of the account. If a Key Administrator encypts data and then deletes the Key, you won’t be able to access that data any longer.

The Resources that will be using a given Key, have to be locted in the same Region where the Key was created.

AWS KMS supports two types of keys - Master Keys and Data Keys. A Data Key is used to encrypt and decrypt the actual data; whereas, the Master Key is used to protect (encrypt and decrypt) the data key as well as some data upto 4Kib.

Even though IAM is a Global service, Encryption keys are stored locally on a given Region.

You will not be able to view an object in an S3 bucket that is encrypted with KMS via the Public URL, but you will be able to view it if you encrypt it with AES256. So, if anyone accidentally makes anything public, nobody will be able to actually view it either way.

You can create up to 1.000 CMKs per account, per region. This is a soft limit.

**SECURE TOKEN SERVICE**

STS enables:

* Web ID Federation
* AD Federation
* Cross Account Access.

**CLOUDTRAIL**

You can enable Log File Integrity Validation for CloudTrail to ensure that if a Log File is changed.

CloudTrail is enabled by Default with a Retention of 7 days.

You can enable LifeCycle policies in the S3 Bucket where CloudTrail logs are stored, so that they are deleted after a given period of time. You can also enable Encryption on the bucket.

**CLOUDWATCH**

RDS can be monitored by Metrics (from CloudWatch) or by Events (from RDS itself, which are based on SNS).

ELB reports only when requests are flowing through it. The metrics are sent to CloudWatch in 60 seconds intervals.

CloudWatch launched High Resolution Custom Metrics on July 26, 2017. This enables you to publish and store custom metrics down to 1-second resolution. Extended retention of metrics was launched on November 1, 2016, and enabled storage of all metrics for customers from the previous 14 days to 15 months.

CloudWatch retains metric data as follows:

* Data points with a period of less than 60 seconds are available for 3 hours. These data points are high-resolution custom metrics
* Data points with a period of 60 seconds (1 minute) are available for 15 days
* Data points with a period of 300 seconds (5 minute) are available for 63 days
* Data points with a period of 3600 seconds (1 hour) are available for 455 days (15 months)

Data points that are initially published with a shorter period are aggregated together for long-term storage. For example, if you collect data using a period of 1 minute, the data remains available for 15 days with 1-minute resolution. After 15 days this data is still available, but is aggregated and is retrievable only with a resolution of 5 minutes. After 63 days, the data is further aggregated and is available with a resolution of 1 hour. If you need availability of metrics longer than these periods, you can use the GetMetricStatistics API to retrieve the datapoints for offline or different storage.

CloudWatch Logs: You can use Amazon CloudWatch Logs to monitor, store, and access your log files from Amazon EC2 instances, AWS CloudTrail, Route 53, and other sources. You can then retrieve the associated log data from CloudWatch Logs. Features:

* Monitor Logs from Amazon EC2 Instances in Real-time: You can use CloudWatch Logs to monitor applications and systems using log data.
* Monitor AWS CloudTrail Logged Events: You can create alarms in CloudWatch and receive notifications of particular API activity as captured by CloudTrail and use the notification to perform troubleshooting.
* Archive Log Data: You can use CloudWatch Logs to store your log data in highly durable storage.
* Log Route 53 DNS Queries: You can use CloudWatch Logs, to log information about the DNS queries that Route 53 receives.

The following services are used in conjunction with CloudWatch Logs:

* AWS CloudTrail
* AWS Identity and Access Management (IAM)
* Amazon Kinesis Data Streams
* AWS Lambda

CloudWatch consist of:

* Dashboard: allow you to add different widgets
* Alarms: you will be notified based on the settings of your alarms
* Events: you can configure Events and Events Rules to be executed when a given Event takes place. Rules match incoming Events and route them to one or more Targets (like Lambda, SNS topics, SQS queues, Kinesis streams, etc.)
* Logs: you can use this to send logs from your applications into CloudWatch. You have to install the CloudWatch agent on the given EC2 instance
* Metrics: you can setup custom metrics

**AWS CONFIG**

It has to be activated/deplouyed in a Region-by-Region basis.

Components are:

* Config Dashboard
* Config Rules
  + Managed
  + Custom
* Resources
* Settings

**CLI COMMANDS**

* mon-disable-alarm-actions will allow to disable CoudWatch alarms
* Ssh ec2-user@52..20.30.40 -i KeyPair.pem = connects via SSH to a linux ECS instance
* aws configure = configures the Access and Secret Access keys, the region and the default Output Format and saves that in the .aws/config in an EC2 instance, but it is much better to use an IAM Role
* aws ec2 describe-instances = list all instances according to the criteria that we defined
* aws ec2 describe-images = list all AMIs available to me to provision
* aws ec2 run-instances = deploys a new instance from an AMI
* aws ec2 start-instance and stop-instance = changes the instance’s status
* aws ec2 terminate-instances = terminates the specified instance/s

**COGNITO**

Amazon Cognito provides authentication, authorization, and user management for your web and mobile apps. Your users can sign in directly with a user name and password, or through a third party such as Facebook, Amazon, or Google. Further information: <https://docs.aws.amazon.com/cognito/latest/developerguide/what-is-amazon-cognito.html>

**S3**

S3 bucket names may only contain lower case letters, periods, numbers, and dashes. Bucket names must not be formatted as an IP address, and they may not begin with a period.

S3 usage is charged based on three pricing components:

* Storage (per GB per month)
* Data transfer (per GB per month)
* Request (per thousand requests per month)

An S3 is a Key-Value object based storage type, that has:

* Key (name)
* Value (data)
* Version ID
* Metadata (data about data)
* ACL

S3 Website URL - [http://mybucketname.s3-website.us-east-1.amazonaws.com](http://mybucketname.s3-website.us-east-1.amazonaws.com/) (check on s3-website in the URL).

S3 Bucket URL - <https://s3.us-east-1.amazonaws.com/mybucketname>.

S3 Common Response Headers: <https://docs.aws.amazon.com/AmazonS3/latest/API/RESTCommonResponseHeaders.html>

S3 Common Request Headers: <https://docs.aws.amazon.com/AmazonS3/latest/API/RESTCommonRequestHeaders.html>

S3 maximum PUT in a single operation: 5 GB.

S3 maximum file size: 5 TB.

S3 Multipart upload is required for files larger than 5GB.

By default, customers can provision up to 100 buckets per AWS account.

An ideal way to backup an S3 bucket is to copy the contents to another S3 bucket owned by another S3 account.

To restrict access to an entire bucket, you use bucket policies; and to restrict access to an individual object, you use access control lists.

Bucket Policies are used to make entire buckets public (like one hosting an S3 website).

Lifecycle management:

S3 transition to IA has to be done 30 days after the creation date and has to be 128 KB in size.

S3 transition to Glacier has to be done 30 days after IA.

You can expire old version (delete marker set) and then Permanently Delete them.

Request rate performance considerations:

* If your workload in an Amazon S3 bucket routinely exceeds 100 PUT/LIST/DELETE requests per second or more than 300 GET requests per second, you will need to check your performance considerations
* If you expect a rapid increase in the request rate for a bucket to more than 300 PUT/LIST/DELETE requests per second or more than 800 GET requests per second, open a support case
* If your requests are typically a mix of GET, PUT, DELETE, or GET Bucket (list objects), choosing appropriate key names for your objects ensures better performance
* If the bulk of your workload are GET-intensive, we recommend using the Amazon CloudFront content delivery service

X.509 Certificates are used to sign SOAP-based requests. You can have AWS create one for you and then download the cert and the private key, or upload into AWS one of your own and use it. X.509 Certificates are ONLY used with S3 for now.

Glacier is designed with the expectation that retrievals are infrequent and unusual, and data will be stored for extended periods of time. You can retrieve up to 5 percent of your average monthly storage (prorated daily) for free each month. If you retrieve more than this amount of data in a month, you are charged an additional (per GB) retrieval fee. A prorated charge (per GB) also applies for items deleted prior to 90 days’ passage.

S3 Access Policies and ACLs:

Explicit Denies on Bucket Poicies, ALWAYS take precendence over any Allow.

If you enable Public Access to an S3 object or to a bucket, the Explicit Denies to IAM users will not be effective if you still use the S3 URL to access the bucket objects. This is because it will be accessing throught Public Internet and not as an authenticated IAM user.

It is a Best Practice to use Cross Region Replication to replicate the Cloud Trail Logs to another AWS account that only the Auditors have access to. This way, if a CloudTrail log is deleted from a given AWS account, the log that was replicated to the Auditing account will remain.

When granting access to an S3 object via a Presigned URL, you don’t need to grant access to the file at the object or bucket level. The Presigned URL will allow you to access the object.

A Glacier Vault is a container which stores one or more Glacier archives

When a Glacier Vault Lock is created it cannot be changed. A vault lock policy is different than a vault access policy. Both policies govern access controls to your vault. However, a vault lock policy can be locked to prevent future changes, providing strong enforcement for your compliance controls.

When configuring Cross Region Replication, the owner of the destination bucket must grant the owner of the source bucket permissions to replicate objects with a bucket policy.

To force your users to access your site using CloudFront and not directly using the S3 URL, you have to:

* Create an origin access identity, which is a special CloudFront user, associate the origin access identity with your distribution
* Change the permissions either on your Amazon S3 bucket or on the files in your bucket so that only the origin access identity has read permission (or read and download permission)
* Configure the bucket policy on your Amazon S3 bucket so that only the origin access identity has read permission for objects in the bucket
* When your users access your Amazon S3 files through CloudFront, the CloudFront origin access identity gets the files on behalf of your users.

Service Control Policies (SCPs) enable you to **restrict**, at the account level of granularity, what services and actions the users, groups, and roles in those accounts can do. So, you can restrict access to S3 across a number of different AWS accounts in your organization.

**STORAGE GATEWAY**

It can be deployed as Virtual Machine that will run on-prem (on VMware or HyperV) or as an EC2 instance, and it has the following functionalities:

* File Interface (Gateway)
  + For flat files stored directly to, and retrieved from, S3
  + It uses NFS to mount the S3 resources as NFS mountpoints
  + It works caching the most accessed S3 objects in the Storage Gateway and transferring new files and updates asynchronously to the S3 buckets
  + File Gateway integrates with on-prem and in-cloud Active Directory
  + Files should only be accessed through the NFS mountpoints (hence, through the Gateway) to avoid miss-behaviors
  + File contents are downloaded when the file is read. You can configure other File Gateways in reead-only mode of a File Share managed by other Gateway, but having multiple Gateways writing to the same File Share, though possible, is neither supported nor recommended
  + The Gateway reports 8EB as the size for any File Share, because it does not limit the total storage.
  + File Gateway makes use of multipart upload, but it does not make use of S3 Transfer Acceration, even if the S3 bucket is setup for S3 Transfer Acceleration

1. Volume Gateway (both are iSCSI based block storage)
   * Gateway-Stored Volumes, entire dataset is stored on site and async backed up to S3. Volumes can be up to **512 TB** in size, and a maximum of **32 volumes** can be stored per Storage Gateway
   * Gateway-Cached Volumes, Entire Dataset is stored on S3 and frequently accessed files are cached. It allows you to store volumes of up to **1024 TB,** for a total of up to **1 PB**
   * You will only be able to access the volumes through the Storage Gateway and **not** through the S3 bucket where the volumes are cached
   * you can snapshot your cached volumes on-demand, or schedule the snapshots
2. Gateway-VTL (iSCSI based virtual tape solution): Virtual Tapes will be stored in a Virtual Tape Library backed by S3 or on a Virtual Tape Shelf (VTS) backed by Glacier. VTL supports up to 1,500 tapes with a maximum aggregate of 1 PB, while VTS supports unlimited tapes. When creating virtual tapes you can set the tape size to one of these available sizes: 100 GB, 200 GB, 400 GB, 800 GB, 1.5 TB, 2.5 TB

Instead of backing up directly to S3, you can backup your data into Storage Gateway and have it replicate it to S3 on the backend automatically.

Storage Gateway supports bandwidth throttling, which means that you can restrict your bandwidth between the Storage Gateway and AWS.

Requirements for the on-premise HW are: 4vCPUs, 16GB of RAM, 80GB VM image and system data

Data in Transit is encrypted by means of SSL and using AES-256.

Storage Gateway traffic can be throttled.

**VMWARE CLOUD ON AWS**

VMware Cloud on AWS SDDC (Software Defined DataCenter) is directly connected to customer’s VPC using Elastic Network Interface(ENI) and therefore has access to AWS services.

You can leverage your existing VMware software investments to secure additional discounts for your VMware Cloud on AWS hybrid environment as part of VMware’s Hybrid Loyalty Program.

VMware Cloud on AWS uses a least privilege security model in which you (and therefore their tools) do not have full administrative access. Tools that require plug-ins or extensive vSphere permissions may not function properly.

There are multiple ways to migrate existing vSphere VMs to VMware Cloud on AWS. You can perform of a live migration of vSphere VMs via a vMotion or by leveraging VMware Hybrid Cloud Extension (HCX).

The VMware Cloud on AWS minimum standard cluster configuration contains 3 dedicated, single-tenant bare metal hosts. Each host is an Amazon EC2 I3.metal instance. These hosts have dual 2.3 GHz CPUs (custom-built Intel Xeon Processor E5-2686 v4 CPU package) with 18 cores per socket (36 cores total), 512 GiB RAM, and 15.2 TB Raw NVMe storage. You can add and/or remove hosts on-demand as long as the minimum cluster size is 3 hosts. The maximum cluster size is 16 ESXi hosts.

VMware Cloud on AWS includes VMware’s vSAN storage technology. VMware vSAN can also utilize Amazon Elastic Block Store (Amazon EBS) with VMware Cloud on AWS running on new Amazon EC2 R5.metal instances to augment existing SDDC for storage-dense environments. Storage per host ranges from 15 to 35 TB in increments of 5 TB.

Customer data at rest will be natively encrypted by vSAN. vSAN will use AWS Key Management Service (KMS) to generate the Customer Master Key (CMK). vSAN encryption at rest cannot be turned on or off for individual clusters; it is a cluster-wide setting that is always on by default when cluster is provisioned in the SDDC.

When you deploy an SDDC using VMware Cloud on AWS, it is configured with two networks: a management network and a compute network. The management network handles network traffic for the SDDC hosts, vCenter Server, NSX Manager, and other management functions. The compute network handles network traffic for your workload VMs.

EIPs are reserved and associated from the VMware Cloud on AWS account and routed to the NSX Edge Gateway.

A stretched cluster is a deployment model in which two or more VMware Cloud on AWS clusters are part of the same logical cluster but are located in separate Availability Zones. Stretched cluster is a deployment time decision. You cannot upgrade a non-stretched cluster to a stretched cluster. The maximum supported size for a stretched cluster is 16 nodes. An SDDC can have either single AZ clusters or stretched clusters.

**SNOWBALL AND SNOWMOBILE**

Snowball capacity: 50 and 80 TB

Snowball Edge capacity: 100 TB

Snowmobile capacity: 100 PB

All AWS Regions have 80 TB Snowballs while US Regions have both 50 TB and 80 TB models.

Snowball might not be the ideal solution if your data can be transferred over the Internet in less than one week. The Snowball appliance is purpose-built for efficient data storage and transfer, including a high-speed, 10 Gbps network connection designed to minimize data transfer times, allowing you to transfer up to 80 TB of data from your data source to the appliance in 2.5 days, plus shipping time. In this case, the end-to-end time to transfer the data into AWS is approximately a week,

The solutions replace Import/Export Disk, which was the old solution.

Import/Export allow to export info from S3 only,  but it allows to import into S3, EBS and Glacier.

Data encryption is optional for Imports only.

Snowball can Import to S3 and Export from S3.

Snowball Edge includes compute capacity, which allows to run lambdas in it.

Snowmobile can be requested with armed Security.

**RDS**

Automated Monitoring Tools for RDS:

* Amazon RDS Events: reports changes to DB instances, clusters, snapshots, cluster snapshots, DB parameters or DB Security Groups
* Database Logfiles: accessed via console or API
* Amazon RDS Enhanced Monitoring: real time metrics of the underlying OS of the RDS instance or cluster
* Amazon CloudWatch metrics: sent every minute
* Amazon CloudWatch Alarms
* Amazon CloudWatch Logs: for MariaDB, MySQL and Aurora MySQL

Manual Monitoring Tools for RDS:

* AWS console: connections, R/W operations, storage, CPU and memory used, Network traffic
* Trusted Advisor
* CloudWatch homepage

Multi-AZ:

* MySQL, Oracle and PostgreSQL use synch physical replication
* SQL Server uses synch logical replication

Read Replicas:

* MySQL, PostgreSQL and MariaDB use Asynch Read Replicas
* Aurora uses SSD-backed virtualized storage layer
* Read Replica is created from a snapshot or an automated backup
* If you Disabled backups of an RDS instance, you must modify the DB to turn them on to be able to create a Read Replica

Read Replicas for MariaDB, MySql and PostgreSQL can be replicated across different Regions. Aurora can replicate an entire DB cluster across different Regions.

A Read Replica can be promoted to be it’s own standalone database. This breaks the link between the Primary and the Secondary.

A Read Replica can now have Multi-AZ enabled (this was introduced very recently), and it can have a Read Replica configured from it (which will require to have a backup or a snapshot first).

In RDS, The maximum storage size for a Microsoft SQL Server instance running SQL Server Express edition is 300 GB.

In RDS, changes to the backup window take effect IMMEDIATELY.

RDS Reserved instances are available for multi-AZ deployments.

With new RDS DB instances, automated backups are enabled by default.

Backups cannot be taken from Read Replicas, but are automatically taken from the replica in a Multi-AZ environment.

RDS maintenance windows are 30 minutes by default.

Non-Relational DB:

* Database
  + Collection (would equal the Table in Relational DB)
  + Document (would equal the Row in Relational DB)
  + Key Value Pairs (would equal the Fields in Relational DB)

ACID is a term that comes from the early 70s, that relates to Relations DBs, which means that it represents what RDS is:

* Atomicity
* Consistency
* Isolation
* Durability

**AMAZON ATHENA**

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL.

**DYNAMODB**

DynamoDB Table -> Partition -> Items (Key + Composite Key, Attribute Names, Attribute Values):

* Partition Key (Hash Key) -> One Attribute
* Partition Key & Sort key (Hash & Range) -> Two Attributes (ie: Product ID + TimeStamp)

The maximum size of a DynamoDB item is 400KB, which includes Attribute Name and Attribute Value lengths.

The maximum length of a partition key value is 2048 bytes and the minimum is 1 byte.

The maximum length of a sort key value is 1024 bytes and the minimum is 1 byte.

DynamoDB Throughput Calculation: Free tier is 25 Read and 25 Write CUs

Read Throughput: Round up to the read items per sec to the next 4K chunk. Then divide by 4. Then multiply by the items to be read per sec. If it is Strongly Consistent Reads, leave as is. If it is Eventual Consistency, divide by 2.

Write Throughput: Since each Write Unit is 1KB of data, simply multiply the writes per second by the size of the items to be written per sec. Write does only allow Strong Consistency

BatchGetItem is the API call to retrieve multiple items from a DynamoDB table.

A single GetBatch operation can retrieve up to 16 MB of data, which may contain up to 100 items, or a partial result if the response size limit is exceeded.

DynamoDB API: <https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/HowItWorks.API.html>

Control plane operations let you create and manage DynamoDB tables. They also let you work with indexes, streams, and other objects that are dependent on tables.

Data plane operations let you perform create, read, update, and delete (also called CRUD) actions on data in a table.

-Local Secondary Index (LSI): Has the same Partition Key but different Sort Key

  LSIs are eventually consistent

* Can only be created when creating a table
* Cannot be deleted unless the table is deleted as well
* Support both, strong and eventual consistency read options
* Read and Write Units are consumed from the ones provisioned to the table that contains the LSI
* Only 5 LSI can be created per table
* A query in an LSI  can return attributes that were specified to be included in the LSI at creation time
* For the attributes not projected, DynamoDB will automatically fetch them from the table
* By default the attributes projected in an LSI are: the table partition key, the index sort key, and the table sort key
* Each table can have up to 20 projected attributes across all 5 possible LSIs

-Global Secondary Index (GSI): Has different Partition Key and different Sort Key

* GSIs are eventually consistent
* Can be created when creating the table or afterwards
* GSIs are limited to 10 GB per partition key value
* Only 5 GSI can be created per table
* A query in a GSI can only return attributes that were specified to be included in the GSI at creation time
* Having not enough Write Units on a GSI can throttle writes to the table even if the table does have enough WUs
* It is recommended to set the same Auto Scaling settings to the GSI as the table

Item Collections are group of items that have the same Partition Key across a table along with all the same Secondary Indexes.

The Default Region for the DynamoDB SDK is Oregon.

DynamoDB offers cross region replication, by creating tables that are automatically replicated across two or more AWS Regions, with full support for multi-master writes. This gives you the ability to build fast, massively scaled applications for a global user base without having to manage the replication process.

Optimistic concurrency depends on checking a value upon save to ensure that it has not changed. Pessimistic concurrency prevents a value from changing by locking the item or row in the database. **DynamoDB does not support item lockin**g, and conditional writes are perfect for implementing optimistic concurrency.

A Scan operation performs eventually consistent reads by default, and it can return up to 1 MB (one page) of data. Therefore, a single Scan request can consume (1 MB page size / 4 KB item size) / 2 (eventually consistent reads) = 128 read operations. If you request strongly consistent reads instead, the Scan operation would consume twice as much provisioned throughput—256 read operations.

DynamoDB Runs exclusively on SSD drives.

DynamoDB Triggers: it allows to trigger actions based on changes/updates to a Table. It relies on Lambda.

DynamoDB Streams will save all updates to the items in a table for a period of 24 hours. It will keep the order in which updates happened on a per item basis, but not among different items in the table.

DAX (DynamoDB Accelerator): it is an in-memory cache for DynamoDB. Only Eventual Reads are retrieves from DAX, since Strongly Consistent Reads will pasa through DAX and be retrieved from the DyanmoDB table.

You can **migrate** your DynamoDB table from one Region to another by using Import/Export to export the data to S3 and then import it back into a DynamoDB table in the target Region.

**REDSHIFT**

Redshift is a Database Engine purposed for BI Analytics.

Backups are enabled by default with a 1-day retention period. Only data that has changed is backed up.

Redshift Workload Management (WLM) allows to manage priorities within workloads, so that fast-running queries won’t get stuck in queues behind long-running queries.

By default, Redshift configures one Queue with a concurrency level of five, plus a Superuser Queue with a concurrency Level of one. The maximum concurrency Level for user-defined queues is 50.

To **migrate** an entire cluster to another Region:

* Use cross-region snapshot functionality to create a snapshot in the target region. Find more details for creating a cross-region snapshot here
* Restore the cluster from the snapshot. When you do, Amazon Redshift creates a new cluster with all the snapshot data on the new cluster. Find more details for restoring a cluster from a snapshot here.45

To **migrate** (move) specific tables:

1. Connect to the Amazon Redshift cluster in the source region and use the Unload command to export data from Amazon Redshift to Amazon S3
2. Copy your S3 data from the source region to the target region using the steps given earlier
3. Create an Amazon Redshift cluster and the required tables in the target region
4. Use the COPY command to load data from Amazon S3 to the required tables.

**CLOUDFRONT**

If a user makes an HTTP request to access a static resource on your server and the requested resource does not exist on the CloudFront server, then CF will QUERY THE ORIGIN SERVER and then CACHE THE RESOURCE on the Edge Location.

The maximum size of a file that can be delivered through CloudFront is 20 GB.

Main components are: Origins (EC2, S3 bucket, non-AWS resource, etc) and Distributions

CloudFront RTMP Distribution: Real Time Message Protocol is aimed for streaming media file. CF users Adobe Flash Media Servers as the streaming server. It accepts requests over ports 1935 and 80.

CloudFront Web Distribution: it is aimed for distributing files over HTTP or HTTPS. It can be used for example for distributing the player that CF RTMP may need to play the streamed media file.

In addition to the Edge Locations, Regional Edge Locations also exist. RELs have a larger cache width that will keep contents. Edge Locations instead, will gradually remove less popular contents to make room for more popular contents. RELs are enabled by default without your intervention.

You can use the GeoRestriction feature to whitelist a list of countries in which your users can access your content. Or you can blacklist those countries where access to your contents won’t be allowed (these last ones will receive a 403 Forbidden message).

You can use SSL with CloudFront, and you can either use a **Dedicated IP Custom SSL** or an **SNI Custom SSL**.

Cached files will be kept at the Edge Location for the time specified by the **Expiration Period**, which by default is 24 hours and can be as low as 0 seconds.

AWS CloudFront has integration with AWS Shield (to mitigate DDoS attacks) and with AWS WAF to protect web applications from SQL Injection and Cross-Site Scripting attacks.

You can configure custom Headers so that your Origin will accept requests only from CloudFront. You can even configure Custom Headers to distinguish origin requests made by each different distribution.

By using cookies, CloudFront supports dynamic content.

CloudFront supports these HTTP methods: GET, HEAD, POST, PUT, PATCH, DELETE and OPTIONS.

A CloudFront distribution can have up to 100 CNAMES.

CloudFront publishes 6 different metrics to CloudWatch at intervals of 1 minute.

CloudFront provides three different price classes according to where your content needs to be distributed. If you don’t need your content to be distributed globally, but only within certain locations such as the US and Europe, you can lower the prices you pay to deliver by choosing a price class that includes only these locations.

**Lambda@Edge** allows you to upload your Lambda functions in Node.js code and configure them to be triggered in response to CloudFront events. This will allow your apps to respond to end users at the lowest network latency. The code will be ready to be executed at every Edge Location and it will scale with the volume of requests.

**ELASTIC BEANSTALK**

AWS Elastic Beanstalk uses:

* Amazon EC2 (Amazon Linux AMI or the Windows Server 2012 R2 AMI)
* Amazon RDS
* Elastic Load Balancing
* Auto Scaling
* Amazon S3
* Amazon SNS

**When updating an application or its configuration on an Elastic Beanstalk environment**, Elastic Beanstalk gives Dev or Ops control over the rate of changes. Where possible, AWS builds and cuts over to the new service before deleting the old service. **The update starts immediately on submission of the instruction**.

You can create YAML config files for non-standard resources definitions. You can also pre-create Elastic Beanstalk environments with the standard elements.

**ELASTICSEARCH**

To ingest data into an ES domain (cluster) you can use:

* Kinesis Firehose
* Logstash
* Elasticsearch API

ES supported storage types are:

* on-instance
* EBS volumes (these can be increased and decreased), up to one 1.5 TB volume per instance associated with a domain

Maximum of datanodes per ES domain: 20 (which result in up to 30 TBs per domain). You can request to have the limit increated up to 100 datanodes (150 TB).

Automatic backups are enabled by default but they can only be restored in Full mode (complete override of the existing cluster). If you do a manual snapshot you can restore the backup into a new domain while keeping the previous one.

The scaling or resizing of an ES domain do not require downtime.

By enabling Zone Awareness, the ES nodes will be spread evenly among two AZs.

**ELASTICACHE**

In-memory cache that supports up to 15 nodes plus up to 5 read replicas, each of up to 6.1 TB of in-memory data.

Creation of Read Replicas of other Read Replicas is not supported.

EC node: the smallest component of a Redis or a Memcached cluster

Redis shard: is a subset of the cluster’s keyspace

EC cluster: all the nodes

When an ElastiCache is deployed within a VPC, it can only be accessed from within the VPC. It cannot even be accessed by nodes in the on-Premise Datacenter.

You cannot move an EC cluster from inside a VPC to the outside or viceversa.

EC Security Groups can only be used with EC clusters that are out of the VPC. For EC clusters within the VPC, access is controlled by means of VPC SGs.

EC uses subnet groups, just like RDS.

Parameter groups: acts as a “container” for engine configuration values that can be applied to one or more clusters. If no Parameter Group is defined, a default one is used.

In EC Redis, scaling down to a smaller node type is not supported.

Redis Replication can be used in conjunction with Multi-AZ. In case of failure or reboot of the Pimary node, the Read Relica with the smaller asynchronous lag will be promoted to Primary.

The parameter that should be adjusted if you find that the overhead pool is less than 50MB is **Memcached\_Connections\_Overhead**

Snapshots of EC clusters can be used to prewarm new nodes with already populated data.

Backups can only be enabled at the cluster level, and you can chose to backup the primary or a replica. Backups are only enabled for EC Redis. Memached **do not** allow backups at all.

Redis Online Cluster allows you to scale your cluster in our out by removing or adding nodes with no downtime.

Encryption in transit and at rest can be enabled on a Redis cluster, but only at the cluster creation time, and not afterwards.

You cannot use your own certs with EC Redis. Certs are provided and automatically updated as part of the service.

To **migrate** your AWS ElastiCache Redis you have to:

* Take a manual backup of the Redis cluster
* Export the backup to S3 using the ElastiCache console, the CLI or the API
* Copy the backup from the S3 in the current region to the target Region
* Restore the ElastiCache cluster from the backup, in the target Region

**AWS DIRECTORY SERVICES**

It has two modes: AD Connector (to connect to an existing AD) and Simple AD (to create a new and very limited and compliant Samba-based AD.

**OPSWORKS**

AWS OpsWorks is a configuration management service that provides managed instances of Chef and Puppet.

Chef and Puppet are automation platforms that allow you to use code to automate the configurations of your servers.

AWS CodeDeploy is a fully managed deployment service that automates software deployments to a variety of compute services such as Amazon EC2, AWS Fargate, AWS Lambda, and your on-premises servers.

Chef turns infrastructure as code (IaC) that allows to deploy and configure your infrastructure quickly.

OpsWorks elements are:

* Stack: Group of resources (ELB, EC2 instances, RDS instances, etc.)
* Layers: They exist within a Stack (like: web application, DB, Load Balancers or Caching layers). When a Layer is created, OpsWorks handles it for you
* Instances (24x7) / Time-based instances / Load-based instances
* Apps: loaded into GitHub, S3, etc.
* Deployments
* Resources
* Permissions

By Default you can have up to 1.000 OpsWorks clusters per region per account.

OpsWorks lifecycle events (each has its own custom recipes)

* Setup
* Configure
* Deploy
* Undeploy
* Shutdown

**KINESIS**

Kinesis allows to process or analyze streaming data for specialized needs. This means that for any kind of scenario where you are streaming large amounts of data that need to be processed quickly, Kinesis will be the option to chose.

Kinesis stores your data for up to 24 hours by default. The retention period can be raised up to 7 days by enabling extended data retention. After that, the data is gone because **Kinesis is not persistent**. If you need persistence, you need to inject the information into S3, DynamoDB, Redshift or EMR.

Data Producers: EC2 instances that have the Java Kinesis client installed on them. There is a Kinesis Client that can be installed on Linux to gather data from the Linux server itself.

Shard is a measure of capacity for Kinesis. It is the base unit of an Amazon Kinesis data stream. One shard provides a **throughput** of 1MB/sec data input and 2Mb/sec data output, and supports up to 1.000 puts (writes) per second. Shards can scale in number with no limits, but by default 10 Shards can be provisioned per Region by default. You can request this limit to be increased.

Record is the unit of data stored in a Kinesis data stream. It is composed of sequence number, partition key and data blob. The maximum size of a data blob is 1MB.

Data Consumers: Kinesis Streams Applications.

Kinesis supports two encryption methods:

* server side encryption: once enabled it is handled automatically as part of the service with AWS KMS
* client side encyption: you can encrypt and descript the data in the client end, and store it in Kinesis already encrypted

**ELASTIC MAP REDUCE**

Allows to process vast amounts of data by using a hosted Hadoop framework running on a web-scale infrastructure of EC2 and S3.

CloudWatch metrics for EMR are updated every five minutes and automatically collected and pushed to CloudWatch for every EMR cluster. This interval is not configurable.

EMR Cluster Node types are:

* Master node: manages the cluster and typically runs master components of distributed applications
* Core nodes: managed by the master node. Core nodes run the Data Node daemon to coordinate data storage as part of the Hadoop Distributed File System (HDFS)
* Task nodes: these are optional and you can use them to add power to perform parallel computation tasks on data, such as Hadoop MapReduce tasks and Spark executors

Once the cluster finishes its processing, it is terminated so that you don’t pay for resources that are no longer necessary. But you can launch your cluster with the -alive flag, which will make it persist until you explicitly terminate it.

By Default you are limited to 20 instances across all of your EMR clusters, but you can place a request to have that limit increased.

EMR runs on EC2 instances built from an Amazon Linux AMI optimized for and configured with Hadoop.

Charges begin when, having requested the launch of the cluster (which has a 10 minute time mark), 90% of the instances are ready to start its processing. Instances could have been booted or checked in (in case they were already running). Charges then are applied per second with a minimum of one minute.

EMR clusters are launched in 2 Security Groups, one for the Master (which connects with the service on the AWS end) and the other for the Slaves. Neither of these SGs can be accessed by other EC2 instances, but you can reconfigure them at will. You can also chose to send your data to EC2 instances or S3 buckets, and in those cases the communication is always established on HTTPS.

EMR relies on the launch of EC2 instances to pull the data stored in S3. Then it processes it and finally places the output into another S3 bucket for the customer to grab it or to serve as input of another EMR cluster.

If the Master of an EMR cluster fails, a message is displayed in the AWS console stating that “The Master node was terminated”. Amazon EMR does not support automatic failover of the Master node, which means that a new cluster has to be started.

If a Slave goes down though, a new Slave node is provisioned to replaced the failed one.

You can SSH into your Master node (on into your Slaves but from your Master) and issue Hadoop commands directly from there.

You can place bootstrap scripts in an S3 bucket and reference them when launching an EMR cluster, to have specific configurations set before the cluster starts processing.

EMR supports Hive, Impala, Pig and Hbase.

EMR also has a built-in Kinesis connector that will allow to easily read from a Kinesis Stream.

To **migrate** an EMR cluster, the entire cluster must be recreated in the target Region. Migration of the data itself, will depend on whether it is stored in S3 or in a Hadoop Distributed File System (HDFS).

**VM IMPORT/EXPORT**

VM Import/Export enables you to import VM images from your existing virtualization environments (either as AMIs or EC2 instances) and then export them back.

For most import needs, it is recommended to use the AWS Server Migration Service (SMS), because it automates the conversion of existing VMs into ready to use AMIs.

VM Import/Export has a connector for VMware that allows you to export your VMware instances and import them directly as EC2 instances,

Encryption is optional for **imports** only.

When using AWS Import/Export to export from a versioned S3 bucket, only the most recent version of each object will be exported.

Import/Export can **export** from S3 only, but it can **import** into S3, EBS and Glacier.

**CLOUDFORMATION**

Templates (no limit), Stacks (200), Nested Stacks and Stack Sets (20).

The **Fn** **GetAtt** will Bring all the details about the resources have just been provisioned with CF.

The **WaitCondition** instructs a CF stack to wait while it processes, to allow applications to be Provisioned.

With the **DeletionPolicy** attribute you can preserve or (in some cases) backup a resource when its stack is deleted. You specify a DeletionPolicy attribute for each resource that you want to control. If a resource has no DeletionPolicy attribute, AWS CloudFormation deletes the resource by default. DeletionPolicy options are:

* Delete
* Retain
* Snapshot

You can have a maximum of 60 parameters in an AWS CloudFormation template. Each parameter must be given a logical name (also called logical ID).

Limits: <https://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/cloudformation-limits.html>

You can use the **Mappings** section to reuse your templates in other Regions, by changing the mapping declarations to substitute region-specific information, such as the unique IDs for AMIs.

CloudFormation does have support for **Chef** and **Puppet**, which allows you to provision your application layer. It also has the ability to utilise the **bootstrap scripts** for EC2 instance provisioning.

CloudFormation **will automatically work out the provisioning order** based on dependencies for each resource.

JSON

{

  "AWSTemplateFormatVersion" : "version date",

  "Description" : "JSON string",

  "Metadata" : {

    template metadata

  },

  "Parameters" : {

    set of parameters

  },

  "Mappings" : {

    set of mappings

  },

  "Conditions" : {

    set of conditions

  },

  "Transform" : {

    set of transforms

  },

  "Resources" : {

    set of resources

  },

  "Outputs" : {

    set of outputs

  }

}

YAML

AWSTemplateFormatVersion: "version date"

Description:

  String

Metadata:

  template metadata

Parameters:  —  60

  set of parameters

Mappings:  —  100

  set of mappings

Conditions:

  set of conditions

Transform:

  set of transforms

Resources:  —  200

  set of resources

Outputs:  —  200

  set of outputs

**CLOUDFORMER**

CloudFormer is a beta tool that creates CloudFormation Templates from existing AWS resources in your account.

CloudFormer consist of a CloudFormation Stack from which you create the CloudFormation Template.

Once you create the CloudFormation Template, you should consider deleting the Stack to avoid incurring in unnecessary charges on your AWS account.

**CLOUDHSM**

AWS CloudHSM will help you meet corporate, contractual and regulatory compliance requirements for data security by using Dedicated Hardware Security Module (HSM) instances within the AWS cloud.

It used to cost 5K per monthm, but now this changed and it is charged by the hour (about 120/hour).

It allows you to securely generate, store and manage cryptographic keys used for data encryption in a way that keys are accessible only by you.

**CLOUDSEARCH**

CloudSearch is a fully managed service that allows you to easily setup, manage and scale a search solution for your website or application.

It is built on top of self-healing clusters, highly available with Multi-AZ. You simply create a search domain and upload the data you want to make searchable, and Amazon CloudSearch will automatically provision the necessary resources and deploy a highly tuned search index.

**WAF, AWS Shield/Inspector/Trusted Advisor**

AWS WAF is tightly integrated with Amazon CloudFront and the Application Load Balancer (ALB). Rules can Allow, Block or Monitor (count).

Rate-based Rules: If an IP address breaches the configured limit over a 5 minute period, new requests will be blocked until the request rate falls below the configured threshold. Managed Rules are pre-configured rules from AWS Marketplace security sellers.

AWS Shield helps protect agains DDoS. There are two tiers: Standard (free) and Advanced (3K/month).

The technologies that can be used to mitigate a DDoS attack are:

* CloudFront
* Route53
* ELBs
* WAFs
* Autoscaling (for both, WAFs and Web Servers)
* CloudWatch

AWS Inspector is agent-based and needs to be installed on top of the client OS of an EC2 instance. Amazon Inspector is an automated security assessment service that helps improve the security and compliance of applications deployed on AWS.

AWS Trusted Advisor advises on Security, Cost, Performance and Fault Tolerance. It has a free version, but for Business and Enterprise accounts, it is more powerful. When it comes to Security, Trusted Advisor primarily focuses on IAM. To unlock all the Trusted Advisor recommendations you need to have a Business or Enterprise subscription.

**EFS**

There are two different performance modes available for Amazon Elastic File System: General Purpose and Max I/O. General Purpose performance mode is the default mode and is appropriate for most file systems. However, if your overall Amazon EFS workload will exceed 7,000 file operations per second per file system, we recommend the files system use Max I/O performance mode. Max I/O performance mode is optimized for applications where tens, hundreds, or thousands of EC2 instances are accessing the file system.

There are three levels of access control to consider when planning your EFS file system security: IAM permissions for API calls; security groups for EC2 instances and mount targets; and Network File System-level users, groups, and permissions.

**ECS**

Kubernetes is an orchestrator for Docker.

OpenShift is also an orchestrator for Docker, but it runs either on RedHat or CentOS (OKD, which is the free version).

ECS has two modes: Fargate launch type and EC2 launch type.

Fargate is a compute engine for Amazon ECS.

With Fargate launch type, all you have to do is package your application in containers, specify the CPU and memory requirements, define networking and IAM policies, and launch the application. EC2 launch type allows you to have server-level, more granular control over the infrastructure that runs your container applications. With EC2 launch type, you can use Amazon ECS to manage a cluster of servers and schedule placement of containers on the servers.

Amazon Elastic Container Service (Amazon ECS) is a highly scalable, fast, container management service that makes it easy to run, stop, and manage Docker containers on a cluster. You can host your cluster on a serverless infrastructure that is managed by Amazon ECS by launching your services or tasks using the Fargate launch type. For more control you can host your tasks on a cluster of Amazon Elastic Compute Cloud (Amazon EC2) instances that you manage by using the EC2 launch type.

The Fargate launch type only supports using container images hosted in Amazon ECR or pub.

Amazon Elastic Container Registry (Amazon ECR) is a managed AWS Docker registry service that is secure, scalable, and reliable. Amazon ECR supports private Docker repositories with resource-based permissions using IAM so that specific users or EC2 instances can access repositories and images.

To prepare your application to run on Amazon ECS, you create a task definition. The task definition is a text file, in JSON format, that describes one or more containers, up to a maximum of ten, that form your application.

A task is the instantiation of a task definition within a cluster. After you have created a task definition for your application within Amazon ECS, you can specify the number of tasks that will run on your cluster. The Amazon ECS task scheduler is responsible for placing tasks within your cluster.

When you run tasks using Amazon ECS, you place them on a cluster, which is a logical grouping of resources. If you use the Fargate launch type with tasks within your cluster, Amazon ECS manages your cluster resources. If you use the EC2 launch type, then your clusters will be a group of container instances you manage. Amazon ECS downloads your container images from a registry that you specify, and runs those images within your cluster.

The container agent runs on each infrastructure resource within an Amazon ECS cluster. It sends information about the resource's current running tasks and resource utilization to Amazon ECS, and starts and stops tasks whenever it receives a request from Amazon ECS.

Which of the following statements are true about Docker containers on AWS?

* To be able to use ECS, you must use the ECS Agent
* ECS allows you to control the scheduling and placement of your containers and tasks
* You can install and manage Kubernetes on AWS, yourself
* ECR can be used to store Docker images (Elastic Container Registry allows to manage your own Docker Container images)

ECS integrates with ALB, because it allows to specify a dynamic port that will be used by the different containers.

MS link that explains what containers are: <https://docs.microsoft.com/en-us/virtualization/windowscontainers/about>

**IAM**

IAM has Users, Groups, Roles and Policies.

When logging with a Web Identity, the AssumeRoleWithWebIdentity is invoked.

When logging with a Federated identity, the AssumeRoleWithSAML is invoked.

The sign in endpoint for SAML-based authentication is <https://signin.aws.amazon.com/saml>.

A successful call to GetFederationToken returns: AccessKeyId, SecretAccessKey, SessionToken and Expiration Further information: <https://docs.aws.amazon.com/cli/latest/reference/sts/get-federation-token.html>

AWS Policy Generator:

<https://awspolicygen.s3.amazonaws.com/policygen.html>

A permissions boundary is an advanced feature in which you use a managed policy to set the maximum permissions that an identity-based policy can grant to an IAM entity. When you set a permissions boundary for an entity, the entity can perform only the actions that are allowed by both its identity-based policies and its permissions boundaries. Further information: [https://docs.aws.amazon.com/IAM/latest/UserGuide/access\_policies\_boundaries.htmlhttps://docs.aws.amazon.com/organizations/latest/userguide/orgs\_manage\_policies\_scp.html](https://docs.aws.amazon.com/IAM/latest/UserGuide/access_policies_boundaries.htmlhttps:/docs.aws.amazon.com/organizations/latest/userguide/orgs_manage_policies_scp.html)

**SQS**

SQS ChangeMessageVisibility: the default visibility time-out is 30 seconds and can increased up to 12 hous. This API call can be used to extend the length of time to process the jobs.

By setting the WaitTimeSeconds value, when the consumer instance polls for new work, the SQS service will allow it to wait a certain time for one or more messages to be available before closing the connection.

Maximum Retention Period for an SQS Message is 14 days and minimum is 1 minute, the default is 4 days.

An SQS Message can be as large as 256 KB.

Amazon SQS supports dead-letter queues, which other queues (source queues) can target for messages that can't be processed (consumed) successfully.

Delay queues let you postpone the delivery of new messages to a queue for a number of seconds. If you create a delay queue, any messages that you send to the queue remain invisible to consumers for the duration of the delay period. The minimum delay for a queue is 0 seconds. The maximum is 15 minutes. Delay queues are similar to visibility timeouts because both features make messages unavailable to consumers for a specific period of time. The difference between the two is that, for delay queues, a *message is hidden when it is first added to queue*, whereas for visibility timeouts a *message is hidden only after it is consumed from the queue*.

Amazon SQS queues exist per region. To **migrate** the data in a queue, you need to drain the queue from the source region and insert it into a new queue in the target region.

**SWF**

The maximum number of simultaneous workflow executions is 100K.

**DATA PIPELINE**

It is an ETL that executes activities such as copying data between data stores, or scheduling chained transforms.

Pipeline: Is the Data Pipeline resource that contains the definition of the dependent chain of data sources, destinations and predefined or custom data processing activities.

Data Node: is the representation of your business data. for example, an S3 path.

Activity: is the action that Data Pipeline initiates on your behalf.

Precondition: must be met before any activities are launched.

Schedule: duh!...

As part of its execution, Data Pipeline can execute SQL queries.

**SUPPORT**

AWS Support Plans: Basic, Developer, Business, Enterprise

* General Guidance: 24 hours
* System impaired: 12 hours
* Production system impaired: 4 hours
* Production system down: 1 hour
* Business-critical system: 15 minutes (Enterprise)

**AWS SERVICE CATALOG**

AWS Service Catalog allows IT Administrators to create, manage and distribute **Catalogs** of approved products to end users, who can then access the products they need in a personalized portal.

You can also create **Portfolios**, that are collections of products, with configuration information that determine who can use those products and how they can use them.

**Products** are created from CloudFormation templates.

You can also subscribe to a Product in the AWS Marketplace and use te *Copy to Service Catalog* action.

**API**

The steps required to **migrate** Amazon API Gateway from the source region to the target region:

1. Export the API from the API Gateway into a Swagger file using the API Gateway Export API
2. Copy the Swagger file to the target region using standard tools like CP, FTP, or rsynch
3. Import the Swagger file to create the API in the API Gateway in the target region

**EC2**

C5 and M5 EC2 instances use the Nitro hypervisor instead of the Xen hypervisor.

EBS boot (root) devices must be SSD, which means that HDD disks can’t be used for an OS disk.

Burst IOPS is limited to 3.000 IOPS. If you need more IOPS than that, you must provision volumes that will provide you with the necessary IOPS.

Initialization of a volume: when restoring a volume from snapshot, all blocks that have data should be read. In Linux volumes this can be done with dd or fio.

IOPS Read and Write operations are processed in 256K increases linearly.

The queue depth is the number of pending I/O requests from your application to your volume. For maximum consistency, a Provisioned IOPS volume must

maintain an average queue depth (rounded to the nearest whole number) of one for

every 500 provisioned IOPS in a minute. For example, for a volume provisioned with

1500 IOPS, the queue depth average must be 3.

HDD-backed volumes process reads and writes in I/O sizes of 1MB. Sequential I/Os are merged and processed as 1 MB units while each non-sequential I/O is processed as 1MB even if the actual I/O size is smaller.

Linux AMIs require GPT partition tables and GRUB 2 for boot volumes 2 TiB.

(2048 GiB) or larger. Many Linux AMIs today use the MBR partitioning scheme, which only supports up to 2047 GiB boot volumes. If your instance does not boot with a boot volume that is 2 TiB or larger, the AMI you are using may be limited to a 2047 GiB boot volume size. Non-boot volumes do not have this limitation on Linux instances.

EBS backed root device volumes can be up to 1 or 2 TB.

Instance Store backed root device volumes can have a maximum size of 10 GB.

Magnetic Standard volumes go from 1GiB to 1TiB and deliver ~100 IOPS.

GP2 (SSD) volumes go from 1GB to 16TB. IO1 (PIOPS) volumes go from 4GB to 16TB.

The ratio for IO1 volumes is 50:1. For example, a 100GB volume can be provisioned with up to 5.000 IOPS.

Recommendation though, is an IOPS-to-GiB greater than 2:1. For example a 2.000 IOPS volume should be smaller than 1.000 GiB.

The uptime SLA for EC2 and EBS within a region is 99.95%

When a new Security Group is created, by default all inbound traffic is DENIED and all outbound traffic is ALLOWED.

An instance's Public IP address is not managed on the instance: It is, instead, an alias applied as a network address translation of the Private IP address.

Placement Groups can NOT span different Availability Zones, but it can span different subnets. There are Cluster PGs (within an AZ) and Spread PGs (different underlying HW and AZs).

When copying an AMI, the following types of information must be manually copied

* S3 Bucket permissions
* Launch permissions
* User-defined tags

Multipart upload provides these benefits:

* Delivers quick recovery from network issues
* Delivers improved throughput
* Delivers the ability to begin an upload before you know the final object size
* Delivers the ability to pause and resume object uploads
* Can be used with S3 Transfer Acceleration

**NETWORKING**

Load Balancer healthcheck: If you are designing to check for loss of contact with the instances you need to use "Evaluate Target Health" to confirm connectivity. The Latency policy will eventually detect the unavailability, however it is not a real time test.

By Default, you can have up to 20 Load Balancers per account, being this a combination of Classic and ALBs. You can raise a ticket to have this limit raised.

For ALBs, in the listeners section you have to define Target Groups.

All subnets in a Default VPC are Internet accessible.

There are at least 2 Availability Zones per AWS Region.

By Default, you can have up to 200 subnets in a Region, but you can raise a ticket with AWS to have this limit increased.

Direct connect: 1Gbps and 10Gbps ports are available. [50/100/200](tel:50%2F100%2F200)/300/400/500 Mbps can be ordered from any APN partners supporting AWS Direct Connect.

Direct Connect Link Aggregation Groups are a way for customers to order and manage multiple direct connect ports as a single larger connection instead. Maximum number of links is 4x.

You can block an IP address with a Network ACL but not with a Security Group. SGs allow traffic but can’t deny it.

VPC Peering: 1 VPC can be peered with up to 50 other VPCs at one single time. This is a soft limit and can be increased to 125. VPC peering is not transitive between VPCs peered with a common VPC.

Inter-region VPC peering can be accomplished as well. Just as VPCs from different accounts can be peered.

Each subnet can be associated with only 1 ACL and Route Table at a time. Every time a subnet is added, it is by default associated with the Default Route Table and with the Default network ACL unless other setting is specified.

The Default Network ACL by default allows all inbound and outbound traffic. Any Custom Network ACL by default denies all inbound and outbound traffic until a rule is added.

When configuring subnets that will have to be publicly accessible, you can enable the “auto-assign public IPv4 address” setting within the properties of the subnet, and that will allow each EC2 instance deployed in that subnet to receive a Public IP by default.

VCP Endpoint (vpce-xxxxx): it is an ENI that servers as an entry point for traffic destined to a service (for example: S3).

Only one virtual private gateway can be attached to a VPC at a time.

VPC limits: <https://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/VPC_Appendix_Limits.html>

VPC Endpoints sends traffic between the VPC and the given service being accessed, to be handled by the AWS Network, without the need of an Internet Gateway, a NAT Gateway or a VPN Connection.

On December 2018 AWS announced that VPCs can now be shared across AWS accounts.

The Internet Gateway enables outgoing access on Public Subnets as well as incoming traffic to those EC2 instances, from Public Internet. You can only have one per VPC.

NAT Instances and NAT Gateways enable outgoing traffic from EC2 instances to Public Internet through an Internet Gateway.

NAT Instances require to have the "Source/Destination Checks” disabled.

NAT Gateways scale automatically up to 10Gbps and are not associated to Security Groups.

Having only one NAT GW on a single AZ is not good enough and you should consider deploying one in each AZ for redundancy.

To establish a successful site-to-site VPN connection from your on-premise network to an AWS Virtual Private Cloud, you must have:

* a VPC with Hardware VPN Access
* an on-premise Customer Gateway
* a Virtual Private Gateway to make the VPN connection work

Route53 do NOT use Key Value pairs and uses ALIAS records. Routing policies are:

* Simple: Round Robin
* Failover: Use when you want to configure active-passive failover
* Geolocation: Use when you want to route traffic based on the location of your users
* Geoproximity: Use when you want to route traffic based on the location of your resources and, optionally, shift traffic from resources in one location to resources in another
* Latency: Use when you have resources in multiple locations and you want to route traffic to the resource that provides the lowest latency for the end user
* Mutivalue answer: Use when you want Route 53 to respond to DNS queries with up to eight healthy records selected at random
* Weighted: Use to route traffic to multiple resources in proportions that you specify over the length of a day (for example: 20% to one zone and 80% to another)

Traffic that is NOT monitored by VPC Flow Logs:

* Traffic generated to the Amazon DNS server (if you use a custom DNS server, then that traffic WILL be logged)
* Traffic generated by a Windows instance for Amazon Windows license activation
* Traffic to and from 169.254.169.254 for instance metadata
* DHCP traffic
* Traffic to the reserved IP address for the default VPC Router

You cannot enable flow logs for VPCs that are peered with your VPC unless they are in the same account, and you cannot change the settings of a VPC flow log once it is created.

**MISC**

Tag Editor can be used to find resources within an AWS account that do not have tags applied to them.

Which key can be used to enforce an MFA authentication requirement in a bucket policy? aws:MultiFactorAuthAge

When using Auto-Scaling you can schedule an Auto-Scaling activity 30 days in advance.

DR Options:

* Backup and Restore: have prebuilt AMIs; also backup to and restore from S3 with bootstrap scripts
* Pilot Light: use pre-allocated IP addresses and/or pre-allocated ENIs, and/or use ELBs to distribute traffic, while also using R53 for DNS resolution. Then consider automating the provisioning of the rest of the infrastructure by means of CloudFormation
* Warm Standby: Keep a standby environment active in the cloud to failover to it in case necessary, but changing the DNS R53 records either manually or automatically based on health checks. The environment can have scaled-down resources that will be scaled up when becoming active
* Multi-Site (active-active): solution that runs on both, the on-prem datacenter and the AWS cloud, in an active-active way, that can have traffic sent either symmetrically or asymmetrically to both sites via R53 or base on weight. You can also use Auto-scaling for EC2 reources

Elasticity: allows to scale out and scale in during short periods of time (auto-scaling).

Scalability: allows you to scale out during long periods of time (change the instance type to a bigger one).

AWS Quick Starts: [https://aws.amazon.com/quickstart](https://aws.amazon.com/quickstart/). Quick Starts are built by AWS solutions architects and partners to help you deploy popular solutions on AWS, based on AWS best practices for security and high availability.

The AWS Concierge will answer questions about billing and the overall AWS account.

AWS Assurance Program: components are “Certifications/Attestations” and “Compliance with Laws and Regulations"

There are three fundamental drivers of cost with AWS: compute, storage, and outbound data transfer. These characteristics vary somewhat, depending on the AWS product and pricing model you choose.

Services with root access credentials:

* EC2
* EMR
* OpsWork
* Elastic Beanstalk

Recovery time objective (RTO): The time it takes after a disruption to restore a business process to its service level, as defined by the operational level agreement (OLA).

Recovery point objective (RPO): The acceptable amount of data loss measured in time.

AWS Cost Management:

AWS Cost Explorer: Cost Explorer is a tool that allows you to view and analyze your costs. You can view data for up to the last 13 months, forecast how much you are likely to spend for the next three months, and get recommendations for what Reserved Instances to purchase. You can use Cost Explorer to see patterns in how much you spend on AWS resources over time, identify areas that need further inquiry, and see trends that you can use to understand your costs. AWS Cost Explorer lets you explore your AWS costs and usage using a number of filtering dimensions (AWS Service, Region, Linked Account, etc.)

When you have an existing reserved instances allocation and you submit a modification request to split the footprint across multiple AZs, new Reserved Instance requests are created in each AZ to match the new footprint requirements.

AWS Budgets: Using AWS Budgets, you can set a budget that alerts you when you exceed (or are forecasted to exceed) your budgeted cost or usage amount. AWS Budgets gives you access to a number of filtering dimensions (i.e., AWS Service, Availability Zone, and Linked Account), and allows you to create budgets that are tracked on a monthly, quarterly, or yearly cadence.

AWS TCO Calculator: [https://aws.amazon.com/tco-calculator](https://aws.amazon.com/tco-calculator/). Our TCO calculators allow you to estimate the cost savings when using AWS and provide a detailed set of reports that can be used in executive presentations.

List of actions (API) by function: <https://docs.aws.amazon.com/AWSEC2/latest/APIReference/OperationList-query.html>

The pillars of the AWS Well Architected Framework are: Cost Optimization, Reliability, Efficiency, Performance and Security (**CREPS**).

PHP Composer and PHP SDK to access AWS resources through SDK scripts in PHP.

The default region for an SDK is "US-EAST-1”.

Available SDKs:

* Android, iOS, JavaScript (Browser)
* Java
* .Net
* Node.js
* PHP
* Python
* Ruby
* Go
* C++

HTTP Status Codes:

* 1xx = Informational (the request was received and understood)
* 2xx = Success (indicates the action requested by the client was received)
* 3xx = Redirects (indicates the client must take additional action to complete the request)
* 4xx = Client errors (situations in which the error seems to have been caused by the client)
* 5xx = Server errors (the server is aware that it has encountered an error or is otherwise incapable of performing the request)

Well-Architected Framework with updated content (all whitepapers have been updated):

[https://aws.amazon.com/architecture/well-architected/](https://urldefense.proofpoint.com/v2/url?u=https-3A__aws.amazon.com_architecture_well-2Darchitected_&d=DwMGaQ&c=eIGjsITfXP_y-DLLX0uEHXJvU8nOHrUK8IrwNKOtkVU&r=2W_Jx7xCk9y6_yOzqor73n4GqsC7HEGWifcFBYjPk02fzPy8k-15F-eCK66ZMV7Q&m=hXVRTJz5nyuwUcGcRlbH7R1_bZm4TXlHVxZAoMmoCsg&s=Wf12IVT7osSIKC2CGOQaBH_QVNSIh2s0SaX-rUW-TkY&e=)