SAP-C03 · CHEAT SHEET · CERTIFYDECK

AWS Certified Solutions Architect – Professional

Nivel: Advanced (Professional) Duración: 180 min Nota de corte: 720/1000 Fuente: guía de estudio oficial de Microsoft Learn

1. Architecture Design and Implementation Built for the Cloud 25–25%

Design architectures built for the cloud

- Las colas estándar de SQS y las asignaciones de origen de eventos de Lambda entregan al menos una vez, por lo que el consumidor debe ser idempotente. Las colas FIFO ordenan por ID de grupo de mensajes y deduplican con un ID o con deduplicación basada en contenido; los temas FIFO de SNS distribuyen a colas SQS, no directamente a Lambda. En Kinesis, la clave de partición fija el fragmento y el orden por clave.
- Una aplicación sin estado guarda la sesión y los datos del usuario fuera de la instancia o tarea, en ElastiCache, DynamoDB, EFS o S3, para poder sustituir y escalar servidores sin perder nada. Las tablas globales de DynamoDB aceptan escrituras en todas las regiones y TTL borra lo caducado; en un almacén global de ElastiCache solo escribe el clúster principal. Las URL prefirmadas de S3 permiten subir archivos sin pasar por el servidor.
- Una canalización sin servidor típica ingiere en Kinesis Data Streams, entrega a S3 con Amazon Data Firehose (que convierte JSON a Parquet con un esquema del Glue Data Catalog y particiona dinámicamente) y consulta con Athena. Los marcadores de trabajo de Glue evitan reprocesar datos; en Athena, los formatos columnares, el particionado y la proyección de particiones reducen los datos examinados y el mantenimiento.
- Los flujos Standard de Step Functions son de exactamente una vez, admiten .sync y .waitForTaskToken y sirven para procesos largos y auditables; los Express son para eventos de gran volumen y breves, de al menos una vez (asíncronos) o como máximo una vez (síncronos), y el tipo no se cambia tras crearlo. Las funciones duraderas de Lambda escriben el flujo en código con pasos con puntos de control y esperas sin coste de cómputo.
- Para recibir eventos de otras cuentas, el bus destino necesita una política basada en recursos que permita cuentas concretas o toda la organización por su ID, y las reglas emisoras apuntan al ARN del bus con un rol de IAM. EventBridge no reenvía a una tercera cuenta los eventos recibidos de otra. Los puntos de enlace globales conmutan a una segunda región con una comprobación de estado de Route 53 y vuelven solos si se activa la replicación.

Integrate AI and ML services into architectures

- La inferencia entre regiones de Amazon Bedrock usa perfiles geográficos (US, EU, APAC), adecuados para la residencia de datos, o globales; si las SCP bloquean regiones, deben permitir todas las regiones de destino del perfil. Los perfiles de inferencia de aplicación con etiquetas permiten imputar costes y medir el uso por equipo, y un punto de enlace de VPC de interfaz para bedrock-runtime evita que el tráfico salga a Internet.
- Amazon Bedrock AgentCore separa las piezas de un agente: Runtime aloja agentes de cualquier marco con sesiones aisladas en microVM, Gateway convierte APIs OpenAPI, funciones de Lambda y servicios en herramientas MCP con autenticación de entrada y salida, Identity gestiona OAuth 2.0 en nombre de los usuarios con un almacén de tokens y Memory guarda contexto a corto y largo plazo. Knowledge Bases aporta RAG con filtros de metadatos.
- SageMaker AI ofrece inferencia en tiempo real (punto de enlace persistente y de baja latencia), serverless (tráfico intermitente, escala a cero, sin GPU ni configuración de VPC), asíncrona (cola, cargas grandes en S3, procesamiento largo, escala a cero y avisos por SNS) y batch transform (lotes sin punto de enlace). Los puntos de enlace multimodelo comparten flota para muchos modelos y SageMaker AI Runtime admite PrivateLink.
- Amazon Bedrock Knowledge Bases aplica RAG: ingiere documentos, los divide en fragmentos (fijo, jerárquico, semántico o sin fragmentar), genera embeddings y los guarda en un almacén vectorial. La sincronización es incremental. En la recuperación se pueden filtrar resultados con metadatos de archivos .metadata.json y usar búsqueda híbrida (vectorial más texto) en RDS, OpenSearch Serverless y MongoDB.

Design hybrid cloud integration patterns

- Para máxima resiliencia, Direct Connect usa conexiones en dispositivos separados en más de una ubicación; alta resiliencia es una conexión por ubicación. En una Transit Gateway, las rutas estáticas ganan a las propagadas y las de Direct Connect gateway a las de VPN, así que una VPN con BGP sirve de respaldo automático. ECMP entre túneles VPN exige enrutamiento dinámico y el mismo ASN en lo que ocupa entre tipos de adjunto distintos.
- DataSync transfiere archivos y objetos entre almacenamiento local (NFS, SMB, HDFS, objetos) y servicios de AWS con tareas programables e incrementales; con un punto de enlace de servicio de VPC el tráfico no pasa por internet. Para modernizar sistemas heredados sin tocarlos: las API REST de API Gateway transforman cuerpos con plantillas de asignación VTL en integraciones sin proxy (las API HTTP solo transforman parámetros) y passthroughBehavior decide si se rechazan tipos de contenido sin plantilla. Amazon MQ es ActiveMQ o RabbitMQ administrado con JMS, AMQP o MQTT sin reescribir código, y Lambda puede consumir y transformar sus mensajes.
- Outposts lleva capacidad administrada por AWS a las instalaciones del cliente como extensión de una región, con service link hacia la región y puerta de enlace local hacia la red local. Las Local Zones acercan recursos a grandes urbes. ECS Anywhere registra servidores propios como instancias externas de un clúster de ECS, EKS Hybrid Nodes los une a EKS con conexión fiable y EKS Anywhere, administrado por el cliente, funciona incluso desconectado.

Design network architectures

- La interconexión de VPC une dos VPC (también entre cuentas y regiones) sin CIDR solapados, no es transitiva ni permite usar la VPN, Direct Connect, NAT o puerta de internet de la otra VPC, y exige rutas en ambos lados.
- Amazon VPC IPAM organiza el espacio IP en ámbitos, grupos jerárquicos por región y entorno y asignaciones a VPC, supervisa solapamientos y cumplimiento, y con AWS Organizations comparte grupos por AWS RAM. Una SCP que deniega ec2:CreateVpc y ec2:AssociateVpcCidrBlock sin ec2:Ipv4IpamPoolId obliga a usarlo. Las subredes públicas enrutan a la puerta de internet, las privadas salen por NAT y el uso compartido de subredes centraliza la inspección para que el tráfico sea simétrico. Gateway Load Balancer inserta de forma transparente dispositivos de terceros para todo el tráfico IP. AWS WAF protege en capa 7 ALB, CloudFront y API Gateway. Los grupos de seguridad tienen estado y solo permiten; las ACL de red, sin estado, admiten denegaciones. Flow Logs solo registra.
- AWS PrivateLink expone un servicio tras un Network Load Balancer a VPC consumidoras con puntos de enlace de interfaz, en un solo sentido y admitiendo CIDR solapados. Transit Gateway aporta enrutamiento transitivo a escala. La inspección centralizada con Transit Gateway necesita el modo de dispositivo en el adjunto de la VPC de inspección para que el tráfico sea simétrico. Gateway Load Balancer inserta de forma transparente dispositivos de terceros para todo el tráfico IP. AWS WAF protege en capa 7 ALB, CloudFront y API Gateway. Los grupos de seguridad tienen estado y solo permiten; las ACL de red, sin estado, admiten denegaciones. Flow Logs solo registra.
- Una VPC IPv4 puede pasar a pila doble asociando bloques CIDR IPv6 a la VPC y a las subredes; IPv4 no se puede desactivar y no hay migración de subred IPv4 a solo IPv6. Para IPv6 solo de salida se usa una puerta de enlace de internet de solo salida. Las cargas solo IPv6 llegan a servicios IPv4 con DNS64 en la subred y NAT64 en una puerta NAT. Un ALB dualstack publica un registro AAAA.
- Route 53 ofrece enrutamiento de conmutación por error (activo-pasivo), por latencia, por geolocalización (continente, país o estado, con registro predeterminado para ubicaciones no asignadas), geoproximidad con sesgo, ponderado, multivalor y por IP; Evaluate Target Health en alias de balanceadores desvía el tráfico de recursos no sanos. Global Accelerator da IP anycast estáticas para TCP y UDP y conmutación regional sin depender de la caché DNS.

Design data, storage, and database architectures

- Un lago de datos en Amazon S3 se cataloga en AWS Glue Data Catalog y se gobierna con AWS Lake Formation: permisos por columna, fila y celda, LF-Tags para escalar y uso compartido entre cuentas con AWS RAM sin copias. Redshift Spectrum consulta S3 desde Redshift con tablas externas de solo lectura. Las tablas Apache Iceberg añaden actualizaciones y borrados por registro, evolución del esquema y aislamiento de instantáneas.
- Amazon MSK ejecuta Apache Kafka administrado y compatible con las aplicaciones existentes, y Redshift puede ingerir en streaming desde MSK o Kinesis Data Streams directamente en vistas materializadas. Amazon AppFlow mueve datos de SaaS como Salesforce a S3 o Redshift sin código por programación o por evento y con flujos privados. En EMR, los nodos principal y central bajo demanda y los de tarea en spot protegen HDFS y abaratan el procesamiento.
- La elección de base de datos sigue el patrón de acceso: DynamoDB para accesos por clave con milisegundos a cualquier escala y modo bajo demanda sin planificar capacidad (sin JOIN), DAX para lecturas eventualmente consistentes en microsegundos, Aurora y RDS para OLTP relacional con transacciones y réplicas de solo lectura, y Redshift para analítica. La integración zero-ETL de Aurora lleva los datos transaccionales a Redshift casi en tiempo real. Elige el almacenamiento según protocolo y patrón de acceso: FSx for Lustre es un sistema POSIX paralelo que se vincula a S3 con importación y exportación automáticas; FSx for NetApp ONTAP sirve los mismos datos por NFS y SMB; EFS (solo NFS, sin Windows) usa rendimiento Elastic para picos impredecibles y ciclo de vida hacia Infrequent Access; EBS Multi-Attach solo vale para io1/io2 en una zona.
- Aurora Global Database replica en la capa de almacenamiento hacia otras regiones; su conmutación por error administrada vuelve a anadir la región antigua como secundaria y el reconfigurar permite escribir desde una secundaria. Las tablas globales de DynamoDB son activo/activo: MREC replica de forma asíncrona y MRSC de forma síncrona con RPO cero. Un clúster Multi-AZ de RDS tiene dos réplicas legibles en una región.

2. Security, Compliance, and Governance

25–25%

Design identity security architectures

- IAM Identity Center conecta un IdP externo con SAML 2.0 para el inicio de sesión y SCIM para aprovisionar usuarios y grupos; el NameID de SAML debe coincidir con el atributo asignado a Username, y los atributos sincronizados pueden pasarse como etiquetas de sesión para ABAC. Para cargas de trabajo externas, como GitHub Actions, la federación OIDC cambia un token por credenciales temporales de un rol.
- El acceso entre cuentas requiere que la política de confianza del rol y la política de identidad del principal lo permitan. Un ID externo único por cliente evita el suplente confundido en accesos de terceros. Los límites de permisos fijan el máximo de un rol y permiten delegar la creación de roles con la condición iam:PermissionsBoundary, y las políticas de sesión restringen una sesión concreta sin conceder permisos.
- IAM Roles Anywhere da credenciales temporales a cargas de trabajo fuera de AWS que se autentican con certificados X.509 de una CA registrada como ancla de confianza (AWS Private CA o una CA externa). El perfil confía en rolesanywhere.amazonaws.com y puede filtrar por atributos del certificado como aws:PrincipalTag/x509Subject/CN; las CRL importadas deben reimportarse y el asistente de credenciales admite credential_process, update, serve y PKCS#11.

Design solutions to implement encryption and key management strategies

- Las claves administradas por el cliente permiten controlar la política de clave y activar la rotación automática con un periodo propio, solo en claves simétricas con material generado por KMS; las de material importado solo rotan bajo demanda. Las claves administradas por AWS rotan cada año y no se pueden compartir. Las claves multirregión comparten material y la rotación se activa en la principal; S3 Bucket Keys reduce las llamadas a KMS.
- AWS CloudHSM ofrece HSM de un solo inquilino con validación FIPS bajo control del cliente. Un almacén de claves de KMS respaldado por CloudHSM combina la gestión de claves de KMS con claves generadas en el clúster, que necesita HSM en dos zonas, pero no admite rotación automática ni claves multirregión; así se rota de forma manual. CloudHSM también permite descargar TLS en servidores web y copias de seguridad a otra región para DR.
- AWS KMS ofrece claves asimétricas postcuánticas ML-DSA (USE_SIGN_VERIFY) para firmas y ML-KEM (uso ENCAPSULATE_DECAPSULATE) para establecer secretos compartidos; la clave privada nunca sale de KMS sin cifrar. Estas claves no admiten material importado ni almacenes personalizados y se rotan manualmente con alias. Las pruebas híbridas a KMS se usan TLS posclásico (X25519MLKEM768), verificable en el campo tlsDetails de CloudTrail.

Design multi-account governance and compliance solutions

- AWS Control Tower crea una zona de aterrizaje sobre AWS Organizations y aplica controles por OU: preventivos (SCP y RCP), de detección (reglas de AWS Config) y proactivos (hooks de CloudFormation). Account Factory aprovisiona cuentas estándar y Account Factory Customization añade blueprints de Service Catalog sin canalizaciones propias. Las OU recomendadas incluyen Security, Infrastructure, Workloads, Sandbox, Policy Staging y Suspended.
- AWS RAM comparte recursos como subredes, Transit Gateway, reglas de Route 53 Resolver, grupos de PrefixLists de prefijos con cuentas, OU o toda la organización. Con Organizations habilitado no hay invitaciones. En una VPC compartida el propietario controla rutas, ACL y NAT, y los participantes sus recursos y grupos de seguridad. Las conexiones compartidas a una Transit Gateway use accepter a menos que cada adjunto automático.
- Las SCP limitan los permisos máximos de las entidades principales de las cuentas miembro y las RCP los de los recursos de esas cuentas, incluso frente a identidades externas. Ninguna concede permisos, ninguna afecta a la cuenta de administración ni a los roles vinculados a servicios, y una acción debe estar permitida en todos los niveles superiores. Combinadas, por ejemplo con aws:ResourceOrgID y aws:PrincipalOrgID, forman un perímetro de datos.
- Las reglas y los paquetes de conformidad de AWS Config se implementan por región en todas las cuentas, incluidas las nuevas, y los miembros no pueden modificarlos; la corrección usa documentos de Systems Manager Automation. Security Hub CSPM evalúa estándares como PCI DSS o CIS con reglas de Config; su configuración central gestiona cuentas y regiones y las reglas de automatización y EventBridge automatizan la respuesta.

Design architectures to meet compliance requirements

- Las claves KMS administradas por el cliente permiten controlar la política y el ciclo de vida de la clave, y su uso queda en CloudTrail; las claves administradas por AWS no se pueden modificar ni compartir. En RDS no se cifra una instancia existente: se usa una implementación azul/verde o una copia cifrada de instantánea. Lake Formation aplica permisos por fila, columna y celda y audita el acceso.
- Para la soberanía de los datos, AWS KMS ofrece material de clave importado, que se puede eliminar al instante y volver a importar, y almacenes de claves externos, donde las claves se usan en un gestor propio fuera de AWS a cambio de asumir la disponibilidad. AWS Artifact entrega los informes de cumplimiento de AWS y permite aceptar acuerdos como el BAA para toda la organización desde la cuenta de administración.
- La validación de integridad de CloudTrail entrega archivos de resumen firmados que prueban si un registro cambió, y S3 Object Lock en modo de cumplimiento impide que nadie lo borre. En Bedrock, InvokeModel es un evento de administración, pero InvokeAgent y Retrieve son eventos de datos que exigen selectores avanzados. Las políticas de protección de datos de CloudWatch Logs enmascaran datos sensibles de los registros de invocaciones.
- Para la residencia de datos, Control Tower ofrece un control de denegación de regiones por OU (CT.MULTISERVICE.PV.1) y otro para toda la zona de aterrizaje. Security Lake permite regiones de consolidación por geografía, y S3 Cross-Region Replication copia objetos nuevos mientras Batch Replication copia los existentes. En Bedrock, los perfiles de inferencia geográficos mantienen el procesamiento dentro de una geografía como la UE.

Design security and responsible AI controls

- Amazon Bedrock Guardrails combina filtros de contenido, temas denegados, filtros de palabras, filtros de información confidencial (bloquear o enmascarar, con regex), comprobaciones de fundamentación contextual y Automated Reasoning. ApplyGuardrail evalúa cualquier texto sin invocar un modelo, también de fuera de Bedrock. Con políticas de Bedrock en AWS Organizations se impone una versión numérica en cuentas y OU sin cambiar código.
- AgentCore Identity gestiona identidades de agentes y credenciales. En la entrada, Runtime y Gateway aceptan IAM SigV4 o un autorizador JWT que valida audiencias, clientes y ámbitos de cualquier proveedor OAuth 2.0. En la salida, los proveedores de credenciales guardan tokens OAuth 2LO o 3LO con consentimiento y claves de API en un almacén cifrado con KMS, fuera del código del agente.
- Las aprobaciones humanas se diseñan con Step Functions y tareas .waitForTaskToken, que esperan sin cómputo hasta que el sistema de aprobación llama a SendTaskSuccess o SendTaskFailure desde la misma cuenta. La espera no tiene tiempo límite predeterminado: TimeoutSeconds y Catch de States.Timeout permiten escalar. Los flujos Express no admiten devoluciones de llamada. Los agentes de Bedrock pueden pedir confirmación al usuario.

3. Cost-Optimized Architecture Design

15–15%

Design architectures based on AWS pricing models

- Los Compute Savings Plans se aplican entre familias, regiones, Fargate y Lambda; los EC2 Instance Savings Plans, a una familia en una región; los Database Savings Plans, entre motores y servicios de bases de datos. Ni los Savings Plans ni las RI regionales reservan capacidad: para eso están las RI zonales o las reservas de capacidad bajo demanda, sin plazo si son de uso inmediato. Spot es para cargas tolerantes a interrupciones.
- Los flujos Standard de Step Functions se facturan por transición de estado y sirven para procesos largos y de exactamente una vez; los Express, por ejecuciones, duración y memoria, para trabajo corto, idempotente y de gran volumen. El tipo no se puede cambiar tras crear la máquina, y un Standard puede iniciar Express anidados. Lambda
- Compute Optimizer da recomendaciones de ajuste de tamaño para EC2, grupos de Auto Scaling, EBS, Lambda, en arm64 (Graviton) mejora el precio-rendimiento si todas las dependencias tienen versión arm64. ECS en Fargate y bases de datos, para toda la organización si se suscribe la cuenta de administración; la memoria de EC2 requiere el agente de CloudWatch. El escalado programado se combina con el seguimiento de destino.
- Amazon Bedrock ofrece los niveles Reserved, Priority, Standard y Flex (descuento para cargas que toleran más Fargate Spot avisa con SIGTERM dos minutos antes; una base en FARGATE garantiza tareas normales. latencia), elegidos con service_tier. La inferencia por lotes procesa archivos de S3 de forma asíncrona a menor precio en modelos compatibles. El caché de prompts abarata contextos largos repetidos si el prefijo estático va primero. SageMaker Asíncrono y sin servidor pueden reducir a cero las instancias.

Optimize data transfer and network costs

- CloudFront Origin Shield añade una capa de caché que consolida las solicitudes de todas las cachés hacia el origen; se habilita en la región con menor latencia hacia el origen y tiene cargos adicionales. La política de caché define la clave de caché (menos valores, más aciertos) y la de solicitudes al origen reenvía datos sin fragmentarla. Los
- Los puntos de enlace de tipo gateway para S3 y DynamoDB no tienen cargo adicional y evitan los gateways NAT, grupos de orígenes solo conmutan GET, HEAD y OPTIONS. que cobran por hora y por GB procesado; los de interfaz cobran por punto de enlace y por GB. La transferencia entre zonas de disponibilidad se cobra en ambos sentidos, así que conviene mantener en la misma zona los recursos
- La replicación de S3 filtra por prefijo o etiqueta, puede escribir las réplicas en una clase más fría y no copia los objetos que más tráfico intercambian y evitar cruzar regiones sin necesidad. existentes: para ellos se usa S3 Batch Replication. S3 RTC añade un objetivo de 15 minutos que solo compensa con RPO exigentes. Las tablas globales de DynamoDB son activo/activo, pero cobran escrituras replicadas en cada
- Transit Gateway cobra por hora de adjunto y por tráfico procesado; la interconexión de VPC no cobra procesamiento región; para RPO de horas basta AWS Backup con copia entre regiones. y es la opción más barata para pocas VPC con CIDR que no se solapan. PrivateLink cobra por hora y por GB al propietario del punto de enlace, admite CIDR solapados y conexiones solo iniciadas por el consumidor. Centralizar
- La inferencia entre regiones de Amazon Bedrock usa perfiles de inferencia: los geográficos (UE, EE. UU., APAC) puntos de enlace de interfaz detrás de Transit Gateway exige una zona alojada privada propia. mantienen el procesamiento dentro de esa geografía y los globales enrutan a cualquier región comercial. No hay coste adicional de enrutamiento, se factura según la región de origen y los datos viajan por la red de AWS. Las SCP

Design cost allocation and visibility solutions

bloquean invocación de destinos o usar bedrock:InferenceProfileArn.

- Las etiquetas de asignación de costes solo aparecen en Cost Explorer y en los informes después de activarlas desde la cuenta de administración, y no son retroactivas salvo que se pida un relleno (backfill). Las etiquetas de cuenta de AWS Organizations se aplican a todo el uso de la cuenta, incluidos créditos y reembolsos, y se combinan
- AWS Data Exports entrega en Amazon S3 el CUR 2.0, de esquema fijo, con ID de recurso, granularidad por hora con las etiquetas de recurso para repartir por unidad, proyecto y entorno. y filtros SQL de filas y columnas; con la integración de Athena se entrega en Parquet con una plantilla que crea el catálogo de Glue. Cost Explorer analiza costes sin código, y la cuenta de administración habilita en sus preferencias
- Las políticas de etiquetas de AWS Organizations normalizan la capitalización y los valores, y con enforced_for los datos por hora y por recurso de los últimos días. impiden etiquetados no conformes, pero no exigen claves ausentes; para eso están las SCP con aws:RequestTag y Null o las claves obligatorias con el hook AWS::TagPolicies::TaggingComplianceValidator en CloudFormation. La
- Las categorías de costes agrupan cuentas, etiquetas y servicios en valores propios y sus reglas de reparto regla required-tags de AWS Config detecta recursos sin etiquetas, pero no impide crearlos. (proporcional, fijo o a partes iguales) distribuyen los costes compartidos, con resultados solo en su página de detalles. Los datos de asignación divididos llevan al CUR el coste por tarea de ECS o pod de EKS. En Bedrock, las
- AWS Budgets avisa en umbrales reales o previstos y puede ejecutar acciones automáticas o con aprobación, como etiquetas de la entidad principal de IAM o de un perfil de inferencia de aplicación imputan costes y tokens. aplicar una política de IAM o, desde la cuenta de administración, una SCP a otra cuenta. AWS Cost Anomaly Detection usa aprendizaje automático para detectar gasto inusual; sus monitores administrados por AWS incluyen todas las cuentas y las alertas combinan umbrales absolutos y porcentuales.

Optimize storage and database costs

- S3 Intelligent-Tiering mueve los objetos entre niveles según su acceso, sin tarifas de recuperación, y sus niveles Archive opcionales son asíncronos; no supervisa objetos de menos de 128 KB. S3 Lifecycle pasa objetos a clases más baratas, pero cada transición se cobra y los archivados llevan metadatos extra, así que conviene agrupar
- En gp2 las IOPS de referencia crecen con el tamaño y dependen de créditos de ráfaga; gp3 cuesta menos por objetos pequeños. En EFS, las políticas de ciclo de vida se aplican a todo el sistema de archivos. GiB y permite aprovisionar IOPS y rendimiento sin añadir capacidad. io2 Block Express ofrece latencia inferior al milisegundo y la mayor durabilidad para bases de datos críticas. Elastic Volumes cambia tipo, IOPS o tamaño (solo
- DynamoDB bajo demanda cobra por solicitud y escala sola; el modo aprovisionado conviene con cargas estables al alza) sin desconectar, y Compute Optimizer recomienda tipo, tamaño e IOPS según el uso. y previsibles. Aurora I/O-Optimized elimina los cargos por E/S cuando esta pesa mucho en la factura, y Aurora Serverless v2 escala con la carga y puede pausarse con 0 ACU. Las instancias reservadas de RDS están ligadas a
- AWS Backup define la retención y puede pasar las copias de recursos compatibles (EFS, DynamoDB con motor, tipo y región, mientras que los Database Savings Plans se mantienen al cambiar de motor, familia o región. características avanzadas, EBS mediante archivado) a almacenamiento en frío, donde deben quedarse al menos 90 días; las políticas de copias de AWS Organizations aplican los planes por herencia. Amazon Data Lifecycle Manager
- El filtrado por nivel de registro de Lambda (con formato JSON) evita ingerir registros que no se necesitan. automatiza instantáneas de EBS y AMI y solo archiva instantáneas con programaciones mensuales o anuales. CloudWatch Logs guarda los datos indefinidamente salvo que se fije una retención, y la clase Essentials (antes Infrequent Access) abarata la ingesta, pero sin filtros de métricas, suscripciones, Live Tail ni EMF y sin poder cambiar la clase después. Para métricas conviene EMF y evitar dimensiones de alta cardinalidad.

4. Resilience, Migration, and Business Continuity

15–15%

Design migration strategies by using AWS frameworks

- AWS Transform sustituye a Application Discovery Service y Migration Hub, cerrados a clientes nuevos. Su herramienta de descubrimiento inventaría VMware vCenter, Hyper-V y servidores importados, y con acceso al sistema operativo recoge bases de datos y conexiones de red; también acepta RVTools, modelizeIT o MPA. La planificación agrupa aplicaciones, genera grupos de traslado con dependencias técnicas y de negocio y crea oleadas.
- AWS Transform MGN (antes AWS Application Migration Service) replica de forma continua a nivel de bloque servidores físicos, virtuales o en la nube y los lanza en EC2 con transiciones de minutos. Sus plantillas de replicación, lanzamiento y posteriores al lanzamiento permiten enviar los datos por IP privada (VPN o Direct Connect), conservar la IP privada, ajustar el tipo de instancia y automatizar acciones como instalar el agente de SSM.
- AWS DMS migra bases de datos con carga completa más CDC para mantener el destino sincronizado y reducir la parada; DMS Serverless evita dimensionar y parchear instancias de replicación, y DMS Schema Conversion evalúa y convierte esquemas entre motores distintos, pero no los datos. AWS DataSync copia NFS, SMB, HDFS u objetos a S3, EFS o FSx con verificación, programación, transferencias incrementales y puntos de enlace de VPC privados.
- Redefinir la plataforma (lift, tinker and shift) traslada la aplicación con optimizaciones, como pasar a un servicio administrado o sin servidor o de máquinas virtuales a contenedores sin cambiar el código. Amazon RDS gestiona copias, parches y recuperación, con Multi-AZ para la conmutación; AWS Fargate ejecuta contenedores de ECS sin servidores que gestionar, y Lambda con S3 y EventBridge Scheduler sustituye trabajos cortos en servidores.
- El patrón strangler fig migra un monolito a microservicios de forma gradual: una capa proxy (por ejemplo, Amazon API Gateway) enruta cada ruta al monolito o al nuevo servicio. Una capa anticorrupción traduce las llamadas internas a las nuevas interfaces sin cambiar a los llamadores, y la bandeja de salida transaccional evita eventos de transacciones revertidas.

Design disaster recovery architectures

- Las cuatro estrategias de DR van de menor a mayor coste: copia de seguridad y restauración, luz piloto (datos replicados siempre activos y servidores apagados), espera semiactiva (copia reducida en ejecución) y multisitio activo/activo. La replicación continua no protege frente a datos dañados, por lo que hay que añadir recuperación a un momento dado y copias en otra cuenta.
- AWS Elastic Disaster Recovery replica servidores de forma continua a nivel de bloque en un área de preparación de bajo coste y lanza instancias de recuperación en minutos, desde el estado más reciente o desde una instantánea anterior. Por defecto replica por Internet; la opción de IP privada usa VPN o Direct Connect. Entre regiones, se vuelve con replicación inversa.
- AWS Resilience Hub define políticas de resiliencia por niveles con RTO y RPO por tipo de interrupción, evalúa cada aplicación frente a su política y recomienda alarmas, SOP y experimentos de FIS. Los SOP se entregan como plantilla de CloudFormation con documentos de SSM Automation, y la evaluación diaria con notificación de desviaciones avisa por SNS.
- Para la conmutación entre regiones, las comprobaciones de estado de Route 53 permiten una conmutación DNS automática, Global Accelerator ofrece IP estáticas anycast y evita la caché DNS, y los controles de enrutamiento de ARC son interruptores del plano de datos para conmutar manualmente con reglas de seguridad. Region switch de ARC orquesta la recuperación multicuenta y mide el tiempo real.
- La validación de DR se automatiza ejecutando evaluaciones de AWS Resilience Hub con StartAppAssessment tras publicar cada versión y bloqueando la canalización si el resultado es PolicyBreached. Los experimentos de FIS recomendados se despliegan como plantilla de CloudFormation y se ejecutan en CI/CD. Para rotaciones regionales planificadas, la transición de Aurora Global Database tiene RPO 0.

Design solutions to meet business continuity and resiliency requirements

- Las políticas de copia de AWS Organizations se heredan de la raíz y las OU y se combinan en un plan efectivo inmutable por cuenta; solo se combinan si comparten nombre de plan. Las copias entre cuentas exigen claves administradas por el cliente para recursos no totalmente administrados por AWS Backup. Un administrador delegado gestiona políticas y supervisa trabajos sin la cuenta de administración.
- Para medir la recuperación, AWS FIS ejecuta acciones como aws:rds:failover-db-cluster y genera informes PDF con paneles de CloudWatch como prueba. Route 53 puede basar comprobaciones de estado en alarmas de CloudWatch cuando los recursos solo tienen IP privadas. El desplazamiento zonal automático de ARC exige ejecuciones de práctica con una alarma de resultado.
- Una arquitectura resiliente se recupera sola: Auto Scaling sustituye instancias que fallan las comprobaciones de ELB si se activan, y la estabilidad estática aprovisiona antes del fallo capacidad para perder una zona sin depender del plano de control. Los reintentos limitados con retroceso exponencial y jitter en una sola capa y los disyuntores evitan fallos en cascada.
- S3 Object Lock en modo de cumplimiento y el bloqueo de instantáneas de EBS en modo de cumplimiento impiden que nadie, ni el usuario raíz, borre los datos antes de que caduque la retención; el modo de gobierno puede anularse con permisos. Un almacén aislado lógicamente de AWS Backup trae Vault Lock en modo de cumplimiento y se comparte por AWS RAM con una cuenta de recuperación.
- La ingeniería del caos con AWS FIS parte de una hipótesis de estado estable que se traduce en alarmas de CloudWatch usadas como condiciones de parada. Las plantillas definen acciones, destinos y rol; la biblioteca de escenarios ofrece eventos predefinidos como AZ Availability: Power Interruption, y los experimentos multicuenta se orquestan desde una cuenta central usando ID de AZ.

5. Operational Excellence and Automation

20–20%

Design infrastructure as code

- CloudFormation es declarativo y AWS CDK permite definir la infraestructura con lenguajes de programación que se sintetizan en CloudFormation; cada entorno de destino se prepara con cdk bootstrap. StackSets con permisos administrados por el servicio despliega en las OU y en las cuentas nuevas automáticamente. El generador de IaC incorpora recursos existentes, y la detección de desviaciones revela cambios hechos fuera.
- La política como código combina capas: los aspectos de CDK validan durante la síntesis, los Hooks de CloudFormation (y los controles proactivos de Control Tower) inspeccionan los recursos antes de aprovisionarlos en modo WARN o FAIL, las SCP bloquean cualquier vía de creación y las reglas de AWS Config con corrección mediante SSM Automation corrigen desviaciones posteriores.
- Las referencias dinámicas de CloudFormation (secretsmanager y ssm-secure) sacan los secretos de las plantillas: el valor se resuelve solo al crear o actualizar el recurso, ssm-secure no lee parámetros de otras cuentas y un secreto de otra cuenta se referencia por su ARN completo con una clave de KMS administrada por el cliente. Para RDS y Aurora, ManageMasterUserPassword deja que RDS cree y rote el secreto.
- Los StackSets implementan una misma plantilla en varias cuentas y regiones. Con permisos administrados por el servicio, CloudFormation crea los roles, admite implementación automática en cuentas nuevas de una OU y puede gestionarse desde un administrador delegado, pero nunca implementa en la cuenta de administración. La tolerancia a errores y los límites de errores por región controlan el radio de impacto.

Design CI/CD and Deploy pipelines

- Las canalizaciones V2 de CodePipeline admiten disparadores filtrados por ramas, rutas de archivo y etiquetas Git, modos de ejecución QUEUED y PARALLEL además de SUPERSEDED, y condiciones de etapa (Entry, On Failure, On Success) con reglas como CloudWatchAlarm y resultados de fallo, omisión o reversión. CodeBuild publica informes de pruebas y CodeDeploy implementa en EC2 y en servidores locales registrados.
- El escaneo básico de ECR cubre vulnerabilidades del sistema operativo; el mejorado usa Amazon Inspector para escanear también paquetes de lenguajes, de forma continua, y emite eventos a EventBridge. En CodePipeline, la acción InspectorScan detiene la canalización si se superan los umbrales de gravedad y genera un SBOM. AWS Signer firma imágenes y paquetes de Lambda, y la política Enforce bloquea código no firmado.
- Una canalización entre cuentas usa una clave de KMS administrada por el cliente en su región, políticas de bucket y de clave para las cuentas de destino y roles en esas cuentas que asume el rol de servicio. Un artefacto no puede pasar entre dos cuentas si ninguna es la de la canalización. Las acciones entre regiones necesitan un bucket y una clave en cada región, en la cuenta de la canalización.
- Amazon Bedrock Evaluations ofrece evaluaciones automáticas con métricas calculadas, evaluaciones con un modelo juez (métricas integradas como Correctness o Harmfulness y métricas personalizadas), evaluaciones con trabajadores humanos propios y evaluaciones RAG de solo recuperación o de recuperación y generación. Se pueden evaluar modelos externos aportando sus respuestas en el conjunto de datos.
- En ECS, las actualizaciones graduales se protegen con el disyuntor de implementación y con alarmas de CloudWatch que revierten a la última implementación completada; las estrategias blue/green, canary y lineal desplazan tráfico con tiempos de espera y mantienen la revisión anterior para revertir. CodeDeploy ofrece configuraciones similares para el mismo fin, pero NLB solo admite todo a la vez.

Design observability and monitoring solutions

- AWS Distro for OpenTelemetry permite instrumentar una vez y enviar trazas y métricas correlacionadas a X-Ray, CloudWatch y Managed Service for Prometheus, con un recopilador sidecar en ECS y una capa administrada en Lambda. X-Ray recomienda OpenTelemetry frente a su SDK. Transaction Search ingiere todos los spans como registros, y el formato de métricas integradas extrae métricas de los registros.
- Container Insights recopila métricas y registros de ECS y EKS, también en Fargate; la observabilidad mejorada de ECS baja hasta el nivel de contenedor y se activa con un ajuste de cuenta para clústeres nuevos y con update-cluster-settings para los existentes. En EKS, el complemento CloudWatch Observability instala el agente y Fluent Bit y activa Container Insights y Application Signals. El agente también puede extraer métricas de Prometheus.
- Las suscripciones entre cuentas envían eventos a un destino de CloudWatch Logs que encapsula en flujo de Kinesis o Firehose; el destino debe estar en la misma región que el grupo de registros y su política puede permitir toda la organización con aws:PrincipalOrgID. Las reglas de centralización copian registros de toda la organización a una cuenta central. Para archivar años de registros, S3 Lifecycle los pasa a S3 Glacier Deep Archive.
- Las alarmas de CloudWatch pueden notificar a SNS, ejecutar acciones de EC2 y Auto Scaling o crear OpsItems de Systems Manager. Las alarmas compuestas combinan estados con AND, OR y NOT para reducir el ruido y admiten alarmas supresoras durante implementaciones. La detección de anomalías sigue la estacionalidad de la métrica, y el tratamiento de datos que faltan (breaching, notBreaching, ignore, missing) evita falsos positivos.
- Los endpoints de SageMaker AI publican ModelLatency y OverheadLatency para separar la latencia del modelo de la del servicio. Amazon Bedrock publica en AWS/Bedrock métricas de latencia, limitación y tokens por ModelId, y el registro de invocaciones guarda solicitudes y respuestas en destinos de la misma cuenta y región. AgentCore Observability añade métricas, sesiones, tokens, errores y trazas de agentes con ADOT.

Design operability and automation for resilient systems

- Las plantillas de solicitud de Service Quotas piden aumentos automáticamente para cada cuenta nueva de la organización, y Service Quotas crea alarmas de CloudWatch cerca del límite. El escalado predictivo prevé patrones diarios y semanales y prelanza instancias, pero no reduce capacidad. Los eventos de AWS Health llegan a EventBridge, que puede iniciar manuales de Systems Manager Automation como respuesta.
- Las remediaciones automáticas necesitan frenos: en Systems Manager Automation, la simultaneidad limita cuántos recursos se tocan a la vez y el umbral de errores detiene el envío a nuevos destinos (con 0, tras el primer fallo). El disyuntor de ECS revierte a la última implementación COMPLETED, y los reintentos deben usar retroceso exponencial con fluctuación, un máximo de intentos y una sola capa.
- El seguimiento de destino mantiene una métrica que varía en proporción a la capacidad (para colas, la acumulación por instancia con metric math). El escalado predictivo prevé patrones diarios o semanales, empieza en modo de solo previsión, puede prelanzar instancias y no reduce capacidad, por lo que se combina con una política dinámica.
- Auto Scaling ignora por defecto las comprobaciones de Elastic Load Balancing: hay que activarlas en el grupo y fijar un periodo de gracia para no terminar instancias que arrancan. Las acciones programadas cambian la capacidad mínima y máxima a horas fijas. ECS sustituye las tareas cuyo contenedor esencial falla la comprobación de la definición de tarea (no la del Dockerfile). En flujos de Kinesis, Lambda limita reintentos, divide lotes y envía lo descartado a un destino en caso de error.
- Las políticas de parches de Quick Setup aplican a OU y regiones elegidas con calendarios separados de análisis e instalación. Una regla de EventBridge puede iniciar un runbook de Automation, y aws:approve lo pausa hasta que aprueben las entidades designadas, aunque no admite ejecuciones multicuenta. Estas requieren un administrador delegado, un rol de administración central y un rol de ejecución en cada cuenta.