Notes from documentation for AWS Certified Solutions Architect Associate Exam

**S3**

* Data transfer within the same region is not charged (S3 > S3 or EC2 > S3), but it IS charged across regions
* Normal rates apply to versioning > this is why lifecycle rules can help manage costs
* Encryption
  + SSE (server side encryption) can be used to encrypt data stored at rest
  + Can configure S3 buckets to automatically encrypt objects before storing them
  + Can use your own encryption libraries to encrypt before storing
* Controlling access
  + IAM
  + Bucket policies
  + Access Control Lists (ACLs)
  + Query string authentication (URL to S3 object that’s only available for a limited time)
* SSE types
  + SSE-S3: Amazon handles everything
  + SSE-C: you leverage S3 to perform encryption & decryption, while you control the keys. You need to manage your own encryption keys but don’t want to implement a client-side encryption library.
  + SSE-KMS: You get to configure everything & can provide keys, AWS executes all of it. Other pros are audit trails & separate permissions for use of master key
* Amazon Macie is an AI-powered security service that helps prevent data loss by automatically discovering, classifying, and protecting sensitive data stored in S3.
  + Recognizes PII, intellectual property, assigns business value, provides visibility into where data is stored and how it’s being used. Monitors for anomalies and alerts when detects risk of unauthorized access or inadvertent data leaks.
* How is 99.(11 9s)% durability achieved?
  + S3 Standard, IA, and Glacier store objects on multiple devices across a minimum of 3 AZs. One Zone-IA stores objects on multiple devices in only 1 AZ.
* S3 One Zone-IA is ~20% less $$ than Standard IA. Availability is 99% rather than 99.9%.
* S3 Glacier retrieval is minutes to hours. Glacier Deep Archive is hours to days.
* S3 has “query in place” functionality that allows you to run queries off of data as exists rather than having to move it to a separate analytics platform. S3 Select, Amazon Athena, Amazon Redshift Spectrum are all ways to do this.
  + S3 Select allows you to retrieve specific data with simple queries for CSV, JSON, and Parquet files
  + Athena lets you query within Athena Management Console, directly on any S3 object (CSV, JSON, ORC Parquet, Avro)
  + Redshift Spectrum allows you to run queries against exabytes of data.
* S3 Transfer acceleration leverages CloudFront to make faster transfers over long distances. Sends data to edge locations and then routed to s3 bucket (can use speed comparison tool to see difference).
  + If it’s not faster, AWS will bypass the transfer acceleration option and not charge for it and just do the normal route.
* Object tags
  + Can be used for IAM policies, lifecycle policies, and customize storage metrics.
  + Can be replicated using cross-region replication, but new permissions must be set up.
* Storage Class Analysis – seems like this is a descriptive tool that allows you to see patterns so that you can create lifecycle policies and whatnot, rather than Tiered Intelligence that does it for you.
* S3 Inventory report > can configure this to output a list of everything on a daily or weekly basis.
* S3 Object Lock is a feature that blocks object version deletion during a customer-defined retention period so you can enforce retention policies as an added layer of data protection or for regulatory compliance.
* S3 lets applications connect to S3 using IPv6 without any additional translation software or systems.

**EC2**

* EC2 = Elastic Cloud Compute
* Initial limit in a region is 20 on demand instances. You have to request more.
  + Also limit of 20 reserved instances. Some on-demand instance limits are lower (generally for bigger instances?).
  + Spot instances are limited by dynamic spot limit per region (probably mostly availability)
* When you use Elastic Block Store (EBS), data on the root device will persist separately from the instance. Local instance store only persists through the lifetime of the instance.
* Instance types & notes:
  + **Accelerated Computing** instance family is a family of instances which use hardware accelerators, or co-processors, to perform some functions, such as floating-point number calculation and graphics processing, more efficiently than is possible in software running on CPUs.
    - **GPU** instances work best for applications with massive parallelism such as workloads using thousands of threads.
    - **P3** instances with their high computational performance will benefit users in artificial intelligence (AI), machine learning (ML), deep learning (DL) and high performance computing (HPC) applications.
    - **FPGAs** are programmable integrated circuits that you can configure using software. By using FPGAs you can accelerate your applications up to 30x when compared with servers that use CPUs alone.
  + **Compute Optimized** instances are designed for applications that benefit from high compute power. These applications include compute-intensive applications like high-performance web servers, high-performance computing (HPC), scientific modelling, distributed analytics and machine learning inference.
    - C5 (newest, latest generation), C4
  + **General Purpose**
    - A1 instances are ideal for scale-out applications such as web servers, containerized microservices, caching fleets, and distributed data stores.
    - M5 instances offer a good choice for running development and test environments, web, mobile and gaming applications, analytics applications, and business critical applications including ERP, HR, CRM, and collaboration apps.
    - Customers should consider using the M5a and M5ad instance types if they are looking to save money on price when their workloads do not fully utilize the compute resources of their chosen instance, resulting in them paying for performance that they don’t actually need.
    - T2 instances provide a cost-effective platform for a broad range of general purpose production workloads.
  + **High Memory**
    - Five High Memory instances are available. u-6tb1.metal offers 6 TB memory; u-9tb1.metal offers 9 TB memory; u-12tb1.metal offers 12 TB memory; u-18tb1.metal offers 18 TB memory; and u-24tb1.metal offers 24 TB memory.
  + **Previous Generation**
    - C1, C3, CC2, CR1, G2, HS1, M1, M2, M3, R3 and T1
    - Still fully functional and supported, but previous generation.
  + **Memory Optimized**
    - X1 instances are ideal for running in-memory databases like SAP HANA, big data processing engines like Apache Spark or Presto, and high performance computing (HPC) applications.
    - X1e instances offer 32GB of memory per vCPU whereas X1 instances offer 16GB of memory per vCPU.
  + **Storage Optimized**
    - Dense-storage instances (D2) and HDD-storage instances (H1) are optimized for applications that require high sequential read/write access and low cost storage for very large data sets
    - I3 and I3en
* **Volumes** 
  + ST1 (Throughput Optimized HDD) volumes are backed by hard disk drives (HDDs) and are ideal for frequently accessed, throughput intensive workloads with large datasets and large I/O sizes, such as MapReduce, Kafka, log processing, data warehouse, and ETL workloads.
  + SC1 (Cold HDD) volumes are backed by hard disk drives (HDDs) and provides the lowest cost per GB of all EBS volume types. It is ideal for less frequently accessed workloads with large, cold datasets.
  + Can attach multiple volumes to one instance, cannot attach multiple instances to one volume.
  + Recommended to detatch or shut down machine before taking snapshot.
* **Elastic File System (EFS)**
  + Can have multiple instances connected to same EFS.
* **Elastic IP** : static IPv4 address associated with account. You can mask failure of instance by rapidly remapping address to another instance in account.
  + Static IP address that doesn’t change when you stop/start an instance, so you don’t have to worry about the address changing.
* Hibernate vs stop: Hibernate retains memory across stop/start cycles. Stop removes all memory from root volume.
* Reserved instances
  + Standard RIs offer a significant discount on EC2 instance usage when you commit to a particular instance family. Convertible RIs offer you the option to change your instance configuration during the term, and still receive a discount on your EC2 usage.
* With a single API call, EC2 Fleet lets you provision compute capacity across different instance types, Availability Zones and across On-Demand, Reserved Instances (RI) and Spot Instances purchase models to help optimize scale, performance and cost.
* On Demand Capacity Reservation: reserving specific capacity for a time period.
  + Different from reserved instances – RIs are a 1 or 3 year time period. Capacity reservations can be created for any duration.
* The Reserved Instance Marketplace is an online marketplace that provides AWS customers the flexibility to sell their Amazon Elastic Compute Cloud (Amazon EC2) Reserved Instances to other businesses and organizations.
* While running, Spot Instances are exactly the same as On-Demand or Reserved instances. The main differences are that Spot Instances typically offer a significant discount off the On-Demand prices, your instances can be interrupted by Amazon EC2 for capacity requirements with a 2-minute notification, and Spot prices adjust gradually based on long term supply and demand for spare EC2 capacity.
* Amazon EC2 instances are grouped into 5 families: General Purpose, Compute Optimized, Memory Optimized, Storage Optimized and Accelerated Computing instances. General Purpose Instances have memory to CPU ratios suitable for most general purpose applications and come with fixed performance (M5, M4) or burstable performance (T2); Compute Optimized instances (C5, C4) have proportionally more CPU resources than memory (RAM) and are well suited for scale out compute-intensive applications and High Performance Computing (HPC) workloads; Memory Optimized Instances (X1e, X1, R4) offer larger memory sizes for memory-intensive applications, including database and memory caching applications; Accelerating Computing instances (P3, P2, G3, F1) take advantage of the parallel processing capabilities of NVIDIA Tesla GPUs for high performance computing and machine/deep learning; GPU Graphics instances (G3) offer high-performance 3D graphics capabilities for applications using OpenGL and DirectX; F1 instances deliver Xilinx FPGA-based reconfigurable computing; Storage Optimized Instances (H1, I3, I3en, D2) that provide very high, low latency, I/O capacity using SSD-based local instance storage for I/O-intensive applications, with D2 or H1, the dense-storage and HDD-storage instances, provide local high storage density and sequential I/O performance for data warehousing, Hadoop and other data-intensive applications. When choosing instance types, you should consider the characteristics of your application with regards to resource utilization (i.e. CPU, Memory, Storage) and select the optimal instance family and instance size.

**IAM**

* Can create “federated users” which are users outside of AWS
* Can create and manage using IAM APIs, AWS CLI, or IAM console
* IAM user can be individual, system, or application
* Groups are collections of IAM users
  + Users can belong to multiple groups
  + Groups can’t belong to other groups
  + Can be granted permissions using access control policies
* Users are global entities
* Can enforce automatic password expiration, prevent re-use of old passwords, etc
* Cannot set usage quotas on individual users
* IAM **ROLE** is an entity that defines a permissions set for making aws service request. Not associated with a specific user or group, but users, applications, and aws services can be assigned roles
* Can associate an IAM role with an EC2 instance, so the applications running on the instance can use credentials to access aws services defined in the IAM policy
* A service-linked-role links an AWS service so that only the linked service can assume the role.
* Managed policies can be attached to multiple IAM users, groups, or roles
* For granular permsisions, you can define service actions you want to allow or deny access to. Can also define aws specific resources the actions can be performed on. And can define conditions to specify when a policy is in effect (i.e. if mfa is enabled or not)
* Can use policy simulator to test policy changes, validate existing policies, etc.
* Temporary security credentials consist of aws access key id, secret access key, security token.

**RDS**

**DYNAMODB**