ASSOCIATE DATA PRACTITIONER · CHEAT SHEET · CERTIFYDECK

Google Cloud Certified Associate Data Practitioner

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

Nivel: Intermediate (Associate) Duración: 120 min Fuente: guía de estudio oficial de Microsoft Learn

1. Data Preparation and Ingestion

30–30%

Prepare and process data

- En ETL los datos se transforman antes de llegar a BigQuery (Dataflow, Cloud Data Fusion), útil si ya existe un proceso o se quiere ahorrar recursos de BigQuery; en ELT se cargan sin procesar y se transforman dentro con SQL y Dataform, el enfoque que Google recomienda en general. ETLT combina una limpieza previa, por ejemplo de datos sensibles, con transformaciones SQL tras la carga; exportar resultados a otros sistemas es ETL inverso.
- Storage Transfer Service es el servicio gestionado para copias en línea desde Amazon S3, Azure, HDFS, sistemas de archivos locales con agentes u otros buckets, programadas o por eventos. Transfer Appliance es un dispositivo para grandes volúmenes sin ancho de banda suficiente; gcloud storage, para copias pequeñas. BigQuery Data Transfer Service carga en BigQuery datos de SaaS o Cloud Storage, y Database Migration Service migra bases de datos.
- Knowledge Catalog (antes Dataplex) ofrece análisis de perfil, con estadísticas por columna como nulos, valores distintos y más frecuentes, y análisis de calidad automática, que validan tablas de BigQuery con reglas integradas o SQL, umbrales, ámbito incremental, puntuaciones y avisos por correo. En Dataform, las aserciones fallan si su consulta devuelve filas; con SQL se pueden contar nulos y duplicados.
- BigQuery data preparation limpia tablas de BigQuery con sugerencias de Gemini, pasos de filtro, deduplicación y validación con tabla de errores; con SQL se usan SAFE_CAST, TRIM, UPPER, UPDATE y ROW_NUMBER. Wrangler de Cloud Data Fusion limpia de forma visual antes de cargar, solo en lotes; Dataflow limpia en streaming o por lotes, con plantillas y UDF o con Apache Beam, y envía filas fallidas a una tabla dead-letter.

Extract and load data into appropriate Google Cloud storage systems

- BigQuery carga datos Avro, CSV, JSON, ORC y Parquet. Avro, Parquet y ORC son autodescriptivos: BigQuery obtiene el esquema de los propios datos y se leen en paralelo aunque usen códecs de compresión internos. El CSV no admite datos anidados ni repetidos, el JSON debe ir delimitado por saltos de línea, y para CSV y JSON hay que activar la detección automática del esquema o declararlo; gzip es su única compresión admitida y frena la carga.
- BigQuery Data Transfer Service programa cargas gestionadas en BigQuery desde aplicaciones SaaS, Cloud Storage, otras nubes y almacenes de datos, sin código y con backfills. Database Migration Service migra bases de datos a Cloud SQL y AlloyDB para PostgreSQL, de forma continua o única. Dataflow ejecuta canalizaciones de Apache Beam o plantillas de Google, y Cloud Data Fusion diseña canalizaciones visuales sin código.
- BigQuery sirve para analizar con SQL grandes volúmenes de datos, y Cloud Storage guarda objetos no estructurados de cualquier formato. Cloud SQL es la base de datos relacional gestionada para MySQL, PostgreSQL y SQL Server; Spanner es relacional con coherencia transaccional a escala global; Bigtable es un almacén clave-valor de baja latencia para series temporales e IoT; Firestore es la base de datos de documentos sin servidor.
- Un bucket regional guarda los datos en varias zonas de una región y no tiene cargos de replicación; birregión y multirregión replican de forma asíncrona entre regiones y siguen sirviendo si cae una región, y la birregión permite elegir las regiones y activar la replicación turbo. En BigQuery la multirregión no da redundancia regional; Cloud SQL con alta disponibilidad protege frente a la caída de una zona, no de una región.
- Los datos estructurados siguen un esquema fijo y encajan en bases relacionales como Cloud SQL o en tablas de BigQuery; los semiestructurados, como JSON, llevan claves pero no un esquema rígido y encajan en Firestore o en columnas JSON de BigQuery. Los no estructurados (imágenes, audio, vídeo, texto libre) se guardan como objetos en Cloud Storage y se analizan desde BigQuery con tablas de objetos.

2. Data Analysis and Presentation

27–27%

Identify data trends, patterns, and insights by using BigQuery and Jupyter notebooks

- En GoogleSQL, WHERE filtra filas antes de agregar, GROUP BY con SUM, COUNT o COUNTIF resume por grupos, HAVING filtra los grupos y GROUP BY ROLLUP añade subtotales. Las funciones de ventana con OVER y PARTITION BY (RANK, ROW_NUMBER, LAG) calculan rankings y variaciones por fila, y QUALIFY filtra su resultado; para explorar una tabla conviene usar la vista previa en lugar de SELECT * con LIMIT.
- Colab Enterprise ofrece cuadernos gestionados dentro de BigQuery Studio, con acceso controlado por IAM, entornos de ejecución que se aprovisionan y se apagan solos, celdas SQL cuyo resultado se guarda como DataFrame y celdas de visualización para crear gráficos sin código. BigQuery DataFrames ejecuta código de estilo pandas en BigQuery sin mover los datos, y los cuadernos pueden programarse con resultados en Cloud Storage.
- Para responder preguntas de negocio en BigQuery, el análisis de contribución de BigQuery ML compara datos de prueba y de control para encontrar los segmentos que explican un cambio, data insights y data canvas usan Gemini para sugerir preguntas y explorar con lenguaje natural, y Connected Sheets lleva los datos a Sheets sin SQL. En SQL, SAFE_DIVIDE evita errores en ratios y SUM OVER () calcula cuotas sobre el total.

Visualize data and create dashboards in Looker given business requirements

- Los paneles definidos por el usuario de Looker se crean desde Create, una carpeta, un Look o un explore, y combinan mosaicos de consulta o vinculados a un Look. Los filtros de panel, obligatorios o vinculados, y el filtrado cruzado permiten acotar los datos; se comparten con los niveles de acceso View o Manage Access, Edit de la carpeta y se envían con Schedule delivery (Weekly, Datagroup update, Run schedule as recipient).
- Looker usa un modelo semántico en LookML, versionado con Git y con modo de desarrollo, para que el negocio consulte con definiciones gobernadas sin escribir SQL. Data Studio (antes Looker Studio) es la herramienta sin coste de informes de arrastrar y soltar; Data Studio Pro, la de pago, añade contenido propiedad de la organización en un proyecto de Google Cloud con IAM, espacios de trabajo de equipo y CMEK.
- En una vista de LookML, sql_table_name indica la tabla consultada (en BigQuery, proyecto.dataset.tabla), label cambia el nombre visible sin romper referencias y hidden: yes oculta un campo del selector sin ser una medida de seguridad. Las dimensiones usan tipos como string, number, yesno o tier y las medidas sum, count o count_distinct; los cambios se prueban en Development Mode, en una rama de Git propia.

Define, train, evaluate, and use ML models

- BigQuery ML entrena modelos con SQL sin mover los datos: regresión lineal para valores numéricos, logística para clasificar, k-means para segmentar sin etiquetas, factorización de matrices para recomendar, ARIMA_PLUS para series temporales y análisis de contribución para explicar cambios. AutoML en Gemini Enterprise Agent Platform (antes Vertex AI AutoML) crea modelos sin código, y BigQuery ML ofrece AUTOML_CLASSIFIER y AUTOML_REGRESSOR.
- Para usar un LLM preentrenado de Gemini Enterprise Agent Platform (antes Vertex AI) desde BigQuery se crea una conexión de recursos de Cloud en la misma ubicación que el dataset, se concede a su cuenta de servicio el rol Agent Platform User y se ejecuta CREATE MODEL ... REMOTE WITH CONNECTION con ENDPOINT. Después, AI.GENERATE_TEXT envía la columna prompt al modelo y devuelve result y status junto a las columnas de entrada.
- Un proyecto de ML sigue el flujo estándar: definir el resultado buscado, recopilar y preparar los datos, entrenar, evaluar y, por último, desplegar y predecir. Al preparar los datos hay que limpiar incoherencias, evitar la fuga de datos y el sesgo entre entrenamiento y servicio, y evaluar siempre con datos que el modelo no vio al entrenar.
- En BigQuery ML se entrena con CREATE MODEL, eligiendo model_type (LINEAR_REG, LOGISTIC_REG, KMEANS, ARIMA_PLUS…) e input_label_cols, y opciones como DATA_SPLIT_METHOD, AUTO_CLASS_WEIGHTS o la cláusula TRANSFORM. ML.EVALUATE sin datos devuelve las métricas calculadas al entrenar, y con una tabla evalúa datos nuevos; ML.CONFUSION_MATRIX y ML.ROC_CURVE completan la evaluación de clasificadores.
- ML.PREDICT ejecuta la predicción por lotes en BigQuery y devuelve las columnas de entrada más predicted_<etiqueta> (y _probs en clasificación), con THRESHOLD para el corte binario. ML.FORECAST sirve para ARIMA_PLUS, ML.EXPLAIN_PREDICT añade atribuciones por fila y ML.DETECT_ANOMALIES detecta anomalías; para predicción online se registra el modelo y se despliega en un endpoint.

3. Data Pipeline Orchestration

18–18%

Design and implement simple data pipelines

- La herramienta depende del requisito: Managed Service for Apache Spark para trabajos de Spark y Hadoop existentes (también serverless, sin clústeres); Dataflow para canalizaciones de Apache Beam por lotes y en streaming con plugins; Cloud Data Fusion para diseñar canalizaciones de forma visual; Dataform para transformaciones SQL versionadas dentro de BigQuery; y Managed Service for Apache Airflow para orquestar DAG.
- Para transformar datos ya cargados en BigQuery con SQL versionado y con aserciones se usa Dataform; para lotes y streaming con código Apache Beam, Dataflow (o sus plantillas) y Job Builder si no quieres programar); para canalizaciones visuales con plugins, Cloud Data Fusion; y para trabajos de Spark existentes, Managed Service for Apache Spark en clúster o serverless.

Schedule, automate, and monitor basic data processing tasks

- Las consultas programadas de BigQuery ejecutan GoogleSQL de forma recurrente con @run_date y @run_time, escriben con WRITE_TRUNCATE o WRITE_APPEND, admiten backfill, se pausan inhabilitándolas y pueden ejecutarse con una cuenta de servicio. Cloud Scheduler lanza trabajos por HTTP o Pub/Sub con entrega al menos una vez y reintentos configurables; Managed Service for Apache Airflow define schedule, catchup y retries. Es de hotel. Execution details (Stage progress, Stage workflow y Worker progress) solo para jobs; esto
- El Logs Explorer filtra entradas por consulta, gravedad e intervalo de tiempo. Las métricas basadas en registros cuentan o distribuyen valores y alimentan gráficos y alertas; las políticas de alertas basadas en registros avisan cuando aparece un mensaje. Admin Activity siempre se registra y Data Access viene activo en BigQuery, los sumideros (sinks) enrutan entradas a Pub/Sub, Cloud Storage o BigQuery, y los exclusores descartan lo que no quieres conservar.
- Una consulta programada para SQL reside en BigQuery, BigQuery y pipelines activados por eventos, Workflows encadena servicios HTTP de forma serverless con latencia de milisegundos, y los Workflow Templates de Managed Service for Apache Spark ejecutan grafos de trabajos DAG con YAML.
- El acceso unificado a nivel de bucket (recomendado) inhabilita las ACL y deja solo IAM, y es necesario para las carpetas gestionadas y las condiciones de IAM en el bucket; el acceso detallado con ACL sirve para permisos por objeto y migraciones desde Amazon S3. La prevención del acceso público anula allUsers y allAuthenticatedUsers, y las URL firmadas dan acceso temporal sin cuenta.
- Una suscripción de BigQuery escribe los mensajes de Pub/Sub en una tabla según llegan, sin suscriptor propio, cuando no hace falta procesarlos, admite esquema, metadatos, DLQ con _CHANGE_TYPE y dead-letter topic, y entrega al menos una vez. Para ventanas, agregaciones, uniones o entrega exactly-once se usa Dataflow, por ejemplo con la plantilla Pub/Sub to BigQuery.

4. Data Management

25–25%

Configure access control and governance

- El mínimo privilegio en IAM consiste en conceder el rol predefinido más limitado en el ámbito más pequeño (un conjunto de datos o un bucket en vez del proyecto), usar una cuenta de servicio por componente y evitar los roles básicos y las claves de cuenta de servicio en producción. IAM Conditions permite que el acceso caduque solo y un permiso (SERVICE.RESOURCE.VERB) permite una acción y un rol agrupa permisos. Los roles básicos dan acceso amplio a todos los servicios, en Cloud Storage solo se conceden en el proyecto y en BigQuery dan READER, WRITER u OWNER en los conjuntos de datos nuevos; los predefinidos los mantiene Google, los personalizados los mantienes tú y no cambian sin tu control.
- BigQuery Sharing (antes Analytics Hub) comparte conjuntos de datos y temas de Pub/Sub sin replicarlos: el publicador crea listings en un intercambio de datos y el suscriptor obtiene un conjunto de datos vinculado de solo lectura o una suscripción vinculada. Conviene frente a exportar copias cuando se quieren datos siempre actuales, permisos de uso o restricciones de salida; las data clean rooms añaden reglas de análisis para datos sensibles.

Configure lifecycle management

- La clase se elige por la frecuencia de acceso: Standard para datos de acceso frecuente, Nearline para una vez al mes o menos, Coldline para como mucho una vez al trimestre y Archive para menos de una vez al año, como registros normativos; todas dan acceso en milisegundos. Autoclass mueve cada objeto según su uso cuando el patrón es imprevisible y con Nearline o Archive como clase terminal.
- En Cloud Storage, una regla de Object Lifecycle Management con la acción Delete y condiciones como age o daysSinceNoncurrentTime elimina objetos automáticamente, no actúa mientras haya una retención sin cumplir. En BigQuery, la caducidad de la tabla, la de las particiones y la caducidad predeterminada del dataset eliminan datos; el valor del dataset no afecta a las tablas nuevas.
- Para archivar en Cloud Storage se elige la clase según la frecuencia de lectura: Nearline (mensual), Coldline (trimestral) o Archive (menos de una vez al año). Autoclass sirve cuando el acceso es imprevisible. Bucket Lock y las retenciones de objetos impiden eliminar datos, y BigQuery y Cloud SQL pueden exportar datos a Cloud Storage para archivarlos.

Identify high availability and disaster recovery strategies for data in Cloud Storage and Cloud SQL

- Cloud SQL ofrece copias automáticas y bajo demanda, copias finales y conservadas tras eliminar la instancia, y recuperación a un momento dado que crea una instancia nueva; las copias mejoradas usan Backup and DR Service con un backup vault central. En Cloud Storage se recomienda soft delete frente a Object Versioning, y BigQuery ofrece time travel y instantáneas de tabla.
- En Cloud SQL, la alta disponibilidad replica de forma síncrona en una standby de otra zona con conmutación automática, las réplicas de lectura, asíncronas y de solo lectura, descargan consultas, y las réplicas entre regiones sirven para DR regional y migraciones. En Cloud Storage existen la turbo replication birregional y la replicación entre buckets; en BigQuery, la replicación de datasets entre regiones.
- Un bucket regional replica los datos en varias zonas de una región, una birregión usa el par de regiones que eliges y una multirregión conjuntos de datos que elige Cloud Storage, ambas activo-activo sin cambiar rutas. En Cloud SQL, la alta disponibilidad usa una zona principal y otra secundaria dentro de la región, y las ubicaciones multirregionales sobran para copias de seguridad.

Apply security measures and ensure compliance with data privacy regulations

- El cifrado predeterminado de Google (la GMEK de la guía) cifra todo en reposo sin configuración. CMEK guarda las claves en Cloud KMS, manuales o con Autokey, para controlar rotación, acceso y registros; lo admiten BigQuery, Cloud SQL y Cloud Storage. CSEK, en Cloud Storage, aporta tu propia clave en cada petición y Google no la almacena.
- Cloud KMS crea y gestiona las claves de cifrado gestionadas por el cliente (CMEK) que protegen en reposo datos de BigQuery, Cloud Storage o Cloud SQL: tú controlas ubicación, rotación, acceso con IAM y el estado de cada versión. El agente de servicio necesita el CryptoKey Encrypter/Decrypter, el conjunto de claves debe incluir la ubicación de los datos y Autokey crea las claves al crear los recursos.
- El cifrado en reposo protege los datos almacenados en disco; Google cifra todo el contenido con AES-256 sin configuración, y CMEK o CSEK solo cambia quién controla las claves. El cifrado en tránsito protege los datos mientras viajan: TLS en el Google Front End para las API, cifrado automático del tráfico por IP interna en la VPC y, en Cloud SQL, modo SSL o Cloud SQL Auth Proxy.