AWS CSA-A Handbook

A Guide for Pursuing and Existing Amazon Web Services
Certified Solutions Architects at the Associate Level

BY
LUCAS C. PICHETTE

INDEX

Figures

Page numbers not finalized

AWS CSA - Associate Exam Information

Time Restraint:

- ▹ 130 minutes in length

Content:

- ▹ About 65 Questions; 60% relating to designing solutions, 30% relating to security and troubleshooting, and 10% relating to deployment and implementation. *These are not guaranteed figures.*
- ▹ Entirely multiple choice
- ▹ Scenario-based questions

Passing Condition:

- ▹ Results are within scores of 100-1000, and a passing score is at least 720.

Lifespan of Certification:

- ▹ Qualification is valid for 3 years

Entry Fee:

- ▹ $150

Tips

Obtaining the AWS CSA - Associate certification does not guarantee a job; while the certification helps, you should also find other ways to bolster your (non-physical, but good hygiene and proper attire helps) attractiveness as an applicant. Create personal projects to solidify learned topics and show employers that you can turn concepts into products. Additionally, an extra resource I have composed for those who like flash cards is a Quizlet folder with multiple sets.

Why is the Public Cloud so Powerful?

Public cloud allows organizations to try out new ideas, new approaches and experiment with little upfront commitment. If it doesn't work out, organizations have the ability to terminate the resources and stop paying for them.

The Well-Architected Framework

The Well-Architected Framework is a guide that proposes a set of questions that you can use to evaluate how well your architecture is aligned to AWS practices. It will help you design architectures that can achieve (known as the "5 pillars"):

- ▹ Operational Excellence
- ▹ Security
- ▹ Reliability
- ▹ Performance Efficiency
- ▹ Cost Optimization

AMAZON SERVICES

AVAILABILITY

GLOBAL INFRASTRUCTURE

Consists of regions, Availability Zones, and Edgepoints.

REGION

Consists of a multitude of availability zones (AZs).

AVAILABILITY ZONE

One or more discrete centers, each with redundant power, networking and connectivity, housed in separate facilities.

EDGEPOINTS

Used for caching content. Typically this consists of CloudFront, Amazon's content delivery network (CDN). NOT a region. Always more Edgepoints/edgelocations than regions.

STORAGE

ELASTIC BLOCK STORAGE (EBS)

Description:

- ▹

Provides:

- ▹ Persistent block storage volumes for use with Amazon EC2 instances in the AWS Cloud.

- ▹ Each Amazon EBS volume is automatically replicated within its Availability Zone to protect you from component failure, offering high availability and durability.

5 Types:

- ▹ General Purpose (SSD; Solid State Drive)
    - ▸ Most workloads
    - ▸ API Name: gp2
    - ▸ Volume Size: 1 GiB - 16 TiB
    - ▸ Max. IOPS**/Volume: 16,000
- ▹ Provisioned IOPS (SSD)
    - ▸ Databases
    - ▸ API Name: io1
    - ▸ Volume Size: 4 GiB - 16 TiB
    - ▸ Max. IOPS**/Volume: 64,000
- ▹ Throughput Optimised Hard Disk Drive
    - ▸ Big Data & Data Warehouses
    - ▸ API Name: st1
    - ▸ Volume Size: 500 GiB - 16 TiB
    - ▸ Max. IOPS**/Volume: 500
- ▹ Cold Hard Disk Drive (HDD)
    - ▸ File Servers
    - ▸ API Name: sc1
    - ▸ Volume Size: 500 GiB - 16 TiB
    - ▸ Max. IOPS**/Volume: 250
- ▹ Magnetic (HDD)
    - ▸ Workloads where data is infrequently accessed
    - ▸ API Name: Standard
    - ▸ Volume Size: 1 GiB - 1 TiB
    - ▸ Max. IOPS**/Volume: 40-200

Notes:

- ▹ Whatever availability zone your instance is in, the volume for the instance will also exist there.
- ▹ Changing device volume to another availability zone: Go to actions, create new snapshot. Now, go to your snapshots, select the snapshot you created, go to actions, create a new image. For more instance types, make sure that when you create this EBS snapshot image, you change virtualization type to Hardware-assisted virtualization. Then, just change the subnet when configuring instance details (step 3) to change your availability zone of the image.

▹ When you terminate an EC2 instance by default the root device volume will also be terminated; however, additional volumes attached to that EC2 instance will persist.

ELASTIC FILE SYSTEM (EFS)

Description:

▹ A file storage service for Amazon Elastic Compute Cloud (Amazon EC2) instances. Amazon EFS is easy to use and provides a simple interface that allows you to create and configure file systems quickly and easily. With Amazon 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.

Notes:

▹ Different from EBS. There can only be one EBS per Instance, however, there can be multiple EFSs per Instance.

▹ Supports the Network File System (NFS) version 4 (NFSv4) protocol.

▹ You only pay for the storage you use (no pre-provisioning required).

▹ Can scale up to the petabytes

▹ Can support thousands of concurrent NFS connections

▹ Data is stored across multiple AZs within a region

▹ Read After Write Consistency

▹ Can support thousands of connections

SIMPLE STORAGE SERVICE (S3):

Important Tidbits:

▹ A safe place to store your files.

▹ S3 is a universal namespace, so buckets must be unique names. Example: https://s3-eu-west-1.amazonaws.com/lucas-pichette

▹ Object-based storage (JSON); Consists of a key, value, and version ID. (Key is the name of the object, value is the bytes associated with the name, and the version is literally just the version. The Metadata is just info about the data you're storing).

▹ You can upload 0 Bytes to 5 TB at once.

▹ First byte latency: How quickly you will be able to access/retrieve your data.

Promises the following features:

▹ Tiered Storage Available

▹ Lifecycle Management

▹ Versioning

▹ Encryption

▹ MFA Delete (Multi-factor-authentication for deleting objects)

▹ Securing data using Access Control Lists (ACL) and Bucket Policies (BP)

- ▹ 100 buckets per account by default
- ▹ There is unlimited storage in S3
- ▹ S3 is not suitable to install an operating system on.
- ▹ S3 cannot be used to host a database.
- ▹ Remember to turn on the MFA delete to protect your objects.
- ▹ An object is composed of: Key, value, version ID, metadata, and subresources (Access Control Lists, Torrent).
- ▹ Consistency Model: Read after Write consistency for PUTS of new objects. Could get Eventual Consistency for overwrite PUTS and DELETES.

Storage Classes:

- ▹ S3 Standard
  - ▸ 99.99% availability
  - ▸ 99.999999999% durability (11 9's, the "11 9s")
  - ▸ Stored redundantly across multiple devices in multiple facilities, and is designed to sustain the loss of 2 facilities concurrently
- ▹ S3 - RRS (Reduced Redundancy Storage)
- ▹ 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
- ▹ S3 One Zone - IA
  - ▸ For a really low-cost option for infrequently accessed data
  - ▸ For when you don't require the multiple availability zone data resilience.
  - ▸ 99.50% Availability
- ▹ S3 - Intelligent Tiering
  - ▸ Designed to optimize costs by automatically moving data to the most cost-effective access tier, without performing impact or operational overhead.
- ▹ S3 - Glacier
  - ▸ For data archiving. (Data you need to hold on to but don't need to access often or soon when needed; Retrieval times can take minutes to hours depending upon how you configure it)
- ▹ S3 - Glacier Deep Archive
  - ▸ Amazon's S3 lowest-cost storage class where retrieval time of around 12 hours is acceptable.
  - ▸ You'll get a 200 HTML code when you successfully upload

Billing method for S3:

- ▹ Storage
- ▹ Requests
- ▹ Storage Management Pricing

- Data Transfer Pricing
- Transfer Acceleration
    - Uploads files to edge locations, then goes through amazon's backbone service, and really increases your upload time. (costs more of course)
- Cross Region Replication Pricing
    - Bucket in US-East, and you want to replicate your objects in this Bucket to a Bucket in AUS-Sydney, then you will be feed more

Encryption:

- Encryption In Transit: SSL/TLS
- Encryption at Rest (Server Side) is achieved by:
    - S3 Managed Keys - SSE-S3 (Server Side Encryption-S3)
    - AWS Key Management Service, Managed Kyes-SSE-KMS
    - Server Side Encryption with Customer Provided Keys-SSE-C
- Client Side Encryption
- Buckets are private by default. You can change permissions for individual files, whole objects, or categories in a bucket.
- Once Versioning is enabled on a bucket it can't be disabled, it can only be suspended.
- Versioning integrates with Lifecycle rules.
- Versioning's MFA Delete capability
- If you delete an object in a bucket with versioning enabled, the file won't be deleted and there will be a "deleted" marker that shows up if you click on the "Show" button in your bucket GUI in S3. The way to restore the file/object is to delete the "deleted" marker. You can completely delete a file by selecting individual versions.

Lifecycle Management:

- Automates moving your objects between different storage tiers.
- Can be used in conjunction with versioning
- Can be applied to current and previous versions

Cross Region Replication:

- Requires versioning to be enabled.
- Cannot replicate to the region you're in
- Can change the storage class and ownership when replicating
- Replicates permissions and tags
- Can't see other files/versions in the new replication. Need to update files in the original bucket to update files in replicated version. (Only bucket itself is replicated)
- Adding a delete marker in one replicated bucket won't add a delete marker in another. Actually deleting a file from a bucket does not actually delete or add a delete marker in any other replication.

Transfer Acceleration:

▹ Utilizes the CloudFront Edge Network to accelerate uploads to S3. Instead of uploading directly to S3 bucket, you can use a distinct URL to upload directly to an edge location which will then transfer that file using Amazon’s backbone network to S3. You will get a distinct URL to upload to: lucas-pichette.s3-accelerate.amazonlaws.com

COMPUTING

ELASTIC COMPUTE CLOUD (EC2)

Description:

▹ Web service that provides resizable compute capacity in the cloud.
▹ Virtual Machines in the cloud

Provides:

▹ Reduces the time required to obtain and boot new server instances to minutes.

Features:

▹ Root volume can be encrypted, even after initially not.
▹ Making a change to a security group of an EC2 instance will occur immediately.
▹ You can have any number of EC2 instances within a security group
▹ You can have multiple security groups on an EC2 instance

Pricing Modules:

▹ On Demand: allows you to pay a fixed rate by the hour (or by the second) with no commitment

▹ Reserved: Provides you with a capacity reservation, and offers a significant discount on the hourly charge for an instance. Contract Terms are 1 year or 3 Year Terms.

Reserved Pricing Types:

▸ Standard Reserved Instances: Offer up to 75% off on demand instances. The more you pay up front, and the longer the contract, the greater the discount.

◇ Convertible Reserved Instances: These offer up to 54% off on demand capability to change the attributes of the RI (reserved instance) as long as the exchange results in the creation of Reserved Instances of equal or greater value

◇ Scheduled Reserved Instances: These are available to launch within the time windows you reserve. This option allows you to match your capacity reservation to a predictable recurring schedule that only requires a fraction of a day, week, or month.

▸ Spot: Enables you to bid whatever price you want for instance capacity, providing for even greater savings if your applications have flexible start and end times.

▹ Dedicated Hosts: Physical EC2 servers dedicated for your use. Dedicated Hosts can help you reduce costs by allowing you to use your existing server-bound software licenses. Can purchase this by the hour. Can be reserved.

Instance Metadata:

- ▹ Used to get information about an instance (such as public ip)
    - ▸ Curl http://169.254.169.254/latest/meta-data

Placement Groups:

- ▹ Way of placing EC2 instances. The 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 (Compute Optimized, GPU, Memory Optimized, Storage Optimized). AWS recommends homogeneous instances within clustered placement groups. You can't merge placement groups. You can move an existing instance into a placement group. Before you move the instance, the instance must be in the stopped state. You can move or remove an instance by using the AWS CLI or tan AWS SDK, you can't do it via the console yet.
- ▹ Types:
    - ▸ Clustered Placement Group: A grouping of instances within a single Availability Zone. Placement groups are recommended for applications that need **low network latency, high network throughput**, or both. Only certain instances can be launched into a Clustered Placement Group. (Way of putting EC2 instances very very close together within one AZ).
    - ▸ Spread Placement Group: A spread placement group is a group of instances that are each placed on distinct underlying hardware. Spread placement groups are **recommended for applications that have a small number of critical instances that should be kept separate from each other.** Think individual instances. (This is the opposite of a Clustered Placement Group).
    - ▸ Partitioned: When using Partition Placement Groups, Amazon EC2 divides each group into logical segments called partitions. Amazon EC2 ensures that each partition within a placement group has its own set of racks. Each rack has its own network and power source. No two partitions within a placement group share the same racks, allowing you to isolate the impact of hardware failure within your application. Think **multiple instances** (HDFS, HBase, and Cassandra). Very similar to Spread Placement Group.

Notes:

- ▹ On an EBS-backed instance, the default actions are for the root EBS volume to be deleted when the instance is terminated.
- ▹ Termination Protection is turned off by default, so you must turn it on.
- ▹ All Outbound traffic is allowed.
- ▹ All Inbound traffic is blocked by default
- ▹ Security groups are STATEFUL; rules for inbound traffic are automatically applied to outbound traffic implicitly.
- ▹ You can specify allow rules, but not deny rules (as everything is automatically denied)

▹ If the Spot instance is terminated by Amazon EC2, you will not be charged for a partial hour of usage. However, if you terminate the instance yourself, you will be charged for any hour in which the instance ran.

▹ You cannot block specific IP addresses using Security Groups, instead use Network Access Control Lists.

▹ Volumes exist on EBS. Think of EBS as a virtual hard disk.

▹ Snapshots exist on S3. Think of snapshots as a photograph of the disk.

▹ You can create AMIs from both Volumes and Snapshots.

▹ For the best performance, it is recommended that you use current generation instance types and Hardware Virtual Machine (HVM) AMIs when you launch your instances.

▹ Paravirtual guests traditionally performed better with storage and network operations than HVM guests because they could leverage special drivers for I/O that avoided the overhead of emulating network and disk hardware, whereas HVM guests had to translate these instructions to emulated hardware. Now PV drivers are available for HVM guests, so operating systems that cannot be ported to run in a paravirtualized environment can still see performance advantages in storage and network I/O by using them. With these PV on HVM drivers, HVM guests can get the same, or better, performance than paravirtual guests.

▹ To help you manage your Amazon EC2 instances, you can assign your own metadata in the form of Tags.

▹ Underlying Hypervisor for EC2: Nitro and Xen

MONITORING

CLOUDTRAIL

Description:

▹ Increases visibility into your user and resource activity by recording AWS Management Console actions and API calls. You can identify which users and which accounts called AWS, the source IP address from which the calls were made, and when the calls occurred. **Activity monitoring**. Think of a CCTV camera.

CLOUDWATCH

Description:

▹ A monitoring service to monitor your AWS resources, as well as the applications that you run on AWS; **Performance monitoring**. Think of a gym instructor.

Features:

▹ Can monitor most of AWS as well as your applications that run on AWS.

▹ CloudWatch with EC2 will monitor events every **5 minutes by default.**

▹ You can have **1 minute intervals by turning on detailed monitoring**.

- ▹ You can create CloudWatch alarms which trigger notifications.
- ▹ You *cannot* combine blacklisting and whitelisting in CloudFront.

Provides:

- ▹ Dashboards: Creates dashboards to see what is happening with your AWS environment
- ▹ Alarms: Allows you rot set Alarms that notify you when particular thresholds are hit
- ▹ Events: CloudWatch Events help you to respond to state changes in your AWS resources
- ▹ Logs: CloudWatch Logs helps you to aggregate, monitor, and store logs.

Can Monitor:

- ▹ Compute
  - ▸ EC2 Instances
  - ▸ Autoscaling Groups
  - ▸ Elastic Load Balances
  - ▸ Route53 Health Checks
- ▹ Storage & Content Delivery
  - ▸ EBS Volumes
  - ▸ Storage Gateways
  - ▸ CloudFront

Host Level Metrics Consist of:

- ▹ CPU
- ▹ Network
- ▹ Disk
- ▹ Status Check

SECURITY

BASTIONS

Description:

- ▹ Bastion hosts are instances that sit within your public subnet and are typically accessed using SSH or RDP. Once remote connectivity has been established with the bastion host, it then acts as a 'jump' server, allowing you to use SSH or RDP to log in to other instances (within private subnets) deeper within your VPC. When properly configured through the use of security groups and Network ACLs (NACLs), the bastion essentially acts as a bridge to your private instances via the internet.

Features:

- ▹

Notes:

- ▹ Basic steps to creating a bastion host for AWS:

▸ 1.) Launch an EC2 instance as you normally would for any other instance.

▸ 2.) Apply OS hardening as required.

▸ 3.) Set up the appropriate security groups (SG).

▸ 4.) Implement either SSH-agent forwarding (Linux connectivity) or Remote Desktop Gateway (Windows connectivity).

▸ 5.) Deploy an AWS bastion host in each of the Availability Zones you're using.

IDENTITY ACCESS MANAGEMENT (IAM):

Description:

▹ Allows you to manage users and their level of access to the AWS console.

Provides:

▹ Centralized control of your AWS account

▹ Shared access to your AWS account

▹ Granular Permissions (Limiting permissions for specific users)

▹ Identity Federation (services like logging in w/ Facebook)

▹ Multi-factor Authentication

▹ Temporary Access for Users/Devices and Services where Necessary

▹ Allows you to set up password rotation policy

▹ Integrates with many AWS services

▹ Supports PCI DSS Compliance

Consists of:

▹ Users: End Users such as employees

▹ Groups: A collection of users

▹ Roles: You can create roles and assign them to AWS resources. (Like allowing a VM to write files with S3; enables communication between AWS services)

▹ Policies: Made up of documents in a JSON format. They provide permissions for specific groups/users/roles.

Root Account:

▹ The account with the email used to create the AWS account. When you log into AWS with the root account you have a kind of "god mode". No restrictions! This account NEEDS to be secure.

NETWORKING & CONTENT DELIVERY

CLOUDFRONT

Description:

- CloudFront is a Content Delivery Network (CDN); A system of distributed servers (network) that delivers web pages and other web content to a user based on the geographic locations of the user, the origin of the webpage, and a content delivery server.

Features:

- You can restrict access to using signed URLs or cookies

Notes:

- Can take a while (up to an hour) to enable (deploy distribution).
- Needs to be disabled (which can take up to an hour) before removing.
- You can invalidate folders and content (You can go to CloudFront and create an “invalidation”).

ROUTE 53

Description:

- A scalable and highly available Domain Name System service.

Features:

- Amazon’s DNS Service
- With Route 53, there is a default limit of 50 domain names. However, this limit can be increased by contacting AWS support.
- Following Routing Policies are available (All are A-Records):
  - Simple Routing
    - Use for a single resource that performs a given function for your domain, for example, a web server that serves content for the example.com website.
    - If this is chosen, you can only have one record with multiple IP addresses. If you specify multiple values in a record, Route 53 returns all values to the user in a random order.
    - If you go to lucas-pichette.com you might be taken to any of the IP addresses you specified in the record. These IP addresses could contain different versions of lucas-pichette.com.
    - Does not take into account whether a resource is online or slow.
  - Weighted Routing
    - Use to route traffic to multiple resources in proportions that you specify.
    - Allows you to split your traffic based on different weights assigned.
    - For example, you can set 10% of traffic to us-east-1 and 90% to eu-west-1.
    - Does not take into account whether a resource is online or slow.
  - Latency-based Routing
    - Use when you have resources in multiple AWS Regions and you want to route traffic to the region that provides the best latency.

        - Allows you to route your traffic based on the lowest network latency for your end user (i.e. whichever region will provide them the fastest response time).
        - To use latency-based routing, you create a latency resource record set for the Amazon EC2 (or ELB) resource in each region that hosts your website. When Amazon Route 53 receives a query for your site, it selects the latency resource record set for the region that gives the user the lowest latency. Route 53 them responds with the value associated with that resource record set.
    - Failover Routing
        - Use when you want to configure active-passive failover.
        - For example, you may want your primary site to be in eu-west-2 and your secondary DR Site in ap-southeast-2. Route 53 will monitor the health of your primary site using a health check.
        - Designed for active-passive failover (2 and only 2)
    - Geolocation Routing
        - Use when you want to route traffic based on the location of your users.
        - For example, you might want all queries from Europe to be routed to a fleet of EC2 instances that are specifically configured for your European customers. These servers may have the local language of your European customers and all prices are displayed in Euros.
        - Does not take into account whether a resource is online or slow.
    - Geoproximity Routing (Traffic flow only)
        - 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.
        - Geoproximity routing lets Amazon Route 53 route traffic to your resources based on the geographic location of your users and your resources. You can also optionally choose to route more traffic or less to a given resource by specifying a value, known as a bias. A bias expands or shrinks the size of the geographic region from which traffic is routed to a resource.
        - *To use geoproximity routing you must use Route 53 traffic flow.*
        - Does not take into account whether a resource is online or slow.
    - Multivalue Answer Routing
        - Use when you want Route 53 to respond to DNS queries with up to eight healthy records selected at random.
- You can set health checks on individual record sets.
- If a record set fails a health check it will be removed from Route 53 until it passes the health check.
- You can set SNS notifications to alert you if a health check is failed.

Notes:

▹ You can only have one Resource Record, which means that you could not have both Failover Routing and Latency-based Routing (for example). To accomplish something similar to the latter you would need to use Route 53 Traffic Flows, which can be complex and are beyond the CSA-A. They are expensive at $50/month per flow, but this is a prorated cost that isn't charged until you attach it to a DNS through a policy record.

▹ Route 53 gets its name from the fact that DNS is on port 53, and the first ever interstate was Route 66 which traveled across the United States.

▹ Elastic Load Balancers (ELBs) never have a pre-defined IPv4 address. You resolve them using a DNS name.

▹ Given a choice between a CName and an Alias, choose an Alias record.

VIRTUAL PRIVATE CLOUD

Description:

▹ Essentially, a datacenter in the cloud

▹ A virtual network dedicated to a single AWS account. It is logically isolated from other virtual networks in the AWS cloud, providing compute resources with security and robust networking functionality.

Features:

▹ VPC allows you to provision a logically isolated section of the AWS where you can launch AWS resources in a virtual network.

▹ You can launch instances into a subnet of your choosing

▹ You can assign custom IP address ranges in each subnet

▹ You can configure route tables between subnets

▹ You can create internet an internet gateway and attach it to your VPC

▹ Much better security control over your AWS resources with a VPC

▹ You can create instance Security Groups.

▹ You can create subnet network access control lists (ACLs)

▹ Default VPC is super user-friendly. All subnets have a route out to the internet. Each EC2 instance has both a public and private IP address.

▹ A custom VPC only comes with a Route Table, Network ACL, and a Security Group by default.

▹ VPC Peering

‣ Allows you to connect one VPC with another via a direct network route using private IP addresses.

‣ Instances behave as if they were on the same private network.

‣ You can peer VPC's with other AWS accounts as well as with other VPCs in the same account.

- ‣ Always in a star configuration. I.e. 1 central VPC peers with 4 others. No transitive peering. If you have VPC A talking to VPC B and VPC B talking to VPC C, you cannot have VPC A directly talk to VPC C. You would need to create a peering connection between VPC A and VPC C.

Notes:

- ▹ 1 subnet is 1 availability zone. You cannot have one subnet spread across one availability zone, but you can have many subnets in one availability zone.
- ▹ Only one IGW per VPC. IGWs are highly available, though.
- ▹ Security groups do not span VPCs. Every VPC is a separate entity, like two EC2 instances.

DATABASES

AURORA

Description:

- ▹ A MySQL and PostgreSQL-compatible relational database engine that combines the speed and availability of high-end commercial databases with the simplicity and cost-effectiveness of open source databases. Amazon Aurora provides up to five times better performance than MySQL and three times better than PostgreSQL databases at a much lower price point, whilst delivering similar performance and availability.

Features:

- ▹ Start with 10GB, scales in 10GB increments to 64TB (Storage Autoscaling)
- ▹ Compute resources can scale up to 32vCPUs and 244 GB of Memory.
- ▹ 2 copies of your data is contained in each availability zone, with a minimum of 3 availability zones. 6 copies of your data.
- ▹ Scaling Aurora:
    - ‣ Aurora is designed to transparently handle the loss of up to two copies of data without affecting database write availability and up to three copies without affecting read ability.
    - ‣ Aurora storage is also self-healing. Data blocks and disks are continuously scanned for errors and repaired automatically.
- ▹ Two Types of Aurora Replicas:
    - ‣ Aurora Replicas (currently up to 15)
    - ‣ MySQL Read Replicas (currently up to 5)
- ▹ Backups:
    - ‣ Automated backups are always enabled on Amazon Aurora DB Instances. Backups do not impact database performance.
    - ‣ You can also take snapshots with Aurora. This also does not impact on performance.
    - ‣ You can share Aurora Snapshots with other AWS accounts.

Notes:

- ▹

DynamoDB

Description:

▹ A fast and flexible NoSQL database service for all applications 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.

Features:

- ▹ Stored on SSD storage (fast)
- ▹ Spread across 3 geographically distinct data centres
- ▹ Eventual Consistent Reads (Default)
  - ▸ Consistency across all copies of data is usually reached within a second. Repeating a read after a short time should return the updated data (usually within a second; the second rule). (Best read performance, only pick this if you can wait a second before reading updated data.)
- ▹ Strongly Consistent Reads
  - ▸ Returns a result that reflects all writes that received a successful response prior to the reading. (Only pick this if you need to view updated data within or less than 1 second)

Notes:

- ▹

ElastiCache

Description:

▹ A web service that makes it easy to deploy, operate, and scale an in-memory cache in the cloud. The service improves the performance of web applications by allowing you to retrieve information from fast, manged, in-memory caches, instead of relying entirely on slower disk-based databases. ElastiCache supports two open-source in-memory caching machines. Used to speed up performance of existing databases (Frequent identical queries).

Features:

▹ Use Elasticache to increase database and web application

▹ Redis, one of the two caching engines, offers multi-az, backups and restores, horizontal scaling, and much more. It does NOT offer Multi-threaded performance, unlike Memcached which only offers the latter, horizontal scalability, and simple cache to offload DB.

Notes:

- ▹

Relational Database Service (RDS)

Description:

- Relational Databases are:
    - A hierarchical version of data-management which organizes data based on relation.
    - Can be documents, key/value pairs, or other; Has a dynamic structure.
    - Horizontally-scalable
    - Not-Only/No Structured Query Language (NoSQL)
- Non-Relational Databases are:
    - A row-major (rows are “fields”) where columns are “attributes” of the “field”. Also, these relationships can be called: Collection (Table), Document (Row), and Key-Value Pairs (Columns)
    - Static structured data which is based on a row-column basis.
    - Vertically
    - Structured Query Language (SQL)

Features:

- Multi-AZ - For Disaster Recovery
    - Multi-AZ is for Disaster Recovery only
    - Available for:
        - SQL Server
        - Oracle
        - MySQL Server
        - PostgreSQL
        - MariaDB
- Read Replicas - For Performance (Up to 5 copies)
- Automated Backups (*Not* one of the two main features)
    - Automated Backups allow you to recover your database to any point in time within a “retention period”. The retention period can be between 1-35 days. Automated Backups will take a full daily snapshot and will also store transaction logs throughout the day. When you do a recovery, AWS will first choose the most recent daily backup, and then apply transaction logs relevant to that day. This allows you to do a point in time recovery down to a second, within the retention period.
    - Enabled by default. The backup data is stored in S3 and you get free storage space equal to the size of your database. So if you have an RDS instance of 10Gb, you will get 10Gb worth of storage.
    - Backups are taken within a defined window. During the defined window, storage I/O may be suspended while your data is being backed up and you may experience elevated latency.

- Whenever you restore either an Automatic Backup or a manual Snapshot, the restored version of the database will be a new RDS instance with a new DNS endpoint.

Notes:

- For OLTP
- Runs on virtual machines
- You cannot log in to these operating systems (can’t ssh and connect to the virtual machines)
- Patching of the RDS Operating System and DB is Amazon’s responsibility
- RDS is NOT Serverless (besides Aurora Serverless)
- Encryption at rest is supported for MySQL, Oracle, SQL Server, PostgreSQL, MariaDB, and Aurora. Encryption is done using the AWS Key Management Service (KMS). Once your RDS instance is encrypted, the data stored at rest in the underlying storage is encrypted, as are its automated backups, read replicas, and snapshots.

REDSHIFT

Description:

- A fast and powerful, fully managed, petabyte-scale data warehouse service in the cloud. Customers can start small for just $0.25 per hour with no commitments or upfront costs and scale to a petabyte or more for a $1,000 per terabyte per year, less than a tenth of most other data warehousing solutions. *Business intelligence*.

Features:

- Single Node (160Gb)
- Multi-Node
  - Leader Node (manages client connections and receives queries).
  - Compute Nodes (store data and perform queries and computations). Up to 128 Compute Nodes.
- Advanced Compression
  - Columnar data stores can be compressed much more than row-based data stored because similar data is stored sequentially on disk. Amazon Redshift employs multiple compression techniques and can often achieve significant compression relative to traditional relational data stores. In addition, Amazon Redshift doesn’t require indexes or materialized views, and so uses less space than traditional relational database systems. When loading data into an empty table, Amazon Redshift automatically samples your data and selects the most appropriate compression scheme.
- Massively Parallel Processing (MPP)
  - Amazon Redshift automatically distributes data and query load across all nodes. Amazon Redshift makes it easy to add nodes to your data warehouse and enables you to maintain fast query performance as your data warehouse grows.

- ▹ Backups
  - ▸ Enabled by default with a 1 day retention period.
  - ▸ Maximum retention period is 35 days.
  - ▸ Redshift always attempts to maintain at least three copies of your data (the original and replica on the Compute Nodes and a backup in Amazon S3).
  - ▸ Redshift can also asynchronously replicate your snapshots to S3 in another region for disaster recovery.

Pricing:

- ▹ Compute Node hours (total number of hours you run across all your Compute Nodes for the billing period. You are billed for 1 unit per node per hour, so a 3-node data warehouse cluster running persistently for an entire month would incur 2,160 instance hours. You would not be charged for Leader Node hours; only Compute Nodes will incur charges.)
- ▹ Backups
- ▹ Data Transfers (only within a VPC, not outside of it)

Notes:

- ▹ Always encrypted in transit using SSL
- ▹ Encrypted at rest using AES-256 encryption
- ▹ By default Redshift takes care of key management; You can manage your own keys through a Hardware Security Module (HSM) if you would like, or through AWS Key Management Service (KMS).
- ▹ Currently only available in 1 AZ; multi-az not offered yet.
- ▹ Can restore snapshots to new AZs in the event of an outage.

SNOWBALL

Description:

Can import and export to/from S3. Encouraged when dealing with large amounts of data, as it is cost effective (cheaper) and saves time (faster).

SUPPLEMENTARY INFORMATION

CONTENT DELIVERY NETWORK (CDN)

- No CDN would result in users pulling directly from the server (literally across oceans at times).
- Edge Location: This is the location where content will be cached. This is separate to an AWS region/AZ.
- Origin: This is the origin of all the files that the CDN will distribute. This can be an S3 Bucket, an EC2 Instance, an Elastic Load Balancer, or Route53.

- Distribution: Name given to the CDN; consists of a collection of Edge Locations.
- Two kinds of distribution:
    - Web Distribution: Typically used for websites
    - RTMP: Used for media streaming

DOMAIN NAME SYSTEM (DNS)

Description:

▹ DNS is used to convert human friendly domain names (such as https://lucas-pichette.github.io) into an Internet Protocol (IP) address (such as http://82.124.53.1)

Features:

▹ Offers two kinds of connections, IPv4 and IPv6.

▹ Holds records, including:

▸ A-Record: Fundamental type of DNS record. The “A” in A record stands for “Address”. The A record is used by a computer to translate the name of the domain to an IP address. (www.google.com to http://123.45.67.89)

▸ CName: A Canonical Name (CName) can be used to resolve one domain name to another. For example, you may have a mobile website with the domain name http://m.acloud.guru that is used for when users browse to your domain name on their mobile devices. You may also want the name http://mobile.acloud.guru to resolve this same address.

▸ Alias Records: Alias Records are used to map resource record sets in your hosted zone to Elastic Load Balancers (ELBs), CloudFront distributions, or S3 buckets that are configured as websites.

▸ Alias Records have special functions that are not present in other DNS servers. Their main function is to provide special functionality and integration into AWS services

▸ PTR Record: Opposite of an A Record; Uses IP address to determine the domain (name).

▹ Various TTLs (Time To Live). A DNS has a standard TTL of 48 hours; An update may take 48 hours or longer to go live.

Notes:

▹ IPv4

▸ Space is 32 bit field and has over 4 billion different addresses (4,294,967,296)

▹ IPv6

▸ Space has 128 bits and has over 340 undecillion addresses (340,282,366,920,938,463,463,374,607,431,768,211,456)

▹ Alias Record vs. CName: A CName can’t be used for naked domain names (zone apex record). You can’t have a CName for http://acloud.guru, it must be either an A Record or an Alias.

Network Address Translation (NAT)

Description:

▹ You can use a NAT device to enable instances in a private subnet to connect to the internet (for example, for software updates) or other AWS services, but prevent the internet from initiating connections with the instances. A NAT device forwards traffic from the instances in the private subnet to the internet or other AWS services, and then sends the response back to the instances. When traffic goes to the internet, the source IPv4 address is replaced with the NAT device’s address and similarly, when the response traffic goes to those instances, the NAT device translates the address back to those instances’ private IPv4 addresses.

Features:

▹ Two main kinds, NAT Instance and NAT Gateway.

▹ NAT Instance:

‣ You can use a network address translation (NAT) instance in a public subnet in your VPC to enable instances in the private subnet to initiate outbound IPv4 traffic to the Internet or other AWS services, but prevent the instances from receiving inbound traffic initiated by someone on the Internet.

‣ Cheaper, more legacy. Can bottleneck flows easily [*fig.1*]. If you’re bottlenecking just increase the NAT Instance size.

‣ To use a NAT Instance, remember that the NAT acts as a bridge between your EC2 Instance (which is pretty much what a NAT Instance consists of), through your private and public subnets, and to your IGW. To create this bridge source/destination checks must be disabled on the NAT Instance.

▹ NAT Gateway:

‣ You can use a network address translation (NAT) gateway to enable instances in a private subnet to connect to the internet or other AWS services, but prevent the internet from initiating a connection with those instances.

‣ Provides better availability, higher bandwidth, and requires less administrative effort.

Notes:

▹ NAT is not supported for IPv6 traffic—use an egress-only Internet gateway instead.

OnLine Transaction/Analytic Processing (OLTP/OLAP)

Description:

▹ OLAP

‣ Runs queries across database to return certain analytics

- ▸ Example: Net Profit for EMEA and Pacific for the Digital Radio Product. Pulls in large numbers of records, such as: Sum of Radios Sold in EMEA, Sum of Radios Sold in Pacific, Unit Cost of Radio in each region. Sales price of each radio, sales price - unit cost.

- ▹ OLTP
  - ▸ A single row per transaction to track various information regarding a certain transaction
  - ▸ Example: Transaction Number: 1780122, Product ID: F2003CA, Date of Transaction: 02022000 (or some other structure)

Features:

- ▹

Notes:

- ▹

SIMPLE NOTIFICATION SERVICE (SNS)

Description:

- ▹ Simple Notification Service (Such as a text reminder to notify you that you've gone over the allotted space for your cloud service). - This is in CloudWatch, a Billing Alarm is a kind of SNS which utilizes email notifications.

START OF AUTHORITY RECORD (SOA)

Description:

- ▹ A type of resource record in the Domain Name System containing administrative information about the zone, especially regarding zone transfers.

Features:

- ▹ Stores information about:
  - ▸ The name of the server that supplied the data for the zone.
  - ▸ The administrator of the zone
  - ▸ The current version of the data file
  - ▸ The default number of second for the time-to-live (TTL) file on resource records

Notes:

- ▹

NAME SERVER RECORDS (NS)

Description:

- ▹ This record indicates which DNS server is authoritative for that domain (which server contains the actual DNS records). A domain will often have multiple NS records which can indicate primary and backup name servers for that domain.

Features:

▹ Used by Top Level Domain servers to direct traffic to the Content DNS server which contains the authoritative DNS records.

Notes:

▹

STORAGE GATEWAY

Description:

A service that connects an on-premises software appliance with cloud-based storage to provide seamless and secure storage to provide seamless and secure integration between an organization's on-premises IT environment and AWS's storage infrastructure. The service enables you to securely store data to the AWS cloud for scalable and cost-effective storage.

- File Gateway (NFS & SMB)
    - Files are stored as objects in your S3 buckets, accessed through a Network File System (NFS) mount point. Ownership, permissions, and timestamps are durably stored in S3 in the user-metadata of the object associated with the file. Once objects are transferred to S3, they can be managed as native S3 objects, and bucket policies such as versioning, lifecycle management, and cross-region replication apply directly to objects stored in your bucket.
    - Flat files, stored directly on S3
- Volume Gateway (iSCI) - Actual disks
    - Stored Volumes
        - Entire dataset stored on site, and is asynchronously backed up to S3.
    - Cached Volumes
        - Entire dataset is stored on site and the most frequently accessed data is cached on site.
- Tape Gateway (VTL) - Virtual Tape Library

TOP LEVEL DOMAIN (TLD)

Description:

▹ One of the domains at the highest level in the hierarchical Domain Name System of the Internet. The top-level domain names are installed in the root zone of the name space.

Features:

▹ Used to help identify the website that it belongs to.

▹ Some examples include: .com, .org, .net, .edu, .gov, .mil, *.anything*

Notes:

▹ http://www.google.com ; Protocol (HyperText Transfer), Subdomain, Domain (name), Top Level Domain (TLD)

You can select your (Amazon Machine Image, think of Linux) AMI based on:

- Region
- Operating Systems
- Architecture (32-bit or 64-bit)
- Launch Permissions
- Storage for the Root Device (Root Device Volume)
  - Instance Store (Ephemeral Storage) - If for some reason they're stopped, all data is terminated. Ephemeral Storage does not come back or persist, it lives with the instance. You can reboot it though without losing data.
  - EBS Backed Volumes

Encrypted Root Device Volumes (RDV) & Snapshots:

- Snapshots of encrypted volumes are encrypted automatically
- Volumes restored from encrypted snapshots are encrypted automatically
- You can share snapshots, but only if they are unencrypted.
- Snapshots can be shared with other AWS accounts or made public (have to be unencrypted)
- You can now encrypt root device volumes upon creation of the EC2 instance.
- If you have an unencrypted root device volume (unencrypted RDV) that you need encrypted then:
  - Create a snapshot of the unencrypted rdv
  - Create a copy of the Snapshot and select the encrypt option
  - Create an AMI from the Snapshot
  - Use that AMI to launch new encrypted instances.

Misc:

- You cannot log in to the AWS console using the Access Key ID / Secret Access Key pair. Instead, you must generate a password for the user, and supply the user with this password and your organization's unique AWS console login URL.
- Using SAML (Security Assertion Markup Language 2.0), you can give your federated users single sign-on (SSO) access to the AWS Management Console.

FIGURES

Figure 1

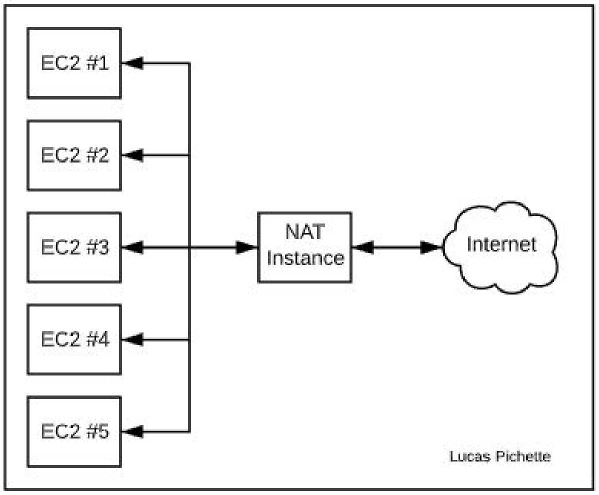

Figure 2