

Nivel: Intermediate (Associate) Duración: 130 min Nota de corte: 720/1000 Fuente: guía de estudio oficial de Microsoft Learn

1. Data Ingestion and Transformation

34–34%

Perform data ingestion

- Amazon Redshift puede leer directamente de Kinesis Data Streams o Amazon MSK hacia una vista materializada, sin pasar por S3, y AUTO REFRESH la mantiene al día. DynamoDB Streams guarda los cambios 24 horas; Kinesis Data Streams para DynamoDB permite más retención y varios consumidores. AWS Glue ofrece trabajos de ETL de streaming continuos sobre Kinesis, Kafka y MSK.
- AWS DMS ofrece carga completa, carga completa con CDC o solo cambios, y con destino S3, escribe CSV o Parquet. Redshift COPY carga en paralelo varios archivos comprimidos de tamaño similar y en Redshift, un manifiesto enumera los archivos exactos que debe cargar COPY, incluso de buckets distintos, y mandatory por prefijo común o manifiesto.
- Amazon AppFlow transfiere datos de aplicaciones SaaS como Salesforce o Zendesk a servicios de AWS sin escribir código, bajo demanda, por evento o con programación. La Amazon Redshift Data API permite ejecutar SQL mediante un punto de enlace HTTP sin controladores ni conexiones persistentes, es asíncrona y se autentica con Secrets Manager o credenciales temporales.
- Amazon EventBridge Scheduler crea programaciones cron, rate o únicas, en la zona horaria elegida y con ajuste al horario de verano, y puede llamar a APIs de AWS. Amazon MWAA ejecuta DAG

Transform and process data

- Amazon EMR on EKS ejecuta Spark y otros marcos en un clúster de EKS existente y administra sus contenedores, compartiendo recursos con otras aplicaciones. En Amazon ECS, Service Auto Scaling ajusta las tareas según CPU, memoria, horarios, patrones históricos o mensajes pendientes por tarea en SQS. Fargate Spot ejecuta tareas tolerantes a interrupciones con descuento y avisa con SIGTERM antes de detenerlas.
- Athena en paralelo. Athena ofrece controladores JDBC y ODBC para herramientas de BI y aplicaciones, y Amazon Redshift ofrece controladores JDBC y ODBC propios.
- Las integraciones zero-ETL replican Aurora, RDS, DynamoDB o aplicaciones SaaS en Redshift, sin canalización propia. En AWS Glue, la clase de ejecución flexible abarata trabajos no urgentes (pruebas, cargas puntuales) y Auto Scaling, desde Glue 3.0, añade y quita trabajadores según el paralelismo de cada etapa. En Amazon EMR, las instancias de spot convienen en los nodos de tarea, que no guardan datos de HDFS; los nodos principal y básicos deben ser bajo demanda si el clúster no puede perderse.
- Streaming y Python shell. Lambda encaja con transformaciones pequeñas activadas por eventos de S3, y en Amazon Redshift las vistas materializadas guardan resultados precalculados de combinaciones y agregaciones

Orchestrate data pipelines

- Amazon MWAA ejecuta DAG de Apache Airflow en Python sin gestionar infraestructura. Los flujos de trabajo de AWS Glue encadenan trabajos y rastreadores y pueden iniciarse por programación, bajo demanda o con eventos de EventBridge agrupados por tamaño y ventana de lote. En Step Functions, el patrón sync espera a que termine el trabajo integrado, mientras que Request Response continúa al recibir la respuesta.
- No saturar servicios posteriores. En Amazon SQS, una cola de mensajes fallidos con maxReceiveCount aísla los mensajes que fallan siempre sin detener el resto.
- Step Functions ofrece flujos estándar, de larga duración y con ejecución exactamente una vez para acciones no idempotentes, y flujos Express, de al menos una vez y para eventos de gran volumen con acciones idempotentes; el tipo no se puede cambiar tras creado. El estado Map distribuido, solo en flujos estándar, procesa en paralelo grandes conjuntos de S3 con ItemReader y ResultWriter.
- EventBridge cambia de estado de trabajos (FAILED, TIMEOUT...) y, con un umbral de retraso, eventos de

Apply programming concepts

- Para acelerar trabajos de Spark en AWS Glue conviene leer menos datos: podar particiones con push_down_predicate, seleccionar columnas y filtrar filas antes de combinar para aprovechar la delegación en Parquet, y evitar reorganizaciones con combinaciones por difusión (broadcast) cuando una tabla cabe en la memoria del ejecutor.
- ParallelizationFactor procesa varios lotes por partición manteniendo el orden por clave, y con SQS la simultaneidad máxima limita cada origen de eventos por separado.
- Las buenas prácticas de ingeniería de software también se aplican a los pipelines de datos. AWS Glue Studio se integra con Git (GitLab, GitHub, Bitbucket, CodeCommit) para versionar trabajos; la imagen Docker de AWS Glue permite probar al local con pytest sin coste de Glue: glueContext.get_logger() escribe en CloudWatch y las métricas de observabilidad muestran antes de ejecutar si un recurso se modificará, reemplazará o eliminará; la detección de desviaciones encuentra cambios hechos fuera de CloudFormation; y el atributo DeletionPolicy (Delete, Retain, RetainExceptOnCreate, Snapshot) decide qué ocurre con cada recurso al eliminar la pila.

2. Data Store Management

26–26%

Choose a data store

- El almacén correcto depende del coste y del rendimiento exigidos. MSK Serverless ejecuta Kafka sin dimensionar el clúster y Kinesis Data Streams ofrece modos bajo demanda y aprovisionado. Los nodos RA3 de Redshift escalan cómputo y almacenamiento por separado. En DynamoDB, la clase Standard-IA abarata tablas dominadas por almacenamiento y se configura según su patrón de acceso. En DynamoDB, los índices secundarios globales permiten consultar por atributos distintos de la clave. En Amazon Redshift, la distribución KEY coloca juntas las filas que se combinan, EVEN conviene a tablas sin combinaciones y las claves de ordenación aceleran filtros por rango. En cada caso de uso tiene un almacén adecuado. Aurora PostgreSQL con pgvector guarda embeddings y acelera búsquedas de vecinos aproximados con índices HNSW; Amazon MemoryDB es una base de datos en memoria duradera, compatible con Valkey y Redis OSS, con lecturas en microsegundos; Keyspaces sirve cargas de AWS Transfer Family ofrece puntos de conexión SFTP, FTPS, FTP y AS2 totalmente administrados que guardan los archivos en Amazon S3 o Amazon EFS sin cambiar los clientes de los socios. Sus flujos de trabajo administrados descifran, etiquetan, copian y llaman a Lambda tras cada subida, y los conectores SFTP recuperan o envían archivos a servidores SFTP externos. AWS DataSync migra datos de NFS, SMB, HDFS y almacenamiento de objetos. MySQL de RDS y Aurora, y Redshift Spectrum consulta archivos de Amazon S3 mediante esquemas externos sin cargarlos. Las vistas materializadas guardan resultados precalculados, se pueden actualizar automáticamente con

Understand AWS Data Catalog systems

- AWS Glue Data Catalog es un repositorio de metadatos que comparten Athena, Amazon EMR, Redshift Spectrum y los trabajos de AWS Glue. Redshift lo usa con CREATE EXTERNAL SCHEMA FROM DATA CATALOG y GRANT USAGE sobre el esquema; EMR lo configura como metastore de Hive; y Glue lee tablas con create_dynamic_frame_from_catalog y push_down_predicate para filtrar particiones. lugar de un metastore local, incluso en otra cuenta con hive.metastore.glue.catalogid. Las tablas deben apuntar a Amazon S3 y no a HDFS, que es transitorio. Athena puede leer un metastore de Hive existente con su conector para rastreadores externos, sin migrar los metadatos. después con los integrados), agrupa los archivos en tablas o particiones y escribe los metadatos en el Data Catalog. CombineCompatibleSchemas crea una sola tabla por ruta de S3 cuando los datos son compatibles, y las particiones nuevas deben registrarse en el catálogo para poder consultarlas: MSCK REPAIR TABLE añade UpdateBehavior LOG o MergeNewColumns limitan los cambios de esquema. particiones de estilo Hive pero no las elimina (para eso está ALTER TABLE DROP PARTITION), los rastreos incrementales solo añaden carpetas nuevas y los trabajos de Glue pueden registrarlas con enableUpdateCatalog. Una conexión de AWS Glue es un objeto del Data Catalog que guarda la URL, las credenciales (mejor en AWS Secrets Manager) y los datos de VPC, y la reutilizan rastreadores y trabajos. Para llegar a un almacén JDBC en una VPC, Glue crea interfaces de red que requieren un grupo de seguridad con regla autorreferenciada y DNS activado;

Manage the lifecycle of data

- COPY carga datos de Amazon S3 en Redshift en paralelo cuando hay varios archivos de tamaño similar o archivos divisibles (CSV sin comprimir, Parquet, ORC); conviene un solo COPY por tabla y un manifiesto para listar archivos concretos. UNLOAD exporta resultados a S3 en texto, JSON o Parquet, con PARTITION BY al estilo Hive, y la copia automática (COPY JOB) carga archivos nuevos sola. Intelligent-Tiering, Glacier Instant Retrieval, Flexible Retrieval y Deep Archive) y pueden filtrar por prefijo, etiquetas o tamaño. Intelligent-Tiering sirve para accesos desconocidos sin tarifas de recuperación; Glacier Instant Retrieval da acceso en milisegundos y Flexible Retrieval y Deep Archive requieren restaurar. versiones solo añade un marcador de borrado y hace falta NoncurrentVersionExpiration para eliminar las versiones antiguas. ExpiredObjectDeleteMarker quita marcadores sin versiones y AbortIncompleteMultipartUpload limpia. El control de versiones de S3 guarda todas las versiones de cada objeto: un DELETE simple solo añade un marcador de borrado, el bucket no puede volver a sin versiones (solo suspenderse) y MFA delete protege los borrados de versiones. El TTL de DynamoDB borra en segundo plano, sin consumir escritura, los elementos cuyo atributo Number en segundos de época ha caducado, y esos borrados aparecen en Streams. de escritura; en S3 con control de versiones, Expiration solo añade un marcador de eliminación y hace falta NoncurrentVersionExpiration para borrar versiones antiguas. Object Lock impide el borrado: la retención legal dura

Design data models and schema evolution

- En Redshift, la distribución KEY sobre la columna de combinación coloca juntas las filas que se combinan, y el tipo SUPER guarda datos semiestructurados que se consultan con PartiQL. En DynamoDB, una clave de ordenación cuando cambia la forma de los datos: AWS Glue Schema Registry valida versiones con modos de compatibilidad (BACKWARD, FORWARD, FULL...); un rastreador de Glue con UpdateBehavior en LOG no toca el esquema existente; y en Athena, Parquet (por nombre) y Avro admiten eliminar y reordenar columnas, a diferencia de CSV y TSV. DMS Schema Conversion es la opción gestionada y web para convertir esquemas a Aurora, RDS o Redshift: evalúa qué se convierte solo, convierte y aplica el código, y añade paquetes de extensión. AWS SCT, que se instala en SageMaker. Más imágenes y código registran el linaje de los flujos de datos y se consultan automáticamente (trabajo AWS DMS de datos, modelos y puntos de conexión) y se consulta con QueryLineage. El linaje de datos de SageMaker Catalog es compatible con OpenLineage: se captura desde conexiones de Glue y Redshift, desde Spark o con la API PostLineageEvent. En Athena conviene particionar según las consultas más habituales, usar claves de partición STRING, formatos columnares como Parquet y la proyección de particiones en tablas muy particionadas para evitar llamadas a GetPartitions. En Redshift, ENCODE AUTO gestiona la compresión salvo que se fije la codificación de alguna columna.

3. Data Operations and Support

22-22%

Automate data processing by using AWS services

- Amazon MWAA ejecuta DAG de Apache Airflow como servicio gestionado con escalado de trabajadores. AWS Step Functions orquesta servicios con patrones de integración: Request Response, Run a Job (.sync), que espera a que termine el trabajo, y .waitForTaskToken; el estado Map distribuido procesa grandes conjuntos de S3 con ejecuciones
- Para depurar Amazon MWAA se activan los grupos de registros de Airflow en CloudWatch Logs (programador, hijo, procesamiento de DAG, tareas, trabajadores y servidor web); las colas largas se resuelven escalando tareas o subiendo el mínimo de trabajadores. En Step Functions, Retry y Catch gestionan errores, pero States.Runtime no se captura con States.ALL.
- Cada servicio aporta funciones de procesamiento para marcos de trabajos de AWS Glue: evita procesar datos antiguos; Amazon EMR Serverless ejecuta Spark o Hive sin gestionar clústeres y libera recursos al terminar; y las vistas materializadas de Redshift guardan resultados precalculados con actualización automática y reescritura de consultas.
- Para consumir API de datos sin gestionar infraestructura: la Redshift Data API ejecuta SQL de forma asíncrona por HTTP, sin controladores ni conexiones y con Secrets Manager o credenciales temporales; Amazon AppFlow trae datos de aplicaciones SaaS a S3 o Redshift sin código, y su Custom Connector SDK cubre API privadas.

Analyze data by using AWS services

- Amazon Quick Sight, la función de BI de Amazon Quick, visualiza datos de orígenes como Athena o Redshift por consulta directa o importándolos en SPICE, su motor en memoria, que acelera las consultas y permite reutilizar los datos sin coste adicional. SPICE admite actualizaciones programadas e incrementales con ventana retrospectiva.
- Para verificar y limpiar datos, Athena permite consultar con SQL y localizar el archivo de S3 de cada fila con " \$path". AWS Glue DataBrew genera perfiles visuales con distribuciones, correlaciones y duplicados.
- SageMaker Data Wrangler, ahora integrado en SageMaker Canvas, genera un informe de calidad con valores que faltan, no válidos, atípicos y avisos como la fuga de la variable objetivo, y ofrece transformaciones para imputar o eliminar valores que faltan y recortar o eliminar atípicos.
- En Amazon Redshift, una vista estándar ejecuta su consulta en cada referencia; con WITH NO SCHEMA BINDING se crea una vista de enlace tardío, necesaria para tablas externas, y una vista materializada guarda resultados precalculados que se actualizan con REFRESH MATERIALIZED VIEW. En Athena, CREATE VIEW crea vistas
- Principio de Athena para CREATE SPARK ETL para el motor de procesamiento de datos de Redshift, de datos compartidos con Redshift hasta los máximos indicados. Requiere un grupo de trabajo con motor Apache Spark, cuyo motor no se puede cambiar después; al abrir un cuaderno se inicia una sesión con parámetros como DPU simultáneas máximas
- Los servicios sin servidor (Redshift Serverless, EMR Serverless, MSK Serverless) aprovisionan y escalan la capacidad automáticamente y cobran por uso, sin tipos de nodo, redimensionado ni pausas manuales. Los aprovisionados dan más control a cambio de gestionar la capacidad. Hay matices: la capacidad preinicializada de Amazon SageMaker Data Wrangler está inactiva, y MSK Serverless exige IAM y no admite ACL de Kafka.

Monitor and monitor data pipelines

- Para extraer registros de auditoría, Athena consulta con SQL los archivos de CloudTrail guardados en S3, mientras que el historial de eventos solo cubre 90 días de eventos de administración. CloudWatch Logs Insights consulta de forma interactiva los grupos de registros, y una tarea de exportación copia un intervalo concreto a S3, incluso a otra cuenta o a un bucket con Object Lock.
- Para la trazabilidad, CloudTrail registra eventos de administración y, si se configuran, eventos de datos como GetObject en S3, filtrables con selectores avanzados; la validación de integridad con archivos de resumen firmados demuestra si un registro se alteró. Redshift ofrece registros de conexiones, usuarios y actividad de usuarios hacia
- Las alertas de una canalización se envían normalmente a temas de Amazon SNS. Las reglas de EventBridge capturan eventos de servicio, como Glue Job State Change en FAILED o TIMEOUT, o Glue Job Run Status cuando una ejecución supera el umbral de notificación de retraso. Las alarmas de CloudWatch notifican al cambiar de
- En AWS Glue, las métricas de memoria y la Spark UI ayudan a localizar el cuello de botella: un driver sin memoria estado, y las alarmas compuestas solo pasan a ALARM cuando se cumple su regla, lo que reduce el ruido.
- Al leer muchos archivos pequeños se corrige con la agrupación (group-files), y Auto Scaling ajusta los workers a cada etapa. En Athena, la proyección de particiones evita buscar miles de particiones en el catálogo y elimina su mantenimiento.
- CloudTrail registra de serie solo eventos de administración, visibles en el historial de eventos durante un periodo limitado; para un registro continuo hay que crear un trail, preferiblemente multirregión. Los eventos de datos, como DeleteObject en S3, deben activarse expresamente. CloudTrail Insights detecta tasas anómalas de llamadas o

Ensure data quality

- AWS Glue Data Quality comprueba los datos en tránsito dentro de los trabajos ETL con la transformación Evaluate Data Quality y reglas DQDL como IsComplete (valores no nulos) o IsUnique. Puede detener el trabajo sin cargar el destino si falla el conjunto de reglas, o añadir columnas por fila (DataQualityEvaluationResult) para separar los
- En AWS Glue DataBrew, las reglas de calidad se agrupan en conjuntos de reglas que comparan métricas (valores registros erróneos en una ubicación de cuarentena, ausentes, duplicados, rangos) con valores esperados y se validan en trabajos de perfil, que generan un informe de validación. Una regla puede aplicarse a varias columnas mediante selectores por nombre o expresión regular, y
- Los trabajos de perfil de AWS Glue DataBrew ayudan a investigar la coherencia de los datos: a nivel de conjunto cada validación emite un evento de EventBridge con su resultado.
- calculan filas duplicadas y la correlación de Pearson entre columnas numéricas, y a nivel de columna cuentan valores ausentes (nulos y cadenas vacías), distintos y únicos, y detectan valores atípicos con la puntuación Z. La
- El muestreo permite analizar datos grandes con menos coste: los proyectos de DataBrew trabajan sobre una configuración permite elegir columnas por nombre o expresión regular.
- muestra de las primeras filas, las últimas o filas aleatorias, y los trabajos de perfil usan por defecto un número fijo de filas salvo que se elija el conjunto completo. En Athena, TABLESAMPLE BERNOULLI elige cada fila de forma
- El sesgo de datos concentra el trabajo en pocas particiones. En Spark sobre AWS Glue se detecta en la Spark UI independiente y SYSTEM toma segmentos lógicos completos.
- y se mitiga con Adaptive Query Execution y claves de combinación de alta cardinalidad. En Kinesis Data Streams, una clave de partición con muchos valores reparte las escrituras entre particiones, y en Redshift una clave de distribución de alta cardinalidad evita segmentos sobrecargados.

