PROFESSIONAL DATA ENGINEER · CHEAT SHEET · CERTIFYDECK

Google Cloud Certified Professional Data Engineer

Dominios, pesos oficiales y habilidades clave del temario de Professional Data Engineer, para repasar antes del examen.

Nivel: Advanced (Professional) Duración: 120 min Fuente: guía de estudio oficial de Microsoft Learn

1. Designing data processing systems

22–22%

Designing for security and compliance

- Concede roles predefinidos en el recurso más bajo posible: BigQuery Data Viewer en el conjunto de datos y Job User en el proyecto, y usa cuentas de servicio dedicadas para los workers de Dataflow. Las condiciones de IAM limitan el acceso en el tiempo, y las políticas de organización, heredadas por todos los descendientes, imponen límites como el uso compartido restringido por dominio o la prevención del acceso público.
- Google cifra los datos en reposo por defecto; con CMEK controlas en Cloud KMS las claves de encriptación de claves, en la misma ubicación que el conjunto de datos, y con Cloud EKM el material de clave se queda en un gestor externo. Autokey crea claves bajo demanda, gcp.restrictNonCmekServices obliga a usar CMEK, y rotar una clave no vuelve a cifrar las tablas existentes de BigQuery.
- El descubrimiento de Sensitive Data Protection localiza la PII de forma continua y la desidentificación la transforma: AES-SIV y el cifrado con conservación de formato son reversibles con la clave, HMAC-SHA-256 no lo es, y el desplazamiento de fechas mantiene los intervalos. En BigQuery, el enmascaramiento dinámico con etiquetas de política muestra valores enmascarados a Masked Reader sin cambiar las consultas.
- La restricción gcp.resourceLocations limita dónde se crean los recursos nuevos, y la ubicación de un conjunto de datos de BigQuery no cambia tras crearlo. Las políticas de almacenamiento de mensajes de Pub/Sub, la región del trabajo de Dataflow y los buckets de registros regionales mantienen los datos en una región; BigQuery Omni consulta datos de S3 sin sacarlos de AWS.
- Para retención normativa, Bucket Lock y Object Retention Lock en Cloud Storage y los buckets de registros bloqueados impiden borrar o acortar la retención. Los registros de auditoría de acceso a datos están desactivados por defecto salvo en BigQuery; Access Transparency y Access Approval registran y aprueban el acceso del personal de Google, y Assured Workloads aplica paquetes de controles como HIPAA.

Designing for reliability and fidelity

- Para limpiar datos ya cargados en BigQuery, Dataform gestiona transformaciones SQLX con Git y árbol de dependencias, y BigQuery data preparation aplica sugerencias de Gemini con pasos de deduplicación y validación hacia una tabla de errores. Antes de la carga, Wrangler de Cloud Data Fusion limpia de forma visual y Dataflow envía las filas fallidas a una tabla dead-letter. Gemini in BigQuery genera SQL a partir de lenguaje natural.
- Managed Airflow orquesta DAG de Python que abarcan nubes y centros locales; Workflows es la opción sin servidor y sin coste en reposo para pocos pasos, y Dataform programa sus ejecuciones con configuraciones de flujo de trabajo. En Airflow, los sensores en modo reschedule liberan el worker. Dataflow se supervisa con su interfaz de trabajos y con alertas de Cloud Monitoring sobre las métricas Failed y System lag.
- La recuperación ante desastres gestionada de BigQuery ofrece conmutación suave, sin pérdida de datos, y forzada, que puede perder datos no replicados; las consultas programadas deben recrearse tras la conmutación. Cloud SQL combina alta disponibilidad entre zonas con réplicas entre regiones que se promueven, Spanner birregional da RPO cero en un país y los buckets birregionales con replicación turbo sirven datos sin cambiar rutas.
- Spanner ofrece transacciones de lectura y escritura atómicas entre tablas con coherencia externa, y de solo lectura sin bloqueos. BigQuery admite transacciones de varias sentencias entre tablas, pero con bloqueo a nivel de tabla y sin cola. Bigtable solo es atómico por fila; con varios clústeres es eventualmente coherente, y la lectura de las propias escrituras exige enrutar al mismo clúster.
- Dataform valida cada ejecución con aserciones (nonNull, uniqueKey, rowConditions o SQLX manual) y prueba la lógica con pruebas unitarias sobre entradas simuladas. La calidad de datos automática de Knowledge Catalog valida tablas sin infraestructura con reglas integradas y SQL personalizadas, ámbito incremental, puntuaciones, avisos por correo, exportación a BigQuery y publicación de resultados en el catálogo.

Designing for flexibility and portability

- Conviene elegir servicios que admitan los requisitos futuros sin rediseño: Datastream para pasar de exportaciones nocturnas a CDC casi en tiempo real, Managed Service for Apache Kafka para conservar los clientes de Kafka, Spark sin servidor para pagar solo al ejecutar, Dataflow para usar la misma lógica en lotes y streaming, consultas continuas para reaccionar con SQL y nuevas suscripciones de Pub/Sub para añadir consumidores.
- BigQuery Omni consulta datos de S3 o Blob Storage con BigLake tables donde residen y puede escribir los resultados en la otra nube. Las BigLake tables delegan el acceso y aplican seguridad de fila y columna también a Spark; las Iceberg managed tables guardan una copia abierta en tu bucket con DML y streaming. Beam en Dataflow cambia de runner sin reescribir, y gcp.resourceLocations limita dónde se crean los recursos.
- Knowledge Catalog (antes Dataplex Universal Catalog) descubre archivos de Cloud Storage con análisis de descubrimiento que publican BigLake tables si indicas una conexión, perfila tablas con análisis de perfil (completos o incrementales) y organiza metadatos con aspectos y tipos de entrada. Su búsqueda combina lenguaje natural y palabras clave y respeta los permisos; en una malla de datos, cada dominio tiene zonas raw y curated.

Designing data migrations

- Para planificar una migración, primero se inventarían cargas, dependencias, equipos y requisitos; en un almacén de datos, la evaluación de migración de BigQuery analiza metadatos y registros de consultas y propone el estado de destino y el plan. Se eligen como primeras cargas las poco críticas y poco acopladas, y la estrategia de cada carga depende de cuánta inactividad tolera.
- Cada herramienta tiene su caso: BigQuery Data Transfer Service carga almacenes como Teradata o Redshift en BigQuery, también de forma incremental; Database Migration Service migra a Cloud SQL y AlloyDB con migraciones continuas; Datastream replica cambios por CDC hacia BigQuery con conectividad privada por VPN o Interconnect; Transfer Appliance cubre la falta de ancho de banda y la herramienta de validación compara origen y destino.

2. Ingesting and processing the data

25–25%

Planning the data pipelines

- Pub/Sub ofrece suscripciones de exportación a BigQuery, Cloud Storage y Bigtable cuando los mensajes no necesitan procesamiento, y temas de importación para orígenes externos como Amazon Kinesis Data Streams.
- Las transformaciones ligeras por mensaje caben en Single Message Transforms de Pub/Sub, el ELT en SQL dentro de BigQuery, con dependencias, aserciones y tablas incrementales, se hace con Dataform, y Cloud Data Fusion ofrece canalizaciones visuales sobre clusters efímeros de Spark. Para orquestar, Managed Airflow encaja con ETL/ELT por DAG con rellenos y workflows con servicios HTTP sin servidor y baja latencia.
- Google cifra los datos en reposo y en tránsito por defecto; con CMEK controlas las claves en Cloud KMS, en hardware con Cloud HSM o fuera de Google con Cloud EKM. Dataflow exige claves regionales en su región y Pub/Sub recomienda claves globales; en BigQuery la clave predeterminada del dataset solo afecta a tablas nuevas y una política de organización puede exigir CMEK. El enlace de Interconnect se cifra con HA VPN o MACsec.

Building the pipelines

- Si hace falta transformar o entregar exactamente una vez, se usa Dataflow con plantillas de Google o managed I/O para Kafka. Iceberg y BigQuery: Managed Service for Apache Kafka encaja cuando la API de Kafka debe ser de BigQuery, con dependencias, aserciones y tablas incrementales, se hace con Dataform, y Cloud Data Fusion
- Los workers de Dataflow sin IP externas llegan a las API de Google con Private Google Access o Private Service Connect; Cloud NAT solo hace falta para servicios externos. Sin Dataflow Shuffle ni Streaming Engine, una regla de entrada con la etiqueta de red de Dataflow debe permitir los puertos 12345 y 12346. En VPC compartida se usa la URL completa de la subred, y VPC Service Controls crea perímetros contra la exfiltración.
- En Dataflow, los registros que no se pueden analizar se envían a una salida etiquetada y a un destino de mensajes no procesables, sin usar dead-letter topics de Pub/Sub. En BigQuery, SAFE_CAST convierte los errores en NULL; las aserciones de Dataform (nonNull, uniqueKey) validan cada ejecución y Wrangler de Cloud Data Fusion limpia datos por lotes de forma visual.
- Dataflow agrupa los flujos con ventanas fijas, deslizantes o de sesión; la marca de agua marca los datos tardíos, que se descartan salvo que se configure un retraso permitido, y los activadores con modo de acumulación o descarte deciden qué emite cada panel. Para enriquecer, Beam ofrece Enrichment con BigtableEnrichmentHandler y BigQuery ofrece AI.GENERATE_TEXT sobre un modelo remoto Gemini.
- BigQuery Data Transfer Service programa cargas gestionadas en BigQuery desde Amazon S3, Google Ads y otras fuentes, con recargas. Storage Transfer Service copia archivos entre nubes, local y Cloud Storage; Transfer Appliance traslada sin conexión cuando la red limita. Datastream replica cambios de bases de datos casi en tiempo real, y los trabajos de carga son la opción más barata para datos que cambian poco.

Deploying and operationalizing the pipelines

- Los temas de importación de Pub/Sub ingieren de forma continua desde Amazon Kinesis o Cloud Storage, y los esquemas Avro o Protocol Buffer de un tema rechazan los mensajes que no cumplen. BigQuery consulta bases operativas con EXTERNAL_QUERY y datos de S3 con BigQuery Omni y tablas de BigLake sin copiarlos. Kafka Connect conecta clusters con MirrorMaker 2.0 y Pub/Sub, y el Hub de Cloud Data Fusion ofrece plugins.
- Workflows orquesta servicios HTTP y conectores sin servidor, con latencia de milisegundos y pago por paso; Managed Airflow (antes Cloud Composer) orquesta flujos ETL por lotes en DAG de Python con operadores, sensores, reintentos, operadores aplazables y disparo por eventos mediante la API REST. Dataform programa sus ejecuciones con configuraciones de flujo de trabajo.
- Cloud Build ejecuta las pruebas en cada pull request y despliega los artefactos. En Dataflow, las Flex Templates empaquetan la canalización como imagen de Docker en Artifact Registry y graph_validate_only valida una actualización sin detener el trabajo. Dataform se conecta a Git, con Developer Connect para redes privadas, y usa configuraciones de versión por entorno con Git commitish y anulación de proyecto.

3. Storing the data

20–20%

Selecting storage systems

- Analiza primero cómo se leerán y escribirán los datos: en Bigtable la clave de fila se diseña para lecturas por clave, prefijo o rango y evitando claves que empiecen por marca de tiempo; en BigQuery, particiones por fecha más clustering por columnas de alta cardinalidad, índices de búsqueda para localizar filas y la Storage Read API para lecturas masivas en paralelo; en Cloud Storage, Autoclass si el acceso es impredecible.
- Elige el servicio por el patrón de acceso y la consistencia: Spanner para SQL relacional con consistencia externa entre regiones; AlloyDB para PostgreSQL con analítica sobre datos transaccionales en vivo; Cloud SQL para MySQL, PostgreSQL o SQL Server gestionados; Bigtable para datos con una sola clave a gran escala; Firestore para documentos (también con compatibilidad MongoDB); Memorystore para caché en memoria; BigQuery para cómputo. En BigQuery, las particiones permiten el almacenamiento a largo plazo por partición, y con facturación física conviene cargar de forma incremental en vez de sobrescribir.
- En Cloud Storage, la clase debe seguir la frecuencia de acceso (Standard para datos calientes, Nearline mensual, Coldline trimestral, Archive menos de una vez al año) y los datos calientes deben estar en la misma región que el cómputo. Object Lifecycle Management aplica reglas Delete y SetStorageClass según la antigüedad, las versiones no actuales u otras condiciones; una política de retención bloqueada con Bucket Lock impide eliminar antes de tiempo.
- En BigQuery, la caducidad de partición o de tabla borra datos automáticamente, y time travel permite restaurar tablas eliminadas dentro de su ventana, tras la cual empieza el periodo de fail-safe.

Planning for using a data warehouse

- Al migrar, transfiere primero el esquema sin cambios y optimízalo después con clustering y campos anidados y repetidos (ARRAY de STRUCT) para relaciones padre-hijo que se consultan juntas. Las vistas de fachada aíslan a los consumidores de los cambios de esquema, las claves INT64 abaratan los joins y las claves principales y externas NOT ENFORCED permiten al optimizador eliminar joins.
- En BigQuery conviene desnormalizar, porque el coste de almacenamiento extra compensa la mejora de rendimiento, preferiblemente con campos anidados y repetidos en vez de aplanar del todo, que provoca shuffling; los esquemas en estrella ya suelen estar optimizados. Las actualizaciones de una sola fila van a una base de datos OLTP como Cloud SQL, y, en Bigtable cada entidad se guarda en una sola fila.
- Traduce cada requisito de negocio en una decisión concreta: residencia en un país, una región de BigQuery de ese país y no una multirregión; frescura acotada en tablas con CDC, max_staleness; aislamiento y atribución de costes, reservas con asignaciones, formato abierto con DML, Apache Iceberg managed tables, copias de prueba baratas, clones, copias de auditoría, instantáneas.
- La arquitectura del almacén se adapta a cada patrón de acceso: vistas materializadas incrementales para agregaciones repetidas y siempre actualizadas, BI Engine con tablas preferidas y particiones para paneles, consultas continuas con EXPORT DATA a Pub/Sub para procesos dirigidos por eventos, ETL inverso a Bigtable para búsquedas de baja latencia, y tablas de destino para no repetir transformaciones caras.

Using a data lake

- Para gestionar el lago, los análisis de descubrimiento de Knowledge Catalog catalogan los archivos de Cloud Storage y los registran como BigLake tables; el acceso se delega en la conexión de la tabla y los analistas no deben leer los objetos directamente. Los costes se controlan con cuotas personalizadas de consulta por proyecto o usuario y con Requester Pays cuando terceros descargan los datos; las zonas raw y curated separan datos por estabilidad.
- Para procesar un data lake en Cloud Storage, las tablas Apache Iceberg gestionadas por BigQuery admiten DML y streaming en tu propio bucket, mientras que las BigLake tables son de solo lectura y las tablas de objetos permiten inferencia sobre datos no estructurados. La caché de metadatos evita listar millones de archivos, y Managed Service for Apache Spark sin servidor ejecuta PySpark sin gestionar clústeres.
- Para supervisar un data lake, Knowledge Catalog ofrece análisis de perfil de datos, análisis de calidad automática con reglas, puntuaciones y avisos por correo, y linaje de datos para rastrear errores e impacto. En Cloud Storage, Storage Insights da visibilidad de toda la organización, los registros de acceso a datos se habilitan explícitamente y Pub/Sub notifica los cambios de objetos.

Designing for a data platform

- Knowledge Catalog (antes Dataplex Universal Catalog) es la capa de metadatos de la plataforma: cataloga automáticamente BigQuery, registra datos de Cloud Storage como BigLake tables o tablas de objetos mediante análisis de descubrimiento, importa metadatos externos con conectividad gestionada y organiza el contexto con grupos de entradas, tipos de entrada, aspectos y glosarios. Los agentes de IA lo consultan con su servidor MCP remoto.
- En una malla de datos, los dominios son propietarios de sus datos y los publican como productos, mientras equipos centrales definen la gobernanza y ofrecen una plataforma de autoservicio con plantillas de IaC seguras. En Knowledge Catalog, un lake representa un dominio, las zonas sin procesar y depuradas organizan a los equipos y los productos de datos añaden grupos de acceso y contratos. Las vistas autorizadas y BigQuery sharing evitan copias.

4. Preparing and using data for analysis

15–15%

Preparing data for visualization

- BigQuery se conecta con Connected Sheets para analizar tablas grandes en hojas de cálculo, con Data Studio mediante su conector y credenciales del propietario, del lector o de una cuenta de servicio, con Looker mediante conexiones y LookML, con otras herramientas mediante controladores JDBC y ODBC que usan la Storage Read API, con CMEK integrada, y con Dataform programada para mantener los resultados actualizados.
- BI Engine es un servicio de análisis en memoria que acelera las consultas de las herramientas de BI sin cambiarlas; se reserva por región y las tablas preferidas limitan la aceleración a lo importante. Las vistas materializadas incrementales precalculan agregados, devuelven datos actualizados y permiten el ajuste inteligente, y la opción de fecha y hora reduce coste cuando se acepta datos algo antiguos.
- En Looker, las etiquetas de acceso a datos y los atributos de usuario filtran qué filas ve cada grupo, y Sensitive Data Protection genera perfiles para localizar las columnas sensibles antes de etiquetarlas.
- Las tablas de objetos exponen en BigQuery archivos no estructurados de Cloud Storage con delegación de acceso y seguridad a nivel de fila. Para RAG, AI.CHUNK, DOCUMENT divide los PDF, AI.EMBED o AI.GENERATE_EMBEDDING crean los embeddings y VECTOR_SEARCH recupera el contexto para AI.GENERATE. Consultas y documentos deben usar el mismo modelo de embedding, y el prompt requiere columnas almacenadas en el agente.
- Para publicar datos sin copiarlos, BigQuery sharing ofrece intercambios privados o públicos (allAuthenticatedUsers con Analytics Hub Viewer y Subscriber), fichas comerciales con Cloud Marketplace y temas de Pub/Sub compartidos. En Data Studio, las credenciales del lector leen los datos. Owner's Credentials usa la autorización del propietario. En la organización, los conjuntos de datos autorizados publican vistas y los productos de datos de Knowledge Catalog añaden contratos y solicitudes de acceso. Requester Pays traslada los cargos de acceso en Cloud Storage. BigQuery sharing (antes Analytics Hub) comparte conjuntos de datos y temas de Pub/Sub sin replicarlos; los publicadores crean fichas en intercambios de datos y los suscriptores obtienen conjuntos de datos vinculados de solo lectura. Los roles roles/analyticshub.* separan administración, publicación y suscripción; las opciones de salida limitan copias y exportaciones, y SCHEMATA_LINKS y SHARED_DATASET_USAGE muestran suscripciones y uso.

Preparing data for AI and ML

- BigQuery ML aplica preprocesamiento automático y permite definir el manual en la cláusula TRANSFORM, que se reaplica en ML.EVALUATE y ML.PREDICT; un modelo TRANSFORM_ONLY reutiliza las reglas por lotes. ML.FEATURES_AT_TIME y ML.ENTITY_FEATURES_AT_TIME evitan la fuga de datos, y Feature Store on Gemini Enterprise Agent Platform sirve online características guardadas en BigQuery.

Sharing data

- Las reglas de análisis de BigQuery se definen en vistas (row_policy) o en las tablas (row_access_policy) y se aplican en la consulta. Las plantillas de consulta limitan la ejecución a consultas aprobadas.
- Para BigQuery, Looker y Looker Studio, las credenciales del lector leen los datos. Owner's Credentials usa la autorización del propietario. En Dataflow se fija un mínimo de workers con min_num_workers, en Managed Airflow se sube el máximo de workers si la cola crece, y en Spark se usan workers secundarios no interrumpibles. Cada uno con una plantilla o clúster.

Designing automation and repeatability

- Un DAG de Managed Airflow es un archivo de Python con la definición del DAG, los operadores y sus dependencias, declaradas con >> y <<. Cada tarea debe ser una unidad idempotente, con reintentos habilitados y pools si hay límites de API. Las ramas se unen con reglas de activación como none_failed_min_one_success, un DAG activa otro con TriggerDagRunOperator y las tareas pesadas o con GPU van a un clúster de GKE aparte. Scheduler para orquestar API HTTP sin servidor, y Managed Airflow para canalizaciones complejas. Las ejecuciones con reintentos o backfill deben ser idempotentes.

Organizing workloads based on business requirements

- BigQuery asigna ranuras a nivel de organización, carpeta o proyecto; las reservas dividen la capacidad en grupos separados para cada prioridad y la columna priority de INFORMATION_SCHEMA.JOBS registra cuál se usó. Para vigilar el uso previsto se combinan presupuestos de Cloud Billing con umbrales de gasto real y previsto, que avisan sin bloquear, y el estimador de slots, que usa el historial de uso para dimensionar reservas. La página de Monitoring ofrece métricas como el tiempo de ejecución y el número de slots para crear alertas. En BigQuery, quotaExceeded es un límite a corto plazo que se resuelve con retroceso exponencial, y quotaExceeded es un límite más largo que exige revisar la carga; INFORMATION_SCHEMA.JOBS y Cloud Audit Logs ayudan a diagnosticarlo. Las cuotas personalizadas QueryUsagePerDay y QueryUsagePerUserPerDay limitan el gasto bajo demanda, mientras que un presupuesto de Cloud Billing solo de alertas no lo limita.

Monitoring and troubleshooting processes

- La observabilidad combina varias fuentes. En BigQuery, las vistas INFORMATION_SCHEMA.JOBS recogen todos los trabajos, incluidos los fallidos, con slots y bytes; Cloud Logging guarda los registros de auditoría; y los gráficos de estado operativo y de uso de recursos resumen slots, concurrencia y errores por reserva. En Dataflow, Cloud Monitoring ofrece métricas como el retraso del sistema y el número de colas para crear alertas. En BigQuery, quotaExceeded es un límite a corto plazo que se resuelve con retroceso exponencial, y quotaExceeded es un límite más largo que exige revisar la carga; INFORMATION_SCHEMA.JOBS y Cloud Audit Logs ayudan a diagnosticarlo. Las cuotas personalizadas QueryUsagePerDay y QueryUsagePerUserPerDay limitan el gasto bajo demanda, mientras que un presupuesto de Cloud Billing solo de alertas no lo limita.

Maintaining awareness of failures and mitigating impact

- En streaming, Dataflow reintenta indefinidamente los paquetes con elementos que fallan, así que conviene una cola de mensajes fallidos con salidas ramificadas; las instantáneas guardan el estado para revertir actualizaciones.
- Dataflow usa por defecto todas las zonas de la región elegida, lo que hace los trabajos más resistentes a la falta de capacidad zonal, y conviene ejecutarlo en la región de sus orígenes y destinos. Un clúster de Managed Service for Apache Spark vive en una sola zona, incluso en alta disponibilidad. Managed Airflow ofrece entornos de alta disponibilidad y, para cargas replicadas, conmutación por región con instantáneas.
- Cloud Storage ofrece replicación dual y multirregional, BigQuery copias de tablas y Cloud SQL réplicas de lectura entre regiones, pero las réplicas entre regiones se promueven a mano para la recuperación regional. Memorystore for Redis Cluster necesita al menos una réplica por shard para ser multizona, Bigtable conmuta automáticamente con enrutamiento multiclúster y la recuperación gestionada de BigQuery ofrece conmutación forzada o suave.